Inference about two groups

Differences between means

Suppose I measure the treatment response for 10 subjects getting treatment A and 10 subjects getting treatment B.

How different are the responses of the two treatments?

→ I am not interested in these particular subjects, but in the treatments generally.
Suppose that
- \(X_1, X_2, \ldots, X_n\) are iid Normal(mean=\(\mu_A\), SD=\(\sigma\)), and
- \(Y_1, Y_2, \ldots, Y_m\) are iid Normal(mean=\(\mu_B\), SD=\(\sigma\)).

Then
\[ E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_A - \mu_B \]
\[ SD(\bar{X} - \bar{Y}) = \sqrt{SD(\bar{X})^2 + SD(\bar{Y})^2} = \sigma \sqrt{\frac{1}{n} + \frac{1}{m}} \]

Note: If \(n = m\), then \(SD(\bar{X} - \bar{Y}) = \sigma \sqrt{2/n}\).

**Pooled estimate of the population SD**

We have two different estimates of the populations’ SD, \(\sigma\):
\[ \hat{\sigma}_A = S_A = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} \quad \hat{\sigma}_B = S_B = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{m-1}} \]

We can use all of the data together to obtain an improved estimate of \(\sigma\), which we call the “pooled” estimate.
\[ \hat{\sigma}_{pooled} = \sqrt{\frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{n + m - 2}} \]
\[ = \sqrt{\frac{S_A^2(n-1) + S_B^2(m-1)}{n + m - 2}} \]

Note: If \(n = m\), then \(\hat{\sigma}_{pooled} = \sqrt{\left(\frac{S_A^2 + S_B^2}{2}\right)}\)
Estimated SE of $(\bar{X} - \bar{Y})$

$$\hat{\text{SD}}(\bar{X} - \bar{Y}) = \hat{\sigma}_{\text{pooled}} \sqrt{\frac{1}{n} + \frac{1}{m}}$$

$$= \sqrt{\left[ \frac{S_A^2(n - 1) + S_B^2(m - 1)}{n + m - 2} \right] \cdot \left[ \frac{1}{n} + \frac{1}{m} \right]}$$

In the case $n = m$,

$$\hat{\text{SD}}(\bar{X} - \bar{Y}) = \sqrt{\frac{S_A^2 + S_B^2}{n}}$$

Cl for the difference between the means

$$\frac{(\bar{X} - \bar{Y}) - (\mu_A - \mu_B)}{\hat{\text{SD}}(\bar{X} - \bar{Y})} \sim t(df = n + m - 2)$$

The procedure:

1. Calculate $(\bar{X} - \bar{Y})$.
2. Calculate $\hat{\text{SD}}(\bar{X} - \bar{Y})$.
3. Find the 97.5 percentile of the $t$ distr’n with $n + m - 2$ d.f. $\rightarrow t$
4. Calculate the interval: $(\bar{X} - \bar{Y}) \pm t \cdot \hat{\text{SD}}(\bar{X} - \bar{Y})$. 
Example

Treatment A:

\[2.67 \ 2.86 \ 2.87 \ 3.04 \ 3.09 \ 3.09 \ 3.13 \ 3.27 \ 3.35\]
\[n = 9, \bar{x} \approx 3.04, s_A \approx 0.214\]

Treatment B:

\[3.78 \ 3.06 \ 3.64 \ 3.31 \ 3.31 \ 3.51 \ 3.22 \ 3.67\]
\[m = 8, \bar{y} \approx 3.44, s_B \approx 0.250\]

\[\hat{\sigma}_{\text{pooled}} = \sqrt{\frac{s_A^2(n - 1) + s_B^2(m - 1)}{n + m - 2}} = \ldots \approx 0.231\]

\[\text{SD}(\bar{X} - \bar{Y}) = \hat{\sigma}_{\text{pooled}}\sqrt{\frac{1}{n} + \frac{1}{m}} = \ldots \approx 0.112\]

97.5 percentile of t(df=15) \approx 2.13

Example

95% confidence interval:

\[(3.04 - 3.44) \pm 2.13 \cdot 0.112 \approx -0.40 \pm 0.24 = (-0.64, -0.16).\]
Example

Treatment A:  
n = 10  
sample mean: \( \bar{x} = 55.22 \)  
sample SD: \( s_A = 7.64 \)  
t value = \( qt(0.975, 9) = 2.26 \)  

\[ \rightarrow \text{95% CI for } \mu_A: \]
\[ 55.22 \pm 2.26 \times 7.64 / \sqrt{10} = 55.2 \pm 5.5 = (49.8, 60.7) \]

Treatment B:  
n = 16  
sample mean: \( \bar{x} = 68.2 \)  
sample SD: \( s_B = 18.1 \)  
t value = \( qt(0.975, 15) = 2.13 \)  

\[ \rightarrow \text{95% CI for } \mu_B: \]
\[ 68.2 \pm 2.13 \times 18.1 / \sqrt{16} = 68.2 \pm 9.7 = (58.6, 77.9) \]

Example

\[ \hat{\sigma}_{\text{pooled}} = \sqrt{\frac{(7.64)^2 \times (10-1) + (18.1)^2 \times (16-1)}{10+16-2}} = 15.1 \]

\[ \hat{\text{SD}}(\bar{X} - \bar{Y}) = \hat{\sigma}_{\text{pooled}} \times \sqrt{\frac{1}{n} + \frac{1}{m}} = 15.1 \times \sqrt{\frac{1}{10} + \frac{1}{16}} = 6.08 \]

t value: \( qt(0.975, 10+16-2) = 2.06 \)  

\[ \rightarrow \text{95% confidence interval for } \mu_A - \mu_B: \]
\[ (55.2 - 68.2) \pm 2.06 \times 6.08 = -13.0 \pm 12.6 = (-25.6, -0.5) \]
One problem

What if the two populations really have different SDs, $\sigma_A$ and $\sigma_B$?

Suppose that

- $X_1, X_2, \ldots, X_n$ are iid Normal($\mu_A, \sigma_A$),
- $Y_1, Y_2, \ldots, Y_m$ are iid Normal($\mu_B, \sigma_B$).

Then

$$SD(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}}$$

$$\widehat{SD}(\overline{X} - \overline{Y}) = \sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{m}}$$

The problem:

$$\frac{\overline{X} - \overline{Y} - (\mu_A - \mu_B)}{SD(\overline{X} - \overline{Y})}$$

does not follow a t distribution.
An approximation

In the case that $\sigma_A \neq \sigma_B$:

Let $k = \frac{\left(\frac{s_A^2}{n} + \frac{s_B^2}{m}\right)^2}{\frac{(s_A^2/n)^2}{n-1} + \frac{(s_B^2/m)^2}{m-1}}$

Let $t^*$ be the 97.5 percentile of the t distribution with k d.f.

$\rightarrow$ Use $\overline{X} - \overline{Y} \pm t^* \hat{SD}(\overline{X} - \overline{Y})$ as a 95% confidence interval.

Example

$k = \frac{\left[\left(7.64\right)^2/10 + \left(18.1\right)^2/16\right]^2}{\left[\left(7.64\right)^2/10\right]^2/9 + \left[\left(18.1\right)^2/16\right]^2/15} = \frac{(5.84 + 20.6)^2}{9 + 15} = 21.8.$

t value = $qt(0.975, 21.8) = 2.07.$

$\hat{SD}(\overline{X} - \overline{Y}) = \sqrt{\frac{s_A^2}{n} + \frac{s_B^2}{m}} = \sqrt{\frac{(7.64)^2}{10} + \frac{(18.1)^2}{16}} = 5.14.$

$\rightarrow$ 95% CI for $\mu_A - \mu_B$:

$-13.0 \pm 2.07 \times 5.14 = -13.0 \pm 10.7 = (-23.7, -2.4)$
### Degrees of freedom

- **One sample of size** $n$:  
  
  \[
  X_1, X_2, \ldots, X_n \to (\bar{X} - \mu)/(S/\sqrt{n}) \sim t(df = n - 1)
  \]

- **Two samples, of size** $n$ and $m$:  
  
  \[
  X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m \to \frac{(\bar{X} - \bar{Y}) - (\mu_A - \mu_B)}{\hat{\sigma}_{\text{pooled}} \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(df = n + m - 2)
  \]

**What are these “degrees of freedom”?**
Degrees of freedom

The degrees of freedom concern our estimate of the population standard deviation

We use the residuals $(X_1 - \overline{X}), \ldots, (X_n - \overline{X})$ to estimate $\sigma$.

But we really only have $n - 1$ independent data points (“degrees of freedom”), since $\sum (X_i - \overline{X}) = 0$.

In the two-sample case, we use $(X_1 - \overline{X}), (X_2 - \overline{X}), \ldots, (X_n - \overline{X})$ and $(Y_1 - \overline{Y}), \ldots, (Y_m - \overline{Y})$ to estimate $\sigma$.

But $\sum (X_i - \overline{X}) = 0$ and $\sum (Y_i - \overline{Y}) = 0$, and so we really have just $n + m - 2$ independent data points.

Testing the difference between two means

Treatment A: $X_1, \ldots, X_n \sim \text{iid Normal}(\mu_A, \sigma_A)$

Treatment B: $Y_1, \ldots, Y_m \sim \text{iid Normal}(\mu_B, \sigma_B)$

Test $H_0 : \mu_A = \mu_B$ vs $H_a : \mu_A \neq \mu_B$

Test statistic: $T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{m}}}$

Reject $H_0$ if $|T| > t_{\alpha/2}$

If $H_0$ is true, then $T$ follows (approximately) a t distr’n with $k$ d.f.

$k$ according to the nasty formula shown previously
Example

Treatment A: \( n = 12 \), sample mean = 103.7, sample SD = 7.2  
Treatment B: \( n = 9 \), sample mean = 97.0, sample SD = 4.5  

\[
\hat{SD}(\overline{X} - \overline{Y}) = \sqrt{\frac{7.2^2}{12} + \frac{4.5^2}{9}} = 1.80
\]

\[T = \frac{103.7 - 97.0}{1.80} = 2.60.\]

\( k = \ldots = 18.48 \), so \( C = 2.10 \). Thus we reject \( H_0 \) at \( \alpha = 0.05 \).

Always give a confidence interval!

\[P = 0.019\]

95% CI: (–34.9, −1.2)

\[P = 0.019\]

95% CI: (–13.6, −0.5)

→ Make a statistician happy: draw a picture of the data.
What to say

When rejecting $H_0$:

- The difference is statistically significant.
- The observed difference can not reasonably be explained by chance variation.

When failing to reject $H_0$:

- There is insufficient evidence to conclude that $\mu_A \neq \mu_B$.
- The difference is not statistically significant.
- The observed difference could reasonably be the result of chance variation.
What about a different significance level?

Recall $T = 2.60 \quad k = 18.48$

If $\alpha = 0.10$, $C = 1.73 \implies$ Reject $H_0$
If $\alpha = 0.05$, $C = 2.10 \implies$ Reject $H_0$
If $\alpha = 0.01$, $C = 2.87 \implies$ Fail to reject $H_0$
If $\alpha = 0.001$, $C = 3.90 \implies$ Fail to reject $H_0$

P-value: the smallest $\alpha$ for which you would still reject $H_0$ with the observed data.

With these data, $P = 2 \times (1 - pt(2.60, 18.48)) = 0.018$.

Another example

Suppose I measure the blood pressure of 6 subjects on a low salt diet and 6 subjects on a high salt diet. We wish to prove that the high salt diet causes an increase in blood pressure.

We imagine $X_1, \ldots, X_n \sim iid \text{ Normal}(\mu_L, \sigma_L)$ low salt
$Y_1, \ldots, Y_m \sim iid \text{ Normal}(\mu_H, \sigma_H)$ high salt

We want to test $H_0: \mu_L = \mu_H$ versus $H_a: \mu_L < \mu_H$

→ Are the data compatible with $H_0$?
A one-tailed test

Test statistic: \[ T = \frac{\bar{X} - \bar{Y}}{SD(X - Y)} \]

Since we seek to prove that \( \mu_L \) is smaller than \( \mu_H \), only large negative values of the statistic are interesting.

Thus, our rejection region is \( T < C \) for some critical value \( C \).

We choose \( C \) so that \( Pr( T < C \mid \mu_L = \mu_H ) = \alpha \).

The example

Low salt: \( n = 6 \); sample mean = 51.0, sample SD = 10.0
High salt: \( n = 6 \); sample mean = 69.1, sample SD = 15.1

\[ \bar{x} - \bar{y} = -18.1 \quad SD(X - Y) = 7.40 \quad T = -18.1 / 7.40 = -2.44 \]

\( k = 8.69 \). If \( \alpha = 0.05 \), then \( C = -1.84 \).

Since \( T < C \), we reject \( H_0 \) and conclude that \( \mu_L < \mu_H \).

Note: P-value = pt \((-2.44, 8.69) = 0.019\).
Suppose I do some pre/post measurements.

I make some measurement on each of 5 subjects before and after some treatment.

Question: Does the treatment have any effect?

<table>
<thead>
<tr>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>18.6</td>
<td>14.3</td>
<td>21.4</td>
<td>19.3</td>
<td>24.0</td>
</tr>
<tr>
<td>After</td>
<td>17.8</td>
<td>24.1</td>
<td>31.9</td>
<td>28.6</td>
<td>40.0</td>
</tr>
</tbody>
</table>

In this sort of pre/post measurement example, study the differences as a single sample.

Why? The pre/post measurements are likely associated, and as a result one can more precisely learn about the effect of the treatment.

<table>
<thead>
<tr>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>18.6</td>
<td>14.3</td>
<td>21.4</td>
<td>19.3</td>
<td>24.0</td>
</tr>
<tr>
<td>After</td>
<td>17.8</td>
<td>24.1</td>
<td>31.9</td>
<td>28.6</td>
<td>40.0</td>
</tr>
<tr>
<td>Difference</td>
<td>-0.8</td>
<td>9.8</td>
<td>10.5</td>
<td>9.3</td>
<td>16.0</td>
</tr>
</tbody>
</table>

n = 5; mean difference = 8.96; SD difference = 6.08.

95% CI for underlying mean difference = \ldots = (1.4, 16.5)

P-value for test of \( \mu_{\text{before}} = \mu_{\text{after}} \): 0.03.
Summary

• Tests of hypotheses → answering yes/no questions regarding population parameters.

• There are two kinds of errors:
  ◦ Type I: Reject $H_0$ when it is true.
  ◦ Type II: Fail to reject $H_0$ when it is false.

• If we fail to reject $H_0$, we do not “accept $H_0$”.

• P-value: the probability, if $H_0$ is true, of obtaining data as extreme as was observed. $Pr(\text{data | no effect})$ rather than $Pr(\text{no effect | data})$.

• P-values are a function of the data (and thus, they are random).

• Power: the probability of rejecting $H_0$ when it is false.

• Always look at the confidence interval as well as the P-value.

Example

\[
\begin{align*}
\bar{X} &= 47.5 & s_A &= 10.5 & n &= 6 \\
\bar{Y} &= 74.3 & s_B &= 20.6 & m &= 9 \\
\bar{s}_p &= 17.4 & T &= -2.93 \\
\rightarrow \quad P &= 2 \times pt(-2.93, 6 + 9 - 2) = 0.011.
\end{align*}
\]
Wilcoxon rank-sum test

Rank the X’s and Y’s from smallest to largest (1, 2, . . . , n+m)

\[ R = \text{sum of ranks for X’s} \]

(Also known as the Mann-Whitney Test)

\[
\begin{array}{ccc}
\text{X} & \text{Y} & \text{rank} \\
35.0 & 1 & \\
38.2 & 2 & \\
43.3 & 3 & \\
46.8 & 4 & \\
49.7 & 5 & \\
50.0 & 6 & \\
51.9 & 7 & \\
57.1 & 8 & \\
61.2 & 9 & \\
74.1 & 10 & \\
75.1 & 11 & \\
84.5 & 12 & \\
90.0 & 13 & \\
95.1 & 14 & \\
101.5 & 15 & \\
\end{array}
\]

\[ R = 1 + 2 + 3 + 6 + 8 + 9 = 29 \]

P-value = 0.026

Note: The distribution of R (given that X’s and Y’s have the same dist’n) is calculated numerically

Permutation test

Compare the observed t-statistic to the distribution obtained by randomly shuffling the group status of the measurements.
**Permutation distribution**

![Permutation distribution graph]

P-value = \( \text{Pr}(|T^*| \geq |T_{\text{obs}}|) \)

\[\rightarrow\] Small \( n \) & \( m \): Look at all \( \binom{n+m}{n} \) possible shuffles

\[\rightarrow\] Large \( n \) & \( m \): Look at a sample (w/ repl) of 1000 such shuffles

Example data:
All 5005 permutations: \( P = 0.015 \); sample of 1000: \( P = 0.013 \).

**Estimating the permutation P-value**

Let \( P \) be the true P-value (if we do all possible shuffles).

Do \( N \) shuffles, and let \( X \) be the number of times the statistic after shuffling is bigger or equal to the observed statistic.

\[\rightarrow\] \( \hat{P} = \frac{X}{N} \) where \( X \sim \text{Binomial}(N, P) \)

\[\rightarrow\] \( E(\hat{P}) = P \quad \text{SD}(\hat{P}) = \sqrt{\frac{P(1-P)}{N}} \)

If the “true” P-value was \( P = 5\% \), and we do \( N=1000 \) shuffles: \( \text{SD}(\hat{P}) = 0.7\% \).
Summary

The t-test relies on a normality assumption. If this is a worry, consider:

- **Paired data:**
  - Sign test
  - Signed rank test
  - Permutation test

- **Unpaired data:**
  - Rank-sum test
  - Permutation test

→ The crucial assumption is independence!

The fact that the permutation distribution of the t-statistic is often closely approximated by a t distribution is good support for just doing t-tests.

Maximum Likelihood Estimation
Estimation

Goal: Estimate a population parameter $\theta$.

Data: $X_1, X_2, \ldots, X_n \sim$ iid with distribution depending on $\theta$.

If one has many estimators to choose from, pick

- That with the smallest SE, among all unbiased estimators
- That with the smallest RMS error, even if biased

$\rightarrow$ Sometimes it is not clear how to form even one good estimator.

Maximum likelihood estimation

Likelihood function: $L(\theta) = Pr(\text{data} \mid \theta)$

Log likelihood: $l(\theta) = \log Pr(\text{data} \mid \theta)$

Maximum likelihood estimate:

Choose, as the estimate of $\theta$, the value of $\theta$ for which the likelihood function $L(\theta)$ (or equivalently, the log likelihood function) is maximized.

$\rightarrow$ You need to solve these equations analytically or numerically.
**Example 1**

Suppose \( X \sim \text{Binomial}(n, p) \).

log likelihood function:  
\[
    l(p) = \log \left\{ \binom{n}{x} p^x (1 - p)^{n-x} \right\} \\
    = x \log(p) + (n - x) \log(1 - p) + \text{constant}
\]

MLE: the obvious thing:  
\[
    \hat{p} = \frac{x}{n}
\]

![Graph of log likelihood function for Example 1](image-url)

**Example 2**

Suppose \( X_1, \ldots, X_{20} \sim \text{iid Poisson}(\lambda) \).

log likelihood function:  
\[
    l(\lambda) = \log \left\{ \prod_i e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right\} \\
    = \ldots = -20 \lambda + (\sum x_i) \log\lambda + \text{constant}
\]

MLE: the obvious thing:  
\[
    \hat{\lambda} = \bar{x}
\]

![Graph of log likelihood function for Example 2](image-url)
Example 3

Suppose $X_1, \ldots, X_n \sim \text{iid } N(\mu, \sigma)$

log likelihood function: $l(\mu, \sigma) = \log \left\{ \prod_i \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x_i - \mu}{\sigma} \right)^2 \right] \right\}$

MLEs: almost the obvious things:

$\hat{\mu} = \bar{x} \quad \hat{\sigma} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$

Example 3: the log likelihood surface
About MLEs

Maximum likelihood estimation is a general procedure for finding a reasonable estimator

- In many cases, the MLE turns out to be the obvious thing.
- MLEs are often very good (but not necessarily the best) possible estimators:
  - unbiased or nearly unbiased
  - small standard errors
- Sometimes obtaining the MLEs requires hefty computation!

Example 4: ABO blood groups

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Genotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>OO</td>
<td>( p_O^2 )</td>
</tr>
<tr>
<td>A</td>
<td>AA or AO</td>
<td>( p_A^2 + 2p_Ap_O )</td>
</tr>
<tr>
<td>B</td>
<td>BB or BO</td>
<td>( p_B^2 + 2p_Bp_O )</td>
</tr>
<tr>
<td>AB</td>
<td>AB</td>
<td>( 2p_Ap_B )</td>
</tr>
</tbody>
</table>

Frequencies under the assumption of Hardy-Weinberg equilibrium.
Example 4: Data

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>No. subjects</th>
<th>% subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>117</td>
<td>46.8%</td>
</tr>
<tr>
<td>A</td>
<td>98</td>
<td>39.2%</td>
</tr>
<tr>
<td>B</td>
<td>29</td>
<td>11.6%</td>
</tr>
<tr>
<td>AB</td>
<td>6</td>
<td>2.4%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>250</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

What are the estimates of $p_A$, $p_B$, $p_O$?

Example 4: Estimates

Simple estimates:

$\tilde{p}_O = \sqrt{0.468} = 0.684$

$\tilde{p}_A^2 + 2\tilde{p}_A 0.684 = 0.392 \quad \rightarrow \quad \tilde{p}_A = 0.243$

$\tilde{p}_B = 0.024 / (2\tilde{p}_A) = 0.072$

Log likelihood:

$I(p_O, p_A, p_B) =$

$117 \log(p_O^2) + 98 \log(p_A^2 + 2p_Ap_O) + 29 \log(p_B^2 + 2p_Bp_O) + 6 \log(2p_Ap_B)$
Example 5: log likelihood

\[ p_B = 0.073 \]
\[ p_A = 0.237 \]
\[ p_D = 0.690 \]