Final Exam 4800

DANA 4800

2023-12-06

Name: An Phan Hoang Nguyen

Student ID: 100404103

YOU GET 5 MARKS FOR A NEAT PRESENTATION. DEDUCTED FOR UNNECESSARY INFORMATION PRESENTED.

QUESTION 1 - Concepts and Application

The first portion of the exam tests your critical thinking skills when reviewing data in the primary literature context.

a) List and explain two reasons why we should view a data set after reading it into the software. [6 points]

WRITE YOUR ANSWER BELOW

1. Check for missing data (Null/Na) and determine what type of missing data we're dealing with (Missing completely at random (MCAR), Missing at random (MAR), or missing not at random (MNAR))
2. Detect any abnormality that could present in the dataset that could let to error while running codes (example: different format of date at some rows)

b) A researcher wants to summarize a number of numerical variables in his data set. What measures of variation/spread can be used to summarize the variables? When is it appropriate to use these measures of variation/spread you have mentioned? [6 points]

WRITE YOUR ANSWER BELOW

Histogram or boxplot for starters can be a good approach here to check, as well as statistical summary of Mean, Median, IQR, STDEV, Min, Max, etc. For variation/spread, if there are too many outliers, STDEV might be not a good choice here so IQR can be used to determine how spread out the values are from the Mean.

c) What is the difference between descriptive statistics and inferential statistics? [6 points]

WRITE YOUR ANSWER BELOW

Descriptive statistics is used to summarize and describe the features about the sample or population being studied through measures of central tendency (mean, median,…), measures of dispersion (range, variance, STDEV), and visualizations (histograms, boxplots,…). No generalizations beyond the data being analyzed but only aim to describe what already there in the dataset.

Inferential statistics involve making inferences or predictions about a population based on a sample taken from that population, using probability theory or linear regression equation to draw conclusions, make predictions, or test hypotheses about a larger population. In other words, it takes a step further from descriptive statistics.

QUESTION 2

Records in a small city show that the distribution of prices of homes in the city has a mean of $140000 and a standard deviation of $55000. The interquartile range for the prices of homes in the city is $40000 with a third quartile of $125000. If a random sample of 100 homes are randomly selected with their prices recorded,

a) What do you think is the shape of the distribution of the prices of homes in the city? Explain your answer.

[5 points]

WRITE YOUR ANSWER BELOW

It is possible that the shape of distribution here is right-skewed, where the tail of the distribution extends towards higher prices, because the mean is higher than the third quartile.

b) What is the probability that the average price of the 100 homes is between $138000 and $150000?

[6 points]

WRITE YOUR ANSWER BELOW

```
# n = sample size = 100 >= 30
# n*p = sample size*prob of success >= 5
# n*(1-p) = sample size*(1-prob of success) >= 5
# We can consider this is normal distribution sample, aka qualified to use
pnorm
# Standard error = sd = STDEV/sqrt(10) = 55000/sqrt(100)
pnorm(150000, 140000, 55000/sqrt(100)) - pnorm(138000, 140000,
55000/sqrt(100))

## [1] 0.607417
```

The probability that the average price of the 100 homes between $138000 and $150000 is 60.74%

c) What is the probability that more than 22.3% of the 100 homes will have a price less than $85000?

```
                                                            [6 points]
```

WRITE YOUR ANSWER BELOW

```
# Find the probability in the population that have a house less than $85000
pnorm(85000, 140000, 55000)

## [1] 0.1586553
```

We get 15.86% here as p

```
# n = sample size >= 30
# n*p = sample size*prob of success >= 5
# n*(1-p) = sample size*(1-prob of success) >= 5
# Meet 2/3 of the above then we can consider this is normal distribution
sample, aka qualified to use pnorm
# STDEV = sqrt(p*(1-p)/n) = sqrt(0.158*(1-0.158)/100)
pnorm(0.223, 0.158, sqrt(0.158*(1-0.158)/100), lower.tail = FALSE)

## [1] 0.03736757
```

The probability that more than 22.3% of the 100 homes that have a price less than $85000 is 3.73%

QUESTION 3

The file ‘weather.csv’ contains various weather-related measurements (DO NOT WORRY ABOUT THE UNITS). Within the dataset, there are different cities. Import the dataset, and answer the following questions:

a) How many observations and variables are there in the dataset? Describe the types of variables in the dataset. [5 points]

WRITE YOUR ANSWER BELOW

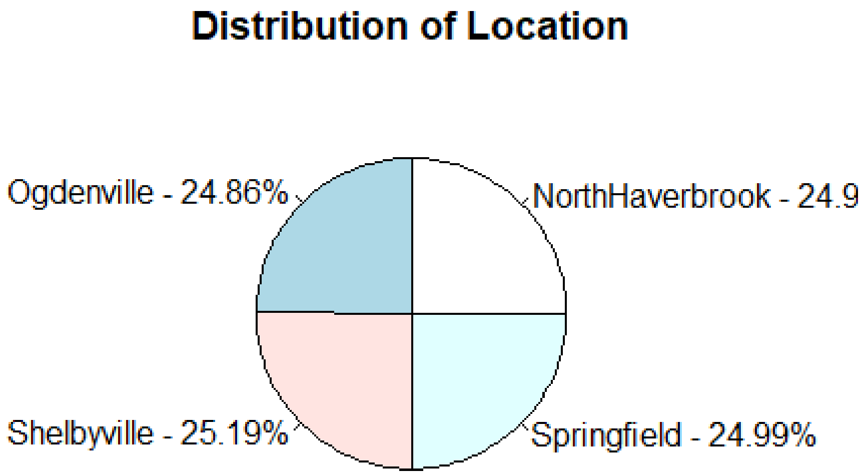
```
# View the dataset first
weather <- read.csv('weather.csv', header = TRUE)
View(weather)

# Structure of the dataset
str(weather)

## 'data.frame':    11987 obs. of  7 variables:
##  $ Location: chr  "Shelbyville" "Shelbyville" "Shelbyville" "Shelbyville"
...
##  $ MinTemp : num  8.8 12.7 6.2 5.3 7.6 5.3 8.4 9.5 10 11.1 ...
##  $ MaxTemp : num  15.7 15.8 15.1 15.9 11.2 13.5 14.3 13.1 15.4 14.7 ...
##  $ Rainfall: num  5 0.8 0 0 16.2 17 1.8 9 3.8 4 ...
##  $ Humidity: num  92 75 81 71 83 73 90 54 85 91 ...
##  $ Pressure: num  1017 1022 1028 1029 1016 ...
##  $ Month   : int  7 7 7 7 7 7 7 7 7 7 ...
```

Comment: There are 7 variables in total and 11987 observations (aka rows). The types of variables are as above: categorical variables are Location and Month while numerical variables are MinTemp, MaxTemp, Rainfall, Humidity, Pressure.

b) How many Location levels are found within this dataset? Summarize the location variable with ONE statistic and ONE graph. Comment on your output. [8 points]

WRITE YOUR ANSWER BELOW

```
# Relative frequency table
prop.table(table(weather$Location))

##
## NorthHaverbrook      Ogdenville     Shelbyville     Springfield
##       0.2495203       0.2486027       0.2519396       0.2499374

# Plot pie chart
pie(prop.table(table(weather$Location)), labels = c("NorthHaverbrook -
24.95%", "Ogdenville - 24.86%", "Shelbyville - 25.19%", "Springfield -
24.99%"), main = "Distribution of Location")
```



Comment: There are 4 location levels in total in the weather dataset. Furthermore, the graph demonstrates similar proportions of nearly a quarter for every location in the dataset.

c) Describe the Rainfall and Humidity variables for each Location with ONLY statistics.

[16 points]

WRITE YOUR ANSWER BELOW

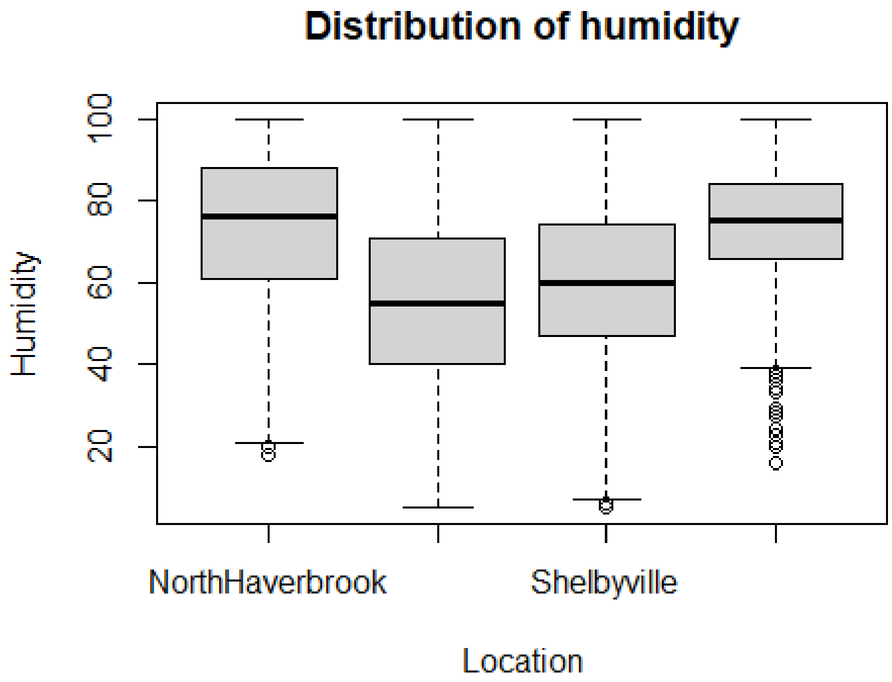
```
# Subset the data first
NorthHaverbrook <- subset(weather, weather$Location == "NorthHaverbrook")

Ogdenville <- subset(weather, weather$Location == "Ogdenville")

Shelbyville <- subset(weather, weather$Location == "Shelbyville")

Springfield <- subset(weather, weather$Location == "Springfield")

# Rainfall and Humidity variables in NorthHaverbrook
summary(NorthHaverbrook$Rainfall)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.926   0.400 104.200

summary(NorthHaverbrook$Humidity)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   61.00   76.00   74.11   88.00  100.00

# Rainfall and Humidity variables in Ogdenville
summary(Ogdenville$Rainfall)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.129   0.000  68.000

summary(Ogdenville$Humidity)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   40.00   55.00   55.46   71.00  100.00

# Rainfall and Humidity variables in Shelbyville
summary(Shelbyville$Rainfall)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.572   0.800  75.200

summary(Shelbyville$Humidity)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     5.0    47.0    60.0    59.6    74.0   100.0

# Rainfall and Humidity variables in Springfield
summary(Springfield$Rainfall)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   2.255   1.800  49.400

summary(Springfield$Humidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   16.00   66.00   75.00   74.74   84.00  100.00
```

d) Graphically compare the distribution of Humidity for the Locations. What are the key observations? [6 points]

WRITE YOUR ANSWER BELOW

```
boxplot(weather$Humidity ~ weather$Location, horizontal = FALSE, main
="Distribution of humidity", xlab = 'Location', ylab = 'Humidity')
```



Comment: NorthHaverbrook seems to have the highest average humidity among 4 locations while Ogdenville is the lowest. Lots of outliers in Springfield. Mean and Median of them are roughly the same so it might be possible for normal distribution of humidity (bell-shaped) in each location, except for Springfield since it has many outliers on the lower end values.

e) Perform a test on the data to determine if Springfield is more Humid than Ogdenville at 5% level of significance. Assume the population variances are unequal. Do your results match your observations in part (d)? [10 points]

WRITE YOUR ANSWER BELOW

Let mu1 represent the population mean Humidity value in Springfield

Let mu2 represent the population mean Humidity value in Ogdenville

1. H0: mu1 <= mu2 OR mu1-mu2=0 H1: mu1 > mu2 OR mu1-mu2>0

2. Level of significance = 0.05

3. One sided test (because the H1 is greater than)

```
t.test(Springfield$Humidity, Ogdenville$Humidity, mu = 0, alternative =
"greater",var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Springfield$Humidity and Ogdenville$Humidity
## t = 42.525, df = 4803.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  18.53868      Inf
## sample estimates:
## mean of x mean of y
##  74.74129  55.45654
```

Comment: Because p-value is smaller than 0.05 so we reject the Null hypothesis, in other words, true mean Humidity of Springfield is greater than in true mean Humidity in Ogdenville at 5% level of significance. Additionally, it also matches with the observation in (d)

f) A meteorologist in Ogdenville claims that true proportion of times that the amount rainfall in the year exceeds 1.129 is more than 0.12. Perform at statistical test for this claim about Rainfall in the city at a 5% level of significance. Hint: Midterm 2 [10 points]

WRITE YOUR ANSWER BELOW

```
# Define a new column of rainfall > 1.129 as High in Ogdenville
Ogdenville$rainfall_cat <- ifelse(Ogdenville$Rainfall > 1.129 ,"High", "Low")

# Probability of Rainfall higher
table(Ogdenville$rainfall_cat)
```

```
##
## High  Low
##  384 2596
```

```
prop.test(x = 384/2596, n = 2980, p = 0.12, alternative = "greater", correct
= FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  384/2596 out of 2980, null probability 0.12
## X-squared = 406.03, df = 1, p-value = 1
## alternative hypothesis: true p is greater than 0.12
## 95 percent confidence interval:
##  2.452308e-06 1.000000e+00
## sample estimates:
##            p
## 4.963754e-05
```

g) What type of error are you at risk of committing from the test in (f)? Explain in the context of the question. [5 points]

WRITE YOUR ANSWER BELOW

We are at the risk of committing a Type I error. This means we are rejecting the Null hypothesis that the Humidity of Springfield is smaller or equal to Humidity in Ogdenville, when this Null hypothesis might be true in reality.