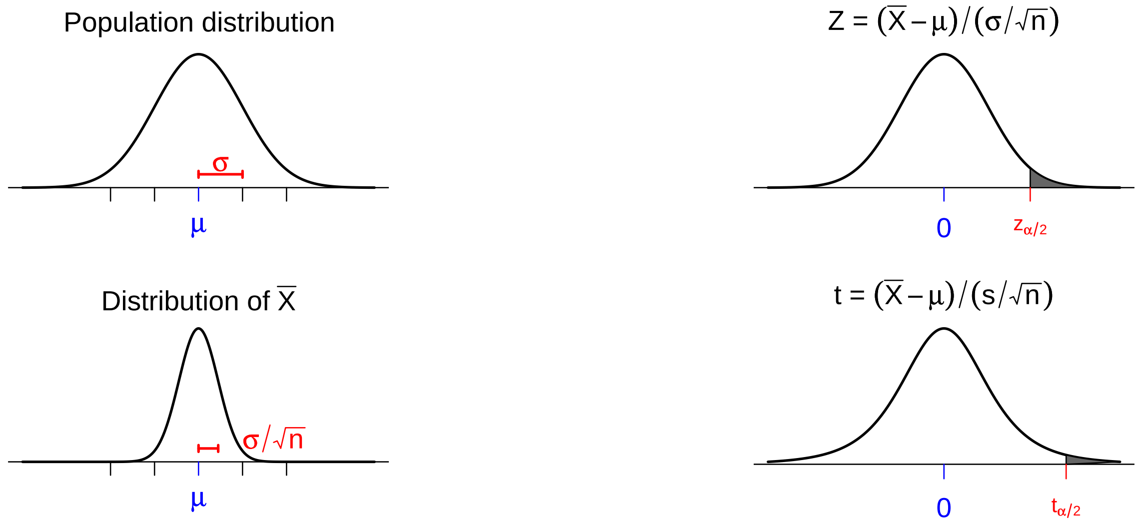
Review



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

95% confidence interval for μ: $\bar{X} \pm t\ s/\sqrt{n}$

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

Data

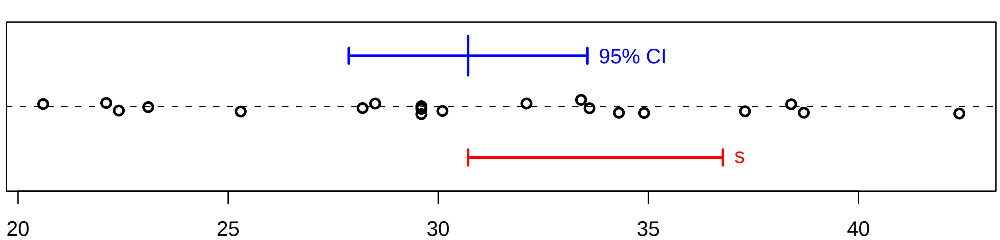
34.9	28.5	34.3	38.4	29.6
28.2	25.3	...	...	32.1

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

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

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

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



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).

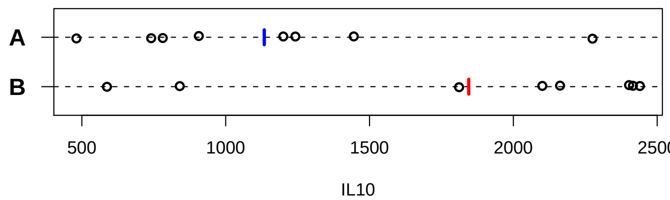
But this is beyond the scope of the current course.

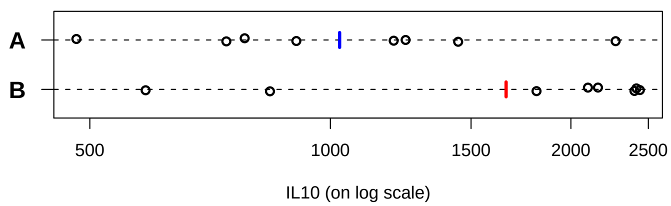
Differences between means

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*.





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

Suppose $X_1, X_2, \ldots, X_n$ are indep. normal(mean=μ_A, SD=σ)

and $Y_1, Y_2, \ldots, Y_m$ are indep. normal(mean=μ_B, SD=σ)

$$\begin{aligned} \mathrm{E}(\bar{X} - \bar{Y}) &= \mathrm{E}(\bar{X}) - \mathrm{E}(\bar{Y}) \\ &= \mu_A - \mu_B \end{aligned}$$

$$\begin{aligned} \mathrm{SD}(\bar{X} - \bar{Y}) &= \sqrt{\mathrm{SD}(\bar{X})^2 + \mathrm{SD}(\bar{Y})^2} \\ &= \sqrt{\left(\frac{\sigma}{\sqrt{n}}\right)^2 + \left(\frac{\sigma}{\sqrt{m}}\right)^2} = \sigma\sqrt{\frac{1}{n} + \frac{1}{m}} \end{aligned}$$

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

Pooled estimate of pop'n SD

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

$$\hat{\sigma}_A = s_A = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} \qquad \hat{\sigma}_B = s_B = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{m-1}}$$

We can use all of the data together to obtain an improved estimate of σ, 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, $\hat{\sigma}_{\text{pooled}} = \sqrt{(s_A^2 + s_B^2)/2}$

Est'd SE of $(\bar{X} - \bar{Y})$

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

In the case n = m,

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

CI for difference between means

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

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

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

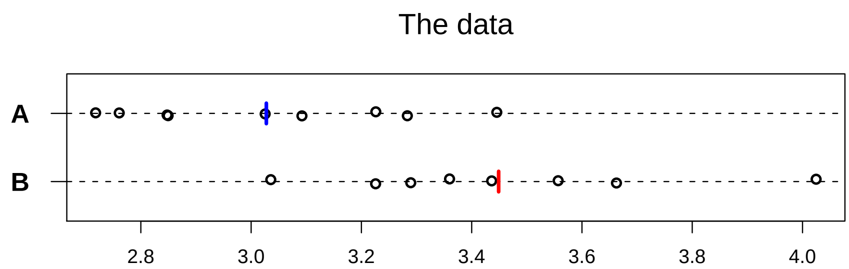
$$\widehat{SD}(\bar{X} - \bar{Y}) = \hat{\sigma}_{\text{pooled}} \sqrt{\frac{1}{n} + \frac{1}{m}} = \ldots \approx 0.112$$

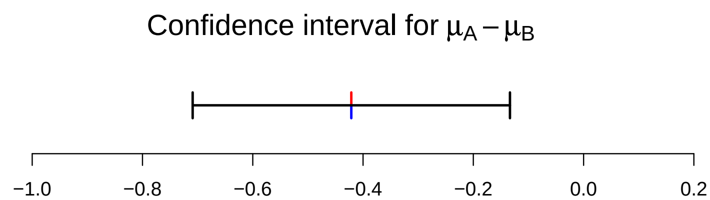
97.5 percentile of t(df=15) $\approx$ 2.13

Example

95% confidence interval:

$$(3.04 - 3.44) \pm \color{orchid}{2.13} \cdot 0.112$$
$$\approx -0.40 \pm 0.24$$
$$= \mathbf{(-0.64, -0.16)}.$$





Example

Strain A:

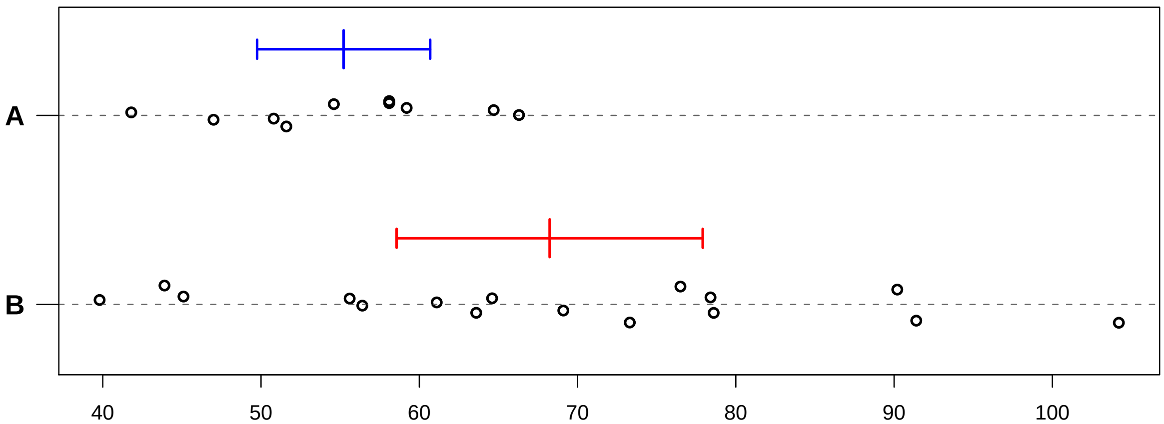
n = 10
sample mean: $\bar{X} = 55.22$
sample SD: $s_A = 7.64$
t value = `qt(0.975, 9)` = 2.26

95% CI for μ_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 μ_B: $68.2 \pm 2.13 \times 18.1 / \sqrt{16}$
$= 68.2 \pm 9.7 = (58.6, 77.9)$



Example

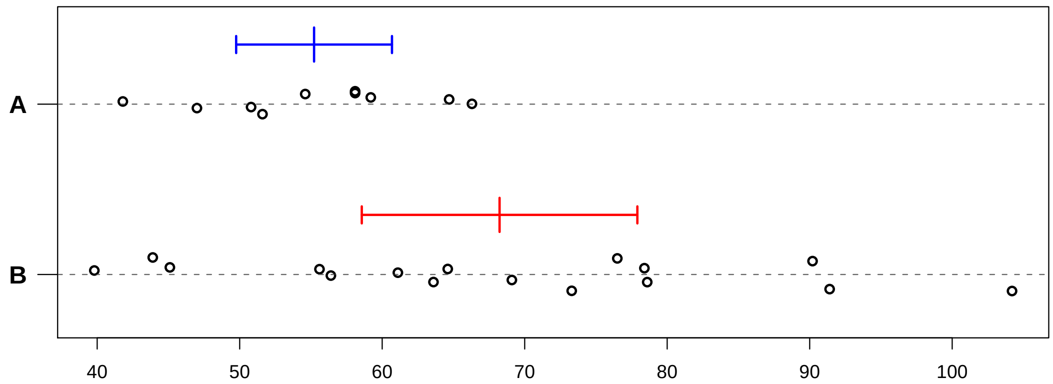
$\hat{\sigma}_{\text{pooled}} = \sqrt{\frac{(7.64)^2 \times (10-1) + (18.1)^2 \times (16-1)}{10+16-2}} = 15.1$

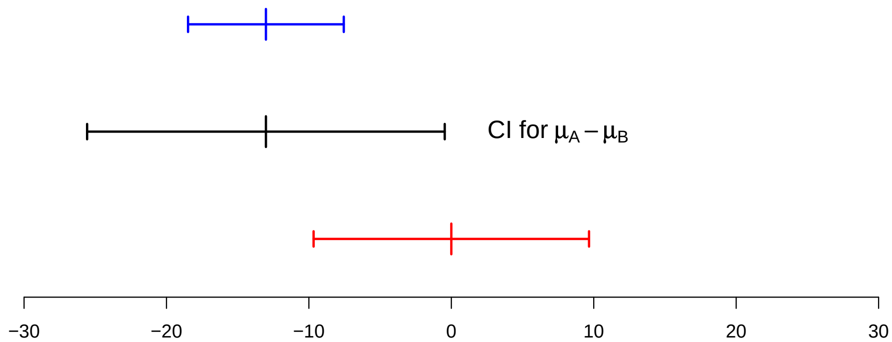
$\widehat{\text{SD}}(\bar{X} - \bar{Y}) = \hat{\sigma}_{\text{pooled}} \times \sqrt{\frac{1}{\text{n}} + \frac{1}{\text{m}}} = 15.1 \times \sqrt{\frac{1}{10} + \frac{1}{16}} = 6.08$

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

95% confidence interval for $\mu_{\text{A}} - \mu_{\text{B}}$:

(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, σ_A and σ_B?

If $X_1, X_2, \ldots, X_n$ are iid normal(μ_A, σ_A)
and $Y_1, Y_2, \ldots, Y_m$ are iid normal(μ_B, σ_B),

$$\mathrm{SD}(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}} \qquad \widehat{\mathrm{SD}}(\bar{X} - \bar{Y}) = \sqrt{\frac{s_A^2}{n} + \frac{s_B^2}{m}}$$

The problem:

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

An approximation

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

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

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

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

Example

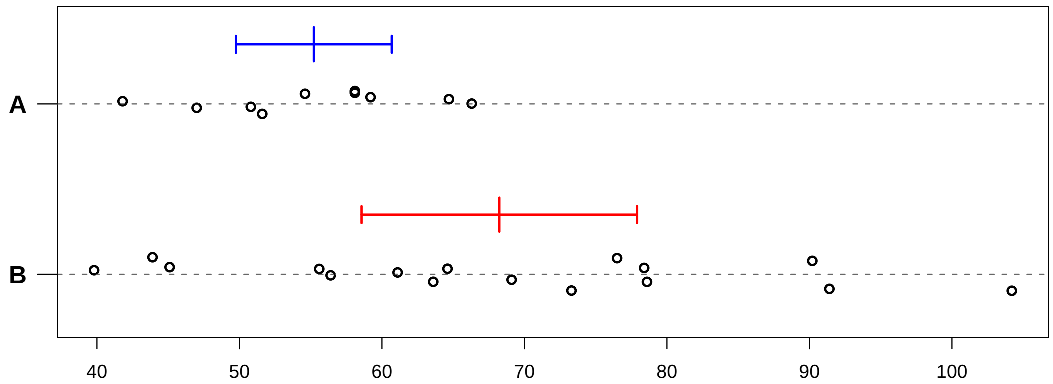
$$k = \frac{[(7.64)^2/10 + (18.1)^2/16]^2}{\frac{[(7.64)^2/10]^2}{9} + \frac{[(18.1)^2/16]^2}{15}} = \frac{(5.84 + 20.6)^2}{\frac{(5.84)^2}{9} + \frac{(20.6)^2}{15}} = 21.8.$$

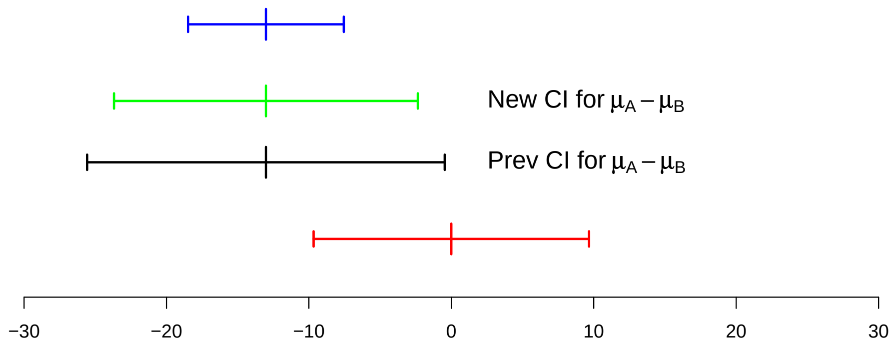
t value = `qt(0.975, 21.8)` = 2.07.

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

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

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





Degrees of freedom

One sample of size n:

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

Two samples, of size n and m:

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

What are these "degrees of freedom"?

Degrees of freedom

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

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

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 σ.

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

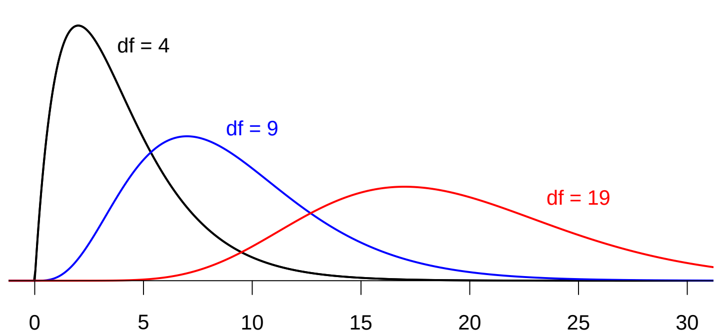
Suppose we observe $X_1, X_2, \ldots, X_n$ iid normal(μ, σ).

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

Our estimate of σ is, of course, the sample SD, s.

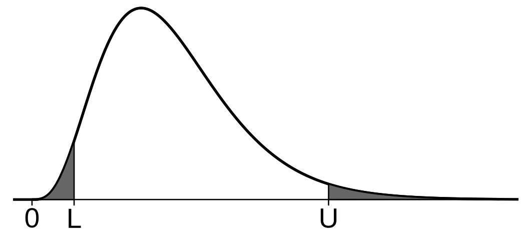
The sampling distribution of s is such that

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



Choose L and U such that

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



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

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

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

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

Example

Strain A:

$n = 10$ sample SD: $s_A = 7.64$

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

95% CI for σ_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:

$n = 16$ sample SD: $s_B = 18.1$

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

95% CI for σ_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)