LOG718 - Preparatory Course in Basic Statistics

Tassew Tolcha

August 2019

Outlines

Lecture 1: Descriptive statistics

Lecture 2 : Probabilities

Lecture 3 : Random variables and probability distributions

Lecture 4 : Sampling and sampling distributions

Lecture 5 : Statistical inference

Textbook:

Johnson, Richard A., and Gouri K. Bhattacharyya. ***Statistics : Principles and Methods***. 8th Ed. Emea ed. Hoboken, NJ: Wiley, 2019.

For further information, use the following link:

https://himolde.instructure.com/courses/1180/modules

Introduction (1)

What is *Statistics*?

- Statistics is a discipline of ***data-based reasoning.***

> **Statistics** as a subject provides a body of principles and methodology for *designing the process* of data collection, *summarizing* and *interpreting* the data, and *drawing conclusions* or generalities.

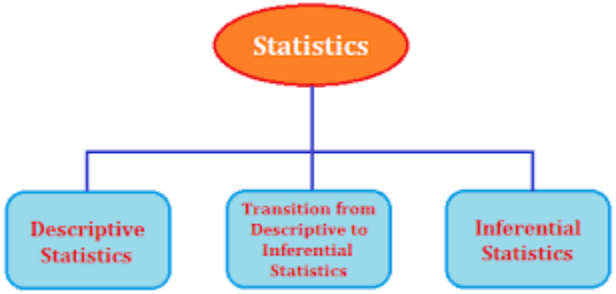
- The principles and methodology of statistics are useful in answering questions such as,
  - What *kind* and *how much* data need to be collected?
  - How should we *organize* and *interpret* the data?
  - How can we *analyze* the data and *draw* conclusions?

Introduction (2)

Statistical methods

- Statistical methods are mathematical *formulas*, *models*, and *techniques* that are used in statistical analysis.

1. **Descriptive statistics –** summarize and describe the *prominent features of data*.
   - Tables and graphs,
   - Measurement of central tendency and location,
   - Measurement of variability/variation.



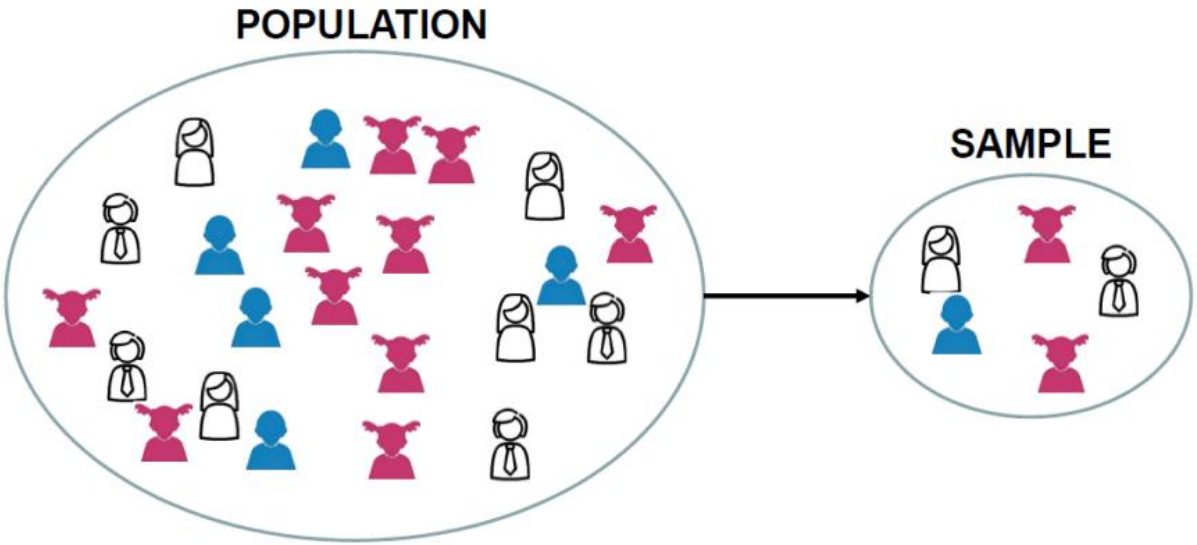
2. **Inferential statistics -** *evaluation of information* present in data and the *assessment of the new learning* gained from this information.
   - Estimations and forecasts
   - Testing

Introduction (3)

Population and Sample

Population is the c*omplete set of all* items that interest an investigator. It represents the *target of an investigation*.

Sample is an *observed subset* (or portion) of a population. It constitutes a part of a far larger collection about which we wish to make inferences.



Descriptive Statistics

- ✓ **Main types of data**
- Describing data by tables and graphs
- Measures of center
- Measures of variation

Main types of data (1)

- Mainly *two* basic types

1. **Qualitative** or **categorical** data - When the characteristic under study concerns a *qualitative trait* that is only classified in categories and *not numerically measured*.

   - *Nominal*- if there is *no natural order* between the categories.
     eg Eye colour - blue, green, brown etc)
     Gender – Male, Female

   - *Ordinal* - if an ordering exists
     eg exam results – pass or fail
     socio-economic status - low, middle or high
     Product quality – poor, average, good

Main types of data (2)

2. Numerical, quantitative or **measurement** data - the characteristic is measured on a numerical scale and the resulting data consist of a set of numbers.

- *Discrete* (often, integer) - if the measurements are integers, the scale is made up of distinct numbers with gaps in between.

  eg. number of people in a household, count of traffic fatalities, number of students in the class, etc

- *Continuous* - the measurements can take on *any value* within the interval.

  eg. height, weight, time to run a race, the temperature, distance, etc

- ✓ **Main types of data**
- ✓ **Describing data by tables and graphs**
- Measures of center
- Measures of variation

Describing data by *tables* and *graphs*

- Describing *categorical* data
- Describing *quantitative* data
  - Describing discrete data,
  - Describing continuous data.

Describing *categorical* data (1)

- *Categorical* data can be described using;
  - Relative frequency,
  - Pie chart,
  - Pareto diagram.
- Categorical data are readily organized in the form of a *frequency table* that shows the *counts (**frequencies**)* of the individual categories.
- The understanding of the data is further enhanced by calculation of the *proportion* (also called ***relative frequency***) of observations in each category.

$$\text{Relative frequency of a category} = \frac{\text{Frequency in the category}}{\text{Total number of observations}}$$

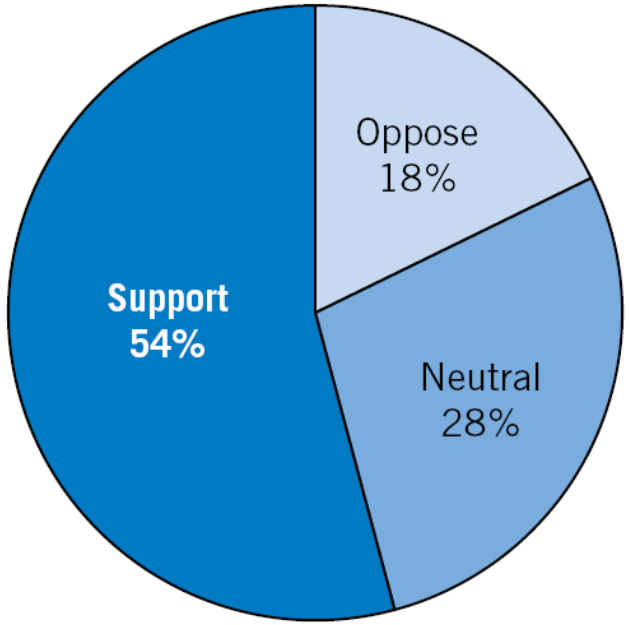
Describing *categorical* data (2)

Example

- A campus press polled a sample of 280 undergraduate students in order to study student attitude toward a proposed change in the dormitory regulations. The numbers were 152 support, 77 neutral, and 51 opposed.

Summary Results of an Opinion Poll

Responses	Frequency	Relative Frequency
Support	152	$\frac{152}{280} = .543$
Neutral	77	$\frac{77}{280} = .275$
Oppose	51	$\frac{51}{280} = .182$
Total	280	1.000

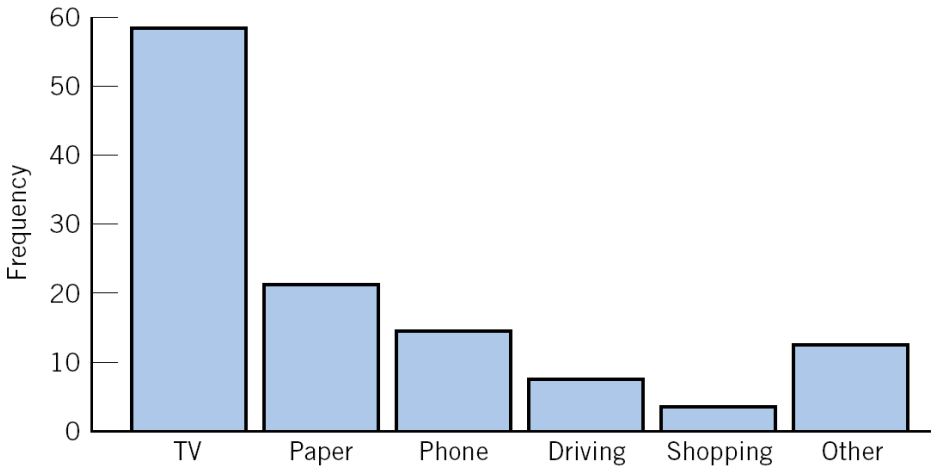


To obtain the angle for any category, we multiply the relative frequency by 360 degrees.

Describing *categorical* data (3)

- **Pareto diagram:** is a powerful graphical technique for displaying events according to their *frequency*.
- According to ***Pareto's empirical law***, any collection of events consists of only a few that are major in that they are the ones that occur most of the time.
  - Pareto's empirical law sometimes **referred as 80-20 rule**. This rule was noted by Italian Economist (Vilfredo Pareto) that in most cases a *small number of factors* are *responsible for most of the problem*.
- The following example shows the r/ship between nail biting and types of activity compiled by some graduate students.

Frequency	Activity
58	Watching television
21	Reading newspaper
14	Talking on phone
7	Driving a car
3	Grocery shopping
12	Other



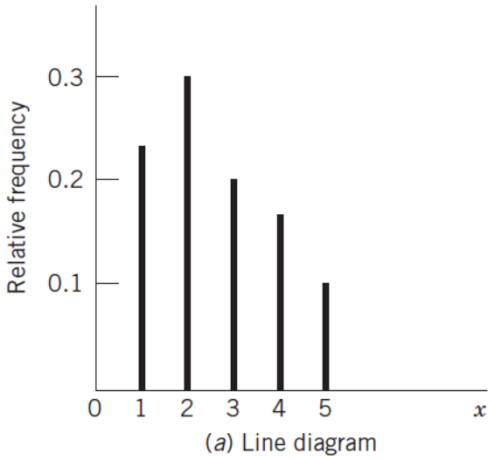
Pareto diagram for nail biting example

Describing discrete data

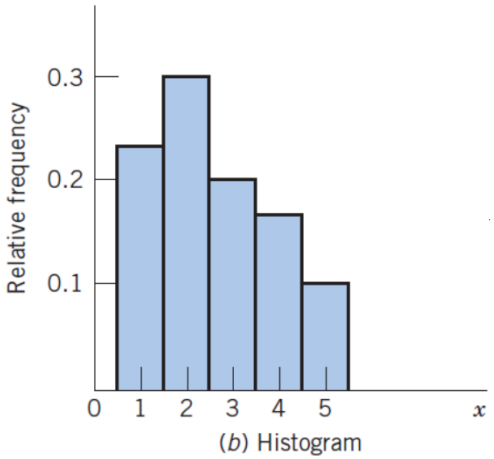
- A **discrete data set** can be summarized/described using;
  - Frequency table,
  - Line diagram,
  - Histogram
- **Example:** The following table shows a sample of 30 people who returned items to Y retail store on December 26 and December 27.

Number of items returned

1	4	3	2	3	4	5	1	2	1
2	5	1	4	2	1	3	2	4	1
2	3	2	3	2	1	4	3	2	5



(a) Line diagram



(b) Histogram

Frequency Distribution for Number (x) of Items Returned

Value x	Frequency	Relative Frequency
1	7	.233
2	9	.300
3	6	.200
4	5	.167
5	3	.100
Total	30	1.000

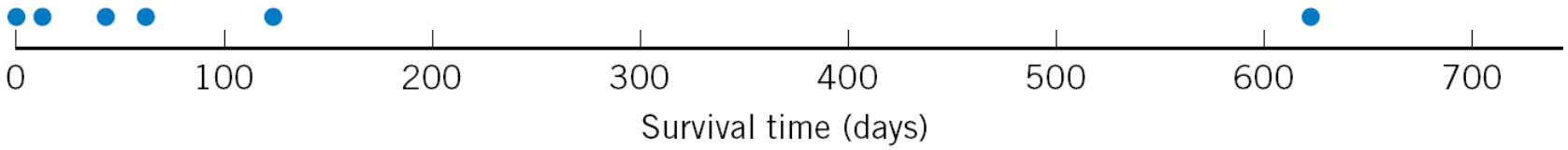
Note: unlike Pareto charts, the bars of a histograms do touch each other.

Describing continuous data (1)

- The appropriate tabular and graphical presentations of continuous data sets include;
  - *Dot diagram* – used for relatively few observations (say, less than 20 or 25)
  - *Frequency distribution* on intervals
  - *Histogram* – used with a larger number of observations
  - *Stem-and-Leaf* Display
  - *Scatter plots*

Example - Dot diagram –

The number of days the first six heart transplant patients at Stanford survived after their operations were 15, 3, 46, 623, 126, 64.



Describing continuous data (2)

Frequency distribution on intervals

Constructing a Frequency Distribution for a Continuous Variable

1. Find the minimum and the maximum values in the data set.
2. Choose intervals or cells of equal length that cover the range between the minimum and the maximum without overlapping. These are called **class intervals,** and their endpoints **class boundaries.**
3. Count the number of observations in the data that belong to each class interval. The count in each class is the **class frequency** or **cell frequency.**
4. Calculate the **relative frequency** of each class by dividing the class frequency by the total number of observations in the data:

$$\text{Relative frequency} = \frac{\text{Class frequency}}{\text{Total number of observations}}$$

Frequency Distribution for Hours of Sleep Data (left endpoints included but right endpoints excluded)

Class Interval	Frequency	Relative Frequency
4.3–5.5	5	$\frac{5}{59} = .085$
5.5–6.7	15	$\frac{15}{59} = .254$
6.7–7.9	20	$\frac{20}{59} = .339$
7.9–9.1	16	$\frac{16}{59} = .271$
9.1–10.3	3	$\frac{3}{59} = .051$
Total	59	1.000

Note: we choose class interval of length ***1.2*** hours.

Example: Students require different amounts of sleep (a sample of 59 students).

4.5	4.7	5.0	5.0	5.3	5.5	5.5	5.7	5.7	5.7
6.0	6.0	6.0	6.0	6.3	6.3	6.3	6.5	6.5	6.5
6.7	6.7	6.7	6.7	7.0	7.0	7.0	7.0	7.3	7.3
7.3	7.3	7.5	7.5	7.5	7.5	7.7	7.7	7.7	7.7
8.0	8.0	8.0	8.0	8.3	8.3	8.3	8.5	8.5	8.5
8.5	8.7	8.7	9.0	9.0	9.0	9.3	9.3	10.0	

Describing continuous data (3)

Histogram:

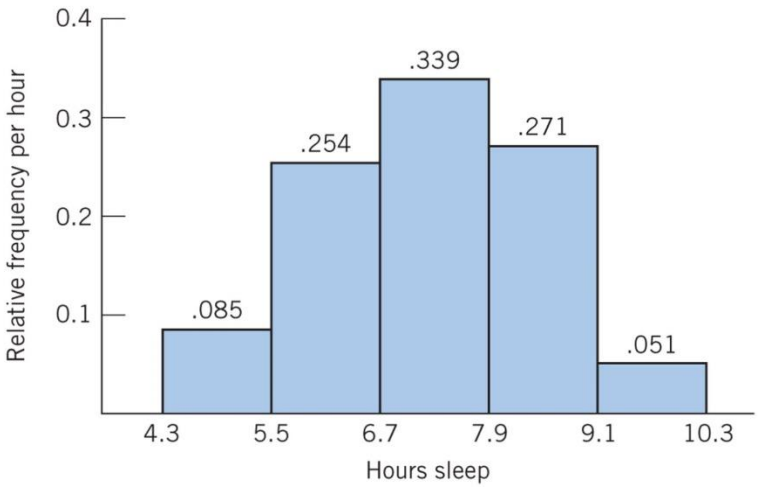
- A frequency distribution can be graphically presented as a histogram.
- Mark the class intervals on the horizontal axis.
- A vertical rectangle represents the proportion of the observations occurring in that class interval.
- To create rectangles whose area is equal to relative frequency, use the rule

$$\text{Height} = \frac{\text{Relative Frequency}}{\text{Width of interval}}$$

The total area of a histogram is 1.

Example:

Histogram of the sleep for 59 students



Describing continuous data (4)

Stem-and-Leaf display

- A **stem-and-leaf display** provides a more efficient variant of the histogram for displaying data, especially when the observations are *two-digit* numbers.
- To make this display:
  - List the digits 0 through 9 in a column and draw a vertical line. These correspond to the *leading digit*.
  - For each observation, *record its second digit* to the right of this vertical line in the row where the first digit appears.
  - Finally, arrange the second digits in each row so they are in *increasing order*.

Example: examination scores of 50 students

75	98	42	75	84	87	65	59	63
86	78	37	99	66	90	79	80	89
68	57	95	55	79	88	76	60	77
49	92	83	71	78	53	81	77	58
93	85	70	62	80	74	69	90	62
84	64	73	48	72				

Stem-and-Leaf Display for the Examination Scores

0	
1	
2	
3	7
4	289
5	35789
6	022345689
7	01234556778899
8	00134456789
9	0023589

- ✓ **Main types of data**
- ✓ **Describing data by tables and graphs**
- ✓ **Measures of center**
- Measures of variation

Measures of center (1)

- The distribution of a sample measurements locate the position of a central/location value about which the measurements are distributed.

- The commonly used *indicators of center*:
  - Mean
  - Median

- The common indicators of *location/position*:
  - Percentiles
  - Quartiles

Measures of center (2)

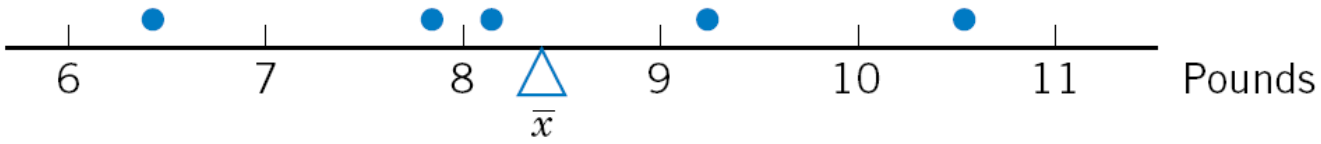
- **Mean**: is the sum of the *data values* divided by the *number of observations*.

The **sample mean** of a set of n measurements $x_1, x_2, \ldots, x_n$ is the sum of these measurements divided by n. The sample mean is denoted by $\bar{x}$.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{or} \quad \frac{\sum x_i}{n}$$

- **Example:** The birth weights in pounds of five babies born one day in the same hospital are 9.2, 6.4, 10.5, 8.1, and 7.8.

$$\bar{x} = \frac{9.2 + 6.4 + 10.5 + 8.1 + 7.8}{5} = \frac{42}{5} = 8.4\ pounds$$



Measures of center (3)

The **sample median** of a set of n measurement $x_1,\ldots, x_n$ is the *middle value* when the measurements are arranged from smallest to largest.

- If the sample size, *n*, is an *odd number*, the median is the *middle observation*.
- If the sample size, *n*, is an *even number*, the median is the *average* of the two middle observations.
- The median will be the number located in the,

$$\mathbf{0.5(n+1)^{th}}\ \textbf{ordered position}$$

Example: Find the median of the birth-weight data given in mean example. The measurements, ordered from smallest to largest, are:

6.4 7.8 **8.1** 9.2 10.5

Measures of center (4)

Mean and/or median?

Example: The number of days the first six heart transplant patients at Stanford survived after their operations were 15, 3, 46, 623, 126, 64.

Mean: $\bar{x} = \frac{3+15+46+64+126+623}{5} = \frac{877}{6} = \mathbf{146.2}\ days$

Median: 3 15 46 64 126 623

$$median = \frac{46+64}{2} = \mathbf{55}\ days$$

- ✓ Median is not affected by a few outliers (small or very large),
- ✓ Outliers have a considerable effect on the mean,
- ✓ For extremely asymmetrical distributions, the median is a likely to be a more sensible measure of center than the mean.

Measures of center (5)

Percentile and Quartiles

- Percentiles and quartiles measures indicate the *location*, or *position*, of a value relative to the entire set of data.
- The sample 100 p^{th} percentile is a value such that after the data are ordered from smallest to largest, at least 100p% of the observations are *at or below* that number and at least 100(1-p)% are *at or above* this value.
- Percentile separate large ordered data sets into 100^{th} . The 50^{th} percentile is the median.

Calculating the Sample 100*p*-th Percentile

1. Order the data from smallest to largest.
2. Determine the product (*sample size*) × (*proportion*) = *np*.

If *np* is not an integer, round it up to the next integer and find the corresponding ordered value.

If *np* is an integer, say *k*, calculate the average of the *k*th and (*k* + 1)st ordered values.

Measures of center (6)

- Quartiles are simply the 25^{th} , 50^{th} and 75^{th} percentiles.

Sample Quartiles

Lower (first) quartile	Q_1 = 25th percentile
Second quartile (or median)	Q_2 = 50th percentile
Upper (third) quartile	Q_3 = 75th percentile

$$Q_1 = \textit{the value in the } 0.25(n+1)^{th} \textit{ ordered position}$$
$$Q_2 = \textit{the value in the } 0.50(n+1)^{th} \textit{ ordered position}$$
$$Q_3 = \textit{the value in the } 0.75(n+1)^{th} \textit{ ordered position}$$

Five number summary;

The five number summary refers to the five descriptive measures: minimum, first quarter, median, third quarter, and maximum.

$$Minimum < Q_1 < median < Q_3 < maximum$$

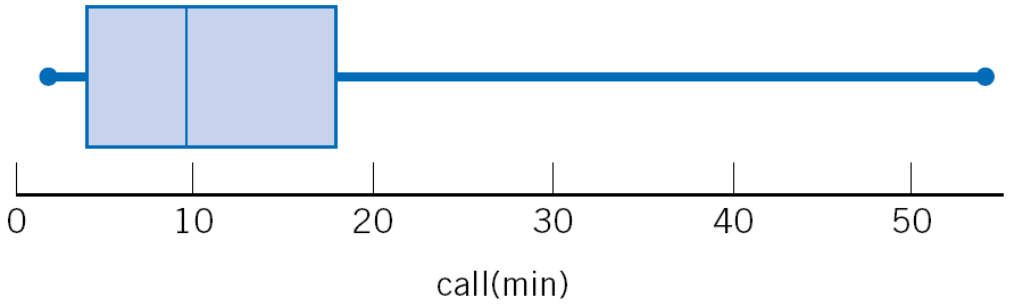
Measures of center (7)

Example - The lengths of long-distance phone calls in minutes (38 calls)

1.6	1.7	1.8	1.8	1.9	2.1	2.5	3.0	3.0	4.4
4.5	4.5	5.9	7.1	7.4	7.5	7.7	8.6	9.3	9.5
12.7	15.3	15.5	15.9	15.9	16.1	16.5	17.3	17.5	19.0
19.4	22.5	23.5	24.0	31.7	32.8	43.5	53.3		

- The 90th percentile $= 0.9(38+1) = 35.1$ $i.e$, 35th ordered observation $=$ $\mathbf{31.7\ minutes}$
- $Q_1 = 0.25(38+1)^{th}\ ordered\ position = 9.75$ i. e, 10th ordered observation, $\boldsymbol{Q_1 = 4.4}$ **minutes**
- $Q_2 = 0.50(38+1)^{th}\ ordered\ position = 19.5$ i.e, 20th ordered observation, $\boldsymbol{Q_2 = 9.5}$ **minutes**
- $Q_3 = 0.75(38+1)^{th}\ ordered\ position = 29.25$ i.e, 29th ordered observation, $\boldsymbol{Q_3 = 17.5}$ **minutes**

Boxplot



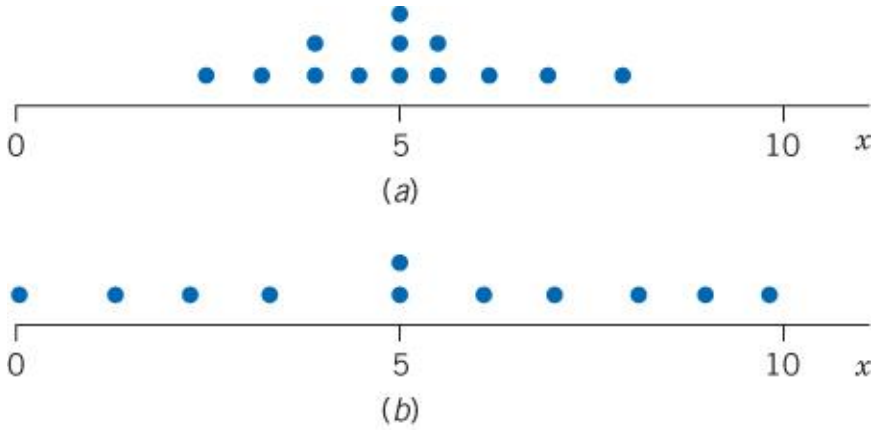
- ✓ **Main types of data**
- ✓ **Describing data by tables and graphs**
- ✓ **Measures of center**
- ✓ **Measures of variation**

Measures of variation

- Measures of variation indicates the extent of variation around the center.
- Variation could be measured by;
  - Variance and standard deviation,
  - Z-score,
  - Range and interquartile range.

Measures of variation (1)

- Measures of variation numerically measure the *extent of variation around the center*.
- Two data sets may exhibit *similar positions* of center may be remarkably different with respect to variability



Measures of variation (2)

Sample variance

- The variation of the individual data points about measurement of center could be reflected in their deviation from the mean.

$$\boldsymbol{Deviation = Observation - sample\ mean = x - \bar{x}}$$

- But the total deviation is zero

$$\sum(\text{Deviations}) = \sum(x_i - \bar{x}) = 0$$

Example:

Observation x	Deviation $x - \bar{x}$
3	−3
5	−1
7	1
7	1
8	2

- To obtain a measure of spread, *we must eliminate the signs of deviations* before averaging by squaring the numbers, ***variance*.**

Measures of variation (3)

Sample variance

Sample variance of n observations:

$$s^2 = \frac{\text{sum of squared deviations}}{n - 1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Example: Calculate the sample variance of the data 3 5 7 7 8.

	Observation x	Deviation $x - \bar{x}$	(Deviation)^2 $(x - \bar{x})^2$
	3	−3	9
	5	−1	1
	7	1	1
	7	1	1
	8	2	4
Total	30	0	16
	$\sum x$	$\sum (x - \bar{x})$	$\sum (x - \bar{x})^2$

$$\bar{x} = \frac{30}{5} = 6$$

$$\text{Sample variance } s^2 = \frac{16}{5 - 1} = 4$$

Measures of variation (4)

Sample standard deviation:

- The variance involves a sum of squares, its unit is the square of the unit in which the measurements are expressed.
- To obtain a measure of variability *in the same unit as the data*, we take the positive square root of the variance, called the **sample standard deviation.**
- The standard deviation rather than the variance serves as a *basic measure of variability*.

Sample Standard Deviation

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

*An **alternative formula** for the sample variance is*

$$s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]$$

Example: Calculate the standard deviation for the data 3 5 7 7 8.

We already calculated the variance $s^2 = 4$ so the standard deviation is $\boldsymbol{s = \sqrt{4} = 2}$

Measures of variation (5)

- **Sample Z-score (standard scale)**: measures the *position of a value relative* to the sample mean in units of standard deviation.

$$\textbf{z-score of measurement} = \frac{Measurement - \bar{x}}{s}$$

Example

Loins typically have babies in twos and threes but sometimes four or five. To protect the very young, the mother will take the babies away from the pride for the first 6 weeks. The size of eight litters born to one pride of lions are: 3 5 3 3 2 3 3 1.

a. Find sample mean, variance and standard deviation

b. Calculate z-score for a liter of size 2.

Solution

$$\bar{x} = \frac{3+5+3+3+2+3+3+1}{8} = 2.88$$

Using alternative formula, $s^2 = \frac{(3^2+5^2+3^2+3^2+2^2+3^2+3^2+1^2)/8}{8-1} = \mathbf{1.268}$

$$\boldsymbol{s} = \sqrt{1.268} = \mathbf{1.126} \text{ cubs}$$

Z-score for the value 2 is $(2 - 2.88)/1.125 = -0.78$*, so it is* $\mathbf{0.78}$ *standard deviation below the sample mean of cubs.*

Measures of variation (6)

Note:1

Z-score > 0, the value is greater than mean,

Z-score < 0, the value is less than mean,

Z-score $= 0$, the value is equal to mean.

Note:2

- For **bell-shaped distributions**, an empirical rule relates the standard deviation to the proportion of the data that lie in an interval around $\bar{x}$.

Empirical Guidelines for Symmetric Bell-Shaped Distributions

Approximately	68%	of the data lie within $\bar{x} \pm s$
	95%	of the data lie within $\bar{x} \pm 2s$
	99.7%	of the data lie within $\bar{x} \pm 3s$

Example: Examine the 59 hours of sleep from earlier example.

$\bar{x} = 7.18,\ \ s = 1.28,\ \ 2s = 2(1.28) = 2.56$, interval $= 7.18 \pm 1.28 = 5.90\ to\ 8.46$ which contains 40 observations, $40/59 = 67.8\%$.

Interval $= 7.18 \pm 2.56 = 4.62\ to\ 9.74$ which contains 57 observations, $57/59 = 96.6\%$.

Measures of variation (7)

Range

$$\text{Sample range} = \text{Largest observation} - \text{Smallest observation}$$

- The range gives the length of the interval spanned by the observations.

Example: Calculate the range for the hours of sleep data (earlier example).

$$Smallest\ observation = \mathbf{4.5}$$
$$Largest\ observation = \mathbf{10.0}$$
$$Sample\ range = 10.0 - 4.5 = \mathbf{5.5\ hours}$$

- Two attractive features of range
  a. Simple to compute and interpret
  b. Too sensitive to the existence of outliers

Measures of variation (8)

Interquartile range

Sample interquartile range = Third quartile − First quartile

- Interquartile range represents the length of the interval covered by the center half of the observations.
- Interquartile range is not disturbed by outliers (if a small fraction of the observations are very large or very small).

Example: Calculate the sample interquartile range for the length of long distance phone calls data (ealier example),

The quartiles were $Q_1 = 4.4$ and $Q_3 = 17.5$

Interquartile range = $Q_3 - Q_1 = 17.5 - 4.4 = 13.1\ minutes$

Nearly 50% of the middle calls are within an interval of length 13.1 minutes.