Homework 9

Marwin Carmo

Winter 2024

Contents

Question 1 [1 pt.]

Simulate 1000 values from the following Normal distributions

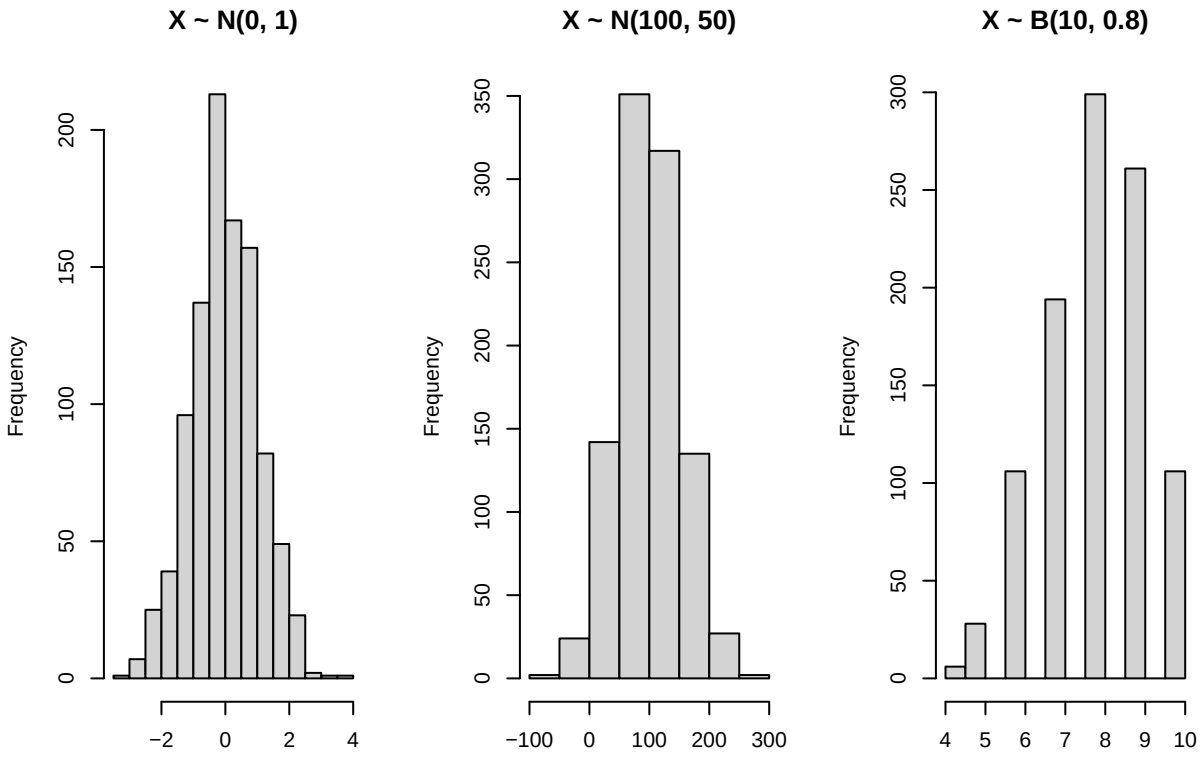
- mean 0 and standard deviation of 1
- mean 100 and standard deviation of 50
- binomial distribution of 10 trials and probability of 0.8

Plot the resulting values in a 1 by 3 plot. Make sure you `set.seed` to 1.

```
set.seed(1)

df_sims <- data.frame(
  simN_0 = rnorm(1000, 0, 1),
  simN_100 = rnorm(1000, 100, 50),
  sim_bin = rbinom(1000, 10, 0.8)
)

par(mfrow = c(1, 3))
hist(df_sims$simN_0, main = "X ~ N(0, 1)", xlab=NULL)
hist(df_sims$simN_100, main = "X ~ N(100, 50)", xlab=NULL)
hist(df_sims$sim_bin, main = "X ~ B(10, 0.8)", xlab=NULL)
```



Question 2

For this question, you will use a ficititious dataset of measures of ability, motivation, and performance. The variables measure:

- GRE Verbal
- GRE Quantitative
- GRE Advanced
- Need for Achievement
- Anxiety
- Graduate GPA
- Rated performance on preliminary exams
- Quality of the Masters Thesis

```
abilities <- read.csv("abilities.csv")
```

2a [1 pt.]

Use a linear regression to predict performance on preliminary exams based on standardized GRE Verbal score and standardized need for achievement. Interpret the meaning of the intercept and regression coefficients. Based on this analysis, which is the better predictor?

```r
abilities$grev_s <- scale(abilities$grev, center = TRUE, scale = TRUE)
abilities$ach_s <- scale(abilities$ach, center = TRUE, scale = TRUE)

mod1 <- lm(prelim ~ grev_s + ach_s, data = abilities)
summary(mod1)
```

```
##
## Call:
## lm(formula = prelim ~ grev_s + ach_s, data = abilities)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89317 -0.57139 -0.00612  0.59573  3.05273
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.02500    0.02854  351.28   <2e-16 ***
## grev_s       0.45132    0.02855   15.81   <2e-16 ***
## ach_s        0.31842    0.02855   11.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9025 on 997 degrees of freedom
## Multiple R-squared:  0.274,  Adjusted R-squared:  0.2725
## F-statistic: 188.1 on 2 and 997 DF,  p-value: < 2.2e-16
```

- Intercept: The intercept tells us that 10.03 is the expected score on the preliminary exam when the GRE verbal score and the need for achievement are at their mean value.
- Coefficient for `grev`: Holding need for achievement constant, we expect that preliminary exam will increase by 0.45 units for a one standard deviation change in GRE Verbal scores.
- Coefficient for `ach`: Holding GRE Verbal scores constant, we expect that preliminary exam will increase by 0.32 units for a one standard deviation change in need for achievement scores.
- Better predictor: `grev`

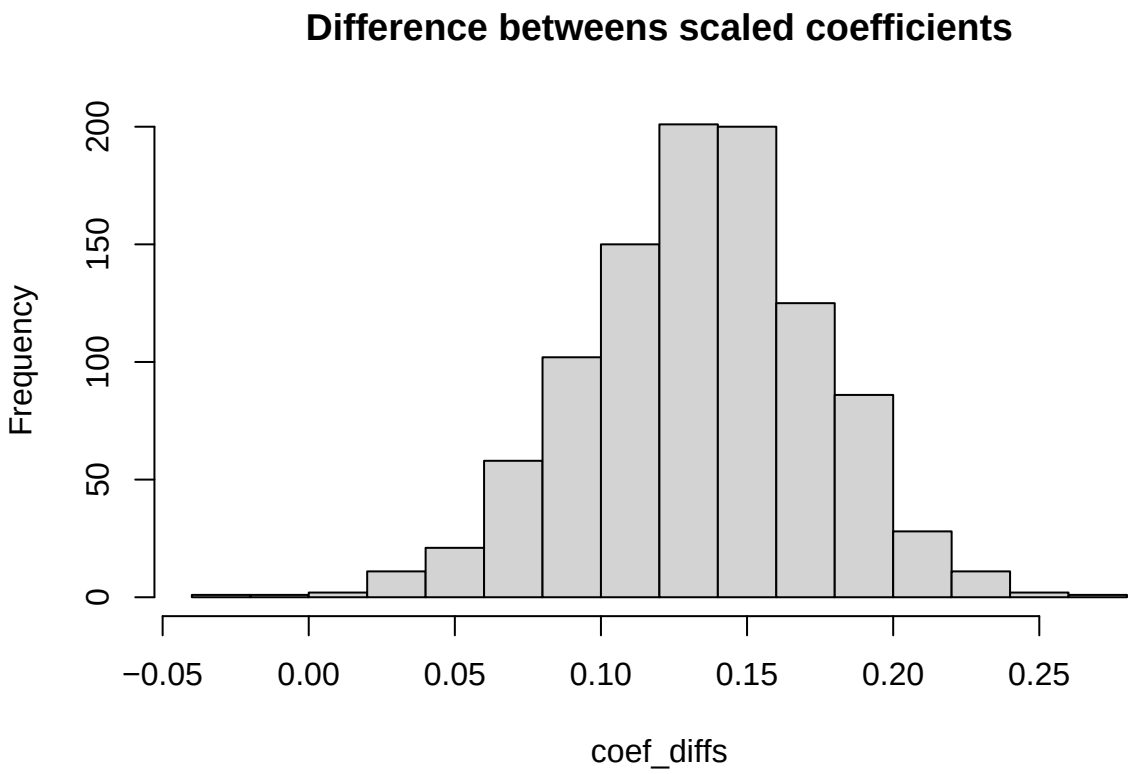
2b [2 pts.]

Now, use a simulation approach to create a distribution of plausible differences between the coefficients for standardized GREV and need for achievement. Plot a histogram of the plausible differences.

```r
set.seed(1)
sim_fit <- sim(mod1, n.sims = 1000)
coef_sims <- sim_fit@coef

coef_diffs <- coef_sims[, 2] - coef_sims[, 3]

hist(coef_diffs, main = "Difference betweens scaled coefficients")
```



2c [1 pt.]

Based on the distribution of differences you created in 2b, calculate a 95% confidence interval. Does this interval support your original conclusion about which predictor is better?

```
quantile(coef_diffs, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## 0.05070244 0.20632208
```

- Answer: The 95% confidence interval of 1000 simulations of the difference between regression coefficients does not contain zero. This suggests that the difference between the two should be greater than zero, supporting the initial conclusions.

Question 3

For this problem, we are going to address the same question as before, but take a bootstrapping approach instead of simulation.

Question 3a [2 pts.]

To use the `boot` function, we need to write a function that computes the statistic we are interested in. I've started it below for you so that it bootstraps the regression. Complete it so that it computes the difference

between the coefficients for `ach` and `grev`. Save and return the difference in an object called `bdiff`. The `coef` function may be helpful in helping you do this.

```
boot_coef_diff <- function(data, i) {
    ach <- scale(data$ach)[i]
    grev <- scale(data$grev)[i]
    prelim <- data$prelim[i]

    bfit <- lm(prelim ~ grev + ach)
    bdiff <- coef(bfit)[2] - coef(bfit)[3]


    return(bdiff)
}
```

Question 3b [2 pts.]

Use the `boot_coef_diff` function to obtain 10000 bootstrapped samples of the diffence between the coefficients for `ach` and `grev`, and compute the bootstrapped confidence interval for this difference. How does it compare to the confidence interval you obtained in question 2c?

```
set.seed(1)
boot_diffs <- boot(abilities, boot_coef_diff, 10000)

boot.ci(boot_diffs)
```

```
## Warning in boot.ci(boot_diffs): bootstrap variances needed for studentized
## intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_diffs)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.0569,  0.2096 )   ( 0.0577,  0.2086 )
##
## Level     Percentile            BCa
## 95%   ( 0.0572,  0.2081 )   ( 0.0572,  0.2081 )
## Calculations and Intervals on Original Scale
```

- Answer: The 95% confidence intervals obtained by simulation are very close in value to those obtained by bootstrapping the differences. Both tell us that the difference between slopes is larger than zero more than 95% of the time.

Question 4 [1 pt.]

In 2 to 4 sentences, compare and contrast the simulation approach to obtain a confidence interval vs. the bootstrapped approach. The lecture slides are likely to be helpful here.

- Answer: The simulation approach assumes a known theoretical distribution for the data. We simulate many samples from this distribution, calculate the statistic of interest for each, and use those to form the confidence interval. In the bootstrapping approach, new samples are created by repeatedly drawing with replacements from the original data, and statistics are calculated on each resample to build the confidence interval.