MA678 Homework 3

Jing Xu

9/27/2022

Disclaimer (remove after you've read)!

A few things to keep in mind : 1) Use `set.seed()` to make sure that the document produces the same random simulation as when you ran the code. 2) Use `refresh=0` for any `stan_glm()` or Stan-based model. `lm()` or non-Stan models don't need this! 3) You can type outside of the R chunks and make new R chunks where it's convenient. Make sure it's clear which questions you're answering. 4) Even if you're not too confident, please try giving an answer to the text responses! 5) Please don't print data in the document unless the question asks. It's good for you to do it to look at the data, but not as good for someone trying to read the document later on. 6) Check your document before submitting! Please put your name where "Your Name" is by the author!

4.4 Designing an experiment

You want to gather data to determine which of two students is a better basketball shooter. You plan to have each student take N shots and then compare their shooting percentages. Roughly how large does N have to be for you to have a good chance of distinguishing a 30% shooter from a 40% shooter?

```
# Considering the 95% confidence interval
#se = sqrt(p_hat*(1-p_hat)/N) z-value = 1.96 depended on normal table
p1 = 0.3
p2 = 0.4
p_hat = abs(p1 - p2)
z = 1.96
N = z^2*(1-p_hat)/p_hat
ceiling(N)
```

```
## [1] 35
```

4.6 Hypothesis testing

The following are the proportions of girl births in Vienna for each month in girl births 1908 and 1909 (out of an average of 3900 births per month):

```
birthdata <- c(.4777,.4875,.4859,.4754,.4874,.4864,.4813,.4787,.4895,.4797,.4876,.4859,
               .4857,.4907,.5010,.4903,.4860,.4911,.4871,.4725,.4822,.4870,.4823,.4973)
```

The data are in the folder `Girls`. These proportions were used by von Mises (1957) to support a claim that that the sex ratios were less variable than would be expected under the binomial distribution. We think von Mises was mistaken in that he did not account for the possibility that this discrepancy could arise just by chance.

(a)

Compute the standard deviation of these proportions and compare to the standard deviation that would be expected if the sexes of babies were independently decided with a constant probability over the 24-month period.

```
sd(birthdata)
```

```
## [1] 0.006409724
```

```
sqrt(mean(birthdata)*(1-mean(birthdata))/length(birthdata))
```

```
## [1] 0.1020202
```

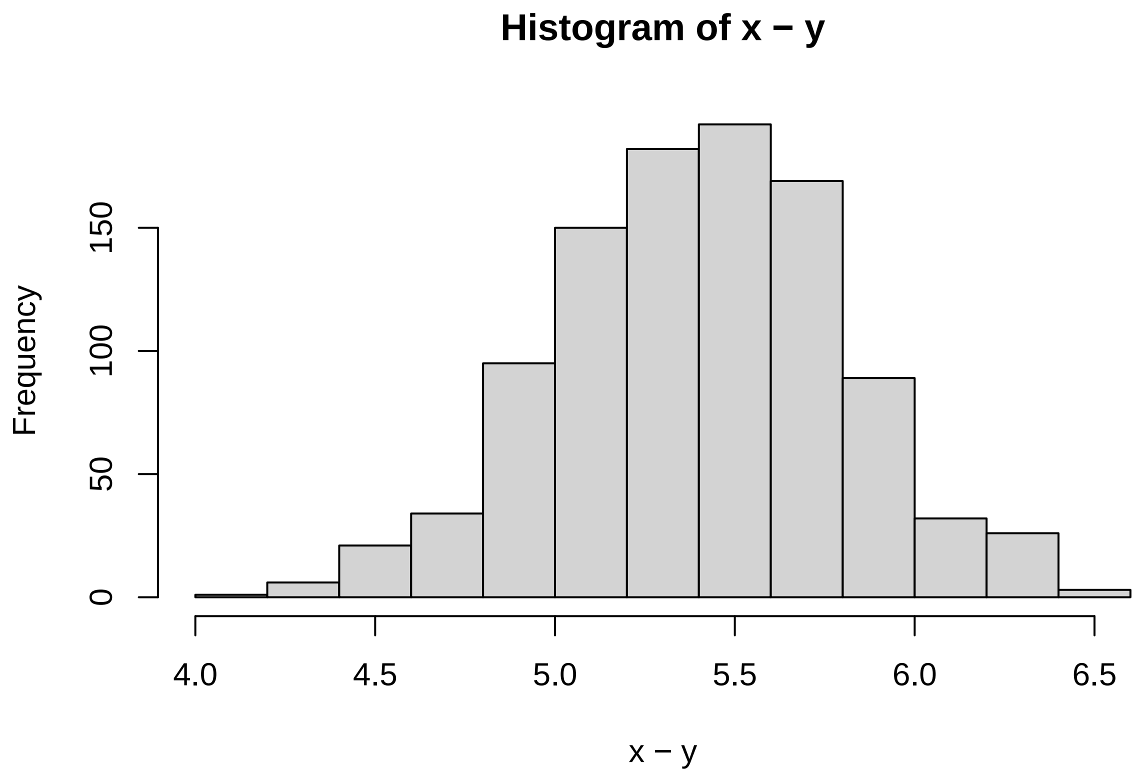
(b)

The observed standard deviation of the 24 proportions will not be identical to its theoretical expectation. In this case, is this difference small enough to be explained by random variation? Under the randomness model, the actual variance should have a distribution with expected value equal to the theoretical variance, and proportional to a χ^2 random variable with 23 degrees of freedom; see page 53.

5.5 Distribution of averages and differences

The heights of men in the United States are approximately normally distributed with mean 69.1 inches and standard deviation 2.9 inches. The heights of women are approximately normally distributed with mean 63.7 inches and standard deviation 2.7 inches. Let x be the average height of 100 randomly sampled men, and y be the average height of 100 randomly sampled women. In R, create 1000 simulations of $x - y$ and plot their histogram. Using the simulations, compute the mean and standard deviation of the distribution of $x - y$ and compare to their exact values.

```
set.seed(100)
size = 100
n = 1000
men_m = 69.1
men_sd = 2.9
women_m = 63.7
women_sd = 2.7
x <- rep(0, n)
y <- rep(0, n)
for (i in 1:1000){
  x[i] <- mean(rnorm(100, men_m, men_sd))
  y[i] <- mean(rnorm(100, women_m, women_sd))
  i+1
}
hist(x-y)
```




```
cat("simulated mean = ", mean(x-y),fill = TRUE)
```

```
## simulated mean =  5.403883
```

```
cat("simulated standard deviation = ", sd(x-y), fill = TRUE)
```

```
## simulated standard deviation =  0.3979568
```

```
cat("actual mean = ", men_m - women_m,fill = TRUE)
```

```
## actual mean =  5.4
```

```
cat("actual standard deviation = ", sqrt(men_sd^2 + women_sd^2)
, fill = TRUE)
```

```
## actual standard deviation =  3.962323
```

5.8 Coverage of confidence intervals:

On page 15 there is a discussion of an experimental study of an education-related intervention in Jamaica, in which the point estimate of the treatment effect, on the log scale, was 0.35 with a standard error of 0.17. Suppose the true effect is 0.10—this seems more realistic than the point estimate of 0.35—so that the treatment on average would increase earnings by 0.10 on the log scale. Use simulation to study the statistical properties of this experiment, assuming the standard error is 0.17.

(a)

Simulate 1000 independent replications of the experiment assuming that the point estimate is normally distributed with mean 0.10 and standard deviation 0.17.

```
set.seed(100)
n_child <- 127
n <- 1000
exp <- list()
for(i in 1:n){
  exp[[i]] <- rnorm(n_child, 0.1, 0.17)

}
```

(b)

For each replication, compute the 95% confidence interval. Check how many of these intervals include the true parameter value.

```
sample_mean <- list()
sample_sd <- list()
ci<- list()
df <- n_child -1
for (i in 1:n) {
  sample_mean[i] <- mean(exp[[i]])
  sample_sd[i] <- sd(exp[[i]])/sqrt(n_child)
  ci[[i]] <- sample_mean[[i]] + qt(c(.025,.975),n_child -1)*sample_sd[[i]]

}
a = 0
for (i in 1:n){
if (ci[[i]][1] <= 0.1 & ci[[i]][2]>=0.1){
    a = a+1
  }
}
cat("# of intervals include the true parameter value = ", a)
```

```
## # of intervals include the true parameter value =  944
```

(c)

Compute the average and standard deviation of the 1000 point estimates; these represent the mean and standard deviation of the sampling distribution of the estimated treatment effect.

```
avg_mean = 0
avg_sd = 0
for( i in 1:1000){
  avg_mean = avg_mean + sum(sample_mean[[i]])
  avg_sd = avg_sd + sum(sample_sd[[i]])
}
avg_mean = avg_mean/n
avg_sd = avg_sd/n
cat("average mean = ", avg_mean, fill = TRUE)
```

```
## average mean =  0.1003613
```

```
cat("average standard deviation = ", avg_sd, fill = TRUE)
```

```
## average standard deviation =  0.01501911
```

10.3 Checking statistical significance

In this exercise and the next, you will simulate two variables that are statistically independent of each other to see what happens when we run a regression to predict one from the other. Generate 1000 data points from a normal distribution with mean 0 and standard deviation 1 by typing `var1 <- rnorm(1000,0,1)` in R. Generate another variable in the same way (call it `var2`). Run a regression of one variable on the other. Is the slope coefficient "statistically significant"? We do not recommend summarizing regressions in this way, but it can be useful to understand how this works, given that others will do so.

```
set.seed(100)
var1 <- rnorm(1000,0,1)
var2 <- rnorm(1000,0,1)
data = data.frame(var1,var2)
fit_1 <- lm(var1 ~ var2)
summary(fit_1)
```

```
##
## Call:
## lm(formula = var1 ~ var2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3284 -0.6678  0.0149  0.6952  3.3141
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01674    0.03260   0.514    0.608
## var2         0.01480    0.03325   0.445    0.656
##
## Residual standard error: 1.031 on 998 degrees of freedom
## Multiple R-squared:  0.0001986,  Adjusted R-squared:  -0.0008032
## F-statistic: 0.1982 on 1 and 998 DF,  p-value: 0.6562
```

```
#we can hardly say so
```

11.3 Coverage of confidence intervals

Consider the following procedure:

- Set $n = 100$ and draw n continuous values x_i uniformly distributed between 0 and 10. Then simulate data from the model $y_i = a + bx_i + \text{error}_i$, for $i = 1, \dots, n$, with $a = 2$, $b = 3$, and independent errors from a normal distribution.
- Regress y on x. Look at the median and mad sd of b. Check to see if the interval formed by the median $\pm$ 2 mad sd includes the true value, $b = 3$.
- Repeat the above two steps 1000 times.

(a)

True or false: the interval should contain the true value approximately 950 times. Explain your answer.

Exactly, that's how a confidence interval being defined.

(b)

Same as above, except the error distribution is bimodal, not normal. True or false: the interval should contain the true value approximately 950 times. Explain your answer.

As for a relativity large n, it will carry out the same result as in (a).