**Biostats: Quiz 1**

Benjamin Gazeau

3732892

1. Quantitative data: This is data that one consists of counts or can be counted and is non-words. The various types of quantitative data are numerical data, continuous data and dates.

**Numerical data**: Numerical data are quantitative in nature. They represent things that can be objectively counted, measured or calculated.

**Continuous data:** is generally rational, decimals or infinite values. This type of data is dependent on how accurately the values have been rounded and often have an upper and lower bound. Summaries that lend themselves to continuous data are frequency distributions, relative frequency distributions, cumulative frequency distributions, scatter plots, boxplots and bar graphs. In R, continuous data is denoted as <dbl>.

Dates are also a type of quantitative data that is generally separated by a dash and provides an understanding of time.

**Nominal data**: Integer data (discrete numbers or whole numbers), such as counts. For example, family A may have 3 children and family B may have 1 child, neither may have 2.3 children. Integer data usually answer the question, “how many?” In R integer data are called int or <int>.

**Qualitative data:** This data is often used descriptive words for columns or rankings. The various types of qualitative data are categorical data, ordinal data and binary data.

**Categorical data:** is a type of data thatdivides the dataset into a set of categories. These categories cannot be measured relative to one another as they do not show any ranking. Example of this is red beetles = 10 and green beetles = 2. Green beetles are not necessarily worse than red just because red is higher in value than green. Categorical data, in R, is denoted as factor or <fctr>. Summaries that lend themselves towards categorical data are, frequency distributions, relative frequency distributions, category statistics pie charts and so on. Ordinal data consists of word/categorical data but is listed in a particular ranking and this is generally read from lowest to highest values such as, good, better, best. Etc.  Ordinal data, in R, is denoted as Ord.factor.

Binary data show one of any two mutually exclusive outcomes such as, true or false, yes or no. In R, this type of data is referred to as “Logical” data and is either TRUE or FALSE, where a value of 1 is given to TRUE and 0 is given to FALSE.

**Binary data:** is denoted as logi or <lgl>, in R.

min()

max()

summary()

summarise()

dim()

str()

head()

tail()

glimpse()

Mean which is the average value of the dataset – mean()

Median which is the middle value of the dataset, specifically the 50% value of the dataset– median()

Skewness – skew()

Kurtosis -kur()

Skewness refers to the symmetry of the data. Data can either have positive skewness or negative skewness. Positively skewed data is data that is skewed to the right. This is when the mean is greater than the median. Whereas, negatively skewed data is skewed to the left and this is when the mean is smaller than the median.

Kurtosis refers to the shape of the tail of the data. Normal distribution has zero kurtosis (mesokurtic data). Negative kurtosis (platykurtic) is data that has a thin shaped tail and only covers a small area. Positive kurtosis (leptokurtic) is data that has a thick shaped tail and covers a larger area.

1. This is numerical data.
2. Mutate- a function that creates or transforms variables.

Select- a function used to highlight or isolate certain variables

Group by- a function used to group different variables together

Filter- used to isolate and remove certain variables

Separate- Separate a character column into multiple columns using a regular expression separator