PLAN

- Lecture 1: estimation
- Lecture 2: hypothesis testing
- Somewhere: statistical tests and when to use what
Estimation

Colm O’Dushlaine
Estimation

Estimation and hypothesis testing (next lecture) are procedures used to make statistical inferences
Overview of lecture

- Point estimates
- Sampling distribution of the mean
- Standard error of the mean (SEM)
- The t distribution
- Odds ratios
- Correlation
Point estimates

- **Point** = a single value used as an estimate of the population parameter.
  The sample mean is a point estimate of the population mean
  A limitation is that it is unlikely to estimate population parameter exactly – it is just a sample
  An *Interval* can be more informative – based on the sampling distribution and therefore incorporates a measure of uncertainty
  The *estimate* is the particular numeric value of the *estimator* of a population parameter, e.g. sample mean

Can use mean and standard error of mean to get an interval
Sampling distribution of the mean

- 10 people from population, mean is not identical
- 1000 * 10 random sets of people → relative frequency distribution
- This is a good approximation to the sampling distribution of the mean
- Best estimate of what is true for the relevant population
- So, estimation is the process of determining a likely value for a variable in the survey population, based on information collected from the sample
Sampling distribution of the mean (2)

- Variability of a sample means:
  - Will be less among the means of large samples than small samples
  - Will be less than the variability of the individual observations in the population
  - Will increase with more variability (stdev) among the individual values in the population

- Central limit theorem: the distribution of the sample means will be nearly normal whatever the distribution of the variable in the population, as long as the samples are large enough (also, c.f. Law of large numbers)
C.L.T.

Law of large numbers
(long term stability of the mean of a random variable)
The sampling distribution of a statistic will be normal or nearly normal, if any of the following conditions apply:

- The population distribution is normal
- The sampling distribution is symmetric, unimodal, without outliers, and the sample size is 15 or less
- The sampling distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40.
- The sample size is greater than 40, without outliers
Standard error of the mean

- So, have a sample and want to make an estimate about the mean of the population
- The S.E.M is the standard deviation of the sampling distribution of the mean
- Helps to quantify how good our estimate of the mean is of the true, unknown, population mean
- s.d. of many samples = $\sigma / \sqrt{n}$, where $\sigma$ is the s.d. in the population
- This can be used to construct a confidence interval
Confidence interval

- An interval estimate of a population parameter. Instead of estimating the parameter by a single value, an interval likely to include the parameter is given.

- Confidence intervals are used to indicate the reliability of an estimate. How likely the interval is to contain the parameter is determined by the confidence level or confidence coefficient.

- Increasing the desired confidence level will widen the confidence interval.
General Form of Confidence Interval

Estimate $\pm$ (critical value from distribution).(standard error)

Were the procedure to be repeated on multiple samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter e.g. 95% of the time.
Confidence interval

Fifty 95% confidence intervals based on 50 samples from a hypothetical normal distribution with a mean of 37°C and SD of 0.407°C
Sampling distribution of $\bar{X}$

95% of the $\bar{X}$'s lie between $\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$. 

$\mu - 1.96 \frac{\sigma}{\sqrt{n}} \quad \mu \quad \mu + 1.96 \frac{\sigma}{\sqrt{n}} \quad \bar{X}$
Confidence Interval for $\mu$ when $\sigma$ is known

- A 95% confidence interval for $\mu$ if $\sigma$ is known is given by:

$$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$
Rationale for Confidence Interval

- From the sampling distribution of $\bar{X}$ conclude that $\mu$ and $\bar{X}$ are within 1.96 standard errors ($\frac{\sigma}{\sqrt{n}}$) of each other 95% of the time.

- Otherwise stated, 95% of the intervals contain $\mu$.

- So, the interval $\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$ can be taken as an interval that typically would include $\mu$. 

Example

- A random sample of 80 tablets had an average potency of 15mg. Assume \( \sigma \) (population stdev) is known to be 4mg.
- \( \bar{X} = 15, \sigma = 4, n=80 \)
- A 95% confidence interval for \( \mu \) is

\[
15 \pm 1.96 \times \frac{4}{\sqrt{80}}
\]

\[
= (14.12, 15.88)
\]
Other notes on C.I.

- 95% C.I. for the sample mean is the range of values which contains the true population mean with probability 0.95
- We assume that the sample, e.g. 200 patients is reflective of all patients with the disease
- If distribution of mean is e.g. lognormal, need to do other things, e.g. log transform the original values and calculate the C.I. and then back-transform to get C.I. for the geometric mean, OR use median
- This example uses the normal distribution but the t-distribution is generally used instead...
Confidence Interval for $\mu$ when $\sigma$ is unknown

- Nearly always $\sigma$ (population stdev) is unknown and is estimated using sample standard deviation $s$
- The value 1.96 in the confidence interval is replaced by a new quantity, $t_{0.025}$
- The 95% confidence interval when $\sigma$ is unknown is: $\bar{x} \pm t_{0.025} \times \frac{s}{\sqrt{n}}$
The t distribution
Student’s $t$ Distribution

- Mean of a sample from a Normal distribution with unknown variance has a distribution similar to, but not the same as, a Normal distribution ($Z$)
  - Symmetric and bell-shaped
  - Has mean = 0 but has a larger standard deviation

- Exact shape depends on a parameter called **degrees of freedom** (df) which is related to sample size – the one parameter of $t$ distribution

- It is used to cope with uncertainty resulting from estimating the standard deviation from a sample, whereas if the population standard deviation were known, a normal distribution could be used
\[ \bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}} \]
\[ \bar{x} \pm t_{0.025} \times \frac{s}{\sqrt{n}} \]
The derivation of the $t$-distribution was first published in 1908 by William Sealy Gosset, while he worked at a Guinness Brewery in Dublin. He was prohibited from publishing under his own name, so the paper was written under the pseudonym *Student*. The $t$-test and the associated theory became well-known through the work of R.A. Fisher, who called the distribution "Student's distribution".

In 1908 (1876-1937)
Student’s $t$ distribution

- When $N$ is e.g. $>100$, can use normal distribution, but no point as effectively the same
- Therefore easier to just use the $t$ distribution for all samples as it handles all sizes
- Used in confidence intervals and hypothesis tests
Student’s $t$ distribution for 3, 10 df and standard Normal distribution
Definition of $t_{0.025}$ values

![Diagram showing the definition of $t_{0.025}$ values with the critical t-values at -4.5, -3.0, -1.5, 0.0, 1.5, 3.0, and 4.5. The shaded areas represent the 0.025 and 0.95 confidence levels.](image-url)
Example

26 measurements of the potency of a single batch of tablets in mg per tablet are as follows

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>498.38</td>
<td>489.31</td>
<td>505.50</td>
<td>495.24</td>
<td>490.17</td>
<td>483.2</td>
</tr>
<tr>
<td>488.47</td>
<td>497.71</td>
<td>503.41</td>
<td>482.25</td>
<td>488.14</td>
<td></td>
</tr>
<tr>
<td>492.22</td>
<td>483.96</td>
<td>473.93</td>
<td>463.40</td>
<td>493.65</td>
<td></td>
</tr>
<tr>
<td>499.48</td>
<td>496.05</td>
<td>494.54</td>
<td>508.58</td>
<td>488.42</td>
<td></td>
</tr>
<tr>
<td>463.68</td>
<td>492.46</td>
<td>489.45</td>
<td>491.57</td>
<td>489.33</td>
<td></td>
</tr>
</tbody>
</table>
Example (contd.)

- $\bar{x} = 490.096$, and $s = 10.783$ mg per tablet
- $t_{0.025}$ with df = 25 is 2.06
  \[
  \bar{x} \pm t_{0.025} \times \frac{s}{\sqrt{n}} = 490.096 \pm 2.06 \times \frac{10.783}{\sqrt{26}} \\
  = 490.096 \pm 4.356
  \]
- So, the batch potency lies between 485.74 and 494.45 mg per tablet
Odds ratios and Correlation
Comparing groups 1 - Odds ratio

- Measure of effect size
- Ratio of the odds of an event occurring in one group to the odds of it occurring in another group
- If the probabilities of the event in each of the groups are \( p \) (first group) and \( q \) (second group), then the odds ratio is:

\[
\frac{p/(1-p)}{q/(1-q)} = \frac{p(1-q)}{q(1-p)}. 
\]

- \( 1 \) \( \rightarrow \) equally likely in both groups
- 100 men, 90 have drunk wine in the previous week. 100 women, 20 have drunk wine in the same period. The odds of a man drinking wine are 9:1, while the odds of a woman drinking wine are 1:4 = 0.25:1. The odds ratio is thus 9/0.25, or 36, showing that men are much more likely to drink wine than women.

\[
\frac{0.9}{0.2} = \frac{0.9 \times 0.8}{0.1 \times 0.2} = \frac{0.72}{0.02} = 36. 
\]
### Comparing groups 1 - Odds ratio (2)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Total number of patients treated</th>
<th>Number who achieved at least 50% pain relief</th>
<th>Number who did not achieve at least 50% pain relief</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ibuprofen 400 mg</td>
<td>40</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>Placebo</td>
<td>40</td>
<td>7</td>
<td>33</td>
</tr>
</tbody>
</table>

#### Calculations made from these results

<table>
<thead>
<tr>
<th>Calculation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental event rate (EER, event rate with ibuprofen)</td>
<td>$22/40 = 0.55$ or $55%$</td>
</tr>
<tr>
<td>Control event rate (CER, event rate with placebo)</td>
<td>$7/40 = 0.18$ or $18%$</td>
</tr>
<tr>
<td>Experimental event odds</td>
<td>$22/18 = 1.2$</td>
</tr>
<tr>
<td>Control event odds</td>
<td>$7/33 = 0.21$</td>
</tr>
<tr>
<td>Odds ratio</td>
<td>$1.2/0.21 = 5.7$</td>
</tr>
<tr>
<td>Relative risk (EER/CER)</td>
<td>$0.55/0.18 = 3.1$</td>
</tr>
<tr>
<td>Relative risk increase (100(EER-CER)/CER) as a percentage</td>
<td>$100((0.55-0.18)/0.18) = 206%$</td>
</tr>
<tr>
<td>Absolute risk increase or reduction (EER-CER)</td>
<td>$0.55 - 0.18 = 0.37$ (or $37%$)</td>
</tr>
<tr>
<td>NNT (1/(EER-CER))</td>
<td>$1/(0.55 - 0.18) = 2.7$</td>
</tr>
</tbody>
</table>
Comparing groups 1 - Odds ratio (3)

- **Relative risk/risk ratio (RR)** is the risk of an event (or of developing a disease) relative to exposure. A ratio of the probability of the event occurring in the exposed group versus a non-exposed group.

- Probability of developing lung cancer among smokers was 20% and among non-smokers 1%, then the relative risk of cancer associated with smoking would be 20. Smokers would be twenty times as likely as non-smokers to develop lung cancer.

- Asymptotically approaches odds ratio for small probabilities.
Comparing groups 1 - Odds ratio (4)

- In medical research, the odds ratio is favoured for case-control studies and retrospective studies. Relative risk is used in randomized controlled trials and cohort studies.

- Approaches like poisson regression (counts of events per unit exposure) have relative risk interpretations: estimated effect of an explanatory variable is multiplicative on the rate, and thus leads to a risk ratio or relative risk.

- Logistic regression (for binary outcomes, or counts of successes out of a number of trials) are interpreted in odds-ratio terms: effect of an explanatory variable is multiplicative on the odds and thus leads to an odds ratio.
Comparing groups 1 - Odds ratio (7)

- RR not valid in C-C studies because can easily vary RR by changing case/control numbers
- Need a method based on calculations within each group
<table>
<thead>
<tr>
<th>Treatment</th>
<th>Total number of patients treated</th>
<th>Number who achieved at least 50% pain relief</th>
<th>Number who did not achieve at least 50% pain relief</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ibuprofen 400 mg</td>
<td>40</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>Placebo</td>
<td>40</td>
<td>7</td>
<td>33</td>
</tr>
</tbody>
</table>

Calculations made from these results

<table>
<thead>
<tr>
<th>Event Rate and Odds</th>
<th>Calculation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental event rate (EER, event rate with ibuprofen)</td>
<td>22/40 = 0.55 or 55%</td>
<td></td>
</tr>
<tr>
<td>Control event rate (CER, event rate with placebo)</td>
<td>7/40 = 0.18 or 18%</td>
<td></td>
</tr>
<tr>
<td>Experimental event odds</td>
<td>22/18 = 1.2</td>
<td></td>
</tr>
<tr>
<td>Control event odds</td>
<td>7/33 = 0.21</td>
<td></td>
</tr>
<tr>
<td>Odds ratio</td>
<td>1.2/0.21 = 5.7</td>
<td></td>
</tr>
<tr>
<td>Relative risk (EER/CER)</td>
<td>0.55/0.18 = 3.1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relative Risk Increase (100(EER-CER)/CER) as a percentage</th>
<th>Calculation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative risk increase or reduction (EER-CER)</td>
<td>100((0.55-0.18)/0.18) = 206%</td>
<td></td>
</tr>
<tr>
<td>NNT (1/(EER-CER))</td>
<td>1/(0.55 - 0.18) = 2.7</td>
<td></td>
</tr>
</tbody>
</table>

Odds ratio

Relative risk (EER/CER)
Comparing groups 1 - Odds ratio (8)

- A bit misleading sometimes, especially in genetics

- Recall *Psynomics*: gene test for bipolar, GRK3...200-300% increase in risk...WOW!...but in reality the absolute risk only goes from 1% to 2/3%

- Other examples?
Comparing groups 2 - correlation

- 2 continuous variables, test association
- Pearson (assumes normality) - ‘r’ (range: -1 → 1)
- ‘r’ measures the scatter around a linear trend line
- N.B. test normality beforehand using a Shapiro-Wilk W test + examine with scatterplot
- Transform either or both variables if necessary OR, use a non-parametric test
Examples of correlations and coefficients
v. significant correlation but no major trend

\[ r = 0.2, \quad p < 0.0005 \]
Comparing groups 2 - correlation (2)

- **Misuse 1**: e.g. use correlation to relate blood pressure and oestrogen levels in pregnant women with varying numbers of observations at different gestational ages. All observations need to be independent.

- **Misuse 2**: e.g. 10 variables can generate 45 correlations. This means we can “pick and choose”...wrong!

- **Misuse 3**: spurious correlations involving time, e.g. divorce rate and cost of oil.

- **Misuse 4**: selective sampling. If looking at hypertensive men, adding a few new samples that are all over-weight or in stressful jobs to “improve correlation” is wrong as does not represent random sampling from a hypertensive cohort.
Non-parametric correlation

- Less power but less likely to give distorted results when assumptions about the data fail (e.g. normality)
- Chi-square, Point biserial correlation, Spearman's $\rho$, Kendall's $\tau$, and Goodman and Kruskal's lambda
Correlation and causality

- Correlation does not imply causality
- Correlation cannot be validly used to infer a causal relationship between the variables
- Causes underlying the correlation, if any, may be indirect and unknown. Consequently, establishing a correlation between two variables is not a sufficient condition to establish a causal relationship
- Correlation between age and height 😊
- Correlation between mood and health 😞
Correlation and linearity

- Anscombe's quartet. The four y variables have the same mean (7.5), standard deviation (4.12), correlation (0.81) and regression line (y = 3 + 0.5x).

- *** Look at your data! ***
Summary

- Point estimates
  - Terms and definitions
- Sampling distribution of the mean
  - Basic concept of a sample and inferences we can make about population at large
- Standard error of the mean
  - How it relates to C.I.
- Confidence intervals
  - How they relate to sample estimates about the population
- \( t \) distribution
  - When used and why
- Odds ratios
  - What they are
  - Relationship with relative risk
- Correlation
  - Uses, misuses
Resources

- *Practical statistics for medical research* (Douglas Altman)
- Wikipedia (as a *secondary resource!*)
- Oxford stats pages
- UCLA stats pages
- *Teaching statistics* (Gelman and Nolan)
In 2 weeks time...

- Hypothesis testing
- P-values and what they mean, and limitations
- 1 vs 2-tailed tests
- Estimation vs. Hypothesis testing