STAT 518 --- Section 5.8: Block Designs

• Recall that in paired-data studies, we match up pairs of subjects so that the two subjects in a pair are alike in some sense.

• Then we randomly assign, say, treatment A to one subject within a pair and, say, treatment B to the other subject. Picture:

• The rationale is that any difference we see in the measurements will likely be due to the treatments rather than intrinsic differences in the subject.

• When there are $k > 2$ treatments, we can separate the subjects into blocks, each having k subjects that are similar in some way (age, weight, health status, etc.).

Picture:

<u>Example 1</u>: We compare mean crop yields for 4 different fertilizers, where the blocks are contiguous regions of a field.
<u>Example 2</u>: We compare mean blood pressure reductions for 3 different drugs, where the blocks are formed of patients of similar age and health.
<u>Example 3</u>: We compare mean aggressiveness for mice placed in 5 different environments, where the blocks are the litters the mice came from.

• Assume there are *b* blocks, and each of the *k* treatments appears exactly once in each block.

• This experimental design is called a ________________ ________________ ______________ design.

• We are primarily interested in testing whether the *k* treatment groups have the same population mean.

The Friedman Test

The data are arranged as follows:

• We rank the data from 1 to k separately within each block (use midranks in case of ties).
• Let $R(X_{ij})$ be the rank within block i of the subject receiving treatment j.

• Then the sum of ranks for each treatment is

• If the b multivariate observations $(X_{11}, \ldots, X_{1k})$, …, $(X_{b1}, \ldots, X_{bk})$ are mutually independent (i.e., independent from block to block), and the data are at least ordinal so that we may rank data with each block, then Friedman's Test can test the hypotheses:

• Friedman suggested the statistic

which has an approximate null distribution.

• But it is often preferred to use

which has an approximate

null distribution, because the approximation is better.

• If H_0 is rejected, a multiple comparisons procedure is available to determine which pairs of treatments differe significantly.

Conclude treatments *i* and *j* are different if:

where

<u>Example 1</u> (Grass data): In a horticultural experiment for compare 4 grass types, 12 homeowners each grew the 4 grasses in parts of their yards. Since the same homeowner would provide the same degree of care to each grass, the homeowners were treated as blocks. At the end of the study, the grasses were ranked from 1 (worst) to 4 (best), separately within each block. The data are on page 372. Is there a significant difference in mean quality among the 4 grasses? (Use $\alpha = 0.05$.)

• The sum of ranks for each grass is:

Test statistic:

Decision rule:

Multiple comparisons:

Related Tests

• The Quade Test tests the same hypotheses as Friedman’s Test.

• Quade’s test gives more weight to blocks whose sample range (maximum – minimum value) is largest.

• It is sometimes more powerful than Friedman’s test, but it requires that we have actual data values, not just ranks as in Example 1.

• Page's Test is appropriate when our alternative hypothesis specified a particular ordering of the population means. The hypotheses are:

• A test statistic is

where the ordering of R_1, …, R_k correspond to the ordering specified in H_1.

• The decision rule is based on a function of the test statistic that has a normal asymptotic null distribution:

Example 1 again: Suppose we suspected that the ordering of the grasses was 2, 3, 4, 1. Test whether there is evidence for this ordering.

<u>Note</u>: If each treatment appears *m* > 1 times (i.e., replication) within each block, Friedman's test can be adjusted to account for this. See pp. 383-384 for details.

Comparison to Competing Tests

• Note: For *k* = 2, Friedman's test is equivalent to the ____________ test and Quade's test is equivalent to the _____________________________ test.

• In general, for *k* samples, the A.R.E. of Friedman's test relative to the classical F-test depends on *k*.

• When the alternative is a simple shift in treatment means, the A.R.E. of Friedman's test relative to the classical F-test is never less than _______________.

Efficiency of Friedman's Test

Population	A.R.E.(Friedman vs. F)
Normal	
Uniform (light tails)	
Double exponential (heavy tails)	

• Quade's test is more powerful than Friedman's when *k* is 3 or 4, but tends to be less powerful than Friedman's when $k \geq 5$.

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:

• 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 total $n + m$ 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 <u>combined</u> 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:

• 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:

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

Example: Random samples of games in which an American League team played with a designated hitter and without 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 $\alpha = 0.05$.)

Randomization Test with Paired Data

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

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

• 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:

• 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:

• 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?

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
we let all $X_i = m$, i.e., form the pairs $(m, Y_1), (m, Y_2), \ldots, (m, Y_n)$ 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 __________ power than parametric tests and ________ 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 ___________ 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 <u>multiple regression</u> and <u>discriminant analysis</u> can be adapted to data having outliers by:

(1)

and

(2)

• This produces more ____________ 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.