ROS: Chapter4 Statistical Inference: Part2

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.

## 4.3 Bias and unmodeled uncertainty

The inferences discussed above are all consistent on the model being true, with unbiased measurements, random samples, and randomized experiments. But real data collection is imperfect, and where possible we should include the possibility of model error in our inferences and predictions. (p. 55)

### 4.3.1 Bias in estimation.

Roughly speaking, we say that an estimate is *unbiased* if it is correct on average. For a simple example, consider a survey, a simple random sample of adults in the United States, in which each respondent is asked the number of hours he or she spends watching television each day. Assuming responses are complete and accurate, the average response in the *sample* is an unbiased estimate of the average number of hours watched in the *population.* (p. 55, *emphasis* in the original)

If the long run average is not equal to what we are estimating we say our estimate is biased.

Let’s check that out in code. For simplicity, let’s say our survey only measures in 1-hour units and that the average number of hours spent watching television is 2. We’ll make a custom function that will take the means of rand0m samples of the Poisson distribution of a set and .

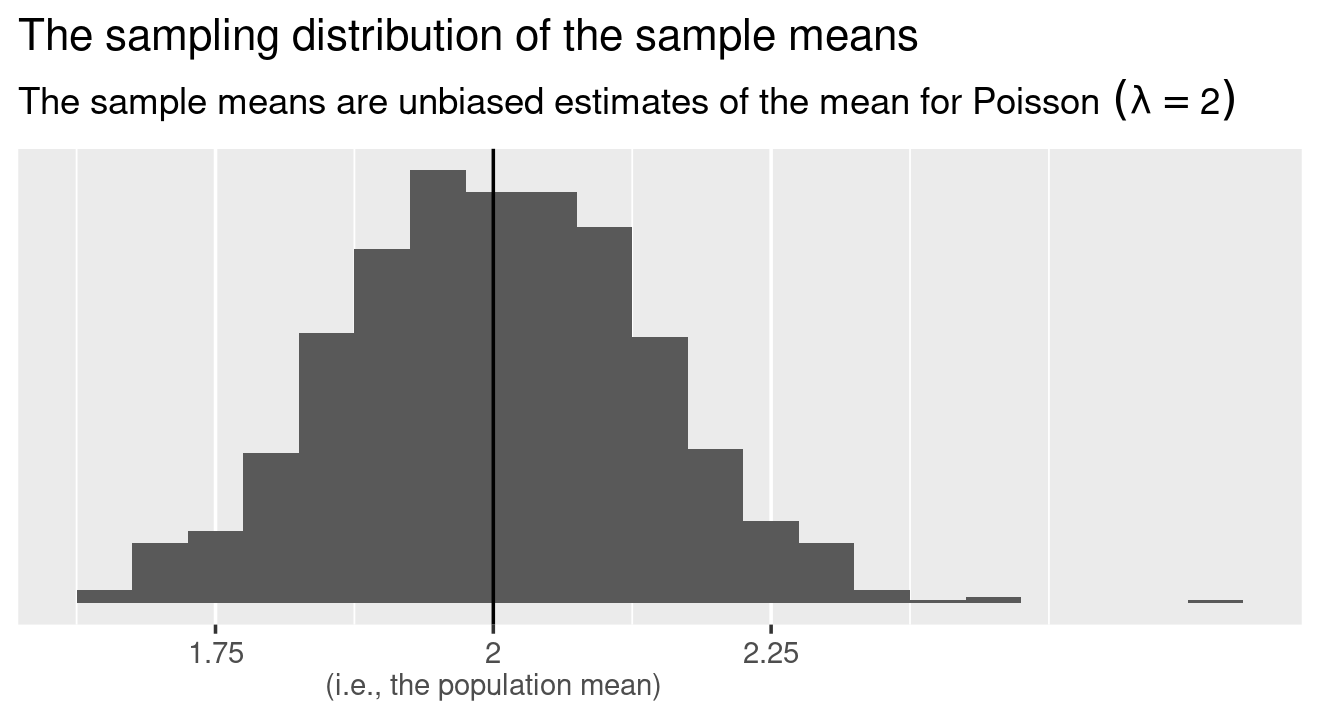
We will take 1,000 random samples for which and . Every simulated sample is a simulated data set with a sample size of 100.

# how many simulations would you like?  
nsims <- 1000  
  
# set the true data-generating parameters  
lambda = 2  
n = 100  
  
set.seed(40)  
  
# simulate  
sims <- matrix(rpois(n = n\*nsims, lambda = lambda),nsims,n)  
ybar <- rowMeans(sims)  
yMedian <-rowMedians(sims)  
  
d <-  
 tibble(i = 1:nsims,ybar,yMedian)   
  
glimpse(d)

## Rows: 1,000  
## Columns: 3  
## $ i <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,…  
## $ ybar <dbl> 1.92, 2.06, 1.78, 1.88, 2.01, 2.00, 2.01, 2.08, 2.03, 2.03, 2.…  
## $ yMedian <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,…

Here is the sample distribution of our means.

d %>%   
 ggplot(aes(x = ybar)) +  
 geom\_histogram(binwidth = 0.05) +  
 geom\_vline(xintercept = 2) +  
 scale\_x\_continuous(NULL,   
 breaks = c(1.75, 2, 2.25),  
 labels = c(1.75, "2\n(i.e., the population mean)", 2.25)) +  
 scale\_y\_continuous(NULL, breaks = NULL) +  
 labs(title = "The sampling distribution of the sample means",subtitle = expression("The sample means are unbiased estimates of the mean for Poisson"~(lambda==2)))



Now formally check how we did by computing the mean of our sample of means.

d %>%   
 summarise(mean = mean(ybar))

## # A tibble: 1 x 1  
## mean  
## <dbl>  
## 1 2.00

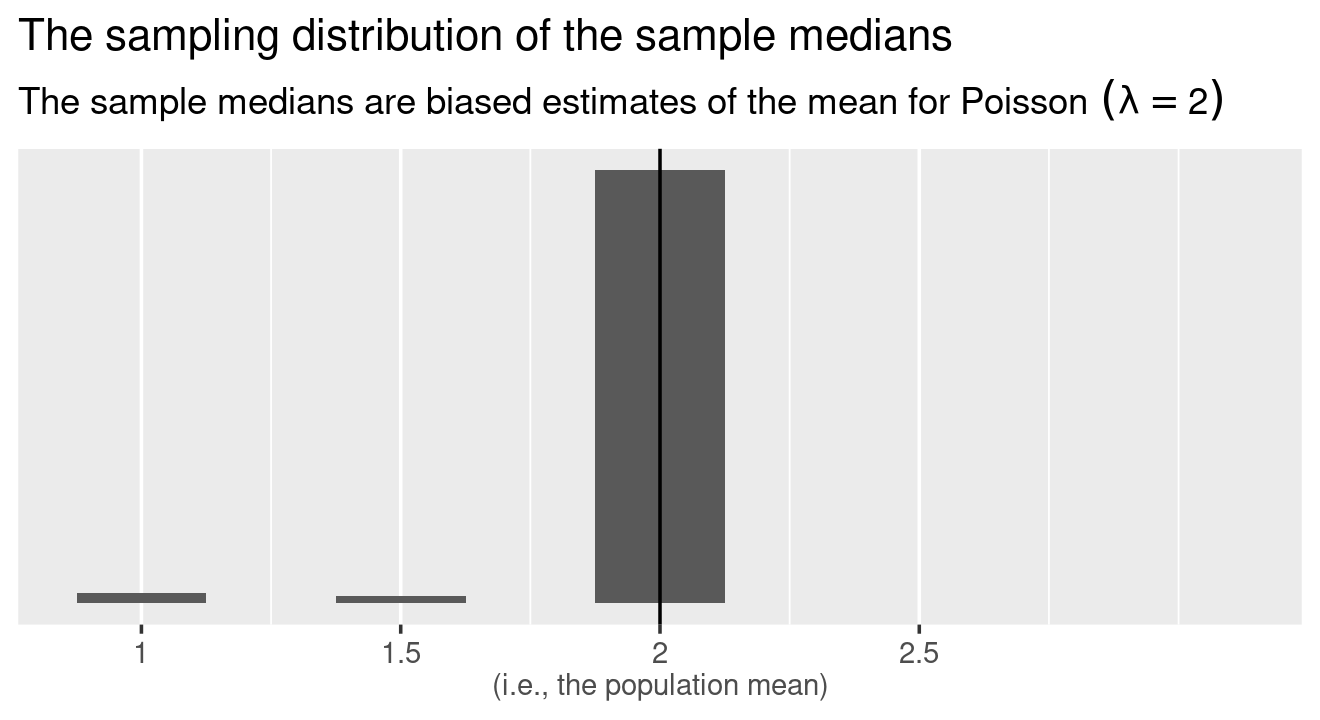
Yep, we got within rounding error of the true mean, 2.

Is the median unbiased???

d %>%   
 summarise(meanofMedians = mean(yMedian))

## # A tibble: 1 x 1  
## meanofMedians  
## <dbl>  
## 1 1.97

d %>%   
 ggplot(aes(x = yMedian)) +  
 geom\_histogram(binwidth=.25) +  
 geom\_vline(xintercept = 2) +  
 scale\_x\_continuous(NULL,   
 breaks = c(.5, 1,1.5, 2, 2.5),  
 labels = c(" ", 1, 1.5, "2\n(i.e., the population mean)" ,2.5)) +  
 scale\_y\_continuous(NULL, breaks = NULL) +  
 labs(title = "The sampling distribution of the sample medians", subtitle = expression("The sample medians are biased estimates of the mean for Poisson"~(lambda==2)))

 Here we see that the median is a slightly biased estimate of lambda because the mean of the sample medians is not equal to lambda.

If we have data that is Normally distributed then the median is an unbiased estimate of the pupulation mean, however, it is more variable, so the mean is better.

Now suppose that women are more likely than men to answer the survey, with nonresponse depending only on sex. In that case, the sample will, on average, overrepresent women, and women on average watch less television than men; hence, the average number of hours watched in the sample is now a *biased* estimate of the proportion in the population. It is possible to correct for this bias by reweighting the sample as in Section 3.1; recognizing the existence of the bias is the first step in fixing it. (p. 55, *emphasis* in the original)

Let’ see if we can work this one out with a mini simulation, too. Sticking with an overall population mean of 2 hours, let’s presume women watch 1.5 hours of television, on average, and men watch 2.5 hours, on average. For simplicity, we’ll further presume the overall population is composed 50%/50% of men and women. The catch is we’ll simulate out samples such that they’re 60% women.

# total sample  
n <- 1000  
  
mu\_women <- 1.5  
mu\_men <- 2.5  
  
set.seed(4)  
  
d <-  
 tibble(hours = c(rpois(n = n \* 0.6, lambda = mu\_women),  
 rpois(n = n \* 0.4, lambda = mu\_men)),  
 sex = rep(c("women", "men"), times = n \* c(.6, .4)))  
  
glimpse(d)

## Rows: 1,000  
## Columns: 2  
## $ hours <int> 2, 0, 1, 1, 3, 1, 2, 3, 4, 0, 2, 1, 0, 4, 1, 1, 4, 2, 4, 2, 2, 6…  
## $ sex <chr> "women", "women", "women", "women", "women", "women", "women", "…

Here are the means, grouped by sex.

d %>%   
 group\_by(sex) %>%   
 summarise(mean = mean(hours), n=n(), var=var(hours))

## # A tibble: 2 x 4  
## sex mean n var  
## <chr> <dbl> <int> <dbl>  
## 1 men 2.36 400 2.17  
## 2 women 1.47 600 1.55

When divided by sex, the simulation appeared to produce unbiased estimates of the subpopulation means. However, it is indeed biased for the overall population.

Notice as well the variance here, since we have a Poisson RV the mean and the variance are expected to be close. Why?

d %>%   
 summarise(mean = mean(hours))

## # A tibble: 1 x 1  
## mean  
## <dbl>  
## 1 1.83

See? It’s a little low. Now we can use the weighting strategy from Section 3.1 where we compute the weighted average following the formula

where indexes groups (sex in this example), stands for the group-specific means, and stands for the number (or percent, in our example) of each group *in the population*.

d %>%   
 group\_by(sex) %>%   
 summarise(mean = mean(hours)) %>%   
 mutate(percent = c(50, 50)) %>%   
 summarise(weighted\_average = sum(percent \* mean) / sum(percent))

## # A tibble: 1 x 1  
## weighted\_average  
## <dbl>  
## 1 1.92

Happily, our weighted average now returns an unbiased estimate of the population average for hours spent each day watching television.

### 4.3.2 Adjusting inferences to account for bias and unmodeled uncertainty.

How can we account for sources of error that are not in our statistical model? In general, there are three ways to go: improve data collection, expand the model, and increase stated uncertainty. (p. 56)

## 4.4 Statistical significance, hypothesis testing, and statistical errors

One concern when performing data analysis is the possibility of mistakenly coming to strong conclusions that do not replicate or do not reflect real patterns in the underlying population. Statistical theories of hypothesis testing and error analysis have been developed to quantify these possibilities in the context of inference and decision making. (p. 57)

### 4.4.1 Statistical significance.

Statistical significance is conventionally defined as a -value less than 0.05, relative to some *null hypothesis* or prespecified value that would indicate no effect present, as discussed below in the context of hypothesis testing. For fitted regressions, this roughly corresponds to coefficient estimates being labeled as statistically significant if they are at least two standard errors from zero, or not statistically significant otherwise. (p. 57, *emphasis* in the original)

The authors then considered a case where you flip a coin 20 times, with eight of the trials coming up heads. The conventional null hypothesis for a fail coin is that you’d have for heads or tails. Using our skills from Section 4.2.3, we know the standard error for a proportion is .

We do the experiment once and we get 8 successes.

n <- 20  
y <- 8  
  
# the estimated probability  
(p <- y / n)

## [1] 0.4

# the standard error  
(se <- sqrt(p \* (1 - p) / n))

## [1] 0.1095445

# the 95% CIs  
p + c(-2 \* se, 2 \* se)

## [1] 0.180911 0.619089

#a better method, use prop.test to get the wilson intervals  
prop.test(y,n)

##   
## 1-sample proportions test with continuity correction  
##   
## data: y out of n, null probability 0.5  
## X-squared = 0.45, df = 1, p-value = 0.5023  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.1997709 0.6358833  
## sample estimates:  
## p   
## 0.4

Those confidence intervals clearly contain within their bounds, leading us to conclude our results are not statistically significantly different from the null hypothesis. We do not have evidence in the data that the true proportion is different than 0.5.

An interpretation of a confidence interval is the values of a parameter that would not be rejected using a 1-CL% alpha level.

### 4.4.2 Hypothesis testing for simple comparisons.

We shall review the key concepts of conventional hypothesis testing with a simple hypothetical example. A randomized experiment is performed to compare the effectiveness of two drugs for lowering cholesterol. The mean and standard deviation of the post-treatment cholesterol levels are and for the people in the treatment group, and and for the people in the control group. (p. 57)

#### 4.4.2.1 Estimate, standard error, and degrees of freedom.

The parameter of interest here is , the expectation (population mean) of the post-test difference in cholesterol between the two groups. Assuming the experiment has been done correctly, the estimate is $ = y\_T - y\_C$ and the standard error is . The approximate 95% interval is then , where is the 97.5th percentile of the unit distribution with degrees of freedom. (p. 57)

#### 4.4.2.2 Null and alternative hypotheses.

To frame the above problem as a hypothesis test problem, one must define *null* and *alternative* hypotheses. The null hypothesis is , that is, , and the alternativeis , thatis, .

The hypothesis test is based on a *test statistic* that summarizes the deviation of the data from what would be expected under the null hypothesis. The conventional test statistic in this sort of problem is the absolute value of the -score, , with the absolute value representing a “two-sided test,” so called because either positive or negative deviations from zero would be noteworthy. (p. 57, *emphasis* in the original)

#### 4.4.2.3 -value.

In a hypothesis test, the deviation of the data from the null hypothesis is summarized by the -*value*, the probability of observing something at least as extreme as the observed test statistic. For this problem, under the null hypothesis the test statistic has a unit distribution with degrees of freedom. (p. 57, *emphasis* in the original)

If we let theta\_hat = , se\_theta = n\_C = , and n\_T = , we can use base **R** to compute our -value like this.

2 \* (1 - pt(abs(theta\_hat) / se\_theta, df = n\_C + n\_T, ncp = 2))

### 4.4.3 Hypothesis testing: general formulation.

In the simplest form of hypothesis testing, the null hypothesis represents a particular probability model, , with potential replication data . To perform a hypothesis test, we must define a test statistic , which is a function of the data. For any given data , the -value is then : the probability of observing, under the model, something as or more extreme than the data.

In regression modeling, testing is more complicated. The model to be fit can be written as , where represents a set of parameters including coefficients, residual standard deviation, and possibly other parameters, and the null hypothesis might be that some particular coefficient of interest equals zero. (p. 58)

### 4.4.4 Comparisons of parameters to fixed values and each other: interpreting confidence intervals as hypothesis tests.

The hypothesis that a parameter equals zero (or any other fixed value) can be directly tested by fitting the model that includes the parameter in question and examining the corresponding 95% interval. If the interval excludes zero (or the specified fixed value), then the hypothesis is said to be rejected at the 5% level.

Testing whether two parameters are equal is equivalent to testing whether their difference equals zero. We can do this by including both parameters in the model and then examining the 95% interval for their difference. As with inference for a single parameter, the confidence interval is commonly of more interest than the hypothesis test. (p. 58)

### 4.4.5 Type 1 and type 2 errors and why we don’t like talking about them.

Statistical tests are typically understood based on *type 1 error*–the probability of falsely rejecting a null hypothesis, if it is in fact true–and *type 2 error*–the probability of not rejecting a null hypothesis that is in fact false. But this paradigm does not match up well with much of social science, or science more generally. (pp. 58–59, *emphasis* in the original)

### 4.4.6 Type M (magnitude) and type S (sign) errors.

A *type S error* occurs when the sign of the estimated effect is of the opposite direction as the true effect. A *type M error* occurs when the magnitude of the estimated effect is much different from the true effect. A statistical procedure can be characterized by its type S error rate–the probability of an estimate being of the opposite sign of the true effect, conditional on the estimate being statistically significant–and its expected exaggeration factor–the expected ratio of the magnitude of the estimated effect divided by the magnitude of the underlying effect. (p. 59, *emphasis* in the original)

For more on this, check out the [**PRDA** package](https://CRAN.R-project.org/package=PRDA), which is designed to assess type S and M errors for a given study design.

### 4.4.7 Hypothesis testing and statistical practice.

We do not generally use null hypothesis significance testing in our own work. In the fields in which we work, we do not generally think null hypotheses can be true: in social science and public health, just about every treatment one might consider will have *some* effect, and no comparisons or regression coefficient of interest will be *exactly* zero. We do not find it particularly helpful to formulate and test null hypotheses that we know ahead of time cannot be true. Testing null hypotheses is just a matter of data collection: with sufficient sample size, any hypothesis can be rejected, and there is no real point to gathering a mountain of data just to reject a hypothesis that we did not believe in the first place. (p. 59, *emphasis* in the original)

## 4.5 Problems with the concept of statistical significance

A common statistical error is to summarize comparisons by statistical significance and to draw a sharp distinction between significant and nonsignificant results. The approach of summarizing by statistical significance has five pitfalls: two that are obvious and three that are less well understood. (p. 60)

### 4.5.1 Statistical significance is not the same as practical importance.

Small effects can be statistically significant, say a treatment results in an increase in annual income of $10 with a SE of $2 in the USA.

### 4.5.2 Non-significance is not the same as zero.

We are told to not accept the null hypothesis but it can be very tempting to do just that. <https://statmodeling.stat.columbia.edu/2020/09/17/we-want-certainty-even-when-its-not-appropriate/> is a recent blog post that might be of interest.

### 4.5.3 The difference between “significant” and “not significant” is not itself statistically significant.

If you pay attention to the substantive literature, you’ll likely find a lot of examples of authors making this mistake.

I(Prof K) have made these mistakes in the past with how I wrote something up. Xeffect is statistically diff, Yeffect is not… then talk about Xeffect, maybe think up some reason why Yeffect is different than Xeffect… oops…. , then worst of all talk about the difference between Data stories can be good, communication is important, but it is very tricky to not over or under sell our results. There are not hard and fast rules, you must have curiosity and ask a lot of questions to understand your models.

Consider two independent studies with effect estimates and standard errors of and . The first study is statistically significant at the 1% level, and the second is not at all significant at 1 standard error away from zero. Thus it would be tempting to conclude that there is a large difference between the two studies. In fact, however, the difference is not even close to being statistically significant: the estimated difference is 15, with a standard error of (p. 61)

### 4.5.4 Researcher degrees of freedom, -hacking, and forking paths.

Another problem with statistical significance is that it can be attained by multiple comparisons, or multiple potential comparisons. When there are many ways that data can be selected, excluded, and analyzed in a study, it is not difficult to attain a low -value even in the absence of any true underlying pattern. The problem here is *not* just the “file-drawer effect” of leaving non-significant findings unpublished, but also that any given study can involve a large number of “degrees of freedom” available to the researcher when coding data, deciding which variables to include in the analysis, and deciding how to perform and summarize the statistical modeling. (p. 61, *emphasis* in the original)

### 4.5.5 The statistical significance filter.

A big and very important point:

A final concern is that statistically significant estimates tend to be overestimates. This is the type M, or magnitude, error problem discussed in Section 4.4. Any estimate with is by necessity at least two standard errors from zero. If a study has a high noise level, standard errors will be high, and so statistically significant estimates will automatically be large, no matter how small the underlying effect. Thus, routine reliance on published, statistically significant results will lead to systematic overestimation of effect sizes and a distorted view of the world. (p. 62)

### 4.5.6 Example: A flawed study of ovulation and political attitudes.

*Note if time is short I might skip going over this example as it is rather involved.*

Herein the authors discussed the flaws in the paper by Durante et al (2013), [*The fluctuating female vote: Politics, religion, and the cvulatory cycle*](https://journals.sagepub.com/doi/abs/10.1177/0956797612466416).

## 4.7 Moving beyond hypothesis testing

Null hypothesis significance testing has all sorts of problems, but it addresses a real concern in quantitative research: we want to be able to make conclusions without being misled by noisy data, and hypothesis testing provides a check on overinterpretation of noise. How can we get this benefit of statistical reasoning while avoiding the overconfidence and exaggerations that are associated with conventional reasoning based on statistical significance? (p. 66)

* analyze all your data
* present all your comparisons
* make your data and code public