Review

Population distribution

\[ Z = \frac{(\bar{X} - \mu)}{(\sigma / \sqrt{n})} \]

Distribution of \( \bar{X} \)

\[ t = \frac{(\bar{X} - \mu)}{(s / \sqrt{n})} \]

\( X_1, X_2, \ldots, X_n \) independent normal(\( \mu, \sigma \)).

95% confidence interval for \( \mu \):

\[ \bar{X} \pm t \times \frac{s}{\sqrt{n}} \]

where \( t = 97.5 \) percentile of \( t \) distribution with \((n-1)\) d.f.

Example

Suppose we have weighed the mass of tumor in 20 mice, and obtained the following numbers

<table>
<thead>
<tr>
<th>Data</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>34.9</td>
<td>28.5</td>
<td>34.3</td>
<td>38.4</td>
<td>29.6</td>
</tr>
<tr>
<td>28.2</td>
<td>25.3</td>
<td>. . . . . .</td>
<td>. . . . . .</td>
<td>32.1</td>
</tr>
</tbody>
</table>

\( \bar{x} = 30.7 \quad n = 20 \) \quad s = 6.06 \quad qt(0.975,19) = 2.09

95% confidence interval for \( \mu \) (the population mean):

\[ 30.7 \pm 2.09 \times 6.06 / \sqrt{20} \approx 30.7 \pm 2.84 = (27.9, 33.5) \]
What is a confidence interval?

A confidence interval is the result of a procedure that 95% of the time produces an interval containing the population parameter.

In advance, there is a 95% chance that the confidence interval that you obtain will contain the parameter of interest.

After the fact, your particular 95% CI either contains the parameter or it doesn’t; we’re not allowed to talk about chance anymore.

What’s the deal?

Why this wacky confidence interval business?

We can talk about Pr(data | µ).

But we can’t talk about Pr(µ | data).

Actually, a portion of modern (and even rather non-modern) statistics (called Bayesian statistics—remember Bayes’s rule?) concerns inferential statements like Pr(µ | data).

But this is beyond the scope of the current course.
Differences between means

Suppose I measure the treatment response on 10 mice from strain A and 10 mice from strain B.

How different are the responses of the two strains?

Again, I’m not interested in these particular mice, but in the strains generally.

Suppose $X_1, X_2, \ldots, X_n$ are indep. normal(mean=$\mu_A$, SD=$\sigma$) and $Y_1, Y_2, \ldots, Y_m$ are indep. normal(mean=$\mu_B$, SD=$\sigma$)

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y})$$

$$= \mu_A - \mu_B$$

$$SD(\bar{X} - \bar{Y}) = \sqrt{SD(\bar{X})^2 + SD(\bar{Y})^2}$$

$$= \sqrt{\left( \frac{\sigma}{\sqrt{n}} \right)^2 + \left( \frac{\sigma}{\sqrt{m}} \right)^2} = \sigma \sqrt{\frac{1}{n} + \frac{1}{m}}$$

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

\[ \bar{X} - \bar{Y} \]
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 $n = m$, $\hat{\sigma}_{pooled} = \sqrt{\frac{2}{n + 2}}$.

### Est’d SE of $(\bar{X} - \bar{Y})$

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

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

In the case $n = m$,

$$\hat{SD}(\bar{X} - \bar{Y}) = \sqrt{\frac{s_A^2 + s_B^2}{n}}$$
CI for difference between means

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

The procedure:

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

Example

Strain A:

2.67 2.86 2.87 3.04 3.09 3.09 3.13 3.27 3.35

\(n = 9, \bar{X} \approx 3.04, s_A \approx 0.214\)

Strain B:

3.78 3.06 3.64 3.31 3.31 3.51 3.22 3.67

\(m = 8, \bar{Y} \approx 3.44, s_B \approx 0.250\)

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

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

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

95% confidence interval:

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

Strain A:

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

95% CI for \(\mu_A\):

\[55.22 \pm 2.26 \times 7.64 / \sqrt{10}\]
\[= 55.2 \pm 5.5 = (49.8, 60.7)\]

Strain B:

n = 16
sample mean: \(\bar{X} = 68.2\)
sample SD: \(s_A = 18.1\)
t value = \(qt(0.975, 15) = 2.13\)

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**

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

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

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

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 $\sigma_B$?

If $X_1, X_2, \ldots, X_n$ are iid normal($\mu_A, \sigma_A$) and $Y_1, Y_2, \ldots, Y_m$ are iid normal($\mu_B, \sigma_B$),

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

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

The problem:

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

does not follow a t distribution.
In the case that $\sigma_A \neq \sigma_B$:

Let $k = \left( \frac{s^2_n + s^2_m}{n} \right)^2 \leq \frac{(s^2_A/n)^2/n - 1}{m - 1} + \frac{(s^2_B/m)^2}{m - 1}$

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

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

Example

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

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

$\hat{SD}(\bar{X} - \bar{Y}) = \sqrt{\frac{s^2_n}{n} + \frac{s^2_m}{m}} = \sqrt{\frac{(7.64)^2/10 + (18.1)^2/16}{10} + \frac{(18.1)^2/16}{15}} = 5.14.$

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 $n$:

$$X_1, X_2, \ldots, X_n \rightarrow \frac{(\bar{X} - \mu)}{(s/\sqrt{n})} \sim t(df = n - 1)$$

Two samples, of size $n$ and $m$:

$$X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m \rightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_A - \mu_B)}{\hat{\sigma}_{\text{pooled}} \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(df = n + m - 2)$$

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

The **degrees of freedom** concern our estimate of the population SD.

We use the residuals \((X_1 - \bar{X}), (X_2 - \bar{X}), \ldots, (X_n - \bar{X})\) to estimate \(\sigma\).

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

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

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

Confidence interval for population SD

Suppose we observe \(X_1, X_2, \ldots, X_n\) iid normal(\(\mu, \sigma\)).

Suppose we wish to create a 95% CI for the population SD, \(\sigma\).

Our estimate of \(\sigma\) is, of course, the sample SD, \(s\).

The sampling distribution of \(s\) is such that

\[
\frac{(n - 1)s^2}{\sigma^2} \sim \chi^2(\text{df} = n - 1)
\]

![Chi-squared distribution](image)
Choose $L$ and $U$ such that

\[ \Pr \left( L \leq \frac{(n-1)s^2}{\sigma^2} \leq U \right) = 95\%. \]

\[ \Rightarrow \Pr \left( \frac{1}{U} \leq \frac{\sigma^2}{(n-1)s^2} \leq \frac{1}{L} \right) = 95\% \]

\[ \Rightarrow \Pr \left( \frac{(n-1)s^2}{U} \leq \sigma^2 \leq \frac{(n-1)s^2}{L} \right) = 95\% \]

\[ \Rightarrow \Pr \left( \sqrt{\frac{n-1}{U}} \leq \sigma \leq \sqrt{\frac{n-1}{L}} \right) = 95\% \]

\[ \Rightarrow \left( \sqrt{\frac{n-1}{U}}, \sqrt{\frac{n-1}{L}} \right) \text{ is a 95\% CI for } \sigma. \]

**Example**

**Strain A:**

\[ n = 10 \quad \text{sample SD: } s_A = 7.64 \]

\[ L = \text{qchisq}(0.025, 9) = 2.70 \]

\[ U = \text{qchisq}(0.975, 9) = 19.0 \]

95\% CI for $\sigma_A$: \( (7.64 \times \sqrt{\frac{9}{19.0}}, 7.64 \times \sqrt{\frac{9}{2.70}}) \)

\[ = (7.64 \times 0.688, 7.64 \times 1.83) \]

\[ = (5.3, 14.0) \]

**Strain B:**

\[ n = 16 \quad \text{sample SD: } s_B = 18.1 \]

\[ L = \text{qchisq}(0.025, 15) = 6.25 \]

\[ U = \text{qchisq}(0.975, 15) = 27.5 \]

95\% CI for $\sigma_B$: \( (18.1 \times \sqrt{\frac{15}{27.5}}, 18.1 \times \sqrt{\frac{15}{6.25}}) \)

\[ = (18.1 \times 0.739, 18.1 \times 1.55) \]

\[ = (13.4, 28.1) \]