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

- **We now examine the situation when our data consist of two <u>independent samples</u>.**

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

- **There is no natural <u>pairing</u> in the data: We simply have two separate <u>independent</u> 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 <u>first</u> sample by $X_1, X_2, \ldots, X_n$ and the <u>second</u> sample by $Y_1, Y_2, \ldots, Y_m$.**

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

$$H_0: F(x) = G(x) \text{ for all } x$$

where $F(\cdot)$ is the cdf corresponding to X_i's and $G(\cdot)$ is the cdf corresponding to Y_i's.

- **The alternative hypothesis could be any of these three:**

$H_1: F(x) \neq G(x)$ for some x	$H_1: F(x) > G(x)$ for all x	$H_1: F(x) < G(x)$ for all x

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

$$H_0: P(X > Y) = P(X < Y)$$

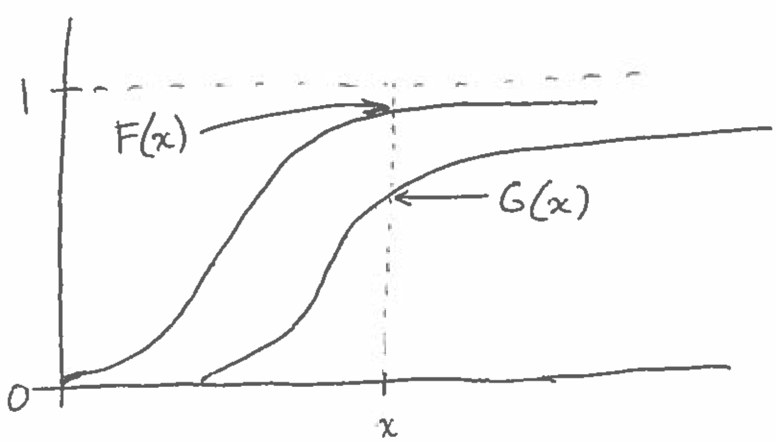
Two-tailed	**Lower-tailed**	**Upper-tailed**
$H_1: P(X>Y) \neq P(X<Y)$	$H_1: P(X>Y) < P(X<Y)$ "Y tends to be larger than X"	$H_1: P(X>Y) > P(X<Y)$ "X tends to be larger than Y"

- **This test could also be used simply as a comparison of two means:** (or medians) $H_0: E(X) = E(Y)$

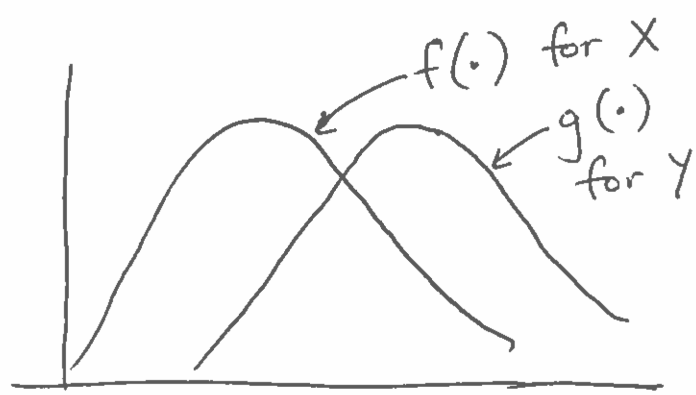
Two-tailed	**Lower-tailed**	**Upper-tailed**
$H_1: E(X) \neq E(Y)$	$H_1: E(X) < E(Y)$	$H_1: E(X) > E(Y)$

- **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:**

Cdf's:



Densities:



• 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 <u>smallest</u> having rank 1 and the <u>largest</u>, $n + m$.
• If there are ties, midranks are used.

• The <u>test statistic</u> is $T = \sum_{i=1}^{n} R(X_{ij})$,

the sum of the ranks assigned to observations in the sample from population 1 (the X's)

• 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:

$$w_p = n(n+m+1) - w_{1-p}$$

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

$$T' = n(N+1) - T$$

along with the corresponding lower-tail quantile.

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

$$T_1 = \frac{T - \frac{n(N+1)}{2}}{\sqrt{\frac{nm}{N(N-1)} \sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}$$

where $\sum_{i=1}^{N} R_i^2$ = the sum of squares of all N ranks from both samples.

Decision Rules

Two-tailed	**Lower-tailed**	**Upper-tailed**
Reject Ho if $T < W_{\alpha/2}$ or if $T > W_{1-\alpha/2}$ (these quantiles are found in Table A7)	Reject Ho if $T < W_{\alpha}$ ← in Table A7	Reject Ho if $T > W_{1-\alpha}$ (or could do: Reject Ho if $T' < W_{\alpha}$) ← in Table A7

• **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?** Use $\alpha = .05$

Data

X	**Cell:**	**456,**	**468,**	**482,**	**501,**	**672,**	**679,**	**688,**	**960**
rank:		5	6	7	8	12	13	14	15 ⇒ $T = 80$
Y	**Radio:**	**426,**	**436,**	**444,**	**449,**	**626,**	**626,**	**642**	
rank:		1	2	3	4	9.5	9.5	11	

Hypotheses: $H_0: E(X) \leq E(Y)$

$H_1: E(X) > E(Y)$

Decision rule: Reject H_0 **if** $T' < w_{.05}$

$n = 8$, $m = 7$

⇒ Reject H_0 if $T' < 50$ ← Table A7

(Equivalent: Reject H_0 if $T > w_{.95} = 8(15+1) - 50 = 78$)

Test statistic: $T' = 8(15+1) - 80 = 48$

P-value = .0363 from R

Conclusion: Reject H_0 since $48 < 50$. Conclude the mean reaction time is greater for the cell-phone group than for the radio group.

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?

Data

X	**Bidding:**	**199**	**210**	**228**	**232**	**245**	**246**	**246**	**249**	**255**
rank:		1	2.5	6	7	10	11.5	11.5	13	16
Y	**BIN:**	**210**	**225**	**225**	**235**	**240**	**250**	**251**		
rank:		2.5	4.5	4.5	8	9	14	15		

$\Rightarrow T = 78.5$

Hypotheses: $H_0: E(X) = E(Y)$

$H_1: E(X) \neq E(Y)$

Decision rule: Reject H_0 if $T < w_{.025}$ or if $T > w_{.975}$ $\quad n = 9$, $m = 7$

Reject H_0 if $T < 58$ or if $T > 9(16+1) - 58 = 95$

(58 from Table A7)

Test statistic: $T = 78.5$, so we fail to reject H_0.

P-value = 0.8736 from R.

Conclusion: We cannot conclude the mean selling price is different for bidding and buy-it-now methods.

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:**

$$H_0: E(X) - E(Y) = d$$
$$H_1: E(X) - E(Y) \neq d$$

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 not rejected at level α.**

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

- **Calculate** $k = w_{\alpha/2} - \frac{n(n+1)}{2}$ using Table A7 and the appropriate n and m.
- **Find all 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. $n=8, \ m=7$

$W_{\alpha/2} = W_{.05} = 50$ from Table A7.

$$k = 50 - \frac{8(9)}{2} = 50 - 36 = 14$$

90% CI from R: [12, 239].

With 90% confidence, the mean cell-phone reaction time is between 12 and 239 milliseconds <u>more</u> than the mean radio reaction time.

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.

$n=9, \ m=7$

$W_{\alpha/2} = W_{.025} = 58$ from Table A7.

$$K = 58 - \frac{9(10)}{2} = 58 - 45 = 13$$

95% CI from R: [−19, 21]

With 95% confidence, the mean selling price for bidding method is between 19 dollars <u>less</u> than and 21 dollars <u>greater</u> than the mean price for the BIN method.

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 outliers.

• 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	0.955	1.5
Uniform (light tails)	1.0	3.0
Double exponential (heavy tails)	1.5	0.75

• The A.R.E. of the M-W test relative to the t-test is never lower than 0.864 but may be as high as ∞.

• For <u>small</u> samples coming from heavy-tailed distributions, the M-W test may be much more powerful 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.