Application

Example: Nine observations of surface soil pH were made at 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.

<table>
<thead>
<tr>
<th>Location 1</th>
<th>Location 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.10</td>
<td>7.85</td>
</tr>
<tr>
<td>7.89</td>
<td>7.05</td>
</tr>
<tr>
<td>8.00</td>
<td>7.30</td>
</tr>
<tr>
<td>7.85</td>
<td>7.15</td>
</tr>
<tr>
<td>8.01</td>
<td>7.54</td>
</tr>
<tr>
<td>7.82</td>
<td>7.27</td>
</tr>
<tr>
<td>7.99</td>
<td>7.50</td>
</tr>
<tr>
<td>7.80</td>
<td>7.23</td>
</tr>
<tr>
<td>7.93</td>
<td>7.41</td>
</tr>
</tbody>
</table>

Application (cont’)

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 (cont’)

• #1 Formulate hypotheses
  
  $H_0$: $\mu_1 = \mu_2$ (there is no difference between the true mean soil pH of location1 and location2)
  
  $H_a$: $\mu_1 \neq \mu_2$ (there is a difference between the true mean soil pH of location1 and location2)

Application (cont’)

• #2 Calculate the test statistic

$$SE_{\bar{x}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{0.1005^2 + 0.2220^2}{9 + 9 - 2}} = 0.081$$

$$t_y = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{x}}} = \frac{7.9322 - 7.4600}{0.081} = 5.827$$

Application (cont’)

• #3 Calculate the p-value

$$p = \frac{2(0.0005)}{2} = 0.001$$
Application (cont’)

#4 Conclusion

Because $p < 0.001 < 0.05$, we will reject $H_0$.

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 (cont’)

• Corresponding Minitab output:

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

Two-sample T for Location 1 vs Location 2

<table>
<thead>
<tr>
<th></th>
<th>Location 1</th>
<th>Location 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Mean</td>
<td>7.932</td>
<td>7.460</td>
</tr>
<tr>
<td>StDev</td>
<td>0.100</td>
<td>0.222</td>
</tr>
<tr>
<td>SE</td>
<td>0.033</td>
<td>0.074</td>
</tr>
</tbody>
</table>

Difference = x̄ (Location 1) - x̄ (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

Application (cont’)

• Confidence interval for $\mu_1 - \mu_2$

- Suppose we calculated a 95% confidence interval to be:

$$(-0.472 \pm 2.030(0.081)) = (0.294, 0.650)$$

- Does this interval surprise you?

CI and HT relationship (cont’)

• With a t test we reject $H_0$ if the p-value is less than $\alpha$
then we reject $H_0$, and fail to reject otherwise

**this is the same thing as saying we reject if $t_0$ is beyond $t_{\alpha/2}$ and fail to reject otherwise**

CI and HT relationship (cont’)

- Further manipulation gives us:

$$(-\bar{y}_1 + \bar{y}_2 - \mu_0) \div \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

- Therefore, we fail to reject $H_0: \mu_1 = \mu_2$ (for the not equal to alternative), if the confidence interval contains 0.
CI and HT relationship (cont')

• 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 \( \mu_1 \) and \( \mu_2 \)
    • T test: has p-value which describes the strength of evidence that \( \mu_1 \) and \( \mu_2 \) are really different.

More on the significance level \( \alpha \)

• We need to choose a significance level BEFORE we analyze the data

Example: Say df = 15 and \( \alpha = 0.05 \)

• If \( t_s \) is in either tail we will reject \( H_0 \). The chance of this happening is 0.05 -- \( P(\text{reject } H_0) = 0.05 \), if \( H_0 \) is true.
• Because we are assuming that \( H_0 \) 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 \( H_0 \)
  - 2.5% would reject \( H_0 \) (-\( t_s \))
  - 2.5% would reject \( H_0 \) (\( t_s \))

In other words, a total of 5% would reject \( H_0 \) when \( H_0 \) is actually true. This is an incorrect conclusion just because of sampling error!

More on the significance level \( \alpha \) (cont')

• When we are analyzing one data set in real life at the 0.05 level and our conclusion is to reject \( H_0 \) there are two possible scenarios:
  1. \( H_0 \) is in fact false
  2. \( H_0 \) 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 \( H_0 \) and \( H_0 \) was true
    • \( P(\text{type I error}) = \alpha \)
    • When we choose \( \alpha \) 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 \( H_0 \) and \( H_0 \) is false
    • \( P(\text{type II error}) = \beta \)
    • \( \beta \) 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 (cont')

• Table (below) is the best way to summarize

<table>
<thead>
<tr>
<th>Reality</th>
<th>( H_0 ) True</th>
<th>( H_0 ) False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision</td>
<td>Correct</td>
<td>Correct</td>
</tr>
</tbody>
</table>

• 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 FTR conclusion)

Type I and Type II Errors (cont')

• Analogy: Think of a car with a car alarm being broken into
  – If the alarm goes off for no reason (reject \( H_0 \) when \( H_0 \) is true) - type I error
  – If the car gets broken into and the alarm doesn't go off (fail to reject \( H_0 \) when \( H_0 \) is false) - type II error
  – Also consider the sensitivity of the alarm
    • REMEMBER: FTR reject \( H_0 \) 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 (cont')

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.
Ho: No significant pollution
Ha: Significant pollution

In this case a type II error would be much worse (probability of failing to reject Ho when Ho 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 $\beta$ is small

Type I and Type II Errors (cont')

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%.

Ho: No significant response (Proportion responding to TX is $\leq 0.5$)
Ha: Significant response (Proportion responding to TX is $> 0.5$)

In this case a type I error is much worse (probability of rejecting Ho when Ho is true – like saying that the TX does something when it really doesn’t)
An ethical researcher would want to make sure they keep $\alpha$ small before collecting and analyzing the data

Type I and Type II Errors (cont')

• Because $\alpha$ 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)
• $\beta$ = P(fail to reject Ho) when Ho is false.
  – The chance of rejecting Ho when it is actually false is called the power of our test
    • Power = 1 - $\beta$ = P(reject Ho) when Ho 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 Ho was rejected if $t_s$ fell in either tail
  – In some analyses it is reasonable that there will be a certain direction of a deviation from Ho
    – 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:
    Ho: $\mu_1 - \mu_2 = 0$
    Ha: $\mu_1 - \mu_2 < 0$
    OR
    Ha: $\mu_1 - \mu_2 > 0$
    – Note: the null hypothesis doesn’t change

One Tailed t Tests (cont’)

Example: 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
Ho: $\mu_1 - \mu_2 = 0$
Ha: $\mu_1 - \mu_2 < 0$

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 Ha.
One Tailed t Tests (cont’)

Step 2: 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 (cont’)

Example: 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

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location 1</td>
<td>9</td>
<td>7.932</td>
<td>0.100</td>
</tr>
<tr>
<td>Location 2</td>
<td>9</td>
<td>7.460</td>
<td>0.222</td>
</tr>
</tbody>
</table>

Difference = mu (Location 1) - mu (Location 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

The Wilcoxon-Mann-Whitney

• WMW
  • Also known as the rank sum test
  • 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 meet the required normality assumption of the t test
  • WMW is called a “distribution-free” type of test or a 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 (cont’)

• Keep in mind that this is another hypothesis test, so there are still four major parts to consider
  • #1 The hypotheses:
    – Ho: The population distributions of Y1 and Y2 are the same
    – Ha: The population distributions of Y1 and Y2 are the different
      • This could also be directional: distribution of Y1 is less than Y2; OR distribution of Y1 is greater than Y2
  • #2 The test statistic:
    – denoted by U
    – measures the degree of separation between the two samples
      • a large value of U indicates that the two samples are well separated with little overlap
      • a small value of U indicates that the two samples are not well separated with much overlap

The Wilcoxon-Mann-Whitney (cont’)

• #3 The p-value:
  – New table!
  – Critical Values are 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 (cont')

• 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.
    * \( K_2 \): for each observation in group 2, count the number of observations in the first group that are smaller. Use 1/2 for tied observations.
  – CHECK: if you have done the procedure correctly \( K_1 + K_2 = n_1 n_2 \)
  – 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

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 than for the unpolluted region? Test using \( \alpha = 0.01 \).

Check Normality:

Let Polluted be group 1, and Unpolluted be group 2

Step 1: arrange the data in increasing order

\[
\begin{array}{c|c|c|c}
\text{# Unpolluted Below} & \text{Polluted} & \text{Unpolluted} & \text{# Polluted Below} \\
1 & 11.1 & 10.1 & 0 \\
1 & 17.1 & 17.2 & 2 \\
4 & 18.7 & 18.3 & 2 \\
5 & 19.7 & 18.4 & 2 \\
5 & 20.9 & 20.0 & 4 \\
5 & 21.3 & 21.3 & 4 \\
5 & 21.4 & 21.4 & 4 \\
\end{array}
\]

\( K_1 = 25 \quad K_2 = 10 \)

Step 2: Determine \( K_1 \) and \( K_2 \)

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

Step 3: \( H_a: \) Polluted (\( Y_1 \)) > Unpolluted (\( Y_2 \)) so \( U_s \) is \( K_1 \)

\( K_1 = 25 = U_s \)

\( K_2 = 10 \)

Step 4:

\( n = 7 \)
\( n' = 5 \)
\( \alpha = 0.01 \)

Step 5: Bracket the p-value

\( p > 0.1 \)
The Wilcoxon-Mann-Whitney (cont’)

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

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Polluted</td>
<td>7 19.700</td>
<td></td>
</tr>
<tr>
<td>Unpolluted</td>
<td>8 18.300</td>
<td></td>
</tr>
</tbody>
</table>

Point estimate for ETA1-ETA2 is 1.400

96.5 Percent CI for ETA1-ETA2 is (-2.897, 8.602)

W = 53.0

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.

38

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 (cont’)

• 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

  K_1 = 0 + 1 + 2 = 3
  K_2 = 0 + 0 + 0
  ∙ ∙ ∙
  Y_1
  ∙ ∙ ∙
  Y_1
  K_1 = 3 + 2 + 1

  K_2 = 0 + 3 + 0
  ∙ ∙ ∙
  Y_2
  ∙ ∙ ∙
  Y_2

The WMW vs. Independent T

• Both try to answer the same question, but treat data differently.

  – WMW 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

  • So...

    – If your data are normally distributed use the t-test
    – If your data do not meet the normally assumption use the WMW test

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