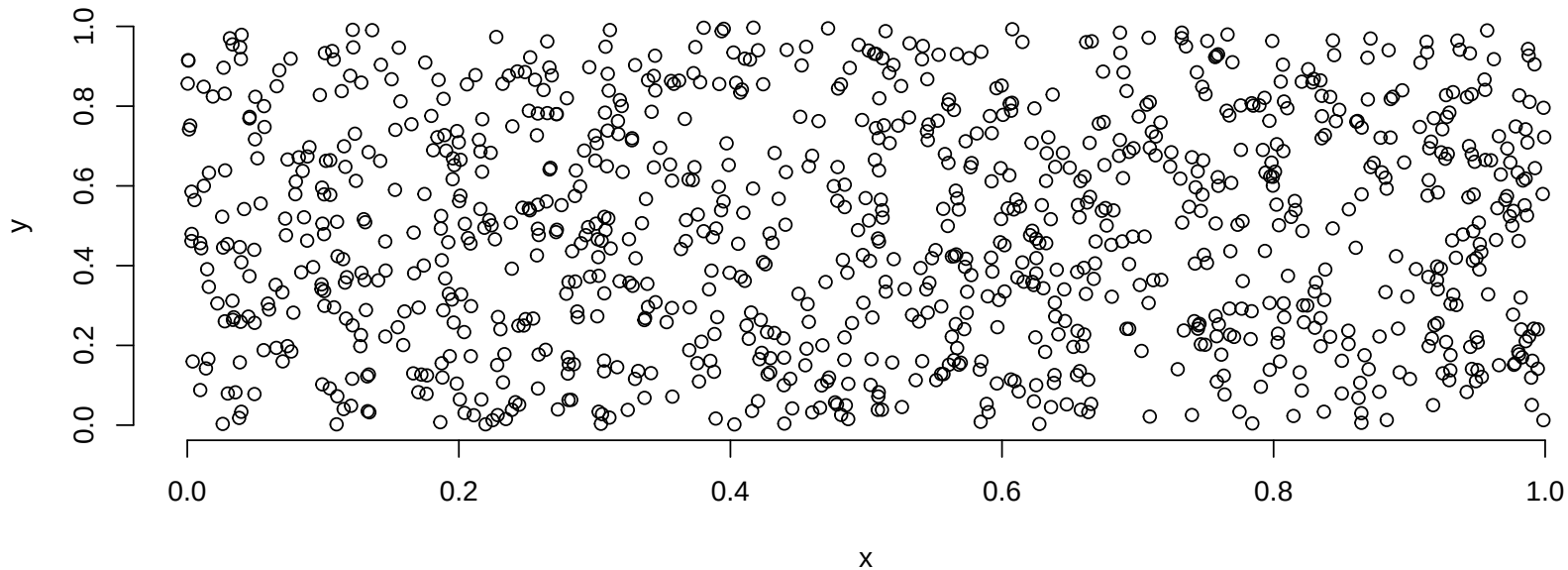
Confidence Interval Estimation

Ryan Miller

Grinnell College
Statistics

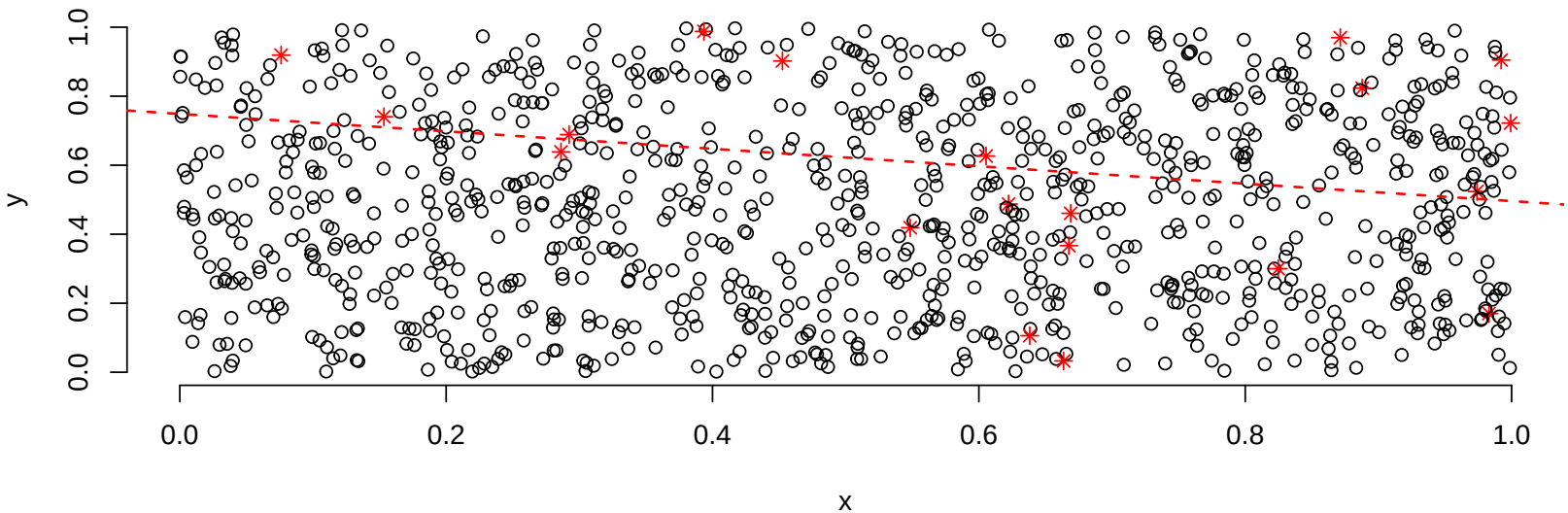
Motivating Example

Consider a *population* containing $N = 1000$ members, with variables X and Y that are *not related* ($\rho = 0$):



Motivating Example (cont.)

Here is a *random sample* ($n = 20$) from this population (sampled cases are colored in red), the sample correlation is $r = -0.245$:



Motivating Example (cont.)

We observe a *weak negative correlation* in the sample, but to reach a conclusion about the population we need to test $H_0 : \rho = 0$, which we can do using a T-test:

$$T = \frac{r - \rho_0}{SE} = \frac{-0.245 - 0}{SE}$$

1. Suppose this test yields a *statistically significant result*, what are two factors that could have driven this result?
2. Suppose the test *fails to provide evidence against* H_0, what are two factors that could have driven this result?

Effect Size vs. Statistical Significance

- A statistically significant result occurs when the difference between the observed effect and the null hypothesis is large relative to the amount of sampling variability
  - The p-value alone doesn't tell us whether the size of the observed effect was large, or whether the amount of sampling variability was small
- It is possible for a small effect to produce a statistically significant result if the SE is small
  - Similarly, it is possible for a large effect to not reach statistical significance if the SE is large

Effect Size and Decision Making

Suppose you are evaluating a new policy that is extremely expensive. What would you recommend as a next step if each of the following results were observed?

- Statistically significant result ($p \leq 0.05$)
  - Scenario 1: The effect size is small
  - Scenario 2: The effect size is large
- Lack of statistical significance ($p > 0.05$)
  - Scenario 1: The effect size is small
  - Scenario 2: The effect size is large

Effect Size and Decision Making (solution)

- Statistically significant result ($p \leq 0.05$)
  - Scenario 1: The effect size is small - *abandon the policy, costs likely outweigh the benefits*
  - Scenario 2: The effect size is large - *adopt the policy, it has evidence supporting it and a large benefit*
- Lack of statistical significance ($p > 0.05$)
  - Scenario 1: The effect size is small - *abandon the policy, its effect seems inconsequential*
  - Scenario 2: The effect size is large - *continue studying the policy with a large sample size*

Point vs. Interval Estimation

When reporting an effect size, such as a difference in means or a correlation, there are two types of estimates that statisticians can provide:

1) **Point estimate** - a *single number* that is the *best guess* of population-level effect. For example, the sample mean $\overline{x}$ is a point estimate for the population's mean, μ.

Point vs. Interval Estimation

When reporting an effect size, such as a difference in means or a correlation, there are two types of estimates that statisticians can provide:

1) **Point estimate** - a *single number* that is the *best guess* of population-level effect. For example, the sample mean $\overline{x}$ is a point estimate for the population's mean, μ.
2) **Interval estimate** - a *range of numbers* that represent *plausible values* of the population parameter. Interval estimates usually have the form: Point Estimate $\pm$ Margin of Error

For interval estimates, is it possible to use a margin of error such that the interval always contains the truth? What would the downside of this be?

Interval Estimation

When developing an interval estimate there are two competing goals:

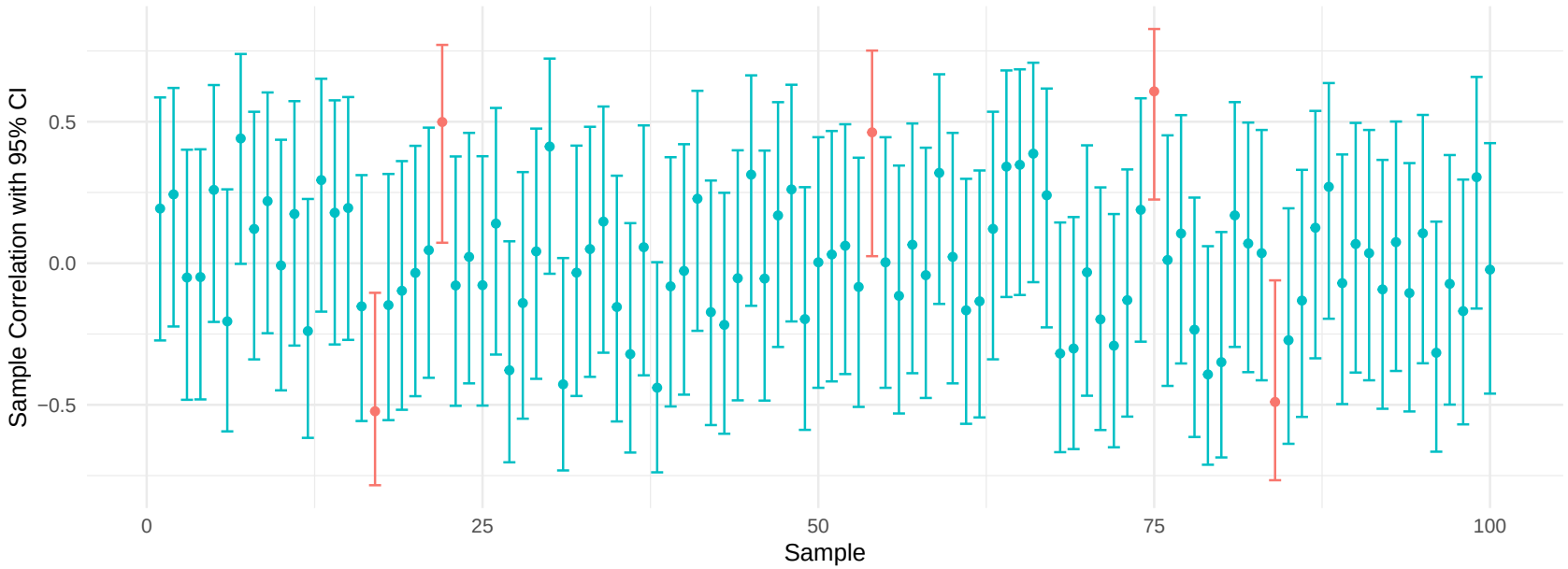
1. **Accuracy**: We want the interval to have a high chance of containing the true population parameter, something that is usually achieved by a wider interval
2. **Precision**: We want the interval to have a small enough margin of error to be useful - An interval suggesting a candidate could receive between 30% and 70% of the vote in an election isn't very useful.

Confidence intervals

- **Confidence intervals** are a special type of interval estimate whose margin of error is calibrated to achieve a long-run "success rate" known as the *confidence level*
  - Here "success" is defined as the interval containing the true value of the parameter being estimated
- For example, the procedure used to construct the margin of error for a 95% confidence interval will make the interval wide enough to contain the parameter of interest in 95% of different random samples (or study replications)

Confidence intervals

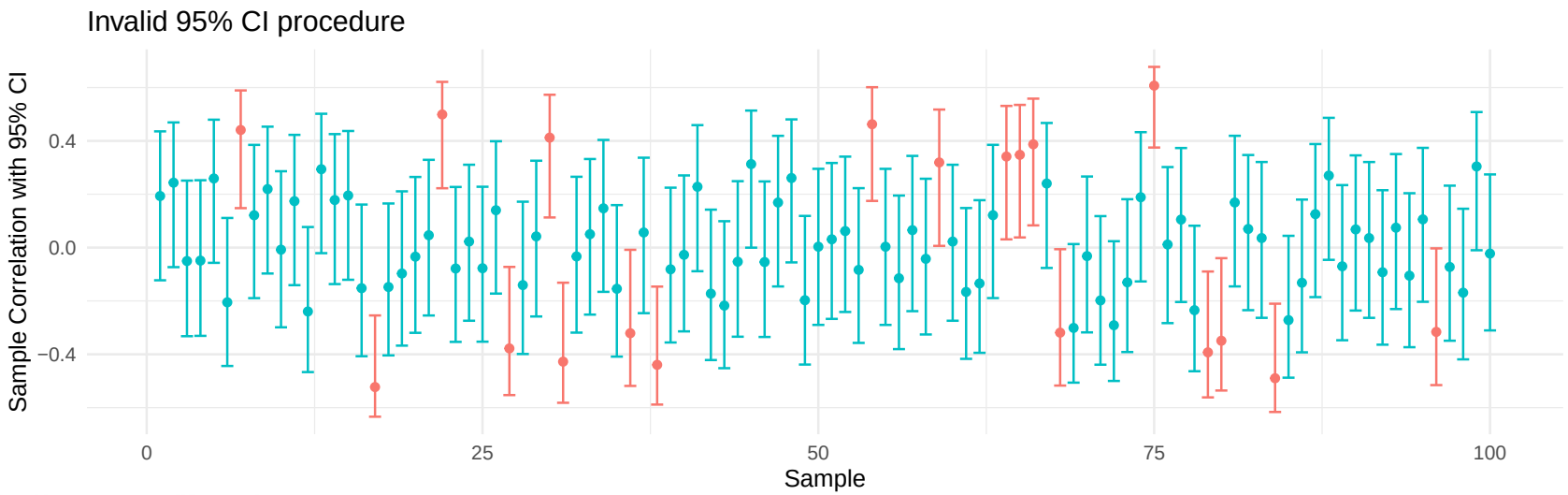
Below are 100 different 95% CI estimates that were made using data from a different random sample of $n = 20$ from a population with $\rho = 0$:



Notice how 5 of 100 samples produced a 95% CI that failed to contain the true population-level correlation!

Confidence Interval Validity

- A **valid** confidence interval procedure will produce intervals where the stated confidence level matches the procedure's long-run success rate
  - The procedure used to make the intervals on the previous slide was valid
- An **invalid** procedure will produce intervals that fail to contain the true parameter more often than they should, for example the procedure shown below:



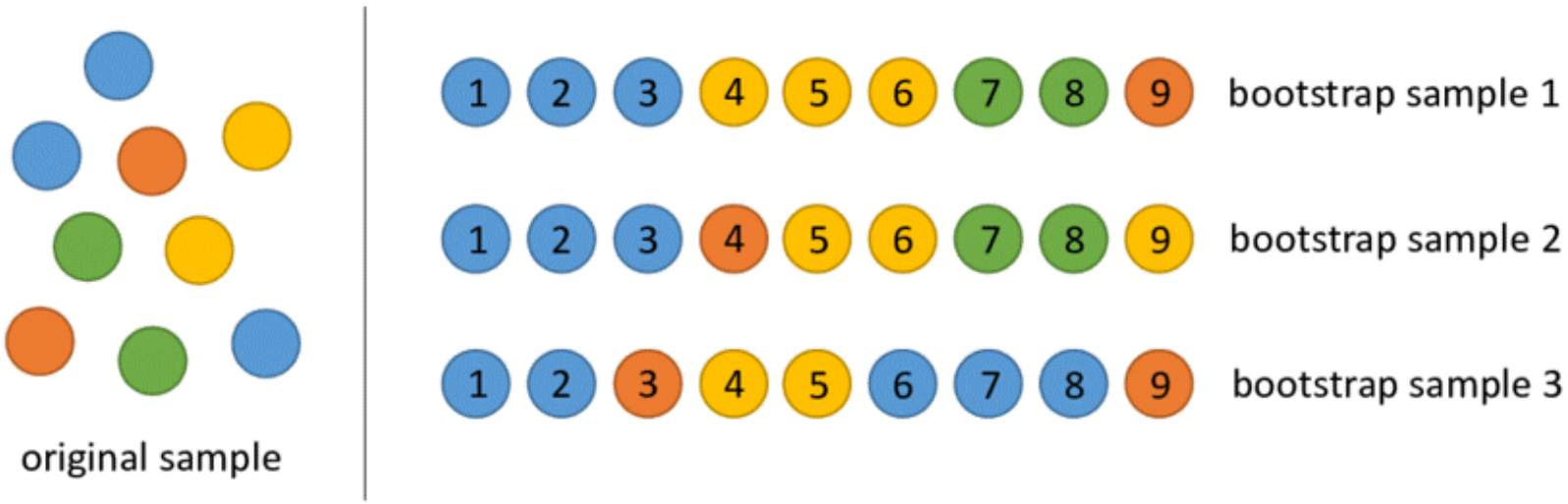
Confidence Interval Estimation

In this class, we will study two methods for making a valid confidence interval using an observed effect:

1. **Bootstrapping** - Here we mimic repeating our study thousands of times to judge the distribution of likely effects
2. **Probability Models** - Here we rely on statistical theory (Central Limit theorem) to add a calibrated margin of error to our point estimate

Method #1 - Bootstrapping

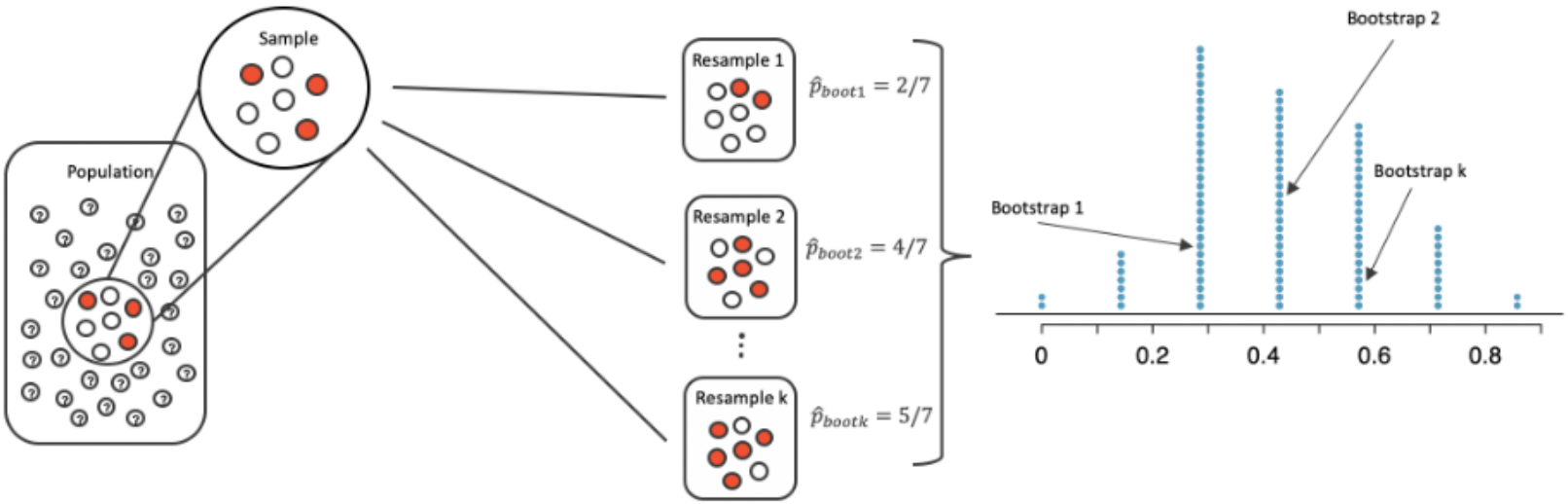
In bootstrapping, we sample from our original sample with replacement to create *bootstrap samples*:



Notice that some cases will appear several times in a bootstrap sample, while others will be absent entirely.

Bootstrapping

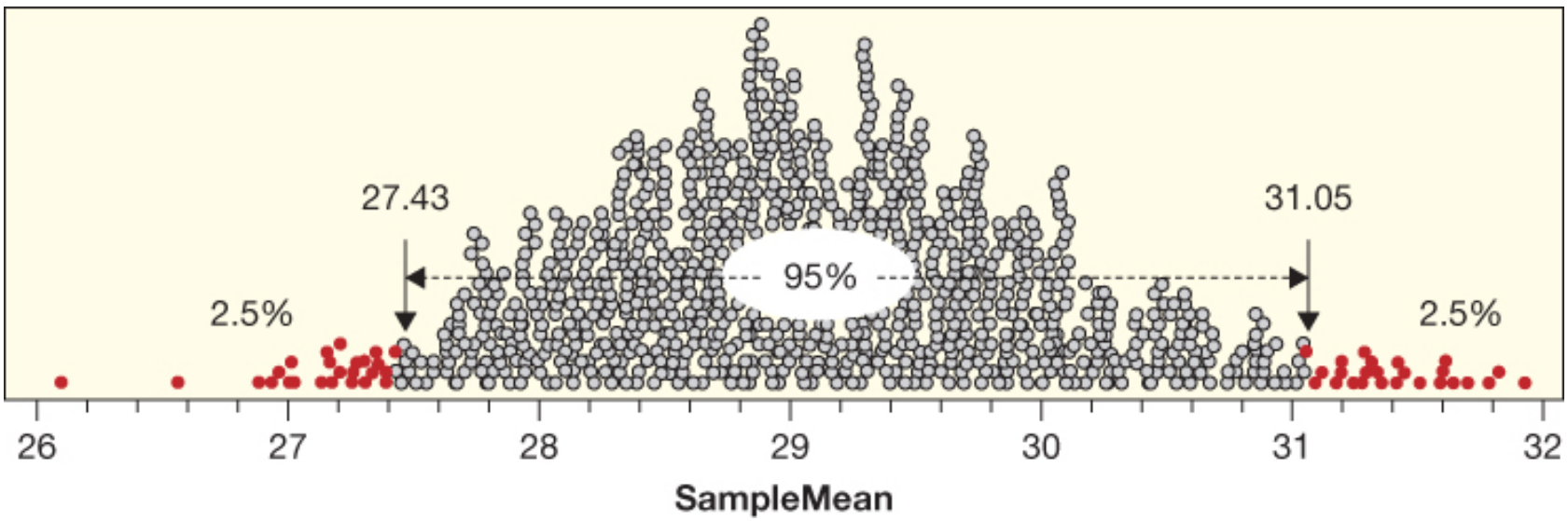
We then calculate the effect in each bootstrap sample, thereby giving us a distribution of plausible effects:



This is known as the **bootstrap distribution**.

Bootstrapping

We can construct a *valid* $P\%$ confidence interval using the *bootstrap distribution* by taking the middle $P\%$ of bootstrapped estimates:



Practice #1

The data below were recorded by a commuter in the greater Toronto area:

https://remiller1450.github.io/data/CommuteTracker.csv

1. Use StatKey to find a 99% percentile bootstrap interval for the correlation between average moving speed and total commute time.
2. Considering the shape of the bootstrap distribution, why might bootstrapping be a good idea here? *Hint*: Think about the probability distribution we use when testing whether a correlation is zero. How does its shape compare to the bootstrap distribution?

Practice #1 - solution

1. The 99% percentile bootstrap confidence interval is approximately (-0.929, -0.760).
2. The bootstrap distribution shows some amount of right-skew, suggesting the sampling distribution isn't symmetric, so a Normal or t-distribution wouldn't be a good model.

Method #2 - Probability Models

To calculate a *valid* confidence interval using a probability model, we must add a properly calibrated margin of error to the point estimate. We can do this by decomposing the margin of error into the *standard error* of the point estimate and a *calibration value*:

$$\text{Point Estimate} \pm c \cdot SE$$

- Central Limit theorem has given us the standard error of most point estimates (the table of our exam coversheets)
- The calibration value, c, can be found using an appropriate probability model

Confidence Intervals using Probability Models

For a **single proportion** or **difference in proportions** the value of c can be chosen from a Standard Normal distribution:

- $c = 1.96$ corresponds to the middle 95% of the $N(0,1)$ distribution
  - This choice of c will produce a valid 95% confidence interval (when the Normal distribution is an accurate probability model)
- $c = 1.28$ corresponds to the middle 80% of the $N(0,1)$ distribution, which can be used to calibrate an 80% confidence interval
  - Notice how this will lead to a smaller margin of error (relative to $c = 1.96$)

Practice #2

We previously heard about a study by Johns Hopkins University Hospital where 31 of 39 babies born in their facilities at 25 weeks gestation (15 weeks early) went on to survive. Suppose our goal is to find an interval estimate for the proportion of babies born under similar circumstances in similar hospitals that will survive.

1) In this application we are trying to estimate p, a population proportion, using $\hat{p} = 31/39$. What is the Standard Error of $\hat{p}$? *Hint*: The formula is on our exam coversheet.
2) Use StatKey to determine the value of c used to calibrate a 99% CI involving these data.
3) Find the endpoints of the 99% CI.
4) If the confidence level is decreased from 99% to 95% what will happen to the width of the interval?

Practice #2 - solution

1) $SE = \sqrt{\frac{(31/39)\cdot(1-31/39)}{39}} = 0.0646$
2) $c = 2.576$
3) $31/39 \pm 2.576 \cdot 0.0646 = (0.628, 0.961)$
4) The width would decrease, but we'd be less confident that the interval contains the truth.

Practice #2 (cont.)

Recall that we checked the *large count condition* when deciding whether we could use the one-sample Z-test. Notice that for these data, we only observed 8 deaths.

1) Use StatKey to find a 99% bootstrap confidence interval using at least 2000 bootstrap replications.
2) How do the endpoints of this interval compare with those you found using a Normal probability model? Why might they differ?

Practice #2 (cont.) - solution

1. 99% CI estimate of (0.615, 0.936) using bootstrapping
2. The bootstrap distribution is skewed left, so taking the middle 99% takes us further into the left tail and not as far into the right tail (relative to the Normal distribution, which is symmetric)

One and Two-Sample T-intervals

For a **single mean** or a **difference in means** the value of c must be chosen from a T-distribution because the SE of these statistics requires us to estimate the *population's standard deviation* using the sample standard deviation, thereby introducing additional uncertainty

- If $df = 5$, $c = 2.571$ corresponds to 95% confidence
  - Notice how this is a much larger multiplier than $c = 1.96$, which corresponded with 95% confidence when the Normal distribution was an accurate probability model
- If $df = 30$, $c = 2.042$ corresponds to 95% confidence

Practice #3

- We previously looked at a study of $n = 12$ professional swimmers who swam a 1500m time trial with and without a scientifically designed wetsuit.
  - Suppose we'd like to use these results to estimate the expected improvement in 1500m swim velocity due to the wetsuit.
- In this study, the average *paired difference* was 0.0775 m/s and the standard deviation of these differences was 0.022.
  - Use the probability model approach to find a 95% confidence interval estimate for the expected difference in velocity in the population.

Practice #3 - solution

To calculate the interval we need the two components of the margin of error, c and SE:

- c is given by the middle 95% of a T-distribution with $df = n - 1 = 11$, which is 2.201 according to StatKey
- $SE = s/\sqrt{n}$ for a single mean, which is $0.022/\sqrt{12} = 0.00635$

So, the 95% CI is given by

$$0.0775 \pm 2.201 \cdot 0.00635 = (0.064, 0.091)$$

Thus, we conclude with 95% confidence that the expected improvement in swim velocity when wearing the wetsuit is between 0.064 m/s and 0.091 m/s. Since the entire interval is above zero, it's implausible that the wetsuit makes no difference.

Confidence Intervals as Hypothesis Tests

- In hypothesis testing, we saw there could be two explanations for a significant result:
  - Large effect size, or low sampling variability
- Using confidence intervals, we do not need to perform a follow-up analysis to distinguish between these explanations
  - The interval's width reflects sampling variability, and the interval's values reflect plausible effect sizes
- Furthermore, confidence intervals and hypothesis tests often use the same underlying probability model in different ways, thus:
  - If a 95% CI doesn't contain $H_0 : p = 0.5$ the p-value is less than 0.05
  - Similarly, if a 99% CI doesn't contain $H_0 : p = 0.5$ the p-value is less than 0.01

Confidence Interval Misconceptions

The following statements are all **false**. Identify the problem with each:

1. If we calculate a 95% CI of (2.5, 4), there is a 95% chance that the true population-level effect is between 2.5 and 4.
2. If we calculate separate 95% confidence intervals for the means of two different groups, and the intervals overlap, then the difference in means is not statistically significant (at $\alpha = 0.05$).
3. If we calculate a 95% CI of (-3, 3), we estimate that 95% of the sample data is between -3 and 3.

Confidence Interval Misconceptions - solutions

1. The confidence level is a long-run success rate of the interval construction procedure, it doesn't tell you the capture rate of any individual interval
2. Values near the edge of an interval are less likely than values near the center, so some overlap doesn't imply that the corresponding combination of values is statistically plausible. To determine significance, we need a confidence interval or hypothesis test for the difference itself.
3. Confidence intervals always pertain to an unknown population parameter, they do not describe the spread of individual data-points