Assaf Bitton

We will calculate some basic summary statistics about the sahp dataset in the **r02pro** package. Please answer the following questions.

You can run the following code to prepare the analysis.

```
library(r02pro)     #INSTALL IF NECESSARY: install.packages("r02pro")

## Warning: package 'r02pro' was built under R version 4.4.3

data("sahp")
```

Q1

a. Use an appropriate R function to find n (the number of observations) and p (the number of variables) for this dataset.

```
dim(sahp)

## [1] 165  12
```

- **Answer (Q1, a):** I used the function **dim()** to print the number of rows and columns. Since rows correspond to the number of observations and columns correspond to the number of variables… the answer is that there are **165 observations and 12 variables.**

b. Which variables in this dataset are categorical? You do not need an R code block for this question. (Hint: Use ? or help function to read the documentation of this dataset where you can learn the details about all variables)

- **Answer (Q1, b):** The answer to this is **nuanced.** The four (4) obvious categorical variables are **house_style, kit_qual, heat_qual, and central_air** – however, **oa_qual** is also a categorical variable, even though it's saved as a numeric **(num)**. You could argue that there are, in fact, **five (5)** categorical variables in this data set.

c. What is the value of the oa_qual variable for the 79th observation in this dataset? Explain the meaning of this value in plain English to someone who is not familiar with this dataset. (Hint: Read the documentation of this dataset)

```
sahp[79, "oa_qual"]

## [1] 3
```

- **Answer (Q1, c):** The *value* of the **oa_qual** variable for the *79th observation* is **3**.

- For starters, this data set consists of sales data for various houses. **oa_qual** is a variable within this data set that represents the "Overall material and finish quality" of these houses. A *value* of **3** means **Fair**, which suggests that the *79th observation* has an *"Overall material and finish quality"* of **Fair.**

Q2

a. What are the values for the last 5 observations of the `bathroom` variable? Explain the meaning of these values to someone who is not familiar with this dataset. (Hint: Read the documentation of this dataset)

```
sahp[161:165, "bathroom"]

## [1] 3.0 1.5 2.0 1.5 2.0
```

- **Answer (Q2, a):** The values for the *last 5 observations* are **3.0, 1.5, 2.0, 1.5, and 2.0.**

- The variable **bathroom** represents the **number of bathrooms** in each of the houses observed in this data set. For example, since the *value* for the *165th observation* is **2**, that would mean there are **2 bathrooms** for that house *(the 165th observation of the data set)*.

b. Fill in the missing part (...) of the following partial `R` program to calculate the population variance of the 5 observations in part a.

```
mean <- (3 + 1.5 + ... )/5
variance <- ((3 - mean)^2 + ... )/5
variance
```

Note that the definition of population variance is:

$$variance = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2.$$

Write the completed program here:

```
mean <- (3 + 1.5 + 2.0 + 1.5 + 2.0)/5
variance <- ((3 - mean)^2 + (1.5 - mean)^2 + (2.0 - mean)^2 + (1.5 - me
an)^2 + (2.0 - mean)^2)/5
variance

## [1] 0.3
```

- **Answer (Q2, b):** Filling in the missing parts was only a matter of following the pattern that follows *what was already provided in the question* and understanding that **variance is the sum of the squared differences between the observations and the mean divided by the number of observations.**

c. Use the `var` function to calculate the sample variance. Is the result the same as the population variance calculated in part b? If not, what could be the reason? (Hint: Use `?` or `help` function to read the documentation of `var` function)

```
bathroom_sample_var <- sahp[161:165, "bathroom"]
var(bathroom_sample_var)
```

```
## [1] 0.375
```

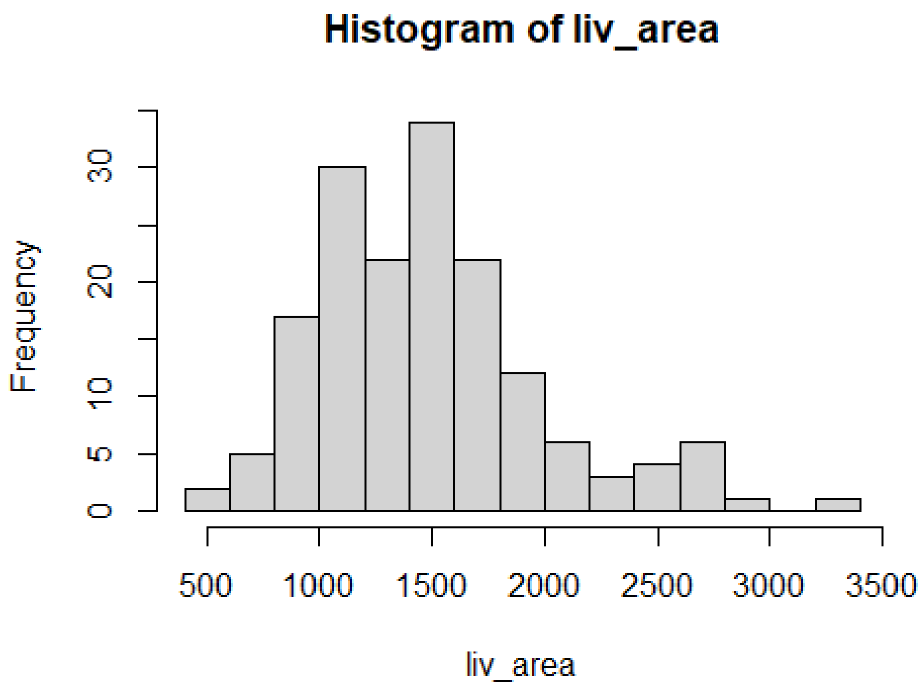
- **Answer (Q2, c):** The answer **is not the same** as the *population variance* calculated in **part b.**

- The **sample variance** function **var()** is described in the documentation as having a "**… denominator [of] *n-1*.**" This is *different* from the denominator of the *population variance*, which has a denominator of *n*. This difference accounts for the difference between the values **0.3** in part *b* and **0.375** in part *c*.

Q3

Plot a histogram of the `liv_area` variable with 15 bins (also known as cells or buckets).

```
attach(sahp)
liv_area_hist <- hist(liv_area, breaks=15)
```



- **Answer (Q3):** I used **?hist** and discovered that **breaks** is an argument which determines the number of cells (bins) for the histogram. I also learned **attach()** *from the textbook*, which allowed me to call **liv_area** directly without me having to write out *sahp$liv_area*.

Q4

Generate 90 random numbers from the standard normal/Gaussian distribution. Fill the numbers **by row** into a 9 × 10 matrix (with 9 rows and 10 columns). What is the value of the entry on the 7th row and 8th column? Use `set.seed(1)` to ensure reproducibility.

```
set.seed(1)
x <- rnorm(90)
x_mat <- matrix(x, 9, 10, T)
x_mat[7,8]

## [1] 1.465555

x[68]

## [1] 1.465555
```

- **Answer (Q4): NOTE:** I used *T* instead of *TRUE* because I learned this was possible in the textbook. I also used **x[68]** to double-check if *1.465555* was truly at position *row = 7* and *col = 8*, since that should be the *68th* ((7-1)x10+8)) position on vector ***x***.

Q5

What do you think of the current pace of the course? Use `R` to print one of the three answers: “too slow”, “about right”, or “too fast”.

```
print("about right")

## [1] "about right"
```