MTH 324, Final Exam, At Home, Fall 2023 Name KEY

Instructions: This exam is in two parts: Part I is to be completed partly at home using the materials posted in the course for the at-home portion and you will answer questions about that work during the in-class portion of the exam; Part II is to be completed entirely in class. You may not use cell phones, and you may only access internet resources you are specifically directed to use.

At home, prepare for questions in Part I using R. Open the data file entitled **324final_data.xlsx** posted in Blackboard. (Note: this file has multiple sheets of data. You may want to separate the data into separate files to upload to R, or look up how to access multiple sheets in R from a single upload.) Complete the calculations noted below. You will be asked for additional analysis and interpretation of this data in the in-class portion of the test. Print out the results of your analysis and code, and bring the pages with you to the exam. You will submit all this work along with the in-class exam.

1. Using the data on Sheet 1, conduct appropriate hypotheses of the data on Sales of two products. Conduct a parametric version of the test, and a nonparametric version. Is there sufficient reason to think one product sales is more than the other? Test any assumptions of the hypothesis test you choose, and be prepared to explain your reasoning why you chose the test you did.

2. The data on Sheet 2 is complex with 12 variables (Person is not a variable).
   a. Explore the data. Make a graph of each variable. Use an appropriate graph for each variable type. For numerical variables, you'll need to be able to describe the shape of the distribution. Bar graphs or pie charts are appropriate for categorical data, or you can experiment with mosaic charts or others that you think display the individual variables well.

   b. Create summary statistics for the numerical variables.

   c. Create a Two-way table for Dwell Type and Pay Type. Conduct a test of independence on this data.

   d. Create a comparative box plot of Weight and Neighborhood. Conduct an ANOVA test on this data. If you reject the null, apply Tukey's method or pairwise comparisons as appropriate.

   e. Use bootstrapping to test whether there is good reason to think average Income is more than $45,000. Construct a confidence interval with this method, and using a traditional one-sample method. Compare the results.

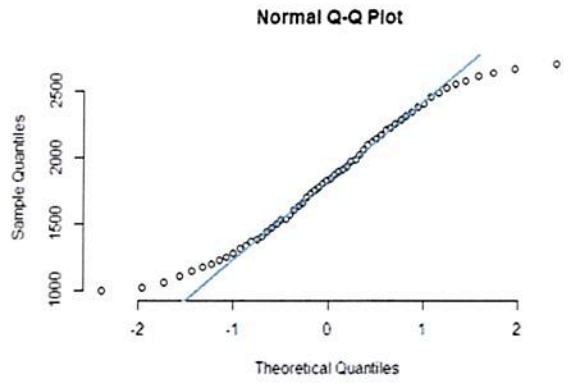
   f. Construct a two-way table of Gender and Pay Type. Conduct a two-sample proportion test to see if the proportion of people living alone differs by gender.

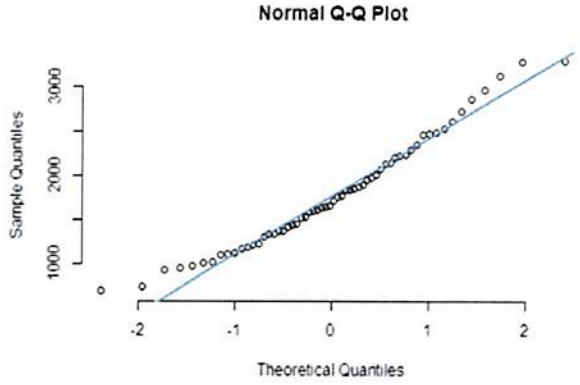
3. The table below shows the results of a simulation of the sum of rolls of two pairs of dice using 492 trials.

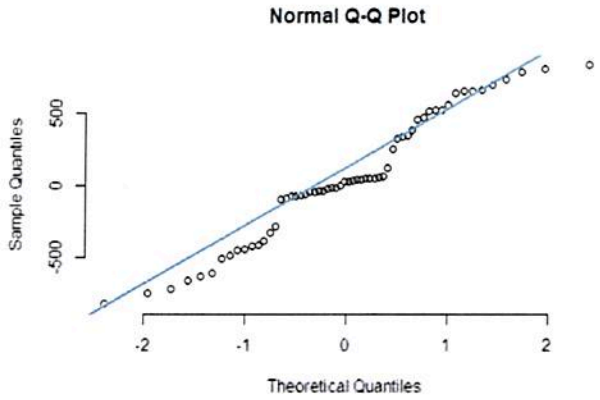
	2	3	4	5	6	7	8	9	10	11	12
Count	14	25	38	60	68	91	61	64	34	19	18

Dice rolls follow a triangle distribution, with $P(X = 2) = \frac{1}{36}$, which increases linearly to $P(X = 7) = \frac{6}{36}$, and then decreases linearly to $P(X = 12) = \frac{1}{36}$. Conduct a goodness-of-fit test to see if this data appears to follow this distribution.

MTH 324 Final Exam At-home Analysis, Fall 2023





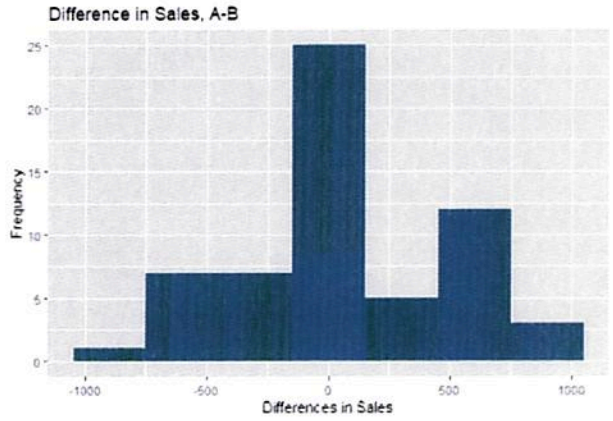


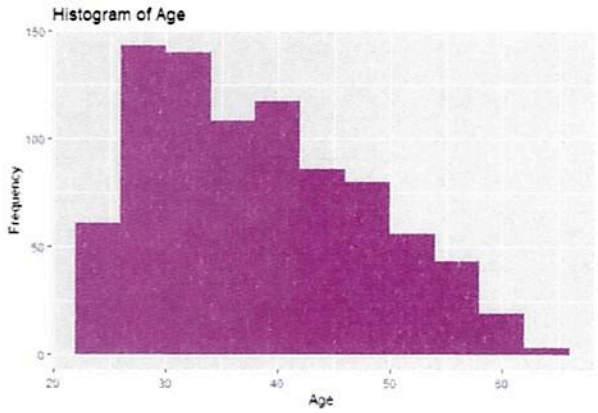
Paired t-test

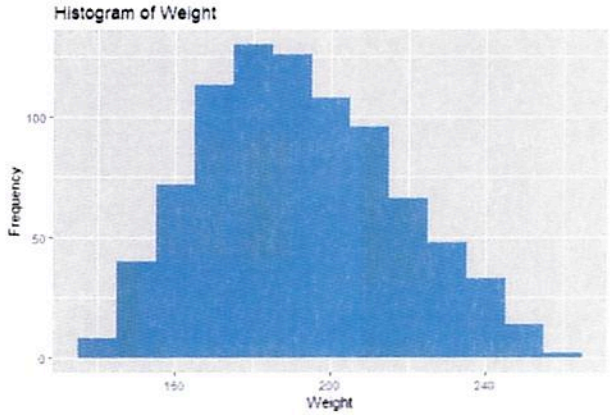
```
data:  x and y
t = 0.95244, df = 59, p-value = 0.3448
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -58.64219 165.17553
sample estimates:
mean difference
       53.26667
```

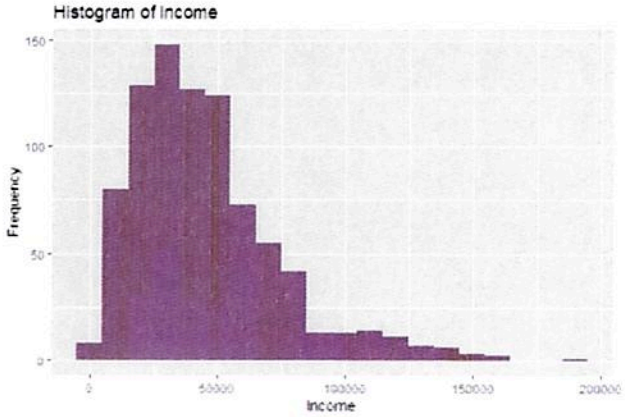
Wilcoxon signed rank test with continuity correction

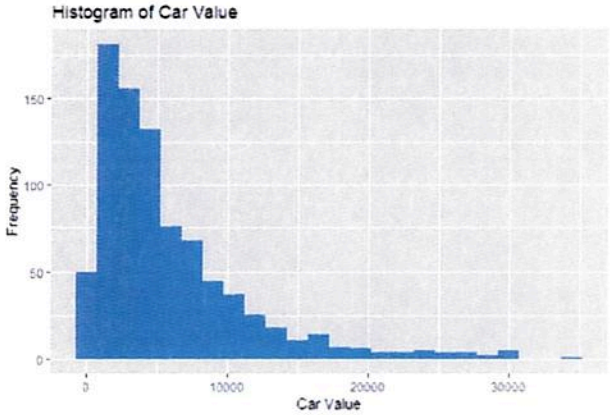
```
data:  x and y
V = 1032, p-value = 0.3911
alternative hypothesis: true location shift is not equal to 0
```

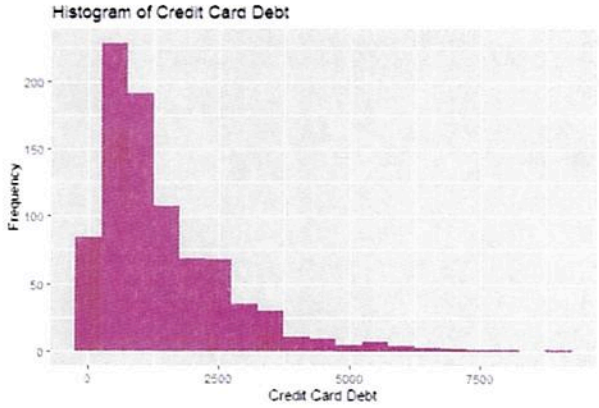

Difference in Sales, A-B
Frequency
Differences in Sales

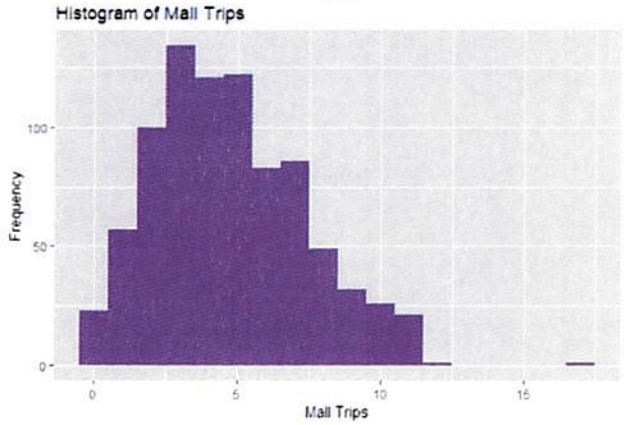

Histogram of Age
Frequency
Age

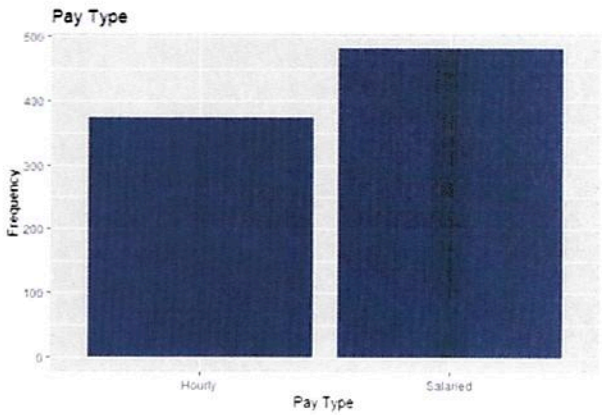

Histogram of Weight
Frequency
Weight

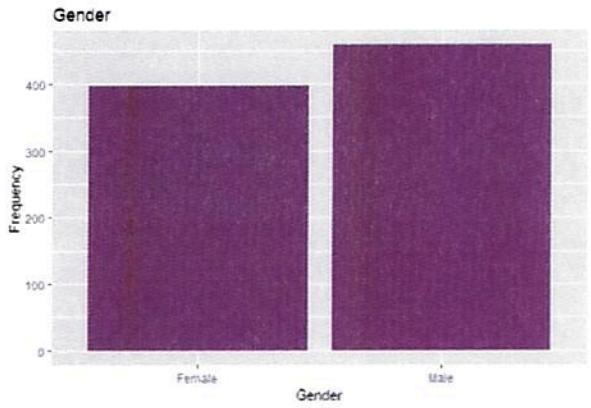

Histogram of Income
Frequency
Income

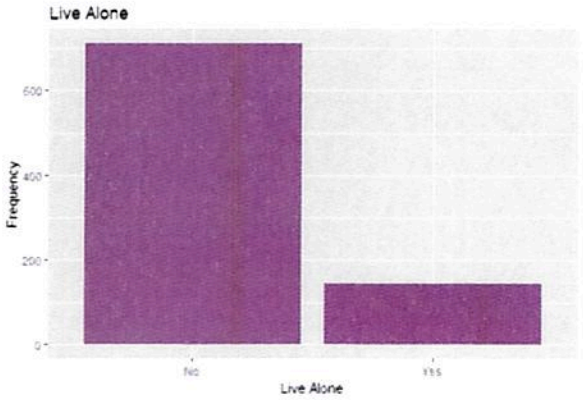

Histogram of Car Value
Frequency
Car Value

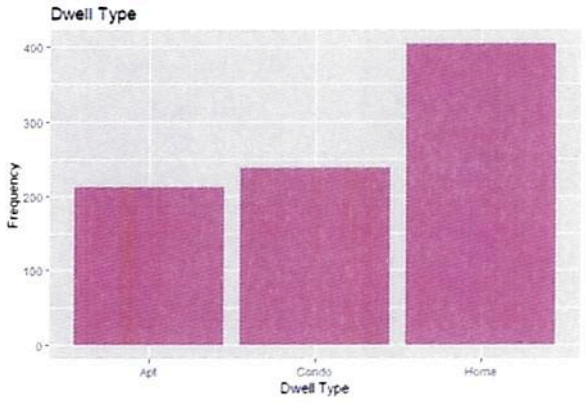

Histogram of Credit Card Debt
Frequency
Credit Card Debt

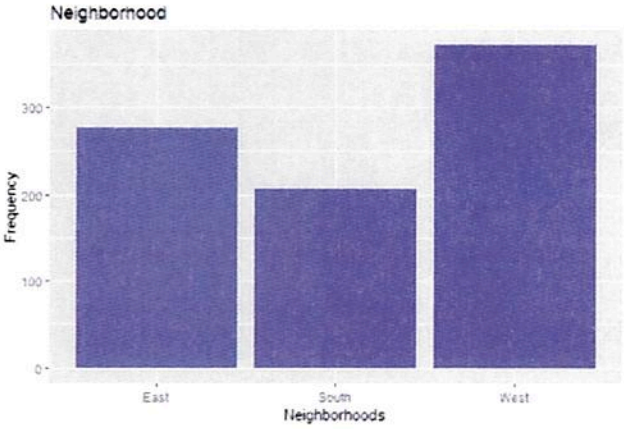

Histogram of Mall Trips
Frequency
Mall Trips

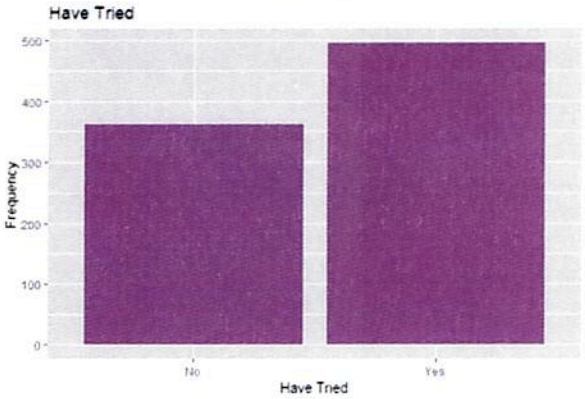

Pay Type
Frequency
Hourly
Salaried
Pay Type









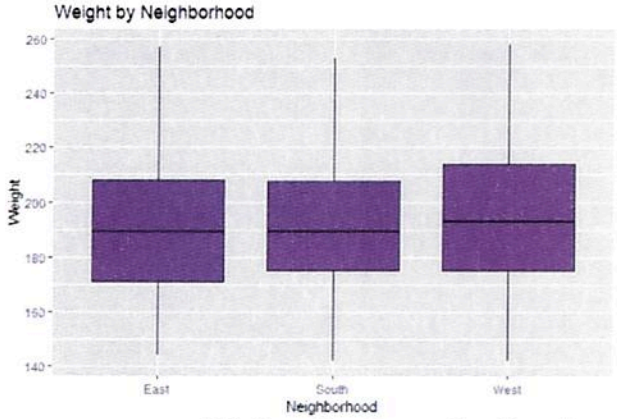



```
data2$`Car Value`
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
     856        0      613        1     5908     5462    677.5   1090.0   2110.0
     .50      .75      .90      .95
  4175.0   7717.5  12605.0  17135.0

lowest :   130   150   210   280   290, highest: 29780 29910 29970 30710 33870
> Hmisc::describe(data2$`CC Debt`)
data2$`CC Debt`
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
     856        0      337        1     1431     1294      130      265      560
     .50      .75      .90      .95
    1020     1972     3165     3785

lowest :    0   70   80   90  100, highest: 6970 7000 7460 8080 8960
```

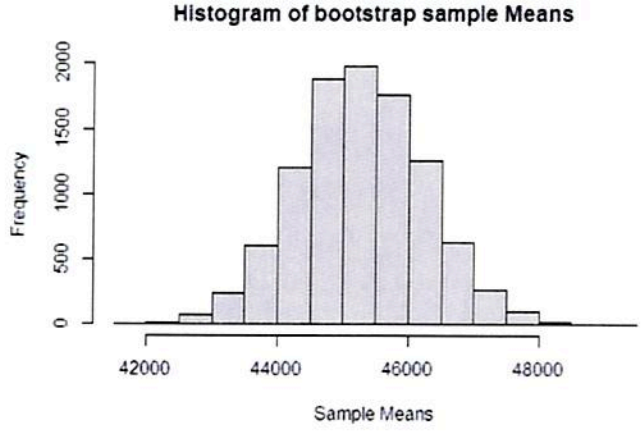
```
data2$Age
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
     856        0       42    0.999    38.78    10.94     25.0     27.0     31.0
     .50      .75      .90      .95
    37.5     46.0     53.0     57.0

lowest : 22 23 24 25 26, highest: 59 60 61 62 64
> Hmisc::describe(data2$Weight)
data2$Weight
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
     856        0      112        1    192.7    28.21      155      162      174
     .50      .75      .90      .95
     190      210      228      237

lowest : 142 144 145 146 147, highest: 253 254 255 257 258
> Hmisc::describe(data2$Income)
data2$Income
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
     856        0      546        1    45267    30438     9975    14850    24475
     .50      .75      .90      .95
   39950    58225    80800   106950

lowest :   2600   3100   3200   3800   4400, highest: 149600 152800 155600 161600 19050
0

data2$`Mall Trips`
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
     856        0       14    0.986    4.735    2.959        1        2        3
     .50      .75      .90      .95
       4        7        8       10

Value       0.00  0.85  1.87  2.89  3.91  4.93  5.95  6.97  7.99  8.84  9.86 10.88
Frequency     23    57   100   134   121   122    83    86    49    32    26    21
Proportion 0.027 0.067 0.117 0.157 0.141 0.143 0.097 0.100 0.057 0.037 0.030 0.025

Value      11.90 17.00
Frequency      1     1
Proportion 0.001 0.001

For the frequency table, variable is rounded to the nearest 0.17
> psych::describe(data2$Age)
   vars   n  mean   sd median trimmed   mad min max range skew kurtosis   se
X1    1 856 38.78 9.61   37.5   38.21 11.12  22  64    42 0.45    -0.69 0.33
> psych::describe(data2$Weight)
   vars   n   mean    sd median trimmed  mad min max range skew kurtosis   se
X1    1 856 192.66 24.75    190  191.79 25.2 142 258   116  0.3    -0.55 0.85
> psych::describe(data2$Income)
   vars   n     mean       sd median  trimmed      mad  min    max  range skew
X1    1 856 45266.94 28631.29  39950 41744.46 24388.77 2600 190500 187900 1.33
   kurtosis    se
X1     2.27 978.6
> psych::describe(data2$`Car Value`)
   vars   n    mean      sd median trimmed     mad min   max range skew kurtosis
X1    1 856 5908.48 5533.46   4175 4920.73 3587.89 130 33870 33740 1.97     4.53
       se
X1 189.13
> psych::describe(data2$`CC Debt`)
   vars   n   mean      sd median trimmed    mad min  max range skew kurtosis    se
X1    1 856 1431.2 1278.04   1020 1232.46 889.56   0 8960  8960 1.84     4.56 43.68
> psych::describe(data2$`Mall Trips`)
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 856 4.73 2.64      4    4.59 2.97   0  17    17 0.53     0.02 0.09
```

Count of Pay Type	Column Labels		
Row Labels	Hourly	Salaried	Grand Total
Apt	81	132	213
Condo	97	142	239
Home	197	207	404
Grand Total	**375**	**481**	**856**

```
      Hourly Salaried
Apt       81      132
Condo     97      142
Home     197      207
> chisq.test(observed_table)

        Pearson's Chi-squared test

data:  observed_table
X-squared = 7.927, df = 2, p-value = 0.019
```




```
             Df Sum Sq Mean Sq F value Pr(>F)
Nbhd          2   3274    1637   2.683 0.0689 .
Residuals   853 520314     610
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
means_sorted[250]
[1] 43398.13
> means_sorted[9751]
[1] 47196.26

mean(data2$Income)-qt(0.975,855)*sd(data2$Income)/sqrt(856)
[1] 43346.2
> mean(data2$Income)+qt(0.975,855)*sd(data2$Income)/sqrt(856)
[1] 47187.67
        One Sample t-test

data:  data2$Income
t = 0.27278, df = 855, p-value = 0.7851
alternative hypothesis: true mean is not equal to 45000
95 percent confidence interval:
 43346.20 47187.67
sample estimates:
mean of x
 45266.94
```

Count of Pay Type	Column Labels		
Row Labels	Hourly	Salaried	Grand Total
Female	186	212	398
Male	189	269	458
Grand Total	**375**	**481**	**856**

```
        2-sample test for equality of proportions with continuity correction

data:  c(186, 189) out of c(398, 458)
X-squared = 2.3684, df = 1, p-value = 0.1238
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01427512  0.12362098
sample estimates:
   prop 1    prop 2
0.4673367 0.4126638

        Chi-squared test for given probabilities

data:  observed
X-squared = 9.4268, df = 10, p-value = 0.4921
```

MTH 324, Final Exam, Fall 2023 Name KEY

Instructions: Answer each question thoroughly. For questions in Part 1, use the work you did at home to answer the questions. Be sure to answer each part of each question. In Part 2, report exact answers unless directed to round.

Part I:
Use the work you did at home to answer these questions about tax paid and the neighborhoods in our dataset.

1. Based on the data from sheet 1 on sales data, is the data distributed normally or approximately so? Explain.

approximately so, yes

2. Report the results of your non-parametric test of the sales data. Which test did you use and why? Clearly state your hypotheses, the conclusion in the context of the problem, and explain why you came to that conclusion.

$H_0: \mu_1 = \mu_2$
$H_a: \mu_1 \neq \mu_2$
fail to reject

p-value : 0.3911

the sales values are approximately equal

3. Based on the graphs of second data set, which of the numerical variables was most right skewed?

Car value

CC debt is a close 2nd

4. Based on the graphs of the second data set, which of the categorical variables were the most uniform?

gender

5. Based on the numerical summary of car value, provide the 5-number summary.

min - 130
Q1 - 2110
med - 4175
Q3 - 7717.5
max - 33870

6. Describe the results of your test of Independence for Dwell Type and Pay Type. State your hypotheses, your conclusion in the context of the problem and explain your reasoning.

H_0: independent
H_a: dependent

p-value = 0.019

reject null

data is dependent

7. Describe the results of your ANOVA test for Weight and Neighborhood. State your hypotheses, your conclusion in the context of the problem and explain your reasoning. Does your result agree with your boxplot?

$H_0: \mu_i = \mu_j \quad \forall\, i \neq j$
$H_a: \mu_i \neq \mu_j$ for some $i \neq j$

p-value = 0.0689

fail to reject null

there is no weight diffs by neighborhood

yes, it agrees.

8. Give your confidence interval from your bootstrap sample of Income. Give your confidence interval (95%) for the one-sample t- or z-interval. How do they compare?

answers will vary

($43,398.13, $47,196.26) bootstrap

t-interval ($43,346.20, $47,189.67)

9. Describe the results of your test of two proportions for Pay Type by Gender. State your hypotheses, your conclusion in the context of the problem and explain your reasoning.

$H_0: p_1 = p_2$

$H_a: p_1 \neq p_2$

p-value = 0.1238

fail to reject null

not sufficient evidence to there pay type varies significantly by gender

10. Based on the table in problem #3 on the at-home portion, describe the results of your goodness-of-fit test. State your hypotheses, your conclusion in the context of the problem and explain your reasoning.

Ho: fits the triangle distribution

Ha: does not fit

p-value = 0.4921

fail to reject null

this does follow the distribution

Part II:

11. If you needed to create a stratified sample in R (let's say on Pay Type), explain how you would go about doing that. (I don't need the code, but explain your steps in words.)

Separate the data by categorical values.

determine the proportion of population in each group.

Proportion the sample to be collected accordingly

Sample the required number from each strata

recombine into a single sample

12. A two-way table of Dwell Type and Neighborhood is shown below. Use it to answer the following questions.

	East	South	West	Grand Total
Apt	69	50	94	213
Condo	80	57	102	239
Home	128	100	176	404
Grand Total	277	207	372	856

a. What is the probability that a random person selected from this data set is from the South neighborhood?

$207/856$

b. What is the probability that a random person selected from this data set lives in an Apartment?

$213/856$

c. What is the probability that a random person selected from this data set is from the South neighborhood and lives in an Apartment?

$50/856$

d. What is the probability that a random person selected from this data set is from the South neighborhood or lives in an Apartment?

$$\frac{207+213-50}{856} = \frac{370}{856}$$

e. What is the probability that a random person selected from this data set is from the South neighborhood given that they live in an apartment?

$50/213$

f. Are the variables Neighborhood and Dwell Type independent or dependent? Does your answer differ if you consider only the descriptive properties of the table, or if you infer the answer from the hypothesis you conducted in Part 1? If it does, explain why.

$P(\text{South}) = \frac{207}{856} = 0.2418\ldots$

$P(\text{South} \mid \text{Apt}) = \frac{50}{213} = 0.2347\ldots$

these are not the same so they are dependent in the data set

13. The proportion of women in the sample on sheet 2 of the data set from the at-home portion of the exam is 0.465. If we were to randomly select 23 subjects from that data set, answer the following questions about the probability of possible outcomes.

a. What is the probability of getting exactly 16 women in the sample?

0.01469

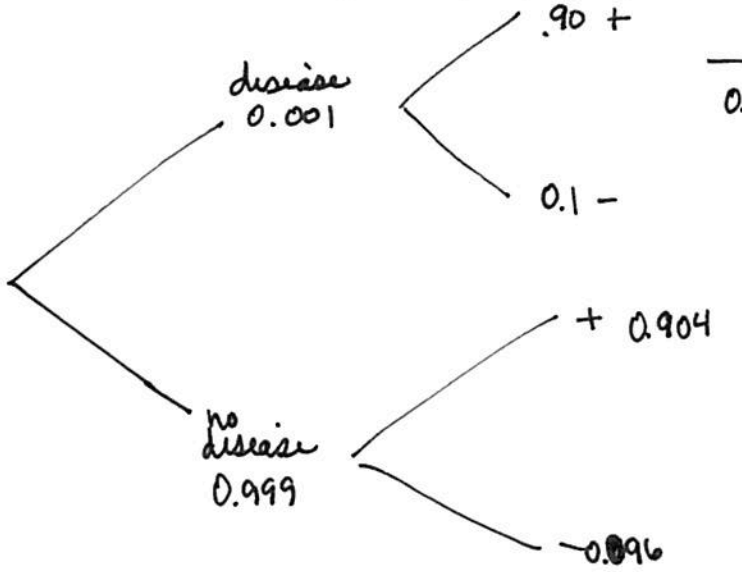
b. What is the probability of having fewer than 5 women in the sample?

0.0036327...

c. What is the expected number of women in the sample?

$23(0.465) = 10.695$

14. Suppose that 0.1% of people have a certain genetic defect. Further suppose that 90% of tests for the gene detect the defect (true positives), and 9.6% of the tests are false positives. If a person gets a positive test result, what is the probability they actually have the genetic defect?



$$\frac{0.001 \times 0.9}{0.001 \times 0.9 + 0.999 \times 0.096} = 0.009297\ldots$$

15. Consider the probability density function $f(x) = \frac{21\sqrt{x}}{8}(1 - x^2), 0 \leq x \leq 1$ (it is equal to 0 everywhere else). Use this information to answer the questions that follow.

a. Verify that this function represents a valid probability distribution.

$$\int_0^1 \frac{21\sqrt{x}}{8}(1-x^2)\,dx = 1 \quad \checkmark$$

b. Find $P\left(\frac{1}{9} \leq X \leq \frac{1}{4}\right)$

$$\int_{1/9}^{1/4} \frac{21\sqrt{x}}{8}(1-x^2)\,dx = 0.1484\ldots$$

c. Find the mean (expected value) of the distribution.

$$E(X) = \int_0^1 \frac{21x\sqrt{x}}{8}(1-x^2)\,dx = 0.46\overline{6}$$

$$7/15$$