Confidence Intervals

Part 1 - Standard Error and Bootstrapping

Ryan Miller

Grinnell College
Statistics

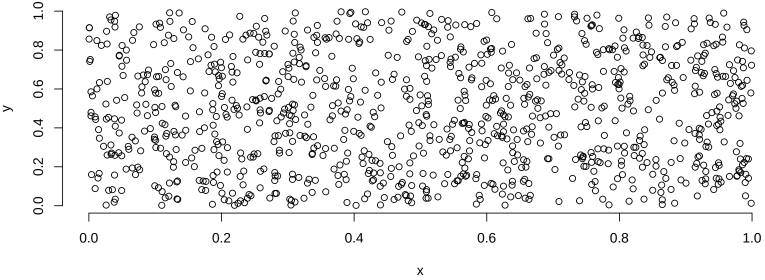
Introduction

To begin today we'll define two related but different terms:

- **Standard deviation**: A description of the variability of *cases* around their mean (for a quantitative variable)
  - We can calculate the sample standard deviation directly from our data with something like: `sd(data$variable)` in `R`
- **Standard error**: A description of the variability of *estimates* around a true parameter (ie: the population parameter of interest)
  - We cannot calculate standard error in a similar manner because our sample data provides a only single *point estimate* of the population parameter of interest
    - For example, the sample mean is just a single estimate the population's mean

Sampling Variability

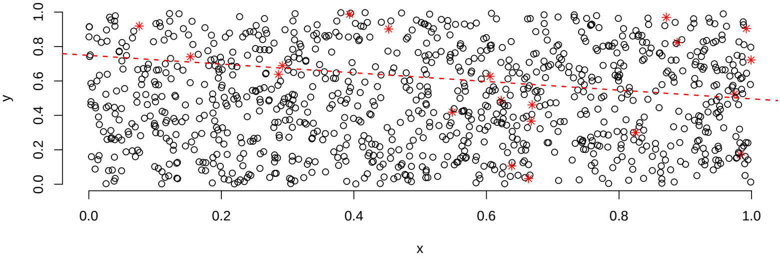
The scatterplot below depicts a *population* ($N = 1000$) where the variables X and Y are *not related* (ie: $\rho = 0$):



Grinnell College
Statistics

Sampling variability

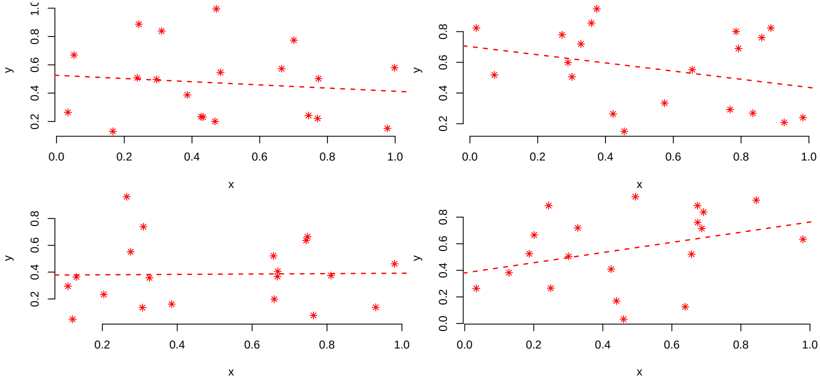
Here is a *random sample* ($n = 20$) from this population (sampled cases are colored in red), the sample correlation is $r = -0.245$:



So, the sample data suggest a *weak negative correlation* despite these variables having *no correlation* in the population

Sampling variability

Below are another four random samples (each using $n = 20$):



Across these samples, the observed sample correlations range from $r = -0.31$ (top right) to $r = 0.35$ (bottom right)

Grinnell College
Statistics

A few questions

Now consider a population where you *do not know* the true correlation between X and Y:

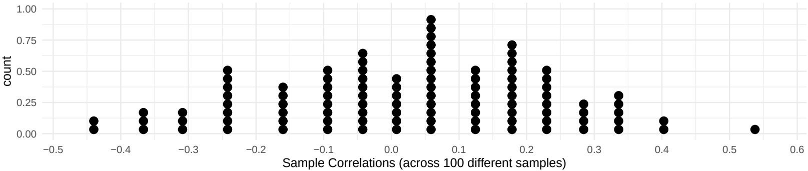
1) If the you take a *single sample* and find a sample correlation of $r = 0.33$, how *confident* can you be that X and Y are related in the population?
2) If you take *100 different samples* and find sample correlations ranging from $r = 0.25$ to $r = 0.35$, how *confident* can you be that X and Y are related in the population?
3) If you take *100 different samples* and find sample correlations ranging from $r = -0.15$ to $r = 0.45$, how *confident* can you be that X and Y are related in the population?

Answers

1) Not confident, it seems entirely possible that $\rho = 0$ but we happened to see an unusual sample where $r = 0.33$ purely by chance.
2) Very confident, none of the samples had correlations near zero and it'd be very unlikely for all 100 of our samples to have sample correlations this large by chance.
3) Not confident, it seems like $\rho = 0$ is plausible since some of the samples had correlations near zero.

Sampling distributions

The distribution of *all possible estimates* that could be observed when sampling is known as the **sampling distribution**:



- **Standard error** is the *standard deviation* of the values in a sampling distribution
  - It allows us to more confidently assess whether a point estimate reflects a real characteristic of the population
    - For example, if we observe $r = 0.4$ with a standard error of 0.1, we'd be confident a relationship exists, but if we observe $r = 0.4$ with a standard error of 0.3 we would not be confident

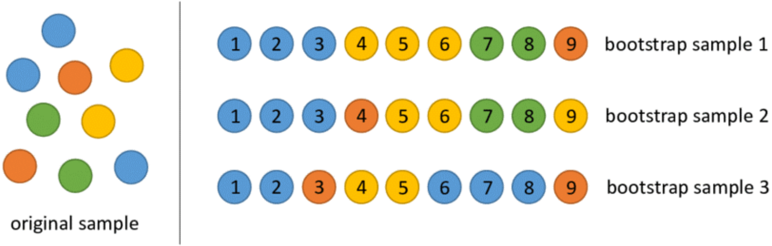
Grinnell College
Statistics

Bootstrapping

- Getting an accurate picture of a sampling distribution requires many different random samples from the population
  - It's not feasible to re-do a study hundreds of times just to assess sampling variability
- However, **bootstrapping**, or re-sampling cases from the original sample with replacement can be used to estimate sampling variability
  - Bootstrapping will only work when your sample size is large enough ($n \geq 30$) for your the cases in your sample to exhibit a similar amount of variability as the cases in your population
  - It's a common misconception that bootstrapping is great when you have a small sample size

Bootstrapping

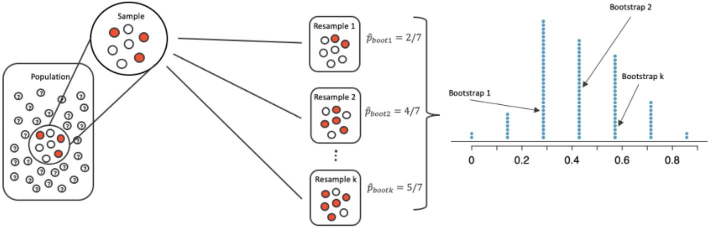
Re-sampling with replacement means that some cases in the original sample will appear more than once in a bootstrap sample:



This is necessary to generate synthetic samples of size n (you should recall that n is an important contributor to *sampling variability*)

Bootstrapping

The purpose of bootstrapping is to use your original sample to approximate the sampling distribution:



- Each *bootstrap sample* produces a *bootstrap estimate*, and these are used to form the **bootstrap distribution**
  - The bootstrap distribution is a direct analog of the sampling distribution, though it will be centered at the original estimate rather than the population parameter

Practice

- Basketball pundits will sometimes describe certain games with phrases like "the teams got into a shootout"
  - But is there a correlation between the number of 3-pt attempts of one team and their opponent?
    - We'll load data from the Golden State Warrior's record setting 2015-16 season into an online app named "StatKey"

Click here to download that data as a CSV
Click here to open StatKey

Question: Using the Bootstrap CI for correlation StatKey menu, find the standard error of correlation coefficient between "FG3A" and "OppFGA3A". Do you think the weak correlation observed in these data can be explained by *sampling variability*?

Statisticians consider two types of estimates for an unknown population parameter:

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

Statisticians consider two types of estimates for an unknown population parameter:

1) **Point estimate** - a *single number* that is the *best guess* for what the population parameter is. 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

Confidence intervals

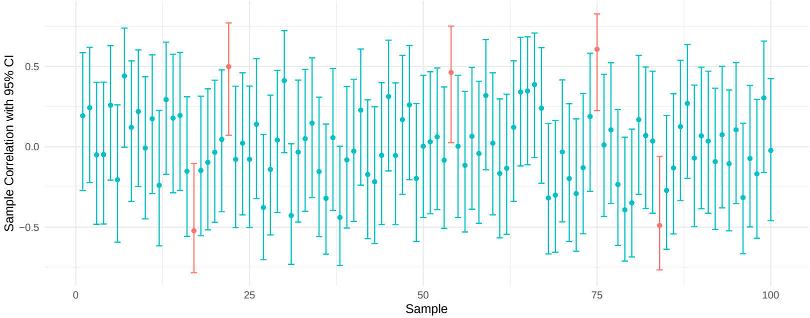
- A **confidence interval** is an interval estimate whose margin of error is based upon a procedure with a long-run "success rate" known as a *confidence level*

Confidence intervals

- A **confidence interval** is an interval estimate whose margin of error is based upon a procedure with a long-run "success rate" known as a *confidence level*
  - A 95% confidence interval was created using a procedure that will succeed in containing the true population parameter in 95% of different random samples (or study replications)
  - The confidence level does not describe the likelihood that particular interval succeeds, instead it describes the estimation procedure's long-run success rate

Confidence intervals

Shown below are 95% CI estimates from 100 different random samples ($n = 20$) drawn from a population with correlation of $\rho = 0$



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

Bootstrapping and the 2-SE method

A *valid* 95% CI requires a margin of error that calibrated to capture the truth in 95% of different random samples:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

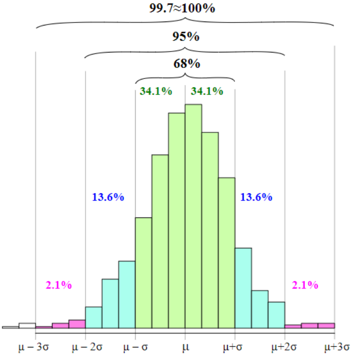
When the *sampling distribution* is symmetric and bell-shaped, the 95% rule produces valid 95% CIs:

$$\text{Point Estimate} \pm 2 * SE$$

- The **standard error**, or SE, can be found by bootstrapping
- Bootstrapping can also help us judge whether the sampling distribution is likely to be bell-shaped

The 95% rule

The SE multiplier of 2 and the 95% rule arise from the distribution of values in the **Normal distribution**:



What SE multiplier can be used for a 99% confidence interval?

Practice

A study conducted by Johns Hopkins University Hospital found that 31 of 39 babies born in their facilities at 25 weeks gestation (15 weeks early) went on to survive. Our goal is to estimate 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 $\hat{p} = 31/39$. Use StatKey to find the bootstrapped standard error of $\hat{p}$
2) Using the bootstrapped SE and the point estimate, construct a 95% confidence interval estimate for p
3) Interpret your 95% confidence interval estimate

Practice (solution)

1) The bootstrapped SE is approximately 0.066
2) The 95% CI for p is $\hat{p} \pm 2 * SE =$
   $31/39 \pm 2 * 0.066 = (0.663, 0.927)$
3) We can be 95% confident the survival rate for babies born in comparable circumstances is between 0.663 and 0.927. This represents a range of plausible values that we are confident will contain the true proportion.

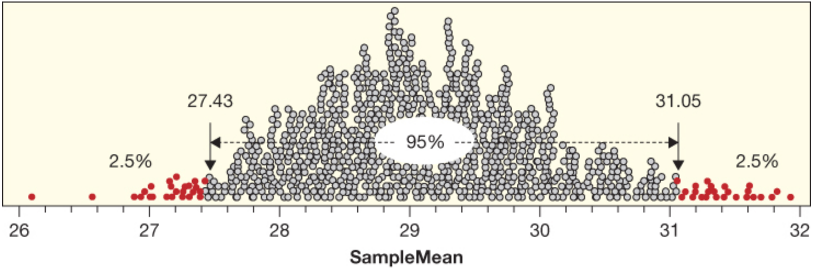
The percentile bootstrap method

- If the sampling distribution is *not bell-shaped*, the 95% rule might produce to *invalid* confidence intervals
  - An invalid CI procedure systematically fails to capture the true population parameter as often as the confidence level advertises (ie: 95% of the time)

The percentile bootstrap method

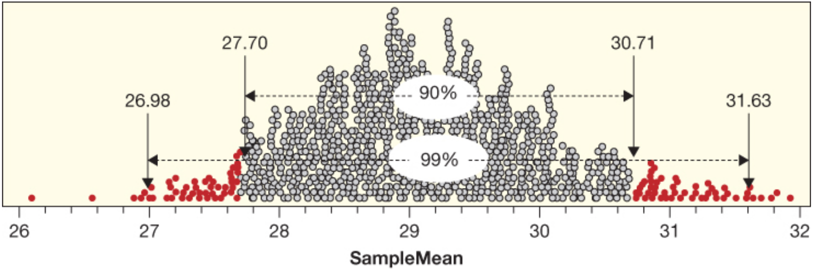
- If the sampling distribution is *not bell-shaped*, the 95% rule might produce to *invalid* confidence intervals
  - An invalid CI procedure systematically fails to capture the true population parameter as often as the confidence level advertises (ie: 95% of the time)
- The **percentile bootstrap method** does not assume any distributional shape, and can be used in a wider variety of situations
  - Finding a 95% percentile bootstrap CI is done by excluding the most extreme 2.5% of the bootstrap samples on each side of the point estimate

The diagram below illustrates the percentile bootstrap method for a 95% CI estimate of a population's mean:



The percentile bootstrap method

The percentile bootstrap method can be used to produce intervals with confidence levels other than 95%:



Practice

A study conducted by Johns Hopkins University Hospital found that 31 of 39 babies born in their facilities at 25 weeks gestation (15 weeks early) went on to survive. Our goal is to estimate the proportion of babies born under similar circumstances in similar hospitals that will survive.

1) Use StatKey and the percentile bootstrap method to find the 95% CI estimate for p.
2) Compare this 95% CI with the interval we previously found using the 2-SE method

Practice (solution)

1) I got (0.641, 0.923) - your answer might be slightly different (depending on your bootstrap samples)
2) Recall the 95% CI from the 2-SE method was (0.663, 0.927). The percentile bootstrap interval is somewhat wider, which is likely needed to achieve 95% confidence given the skew seen in the bootstrap distribution.

Conclusion

- Trends observed in sample data are not a perfect reflection of the population we're studying
  - **Confidence intervals** provide a meaningful way to quantify how much uncertainty exists when generalizing our sample results to a broader population
  - Confidence intervals take the form:
    Point Estimate $\pm$ Margin of Error
- **Bootstrapping** is a method used to find the amount of sampling variability present in our data
  - If the bootstrap distribution is reasonably bell-shaped, we can use the *2-SE method* to come up with a 95% CI estimate
  - More generally, we can use the *percentile bootstrap method* to find confidence intervals for a variety of confidence levels and sampling distribution shapes