Nonparametric tests

Adapted from Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Nonparametric one and two-sample tests

If data do not come from a normal population (and if the sample is not large), we cannot use a t-test. One useful approach to creating test statistics is through the use of *rank statistics*.

Resampling methods provide alternative approaches for testing simple hypotheses and obtaining confidence intervals. For example, the t approach can be used with a permutation test to test $H_0 : \mu_1 = \mu_2$ versus any of the alternatives, *regardless of whether the data are normal or not*. This is covered in Section 16.9 (pp. 712–716) and available in `proc multtest` in SAS; R packages `coin` and `perm` conduct permutation tests too.

Sign test for one population

The sign test assumes the data come from from a continuous distribution with model

$$Y_i = \eta + \epsilon_i, \ i = 1, \ldots, n.$$

- η is the population median and ϵ_i follow an unknown, continuous distribution.
- Want to test $H_0 : \eta = \eta_0$ where η_0 is known versus one of the three common alternatives: $H_a : \eta < \eta_0$, $H_a : \eta \neq \eta_0$, or $H_a : \eta > \eta_0$.

Sign test

- Test statistic is $B^* = \sum_{i=1}^{n} I\{y_i > \eta_0\}$, the number of $y_1, \ldots, y_n$ larger than η_0.
- Under H_0, $B^* \sim \text{bin}(n, 0.5)$.
- Reject H_0 if B^* is "unusual" relative to this binomial distribution.

Question: How would you form a "large sample" test statistic from B^*? You would not need to do that here, but this is common with more complex test statistics with non-standard distributions.

Example: eye relief data

- Data are time in minutes that a drug takes to relieve $n = 20$ irritated eyes, measured redness.
- Rao (1998) page 178.
- `proc univariate` gives the sign test (and the Wilcoxon signed-rank test), but for a two-sided alternative. How do we get the p-value for a one-sided alternative (i.e., $H_a : \eta < 5$)

Data:

0.4	4.6	2.2	1.2	4.5	5.7	8.0	2.1	4.8	3.0
8.8	11.4	1.3	1.4	2.1	1.3	12.5	2.4	4.6	2.8

SAS code

```
data relief;
input time @@;
datalines;
 0.4  4.6  2.2  1.2  4.5  5.7  8.0  2.1  4.8  3.0
 8.8 11.4  1.3  1.4  2.1  1.3 12.5  2.4  4.6  2.8
;
proc univariate plot data=relief mu0=5; * hypothesized value is 5 minutes;
var time;
run;
```

Wilcoxon signed rank test

- Again, test $H_0 : \eta = \eta_0$. However, this method *assumes a symmetric pdf* around the median η.
- Test statistic built from *ranks* of $\{|y_1 - \eta_0|, |y_2 - \eta_0|, \ldots, |y_n - \eta_0|\}$, denoted $R_1, \ldots, R_n$.
- The signed rank for observation i is

$$R_i^+ = \left\{ \begin{array}{ll} R_i & y_i > \eta_0 \\ 0 & y_i \leq \eta_0 \end{array} \right\}.$$

- The "signed rank" statistic is $W^+ = \sum_{i=1}^n R_i^+$.
- If W^+ is large, this is evidence that $\eta > \eta_0$.
- If W^+ is small, this is evidence that $\eta < \eta_0$.

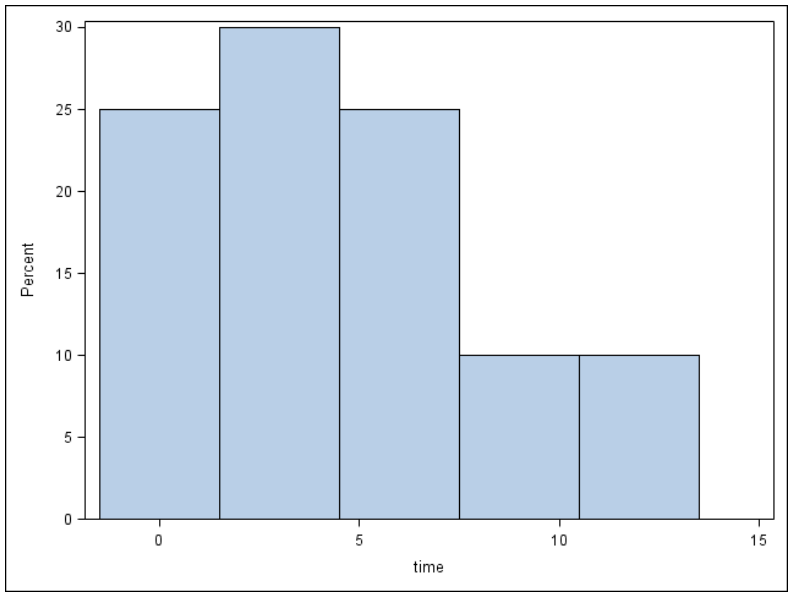
Discussion

- Both the sign test and the signed-rank test can be used with paired data (e.g. we could test whether the median difference is zero).
- **When to use what?** Use t-test when data are approximately normal, or in large sample sizes. Use sign test when data are highly skewed, multimodal, etc. Use signed rank test when data are approximately symmetric but non-normal (e.g. heavy or light-tailed, multimodal yet symmetric, etc.)

Note: The sign test and signed-rank test are more flexible than the t-test because they require less strict assumptions, but the t-test has more power when the data are approximately normal.

Eye relief

```
proc sgplot data=relief;
 histogram time;
```



Which of the three tests (t-test, sign, signed rank) is most appropriate? How do the p-values differ for the latter two?

Mann-Whitney-Wilcoxin test (pp. 795–796)

The Mann-Whitney test assumes

$$Y_{11}, \ldots Y_{1n_1} \overset{iid}{\sim} F_1 \text{ independent } Y_{21}, \ldots, Y_{2n_2} \overset{iid}{\sim} F_2,$$

where F_1 is the cdf of data from the first group and F_2 is the cdf of data from the second group. The null hypothesis is $H_0 : F_1 = F_2$, i.e. that the distributions of data in the two groups are identical.

The alternative is commonly taken to be $H_1 : F_1 \neq F_2$.
One-sided tests can also be performed.

Inference for location shift Δ

Although the test statistic is built assuming $F_1 = F_2$, the alternative is often taken to be that the population *medians* are unequal. This is a fine way to report the results of the test.

Additionally, a CI will give a plausible range for Δ in the shift model $F_2(x) = F_1(x - \Delta)$. Δ can be the difference in medians or means.

An aside...

The test is *consistent* for $H_0 : P(X < Y) = 0.5$ versus $H_a : P(X < Y) \neq 0.5$ where $X \sim F_1$ independent of $Y \sim F_2$. Consistent means the probability of rejecting goes to one if H_a is true. This alternative implies $H_1 : F_1 \neq F_2$ but not vice-versa. When using this form of the test, there are essentially no assumptions on either F_1 or F_2.

If one rather assumes the model $F_2(x) = F_1(x - \Delta)$, then the test reduces to $H_0 : \Delta = 0$ versus $H_a : \Delta \neq 0$. Under this scenario the test statistic can be inverted to provide a CI for Δ, which is the mean or median difference. One must assume that the distributions have the same overall *shape* but not location. The confidence interval for Δ is called the "Hodges-Lehmann" estimate.

Building the test statistic

The Mann-Whitney test is intuitive. The data are

$$y_{11}, y_{12}, \ldots, y_{1n_1} \quad \text{and} \quad y_{21}, y_{22}, \ldots, y_{2n_2}.$$

For each observation j in the first group count the number of observations in the second group c_j that are smaller; ties result in adding 0.5 to the count.
Assuming H_0 is true, on average half the observations in group 2 would be above Y_{1j} and half would be below if they come from the same distribution. That is $E(c_j) = 0.5n_2$.
The sum of these guys is $U = \sum_{j=1}^{n_1} c_j$ and has mean $E(U) = 0.5n_1n_2$. The variance is a bit harder to derive, but is $\text{Var}(U) = n_1n_2(n_1 + n_2 + 1)/12$.

Large sample inference

Something akin to the CLT tells us

$$Z_0 = \frac{U - E(U)}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \dot{\sim} N(0, 1),$$

when H_0 is true. Seeing a U far away from what we expect under the null gives evidence that H_0 is false; U is then standardized as usual (subtract off then mean we expect under the null and standardize by an estimate of the standard deviation of U).

A p-value can be computed as usual as well as a CI.

Note: This test essentially boils down to replacing the observed values with their ranks and carrying out a simple pooled t-test!

proc npar1way

Gives *five* different nonparametric two-sample tests, including Mann-Whitney-Wilcoxon.

```
*******************************
* Mann-Whitney-Wilcoxon
*******************************;
proc npar1way data=teaching hl; * hl adds Hodges-Lehmann confidence interval for delta;
 class attend; var rating; run;

*******************************
* permutation test (if interested)
*******************************;
proc multtest data=teaching permutation nsample=10000;
 test mean(rating);
 class attend; run;
```

proc npar1way output

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable rating
Classified by Variable attend

attend	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Attended	63	4128.50	3496.50	165.227042	65.531746
NotAtten	47	1976.50	2608.50	165.227042	42.053191

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic	1976.5000
Normal Approximation	
Z	-3.8220
One-Sided Pr < Z	<.0001
Two-Sided Pr > |Z|	0.0001
t Approximation	
One-Sided Pr < Z	0.0001
Two-Sided Pr > |Z|	0.0002

Z includes a continuity correction of 0.5.

Hodges-Lehmann Estimation

Location Shift	-0.5000		
95% Confidence Limits		Interval Midpoint	Asymptotic Standard Error
-0.7000	-0.2000	-0.4500	0.1276