**IHP 525 Milestone Two Template**

**READ ENTIRE TEMPLATE DIRECTIONS BEFORE FILLING IN TEMPLATE**

**Do NOT delete anything from this template.**

1. **Student Name: Irina Sillanpaa**

State the question you will pursue. (This should be copied and pasted from the list of questions and should be the same question submitted in week 2 unless you have changed your question.)

1. **Question of Interest (paste):** To what extent does **gender** influence **length of hospital stay** for MI patients?

**Directions for following page(s):**

* Fill out the table below for **EACH** variable of interest for **YOUR QUESTION**. (Your possible variables of interest are gender, length of stay, age, BMI, follow-up time, and survival.)
* Include **ONLY** the variables that are relevant to your question of interest. (**If it’s not mentioned in your question directly, it’s not relevant.**)
* Each variable should take up **ONE** row.
* Only use the rows you need; you should **NOT** fill every row of the table.
* You will have more than one variable; it will be quantitative or categorical. Choose an appropriate row for your variable type.
* Data set comes from the file uploaded to StatCrunch in Module One.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable type (categorical or quantitative)  Do **NOT** change anything in this column. | Variable name  (**one variable per row, don’t use all rows**)  Your choices for names are gender, length of stay, age, BMI, follow-up time, and survival. | Descriptive statistics  \*For THIS part of the final project, do not subdivide the data for any quantitative variable in each row based on any other variable.  **For example, do NOT find the mean length of stay separately for males and females for this part of the final project.** | | Key features  Creating a histogram or bar chart may help, but do NOT include either of these in this table. |
| Categorical | Variable name:  gender | Frequency of each value of the variable:  Male: 65  Female: 35  Relative Frequencies (percentages):  Male: 0.65  Female: 0.35 | Frequency Definition/formula:  Frequency means the number of times something occurs. A frequency is the number of times that a particular data value occurs in a data set. Frequency = count  Relative Frequency Definition/formula:  A relative frequency is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. Relative frequency = proportion/ *Relative Frequency = f / n*  f = the number of times the data occurred in an observation n = total frequencies | Is there an equal number of data points for each value of the variable (e.g., is there an equal number of males and females)?  No, there are more males than females |
| Quantitative | Variable name:  length of stay | Mean value: 6.84  Median value: 5  Range value: 55  Standard Deviation value: 5.9181624 | Mean Definition/formula:  Mean refers to the arithmetic average of a data set/  *n* = the sample size  *xi* = the value of the *ith* observation in the data set  *x* (“x-bar”) = is the sample mean  Median Definition/formula:  The median is the point that divides the data set into a top half and bottom half; it is halfway up (or down) the ordered list/ The**median formula** is *{(n + 1) / 2}th*,  *n* = the number of items in the set *th* just means the (n)th number  Range Definition/formula:  The range is the difference between the lowest and highest values/ *Range****= Maximum Value – Minimum Value***  Standard Deviation Definition/formula:  The standard deviation is a measure of the spread of scores within a set of data/  s = sample standard deviation Sum of = sum of... Sample mean = sample mean n = number of scores in sample.    σ = population standard deviation  ∑ = sum of…  X = each value  μ = population mean  N = number of values in the population | Is this variable symmetric or skewed?  Data are skewed to the right.  Describe any outliers:  There is an outlier of 56. |
| Categorical | Variable name: | Frequency of each value of the variable:  Relative Frequencies (percentages): | Frequency Definition/formula:  Relative Frequency Definition/formula: | Is there an equal number of data points for each value of the variable (e.g., is there an equal number of males and females)? |
| Quantitative | Variable name: | Mean value:  Median value:  Range value:  Standard Deviation value: | Mean Definition/formula:  Median Definition/formula:  Range Definition/formula:  Standard Deviation Definition/formula: | Is this variable symmetric or skewed? (Check histogram but do not include histogram.)  Describe any outliers: |
| Categorical | Variable name: | Frequency of each value of the variable:  Relative Frequencies (percentages): | Frequency Definition/formula:  Relative Frequency Definition/formula: | Is there an equal number of data points for each value of the variable (e.g., is there an equal number of males and females)? |
| Quantitative | Variable name: | Mean value:  Median value:  Range value:  Standard Deviation value: | Mean Definition/formula:  Median Definition/formula:  Range Definition/formula:  Standard Deviation Definition/formula: | Is this variable symmetric or skewed? (Check histogram but do not include histogram.)  Describe any outliers: |

DO NOT list or discuss descriptive statistics in this space. Use the table above, as directed.

DATA COLLECTION:

Examine the statistical report description (found in module 3 milestone rubric or announcements) to find the source of the data. The data we are using is a smaller subset of the original data. Do you think this smaller data set (100 observations) was randomly selected from the larger data set in the original study? Please give a reason for your answer.

Answer here: This statistical report will be created based on the data provided by Hosmer, Lemeshow, and May's data set (2008). This data set contains data that has been collected during thirteen 1-year periods. Other data are provided; nevertheless, this report would not benefit as it relates only to the Length of Hospital Stay and gender of MI patients. Since this data set has been collected over a long timeframe of thirteen years, it can provide rich data for the question to be addressed in the Statistical Report. This data set provides access to Hospital Length Stay for 100 patients and the gender of each patient. This information makes it possible to compare them statistically. The idea is that a greater sampling size would provide an accurate estimation of the parameter for the question. A clear answer should be given at the conclusion of this statistical report as to the effects of gender on the Length of Hospital Stay of MI patients.

Analyze the **limitations** of the data set you were provided and how those limitations might affect your findings.

* **Limit your response to the data relevant to your question of interest.**
  + For example, only using two variables in the data set is NOT a limitation of the data in your question of interest. It may be a limitation of the study or question of interest, but it is NOT a limitation of the data you have been provided for your question of interest. Don’t discuss this here.
  + **Comorbidities or missing patient information are NOT appropriate to discuss here.** Not including these elements might be a limitation of the study or question of interest, but it is NOT a limitation of the data you have been provided for your question of interest.
  + **Other treatment information is NOT appropriate to discuss here.** Not including these elements might be a limitation of the study or question of interest, but it is NOT a limitation of the data you have been provided for your question of interest.
* **Possible limitations:**
  + Consider any **outliers** and how they could affect the analysis (if there are any).
  + Consider the **scope** of data collection (time frame, location) and the question of interest. Is the sample adequate to answer the general question of interest?
* **ONE clear limitation is sufficient.**

Write limitation discussion here: