5. Sampling Distributions

LPO.8800: Statistical Methods in Education Research

Sean P. Corcoran

Last time

- Probability concepts (sample space, events, complements, $\cup$, $\cap$)
- Probability as relative frequency in the population
- Rules, e.g. $P(A \cup B)$ when A and B are mutually exclusive or not
- $P(A \cap B)$ when A and B are independent events
- Conditional probability, Bayes Rule
- Discrete probability distributions (e.g., Bernoulli, binomial)
- Continuous probability distributions (uniform, normal)
- Expectations and variance

Sampling distributions

- Probability distributions are used to describe the possible outcomes of a random variable, and their probabilities
- Statistics calculated from random draws from the population (such as the sample mean $\bar{x}$) **are also random variables**
- That is, they take on different values from one sample to the next—**sampling variation**

Sampling distributions

- As random variables, sample statistics (like $\bar{x}$) *also* have probability distributions that describe the relative frequency of its outcomes
- This probability distribution has central tendency (e.g. a mean or expected value) and variation (variance, standard deviation)
- The probability distribution of a sample statistic (like $\bar{x}$) is called a **sampling distribution**
- The sampling distribution is a *popluation* probability distribution since it characterizes all possible samples from a source population.

Sampling distributions: example

Consider a population for x consisting of 10 values:

$$0\ 1\ 3\ 3\ 5\ 7\ 7\ 7\ 8\ 10$$

- We can easily calculate the *population* mean $\mu = E(x) = 5.1$ and population standard deviation $\sigma = 3.08$
- Now imagine we don't observe this population, but take a random sample of $n = 2$ from it, with replacement, and calculate the sample mean ($\bar{x}$)

Sampling distributions: example

The realized value of $\bar{x}$ *depends on the sample drawn*:

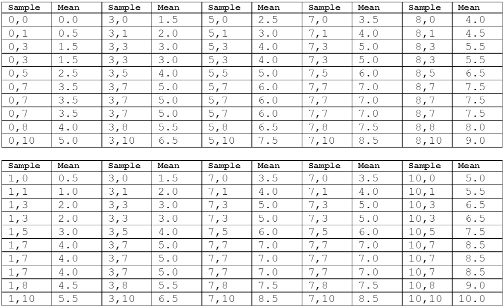
- (3, 3): $\bar{x} = 3$
- (1, 8): $\bar{x} = 4.5$
- (8, 10): $\bar{x} = 9$

Now imagine *all possible samples* of $n = 2$ one could draw, and their associated values of $\bar{x}$. (There are 100 such combinations).

This is relatively easy to do since the population is small (10 outcomes) and the sample size is small (n=2). This list gets dramatically larger with bigger populations and *n*s.

Sampling distributions: example

A complete list of all possible samples of size $n = 2$ and their $\bar{x}$:

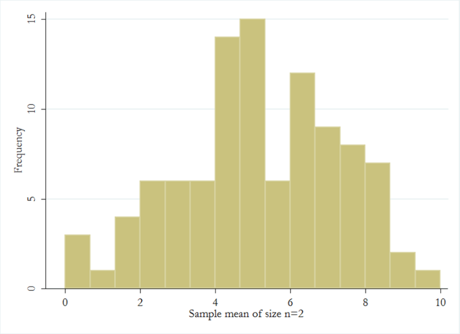


Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean
0,0	0.0	3,0	1.5	5,0	2.5	7,0	3.5	8,0	4.0
0,1	0.5	3,1	2.0	5,1	3.0	7,1	4.0	8,1	4.5
0,3	1.5	3,3	3.0	5,3	4.0	7,3	5.0	8,3	5.5
0,3	1.5	3,3	3.0	5,3	4.0	7,3	5.0	8,3	5.5
0,5	2.5	3,5	4.0	5,5	5.0	7,5	6.0	8,5	6.5
0,7	3.5	3,7	5.0	5,7	6.0	7,7	7.0	8,7	7.5
0,7	3.5	3,7	5.0	5,7	6.0	7,7	7.0	8,7	7.5
0,7	3.5	3,7	5.0	5,7	6.0	7,7	7.0	8,7	7.5
0,8	4.0	3,8	5.5	5,8	6.5	7,8	7.5	8,8	8.0
0,10	5.0	3,10	6.5	5,10	7.5	7,10	8.5	8,10	9.0

Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean
1,0	0.5	3,0	1.5	7,0	3.5	7,0	3.5	10,0	5.0
1,1	1.0	3,1	2.0	7,1	4.0	7,1	4.0	10,1	5.5
1,3	2.0	3,3	3.0	7,3	5.0	7,3	5.0	10,3	6.5
1,3	2.0	3,3	3.0	7,3	5.0	7,3	5.0	10,3	6.5
1,5	3.0	3,5	4.0	7,5	6.0	7,5	6.0	10,5	7.5
1,7	4.0	3,7	5.0	7,7	7.0	7,7	7.0	10,7	8.5
1,7	4.0	3,7	5.0	7,7	7.0	7,7	7.0	10,7	8.5
1,7	4.0	3,7	5.0	7,7	7.0	7,7	7.0	10,7	8.5
1,8	4.5	3,8	5.5	7,8	7.5	7,8	7.5	10,8	9.0
1,10	5.5	3,10	6.5	7,10	8.5	7,10	8.5	10,10	10.0

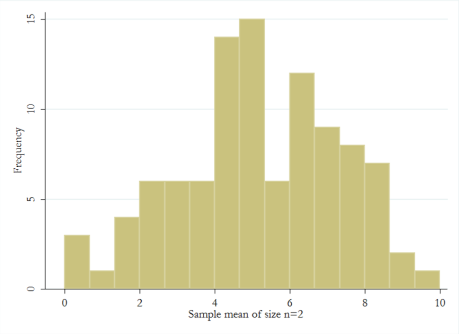
Sampling distributions: example

Plotting a histogram of these sample means shows the relative frequency (sampling distribution) of $\bar{x}$.



Sampling distributions: example

With a sample of $n = 2$, $\bar{x}$ ranges from 0 to 10, but is often between 1 and 9. The mean of these 100 $\bar{x}$'s is **5.1** and standard deviation **2.19**.



Sampling distributions: example

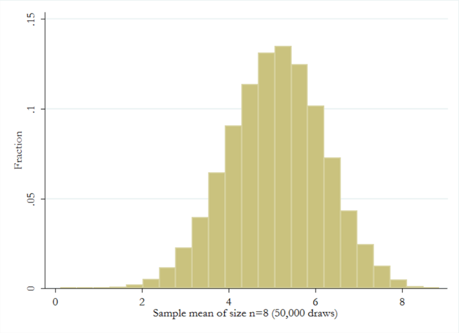
Now consider a simulation in which we *repeatedly* draw independent random samples of size $n = 8$ from the same population.

- Do this 50,000 times
- Calculate the sample mean each time
- Produce a histogram showing the relative frequency of these 50,000 sample means

Note: syntax for conducting simulations in Stata is covered at the end of this lecture.

Sampling distributions: example

With a sample of $n = 8$, $\bar{x}$ is often between 3 and 7. The mean of these 50,000 $\bar{x}$'s is **5.1** and standard deviation **1.09**.



Standard error

- These sampling distributions allow us to see the probabilities that our sample statistic will take on certain values

- Ex: with a sample size of $n = 8$, how likely is it that we obtain a sample mean of 6.1 or higher? (About 16%)

- The standard deviation of a sampling distribution is called the **standard error**

- The standard error is a *population* quantity. It is a measure of variation in the statistic across many (repeated) samples.

Standard error of $\bar{x}$

The **standard error** of $\bar{x}$ represents how much the mean of a random sample of size n deviates, on average, from the population mean, over repeated samples.

It is $\sqrt{Var(\bar{x})}$

Denote $se(\bar{x})$ or $\sigma_{\bar{x}}$

Mean, variance, and standard error of $\bar{x}$

The population mean, variance, and standard error of $\bar{x}$ are easy to find, applying some basic rules of expectations and variance:

- The expectation of a sum of random variables is the sum of the expectations: $E\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} E(x_i)$
- The variance of a sum of *independent* random variables is the sum of the variances: $Var\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} Var(x_i)$
- With *independent* random variables x_i: $Var\left(\sum_{i=1}^{n} ax_i\right) = \sum_{i=1}^{n} a^2 Var(x_i)$

Mean, variance, and standard error of $\bar{x}$

$\bar{x}$ is $\frac{1}{n}\sum_{i=1}^{n} x_i$, so:

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} n\mu = \mu$$

$$Var(\bar{x}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var(x_i) = \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}$$

$$se(\bar{x}) = \sqrt{Var(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

These apply no matter the underlying distribution of x.

Central limit theorem part 1

- The **central limit theorem** tells us something important about the sampling distribution of $\bar{x}$
- If x has a **normal** distribution with mean μ and standard deviation σ, i.e. $x \sim N(\mu, \sigma)$, then the sampling distribution of $\bar{x}$ is **normal** with a mean of μ and a standard error of $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
  - $E(\bar{x}) = \mu$ and $Var(\bar{x}) = \sigma^2/n$
- The notation $\sigma_{\bar{x}}$ is one way to write the standard error of $\bar{x}$. Another way is $se(\bar{x})$.
- Think about what happens as $n \to \infty$

Central limit theorem part 2

- Even if x does *not* have a normal distribution, with a large enough n, the sampling distribution of $\bar{x}$ is **approximately normal** with a mean of μ and a standard error of $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

- This is a remarkable and very useful result. It helps us easily "quantify our uncertainty" about $\bar{x}$, since we can determine from the normal distribution the likelihood of $\bar{x}$ taking on certain values.

- As a rule of thumb, $n \geq 30$ is often thought of as "large" for the sampling distribution of $\bar{x}$

Standard error of $\bar{x}$

There is sampling variation in $\bar{x}$ for two reasons:

- x itself varies in the population (σ). Different random samples of size n will produce different answers. Holding n constant, the more variability in the population (σ), the more sampling variation.

- The sample size. A smaller sample (n) will result in greater sampling variation than a larger one. Holding σ constant, the more data that is used, the more *precise* the estimate.

Put another way, the standard error $\sigma_{\bar{x}}$ increases with σ and decreases with n.

Central limit theorem: example

Consider a population: all births in the state of Maryland in 1994. The population size is 73,971

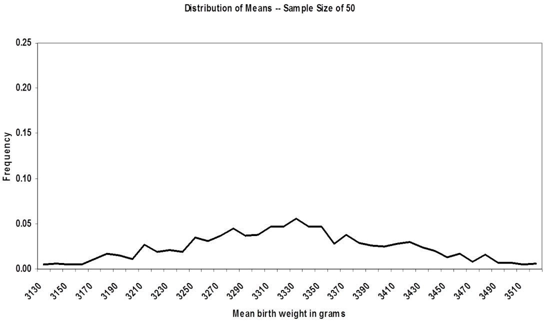
- The outcome of interest (x) is the birth weight of a newborn
- Suppose we know from the state census of births that the true population mean and standard deviation for birth weight in this year are: $\mu = 3320$g and $\sigma = 681.7$g
- Note: in the real world, we *don't know* the population parameters, but rather are trying to estimate them. For the sake of illustration, let's pretend we know this information!

Central limit theorem: example

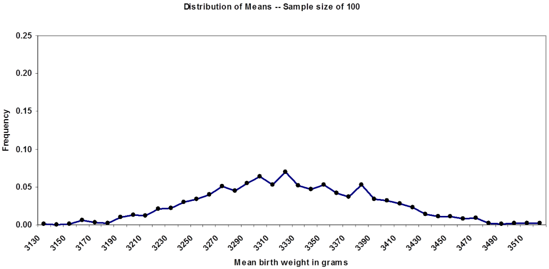
Now draw 1,000 samples of newborns at random, of size n:

- $n = 50$ births
- $n = 100$ births
- $n = 500$ births
- $n = 1000$ births
- $n = 2000$ births

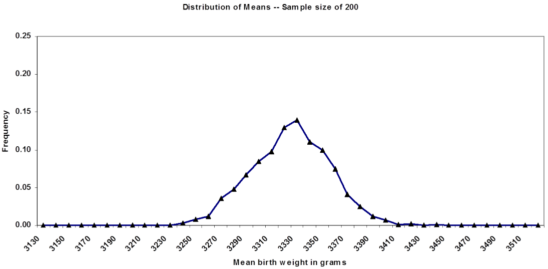
Central limit theorem: example



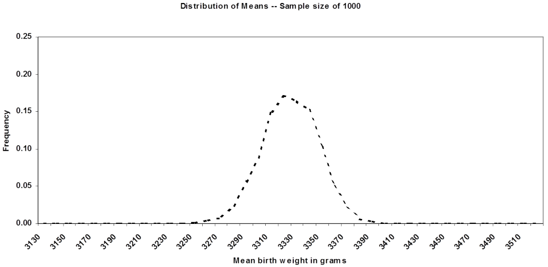
Central limit theorem: example



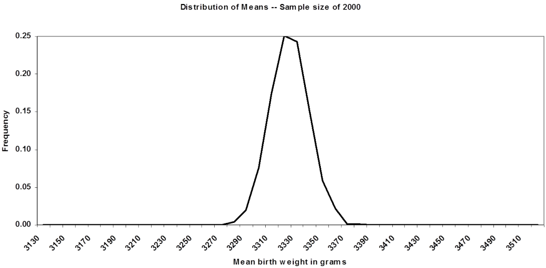
Central limit theorem: example



Central limit theorem: example

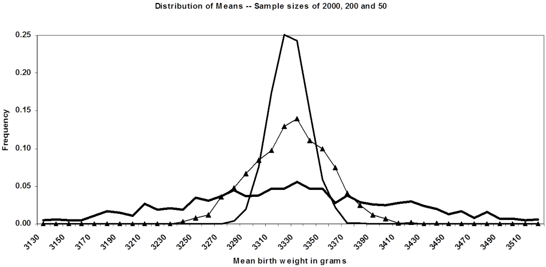


Central limit theorem: example



Distribution of Means -- Sample size of 2000

Central limit theorem: example



Distribution of Means -- Sample sizes of 2000, 200 and 50

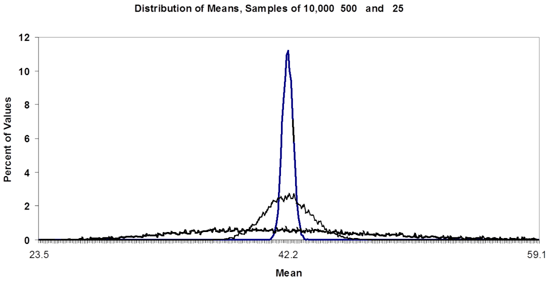
Central limit theorem: example

In this example where we know the population mean μ and σ^2:

- $E(\bar{x}) = \mu = 3320.$
- $Var(\bar{x}) = \sigma^2/n = 681.7^2/n$
- $se(\bar{x}) = \sqrt{Var(\bar{x})} = \sigma/\sqrt{n} = 681.7/\sqrt{n}$

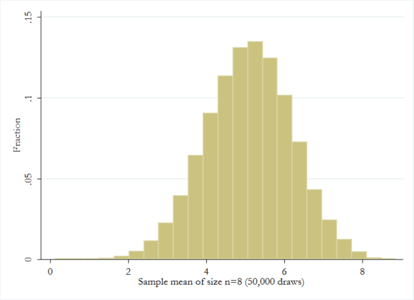
Central limit theorem: example

Another example using random samples of household income from the 1990 Census:



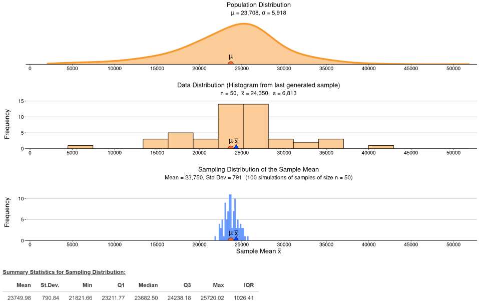
Central limit theorem: example

In the first example x was not even close to normally distributed, and n was only 8. Yet the sampling distribution of $\bar{x}$ looks very normal:



Central limit theorem: web apps

See the Github site for links to some great web apps for visualizing sampling distributions and the CLT. For example:


Population Distribution
μ = 23,708, σ = 5,918
μ
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
50000
Data Distribution (Histogram from last generated sample)
n = 50, x̄ = 24,350, s = 6,813
Frequency
15
10
5
0
μ x̄
Sampling Distribution of the Sample Mean
Mean = 23,750, Std Dev = 791 (100 simulations of samples of size n = 50)
Frequency
10
5
0
μ x̄
Sample Mean x̄

Summary Statistics for Sampling Distribution:

Mean	St.Dev.	Min	Q1	Median	Q3	Max	IQR
23749.98	790.84	21821.66	23211.77	23682.50	24238.18	25720.02	1026.41

Application 1

- We sample x from a *normally distributed population* in which $\mu = 15$ and $\sigma = 3$
- Our sample size is $n = 16$
- What will the sampling distribution of $\bar{x}$ look like?

Application 1

From the CLT we know that $\bar{x}$

- Has a **normal** distribution
- Has a mean of **15**
- Has a standard error of $\sigma/\sqrt{n} = 3/\sqrt{16} = \mathbf{0.75}$

Critically, we did *not* have to actually take repeated samples to know this!

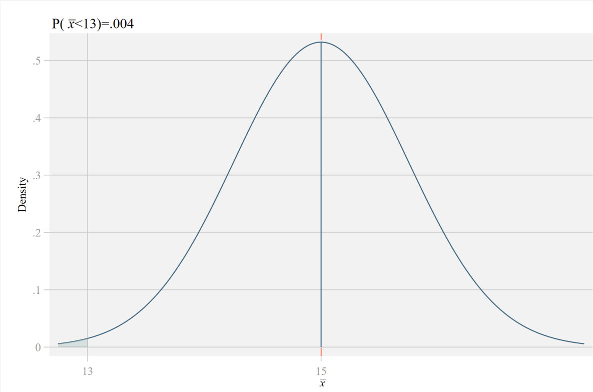
In practice, we don't know the specific values of μ or σ but we do know that $E(\bar{x}) = \mu$ and $se(\bar{x}) = \sigma/\sqrt{n}$, and that $\bar{x}$ has a normal distribution.

Application 2[a]

Using the same information from Application 1, what is the probability that a random sample of $n = 16$ is drawn with an $\bar{x}$ of **13 or less**?

- We know that $\bar{x} \sim N(15, 0.75)$
- So $z = \frac{\bar{x}-15}{0.75} \sim N(0,1)$
- $Pr(\bar{x} \leq 13) = Pr\left(\frac{\bar{x}-15}{0.75} \leq \frac{13-15}{0.75}\right) = Pr(z < -2.67) = 0.0038$
- Stata: `display normal(-2.67)`

Application 2[a]

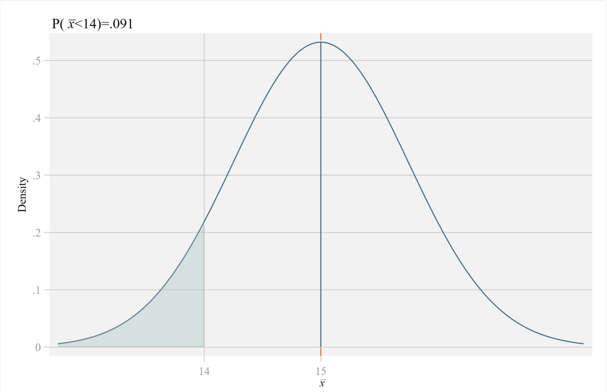


Application 2[b]

What is the probability that one would draw a random sample of $n = 16$ and obtain an $\bar{x}$ of **14 or less**?

- $Pr(\bar{x} \leq 14) = Pr\left(\frac{\bar{x}-15}{0.75} \leq \frac{14-15}{0.75}\right) = Pr(z < -1.33) = 0.0918$
- Stata: `display normal(-1.33)`

Application 2[b]

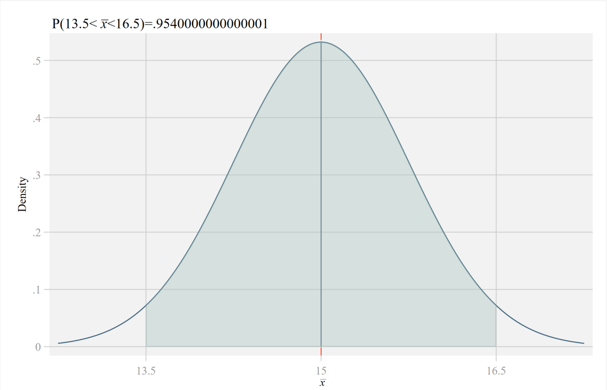


Application 2[c]

What is the probability that one would draw a random sample of $n = 16$ and obtain an $\bar{x}$ **between 13.5 and 16.5**?

- The z-values for 13.5 and 16.5 are:
- $\frac{(13.5-15)}{0.75} = -2$
- $\frac{(16.5-15)}{0.75} = +2$
- $Pr(-2 \leq z \leq 2) \approx 0.95$
- Stata: `display normal(2) - normal(-2)` $= 0.954$

Application 2[c]



Application 3

Use the CLT to determine the necessary sample size for a given amount of precision.

- Suppose we know that family income in 1997 averaged $32K with a standard deviation of $17K ($\mu$ and σ)
- We plan to survey families about their income, and do not want $\bar{x}$ to vary too much
- How many families do we have to survey so that $\bar{x}$ is within $4K of μ 95% of the time?

Application 3

- We know that family income $\mu = 32$ and $\sigma = 17$
- By the CLT, $\bar{x} \sim N(32, 17/\sqrt{n})$
- Find the n so that $Pr(28 \leq \bar{x} \leq 36) = 0.95$
- In other words, $\bar{x}$ can be expected to fall between 28 and 36 ($4K from μ) 95% of the time.

Application 3

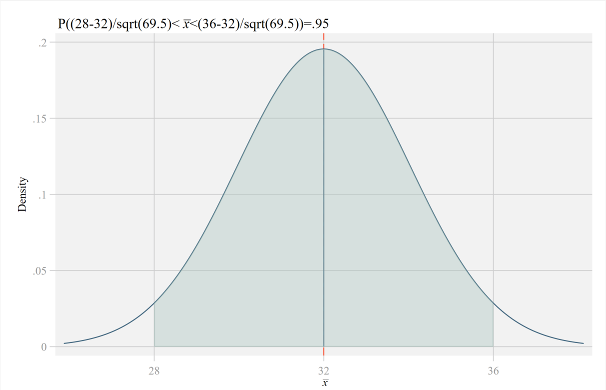
Expressing using z-scores:

$$P\left(\frac{28-32}{17/\sqrt{n}} < z < \frac{36-32}{17/\sqrt{n}}\right) = 0.95$$

We know the probability that z falls between -1.96 and 1.96 is 0.95. So:

$$\left(\frac{36-32}{17/\sqrt{n}} = 1.96\right) \text{ or } n = 69.4$$

Application 3



Estimators

Because $\bar{x}$ was used with a random sample to estimate μ, it is called an **estimator**.

- What makes $\bar{x}$ a *good* estimator of μ?
- Is $\bar{x}$ the *best* estimator of μ?

Estimators

- An **unbiased estimator** of a population parameter (like μ, for example) has a sampling distribution with a mean equal to the parameter being estimated.
- In other words, *in repeated samples* the average realized value of the estimator is the parameter value you're looking for

Estimators

- The CLT tells us that the mean of the sampling distribution for $\bar{x}$ is μ. So, $\bar{x}$ is an *unbiased* estimator of μ. $E(\bar{x}) = \mu$

- When asking whether an estimator is "biased" you are asking whether there is anything that causes it to *systematically* over or under-estimate the true population quantity.

- Critical: unbiased estimates are not necessarily *correct* in any specific sample! In fact, they probably won't be. They just use a *procedure* that will, on average, yield the correct estimate.

Estimators

- In many applications of statistics there are multiple estimators available for estimating a population parameter.

- How can we choose among these? Which estimator is the "best?"

- Unbiasedness is one good quality of an estimator.

- **Efficiency** is another. One estimator $\tilde{x}$ is more *efficient* than another $\bar{x}$ if the variance of the sampling distribution of $\tilde{x}$ is less than the variance of the sampling distribution of $\bar{x}$.
  - We've seen one example of this already: $\bar{x}$ estimated with $n = 100$ is more efficient than $\bar{x}$ estimated with $n = 10$.

Sampling distributions for proportions

Consider a Bernoulli random variable x. We know the mean of x represents the *proportion* of cases in which $x = 1$.

- $E(x) = \pi$, the proportion of cases in the *population* where $x = 1$. (π is just a special name for μ whenever x is dichotomous).
- $1 - \pi$ is the proportion of cases in the population where $x = 0$.
- $\hat{\pi}$ is the *sample* proportion of cases where $x = 1$.
- $\hat{\pi}$ is an estimator of π

Sampling distributions for proportions

When x is a Bernoulli variable, the population standard deviation of x is: $\sqrt{\pi(1-\pi)}$ (see Lecture 4). Applying the CLT:

- The sampling distribution of $\hat{\pi}$ is **approximately normal** when n is sufficiently large.
- This is a good example of the CLT when the underlying variable x does *not* have a normal distribution
- The mean of $\hat{\pi}$ will be π, and its standard error $\sqrt{\frac{\pi(1-\pi)}{n}}$
- $\hat{\pi} \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

Application 4

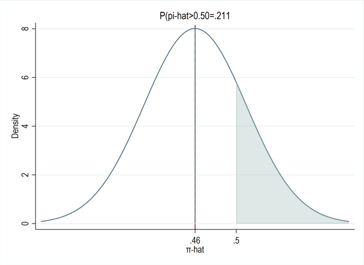
Suppose we know that 46% of the voting age population approves of the way Joe Biden is handling his job as president. If 100 adults are polled at random and asked whether or not they approve of Biden, the sample proportion in favor is $\hat{\pi}$.

- The sampling distribution of $\hat{\pi}$ is **approximately normal**
- With a mean of $\pi = 0.46$ (the population proportion)
- And a standard error of $\sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.46(1-0.46)}{100}} = 0.0498$
- Notice the notation $\sigma_{\hat{\pi}}$ for the standard error of the estimator $\hat{\pi}$

Application 4

What is the probability in a random sample of 100 adults that a *majority* approve of the way Joe Biden is handling his job as president?

- $Pr(\hat{\pi} > 0.50) = Pr(\frac{\hat{\pi}-0.46}{0.046} > \frac{0.50-0.46}{0.0498}) = Pr(z > 0.803)$
- In Stata: `1-normal(0.803)` $= 0.211$



Application 5

Continuing with the same example, how many adults would we need to sample so that $\hat{\pi}$ is within 2 points (0.02) of $\pi = 0.46$ 95% of the time?

- By the CLT, $\hat{\pi} \sim N\left(0.46, \sqrt{\frac{0.46(1-0.46)}{n}}\right)$
- Find the n so that $Pr(0.44 \leq \hat{\pi} \leq 0.48) = 0.95$
- In other words, $\hat{\pi}$ can be expected to fall between 0.44 and 0.48 95% of the time.

Application 5

Expressing using z-scores:

$$P\left(\frac{0.44-0.46}{\sqrt{\frac{0.46(1-0.46)}{n}}} < z < \frac{0.48-0.46}{\sqrt{\frac{0.46(1-0.46)}{n}}}\right) = 0.95$$

We know the probability that z falls between -1.96 and 1.96 is 0.95. So:

$$\left(\frac{0.48-0.46}{\sqrt{\frac{0.46(1-0.46)}{n}}} = 1.96\right) \text{or } n = 2,385$$

Sampling distributions

You may have noticed something unusual about the examples in this lecture:

- Describing the sampling distribution of $\bar{x}$ requires knowing what σ is in the population.
- Describing the sampling distribution of $\hat{\pi}$ requires knowing what π is in the population.
- *If we have this information about the population, why are we taking a sample to begin with?*
- Good question. When the population quantities are unknown, they can be estimated. The standard error we use in practice is an *estimated* standard error, not the population standard error.

Simulating sampling distributions in Stata

Monte Carlo simulation is the repeated execution of a random process. It is used to better understand the data generating process, solve probability problems, and evaluate estimators. They are especially useful for inspecting the sampling distribution of estimators.

Refer to the handout "Useful Stata commands for simulation" on Github.

- Loops
- `simulate` command
- Bootstrapping
- `postfile` and related commands

Again, it's good practice to `set seed` before doing simulations.

Simulating sampling distributions in Stata

Bootstrapping involves repeated sampling *with replacement*. Recall the example that drew 50,000 samples of size $n = 8$ from a population. The following code repeatedly samples 8 observations from the data, executes `summarize` on variable *x1*, and retains the result stored in `r(mean)`:

```
bootstrap r(mean), reps(50000) size(8) saving(draws50k,
replace):  summ x1
```

After 50,000 replications, the results are saved in a file called *draws50k*. `bootstrap` is a prefix that comes before the main Stata command (`summ` in this case).

Useful when sampling from existing data.

Simulating sampling distributions in Stata

`simulate` command example: first write a little program that draws a random sample and calculates statistics, then replicate it many times.

```
// Method 1 - program and simulate
clear
capture program drop app1

program app1, rclass
   version 15.1
   drop _all
   set obs 16
   gen x=rnormal(15, 3)
   summ x
   return scalar mean=r(mean)
   return scalar Var =r(Var)
end

set seed 4321
simulate mean=r(mean) var=r(Var), reps(1000) nodots: app1
summ mean var
```

Simulating sampling distributions in Stata

Same thing using `postfile` and related commands:

```
// Method 2 - postfile commands
// tempname is temporary holding place where results are held
// tempfile is temporary file for collecting results
clear
tempname results
tempfile meantable

// postfile tells Stata where results will be iteratively saved, which items
// will be saved, and where the final results will be collected
postfile ‘results’ xmean xvar using ‘meantable’

forvalues j=1/1000 {
   clear
   drawnorm x, n(16) mean(15) sds(3)
   quietly summ
   post ‘results’ (r(mean)) (r(Var))
   }
postclose ‘results’

use ‘meantable’, clear
summ xmean xvar
```