IMS 13 (ish): Normal Distribution and Confidence Intervals

Recall: FTS

Fundamental Theorem of Statistics (overview)

Suppose we have an independent random sample taken from a given population. Then estimated mean/proportion will converge to the true mean/proportion as the sample size (n) increases

$$\overline{x} \to \mu$$

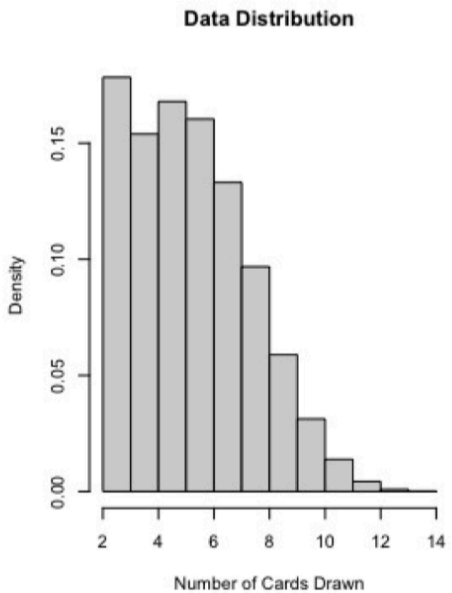
$$\hat{p} \to p$$

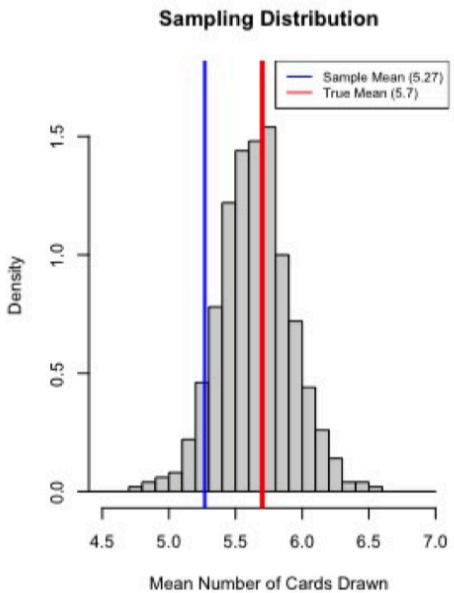
Furthermore, the sampling distribution of $\overline{x}/\hat{p}$ convervges to a **normal distribution** as n increases.

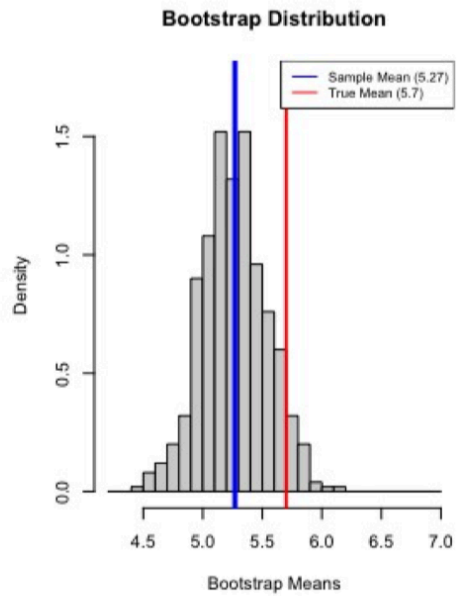
What is the mean and standard deviation of the normal distribution for the sample mean? That is, $\overline{X} \sim N(?, ?)$?

Distribution for $\overline{x}$

- The shape/spread of a *data distribution* and the *sampling distribution* is usually different.
- The shape/spread of a *sampling distribution* and the *bootstrap distribution* is the same (the centers can differ though).







Mean	μ	μ	μ
Std. Dev	σ_x	$\sigma_{\overline{x}}$	$\sigma_{\overline{x},boot}$

Distribution for $\overline{x}$

Sampling Distribution of the Sample Mean $\overline{X}$

For a independent random sample: $X_1, X_2, \ldots, X_n$, with a population mean μ and standard deviation σ_x then as $n \to \infty$,

$$\overline{X} \sim N\left(\mu, \sigma_{\overline{x}}\right)$$

where $\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$ is called the **standard error**.

- The value $\sigma_{\overline{x}}$ is the standard deviation of sample means.
- The value σ_x is the standard deviation of the original data.
- We refer to $\sigma_{\overline{x}}$ as *standard error* and as σ_x is the *standard deviation* for simplicity.
- $\sigma_{\overline{x}}$ and $\sigma_{\overline{x},boot}$ are the same

Finding $\sigma_{\overline{x}}$

$$
\begin{aligned}
Var(\overline{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) && \text{Def. of } \overline{X} \\
&= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} x_i\right) && \text{Linear Properties of Var} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(x_i\right) && \text{Independence} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma_x^2 && \text{Same Distr.} \\
&= \frac{1}{n^2} n\sigma_x^2 \\
\rightarrow SD(\overline{X}) &= \sqrt{Var(\overline{X})} = \frac{\sigma_x}{\sqrt{n}}
\end{aligned}
$$

Approximately Normal

- Are the distributions for $\overline{x}$ and $\hat{p}$ always a normal distribution? Not quite.
- The FTS is specifically regarding the behavior of the statistic $\overline{x}$ and $\hat{p}$ when the sample size is large.
- The real sampling distribution (or data distribution) is never exactly normal. Instead we look for *approximately normal*.
- We have rules of thumb for when $\overline{x}$ and $\hat{p}$ are approximately normal.

Approximately Normal

Rule-of-thumb: How big of a sample do we need for $\overline{x}$ and $\hat{p}$ to be approximately normal?

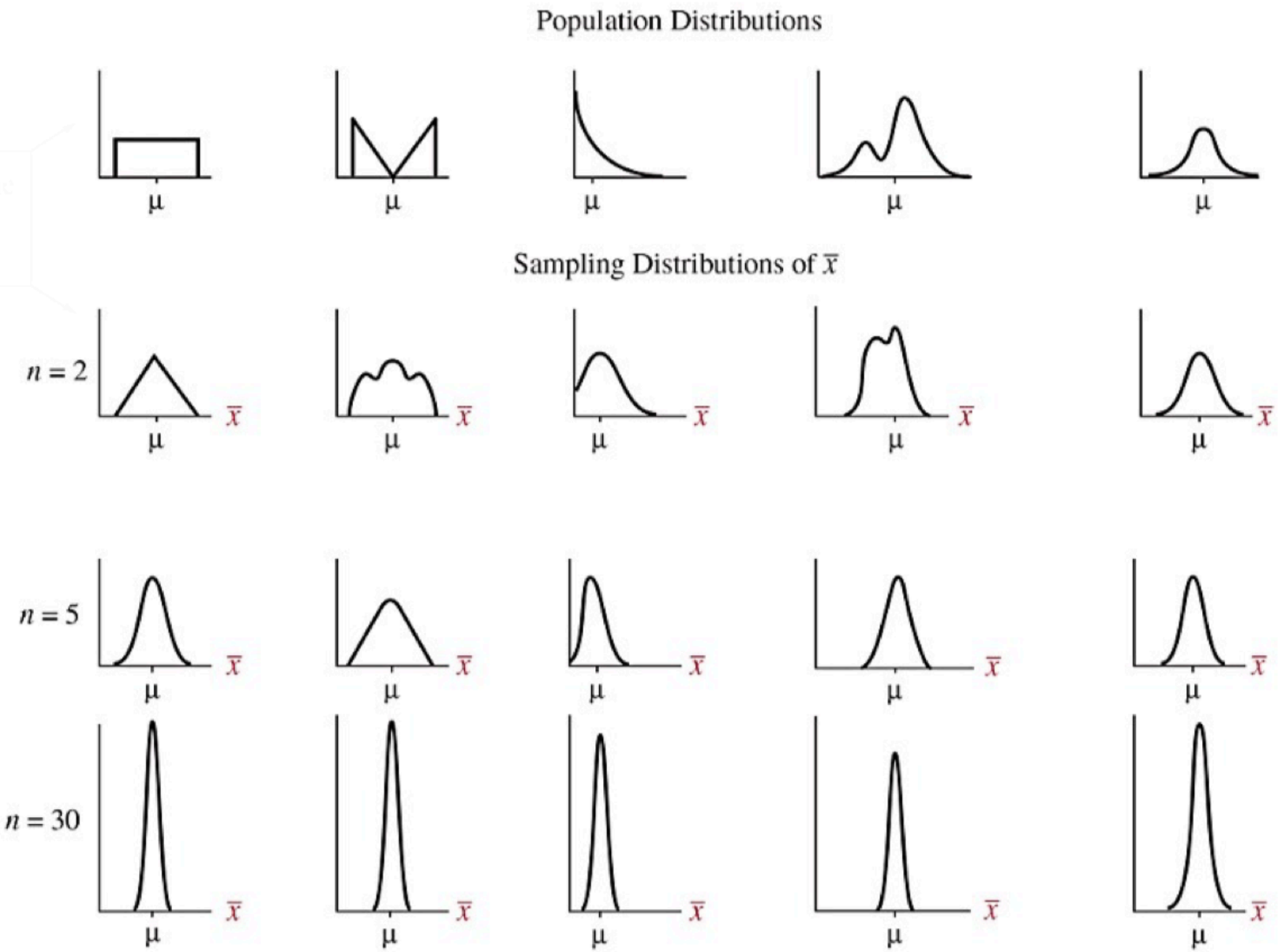
If $X_1, \ldots, X_n$ is a binary (indicator) variable with $P(X = 1) = p$, then $\hat{p}$ is approximately normally distributed when

$$np \geq 10 \text{ and } n(1-p) \geq 10$$

If $X_1, \ldots, X_n$ is a numeric variable, then $\overline{x}$ is approximately normally distributed when

$$n \geq 30$$

Approximately Normal



Practice Problem

Each winter season, Peter works numerous shifts as a waiter in a mountain resort in Austria. His salary (including tips) varies from shift to shift depending on the time or day. The resort manager told Peter that he can expect to earn €150 per shift, but that the salary may vary a lot from shift to shift, as evidenced by a standard deviation of €60. At the end of each season, Peter randomly selects pay stubs from 32 shifts he worked that season and computes the average salary per shift.

a) Around which euro value would you expect his average salary per shift to fluctuate?

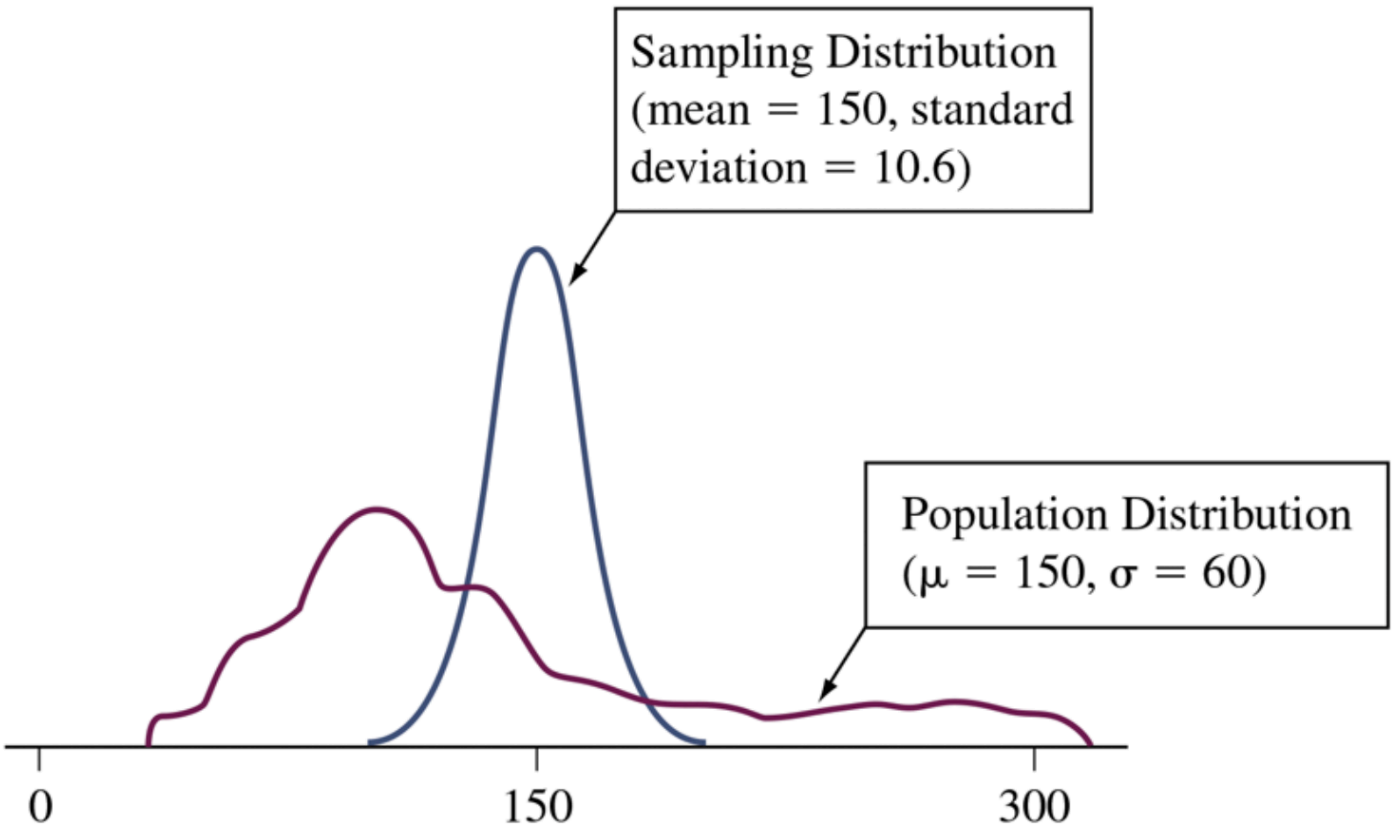
b) How much variability would you expect in the average salaries from one season to the next? Find the standard deviation of the sampling distribution of the sample mean (standard error).

c) Do we suspect the distribution for the average salary he calculated using 32 shifts to be approximately normal?

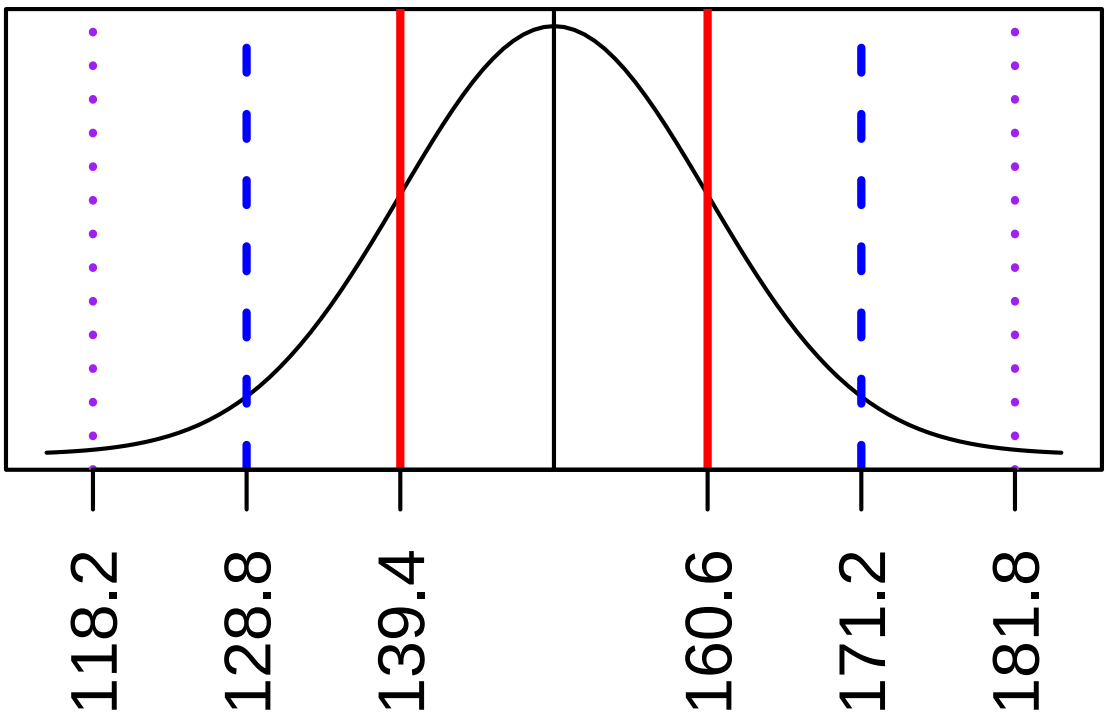
d) Between what values do you expect about 95% of his average salaries to fall over many seasons?

e) What is the probability that there is a season when his average salary falls below €129?

Practice Problem

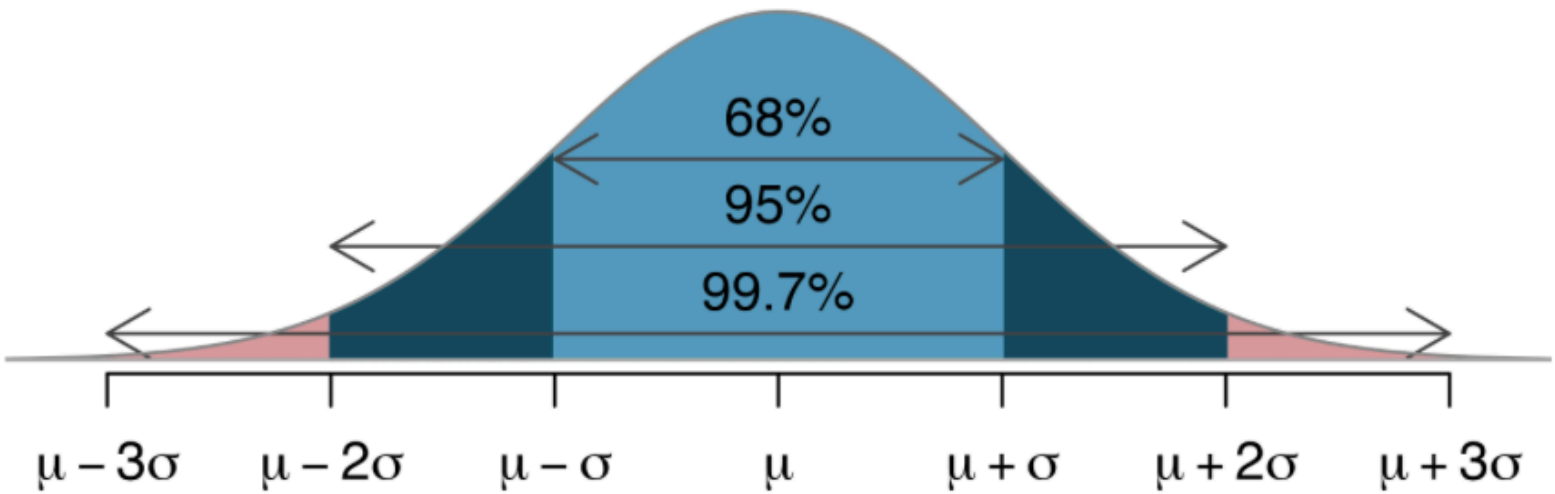


Practice Problem



Confidence Intervals

- We previously found confidence intervals and sampling distributions using *percentiles* for the bootstrap distribution.
- Because we know that $\overline{x} \sim N(\mu, \sigma_{\overline{x}})$, we can also find confidence intervals mathematically.



Confidence Intervals

Constructing Confidence Intervals Mathematically

When the sampling distribution of a point estimate can reasonably be modeled as normal, the point estimate we observe will be within $z_{\alpha/2}$ standard errors of the true value of interest about $(1-\alpha)100\%$ of the time. Thus, a $(1-\alpha)100\%$ confidence interval for such a point estimate can be constructed. The lower and upper bounds of the CI are:

$$\text{lower} = \text{point estimate} - z_{\alpha/2} \times SE$$

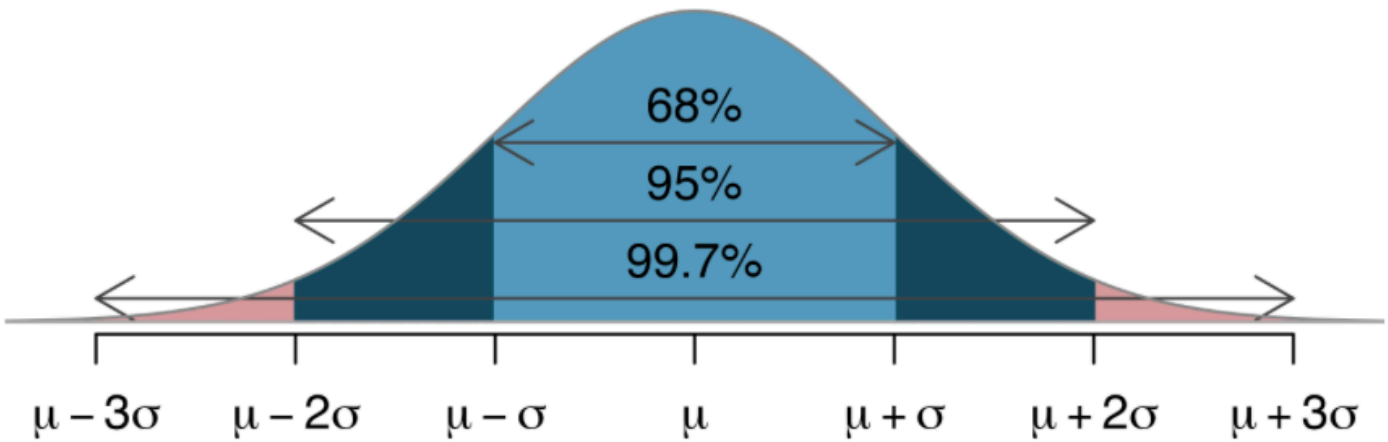
$$\text{upper} = \text{point estimate} + z_{\alpha/2} \times SE$$

Interpretation does not change (it is in the notes on Oct 18)!

Confidence Intervals

That is, we expect a point estimate to fall within

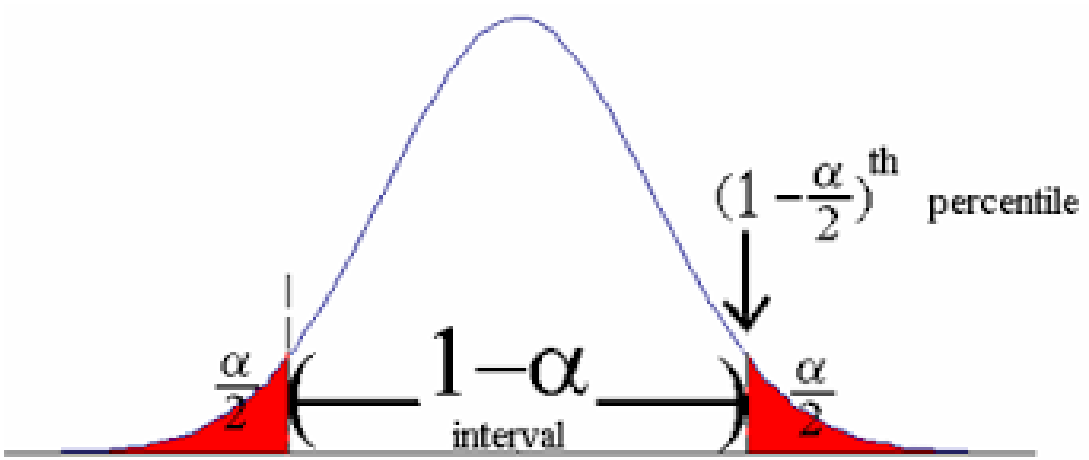
- $z_{\alpha/2} = 2$ (or 1.96) standard errors from the true mean about 95% of the time.
  - $\alpha = .05$
- $z_{\alpha/2} = 2.576$ standard errors from the true mean about 99% of the time.
  - $\alpha = .01$



Confidence Intervals

The value $z_{\alpha/2}$ is the $(1-\frac{\alpha}{2})100^{th}$ percentile of the standard normal distribution. Because of symmetry $z_{\alpha/2}$ corresponds to

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



Confidence Intervals

- For a 90% confidence interval we use $z_{.05} = 1.645$
- For a 95% confidence interval we use $z_{.025} = 1.96$
- For a 99% confidence interval we use $z_{.005} = 2.576$

Practice Problems

Researchers where interested in houses that were recently sold in the Duke Forest neighborhood of Durham, NC. They recorded 98 home sales in 2020. The data was recorded in units of 10,000 USD. The summary statistics for the data distribution is below.

min	1st Qu	Median	Mean	3rd Qu	Max	Var
9.50	45.06	54.00	55.99	64.38	152.00	508.27

a) Do you think that the data distribution is normal? Why or why not?

b) Can you construct a 95% confidence interval for the population mean? If not,explain why not. If so, do so and interpret.

Practice Problems

[IMS 13.2 Adjusted] Twitter users and news. A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter. The standard error for this estimate was 2.4%, and a normal distribution may be used to model the sample proportion.

a) Construct a 99% confidence interval for the fraction of U.S. adult Twitter users who get some news on Twitter, and interpret the confidence interval in context.

b) Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

   i) The confidence interval provides evidence that more than half of U.S. adult Twitter users get some news through Twitter.

   ii) Since the standard error is 2.4%, we can conclude that 97.6% of all U.S. adult Twitter users were included in the study.

   iii) If we want to reduce the standard error of the estimate, we should collect less data.

   iv) If we construct a 90% confidence interval for the percentage of U.S. adults Twitter users who get some news through Twitter, this confidence interval will be wider than a corresponding 99% confidence interval.

Practice Problems

[IMS 12.7] Cyberbullying rates. Teens were surveyed about cyberbullying, and 54% to 64% reported experiencing cyberbullying (95% confidence interval). Answer the following questions based on this interval. (Pew Research Center 2018)

a. A newspaper claims that a majority of teens have experienced cyberbullying. Is this claim supported by the confidence interval? Explain your reasoning.

b. A researcher conjectured that 70% of teens have experienced cyberbullying. Is this claim supported by the confidence interval? Explain your reasoning.

c. Without actually calculating the interval, determine if the claim of the researcher from part (b) would be supported based on a 90% confidence interval?

Connecting Bootstrap CI and Traditional CIs

- We have learned two different ways to construct confidence intervals:
  - Bootstrap distribution using percentiles
  - Mathematical properties with the CLT and $z_{\alpha/2}$
- These methods should return the same confidence interval (or at least extremely similar)!

Researchers where interested in houses that were recently sold in the Duke Forest neighborhood of Durham, NC. They recorded 98 home sales in 2020. The data was recorded in units of 10,000 USD.

a) Calculate a bootstap CI.
b) Plot vertical lines indicating the bounds for a 95% bootstrap CI (blue) and traditional CI (red) bounds on the same histogram.

Recall: Bootstrap CI

Create 10,000 bootstrap samples, and estimate the mean for each sample.

Lower and upper bounds for a 95% bootstrap CI.

```
# Change scale (for ease of reading)
duke_forest$price <- duke_forest$price/10000

# Bootstrap means
bootstrap_means <- duke_forest %>%
                   rep_sample_n(size=98, reps=10000, replace =TRUE) %>%
                   summarize(boot_mean = mean(price))

# Bounds
lower <- quantile(bootstrap_means$boot_mean, 0.025)
upper <- quantile(bootstrap_means$boot_mean, 0.975)
c(lower, upper)
```

```
    2.5%    97.5%
51.74429 60.59210
```

Confidence Intervals using the formula

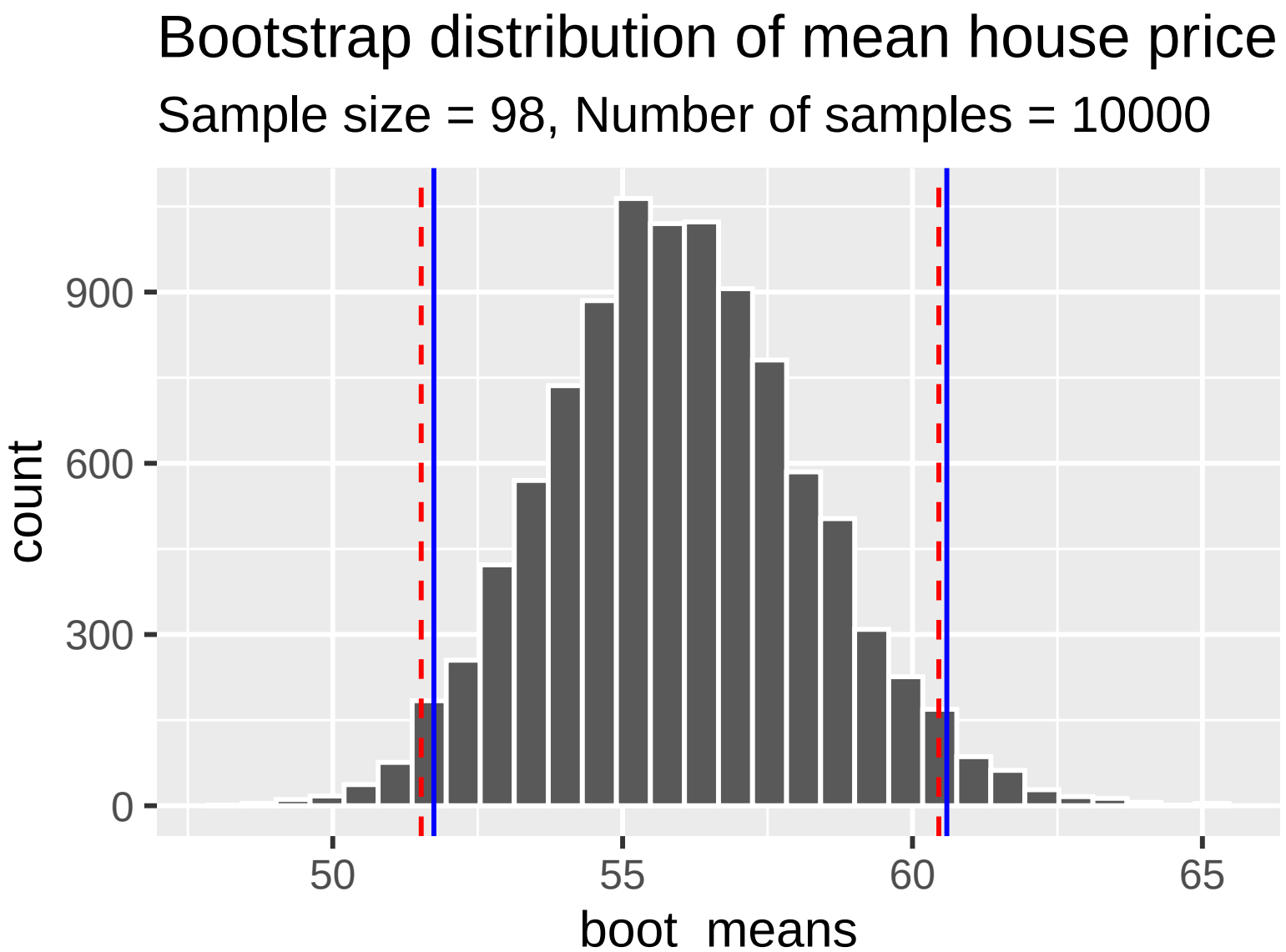
```
# Values needed for CI
SE <- sd(duke_forest$price)/sqrt(98)
sample_mean <- mean(duke_forest$price)

# CI bounds
upper_z <- sample_mean + 1.96*SE
lower_z <- sample_mean - 1.96*SE
c(lower_z, upper_z)
```

```
[1] 51.52622 60.45351
```

Plotting Both

```
ggplot(data = bootstrap_means, aes(x = boot_mean)) +
  geom_histogram( color = "white") +
  labs(
    x = "boot_means",
    title = "Bootstrap distribution of mean house price in 10,000 USD",
    subtitle = "Sample size = 98, Number of samples = 10000"
  ) +
  geom_vline(xintercept = lower, colour = "blue")+
  geom_vline(xintercept = upper, colour = "blue")+
  geom_vline(xintercept = lower_z, colour = "red", lty = 2)+
  geom_vline(xintercept = upper_z, colour = "red", lty = 2)
```

Plotting Both



Another Example

- Recall the in class activity from Lecture 16.
- Create a 99% confidence interval for this data set using traditional CI method.