Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture 19

Confidence Intervals IV

Today: 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?

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

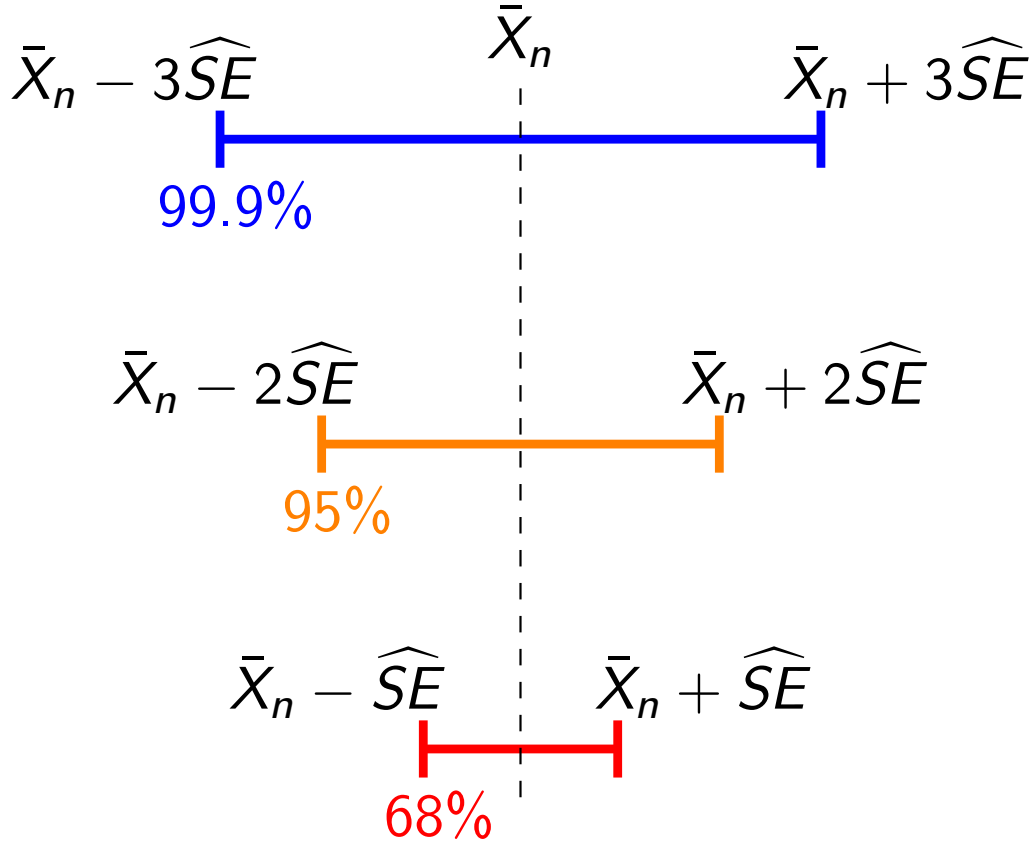


Figure : Each CI gives a range of "plausible" values for the population mean μ, centered at the sample mean $\bar{X}_n$. Values near the middle are "more plausible" in the sense that a small reduction in confidence level gives a much shorter interval centered in the same place. This is because the sample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

CI for Difference of Population Means Using CLT

Last Time

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 last time 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.

The Anchoring Experiment

At the beginning of the semester you were each shown a "random number." In fact the numbers weren't random: there was a "Hi" group that was shown 65 and a "Lo" group that was shown 10. You were randomly assigned to one of these two groups and shown your "random" number. You were then asked what proportion of UN member states are located in Africa. Let's take a look at the results...

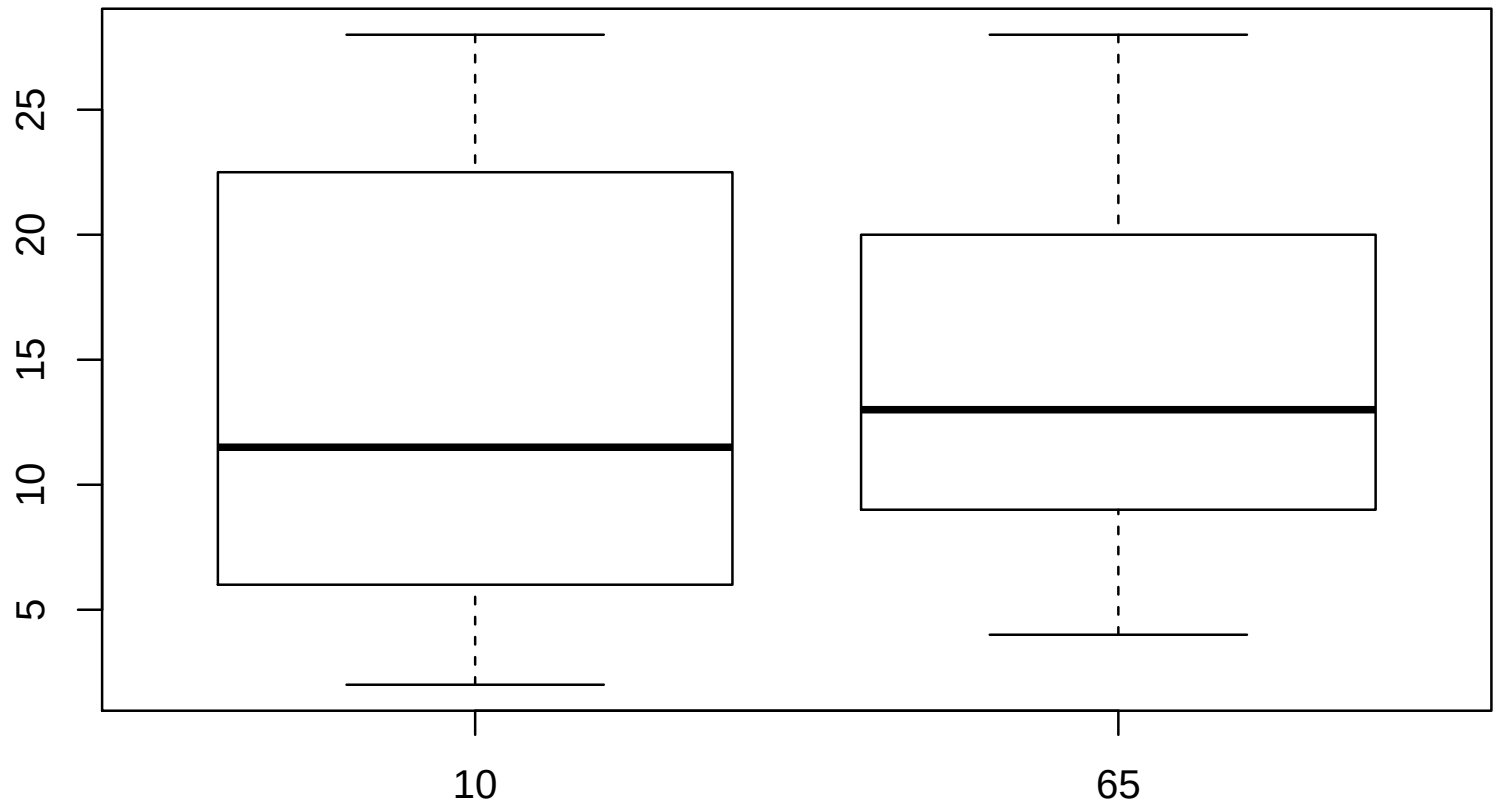
Load Data for Anchoring Experiment

```
data.url <- "http://www.ditraglia.com/econ103/survey_clean.csv"
survey <- read.csv(data.url)
anchoring <- survey[,c("rand.num", "africa.percent")]
head(anchoring)

##   rand.num africa.percent
## 1       65              5
## 2       10             21
## 3       65             13
## 4       65             17
## 5       65              5
## 6       65             14
```

Boxplot of Anchoring Experiment

```
boxplot(africa.percent ~ rand.num, data = anchoring)
```



Anchoring Experiment

From what population is our sample drawn?

US College Students? Penn Students? Penn Econ Majors?

Do We Have a Random Sample?

Definitely not a random sample of US College Students. Possibly a random sample of Penn Econ Majors since Econ 103 is required.

Observational or Experimental Data?

Randomized Experiment drew from a bag of "random" numbers

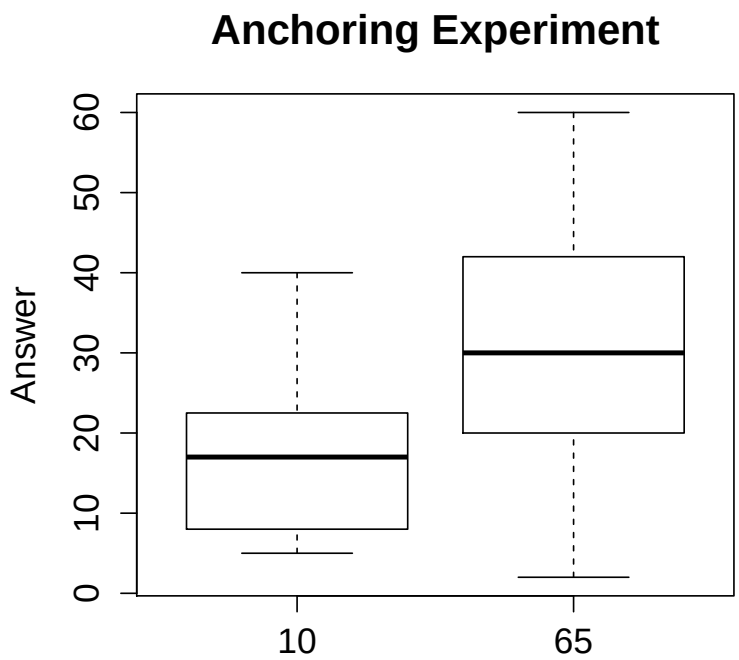
Are the two samples independent?

Yes: I told you not to show your number to any other students or consult with them in any way.

What is the Research Question?

Does "anchoring" cause of bias in decision-making?

Past Semester's Anchoring Experiment



"Lo" Group – Shown 10

$m_{Lo} = 43$

$\bar{y}_{Lo} = 17.1$

$s^2_{Lo} = 86$

"Hi" Group – Shown 65

$n_{Hi} = 46$

$\bar{x}_{Hi} = 30.7$

$s^2_{Hi} = 253$

ME for approx. 95% for Difference of Means

"Lo" Group

$$\begin{aligned} \bar{y}_{Lo} &= 17.1 \\ m_{Lo} &= 43 \\ s^2_{Lo} &= 86 \\ \widehat{SE}(\bar{y}_{Lo})^2 &= \frac{s^2_{Lo}}{m_{Lo}} = 2 \end{aligned}$$

"Hi" Group

$$\begin{aligned} \bar{x}_{Hi} &= 30.7 \\ n_{Hi} &= 46 \\ s^2_{Hi} &= 253 \\ \widehat{SE}(\bar{x}_{Hi})^2 &= \frac{s^2_{Hi}}{n_{Hi}} = 5.5 \end{aligned}$$

$$\begin{aligned} \bar{X}_{Hi} - \bar{Y}_{Lo} &= 30.7 - 17.1 = 13.6 \\ \widehat{SE}(\bar{X}_{Hi} - \bar{Y}_{Lo}) &= \sqrt{\widehat{SE}(\bar{X}_{Hi})^2 + \widehat{SE}(\bar{Y}_{Lo})^2} = \sqrt{7.5} \approx 2.7 \Rightarrow ME \approx 5.4 \end{aligned}$$

Approximate 95% CI $(8.2, 19)$

What can we conclude?

Which is the Harder Exam?

Last fall I gave 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 : 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

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

Solving this as a One-Sample Problem

Let $D_i = X_i - Y_i$ be the difference of student i'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:

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

How are the Independent Samples and Matched Pairs Problems Related?

Difference of Means = Mean of Differences?

Let $D_i = X_i - Y_i$ be the difference of student i'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:

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

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

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	?
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 CIs for Proportions:

"Add Two Successes and Failures"

Refined CI for Population Proportion

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

Textbook Confidence Interval

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

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

Refined Confidence Interval

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

$$\widetilde{p} \pm \texttt{qnorm}\left(1 - \frac{\alpha}{2}\right) \times \sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{n+4}}$$

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

Refined CI for Difference of Population Proportions

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

Textbook Confidence Interval

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

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

Refined Confidence Interval

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

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

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.