# APS606 - Lecture Note (Week 1)

Norman Lo, M.S. in Quantitative Economics

9/30/2020

## Introduction to Quantitative Methods

I. What is quantitative research?
II. Quantitative Research vs. Qualitative Research
III. Introduction to Fundamental Statistics
IV. Descriptive Statistics

### I. What is Quantitative Research?

**Quantitative Research** is a systematic approach to collect data through sampling method like online polls, online surveys, Questionnaires etc. Quantitative Research is generally used in the field of community health, marketing, sociology, economics, psychology, demographics, gender studies, political science. The objective of Quantitative Research is to employ mathematical theories in relation to phenomena. Quantitative Research is a methods to measure variables, analyse them and report among the studied variables through a **numerical system**. The objective is to understand, analyse, describe and give a prediction of a product or service. Quantitative Research will predict about products and services , can make the people understand the dynamics of changes to do in their products or services respectively. The questions about the research study are defined and all its aspects are designed , so that the data collected is reliable and accurate. The quantitative Research uses different tools to gather numerical data that is in the form of statistics and numbers are arranged in non-textual manner like figures , charts and tables. The Importance of Quantitative Research is that it helps tremendous help in studying samples and populations. It discusses detailed relevant questions like, where is the data come from, gap in the data, how robust is it and what are the exclusions in the data research. it is vital to describe the process of selection and describes the methods and tools that are being used by the researcher to collect the data. The quantitative research identifies variables that are being measured , gives a detailed description about applicable methods that is used in obtaining relevant data, notes down important criteria about the fact that the data was already in existence or the researcher has collected of its own. The importance of quantitative research is that it helps to describe the process or method for both processing and analyzing the data in detail, specific tools used for studying the research objective etc. In a quantitative method , the findings of research is written in a precise form that is entirely objective. The non-textual elements like graphs, charts, tables are there to give the overall description of available results. It also clarifies important points , so that the readers can understand the data in proper manner, as it was intend by.

The goal in conducting quantitative research study is to determine the relationship between one thing [an independent variable] and another [a dependent or outcome variable] within a population. Quantitative research designs are either descriptive [subjects usually measured once] or experimental [subjects measured before and after a treatment]. A descriptive study establishes only associations between variables; an experimental study establishes causality.

Quantitative research deals in numbers, logic, and an objective stance. Quantitative research focuses on numeric and unchanging data and detailed, convergent reasoning rather than divergent reasoning [i.e., the generation of a variety of ideas about a research problem in a spontaneous, free-flowing manner].

Its main characteristics are:

1. The data is usually gathered using structured research instruments.

2. The results are based on larger sample sizes that are representative of the population.

3. The research study can usually be replicated or repeated, given its high reliability.

4. Researcher has a clearly defined research question to which objective answers are sought.

5. All aspects of the study are carefully designed before data is collected.
6. Data are in the form of numbers and statistics, often arranged in tables, charts, figures, or other non-textual forms.
   Project can be used to generalize concepts more widely, predict future results, or investigate causal relationships.

7. Researcher uses tools, such as questionnaires or computer software, to collect numerical data.

8. The overarching aim of a quantitative research study is to classify features, count them, and construct statistical models in an attempt to explain what is observed.

Things to keep in mind when reporting the results of a study using quantitative methods:

- Explain the data collected and their statistical treatment as well as all relevant results in relation to the research problem you are investigating. Interpretation of results is not appropriate in this section.

- Report unanticipated events that occurred during your data collection. Explain how the actual analysis differs from the planned analysis. Explain your handling of missing data and why any missing data does not undermine the validity of your analysis.

- Explain the techniques you used to "clean" your data set.

- Choose a minimally sufficient statistical procedure; provide a rationale for its use and a reference for it. Specify any computer programs used.

- Describe the assumptions for each procedure and the steps you took to ensure that they were not violated.

- When using inferential statistics, provide the descriptive statistics, confidence intervals, and sample sizes for each variable as well as the value of the test statistic, its direction, the degrees of freedom, and the significance level [report the actual p value].

- Avoid inferring causality, particularly in nonrandomized designs or without further experimentation.
  Use tables to provide exact values; use figures to convey global effects. Keep figures small in size; include graphic representations of confidence intervals whenever possible.

- Always tell the reader what to look for in tables and figures.

NOTE: When using pre-existing statistical data gathered and made available by anyone other than yourself [e.g., government agency], you still must report on the methods that were used to gather the data and

describe any missing data that exists and, if there is any, provide a clear explanation why the missing data does not undermine the validity of your final analysis.

## II. Quantitative Research vs. Qualitative Research

When collecting and analyzing data, **quantitative research** deals with numbers and statistics, while **qualitative research** deals with words and meanings. Both are important for gaining different kinds of knowledge.

Quantitative research:

Quantitative research is expressed in numbers and graphs. It is used to test or confirm theories and assumptions. This type of research can be used to establish generalizable facts about a topic.

Common quantitative methods include experiments, observations recorded as numbers, and surveys with closed-ended questions.

Qualitative research:

Qualitative research is expressed in words. It is used to understand concepts, thoughts or experiences. This type of research enables you to gather in-depth insights on topics that are not well understood.

Common qualitative methods include interviews with open-ended questions, observations described in words, and literature reviews that explore concepts and theories.

Difference between quantitative and qualitative research:

| Quantitative Research | Qualitative Research |
|---|---|
| Focuses on testing theories and hypothesis | Focuses on exploring ideas and formulating a theory or hypothesis |
| Analyzed through math and statistical analysis | Analyzed by summarizing, categorizing, and interpreting |
| Mainly expressed in numbers, graphs, and tables | Mainly expressed in words |
| Requires many respondents | Requires few respondents |
| Closed (multiple choice) questions | Open-ended questions |
| Key terms: testing, measurement, objectivity, replicability | Key terms: understanding, context, complexity, subjectivity |

Data Collection Methods:

Quantitative and qualitative data can be collected using various methods. It is important to use a data collection method that will help answer your research question(s). Many data collection methods can be either qualitative or quantitative. For example, in surveys, observations or case studies, your data can be represented as numbers (e.g. using rating scales or counting frequencies) or as words (e.g. with open-ended questions or descriptions of what you observe). However, some methods are more commonly used in one type or the other.

Quantitative Data Collection Methods:

**Surveys:** List of closed or multiple choice questions that is distributed to a sample (online, in person, or over the phone).
**Experiments:** Situation in which variables are controlled and manipulated to establish cause-and-effect relationships.
**Observations:** Observing subjects in a natural environment where variables can't be controlled.

Qualitative Data Collection Methods:

**Interviews:** Asking open-ended questions verbally to respondents.
**Focus groups:** Discussion among a group of people about a topic to gather opinions that can be used for

further research.
**Ethnography:** Participating in a community or organization for an extended period of time to closely observe culture and behavior.
**Literature review:** Survey of published works by other authors.

When to use qualitative vs. quantitative research?

A rule of thumb for deciding whether to use qualitative or quantitative data is:

Use quantitative research if you want to confirm or test something (a theory or hypothesis)

Use qualitative research if you want to understand something (concepts, thoughts, experiences)

For most research topics you can choose a qualitative, quantitative or mixed methods approach. Which type you choose depends on, among other things, whether you're taking an inductive vs. deductive research approach; your research question(s); whether you're doing experimental, correlational, or descriptive research; and practical considerations such as resources, time, availability of data, and access to respondents.

## III. Fundamental Statistics

**Statistics** is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data. Statistics is a highly interdisciplinary field; research in statistics finds applicability in virtually all scientific fields and research questions in the various scientific fields motivate the development of new statistical methods and theory. In developing methods and studying the theory that underlies the methods statisticians draw on a variety of mathematical and computational tools.

When analysing data, such as the marks achieved by 100 students for a piece of coursework, it is possible to use both **descriptive** and **inferential** statistics in your analysis of their marks. Typically, in most research conducted on groups of people, you will use both descriptive and inferential statistics to analyse your results and draw conclusions. So what are descriptive and inferential statistics? And what are their differences?

### Descriptive Statistics:

**Descriptive statistics** is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analysed or reach conclusions regarding any hypotheses we might have made. They are simply a way to describe our data.

Descriptive statistics are very important because if we simply presented our raw data it would be hard to visualize what the data was showing, especially if there was a lot of it. Descriptive statistics therefore enables us to present the data in a more meaningful way, which allows simpler interpretation of the data. For example, if we had the results of 100 pieces of students' coursework, we may be interested in the overall performance of those students. We would also be interested in the distribution or spread of the marks. Descriptive statistics allow us to do this. How to properly describe data through statistics and graphs is an important topic and discussed in other Laerd Statistics guides. Typically, there are two general types of statistic that are used to describe data:

**Measures of central tendency:** these are ways of describing the central position of a frequency distribution for a group of data. In this case, the frequency distribution is simply the distribution and pattern of marks scored by the 100 students from the lowest to the highest. We can describe this central position using a number of statistics, including the mode, median, and mean. You can learn more in our guide: Measures of Central Tendency.

**Measures of spread:** these are ways of summarizing a group of data by describing how spread out the scores are. For example, the mean score of our 100 students may be 65 out of 100. However, not all students will have scored 65 marks. Rather, their scores will be spread out. Some will be lower and others higher. Measures of spread help us to summarize how spread out these scores are. To describe this spread, a number

of statistics are available to us, including the range, quartiles, absolute deviation, variance and standard deviation.

When we use descriptive statistics it is useful to summarize our group of data using a combination of tabulated description (i.e., tables), graphical description (i.e., graphs and charts) and statistical commentary (i.e., a discussion of the results).

**Inferential Statistics:**

We have seen that descriptive statistics provide information about our immediate group of data. For example, we could calculate the mean and standard deviation of the exam marks for the 100 students and this could provide valuable information about this group of 100 students. Any group of data like this, which includes all the data you are interested in, is called a **population**. A population can be small or large, as long as it includes all the data you are interested in. For example, if you were only interested in the exam marks of 100 students, the 100 students would represent your population. Descriptive statistics are applied to populations, and the properties of populations, like the mean or standard deviation, are called **parameters** as they represent the whole population (i.e., everybody you are interested in).

Often, however, you do not have access to the whole population you are interested in investigating, but only a limited number of data instead. For example, you might be interested in the exam marks of all students in the UK. It is not feasible to measure all exam marks of all students in the whole of the UK so you have to measure a smaller **sample** of students (e.g., 100 students), which are used to represent the larger population of all UK students. Properties of samples, such as the mean or standard deviation, are not called parameters, but **statistics**. Inferential statistics are techniques that allow us to use these samples to make generalizations about the populations from which the samples were drawn. It is, therefore, important that the sample accurately represents the population. The process of achieving this is called **sampling**. Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected to perfectly represent the population. The methods of inferential statistics are (1) the estimation of parameter(s) and (2) testing of statistical hypotheses.

**Types of Data:**

When working with statistics, it's important to recognize the different types of data. Data are the actual pieces of information that you collect through your study. For example, if you ask five of your friends how many pets they own, they might give you the following data: 0, 2, 1, 4, 18. (The fifth friend might count each of her aquarium fish as a separate pet.) Not all data are numbers; let's say you also record the gender of each of your friends, getting the following data: male, male, female, male, female. Most data fall into one of two groups: **Numerical** or **Categorical**.

Numerical data have meaning as a measurement, such as a person's height, weight, IQ, or blood pressure; or they're a count, such as the number of stock shares a person owns, how many teeth a dog has, or how many pages you can read of your favorite book before you fall asleep. (Statisticians also call numerical data quantitative data.)

Numerical data can be further broken into two types: **discrete** and **continuous**.

**Discrete data** represent items that can be counted; they take on possible values that can be listed out. The list of possible values may be fixed (also called finite); or it may go from 0, 1, 2, on to infinity (making it countably infinite). For example, the number of heads in 100 coin flips takes on values from 0 through 100 (finite case), but the number of flips needed to get 100 heads takes on values from 100 (the fastest scenario) on up to infinity (if you never get to that 100th heads). Its possible values are listed as 100, 101, 102, 103, . . . (representing the countably infinite case).

**Continuous data** represent measurements; their possible values cannot be counted and can only be described using intervals on the real number line. For example, the exact amount of gas purchased at the pump for cars with 20-gallon tanks would be continuous data from 0 gallons to 20 gallons, represented by

the interval [0, 20], inclusive. You might pump 8.40 gallons, or 8.41, or 8.414863 gallons, or any possible number from 0 to 20. In this way, continuous data can be thought of as being uncountably infinite. For ease of recordkeeping, statisticians usually pick some point in the number to round off. Another example would be that the lifetime of a C battery can be anywhere from 0 hours to an infinite number of hours (if it lasts forever), technically, with all possible values in between. Granted, you don't expect a battery to last more than a few hundred hours, but no one can put a cap on how long it can go (remember the Energizer Bunny?).

**Categorical data** represent characteristics such as a person's gender, marital status, hometown, or the types of movies they like. Categorical data can take on numerical values (such as "1" indicating male and "2" indicating female), but those numbers don't have mathematical meaning. You couldn't add them together, for example. (Other names for categorical data are qualitative data, or Yes/No data.)

**Measurement of Data:**

Normally, when one hears the term measurement, they may think in terms of measuring the length of something (ie. the length of a piece of wood) or measuring a quantity of something (ie. a cup of flour). This represents a limited use of the term measurement. In statistics, the term measurement is used more broadly and is more appropriately termed scales of measurement. Scales of measurement refer to ways in which variables/numbers are defined and categorized. Each scale of measurement has certain properties which in turn determines the appropriateness for use of certain statistical analyses. The four scales of measurement are nominal, ordinal, interval, and ratio.

**Nominal:** Categorical data and numbers that are simply used as identifiers or names represent a nominal scale of measurement. Numbers on the back of a baseball jersey (St. Louis Cardinals 1 = Ozzie Smith) and your social security number are examples of nominal data. If I conduct a study and I'm including gender as a variable, I will code Female as 1 and Male as 2 or visa versa when I enter my data into the computer. Thus, I am using the numbers 1 and 2 to represent categories of data.

**Ordinal:** An ordinal scale of measurement represents an ordered series of relationships or rank order. Individuals competing in a contest may be fortunate to achieve first, second, or third place. First, second, and third place represent ordinal data. If Roscoe takes first and Wilbur takes second, we do not know if the competition was close; we only know that Roscoe outperformed Wilbur. Likert-type scales (such as "On a scale of 1 to 10 with one being no pain and ten being high pain, how much pain are you in today?") also represent ordinal data. Fundamentally, these scales do not represent a measurable quantity. An individual may respond 8 to this question and be in less pain than someone else who responded 5. A person may not be in half as much pain if they responded 4 than if they responded 8. All we know from this data is that an individual who responds 6 is in less pain than if they responded 8 and in more pain than if they responded 4. Therefore, Likert-type scales only represent a rank ordering.

**Interval:** A scale which represents quantity and has equal units but for which zero represents simply an additional point of measurement is an interval scale. The Fahrenheit scale is a clear example of the interval scale of measurement. Thus, 60 degree Fahrenheit or -10 degrees Fahrenheit are interval data. Measurement of Sea Level is another example of an interval scale. With each of these scales there is direct, measurable quantity with equality of units. In addition, zero does not represent the absolute lowest value. Rather, it is point on the scale with numbers both above and below it (for example, -10 degrees Fahrenheit).

**Ratio:** The ratio scale of measurement is similar to the interval scale in that it also represents quantity and has equality of units. However, this scale also has an absolute zero (no numbers exist below the zero). Very often, physical measures will represent ratio data (for example, height and weight). If one is measuring the length of a piece of wood in centimeters, there is quantity, equal units, and that measure can not go below zero centimeters. A negative length is not possible.

The table below will help clarify the fundamental differences between the four scales of measurement,

| | Indications Difference | Indicates Direction of Difference | Indicates Amount of Difference | Absolute Zero |
|---|---|---|---|---|
| Nominal | Yes | No | No | No |
| Ordinal | Yes | Yes | No | No |
| Interval | Yes | Yes | Yes | No |
| Ratio | Yes | Yes | Yes | Yes |

You will notice in the above table that only the ratio scale meets the criteria for all four properties of scales of measurement.

Interval and Ratio data are sometimes referred to as parametric and Nominal and Ordinal data are referred to as nonparametric. Parametric means that it meets certain requirements with respect to parameters of the population (for example, the data will be normal - the distribution parallels the normal or bell curve). In addition, it means that numbers can be added, subtracted, multiplied, and divided. Parametric data are analyzed using statistical techniques identified as Parametric Statistics. As a rule, there are more statistical technique options for the analysis of parametric data and parametric statistics are considered more powerful than nonparametric statistics.

Nonparametric data are lacking those same parameters and can not be added, subtracted, multiplied, and divided. For example, it doesnot make sense to add Social Security numbers to get a third person. Nonparametric data are analyzed by using Nonparametric Statistics. As a rule, ordinal data is considered nonparametric and can not be added, etc.. Again, it does not make sense to add together first and second place in a race - one does not get third place. However, many assessment devices within the behavioral and social sciences (for example, intelligence scales) as well as Likert-type scales represent ordinal data but are often treated as if they are interval data. For example, the "average" amount of pain that a person reports on a Likert-type scale over the course of a day would be computed by adding the reported pain levels taken over the course of the day and dividing by the number of times the question was answered.

As stated above, many measures (ie. personality, intelligence, psycho-social, etc.) within the behavioral and social sciences represent ordinal data. IQ scores may be computed for a group of individuals. They will represent differences between individuals and the direction of those differences but they lack the property of indicating the amount of the differences. Psychologists have no way of truly measuring and quantifying intelligence. An individual with an IQ of 70 does not have exactly half of the intelligence of an individual with an IQ of 140. Therefore, IQ scales should theoretically be treated as ordinal data.

In both of the above illustrations, the statement is make that they should be theoretically treated as ordinal data. In practice, however, they are usually treated as if they represent parametric (interval or ratio) data. This opens up the possibility for use of parametric statistical techniques with these data and the benefits associated with the use of techniques.

## IV. Descriptive Statistics

Descriptive statistics allow you to characterize your data based on its properties. Generally speaking, there are four major types of descriptive statistics:

1. Measures of Frequency

- Frequency, Related Frequency, and Percent Frequency
- Shows how often an outcome occurs
- Use this when you want to show how often a response is given

2. Measures of Central Tendency

- Mean, Median, and Mode
- Locates the distribution by various points
- Use this when you want to show how an average or most commonly indicated response

3. Measures of Variability

- Range, Variance, Standard Deviation
- Identifies the spread of scores by stating intervals
- Range = Minimum and Maximum points
- Variance or Standard Deviation = difference between observed score and mean
- Use this when you want to show how "spread out" the data are. It is helpful to know when your data are so spread out that it affects the mean

4. Measures of Position

- Percentile Ranks, Quartile Ranks
- Describes how scores fall in relation to one another. Relies on standardized scores
- Use this when you need to compare scores to a normalized score (e.g. a national norm)

**Practice with R**

In this section, we will practice using the software R (one of the most popular data science softwares for both academic and private research) for descriptive statistics. Base R provides several useful dataset for users to practice coding. We are using a dataset name "USArrests" to practice the tools in R and find some interesting insights from the data.

Data Set: **Violent Crime Rates by US State**

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Four variables are included in this dataset.

1. Murder: numeric Murder arrests (per 100,000)
2. Assault: numeric Assault arrests (per 100,000)
3. UrbanPop: numeric Percent urban population
4. Rape: numeric Rape arrests (per 100,000)
5. HighMurder: murder ratio greater than 8 is "High", otherwise is "Low". (Added Variable)
6. HighMurder: assault ratio greter than 150 is "1", otherwise is "0". (Added Variable)

```
## [1] "Murder"      "Assault"     "UrbanPop"    "Rape"        "HighMurder"
## [6] "HighAssault"
```

```
##            Murder Assault UrbanPop Rape Murder Assault
## Alabama      13.2     236       58 21.2   High       1
## Alaska       10.0     263       48 44.5   High       1
## Arizona       8.1     294       80 31.0   High       1
## Arkansas      8.8     190       50 19.5   High       1
## California    9.0     276       91 40.6   High       1
## Colorado      7.9     204       78 38.7    Low       1
```

**Tabular Descriptive Statistics:**

Frequency Tables:

```
# Summary statistics of the dataframe
summary(data)
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
##  HighMurder.Murder HighAssault.Assault
##  High:22           0:22
##  Low :28           1:28
##
##
##
##
```

```
# Frequency Tables for Discrete or Categorical Variable
table(data$HighMurder)
```

```
##
## High  Low
##   22   28
```

```
table(data$HighAssault)
```

```
##
##  0  1
## 22 28
```

```
# Frequency Tables for Continuous Variable
range(data$Murder)
```

```
## [1]  0.8 17.4
```

```
breaks <- seq(0, 18, by = 3)
murderCut <- cut(data$Murder, breaks, right = FALSE)
table(murderCut)
```

```
## murderCut
##   [0,3)   [3,6)   [6,9)  [9,12) [12,15) [15,18)
##       8      11      12       8       7       4
```

```r
# Relative Frequency Tables
table(data$HighMurder) / nrow(data)
```

```
##
## High  Low
## 0.44 0.56
```

```r
table(data$HighAssault) / nrow(data)
```

```
##
##    0    1
## 0.44 0.56
```

```r
table(murderCut) / nrow(data)
```

```
## murderCut
##   [0,3)   [3,6)   [6,9)  [9,12) [12,15) [15,18)
##    0.16    0.22    0.24    0.16    0.14    0.08
```

Measure of Central Tendency:

```r
# Calculate the average murder ratio
mean(data$Murder)
```

```
## [1] 7.788
```

```r
# Calculate the median murder ratio
median(data$Murder)
```

```
## [1] 7.25
```

```r
# Find the mode of the murder ratio
library(lsr)
modeOf(data$Murder)
```

```
## [1] 13.2  9.0 15.4  2.6  2.2  6.0  2.1
```

```r
# Create an aggregate table to identify the means of murder ratio by High / Low assualt category
aggregate(data$Murder ~ data$HighAssault, FUN = mean)
```

```
##   Assault data$Murder
## 1       0    4.263636
## 2       1   10.557143
```

```r
# Create an aggregate table to identify the means of assault ratio by High / Low murder category
aggregate(data$Assault ~ data$HighMurder, FUN = mean)
```

```
##   Murder data$Assault
## 1   High     243.4545
## 2    Low     113.6429
```

Measure of Variability:

```r
# Find the range of assault ratio
range(data$Assault)
```

```
## [1]  45 337
```

```r
# Calculate the population variance of assault ratio
m <- mean(data$Assault)
v <- sum((data$Assault - m)^2) / nrow(data)

# Calculate the population standard deviation of assault ratio
s <- sqrt(v)

# Calculate the sample variance of assault ratio
vA <- var(data$Assault)

# Calculate the sample standard deviation of assault ratio
sD <- sd(data$Assault)

# Create an aggregate table to identify the variance of assault ratio by High / Low murder category
aggregate(data$Assault ~ data$HighMurder, FUN = var)
```

```
##   Murder data$Assault
## 1   High     3180.450
## 2    Low     2441.423
```

```r
# Create an aggregate table to identify the standard deviation of assault ratio by High / Low murder ca
aggregate(data$Assault ~ data$HighMurder, FUN = sd)
```

```
##   Murder data$Assault
## 1   High     56.39548
## 2    Low     49.41076
```

Measure of Position:

```r
# Identify the quartiles of murder ratio
quantile(data$Murder)
```

```
##     0%    25%    50%    75%   100%
##  0.800  4.075  7.250 11.250 17.400
```

```r
# Find the 85th percentile of murder ratio
quantile(data$Murder, 0.85)
```

```
##    85%
## 12.895
```

```r
# Find the 95th percentile of assault ratio
quantile(data$Assault, 0.95)
```

```
##   95%
## 297.3
```

```r
# Find the 30th, 45th, and 65th percentile of assault ratio
quantile(data$Assault, c(0.3, 0.45, 0.65))
```

```
##    30%    45%    65%
## 112.10 151.25 203.55
```

Cross Tabulation: (Multi-Variable Description)

Usually applies to two categorical or discrete variables.

```r
# Contingency Table for two categorical varaibles
t <- table(data$HighMurder, data$HighAssault)
t
```

```
##
##         0  1
##   High  1 21
##   Low  21  7
```

```r
# Margin Table of the contingency table
margin.table(t, 1) # HighMurder frequencies summed over HighAssault
```

```
##
## High  Low
##   22   28
```

```r
margin.table(t, 2) # HighAssault frequencies summed over HighMurder
```

```
##
##  0  1
## 22 28
```

```r
# Proportion Table of the contingency table
prop.table(t) # cell percentages
```

```
##
##           0    1
##   High 0.02 0.42
##   Low  0.42 0.14
```

```r
prop.table(t, 1) # row percentages
```

```
##
##                  0          1
##   High 0.04545455 0.95454545
##   Low  0.75000000 0.25000000
```

```r
prop.table(t, 2) # column precentages
```

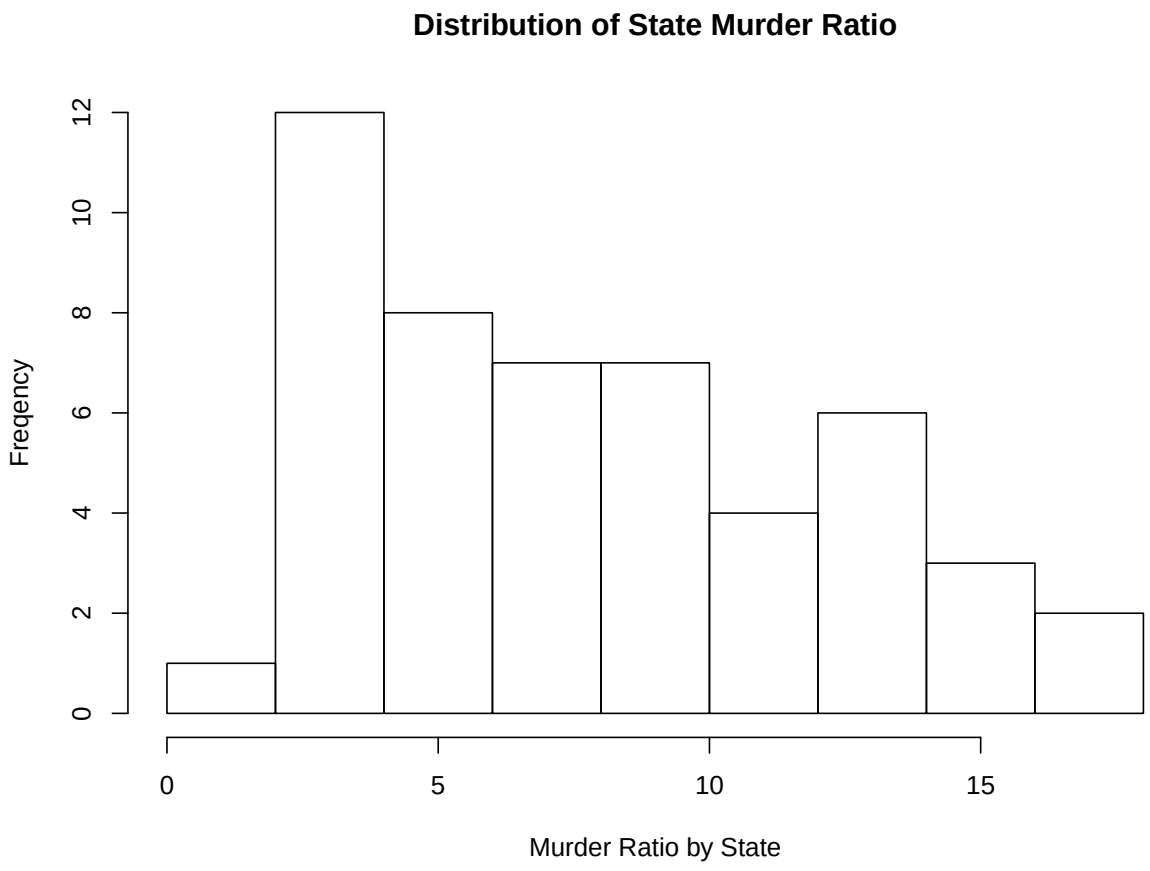
```
##
##                 0          1
##   High 0.04545455 0.75000000
##   Low  0.95454545 0.25000000
```

### Data Visualization

Other than tabulation, most researchers includes data visualization in their report or paper to engage audiences interest. Data visualization is a great tool for story telling even for a complex academic research study. Here we are going to cover a few popular visualization tools in R, which generally applies to most scenario in academic research study.

Distribution for Single Continuous Varialbe: (Histogram)

```r
# Plot a histogram for murder ratio
hist(data$Murder, main = "Distribution of State Murder Ratio",
     xlab = "Murder Ratio by State",
     ylab = "Freqency")
```

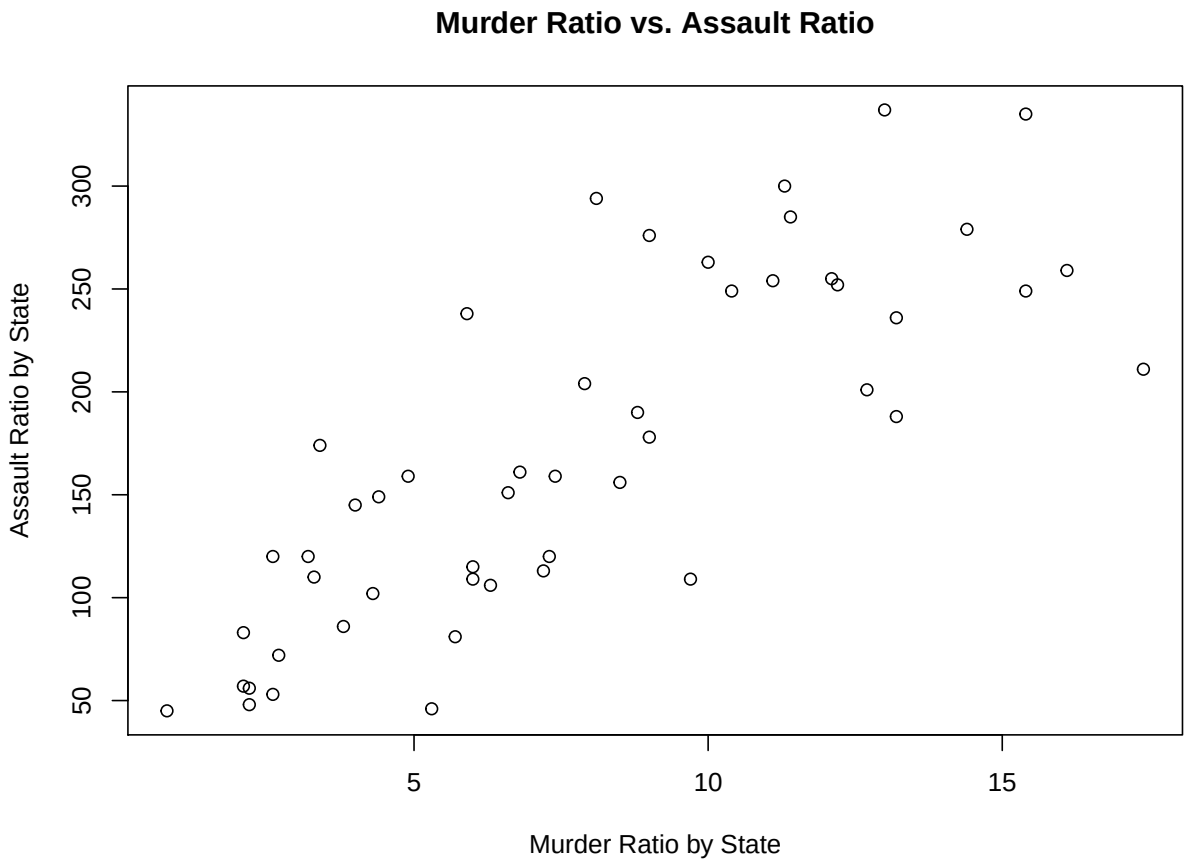



Distribution for Single Discrete / Categorical Variable: (Bar Chart)

```r
# Plot a bar chart for high and low murder states
barplot(table(data$HighMurder),
        main = "Distribution of High / Low Murder",
        xlab = "High / Low Murder Ratio by Category",
        ylab = "Frequency")
```

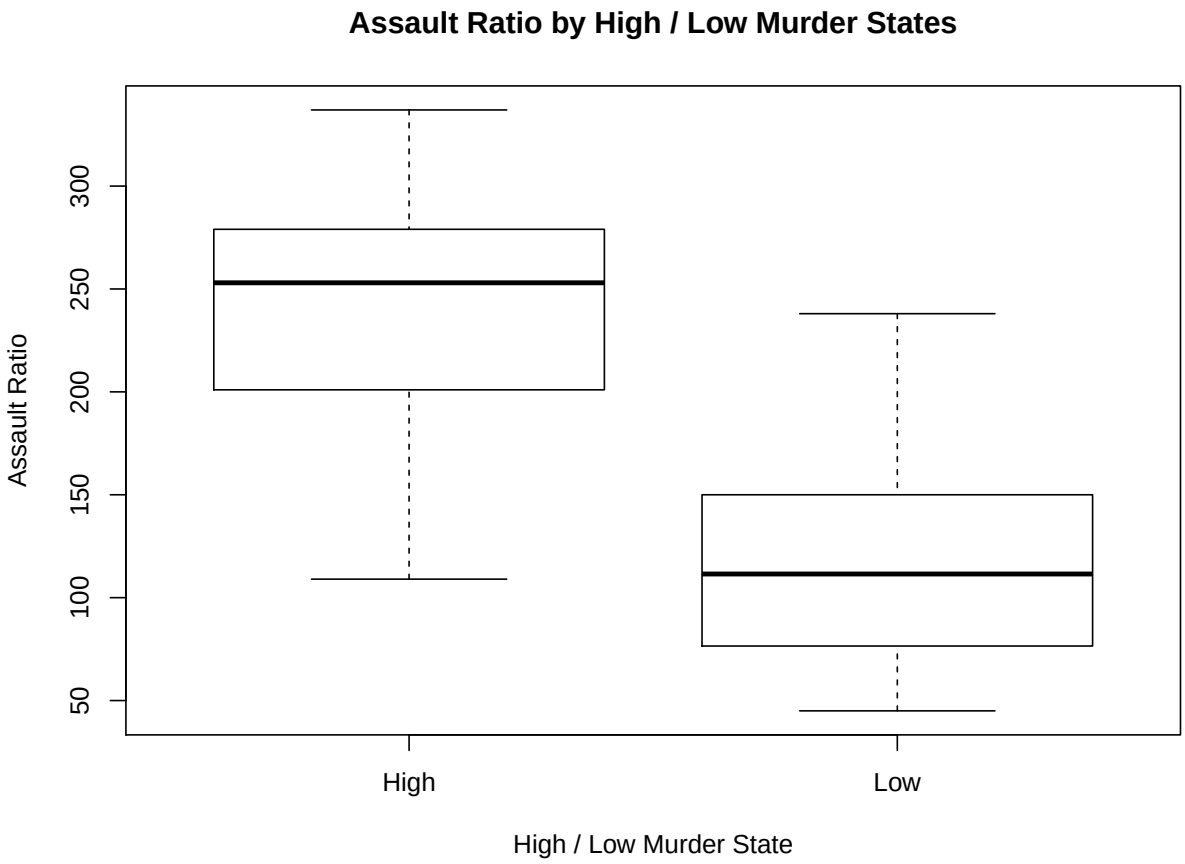

Correlation Between Two Continuous Variables: (Scatter Plot)

```r
# Plot a scatter plot displays correlation between murder ratio and assault ratio
plot(data$Murder, data$Assault, main = "Murder Ratio vs. Assault Ratio",
     xlab = "Murder Ratio by State", ylab = "Assault Ratio by State")
```




Correlation Between a Continuous and a Discrete / Categorical Variable: (Box Plot)

```
# Plot the correlation between High Murder State and Assault Ratio
plot(factor(data$HighMurder), data$Assault,
     main = "Assault Ratio by High / Low Murder States",
     xlab = "High / Low Murder State",
     ylab = "Assault Ratio")
```


Assault Ratio by High / Low Murder States
Assault Ratio
50
100
150
200
250
300
High
Low
High / Low Murder State