Section 5.11: Randomization Tests

• R.A. Fisher introduced the idea of using random permutations of the data themselves as the null distribution for a variety of tests.

• This idea is purely distribution-free and is fairly easy to perform (at least approximately) with today's computing power.

• This type of test is especially appropriate when the data do not represent a sample from some large population, but rather represent the entire population.

Randomization Test with Two Independent Samples

• This is another way to test the hypotheses of the Mann-Whitney Test: $H_0: E(X) = E(Y)$ vs.

$H_1: E(X) \neq E(Y)$ or $H_1: E(X) < E(Y)$ or $H_1: E(X) > E(Y)$

• There are two mutually independent random samples, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$. We assume the data are at least interval in measurement scale.

• Note that if the two samples both have the same distribution, any n of the <u>total</u> $n + m$ observations might as well serve as the first sample.

• **So we could consider all possible selections of n values from the $n + m$ observations in the combined sample.**

• **If the number of these possible selections is very large, we could repeatedly pick n values at random from the $n + m$ observations, many times.**

• **The test statistic is the sum of the n values in the first (X) sample:**

$$T_1 = \sum_{i=1}^{n} X_i$$

• **By considering all (or very many) ways of selecting the n values and calculating T_1 each time, we obtain the null distribution of T_1.**

• **If our observed T_1 is very "unusual" or "extreme" relative to this null distribution, we would reject H_0.**

• **The P-value is defined differently depending on H_1:**

Two-tailed: $2\left[\min\left\{\hat{P}(T_1 \geq T_1^{obs}), \hat{P}(T_1 \leq T_1^{obs})\right\}\right]$

Lower-tailed: $\hat{P}(T_1 \leq T_1^{obs})$

Upper-tailed: $\hat{P}(T_1 \geq T_1^{obs})$

where $\hat{P}$ is a ~~sample~~ empirical proportion based on the random permutations.

• **Typically it is easiest to implement this method with R.**

Example: Random samples of games in which an American League team played <u>with</u> a designated hitter and <u>without</u> a DH were taken. Is there evidence that the mean number of runs scored by the team is greater in the games with the DH? (Use $\alpha = 0.05$.)

X = with DH, Y = without DH

$H_0: E(X) \leq E(Y)$ vs. $H_1: E(X) > E(Y)$

$T_1^{obs} = \sum X_i = 179$

P-value: $\hat{P}(T_1 > 179) = .0251$ from R

↑ will vary slightly for different runs.

At $\alpha = .05$, reject H_0 and conclude the mean number of runs is greater in games <u>with</u> the DH.

Randomization Test with Paired Data

- **This is another way to test the hypotheses of the Wilcoxon Signed-Ranks Test:**

$H_0: E(D) = 0$ (where $D = Y - X$)

vs.

$H_1: E(D) \neq 0$ or $H_1: E(D) < 0$ or $H_1: E(D) > 0$

- **Suppose we have n' paired observations (X_1, Y_1), (X_2, Y_2), …, $(X_{n'}, Y_{n'})$, and calculate the n nonzero differences via**

$D_i = Y_i - X_i, \quad i = 1, \ldots, n$

- **We assume the differences have symmetric distributions, are mutually independent with the same mean, and are at least interval in measurement scale.**

- **The test statistic is the sum of the positive differences:**

$$T_2 = \sum_i D_i \quad \underline{\text{only}} \text{ for those } D_i > 0$$

- **If the null hypothesis is true, then each difference has an equal chance to be positive or negative.**

- **Therefore we obtain all (or many) possible ordered combinations of "+" and "–" signs and attach these to our observed list of differences.**

- **We calculate T_2 for each of these combinations of signed differences, and this serves as our null distribution of T_2.**

- **If our observed T_2 is very "unusual" or "extreme" relative to this null distribution, we would reject H_0.**

- **The P-value is defined differently depending on H_1:**

Two-tailed: $2\left[\min\left\{\hat{P}(T_2 \geq T_2^{obs}), \hat{P}(T_2 \leq T_2^{obs})\right\}\right]$

Lower-tailed: $\hat{P}(T_2 \leq T_2^{obs})$

Upper-tailed: $\hat{P}(T_2 \geq T_2^{obs})$

where $\hat{P}$ is found based on the random permutations.

• **Again, we implement this method with R.**

Example: For 17 pairs of matched patients, each member of the pair was given either Drug A or Drug B in an attempt to reduce their cholesterol. The cholesterol reductions were recorded. Do the two drugs differ in terms of mean cholesterol reduction?

Y = chol. reduction for drug A, X = chol. reduc. for drug B

$D = Y - X \Rightarrow$ $H_0: E(D) = 0$ vs. $H_1: E(D) \neq 0$

$T_2^{obs} = 148$, and p-value = .023 from R.

At $\alpha = .05$, reject H_0 and conclude the drugs differ in mean chol. reduc.

Note: This approach can also work on a single sample $Y_1, Y_2, \ldots, Y_n$, to test whether the median of Y equals some constant number m.

• **To test** $H_0: \text{med}(Y) = m$
we let all $X_i = m$, i.e., form the pairs (m, Y_1), (m, Y_2), …, (m, Y_n) and carry out the randomization test as before.

Comparison to Competing Tests

• **Randomization tests work well in many situations, and permit other reasonable choices of test statistic.**

• **For heavy-tailed population distributions, randomization tests tend to have** more **power than parametric tests and** less **power than rank-based tests.**
• **For large sample sizes, the power of the randomization tests resembles the power of the parametric tests.**

Section 5.12: The Rank Transformation

• Many procedures in Chapter 5 are based on using ranks instead of raw data values.

• Several of these test procedures (Signed-rank, M-W, K-W) actually produce equivalent results to simply performing the respective classical parametric test on the ranks **rather than on the actual data.**

• In general, when data are clearly nonnormal or have outliers, it is usually a reasonable approach to rank all the data and then perform the usual parametric procedure on the ranks.

• Advanced multivariate methods such as multiple regression and discriminant analysis can be adapted to data having outliers by:

(1) ranking each variable separately

and

(2) performing the usual procedure on the ranks.

• This produces more robust **procedures.**

• In multiple regression, prediction can be accomplished by predicting the rank for a given individual and transforming this back onto the data scale via interpolation within the observed *Y*-values.