Review

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

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

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

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

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

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

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

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 (μ) is

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

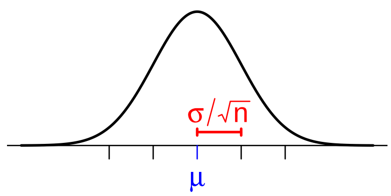
What does this mean?

What is the chance that (3.20, 3.84) contains μ?

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

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

How close is $\bar{X}$ to μ?



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

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

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

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.

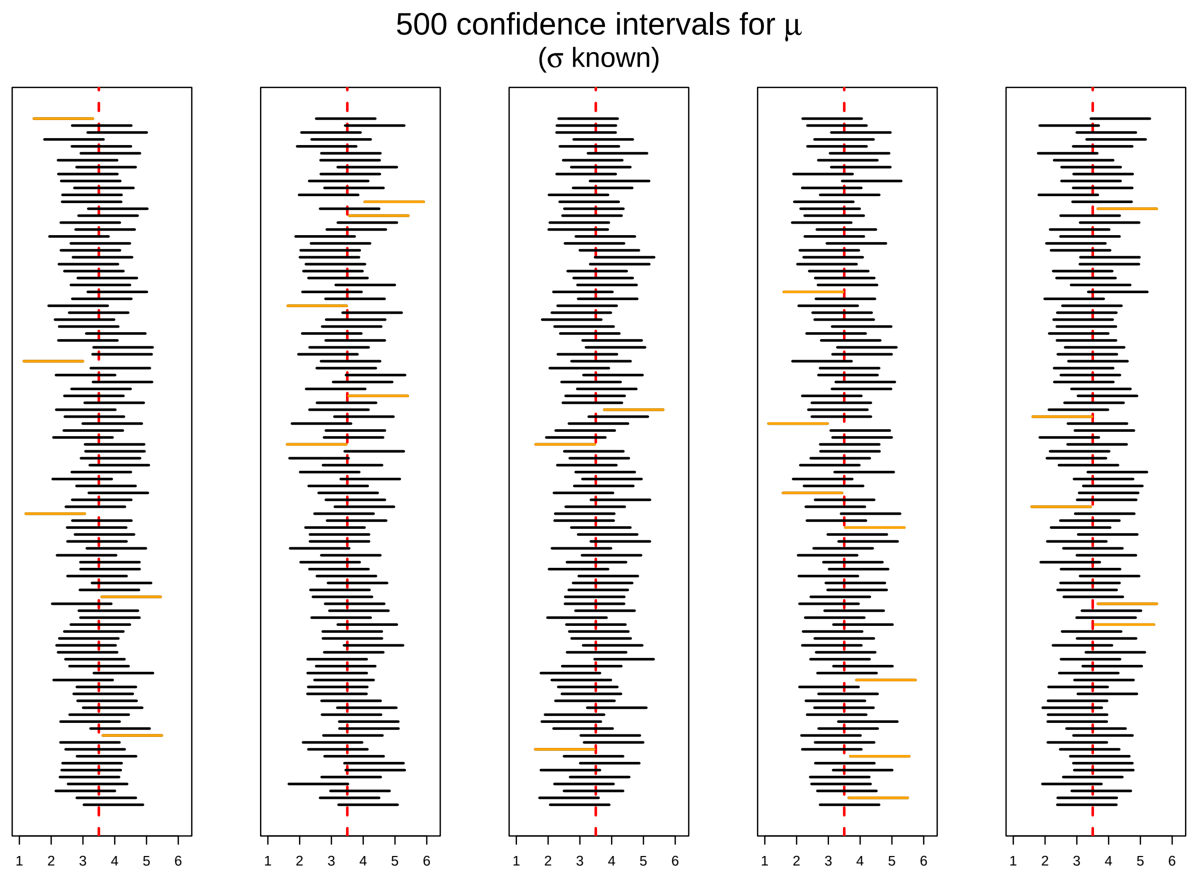
In advance, $\bar{X} \pm 1.96\sigma/\sqrt{n}$ has a 95% chance of covering μ.

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

Note that, after the data is gathered (for instance, n=100, $\bar{X} = 3.52$, s $= 1.61$), 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.



Longer and shorter intervals

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

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 μ.

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

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 μ.

What is a confidence interval?

A 95% confidence interval is a 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 μ requires knowledge of σ.

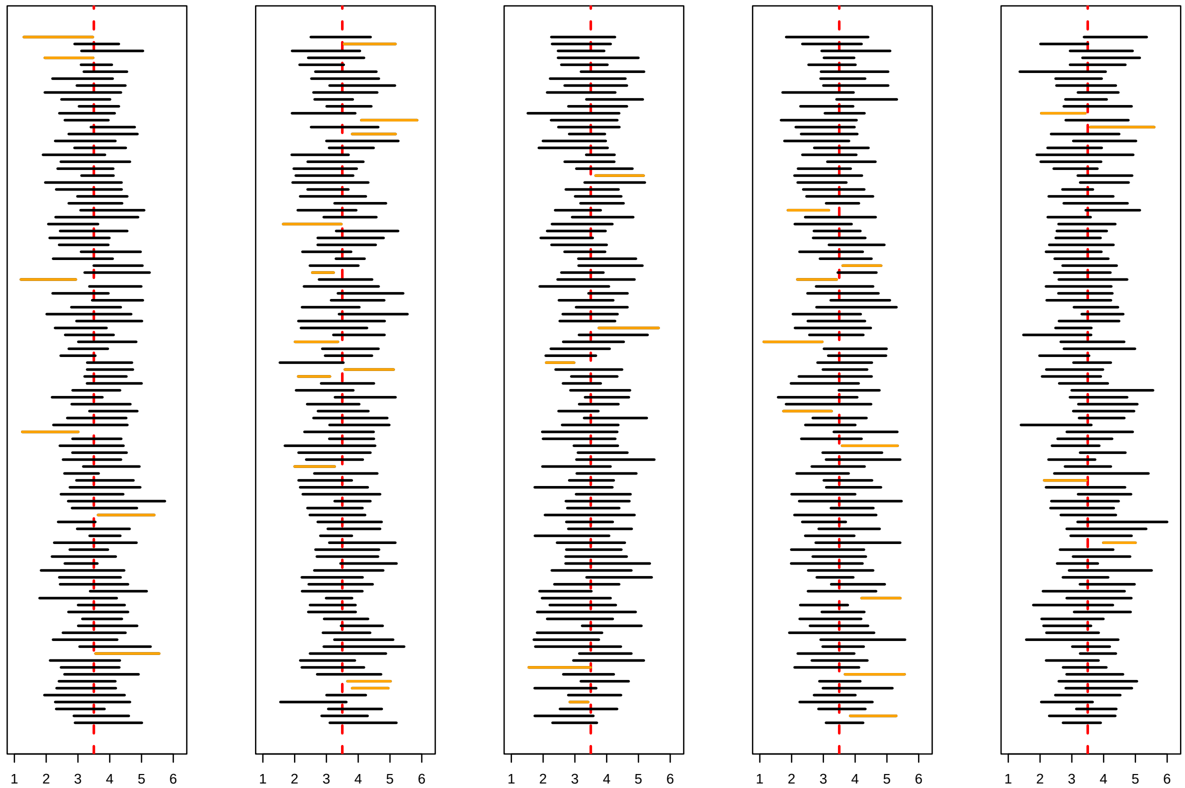
That the above is a 95% confidence interval for μ is a result of the following:

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

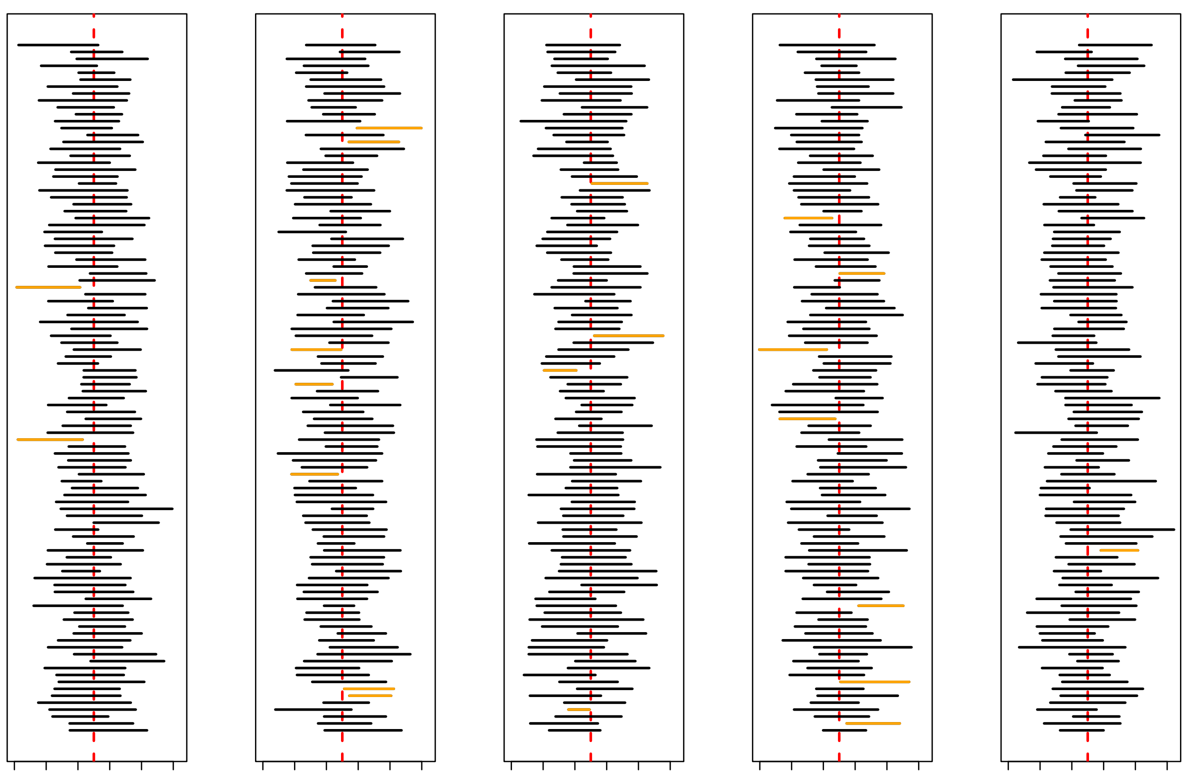
What if we don't know σ?

We plug in the sample SD (s), but then we need to widen the intervals to account for the uncertainty in s.

500 BAD confidence intervals for μ
(σ unknown)



500 confidence intervals for μ
(σ unknown)



The Student t distribution

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

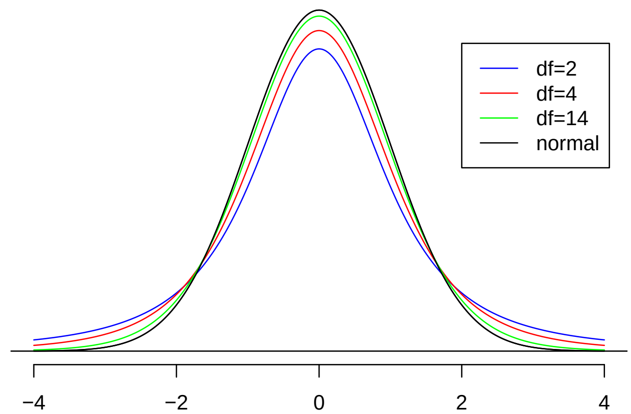
$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(df = n - 1)$$

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

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

e.g., `qt(0.975,9)` returns `2.26` (cf 1.96)

`pt(1.96,9)-pt(-1.96,9)` returns `0.918` (cf 0.95)

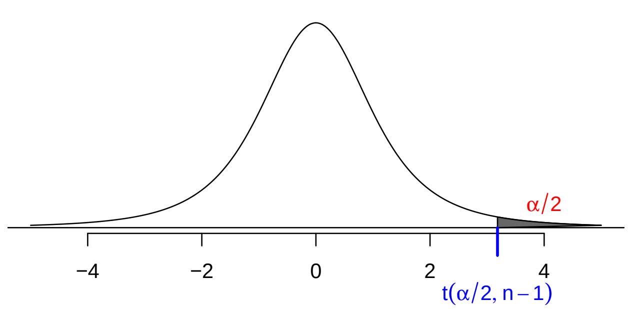


The t interval

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

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

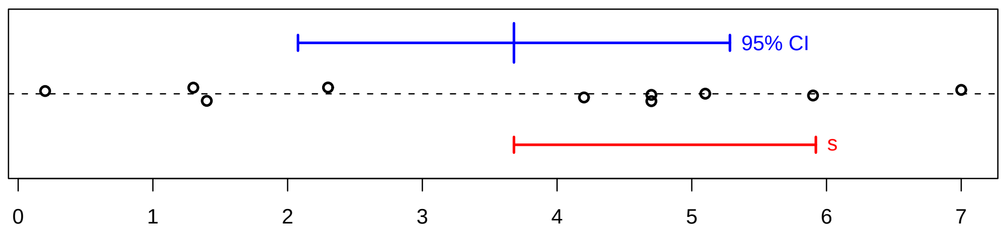
Data

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 RealTime-PCR) the $\log_{10}$ expression of a gene in 3 tissue samples, and obtained the following numbers:

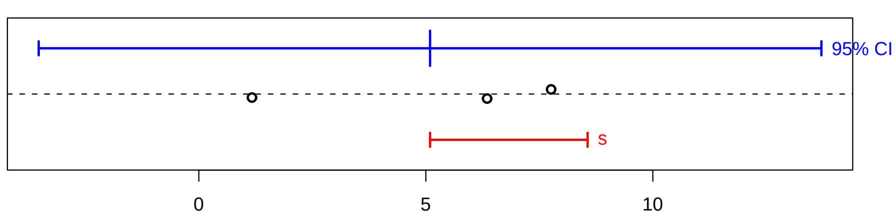
Data

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):

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