ROS: Chapter4 Statistical Inference: Part1

Prof. Kapitula

9/23/2021

### Needed packages

library(tidyverse)  
library(infer)  
library(janitor)  
library(matrixStats)

Working through examples in *Regression and Other Stories* by Gelman, Hill and Vehtari. Original RMD files were Solomon Kurz’s versions of examples from ROS that he originally edited but they have been heavily edited by Professor Kapitula.

## Sampling distributions and generative models

### Sampling, measurement error, and model error

Standard paradigms for thinking about the role of inference:

* In the sampling model, we are interested in learning some characteristics of a population (for example, the mean and standard deviation of the heights of all women in the United States), which we must estimate from a sample, or subset, of that population.
* In the measurement error model, we are interested in learning aspects of some underlying pattern or law, simple example:

So the is error.

* Model error refers to the inevitable imperfections of the models that we apply to real data.

In practice, we often consider all three issues when constructing and working with a statistical model. We will consider the standard set up for regression where the data can be thought of as a sample from some population or distribution. The will be thought of typically as independent, identically distributed with some some distribution, and zero mean. For example,

## The mean only model

Suppose we have

This is sometimes called the null model, or the grand mean only model.

### The sampling distribution.

The *sampling distribution* is the set of possible datasets that could have been observed if the data collection process had been re-done, along with the probabilities of these possible values….

the sampling distribution in general will not typically be known, as it depends on aspects of the population, not merely on the observed data (p. 50, *emphasis* in the original).

## 4.2 Estimates, standard errors, and confidence intervals

### 4.2.1 Parameters, estimands, and estimates.

In statistics jargon, *parameters* are the unknown numbers that determine a statistical model. For example, consider the model , in which the errors are normally distributed with mean 0 and standard deviation . The parameters in this model are , , and . The parameters and are called *coefficients*, and is a called a *scale* or *variance parameter*….

An *estimand*, or *quantity of interest*, is some summary of parameters or data that somebody is interested in estimating. For example, in the regression model, , the parameters and might be of interest– is the intercept of the model, the predicted value of when ; and is the slope, the predicted difference in , comparing two data points that differ by 1 in . Other quantities of interest could be predicted outcomes for particular new data points, or combinations of predicted values such as sums, differences, averages, and ratios.

We use the data to construct estimates of parameters and other quantities of interest. (pp. 50–51, *emphasis* in the original)

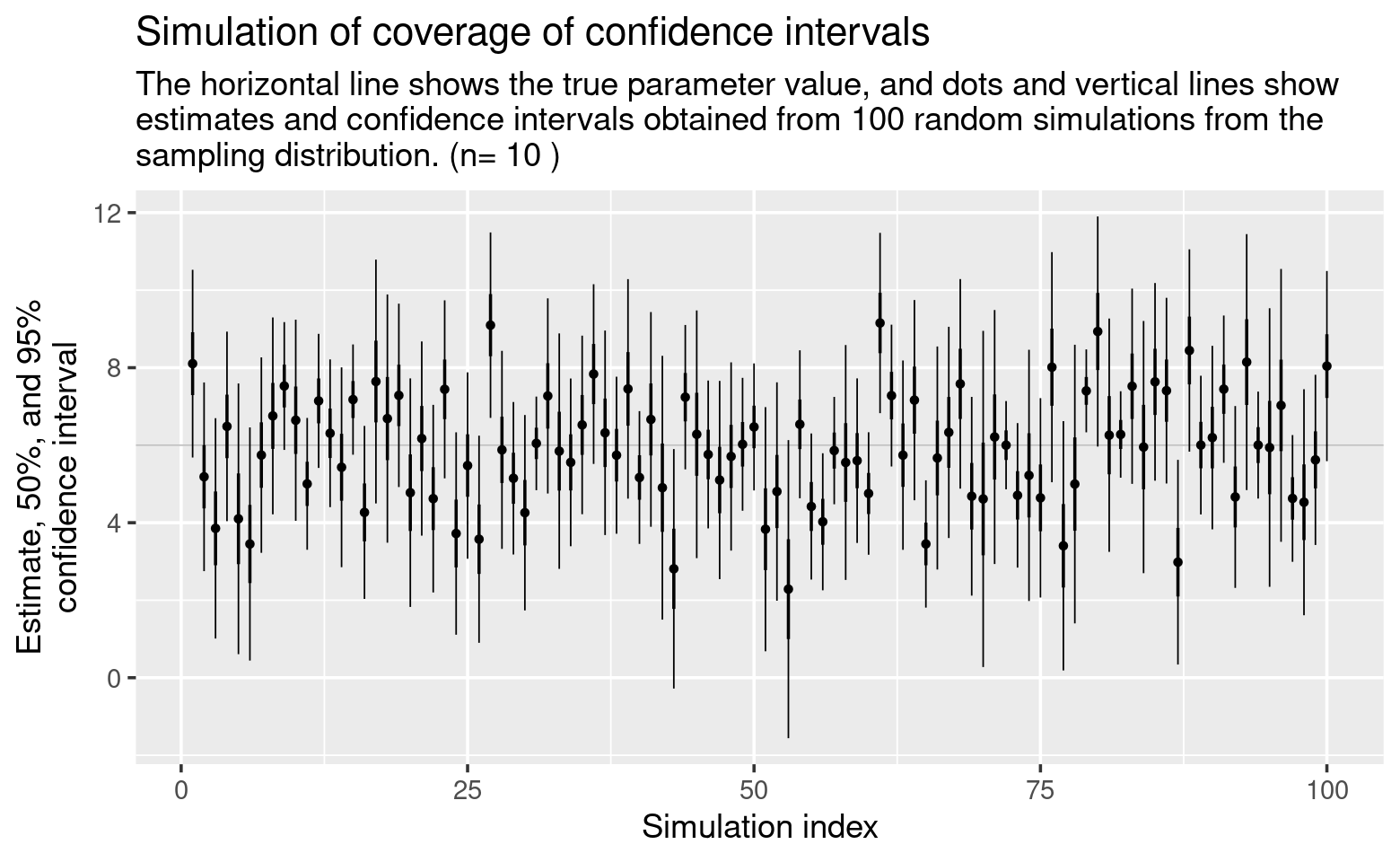
### 4.2.2 Standard errors, inferential uncertainty, and confidence intervals.

The standard error is a measure of the variation in an estimate and gets smaller as sample size gets larger, converging on zero as the sample increases in size. > > The *confidence interval* represents a range of values of a parameter or quantity of interest that are roughly consistent with the data, given the assumed sampling distribution. If the model is correct, then in repeated applications the 50% and 95% confidence intervals will include the true value 50% and 95% of the time. (p. 51, *emphasis* in the original)

You can find the authors’ simulation code for Figure 4.2 in the coverage.Rmd file within the Coverage folder. Here we adjust it a bit for a **tidyverse**-style work flow.

# how many simulations would you like?  
nsims <- 100  
n <-10  
# set the true data-generating parameters  
mu <- 6  
sigma <- 4  
set.seed(410)  
  
# simulate  
sims <- matrix(rnorm(n\*nsims, mean = mu, sd = sigma),nsims,n)  
# sims  
ybar <- rowMeans(sims)  
s <- rowSds(sims)

d <-  
 tibble(i = 1:nsims,ybar,s) %>%   
 mutate(ll95 = ybar - 2 \* s/sqrt(n),  
 ll50 = ybar - 0.67 \* s/sqrt(n),  
 ul50 = ybar + 0.67 \* s/sqrt(n),  
 ul95 = ybar + 2 \* s/sqrt(n))   
  
# plot  
d %>%   
 ggplot(aes(x = i, y = ybar)) +  
 geom\_hline(yintercept = mu, color = "grey75", size = 1/4) +  
 geom\_pointrange(aes(ymin = ll95, ymax = ul95),  
 size = 1/4, fatten = 2/3) +  
 geom\_linerange(aes(ymin = ll50, ymax = ul50),  
 size = 1/2) +  
 labs(title = "Simulation of coverage of confidence intervals",  
 subtitle = paste("The horizontal line shows the true parameter value, and dots and vertical lines show\nestimates and confidence intervals obtained from 100 random simulations from the\nsampling distribution. (n=",n,")"),  
 x = "Simulation index",  
 y = "Estimate, 50%, and 95%\nconfidence interval")



To check, here’s the percentage of 95% intervals containing the data-generating population mean, mu.

# when we sum a logical, we count  
d %>%   
 summarise(percent = sum(ll95 <= mu & mu <= ul95 )/nsims)

## # A tibble: 1 x 1  
## percent  
## <dbl>  
## 1 0.93

And here’s the percentage of 50% intervals containing the data-generating population mean, mu.

d %>%   
 summarise(percent = sum(ll50 <= mu & ul50 >= mu)/nsims)

## # A tibble: 1 x 1  
## percent  
## <dbl>  
## 1 0.45

Both showed good coverage.

The confidence level is the success rate of the method for calculating the confidence interval.

### 4.2.3 Standard errors and confidence intervals for averages and proportions.

When estimating the mean of an infinite population, given a simple random sample of size , the standard error is , where is the standard deviation of the measurements in the population. This property holds regardless of any assumption about the shape of the sampling distribution, but the standard error might be less informative for sampling distributions that are far from normal.

A proportion is a special case of an average in which the data are 0’s and 1’s. Consider a survey of size with Yes responses and No responses. The estimated proportion of the population who would answer Yes to this survey is , and the standard error of this estimate is . If is near 0.5, we can approximate this by . (pp. 51–52)

Consider a case where out of a random sample of 1,000, 700 opposed the death penalty and 300 supported it, we can use the formula to compute the 95% confidence intervals in **R** like so.

Below is a confidence interval comuted using a normal approximation.

n <- 1000  
y <- 700  
  
estimate <- y/n  
  
se <- sqrt(estimate \* (1 - estimate) / n)  
  
estimate + qnorm(c(.025, .975), mean = 0, sd = 1) \* se

## [1] 0.6715974 0.7284026

### 4.2.4 Standard error and confidence interval for a proportion when or .

As a proportion approaches zero or one, the method used above tends to break down. “A standard and reasonable quick correction for constructing a 95% interval when or is near zero is to use the estimate with standard error ” (p. 52). Here’s how this would work when and .

n <- 75  
y <- 0  
  
# probability, when we use the +4 method we add two successes  
# and two failures  
(estimate <- (y + 2) / (n + 4))

## [1] 0.02531646

# se  
(sqrt(estimate \* (1 - estimate) / (n + 4)))

## [1] 0.01767338

# 95% CI  
estimate + qnorm(c(.025, .975), mean = 0, sd = 1) \* se

## [1] -0.003086121 0.053719032

However, since “it makes no sense for the interval for a proportion to contain negative values, so we truncate the interval to obtain [0, 0.054][^1]” (p. 52).

Another, method to use to get the CI is to use prop.test. prop.test uses the Wilson Interval, which will work fairly well with small y or n-y. <https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval#Wilson_score_interval_with_continuity_correction>

n <- 75  
y <- 0  
  
prop.test(y,n)

##   
## 1-sample proportions test with continuity correction  
##   
## data: y out of n, null probability 0.5  
## X-squared = 73.013, df = 1, p-value < 2.2e-16  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.00000000 0.06071113  
## sample estimates:  
## p   
## 0

### 4.2.5 Standard error for a comparison.

The formula to compute the standard error of the difference of two independent quantities follows the form

Consider a survey of 400 men, 57% of whom said they’d vote Republican, and 600 women, 45% of whom said they’d vote Republican.

# men  
n <- 400  
y <- n \* .57  
  
estimate\_men <- y/n  
  
se\_men <- sqrt(estimate\_men \* (1 - estimate\_men) / n)  
  
# women  
n <- 600  
y <- n \* .45  
  
estimate\_women <- y/n  
  
se\_women <- sqrt(estimate\_women \* (1 - estimate\_women) / n)  
  
# estimated gender gap  
estimate\_men - estimate\_women

## [1] 0.12

# se difference  
sqrt(se\_men^2 + se\_women^2)

## [1] 0.03201953

# can use prop.test to get CI for difference  
prop.test(matrix(c(.57\*400,.45\*600,400-.57\*400,600-.45\*600),2,2))

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: matrix(c(0.57 \* 400, 0.45 \* 600, 400 - 0.57 \* 400, 600 - 0.45 \* 600), 2, 2)  
## X-squared = 13.348, df = 1, p-value = 0.0002586  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.05515955 0.18484045  
## sample estimates:  
## prop 1 prop 2   
## 0.57 0.45

### 4.2.6 Sampling distribution of the sample mean and standard deviation; normal and distributions.

Suppose you draw n data points from a normal distribution with mean and standard deviation , and then compute the sample mean , and standard deviation . These two statistics have a sampling distribution that can be derived mathematically from the properties of independent samples from the normal. The sample mean, , is normally distributed with mean and standard deviation . The sample standard deviation has a distribution defined as follows: has a distribution with degrees of freedom.

### 4.2.7 Degrees of freedom.

The concept of *degrees of freedom* arises with the distribution and several other places in probability and statistics. Without going into the technical details, we can briefly say that degrees of freedom relate to the need to correct for overfitting when estimating the error of future predictions from a fitted model… Roughly speaking, we can think of observed data as supplying “degrees of freedom” that can be used for parameter estimation, and a regression with coefficients is said to use up of these degrees of freedom. (p. 53, *emphasis* in the original)

### 4.2.8 Confidence intervals from the distribution.

The distribution is a family of symmetric distributions with heavier tails (that is, a greater frequency of extreme values) compared to the normal distribution. The is characterized by a center, a scale, and a degrees of freedom parameter that can range from 1 to . Distributions in the family with low degrees of freedom have very heavy tails; in the other direction, in the limit as degrees of freedom approach infinity, the distribution approaches the normal.

When a standard error is estimated from n data points, we can account for uncertainty using the distribution with degrees of freedom, calcuated as data points minus 1 because of the mean is being estimated from the data. (p. 53)

Take the case where . Here are our , mean, and standard deviation.

y <- c(35, 34, 38, 35, 37)  
  
(n <- length(y))

## [1] 5

(estimate <- mean(y))

## [1] 35.8

(s <- sd(y))

## [1] 1.643168

Now compute the standard error, along with the 50% and 95% intervals.

(se <- s / sqrt(n))

## [1] 0.7348469

# 50% CIs  
estimate + qt(c(0.25, 0.75), df = n - 1) \* se

## [1] 35.2557 36.3443

# 95% CIs  
estimate + qt(c(0.025, 0.975), df = n - 1) \* se

## [1] 33.75974 37.84026

# use t.test  
  
t.test(y)

##   
## One Sample t-test  
##   
## data: y  
## t = 48.718, df = 4, p-value = 1.062e-06  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 33.75974 37.84026  
## sample estimates:  
## mean of x   
## 35.8