

Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture 17

Last Time

Confidence Interval for Population Mean:

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \times \sigma / \sqrt{n}$$

Based on Assumptions:

1. The population standard deviation σ was known.
2. The population is normally distributed (bell-shaped).

Today

What if population is normal but σ is unknown?

We Don't know σ . What to use instead?

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \times \sigma / \sqrt{n}$$

What about Sample Standard Deviation S ?

$$P\left(-2 \leq \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \leq 2\right) = 0.95 \text{ ???}$$

Not Quite!

Although $(\bar{X}_n - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$, $S \neq \sigma$. In fact, S is an **estimator** of σ so it is a **random variable**!

What is the sampling distribution?

Suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

$$\boxed{\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim ???}$$

First Step

What is the sampling distribution of S ?

What is the Distribution?



Suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. What is the distribution of this sum?

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2$$

- (a) $\chi^2(n)$
- (b) $N(\mu, \sigma^2)$
- (c) $N(0, 1)$
- (d) $N(\mu, \sigma^2/n)$
- (e) $\chi^2(1)$

Towards the Sampling Dist. of S

If $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, then

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

Now:

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \left(\frac{n-1}{\sigma^2} \right) \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \mu)^2 \right] \sim \chi^2(n)$$

Anything look familiar?

Sampling Distribution of Sample Variance

Suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Then whereas

$$\left(\frac{n-1}{\sigma^2}\right) \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \mu)^2 \right] \sim \chi^2(n)$$

Replacing μ with \bar{X} “loses” a degree of freedom

$$\left(\frac{n-1}{\sigma^2}\right) \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right] = \left(\frac{n-1}{\sigma^2}\right) S^2 \sim \chi^2(n-1)$$

Ultimately, we will use this fact to work out the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)/S$, but for now let's take a detour...

95% CI for Variance of Normal Population

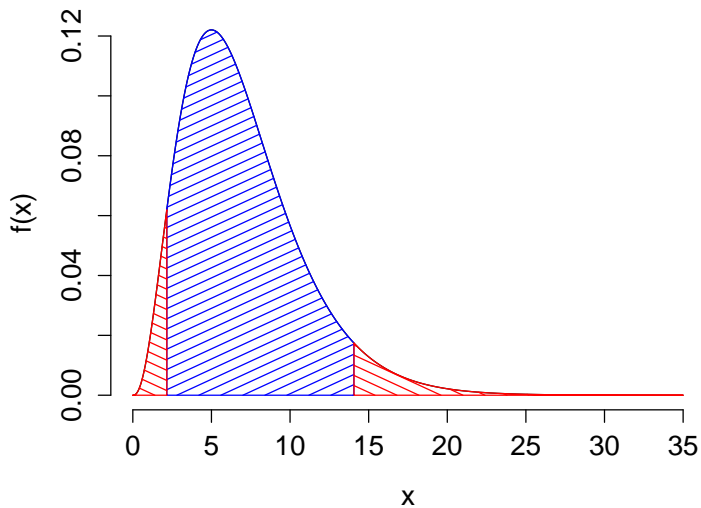
We know that:

$$\left(\frac{n-1}{\sigma^2} \right) S^2 \sim \chi^2(n-1)$$

First Step: find a, b such that

$$P \left[a \leq \left(\frac{n-1}{\sigma^2} \right) S^2 \leq b \right] = 0.95$$

Although there are many choices for a, b that would work, a sensible idea is to put 2.5% in each tail...



What R command should I use to calculate a ?



$$P \left[a \leq \left(\frac{n-1}{\sigma^2} \right) S^2 \leq b \right] = 0.95$$

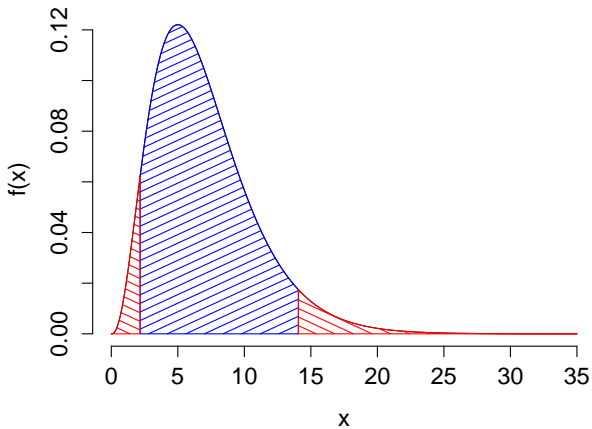
- (a) `qchisq(0.95, df = n - 1)`
- (b) `qchisq(0.025, df = n)`
- (c) `qchisq(0.975, df = n - 1)`
- (d) `qchisq(0.025, df = n - 1)`
- (e) `qchisq(0.975, df = n)`

What R command should I use to calculate b ?



$$P \left[a \leq \left(\frac{n-1}{\sigma^2} \right) S^2 \leq b \right] = 0.95$$

- (a) `qchisq(0.95, df = n - 1)`
- (b) `qchisq(0.025, df = n)`
- (c) `qchisq(0.975, df = n - 1)`
- (d) `qchisq(0.025, df = n - 1)`
- (e) `qchisq(0.975, df = n)`



```
a = qchisq(0.025, df = n - 1)
```

```
b = qchisq(0.975, df = n - 1)
```

Step 2: After Finding a, b Rearrange

$$P \left[a \leq \left(\frac{n-1}{\sigma^2} \right) S^2 \leq b \right] = 0.95$$

$$P \left[\frac{a}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{b}{(n-1)S^2} \right] = 0.95$$

$$P \left[\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a} \right] = 0.95$$

This CI is *not* symmetric: it *doesn't* take the form $\hat{\theta} \pm ME$!

Example: 95% Confidence Interval for Normal Variance

$X_1, \dots, X_{100} \sim N(\mu, \sigma^2)$. Here $n - 1 = 99$, hence

$$a = \text{qchisq}(0.025, \text{df} = 99) \approx 73$$

$$b = \text{qchisq}(0.975, \text{df} = 99) \approx 128$$

From the sample data, $s^2 = 4.3$

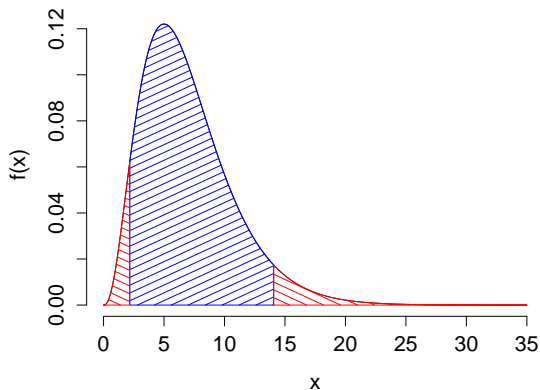
$$LCL = (n - 1)s^2/b = 99 \times 4.3/128 \approx 3.3$$

$$UCL = (n - 1)s^2/a = 99 \times 4.3/73 \approx 5.8$$

95% CI for σ^2 is [3.3, 5.8]. What values are plausible?

The actual population variance in this case was 4

Arbitrary Confidence Level: $(1 - \alpha)$



```
a = qchisq( $\alpha/2$ , df = n - 1)
```

```
b = qchisq( $1 - \alpha/2$ , df = n - 1)
```

CI for Normal Variance

`a = qchisq($\alpha/2$, df = n - 1)`

`b = qchisq($1 - \alpha/2$, df = n - 1)`

$$P \left[a \leq \left(\frac{n-1}{\sigma^2} \right) S^2 \leq b \right] = 1 - \alpha$$

$$P \left[\frac{a}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{b}{(n-1)S^2} \right] = 1 - \alpha$$

$$P \left[\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a} \right] = 1 - \alpha$$

CI for Normal Variance

Suppose $X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$ and let:

$$a = \text{qchisq}(\alpha/2, \text{df} = n - 1)$$

$$b = \text{qchisq}(1 - \alpha/2, \text{df} = n - 1)$$

Then,

$$\left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a} \right]$$

is a $100 \times (1 - \alpha)\%$ confidence interval for σ^2 .

End of Detour

We want to know the Sampling Distribution of $\sqrt{n}(\bar{X}_n - \mu)/S$ and we just saw that:

$$\left(\frac{n-1}{\sigma^2} \right) S^2 \sim \chi^2(n-1)$$

How can we use this fact to help us?

What is the Sampling Distribution of $\sqrt{n}(\bar{X}_n - \mu)/S$?

This slide is just algebra:

$$\begin{aligned}\frac{\bar{X}_n - \mu}{S/\sqrt{n}} &= \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \cdot \left(\frac{\sigma/\sqrt{n}}{\sigma/\sqrt{n}} \right) = \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right) \left(\frac{\sigma/\sqrt{n}}{S/\sqrt{n}} \right) \\&= \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right) \left(\frac{\sigma}{S} \right) = \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right) \left(\sqrt{\frac{n-1}{n-1}} \cdot \sqrt{\frac{\sigma^2}{S^2}} \right) \\&= \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right) \left(\sqrt{\frac{(n-1)\sigma^2}{(n-1)S^2}} \right) \\&= \frac{\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right)}{\sqrt{\left[\frac{(n-1)S^2}{\sigma^2} \right] / (n-1)}}\end{aligned}$$

Distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$



Suppose $X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$ and \bar{X}_n is the sample mean.

Then the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is

- (a) $t(n)$
- (b) $t(n - 1)$
- (c) $\chi^2(n)$
- (d) $\chi^2(n - 1)$
- (e) $N(\mu, \sigma^2)$
- (f) $N(0, 1)$
- (g) $N(\mu, \sigma^2/n)$
- (h) $F(n, n - 1)$

Distribution of $(n - 1)S^2/\sigma^2$



Suppose $X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$ and S^2 is the sample variance.

Then the sampling distribution of $(n - 1)S^2/\sigma^2$ is

- (a) $t(n)$
- (b) $t(n - 1)$
- (c) $\chi^2(n)$
- (d) $\chi^2(n - 1)$
- (e) $N(\mu, \sigma^2)$
- (f) $N(0, 1)$
- (g) $N(\mu, \sigma^2/n)$
- (h) $F(n, n - 1)$

What is the Sampling Distribution?



Suppose $Z \sim N(0, 1)$ independent of $Y \sim \chi^2(n - 1)$. Then the sampling distribution of $Z / \sqrt{Y / (n - 1)}$ is

- (a) $t(n)$
- (b) $t(n - 1)$
- (c) $\chi^2(n)$
- (d) $\chi^2(n - 1)$
- (e) $N(\mu, \sigma^2)$
- (f) $N(0, 1)$
- (g) $N(\mu, \sigma^2/n)$
- (h) $F(n, n - 1)$

What is the Sampling Distribution of $\sqrt{n}(\bar{X}_n - \mu)/S$?

From three slides back:

$$\begin{aligned}\frac{\bar{X}_n - \mu}{S/\sqrt{n}} &= \frac{\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\left[\frac{(n-1)S^2}{\sigma^2}\right]/(n-1)}} \\ &= \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \\ &\sim t(n-1)\end{aligned}$$

Strictly speaking, need to show that numerator and denominator are independent, but you can take my word for it!

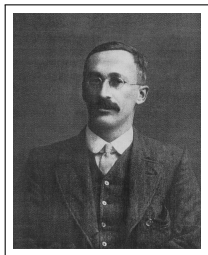
Punchline: Sampling Distribution of $\sqrt{n}(\bar{X}_n - \mu)/S$

If $X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$, then

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

Who was “Student?”

“Guinnessometrics: The Economic Foundation of Student's t”



“Student” is the pseudonym used in 19 of 21 published articles by William Sealy Gosset, who was a chemist, brewer, inventor, and self-trained statistician, agronomer, and designer of experiments ... [Gosset] worked his entire adult life ... as an experimental brewer for one employer: Arthur Guinness, Son & Company, Ltd., Dublin, St. James's Gate. Gosset was a master brewer and rose in fact to the top of the top of the brewing industry: Head Brewer of Guinness.

Three Key Sampling Distributions

Suppose that $X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. Then:

$$\left(\frac{n-1}{\sigma^2} \right) S^2 \sim \chi^2(n-1)$$

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

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

CI for Mean of Normal Distribution, Popn. Var. Unknown

Same argument as we used when the variance was known, except with $t(n - 1)$ rather than standard normal distribution:

$$P\left(-c \leq \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \leq c\right) = 1 - \alpha$$

$$P\left(\bar{X}_n - c\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X}_n + c\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

$$c = \text{qt}(1 - \alpha/2, \text{df} = n - 1)$$

$$\boxed{\bar{X}_n \pm \text{qt}(1 - \alpha/2, \text{df} = n - 1) \frac{S}{\sqrt{n}}}$$

Comparison of CIs for Mean of Normal Distribution

$100 \times (1 - \alpha)\%$ Confidence Level

$$X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

Known Population Std. Dev. (σ)

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \frac{\sigma}{\sqrt{n}}$$

Unknown Population Std. Dev. (σ)

$$\bar{X}_n \pm \text{qt}(1 - \alpha/2, \text{df} = n - 1) \frac{S}{\sqrt{n}}$$

Standard Error vs. Estimator of Standard Error

Standard Error

Recall that the standard deviation of the sampling distribution of an estimator is called the *standard error* (SE) of that estimator.

Example: Standard Error of the Mean

$$SE(\bar{X}_n) = \sqrt{\text{Var}(\bar{X}_n)} = \sigma/\sqrt{n}$$

Estimator of Standard Error of the Mean

Whereas σ/\sqrt{n} *is* the standard error of the mean, S/\sqrt{n} is an *estimator* of this quantity: $\widehat{SE}(\bar{X}_n) = S/\sqrt{n}$

Writing the CIs in terms of Actual and Estimated SE

$100 \times (1 - \alpha)\%$ Confidence Level

$$X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

Known Population Std. Dev. (σ)

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \textcolor{red}{SE}(\bar{X}_n)$$

Unknown Population Std. Dev. (σ)

$$\bar{X}_n \pm \text{qt}(1 - \alpha/2, \text{df} = n - 1) \textcolor{red}{\widehat{SE}}(\bar{X}_n)$$

Comparison of Normal and t CIs

Table: Values of $qt(1 - \alpha/2, df = n - 1)$ for various choices of n and α .

n	1	5	10	30	100	∞
$\alpha = 0.10$	6.31	2.02	1.81	1.70	1.66	1.64
$\alpha = 0.05$	12.71	2.57	2.23	2.04	1.98	1.96
$\alpha = 0.01$	63.66	4.03	3.17	2.75	2.63	2.58

Recall that as $n \rightarrow \infty$, $t(n - 1) \rightarrow N(0, 1)$

In a sense, using the t -distribution involves making a “small-sample correction.” In other words, it is only when n is fairly small that this makes a practical difference for our confidence intervals.

Am I Taller Than The Average American Male?

Source: Centers for Disease Control (pg. 16)

Sample Mean	69 inches
Sample Std. Dev.	6 inches
Sample Size	5647
My Height	73 inches

$$\begin{aligned}\widehat{SE}(\bar{X}_n) &= s/\sqrt{n} \\ &= 6/\sqrt{5647} \\ &\approx 0.08\end{aligned}$$

Assuming the population is normal,

$$\bar{X}_n \pm qt(1 - \alpha/2, df = n - 1) \widehat{SE}(\bar{X}_n)$$

What is the approximate value of
 $qt(1-0.05/2, df = 5646)$?

For large n , $t(n - 1) \approx N(0, 1)$, so the
answer is approximately 2

What is the ME for the 95% CI?

$$ME \approx 0.16 \implies 69 \pm 0.16$$