Biochemistry Prelims Statistics

Lecture III:

Perils of the $t$-test

Jack J. Miller, DPhil
jack.miller@dpag.ox.ac.uk

Hilary Term 2018

$p < 0.05 \Rightarrow$ I'm right!
Perils of the $t$-test
Last time...

- We talked quite a lot about sampling, and saw how the sample mean (and standard deviation) are fundamentally uncertain *estimates* of the population mean (and standard deviation).
- **If** the samples are drawn from a normal distribution, then the sample mean is related to the $t$ distribution, and we can use this to both place confidence limits on the location of the population mean (given the sample), and see if the sample mean is different from a specified value.
Last time...

- We talked quite a lot about sampling, and saw how the sample mean (and standard deviation) are fundamentally uncertain estimates of the population mean (and standard deviation).

- If the samples are drawn from a normal distribution, then the sample mean is related to the *t* distribution, and we can use this to both place confidence limits on the location of the population mean (given the sample), and see if the sample mean is different from a specified value.

- One can generalise this, and use *t*-tests to investigate whether or not samples drawn from two groups have sample means that are likely to be from different populations.
Interpretation

- A statistical test gives us a $p$-value that represents the probability of observing a test statistic at least as extreme as that observed assuming the null hypothesis is true.

- If this value is below some arbitrary cut-off level, which we call the significance level, we reject the null hypothesis (and usually conclude that the two groups have different means).
Errors we can always make with hypothesis testing

- Note that *by definition* the significance level means that *if* the null hypothesis is true, we will *still* be wrong some of the time.
Errors we can always make with hypothesis testing

- Note that *by definition* the significance level means that *if* the null hypothesis is true, we will *still* be wrong some of the time.
- This is analogous to a *false positive* result – and is also known as a **Type I error**.
Errors we can always make with hypothesis testing

- Note that *by definition* the significance level means that *if* the null hypothesis is true, we will *still* be wrong some of the time.
- This is analogous to a **false positive** result – and is also known as a **Type I error**.
- The rate at which we *will* have false positives is the significance level.
Errors we can always make with hypothesis testing

- The reverse is also true – we can miss a positive result.

This is analogous to a false negative result, and is known as a Type II error. This is harder to control, and is related to the statistical power of the test – if type II errors occur at a rate, then power is 1. I'll talk about power later, but it is a function of three variables:

- The sample sizes in each group
- The standard deviation of each group
- The type of test being used.
Errors we can always make with hypothesis testing

- The reverse is also true – we can miss a positive result.
- This is analogous to a **false negative** result, and is known as a **Type II error**.
Errors we can always make with hypothesis testing

- The reverse is also true – we can miss a positive result.
- This is analogous to a **false negative** result, and is known as a **Type II error**.
- This is harder to control, and is related to the **statistical power** of the test – if type II errors occur at a rate $\beta$, then power is $1 - \beta$. I’ll talk about power later, but it is a function of three variables:
  - The sample sizes in each group
  - The standard deviation of each group
  - The type of test being used.
Type I and Type II errors

<table>
<thead>
<tr>
<th>Null hypothesis actually is:</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$:</td>
<td>False positive</td>
<td>Correct inference</td>
</tr>
<tr>
<td>Accept $H_0$:</td>
<td>Correct inference</td>
<td>False negative</td>
</tr>
</tbody>
</table>
Multiple comparisons

Previously we’ve only looked at comparing two groups with a $t$-test.

In reality, we often want to compare more than two groups – sometimes a small number of groups (e.g. what happens if I replace valine at a particular location by either leucine, isoleucine, alanine or glycine in my protein of interest?) and sometimes a positively huge number (e.g. microarrays).

One way to proceed is by doing lots of $t$-tests pairwise – we’ll cover more general methods for examining groups of many members (ANOVA) next year.
Multiple comparisons

*Every* time we do a test, we run the risk of a Type I error.
Multiple comparisons

*Every* time we do a test, we run the risk of a Type I error.
Multiple comparisons

_Every_ time we do a test, we run the risk of a Type I error.
Multiple comparisons

*Every* time we do a test, we run the risk of a *Type I* error.

If each failure is independent, I can’t cycle home about

\[ 1 - (0.98)^7 \approx 13\% \] of weekdays!

(Increasing 2\% to 5\% makes this \( \sim 30\% \))
Multiple comparisons

- Doing lots of tests *quickly* generates plenty of opportunity for this to happen.
Multiple comparisons

- Doing lots of tests *quickly* generates plenty of opportunity for this to happen.
- In $k$ groups, there are $\frac{k(k-1)}{2}$ different possible pairwise tests.

\[ \frac{k(k-1)}{2} \]
Multiple comparisons

- Doing lots of tests *quickly* generates plenty of opportunity for this to happen
- In $k$ groups, there are $\frac{k(k-1)}{2}$ different possible pairwise tests.
- E.g., in 7 groups there are 21 different possible tests.
Multiple comparisons

- Doing lots of tests *quickly* generates plenty of opportunity for this to happen.
- In $k$ groups, there are $\frac{k(k - 1)}{2}$ different possible pairwise tests.
- E.g., in 7 groups there are 21 different possible tests.
- If we pick $\alpha = 0.05$, we’d therefore expect $21 \times 0.05 \approx 1$ of them to be a false positive.
Correcting for multiple comparisons

This is clearly a problem. One way it can be dealt with was originally proposed by an Italian mathematician – Carlo Emilio Bonferroni (1892 – 1960) — who worked on many different problems in probability.
Correcting for multiple comparisons

This is clearly a problem. One way it can be dealt with was originally proposed by an Italian mathematician – Carlo Emilio Bonferroni (1892 – 1960) — who worked on many different problems in probability.

He’s probably best known for Bonferroni’s correction for multiple comparisons:
Correcting for multiple comparisons

This is clearly a problem. One way it can be dealt with was originally proposed by an Italian mathematician – Carlo Emilio Bonferroni (1892 – 1960) — who worked on many different problems in probability.

He’s probably best known for **Bonferroni’s correction** for multiple comparisons:

If I do $m$ hypothesis tests, then to control the rate of Type I error, I **divide my p-value by** $m$. 
Correcting for multiple comparisons

It turns out that this method is \textbf{somewhat conservative} – there are other methods that are slightly more complex but are more powerful:

\begin{itemize}
  \item Holm-Sidak aka Holm-Bonferroni which tests the lowest $p$ value in a group most strictly, and others progressively less so.
  \item Hochberg’s and Hommel’s methods, which are valid when the hypothesis tests are independent. Hommel’s method is more powerful than Hochberg’s, but the difference is usually small and the Hochberg $p$-values are faster to compute.
\end{itemize}
Fortunately...

In R, everything is done with `pairwise.t.test()`, which offers a variety of correction methods (e.g. `pairwise.t.test(data, groups, p.adj="bonferroni")`)
Assumptions specific to the $t$-test

- Every $t$-test has a set of assumptions behind it:
Assumptions specific to the $t$-test

- Every $t$-test has a set of assumptions behind it:
  - The underlying populations are normally distributed and are continuous
Assumptions specific to the $t$-test

- Every $t$-test has a set of assumptions behind it:
  - The underlying populations are normally distributed and are continuous
  - The samples are drawn randomly from these populations and each observation is independent
Assumptions specific to the $t$-test

- Every $t$-test has a set of assumptions behind it:
  - The underlying populations are normally distributed and are continuous
  - The samples are drawn randomly from these populations and each observation is independent
  - These populations have the same (Student’s $t$-test) or different (Welch’s $t$-test) variance
Some common ways in which these assumptions are broken

- Independence — e.g. time-series data are usually highly correlated.
- Samples drawn randomly — this does not mean that continue doing the experiment until you get the “right answer” and something becomes significant.
- Normality — I’ll talk about this next.
The need for non-parametric tests
Parametric vs non-parametric tests

- The $t$-test is one example of a family of statistical tests called **parametric tests**
- There are many others that you’ll encounter later – $F$-tests for ANOVA and $\chi^2$ tests are among the more famous.
Parametric vs non-parametric tests

▶ The $t$-test is one example of a family of statistical tests called **parametric tests**
▶ There are many others that you’ll encounter later – $F$-tests for ANOVA and $\chi^2$ tests are among the more famous.
▶ Parametric tests fundamentally **assume** something, usually boiling down to the form of the PDF of the data – e.g. the $t$-test assumes the data are drawn from a normal distribution
Parametric vs non-parametric tests

- The *t*-test is one example of a family of statistical tests called **parametric tests**
- There are many others that you’ll encounter later – *F*-tests for ANOVA and $\chi^2$ tests are among the more famous.
- Parametric tests fundamentally **assume** something, usually boiling down to the form of the PDF of the data – e.g. the *t*-test assumes the data are drawn from a normal distribution.
- Non-parametric methods are a family of statistical tools that assume *nothing* about the PDF of the data.
Is it fair to assume everything is normally distributed?

Before you lies a small packet of smarties. Eat it, and count the number of yellow ones. (NB: Smarties are vegetarian [not vegan], contain small amount of gluten, and may contain traces of nuts)

What do you think the PDF for the number of yellow smarties per box is?
Experiment time!
The Poisson distribution

If the probability of an event occurring per unit time (or distance, or...) is constant and events occur at an average rate $\lambda$, and each event is independent, then we get a distribution of events known as the Poisson distribution, whose PDF is:

$$p(k \text{ yellow smarties}) = e^{-\lambda} \frac{\lambda^k}{k!}.$$ 

This is roughly true for the yellow smarties: one being in a box has only a small effect on the probability that another will join it, and $p(\text{yellow smartie})$ is constant over time.
The Poisson distribution

If the probability of an event occurring per unit time (or distance, or...) is constant and events occur at an average rate \( \lambda \), and each event is independent, then we get a distribution of events known as the Poisson distribution, whose PDF is:

\[
p(k \text{ yellow smarties}) = e^{-\lambda} \frac{\lambda^k}{k!}.
\]

This is roughly true for the yellow smarties: one being in a box has only a small effect on the probability that another will join it, and \( p(\text{yellow smartie}) \) is constant over time.

Perhaps “more scientifically”, it’s also true for the number of radioactive decays per unit time and for the number of mutations on a strand of DNA per unit length.
The Poisson distribution
Sampling from a non-normal PDF

What happens to the *sampling distribution* of the mean?
Sampling from a non-normal PDF

What happens to the sampling distribution of the mean?

n=2 samples
Sampling from a non-normal PDF

What happens to the *sampling distribution* of the mean?

![Histogram of number of smarties](image1)

![Histogram of mean smarties](image2)

n=5 samples
Sampling from a non-normal PDF

What happens to the sampling distribution of the mean?

n=10 samples
Sampling from a non-normal PDF

What happens to the \textit{sampling distribution} of the mean?

\begin{itemize}
  \item Count
  \begin{itemize}
    \item 0
    \item 0.25
    \item 0.50
    \item 0.75
    \item 1.00
  \end{itemize}
\end{itemize}
Sampling from a non-normal PDF

What happens to the sampling distribution of the mean?

![Histogram of number of smarties](image1)

n=100 samples

![Histogram of mean smarties](image2)
The Central Limit Theorem

This is an example of a wonderfully powerful piece of maths, called the **Central Limit Theorem**, or CLT.
The Central Limit Theorem

This is an example of a wonderfully powerful piece of maths, called the **Central Limit Theorem**, or CLT.

The CLT states that the sum of lots of independently and randomly sampled random variables (i.e. uncertain things) tends to a normal distribution, *whatever* the shape of the PDF that the random variables are drawn from.
The Central Limit Theorem

This is an example of a wonderfully powerful piece of maths, called the Central Limit Theorem, or CLT.

The CLT states that the sum of lots of independently and randomly sampled random variables (i.e. uncertain things) tends to a normal distribution, whatever the shape of the PDF that the random variables are drawn from.

This means that in the limit as the number of samples tends to infinity the distribution of their mean tends to the normal one.
Non-normality

There’s quite a large jump between $\infty$ and any experiment you’d like to do in the lab. There are several methods to test the hypothesis that your data are normal (e.g. the Kolmogorov–Smirnov, or Shapiro–Wilk test) but they tend to require comparatively large sample sizes.

Alternatively, a common “rule of thumb” is that if you draw $n \gtrsim 30$ independent and random samples from a unimodal distribution the CLT is valid.
Transforming data

Sometimes, the data that we acquire isn’t normally distributed, but some function of it is. We can then transform the data we acquire – e.g. by taking logs – into something that is normally distributed, on which we use statistics. This is usually done for a reason known before hand.
Outliers

- **Outliers** are measurements or observations that are distant from all the others you’ve measured thus far.

How far is “far”? There is no rigorous mathematical definition of what an outlier actually is, but people typically say that they are points more than 3 or 5 the standard deviation away from the population as a whole.

Roughly speaking, outliers can mean one of three things: (a) you’ve just watched someone win the lottery, (b) experimental error, or (c) your conceptual understanding of what you’re looking at is wrong.

From the point of view of $t$-tests, outliers are bad news: the mean is not robust to them, and one dodgy piece of data can change “your answer” substantially.
Outliers

- Outliers are measurements or observations that are distant from all the others you’ve measured thus far.

- How far is “far”? There is no rigorous mathematical definition of what an outlier actually is, but people typically say that they are points more than $3\times$ or $5\times$ the standard deviation away from the population as a whole.
Outliers

- **Outliers** are measurements or observations that are distant from all the others you’ve measured thus far.

- How far is “far”? There is no rigorous mathematical definition of what an outlier actually is, but people typically say that they are points more than $3\times$ or $5\times$ the standard deviation away from the population as a whole.

- Roughly speaking, outliers can mean one of three things: (a) you’ve just watched someone win the lottery, (b) experimental error, or (c) your conceptual understanding of what you’re looking at is wrong.
Outliers

- **Outliers** are measurements or observations that are distant from all the others you’ve measured thus far.

- How far is “far”? There is no rigorous mathematical definition of what an outlier actually is, but people typically say that they are points more than $3\times$ or $5\times$ the standard deviation away from the population as a whole.

- Roughly speaking, outliers can mean one of three things: (a) you’ve just watched someone win the lottery, (b) experimental error, or (c) your conceptual understanding of what you’re looking at is wrong.

- From the point of view of $t$-tests, outliers are **bad news**: the mean is not robust to them, and one dodgy piece of data can change “your answer” substantially.
An example

Consider an mRNA microarray reporting on the relative expression of many thousands of different genes.
An example

Golub et al. (Science, Oct. 1999 vol. 15; 286(5439):531-7) measured the gene activity of 3500 genes for 67 patients. 42 patients had subtype A of leukaemia and 25 had subtype B. Does the gene activity differ between these subtypes? If so, for which genes? (We perhaps don’t expect many changes – both cells are leukaemia cells, after all).
An example

The actual data is the logarithm of the ratio of fluorescence between a test sample and the sample containing the real RNA, e.g.:

<table>
<thead>
<tr>
<th>Patient:</th>
<th>Gene 1</th>
<th>Subtype A</th>
<th></th>
<th></th>
<th>Gene 2</th>
<th>Subtype B</th>
<th></th>
<th></th>
<th>Gene 3500</th>
<th>Subtype A</th>
<th></th>
<th></th>
<th>Subtype B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0.79</td>
<td></td>
<td></td>
<td>-0.75</td>
<td>1.34</td>
<td></td>
<td></td>
<td>0.99</td>
<td>1.34</td>
<td></td>
<td></td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.34</td>
<td></td>
<td></td>
<td>-1.26</td>
<td>1.32</td>
<td></td>
<td></td>
<td>0.33</td>
<td>1.11</td>
<td></td>
<td></td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>…</td>
<td>…</td>
<td></td>
<td></td>
<td>…</td>
<td>…</td>
<td></td>
<td></td>
<td>…</td>
<td>…</td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>42</td>
<td>1.42</td>
<td>0.27</td>
<td></td>
<td></td>
<td>1.32</td>
<td>0.55</td>
<td></td>
<td></td>
<td>1.11</td>
<td>1.14</td>
<td></td>
<td></td>
<td>1.06</td>
</tr>
<tr>
<td>1</td>
<td>0.44</td>
<td>…</td>
<td></td>
<td></td>
<td>0.58</td>
<td>…</td>
<td></td>
<td></td>
<td>1.06</td>
<td>…</td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>2</td>
<td>…</td>
<td>…</td>
<td></td>
<td></td>
<td>…</td>
<td>…</td>
<td></td>
<td></td>
<td>1.22</td>
<td>…</td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>25</td>
<td>1.42</td>
<td>…</td>
<td></td>
<td></td>
<td>…</td>
<td>0.99</td>
<td></td>
<td></td>
<td>1.22</td>
<td>…</td>
<td></td>
<td></td>
<td>…</td>
</tr>
</tbody>
</table>
Let’s look at a specific row – gene 3 (which has the wonderful name STAT1 — Signal Transducer And Activator Of Transcription — seriously!)

<table>
<thead>
<tr>
<th>Patient</th>
<th>Subtype A</th>
<th></th>
<th>Subtype B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>\cdots</td>
<td>42</td>
</tr>
<tr>
<td>Gene 3</td>
<td>1.79</td>
<td>0.24</td>
<td>\cdots</td>
<td>1.62</td>
</tr>
</tbody>
</table>
An example

Let’s look at a specific row – gene 3 (which has the wonderful name STAT1 — Signal Transducer And Activator Of Transcription — seriously!)

<table>
<thead>
<tr>
<th>Patient:</th>
<th>Subtype A</th>
<th>Subtype B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>42</td>
</tr>
<tr>
<td>Gene 3</td>
<td>1.79</td>
<td>0.24</td>
</tr>
</tbody>
</table>
An example

Would we notice this difference, which might be due to measurement or human error?

We’d like to find a way to limit the damage that this mistake can do.
One example of a non-parametric test that is robust to these sorts of problems is the Wilcoxon Rank Sum test, aka the Mann-Whitney U test.

We take a totally different approach to the problem: rather than looking at “actual numbers” (like the values in the dataset) we rank them, from smallest (1) to largest ($n = 42 + 25$ above). If there are ties, we take their average.
The Wilcoxon test

▶ One example of a **non-parametric test** that is **robust** to these sorts of problems is the **Wilcoxon Rank Sum test**, aka the Mann-Whitney U test.

▶ We take a totally different approach to the problem: rather than looking at “actual numbers” (like the values in the dataset) we **rank** them, from smallest (1) to largest ($n = 42 + 25$ above). If there are ties, we take their average.

▶ The distribution of the **sum of the ranks** becomes the test statistic. If there is no difference between the groups, then it shouldn’t matter what order they’ve happened to come in.
The Wilcoxon test

- One example of a **non-parametric test** that is **robust** to these sorts of problems is the **Wilcoxon Rank Sum test**, aka the Mann-Whitney U test.

- We take a totally different approach to the problem: rather than looking at “actual numbers” (like the values in the dataset) we **rank** them, from smallest (1) to largest ($n = 42 + 25$ above). If there are ties, we take their average.

- The distribution **of the** sum of the **ranks** becomes the test statistic. If there is no difference between the groups, then it shouldn’t matter what order they’ve happened to come in.

- Note that the fact that we only care about the ranks means that we’re already somewhat immune to many physical sources of error (e.g. quadratic calibration issues) **and** we can use it when we can **only** rank the data.
The Wilcoxon test

- So, how does it work?
  (In R, very simply – `wilcox.test(group1, group2)`)
- Let’s imagine a subset of the big gene table above:
The Wilcoxon test

- So, how does it work?
  (In R, very simply – `wilcox.test(group1, group2)`)
- Let’s imagine a subset of the big gene table above:

<table>
<thead>
<tr>
<th>Patient</th>
<th>Subtype A</th>
<th>Subtype B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Gene 3500:</td>
<td>0.99</td>
<td>0.33</td>
</tr>
<tr>
<td>Rank</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.14</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>
The Wilcoxon test

- So, how does it work?
  (In R, very simply – `wilcox.test(group1, group2)`)  
- Let’s imagine a subset of the big gene table above:

<table>
<thead>
<tr>
<th>Patient</th>
<th>Subtype A</th>
<th>Subtype B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 3500:</td>
<td>0.99</td>
<td>0.33</td>
</tr>
<tr>
<td>Rank</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

- The sum of the ranks in one group (call it B) is \( W = 7 \). Note that this determines the sum of the ranks in the other group.
The Wilcoxon test

- So, how does it work?
  (In R, very simply – `wilcox.test(group1, group2)`)
- Let’s imagine a subset of the big gene table above:

<table>
<thead>
<tr>
<th>Patient:</th>
<th>Subtype A</th>
<th>Subtype B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 3500:</td>
<td>0.99 0.33</td>
<td>1.14 1.06</td>
</tr>
<tr>
<td>Rank</td>
<td>2 1 4 3</td>
<td></td>
</tr>
</tbody>
</table>

- The sum of the ranks in one group (call it B) is \( W = 7 \). Note that this determines the sum of the ranks in the other group.
- Is \( W = 7 \) “suspiciously” low or high?
Wilcoxon test

What is the distribution of $W$ under the null hypothesis? All of the following outcomes are equally likely if there is no difference between the groups (leaving out permutations within groups):
Wilcoxon test

What is the distribution of $W$ under the null hypothesis? All of the following outcomes are equally likely if there is no difference between the groups (leaving out permutations within groups):

<table>
<thead>
<tr>
<th>Patient:</th>
<th>Subtype A</th>
<th>Subtype B</th>
<th>$W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank:</td>
<td>1 2</td>
<td>3 4</td>
<td>7</td>
</tr>
<tr>
<td>Rank:</td>
<td>1 3 2 4</td>
<td>2 4</td>
<td>6</td>
</tr>
<tr>
<td>Rank:</td>
<td>1 4 2 3</td>
<td>2 3</td>
<td>5</td>
</tr>
<tr>
<td>Rank:</td>
<td>2 3 1 4</td>
<td>1 4</td>
<td>5</td>
</tr>
<tr>
<td>Rank:</td>
<td>2 4 1 3</td>
<td>1 2</td>
<td>4</td>
</tr>
<tr>
<td>Rank:</td>
<td>3 4 1 2</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>
Wilcoxon test

- So, \( P_{H_0}(W \geq 7) = 1/6 \approx 0.167 \). This is the most extreme outcome, and is indeed “highly suspicious”, but it’s also \( > 0.05 \). In fact, we can’t reject \( H_0 \) at the commonly used \( \alpha = 0.05 \) with two samples in each group.
- Let’s add two more patients to the group:
Wilcoxon test

- So, \( P_{H_0}(W \geq 7) = 1/6 \approx 0.167 \). This is the most extreme outcome, and is indeed “highly suspicious”, but it’s also \( > 0.05 \). In fact, we can’t reject \( H_0 \) at the commonly used \( \alpha = 0.05 \) with two samples in each group.

- Let’s add two more patients to the group:

<table>
<thead>
<tr>
<th>Patient</th>
<th>Subtype A</th>
<th></th>
<th>Subtype B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Gene 3500</td>
<td>0.99</td>
<td>0.33</td>
<td>0.41</td>
<td>1.01</td>
</tr>
<tr>
<td>Rank</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

This gives \( W = 26 \). Is this significant?
Wilcoxon test

- With $n = m = 4$ patients in each group, there are $8! = 40320$ different possible combinations of rank.

- Working out all of these is somewhat difficult, particular if $n$ (the number of patients in one group) and/or $m$ (the other group) are large.
Wilcoxon test

- With $n = m = 4$ patients in each group, there are $8! = 40320$ different possible combinations of rank.
- Working out all of these is somewhat difficult, particularly if $n$ (the number of patients in one group) and/or $m$ (the other group) are large.
- It turns out that

$$P_{H_0}(W \geq 26) = \frac{1}{\binom{n+m}{m}} = \frac{m! n!}{(n+m)!} \approx 0.0142$$
Wilcoxon test

- With \( n = m = 4 \) patients in each group, there are \( 8! = 40,320 \) different possible combinations of rank.
- Working out all of these is somewhat difficult, particular if \( n \) (the number of patients in one group) and/or \( m \) (the other group) are large.
- It turns out that

\[
P_{H_0}(W \geq 26) = \frac{1}{\binom{n+m}{m}} = \frac{m!n!}{(n+m)!} \approx 0.0142
\]

- If \( \Delta \) is our “effect” (where \( H_0 \) is \( \Delta = 0 \)), we then conclude that the \( p \)-value of \( H_1 : \Delta > 0 \) is 0.0142, and the \( p \)-value of \( H_1 : \Delta \neq 0 \) is \( 2 \times 0.0142 \).
Wilcoxon test

Both of these are below our usual $\alpha < 0.05$, and we can reject the null.
What happens if I corrupt the data?
Robustness

Here are confidence intervals on $\Delta$ for 200 randomly selected genes, while the right panel shows corresponding confidence intervals obtained by $t$-tests. If these exclude 0, then there’s a significant difference in that gene between the disease groups.
Wilcoxon test

What happens if I corrupt the dataset – by adding 100 to the first point in every experiment?
What happens if I corrupt the dataset – by adding 100 to the first point in every experiment? \textit{t}-tests:
What happens if I corrupt the dataset – by adding 100 to the first point in every experiment? Wilcoxon test:

Example with thanks to Jennifer Rogers, Dept of Statistics

Gene
Conf. Interval
Gene
Conf. Interval

Example with thanks to Jennifer Rogers, Dept of Statistics
The price of being robust is power

- The Wilcoxon test (and other non-parametric tests more generally) are much more robust in the presence of outliers. They also tend to work under many transformations of the data (e.g. if I multiply every data point by 10, their rank won’t change).

- The price of this is power. It turns out that if the data are perfectly normal the Wilcoxon test is \(3/\pi \approx 95\%\) as efficient as a corresponding \(t\)-test.
Effect size and power
Effect size

All statistical tests provide a handle on seeing if estimators computed on samples drawn from two groups of samples are likely to be from the same distribution – e.g. the $t$-test looks for a difference in means.

Clearly this is sometimes going to be easier to spot than others – sometimes the difference is huge and other times it is tiny!
Effect size

Jacob Cohen (1923 – 1998) was an American statistician who was the first person to formalise this concept. He proposed a simple measure to quantify the effect size that we’re interested in, defined as

\[ d = \frac{\bar{x}_A - \bar{x}_B}{s_p} \]

where \( s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}} \)

is the pooled estimate of standard deviation (as in the last lecture).
Cohen’s $d$

The nice thing about $d$ is that it gives a **completely situation independent** measure of how “subtle” an effect is:

- $d=0.5$
- $d=1$
- $d=2$
- $d=3$
Cohen’s $d$

- There are a few other measures of effect size which are subtly different – e.g. Hedge’s $g^*$ which corrects for a small bias in $d$ that occurs if the sample sizes are small

\[
g^* \approx \left( 1 - \frac{3}{4(n_A + n_B) - 9} \right) d
\]

- Because the effect size is independent of the thing being studied, people have proposed a series of standard names for them – e.g. $d = 0.2$ is a small effect, $d = 0.5$ a medium one, $d = 0.8$ a large effect (Cohen, 1988) and $d = 2.0$ huge (Sawilowsky, 2009). Unfortunately, this doesn’t seem to really have caught on.
$p$-values and effect size

As $n_{A, B} \to \infty$ any difference in means will become statistically significant with a $t$-test, and moreover the $p$-value will tend to zero.

Many authors (erroneously) believe that very small $p$-values are indicative that they are very right.

The effect size is really the thing that you almost certainly actually care about – how different the two groups are. The $p$-value tells you if the drug appears to work; the effect size tells you “how well”.
Power

- Power is something I mentioned above, but will now define as

  \[ \text{Power} = P(\text{Reject } H_0 \text{ given that } H_1 \text{ is true}) \]

- As power increases the Type II error rate goes down.
Power

- Power is something I mentioned above, but will now define as

\[
\text{Power} = P(\text{Reject } H_0 \text{ given that } H_1 \text{ is true})
\]

- As power increases the Type II error rate goes down.
- Remember, **absence of evidence is not evidence of absence**.
Power

- Power is something I mentioned above, but will now define as

\[
\text{Power} = P(\text{Reject } H_0 \text{ given that } H_1 \text{ is true})
\]

- As power increases the Type II error rate goes down.
- Remember, **absence of evidence is not evidence of absence**.
- So, if we can work out power, we calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size.
Power

- Power is something I mentioned above, but will now define as

  \[ \text{Power} = P(\text{Reject } H_0 \text{ given that } H_1 \text{ is true}) \]

- As power increases the Type II error rate goes down.
- Remember, \textit{absence of evidence is not evidence of absence}.
- So, if we can work out power, we calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size.
- This is called “a power calculation” and is often part of the grant-writing process.
Power calculations

Power is a bit of a pain to calculate ‘by hand’. Fortunately, in R, we can just specify $d$ and ask what sample size we’d need to see it:

```
> library("pwr")  #Load package
> pwr.t.test(d=0.5,power=0.8,sig.level=0.05)
   Two-sample t test power calculation

   n = 63.76561
   d = 0.5
   sig.level = 0.05
   power = 0.8
   alternative = two.sided
```

NOTE: n is the number in *each* group
Power calculations

Note that proving a negative is very hard – but that doing this calculation lets us put a bound on the minimum effect size we could have seen (but didn’t) at an appropriate level of power.
Power calculations

Typically 80% is accepted as being reasonable for power.

Two-sample t-test power calculation

<table>
<thead>
<tr>
<th>Test power = $1 - \beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tails = two.sided</td>
</tr>
<tr>
<td>Effect size $d = 0.5$</td>
</tr>
<tr>
<td>Alpha = 0.05</td>
</tr>
</tbody>
</table>

Optimal sample size:

$n = 64$

$n$ is number in *each* group
Power calculations

Sample Size Estimation for t–tests

Alpha=0.05 (Two–tailed)

Effect size, $d$

Sample Size (n)

Power

- 0.4
- 0.5
- 0.6
- 0.7
- 0.8
- 0.9

HT 2018  Statistics Lecture 3  Effect size and power
An example power calculation

- The Zucker Diabetic Fatty (ZDF) Rat is an often used genetic model of Type 2 diabetes.
- An investigator is planning to use a novel biologic drug to double the rate of fatty acid oxidation, with the hope of ameliorating diabetes.
- For now, let us look only at one aspect of diabetes – the proportion of haemoglobin that is (non enzymatically) glycated at the terminal lysine, known as HbA1c, compared to the proportion of non-glycated haemoglobin.
An example power calculation

Glycated hemoglobin, PDB entry 3b75

Close-up view of glycated lysine, PDB entry 3b75
An example power calculation

- Previous work has shown that the proportion of $Hba1c$ to total haemoglobin in ZDF rats is $(60 \pm 15)$ mmol/mol ($\bar{x} \pm s$; $n = 10$), and in healthy controls it is far lower – typically around 10 mmol/mol. (In clinical biochemistry, an $Hba1c$ of $> 6.5\%$ is a diagnostic criteria for type-2 diabetes)

- How many rats are required?
An example power calculation

- *If* the drug doubled fatty acid oxidation and halved the long term plasma glucose concentration, we may expect to see a reduction in *Hba1c*. 

Assume for now that we would be satisfied if it halved compared to the untreated group, and that we would expect the distribution to remain unchanged – i.e. we'd assume that $s$ was 15 based on an $n = 10$. If this is true, how many rats are required?

This is enough information to compute Cohen's $d = \frac{x_A - x_B}{s_p} = \frac{60 - 30}{15} = 2$. 

HT 2018 Statistics Lecture 3 — Effect size and power
An example power calculation

- If the drug doubled fatty acid oxidation and halved the long term plasma glucose concentration, we may expect to see a reduction in $Hba1c$.

- Assume for now that we would be satisfied if it halved compared to the untreated group, and that we would expect the distribution to remain unchanged – i.e. we’d assume that $s$ was 15 based on an $n = 10$. If this is true, how many rats are required?
An example power calculation

- *If* the drug doubled fatty acid oxidation and halved the long term plasma glucose concentration, we may expect to see a reduction in *Hba1c*.

- Assume for now that we would be satisfied if it halved compared to the untreated group, and that we would expect the distribution to remain unchanged – i.e. we’d assume that $s$ was 15 based on an $n = 10$. *If* this is true, how many rats are required?

- This is enough information to compute Cohen’s

$$d = \frac{\bar{x}_A - \bar{x}_B}{s_p} = \frac{60 - 30}{15} = 2$$
An example power calculation

- With some rough estimate of effect size, we can then work out the sample size required for 80% power:

```r
> pwr.t.test(n=NULL, d=2, sig.level = 0.05, power=0.8)

Two-sample t test power calculation

n = 5.089995
d = 2
sig.level = 0.05
power = 0.8
alternative = two.sided

NOTE: n is number in *each* group
Quick summary

- Any hypothesis test is prone to Type I (false positive) and Type II (false negative) errors.
Quick summary

- Any hypothesis test is prone to Type I (false positive) and Type II (false negative) errors.
- We have to correct for performing multiple comparisons to avoid false positives.

\[ \text{t-test assumes quite a lot of things about your data, including normality} \]

\[ \text{Fortunately the CLT means that for } n \gtrapprox 30 \text{ normality (of the mean's sampling distribution!) is probably true, but below this we should be careful.} \]
Quick summary

- Any hypothesis test is prone to Type I (false positive) and Type II (false negative) errors.
- We have to correct for performing multiple comparisons to avoid false positives.
- The $t$-test assumes quite a lot of things about your data, including normality.
Quick summary

- Any hypothesis test is prone to Type I (false positive) and Type II (false negative) errors.
- We have to correct for performing multiple comparisons to avoid false positives.
- The $t$-test assumes quite a lot of things about your data, including normality.
- Fortunately the CLT means that for $n \gtrsim 30$ normality (of the mean’s sampling distribution!) is probably true, but below this we should be careful.
Quick summary

- Data may themselves be occasionally ‘corrupted’ (particularly when machines are involved and the dataset is too large to fit into a human’s brain)
Quick summary

- Data may themselves be occasionally ‘corrupted’ (particularly when machines are involved and the dataset is too large to fit into a human’s brain)

- **Non-parametric** tests are, as a rule, far more robust to this problem (and when the assumptions of the $t$-test are violated), but this comes at the price of power.
Quick summary

- Data may themselves be occasionally ‘corrupted’ (particularly when machines are involved and the dataset is too large to fit into a human’s brain)
- **Non-parametric** tests are, as a rule, far more robust to this problem (and when the assumptions of the *t*-test are violated), but this comes at the price of power.
- Power and effect size tell us what hypothesis tests can (and can’t) reasonably detect.
Some recommended links

▶ [https://xkcd.com/882/](https://xkcd.com/882/)
The end!