



BITS Pilani
Pilani Campus

Descriptive Statistics

Akanksha Bharadwaj
Asst. Professor, BITS Pilani



BITS Pilani
Pilani Campus



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS Contact session 1

Books



No	Author(s), Title, Edition, Publishing House
T1	Probability and Statistics for Engineering and Sciences, 8 th Edition, Jay L Devore, Cengage Learning
T2	Applied Logistic Regression, Hosmer and Lemeshow, 3 rd Edition, Wiley
T3	Introduction to Time Series and Forecasting, Second Edition, Peter J <u>Brockwell</u> , Richard A Davis, Springer.

No	Author(s), Title, Edition, Publishing House
R1	Miller and Freund's Probability and statistics for Engineers, 8 th Edition, PHI
R2	Statistics for Business and Economics by Anderson, Sweeney and <u>Williams</u> , CENAGE learning

Session Plan



⊕ Contact hour-1, Module 1:

Time	Type	Description	References
Pre- CH-1	RL	RL – 1.1.1 (Data representation)	
During CH-1	CH-1	Discussion on data representation	T1:Chapter 1
Post-CH-1	HW	T1: Chapter 1	T1:Chapter 1
Lab			

Contact hour-2, Module 1

Time	Type	Description	References
Pre-CH-2	RL	RL – 1.1.2 (Data visualization)	
During CH-2	CH-2	Measures of Central Tendency, Measures of Variability	T1:Chapter 1
Post-CH-2	HW	T1: Chapter 1	T1:Chapter 1
Lab			

What is Statistics?



- a form of knowledge- a mode of arranging and stating facts which belong to various sciences (Lond. And Westn. Rev, 1838)
- Science dealing with collection, analysis, interpretation, and presentation of masses of numerical data (Webster dictionary, 1966)
- Science of collecting and analysing numerical data (Oxford dictionary, 1996)

Population Vs Sample

- An investigation will typically focus on a well-defined collection of objects constituting a **population** of interest
- When desired information is available for all objects in the population, we have what is called a **census**.
- Constraints on time, money, and other scarce resources usually make a census impractical or infeasible.
- Instead, a subset of the population—a **sample**—is selected in some prescribed manner.

Parameter Vs Statistic



- A descriptive measure of the population is called **parameter**
- A descriptive measure of the sample is called **statistic**

Branches of statistics



- Descriptive statistics
- Inferential statistics

Descriptive statistics



- If a business analyst is using data gathered on a group to describe and reach conclusions about the same group, the statistics are called descriptive statistics.
- Example- if an instructor produces statistics to summarize a class's examination efforts and uses those statistics to reach conclusions about that class only.

Inferential statistics



- If a researcher gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken
- Example- pharmaceutical research



BITS Pilani
Pilani Campus

Terminologies

Variable



- A **variable** is any characteristic whose value may change from one object to another in the population.
- E.g. age the patients, number of visits to a particular website , etc.

Types of variable



- **Categorical/ qualitative variables:**
 - Take category or label values and place an individual into one of several groups.
 - Each observation can be placed in only one category, and the categories are mutually exclusive.
- **Quantitative variables:**
 - Take numerical values and represent some kind of measurement.

Example: Indian census data 2010



	State	Zip code	Family size	Annual income
1	U.P	201001	5	10,00,000
2	Delhi	110092	10	25,00,000
3	Gurgaon	122503	12	40,00,000
4	Delhi	110091	4	8,00,000
5	U.P	201003	2	2,00,000
6	Gurgaon	122004	1	5,00,000

Data Sets



- A **dataset** is a set of data identified with particular circumstances. Datasets are typically displayed in tables, in which rows represent individuals and columns represent variables.
- A **univariate** data set consists of observations on a single variable.
- **Bivariate** data sets have observations made on two variables
- **Multivariate** data arises when observations are made on more than one variable (so bivariate is a special case of multivariate)

Data Measurement



Nominal level-

- It is the lowest level of data measurement.
- Numbers representing nominal level data can be used only to classify or categorize
- Example- Employee ID

Data Measurement



Ordinal Level-

- In addition to nominal level capabilities, it can be used to rank or order objects
- The categories for each of these ordinal variables show order, but not the magnitude of difference between two adjacent points.
- Example- a supervisor can rank the productivity of employees from 1 to 5

Data Measurement



Interval level-

- In this distances between consecutive numbers have meaning and the data are always numerical.
- the distance between pairs of consecutive numbers is assumed to be equal.
- Zero is just another point on scale and does not mean the absence of phenomenon
- Example- temperature in Fahrenheit

Data Measurement



Ratio level-

- It has same properties as interval data, but ratio data have an absolute zero, and the ratio of two numbers is meaningful
- Example- height, weight, time, volume, production cycle time, etc.
- For instance, we know that someone who is forty years old is twice as old as someone who is twenty years old.
- There is a meaningful zero point – that is, it is possible to have the absence of age.

Exercise- Healthcare industry



The following type of questions are sometimes asked in the survey. These question will result in what level of data measurement

- How long ago were you released from the hospital?
- Which type of unit were you in for most of your stay?
 - Intensive care
 - Maternity care
 - Surgical unit
- How serious was your condition when you were first admitted to the hospital
 - Critical
 - Moderate
 - Minor
 - 1- it is a time measurement with absolute zero and is therefore ratio level measurement
 - 2- nominal data as it is used only to categorize
 - 3- it can be ranked by selection so ordinal data



BITS Pilani
Pilani Campus



Data Visualization



Pie charts and bar chart for a categorical data



- Refer to BMI data
- Both the pie chart and the bar chart help us visualize the distribution of a categorical variable

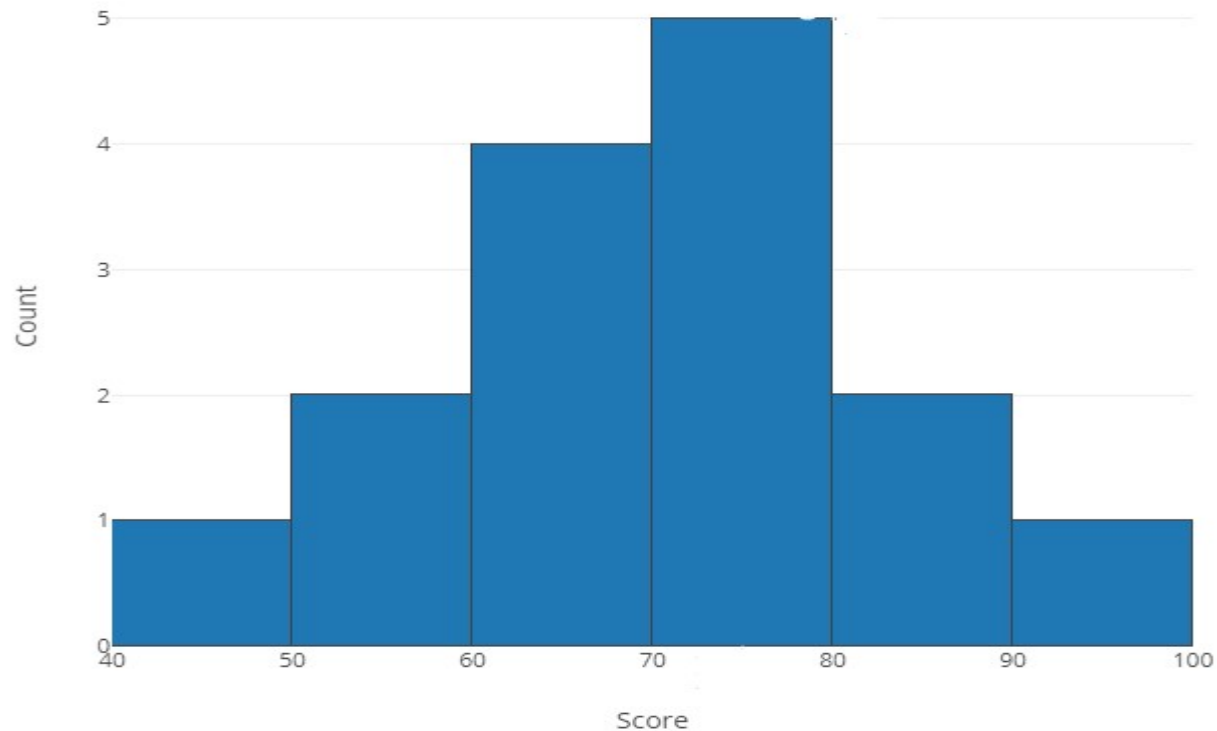
Histograms for quantitative data



Here are the score in mathematics for 15 students:

78, 58, 65, 71, 57, 74, 79, 75, 87, 92, 81, 69, 66, 43, 63

Score	Count
✓ [40-50]	1 ✓
✓ [50-60]	2
✓ [60-70]	4
✓ [70-80]	5 ✓
✓ [80-90]	2
✓ [90-100]	1



Ques. What percentage of students earned less than a grade of 70 on the exam?

$$7/15 * 100$$

Example



A survey was conducted to see how many video calls people made daily. The results are displayed in the table below:

Number of calls made	Frequency
1 – 3	10
4 - 7	7
8 – 11	4
12 - 15	1
16 - 19	1

Ques1. Tell how many of the people surveyed make less than 4 video calls daily? 10

Ques2. How many people were surveyed? 23

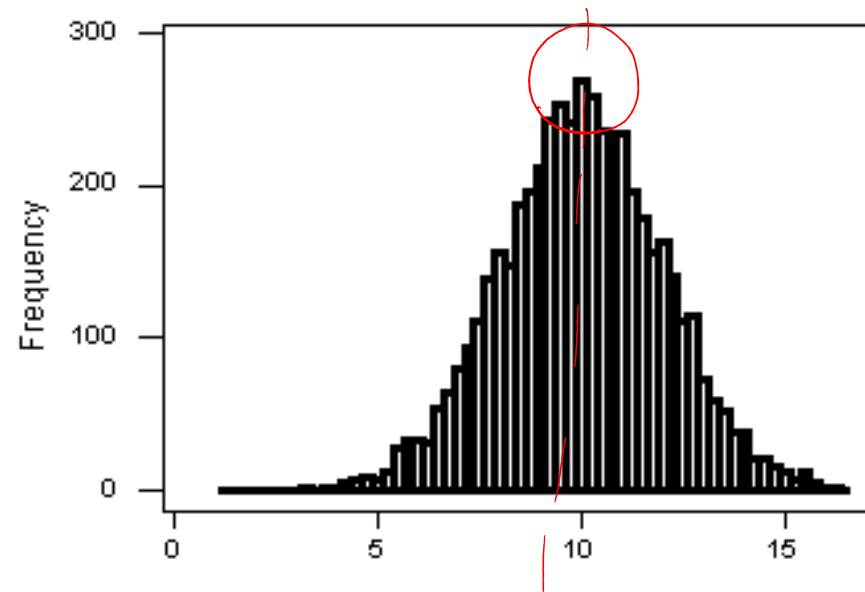
Shape of histograms



When describing the shape of a distribution, we should consider:

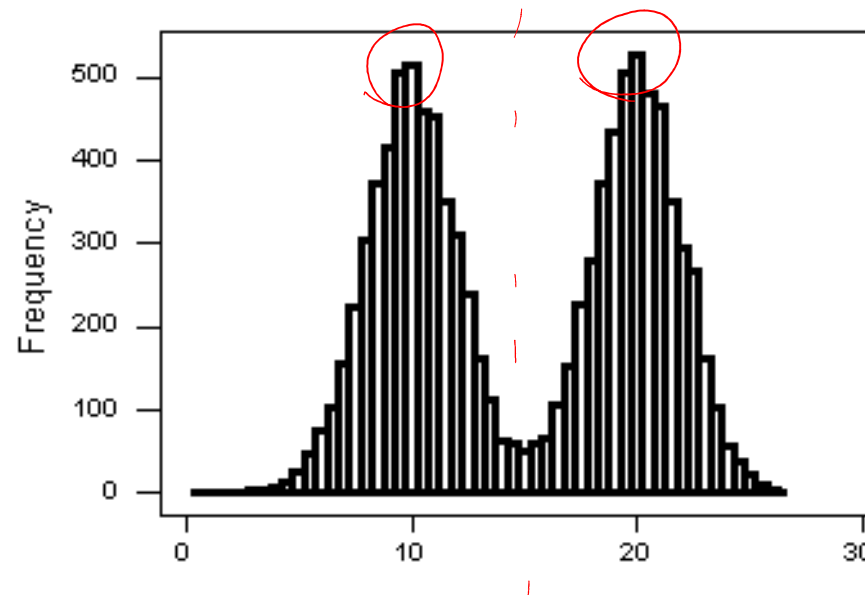
- **Symmetry/skewness** of the distribution.
- **Peakedness (modality)**—the number of peaks (modes) the distribution has.

Symmetric and single peaked distribution



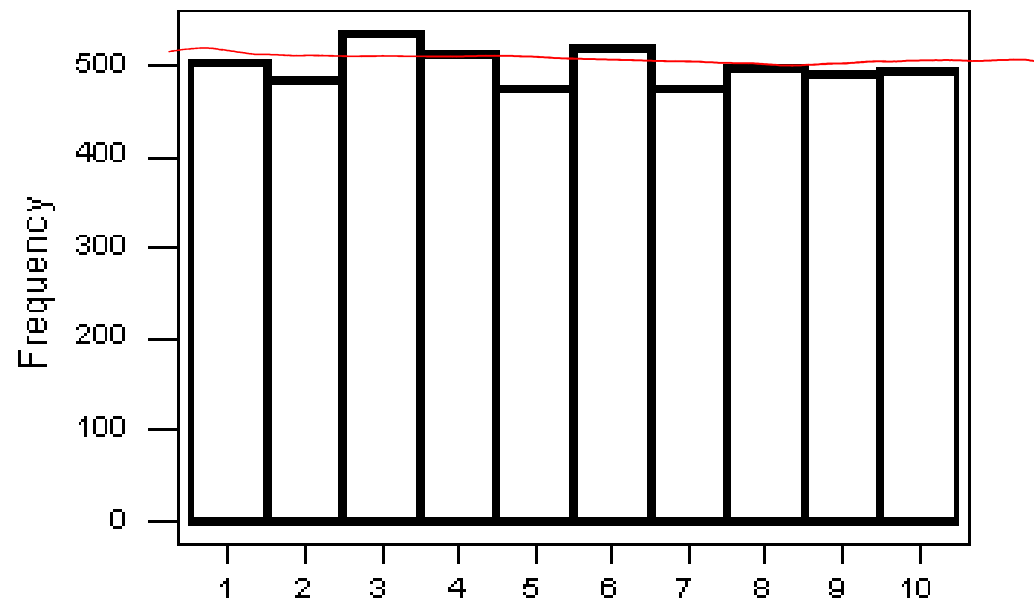
<https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/>

Symmetric and double peaked distribution



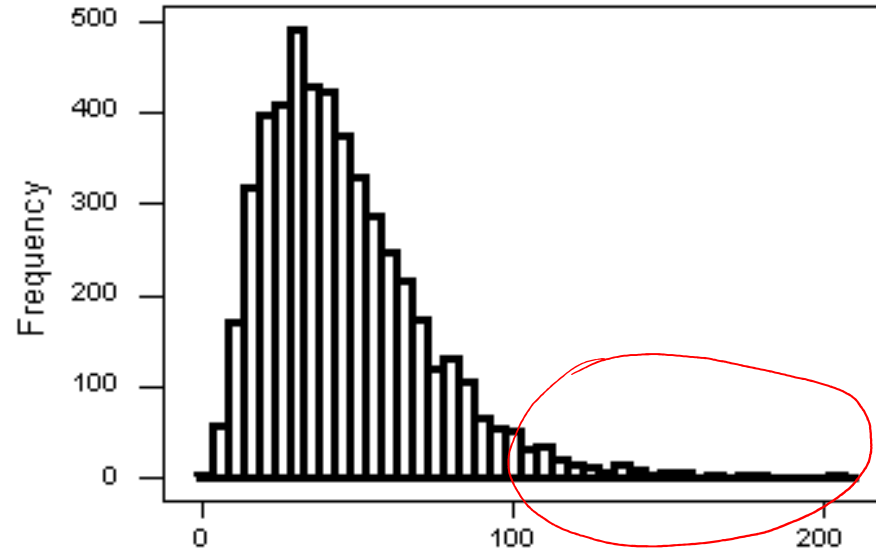
<https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/>

Symmetric and flat distribution



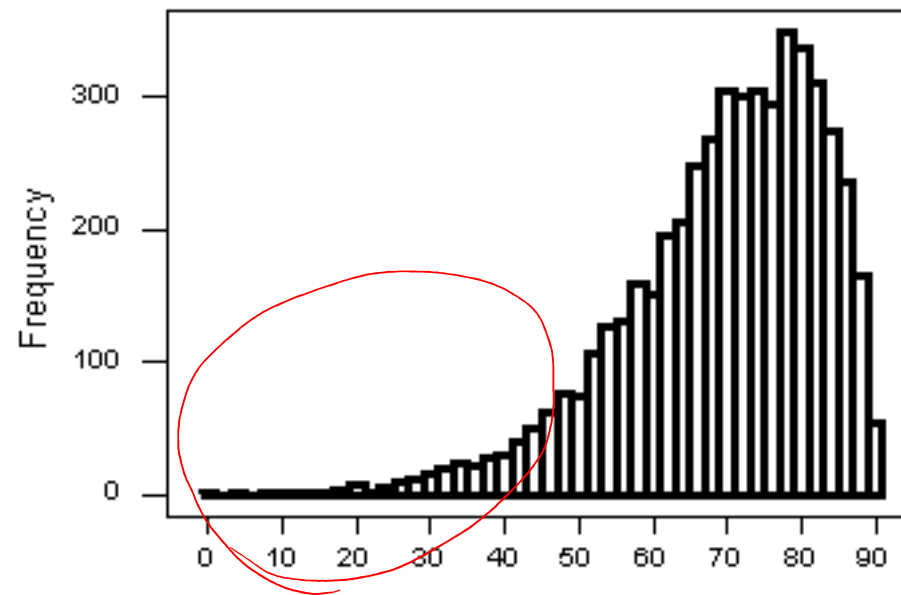
<https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/>

Right skewed distribution



<https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/>

Left skewed distribution

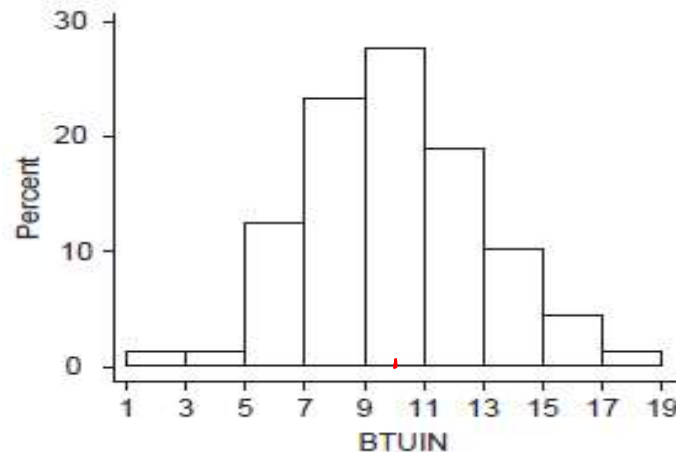


<https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/>

Center



The center of the distribution is its **midpoint**—the value that divides the distribution so that approximately half the observations take smaller values, and approximately half the observations take larger values.



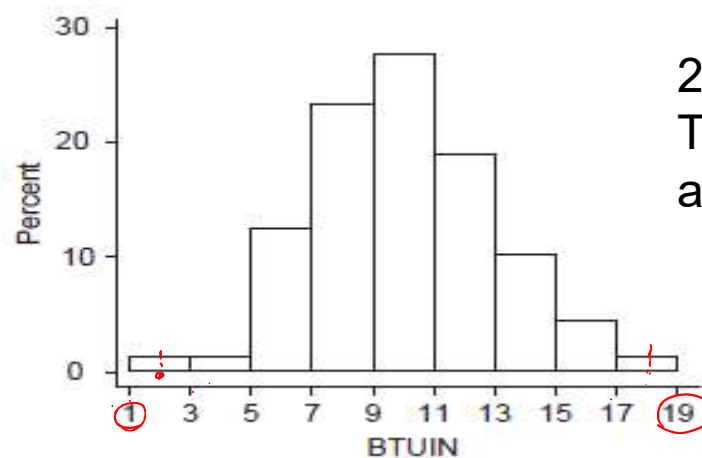
Histogram of the energy consumption data

Image: Book (Probability and statistics for the engineering and sciences by Devore

Spread



The **spread** (also called **variability**) of the distribution can be described by the approximate range covered by the data. From looking at the histogram, we can approximate the smallest observation (min), and the largest observation (max), and thus approximate the range.



2 to 18

Take the middle of the lowest and highest interval of scores

Histogram of the energy consumption data

Image: Book (Probability and statistics for the engineering and sciences by Devore

Stem and Leaf Plot



The stemplot (also called stem and leaf plot) is another graphical display of the distribution of quantitative data.

Separate each data point into a stem and leaf, as follows:

- The leaf is the right-most digit. *222 3*
- The stem is everything except the right-most digit. *222 5*
- So, if the data point is 54, then 5 is the stem and 4 is the leaf. *222 7*
- If the data point is 5.35, then 5.3 is the stem and 5 is the leaf.

Example

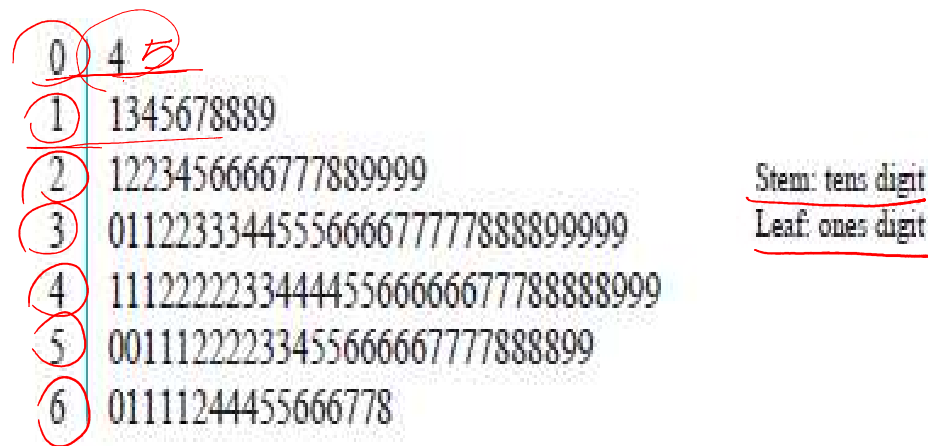


Figure 1.4 Stem-and-leaf display for the percentage of binge drinkers at each of the 140 colleges

0 5
0 4
1 3
1 3
1 4
1 5
1 6
1 7
:
:
:
2 1
2 2
2 2
2 3

2 2 2 3
2 2 2 5
2 2 2 7

2 2 2 | 3 5 7

Dotplot



- Used to summarize a quantitative variable graphically. The dotplot, like the stemplot, shows each observation, but displays it with a dot rather than with its actual value.
- When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically.

Example



A dot plot of 50 random values from 0 to 9.

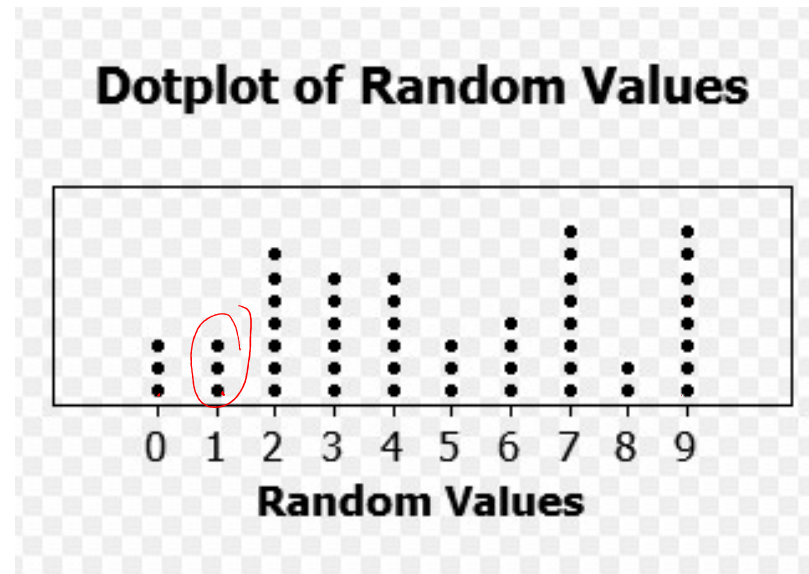


Image: Wikipedia



BITS Pilani
Pilani Campus



Measures of Location/Central Tendency: Ungrouped Data

Introduction



Here are the number of hours that 9 students spend on social media on a typical day:

11 6 7 5 2 8 11 12 15

Summarize the data using a single digit

Mean



Mean is the sum of the observations divided by the number of observations

11 6 7 5 2 8 11 12 15

Ques1. Mean is?

$$\underline{77/9} = 8.55$$

Sample Mean	Population Mean
$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum x}{N}$

where $\sum X$ is sum of all data values

N is number of data items in population

n is number of data items in sample

Mean



When to Use the Mean

- Sampling stability is desired.
- Other measures are to be computed such as standard deviation, coefficient of variation and skewness

Median



The median **M** is the midpoint of the ordered distribution.

Steps:

- Order the data from smallest to largest.
- Consider whether n , the number of observations, is even or odd.
 - If n is odd, the median M is the center observation in the ordered list. This observation is the one "sitting" in the $(n + 1) / 2$ spot in the ordered list.
 - If n is even, the median M is the mean of the two center observations in the ordered list. These two observations are the ones "sitting" in the $n / 2$ and $n / 2 + 1$ spots in the ordered list.

Example



11 6 7 5 2 8 11 12 15

Ques1. Median is?

* 2 5 6 7 8 11 11 12 15

Location is $(9+1)/2 = 5^{\text{th}}$ element

So median is 8

Mode



the mode is the most commonly occurring value in a distribution.

11 6 7 5 2 8 11 12 15 12

Ques1. Mode is?

Mode will be 11

Ques2. What kind of distribution is formed by the data from the above 9 students?

Unimodal

Comparing the Mean and the Median



The mean is very sensitive to outliers, while the median is resistant to outliers?

TRUE or FALSE?

Use the below data to analyze

Data set A → 54 55 56 68 70 71 73

Data set B → 54 55 56 68 70 71 730

Comparing the Mean and the Median: Interpretations



- For symmetric distributions with no outliers: mean is approximately equal to median



- For skewed right distributions and/or datasets with high outliers: mean $>$ median



- For skewed left distributions and/or datasets with low outliers: mean $<$ median

Ques. The Current Population Survey conducted by the Census Bureau records the incomes of a large sample of Indian households each month. What will be the relationship between the mean and median of the collected data?

Ans. mean $>$ median since data will be right skewed

The mean is an appropriate measure of center only for symmetric distributions with no outliers. In all other cases, the median should be used to describe the center of the distribution.

Quartiles



- Quartiles in statistics are values that divide your data into quarters.
- However, quartiles aren't shaped like pizza slices; Instead they divide your data into four segments according to where the numbers fall on the number line.

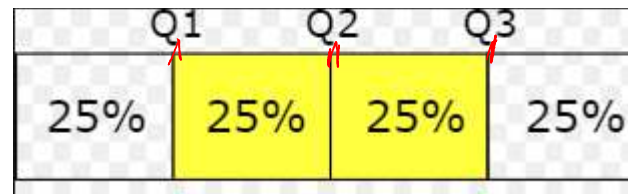
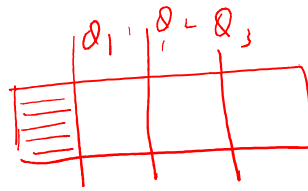


Image: Wikipedia

Quartiles



There are three quartiles: the first quartile (Q_1), the second quartile (Q_2), and the third quartile (Q_3).

- The first quartile (lower quartile, Q_L), is equal to the 25th percentile of the data.
- The second (middle) quartile or median of a data set is equal to the 50th percentile of the data
- The third quartile, called upper quartile (Q_U), is equal to the 75th percentile of the data.

Example

$$\left(\frac{n}{2}\right) \left(\frac{n}{2} + 1\right)$$

Ordered Data Set: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Q2 will be $(n+1)/2 = 12/2 = 6^{\text{th}}$ element i.e. **40**

Q1 will be $(15+36)/2 = \mathbf{25.5}$

Q3 will be $(42+43)/2 = \mathbf{42.5}$



BITS Pilani
Pilani Campus



Measures of variability: Ungrouped Data

Range




- The range is exactly the distance between the smallest data point (min) and the largest one (Max).
- Here are the number of hours that 9 students spend on social media on a typical day:
11 6 7 5 2 8 11 12 15

Ques. Range for above case is ?

2 5 6 7 8 11 11 12 15

Range is $15 - 2 = 13$

Inter-Quartile Range (IQR)

- While the range quantifies the variability by looking at the range covered by *ALL* the data, 
- The IQR measures the variability of a distribution by giving us the range covered by the *MIDDLE* 50% of the data.
- The middle 50% of the data falls between Q1 and Q3, and therefore: $IQR = Q3 - Q1$
- The IQR should be used as a measure of spread of a distribution only when the median is used as a measure of center.

Example



Ordered Data Set: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

✓ Q2 will be $(n+1)/2 = 12/2 = 6^{\text{th}}$ element i.e. **40**

✓ Q1 will be $(15+36)/2 = \mathbf{25.5}$

✓ Q3 will be $(42+43)/2 = \mathbf{42.5}$

Ques. IQR range will be?

$$Q3 - Q1 = 42.5 - 25.5 = 17$$

Using the IQR to Detect Outliers



- The IQR is used as the basis for a rule of thumb for identifying outliers.
- An observation is considered a suspected outlier if it is:
 - below $Q1 - 1.5(\text{IQR})$ or
 - above $Q3 + 1.5(\text{IQR})$

Boxplot



- The boxplot graphically represents the distribution of a quantitative variable by visually displaying the five-number summary and any observation that was classified as a suspected outlier using the $1.5(IQR)$ criterion.

Steps



- The central box spans from Q1 to Q3.
- A line in the box marks the median
- Lines go from the edges of the box to the smallest and largest observations that are not classified as outliers.

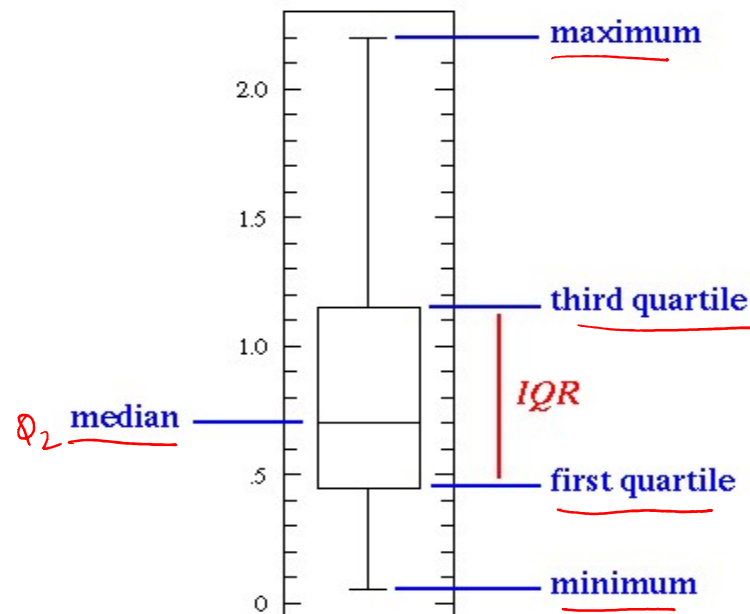
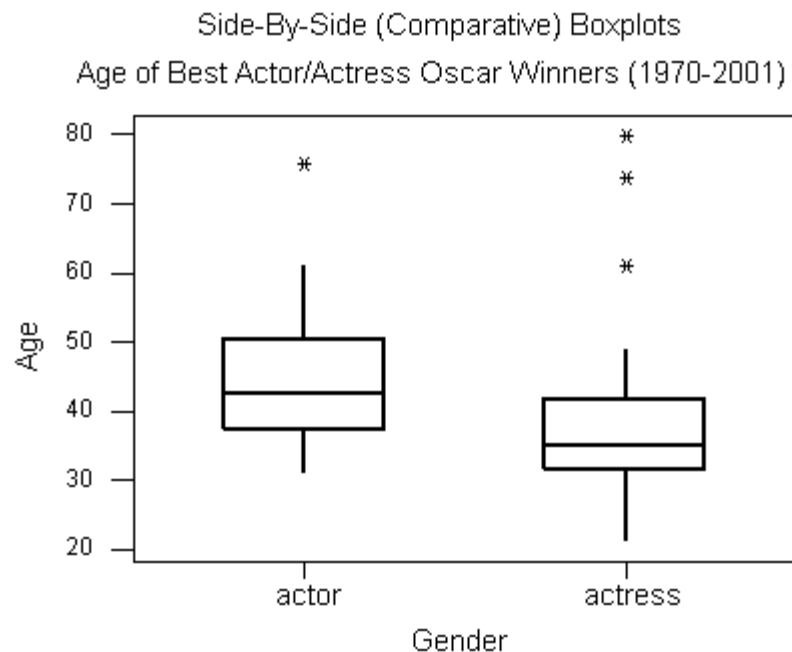


Image: google

Side by side Box plot

- Boxplots are most useful when presented side-by-side for comparing and contrasting distributions from two or more groups.



- Actors: min = 31, Q1 = 37.25, M = 42.5, Q3 = 50.25, Max = 76
- Actresses: min = 21, Q1 = 32, M = 35, Q3 = 41.5, Max = 80

Variance



- Its is the average of squared deviations about the arithmetic mean for a set of numbers

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} \quad \text{Population Variance}$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \quad \text{Sample Variance}$$

degree of freedom

Standard Deviation



- The idea behind the standard deviation is to quantify the spread of a distribution by measuring how far the observations are from their mean.
- The standard deviation gives the average (or typical distance) between a data point and the mean.
- It is the square root of variance

Example



Here are the number of hours that 9 students spend on social media on a typical day:

11 6 7 5 2 8 11 12 15

Find the standard deviation.

Mean = 8.55

- Deviations from the mean

11-8.55, 6-8.55, 7-8.55, 5-8.55, 2-8.55, 8-8.55, 11-8.55, 12-8.55, 15-8.55

2.45, -2.55, -1.55, -3.55, -6.55, -0.55, 2.45, 3.45, 6.45

Example



11 6 7 5 2 8 11 12 15

Square of each of the deviation

6.0025, 6.5025, 2.4025, 12.6025, 42.9025, 0.3025, 6.0025,
11.9025, 41.6025

- Average the square deviations by adding them up, and dividing by $n - 1$

16.277

This average of the squared deviations is called the variance of the data.

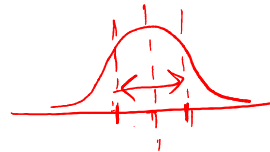
- The SD of the data is the square root of the variance

4.034

The Empirical Rule



Consider a symmetric mound-shaped distribution, the following rule applies:



- Approximately 68% of the observations fall within 1 standard deviation of the mean.
- Approximately 95% of the observations fall within 2 standard deviations of the mean.
- Approximately 99.7% of the observations fall within 3 standard deviations of the mean.

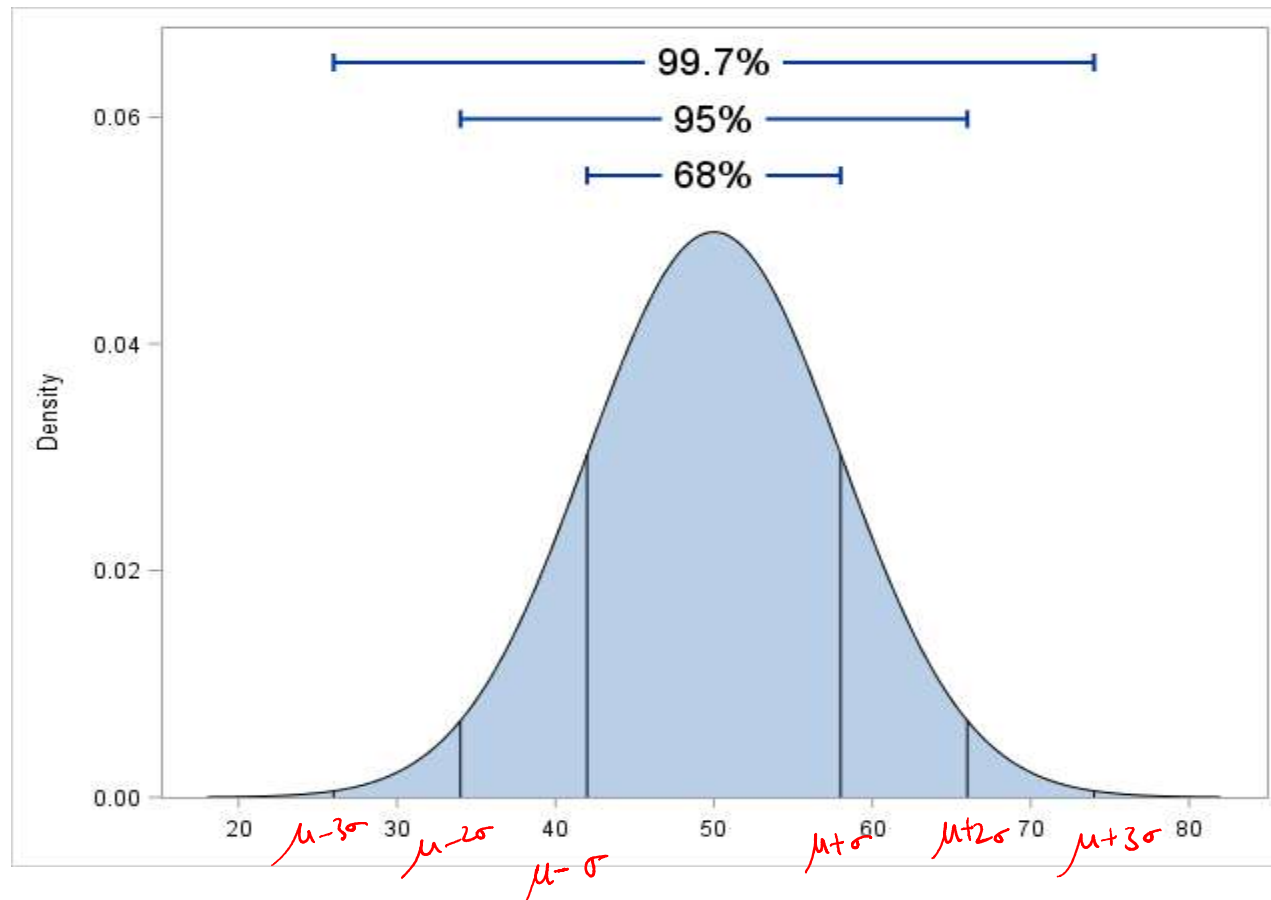


Image: google

Coefficient of variation

- It is the ratio of the standard deviation to the mean expressed in percentage and denoted by CV

CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$



BITS Pilani
Pilani Campus



Measures of central tendency: Grouped Data

Mean



Here M_i represents class mid-point

$$\mu_{\text{grouped}} = \frac{\Sigma fM}{N} = \frac{\Sigma fM}{\Sigma f} = \frac{f_1M_1 + f_2M_2 + \dots + f_iM_i}{f_1 + f_2 + \dots + f_i}$$

where

i = the number of classes

f = class frequency

N = total frequencies

10 - 20

20 - 30

30 - 40

Median



$$\text{Median} = \left(L + \frac{\frac{N}{2} - cf_p}{f_{med}} \right) (W)$$

where:

L = the lower limit of the median class interval

cf_p = a cumulative total of the frequencies up to but not including the frequency of the median class

f_{med} = the frequency of the median class

W = the width of the median class interval

N = total number of frequencies

10 - 25	10
25 - 50	5
50 - 75	5
75 - 100	5
100 - 125	4

Example

(HW)



- Frequency Distribution of 60 Years of Unemployment Data for Canada (Grouped Data)

Class Interval	Frequency	Cumulative Frequency
1-under 3	4	4
3-under 5	12	16
5-under 7	13	29
7-under 9	19	48
9-under 11	7	55
11-under 13	5	60

Mode



- The mode for grouped data is the class midpoint of the modal class. The modal class is the class interval with the greatest frequency.

Ques. What will be the mode for the previous example data?

Class Interval	Frequency	Cumulative Frequency
1-under 3	4	4
3-under 5	12	16
5-under 7	13	29
<u>7-under 9</u>	<u>19</u>	48
9-under 11	7	55
11-under 13	5	60



BITS Pilani
Pilani Campus



Measures of variability: Grouped Data

Population Variance and Standard deviation



FORMULAS FOR POPULATION VARIANCE AND STANDARD DEVIATION OF GROUPED DATA

Original Formula	Computational Version
$\sigma^2 = \frac{\sum f(\underline{M} - \mu)^2}{N}$ $\sigma = \sqrt{\sigma^2}$	$\sigma^2 = \frac{\sum fM^2 - \frac{(\sum fM)^2}{N}}{N}$

where:

f = frequency

M = class midpoint

$N = \sum f$, or total frequencies of the population

μ = grouped mean for the population

Ques. Calculate Variance and Standard deviation for previous example (HW)

Sample Variance and Standard deviation



FORMULAS FOR SAMPLE VARIANCE AND STANDARD DEVIATION OF GROUPED DATA

Original Formula	Computational Version
$s^2 = \frac{\sum f(M - \bar{x})^2}{n - 1}$ $s = \sqrt{s^2}$	$s^2 = \frac{\sum fM^2 - \frac{(\sum fM)^2}{n}}{n - 1}$

where:

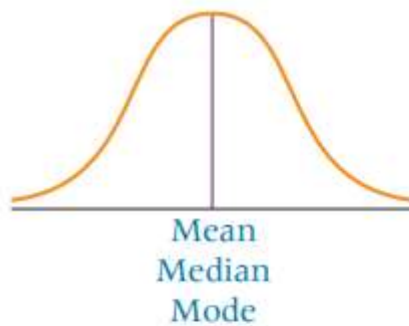
f = frequency

M = class midpoint

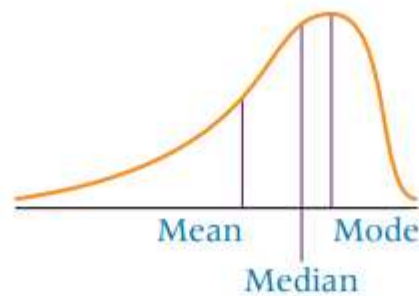
$n = \sum f$, or total of the frequencies of the sample

\bar{x} = grouped mean for the sample

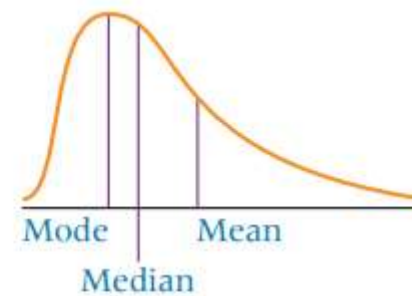
Relation between mean, median and mode



(a)
Symmetric distribution
(no skewness)



(b)
Negatively
skewed



(c)
Positively
skewed

Questions



References

- Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning
- Applied Business Statistics, Ken Black
- <http://www2.isye.gatech.edu/~jeffwu/presentations/datasience.pdf>
- <https://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/>
- <https://link.springer.com/article/10.1007/s41060-018-0102-5>
- <https://link.springer.com/article/10.1007/s42081-018-0009-3#Sec2>