Economics 103 – Statistics for Economists

Minsu Chang

University of Pennsylvania

Lecture 17

Confidence Intervals – Part III

Writing the CIs in terms of Actual and Estimated SE

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

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

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: 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.

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

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,

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

What is the approximate value of

`qt(1-0.05/2, df = 5646)`?

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

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

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

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

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

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?

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

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

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

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

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

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

(a) χ^2

(b) t

(c) F

(d) Normal

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 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.

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.

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.

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.

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

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.

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

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

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.

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

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

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 $X_1, \ldots, X_n \sim$ 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 underlying population is non-normal.

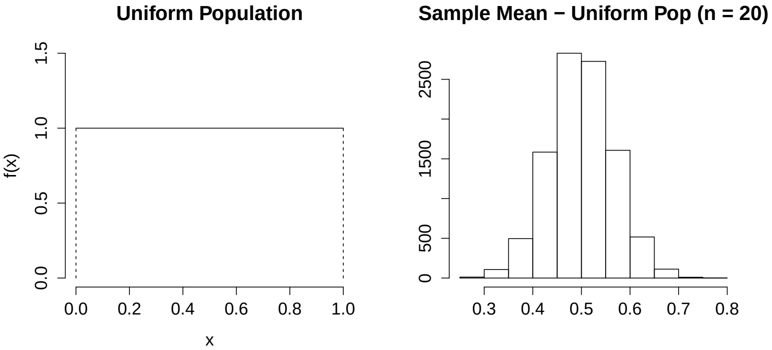
In Other Words...

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

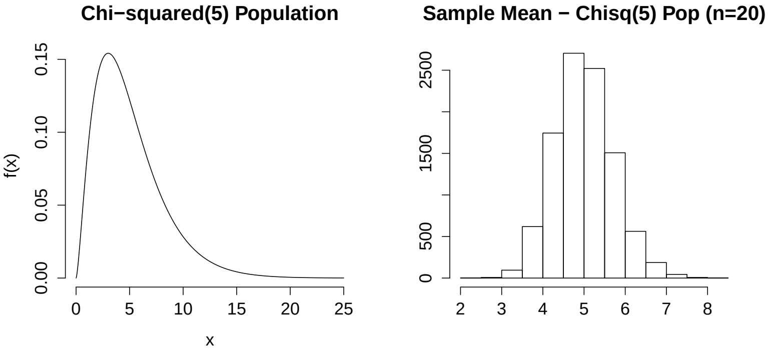
Use this to create *approximate* CIs for population mean!

You should be amazed by this.

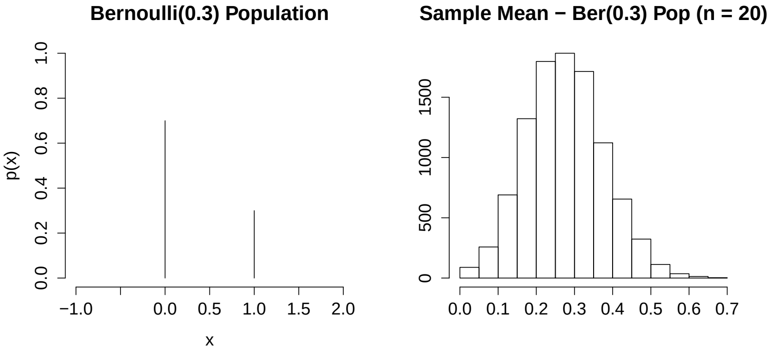
Example: Uniform(0,1) Population, $n = 20$



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



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



Who is the Chief Justice of the US Supreme Court?

(a) Harry Reid

(b) John Roberts

(c) William Rehnquist

(d) Stephen Breyer

Are US Voters Really That Ignorant?

Pew: "What Voters Know About Campaign 2012"

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)

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

Sample Proportion *is* the Sample Mean!

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

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

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

Sample Proportion *is* the Sample Mean!

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

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

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

Sample Proportion *is* the Sample Mean!

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

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

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

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

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

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

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

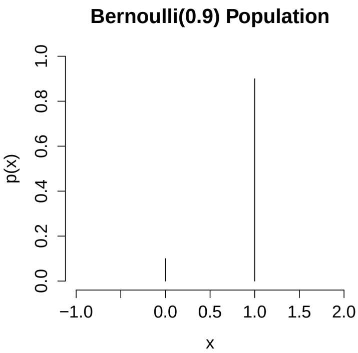
$100 \times (1 - \alpha)$ CI for Population Proportion (p)

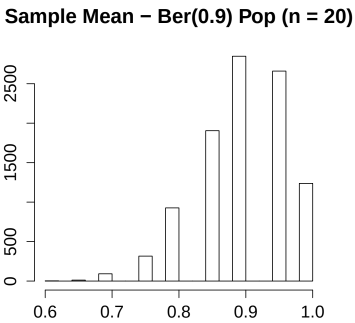
$X_1, \ldots, X_n \sim$ iid Bernoulli(p)

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

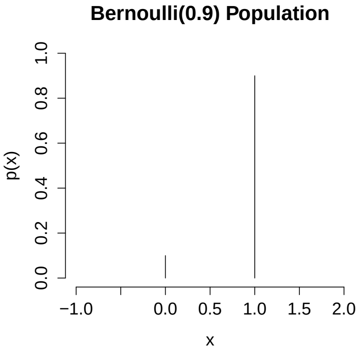
Approximation based on the CLT. Works well provided n is large and p isn't too close to zero or one.

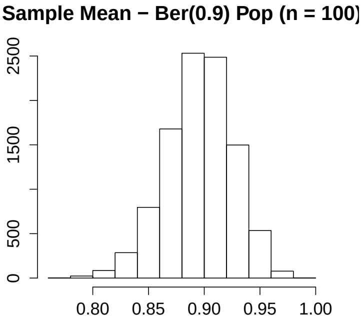
Example: Bernoulli(0.9) Population, $n = 20$





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?

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?

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

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?

$\boxed{\text{Approximate 95\% CI: } (0.35, 0.43)}$