Confidence Intervals

Review

$\longrightarrow$ If $X_1, \ldots, X_n$ have mean μ and SD σ, then

E($\overline{X}$) = μ no matter what

SD($\overline{X}$) = $\sigma/\sqrt{n}$ if the X's are independent

$\longrightarrow$ If $X_1, \ldots, X_n$ are iid Normal(mean=μ, SD=σ), then

$$\overline{X} \sim \text{Normal}(\text{mean} = \mu, \text{SD} = \sigma/\sqrt{n}).$$

$\longrightarrow$ If $X_1, \ldots, X_n$ are iid with mean μ and SD σ and the sample size n is large, then

$$\overline{X} \sim \text{Normal}(\text{mean} = \mu, \text{SD} = \sigma/\sqrt{n}).$$

Confidence intervals

Suppose we measure the $\log_{10}$ cytokine response in 100 male mice of a certain strain, and find that the sample average ($\bar{x}$) is 3.52 and sample SD (s) is 1.61.

Our estimate of the SE of the sample mean is $1.61/\sqrt{100} = 0.161$.

A 95% confidence interval for the population mean (μ) is roughly

$$3.52 \pm (2 \times 0.16) = 3.52 \pm 0.32 = (3.20, 3.84).$$

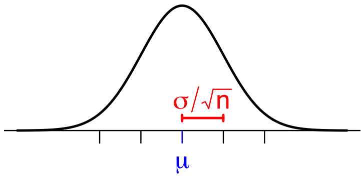
What does this mean?

Confidence intervals

Suppose that $X_1, \dots, X_n$ are iid Normal(mean=μ, SD=σ).
Suppose that we actually know σ.

Then $\bar{X} \sim$ Normal(mean=μ, SD=$\sigma/\sqrt{n}$) σ is known but μ is not!

$\longrightarrow$ How close is $\overline{X}$ to μ?



$$\Pr\left(-1.96 \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right) = 95\%$$

$$\Pr\left(\frac{-1.96\,\sigma}{\sqrt{n}} \le \overline{X} - \mu \le \frac{1.96\,\sigma}{\sqrt{n}}\right) = 95\%$$

$$\Pr\left(\overline{X} - \frac{1.96\,\sigma}{\sqrt{n}} \le \mu \le \overline{X} + \frac{1.96\,\sigma}{\sqrt{n}}\right) = 95\%$$

What is a confidence interval?

A 95% confidence interval is an interval calculated from the data that in advance has a 95% chance of covering the population parameter.

In advance, $\overline{X} \pm 1.96\sigma/\sqrt{n}$ has a 95% chance of covering μ.

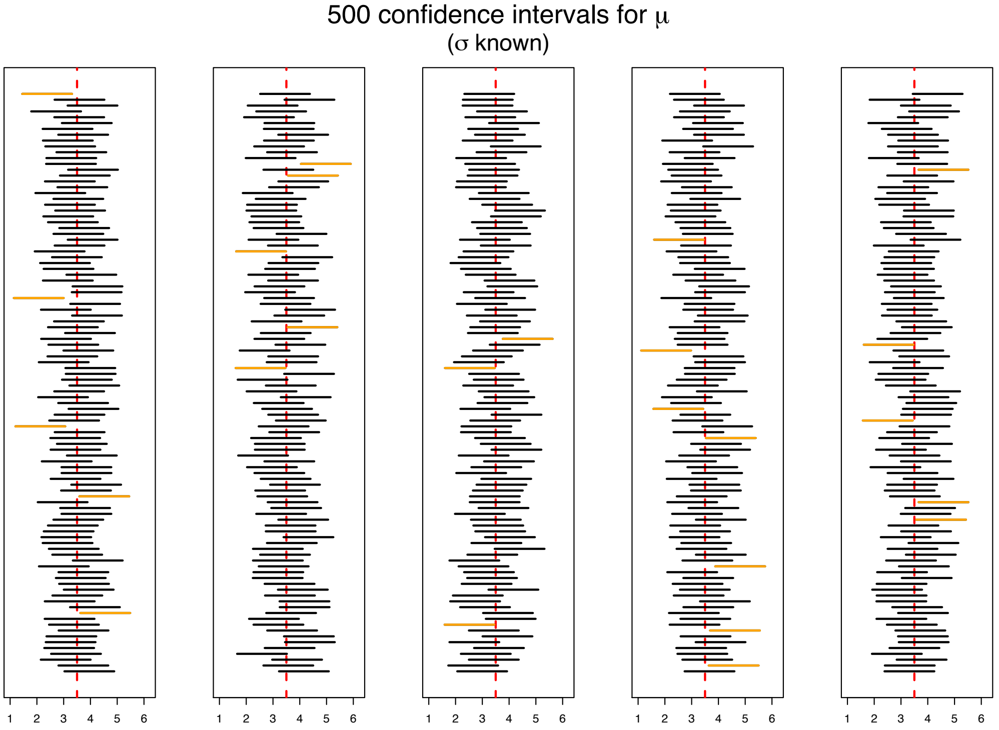
Thus, it is called a 95% confidence interval for μ.

Note that, after the data is gathered (for instance, n=100, $\bar{x} = 3.52$, $\sigma = 1.61$), the interval becomes fixed:

$$\bar{x} \pm 1.96\sigma/\sqrt{n} = 3.52 \pm 0.32.$$

We can't say that there's a 95% chance that μ is in the interval 3.52 ± 0.32. It either is or it isn't; we just don't know.

What is a confidence interval?



Longer and shorter intervals

⟶ If we use 1.64 in place of 1.96, we get shorter intervals with lower confidence.

Since $\Pr\left(\dfrac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \leq 1.64\right) = 90\%$,

$\overline{X} \pm 1.64\sigma/\sqrt{n}$ is a 90% confidence interval for μ.

⟶ If we use 2.58 in place of 1.96, we get longer intervals with higher confidence.

Since $\Pr\left(\dfrac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \leq 2.58\right) = 99\%$,

$\overline{X} \pm 2.58\sigma/\sqrt{n}$ is a 99% confidence interval for μ.

What is a confidence interval? (cont)

A 95% confidence interval is obtained from a procedure for producing an interval, based on data, that 95% of the time will produce an interval covering the population parameter.

In advance, there's a 95% chance that the interval will cover the population parameter.

After the data has been collected, the confidence interval either contains the parameter or it doesn't.

Thus we talk about confidence rather than probability.

But we don’t know the SD

Use of $\overline{X} \pm 1.96\,\sigma/\sqrt{n}$ as a 95% confidence interval for μ requires knowledge of σ.

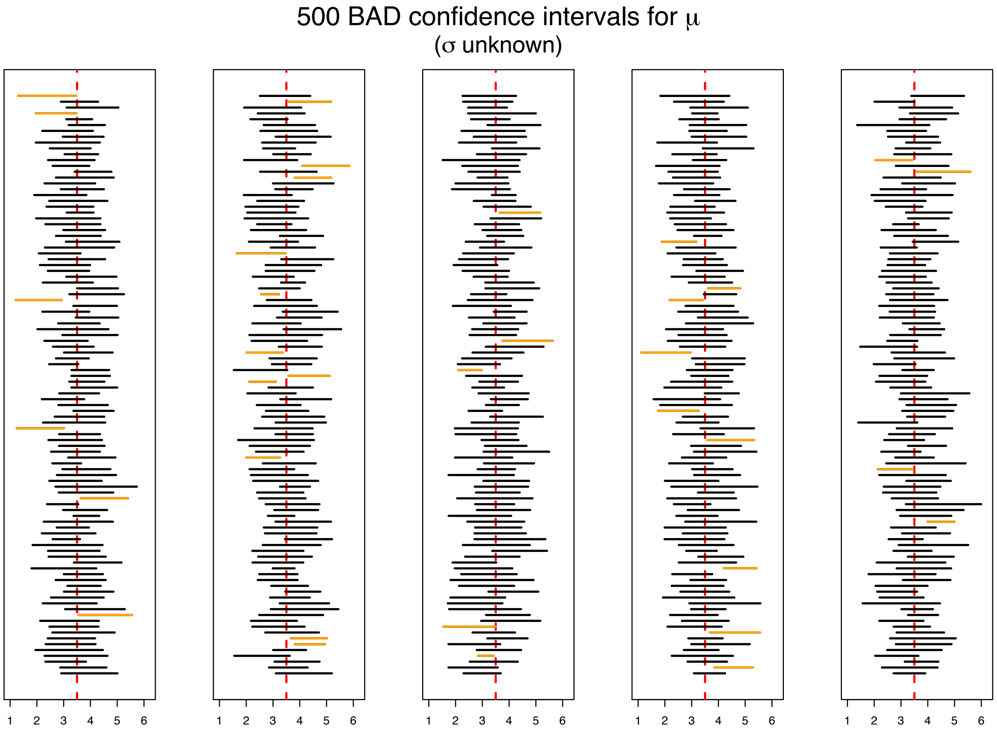
That the above is a 95% confidence interval for μ is a result of the following:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal(0,1)}$$

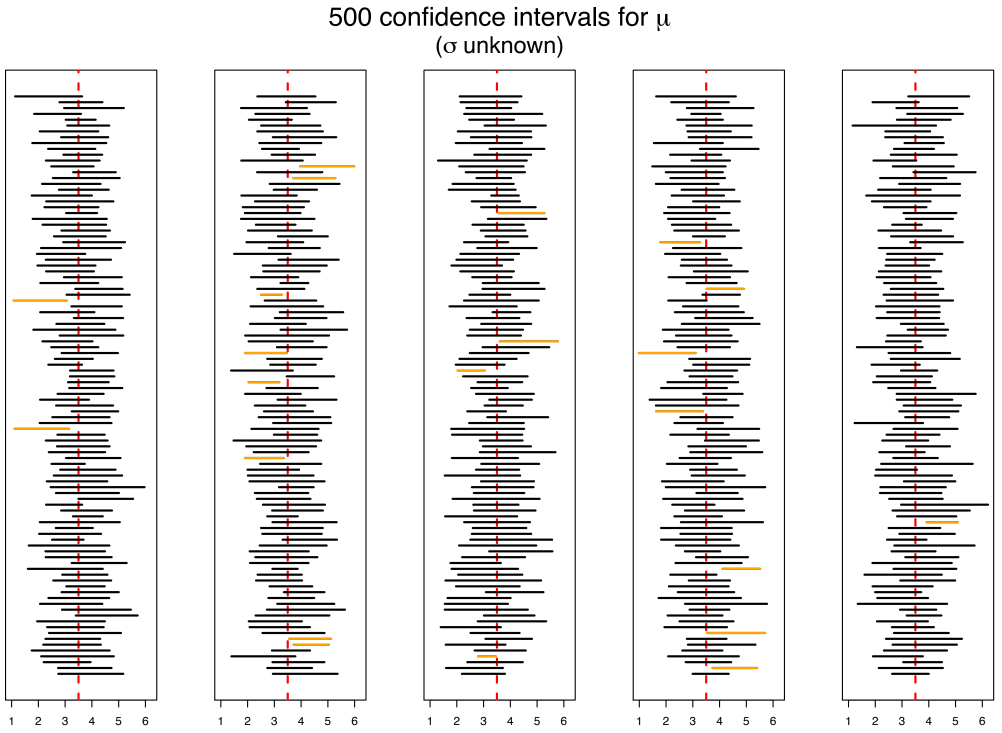
What if we don’t know σ?

$\longrightarrow$ We plug in the sample SD S, but then we need to widen the intervals to account for the uncertainty in S.

What is a confidence interval? (cont)



What is a confidence interval? (cont)



The Student t distribution

If $X_1, X_2, \ldots X_n$ are iid Normal(mean=μ, SD=σ), then

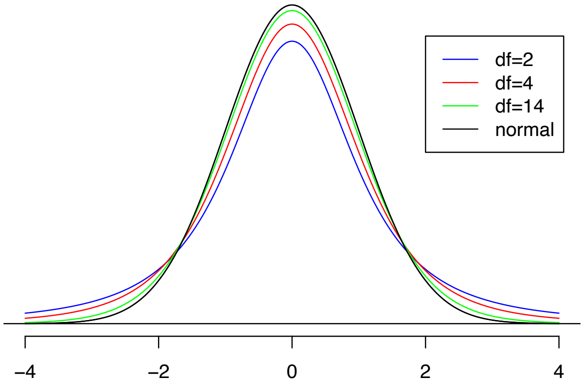
$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(df = n - 1)$$

Discovered by William Gossett (“Student”) who worked for Guinness.

In R, use the functions `pt()`, `qt()`, and `dt()`.

⟶ `qt(0.975,9)` returns `2.26` (compare to 1.96)

⟶ `pt(1.96,9)-pt(-1.96,9)` returns `0.918` (compare to 0.95)



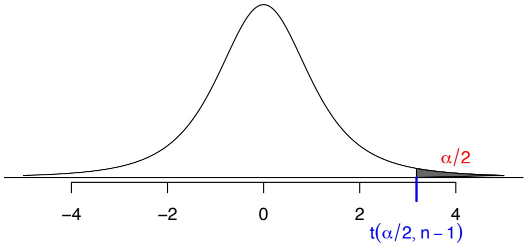
The t interval

If $X_1, \ldots, X_n$ are iid Normal(mean=μ, SD=σ), then

$$\overline{X} \pm t(\alpha/2, n-1)\ S/\sqrt{n}$$

is a $1-\alpha$ confidence interval for μ.

⟶ $t(\alpha/2, n-1)$ is the $1-\alpha/2$ quantile of the t distribution with $n-1$ "degrees of freedom."



⟶ In R: `qt(0.975,9)` for the case n=10, α=5%.

Example 1

Suppose we have measured the $\log_{10}$ cytokine response of 10 mice, and obtained the following numbers:

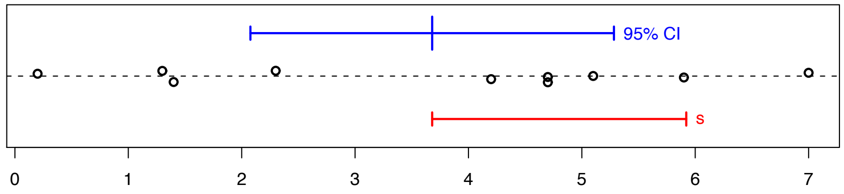
Data

0.2 1.3 1.4 2.3 4.2
4.7 4.7 5.1 5.9 7.0

$\bar{x} = 3.68$ n = 10
s = 2.24 `qt(0.975,9)` = 2.26

⟶ 95% confidence interval for μ (the population mean):

$$3.68 \pm 2.26 \times 2.24\ /\ \sqrt{10} \approx 3.68 \pm 1.60 = (2.1, 5.3)$$



Example 2

Suppose we have measured (by RT-PCR) the $\log_{10}$ expression of a gene in 3 tissue samples, and obtained the following numbers:

Data

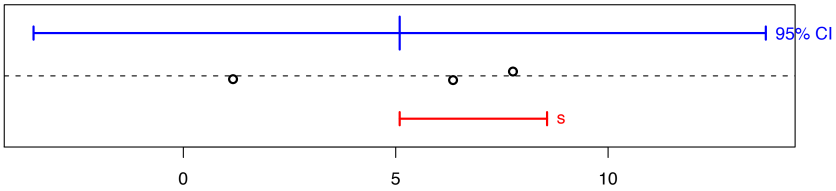
1.17 6.35 7.76

$\bar{x}$ = 5.09 n = 3

s = 3.47 `qt(0.975,2)` = 4.30

⟶ 95% confidence interval for μ (the population mean):

$$5.09 \pm 4.30 \times 3.47 \,/\, \sqrt{3} \approx 5.09 \pm 8.62 = (-3.5, 13.7)$$



Example 3

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

Data

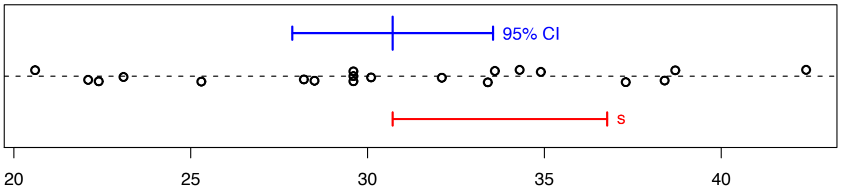
34.9 28.5 34.3 38.4 29.6
28.2 25.3 . . . . . . 32.1

$\bar{x}$ = 30.7 n = 20

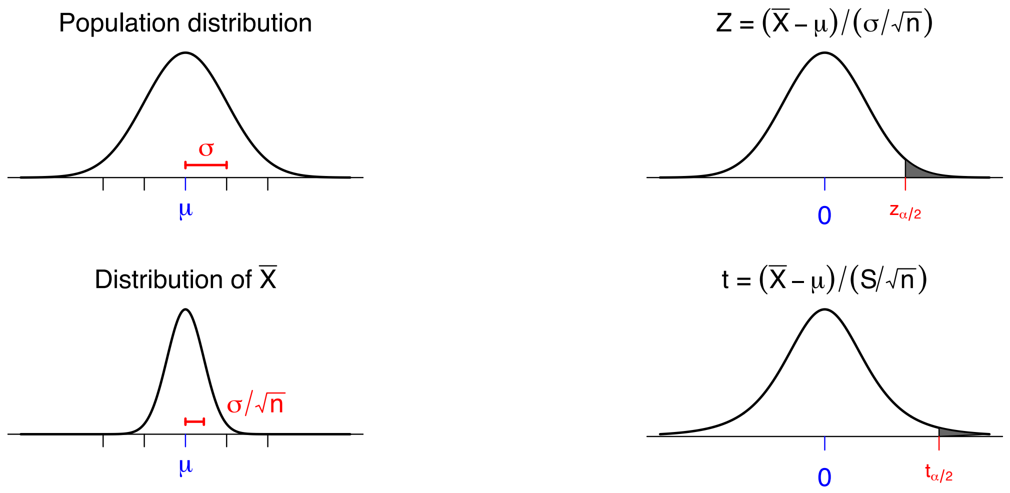
s = 6.06 `qt(0.975,19)` = 2.09

⟶ 95% confidence interval for μ (the population mean):

$$30.7 \pm 2.09 \times 6.06 \,/\, \sqrt{20} \approx 30.7 \pm 2.84 = (27.9, 33.5)$$



Confidence interval for the mean



$X_1, X_2, \ldots, X_n$ independent Normal(μ, σ).

95% confidence interval for μ:

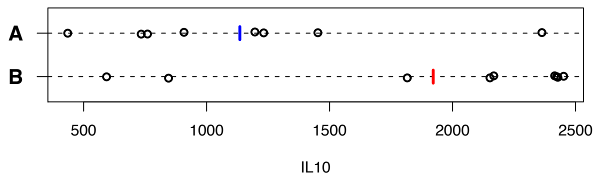
$\overline{X} \pm t\, S/\sqrt{n}$ where t = 97.5 percentile of t distribution with (n – 1) d.f.

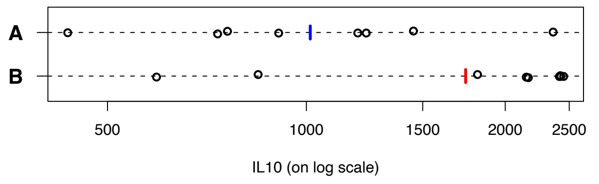
Differences between means

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?

$\longrightarrow$ I am not interested in these *particular* mice, but in the strains *generally*.





$\overline{X} - \overline{Y}$

Suppose that

- $\circ$ $X_1, X_2, \ldots, X_\text{n}$ are iid Normal(mean=μ_A, SD=σ), and
- $\circ$ $Y_1, Y_2, \ldots, Y_\text{m}$ are iid Normal(mean=μ_B, SD=σ).

Then

$\longrightarrow$ $\text{E}(\overline{X} - \overline{Y}) = \text{E}(\overline{X}) - \text{E}(\overline{Y}) = \mu_\text{A} - \mu_\text{B}$

$\longrightarrow$ $\text{SD}(\overline{X} - \overline{Y}) = \sqrt{\text{SD}(\overline{X})^2 + \text{SD}(\overline{Y})^2} =$

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

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

Pooled estimate of the population SD

We have two different estimates of the populations' SD, σ:

$$\hat{\sigma}_\text{A} = S_\text{A} = \sqrt{\frac{\sum(X_\text{i} - \overline{X})^2}{\text{n} - 1}} \qquad \hat{\sigma}_\text{B} = S_\text{B} = \sqrt{\frac{\sum(Y_\text{i} - \overline{Y})^2}{\text{m} - 1}}$$

We can use all of the data together to obtain an improved estimate of σ, which we call the "pooled" estimate.

$$\begin{aligned} \hat{\sigma}_\text{pooled} &= \sqrt{\frac{\sum(X_\text{i} - \overline{X})^2 + \sum(Y_\text{i} - \overline{Y})^2}{\text{n} + \text{m} - 2}} \\ &= \sqrt{\frac{S_\text{A}^2(\text{n} - 1) + S_\text{B}^2(\text{m} - 1)}{\text{n} + \text{m} - 2}} \end{aligned}$$

Note: If n = m, then $\hat{\sigma}_\text{pooled} = \sqrt{\left(S_\text{A}^2 + S_\text{B}^2\right)/2}$

Estimated SE of $(\overline{X} - \overline{Y})$

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

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

In the case n = m,

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

CI for the difference between the means

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

The procedure:

1. Calculate $(\overline{X} - \overline{Y})$.
2. Calculate $\widehat{\text{SD}}(\overline{X} - \overline{Y})$.
3. Find the 97.5 percentile of the t distr'n with n + m – 2 d.f.
   $\longrightarrow$ t
4. Calculate the interval: $(\overline{X} - \overline{Y}) \pm t \cdot \widehat{\text{SD}}(\overline{X} - \overline{Y})$.

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, $\bar{y} \approx 3.44$, $s_B \approx 0.250$

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

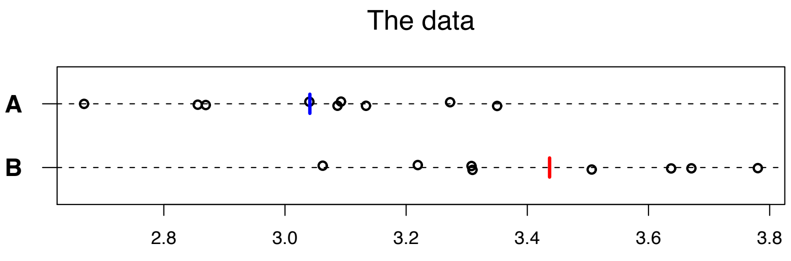
$$\widehat{\text{SD}}(\overline{X} - \overline{Y}) = \hat{\sigma}_{\text{pooled}}\sqrt{\frac{1}{n} + \frac{1}{m}} = \ldots \approx 0.112$$

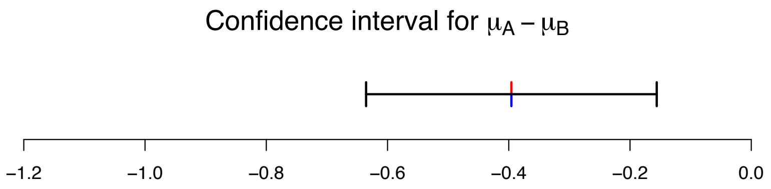
97.5 percentile of t(df=15) ≈ 2.13

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

Strain A:

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

$\longrightarrow$ 95% CI for μ_A:

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

Strain B:

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

$\longrightarrow$ 95% CI for μ_B:

$$68.2 \pm 2.13 \times 18.1 \,/\, \sqrt{16} \quad = \quad 68.2 \pm 9.7 \quad = \quad (58.6, 77.9)$$

Example

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

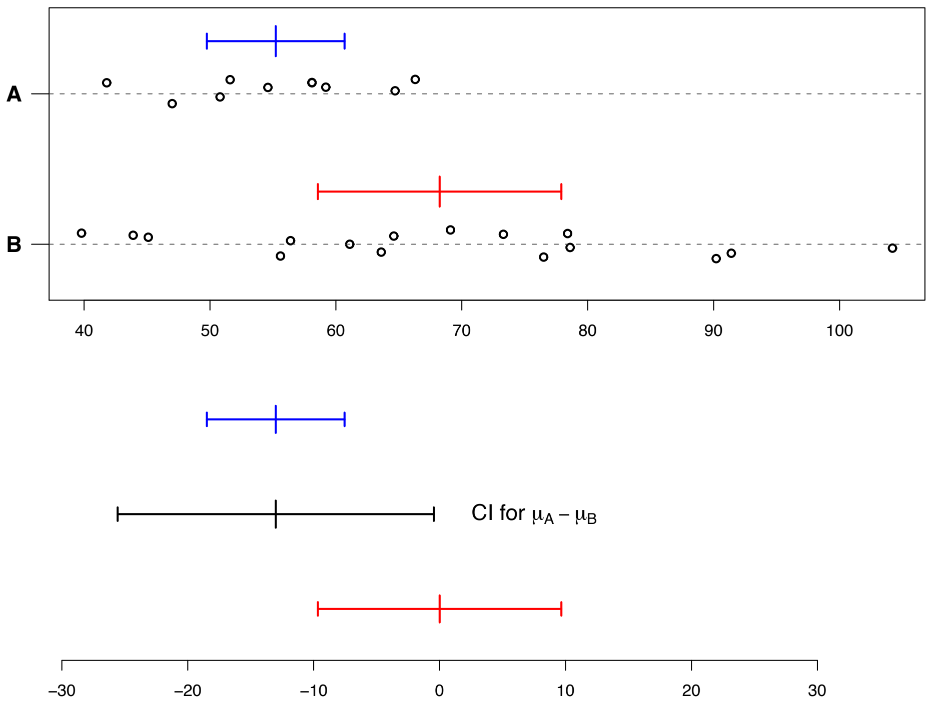
$$\widehat{\text{SD}}(\overline{X} - \overline{Y}) = \hat{\sigma}_{\text{pooled}} \times \sqrt{\tfrac{1}{n} + \tfrac{1}{m}} = 15.1 \times \sqrt{\tfrac{1}{10} + \tfrac{1}{16}} = 6.08$$

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

$\longrightarrow$ 95% confidence interval for $\mu_A - \mu_B$:

$$(55.2 - 68.2) \pm 2.06 \times 6.08 \quad = \quad -13.0 \pm 12.6 \quad = \quad (-25.6, -0.5)$$

Example



One problem

What if the two populations really have different SDs, σ_A and σ_B?

Suppose that

- ○ $X_1, X_2, \ldots, X_\mathrm{n}$ are iid Normal(μ_A, σ_A),
- ○ $Y_1, Y_2, \ldots, Y_\mathrm{m}$ are iid Normal(μ_B, σ_B).

Then

$$\mathrm{SD}(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_\mathrm{A}^2}{\mathrm{n}} + \frac{\sigma_\mathrm{B}^2}{\mathrm{m}}} \qquad \widehat{\mathrm{SD}}(\overline{X} - \overline{Y}) = \sqrt{\frac{S_\mathrm{A}^2}{\mathrm{n}} + \frac{S_\mathrm{B}^2}{\mathrm{m}}}$$

The problem:

$\longrightarrow$ $\dfrac{(\overline{X} - \overline{Y}) - (\mu_\mathrm{A} - \mu_\mathrm{B})}{\widehat{\mathrm{SD}}(\overline{X} - \overline{Y})}$ does not follow a t distribution.

An approximation

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

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

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

$\longrightarrow$ Use $(\overline{X} - \overline{Y}) \pm t^\star \, \widehat{SD}(\overline{X} - \overline{Y})$ as a 95% confidence interval.

Example

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

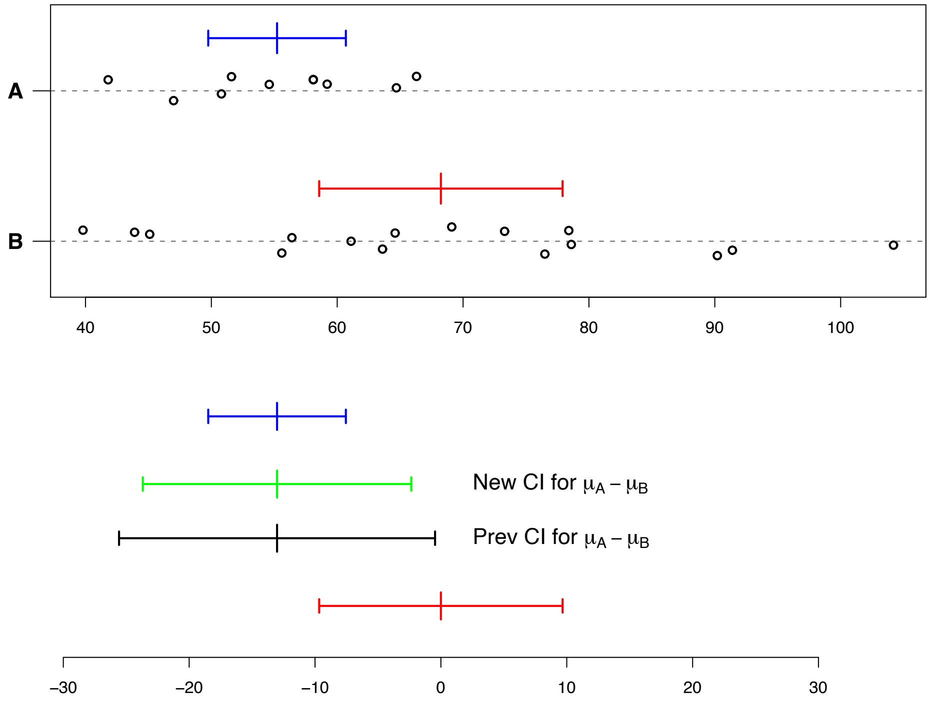
t value = `qt(0.975, 21.8)` = 2.07.

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

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

$$-13.0 \pm 2.07 \times 5.14 \quad = \quad -13.0 \pm 10.7 \quad = \quad (-23.7, -2.4)$$

Example



Degrees of freedom

- ○ One sample of size n:

$$X_1, X_2, \ldots, X_n \longrightarrow (\overline{X} - \mu)/(S/\sqrt{n}) \sim \mathrm{t}(\mathrm{df} = n - 1)$$

- ○ Two samples, of size n and m:

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

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 $(X_1 - \overline{X}), \dots, (X_n - \overline{X})$ to estimate σ.

$\longrightarrow$ 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 $(X_1 - \overline{X}), (X_2 - \overline{X}), \dots, (X_n - \overline{X})$ and $(Y_1 - \overline{Y}), \dots, (Y_m - \overline{Y})$ to estimate σ.

$\longrightarrow$ 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.

Confidence interval for the population SD

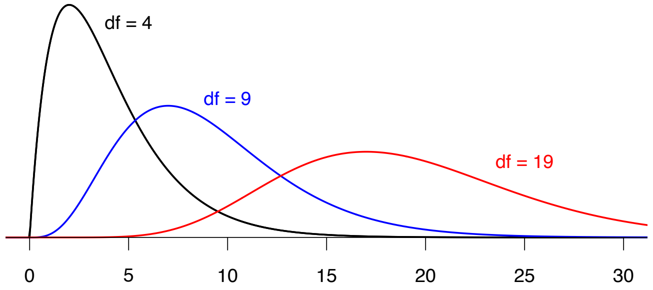
Suppose we observe $X_1, X_2, \dots, X_n$ iid Normal(μ, σ).

Suppose we wish to create a 95% CI for the population SD, σ.

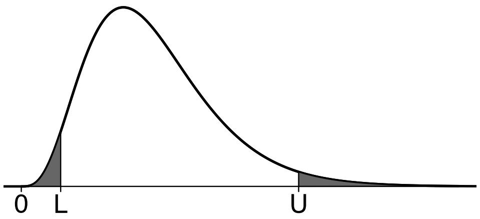
Our estimate of σ is the sample SD, S.

The sampling distribution of S is such that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(\text{df} = n - 1)$$



Confidence interval for the population SD



Choose L and U such that

$$\Pr\left(\mathrm{L} \le \frac{(\mathrm{n}-1)S^2}{\sigma^2} \le \mathrm{U}\right) = 95\%.$$

$$\Pr\left(\frac{1}{\mathrm{U}} \le \frac{\sigma^2}{(\mathrm{n}-1)S^2} \le \frac{1}{\mathrm{L}}\right) = 95\%.$$

$$\Pr\left(\frac{(\mathrm{n}-1)S^2}{\mathrm{U}} \le \sigma^2 \le \frac{(\mathrm{n}-1)S^2}{\mathrm{L}}\right) = 95\%.$$

$$\Pr\left(S\sqrt{\frac{\mathrm{n}-1}{\mathrm{U}}} \le \sigma \le S\sqrt{\frac{\mathrm{n}-1}{\mathrm{L}}}\right) = 95\%.$$

$\longrightarrow$ $\left(S\sqrt{\frac{\mathrm{n}-1}{\mathrm{U}}}\ ,\ S\sqrt{\frac{\mathrm{n}-1}{\mathrm{L}}}\right)$ is a 95% CI for σ.

Example

Strain A: n = 10; sample SD: $\mathrm{s}_\mathrm{A} = 7.64$

L = `qchisq(0.025,9)` = 2.70
U = `qchisq(0.975,9)` = 19.0

$\longrightarrow$ 95% CI for σ_A:

$$(7.64 \times \sqrt{\tfrac{9}{19.0}}, 7.64 \times \sqrt{\tfrac{9}{2.70}}) = (7.64 \times 0.69, 7.64 \times 1.83) = (5.3, 14.0)$$

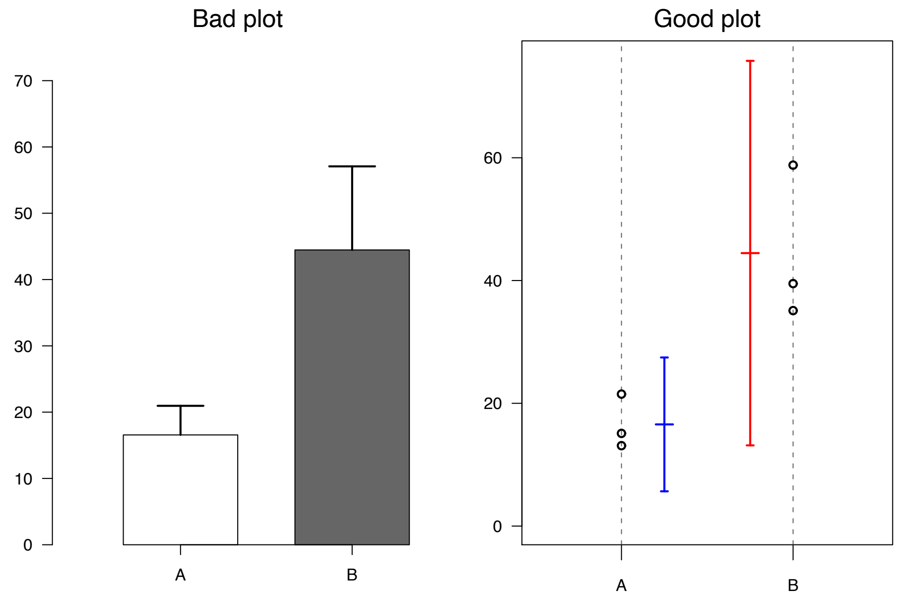
Strain B: n = 16; sample SD: $\mathrm{s}_\mathrm{B} = 18.1$

L = `qchisq(0.025,15)` = 6.25
U = `qchisq(0.975,15)` = 27.5

$\longrightarrow$ 95% CI for σ_B:

$$(18.1 \times \sqrt{\tfrac{15}{27.5}}, 18.1 \times \sqrt{\tfrac{15}{6.25}}) = (18.1 \times 0.74, 18.1 \times 1.55) = (13.4, 28.1)$$

Summarizing data


Bad plot
A
B
Good plot
A
B