STAT 518 --- Section 5.1: The Mann-Whitney Test

• We now examine the situation when our data consist of two independent samples.

Example 1: We want to compare urban versus rural high school seniors on the basis of their test scores.
Example 2: We want to estimate the difference between the median BMIs for females and males.
Example 3: We want to compare the housing markets in New York and California in terms of median selling price.

• There is no natural pairing in the data: We simply have two separate independent samples.
• The sizes of the two samples, say n and m, could be different.

• Assume we have independent random samples from two populations.
• The measurement scale of the data is at least ordinal.

• Denote the first sample by $X_1, X_2, \ldots, X_n$ and the second sample by $Y_1, Y_2, \ldots, Y_m$.

• The null hypothesis of the Mann-Whitney test (also called the Wilcoxon Rank Sum test) can be stated in terms of the cumulative distribution functions:

• The alternative hypothesis could be any of these three:

• However, it is more interpretable to state the null and alternative hypotheses in terms of probabilities:

Two-tailed	**Lower-tailed**	**Upper-tailed**

• This test could also be used simply as a comparison of two means:

Two-tailed	**Lower-tailed**	**Upper-tailed**

• If the M-W test is used to compare two means, we should assume that the c.d.f.'s of the two populations are the same except for a potential shift. Picture:

• We first combine the X's and Y's into a combined set of N values, where $N = n + m$.

• We rank the observations in the combined sample, with the smallest having rank 1 and the largest, $n + m$.
• If there are ties, midranks are used.

• The test statistic is $T =$

• Table A7 tabulates null distribution of T for selected sample sizes (for $n \leq 20$ and $m \leq 20$).

• This is exact if there are no ties.

• Upper quantiles of T are found via the formula:

• Or, for an upper-tailed situation, we could equivalently use the statistic:

along with the corresponding lower-tail quantile.

• For examples with many ties, or with larger sample sizes, we can use another test statistic:

where

<u>Decision Rules</u>

Two-tailed Lower-tailed Upper-tailed

• If the test is performed using T_1, then standard normal quantiles are used rather than the values in Table A7.

• Approximate <u>P-values</u> can be obtained from the normal distribution using one of equations (6)-(10) on pp. 274-275, or by interpolating within Table A7, but we will typically use software to get approximate P-values.

Example 1: In a simulated-driving experiment, subjects were asked to react to a red "brake" light. Their reaction time (in milliseconds) was recorded. Some of the subjects were conversing on cell phones while "driving" while another group was listening to a radio broadcast. Is mean reaction time significantly greater for the cell-phone group?

Data

Cell: 456, 468, 482, 501, 672, 679, 688, 960

Radio: 426, 436, 444, 449, 626, 626, 642

Hypotheses:

Decision rule: Reject H_0 if

Test statistic:

P-value =

Conclusion:

On computer: Use `wilcox.test` function in R (see example code on course web page)

Example 2: Samples of sale prices for a handheld computing device on eBay were collected for two different auction methods (bidding and buy-it-now). At $\alpha = .05$, are the mean selling prices significantly different for the two groups?

<u>Data</u>

Bidding: 199, 210, 228, 232, 245, 246, 246, 249, 255

BIN: 210, 225, 225, 235, 240, 250, 251

Hypotheses:

Decision rule: Reject H_0 if

Test statistic:

P-value =

Conclusion:

On computer: Use `wilcox.test` function in R (see example code on course web page.

• The M-W test can be used to test hypotheses like:

where *d* is some specific number of interest.
• In this case, simply add *d* to each *Y* value and carry out the M-W test on the *X*'s and the adjusted *Y*'s.

• When estimating the difference between E(*X*) and E(*Y*) is of interest, a CI can be obtained.

Confidence Interval for the Difference in Two Population Means

• The values in the (1 – α)100% CI are all numbers *d* such that the above null hypothesis is <u>not</u> rejected at level α.

• To find this CI for E(*X*) – E(*Y*):

• Calculate

• Find <u>all</u> differences $X_i - Y_j$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

• The CI endpoints are the *k*-th smallest and the *k*-th largest of these differences.

• Note: Computing and sorting the differences is most easily done via software.

Example 1 again: Find a 90% CI for the difference between the mean reaction times for the cell-phone drivers and the radio drivers.

Example 2 again: Find a 95% CI for the difference between the population mean selling prices for the bidding group and the buy-it-now group.

Comparison of M-W test to Competing Tests

• If both populations are normal, the 2-sample t-test is most powerful for comparing two means.

• However, the 2-sample t-test lacks power when one or both samples contain ______________.

• The median test (covered in Chapter 4) is another distribution-free test in this situation.

Efficiency of the Mann-Whitney Test

Population	A.R.E.(M-W vs. t)	A.R.E.(M-W vs. median)
Normal		
Uniform (light tails)		
Double exponential (heavy tails)		

• The A.R.E. is of the M-W test relative to the t-test is never lower than _______ but may be as high as _____.

• For <u>small</u> samples coming from heavy-tailed distributions, the M-W test may be _________________ than the median test.
• But the median test is more <u>flexible</u> --- it does not require the distributions of *X* and *Y* to be identical under H_0.

Section 5.2: Analyzing Several Independent Samples

• The M-W test is designed to compare two populations.

• Sometimes we have *k* independent samples from *k* populations.

• We wish to test whether all *k* populations are identical in distribution.

Kruskal-Wallis Test

• We assume the *k* random samples are all mutually independent and that the measurement scale is at least <u>ordinal</u>.

• The K-W test is again based on the <u>ranks</u>.

• Denote Sample 1 as

Sample 2 as

Sample *k* as

• We combine all *k* samples and rank the observations in the combined sample from 1 (smallest) to *N* (largest).

• Let

Hypotheses:

Test Statistic:

Null Distribution of *T*

Note:

• So the asymptotic null distribution of T is χ^2 with (k – 1) degrees of freedom.

Decision Rule

• T is large when the R_i's are fairly different from each other.
• This is evidence in favor of

So:

Example 1: In an experiment, 43 newborn chicks were each given one of 4 diets. Weight gain in the first 21 days was measured (in grams). Is there evidence (at α = 0.05) that the four diets produce different mean weight gains?

On computer: Use `kruskal.test` function in R (see example code on course web page.

• If H_0 is rejected, we use <u>multiple comparisons</u> to infer <u>which</u> population means seem to differ.

• Populations *i* and *j* are significantly different if:

• This can be checked readily in R.

Example 1 again:

• The K-W test can be used with categorical data (e.g., data in contingency tables) as long as the variable observed on each individual is <u>ordinal</u> so that the categories can be ranked in order.

Example 2: The grade distributions for 3 instructors were compared to see whether students tended to get similar grade distributions across instructor. The data are given on page 293.

• If we score A, B, C, D, F numerically as 4, 3, 2, 1, 0, then we can perform the K-W test on the data:

Comparison to Other Tests

• When all *k* populations are normal, the usual parametric procedure to compare the *k* population means is the (one-way) analysis of variance (ANOVA) F-test.

• The F-test is robust against the normality assumption in terms of the actual significance level.

• But the F-test can have ________ _______ power when the data are nonnormal (especially when heavy-tailed).

• The A.R.E. of the K-W test relative to the F-test and relative to the median test is very similar to the A.R.E. of the M-W test relative to its competitors.