Basic Probability and Statistics

1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries

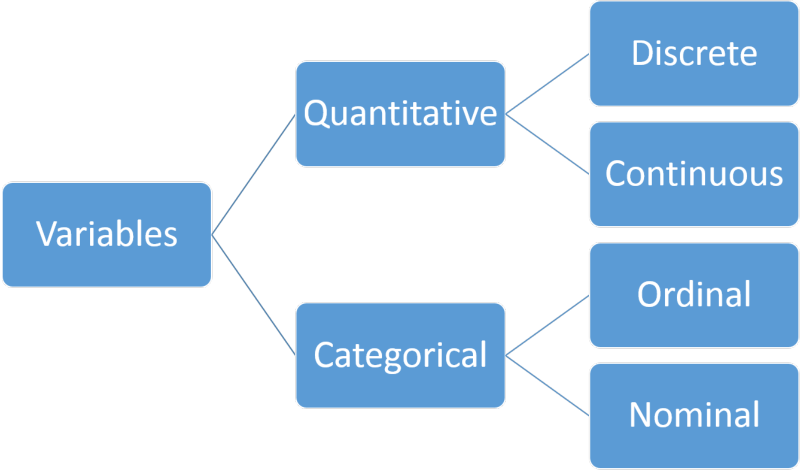
3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

1. Introduction

2. Single Quantitative Variable Exploration
    - Numerical Summaries
    - Graphical Summaries

3. Association Between Two Variables
    - Two Quantitative Variables
    - One Categorical and One Quantitative Variable
    - Two Categorical Variables

Types of Data



Descriptive Statistics

- There are two major ways of describing data descriptively: numerical and graphical summaries.

- One variable: the numerical and graphical summaries will be covered.

- For two variables: association between two variables will be covered.

1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

Numerical and Graphical Summaries

- **Numerical summaries**/descriptive measures: number of observations (sample size), location, variability and other measures.

- **Graphical summaries**: histogram, boxplot, QQ plot (for checking normality of a dataset), scatter plot for bivariate data.

1. Introduction

2. Single Quantitative Variable Exploration
    - Numerical Summaries
    - Graphical Summaries

3. Association Between Two Variables
    - Two Quantitative Variables
    - One Categorical and One Quantitative Variable
    - Two Categorical Variables

An Example: Yearly Sales

-
  ```
  > sales <- read.csv("C:/Data/yearly_sales.csv")
  ```

- The function head() displays the first few records in the data set

  ```
  > head(sales)
    cust_id sales_total num_of_orders gender
  1  100001      800.64             3      F
  2  100002      217.53             3      F
  3  100003       74.58             2      M
  4  100004      498.60             3      M
  5  100005      723.11             4      F
  6  100006       69.43             2      F
  > total = sales$sales_total
  ```

Summary of the Center

- Center of data should include the information on: mean, median and mode.
- About the total sales, we roughly can have

```
> n = length(total); n
[1] 10000
> summary(total)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.02   80.29  151.65  249.46  295.50 7606.09
```

Summary of the Variability

```
> range(total)
[1]   30.02 7606.09

> var(total)
[1] 101793.4

> sd(total)
[1] 319.0508

> IQR(total)
[1] 215.21
```

A Note on Numerical Summaries

- For a sample, if the mean is the same or approximately the same as the median, then the sample is close to symmetric.

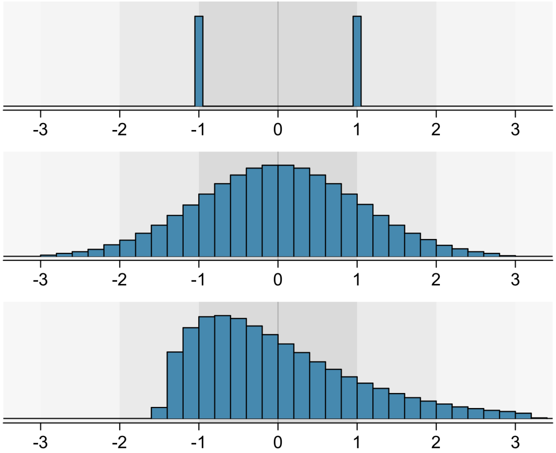
- Mean is sensitive to the outlier(s) while median is not.

- When the mean is much larger than the median, sample is right skewed; while when the mean is much smaller than the median then sample is left skewed.

1. Introduction
2. Single Quantitative Variable Exploration
    - Numerical Summaries
    - Graphical Summaries
3. Association Between Two Variables
    - Two Quantitative Variables
    - One Categorical and One Quantitative Variable
    - Two Categorical Variables

Numerical Summaries Are Not Enough

- All 3 samples below had a sample mean of 0 and a sample variance of 1.
- No matter how many of the summary measures we report, nothing beats a picture.



Histogram and Density Plot

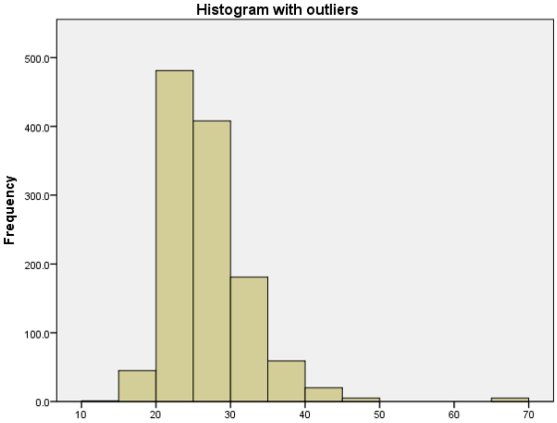
- A histogram is a graph that uses bars to portray the frequencies or relative frequencies of the possible outcomes for a quantitative variable.

- Density plots can be thought of as plots of smoothed histograms.

Histogram

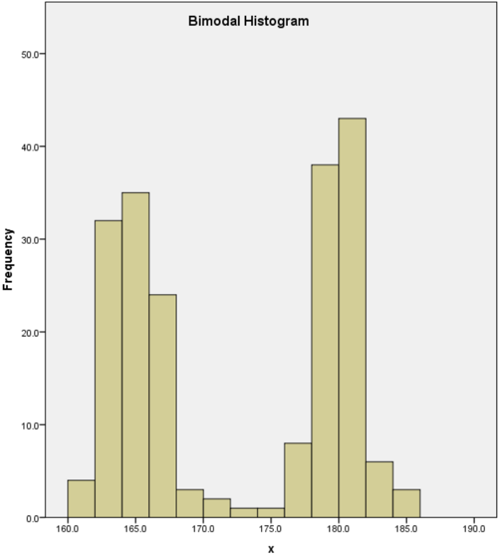
- What do we look for in a histogram?
    - The overall pattern. Do the data cluster together, or is there a gap such that one or more observations deviate from the rest?
    - Do the data have a single mound? This is known as a unimodal distribution. Data with two mound are known as bimodal, and data with many mounds are referred to as multimodal.
    - Is the distribution symmetric or skewed? Any suspected outliers?

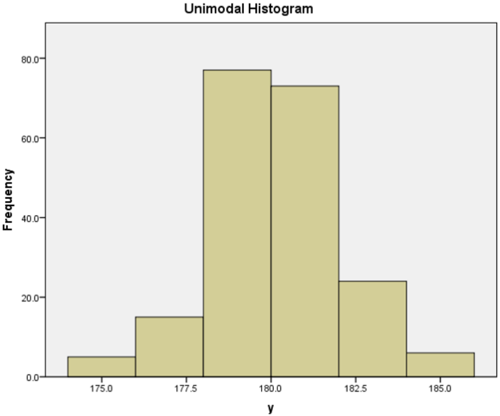
A Histogram With Suspected Outliers



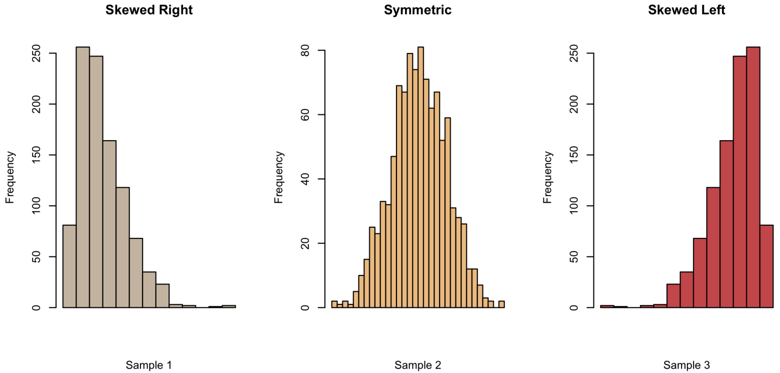
- This histogram is unimodal, but it has suspected outliers on the right.

Unimodal and Bimodal Histograms





Skewness of Histograms

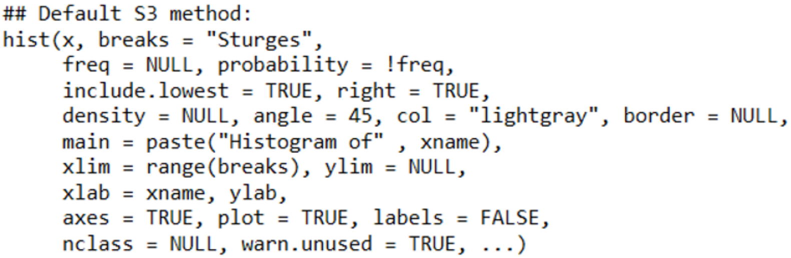


- Income is typically right-skewed.
- IQ is typically symmetric.
- Life-span is typically left-skewed.

Histogram and Density Plot in R

There are many ways to plot histograms in R:

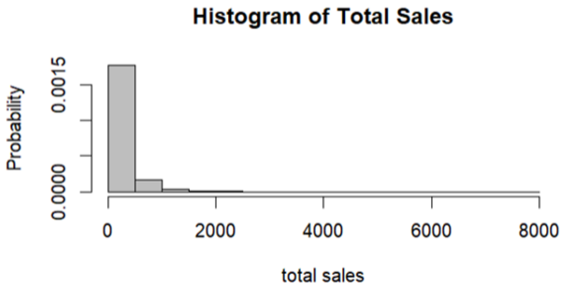
- The hist function in the base graphics package;
- truehist in package MASS;
- histogram in package lattice;
- geom_histogram in package ggplot2.



```
## Default S3 method:
hist(x, breaks = "Sturges",
     freq = NULL, probability = !freq,
     include.lowest = TRUE, right = TRUE,
     density = NULL, angle = 45, col = "lightgray", border = NULL,
     main = paste("Histogram of" , xname),
     xlim = range(breaks), ylim = NULL,
     xlab = xname, ylab,
     axes = TRUE, plot = TRUE, labels = FALSE,
     nclass = NULL, warn.unused = TRUE, ...)
```

Histogram and Normal Density Plot in R

```
> hist(total, freq=FALSE, main = paste("Histogram of Total Sales"),
+       xlab = "total sales", ylab="Probability", col = "grey")
```



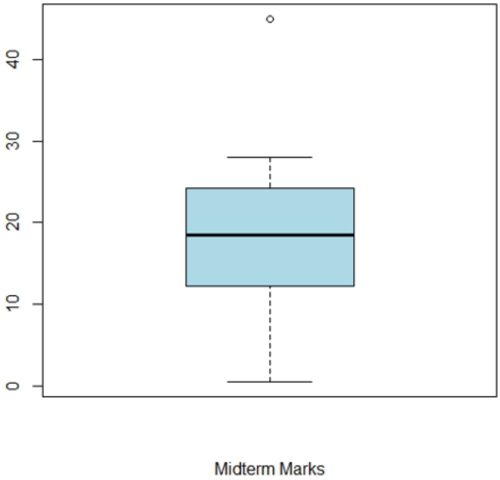
The histogram is highly right skewed.

Boxplots

- Boxplots provide a skeletal representation of a distribution, and they are very well suited for showing distributions for multiple variables.

- A boxplot helps us to identify median, lower and upper quantiles, IQR, and outlier(s).

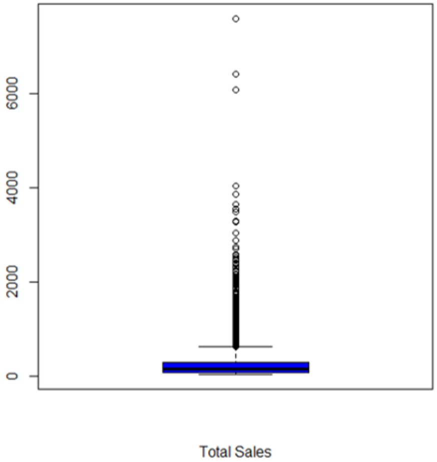
Boxplot



Boxplots in R

The code should be

```
> boxplot(total, xlab = "Total Sales", col = "blue")
```



- The median is very low, close to 200. Box plot shows many outliers and extreme outliers.
- If the sample is unimodal then the distribution is highly right skewed.

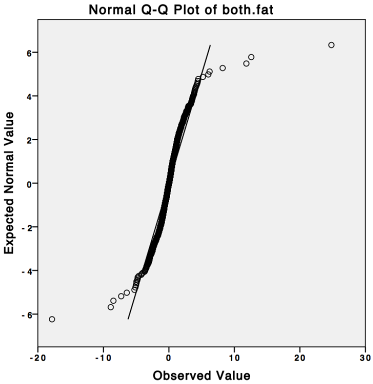
QQ Plots

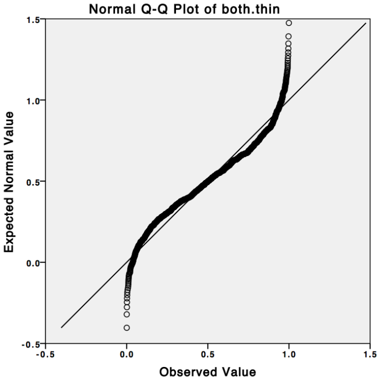
- **The purpose of plotting a QQ plot of a sample is to see if the sample follows (approximately) a normal distribution or not.**

- A QQ-plot matches the standardized sample quantiles against the theoretical quantiles of a N(0, 1) distribution.

- From the points on the plot, we can usually tell whether our sample has longer or shorter tail than normal.

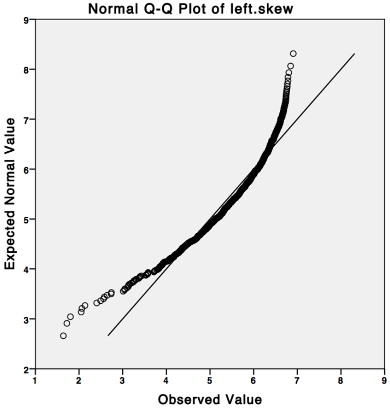
QQ plots (1)

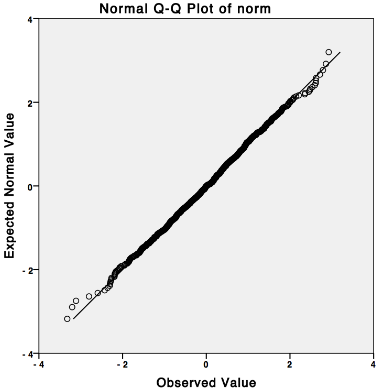




- Figure on the left is a data with both longer tails than normal.
- Figure on the right is a data with both shorter tails than normal.

QQ plots (2)



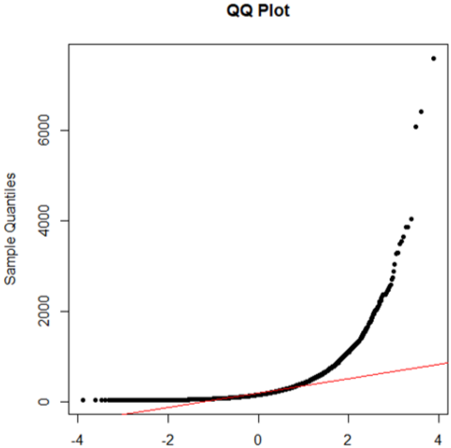


- Figure on the left is a data with left tail longer than normal but right tail is shorter than normal.
- Figure on the right is a data with both tails are normal.

QQ Plots in R

The code should be

```
> qqnorm(total, main = "QQ Plot", pch = 20)
> qqline(total, col = "red")
```



- The QQ plot of the sample has the right tail much longer than normal while the left tail is much shorter than normal.

1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

Quantifying the Association: Correlation Value

- Let X and Y are two features from a set of n points.

- The correlation of these two is defined as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\frac{X_i - \bar{X}}{s_X}\Big)\Big(\frac{Y_i - \bar{Y}}{s_Y}\Big)$$

  where $\bar{X}, \bar{Y}$ are the sample means, s_X, s_Y are the sample standard deviations of the two features.

- r is always between -1 and 1.

Correlation Value

- A positive value for r indicates a positive association and a negative value of r indicates a negative association.

```
> order = sales$num_of_orders
> cor(total, order)
[1] 0.7508015
```

Visualization the Association: Scatterplots

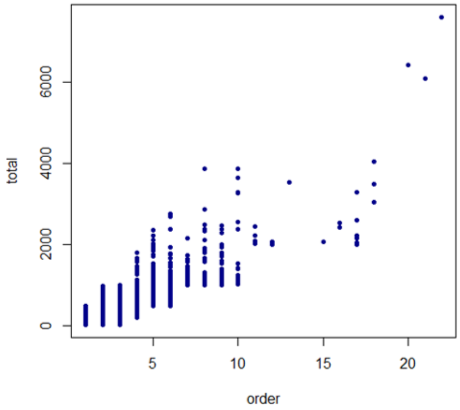
- Scatterplot can help to visualize the association between two quantitative features well.

  What to say given a scatterplot:

- Is there any (possible) relationship between the 2 variables?
- If yes, is the association positive or negative?
- If there is association, is it linear or non-linear type?
- Are some observations unusual, departing from the overall trend?

Scatterplots in R

```
> plot(order,total, pch = 20, col = "darkblue")
```

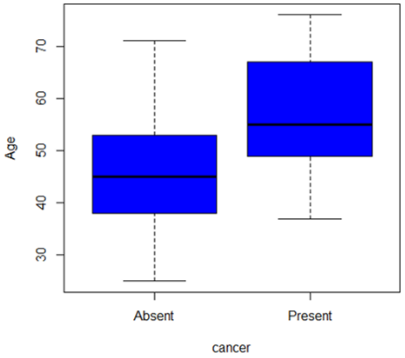


1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

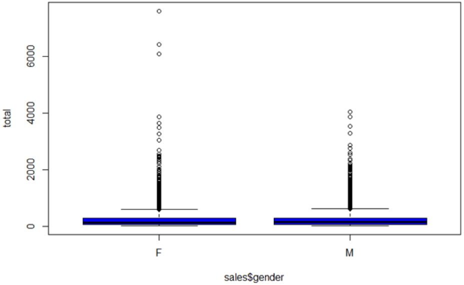
Boxplots of Multiple Groups



Categorical variable “cancer” has two categories: male and female. Variable “Age” is quantitative. One would check if any relationship between these two variables.

Boxplots of Multiple Groups in R

```
> attach(sales)
> boxplot(total ~ gender)
```



There is no obvious difference in the total sales of the customer's gender. The median of two groups are similar, and the IRQ are about the same.

Association of 3 Variables

- Can you figure out a way to visualize the association of the three features: total sales, number of orders and the gender of the customers?

1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

Summary of a Categorical Variable

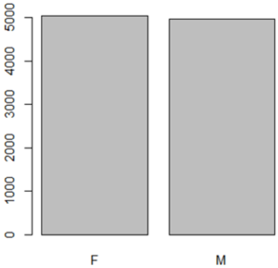
- For a single categorical variable, we can use **frequency table** (which also can produce the proportion or percentage) as numerical summaries.

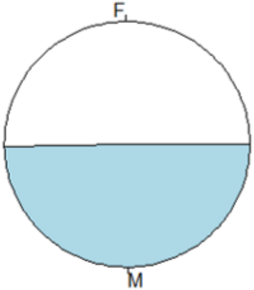
- The category with the highest frequency is reported as the **modal category**.

- Common graphical to display a categorical variable is **bar plot** or pie chart.

Barplot and Pie Chart

```
> count = table(gender)
> count # frequency table
gender
   F    M
5035 4965
> barplot(count)
> pie(count)
```





Two Categorical Variables

- Contingency table is often used to summarize the two categorical variables.

- Odds ratio is useful too.

Two Categorical Variables

- Categorizing the number of orders into two categories: small and large size.

```
> order.size = ifelse(order<=5, "small", "large")
> table(order.size)
order.size
large small
  324  9676
```

- Contingency table of frequency

```
> table = table(gender,order.size);table
      order.size
gender large small
     F   142  4893
     M   182  4783
```

Contingency Tables

- Contingency table of joint proportion

```
> prop.table(table)
      order.size
gender  large  small
     F 0.0142 0.4893
     M 0.0182 0.4783
```

Contingency Tables

- Contingency table of proportion by gender

```
> tab = prop.table(table, "gender") # proportion by gender
> tab
      order.size
gender      large      small
     F 0.02820258 0.97179742
     M 0.03665660 0.96334340
```

Among orders by females, 2.82% are large orders while 3.67% of orders by males are large.

Odds of Success

- For a probability of success π, the **odds of success** is defined as $odds = \pi/(1-\pi)$.

- If we consider having a large order is a success, then for the female groups, the odds of success, or the odds of large order, is 0.029.

  ```
  > tab[1]/(1-tab[1])
  [1] 0.02902105
  ```

- For the male group, the odds of having large order is 0.038.

  ```
  > tab[2]/(1-tab[2])
  [1] 0.03805143
  ```

Odds Ratio

- Odds ratio is the ratio of two odds of success: odds of larger orders in the female group (0.029), and odds of larger orders in the male group (0.038).

$$OR = \frac{0.029}{0.038} = 0.76.$$

What does this value mean?