STAT 515 fa 2023 Lec 10 slides

Confidence intervals for the mean and proportion

Karl B. Gregory

University of South Carolina

Wed: Up to 2 bonus points on exam.

Quiz

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

SC poll from Sep, 2020: From a sample of 824 SC voters, 47% of them said they would vote for Trump, 43% for Biden, 1% for Jo Jorgensen, 1% for Howie Hawkins, and 8% are not sure.

95%: $\hat{p}_n \pm 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} = .47 \pm 1.96\sqrt{\frac{.47(1-.47)}{824}} = 0.034$

Margin of error reported as 3.4%, so if p is the proportion for Trump, poll says

$$p \in (0.47 - \underbrace{0.034}_{ME}, 0.47 + \underbrace{0.034}_{ME}) = \underbrace{(0.436, 0.504)}_{L \quad U}.$$

- This type of interval is called a *confidence interval (CI)*.
- We like to calibrate CIs so they capture their target with probability $1-\alpha$.
- The value $\alpha \in (0,1)$ is called the *confidence level*.

alpha

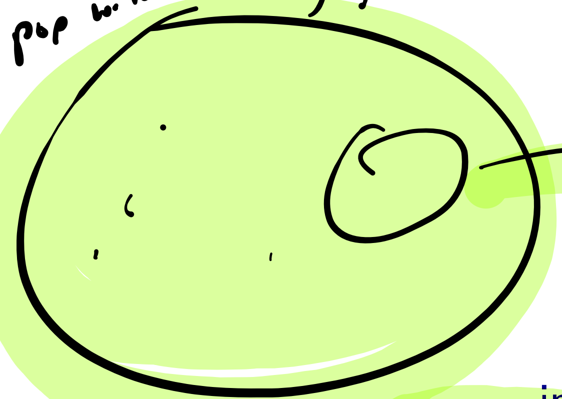
α is how often we permit ourselves to miss the target.

The idea of CIs is to find lower and upper bounds L and U such that

$$\overset{\hat{p}_n \pm ME}{p \in (L, U)} \quad \text{or} \quad \boxed{\overset{\bar{x}_n \pm ME}{\mu \in (L, U)}} \quad \text{or} \quad \sigma^2 \in (L, U),$$

for example, with probability $1-\alpha$.

pop with mean μ, variance σ^2



$$Z = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$x_1, \dots, x_n$, $\bar{x}_n$. sample.

Want to make an interval $\bar{x}_n \pm ME$ which "captures" μ w/ prob. 0.95.

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$ and use the fact that

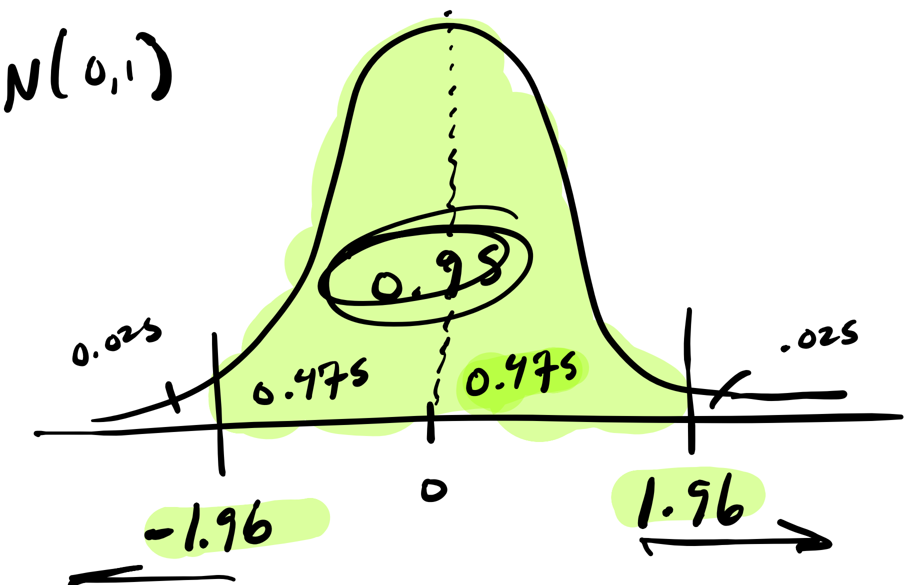
$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

to find a 95% confidence interval for μ.

$$P\left(-\frac{\sigma}{\sqrt{n}} 1.96 < \bar{x} - \mu < \frac{\sigma}{\sqrt{n}} 1.96\right) = 0.95 \qquad Z \sim N(0,1)$$

$$\Leftrightarrow P\left(-\bar{x}_n - \frac{\sigma}{\sqrt{n}} 1.96 < -\mu < -\bar{x}_n + \frac{\sigma}{\sqrt{n}} 1.96\right) = 0.95$$

$$\Leftrightarrow P\left(\underbrace{\bar{x}_n + \frac{\sigma}{\sqrt{n}} 1.96}_{U} > \mu > \underbrace{\bar{x}_n - \frac{\sigma}{\sqrt{n}} 1.96}_{L}\right) = 0.95$$



So the interval $\left(\underbrace{\bar{X}_n - \frac{\sigma}{\sqrt{n}} 1.96}_{L},\ \underbrace{\bar{X}_n + \frac{\sigma}{\sqrt{n}} 1.96}_{U}\right)$

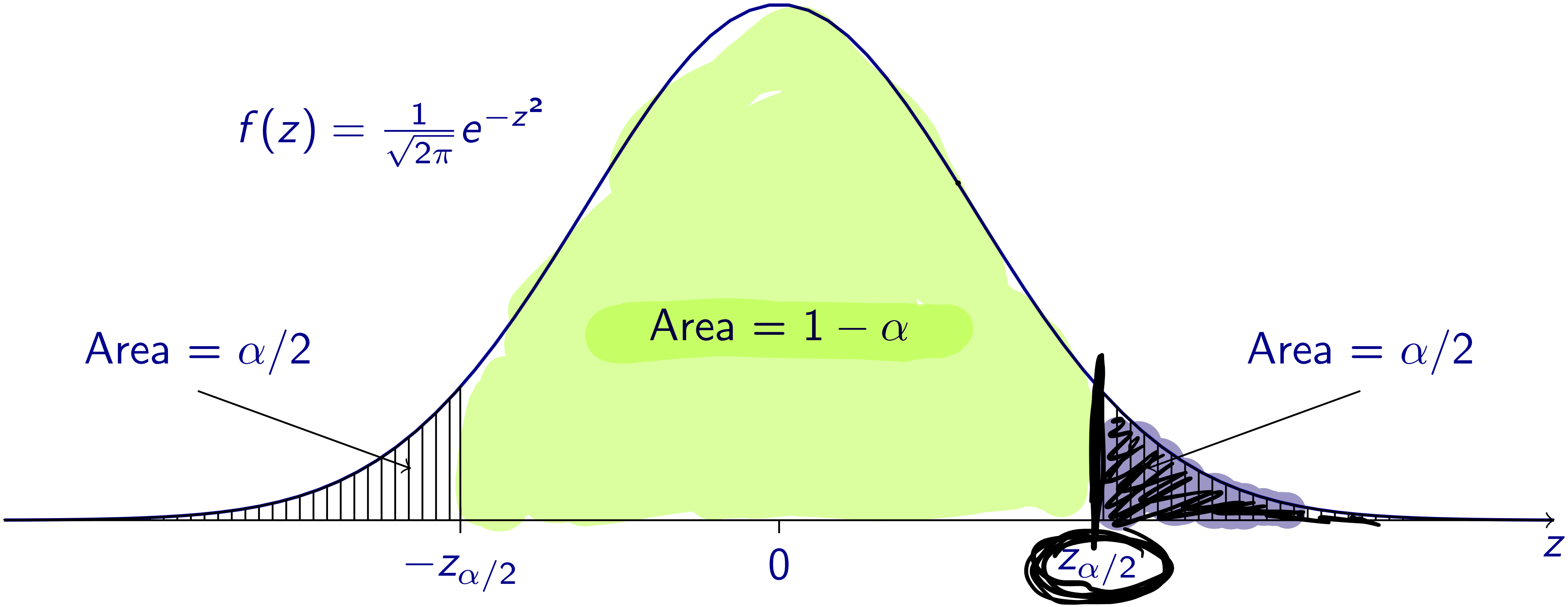
contain μ with prob 0.95.

$$\boxed{\text{A } \underline{95\%} \text{ confidence Interval for } \mu \text{ is } \quad \bar{X}_n \pm \frac{\sigma}{\sqrt{n}} \underline{1.96}.}$$

$\underline{95\%}$ corresponds to $\underline{1-\alpha} = .95$, $\underline{\alpha = 0.05}$.

$95\% = (1 - \underbrace{0.05}_{\alpha})100\%$

What about a general $(1-\alpha)\times 100\%$ CI for any $\alpha \in (0,1)$?



$z_{\alpha/2}$ the "upper" $\alpha/2$ quantile

Confidence interval for the mean of a Normal population with σ known

Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$. Then a $(1-\alpha) \times 100\%$ CI for μ is

$$\bar{X}_n \pm \underbrace{z_{\alpha/2}}_{\text{comes from z-table}} \frac{\sigma}{\sqrt{n}}.$$

Example: These are the commute times (sec) to class of a sample of students.

1832	1440	1620	1362	577	934	928	998	1062	900
1380	913	654	878	172	773	1171	1574	900	900

Assume the population is Normal with $\sigma = 400$.

x <- c(1832, 1440, ..., 900)
mean(x)

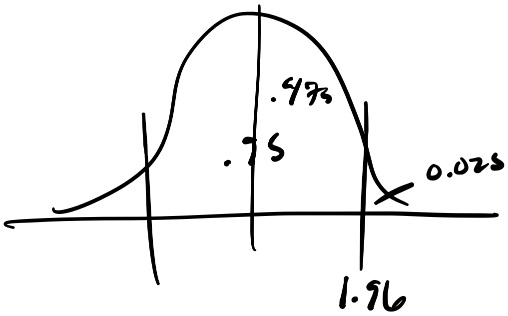
1. Construct a 95% confidence interval for the mean commute time of all students.
2. Construct a 99% confidence interval for the mean commute time of all students.
3. Give an interpretation of the two confidence intervals.
4. Which confidence interval is wider? Does it make sense why??

① $\bar{X}_n = 1048.4$, $\sigma = 400$ $n = 20$

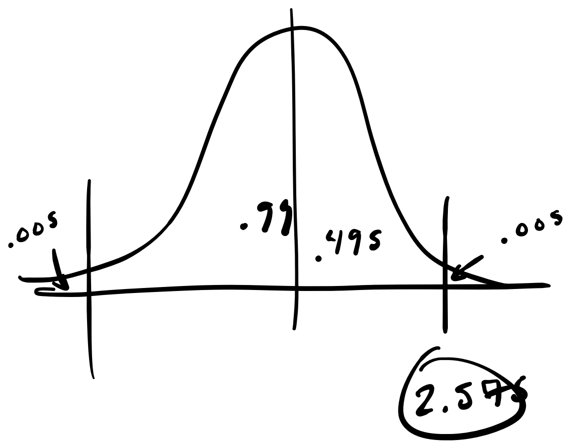
$(1-\alpha)100\%$ C.I. is $\bar{x}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

So a 95% C.I.

$$\boxed{1048.4} \pm 1.96 \frac{400}{\sqrt{20}} = (873.09, 1223.7)$$



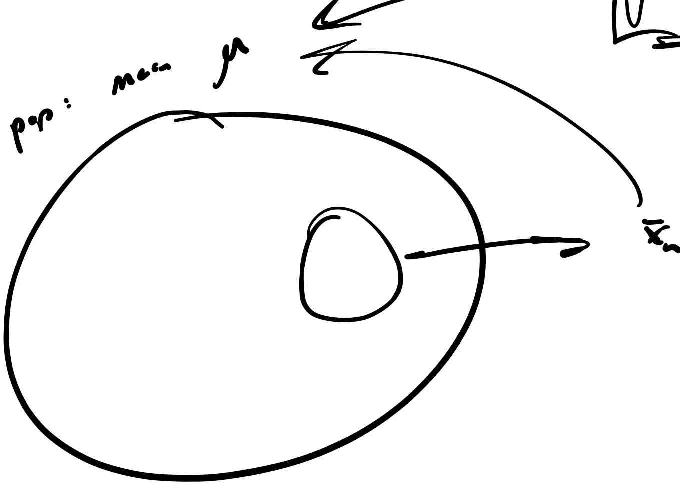
② 99% C.I. now: $\alpha = 0.01$



$$1048.4 \pm 2.575 \frac{400}{\sqrt{20}}$$

$$= (818.1, 1278.7)$$

$$\boxed{\bar{x}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}$$



Confidence interval for mean of a non-Normal pop. with σ known

Let $X_1, \dots, X_n$ be a rs from a non-Normal dist. with mean μ and var. $\sigma^2 < \infty$. Then

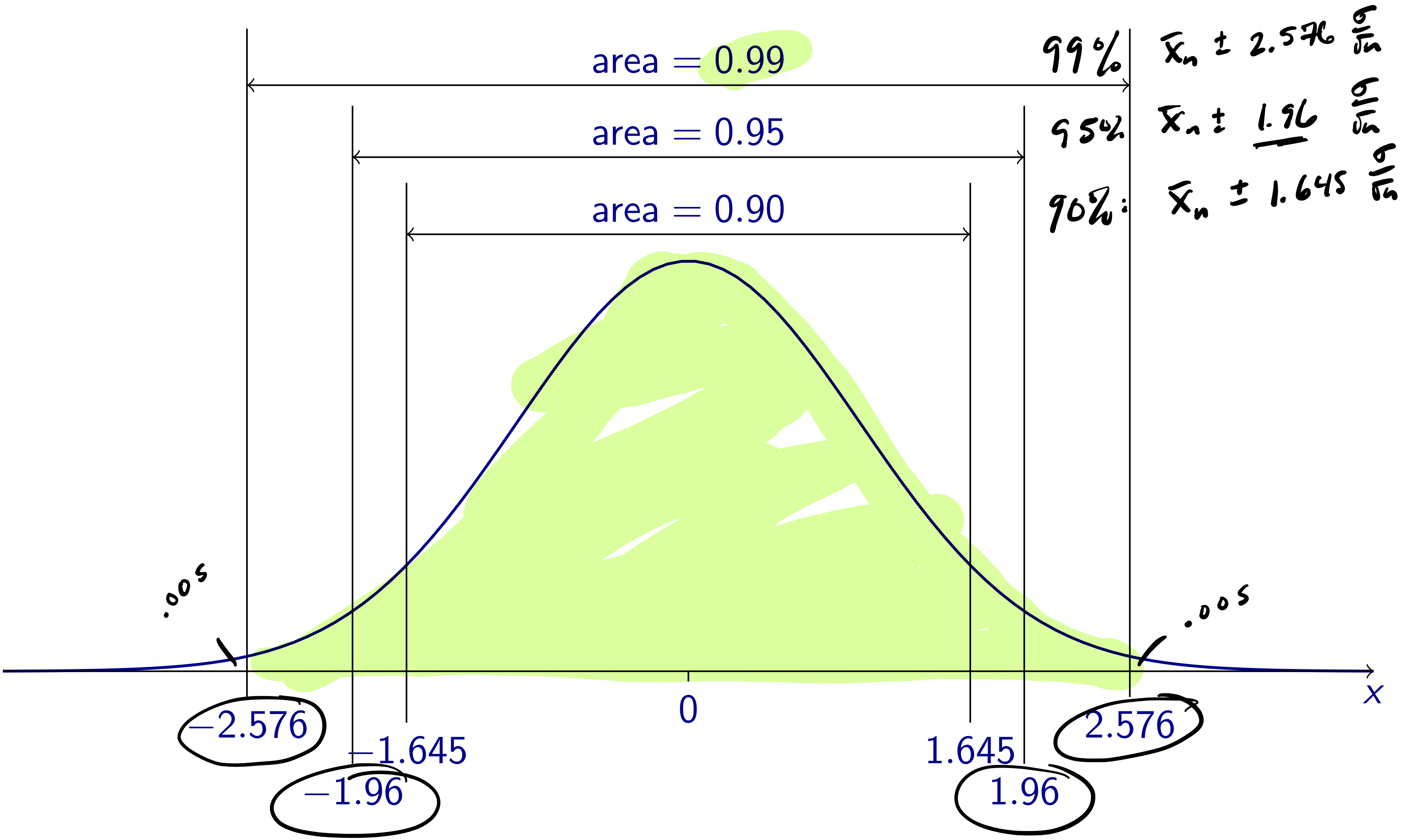
$$\bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

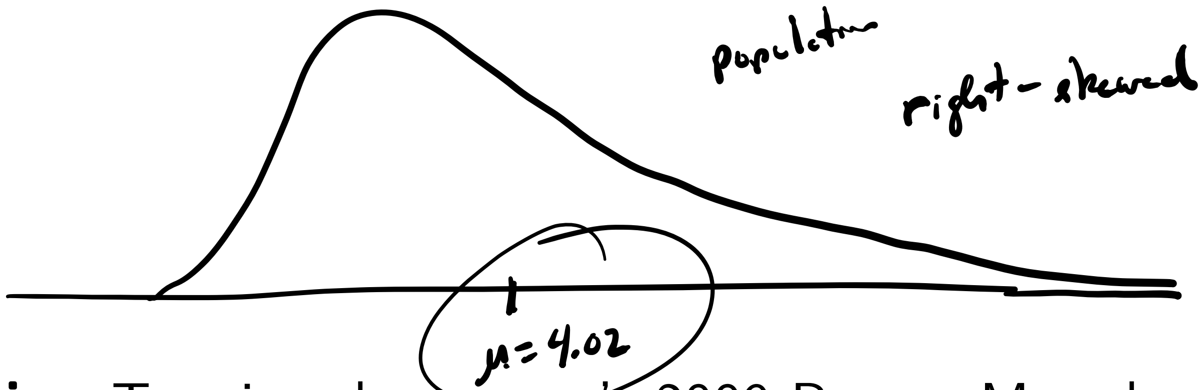
contains μ with probability closer and closer to $1 - \alpha$ for larger and larger n.

This is a $\boxed{\text{Central Limit Theorem}}$ result.

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} \text{Normal}(0,1) \quad \text{for large } \underline{\underline{n}}$$

We often construct 90%, 95%, and 99% confidence intervals:





$\sigma = 0.555$.

Exercise: Treating the women's 2009 Boston Marathon finishing times, which have mean $\mu = 4.02$ and standard deviation $\sigma = 0.555$,

1. Draw 100 samples of size $n = 30$ and each time build a 90% CI for the mean.
2. Check for what proportion of samples the CI captured the true mean μ.

link to women's data.

$$\bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Draw sample of $n = 30$ finishing times.

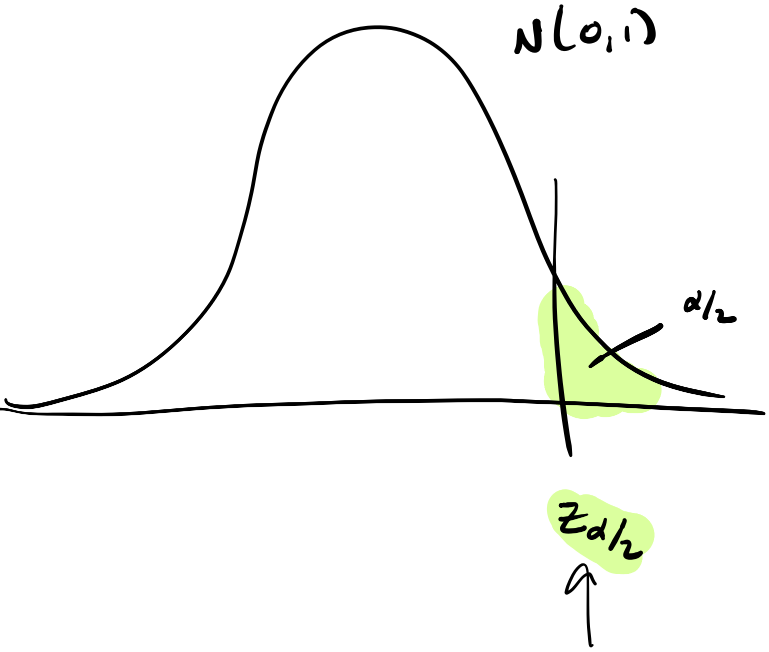
90% C.I.

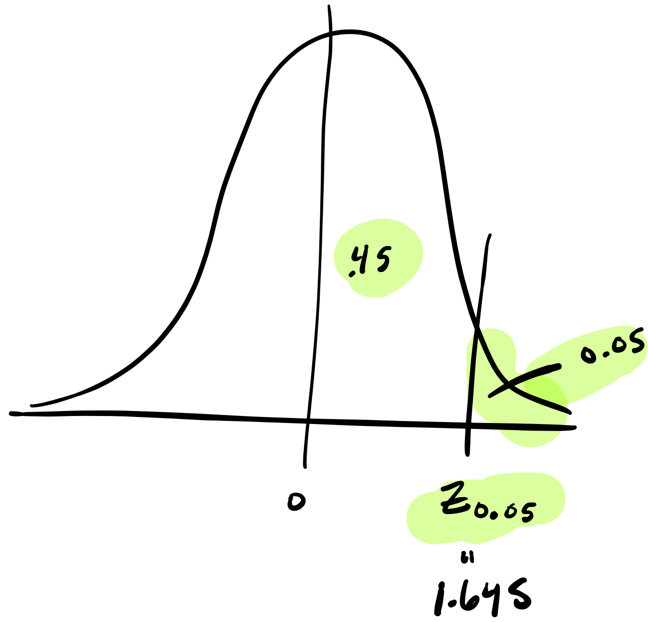
$$\bar{X}_n \pm z_{\frac{0.1}{2}} \frac{0.555}{\sqrt{30}}$$

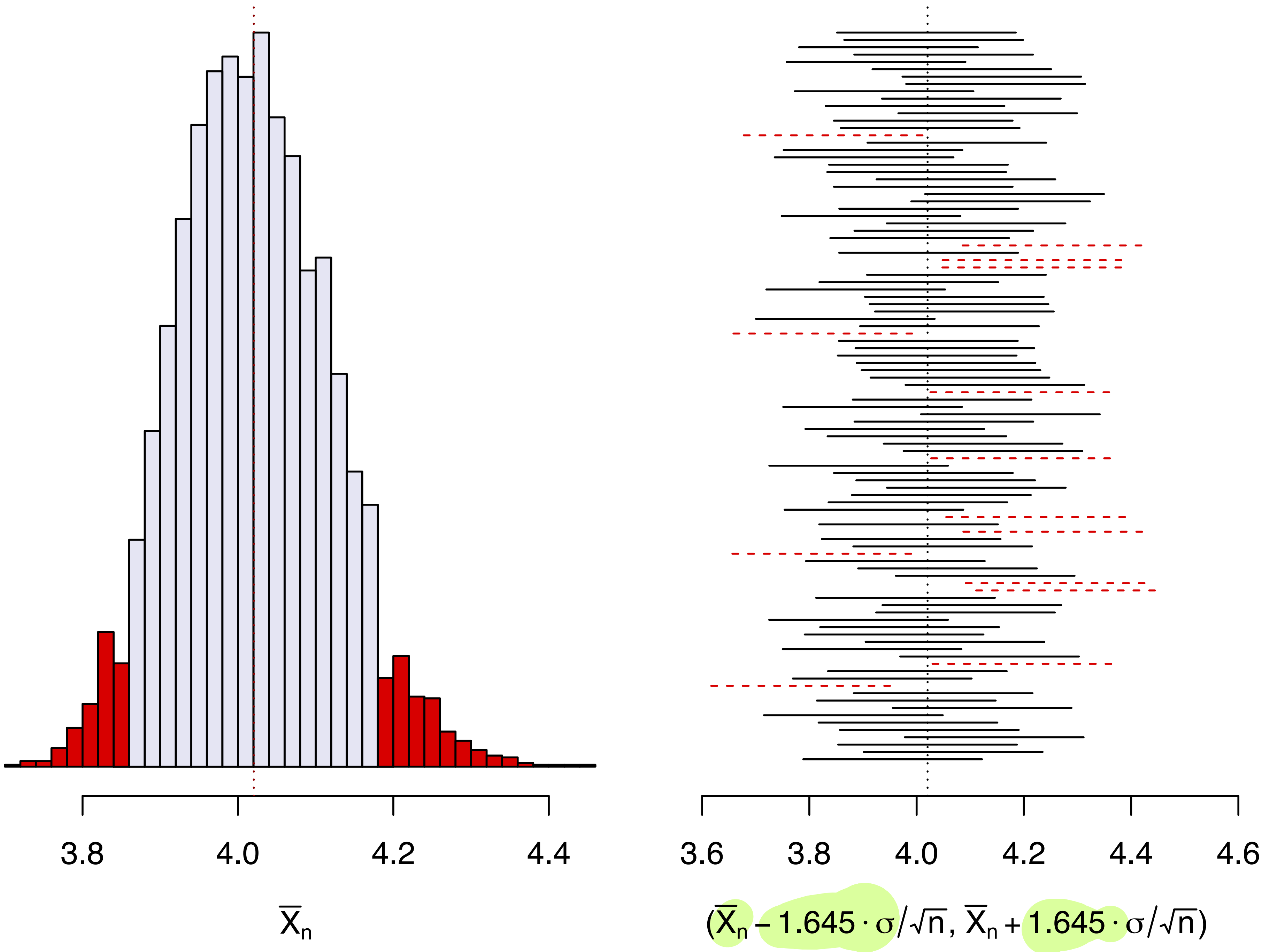
$\boxed{\alpha = 0.10}$

$z_{.05} = 1.645$

$$\bar{X}_n \pm 1.645 \frac{0.555}{\sqrt{30}}.$$


N(0,1)
α/2
Z α/2


.45
0.05
0
Z 0.05
=
1.645

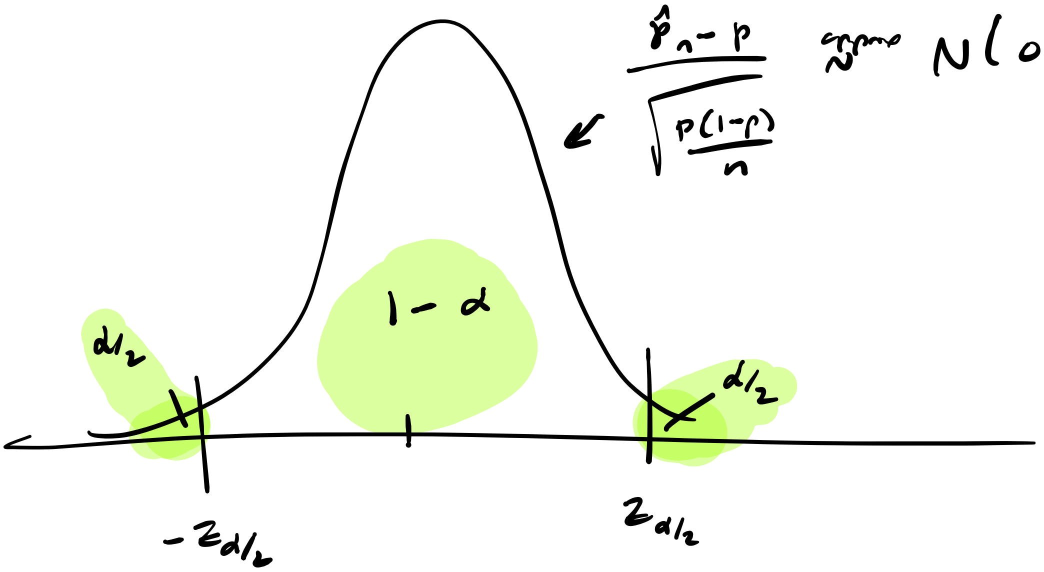

3.8
4.0
4.2
4.4
$\bar{X}_n$
3.6
3.8
4.0
4.2
4.4
4.6
$(\bar{X}_n - 1.645 \cdot \sigma/\sqrt{n}, \bar{X}_n + 1.645 \cdot \sigma/\sqrt{n})$

rearrange this

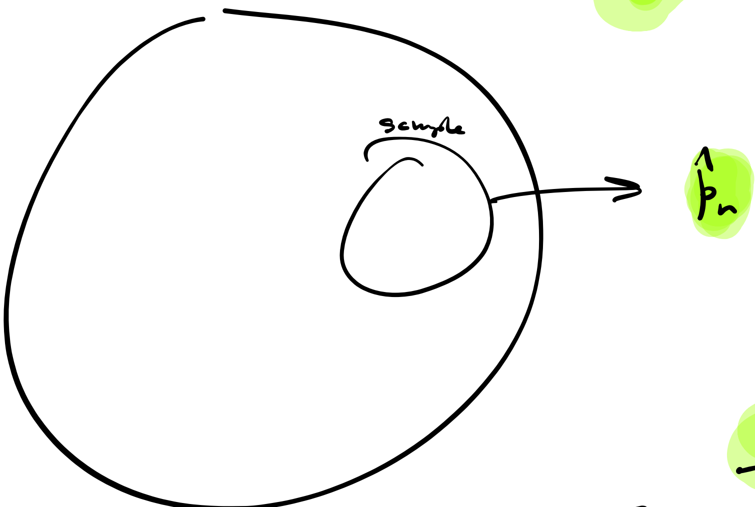
$$P\left(-z_{\alpha/2} < \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2} \right) = 1-\alpha$$

$$P\Bigg(\underbrace{\hat{p}_n - z_{\alpha/2}\sqrt{\tfrac{p(1-p)}{n}}}_{L} < p < \underbrace{\hat{p}_n + z_{\alpha/2}\sqrt{\tfrac{p(1-p)}{n}}}_{U} \Bigg) = 1-\alpha.$$

$$\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\text{approx}}{\sim} N(0,1)$$



pop: proportion of 1s is p.



$$Z \sim \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\text{approx}}{\sim} N(0,1)$$

if n large.

$$\mathbb{E}\,\hat{p}_n = p$$

$$\text{Var}\,\hat{p}_n = \frac{p(1-p)}{n}$$

$P($

Confidence interval (Wald-type) for a proportion

If $\underbrace{n\hat{p}_n}_{\#\text{ successes}} \geq 15$ and $\underbrace{n(1-\hat{p}_n)}_{\#\text{ failures}} \geq 15$, then

$$\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

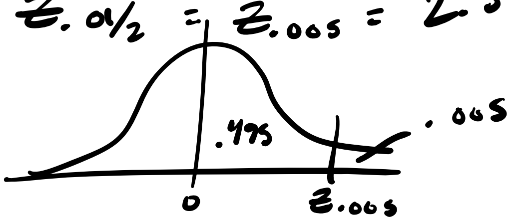
is an approximate $(1-\alpha) \times 100\%$ CI for p.

Explain how we arrive at this confidence interval.

Exercise: From a sample of 1,000 voters, 478 say they will vote for candidate A. Let p be the true proportion of voters who will vote for candidate A.

1. Build a 95% CI for p. $\alpha = 0.05 \rightarrow z_{\alpha/2} = z_{.05/2} = z_{0.025} = 1.96$
2. Build a 99% CI for p. $\alpha = 0.01 \rightarrow z_{\alpha/2} = z_{.01/2} = z_{.005} = 2.575$

$n = 1000$



(1) $\hat{p}_n = \frac{478}{1000}$

$$\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} = \frac{478}{1000} \pm 1.96\sqrt{\frac{\frac{478}{1000}\left(1-\frac{478}{1000}\right)}{1000}}$$

$$= .478 \pm 0.031$$

$$= (0.447,\ 0.509)$$

$$47.8\% \pm 3.1\%$$

$$= \frac{478}{1000} \pm 2.576\sqrt{\frac{\frac{478}{1000}\left(1-\frac{478}{1000}\right)}{1000}} = (0.437,\ 0.519).$$

Sample of size n

$\hat{p}_n = \frac{\#\text{successes}}{n}$

Confidence interval (Agresti-Coull) for a proportion

If $\underbrace{n\hat{p}_n}_{\#\text{successes}} \geq 5$ and $\underbrace{n(1-\hat{p}_n)}_{\#\text{failures}} \geq 5$, then

$$\tilde{p}_n \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_n(1-\tilde{p}_n)}{n+4}}, \quad \text{where} \quad \tilde{p}_n = \frac{\#\{\text{successes}\}+2}{n+4},$$

is an approximate $(1-\alpha)\times 100\%$ CI for p.

$n = 50$ $\quad \hat{p}_n = \frac{5}{50}$ $\quad \tilde{p}_n = \frac{7}{54}$

Exercise: From a sample of $\boxed{\text{50 students, 5}}$ say they hang-dry their laundry. Let p be the true proportion of students who hang-dry their laundry.

1. Build a 95% $\boxed{\text{Wald-type}}$ CI for p. $\hat{p}_n \pm 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} = (0.017, 0.183)$
2. Build a 95% Agresti-Coull CI for p. $\tilde{p}_n \pm 1.96\sqrt{\frac{\tilde{p}_n(1-\tilde{p}_n)}{n+4}} = (0.04, 0.219)$
3. Do the same if 50 out of 500 students say they hang-dry their laundry.

Return to 2020 poll on first slide and check what α is...