Problem Set 1 - Answer Key

MaCCS 201 - Fall 2025

2025-10-05

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 102 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.

$$H_0 : \mu \leq 12,500$$
$$H_a : \mu > 12,500$$

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

This is a one tailed test. The critical value is $t_{0.05,101} = 1.66008$. You can get this by

```
tst <-qt(0.95,101)
tst
```

```
## [1] 1.660081
```

3. The file "sales.csv' ' has the sales for the 102 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.)

```
sales <- read.csv("sales.csv", header = FALSE)
xbar <- mean(sales$V1)
sd <-sd(sales$V1)/sqrt(102)
dr <- 12500 + tst*sd
dr
```

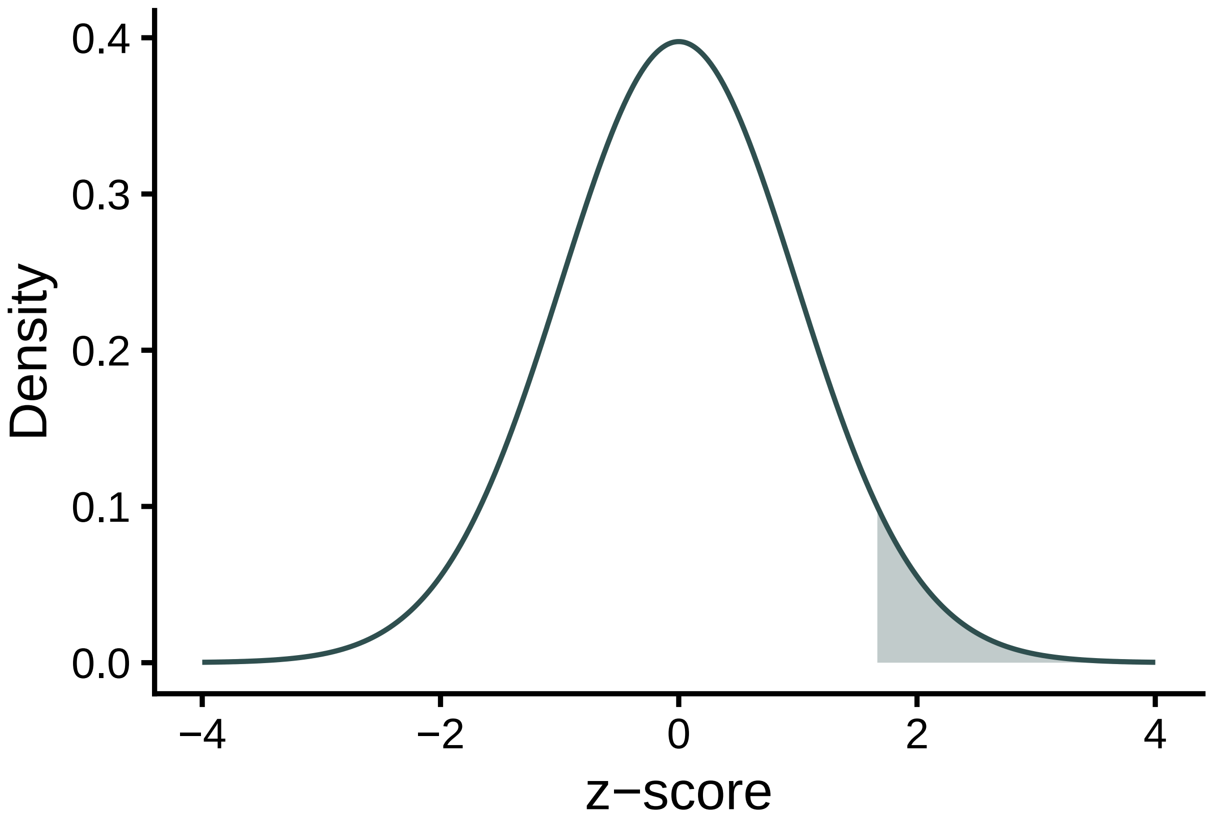
```
## [1] 12816.61
```

You would need to sell more than $12816.61 worth of stuff.

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

The grey area is the reject region (greater than 1.66 t-scores)

```
ggplot(data = tibble(x = c(-4, 4)), aes(x)) +
  stat_function(
    fun = dt, args= list(df=71), n = 1e3,
    color = "darkslategrey"
  ) +
   geom_area(
    stat = "function", fun = dt, args= list(df=71),
    fill = "darkslategrey", alpha = 0.3,
    xlim = c(1.667, 4)
  ) +
  labs(x = "z-score", y = "Density") +
  theme_classic(base_size = 20)
```



5. Calculate the relevant test statistic.

The relevant statistic is $t = \frac{\bar{x} - 12500}{s/\sqrt{n}}$, using the numbers above the test statistic is $t = 1.38217$.

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

You fail to reject the null and Dave doesn't get funded!

7. Calculate the Type II error probability if the true average sales are \$13000 per store and standard deviation is \$2000.

This is equivalent to asking what the probability mass to the left of \$12816.61 (your decision rule) is for a t with mean \$13,000 and standard deviation \$2000/sqrt(102).

```r
t1 <- (xbar - 12500)/(sd(sales$V1)/sqrt(102))
t1
```

```
## [1] 1.38217
```

```r
t2 <- (dr-13100)/(2000/sqrt(102))
t2
```

```
## [1] -1.43106
```

```r
pt(t2,101)
```

```
## [1] 0.07775
```

So the Type II error probability here is 7.78%.

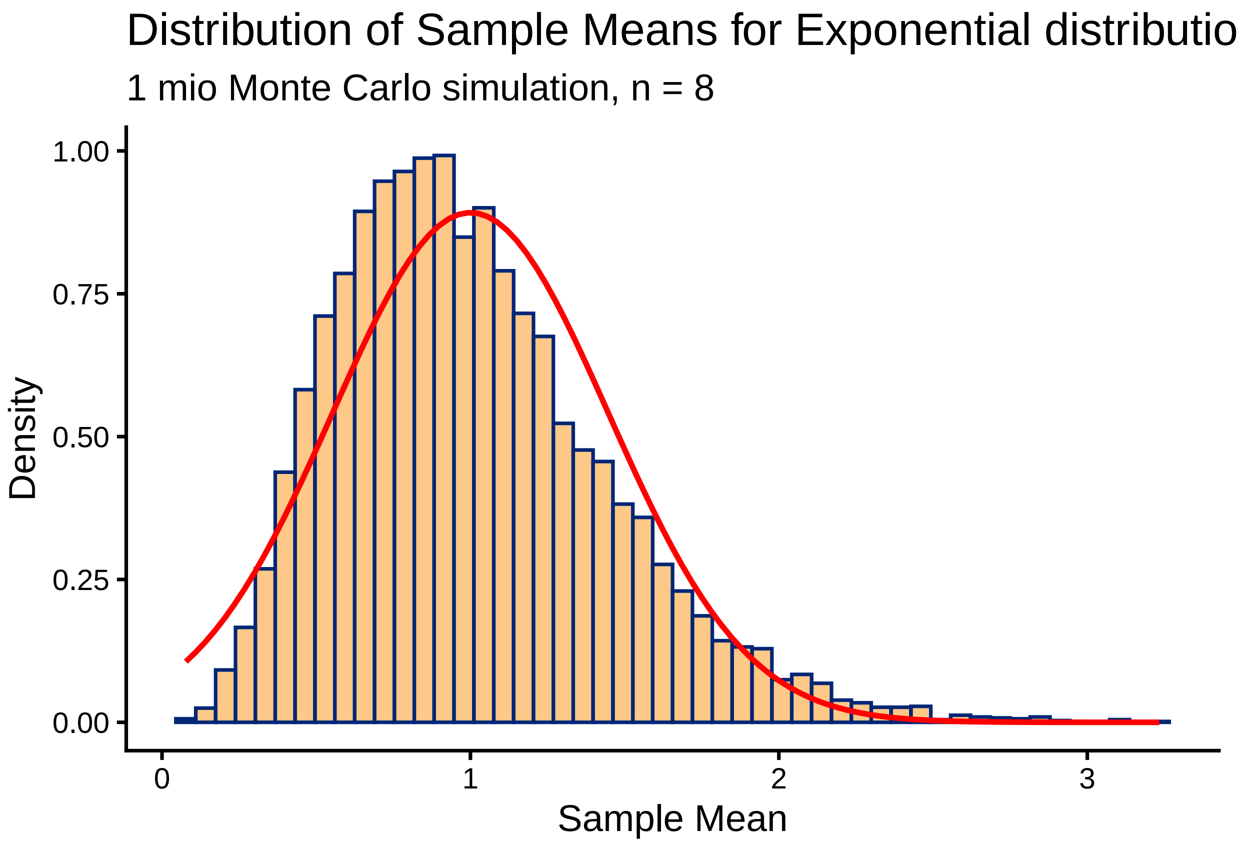
##Part B: Monte Carlo is not just a city in Monaco.

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. $n > 10 * |K4|$. 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.

```r
set.seed(22092008) # set random number generator seed
numloop <- 10000 # number of draws
n <- 5 #sample size

g <- numeric(numloop) # empty vector to hold sample mean for each iteration
k <- numeric(numloop) # empty vector to hold kurtosis for each iteration

for(i in 1:numloop) {
  r <- r(distr::Exp(rate = 1))(n) # Generates random draw from exponential distribution with rate param

  g[i] <- mean(r) # Store mean of drawn sample
  k[i] <- kurtosis(r) # Store kurtosis of drawn sample
}

clt <- data.frame(g,k)
ggplot(clt, aes(x=g)) +
  geom_histogram(aes(y =after_stat(density)),
                 alpha=0.5, fill="#FC9313", color="#002676", position="identity", bins=50)+
  stat_function(fun = dnorm,
                args = list(mean = 1,
                            sd   = 1/sqrt(n)),
                color = "red", size = 1) +
  labs(title="Distribution of Sample Means for Exponential distribution",
       subtitle = "1 mio Monte Carlo simulation, n = 8",
       x="Sample Mean", y = "Density")+

  theme_classic()+
  theme_classic(base_size = 14)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



We created here a Montecarlo simulation using a random draw with sample size n = 5 from an exponential distribution with rate parameter $\lambda = 1$. This distribution has a mean of 1, a standard deviation of 1 and a kurtosis of 9. Therefore, $n < 10/times$kurtosis. We see that a normal distribution $N(1, 1/\sqrt{n})$ does not approximate the sampling distribution of the sample mean.