Stat 243 – GS01

Richmond Yevudza

February 02, 2022

```
countries <- data.frame(
    Variable = c("Country", "Code", "LandArea", "Population", "Energy", "Rural",
                 "Military", "Health", "HIV", "Internet", "Developed", "BirthRate",
                 "ElderlyPop", "LifeExpectancy", "CO2", "GDP", "Cell", "Electricity"),
    VariableType = c("Categorical", "Categorical", "Quantitative", "Quantitative",
                     "Quantitative", "Quantitative", "Quantitative", "Quantitative",
                     "Quantitative", "Quantitative", "Categorical", "Quantitative",
                     "Quantitative", "Quantitative", "Quantitative", "Quantitative",
                     "Categorical", "Quantitative")
)
countries
```

1.12a

```
##          Variable VariableType
## 1         Country  Categorical
## 2            Code  Categorical
## 3        LandArea Quantitative
## 4      Population Quantitative
## 5          Energy Quantitative
## 6           Rural Quantitative
## 7        Military Quantitative
## 8          Health Quantitative
## 9             HIV Quantitative
## 10       Internet Quantitative
## 11      Developed  Categorical
## 12      BirthRate Quantitative
## 13     ElderlyPop Quantitative
## 14 LifeExpectancy Quantitative
## 15            CO2 Quantitative
## 16            GDP Quantitative
## 17           Cell  Categorical
## 18    Electricity Quantitative
```

1.12b

1. What is the average population of all countries around the world?
2. Which country has the highest GDP?

1.12c

1. Do the countries with larger area have more access to the internet?

2. Do the countries with smaller area have a lower birth rate?

1.20a The given dataset is about the number of days to row alone the Atlantic Ocean. The number of cases are 8. There are two variables in the given study, namely gender and number of days to cross the Atlantic Ocean. For the Gender variable, the subjects are classified into two groups (male and female). Therefore it is categorical. For the number of days to cross the Atlantic Ocean, the data represents a numerical value. Therefore, it is quantitative.

```
Atlantic <- data.frame(
    Gender = c("M", "M", "M", "M", "M", "F", "F", "F"),
    DaysToRow = c(40, 87, 78, 106, 67, 70, 153, 81)
)
Atlantic
```

1.20b

```
##   Gender DaysToRow
## 1      M        40
## 2      M        87
## 3      M        78
## 4      M       106
## 5      M        67
## 6      F        70
## 7      F       153
## 8      F        81
```

1.22a The information is obtained from 41 people who participated in the study. Therefore there are 41 cases in this study.

1.22b The variables are as following:

- Whether or not the participants were involved in mediation program (categorical)
- Brain wave activity across the front of the left hemisphere measured before study (quantitative)
- Brain wave activity across the front of the left hemisphere measured after study (quantitative)
- Brain wave activity across the front of the left hemisphere measured 4 months later (quantitative)
- Immune response to the vaccine after 1 month (quantitative)
- Immune response to the vaccine after 2 months (quantitative)
- Positive emotions before (quantitative)
- Negative emotions before (quantitative)
- Positive emotions after (quantitative)
- Negative emotions after (quantitative)

1.22c The explanatory variable is a factor that has been manipulated in the experiment by the researcher. In this study, **whether or not the participants were involved in mediation program** is the explanatory variable.

1.22d If it is assumed that each case represents a row, then the dataset will contain **41 rows**. If each variable is a column, then the dataset will contain **at least 11 columns**.

1.24 Data might be collected from a sample of people who are eligible to vote and the following procedure could be followed:

- Ask about the person's political party.
- Ask whether they voted in the last election.

In this dataset, the cases are people who are eligible to vote. The variables are the person's political party, and whether or not they voted in the last election.

1.26 Question: To what extent does your college GPA affect your work salary in the future?

To be able to come up with data for this question I will have to gather a sample of earners from different earning classes (i.e 60-99k, 100-199k, 200-299k...) and find out their GPA. By comparing these two quantitative variables, I might be able to come up with a conclusion.

The cases in this dataset will be the workers. The variables will be GPA and salary.

1.50 This method of data collection is **not biased**.

This is mainly because in this sampling method, a random sample of one type of typewriter is selected and testes for the number of pages that can be printed before the ink runs out. Then after that, teh average number of pages for that can be printed before the ink turns out is estimated for that selected type of typewriter. It is quite clear that the selection is random and the results that are gotten can also be trusted.

1.52 This statistic can be generalized to **all parents in Kansas City**.

1.56 The sample is **not the representative** of all Australians.

This is mainly because the sample is a volunteer sample and therefore the participants are not a representative of the population. In addition, the advertisements for the study were on two rock radio stations in Sydney. This means that it is only the people who hear about it that have the chance to participate and clearly not the whole population hear about it. Also, there are participants who choose because they like alcohol and marijuana. These people tend to know more about the substances than the overall population.

1.58a 300 tanning salons.

1.58b **Yes**, the sample is representative of all tanning salons in the US because the sample is selected in such a way that shows this.

1.58c Although the sample is random, the results don't give an accurate picture of the dangers of tanning because most of the salons are concerned more about marketing what they offer than actually showing facts. So quite a number of the responses given by the salons are not true.

1.58d **Yes**, the study accurately portrays the messages tanning salons give to teenage girls because the sample is selected well.

1.60a All US residents.

1.60b All the emergency room patients in the US.

1.60c

i. NHANES
ii. NHAMCS
iii. NHAMCS
iv. NHANES

2.13a **Sample**: The random 119 players who were observed playing Rock-Paper-Scissors.

Population: All people who play Rock-Paper-Scissors.

The variable measures the option selected (rock, paper or scissors) which is categorical data. It also measures the frequency which is quantitative data.

```
RockPaperScissors <- data.frame(
    OptionSelected = c("Rock", "Paper", "Scisors"),
    Frequency = c(66, 39, 14)
)
data.frame(
  OptionSelected = RockPaperScissors$OptionSelected,
  Frequency = RockPaperScissors$Frequency,
  RelativeFrequency = RockPaperScissors$Frequency/sum(RockPaperScissors$Frequency)
  )
```

2.13b

```
##   OptionSelected Frequency RelativeFrequency
## 1           Rock        66         0.5546218
## 2          Paper        39         0.3277311
## 3        Scisors        14         0.1176471
```

2.13c For the odds to be in your favor, your best bet would be to play **paper** as rock's relative frequency is 0.55 (the highest of them all).

2.13d The option picked for the next round should be **scissors** because the players are more likely to repeat the option paper.

```
TwoWayTable <- matrix(c(363, 176, 196, 735, 557, 466, 789, 1812, 20, 26, 32, 78, 940,
                        668, 1017, 2625), ncol=4, byrow=TRUE)
rownames(TwoWayTable) <- c("Agree", "Disagree", "Don't know", "Total")
colnames(TwoWayTable) <- c("HS", "Some", "College", "Total")
TwoWayTable
```

2.18a

```
##             HS Some College Total
## Agree      363  176     196   735
## Disagree   557  466     789  1812
## Don't know  20   26      32    78
## Total      940  668    1017  2625
```

2.18b Required percentage $= \frac{363}{940} = 0.386 = 38.6\%$

Therefore, the percentage of agree with high school degree or less is **38.6%**.

Required percentage $= \frac{176}{668} = 0.263 = 26.3\%$

Therefore, the percentage of agree with some college is **26.3%**.

Required percentage $= \frac{196}{1017} = 0.193 = 19.3\%$

Therefore, the percentage of agree with college graduate or higher is **19.3%**.

2.18c Required percentage $= \frac{1017}{2625} = 0.387 = 38.7\%$

Therefore, **38.7%** of the people participating in the survey have a college degree or higher.

2.18d Required percentage $= \frac{557}{1812} = 0.307 = 30.7\%$

Therefore, **30.7%** of people who disagreee with the statement have a college degree or higher.

2.19a Percentage of female smokers in female sample,

$\frac{16}{169} = 0.095 = 9.5\%$

Percentage of male smokers in male sample,

$\frac{27}{193} = 0.140 = 14.0\%$

Therefore, the gender with the higher percentage of smokers is the **male**.

2.19b Required proportion $= \frac{43}{362} = 0.119 = 11.9\%$

Therefore, the proportion of smokers for the whole sample is **11.9%**.

2.19c Required proportion $= \frac{16}{43} = 0.372 = 37.2\%$

Therefore, the proportion of smokers in the sample who are female is **37.2%**.

2.20a This experiment is an observational study. The two variables are.**Does the person have dyslexia or not** and **Does the person have the DYXCI break or not**.

2.20b The total number of "Dyslexia group" is 109 and the total nuber of "Control group" is 195. Therefore, the total number of rows (cases) in the study is **304**.

The total number of columns is**2** as there are two variables in this data set.

```
TwoWayTable <- matrix(c(10, 99, 109, 5, 190, 195, 15, 289, 304), ncol=3, byrow=TRUE)
rownames(TwoWayTable) <- c("Dyslexia group", "Control group", "Total")
colnames(TwoWayTable) <- c("Gene break", "No break", "Total")
TwoWayTable
```

2.20c

```
##                Gene break No break Total
## Dyslexia group         10       99   109
## Control group           5      190   195
## Total                  15      289   304
```

2.20d Proportion of "Dyslexia group" with "Gene break" $= \frac{10}{109} = 0.902$

Proportion of "Control group" with "Gene break" $= \frac{5}{195} = 0.026$

Therefore, looking at the proportion of each group who have the break on DYXCI gene, a conclusion can be drawn that there appears to be a very substantial difference between the group with dyslexia and the control group.

2.20e There appears to be an association between the genetic marker and dyslexia for the people in this sample.

2.20f If the association appears to be strong, then it can't be assumed that the gene distribution causes dyslexia because the data set comes from the observational study but not from an experiment.

2.55a The shape of the distribution of ratings of word is **left skewed** because the left side of the figure is extended larger than the right side.

2.55b Looking at Figure 2.14, the value that separates the area into almost two equal parts is 6.5. Therefore, the median value is around **6.5**.

2.55c The value of the mean will be **smaller** than the median because the shape of the shape of the distribution of ratings of word is skewed to the left which shows that the mean is smaller than the median.

2.60a $\overline{x}_f = \frac{\sum(x)}{n} = \frac{64}{10} = 6.4$

Therefore, the mean number of hours spent exercising by the females is **6.4**.

2.60b $\overline{x}_m = \frac{\sum(x)}{n} = \frac{177}{26} = 6.81$

Therefore, the mean number of hours spent exercising by the males is **6.81**.

2.60c $\overline{x}_m - \overline{x}_f = 6.81 - 6.4 = 0.41$

This shows that the average hours spent on exercise by the males is higher than the average hours spent on exercise by the females by **0.41**.

2.63a Data set: 1, 49, 50, 52, 56

$\overline{x} = \frac{\sum(x)}{n} = \frac{208}{5} = 41.6$

Median: 50

Mean: 41.6

2.63b Data set: 1, 20, 30, 80, 100

$\overline{x} = \frac{\sum(x)}{n} = \frac{231}{5} = 46.2$

Median: 30

Mean: 46.2

2.82

- Minimum: 58
- First quartile: 65
- Median: 68
- Third quartile: 70
- Maximum: 77

2.91 The required interval is $\overline{x} \pm 2s$, where $\overline{x}$ is the mean and s is the standard deviation.

$\overline{x} \pm 2s = 200 \pm 2(25) = 200 \pm 50 = (150, 250)$

Therefore, the 95% of the data values lies between **150 and 250**.