MA678 Homework 3

Paul Moon

10/3/2024

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?

```
N <- ((0.3 * 0.7) + (0.4 * 0.6)) * (2.8 / (0.3 - 0.4)) ^ 2
cat("There needs to be about", N, "shots needed to distinguish a 30% shooter from a 40% shooter.")
```

```
## There needs to be about 352.8 shots needed to distinguish a 30% shooter from a 40% shooter.
```

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.

```
set.seed(100)
samSD <- sd(birthdata)
conSD <- sd(rbinom(24, 3900, 0.5)/3900) #24 samples given in birthdata, average of 3900, male/female ra
cat("The standard deviation of these proportions is:", samSD)
```

```
## The standard deviation of these proportions is: 0.006409724
```

```
cat("\nThe independently decided with a constant probability standard deviation is:", conSD)
```

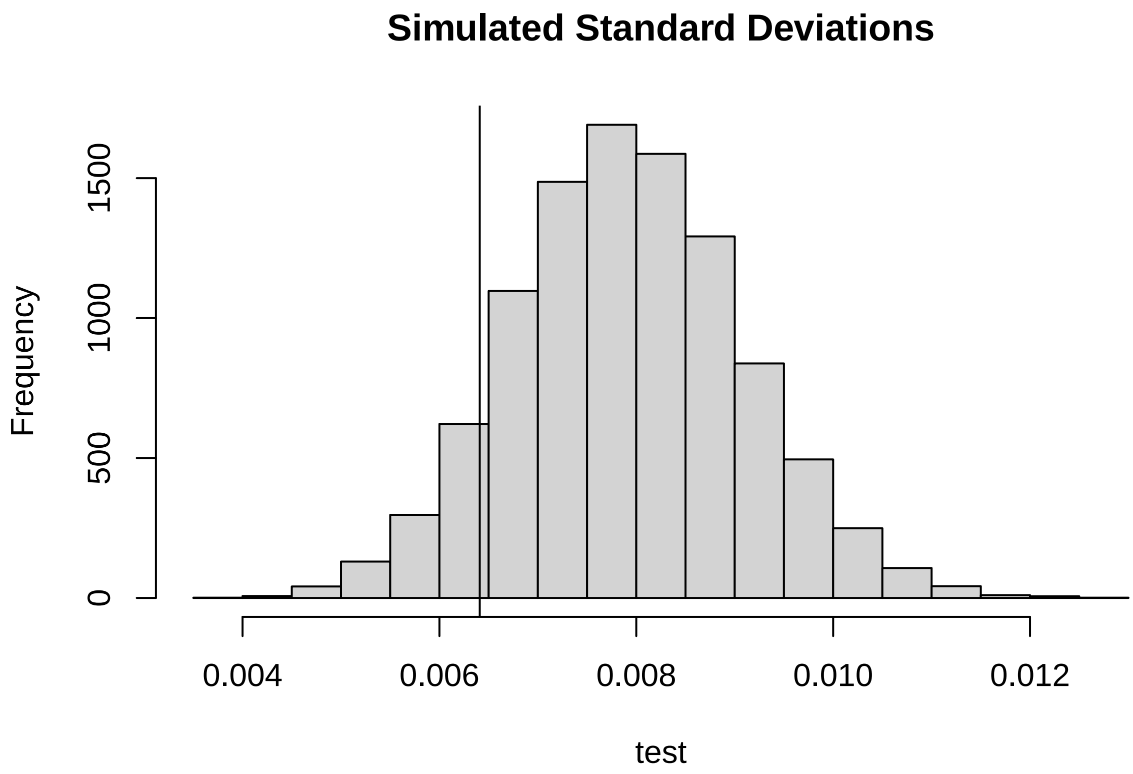
```
##
## The independently decided with a constant probability standard deviation is: 0.005521305
```

(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.

```
set.seed(100)
test <- 1:10000 #used a large number to get a better sense of the distribution (tested numerous differe

for(i in 1:10000){
  test[i] = sd(rbinom(24, 3900, mean(birthdata)) / 3900)
}
hist(test, main = "Simulated Standard Deviations")
abline(v = samSD) #we compared the sample to the repeated tests
```



```
cat("Since the data's standard deviation is not far from the center of the normal distribution, we can s
```

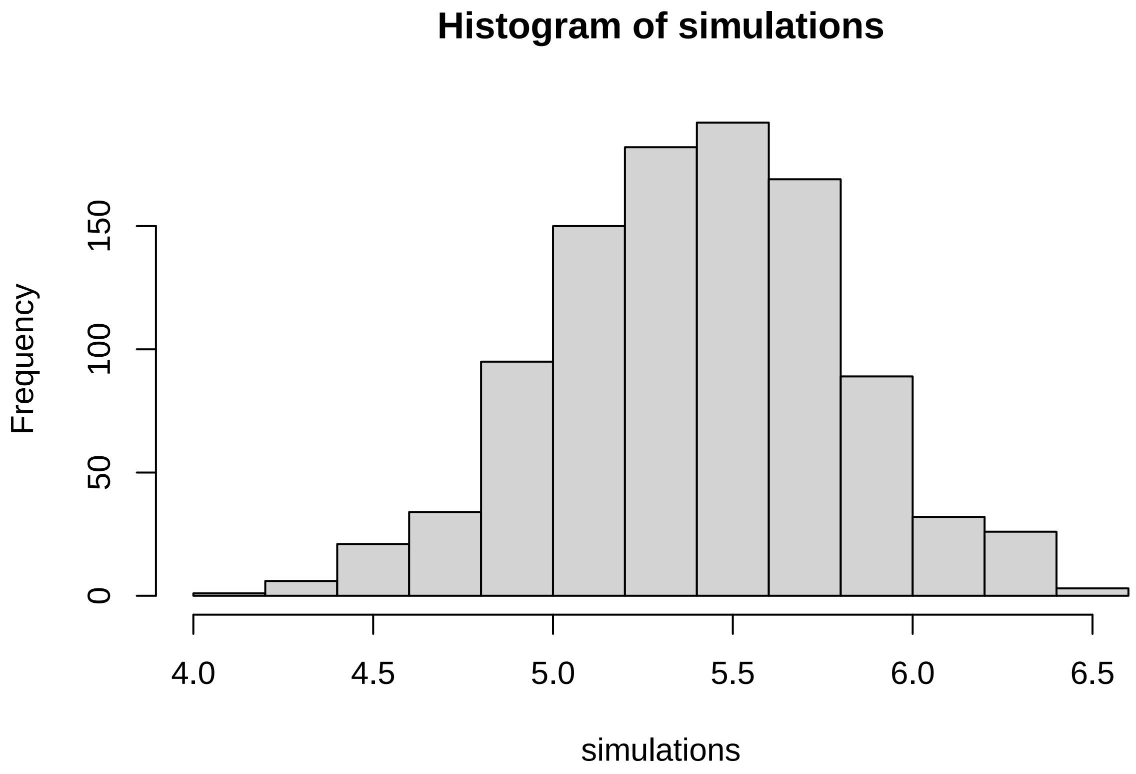
```
## Since the data’s standard deviation is not far from the center of the normal distribution, we can st
```

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.

```r
set.seed(100)
simulations <- 1:1000

for(i in 1:1000){
  men <- mean(rnorm(100, 69.1, 2.9))
  women <- mean(rnorm(100, 63.7, 2.7))
  simulations[i] = men - women #we find the difference in rnorms
}
hist(simulations)
```



```r
cat("Simulated Mean Difference:", mean(simulations))
```

```
## Simulated Mean Difference: 5.403883
```

```
cat("\nSimulated Standard Deviation:", sd(simulations))
```

```
##
## Simulated Standard Deviation: 0.3979568
```

```
cat("\nTrue Mean Difference:", (69.1 - 63.7))
```

```
##
## True Mean Difference: 5.4
```

```
cat("\nTrue Standard Deviation:", sqrt((2.9 ^ 2 / 100) + (2.7 ^ 2 / 100)))
```

```
##
## True Standard Deviation: 0.3962323
```

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.

```
n1 <- 127
pEST <- 0.35
M <- 0.1
stdE <- 0.17
sd <- stdE * sqrt(127) #given numbers go here
```

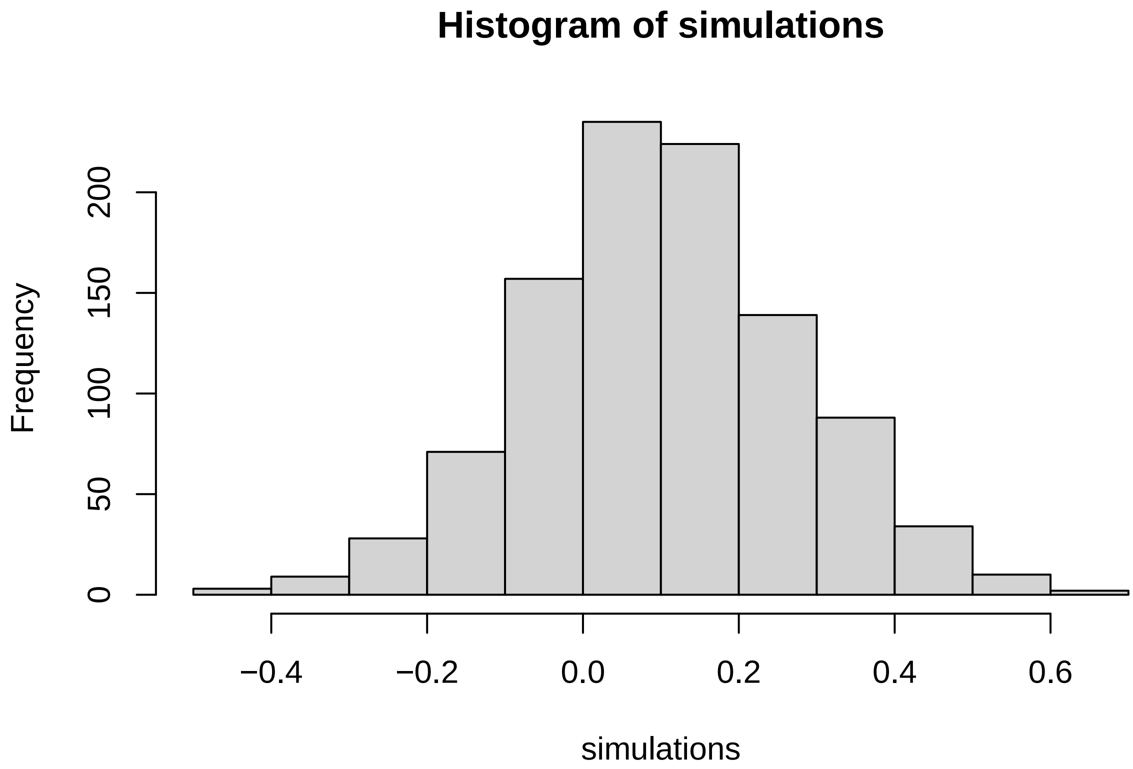
(a)

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)
simulations <- rep(NA, 1000)

for(i in 1:1000){
  sample = rnorm(n1, M, sd)
  simulations[i] = mean(sample)
}

hist(simulations)
```



(b)

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

```
set.seed(100)
CI = rep(NA, 1000)

for(i in 1:1000){
  sample = rnorm(n1, M, sd)
  lower = mean(sample) + qt(0.025, 126) * sd(sample) / sqrt(n1) #lower bound
  upper = mean(sample) + qt(0.975, 126) * sd(sample) / sqrt(n1) #upper bound

  CI[i] = ifelse(lower < 0.1 & upper > 0.1, 1, 0) #confidence interval range
}
sum(CI)
```

```
## [1] 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.

```
cat("Mean of the sampling distribution:", mean(simulations))
```

```
## Mean of the sampling distribution: 0.1040718
```

```
cat("\nStandard Deviation of the sampling distribution:", sd(simulations))
```

```
##
## Standard Deviation of the sampling distribution: 0.1720567
```

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)

regModel <- lm(var1 ~ var2)
summary(regModel)
```

```
##
## 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
```

```
cat("The slope coefficient is NOT statistically significant since the value of 0 is within the 95% confi
```

```
## The slope coefficient is NOT statistically significant since the value of 0 is within the 95% confide
```

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 $i = 1, \ldots, n$, 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 ± 2 mad sd includes the true value, $b = 3$.
- Repeat the above two steps 1000 times.

```
set.seed(100)
q <- c(1:1000)

for(i in 1:1000){
  n = 100
  x <- runif(n, 0, 10)
  y = 2 + 3 * x + rnorm(n, 0, 3)
  model <- lm(y ~ x)

  lower = summary(model)$coefficients[2, 1] - 2 * summary(model)$coefficients[2, 2] #lower bound
  upper = summary(model)$coefficients[2, 1] + 2 * summary(model)$coefficients[2, 2] #upper bound
  q[i] = ifelse(lower < 3 & upper > 3, 1, 0) #confidence interval range
}
mean(q) * 1000 #total number out of 1000
```

```
## [1] 956
```

(a)

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

TRUE. Based on the simulations above, the interval should contain the true value approximately 950 times because we are looking at a normal distribution with values within 2 standard deviations of the mean, which should give us the 95% confident interval. And 95% of 1000 attempts is around 950.

(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.

TRUE. When it comes to a bimodal distribution, if we assume that the errors are independent, the fitted line should just be a shift and therefore still have similar slopes resulting in the same results as before of 950 out of 1000 through the 95% confidence interval.