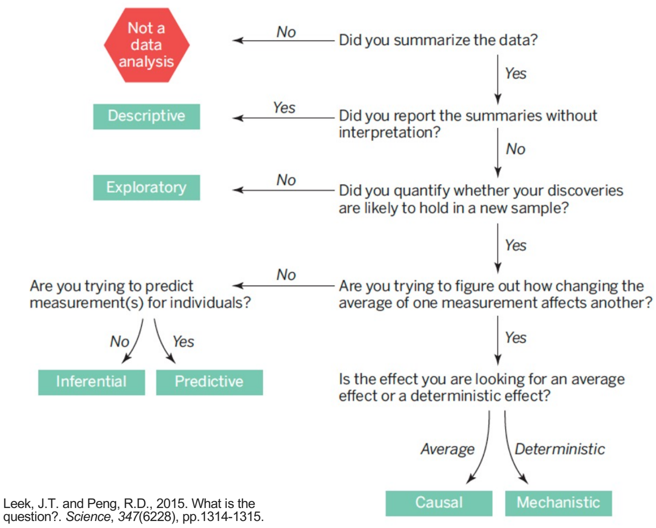
Data Science for Researchers and Scholars

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Data Science Starts with a Question



Leek, J.T. and Peng, R.D., 2015. What is the question?. *Science*, *347*(6228), pp.1314-1315.

Data science begins with a question

- Questions come in many forms

Question type	Description	Example
Descriptive	A question about summary characteristics of a data set without interpretation (i.e., report a fact).	How many students are enrolled at Penn State in Fall 2023?
Exploratory	A question about patterns, trends, or relationships within a single data set. Often used to propose hypotheses for future study.	Do political party preferences change with indicators of wealth in a collected sample of 2000 individuals US?

Data science begins with a question

- Questions come in many forms

Question type	Description	Example
Predictive	A question about prediction of an outcome of interest, but not what causes the outcome.	What political party will Joe Sixpack vote for in the next US Presidential election?
Inferential	A question about patterns, trends, or relationships in a single data set **and** quantification of how applicable these findings are to the wider population.	Do political party preferences change with indicators of wealth for all people living in the US?

Data science begins with a question

- Questions come in many forms

Question type	Description	Example
Causal	A question about whether changing one factor will lead to a change in another factor, on average, in the wider population.	Does college education causally impact voting for a certain political party in the US elections?
Mechanistic	A question about the underlying mechanism of the observed patterns, trends, or relationships (i.e., how does it happen?)	How do wealth lead to voting for a certain political party in the US elections?

- Mechanistic questions are beyond the scope of this course

Review: Statistics and Probability

Statistics

- Statistics is the science of collecting, organizing, analyzing and interpreting data.
- “We can no more escape data than we can avoid the use of words”.

A few headlines

- About 1 in 9 people age 65 and older (10.7%) has Alzheimer's.
- Lingering inflation worries keep Biden approval stagnant at 40%
- US families of two persons had an annual median income of $75,143 as of 2023.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.

- There is a 70 percent chance that a large earthquake will strike San Francisco by 2030.
- There is 56% chance of rain tomorrow (in State College).

Population and Sample

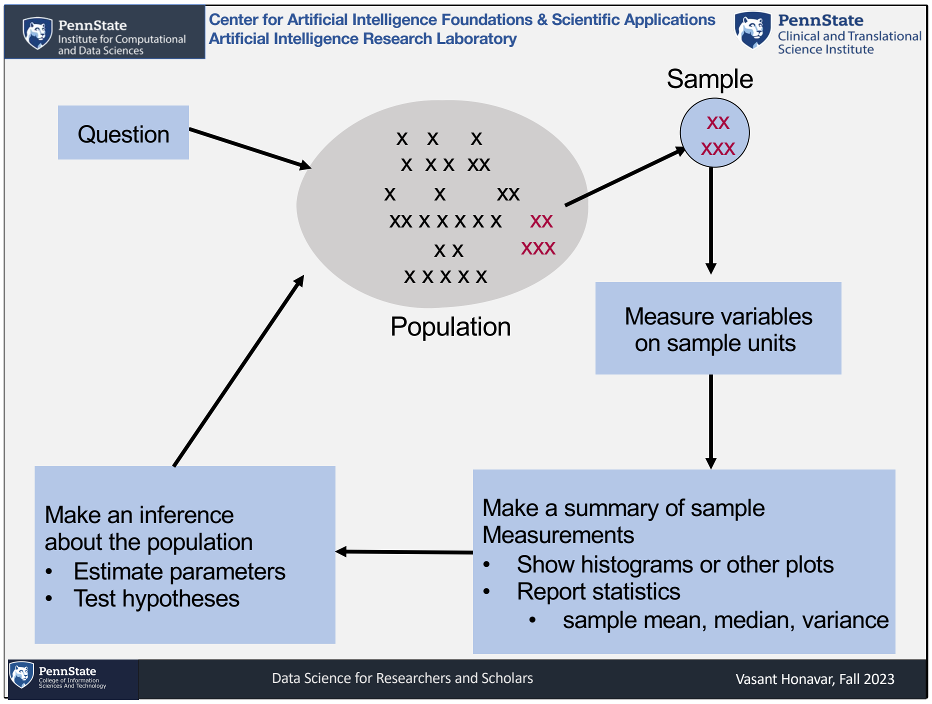
- **Population** The population is the set representing all entities of interest to the investigator.
- A population can be an entire collection of people, animals, plants or things from which we may collect data.
- It is the entire group we are interested in, which we wish to describe or draw conclusions about.
- To make any generalizations about a population, we often study a sample, that is representative of the population.
- A sample could be the whole population, e.g. US Census
- In many cases the population is conceptual, e.g. daily precipitation over the next year.

Examples – Population versus Sample

- President's approval rating:
  - A CNN poll of 1500 adult Americans conducted on January 7, 2002 showed that 618 said they "approve", 702 "disapprove" and 180 had no opinion.
  - Population: 150-plus million adult Americans.
  - Sample: 1500 interviewed.
- The ratio of the mass of the earth to that of the moon:
  - Measured during different space flights: 81.3001, 81.3015, 81.3006, 81.3011, 81.299, 81.3015, 81.3005, 81.3021.
  - These number differ form one another (and presumably from the true ratio) because of measurement error.
  - Population: all possible measurements that might be made under similar experimental conditions: a conceptual (hypothetical) population since it does not actually exist.
  - Sample: eight measurements.

The Six Step Inference Process

- Step 1: What is the question of interest?
- Step 2: What is population associated with the question?
- Step 3: Take a representative sample from the population
- Step 4: Measure one or more properties or variables on each unit in the sample.
- Step 5: Make a summary of the measured variables (data)
- Step 6: Make an inference about the population from the data.


Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory
PennState Institute for Computational and Data Sciences
PennState Clinical and Translational Science Institute
Question
Sample
XX
XXX
Population
Measure variables on sample units
Make a summary of sample Measurements
• Show histograms or other plots
• Report statistics
• sample mean, median, variance
Make an inference about the population
• Estimate parameters
• Test hypotheses
PennState College of Information Sciences And Technology
Data Science for Researchers and Scholars
Vasant Honavar, Fall 2023

Example: Pollution at an oil refinery

- Environmental Protection Agency (EPA) has accused Shell Oil Company of violation environmental regulations at its refinery located in Huston during the year 2001.
- The regulations state that the average petroleum leaked into the ground at the refinery must not exceed 100 gallons per day during any calendar year.
- Fine for violating the regulations is $1,000,000.
- EPA regulators visited the refinery on eight days in December, 2001 and measured the petroleum leaked as 110, 96, 104, 101, 87, 99, 116, 108 gallons.
- Sample average: 102.625
- What is the question? population? sample? variable? summary? inference?

Example: Pollution at an oil refinery

- Question: Does average leakage exceed 100 gallons per day?
- Population: Every single day of that year 2001.
- Sample: Measurements on 8 days in December.
- Variable: Leakage in gallons; data as given.
- Summary: Average leakage = 102.625 gallon/day.
- Inference: The average leakage for the year 2001 exceeds 100 gallons. The company should be fined $1 million.

Example: Pollution at an oil refinery

Follow-up questions

- Is the sample representative of the population?
- What argument might Shell use to contest EPA's conclusions?
- How could EPA defend its conclusions?
- Answers to these questions are statistical in nature
  - What is the error of the estimate?
  - What is the probability that the population mean is less than 100 given that the sample mean is 102.625?

Descriptive versus Inferential Statistics

- Descriptive Statistics
  - Summarizing data, visualizing data
- Inferential Statistics
  - Making decisions or predictions about a population based on sampled data.

Terminology

- A variable is a characteristic that varies for different individuals or units in the population
- An experimental unit may be an individual or object on which a variable is measured, yielding a measurement
- Data is a set of measurements from a sample
- Examples:
  - Hair color
  - White blood cell count
  - PM2.5 in the air
  - Gene expression
  - GPA
  - Annual Income
  - Word count

Experimental unit: Person
Variable: Hair Color
Measurements: Black, Brown, Red, Blonde

Variables

- Univariate (one) or multi-variate (many)
- Qualitative (categorical) – denote a quality or characteristic
  - Gender, Hair Color, Amino Acid type, Species, Movie type
- Ordinal – (where only relative ordering matters)
  - Letter grade (A, B, C, D, E, F)
- Quantitative (numeric)
  - Discrete – assume a countable number of values
    - Number of students in a class, number of amino acids in a protein, number of votes received by a candidate
  - Continuous – assume infinitely many values corresponding to points on a line interval
    - Temperature, Binding strength, Melting point

Examples

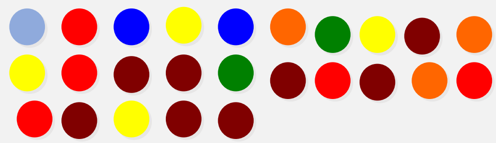
- The faculty size for each department at Penn State
  - Quantitative discrete
- Time until each light bulb burns out
  - Quantitative continuous
- Letter grade received by each student
  - Ordinal
- Blood type of individuals
  - Qualitative

Graphing Qualitative Variables

- Use a data distribution to describe:
  - What values of the variable have been observed
  - How often each value has shows up in the sample
- "How often" can be measured 3 ways:
  - Frequency
  - Relative frequency = Frequency/number of samples
  - Percent = 100 x Relative frequency

Example

- A sample of 25 M&Ms:
- Raw Data

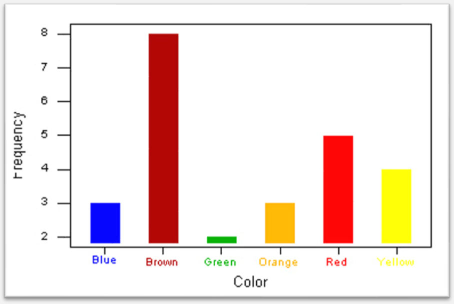


Statistical Table

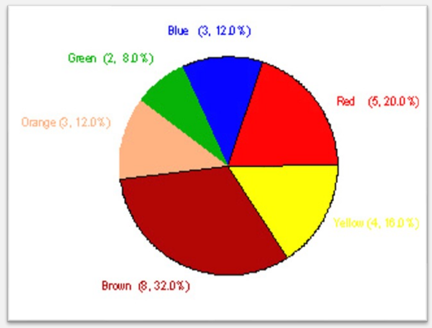
Color	Tally	Frequency	Relative Frequency	Percent
Red		5	5/25 = .20	20%
Blue		3	3/25 = .12	12%
Green		2	2/25 = .08	8%
Orange		3	3/25 = .12	12%
Brown		8	8/25 = .32	32%
Yellow		4	4/25 = .16	16%

Graphs

Bar Chart

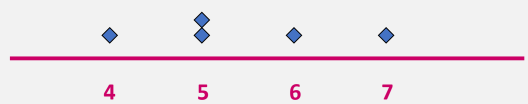


Pie Chart

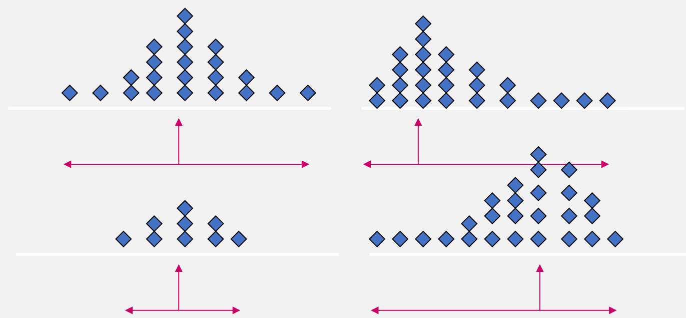


Scatterplots

- The simplest graph for quantitative data
- Plots the measurements as points on a horizontal axis, stacking the points that duplicate existing points.
  - Example data: 4, 5, 5, 7, 6

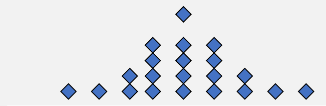


Interpreting Graphs: Location and Spread

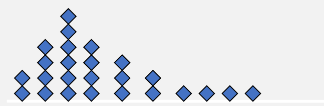


Where is the data centered on the horizontal axis, and how does it spread out from the center?

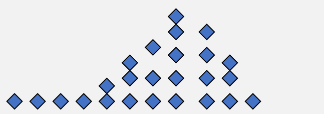
Interpreting Graphs: Shapes



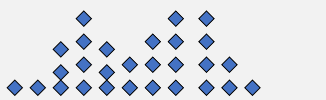
Mound shaped and symmetric (mirror images)



Skewed right: a few unusually large measurements



Skewed left: a few unusually small measurements

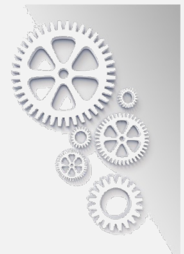


Bimodal: two local peaks

Interpreting Graphs: Outliers

- Are there any strange or unusual measurements that stand out in the data set?

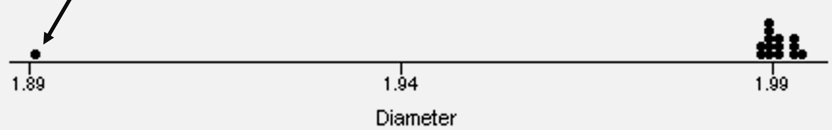
Example



- A quality control process measures the diameter of a gear being made by a machine (cm).
- The technician records 15 diameters, but inadvertently makes a typing mistake on the second entry.

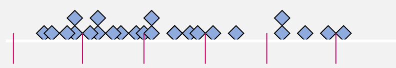
1.991 1.891 1.991 1.988 1.993 1.989 1.990 1.988

1.988 1.993 1.991 1.989 1.989 1.993 1.990 1.994

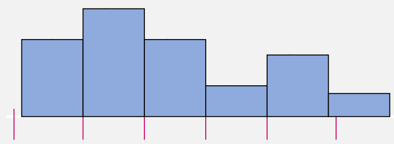


Relative Frequency Histograms

- A relative frequency histogram for a quantitative data set is a bar graph in which the height of the bar shows "how often" (as quantified by relative frequency) measurements fall in a particular group or subinterval.



Create intervals



Calculate relative frequency in each sub-interval

How to Draw Relative Frequency Histograms

- Divide the range of the data into 5-10 subintervals of equal length.
- Calculate the approximate width of the subinterval as Range/number of subintervals.
- Round the approximate width up to a convenient value.
- Use left inclusion – include the left endpoint of the interval, but not the right, in your tally.
- Create a statistical table including the subintervals, their frequencies and relative frequencies.

How to Draw Relative Frequency Histograms

- Draw the relative frequency histogram, plotting the subintervals on the horizontal axis and the relative frequencies on the vertical axis.
- The height of the bar represents
  - The proportion of measurements falling in that class or subinterval.
  - The probability that a single measurement, drawn at random from the set, will belong to that class or subinterval.

Example

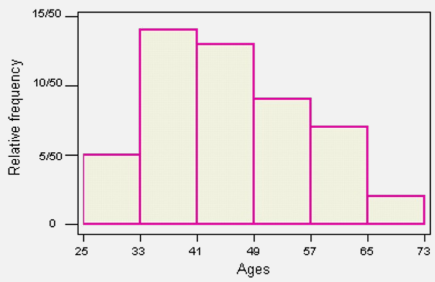
The ages of 50 tenured faculty at PSU

34 48 **70** 63 52 52 35 50 37 43 53 43
52 44 42 31 36 48 43 **26** 58 62 49 34
48 53 39 45 34 59 34 66 40 59 36 41
35 36 62 34 38 28 43 50 30 43 32 44
58 53

- We choose to use 6 intervals.
- Minimum interval width = (70 – 26)/6 = 7.33
- Convenient interval width = 8
- Use 6 intervals of length 8 each, starting at 25.

Example Continued

Age	Frequency	Relative Frequency	Percent
25 to < 33	5	5/50 = .10	10%
33 to < 41	14	14/50 = .28	28%
41 to < 49	13	13/50 = .26	26%
49 to < 57	9	9/50 = .18	18%
57 to < 65	7	7/50 = .14	14%
65 to < 73	2	2/50 = .04	4%



Shape – skewed right

Outliers – No

What proportion of the tenured faculty are younger than 41? (14 + 5)/50 = 19/50 = .38

What is the probability that a randomly selected faculty member is 49 or older?

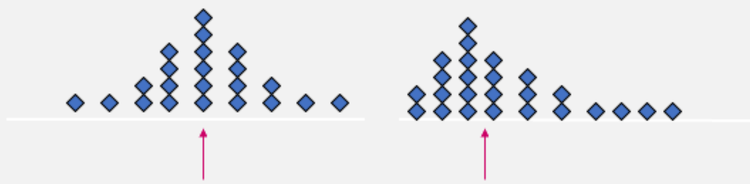
(9+ 7 + 2)/50 = 18/50 = .36

Describing Data with Numerical Measures

- Plots may not always be sufficient for describing data.
- Numerical measures can be obtained for both populations and samples.
  - A parameter is a numerical descriptive measure of a population.
  - A statistic is a numerical descriptive measure of a sample.

Descriptive Statistics

- Descriptive statistics provides ways to capture the properties of a given data or sample.
  - Central tendency measures describe the center around the data is distributed.
  - Variation or variability measures describe data spread, i.e. how far the measurements lie from the center.



Some Notations

- A little notation goes a long way
- Suppose we have measurements on n samples
- We denote them by $x_1, x_2, \cdots, x_n$

Example

- Suppose we ask five people how many hours of they spend on the internet in a week and get the following answers: 2, 9, 11, 5, 6.
- Then $n = 5$
- $x_1 = 2,\ x_2 = 9,\ x_3 = 11,\ x_4 = 5,\ x_5 = 6$

Arithmetic Mean or Average

The mean of a set of measurements is the sum of the measurements divided by the total number of measurements.

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

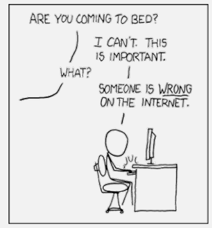
where

- n = number of measurements and
- $\sum_{i=1}^{n} x_i$ denotes the sum of all measurements

Example

Time spent on internet:

2, 9, 11, 5, 6 hours



$$\bar{x} = \frac{\sum x_i}{n} = \frac{2+9+11+5+6}{5} = \frac{33}{5} = 6.6$$

If we were able to enumerate the whole population, the population mean would be called μ (the Greek letter "mu").

Median

- When the measurements are ranked from smallest to largest the median of a set of measurements is
  - the middle measurement when the number of measurements is odd
  - the mean of the two middle measurements when the number of measurements is even
- Example
  - What is the median of 4, 2, 8, 9, 6?
    - Sort the list: 2, 4, 6, 8, 9
    - Median is 6
  - What is the median of 4, 2, 8, 7, 9, 6?
    - Sort the list: 2, 4, 6, 7, 8, 9
    - Median is 6.5

Which Central Tendency Measure to Use?

- Mean is meaningful for symmetric distributions without outliers: e.g. height and weight.
- Median is better for skewed distributions or data with outliers: e.g. wealth and income.
- Bill Gates / Warren Buffet / Larry Ellison[$] would each add somewhere between* $300 and $400
  - to the mean per capita wealth in the United States
  - but nothing to the median.

[$]Net worth about $100 million each
*Uncertainty due to numbers being inexact and fluctuations of the stock market

Other Central Tendency Measures

- The geometric mean is the nth root of the product of n values
- The geometric mean is always $\leq$ arithmetic mean, and more sensitive to values near zero.
- Geometric means make sense with ratios

Mode

- The **mode** is the measurement which occurs most frequently.
- Data: 2, 4, 9, 8, 8, 5, 3
  - The mode is 8, which occurs twice
- Data: 2, 2, 9, 8, 8, 5, 3
  - There are two modes—8 and 2 (bimodal)
- Data: 2, 4, 9, 8, 5, 3
  - There is no mode (each value is unique).

Example

Data: The number of quarts of milk purchased by 25 households:

0 0 1 1 1 1 1 2 2 2 2 2 2
2 2 2 3 3 3 3 3 4 4 4 5

- Mean?

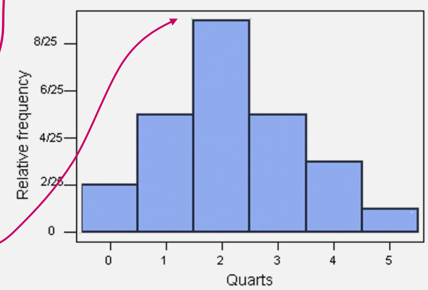
$$\bar{x} = \frac{\sum x_i}{n} = \frac{55}{25} = 2.2$$

- Median?

$$m = 2$$

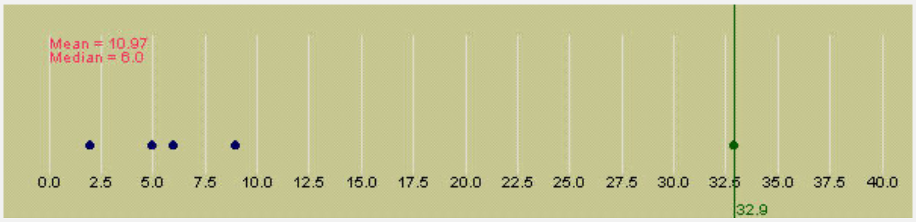
- Mode? (Highest peak)

$$\text{mode} = 2$$



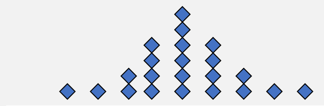
Extreme Values

- The mean is more easily affected by extremely large or small values than the median.

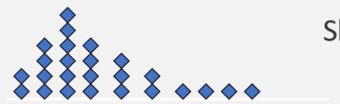


- The median is often used as a centrality when the distribution is skewed

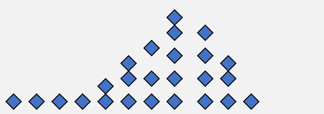
Extreme Values



Symmetric: Mean = Median



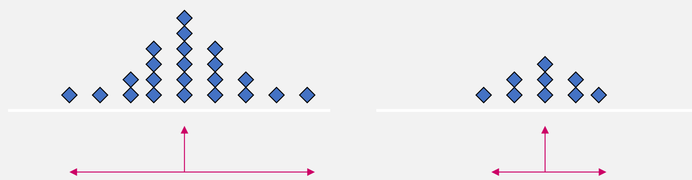
Skewed right: Mean > Median



Skewed left: Mean < Median

Measures of Variability

- A measure along the horizontal axis of the data distribution that describes the spread of the distribution from the center.



The Range

- The range, R, of a set of n measurements is the difference between the largest and smallest measurements.
- Example:
  - A botanist records the number of petals on 5 flowers:

5, 12, 6, 8, 14

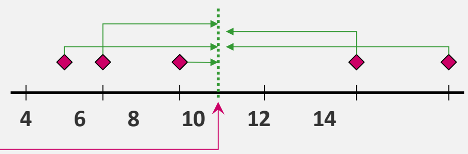
- The range is $R = 14 - 5 = 9$.

Quick and easy to compute but only uses 2 of the 5 measurements.

The Variance

- The variance measures the average deviation of the measurements around their mean.
- Flower petals: 5, 12, 6, 8, 14

$$\bar{x} = \frac{45}{5} = 9$$



The Variance

- The variance of a population of N measurements is the average of the squared deviations of the measurements about their mean μ.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

- The variance of a sample of n measurements is the sum of the squared deviations of the measurements about their mean, divided by $(n-1)$.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The Standard Deviation

- In calculating the variance, we squared all of the deviations, and in doing so changed the scale of the measurements.
- To return this measure of variability to the original units of measure, we calculate the standard deviation, the positive square root of the variance.

Population standard deviation : $\sigma = \sqrt{\sigma^2}$

Sample standard deviation : $s = \sqrt{s^2}$

Calculating Sample Variance

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	5	-4	16
	12	3	9
	6	-3	9
	8	-1	1
	14	5	25
Sum			60

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$= \frac{60}{4} = 15$$

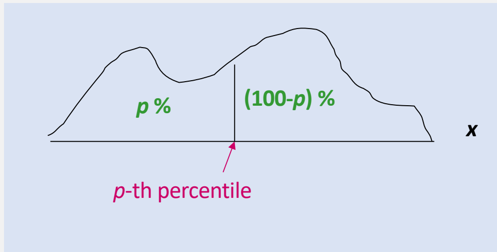
$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

Some Notes

- The value of s is always positive.
- The larger the value of s^2 or s, the larger the variability of the data.
- Why divide by $n - 1$?
  - The sample standard deviation s is often used to estimate the population standard deviation s.
  - Dividing by $n - 1$ gives us a better estimate of s.
- Distributions with the same mean can look very different. But together, the mean and standard deviation fairly well characterize any distribution.

Measures of Relative Standing

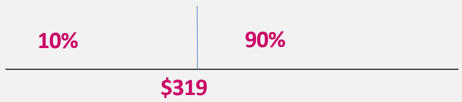
How many measurements lie below the measurement of interest? This is measured by the p^{th} percentile.



Examples

90% of all men (16 and older) earn more than \$319 per week.

BUREAU OF LABOR STATISTICS 2002



\$319 is the 10^{th} percentile

50^{th} Percentile	≡ Median
25^{th} Percentile	≡ Lower Quartile (Q_1)
75^{th} Percentile	≡ Upper Quartile (Q_3)

Quartiles and the IQR

- The lower quartile (Q_1) is the value of x which is larger than 25% and less than 75% of the ordered measurements.
- The upper quartile (Q_3) is the value of x which is larger than 75% and less than 25% of the ordered measurements.
- The range of the “middle 50%” of the measurements is the interquartile range, $IQR = Q_3 - Q_1$

Calculating Sample Quartiles

- The lower and upper quartiles (Q_1 and Q_3), can be calculated as follows:
- The position of Q_1 is $0.25(n + 1)$
  - The position of Q_3 is $0.75(n+1)$

once the measurements have been ordered.

- If the positions are not integers, find the quartiles by interpolation.

Example

The prices ($) of 18 brands of walking shoes:

40 60 65 65 65 68 68 70 70

70 70 70 70 74 75 75 90 95

Position of Q_1 = .25(18 + 1) = 4.75

Position of Q_3 = .75(18 + 1) = 14.25

- Q_1 is 3/4 of the way between the 4th and 5th ordered measurements, or
  Q_1 = 65 + .75(65 - 65) = 65.

Example

The prices ($) of 18 brands of walking shoes:

40 60 65 65 65 68 68 70 70

70 70 70 70 74 75 75 90 95

Position of Q_1 = .25(18 + 1) = 4.75

Position of Q_3 = .75(18 + 1) = 14.25

- Q_3 is 1/4 of the way between the 14th and 15th ordered measurements, or
  Q_3 = 75 + .25(75 - 74) = 75.25

  IQR = $Q_3 - Q_1$ = 75.25 - 65 = 10.25

Using Measures of Center and Spread: The Box Plot

The Five-Number Summary: Min, Q_1, Median, Q_3, Max

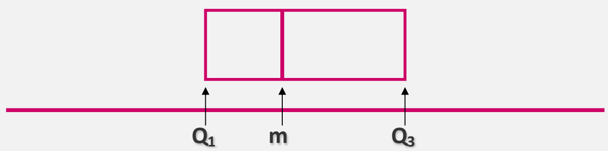
- Divides the data into 4 subsets containing an equal number of measurements.
- To get a quick summary of the data distribution.
- Construct a box plot to describe the shape of the distribution and to detect outliers.

Constructing a Box Plot

Calculate Q_1, the median, Q_3 and IQR.

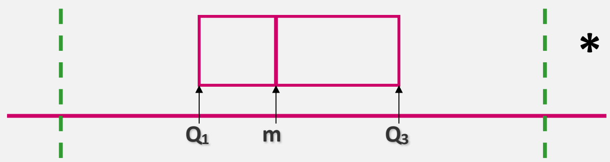
✓Draw a horizontal line to represent the scale of measurement.

✓Draw a box using Q_1, the median, Q_3.



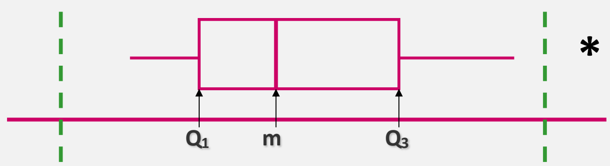
Constructing a Box Plot

- ✓Isolate outliers by calculating
  - ✓Lower fence: Q_1-1.5 IQR
  - ✓Upper fence: Q_3+1.5 IQR
- ✓Measurements beyond the upper or lower fence is are outliers and are marked (*).



Constructing a Box Plot

- Draw “whiskers” connecting the largest and smallest measurements that are NOT outliers to the box.

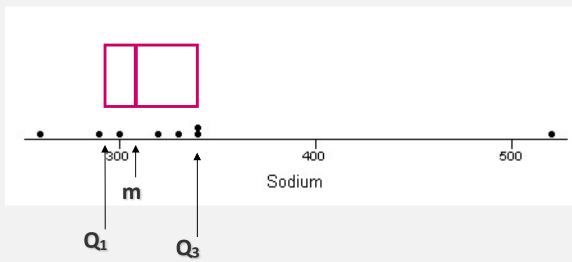


Example

Amount of sodium in 8 brands of cheese:

260 290 300 320 330 340 340 520

Q_1 = 292.5 m = 325 Q_3 = 340



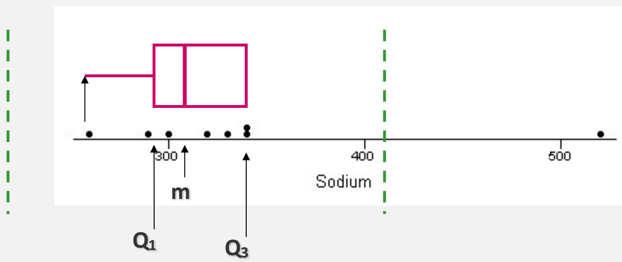
Example

$$IQR = 340 - 292.5 = 47.5$$

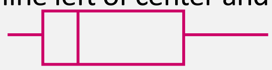
Lower fence $= 292.5 - 1.5(47.5) = 221.25$

Upper fence $= 340 + 1.5(47.5) = 411.25$

Outlier: x = 520



Interpreting Box Plots

- Median line in center of box and whiskers of equal length—symmetric distribution 
- Median line left of center and long right whisker—skewed right 
- Median line right of center and long left whisker—skewed left