Problem Set 1

MaCCS 201 - Fall 2024

2024-09-30

Tentative Due Date: September 25

Please submit markdown file named [last_name]_[first_name]_ps1.Rmd on git (instructions to follow)

Part A: Pirates, Astronauts and learning how to express yourself. . . . . .

Dave Eggers is a famous author. Prof. Max likes his books. Turns out, Dave Eggers also tries to help kids by providing places with after school support by offering writing and arts programs. If you have never been to 826 Valencia in San Francisco, you should visit it sometime. Running these programs is costly. So in order to make this operation work, Dave had the genius idea to put a store at the front end of these writing/arts centers, where nerds like Prof. Max buy things they do not need (like 42 sided dice, doubloons etc.). Dave is talking to a large funding agency about possibly providing a grant to support his operations. They want evidence of revenues for these stores. Dave has his accountant draw a random sample of monthly sales from his store network. He has data on 72 monthly sales in dollars. You can assume that sales are not seasonal and i.i.d for this. The funding agency will fund the grant if he can show that average monthly sales (q) are greater than $12,500 per store.

1. Write down the null and alternate hypothesis.

$$H_0 : q \leq 12,500 H_1 : q > 12,500$$

2. Assuming a type 1 error probability of 5%, calculate the critical value (in the relevant t or z score) to the fifth decimal.

$$df = n - 1 = 72 - 1 = 71$$

```
rm(list = ls())

alpha <- 0.05

n <- 72
df <- n-1

t_critical <- qt(1 - alpha, df)
t_critical
```

```
## [1] 1.6666
```

3. The file 'sales.csv' has the sales for the 72 months he supplies to the foundation. Using these numbers, calculate the decision rule (Hint: The $ amount in sales that you would have to exceed to convince the funding agency to fund you.)

If $t \leq t_\alpha$, *fail to* reject null hypothesis;
If $t > t_\alpha$, reject null hypothesis.

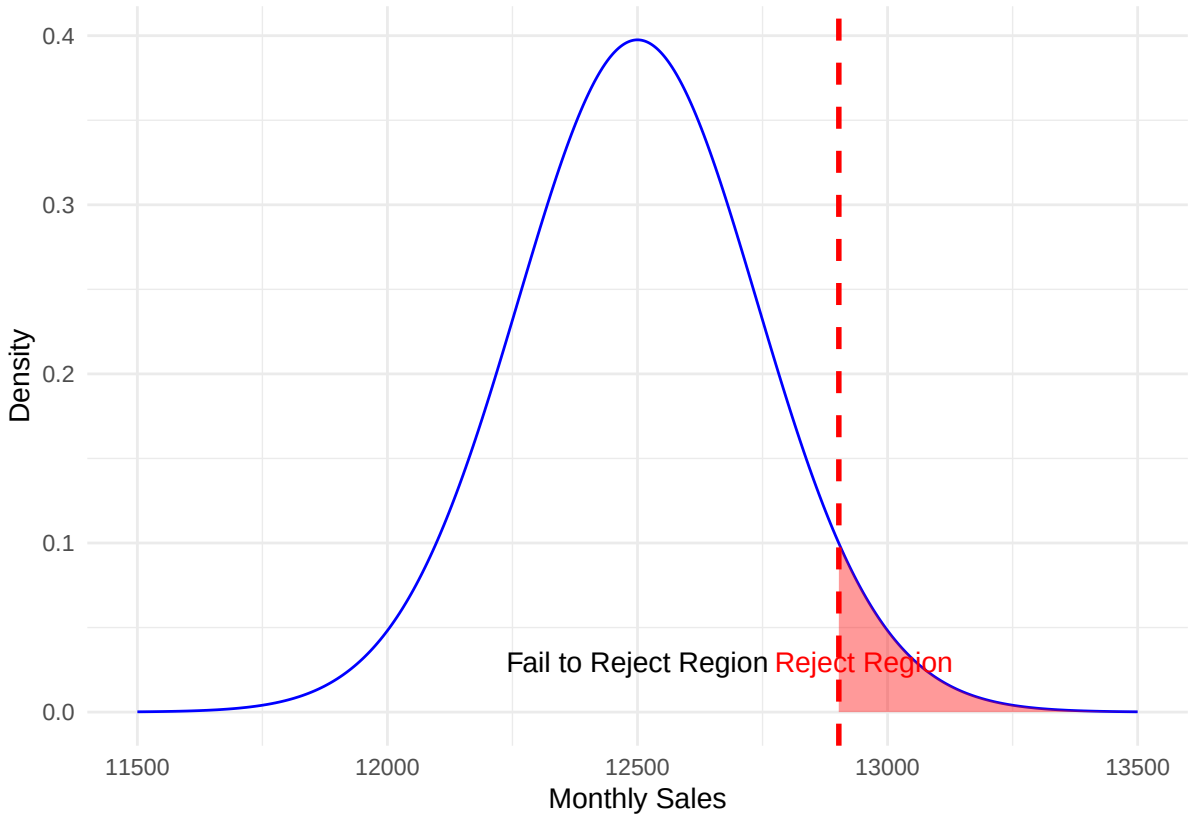
$$t = \frac{\bar{x} - 12,500}{\frac{s}{\sqrt{n}}} > t_\alpha \bar{x} > t_\alpha \frac{s}{\sqrt{n}} + 12,500$$

```
sales <- read.csv(file = "sales.csv", header = FALSE)
colnames(sales) <- c('monthly_sales')

mu_0 <- 12500

mean <- mean(sales$monthly_sales)
std <- sd(sales$monthly_sales)

minimum_mean <- mu_0 + t_critical*(std/sqrt(n))

cat("Minimum mean value rejects the hull hypothesis:", minimum_mean)
```

```
## Minimum mean value rejects the hull hypothesis: 12902.52
```

4. Like in class, draw the sampling distribution under the null hypothesis being true and clearly identify the Reject and Fail to Reject regions.

```
library(ggplot2)

x_values <- seq(mu_0 - 1000, mu_0 + 1000, length = 2000)
sampling_distribution <- dt((x_values - mu_0) / (std / sqrt(n)), df)

ggplot(data.frame(x_values, sampling_distribution), aes(x = x_values, y = sampling_distribution)) +

  geom_line(color = "blue") +

  geom_vline(xintercept = mu_0 + t_critical * (std / sqrt(n)),
             linetype = "dashed", color = "red", linewidth = 1) +

  geom_area(data = data.frame(x_values = x_values
                              [x_values >= (mu_0 + t_critical * (std / sqrt(n)))],
                              sampling_distribution = sampling_distribution
                              [x_values >= (mu_0 + t_critical * (std / sqrt(n)))]),
            aes(x = x_values, y = sampling_distribution), fill = "red", alpha = 0.4) +

  annotate("text", x = mu_0 + t_critical * (std / sqrt(n)) + 50, y = 0.03, label = "Reject Region",
  annotate("text", x = mu_0, y = 0.03, label = "Fail to Reject Region", color = "black") +

  labs(x = "Monthly Sales", y = "Density") +
  theme_minimal()
```



5. Calculate the relevant test statistic.

```
t_stat <- (mean - mu_0) / (std/sqrt(n))
print(t_stat)
```

```
## [1] 1.812827
```

6. Conduct the hypothesis test and tell the funding agency what to do. Should they fund Dave or not?

```
reject_null <- t_stat > t_critical
cat('Reject Null Hypothesis:', reject_null)
```

```
## Reject Null Hypothesis: TRUE
```

Reject null hypothesis, the funding agency should fund Dave.

7. Calculate the Type II error probability if the true average sales are $13000 per store and standard deviation is $2050.

$$\mu_0 = 12,500 \mu_1 = 13,000 \sigma = 2,050$$

$$H_0 : \mu = \mu_0 \Rightarrow \mu = 12,500 H_1 : \mu = \mu_1 \Rightarrow \mu = 13,000$$

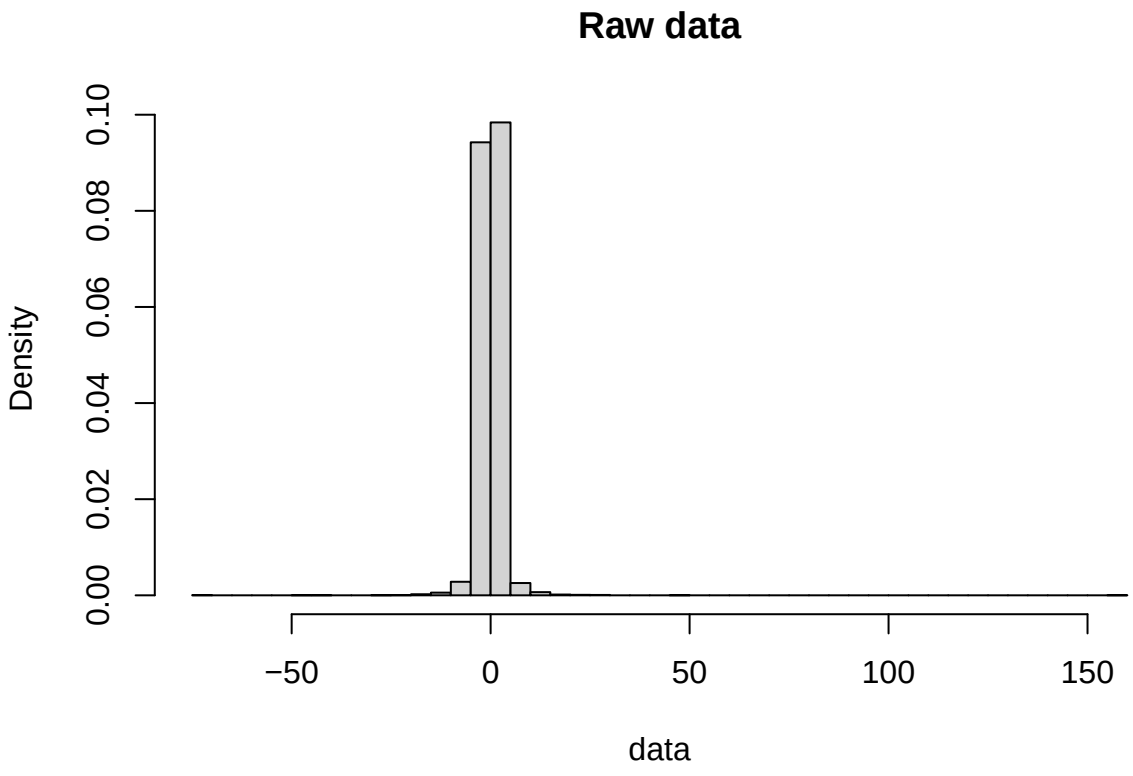
```
mu_1 <- 13000
sd <- 2050

se <- sd/sqrt(n)
t_stat <- (mu_1 - mu_0)/se

beta <- pt(t_critical, df, ncp = t_stat, lower.tail = TRUE)
print(beta)
```

```
## [1] 0.3427597
```

Part B: Monte Carlo is not just a city in Monaco.

In class Max showed a Monte Carlo experiment of how a sample proportion converges to a normal distribution if the sample size condition is met ($n * p > 10$ and $(1 - p) * n > 10$). He also showed that things break down if that condition is not met. He also gave a condition necessary in the case of a continuous random variable, which required that a normal model provides an accurate approximation to the sampling distribution of the sample mean if the sample size n is larger than 10 times the absolute value of the kurtosis $n > 10 * |K_4|$. Using the Monte Carlo code provided in lecture 2 as a point of departure, design a Monte Carlo for a setting that violates this condition and show us what the resulting sampling distribution looks like as a histogram. I would run a Monte Carlo with 10,000 reps and a histogram with 50 bins.

```
rm(list = ls())

set.seed(20030121)

N <- 100000
reps <- 10000

df <- 2.1

# data
data <- rt(10000, df)

# calculate K_4
kurtosis_t <- 6 / (df - 2)

# Plot population
hist(data, prob=TRUE, breaks = 50, main = "Raw data")
```



```r
n <- 20

sample_means <- replicate(reps, mean(sample(data, n, replace = TRUE)))

hist(sample_means, prob=TRUE, breaks = 50, main = "Sampling Distribution of sample mean")

curve(dnorm(x, mean=0, sd=sqrt(var(data)/n)),
      min(sample_means), max(sample_means),
      add=TRUE, lwd=2, col="red")
```