Some Reminders:

- No class or tutorial next week (Reading week)
  - No OH, but Piazza will be regularly checked
  - New College Stats Aid Centre will be open during reading week
- The midterm is the following week during your usual tutorial time (March 1st)
  - You MUST attend the correct section's midterm.
  - AM section: EX200
  - PM Section:
    - MS 3154: Last names from A – Lo
    - WB 116: Last names from Lu – Z
  - Includes all material up to & including Feb 25th (mostly a review class)
  - Format: Multiple choice, fill in the blanks, written answers (make sure to write complete sentences)
- Example midterms have been posted to Quercus

Topics for today

- Sampling / Bootstrapping
- Confidence interval

Sampling and Sampling distribution

- Recall that: sampling distribution is the distribution for the population parameter such as sample mean and sample variance.
- **Bootstrapping approach:** estimate sampling distribution by re-sampling from the **original** sample.
- Note that bootstrapping does not create new data.
- It simply is a tool to allow us to explore the variability of estimates from our original sample

Question 1 (c)

- Variable we will be working on: PAID (paid claims)
- First, we take a look at our data set.

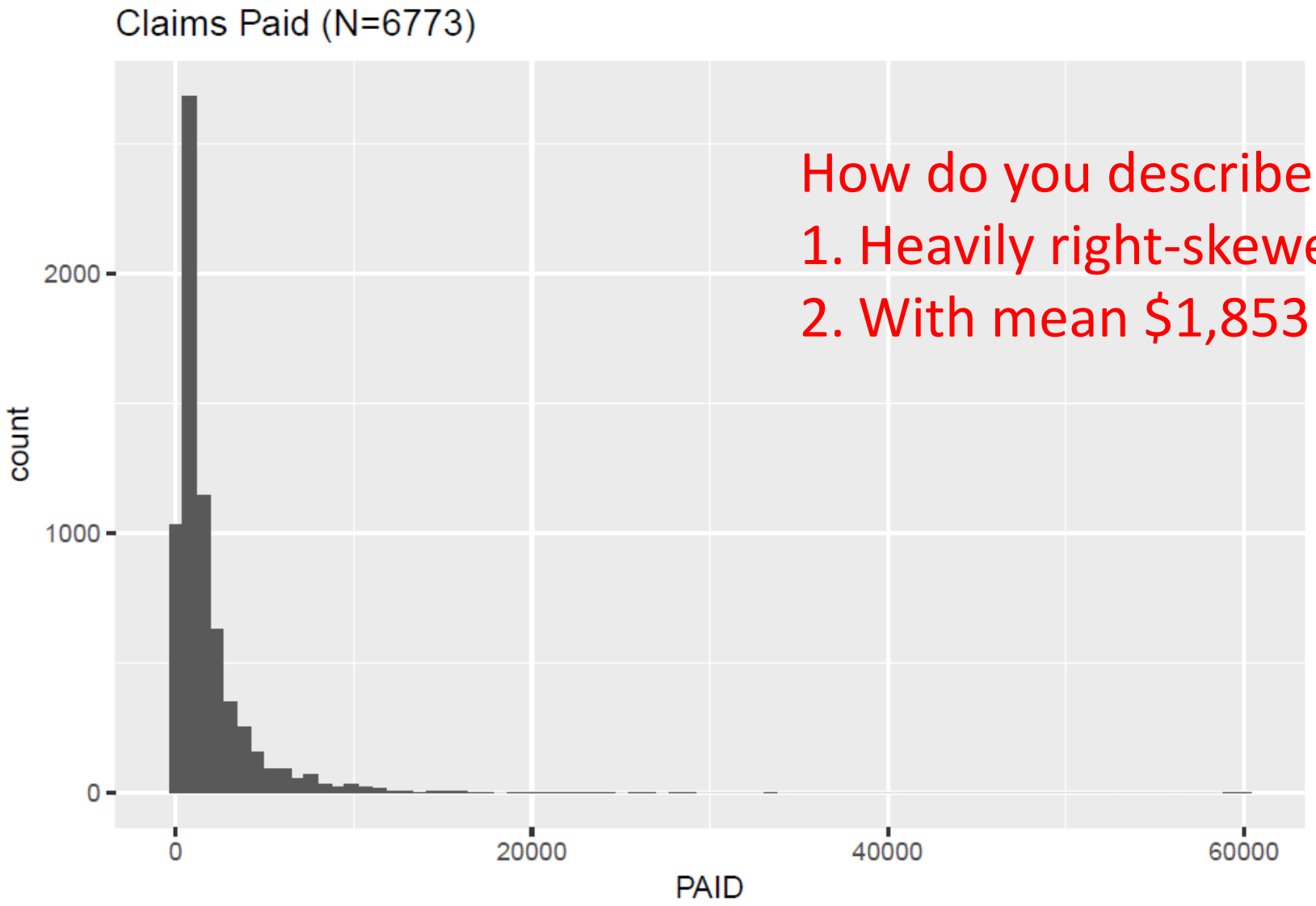
```
## $ PAID   <dbl> 1134.44, 3761.24, 7842.31, 2384.67, 650.00, 391.12, 377...
```

```
summarise(AutoClaims,
          min=min(PAID),
          mean = mean(PAID),
          median = median(PAID),
          max=max(PAID),
          sd = sd(PAID),
          n=n())
```

```
##   min     mean median   max       sd    n
## 1 9.5 1853.035 1001.7 60000 2646.909 6773
```

Distribution of PAID

```
ggplot(AutoClaims, aes(x=PAID)) + geom_histogram(bins=80) + labs(title="Claims Paid (N=6773)")
```



How do you describe the shape?

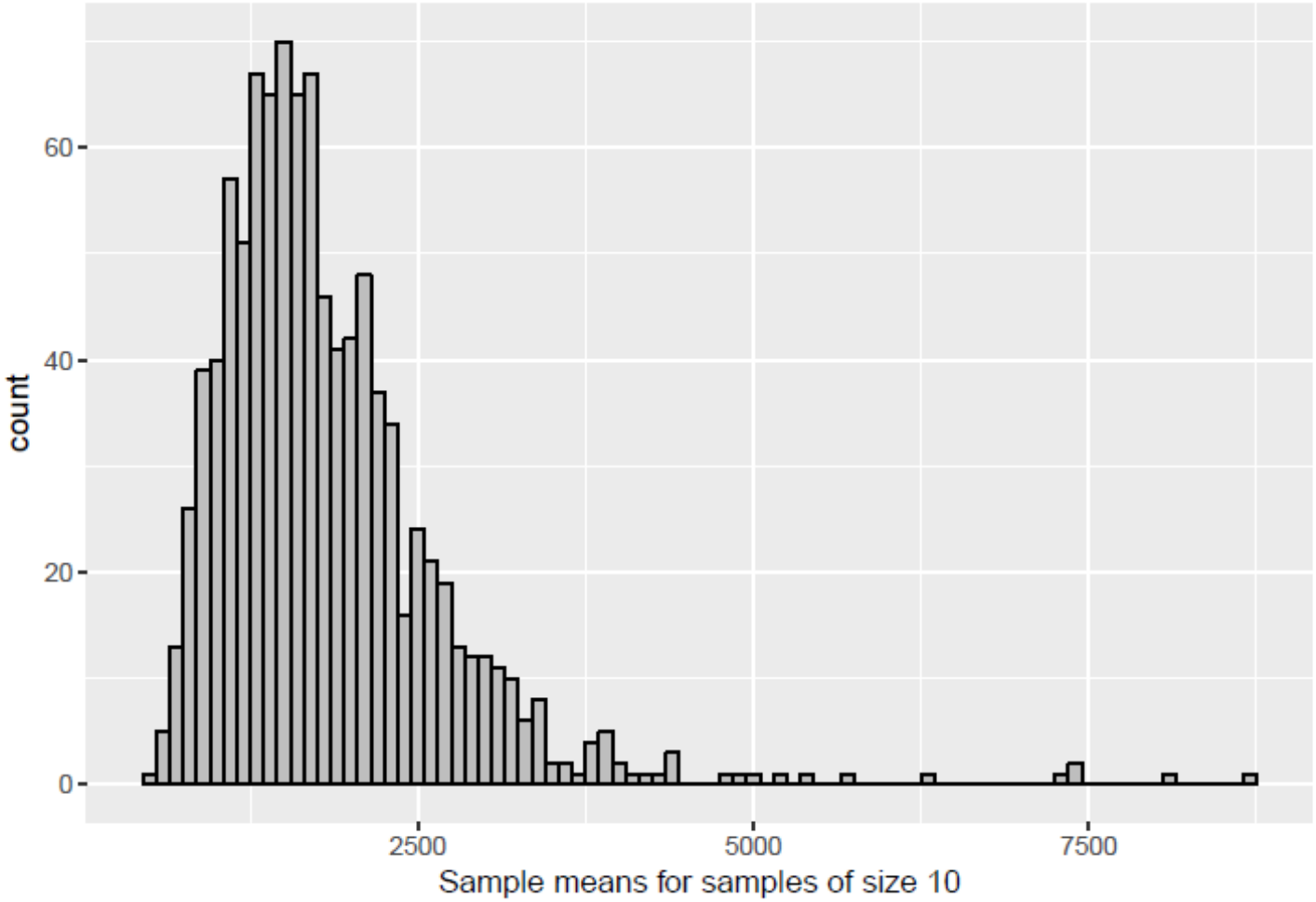
1. Heavily right-skewed.
2. With mean $1,853

Sampling distribution: Distribution of the sample mean.

(c) Estimate the sampling distributions of sample mean of paid claims by taking 1000 random samples of (i) size n=10 and (ii) size n=100 from the distribution in (a) and produce appropriate data summaries.

```
repetitions <- 1000
sim10 <- rep(NA, repetitions)
for (i in 1:repetitions)
{
  new_sim <- sample(AutoClaims$PAID,size = n)
  sim_mean <- mean(new_sim)
  sim10[i] <- sim_mean
}
sim10 <- data_frame(mean = sim10)
sim10 %>% ggplot(aes(x = mean)) +
  geom_histogram(binwidth = 100, colour = "black", fill = "grey") +
  labs(x="Sample means for samples of size 10")
```

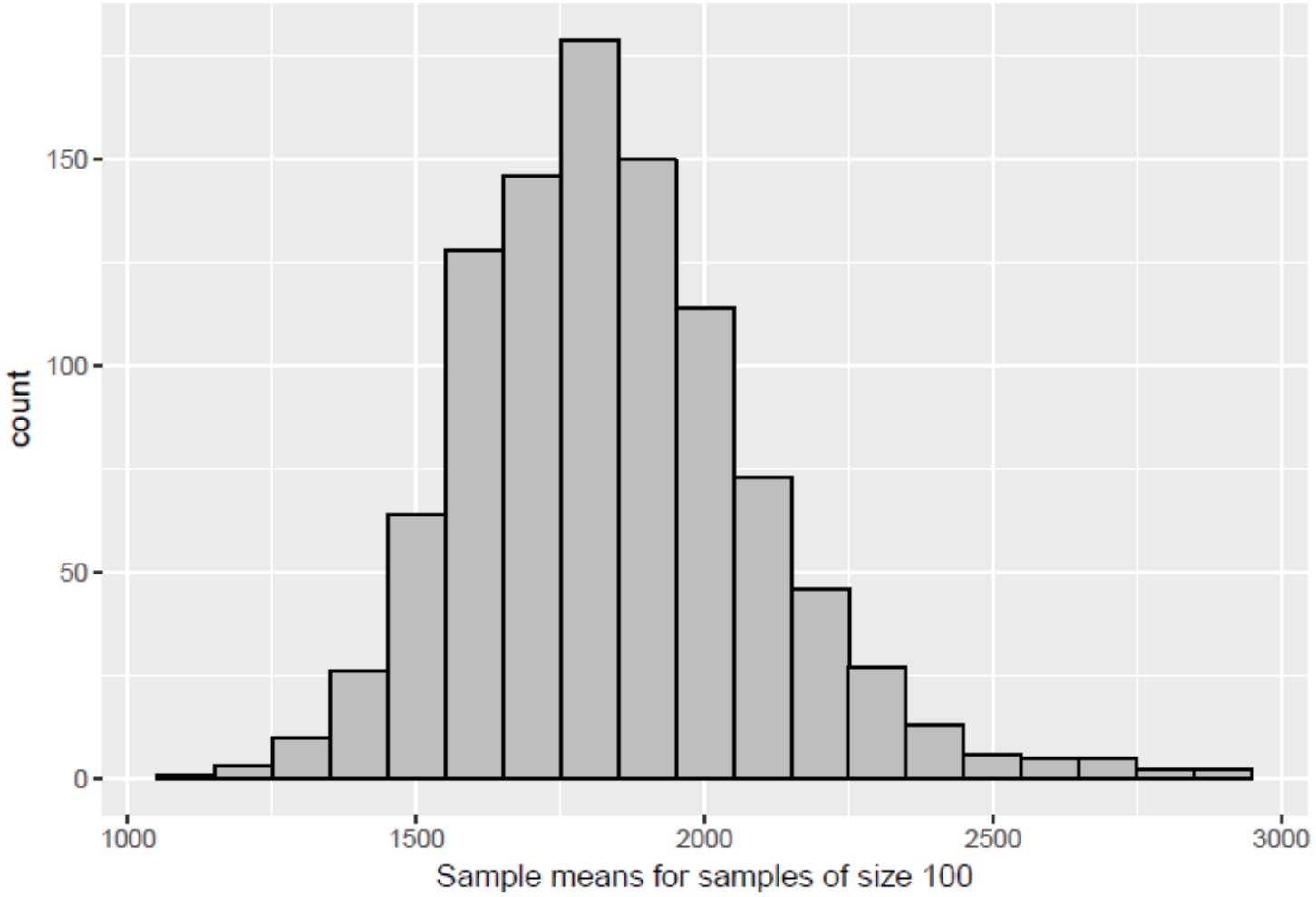
Sample size: n = 10




```
summary(sim10$mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   470.6  1275.0  1652.0  1823.4  2168.4  8707.0
```

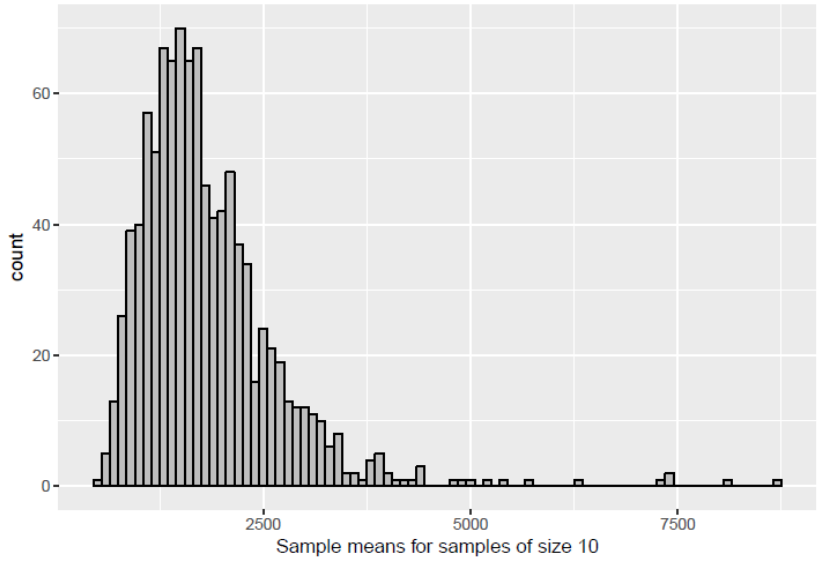
Sample size: n = 100




```
summary(sim100$mean)
```

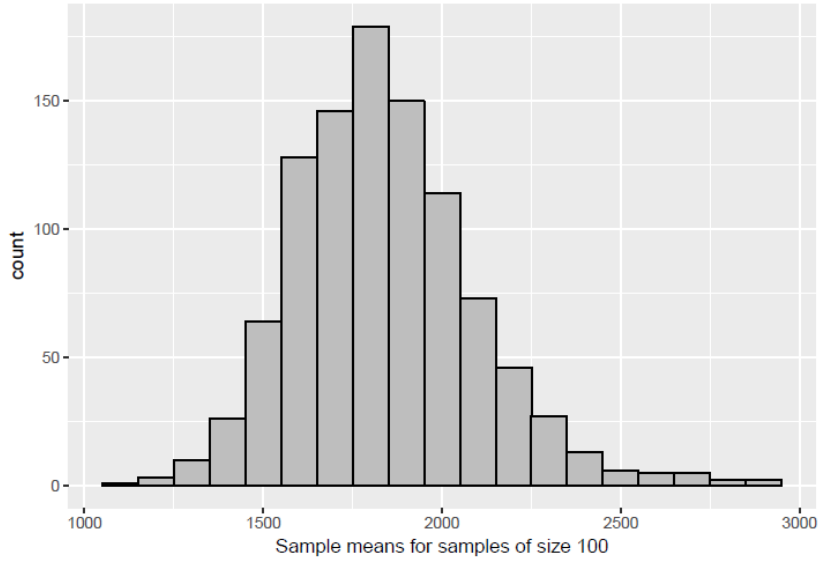
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1127    1664    1817    1840    1985    2882
```

- What do you notice between this two distributions?




```
summary(sim10$mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   470.6  1275.0  1652.0  1823.4  2168.4  8707.0
```




```
summary(sim100$mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1127    1664    1817    1840    1985    2882
```

1. They are both right-skewed.
2. The one for means of random samples of size 100 is not as right-skewed as the one for means of random samples of size 10.
3. Their means are close (i.e., mid $1,800s)
4. The distribution of means based on larger random samples (i.e., n=100) has much less variation than the distribution of means based on smaller random samples (i.e., n=10).

For those of you who are interested:

- If you keep increasing your sample size from n=1 to 100 to 1,000, etc.
- You will find that the sampling distribution you generated will become more and more symmetry.
- And with the center close to the population mean, in this case, around mid $1,800.
- This is true in general no matter how is your original population distributed.
- The property provides the intuition for the Central Limit Theorem, for which you will learn more about it in the second-year STA courses.

Confidence interval

- Before CI, how do we estimate the parameters:
- By **points estimate**: e.g. estimating the population mean by **sample mean**, estimating the population variance by the **sampling variance**.
- $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

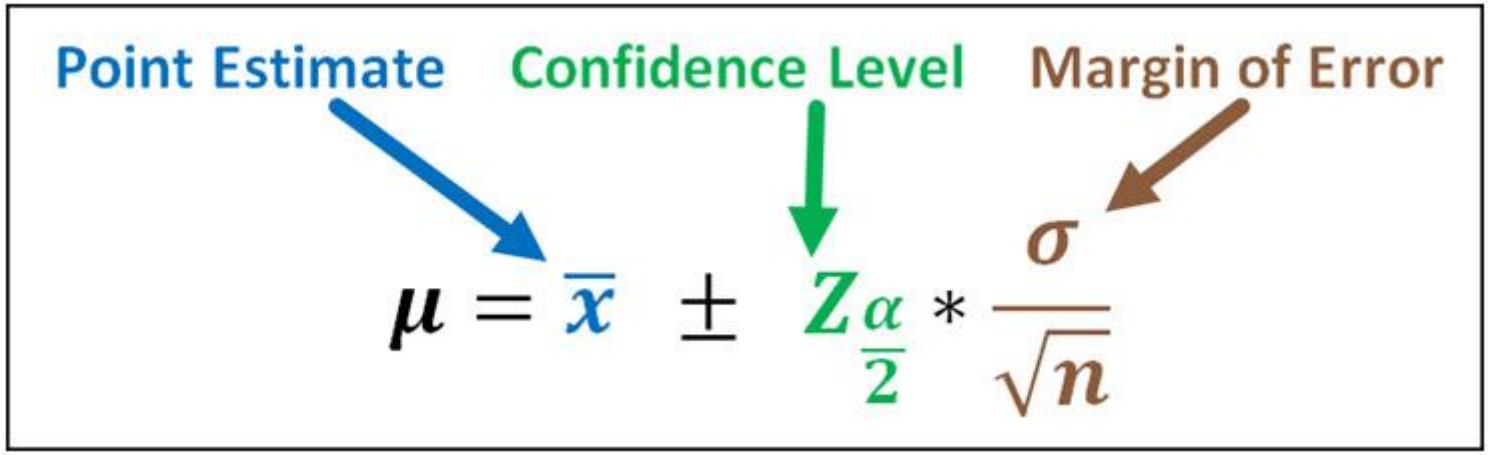
- Point estimate is a helpful way to estimate your population parameters, but it only gives you a limit amount of information on those population parameters.

Confidence interval

- **Purpose of CIs:** to obtain an estimate the **parameter** of your population that reflects sampling variability.
- E.g. Wish to estimate the proportion of people living in Toronto who use the TTC, number of coffees people in this class drink each week, etc.

Confidence interval

- An extension of point estimates.
- For now, you don't have to worry about the math, but you should learn how to interpret CI correctly.

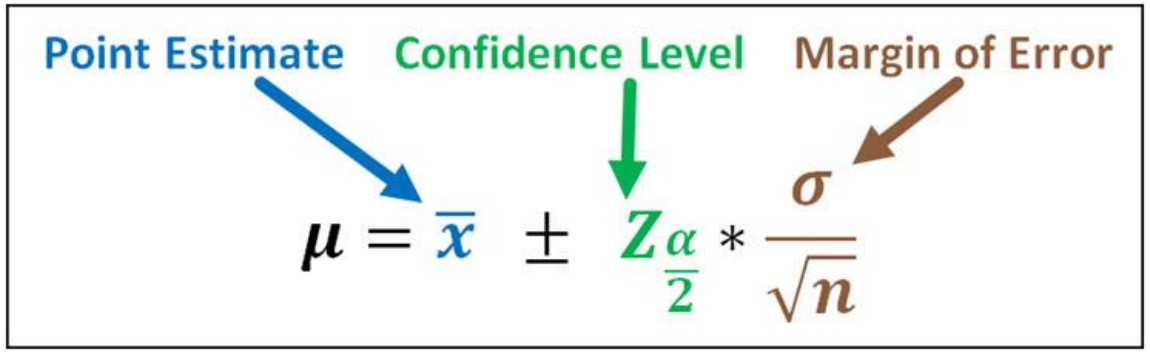


TIPS:

This formula can provide you a way to narrow your CI.

$$\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right)$$

- In general, the narrower you CI is, the more precise estimation it gives you.
- So, in the situation that you want to have a more precise estimation.
- You can:
- 1. Increase the sample size n
- 2. decrease the confidence level



say from 99% confidence to 90% confidence, then you will have a narrower CI.

The more confidence you are, the smaller the z-score is.

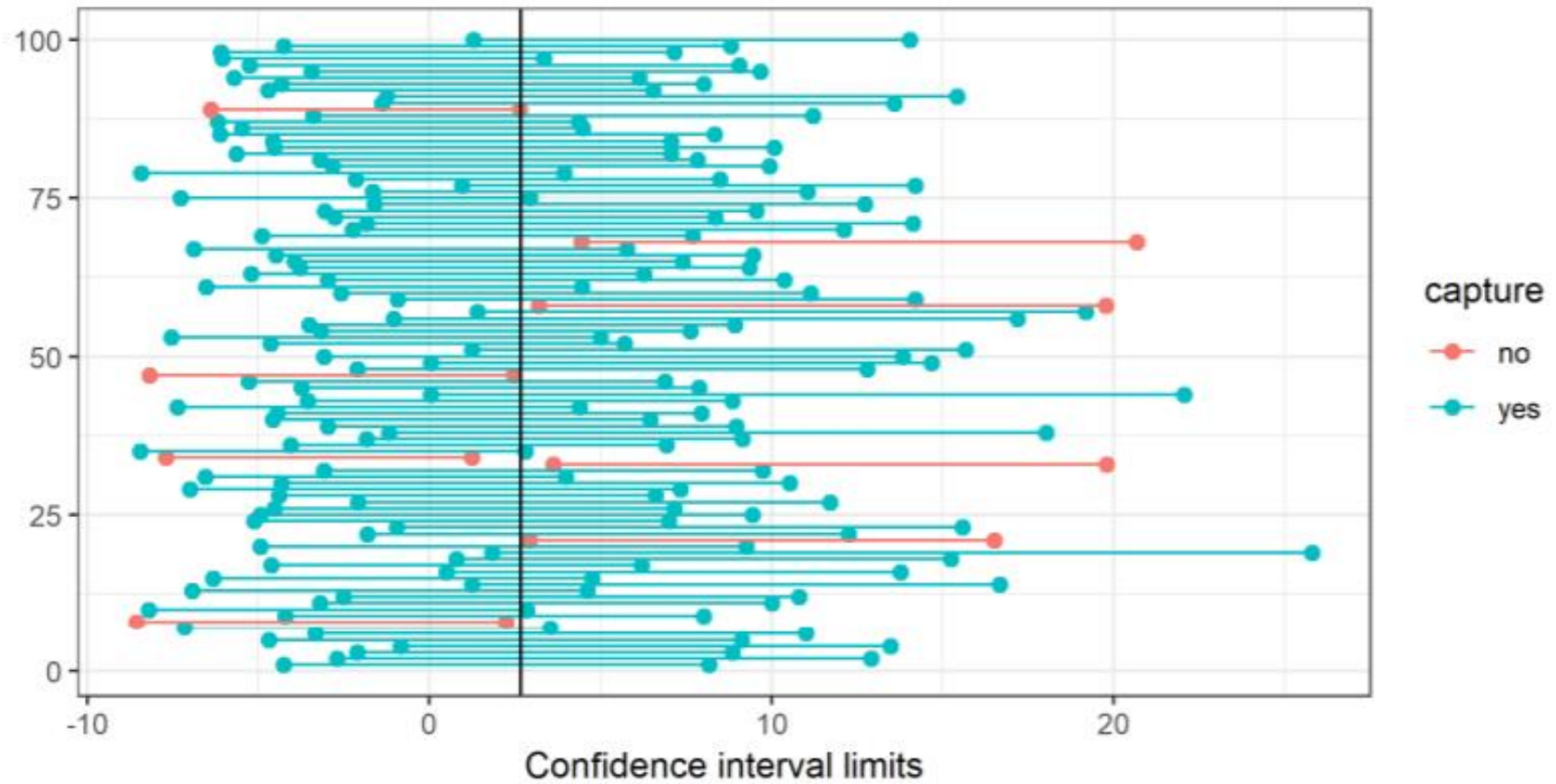
C	z*
99%	2.576
98%	2.326
95%	1.96
90%	1.645

Interpreting CI

- Suppose you have a 95% confidence interval: [165cm, 182cm], how do you interpret it?
- Do the following statements correct?
- 1. There is 95% chance that the true mean of students' height in UofT is in [165cm, 182cm].
- 2. The probability that the true mean of students' height in UofT is in [165cm, 182cm] is 0.95

- Both are **wrong**, it is very temping to make such statements.
- We don’t make probabilistic statement on CI!
- What is the correct interpretation?
- The true mean is either in [165cm, 182cm] or not in.
- So, the probability is either 0 or 1, but we don’t know about it yet.

100 bootstrap confidence intervals for the mean, each calculated from a random sample from the population of size 200



How many of the confidence intervals capture the population mean?

- What does a 95% confidence interval of students' height in Uoft mean?
- Each time you sample from your population, you get a subset of student in UofT, you calculate the sample mean, using the formula shown above, you calculate a range of prediction of average height, which is your CI.
- Suppose you repeat the process 100 times, approximately 95 times, the prediction interval you got will capture the true mean of students' height in UofT.

General steps to calculate the bootstrap confidence interval

- 1. Take a bootstrap sample of the data by sampling with replacement, the same number of observations as the original data.
- 2. For the bootstrap sample, calculate the statistic that estimates the parameter you are interested in.
- 3. Repeat steps 1 and many times to get a distribution of bootstrap statistics.
- 4. A 95% confidence interval for the parameter is the middle 95% of values of the bootstrap statistics.
- First three steps: calculate the sampling distribution, last step: calculate CI

How do we compute the CI in R

- A 95% confidence interval for the parameter is the middle 95% of values of your sampling distribution.
- In R, you first generate your sampling distribution (i.e. bootstrapping), then by using the function: **quantile(sample_means, c(0.025, 0.975))** you can get your 95% CI.

Q3 (c) (i)

(c) (i) Use R to find a 99% bootstrap confidence interval for the mean of mother's age. Use 5000 bootstrap samples. *NOTE:* More bootstrap samples is better, but if you find this times out or takes too long in RStudio Cloud, try using 1000 boostrap samples instead.

```
boot_means <- rep(NA, 5000)  # where we'll store the bootstrap means
sample_size <- as.numeric(Gestation %>% summarize(n()))
set.seed(50)
for (i in 1:5000)
{
  boot_samp <- Gestation %>% sample_n(size = sample_size, replace=TRUE)
  boot_means[i] <- as.numeric(boot_samp %>% summarize(mean(age)))
}
quantile(boot_means,c(.005,.995))
```

```
##     0.5%    99.5%
## 26.83871 27.69045
```

(ii) Suppose your confidence interval was (26.8, 27.7). Explain why the interpretation *"There is a 99% chance that the true mean of a mother's age at the time this sample was taken is between 26.8 and 27.7 years."* is *INCORRECT*. What is a correct interpretation?

Correct way on interpreting the results

The true mean age of mothers in 1961/62 is unknown, but it's not random. In other words, it's a fixed, but unknown constant. Therefore it either is or isn't in this interval (i.e., the chance is either 0% or 100%). We just do not know either way.

We can conclude that we are 99% *confident* that the true mean mother's age in 1961/62 was between 26 and 27 years. We are confident in this because the method we used to obtain the interval will produce intervals that do include the true value of the parameter of interest for 99% of the possible samples we could take.

After you compute your CI

- Always check that your CI range makes sense.
  - E.g. if reported the CI for a proportion, it needs to be bounded by zero and one. You can’t have a probability less than zero or greater than 1!

Presentation exercise

- Regarding **question 3** on your homework, pick one of the following topics and prepare a **5** minutes presentation.
- Hand in your draft for your presentation.

- ***Topic one:***
- Describe the content of the graph (mention the variable of interest)
  - What does the x-, y-axis represent?
  - Describe the distribution (range, center, symmetry, skewness, number of points).
- How was one dot calculated:
  - how a single bootstrap sample is produced (e.g. size, with/without replacement)?
  - what statistic did you calculate from this particular bootstrap sample?
- How can the generated distribution of mean age be used for inference?
  - Since the observed data were generated … (this was a key word from last week), we can … (a key word from today's vocab list) from the observed data by sampling with replacement.
  - In other words, if the data resemble the … (a key word from today's vocab list), the bootstrap samples will also resemble the (the same key word from today's list).
  - Using the statistic calculated from each bootstrap sample, we can obtain a distribution of sample statistic and it gives an estimate of the … (a key phrase from today's vocab list) of the statistic.

- ***Topic two:***
- Rationale of using the bootstrap sampling distributions:
  - Does distribution of bootstrap sample statistic tend to capture the population value (mean/median)?
  - Where does the population value tend to be in the range of the bootstrap distribution?
- Construct the confidence interval
  - What is the range of values (in terms of percentile) taken to construct the 90% CI?
  - Describe how you did the above in R (are there ties? Was it easy or difficult to do this in R and why?)
- State the interval with reference to the data and variable (E.g. “A 90% CI for the mean (median) of the mother’s age is ….”).
- Interpret the interval you produced.
- (Optional) Could you check if the interval produce is indeed the 90% CI? Why or why not?
  - Do you know the population parameter value (mean)?
  - Do you need to do more calculations?
  - How many times do you check if the population mean is captured by the CI?

- ***Topic three:***
- Describe how to produce the plot
  - How a single bootstrap sample is produced (e.g. number of data points used, with/without replacement)?
  - What statistic did you calculate from each bootstrap sample?
- Describe the content of the graph (mention the variable of interest)
  - What does the x-, y-axis represent?
  - Describe the shape of the distribution (range, center, symmetry, skewness, number of points).
- Rationale of using the bootstrap sampling distributions:
  - Does distribution of bootstrap sample statistic tend to capture the population value (median)?
  - Where does the population value tend to be in the range of the bootstrap distribution?
- Construct the confidence interval
  - What is the range of values (in terms of percentile) taken to construct the 99% CI?
  - Describe how you did the above in R (are there ties? Was it easy or difficult to do this in R and why?)
- State the interval with reference to the data and variable (E.g. “A 99% CI for the median of the mother’s age is ….”).
- Interpret the interval you produced.