Review for Exam II

Administrative Items

Midterm Exam #2

Tomorrow Tuesday April 4, from 6-8 p.m. in Annenberg 110.

Getting help

– See me Monday 3-5:30 or tomorrow after 2:30.
– Send me an e-mail with your question. (stine@wharton)
– Visit the StatLab/TAs, *particularly for help using the computer.*

Preparing for the exam

– Review old exams (no prediction intervals, few "calculation" formulas)
– Review assignments, emphasis on material discussed in class.
– Sample regression problems from text are 44, 46, 47a:
Interpret slope and intercept, use the equation to predict,
Know the relationship to correlation, answer if slope is ≠ 0.

Tests Derived from Chi Square

Keys to chi-square

– All use the formula

$$\chi^2 = \sum \frac{(obs - \exp ected)^2}{\exp ected}$$

– Expected counts depend on the null hypothesis being tested.
– Degrees of freedom = (#rows – 1)(#cols – 1) or (#cells – 1) if only one row.

Testing for independence

– Fill in the cells using the marginal totals, computing the expected counts so that the expected counts are "independent." (p 403)

Testing for goodness of fit

– Fill in the cells using the conjectured relationship. (p 401)

Inference

Given the assumptions, reject H_0 (independence or the claim of some distribution for the counts) if χ^2 is *larger* than the tabled value with the computed number of degrees of freedom.

Expected counts

Make sure you know how to compute the expected counts and the chi-square test statistic.

Measures of the Strength of a Relationship in a Table

Predictive index λ

– Using one variables predictor and the other as response (as in regression) λ measures the reduction in the number of errors made if you predict the response as the most likely category. (p 411-412)

$$\lambda = \frac{(\#\,errors\,without\,predictor) - (\#\,errors\,with\,predictor)}{(\#\,errors\,without\,predictor)}$$

– The number of errors without the predictor is the sum of all but the largest marginal cell for the response.

– The number of errors with the predictor is the sum of all but the largest cell for the response, *added up for every column*.

Odds ratio

– Odds ratio is just the ratio of the odds of an event, typically computed for two columns or rows of a table (p 414).

– Inference is done (under same assumptions as χ^2) by finding the confidence interval for the log of the odds ratio (p 415).

One-Way Analysis of Variance

Objective

Are the averages of these groups different?

Design and assumptions

– Randomization is key.

– Usual 3 others: independence, constant variance, and normality.

Are the means of the groups significantly different?

Test via the overall F-ratio appearing in the Anova summary, with the computations as described in the text (p 428).

What to do if the data are not normal?

Use the nonparametric version of the usual one-way anova, known as the Wilcoxon or Kruskal-Wallis test. The test statistic is a χ^2 statistic, with the calculations described on page 431.

Is the largest (smallest) mean significantly different from the others?

Answer using Hsu's multiple comparison procedure. You will always use JMP for these, since the text lacks the needed critical values.

Which pairs of means are significantly different?

Answer using Tukey's multiple comparison procedure, with the formula given by

$$(difference\,in\,means) \pm q_{\alpha}(number\,of\,means, error\,df)\sqrt{\hat{\sigma}^2 / num\,per\,group}$$

with q_α from Table 8 and the mean squared error estimate (sigma-hat squared) from the mean square for error or mean squared error in the anova table.

Two-Way Analysis of Variance

Assumptions

Same as in one-way anova.

Beyond one-way anova: goals

– Explore interactions between two factors.
– Find the combination of factors that yields the best response.
– Reduce background variation so that effects become more clear (i.e., make the error smaller so that the mean differences become more apparent).

Interaction is *the* special feature

– Interaction implies that the differences in the means defined by one factor depend on the value of the other factor.
– Test for presence of interaction by considering the F-ratio for interaction (factor1 * factor2) in the anova summary.
– Use caution when interpreting the effects of the two factors when interaction is present, since their effects are not consistent.

Profile plot

– The profile plot plots the averages of the cells defined by the two-way layout, joining the averages in one row, say, of the table (p 446).
– The profile plot indicates interaction by a lack of parallel lines joining the cell averages.

Multiple comparisons

You can still use Tukey's method, noting that the generic formula offered above still works. Note that the text formula (p 450) has an error (there should not be a "+" sign between the means on the right – it ought to be the difference).

Linear Regression with One Predictor

Objective

– Predict the response Y using a predictor X.
– Understand how changes in the values of the predictor affect the response.

What do the slope and intercept mean?

The intercept is the predicted value when the predictor is zero and has the units of the response. The slope describes how changes in the predictor convert to changes in the response, on average, and has units (units of Y)/(units of X).

Using the equation to predict a new value

Simple: substitute for the value of the predictor in the fitted equation.

Assumptions

In addition to the usual three (independent, constant variance, normality), regression adds the assumption that one has fit the right predictor in the right equation (i.e., gotten the right transformation - such as log).

Relationship to correlation

Correlation provides a measure of the "strength" of the relationship between the predictor and response, in the form of

$$R^2 = \frac{\textit{Fitted Variation}}{\textit{Total Variation}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

The predicted values ("y – hat") are just the values from the fitted regression line, namely

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

The value of R^2 is the squared correlation between X and Y.

Testing for a significant relationship

To see if the fit is significant, test whether the slope differs significantly from zero. One way to do this is to use a confidence interval for the slope, using a formula like that used for the CI for a mean:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} SE(\hat{\beta}_1)$$

where the SE for the slope estimate is roughly computed as

$$SE(\hat{\beta}_1) \approx \frac{\sigma}{\sqrt{n}} \times \frac{1}{SD(X)}$$

If the confidence interval does not include zero, then the population slope differs from zero with confidence α. (Alternatively, just count how many SEs separate the fitted slope from zero. If the distance – the t-statistic or t-ratio – exceeds $t_{\alpha/2,n-2} \approx 2$, then conclude the slope is significantly different from zero.)

Next Class on Wednesday

Simple regression in finance

The term " beta" in finance refers to a regression coefficient. On Weds, we'll take a look at what makes this coefficient so interesting.