assignment_01

July 30, 2025

1 Assignment 1

---

1. **Data Classification** *(5)*

Consider the following R dataset detailing the attributes for different vehicles including vehicle make, model, year, class, transmission, drive type, number of engine cylinders, total engine displacement, fuel type, and mileage (highway and city). Classify each variable in the dataset as one of the following: Discrete Quantitative, Continuous Quantitative, Qualitative, and Categorical.

[ ]:

```r
# Loading the packages
library(dplyr)
library(fueleconomy)
```

[ ]:

```r
# Loading the dataset
data <- fueleconomy::vehicles
```

[ ]:

```r
# Dataset Structure
str(data)
```

---

2. **Data Summary** *(10)*

a. Using the vehicles dataset filtered out for Renault vehicles, summarise measure of location (mean, median, mode), dispersion (range, inter-quartile range, standard deviation), and shape (skewness, kurtosis) for highway as well as city miles per galon. *(8)*

[ ]:

```r
# Renault data
fueleconomy::vehicles %>% filter(make=="Renault")
```

b. Based on these statistics, draw inferences for highway and city mileage *(2)*

---

3. **Probability Analysis** *(5)*

a. Using the vehicles dataset filtered out for Honda vehicles, verify the axioms of probability for vehicle classes and engine cylinders. *(1)*

[ ]:

```r
# Honda data
data <- fueleconomy::vehicles %>% filter(make=="Honda")
```

[ ]:

```r
# Honda vehicle classes
as.data.frame(table(class=data$class))
```

[ ]:

```r
# Honda engine cylinders
as.data.frame(table(cyl=data$cyl))
```

b. Using the vehicles dataset filtered out for Honda vehicles, employ conditional probability formula to evaluate the probability of a compact car having a 4-cylinder engine and consequently, employ the Bayes' rule to evaluate the probability a 4-cylinder engine vehicle being a compact car. *(4)*

[ ]:

```r
# Honda data
data <- fueleconomy::vehicles %>% filter(make == "Honda")
```

[ ]:

```r
# Honda vehicle classes and engine cylinders
as.data.frame(table(class=data$class, cyl=data$cyl))
```

---

4. **Data Sampling** *(8)*

a. For the following randomly sampled data from the vehicles dataset, compute bias and standard error for the estimator on highway mileage. *(5)*

[ ]:

```r
library(ggplot2)

P <- fueleconomy::vehicles$hwy
m <- 1000
n <- 100

# TODO: replace with population parameter
z <- ...

Z <- vector("numeric", m)
for (i in 1:m) {
  # TODO: replace with last three digits of your roll number
  set.seed(...)
  I <- order(runif(length(P)))[1:n]
  S <- P[I]
  # TODO: sample parameter
  Z[i] <- ...
}

# TODO: replace with bias formula
b <- ...
message("Bias: ", round(b, 3))
```

```r

# TODO: replace with error formula
E <- ...

# TODO: replace with standard error formula
se <- ...
message("SE: ", round(se, 3))

# plot sample parameters and population parameter
df <- data.frame(Sample = 1:m, SampleMean = Z)
options(repr.plot.width = 12, repr.plot.height = 8)
ggplot(df, aes(x = Sample, y = SampleMean)) +
geom_point(color = "steelblue", size = 3) +
geom_hline(yintercept = z, color = "red", linetype = "dashed", linewidth = 1.2)␣
  ↪+
labs(title = "Sample Means vs Population Mean",
      subtitle = "Each point is the mean highway mileage from the sample",
      y = "Sample Mean",
      x = "Sample Number") +
theme(
  plot.title = element_text(size = 18, face = "bold"),
  plot.subtitle = element_text(size = 14),
  axis.title = element_text(size = 14),
  axis.text = element_text(size = 12)
)
```

b. Using the Archery analogy discussed in the class, draw a representative target board to comment upon the accuracy and precision of the estimator. *(3)*

---

5. **Hypothesis Testing** *(12)*

Test the following hypothesis

a. city mileage is greater than 17.5 mpl

b. highway mileage is less than 23.5 mpl

c. highway mileage is same as the city mileage

Note, make appropriate assumptions, develop the null and alternate hypotheses, evaluate the test statistic, present the threshold value and consequently, make appropriate inferences.