Nonparametric tests, Bootstrapping

http://www.isrec.isb-sib.ch/~darlene/EMBnet/

Hypothesis testing review

- 2 'competing theories' regarding a population parameter:
  - *NULL* hypothesis *H* ('straw man')
  - *ALTERNATIVE* hypothesis *A* ('claim', or theory you wish to test)
- *H:* NO DIFFERENCE
  - any observed deviation from what we expect to see is due to *chance variability*
- *A:* THE DIFFERENCE IS *REAL*

Test statistic

- Measure how far the observed data are from what is expected *assuming the NULL H* by computing the value of a *test statistic* (TS) from the data
- The particular TS computed depends on the parameter
- For example, to test the population mean, the TS is the *sample mean* (or standardized sample mean)

Testing a population mean

- We have already learned how to test the
  mean
  of a population for a variable with a
  normal distribution
  when the sample size is
  small
  and the population
  SD is unknown
- *What test is this??*

t-test assumption of normality

- The *t*-test was developed for samples that have *normally distributed* values
- This is an example of a *parametric test* – a (parametric) form of the distribution is assumed (here, a normal distribution)
- The *t*-test is fairly robust against departures from normality if the sample size is not too small
- *BUT* if the values are extremely non-normal, it might be better to use a procedure which does not make this assumption

Nonparametric hypothesis tests

- *Nonparametric* (or *distribution-free*) hypothesis tests do not make assumptions about the *form* of the distribution of the data values
- These tests are usually based on the *ranks* of the values, rather than the actual values themselves
- There are nonparametric analogues of many parametric test procedures

One-sample Wilcoxon test

- Nonparametric alternative to the *t*-test
- Tests value of the center of a distribution
- Based on sum of the (positive or negative) ranks of the differences between observed and expected center
- Test statistic corresponds to selecting each number from 1 to n with probability $\frac{1}{2}$ and calculating the sum
- In R: `wilcox.test()`

Two-sample Wilcoxon test

- Nonparametric alternative to the 2-sample *t*-test
- Tests for differences in location (center) of 2 distributions
- Based on replacing the data values by their ranks (without regard to grouping) and calculating the sum of the ranks in a group
- Corresponds to sampling n_1 values without replacement from 1 to $n_1 + n_2$
- In R: `wilcox.test()`

Matched-pairs Wilcoxon

- Nonparametric alternative to the paired *t*-test
- Analogous to paired *t*-test, same as one-sample Wilcoxon but on the *differences* between paired values
- In R: **`wilcox.test()`**

ANOVA and the Kruskal-Wallis test

- Nonparametric alternative to one-way ANOVA
- Mechanics similar to 2-sample Wilcoxon test
- Based on between group sum of squares calculated from the average ranks
- In R: **`kruskal.test()`**

Issues in nonparametric testing

- Some (mistakenly) assume that using a nonparametric test means that you don't make any assumptions at all
- THIS IS **NOT** TRUE!!
- In fact, there is really only one assumption that you are relaxing, and that is of the *form* that the distribution of sample values takes
- A major reason that nonparametric tests are avoided if possible is their relative *lack of power* compared to (appropriate) parametric tests

Parameter estimation

- Have an unknown *population parameter* of interest
- Want to use a sample to make a guess (*estimate*) for the value of the parameter
- *Point estimation* : Choose a *single value* (a 'point') to estimate the parameter value
- Methods of point estimation include: ML, MOM, Least squares, Bayesian methods...
- *(Confidence) Interval estimation* : Use the data to find a *range of values* (an interval) that seems likely to contain the true parameter value

CI mechanics

- When the CLT applies, a CI for the *population mean* looks like

  sample mean +/- $z^* \sigma/\sqrt{n}$,

  where *z* is a number from the *standard normal* chosen so the *confidence level* is a specified size (*e.g.* 95%, 90%, *etc.*)
- For small samples from a normal distribution, use CI based on *t*-distribution

  sample mean +/- $t^* s/\sqrt{n}$

Example

- To set a standard for what is to be considered a 'normal' calcium reading, a random sample of 100 apparently healthy adults is obtained. A blood sample is drawn from each adult. The variable studied is X = number of mg of calcium per dl of blood.
  - sample mean = 9.5
  - sample SD = 0.5
- Find an approximate 95% CI for the (population) average number of mg of calcium per dl of blood ...

Russian dolls analogy*

- Père Noël dolls … Outermost is 'doll 0', next is 'doll 1', *etc.*
- We are *not allowed to observe doll 0*, which represents the *population* in a sampling scheme)
- Want to estimate some characteristic of doll 0 (*e.g.* number of points on the beard)
- *Key assumption* : the relationship (*e.g.* ratio) between dolls 1 and 2 *is the same* as that between dolls 0 and 1

* from *The Bootstrap and Edgeworth Expansion*, by Peter Hall, Springer 1992

From dolls to statistics

- Say you want to estimate some function of a *population* distribution – e.g. the population mean
- It makes sense, when possible, to use the same function of the *sample* distribution
- We can do this same thing for many other types of functions
- A common example is that we might wish to obtain the *sampling distribution* of an estimator in order to make a CI, say, in cases where large sample approximations might not hold

An idea

- Where exact calculations are difficult to obtain, they may be approximated by *resampling* from the observed distribution of sample values
- That is, *pretend* that the sample is the 'population'
- The *bootstrap procedure* is to draw some number (*R*) of samples *with replacement* from the 'bootstrap population' (*i.e.* the original sample values)
- You need a computer to do this!

Bootstrap procedure

- For each *bootstrap sample*, compute the value of the desired statistic
- At the end, you will have *R* values of the statistic
- You can use standard data summary procedures to summarize or explore the distribution of the statistic (histogram, QQ plot, compute the mean, SD, *etc.*)
- For example, to make a bootstrap CI for the sample mean based on the normal distribution, you could use the bootstrap SD (instead of the sample SD) ...

Versions of the bootstrap

- *Nonparametric Bootstrap* : as just described, draw bootstrap samples from the original data
- *Parametric Bootstrap* : assume that your original data came from some *particular distribution* (for example, a normal distribution, or exponential, *etc.*)
- In this case, samples are *simulated* from that *assumed distribution*
- Distribution parameters (for example, the mean and SD for the normal) are *estimated from the original sample*

R: bootstrap demo

- You will have some practice with this in the TP
- Let's go to the demo ...