|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PM592: Regression Analysis for Data Science** | | | Name: |  |  |
| **HW1** |  |  |  | Flemming Wu |  |
| *Distributions, EDA, Statistical Tests, Sampling Distribution* | | | | |  |

**Instructions**

* Answer questions directly within this document.
* Upload to Blackboard by the due date & time.
* Clearly indicate your answers to all questions.
* If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
* There are 4 questions and 30 points possible.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 1** |  |  |  |  | [6 points] |  |

Solve these problems using the probability functions in R. Create 8 separate variables that contains the answer to each, and store the variables in a tibble. Attach the code you used to create the tibble and the tibble output. Assume Z~N(0,1).

|  |  |
| --- | --- |
|  | 1a. [1 point]. P(Z ≤ ?) = 0.01 |

|  |  |
| --- | --- |
|  | 1b. [1 point] P(|Z| ≥ 1.96) |

|  |  |
| --- | --- |
|  | 1c. [1 point] P( ≤ ?) = 0.95 |

|  |  |
| --- | --- |
|  | 1d. [1 point] P(≤ 10) = ? |

|  |  |
| --- | --- |
|  | 1e. [1 point] P(|T16| ≤ ?) = 0.15 |

|  |  |
| --- | --- |
|  | 1f. [1 point] P(F7,30 ≤ 1.9) = ? |

> # 1a

> a <- qnorm(p=0.01, mean=0, sd=1)

>

> # 1b

> b <- 2\*pnorm(q=1.96, lower.tail=F)

>

> # 1c

> c <- qchisq(p=0.95, df=7)

>

> # 1d

> d <- pchisq(q=10, df=12)

>

> # 1e

> e <- abs(qt(p=.15/2, df=16))

>

> # 1f

> f <- pf(q=1.9, df1=7, df2=30)

>

> ans <- tibble(

+ "1a" = a, "1b" = b, "1c" = c, "1d" = d, "1e" = e, "1f" = f

+ )

> ans

# A tibble: 1 × 6

`1a` `1b` `1c` `1d` `1e` `1f`

*<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 -2.33 0.0500 14.1 0.384 1.51 0.895

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 2** |  |  |  |  | [9 points] |  |

Use the “wcgs.csv” data set.

* Load the data into an object called “wcgs\_raw”.
* Create a new data set called “wcgs” for the following modifications. When you are done with the problem, save the data as an R data set.

|  |  |
| --- | --- |
|  | 2a. [1 point] Explore this data set. What information d the “str”, “names”, and “dim” functions provide you? |

`str` provides the shape of the data frame, each column’s data type, and first few values of each column.

`names` provides a character vector containing the column names.

`dim` provides a vector of length 2, the first element being the number of rows, and the second being the number of columns.

|  |  |
| --- | --- |
|  | 2b. [1 point] Create an ordinal factor variable for weight category that consists of four categories: <140, 140-170, 170-200, >200. Find the frequency of each of these values in the sample. |

> wcgs <- wcgs\_raw %>%

+ mutate(weight\_cat = cut(weight, breaks=c(-Inf, 140, 170, 201, Inf), labels=c("<140", "140-170", "170-200", ">200"), include.lowest=F, right=F))

>

> wcgs %>%

+ count(weight\_cat) %>%

+ mutate(pct = n / sum(n))

# A tibble: 4 × 3

weight\_cat n pct

*<fct>* *<int>* *<dbl>*

1 <140 165 0.0523

2 140-170 1391 0.441

3 170-200 1385 0.439

4 >200 213 0.0675

|  |  |
| --- | --- |
|  | 2c. [1 point] Create an ordinal factor variable for age category that consists of five categories: 35-40, 41-45, 46-50, 51-55, 56-60. Find the frequency of each of these values in the sample. |

> wcgs <- wcgs %>%

+ mutate(age\_cat = cut(age, breaks=c(35, 41, 46, 51, 56, 60), labels=c("35-40", "41-45", "46-50", "51-55", "56-60"), right=F))

>

> wcgs %>%

+ count(weight\_cat) %>%

+ mutate(pct = n / sum(n))

# A tibble: 4 × 3

weight\_cat n pct

*<fct>* *<int>* *<dbl>*

1 <140 165 0.0523

2 140-170 1391 0.441

3 170-200 1385 0.439

4 >200 213 0.0675

|  |  |
| --- | --- |
|  | 2d. [1 point] Create a variable for BMI (you will have to look up the equation). In the WCGS data, height is measured in inches and weight is measured in pounds. Provide summary statistics for BMI using the package of your choice. |

|  |  |
| --- | --- |
|  | 2e. [1 point] Provide histograms for age, SBP (systolic blood pressure), DBP (diastolic blood pressure), weight, height, and cholesterol. Comment on whether these variables appear normally distributed. |

|  |  |
| --- | --- |
|  | 2f. [1 point] Create a variable that is the natural log of SBP. Provide summary statistics for this variable using the package of your choice. |

|  |  |
| --- | --- |
|  | 2g. [1 point] Create a scatterplot of SBP as a function of BMI. Label your axes appropriately. What are your impressions of the relationship between SBP and BMI from this figure? |

|  |  |
| --- | --- |
|  | 2h. [1 point] Calculate the total number of cigarettes that are smoked per day by all subjects in the data set (i.e., the sum of ncigs). |

|  |  |
| --- | --- |
|  | 2i. [1 point] Provide a cross-tabulation of personality type (dibpat) by smoking status. What percent of Type A personalities smoke? What percent of Type B personalities smoke? |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 3** |  |  |  |  | [7 points] |  |

Continue to use your WCGS file.

|  |  |
| --- | --- |
|  | 3a. [1 point] Consider a SBP of 125 as “normal”. Is there any evidence that the mean SBP of individuals in this sample is different from 125? |

|  |  |
| --- | --- |
|  | 3b. [1 point] Create boxplots of SBP by smoking status. What is your impression of how SBP relates to smoking status? |

|  |  |
| --- | --- |
|  | 3c. [1 point] Perform a parametric statistical test to determine if mean SBP differs by smoking status. |

|  |  |
| --- | --- |
|  | 3d. [1 point] Perform a non-parametric statistical test to determine if mean SBP differs by smoking status. |

|  |  |
| --- | --- |
|  | 3e. [3 points] Provide a conclusion about the hypothesis that there is a difference in mean SBP between smokers and nonsmokers. Provide a brief written summary of your methods and results. Make sure to justify your choice of statistical test. Include a small table that shows descriptive statistics (mean and 95% CI) for SBP overall and for SBP stratified by smokers and non-smokers. Include the p-value for comparing the smoking groups. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 4** |  |  |  |  | [8 points] |  |

In this problem we will simulate drawing samples from two different populations.

Use the “help” (?) function to determine how the “rnorm” function works.

|  |  |
| --- | --- |
|  | 4a. [2 points] Use the “rnorm” function to create a vector named “pop1” that contains of 30 samples from a population with . Create a vector named “pop2” that contains 30 samples from a population with . Would the results of a t-test lead us to conclude that these samples are from populations with different means? |

|  |  |
| --- | --- |
|  | 4b. [4 points] Re-run the program several (>6) times, each time altering the value of the means, standard deviations, OR sample sizes (only alter one parameter at a time). Fill the values you obtain into the table below. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Population 1 | | Population 2 | | Sample 1 | | | Sample 2 | | |  |  |
| Model |  |  |  |  | *N* |  | *SD* | *N* |  | *SD* | t | p |
| 1 | 100 | 20 | 105 | 20 | 30 |  |  | 30 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |
| --- | --- |
|  | 4c. [2 points] Summarize the effect of changing the population mean, standard deviation, and N has on the t and p-values. What general trends do you observe? |