Homework for Chapter 3: Describing Variables

1. What is a variable?

A variable is the attribute of the analysis units that may take different possible values or categories over cases and time. In the context of empirical research, a variable is a set of observations of the same operationalized measurement.

1. For each of the following variables, what types of variables are they (continuous, count, ordinal, categorical, qualitative)?
   1. Age: continuous variable
   2. Gender: categorical variable
   3. The number of times that the President has tweeted in the past day: count variable
   4. Income: continuous variable
   5. Number of Instagram posts about statistics in the past month: count variable
   6. The number of unemployment claims filed in US last week: count variable
   7. The university or college that a student attends: qualitative variable

\*It is not a usual practice to list all academic affiliations while categorical variable should be as exhaustive as possible in that all the data must go into one of the categories, except for some special survey targeted at the population of students from certain colleges or universities, the university or college that a student attends can be a categorical variable. Actually, according to *Approaches to Social Research* (Singleton, A., Straits, & Bruce, C., 2017), a qualitative variable is also called categorical variable. They are often used interchangeably.

* 1. A therapist’s written assessment of a patient’s symptoms of depression: qualitative variable
  2. Whether a soccer team is in its league’s A division (highest), B division (next highest), or C division (lowest): ordinal variable

1. Which of the following provides a description of the probability that each possible value of a variable will occur? b
   1. Variation
   2. Distribution
   3. Range
   4. Mean
2. Graphical user interface, text, application

   Description automatically generatedBelow is a frequency table depicting the salaries of Economics professors employed at a university. The column named Salary contains the salary, and the column named Frequency contains the number of professors who earn the stated salary.

|  |  |
| --- | --- |
| Salary | Frequency |
| $85,000 | 5 |
| $90,000 | 4 |
| $100,000 | 1 |
| $120,000 | 2 |
| $125,000 | 3 |
| $130,000 | 2 |

* 1. Calculate the average salary earned by professors in the Economics department.

$103,529.4

* 1. Calculate the median. $ 90,000
  2. Calculate the minimum and maximum.

The minimum salary is $ 85,000. The maximum salary is $ 130,000.

* 1. Calculate the interquartile range. $ 40,000

1. Which of the following commonly represents the *truth* we are trying to estimate in statistics? c
   1. English/Latin letters like and
   2. Modifications of English letters like
   3. Greek letters like and
   4. Modifications of Greek letters like
2. Which of the following reduces impact of extreme values if the distribution of a variable is skewed? a
   1. Log transforming the variable
   2. Mean centering the variable
   3. Standardizing the variable
   4. Ignoring the variable
3. Order the following distributions from the one with lowest to the one with highest variability:  
   Shape, histogram

   Description automatically generatedd, a, c, b
4. The following graph represents the final exam scores for 1000 students who took an Introduction to Statistics course at a university.  
   Chart, line chart

   Description automatically generated
   1. Describe the distribution.

The distribution has three peaks, one roughly at 77.5, one roughly at 79, and another at roughly 82, around which the observations are concentrated. But the second one is larger than the first one, and the first one is larger than the third one (the most frequent score, the mode is around 79). The mean of the distribution is approximately between 80 and 82.5. The median is between the mode and the mean. Most of the distribution ranges from 75 to 90 (we cannot tell the lowest score or the highest score from the graph shown). The distribution is right skewed with few observations of small values under 75 and a few observations of big values over 85.

* 1. Is there skewness to the data?

Yes, the distribution has a right/positive skew — its right tail stretches more far out than its left tail does.

* 1. Would the mean or the median be a better measure to describe the center of the distribution?

With the skewed distribution, the median would be a better measure to describe the center of the distribution since unlike the mean which is more easily susceptible to the big observations. The median in this case can give a representative observation instead of a representative value to measure the central tendency of the data.

* 1. What measure would you use to describe the variability in the distribution?

Theoretically, we can use the smallest observation (minimum) and the largest observation (maximum) to approximate the range from the distribution. We can also calculate the standard deviation to quantifies the spread of the distribution by measuring how far the observations are from the mean. It gives the average/typical distance between a data point and the mean (subtracting the mean from each data point; squaring each of the results; adding up the results and dividing the sum by the total number of data points minus 1; taking the square root of the result). But interquartile range (IQR) is more appropriate to get the variability of the skewed distribution than the range or the standard deviation since it isn’t very strongly affected by big tail observations. The IQR (difference between the 75th percentile and the 25th percentile) gives us the range of the middle 50% of the data and a good sense of how close to the center the sample data are.

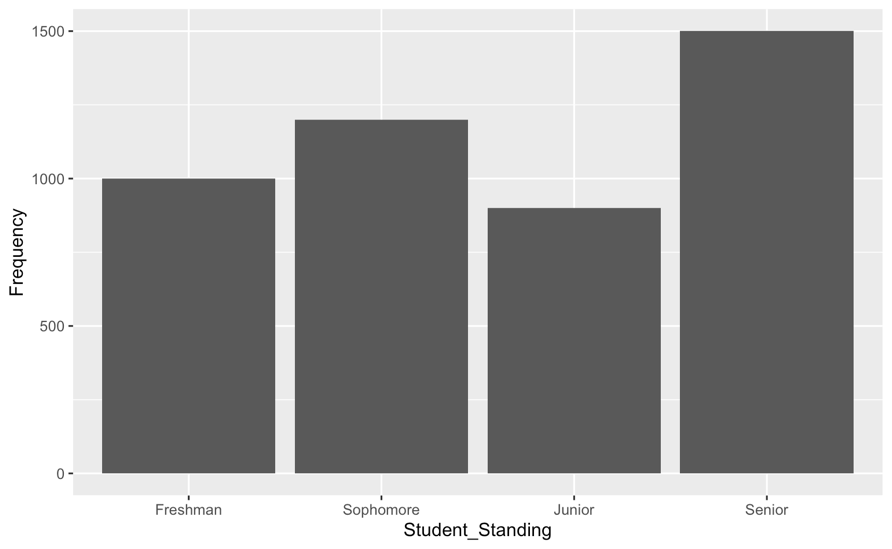
1. Which of the following statement is correct regarding a *left-skewed distribution*? b
   1. The mean is greater than the median
   2. The mean is less than the median
   3. The mean and median are equal
2. The table below shows data on how many students in a university are Freshman, Sophomore, Junior and Senior.

|  |  |
| --- | --- |
| Student Standing | Frequency |
| Freshman | 1000 |
| Sophomore | 1200 |
| Junior | 900 |
| Senior | 1500 |

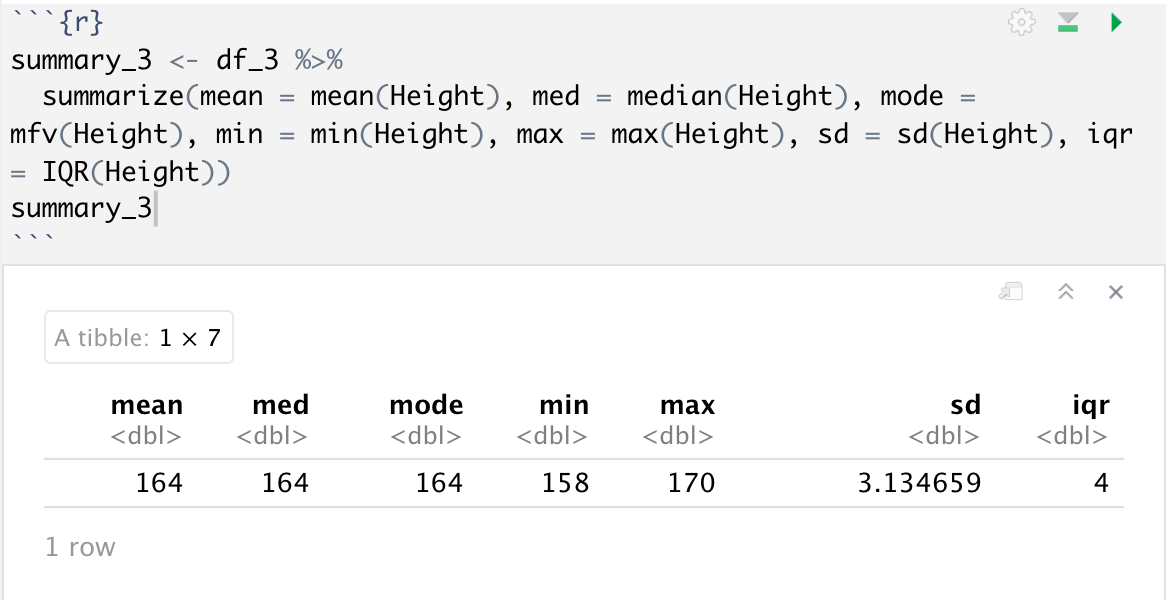
* 1. For an individual student, what kind of variable is their student standing (continuous, count, ordinal, categorical, qualitative)? Categorical variable (no comparison or rank order is necessarily made for an individual student).
  2. For the university itself, what kind of variable is the number of students with a given student standing (continuous, count, ordinal, categorical, qualitative)? The number of a specific group of students is a count variable.

\*I’m not quite sure about the intended meaning of the question here. If the question means, at the level of a university, what kind of variable is the student standing of the undergraduates, the answer is the ordinal variable since here the order matters—the values of the variable in question are ranked from lowest to highest (freshman to senior) or vice versa.

* 1. What kind of graph would best visualize this distribution for the university? Create the graph either by hand or with software.

A barplot would best visualize the distribution of the student standing.

1. Collect any continuous data from your classmates, family members, and/or friends. For example, you can collect data about their age or height.
   1. What is the distribution of the data? Describe the shape of the distribution. Summarize the data using measures like mean, median, range, standard deviation.

Chart, line chart

Description automatically generatedThe distribution of my friends’ heights in centimeters is a symmetric normal distribution. The distribution centers around the mean 164 and the median and mode are also 164. It ranges from 158 to 170 and has a standard deviation of 3.134659 as well as a IQR of 4.

* 1. What theoretical distribution may have generated the data that you sampled?  
     Normal distribution. The height of adult female is a continuous variable that theoretically can adopt any value (still with a most likely range). Women can be taller or shorter and there can be a typical height for women.
  2. Does your data give you a good idea about what the underlying theoretical distribution might be? Explain your reasoning.

The data indicates that the adult female heights might follow a normal distribution given that the distribution appears unimodal and symmetrical around its center, and the mean, median, and mode are all equal. However, my sample is a convenience sample, and the sample size is too small, so I cannot make a sound inference just based on the data I collected.