Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 21 (MWF) The paired t-test for dependent ‘paired’ samples

Suhasini Subba Rao
Previous lecture: Comparing populations

We have two populations which have the same variance. Suppose $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are samples from populations 1 and 2 respectively, our aim is to make inference about the mean difference $\mu_X - \mu_Y$ based on the sample populations.

$$H_0 : \mu_X - \mu_Y = 0 \text{ against } H_A : \mu_X - \mu_Y \neq 0.$$  

• If the sample sizes $n$ and $m$ are relatively large say both over 20-30 (depending on how close to normal the original distribution) and the standard deviations of both populations are almost the same (spread of both samples are about the same), then we construct confidence intervals for the mean difference, do an independent sample t-test (test $H_0 : \mu_X - \mu_Y = 0 \text{ against } H_A : \mu_X - \mu_Y \neq 0$ or $H_0 : \mu_X - \mu_Y \geq -0.3$ against $H_A : \mu_X - \mu_Y < -0.3$ (for example)).

• On the other hand, if the sample sizes $n$ and $m$ are quite small, or the
sample has many outliers, we need to use a (nonparametric) Wilcoxon rank sum test.

The Wilcoxon rank sum test is a nonparametric test. This means we cannot use it to construct CI or test $H_0 : \mu_X - \mu_Y \geq -0.3$ against $H_A : \mu_X - \mu_Y < -0.3$ (test shifts away from zero).

But we can test whether the distribution of one population is a shift of the other, hence test $H_0 : \mu_X - \mu_Y = 0$ against $H_A : \mu_X - \mu_Y \neq 0$ or $H_0 : \mu_X - \mu_Y \leq 0$ against $H_A : \mu_X - \mu_Y > 0$. In other words, $H_0 :$ distributions of $X$ and $Y$ are the same. $H_A :$ distribution of $X$ and $Y$ are a shift of each other.

(in this case we need to distribution of both populations have the same shape).

For small sample sizes, we need to use the t-distribution to construct confidence intervals or do tests that do not involve zero, ie. $H_0 :$
\[ \mu_X - \mu_Y \geq -0.3 \] against \( H_A : \mu_X - \mu_Y < -0.3 \). However, we need to keep in mind that the CI and the results of the test may not be accurate.
Comparing population, when the independence assumptions are not satisfied: Friday the 13th

- Is Friday the 13th an unusually unlucky day, or is this just another superstition? Does the behaviour of people change on this day?

- Researchers (Scanlon, et al. (1993)) analysed accident patterns on past Friday the 13ths.

- To make the comparison fair they compared the data with what happened the previous week - Friday 6th.

Why do you think they chose this date to make the comparison?

- Answer to make the fair comparison. They need to compare what happens on the 13th with another date where almost all the factors are the same, except the 13th.
Thursday the 12th is not suitable because, patterns on Thursday could be different to those on Friday. In the UK more accidents happen on Friday and Saturday compared to other days of the week. A completely different time is not fair because more accidents may happen in one season than another. The previous Friday seemed the closest timewise.

The design of an experiment is extremely important!

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>6th</th>
<th>13th</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989</td>
<td>October</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>1990</td>
<td>July</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>1991</td>
<td>September</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>1991</td>
<td>December</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>1992</td>
<td>March</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1992</td>
<td>November</td>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>
Friday 13th: Method 1: The independent t-test

- Let $\mu_{13}$ be the mean number of accidents which happen on Friday the 13th and $\mu_6$ the mean number of accidents which happen on Friday the 6th. We want to test (with $\alpha = 0.05$)
  
  - $H_0 : \mu_{13} - \mu_6 \leq 0$
  - $H_A : \mu_{13} - \mu_6 > 0$.

- The sample mean for the 6th is $\bar{X} = \frac{1}{6} \sum_{i=1}^{6} X_i = 7.5$ and the sample mean for the 13th is $\bar{Y} = \frac{1}{6} \sum_{i=1}^{6} Y_i = 10.83$. 
THE JMP output

Oneway Analysis of Column 1 By Column 2

Column 1

Column 2

1.0
Assuming unequal variances
Difference  3.3333  t Ratio  1.664357
Std Err Dif  2.0028  DF  9.940201
Upper CL Dif  7.7394  Prob > U  0.1272
Lower CL Dif  -1.1328  Prob > L  0.0636
Confidence  0.95  Prob < t  0.9364
**Friday 13th: Method 2: The Wilcoxon sum rank test**

- The null hypothesis: $H_0$: The populations are identical (the distribution of accidents on Friday 13th and other Fridays in the same). The alternative hypothesis $H_A$: The distribution of accidents on Friday 13th is a right shift of the distribution of accidents on other Fridays.

<table>
<thead>
<tr>
<th>Sam. 1</th>
<th>Sam. 2</th>
<th>Order Sam. 1</th>
<th>Order Sam. 2</th>
<th>Rank 1</th>
<th>Rank 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>13</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>5</td>
<td>10</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>6</td>
<td>12</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>9</td>
<td>12</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>11</td>
<td>13</td>
<td>7.5</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>11</td>
<td>14</td>
<td>7.5</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>T=28</td>
<td></td>
</tr>
</tbody>
</table>
We choose $T = 28$ (which corresponds to the ‘other Fridays’).

Check tables: with $\alpha = 5\%$ we get the upper bound 28.

Since 28 is not less than 28, the p-value is exactly 5% and there is not enough evidence to reject $H_0$ at $\alpha = 5\%$.

This gives us the same result as using the t-test.

May be that's it. There does not appear to be enough evidence to suggest that peoples behaviour is different on the 13th?
Lecture 21 (MWF) The paired t-test

**Oneway Analysis of Column 1 By Column 2**

**t Test**

- Assuming unequal variances
  - Difference: 3.3333, t Ratio: 1.664357
  - Std Err Dif: 2.0028, DF: 9.940201
  - Upper CL Diff: 7.7994, Prob > t: 0.1272
  - Lower CL Diff: -1.1328, Prob > t: 0.0636
  - Confidence: 0.95, Prob < t: 0.9364

**Wilcoxon / Kruskal-Wallis Tests (Rank Sums)**

<table>
<thead>
<tr>
<th>Level</th>
<th>Count</th>
<th>Score Sum</th>
<th>Expected Score</th>
<th>Score Mean</th>
<th>(Mean-Mean0)/Std0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>28.000</td>
<td>39.000</td>
<td>4.66667</td>
<td>-1.687</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>50.000</td>
<td>39.000</td>
<td>8.33333</td>
<td>1.687</td>
</tr>
</tbody>
</table>

**2-Sample Test, Normal Approximation**

| S     | Z     | Prob>|Z| |
|-------|-------|------|
| 50    | 1.68726| 0.0916|

**1-way Test, ChiSquare Approximation**

| ChiSquare | DF | Prob>|ChiSq| |
|-----------|----|------|---|
| 3.1244    | 1  | 0.0771| |
• There is a lot evidence that peoples behaviour changes on so called ‘bad’ days. But our tests so far does not support this.

• Why not? It could be that we do not have enough data.

May be, but it could be something more fundamental: some of the assumptions are being violated and we are not using the correct test.
Did friday 13th data satisfy the usual assumptions?

• Lets look at the data again and consider how it was collected.

<table>
<thead>
<tr>
<th></th>
<th>6th</th>
<th>13th</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>diff</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

• From the differences it appears plausible that the number of accidents on the 13th are more than those on the 6th.

• We have done both the t-test on the data and the Wilcoxon rank sum test, for both tests we do not find enough evidence to reject the null (contradicting what we see in the data).
• But recall the assumptions. Do you think the two data sets are independent? Within each sample we can assume data is close to independent, the data was taken at different times of the year. But do you think each pair is independent and that the data is identically distributed?

• Its highly plausible that more accidents happen at certain times of the year. For example, there are more accidents in December than September.
Plot of Friday 13th against Friday 6th

There seems to be a trend; if the number of accidents on the 6th is large it is followed by a large number of accidents on the 13th.
Problems with paired data

• This is a plot of the number of accidents which happen on the 6th against those which happen on the 13th.

• As each 6th of the month can be paired with the following 13th of the month. We called this paired data.

• Looking at the plot, we see that the variation within each sample (6th and 13th) is a lot.

  In particular, the range of each sample - they are both wide. So any significant difference between the sample means is ‘swamped’ by the variability in the data. The huge range gives us a large standard deviation and thus a non-rejection region which is large. Making it hard for us to reject the null.
• Recall for the t-test it is hard to reject the null, if the variance is large.
• We see that in general if the number of accidents on the 6th are high the number of accidents on the following 13th will also be high.

Why? The number of accidents really depend on the season. For example more accidents may happen around Christmas because of the large number of parties.

• In such a case it is also hard to reject the null doing a Wilcoxon sum rank test. Because pairs of observations tend to be clustered so it is hard to see a ‘nice’ groupings in the ranks.

• What does this tell us about our assumptions for doing an independent test?

The main thing is that the data on the 6th and the following 13th are not independent, that is the paired data \((X_i, Y_i)\) are not independent.
• Because of non-independence between the pairs in the sample, the test as it stands is not appropriate. Instead, we need to do a matched paired t-test.
The paired t-test is used when there is dependence within each pair of observations

When there is a dependence between the pairs take the difference between each pair, and define a new random variable $D_i = X_i - Y_i$. By taking the difference we are reducing the amount of variation (reduce the variance), which makes it easier to detect an effect.

<table>
<thead>
<tr>
<th>6th</th>
<th>9</th>
<th>6</th>
<th>11</th>
<th>11</th>
<th>3</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>13th</td>
<td>13</td>
<td>12</td>
<td>14</td>
<td>10</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>$D_i$</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

From a plot of the differences, is it likely this data could have arisen if there are no differences (or there is a tendency for less accidents to happen on the 13th).
By considering the difference $D_i$ rather than the individual $X_i$ and $Y_i$ we remove a common factors between the pair - for example, more accidents happen in December, or that some individuals have a baseline ability at sport (while other do not) - taking the difference removes this, and only looks at the change.

Now we treat $D_i$ as a new random variable.

The sample mean is $\frac{1}{6}(4 + 6 + 3 - 1 + 1 + 7) = 3.33$.

The sample standard deviation of the difference is

$$s_d = \sqrt{\frac{1}{5}((-4 + 3.33)^2(-6 + 3.33)^2(-3 + 3.33)^2 + (1 + 3.33^2)(-1 + 3.33)^2(-7 + 3.33)^2} = \sqrt{9.01}.$$  

If $X_i$ and $Y_i$ come from the same population then the mean of $D_i$ is 0. Let $\mu_d = \mu_X - \mu_Y$ be the population mean of $D_i$. 

20
The paired t-test

• To test whether the means of $\mu_X$ and $\mu_Y$ are the same is equivalent to testing $H_0 : \mu_d = \mu_{13} - \mu_6 \leq 0$ against $H_A : \mu_d = \mu_{13} - \mu_6 > 0$.

• Now do the ‘usual’ t-test using the difference observations $D_i$ with $(4, 6, 3, -1, 1, 7)$.

• Use the average difference $\bar{D} = \frac{1}{6} \sum_{i=1}^{6} D_i = 3.33$ and the sample standard deviation of the differences: $s = \sqrt{9.01}$.

• We now do the z-transform under the null that $H_0 : \mu_d = 0$. That is, under the null $\bar{D} \sim N(0, \frac{\sigma^2}{n})$. Since we have to estimate the standard
deviation, we evaluate

\[
\frac{\bar{D}}{\sqrt{\frac{s^2}{n}}} \sim t(n - 1).
\]

We now evaluate the t-transform:

\[
t = \frac{\bar{D}}{\sqrt{\frac{s^2}{n}}} = \frac{3.33}{\sqrt{\frac{9.01}{6}}} = 2.717.
\]

• Under the null hypothesis that both means are the same \( \frac{\bar{D}}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1} \) (it is a t-distribution with \( n - 1 \) degrees of freedom - where \( n \) are the number of pairs). In our example, under the null it has \( t(5) \)-distribution.
• Look up $t_{0.025}(5) = 2.015$ (because there are 6 pairs, $n = 6$). Since $2.717 > 2.57$, there is enough evidence to reject the null.

• Alternative (but equivalently) calculate the p-value (hard without a computer!) Using 2.7 the p-value is:

$$P(t(5) \leq 3.37/\sqrt{9.01/6}) = P(t(5) \leq 2.717) = 0.0209.$$  

We see that 0.0209 is less than 0.05, hence there is enough evidence to reject the null hypothesis with $\alpha = 0.05$.

• Either way there is enough evidence to reject the null.
Friday 13th Data: The matched t-test in JMP

Matched Pairs

Difference: 13th-6th

<table>
<thead>
<tr>
<th>Difference</th>
<th>10.8333</th>
<th>t-Ratio</th>
<th>2.711631</th>
</tr>
</thead>
<tbody>
<tr>
<td>6th</td>
<td>7.5</td>
<td>DF</td>
<td>5</td>
</tr>
<tr>
<td>Mean Difference</td>
<td>3.33333</td>
<td>Prob &gt;</td>
<td>0.0422*</td>
</tr>
<tr>
<td>Std Error</td>
<td>1.22927</td>
<td>&gt; t</td>
<td>0.0211*</td>
</tr>
<tr>
<td>Upper 95%</td>
<td>6.49328</td>
<td>Prob &lt;</td>
<td>0.9789</td>
</tr>
<tr>
<td>Lower 95%</td>
<td>0.17339</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td>0.62515</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We already came across this test in the Red wine and Polyphenol example!

The paired t-test is very obvious. In fact we came across it in the wine and polyphenol example considered in Lecture 16. Recall that we wanted to test whether polyphenol levels had increased after taking red wine for two weeks. We tested the hypothesis $H_0 : \mu \leq 0$ against $H_A : \mu > 0$ using the data:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-B</td>
<td>0.7</td>
<td>3.5</td>
<td>4.9</td>
<td>5.5</td>
<td>7</td>
<td>7.4</td>
<td>8.1</td>
<td>8.4</td>
<td>3.2</td>
<td>0.8</td>
<td>4.3</td>
<td>-0.2</td>
<td>-0.6</td>
<td>7.5</td>
<td></td>
</tr>
</tbody>
</table>
We choose the p-value pointing to the right, which is less than 0.001 (less than 0.1%), therefore there is evidence to reject the null (the data suggests that drinking wine increases average polyphenol levels).
A paired t-test in disguise!
What exactly is \( \mu \) and how exactly was the data collected?

- \( \mu \) is the population mean *difference* in polyphenol levels before and after taking red wine.

- The data has been ‘processed’, it is the difference in the polyphenol level for each individual, before and after they took the red wine. The pre-processed data looks like this:

- There is a clear ‘matching’ between the before and after observations (since the involve the same individual).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>42.7</td>
<td>43.6</td>
<td>42.9</td>
<td>43.3</td>
<td>36.8</td>
<td>38.6</td>
<td>37.3</td>
<td>37.9</td>
<td>41.9</td>
<td>43.0</td>
<td>40.2</td>
<td>32.4</td>
<td>31.8</td>
<td>36.7</td>
<td>45.0</td>
</tr>
<tr>
<td>A</td>
<td>43.4</td>
<td>47.1</td>
<td>46.9</td>
<td>48.2</td>
<td>42.3</td>
<td>45.6</td>
<td>44.7</td>
<td>46.0</td>
<td>50.3</td>
<td>46.2</td>
<td>41.0</td>
<td>36.7</td>
<td>31.6</td>
<td>36.1</td>
<td>52.5</td>
</tr>
<tr>
<td>A-B</td>
<td>0.7</td>
<td>3.5</td>
<td>4</td>
<td>4.9</td>
<td>5.5</td>
<td>7</td>
<td>7.4</td>
<td>8.1</td>
<td>8.4</td>
<td>3.2</td>
<td>0.8</td>
<td>4.3</td>
<td>-0.2</td>
<td>-0.6</td>
<td>7.5</td>
</tr>
</tbody>
</table>
Plotting the differences

From a plot of the differences, is it likely this data could have arisen if red wine had not positive effect? What do you think the p-value would be for $H_0 : \mu_A - \mu_B \leq 0$ vs. $H_A : \mu_A - \mu_B > 0$?
The paired t-test using the unprocessed data

Of course the output for the matched paired t-test and the t-test of the differences give the same result - they are the same test!
A plot of each individual’s polyphenol concentration before and after shows a clear dependence (which is obvious since they are the same individual).
Example 1: Runners at altitude

Runners were compared at a high and low altitude. For each runner, the running time was measured at a high altitude and then again at a low altitude. 12 runners were used.

<table>
<thead>
<tr>
<th>Runner</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>9.4</td>
<td>9.8</td>
<td>9.9</td>
<td>10.3</td>
<td>8.9</td>
<td>8.8</td>
<td>9.8</td>
<td>8.2</td>
<td>9.4</td>
<td>9.9</td>
<td>12.2</td>
<td>9.3</td>
</tr>
<tr>
<td>Low</td>
<td>8.7</td>
<td>7.9</td>
<td>8.3</td>
<td>8.4</td>
<td>9.2</td>
<td>9.1</td>
<td>8.2</td>
<td>8.1</td>
<td>8.9</td>
<td>8.2</td>
<td>8.9</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Does altitude influence increase running time?
A plot

It is clear that there is a matching between the high and altitude times, because the same runner is doing both. This can be seen in the plot of high against low. However, it is worth bearing in mind, even if you do not see a clear linear dependence between the two, if there is a clear matching of data a matched test must be done.
Runners and the JMP output

The JMP output for the matched t-test is given below.
Analysing the JMP output

• We want to test $H_0 : \mu_L - \mu_H \geq 0$ against $H_A : \mu_L - \mu_H < 0$. We do the test at the 5% significance level.

• From the output it is clear the p-value for this test is 0.13%. As this is less than 5% there is evidence to suggest that running at high altitudes increases ones running times.

• Using the output we can obtain more information, for example the output gives the 95% confidence interval for the mean difference.

• The output can also be used to obtain confidence intervals at other levels, for example a 99% confidence interval.

• We can also use the output to test $H_0 : \mu_L - \mu_H \geq -0.8$ against $H_A : \mu_L - \mu_H < -0.8$ (ie. test on the magnitudes of the mean differences).
differences). To do this test we make a t-transform

\[ t = \frac{-1.2 - (-0.8)}{0.31} = -1.29. \]

Since the alternative is pointing to the left we use the t-tables to deduce the area to the left of -1.29.
Is the number of blue and yellow M&Ms different?

• To answer this question, it only makes sense to ask whether the mean number of blue and yellow M&Ms in a bag different?

• Here we are testing $H_0 : \mu_B - \mu_Y = 0$ against $H_A : \mu_B - \mu_Y \neq 0$.

• We collect a sample of 170 M&M bags and count the number of blues and yellows in each one. The average number of blues in the 170 bags is 2.71 and the average number of yellow is 2.01. This gives an average difference of 0.7 M&Ms between blues and yellows. Without any additional information on the reliability of these estimates, it is impossible to say whether 0.7 is sufficiently large to say that it is statistically significant and suggests that the overall means are different.

• This is when statistical inference comes in useful.
**M&M scatterplot**

Scatter plot of blue against yellow. A linear line is not clear, however, there is clear matching since a bag is common between the yellow and blues. Therefore we will use a matched paired t-test.
M&M JMP output

Matched Pairs

Difference: Yellow-Blue

Yellow 2.01765 t-Ratio -3.99582
Blue 2.71765 DF 169
Mean Difference -0.7 Prob > ltl <.0001*
Std Error 0.17518 Prob > t 1.0000
Upper 95% -0.3542 Prob < t <.0001*
Lower 95% -1.0458
N 170
Correlation 0.01796
The paired t-test

Analysing JMP output for M&M data

• If we test $H_0 : \mu_B - \mu_Y = 0$ against $H_A : \mu_B - \mu_Y \neq 0$ at the 5% level, we see from the output that the p-value is less than 0.01%, therefore there is strong evidence to reject the null and suppose that the mean number of blue and yellow M&Ms in a bag are different.

• With 95% confidence the mean differences lies in the interval $[-1.04, -0.35]$. So the mean difference lies anywhere between 0.3 M&M to one M&M (with 95% confidence).

• One could argue that using an independent sample t-test for the M&M data would have been more appropriate. If we had done that, and the blues and yellows are indeed independent we get a very similar answer (indeed the standard errors for both tests would have been quite similar). The only thing that would have been different is the number of degrees
of freedom. The matched pair t-test tends to have a lower number of degrees of freedom than an independent sample t, this makes it a little harder to reject the null if the alternative is true.
The full moon and disruptive patients

A hospital wants to know whether disruptive patients become even more disruptive during the full moon. They monitored disruptive patients over a period of time noting the average number of disruptive instances when there wasn’t a full moon and when there was a full moon. The data can be found on my website. There is a clear matching in this data set as the number of disruptive incidences of each patient is monitored.
Plot of moon against non-full moon

There isn’t a clear line (there rarely ever is), but we do see some ‘evidence’ of a linear trend.
**Moon JMP output**

**Matched Pairs**

**Difference: Other-Moon**

<table>
<thead>
<tr>
<th></th>
<th>Other</th>
<th>Moon</th>
<th>t-Ratio</th>
<th>DF</th>
<th>Prob &gt;</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
<th>Std Error</th>
<th>Prob &gt;</th>
<th>Prob &lt;</th>
<th>N</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-2.3033</td>
<td>3.13889</td>
<td>0.83556</td>
<td>17</td>
<td>.0001*</td>
<td>-3.027</td>
<td>-1.5796</td>
<td>0.34302</td>
<td>1.0000</td>
<td>.0001*</td>
<td>18</td>
<td>0.25111</td>
</tr>
</tbody>
</table>
The Analysis of Full moon data

• We are testing \( H_0 : \mu_{NM} - \mu_M \geq 0 \) against \( H_A : \mu_{NM} - \mu_M < 0 \).

• From the output we see that the sample mean difference is -2.3 and this difference is statistically significant with a p-value of less than 0.01%. This means there appears to be more disruptive events during the full moon compared to other times. As this is significant it suggests that more staff should be brought in.

• Next we need to decide on how many extra staff is required, and this depends on the mean number of additional disruptive events. The 95% confidence of \([-3.03, -1.57]\) tells us (with 95% confidence) how many more disruptive events there is likely to be. If you want to be cautious you may take the upper bound and use the upper bound for the mean of 3.03 extra disruptive events per person and calculate the additional staff.
based on this number. Else you may want to save money and base your calculation on 1.57.

- Of course, this is only a 95% confidence interval, if you want even more confidence you could use a 99% confidence interval 
  \[ [-2.3 - 2.89 \times 0.34, -2.3 + 2.89 \times 0.34] = [-3.28, -1.31]. \]
When to use an independent or a matched paired test

• The independent and matched paired t-test are testing the same hypothesis. The only difference is how the data is collected from both samples.

• If the two samples appear to be completely independent of each other do an independent test

• If there appears a natural ‘pairing’ in the data do a matched paired test, ie. The SAME person before and after a intervention (red wine example, full moon example, running high and low altitude example), the SAME month (Friday 13th example), the SAME bag of M&Ms.

• However, don’t be fooled by a spread sheet where there ‘appears’ to be a natural pairing but isn’t.
Example 10 randomly sampled people were put on a diet. 5 were randomly allocated to diet 1 and the other 5 to diet 2. These are two clearly independent samples. However, if the data is collected and displayed in two columns it may give the false impression of pairing, where no real pairing exists.

<table>
<thead>
<tr>
<th>Diet 1</th>
<th>Diet 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.6</td>
</tr>
<tr>
<td>2</td>
<td>2.2</td>
</tr>
<tr>
<td>3</td>
<td>3.1</td>
</tr>
<tr>
<td>4</td>
<td>6.6</td>
</tr>
<tr>
<td>5</td>
<td>1.2</td>
</tr>
</tbody>
</table>

- If there logically seems to be a pairing in the data but no clear line is seen in the scatterplot, doing a matched paired t-test still makes sense.

Your results will still be relatively reliable if you do a matched paired
t-test and the samples are completely independent. On the other hand, if you do an independent sample t-test when there is matching, the results will not be reliable. To those who like mathematics, this is because the matched paired t-test gives the correct standard error when there is completely independence. However the independent sample t-test does not give the correct standard error when there is pairing in the data.
For those who like to understand with mathematical models

- If you are familiar with mathematical notation and random variables, in a matched set-up each pairing satisfies the ‘model’

\[ X_{1i} = Z_i + \varepsilon_{1i} \quad \text{and} \quad X_{2i} = Z_i + \varepsilon_{2i} \]

- So when we are testing if the means in each population are the same, we are testing if the means of \( \varepsilon_{1i} \) and \( \varepsilon_{2i} \) are the same.

- Clearly, there is a dependence between the groups (since there is a dependence between these pairings), which violates the assumption of independence (in the independent sample t-test). The dependence discussed above can usually (but not always) be seen in plots of \( X_{1i} \) against \( X_{2i} \).
• Using the above model it is ‘clear’ that taking differences $X_{1i} - X_{2i}$ will eliminate the ‘nuisance’ $Z_i$ and allow us to directly test if the means $\varepsilon_{1i}$ and $\varepsilon_{2i}$ are the same.

• Note that if we used the independence t-test the standard error is $\sqrt{\frac{\sigma_Z^2 + \sigma_1^2}{n} + \frac{\sigma_Z^2 + \sigma_2^2}{n}}$, whereas if we use a paired t-test the standard error is $\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}$. The latter standard deviation is smaller, and it is this that makes the matched t-test test more powerful than the independent sample t-test covered in lecture 19 (power will be defined in a later lecture).