Sampling Distributions, Bootstrapping, and Confidence Intervals

Ryan Miller

Outline

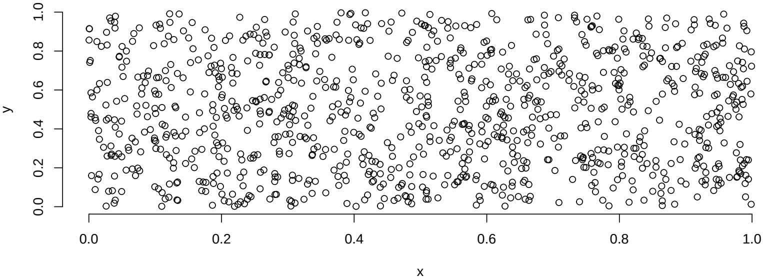
1. Sources of variability and sampling distributions
   - Sampling distributions, bootstrapping
2. Confidence intervals
   - Definition and interpretation, the standard error method, the percentile bootstrap method

Introduction

- Lately we've been discussing possible explanations for trends observed in sample data
  - *Study design* can be used to rule out explanations like confounding variables or bias
  - In well-designed study, *random chance* (ie: sampling variability) is the final factor that must be considered when evaluating an observed association

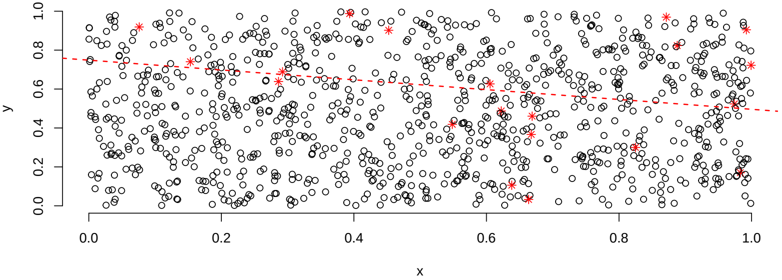
Sampling variability

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



Sampling variability

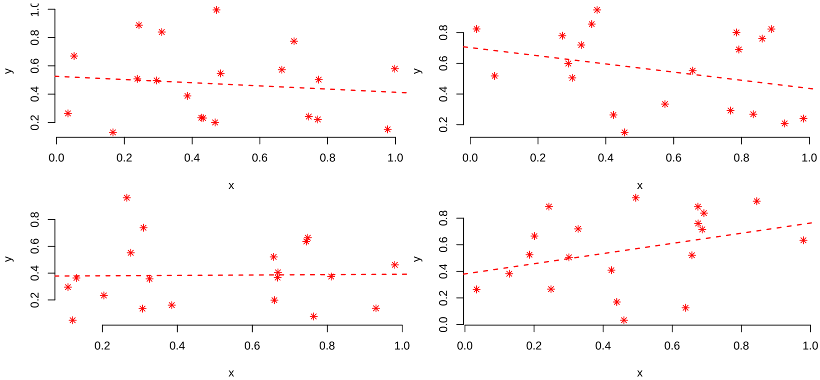
Here is a *random sample* ($n = 20$) from this population (sampled cases are colored in red), the sample correlation is $r = -0.245$:



A *weak negative correlation* is seen in the sample data despite these variables having *no correlation* in the population

Sampling variability

Shown below are another four random samples (each $n = 20$):



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

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?

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?

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?

Questions

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

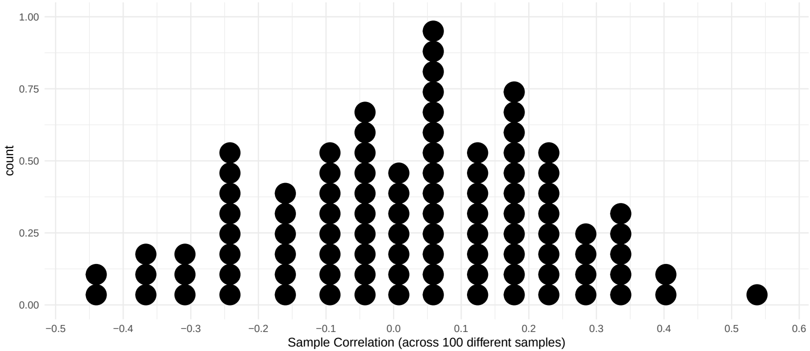
1) If the you take a *single sample* and find a sample correlation of $r = 0.33$, how *confident* can you be that X and Y are related in the population?
2) If you take *100 different samples* and find sample correlations ranging from $r = 0.25$ to $r = 0.35$, how *confident* can you be that X and Y are 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) Confident, none of the samples had correlations near zero and it'd be very unlikely for all 100 to have sample correlations this high by chance.
3) Not confident, it seems like $\rho = 0$ is plausible since many of the samples had correlations near zero.

Sampling distributions

The distribution of *all possible sample outcomes* that could be observed in a study is known as a **sampling distribution**:



The variability seen in the sampling distribution relates to how confident we can be in any observed sample estimate

Bootstrapping

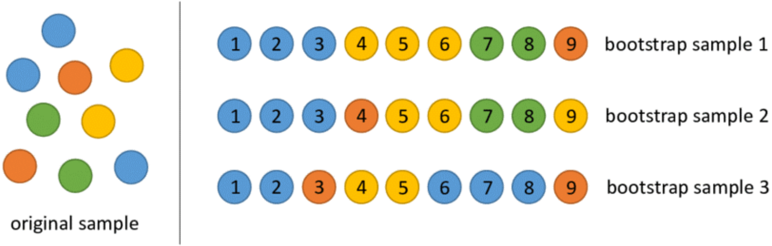
- Obtaining a sampling distribution can be difficult
  - It's not feasible to repeat a study hundreds times to assess the amount of sampling variability

Bootstrapping

- Obtaining a sampling distribution can be difficult
  - It's not feasible to repeat a study hundreds times to assess the amount of sampling variability
- **Bootstrapping** is an ingenious method designed to mimic the process of repeating a study
  - Instead of repeatedly collecting new samples from the population, *bootstrap samples* are obtained by randomly drawing cases from the original sample *with replacement*

Bootstrapping

Notice how most bootstrap samples will contain replicates of data-points that only appeared once in the original sample:



Statisticians have *mathematically proven* that bootstrapping provides an accurate estimate of the true amount of sampling variability in most applications

Bootstrapping in StatKey

- 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 generalize these findings to other comparable hospitals while being mindful of sampling variability

Using the "Bootstrap for a single proportion" menu of StatKey, we'll enter the "count" as 31 and the "sample size" as 39, then generate 1000 bootstrap samples:

1) What does each dot depicted on the "Bootstrap dotplot" represent?
2) How would you express the amount of sampling variability present in this study?

Bootstrapping in StatKey (solution)

1) Each dot represents the proportion of babies who survived in a *different bootstrap sample*.
2) The standard error (which is the standard deviation of a sample estimate) is approximately 0.066, this describes how much we'd expect different sample estimates to deviate from their expected value (on average).

Interval estimation

Generally speaking, there are two different types of estimates (of a population parameter) that statisticians derive and report using sample data:

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

Interval estimation

Generally speaking, there are two different types of estimates (of a population parameter) that statisticians derive and report using sample data:

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

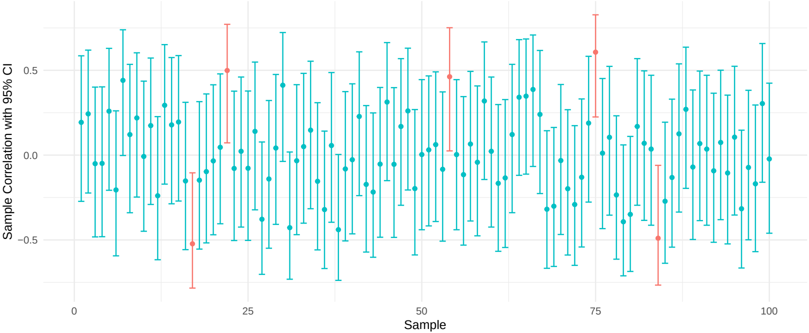
- A **confidence interval** is an interval estimate whose margin of error is based upon a procedure with a long-run "success rate" known as a *confidence level*

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 to come up with a 95% CI:

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

- 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

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

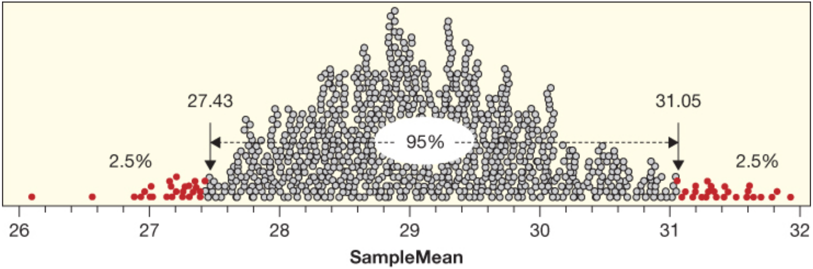
- If the sampling distribution is *not bell-shaped*, the 95% rule could produce to *invalid* confidence intervals
  - An invalid CI procedure systematically fails to capture the true population parameter as often as the confidence level advertises (ie: 95% of the time)

The percentile bootstrap method

- If the sampling distribution is *not bell-shaped*, the 95% rule could 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

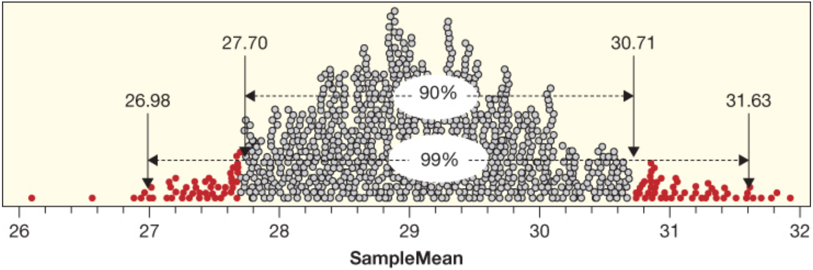
The percentile bootstrap method

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