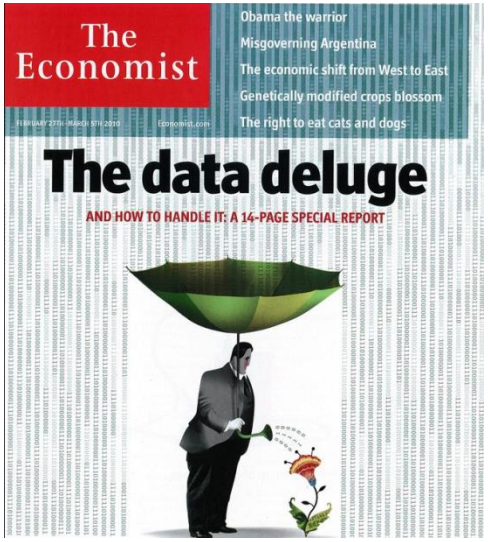
APPLIED STATISTICS
Wroclaw U, Fall 2015

Lecture 2.1: Some Motivation to Learn Stats

Lecture 2.2: Graphical an Numerical Data Descriptions/Summaries



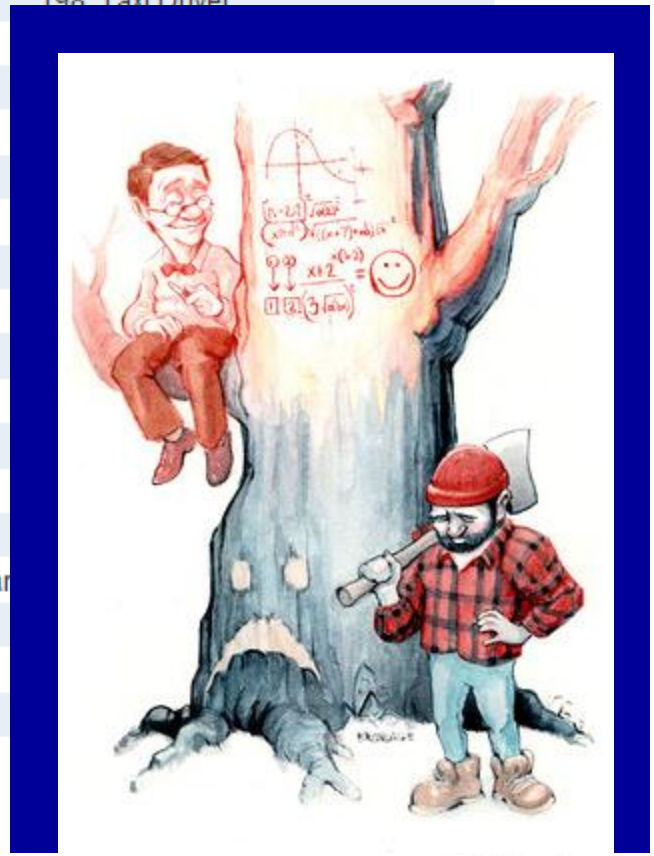
"Chief information officers (CIOs) have become somewhat more prominent in the executive suite, and a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data.

Hal Varian, Google's chief economist, predicts that the job of statistician will become the "sexiest" around. Data, he explains, are widely available; what is scarce is the ability to extract wisdom from them."

The Best and Worst Jobs

Of 200 Jobs studied, these came out on top -- and at the bottom:

The Best	The Worst
1. Mathematician	200. Lumberjack
2. Actuary	199. Dairy Farmer
3. Statistician	198. Taxi Driver
4. Biologist	
5. Software Engineer	
6. Computer Systems Analyst	
7. Historian	
8. Sociologist	
9. Industrial Designer	
10. Accountant	
11. Economist	
12. Philosopher	
13. Physicist	
14. Parole Officer	
15. Meteorologist	
16. Medical Laboratory Technician	
17. Paralegal Assistant	
18. Computer Programmer	
19. Motion Picture Editor	
20. Astronomer	



Scott Brundage

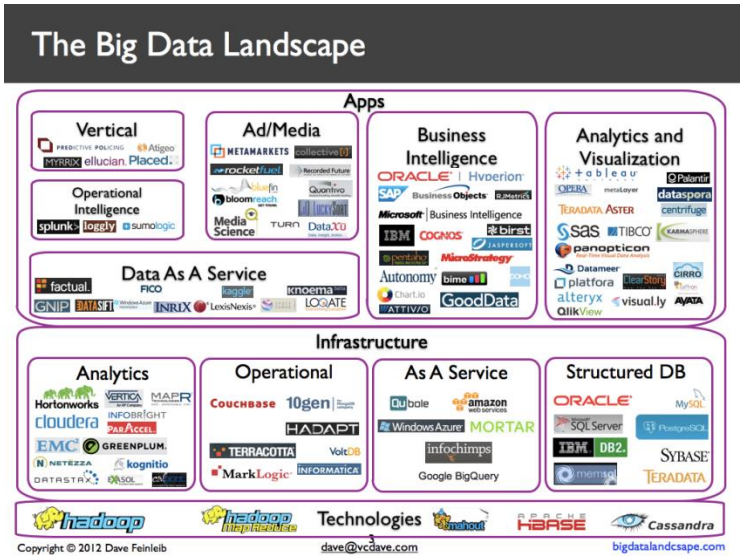
The New York Times
Expect the World®

"GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

A report last year by the McKinsey Global Institute,…, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers."

"It's a revolution," says Gary King, director of Harvard's Institute for Quantitative Social Science. "We're really just getting under way. But the march of quantification, made possible by enormous new sources of data, will sweep through academia, business and government. There is no area that is going to be untouched."

Welcome to the Age of Big Data."



What is Statistics ?

Collect data

(census, polls, questionnaire, elections, physical experiments, *etc.*)

Analyze data

(summarize, visualize, construct models)

Interpret data = make conclusions

(tell what is happening now and what may happen in the future)

Example: Engineering Math/Stat class performance

Data from MATH/STAT, Fall 2012 (86 people)

I Midterm 2	HW average	Minitab average	Quiz average	Attendance	PRELIM COURSE S	PRELIM. LI	Final exam	Score afte	Final course score
	1.66666667	3.4		4	14.93	F		12.77	14.93
				0	2.5	F		2	2.5
49	9.08333333	6.6	100	8	84.57	B		66.78	84.57
				1	3.5	F		3	3.5
50	8.16666667	9.8	100	10	88.53	B+	48	91.27	91.27
44	4.66666667	8	90	9	78.83	C+	49	84.97	84.97
47	4.91666667	3.8	100	10	61.53	D	35	64.42	64.42
49	8.75	7	50	10	84.5	B		67.15	84.5
32	6.83333333	8.6	50	8	69.57	D+	31	68.83	68.83
49	9.83333333	10	100	10	87.67	B+	36	84.83	84.83
47	9.08333333	10	100	10	89.67	B+		72.68	89.67
48	8.91666667	2.2	68	10	67.03	D+	40	67.62	67.62
				1	3.5	F		3	3.5

Historically, the goal of statistics was to collect and analyze data for government

New Latin -- *statisticum collegium* = "council of state"

Italian -- *statista* = "statesman" or "politician"

German – *Statistik* = *"science of state"*

Census: the oldest data collection method

- A census is the process of obtaining information about every member of a population.
  Examples: *nation population census, agriculture census, voting, …*

- The first census was taken by the Babylonians in 3800 BC, nearly 6000 years ago. It counted the number of people, livestock, quantities of butter, honey, milk, wool and vegetables.

劉

The oldest existing data are on the census in China during the Han Dynasty, in the fall of 2 AD.

Census: the most expensive data collection method

- The cost of a census in a developing country is about $3 per person
- The last US census happened in March 2010
- The cost for the 2010 census in the US was estimated to be $14 billion, which is about $45 per person
- Denmark, Finland and Norway use administrative registers.
- France and Germany use partial censuses: so-called ‘Micro censuses’ or ‘Sample censuses'.

Goal of statistics?

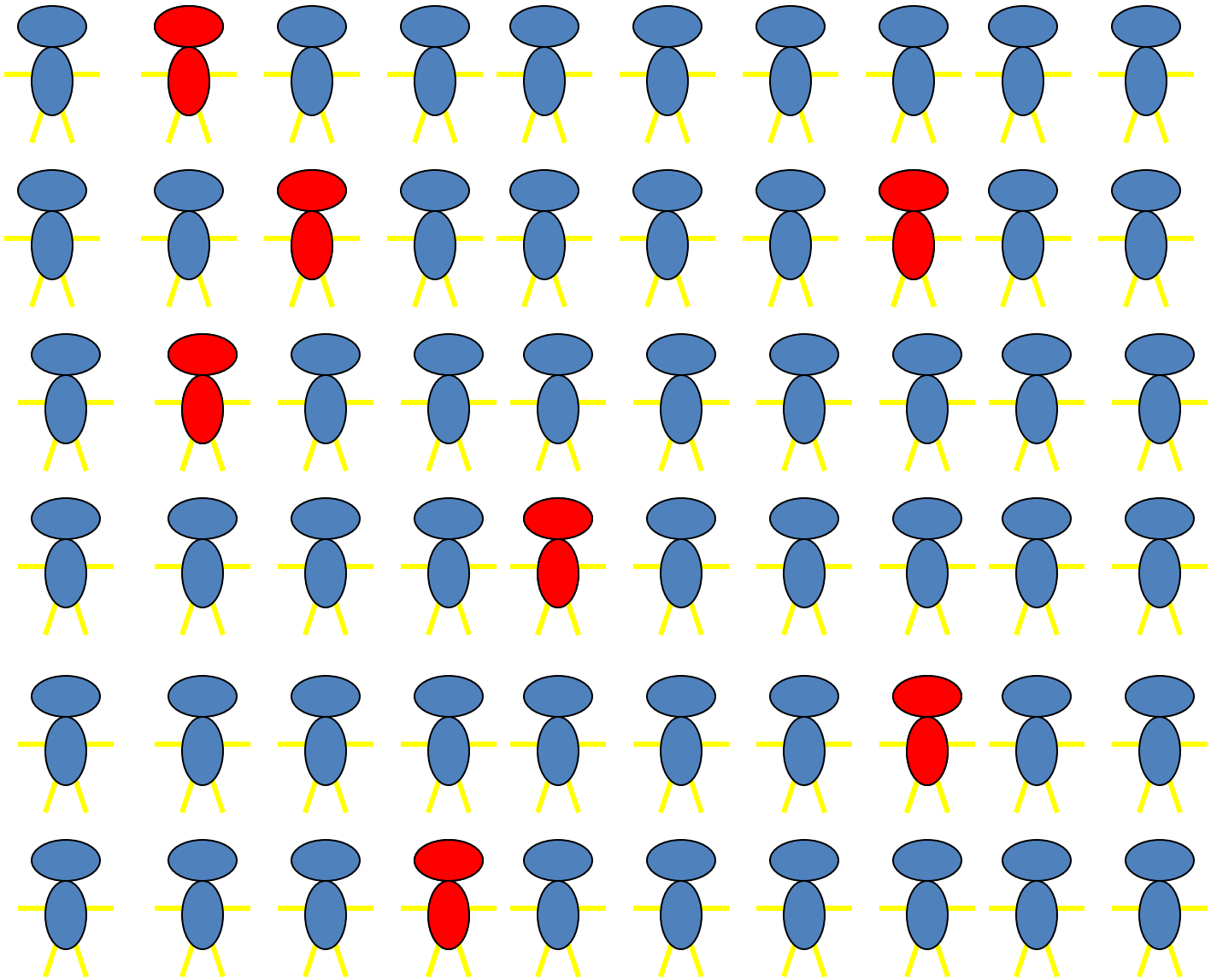
Statistics tries

to *obtain conclusions* about the population

by analyzing data from a

small part of the population called *sample*.

Population and sample



Currently, statistics is among disciplines fundamental for numerous aspects of human life

Important applications include:

- **Public health (biostatistics, epidemiology, etc.)**
- **Socio-economics (unemployment, econometry, etc.)**

- **Medicine**
- **Engineering**
- **Natural sciences**
- **Finance**
- **Marketing**
- **Education**
- **Industrial quality control**
- **Gaming industry**
- **…**

Three main problems of Statistics

Estimation (point or interval)

- What proportion of the voters do support Mr. Green's new CO2 reduction policy?

Hypothesis testing

- Is it true that the proportion of Mr. Green's policy supporters is above 60%?

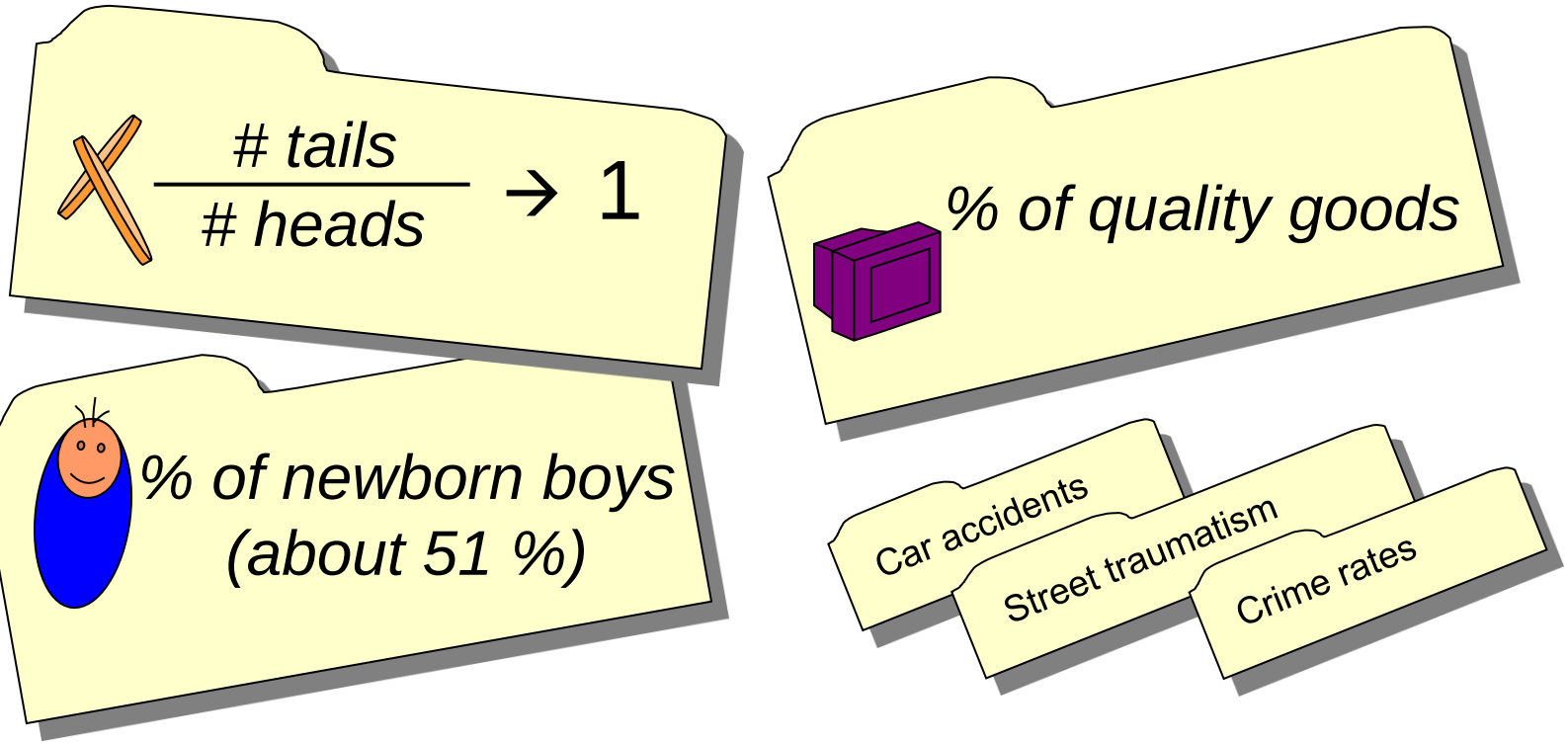
Correlation analysis

- How the Earth temperature is related to the amount of CO2 in the atmosphere?

Why does this work?

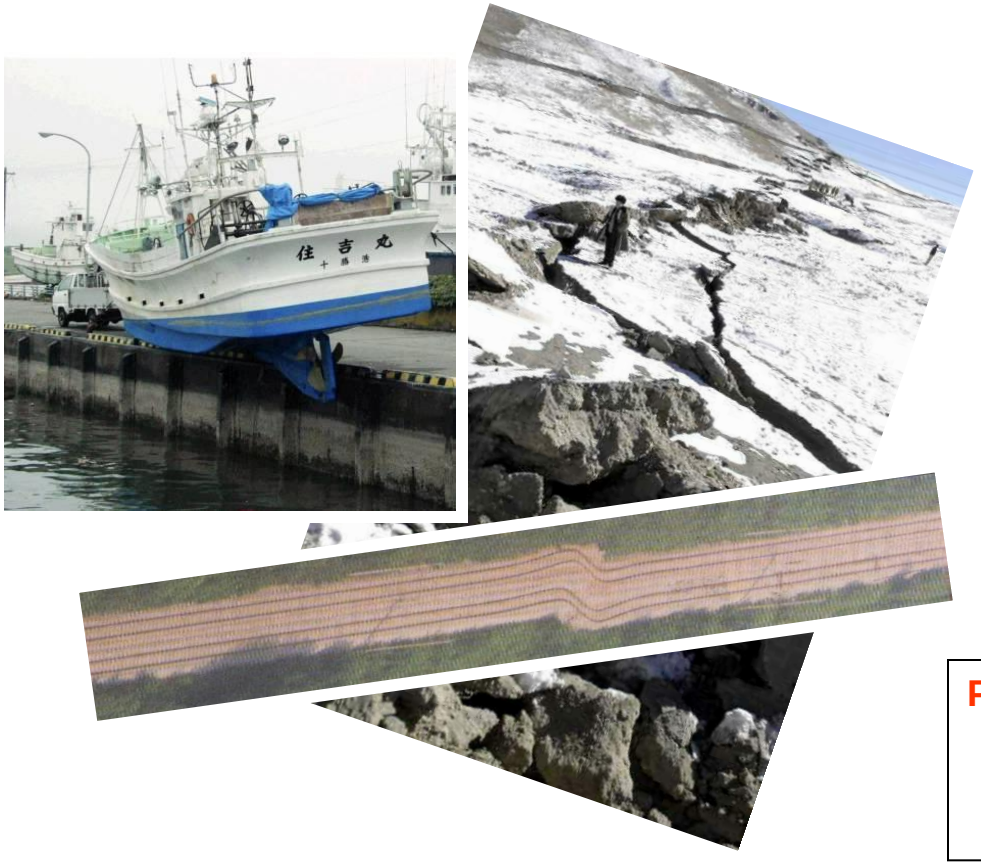
Stability of Frequencies

The observed phenomenon of **stability of frequencies**:
Observing a large number of similar random events,
we often detect amazing regularities

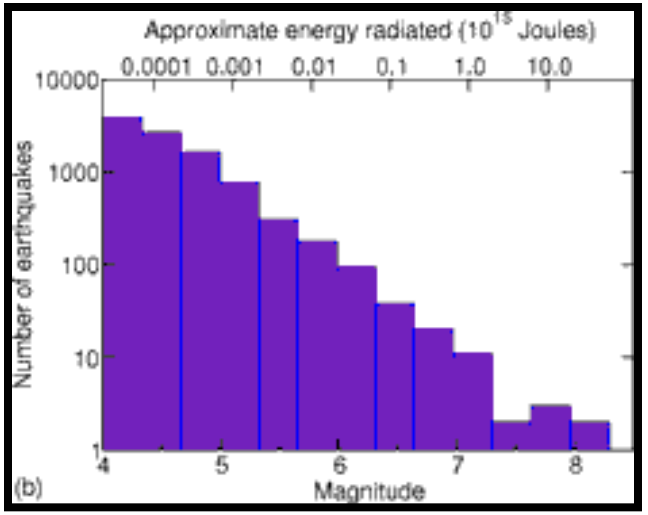


Stability of Frequencies

Earthquakes: the most unpredictable Natural disaster



Gutenberg-Richter law



Prediction:

In 2012 there will be about 1000 EQs with magnitude 5, and 10 EQs with M7

Individual vs. collective behavior

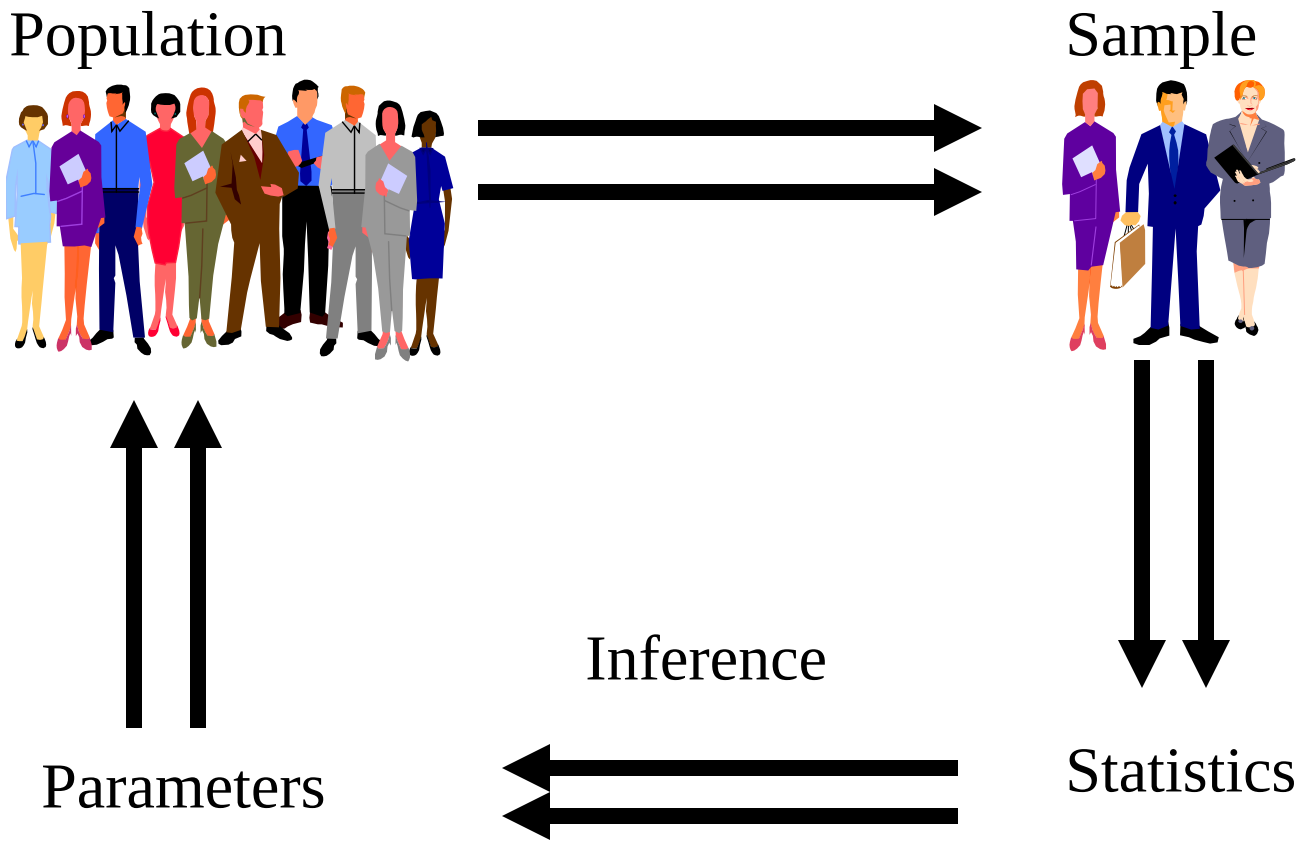
$$$ = ???

$$$ = !!!

Lecture 2.2

Graphical an Numerical Data Descriptions/Summaries

STATISTICAL PROBLEMS



Sampling

Definitions:

- **A population is the entire collection of objects or outcomes about which information is sought.**

- **A sample is a subset of a population, containing the objects or outcomes that are actually observed.**

- **A simple random sample (SRS) of size *n* is a sample chosen by a method in which each collection of *n* population items is equally likely to comprise the sample, just as in the lottery.**

Types of Data

- **A subset of the data from a study of a series of male patients from Greenlane Hospital in Aukland after a heart attack**
- Goal of the study:
  How long will the patient live after the heart attack?

TABLE 2.1.1 Data on Male Heart Attack Patients

ID	EJEC	SYS-VOL	DIA-VOL	OCCLU	STEN	TIME	OUT-COME	AGE	SMOKE	BETA	CHOL[a]	SURG
390	72	36	131	0	0	143	0	49	2	2	59	0
279	52	74	155	37	63	143	0	54	2	2	68	1
391	62	52	137	33	47	16	1	56	2	2	52	0
201	50	165	329	33	30	143	0	42	2	2	39	0
202	50	47	95	0	100	143	0	46	2	2	74	1
69	27	124	170	77	23	143	0	57	2	2	NA	2
310	60	86	215	7	50	40	0	51	2	2	58	0
392	72	37	132	40	10	9	1	56	2	2	75	0
311	60	65	163	0	40	142	0	45	2	2	72	0
393	63	52	140	0	10	142	0	46	2	2	90	0
70	29	117	164	50	0	142	0	48	2	2	72	0
203	48	69	133	0	27	142	0	54	2	2	NA	0
394	59	54	133	30	13	142	0	39	2	1	NA	0
204	50	67	135	37	63	141	0	49	2	2	86	2
280	53	65	138	0	33	140	0	58	2	1	49	0
55	17	184	221	57	13	5	1	50	2	2	70	2
79	37	88	140	37	47	118	1	58	2	2	NA	0
205	45	106	193	33	43	140	0	47	1	1	38	1
206	43	85	150	0	50	23	1	51	2	2	61	0
312	60	59	149	7	37	139	0	43	2	1	56	0
80	38	103	168	47	43	100	1	55	2	2	62	1
281	57	53	124	0	57	140	0	58	2	1	93	0
207	44	68	121	27	60	139	0	55	2	2	63	1
282	51	53	109	0	77	139	0	41	2	2	45	4
396	63	58	157	0	73	139	0	51	2	2	60	0
208	49	81	157	13	13	139	0	49	2	2	60	0
209	48	58	112	0	0	72	1	56	2	2	57	0
283	58	71	167	27	0	138	0	45	2	1	46	0
210	42	92	159	0	0	139	0	57	2	2	58	0
397	68	50	156	0	100	138	0	51	2	1	NA	0
211	43	146	259	47	33	3	1	56	2	2	70	0
398	67	43	130	0	70	138	0	49	2	2	NA	3
284	52	70	146	0	23	137	0	47	1	2	NA	0
399	63	73	195	27	0	136	0	36	1	1	61	0
285	54	62	133	33	23	137	0	38	2	2	NA	0
71	37	93	148	47	0	137	0	59	2	2	NA	0
286	51	65	133	43	7	136	0	54	2	2	NA	0
212	42	95	163	40	10	109	1	57	2	2	NA	4
400	66	49	144	10	50	65	1	52	2	2	55	0
287	54	66	145	7	40	136	0	47	2	2	62	0
81	39	144	237	13	87	136	0	39	2	2	56	3
813	63	52	141	0	47	43	1	48	2	2	NA	0
68	30	219	314	33	45	76	1	53	1	2	NA	0
288	59	39	94	0	0	135	0	47	1	2	63	0
407	67	39	117	0	73	53	1	57	2	2	62	2

[a] *NA* = Not Available(missing data code).

Types of variables

Quantitative

(Age, time, weight, etc.)

Qualitative

(Color, surgery outcome, smoking, etc.)

Continuous

May take any value from some interval (weight)

Discrete

May take values from some grid (age in years)

Categorical

No order (surgery outcome)

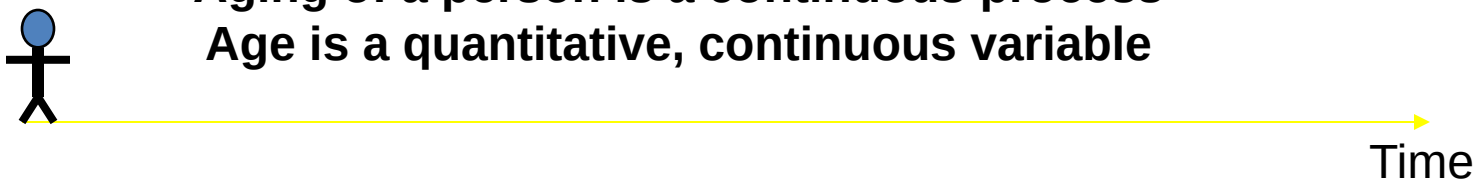
Ordinal

Order (Letter grade)

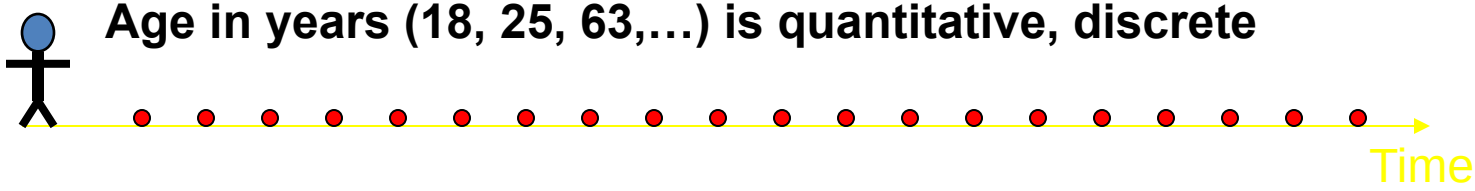
Type of variable

is determined by data you have and problem you consider

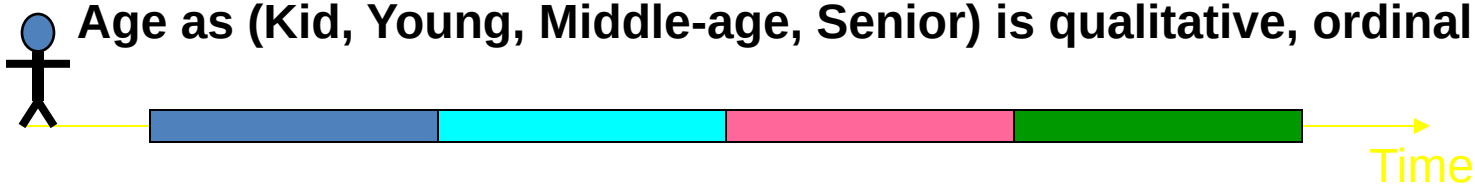
Aging of a person is a continuous process
Age is a quantitative, continuous variable



Age in years (18, 25, 63,…) is quantitative, discrete



Age as (Kid, Young, Middle-age, Senior) is qualitative, ordinal



Quantitative Data: Important Characteristics

1. **Center:** **A representative or average value that indicates where the middle of the data set is located.**

2. **Variation: A measure of the amount that the values vary among themselves.**

3. **Distribution:** **The nature or shape of the distribution of data (such as bell-shaped, uniform, or skewed).**

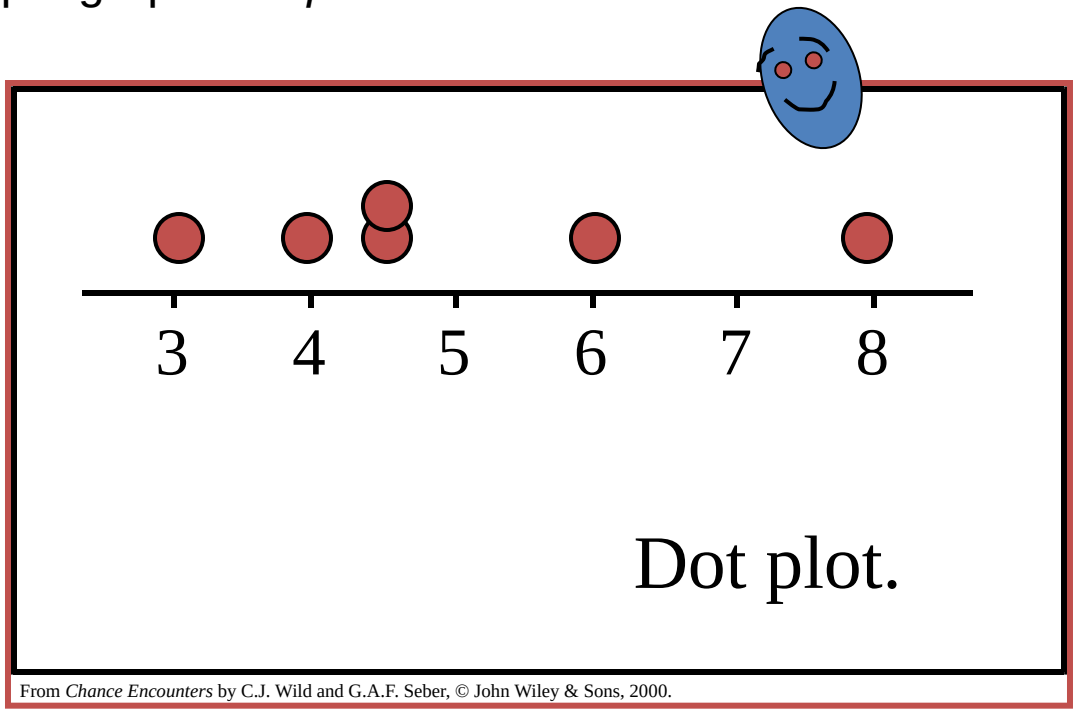
4. **Outliers:** **Sample values that lie very far away from the vast majority of other sample values.**

5. **Time: Changing characteristics of the data over time.**

Plots- Graphical Summaries of Data

Plots: dot plot

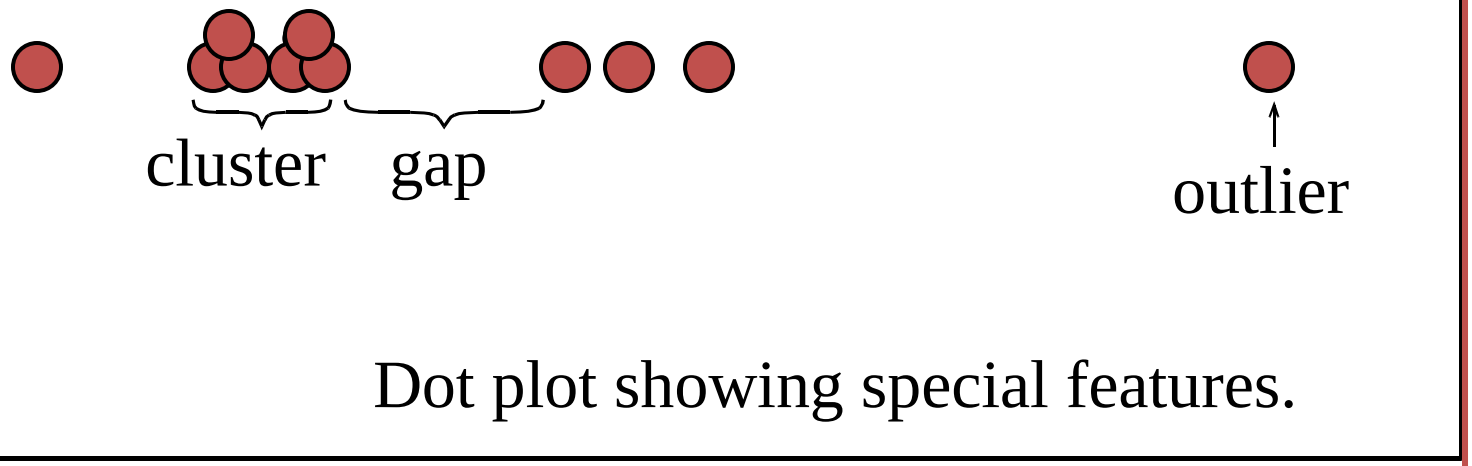
- Original data: {8 3 6 4.5 4 4.5}
- Sorted data: {3 4 4.5 4.5 6 8}
- Simple graph: *dot plot*



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Plots: dot plot

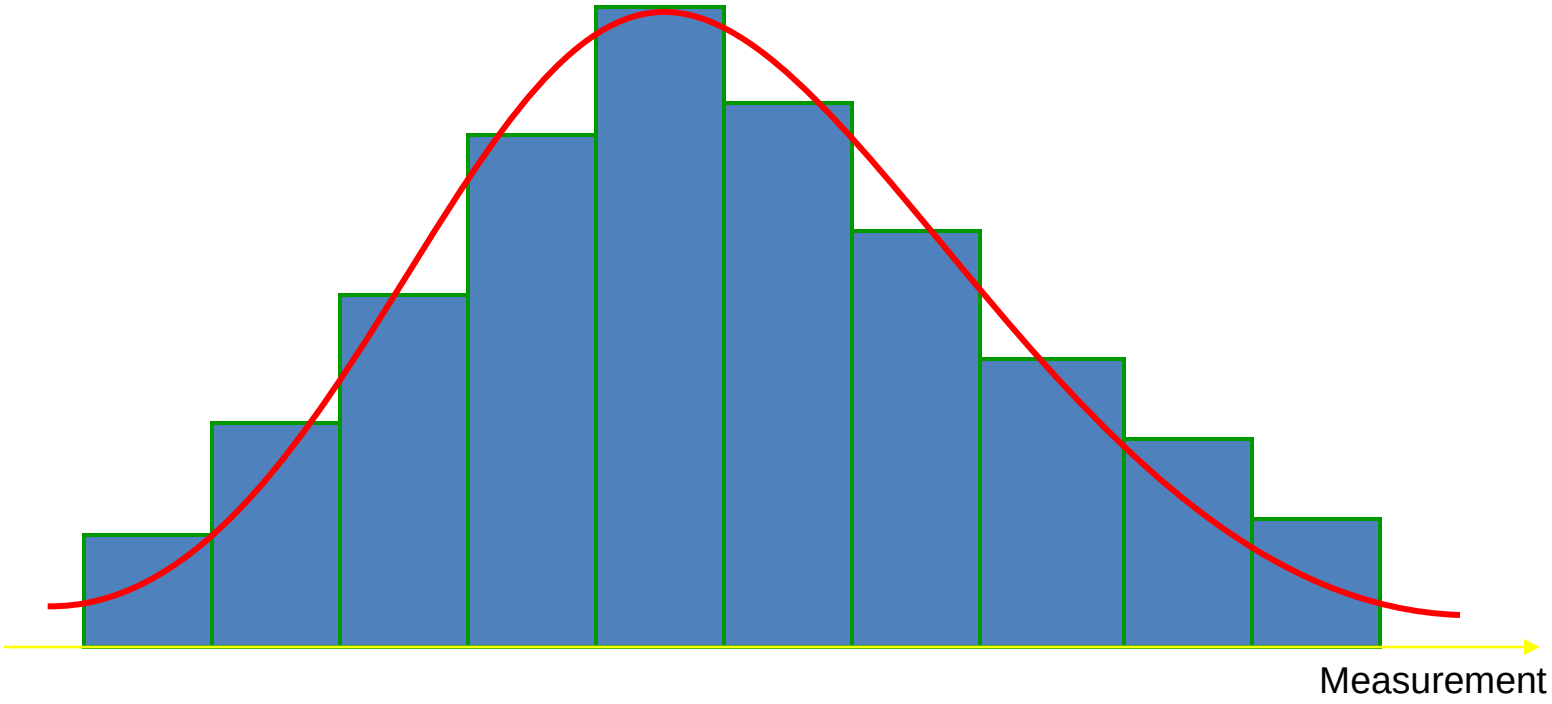
Interesting features of the data emphasized by the dot plot



Dot plot showing special features.

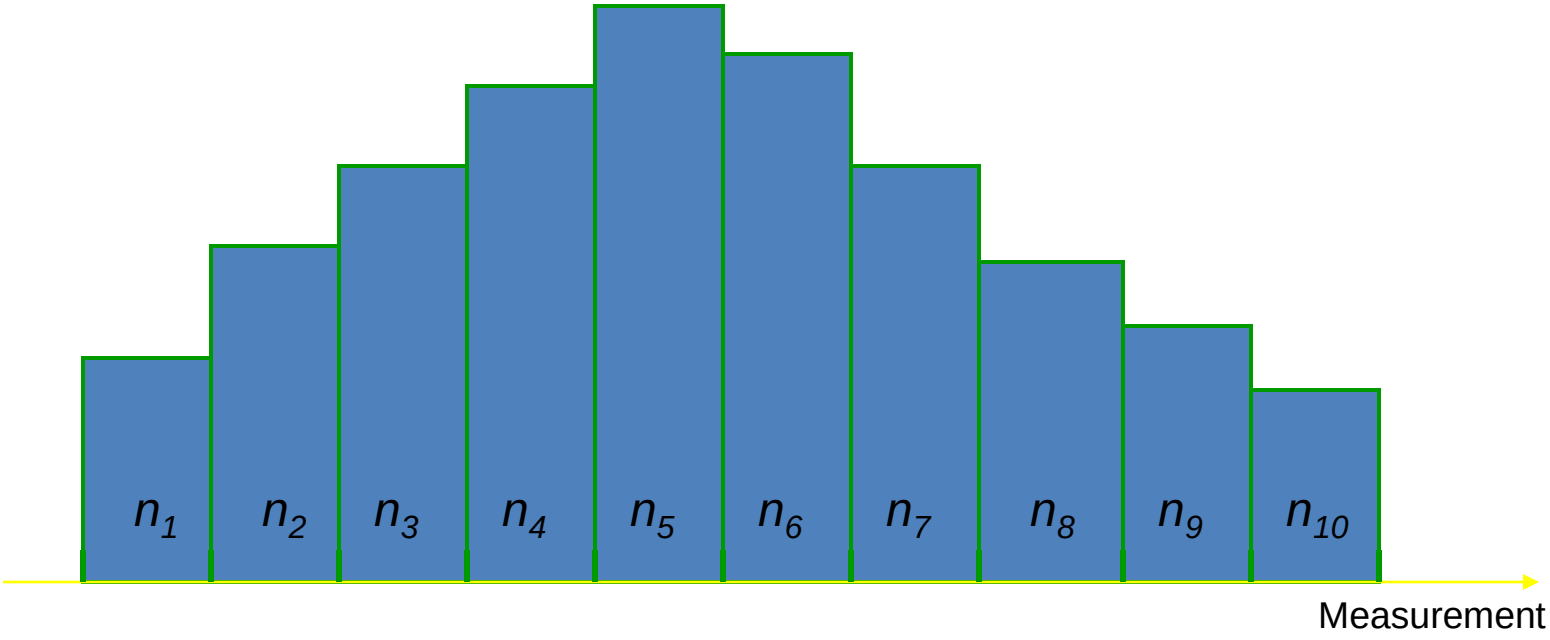
Plots: histogram

- Histogram is the most widely used statistical graph – shows shape of the data.



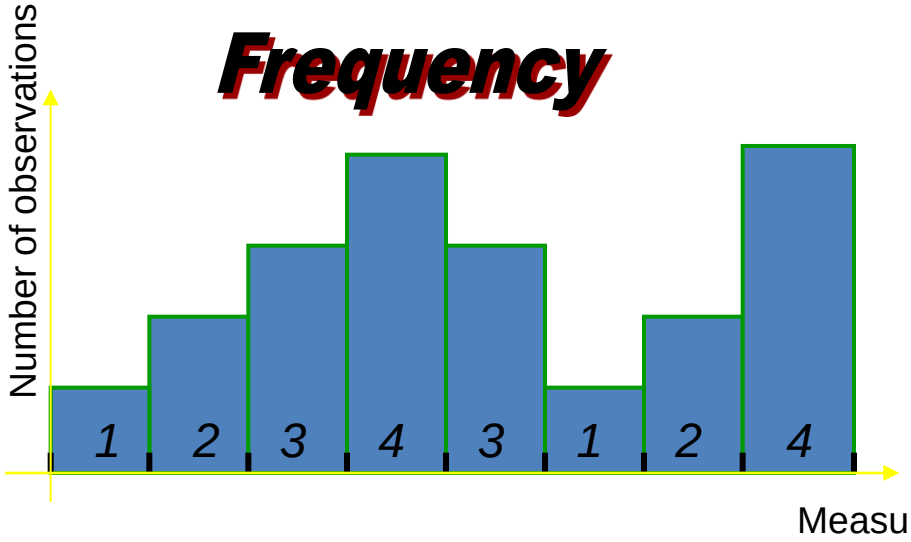
Plots: histogram

- Divide observational interval into subintervals,
  (also called *bins, class intervals*)
- Calculate number of observation within each bin
- Draw a rectangle w/heigth = number of observations = *frequency*
- *Relative frequency* is the number of observations within a bin divided by the total number of observations



Frequency vs. relative frequency

Frequency



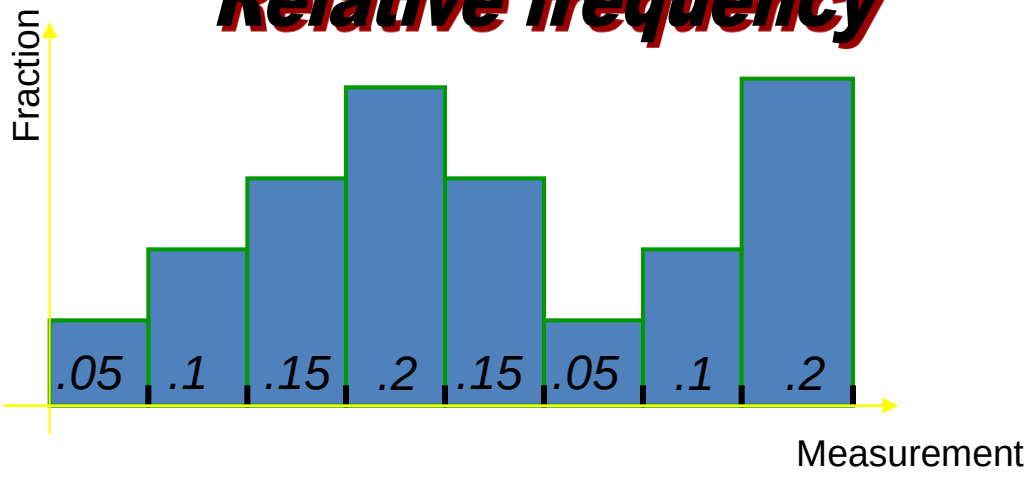
$$\sum_{i=1}^{k} n_i = n = 20$$

- n=20 (sample size)
  - k=8 (# of bins)

Relative frequency

$$\sum_{i=1}^{k} \frac{n_i}{n} = \frac{n}{n} = 1$$

- n=20 (sample size)
- k=8 (# of bins)



Histogram: Essential terminology

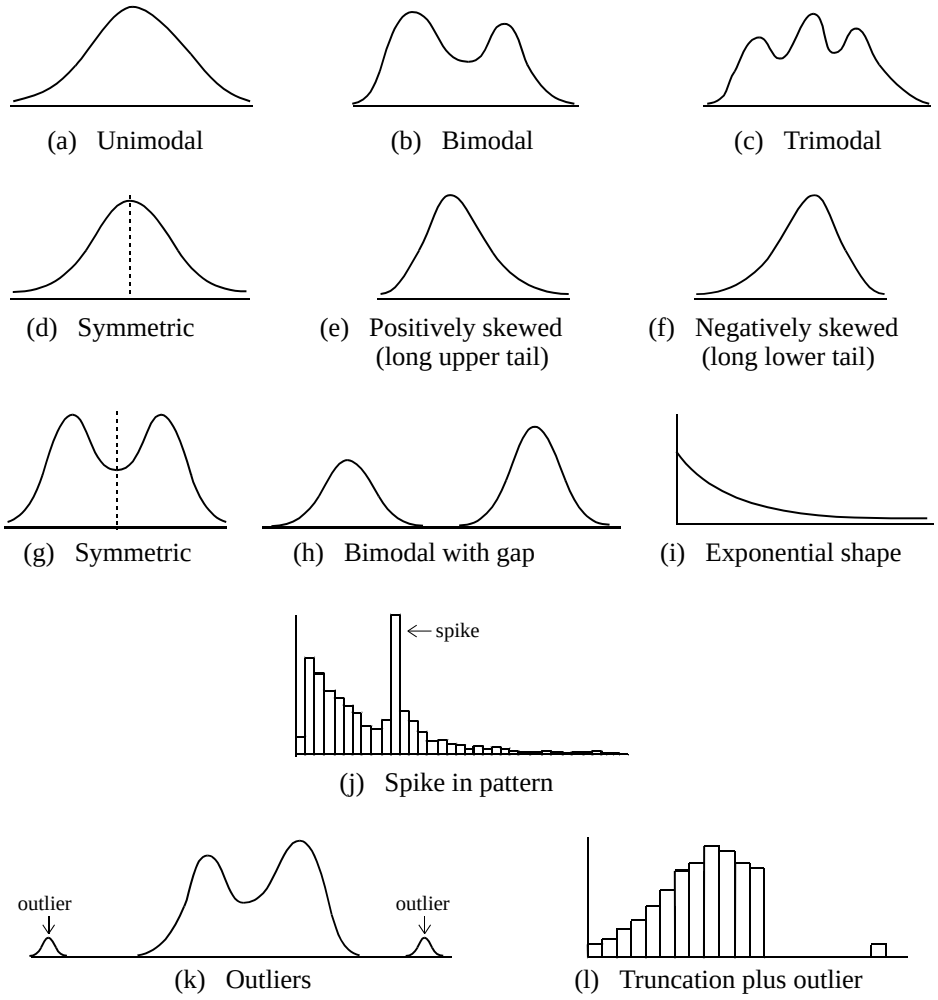
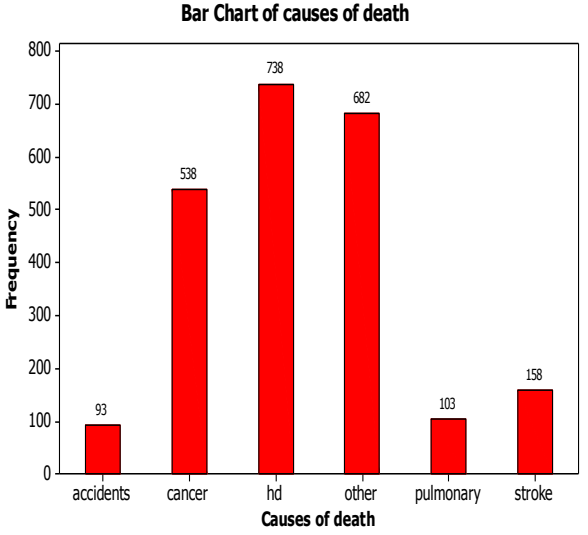


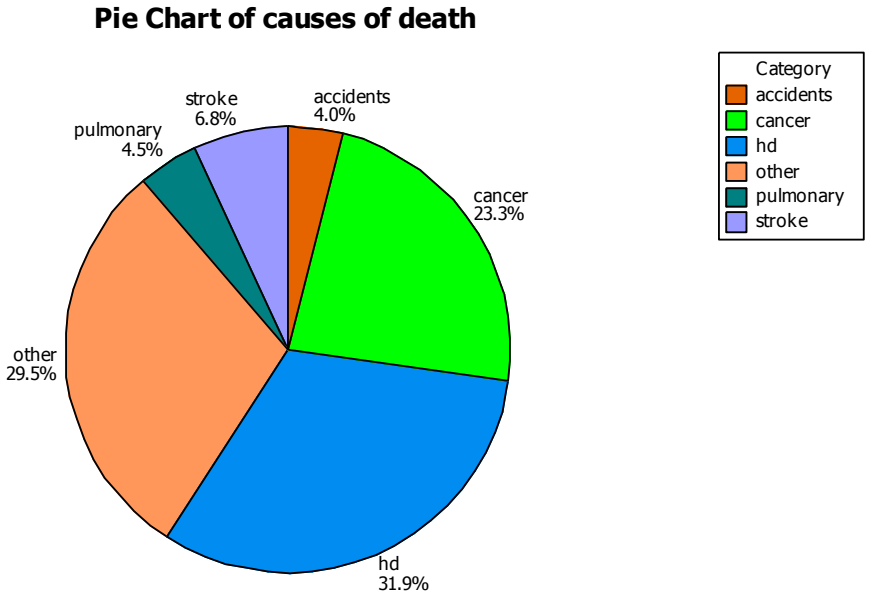
Figure 2.3.10 Features to look for in histograms and stem-and-leaf plots.

GRAPHICAL SUMMARIES: qualitative variables

Example. **According to the National Center for Health Statistics, the 6 leading causes of death in 1995 are: heart disease, cancer, stroke, pulmonary diseases, accidents, and others.**



Cause of death	Count (k)	percent
heart diseases	738	31.92
cancer	538	23.27
stroke	158	6.83
pulmonary diseases	103	4.46
accidents	93	4.02
others	682	29.5
All causes	2,312	100



Sample Statistics and Population Parameters

- **A numerical summary of a sample is called a statistic.**
- **A numerical summary of a population is called a parameter.**
- **Statistics are often used to estimate parameters.**

Descriptive Statistics

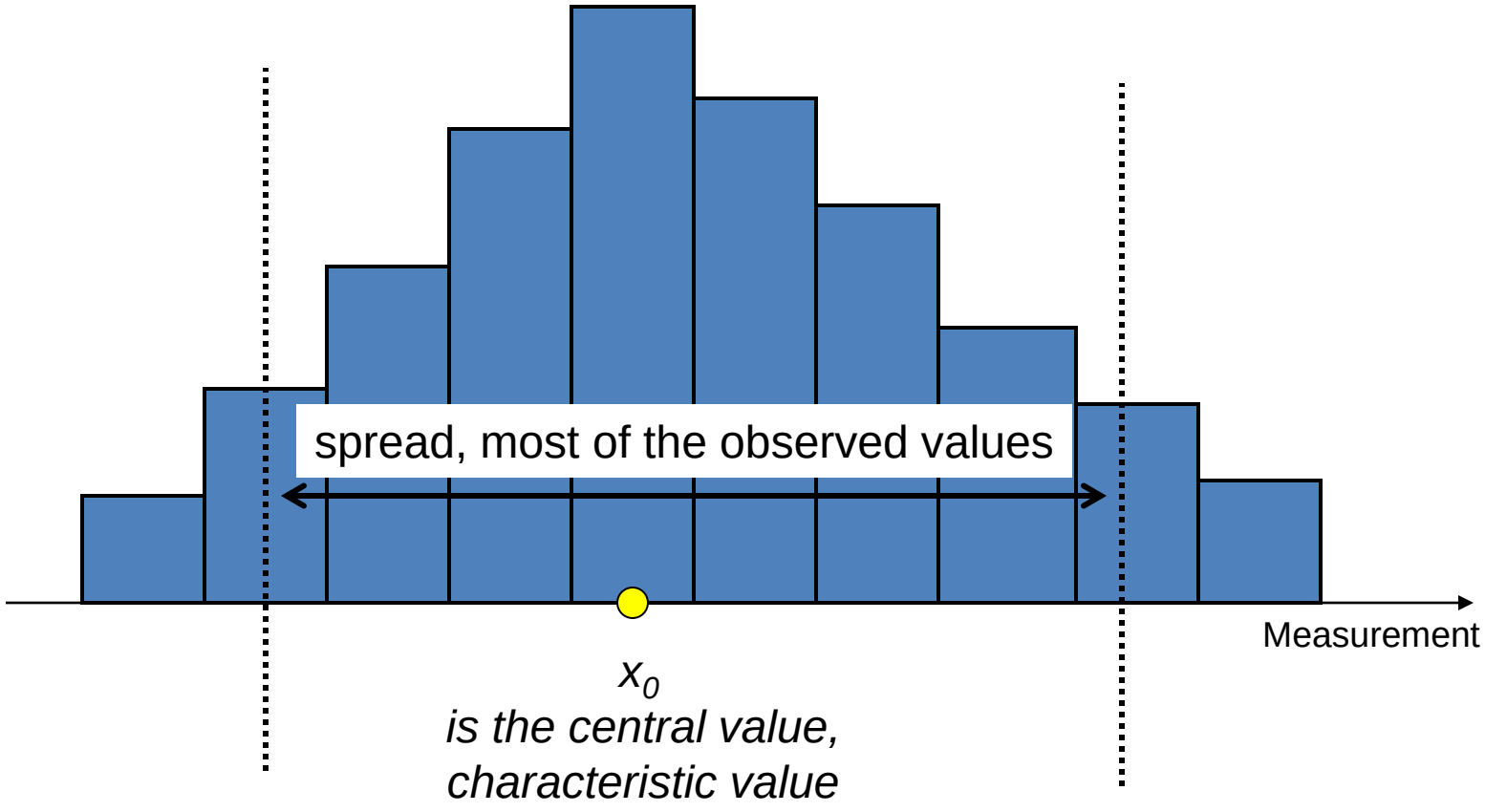
Descriptive Statistics

summarize or describe the important characteristics of a known set of data

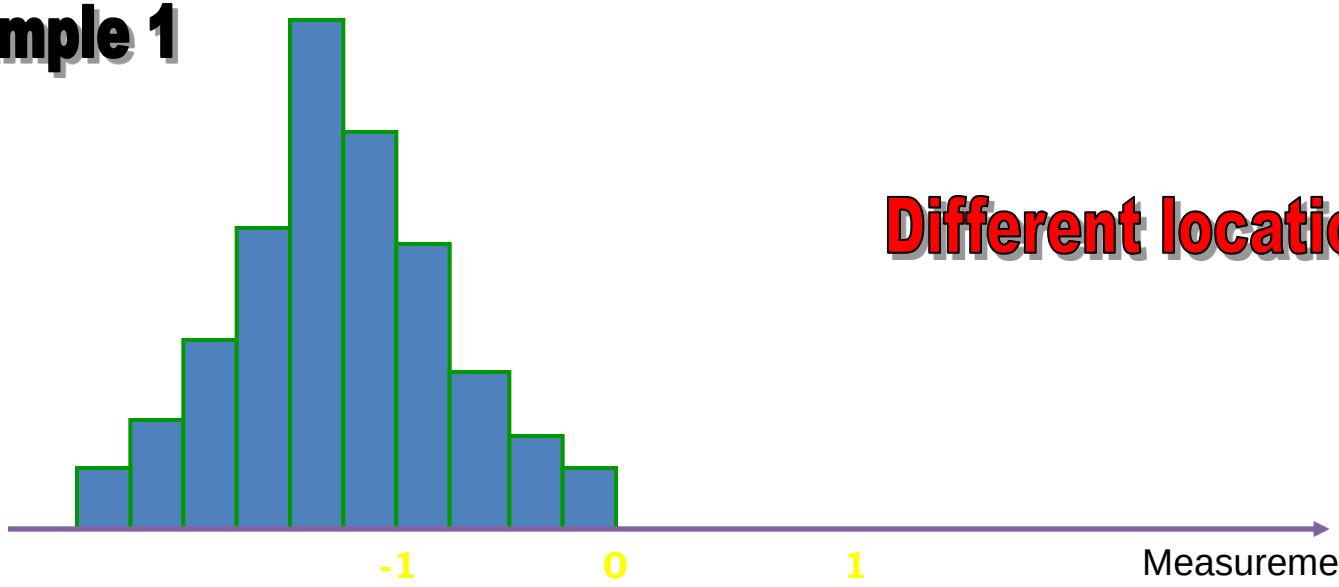
Inferential Statistics

use sample data to make inferences (or generalizations) about a population

Numerical Summaries:
Central Value and Spread/Variability/Dispersion

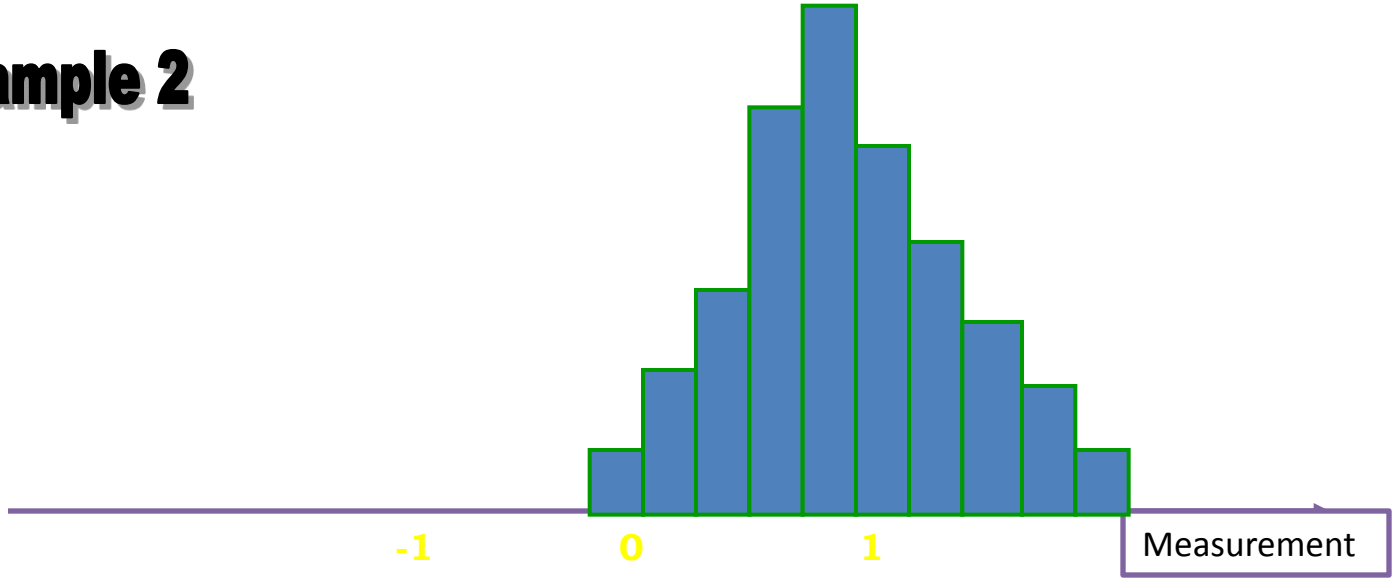


Sample 1

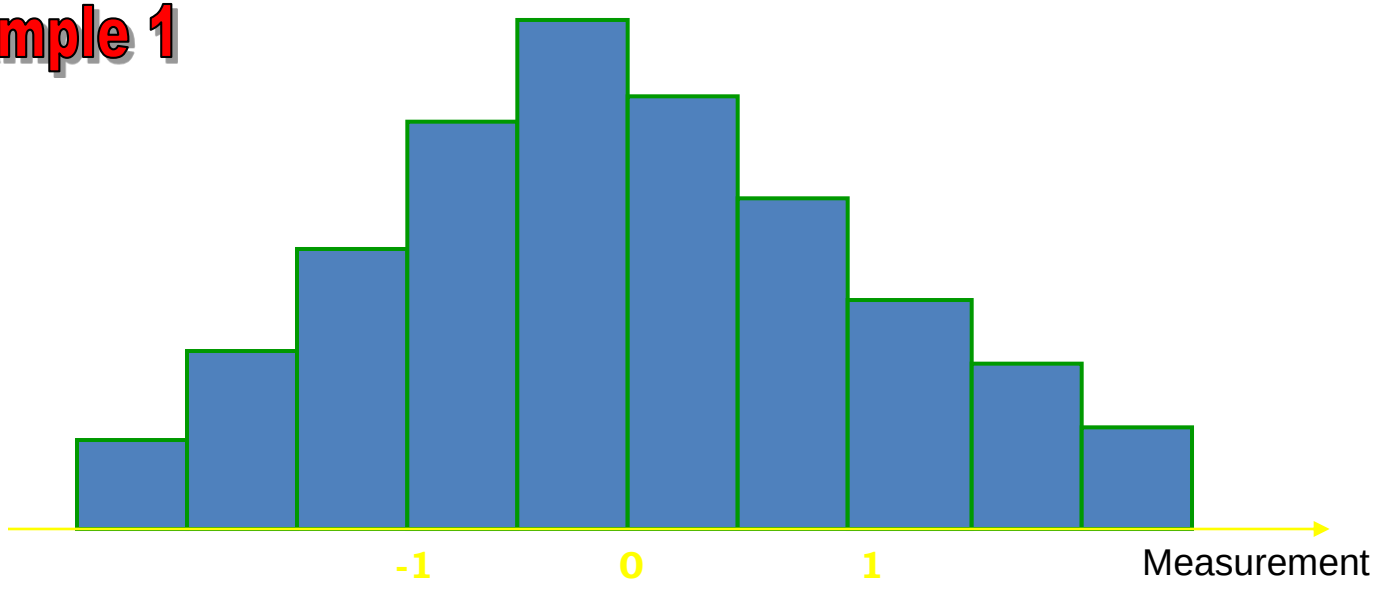


Different locations!

Sample 2

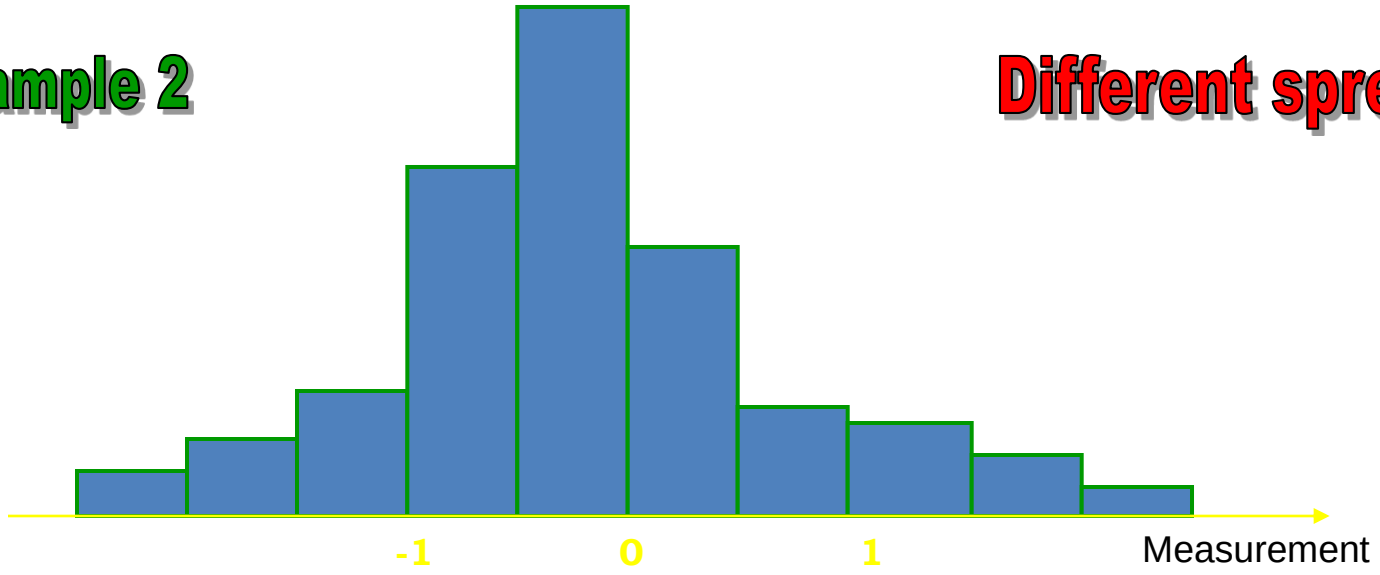


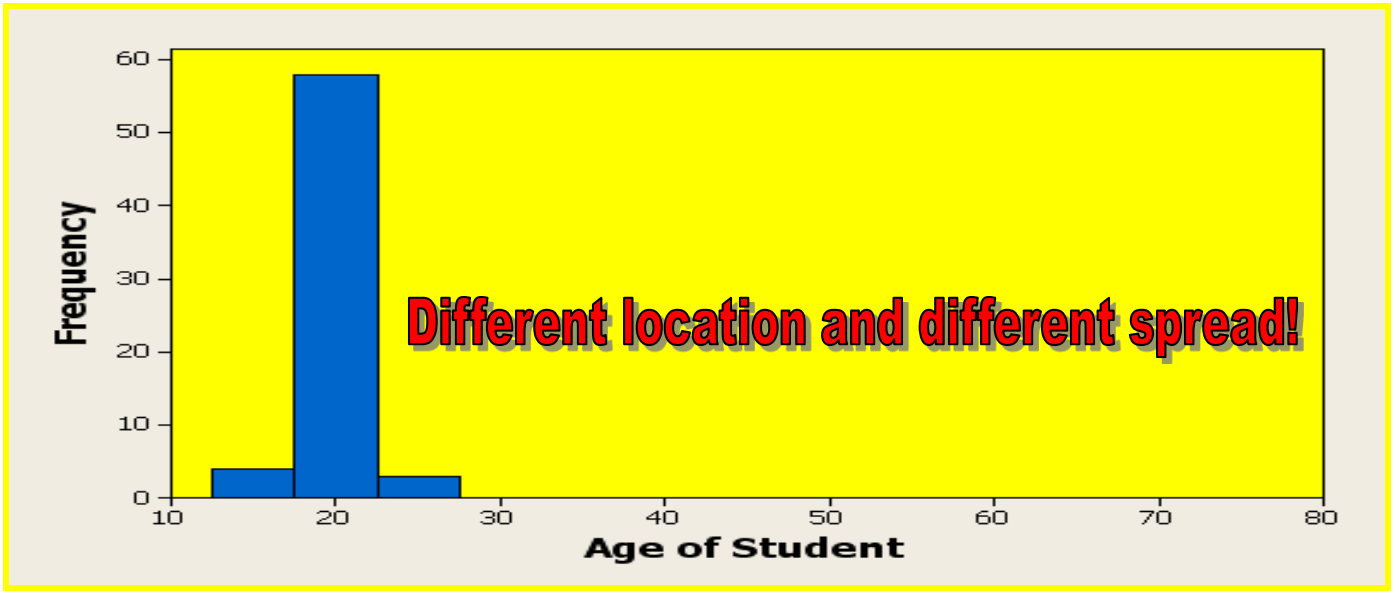
Sample 1

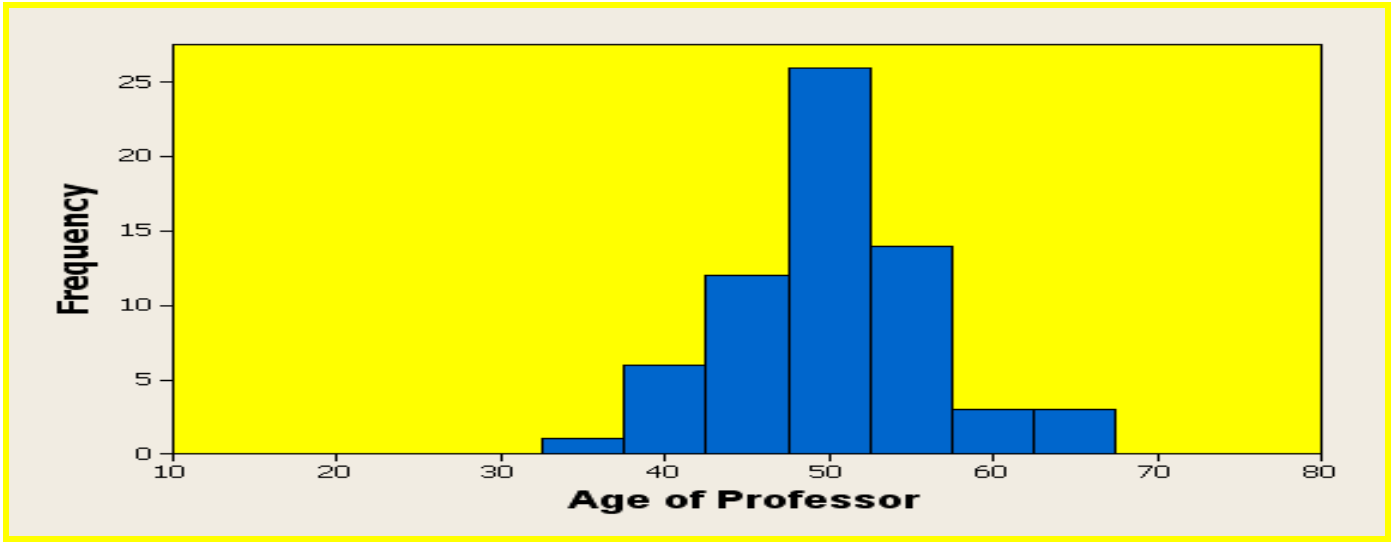


Sample 2

Different spread!







Data from “Smart Student” Questionnaire

Quantitative measures of location and spread

Measuring Center of the Data Set

- **Example. You want to buy a 3-4 bedroom (5-6 room) house. Need info on real estate sales in Reno, say on 200 recently sold 3-4bdrm houses.**
- **Data: $325,300, $287,650, $589,900, $230,900, …, $455,800.**
- **Q: What is the "average" selling price for a 3-4 bdrm house?**

  What does AVERAGE mean?

- **Most common? Most frequent? Mode**
- **Dividing selling prices in half, i.e. half are lower and higher that the "average"? Median**
- **Arithmetic average of all selling prices. Mean**

Most Common Summary Statistics: measuring location/center and spread/variability of the data

Let $X_1, \ldots, X_n$ **be a sample.**

- **Sample Mean- measure of center:** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- **Sample Variance- measure of variability:**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

- **Sample standard deviation is the square root of the sample variance.**

More on Summary Statistics: effect of rescaling and shifting data.

- If $X_1, \ldots, X_n$ is a sample, and $Y_i = a + bX_i$,where *a* and *b* are constants, then

$$\bar{Y} = a + b\bar{X}$$

- If $X_1, \ldots, X_n$ is a sample, and $Y_i = a + bX_i$,where *a* and *b* are constants, then

$$s_y^2 = b^2 s_x^2, \text{ and } s_y = |b| s_x.$$

Examples of use: Changing scales from Fahrenheit to Celsius, or from meters to feet, etc.

Summary Statistics: Median

SAMPLE MEDIAN is the "middle value" when the data is arranged in an increasing (or decreasing) order. Equal numbers of observations are larger and smaller than median.

- **SORT the data**, then
- Odd number of observations - the median is the middle observation.

- Even number of observations - the median is the average of the two middle values.

Example. Quiz scores: 8, 5, 7, 3, 7. Find median of the quiz scores.

Step 1. Sort the data: 3, 5, 7, 7, 8.
Step 2. Even or odd n? Odd.
Step 3. Median is the middle observation =7.

Let's add an observation: New quiz data: 8, 5, 7, 6, 3, 7. Find median of the quiz scores.

Step 1. Sort the data: 3, 5, 6, 7, 7, 8.
Step 2. Even or odd n? Even.
Step 3. Median is the average of the two middle observations. (6+7)/2=6.5=median.

Summary Statistics: Mode

SAMPLE MODE is the most frequent value in the data set.

Example. Quiz scores: 8, 5, 7, 3, 7. Find mode of the quiz scores.

Answer: Mode is 7 because 7 is most frequent.

- **Note that mode may not always be unique. Why?**

Example: 1,2,4, 1, 5, 2, 7. Mode: 1 and 2: bimodal data

- **Note, that mode does not always exists.**

Example: 1, 2, 5, 7, -1, 9, 3. No mode, all observations are different.

Summary statistics: Range

RANGE is the difference between the largest and the smallest observations.

Range = maximum value – minimum value

The more variability or spread is in the data, the larger the difference between the min and max, the larger the range.

Measures of Relative Standing

Q: An average quiz score in math was 85 with a standard deviation of 5. An average quiz score in physics was 60 with a standard deviation of 1. A student got 97 in math and 65 in physics. Which score is better?

Measures of Relative Standing and

Exploratory Data Analysis (EDA)

Problem: **The average weekly sales of a small company are $10,000 with a standard deviation of $450. This week their sales were $9050. Is this week unusually low?**

Measures of relative standing are used to:

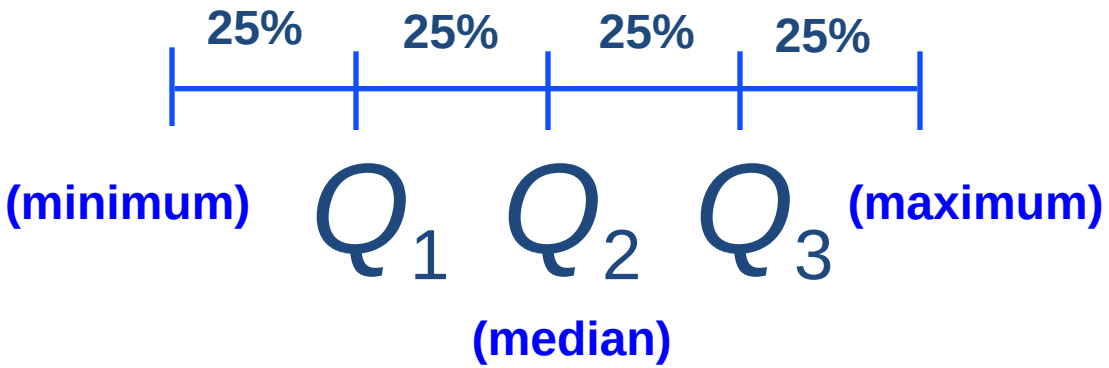
- **compare values from different data sets, or**
- **compare values within the same data set.**

Measures of relative standing: percentiles.

Measures of Relative Standing: Quartiles and Percentiles

- **Q_1 (First Quartile) separates the bottom (smallest) 25% of sorted values from the top (largest) 75%.**
- **Q_2 (Second Quartile) same as the median; separates the bottom (smallest) 50% of sorted values from the top (largest) 50%.**
- **Q_1 (Third Quartile) separates the bottom (smallest) 75% of sorted values from the top (largest) 25%.**

divide ranked/sorted scores into four equal parts



- **Interquartile Range (or IQR): $Q_3 - Q_1$**

Percentiles

99 percentiles denoted P_1, P_2, . . . P_{99}, partition/divide the data into 100 groups.

Finding the Percentile of a Given Score

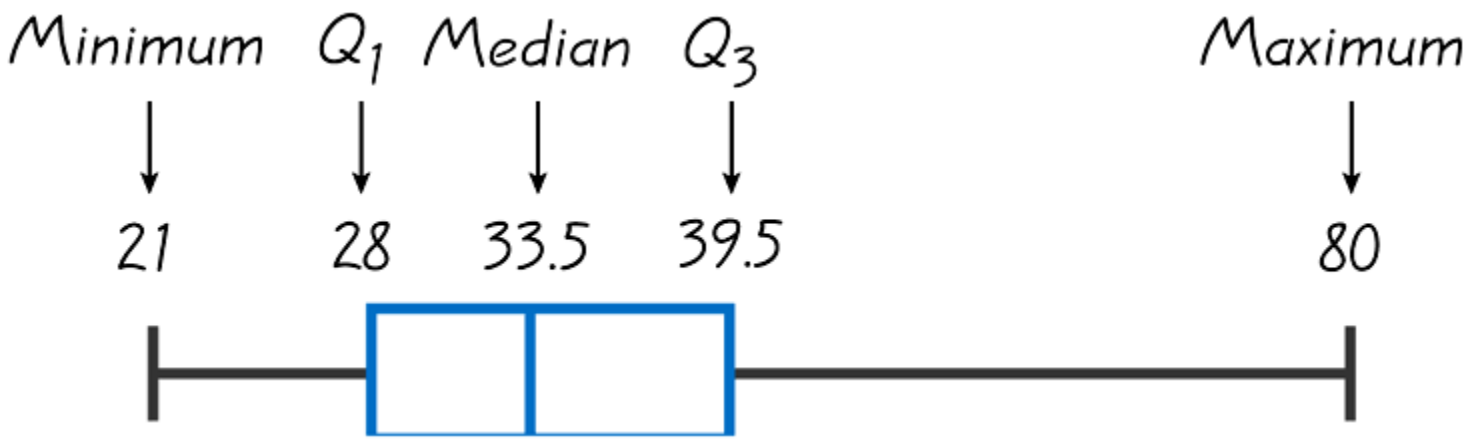
$$\text{Percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

5-number summary

- The 5-number summary of a data set consists of (1) the minimum value; (2) Q_1; (3) the median (Q_2); (4) Q_3; and (5) the maximum value.

- **A boxplot (or box-and-whisker-diagram) is a graph of a data set that visualizes the 5-point summary.**

Boxplot of Ages of Best Actresses



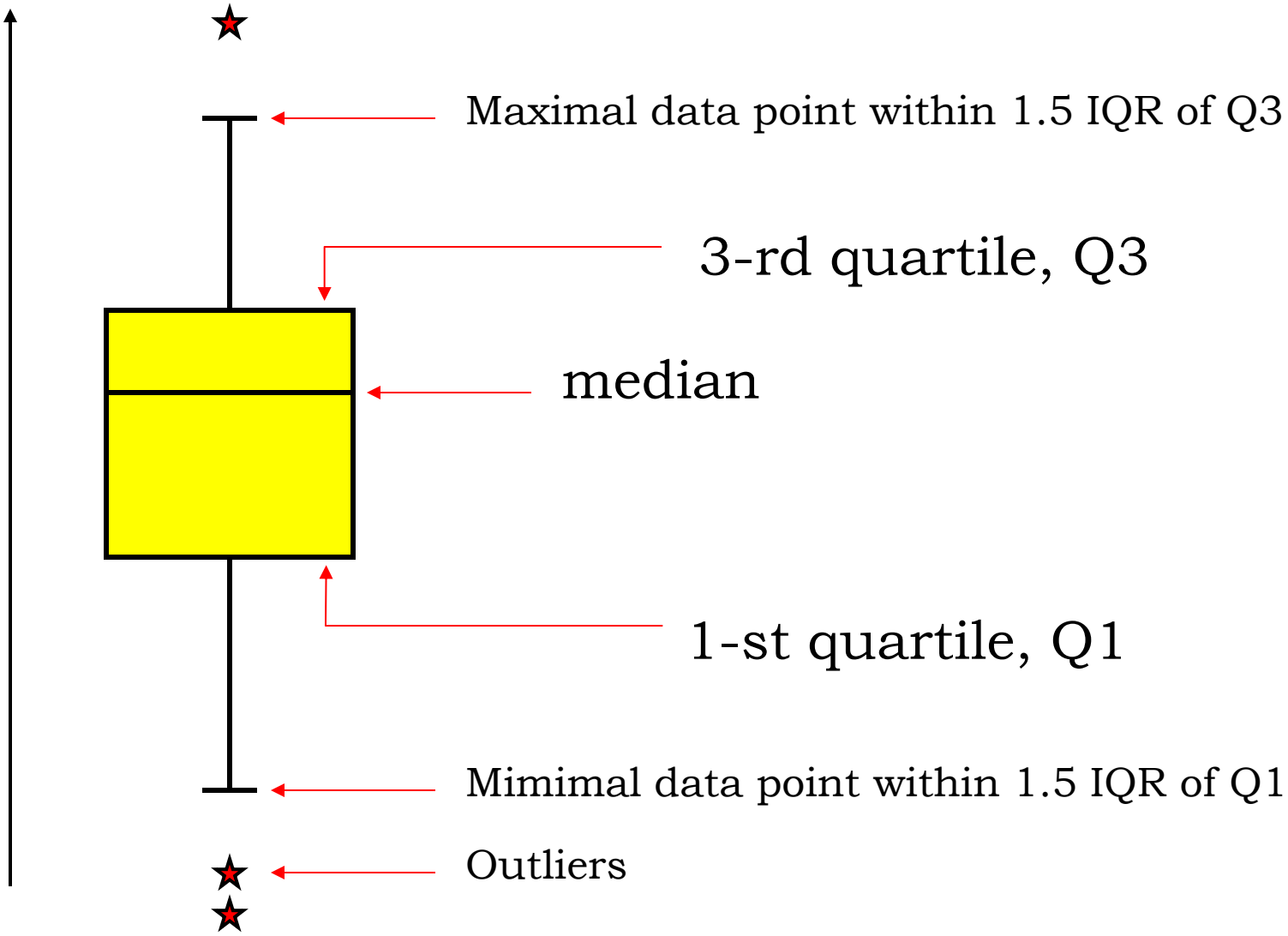
Boxplots

- **A boxplot is a graphic that presents the median, the first and third quartiles, and any outliers present in the sample.**

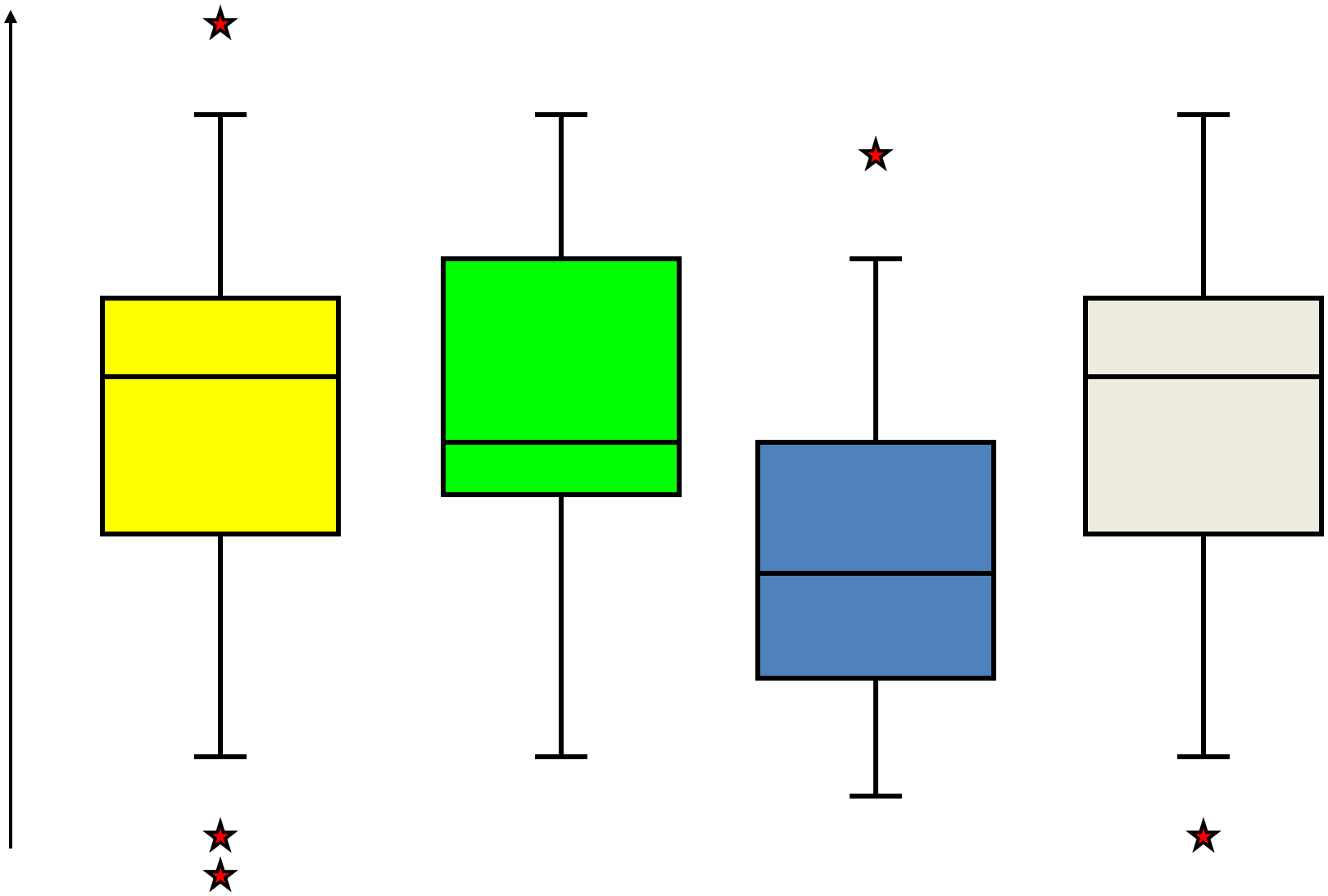
- **The interquartile range (IQR) is the difference between the third quartile and the first quartile. This is the distance needed to span the middle half of the data.**

- **Points more than 1.5 IQR above the third quartile, or more than 1.5 IQR below the first quartile are designated as outliers. Outliers are ploted individually.**

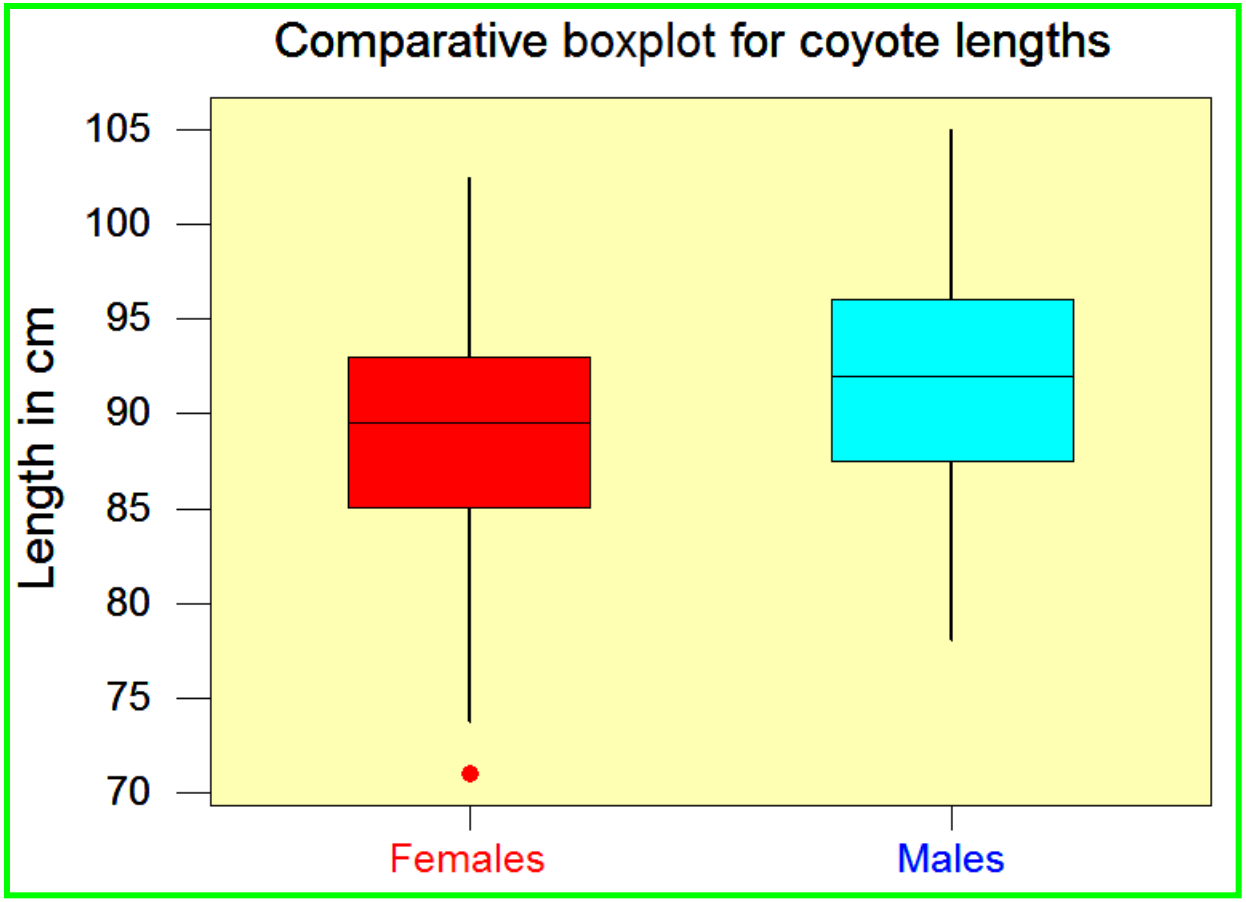
Anatomy of a Boxplot



Plots: comparative boxplot



Simple plots: comparative boxplot



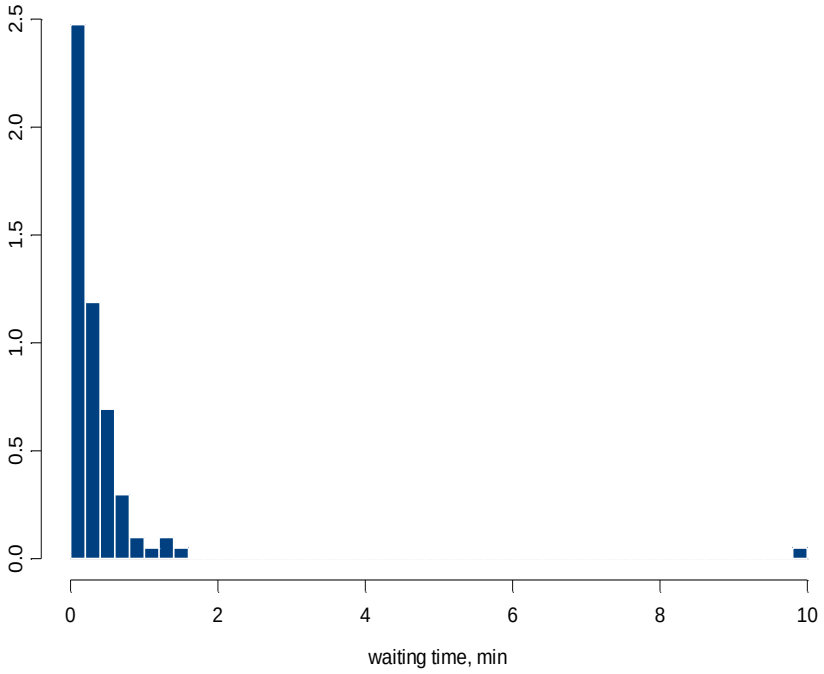
OUTLIERS

OUTLIERS – observations FAR outside the regular pattern of the data.

Where do outliers come from?

- **Errors of measurements or recording – in those cases, people tend to disregard them.**
- **Natural order of things – they point to very important phenomena like floods, heat waves, hurricanes, etc. Should not be discarded but studied.**

Example. Waiting times (in minutes) for a bus, 100 observations.



OUTLIERS AND MEASURES OF CENTER

Example. Take quiz scores. 3, 5, 7, 7, 8. Suppose I made a recording error and instead of 8 recorded 88. New data: 3, 5, 7, 7, 88.

New median= 7 = old median NO change,

New mode = 7 = old mode NO change,

New mean = (88+5+7+3+7)/5=22 LARGE change, old mean=6.

MEDIAN AND MODE are RESISTANT (ROBUST) TO OUTLIERS i.e. do not change if we add outliers to a data set.

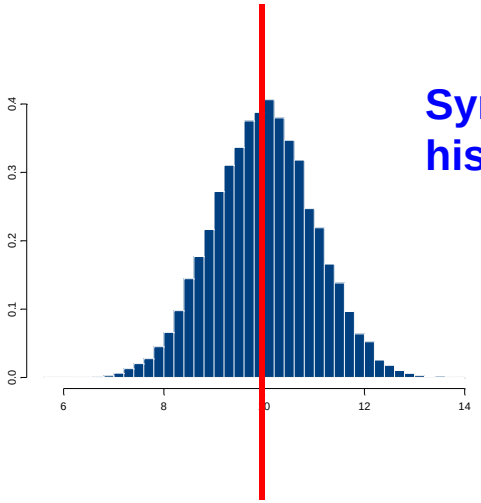
MEAN IS SENSITIVE TO OUTLIERS i.e. changes if we add an outlier to a data set.

Summary: If you do not want the very large or the very small (outliers) observations to affect the information you are getting about the "center" of the data ask for median rather than the mean.

Symmetry and Skewness of a Distribution

Symmetric

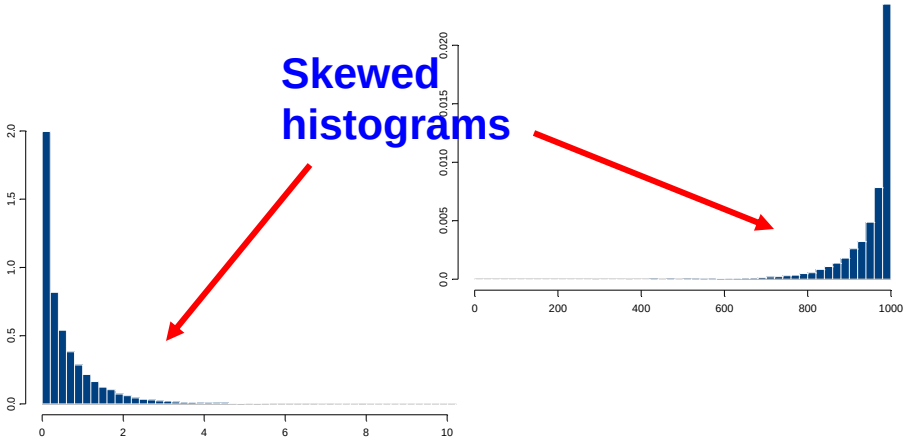
distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half



Symmetric histogram

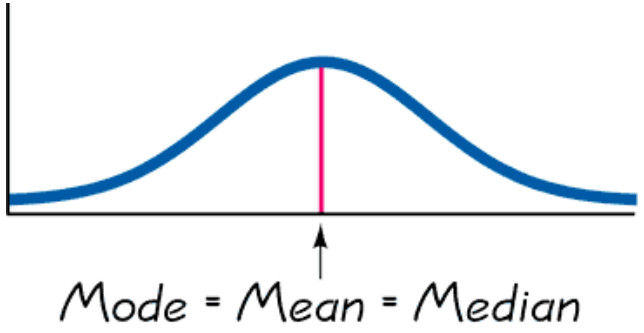
Skewed

distribution of data is skewed if it is not symmetric and if it extends more to one side than the other



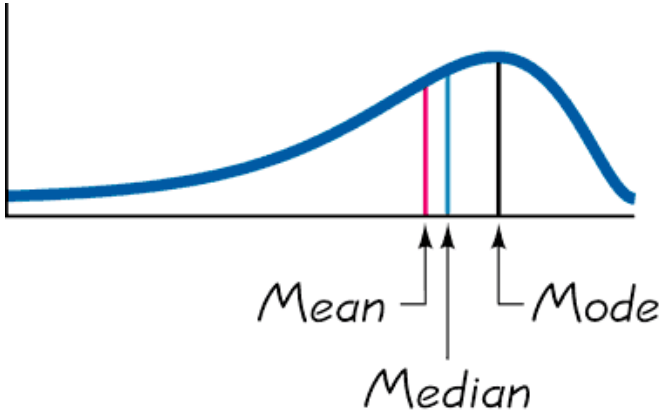
Skewed histograms

Skewness



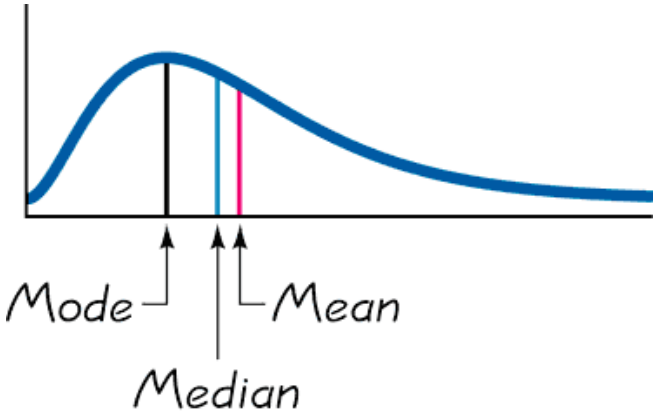
MEAN = MEDIAN

(b) Symmetric



(a) Skewed to the Left (Negatively)

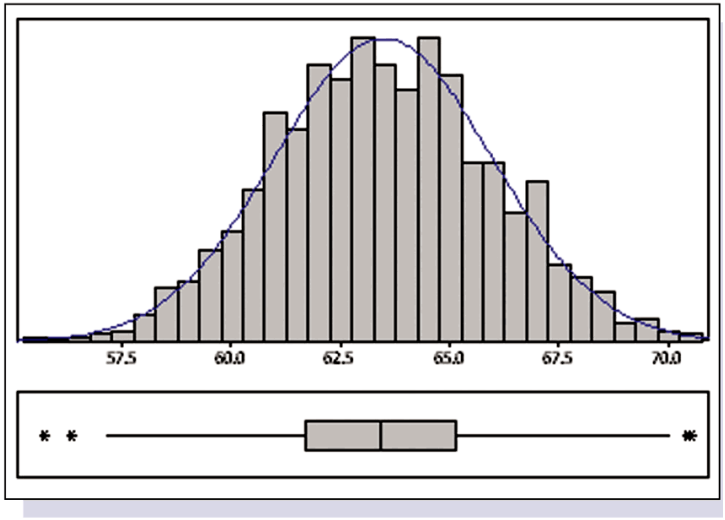
MEAN < MEDIAN



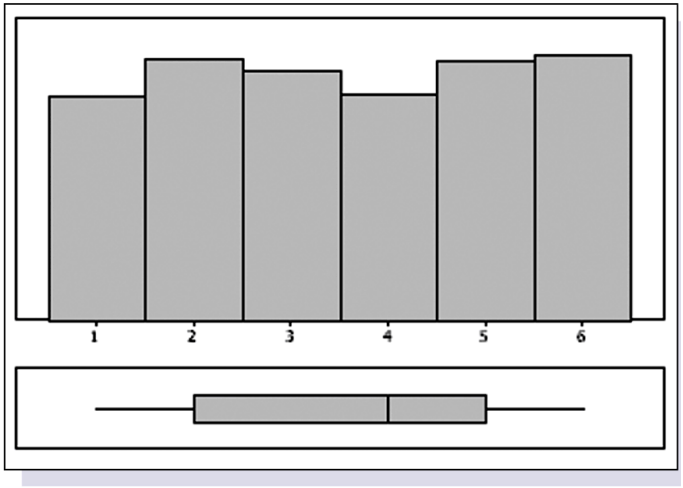
(c) Skewed to the Right (Positively)

MEAN > MEDIAN

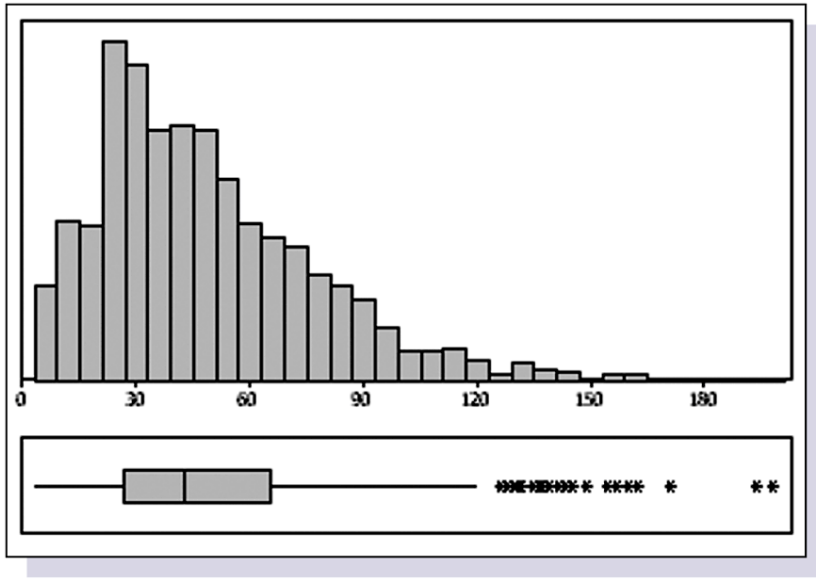
Boxplots , histograms and symmetry



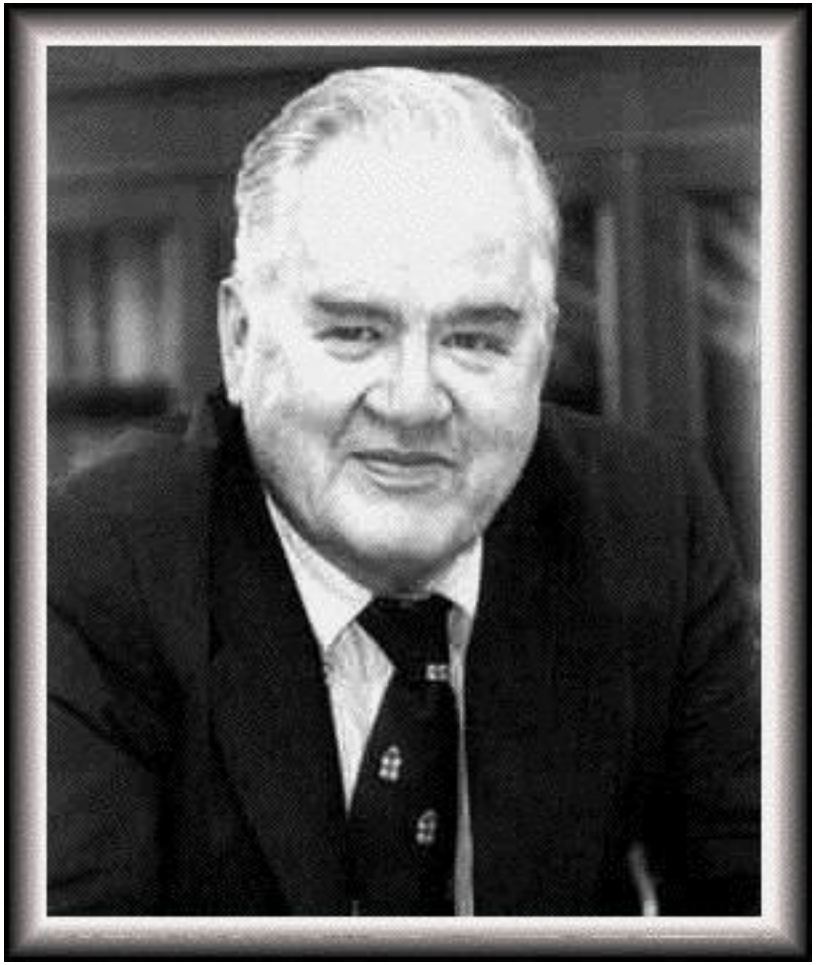
(a) Normal (bell-shaped) distribution
1000 heights (in.) of women



(b) Uniform distribution
1000 rolls of a die



(c) Skewed distribution
Incomes (thousands of dollars) of 1000 statistics professors



John Wilder Tukey

(June 16, 1915 - July 26, 2000)