Statistical Tests

Review

Confidence intervals

– General form of 95% interval is (estimate) ± 2 (SE of estimate)

– X-bar ± 2 s/√n interval includes values within 2 SEs of X-bar.

Key assumptions

– Independent observations (random sample from population).

– Equal precision for all measurements (constant σ)

– Normality

Implication of SE & CI on sample design

– Use of the ± 2 SE in length implies how to set size for survey.

– Margin for error in opinion polls.

New Application for Today

Making a choice

– Comparison of status quo to an alternative

– Different costs for choices

– Legal analogy: presume innocent versus guilty

– Asymmetry of the procedure in favor of status quo.

Definitions and Concepts

Blackboard example: pain relief

– Old method scored 40 points on pain relief scale, new averages 50.
Is the new method really better – in the population?

– Assuming the old one is better, would you get sample results as far from the old method's results *by chance alone*?
Can we "explain away" this difference as simply chance variation.

Hypothesis testing

– Statements about some feature of the population (like μ)

– Null hypothesis (presume status quo, "innocent")

– Alternative hypothesis (change from status quo, "guilty")

– Statistical assessment of the size of the difference

– Same assumptions as required for confidence intervals

Types of errors that can be made

– 2 by 2 table, with "Type I" and "Type II" errors.

– Falsely rejecting the null or status quo (unnecessary change)

– Failing to reject the null (missed opportunity to improve)

– Testing controls Type I error rate, but often ignores Type II.

Statistical basis for test: counting standard errors

– How to measure the "distance" of the observed statistic (e.g. sample average) from the hypothesis value (e.g. μ)

– Two or more standard errors is "a long way". Far enough to reject H_0.

– Number of SEs away from H0 = t-statistic or t-ratio.

p-value

– Computed as a probability under the null hypothesis
In context of blackboard example

– Interpretation: Plausibility of H_0 given information in sample.

– Common rule: "reject" null if p-value < 0.05. (more than 2 SE)

– Serves to enhance the confidence interval
Inside/outside the interval versus "How far outside?"

Summary of procedure

– Assume null hypothesis (i.e., presume innocent)
– Gather data, compute test statistic and p-value *given assumptions*
– Is p-value small (e.g., p-value < 0.05, or more than 2 SE away)?
– Make choice, weighing importance of two errors
    I. Incorrectly dropping the status quo.
    II. Failing to adopt new method when in fact better.

Assumptions for two-sample t-test

– independence

– equal variance <– we can avoid this one in some ways

– normality <– may not always be appropriate

Complication: multiple comparisons

– What happens if you do lots and lots of tests?

– Have a 5% chance for error on each… So the chance for an error somewhere along the way becomes large.

– Tukey-Kramer method protects from finding false positives among many alternatives.

One or two-sided test

– Ignore this distinction and only use two-sided tests.
– FDA only allows two-sided testing.

New Examples for Today

Context

– Do the two samples come from the same population?

– How might the populations differ?
  • mean
  • variance

– Improved tests allow you to compare the means, even if the variances might differ from each other.

– So-called nonparametric tests avoid the assumption of normality.

Traffic Pulse

Is targeted marking necessary? Do men react differently to a new information service, Traffic Pulse, than women?

– Initial comparison of the two groups using linked histograms.

– Side-by-side comparison using boxplots to facilitate comparison.

– Picking the null hypothesis

– Role for assumptions

– Two-sample t-test and comparison via confidence intervals.

– Find no significant different, little reason to reject null that men and women react similarly to Traffic Pulse.

Do employees from different fields react differently to Traffic Pulse?

– More than two things to compare, as now have 10 occupations.

– Might we find some difference by chance alone?

– Multiple comparisons procedure via the Tukey-Kramer method.

– Obtain using the Fit Model command.

Selecting a painting process (p 131)

The *target* thickness of an automatically applied primer is 1.2 mils. Which application method applies the paint at the right thickness?

– Hypothesis test of new method compared to old favors the installed method at the expense of missing some better new processes.
  • i.e. the test is a "conservative" procedure.

– Comparison of both to target (p 136)
  • CI for method *b* contains target.

– Comparison of the two to each other: Which has smaller error? (p 138)
  • Compare absolute deviation from the target value.
  • Confidence intervals are distinct, so significant difference.

– Two-sample comparison of means also significant via t-test or a confidence interval for the difference in means:
  • If the confidence interval for the difference in means μ_a-μ_b holds zero, then zero is a plausible value for the difference.

– Checking assumptions: (p 139)
  Are the variances in two groups comparable?

Re-engineering a food processing line (supplemental, p 141)

Is the new change-over method faster than the old method?

– Comparison based on the average speed of change-over swap.

– Comparison by groups using two intervals (p145) misses some differences that are indeed significant. This simple comparison is inferior to direct comparison of mean values using the t-test.

– Both t-test and confidence interval for the difference indicate a significant difference in population mean values. (p 146)

– Use of confidence interval to obtain cost differential (p 147)

Analysis of time for service calls (p 148)

Which service method is faster for handling the typical call?

– Impact of skewness (as in credit card data):
  • Two-sample t-test finds no significant difference.

– Transformed to normal does find significant difference
  • Transform to log scale
  • Use a nonparametric method that avoids the assumption of normal and is not so influenced by outliers (p 152).

– Are such transformations appropriate, particularly when dealing with $.

Paired Tests: Exploiting Dependence

Sales force comparison (PharmSal.jmp, p163)

Which sales force should be kept in merged pharmaceutical firm?

(generally not a statistical issue in any merger I have seen)

– Comparison without pairing finds no difference nothing. (p 164)

– More subtle form of pairing
Paired by region, the data are clearly very *dependent* (p 165)

– With pairing, a clear difference emerges (p 166).

– Easier to do this test by subtraction, converting to a one-sample test with differences within pairs. (p 168)

–Again, always check the assumptions before acting on the results of a statistical test. Make sure that the differences are caused by a deviation from the null hypothesis, not a deviation from the assumptions.