



IT302 PROBABILITY AND STATISTICS

Lec-1

Dr. Anand Kumar M

Department of Information Technology
National Institute of Technology-Karnataka (NITK)
Surathkal, Mangalore.

Outline

- Statistics - Data Collection
- Population Sample
- Descriptive and Inferential Statistics
- Types of Variables
- Describing Dataset - Descriptive
 - Freq Table – Relative Freq Table – Data Grouping – Ogives – Stem and Leaf
- Summarizing Dataset
 - Mean – Median-Mode
 - Variance and SD
 - Percentiles and Box

Statistics - Data Collection

- Data is important to learn about something.
- “Statistics is the art of *Learning from Data* “
- *Describe-Analyze-Summarize* the data.
- Data is Not Available
 - Stat Theory used to design appropriate Experiment to generate data
 - *Example* – Class – Online/Offline
 - Not to be biased (Random)

Population

- The **entire group** of elements is called the *population*.
- For example, a researcher may be interested in the relation between class size (variable 1) and academic performance (variable 2) for the population of third-grade children.

Sample

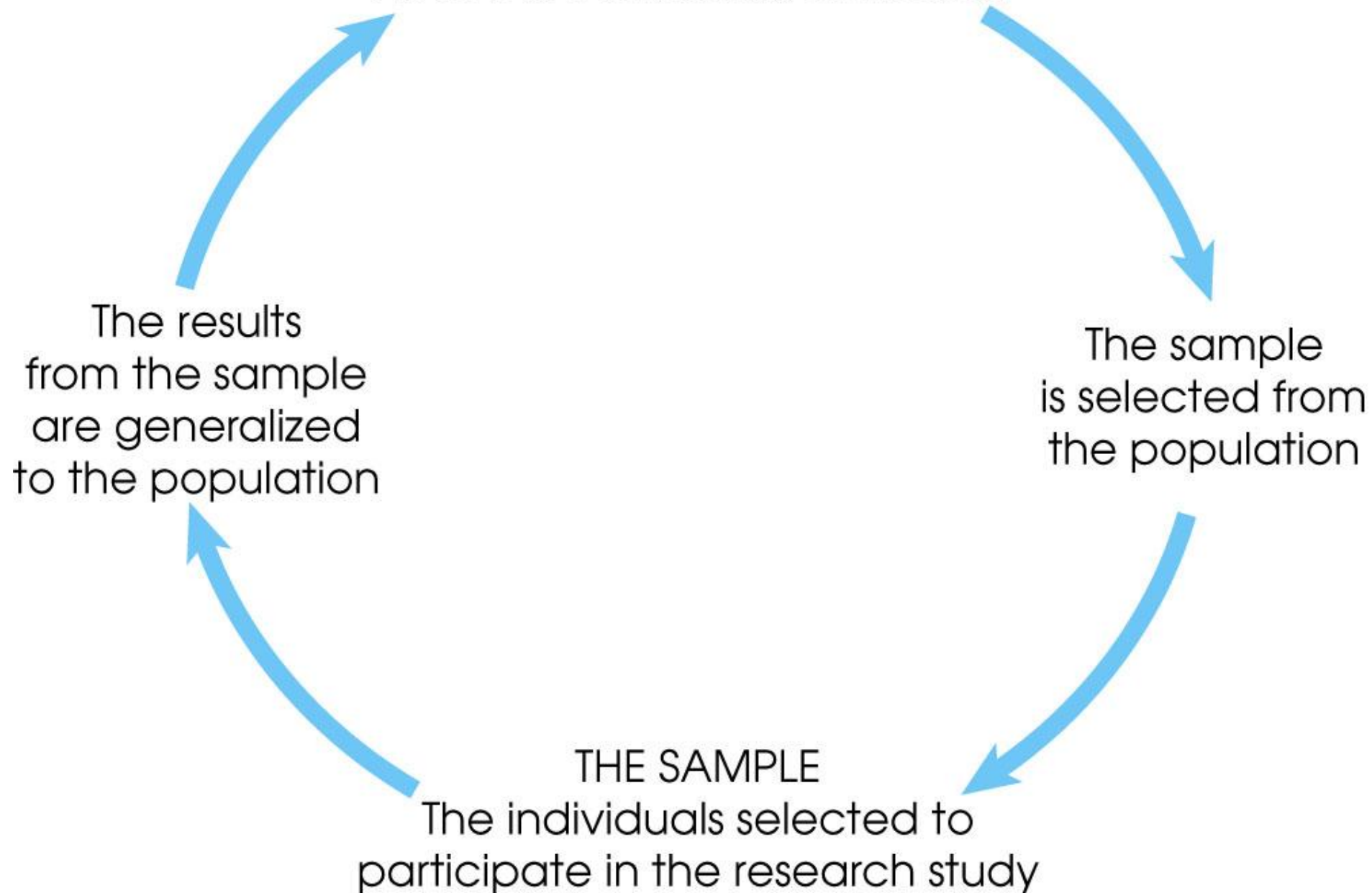
- Usually populations are so large that a researcher **cannot examine the entire group**.
- Therefore, a **sample** is selected to represent the population in a research study.
- The goal is to use the *results obtained from the sample to help answer questions about the population*.

THE POPULATION
All of the individuals of interest

The sample
is selected from
the population

THE SAMPLE
The individuals selected to
participate in the research study

The results
from the sample
are generalized
to the population



Descriptive and Inferential Statistics

- End of Experiment data need to be described and summarized, i.e DS.
- Method concerned about drawing conclusions are IS.
- Ex: With 10 toss – 7 FAIR OR NOT
- With 100 toss- 92 FAIR OR NOT
- Assumptions required to draw conclusion – the assumptions are called Probability Model

Descriptive statistics

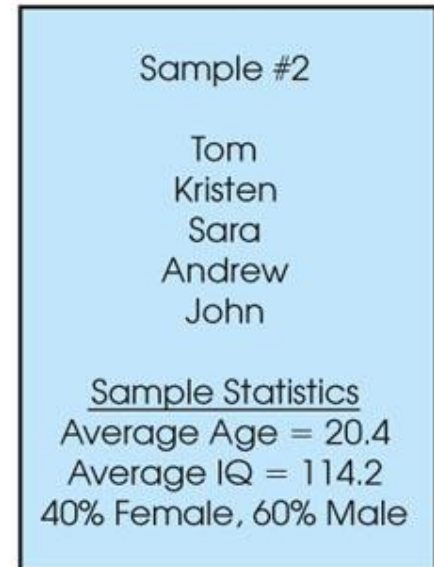
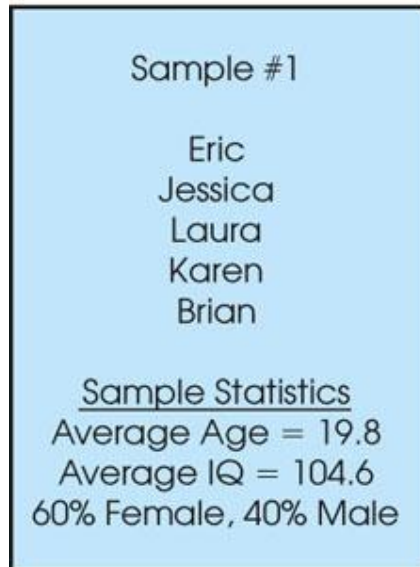
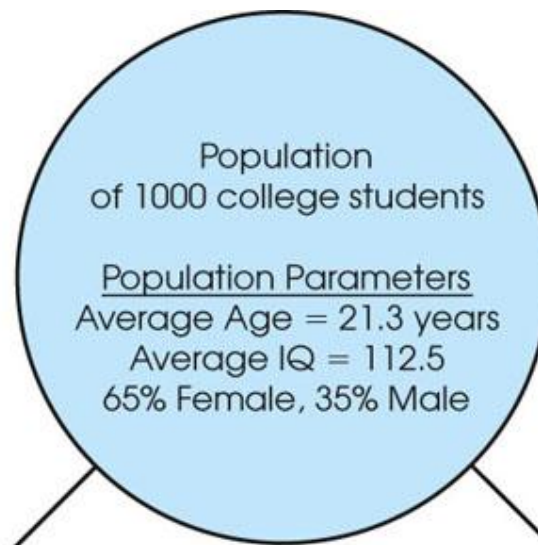
- **Descriptive statistics** are methods for organizing and summarizing data.
- For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.
- A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**

Inferential Statistics

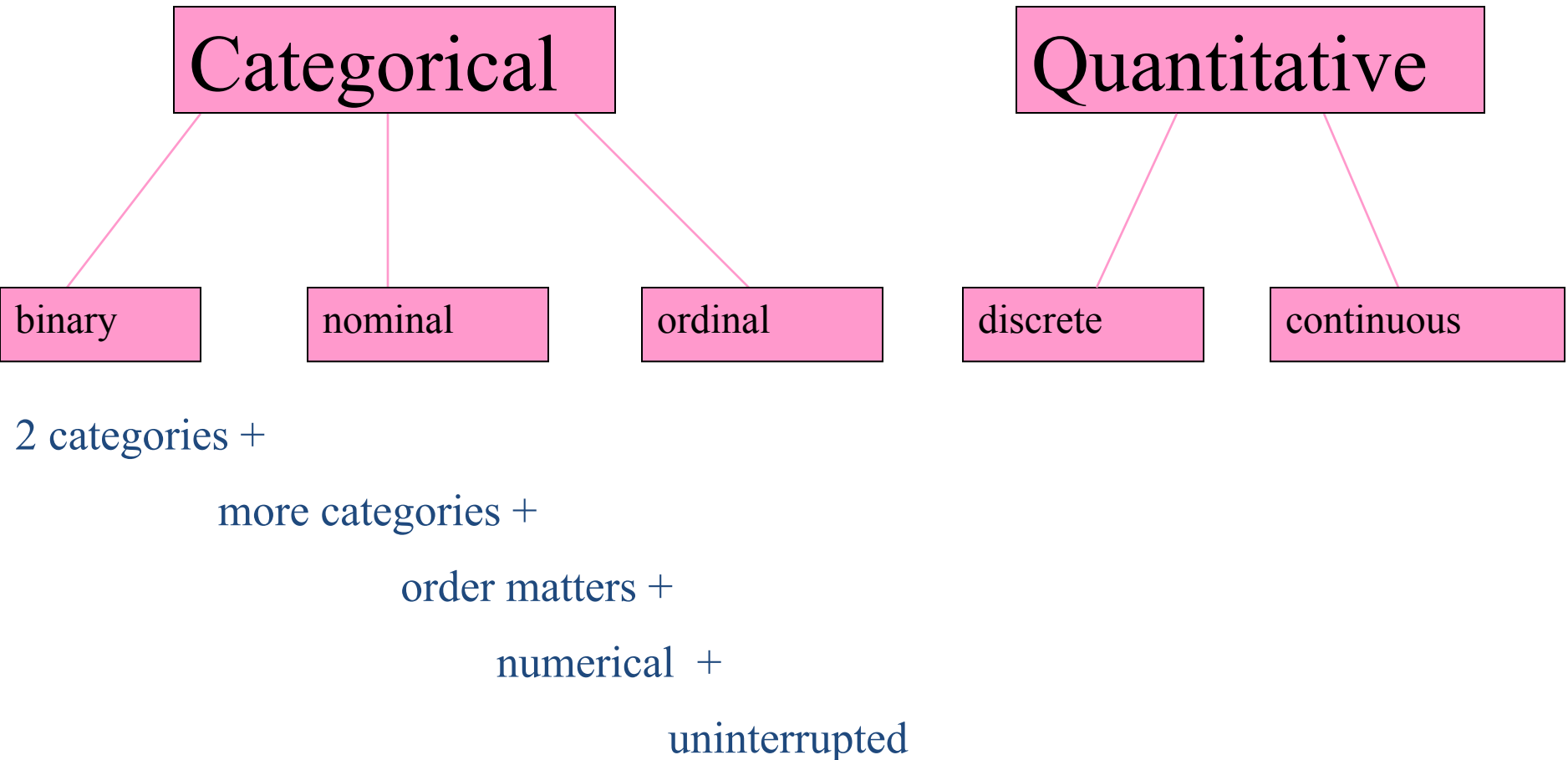
- **Inferential statistics** are methods for using sample data to make general conclusions (inferences) about populations.
- Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

Sampling Error

- The discrepancy between a sample statistic and its population parameter is called **sampling error**.
- Defining and measuring sampling error is a large part of inferential statistics.



Types of Variables: Overview



Categorical Variables

- Also known as “qualitative.”
- Categories.
 - treatment groups
 - exposure groups
 - disease status

Categorical Variables

- Dichotomous (binary) – two levels
 - Dead/alive
 - Treatment/placebo
 - Disease/no disease
 - Exposed/Unexposed
 - Heads/Tails
 - Pulmonary Embolism (yes/no)
 - Male/female

Categorical Variables

- Nominal variables – Named categories
Order doesn't matter!
 - The blood type of a patient (O, A, B, AB)
 - Marital status
 - Occupation

Categorical Variables

- Ordinal variable – Ordered categories. Order matters!
 - Staging in breast cancer as I, II, III, or IV
 - Birth order—1st, 2nd, 3rd, etc.
 - Letter grades (A, B, C, D, F)
 - Ratings on a scale from 1-5
 - Ratings on: always; usually; many times; once in a while; almost never; never
 - Age in categories (10-20, 20-30, etc.)
 - Shock index categories (Kline et al.)

Quantitative Variables

- Numerical variables; may be arithmetically manipulated.
 - Counts
 - Time
 - Age
 - Height

Quantitative Variables

- Discrete Numbers – a limited set of distinct values, such as whole numbers.
 - Number of new AIDS cases in CA in a year (counts)
 - Years of school completed
 - The number of children in the family (cannot have a half a child!)
 - The number of deaths in a defined time period (cannot have a partial death!)
 - Roll of a die

Quantitative Variables

- Continuous Variables - Can take on any number within a defined range.
 - Time-to-event (survival time)
 - Age
 - Blood pressure
 - Serum insulin
 - Speed of a car
 - Income
 - Shock index (Kline et al.)

Looking at Data

- ✓ How are the data distributed?
 - Where is the center?
 - What is the range?
 - What's the shape of the distribution (e.g., Gaussian, binomial, exponential, skewed)?
- ✓ Are there “outliers”?
- ✓ Are there data points that don't make sense?

The first rule of statistics:
USE COMMON SENSE!

90% of the information is contained
in the graph.

Frequency Plots (univariate)

Categorical variables

- Bar Chart

Continuous variables

- Box Plot
- Histogram

Describing Dataset – Descriptive

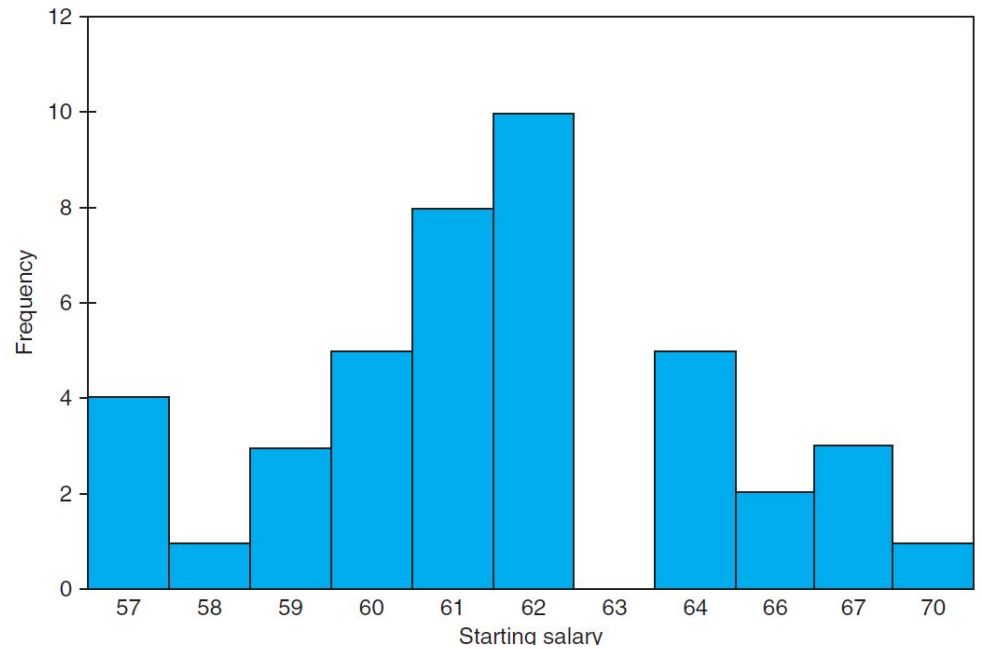
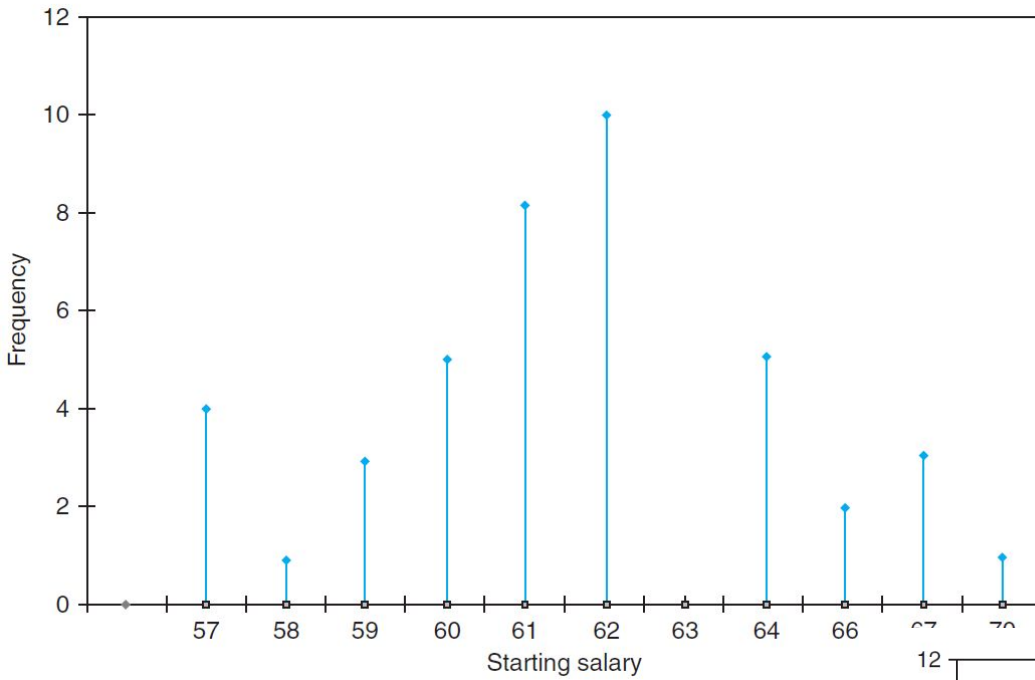
- Freq Table
- Relative Freq Table
- Data Grouping
- Ogives
- Stem and Leaf

Freq Table AND Graphs

TABLE 2.1 *Starting Yearly Salaries*

Starting Salary	Frequency
57	4
58	1
59	3
60	5
61	8
62	10
63	0
64	5
66	2
67	3
70	1

Relative Frequency Graphs



Polygon

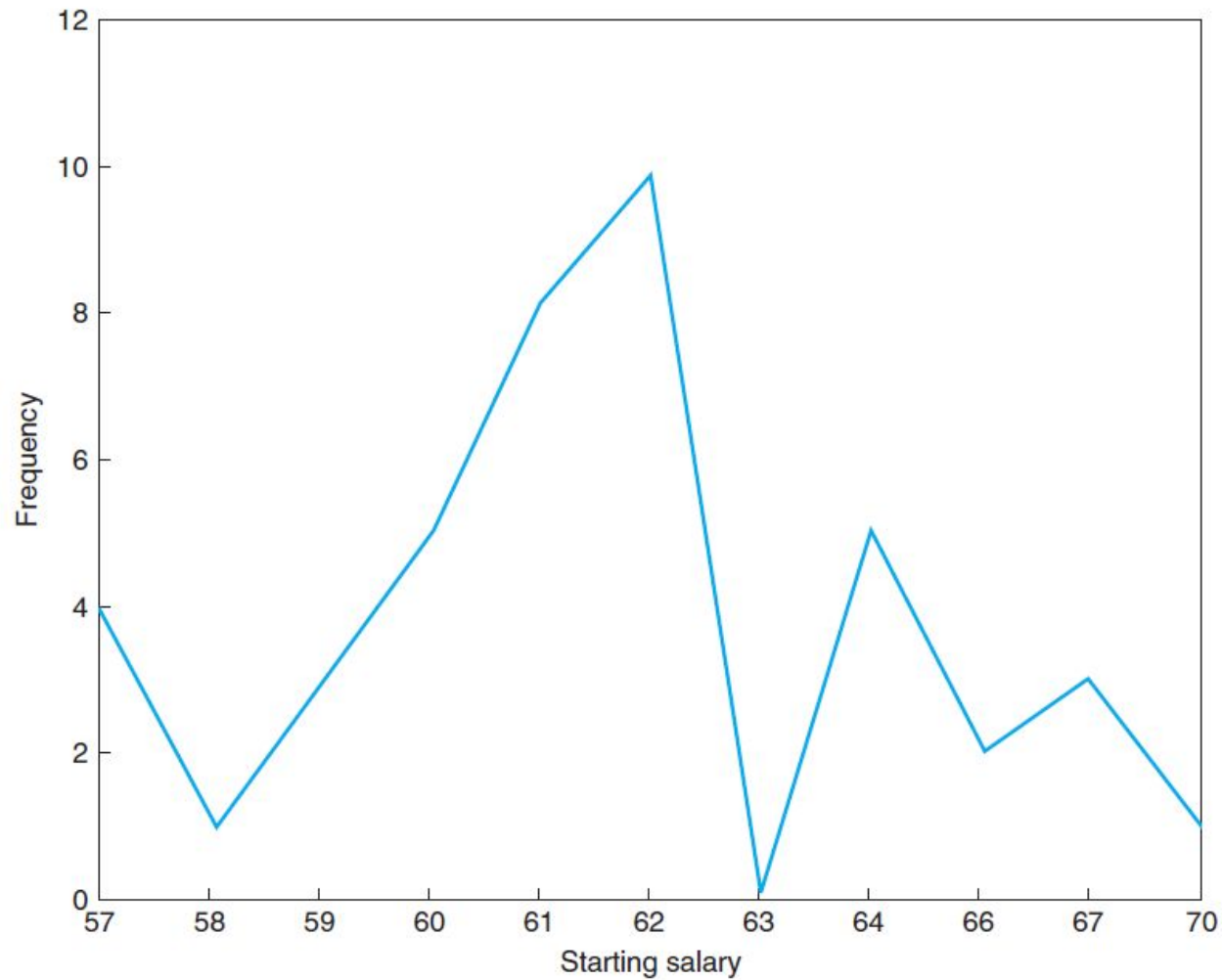
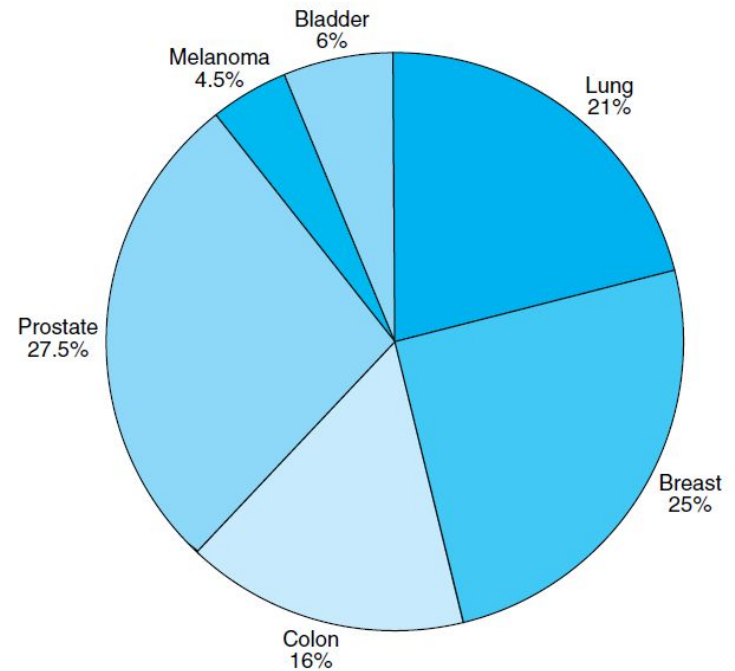


FIGURE 2.3 *Frequency polygon for starting salary data.*

TABLE 2.2

Starting Salary	Frequency
47	$4/42 = .0952$
48	$1/42 = .0238$
49	$3/42$
50	$5/42$
51	$8/42$
52	$10/42$
53	0
54	$5/42$
56	$2/42$
57	$3/42$
60	$1/42$



Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06

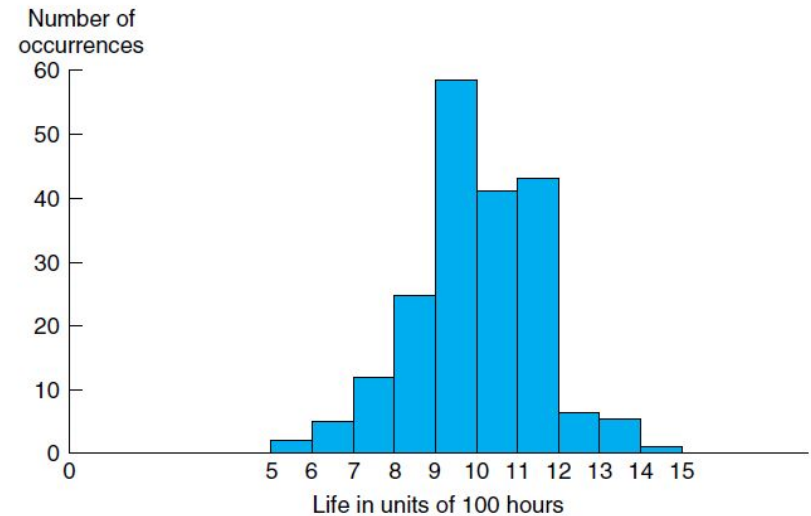
Data Grouping -Histograms

TABLE 2.3 *Life in Hours of 200 Incandescent Lamps*

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

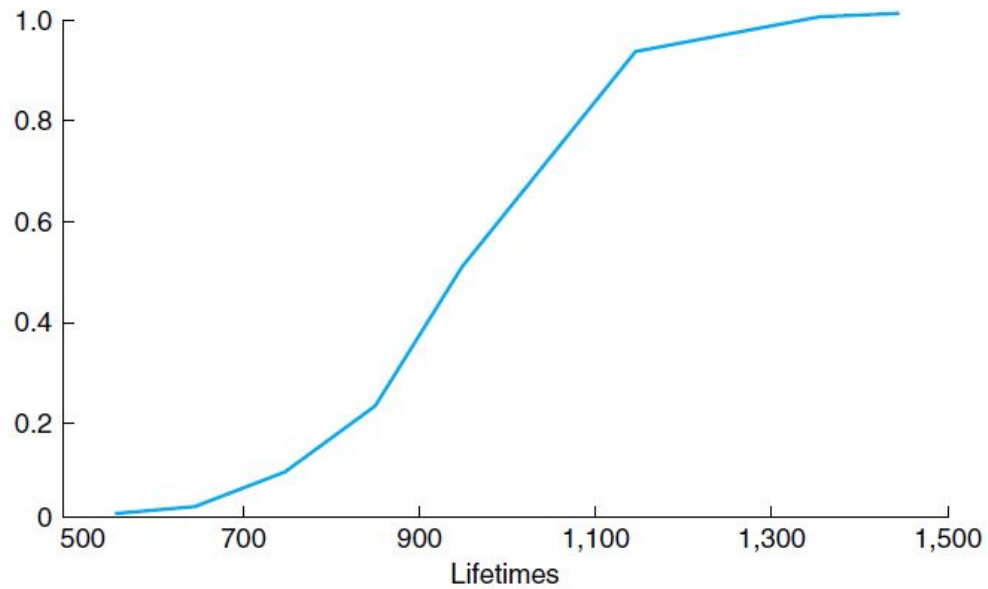
TABLE 2.4 *A Class Frequency Table*

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1




- The number of class intervals chosen should be a trade-off
- between (1) choosing too few classes at a cost of losing too much information about the
- actual data values in a class and (2) choosing too many classes, which will result in the
- frequencies of each class being too small for a pattern to be discernible

Ogives



Stem and Leaf

7		0.0
6		9.0
5		1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
4		0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
3		3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.5, 9.5
2		9.0, 9.8 

Summarizing the datasets

Sample Mean, Sample Median, and Sample Mode

- Statistics that are used for describing the center of a set of data values.
- Suppose that we have a data set consisting of the *n numerical values* x_1, x_2, \dots, x_n .
- *The sample mean is the arithmetic average of these values.*

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

Mean: example

Some data:

Age of participants: 17 19 21 22 23 23 23 38

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{17 + 19 + 21 + 22 + 23 + 23 + 23 + 38}{8} = 23.25$$

Mean

- $y_i = ax_i + b$

EXAMPLE 2.3a The winning scores in the U.S. Masters golf tournament in the years from 2004 to 2013 were as follows:

280, 278, 272, 276, 281, 279, 276, 281, 289, 280

Find the sample mean of these scores.

SOLUTION Rather than directly adding these values, it is easier to first subtract 280 from each one to obtain the new values $y_i = x_i - 280$:

0, -2, -8, -4, 1, -1, -4, 1, 9, 0

Because the arithmetic average of the transformed data set is

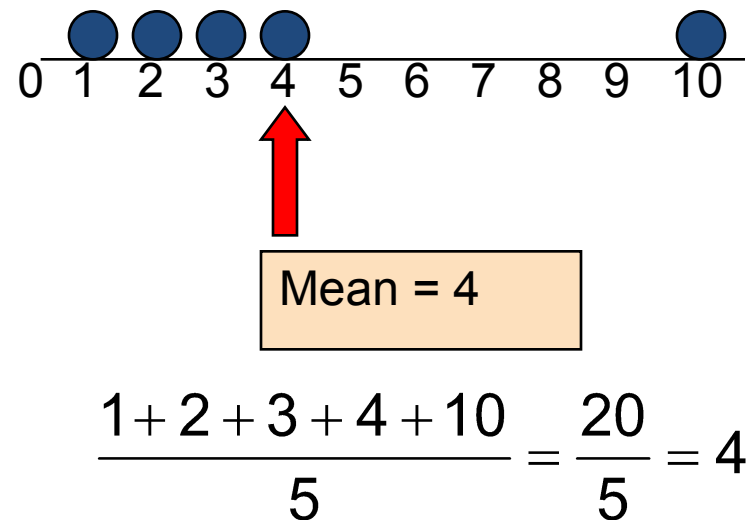
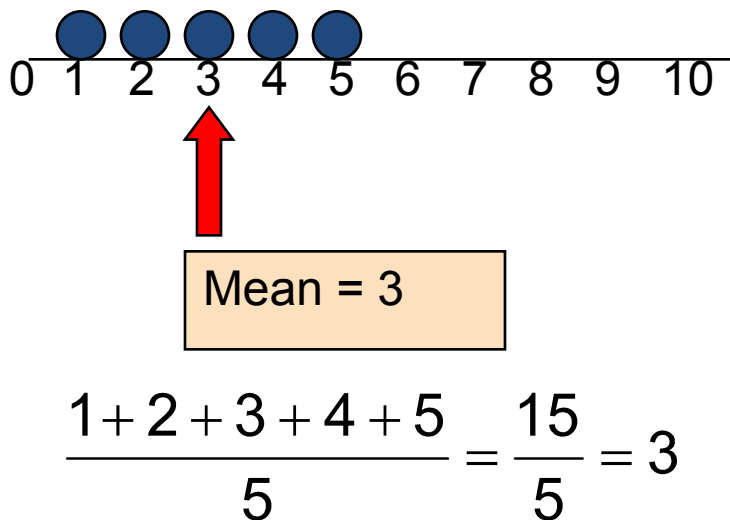
$$\bar{y} = -8/10$$

it follows that

$$\bar{x} = \bar{y} + 280 = 279.2 \quad \blacksquare$$

Mean

- The mean is affected by extreme values (outliers)



Mean?

Age	Frequency
15	2
16	5
17	11
18	9
19	14
20	13

Central Tendency

- Median – the exact middle value

Calculation:

- If there are an odd number of observations, find the middle value
- If there are an even number of observations, find the middle two values and average them.

Median: example

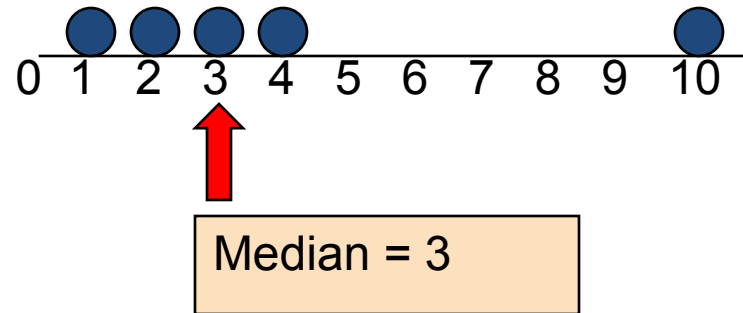
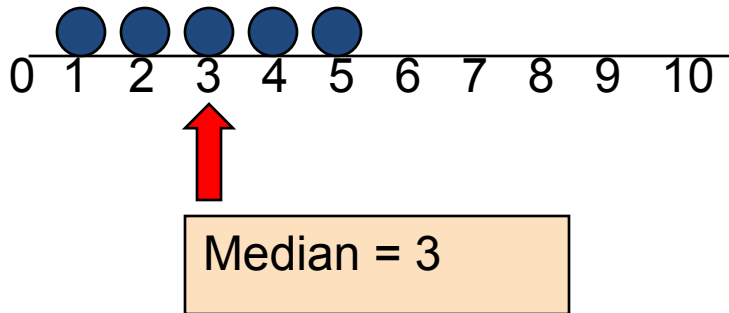
Some data:

Age of participants: 17 19 21 22 23 23 23 38

$$\text{Median} = (22+23)/2 = 22.5$$

Median

- The median is not affected by extreme values (outliers).



Central Tendency

- Mode – the value that occurs most frequently

Mode: example

Some data:

Age of participants: 17 19 21 22 23 23 23 38

Mode = 23 (occurs 3 times)

Value	Frequency
1	9
2	8
3	5
4	5
5	6
6	7

Practice

- Find the median of a series of all the even terms from 4 to 296.

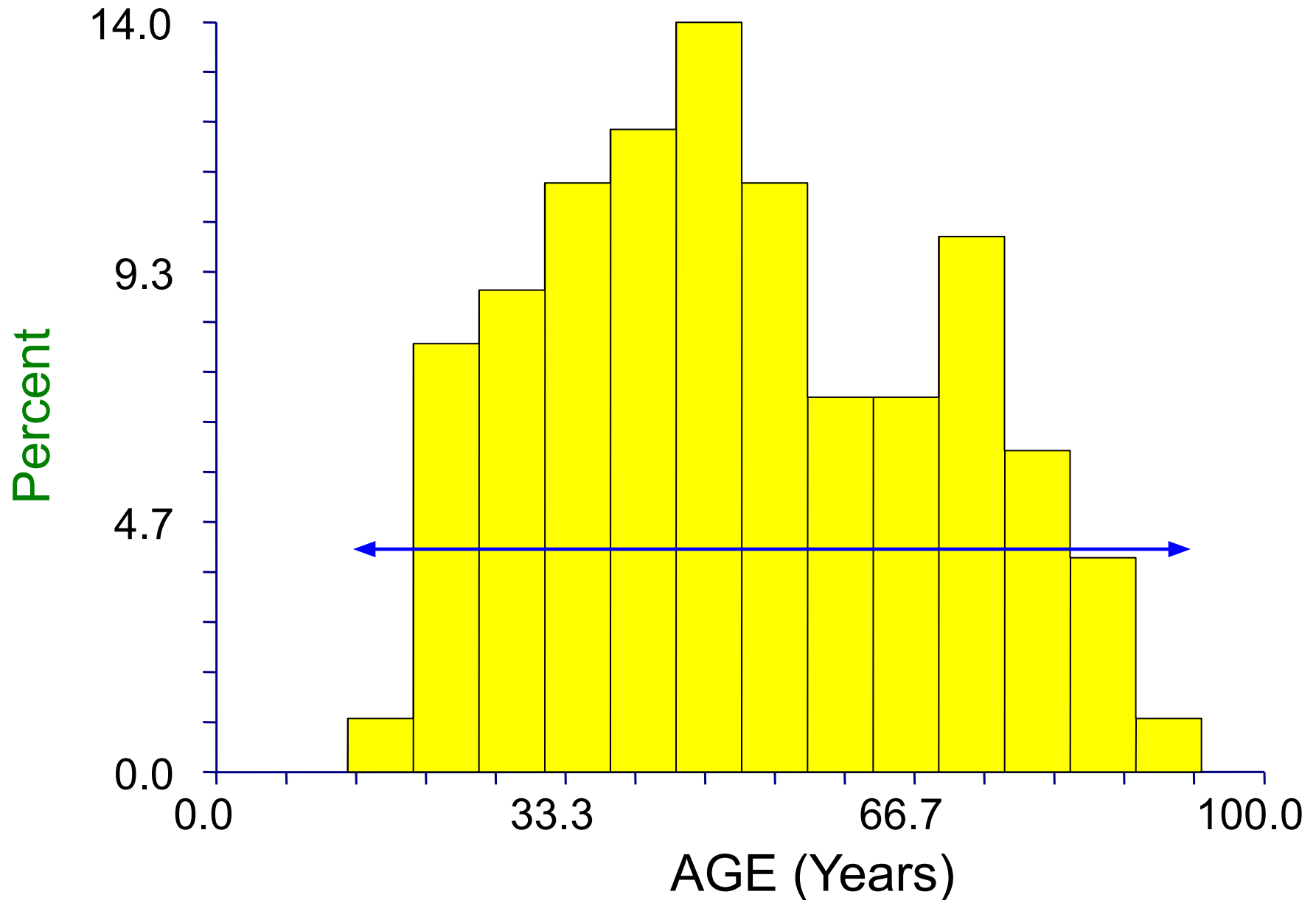
Measures of Variation/Dispersion

- Range
- Percentiles/quartiles
- Interquartile range
- Standard deviation/Variance

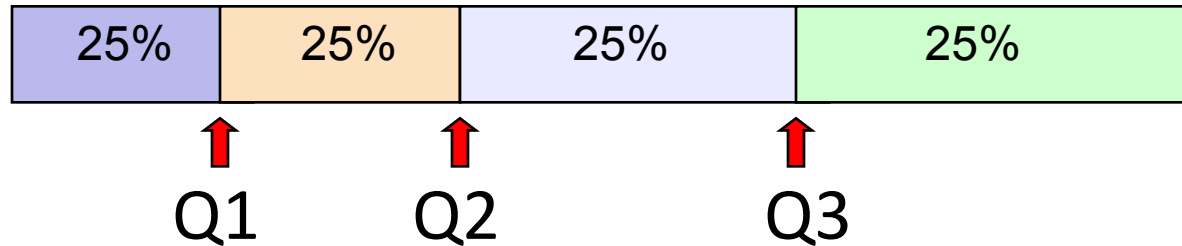
Range

- Difference between the largest and the smallest observations.

Range of age: 94 years-15 years = 79 years



Quartiles

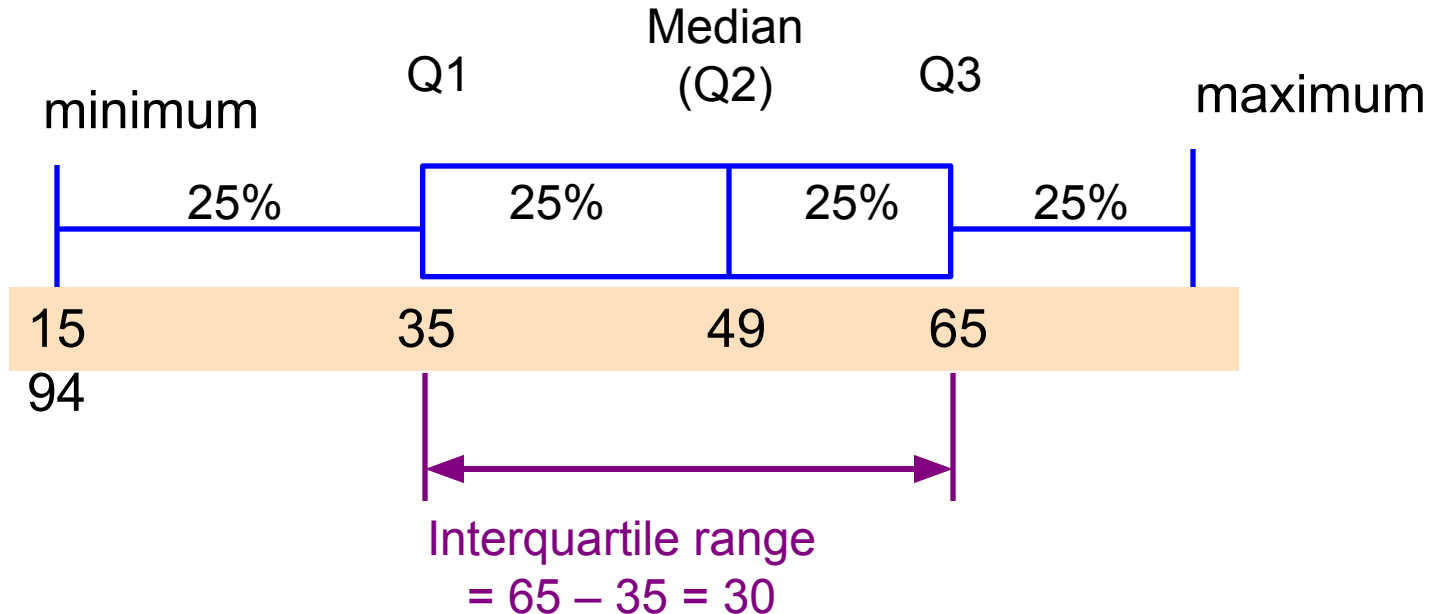


- The first quartile, Q_1 , is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Interquartile Range

- Interquartile range = 3rd quartile – 1st quartile
= $Q_3 - Q_1$

Interquartile Range: age



Variance

- Average (roughly) of squared deviations of values from the mean
- Describe the spread or variability of the data values

$$S^2 = \frac{\sum_i^n (x_i - \bar{X})^2}{n-1}$$

Why squared deviations?

- Adding deviations will yield a sum of 0.
- Absolute values are tricky!
- Squares eliminate the negatives.
- Result:
 - Increasing contribution to the variance as you go farther from the mean.

Practice

Find the sample variances of the data sets **A** and **B** given below.

A: 3, 4, 6, 7, 10 **B**: -20, 5, 15, 24

Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

$$S = \sqrt{\frac{\sum_i^n (x_i - \bar{X})^2}{n - 1}}$$

Calculation Example: Sample Standard Deviation

Age data (n=8) : 17 19 21 22 23 23 23 38

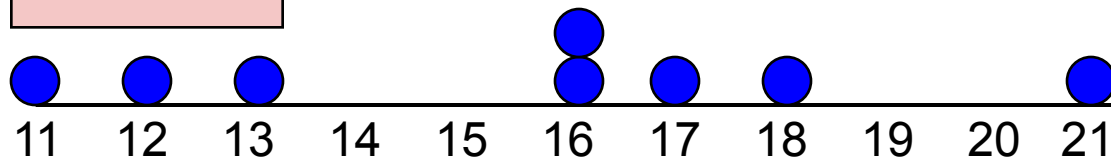
n = 8 Mean = $\bar{X} = 23.25$

$$S = \sqrt{\frac{(17 - 23.25)^2 + (19 - 23.25)^2 + \square + (38 - 23.25)^2}{8 - 1}}$$
$$= \sqrt{\frac{280}{7}} = 6.3$$

standard deviation gets bigger when numbers
are more spread out.

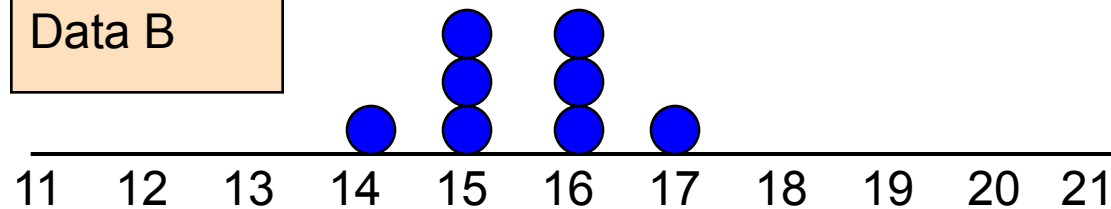
Comparing Standard Deviations

Data A

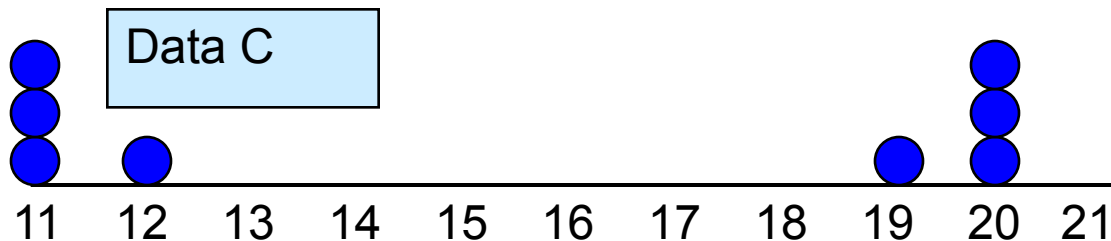


Mean = 15.5
 $S = 3.338$

Data B

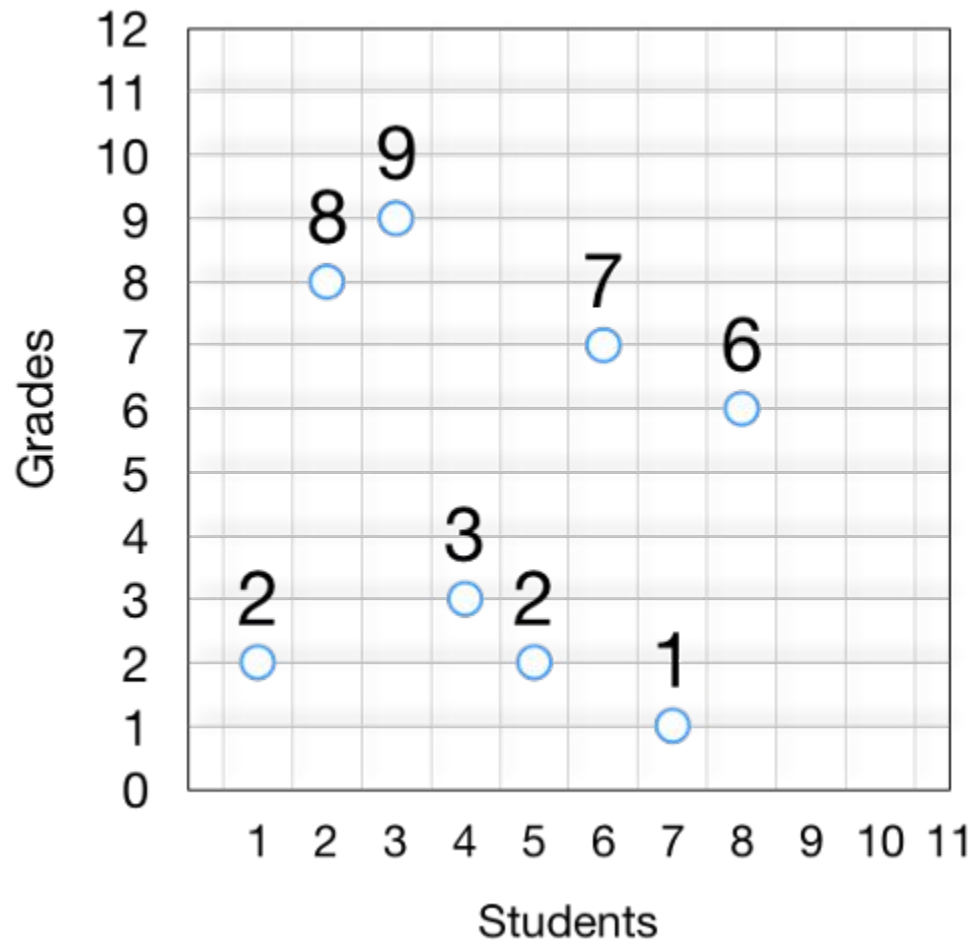


Mean = 15.5
 $S = 0.926$

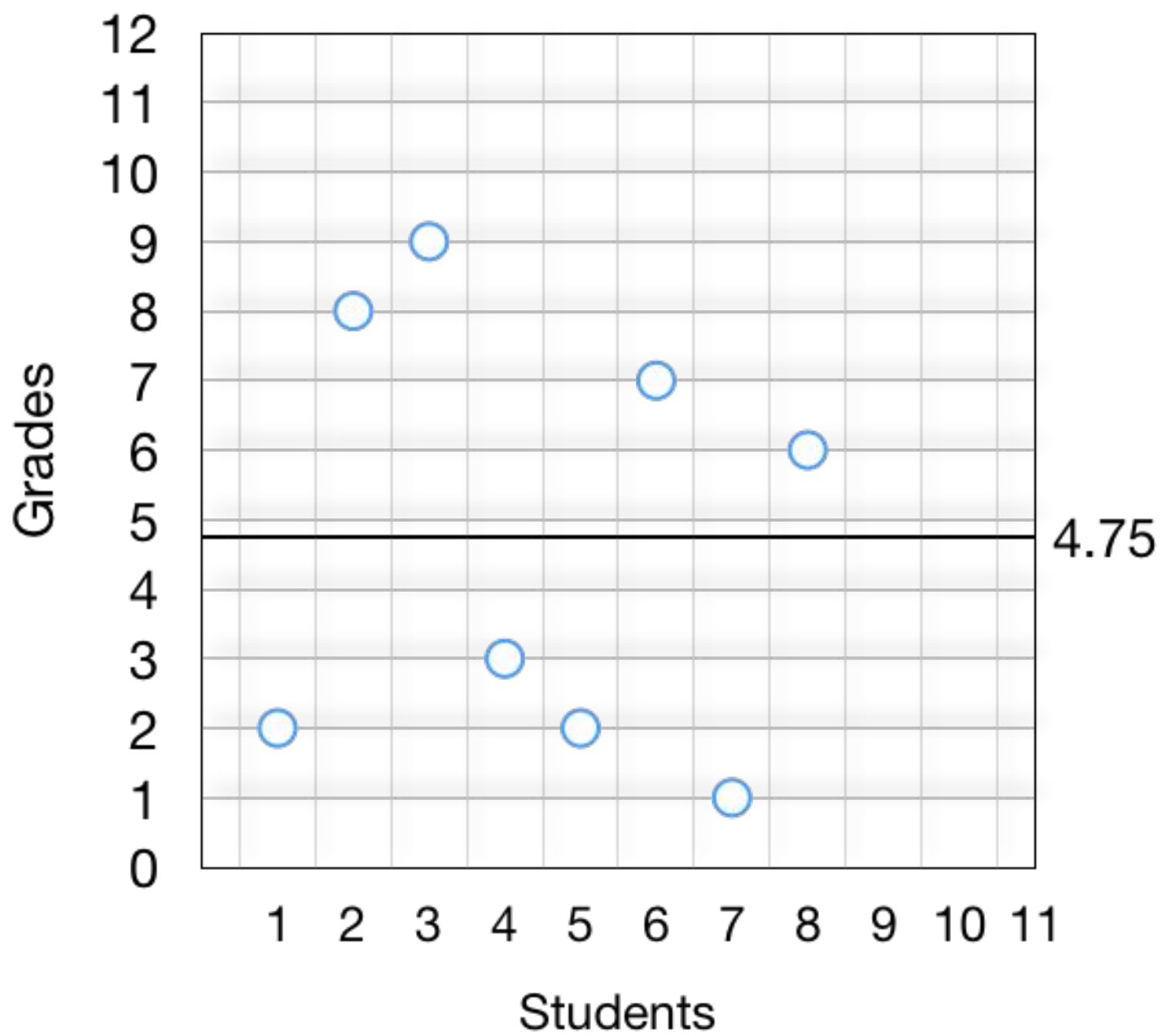


Mean = 15.5
 $S = 4.570$

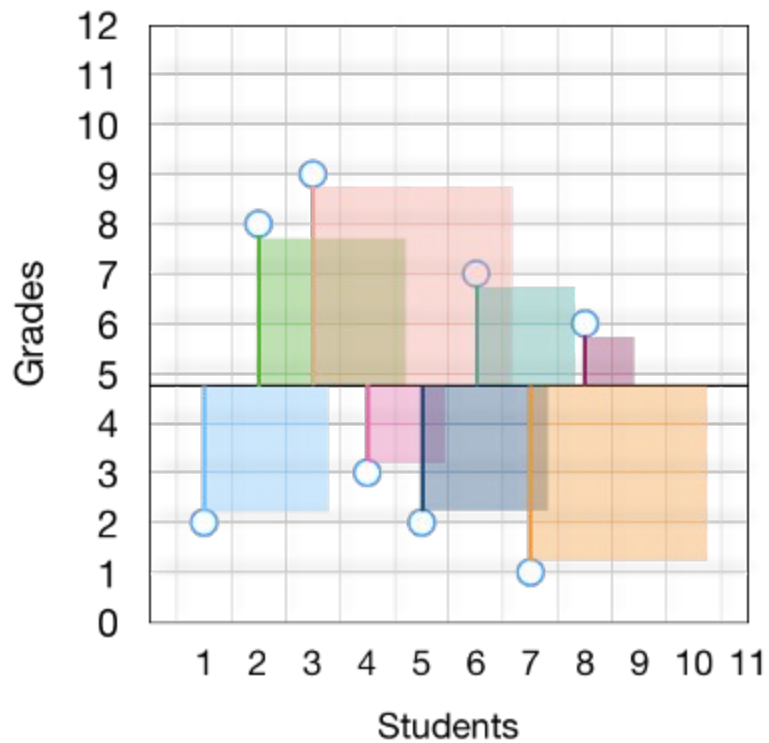
Visual Meaning



Student	Grade
1	2
2	8
3	9
4	3
5	2
6	7
7	1
8	6







$$\sum (x_n - \bar{x})^2 =$$

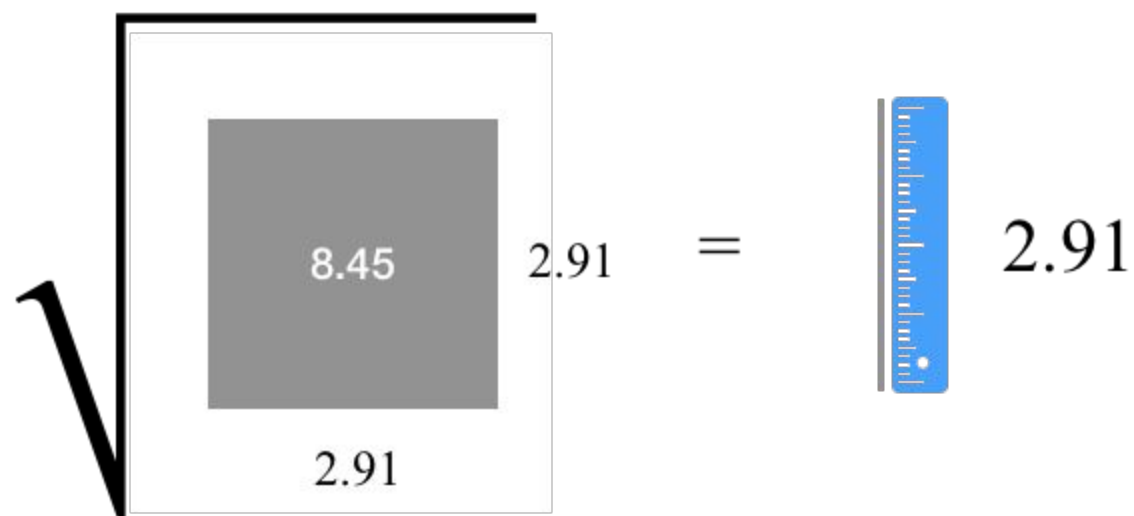
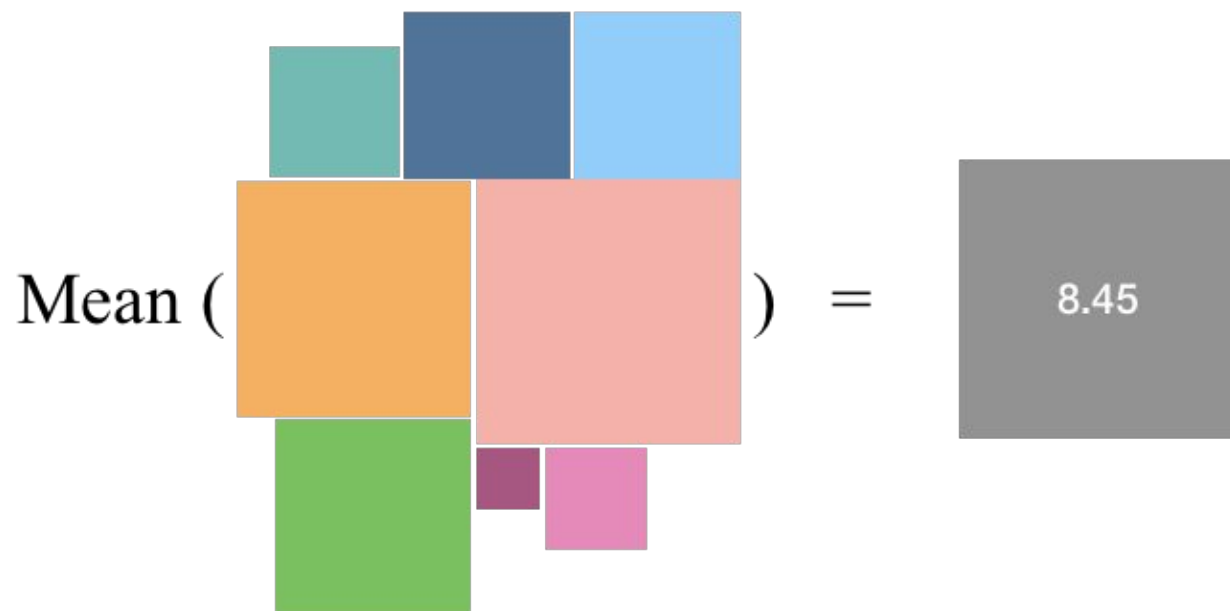
$$7.5625 + 10.5625$$

$$+ 18.0625 + 3.0625$$

$$+ 7.5625 + 5.0625$$

$$+ 14.0625 + 1.5625$$

$$= 67.5$$



Bienaymé-Chebyshev Rule

- Regardless of how the data are distributed, a certain percentage of values must fall within K standard deviations from the mean:

Note use of μ (mu) to represent “mean”.

Note use of σ (sigma) to represent “standard deviation.”

At least	within
$(1 - 1/1^2) = 0\%$	$k=1 \ (\mu \pm 1\sigma)$
$(1 - 1/2^2) = 75\%$	$k=2 \ (\mu \pm 2\sigma)$
$(1 - 1/3^2) = 89\%$	$k=3 \ (\mu \pm 3\sigma)$

Symbol Clarification

- S = Sample standard deviation (example of a “sample statistic”)
- σ = Standard deviation of the entire population (example of a “population parameter”) or from a theoretical probability distribution
- \bar{X} = Sample mean
- μ = Population or theoretical mean

****The beauty of the normal curve:**

No matter what μ and σ are, the area between $\mu - \sigma$ and $\mu + \sigma$ is about 68%; the area between $\mu - 2\sigma$ and $\mu + 2\sigma$ is about 95%; and the area between $\mu - 3\sigma$ and $\mu + 3\sigma$ is about 99.7%. Almost all values fall within 3 standard deviations.

68-95-99.7 Rule

