Lecture 9: Central Limit Theorem

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Previously on STAT 324...

Confidence Intervals

Scenario 1:

The data (i.e. $X_1, \ldots, X_n$) are normally distributed, independent, and **we know** σ. Then

$$\bar{X} \sim N(\mu, \sigma^2/n) \qquad \text{and} \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

so

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

where $z_{\alpha/2}$ is the number such that $P(Z > z_{\alpha/2}) = \alpha/2$

Scenario 2:

The data (i.e. $X_1, \ldots, X_n$) are normally distributed, independent, but we **do not** know σ. Then

$$\bar{X} \sim N(\mu, \sigma^2/n) \qquad \text{and} \qquad \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

so

$$P\left(\bar{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

where $t_{n-1,\alpha/2}$ is the number such that $P(T_{n-1} > t_{n-1,\alpha/2}) = \alpha/2$

Confidence Intervals

```r
library(tidyverse); library(distributions3)
paint_thickness <- tibble(
  thickness = c(1.29, 1.12, 0.88, 1.65, 1.48, 1.59, 1.04, 0.83,
                1.76, 1.31, 0.88, 1.71, 1.83, 1.09, 1.62, 1.49))
```

Create 95% confidence interval WITHOUT assuming we know the true SD. Need mean, sd, and critical value.

Mean and SD:

```r
paint_thickness %>%
  summarize(mean = mean(thickness), sd = sd(thickness))
```

```
## # A tibble: 1 x 2
##    mean    sd
##   <dbl> <dbl>
## 1  1.35 0.339
```

Critical value (recall, df = n - 1):

```r
T <- StudentsT(df = 16-1)

(t_crit <- quantile(T, 0.975))
```

```
## [1] 2.13145
```

Confidence Intervals

We can find the confidence interval as

$$\bar{x} \pm t_{15,0.025} \cdot \frac{s}{\sqrt{16}}$$

```
paint_thickness %>%
  summarize(mean = mean(thickness),
            sd = sd(thickness),
            LL = mean - t_crit*sd/sqrt(n()),
            UL = mean + t_crit*sd/sqrt(n()))
```

```
## # A tibble: 1 x 4
##    mean    sd    LL    UL
##   <dbl> <dbl> <dbl> <dbl>
## 1  1.35 0.339  1.17  1.53
```

Confidence Intervals

Some vocabulary and intuition

Our *estimate* of the true mean paint thickness is 1.348 - also call this the *point estimate*.

The interval 1.168 to 1.529 is a 95% confidence interval - also call this an *interval estimate*.

We are 95% *confident* the true paint thickness is between 1.168 and 1.529.

Compare to 95% CI when knowing SD: 1.182 to 1.515. When SD unknown, CI larger.

Intuitively, we know less, so less confident.

What if we do not know if $X_1, \ldots, X_n$ are normally distributed?

Or maybe we know they are in fact NOT normally distributed. Then what?

Central Limit Theorem

If $X_1, \ldots, X_n$ are iid random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. For "large enough" n,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{(approximately)}$$

n "large enough" depends on the true distribution of X_i's. If "close to normal", smaller n needed.

Generally, $n \geq 30$ is a good rule of thumb for "large enough".

I, personally, find it easier to remember that $\bar{X} \sim N(E(\bar{X}), \text{Var}(\bar{X}))$. (Note: this is the same, since $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$.)

Confidence Intervals: Population Proportion

- An accounting firm has a large list of clients, each client has a file with information.
- Noticed some files contain errors
- What proportion of all files contain errors?

Parameter of interest: π = true proportion of files containing errors.

Don't want to go through all files, so take a simple random sample of size 50. Let $X_1, \ldots, X_{50}$ be the random variables denoting if the files have an error or not. If file number i has an error, $X_i = 1$. Otherwise, $X_i = 0$.

Distribution of X_i is $\text{Bernoulli}(\pi)$.

A good estimator of the true proportion of files with errors is $P = \frac{\sum_{i=1}^{50} X_i}{n}$ the sample proportion.

Confidence Intervals: Population Proportion

$E(P) = \pi$

$\text{Var}(P) = \frac{\pi(1-\pi)}{n}$

To find a CI for π, we need the distribution of P. Since $P = \frac{1}{n}\sum_{i=1}^{50} X_i$, P is an average, CLT tells us that, for n large enough, $P \sim N(E(P), \text{Var}(P))$, or equivalently $P \sim N(\pi, \pi(1-\pi)/n)$.

Since we do not know $\text{Var}(P)$, we will use $\widehat{\text{SD}}(P) = \sqrt{P(1-P)/n}$.

Confidence Intervals: Population Proportion

$P \sim N(\pi, \pi(1-\pi)/n)$. So what is the distribution of $\frac{P-\pi}{\sqrt{P(1-P)/n}}$?

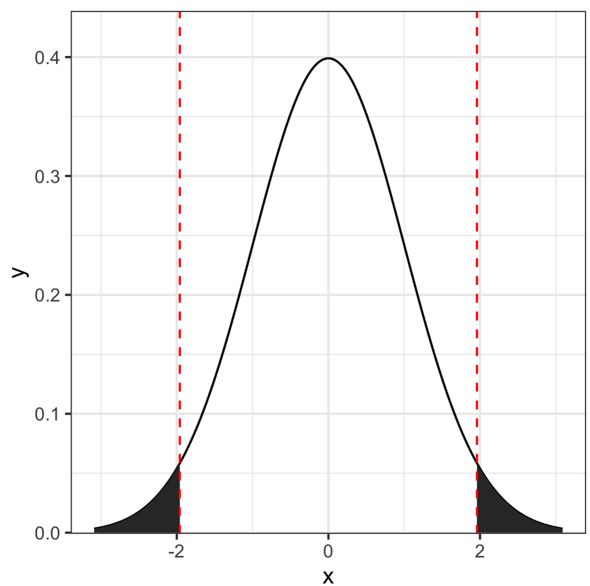
$$\frac{P-\pi}{\sqrt{P(1-P)/n}} = \frac{P-E(P)}{\widehat{\mathrm{SD}}(P)} = Z \sim N(0,1).$$

Why not t? Because estimating π with P gives us $\widehat{SD}(P)$ for free! No extra estimation required.

Now, we can find values x_1, x_2 such that $P(Z \leq x_1) + P(Z \geq x_2) = \alpha$. Let's for simplicity use $\alpha = 0.05$. I.e. we want to find x_1, x_2 such that this area is 0.05.

If we decide the two areas in the tails are the same, $x_1 = -x_2$.

x_2 is by definition the $\alpha/2$ (0.025 in this case) critical value, $z_{\alpha/2}$ - it cuts off $\alpha/2$ (0.025) to the right!



Confidence Intervals: Population Proportion

So,

$$1-\alpha = P(-z_{\alpha/2} \le Z \le z_{\alpha/2})$$

$$= P\left(-z_{\alpha/2} \le \frac{P-\pi}{\sqrt{P(1-P)/n}} \le z_{\alpha/2}\right)$$

$$= P\left(-z_{\alpha/2}\sqrt{P(1-P)/n} \le P-\pi \le z_{\alpha/2}\sqrt{P(1-P)/n}\right)$$

$$= P\left(-P-z_{\alpha/2}\sqrt{P(1-P)/n} \le -\pi \le -P+z_{\alpha/2}\sqrt{P(1-P)/n}\right)$$

$$= P\left(P+z_{\alpha/2}\sqrt{P(1-P)/n} \ge \pi \ge P-z_{\alpha/2}\sqrt{P(1-P)/n}\right)$$

$$= P\left(P-z_{\alpha/2}\sqrt{P(1-P)/n} \le \pi \le P+z_{\alpha/2}\sqrt{P(1-P)/n}\right)$$

Confidence Intervals: Population Proportion

So, a $(1-\alpha)\cdot 100\%$ Confidence Interval for the true population proportion π is $[P - z_{\alpha/2}\sqrt{P(1-P)/n}, P + z_{\alpha/2}\sqrt{P(1-P)/n}]$.

Confidence Intervals: Population Proportion

Say we observe 10 files with errors and 40 files without.

Our estimate would be $p = \frac{1}{50}\sum_{i=1}^{n} x_i = \frac{1}{50}(10 \cdot 1 + 40 \cdot 0) = 0.2$.

The estimated SD would be $\sqrt{p(1-p)/n} = \sqrt{0.2 \cdot 0.8/50} \approx 0.057$.

So a 95% CI for the true population proportion has lower limit

$$p - z_{\alpha/2}\widehat{\mathrm{SD}}(P) = 0.2 - 1.96 \cdot 0.057 = 0.089$$

and upper limit

$$p + z_{\alpha/2}\widehat{\mathrm{SD}}(P) = 0.2 + 1.96 \cdot 0.057 = 0.311$$

Confidence Intervals: Population Proportion

There's a pretty strong pattern here: if $\bar{X}$ is normally distributed, then a $(1-\alpha)\cdot 100\%$ CI for the true value of $E(\bar{X})$ (which is also the true value of $E(X_i)$) is

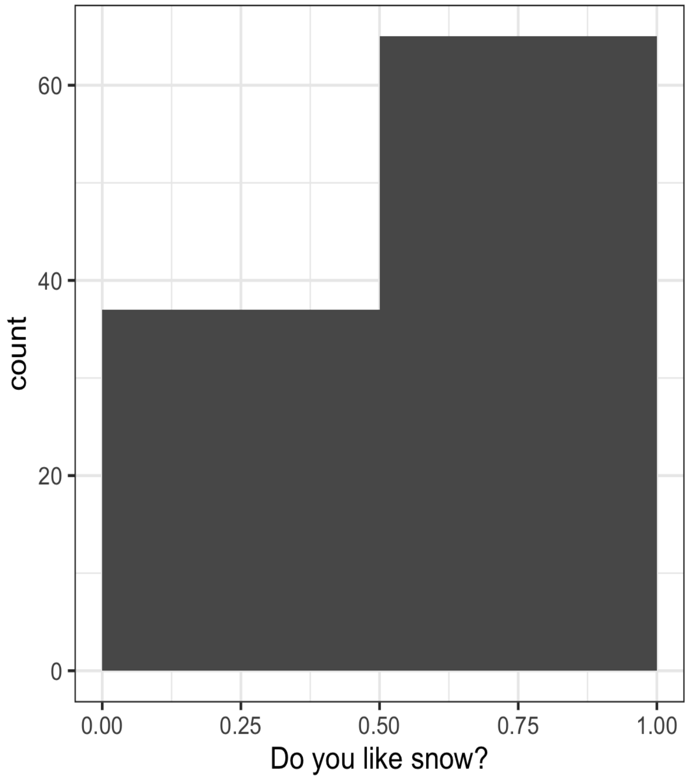
- $\bar{X} \pm z_{\alpha/2}\widehat{\mathrm{SD}}(\bar{X})$ if calculating $\bar{X}$ gives us $\widehat{\mathrm{SD}}(\bar{X})$ "for free",
- $\bar{X} \pm t_{\alpha/2}\widehat{\mathrm{SD}}(\bar{X})$ if we still need to estimate $\widehat{\mathrm{SD}}(\bar{X})$.

This "average $\pm$ critical value $\times$ standard deviation" pattern comes up all the time.

Confidence Interval: CLT Examples

Do you like snow?

True distribution:



Not normal because:

- not symmetrical
- not even continuous!!!

Confidence Interval: CLT Examples

Do you like snow?

Say we didn't have the entire population data. Just a sample of 20 students:

$$1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1$$

Estimated propotion: $p = \frac{10}{20} = 0.5$.

Estimated standard deviation of P:
$\widehat{\mathrm{SD}}(P) = \sqrt{p(1-p)/n} = \sqrt{0.5 \cdot 0.5/20} = 0.112$.

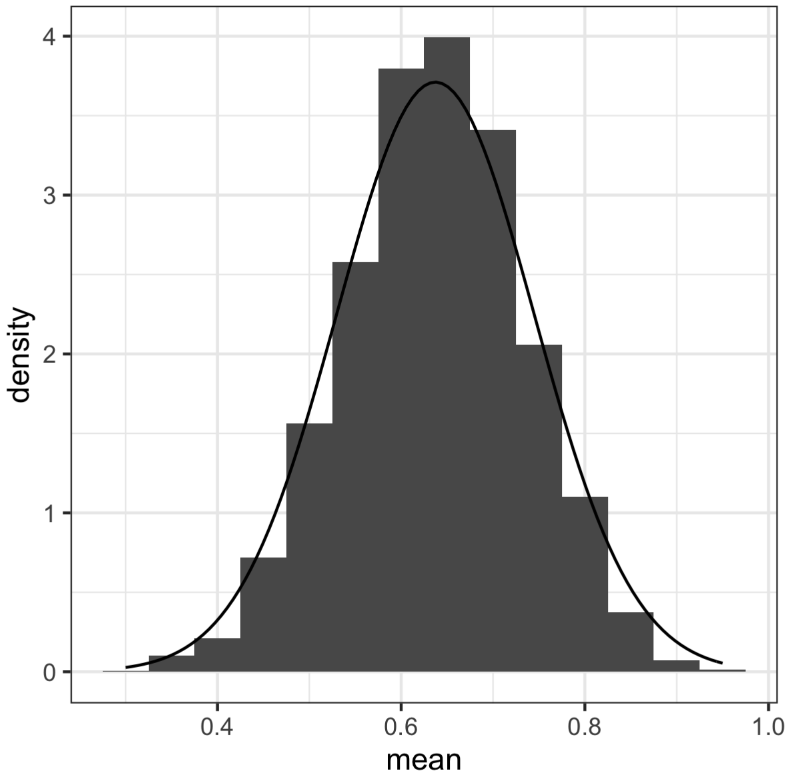
So a 95% CI is $0.5 \pm 1.96 \cdot 0.112 = [0.28, 0.72]$.

For once, we know the truth: 0.637.

Confidence Interval: CLT Examples

Do you like snow?

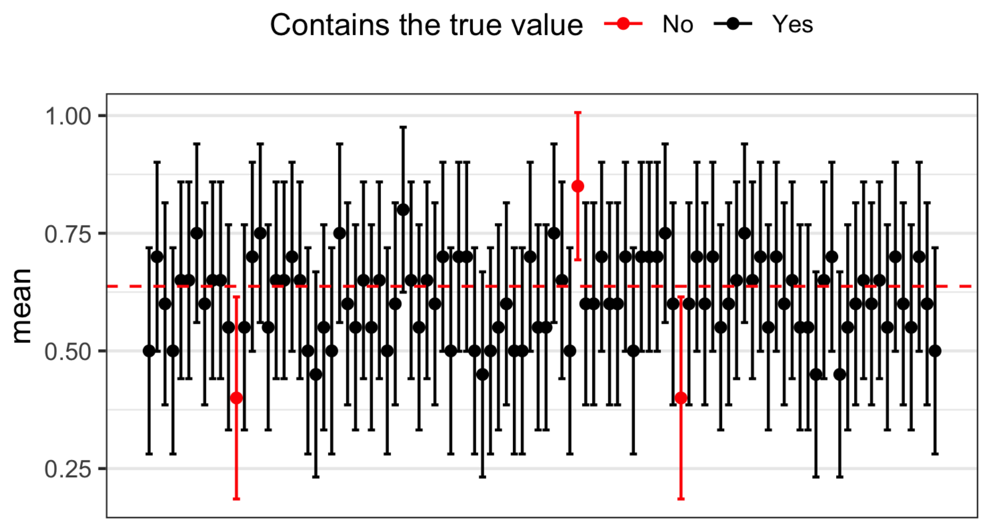
Let's repeat the process many, many times. Actually, I redo this 5000 times!



Confidence Interval: CLT Examples

Do you like snow?

If the confidence interval correct, it should contain the true value 95% of the time. Here are the first 100:

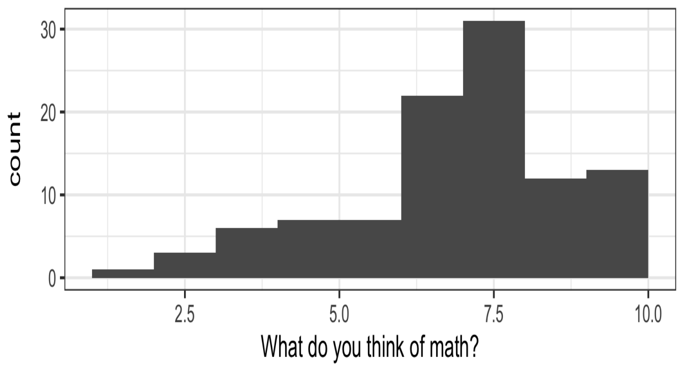


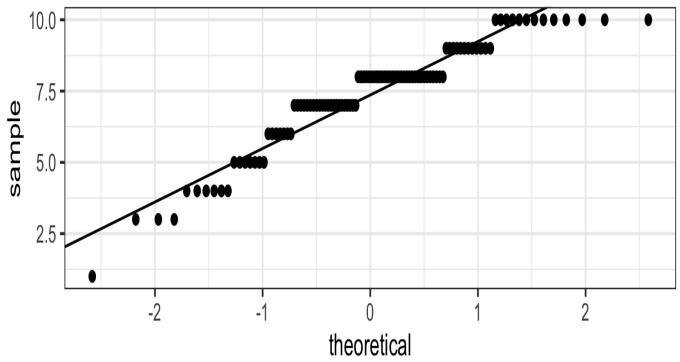
Proportion of all 5000 CIs containing the true value: 0.961. Pretty good!

Confidence Interval: CLT Examples

What do you think of math?

True distribution:





Not normal because:

- not symmetrical (left skewed)
- not even continuous!!!

Confidence Interval: CLT Examples

What do you think of math?

Say we didn't have the entire population data. Just a sample of 20 students:

$$7, 7, 8, 7, 1, 5, 7, 9, 9, 4, 10, 8, 7, 8, 8, 10, 3, 7, 4, 8$$

Estimated mean: $\bar{x} = 6.85$.

Estimated standard deviation of $\bar{X}$: $\widehat{\mathrm{SD}}(\bar{X}) = 0.5245815$.

So a 95% CI is

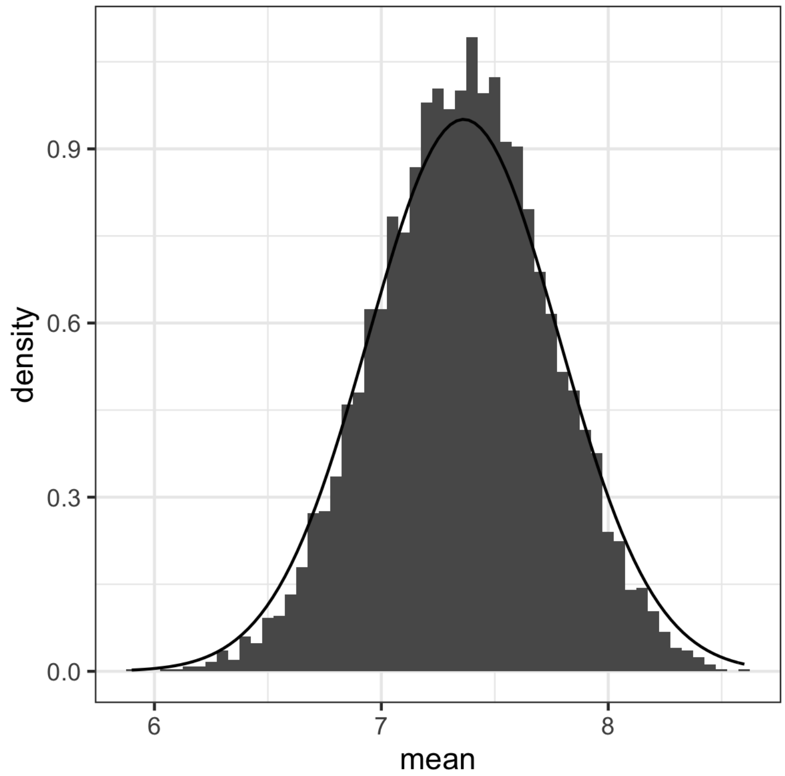
$$\begin{aligned} \bar{x} \pm t_{19,0.025} s/\sqrt{20} &= 6.85 \pm 2.0930241 \cdot 0.5245815 \\ &= [5.822, 7.878] \end{aligned}$$

For once, we know the truth: 7.363.

Confidence Interval: CLT Examples

What do you think of math?

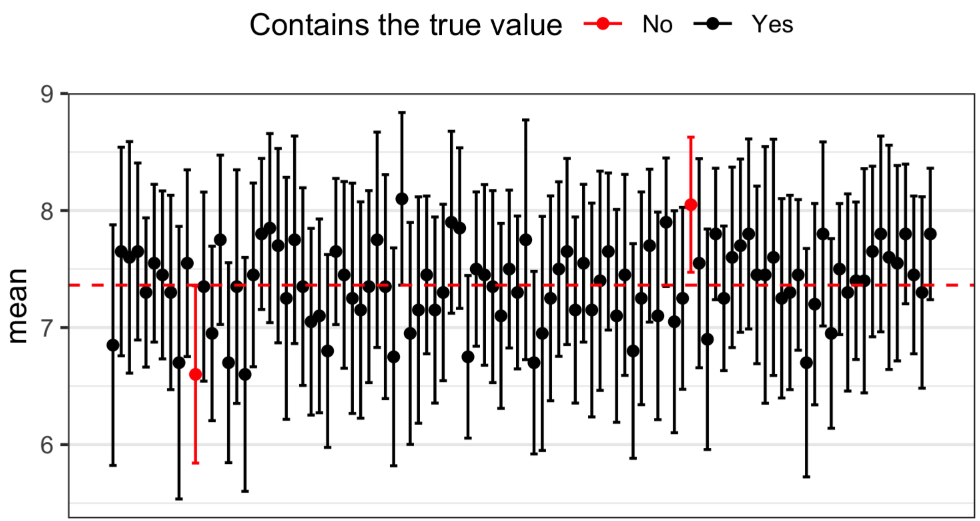
Let's repeat the process many, many times. Actually, I redo this 5000 times!



Confidence Interval: CLT Examples

What do you think of math?

If the confidence interval correct, it should contain the true value 95% of the time. Here are the first 100:



Proportion of all 5000 CIs containing the true value: 0.947. Pretty good!

Confidence Intervals: Summary I

It is all about finding an estimator for parameter of interest, and finding the distribution of that estimator. To find the distribution, the Central Limit Theorem is a powerful ally.

- If data are from a normal distribution and σ known: $\bar{X} \sim N$ and $\bar{X} \pm z_{\alpha/2}\mathrm{SD}(\bar{X})$ contains the true value of $E(X_i)$ $(1-\alpha)\cdot 100\%$ of the time
  - $\mathrm{SD}(\bar{X}) = \sigma/\sqrt{n}$
- If data are from a normal distribution and σ unknown: $\bar{X} \sim N$ and $\bar{X} \pm t_{n-1,\alpha/2}\widehat{\mathrm{SD}}(\bar{X})$ contains the true value of $E(X_i)$ $(1-\alpha)\cdot 100\%$ of the time
  - $\widehat{\mathrm{SD}}(\bar{X}) = S/\sqrt{n} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}}{\sqrt{n}}$

Confidence Intervals: Summary II

- If data are binary, and n "large enough": $P \sim N$ and $P \pm z_{\alpha/2}\widehat{\mathrm{SD}}(P)$ contains the true value of $E(X_i)$ $(1-\alpha)\cdot 100\%$ of the time
  - $\widehat{\mathrm{SD}}(P) = \sqrt{P(1-P)/n}$
  - n is large enough if $n \cdot \pi > 5$ and $n \cdot (1-\pi) > 5$.
  - We check this using the estimated value of π, i.e. p. So we check if $n \cdot p > 5$ and $n \cdot (1-p) > 5$.

- If data are NOT from a normal distribution, and n "large enough": $\bar{X} \sim N$ and $\bar{X} \pm t_{n-1,\alpha/2}\widehat{\mathrm{SD}}(\bar{X})$ contains the true value of $E(X_i)$ $(1-\alpha)\cdot 100\%$ of the time
  - $\widehat{\mathrm{SD}}(\bar{X}) = S/\sqrt{n} = \dfrac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}}{\sqrt{n}}$
  - usually, $n \geq 30$ satisfies n "large enough"