UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

Instructor: Ivo Dinov,
Asst. Prof. of Statistics and Neurology

Teaching Assistants:
Brandi Shanata & Tiffany Head

University of California, Los Angeles, Fall 2007
http://www.stat.ucla.edu/~dinov/courses_students.html

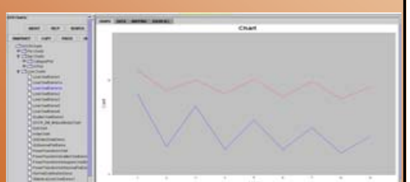
Lecture Set 8
The T Test
Wilcoxon-Mann-Whitney Test

Application

Example: Nine observations of surface soil pH were made two different locations. Does the data suggest that the true mean soil pH values differ for the two locations? Test using $\alpha = 0.05$, and be sure to check any necessary assumptions for the validity of your test.

LineChartDemo1b
http://socr.ucla.edu/htmls/SOCR_Charts.html

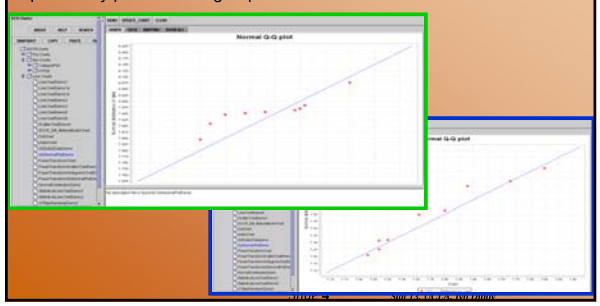
Location 1	Location 2
8.10	7.85
7.89	7.30
8.00	7.73
7.85	7.27
8.01	7.58
7.82	7.27
7.99	7.50
7.80	7.23
7.93	7.41



Application

QQNormalPlotDemo: http://socr.ucla.edu/htmls/SOCR_Charts.html

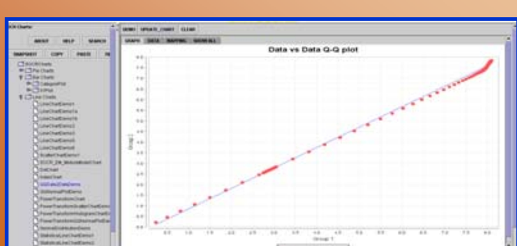
To meet the assumption of normality (necessary for the t-test with such a small sample size in each group), we will calculate a normal probability plot for each group.



Application

QQData2DataDemo: http://socr.ucla.edu/htmls/SOCR_Charts.html

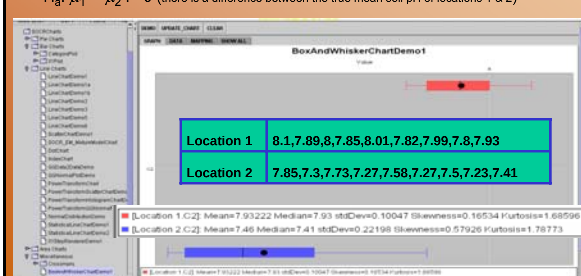
Equ-distributed samples in the two groups? Calculate a QQ probability plot of one group against the other.



Application

BoxAndWhiskerCHartDemo1: http://socr.ucla.edu/htmls/SOCR_Charts.html

- #1 Formulate hypotheses
  - H_o: $\mu_1 - \mu_2 = 0$ (no difference between the true mean soil pH of locations 1 & 2)
  - H_a: $\mu_1 - \mu_2 \ != 0$ (there is a difference between the true mean soil pH of locations 1 & 2)



Location 1	8.1,7.89,8,7.85,8.01,7.82,7.99,7.8,7.93
Location 2	7.85,7.3,7.73,7.27,7.58,7.27,7.5,7.23,7.41

Application

- #2 Calculate the test statistic

Descriptive Statistics: Location 1, Location 2

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Location 1	9	0	7.9322	0.0335	0.1005	7.8000	7.8350	7.9300	8.0050
Location 2	9	0	7.4600	0.0740	0.2220	7.2300	7.2700	7.4100	7.6550

Variable	Maximum
Location 1	8.1000
Location 2	7.8500

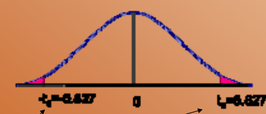
$$SE_{\bar{y}_1-\bar{y}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \sqrt{\frac{0.1005^2}{9}+\frac{0.222^2}{9}} = 0.081$$

$$t_s = \frac{\bar{y}_1-\bar{y}_2-0}{SE_{\bar{y}_1-\bar{y}_2}} = \frac{7.9322-7.460-0}{0.081} = 5.827$$

Application

- #3 Calculate the p-value

$$df = \frac{\left(SE_1^2+SE_2^2\right)^2}{\frac{SE_1^4}{n_1-1}+\frac{SE_2^4}{n_2-1}} = \frac{\left(0.0335^2+0.074^2\right)^2}{\frac{0.0335^4}{9-1}+\frac{0.074^4}{9-1}} = 11.03 \approx 11 df$$



$p < 2(0.0005) = 0.001$ (SOCR)

Application

- #4 Conclusion

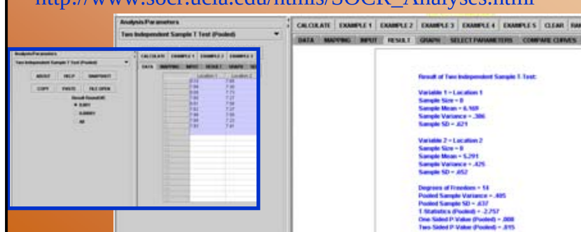
Because $p < 0.001 < 0.05$, we will reject H_o.

CONCLUSION: These data show that there is a statistically significant true mean difference in the pH of Location 1 and Location 2 ($P < 0.001$).

Application

- #5 SOCR Analysis:

http://www.socr.ucla.edu/htmls/SOCR_Analyses.html



Application

- #5 SOCR Analysis:

http://www.socr.ucla.edu/htmls/SOCR_Analyses.html

Result of Two Independent Sample T-Test:

Variable 1 = Location 1
Sample Size = 8
Sample Mean = 6.169
Sample Variance = .386
Sample SD = .621

Variable 2 = Location 2
Sample Size = 8
Sample Mean = 5.291
Sample Variance = .425
Sample SD = .652

Degrees of Freedom = 14
Pooled Sample Variance = .405
Pooled Sample SD = .637
T-Statistics (Pooled) = -2.757
One-Sided P-Value (Pooled) = .008
Two-Sided P-Value (Pooled) = .015

Application

- Confidence interval for $\mu_1 - \mu_2$
  - Suppose we calculated a 95% confidence interval to be:

$$\left(\bar{y}_1-\bar{y}_2\right) \pm t(df)_{0.025}\left(SE_{\bar{y}_1-\bar{y}_2}\right) = (7.932-7.460) \pm t(11)_{0.025}(0.081)$$
$$= (0.472) \pm 2.201(0.081)$$
$$= (0.294,\ 0.650)$$

  - Does this interval surprise you?

Application

- Corresponding computer output:

```
Two-Sample T-Test and CI: Location 1, Location 2

Two-sample T for Location 1 vs Location 2

             N   Mean  StDev  SE Mean
Location 1   9  7.932  0.100    0.033
Location 2   9  7.460  0.222    0.074


Difference = mu (Location 1) - mu (Location 2)
Estimate for difference:  0.472222
95% CI for difference:  (0.293459, 0.650985)
T-Test of difference = 0 (vs not =): T-Value = 5.81
P-Value = 0.000  DF = 11
```


CI and Hypothesis-Testing relationship

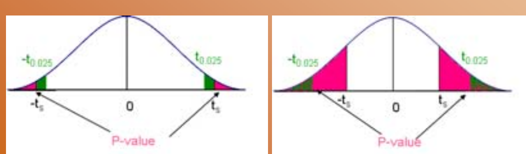
- Consider a 95% confidence interval for $\mu_1 - \mu_2$ and it's relationship to the t test at $\alpha = 0.05$
  - Both use $\bar{y}_1 - \bar{y}_2$ and $SE_{\bar{y}_1-\bar{y}_2}$ in their calculations

CI: $$(\bar{y}_1 - \bar{y}_2) \pm t(df)_{\alpha/2}\left(SE_{\bar{y}_1-\bar{y}_2}\right)$$

T_s: $$t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{SE_{\bar{y}_1-\bar{y}_2}}$$

CI and Hypothesis-Testing relationship

- With a t test we reject H_o if the p-value is less than α and fail to reject otherwise
  - this is the same thing as saying we reject if t_s is beyond $\pm t_{0.025}$, and fail to reject otherwise



CI and Hypothesis-Testing relationship

- Focusing on the upper half of the distribution and remembering the symmetry: we fail to reject when

$$|T_s| = \frac{|\bar{y}_1 - \bar{y}_2|}{SE_{\bar{y}_1-\bar{y}_2}} < t_{0.025}$$

  - Further manipulation gives us:

$$\begin{aligned}
&|\bar{y}_1 - \bar{y}_2| < t_{0.025}(SE_{\bar{y}_1-\bar{y}_2})\\
&= -t_{0.025}(SE_{\bar{y}_1-\bar{y}_2}) < \bar{y}_1 - \bar{y}_2 < t_{0.025}(SE_{\bar{y}_1-\bar{y}_2})\\
&= -(\bar{y}_1 - \bar{y}_2) - t_{0.025}(SE_{\bar{y}_1-\bar{y}_2}) < 0 < -(\bar{y}_1 - \bar{y}_2) + t_{0.025}(SE_{\bar{y}_1-\bar{y}_2})\\
&= (\bar{y}_1 - \bar{y}_2) + t_{0.025}(SE_{\bar{y}_1-\bar{y}_2}) > 0 > (\bar{y}_1 - \bar{y}_2) - t_{0.025}(SE_{\bar{y}_1-\bar{y}_2})
\end{aligned}$$

- Therefore, we **fail to reject H_o**: $\mu_1 - \mu_2 = 0$ (for the not equal to alternative), if the **confidence interval contains 0**.

CI and Hypothesis-Testing relationship

- If a two-tailed t test and a confidence interval give us the same result, why learn both?
  - There are advantages to each one
    - **Confidence interval**: shows magnitude of difference between μ_1 and μ_2
    - **T test**: has p-value which describes the strength of evidence that μ_1 and μ_2 are really different.

More on the significance level α

- **Choose a significance level BEFORE analyzing the data**

Example: Say *df* = 15 and α = 0.05

- If t_s is in either tail we will reject H_o. The chance of this happening due to **random variation** is 0.05. I.e., P(reject H_o) = 0.05, if H_o is true.
- Because we are assuming that Ho is true, all t_s values on the t curve would only deviate from 0 because of **sampling error**.
- This means:
  - 95% would fail to reject H_o
  - 2.5% would reject H_o ($-t_s$)
  - 2.5% would reject H_o (t_s)

In other words, a total of 5% would reject H_o when H_o is actually true. This is an incorrect conclusion just because of sampling error!

More on the significance level α

• When we are analyzing one data set in real life at the 0.05 level and our conclusion is to reject H_o there are two possible scenarios:

1. H_o is in fact false
2. H_o is true, but we were unlucky (5%)

Type I and Type II Errors

- There are two possible mistakes that can be made when conducting a hypothesis test:
  - A type I error is when we reject H_o and H_o was true
    - **P(type I error) = α**
    - When we choose α before we conduct our test, we are actually protecting ourselves against a type I error
    - This choice will depend on your experiment
  - A type II error is when we fail to reject H_o and H_o is false
    - **P(type II error) = β**
    - β can also be specified before we collect our data
    - will have more to do with the number of observations in our sample

Type I and Type II Errors

- Table (below) is the best way to summarize

		Reality	
		H_o True	H_o False
Decision	Fail To Reject H_o	Correct TN	Type II FN
	Reject H_o	Type I FP	Correct TP

- You cannot make both errors at the same time
  - Once you have reached a conclusion (reject or fail to reject) based on the data from your experiment you've either made a correct decision or you've made an error (type I for a reject conclusion and type II for a failing-to-reject conclusion)

Type I and Type II Errors

- Analogy: Think of a car with a car alarm being broken into
  - If the alarm goes off for no reason (reject H_o when H_o is true) - type I error
  - If the car gets broken into and the alarm doesn't go off (fail to reject Ho when Ho is false) - type II error
  - Also consider the sensitivity of the alarm
  - REMEMBER: Fail To Reject H_o means "nothing is going on" or the data do not show otherwise
- Consequences of Type I / II errors are quite different

Type I and Type II Errors

Example: Measuring pollution in a lake. Say the EPA institutes a rule that companies near bodies of water must test their pollution output. If the company doesn't find any statistical significance in their results, they may continue their current practices.

H_o: No significant pollution

H_a: Significant pollution

In this case a type II error would be much worse (probability of failing to reject H_o when H_o is false – saying no significant pollution when there really is)

An "ethical" company would want to make sure they tested enough samples to guarantee that β is small

Type I and Type II Errors

Example: Drug Treatments. Say a doctor would like to study a new highly toxic drug treatment for cancer patients. There are many risks and side effects of the new drug, but would be of benefit if the proportion of patients responding is greater than 50%.

H_o: No significant response (Proportion responding to TX is $\le$ 0.5)

H_a: Significant response (Proportion responding to TX is > 0.5)

In this case a type I error is much worse (probability of rejecting H_o when H_o is true – like saying that the TX does something when it really doesn't)

An ethical researcher would want to make sure they keep α small before collecting and analyzing the data

Type I and Type II Errors

- Because α is chosen beforehand, we are protected against type I errors. However, type II errors depend on many things, such as sample size (section 7.8)
- β = P(fail to reject H_o) when H_o is false.
  - The chance of rejecting H_o when it is actually false is called the power of our test
    - Power = 1 - β = P(reject H_o) when H_o is false
    - measures the ability of the test to detect a difference when a difference really does exist
    - Power depends on sample size. A larger sample gives more information and also increases power.
    - When you plan an analysis you always need to take power into account (ie plan for n):
      - decide desired SE and calculate n
      - analysis of power (7.8)

One Tailed t Tests

- The previous hypothesis test was called a two-tailed (or non-directional) test because H_o was rejected if t_s fell in <u>either</u> tail
- In some analyses it is reasonable that there will be a certain direction of a deviation from H_o
  - This means that we are looking to see if one group mean is smaller/larger than the other
- The hypotheses for a one-tailed (or directional) test are:

  H_o: $\mu_1 - \mu_2 = 0$
  H_a: $\mu_1 - \mu_2 > 0$
  OR
  Ha: $\mu_1 - \mu_2 < 0$
  - Note: the null hypothesis doesn't change

One Tailed t Tests

- One-Tailed Test Procedure:
  - Step 1: Check direction to see if data deviate from H_o in the direction specified by H_a
    (If $\mu_1 < \mu_2$ then we expect t_s to be negative, if $\mu_1 > \mu_2$ then we expect t_s to be positive.)
    a. If no, then p-value > 0.5
    b. If yes, then proceed to step 2
  - Step 2: The p-value of the data is the one-tailed area beyond t_s

One Tailed t Tests

<u>Example</u>: Cholesterol (cont')

RECALL: Group 1 = Medication, Group 2 = Placebo

Suppose it is reasonable to assume that $\mu_1 < \mu_2$, in other words the researcher is hoping to show that this new medication lowers cholesterol

The appropriate hypotheses would be

H_o: $\mu_1 - \mu_2 = 0$

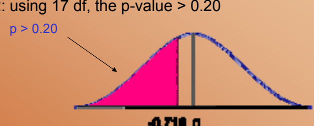
H_a: $\mu_1 - \mu_2 < 0$

Calculate the p-value for this test

Step 1: t_s was calculated as -0.716. Check that the data deviate in the direction of H_a.

One Tailed t Tests

Step2: using 17 df, the p-value > 0.20



What would the p-value be if the researchers had specified an upper tailed hypothesis?

One Tailed t Tests

<u>Example</u>: Soil pH (cont')

Suppose researchers had reason to believe that the soil pH for Location 1 was greater than Location 2.

Two-Sample T-Test and CI: Location 1, Location 2

Two-sample T for Location 1 vs Location 2

	N	Mean	StDev	SE Mean
Location 1	9	7.932	0.100	0.033
Location 2	9	7.460	0.222	0.074

Difference = $\mu_1 - \mu_2$
Estimate for difference: 0.472222
95% lower bound for difference: 0.326361
T-Test of difference = 0 (vs >): T-Value = 5.81
P-Value = 0.000 DF = 11

One Tailed t Tests

- P-values for a directional alternative are 1/2 of a non-directional
  - *assuming the direction matched H_a*
- It is easier to reject H_o with a one-tailed alternative
  - However it is important that we decide on the direction of H_a before the data is collected
- If the data do not match the direction of H_a we conclude that the data do not indicate that the true means differ
  - However t_s may be statistically significant in the other tail
    - In this case we would want to look for methodological errors in the experiment

The Wilcoxon-Mann-Whitney

- Also known as the rank sum test
- http://www.socr.ucla.edu/htmls/SOCR_Analyses.html
- http://socr.stat.ucla.edu/Applets.dir/WilcoxonRankSumTable.html
- This hypothesis test is **also used to compare two independent samples**
  - This procedure is different from the independent t test because it is **valid even if the population distributions are not normal**
  - In other words, we can use this test as a fair substitute when we cannot not meet the required normality assumption of the t test
- WMW is called **distribution-free** or non-parametric test
  - This test doesn't focus on a parameter like the mean, instead it examines the distributions of the two groups

The Wilcoxon-Mann-Whitney

- Keep in mind that this is another hypothesis test, so there are still four major parts to consider
- #1 The hypotheses:
  - H_o: The population distributions of Y_1 and Y_2 are the same
  - H_a: The population distributions of Y_1 and Y_2 are the different
    - This could also be directional: distribution of Y_1 is less than Y_2; OR distribution of Y_1 is greater than Y_2
- #2 The test statistic:
  - denoted by U_s
  - measures the degree of separation between the two samples
    - a large value of U_s indicates that the two samples are well separated with little overlap
    - a small value of U_s indicates that the two samples are not well separated with much overlap

The Wilcoxon-Mann-Whitney

- #3 The p-value:
  - http://socr.stat.ucla.edu/Applets.dir/WilcoxonRankSumTable.html
  - http://www.socr.ucla.edu/htmls/SOCR_Analyses.html
  - Critical Values are also in table 6 on p.680
  - Method very similar to using the t table
    - find the appropriate row and then search for a number closest to the test statistic
  - don't need to worry about doubling the p-value for a two-tailed test (assuming we go to the right row header)
- #4 Conclusion:
  - Similar to the conclusion of an independent t test, but not linked to any parameter (for example the difference in means)

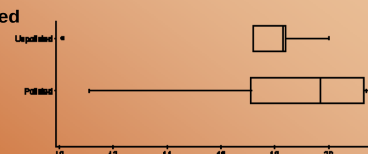
The Wilcoxon-Mann-Whitney

- The Method:
  - Step 1: Arrange the data in increasing order
  - Step 2: Determine K_1 and K_2
    - K_1: for each observation in group 1, count the number of observations in the second group that are smaller. Use 1/2 for tied observations. K_1 is the sum of these ranks.
    - K_2: for each observation in group 2, count the number of observations in the first group that are smaller. Use 1/2 for tied observations. K_2 is the sum of these ranks.
    - CHECK: if you have done the procedure correctly $K_1 + K_2 = n_1n_2$
  - Step 3: If the test is non-directional then U_s is the larger of K_1 and K_2. If the test is directional then U_s is the K that jives with the direction of H_a (**if H_a is $Y_1>Y_2$ then $U_s = K_1$**, **if H_a is $Y_1<Y_2$ then $U_s = K_2$**)
  - Step 4: Determine the critical value
    - n = larger of n_1 and n_2
    - n' = smaller of n_1 and n_2
  - Step 5: Bracket the p-value

The Wilcoxon-Mann-Whitney

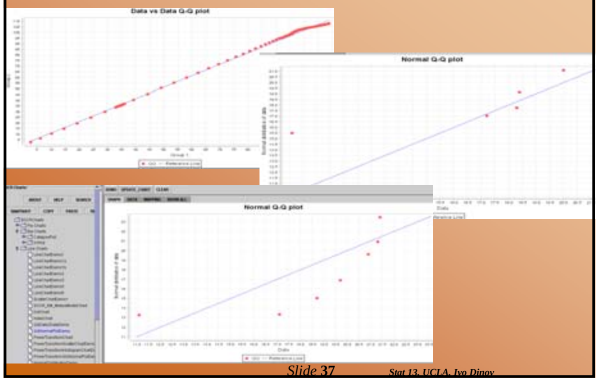
Example: The urinary fluoride concentration (ppm) was measured both for a sample of livestock grazing in an area previously exposed to fluoride pollution and also for a similar sample of livestock grazing in an unpolluted area.

Polluted	Unpolluted
21.3	10.1
18.7	18.3
21.4	17.2
17.1	18.4
11.1	20.0
20.9	
19.7	



Does the data suggest that the fluoride concentration for livestock grazing in the polluted region is larger that for the unpolluted region? Test using α = 0.01.

The Wilcoxon-Mann-Whitney – Normality Check
http://socr.ucla.edu/htmls/SOCR_Charts.html



The Wilcoxon-Mann-Whitney

#1 The hypotheses:

H_o: urinary fluoride values do not differ between the polluted and unpolluted regions.

H_a: the polluted region has a higher livestock urinary fluoride than the unpolluted region.

#2 The test statistic:

For this we need to deploy the WMW method shown a few slides earlier.

The Wilcoxon-Mann-Whitney

Let Polluted be group 1, and Unpolluted be group 2

Step 1: arrange the data in increasing order

# Unpolluted Below	Polluted N_1=7	Unpolluted N_2=5	# Polluted Below
1	11.1	10.1	0
1	17.1	17.2	2
4	18.7	18.3	2
4	19.7	18.4	2
5	20.9	20.0	4
5	21.3		
5	21.4		
K_1 = 25			K_2 = 10

Step 2: Determine K_1 and K_2

CHECK: 25 + 10 = 35 = (7)(5)

The Wilcoxon-Mann-Whitney

Step 3: H_a: Polluted (Y_1) > Unpolluted (Y2) so U_s is K_1

$K_1 = 25 = U_s$

$K_2 = 10$

Step 4:

n = 7

n' = 5

α = 0.01

#3 The p-value:

0.2 > p > 0.1

The Wilcoxon-Mann-Whitney

#4 CONCLUSION: These data show that the urinary fluoride concentration ppm for livestock grazing in polluted region is not greater than in the unpolluted region (P>0.1)

NOTE: No mention of the population means!

Corresponding Minitab output:

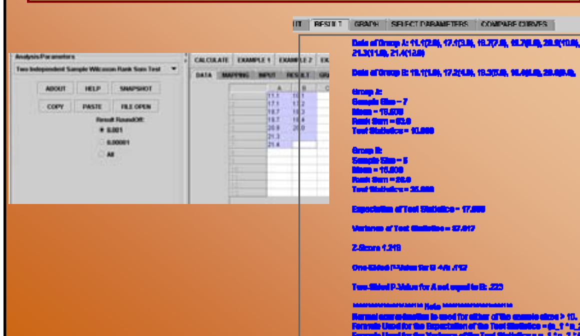
```
Mann-Whitney Test and CI: Polluted, Unpolluted
              N   Median
Polluted      7   19.700
Unpolluted    5   18.300

Point estimate for ETA1-ETA2 is 1.400
96.5 Percent CI for ETA1-ETA2 is (-2.897,8.602)
W = 53.0
Test of ETA1 = ETA2 vs ETA1 > ETA2
is significant at 0.1278
```

Test statistic is calculated using a different formula than our text, but W is the test statistic from the output

The p-value is calculated using the computer, but is not labeled well.

The Wilcoxon-Mann-Whitney
SOCR Analysis: http://socr.ucla.edu/htmls/SOCR_Analyses.html



The Wilcoxon-Mann-Whitney
SOCR Analysis: http://socr.ucla.edu/htmls/SOCR_Analyses.html

Data of Group A: 11.1(2.0), 17.1(3.0), 18.7(7.0), 19.7(8.0), 20.9(10.0), 21.3(11.0), 21.4(12.0)
Data of Group B: 10.1(1.0), 17.2(4.0), 18.3(5.0), 18.4(6.0), 20.0(9.0),

Group A:
Sample Size = 7
Mean = 18.600
Rank Sum = 53.0
Test Statistics = 10.000

Group B:
Sample Size = 5
Mean = 16.800
Rank Sum = 25.0
Test Statistics = 25.000

Expectation of Test Statistics = 17.500
Variance of Test Statistics = 37.917
Z-Score 1.218
One-Sided P-Value for B < A: 0.112
Two-Sided P-Value for A not equal to B: .223
*********************** Note ***********************
Normal approximation is used for either of the sample sizes > 10.
Formula Used for the Expectation of the Test Statistics = (n_1 * n_2)/ 2
Formula Used for the Variance of the Test Statistics = n_1 * n_2 * (n_1 + n_2 + 1) / 12

The Wilcoxon-Mann-Whitney

Practice:

Say n = 7, n' = 5 and $U_s = 32$

Two-tailed p-value: $0.01 < p < 0.02$

One-tailed p-value: $0.005 < p < 0.01$

Say n = 7, n' = 5 and $U_s = 36$

Impossible, for these sample sizes U_s cannot be larger than 35!

The Wilcoxon-Mann-Whitney

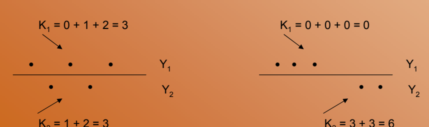
- Why does this procedure make sense?

Suppose $n_1 = 3$ and $n_2 = 2$

$K_1 + K_2 = (3)(2) = 6$

we know that $K_1 + K_2$ should sum to 6

The relative magnitudes of K_1 and K_2 indicate the overlap in Y_1 and Y_2



The Wilcoxon-Mann-Whitney

- Conditions for the WMW:
  - Data are from random samples
  - Observations are independent
  - Samples are independent
- Remember: normality will not matter for this test

Wilcoxon-Mann-Whitney vs. Independent Test

- Both answer the same question, but treat data differently.
  - W-M-W uses rank ordering
    - Positive: doesn't depend on normality or population parameters
    - Negative: distribution free lacks power because it doesn't use all the info in the data
  - T-test uses actual Y values
    - Positive : Incorporates all of the data into calculations
    - Negative : Must meet normality assumption
  - neither is superior
  - If your data are normally distributed use the T-test
  - If your data are not normal use the WMW test
- http://www.socr.ucla.edu/htmls/SOCR_Analyses.html