Statistics 501

Introduction to Nonparametrics & Log-Linear Models

Paul Rosenbaum, 473 Jon Huntsman Hall, 8-3120

[rosenbaum@wharton.upenn.edu](mailto:rosenbaum@wharton.upenn.edu) Office Hours Tuesdays 1:30-2:30.

# BASIC STATISTICS REVIEW

# NONPARAMETRICS

Paired Data

Two-Sample Data

Anova

Correlation/Regression

Extending Methods

# LOG-LINEAR MODELS FOR DISCRETE DATA

Contingency Tables

Markov Chains

Square Tables

Incomplete Tables

Logit Models

Conditional Logit Models

Ordinal Logit Models

Latent Variables

Some abstracts

# PRACTICE EXAMS

Old Exams (There are no 2009 exams)

**Get Course Data in an R workspace**

<http://www-stat.wharton.upenn.edu/~rosenbap/index.html>

**or in a plain file if you are not using R**

<http://stat.wharton.upenn.edu/statweb/course/Spring-2008/stat501>

The one file for R is Rst501.RData It contains several data sets. Go back to the web page to get the latest version of this file.

## Get R for Free: <http://www.r-project.org/>

## Statistics Department

<http://www-stat.wharton.upenn.edu/> (Note: “www-“ not “www.”)

**Paul Rosenbaum’s Home Page**

<http://www-stat.wharton.upenn.edu/~rosenbap/index.html>

**Course Materials**: Hollander and Wolfe: *Nonparametric Statistical Methods* and Fienberg: *Analysis of Cross-Classified Categorical Data*. For R users, suggested: Maindonald and Braun *Data Analysis and Graphics Using R* and/or Dalgaard *Introductory Statistics with R*. The recommended new (2014) third edition of *Nonparametric Statistical Methods* now uses R (the second edition did not), and there is an R package for the book, NSM3, freely available from cran, and described in the R Program Index at the back of the textbook.

Common Questions

## How do I get R for Free?

## <http://www.r-project.org/>

Where is the R workspace for the course?

<http://www-stat.wharton.upenn.edu/~rosenbap/index.html>

The R workspace I just downloaded doesn’t

have the new object I need.

Sometimes, when you download a file, your web browser things you have it already, and opens the old version on your computer instead of the new version on the web. You may need to clear your web browsers cache.

I don’t want to buy an R book – I want a free introduction.

## Go to <http://www.r-project.org/>, click manuals, and take:

**An Introduction to R**

**(The R books you buy teach more)**

**I use a MAC and I can’t open the R workspace from your web page.**

Right-click on the workspace on your webpage and select "Save file/link as" and save the file onto the computer.

I want to know many R tricks.

cran.r-project.org/doc/contrib/Paradis-rdebuts\_en.pdf

## (search for this at <http://www.r-project.org/>)

Statistics Department Courses (times, rooms)

<http://www.upenn.edu/registrar/>

Final Exams (dates, rules)

<http://www.upenn.edu/registrar/>

# When does the the course start?

# When does it end? Holidays?

<http://www.upenn.edu/almanac/3yearcal.html>

Does anybody have any record of this?

<http://www.upenn.edu/registrar/>

**Grades/Cheating/Class Attendance**

**There is a take-home mid-term covering nonparametrics and a final covering categorical data**. The final may be in-class or take-home. In either case, both exams are open-book, open-notebook. Take-home exams must be your own work, with no communication with other people. **If you communicate with anyone in any way about the midterm or the final, then you have cheated on exam.** Cheating on an exam is the single stupidest thing a PhD student at Penn can do.

**Copies of old midterms and finals are at the end of this bulk pack. You should do several of each for practice**, ideally working on old exams all semester long as topics are covered. In working on old practice exams, you may work with other students. The exams involve working with data, and understanding statistical methods requires using them with data. If you want to learn the material in the course, do lots of practice exams.

**You are expected to attend class**. It is no problem at all if you miss one or two classes because of illness or family issues or transportation problems or a conference or job talk or whatever. If you miss a substantial number of classes, much more than one or two classes, then your grade in the class will be substantially reduced regardless of exam performance, and I may contact your departmental advisor to discuss your situation.

Review of Basic Statistics – Some Statistics

* The review of basic statistics is a quick review of ideas from your first course in statistics.
* n measurements: 
* **mean** (or average): 
* **order statistics** (or data sorted from smallest to largest): Sort  placing the smallest first, the largest last, and write , so the smallest value is the first order statistic, , and the largest is the nth order statistic, . If there are n=4 observations, with values , then the n=4 order statistics are .
* **median** (or middle value): If n is odd, the median is the middle order statistic – e.g.,  if n=5. If n is even, there is no middle order statistic, and the median is the average of the two order statistics closest to the middle – e.g.,  if n=4. Depth of median is  where a “half” tells you to average two order statistics – for n=5, , so the median is , but for n=4, , so the median is . The median cuts the data in half – half above, half below.
* **quartiles**: Cut the data in quarters – a quarter above the upper quartile, a quarter below the lower quartile, a quarter between the lower quartile and the median, a quarter between the median and the upper quartile. The **interquartile range** is the upper quartile minus the lower quartile.
* **boxplot**: Plots median and quartiles as a box, calls attention to extreme observations.

median

quartile



quartile



extreme



* **sample standard deviation**: square root of the typical squared deviation from the mean, sorta,



however, you don’t have to remember this ugly formula.

* **location**: if I add a constant to every data value, a measure of location goes up by the addition of that constant.
* **scale**: if I multiply every data value by a constant, a measure of scale is multiplied by that constant, but a measure of scale does not change when I add a constant to every data value.

**Check your understanding:** What happens to the mean if I drag the biggest data value to infinity? What happens to the median? To a quartile? To the interquartile range? To the standard deviation? Which of the following are measures of location, of scale or neither: median, quartile, interquartile range, mean, standard deviation? In a boxplot, what would it mean if the median is closer to the lower quartile than to the upper quartile?

Topic: Review of Basic Statistics – Probability

* **probability space**: the set of everything that can happen, . Flip two coins, dime and quarter, and the sample space is = {HH, HT, TH, TT} where HT means “head on dime, tail on quarter”, etc.
* **probability**: each element of the sample space has a probability attached, where each probability is between 0 and 1 and the total probability over the sample space is 1. If I flip two fair coins: prob(HH) = prob(HT) = prob(TH) = prob(TT) = ¼.
* **random variable**: a rule **X** that assigns a number to each element of a sample space. Flip to coins, and the number of heads is a random variable: it assigns the number **X**=2 to HH, the number **X**=1 to both HT and TH, and the number **X**=0 to TT.
* **distribution of a random variable**: The chance the random variable **X** takes on each possible value, x, written prob(**X**=x). Example: flip two fair coins, and let **X** be the number of heads; then prob(**X**=2) = ¼, prob(**X**=1) = ½, prob(**X**=0) = ¼.
* **cumulative distribution of a random variable**: The chance the random variable **X** is less than or equal to each possible value, x, written prob(**X**x). Example: flip two fair coins, and let **X** be the number of heads; then prob(**X**0) = ¼, prob(**X**1) = ¾, prob(**X**2) = 1. Tables at the back of statistics books are often cumulative distributions.
* **independence of random variables**: Captures the idea that two random variables are unrelated, that neither predicts the other. The formal definition which follows is not intuitive – you get to like it by trying many intuitive examples, like unrelated coins and taped coins, and finding the definition always works. Two random variables, **X** and **Y**, are independent if the chance that simultaneously **X**=x and **Y**=y can be found by multiplying the separate probabilities

prob(**X**=x and **Y**=y) = prob(**X**=x) prob(**Y**=y) for every choice of x,y**.**

**Check your understanding**: Can you tell exactly what happened in the sample space from the value of a random variable? Pick one: Always, sometimes, never. For people, do you think **X**=height and **Y**=weight are independent? For undergraduates, might **X**=age and **Y**=gender (1=female, 2=male) be independent? If I flip two fair coins, a dime and a quarter, so that prob(HH) = prob(HT) = prob(TH) = prob(TT) = ¼, then is it true or false that getting a head on the dime is independent of getting a head on the quarter?

Topic: Review of Basics – Expectation and Variance

* **Expectation**: The expectation of a random variable **X** is the sum of its possible values weighted by their probabilities,

****

* **Example**: I flip two fair coins, getting **X**=0 heads with probability ¼, **X**=1 head with probability ½, and **X**=2 heads with probability ¼; then the expected number of heads is , so I expect 1 head when I flip two fair coins. Might actually get 0 heads, might get 2 heads, but 1 head is what is typical, or expected, on average.
* **Variance and Standard Deviation**: The standard deviation of a random variable **X** measures how far **X** typically is from its expectation *E*(**X**). Being too high is as bad as being too low – we care about errors, and don’t care about their signs. So we look at the squared difference between **X** and *E*(**X**), namely , which is, itself, a random variable. The variance of **X** is the expected value of **D** and the standard deviation is the square root of the variance,  and .
* **Example**: I independently flip two fair coins, getting **X**=0 heads with probability ¼, **X**=1 head with probability ½, and **X**=2 heads with probability ¼. Then *E*(**X**)=1, as noted above. So  takes the value **D** =  with probability ¼, the value **D** =  with probability ½, and the value **D** =  with probability ¼. The variance of **X** is the expected value of **D** namely: var(**X**) = . So the standard deviaiton is . So when I flip two fair coins, I expect one head, but often I get 0 or 2 heads instead, and the typical deviation from what I expect is 0.707 heads. This 0.707 reflects the fact that I get exactly what I expect, namely 1 head, half the time, but I get 1 more than I expect a quarter of the time, and one less than I expect a quarter of the time.

**Check your understanding**: If a random variance has zero variance, how often does it differ from its expectation? Consider the height **X** of male adults in the US. What is a reasonable number for *E*(**X**)? Pick one: 4 feet, 5’9”, 7 feet. What is a reasonable number for *st.dev.*(**X**)? Pick one: 1 inch, 4 inches, 3 feet. If I independently flip three fair coins, what is the expected number of heads? What is the standard deviation?

Topic: Review of Basics – Normal Distribution

* **Continuous random variable**: A continuous random variable can take values with any number of decimals, like 1.2361248912. Weight measured perfectly, with all the decimals and no rounding, is a continuous random variable. Because it can take so many different values, each value winds up having probability zero. If I ask you to guess someone’s weight, not approximately to the nearest millionth of a gram, but rather exactly to all the decimals, there is no way you can guess correctly – each value with all the decimals has probability zero. But for an interval, say the nearest kilogram, there is a nonzero chance you can guess correctly. This idea is captured in by the density function.
* **Density Functions**: A density function defines probability for a continuous random variable. It attaches zero probability to every number, but positive probability to ranges (e.g., nearest kilogram). The probability that the random variable **X** takes values between 3.9 and 6.2 is the area under the density function between 3.9 and 6.2. The total area under the density function is 1.
* **Normal density**: The Normal density is the familiar “bell shaped curve”.

The standard Normal distribution has expectation zero, variance 1, standard deviation 1 = . About 2/3 of the area under the Normal density is between –1 and 1, so the probability that a standard Normal random variable takes values between –1 and 1 is about 2/3. About 95% of the area under the Normal density is between –2 and 2, so the probability that a standard Normal random variable takes values between –2 and 2 is about .95. (To be more precise, there is a 95% chance that a standard Normal random variable will be between –1.96 and 1.96.) If **X** is a standard Normal random variable, and  and  are two numbers, then  has the Normal distribution with expectation , variance  and standard deviation , which we write N(,). For example,  has expectation 3, variance 4, standard deviation 2, and is N(3,4).

* **Normal Plot**: To check whether or not data,  look like they came from a Normal distribution, we do a Normal plot. We get the order statistics – just the data sorted into order – or  and plot this ordered data against what ordered data from a standard Normal distribution should look like. The computer takes care of the details. A straight line in a Normal plot means the data look Normal. A straight line with a couple of strange points off the lines suggests a Normal with a couple of strange points (called outliers). Outliers are extremely rare if the data are truly Normal, but real data often exhibit outliers. A curve suggest data that are not Normal. Real data wiggle, so nothing is ever perfectly straight. In time, you develop an eye for Normal plots, and can distinguish wiggles from data that are not Normal.

Topic: Review of Basics – Confidence Intervals

* Let  be n independent observations from a Normal distribution with expectation  and variance . A compact way of writing this is to say  are iid from N(,). Here, iid means independent and identically distributed, that is, unrelated to each other and all having the same distribution.
* How do we know  are iid from N(,)? We don’t! But we check as best we can. We do a boxplot to check on the shape of the distribution. We do a Normal plot to see if the distribution looks Normal. Checking independence is harder, and we don’t do it as well as we would like. We do look to see if measurements from related people look more similar than measurements from unrelated people. This would indicate a violation of independence. We do look to see if measurements taken close together in time are more similar than measurements taken far apart in time. This would indicate a violation of independence. Remember that statistical methods come with a warrantee of good performance if certain assumptions are true, assumptions like  are iid from N(,). We check the assumptions to make sure we get the promised good performance of statistical methods. Using statistical methods when the assumptions are not true is like putting your CD player in washing machine – it voids the warrantee.
* To begin again, having checked every way we can, finding no problems, assume  are iid from N(,). We want to estimate the expectation . We want an interval that in most studies winds up covering the true value of . Typically we want an interval that covers  in 95% of studies, or a **95% confidence interval**. Notice that the promise is about what happens in most studies, not what happened in the current study. If you use the interval in thousands of unrelated studies, it covers  in 95% of these studies and misses in 5%. You cannot tell from your data whether this current study is one of the 95% or one of the 5%. All you can say is the interval usually works, so I have confidence in it.
* If are iid from N(,), then the confidence interval uses the sample mean, , the sample standard deviation, *s*, the sample size, *n*, and a critical value obtained from the t-distribution with *n-1* degrees of freedom, namely the value, , such that the chance a random variable with a t-distribution is above  is 0.025. If *n* is not very small, say n>10, then  is near 2. The 95% confidence interval is:

 = 

Topic: Review of Basics – Hypothesis Tests

* **Null Hypothesis**: Let  be n independent observations from a Normal distribution with expectation  and variance . We have a particular value of  in mind, say , and we want to ask if the data contradict this value. It means something special to us if  is the correct value – perhaps it means the treatment has no effect, so the treatment should be discarded. We wish to test the null hypothesis, . Is the null hypothesis plausible? Or do the data force us to abandon the null hypothesis?
* **Logic of Hypothesis Tests**: A hypothesis test has a long-winded logic, but not an unreasonable one. We say: Suppose, just for the sake of argument, not because we believe it, that the null hypothesis is true. As is always true when we suppose something for the sake of argument, what we mean is: Let’s suppose it and see if what follows logically from supposing it is believable. If not, we doubt our supposition. So suppose  is the true value after all. Is the data we got, namely , the sort of data you would usually see if the null hypothesis were true? If it is, if  are a common sort of data when the null hypothesis is true, then the null hypothesis looks sorta ok, and we *accept* it. Otherwise, if there is no way in the world you’d ever see data anything remotely like our data, , if the null hypothesis is true, then we can’t really believe the null hypothesis having seen , and we *reject* it. So the basic question is: Is data like the data we got commonly seen when the null hypothesis is true? If not, the null hypothesis has gotta go.
* **P-values or significance levels**: We measure whether the data are commonly seen when the null hypothesis is true using something called the P-value or significance level. Supposing the null hypothesis to be true, the P-value is the chance of data at least as inconsistent with the null hypothesis as the observed data. If the P-value is ½, then half the time you get data as or more inconsistent with the null hypothesis as the observed data – it happens half the time by chance – so there is no reason to doubt the null hypothesis. But if the P-value is 0.000001, then data like ours, or data more extreme than ours, would happen only one time in a million by chance if the null hypothesis were true, so you gotta being having some doubts about this null hypothesis.
* **The magic 0.05 level:** A convention is that we “reject” the null hypothesis when the P-value is less than 0.05, and in this case we say we are testing at **level** 0.05. Scientific journals and law courts often take this convention seriously. It is, however, only a convention. In particular, sensible people realize that a P-value of 0.049 is not very different from a P-value of 0.051, and both are very different from P-values of 0.00001 and 0.3. It is best to report the P-value itself, rather than just saying the null hypothesis was rejected or accepted.
* **Example**: You are playing 5-card stud poker and the dealer sits down and gets 3 royal straight flushes in a row, winning each time. The null hypothesis is that this is a fair poker game and the dealer is not cheating. Now, there are  or 2,598,960 five-card stud poker hands, and 4 of these are royal straight flushes, so the chance of a royal straight flush in a fair game is . In a fair game, the chance of three royal straight flushes in a row is 0.000001539x0.000001539x0.000001539 = . (Why do we multiply probabilities here?) Assuming the null hypothesis, for the sake of argument, that is assuming he is not cheating, the chance he will get three royal straight flushes in a row is very, very small – that is the P-value or significance level. The data we see is highly improbable if the null hypothesis were true, so we doubt it is true. Either the dealer got very, very lucky, or he cheated. This is the logic of all hypothesis tests.
* **One sample t-test**: Let  be n independent observations from a Normal distribution with expectation  and variance . We wish to test the null hypothesis, . We do this using the one-sample t-test:

t = 

looking this up in tables of the t-distribution with *n-1* degrees of freedom to get the P-value.

* **One-sided vs Two-sided tests**: In a two-sided test, we don’t care whether  is bigger than or smaller than , so we reject at the 5% level when |t| is one of the 5% largest values of |t|. This means we reject for 2.5% of t’s that are very positive and 2.5% of t’s that are very negative:

Reject

Reject

In a two sided test we reject when t is big positive or big negative. If we reject when the P-value is less than 0.05, then each tail has probability 0.025.

# In a one sided test, we do care, and only want to reject when is on one particular side of, say when is bigger than , so we reject at the 5% level when t is one of the 5% largest values of t. This means we reject for the 5% of t’s that are very positive:

Reject

In a one sided test we reject on just one side, say big positive. If we reject when the P-value is less than 0.05, the tail on the right has probability 0.05.

* **Should I do a one-sided or a two-sided test**: Scientists mostly report two-sided tests.

Some Aspects of Nonparametrics in R

Script is my commentary to you. **Bold Courier is what I type in R.** Regular Courier is what R answered.

What is R?

R is a close relative of Splus, but R is available for free. You can download R from

<http://cran.r-project.org/> . R is very powerful and is a favorite (if not the favorite) of statisticians; however, it is not easy to use. It is command driven, not menu driven. You can add things to R that R doesn’t yet know how to do by writing a little program. R gives you fine control over graphics. Most people need a book to help them, and so Mainland & Braun’s book, Data Analysis and Graphics Using R, Cambridge University Press, 2003, is in the book store as an OPTIONAL book.

This is the cadmium example, paired data, Wilcoxon’s signed rank test.

First, enter the data.

> **cadmium<-c(30,35,353,106,-63,20,52,9966,106,24146,51,106896)**

This command asks for Wilcoxon’s signed rank test with the confidence interval and the Hodges-Lehmann estimate.

> **wilcox.test(cadmium,conf.int=T)**

Wilcoxon signed rank test with continuity correction

data: cadmium

V = 72, p-value = 0.01076

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

35.00005 12249.49999

sample estimates:

(pseudo)median

191.4999

Warning messages:

1: cannot compute exact p-value with ties in: wilcox.test.default(cadmium, conf.int = T)

2: cannot compute exact confidence interval with ties in: wilcox.test.default(cadmium, conf.int = T)

You can teach R new tricks. This is a little program to compute Walsh averages. You enter the program. Then R knows how to do it. You can skip this page if you don’t want R to do new tricks.

**> walsh<-function(data)**

**{**

**w <- outer(data, data, "+")/2.**

**n <- length(data)**

**w <- w[outer(1.:n, 1.:n, "<=")]**

**sort(w)**

**}**

Now we try the program on the cadmium data. It returns all the Walsh averages.

> **walsh(cadmium)**

[1] -63.0 -21.5 -16.5 -14.0 -6.0 -5.5 20.0 21.5 21.5 25.0 27.5 30.0 32.5 35.0 35.5 36.0 40.5

[18] 41.0 43.0 43.5 51.0 51.5 52.0 63.0 63.0 68.0 68.0 70.5 70.5 78.5 78.5 79.0 79.0 106.0

[35] 106.0 106.0 145.0 186.5 191.5 194.0 202.0 202.5 229.5 229.5 353.0 4951.5 4993.0 4998.0 5000.5 5008.5 5009.0

[52] 5036.0 5036.0 5159.5 9966.0 12041.5 12083.0 12088.0 12090.5 12098.5 12099.0 12126.0 12126.0 12249.5 17056.0 24146.0 53416.5 53458.0

[69] 53463.0 53465.5 53473.5 53474.0 53501.0 53501.0 53624.5 58431.0 65521.0 106896.0

The Hodges-Lehmann estimate is the median of the Walsh averages. This agrees with what wilcox.test just told us.

> **median(walsh(cadmium))**

[1] 192.75

This does the Wilcoxon rank sum test with confidence interval and Hodges-Lehmann estimate.

First we enter the PTT times.

> **pttRecan<-c(41,86,90,74,146,57,62,78,55,105,46,94,26,101,72,119,88)**

> **pttControl<-c(34,23,36,25,35,24,87,48)**

Then we compare the two groups.

> wilcox.test(pttRecan,pttControl,conf.int=T)

Wilcoxon rank sum test

data: pttRecan and pttControl

W = 120, p-value = 0.00147

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

18 63

sample estimates:

difference in location

40

Instead, we can take square roots first.

> wilcox.test(sqrt(pttRecan),sqrt(pttControl),conf.int=T)

Wilcoxon rank sum test

data: sqrt(pttRecan) and sqrt(pttControl)

W = 120, p-value = 0.00147

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

1.416198 4.218924

sample estimates:

difference in location

2.769265

This is the program that does both Wilcoxon tests.

**help(wilcox.test)**

wilcox.test package:stats R Documentation

Wilcoxon Rank Sum and Signed Rank Tests

Description:

Performs one and two sample Wilcoxon tests on vectors of data; the

latter is also known as 'Mann-Whitney' test.

Usage:

wilcox.test(x, ...)

## Default S3 method:

wilcox.test(x, y = NULL,

alternative = c("two.sided", "less", "greater"),

mu = 0, paired = FALSE, exact = NULL, correct = TRUE,

conf.int = FALSE, conf.level = 0.95, ...)

## S3 method for class 'formula':

wilcox.test(formula, data, subset, na.action, ...)

Arguments:

x: numeric vector of data values. Non-finite (e.g. infinite or

missing) values will be omitted.

y: an optional numeric vector of data values.

alternative: a character string specifying the alternative hypothesis,

must be one of '"two.sided"' (default), '"greater"' or

'"less"'. You can specify just the initial letter.

mu: a number specifying an optional location parameter.

paired: a logical indicating whether you want a paired test.

exact: a logical indicating whether an exact p-value should be

computed.

correct: a logical indicating whether to apply continuity correction

in the normal approximation for the p-value.

conf.int: a logical indicating whether a confidence interval should be

computed.

conf.level: confidence level of the interval.

formula: a formula of the form 'lhs ~ rhs' where 'lhs' is a numeric

variable giving the data values and 'rhs' a factor with two

levels giving the corresponding groups.

data: an optional data frame containing the variables in the model

formula.

subset: an optional vector specifying a subset of observations to be

used.

na.action: a function which indicates what should happen when the data

contain 'NA's. Defaults to 'getOption("na.action")'.

...: further arguments to be passed to or from methods.

Details:

The formula interface is only applicable for the 2-sample tests.

If only 'x' is given, or if both 'x' and 'y' are given and

'paired' is 'TRUE', a Wilcoxon signed rank test of the null that

the distribution of 'x' (in the one sample case) or of 'x-y' (in

the paired two sample case) is symmetric about 'mu' is performed.

Otherwise, if both 'x' and 'y' are given and 'paired' is 'FALSE',

a Wilcoxon rank sum test (equivalent to the Mann-Whitney test: see

the Note) is carried out. In this case, the null hypothesis is

that the location of the distributions of 'x' and 'y' differ by

'mu'.

By default (if 'exact' is not specified), an exact p-value is

computed if the samples contain less than 50 finite values and

there are no ties. Otherwise, a normal approximation is used.

Optionally (if argument 'conf.int' is true), a nonparametric

confidence interval and an estimator for the pseudomedian

(one-sample case) or for the difference of the location parameters

'x-y' is computed. (The pseudomedian of a distribution F is the

median of the distribution of (u+v)/2, where u and v are

independent, each with distribution F. If F is symmetric, then

the pseudomedian and median coincide. See Hollander & Wolfe

(1973), page 34.) If exact p-values are available, an exact

confidence interval is obtained by the algorithm described in

Bauer (1972), and the Hodges-Lehmann estimator is employed.

Otherwise, the returned confidence interval and point estimate are

based on normal approximations.

Value:

A list with class '"htest"' containing the following components:

statistic: the value of the test statistic with a name describing it.

parameter: the parameter(s) for the exact distribution of the test

statistic.

p.value: the p-value for the test.

null.value: the location parameter 'mu'.

alternative: a character string describing the alternative hypothesis.

method: the type of test applied.

data.name: a character string giving the names of the data.

conf.int: a confidence interval for the location parameter. (Only

present if argument 'conf.int = TRUE'.)

estimate: an estimate of the location parameter. (Only present if

argument 'conf.int = TRUE'.)

Note:

The literature is not unanimous about the definitions of the

Wilcoxon rank sum and Mann-Whitney tests. The two most common

definitions correspond to the sum of the ranks of the first sample

with the minimum value subtracted or not: R subtracts and S-PLUS

does not, giving a value which is larger by m(m+1)/2 for a first

sample of size m. (It seems Wilcoxon's original paper used the

unadjusted sum of the ranks but subsequent tables subtracted the

minimum.)

R's value can also be computed as the number of all pairs '(x[i],

y[j])' for which 'y[j]' is not greater than 'x[i]', the most

common definition of the Mann-Whitney test.

References:

Myles Hollander & Douglas A. Wolfe (1999)Or second edition (1999).

David F. Bauer (1972), Constructing confidence sets using rank

statistics. \_Journal of the American Statistical Association\_

\*67\*, 687-690.

See Also:

'psignrank', 'pwilcox'.

'kruskal.test' for testing homogeneity in location parameters in

the case of two or more samples; 't.test' for a parametric

alternative under normality assumptions.

Examples:

## One-sample test.

## Hollander & Wolfe (1973), 29f.

## Hamilton depression scale factor measurements in 9 patients with

## mixed anxiety and depression, taken at the first (x) and second

## (y) visit after initiation of a therapy (administration of a

## tranquilizer).

x <- c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)

y <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)

wilcox.test(x, y, paired = TRUE, alternative = "greater")

wilcox.test(y - x, alternative = "less") # The same.

wilcox.test(y - x, alternative = "less",

exact = FALSE, correct = FALSE) # H&W large sample

# approximation

## Two-sample test.

## Hollander & Wolfe (1973), 69f.

## Permeability constants of the human chorioamnion (a placental

## membrane) at term (x) and between 12 to 26 weeks gestational

## age (y). The alternative of interest is greater permeability

## of the human chorioamnion for the term pregnancy.

x <- c(0.80, 0.83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, 0.73, 1.46)

y <- c(1.15, 0.88, 0.90, 0.74, 1.21)

wilcox.test(x, y, alternative = "g") # greater

wilcox.test(x, y, alternative = "greater",

exact = FALSE, correct = FALSE) # H&W large sample

# approximation

wilcox.test(rnorm(10), rnorm(10, 2), conf.int = TRUE)

## Formula interface.

boxplot(Ozone ~ Month, data = airquality)

wilcox.test(Ozone ~ Month, data = airquality,

subset = Month %in% c(5, 8))

You can get information about other nonparametric procedures.

help(kruskal.test)

help(friedman.test)

help(cor.test)

help(cor)

**Binomial in R**

Binomial probabilities

Chance of 2 heads in 3 independent trial with probability 1/3 of a head:

> **dbinom(2,3,1/3)**

[1] 0.2222222

Numbers from 0 to 3

> **0:3**

[1] 0 1 2 3

Chances of 0, 1, 2 or 3 heads in 3 independent trials with probability 1/3 of a head

> **dbinom(0:3,3,1/3)**

[1] 0.29629630 0.44444444 0.22222222 0.03703704

Cumulative probabilities: chance of 1 or fewer heads in 3 independent trials with probability 1/3 of a head

> **pbinom(1,3,1/3)**

[1] 0.7407407

Compare with dbinom result above:

> **0.29629630+0.44444444**

[1] 0.7407407

Probability of 24 or fewer heads in 50 trials with probability 1/3 of a head:

> **pbinom(24,50,1/3)**

[1] 0.9891733

Probability of 25 or more heads in 50 trials with probability 1/3 of a head:

> **1-pbinom(24,50,1/3)**

[1] 0.01082668

So of course

> **0.01082668+0.9891733**

[1] 1

One sided test and confidence interval

> **binom.test(25,50,p=1/3,alternative="greater")**

Exact binomial test

data: 25 and 50

number of successes = 25, number of trials = 50, p-value = 0.01083

alternative hypothesis: true probability of success is greater than 0.3333333

95 percent confidence interval:

0.3762459 1.0000000

sample estimates:

probability of success

0.5

Two sided test and confidence interval

> **binom.test(25,50,p=1/3)**

Exact binomial test

data: 25 and 50

number of successes = 25, number of trials = 50, p-value = 0.01586

alternative hypothesis: true probability of success is not equal to 0.3333333

95 percent confidence interval:

0.355273 0.644727

sample estimates:

probability of success

0.5

Get help

> **help(rbinom)**

or

> **help(binom.test)**

**Looking at Densities**

**Sampling from Distributions**

**In R**

This creates equally spaced numbers between -5 and 5. They will be plotting positions.

> space<-(-500):500/100

dnorm(x) gives you the Normal density function, rnorm(n) gives you n random draws from the Normal, pnorm gives you the Normal cumulative distribution, qnorm gives you the Normal quantiles. The same idea works for the Cauchy distribution (eg rcauchy(n)) or the logistic distribution (eg rlogis(n)).

> **pnorm(-1.96)**

[1] 0.02499790

> **pnorm(1.96)**

[1] 0.9750021

> **qnorm(.025)**

[1] -1.959964

> **rnorm(5)**

[1] 0.9154958 0.5835557 0.3850987 -1.1506946

0.5503568

This sets you up to do a 2x2 four panel plot

> **par(mfrow=c(2,2))**

> **plot(space,dnorm(space))**

> **plot(space,dcauchy(space))**

> **plot(space,dlogis(space))**

> **boxplot(rnorm(500),rlogis(500),rcauchy(500))**

**Bloodbags Data**

> **bloodbags2**

id acdA acd dif

1 1 63.0 58.5 4.5

2 2 48.4 82.6 -34.2

3 3 58.2 50.8 7.4

4 4 29.3 16.7 12.6

5 5 47.0 49.5 -2.5

6 6 27.7 26.0 1.7

7 7 22.3 56.3 -34.0

8 8 43.0 35.7 7.3

9 9 53.3 37.9 15.4

10 10 49.5 53.3 -3.8

11 11 41.1 38.2 2.9

12 12 32.9 37.1 -4.2

If you attach the data, then you can refer to variables by their names. Remember to detach when done.

> **attach(bloodbags2)**

Plot data!

> **par(mfrow=c(1,2))**

> **boxplot(dif,ylim=c(-40,20))**

> **qqnorm(dif,ylim=c(-40,20))**

Data do not look Normal in Normal plot, and Shapiro-Wilk test confirms this.

> **shapiro.test(dif)**

Shapiro-Wilk normality test

data: dif

W = 0.8054, p-value = 0.01079

Wilcoxon signed rank test, with Hodges-Lehmann point estimate and confidence interval using Walsh averages.

> **wilcox.test(dif,conf.int=T)**

Wilcoxon signed rank test

data: dif

V = 44, p-value = 0.7334

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

-14.85 7.35

sample estimates:

(pseudo)median

1.575

> **detach(bloodbags2)**

**Sign Test Procedures in R**

> **attach(cadmium)**

> **dif**

30 35 353 106 -63 20 52 9966 106 24146 51 106896

The sign test uses just the signs, not the ranks.

> **1\*(dif<0)**

0 0 0 0 1 0 0 0 0 0 0 0

There was 1 negative differences in 12 pairs.

> **sum(1\*(dif<0))**

[1] 1

Compare to the binomial with 12 trials, 1 tail, probability of head ½: One sided p-value

> **pbinom(1,12,1/2)**

[1] 0.003173828

Usual two sided p-value

> **2\*pbinom(1,12,1/2)**

[1] 0.006347656

Because the distribution is very long tailed, the sign test is better than the signed rank for these data. This is the binomial for n=12:

> **rbind(0:12,round(pbinom(0:12,12,.5),3))**

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]

[1,] 0 1.000 2.000 3.000 4.000 5.000 6.000 7.000 8.000 9.000 10.000 11 12

[2,] 0 0.003 0.019 0.073 0.194 0.387 0.613 0.806 0.927 0.981 0.997 1 1

Two of the sorted observations (order statistics) form the confidence interval for the population median

> **sort(dif)**

[1] -63 20 30 35 51 52 106 106 353 9966 24146 106896

At the 0.025 level, you can reject for a sign statistic of 2, but not 3,

> **pbinom(3,12,1/2)**

[1] 0.07299805

> **pbinom(2,12,1/2)**

[1] 0.01928711

So, it is #3 and #10 that form the confidence interval:

> **sort(dif)[c(3,10)]**

[1] 30 9966

> sum(1\*(dif-30.001)<0)

[1] 3

> sum(1\*(dif-29.9999)<0)

[1] 2

> **2\*pbinom(sum(1\*(dif-29.9999)<0),12,1/2)**

[1] 0.03857422

> **2\*pbinom(sum(1\*(dif-30.001)<0),12,1/2)**

[1] 0.1459961

**Rank Sum & Transformations**

Model Y = 2, X = 2, or Y = (2)(2) = (2)X,

so log2(Y) =  , log2(X) = 

> **wilcox.test(log2(pttRecan),log2(pttControl),conf.int=T)**

Wilcoxon rank sum test

data: log2(pttRecan) and log2(pttControl)

W = 120, p-value = 0.00147

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

0.5849625 1.6415460

sample estimates:

difference in location

1.172577

Transform back to estimate multiplier 2

> **2^0.5849625**

[1] 1.5

> **2^1.6415460**

[1] 3.12

> **2^1.172577**

[1] 2.25414

95% Confidence interval for multiplier 2 is [1.5, 3.12] and point estimate is 2.25.

Two Sample Comparisons in Stata

(Commands are in **bold**)

**. kwallis PTT, by( Recanal)**

Test: Equality of populations (Kruskal-Wallis test)

Recanal \_Obs \_RankSum

0 8 52.00

1 17 273.00

chi-squared = 9.176 with 1 d.f.

probability = 0.0025

chi-squared with ties = 9.176 with 1 d.f.

probability = 0.0025

**. generate rt = sqrt(PTT)**

**. generate lg2Ptt =ln( PTT)/0.693147**

**. npshift PTT, by(Recanal)** *Bad idea! Not a shift!*

Hodges-Lehmann Estimates of Shift Parameters

-----------------------------------------------------------------

Point Estimate of Shift : Theta = Pop\_2 - Pop\_1 = 40

95% Confidence Interval for Theta: [17 , 64]

-----------------------------------------------------------------

**. npshift rt, by(Recanal)** *Better idea. Hard to interpret!*

Hodges-Lehmann Estimates of Shift Parameters

-----------------------------------------------------------------

Point Estimate of Shift : Theta = Pop\_2 - Pop\_1 = 2.769265

95% Confidence Interval for Theta: [1.403124 , 4.246951]

-----------------------------------------------------------------

**. npshift lg2Ptt, by(Recanal)** *Best idea. Correct, interpretable.*

Hodges-Lehmann Estimates of Shift Parameters

-----------------------------------------------------------------

Point Estimate of Shift : Theta = Pop\_2 - Pop\_1 = 1.172577

95% Confidence Interval for Theta: [.4518747 , 1.646364]

-----------------------------------------------------------------

*21.1726 = 2.25*

*2.4519 = 1.37 21.6464 = 3.13*

**Ansari Bradley Test**

> **help(ansari.test)**

**Example from book, page 147. Two methods of determining level of iron in serum. True level was 105m grams/100ml. Which is more accurate? (Data in R help)**

> ramsay <- c(111, 107, 100, 99, 102, 106, 109, 108, 104, 99,101, 96, 97, 102, 107, 113, 116, 113, 110, 98)

> jung.parekh <- c(107, 108, 106, 98, 105, 103, 110, 105, 104,100, 96, 108, 103, 104, 114, 114, 113, 108, 106, 99)

> **ansari.test(ramsay, jung.parekh)**

Ansari-Bradley test

data: ramsay and jung.parekh

AB = 185.5, p-value = 0.1815

alternative hypothesis: true ratio of scales is not equal to 1

> **ansari.test(pttControl,pttRecan)**

Ansari-Bradley test

data: pttControl and pttRecan

AB = 42, p-value = 0.182

alternative hypothesis: true ratio of scales is not equal to 1

>**ansari.test(pttControl-median(pttControl),pttRecan-median(pttRecan))**

Ansari-Bradley test

data: pttControl - median(pttControl) and pttRecan - median(pttRecan)

AB = 68, p-value = 0.1205

alternative hypothesis: true ratio of scales is not equal to 1

**Kolmogorov-Smirnov Test in R**

Tests whether distributions differ in any way.

Mostly useful if you are not looking for a change in level or dispersion.

Two simulated data sets

> one<-rexp(1000)-1

> two<-1-rexp(1000)

Similar means and variances (would be the same if n were very large)

> mean(one)

[1] 0.01345924

> mean(two)

[1] -0.0345239

> sd(one)

[1] 0.9891292

> sd(two)

[1] 1.047116

Yet they look very different!

> boxplot(one,two)

The K-S test compares the empirical cumulative distributions:

> par(mfrow=c(1,2))

> plot(ecdf(one),ylab="Proportion <= x",main="one")

> plot(ecdf(two),ylab="Proportion <= x",main="two")

Very small p-value – distributions clearly different!

> ks.test(one,two)

Two-sample Kolmogorov-Smirnov test

data: one and two

D = 0.272, p-value < 2.2e-16

alternative hypothesis: two.sided

**Ideas from Chapter 5 of Hollander et al.**

**Using Their R-package NSM3**

**Lepage in NSM3**

> pLepage(pttRecan,pttControl,method="Asymptotic")

Number of X values: 17 Number of Y values: 8

Lepage D Statistic: 11.1479

Asymptotic upper-tail probability: 0.0038

**Kolmogorov-Smirnov**

> pKolSmirn(pttRecan,pttControl)

Number of X values: 17 Number of Y values: 8

Kolmogorov-Smirnov J Statistic: 0.6985

Exact upper-tail probability: 0.0052

> ks.test(pttRecan,pttControl)

Two-sample Kolmogorov-Smirnov test

data: pttRecan and pttControl

D = 0.6985, p-value = 0.005231

alternative hypothesis: two-sided

**Confidence band**

cbind(ecdf.ks.CI(meyerdif)$lower,ecdf.ks.CI(meyerdif)$upper,sort(unique(meyerdif)))

Quantile

qKolSmirnLSA(.05)

**Kruskal Wallis Test in R**

Data from: Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M., and Saah, M. (1982) Lead absorption in children of employees in a lead-related industry. *American Journal of Epidemiology*, **115**, 549-555.

> lead

control exposed level hyg

1 13 14 high good

2 16 13 high good

3 11 25 high good

4 18 41 high mod

5 24 18 high mod

6 7 49 high mod

7 16 38 high poor

8 18 23 high poor

9 19 37 high poor

10 15 62 high poor

11 18 24 high poor

12 9 45 high poor

13 14 39 high poor

14 18 48 high poor

15 19 44 high poor

16 19 35 high poor

17 11 43 high poor

18 18 34 high poor

19 13 73 high poor

20 22 39 medium good

21 NA 29 medium mod

22 16 31 medium poor

23 25 34 medium poor

24 16 20 medium poor

25 21 22 medium poor

26 12 35 medium poor

27 16 16 low mod

28 11 36 low poor

29 10 23 low poor

30 19 21 low poor

31 10 17 low poor

32 13 27 low poor

33 24 15 low poor

34 13 10 low poor

-- level and hyg are factors. Type: help(factor)

Type: help(kruskal.test)

First we do the test by level.

> **kruskal.test(exposed~level,data=lead)**

Kruskal-Wallis rank sum test

data: exposed by level

Kruskal-Wallis chi-squared = 8.5172, df = 2, p-value = 0.01414

Now we do the test by hyg for kids with level=high..

>**kruskal.test(exposed~hyg,subset=(level=="high"),data=lead)**

Kruskal-Wallis rank sum test

data: exposed by hyg

Kruskal-Wallis chi-squared = 5.5611, df = 2, p-value = 0.062

Now we do the test by level for kids with hyg=poor.

>**kruskal.test(exposed~level,subset=(hyg=="poor"),data=lead)**

Kruskal-Wallis rank sum test

data: exposed by level

Kruskal-Wallis chi-squared = 12.5104, df = 2, p-value = 0.001920

**Jonckheere-Terpstra Test in R (almost)**

You can get R to do most of the work in the Jonckheere-Terpstra test, but you have to do some by hand.

> **high<-lead$exposed[lead$level=="high"]**

> **medium<-lead$exposed[lead$level=="medium"]**

> **low<-lead$exposed[lead$level=="low"]**

> **high**

[1]14 13 25 41 18 49 38 23 37 62 24 45 39 48 44 35 43 34 73

> **medium**

[1] 39 29 31 34 20 22 35

> **low**

[1] 16 36 23 21 17 27 15 10

You do 3 wilcox.test commands, and add them up. Order matters in the command! Not wilcox.test(low,medium)!

> **wilcox.test(medium,low)**

Wilcoxon rank sum test

data: medium and low

W = **45**, p-value = 0.05408

alternative hypothesis: true mu is not equal to 0

> **wilcox.test(high,low)**

Wilcoxon rank sum test with continuity correction

data: high and low

W = **125.5**, p-value = 0.00926

alternative hypothesis: true mu is not equal to 0

> **wilcox.test(high,medium)**

Wilcoxon rank sum test with continuity correction

data: high and medium

W = **90.5**, p-value = 0.1741

alternative hypothesis: true mu is not equal to 0

J = 45 + 125.5 + 90.5 = 261

Compute J\* in expression (6.17) on page 203 in H&W

**Kruskal-Wallis Test in R**

Lead Data.

> **lead[1:3,]**

control exposed level hyg

1 13 14 high good

2 16 13 high good

3 11 25 high good

Kruskal Wallis test of no difference

> **kruskal.test(exposed~level)**

Kruskal-Wallis rank sum test

data: exposed by level

Kruskal-Wallis chi-squared = 8.5172, df = 2, p-value = 0.01414

To do multiple comparisons, get Mann-Whitney statistic from Wilcox.test and convert to Wilcoxon statistic

> **wilcox.test(exposed[level=="high"],exposed[level=="low"])**

Wilcoxon rank sum test with continuity correction

data: exposed[level == "high"] and exposed[level == "low"]

W = 125.5, p-value = 0.00926

alternative hypothesis: true mu is not equal to 0

Warning message:

cannot compute exact p-value with ties in: wilcox.test.default(exposed[level == "high"], exposed[level ==

> **125.5+((19+1)\*19/2)**

[1] 315.5 This is one rank sum

> **wilcox.test(exposed[level=="high"],exposed[level=="medium")**

Wilcoxon rank sum test with continuity correction

data: exposed[level == "high"] and exposed[level == "medium"]

W = 90.5, p-value = 0.1741

alternative hypothesis: true mu is not equal to 0

Warning message:

cannot compute exact p-value with ties in: wilcox.test.default(exposed[level == "high"], exposed[level ==

> **90.5+((19+1)\*19/2)**

[1] 280.5 This is one rank sum

> **wilcox.test(exposed[level == "medium"],exposed[level == "low"])**

Wilcoxon rank sum test

data: exposed[level == "medium"] and exposed[level == "low"]

W = 45, p-value = 0.05408

alternative hypothesis: true mu is not equal to 0

> **45+((7+1)\*7/2)**

[1] 73 This is one rank sum

**Jonckheere/Terpstra Test for Ordered Alternatives (6.2)**

Illustrates the use of NSM3 R-package from Hollander/Wolfe/

> library(NSM3)

> help(pJCK)

> head(lead)

control exposed level hyg both

1 13 14 high good high.ok

2 16 13 high good high.ok

3 11 25 high good high.ok

4 18 41 high mod high.ok

5 24 18 high mod high.ok

6 7 49 high mod high.ok

> attach(lead)

Does father’s exposure predict the child’s blood lead level?

> pJCK(exposed,g=as.integer(level),method="Asymptotic")

Ties are present, so p-values are based on conditional null distribution.

Group sizes: 8 7 19

Jonckheere-Terpstra J\* Statistic: 3.0078

Asymptotic upper-tail probability: 0.0013

> kruskal.test(exposed~level)

Kruskal-Wallis rank sum test

data: exposed by level

Kruskal-Wallis chi-squared = 8.5172, df = 2, p-value = 0.01414

For father’s with high exposure, does father’s hygiene predict the child’s blood lead level?

> pJCK(exposed[level=="high"],g=as.integer(hyg[level=="high"]),

method="Asymptotic")

Group sizes: 3 3 13

Jonckheere-Terpstra J\* Statistic: 1.9347

Asymptotic upper-tail probability: 0.0265

> kruskal.test(exposed[level=="high"]~as.integer(hyg[level=="high"]))

Kruskal-Wallis rank sum test

data: exposed[level == "high"] by as.integer(hyg[level == "high"])

Kruskal-Wallis chi-squared = 5.5611, df = 2, p-value = 0.062

**Multiple Comparisons**

> **help(pairwise.wilcox.test)**

Bonferroni

Bonferroni inequality says Pr(A or B) <= Pr(A)+Pr(B), or more generally, the

Bonferroni inequality says Pr(A1 or A2 or … or AL) <= Pr(A1)+ Pr(A2)+…+ Pr(AL).

Know that if H0 is true, Pr(p-value<0.05)<=0.05. More generally, Pr(p-value<=)<=, for all 0<<1.

If we have L hypotheses, H01,… H0L with p-values p1,…,pL and we reject H0k when pk<=0.05/L, then the chance that we falsely reject any H0k is <=0.05.

Pick one hypothesis. If H0k is false, then we cannot falsely reject it, so the chance that we falsely reject it is zero. If H0k is true, we falsely reject it when pk<=0.05/L, which happens with probability <=0.05/L.

Then the chance that we falsely reject any H0k is

Pr(Falsely reject H01 or …or Falsely reject H0L) <=

Pr(Falsely reject H01)+…+Pr(Falsely reject H0L)<= 0.05/L+…+0.05/L = 0.05.

Because 0.0093 < 0.05/3 = 0.016667, we reject it by the Bonferroni method.

> **pairwise.wilcox.test(exposed,level,p.adjust.method="bonferroni")**

Pairwise comparisons using Wilcoxon rank sum test

data: exposed and level

low medium

medium 0.162 -

high 0.028 0.522

P value adjustment method: bonferroni

Holm

Order p-values, smallest to largest, p(1)<…<p(L)

P(1) = 0.0093, P(2) = 0.0541, P(3) = 0.1741.

If P(1) <= 0.05/L = 0.01667, then I can safely reject the corresponding hypothesis by Bonferroni.

Holm’s idea is that it is now safe to assume this hypothesis is false, and test the rest assuming there are only L-1 possible true hypotheses.

This means that P(2) is significant if less than 0.05/(L-1) = 0.05/(3-1) = 0.025.

If we reject this hypothesis, we continue and compare P(3) to 0.05/(L-2), etc.

Holm showed that the chance that we falsely reject any true hypothesis among the L hypotheses by this method is at most 0.05.

Notice that it gets easier to reject hypotheses as more hypotheses are rejected.

Notice that you must stop at the first j such that P(j) > 0.05/(L-j+1). You cannot skip forward.

> **pairwise.wilcox.test(lead$exposed,lead$level)**

Pairwise comparisons using Wilcoxon rank sum test

data: lead$exposed and lead$level

low medium

medium 0.108 -

high 0.028 0.174

P value adjustment method: **holm**

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.

Wright, S. P. (1992). Adjusted P-values for simultaneous

inference. *Biometrics*, 48, 1005-1013.

**Rapid Eye Movement Example**

> **rapideye**

eyetrack fixation

1 980.8 4.85

2 926.4 4.41

3 892.9 3.80

4 870.2 4.53

5 854.6 4.33

6 777.2 3.81

7 772.6 3.97

8 702.4 3.68

9 561.7 3.43

> **plot(rapideye$eyetrack,rapideye$fixation)**

> **cor.test(rapideye$eyetrack,rapideye$fixation,method="kendall")**

Kendall's rank correlation tau

data: rapideye$eyetrack and rapideye$fixation

T = 30, p-value = 0.01267

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.6666667

> **round(theil(rapideye$eyetrack,rapideye$fixation),5)**

[1] -0.03478 -0.03216 -0.01384 -0.00214 -0.00141 -0.00009 0.00063 0.00111

[9] 0.00112 0.00174 0.00176 0.00178 0.00256 0.00269 0.00286 0.00289

[17] 0.00307 0.00326 0.00339 0.00357 0.00402 0.00412 0.00413 0.00420

[25] 0.00423 0.00427 0.00439 0.00507 0.00511 0.00574 0.00672 0.00774

[33] 0.00809 0.01195 0.01282 0.01821

> **median(theil(rapideye$eyetrack,rapideye$fixation))**

[1] 0.003323571

**Hoeffding’s Test of Independence in R**

Hollander and Wolfe (1999) Section 8.6

It’s not in “standard R”. So you search:

> **help.search("hoeff")**

And you find it.

> **help(hoeffd,pack=Hmisc)**

hhoeffd package:Hmisc R Documentation

Matrix of Hoeffding's D Statistics

Description:

Computes a matrix of Hoeffding's (1948) 'D' statistics for all

possible pairs of columns of a matrix. 'D' is a measure of the

distance between 'F(x,y)' and 'G(x)H(y)', where 'F(x,y)' is the

joint CDF of 'X' and 'Y', and 'G' and 'H' are marginal CDFs.

Missing values are deleted in pairs rather than deleting all rows

of 'x' having any missing variables. The 'D' statistic is robust

against a wide variety of alternatives to independence, such as

non-monotonic relationships. The larger the value of 'D', the more

dependent are 'X' and 'Y' (for many types of dependencies). **'D'**

**used here is 30 times Hoeffding's original 'D'**, and ranges from

-0.5 to 1.0 if there are no ties in the data. 'print.hoeffd'

prints the information derived by 'hoeffd'. The higher the value

of 'D', the more dependent are 'x' and 'y' ...

Go to “Packages” menu, “Load Package” option, and pick “Hmisc”

Then it’s yours. This is the example from the book:

> eg8.5

collagen proline

[1,] 7.1 2.8

[2,] 7.1 2.9

[3,] 7.2 2.8

[4,] 8.3 2.6

[5,] 9.4 3.5

[6,] 10.5 4.6

[7,] 11.4 5.0

> hoeffd(eg8.5)

D

[,1] [,2]

[1,] 1.00 0.19

[2,] 0.19 1.00

n= 7

P

[,1] [,2]

[1,] 0.0215

[2,] 0.0215

30 times the books answer is 0.19. The large sample p-value is given, close to the book’s large sample value.

**Bootstrap Confidence Interval for Kendall’s Correlation**

(Hollander and Wolfe 1999 section 8.4)

The original data.

> **rapideye**

eyetrack fixation

1 980.8 4.85

2 926.4 4.41

3 892.9 3.80

4 870.2 4.53

5 854.6 4.33

6 777.2 3.81

7 772.6 3.97

8 702.4 3.68

9 561.7 3.43

Kendall’s correlation for the original data

> **cor.test(rapideye$eyetrack,rapideye$fixation,method="kendall")**

Kendall's rank correlation tau

data: rapideye$eyetrack and rapideye$fixation

T = 30, p-value = 0.01267

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.6666667

A bootstrap sample is a sample of size n=9 WITH REPLACEMENT from the n=9 observations

> **j<-sample(1:9,9,replace=T)**

> j

[1] 2 1 9 9 3 9 4 4 9

> **rapideye[j,]**

eyetrack fixation

2 926.4 4.41

1 980.8 4.85

9 561.7 3.43

9.1 561.7 3.43

3 892.9 3.80

9.2 561.7 3.43

4 870.2 4.53

4.1 870.2 4.53

9.3 561.7 3.43

Kendall’s correlation for the first bootstrap sample

> **cor.test(rapideye$eyetrack[j],rapideye$fixation[j],method="kendall")**

Kendall's rank correlation tau

data: rapideye$eyetrack[j] and rapideye$fixation[j]

z = 2.7179, p-value = 0.00657

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.724138

Little function bootkendall computes B bootstrap samples Kendall’s tau’s, sorts them, and gives the lower and upper 2.5% as the confidence interval.

> **bootkendall**

function(Z,B){

tauhat<-rep(NA,B)

n<-dim(Z)[1]

for (i in 1:B){

j<-sample(1:n,n,replace=T)

tauhat[i]<cor.test(Z[j,1],Z[j,2],method="kendall")$estimate

}

tauhat<-sort(tauhat)

k<-floor(B\*0.025)

c(tauhat[k],tauhat[B+1-k])

}

Let’s try it.

> **bootkendall(rapideye,1000)**

[1] 0.03225806 1.00000000

There were 50 or more warnings (use warnings() to see the first 50)

Because they are samples, you get different answers each time. Need to set B large!

> **bootkendall(rapideye,1000)**

[1] 0.1724138 1.0000000

> **bootkendall(rapideye,5000)**

[1] 0.125 1.000

There were 50 or more warnings (use warnings() to see the first 50)

> **bootkendall(rapideye,5000)**

[1] 0.1111111 1.0000000

There were 50 or more warnings (use warnings() to see the first 50)

**Nonparametric Rank Based Multiple Regression (Section 9.6 in H&W)**

> Fuel[1:3,]

id state fuel tax lic inc road

1 1 ME 541 9 52.5 3.571 1.976

2 2 NH 524 9 57.2 4.092 1.250

3 3 VT 561 9 58.0 3.865 1.586

You must install the Rfit package using the packages menu.

> library(Rfit)

> help(rfit)

> attach(Fuel)

**Fitting a model**

> out<-rfit(fuel~tax+lic)

> summary(out)

Call: rfit.default(formula = fuel ~ tax + lic)

Coefficients:

Estimate Std. Error t.value p.value

130.4176 181.9920 0.7166 0.47740

tax -29.8342 12.8729 -2.3176 0.02519 \*

lic 11.7820 2.2064 5.3398 3.116e-06 \*\*\*

Multiple R-squared (Robust): 0.4534957

Reduction in Dispersion Test: 18.67077 p-value: 0

**Residuals from this fit**

> res<-as.vector(out$residual)

> res

[1] 60.53645424 -11.83879810 15.73562704 …

Least squares minimizes the sum of the squares of the residuals. In contrast, rfit minimizes a rank based measure of the dispersion of the residuals due to Jaeckel (1972).

> outlm<-lm(fuel~tax+lic)

> reslm<-outlm$residual

So reslm has the least squares residuals.

> JaeckelD(res)

[1] 3313.408

> JaeckelD(reslm)

[1] 3322.443

> sum(res^2)

[1] 262540

> sum(reslm^2)

[1] 260834

So sum(reslm^2) < sum(res^2) but JaeckelD(res) < JaeckelD(reslm).

> Fuel2<-cbind(Fuel,res,reslm)

> Fuel2[39:41,]

id state fuel tax lic inc road res reslm

39 39 ID 648 8.5 66.3 3.635 3.274 -9.97179 -18.0659

40 40 WY 968 7.0 67.2 4.345 3.905 254.67321 242.5577

41 41 CO 587 7.0 62.6 4.449 4.639 -72.12974 -80.8739

Least squares made the residual for WY smaller but the residuals for ID and CO larger than did rfit.

**Testing whether several regression coefficients are zero**

> help(drop.test)

drop.test is the analog to the general linear hypothesis F-test asking whether additional terms in a full model (fitF) are needed or whether a reduced model (fitR) is plausible.

> out<-rfit(fuel~tax+lic)

> out2<-rfit(fuel~tax+lic+inc+road)

> JaeckelD(out$residual)

[1] 3313.408

> JaeckelD(out2$residual)

[1] 2700.551

So of course, the full model (out2) fits better than the reduced model (out), but could the improvement just be chance?

> drop.test(fitF=out2,fitR=out)

Drop in Dispersion Test

F-Statistic p-value

1.0191e+01 2.3847e-04

So the improvement in fit from inc and road is not plausibly due to chance.

Hettmansperger, T.P. and McKean J.W. (2011), Robust Nonparametric Statistical Methods, 2nd ed., New York: Chapman-Hall.

Hettmansperger, T.P. and McKean J.W. (1977) A robust alternative based on ranks to least squares in analyzing linear models. Technonmetrics, 19, 275-284. In JSTOR on the library web-page.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. Annals of Mathematical Statistics, 43, 1449 - 1458. In JSTOR on the library web-page.

**Robust Regression in R**

Robust and nonparametric regression are not quite as standardized as most topics in Hollander and Wolfe. In part, there are many competing proposals, and the dust has not settled to yield one conventional proposal.

The more commonly used procedure is m-estimation. It is available in R, Splus, SAS and Stata.

In R, you need the MASS package, which may be available with your version of R, or can be obtained from the usual location <http://www.r-project.org/>

Once you have the MASS package, type:

>**library(MASS)**

to make its features available. The specific routine you want is rlm (for robust linear model) so type

>**help(rlm)**

to obtain documentation.

The first five rows of the fuel data (from stat 500) are:

> **dim(fuel)**

[1] 48 7

> **fuel[1:5,]**

id state fuel tax lic inc road

1 1 ME 541 9.0 52.5 3.571 1.976

2 2 NH 524 9.0 57.2 4.092 1.250

3 3 VT 561 9.0 58.0 3.865 1.586

4 4 MA 414 7.5 52.9 4.870 2.351

5 5 RI 410 8.0 54.4 4.399 0.431

The call for a robust regression is

> **rlm(fuel~tax+lic,fuel)**

but that generates only a little output, so instead type

> **summary(rlm(fuel~tax+lic,fuel))**

Call: rlm(formula = fuel ~ tax + lic, data = fuel)

Residuals:

Min 1Q Median 3Q Max

-122.3625 -54.8412 0.5224 47.3381 256.4014

Coefficients:

Value Std. Error t value

(Intercept) 143.1360 165.7618 0.8635

tax -30.4104 11.7694 -2.5838

lic 11.6270 2.0173 5.7636

Residual standard error: 81.85 on 45 degrees of freedom

Correlation of Coefficients:

(Intercept) tax

tax -0.7444

lic -0.8509 0.2880

You interpret this much as you do a least squares regression. However, outliers get gradually down-weighted, rather than tested-and-deleted.

> **w<-rlm(fuel~tax+lic,fuel)$w**

> **resid<-rlm(fuel~tax+lic,fuel)$resid**

Look at Wyoming on the next page: without rejecting it as an outlier, it got down-weighted to have weight w=.43 compared to 1 for most observations.> **cbind(fuel,w,resid)**

id state fuel tax lic inc road w resid

1 1 ME 541 9.00 52.5 3.571 1.976 1.0000000 61.1391882

2 2 NH 524 9.00 57.2 4.092 1.250 1.0000000 -10.5077687

3 3 VT 561 9.00 58.0 3.865 1.586 1.0000000 17.1906216

4 4 MA 414 7.50 52.9 4.870 2.351 0.9480140 -116.1271732

5 5 RI 410 8.00 54.4 4.399 0.431 0.8997146 -122.3625059

6 6 CN 457 10.00 57.1 5.342 1.333 1.0000000 -45.9346965

7 7 NY 344 8.00 45.1 5.319 11.868 1.0000000 -80.2312932

8 8 NJ 467 8.00 55.3 5.126 2.138 1.0000000 -75.8268168

9 9 PA 464 8.00 52.9 4.447 8.577 1.0000000 -50.9219877

10 10 OH 498 7.00 55.2 4.512 8.507 1.0000000 -74.0744866

11 11 IN 580 8.00 53.0 4.391 5.939 1.0000000 63.9153111

12 12 IL 471 7.50 52.5 5.126 14.186 1.0000000 -54.4763684

13 13 MI 525 7.00 57.4 4.817 6.930 1.0000000 -72.6539132

14 14 WI 508 7.00 54.5 4.207 6.580 1.0000000 -55.9355781

15 15 MN 566 7.00 60.8 4.332 8.159 1.0000000 -71.1857544

16 16 IA 635 7.00 58.6 4.318 10.340 1.0000000 23.3936723

17 17 MO 603 7.00 57.2 4.206 8.508 1.0000000 7.6714892

18 18 ND 714 7.00 54.0 3.718 4.725 0.7063010 155.8779280

19 19 SD 865 7.00 72.4 4.716 5.915 1.0000000 92.9409052

20 20 NE 640 8.50 67.7 4.341 6.010 1.0000000 -31.7965814

21 21 KS 649 7.00 66.3 4.593 7.834 1.0000000 -52.1343210

22 22 DE 540 8.00 60.2 4.983 0.602 1.0000000 -59.7991761

23 23 MD 464 9.00 51.1 4.897 2.449 1.0000000 0.4170051

24 24 VA 547 9.00 51.7 4.258 4.686 1.0000000 76.4407979

25 25 WV 460 8.50 55.1 4.574 2.619 1.0000000 -65.2962288

26 26 NC 566 9.00 54.4 3.721 4.746 1.0000000 64.0478652

27 27 SC 577 8.00 54.8 3.448 5.399 1.0000000 39.9866893

28 28 GA 631 7.50 57.9 3.846 9.061 1.0000000 42.7377662

29 29 FA 574 8.00 56.3 4.188 5.975 1.0000000 19.5461711

30 30 KY 534 9.00 49.3 3.601 4.650 1.0000000 91.3456269

31 31 TN 571 7.00 51.8 3.640 6.905 1.0000000 38.4573546

32 32 AL 554 7.00 51.3 3.333 6.594 1.0000000 27.2708606

33 33 MS 577 8.00 57.8 3.063 6.524 1.0000000 5.1056530

34 34 AR 628 7.50 54.7 3.357 4.121 1.0000000 76.9442050

35 35 LA 487 8.00 48.7 3.528 3.495 1.0000000 20.9114632

36 36 OK 644 6.58 62.9 3.802 7.834 1.0000000 -30.3748357

37 37 TX 640 5.00 56.6 4.045 17.782 1.0000000 -9.1730456

38 38 MT 704 7.00 58.6 3.897 6.385 1.0000000 92.3936723

39 39 ID 648 8.50 66.3 3.635 3.274 1.0000000 -7.5187644

40 40 WY 968 7.00 67.2 4.345 3.905 0.4293990 256.4013681

41 41 CO 587 7.00 62.6 4.449 4.639 1.0000000 -71.1143762

42 42 NM 699 7.00 56.3 3.656 3.985 0.9646353 114.1358001

43 43 AZ 632 7.00 60.3 4.300 3.635 1.0000000 0.6277517

44 44 UT 591 7.00 50.8 3.745 2.611 1.0000000 70.0843667

45 45 NV 782 6.00 67.2 5.215 2.302 1.0000000 39.9909971

46 46 WN 510 9.00 57.1 4.476 3.942 1.0000000 -23.3450675

47 47 OR 610 7.00 62.3 4.296 4.083 1.0000000 -44.6262726

48 48 CA 524 7.00 59.3 5.002 9.794 1.0000000 -95.7452362

The ideas in section 9.6 of Hollander and Wolfe are now available in R.

Install the Rreg package. Then load it.

> library(Rreg)

> help(rfit)

> help(drop.test)

rfit fits a multiple regression by ranks, whereas drop.test is used to test that a subset of regression coefficients have values zero.

> attach(Fuel)

> Fuel[1:3,]

id state fuel tax lic inc road

1 1 ME 541 9 52.5 3.571 1.976

2 2 NH 524 9 57.2 4.092 1.250

3 3 VT 561 9 58.0 3.865 1.586

Syntax is similar to linear models, lm, in Stat 500.

> rfit(fuel~tax+lic)

Call:

rfit.default(formula = fuel ~ tax + lic)

Coefficients:

tax lic

135.91506 -30.22136 11.73888

> summary(rfit(fuel~tax+lic))

Call:

rfit.default(formula = fuel ~ tax + lic)

Coefficients:

Estimate Std. Error t.value p.value

135.915 182.842 0.7433 0.46122

tax -30.221 12.935 -2.3364 0.02409 \*

lic 11.739 2.217 5.2948 3.621e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple R-squared (Robust): 0.4523139

Reduction in Dispersion Test: 18.58193 p-value: 0

> rfit(fuel~tax+lic)$residual[1:5]

[1] 60.78579 -11.38696 16.22193 -116.24181 -122.73945

> rfit(fuel~tax+lic)$fit[1:5]

[1] 480.2142 535.3870 544.7781 530.2418 532.7395

> modf<-rfit(fuel~tax+lic+road)

> modr<-rfit(fuel~tax)

> drop.test(modf,modr)

Test whether lic and road have zero coefficients in the model with predictors tax, lic, road.

Drop in Dispersion Test

F-Statistic p-value

1.2385e+01 5.4084e-05

For mor info, go o to the Journal of Statistical Software

|  |  |
| --- | --- |
| Authors: | [Jeff T. Terpstra](http://www.jstatsoft.org/index.php?vol=14##) and [Joseph W. McKean](http://www.jstatsoft.org/index.php?vol=14##) |
| Title: | **Rank-Based Analysis of Linear Models Using R** |
| Reference: | Volume 14, 2005, Issue 7 |
| Acquire: | [paper](http://www.jstatsoft.org/counter.php?id=132&url=v14/i07/v14i07.pdf&ct=1) [1816]  [browse files](http://www.jstatsoft.org/counter.php?id=132&url=v14/i07&ct=2) [749] |
| Dates: | *submitted*: 2004-04-06    *accepted*: 2005-07-01 |

which is at:

<http://www.jstatsoft.org/index.php?vol=14>

with documentation

<http://www.jstatsoft.org/v14/i07/v14i07.pdf>

Log-Linear Models in R

Script is my commentary to you. **Bold Courier is what I type in R.** Regular Courier is what R answered.

What is R?

R is a close relative of Splus, but R is available for free. You can download R from

<http://cran.r-project.org/> . R is very powerful and is a favorite (if not the favorite) of statisticians; however, it is not easy to use. It is command driven, not menu driven. You can add things to R that R doesn’t yet know how to do by writing a little program. R gives you fine control over graphics. Most people need a book to help them, and so Mainland & Braun’s book, Data Analysis and Graphics Using R, Cambridge University Press, 2003, is in the book store as an OPTIONAL book.

Who should use R?

If statistics and computers terrify you, stay away from R. On the other hand, if you want a very powerful package for free, one you won’t outgrow, then R worth a try. For some people, getting stuck is a minor challenge, like a cross-word puzzle; people like that like R. For other people, getting stuck is an ulcer; people like that hate R. If you find you need lots of help to install R or make R work, then R isn’t for you.

This is the crabmeat-potato salad data. R is case-sensitive, so crabpot.tab is an object, a table, but Crabpot.tab is an error.

**> crabpot.tab**

, , Illness = Ill

Crabmeat

Potato CM NoCM

PS 120 22

NoPS 4 0

, , Illness = NotIll

Crabmeat

Potato CM NoCM

PS 80 24

NoPS 31 23

This is the way you fit the potato-crabmeat = c(1,2) margin and the potato-illness =c(1,3) margin. Because I did not request any options, all I got back were the likelihood ratio chi square ($lrt), the Pearson chi square ($pearson) and the degrees of freedom ($df). Note carefully the placement of the (); they matter. The form is always loglin(name-of-your-table, list(c(first margin),c(second margin),…,(last margin)), XXXXX) where XXXXX are requests for optional output.

**>loglin(crabpot.tab,list(c(1,2),c(1,3)))**

2 iterations: deviation 0

$lrt

[1] 6.481655

$pearson

[1] 5.094513

$df

[1] 2

$margin

$margin[[1]]

[1] "Potato" "Crabmeat"

$margin[[2]]

[1] "Potato" "Illness"

This is the way you fit the potato-crabmeat = c(1,2) margin and the potato-illness =c(1,3) margin and the crabmeat-illness margin =c(2,3). Because I did not request any options, all I got back were the likelihood ratio chi square ($lrt), the Pearson chi square ($pearson) and the degrees of freedom ($df).

> **loglin(crabpot.tab,list(c(1,2),c(1,3),c(2,3)))**

4 iterations: deviation 0.07563798

$lrt

[1] 2.742749

$pearson

[1] 1.702133

$df

[1] 1

$margin

$margin[[1]]

[1] "Potato" "Crabmeat"

$margin[[2]]

[1] "Potato" "Illness"

$margin[[3]]

[1] "Crabmeat" "Illness"

This is the same fit, with the potato-crabmeat = c(1,2) margin and the potato-illness =c(1,3) margin and the crabmeat-illness margin =c(2,3). Here, I requested optional output, fit=T. That gives the fitted values.

**> loglin(crabpot.tab,list(c(1,2),c(1,3),c(2,3)),fit=T)**

4 iterations: deviation 0.07563798

$lrt

[1] 2.742749

$pearson

[1] 1.702133

$df

[1] 1

$margin

$margin[[1]]

[1] "Potato" "Crabmeat"

$margin[[2]]

[1] "Potato" "Illness"

$margin[[3]]

[1] "Crabmeat" "Illness"

$fit

, , Illness = Ill

Crabmeat

Potato CM NoCM

PS 121.084527 20.916258

NoPS 2.915473 1.083742

, , Illness = NotIll

Crabmeat

Potato CM NoCM

PS 78.922805 25.071857

NoPS 32.077195 21.928143

This is the same fit, with the potato-crabmeat = c(1,2) margin and the potato-illness =c(1,3) margin and the crabmeat-illness margin =c(2,3). Here, I requested optional output, param=T. That gives parameter estimates.

> **loglin(crabpot.tab,list(c(1,2),c(1,3),c(2,3)),param=T)**

4 iterations: deviation 0.07563798

$lrt

[1] 2.742749

$pearson

[1] 1.702133

$df

[1] 1

$margin

$margin[[1]]

[1] "Potato" "Crabmeat"

$margin[[2]]

[1] "Potato" "Illness"

$margin[[3]]

[1] "Crabmeat" "Illness"

$param

$param$"(Intercept)"

[1] 2.8917

$param$Potato

PS NoPS

0.965108 -0.965108

$param$Crabmeat

CM NoCM

0.5340841 -0.5340841

$param$Illness

Ill NotIll

-0.6448331 0.6448331

$param$Potato.Crabmeat

Crabmeat

Potato CM NoCM

PS 0.1915874 -0.1915874

NoPS -0.1915874 0.1915874

$param$Potato.Illness

Illness

Potato Ill NotIll

PS 0.706533 -0.706533

NoPS -0.706533 0.706533

$param$Crabmeat.Illness

Illness

Crabmeat Ill NotIll

CM 0.1523095 -0.1523095

NoCM -0.1523095 0.1523095

> **names(dimnames(crabpot.tab))**

[1] "Potato" "Crabmeat" "Illness"

You can refer to the margins of a table by name rather than by number.

>**loglin(crabpot.tab,list(c("Potato","Crabmeat"),c("Potato","Illness")))**

2 iterations: deviation 0

$lrt

[1] 6.481655

$pearson

[1] 5.094513

$df

[1] 2

$margin

$margin[[1]]

[1] "Potato" "Crabmeat"

$margin[[2]]

[1] "Potato" "Illness"

There are several ways to enter a contingency table in R. This is one way. Use the command “array”. Do help(array) to learn how array works. Essentially, it is array(c(counts),c(dimensions-of-table),list(c(labels for dimension 1),…,c(labels for last dimension))). The example below is from the Srping 2005 final exam for Stat 501 in your bulk pack. I made up the name binge.tab, but you can make up any name you want. I said binge.tab <- something, which creates a new object called binge.tab, defined to be equal to something. The something is an array. The counts go down the first column, then down the second. The first variable is the one changing fastest, namely DrinkW. The third variable is the one changing slowest, namely CC. You have to be very, very, very careful to make sure that the order of the numbers agrees with the order of the variables.

**>binge.tab<-array(c(1912,191,42,110,82,134,150,22,8,18,20,34),c(3,2,2), dimnames=list(c("0 to 6","7 to 13","14+"), c("0 to 1","2+"), c("No","Yes")))**

Usually, you want to give names to the variables too, and this is how you do that.

> **names(dimnames(binge.tab))<-c("DrinkW","BingeM","CC")**

You have just created the following object.

> **binge.tab**

, , CC = No

BingeM

DrinkW 0 to 1 2+

0 to 6 1912 110

7 to 13 191 82

14+ 42 134

, , CC = Yes

BingeM

DrinkW 0 to 1 2+

0 to 6 150 18

7 to 13 22 20

14+ 8 34

> **loglin(binge.tab,list(c(1,2),c(3)))**

2 iterations: deviation 2.842171e-14

$lrt

[1] 44.75657

$pearson

[1] 53.30375

$df

[1] 5

$margin

$margin[[1]]

[1] "DrinkW" "BingeM"

$margin[[2]]

[1] "CC"

You can always ask for help. Type help(command-name). Sometimes, when you ask R for help, it helps you. Sometimes not. Life is like that. Or, at least, R is like that. This is what happens when you ask for help about logliin.

**>help(loglin)**

loglin package:stats R Documentation

Fitting Log-Linear Models

Description:

'loglin' is used to fit log-linear models to multidimensional

contingency tables by Iterative Proportional Fitting.

Usage:

loglin(table, margin, start = rep(1, length(table)), fit = FALSE,

eps = 0.1, iter = 20, param = FALSE, print = TRUE)

Arguments:

table: a contingency table to be fit, typically the output from

'table'.

margin: a list of vectors with the marginal totals to be fit.

(Hierarchical) log-linear models can be specified in terms of

these marginal totals which give the "maximal" factor subsets

contained in the model. For example, in a three-factor

model, 'list(c(1, 2), c(1, 3))' specifies a model which

contains parameters for the grand mean, each factor, and the

1-2 and 1-3 interactions, respectively (but no 2-3 or 1-2-3

interaction), i.e., a model where factors 2 and 3 are

independent conditional on factor 1 (sometimes represented as

'[12][13]').

The names of factors (i.e., 'names(dimnames(table))') may be

used rather than numeric indices.

start: a starting estimate for the fitted table. This optional

argument is important for incomplete tables with structural

zeros in 'table' which should be preserved in the fit. In

this case, the corresponding entries in 'start' should be

zero and the others can be taken as one.

fit: a logical indicating whether the fitted values should be

returned.

eps: maximum deviation allowed between observed and fitted

margins.

iter: maximum number of iterations.

param: a logical indicating whether the parameter values should be

returned.

print: a logical. If 'TRUE', the number of iterations and the final

deviation are printed.

Details:

The Iterative Proportional Fitting algorithm as presented in

Haberman (1972) is used for fitting the model. At most 'iter'

iterations are performed, convergence is taken to occur when the

maximum deviation between observed and fitted margins is less than

'eps'. All internal computations are done in double precision;

there is no limit on the number of factors (the dimension of the

table) in the model.

Assuming that there are no structural zeros, both the Likelihood

Ratio Test and Pearson test statistics have an asymptotic

chi-squared distribution with 'df' degrees of freedom.

Package 'MASS' contains 'loglm', a front-end to 'loglin' which

allows the log-linear model to be specified and fitted in a

formula-based manner similar to that of other fitting functions

such as 'lm' or 'glm'.

Value:

A list with the following components.

lrt: the Likelihood Ratio Test statistic.

pearson: the Pearson test statistic (X-squared).

df: the degrees of freedom for the fitted model. There is no

adjustment for structural zeros.

margin: list of the margins that were fit. Basically the same as the

input 'margin', but with numbers replaced by names where

possible.

fit: An array like 'table' containing the fitted values. Only

returned if 'fit' is 'TRUE'.

param: A list containing the estimated parameters of the model. The

"standard" constraints of zero marginal sums (e.g., zero row

and column sums for a two factor parameter) are employed.

Only returned if 'param' is 'TRUE'.

Author(s):

Kurt Hornik

References:

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) \_The New S

Language\_. Wadsworth & Brooks/Cole.

Haberman, S. J. (1972) Log-linear fit for contingency

tables-Algorithm AS51. \_Applied Statistics\_, \*21\*, 218-225.

Agresti, A. (1990) \_Categorical data analysis\_. New York: Wiley.

See Also:

'table'

Examples:

## Model of joint independence of sex from hair and eye color.

fm <- loglin(HairEyeColor, list(c(1, 2), c(1, 3), c(2, 3)))

fm

1 - pchisq(fm$lrt, fm$df)

## Model with no three-factor interactions fits well.

You can use array to enter a contingency table.

> **help(array)**

array package:base R Documentation

Multi-way Arrays

Description:

Creates or tests for arrays.

Usage:

array(data = NA, dim = length(data), dimnames = NULL)

as.array(x)

is.array(x)

Arguments:

data: a vector (including a list) giving data to fill the array.

dim: the dim attribute for the array to be created, that is a

vector of length one or more giving the maximal indices in

each dimension.

dimnames: the names for the dimensions. This is a list with one

component for each dimension, either NULL or a character

vector of the length given by 'dim' for that dimension. The

list can be names, and the names will be used as names for

the dimensions.

x: an R object.

Value:

'array' returns an array with the extents specified in 'dim' and

naming information in 'dimnames'. The values in 'data' are taken

to be those in the array with the leftmost subscript moving

fastest. If there are too few elements in 'data' to fill the

array, then the elements in 'data' are recycled. If 'data' has

length zero, 'NA' of an appropriate type is used for atomic

vectors ('0' for raw vectors) and 'NULL' for lists.

'as.array()' coerces its argument to be an array by attaching a

'dim' attribute to it. It also attaches 'dimnames' if 'x' has

'names'. The sole purpose of this is to make it possible to access

the 'dim'[names] attribute at a later time.

'is.array' returns 'TRUE' or 'FALSE' depending on whether its

argument is an array (i.e., has a 'dim' attribute) or not. It is

generic: you can write methods to handle specific classes of

objects, see InternalMethods.

References:

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) \_The New S

Language\_. Wadsworth & Brooks/Cole.

See Also:

'aperm', 'matrix', 'dim', 'dimnames'.

Examples:

dim(as.array(letters))

array(1:3, c(2,4)) # recycle 1:3 "2 2/3 times"

# [,1] [,2] [,3] [,4]

#[1,] 1 3 2 1

#[2,] 2 1 3 2

**Creating a Contingency Table in R from a Vector of Counts**

This concerns the mechanics of creating a contingency table from a vector of counts. You use “array”. The example is from the Spring 2008 Final. The counts are in d2. In array, you tell it the dimensions, here 2x3x2, and the dimnames, that is, the levels of the variables. As a second step, using “names(dimnames())” you tell it the names of the variables.

> **help(array)**

> **d2**

653 1516 4307 8963 331 884 27 78 176 592 53 136

> **TurnCrash <- array(data=d2,**

**dim=c(2,3,2),dimnames=list(c("KSI","Other"),**

**c("Uncon","Sign","Signal"),c("A","B")))**

> **TurnCrash**

, , A

Uncon Sign Signal

KSI 653 4307 331

Other 1516 8963 884

, , B

Uncon Sign Signal

KSI 27 176 53

Other 78 592 136

> **names(dimnames(TurnCrash))<-c("Injury","Control","CrashType")**

> **TurnCrash**

, , CrashType = A

Control

Injury Uncon Sign Signal

KSI 653 4307 331

Other 1516 8963 884

, , CrashType = B

Control

Injury Uncon Sign Signal

KSI 27 176 53

Other 78 592 136

**2x2 Tables in R**

**(Many ways of doing one thing)**

> **sdsyaf**

SDS YAF

Auth 29 33

Dem 131 78

Who joins the SDS? Two binomials in the rows.

> **29/(29+33)**

[1] 0.4677419

> **131/(131+78)**

[1] 0.6267943

Compares SDS membership for Auth and Dem homes:

> prop.test(sdsyaf)

2-sample test for equality of proportions with continuity correction

data: sdsyaf

X-squared = 4.3659, df = 1, p-value = 0.03666

alternative hypothesis: two.sided

95 percent confidence interval:

-0.309954304 -0.008150341

sample estimates:

prop 1 prop 2

0.4677419 0.6267943

Who came from an authoritarian home? Two binomials in the columns.

> **29/(29+131)**

[1] 0.18125

> **33/(33+78)**

[1] 0.2972973

To interchange rows and columns, use transpose t(.)

> **t(sdsyaf)**

Auth Dem

SDS 29 131

YAF 33 78

Compares SDS membership for Auth and Dem homes:

> **prop.test(t(sdsyaf))**

2-sample test for equality of proportions with continuity correction

data: t(sdsyaf)

X-squared = 4.3659, df = 1, p-value = 0.03666

alternative hypothesis: two.sided

95 percent confidence interval:

-0.227565559 -0.004529036

sample estimates:

prop 1 prop 2

0.1812500 0.2972973

**2x2 Tables in R, Continued**

> **sum(sdsyaf)**

[1] 271

Who done what? One multinomial.

> **sdsyaf/271**

SDS YAF

Auth 0.1070111 0.1217712

Dem 0.4833948 0.2878229

> **chisq.test(sdsyaf)**

Pearson's Chi-squared test with Yates' continuity correction

data: sdsyaf

X-squared = 4.3659, df = 1, p-value = 0.03666

Chi Square test compares observed and “expected” = fitted counts under independence

> **chisq.test(sdsyaf)$expected**

SDS YAF

Auth 36.60517 25.39483

Dem 123.39483 85.60517

Fitted counts have the same total as observed counts

> **sum(chisq.test(sdsyaf)$expected)**

[1] 271

But the sample proportions satisfy independence

> **chisq.test(sdsyaf)$expected/271**

SDS YAF

Auth 0.1350744 0.09370787

Dem 0.4553315 0.31588622

A key idea: the odds ratio

> **sdsyaf**

SDS YAF

Auth 29 33

Dem 131 78

Kids from authoritarian homes are half as likely to join the SDS

> **(29\*78)/(131\*33)**

[1] 0.5232477

Kids from democratic homes are twice as likely to join the SDS

> **(131\*33)/(29\*78)**

[1] 1.911141

> **fisher.test(sdsyaf)**

Fisher's Exact Test for Count Data

data: sdsyaf

p-value = 0.02809

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

0.2833216 0.9659237

sample estimates:

odds ratio

0.5245336

**Loglinear Model for a 2x2 Table**

(Big weapon, little target!)

xij

> **sdsyaf**

SDS YAF

Auth 29 33

Dem 131 78

Independence model

> **loglin(sdsyaf,list(1,2),fit=T,param=T)**

2 iterations: deviation 0

$lrt

[1] 4.937627

The familiar chi square

$pearson

[1] 5.002007

Degrees of freedom

$df

[1] 1

$margin

$margin[[1]]

[1] 1

$margin[[2]]

[1] 2

Estimates of mij = E(xij)

$fit

SDS YAF

Auth 36.60517 25.39483

Dem 123.39483 85.60517

Estimates of model parameters

log(mij) = u + u1(i) + u2(j)

$param

$param$"(Intercept)"

[1] 4.024968

$param$"1"

Auth Dem

-0.6075999 0.6075999

$param$"2"

SDS YAF

0.1828218 -0.1828218

Saturated model

> **loglin(sdsyaf,list(c(1,2)),fit=T,param=T)**

2 iterations: deviation 0

$lrt

[1] 0

The familiar chi square

$pearson

[1] 0

$df

[1] 0

$margin

$margin[[1]]

[1] 1 2

Estimates of mij = E(xij)

$fit

SDS YAF

Auth 29 33

Dem 131 78

Estimates of model parameters

log(mij) = u + u1(i) + u2(j) + u12(ij)

$param

$param$"(Intercept)"

[1] 4.023927

$param$"1"

Auth Dem

-0.5920257 0.5920257

$param$"2"

SDS YAF

0.0973192 -0.0973192

$param$"1.2"

SDS YAF

Auth -0.1619251 0.1619251

Dem 0.1619251 -0.1619251

**Hierarchical Models Preserve Marginal Totals**

Our data

> crabpot.tab

, , Illness = Ill

Crabmeat

Potato CM NoCM

PS 120 22

NoPS 4 0

, , Illness = NotIll

Crabmeat

Potato CM NoCM

PS 80 24

NoPS 31 23

Fit independence of illness and save fitted counts in “fit”

> fit<-loglin(crabpot.tab,list(c(1,2),3),fit=T)$fit

2 iterations: deviation 2.842171e-14

Our fitted counts

> fit

, , Illness = Ill

Crabmeat

Potato CM NoCM

PS 96.05263 22.09211

NoPS 16.80921 11.04605

, , Illness = NotIll

Crabmeat

Potato CM NoCM

PS 103.94737 23.90789

NoPS 18.19079 11.95395

Margin from data

> apply(crabpot.tab,c(1,2),sum)

Crabmeat

Potato CM NoCM

PS 200 46

NoPS 35 23

Margin from fit -- they are equal because 12 is in model

> apply(fit,c(1,2),sum)

Crabmeat

Potato CM NoCM

PS 200 46

NoPS 35 23

Margin from data

> apply(crabpot.tab,c(1,3),sum)

Illness

Potato Ill NotIll

PS 142 104

NoPS 4 54

Margin from fit – they are unequal because 13 is not in model

> apply(fit,c(1,3),sum)

Illness

Potato Ill NotIll

PS 118.14474 127.85526

NoPS 27.85526 30.14474

**Breast Cancer Survival**

Based on Morrison, et al. (1973) International Journal of Cancer, 11, 261-267.

> cancer.tab

, , Age = 50, Center = Boston

appear

Surv3y Benign Malignant

Alive 24 15

Dead 7 12

, , Age = 60, Center = Boston

appear

Surv3y Benign Malignant

Alive 61 28

Dead 22 11

, , Age = 70, Center = Boston

appear

Surv3y Benign Malignant

Alive 27 16

Dead 18 12

, , Age = 50, Center = Glamorgn

appear

Surv3y Benign Malignant

Alive 21 24

Dead 7 19

, , Age = 60, Center = Glamorgn

appear

Surv3y Benign Malignant

Alive 43 37

Dead 12 17

, , Age = 70, Center = Glamorgn

appear

Surv3y Benign Malignant

Alive 12 16

Dead 7 6

, , Age = 50, Center = Tokyo

appear

Surv3y Benign Malignant

Alive 77 51

Dead 10 13

, , Age = 60, Center = Tokyo

appear

Surv3y Benign Malignant

Alive 51 38

Dead 11 20

, , Age = 70, Center = Tokyo

appear

Surv3y Benign Malignant

Alive 7 6

Dead 3 3

> **help(margin.table)**

> **margin.table(cancer.tab,margin=c(1,4))**

Center

Surv3y Boston Glamorgn Tokyo

Alive 171 153 230

Dead 82 68 60

> **loglin(cancer.tab,list(c(1,2,3),c(2,3,4)))**

2 iterations: deviation 2.842171e-14

$lrt

[1] 16.46446

$pearson

[1] 16.30529

$df

[1] 12

$margin

$margin[[1]]

[1] "Surv3y" "appear" "Age"

$margin[[2]]

[1] "appear" "Age" "Center"

> **loglin(cancer.tab,list(c(1,2,3),c(2,3,4),c(1,4)))**

4 iterations: deviation 0.04292951

$lrt

[1] 9.130212

$pearson

[1] 9.142773

$df

[1] 10

$margin

$margin[[1]]

[1] "Surv3y" "appear" "Age"

$margin[[2]]

[1] "appear" "Age" "Center"

$margin[[3]]

[1] "Surv3y" "Center"

> **16.46446-9.130212**

[1] 7.334248

> **12-10**

[1] 2

> **1-pchisq(7.334248,2)**

[1] 0.02554985

**A 24 Table: How Are Symptoms Related?**

A 2x2x2x2 table recording 4 psychiatric symptoms. Originally from Coppen, A (1966) The Mark-Nyman temperament scale: an English translation, British Journal of Medical Psychology, 33, 55-59; used as an example in Wermuth (1976) Model search in multiplicative models, Biometrics 32, 253-263.

> **symptoms.tab**

, , Stability = introvert, Depression = depressed

Validity

Solidity energetic psychasthenic

rigid 15 30

hysteric 9 32

, , Stability = extrovert, Depression = depressed

Validity

Solidity energetic psychasthenic

rigid 23 22

hysteric 14 16

, , Stability = introvert, Depression = not depressed

Validity

Solidity energetic psychasthenic

rigid 25 22

hysteric 46 27

, , Stability = extrovert, Depression = not depressed

Validity

Solidity energetic psychasthenic

rigid 14 8

hysteric 47 12

Independence is a poor fit

> **loglin(symptoms.tab,list(1,2,3,4))**

2 iterations: deviation 2.842171e-14

$lrt

[1] 68.89475

$df

[1] 11

Constant association is a plausible fit

>**loglin(symptoms.tab,list(c(1,2),c(1,3),c(1,4),c(2,3),c(2,4),c(3,4)))**

5 iterations: deviation 0.02118533

$lrt

[1] 8.476963

$df

[1] 5

> 1-pchisq(8.477,5)

[1] 0.131833

**A 24 Table, Continued: Looking for a Simpler Model**

Remove (3,4): fit is poor

> 1-pchisq(loglin(symptoms.tab,list(c(1,2),c(1,3),c(1,4),c(2,3),c(2,4)))$lrt,6)

4 iterations: deviation 0.02907685

[1] 0.0348279

Remove (2,4): fit is poor

> 1-pchisq(loglin(symptoms.tab,list(c(1,2),c(1,3),c(1,4),c(2,3),c(3,4)))$lrt,6)

3 iterations: deviation 0.01097763

[1] 1.416786e-06

Remove (2,3): fit is poor

> 1-pchisq(loglin(symptoms.tab,list(c(1,2),c(1,3),c(1,4),c(2,4),c(3,4)))$lrt,6)

3 iterations: deviation 0.05751916

[1] 0.001317691

Remove (1,4): fit is poor

> 1-pchisq(loglin(symptoms.tab,list(c(1,2),c(1,3),c(2,3),c(2,4),c(3,4)))$lrt,6)

4 iterations: deviation 0.08614444

[1] 0.0007099575

Remove (1,3): fit is ok

> 1-pchisq(loglin(symptoms.tab,list(c(1,2),c(1,4),c(2,3),c(2,4),c(3,4)))$lrt,6)

4 iterations: deviation 0.09620685

[1] 0.1892320

Remove (1,2): fit is ok

> 1-pchisq(loglin(symptoms.tab,list(c(1,3),c(1,4),c(2,3),c(2,4),c(3,4)))$lrt,6)

5 iterations: deviation 0.01863555

[1] 0.1930569

Remove (1,2), (1,3): fit is ok

> 1-pchisq(loglin(symptoms.tab,list(c(1,4),c(2,3),c(2,4),c(3,4)))$lrt,7)

4 iterations: deviation 0.0846266

[1] 0.2505973

Remove (1,4): fit is poor

> 1-pchisq(loglin(symptoms.tab,list(c(2,3),c(2,4),c(3,4)))$lrt,8)

4 iterations: deviation 0.0846266

[1] 0.0001212491

Remove (2,3): fit is poor

> 1-pchisq(loglin(symptoms.tab,list(c(1,4),c(2,4),c(3,4)))$lrt,8)

2 iterations: deviation 2.842171e-14

[1] 0.004141092

Remove (2,4): fit is poor

> 1-pchisq(loglin(symptoms.tab,list(c(1,4),c(2,3),c(3,4)))$lrt,8)

2 iterations: deviation 1.421085e-14

[1] 2.626477e-06

Remove (3,4): fit is poor

> 1-pchisq(loglin(symptoms.tab,list(c(1,4),c(2,3),c(2,4)))$lrt,8)

2 iterations: deviation 2.842171e-14

[1] 0.0852286

**See-Buy Data**

> **seebuy**

, , 1st Buy = Buy1, 1st See = See1

2nd See

2nd Buy See2 NoSee2

Buy2 83 35

NoBuy2 8 7

, , 1st Buy = NoBuy1, 1st See = See1

2nd See

2nd Buy See2 NoSee2

Buy2 22 11

NoBuy2 68 28

, , 1st Buy = Buy1, 1st See = NoSee1

2nd See

2nd Buy See2 NoSee2

Buy2 25 95

NoBuy2 10 15

, , 1st Buy = NoBuy1, 1st See = NoSee1

2nd See

2nd Buy See2 NoSee2

Buy2 8 6

NoBuy2 32 493

Only saturated model fits:

> **loglin(seebuy,list(c(1,2,3),c(1,2,4),c(1,3,4),c(2,3,4)))**

9 iterations: deviation 0.06642452

$lrt

[1] 21.87360

$df

[1] 1

Didn’t see, Didn’t buy:

> **or(seebuy[,,2,2])**

[1] 20.54167

Saw, bought:

> **or(seebuy[,,1,1])**

[1] 2.075

Saw, didn’t buy:

> **or(seebuy[,,2,1])**

[1] 0.8235294

Didn’t see, bought:

> **or(seebuy[,,1,2])**

[1] 0.3947368

**Log-Linear Models with Structural Zeros in R**

To fit a log-linear model with structural zeros, you use the “start=” option in loglin:

loglin(table, margin, start = rep(1, length(table)), fit = FALSE,

eps = 0.1, iter = 20, param = FALSE, print = TRUE)

start: a starting estimate for the fitted table. This optional

argument is important for incomplete tables with structural

zeros in 'table' which should be preserved in the fit. In

this case, the corresponding entries in 'start' should be

zero and the others can be taken as one.

This is from page I1, Table 5.2-3 “Classification of Purum Marriages,” from Bishop, Fienberg and Holland, *Discrete Multivariate Analysis*.

> Marriages

Marrim Makan Parpa Thao Keyang

Marrim 0 5 17 0 6

Makan 5 0 0 16 2

Parpa 0 2 0 10 11

Thao 10 0 0 0 9

Keyang 6 20 8 0 1

Some of the zeros are “random” (meaning “didn’t happen”) and others are “structural” (meaning “can’t happen”). In the start table, you put a 0 for the structural zeros and a 1 for everything else, including the random zeros.

> MarriagesS

Marrim Makan Parpa Thao Keyang

Marrim 0 1 1 0 1

Makan 1 0 1 1 1

Parpa 0 1 0 1 1

Thao 1 0 0 0 1

Keyang 1 1 1 1 1

**This fits “quasi-independence:”**

> loglin(Marriages,list(1,2),start=MarriagesS,fit=T)

5 iterations: deviation 0.09330366

$lrt

[1] 76.2508

$pearson

[1] 66.54045

$df

[1] 16

$margin

$margin[[1]]

[1] 1

$margin[[2]]

[1] 2

$fit

Marrim Makan Parpa Thao Keyang

Marrim 0.000000 10.786776 10.682608 0.000000 6.548952

Makan 4.866051 0.000000 6.562317 7.543237 4.023015

Parpa 0.000000 8.382528 0.000000 9.542478 5.089266

Thao 10.383451 0.000000 0.000000 0.000000 8.584534

Keyang 5.750498 7.830696 7.755075 8.914285 4.754233

**This fit satisfies quasi-independence**, for intstance:

> mfit<-loglin(Marriages,list(1,2),start=MarriagesS,fit=T)$fit

5 iterations: deviation 0.09330366

**The odds ratio is 1 in any complete piece**:

> mfit[2:3,4:5]

Thao Keyang

Makan 7.543237 4.023015

Parpa 9.542478 5.089266

> mfit[2,4]\*mfit[3,5]/(mfit[3,4]\*mfit[2,5])

[1] 1

**Log-Linear and Logit Models for the Same Data**

We look at this data set before: a 2x2x2x2 table recording 4 psychiatric symptoms. Originally from Coppen, A (1966) The Mark-Nyman temperament scale: an English translation, British Journal of Medical Psychology, 33, 55-59; used as an example in Wermuth (1976) Model search in multiplicative models, Biometrics 32, 253-263.

> **symptoms.tab**

, , Stability = introvert, Depression = depressed

Validity

Solidity energetic psychasthenic

rigid 15 30

hysteric 9 32

, , Stability = extrovert, Depression = depressed

Validity

Solidity energetic psychasthenic

rigid 23 22

hysteric 14 16

, , Stability = introvert, Depression = not depressed

Validity

Solidity energetic psychasthenic

rigid 25 22

hysteric 46 27

, , Stability = extrovert, Depression = not depressed

Validity

Solidity energetic psychasthenic

rigid 14 8

hysteric 47 12

If you recall, we came to like the model in which variable 1, Solidity was conditionally independent of variables 2 and 3, Validity and Stability, given variable 4, Depression:

> 1-pchisq(loglin(symptoms.tab,list(c(1,4),c(2,3),c(2,4),c(3,4)))$lrt,7)

4 iterations: deviation 0.0846266

[1] 0.2505973

This is an nice model in some ways, but not one that you could fit with a standard logit model. In a standard logit model, one of the variables is a binary “dependent variable” and the others are “independent variables.” There are many kinds of logit models, but this is the typical kind. Let’s take variable 4, Depression as the dependent variable, and predict it from the other symptoms. This means we only fit models which preserve the relationships among the independent variables, 1,2 and 3, and only model relationships that involve the dependent variable. This means we always include the c(1,2,3) term. This means we can’t discover a simple relationship among the independent variables, because we have declared we are not interested in such things; we are only interested in predicting depression from other symptoms.

> loglin(symptoms.tab,list(c(1,2,3),c(1,4),c(2,4),c(3,4)))

5 iterations: deviation 0.03387941

$lrt $df

[1] 7.762877 4

$margin[[1]]

[1] "Solidity" "Validity" "Stability"

$margin[[2]]

[1] "Solidity" "Depression"

$margin[[3]]

[1] "Validity" "Depression"

$margin[[4]]

[1] "Stability" "Depression"

**Log-Linear and Logit Models, Continued**

For the logit model, we rearrange the data to look more like regression, with coded variables, etc.

> **symptoms.X**

depressed notdepressed Solidity Validity Stability

1 15 25 0.5 0.5 0.5

2 9 46 -0.5 0.5 0.5

3 30 22 0.5 -0.5 0.5

4 32 27 -0.5 -0.5 0.5

5 23 14 0.5 0.5 -0.5

6 14 47 -0.5 0.5 -0.5

7 22 8 0.5 -0.5 -0.5

8 16 12 -0.5 -0.5 -0.5

> **y<-as.matrix(symptoms.X[,1:2])**

> **y**

depressed notdepressed

1 15 25

2 9 46

3 30 22

4 32 27

5 23 14

6 14 47

7 22 8

8 16 12

> **attach(symptoms.X)**

The glm program fits many types of “generalized” linear models. If you say, “family=binomial” it fits a logit model.

> **summary(glm(y~Solidity+Validity+Stability,family=binomial))**

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.1197 0.1148 -1.043 0.296957

Solidity 0.8658 0.2276 3.804 0.000142 \*\*\*

Validity -1.2212 0.2330 -5.242 1.59e-07 \*\*\*

Stability -0.5246 0.2351 -2.231 0.025655 \*

---

Null deviance: 55.4057 on 7 degrees of freedom

Residual deviance: 7.7629 on 4 degrees of freedom

AIC: 48.304

Number of Fisher Scoring iterations: 3

Although it may not look like it, this is actually the “same” model. For instance, the likelihood ratio chi square from the log-linear model was

$lrt $df

[1] 7.762877 4

which is the same as the “residual deviance” from the logit model

Residual deviance: 7.7629 on 4 degrees of freedom

Similarly, the “Null deviance” in the logit model is

Null deviance: 55.4057 on 7 degrees of freedom

which is the same as the likelihood ratio chi square from the loglinear model which says Depression (#4) is independent of the other three variables, but the other three variables can have any relationship.

> loglin(symptoms.tab,list(c(1,2,3),4))

2 iterations: deviation 2.842171e-14

$lrt $df

[1] 55.40572 7

**Log-Linear and Logit Models, Fitted Values**

They also give the same “fitted values” or fitted probabilities of degression:

> **glm(y~Solidity+Validity+Stability,family=binomial)$fitted.values**

1 2 3 4 5 6 7 8

**0.3636** 0.1938 0.6595 0.4491 0.4912 0.2888 0.7660 0.5793

We can compute the same thing from the fitted counts for the log-linear model:

> **14.54564/(14.54564+ 25.45283)**

[1] 0.3636549

> **loglin(symptoms.tab,list(c(1,2,3),c(1,4),c(2,4),c(3,4)),fit=T)**

, , Stability = introvert, Depression = depressed

Validity

Solidity energetic psychasthenic

rigid **14.54564** 34.29702

hysteric 10.66283 26.49452

, , Stability = extrovert, Depression = depressed

Validity

Solidity energetic psychasthenic

rigid 18.17336 22.98440

hysteric 17.62069 16.22155

, , Stability = introvert, Depression = not depressed

Validity

Solidity energetic psychasthenic

rigid **25.45283** 17.70196

hysteric 44.34483 32.50037

, , Stability = extrovert, Depression = not depressed

Validity

Solidity energetic psychasthenic

rigid 18.82339 7.021948

hysteric 43.37632 11.778337

log(mhijk) = u + uR(h) + uV(i) + uS(j) + uD(k) + … + uSD(jk)

log(mhij1/mhij2) = log(mhij1)-log(mhij2) = uD(1)- uD(2) + uRD(h1)- uRD(h2)

+ uVD(i1)- uVD(i2)  + uSD(h1)- uSD(j2)

= w + wR(h) + wV(i) + wS(j)

**Log-Linear and Logit Models, Model Parameters**

They also give the “same” parameters. If you type:

> **loglin(symptoms.tab,list(c(1,2,3),c(1,4),c(2,4),c(3,4)),param=T)**

you get many parameters, including:

$param$Solidity.Depression

Depression

Solidity depressed not depressed

rigid **0.2164241** -0.2164241

hysteric -0.2164241 0.2164241

$param$Validity.Depression

Depression

Validity depressed not depressed

energetic **-0.3052297** 0.3052297

psychasthenic 0.3052297 -0.3052297

$param$Stability.Depression

Depression

Stability depressed not depressed

introvert **-0.1310981** 0.1310981

extrovert 0.1310981 -0.1310981

whereas glm gives

> **glm(y~Solidity+Validity+Stability,family=binomial)**

Coefficients:

(Intercept) Solidity Validity Stability

-0.1197 0.8658 -1.2212 -0.5246

but they are the same once you multiply by 4:

> **4\*0.2164241**

[1] 0.8656964

> **4\*-0.3052297**

[1] -1.220919

> **4\*-0.1310981**

[1] -0.5243924

**Fitting Logit Models**

The data are from DC\*MADS which is study #2347 at <http://www.icpsr.umich.edu/> available from the Penn Library web page. These are 986 babies born in Washington DC hospitals. From NIDA. Abuse The DCBaby data.frame includes a few incomplete cases excluded from DCBaby.complete

> **dim(DCBaby)**

[1] 986 8

> **dim(DCBaby.complete)**

[1] 974 8

First two babies:

> **DCBaby[1:2,]**

ID Bweight low15 low25 BWgroup cigs alcoh momage

1 1 2438 0 1 2 0 0 18

2 2 2296 0 1 2 0 0 18

Birth weight appears four ways, in grams (Bweight), as a binary variable <1500 grams or not, as a binary variable <2500 grams or not, and as an ordinal variable <1500 grams, 1500-2500 grams, > 2500 grams as 3, 2 or 1. Notice that 3 is very under weight. Cigs = 1 if mom smoked during pregnancy, = 0 otherwise. Alcoh = 1 if mom drank alcohol during pregnancy, =0 otherwise. Here, use is “at least once a week”. Momage is mom’s age.

> **attach(DCBaby.complete)**

> **table(factor(cigs),factor(alcoh))**

0 1

0 615 104

1 107 148

> **or(table(factor(cigs),factor(alcoh)))**

[1] 8.179367

Mom’s who smoked were 8 times more likely to drink. Older moms smoked and drank somewhat more.

**Fitting Logit Models, continued**

Let’s predict birthweight < 2500 grams using cigs, alcoh, momage:

> **summary(glm(low25~cigs+alcoh+momage,family=binomial))**

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.72283 0.34612 -2.088 0.0368 \*

cigs 1.26114 0.17982 7.013 2.33e-12 \*\*\*

alcoh 0.17240 0.18710 0.921 0.3568

momage -0.03030 0.01309 -2.315 0.0206 \*

---

Null deviance: 1103.1 on 973 degrees of freedom

Residual deviance: 1033.3 on 970 degrees of freedom

In sparse problems, like this one, you **cannot** use G2 to test goodness of fit, but you can compare models. Looks like cigs = 1 smoking is bad, z=7.0, p<0.0001. We can use the estimate of cigs , namely 1.26, to estimate how bad:

> **exp(1.26114)**

[1] 3.529443

So we estimate that the odds of a small baby, <2500 grams, are 3.5 times greater for a mom who smokes. Might want a confidence interval. Build a confidence interval for cigs, then take antilogs. Confidence interval is estimate plus or minus 1.96 x std.error.

> **exp(c(1.26114-1.96\*0.17982,1.26114+1.96\*0.17982))**

[1] 2.481077 5.020790

So our point estimate is 3.5 times, but the 95% confidence interval is [2.5, 5.0] times greater risk of a small baby for smoking moms.

Let’s add some interactions and see if they improve the fit.

> **glm(low25~cigs+alcoh+momage+cigs\*alcoh+cigs\*momage**

**+alcoh\*momage,family=binomial)**

Coefficients:

(Intercept) cigs alcoh momage

-0.47840 -2.13229 2.92436 -0.04002

cigs:alcoh cigs:momage alcoh:momage

0.10041 0.12093 -0.10045

Does the new model fit significantly better than the simpler model? They are nested, so we use the change in G2 = residual deviance to compare the models. Reduced model G2 =1033.3 on 970 degrees of freedom, versus full model G2 = 1018 on 967 because 3 parameters were added.

> **1033.3-1018**

[1] 15.3

> **1-pchisq(15.3,3)**

[1] 0.001577423

So the full model with interactions fits significantly better: The null hypothesis that the three interaction coefficients are all zero is rejected at the 0.0016 level. Look at fit:

> boxplot(p[cigs==1&alcoh==1],p[cigs==1&alcoh==0],

p[cigs==0&alcoh==1],p[cigs==0&alcoh==0], ylab =

"Prob < 2500 grams",names=c("S&D","S","D","Neither"))



Leukemia Logit

> **boxplot(wbc)**

> **boxplot(log(wbc))**

> **summary(glm(year~ag10+log(wbc),family=binomial))**

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 8.0964 4.0537 1.997 0.0458 \*

ag10 2.5196 1.0907 2.310 0.0209 \*

log(wbc) -1.1088 0.4609 -2.405 0.0162 \*

---

Null deviance: 42.010 on 32 degrees of freedom

Residual deviance: 26.833 on 30 degrees of freedom

AIC: 32.833

>**summary(glm(year~ag10+log(wbc)+ag10\*log(wbc),family=binomial))**

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 11.4603 8.9559 1.280 0.201

ag10 -2.0078 10.1840 -0.197 0.844

log(wbc) -1.5039 1.0574 -1.422 0.155

ag10:log(wbc) 0.5181 1.1732 0.442 0.659

Null deviance: 42.010 on 32 degrees of freedom

Residual deviance: 26.615 on 29 degrees of freedom

AIC: 34.615

Number of Fisher Scoring iterations: 6

> **p<-glm(year~ag10+log(wbc),family=binomial)$fitted.values**

> **cbind(leukemia,round(p,3))**

> summary(wbc)

Min. 1st Qu. Median Mean 3rd Qu. Max.

750 5300 10500 29170 32000 100000

> l<-8.096+2.520+(-1.109\*log(5300))

> exp(l)/(1+exp(l))

[1] 0.7513476

> l<-8.096+2.520+(-1.109\*log(32000))

> exp(l)/(1+exp(l))

[1] 0.2914811

> l<-8.096+(-1.109\*log(5300))

> exp(l)/(1+exp(l))

[1] 0.1955744

> l<-8.096+(-1.109\*log(32000))

> exp(l)/(1+exp(l))

[1] 0.03204012

**McNemar Test via Conditional Logit Regression**

> **library(survival)**

> **sartwell[1:8,]**

pair thromb ocuse

1 1 0 1

2 1 1 1

3 2 0 1

4 2 1 1

5 3 0 1

6 3 1 1

7 4 0 1

8 4 1 1

> **dim(sartwell)**

[1] 350 3

> **attach(sartwell)**

> **table(ocuse[thromb==1],ocuse[thromb==0])**

0 1

0 95 13

1 57 10

> **pbinom(13,13+57,1/2)**

[1] 5.144971e-08

> **2\* pbinom(13,13+57,1/2)**

[1] 1.028994e-07

> **57/13**

[1] 4.384615

> **clogit(ocuse~thromb+strata(pair),data=sartwell)**

Call:

clogit(ocuse ~ thromb + strata(pair), data = sartwell)

coef exp(coef) se(coef) z p

thromb 1.48 4.38 0.307 4.81 1.5e-06

Likelihood ratio test=29.9 on 1 df, p=4.67e-08 n= 350

**Conditional Logit Regression**

In this example, there are 59 matched sets, i=1,2,…,59, and each set contains 3 boys, j=1,2,3 who were matched on many variables. In each set, the first boy, boy j=1, joined a gang for the first time at age 14, while the other two boys had not yet joined gangs. (Look at newgang14, which is 1 for the joiner and 0 for the controls.) gang17 indicates whether the boy is in a gang at age 14. vio13 measures violence at age 13, iqC is a rough iq measure, and nbp13 is self-reported # of sexual partners at age 13.

Example adapted from Haviland, et al. (2007) *Psychological Methods*. Data from “Montréal Longitudinal Study of Boys,” Tremblay, R. E., et al (1987), *International Journal of Behavioral Development*, 10, 467-484.

> **dim(gangEG)**

[1] 177 6

> **gangEG[1:21,]**

mset newgang14 gang17 vio13 iqC nbp13

26 235 1 0 0 10 0

212 235 0 NA 1 11 0

298 235 0 0 0 10 0

274 236 1 0 0 11 0

166 236 0 0 0 8 0

304 236 0 0 0 10 0

280 237 1 1 1 7 0

16 237 0 0 1 9 0

487 237 0 0 1 9 0

285 238 1 0 0 11 0

52 238 0 0 0 11 0

275 238 0 0 1 10 0

311 239 1 0 1 11 0

114 239 0 0 1 9 0

362 239 0 0 1 10 0

355 240 1 0 0 11 0

95 240 0 NA 1 10 0

328 240 0 0 0 11 0

357 241 1 1 0 7 NA

218 241 0 0 0 12 0

461 241 0 1 NA 0 NA

**Conditional Logit Regression, Continued**

Will predict gang17 from other variables. Here, pij=Prob(gang17=1).

log{pij/(1-pij)} = i + 1 xij + 2 wij

Notice that each matched set has its own parameter, i, so this model looks tiny, but it has 59+2 = 61 parameters. The conditional logit model eliminates 59 of the parameters by conditioning on the total number of boys in gangs at age 17 in each set, i=1,2,…,59, where that total can be 0, 1, 2, or 3. The model then just has the betas.

> **library(survival)**

> **help(clogit)**

> **clogit(gang17~newgang14+iqC+strata(mset),data=gangEG)**

Call:

clogit(gang17 ~ newgang14 + iqC + strata(mset), data = gangEG)

coef exp(coef) se(coef) z p

newgang14 0.562 1.755 0.522 1.08 0.28

iqC -0.402 0.669 0.196 -2.05 0.04

Likelihood ratio test=6.65 on 2 df, p=0.0360 n=157 (20 observations deleted due to missing)

> **clogit(gang17~newgang14+iqC+vio13+strata(mset),data=gangEG)**

Call:

clogit(gang17 ~ newgang14 + iqC + vio13 + strata(mset), data = gangEG)

coef exp(coef) se(coef) z p

newgang14 0.702 2.017 0.577 1.22 0.220

iqC -0.449 0.638 0.224 -2.01 0.045

vio13 -0.546 0.579 0.355 -1.54 0.120

Likelihood ratio test=8.47 on 3 df, p=0.0373 n=155 (22 observations deleted due to missing)

**Proportional Odds Model for Ordinal Data**

DCBaby contains data on 986 babies born in 8 Washington DC hospitals. > **dim(DCBaby)**

[1] 986 8

This is data for baby 1 and baby 2:

> **DCBaby[1:2,]**

ID Bweight low15 low25 BWgroup cigs alcoh momage

1 1 2438 0 1 2 0 0 18

2 2 2296 0 1 2 0 0 18

Birth weight appears four ways, in grams (Bweight), as a binary variable <1500 grams or not, as a binary variable <2500 grams or not, and as an ordinal variable <1500 grams, 1500-2500 grams, > 2500 grams as 3, 2 or 1. Notice that 3 is very under weight, and 1 is much heavier.

> **table(low15,low25)**

low25

low15 0 1

0 734 195

1 0 55

> **table(low15,BWgroup)**

BWgroup

low15 1 2 3

0 734 195 0

1 0 0 55

> **table(low25,BWgroup)**

BWgroup

low25 1 2 3

0 734 0 0

1 0 195 55

You need to get the polr program in the MASS library.

> **library(MASS)**

> **help(polr)**

To use polr, the outcome, here BWgroup, must be an ordered factor. If BWgroup were entered as 1, 2, 3, it becomes an ordered factor by setting BWgroup<- factor(BWgroup, ordered=T).

**Proportional Odds Model for Ordinal Data, continued**

Ordinal logit fits log{pr(Y < j)/pr(Y>j)} = j –  x simultaneously for all j, whereas a binary logit regression of a similar sort fits log{pr(Y > j)/pr(Y<j)} = j +  x for one j at a time. Notice that there are small changes in the model which mix the signs.

> **glm(low15~cigs,family=binomial)**

Coefficients:

(Intercept) cigs

-3.215 1.064

> **glm(low25~cigs,family=binomial)**

Coefficients:

(Intercept) cigs

-1.484 1.275

> **polr(BWgroup~cigs)**

Coefficients:

cigs

1.252373

Intercepts:

1|2 2|3

1.479582 3.311650

Notice that polr gave you one slope, two intercepts, whereas binary logit regression gave you two of each.

Let’s get the “fitted values” from the ordinal logit model, and round them to 3 decimals.

> **fit<-polr(BWgroup~cigs)$fitted.values**

> **fit<-round(fit,3)**

Let’s add the fitted values to the data set and print the first 7 babies.

> cbind(DCBaby[1:7,],fit[1:7,])

ID Bweight low15 low25 BWgroup cigs alcoh momage 1 2 3

1 1 2438 0 1 2 0 0 18 0.815 0.15 0.035

2 2 2296 0 1 2 0 0 18 0.815 0.15 0.035

3 3 3020 0 0 1 0 0 30 0.815 0.15 0.035

4 4 2385 0 1 2 0 0 17 0.815 0.15 0.035

5 5 3016 0 0 1 0 1 35 0.815 0.15 0.035

6 6 4175 0 0 1 0 0 23 0.815 0.15 0.035

7 7 3725 0 0 1 1 1 27 0.557 0.33 0.113

Look at baby 6 and baby 7. Baby 6 had a mom who did not smoke. Baby 7 had a mom who smoked. So baby 7 had a much higher chance of being in

**Latent Class Model**

> **army.tab**

> **dim(army)**

[1] 1000 4

> **army[1:2,]**

Well run Favorable Square Deal Enlisted

1 1 1 1 1

2 1 1 1 1

> **summary(army)**

Well run Favorable Square Deal Enlisted

Min. :0.000 Min. :0.000 Min. :0.0 Min. :0.000

1st Qu.:0.000 1st Qu.:0.000 1st Qu.:0.0 1st Qu.:0.000

Median :1.000 Median :0.000 Median :0.0 Median :0.000

Mean :0.641 Mean :0.374 Mean :0.3 Mean :0.254

3rd Qu.:1.000 3rd Qu.:1.000 3rd Qu.:1.0 3rd Qu.:1.000

Max. :1.000 Max. :1.000 Max. :1.0 Max. :1.000

> **library(e1071)**

Loading required package: class

> **help(lca)**

> **lca(as.matrix(army),2,niter=100)**

LCA-Result

----------

Datapoints: 1000

Classes: 2

Probability of classes

[1] 0.492 0.508

Itemprobabilities

1 2 3 4

1 0.89 0.59 0.53 0.47

2 0.40 0.16 0.08 0.04

> summary(lca(as.matrix(army),2,niter=100))

LCA-Result

----------

Datapoints: 1000

Classes: 2

Goodness of fit statistics:

Number of parameters, estimated model: 9

Number of parameters, saturated model: 15

Log-Likelihood, estimated model: -2348.906

Log-Likelihood, saturated model: -2343.960

Information Criteria:

BIC, estimated model: 4759.982

BIC, saturated model: 4791.537

TestStatistics:

Likelihood ratio: 9.891697 p-val: 0.1292876

Pearson Chi^2: 9.04448 p-val: 0.171092

Degress of freedom: 6

**Using glm for Poisson Regression**

> **crabpot.tab**

, , Illness = Ill

Crabmeat

Potato CM NoCM

PS 120 22

NoPS 4 0

, , Illness = NotIll

Crabmeat

Potato CM NoCM

PS 80 24

NoPS 31 23

> **loglin(crabpot.tab,list(c(1,2),c(1,3)))**

2 iterations: deviation 0

$lrt

[1] 6.481655

$df

[1] 2

$margin

$margin[[1]]

[1] "Potato" "Crabmeat"

$margin[[2]]

[1] "Potato" "Illness"

> **crabpot.X**

count ill potato crabmeat ip ic cp

1 120 1 1 1 1 1 1

2 4 1 0 1 0 1 0

3 22 1 1 0 1 0 0

4 0 1 0 0 0 0 0

5 80 0 1 1 0 0 1

6 31 0 0 1 0 0 0

7 24 0 1 0 0 0 0

8 23 0 0 0 0 0 0

> glm(count~ill+potato+crabmeat+ip+cp,family=poisson)

Coefficients:

(Intercept) ill potato crabmeat ip cp

3.06404 -2.60269 -0.09633 0.41985 2.91413 1.04982

Degrees of Freedom: 7 Total (i.e. Null); 2 Residual

Null Deviance: 295.3

Residual Deviance: 6.482 AIC: 54.81

**Measuring Agreement Using Kappa**

Flip a dime and a quarter and they agree with probability ½:

> ((1/2)^2)+((1/2)^2)

[1] 0.5

> dime<-sample(c("head","tail"),10000,replace=T)

> quarter<-sample(c("head","tail"),10000,replace=T)

> table(dime,quarter)

quarter

dime head tail

head 2564 2467

tail 2506 2463

> (2564+2463)/10000

[1] 0.5027

So agreeing half the time does not mean much.

Roll a red die and a blue die and record 1 or other. They agree with probability:

> ((1/6)^2)+((5/6)^2)

[1] 0.7222222

> red<-rbinom(10000,1,1/6)

> blue<-rbinom(10000,1,1/6)

> table(red,blue)

blue

red 0 1

0 6950 1327

1 1428 295

> (295+6950)/10000

[1] 0.7245

So .72 agreement does not mean much, nor does the fact that dice agree more than coins – it is just luck.

Cohen’s kappa asks about agreement above chance.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Fleiss, J. L., et al. (2003) Statistical Methods for Rates and Proportions. NY: Wiley.

Find the expected counts under independence:

> chisq.test(table(red,blue))$expected

blue

red 0 1

0 6934.471 1342.5294

1 1443.529 279.4706

> (6934.471+279.4706)/10000

[1] 0.7213942

So although we got 72% agreement, we also expected 72% agreement by chance.

Kappa is the percent agreement in excess of change,

(actual – expected)/(1 – expected)

> (0.7245-0.7213942)/(1-0.7213942)

[1] 0.01114765

So it is just 1% better than expected by chance.

> library(irr)

> help(package=irr)

> kappa2(cbind(red,blue))

Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 10000

Raters = 2

Kappa = 0.0111

z = 1.12

p-value = 0.265

> kappa2(cbind(dime,quarter))

Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 10000

Raters = 2

Kappa = 0.00531

z = 0.531

p-value = 0.595

Erie County Ohio

> attach(erieAgree)

> kappa2(cbind(MayFrom,MayTo))

Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 445

Raters = 2

Kappa = 0.756

z = 22.5

p-value = 0

> kappa2(cbind(JuneFrom,JuneTo))

Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 445

Raters = 2

Kappa = 0.762

z = 22.7

p-value = 0

> kappa2(cbind(JulyFrom,JulyTo))

Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 445

Raters = 2

Kappa = 0.692

z = 21.3

p-value = 0

> kappa2(cbind(AugFrom,AugTo))

Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 445

Raters = 2

Kappa = 0.864

z = 25.3

p-value = 0

> kappa2(cbind(SeptFrom,SeptTo))

Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 445

Raters = 2

Kappa = 0.872

z = 25.2

p-value = 0

So people changed least in August and Sept, and most from July to August.

**Is There a One-Dimensional Latent Variable?**

If there is a one-dimensional latent variable positively related to all the variables, then the partial association between any two variables, say S and E, given the sum of all other variables, here WF12, is positive – that is, the odds ration is at least 1.

> **table(S,E,WF12)**

, , WF12 = 0

E

S 0 1

0 229 16

1 25 10

, , WF12 = 1

E

S 0 1

0 251 53

1 76 45

, , WF12 = 2

E

S 0 1

0 96 55

1 69 75

> **table(S,E,WF12)[,,1]**

E

S 0 1

0 229 16

1 25 10

> **or(table(S,E,WF12)[,,1])**

[1] 5.725

> **table(S,E,WF12)[,,2]**

E

S 0 1

0 251 53

1 76 45

> **or(table(S,E,WF12)[,,2])**

[1] 2.804121

> **table(S,E,WF12)[,,3]**

E

S 0 1

0 96 55

1 69 75

> **or(table(S,E,WF12)[,,3])**

[1] 1.897233

> **mantelhaen.test(S,E,WF12)**

Mantel-Haenszel chi-squared test with continuity correction

data: S and E and WF12

Mantel-Haenszel X-squared = 33.6202, df = 1, p-value = 6.7e-09

alternative hypothesis: true common odds ratio is not equal to 1

95 percent confidence interval:

1.819444 3.404577

sample estimates:

common odds ratio

2.488863

**Statistics 501, Spring 2015, Midterm: Data Page #1**

**This is an exam. Do not discuss it with anyone**. If you discuss the exam in any way with anyone, then you have cheated on the exam. Cheating on an exam is the dumbest thing a PhD student at Penn can do.

Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question.

The data are from a paper by Chen Zhijian et al. (2006), Evaluating the genotoxic effects of workers exposed to lead using micronucleus assay, comet assay and TCR gene mutation test, *Toxicology*, 223(3), 219-226. There is no need to consult the paper unless you want to. The data are in an object, **storagebattery**, in the course workspace. You will need to download the workspace again. If it is not there, then you need to clear your browser’s memory and download again. The first few lines are below.

Alternatively, there is a csv file of the data on my webpage at the link **data.csv**.

> head(storagebattery)

pair wsex csex wage cage wyears wlead clead wmnr cmnr wmtm cmtm wmftcr cmftcr leadgrp leadgrpi

1 1 M M 20 20 2.0 560 38 8 1 1.24 0.16 1.89 1.22 high 3

2 2 M M 22 22 3.0 250 28 14 2 0.82 0.14 1.21 2.19 medium 2

3 3 M M 23 22 2.0 620 18 38 3 0.73 0.26 1.07 2.12 high 3

4 4 F F 28 25 3.0 350 20 14 0 0.70 0.61 1.39 1.00 medium 2

5 5 M M 28 28 2.5 340 47 8 2 0.76 0.39 1.31 0.75 medium 2

6 6 M M 28 32 3.0 160 2 3 1 0.97 0.21 1.07 1.20 low 1

The data describe 50 people in 25 matched pairs. The worker (w) was exposed to lead while involved in the production of storage batteries. The paper concerns the possibility of genotoxic effects of lead exposure. The controls (c) were not known to be exposed to lead. The pairs were matched for gender (wsex or ccex) and age (wage or cage). Workers were exposed for wyears. Blood lead levels are wlead and clead recorded in g/l. There are three measures of genetic damage based on lymphocytes in a blood sample. The micronucleus rate (wmnr and cmnr) is the number of micronuclei per 1000 binucleated lymphocytes, that is, a measure of the extent to which cell divisions went wrong, producing not two intact nuclei but rather additional micronuclei containing genetic material. The mean tail moment (wmtm and cmtm) of the comet assay is a fairly direct measure of damage to DNA with larger values signifying greater damage. Also, wmftcr and cmftcr are results of the T-cell receptor gene mutation test. The variable leadgrp cuts wlead at its thirds into low, medium, high, and leadgrpi is as.integer(leadgrp).

In answering questions, please remember: (i) the workers and controls are paired to have the same gender and similar age, (ii) if an ordered alternative is considered, it is natural to look for greater genetic damage in workers with more lead in their blood. You should plot the data in various ways. Do not submit the plots. Question 1 asks you to compare workers (wmtm) and controls (cmtm) in terms of the mean tail moment of the comet assay. Question 2 asks you to set aside the controls and to look at the mean tail moment of the comet assay for workers (wmtm) in relation to the lead groups for workers (either leadgrp or leadgrpi). Question 3 asks you to set aside the controls and to look at the micronucleus rate for workers (wmnr) in relation to the lead levels for workers (wlead). So Question 2 refers to lead groups for workers, but question 3 refers to the numeric lead levels for workers. Question 3 asks you to test the null hypothesis of zero correlation against a one-sided alternative of a correlation greater than zero. Most journals would require a two-sided test. Of course, the authors are looking for genetic damage from higher lead levels.

**Statistics 501, Midterm, Answer Page #1. This is an exam. Do not discuss it with anyone**.

|  |  |
| --- | --- |
| **Questions in part 1 refer to the mean tail moments (mtm) of the comet assay.** | FILL IN OR **CIRCLE** THE CORRECT ANSWER |
| 1.1 In a Normal-quantile plot, do the worker-minus-control pair differences in mtm look Normal? This Normal quantile plot shows a positive outlier (true or false). If the quantile plot shows any outliers, give the pair number of the one most extreme outlier. If none, write “none”. | Look Normal? YES NO  Positive outlier? TRUE FALSE  Identify one outlier: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 1.2 Test the null hypothesis that the worker-minus-control pair differences in mtm are Normal. Give the P-value.  Is the null hypothesis plausible? | P-value: \_\_\_\_\_\_\_\_\_\_\_\_  PLAUSIBLE NOT PLAUSIBLE |
| 1.3 Do an appropriate two-sided Student’s t-test to compare mean tail moments (mtm) of the comet assay for workers and matched controls. Give the two-sided P-value, the two-sided 95% confidence interval, and the associated point estimate of the typical difference. | P-value: \_\_\_\_\_\_\_\_\_\_\_\_  95% CI: [ , ]  Estimate:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 1.4 Do an appropriate two-sided Wilcoxon test to compare mean tail moments (mtm) of the comet assay for workers and matched controls. Give the two-sided P-value, the two-sided 95% confidence interval, and the associated point estimate of the typical difference. | P-value: \_\_\_\_\_\_\_\_\_\_\_\_  95% CI: [ , ]  Estimate:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 1.5 Divide the length of the 95% confidence interval from the t-test (numerator) by the length of the 95% confidence interval from the Wilcoxon test (denominator). | Ratio of lengths: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 1.6 The exact t-test confidence interval in problem 1.3 and the Wilcoxon confidence interval in 1.4 both assume the differences in problem 1.2 are symmetrically distributed about their population median, but the t-test assumes more than this. | TRUE FALSE |
| 1.7 Use the Randles-Fligner-Policello-Wolfe test to test that the null hypothesis that the worker-minus-control differences in question 1.2 are symmetrically distributed about the population median. Give the P-value and state whether symmetry is plausible. | P-value: \_\_\_\_\_\_\_\_\_\_\_\_  PLAUSIBLE NOT PLAUSIBLE |
| 1.8 If you used the t-test under Normality assumptions to test the null hypothesis H0 that the differences in problem 1.2 are symmetric about 0=0.4 then you would accept H0 at the two-sided 0.05 level, but if you tested H0 using the appropriate Wilcoxon test, then you would reject H0 at the two-sided 0.05 level. | TRUE FALSE |

Print LAST name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, First:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_

**Statistics 501, Spring 2014, Midterm, Answer Page #2.**

**This is an exam. Do not discuss it with anyone**. Read the data page. Due noon March 31, 2015.

|  |  |
| --- | --- |
| Question 2 refers to wmtm and leadgrp or leadgrpi; i.e., set aside the controls; see the data page. | FILL IN OR **CIRCLE** THE CORRECT ANSWER |
| 2.1 Test the null hypothesis that wmtm has the same distribution in the three lead groups using the Kruskal Wallis test. Give the P-value and state whether the null hypothesis is plausible. | P-value: \_\_\_\_\_\_\_\_\_\_\_\_  PLAUSIBLE NOT PLAUSIBLE |
| 2.2 Test the null hypothesis that wmtm has the same distribution in the three lead groups against the ordered alternative that the ordering that a higher lead group predicts a higher wmtm. Give the exact and asymptotic P-values. (R computes the exact P-value in this case.) | Exact P-value: \_\_\_\_\_\_\_\_\_\_\_\_  Asymptotic P-value: \_\_\_\_\_\_\_\_\_\_\_\_ |
| 2.3 Use pairwise two-sided Wilcoxon tests to compare wmtm in all three pairs of two groups defined by leadgrp. Which two groups have the smallest unadjusted P-value? What is this smallest unadjusted P-value? What is the corresponding P-value adjusted by the Bonferroni method? What is the corresponding P-value adjusted by the Holm method? What is the corresponding P-value adjusted by the Shaffer method? | Two groups: \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_  Unadjusted: \_\_\_\_\_\_\_\_\_\_ Bonferroni: \_\_\_\_\_\_\_\_\_  Holm: \_\_\_\_\_\_\_\_\_\_\_ Shaffer: \_\_\_\_\_\_\_\_\_\_\_\_ |
| 2.4 If you strongly control the familywise error rate at 0.05 in a multiple testing problem (like 2.3), then in the experiment as a whole, the probability that at least one true alternative hypothesis is not rejected is at most 5%. | TRUE FALSE |

|  |  |
| --- | --- |
| 3. Question 3 refers to **wmnr** and **wlead** for workers only; see the data page. Not wmtm! | FILL IN OR **CIRCLE** THE CORRECT ANSWER |
| 3.1 Plot y=wmnr against x=wlead. Think about it. | Free points. (Happiness from plotting and thinking) |
| 3.2 Test the null hypothesis of zero correlation against the one-sided alternative of positive correlations using Pearson’s (the usual) correlation. Give the correlation and one-sided P-value. | One-sided  Correlation: \_\_\_\_\_\_\_\_\_\_\_\_ P-value:\_\_\_\_\_\_\_\_\_\_\_ |
| 3.3 Test the null hypothesis of zero correlation against the one-sided alternative of positive correlations using Kendalls’s correlation. Give the correlation and one-sided P-value. | One-sided  Correlation: \_\_\_\_\_\_\_\_\_\_\_\_ P-value:\_\_\_\_\_\_\_\_\_\_\_ |
| 3.4 Fit the model wmnr = 0 + 1 wlead + e with symmetric independent errors using M-estimation with the default settings for rlm in the MASS. Give the estimate of the slope, 1, its estimated standard error and the t-value (aka z value) formed as the ratio of the estimate to its standard error. | Estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_  Estimated standard error: \_\_\_\_\_\_\_\_\_\_  t = z value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

**Answers: Statistics 501, Spring 2014, Midterm, Answer Page #1.**

**This is an exam. Do not discuss it with anyone**. Read the data page.

|  |  |
| --- | --- |
| **Questions in part 1 refer to the mean tail moments (mtm) of the comet assay.** | FILL IN OR **CIRCLE** THE CORRECT ANSWER |
| 1.1 In a Normal-quantile plot, do the worker-minus-control pair differences in mtm look Normal? This Normal quantile plot shows a positive outlier (true or false). If the quantile plot shows any outliers, give the pair number of the one most extreme outlier. If none, write “none”. | Look Normal? YES NO  Positive outlier? TRUE FALSE  Identify one outlier: #11 |
| 1.2 Test the null hypothesis that the worker-minus-control pair differences in mtm are Normal. Give the P-value.  Is the null hypothesis plausible? | P-value: 4.024 x 10-5  PLAUSIBLE NOT PLAUSIBLE |
| 1.3 Do an appropriate two-sided Student’s t-test to compare mean tail moments (mtm) of the comet assay for workers and matched controls. Give the two-sided P-value, the two-sided 95% confidence interval, and the associated point estimate of the typical difference. | P-value: 1.007 x 10-5  95% CI: [ 0.348, 0.758 ]  Estimate: 0.553 |
| 1.4 Do an appropriate two-sided Wilcoxon test to compare mean tail moments (mtm) of the comet assay for workers and matched controls. Give the two-sided P-value, the two-sided 95% confidence interval, and the associated point estimate of the typical difference. | P-value: 0.000227  95% CI: [ 0.465, 0.740 ]  Estimate: 0.595 |
| 1.5 Divide the length of the 95% confidence interval from the t-test (numerator) by the length of the 95% confidence interval from the Wilcoxon test (denominator). | Ratio of lengths: 1.49  t-interval is 50% longer! |
| 1.6 The exact t-test confidence interval in problem 1.3 and the Wilcoxon confidence interval in 1.4 both assume the differences in problem 1.2 are symmetrically distributed about their population median, but the t-test assumes more than this. | TRUE FALSE |
| 1.7 Use the Randles-Fligner-Policello-Wolfe test to test that the null hypothesis that the worker-minus-control differences in question 1.2 are symmetrically distributed about the population median. Give the P-value and state whether symmetry is plausible. | P-value: 0.785  PLAUSIBLE NOT PLAUSIBLE |
| 1.8 If you used the t-test under Normality assumptions to test the null hypothesis H0 that the differences in problem 1.2 are symmetric about 0=0.4 then you would accept H0 at the two-sided 0.05 level, but if you tested H0 using the appropriate Wilcoxon test, then you would reject H0 at the two-sided 0.05 level. | TRUE FALSE  You can just look at the 95% confidence intervals, or you can do the tests with 0 = .4 |

**Answers**

**Statistics 501, Spring 2014, Midterm, Answer Page #2.**

**This is an exam. Do not discuss it with anyone**. Read the data page. Due noon March 31, 2015.

|  |  |
| --- | --- |
| Question 2 refers to wmtm and leadgrp or leadgrpi; i.e., set aside the controls; see the data page. | FILL IN OR **CIRCLE** THE CORRECT ANSWER |
| 2.1 Test the null hypothesis that wmtm has the same distribution in the three lead groups using the Kruskal Wallis test. Give the P-value and state whether the null hypothesis is plausible. | P-value: 0.2527  PLAUSIBLE NOT PLAUSIBLE |
| 2.2 Test the null hypothesis that wmtm has the same distribution in the three lead groups against the ordered alternative that the ordering that a higher lead group predicts a higher wmtm. Give the exact and asymptotic P-values. (R computes the exact P-value in this case.) | Exact P-value: 0.053  Asymptotic P-value: 0.05  Remember: you cannot shop for p-values, doing all three tests in 2.1 & 2.2. Most appropriate here is the exact ordered test with p-value 0.053. |
| 2.3 Use pairwise two-sided Wilcoxon tests to compare wmtm in all three pairs of two groups defined by leadgrp. Which two groups have the smallest unadjusted P-value? What is this smallest unadjusted P-value? What is the corresponding P-value adjusted by the Bonferroni method? What is the corresponding P-value adjusted by the Holm method? What is the corresponding P-value adjusted by the Shaffer method? | Two groups: low high  The smallest adjusted p-values are all 3 times the unadjusted p-value, or approximately 0.34~3x.11  Unadjusted: 0.11 Bonferroni: 0.34  Holm: 0.34 Shaffer: 0.34  Holm and Shaffer can win over Bonferroni only if the smallest p-value meets the Bonferroni standard of 0.05/(number of tests), but that did not happen here. |
| 2.4 If you strongly control the familywise error rate at 0.05 in a multiple testing problem (like 2.3), then in the experiment as a whole, the probability that at least one true alternative hypothesis is not rejected is at most 5%. | TRUE FALSE  The promise is about falsely rejecting true null hypotheses. The promise does not refer to the alternative hypothesis. |

|  |  |
| --- | --- |
| 3. Question 3 refers to **wmnr** and **wlead** for workers only; see the data page. Not wmtm! | FILL IN OR **CIRCLE** THE CORRECT ANSWER |
| 3.1 Plot y=wmnr against x=wlead. Think about it. | Free points. (Happiness from plotting and thinking) |
| 3.2 Test the null hypothesis of zero correlation against the one-sided alternative of positive correlations using Pearson’s (the usual) correlation. Give the correlation and one-sided P-value. | One-sided  Correlation: 0.358 P-value: 0.039  But it is all one outlier. |
| 3.3 Test the null hypothesis of zero correlation against the one-sided alternative of positive correlations using Kendalls’s correlation. Give the correlation and one-sided P-value. | One-sided  Correlation: 0.041 P-value: 0.389 |
| 3.4 Fit the model wmnr = 0 + 1 wlead + e with symmetric independent errors using M-estimation with the default settings for rlm in the MASS. Give the estimate of the slope, 1, its estimated standard error and the t-value (aka z value) formed as the ratio of the estimate to its standard error. | Estimate: 0.0072  Estimated standard error: 0.0090  t = z value: 0.8033 |

**Doing the Problem Set in R**

Midterm Spring 2015 Statistics 501

Problem 1:

1.1

> dmtm<-wmtm-cmtm

> qqnorm(dmtm)

FALSE because there is a negative, not a positive, outlier.

> which.min(dmtm)

[1] 11

> storagebattery[11,]

pair wsex csex wage cage wyears wlead clead wmnr cmnr wmtm cmtm wmftcr cmftcr leadgrp leadgrpi

11 11 F F 39 35 3 260 2 4 0 1.05 2.47 1.37 0.9 medium 2

In this one pair, the control had a high mtm.

1.2

> shapiro.test(dmtm)

Shapiro-Wilk normality test. data: dmtm

W = 0.7524, p-value = 4.024e-05

1.3

> t.test(dmtm)

One Sample t-test data: dmtm

t = 5.5629, df = 24, p-value = 1.007e-05

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

0.3477058 0.7578942

sample estimates: mean of x

0.5528

1.4

> wilcox.test(dmtm,conf.int=T)

Wilcoxon signed rank test with continuity correction data: dmtm

V = 300, p-value = 0.0002273

alternative hypothesis: true location is not equal to 0

95 percent confidence interval:

0.4650308 0.7400173

sample estimates: (pseudo)median

0.5950558

1.5

> (0.7578942-0.3477058)/(0.7400173-0.4650308)

[1] 1.491667

The t-interval is 50% longer!

1.6 The t-test is an exact test if the data are Normal, hence symmetric about their population median.

1.7

> library(NSM3)

> RFPW(dmtm)

$obs.stat

[1] -0.2734514

$p.val

[1] 0.7845063

1.8 Look at the 95% confidence intervals.

Question 2.

2.1

> kruskal.test(wmtm,leadgrp)

Kruskal-Wallis rank sum test. data: wmtm and leadgrp

Kruskal-Wallis chi-squared = 2.7508, df = 2, p-value = 0.2527

2.2

> pJCK(wmtm,leadgrpi) #exact because of small samples, no ties

Group sizes: 9 8 8 Jonckheere-Terpstra J Statistic: 137

Exact upper-tail probability: 0.053

> pJCK(wmtm,leadgrpi,method="Asymptotic")

Group sizes: 9 8 8 Jonckheere-Terpstra J\* Statistic: 1.6445

Asymptotic upper-tail probability: 0.05

2.3

> pairwise.wilcox.test(wmtm,leadgrp,p.adjust.method="none")

low medium

medium 0.54 -

high 0.11 0.38

> pairwise.wilcox.test(wmtm,leadgrp,p.adjust.method="bonf").34~.11x3

> pairwise.wilcox.test(wmtm,leadgrp,p.adjust.method="holm").34~.11x3

Question 3.

3.2

> cor.test(wmnr,wlead,alternative="greater")

Pearson's product-moment correlation data: wmnr and wlead

t = 1.8399, df = 23, p-value = 0.03936

alternative hypothesis: true correlation is greater than 0

95 percent confidence interval: 0.0241278 1.0000000

sample estimates: cor 0.3581974

3.3

> cor.test(wmnr,wlead,alternative="greater",method="k")

Kendall's rank correlation tau data: wmnr and wlead

z = 0.2819, p-value = 0.389

alternative hypothesis: true tau is greater than 0

sample estimates: tau 0.04131109

3.4

> library(MASS)

> summary(rlm(wmnr~wlead))

Call: rlm(formula = wmnr ~ wlead)

Coeff: Value Std. Error t value

(Intercept) 5.8644 3.0764 1.9063

wlead 0.0072 0.0090 0.8033

**Statistics 501 Spring 2015 Final Exam: Data Page 1**

**This is an exam. Do not discuss it with anyone. Due May 7, 2015, noon.**

The data are from a paper, Benson, P. (1981) “Political alienation and public satisfaction with police services,” *Pacific Sociological Review*, 24, 45-64. The paper is available in jstor, but there is no need to look at it unless you want to – they used different methods, and I have reduced the number of variables to make the problem set simpler. The data were from a telephone survey of the St. Louis SMSA during summer 1977. The table used here is a 2x2x2x2 table with dimensions PIntegrity, Evaluation, Alienation, and Race. PIntegrity concerned the respondent’s opinion of the integrity of the police, either “not low” or “low”, and it was based on the response to two statements, “Policemen in your neighborhood are basically honest” and “The police in your neighborhood treat all citizens equally according to the law.” Evaluation concerned the respondent’s evaluation of police performance, and it was either “Positive”, or “Negative”. Alienation was an indicator of the respondent’s political alienation, either “Alienated” or “NotA” for “not alienated”, and it was based on the response to two statements, “The local government is concerned about your neighborhood” and “A person can’t get any satisfaction out of talking to the public officials in your community.” Race was either “White” or “Other”, where “Other” included black, Latino, native American, and others. Remember that I, E and A refer to opinions expressed by respondents to a survey.

The data are in an object, policeStL, in the course workspace. You will need to download the work space again. You may need to clear your web browser’s memory to download the workspace. The table appears on the second data page, consists of 16 numbers, and may be easily entered into any program by hand. The first cell of the table indicates that 1345 respondents had a positive view of police performance, thought police integrity was not low, were not politically alienated, and were white.

**Important**: In referring to models, use the short form I for PIntegrity, E for evaluation, A for alienation, and R for race, so the standard hierarchical notation for the saturated model is [IEAR]. If you mess up this **notation** you may lose many points for no good reason, so don’t mess up the notation. Make sure you know how to use the standard hierarchical notation – e.g., the model [IE] [AR] and how it differs from model [I] [EAR]. **Important**: All chi-square tests should use the **likelihood ratio chi-square, not the Pearson chi-square**. If you mess up and use the wrong chi-square, you may lose many points for no good reason, so don’t mess up. Remember that a goodness-of-fit test is one that tests a model H0 against the alternative that H0 is false, whereas some other tests have more specific alternative hypotheses.

> dimnames(policeStL)

$PIntegrity (**Short form I**)

[1] "NotLow" "Low"

$Evaluation (**Short form E**)

[1] "Positive" "Negative"

$Alienation (**Short form A**)

[1] "NotA" "Alienated"

$Race (**Short form R**)

[1] "White" "Other"

**Statistics 501 Spring 2015 Final Exam: Data Page 2**

**This is an exam. Do not discuss it with anyone. Due May 7, 2015, noon.**

> policeStL

, , Alienation = NotA, Race = White

Evaluation

PIntegrity Positive Negative

NotLow 1345 18

Low 135 23

, , Alienation = Alienated, Race = White

Evaluation

PIntegrity Positive Negative

NotLow 69 11

Low 36 22

, , Alienation = NotA, Race = Other

Evaluation

PIntegrity Positive Negative

NotLow 599 31

Low 203 64

, , Alienation = Alienated, Race = Other

Evaluation

PIntegrity Positive Negative

NotLow 64 16

Low 68 44

**Make and keep a photocopy of your answer page**. **The exam is due in my office, 473 Huntsman, on Thursday May 7, 2015, noon.** You may turn in the exam early at my mail box in the Statistics Department, 4th floor, Huntsman or by giving it to Noelle at the front desk in statistics, but if you turn in the exam early, place it in an envelope addressed to me. When all of the exams are graded, I will add an **answer key** to the on-line bulk-pack for the course. You can compare the answer key to your photocopy of your exam. Your course grade will be available from the Registrar. I no longer distribute answer keys and graded exams by US Mail. **Turn in only the answer page**. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question

**This is an exam. Do not discuss it with anyone.**

**Have a great summer!**

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2015 Final Exam: Answer Page 1 This is an exam. Do not discuss it with anyone. Due: Due May 7, 2015, noon.**

|  |  |
| --- | --- |
| **Before you do anything else, read the important notes on Data Page 1.** | **Fill in or Circle the Correct Answer** |
| 1.1 Test the one null hypothesis H0 that I, E, A and R are all independent against the alternative hypothesis that H0 is false. Give the standard notation for the log-linear model corresponding with H0. Give the value of the test statistic, the degrees of freedom, the P-value. Is H0 plausible? | Model in standard notation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_  Value: \_\_\_\_\_\_\_\_ DF: \_\_\_\_\_ P-value:\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 1.2 Test the null hypothesis H0 that the evaluation E of the police is independent of the other variables, permitting any kind of dependence among the other three variables. Test against the alternative that H0 is false. Give the standard notation for the log-linear model corresponding with H0. Give the value of the test statistic, the degrees of freedom, the P-value. Is H0 plausible? | Model in standard notation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_  Value: \_\_\_\_\_\_\_\_ DF: \_\_\_\_\_ P-value:\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 1.3 Test the null hypothesis H0 that police integrity I is conditionally independent of the evaluation E of the police given both the level of political alienation A and the race of the respondent R. Test against the alternative that H0 is false. Give the standard notation for the log-linear model corresponding with H0. Give the value of the test statistic, the degrees of freedom, the P-value. Is H0 plausible? | Model in standard notation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_  Value: \_\_\_\_\_\_\_\_ DF: \_\_\_\_\_ P-value:\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 1.4 Test the null hypothesis H0 that police integrity I and evaluation E of the police are related in the same way (i.e., same odds ratio) for all categories of Alienation A and Race R, where I, A and R can have any possible relationship and E, A and R can have any possible relationship. Test against the alternative that H0 is false. Give the notation for the log-linear model corresponding with H0. Give the value of the statistic, the degrees of freedom, the P-value. Is H0 plausible? | Model in standard notation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_  Value: \_\_\_\_\_\_\_\_ DF: \_\_\_\_\_ P-value:\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2015 Final Exam: Answer Page 1 This is an exam. Do not discuss it with anyone. Due: Due May 7, 2015, noon.**

|  |  |
| --- | --- |
| **Before you do anything else, read the important notes on Data Page 1.** | **Fill in or Circle the Correct Answer** |
| 2.1 The model [IEA] [IAR] [EAR] permits the relationship between perceieved police integrity I and evaluation E to be related in a different way depending upon political alienation A. | TRUE FALSE |
| 2.2 Test the goodness-of-fit of the model in 2.1. Give the value of the test statistic, the degrees of freedom, the P-value, and state whether the model is plausible based on the test. | Value: \_\_\_\_\_\_\_\_ DF: \_\_\_\_\_ P-value:\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 2.3 Test the null hypothesis H0 that the model in question 1.4 is the correct model against the alternative that the model in 2.1 is the correct model. Give the value of the test statistic, the degrees of freedom, the P-value, and state whether H0 is plausible. | Value: \_\_\_\_\_\_\_\_ DF: \_\_\_\_\_ P-value:\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 2.4 Fit the model [IEA] [IAR] [EAR] in question 2.1 with eps=0.0001 keeping the fitted counts. From the fitted counts, compute 4 odds ratios linking I and E, one at each level of A and R. Put I=notlow and E=positive in the **numerator** of the odds ratio. | Put odds ratios in the table.   |  |  |  | | --- | --- | --- | |  | A=notA | A=alienated | | R=white |  |  | | R=other |  |  | |
| 2.5 Fit the model [IEA] [IAR] [EAR] in question 2.1 with eps=0.0001 keeping the fitted counts. From the fitted counts, compute the 8 fitted probabilities of a **negative** evaluation of police performance. | Put estimated probabilities in the table.   |  |  |  | | --- | --- | --- | |  | I=not low | I=low | | R=white  A=notA |  |  | | R=white  A=alienated |  |  | | R=other  A=notA |  |  | | R=other  A=alienated |  |  | |
| 2.6 The model [IEA] [IAR] [EAR] preserves the [IER] marginal table and is the smoothest (i.e., maximum entropy) table that dose this. | TRUE FALSE |

**Answers: Stat 501 S-2015 Final Exam: Answer Page 1 This is an exam. Do not discuss it with anyone.**

|  |  |
| --- | --- |
| **Before you do anything else, read the important notes on Data Page 1.** | **Fill in or Circle the Correct Answer** |
| 1.1 Test the one null hypothesis H0 that I, E, A and R are all independent against the alternative hypothesis that H0 is false. Give the standard notation for the log-linear model corresponding with H0. Give the value of the test statistic, the degrees of freedom, the P-value. Is H0 plausible? | Model in standard notation: [I] [E] [A] [R]  Value: 709.98 DF: 11 P-value: <0.0001  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 1.2 Test the null hypothesis H0 that the evaluation E of the police is independent of the other variables, permitting any kind of dependence among the other three variables. Test against the alternative that H0 is false. Give the standard notation for the log-linear model corresponding with H0. Give the value of the test statistic, the degrees of freedom, the P-value. Is H0 plausible? | Model in standard notation: [E] [IAR]  Value: 341.32 DF: 7 P-value: <0.0001  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 1.3 Test the null hypothesis H0 that police integrity I is conditionally independent of the evaluation E of the police given both the level of political alienation A and the race of the respondent R. Test against the alternative that H0 is false. Give the standard notation for the log-linear model corresponding with H0. Give the value of the test statistic, the degrees of freedom, the P-value. Is H0 plausible? | Model in standard notation: [IAR] [EAR]  Value: 138.53 DF: 4 P-value: <0.0001  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 1.4 Test the null hypothesis H0 that police integrity I and evaluation E of the police are related in the same way (i.e., same odds ratio) for all categories of Alienation A and Race R, where I, A and R can have any possible relationship and E, A and R can have any possible relationship. Test against the alternative that H0 is false. Give the notation for the log-linear model corresponding with H0. Give the value of the statistic, the degrees of freedom, the P-value. Is H0 plausible? | Model in standard notation: [IE] [IAR] [EAR]  Value: 12.00 DF: 3 P-value: 0.00738  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |

**Answers: Stat 501 S-2015 Final Exam: Answer Page 1 This is an exam. Do not discuss it with anyone. Due:**

|  |  |
| --- | --- |
| **Before you do anything else, read the important notes on Data Page 1.** | **Fill in or Circle the Correct Answer** |
| 2.1 The model [IEA] [IAR] [EAR] permits the relationship between perceived police integrity I and evaluation E to be related in a different way depending upon political alienation A. | TRUE FALSE |
| 2.2 Test the goodness-of-fit of the model in 2.1. Give the value of the test statistic, the degrees of freedom, the P-value, and state whether the model is plausible based on the test. | Value: 3.88 DF: 2 P-value: 0.144  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 2.3 Test the null hypothesis H0 that the model in question 1.4 is the correct model against the alternative that the model in 2.1 is the correct model. Give the value of the test statistic, the degrees of freedom, the P-value, and state whether H0 is plausible. | Value: 12.00068-3.880125=8.120555 DF: 3-2=1 P-value: 0.0044  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 2.4 Fit the model [IEA] [IAR] [EAR] in question 2.1 with eps=0.0001 keeping the fitted counts. From the fitted counts, compute 4 odds ratios linking I and E, one at each level of A and R. Put I=notlow and E=positive in the **numerator** of the odds ratio. | Put odds ratios in the table.   |  |  |  | | --- | --- | --- | |  | A=notA | A=alienated | | R=white | 7.87 | 3.03 | | R=other | 7.87 | 3.03 | |
| 2.5 Fit the model [IEA] [IAR] [EAR] in question 2.1 with eps=0.0001 keeping the fitted counts. From the fitted counts, compute the 8 fitted probabilities of a **negative** evaluation of police performance. **Round** to 2 digits after the decimal, as in 0.99. | Put estimated probabilities in the table.   |  |  |  | | --- | --- | --- | |  | I=not low | I=low | | R=white  A=notA | 0.02 | 0.12 | | R=white  A=alienated | 0.15 | 0.36 | | R=other  A=notA | 0.04 | 0.26 | | R=other  A=alienated | 0.18 | 0.40 | |
| 2.6 The model [IEA] [IAR] [EAR] preserves the [IER] marginal table and is the smoothest (i.e., maximum entropy) table that dose this. | TRUE FALSE |

Statistics 510 Final 2015 Answer Page

1.1

> loglin(policeStL,list(1,2,3,4))

2 iterations: deviation 4.547474e-13

$lrt

[1] 709.9806

$df

[1] 11

> 1-pchisq(709.9806,11)

[1] 0

1.2

> loglin(policeStL,list(2,c(1,3,4)))

2 iterations: deviation 7.105427e-15

$lrt

[1] 341.3207

$df

[1] 7

> 1-pchisq(341.3207,7)

[1] 0

1.3

> loglin(policeStL,list(c(1,3,4),c(2,3,4)))

2 iterations: deviation 0

$lrt

[1] 138.5315

$df

[1] 4

> 1-pchisq(138.5315,4)

[1] 0

1.4

> loglin(policeStL,list(c(1,2),c(1,3,4),c(2,3,4)))

5 iterations: deviation 0.09289019

$lrt

[1] 12.00068

$df

[1] 3

> 1-pchisq(12.00068,3)

[1] 0.007380831

2.2

> loglin(policeStL,list(c(1,2,3),c(1,3,4),c(2,3,4)))

5 iterations: deviation 0.01297741

$lrt

[1] 3.880125

$df

[1] 2

> 1-pchisq(3.880125,2)

[1] 0.143695

2.3 Compare the full and reduced models in terms of the likelihood ratio chi-square.

> 12.00068-3.880125

[1] 8.120555

> 3-2

[1] 1

> 1-pchisq(8.120555,1)

[1] 0.004376616

2.4

> ft<-loglin(policeStL,list(c(1,2,3),c(1,3,4),c(2,3,4)),

eps=0.0001,fit=T)$fit

7 iterations: deviation 6.614585e-05

> ft

, , Alienation = NotA, Race = White

Evaluation

PIntegrity Positive Negative

NotLow 1340.4677 22.53229

Low 139.5323 18.46771

, , Alienation = Alienated, Race = White

Evaluation

PIntegrity Positive Negative

NotLow 67.65754 12.34246

Low 37.34246 20.65754

, , Alienation = NotA, Race = Other

Evaluation

PIntegrity Positive Negative

NotLow 603.5323 26.46771

Low 198.4677 68.53229

, , Alienation = Alienated, Race = Other

Evaluation

PIntegrity Positive Negative

NotLow 65.34246 14.65754

Low 66.65754 45.34246

> 1340.4677\*18.46771/(139.5323\*22.53229)

[1] 7.873889

> 67.65754\*20.65754/(37.34246\*12.34246)

[1] 3.032426

> 603.5323\*68.53229/(198.4677\*26.46771)

[1] 7.873894

> 65.34246\*45.34246/(66.65754\*14.65754)

[1] 3.032426

2.5

> prop.table(ft,c(1,3,4))

, , Alienation = NotA, Race = White

Evaluation

PIntegrity Positive Negative

NotLow 0.9834686 0.01653139

Low 0.8831157 0.11688425

, , Alienation = Alienated, Race = White

Evaluation

PIntegrity Positive Negative

NotLow 0.8457193 0.1542807

Low 0.6438355 0.3561645

, , Alienation = NotA, Race = Other

Evaluation

PIntegrity Positive Negative

NotLow 0.9579878 0.04201224

Low 0.7433248 0.25667524

, , Alienation = Alienated, Race = Other

Evaluation

PIntegrity Positive Negative

NotLow 0.8167807 0.1832193

Low 0.5951566 0.4048434

**Statistics 501, Spring 2014, Midterm: Data Page #1**

Due in class, 1 April 2014 at noon.

**This is an exam. Do not discuss it with anyone**. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Cheating on an exam is the single dumbest thing a PhD student at Penn can do.

Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question. **Due in class Tuesday 1 April, 2014.**

The data for this problem are at in the latest Rst501.RData for R users as the object nhanesInsurance and in the nhanesInsurance.csv file at <http://stat.wharton.upenn.edu/statweb/course/Spring-2008/stat501> The list is case sensitive, so nhanesInsurance.csv is with lower case items.

The data are from NHANES 2011-2012. They described 801 people who report they have no health insurance and 801 paired people who report they do have health insurance. Everyone is at least 25 years old and under 65. The pairing if for age and gender.

SEQN = NHANES identification number

age = age in years (RIDAGEYR)

female = 1 for female, 0 for male (RIAGENDR)

educ = education categories (DMDEDUC2)

povertyratio = ratio of income to the poverty level, capped at 5 times (INDFMPIR)

bmi = body mass index (BMXBMI)

sad1, sad2 = Sagittal abdominal diameter measured twice in cm (BMXSAD1, BMXSAD2)

noplace = 1 if respondent said he/she had no routine place to go for healthcare (HUQ030==2)

nocare = 1 if respondent said he/she received no healthcare in the last 12 month (HUQ050 == 0)

noinsurance = 1 if respondent said he/she had no health insurance (including no Medicaid) (HIQ011 == 2) systolic blood pressure measured 3 times, systolic1, systolic2, systolic3, (BPXSY1, BPXSY2, BPXSY3)

diastolic blood pressure measured 3 times, diastolic1, diastolic2, diastolic3, (BPXDI1, BPXDI2, BPXDI3)

bpmedication = respondent says he/she is on blood pressure medication (BPQ050A)

pair = 1, 2, … , 801, for 801 pairs of two people, one with health insurance, the other without.

The Sagittal abdominal diameter (cm) attempts to measure waist size as it related to abdominal fat. <http://en.wikipedia.org/wiki/Sagittal_Abdominal_Diameter> BMI is an index of obestity <https://www.nhlbi.nih.gov/guidelines/obesity/BMI/bmicalc.htm> For information about blood pressure, see

<http://www.heart.org/HEARTORG/Conditions/HighBloodPressure/AboutHighBloodPressure/Understanding-Blood-Pressure-Readings_UCM_301764_Article.jsp>

The NHLBI writes “Metabolic syndrome is the name for a group of risk factors that raises your risk for heart disease and other health problems, such as diabetes and stroke.” Among the risk factors are a large waistline, high blood pressure (or being treated for high blood pressure). The write “Successfully controlling metabolic syndrome requires long-term effort and teamwork with your health care providers.” <http://www.nhlbi.nih.gov/health/health-topics/topics/ms/>

In that spirit, you will be looking at whether having health insurance is related to better markers for metabolic syndrome, controlling for age and sex. There are many possibilities here because having health insurance is related to SES, education, income and many other things.

The data you have consists of 801 matched pairs, indicated by the pair variable, so the first two people are paired. The first person in the pair has no health insurance, while the second person does have health insurance. The pairs were matched for age and gender. It is important that you take appropriate account of the pairing in selecting and using statistical methods.

**Statistics 501, Spring 2014, Midterm: Data Page #2**

You should look at the data to understand its structure before starting the exam. For example:

> nhanesInsurance[1:4,c(1,2,3,6,19)]

SEQN age female bmi pairID

1 62208 38 0 22.2 1

2560 70910 38 0 32.1 1

107 63401 49 1 21.0 2

2735 71657 49 1 29.4 2

So pair 1 consists of two men aged 38, while pair 2 consists of two women aged 49.

> attach(nhanesInsurance)

Consecutive people are paired:

> table(pairID[noinsurance==1]==pairID[noinsurance==0])

TRUE

801

Age is balanced by pairing:

> boxplot(age[noinsurance==1],age[noinsurance==0])

Age differs within a pair by at most one year:

> boxplot(age[noinsurance==1]-age[noinsurance==0])

> table(female[noinsurance==1],female[noinsurance==0])

0 1

0 445 0

1. 0 356

> plot(density(bmi))

> abline(v=25)

> abline(v=30)

> abline(v=35)

Pair difference in bmi, noinsurance-minus-insurance:

> bmidif<-bmi[noinsurance==1]-bmi[noinsurance==0]

If you are not sure what this does, look at bmi, at noinsurance==1, at bmi[noinsurance==1], etc so that you do kwno what this does.

> summary(bmidif)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-43.9000 -5.2000 0.2000 0.3839 6.0000 36.1000

Models and terminology

**Model 1**: Zi are independent, identically distributed (iid), with a continuous distribution symmetric about , or Zi =  + ei where ei are iid with continuous distribution symmetric about 0.

**Model 2**: Zi are independent distributed, with continuous distributions with median , or Zi =  + ei where ei are independent with continuous distributions with median 0. The observations need not have identical distributions.

**Model 3**: Xi are independent, identically distributed with a continuous distribution. Yi are independent, identically distributed with a continuous distribution, possibly different from the distribution of the X’s. The X’s and Y’s are independent.

**Model 4**: Xi = ei i=1,…,m, and Yj =  + em+j j=1,…,n, where the ek are m+n independent, identically distributed observations with a continuous distribution.

**Model 5**: (Xi, Yi) are independent bivariate observations.

**Pair differences** always mean noinsurance-minus-insurance and there are 801 pair differences.

Print **LAST name**: \_\_\_\_\_\_\_\_\_\_\_, first:\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2014, Midterm, Answer Page #1 Due in class, 1 April2014, noon.

**This is an exam. Do not discuss it with anyone**. Read the data page.

|  |  |
| --- | --- |
| Models are defined on the data page. Also, see the definition of a pair difference on the data page. | Fill in or CIRCLE the correct answer |
| 1.1 Under model 1, test the null hypothesis H0 that the pair differences in (first) systolic blood pressure (**systolic1**) are symmetric about zero using the appropriate Wilcoxon test. Give the value of name of the test, the value the test statistic (as reported by R) and the two-sided P-value. Is H0 plausible? | Name of test:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Value: \_\_\_\_\_\_\_\_\_\_\_ P-value:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 1.2 Under model 1, give the 95% confidence interval and the Hodges-Lehmann point estimate for the typical pair difference in (first) systolic blood pressure (systolic1), that is, the center of symmetry. | 95% interval: [ , ]  Point estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 1.3 Repeat questions 1.1 and 1.2 using (first) Sagittal abdominal diameter (**sad1**) in place of blood pressure. | Value: \_\_\_\_\_\_\_\_\_\_\_ P-value:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE  95% interval: [ , ]  Point estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 1.4 Under model 2, test the null hypothesis that the population median pair difference in **bmi** is zero. Give the name of the test, and the two-sided P-value, correcting appropriately for ties. How many of the 801 pairs contribute to the test? | Name of test:\_\_\_\_\_\_\_\_\_ P-value:\_\_\_\_\_\_\_\_\_\_  How many pairs contribute? \_\_\_\_\_\_\_\_\_\_ |
| 1.5 Under model 2, give the 95% confidence interval for the median pair difference in **bmi** and the associated point estimate. (These are not adjusted for ties.) | 95% interval: [ , ]  Point estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 1.6 You cannot appropriately use the method in question 1.1 with the 801 pair differences in bmi, because it is clear that the 1602 bmi values are skewed right. | CIRCLE ONE  TRUE FALSE |

|  |  |
| --- | --- |
| Question 2 asks about the relationship between the 801 Sagittal abdominal diameter (**sad1**) and the 801 (first) systolic blood pressures (**systolic1**) for the 801 people without health insurance. **Do not use** their paired controls with health insurance. | Fill in or CIRCLE the correct answer |
| 2.1 Use Kendall’s rank correlation to test the null hypothesis that sad1 and systolic1 are independent for the 801 people without health insurance. Give the value of the correlation coeffient and the two-sided P-value. Is the null hypothesis plausible? | Value: \_\_\_\_\_\_\_\_\_\_\_ P-value:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 2.2 Give the 95% (asymptotic) confidence interval for Kendall’s rank correlation for the data in 2.1. | 95% interval: [ , ] |
| Use the results in 2.1 and 2.2 to give the estimate of the probability of concordance and the 95% confidence interval for that probability. | Estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  95% interval: [ , ] |

Print **LAST name**, then first: \_\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2014, Midterm, Answer Page #2 Due in class, 1 April 2014, noon.

**This is an exam. Do not discuss it with anyone**. Read the data page

|  |  |
| --- | --- |
| For this question, you should boxplot the pair difference in sad1 for pairs containing men versus pairs containing women. Do not turn in the plot. | Fill in or CIRCLE the correct answer |
| 3.1 Of the 801 pairs, how many contain two women? How many contain two men? | 2 women:\_\_\_\_\_\_\_\_\_\_\_ + 2 men \_\_\_\_\_\_\_\_ = 801 |
| 3.2 Test the null hypothesis that the pair difference in **sad1** have the same distribution for female pairs and for male pairs using the appropriate Wilcoxon test. Give the name of the test and the two-sided P-value. Is the null hypothesis plausible? | Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 3.3 Assuming model 4, build a 95% confidence interval and point estimate for the shift, , female-minus-male difference in noinsurance-minus-insurance pair difference in **sad1**. What are the units of measurement (miles, weeks, dollars, etc.)? | Estimate: \_\_\_\_\_\_\_\_\_\_  95% interval: [ , ]  Units of measurement: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 3.4 Estimate the probability that the pair difference in sad1, noinsurance-minus-insurance, will be larger a female pair than for a male pair. | Estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 3.5 The test in 3.2 and the estimate in 3.4 are not valid if model 4, the shift model, is false. | CIRCLE ONE  TRUE FALSE |
| 3.6 Test the same hypothesis as in 3.2 but use LePage’s test. Give the two-sided asymptotic P-value. | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 3.7 Use the Kolomogorov-Smirnov test to test the same hypothesis as in 3.2. What is the two-sided P-value? | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

Question 4 looks just at the bmi for the 801 people without insurance by whether they no place to go for health care (noplace=1) and whether they received no health care in the previous year (nocare=1), making 4 groups.

> bmiNoI<-bmi[noinsurance==1]

> grp<-(factor(noplace):factor(nocare))[noinsurance==1]

> table(grp)

|  |  |
| --- | --- |
| Question 4 refers only to the 801 people with no health insurance. See above. | Fill in or CIRCLE the correct answer |
| Use the Kruskal Wallis test the null hypothesis that the distribution of bmi is the same over the four groups defined by grp. Give the P-value. Is the null hypothesis plausible? | P-value: \_\_\_\_\_\_\_\_\_  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| Compare all four groups using Wilcoxon two sample tests adjusting using Holm’s procedure. List all **pairs** of groups that differ significantly at the 0.05 level after adjustment. If none, write “none”. Example: listing (0,1):(0,0) means (noplace,nocare) = (0,1) differs from (noplace,nocare) = (0,0). |  |
| The chance that Holm’s procedure will reject at least one false null hypothesis at the 5% level is at most 5%. | CIRCLE ONE  TRUE FALSE |

**Statistics 501 Spring 2104 Midterm Answer page**

|  |  |
| --- | --- |
| Models are defined on the data page. Also, see the definition of a pair difference on the data page. | Fill in or CIRCLE the correct answer |
| 1.1 Under model 1, test the null hypothesis H0 that the pair differences in (first) systolic blood pressure (**systolic1**) are symmetric about zero using the appropriate Wilcoxon test. Give the value of name of the test, the value the test statistic (as reported by R) and the two-sided P-value. Is H0 plausible? | Name of test Wilcoxon signed rank test  Value:152295.5 P-value: 0.31  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 1.2 Under model 1, give the 95% confidence interval and the Hodges-Lehmann point estimate for the typical pair difference in (first) systolic blood pressure (systolic1), that is, the center of symmetry. | 95% interval: [ -1.00, 2.00]  Point estimate: 1.00 |
| 1.3 Repeat questions 1.1 and 1.2 using (first) Sagittal abdominal diameter (**sad1**) in place of blood pressure. | Value162915.2 P-value: 0.36  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE  95% interval: [ -0.25, 0.60 ]  Point estimate: 0.20 |
| 1.4 Under model 2, test the null hypothesis that the population median pair difference in **bmi** is zero. Give the name of the test, and the two-sided P-value, correcting appropriately for ties. How many of the 801 pairs contribute to the test? | Name of test: sign test P-value: 0.697  How many pairs contribute? 392+404 = 796 |
| 1.5 Under model 2, give the 95% confidence interval for the median pair difference in **bmi** and the associated point estimate. (These are not adjusted for ties.) | 95% interval: [ -0.5, 1.1 ]  Point estimate: 0.20 |
| 1.6 You cannot appropriately use the method in question 1.1 with the 801 pair differences in bmi, because it is clear that the 1602 bmi values are skewed right. | CIRCLE ONE  TRUE FALSE |

|  |  |
| --- | --- |
| Question 2 asks about the relationship between the 801 Sagittal abdominal diameter (**sad1**) and the 801 (first) systolic blood pressures (**systolic1**) for the 801 people without health insurance. Do not use their paired controls with health insurance. | Fill in or CIRCLE the correct answer |
| 2.1 Use Kendall’s rank correlation to test the null hypothesis that sad1 and systolic1 are independent for the 801 people without health insurance. Give the value of the correlation coeffient and the two-sided P-value. Is the null hypothesis plausible? | Value: 0.206 P-value: 2.2 x 10-16  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 2.2 Give the 95% (asymptotic) confidence interval for Kendall’s rank correlation for the data in 2.1. | 95% interval: [ 0.159, 0.253 ] |
| Use the results in 2.1 and 2.2 to give the estimate of the probability of concordance and the 95% confidence interval for that probability. | Estimate: 0.603  95% interval: [ 0.5795, 0.6265 ] |

**Statistics 501 Spring 2104 Midterm Answer page**

|  |  |
| --- | --- |
| For this question, you should boxplot the pair difference in sad1 for pairs containing men versus pairs containing women. Do not turn in the plot. | Fill in or CIRCLE the correct answer |
| 3.1 Of the 801 pairs, how many contain two women? How many contain two men? | 2 women: 356 + 445 = 801 |
| 3.2 Test the null hypothesis that the pair difference in **sad1** have the same distribution for female pairs and for male pairs using the appropriate Wilcoxon test. Give the name of the test and the two-sided P-value. Is the null hypothesis plausible? | Name: Wilcoxn’s rank sum P-value: 0.009372  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| 3.3 Assuming model 4, build a 95% confidence interval and point estimate for the shift, , female-minus-male difference in noinsurance-minus-insurance pair difference in **sad1**. What are the units of measurement (miles, weeks, dollars, etc.)? | Estimate: 1.10  95% interval:[ 0.30, 1.90]  Units of measurement: cm |
| 3.4 Estimate the probability that the pair difference in sad1, noinsurance-minus-insurance, will be larger a female pair than for a male pair. | Estimate: 0.553 |
| 3.5 The test in 3.2 and the estimate in 3.4 are not valid if model 4, the shift model, is false. | CIRCLE ONE  TRUE FALSE |
| 3.6 Test the same hypothesis as in 3.2 but use LePage’s test. Give the two-sided asymptotic P-value. | P-value: 0.0226 |
| 3.7 Use the Kolomogorov-Smirnov test to test the same hypothesis as in 3.2. What is the two-sided P-value? | P-value: 0.006937 |

Question 4 looks just at the bmi for the 801 people without insurance by whether they no place to go for health care (noplace=1) and whether they received no health care in the previous year (nocare=1), making 4 groups.

> bmiNoI<-bmi[noinsurance==1]

> grp<-(factor(noplace):factor(nocare))[noinsurance==1]

> table(grp)

|  |  |
| --- | --- |
| Question 4 refers only to the 801 people with no health insurance. See above. | Fill in or CIRCLE the correct answer |
| Use the Kruskal Wallis test the null hypothesis that the distribution of bmi is the same over the four groups defined by grp. Give the P-value. Is the null hypothesis plausible? | P-value: 0.01107  CIRCLE ONE  PLAUSIBLE NOT PLAUSIBLE |
| Compare all four groups using Wilcoxon two sample tests adjusting using Holm’s procedure. List all **pairs** of groups that differ significantly at the 0.05 level after adjustment. If none, write “none”. Example: listing (0,1):(0,0) means (noplace,nocare) = (0,1) differs from (noplace,nocare) = (0,0). | (0,0):(1,1)  or equivalently  (someplace,somecare) differs from (noplace,nocare) |
| The chance that Holm’s procedure will reject at least one false null hypothesis at the 5% level is at most 5%. | CIRCLE ONE  TRUE FALSE |

**Stat 501 Spring 2104 Midterm: Doing the problem set in R**

1.1-1.2

> difsystolic1<-systolic1[noinsurance==1]-

systolic1[noinsurance==0]

> boxplot(difsystolic1)

> wilcox.test(difsystolic1,conf.int=T)

Wilcoxon signed rank test with continuity correction

data: difsystolic1

V = 152295.5, p-value = 0.3109

alternative hypothesis: true location is not equal to 0

95 percent confidence interval:

-0.9999988 2.0000214

sample estimates:

(pseudo)median

0.9999535

1.3

> difsad<-sad1[noinsurance==1]-sad1[noinsurance==0]

> wilcox.test(difsad,conf.int=T)

Wilcoxon signed rank test with continuity correction

data: difsad

V = 162912.5, p-value = 0.3598

alternative hypothesis: true location is not equal to 0

95 percent confidence interval:

-0.2499814 0.6000532

sample estimates:

(pseudo)median

0.1999492

1.4

> difbmi<-bmi[noinsurance==1]-bmi[noinsurance==0]

> table(sign(difbmi))

-1 0 1

392 5 404 Five 0’s (ties) are removed.

> prop.test(392,392+404,p=1/2)

1-sample proportions test with continuity correction

data: 392 out of 392 + 404, null probability 1/2

X-squared = 0.152, df = 1, p-value = 0.6966

alternative hypothesis: true p is not equal to 0.5

95 percent confidence interval:

0.4572270 0.5277714

sample estimates:

p 0.4924623

1.5

> signCI(difbmi)

$CI

[1] -0.5 1.1

$coverage

[1] 0.9522137

> median(difbmi)

[1] 0.2

2.1

> cor.test(systolic1[noinsurance==1],sad1[noinsurance==1],method="k") Kendall's rank correlation tau

data: systolic1[noinsurance == 1] and sad1[noinsurance == 1]

z = 8.5379, p-value < 2.2e-16

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau 0.2060076

2.2

> library(NSM3)

> kendall.ci(systolic1[noinsurance==1],sad1[noinsurance==1])

1 - alpha = 0.95 two-sided CI for tau:

0.159, 0.253

2.3 Just transform the correlation.

> (c(0.206,0.159, 0.253)+1)/2

[1] 0.6030 0.5795 0.6265

3.1 – 3.4

> fem<-1==female[noinsurance==1]

> table(fem)

fem

0 1

445 356

> wilcox.test(difsad[fem==1],difsad[fem==0],conf.int=T)

Wilcoxon rank sum test with continuity correction

data: difsad[fem == 1] and difsad[fem == 0]

W = 87664.5, p-value = 0.009372

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval: 0.2999803 1.9000161

sample estimates difference in location 1.10002

> 87664.5/(445\*356)

[1] 0.5533676

3.6

> pLepage(difsad[fem==1],difsad[fem==0],method="Asymptotic")

Ties are present, so p-values are based on conditional null distribution.

Number of X values: 356 Number of Y values: 445

Lepage D Statistic: 7.5763

Asymptotic upper-tail probability: 0.0226

3.7

> ks.test(difsad[fem==1],difsad[fem==0])

Two-sample Kolmogorov-Smirnov test

data: difsad[fem == 1] and difsad[fem == 0]

D = 0.1197, p-value = 0.006937

alternative hypothesis: two-sided

4

> bmiNoI<-bmi[noinsurance==1]

> grp<-(factor(noplace):factor(nocare))[noinsurance==1]

> kruskal.test(bmiNoI~grp)

Kruskal-Wallis rank sum test

data: bmiNoI by grp

Kruskal-Wallis chi-squared = 11.125, df = 3, p-value = 0.01107

> pairwise.wilcox.test(bmiNoI,grp)

Pairwise comparisons using Wilcoxon rank sum test

data: bmiNoI and grp

0:0 0:1 1:0

0:1 0.363 - -

1:0 0.114 1.000 -

1:1 0.019 1.000 1.000

P value adjustment method: holm

**Statistics 501 Spring 2014 Final Exam: Data Page 1**

**This is an exam. Do not discuss it with anyone.**

**Due: Monday, May 12, at 12:00am**

The data are from NHANES 2011-2012. It is a 24 table. It was built from four questions, WHQ030M, WHQ500, RIAGENDR, BMXBMI, with some categories changed to make them binary. The variable “thinkfat” asks about a person’s weight, and people either think they are “fat or overweight” or “about the right weight”; there were other categories, but they are not represented here. The variable “trying to” asks whether a person is trying to lose weight or is not trying to make a change in weight; again there were other categories. The variable “bmi25” is based on measurements of body mass index, and is <25 or >=25, where >=25 is the conventional cut for overweight. The last variables is gender, male or female. The table is in nhanesWeight in the R workspace for the course. You will need to download the workspace again, and may need to clear your web browser’s memory. If you are using some other software, you can enter the following 16 numbers by hand.

> nhanesWeight

, , bmi25 = <25, gender = male

tryingto

thinkfat lose.weight no.change

fat/overweight 38 4

about.right 85 295

, , bmi25 = >=25, gender = male

tryingto

thinkfat lose.weight no.change

fat/overweight 65 6

about.right 43 16

, , bmi25 = <25, gender = female

tryingto

thinkfat lose.weight no.change

fat/overweight 32 3

about.right 114 318

, , bmi25 = >=25, gender = female

tryingto

thinkfat lose.weight no.change

fat/overweight 82 5

about.right 41 11

**Very important**: There are 4 variables in this table. Variable 1 is thinkfat and is abbreviated in this exam as variable F. Variable 2 is tryingto and is abbreviated as T. Variable 3 is bmi25 and is abbreviated as B. Variable 4 is gender ad is abbreviated as G. If you mess up and use T to refer to thinkfat (wrong) instead of tryingto (right), then you will get many questions wrong for a tiny error. **Don’t mess up** that way: check that you are using the correct abbreviations.

**Very important**: As always, model [FT][BG] is the abbreviation for the log linear model

log(mijkn) = u + uF(i) + uT(j) + uB(k) + uG(n) + uFT(ij) + uBG(kn)

Make sure you understand this notation in terms of variables before attempting the exam. The notation is used in Fienberg’s book (and everywhere else). When a question asks for a model, give the model in this abbreviated notation; eg [FT][BG]. If a question speaks of a model it will use this notation. Keep in mind that [F][T][B][G] and [FTBG] are very different models, and FTBG does not denote a model: brackets matter. Don’t invent a new notation.

**Very important**: Always use the likelihood ratio chi-square, not Pearson’s chi-square.

> dimnames(nhanesWeight)

$thinkfat (F)

[1] "fat/overweight" "about.right"

$tryingto (T)

[1] "lose.weight" "no.change"

$bmi25 (B)

[1] "<25" ">=25"

$gender (G)

[1] "male" "female"

**Claim 3.2 is related to question 3.2.**

Claim 3.2: Thinking you are fat/overweight is positively related to trying to lose weight, but the relationship is much stronger for people with BMI<25 than for people with BMI>=25, and this is true for both men and women.

**Make and keep a photocopy of your answer page**. **The exam is due in my office, 473 Huntsman, on Monday 12 May 2014 at noon.** You may turn in the exam early at my mail box in the Statistics Department, 4th floor, Huntsman or by giving it to Noelle at the front desk in statistics, but if you turn in the exam early, place it in an envelope addressed to me. When all of the exams are graded, I will add an **answer key** to the on-line bulk-pack for the course. You can compare the answer key to your photocopy of your exam. Your course grade will be available from the Registrar. I no longer distribute answer keys and graded exams by US Mail. **Turn in only the answer page**. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question

**This is an exam. Do not discuss it with anyone.**

**Have a great summer!**

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2014 Final Exam: Answer Page 1 This is an exam. Do not discuss it with anyone. Due: Monday, May 12, at 12:00am**

|  |  |
| --- | --- |
| Use the abbreviations listed as very important on the data page, eg, F, T and [FT] | Fill in or circle the correct answer. |
| 1.1 Use the [] notation to give the log-linear model that corresponds with the four variables, F, T, B and G being independent. | Model: |
| 1.2 Using the likelihood ratio test of goodness of fit, test the null hypothesis that the model in 1.1 is correct. Give the value of the statistic, the DF and P-value. Is the model plausible? | Value (one number):  Degrees of freedom:  P-value:  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 1.3 Under model [FB][TG], gender (G) is independent of body mass index <25 (B). | Circle one:  TRUE FALSE |
| 1.4 Under model [FB][TG][TB], the odds ratio linking F and B is the same for males and females. | Circle one:  TRUE FALSE |
| 1.5 In model [FTB][TBG][FG], feeling fat/overweight (F) is conditionally independent of gender (G) given T and B. | Circle one:  TRUE FALSE |
| 1.6 Model [FTB][TBG][FG] can be collapsed over T without changing the odds ratio linking F and G. | Circle one:  TRUE FALSE |
| 1.7 In model [FTB][TBG],  feeling fat/overweight (F) is independent of gender (G). | Circle one:  TRUE FALSE |
| 1.8 In model [FTB][TBG], at each of the 4 levels of  T-and-B, the odds ratio linking feeling gender (G) and feeling fat/overweight (F) equals 1. | Circle one:  TRUE FALSE |
| 1.9 In model [FTB][G], G and B are **not** independent but **are** conditionally independent given F-and-T. | Circle one:  TRUE FALSE |

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2014 Final Exam: Answer Page 2 This is an exam. Do not discuss it with anyone. Due: Monday, May 12, at 12:00am**

|  |  |
| --- | --- |
|  | Fill in/circle the answer. |
| 2.1 Test the goodness of fit of model [FTB][G]. Give the value of the likelihood ratio chi-square, the degrees of freedom, the P-value. Based on this test alone, is the null hypothesis that this model is correct plausible? | Value (one number):  Degrees of freedom:  P-value:  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 2.2 Test the null hypothesis that [FTB][G] is the correct model against the alternative hypothesis that [FTB][TG] is the correct model. Give the value of the likelihood ratio chi-square test statistic, its degrees of freedom, and P-value and state whether the null hypothesis is plausible. | Value (one number):  Degrees of freedom:  P-value:  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 2.3 Test the null hypothesis that [FT][FB][TB][G] is the correct model against the alternative hypothesis that [FTB][G] is the correct model. Answer the same questions for this comparison as in 2.2. | Value (one number):  Degrees of freedom:  P-value:  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 2.4 Model [FTB][G] can be collapsed over gender G without changing the odds ratios linking the other three variables, F, T and B. | Circle one:  TRUE FALSE |

|  |  |
| --- | --- |
| Use fitted counts from [FTB][G] for question 3. | Fill in or circle the correct answer. |
| 3.1 Give the four estimated odds ratios linking F and T at each level of B-and-G. You are to enter 4 numbers, and all 4 are >=1. | |  |  |  |  | | --- | --- | --- | --- | |  | G=Male | G=Female | | | B is <25 |  |  | | B is >=25 |  |  | |
| 3.2 Claim 3.2 on the data page is a reasonable summary of the odds ratios in 3.1. | Circle one:  TRUE FALSE |

**Statistics 501 Spring 2014 Final Exam: Answers**

|  |  |
| --- | --- |
| Use the abbreviations listed as very important on the data page, eg, F, T and [FT] | Fill in or circle the correct answer. |
| 1.1 Use the [] notation to give the log-linear model that corresponds with the four variables, F, T, B and G being independent. | Model: [F][T][B][G] |
| 1.2 Using the likelihood ratio test of goodness of fit, test the null hypothesis that the model in 1.1 is correct. Give the value of the statistic, the DF and P-value. Is the model plausible? | Value (one number): 715.5127  Degrees of freedom: 11  P-value: <0.0001  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 1.3 Under model [FB][TG], gender (G) is independent of body mass index <25 (B). | Circle one:  TRUE FALSE |
| 1.4 Under model [FB][TG][TB], the odds ratio linking F and B is the same for males and females. | Circle one:  TRUE FALSE |
| 1.5 In model [FTB][TBG][FG], feeling fat/overweight (F) is conditionally independent of gender (G) given T and B. | Circle one:  TRUE FALSE |
| 1.6 Model [FTB][TBG][FG] can be collapsed over T without changing the odds ratio linking F and G. | Circle one:  TRUE FALSE |
| 1.7 In model [FTB][TBG],  feeling fat/overweight (F) is independent of gender (G). | Circle one:  TRUE FALSE |
| 1.8 In model [FTB][TBG], at each of the 4 levels of  T-and-B, the odds ratio linking feeling gender (G) and feeling fat/overweight (F) equals 1. | Circle one:  TRUE FALSE |
| 1.9 In model [FTB][G], G and B are **not** independent but **are** conditionally independent given F-and-T. | Circle one:  TRUE FALSE |

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2014 Final Exam: Answer Page 2 This is an exam. Do not discuss it with anyone. Due: Monday, May 12, at 12:00am**

|  |  |
| --- | --- |
|  | Fill in/circle the answer. |
| 2.1 Test the goodness of fit of model [FTB][G]. Give the value of the likelihood ratio chi-square, the degrees of freedom, the P-value. Based on this test alone, is the null hypothesis that this model is correct plausible? | Value (one number): 6.283948  Degrees of freedom: 7  P-value: 0.507013  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 2.2 Test the null hypothesis that [FTB][G] is the correct model against the alternative hypothesis that [FTB][TG] is the correct model. Give the value of the likelihood ratio chi-square test statistic, its degrees of freedom, and P-value and state whether the null hypothesis is plausible. | Value (one number): 0.760835  Degrees of freedom: 1  P-value: 0.3830673  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 2.3 Test the null hypothesis that [FT][FB][TB][G] is the correct model against the alternative hypothesis that [FTB][G] is the correct model. Answer the same questions for this comparison as in 2.2. | Value (one number): 12.7778  Degrees of freedom: 1  P-value: 0.0003507572  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 2.4 Model [FTB][G] can be collapsed over gender G without changing the odds ratios linking the other three variables, F, T and B. | Circle one:  TRUE FALSE |

|  |  |
| --- | --- |
| Use fitted counts from [FTB][G] for question 3. | Fill in or circle the correct answer. |
| 3.1 Give the four estimated odds ratios linking F and T at each level of B-and-G. You are to enter 4 numbers, and all 4 are >=1. | |  |  |  |  | | --- | --- | --- | --- | |  | G=Male | G=Female | | | B is <25 | 30.8 | 30.8 | | B is >=25 | 4.3 | 4.3 | |
| 3.2 Claim 3.2 on the data page is a reasonable summary of the odds ratios in 3.1. | Circle one:  TRUE FALSE |

Doing the Problem Set in R

**1.**

> loglin(nhanesWeight,list(1,2,3,4))

2 iterations: deviation 1.136868e-13

$lrt

[1] 715.5127

$pearson

[1] 978.587

$df

[1] 11

> 1-pchisq(715.5127,11)

[1] 0

**2.1**

> loglin(nhanesWeight,list(c(1,2,3),4))

2 iterations: deviation 2.273737e-13

$lrt

[1] 6.283948

$df

[1] 7

> 1-pchisq(6.283948,7)

[1] 0.507013

**2.2**

> loglin(nhanesWeight,list(c(1,2,3),c(2,4)))

2 iterations: deviation 5.684342e-14

$lrt

[1] 5.523113

$df

[1] 6

> 6.283948-5.523113

[1] 0.760835

> 7-6

[1] 1

> 1-pchisq(0.760835,1)

[1] 0.3830673

**2.3**

> loglin(nhanesWeight,list(c(1,2),c(1,3),c(2,3),4))

6 iterations: deviation 0.03227441

$lrt

[1] 19.06175

$df

[1] 8

> 19.06175-6.283948

[1] 12.7778

> 1-pchisq(12.7778,1)

[1] 0.0003507572

3.

> ft<-loglin(nhanesWeight,list(c(1,2,3),4),fit=T)$fit

2 iterations: deviation 2.273737e-13

> ft

, , bmi25 = <25, gender = male

tryingto

thinkfat lose.weight no.change

fat/overweight 33.367876 3.336788

about.right 94.860104 292.207254

, , bmi25 = >=25, gender = male

tryingto

thinkfat lose.weight no.change

fat/overweight 70.072539 5.243523

about.right 40.041451 12.870466

, , bmi25 = <25, gender = female

tryingto

thinkfat lose.weight no.change

fat/overweight 36.632124 3.663212

about.right 104.139896 320.792746

, , bmi25 = >=25, gender = female

tryingto

thinkfat lose.weight no.change

fat/overweight 76.927461 5.756477

about.right 43.958549 14.129534

> 33.367876\*292.207254/(94.860104\*3.336788)

[1] 30.80402

> 70.072539\*12.870466/(40.041451\*5.243523)

[1] 4.295455

> 36.632124\*320.792746/(104.139896\*3.663212)

[1] 30.80402

> 76.927461\*14.129534/(43.958549\*5.756477)

[1] 4.295454

**Statistics 501, Spring 2013, Midterm: Data Page #1**

Due in class, noon, Tuesday March 26, 2013

**This is an exam. Do not discuss it with anyone**. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Cheating on an exam is the single dumbest thing a PhD student at Penn can do.

Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question. **Due in class Tuesday March 26, 2013.**

The data for this problem are at in the latest Rst501.RData for R users as the object garki and in the garki.csv file at <http://stat.wharton.upenn.edu/statweb/course/Spring-2008/stat501> The list is case sensitive, so garki.csv is with lower case items.

The data are from a study by Molineaux and Gramiccia (1980) The GARKI project: Research on the epidemiology and control of malaria in the Sudan Savanna of West Africa, Geneva: World Health Organization. The study looked at several treatments and a control. We will look at one treatment and the control. The treatment involved spraying an insecticide, propoxur, and administering a drug, sulfalene-pyrimethamine at high frequency. The controls did not receive these interventions. In the example here, there are 1560 treated individuals matched in pairs to 1560 controls, the matching being for age and gender. The outcome is the frequency of Plasmodium falciparum in blood samples, that is the frequency of a protozoan parasite that causes malaria. A slide containing blood is divided into 200 fields and the outcome is the number of fields with the parasite, 0-200. Low numbers are better. Each person has two measures responses, one before the treatment period started, the other after the treatment period. Each response is the average of 2 to 4 blood samples. Each row of data contains a treated person (treated) and a match control person (control), their response before and after, their ages, their genders, and id numbers. The first 3 lines of data are below. In the first line, there are two men aged 35 years, the treated man declining from .50 before treatment to 0 after treatment, the control staying the same from .5 before to .5 after. In the control group, nothing happened between before and after, just the passage of time. You are to assume that the 1560 distinct pairs are independent, although of course the pairing may make the two people within a pair dependent.

> dim(garki)

[1] 1560 11

> garki[1:3,]

matched.id treated.before treated.after control.before

1 1 0.50 0 0.50

2 2 1.75 0 6.25

3 3 1.50 0 20.00

control.after treated.age control.age treated.male control.male

1 0.5 35 35 1 1

2 0.0 30 30 0 0

3 5.5 10 10 1 1

control treated

1 12059 6242

2 6209 6243

3 6109 6244

In your analyses of the garki data, please (i) assume that different rows of the garki data frame are independent, (ii) act as if the parasite levels were untied (that is, ignore ties, letting R do its thing with ties). The matching is very close but not perfect. In two of 1560 pairs, a male is paired with a female. Use treated.male=1 for a male pair, and treated.male=0 for a female pair, ignoring the two mismatches. STATISTICS 501, SPRING 2012, MIDTERM DATA PAGE #2

Due in class, noon, Tuesday March 26, 2013

Define a change as after-minus-before, for instance,

tamb <- treated.after-treated.before

camb <- control.after-control.before

Define the difference-in-differences to be dind <- tamb- camb or

dind <- (treated.after-treated.before)-(control.after-control.before)

Define three factors

> young<-factor(treated.age<=10,levels=c(F,T),labels=c("NotYoung","Young"))

> table(young)

young

NotYoung Young

1113 447

> male<-factor(treated.male,levels=c(0,1),labels=c("Female","Male"))

> table(male)

male

Female Male

766 794

> group<-young:male

> table(group)

group

NotYoung:Female NotYoung:Male Young:Female Young:Male

570 543 196 251

Question 3 asks you to look at this boxplot.

> boxplot(dind~young)

Question 4 refers to a variable, avage, which is the average age in a treated-versus-control pair.

> avage<-(treated.age+control.age)/2

Question 4 refers to a variable, young10, which is a binary version of the factor young.

> young10<-1\*(treated.age<=10)

Print Name **LAST name**, then first: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2013, Midterm, Answer Page #1 Due in class, noon, Tuesday March 26

**This is an exam. Do not discuss it with anyone**.

|  |  |
| --- | --- |
| Use the appropriate Wilcoxon test to answer the questions in part 1. | Fill in or CIRCLE the correct answer |
| 1.1 Apply the appropriate Wilcoxon test to the change in outcome among controls, camb. Give the two sided P-value, point estimate and 95% confidence interval. Is no change plausible? | P-value: \_\_\_\_\_ Estimate: \_\_\_\_\_\_ CI:[ , ]  Circle one:  Plausible Not plausible |
| 1.2 Apply the appropriate Wilcoxon test to the change in outcome among treated subjects, tamb. Give the two sided P-value, point estimate and 95% confidence interval. Is no change plausible? | P-value: \_\_\_\_\_ Estimate: \_\_\_\_\_\_ CI:[ , ]  Circle one:  Plausible Not plausible |
| 1.3 Apply the appropriate Wilcoxon test to whether the typical difference-in-differences is zero, dind. Give the two sided P-value, point estimate and 95% confidence interval. Is zero plausible? | P-value: \_\_\_\_\_ Estimate: \_\_\_\_\_\_ CI:[ , ]  Circle one:  Plausible Not plausible |
| 1.4 In question 1.3, the appropriate Wilcoxon test is Wilcoxon’s rank sum test (from chapter 4 in H&W) because treated and control groups are unrelated. | Circle one:  TRUE FALSE |
| 1.5 In question 1.2, the boxplot of changes for the 1560 treated subjects does not look symmetric about its center. So, the P-value in 1.2 has no meaning as a test of the null hypothesis of symmetry of changes about zero.. | Circle one:  Doesn’t look symmetric TRUE FALSE  P-value meaningless TRUE FALSE |
| 1.6 In question 1.1, there are 1,217,580 Walsh averages for the 1560 changes in the control group, and more than 80% of these are negative. | Circle one:  TRUE FALSE |
| 1.7 Based on your answer to 1.3, the treatment was associated with a greater increase parasites in the blood of treated subjects than in controls. | Circle one:  TRUE FALSE |

|  |  |
| --- | --- |
| 2. Define the group variable as on the data page. It refers to age and gender of the treated person in a pair. Use it to study how dind varies among the four groups. Use appropriate nonparametric tests for all comparisons. | Fill in or CIRCLE the correct answer |
| 2.1 Is it plausible that dind has the same distribution in the four groups defined by group? What is the name of the appropriate nonparametric test? What is the value of the test statistic? What is the P-value? Is the null hypothesis of no difference plausible? | Name of test:  Value: \_\_\_\_\_\_\_\_\_ P-value:\_\_\_\_\_\_\_\_\_\_\_\_\_  Circle one:  Plausible Not plausible |
| 2.2 Compare all six pairs of two of the four groups from 2.1. Use Holm’s method with an appropriate nonparametric test. List all pairs of groups as (A,B) that do not differ significantly at the 0.05 level, for instance (young.male,notyoungmale). List up to 6 pairs. If none, write none. |  |
| 2.3 In all data sets, Holm’s procedure rejects each hypothesis rejected by the Bonferroni method and may reject additional hypotheses. | Circle one:  TRUE FALSE |

Print Name **LAST name**, then first: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2013, Midterm, Answer Page #2 Due in class, noon, Tuesday March 26

**This is an exam. Do not discuss it with anyone**.

|  |  |
| --- | --- |
| 3. Compare dind by young on the data page, starting with the boxplot boxplot(dind~young) | Fill in or CIRCLE the correct answer |
| 3.1 The boxplot above suggests the quantity dind is lower in pairs with a treated subject 10 years or younger, but dind is also more dispersed in the young group. | Circle one:  TRUE FALSE |
| 3.2 Given what you saw in the boxplot, you cannot appropriately use Wilcoxon’s rank sum test to test the null hypothesis that dind has the same distribution in young and notyoung groups. | Circle one:  TRUE FALSE |
| 3.3 If you compare Young and NotYoung groups in terms of dind in all 1113x447=497511 possible ways, in more than 70% of such comparisons, the dind value is a larger number for the person in the NotYoung group. | Circle one:  TRUE FALSE |
| 3.4 If you assumed dind was symmetric about its medians in Young and NotYoung groups but the group dispersions were different, then you could not appropriately test that the two groups had equal medians (with possibly unequal dispersions) using Wilcoxon’s rank sum test in section 4.1 of Hollander and Wolfe (2nd ed) but you could use the method of Fligner and Policello in section 4.4. | Circle one:  TRUE FALSE |

|  |  |
| --- | --- |
| 4. Use the variables avage and young10 defined on the data page. You also need to install, then load the Rfit package. | Fill in or CIRCLE the correct answer |
| 4.1 What is the Kendall correlation between avage and dind? What is the P-value testing independence? What is the estimate of the probability of concordance? | Cor:\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P-value:\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Prob Concordance: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 4.2 Use rfit to fit a rank regression of dind on three predictors, avage, young10, treated.male. What is the estimated coefficient of avage in this regression? What is the P-value for testing the null hypothesis that the coefficient is zero. | Estimate:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  P-value:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 4.3 In the model you fitted using rfit in 4.2, test the one hypothesis that both the coefficient of avage and the coefficient of young 10 are simultaneously zero. Do this using the methods for a rank regression fitted by rfit. What is the value of the test statistic? What is the P-value? | Value:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  P-value:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 4.4 Appropriately used, the rfit function assumes that the errors around the linear model are Normal, independent, with equal dispersion. | Circle one:  TRUE FALSE |

Answers

Statistics 501, Spring 2013, Midterm, Answer Page #1

(H&W refers to the text, Hollander and Wolfe 1999 2nd Ed)

|  |  |
| --- | --- |
| Use the appropriate Wilcoxon test to answer the questions in part 1. | Fill in or CIRCLE the correct answer  6 points each |
| 1.1 Apply the appropriate Wilcoxon test to the change in outcome among controls, camb. Give the two sided P-value, point estimate and 95% confidence interval. Is no change plausible? | P-value: 2.2 x 10-16  Estimate: -4.79 CI: [-5.87, -3.87]  Circle one:  Plausible Not plausible |
| 1.2 Apply the appropriate Wilcoxon test to the change in outcome among treated subjects, tamb. Give the two sided P-value, point estimate and 95% confidence interval. Is no change plausible? | P-value: 2.2 x 10-16  Estimate: -17.5 CI:[-20.08, -15.12]  Circle one:  Plausible Not plausible |
| 1.3 Apply the appropriate Wilcoxon test to whether the typical difference-in-differences is zero, dind. Give the two sided P-value, point estimate and 95% confidence interval. Is zero plausible? | P-value: 2.2 x 10-16  Estimate: -5.52 CI:[-7.04, -4.25]  Circle one:  Plausible Not plausible |
| 1.4 In question 1.3, the appropriate Wilcoxon test is Wilcoxon’s rank sum test (from chapter 4 in H&W) because treated and control groups are unrelated. | Circle one:  TRUE FALSE  Signed rank, because of matched pairs |
| 1.5 In question 1.2, the boxplot of changes for the 1560 treated subjects does not look symmetric about its center. So, the P-value in 1.2 has no meaning as a test of the null hypothesis of symmetry of changes about zero. | Circle one:  Doesn’t look symmetric TRUE FALSE  P-value meaningless TRUE FALSE  H&W page 49, comment #14. |
| 1.6 In question 1.1, there are 1,217,580 Walsh averages for the 1560 changes in the control group, and more than 80% of these are negative. | I did not grade this question. There were lots of ties. A student pointed out that R’s (correct) handling of ties was surprising. I did not intend this to be a complex question, so I did not grade it. |
| 1.7 Based on your answer to 1.3, the treatment was associated with a greater increase parasites in the blood of treated subjects than in controls. | Circle one:  TRUE FALSE |

|  |  |
| --- | --- |
| 2. Define the group variable as on the data page. It refers to age and gender of the treated person in a pair. Use it to study how dind varies among the four groups. Use appropriate nonparametric tests for all comparisons. | Fill in or CIRCLE the correct answer  6 points each |
| 2.1 Is it plausible that dind has the same distribution in the four groups defined by group? What is the name of the appropriate nonparametric test? What is the value of the test statistic? What is the P-value? Is the null hypothesis of no difference plausible? | Name of test: Kruskal Wallis test.  Value: 200.65 P-value: 2.2 x 10-16  Circle one:  Plausible Not plausible |
| 2.2 Compare all six pairs of two of the four groups from 2.1. Use Holm’s method with an appropriate nonparametric test. List all pairs of groups as (A,B) that do **not** differ significantly at the 0.05 level, for instance (young.male,notyoungmale). List up to 6 pairs. If none, write none. | (NotYoung:Male, NotYoung:Female)  (Young:Male, Young:Female) |
| 2.3 In all data sets, Holm’s procedure rejects each hypothesis rejected by the Bonferroni method and may reject additional hypotheses. | Circle one:  TRUE FALSE |

Answers, continued

Statistics 501, Spring 2013, Midterm, Answer Page #2 Due in class, noon, Tuesday March 26

**This is an exam. Do not discuss it with anyone**.

|  |  |
| --- | --- |
| 3. Compare dind by young on the data page, starting with the boxplot boxplot(dind~young) | Fill in or CIRCLE the correct answer  6 points each |
| 3.1 The boxplot above suggests the quantity dind is lower in pairs with a treated subject 10 years or younger, but dind is also more dispersed in the young group. | Circle one:  TRUE FALSE |
| 3.2 Given what you saw in the boxplot, you cannot appropriately use Wilcoxon’s rank sum test to test the null hypothesis that dind has **the same distribution** in young and notyoung groups. | Circle one:  TRUE FALSE  H&W page 123, comment #14. |
| 3.3 If you compare Young and NotYoung groups in terms of dind in all 1113x447=497511 possible ways, in more than 70% of such comparisons, the dind value is a larger number for the person in the NotYoung group. | Circle one:  TRUE FALSE  H&W page 117, comment #7 |
| 3.4 If you assumed dind was symmetric about its medians in Young and NotYoung groups but the group dispersions were different, then you could not appropriately test that the **two groups had equal medians (with possibly unequal dispersions)** using Wilcoxon’s rank sum test in section 4.1 of Hollander and Wolfe (2nd ed) but you could use the method of Fligner and Policello in section 4.4. | Circle one:  TRUE FALSE  H&W page 135, section 4.4, paragraph 1 and H&W page 120, comment #11.  Wilcoxon’s test can test that the null hypothesis that two distributions are **the same** but not that they have **the same medians with different dispersions**. |

|  |  |
| --- | --- |
| 4. Use the variables avage and young10 defined on the data page. You also need to install, then load the Rfit package. | Fill in or CIRCLE the correct answer  6 points each, except 4.4 for 4 points |
| 4.1 What is the Kendall correlation between avage and dind? What is the P-value testing independence? What is the estimate of the probability of concordance? | Cor: 0.237 P-value: 2.2 x 10-16  Prob Concordance: 0.618 |
| 4.2 Use rfit to fit a rank regression of dind on three predictors, avage, young10, treated.male. What is the estimated coefficient of avage in this regression? What is the P-value for testing the null hypothesis that the coefficient is zero. | Estimate: 0.079  P-value: 0.000870 |
| 4.3 In the model you fitted using rfit in 4.2, test the one hypothesis that both the coefficient of avage and the coefficient of young 10 are simultaneously zero. Do this using the methods for a rank regression fitted by rfit. What is the value of the test statistic? What is the P-value? | Value: 373.77  P-value: 0.00  Use the drop.test function in the Rfit package:  it is analogous to the F-test of a general linear hypothesis |
| 4.4 Appropriately used, the rfit function assumes that the errors around the linear model are **Normal**, independent, with equal dispersion. | TRUE FALSE  Errors are not assumed Normal.  H&W page 439 section 9.6, assumption C2. |

Doing the Problem Set in R

Stat 501, Spring 2013, Midterm

> attach(garki)

> tamb<-treated.after-treated.before

> camb<-control.after-control.before

> dind<-tamb-camb

Question 1.1

> wilcox.test(camb,conf.int=T)

Wilcoxon signed rank test with continuity correction

data: camb

V = 216499, p-value < 2.2e-16

alternative hypothesis: true location is not equal to 0

95 percent confidence interval:

-5.874995 -3.874996

sample estimates:

(pseudo)median

-4.791616

Question 1.2

> wilcox.test(tamb,conf.int=T)

Wilcoxon signed rank test with continuity correction

data: tamb

V = 3269.5, p-value < 2.2e-16

alternative hypothesis: true location is not equal to 0

95 percent confidence interval:

-20.08338 -15.12499

sample estimates:

(pseudo)median

-17.49996

Question 1.3

> wilcox.test(dind,conf.int=T)

Wilcoxon signed rank test with continuity correction

data: dind

V = 336360.5, p-value < 2.2e-16

alternative hypothesis: true location is not equal to 0

95 percent confidence interval:

-7.041657 -4.250007

sample estimates:

(pseudo)median

-5.517102

Question 1.6

> 1560\*1561/2

[1] 1217580

> 216499/1217580

[1] 0.1778109

> 1-0.1778109

[1] 0.8221891

> 0.8221891>.8

[1] TRUE

Question 2.1

> kruskal.test(dind~group)

Kruskal-Wallis rank sum test

data: dind by group

Kruskal-Wallis chi-squared = 200.6542, df = 3, p-value < 2.2e-16

Question 2.2

> pairwise.wilcox.test(dind,group)

Pairwise comparisons using Wilcoxon rank sum test

data: dind and group

NotYoung:Female NotYoung:Male Young:Female

NotYoung:Male 0.990 - -

Young:Female 2.9e-12 2.9e-12 -

Young:Male < 2e-16 < 2e-16 0.083

P value adjustment method: holm

Question 4.1

> cor.test(avage,dind,method="k")

Kendall's rank correlation tau

data: avage and dind

z = 13.7758, p-value < 2.2e-16

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.2366609

> (0.2366609+1)/2

[1] 0.6183304

Question 4.2

> md<-rfit(dind~treated.age+young10+treated.male)

> summary(md)

Coefficients:

Estimate Std. Error t.value p.value

-3.276001 0.868523 -3.7719 0.0001681 \*\*\*

treated.age 0.079314 0.023775 3.3360 0.0008700 \*\*\*

young10 -30.924315 0.921458 -33.5602 < 2.2e-16 \*\*\*

treated.male -0.982663 0.547096 -1.7961 0.0726660 .

---

Multiple R-squared (Robust): 0.3260413

Reduction in Dispersion Test: 250.9156 p-value: 0

Question 4.3

Compare the full model to the reduced model.

> mdr<-rfit(dind~treated.male)

rfit.default(formula = dind ~ treated.male)

treated.male

-1.000000 -1.249973

> drop.test(md,mdr)

Drop in Dispersion Test

F-Statistic p-value

373.77 0.00

**Statistics 501 Spring 2013 Final Exam: Data Page 1**

**This is an exam. Do not discuss it with anyone.**

**Due: Monday, April 29, at 11:00am**

The data are from NHANES 2009-2010. It is a 25 table. The data are in the nhanesD object in the workspace, and the table is below. As suggested by NHANES, a person is judged depressed if their score on the 9 item depression screener (DPQ) is 10 or more. The other variables are alcohol last year, age, married, and gender.

> dimnames(nhanesD)

$depressed

[1] "Depressed" "Not Depressed"

$alcohol

[1] "<12 drinks last year" ">=12 drinks last year"

$age

[1] "<50" ">=50"

$married

[1] "married" "other"

$gender

[1] "male" "female"

**IMPORTANT**: **Please refer to the variables with the letters** d=depressed, b=alcohol (booze), a=age, m=married and g=gender. Use the margin-preservation notation with these letters to refer to log-linear models. For example, you would refer to model of independence as [d][b][a][m][g]. **Use the likelihood ratio chi-square**.

d-1 b-2 a-3 m-4 g-5

"depressed" "alcohol" "age" "married" "gender"

**Questions 2.3-2.6** asks you to calculate two odds ratios or probabilities from fitted counts. This refers to the fitted counts in the full nhanesD table as that table is currently structured, with the 11 and 22 cells in the numerator. The program, loglin, fits iteratively, and the question asks you to set eps=0.000001 in the loglin call and report odds ratios to 2 significant digits. If the odds are twice as great, we speak in English as twice as likely.

**Make and keep a photocopy of your answer page**. **The exam is due in my office, 473 Huntsman, on Monday April 29 at 11:00am.** You may turn in the exam early at my mail box in the Statistics Department, 4th floor, Huntsman or by giving it to Adam at the front desk in statistics, but if you turn in the exam early, place it in an envelope addressed to me. When all of the exams are graded, I will add an **answer key** to the on-line bulk-pack for the course. You can compare the answer key to your photocopy of your exam. Your course grade will be available from the Registrar. I no longer distribute answer keys and graded exams by US Mail. **Turn in only the answer page**. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question

**Have a great summer!**

**This is an exam. Do not discuss it with anyone.**

> nhanesD

, , age = <50, married = married, gender = male

alcohol

depressed <12 drinks last year >=12 drinks last year

Depressed 5 27

Not Depressed 70 499

, , age = >=50, married = married, gender = male

alcohol

depressed <12 drinks last year >=12 drinks last year

Depressed 6 36

Not Depressed 154 695

, , age = <50, married = other, gender = male

alcohol

depressed <12 drinks last year >=12 drinks last year

Depressed 10 53

Not Depressed 71 559

, , age = >=50, married = other, gender = male

alcohol

depressed <12 drinks last year >=12 drinks last year

Depressed 6 35

Not Depressed 61 316

, , age = <50, married = married, gender = female

alcohol

depressed <12 drinks last year >=12 drinks last year

Depressed 19 43

Not Depressed 201 374

, , age = >=50, married = married, gender = female

alcohol

depressed <12 drinks last year >=12 drinks last year

Depressed 25 24

Not Depressed 235 328

, , age = <50, married = other, gender = female

alcohol

depressed <12 drinks last year >=12 drinks last year

Depressed 33 81

Not Depressed 182 458

, , age = >=50, married = other, gender = female

alcohol

depressed <12 drinks last year >=12 drinks last year

Depressed 42 58

Not Depressed 270 293

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2013 Final Exam: Answer Page 1 This is an exam. Do not discuss it with anyone. Due Monday, April 29, 2013, at 11:00am.**

|  |  |
| --- | --- |
| Use letters d, b, a, m and g to refer to variables and the [] notation to refer to log-linear models. In question 1.1, this is done for you. See the data page. | Fill in or CIRCLE the correct answer |
| 1.1 Which log-linear model says the five variables are independent? Give the numerical value of the likelihood ratio test of fit, its degrees of freedom (df), the P-value (P). Is independence of all 5 variables plausible? | Model: [d][b][a][m][g]  Value: \_\_\_\_ df:\_\_\_\_ P:\_\_\_\_\_\_  Circle One:  Plausible Not Plausible |
| 1.2 Which log-linear model says that alcohol last year (b) and marital status (m) are conditionally independent given the other three variables? Write in the model, as in 1.1. Give the numerical value of the likelihood ratio test of fit, its degrees of freedom (df), the P-value (P). Based just on the this test of fit, is this model plausible? | Model: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Value: \_\_\_\_ df:\_\_\_\_ P:\_\_\_\_\_\_  Circle One:  Plausible Not Plausible |
| 1.3 Which log-linear model says being depressed (d) and alcohol (b) are related, but related in a simple way, specifically with the same odds ratio at all values of the other variables, and subject to that condition, the other variables may have any relationship at all? Write in the model. Give the numerical value of the likelihood ratio test of fit, its degrees of freedom (df), the P-value (P). Based just on the this test of fit, is this model plausible? | Model:  Value: \_\_\_\_ df:\_\_\_\_ P:\_\_\_\_\_\_  Circle One:  Plausible Not Plausible |
| 1.4 Consider [dm][dg][ba][bg][amg] as the model. Give the value of the likelihood ratio test of fit, its degrees of freedom (df), the P-value (P). Based just on the this test of fit, is this model plausible? | Value: \_\_\_\_\_\_\_ df:\_\_\_\_\_\_ P:\_\_\_\_\_\_\_\_\_  Circle One:  Plausible Not Plausible |
| 1.5 The model [dm][dg][ba][bg][amg] in question 1.4 says being depressed (d) is independent of age (a). | Circle One:  TRUE FALSE |
| 1.6 The model [dm][dg][ba][bg][amg] in question 1.4 says the odds ratio (in the full nhanesD table) linking being depressed (d) with being married (m) is different for men and women (g). | Circle One:  TRUE FALSE |

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2013 Final Exam: Answer Page 2 This is an exam. Do not discuss it with anyone. Due Monday, April 29, 2013, at 11:00am.**

|  |  |
| --- | --- |
| See the data page. | Fill in or CIRCLE the correct answer |
| 2.1 Test the null hypothesis that the model [dm][dg][ba][bg][amg] in 1.4 is adequate against the alternative hypothesis that the [dmg] u-term needs to be added to the model. Give the value of chi-square for this test, its degrees of freedom, and indicate whether the null hypothesis is plausible. | Value: \_\_\_\_\_\_\_ df:\_\_\_\_\_\_ P:\_\_\_\_\_\_\_\_\_  Circle One:  Plausible Not Plausible |
| 2.2 Are there any u-terms in the model [dm][dg][ba][bg][amg] that may be removed without a statistically significant degradation in fit at the 0.05 level. If yes, list any and all u-terms that may be removed (one at a time) without degradation of fit. If none, write “none”. |  |
| 2.3 In the model [dm][dg][ba][bg][amg], give the fitted odds ratio linking being depressed with being married for men, under 50, who had fewer than 12 alcoholic drinks last year. Repeat for women, under 50, who had fewer than 12 alcoholic drinks last year. See the data page. | Set eps=0.000001 in the loglin call. Report odds ratios to 2 significant digits.  Men, <50, <12 drinks: \_\_\_\_\_\_\_\_\_\_\_\_\_  Women, <50, <12 drinks: \_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 2.4 Consider just individuals who are under 50 and had fewer than 12 alcoholic drinks last year. For these individuals, based on your answer to question 2.3, the point estimates of odds ratios suggest that married men are about half as likely as unmarried men to be depressed, but women in this group are twice as likely to be depressed. | Circle One:  TRUE FALSE |
| 2.5 In the model [dm][dg][ba][bg][amg], under age 50, with fewer than 12 alcoholic drinks last year, married men are estimated to be about 5 times more likely than married women to be depressed. (Use fitted odds ratios). | Circle One:  TRUE FALSE |
| 2.6 In model [dm][dg][ba][bg][amg], married men under age 50 with fewer than 12 drinks last year are estimated to have a probability of depression of about 0.05. | Circle One:  TRUE FALSE |

**Stat 501 S-2013 Final Exam: Answer Page 1 Answers.**

|  |  |
| --- | --- |
| Use letters d, b, a, m and g to refer to variables and the [] notation to refer to log-linear models. In question 1.1, this is done for you. See the data page. | Fill in or CIRCLE the correct answer  8 points each except as noted |
| 1.1 Which log-linear model says the five variables are independent? Give the numerical value of the likelihood ratio test of fit, its degrees of freedom (df), the P-value (P). Is independence of all 5 variables plausible? | Model: [d][b][a][m][g]  Value: 729.894 df: 26 P: ~0  Circle One:  Plausible Not Plausible |
| 1.2 Which log-linear model says that alcohol last year (b) and marital status (m) are conditionally independent given the other three variables? Write in the model, as in 1.1. Give the numerical value of the likelihood ratio test of fit, its degrees of freedom (df), the P-value (P). Based just on the this test of fit, is this model plausible? (10 points) | Model: [dbag][damg]  Value: 12.505 df: 8 P: 0.13  Circle One:  Plausible Not Plausible |
| 1.3 Which log-linear models says being depressed (d) and alcohol (b) are related, but related in a simple way, specifically with the same odds ratio at all values of the other variables, and subject to that condition, the other variables may have any relationship at all? Write in the model, as in 1.1. Give the numerical value of the likelihood ratio test of fit, its degrees of freedom (df), the P-value (P). Based just on the this test of fit, is this model plausible? (10 points) | Model: [db][damg][bamg]  Value: 5.11 df: 7 P: 0.64  Circle One:  Plausible Not Plausible |
| 1.4 Consider [dm][dg][ba][bg][amg] as the model. Give the value of the likelihood ratio test of fit, its degrees of freedom (df), the P-value (P). Based just on the this test of fit, is this model plausible? | Value: 17.6 df: 18 P: 0.48  Circle One:  Plausible Not Plausible |
| 1.5 The model [dm][dg][ba][bg][amg] in question 1.4 says being depressed (d) is independent of age (a). | Circle One:  TRUE FALSE |
| 1.6 The model [dm][dg][ba][bg][amg] in question 1.4 says the odds ratio (in the full nhanesD table) linking being depressed (d) with being married (m) is different for men and women (g). | Circle One:  TRUE FALSE |

**Stat 501 S-2013 Final Exam: Answer Page 2 Answers.**

|  |  |
| --- | --- |
| See the data page. | Fill in or CIRCLE the correct answer |
| 2.1 Test the null hypothesis that the model [dm][dg][ba][bg][amg] in 1.4 is adequate against the alternative hypothesis that the [dmg] u-term needs to be added to the model. Give the value of chi-square for this test, its degrees of freedom, and indicate whether the null hypothesis is plausible. | Value: 0.166 df: 1 P: 0.68  Circle One:  Plausible Not Plausible |
| 2.2 Are there any u-terms in the model [dm][dg][ba][bg][amg] that may be removed without a statistically significant degradation in fit at the 0.05 level. If yes, list any and all u-terms that may be removed (one at a time) without degradation of fit. If none, write “none”. | none |
| 2.3 In the model [dm][dg][ba][bg][amg], give the fitted odds ratio linking being depressed with being married for men, under 50, who had fewer than 12 alcoholic drinks last year. Repeat for women, under 50, who had fewer than 12 alcoholic drinks last year. See the data page. | Set eps=0.000001 in the loglin call. Report odds ratios to 2 significant digits.  Men, <50, <12 drinks: 0.48  Women, <50, <12 drinks: 0.48 |
| 2.4 Consider just individuals who are under 50 and had fewer than 12 alcoholic drinks last year. For these individuals, based on your answer to question 2.3, the point estimates of odds ratios suggest that married men are about half as likely as unmarried men to be depressed, but women in this group are twice as likely to be depressed. | Circle One:  TRUE FALSE  Married men are half as likely as unmarried men to be depressed, but the same is true for married women. |
| 2.5 In the model [dm][dg][ba][bg][amg], under age 50, with fewer than 12 alcoholic drinks last year, married men are estimated to be about 5 times more likely than married women to be depressed. (Use fitted odds ratios). | Circle One:  TRUE FALSE  1/5 as likely, not 5 times as likely. |
| 2.6 In model [dm][dg][ba][bg][amg], married men under age 50 with fewer than 12 drinks last year are estimated to have a probability of depression of about 0.05. | Circle One:  TRUE FALSE |

Doing the Problem Set in R

(Final Spring 2013 Statistics 501)

**1.1**

> loglin(nhanesD,list(1,2,3,4,5))

2 iterations: deviation 9.094947e-13

$lrt

[1] 729.894

$df

[1] 26

> 1-pchisq(729.894,26)

[1] 0

**1.2**

> loglin(nhanesD,list(c(1,2,3,5),c(1,3,4,5)))

2 iterations: deviation 1.136868e-13

$lrt

[1] 12.5054

$df

[1] 8

$margin

$margin[[1]]

[1] "depressed" "alcohol" "age" "gender"

$margin[[2]]

[1] "depressed" "age" "married" "gender"

> 1-pchisq(12.5054,8)

[1] 0.1300384

**1.3**

> loglin(nhanesD,list(c(1,2),c(1,3,4,5),c(2,3,4,5)))

4 iterations: deviation 0.03090209

$lrt

[1] 5.111691

$df

[1] 7

$margin

$margin[[1]]

[1] "depressed" "alcohol"

$margin[[2]]

[1] "depressed" "age" "married" "gender"

$margin[[3]]

[1] "alcohol" "age" "married" "gender"

> 1-pchisq(5.111691,7)

[1] 0.6463351

**1.4**

> loglin(nhanesD,list(c(1,4),c(1,5),c(2,3),c(2,5),c(3,4,5)))

5 iterations: deviation 0.01719219

$lrt

[1] 17.5982

$df

[1] 18

$margin

$margin[[1]]

[1] "depressed" "married"

$margin[[2]]

[1] "depressed" "gender"

$margin[[3]]

[1] "alcohol" "age"

$margin[[4]]

[1] "alcohol" "gender"

$margin[[5]]

[1] "age" "married" "gender"

> 1-pchisq(17.5982,18)

[1] 0.4824019

**2.1** Compare two nested models, the following model and the one in 1.4.

> loglin(nhanesD,list(c(1,4,5),c(2,3),c(2,5),c(3,4,5)))

5 iterations: deviation 0.01679229

$lrt

[1] 17.43217

$df

[1] 17

$margin

$margin[[1]]

[1] "depressed" "married" "gender"

$margin[[2]]

[1] "alcohol" "age"

$margin[[3]]

[1] "alcohol" "gender"

$margin[[4]]

[1] "age" "married" "gender"

> 17.5982-17.43217

[1] 0.16603

> 18-17

[1] 1

> 1-pchisq(0.16603,1)

[1] 0.6836644

**2.3** You set eps==0.000001 to ensure you are close to convergence.

> ft<-loglin(nhanesD,list(c(1,4),c(1,5),c(2,3),c(2,5),c(3,4,5)),fit=T,

eps=0.000001,iter=30)$fit

10 iterations: deviation 1.730364e-07 (Notice: 10 iterations)

> ft[,1,1,,1]

married

depressed married other

Depressed 3.500608 7.269726

Not Depressed 65.192347 71.938623

> or(ft[,1,1,,1])

[1] 0.4815323

> ft[,1,1,,2]

married

depressed married other

Depressed 17.70299 36.61797

Not Depressed 185.21136 203.56637

> or(ft[,1,1,,2])

[1] 0.4834511

**2.6**

> ft[,1,1,,1]

married

depressed married other

Depressed 3.500608 7.269726

Not Depressed 65.192347 71.938623

> 3.500608/(3.500608+65.192347)

[1] 0.05096022

**Statistics 501, Spring 2012, Midterm: Data Page #1**

Due in class, noon, Tuesday March 27, 2012

**This is an exam. Do not discuss it with anyone**. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Cheating on an exam is the single dumbest thing a PhD student at Penn can do.

Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question. **Due in class Tuesday 29 March 2011.**

The data for this problem are at in the latest Rst501.RData for R users as the object bmi501 and in the bmi501 file at <http://stat.wharton.upenn.edu/statweb/course/Spring-2008/stat501> The list is case sensitive, so nhanes501.txt is with lower case items.

The data are from the US National Health and Nutrition Examination Survey (NHANES) for 2007-2008. You can obtain the complete survey from ICPSR via the Penn library web page or directly from the CDC, but there is no reason to do this for the current exam. The data consist of 676 matched pairs of one daily smoker and one nonsmoker. A daily smoker reported smoking on every day of the past 30 days (SMD641=30) and having smoked at least 100 cigarettes in his or her lifetime (SMQ020=YES). A nonsmoker reports having smoked fewer than 100 cigarettes in his or her lifetime (SMQ020=NO) and has no reported smoking in the previous 30 days (SMD641=missing). Data for a smoker begins with S, while data for a nonsmoking control begins with C. The pairs were matched for education (Educ, higher=more), Income (ratio to poverty level), Black (1=yes), Female (1=yes), Married (1=yes), and Age. In the first row of bmi501, an unmarried female smoker aged 77 is paired with an unmarried female nonsmoker aged 79.

It is often said that smoking depresses appetite. People sometimes say that they are reluctant to quit smoking for fear of gaining weight. What do data say about this? The dataset also contains BMI for smokers and controls. See <http://www.nhlbisupport.com/bmi/> In the first row of bmi501, the smoker weighed a little less, BMI = 19.96 for the smoker versus BMI = 22.71 for the control.

> dim(bmi501)

[1] 676 14

> bmi501[1,]

Seduc Sincome Sblack Sfemale Smarried Sage

1 2 1.57 0 1 0 77

SBMI Ceduc Cincome Cblack Cfemale Cmarried

1 19.96 2 1.67 0 1 0

Cage CBMI

1. 79 22.71

You will need to calculate the matched pair difference in BMI

>attach(bmi501)

> dif<-SBMI-CBMI

You will need the variable grp, which equals the variable grp2. Spend some time to make sure you understand what the levels of grp means and what it means that grp = grp2.

> grp<-factor(SMarried):factor(SFemale)

> grp2<-factor(CMarried):factor(CFemale)

> table(grp,grp2)

grp2

grp 0:0 0:1 1:0 1:1

0:0 227 0 0 0

0:1 0 174 0 0

1:0 0 0 162 0

1:1 0 0 0 113

STATISTICS 501, SPRING 2012, MIDTERM DATA PAGE #2

Due in class, noon, Tuesday March 27, 2012

Please assume that the 676 matched pairs are independent for distinct pairs and that they represent 676 independent draws from a single multivariate (ie many variable) distribution.

The model for **question 2.2-2.4** has the difij =  + j + eij where there are groups j = 1, 2, 3, 4. You are asked to test H0: 1 = 2 = 3= 4 against a general alternative, and six hypotheses of the form H12: 1 = 2 , H13: 1 = 3 , H14: 1 = 4 , H23: 2 = 3, H24: 2 = 4 , H34: 3 = 4.

Question 3 asks you to use the rfit function in the Rfit package to fit and compare two regressions using just data on the 676 controls.

Model 1: CBMI = 0 + age CAge + female CFemale + e with e iid, symmetric about 0, continuous.

Model 2: CBMI = 0 + age CAge + female CFemale + educ CEduc + income CIncome + u with u iid, symmetric about 0, continuous.Print Name **Last name**, then First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2012, Midterm, Answer Page #1 Due noon, Tuesday March 27, 2012

**This is an exam. Do not discuss it with anyone**.

|  |  |
| --- | --- |
| Use bmi501 to answer these questions. | Fill in or CIRCLE the Correct Answer |
| 1.1 The median smoker and the median nonsmoker are both overweight (BMI 25-29.9). | TRUE FALSE |
| 1.2 More than a quarter of smokers and more than a quarter of nonsmokers are obese (BMI of 30 or more). | TRUE FALSE |
| 1.3 In the dataset, men are always paired with men and women are always paired with women. | TRUE FALSE |
| 1.4 Use the Shapiro-Wilk test to test the null hypothesis that the matched pair differences “dif” in BMI are Normally distributed. Give the P-value. Is the null hypothesis plausible? | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 1.5 Use the appropriate Wilcoxon procedure to test the null hypothesis that the smoker-minus-control matched pair differences are symmetrically distributed about zero. Give the 2-sided P-value. Is the null hypothesis plausible? | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 1.6 For the test you did in 1.5, give the corresponding point estimate of the center of symmetry of the smoker-minus-control matched pair differences. Is this point estimate the median of the choose(676,2)= 228150 pairwise differences between a smoker and a matched control? (Yes or No). | Point estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Circle one:  YES NO |
| 1.7 Give the 2-sided 95% confidence interval for the center of symmetry of the smoker-minus-control pair differences based on the test you did in 1.5. Give the interval. Smoking is associated with an increase in BMI. (True or false). | Confidence interval: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Circle one:  TRUE FALSE |
| 1.8 Use an appropriate t-test to construct the confidence interval for the center of symmetry of the smoker-minus-control pair differences. Give the interval. The t-interval is more than 8% longer than the Wilcoxon interval. (True or false). | Confidence interval: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Circle one:  TRUE FALSE |
| 1.9 A central limit theorem says t has more power than Wilcoxon in large samples. | Circle one:  TRUE FALSE |

Print Name Clearly, **Last**, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2012, Midterm, Answer Page #2 Due noon, Tuesday March 27, 2012

This is an exam. Do not discuss it with anyone.

|  |  |
| --- | --- |
| 2.1 Use an appropriate Wilcoxon procedure to test the null hypothesis that the smoker-minus-control pair difference “dif” in BMI has the same distribution for men and women. Give the two-sided P-value and the associated point estimate for the difference. Is the hypothesis plausible? | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Point estimate:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 2.2 Use an appropriate nonparametric analog of the F-test (from H&W) to test the null hypothesis that the four levels of “grp” have the same distribution of “dif”. (See the data page for definitions.) What is the name of the test? What is the P-value? Is the null hypothesis plausible? | Name of test:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 2.3 Use Holm’s procedure with an appropriate Wilcoxon procedure to test all pairwise comparisons in 2.2. What is the smallest of the 6 adjusted P-values from Holm’s procedure? | Smallest of 6 pairwise P-values after adjustment using Holms method:  P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 2.4 When using Holm’s procedure, it is logically possible that there is exactly one false hypothesis, namely H14: 1 = 4 . | Circle one:  TRUE FALSE |
| 2.5 If one looks at two of the 676 matched pairs picked at random, what is the estimate of the probability that the pair with the higher CEduc also has the higher SEduc? | Estimated probability: \_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

|  |  |
| --- | --- |
| Question 3 refers to models 1 and 2 on the data page and asks you to fit them using the rfit function in the Rfit package. | Fill in or CIRCLE the Correct Answer |
| 3.1 In model 1 on the data page, test the null hypothesis H0: age = 0 using the nonparametric analog of the partial t-test in regression. Give the 2-sided P-value. Is H0 plausible as judged by the conventional 0.05 standard? | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 3.2 In model 2 on the data page, test the null hypothesis H0: educ = income = 0 using the nonparametric analog of the partial F-test (aka general linear hypothesis). Give the P-value. Is H0 plausible as judged by the conventional 0.05 standard? | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Circle one:  PLAUSIBLE NOT PLAUSIBLE |

Statistics 501, Spring 2012, Midterm, Answer Page #1, ANSWERS

**This is an exam. Do not discuss it with anyone**.

|  |  |
| --- | --- |
| Use bmi501 to answer these questions. | Fill in or CIRCLE (6 points each) |
| 1.1 The median smoker and the median nonsmoker are both overweight (BMI 25-29.9). | TRUE FALSE |
| 1.2 More than a quarter of smokers and more than a quarter of nonsmokers are obese (BMI of 30 or more). | TRUE FALSE |
| 1.3 In the dataset, men are always paired with men and women are always paired with women. | TRUE FALSE |
| 1.4 Use the Shapiro-Wilk test to test the null hypothesis that the matched pair differences “dif” in BMI are Normally distributed. Give the P-value. Is the null hypothesis plausible? | P-value: 1.765e-07  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 1.5 Use the appropriate Wilcoxon procedure to test the null hypothesis that the smoker-minus-control matched pair differences are symmetrically distributed about zero. Give the 2-sided P-value. Is the null hypothesis plausible? | P-value: 9.045e-08  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 1.6 For the test you did in 1.5, give the corresponding point estimate of the center of symmetry of the smoker-minus-control matched pair differences. Is this point estimate the median of the choose(676,2)= 228150 pairwise differences between a smoker and a matched control? (Yes or No). | Point estimate: -1.92  Circle one:  YES NO  It is the median of the pairwise averages, not the pairwise differences. |
| 1.7 Give the 2-sided 95% confidence interval for the center of symmetry of the smoker-minus-control pair differences based on the test you did in 1.5. Give the interval. Smoking is associated with an increase in BMI. (True or false). | Confidence interval: [-2.60, -1.23]  Circle one:  TRUE FALSE |
| 1.8 Use an appropriate t-test to construct the confidence interval for the center of symmetry of the smoker-minus-control pair differences. Give the interval. The t-interval is more than 8% longer than the Wilcoxon interval. (True or false). | Confidence interval: [-2.77, -1.29]  Circle one:  TRUE FALSE |
| 1.9 A central limit theorem says t has more power than Wilcoxon in large samples. | Circle one:  TRUE FALSE |

For 1.9, t has best power if the data are Normal, but not in general.Statistics 501, Spring 2012, Midterm, Answer Page #2

This is an exam. Do not discuss it with anyone.

|  |  |
| --- | --- |
| 2.1 Use an appropriate Wilcoxon procedure to test the null hypothesis that the smoker-minus-control pair difference “dif” in BMI has the same distribution for men and women. Give the two-sided P-value and the associated point estimate for the difference. Is the hypothesis plausible? | P-value: 0.613  Point estimate: 0.360  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 2.2 Use an appropriate nonparametric analog of the F-test (from H&W) to test the null hypothesis that the four levels of “grp” have the same distribution of “dif”. (See the data page for definitions.) What is the name of the test? What is the P-value? Is the null hypothesis plausible? | Name of test: Kruskal-Wallis  P-value: 0.607  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 2.3 Use Holm’s procedure with an appropriate Wilcoxon procedure to test all pairwise comparisons in 2.2. What is the smallest of the 6 adjusted P-values from Holm’s procedure? | Smallest of 6 pairwise P-values after adjustment using Holms method:  P-value: 1 |
| 2.4 When using Holm’s procedure, it is logically possible that there is exactly one false hypothesis, namely H14: 1 = 4 . | TRUE FALSE  If 1 and 4 are unequal, they can’t both equal 2, so there can’t just be 1 false hypothesis. |
| 2.5 If one looks at two of the 676 matched pairs picked at random, what is the estimate of the probability that the pair with the higher CEduc also has the higher SEduc? | Estimated probability: 0.972  This is the probability of concordance from Kendall’s correlation. |

|  |  |
| --- | --- |
| Question 3 refers to models 1 and 2 on the data page and asks you to fit them using the rfit function in the Rfit package. | Fill in or CIRCLE (8 points each) |
| 3.1 In model 1 on the data page, test the null hypothesis H0: age = 0 using the nonparametric analog of the partial t-test in regression. Give the 2-sided P-value. Is H0 plausible as judged by the conventional 0.05 standard? | P-value: 0.04996  Circle one:  PLAUSIBLE NOT PLAUSIBLE |
| 3.2 In model 2 on the data page, test the null hypothesis H0: educ = income = 0 using the nonparametric analog of the partial F-test (aka general linear hypothesis). Give the P-value. Is H0 plausible as judged by the conventional 0.05 standard? | P-value: 0.779  Circle one:  PLAUSIBLE NOT PLAUSIBLE |

**Doing the Problem Set in R: Spring 2012 Midterm St 501**

1.1 and 1.2

> attach(bmi501)

> boxplot(SBMI,CBMI)

> summary(SBMI)

Min. 1st Qu. Median Mean 3rd Qu. Max.

14.20 22.99 26.48 27.63 31.02 63.95

> summary(CBMI)

Min. 1st Qu. Median Mean 3rd Qu. Max.

15.69 25.31 28.48 29.66 32.64 73.43

1.3

> table(SFemale,CFemale)

CFemale

SFemale 0 1

0 389 0

1. 0 287

1.4

> dif<-SBMI-CBMI

> shapiro.test(dif)

Shapiro-Wilk normality test

data: dif

W = 0.9817, p-value = 1.765e-07

1.5-1.7

> wilcox.test(dif,conf.int=T)

Wilcoxon signed rank test: dif

V = 87263.5, p-value = 9.045e-08

95 percent confidence interval:

-2.600036 -1.230013

sample estimates:

-1.919962

1.8

> t.test(dif)

One Sample t-test data: dif

t = -5.3672, df = 675, p-value = 1.101e-07

95 percent confidence interval:

-2.769503 -1.285911

sample estimates: mean of x

-2.027707

> (-1.285911)-(-2.769503)

[1] 1.483592

> (-1.230013)-(-2.600036)

[1] 1.370023

> 1.483592/1.370023

[1] 1.082896

> wilcox.test(dif~SFemale,conf.int=T)

Wilcoxon rank sum test data: dif by SFemale

W = 57091.5, p-value = 0.613

95 percent confidence interval:

-1.049956 1.770011

sample estimates: difference in location

0.3599914

2.2

> kruskal.test(dif~grp)

Kruskal-Wallis rank sum test data: dif by grp

Kruskal-Wallis chi-squared = 1.8368, df = 3, p-value = 0.607

2.3

> pairwise.wilcox.test(dif,grp)

Pairwise comparisons using Wilcoxon rank sum test

data: dif and grp P value adjustment method: holm

0:0 0:1 1:0

0:1 1 - -

1:0 1 1 -

1:1 1 1 1

2.5

> cor.test(SEduc,CEduc,method="kendall")

Kendall's rank correlation tau data: SEduc and CEduc

z = 29.5666, p-value < 2.2e-16

sample estimates: tau

0.9446568

> (0.94465685+1)/2

[1] 0.9723284

3.1-2

> out<-rfit(CBMI~CAge+CFemale)

> summary(out)

Coefficients:

Estimate Std. Error t.value p.value

26.985321 0.691313 39.0349 < 2e-16 \*\*\*

CAge 0.026457 0.013472 1.9638 0.04996 \*

CFemale 0.691179 0.441895 1.5641 0.11826

> out2<-rfit(CBMI~CAge+CFemale+CIncome+CEduc)

> summary(out2)

Coefficients:

Estimate Std. Error t.value p.value

27.316813 0.921498 29.6439 < 2e-16 \*\*\*

CAge 0.025602 0.013506 1.8956 0.05844 .

CFemale 0.734714 0.449071 1.6361 0.10229

CIncome 0.028365 0.164397 0.1725 0.86307

CEduc -0.151863 0.221289 -0.6863 0.49278

> drop.test(out2,out)

Drop in Dispersion Test

F-Statistic p-value

0.25046 0.77851

**Statistics 501, Spring 2011, Midterm: Data Page #1**

**This is an exam. Do not discuss it with anyone**. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Cheating on an exam is the single dumbest thing a PhD student at Penn can do.

Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question. **Due in class Tuesday 29 March 2011.**

The data for this problem are at in the latest Rst501.RData for R users as the object nhanes501 and in nhanes501.txt as a text file at <http://stat.wharton.upenn.edu/statweb/course/Spring-2008/stat501> The list is case sensitive, so nhanes501.txt is with lower case items.

The data are from the US National Health and Nutrition Examination Survey (NHANES) for 2007-2008. You can obtain the complete survey from ICPSR via the Penn library web page, but there is no reason to do this for the current exam. The data consist of 250 matched pairs of one daily smoker and one nonsmoker. A daily smoker reported smoking on every day of the past 30 days (SMD641=30) and having smoked at least 100 cigarettes in his or her lifetime (SMQ020=YES). A nonsmoker reports having smoked fewer than 100 cigarettes in his or her lifetime (SMQ020=NO) and has no reported smoking in the previous 30 days (SMD641=missing). The pairs were matched for gender, age, Hispanic or black or other, education level, household income level, and missing value indicators for education and income.

LBXBCD is the blood level of cadmium in g/dL, and LBXBPB is is the blood level of lead in g/dL, where LBXBCDsmk is for the smoker in a pair, LBXBCDcont is for the nonsmoker (control) in the pair, and LBXBCDdif is the difference, .82-.37=.45 for pair 1. SMD650smk is for the smoker in the pair: it is the answer to “During the past 30 days, on the days that you smoked, about how many cigarettes did you smoke per day? 1 pack = 20 cigarettes. If >95, enter 95”. There is one missing value for SMD650smk – it is an NA. The variable female = 1 if the smoker is female, but in almost all pairs, the two individuals are of the same gender.

> **dim(nhanes501)**

[1] 250 9

> **round(nhanes501,2)[1:3,]**

id LBXBCDsmk LBXBCDcont LBXBCDdif LBXBPBsmk LBXBPBcont LBXBPBdif SMD650smk female

1 1 0.82 0.37 0.45 0.86 1.30 -0.44 1 1

2 2 3.00 0.27 2.73 2.60 0.82 1.78 3 1

3 3 0.44 0.53 -0.09 1.71 3.40 -1.69 3 1

STATISTICS 501, SPRING 2011, MIDTERM DATA PAGE #2

For problem 1.2, the codes are:

1. The matched pair differences in blood cadmium levels are approximately Normal.
2. The matched pair differences in blood cadmium levels have a thick right tail compared to the Normal (i.e., too many large positive values)
3. The matched pair differences in blood cadmium levels have a thick left tail compared to the Normal (i.e., too many large negative values)
4. The matched pair differences in blood cadmium levels have thick symmetric tails compared to the Normal (i.e., extreme values occur too often for the Normal, but they are equally likely to be positive or negative)
5. There are three large outliers, but otherwise the differences look Normal.

For question 2.

Hypotheses:

1. H0: Zi are iid, continuous and symmetric about zero versus HA: Zi are iid, continuous but not symmetric about 0.
2. H0: Zi are iid, continuous and symmetric about zero versus HA: Zi are iid, continuous symmetric about .
3. H0: Zi are independent, continuous and with median zero versus HA: Zi are iid, continuous with common median .

For **question 3**, use the model (Zi, Vi) are iid bivariate observations from a continuous distribution.

**Question 4** asks you to construct 3 groups based on the number of cigarettes smoked per day, SMD650smk. The groups are less than 10 (less than half a pack), half a pack to less than a pack (10 to less than 20), and a pack or more. You do this in R with the command cut. Notice that you need to use right=F to exclude 10 from the first interval and exclude 20 from the second. **An easy way to mess up on question 4** is to make the wrong groups. Check that you have the right groups by making sure you have the correct numbers in each group. Remember there is one NA, so 74+82+93 = 249.

> **pack<-cut(SMD650smk,c(0,10,20,99),right=F)**

> **table(pack)**

pack

[0,10) [10,20) [20,99)

74 82 93

as.numeric(pack) makes pack into 1, 2, 3.

> **table(as.numeric(pack),pack)**

pack

[0,10) [10,20) [20,99)

1 74 0 0

2 0 82 0

3 0 0 93

The model for **question 4** has the Zij =  + j = eij where there are groups j = 1, 2, 3, and i goes from 1 to 74 in group 1, from 1 to 82 in group 2, and from 1 to 93 in group 3, where the eij are iid from a continuous distribution. Here, j=1 for <half a pack, j=2 for at least half a pack but less than a pack, and j=3 for a pack or more per day. You are asked to test H0: 1 = 2 = 3 against a general alternative, and H12: 1 = 2 , H13: 1 = 3 , and H23: 2 = 3.Print Name **Last name**, then First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2011, Midterm, Answer Page #1

**This is an exam. Do not discuss it with anyone**.

1. Question 1 refers to the matched pair differences in blood levels of cadmium, LBXBCDdif.

|  |  |
| --- | --- |
| All parts of question 1 use LBXBCDdif | Fill in or CIRCLE the correct answer |
| 1.1 Test the null hypothesis that the differences in cadmium levels are Normally distributed using the Shapiro-Wilk test. Give the P-value and state whether the null hypothesis is plausible. | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Null hypothesis is:  PAUSIBLE NOT PLAUSIBLE |
| 1.2 Do a normal quantile plot of the differences in cadmium levels. Circle the ONE best interpretation of that plot; see the data page for the descriptions. (Do not turn in the plot.) | A B C D E |
| 1.3 Use the appropriate version of Wilcoxon’s test to test the null hypothesis that the differences in cadmium levels are symmetric about zero. Give the full name of the test, the two-sided P-value, and state whether the null hypothesis is plausible. | Full name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  PAUSIBLE NOT PLAUSIBLE |
| 1.4 Give the 95% confidence interval for the center of symmetry of the matched pair differences associated with the test in 1.3. Give the 95% confidence interval from the paired t-test. Is it true or false that the t-test interval is about 20% longer than the nonparametric interval? | Nonparametric interval: [ , ]  t-test interval: [ , ]  TRUE FALSE |
| 1.5 What is the Hodges-Lehmann point estimate of the center of symmetry of the differences in cadmium levels. | Estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

2. Question 2 refers to the analyses you did in question 1 and to the three hypotheses, I, II, and III listed on the data page, where the Zi are the matched pair differences in cadmium levels, i=1,2,…,250.

|  |  |
| --- | --- |
|  | Fill in or CIRCLE the correct answer |
| 2.1 Wilcoxon’s test can be used to test hypothesis I, but it will have meaningful power only if Prob(Zi+Zj >0) is not close to ½. Here, “can be used” means that it falsely rejects the null hypothesis at the stated level, conventionally 0.05. | TRUE FALSE |
| 2.2 The Hodges-Lehmann point estimate and confidence interval are not valid under the alternative hypothesis HA of I but are value under HA of II. Here, valid means that the point estimate is consistent for the center of symmetry of the Zi and the 95% confidence interval covers the center of symmetry in 95% of studies. | TRUE FALSE |
| 2.3 You cannot correctly use Wilcoxon’s test to test hypothesis III, but you can use the sign test. Here, correctly means that the test falsely rejects a true null hypothesis at the nominal rate, conventionally 0.05. | TRUE FALSE |
| 2.4 For the matched pair differences in cadmium levels in question 1, under the alternative hypothesis HA of I, give a consistent estimate of Prob(Zi+Zj >0). Give the numerical value of the estimate. | Estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

Print Name Clearly, **Last**, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2011, Midterm, Answer Page #2

This is an exam. Do not discuss it with anyone.

3. Question 3 asks you to relate the pair differences in cadmium LBXBCDdif to the pair differences in lead LBXBPBdif.

|  |  |
| --- | --- |
|  | Fill in or CIRCLE the correct answer |
| 3.1 Use Pearson’s Normal theory correlation to test the null hypothesis that cadmium differences are unrelated to lead differences. Give the two-sided P-value and the point estimate of the correlation. By this standard, is it plausible that LBXBCDdif and LBXBPBdif are unrelated? | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Correlation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  PLAUSIBLE NOT PLAUSIBLE |
| 3.2 Use Kendall’s nonparametric correlation to test the null hypothesis that cadmium differences are unrelated to lead differences. Give the two-sided P-value and the point estimate of the correlation. By this standard, is it plausible that LBXBCDdif and LBXBPBdif are unrelated? | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Correlation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  PLAUSIBLE NOT PLAUSIBLE |
| 3.3 Use the results in 3.2 to estimate the probability that, in two pairs, the higher cadmium difference will occur in the same pair as the higher lead difference (ie the probability of concordance. | Estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

4. Question 4 asks you to relate the pair differences in cadmium LBXBCDdif to the number of packs per day smoked by the smoker in each pair, in three groups, less than half a pack, at least half a pack but less than a pack, and a pack or more. See the data page for construction of the variable pack. Do this step carefully, or everything will be wrong – make sure the groups have 74, 82, and 93 pairs. Use the notation on the data page for question 4 (eg. H0 or H12) to refer to null hypotheses – do not invent a new notation.

|  |  |
| --- | --- |
|  | Fill in or CIRCLE the correct answer |
| 4.1 Use an appropriate nonparametric test to test H0, the hypothesis of no difference against the alternative of any pattern of differences among the ’s. Give the name of the test, the P-value and state whether the null hypothesis is plausible. | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Name of test: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  PLAUSIBLE NOT PLAUSIBLE |
| 4.2 Use Kendall’s correlation to correlate SMD650smk and LBXBCDdif . Use it again to correlate as.numeric(pack) and LBXBCDdif. Give both correlations and two-sided P-values. | SMD650smk:  P-value\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Correlation:\_\_\_\_\_\_\_\_\_  as.numeric(pack):  P-value\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Correlation:\_\_\_\_\_\_\_\_\_ |
| 4.3 As discussed in class, use Holm’s method to correct the pairwise Wilcoxon P-values. Give the P-values. | H12: \_\_\_\_\_\_\_\_\_ H13: \_\_\_\_\_\_\_\_\_ H23:\_\_\_\_\_\_\_\_\_ |
| 4.4 As discussed in class, use Bonferroni’s method to correct the pairwise Wilcoxon P-values. Give the P-values. | H12: \_\_\_\_\_\_\_\_\_ H13: \_\_\_\_\_\_\_\_\_ H23:\_\_\_\_\_\_\_\_\_ |
| 4. Extra-credit: As discussed in class, use Shaffer’s method to correct the pairwise Wilcoxon P-values. Give the P-values. (R won’t do this, so it takes more thinking, although it is not difficult.) | H12: \_\_\_\_\_\_\_\_\_ H13: \_\_\_\_\_\_\_\_\_ H23:\_\_\_\_\_\_\_\_\_ |

Statistics 501, Spring 2011, Midterm, Answer Page #1 ANSWERS

1. Question 1 refers to the matched pair differences in blood levels of cadmium, LBXBCDdif.

|  |  |
| --- | --- |
| All parts of question 1 use LBXBCDdif | Fill in or CIRCLE the correct answer |
| 1.1 Test the null hypothesis that the differences in cadmium levels are Normally distributed using the Shapiro-Wilk test. Give the P-value and state whether the null hypothesis is plausible. (6 points) | P-value: 1.3 x 10-15  Null hypothesis is:  PAUSIBLE NOT PLAUSIBLE |
| 1.2 Do a normal quantile plot of the differences in cadmium levels. Circle the ONE best interpretation of that plot; see the data page for the descriptions. (Do not turn in the plot.) (6 points) | A B C D E |
| 1.3 Use the appropriate version of Wilcoxon’s test to test the null hypothesis that the differences in cadmium levels are symmetric about zero. Give the full name of the test, the two-sided P-value, and state whether the null hypothesis is plausible.  (6 points) | Full name: Wilcoxon’s signed rank test  P-value: 2.2 x 10-16  PAUSIBLE NOT PLAUSIBLE |
| 1.4 Give the 95% confidence interval for the center of symmetry of the matched pair differences associated with the test in 1.3. Give the 95% confidence interval from the paired t-test. Is it true or false that the t-test interval is about 20% longer than the nonparametric interval?  (10 points) | Nonparametric interval: [ 0.79, 0.99 ]  t-test interval: [ 0.92, 1.16 ]  TRUE FALSE |
| 1.5 What is the Hodges-Lehmann point estimate of the center of symmetry of the differences in cadmium levels. (6 points) | Estimate: 0.885 |

2. Question 2 refers to the analyses you did in question 1 and to the three hypotheses, I, II, and III listed on the data page, where the Zi are the matched pair differences in cadmium levels, i=1,2,…,250.

|  |  |
| --- | --- |
| 6 points each | Fill in or CIRCLE the correct answer |
| 2.1 Wilcoxon’s test can be used to test hypothesis I, but it will have meaningful power only if Prob(Zi+Zj >0) is not close to ½. Here, “can be used” means that it falsely rejects the null hypothesis at the stated level, conventionally 0.05. | See comment 14, age 49 in H&W.  TRUE FALSE |
| 2.2 The Hodges-Lehmann point estimate and confidence interval are not valid under the alternative hypothesis HA of I but are value under HA of II. Here, valid means that the point estimate is consistent for the center of symmetry of the Zi and the 95% confidence interval covers the center of symmetry in 95% of studies. (6 points) | You can’t estimate the center of symmetry unless the distribution is symmetric.  TRUE FALSE |
| 2.3 You cannot correctly use Wilcoxon’s test to test hypothesis III, but you can use the sign test. Here, correctly means that the test falsely rejects a true null hypothesis at the nominal rate, conventionally 0.05. (6 points) | Contrast the assumptions of the sign test (p60) with those of the signed rank test.  TRUE FALSE |
| 2.4 For the matched pair differences in cadmium levels in question 1, under the alternative hypothesis HA of I, give a consistent estimate of Prob(Zi+Zj >0). Give the numerical value of the estimate. (6 points) | Estimate: 0.983  98.3% of the time, when you look at two pairs, the more affected pair (larger | Zi |) is positive, so positive results offset other results 98.3% of the time. |

Statistics 501, Spring 2011, Midterm, Answer Page #2 ANSWERS

3. Question 3 asks you to relate the pair differences in cadmium LBXBCDdif to the pair differences in lead LBXBPBdif.

|  |  |
| --- | --- |
|  | Fill in or CIRCLE the correct answer |
| 3.1 Use Pearson’s Normal theory correlation to test the null hypothesis that cadmium differences are unrelated to lead differences. Give the two-sided P-value and the point estimate of the correlation. By this standard, is it plausible that LBXBCDdif and LBXBPBdif are unrelated? (6 points) | P-value: 0.28  Correlation: 0.069  PLAUSIBLE NOT PLAUSIBLE |
| 3.2 Use Kendall’s nonparametric correlation to test the null hypothesis that cadmium differences are unrelated to lead differences. Give the two-sided P-value and the point estimate of the correlation. By this standard, is it plausible that LBXBCDdif and LBXBPBdif are unrelated? (6 points) | P-value: 0.00037  Correlation: 0.151  PLAUSIBLE NOT PLAUSIBLE |
| 3.3 Use the results in 3.2 to estimate the probability that, in two pairs, the higher cadmium difference will occur in the same pair as the higher lead difference (ie the probability of concordance.) (6pts) | Estimate: 0.576  versus 0.500 for chance agreement |

4. Question 4 asks you to relate the pair differences in cadmium LBXBCDdif to the number of packs per day smoked by the smoker in each pair, in three groups, less than half a pack, at least half a pack but less than a pack, and a pack or more. See the data page for construction of the variable pack. Do this step carefully, or everything will be wrong – make sure the groups have 74, 82, and 93 pairs. Use the notation on the data page for question 4 (eg. H0 or H12) to refer to null hypotheses – do not invent a new notation.

|  |  |
| --- | --- |
|  | Fill in or CIRCLE the correct answer |
| 4.1 Use an appropriate nonparametric test to test H0, the hypothesis of no difference against the alternative of any pattern of differences among the ’s. Give the name of the test, the P-value and state whether the null hypothesis is plausible. (6 points) | P-value: 0.008776  Name of test: Kruskal-Wallis test  PLAUSIBLE NOT PLAUSIBLE |
| 4.2 Use Kendall’s correlation to correlate SMD650smk and LBXBCDdif . Use it again to correlate as.numeric(pack) and LBXBCDdif. Give both correlations and two-sided P-values.  (6 points) | SMD650smk:  P-value 0.0003145 Correlation: 0.161  as.numeric(pack):  P-value 0.00374 Correlation: 0.142 |
| 4.3 As discussed in class, use Holm’s method to correct the pairwise Wilcoxon P-values. Give the P-values. (6 points) | H12: 0.0507 H13: 0.0086 H23: 0.4841 |
| 4.4 As discussed in class, use Bonferroni’s method to correct the pairwise Wilcoxon P-values. Give the P-values. (6 points) | H12: 0.0761 H13: 0.0086 H23: 1.000 |
| 4. Extra-credit: As discussed in class, use Shaffer’s method to correct the pairwise Wilcoxon P-values. Give the P-values. (R won’t do this, so it takes more thinking, although it is not difficult.) (3 points) | H12: 0.0254 H13: 0.0086 H23: 0.4841 |

**STATISTICS 501 SPRING 2011 MIDTERMDOING THE PROBLEM SET IN R**

1.1 and 1.2> par(mfrow=c(1,2))> boxplot(LBXBCDdif)> qqnorm(LBXBCDdif)> shapiro.test(LBXBCDdif) Shapiro-Wilk normality testdata: LBXBCDdif W = 0.8351, p-value = 1.324e-15 1.3-1.5> wilcox.test(LBXBCDdif,conf.int=T) Wilcoxon signed rank test with continuity correctiondata: LBXBCDdif V = 30836.5, p-value < 2.2e-16alternative hypothesis: true mu is not equal to 0 95 percent confidence interval: 0.7900248 0.9899250 sample estimates:(pseudo)median 0.885041 > t.test(LBXBCDdif) One Sample t-testdata: LBXBCDdif t = 17.0183, df = 249, p-value < 2.2e-16alternative hypothesis: true mean is not equal to 0 95 percent confidence interval: 0.9162798 1.1561202 sample estimates:mean of x 1.0362> 1.1561202-0.9162798[1] 0.2398404> 0.9899250-0.7900248[1] 0.1999002> 0.2398404/0.1999002[1] 1.1998012.4Refer to the output above from wilcox.test.> 30836.5/(250\*(250+1)/2)[1] 0.98283673.1> cor.test(LBXBCDdif,LBXBPBdif) Pearson's product-moment correlationdata: LBXBCDdif and LBXBPBdif t = 1.0816, df = 248, p-value = 0.2805alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: -0.05602143 0.19096538 sample estimates: cor 0.068521833.2> cor.test(LBXBCDdif,LBXBPBdif,method="kendall") Kendall's rank correlation taudata: LBXBCDdif and LBXBPBdif z = 3.5588, p-value = 0.0003725alternative hypothesis: true tau is not equal to 0 sample estimates: tau 0.15125103.2> (0.1512510+1)/2[1] 0.57562554.1> kruskal.test(LBXBCDdif,pack) Kruskal-Wallis rank sum testdata: LBXBCDdif and pack Kruskal-Wallis chi-squared = 9.4715, df = 2, p-value = 0.0087764.2 > cor.test(SMD650smk,LBXBCDdif,method="kendall") Kendall's rank correlation taudata: SMD650smk and LBXBCDdif z = 3.6031, p-value = 0.0003145alternative hypothesis: true tau is not equal to 0 sample estimates: tau 0.1614722> cor.test(LBXBCDdif,as.numeric(pack),method="kendall") Kendall's rank correlation taudata: LBXBCDdif and as.numeric(pack) z = 2.8992, p-value = 0.003742alternative hypothesis: true tau is not equal to 0 sample estimates: tau 0.14219394.3> pairwise.wilcox.test(LBXBCDdif,pack) Pairwise comparisons using Wilcoxon rank sum test data: LBXBCDdif and pack [0,10) [10,20)[10,20) 0.0507 - [20,99) 0.0086 0.4841 P value adjustment method: holm4.4 pairwise.wilcox.test(LBXBCDdif,pack,p.adjust.method="bonf") Pairwise comparisons using Wilcoxon rank sum test data: LBXBCDdif and pack [0,10) [10,20)[10,20) 0.0761 - [20,99) 0.0086 1.00004. Extra creditpairwise.wilcox.test(LBXBCDdif,pack,p.adjust.method="none") Pairwise comparisons using Wilcoxon rank sum test data: LBXBCDdif and pack [0,10) [10,20)[10,20) 0.0254 - [20,99) 0.0029 0.4841In Shaffer’s method with 3 groups, 0.0029 is adjusted to 3x0.0029 = 0.0086, but if this is less than 0.05, then the other two p-values are not adjusted.

**Statistics 501 Spring 2012 Final Exam: Data Page 1**

**This is an exam. Do not discuss it with anyone.**

**Due: Wednesday, May 2, 2012 at 11:00am**

Table MaritalMarijuana is from a paper by Kazuo Yamaguchi and Denise Kandel, “Marital homophily on illicit drug use among young adults,” *Social Forces*, 1993, 72, 505-528. It is in JSTOR if you want to look at it, but there is no need to do that for this exam. The data were extracted from a repeated survey. The table describes illicit drug use before marriage (time 1) and after marriage (time 2) for couples consisting of a husband and a wife. A + indicates use of illicit drugs, including marijuana, psychedelics, cocaine, heroin and nonprescribed pills. A – indicates no use of illicit drugs. For example, in 44 instances, a husband and wife who had both used drugs before marriage (Wife1=+, Husband1=+) were both not using drugs in the final survey when married (Wife2=-, Husband 2=-). MaritalMarijuana is in the R workspace for the course. For other programs, you will have to enter the 16 numbers.

**IMPORTANT**: Refer to the four variables in this table by their letter/number pairs, W1 = Wife1, H1 = Husband1, W2 = Wife2, H2 = Husband 2. Fit only hierarchical log-linear models and refer to them by the highest order u-terms they contain, so [W1,H1] [W2,H2] has a interaction u-term linking W1 and H1, W2 and H2, and main effect u-terms for W1, H1, W2, H2, separately and a constant term.

> MaritalMarijuana

, , Wife1 = +, Husband1 = +

Husband2

Wife2 + -

+ 92 16

- 46 44

, , Wife1 = -, Husband1 = +

Husband2

Wife2 + -

+ 5 2

- 41 66

, , Wife1 = +, Husband1 = -

Husband2

Wife2 + -

+ 1 13

- 0 42

, , Wife1 = -, Husband1 = -

Husband2

Wife2 + -

+ 0 4

- 1 156

**This is an exam. Do not discuss it with anyone.**

**Statistics 501 Spring 2012 Final Exam: Data Page 2**

**This is an exam. Do not discuss it with anyone.**

**Save yourself some arithmetic** by learning to use [ ] in R. See what happens when you type MaritalMarijuana[,,1,1]**.** Also, type help(margin.table)

Some questions are “true or false”. Such a question says: “blah and blah and blah”, where the options as answers are “true” and “false”. Circle “true” if “blah and blah and blah” makes sense and is true, but circle false if “blah and blah and not blah” is true. Circle false if “blah and blah and blah” makes no sense.

Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question.

Please write your name on both sides of the exam, last name first. Please include your Penn ID number.

**This problem set is an exam**. If you discuss or communicate with anyone about this exam, then you have cheated on an exam. Cheating on an exam is the single dumbest thing a doctoral student at Penn can do.

**Make and keep a photocopy of your answer page**.Place the exam in an envelope with ‘Paul Rosenbaum, Statistics Department’ on it**. The exam is due in my office, 473 Huntsman, on Wednesday, May 2, 2012 at 11:00am.** You may turn in the exam early at my mail box in the Statistics Department, 4th floor, Huntsman or by giving it to Adam at the front desk in statistics, but if you turn in the exam early, place it in an envelope addressed to me. When all of the exams are graded, I will add an **answer key** to the on-line bulk-pack for the course. You can compare the answer key to your photocopy of your exam. Your course grade will be available from the Registrar. I no longer distribute answer keys and graded exams by US Mail, but you may stop in the pick up your graded exam if you wish.

**Have a great summer!**

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2012 Final Exam: Answer Page 1 This is an exam. Do not discuss it.**

|  |  |
| --- | --- |
| **Due Wednesday, May 2 at 11:00am** | Fill in or CIRCLE the correct answer |
| 1.1 The table MaritalMarijuana describes how many people? How many married couples? | People: \_\_\_\_\_\_\_\_\_ Couples: \_\_\_\_\_\_\_\_\_\_ |
| 1.2 What percent of wives used illicit drugs before marriage? Of husbands before marriage? Of wives after marriage? Of husbands after marriage? | Before After  Wife \_\_\_\_\_\_\_% \_\_\_\_\_\_\_\_%  Husband \_\_\_\_\_\_\_% \_\_\_\_\_\_\_\_% |
| 1.3 Ignoring the data after marriage, what is the odds ratio linking illicit drug use before marriage by people who would later become husband and wife? (This is the W1-H1 odds ratio collapsing over W2-H2.)  Give the odds ratio and the two-sided 95% confidence interval (CI). Repeat this for the husband/wife odds ratio after marriage (W2-H2) ignoring data before marriage (i.e., collapsing over W1-H1). | Collapsed Husband/Wife Odds Ratio (OR)  **Before Marriage**  OR: \_\_\_\_\_\_\_\_\_ CI: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **After Marriage**  OR: \_\_\_\_\_\_\_\_\_ CI: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 1.4 A multinomial model for table MaritalMarijuana assumes independence of drug use by all the individual people in the table, including independence of husbands and wives. | True False |
| 1.5 Only 2 husbands switched from not using illicit drugs prior to marriage to using illicit drugs after marriage. | True False |

|  |  |
| --- | --- |
| 2 Fit the hierarchical log-linear model with all 2-variable interactions and no three-factor or four factor interactions. Use this model for the questions in 2. | Fill in or CIRCLE the correct answer |
| 2.1 What is the likelihood ratio goodness of fit (LRgof) statistic for this model? What are the degrees of freedom (DF)? What is the p-value? Based just on the LRgof: Is this an acceptable fit? | LRgof=\_\_\_\_\_\_\_\_\_\_ DF=\_\_\_\_\_\_\_\_  P-value: \_\_\_\_\_\_\_\_\_\_\_  Acceptable Not Acceptable |
| 2.2 Use the fitted counts from this model to estimate the odds ratio linking after marriage drug use by husbands and wives  (Husband2 and Wife2) at each level of use before marriage. Write in 4 odds ratios. | W1=+,H1=+:\_\_\_\_\_\_W1=-,H1=+:\_\_\_\_\_\_  W1=+,H1=-:\_\_\_\_\_\_W1=-,H1=-:\_\_\_\_\_\_ |

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2012 Final Exam: Answer Page 2 This is an exam. Do not discuss it.**

|  |  |
| --- | --- |
| 3.1 The model [W1,H1] [W2,H2] says drug use by husbands and wives may be related at a fixed time, but the couples’ patterns of use before marriage are independent of the patterns of use after marriage. | True False |
| 3.2 Test the fit of the model [W1,H1] [W2,H2] using the likelihood ratio goodness of fit statistic (LRgof). What is its value? What are the degrees of freedom (DF)? What is the p-value? Based just on the LRgof: Is this an acceptable fit? | LRgof=\_\_\_\_\_\_\_\_\_\_ DF=\_\_\_\_\_\_\_\_  P-value: \_\_\_\_\_\_\_\_\_\_\_  Acceptable Not Acceptable |
| 3.3 The model [W1,W2] [H1,H2] says that, for a person, drug use before marriage may be related to drug use after marriage, but drug use by husbands is independent of drug use by wives. (True/False) Is the fit of this model acceptable? | True False  Acceptable Not Acceptable |

|  |  |
| --- | --- |
| 4 The model (called **mq4**) in question 4 is  [W1,W2],[W1,H1],[H1,H2],[W2,H2] | Fill in or CIRCLE the correct answer |
| 4.1 Test the hypothesis that the model mq4 is acceptable against the alternative that the model in question 2 (which adds [W1,H2] and [W2,H1]) is required. What is the likelihood ratio chi square? What are the degrees of freedom? What is the p-value? Is mq4 acceptable by this test? | LR chi square=\_\_\_\_\_\_\_\_\_\_ DF=\_\_\_\_\_\_\_\_  P-value: \_\_\_\_\_\_\_\_\_\_\_  Acceptable Not Acceptable |
| 4.2 There are four hierarchical models that delete one two-factor u-term from mq4, but none of these four models is an acceptable fit. (True or false) | True False |
| 4.3 Model mq4 says that a wife’s drug use before marriage (W1) is independent of her husband’s drug use after marriage (H2). | True False |
| 4.4 Model mq4 says that a wife’s drug use before marriage (W1) is conditionally independent of her husband’s drug use after marriage (H2) given (W2,H1). | True False |
| 4.5 Use the **fitted counts** from model mq4 to estimate the odds ratio linking W1 and H2 for a couple with (W2=+,H1=+). | Odds ratio = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

**Stat 501 S-2012 Final Exam: Answer Page 1. Answers**

|  |  |
| --- | --- |
| All questions 7 points, except 1.4, 2 points. | Fill in or CIRCLE the correct answer |
| 1.1 The table MaritalMarijuana describes how many people? How many married couples? | People: 1058 Couples: 529 |
| 1.2 What percent of wives used illicit drugs before marriage? Of husbands before marriage? Of wives after marriage? Of husbands after marriage? | Before After  Wife 48.0% 25.1%  Husband 59.0% 35.2% |
| 1.3 Ignoring the data after marriage, what is the odds ratio linking illicit drug use before marriage by people who would later become husband and wife? (This is the W1-H1 odds ratio collapsing over W2-H2.)  Give the odds ratio and the two-sided 95% confidence interval (CI). Repeat this for the husband/wife odds ratio after marriage (W2-H2) ignoring data before marriage (i.e., collapsing over W1-H1). | Collapsed Husband/Wife Odds Ratio (OR)  **Before Marriage**  OR: 4.98 CI: [3.35, 7.46]  **After Marriage**  OR: 9.75 CI: [6.10, 15.87] |
| 1.4 A multinomial model for table MaritalMarijuana assumes independence of drug use by all the individual people in the table, including independence of husbands and wives. | True False |
| 1.5 Only 2 husbands switched from not using illicit drugs prior to marriage to using illicit drugs after marriage. | True False |

|  |  |
| --- | --- |
| 2 Fit the hierarchical log-linear model with all 2-variable interactions and no three-factor or four factor interactions. Use this model for the questions in 2. | Fill in or CIRCLE the correct answer |
| 2.1 What is the likelihood ratio goodness of fit (LRgof) statistic for this model? What are the degrees of freedom (DF)? What is the p-value? Based just on the LRgof: Is this an acceptable fit? | LRgof= 1.346 DF= 5  P-value: 0.93  Acceptable Not Acceptable |
| 2.2 Use the fitted counts from this model to estimate the odds ratio linking after marriage drug use by husbands and wives  (Husband2 and Wife2) at each level of use before marriage. Write in 4 odds ratios. | W1=+,H1=+: 5.39 W1=-,H1=+: 5.39  W1=+,H1=-: 5.39 W1=-,H1=-: 5.39 |

**Stat 501 S-2012 Final Exam: Answer Page 2 .** **Answers**

|  |  |
| --- | --- |
| 3.1 The model [W1,H1] [W2,H2] says drug use by husbands and wives may be related at a fixed time, but the couples’ patterns of use before marriage are independent of the patterns of use after marriage. | True False |
| 3.2 Test the fit of the model [W1,H1] [W2,H2] using the likelihood ratio goodness of fit statistic (LRgof). What is its value? What are the degrees of freedom (DF)? What is the p-value? Based just on the LRgof: Is this an acceptable fit? | LRgof= 358.05 DF= 9  P-value: <0.0001  Acceptable Not Acceptable |
| 3.3 The model [W1,W2] [H1,H2] says that, for a person, drug use before marriage may be related to drug use after marriage, but drug use by husbands is independent of drug use by wives. (True/False) Is the fit of this model acceptable? | True False  Acceptable Not Acceptable |

|  |  |
| --- | --- |
| 4 The model (called **mq4**) in question 4 is  [W1,W2],[W1,H1],[H1,H2],[W2,H2] | Fill in or CIRCLE the correct answer |
| 4.1 Test the hypothesis that the model mq4 is acceptable against the alternative that the model in question 2 (which adds [W1,H2] and [W2,H1]) is required. What is the likelihood ratio chi square? What are the degrees of freedom? What is the p-value? Is mq4 acceptable by this test? | LR chi square= 4.06 DF= 2  P-value: 0.13  Acceptable Not Acceptable |
| 4.2 There are four hierarchical models that delete one two-factor u-term from mq4, but none of these four models is an acceptable fit. (True or false) | True False |
| 4.3 Model mq4 says that a wife’s drug use before marriage (W1) is independent of her husband’s drug use after marriage (H2). | True False |
| 4.4 Model mq4 says that a wife’s drug use before marriage (W1) is conditionally independent of her husband’s drug use after marriage (H2) given (W2,H1). | True False |
| 4.5 Use the **fitted counts** from model mq4 to estimate the odds ratio linking W1 and H2 for a couple with (W2=+,H1=+). | Odds ratio = 1.000 |

Doing the Problem Set in R: Spring 2012 Final (Page 1)

1.1

> sum(m)

[1] 529

> 2\*sum(m)

[1] 1058

1.2

> margin.table(m,1)/529

Wife2

+ -

0.2514178 0.7485822

> margin.table(m,2)/529

Husband2

+ -

0.3516068 0.6483932

> margin.table(m,3)/529

Wife1

+ -

0.4801512 0.5198488

> margin.table(m,4)/529

Husband1

+ -

0.5897921 0.4102079

1.3

> fisher.test(margin.table(m,c(3,4)))

Fisher's Exact Test for Count Data

p-value < 2.2e-16

95 percent confidence interval:

3.353750 7.459505

sample estimates of odds ratio: 4.977578

> fisher.test(margin.table(m,c(1,2)))

Fisher's Exact Test for Count Data

p-value < 2.2e-16

95 percent confidence interval:

6.09626 15.87163

sample estimates of odds ratio: 9.745102

2.1

> loglin(m,list(c(1,2),c(1,3),c(1,4),c(2,3),c(2,4),c(3,4)),fit=T)

$lrt

[1] 1.345795

$df

[1] 5

> 1-pchisq(1.345795,5)

[1] 0.9301496

Doing the Problem Set in R: Spring 2012 Final (Page 1)

2.2

> ft<-loglin(m,list(c(1,2),c(1,3),c(1,4),c(2,3),c(2,4),c(3,4)),fit=T)$fit

11 iterations: deviation 0.07459749

> or(ft[,,1,1])

[1] 5.389931

> or(ft[,,1,2])

[1] 5.389931

> or(ft[,,2,1])

[1] 5.389931

> or(ft[,,2,2])

[1] 5.389931

3.2

> loglin(m,list(c(1,2),c(3,4)))

$lrt

[1] 358.0547

$df

[1] 9

3.3

> loglin(m,list(c(1,3),c(2,4)))

$lrt

[1] 152.9353

$df

[1] 9

4.1 Compare two nested models using change in LR chi square.

> loglin(m,list(c(1,2),c(1,3),c(2,4),c(3,4)))

$lrt

[1] 5.405324

$df

[1] 7

> 5.405324-1.345795

[1] 4.059529

> 1-pchisq(5.405324-1.345795,2)

[1] 0.1313665

4.2 Four tests, of which the first is:

> loglin(m,list(c(1,2),c(1,3),c(2,4)))

$lrt

[1] 39.72262

$df

[1] 8

> 1-pchisq(39,8)

[1] 4.915382e-06

> ft<-loglin(m,list(c(1,2),c(1,3),c(2,4),c(3,4)),fit=T)$fit

> or(ft[1,,,1])

[1] 1

**Statistics 501 Spring 2011 Final Exam: Data Page 1**

**This is an exam. Do not discuss it with anyone.**

The data in the contingency table smoke are from the US National Health and Nutrition Examination Survey (NHANES) for 2007-2008. You can obtain the complete survey from ICPSR via the Penn library web page, but there is no reason to do this for the current exam. A daily smoker reported smoking on every day of the past 30 days (SMD641=30) and having smoked at least 100 cigarettes in his or her lifetime (SMQ020=YES). A nonsmoker reports having smoked fewer than 100 cigarettes in his or her lifetime (SMQ020=NO) and has no reported smoking in the previous 30 days (SMD641=missing). The table also classifies individuals by gender (RIAGENDR), whether the individual served in the military (DMQMILIT), and whether family income is at least twice the poverty level (INDFMPIR>=2). During WWII, the military supplied cigarettes to soldiers, which has led to various studies of the relationship between smoking and military service.

The table smoke is now in the Rworkspace for the course available at my web page

<http://www-stat.wharton.upenn.edu/~rosenbap/index.html> . You will need to download it again. You may need to clear your web-browser’s memory so it forgets the old version and downloads the new one – if you can’t find smoke, that’s probably the reason. If you are not using R, you will need to enter the 16 numbers by hand into some other log-linear program.

**IMPORTANT**: Refer to the four variables in this table by their first letters, S=SmokeDaily, M=Military, T=TwicePoverty, G=Gender. Fit only hierarchical log-linear models and refer to them by the highest order u-terms they contain, so [SM] [TG] has a u-term linking S=SmokeDaily and M=Military, and another linking T=TwicePoverty and G=Gender, contains all four main effects and a constant. This is the **COMPACT NOTATION**.

> smoke

, , TwicePoverty = >= 2xPoverty, Gender = Male

Military

SmokeDaily Served Did not

Nonsmoker 149 509

Smokes Daily 48 147

, , TwicePoverty = < 2xPoverty, Gender = Male

Military

SmokeDaily Served Did not

Nonsmoker 54 376

Smokes Daily 66 286

, , TwicePoverty = >= 2xPoverty, Gender = Female

Military

SmokeDaily Served Did not

Nonsmoker 16 820

Smokes Daily 5 123

, , TwicePoverty = < 2xPoverty, Gender = Female

Military

SmokeDaily Served Did not

Nonsmoker 7 780

Smokes Daily 2 292

**Statistics 501 Spring 2011 Final Exam: Data Page 2**

**This is an exam. Do not discuss it with anyone.**

**Save yourself some arithmetic** by learning to use [ ] in R. See what happens when you type smoke[,,1,1]**.** Also, type help(margin.table)

Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question.

**Make and keep a photocopy of your answer page**.Place the exam in an envelope with ‘Paul Rosenbaum, Statistics Department’ on it**. The exam is due in my office, 473 Huntsman, on Friday, May 6 at 10:00am.** You may turn in the exam early at my mail box in the Statistics Department, 4th floor, Huntsman or by giving it to Adam at the front desk in statistics. When all of the exams are graded, I will add an **answer key** to the on-line bulk-pack for the course. You can compare the answer key to your photocopy of your exam. Your course grade will be available from the Registrar. I no longer distribute answer keys and graded exams by US Mail, but you may stop in the pick up your graded exam if you wish.

**Have a great summer!**

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2011 Final Exam: Answer Page 1 This is an exam. Do not discuss it.**

1. From the 2x2x2x2 table smoke, compute the 2x2 marginal table relating smoking (S) to military service (M). Give the counts and the marginal totals (fill in 9 numbers).

|  |  |  |  |
| --- | --- | --- | --- |
| S x M margin table | Served in Military | Did not serve | Total |
| Nonsmoker |  |  |  |
| Smokes Daily |  |  |  |
| Total |  |  |  |

2. Use the marginal table in question 1 to answer question 2. (R-users, please use the fisher.test command in R.) (Fill in or circle the correct answer.)

|  |  |
| --- | --- |
| 2a. Test the hypothesis of independence in the 2x2 table in question 1, smoking x military service. Give the p-value. Is the null hypothesis of independence plausible? | P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Plausible Not Plausible |
| 2b. What is the (point) estimate of the **odds ratio** of the table in question 1? Are people who served in the military **more likely** than others to smoke daily? Base your answer on the table in question 1. | Odds ratio: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  More likely Not more likely |
| 2c. What is the 95% confidence interval for the odds ratio in the table in question 1? | Conf. Interval = [ , ] |

3. Questions 3-6 return to the 2x2x2x2 table smoke and asks you to answer by fitting log-linear models. Always **use the likelihood ratio chi-square**, not the Pearson chi-square. Refer to models by the COMPACT NOTATION described on the data page.

|  |  |
| --- | --- |
| You want to test the **null hypothesis** that smoking (S) is independent of the other three variables, allowing the other three variables to have any relationship at all. | Fill in or circle the correct answer |
| 3a. **Circle the one model** which best expresses the null hypothesis (i.e., the null hypothesis is true if the model is true). | [S][M][T][G] [SM][ST][SG]  [S][MT][MG][TG] [S][MTG] |
| 3b. Test the goodness of fit of the one selected model in 3a. Give the value of the test statistic, the degrees of freedom (DF), the p-value. Is the null hypothesis of problem 3 that S is independent of M, T, and G plausible? | Value: \_\_\_\_\_\_ DF: \_\_\_\_\_ p-value: \_\_\_\_\_\_  Plausible Not Plausible |

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2011 Final Exam: Answer Page 2 This is an exam. Do not discuss it.**

4. In the 2x2x2x2 table smoke, fit the hierarchical model with all two-variable u-terms and no three-variable u-terms. Use this model to answer the following questions.

|  |  |
| --- | --- |
| 4.1 Use the compact notation to express the model. | **Model is**: |
| 4.2 Does this model fit the data reasonably well as judged by the likelihood ratio test of goodness of fit. Give the value of the statistic, its degrees of freedom, p-value and indicate whether this test **alone** suggests the fit is ok. | Value: \_\_\_\_\_\_ DF: \_\_\_\_\_ p-value: \_\_\_\_\_\_  CIRCLE ONE:  Fit looks OK Definitely not OK |
| 4.3 **Use the fitted counts** from the model in question 4 to estimate four odds ratios linking smoking (S) and military service (M) for the four categories of Twice Poverty and Gender. Fill in the four odds ratios. | |  |  |  | | --- | --- | --- | |  | Male | Female | | >=2xPoverty |  |  | | <2xPoverty |  |  | |
| 4.3 Test each of the 2-variable u-terms in the model in 4.1 **one at a time** to see if you can simplify the model. That is, test the null hypothesis that each 2-variable u-term is zero in a model that retains all the other 2-variable u-terms. So you are thinking about models that differ from the model in 4.1 by one u-term. **Do not do a goodness of fit test**. **List only** those u-terms for which the null hypothesis is plausible, so that a model without that u-term is plausible. **If none, write none**. | List **only** u-terms that are plausibly zero. Here [MG] is the name of the u-term linking Military service and gender. Give the chi-square, degrees of freedom, p-value. You will lose points if you list u-terms that are not plausibly zero.  u-term Chi-Square DF p-value |

5.

|  |  |
| --- | --- |
| 5.1 Which log-linear model says that S = smoking daily is conditionally independent of M=military service given both of the other variables, G=Gender, T =TwicePoverty. Use the compact notation. |  |
| 5.2 Test the goodness of fit of the model in 5.1. Give the likelihood ratio goodness of fit test statistic, degrees of freedom, p-value and state whether the model is rejected at the 0.05 level based on this test. | Value: \_\_\_\_\_\_ DF: \_\_\_\_\_ p-value: \_\_\_\_\_\_  CIRCLE ONE:  Reject at 0.05 Do not reject |

**Answers**

**Stat 501 S-2011 Final Exam: Answer Page 1 This is an exam. Do not discuss it.**

1. From the 2x2x2x2 table smoke, compute the 2x2 marginal table relating smoking (S) to military service (M). Give the counts and the marginal totals (fill in 9 numbers).

|  |  |  |  |
| --- | --- | --- | --- |
| S x M margin table | Served in Military | Did not serve | Total |
| Nonsmoker | 226 | 2485 | 2711 |
| Smokes Daily | 121 | 848 | 969 |
| Total | 347 | 3333 | 3680 |

2. Use the marginal table in question 1 to answer question 2. (R-users, please use the fisher.test command in R.) (Fill in or circle the correct answer.)

|  |  |
| --- | --- |
| 2a. Test the hypothesis of independence in the 2x2 table in question 1, smoking x military service. Give the p-value. Is the null hypothesis of independence plausible? | P-value: 0.0002495  Plausible Not Plausible |
| 2b. What is the (point) estimate of the **odds ratio** of the table in question 1? Are people who served in the military **more likely** than others to smoke daily? Base your answer on the table in question 1. | Odds ratio: 0.637  More likely Not more likely |
| 2c. What is the 95% confidence interval for the odds ratio in the table in question 1? | Conf. Interval = [ 0.502 , 0.813 ] |

3. Questions 3-6 return to the 2x2x2x2 table smoke and asks you to answer by fitting log-linear models. Always **use the likelihood ratio chi-square**, not the Pearson chi-square. Refer to models by the COMPACT NOTATION described on the data page.

|  |  |
| --- | --- |
| You want to test the **null hypothesis** that smoking (S) is independent of the other three variables, allowing the other three variables to have any relationship at all. | Fill in or circle the correct answer |
| 3a. **Circle the one model** which best expresses the null hypothesis (i.e., the null hypothesis is true if the model is true). | [S][M][T][G] [SM][ST][SG]  [S][MT][MG][TG] [S][MTG] |
| 3b. Test the goodness of fit of the one selected model in 3a. Give the value of the test statistic, the degrees of freedom (DF), the p-value. Is the null hypothesis of problem 3 that S is independent of M, T, and G plausible? | Value: 237.1 DF: 7 p-value: <0.0001  Plausible Not Plausible |

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2011 Final Exam: Answer Page 2 This is an exam. Do not discuss it.**

4. In the 2x2x2x2 table smoke, fit the hierarchical model with all two-variable u-terms and no three-variable u-terms. Use this model to answer the following questions.

|  |  |
| --- | --- |
| 4.1 Use the compact notation to express the model. | **Model is**:  [S,M][ST][SG][MT][MG][TG] |
| 4.2 Does this model fit the data reasonably well as judged by the likelihood ratio test of goodness of fit. Give the value of the statistic, its degrees of freedom, p-value and indicate whether this test **alone** suggests the fit is ok. | Value: 5.02 DF: 5 p-value: 0.41  CIRCLE ONE:  Fit looks OK Definitely not OK |
| 4.3 **Use the fitted counts** from the model in question 4 to estimate four odds ratios linking smoking (S) and military service (M) for the four categories of Twice Poverty and Gender. Fill in the four odds ratios. | |  |  |  | | --- | --- | --- | |  | Male | Female | | >=2xPoverty | 0.752 | 0.752 | | <2xPoverty | 0.752 | 0.752 | |
| 4.3 Test each of the 2-variable u-terms in the model in 4.1 **one at a time** to see if you can simplify the model. That is, test the null hypothesis that each 2-variable u-term is zero in a model that retains all the other 2-variable u-terms. So you are thinking about models that differ from the model in 4.1 by one u-term. **Do not do a goodness of fit test**. **List only** those u-terms for which the null hypothesis is plausible, so that a model without that u-term is plausible. **If none, write none**. | List **only** u-terms that are plausibly zero. Here [MG] is the name of the u-term linking Military service and gender. Give the chi-square, degrees of freedom, p-value. You will lose points if you list u-terms that are not plausibly zero.  u-term Chi-Square DF p-value  None |

5.

|  |  |
| --- | --- |
| 5.1 Which log-linear model says that S = smoking daily is conditionally independent of M=military service given both of the other variables, G=Gender, T =TwicePoverty. Use the compact notation. | [SGT] [MGT] |
| 5.2 Test the goodness of fit of the model in 5.1. Give the likelihood ratio goodness of fit test statistic, degrees of freedom, p-value and state whether the model is rejected at the 0.05 level based on this test. | Value: 7.87 DF: 4 p-value: 0.097  CIRCLE ONE:  Reject at 0.05 Do not reject |

**Doing the Problem Set in R**

**Final, Spring 2011**

Question 1.

> margin.table(smoke,c(1,2))

Military

SmokeDaily Served Did not

Nonsmoker 226 2485

Smokes Daily 121 848

Question 2.

> fisher.test(margin.table(smoke,c(1,2)))

Fisher's Exact Test for Count Data

data: margin.table(smoke, c(1, 2))

p-value = 0.0002495

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

0.5016968 0.8128271

sample estimates:

odds ratio

0.637456

Question 3.

> loglin(smoke,list(1,c(2,3,4)))

2 iterations: deviation 1.136868e-13

$lrt

[1] 237.1712

$pearson

[1] 240.0149

$df

[1] 7

> 1-pchisq(237.1712,7)

[1] 0

$margin

$margin[[1]]

[1] "SmokeDaily"

$margin[[2]]

[1] "Military" "TwicePoverty" "Gender"

Question 4.1

> loglin(smoke,list(c(1,2),c(1,3),c(1,4),c(2,3),c(2,4),c(3,4)))

5 iterations: deviation 0.06586971

$lrt

[1] 5.021628

$pearson

[1] 5.039157

$df

[1] 5

> 1-pchisq(5.021628,5)

[1] 0.4132464

Question 4.2

ft<-loglin(smoke,list(c(1,2),c(1,3),c(1,4),c(2,3),c(2,4),c(3,4)),fit=T)$fit

> or(ft[,,1,1])

[1] 0.7522604

> or(ft[,,1,2])

[1] 0.7522604

> or(ft[,,2,1])

[1] 0.7522604

> or(ft[,,2,2])

[1] 0.7522604

Question 4.3

None of the terms can be deleted. For example:

> loglin(smoke,list(c(1,3),c(1,4),c(2,3),c(2,4),c(3,4)))

5 iterations: deviation 0.02724409

$lrt

[1] 9.778795

$pearson

[1] 10.62873

$df

[1] 6

> 9.778795-5.021628

[1] 4.757167

> 1-pchisq(4.757167,1)

[1] 0.02917653

Question 5.

> loglin(smoke,list(c(1,3,4),c(2,3,4)))

2 iterations: deviation 1.136868e-13

$lrt

[1] 7.868466

$pearson

[1] 8.222663

$df

[1] 4

$margin

$margin[[1]]

[1] "SmokeDaily" "TwicePoverty" "Gender"

$margin[[2]]

[1] "Military" "TwicePoverty" "Gender"

> 1-pchisq(7.868466,4)

[1] 0.09651703

**Statistics 501, Spring 2010, Midterm: Data Page #1**

This is an exam. Do not discuss it with anyone. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question. Due in class Tuesday 30 March. The data for this problem are at in the latest Rst501.RData for R users as the object katzkrueger and in katzkrueger.txt as a text file at <http://stat.wharton.upenn.edu/statweb/course/Spring-2008/stat501> The list is case sensitive, so katzkrueger.txt is with lower case items.

The data are from a paper by Lawrence Katz and Alan Krueger (1992), The effect of the minimum wage on the fast food industry, *Industrial and Labor Relations Review*, 46, 6-21, specifically the subset of data in Table 5 of that paper. (You do not need to look at the paper to do the problem set. If you wish to look at the paper, it in JSTOR on the UPenn library web page at <http://www.jstor.org/stable/2524735>.) The paper concerns an increase in the Federal minimum wage from $3.80 to $4.25 per hour that took place on April 1, 1991, and its effects on employment in the fast food industry in Texas. The data are based on two surveys of the same 100 fast food restaurants, one in December 1990 before the increase in the minimum wage, the other in July/August 1991 after the increase in the minimum wage. The restaurant chains were Burger King, KFC and Wendy’s (McDonald’s refused). Economic theory typically predicts that an increase in the minimum wage will reduce employment essentially because some workers are worth employing at $3.80 per hour but not at $4.25 per hour (e.g., George J. Stigler (1946) **The Economics of Minimum Wage Legislation,** American Economic Review, 36, 358-365, <http://www.jstor.org/stable/1801842>). Katz and Krueger looked at this in various ways. In particular, they looked at the change in full-time-equivalent (fte) employment, after-minus-before, which is fte91-fte90. (Some of the numbers look a little odd – they are survey responses.) They also looked at the gap in starting wages that the restaurant needed to close to comply with the new minimum age. For instance, the first restaurant (id=1) below was a Burger King (bk=1, kfc=0) that was not company owned (co\_owned=0) paying $3.85 per hour as a starting wage in December 1990, so to reach the new minimum wage it had to close a gap of $4.25-$3.85 = $0.40. Notice that the second restaurant below (id=3) had the same gap, but raised the starting wage to $4.40, that is, an increase of $0.65, rather than the required $0.40. Notice that the eighth restaurant (id=27) was paying $4.60 before the increase, so its gap is zero, as the law did not require it to raise wages. Katz and Krueger argued that if conventional theory were correct, the decline in employment should be larger if gap is larger. The variable grp forms groups using gap, while twogrp makes just two groups, >$0.40 and <=$0.25 with the rest as missing (NA).

> dim(katzkrueger)

[1] 100 11

> katzkrueger[1:10,]

id bk kfc co\_owned paydec payjul gap fte90 fte91 grp twogrp

1 1 1 0 0 3.85 4.25 0.40 10.13 3.57 (0.25,0.4] NA

2 3 1 0 0 3.85 4.40 0.40 25.70 23.55 (0.25,0.4] NA

3 4 0 1 1 4.15 4.25 0.10 11.98 12.70 (0,0.24] 1

4 14 0 0 0 4.00 4.25 0.25 21.66 29.95 (0.24,0.25] 1

5 20 1 0 0 3.80 4.25 0.45 31.40 18.54 (0.4,0.45] 0

6 24 1 0 0 3.80 4.25 0.45 14.25 12.14 (0.4,0.45] 0

7 25 1 0 1 3.80 4.25 0.45 11.40 19.25 (0.4,0.45] 0

8 27 1 0 0 4.60 4.25 0.00 23.55 7.41 (-Inf,0] 1

9 30 1 0 1 4.20 4.25 0.05 26.99 27.00 (0,0.24] 1

10 32 0 0 0 3.80 4.32 0.45 8.55 7.70 (0.4,0.45] 0

…

> table(twogrp,grp)

grp

twogrp (-Inf,0] (0,0.24] (0.24,0.25] (0.25,0.4] (0.4,0.45]

1 10 10 23 0 0

0 0 0 0 0 37

STATISTICS 501, SPRING 2010, MIDTERM DATA PAGE #2

Define three new variables as:

> attach(katzkrueger)

> dife<-fte91-fte90

> difer<-fte90-fte91

> difp<-payjul-paydec

Notice that difer = -dife.

Model 1: *Xi =*  + i, *i=1,2,…,m*, *Yi =*  +  + j+m, *j=1,…,n*, where k ~ iid, *k=1,2,…,n+m*, with a continuous distribution.

Model 2: *(X1,Y1), …, (Xn,Yn)* are n iid observations from a continuous bivariate distribution.

Model 3: *Xi*, *i=1,2,…,m*, are iid from one continuous distribution, and *Yi*, *j=1,…,n*, are iid from another continuous distribution, and the *X*’s and *Y*’s are independent of each other.

Model 4: *Xi*, *i=1,2,…,m*, and *Yi*, *j=1,…,n*, are *n+m iid* observations from the same continuous distribution.

Model 5: *Yi* - *Xi* = i where i ~ iid, with a continuous distribution symmetric about 0, i=1,…,n.

Model 6: *Yi* =  + *Xi* + *ei*, …, where the *ei* are n iid observations from a continuous distribution with median zero independent of the *X*i which are untied.

Model 7: *Yi* - *Xi* = i where i are independent, with possibly different continuous distributions each having median zero.

Model 8: *Xij =*  + j + ij, *i=1,2,…,nj*, *j=1,…,K* where the ij’s are iid from a continuous distribution, with 0 = 1+…+K.

Model 9 *Xij =*  + j + ij, *i=1,2,…,nj*, *j=1,…,K* where the ij’s are iid from a continuous distribution, with 0 = 1+…+K, with j>j+1  or j=j+1 with at least one strict inequality.

Model 10: *Xij =*  + j + ij, *i=1,2,…,nj*, *j=1,…,K* where the ij’s are iid from a continuous distribution, with 0 = 1+…+K, j<j+1  or j=j+1 with at least one strict inequality.

Model 11: *Xij =*  + ij, *i=1,2,…,nj*, *j=1,…,K* where the ij’s are iid from a continuous distribution.

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2010, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. Use abbreviations and model #’s from the data page.

1. Use the 100 observations on dife to answer question 1. dife is defined on the data page.

|  |  |
| --- | --- |
| 1.A. Do a boxplot, a Normal quantile plot and a Shapiro-Wilk test of Normality to determine whether the changes in employment (dife) look like observations from a Normal distribution. Do not turn in the plots. What is the P-value for the Shapiro-Wilk test? Do the changes in employment look like observations from a Normal distribution? | Shapiro-Wilk P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  Looks Normal Does not look Normal |
| 1.B Use Wilcoxon’s signed rank test to test the hypothesis that the changes in employment (dife) are symmetric about zero difference. What is the value of the test statistic? What is the two-sided P-value? Is the null hypothesis rejected at the conventional 0.05 level? | Test statistic: \_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  Rejected at 0.05 Not rejected at 0.05 |
| 1.C Use procedures developed from Wilcoxon’s signed rank test to find a two-sided 95% confidence interval and a point estimate for the center of symmetry of the changes in employment (dife). Taking the point estimate at naively, at face value, roughly (to the nearest integer) how many fte employees were gained or lost following the increase in the minimum wage? | 95% Interval: [ \_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_\_\_ ]  Point estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Lost \_\_\_\_\_\_\_\_\_\_ or gained \_\_\_\_\_\_\_\_\_ employees |
| 1.D Do a two-sided test of the null hypothesis that the center of symmetry of the changes in employment (dife) reflect a typical decline of ½ of an full time equivalent (fte) employee. State briefly how you did the test, give the value of the test statistic, the P-value, and say whether an ½ employee decline is plausible. | How you did it:  Test statistic \_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  Plausible Not plausible |
| 1.E Of the models on the data page, which one model underlies Wilcoxon’s signed rank test as you used it in 1.D? Give one model number. | Model number: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

2. Use the 100 observations on (dife, gap) to answer question 2.

|  |  |
| --- | --- |
| 2.A. Test the null hypothesis that dife and gap are independent using Kendall’s correlation. Give: the correlation estimate, the two-sided P-value, and the estimated probability of concordance. Is it plausible that dife and gap are unrelated? Does this result suggest that declines in employment are typically larger when an increase in the minimum wage requires a larger increase in starting wages to comply with the new minimum wage? | Correlation: \_\_\_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_\_\_  Probability of concordance: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  Plausible Not plausible  CIRCLE ONE  Does suggest Does not suggest |
| 2.B. Under which two models on the data page would the test in 2A be appropriate? Write two model numbers. | Model numbers: \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_ |
| 2C. In model 6, with Yi=dife, Xi=gap, use Kendall’s correlation to test the hypothesis that H0:= -1, so that a restaurant with the maximum gap of gap=0.45 would experience about ½ (exactly 0.45) larger decline than a restaurant with gap=0. | Two-sided p-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  H0:= -1 is plausible Not plausible |

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2010, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. Use abbreviations and model #’s from the data page.

3. Question 3 asks you to compare the changes in employment, dife, for the restaurants with a large (gap>.4) or small (gap<=0.25) gap as defined by twogrp. There are 37 large gaps, and 43 small ones. You are looking at the large-minus-small differences in the change in employment, that is, in a difference-in-differences.

|  |  |
| --- | --- |
| 3.A Use an appropriate two-sided nonparametric test to see if the change in employment, after-minus-before, is higher or lower with twogrp=0 versus twogrp=1. What is the name of the test statistic? What is the two-sided P-value? Is the null hypothesis of no difference plausible? | Name of test: \_\_\_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  Plausible Not plausible |
| 3.B Give a two-sided 95% confidence interval and point estimate of shift associated with the test in 3A. Orient the difference so it is high gap minus low gap, or twogrp=0 minus twogrp=1. Does this calculation suggest that a large gap to meet the new minimum wage is associated with a larger decline in employment? | 95% Confidence Interval: [ \_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_]  Point estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  Does suggest Does not suggest |
| 3C. Which model underlies the procedure in 3B. Give one model number. | Model number:  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 3D. If Y=dife for twogrp=1 and X=dife for twogrp=0, give an estimate of Pr(Y>X) based on the procedure in 3A. Does this calculation suggest that a large gap to meet the new minimum wage is associated with a larger decline in employment? | Point estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  Does suggest Does not suggest |
| 3E. Is the model in 3C needed for the estimate in 3D or could a more general model be used instead? If a more general model would suffice, give its model number. | CIRCLE ONE  Needed More general would suffice  Model #, if applicable:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

4. Question 4 asks you to use either dife or difer with grp to compare levels of changes in employment for groups defined by gap. There are 100 restaurants in 5 groups.

|  |  |
| --- | --- |
| 4A. Test the null hypothesis that the five groups do not differ in level. What is the name of the appropriate nonparametric test? What is the P-value? Which one model is the null hypothesis in this test and which other one model is the alternative hypothesis? (Pick the best choices and give model numbers.) | Name of test: \_\_\_\_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_  Null Model: \_\_\_\_\_\_\_\_ Alternative: \_\_\_\_\_\_\_\_\_\_\_\_ |
| 4B. Use Holm’s procedure with the Wilcoxon test to compare all pairs of groups. List the pairs of groups that differ significantly as (grp1, grp2). | List pairs. If none, write “none”. |
| 4C. Stigler’s analysis would lead you to expect that a larger gap would lead to a greater decline in employment. Test no difference against Stigler’s prediction using the Jonckheere-Terpstra test and the jonck.test function in the course workspace. Give the one-sided p-value. Be careful and think: you must orient the test and calculations so it aims at Stigler’s prediction. Is no difference rejected at the 0.05 level in the direction that Stigler predicted? | One-sided p-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  CIRCLE ONE  Rejected at 0.05 Not rejected at 0.05 |

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2010 Midterm Answer Page 1

This is an exam. Do not discuss it with anyone. Use abbreviations and model #’s from the data page.

1. Use the 100 observations on dife to answer question 1. dife is defined on the data page.

|  |  |
| --- | --- |
| 1.A. Do a boxplot, a Normal quantile plot and a Shapiro-Wilk test of Normality to determine whether the changes in employment (dife) look like observations from a Normal distribution. Do not turn in the plots. What is the P-value for the Shapiro-Wilk test? Do the changes in employment look like observations from a Normal distribution? | Shapiro-Wilk P-value: 0.0488  CIRCLE ONE  Looks Normal Does not look Normal |
| 1.B Use Wilcoxon’s signed rank test to test the hypothesis that the changes in employment (dife) are symmetric about zero difference. What is the value of the test statistic? What is the two-sided P-value? Is the null hypothesis rejected at the conventional 0.05 level? | Test statistic: 2966 P-value: 0.0557  CIRCLE ONE  Rejected at 0.05 Not rejected at 0.05 |
| 1.C Use procedures developed from Wilcoxon’s signed rank test to find a two-sided 95% confidence interval and a point estimate for the center of symmetry of the changes in employment (dife). Taking the point estimate at naively, at face value, roughly (to the nearest integer) how many fte employees were gained or lost following the increase in the minimum wage? | 95% Interval: [ -0.015, 2.345 ]  Point estimate: 1.08  Lost \_\_\_\_\_\_\_\_\_\_ or gained 1 employees |
| 1.D Do a two-sided test of the null hypothesis that the center of symmetry of the changes in employment (dife) reflect a typical decline of ½ of an full time equivalent (fte) employee. State briefly how you did the test, give the value of the test statistic, the P-value, and say whether an ½ employee decline is plausible. | How you did it: Subtract -1/2 (or add ½) and test no difference.  Test statistic 3312.5 P-value: 0.00681  CIRCLE ONE  Plausible Not plausible |
| 1.E Of the models on the data page, which one model underlies Wilcoxon’s signed rank test as you used it in 1.D? Give one model number. | Model number: 5 |

2. Use the 100 observations on (dife, gap) to answer question 2.

|  |  |
| --- | --- |
| 2.A. Test the null hypothesis that dife and gap are independent using Kendall’s correlation. Give: the correlation estimate, the two-sided P-value, and the estimated probability of concordance. Is it plausible that dife and gap are unrelated? Does this result suggest that declines in employment are typically larger when an increase in the minimum wage requires a larger increase in starting wages to comply with the new minimum wage? | Correlation: 0.177 P-value: 0.01664  Probability of concordance: .59  CIRCLE ONE  Plausible Not plausible  CIRCLE ONE  Does suggest Does not suggest |
| 2.B. Under which two models on the data page would the test in 2A be appropriate? Write two model numbers. | Model numbers: 2 and 6 |
| 2C. In model 6, with Yi=dife, Xi=gap, use Kendall’s correlation to test the hypothesis that H0:= -1, so that a restaurant with the maximum gap of gap=0.45 would experience about ½ (exactly 0.45) larger decline than a restaurant with gap=0. | Two-sided p-value: 0.0087  CIRCLE ONE  H0:= -1 is plausible Not plausible |

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2010, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. Use abbreviations and model #’s from the data page.

3. Question 3 asks you to compare the changes in employment, dife, for the restaurants with a large (gap>.4) or small (gap<=0.25) gap as defined by twogrp. There are 37 large gaps, and 43 small ones. You are looking at the large-minus-small differences in the change in employment, that is, in a difference-in-differences.

|  |  |
| --- | --- |
| 3.A Use an appropriate two-sided nonparametric test to see if the change in employment, after-minus-before, is higher or lower with twogrp=0 or twogrp=1. What is the name of the test statistic? What is the two-sided P-value? Is the null hypothesis of no difference plausible? | Name: Wilcoxon rank sum P-value: 0.0268  CIRCLE ONE  Plausible Not plausible |
| 3.B Give a two-sided 95% confidence interval and point estimate of shift associated with the test in 3A. Orient the difference so it is high gap minus low gap, or twogrp=0 minus twogrp=1. Does this calculation suggest that a large gap to meet the new minimum wage is associated with a larger decline in employment? | 95% Confidence Interval: [ 0.41 , 5.41 ]  Point estimate: 2.57  CIRCLE ONE  Does suggest Does not suggest |
| 3C. Which model underlies the procedure in 3B. Give one model number. | Model number: 1  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 3D. If Y=dife for twogrp=0 and X=dife for twogrp=1, give an estimate of Pr(Y>X) based on the procedure in 3A. Does this calculation suggest that a large gap to meet the new minimum wage is associated with a larger decline in employment? | Point estimate: 0.645 (accepted .35 = 1-.645)  CIRCLE ONE  Does suggest Does not suggest |
| 3E. Is the model in 3C needed for the estimate in 3D or could a more general model be used instead? If a more general model would suffice, give its model number. | CIRCLE ONE  Needed More general would suffice  Model #, if applicable:\_\_\_3\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

4. Question 4 asks you to use either dife or difer with grp to compare levels of changes in employment for groups defined by gap. There are 100 restaurants in 5 groups.

|  |  |
| --- | --- |
| 4A. Test the null hypothesis that the five groups do not differ in level. What is the name of the appropriate nonparametric test? What is the P-value? Which one model is the null hypothesis in this test and which other one model is the alternative hypothesis? (Pick the best choices and give model numbers.) | Name of test: Kruskal-Wallis P-value: .135  Null Model: 11 Alternative: 8 |
| 4B. Use Holm’s procedure with the Wilcoxon test to compare all pairs of groups. List the pairs of groups that differ significantly as (grp1, grp2). | List pairs. If none, write “none”.  None |
| 4C. Stigler’s analysis would lead you to expect that a larger gap would lead to a greater decline in employment. Test no difference against Stigler’s prediction using the Jonckheere-Terpstra test and the jonck.test function in the course workspace. Give the one-sided p-value. Be careful and think: you must orient the test and calculations so it aims at Stigler’s prediction. Is no difference rejected at the 0.05 level in the direction that Stigler predicted? | One-sided p-value: 0.987  CIRCLE ONE  Rejected at 0.05 Not rejected at 0.05  The direction is backwards, so you do not reject in this direction. Had you predicted the opposite direction, you would have rejected. |

**Doing the Problem Set in R (Spring 2010)**

1.A

> **par(mfrow=c(1,2))**

> **boxplot(dife,ylab="change in fte")**

> **qqnorm(dife,ylab="change in fte")**

> **qqline(dife)**

> **shapiro.test(dife)**

Shapiro-Wilk normality test

data: dife

W = 0.9745, p-value = 0.04888

1.B, 1.C

> **wilcox.test(dife,conf.int=T)**

Wilcoxon signed rank test with continuity correction

data: dife

V = 2966, p-value = 0.05568

alternative hypothesis: true location is not equal to 0

95 percent confidence interval:

-0.01494096 2.34500504

sample estimates:

(pseudo)median

1.080024

1.D

> **wilcox.test(dife+.5)**

Wilcoxon signed rank test with continuity correction

data: dife + 0.5

V = 3312.5, p-value = 0.00681

alternative hypothesis: true location is not equal to 0

*or equivalently*

> **wilcox.test(dife-(-.5))**

data: dife - (-0.5)

V = 3312.5, p-value = 0.00681

alternative hypothesis: true location is not equal to 0

2.A

> **cor.test(dife,gap,method="kendall")**

Kendall's rank correlation tau

data: dife and gap

z = 2.3945, p-value = 0.01664

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.1771132

> **(0.1771132+1)/2**

[1] 0.5885566

2.C

> **cor.test(dife-(-1\*gap),gap,method="kendall")**

Kendall's rank correlation tau

data: dife - (-1 \* gap) and gap

z = 2.6229, p-value = 0.00872

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.1939636

3A, 3B, 3C.

> wilcox.test(dife~twogrp,conf.int=T)

Wilcoxon rank sum test with continuity correction

data: dife by twogrp

W = 1025.5, p-value = 0.02678

alternative: true location shift is not equal to 0

95 percent confidence interval:

0.4100318 5.4100372

sample estimates:

difference in location

2.57228

> **1025.5/(43\*37)**

[1] 0.6445632

4A.

> **kruskal.test(dife,grp)**

Kruskal-Wallis rank sum test

data: dife and grp

Kruskal-Wallis chi-squared = 7.0116, df = 4, p-value = 0.1353

4B.

> **pairwise.wilcox.test(dife,grp)**

Pairwise comparisons using Wilcoxon rank sum test

data: dife and grp

(-Inf,0] (0,0.24] (0.24,0.25] (0.25,0.4]

(0,0.24] 1.00 - - -

(0.24,0.25] 1.00 1.00 - -

(0.25,0.4] 1.00 1.00 1.00 -

(0.4,0.45] 0.62 0.79 1.00 0.62

P value adjustment method: holm

4C. *Must use difer to get direction right.*

> jonck.test(difer,grp)

$pval

0.986594

**Statistics 501 Spring 2010 Final Exam: Data Page 1**

**This is an exam. Do not discuss it with anyone.**

The data are from a survey conducted in 2007 by the CDC: “The Youth Risk Behavior Surveillance System (YRBSS) monitors priority health-risk behaviors and the prevalence of obesity and asthma among youth and young adults. The YRBSS includes a national school-based survey conducted by the Centers for Disease Control and Prevention (CDC) and state, territorial, tribal, and local surveys conducted by state, territorial, and local education and health agencies and tribal governments.” Strictly speaking, specialized methods should be used for data from complex sample surveys, but for the current exam this issue will be ignored. <http://www.cdc.gov/HealthyYouth/yrbs/data/index.htm>

The data are a 25 contingency table, yrbs2007, described kids 15-18, describing smoking (S), cocaine use (C), alcohol use (A), age in years (Y) and gender (G). **Use S, C, A, Y, and G to refer to the variables.** The table yrbs2007.2 is the same as yrbs2007 except it uses the letters S, C, A, Y and G; use either table.

> yrbs2007

, , alcohol\_Q42 = 0 times, age = 15-16, Q2 = Female

cocaine\_Q50

smoke\_Q34 0 times >0 times

No 2318 5

Yes 111 4

, , alcohol\_Q42 = >0 times, age = 15-16, Q2 = Female

cocaine\_Q50

smoke\_Q34 0 times >0 times

No 520 27

Yes 130 32

, , alcohol\_Q42 = 0 times, age = 17-18, Q2 = Female

cocaine\_Q50

smoke\_Q34 0 times >0 times

No 1715 6

Yes 144 2

, , alcohol\_Q42 = >0 times, age = 17-18, Q2 = Female

cocaine\_Q50

smoke\_Q34 0 times >0 times

No 496 29

Yes 155 44

, , alcohol\_Q42 = 0 times, age = 15-16, Q2 = Male

cocaine\_Q50

smoke\_Q34 0 times >0 times

No 2133 15

Yes 91 2

, , alcohol\_Q42 = >0 times, age = 15-16, Q2 = Male

cocaine\_Q50

smoke\_Q34 0 times >0 times

No 508 49

Yes 142 43

, , alcohol\_Q42 = 0 times, age = 17-18, Q2 = Male

cocaine\_Q50

smoke\_Q34 0 times >0 times

No 1520 12

Yes 118 7

, , alcohol\_Q42 = >0 times, age = 17-18, Q2 = Male

cocaine\_Q50

smoke\_Q34 0 times >0 times

No 641 45

Yes 212 75

**Statistics 501 Spring 2010 Final Exam: Data Page 2**

**This is an exam. Do not discuss it with anyone.**

> **dimnames(yrbs2007)**

$smoke\_Q34

[1] "No" "Yes"

**S** = smoke\_Q34 is “Have you ever smoked cigarettes daily, that is, at least one cigarette every day for 30 days?”

$cocaine\_Q50

[1] "0 times" ">0 times"

**C** = cocaine\_Q50 is: “During the past 30 days, how many times did you use any form of cocaine, including powder, crack or freebase?”

$alcohol\_Q42

[1] "0 times" ">0 times"

**A** = alcohol\_Q42 is: “During the past 30 days, on many days did you have 5 or more drinks of alcohol in a row, that is, within a couple of hours?

$age

[1] "15-16" "17-18"

**Y** for years. (Younger kids are excluded.)

$Q2

[1] "Female" "Male"

**G** for gender.

**Save yourself some arithmetic** by learning to use [ ] in R. See what happens when you type yrbs2007[,,1,1,1] or yrbs2007[,2,,,]**.** Also, type help(round)

**IMPORTANT**

The only log-linear models considered are hierarchical models. Refer to such a model using the **compact notation** that indicates the highest order u-terms that are included. Example: log(mijklm) = u + uS(i) + uC(j) + uA(k) + uY(l) + uG(m) + uSC(ij) + uYG(lm) is

[SC] [A] [YG]. Use the S, C, A, Y, G letters and brackets [ ].

Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question.

**Make and keep a photocopy of your answer page**.Place the exam in an envelope with ‘Paul Rosenbaum, Statistics Department’ on it**. The exam is due in my office, 473 Huntsman, on Tuesday, May 11 at 11:00am.** You may turn in the exam early at my mail box in the Statistics Department, 4th floor, Huntsman. When all of the exams are graded, I will add an answer key to the on-line bulk-pack for the course.

**This is an exam. Do not discuss it with anyone.**

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2010 Final Exam: Answer Page 1 This is an exam. Do not discuss it.**

|  |  |
| --- | --- |
| 1 Answer this question using ONLY the likelihood ratio goodness-of-fit chi-square for the one model in this question. | CIRCLE ONE or FILL IN |
| 1.1. Does the hierarchical log-linear model with all 2-factor interactions (and no 3 factor interactions) provide an adequate fit to the data? | adequate not adequate |
| 1.2. What is the value of the likelihood ratio chi-square for the model in 1.1? What are its degrees of freedom? What is the p-value? | chi square: \_\_\_\_\_\_\_\_\_\_\_ df: \_\_\_\_\_\_\_\_\_  p-value: \_\_\_\_\_\_\_\_\_\_\_\_\_ |

|  |  |
| --- | --- |
| 2 Answer this question using ONLY the likelihood ratio goodness-of-fit chi-square for the one model in this question. | CIRCLE ONE or FILL IN |
| 2.1 Which hierarchical log-linear model says smoking (S) is conditionally independent of gender (G) given the other three variables (C & A & Y)? The question asks for the largest or most complex model which has this condition. |  |
| 2.2 Does the hierarchical log-linear model in 2.1 provide an adequate fit to the data? | adequate not adequate |
| 2.3. What is the value of the likelihood ratio chi-square for the model in 2.1? What are its degrees of freedom? What is the p-value? | chi square: \_\_\_\_\_\_\_\_\_\_\_ df: \_\_\_\_\_\_\_\_\_  p-value: \_\_\_\_\_\_\_\_\_\_\_\_\_ |

|  |  |
| --- | --- |
| 3 Answer this question using ONLY the likelihood ratio goodness-of-fit (**lrgof**) chi-square for the one model in this question. | CIRCLE ONE or FILL IN |
| 3.1 Does the model [SC] [CA] [CG] [SAY] [AYG] provide an adequate fit based on the **lrgof**? | adequate not adequate |
| 3.2 What is the value of the likelihood ratio chi-square for the model in 3.1? What are its degrees of freedom? What is the p-value? | chi square: \_\_\_\_\_\_\_\_\_\_\_ df: \_\_\_\_\_\_\_\_\_  p-value: \_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 3.3. If the model in 3.1 were true, would smoking and gender be conditionally independent give the other three variables? | yes no |

**Last** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2010 Final Exam: Answer Page 2 This is an exam. Do not discuss it.**

|  |  |  |
| --- | --- | --- |
| 4 Question 4 asks you to compare the simpler model [SC] [CA] [CG] [SAY] [AYG] and the more complex model [SC] [CA] [CG] [SAY] [AYG] [CAG] to see whether the added complexity is needed. | | CIRCLE ONE or FILL IN |
| 4.1 Is the fit of the simpler model adequate or is the CAG term needed. In this question, use the 0.05 level as the basis for your decision. | adequate not adequate | |
| 4.2 What is the value of the likelihood ratio chi-square for the test in 4.1? What are its degrees of freedom? What is the p-value? | chi square: \_\_\_\_\_\_\_\_\_\_\_ df: \_\_\_\_\_\_\_\_\_  p-value: \_\_\_\_\_\_\_\_\_\_\_\_\_ | |
| 4.3. If CAG were needed, would the odds ratio linking cocaine use (C) and alcohol (A) be different for males and females? | yes no | |

**5**. Fit the model [SC] [CA] [CG] [SAY] [AYG] setting eps=0.01. Use the fitted counts under this model to estimate the eight odds ratios linking smoking (S) with cocaine (C) for fixed levels of alcohol (A), age (Y) and gender (G). Fill in the following table with the eight fitted odds ratios.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Male | Male | Female | Female |
|  | Age 15-16 | Age 17-18 | Age 15-16 | Age 17-18 |
| Alcohol = 0 |  |  |  |  |
| Alcohol > 0 |  |  |  |  |

**6**. Fit the model [SC] [CA] [CG] [SAY] [AYG] setting eps=0.01. Use the fitted counts under this model to estimate the 16 conditional probabilities of cocaine use, cocaine>0, given the levels of the other four variables. Put the values in the table. **Round to 2 digits**, so probability 0.501788 rounds to 0.50. The first cell (upper left) is the estimate of the probability of cocaine use for a male, aged 15-16, who neither smokes nor drinks.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Male | Male | Female | Female |
|  |  | Age 15-16 | Age 17-18 | Age 15-16 | Age 17-18 |
| Smoke = 0 | Alcohol = 0 |  |  |  |  |
| Smoke = 0 | Alcohol > 0 |  |  |  |  |
| Smoke > 0 | Alcohol = 0 |  |  |  |  |
| Smoke > 0 | Alcohol > 0 |  |  |  |  |

**Answer Key: Stat 501 Final, Spring 2010**, Page 1

|  |  |
| --- | --- |
| 1 Answer this question using ONLY the likelihood ratio goodness-of-fit chi-square for the one model in this question. | CIRCLE ONE or FILL IN |
| 1.1. Does the hierarchical log-linear model with all 2-factor interactions (and no 3 factor interactions) provide an adequate fit to the data? | adequate not adequate |
| 1.2. What is the value of the likelihood ratio chi-square for the model in 1.1? What are its degrees of freedom? What is the p-value? | chi square: 32.4 df: 16  p-value: 0.00889 |

|  |  |
| --- | --- |
| 2 Answer this question using ONLY the likelihood ratio goodness-of-fit chi-square for the one model in this question. | CIRCLE ONE or FILL IN |
| 2.1 Which hierarchical log-linear model says smoking (S) is conditionally independent of gender (G) given the other three variables (C & A & Y)? The question asks for the largest or most complex model which has this condition. | [SCAY] [CAYG]  This is the most complex hierarchical model which has no u-term linking S and G, that is, no uSG(im) etc. |
| 2.2 Does the hierarchical log-linear model in 2.1 provide an adequate fit to the data? | adequate not adequate |
| 2.3. What is the value of the likelihood ratio chi-square for the model in 2.1? What are its degrees of freedom? What is the p-value? | chi square: 6.58 df: 8  p-value: 0.58 |

|  |  |
| --- | --- |
| 3 Answer this question using ONLY the likelihood ratio goodness-of-fit (**lrgof**) chi-square for the one model in this question. | CIRCLE ONE or FILL IN |
| 3.1 Does the model [SC] [CA] [CG] [SAY] [AYG] provide an adequate fit based on the **lrgof**? | adequate not adequate |
| 3.2 What is the value of the likelihood ratio chi-square for the model in 3.1? What are its degrees of freedom? What is the p-value? | chi square: 13.99 df: 16  p-value: 0.599 |
| 3.3. If the model in 3.1 were true, would smoking and gender be conditionally independent give the other three variables? | yes no  As in 2.1, there are no u-terms linking S and G. |

**Answer Key: Stat 501 Final, Spring 2010**, Page 2

|  |  |  |
| --- | --- | --- |
| 4 Question 4 asks you to compare the simpler model [SC] [CA] [CG] [SAY] [AYG] and the more complex model [SC] [CA] [CG] [SAY] [AYG] [CAG] to see whether the added complexity is needed. | | CIRCLE ONE or FILL IN |
| 4.1 Is the fit of the simpler model adequate or is the CAG term needed. In this question, use the 0.05 level as the basis for your decision. | adequate not adequate  Barely adequate – p-value is 0.089 | |
| 4.2 What is the value of the likelihood ratio chi-square for the test in 4.1? What are its degrees of freedom? What is the p-value? | chi square: 2.91 df: 1  p-value: 0.089 | |
| 4.3. If CAG were needed, would the odds ratio linking cocaine use (C) and alcohol (A) be different for males and females? | yes no | |

**5**. Fit the model [SC] [CA] [CG] [SAY] [AYG] setting eps=0.01. Use the fitted counts under this model to estimate the eight odds ratios linking smoking (S) with cocaine (C) for fixed levels of alcohol (A), age (Y) and gender (G). Fill in the following table with the eight fitted odds ratios.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Male | Male | Female | Female |
|  | Age 15-16 | Age 17-18 | Age 15-16 | Age 17-18 |
| Alcohol = 0 | 4.59 | 4.59 | 4.59 | 4.59 |
| Alcohol > 0 | 4.59 | 4.59 | 4.59 | 4.59 |

**6**. Fit the model [SC] [CA] [CG] [SAY] [AYG] setting eps=0.01. Use the fitted counts under this model to estimate the 16 conditional probabilities of cocaine use, cocaine>0, given the levels of the other four variables. Put the values in the table. **Round to 2 digits**, so probability 0.501788 rounds to 0.50. The first cell (upper left) is the estimate of the probability of cocaine use for a male, aged 15-16, who neither smokes nor drinks.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Male | Male | Female | Female |
|  |  | Age 15-16 | Age 17-18 | Age 15-16 | Age 17-18 |
| Smoke = 0 | Alcohol = 0 | 0.01 | 0.01 | 0.00 | 0.00 |
| Smoke = 0 | Alcohol > 0 | 0.07 | 0.07 | 0.05 | 0.05 |
| Smoke > 0 | Alcohol = 0 | 0.03 | 0.03 | 0.02 | 0.02 |
| Smoke > 0 | Alcohol > 0 | 0.27 | 0.27 | 0.20 | 0.20 |

**Spring 2010 Final: Doing the Exam in R**

Question 1. This model has all 10 = 5x4/2 pairwise interactions.

> **loglin(yrbs2007.2,list(c(1,2),c(1,3),c(1,4),c(1,5),c(2,3),c(2,4), c(2,5),c(3,4),c(3,5),c(4,5)))**

6 iterations: deviation 0.02655809

$lrt

[1] 32.38944

$df

[1] 16

> **1-pchisq(32.38944,16)**

[1] 0.00889451

Question 2. This model omits the [S,G] or [4,5] u-term and all higher order u-terms that contain it, but includes all other u-terms.

> **loglin(yrbs2007.2,list(c(1,2,3,4),c(2,3,4,5)))**

2 iterations: deviation 0

$lrt

[1] 6.578771

$df

[1] 8

> **1-pchisq(6.578771,8)**

[1] 0.5826842

Question 3.

> **loglin(yrbs2007.2,list(c(1,2),c(2,3),c(2,5),c(1,3,4),c(3,4,5)))**

5 iterations: deviation 0.05294906

$lrt

[1] 13.99041

$df

[1] 16

> **1-pchisq(13.99041,16)**

[1] 0.5994283

> **loglin(yrbs2007.2,list(c(1,2),c(2,3),c(2,5),c(1,3,4),c(3,4,5)))**

5 iterations: deviation 0.05294906

$lrt

[1] 13.99041

$df

[1] 16

> **loglin(yrbs2007.2,list(c(1,2),c(2,3),c(2,5),c(1,3,4),**

**c(3,4,5),c(2,3,5)))**

6 iterations: deviation 0.01950314

$lrt

[1] 11.07689

$df

[1] 15

> **13.9904108-11.076890**

[1] 2.913521

> **1-pchisq(2.914,1)**

[1] 0.08781383

Question 5.

> **mhat<-loglin(yrbs2007.2,list(c(1,2),c(2,3),c(2,5), c(1,3,4),c(3,4,5)),eps=0.01,fit=T)$fit**

6 iterations: deviation 0.005120433

> **mhat[,,1,1,1]**

C

S 0 times >0 times

No 2319.8584 10.135780

Yes 105.8827 2.123171

> **or<-function(tb){tb[1,1]\*tb[2,2]/(tb[1,2]\*tb[2,1])}**

> **or(mhat[,,1,1,1])**

[1] 4.58949

> **or(mhat[,,1,1,2])**

[1] 4.58949

> **or(mhat[,,1,2,1])**

[1] 4.58949

> **or(mhat[,,1,2,2])**

[1] 4.58949

> **or(mhat[,,2,1,1])**

[1] 4.58949

> **or(mhat[,,2,1,2])**

[1] 4.58949

> **or(mhat[,,2,2,1])**

[1] 4.58949

> **or(mhat[,,2,2,2])**

[1] 4.58949

Question 6.

> **round( mhat[,2,,,]/( mhat[,1,,,]+ mhat[,2,,,]),2)**

, , Y = 15-16, G = Female

A

S 0 times >0 times

No 0.00 0.05

Yes 0.02 0.20

, , Y = 17-18, G = Female

A

S 0 times >0 times

No 0.00 0.05

Yes 0.02 0.20

, , Y = 15-16, G = Male

A

S 0 times >0 times

No 0.01 0.07

Yes 0.03 0.27

, , Y = 17-18, G = Male

A

S 0 times >0 times

No 0.01 0.07

Yes 0.03 0.27

**Have a great summer!**

**Statistics 501, Spring 2008, Midterm: Data Page #1**

This is an exam. Do not discuss it with anyone. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question. Due in class Tuesday 25 March 2008. The data for this problem are at in the latest Rst501.RData for R users and in frozenM.txt as a text file at <http://stat.wharton.upenn.edu/statweb/course/Spring-2008/stat501> The list is case sensitive, so frozenM.text is with lower case items, and Rst501.RData is with upper case items.

The data are adapted from a paper by Hininger, et al. (2004), “Assessment of DNA damage by comet assay…” *Mutation Research*, 558-75-80. The paper is available from the library web page if you’d like to look at it, but that is not necessary to do this exam. There are ten nonsmokers (N) and ten smokers (S) in ten pairs matched for gender and approximately for age. For example, pair #1 consists of a female nonsmoker (Ngender=F) of age 24 (Nage=24) matched to a female smoker (Sgender=F) of age 26 (Sage=26). Using samples of frozen blood, the comet tail assay was performed to measure damage to DNA, with value Ndna=1.38 for the first nonsmoker and Sdna=3.07 for the first matched smoker. A photograph of the comet assay is given at <http://www.cometassayindia.org/definitions.htm>, although you do not need to examine this to do the problem. Also, for the smoker, there is a measure of cigarettes per day (CigPerDay) and years of smoking (YearsSm).

> is.data.frame(frozenM)

[1] TRUE

> frozenM

Nid Ngender Nage Ndna Sid Sgender Sage CigPerDay YearsSm Sdna

1 1 F 24 1.38 1 F 26 11 10 3.07

2 4 F 32 1.27 10 F 35 12 20 1.63

3 7 F 33 1.38 6 F 36 15 20 1.09

4 9 F 42 1.04 5 F 38 13 14 2.06

5 3 F 46 1.40 8 F 45 20 28 1.94

6 8 M 27 1.60 9 M 26 9 6 0.88

7 5 M 31 1.25 3 M 30 13 9 2.39

8 10 M 33 0.74 4 M 32 10 15 1.65

9 6 M 35 1.16 7 M 40 11 25 1.61

10 2 M 51 1.07 2 M 50 17 32 2.89

Test abbreviations:

SR = Wilcoxon’s signed rank test (3.1). HLsr = Hodges-Lehmann estimate associated with Wilcoxon’s signed rank test (3.2). RS = Wilcoxon’s rank sum test (4.1). HLrs = Hodges-Lehmann estimate associated with Wilcoxon’s rank sum test (4.2). AB = Ansari-Bradley test (5.1). LE = Lepage’s test (5.3). KS = Kolmogorov-Smirnov test (5.4). KW = Kruskal-Wallis test (6.1). OA = Jonckheere-Terpstra test for ordered alternatives (6.2). KE = Kendall’s test (8.1). TH = Theil’s test for a specified slope (9.1), THe = Theil’s estimate. When a question asks for a name of a test, give one of these abbreviations.

A “best” test should have the correct level when the null hypothesis is true (i.e., it should give P<=0.05 at most 5% of the time when the null hypothesis is true), it should be consistent against the stated alternative hypotheses (i.e., it should be nearly certain to give a P<0.05 if the alternative hypothesis is true and the sample size is very, very large). If more than one test has the correct level and is consistent, the best test will be targeted against the stated alternative, and will have greater power against it.

Be careful in reading and writing to distinguish distinct procedures with similar names. For instance, “the Wilcoxon test” or “the Hodges-Lehmann estimate” are not uniquely defined terms, since more than one procedure can be so defined, so “the Wilcoxon test” is a vague and hence incorrect answer to any question asking for a specific test. Use the test abbreviations given above, and write them clearly.

Some questions ask you to compare males to females, and the order (male-female) or (female-male) may slightly alter the answer. Give either answer. I will recognize either answer and realize what you did.

STATISTICS 501, SPRING 2008, MIDTERM DATA PAGE #2

Model 1: *Xi =*  + i, *i=1,2,…,m*, *Yi =*  +  + j+m, *j=1,…,n*, where k ~ iid, *k=1,2,…,n+m*, with a continuous distribution.

Model 2: *(X1,Y1), …, (Xn,Yn)* are n iid observations from a continuous bivariate distribution.

Model 3: *Xi =*  + i, *i=1,2,…,m*, *Yi =*  + j+m, *j=1,…,n*, where k ~ iid, *k=1,2,…,n+m*, with median zero and a continuous distribution,  >0.

Model 4: *Xi =*  + i, *i=1,2,…,m*, *Yi =*  +  + j+m, *j=1,…,n*, where k ~ iid, *k=1,2,…,n+m*, with median zero and a continuous distribution,  >0.

Model 5: *Yi* - *Xi* = i where i ~ iid, with a continuous distribution symmetric about 0, i=1,…,n.

Model 6: *Yi* =  + *Xi* + *ei*, …, where the *ei* are n iid observations from a continuous distribution with median zero independent of the *X*i which are untied.

Model 7: *Y1,…,Yn* ~ iid with a continuous distribution, *X1,…,Xm* ~ iid with a continuous distribution, with the *Y*’s and *X*’s independent of each other.

Model 8: *Yi* - *Xi* = i where i are independent, with possibly different continuous distributions symmetric each having median zero.

Model 9: *Xij =*  + j + ij, *i=1,2,…,N*, *j=1,…,K* *Yi =*  +  + j+m, *j=1,…,N*, where the NK ij’s are iid from a continuous distribution, with 0 = 1+…+K.

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2008, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. Use abbreviations and model #’s from the data page.

1. For each stated inference problem, insert the abbreviation of the most appropriate or best statistical procedure from those listed on the data page and then indicate the number of the model under which the procedure is appropriate. Do not do the tests, etc – just indicate the procedure and model. (16 points)

|  |  |  |
| --- | --- | --- |
| Problem | Abbreviation of statistical procedure | Model number |
| 1.1 For smokers, test the null hypothesis that years of smoking is independent of Sdna against the alternative that higher values of Sdna tend to be more common for smokers with more years of smoking. |  |  |
| 1.2 The investigator multiplies CigsPerDay and YearsSm to produce an index of smoking intensity, and fours three groups, low, medium and high, consisting of the lowest three smokers, the middle four smokers, and the highest three smokers. Test the null hypothesis that the three groups have the same distribution of Sdna against the alternative that the three groups differ in level in any way. |  |  |
| 1.3 Using Ndna, test the null hypothesis that male and female nonsmokers have the same level and dispersion of the Ndna results against the alternative that either the level or the dispersion or both differ for males and females. |  |  |
| 1.4 Give a point estimate of a shift in the distribution of Sdna when comparing male smokers to female smokers. |  |  |

2. Circle the correct answer. (16 points)

|  |  |
| --- | --- |
|  | CIRCLE ONE |
| 2.1 If the Ansari-Bradley test were used to no difference in Sdna between male and female smokers, the test would have little power to detect a difference in dispersion under model 4 if  were large. | TRUE FALSE |
| 2.2 The signed rank test is the appropriate test of *H0:=0* assuming model 8 is true. | TRUE FALSE |
| 2.3 To test *H0:=3* in model 6, apply Kendall’s rank correlation to test for zero correlation between *Yi-(+ei)* and 3*Xi*. | TRUE FALSE |
| 2.4 Under model 7, the Mann-Whitney U-statistic divided by nm estimates the probability that favorable results offset unfavorable ones in the sense that Pr{(Yi+Xi)/2 > 0}. | TRUE FALSE |

3. Use an appropriate nonparametric statistical procedure from the list on the data page to compare Sdna for smokers to Ndna for nonsmokers, with a view to seeing if the level is typically the same, or if the level is different for smokers, either higher or lower. Give the abbreviation of the test, the number of the model under which this test is appropriate, the numerical value of the test statistic, the two-sided p-value, the abbreviation of the associated point estimate, the numerical value of the point estimate, and the two-sided 95% confidence interval. Is it plausible that smokers and nonsmokers have the same level of the comet tail dna result? (16 points)

Test abbreviation: \_\_\_\_\_\_\_ Model #: \_\_\_\_\_\_\_\_\_\_ Value of statistic: \_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_

Estimate abbreviation: \_\_\_\_\_\_\_\_ Value of Estimate: \_\_\_\_\_\_\_\_\_\_ 95% CI: [ \_\_\_\_\_\_ , \_\_\_\_\_\_ ]

Is the null hypothesis plausible? CIRCLE ONE PLAUSIBLE NOT PLAUSIBLE

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2008, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. Use abbreviations and model #’s from the data page.

4. Use an appropriate nonparametric statistical procedure from the list on the data page to test the null hypothesis that, for smokers, the number of cigarettes per day is independent of the number of years of smoking, against the alternative that more years predicts either higher or lower consumption per day. What is the abbreviation of the test. What is the number of the model under which this test is appropriate? What is the two-sided p-value? What is the value of the associated estimate? What is the estimate of the probability of concordance between years and number of cigarettes? Is the null hypothesis plausible?

(12 points)

Test abbreviation: \_\_\_\_\_\_\_ Model #: \_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_ Numerical estimate: \_\_\_\_\_\_\_\_

Estimate of probability of concordance: \_\_\_\_\_\_\_\_\_\_

Is the null hypothesis plausible? CIRCLE ONE PLAUSIBLE NOT PLAUSIBLE

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5. Under model 6 relating Y=CigPerDay to X=YearsSm, is the null hypothesis *H0:=1* plausible when judged by an appropriate two-sided, 0.05 level nonparametric test? What is the abbreviation of the test? What is the two-sided p-value? BRIEFLY describe how you did the test. Is *H0:=1* plausible? What is the numerical value of the associated estimate of the slope **?

(12 points)

Test abbreviation: \_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_ Estimate of **: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Is the null hypothesis plausible? CIRCLE ONE PLAUSIBLE NOT PLAUSIBLE

Describe how you did the test:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6. Use an appropriate nonparametric statistical procedure from the list on the data page to compare Sdna for male smokers to Sdna for female smokers, with a view to seeing if the distributions are the same, or if the level is different for males than for females, either higher or lower. Give the abbreviation of the test, the number of the model under which this test is appropriate, the numerical value of the test statistic, the two-sided p-value, the abbreviation of the associated point estimate, the numerical value of the point estimate, and the two-sided 95% confidence interval. Is it plausible that male and female smokers have the same distribution of Sdna? (16 points)

Test abbreviation: \_\_\_\_\_\_\_ Model #: \_\_\_\_\_\_\_\_\_\_ Value of statistic: \_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_

Estimate abbreviation: \_\_\_\_\_\_\_\_ Value of Estimate: \_\_\_\_\_\_\_\_\_\_ 95% CI: [ \_\_\_\_\_\_ , \_\_\_\_\_\_ ]

Is the null hypothesis plausible? CIRCLE ONE PLAUSIBLE NOT PLAUSIBLE

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7. Assuming male and female nonsmokers have the same population median of Ndna, test the hypothesis that the distributions are the same against the alternative hypothesis that one group, male or female, is more dispersed than the other. What is the abbreviation of the test. What is the number of the model under which this test is appropriate? What is the value of the test statistic? What is the two-sided p-value? Is the null hypothesis plausible? (12 points)

Test abbreviation: \_\_\_\_\_\_\_ Model #: \_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_

Is the null hypothesis plausible? CIRCLE ONE PLAUSIBLE NOT PLAUSIBLE

Statistics 501, Spring 2008, Midterm: Data Page #1

This is an exam. Do not discuss it with anyone. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question. Due in class Tuesday 25 March 2008.

The data are adapted from a paper by Hininger, et al. (2004), “Assessment of DNA damage by comet assay…” *Mutation Research*, 558-75-80. The paper is available from the library web page if you’d like to look at it, but that is not necessary to do this exam. There are ten nonsmokers (N) and ten smokers (S) in ten pairs matched for gender and approximately for age. For example, pair #1 consists of a female nonsmoker (Ngender=F) of age 24 (Nage=24) matched to a female smoker (Sgender=F) of age 26 (Sage=26). Using samples of frozen blood, the comet tail assay was performed to measure damage to DNA, with value Ndna=1.38 for the first nonsmoker and Sdna=3.07 for the first matched smoker. A photograph of the comet assay is given at <http://www.cometassayindia.org/definitions.htm>, although you do not need to examine this to do the problem. Also, for the smoker, there is a measure of cigarettes per day (CigPerDay) and years of smoking (YearsSm).

> is.data.frame(frozenM)

[1] TRUE

> frozenM

Nid Ngender Nage Ndna Sid Sgender Sage CigPerDay YearsSm Sdna

1 1 F 24 1.38 1 F 26 11 10 3.07

2 4 F 32 1.27 10 F 35 12 20 1.63

3 7 F 33 1.38 6 F 36 15 20 1.09

4 9 F 42 1.04 5 F 38 13 14 2.06

5 3 F 46 1.40 8 F 45 20 28 1.94

6 8 M 27 1.60 9 M 26 9 6 0.88

7 5 M 31 1.25 3 M 30 13 9 2.39

8 10 M 33 0.74 4 M 32 10 15 1.65

9 6 M 35 1.16 7 M 40 11 25 1.61

10 2 M 51 1.07 2 M 50 17 32 2.89

Test abbreviations:

SR = Wilcoxon’s signed rank test (3.1). HLsr = Hodges-Lehmann estimate associated with Wilcoxon’s signed rank test (3.2). RS = Wilcoxon’s rank sum test (4.1). HLrs = Hodges-Lehmann estimate associated with Wilcoxon’s rank sum test (4.2). AB = Ansari-Bradley test (5.1). LE = Lepage’s test (5.3). KS = Kolmogorov-Smirnov test (5.4). KW = Kruskal-Wallis test (6.1). OA = Jonckheere-Terpstra test for ordered alternatives (6.2). KE = Kendall’s test (8.1). TH = Theil’s test for a specified slope (9.1), THe = Theil’s estimate. When a question asks for a name of a test, give one of these abbreviations.

A “best” test should have the correct level when the null hypothesis is true (i.e., it should give P<=0.05 at most 5% of the time when the null hypothesis is true), it should be consistent against the stated alternative hypotheses (i.e., it should be nearly certain to give a P<0.05 if the alternative hypothesis is true and the sample size is very, very large). If more than one test has the correct level and is consistent, the best test will be targeted against the stated alternative, and will have greater power against it.

Be careful in reading and writing to distinguish distinct procedures with similar names. For instance, “the Wilcoxon test” or “the Hodges-Lehmann estimate” are not uniquely defined terms, since more than one procedure can be so defined, so “the Wilcoxon test” is a vague and hence incorrect answer to any question asking for a specific test. Use the test abbreviations given above, and write them clearly.

STATISTICS 501, SPRING 2008, MIDTERM DATA PAGE #2

Model 1**: *Xi =*  + i, *i=1,2,…,m*, *Yi =*  +  + j+m, *j=1,…,n*, where k ~ iid, *k=1,2,…,n+m*, with a continuous distribution.**

Model 2**: *(X1,Y1), …, (Xn,Yn)* are n iid observations from a continuous bivariate distribution.**

Model 3**: *Xi =*  + i, *i=1,2,…,m*, *Yi =*  + j+m, *j=1,…,n*, where k ~ iid, *k=1,2,…,n+m*, with median zero and a continuous distribution,  >0.**

Model 4**: *Xi =*  + i, *i=1,2,…,m*, *Yi =*  +  + j+m, *j=1,…,n*, where k ~ iid, *k=1,2,…,n+m*, with median zero and a continuous distribution,  >0.**

Model 5**: *Yi* - *Xi* = i where i ~ iid, with a continuous distribution symmetric about 0, i=1,…,n.**

Model 6**: *Yi* =  + *Xi* + *ei*, …, where the *ei* are n iid observations from a continuous distribution with median zero independent of the *X*i which are untied.**

Model 7**: *Y1,…,Yn* ~ iid with a continuous distribution, *X1,…,Xm* ~ iid with a continuous distribution, with the *Y*’s and *X*’s independent of each other.**

Model 8**: *Yi* - *Xi* = i where i are independent, with possibly different continuous distributions each having median zero.**

Model 9**: *Xij =*  + j + ij, *i=1,2,…,N*, *j=1,…,K* where the NK ij’s are iid from a continuous distribution, with 0 = 1+…+K.**

Answers Statistics 501, Spring 2008, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. **Use abbreviations and model #’s from the data page.**

1. For each stated inference problem, insert the abbreviation of the most appropriate or best statistical procedure from those listed on the data page and then indicate the number of the model under which the procedure is appropriate. Do not do the tests, etc – just indicate the procedure and model. (16 points)

|  |  |  |
| --- | --- | --- |
| Problem | Abbreviation of statistical procedure | Model number |
| 1.1 For smokers, test the null hypothesis that years of smoking is independent of Sdna against the alternative that higher values of Sdna tend to be more common for smokers with more years of smoking. | KE  Not TH because a line is not assumed in the question. | 2 |
| 1.2 The investigator multiplies CigsPerDay and YearsSm to produce an index of smoking intensity, and fours three groups, low, medium and high, consisting of the lowest three smokers, the middle four smokers, and the highest three smokers. Test the null hypothesis that the three groups have the same distribution of Sdna against the alternative that the three groups differ in level in any way. | KW  Not OA because of the final words “in any way” | 9 |
| 1.3 Using Ndna, test the null hypothesis that male and female nonsmokers have the same level and dispersion of the Ndna results against the alternative that either the level or the dispersion or both differ for males and females. | LE  “level or the dispersion  or both” | 4 |
| 1.4 Give a point estimate of a shift in the distribution of Sdna when comparing male smokers to female smokers. | HLrs | 1 |

2. Circle the correct answer. (16 points)

|  |  |
| --- | --- |
|  | CIRCLE ONE |
| 2.1 If the Ansari-Bradley test were used to no difference in Sdna between male and female smokers, the test would have little power to detect a difference in dispersion under model 4 if  were large. | TRUE FALSE |
| 2.2 The signed rank test is the appropriate test of *H0:=0* assuming model 8 is true. | Need symmetry as in Model 5 for SR  TRUE FALSE |
| 2.3 To test *H0:=3* in model 6, apply Kendall’s rank correlation to test for zero correlation between *Yi-(+ei)* and 3*Xi*. | Close but very missed up!  TRUE FALSE |
| 2.4 Under model 7, the Mann-Whitney U-statistic divided by nm estimates the probability that favorable results offset unfavorable ones in the sense that Pr{(Yi+Xi)/2 > 0}. | Close but very messed up!  TRUE FALSE |

3. Use an appropriate nonparametric statistical procedure from the list on the data page to compare Sdna for smokers to Ndna for nonsmokers, with a view to seeing if the level is typically the same, or if the level is different for smokers, either higher or lower. Give the abbreviation of the test, the number of the model under which this test is appropriate, the numerical value of the test statistic, the two-sided p-value, the abbreviation of the associated point estimate, the numerical value of the point estimate, and the two-sided 95% confidence interval. Is it plausible that smokers and nonsmokers have the same level of the comet tail dna result? (16 points)

Test abbreviation: SR Model #: 5 Value of statistic: 49 P-value: 0.02734

Estimate abbreviation: HLsr Value of Estimate: 0.725 95% CI: [0.095, 1.300 ]

Is the null hypothesis plausible? CIRCLE ONE PLAUSIBLE NOT PLAUSIBLE

Answers Midterm Spring 2008, Page 2

This is an exam. Do not discuss it with anyone. **Use abbreviations and model #’s from the data page.**

4. Use an appropriate nonparametric statistical procedure from the list on the data page to test the null hypothesis that, for smokers, the number of cigarettes per day is independent of the number of years of smoking, against the alternative that more years predicts either higher or lower consumption per day. What is the abbreviation of the test. What is the number of the model under which this test is appropriate? What is the two-sided p-value? What is the value of the associated estimate? What is the estimate of the probability of concordance between years and number of cigarettes? Is the null hypothesis plausible?

(12 points)

Test abbreviation: KE Model #: 2 P-value: 0.06422 Numerical estimate: 0.4598

Estimate of probability of concordance: (0.4598+1)/2 = 0.73

Is the null hypothesis plausible? CIRCLE ONE Barely PLAUSIBLE NOT PLAUSIBLE

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5. Under model 6 relating Y=CigPerDay to X=YearsSm, is the null hypothesis *H0:=1* plausible when judged by an appropriate two-sided, 0.05 level nonparametric test? What is the abbreviation of the test? What is the two-sided p-value? BRIEFLY describe how you did the test. Is *H0:=1* plausible? What is the numerical value of the associated estimate of the slope **?

(12 points)

Test abbreviation: TH P-value: 0.0004377 Estimate of **: 0.2222

Is the null hypothesis plausible? CIRCLE ONE PLAUSIBLE NOT PLAUSIBLE

Describe how you did the test: Do Kendall’s correlation between Y-1X and X.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6. Use an appropriate nonparametric statistical procedure from the list on the data page to compare Sdna for male smokers to Sdna for female smokers, with a view to seeing if the distributions are the same, or if the level is different for males than for females, either higher or lower. Give the abbreviation of the test, the number of the model under which this test is appropriate, the numerical value of the test statistic (as reported by R), the two-sided p-value, the abbreviation of the associated point estimate, the numerical value of the point estimate, and the two-sided 95% confidence interval. Is it plausible that male and female smokers have the same distribution of Sdna? (16 points)

Test abbreviation: RS Model #: 1 Value of statistic: 14 or 11 P-value: 0.84

Estimate abbreviation: HLrs Value of Estimate: 0.18 95% CI: [-1.26 1.42 ]

Is the null hypothesis plausible? CIRCLE ONE PLAUSIBLE NOT PLAUSIBLE

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7. Assuming male and female nonsmokers have the same population median of Ndna, test the hypothesis that the distributions are the same against the alternative hypothesis that one group, male or female, is more dispersed than the other. What is the abbreviation of the test. What is the number of the model under which this test is appropriate? What is the value of the test statistic? What is the two-sided p-value? Is the null hypothesis plausible? (12 points)

Test abbreviation: AB Model #: 3 P-value: 0.8254

Is the null hypothesis plausible? CIRCLE ONE PLAUSIBLE NOT PLAUSIBLE

Doing the problem set in R (Spring 2008)

> frozenM

Nid Ngender Nage Ndna Sid Sgender Sage CigPerDay YearsSm Sdna

1 1 F 24 1.38 1 F 26 11 10 3.07

2 4 F 32 1.27 10 F 35 12 20 1.63

3 7 F 33 1.38 6 F 36 15 20 1.09

4 9 F 42 1.04 5 F 38 13 14 2.06

5 3 F 46 1.40 8 F 45 20 28 1.94

6 8 M 27 1.60 9 M 26 9 6 0.88

7 5 M 31 1.25 3 M 30 13 9 2.39

8 10 M 33 0.74 4 M 32 10 15 1.65

9 6 M 35 1.16 7 M 40 11 25 1.61

10 2 M 51 1.07 2 M 50 17 32 2.89

> attach(frozenM)

Question 3:

> wilcox.test(Sdna-Ndna,conf.int=T)

Wilcoxon signed rank test

data: Sdna - Ndna

V = 49, p-value = 0.02734

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

0.095 1.300

sample estimates:

(pseudo)median

0.725

Question 4:

> cor.test(CigPerDay,YearsSm,method="kendall")

Kendall's rank correlation tau

data: CigPerDay and YearsSm

z = 1.8507, p-value = 0.06422

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.4598005

> (0.4598005+1)/2

[1] 0.7299003

Question 5:

> cor.test(CigPerDay-1\*YearsSm,YearsSm,method="kendall")

Kendall's rank correlation tau

data: CigPerDay - 1 \* YearsSm and YearsSm

z = -3.5163, p-value = 0.0004377

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

-0.873621

> median(theil(YearsSm,CigPerDay))

[1] 0.2222222

Question 6:

> wilcox.test(Sdna[Sgender=="F"],Sdna[Sgender=="M"],conf.int=T)

Wilcoxon rank sum test

data: Sdna[Sgender == "F"] and Sdna[Sgender == "M"]

W = 14, p-value = 0.8413

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

-1.26 1.42

sample estimates:

difference in location

0.18

or

> wilcox.test(Sdna[Sgender=="M"],Sdna[Sgender=="F"],conf.int=T)

Wilcoxon rank sum test

data: Sdna[Sgender == "M"] and Sdna[Sgender == "F"]

W = 11, p-value = 0.8413

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

-1.42 1.26

sample estimates:

difference in location

-0.18

Question 7:

> ansari.test(Sdna[Sgender=="M"],Sdna[Sgender=="F"])

Ansari-Bradley test

data: Sdna[Sgender == "M"] and Sdna[Sgender == "F"]

AB = 14, p-value = 0.8254

alternative hypothesis: true ratio of scales is not equal to 1

or

> ansari.test(Sdna[Sgender=="F"],Sdna[Sgender=="M"])

Ansari-Bradley test

data: Sdna[Sgender == "F"] and Sdna[Sgender == "M"]

AB = 16, p-value = 0.8254

alternative hypothesis: true ratio of scales is not equal to 1

**Statistics 501 Spring 2008 Final Exam: Data Page 1**

This is an exam. Do not discuss it with anyone.

The data are from: Pai and Saleh (2008) Exploring motorcyclist injury severity in approach-turn collisions at T-junctions: Focusing on the effects of driver’s failure to yield and junction control measures, *Accident Analysis and Prevention*, 40, 479-486. The paper is available as an e-journal at the UPenn library, but there is no need to look at the paper unless you want to do so. The data described 17,716 motorcycle crashes involving another vehicle at a T junction. The “injury” to the motorcyclist was either KSI=(killed or seriously injured) or Other=(no injury or slight injury). The intersection was “controlled” by a Sign=(stop, give-way signs or markings) or by Signal=(automatic signals) or it was Uncon=(uncontrolled). There were two types of crash, A and B, depicted in the figure. In A, the motorcyclist collided with a turning car. In B, the car collided with a turning motorcyclist. The variables are I=injury, C=Control, T=CrashType. **Refer to the variables using the letters I, C and T**.

> **TurnCrash**

, , CrashType = A

Control

Injury Uncon Sign Signal

KSI 653 4307 331

Other 1516 8963 884

, , CrashType = B

Control

Injury Uncon Sign Signal

KSI 27 176 53

Other 78 592 136

Pai and Saleh write: “In this study an approach-turn crash is classified into two sub-crashes—approach-turn A: a motorcycle approaching straight collides with a vehicle travelling from opposite direction and turning right into such motorcycle's path; and approach-turn B crash: an approaching vehicle is in a collision with a motorcycle travelling from opposite direction and turning right into such vehicle's path (this categorisation includes either a vehicle or motorcycle making a U-turn onto the same street as the approaching vehicle/motorcycle). The categorisation is schematically illustrated in Figure 1.

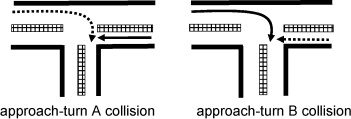


Figure 1. Schematic diagram of approach-turn A/B collisions at T-junctions. *Note*: Pecked line represents the intended path of the vehicle; solid line represents the intended path of the motorcycle.”

**Statistics 501 Spring 2008 Final Exam: Data Page 2**

This is an exam. Do not discuss it with anyone.

The data littlegrogger is based on the grogger data set in Jeffrey Wooldridge’s (2002) book *Econometric Analysis of Cross Section and Panel Data* is due to Jeffrey Grogger. In littlegrogger, there are three variables, farr = 1 if arrested for a felony in 1986, 0 otherwise, pcnv = proportion of prior arrests that resulted in conviction, and durat = recent unemployment duration in months.

> **dim(littlegrogger)**

[1] 2725 3

> **littlegrogger[1:3,]**

farr pcnv durat

1 0 0.38 0

2 1 0.44 0

3 1 0.33 11

> **summary(littlegrogger)**

farr pcnv durat

Min. :0.0000 Min. :0.0000 Min. : 0.000

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.: 0.000

Median :0.0000 Median :0.2500 Median : 0.000

Mean :0.1798 Mean :0.3578 Mean : 2.251

3rd Qu.:0.0000 3rd Qu.:0.6700 3rd Qu.: 2.000

Max. :1.0000 Max. :1.0000 Max. :25.000

Model #1 asserts log{Pr(farr=1)/Pr(farr=0)} =  +  pcnv +  durat

Model #2 asserts log{Pr(farr=1)/Pr(farr=0)} =  +  pcnv +  durat +  pcnv  durat

The data TurnCrash and littlegrogger for this problem set are at in the latest Rst501.RData for R users and in TurnCrash.txt and littlegrogger.txt as a text files at

<http://stat.wharton.upenn.edu/statweb/course/Spring-2008/stat501>Keep in mind that the list is case-sensitive, so upper and lower case files are in different places.

Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question.

**Make and keep a photocopy of your answer page**.Place the exam in an envelope with ‘Paul Rosenbaum, Statistics Department’ on it**. The exam is due in my office, 473 Huntsman, on Wednesday, May 7 at 12:00am.** You may turn in the exam early at my mail box in the Statistics Department, 4th floor, Huntsman. If you would like to receive your graded exam, final grade, and an answer key, then include a stamped, self-addressed, regular envelope. (I will send just two pages, so a regular envelope with regular postage should do it.)

Last Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2008 Final Exam: Answer Page 1 This is an exam. Do not discuss it.**

|  |  |
| --- | --- |
| These questions refer to the TurnCrash data | CIRCLE ONE |
| **1.1** The model [IC] [CT] says Injury is independent of CrashType. | TRUE FALSE |
| **1.2** The model [IT] [CT] says Injury is independent of CrashType. | TRUE FALSE |
| **1.3** The model [IC] [IT] [CT] says that Injury and CrashType are dependent but the relationship is indirect through Control. | TRUE FALSE |
| **1.4** If [IC][CT] were the correct model, then one can collapse over Control without changing the relationship between Injury and CrashType, where relationships are measured by odds ratios. | TRUE FALSE |
| **1.5** The model [IC] [CT] preserves the marginal table of Injury with CrashType. | TRUE FALSE |
| **1.6** The model [IC] [CT] is not hierarchical. | TRUE FALSE |
| **1.7** The model [IC] [CT] is nested within the model [IT] [CT]. | TRUE FALSE |

**2.** Test the null hypothesis that model [IC] [CT] is correct against the alternative model

[IC] [IT] [CT]. What is numerical value of the relevant chi-square statistic? What are its degrees of freedom? What is the p-value? Is the null hypothesis plausible?

Value of chi-square: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ DF:\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_\_\_

The null hypothesis is: (CIRCLE ONE)

PLAUSIBLE NOT PLAUSIBLE

**3.** Fit the model [IC] [IT] [CT] and use the fitted counts to compute the odds ratios linking Injury with CrashType for each of the three levels of Control. Put the odds ratios in the table below:

|  |  |  |
| --- | --- | --- |
| Control=Uncon | Control=Sign | Control=Signal |
|  |  |  |

Last Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#: \_\_\_\_\_

**Stat 501 S-2008 Final Exam: Answer Page 2 This is an exam. Do not discuss it.**

**4.** Fit the model [ICT] and use the fitted counts to compute the odds ratios linking Injury with CrashType for each of the three levels of Control. Put the odds ratios in the table below.

|  |  |  |
| --- | --- | --- |
| Control=Uncon | Control=Sign | Control=Signal |
|  |  |  |

Is the simpler model, [IC] [IT] [CT], an *adequate fit* to the data, or is it *implausible*, so [ICT] should be used instead? Here, implausible means rejected by an appropriate test?

(CIRCLE ONE)

ADEQUATE FIT IMPLAUSIBLE

Is it reasonably *accurate to say* that when the intersection is controlled by a signal, crash type is not much associated with degree of injury, but if the intersection is controlled by a sign then crash type A is more likely to be associated with KSI than crash type B? (CIRCLE ONE)

ACCURATE TO SAY NOT ACCURATE

**5.** Use the littlegrogger data to fit models #1 and #2 on the data page. Use the fit to answer the following questions.

|  |  |
| --- | --- |
| Question | CIRCLE ONE or Write Answer in Space |
| In model #1, give the estimate of , an approximate 95% confidence interval, and the two-sided p-value for testing H0:=0. | Estimate: 95%CI: p-value  \_\_\_\_\_\_\_\_\_ [\_\_\_\_\_\_, \_\_\_\_\_\_] \_\_\_\_\_\_\_\_\_ |
| Consider two individuals with no recent unemployment (durat=0). The estimate of  in model 1 suggests that of these two individuals, the one with a higher proportion of previous convictions is MORE/LESS likely to arrested for a felony than the individual with a lower proportion. | MORE LESS |
| The third individual in littlegrogger has pcnv=0.33 and durat=11. What is the estimated probability that this individual will be arrested for a felony? | Estimated probability: \_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Use the z-value to test the hypothesis that H0:=0 in model #2. What is the z-value? What is the two-sided p-value? Is the hypothesis plausible? | z-value: \_\_\_\_\_\_\_\_ p-value: \_\_\_\_\_\_\_  PLAUSIBLE NOT |
| Use the likelihood ratio chi-square to test the hypothesis that H0:=0 in model #2. What is chi-square? What is the p-value? | Chi square: \_\_\_\_\_\_\_\_\_ p-value: \_\_\_\_\_\_\_\_ |

**Have a great summer!**

**Stat 501 S-2008 Final Exam: Answers**

|  |  |
| --- | --- |
| These questions refer to the TurnCrash data | CIRCLE ONE (3 points each, 21 total) |
| **1.1** The model [IC] [CT] says Injury is independent of CrashType. | TRUE FALSE |
| **1.2** The model [IT] [CT] says Injury is independent of CrashType. | TRUE FALSE |
| **1.3** The model [IC] [IT] [CT] says that Injury and CrashType are dependent but the relationship is indirect through Control. | TRUE FALSE |
| **1.4** If [IC][CT] were the correct model, then one can collapse over Control without changing the relationship between Injury and CrashType, where relationships are measured by odds ratios. | TRUE FALSE |
| **1.5** The model [IC] [CT] preserves the marginal table of Injury with CrashType. | TRUE FALSE |
| **1.6** The model [IC] [CT] is not hierarchical. | TRUE FALSE |
| **1.7** The model [IC] [CT] is nested within the model [IT] [CT]. | TRUE FALSE |

**2.** Test the null hypothesis that model [IC] [CT] is correct against the alternative model

[IC] [IT] [CT]. What is numerical value of the relevant chi-square statistic? What are its degrees of freedom? What is the p-value? Is the null hypothesis plausible? (20 points)

Value of chi-square: 25.86=33.17-7.31\_ DF: 1 =3-2 P-value: 3.6 x 10-7

The null hypothesis is: (CIRCLE ONE)

PLAUSIBLE NOT PLAUSIBLE

**3.** Fit the model [IC] [IT] [CT] and use the fitted counts to compute the odds ratios linking Injury with CrashType for each of the three levels of Control. Put the odds ratios in the table below: (10 points)

|  |  |  |
| --- | --- | --- |
| Control=Uncon | Control=Sign | Control=Signal |
| 1.44 | 1.44 | 1.44 |

**Answers**

**4.** Fit the model [ICT] and use the fitted counts to compute the odds ratios linking Injury with CrashType for each of the three levels of Control. Put the odds ratios in the table below. (19 points)

|  |  |  |
| --- | --- | --- |
| Control=Uncon | Control=Sign | Control=Signal |
| 1.24 | 1.62 | 0.96 |

Is the simpler model, [IC] [IT] [CT], an *adequate fit* to the data, or is it *implausible*, so [ICT] should be used instead? Here, implausible means rejected by an appropriate test?

(CIRCLE ONE)

ADEQUATE FIT IMPLAUSIBLE

Is it reasonably *accurate to say* that when the intersection is controlled by a signal, crash type is not much associated with degree of injury, but if the intersection is controlled by a sign then crash type A is more likely to be associated with KSI than crash type B? (CIRCLE ONE)

ACCURATE TO SAY NOT ACCURATE

**5.** Use the littlegrogger data to fit models #1 and #2 on the data page. Use the fit to answer the following questions. (6 points each, 30 total)

|  |  |
| --- | --- |
| Question | CIRCLE ONE or Write Answer in Space |
| **5.1** In model #1, give the estimate of , an approximate 95% confidence interval, and the two-sided p-value for testing H0:=0. | Estimate: 95%CI: p-value  -0.662 [-0.93, -0.39] 1.4x10-6 |
| **5.2** Consider two individuals with no recent unemployment (durat=0). The estimate of  in model 1 suggests that of these two individuals, the one with a higher proportion of previous convictions is MORE/LESS likely to arrested for a felony than the individual with a lower proportion. | MORE LESS |
| **5.3** The third individual in littlegrogger has pcnv=0.33 and durat=11. What is the estimated probability that this individual will be arrested for a felony? | Estimated probability: 0.25 |
| **5.4** Use the z-value to test the hypothesis that H0:=0 in model #2. What is the z-value? What is the two-sided p-value? Is the hypothesis plausible? | z-value: 1.06 p-value: 0.29  PLAUSIBLE NOT |
| **5.5** Use the likelihood ratio chi-square to test the hypothesis that H0:=0 in model #2. What is chi-square? What is the p-value? | Chi square: 1.1 p-value: 0.29 |

**Doing the Problem Set in R**

**Spring 2008, Final Exam**

Question 2

> **loglin(TurnCrash,list(c(1,2),c(1,3),c(2,3)))**

$lrt

[1] 7.307596

$df

[1] 2

> **loglin(TurnCrash,list(c(1,2),c(2,3)))**

$lrt

[1] 33.16735

$df

[1] 3

> **33.16735-7.307596**

[1] 25.85975

> 3-2

1

> **1-pchisq(25.85975,1)**

[1] 3.671455e-07

Question 3: Compute the odds ratios from the fitted counts

> **loglin(TurnCrash,list(c(1,2),c(1,3),c(2,3)),fit=T)$fit**

Question 4: The saturated model, [ICT] is just the observed data with chi-square of 0 on 0 df.

> **loglin(TurnCrash,list(c(1,2),c(1,3),c(2,3)))**

$lrt

[1] 7.307596

$df

[1] 2

> **1-pchisq( 7.307596,2)**

[1] 0.0258926

Question 5.1-2

> **summary(glm(farr~pcnv+durat,family=binomial))**

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.442034 0.069998 -20.601 < 2e-16 \*\*\*

pcnv -0.662226 0.137211 -4.826 1.39e-06 \*\*\*

durat 0.053424 0.009217 5.797 6.77e-09 \*\*\*

---

Null deviance: 2567.6 on 2724 degrees of freedom

Residual deviance: 2510.7 on 2722 degrees of freedom

95% Confidence interval for 

> **-0.662226+0.137211\*c(-1.96,1.96)**

[1] -0.9311596 -0.3932924

Question 5.3

> **glm(farr~pcnv+durat,family=binomial)$fitted.values[1:5]**

1 2 3 4 5

0.1552925 0.1501515 0.2548513 0.1669234 0.1996298

Question 5.4

> **summary(glm(farr~pcnv+durat+pcnv\*durat,family=binomial))**

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.42358 0.07192 -19.795 < 2e-16 \*\*\*

pcnv -0.73644 0.15521 -4.745 2.09e-06 \*\*\*

durat 0.04571 0.01181 3.869 0.000109 \*\*\*

pcnv:durat 0.02964 0.02801 1.058 0.289980

---

Null deviance: 2567.6 on 2724 degrees of freedom

Residual deviance: 2509.6 on 2721 degrees of freedom

Question 5.5

> **2510.7-2509.6**

[1] 1.1

> **2722- 2721**

[1] 1

> **1-pchisq(1.1,1)**

[1] 0.2942661

**Statistics 501, Spring 2007, Midterm: Data Page #1**

This is an exam. Do not discuss it with anyone. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question. Due in class Tuesday 20 March 2007.

The data are adapted from a paper by Botta, et al. (2006), Assessment of occupational exposure to welding fumes by inductively coupled plasma-mass spectroscopy and by the comet assay, *Environmental and Molecular Mutagenesis*, 27, 284-295. The paper is available from the library web page if you’d like to look at it, but that is not necessary to do this exam. The data are “adapted” only in the sense that the first ten observations are used – this is to simplify life for anyone who prefers to do the computations “by hand.”

The study concerned the possibility that exposure to welding fumes promptly damages to DNA. The data below concern ten welders and ten unrelated controls. The metal arc welders worked in building industries in the south of France, and the controls worked in the same industries, but were not exposed to welding fumes. The outcome measure (OTM) is the median olive tail moment of the comet tail assay. For the ten welders, *i*=1,…,*n*=10, measurements were taken at the beginning of the work week (BoW) and at the end of the work week (EoW). For the unrelated, unexposed controls, *j*=1,…,*m*=10, measurements were taken at the beginning of the work week. (Beginning=Monday, End=Friday). For instance, the first welder had OTM=0.87 at the beginning of the week, and OTM=3.92 at the end of the week. The first control, in no particular order, had OTM=1.73 at the beginning of the work week. (In the original data, there are 30 welders and 22 controls.) Notation: Write *Xi* = BoW for welder *i*, *Yi* = EoW for welder *i*, and *Zj* = Control for control *j*, so *X2* = 1.13, *Y2* = 4.39, and *Z2* = 1.45. The data are in Fume in the latest Rst501 workspace and in a txt file, Fume.txt.

> Fume

EoW BoW Control

1 3.92 0.87 1.73

2 4.39 1.13 1.45

3 5.29 1.61 1.63

4 4.04 0.87 0.96

5 3.06 1.28 1.41

6 6.03 2.60 0.91

7 3.21 0.57 1.93

8 7.90 2.40 0.94

9 3.23 2.50 1.62

10 4.33 2.11 1.57

Test abbreviations:

SR = Wilcoxon’s signed rank test (3.1). HLsr = Hodges-Lehmann estimate associated with Wilcoxon’s signed rank test (3.2). RS = Wilcoxon’s rank sum test (4.1). HLrs = Hodges-Lehmann estimate associated with Wilcoxon’s rank sum test (4.2). AB = Ansari-Bradley test (5.1). LE = Lepage’s test (5.3). KS = Kolmogorov-Smirnov test (5.4). KW = Kruskal-Wallis test (6.1). OA = Jonckheere-Terpstra test for ordered alternatives (6.2). KE = Kendall’s test (8.1). TH = Theil’s test for a specified slope (9.1), THe = Theil’s estimate. When a question asks for a name of a test, give one of these abbreviations.

A “best” test should have the correct level when the null hypothesis is true (i.e., it should give P<=0.05 at most 5% of the time when the null hypothesis is true), it should be consistent against the stated alternative hypotheses (i.e., it should be nearly certain to give a P<0.05 if the alternative hypothesis is true and the sample size is very, very large). If more than one test has the correct level and is consistent, the best test will be targeted against the stated alternative, and will have greater power against it.

Be careful in reading and writing to distinguish distinct procedures with similar names. For instance, “the Wilcoxon test” or “the Hodges-Lehmann estimate” are not uniquely defined terms, since more than one procedure can be so defined, so “the Wilcoxon test” is a vague and hence incorrect answer to any question asking for a specific test. Use the test abbreviations given above, and write them clearly.

STATISTICS 501, SPRING 2007, MIDTERM DATA PAGE #2

Model 1: *Yi* - *Xi* = i where i ~ iid, with a continuous distribution symmetric about 0, i=1,…,n.

Model 2: *Yi* - *Zi* = i where i ~ iid, with a continuous distribution symmetric about 0, i=1,…,n.

Model 3: *Zi* - *Xi* = i where i ~ iid, with a continuous distribution symmetric about 0, i=1,…,n.

Model 4: *Y1,…,Yn* ~ iid with a continuous distribution, *X1,…,Xn* ~ iid with a continuous distribution, with the *Y*’s and *X*’s independent of each other.

Model 5: *Y1,…,Yn* ~ iid with a continuous distribution, *Z1,…,Zm* ~ iid with a continuous distribution, with the *Y*’s and *Z*’s independent of each other.

Model 6: *X1,…,Xn* ~ iid with a continuous distribution, *Z1,…,Zm* ~ iid with a continuous distribution, with the *X*’s and *Z*’s independent of each other.

Model 7: *Y1,…,Yn* ~ iid with a continuous distribution with median , *Z1,…,Zm* ~ iid with a continuous distribution with median , with the *Y*’s and *Z*’s independent of each other, and (*Yj*-) having the same distribution as (*Zi* - ) for each i,j, for some >0.

Model 8: *Y1,…,Yn* ~ iid with a continuous distribution, *X1,…,Xn* ~ iid with a continuous distribution, with the *Y*’s and *X*’s independent of each other, and *Yj* having the same distribution as *Xi* +  for each *i,j*.

Model 9: *Y1,…,Yn* ~ iid with a continuous distribution, *Z1,…,Zm* ~ iid with a continuous distribution, with the *Y*’s and *Z*’s independent of each other, and *Yi* having the same distribution as *Zj* +  for each *i,j*.

Model 10: *X1,…,Xn* ~ iid with a continuous distribution, *Z1,…,Zm* ~ iid with a continuous distribution, with the *X*’s and *Z*’s independent of each other, and *Xi* having the same distribution as *Zj* +  for each *i,j*.

Model 11: *(X1,Y1), …, (Xn,Yn)* are n iid observations from a continuous bivariate distribution.

Model 12: *Yi* =  + *Xi* + *ei*, …, where the *ei* are n iid observations from a continuous distribution with median zero independent of the *X*i which are untied.

Model 13: *Y1,…,Yn* ~ iid with a continuous distribution, *Z1,…,Zm* ~ iid with a continuous distribution, with the Y’s and Z’s independent of each other, and *Yi* having the same distribution as  + *Zj* for each *i,j* for some >0.

Model 14: *Y1,…,Yn* ~ iid with a continuous distribution symmetric about its median, y; *Z1,…,Zm* ~ iid with a continuous distribution symmetric about its median, x, with the *Y*’s and *Z*’s independent of each other.

Reason A: Medians look different. Reason B: Interquartile ranges look different.

Reason C: *Yi* should be independent of *Xi* Reason D: *Yi* should be independent of *Zi*

Reason E: *Xi* should be independent of *Zi* Reason F: Distribution of *Xi* looks asymmetric

Reason G: RS is inappropriate unless distributions are shifted

Reason H: RS is inappropriate unless distributions are symmetric

Reason I: SR is inappropriate unless distribution of differences is shifted

Reason J: SR is inappropriate unless distributions are symmetric

Reason K: Data are paired, not independent Reason L: Data are independent, not paired

Reason M: If the distributions are not shifted, you cannot estimate the amount by which they are shifted.

Reason N: If the distributions are not symmetric, you cannot estimate the center of symmetry.

Reason O: The KS test does not have the correct level if this model is true.

Reason P: The KS test has the correct level if this model is true.

Reason Q: The population mean (that is, the expectation) may not exist if this model is true.

Reason R: This method works with asymmetric distributions, but not antisymmetric distributions.

Reason S: The medians look about the same. Reason T: Need paired data for HLrs

Reason U: When viewed as a U-statistic, RS tests H0: no difference vs H1: Prob(Y>Z) not ½ Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2007, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. See the data page for abbreviations.

1. Plot the data. Think about the design of the study. Circle “true” if the statement is true for these data, and circle “false” if it is false for these data. Give the letter of the one most appropriate reason. 24 points

|  |  |  |
| --- | --- | --- |
| Statement (Is it true or false? Why? Use Reason Letters from data page.) | CIRCLE ONE | One reason letter |
| a. Model 1 is clearly inappropriate for these data. | TRUE FALSE |  |
| b. Model 2 is clearly inappropriate for these data. | TRUE FALSE |  |
| c. Model 8 is clearly inappropriate for these data. | TRUE FALSE |  |
| d. Model 9 is clearly inappropriate for these data. | TRUE FALSE |  |
| e. Under Model 5, the KS test could be used to test whether the EoW=*Yi* measurements have the same distribution as the Control=*Zj* measurements. | TRUE FALSE |  |
| f. Under Model 13, the HLrs estimate could be used to estimate  | TRUE FALSE |  |
| g. It is appropriate to test that the EoW=*Yi* measurements have the same dispersion as the Control=*Zj* measurements by assuming Model 7 is true and applying the AB test. | TRUE FALSE |  |
| h. It is appropriate to test that the EoW=*Yi* have the same distribution as the Control=*Zj* measurements by assuming Model 5 is true and applying the RS test. | TRUE FALSE |  |

2. Plot the data. Think about the design of the study. Which model is more appropriate and why? Circle the more appropriate model. Give the letter of the one most appropriate reason. 6 points.

|  |  |
| --- | --- |
| CIRCLE MORE APPROPRIATE MODEL | GIVE ONE REASON LETTER |
| Model 1 Model 4 |  |
| Model 7 Model 10 |  |
| Model 9 Model 14 |  |

3. Test the hypothesis that the changes in OTM for welders, (end-of-week)-minus-(beginning-of-week) = EoW-BoW, are symmetric about zero. What is the name of the most appropriate test? (Use abbreviations from data page.) What is the number of the model underlying this test? What is the two-sided P-value? What is the name of the associated point estimate of the center of symmetry of the changes? What is the value of the point estimate? What is the value of the 95% confidence interval for the center of symmetry of the changes? Is the null hypothesis of no change plausible? 15 points

Name of test: \_\_\_\_\_\_\_\_\_\_\_\_\_ Model #:\_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of estimate: \_\_\_\_\_\_\_\_\_\_ Value of estimate: \_\_\_\_\_\_\_\_\_ 95% CI: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

No change is: (CIRCLE ONE) PLAUSIBLE NOT PLAUSIBLE

4. Under model 11, use Kendall’s correlation to test that BoW= *Xi* and EoW= *Yi* measurements are independent. What are the values of the estimates of Kendall’s correlation and the probability of concordance? What is the two-sided P-value? Is independence plausible? 10 points.

Kendall’s Correlation: \_\_\_\_\_\_\_\_\_\_\_ Prob(concordant): \_\_\_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_\_\_

Independence is: (CIRCLE ONE) PLAUSIBLE NOT PLAUSIBLE

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2007, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. Use abbreviations from data page.

5. Test the hypothesis that the end of week OTM measurements for welders (EoW) have the same distribution as the OTM measurements for controls against the alternative that the EoW measurements tend to be higher. What is the name of the most appropriate nonparametric test? What is the number of the model underlying this test? What is the two-sided P-value? What would be an appropriate parameter to estimate that is associated with this test? What is the value of the point estimate? Is the null hypothesis of no difference plausible? 15 points

Name of test: \_\_\_\_\_\_\_\_\_\_\_\_\_ Model #:\_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Parameter: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Value of estimate: \_\_\_\_\_\_\_\_\_

No difference is: (CIRCLE ONE) PLAUSIBLE NOT PLAUSIBLE

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6. Test the hypothesis that the beginning of week OTM measurements for welders (BoW) have the same distribution as the OTM measurements for controls against the alternative hypothesis that the BoW measurements have the same distribution as the controls *except* greater dispersion (larger scale). What is the name of the most appropriate nonparametric test? What is the number of the model underlying this test? What is the two-sided P-value? Is the null hypothesis of no difference in dispersion plausible?

10 points.

Name of test: \_\_\_\_\_\_\_\_\_\_\_\_\_ Model #:\_\_\_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

No difference is: (CIRCLE ONE) PLAUSIBLE NOT PLAUSIBLE

7. Use the Kolmogorov-Smirnov test to test whether *Yi* = EoW and *Zi* = Control have the same distribution. What model is assumed when this test is used? What is the two-sided P-value? Is the null hypothesis plausible? Also, give the two-sided P-value comparing *Xi* = BoW and *Zi* = Control from the Kolmogorov-Smirnov test. 10 points.

Model #: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P-value *Yi* vs *Zi*:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P-value *Xi* vs *Zi*:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Plausible that *Yi* and *Zi* have the same distribution: (CIRCLE ONE) Plausible Not Plausible

8. Under model 12, use an appropriate nonparametric procedure from Hollander and Wolfe to test the null hypothesis H0: =2. Give the two-sided P-value and explain very briefly how you did the test. 10 points

P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Briefly how:

Strictly Optional Extra Credit: This question concerns a method we did not discuss, namely the Fligner-Policello test in section 4.4 of Hollander and Wolfe (1999). For extra credit, use this test to compare the medians of *Yi* = EoW and *Zi* = Control. Which model underlies this test? (Give the model # from the data page.) Why is this model better than model #9? (Give a reason letter from the data page.) What is the value of the test statistic (expression 4.53 in H&W). Use Table A.7 to give a two-sided p-value interval for this test (eg, P<0.000001 or P>0.05 or whatever).

Model #: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Reason letter: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistic = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P-value interval: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2007, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. See the data page for abbreviations.

1. Plot the data. Think about the design of the study. Circle “true” if the statement is true for these data, and circle “false” if it is false for these data. Give the letter of the one most appropriate reason. 24 points

|  |  |  |
| --- | --- | --- |
| Statement (Is it true or false? Why? Use Reason Letters from data page.) | CIRCLE ONE | One reason letter |
| a. Model 1 is clearly inappropriate for these data.  after-before on same person look plausibly symmetric | TRUE FALSE | K |
| b. Model 2 is clearly inappropriate for these data.  Controls unrelated – numbering is arbitrary | TRUE FALSE | L |
| c. Model 8 is clearly inappropriate for these data.  Xi and Yi are paired measurements on the same welder | TRUE FALSE | K |
| d. Model 9 is clearly inappropriate for these data.  Boxplots show dispersions are very different | TRUE FALSE | B (or M) |
| e. Under Model 5, the KS test could be used to test whether the EoW=*Yi* measurements have the same distribution as the Control=*Zj* measurements. | TRUE FALSE | P |
| f. Under Model 13, the HLrs estimate could be used to estimate  | TRUE FALSE | M |
| g. It is appropriate to test that the EoW=*Yi* measurements have the same dispersion as the Control=*Zj* measurements by assuming Model 7 is true and applying the AB test. | TRUE FALSE  AB test assumes equal medians – doesn’t look like it | A |
| h. It is appropriate to test that the EoW=*Yi* have the same distribution as the Control=*Zj* measurements by assuming Model 5 is true and applying the RS test. | TRUE FALSE | U |

2. Plot the data. Think about the design of the study. Which model is more appropriate and why? Circle the more appropriate model. Give the letter of the one most appropriate reason. 6 points.

|  |  |
| --- | --- |
| CIRCLE MORE APPROPRIATE MODEL | GIVE ONE REASON LETTER |
| Model 1 Model 4  Xi and Yi are paired measurements on the same welder | K |
| Model 7 Model 10  Xi and Zj seem to have unequal dispersions | Not graded: +2 for everyone |
| Model 9 Model 14  Yi and Zj seem to have unequal dispersions | B |

3. Test the hypothesis that the changes in OTM for welders, (end-of-week)-minus-(beginning-of-week) = EoW-BoW, are symmetric about zero. What is the name of the most appropriate test? (Use abbreviations from data page.) What is the number of the model underlying this test? What is the two-sided P-value? What is the name of the associated point estimate of the center of symmetry of the changes? What is the value of the point estimate? What is the value of the 95% confidence interval for the center of symmetry of the changes? Is the null hypothesis of no change plausible? 15 points

Name of test: SR Model #: 1 P-value: 0.001953

Name of estimate: HLsr Value of estimate: 2.95 95% CI: [2.00, 3.68]

No change is: (CIRCLE ONE) PLAUSIBLE NOT PLAUSIBLE

4. Under model 11, use Kendall’s correlation to test that BoW= *Xi* and EoW= *Yi* measurements are independent. What are the values of the estimates of Kendall’s correlation and the probability of concordance? What is the two-sided P-value? Is independence plausible? 10 points.

Kendall’s Correlation: 0.405 Prob(concordant): 0.702 = (0.405+1)/2 P-value: 0.1035

Independence is: (CIRCLE ONE) PLAUSIBLE NOT PLAUSIBLE

Statistics 501, Spring 2007, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. Use abbreviations from data page.

5. Test the hypothesis that the end of week OTM measurements for welders (EoW) have the same distribution as the OTM measurements for controls against the alternative that the EoW measurements tend to be higher. What is the name of the most appropriate nonparametric test? What is the number of the model underlying this test? What is the two-sided P-value? What would be an appropriate parameter to estimate that is associated with this test? What is the value of the point estimate? Is the null hypothesis of no difference plausible? 15 points Everyone got this question wrong. You can’t use HL to estimate a shift if the distributions are not shifted! You can estimate Pr(Y>Z). I gave credit for HL; it is, nonetheless, wrong.

Name of test: RS Model #: 5, not 9, for Reason B P-value: 1.083 x 10-05

Parameter: Pr(Y>Z) Value of estimate: U/nm = 100/(10x10) = 1 Y’s always bigger than Z’s !

No difference is: (CIRCLE ONE) PLAUSIBLE NOT PLAUSIBLE

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6. Test the hypothesis that the beginning of week OTM measurements for welders (BoW) have the same distribution as the OTM measurements for controls against the alternative hypothesis that the BoW measurements have the same distribution as the controls *except* greater dispersion (larger scale). What is the name of the most appropriate nonparametric test? What is the number of the model underlying this test? What is the two-sided P-value? Is the null hypothesis of no difference in dispersion plausible?

10 points.

Name of test: AB Model #: 7 P-value: 0.02262

No difference is: (CIRCLE ONE) PLAUSIBLE NOT PLAUSIBLE

7. Use the Kolmogorov-Smirnov test to test whether *Yi* = EoW and *Zi* = Control have the same distribution. What model is assumed when this test is used? What is the two-sided P-value? Is the null hypothesis plausible? Also, give the two-sided P-value comparing *Xi* = BoW and *Zi* = Control from the Kolmogorov-Smirnov test. 10 points.

Model #: 5 P-value *Yi* vs *Zi*: 1.083 x 10-05 P-value *Xi* vs *Zi*: 0.40

Plausible that *Yi* and *Zi* have the same distribution: (CIRCLE ONE) Plausible Not Plausible

8. Under model 12, use an appropriate nonparametric procedure from Hollander and Wolfe to test the null hypothesis H0: =2. Give the two-sided P-value and explain very briefly how you did the test. 10 points

P-value: 0.1478 Briefly how: Test zero Kendall’s correlation between Yi – 2Xi and Xi .

Strictly Optional Extra Credit: This question concerns a method we did not discuss, namely the Fligner-Policello test in section 4.4 of Hollander and Wolfe (1999). For extra credit, use this test to compare the medians of *Yi* = EoW and *Zi* = Control. Which model underlies this test? (Give the model # from the data page.) Why is this model better than model #9? (Give a reason letter from the data page.) What is the value of the test statistic (expression 4.53 in H&W). Use Table A.7 to give a two-sided p-value interval for this test (eg, P<0.000001 or P>0.05 or whatever).

Model #: 14 Reason letter: B

Statistic = Infinity – don’t even have to do arithmetic – because U/nm = 1 in question 5.

Two-sided p-value. Table gives Pr(U>=2.770)=0.010, so we double this for a two-sided p-value, obtaining P-value < 0.02.Doing the Spring 2007 Midterm in R

Problem 3:

> boxplot(EoW-BoW)

> wilcox.test(EoW-BoW,conf.int=T)

Wilcoxon signed rank test

data: EoW - BoW

V = 55, p-value = 0.001953

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

2.00 3.68

sample estimates:

(pseudo)median

2.95

Problem 4:

> plot(BoW,EoW)

> cor.test(BoW,EoW,method="kendall")

Kendall's rank correlation tau

data: BoW and EoW

z = 1.6282, p-value = 0.1035

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.4045199

Warning message:

Cannot compute exact p-value with ties in: cor.test.default(BoW, EoW, method = "kendall")

Problem 5:

> boxplot(EoW,Control)

> wilcox.test(EoW,Control)

Wilcoxon rank sum test

data: EoW and Control

W = 100, p-value = 1.083e-05

alternative hypothesis: true mu is not equal to 0

Problem 6:

> boxplot(BoW,Control)

> ansari.test(BoW,Control)

Ansari-Bradley test

data: BoW and Control

AB = 40, p-value = 0.02262

alternative hypothesis: true ratio of scales is not equal to 1

Doing the Spring 2007 Midterm in R, continued

Problem 7:

> ks.test(EoW,Control)

Two-sample Kolmogorov-Smirnov test

data: EoW and Control

D = 1, p-value = 1.083e-05

alternative hypothesis: two.sided

> ks.test(BoW,Control)

Two-sample Kolmogorov-Smirnov test

data: BoW and Control

D = 0.4, p-value = 0.4005

alternative hypothesis: two.sided

Problem 8:

> plot(BoW,EoW-2\*BoW)

> cor.test(EoW-2\*BoW,BoW,method="kendall")

Kendall's rank correlation tau

data: EoW - 2 \* BoW and BoW

z = -1.4473, p-value = 0.1478

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

-0.3595733

Extra credit: You can do this one by eye. However, you could write a general program:

> fp

function(x,y){

#Fligner-Policello test (HW p135)

#x and y are vectors

P<-apply(outer(y,x,"<")+(outer(y,x,"==")/2),2,sum)

Q<-apply(outer(x,y,"<")+(outer(x,y,"==")/2),2,sum)

Pb<-mean(P)

Qb<-mean(Q)

v1<-sum((P-Pb)^2)

v2<-sum((Q-Qb)^2)

U<-(sum(Q)-sum(P))/(2\*sqrt(v1+v2+Pb\*Qb))

pval<-1-pnorm(U)

list(P=P,Q=Q,Pb=Pb,Qb=Qb,v1=v1,v2=v2,U=U,pval=pval)}

> fp(Control,EoW)

$P

[1] 0 0 0 0 0 0 0 0 0 0

$Q

[1] 10 10 10 10 10 10 10 10 10 10

$Pb

[1] 0

$Qb

[1] 10

$v1

[1] 0

$v2

[1] 0

$U

[1] Inf

$pval

[1] 0

Statistics 501, Spring 2007, Final: Data Page #1

This is an exam. Do not discuss it with anyone. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question.

The data are from Tsai, et al. (2001) Betal quid chewing as a risk factor for hepatocellular carcinoma: a case-control study. *British Journal of Cancer*, 84, 709-713. This was a case-control study from Taiwan, comparing cases of hepatocellular carcinoma (liver cancer) to health “controls” who had come for a regular check up, matched for age and gender. Variable H is hepatocellular carcinoma (case of control). You will ignore the matching in this analysis. A popular habit in Taiwan is betal quid chewing (a nut and leaf derived from the areca tree). Variable Q is betal quid chewing (betal-user, nonuser). In addition, it was determined whether each person had antibodies for hepatitis B and hepatitis C; you would have antibodies if you had been infected in the past. Variable B is hepatitis B (HBsAg-, HBsAg+), where + means you have antibodies. Variable C is hepatitis C (antiHCV-, antiHCV+), where + means you have antibodies. Use H, Q, B and C to refer to variables. There were 19 cases of hepatocellular carcinoma who wer betal quid chewers, with no antibodies for hepatitis B or C.

> betal

, , Hepatitis B = HBsAg-, Hepatitis C = antiHCV-

Betal quid

Hepatocelluar Cancer nonuser betal-user

case 19 7

control 180 17

, , Hepatitis B = HBSAg+, Hepatitis C = antiHCV-

Betal quid

Hepatocelluar Cancer nonuser betal-user

case 102 50

control 50 4

, , Hepatitis B = HBsAg-, Hepatitis C = antiHCV+

Betal quid

Hepatocelluar Cancer nonuser betal-user

case 55 11

control 7 1

, , Hepatitis B = HBSAg+, Hepatitis C = antiHCV+

Betal quid

Hepatocelluar Cancer nonuser betal-user

case 16 3

control 4 0

Make and keep a photocopy of your answer page. Place the exam in an envelope with ‘Paul Rosenbaum, Statistics Department’ on it. The exam is due in my office, 473 Huntsman, on Wednesday, May 7 at 12:00am. You may turn in the exam early at my mail box in the Statistics Department, 4th floor, Huntsman. If you would like to receive your graded exam, final grade, and an answer key, then include a stamped, self-addressed, regular envelope. (I will send just two pages, so a regular envelope with regular postage should do it.)

Statistics 501, Spring 2007, Final: Data Page #2

When you compute odds ratios, compute:

(case&user) (control&nonuser)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(case&nonuser) (control&user)

or (7x180)/(19x17) = 3.90 for

Betal quid

Hepatocelluar Cancer nonuser betal-user

case 19 7

control 180 17

Similarly, for hepatitis Q or C, compute:

(user&+) (nonuser&-)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(user&-) (nonuser&+)

As usual, the betal data are in file [Rst501.RData](http://stat.wharton.upenn.edu/statweb/course/Spring-2007/stat501/Rst501.RData) at

<http://stat.wharton.upenn.edu/statweb/course/Spring-2007/stat501/>**Print Name Clearly,** Last**, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Statistics 501, Spring 2007, Final, Answer Page #1**

This is an exam. Do not discuss it with anyone. See the data page. Due May 1 at 11:00am.

1. Consider the model which asserts that the four variables, H, Q, B, and C are independent. What is the likelihood ratio chi square (G2) for testing the goodness of fit of this model? What are the degrees of freedom (DF)? What is the p-value? Based on the goodness of fit test alone: does this model fit the data well?

G2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ DF = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p-value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The model fits the data (CIRCLE ONE): WELL NOT WELL

2. Consider the hierarchical model with all two factor interactions among the four variables, H, Q, B, and C. How many two factor interactions does this model include? What is the likelihood ratio chi square (G2) for testing the goodness of fit of this model? What are the degrees of freedom (DF)? What is the p-value? Based on the goodness of fit test alone: does this model fit the data well?

How many two-factor interactions (give an integer): \_\_\_\_\_\_\_\_\_\_\_\_

G2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ DF = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p-value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The model fits the data (CIRCLE ONE): WELL NOT WELL

3. Consider the hierarchical model with all three factor interactions among the four variables, H, Q, B, and C. How many two factor interactions does this model include? How many three factor interactions does this model include? What is the likelihood ratio chi square (G2) for testing the goodness of fit of this model? What are the degrees of freedom (DF)? What is the p-value? Based on the goodness of fit test alone: does this model fit the data well?

How many 2-factor interactions (give an integer): \_\_\_\_\_\_\_\_ How many 3-factor interactions: \_\_\_\_\_\_\_\_\_\_

G2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ DF = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p-value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The model fits the data (CIRCLE ONE): WELL NOT WELL

4. Return to the hierarchical model in question 2, with all 2 factor interactions. If you could add just one three factor interaction to this model, which three factor interaction would it be? (Obviously, you need to pick the interaction that does the most to improve the fit.) Indicate the interaction by the three letters of the variables in the interaction, as in [X,Y,Z]. Does this new hierarchical model, with all two factor interactions plus your choice of one 3-factor interaction fit well as judged solely by the likelihood ratio goodness of fit chi square? What is the likelihood ratio chi square (G2) for testing the goodness of fit of this model? What are the degrees of freedom (DF)? What is the p-value?

Which 3-factor interaction is best to add? (Give 3 letters): [ , , ]

G2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ DF = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p-value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The model fits the data (CIRCLE ONE): WELL NOT WELL

5. Does the model in question 4 fit better than the model in question 2? Are the models hierarchical? Are the models nested? Test the null hypothesis that the model in question 2 is adequate and that the added model terms (u-terms) for the model in question 4 are zero and hence not needed. Very briefly, in one sentence, explain how to do the test. What is the value of the test statistic? What are its degrees of freedom? What is the p-value? Is null hypothesis plausible?

How do you do the test (one sentence maximum):

Test Statistic = \_\_\_\_\_\_\_\_\_\_\_\_\_ DF = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p-value = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Null hypothesis is (CIRCLE ONE): PLAUSIBLE NOT PLAUSIBLE

**Statistics 501, Spring 2007, Final, Answer Page #2**

6. Give the likehood ratio goodness of fit chi squre, G2, degrees of freedom, and p-value for the following two models. If you *had* to pick *one* of these two models, which one is *more appropriate* for these data and why?

|  |  |  |  |
| --- | --- | --- | --- |
|  | G2 | degrees of freedom | p-value |
| [H, Q] [H, B, C] |  |  |  |
| [H, Q] [Q, C] [H, B, C] |  |  |  |

Which model? CIRCLE ONE [H, Q] [H, B, C] [H, Q] [Q, C] [H, B, C]

In a sentence, why?

|  |  |
| --- | --- |
| 7. STATMENT | CIRCLE ONE |
| Model [H, Q] [H, B, C] implies chewing betal quid, Q, is independent of hepatitis B and C | TRUE FALSE |
| Model [H, Q] [H, B, C] implies hepatitis B and hepatitis C are conditionally independent given hepatocellular carcinoma H | TRUE FALSE |
| Under model [H, Q] [Q, C] [H, B, C], the odd ratio linking H and Q changes with C. | TRUE FALSE |
| Model [H, Q] [H, B, C] implies chewing betel quid Q is conditionally independent of hepatitis B and hepatitis C given hepatocellular carcinoma H | TRUE FALSE |

8. Under Model [H, Q] [Q, C] [H, B, C], give the fitted odds ratio linking hepatocellular carcinoma H with chewing betel quid Q for each combination of hepatitis antibodies. See the data page for the order of computation.

|  |  |  |
| --- | --- | --- |
| Give four H&Q odds ratios | B = + | B = - |
| C = + |  |  |
| C = - |  |  |

In question #9, risk is measured by the odds ratio, and refers to risk of hepatocellular carcinoma H.

|  |  |
| --- | --- |
| 9. STATEMENT | CIRCLE ONE |
| Under model [H, Q] [Q, C] [H, B, C], the increased risk (of H) associated with chewing betel quid Q is much greater for people who have had hepatitis C. | TRUE FALSE |
| Model [H, Q] [Q, C] [H, B, C], implies the risk (of H) associated with hepatitis B is the same whether or not you have antibodies for hepatitis C | TRUE FALSE |
| Under model [H, Q] [Q, C] [H, B, C], odds ratio linking Q and C is the same for all levels of (H,B) and is estimated to be 2.26. | TRUE FALSE |

Have a great summer!

**Statistics 501, Spring 2007, Final, Answer Page #1**

This is an exam. Do not discuss it with anyone. See the data page. Due May 1 at 11:00am.

1. Consider the model which asserts that the four variables, H, Q, B, and C are independent. What is the likelihood ratio chi square (G2) for testing the goodness of fit of this model? What are the degrees of freedom (DF)? What is the p-value? Based on the goodness of fit test alone: does this model fit the data well? 10 points

G2 = 321.2 DF = 11 p-value < 0.0001

It stinks!

The model fits the data (CIRCLE ONE): WELL NOT WELL

2. Consider the hierarchical model with all two factor interactions among the four variables, H, Q, B, and C. How many two factor interactions does this model include? What is the likelihood ratio chi square (G2) for testing the goodness of fit of this model? What are the degrees of freedom (DF)? What is the p-value? Based on the goodness of fit test alone: does this model fit the data well? 10 points

How many two-factor interactions (give an integer): 6

G2 = 19.305 DF = 5 p-value = 0.00169

The model fits the data (CIRCLE ONE): WELL NOT WELL

3. Consider the hierarchical model with all three factor interactions among the four variables, H, Q, B, and C. How many two factor interactions does this model include? How many three factor interactions does this model include? What is the likelihood ratio chi square (G2) for testing the goodness of fit of this model? What are the degrees of freedom (DF)? What is the p-value? Based on the goodness of fit test alone: does this model fit the data well? 10 points

How many 2-factor interactions (give an integer): 6 How many 3-factor interactions: 4

G2 = 0.51 DF = 1 p-value = 0.47

The model fits the data (CIRCLE ONE): WELL NOT WELL

4. Return to the hierarchical model in question 2, with all 2 factor interactions. If you could add just one three factor interaction to this model, which three factor interaction would it be? (Obviously, you need to pick the interaction that does the most to improve the fit.) Indicate the interaction by the three letters of the variables in the interaction, as in [X,Y,Z]. Does this new hierarchical model, with all two factor interactions plus your choice of one 3-factor interaction fit well as judged solely by the likelihood ratio goodness of fit chi square? What is the likelihood ratio chi square (G2) for testing the goodness of fit of this model? What are the degrees of freedom (DF)? What is the p-value? 10 points

Which 3-factor interaction is best to add? (Give 3 letters): [ H, B, C]

G2 = 1.769 DF = 4 p-value = 0.778

The model fits the data (CIRCLE ONE): WELL NOT WELL

5. Does the model in question 4 fit better than the model in question 2? Are the models hierarchical? Are the models nested? Test the null hypothesis that the model in question 2 is adequate and that the added model terms (u-terms) for the model in question 4 are zero and hence not needed. Very briefly, in one sentence, explain how to do the test. What is the value of the test statistic? What are its degrees of freedom? What is the p-value? Is null hypothesis plausible? 10 points

How do you do the test (one sentence maximum): Compare the change in G2 to the chi-square distribution with the change in degrees of freedom as the degrees of freedom

Test Statistic = 19.305-1.769 = 17.536 DF = 5-4 = 1 p-value = 2.819189 x 10-05

Null hypothesis is (CIRCLE ONE): PLAUSIBLE NOT PLAUSIBLE

**Statistics 501, Spring 2007, Final, Answer Page #2**

6. Give the likehood ratio goodness of fit chi squre, G2, degrees of freedom, and p-value for the following two models. If you *had* to pick *one* of these two models, which one is *more appropriate* for these data and why? 21 points

|  |  |  |  |
| --- | --- | --- | --- |
|  | G2 | degrees of freedom | p-value |
| [H, Q] [H, B, C] | 8.814 | 6 | 0.184 |
| [H, Q] [Q, C] [H, B, C] | 1.820 | 5 | 0.873 |

Which model? CIRCLE ONE [H, Q] [H, B, C] [H, Q] [Q, C] [H, B, C]

In a sentence, why? The improvement in fit is significant, 8.814-1.820 = 6.994 on 6-5 = 1 df, p=0.0082.

|  |  |
| --- | --- |
| 7. STATEMENT 12 points | CIRCLE ONE |
| Model [H, Q] [H, B, C] implies chewing betal quid, Q, is independent of hepatitis B and C | TRUE FALSE |
| Model [H, Q] [H, B, C] implies hepatitis B and hepatitis C are conditionally independent given hepatocellular carcinoma H | TRUE FALSE |
| Under model [H, Q] [Q, C] [H, B, C], the odd ratio linking H and Q changes with C. | TRUE FALSE |
| Model [H, Q] [H, B, C] implies chewing betel quid Q is conditionally independent of hepatitis B and hepatitis C given hepatocellular carcinoma H | TRUE FALSE |

8. Under Model [H, Q] [Q, C] [H, B, C], give the fitted odds ratio linking hepatocellular carcinoma H with chewing betel quid Q for each combination of hepatitis antibodies. See the data page for the order of computation. 8 points

|  |  |  |
| --- | --- | --- |
| Give four H&Q odds ratios | B = + | B = - |
| C = + | 4.95 | 4.95 |
| C = - | 4.95 | 4.95 |

In question #9, risk is measured by the odds ratio, and refers to risk of hepatocellular carcinoma H.

|  |  |
| --- | --- |
| 9. STATEMENT 9 points | CIRCLE ONE |
| Under model [H, Q] [Q, C] [H, B, C], the increased risk (of H) associated with chewing betel quid Q is much greater for people who have had hepatitis C. | TRUE FALSE |
| Model [H, Q] [Q, C] [H, B, C], implies the risk (of H) associated with hepatitis B is the same whether or not you have antibodies for hepatitis C | TRUE FALSE |
| Under model [H, Q] [Q, C] [H, B, C], odds ratio linking Q and C is the same for all levels of (H,B) and is estimated to be 2.26. | TRUE FALSE |

Doing the Problem Set in R

> quick

function(tab,mod){

o<-loglin(tab,mod)

list(g2=o$lrt,df=o$df,pval=1-pchisq(o$lrt,o$df))}

> quick(betal,list(1,2,3,4))

2 iterations: deviation 1.136868e-13

$g2

[1] 321.2392

$df

[1] 11

$pval

[1] 0

> quick(betal,list(c(1,2),c(1,3),c(1,4),c(2,3),c(2,4),c(3,4)))

10 iterations: deviation 0.08435694

$g2

[1] 19.30488

$df

[1] 5

$pval

[1] 0.001686281

> quick(betal,list(c(1,2,3),c(1,2,4),c(1,3,4),c(2,3,4)))

4 iterations: deviation 0.09403469

$g2

[1] 0.5119507

$df

[1] 1

$pval

[1] 0.4742956

> quick(betal,list(c(1,2),c(1,3),c(1,4),c(2,3),c(2,4),c(3,4),c(1,2,3)))

10 iterations: deviation 0.05667761

$g2

[1] 18.59602

$df

[1] 4

$pval

[1] 0.0009433637

> quick(betal,list(c(1,2),c(1,3),c(1,4),c(2,3),c(2,4),c(3,4),c(1,2,4)))

9 iterations: deviation 0.08713973

$g2

[1] 18.83938

$df

[1] 4

$pval

[1] 0.0008451545

> quick(betal,list(c(1,2),c(2,3),c(2,4),c(1,3,4)))

8 iterations: deviation 0.09538882

$g2

[1] 1.769219

$df

[1] 4

$pval

[1] 0.7781088

> quick(betal,list(c(1,2),c(1,3,4)))

2 iterations: deviation 1.421085e-14

$g2

[1] 8.814025

$df

[1] 6

$pval

[1] 0.1843105

> quick(betal,list(c(1,2),c(2,4),c(1,3,4)))

5 iterations: deviation 0.02971972

$g2

[1] 1.819888

$df

[1] 5

$pval

[1] 0.8734632

> fit<-loglin(betal,list(c(1,2),c(2,4),c(1,3,4)),fit=T)$fit

5 iterations: deviation 0.02971972

> 1/or(fit[,,1,1])

[1] 4.954774

> 1/or(fit[,,2,1])

[1] 4.954774

> 1/or(fit[,,1,2])

[1] 4.954774

> 1/or(fit[,,2,2])

[1] 4.954774

> 1/(or(fit[1,,1,]))

[1] 2.255539

> 1/(or(fit[1,,2,]))

[1] 2.255539

> 1/(or(fit[2,,1,]))

[1] 2.255539

> 1/(or(fit[2,,2,]))

[1] 2.255539

**Statistics 501, Spring 2006, Midterm: Data Page #1**

This is an exam. Do not discuss it with anyone. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question.

The data are from the following paper, Stanley, M., Viriglio, J., and Gershon, S. (1982) “Tritiated imipramine binding sites are decreased in the frontal cortex of suicides,” *Science*, 216, 1337-1339. There is no need to examine the paper unless you wish to do so. It is available at the library web page via JSTOR.

Imipramine is a drug often used to treat depression. Stanley, et al. obtained brain tissue from the New York City Medical examiners office for nine suicides and for nine age-matched controls who died from other causes. Data for the 9 pairs appears below. They measured imipramine binding (Bmax in fmole per milligram of protein) in samples from the Brodmann’s areas 8 and 9 of the frontal cortex, where high values of Bmax indicate greater binding with imipramine. The data appear below, where SBmax and CBmax are Bmax for Suicide and matched Control, SDtoA and CDtoA are minutes between death and autopsy, and Scause and Ccause are the cause of death. Although Stanley, et al. are interested in imipramine binding as it relates to depression and suicide, they need to rule out other explanations, such as differences in time to autopsy or cause of death. Notice that there were no suicides by myocardial infarction (MI), and no controls who died by hanging or jumping, but some suicides shot themselves and some controls where shot by someone else.

> imipramine

pair SBmax CBmax SDtoA CDtoA Scause Ccause

1 1 464 740 1920 1650 hanging gunshot

2 2 249 707 1140 1190 gunshot gunshot

3 3 345 353 555 750 hanging MI

4 4 328 350 1560 1570 gunshot gunshot

5 5 285 350 1020 880 gunshot MI

6 6 237 531 990 550 hanging auto

7 7 443 1017 2250 1440 hanging gunshot

8 8 136 695 1140 1200 jump MI

9 9 483 544 1320 1455 hanging MI

> i<-imipramine

> wilcox.test(i$CBmax,i$SBmax,conf.int=T)

Wilcoxon rank sum test with continuity correction

data: i$CBmax and i$SBmax

W = 72, p-value = 0.006167

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

64.99999 454.99997

sample estimates:

difference in location

238.0443

Warning message: cannot compute exact p-value and exact confidence intervals with ties in: wilcox.test.default(i$CBmax, i$SBmax)

> wilcox.test(i$CBmax-i$SBmax,conf.int=T)

Wilcoxon signed rank test

data: i$CBmax - i$SBmax

V = 45, p-value = 0.003906

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

41.5 458.0

sample estimates:

(pseudo)median

276

STATISTICS 501, SPRING 2006, MIDTERM DATA PAGE #2

Model 1: Zi = i where i ~ iid, with a continuous distribution symmetric about 0, i=1,…,n.

Model 2: Y1,…,Yn ~ iid with a continuous distribution, X1,…,Xm ~ iid with a continuous distribution, with the Y’s and X’s independent of each other.

Model 3: Y1,…,Yn ~ iid with a continuous distribution with median , X1,…,Xm ~ iid with a continuous distribution with median , with the Y’s and X’s independent of each other, and (Yj-) having the same distribution as (Xi - ) for each i,j, for some >0.

Model 4: Y1,…,Yn ~ iid with a continuous distribution, X1,…,Xm ~ iid with a continuous distribution, with the Y’s and X’s independent of each other, and Yj having the same distribution as Xi +  for each i,j.

Model 5: (X1,Y1), …, (Xn,Yn) are n iid observations from a continuous bivariate distribution.

Model 6: Yi =  + xi + ei, …, where the ei are n iid observations from a continuous distribution with median zero independent of the xi which are untied and fixed.

Model 7: Y1,…,Yn ~ iid with a continuous distribution, X1,…,Xm ~ iid with a continuous distribution, with the Y’s and X’s independent of each other, and Yj having the same distribution as  + Xi for each i,j for some >0.

Test abbreviations:

SR = Wilcoxon’s signed rank test (3.1). HLsr = Hodges-Lehmann estimate associated with Wilcoxon’s signed rank test (3.2). RS = Wilcoxon’s rank sum test (4.1). HLrs = Hodges-Lehmann estimate associated with Wilcoxon’s rank sum test (4.2). AB = Ansari-Bradley test (5.1). LE = Lepage’s test (5.3). KS = Kolmogorov-Smirnov test (5.4). KW = Kruskal-Wallis test (6.1). OA = Jonckheere-Terpstra test for ordered alternatives (6.2). KE = Kendall’s test (8.1). TH = Theil’s test for a specified slope (9.1), THe = Theil’s estimate.

A “best” test should have the correct level when the null hypothesis is true (i.e., it should give P<=0.05 at most 5% of the time when the null hypothesis is true), it should be consistent against the stated alternative hypotheses (i.e., it should be nearly certain to give a P<0.05 if the alternative hypothesis is true and the sample size is very, very large). If more than one test has the correct level and is consistent, the best test will be targeted against the stated alternative, and will have greater power against it.

Be careful in reading and writing to distinguish distinct procedures with similar names. For instance, “the Wilcoxon test” or “the Hodges-Lehmann estimate” are not uniquely defined terms, since more than one procedure can be so defined, so “the Wilcoxon test” is a vague and hence incorrect answer to any question asking for a specific test. You may use the test abbreviations given above, but write them clearly.

The data are available in files:

JMP-IN: imipramine.JMP

MS-Excel: imipramine.xls

and as object imipramine in Rst501.Rdata

at: <http://www-stat.wharton.upenn.edu/>

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2006, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. See the data page.

1. In the Imipramine data on the data page, what is the best test of the null hypothesis of no difference between Bmax for suicides and matched controls against the alternative that the typical level of Bmax is different. What is the name of the test? (Be precise; use given abbreviations on data page.) What is the numerical value of the test statistic (*as it is defined in Hollander and Wolfe*)? What is the two-sided significance level? Is the null hypothesis of no difference plausible? Which model on the data page underlies this test? (Give the model #.)

Name of test: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Numerical value of test statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Significance level: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ H0 is (circle one): Plausible Not Plausible

Model #: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. For the correct procedure in question 1, estimate the magnitude of the shift in level of Bmax, suicides vs matched controls. What is the name of the procedure? (Be precise; use given abbreviations on data page.) What is the numerical value of the point estimate? What is the 95% confidence interval? For the estimate and confidence interval, which model on the data page underlies this test? (Give the model #.)

Name of procedure: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Point estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

95% Confidence Interval: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Model #: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. Setting aside the data on suicides, and setting aside the one control who died from an auto accident, test the null hypothesis that the level of Bmax for the four controls who died from gunshot wounds does not differ in level from the level of Bmax for the four controls who died from MI. What is the name of the test? (Be precise; use given abbreviations on data page.) What is the value of the test statistic (*as it is defined in Hollander and Wolfe*)? What is the two sided significance level? Is the null hypothesis plausible? What model underlies *the null hypothesis* for this test? (Give the model #.)

Name of test: \_\_\_\_\_\_\_\_\_\_\_\_\_ Value of test statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_ Significance level: \_\_\_\_\_\_\_\_\_

Null hypothesis is: (circle one) Plausible Not plausible Model for test: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

4. Consider the test you performed in question 3 with four gunshot deaths compared to four MI’s. Question 4 asks whether the sample size is adequate to yield reasonable power. Suppose you do a two-sided 0.05 level test. Suppose the difference in the population were quite large; specifically, 90% of the time in the population, Bmax is larger for MI’s than for gunshot deaths. Fifty percent power is a low level of power; when H0 is false, you reject only half the time. Would the test you did in question 3 have 50% power to detect the supposed 90% difference? What sample size would you need for 50% power? Assume equal numbers of MI’s and gunshot deaths, and use the approximation in Hollander and Wolfe. Briefly indicate the formula and calculations you used.

Does (4,4) sample size yield 50% power? (Circle one) YES NO

What sample sizes would be needed for 50% power? \_\_\_\_\_\_ MI’s + \_\_\_\_\_ Gunshots = \_\_\_\_\_\_\_ total

Briefly indicate computation of needed sample size in form (abstract formula) = (formula with numbers)

|  |  |  |
| --- | --- | --- |
| Computation of sample size for 50% power | Abstract symbolic formula | Formula with needed numbers in place of abstractions. |
| Sample size = |  | = |

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2006, Midterm, Answer Page #2

5. Use Kendall’s rank correlation to test the null hypothesis that the difference in Bmax (i.e.,

Y=SBmax-CBmax) is independent of the difference in time from death to autopsy (i.e., X=SDtoA-CDtoA). Give rank correlation, the estimated probability of concordance, and the two-sided significance level. Is the null hypothesis plausible? Which model underlies the test (given the model # from the data page)?

Rank correlation: \_\_\_\_\_\_\_\_\_\_ Probability of concordance: \_\_\_\_\_\_\_\_ Significance level: \_\_\_\_\_\_\_\_\_\_\_\_\_

Which model #? \_\_\_\_\_\_\_\_\_\_\_ Null hypothesis is (circle one) Plausible Not plausible

6. Continuing question #5, under model #6, with use a nonparametric procedure to test the null hypothesis

H0: = 1.0. What is the name of the test? (Use the abbreviations on the data page.) What is the value of the test statistic (as defined in Hollander and Wolfe)? What is the two-sided significance level? Is the null hypothesis plausible?

Name of test: \_\_\_\_\_\_\_\_\_\_\_\_\_ Value of statistic: \_\_\_\_\_\_\_\_\_\_\_ Significance level: \_\_\_\_\_\_\_\_\_\_\_\_

Null hypothesis is: (circle one) Plausible Not plausible

­­­­­­­­­­­­­­­­­­­­­

|  |  |
| --- | --- |
| 7. Here, “best procedure” means the most appropriate procedure from the list of options. | CIRCLE ONE BEST PROCEDURE  (Use the test abbreviations on the data page) |
| Given n iid continuous differences Yi-Xi: What is the best test of the null hypothesis that Yi-Xi is symmetrically distributed about zero against the alternative that Yi-Xi is symmetrically distributed about some nonzero quantity? | SR HLsr RS HLrs AB LE KS KW OA KE TH THe |
| Under model 2: What is the best test of the null hypothesis that X and Y have the same distribution against the alternative that Prob(Y>X) is not equal to ½ ? | SR HLsr RS HLrs AB LE KS KW OA KE TH THe |
| Under model 7, what is the best test of the null hypothesis H0: =0, =1 against the alternative that H0 is not true. | SR HLsr RS HLrs AB LE KS KW OA KE TH THe |
| Under model 2: What is the best test of the null hypothesis that X and Y have the same distribution against the alternative that the distributions are different. | SR HLsr RS HLrs AB LE KS KW OA KE TH THe |
| Best estimate of  under Model 4. | SR HLsr RS HLrs AB LE KS KW OA KE TH THe |
| Under model 2: What is the best test of the null hypothesis that X and Y have the same distribution with median  against the alternative that Prob(X>Y>) + Prob(Y>X>) is not ¼. | SR HLsr RS HLrs AB LE KS KW OA KE TH THe |

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2006, Midterm, Answer Page #1

This is an exam. Do not discuss it with anyone. See the data page.

1. In the Imipramine data on the data page, what is the best test of the null hypothesis of no difference between Bmax for suicides and matched controls against the alternative that the typical level of Bmax is different. What is the name of the test? (Be precise; use given abbreviations on data page.) What is the numerical value of the test statistic (*as it is defined in Hollander and Wolfe*)? What is the two-sided significance level? Is the null hypothesis of no difference plausible? Which model on the data page underlies this test? (Give the model #.) 12 points

Name of test: SR = Wilcoxon’s signed rank Numerical value of test statistic: 45

Significance level: 0.0039 H0 is (circle one): Plausible Not Plausible

Model #: 1

2. For the correct procedure in question 1, estimate the magnitude of the shift in level of Bmax, suicides vs matched controls. What is the name of the procedure? (Be precise; use given abbreviations on data page.) What is the numerical value of the point estimate? What is the 95% confidence interval? For the estimate and confidence interval, which model on the data page underlies this test? (Give the model #.) 14 points

Name of procedure: HLsr = Hodges-Lehmann for Wilcoxon’s signed rank Point estimate: 276 lower for suicides

95% Confidence Interval: [41.5, 458.0] Model #: 1

3. Setting aside the data on suicides, and setting aside the one control who died from an auto accident, test the null hypothesis that the level of Bmax for the four controls who died from gunshot wounds does not differ in level from the level of Bmax for the four controls who died from MI. What is the name of the test? (Be precise; use given abbreviations on data page.) What is the value of the test statistic (*as it is defined in Hollander and Wolfe*)? What is the two sided significance level? Is the null hypothesis plausible? What model underlies this test? (Give the model #.) 14 points

Name of test: RS = Wilcoxon’s rank sum Value of test statistic: 22.5 for gunshot Significance level: 0.24

or 13.5 for MI

Null hypothesis is: (circle one) Plausible Not plausible Model for test: 2

4. Consider the test you performed in question 3 with four gunshot deaths compared to four MI’s. Question 4 asks whether the sample size is adequate to yield reasonable power. Suppose you do a two-sided 0.05 level test. Suppose the difference in the population were quite large; specifically, 90% of the time in the population, Bmax is larger for MI’s than for gunshot deaths. Fifty percent power is a low level of power; when H0 is false, you reject only half the time. Would the test you did in question 3 have 50% power to detect the supposed 90% difference? What sample size would you need for 50% power? Assume equal numbers of MI’s and gunshot deaths, and use the approximation in Hollander and Wolfe. Briefly indicate the formula and calculations you used. 14 points

Does (4,4) sample size yield 50% power? (Circle one) YES NO

What sample sizes would be needed for 50% power? 4 MI’s + 4 Gunshots = 8 total

Briefly indicate computation of needed sample size in form (abstract formula) = (formula with numbers)

|  |  |  |
| --- | --- | --- |
| Computation of sample size for 50% power | Abstract symbolic formula | Formula with needed numbers in place of abstractions. |
| Sample size = 8.003 | (z0.025 + z0.5)2  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  12 (c) (1- c) (– 0.5)2 | (1.96 + 0)2  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  12 (½) (1- ½) (0.9 – 0.5)2 |

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2006, Midterm, Answer Page #2

5. Use Kendall’s rank correlation to test the null hypothesis that the difference in Bmax (i.e.,

Y=SBmax-CBmax) is independent of the difference in time from death to autopsy (i.e., X=SDtoA-CDtoA). Give rank correlation, the estimated probability of concordance, and the two-sided significance level. Is the null hypothesis plausible? Which model underlies the test (given the model # from the data page)?

14 points

Rank correlation: -0.44 Probability of concordance: 0.28 Significance level: 0.12

Which model #? 5 Null hypothesis is (circle one) Plausible Not plausible

6. Continuing question #5, under model #6, with use a nonparametric procedure to test the null hypothesis

H0: = 1.0. What is the name of the test? (Use the abbreviations on the data page.) What is the value of the test statistic (as defined in Hollander and Wolfe)? What is the two-sided significance level? Is the null hypothesis plausible? 14 points

Name of test: TH = Theil’s test Value of statistic: -26 (i.e., 9.4 on page 416 in H&W) Significance level: 0.0059

Null hypothesis is: (circle one) Plausible Not plausible

­­­­­­­­­­­­­­­­­­­­­

18 points, 3 each.

|  |  |
| --- | --- |
| 7. Here, “best procedure” means the most appropriate procedure from the list of options. | CIRCLE ONE BEST PROCEDURE  (Use the test abbreviations on the data page) |
| Given n iid continuous differences Yi-Xi: What is the best test of the null hypothesis that Yi-Xi is symmetrically distributed about zero against the alternative that Yi-Xi is symmetrically distributed about some nonzero quantity? | SR |
| Under model 2: What is the best test of the null hypothesis that X and Y have the same distribution against the alternative that Prob(Y>X) is not equal to ½ ? | RS |
| Under model 7, what is the best test of the null hypothesis H0: =0, =1 against the alternative that H0 is not true. | LE |
| Under model 2: What is the best test of the null hypothesis that X and Y have the same distribution against the alternative that the distributions are different. | KS |
| Best estimate of  under Model 4. | HLrs |
| Under model 2: What is the best test of the null hypothesis that X and Y have the same distribution with median  against the alternative that Prob(X>Y>) + Prob(Y>X>) is not ¼. | AB |

>wilcox.test(i$SBmax[i$Ccause=="gunshot"],i$SBmax[i$Ccause=="MI"],conf.int=T)

Wilcoxon rank sum test

data: i$SBmax[i$Ccause == "gunshot"] and i$SBmax[i$Ccause == "MI"]

W = 9, p-value = 0.8857

alternative hypothesis: true mu is not equal to 0

95 percent confidence interval:

-234 328

sample estimates:

difference in location

70.5

> 9/16

[1] 0.5625

> (1.96^2)/3

[1] 1.280533

> ((1.96^2)/3)/((.9-.5)^2)

[1] 8.003333

> cor.test(i$SBmax-i$CBmax,i$SDtoA-i$CDtoA,method="kendall")

Kendall's rank correlation tau

data: i$SBmax - i$CBmax and i$SDtoA - i$CDtoA

T = 10, p-value = 0.1194

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

-0.4444444

> cor.test((i$SBmax-i$CBmax)-(i$SDtoA-i$CDtoA),i$SDtoA-i$CDtoA,method="kendall")

Kendall's rank correlation tau

data: (i$SBmax - i$CBmax) - (i$SDtoA - i$CDtoA) and i$SDtoA - i$CDtoA

T = 5, p-value = 0.005886

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

-0.7222222 = -26/choose(9,2)

Statistics 501, Spring 2006, Final: Data Page #1

This is an exam. Do not discuss it with anyone. If you discuss the exam in any way with anyone, then you have cheated on the exam. The University often expels students caught cheating on exams. Turn in only the answer page. Write answers in the spaces provided: brief answers suffice. If a question asks you to circle the correct answer, then you are correct if you circle the correct answer and incorrect if you circle the incorrect answer. If instead of circling an answer, you cross out an answer, then you are incorrect no matter which answer you cross out. Answer every part of every question.

The data are from a paper by E. Donnell and J. Mason (2006) “Predicting the frequency of median barrier crashes on Pennsylvania interstate highways,” *Accident Analysis and Prevention,* 38, 590-599. It is available via the library web page, but there is no need to consult the paper unless you want to. Some parts of some highways have a “median barrier,” which is a solid barrier separating the left lane heading in one direction from the left lane heading in the other. The median barrier is intended to prevent head-on collisions, in which cars traveling in opposite directions hit each other, but of course this is accomplished by hitting the median barrier instead. Table 1 counts crashes on the Interstate Highway System in Pennsylvania from 1994 to 1998. Crashes were classified by whether there was or was not a fatality (S=SEVERITY). The Interstate Highway System was divided into two parts (R=ROAD): the Interstate-designated portion of the Pennsylvania Turnpike (470 miles) and the remainder of the Interstate Highway System in Pennsylvania (2090 miles). The crashes were also classified into whether the accident involved a collision with a median barrier (T=TYPE). For instance, there were 31 fatal crashes involving the median barrier on the part of the Interstate that does not include the Turnpike. The Turnpike is older than most of the rest of the Interstate in Pennsylvania, and the distance to the barrier is shorter. On the Turnpike the barrier offset is 4 feet or less, and on 16% of the turnpike it is 2 feet or less. In contrast, on 62% of the rest of the Interstate, the offset is 5 feet or more. In this problem, “interstate” refers to “interstate highways other than the turnpike.” Questions 1 to 5 refer to Table 1 and its analysis.

Table 1:observed Frequencies

====================

ROAD$ TYPE$ | SEVERITY$

| fatal nonfatal

----------+---------+-------------------------

interstate barrier | 31 4385

other | 381 25857

+

turnpike barrier | 26 2832

other | 60 6207

**---------- ---------+-------------------------\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**[R] [S] [T] LR ChiSquare 1254.6037 df 4 Probability 0.00000**

**[RS] [T] LR ChiSquare 1244.8218 df 3 Probability 0.00000**

**[RT] [S]** **LR ChiSquare 28.7015 df 3 Probability 0.00000**

**[ST] [R] LR ChiSquare 1236.9130 df 3 Probability 0.00000**

**[RS] [TS] LR ChiSquare 1227.1311 df 2 Probability 0.00000**

**[RT] [TS] LR ChiSquare 11.0109 df 2 Probability 0.00406**

**[RT] [RS] LR ChiSquare 18.9196 df 2 Probability 0.00008**

**[RT] [RS] [TS] LR ChiSquare 5.0613 df 1 Probability 0.02447**

USE THE NOTATION ABOVE TO REFER TO MODELS, FOR INSTANCE [RT][RS]

**Fitted Values from Model [RT] [RS] [TS]**

===============

ROAD$ TYPE$ | SEVERITY$

| fatal nonfatal

---------+---------+-------------------------

interstat barrier | 38.320 4377.680

other | 373.680 25864.320

+

turnpike barrier | 18.680 2839.320

other | 67.320 6199.680

**-------------------+-------------------------**

Statistics 501, Spring 2006, Final: Data Page #2

**Below is systat output for data from a Veteran’s Administration randomized trial for inoperable lung cancer, as described by Kalbfleisch and Prentice (1980) *Statistical Analysis of Failure Time Data*, NY: Wiley, appendix 1. The outcome is SURV100 or survival for 100 days, 1=yes or 0=no. There are three predictors, age in years (not given here), a binary variable “RX” distinguishing the new chemotherapy (RX=1) from the standard chemotherapy (RX=0), and whether the patient had received a previous chemotherapy (PRIORRX=1) or not (PRIORRX=0). The table is just descriptive. The model is log{Pr(Survive)/Pr(Die)} = 0 + 1 RX +2 PRIORRX +3 AGE.**

Observed Frequencies

====================

PRIORRX RX | SURV100

| 0=no 1=yes

-------+---------+-------------------------

0=no 0=standard| 22.000 25.000

1=new | 34.000 13.000

+

1=yes 0=standard| 11.000 9.000

1=new | 11.000 8.000

**-------------------+-------------------------**

Categorical values encountered during processing are:

SURV100 (2 levels) 0, 1

Binary LOGIT Analysis.

Dependent variable: SURV100

Input records: 137

Records for analysis: 133

Records deleted for missing data: 4

Sample split

Category choices

0 (REFERENCE) 78

1 (RESPONSE) 55

Total : 133

Log Likelihood: -87.406

Parameter Estimate S.E. t-ratio p-value

1 CONSTANT 0.704 1.025 0.687 0.492

2 RX -0.772 0.362 -2.135 0.033

3 PRIORRX 0.100 0.394 0.255 0.799

4 AGE -0.012 0.017 -0.721 0.471

95.0 % bounds

Parameter Odds Ratio Upper Lower

2 RX 0.462 0.939 0.228

3 PRIORRX 1.106 2.394 0.511

4 AGE 0.988 1.021 0.955

You do not need the data to do the final; however, the data are available. The crash data is in systat and excel formats as PAbarrier and in the Rworkspace for stat501, namely Rst501.RData. The VA data is in systat and excel formats in VAlungLogit. <http://www-stat.wharton.upenn.edu/>

Make and keep a photocopy of your answer page. Place the exam in an envelope with ‘Paul Rosenbaum, Statistics Department’ on it. The exam is due in my office, 473 Huntsman, on Wednesday, May 3 at noon. You may turn in the exam early at my mail box in the Statistics Department, 4th floor, Huntsman. If you would like to receive your graded exam, final grade, and an answer key, then include a stamped, self-addressed, regular envelope. (I will send just two pages, so a regular envelope with regular postage should do it.)Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2006, Final, Answer Page #1

This is an exam. Do not discuss it with anyone. See the data page. Due May 3 at noon.

1. For each claim, fill in the appropriate model (in the form [RT][RS] or whatever), give the goodness of fit p-value for that model, and state whether the claim is plausible.

|  |  |  |  |
| --- | --- | --- | --- |
| Claim | Model | Goodness of fit p-value | Claim is: (Circle One) |
| The road predicts the severity of injury only indirectly through their separate relationships with crash type. |  |  | Plausible  Not Plausible |
| Road and crash type are related, but injury severity is just luck, unrelated to road and crash type. |  |  | Plausible  Not Plausible |
| Although the road is related to the crash type, and both road and crash type are related to injury severity, barrier crashes are related to injury severity in the same way on both road groups. |  |  | Plausible  Not Plausible |
| Fatal accidents are a relatively larger fraction of all accidents (fatal and nonfatal together) on the turnpike than on the interstate, but that’s just because barrier crashes are more common on the turnpike: if you compare crashes of the same type, there is no association between road and injury severity. |  |  | Plausible  Not Plausible |

2. Test the null hypothesis that the addition of [RT] to the model [RS][TS] is not needed. Give the *value* of the test statistic, the *degrees of freedom*, the *p-value*, and state whether there is *strong evidence* that [RT] should be added to the model. *Explain briefly* how the test statistic is computed.

CIRCLE ONE

Value: \_\_\_\_\_\_\_\_\_\_\_ Degrees of Freedom: \_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_ Strong-Evidence Not-Strong

Explain briefly:

3. What is the simplest model that fits well? Test that your candidate model fits significantly better than the model that is as similar as possible but simpler. What is the simpler model? Give the *value* of the test statistic, the *degrees of freedom*, the *p-value*.

Simplest model that fits well: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Just simpler model that doesn’t fit well: \_\_\_\_\_\_\_\_\_\_\_\_

Value: \_\_\_\_\_\_\_\_\_\_\_ Degrees of Freedom: \_\_\_\_\_\_\_\_ P-value: \_\_\_\_\_\_\_\_

4. Continuing question 3, if the simplest model that fits well were true, would the odds ratio linking crash type and injury severity be the same on the turnpike and the other interstate highways? Use the fitted counts from the simplest model that fits well to estimate the two odds ratios just mentioned.

CIRCLE ONE

Odds ratios would be: The same Not the same

|  |  |  |
| --- | --- | --- |
| Compute the odds ratios from the fitted counts for the simplest model that fits well. | Interstate | Turnpike |
| Estimated odds ratio linking barrier crashes with fatal injury. |  |  |

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2006, Final, Answer Page #2

|  |  |
| --- | --- |
| 5. Use the simplest model that fits well to answer these questions. Here, “likely” refers to odds or odds ratios. | CIRCLE ONE |
| Most crashes result in fatalities. | TRUE FALSE |
| Barrier crashes are somewhat less than half as likely as other crashes to be associated with fatal injuries on the interstate, but that is not true on the turnpike. | TRUE FALSE |
| Barrier crashes are a minority of all crashes, but they are not equally likely on the turnpike and the interstate. Whether you look at fatal crashes or nonfatal ones, the odds of a barrier crash are somewhat greater on the turnpike. | TRUE FALSE |
| The odds ratio linking barrier crashes with fatal injury is higher on the turnpike than on the interstate. | TRUE FALSE |

In the VA Lung Cancer Trial, what is the estimate of the coefficient 3 of AGE? Is the null hypothesis, H0:3 =0 plausible? What is the p-value? Is there clear evidence that patient AGE predicts survival for 100 days? If age were expressed in months rather than years, would the numerical value of 3. The logit model has no interaction between AGE and PRIORRX. Does that mean that the model assumes AGE and prior treatment (PRIORRX) are independent? Does it mean that the model assumes AGE and prior treatment (PRIORRX) are conditionally independent given SURV100 and RX?

CIRCLE ONE

Estimate of 3: \_\_\_\_\_\_\_ p-value: \_\_\_\_\_\_\_\_\_ H0 is PLAUSIBLE NOT PLAUSIBLE

Clear evidence that Age predicts survival for 100 days: YES NO

AGE and PRIORRX assumed independent: TRUE FALSE

AGE and PRIORRX assumed conditionally independent

given SURV100 and RX: TRUE FALSE

7. In the VA Lung Cancer Trial, what is the estimate of the coefficient 1 of RX? Is the null hypothesis, H0:1 =0 plausible? What is the p-value? Is the new treatment better than, perhaps no different from, or worse than the standard treatment if your goal is to survive 100 days? Looking at the point estimate: Is the new treatment, when compared with the standard treatment, associated with a doubling, a halving or no change in your odds of surviving 100 days?

CIRCLE ONE

Estimate of 1: \_\_\_\_\_\_\_ p-value: \_\_\_\_\_\_\_\_\_ H0 is PLAUSIBLE NOT PLAUSIBLE

New treatment is: BETTER PERHAPS NO DIFFERENT WORSE

Odds of survival for 100 days are: DOUBLED HALVED NO CHANGE

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2006, Final, Answer Page #1

This is an exam. Do not discuss it with anyone. See the data page. Due May 3 at noon.

1. For each claim, fill in the appropriate model (in the form [RT][RS] or whatever), give the goodness of fit p-value for that model, and state whether the claim is plausible. 15 points

|  |  |  |  |
| --- | --- | --- | --- |
| Claim | Model | Goodness of fit p-value | Claim is: (Circle One) |
| The road predicts the severity of injury only indirectly through their separate relationships with crash type. | [RT][TS] | 0.00406 | Plausible  Not Plausible |
| Road and crash type are related, but injury severity is just luck, unrelated to road and crash type. | [S][RT] | 0.0000+ | Plausible  Not Plausible |
| Although the road is related to the crash type, and both road and crash type are related to injury severity, barrier crashes are related to injury severity in the same way on both road groups. | [RT][RS][TS] | 0.025 | Plausible  Not Plausible |
| Fatal accidents are a relatively larger fraction of all accidents (fatal and nonfatal together) on the turnpike than on the interstate, but that’s just because barrier crashes are more common on the turnpike: if you compare crashes of the same type, there is no association between road and injury severity. | [RT][TS] | 0.00406 | Plausible  Not Plausible |

2. Test the null hypothesis that the addition of [RT] to the model [RS][TS] is not needed. Give the *value* of the test statistic, the *degrees of freedom*, the *p-value*, and state whether there is *strong evidence* that [RT] should be added to the model. *Explain briefly* how the test statistic is computed. 15 points

CIRCLE ONE

Value: 1222.1 Degrees of Freedom: 1 P-value: <0.0001 Strong-Evidence Not-Strong

Explain briefly:

Change in likelihood ratio chi square 1227.1 – 5.1 = 1222.1 on 2-1=1 degrees of freecom.

3. What is the simplest model that fits well? Test that your candidate model fits significantly better than the model that is as similar as possible but simpler. What is the simpler model? Give the *value* of the test statistic, the *degrees of freedom*, the *p-value*. 15 points

The saturated model always fits perfectly, with G2=0, 0 df, and fitted counts equal to observed counts!

Simplest model that fits well: [RST] Just simpler model that doesn’t fit well: [RT][RS][TS]

Value: 5.06 Degrees of Freedom: 1 P-value: 0.025

4. Continuing question 3, if the simplest model that fits well were true, would the odds ratio linking crash type and injury severity be the same on the turnpike and the other interstate highways? Use the fitted counts from the simplest model that fits well to estimate the two odds ratios just mentioned. 10 points

CIRCLE ONE

Odds ratios would be: The same Not the same

|  |  |  |
| --- | --- | --- |
| Compute the odds ratios from the fitted counts for the simplest model that fits well. | Interstate | Turnpike |
| Estimated odds ratio linking barrier crashes with fatal injury. | 0.48 or about 1/2 | 0.95 or about 1 |

Print Name Clearly, Last, First: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics 501, Spring 2006, Final, Answer Page #2

|  |  |
| --- | --- |
| 5. Use the simplest model that fits well to answer these questions. Here, “likely” refers to odds or odds ratios. 15 points | CIRCLE ONE |
| Most crashes result in fatalities. | TRUE FALSE |
| Barrier crashes are somewhat less than half as likely as other crashes to be associated with fatal injuries on the interstate, but that is not true on the turnpike. | TRUE FALSE |
| Barrier crashes are a minority of all crashes, but they are not equally likely on the turnpike and the interstate. Whether you look at fatal crashes or nonfatal ones, the odds of a barrier crash are somewhat greater on the turnpike. | TRUE FALSE |
| The odds ratio linking barrier crashes with fatal injury is higher on the turnpike than on the interstate. | TRUE FALSE |

In the VA Lung Cancer Trial, what is the estimate of the coefficient 3 of AGE? Is the null hypothesis, H0:3 =0 plausible? What is the p-value? Is there clear evidence that patient AGE predicts survival for 100 days? The logit model has no interaction between AGE and PRIORRX. Does that mean that the model assumes AGE and prior treatment (PRIORRX) are independent? Does it mean that the model assumes AGE and prior treatment (PRIORRX) are conditionally independent given SURV100 and RX? 15 points

CIRCLE ONE

Estimate of 3: -0.012 p-value: 0.471 H0 is PLAUSIBLE NOT PLAUSIBLE

Clear evidence that Age predicts survival for 100 days: YES NO

AGE and PRIORRX assumed independent: TRUE FALSE

AGE and PRIORRX assumed conditionally independent

given SURV100 and RX: TRUE FALSE

The logit model makes no assumptions about the relationships among the predictors!

7. In the VA Lung Cancer Trial, what is the estimate of the coefficient 1 of RX? Is the null hypothesis, H0:1 =0 plausible? What is the p-value? Is the new treatment better than, perhaps no different from, or worse than the standard treatment if your goal is to survive 100 days? Looking at the point estimate: Is the new treatment, when compared with the standard treatment, associated with a doubling, a halving or no change in your odds of surviving 100 days? 15 points

CIRCLE ONE

Estimate of 1: -0.772 p-value: 0.033 H0 is PLAUSIBLE NOT PLAUSIBLE

New treatment is: BETTER PERHAPS NO DIFFERENT WORSE

Odds of survival for 100 days are: DOUBLED HALVED NO CHANGE

Some useful articles (available from the Library Web Page)

* **The Analysis of Repeated Measures: A Practical Review with Examples**
* Author(s): B. S. Everitt
* Source: Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 44, No. 1 (1995), pp. 113-135
* Published by: [Blackwell Publishing](http://www.jstor.org/action/showPublisher?publisherCode=black) for the [Royal Statistical Society](http://www.jstor.org/action/showPublisher?publisherCode=rss)
* Stable URL: <http://www.jstor.org/stable/2348622>

### Abstract

Repeated measures data, in which the same response variable is recorded on each observational unit on several different occasions, occur frequently in many different disciplines. Many methods of analysis have been suggested including $t$-tests at each separate time point and multivariate analysis of variance. In this paper the application of a number of methods is discussed and illustrated on a variety of data sets. The approach involving the calculation of a small number of relevant summary statistics is considered to have advantages in many circumstances.\_kw Compound Symmetry

* **328. Note: The Use of Non-Parametric Methods in the Statistical Analysis of the Two-Period Change-Over Design**
* Author(s): Gary G. Koch
* Source: Biometrics, Vol. 28, No. 2 (Jun., 1972), pp. 577-584
* Published by: [International Biometric Society](http://www.jstor.org/action/showPublisher?publisherCode=ibs)
* Stable URL: <http://www.jstor.org/stable/2556170>

### Abstract

The two-period change-over design is often used in clinical trials in which subjects serve as their own controls. This paper is concerned with the statistical analysis of data arising from such subjects when assumptions like variance homogeneity and normality do not necessarily apply. Test procedures for hypotheses concerning direct effects and residual effects of treatments and period effects are formulated in terms of Wilcoxon statistics as calculated on appropriate within subject linear functions of the observations. Thus they may be readily applied to small sample-data.

* **A Distribution-Free Test for Related Correlation Coefficients**
* Author(s): Douglas A. Wolfe
* Source: Technometrics, Vol. 19, No. 4 (Nov., 1977), pp. 507-509
* Published by: [American Statistical Association](http://www.jstor.org/action/showPublisher?publisherCode=astata) and [American Society for Quality](http://www.jstor.org/action/showPublisher?publisherCode=asq)
* Stable URL: http://www.jstor.org/stable/1267893

### Abstract

Let (X1,X2,X3) be a continuous, trivariate random vector for which it is of interest to compare the correlation between X2 and X1 with that between X3 and X1. This problem is important, for example, when both X2 and X3 are potential linear predictors for X1 or, more generally, whenever the variable (X1,X2,X3) represents a triplet of measurements on the same individual and correlation comparisons are desired. In this note it is shown that an exact distribution-free test for such problems can be based on a single Kendall correlation coefficient.

* **The Analysis of Multidimensional Contingency Tables**
* Author(s): Stephen E. Fienberg
* Source: Ecology, Vol. 51, No. 3 (May, 1970), pp. 419-433
* Published by: [Ecological Society of America](http://www.jstor.org/action/showPublisher?publisherCode=esa)
* Stable URL: <http://www.jstor.org/stable/1935377>

### Abstract

Ecological data often come in the form of multidimensional tables of counts, referred to as contingency tables. During the last decade several new methods of analyzing such tables have been proposed. Here, a class of models analogous to those used in the analysis of variance is discussed, and a method for computing the expected cell counts for the different models is presented. Two differenet tests for checking the goodness-of-fit of a particular model are then examined. The first is the simple generalization of the Pearson chi-square test statistic, while the second is referred to as the likelihood-ratio chi-square test statistic. Both have the same asympototic @g^2 distribution. The likelihood-ratio statistic can be used in the selection of a suitable model, via the technique of partitioning. All of the methods presented are illustrated using data collected by Schoener on lizards from the West Indies.