Exam 1, Thurs Mar 3, 1 hand-written page of notes, calculator
Grades between 0-100, covers everything that we've talked about Chapter 0 - Chapter 2

How do we deal with the large-sample situation?

- We have noted the similarity in procedures thus far as making use of sums of scores associated with treatments.

Assumption of variance, same spread or very similar
Assumptions:

- Observations independent
- Continuity of distribution (maybe)
- Nothing on normality, etc.

Omnibus test = komogorov-smirnoff test

- We will cover a procedure for obtaining a large-sample approximation of the permutation distribution of any statistic that can be computed as the sum of scores associated with one of the two treatments.
- Applies to the Wilcoxon rank-sum statistic, Wilcoxon rank-sum statistic with ties, the van der Waerden......

As usual, let m and n denote the number of observations in treatments 1 and 2

- N = m + n is the combined sample size
- Let A_1, A_2,...,A_N be N numbers representing ranks, normal scores, or other such scores that have been assigned to the combined observations.
- Similar to earlier, let T_1 denote the sum of the A_i's associated with treatment 1.

If there is no difference between the two treatments then:

- Any A_i is as likely to occur among the scores for treatment 1 as any other A_i.
- Therefore, the m scores associated with treatment 1 occur as if they had been randomly selected without replacement from the finite population of scores A_1, A_2,...,A_N

Sampling theory gives us:

$$E(T_1) = m\mu$$

$$var(T_1) = \frac{mn\sigma^2}{N-1}$$

Where

$$\mu = \frac{\sum_{i=1}^{N} A_i}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (A_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} (A_i)^2}{N} - \mu^2$$

Therefore, if m and n are sufficiently large, then T_1 will have an approximate normal distribution. Hence:

$$Z = \frac{T_1 - E(T_1)}{}$$

$$- \sqrt{var(T_1)}$$

Thus, one may use the resulting Z statistic and the associated standard normal distribution to make decisions about the hypothesis in question.

For the Wilcoxon rank-sum test without ties, the A_i's are the ranks 1,2,...N.
- Need to compute mu and sigma^2

In order to compute mu and sigma^2, we make use of the following identities

$$\frac{N(N+1)}{2} = 1+2+\dots+N$$

$$1^2+2^2+\dots+N^2 = \frac{N(N+1)(2N+1)}{6}$$

Therefore,

$$\mu = \frac{1+2+\dots+N}{N} = \frac{N+1}{2}$$

$$\sigma^2 = \frac{1^2+2^2+\dots+N^2}{N} - \left(\frac{(N+1)}{2}\right)^2$$

$$= \frac{(N-1)(N+1)}{12}$$

Therefore,

$$\left.\begin{aligned} E(W) &= \frac{m(N+1)}{2} \\ var(W) &= \frac{mn(N+1)}{12} \end{aligned}\right\} \text{ranks only}$$

Example 2.10.1:
- Find Z and the associated p-value if m=6 and n=6 for P(W>=50)
- Find Z and the associated p-value if a continuity correction is used for P(W>=50) use 49.5
- Compare these approximations to the exact p-value of .0465

Continuity correction = +- 0.5
If W>=50 -> cc(49.5)
If W<=50 -> cc(50.5)

Example 2.10.2:
- For the data in Example 2.4.3 on the dry weights of strawberry plants, the two groups have nine and

seven observations, respectively, and the rank sum is W=84 for the treatment with seven observations.
-Find P(W>=84) without the continuity correction.
-Compare to the exact p-value of .0039.

Though one can apply the preceding sample formulas directly to the adjusted ranks for tied data, there are actually explicit formulas for E(W_ties) and var(W_ties).

- Since average ranks are used for tied observations, E(W_ties) is the same as if the data were not tied.
- However, the variance is adjusted downwards from what it would be if the observations were not tied.

Therefore, let's examine exactly how the explicit variance formula differs.

In the combined data set, put each set of tied observations into its own group
Let k be the number of such groups, and let t_i denote the number of observations (the number of ties) in the ith group,
i=1,2,...,k.
To obtain the variance with ties, we compute:

$$var(W_{ties}) = \frac{mn(N+1)}{12} - AF$$

Where AF is called the adjustment factor and computed by:

$$AF = \frac{mn \sum_{n}}{\quad}$$

The ties are {11,11}, {13,13,13}, and {19,19}. Therefore, t_1 = 2, t_2 =3, and t_3 = 2

Let's use the notation of Section 2.6.
Observations from treatment 1: X1, X2,...,X_m
Observations from treatment 2: Y1,Y2,...,Y_n
The Mann-Whitney statistic, U: the number of pairs of (X_i,Y_j) such that X_i < Y_
W_2: the sum of the ranks of the Y's.

- We remember that:

$$W_2 = \frac{n(n+1)}{2} + U$$

Therefore,

- since E(W_2) =......, it follows that

$$E(U) = \frac{n(N+1)}{2} - \frac{n(n+1)}{2} = \frac{mn}{2}$$

And

$$var(U) = var(W_2)$$

Where var(U) is computed with or without ties as appropriate

Using a normal approximation for the distribution of U, we have:

$$P(k_a \le U \le k_b - 1)$$

$$\approx P\left(\frac{k_A - E(U)}{\sqrt{var(U)}} \le Z \le \frac{k_B - 1 - E(U)}{\sqrt{VAR(U)}}\right) = .90$$

Exam 1 Material
1st half (boring) 50% = application of stuff that we've learned like quiz/HW, one problem one answer
No writing down R code
Study by going over examples, HW problems without R work
Quiz will be sent out
2nd half 50% = given a problem, act as a consultant, absolutely 0 data, give a solution (steps), options possible, whatever choices you make JUSTIFY reasons
Assumptions needed, which test, differences between, why, what results you would be looking for

Know what power is, why power is important (assessing 2 different tests)