

# Data Analytics and Mining

Variables, Frequency Distribution, Central Tendency

Data Analytics and Mining, 2024

Majid Sohrabi

National Research University Higher School of Economics



October 11, 2024

# Variables



# Constructs

- ▶ Some variables, such as height, weight, and eye color are well-defined, concrete entities that **can be observed and measured directly**.
- ▶ Variables like intelligence, anxiety, and hunger are called constructs, and because they are intangible and cannot be directly observed, they are often called **hypothetical constructs**.

# Numerical Variables

- ▶ Some variables, such as height, weight, and eye color are well-defined, concrete entities that **can be observed and measured directly**.
- ▶ Variables like intelligence, anxiety, and hunger are called constructs, and because they are intangible and cannot be directly observed, they are often called **hypothetical constructs**.



When measuring a continuous variable, it should be very rare to obtain identical measurements for two different individuals. Because a continuous variable has an infinite number of possible values, it should be almost impossible for two people to have exactly the same score. -> We use **intervals**.

# Variables' types

- ▶ An **interval scale** consists of ordered categories that are all intervals of exactly the same size. Equal differences between numbers on the scale reflect equal differences in magnitude. However, the zero point on an interval scale is arbitrary and does not indicate a zero amount of the variable being measured.

*E.g. Celsius scale for temperature (0 is arbitrary).*

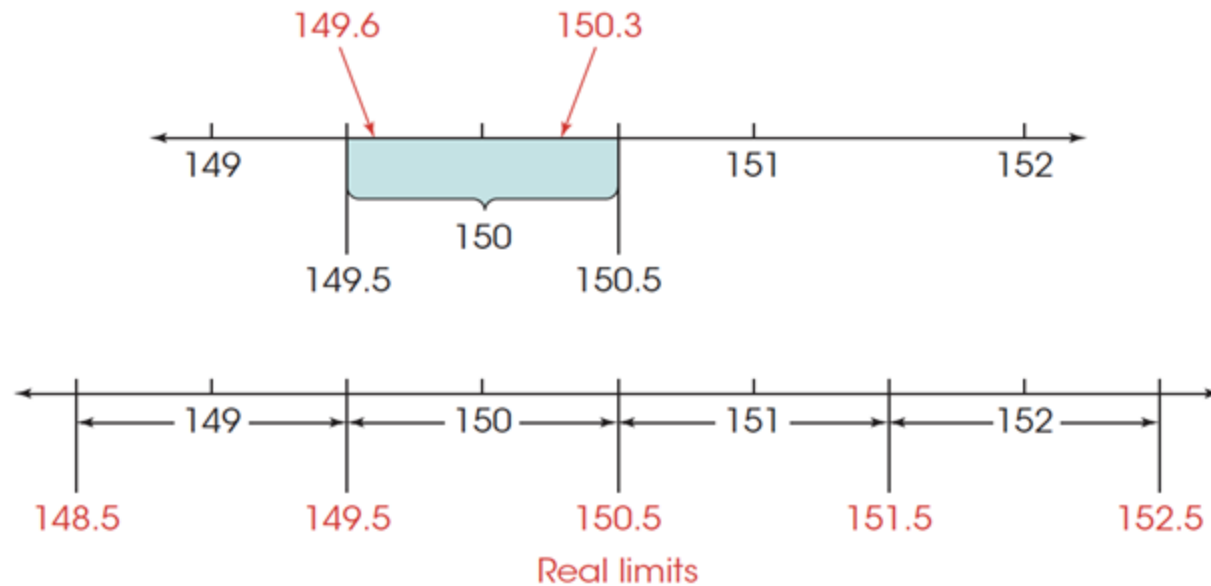
- ▶ A **ratio scale** is an interval scale with the additional feature of an absolute zero point. With a ratio scale, ratios of numbers do reflect ratios of magnitude.

*E.g. weight (0 is the absence of weight).*

*Can compare numbers by ratios:*

*10 kg is twice as heavy as 5 kg*

# Variables' types



**Real limits** are the boundaries of intervals for scores that are represented on a continuous number line. The real limit separating two adjacent scores is located exactly halfway between the scores. Each score has two real limits. The **upper real limit** is at the top of the interval, and the **lower real limit** is at the bottom.

# Categorical Variables

- ▶ A **nominal scale** consists of a set of categories that have different names. Measurements on a nominal scale label and categorize observations, but do not make any quantitative distinctions between observations.
- ▶ An **ordinal scale** consists of a set of categories that are organized in an ordered sequence. Measurements on an ordinal scale rank observations in terms of size or magnitude.



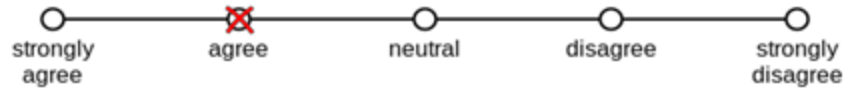
With measurements from an **ordinal scale**, you can determine whether two individuals are different and you can determine the direction of difference. However, ordinal measurements do not allow you to determine the size of the difference between two individuals.

# Variables' types

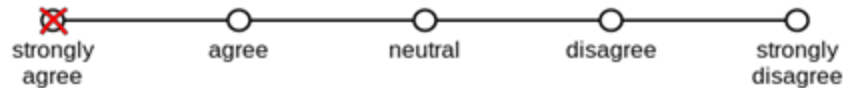
## Likert scale (ordinal scale)

### Website User Survey

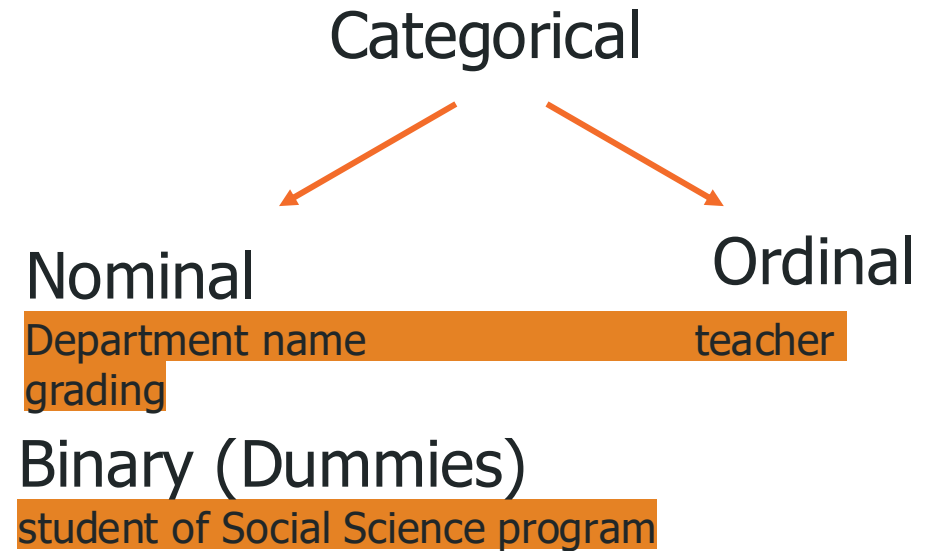
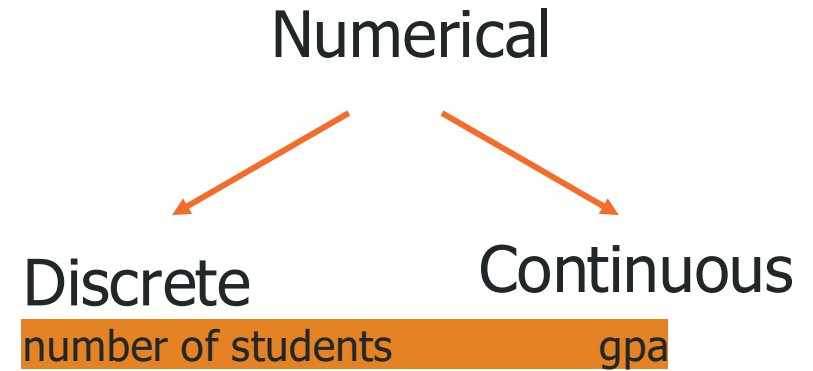
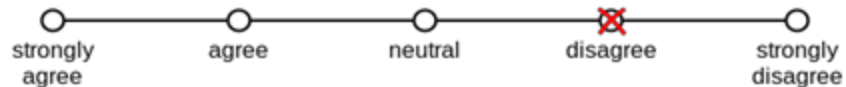
1. The website has a user friendly interface.



2. The website is easy to navigate.



3. The website's pages generally have good images.





# Learning check

- 2.** A researcher studies the factors that determine the number of children that couples decide to have. The variable, number of children, is an example of a \_\_\_\_\_ variable.
- a.** discrete
  - b.** continuous
  - c.** nominal
  - d.** ordinal
- 4.** The teacher in a communications class asks students to identify their favorite reality television show. The different television shows make up a \_\_\_\_\_ scale of measurement.
- a.** nominal
  - b.** ordinal
  - c.** interval
  - d.** ratio


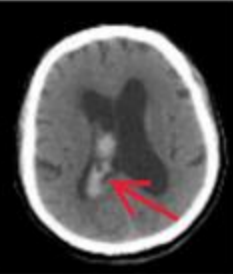
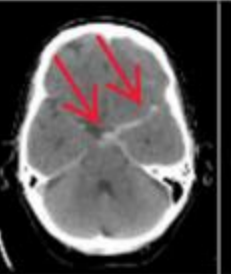
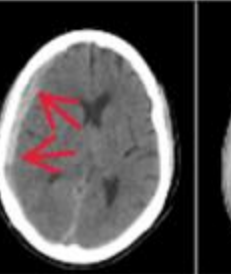
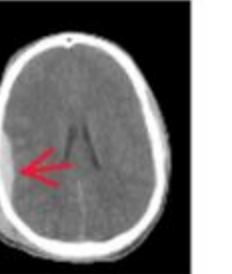
# Variables' types

Variable	Key	Variable	Key
survival	0 = Yes, 1 = No	age	-
pclass	1 = 1st, 2 = 2nd, 3 = 3rd	fare	-
sex	M = male, F = female	cabin	-
embarked	C = Cherbourg, Q = Queenstown, S = Southampton c		

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S

# World of Data

	A	B	C	D	E	F
1	Country ▼	Salesperson ▼	Order Date ▼	OrderID ▼	Units ▼	Order Amount ▼
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60
14	USA	Callahan	10/01/2011	10402	11	2,713.50
15	UK	Rayleigh	13/01/2011	10406	15	1,830.78
16	USA	Callahan	14/01/2011	10408	10	1,622.40
17	USA	Farnham	14/01/2011	10409	19	319.20
18	USA	Farnham	15/01/2011	10410	16	802.00

	Intraparenchymal	Intraventricular	Subarachnoid	Subdural	Epidural
<b>Location</b>	Inside of the brain	Inside of the ventricle	Between the arachnoid and the pia mater	Between the Dura and the arachnoid	Between the dura and the skull
<b>Imaging</b>					
<b>Mechanism</b>	High blood pressure, trauma, arteriovenous malformation, tumor, etc	Can be associated with both intraparenchymal and subarachnoid hemorrhages	Rupture of aneurysms or arteriovenous malformations or trauma	Trauma	Trauma or after surgery
<b>Source</b>	Arterial or venous	Arterial or venous	Predominantly arterial	Venous (bridging veins)	Arterial
<b>Shape</b>	Typically rounded	Conforms to ventricular shape	Tracks along the sulci and fissures	Crescent	Lentiform
<b>Presentation</b>	Acute (sudden onset of headache, nausea, vomiting)	Acute (sudden onset of headache, nausea, vomiting)	Acute (worst headache of life)	May be insidious (worsening headache)	Acute (skull fracture and altered mental status)

# World of Data

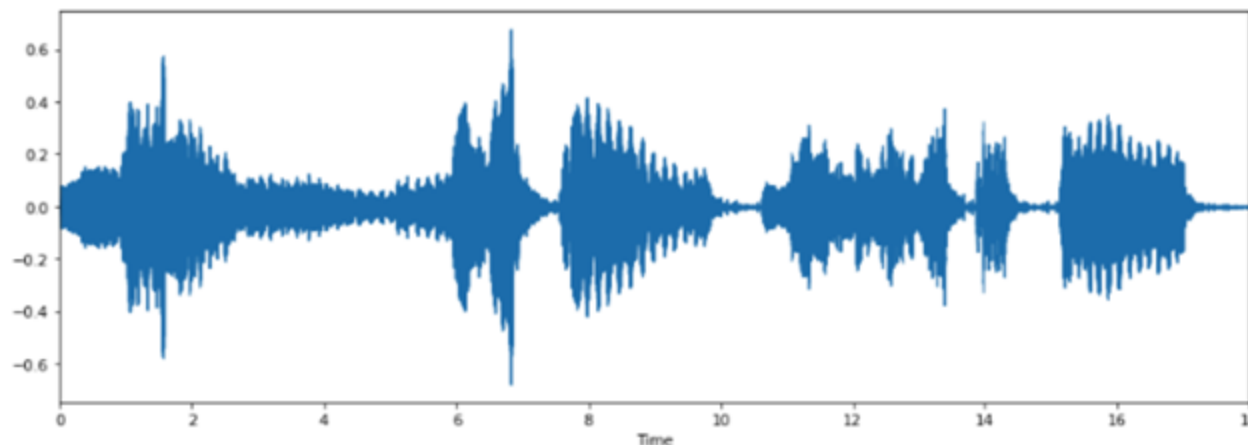
Тональность документа:

позитив 0.12

негатив 0.38

Lenovo Vibe Shot Достоинства: Дизайн, **камера** - супер, удобно держать одной рукой) Недостатки: **Цена** оставляет желать лучшего. В Европе он стоит от 14 до 17 т.р Из Китая его могут прислать от 12 до 16 т.р. в России он от 21000, разница очевидна.... Комментарий: минусы я не буду писать т.к. у каждого смартфона, да и вообще телефона=) будут минусы, если капаться придирчиво. По мне этот смартфон - камерафон. Что по поводу **батареи** она не большая и не маленькая, такая... средняя)) ну это ведь андроид, можно и на 4000mah за день израсходовать, так что смотрите сами)) **камеры** норм) как фронтальная, таки основная)

<https://blog.br-analytics.ru/kak-rabotaet-analiz-tonalnosti-soobshhenij-v-brand-analytics/>



# Distributions



# Statistical notation

## Table:

row – observation  
column – variable

Quiz Scores	Height	Weight
X	X	Y
37	72	165
35	68	151
35	67	160
30	67	160
25	68	146
17	70	160
16	66	133

## Summation notation:

10, 6, 7, 4,  $\Sigma X = 27$  and  $N = 4$ .

X – variable value

N – number of observations in population

\*n – number of observations in sample

***What is the value of  $\Sigma(X - 2)$  for the following scores: 6, 2, 4, 2?***

# Frequency distribution

---

## Visible Tattoo

---

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

---

---

## No Visible Tattoo

---

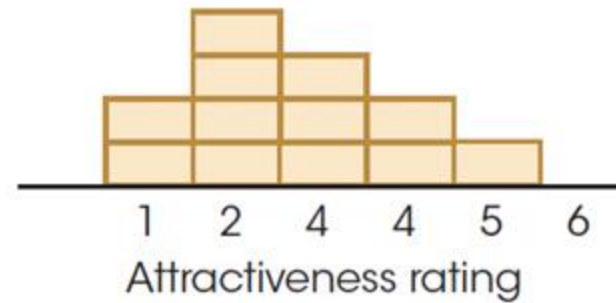
2	4	4	3
5	4	2	4
4	5	3	3

---

# Frequency distribution

## Visible Tattoo

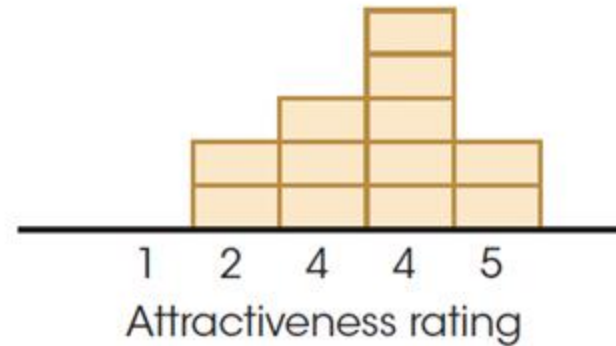
1	2	4	3
2	2	1	3
2	5	4	3



Photograph with visible tattoo

## No Visible Tattoo

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



Photograph with no visible tattoo



# Frequency distribution

- ▶ A **frequency distribution** is an organized tabulation of the number of individuals located in each category on the scale of measurement.

What is the frequency distribution for the following data?

8, 9, 8, 7, 10, 9, 6, 4, 9, 8,  
7, 8, 10, 9, 8, 6, 9, 7, 8, 8

# Frequency distribution

- ▶ A **frequency distribution** is an organized tabulation of the number of individuals located in each category on the scale of measurement.

What is the frequency distribution for the following data?

8, 9, 8, 7, 10, 9, 6, 4, 9, 8,  
7, 8, 10, 9, 8, 6, 9, 7, 8, 8

$X$	$f$
10	2
9	5
8	7
7	3
6	2
5	0
4	1

$$\Sigma f = N$$

# Proportions and Percentage

$X$	$f$
5	1
4	2
3	3
2	3
1	1

# Proportions and Percentage

$X$	$f$	$p = \frac{f}{N}$
5	1	$\frac{1}{10} = 0.10$
4	2	$\frac{2}{10} = 0.20$
3	3	$\frac{3}{10} = 0.30$
2	3	$\frac{3}{10} = 0.30$
1	1	$\frac{1}{10} = 0.10$

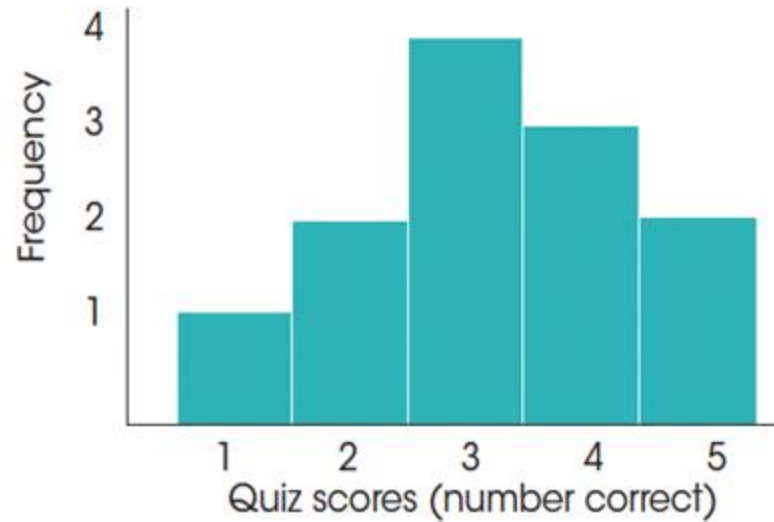
# Proportions and Percentage

$X$	$f$	$p = \frac{f}{N}$	$\% = p(100)$
5	1	$\frac{1}{10} = 0.10$	10%
4	2	$\frac{2}{10} = 0.20$	20%
3	3	$\frac{3}{10} = 0.30$	30%
2	3	$\frac{3}{10} = 0.30$	30%
1	1	$\frac{1}{10} = 0.10$	10%

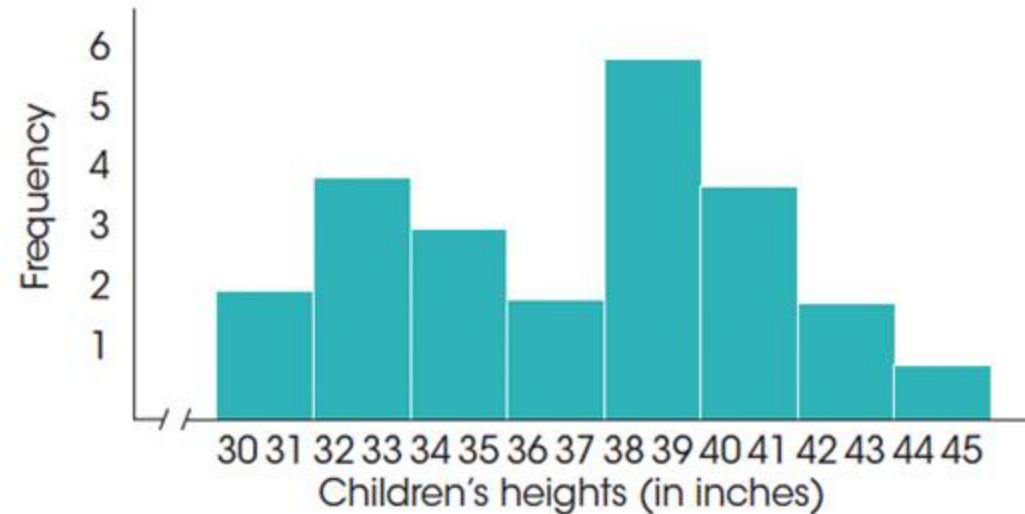
# Grouped Frequency distribution

$X$	$f$
90–94	3
85–89	4
80–84	5
75–79	4
70–74	3
65–69	1
60–64	3
55–59	1
50–54	1

# Frequency distribution graphs: histogram



$X$	$f$
5	2
4	3
3	4
2	2
1	1



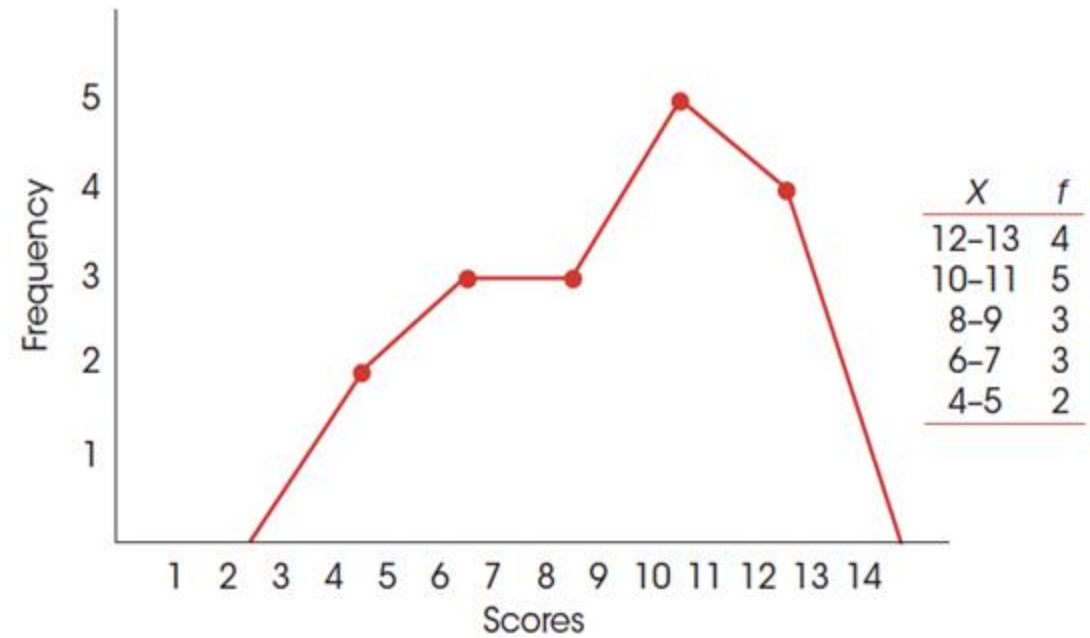
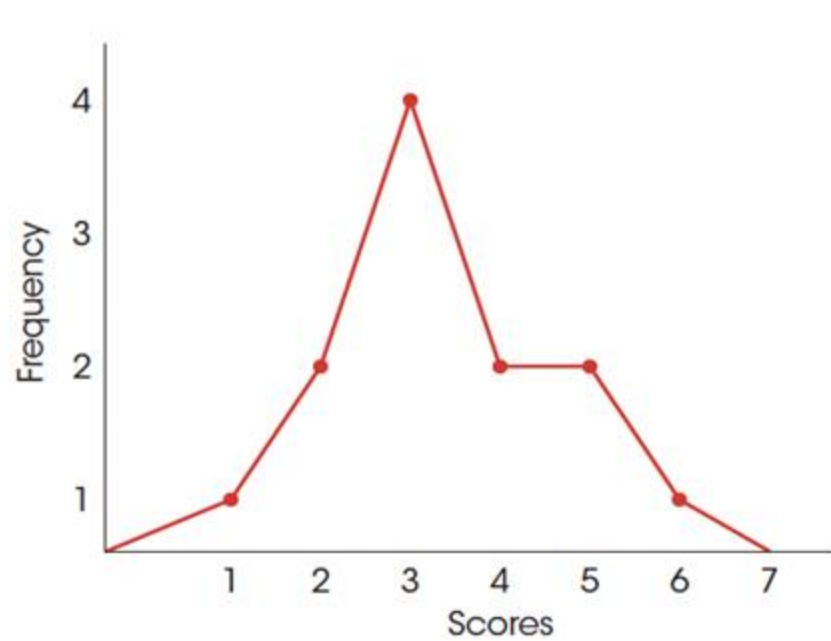
$X$	$f$
44-45	1
42-43	2
40-41	4
38-39	6
36-37	2
34-35	3
32-33	4
30-31	2

# Histogram

- ▶ To construct a histogram, you first list the numerical scores (the categories of measurement) along the X-axis. Then you draw a bar above each X value so that
  - The height of the bar corresponds to the frequency for that category.
  - For continuous variables, the width of the bar extends to the real limits of the category. For discrete variables, each bar extends exactly half the distance to the adjacent category on each side.



# Frequency distribution graphs: polygons



Here dots are placed in the middle of the interval

# Polygons

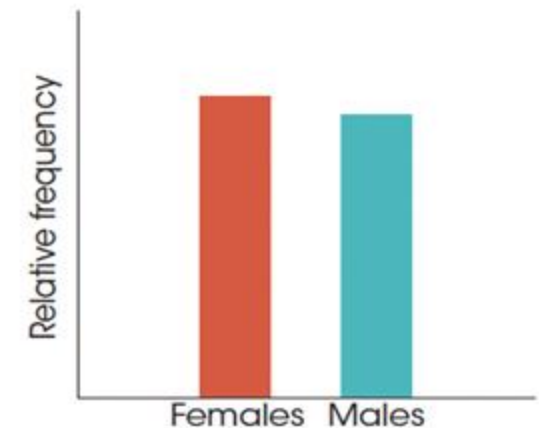
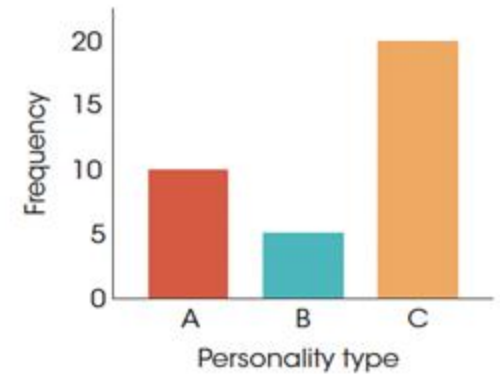
- ▶ To construct a polygon, you begin by listing the numerical scores (the categories of measurement) along the X-axis. Then,
  - ✓ A dot is centered above each score so that the vertical position of the dot corresponds to the frequency for the category.
  - ✓ A continuous line is drawn from dot to dot to connect the series of dots.
  - ✓ The graph is completed by drawing a line down to the X-axis (zero frequency) at each end of the range of scores.

# Bar Graph (Chart)

- ▶ A bar graph is essentially the same as a histogram, except that spaces are left between adjacent bars.
- ▶ For a nominal scale, the space between bars emphasizes that the scale consists of separate, distinct categories. For ordinal scales, separate bars are used because you cannot assume that the categories are all the same size.

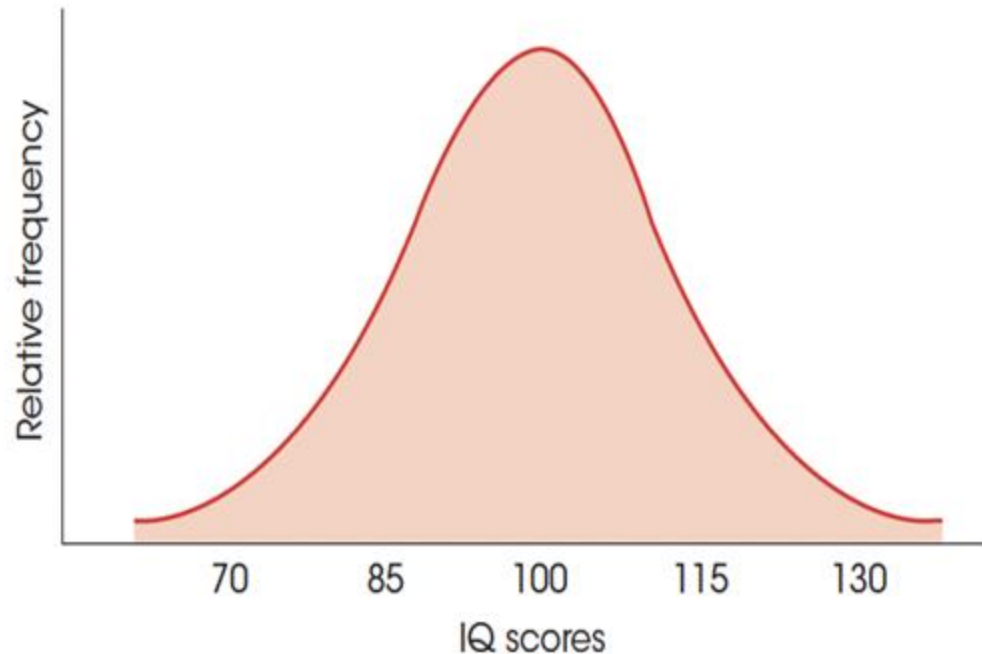
To construct a bar graph, list the categories of measurement along the X-axis and then draw a bar above each category so that the height of the bar corresponds to the frequency for the category.

Frequency distribution graphs:  
bar chart



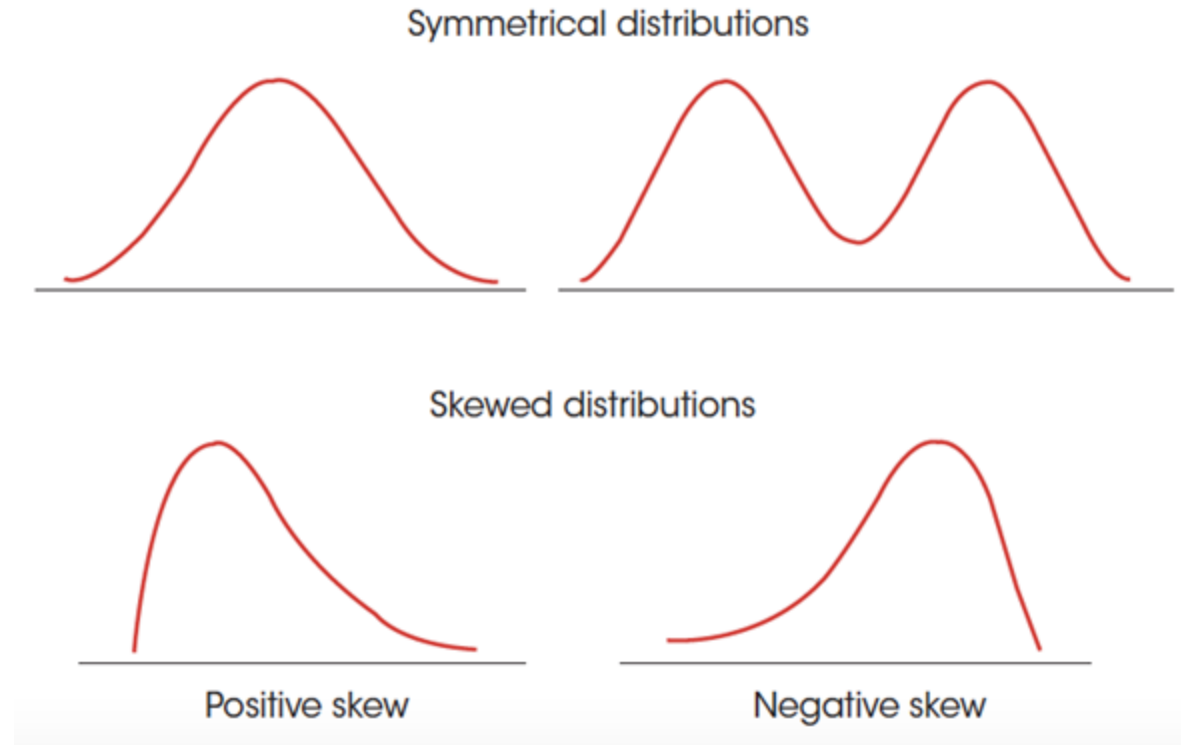
# Smooth curves

When a population consists of numerical scores from an interval or a ratio scale, it is customary to draw the distribution with a smooth curve instead of the jagged, step-wise shapes that occur with histograms and polygons. The smooth curve indicates that you are not connecting a series of dots (real frequencies) but instead are showing the relative changes that occur from one score to the next.

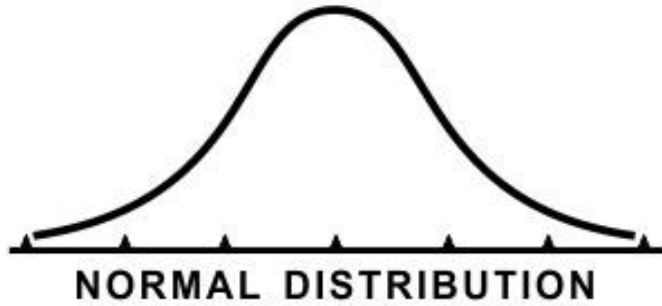


# Shape of a Frequency Distribution

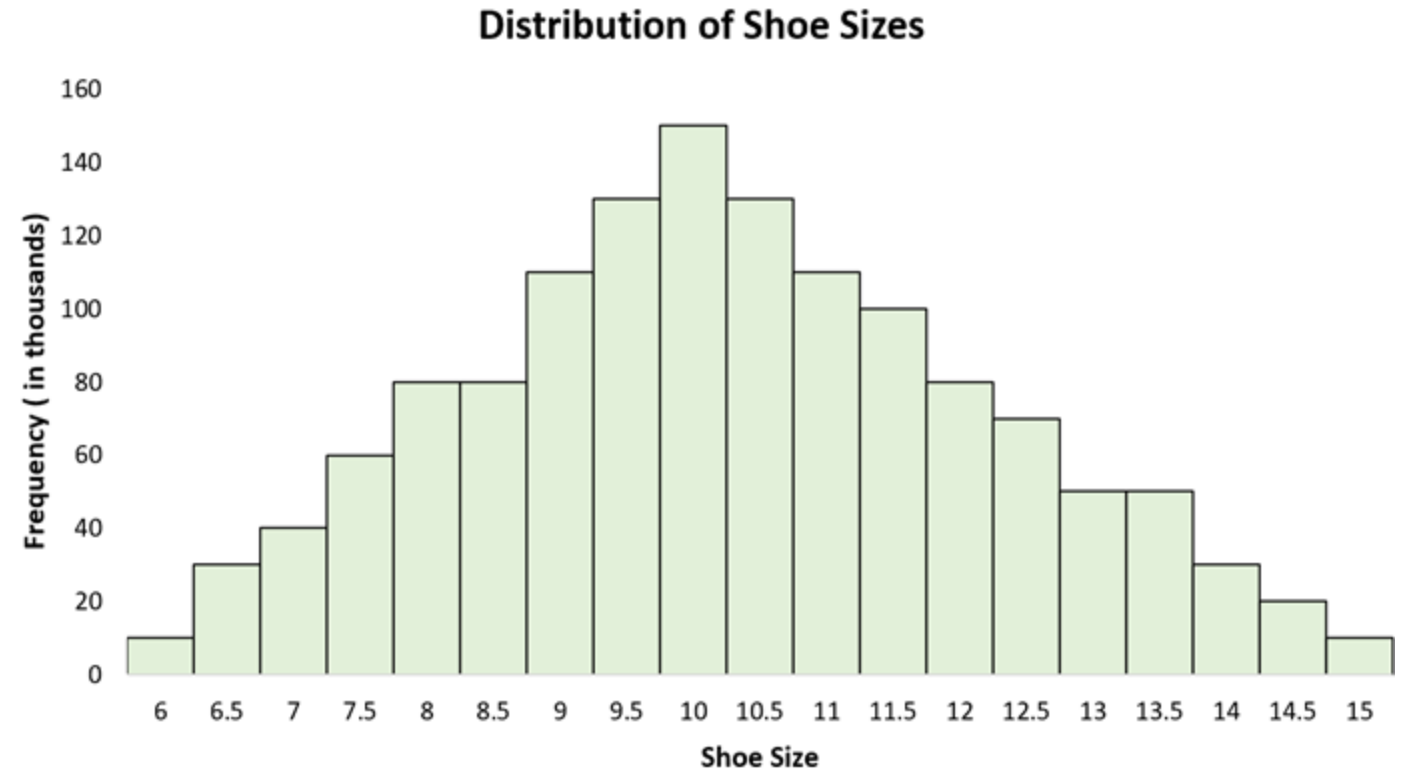
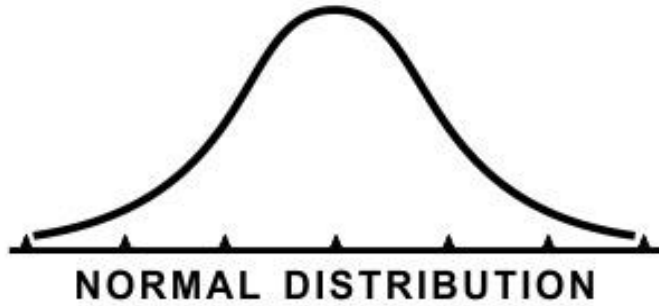
- In a **symmetrical distribution**, it is possible to draw a vertical line through the middle so that one side of the distribution is a mirror image of the other.
- In a **skewed distribution**, the scores tend to pile up toward one end of the scale and taper off gradually at the other end.
- The section where the scores taper off toward one end of a distribution is called the **tail of the distribution**.
- A **skewed distribution** with the tail on the right-hand side is **positively skewed** because the tail points toward the positive (above-zero) end of the X-axis. If the tail points to the left, the distribution is **negatively skewed**.



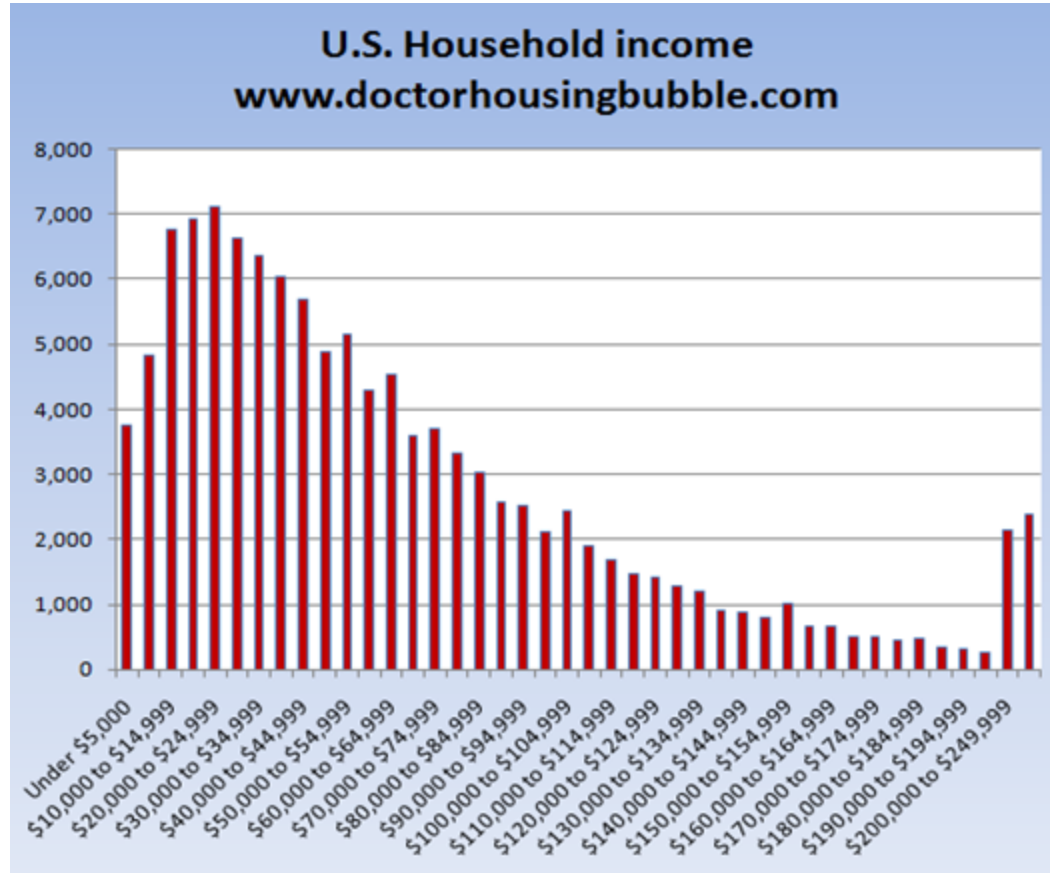
# Normal distribution



# Normal distribution

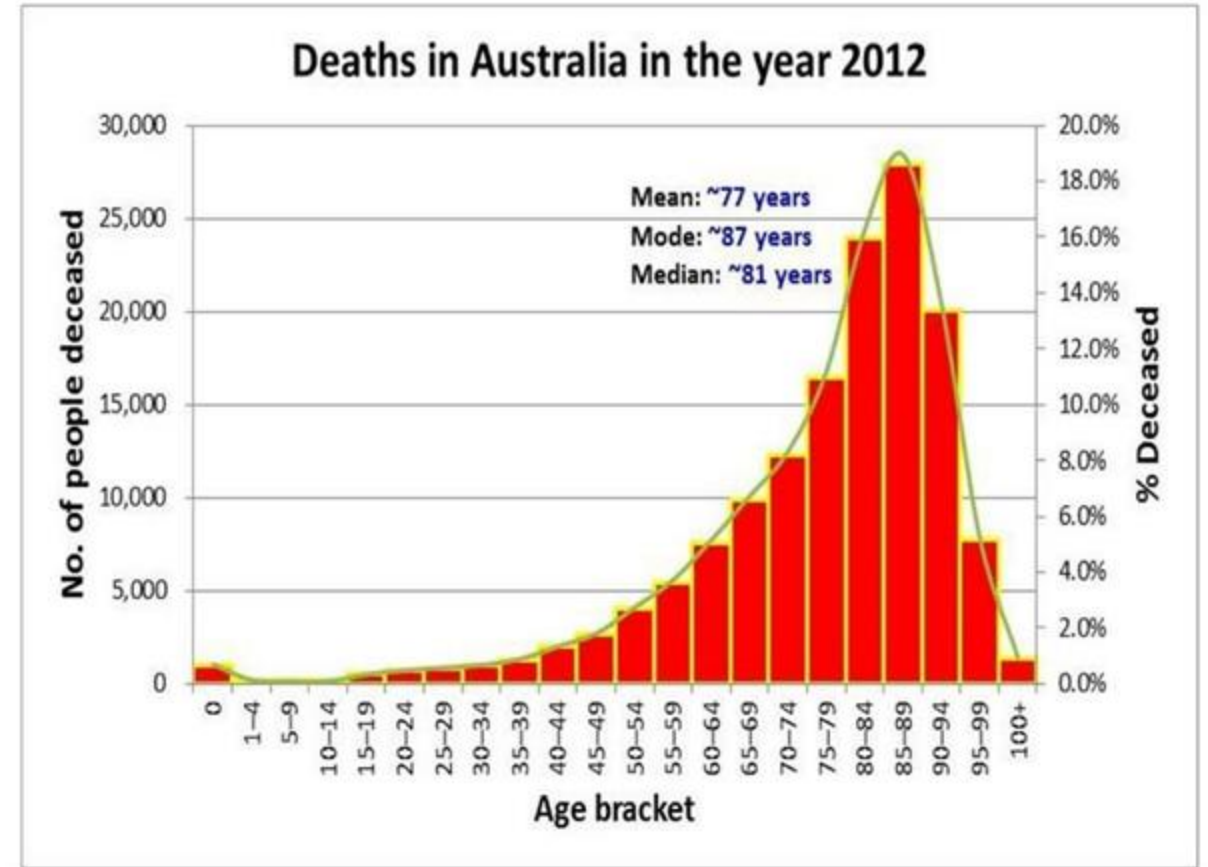
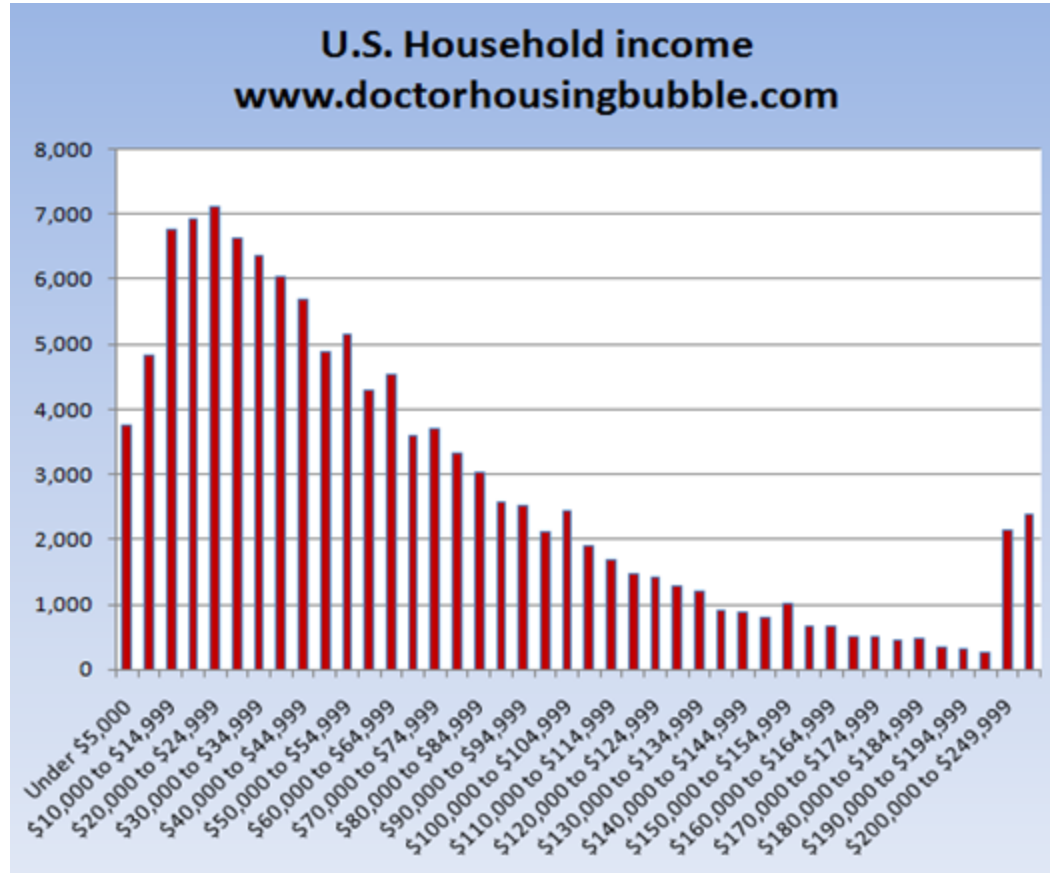


# Skewed distribution





# Skewed distribution



# Learning check

- 1.** The seminar rooms in the library are identified by letters (A, B, C, and so on). A professor records the number of classes held in each room during the fall semester. If these values are presented in a frequency distribution graph, what kind of graph would be appropriate?
  - a.** a histogram
  - b.** a polygon
  - c.** a histogram or a polygon
  - d.** a bar graph

# Learning check

- 2.** A group of quiz scores ranging from 4–9 are shown in a histogram. If the bars in the histogram gradually increase in height from left to right, what can you conclude about the set of quiz scores?
- a.** There are more high scores than there are low scores.
  - b.** There are more low scores than there are high scores.
  - c.** The height of the bars always increases as the scores increase.
  - d.** None of the above

# Measures of central tendency



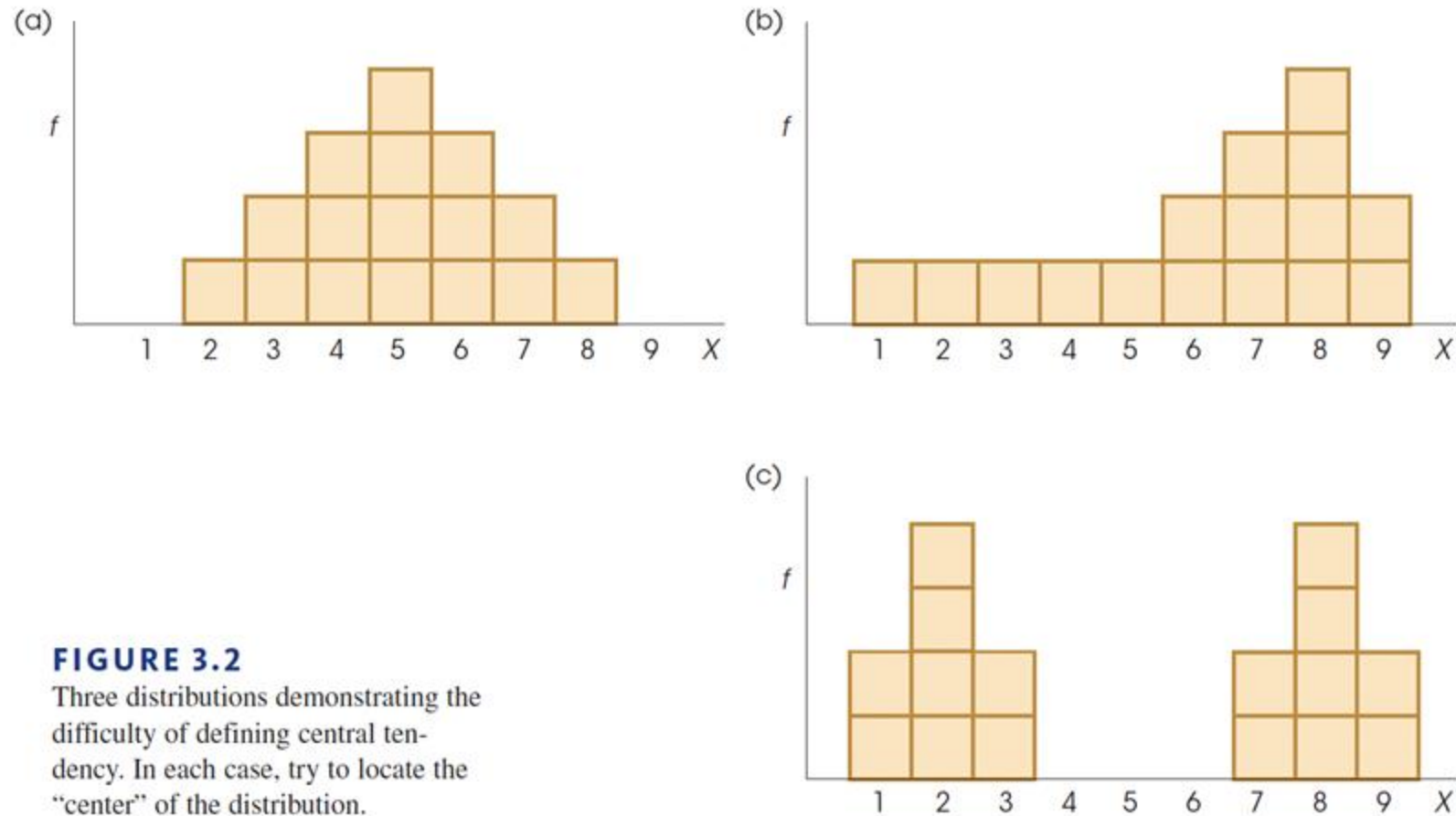
# Central tendency

- ▶ Central tendency is a statistical measure to determine a single score that defines the center of a distribution. The goal of central tendency is to find the single score that is most typical or most representative of the entire group.

Examples of central tendency metrics:

- ✓ Mean
- ✓ Median
- ✓ Mode

# No central tendency measure is good for every situation



# Mean

- ▶ The **mean** for a distribution is the sum of the scores divided by the number of scores.

Population formula:

$$\mu = \frac{\Sigma X}{N}$$

Sample formula:

$$M = \frac{\Sigma X}{n}$$

The mean as a balance point:

- Notice that the mean balances the distances. That is, the total distance below the mean is the same as the total distance above the mean:

**below the mean:**  $4 + 3 = 7$  points

**above the mean:**  $1 + 1 + 5 = 7$  points

Score	Distance from the Mean
$X = 1$	4 points below the mean
$X = 2$	3 points below the mean
$X = 6$	1 point above the mean
$X = 6$	1 point above the mean
$X = 10$	5 points above the mean

# Learning check

**1.** What is the mean for the following sample of scores?

Scores: 1, 2, 5, 4

- a.** 12
- b.** 6
- c.** 4
- d.** 3

**2.** A sample has a mean of  $M = 45$ . If one person with a score of  $X = 53$  is removed from the sample, what effect will it have on the sample mean?

- a.** The sample mean will increase.
- b.** The sample mean will decrease.
- c.** The sample mean will remain the same.
- d.** Cannot be determined from the information given



# Finding the median

3, 5, 8, 10, 11

# Finding the median

3, 5, 8, 10, 11

# Finding the median

3, 5, 8, 10, 11

1, 1, 4, 5, 7, 8

# Finding the median

3, 5, 8, 10, 11

1, 1, 4, 5, 7, 8

$$\text{median} = \frac{4 + 5}{2} = \frac{9}{2} = 4.5$$

# Median

- ▶ If the scores in a distribution are listed in order from smallest to largest, the median is the midpoint of the list. More specifically, the **median** is the point on the measurement scale below which 50% of the scores in the distribution are located.
- ▶ Defining the median as the midpoint of a distribution means that the scores are being divided into two equal-sized groups. We are not locating the midpoint between the highest and lowest X values. To find the median, list the scores in order from smallest to largest. Begin with the smallest score and count the scores as you move up the list. The median is the first point you reach that is greater than 50% of the scores in the distribution.

# Learning check

**2.** What is the median for the following set of scores:

Scores: 8, 10, 11, 12, 14, 15

- a.** 11
- b.** 11.5
- c.** 12
- d.**  $\frac{70}{6} = 11.67$

# Mode

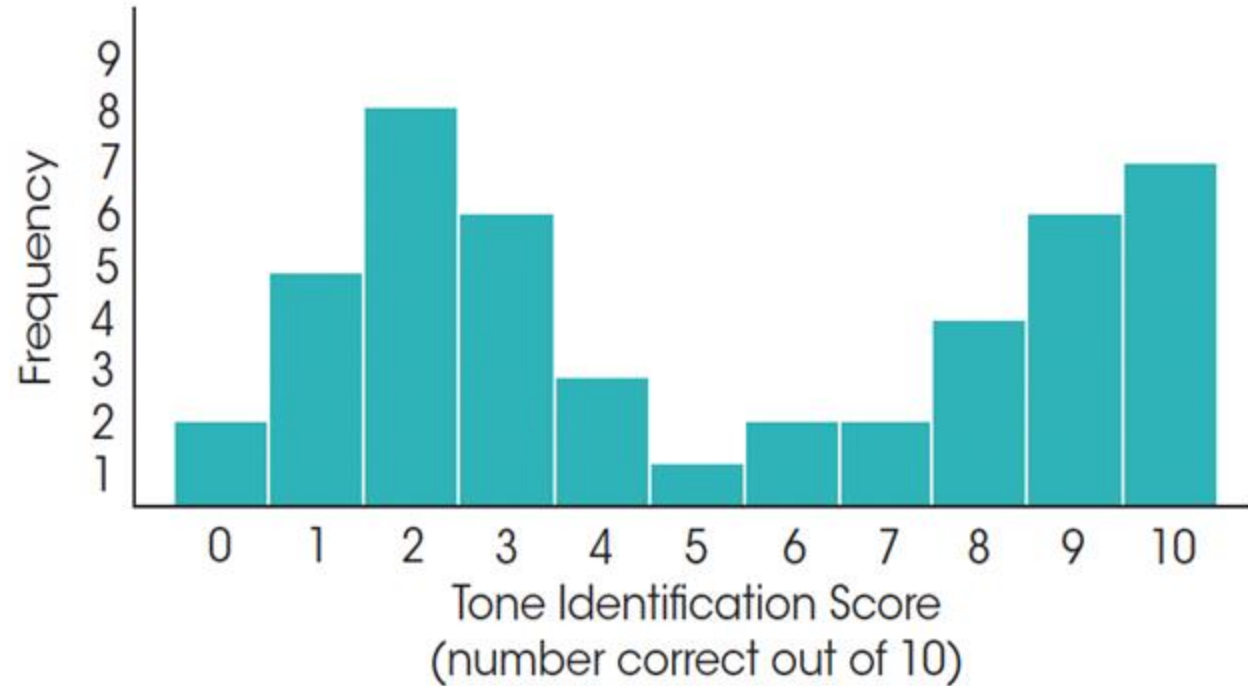
- ▶ In a frequency distribution, the **mode** is the score or category that has the greatest frequency.
- ▶ Although a distribution will have only one mean and only one median, it is possible to have more than one mode. Specifically, it is possible to have two or more scores that have the same highest frequency.
- ▶ In a frequency distribution graph, the different modes will correspond to distinct, equally high peaks. A distribution with two modes is said to be **bimodal**, and a distribution with more than two modes is called **multimodal**. Occasionally, a distribution with several equally high points is said to have no mode.

Restaurant	<i>f</i>
College Grill	5
George & Harry's	16
Luigi's	42
Oasis Diner	18
Roxbury Inn	7
Sutter's Mill	12

# Bimodal distribution

**FIGURE 3.7**

A frequency distribution for tone identification scores. An example of a binomial distribution.



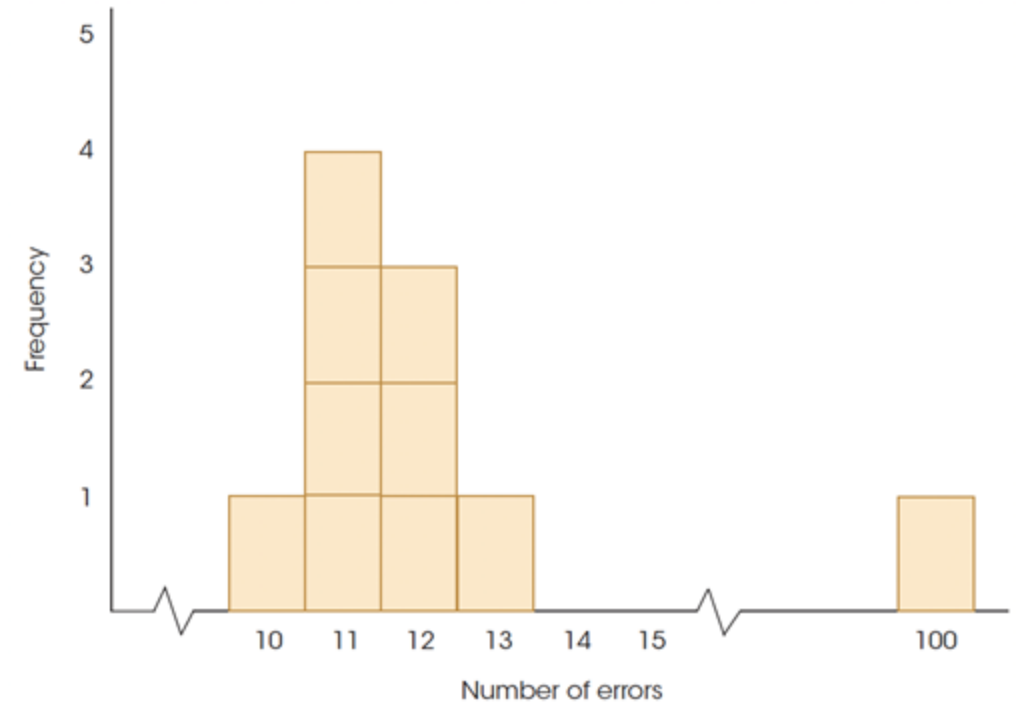


# Selecting a measure of central tendency

You can use **mean** for **numerical features** and for **symmetrical distributions** since it can be easily affected by extreme values and become less representative.

It's better to use the **median** for **skewed distributions**. Also do not use mean but median, when you have **undetermined values** in your data. The median is also suitable for an **ordinal scale** when the mean is not!

The **mode** you can find for any distribution, but most likely you will use it with **nominal data** since it is the only central tendency metric that can be calculated for those. Also, it makes sense for discrete variables (like children's numbers).

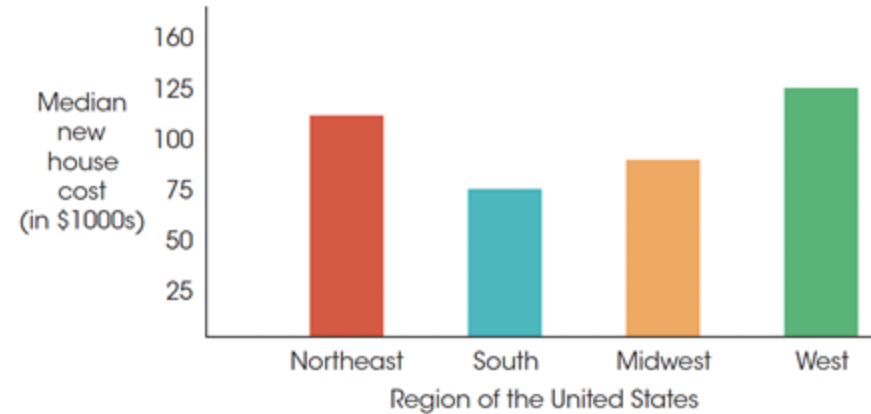
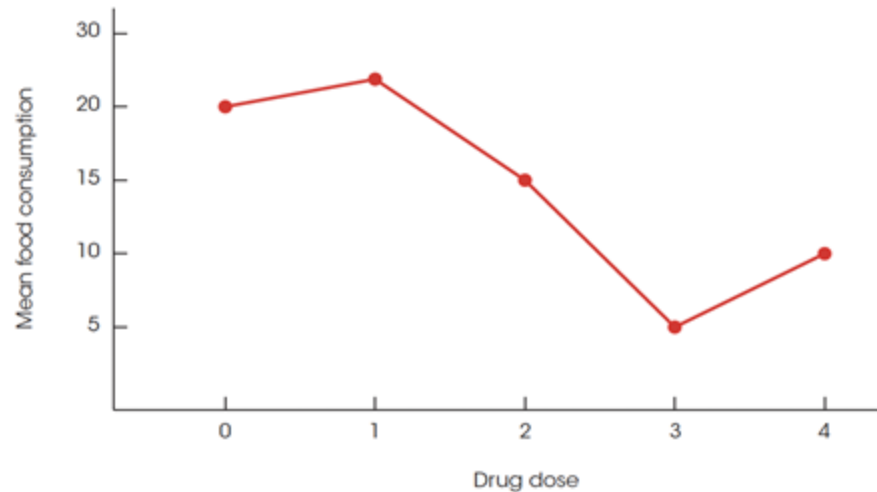


# Undetermined values & Open-ended data

Person	Time (Min.)
1	8
2	11
3	12
4	13
5	17
6	Never finished

Number of Pizzas (X)	f
5 or more	3
4	2
3	2
2	3
1	6
0	4

- Mean and median's in graphs:



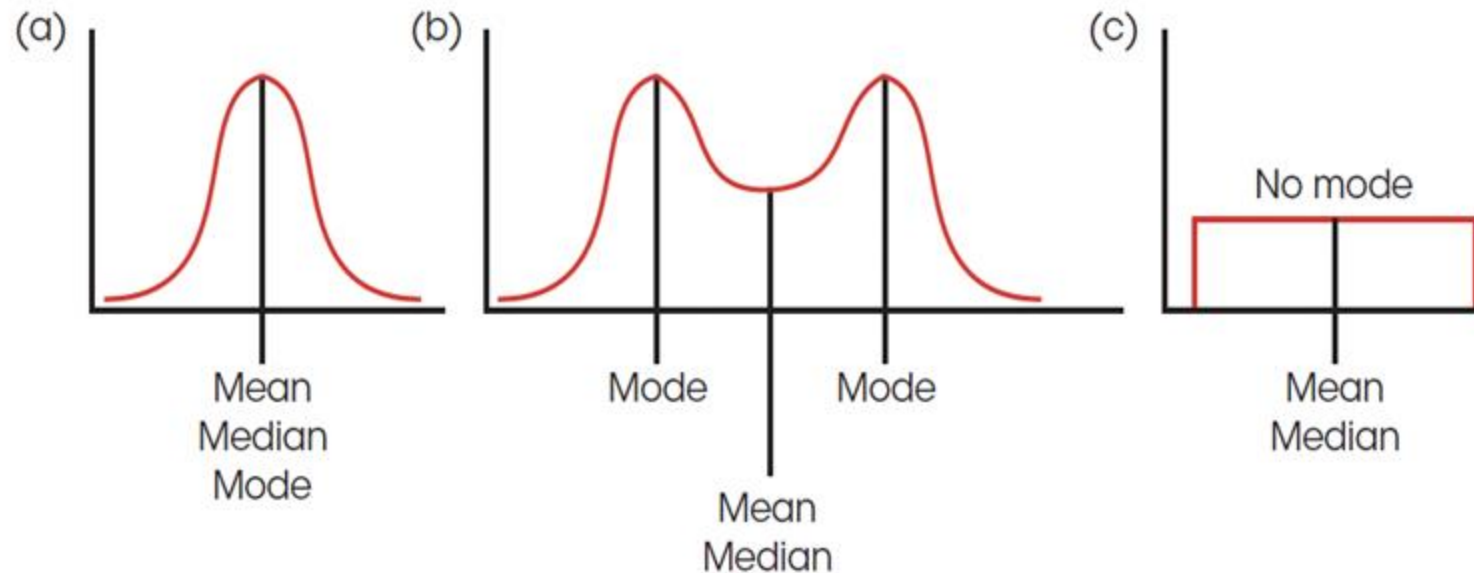
# Central Tendency and the Shape of the Distribution

There are situations in which all three measures will have exactly the same value. On the other hand, there are situations in which the three measures are guaranteed to be different. In part, the relationships among the mean, median, and mode are determined by the shape of the distribution. We will consider two general

# Symmetrical distribution

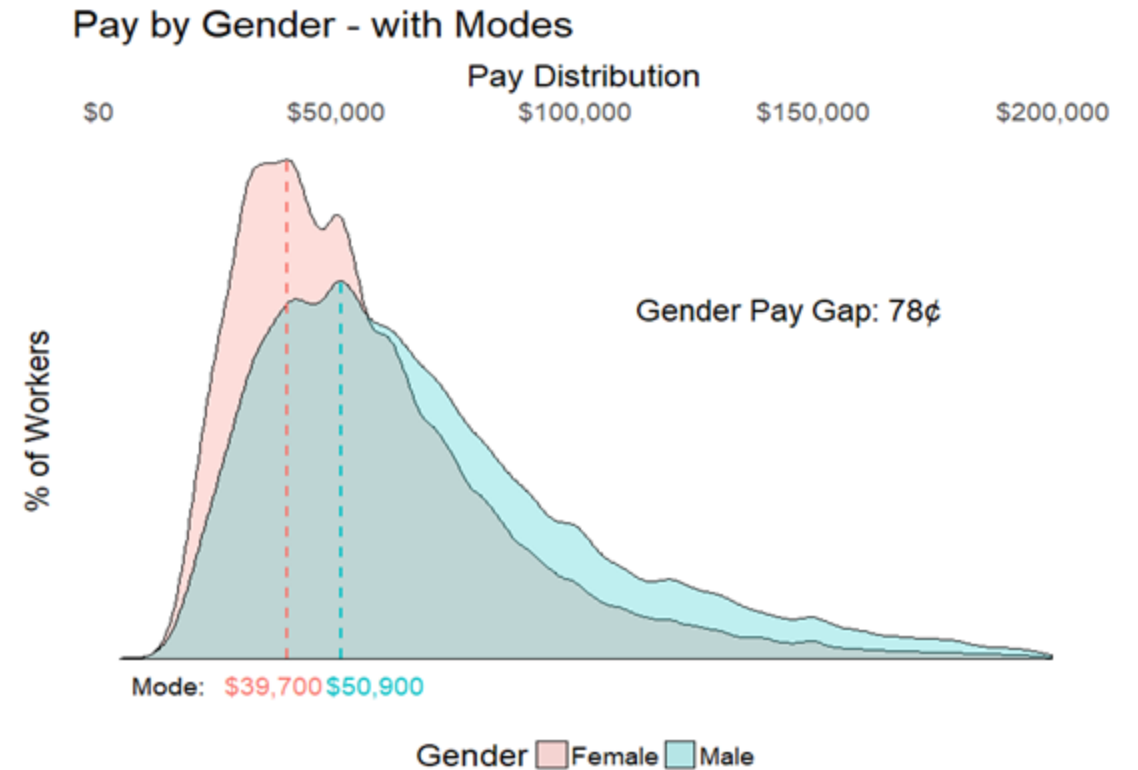
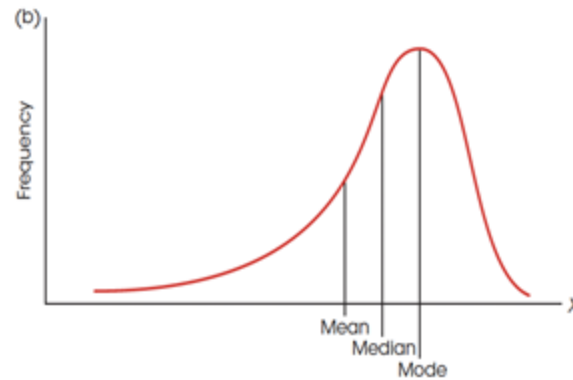
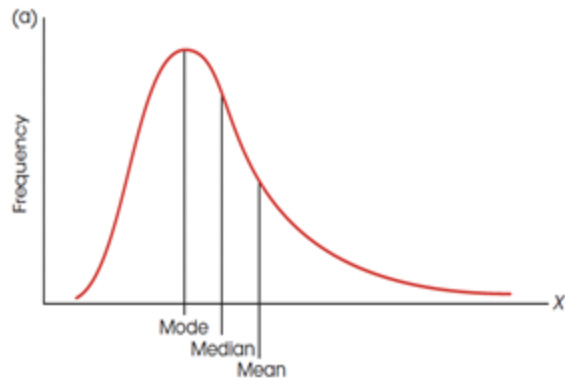
For a perfectly symmetrical distribution with one mode, all three measures of central tendency — the mean, median, and mode — have the same value.

However, there will be cases that are different.



# Skewed distribution

In skewed distributions, especially distributions for continuous variables, there is a strong tendency for the mean, median, and mode to be located in predictably different positions.



# Selecting a measure of central tendency

	<b>Mean</b>	<b>Median</b>	<b>Mode</b>
Nominal variables	No	No	Yes
There are extreme values	No	Yes	Yes
Distribution is symmetrical	Yes (with caution)	Yes	Yes
Distribution is skewed	No	Yes	No
Ordinal variables	No	Yes	Yes

# Learning check

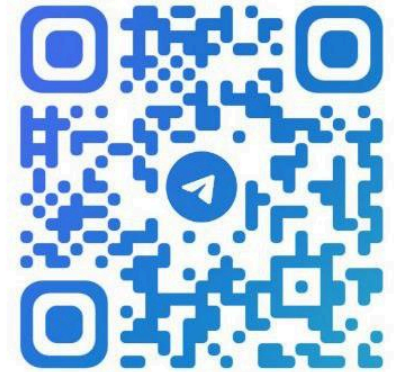
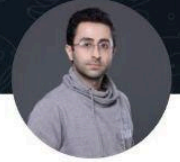
- 1.** Which of the following is true for a symmetrical distribution?
  - a.** the mean, median, and mode are all equal
  - b.** mean = median
  - c.** mean = mode
  - d.** median = mode

# Thank you!



Majid Sohrabi

[msohrabi@hse.ru](mailto:msohrabi@hse.ru)



@MSOHRABI\_CS