Econ 103 – Statistics for Economists

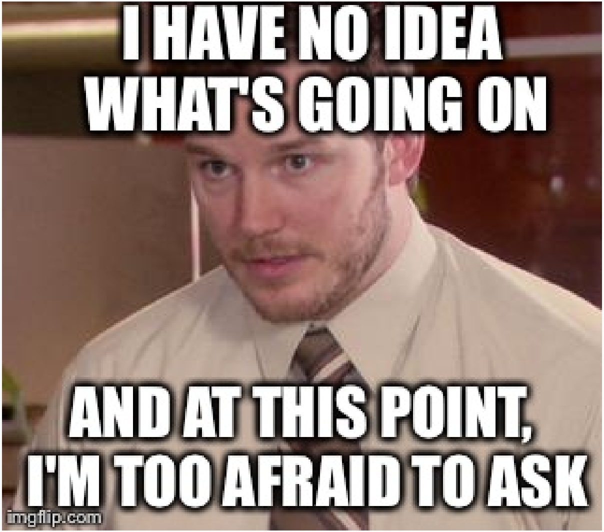
Chapter 6 and 7: Confidence Intervals

Mallick Hossain

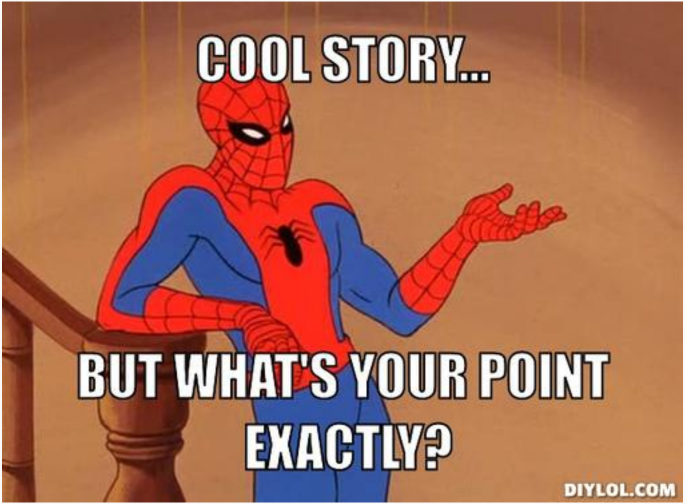
University of Pennsylvania

3 Students

Lost Larry


I HAVE NO IDEA
WHAT'S GOING ON
AND AT THIS POINT,
I'M TOO AFRAID TO ASK
imgflip.com

Smart Spider-Man



Synthesizing Squirrel

What’s the Point?

The goal is to get you closer to the squirrel (or at least Spider-Man)

Recap and Motivation

What We’ve Done So Far (Theory Side)

- We spent the past few weeks covering discrete and continuous random variables
    - You should be very comfortable with each random variable and their associated properties (see random variable handout for a nice (not necessarily exhaustive) summary)
- We dug into the normal distribution and all of its nice properties
    - The more intuitive the normal RV feels, the easier the rest of the semester will be
- Briefly introduced chi-squared, t-, and F-distributions
    - You’ll see why they are so important today! The wait is over!

What We've Done So Far (Practical Side)

- Random Sampling: $X_1, \ldots, X_n \sim$ iid
- Use estimator $\widehat{\theta}$ to learn about population parameter θ_0
- Estimator $\widehat{\theta}$ is a random variable:
    - Distribution of $\widehat{\theta}$ is called *sampling distribution*
    - Bias of an estimator
    - Variance of an estimator
    - Mean-squared Error (MSE) of an estimator
    - Consistency of an Estimator

The Road Ahead

Confidence Intervals

What values of θ_0 are consistent with the data we observed?

Hypothesis Testing

I think that $\theta_0 = 0$. Should I change my mind based on the data?

Motivation

- Do we expect point estimates to be exactly right?
    - No! As we saw last lecture, our estimate is basically a draw from the distribution of a random variable
- If we predicted that the S&P 500 would close at $2150.00 on Monday and it closed at $2150.88, my point estimate was wrong. Does that mean it's worthless though?
    - No! It was "close" which can be very informative!
    - Confidence intervals are instrumental in giving us a better idea of what counts as "close."

Example

(Above?) Average Joe

Joe is 73 inches tall. Based on a sample of US males aged 20 and over, the Centers for Disease Control (CDC) reported a mean height of about 69 inches in a recent report.

Clearly Joe is taller than the average American male!
Do you agree or disagree?

(a) Agree
(b) Disagree
(c) Not Sure

Remember: The Sample Mean is Random!

Just because the sample mean is 69 inches it doesn't follow that the population mean is 69 inches!

What Else Should We Consider?

- How big was the sample?
  - If the sample was very small there's a higher chance that it won't be representative of the population as a whole
  - Why? The variance of the sample mean is *decreasing with sample size* so bigger samples are less noisy.
- How much variability is there in height in the population?
  - If everyone is very similar in height, any sample we take will be representative of the population.
  - Remember: the variance of the sample mean is *increasing* with the population standard deviation.

Am I Taller Than The Average American Male?

Table 1: Height in inches for Males aged 20 and over (approximate)

Sample Mean	69 inches
Sample Std. Dev.	6 inches
Sample Size	5647
Joe's Height	73 inches

We'll return to this example later.

Theoretical Example

For Now – Single Population, Normally Distributed

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)}$$

Later we'll look at more than one population and talk about what happens if Normality doesn't hold.

Suppose $X_1, X_2, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$. What is the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$?

(a) $N(\mu, \sigma^2)$

(b) $N(0, 1)$

(c) $N(0, \sigma)$

(d) $N(\mu, 1)$

(e) Not enough information to determine.

Z-score!

Suppose $X_1, X_2, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$. From above,

$$\begin{aligned} E[\bar{X}_n] &= \mu \\ Var(\bar{X}_n) &= \sigma^2/n \\ &\Rightarrow SD(\bar{X}_n) = \sigma/\sqrt{n} \end{aligned}$$

Thus,

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

Remember that we call the standard deviation of a sampling distribution the standard error, written SE, so

$$\frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \sim N(0,1)$$

Standard Error vs Standard Deviation

- Standard Deviation
  - The square root of the variance
  - Measures the deviation from the mean
- Standard Error
  - A specific kind of standard deviation
  - This is the standard deviation of the estimator
  - For example, if we are estimating the population mean, the standard error tells us how far our estimate is from the actual population mean.

Suppose $X_1, X_2, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$. What is the approximate value of the following?

$$P\left(-2 \leq \frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq 2\right) \approx 0.95$$

What happens if I rearrange?

$$\begin{aligned}
P\left(-2 \leq \frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq 2\right) &= 0.95 \\
P\left(-2 \cdot SE \leq \bar{X}_n - \mu \leq 2 \cdot SE\right) &= 0.95 \\
P\left(-2 \cdot SE - \bar{X}_n \leq -\mu \leq 2 \cdot SE - \bar{X}_n\right) &= 0.95 \\
P\left(\bar{X}_n - 2 \cdot SE \leq \mu \leq \bar{X}_n + 2 \cdot SE\right) &= 0.95
\end{aligned}$$

Confidence Intervals

Confidence Intervals

Confidence Interval (CI)

A confidence interval is a range (A, B) constructed from the sample data that has a specified probability of containing a population parameter:

$$P(A \leq \theta_0 \leq B) = 1 - \alpha$$

Confidence Level

The specified probability, typically denoted $1 - \alpha$, is called the confidence level. For example, if $\alpha = 0.05$ then the confidence level is 0.95 or 95%.

Confidence Interval for Mean of Normal Population

Confidence Interval for Mean of Normal Population

The interval $\boxed{\bar{X}_n \pm 2\sigma/\sqrt{n}}$ has approximately 95% probability of containing the population mean μ, provided that:

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)}$$

But What Does This Mean?

Which quantities are random?

Suppose $X_1, X_2, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$. Which quantities are random variables?

(a) μ only

(b) σ and μ

(c) σ only

(d) σ, μ and $\bar{X}_n$

(e) $\bar{X}_n$ only

What does this mean for our confidence intervals?

Confidence Interval is a Random Variable!

1. $X_1, \ldots, X_n$ are RVs $\Rightarrow \bar{X}_n$ is a RV (repeated sampling)
2. μ, σ and n are constants
3. Confidence Interval $\bar{X}_n \pm 2\sigma/\sqrt{n}$ is also a RV!

Meaning of Confidence Interval

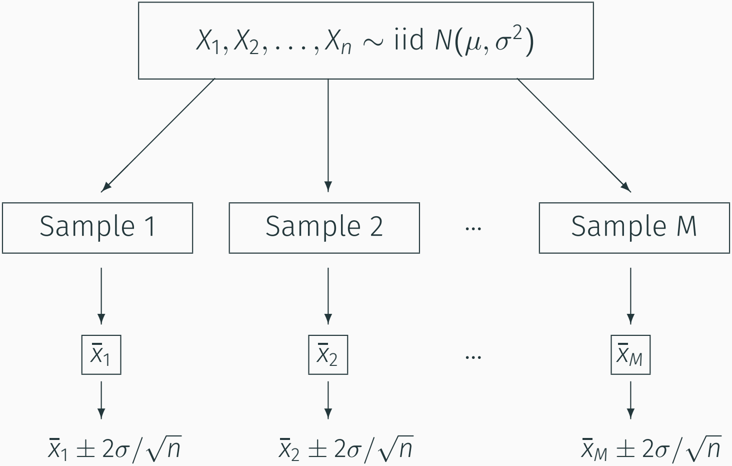
Meaning of Confidence Interval

If we sampled many times we'd get many different sample means, each leading to a different confidence interval. Approximately 95% of these intervals will contain μ.

Rough Intuition

What values of μ are consistent with the data?

CI for Population Mean: Repeated Sampling



$X_1, X_2, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)$

Sample 1 ... Sample 2 ... Sample M

$\bar{x}_1$ ... $\bar{x}_2$ ... $\bar{x}_M$

$\bar{x}_1 \pm 2\sigma/\sqrt{n}$ ... $\bar{x}_2 \pm 2\sigma/\sqrt{n}$ ... $\bar{x}_M \pm 2\sigma/\sqrt{n}$

Repeat M times $\rightarrow$ get M different intervals

Large M $\Rightarrow$ Approx. 95% of these Intervals Contain μ

Simulation Example: $X_1, \ldots, X_5 \sim$ iid $N(0,1)$, $M = 20$

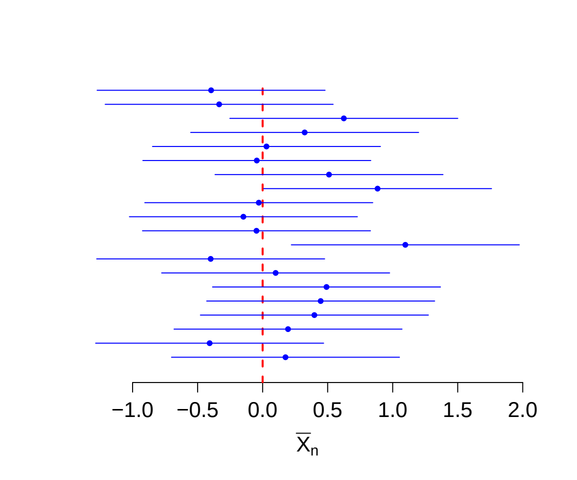


Figure 1: Twenty confidence intervals of the form $\bar{X}_n \pm 2\sigma/\sqrt{n}$ where $n = 5$, $\sigma^2 = 1$ and the true population mean is 0.

Meaning of Confidence Interval for θ_0

$$\boxed{P(A \leq \theta_0 \leq B) = 1 - \alpha}$$

Each time we sample we'll get a different confidence interval, corresponding to different realizations of the random variables A and B. If we sample many times, approximately $100 \times (1 - \alpha)\%$ of these intervals will contain the population parameter θ_0.

True or False?

Suppose

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)}$$

Then the population mean μ has approximately a 95% chance of falling in the interval $\bar{X}_n \pm 2\sigma/\sqrt{n}$.

(a) True

(b) False

FALSE! – μ is a constant!

Confidence Intervals: Some Terminology

Margin of Error

When a CI takes the form $\widehat{\theta} \pm ME$, ME is the Margin of Error.

Lower and Upper Confidence Limits

The lower endpoint of a CI is the lower confidence limit (LCL), while the upper endpoint is the upper confidence limit (UCL).

Width of a Confidence Interval

The distance $|\mathrm{UCL} - \mathrm{LCL}|$ is called the width of a CI. This means exactly what it says.

Margin of Error

What is the Margin of Error

In the preceding example of a 95% confidence interval for the mean of a normal population when the population variance is known, which of these is the margin of error?

(a) $\sigma/\sqrt{n}$

(b) $\bar{X}_n$

(c) σ

(d) $2\sigma/\sqrt{n}$

(e) $1/\sqrt{n}$

$2\sigma/\sqrt{n}$, since the CI is $\bar{X}_n \pm 2\sigma/\sqrt{n}$

What is the Width?

In the preceding example of a 95% confidence interval for the mean of a normal population when the population variance is known, which of these is the width of the interval?

(a) $\sigma/\sqrt{n}$

(b) $2\sigma/\sqrt{n}$

(c) $3\sigma/\sqrt{n}$

(d) $4\sigma/\sqrt{n}$

(e) $5\sigma/\sqrt{n}$

$4\sigma/\sqrt{n}$, since the CI is $\bar{X}_n \pm 2\sigma/\sqrt{n}$

Example: Calculate the Margin of Error

$X_1, \dots, X_{100} \sim$ iid $N(\mu, 1)$ but we don't know μ.
Want to create a 95% confidence interval for μ.

What is the margin of error?

The confidence interval is $\bar{X}_n \pm 2\sigma/\sqrt{n}$ so

$$ME = 2\sigma/\sqrt{n} = 2 \cdot 1/\sqrt{100} = 2/10 = 0.2$$

Example: Calculate the Lower Confidence Limit

$X_1, \ldots, X_{100} \sim N(\mu, 1)$ but we don't know μ. Want to create a 95% confidence interval for μ.

We found that $ME = 0.2$. The sample mean $\bar{x} = 4.9$. What is the lower confidence limit?

$$\text{LCL} = \bar{x} - ME = 4.9 - 0.2 = 4.7$$

Example: Similarly for the Upper Confidence Limit...

$X_1, \ldots, X_{100} \sim N(\mu, 1)$ but we don't know μ. Want to create a 95% confidence interval for μ.

We found that $ME = 0.2$. The sample mean $\bar{x} = 4.9$. What is the upper confidence limit?

$$\text{UCL} = \bar{x} + ME = 4.9 + 0.2 = 5.1$$

Example: 95% CI for Normal Mean, Popn. Var. Known

$$\boxed{X_1, \ldots, X_{100} \sim N(\mu, 1) \text{ but we don't know } \mu.}$$

95% CI for $\mu = [4.7, 5.1]$

What values of μ are plausible?

The data actually came from a $N(5, 1)$ Distribution.

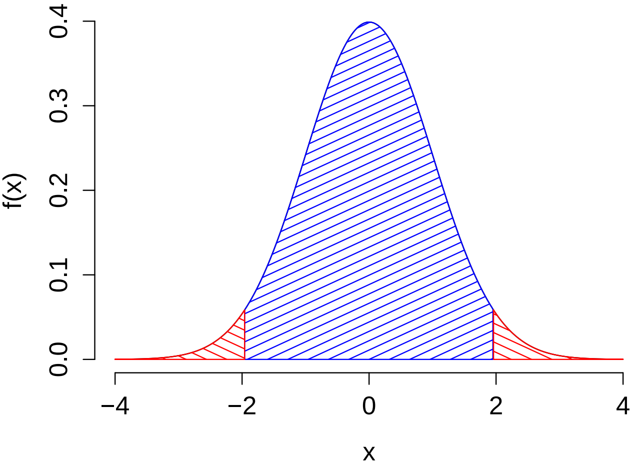
Want to be more certain? Use higher confidence level.

What value of c should we use to get a $100 \times (1-\alpha)\%$ CI for μ?

$$\begin{aligned} P\left(-c \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le c\right) &= 1-\alpha \\ P\left(\bar{X}_n - c\sigma/\sqrt{n} \le \mu \le \bar{X}_n + c\sigma/\sqrt{n}\right) &= 1-\alpha \end{aligned}$$

Take $c = \texttt{qnorm}(1-\alpha/2)$

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


0.0
0.1
0.2
0.3
0.4
f(x)
−4
−2
0
2
4
x

Confidence Interval for a Normal Mean, σ Known

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

What Affects the Margin of Error?

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

Sample Size n

ME decreases with n: bigger sample $\implies$ tighter interval

Population Std. Dev. σ

ME increases with σ: more variable population $\implies$ wider interval

Confidence Level $1-\alpha$

ME increases with $1-\alpha$: higher conf. level $\implies$ wider interval

Conf. Level	90%	95%	99%
α	0.1	0.05	0.01
$\texttt{qnorm}(1-\alpha/2)$	1.64	1.96	2.56

But What if σ is Unknown?

- What we've done so far assumed that σ was known.
- In real applications this is typically not the case.
- So what do we do now?

The Suspense!

We Don’t know σ. What to use instead?

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

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

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

First Step

What is the sampling distribution of S?

What is the Distribution?

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

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

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...

Detour

95% CI for Variance of Normal Population

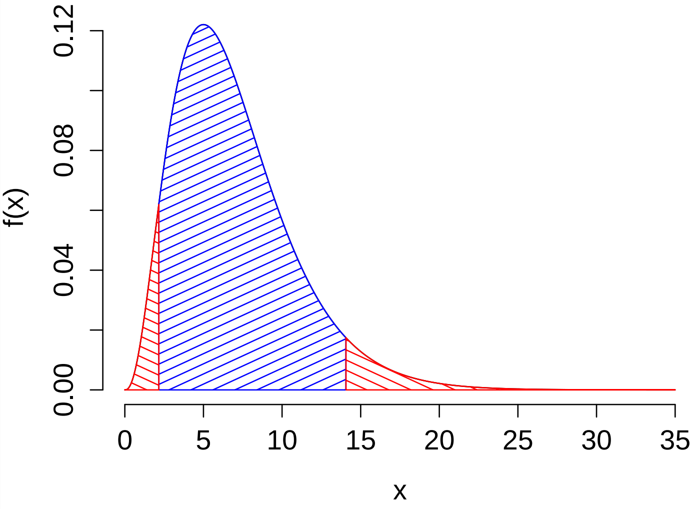
We know that:

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

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...


0.12
0.08
0.04
0.00
f(x)
0
5
10
15
20
25
30
35
x

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?

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

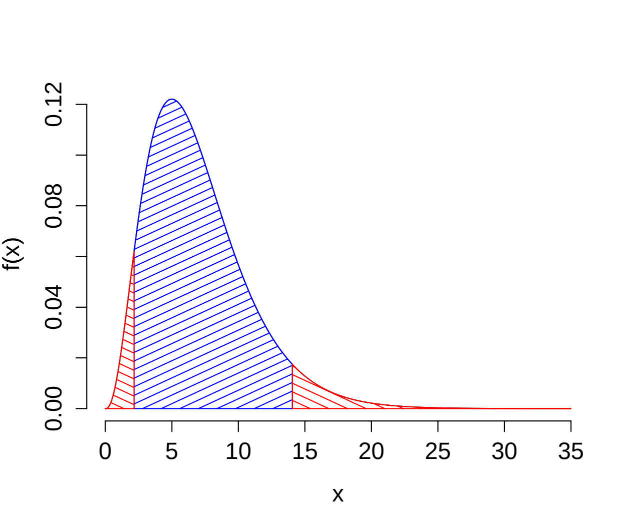
(a) `qchisq`$(0.95,$ `df =` $n-1)$

(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

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

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

Example: 95% Confidence Interval for Normal Variance

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

$$a = \texttt{qchisq(0.025, df = 99)} \approx 73$$
$$b = \texttt{qchisq(0.975, df = 99)} \approx 128$$

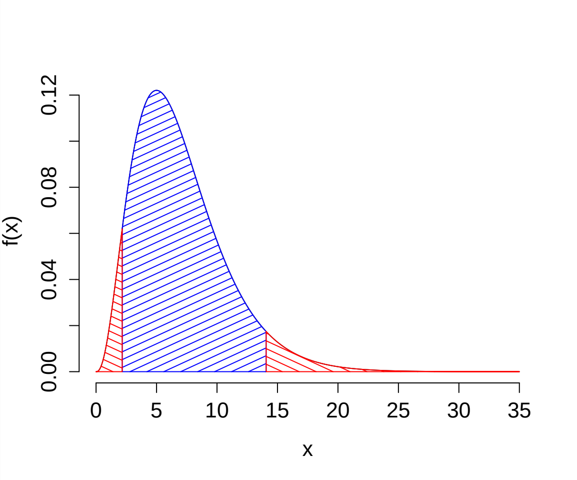
From the sample data, $s^2 = 4.3$

$$\begin{aligned} LCL &= (n-1)s^2/b = 99 \times 4.3/128 \approx 3.3 \\ UCL &= (n-1)s^2/a = 99 \times 4.3/73 \approx 5.8 \end{aligned}$$

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)`

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

CI for Normal Variance

Suppose $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$ and let:

$$\begin{aligned} a &= \texttt{qchisq(}\alpha/2\texttt{, df = n - 1)} \\ b &= \texttt{qchisq(}1-\alpha/2\texttt{, df = n - 1)} \end{aligned}$$

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:

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

How can we use this fact to help us?

Back on Track

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 $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$ and $\bar{X}_n$ is ths 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 $X_1, \ldots, X_n \sim$ 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:

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

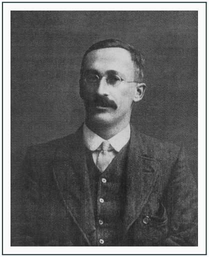
Strictly speaking, need to show that numerator and denominator are independent, but you can take my word for it!

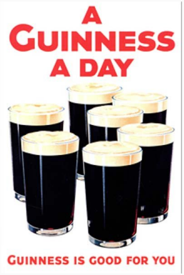
Punchline: Sampling Distribution of $\sqrt{n}(\bar{X}_n - \mu)/S$

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

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

Who was "Student?"





"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. Source

Three Key Sampling Distributions

Suppose that $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$. Then:

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

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:

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

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

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

Comparison of CIs for Mean of Normal Distribution

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

Known Population Std. Dev. (σ)

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

Unknown Population Std. Dev. (σ)

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

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

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

Estimator of Standard Error of the Mean

Whereas $\sigma/\sqrt{n}$ *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

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

Known Population Std. Dev. (σ)

$$\bar{X}_n \pm \texttt{qnorm}(1 - \alpha/2)\; SE(\bar{X}_n)$$

Unknown Population Std. Dev. (σ)

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

Comparison of Normal and t CIs

Table 2: Values of $\texttt{qt}(1-\alpha/2, \texttt{df} = n-1)$ 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 $n \to \infty$, $t(n-1) \to N(0,1)$

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.

Is Joe Taller Than The Average American Male?

Sample Mean	69 inches
Sample Std. Dev.	6 inches
Sample Size	5647
Joe's 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,

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

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?

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

Two-Sample Problem

Two-sample Problem

Suppose $X_1, \ldots, X_n \sim$ iid $N(\mu_x, \sigma_x^2)$ independently of $Y_1, \ldots, Y_m \sim$ iid $N(\mu_y, \sigma_y^2)$. What is $E[\bar{X}_n - \bar{Y}_m]$, the expectation of the sampling distribution of the difference of sample means?

(a) μ_x

(b) $\mu_x - \mu_y$

(c) μ_y

(d) $\mu_x + \mu_y$

(e) 0

$$E[\bar{X}_n - \bar{Y}_m] = E[\bar{X}_n] - E[\bar{Y}_m] = \mu_x - \mu_y$$

Two-sample Problem

Suppose $X_1, \ldots, X_n \sim$ iid $N(\mu_x, \sigma_x^2)$ independently of $Y_1, \ldots, Y_m \sim$ iid $N(\mu_y, \sigma_y^2)$. What is $Var[\bar{X}_n - \bar{Y}_m]$, the variance of the sampling distribution of the difference of sample means?

(a) $\sigma_x^2 - \sigma_y^2$

(b) $\sigma_x^2 + \sigma_y^2$

(c) $\sigma_x^2/n + \sigma_y^2/m$

(d) $\sigma_x^2/n - \sigma_y^2/m$

(e) 1

By independence: $Var[\bar{X}_n - \bar{Y}_m] = Var[\bar{X}_n] + Var[\bar{Y}_m] = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$

Two-sample Problem

Suppose $X_1, \dots, X_n \sim$ iid $N(\mu_x, \sigma_x^2)$ independently of $Y_1, \dots, Y_m \sim$ iid $N(\mu_y, \sigma_y^2)$. What is the sampling distribution of $\bar{X}_n - \bar{Y}_m$, the difference of sample means?

(a) χ^2

(b) t

(c) F

(d) Normal

Normal, by independence and linearity property of normal distributions.

Sampling Distribution of $\bar{X}_n - \bar{Y}_m$

Suppose $X_1, \ldots, X_n \sim$ iid $N(\mu_x, \sigma_x^2)$ independently of $Y_1, \ldots, Y_m \sim$ iid $N(\mu_y, \sigma_y^2)$. Then,

$$(\bar{X}_n - \bar{Y}_m) \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)$$

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0,1)$$

Shorthand: $SE(\bar{X}_n - \bar{Y}_m) = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$

CI for Difference of Population Means, σ_x^2, σ_y^2 Known

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_x - \mu_y)}{SE(\bar{X}_n - \bar{Y}_m)} \sim N(0,1)$$

Thus, we construct a $100 \times (1-\alpha)\%$ CI for $\mu_x - \mu_y$ as follows:

$$(\bar{X}_n - \bar{Y}_m) \pm \texttt{qnorm}(1-\alpha/2)\ SE(\bar{X}_n - \bar{Y}_m)$$

Where $SE(\bar{X}_n - \bar{Y}_m) = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$

Calculate the ME for the Difference of Means

I generated independent random samples of size 25 from two normal distributions in R. One had a population standard deviation of 4 and the other had a population standard deviation of 3. The sample means were approximately 4.2 and 3.1.

Calculate the ME for a 95% confidence interval for the difference of population means.

$$SE = \sqrt{\frac{3^2}{25} + \frac{4^2}{25}} = \frac{\sqrt{9+16}}{5} = 1$$

$$ME = \texttt{qnorm(1 - 0.05/2)} \times SE \approx 2 \times SE = 2$$

Calculate the LCL for the Difference of Means

I generated independent random samples of size 25 from two normal distributions in R. One had a population standard deviation of 4 and the other had a population standard deviation of 3. The sample means were approximately 4.2 and 3.1.

Calculate the LCL for a 95% confidence interval for the difference of population means.

$$LCL = (4.2 - 3.1) - ME = 1.1 - 2 = -0.9$$

Calculate the UCL for the Difference of Means

I generated independent random samples of size 25 from two normal distributions in R. One had a population standard deviation of 4 and the other had a population standard deviation of 3. The sample means were approximately 4.2 and 3.1.

Calculate the UCL for a 95% confidence interval for the difference of population means.

$$UCL = (4.2 - 3.1) + ME = 1.1 + 2 = 3.1$$

$$\boxed{\text{95\% Confidence Interval: } (-0.9, 3.1)}$$

The actual population means were 4 and 3, respectively

What if σ_x^2, σ_y^2 are Unknown?

Suppose $X_1, \ldots, X_n \sim$ iid $N(\mu_x, \sigma_x^2)$ independently of $Y_1, \ldots, Y_m \sim$ iid $N(\mu_y, \sigma_y^2)$. Then,

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \sim t(\nu)$$

Formula for ν is Complicated and You Don't Need to Know it

Two possibilities:

1. Have R find the correct value of ν for us
2. If m, n are large enough, approximately standard normal.

Case of Equal, Unknown Variances

The book considers a case where $\sigma_x^2 = \sigma_y^2 = \sigma^2$, that is a common unknown variance. This is a very dangerous assumption. It is almost certainly false and can throw off our results in a serious way. You are not responsible for this case.

Sampling Distributions Under Normality: One-sample

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

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

Sampling Distributions Under Normality: Two-sample

Suppose $X_1, \ldots, X_n \sim$ iid $N(\mu_x, \sigma_x^2)$ independently of $Y_1, \ldots, Y_m \sim$ iid $N(\mu_y, \sigma_y^2)$. Then:

$$\frac{(\bar{X}_n - \bar{Y}_n) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \quad \sim \quad N(0,1)$$

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \quad \sim \quad t(\nu)$$

But what if the population isn’t Normal?

The Central Limit Theorem

Suppose that $X_1, \ldots, X_n$ are a random sample from a population with unknown mean μ. Then, provided that n is *sufficiently large*, the sampling distribution of $\bar{X}_n$ is approximately $N\left(\mu, \widehat{SE}(\bar{X}_n)^2\right)$, even if the even if the underlying population is non-normal.

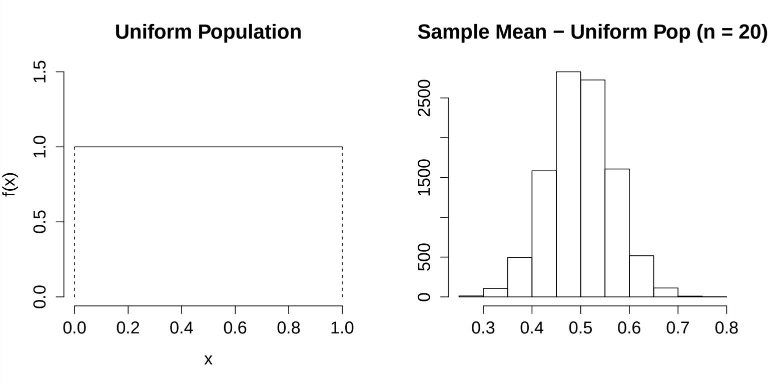
In Other Words...

$$\frac{\bar{X}_n - \mu}{\widehat{SE}(\bar{X}_n)} \approx N(0, 1)$$

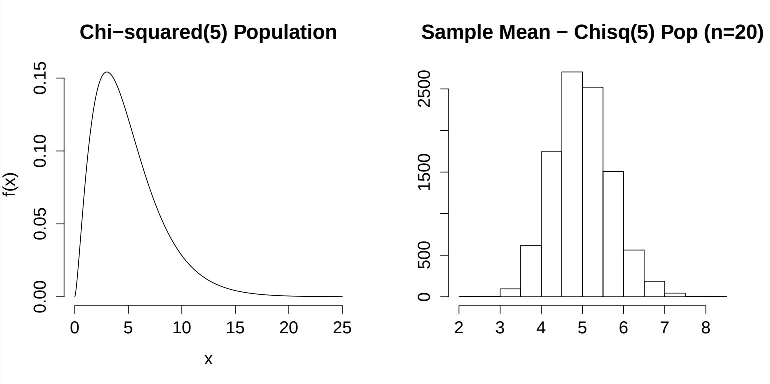
Use this to create *approximate* CIs for population mean!

You should be amazed by this.

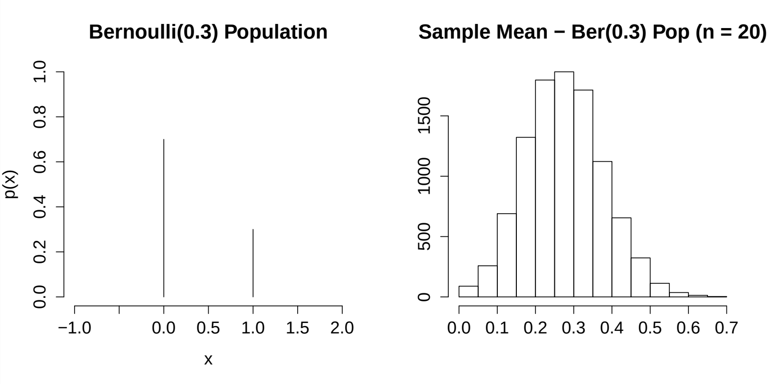
Example: Uniform(0,1) Population, $n = 20$



Example: $\chi^2(5)$ Population, $n = 20$



Example: Bernoulli(0.3) Population, $n = 20$



Sample Proportions

Who is the Chief Justice of the US Supreme Court?

(a) Harry Reid

(b) John Roberts

(c) William Rehnquist

(d) Stephen Breyer

(e) Antonin Scalia (*in absentia*)

Are US Voters Really That Ignorant?

The Data

Of 771 registered voters polled, only 39% correctly identified John Roberts as the current chief justice of the US Supreme Court.

Research Question

Is the majority of voters unaware that John Roberts is the current chief justice, or is this just sampling variation?

Assume Random Sampling...

Confidence Interval for a Proportion

What is the appropriate probability model for the sample?

$X_1, \ldots, X_n \sim$ iid Bernoulli(p), 1 = Know Roberts is Chief Justice

What is the parameter of interest?

p = Proportion of voters *in the population* who know Roberts is Chief Justice.

What is our estimator?

Sample Proportion: $\widehat{p} = (\sum_{i=1}^{n} X_i)/n$

Sample Proportion *is* the Sample Mean!

$X_1, \ldots, X_n \sim$ iid Bernoulli(p)

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

$$\begin{aligned}
E[\widehat{p}] &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{np}{n} = p \\
Var(\widehat{p}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \\
SE(\widehat{p}) &= \sqrt{Var(\widehat{p})} = \sqrt{\frac{p(1-p)}{n}} \\
\widehat{SE}(\widehat{p}) &= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}
\end{aligned}$$

Central Limit Theorem Applied to Sample Proportion

Central Limit Theorem: Intuition

Sample means are approximately normally distributed provided the sample size is large even if the population is non-normal.

CLT For Sample Mean

$$\frac{\bar{X}_n - \mu}{\widehat{SE}(\bar{X}_n)} \approx N(0,1)$$

CLT for Sample Proportion

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \approx N(0,1)$$

In this example, the population is Bernoulli(p) rather than normal. The sample mean is $\widehat{p}$ and the population mean is p.

Approximate 95% CI for Population Proportion

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \approx N(0,1)$$

$$\begin{aligned} P\left(-2 \leq \frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \leq 2\right) &\approx 0.95 \\ P\left(\widehat{p} - 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) &\approx 0.95 \end{aligned}$$

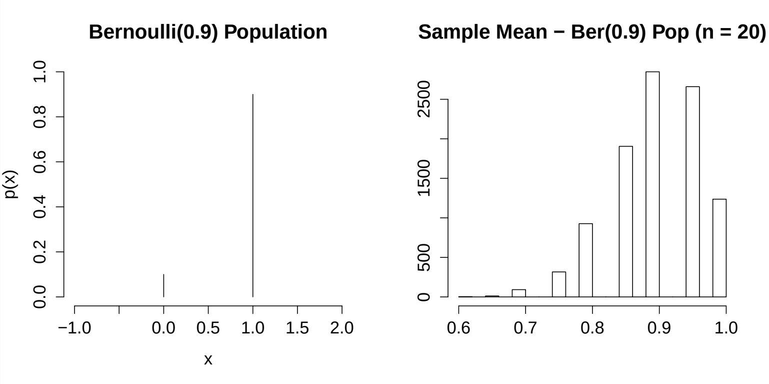
$100 \times (1-\alpha)$ CI for Population Proportion (p)

$X_1, \ldots, X_n \sim$ iid Bernoulli(p)

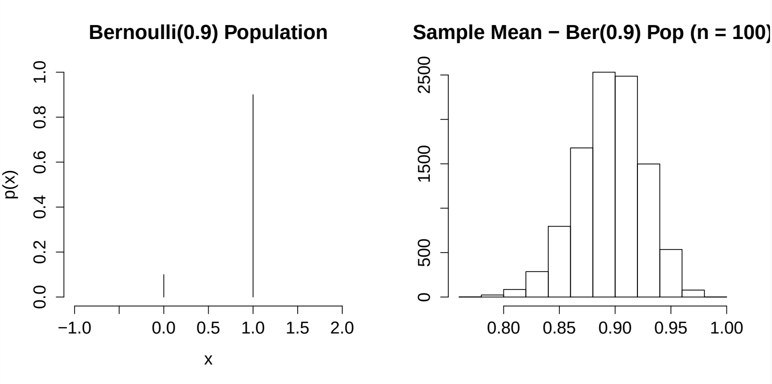
$$\widehat{p} \pm \texttt{qnorm}(1-\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Approximation based on the CLT. Works well provided n is large and p isn't too close to zero or one.

Example: Bernoulli(0.9) Population, $n = 20$



Example: Bernoulli(0.9) Population, $n = 100$



Approximate 95% CI for Population Proportion

39% of 771 Voters Polled Correctly Identified Chief Justice Roberts

$$\begin{aligned}\widehat{SE}(\widehat{p}) &= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \sqrt{\frac{(0.39)(0.61)}{771}} \\ &\approx 0.018\end{aligned}$$

What is the ME for an approximate 95% confidence interval?

$$ME \approx 2 \times \widehat{SE}(\bar{X}_n) \approx 0.04$$

What can we conclude?

Approximate 95% CI: (0.35, 0.43)

Are Republicans Better Informed Than Democrats?

Of the 239 Republicans surveyed, 47% correctly identified John Roberts as the current chief justice. Only 31% of the 238 Democrats surveyed correctly identified him. Is this difference meaningful or just sampling variation?

Again, assume random sampling.

Confidence Interval for a Difference of Proportions

What is the appropriate probability model for the sample?

$X_1, \ldots, X_n \sim$ iid Bernoulli(p) independently of
$Y_1, \ldots, Y_m \sim$ iid Bernoulli(q)

What is the parameter of interest?

The difference of population proportions $p - q$

What is our estimator?

The difference of sample proportions: $\widehat{p} - \widehat{q}$ where:

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \widehat{q} = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

Difference of Sample Proportions $\widehat{p} - \widehat{q}$ and the CLT

What We Have

Approx. sampling dist. for *individual* sample proportions from CLT: $\widehat{p} \approx N\left(p, \widehat{SE}(\widehat{p})^2\right), \quad \widehat{q} \approx N\left(q, \widehat{SE}(\widehat{q})^2\right)$

What We Want

Sampling Distribution of the *difference* $\widehat{p} - \widehat{q}$

Use Independence of the Two Samples

$$\widehat{p} - \widehat{q} \approx N\left(p - q, \widehat{SE}(\widehat{p})^2 + \widehat{SE}(\widehat{q})^2\right)$$

$$\implies \widehat{SE}(\widehat{p} - \widehat{q}) = \sqrt{\widehat{SE}(\widehat{p})^2 + \widehat{SE}(\widehat{q})^2} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{\widehat{q}(1-\widehat{q})}{m}}$$

Approx. 95% CI for Difference of Population Proportions

$$\frac{(\widehat{p}-\widehat{q})-(p-q)}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}+\frac{\widehat{q}(1-\widehat{q})}{m}}} \approx N(0,1)$$

$$P\left(-2 \leq \frac{(\widehat{p}-\widehat{q})-(p-q)}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}+\frac{\widehat{q}(1-\widehat{q})}{m}}} \leq 2\right) \approx 0.95$$

$$(\widehat{p}-\widehat{q}) \pm \texttt{qnorm}(1-\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}+\frac{\widehat{q}(1-\widehat{q})}{m}}$$

$100 \times (1-\alpha)$ CI for Diff. of Popn. Proportions $(p-q)$

$X_1, \ldots, X_n \sim$ iid Bernoulli(p) indep. $Y_1, \ldots, Y_n \sim$ iid Bernoulli(q)

$$(\widehat{p} - \widehat{q}) \pm \texttt{qnorm}(1-\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{\widehat{q}(1-\widehat{q})}{m}}$$

Approximation based on the CLT. Works well provided n, m large and p, q aren't too close to zero or one.

ME for approx. 95% for Difference of Proportions

47% of 239 Republicans vs. 31% of 238 Democrats identified Roberts

Republicans

$$\begin{aligned} \widehat{p} &= 0.47 \\ n &= 239 \\ \widehat{SE}(\widehat{p}) &= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \approx 0.032 \end{aligned}$$

Democrats

$$\begin{aligned} \widehat{q} &= 0.31 \\ m &= 238 \\ \widehat{SE}(\widehat{q}) &= \sqrt{\frac{\widehat{q}(1-\widehat{q})}{m}} \approx 0.030 \end{aligned}$$

Difference: (Republicans - Democrats)

$$\begin{aligned} \widehat{p} - \widehat{q} &= 0.47 - 0.31 = 0.16 \\ \widehat{SE}(\widehat{p} - \widehat{q}) &= \sqrt{\widehat{SE}(\widehat{p})^2 + \widehat{SE}(\widehat{q})^2} \approx 0.044 \implies ME \approx 0.09 \end{aligned}$$

Approximate 95% CI $(0.07, 0.25)$

What can we conclude?

Confidence Interval “Roundup”

1. Values near the middle of a CI are “more plausible.”
2. CI for Difference of Means using the CLT
3. Independent Samples versus Matched Pairs
4. Refined CIs for Population Proportion

Note that we are no longer assuming that the population is normal. Instead, we are constructing confidence intervals based on a large sample approximation using the CLT.

How Much Narrower is a 68% CI?

Suppose we're constructing an approximate confidence interval for a population mean using the CLT:

$$\bar{X}_n \pm \texttt{qnorm}(1 - \alpha/2) \times \widehat{SE}(\bar{X}_n)$$

Approximately what is the value of the *ratio* of the width of a 95% interval divided by the width of a 68% interval based on the above expression?

$$\boxed{\begin{array}{c} \texttt{qnorm}(1 - 0.05/2) \approx 2 \\ \texttt{qnorm}(1 - 0.32/2) \approx 1 \end{array}} \implies \frac{2 \times \texttt{qnorm}(1 - 0.05/2) \times \widehat{SE}(\bar{X}_n)}{2 \times \texttt{qnorm}(1 - 0.32/2) \times \widehat{SE}(\bar{X}_n)} \approx 2$$

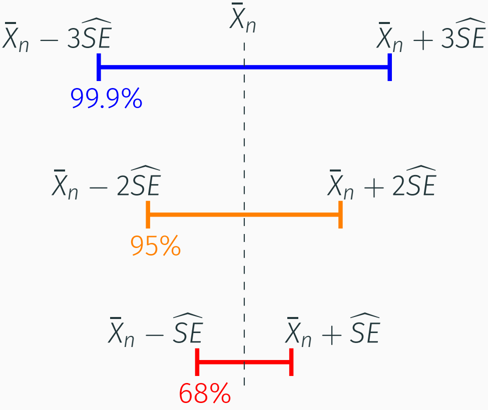

$\bar{X}_n$
$\bar{X}_n - 3\widehat{SE}$
$\bar{X}_n + 3\widehat{SE}$
99.9%
$\bar{X}_n - 2\widehat{SE}$
$\bar{X}_n + 2\widehat{SE}$
95%
$\bar{X}_n - \widehat{SE}$
$\bar{X}_n + \widehat{SE}$
68%

Figure 2: Each CI gives a range of "plausible" values for the population mean μ, centered at the sample mean $\bar{X}_n$. Values near the middle are "more plausible" in the sense that a small reduction in confidence level gives a much shorter interval centered in the same place. This is because the sample mean is unlikely to take on values far from the population mean in repeated sampling (CLT).

CI for Difference of Population Means Using CLT

Earlier

Used CLT to get CI for difference of population proportions based on independent samples.

But Proportions are a Kind of Mean!

Population proportion is mean of Bernoulli random variable, and sample proportion is mean of sample comprised of ones and zeros.

The general problem of constructing a CI for the difference of population means using the CLT is essentially identical to what we did earlier for population proportions.

CI for Difference of Population Means Using CLT

Setup: Independent Random Samples

$X_1, \ldots, X_n \sim$ iid with mean μ_X and variance σ_X^2
$Y_1, \ldots, Y_m \sim$ iid with mean μ_Y and variance σ_Y^2
where each sample is independent of the other

We Do Not Assume the Populations are Normal!

Difference of Sample Means $\bar{X}_n - \bar{Y}_m$ and the CLT

What We Have

Approx. sampling dist. for *individual* sample means from CLT:

$$\bar{X}_n \approx N\left(\mu_X, \widehat{SE}(\bar{X}_n)^2\right), \quad \bar{Y}_m \approx N\left(\mu_Y, \widehat{SE}(\bar{Y}_m)^2\right)$$

What We Want

Sampling Distribution of the *difference* $\bar{X}_n - \bar{Y}_m$

Use Independence of the Two Samples

$$\bar{X}_n - \bar{Y}_m \approx N\left(\mu_X - \mu_Y, \widehat{SE}(\bar{X}_n)^2 + \widehat{SE}(\bar{Y}_m)^2\right)$$

$$\implies \widehat{SE}(\bar{X}_n - \bar{Y}_m) = \sqrt{\widehat{SE}(\bar{X}_n)^2 + \widehat{SE}(\bar{Y}_m)^2} = \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$$

CI for Difference of Pop. Means (Independent Samples)

$X_1, \ldots, X_n \sim$ iid with mean μ_X and variance σ_X^2
$Y_1, \ldots, Y_m \sim$ iid with mean μ_Y and variance σ_Y^2
where each sample is independent of the other

$$(\bar{X}_n - \bar{Y}_m) \pm \texttt{qnorm}(1 - \alpha/2)\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$$

Approximation based on the CLT. Works well provided n, m large.

Which is the Harder Exam?

A previous semester's class had two midterms. Here are the scores:

Student	Exam 1	Exam 2	Difference
1	57.1	60.7	3.6
2	77.1	77.9	0.7
3	83.6	93.6	10.0
⋮	⋮	⋮	⋮
69	75.0	74.3	−0.7
70	96.4	86.4	−10.0
71	78.6	82.9	4.3
Sample Mean:	79.6	81.4	1.8

Was the second exam easier than the first?

What is the population model?

What does it mean to say that one exam is easier?

- Exam partly measures what you know and is partly random
  - You could have a bad day
  - The exam might focus on your weaker areas
- If a very large number of students take the exams, the randomness should *average out*.
- If a small number of students take the exams, they might score lower on the “easier exam” because of bad luck.

Are the two samples independent?

Suppose we treat the scores on the first midterm as one sample and the scores on the second as another. Are these samples independent?

(a) Yes

(b) No

(c) Not Sure

No – Each sample contains exactly the same students!

Dependent Samples – Same Students Took Each Exam

Student	Exam 1	Exam 2	Difference
1	57.1	60.7	3.6
⋮	⋮	⋮	⋮
71	78.6	82.9	4.3
Sample Mean:	79.6	81.4	1.8
Sample Corr.	0.54		

Table 3: The samples are dependent because each includes *exactly the same students*. Indeed, we see that scores on the two exams are strongly positively correlated: students who did well on the first exam tended to do well on the second.

What’s going on here?

We don’t really have two samples: we have a *single* sample of students, each of whom took two exams. This is really a *one sample problem* based on the *difference of individual exam scores*. Such a setup is sometimes referred to as matched pairs data

$$\boxed{\text{Let } D_i = X_i - Y_i \text{ be the difference of student } i\text{'s exam scores.}}$$

Solving this as a One-Sample Problem

$\boxed{\text{Let } D_i = X_i - Y_i \text{ be the difference of student } i\text{'s exam scores.}}$

I calculated the following in R:

$$\begin{aligned} \bar{D}_n &= \frac{1}{n}\sum_{i=1}^{n} D_i \approx 1.8 \\ S_D^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 \approx 124 \\ \widehat{SE}(\bar{D}_n) &= (S_D/\sqrt{n}) \approx \sqrt{124/71} \approx 1.3 \end{aligned}$$

Approximate 95% CI Based on the CLT:

$\boxed{1.8 \pm 2.6 = (-0.8, 4.4)}$ What is our conclusion?

How are the Independent Samples and Matched Pairs Problems Related?

Difference of Means = Mean of Differences?

$\boxed{\text{Let } D_i = X_i - Y_i \text{ be the difference of student } i\text{'s exam scores.}}$

True or False:

$$\bar{D}_n = \bar{X}_n - \bar{Y}_n$$

(a) True

(b) False

(c) Not Sure

Difference of Means Equals Mean of Differences

Let $D_i = X_i - Y_i$ be the difference of student i's exam scores.

Sample mean of differences *equals* difference of sample means

$$\begin{aligned}\bar{D}_n &= \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \bar{X}_n - \bar{Y}_n\end{aligned}$$

Linearity of Expectation holds even under dependence:

$$E[\bar{D}_n] = E[\bar{X}_n - \bar{Y}_n] = E[\bar{X}_n] - E[\bar{Y}_n] = \mu_X - \mu_Y$$

Exam Dataset

Student	Exam 1	Exam 2	Difference
1	57.1	60.7	3.6
⋮	⋮	⋮	⋮
71	78.6	82.9	4.3
Sample Mean:	79.6	81.4	1.8

$$\begin{aligned} \bar{D}_n &= 1.8 \\ \bar{X}_n - \bar{Y}_n &= 81.4 - 79.6 = 1.8 \checkmark \end{aligned}$$

…But Dependence Changes the Variance Calculation

Recall that for any two RVs X, Y and constants a, b

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

From the last slide, $\bar{D}_n = \bar{X}_n - \bar{Y}_n$, hence

$$\begin{aligned} Var(\bar{D}_n) &= Var(\bar{X}_n - \bar{Y}_n) \\ &= Var(\bar{X}_n) + Var(\bar{Y}_n) - 2Cov(\bar{X}_n, \bar{Y}_n) \\ &= \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n} - 2Cov(\bar{X}_n, \bar{Y}_n) \neq \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n} \end{aligned}$$

Since the samples are correlated, $Cov(\bar{X}_n, \bar{Y}_n) \neq 0$! Hence the standard error estimate for independent samples, $\sqrt{S_X^2/n + S_Y^2/n}$, is inappropriate!

Another Way to Calculate Sample Var. of the Differences

Variance of the differences can also be calculated from the sample variance for each exam along with the correlation between them:

$$
\begin{aligned}
S_D^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \bar{D}_n\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[(X_i - Y_i) - \left(\bar{X}_n - \bar{Y}_n\right)\right]^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(X_i - \bar{X}_n\right) - \left(Y_i - \bar{Y}_n\right)\right]^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(X_i - \bar{X}_n\right)^2 + \left(Y_i - \bar{Y}_n\right)^2 - 2\left(X_i - \bar{X}_n\right)\left(Y_i - \bar{Y}_n\right)\right] \\
&= S_X^2 + S_Y^2 - 2S_{XY} \\
&= S_X^2 + S_Y^2 - 2S_X S_Y r_{XY}
\end{aligned}
$$

$$\boxed{r_{XY} > 0 \implies S_D^2 < S_X^2 + S_Y^2}$$

Dependent Samples – Calculating the ME

Student	Exam 1	Exam 2	Difference
1	57.1	60.7	3.6
⋮	⋮	⋮	⋮
71	78.6	82.9	4.3
Sample Var.	117	151	$\boxed{?}$
Sample Corr.	0.54		

$$117 + 151 - 2 \times 0.54 \times \sqrt{117 \times 151} \approx 124 \checkmark$$

This agrees with what we got when we did calculations directly for the differences!

The “Wrong CI” (Assuming Independence) is Too Wide

Student	Exam 1	Exam 2	Difference
Sample Size	71	71	71
Sample Mean	79.6	81.4	1.8
Sample Var.	117	151	124
Sample Corr.	0.54		

Wrong Interval – Assumes Independence

$$1.8 \pm 2 \times \sqrt{117/71 + 151/71} \implies (-2.1, 5.7)$$

Correct Interval – Matched Pairs

$$1.8 \pm 2 \times \sqrt{124/71} \implies (-0.8, 4.4)$$

Top CI is too wide because the exam scores are positively correlated, so the variance of the differences is less than the sum of the variances of the two exams. Both CIs, however, are correctly centered.

Overview: Independent Samples Versus Matched Pairs

- When you see a problem that involves two datasets, think carefully about whether they should be treated as independent samples or if they’re really matched pairs. The CIs differ!
- The matched pairs calculations can be done in two ways:
    1. Direct calculation using the sample mean and standard deviation of the individual differences $D_i = X_i - Y_i$
    2. Indirect calculation using the sample mean and standard deviation of the X’s and Y’s *separately* along with the sample correlation between them.

Refined 95% CI for Proportion: “Add Two Successes and Failures”

Refined 95% CI for Population Proportion

Add four “fake” observations to the dataset: two zeros and two ones.

Textbook Confidence Interval

$$\widehat{p} = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)$$

$$\widehat{p} \pm \texttt{qnorm}(0.975) \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Refined Confidence Interval

$$\widetilde{p} = \frac{1}{n+4}\left(2 + \sum_{i=1}^{n} X_i\right)$$

$$\widetilde{p} \pm \texttt{qnorm}(0.975) \times \sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{n+4}}$$

This is related to problem 7-13 in the textbook...

Refined 95% CI for Difference of Population Proportions

Add four “fake” observations total: two to *each* dataset (a one and a zero).

Textbook Confidence Interval

$$\widehat{p} - \widehat{q} = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) - \frac{1}{m}\left(\sum_{i=1}^{n} Y_i\right)$$

$$(\widehat{p} - \widehat{q}) \pm \texttt{qnorm}(0.975) \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{\widehat{q}(1-\widehat{q})}{m}}$$

Refined Confidence Interval

$$p^* - q^* = \frac{1}{n+2}\left(1 + \sum_{i=1}^{n} X_i\right) - \frac{1}{m+2}\left(1 + \sum_{i=1}^{m} Y_i\right)$$

$$(p^* - q^*) \pm \texttt{qnorm}(0.975) \times \sqrt{\frac{p^*(1-p^*)}{n+2} + \frac{q^*(1-q^*)}{m+2}}$$

What is the point of this?

Recall from Last Time

Our CIs for proportions are *approximations* based on the CLT.

When is the approximation good?

Large sample size (n,m) and true population proportions (p,q) that aren't too close to zero or one.

Why the Refined Intervals?

They work well even when sample sizes (n, m) are small and true population proportions (p, q) are close to zero or one. When the samples are large, the refined intervals are practically identical to the textbook intervals.

Confidence Intervals We've Covered

1. Exact CIs based on assumption of Normality:
   (a) CI for population mean, population variance known (`qnorm`)
   (b) CI for population mean, population variance unknown (`qt`)
   (c) CI for population variance (`qhisq`)
   (d) CI for difference of population means, indep. samples (`qt`)
2. Approximate CIs using CLT (`qnorm`)
   (a) CI for population mean (also matched pairs data)
   (b) CI for difference of population means, independent samples
   (c) CIs for proportions and differences of proportions
      (i) "Textbook" version – use sample proportions
      (ii) "Refined" version – "add two successes and failures (total)"

In nearly all real applications we'll use 2, but you need to understand what's going on in 1 as well.