Descriptive Statistics

STAT-UB.0001 Statistics for Business Control

Ningshan Zhang

IOMS Department
nzhang@stern.nyu.edu

July 3, 2018

Descriptive Statistics

Descriptive Statistics

- Methods of organizing, summarizing and presenting numerical data in a convenient form.
- Descriptive statistics are the foundation for any statistical analysis.
- What factors should you consider when deciding the best way to present your data?

Qualitative vs Quantitative Data

Qualitative: categorical

Examples:

- Level of Education
- Movie genre

Quantitative: numerical

Examples:

- Interest rates
- Temperature

Qualitative Data

In the study of qualitative data, one usually wants to compare the amount of participants in one group relative to another group.

- Frequency: Number of observations falling into a particular group.
- Relative Frequency: Proportions of observations falling into a particular group.

These can be presented numerically (table) or graphically (bar chart or pie chart).

Movies by MPAA Rating in 2012

Frequency Table[1]

Rating	Frequency
G	14
PG	59
PG-13	140
R	210
NC-17	2
Not Rated	36
Total	**461**

[1] Source: `www.the-numbers.com`, Movie2012.mtw

Movies by MPAA Rating in 2012

Frequency Table $\rightarrow$ Relative Frequency Table

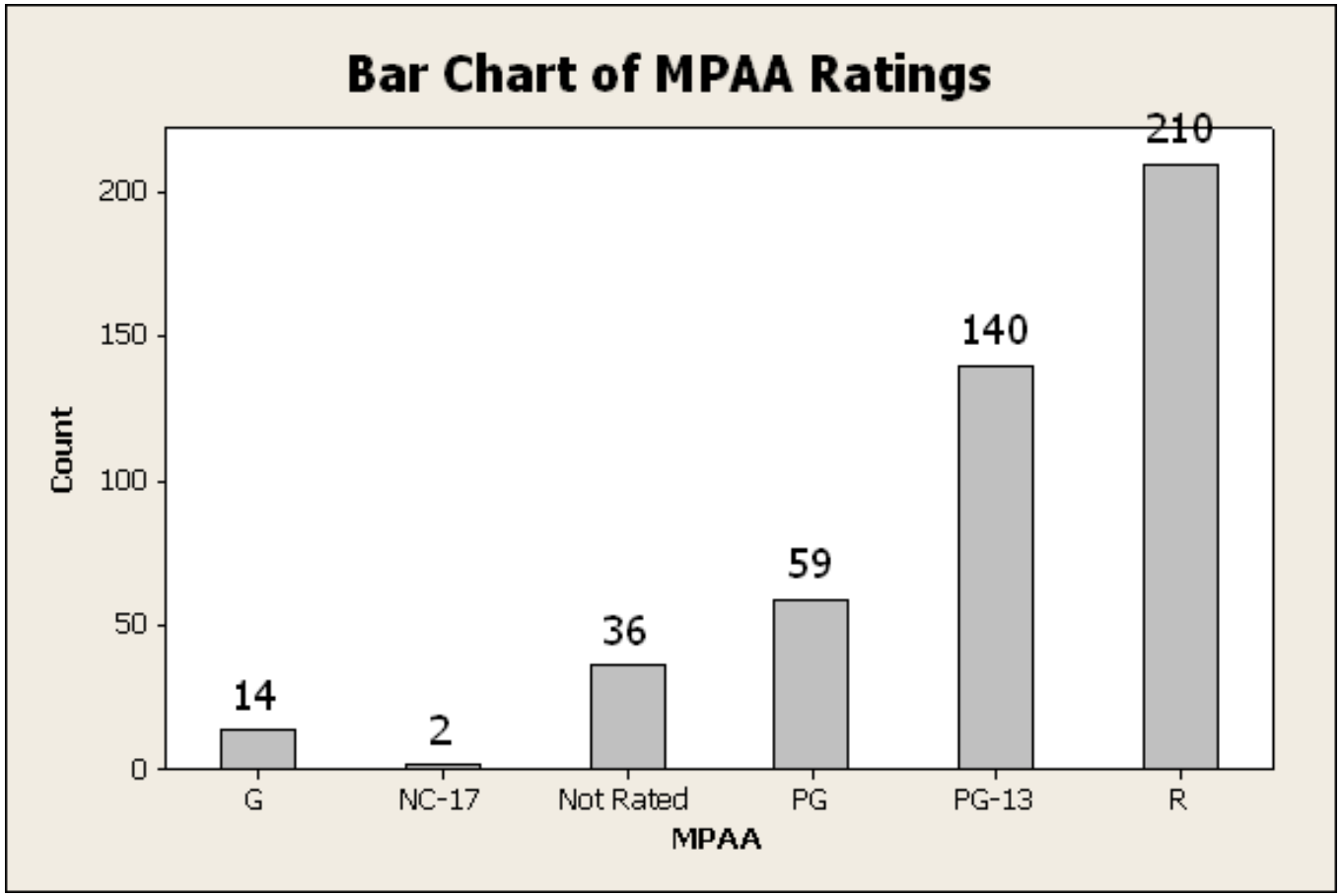
Rating	Frequency
G	14
PG	59
PG-13	140
R	210
NC-17	2
Not Rated	36
Total	**461**

$\rightarrow$

Rating	Rel. Frequency
G	
PG	
PG-13	
R	
NC-17	
Not Rated	
Total	**100%**

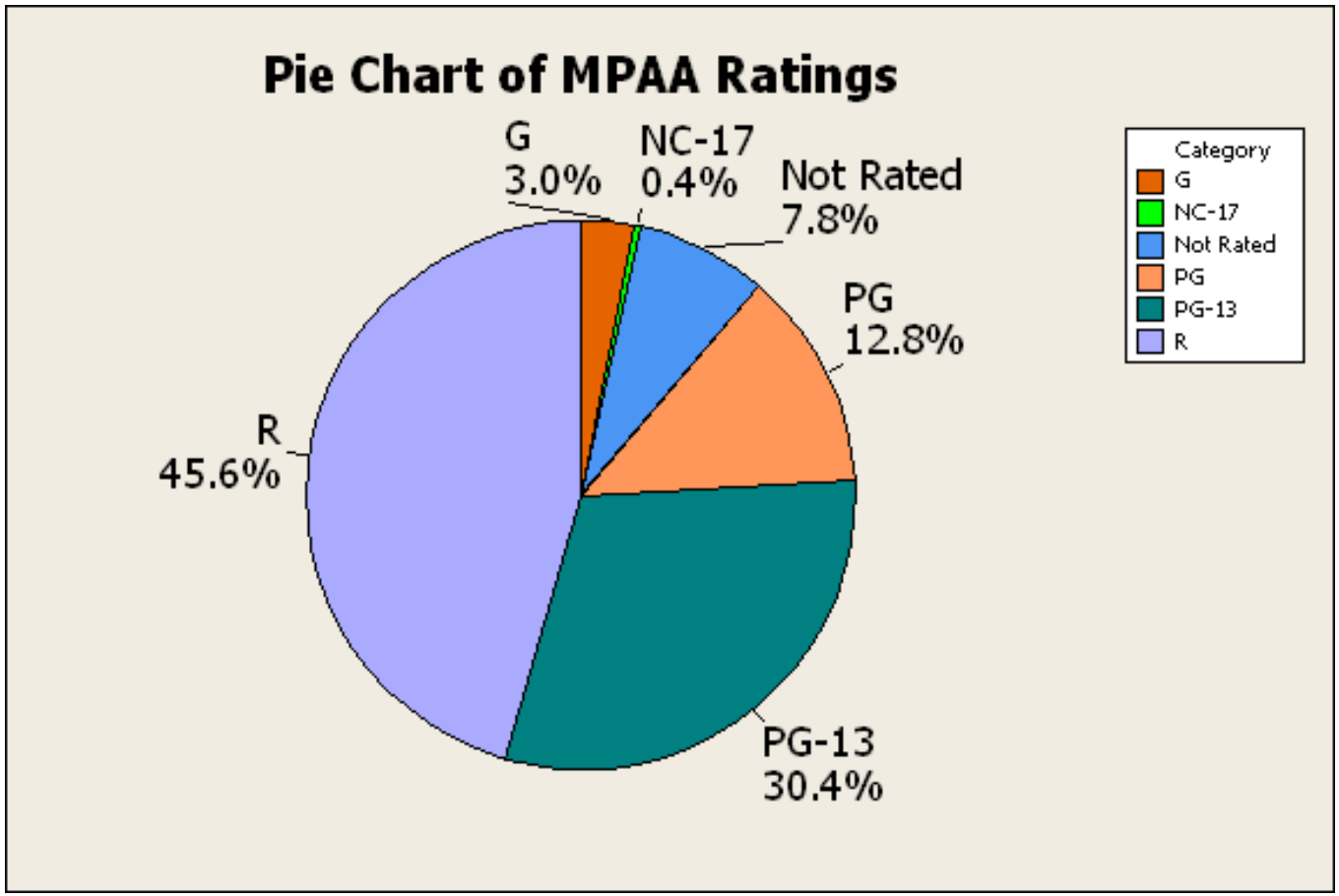
Movies by MPAA Rating in 2012

Bar chart of Frequencies



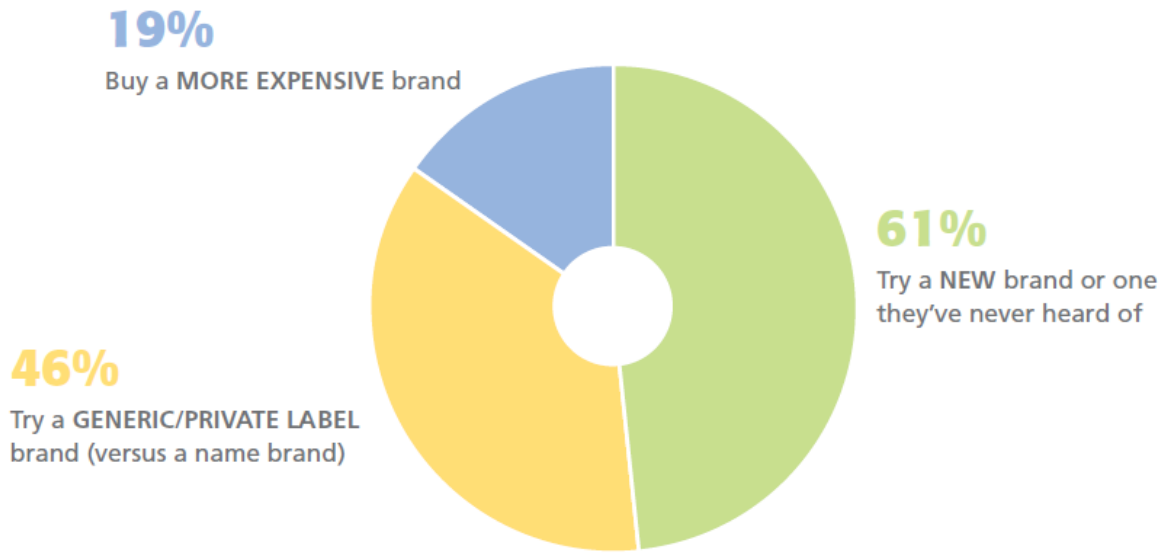
Movies by MPAA Rating in 2012

Pie chart of Relative Frequencies



Charts Gone Bad

According to the 2010 Cause Evolution Study[2], " ... Consumers are willing to:"



What is wrong with this chart?

[2] http://www.conecomm.com/2010-cone-communications-cause-evolution-study-pdf/

Quantitative Data

In the study of quantitative data, one usually wants to find and display certain distributional properties.

Numerical summaries

- Measures of central tendency
- Measures of variability
- Identifying outliers

Graphical summaries

- Histograms
- Boxplots
- Time Series Plot

Measures of Central Tendency: Mean vs Median

Mean

The mean of a sample is the average of the observations:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right).$$

Median

The median is the middle value in a *sorted* dataset.

- When n is odd, take "true" middle value.
- When n is even, take the average of the two middle values.

What is the mean and median of $\{6, 4, 19, 6, 12, 8, 13, 0\}$?

Measures of Central Tendency: Mean vs Median

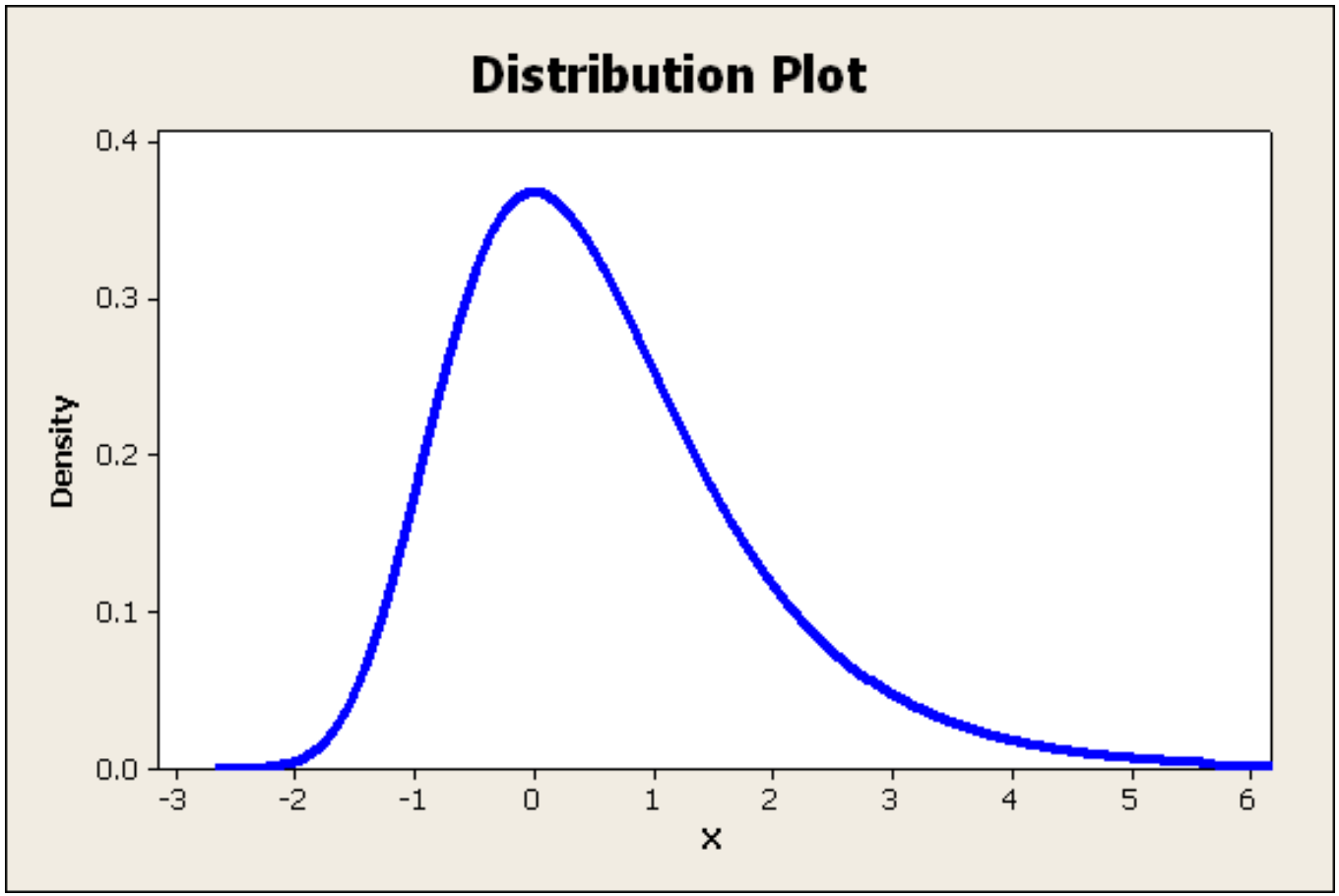
Comparing the mean and the median helps us detect skewness in the data.

(Nonparametric) Skewness

- Positive/right skew: mean - median > 0 , mean is to the right of the median.
- Negative/left skew: mean - median < 0, mean is to the left of the median.

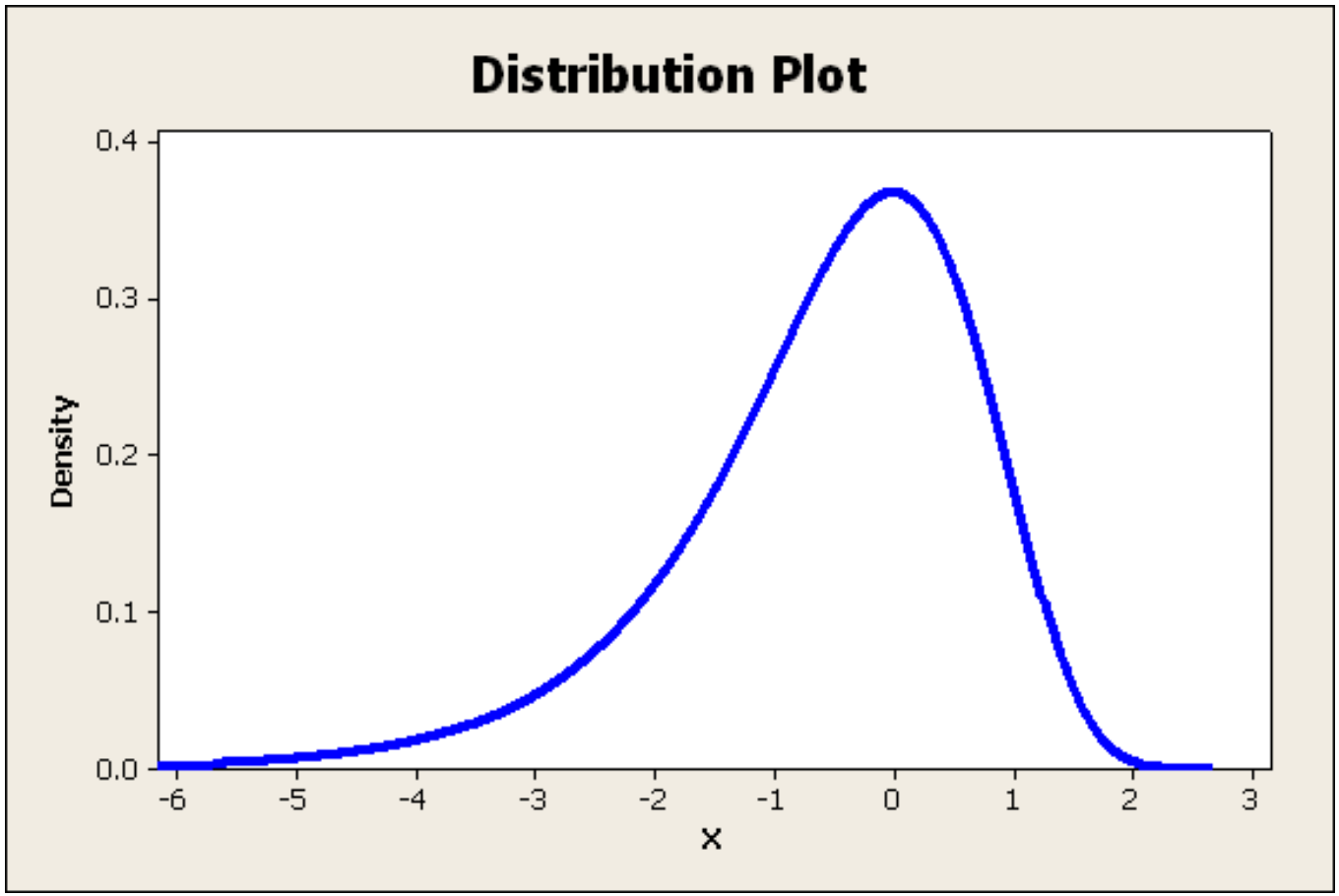
Positive/Right Skewed

Mean $>$ Median. Usually appears as a *left*-leaning curve. Also called right-tailed since right tail is longer.



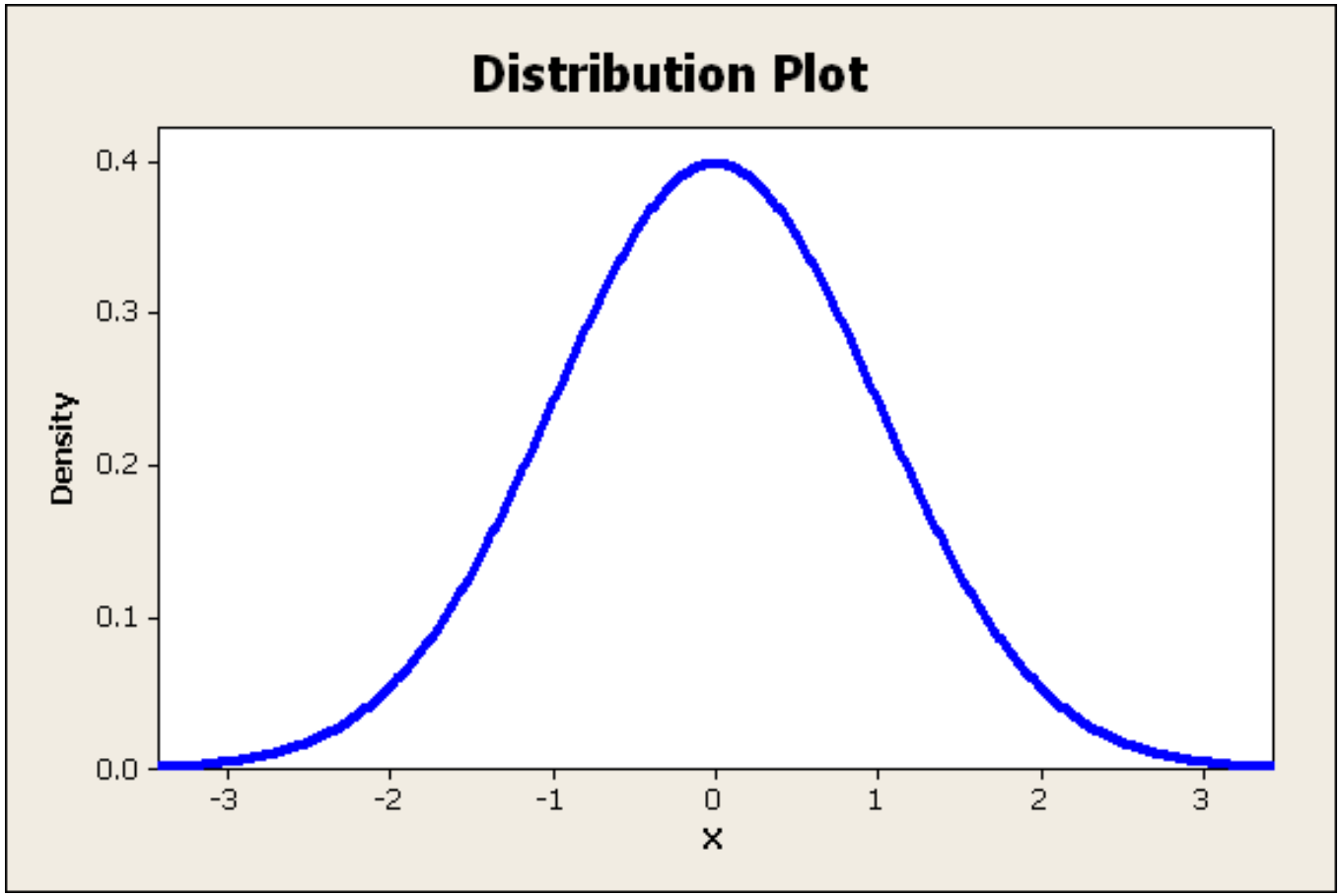
Negative/Left Skewed

Mean $<$ Median. Usually appears as a *right*-leaning curve. Also called left-tailed since left tail is longer.



Not Skewed or Symmetric

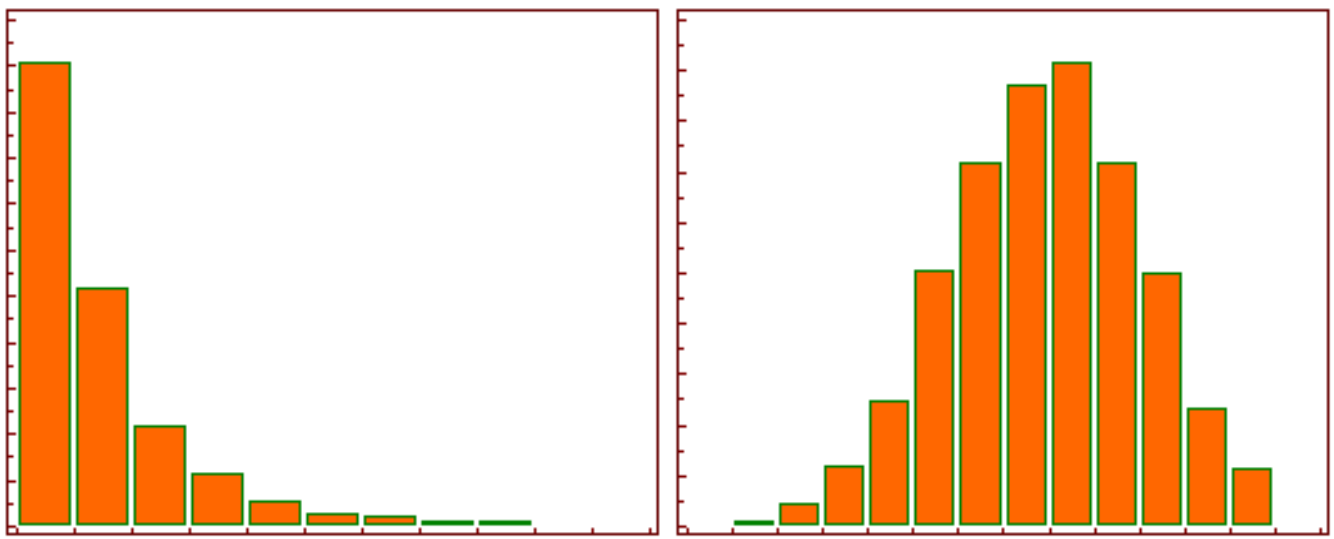
Mean = Median.



Log Transformation

The log transformation usually can make highly skewed distributions, especially right skewed distributions less skewed. (But not always.)

Figure: Left: Histogram of x, the original data. Right: Histogram of $log(x)$.



* Plot from https://www.medcalc.org/manual/log_transformation.php

Measures of Central Tendency: Mode

Mode

The mode is the most common value in a data set.

Sample $= \{6, 4, 19, 6, 12, 8, 13, 0\}$. What is the mode?

If the distribution is both symmetric and unimodal, then

$$\text{Mean} = \text{Median} = \text{Mode}.$$

Measuring Variability

Variability refers to the spread in the data. Common measures:

- Range, or Minimum & Maximum.
- Inter-Quartile.
- Variance or Standard Deviation.

Measuring Variability: Range

The simplest measure of variability is the range of the data:

- Minimum = smallest value in a dataset.
- Maximum = largest value in a dataset.
- Range = Maximum - Minimum.

Example: $\{6, 4, 19, 6, 12, 8, 13, 0\}$.

Measuring Variability: Inter-quartile Range

A more useful quantity is the inter-quartile range (IQR):

- 1st quartile = the $(n+1)/4$-th value in a sorted dataset (aka lower quartile, Q_L , 25th percentile, Q_1).
- 3rd quartile = the $3(n+1)/4$-th value in a sorted dataset (aka upper quartile, Q_U, 75th percentile, Q_3).
- IQR $= Q_U - Q_L$.

Example: $\{6, 4, 19, 6, 12, 8, 13, 0\}$, $Q_L = 4.5$, $Q_U = 12.75$.

Percentile: generalization of quartile, Qth percentile is the number such that Q% of all observations are less to.

Measuring Variability: Variance and Standard Deviation

The most important measures of variability are the sample variance and the sample standard deviation.

- Sample Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - n\bar{x}^2)$$

- Sample standard deviation:

$$s = \sqrt{s^2}$$

Example: $\{6, 4, 19, 6, 12, 8, 13, 0\}$.

Z-score and Outliers

Z-score

Z-score is the number of standard deviations the observation is away from the mean. Formally, the z-score of x is

$$z = \frac{x - \bar{x}}{s},$$

where

- x is an observed value,
- $\bar{x}$ is the sample mean,
- s is the sample standard deviation.

Z-score and Outliers

Outliers are the observations with *unusually* large or *unusually* small values relative to the *other values* in a data set.

In other words, outliers have large absolute z-scores.

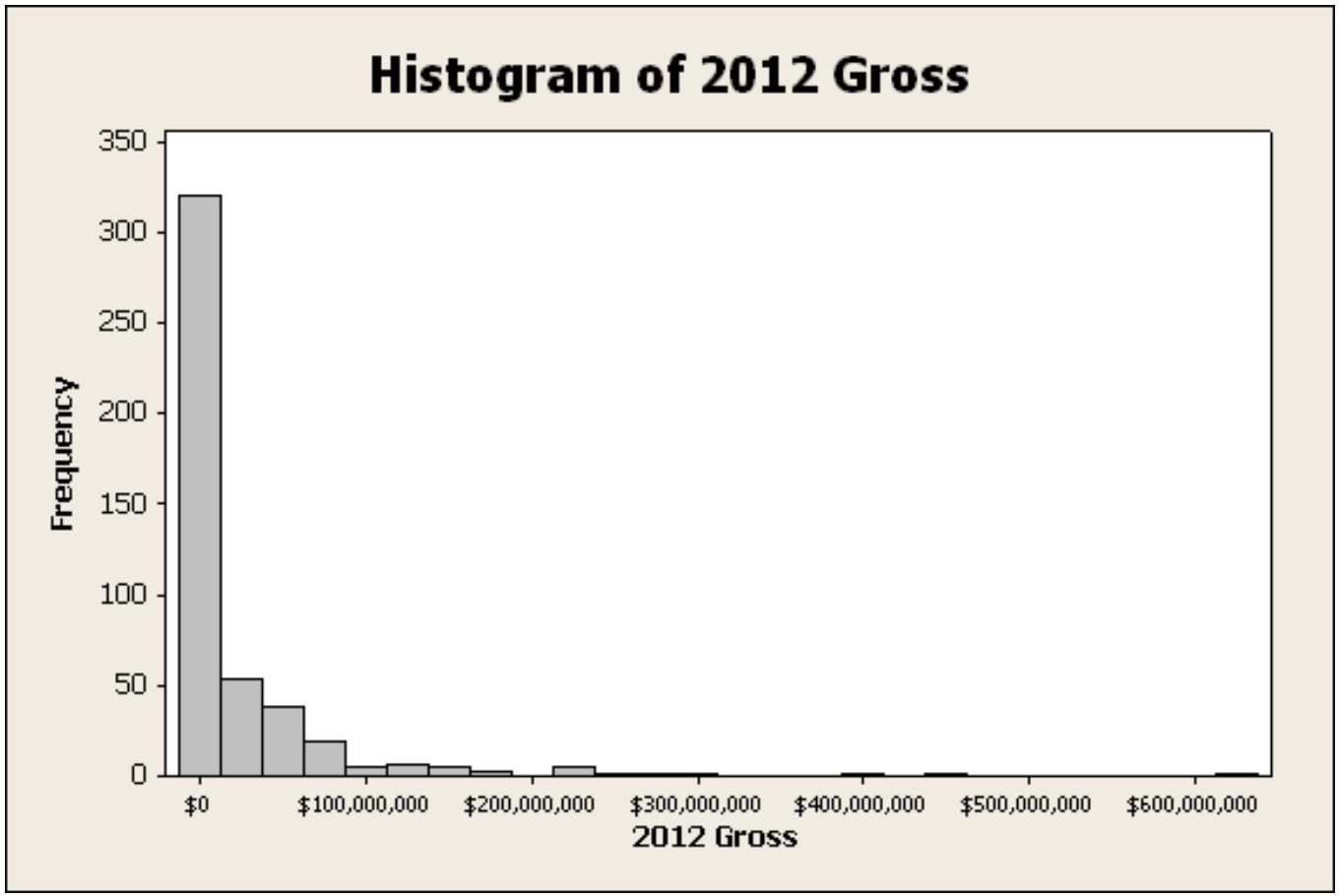
Identifying Outliers: Empirical Rule

For roughly bell-shaped distributions,

- About 68% of data will have z-scores in (-1,1), i.e. within the range $[\bar{x} - s, \bar{x} + s]$.
- About 95% of data will have z-scores in (-2,2), i.e. within the range $[\bar{x} - 2s, \bar{x} + 2s]$.
- About 99.7% of data will have z-scores in (-3,3), i.e. within the range $[\bar{x} - 3s, \bar{x} + 3s]$.

Histograms

Histograms provide a visual representation of the distribution of the data. Example[3]:

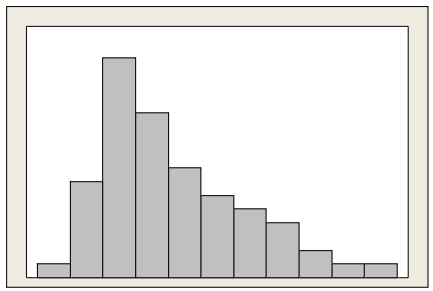


[3] Source: `www.the-numbers.com`, Movie2012.mtw

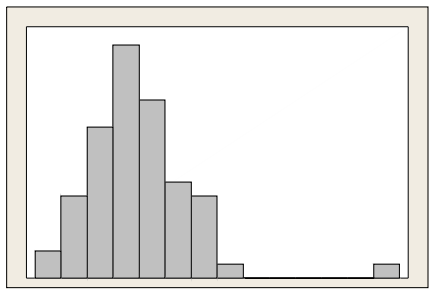
Histograms

With a histogram, we can detect

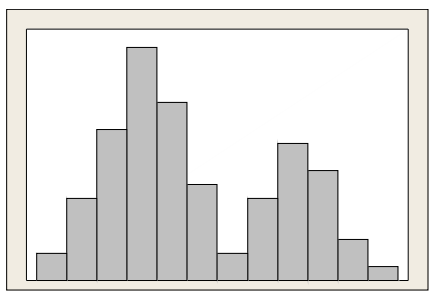
- Skewness



- Outliers

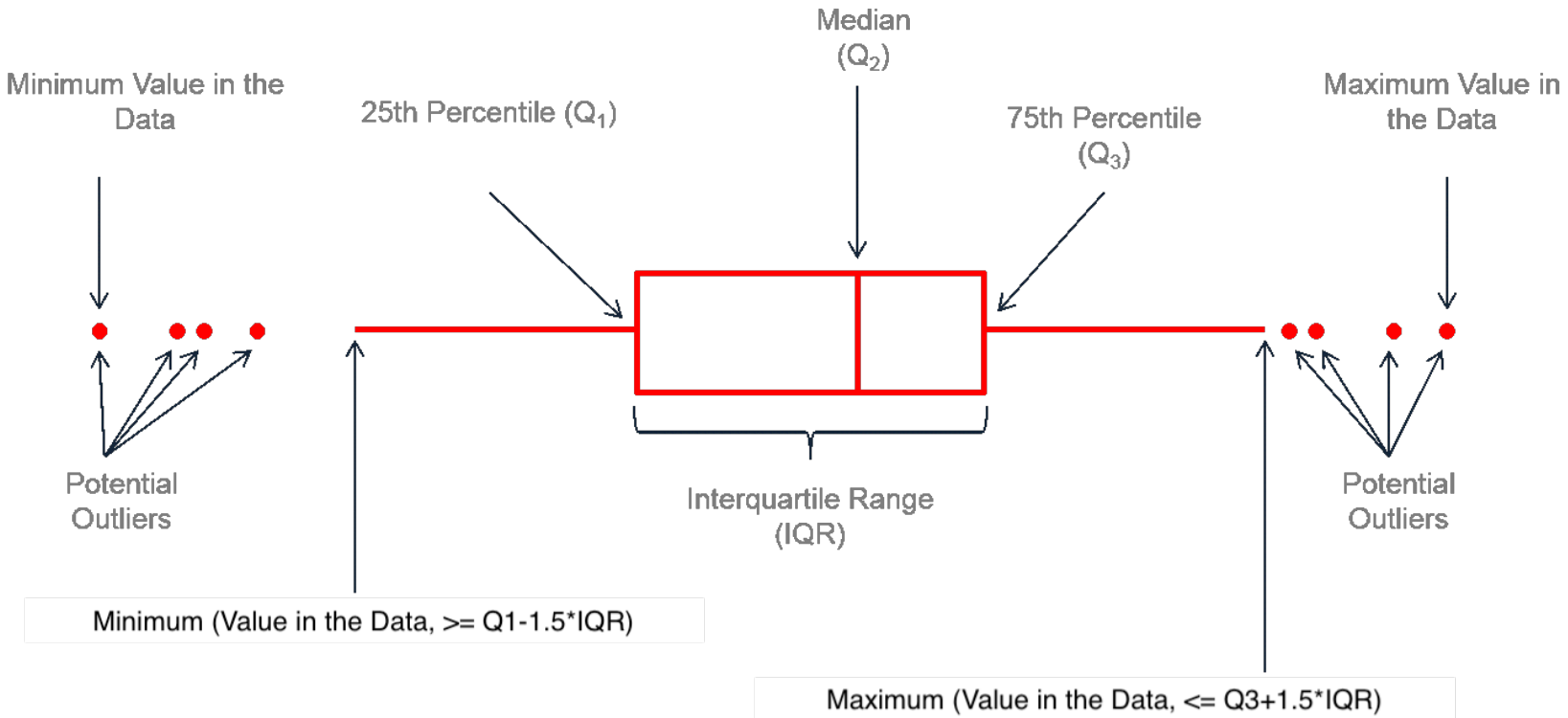


- Bimodal distribution



The Box-and-whisker plot

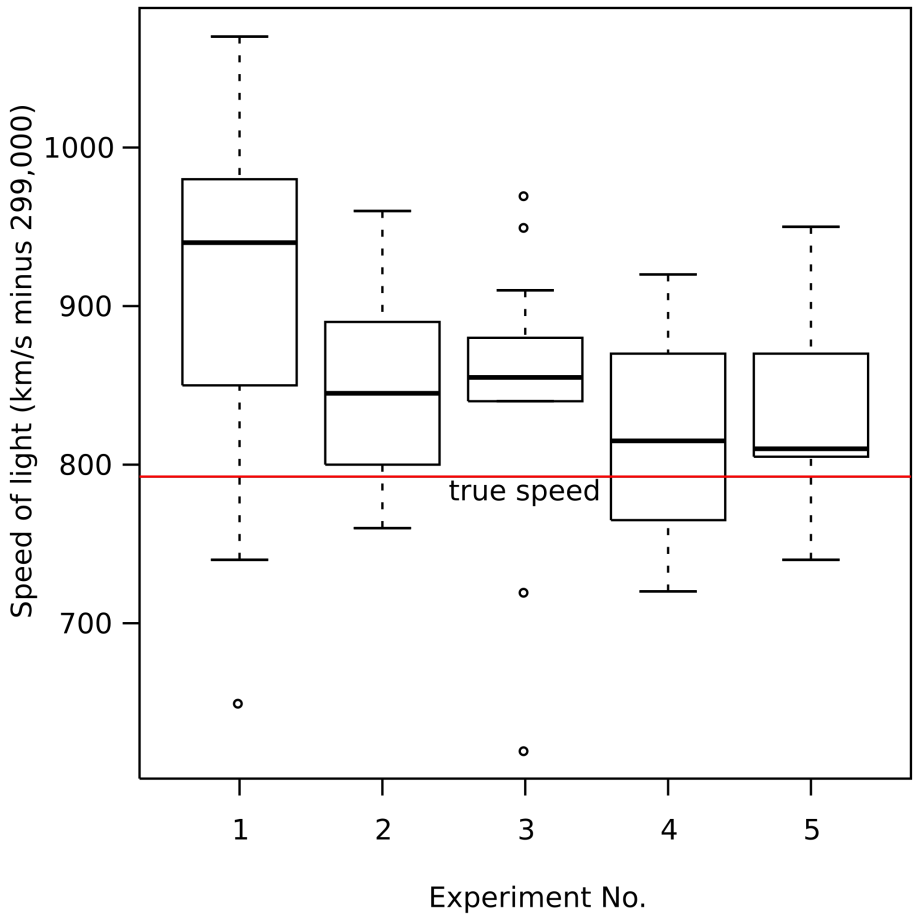
Boxplots also provide a visual representation of the distribution of the data.



* Plot from `https://www.leansigmacorporation.com/box-plot-with-minitab/`

Boxplots are excellent for comparing distributions

Figure: Box plot of data from the MichelsonMorley experiment



* Plot from Wikipedia

Identifying Outliers

Two methods for identifying outliers:

- Z-score method
  - Observations with z-scores outside $(-2, 2)$ are outliers.
  - Stated differently, observations outside $(\bar{x} - 2s, \bar{x} + 2s)$ are outliers.
- Boxplot method
  - Observations outside $(Q_L - 1.5 * \text{IQR}, Q_U + 1.5 * \text{IQR})$ are outliers.
  - Observations outside $(Q_L - 3 * \text{IQR}, Q_U + 3 * \text{IQR})$ are serious outliers.

May produce different results.

Time Series Plot

Time series plots are useful when time sequencing is important.

Figure: Bitcoin Price from Jun 30, 2017 to Jun 30, 2018.

* Plot from Coindesk.com

Summary of Descriptive Statistics

Qualitative (Categorical)

- Numerically: Frequency, Relative Frequency.
- Graphically: Bar chart, Pie chart.

Quantitative (Numerical)

- Numerically
  - Measures of central tendency (mean, median, mode).
  - Measures of variability (range, IQR, standard deviation).
  - Identifying outliers (z-scores, empirical rule).
- Graphically: Histogram, Box plot, Time series plot.