# Introduction

- Data: Categorical / Continuous
- Variables
- Uncertainty/Variations
- Random Variables
- Causes of Variations
  - Chance causes/Natural causes
  - Assignable causes
- Data filtration- Outliers, Leverage points
- Statistics: Science of modeling random variables
- Business Statistics: Data analysis for business application

# Example:

The following sample data set lists the number of minutes of 50 Internet subscribers spent on the Internet during their most recent session.

50 40 41 17 11 7 22 44 28 21 19 23 37 51 54 42 86

41 78 56 72 56 17 7 69 30 80 56 29 33 46 31 39 20

18 29 34 59 73 77 36 39 30 62 54 67 39 31 53 44

# Frequency Distribution

- Grouping of data into ***mutually exclusive*** categories (**classes** or **intervals**) showing the number of observations in each class (**frequency, *f* )**.
- Constructing a Frequency Distribution
  - Find the number of classes (k): *at least 5-6*
  - Find the class width (i)
    - Should be the same for all classes, and
    - lowest value (L) and highest (H) of observation values
    - i ≥ (H-L)/k
  - Round up to the next convenient number.
  - With help of tally find the frequency.

# Solution:

Class width

19 - 7 = 12

| Class | Tally | Frequency, *f* |
|---|---|---|
| 7 – 18 | 𝍸 I | 6 |
| 19 – 30 | 𝍸 𝍸 | 10 |
| 31 – 42 | 𝍸 𝍸 III | 13 |
| 43 – 54 | 𝍸 III | 8 |
| 55 – 66 | 𝍸 | 5 |
| 67 – 78 | 𝍸 I | 6 |
| 79 – 90 | II | 2 |
| | | $\Sigma f = 50$ |

Lower class limits

Upper class limits

**Midpoint of a class =**

$$\frac{(\text{Lower class limit}) + (\text{Upper class limit})}{2}$$

| Class | Midpoint | Frequency, *f* |
|---|---|---|
| 7 – 18 | $\frac{7+18}{2} = 12.5$ | 6 |
| 19 – 30 | $\frac{19+30}{2} = 24.5$ | 10 |
| 31 – 42 | $\frac{31+42}{2} = 36.5$ | 13 |

Class width = 12

# Relative Frequency of a class

- Portion or percentage of the data that falls in a particular class.

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{Sample size}} = \frac{f}{n}$$

$$\boxed{\sum \frac{f}{n} = 1}$$

| Class | Frequency, *f* | Relative Frequency |
|---|---|---|
| 7 – 18 | 6 | $\frac{6}{50} = 0.12$ |
| 19 – 30 | 10 | $\frac{10}{50} = 0.20$ |
| 31 – 42 | 13 | $\frac{13}{50} = 0.26$ |

# Cumulative frequency of a class

The sum of the frequency for that class and all previous classes.

| Class | Frequency, *f* | Cumulative frequency |
|---|---|---|
| 7 – 18 | 6 | 6 |
| 19 – 30 | + 10 | 16 |
| 31 – 42 | + 13 | 29 |

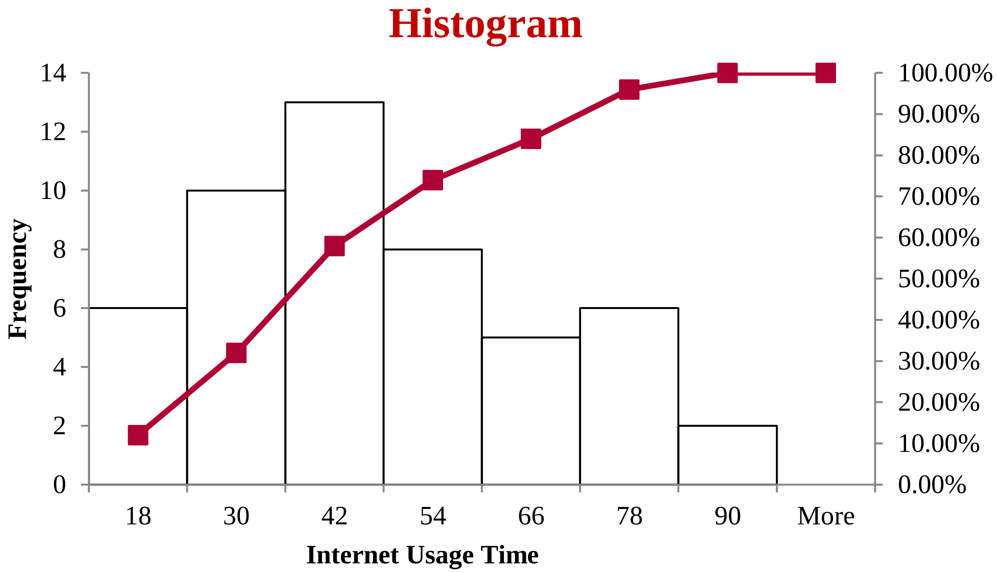
## Class boundaries

- The numbers that separate classes without forming gaps between them. Mostly applicable for continuous variables
- The distance from the upper limit of the first class to the lower limit of the second class is 19 – 18 = 1.
- Half this distance is 0.5.
- First class lower boundary = 7 – 0.5 = 6.5
- First class upper boundary = 18 + 0.5 = 18.5

| Class | Class boundaries | Midpoint | Frequency, f |
|---|---|---|---|
| 7 – 18 | 6.5 – 18.5 | 12.5 | 6 |
| 19 – 30 | 18.5 – 30.5 | 24.5 | 10 |
| 31 – 42 | 30.5 – 42.5 | 36.5 | 13 |
| 43 – 54 | 42.5 – 54.5 | 48.5 | 8 |
| 55 – 66 | 54.5 – 66.5 | 60.5 | 5 |
| 67 – 78 | 66.5 – 78.5 | 72.5 | 6 |
| 79 – 90 | 78.5 – 90.5 | 84.5 | 2 |

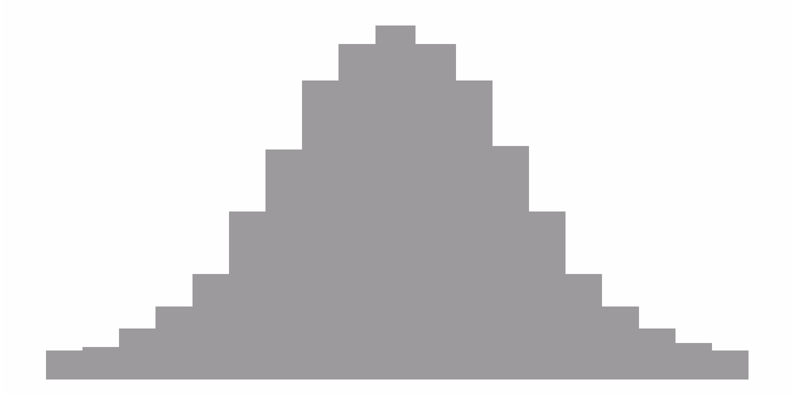
# Frequency Distribution

| Class | Frequency, f | Midpoint x | Relative frequency | Cumulative frequency | Cumulative Relative frequency |
|---|---|---|---|---|---|
| 7 – 18 | 6 | 12.5 | 0.12 | 6 | 0.12 |
| 19 – 30 | 10 | 24.5 | 0.20 | 16 | 0.32 |
| 31 – 42 | 13 | 36.5 | 0.26 | 29 | 0.58 |
| 43 – 54 | 8 | 48.5 | 0.16 | 37 | 0.74 |
| 55 – 66 | 5 | 60.5 | 0.10 | 42 | 0.84 |
| 67 – 78 | 6 | 72.5 | 0.12 | 48 | 0.96 |
| 79 – 90 | 2 | 84.5 | 0.04 | 50 | 1.00 |



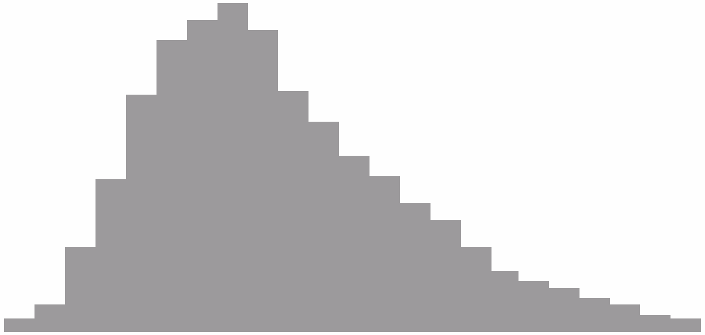
# Typical Histogram Shapes and What They Mean

**Normal.** A common pattern is the bell-shaped curve known as the "normal distribution." In a normal distribution, points are as likely to occur on one side of the average as on the other.



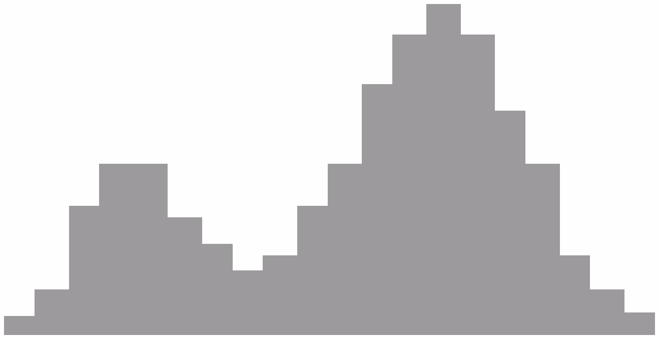
Normal distribution

- **Skewed.** The skewed distribution is asymmetrical because a natural limit prevents outcomes on one side.
- The distribution’s peak is off center toward the limit and a tail stretches away from it.
- For example, a distribution of purity of a product would be skewed, because the product cannot be more than 100 percent pure. Other examples of natural limits are holes that cannot be smaller than the diameter of the drill bit or call-handling times that cannot be less than zero. These distributions are called right- or left-skewed according to the direction of the tail.



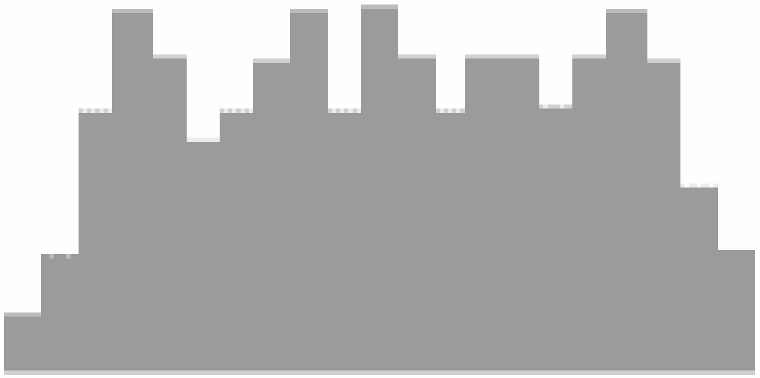
Right-skewed distribution

- **Double-peaked or bimodal.** The bimodal distribution looks like the back of a two-humped camel.
- The outcomes of two processes with different distributions are combined in one set of data.
- For example, a distribution of production data from a two-shift operation might be bimodal, if each shift produces a different distribution of results. Stratification often reveals this problem.



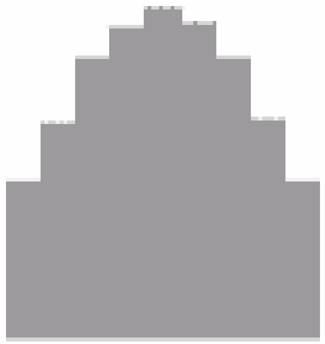
Bimodal (double-peaked) distribution

- **Plateau.** The plateau might be called a “multimodal distribution.”
- Several processes with normal distributions are combined. Because there are many peaks close together, the top of the distribution resembles a plateau.



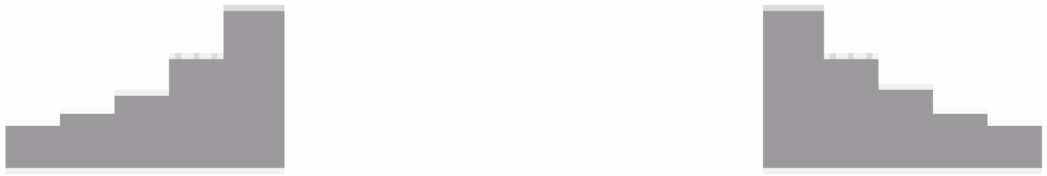
Plateau distribution

- **Truncated.** The truncated distribution looks like a normal distribution with the tails cut off.
- The supplier might be producing a normal distribution of material and then relying on inspection to separate what is within specification limits from what is out of spec. The resulting shipments to the customer from inside the specifications are the truncated.



Truncated or heart-cut distribution

- **Dog food.** The dog food distribution is missing something—results near the average.
- If a customer receives this kind of distribution, someone else is receiving a heart cut, and the customer is left with the “dog food,” the odds and ends left over after the master’s meal. Even though what the customer receives is within specifications, the product falls into two clusters: one near the upper specification limit and one near the lower specification limit. This variation often causes problems in the customer’s process.



Dog food distribution

# Characterizing a distribution

- Data set have a tendency to lie around a particular point
  - Measures of central tendency
- How much is the spread of the data set?
  - Measures of dispersion
- Symmetry of the distribution: Skewness
- Peakedness of the distribution: Kurtosis

# Measures of Central Tendency

- A measures of central of tendency may be defined as single expression of a group of data

- There are two main objectives for the study of measures of central tendency:
  - To get one single value that represent the entire data
  - To facilitate comparison

# Different Measures of Central Tendency

- Mean:
  - Arithmetic Mean
  - Weighted Mean
  - Geometric Mean
  - Harmonic Mean
- Median
- Mode

# Arithmetic Mean

- The arithmetic mean is the sum of a set of all observations, positive, negative or zero, divided by the number of observations. If we have "n" real numbers $x_1, x_2, x_3, \ldots\ldots., x_n$.

- Arithmetic mean

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots\ldots\ldots\ldots + x_n}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

# Arithmetic Mean of Group Data

- if $x_1, x_2, x_3, \ldots\ldots\ldots, x_k$ are the mid-values and $f_1, f_2, f_3, \ldots\ldots\ldots, f_k$ are the corresponding frequencies, where the subscript '*k*' stands for the number of classes, then the mean is

$$\overline{x} = \frac{\sum f_i x_i}{\sum f_i}$$

# Example: Find the Mean of a Frequency Distribution

| Class | Midpoint, $x$ | Frequency, $f$ | $(x \cdot f)$ |
|---|---|---|---|
| 7 – 18 | 12.5 | 6 | 12.5*6 = 75.0 |
| 19 – 30 | 24.5 | 10 | 24.5*10 = 245.0 |
| 31 – 42 | 36.5 | 13 | 36.5*13 = 474.5 |
| 43 – 54 | 48.5 | 8 | 48.5*8 = 388.0 |
| 55 – 66 | 60.5 | 5 | 60.5*5 = 302.5 |
| 67 – 78 | 72.5 | 6 | 72.5*6 = 435.0 |
| 79 – 90 | 84.5 | 2 | 84.5*2 = 169.0 |
| | | n = 50 | $\Sigma(x \cdot f) = 2089.0$ |

$$\bar{x} = \frac{\Sigma(x \cdot f)}{n} = \frac{2089}{50} \approx 41.8 \text{ minutes}$$

# Weighted Mean

- The Weighted mean of the positive real numbers $x_1, x_2, \ldots, x_n$ with their weight $w_1, w_2, \ldots, w_n$ is defined to be

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

# Example: Finding a Weighted Mean

| **Source** | **Score, *x*** | **Weight, *w*** | ***x·w*** |
|---|---|---|---|
| Test Mean | 86 | 0.50 | 86(0.50)= 43.0 |
| Midterm | 96 | 0.15 | 96(0.15) = 14.4 |
| Final Exam | 82 | 0.20 | 82(0.20) = 16.4 |
| Computer Lab | 98 | 0.10 | 98(0.10) = 9.8 |
| Homework | 100 | 0.05 | 100(0.05) = 5.0 |
|  |  | $\Sigma w = 1$ | $\Sigma(x\cdot w) = 88.6$ |

$$\bar{x} = \frac{\Sigma(x\cdot w)}{\Sigma w} = \frac{88.6}{1} = 88.6$$

Your weighted mean performance for the course is 88.6.

# Median

- The implication of this definition is that a median is the middle value of the observations such that the number of observations above it is equal to the number of observations below it.

**If "n" is odd**

$$M_e = X_{\frac{1}{2}(n+1)}$$

**If "n" is Even**

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

# Median of Group Data

$$\textbf{Median} = \textbf{l} + \frac{\textbf{(N/2 – F) * i}}{\textbf{f}}$$

Where,

l = lower limit of Median class

N= Total frequency (total number of observations)

F = Cumulative frequency of the class just preceding to median class

f = Frequency of Median class

i = Size of the class interval

| Class boundaries | Midpoint, x | Frequency, f | Cumulative frequency |
|---|---|---|---|
| 6.5 – 18.5 | 12.5 | 6 | 6 |
| 18.5 – 30.5 | 24.5 | 10 | 16 |
| 30.5 – 42.5 | 36.5 | 13 | 29 |
| 42.5 – 54.5 | 48.5 | 8 | 37 |
| 54.5 – 66.5 | 60.5 | 5 | 42 |
| 66.5 – 78.5 | 72.5 | 6 | 48 |
| 78.5 – 90.5 | 84.5 | 2 | 50 |

N/2 = 50/2 = 25

Median Class: 30.5 – 42.5

Md = 30.5 + (25-16)*12/13 = 38.806

# Example of Median

| CI | f | Cumulative Frequency | |
|---|---|---|---|
| 30.5 – 40.5 | 6 | 6 | |
| 40.5– 50.5 | 8 | 14 | |
| 50.5 – 60.5 | 10 | 24 | N/2=20 |
| 60.5 – 70.5 | 6 | 30 | |
| 70.5 – 80.5 | 4 | 34 | |
| 80.5 – 90.5 | 3 | 37 | |
| 90.5 – 100.5 | 3 | 40 | |

**N=40**

$$\textbf{Median} = l + \frac{(N/2 - F)}{f} * i$$

Here, l = 50.5, F = 14, f = 10, i = 10

$$\text{Median} = 50.5 + \frac{(20 - 14)}{10} * 10$$

$$= 50.5 + 6$$

**Median = 56.5**

# QUARTILES

- The values which divide the given data in to four equal parts when observations are arranged in order.

  obviously there will be three quartiles Q1,Q2 & Q3.

Q1(1st quartile):25%below &75%above

Q2(2nd quartile): same as median 50% above & below

Q3(3rd quartile):75%below &25% above

To calculate nth quartile :

$$\mathbf{l + \frac{(nN/4 - F)}{f} * i}$$

# Example of Q1

| CI | f | Cumulative Frequency |
|---|---|---|
| 30.5 – 40.5 | 6 | 6 |
| 40.5– 50.5 | 8 | 14 |
| 50.5 – 60.5 | 10 | 24 |
| 60.5 – 70.5 | 6 | 30 |
| 70.5 – 80.5 | 4 | 34 |
| 80.5 – 90.5 | 3 | 37 |
| 90.5 – 100.5 | 3 | 40 |

**N/4=10**

**N=40**

$$\mathbf{Q1 = l + \frac{(N/2 - F) * i}{f}}$$

Here, l = 40.5, F = 6, f = 8, i = 10

$$Q1 = 40.5 + \frac{(10 - 6) * 10}{8}$$

$$= 40.5 + 5$$

**Q1 = 45.5**

# QUINTILES & DECILES

- Quintiles : It contains four points so it will divide data in to five equal parts.

- Deciles : it contain 9 points & it will divide data in to ten equal parts.
- To calculate nth deciles:

$$\mathbf{l + \frac{(nN/10 - F)}{f} * i}$$

# Percentile

- Points divide the data set into 100 equal parts of total frequency.

To calculate nth percentile :

$$l + \frac{(nN/100 - F)}{f} * i$$

# Mode

- Mode is the value of a distribution for which the frequency is maximum. In other words, mode is the value of a variable, which occurs with the highest frequency.

- So the mode of the list (1, 2, 2, 3, 3, 3, 4) is 3. The mode is not necessarily well defined.

# Mode of Group Data

$$M_0 = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} i$$

- $L_1$ = Lower boundary of modal class
- $\Delta_1$ = difference of frequency between
  modal class and previous class
- $\Delta_2$ = difference of frequency between
  modal class and following class
- i = class interval

| Class boundaries | Frequency, f |
|---|---|
| 6.5 – 18.5 | 6 |
| 18.5 – 30.5 | 10 |
| 30.5 – 42.5 | 13 |
| 42.5 – 54.5 | 8 |
| 54.5 – 66.5 | 5 |
| 66.5 – 78.5 | 6 |
| 78.5 – 90.5 | 2 |

- Modal Class: 30.5-42.5
- $L_1$ = 30.5
- $\Delta_1$ = 13 – 10 = 3
- $\Delta_2$ = 13 – 8 = 5
- i = 12

$$M_0 = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} i$$

$$= 30.5 + \frac{3 \times 12}{3 + 5} = 35$$

# Example of Mode

| CI | f |
|---|---|
| 30.5 – 40.5 | 6 |
| 40.5– 50.5 | 8 |
| 50.5 – 60.5 | 10 → **Modal Class** |
| 60.5 – 70.5 | 6 |
| 70.5 – 80.5 | 4 |
| 80.5 – 90.5 | 3 |
| 90.5 – 100.5 | 3 |
|  | **N=40** |

- $L_1 = 50.5$
- $\Delta_1 = 10 – 8 = 2$
- $\Delta_2 = 10 – 6 = 4$
- $i = 10$

$$M_0 = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} i$$

$$= 50.5 + \frac{2 \times 10}{2 + 4} = 53.833$$

# Geometric Mean

- Geometric mean is defined as the positive root of the product of observations. Symbolically,

$$G = (x_1 x_2 x_3 \cdots x_n)^{1/n}$$

- It is also often used for a set of numbers whose values are meant to be multiplied together or are exponential in nature, such as data on the growth of the human population or ratios.
- Cannot be used with numbers of value 0 or negative.

# Geometric mean of Group data

- If the "*n*" non-zero and positive

  and values $x_1, x_2, ..., x_n$ occur $f_1, f_2, ..., f_n$ times, respectively, then the geometric mean of the set of observations is defined by:

$$G = \left[ x_1^{\ f_1} \ x_2^{\ f_2} \cdots x_n^{\ f_n} \right]^{\frac{1}{N}} = \left[ \prod_{i=1}^{n} x_i^{\ f_i} \right]^{\frac{1}{N}}$$

Where

$$N = \sum_{i=1}^{n} f_i$$

# Harmonic Mean

- **Harmonic mean** (formerly sometimes called the **subcontrary mean**) is one of several kinds of average.

- Typically, it is appropriate for situations when the average of rates is desired. The harmonic mean is the number of variables divided by the sum of the reciprocals of the variables. Useful for rates such as speed (=distance/time) etc.

# Harmonic Mean Group Data

- The harmonic mean *H* of the positive real numbers $x_1, x_2, \ldots, x_n$ is defined to be

Ungroup Data

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Group Data

$$H = \frac{n}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$