

# Confidence Intervals

## Review

---

→ If  $X_1, \dots, X_n$  have mean  $\mu$  and SD  $\sigma$ , then

$$E(\bar{X}) = \mu \quad \text{no matter what}$$

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

→ If  $X_1, \dots, X_n$  are iid Normal(mean= $\mu$ , SD= $\sigma$ ), then

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

→ If  $X_1, \dots, X_n$  are iid with mean  $\mu$  and SD  $\sigma$  and the sample size  $n$  is large, then

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

## 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 ( $\mu$ ) is roughly  
 $3.52 \pm (2 \times 0.16) = 3.52 \pm 0.32 = (3.20, 3.84)$ .

What does this mean?

## Confidence intervals

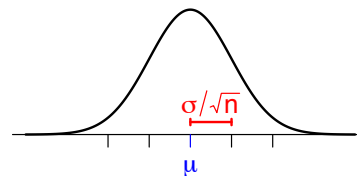
---

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

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

→ How close is  $\bar{X}$  to  $\mu$ ?

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



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

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

# 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,  $\bar{X} \pm 1.96\sigma/\sqrt{n}$  has a 95% chance of covering  $\mu$ .

Thus, it is called a 95% confidence interval for  $\mu$ .

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

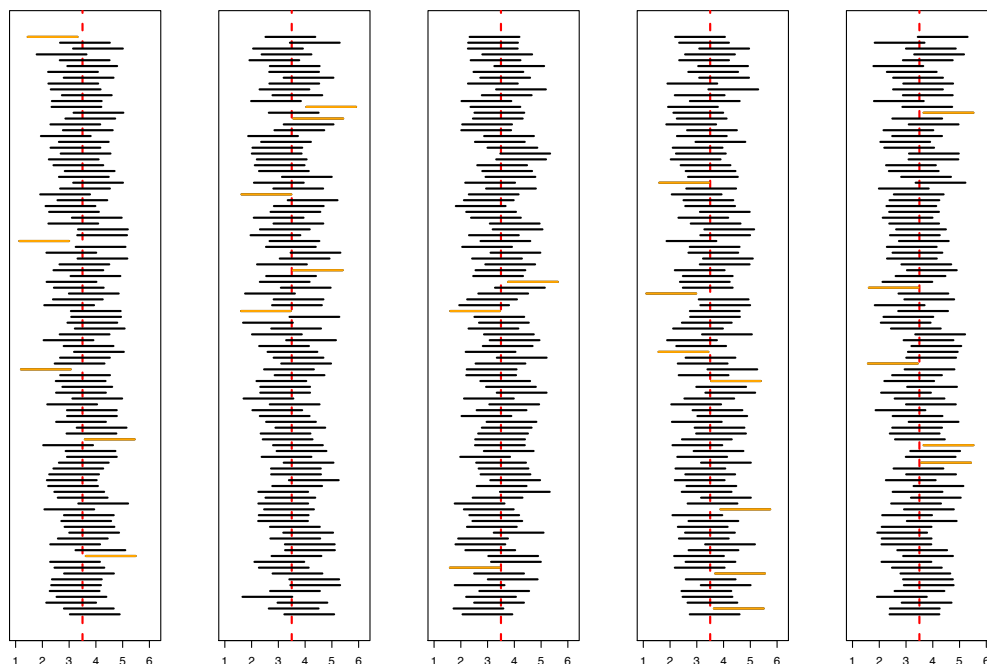
$$\bar{x} \pm 1.96\sigma/\sqrt{n} = 3.52 \pm 0.32.$$

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

# What is a confidence interval?

---

500 confidence intervals for  $\mu$   
( $\sigma$  known)



## Longer and shorter intervals

---

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

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

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

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

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

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

## 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  $\bar{X} \pm 1.96 \sigma / \sqrt{n}$  as a 95% confidence interval for  $\mu$  requires knowledge of  $\sigma$ .

That the above is a 95% confidence interval for  $\mu$  is a result of the following:

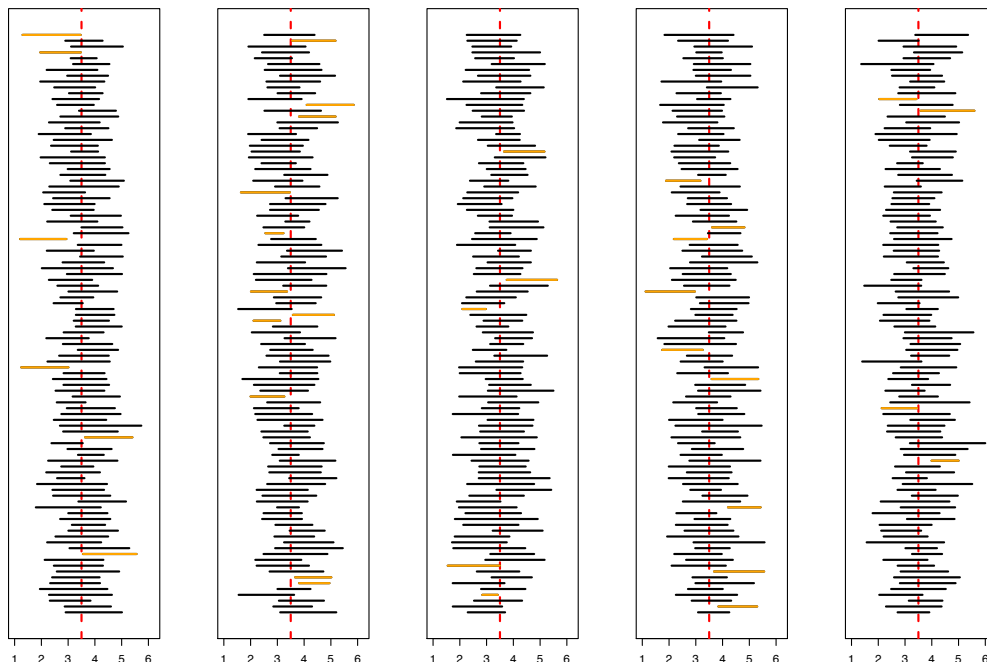
$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \text{Normal}(0,1)$$

What if we **don't know**  $\sigma$ ?

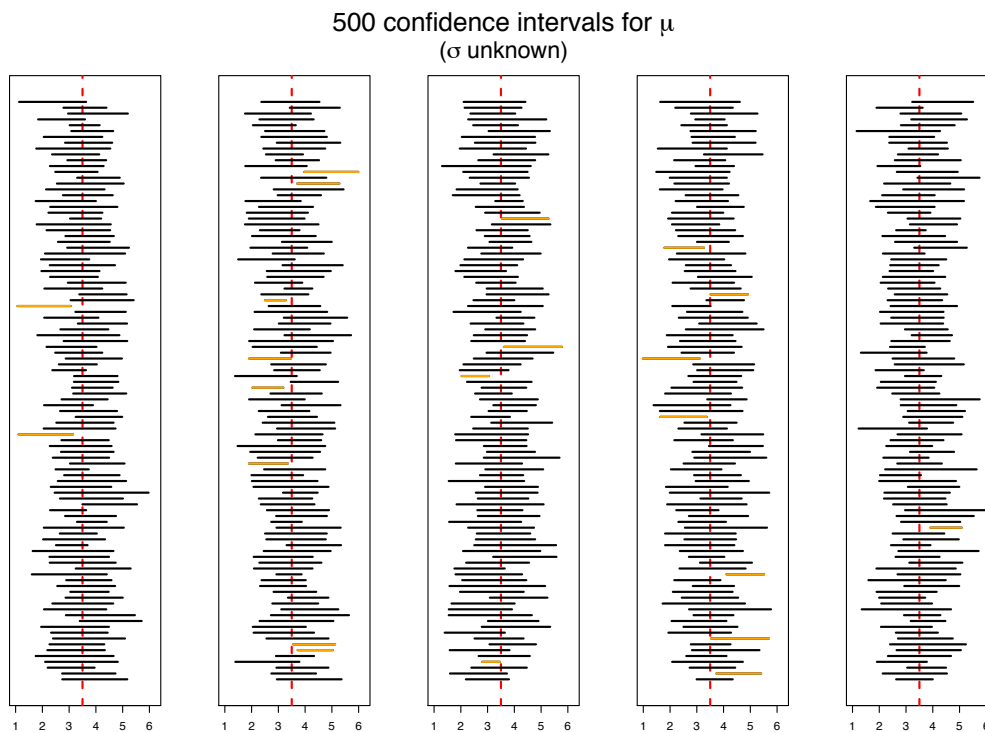
- We plug in the sample SD  $S$ , but then we need to widen the intervals to account for the uncertainty in  $S$ .

## What is a confidence interval? (cont)

500 BAD confidence intervals for  $\mu$   
( $\sigma$  unknown)



# What is a confidence interval? (cont)



## The Student t distribution

If  $X_1, X_2, \dots, X_n$  are iid Normal(mean= $\mu$ , SD= $\sigma$ ), then

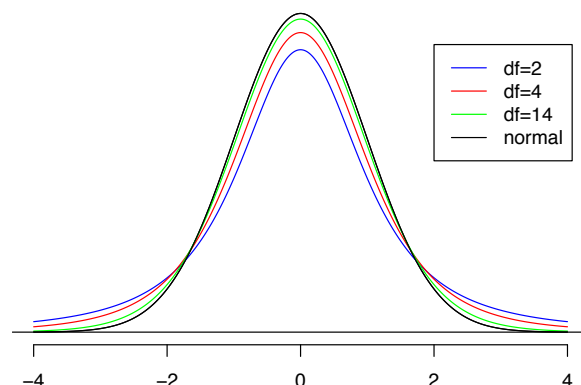
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(\text{df} = n - 1)$$

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)



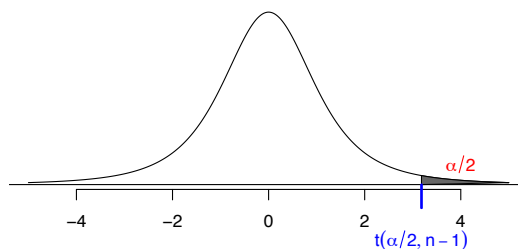
# The t interval

If  $X_1, \dots, X_n$  are iid Normal(mean= $\mu$ , SD= $\sigma$ ), then

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

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

→  $t(\alpha/2, n-1)$  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  $n=10$ ,  $\alpha=5\%$ .

## Example 1

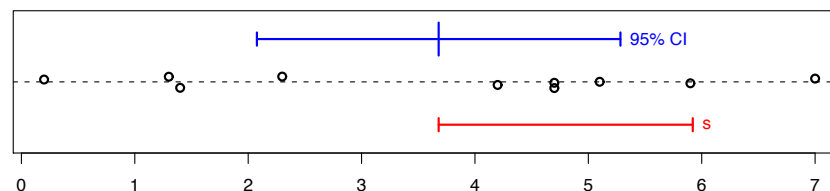
Suppose we have measured the  $\log_{10}$  cytokine response of 10 mice, and obtained the following numbers:

### Data

0.2	1.3	1.4	2.3	4.2	$\bar{x} = 3.68$	$n = 10$
4.7	4.7	5.1	5.9	7.0	$s = 2.24$	$qt(0.975, 9) = 2.26$

→ 95% confidence interval for  $\mu$  (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

$$\bar{x} = 5.09$$

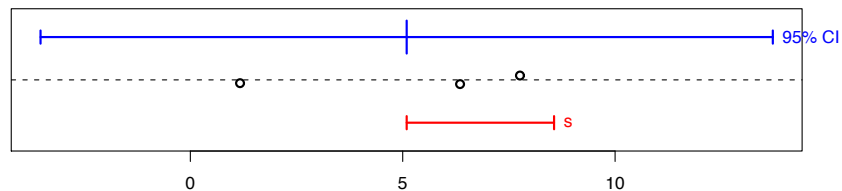
$$n = 3$$

$$s = 3.47$$

$$qt(0.975, 2) = 4.30$$

→ 95% confidence interval for  $\mu$  (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

$$\bar{x} = 30.7$$

$$n = 20$$

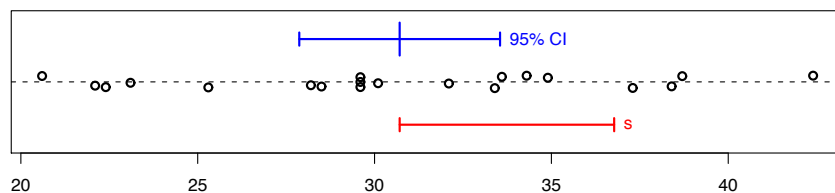
28.2 25.3 ... ... 32.1

$$s = 6.06$$

$$qt(0.975, 19) = 2.09$$

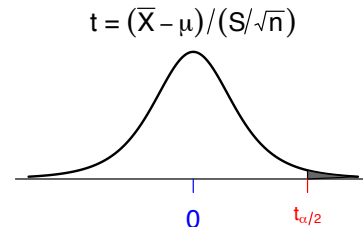
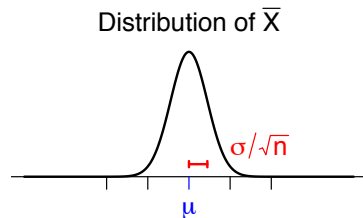
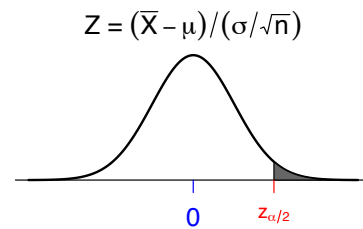
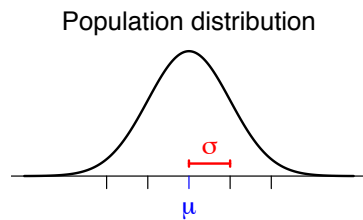
→ 95% confidence interval for  $\mu$  (the population mean):

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





# Confidence interval for the mean



$X_1, X_2, \dots, X_n$  independent  $\text{Normal}(\mu, \sigma)$ .

95% confidence interval for  $\mu$ :

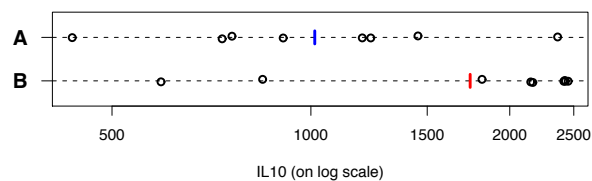
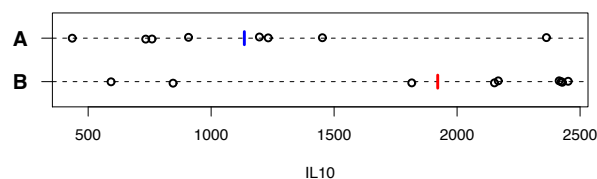
$\bar{X} \pm t S / \sqrt{n}$  where  $t = 97.5$  percentile of  $t$  distribution with  $(n - 1)$  d.f.

# 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?

→ I am not interested in these *particular* mice, but in the strains *generally*.



$$\bar{X} - \bar{Y}$$


---

Suppose that

- $X_1, X_2, \dots, X_n$  are iid Normal(mean= $\mu_A$ , SD= $\sigma$ ), and
- $Y_1, Y_2, \dots, Y_m$  are iid Normal(mean= $\mu_B$ , SD= $\sigma$ ).

Then

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

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

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

Note: If  $n = m$ , 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.

$$\begin{aligned} \hat{\sigma}_{\text{pooled}} &= \sqrt{\frac{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2}{n+m-2}} \\ &= \sqrt{\frac{S_A^2(n-1) + S_B^2(m-1)}{n+m-2}} \end{aligned}$$

Note: If  $n = m$ , then  $\hat{\sigma}_{\text{pooled}} = \sqrt{(S_A^2 + S_B^2) / 2}$

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

---

$$\begin{aligned}\widehat{SD}(\bar{X} - \bar{Y}) &= \hat{\sigma}_{\text{pooled}} \sqrt{\frac{1}{n} + \frac{1}{m}} \\ &= \sqrt{\left[ \frac{S_A^2(n-1) + S_B^2(m-1)}{n+m-2} \right] \cdot \left[ \frac{1}{n} + \frac{1}{m} \right]}\end{aligned}$$

In the case  $n = m$ ,

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

## CI for the difference between the means

---

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

The procedure:

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

## 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

$m = 8$ ,  $\bar{y} \approx 3.44$ ,  $s_B \approx 0.250$

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

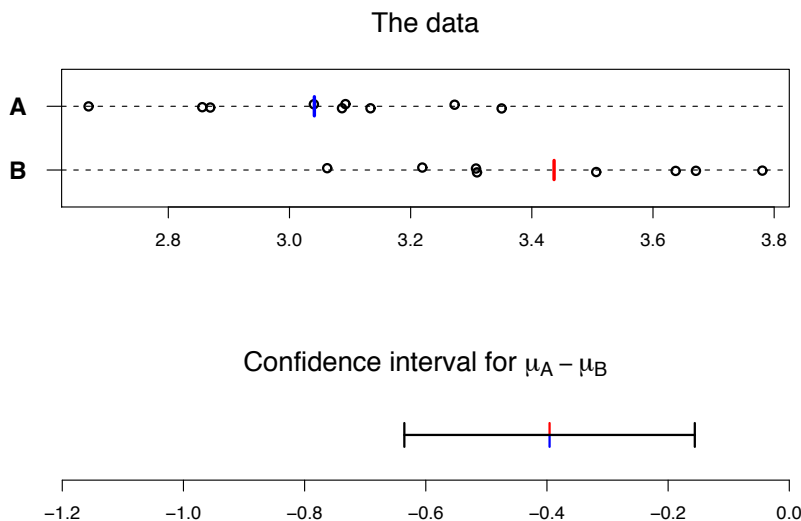
$$\widehat{SD}(\bar{X} - \bar{Y}) = \hat{\sigma}_{\text{pooled}} \sqrt{\frac{1}{n} + \frac{1}{m}} = \dots \approx 0.112$$

97.5 percentile of  $t(\text{df}=15) \approx 2.13$

## 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

---

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:  $s_B = 18.1$   
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

---

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

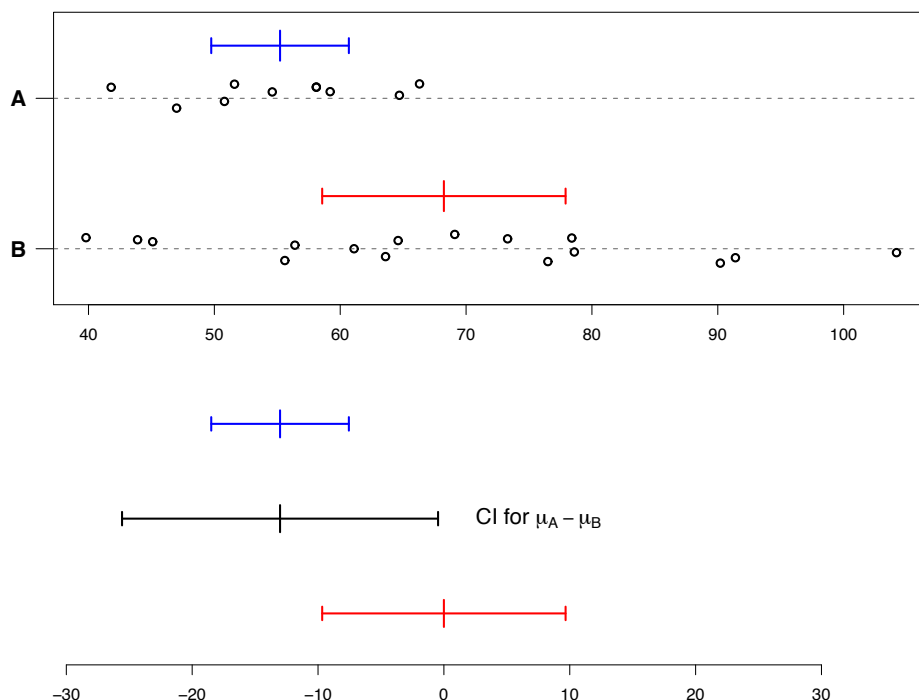
$$\widehat{SD}(\bar{X} - \bar{Y}) = \hat{\sigma}_{\text{pooled}} \times \sqrt{\frac{1}{n} + \frac{1}{m}} = 15.1 \times \sqrt{\frac{1}{10} + \frac{1}{16}} = 6.08$$

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

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

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

## Example



## One problem

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

Suppose that

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

Then

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

The problem:

→  $\frac{(\bar{X} - \bar{Y}) - (\mu_A - \mu_B)}{\widehat{SD}(\bar{X} - \bar{Y})}$  does not follow a t distribution.

## An approximation

---

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

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

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

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

## Example

---

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

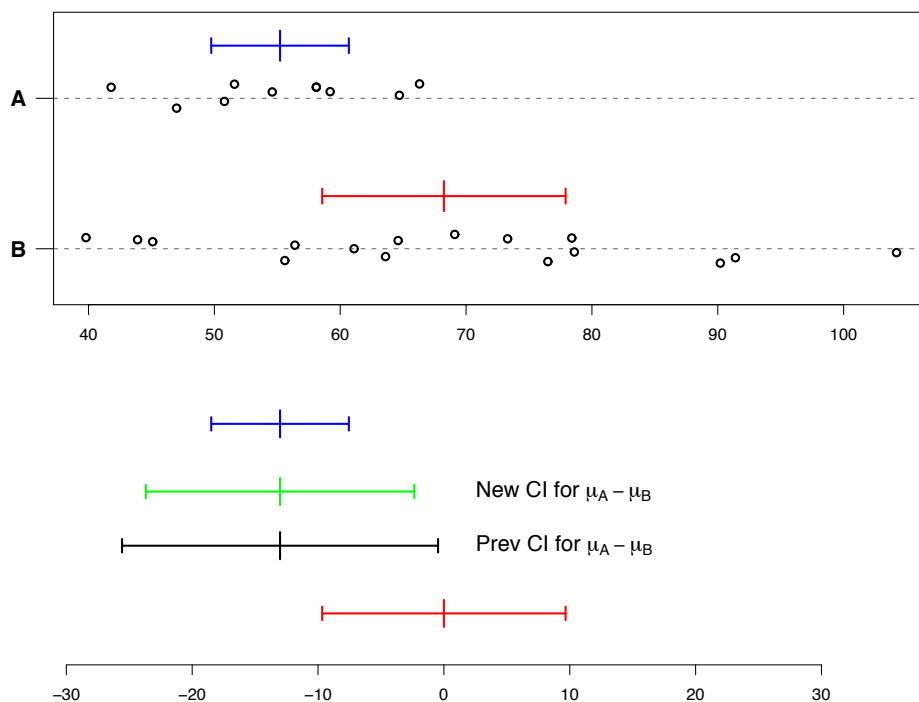
$$t \text{ value} = qt(0.975, 21.8) = 2.07.$$

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

→ 95% CI for  $\mu_A - \mu_B$ :

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

## Example



## Degrees of freedom

- One sample of size  $n$ :

$$X_1, X_2, \dots, X_n \longrightarrow (\bar{X} - \mu) / (S / \sqrt{n}) \sim t(\text{df} = n - 1)$$

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

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

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  $(X_1 - \bar{X}), \dots, (X_n - \bar{X})$  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  $(X_1 - \bar{X}), (X_2 - \bar{X}), \dots, (X_n - \bar{X})$  and  $(Y_1 - \bar{Y}), \dots, (Y_m - \bar{Y})$  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 the population SD

---

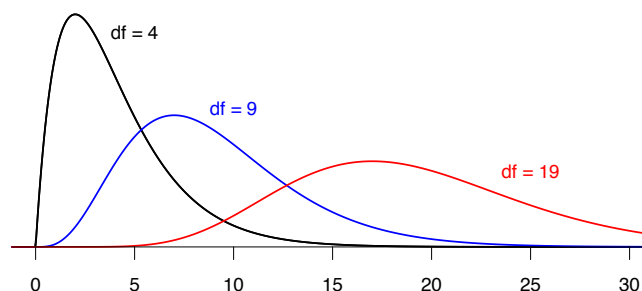
Suppose we observe  $X_1, X_2, \dots, X_n$  iid  $\text{Normal}(\mu, \sigma)$ .

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

Our estimate of  $\sigma$  is the sample SD,  $S$ .

The sampling distribution of  $S$  is such that

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

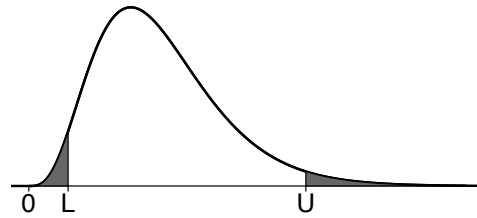


# Confidence interval for the population SD

---

Choose L and U such that

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



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

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

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

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

## Example

---

Strain A:  $n = 10$ ; 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  $\sigma_A$ :

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

Strain B:  $n = 16$ ; sample SD:  $s_B = 18.1$

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

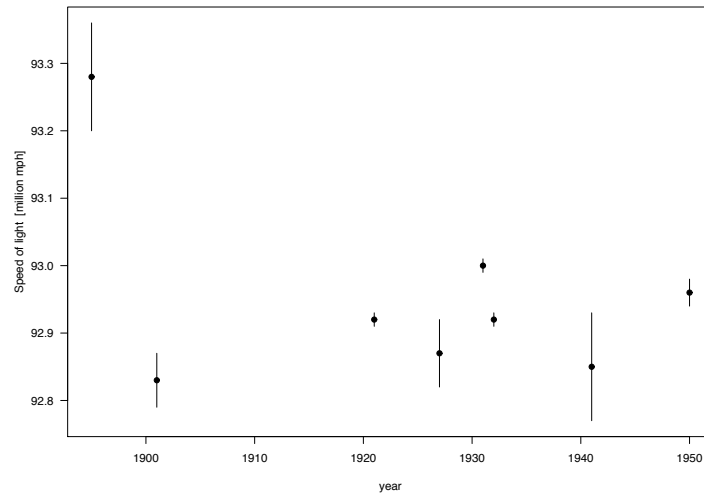
$$U = \text{qchisq}(0.975, 15) = 27.5$$

→ 95% CI for  $\sigma_B$ :

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

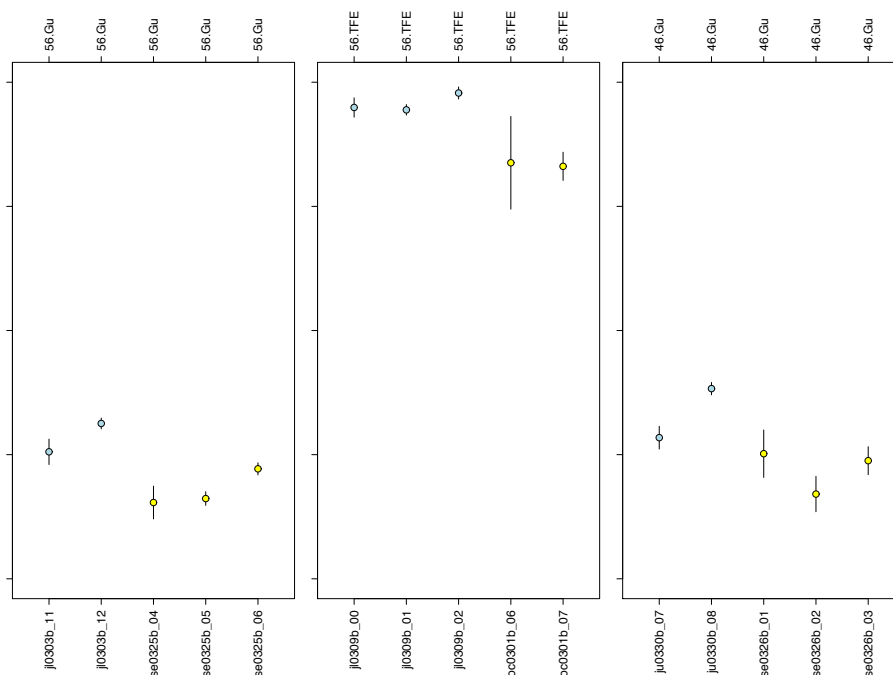
# Confidence interval for what..?

Estimates of the speed of light, with “confidence intervals”.



Youden W (1972). *Technometrics* 14: 1-11.

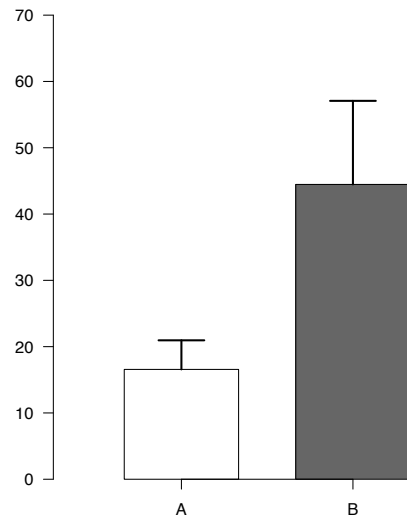
# Confidence interval for what..?



# Summarizing data

---

Bad plot



Good plot

