Hypothesis Testing

Daniel Lawson

Wellcome Trust Sir Henry Dale Research Fellow
IEU, University of Bristol
Visiting researcher at the University of Oxford

October 2014
Gently adapted from David Steinsaltz’s 2013 ‘hypothesis testing’ slides
Table of Contents

Background

The Z test
   The simple Z test
   The Z test for proportions
   The Z test for the difference between means
   Z test for the difference between proportions

The t test
   The simple t test
   The Matched sample t test
   Matched sample t test

The $\chi^2$ test
   The $\chi^2$ test

Non-parametric tests
   Why we need non-parametric tests
   Mann-Whitney U test
   Paired value tests

The menagerie of tests
Section 1

Background
The testing paradigm

Significance testing is about rejecting a null model.

- We have a research hypothesis, which helps define the alternative hypothesis.
- The null model is our best explanation of the data without that hypothesis.
- We see if the null model ‘fits’ the data with a test.
- We will ‘test’ all of the assumptions of the null together!
- i.e. we won’t know which assumption failed.
- If we reject the null, we hope our alternative hypothesis is the explanation!
- Confounding by some unaccounted process is the most common reason for incorrectly accepted alternatives.
- If you can’t account for all reasonable confounders, there is no point doing the test!
Decisions and hypothesis testing

One reason for HT is to **make decisions**. It is biased against the alternative, so that **if the null is rejected** we are fairly certain that is not a mistake.

- Example: A new vaccine.
- Null hypothesis: it doesn’t work.
- Alternative: it does. So use it!

It is important to control the probability that we decide it works! (But we aren’t taking into account the **cost** of getting it wrong.)
Types of error

- Type I: We reject the null hypothesis, but it's really true
- Type II: We retain the null, but it isn’t true

These are very different!
In many fields, no null hypothesis can be true.
This leads to the question: do we have power to reject the null?
In this case, standard hypothesis testing can be useless for demonstrating a specific alternative.
## The Confusion Matrix

<table>
<thead>
<tr>
<th>What is true?</th>
<th>Reject Null</th>
<th>Retain null</th>
</tr>
</thead>
<tbody>
<tr>
<td>null</td>
<td>False positive</td>
<td>True negative</td>
</tr>
<tr>
<td>alternative</td>
<td>True positive</td>
<td>False negative</td>
</tr>
</tbody>
</table>

Type I error: Reject null when true is alternative
Type II error: Retain null when true is alternative

Power = \[
\frac{\text{true positive rate}}{\text{false negative rate}}
\]

for a given false positive rate \(\alpha\).
Example: Spam Filter

Spam filters look at incoming email, and use statistical models to compute the probability that a real message would have certain features. If the probability is too low, it goes directly in the bin.

Null hypothesis: Real message.
Alternative: Spam.

Then a Type I error is

1. A spam message that gets through the filter;
2. An email from your friend that gets junked?
Example: Spam Filter

Spam filters look at incoming email, and use statistical models to compute the probability that a real message would have certain features. If the probability is too low, it goes directly in the bin.

Null hypothesis: Real message.
Alternative: Spam.

Then a **Type I error** is

1. A spam message that gets through the filter;
2. An email from your friend that gets junked?

Type I = falsely rejecting the null
Example: Spam Filter

Spam filters look at incoming email, and use statistical models to compute the probability that a real message would have certain features. If the probability is too low, it goes directly in the bin.

Null hypothesis: Real message.
Alternative: Spam.

Then a **Type I error** is

1. A spam message that gets through the filter;
2. An email from your friend that gets junked?

Type I = falsely rejecting the null
Answer: 2
Example: Criminal Trial

Juries weigh the evidence, and decide how likely the evidence would be if the defendant were innocent. Null hypothesis: Defendant is innocent. Alternative: Defendant is guilty.

Then a Type II error is

1. A guilty defendant set free;
2. An innocent defendant convicted.
Example: Criminal Trial

Juries weigh the evidence, and decide how likely the evidence would be if the defendant were innocent.
Null hypothesis: Defendant is innocent.
Alternative: Defendant is guilty.

Then a **Type II error** is

1. A guilty defendant set free;
2. An innocent defendant convicted.

Type II = falsely rejecting the alternative
Example: Criminal Trial

Juries weigh the evidence, and decide how likely the evidence would be if the defendant were innocent.
Null hypothesis: Defendant is innocent.
Alternative: Defendant is guilty.

Then a **Type II error** is

1. A guilty defendant set free;
2. An innocent defendant convicted.

Type II = falsely rejecting the alternative
Answer: 1
Example: Clinical Trial

A new cancer medication is tested in comparison to an old one. We test how likely the apparent improvement in survival would be, if the new drug were no better than the old one. The medication will be approved if its proved to be better.

Null hypothesis: The new medication is no better than the old one. Alternative: The new medication is better.

Then a Type II error is

1. When we approve the new medication even though its no better than the old one;

2. When we dont approve the new medication, even though it is better.
Example: Clinical Trial

A new cancer medication is tested in comparison to an old one. We test how likely the apparent improvement in survival would be, if the new drug were no better than the old one. The medication will be approved if its proved to be better.

Null hypothesis: The new medication is no better than the old one. Alternative: The new medication is better. Then a Type II error is

1. When we approve the new medication even though its no better than the old one;
2. When we dont approve the new medication, even though it is better.

Answer: 2
Hypothesis testing

- Start with a ‘research hypothesis’ - the ‘alternative hypothesis’
- Form a ‘null hypothesis’ that says nothing interesting happened – it was all chance variation
- Determine the test statistic $T$, measuring how unusual the data is under the null hypothesis
- Define a ‘significance level’ $\alpha$
- Determine the critical value $T_{crit}(\alpha)$
- Compute test statistic $T$ for the observed data
- Compute the p-value: $p(T)$ probability of a test statistic as extreme or more than that observed (under the null)
- Retain the null if $p > \alpha$, or equivalently $T < T_{crit}$. Otherwise reject it in favour of the alternative.
Section 2

The Z test
Subsection 1

The simple Z test
The simple Z test

For $n$ data points $X_i$

- If the mean of the data can be treated as Normal,
- And our null hypothesis is $\bar{X} = \mu$ for known $\mu$...
- And we know the standard deviation $\sigma$...
- Then we compute the test statistic $Z = (\bar{X} - \mu)/\sigma$...
- Under the null, $Z \sim N(0, 1)$

So the Z test computes

- $p(Z' \geq Z)$ (one tailed)
- $p(|Z'| \geq |Z|)$ (two tailed)
- where $Z'$ is from the null
Important properties of the Normal distribution

If $X$ is normal then $aX + b$ is normal
If $X$ and $Y$ are normal and independent, then $X + Y$ is normal. Specifically, $Z = X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. 
Important properties of the Normal distribution

If $X$ is normal then $aX + b$ is normal
If $X$ and $Y$ are normal and independent, then $X + Y$ is normal.
Specifically, $Z = X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Meaning that we add the variance, regardless of whether we add $X$ and $Y$!
Important properties of the Normal distribution

If $X$ is normal then $aX + b$ is normal.

If $X$ and $Y$ are normal and independent, then $X + Y$ is normal. Specifically, $Z = X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Meaning that we add the variance, regardless of whether we add $X$ and $Y$!

More generally, if $X$ and $Y$ are correlated,

$$\text{var}(X + Y) = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$$
Law of large numbers

Let $X_1, \ldots, X_n$ be independent samples from a distribution with mean $\mu$ and variance $\sigma^2$. Then

$$\bar{X} := \frac{1}{n} (X_1 + \ldots + X_n)$$

converges to $\mu$ as $n \to \infty$.

- ‘Estimate’ of $\mu$: $\bar{X}$
- Variance of the estimator: $\text{var}(\bar{X}) = \sigma^2 / n$
- ‘standard error’ $SE$: $\text{SD}(\bar{X}) = \sigma / \sqrt{n}$

So $Z := \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ is ‘standardized’ to have mean 0 and s.d. 1.
Takeaway message:

- You can assume that the mean of a distribution is normal
- ... if \( n \) (sample size) is big enough
- IT DOESNT MATTER if the data you sampled were normally distributed.
- The mean still has to be normally distributed
- How big is ‘big enough’? It depends on the distribution!

1: There are some special distributions that don’t obey the law of large numbers. These have infinite variance.
Z test example

- British birthweights have Mean 3426g and SD 538g from a large sample
- Australian birthweights sampled (1 day)
- Australians have mean 3276
- Null: birthweights have the same mean
- Alternative: Australian babies are smaller

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>Sex</th>
<th>Weight (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>F</td>
<td>3837</td>
</tr>
<tr>
<td>64</td>
<td>F</td>
<td>3334</td>
</tr>
<tr>
<td>78</td>
<td>M</td>
<td>3554</td>
</tr>
<tr>
<td>115</td>
<td>M</td>
<td>3838</td>
</tr>
<tr>
<td>177</td>
<td>M</td>
<td>3625</td>
</tr>
<tr>
<td>245</td>
<td>F</td>
<td>2208</td>
</tr>
<tr>
<td>247</td>
<td>F</td>
<td>1745</td>
</tr>
<tr>
<td>262</td>
<td>M</td>
<td>2846</td>
</tr>
<tr>
<td>271</td>
<td>M</td>
<td>3166</td>
</tr>
<tr>
<td>428</td>
<td>M</td>
<td>3520</td>
</tr>
<tr>
<td>455</td>
<td>M</td>
<td>3380</td>
</tr>
<tr>
<td>492</td>
<td>M</td>
<td>3294</td>
</tr>
<tr>
<td>494</td>
<td>F</td>
<td>2576</td>
</tr>
<tr>
<td>549</td>
<td>F</td>
<td>3208</td>
</tr>
<tr>
<td>635</td>
<td>M</td>
<td>3521</td>
</tr>
<tr>
<td>649</td>
<td>F</td>
<td>3746</td>
</tr>
<tr>
<td>653</td>
<td>F</td>
<td>3523</td>
</tr>
<tr>
<td>693</td>
<td>M</td>
<td>2902</td>
</tr>
<tr>
<td>729</td>
<td>M</td>
<td>2635</td>
</tr>
<tr>
<td>776</td>
<td>M</td>
<td>3920</td>
</tr>
<tr>
<td>785</td>
<td>M</td>
<td>3690</td>
</tr>
<tr>
<td>846</td>
<td>F</td>
<td>3430</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>Sex</th>
<th>Weight (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>847</td>
<td>F</td>
<td>3480</td>
</tr>
<tr>
<td>873</td>
<td>F</td>
<td>3116</td>
</tr>
<tr>
<td>886</td>
<td>F</td>
<td>3428</td>
</tr>
<tr>
<td>914</td>
<td>M</td>
<td>3783</td>
</tr>
<tr>
<td>991</td>
<td>M</td>
<td>3345</td>
</tr>
<tr>
<td>1017</td>
<td>M</td>
<td>3034</td>
</tr>
<tr>
<td>1062</td>
<td>F</td>
<td>2184</td>
</tr>
<tr>
<td>1087</td>
<td>M</td>
<td>3300</td>
</tr>
<tr>
<td>1105</td>
<td>F</td>
<td>2383</td>
</tr>
<tr>
<td>1134</td>
<td>M</td>
<td>3428</td>
</tr>
<tr>
<td>1149</td>
<td>M</td>
<td>4162</td>
</tr>
<tr>
<td>1187</td>
<td>M</td>
<td>3630</td>
</tr>
<tr>
<td>1189</td>
<td>M</td>
<td>3406</td>
</tr>
<tr>
<td>1191</td>
<td>M</td>
<td>3402</td>
</tr>
<tr>
<td>1210</td>
<td>F</td>
<td>3500</td>
</tr>
<tr>
<td>1237</td>
<td>M</td>
<td>3736</td>
</tr>
<tr>
<td>1251</td>
<td>M</td>
<td>3370</td>
</tr>
<tr>
<td>1264</td>
<td>M</td>
<td>2121</td>
</tr>
<tr>
<td>1283</td>
<td>M</td>
<td>3150</td>
</tr>
<tr>
<td>1337</td>
<td>F</td>
<td>3866</td>
</tr>
<tr>
<td>1407</td>
<td>F</td>
<td>3542</td>
</tr>
<tr>
<td>1435</td>
<td>F</td>
<td>3278</td>
</tr>
</tbody>
</table>
Z test example

- Null hypothesis: The Australian birthweights are samples from the same distribution as UK birth weights.
- Let $X_1, \cdots, X_{44}$ be the birth weights in the sample
- If the null hypothesis holds,
  \[
  \bar{X} = \frac{X_1 + \cdots + X_{44}}{44} \sim N(3426, \frac{538^2}{44})
  \]
- What is the probability that we would observe $\bar{X} \leq 3276$?
- Standardise: Observed
  \[
  Z = (\bar{X} - 3426)/81 \sim N(0, 1) = -1.81
  \]
  
  $p(Z \leq -1.81) = 0.035$ can be looked up in a table!
  or computed using:
  - Matlab: cdf('normal',-1.81)
Z test example

- Null hypothesis: The Australian birthweights are samples from the same distribution as UK birth weights.
- Let $X_1, \ldots, X_{44}$ be the birth weights in the sample
- If the null hypothesis holds,
  \[
  \bar{X} = \frac{X_1 + \cdots + X_{44}}{44} \sim N(3426, 538^2/44)
  \]
- What is the probability that we would observe $\bar{X} \leq 3276$?
- Standardise: Observed
  \[
  Z = \frac{\bar{X} - 3426}{81} \sim N(0, 1) = -1.81
  \]
  \[
  p(Z \leq -1.81) = 0.035 \text{ can be looked up in a table!}
  \]
or computed using:
  - Matlab: `cdf('normal',-1.81)`
  - R: `pnorm(-1.81)`
Z test example

- Null hypothesis: The Australian birthweights are samples from the same distribution as UK birth weights.
- Let $X_1, \cdots, X_{44}$ be the birth weights in the sample
- If the null hypothesis holds,

\[ \bar{X} = (X_1 + \cdots + X_{44})/44 \sim N(3426, 538^2/44) \]

- What is the probability that we would observe $\bar{X} \leq 3276$?
- Standardise: Observed

\[ Z = (\bar{X} - 3426)/81 \sim N(0, 1) = -1.81 \]

$p(Z \leq -1.81) = 0.035$ can be looked up in a table! or computed using:

- Matlab: cdf('normal',-1.81)
- R: pnorm(-1.81)
- Excel:
Z test example

- Null hypothesis: The Australian birthweights are samples from the same distribution as UK birth weights.
- Let $X_1, \cdots, X_{44}$ be the birth weights in the sample
- If the null hypothesis holds,
  \[
  \bar{X} = \frac{X_1 + \cdots + X_{44}}{44} \sim N(3426, 538^2/44)
  \]
- What is the probability that we would observe $\bar{X} \leq 3276$?
- Standardise: Observed
  \[
  Z = \frac{\bar{X} - 3426}{81} \sim N(0, 1) = -1.81
  \]
  $p(Z \leq -1.81) = 0.035$ can be looked up in a table!
  or computed using:
  - Matlab: cdf('normal',-1.81)
  - R: pnorm(-1.81)
  - Excel: (I’m not going to encourage this!)
Z test example

- Null hypothesis: The Australian birthweights are samples from the same distribution as UK birth weights.
- Let $X_1, \cdots, X_{44}$ be the birth weights in the sample
- If the null hypothesis holds,
  \[
  \overline{X} = (X_1 + \cdots + X_{44})/44 \sim N(3426, 538^2/44)
  \]
- What is the probability that we would observe $\overline{X} \leq 3276$?
- Standardise: Observed
  \[
  Z = (\overline{X} - 3426)/81 \sim N(0, 1) = -1.81
  \]
  \[
  p(Z \leq -1.81) = 0.035 \text{ can be looked up in a table!}
  \]
or computed using:
  - Matlab: cdf('normal', -1.81)
  - R: pnorm(-1.81)
  - Excel: (I’m not going to encourage this!)
  - Wolfram Alpha, your phone, Google, etc!
Tails of the Normal Distribution

![Graph of the Normal Distribution with shaded area in the tail]
Two tailed test

- P-value 0.035 in one-tailed test
- P-value 0.070 in two-tailed test
- For symmetric distributions, p-value always doubles
Does it matter how many tails?

- Not really… Just make sure you’re clear on why.
- Switching from two-tailed to one-tailed can make a non-significant result significant.
- Hypothesis testing is in the business of being conservative…
- You should therefore have to justify a one-tailed choice.
- Previously, before we looked at the data, did we expect Australian babies to be smaller? Would we not have been interested if they were bigger?
- If that was interesting too, we need a two-tailed test.
Requiem on errors: Power and alternative hypotheses

\[ \text{se} = 1.00 \quad z^* = 1.64 \quad \text{power} = 0.26 \]
\[ n = 1 \quad \text{sd} = 1.00 \quad \text{diff} = 1.00 \quad \alpha = 0.050 \]

Null Distribution

---\> rejection region

\[ \alpha \]

Alternative Distribution

---\> rejection region

\[ \text{Power} \]

Power with 1 sd difference in the alternative
Requiem on errors: Power and alternative hypotheses

se = 1.00  \( z^* = 1.64 \)  power = 0.991
n = 1  sd = 1.00  diff = 4.00  alpha = 0.050

Power with 4 sd difference in the alternative
Subsection 2

The Z test for proportions
Testing the number of successes

- Test: observe $X$ successes from $n$ trials
- H0: $X/n = p$, we are testing the probability of a success
- Let $C_i \sim Bern(p)$
- i.e. Bernoulli with $p=$success probability,
- $X = \sum_{i=1}^{n} C_i \sim Binomial(n, p)$ is the number of successes
- Can compute $p(X \leq x; n, p)$ (too few successes) and $p(X \geq x; n, p)$ (too many successes) explicitly for moderate $n$
- But do we need to?
Normal approximation to the Binomial

If \( X \sim Bin(n, p) \) then \( X \) is approximately distributed as \( N(\mu, \sigma^2) \) when \( n \) is large

- With \( \mu = np \) and \( \sigma^2 = np(1 - p) \)
- What is large? Rule of thumb: \( \mu \geq 3\sigma \)
- What is meant by approximately? \( P(a < X < b) \) is close to \( P(a < \mu + \sigma Z < b) \) for \( Z \sim N(0, 1) \)
Binom(n=3, p=0.5)
Binom(n=10, p=0.5)
Binom\((n=25, p=0.5)\)
Binom(n=100,p=0.5)
Binom(n=3,p=0.1)
Binom(n=10, p=0.1)
Binom(n=25, p=0.1)
Binom(n=100, p=0.1)
Continuity Correction

- Suppose we wanted to know $p(10 \leq X \leq 15; n = 25, p = 0.5)$
- e.g. Probability that we get between 10 and 15 heads from 25 flips of a fair coin
- How do we account for ‘discreteness’ of $X$ when using the normal approximation?
Continuity Correction
Continuity Correction

- Suppose we wanted to know \( p(10 \leq X \leq 15; n = 25, p = 0.5) \)
- e.g. Probability that we get between 10 and 15 heads from 25 flips of a fair coin
- How do we account for ‘discreteness’ of \( X \) when using the normal approximation?
- **Answer**: ‘Continuity correct’ to halfway between discrete values
- \( P(a \leq X \leq b) \approx P(a - 0.5 \leq Y \leq b + 0.5) \), where \( Y \sim N(\mu, \sigma^2) \)
Continuity Correction

- Suppose we wanted to know $p(10 \leq X \leq 15; n = 25, p = 0.5)$
- e.g. Probability that we get between 10 and 15 heads from 25 flips of a fair coin
- How do we account for ‘discreteness’ of $X$ when using the normal approximation?
- **Answer:** ‘Continuity correct’ to halfway between discrete values
- $P(a \leq X \leq b) \approx P(a - 0.5 \leq Y \leq b + 0.5)$, where $Y \sim N(\mu, \sigma^2)$
- Exact calculation: 0.7705
Continuity Correction

- Suppose we wanted to know \( p(10 \leq X \leq 15; n = 25, p = 0.5) \)
- e.g. Probability that we get between 10 and 15 heads from 25 flips of a fair coin
- How do we account for ‘discreteness’ of \( X \) when using the normal approximation?
- **Answer**: ‘Continuity correct’ to halfway between discrete values
- \( P(a \leq X \leq b) \approx P(a - 0.5 \leq Y \leq b + 0.5) \), where \( Y \sim N(\mu, \sigma^2) \)
- Exact calculation: 0.7705
- Normal approximation: 0.7699

Relative error: \( \frac{\text{exact} - \text{normal}}{\text{exact}} = 0.0008 \)
Continuity Correction

- Suppose we wanted to know \( p(10 \leq X \leq 15; n = 25, p = 0.5) \)
- e.g. Probability that we get between 10 and 15 heads from 25 flips of a fair coin
- How do we account for ‘discreteness’ of \( X \) when using the normal approximation?
- **Answer:** ‘Continuity correct’ to halfway between discrete values

\[
P(a \leq X \leq b) \approx P(a - 0.5 \leq Y \leq b + 0.5), \text{ where } Y \sim N(\mu, \sigma^2)
\]

- Exact calculation: 0.7705
- Normal approximation: 0.7699
- Relative error: \( \frac{(\text{exact}-\text{normal})}{\text{exact}} = 0.0008 \)
Z test for proportions

- $n$ independent trials with success probability $p$
- Observe $X$ successes
- H0: Probability of success is $p_0$
- IF H0 is true, $X/n \sim N(p_0, \sqrt{p_0(1-p_0)/n})$
- Z test with $Z = \frac{X/n - p_0}{\sqrt{p_0(1-p_0)/n}}$
Z test for proportions

- $n$ independent trials with success probability $p$
- Observe $X$ successes
- H0: Probability of success is $p_0$
- IF H0 is true, $X/n \sim N(p_0, \sqrt{p_0(1-p_0)/n})$
- Z test with $Z = \frac{X/n - p_0}{\sqrt{p_0(1-p_0)/n}}$
- (Should do continuity correction, but this is not important for large $n$...)
Example: ESP

- Charles Tart (1970s): 7500 (500×15) attempts on ‘Aquarius machine’
- Subjects predict which of four lights will come on
- Signal tells them if they were right.
- 7500 attempts. Expect 7500/4=1875 right.
- Actually observed 2006 correct guesses. Could it be purely by chance?
Example: ESP

- H0: Probability of success $p = 0.25$
- H1: Probability of success $p > 0.25$
- Significance level 0.01, meaning $Z_{crit} = 2.3$
- $X = \frac{2006}{7500} = 0.26747$, $\sigma = \sqrt{0.25 \times 0.75/7500} = 0.005$

$$Z = \frac{0.26747 - 0.25}{0.005} = 3.49$$

- Comfortably above critical value. $p = 0.0002$
ESP conclusions

Did the subjects do so well purely by chance?

- **Almost certainly not.** Under the null this would happen in one experiment out of 5000.
- Should we conclude that some of the subjects had the power to see into the future and predict which light would come on?
- Can you think of other alternatives?

In fact, there was a problem with the machine which made the order of the lights be not independent.

Conclusion: You have to be careful in interpreting the results of statistical tests. Just because you can show it didn't happen 'by chance' doesn't mean your favourite alternative holds.
ESP conclusions

Did the subjects do so well purely by chance?

► **Almost certainly not.** Under the null this would happen in one experiment out of 5000.

► Should we conclude that some of the subjects had the power to see into the future and predict which light would come on?

► Can you think of other alternatives?

► In fact, there was a problem with the machine which made the order of the lights be not independent.

Conclusion: You have to be careful in interpreting the results of statistical tests. Just because you can show it didn’t happen ‘by chance’ doesn’t mean your favourite alternative holds.
Subsection 3

The Z test for the difference between means
Z test for difference between means

- Observations from **two different populations**.
- Means from both are normally distributed.
- **SDs are known**: $\sigma_1$ and $\sigma_2$
- Unknown means $\mu_1$ and $\mu_2$
- Observe mean $\overline{X}_1$ from $n_1$ pop 1 samples and mean $\overline{X}_2$ from $n_2$ pop 2 samples
- Test $H_0 : \mu_1 = \mu_2$
- If $H_0$ is true, $X_1 - X_2 \sim N(0, \sigma^2)$
- with $\sigma = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
- Test statistic $Z = (\overline{X}_1 - \overline{X}_2)/\sigma$
Example: Do tall men get picked first?

Heights of K Husbands, by age of marriage

<table>
<thead>
<tr>
<th>Age of marriage</th>
<th>number</th>
<th>mean height (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>early (&lt; 30)</td>
<td>160</td>
<td>1735</td>
</tr>
<tr>
<td>late (≥ 30)</td>
<td>35</td>
<td>1716</td>
</tr>
</tbody>
</table>

Suppose we know that the standard deviation of height is 70mm.

\[ \sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 13.1 \]

\[ Z = \frac{1735 - 1716}{13.1} = 1.45 \]

One tailed p-value 0.0735

Conclusion: Insufficient evidence to reject the null.
Example: Do tall men get picked first?

Heights of K Husbands, by age of marriage

<table>
<thead>
<tr>
<th></th>
<th>Age of marriage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>early (&lt; 30)</td>
</tr>
<tr>
<td>number</td>
<td>160</td>
</tr>
<tr>
<td>mean height (mm)</td>
<td>1735</td>
</tr>
</tbody>
</table>

- Suppose we know that the standard deviation of height is 70mm

\[ \sigma = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} = 13.1 \]

\[ Z = \frac{1735 - 1716}{13.1} = 1.45 \]

- One tailed p-value 0.0735
- Conclusion: Insufficient evidence to reject the null
Example: Do tall men get picked first?

Heights of K Husbands, by age of marriage

<table>
<thead>
<tr>
<th>Age of marriage</th>
<th>early (&lt; 30)</th>
<th>late (≥ 30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>number</td>
<td>160</td>
<td>35</td>
</tr>
<tr>
<td>mean height (mm)</td>
<td>1735</td>
<td>1716</td>
</tr>
</tbody>
</table>

Suppose we know that the standard deviation of height is 70mm

\[ \sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 13.1 \]

\[ Z = \frac{1735 - 1716}{13.1} = 1.45 \]

One tailed p-value 0.0735

Conclusion: Insufficient evidence to reject the null

Weitzman & Conley: “From Assortative to Ashortative Coupling: Men’s Height, Height Heterogamy, and Relationship Dynamics in the United States”: Short men tend to get married later ... but to stay married longer...
Subsection 4

Z test for the difference between proportions
Z test for difference between proportions

- Observations from two different kinds of trials.
- Probabilities of success are $p_1$ and $p_2$
- Test $H_0 : p_1 = p_2$
- Observe $X_1$ successes from $n_1$ trials from pop 1
- Observe $X_2$ successes from $n_2$ trials from pop 2
- Standardized test statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

- with $\hat{p}_1 = X_1/n_1$, $\hat{p}_2 = X_2/n_2$ and $\hat{p} = (X_1 + X_2)/(n_1 + n_2)$
Example: Circumcision and AIDS

Study in Uganda: 70 circumcised men, 54 controls.

<table>
<thead>
<tr>
<th></th>
<th>circum.</th>
<th>non-circum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>70</td>
<td>54</td>
</tr>
<tr>
<td>infected</td>
<td>11</td>
<td>4</td>
</tr>
</tbody>
</table>

\[ \hat{p}_1 = \frac{11}{70} = 0.157 \]

\[ \hat{p}_2 = \frac{4}{54} = 0.121 \]

\[ \sigma = \sqrt{0.121 \times 0.157 \left( \frac{1}{70} + \frac{1}{54} \right)} = 0.059 \]

\[ Z = \frac{0.157 - 0.121}{0.059} = 1.41 \]

One tailed p-value 0.08
Example: Circumcision and AIDS

Study in Uganda: 70 circumcised men, 54 controls.

<table>
<thead>
<tr>
<th></th>
<th>circum.</th>
<th>non-circum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>70</td>
<td>54</td>
</tr>
<tr>
<td>infected</td>
<td>11</td>
<td>4</td>
</tr>
</tbody>
</table>

- $\hat{p}_1 = 11/70 = 0.157$
- $\hat{p}_2 = 4/54 = 0.121$
- $\sigma = \sqrt{0.121 \times 0.157 \left( \frac{1}{70} + \frac{1}{54} \right)} = 0.059$
- $Z = \frac{0.157 - 0.121}{0.059} = 1.41$
- One tailed p-value 0.08
- Conclusion: Insufficient evidence to reject the null
Section 3

The t test
Subsection 1

The simple $t$ test
Example: Kidney Dialysis

- Phosphate measure in blood of dialysis patients on six successive visits. Known to vary approximately according to normal distribution.
- One patient had the following measures in mg/dl:
  
  5.6, 5.1, 4.6, 4.8, 5.7, 6.4

- Suppose 4.0 or below is a dangerous level.
- Test at the 0.01 level whether the level might be that low.
- \( \bar{X} = 5.4 \text{mg/dl (empirical mean)} \)
- \( s = 0.67 \text{mg/dl (empirical standard deviation)} \)
Example: Kidney Dialysis

- Problem! We don’t know the SD!
- Estimate from the data:

\[ SD = \sigma \approx s = \sqrt{\frac{1}{n-1} \sum (X_i - \bar{X})^2} \]

- \( Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \)
- \( T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \)
- \( T \) has a different distribution to \( Z \)
- For \( s = \sigma \) they are the same...
- But sample variation in \( s \) leads to \( T \) having ‘heavier tails’
- \( n - 1 \) = number of degrees of freedom
Does it matter? Monte Carlo experiment

Critical value for $Z$ at 0.01 level is 2.3.

1. Compute 6 independent samples from $N(4.0, 0.67^2)$
2. With mean $\bar{X}_j$ and empirical variance $s_j$
3. Compute $T_j = (\bar{X}_j - 4.0)/(s_j/\sqrt{6})$
4. Repeat 10000 times
5. Did $T$ exceed 2.3 about 1% of the time?
The t test

- Suppose 4.0 or below is a dangerous level.
- Want a 1% chance of failing to recognise that the level is low
- H0: Average phosphate level = 4.0 mg/dl
- H1: Average phosphate level > 4.0 mg/dl

\[ T = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{5.4 - 4.0}{0.67/\sqrt{6}} = 5.12 \]

- Matlab: \( tinv(0.99,5)=3.36 \)
- i.e. the critical value is 3.36
- Observed \( T \) is bigger - reject H0
The simple t test (Student’s t test)

For \( n \) data points \( X_i \)
- If the **mean** of the data can be treated as Normal ...
- And our null hypothesis is \( \bar{X} = \mu \) for known \( \mu \) ...
- When we **estimate** the standard deviation \( \sigma \) using \( s \) ...
- We compute the standard error \( SE = s/\sqrt{n} \)
- We compute the test statistic \( T = (\bar{X} - \mu)/SE \) ...
- Under the null, \( T \sim t(0, 1, df = n - 1) \)

So the t test (also) computes \( p(T' \geq T) \) (one tailed)
Or \( p(|T'| \geq |T|) \) (two tailed)

**Important note:** more generally, \( df = n - p \) with \( p \) unknown parameters
Z test example

- British birthweights have Mean 3426g and SD 538g from a large sample
- Australian birthweights sampled (1 day)
- Australians have mean 3276
- Null: birthweights have the same mean
- Alternative: Australian babies are smaller
t test example

Now imagine that we were not provided with the standard deviation of the British sample.

- Null hypothesis: The Australian birthweights are samples from the same distribution as UK birth weights.
- With one unknown parameter: \( \sigma \)
- If the null hypothesis holds,

\[
\bar{X} = (X_1 + \cdots + X_{44})/44 \sim N(3426, SE^2)
\]

with \( SE = SD(X_i)/\sqrt{44} = 84.5 \)

- What is the probability that we would observe \( \bar{X} \leq 3276? \)
- Standardise: Observed

\[
t = (\bar{x} - 3426)/84.5 \sim t(df = 44 - 1) = -1.64
\]

\( p(t \leq -1.64) = 0.054 \). Bigger than before:

- Tails of \( t(n) \) larger than tails of \( N(0, 1) \)
- Because of uncertainty in \( \hat{\sigma} \)
Tails of the student t-distribution

5 degrees of freedom
Tails of the student t-distribution

30 degrees of freedom
Students t distribution

df = 1
Students t distribution

![Graph showing two t-distributions: df = 1 and df = 2. The graph displays the probability density function for T vs. df values.](image)
Students t distribution

df = 1

df = 2

df = 4
Students t distribution

![Graph showing multiple t-distributions with different degrees of freedom (df). The graph displays the probability density function for t-distributions with 1, 2, 4, and 10 degrees of freedom. The x-axis represents the T variable, ranging from -4 to 4, and the y-axis represents the probability density, ranging from 0 to 0.4. Each curve represents a different df value, with df = 1 having the widest spread and df = 10 being the narrowest.](image-url)
Students t distribution
Students t distribution

![Graph of Students' t distribution with different degrees of freedom (df).](image-url)
Quiz time!

Let \( t_\alpha(d) \) be the \( \alpha \) quantile of the Student T distribution with \( d \) d.f. i.e. the probability of \( t < \alpha \).

- \( t_{0.95}(2) > t_{0.95}(3) \)?

- \( t_\alpha(100) \) is a little smaller than \( Z_\alpha \)?

- 1% of the measurements of average phosphate will be within 2.3 s.d. of 4.0?
Let $t_\alpha(d)$ be the $\alpha$ quantile of the Student T distribution with $d$ d.f. i.e. the probability of $t < \alpha$.

- $t_{0.95}(2) > t_{0.95}(3)$? ▶ true
- $t_\alpha(100)$ is a little smaller than $Z_\alpha$? ▶ false - it can be very different for small $\alpha$
- 1% of the measurements of average phosphate will be within 2.3 s.d. of 4.0? ▶ false - that is only true for $Z_\alpha$
Quiz time!

Let $t_\alpha(d)$ be the $\alpha$ quantile of the Student T distribution with $d$ d.f. i.e. the probability of $t < \alpha$.

- $t_{0.95}(2) > t_{0.95}(3)$?
  - true

- $t_\alpha(100)$ is a little smaller than $Z_\alpha$?
  - false - it can be very different for small $\alpha$

- 1% of the measurements of average phosphate will be within 2.3 s.d. of 4.0?

false - that is only true for $Z_\alpha$
Let $t_\alpha(d)$ be the $\alpha$ quantile of the Student T distribution with $d$ d.f. i.e. the probability of $t < \alpha$.

- $t_{0.95}(2) > t_{0.95}(3)$?
  - true

- $t_\alpha(100)$ is a little smaller than $Z_\alpha$?
  - false - it can be very different for small $\alpha$

- 1% of the measurements of average phosphate will be within 2.3 s.d. of 4.0?
  - false - that is only true for $Z$
Subsection 2

The Matched sample t test
Example - Schizophrenia

- 15 healthy, 15 Schizophrenia sufferers
- Measure hippocampus volume
- schizophrenic:
  1.27, 1.63, 1.47, 1.39, 1.93, 1.26, 1.71, 1.67, 1.28, 1.85, 1.02, 1.34, 2.02, 1.59, 1.97
- healthy:
  1.94, 1.44, 1.56, 1.58, 2.06, 1.66, 1.75, 1.77, 1.78, 1.92, 1.25, 1.93, 2.04, 1.62, 2.08
- Test for equality of means at 0.05 level.
- Don't know SD.

<table>
<thead>
<tr>
<th></th>
<th>Unaff.</th>
<th>Schiz.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.76</td>
<td>1.56</td>
</tr>
<tr>
<td>SD</td>
<td>0.24</td>
<td>0.30</td>
</tr>
</tbody>
</table>
2-sample t test

- X: Schizophrenic data
- Y: Non-schizophrenic data
- H0: Samples came from the same distribution.
- If H0 true, then we can estimate \( \sigma \) by pooling X and Y
- Pooled sample variance:

\[
s_p^2 = \frac{(n_x - 1)s_X^2 + (n_y - 1)s_Y^2}{n_x + n_y - 2}
\]
2-sample t test

- X: Schizophrenic data
- Y: Non-schizophrenic data
- H0: Samples came from the same distribution.
- If H0 true, then we can estimate $\sigma$ by pooling X and Y
- Pooled sample variance:

$$s_p^2 = \frac{(n_x - 1)s_X^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

- Here $s_p = 0.27$
2-sample t test

- X: Schizophrenic data
- Y: Non-schizophrenic data
- H0: Samples came from the same distribution.
- If H0 true, then we can estimate $\sigma$ by pooling $X$ and $Y$
- Pooled sample variance:

$$s_p^2 = \frac{(n_x - 1)s_X^2 + (n_y - 1)s_Y^2}{n_x + n_y - 2}$$

Here $s_p = 0.27$

- Standard error $SE = s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} = 0.099$
2-sample t test

- $X$: Schizophrenic data
- $Y$: Non-schizophrenic data
- H0: Samples came from the same distribution.
- If H0 true, then we can estimate $\sigma$ by pooling $X$ and $Y$
- Pooled sample variance:

$$sp^2 = \frac{(nx - 1)s_X^2 + (ny - 1)s_y^2}{nx + ny - 2}$$

- Here $sp = 0.27$
- Standard error $SE = sp\sqrt{\frac{1}{nx} + \frac{1}{ny}} = 0.099$
- $T = \frac{X - Y}{SE} = 2.02$
2-sample t test

- $X$: Schizophrenic data
- $Y$: Non-schizophrenic data
- $H_0$: Samples came from the same distribution.
- If $H_0$ true, then we can estimate $\sigma$ by pooling $X$ and $Y$
- Pooled sample variance:

$$s_p^2 = \frac{(n_x - 1)s_X^2 + (n_y - 1)s_Y^2}{n_x + n_y - 2}$$

- Here $s_p = 0.27$
- Standard error $SE = s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} = 0.099$
- $T = \frac{\bar{X} - \bar{Y}}{SE} = 2.02$
- $df = n_x + n_y - 2 = 28$
2-sample t test

- X: Schizophrenic data
- Y: Non-schizophrenic data
- H0: Samples came from the same distribution.
- If H0 true, then we can estimate $\sigma$ by pooling X and Y
- Pooled sample variance:
  \[ s_p^2 = \frac{(n_x - 1)s_X^2 + (n_y - 1)s_Y^2}{n_x + n_y - 2} \]
- Here $s_p = 0.27$
- Standard error $SE = s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} = 0.099$
- $T = \frac{\bar{X} - \bar{Y}}{SE} = 2.02$
- $df = n_x + n_y - 2 = 28$
- Critical value at $p = 0.05$ is $T = 2.05$. Don’t reject.
What does this mean?

- The schizophrenic subjects have smaller hippocampal volume on average.
- BUT there’s a lot of variability overall - samples of 15 individuals can differ by this much purely by chance.
- Can we do anything to reduce this variability within groups, so we can see the difference between the groups more clearly?
Subsection 3

Matched sample t test
Matched case-control study

- Idea: Experiment and control group are in matched pairs, chosen to be similar in ways likely to affect what we’re measuring.

- Why?
- A lot of the variability will disappear (we hope) from the difference, since the matched pairs will vary together.
- Shared variance cancels out!
Example: Schizophrenia

The 30 subjects in the schizophrenia study were 15 matched pairs of monozygotic twins.

- Mean difference $E(X - Y) = 0.20$ (as before, i.e. $E(X) - E(Y)$)
- However, standard deviation $s_{diff} = s(D) = s(X - Y) = 0.238$
- Now a standard t test:
- Test H0: $\mu_{diff} = 0$
- $T = \frac{E(X - Y)}{s_{diff}/\sqrt{15}} = 3.25$
- Critical $T$ for $df = 14$ is 2.15
- Reject H0.
Section 4

The $\chi^2$ test
Subsection 1

The $\chi^2$ test
Example: suicides by birth month

Salib and Cortina-Borja examined death certificates of 26,886 suicides in England and Wales. Tabulated by month of birth.

Does spring birthday predispose to suicide?

<table>
<thead>
<tr>
<th>Month</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>527</td>
<td>1774</td>
<td>2301</td>
</tr>
<tr>
<td>Feb</td>
<td>435</td>
<td>1639</td>
<td>2074</td>
</tr>
<tr>
<td>Mar</td>
<td>454</td>
<td>1939</td>
<td>2393</td>
</tr>
<tr>
<td>Apr</td>
<td>493</td>
<td>1777</td>
<td>2270</td>
</tr>
<tr>
<td>May</td>
<td>535</td>
<td>1969</td>
<td>2504</td>
</tr>
<tr>
<td>Jun</td>
<td>515</td>
<td>1739</td>
<td>2254</td>
</tr>
<tr>
<td>Jul</td>
<td>490</td>
<td>1872</td>
<td>2362</td>
</tr>
<tr>
<td>Aug</td>
<td>489</td>
<td>1833</td>
<td>2322</td>
</tr>
<tr>
<td>Sep</td>
<td>476</td>
<td>1624</td>
<td>2100</td>
</tr>
<tr>
<td>Oct</td>
<td>474</td>
<td>1661</td>
<td>2135</td>
</tr>
<tr>
<td>Nov</td>
<td>442</td>
<td>1568</td>
<td>2010</td>
</tr>
<tr>
<td>Dec</td>
<td>471</td>
<td>1690</td>
<td>2161</td>
</tr>
</tbody>
</table>
Example: suicides by birth month

- Null hypothesis: Suicides are equally likely to have been born any day of the year. The probability of having been born in a given month is proportional to the number of days in the month.

- Test the null hypothesis at the 0.01 level.

- One approach: Divide into two groups, and use the Z test
Example: suicides by birth month

- Null hypothesis: Suicides are equally likely to have been born any day of the year. The probability of having been born in a given month is proportional to the number of days in the month.

- Test the null hypothesis at the 0.01 level.

- One approach: Divide into two groups, and use the Z test

- If we define spring as March–June, there are 122 days

- Under H0 $P($spring birthday$) = \frac{122}{365.25} = 0.34$

- Observed number spring = 9421

- Expected number spring = $0.334 \times 26886 = 8980$

- Standard Error = $\sqrt{0.334 \times 0.666 \times 26886} = 77.3$

- $Z = \frac{9421 - 8980}{77.3} = 5.71$

- Critical value $Z_{crit} = 2.58$

- $p$-value $\approx 10^{-8}$
Example: suicides by birth month

- **Problem:** We cheated!
- We used the largest choice of months
- Was that our hypothesis before we saw the data?
- **No.** We might have instead wondered if there were any months that were unusual
- Can we test all deviations simultaneously?
The $\chi^2$ test

- Test statistic that measures deviations in all $k$ categories simultaneously.

\[ \chi^2 = \sum \frac{(\text{observed}_i - \text{predicted}_i)^2}{\text{expected}_i} \]

- \[ Z_i = \frac{\text{observed}_i - \text{predicted}_i}{\text{standard error}} \]

- \[ \chi^2(k) = \sum_{k-1} Z_i^2 \] where $Z_i$ are independent

- Mathematical fact: If the number of observations is large then this chi-squared statistic has a certain distribution, called the $\chi^2$ distribution.

- How large is large? Rule of thumb: At least 5 expected in each cell of the table.
The $\chi^2$ test

- Test statistic that measures deviations in all $k$ categories simultaneously.

- $\chi^2 = \sum \frac{(\text{observed} - \text{predicted})^2}{\text{expected}}$

- $Z_i = \frac{\text{observed} - \text{predicted}}{\text{standard error}}$

- $\chi^2(k) = \sum_{k-1} Z_i^2$ where $Z_i$ are independent

- Mathematical fact: If the number of observations is large then this chi-squared statistic has a certain distribution, called the $\chi^2$ distribution.

- How large is large? Rule of thumb: At least 5 expected in each cell of the table.

- This distribution also has a ‘degrees of freedom’ parameter

- General rule: $\text{df} = \text{Num. data pts} - \text{parameters estimated} - 1$
The $\chi^2$ test

- Test statistic that measures deviations in all $k$ categories simultaneously.
- $\chi^2 = \sum \frac{(\text{observed}_i - \text{predicted}_i)^2}{\text{expected}}$
- $Z_i = \frac{\text{observed}_i - \text{predicted}_i}{\text{standard error}}$
- $\chi^2(k) = \sum_{k-1} Z_i^2$ where $Z_i$ are independent
- Mathematical fact: If the number of observations is large then this chi-squared statistic has a certain distribution, called the $\chi^2$ distribution.
- How large is large? Rule of thumb: At least 5 expected in each cell of the table.
- This distribution also has a ‘degrees of freedom’ parameter
- General rule: $df = \text{Num. data pts} - \text{parameters estimated} - 1$
- So here: $df = \text{cells} - \text{parameters estimated} - 1$
Calculating with $\chi^2$

- $\chi^2$ with $d$ degrees of freedom has mean $d$ and variance $2d$.
- It has a density proportional to

$$\chi^{d/2-1}e^{-x/2}$$

- We will not use this formula, instead using the computer (as usual!)
The $\chi^2$ Distribution
The $\chi^2$ Distribution
The $\chi^2$ Distribution
The $\chi^2$ Distribution

![Graph showing the $\chi^2$ distribution for different degrees of freedom (df)].

- df = 1
- df = 2
- df = 3
- df = 4
The $\chi^2$ Distribution
The $\chi^2$ Distribution
The $\chi^2$ Distribution

$df = 5$
The $\chi^2$ Distribution

![Graph showing two chi-squared distributions with different degrees of freedom (df = 5 and df = 10).]
The $\chi^2$ Distribution
The $\chi^2$ Distribution

![Graph of the chi-squared distribution with different degrees of freedom (df = 5, df = 10, df = 20, df = 30).]
Simple example: Testing a die

<table>
<thead>
<tr>
<th>side</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>freq</td>
<td>16</td>
<td>15</td>
<td>4</td>
<td>6</td>
<td>14</td>
<td>5</td>
</tr>
</tbody>
</table>

- Test the null hypothesis that all sides are equally likely at the 0.01 level

\[ X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \]  

\[ = \frac{(16 - 10)^2}{10} + \cdots + \frac{(5 - 10)^2}{10} = 15.4 \]
Simple example: Testing a die

- $X^2 = 15.4$ with 5 degrees of freedom
- Matlab: $1 - \text{chi2cdf}(15.4,5) = 0.00885$

![Chi-square distribution graph with critical values at p=0.05 and p=0.01]
Simple example: Testing a die

- $X^2 = 15.4$ with 5 degrees of freedom
- Matlab: 1-chi2cdf(15.4, 5) = 0.00885
- Reject at the 0.01 level
Salib and Cortina-Borja examined death certificates of 26,886 suicides in England and Wales. Tabulated by month of birth. Does spring birthday predispose to suicide?

<table>
<thead>
<tr>
<th>Month</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>527</td>
<td>1774</td>
<td>2301</td>
</tr>
<tr>
<td>Feb</td>
<td>435</td>
<td>1639</td>
<td>2074</td>
</tr>
<tr>
<td>Mar</td>
<td>454</td>
<td>1939</td>
<td>2393</td>
</tr>
<tr>
<td>Apr</td>
<td>493</td>
<td>1777</td>
<td>2270</td>
</tr>
<tr>
<td>May</td>
<td>535</td>
<td>1969</td>
<td>2504</td>
</tr>
<tr>
<td>Jun</td>
<td>515</td>
<td>1739</td>
<td>2254</td>
</tr>
<tr>
<td>Jul</td>
<td>490</td>
<td>1872</td>
<td>2362</td>
</tr>
<tr>
<td>Aug</td>
<td>489</td>
<td>1833</td>
<td>2322</td>
</tr>
<tr>
<td>Sep</td>
<td>476</td>
<td>1624</td>
<td>2100</td>
</tr>
<tr>
<td>Oct</td>
<td>474</td>
<td>1661</td>
<td>2135</td>
</tr>
<tr>
<td>Nov</td>
<td>442</td>
<td>1568</td>
<td>2010</td>
</tr>
<tr>
<td>Dec</td>
<td>471</td>
<td>1690</td>
<td>2161</td>
</tr>
</tbody>
</table>
Suicides by birth month

- **H0:** Suicides are equally likely to have been born any day of the year. The probability of having been born in a given month is proportional to the number of days in the month.
- e.g., \( P(\text{January}) = \frac{31}{365.25} = 0.0849 \)
- \( P(\text{February}) = \frac{28.25}{365.25} = 0.0773 \)

\[
\chi^2 = \sum \frac{(\text{expected} - \text{observed})^2}{\text{expected}}
\]

\[
= \frac{(2281 - 2301)^2}{2281} + \cdots + \frac{(2281 - 2161)^2}{2281}
\]

\[
= 72.4
\]

- df = 12 − 1 = 11
- p-value, from Matlab:
  \( 1 - \text{chi2cdf}(72.4,11) = 0.00000000004262901 \)
- i.e. \( p < 10^{-10} \). Reject H0.
- The variation is not due to chance variation.
Suicides by birth month

Looking at just the data for females:

\[
\chi^2 = \frac{(492 - 527)^2}{492} + \cdots + \frac{(492 - 471)^2}{492}
\]  
\[
= 17.4
\]

Matlab: \text{chi2inv(.95,11)=19.68}

We do not reject H0 at the 5% level....

“The difference in frequency of suicides by birth month among women is NOT statistically significant. It could be explained by chance variation.”
Section 5

Non-parametric tests
Subsection 1

Why we need non-parametric tests
An experiment

“If a newborn infant is held under his arms and his bare feet are permitted to touch a flat surface, he will perform well-coordinated walking movements similar to those of an adult... Normally, the walking and pacing reflexes disappear by about 8 weeks.”

Observation: If the infant exercises this reflex, it does not disappear.

Hypothesis: Maintaining this reflex will help children learn to walk earlier.
How do we test this hypothesis?

- Idea: Do weekly exercises with a newborn. See when he/she starts walking.
- Result: 10 months.
- Problem: Is that long or short?
- New Idea: Do weekly exercises with a newborn. Don’t do weekly exercises with another newborn. See which one starts walking first
- Result: mean with exercise 10.1 months
- without exercise 11.7 months
- Problem: Newborns don’t all start walking at the same age, regardless of exercise.
**t test for walking data**

**Age in months at first walking**

<table>
<thead>
<tr>
<th>Treatment (Exercise)</th>
<th>Control (No Exercise)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.0</td>
<td>11.5</td>
</tr>
<tr>
<td>9.5</td>
<td>12.0</td>
</tr>
<tr>
<td>9.75</td>
<td>9.0</td>
</tr>
<tr>
<td>10.0</td>
<td>11.5</td>
</tr>
<tr>
<td>13.0</td>
<td>13.25</td>
</tr>
<tr>
<td>9.5</td>
<td>13.0</td>
</tr>
</tbody>
</table>

Mean 10.1  11.7  
SD 1.45   SD 1.52
The Treatment numbers are generally smaller, but not always. Could the difference be merely due to chance?

- Two sample \( t \) test
- \( H_0: \mu_T = \mu_C \)
- \( H_1: \mu_T < \mu_C \) (one-tailed test)
- Test at 0.05 signif. level 12-2=10 d.f.
- Critical value 1.81

Pooled sample variance:

\[
sp = \sqrt{\frac{(6-1)1.45^2 + (6-1)1.52^2}{6+6-2}} = 1.48
\]

Standard error \( SE = sp\sqrt{1/6 + 1/6} = 0.85 \)

\[
T = \frac{\bar{X} - \bar{Y}}{SE} = 1.85
\]

Reject Null
What if the distribution isn’t Normal?

- Under the null...
- A bimodal distribution?
- Mean of 6 samples will not be Normal!
Simulation study using replicate data
Simulation study using replicate data
Nonparametric tests

- Idea: Come up with test statistics whose significance level doesn’t depend on the distribution that the data came from.
- Advantage: We reject with the right probability if the null hypothesis is true.
- Drawback: We lose power. That is, we need a larger sample to reject the null if its false.
- We focus on two tests that the median of two distributions is equal
- They are varyingly sensitive to other differences in distribution
Subsection 2

Mann-Whitney U test
Mann-Whitney U test

Also called the Wilcoxon two sample Rank-sum test

- We have samples $X_1, \ldots, X_{n_x}$ and $Y_1, \ldots, Y_{n_y}$ with unknown distributions.
- H0: medians are the same.
- H1: $m_x > m_y$ (one-tailed) or $m_x \neq m_y$ (two-tailed)
Mann-Whitney calculation

- **Step 1:** Put the data in order: \( Z_i = \text{sort}(X, Y) \)
- **Step 2:** Write down the ranks: \( r_i = i \)
- **Step 2.5:** Combine ties: \( r_i = \text{mean}(r_j | Z_j = Z_i) \)
- **Step 3:** Add up the ranks: \( R_X = \sum_{i=1}^{N} \text{s.t. } i \in X r_i \), \( R_Y = \sum_{i=1}^{N} \text{s.t. } i \in Y r_i \)
- **Step 4:** Compute \( R = \min(R_X, R_Y) \)
- **Step 5:** Compute significance. Under H0, each ranking should be random from \((1,N)\) with replacement.

Matlab: \( p = \text{ranksum}(X,Y) \)

This computes all permutations for small \( N \), and uses the normal approximation to the sum for large \( N \).
Computation for walking babies

<table>
<thead>
<tr>
<th>9.0</th>
<th>9.0</th>
<th>9.5</th>
<th>9.5</th>
<th>9.75</th>
<th>10.0</th>
<th>11.5</th>
<th>11.5</th>
<th>12</th>
<th>13.0</th>
<th>13.0</th>
<th>13.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>1.5</td>
<td>3.5</td>
<td>3.5</td>
<td>5</td>
<td>6</td>
<td>7.5</td>
<td>7.5</td>
<td>9</td>
<td>10.5</td>
<td>10.5</td>
<td>12</td>
</tr>
</tbody>
</table>

\[ R_X = 30, \quad R_Y = 48 \]

\[ R = R_X = 30 \text{ since } X \text{ and } Y \text{ have the same size} \]

\[ p=\text{ranksum}(X,Y,'\text{tail'},'\text{left'}) = 0.085 \text{ (recent matlab versions only!)} \]

\[ p=\text{ranksum}(X,Y)/2 = 0.085 \text{ (old versions only implement two-tailed test)} \]

Retain the null hypothesis
Subsection 3

Paired value tests
The sign test

- We have paired samples $X_1, Y_1, \cdots, X_n, Y_n$
- Work only with $I_i = \mathbb{I}(X_i - Y_i > 0)$, i.e. 1 if $X_i > Y_i$, 0 otherwise.
- H0: $S = \sum_i I_i$ is a Binomial RV with success probability $p = 0.5$
- H1: $p > 0.5$
- Matlab: `binocdf(S, n, 0.5)` (left tail)
- Matlab: `binocdf(n - S, n, 0.5)` (right tail)
- Matlab: $2 \min(\text{binocdf}(S, n, 0.5), \text{binocdf}(n - S, n, 0.5))$ (two-tailed)
Example: Schizophrenia

\[ \begin{array}{cccccccccccccccc}
\text{s.} & 1.27 & 1.63 & 1.47 & 1.39 & 1.93 & 1.26 & 1.71 & 1.67 & 1.28 & 1.85 & 1.02 & 1.34 & 2.02 & 1.59 & 1.97 \\
\text{h.} & 1.94 & 1.44 & 1.56 & 1.58 & 2.06 & 1.66 & 1.75 & 1.77 & 1.78 & 1.92 & 1.25 & 1.93 & 2.04 & 1.62 & 2.08 \\
\text{d.} & + & - & + & + & + & + & + & + & + & + & + & + & + & + & + \\
\end{array} \]

- Observed \( X > Y \) for 14 out of 15 twins
- Compute \( P(S = 14, 15) = \binom{15}{14} + \binom{15}{15} \left( \frac{1}{2} \right)^{15} = 0.0005 \)
- Or we can use the Normal approximation, p-value \( p = 0.0004 \): 

\[
Z = \frac{14 - 0.5 \times 15}{\sqrt{0.5 \times 0.5 \times 15}} = 3.36
\]
Idea: It makes sense to consider, not just if \( X > Y \), but whether this happens for big or small numbers.

It might be that there are equal numbers of + and − differences, but the + are bigger

Step 1: list differences ordered by absolute value

Step 2: Calculate \( W = \min(\sum_{i: X > Y} r_i, \sum_{i: X < Y} r_i) \)

Step 3: Compare \( W \) to the distribution that would be given under the null, that the ranks are unrelated to the signs

For small \( n \) this means enumerating all options (enumeration is not fun)
Wilcoxon one sample sign-rank test

For \( n \approx 10 \) or more, \( Z = \frac{W - \mu_W}{\sigma_W} \sim N(0, 1) \), and a \( Z \) test can be used,

Where

\[
\mu_W = \frac{1}{2} + \frac{n(n + 1)}{4}
\]

and

\[
\sigma_W = \sqrt{\frac{n(n + 1)(2n + 1)}{24}}.
\]
Example: Schizophrenia

s. 1.27 1.63 1.47 1.39 1.93 1.26 1.71 1.67 1.28 1.85 1.02 1.34 2.02 1.59 1.97
h. 1.94 1.44 1.56 1.58 2.06 1.66 1.75 1.77 1.78 1.92 1.25 1.93 2.04 1.62 2.08
d. 0.67 -0.19 0.09 0.19 0.13 0.40 0.04 0.10 0.50 0.07 0.23 0.59 0.02 0.03 0.11
Leading to the sorted differences
0.02 0.03 0.04 0.07 0.09 0.10 0.11 0.13 -0.19 0.19 0.23 0.40 0.50 0.59 0.67
1  2  3  4  5  6  7  8  9.5  9.5 11 12 13 14 15

- Sums: \( R_x = 110.5 \), \( R_y = 9.5 \)
- So \( R = 9.5 \)
- \( \mu_R = 0.5 + 15 \times 16/4 = 60.5 \)
- \( \sigma_R = \sqrt{15 \times 16 \times (30 + 1)/24} = 17.6 \)
- \( Z = (R - \mu_R)/\sigma_R = -2.897 \)
- So the p-value is 0.0019
Section 6

The menagerie of tests
But which test should I use???

- These are just a small subset of the possible tests available
- Many of them have different options:
  - How did we make the data look like a Normal?
  - Parameters in model - leads to degrees of freedom
  - Tails of test
  - Pairing data
  - etc
- How to decide?
Model assumptions

The key is to **make appropriate assumptions**

▶ Are your data **independent** and **random** samples from a defined population? (All tests considered here)

▶ Are you primarily testing for a difference in the **location** in the two distributions? ($Z, t$, non-parametric tests)

▶ Or the variance of many random variables? ($\chi^2$)

▶ Is the underlying distribution normal? ($Z$ test, $t$ test, $\chi^2$ test)

▶ Or do we want to avoid assumptions about it, and test the median? (Mann-Whitney, Wilcoxon)

▶ If so, do we want to test the whole distribution? (Wilcoxon)

▶ Are there unknown parameters? ($t$ test, $\chi^2$ test)

Look up the specific assumptions when you use a test!
Z Test

H1 tests for a difference in the mean of two distributions. H0 makes the following assumptions:

- Independent random samples
- Mean is approx. normal
- Continuous variable (recall: continuity correction)
- Known variance
H1 tests for a difference in the mean of two distributions. H0 makes the following assumptions:

- Independent random samples
- Mean is approx. normal
- Continuous variable (recall: continuity correction)
- Unknown variance, estimated using $s$
χ² Test

H1 tests for a difference in the variance of \( n \) distributions. H0 makes the following assumptions:

- Sum of independent random samples
- Whose mean is approx. normal (hence sample size > 5 desirable)
- Continuous variable (recall: continuity correction)
- Unknown variance, target of the test
Sign Test

H1 tests for a difference in the median of two distributions. H0 makes the following assumptions:

- Independent random samples
Mann-Whitney U Test

H1 tests for a difference in the median of two distributions. H0 makes the following assumptions:
  ▶ Independent random samples
  ▶ Continuous variable (recall: tied value correction)
This is ‘just’ the unpaired Wilcoxon Test.
Wilcoxon Test

H1 tests for a difference in either the median of two paired distributions.  
H0 makes the following assumptions:

- Independent random samples
- Continuous variable (recall: tied value correction)
- symmetric distribution of differences
Other tests you might encounter

We have looked at tests for the location of one or more distributions. Other important cases are:

- **F-test**: Compares the variance of two distributions. Used in Analysis of variance (ANOVA).
- **Kolmogorov-Smirnov test**: A non-parametric test for whether two distributions are the same, based on the maximum deviation from the empirical cumulative density functions.
Conceptually different tests

- **Likelihood ratio test:** The most important, because it uses a specific alternative hypothesis. It considers two models, one of which can be more complicated than the other (but nested). It accounts for the difference in complexity. But: you have to define the two models explicitly.

- **Monte carlo tests:** If we don’t know the distribution of the data, but can simulate from it, we can simulate $k-1$ test statistics and report the p-value as the quantile of the true test.

- **Bayesian tests:** A very different paradigm, Bayesian tests usually ask whether a parameter estimate falls outside of some range, given the data and some prior knowledge of the parameter.
Conceptually different tests

- **Likelihood ratio test**: the most important, because it uses a specific alternative hypothesis. It considers two models, one of which can be more complicated than the other (but nested). It accounts for the difference in complexity. But: you have to define the two models explicitly.
Conceptually different tests

- **Likelihood ratio test**: the most important, because it uses a specific alternative hypothesis. It considers two models, one of which can be more complicated than the other (but nested). It accounts for the difference in complexity. But: you have to define the two models explicitly.

- **Monte carlo tests**: If we don’t know the distribution of the data, but can simulate from it, we can simulate $k - 1$ test statistics and report the p-value as the quantile of the true test.
Conceptually different tests

- **Likelihood ratio test**: the most important, because it uses a specific alternative hypothesis. It considers two models, one of which can be more complicated than the other (but nested). It accounts for the difference in complexity. But: you have to define the two models explicitly.

- **Monte carlo tests**: If we don’t know the distribution of the data, but can simulate from it, we can simulate \( k - 1 \) test statistics and report the p-value as the quantile of the true test.

- **Bayesian tests**: A very different paradigm, Bayesian tests usually ask whether a parameter estimate falls outside of some range, given the data and some prior knowledge of the parameter.