IS 709 Introduction to Data Science

Lecture 3 – Understanding Data- Part1

Attributes

Understanding Data

- Before conducting any analysis, you should first understand your own data.
- Better if you manually inspect it but
  - If data is high-dimensional and very large in size, it is not possible.
  - Suggested: Visualization techniques, descriptive statistics of data

The analyst knew he was in for a rough day when his boss came in dressed like Indiana Jones.

Understanding Data

- Before carrying out more complex analysis such as machine learning techniques, it is common to use the tools of statistics:
- Data collection: Collecting relevant data for our research questions.
- Descriptive statistics: We will generate statistics that summarize the data concisely, and evaluate different ways to visualize data.
- Exploratory data analysis: We will look for patterns, differences, and other features that address the questions we are interested in. At the same time we will check for inconsistencies and identify limitations.
- Hypothesis testing: Where we see apparent effects, like a difference between two groups, we will evaluate whether the effect is real, or whether it might have happened by chance.
- Estimation: We will use data from a sample to estimate characteristics of the general population.

Understanding Data

Definitions

- Probability is the study of random or non-deterministic events.
- Statistics is the discipline of using data samples to support claims about populations.
  - Most statistical analysis is based on probability, which is why these pieces are usually presented together.
- Population: A group we are interested in studying, often a group of people, but the term is also used for animals, vegetables and minerals.
- Sample: The subset of a population used to collect data.

Data

- Data is plural of datum which is an abstraction.
- Datum is a single quantity or quality of a real-world entity such as person, object and event.
  - Corresponds to an attribute (also known as feature, variable, dimension) of an entity.
- Each entity is typically described by a number of attributes.
  - For example a book might have attributes such as author, title, topic, genre, publisher, price, date published, word count, number of chapters, number of pages, edition, and ISBN.

Captured and Exhaust Data

- **Captured data/Designed data** are collected through a direct measurement or observation that is designed to gather the data.
  - Experiments or surveys can be used to gather data on a particular topic of interest
- One particular example of designed data is data captured by devices that are classified under the term Internet of Things (IoT).
  - These devices are created by people and the data they collect were pre-specified by producers (e.g. logs).

Captured and Exhaust Data

- **Exhaust data/Organic data/Digital footprint** are a by-product of a process whose primary purpose is something other than data capture.
  - For example, collecting data such as items browsed, items put into cart and time spent while a customer is buying products on a web site.
- It refers to passively generated data
- Metadata (i.e., data that describe other data) is also one of the most common types of exhaust data
- **Big data** are highly detailed exhaust data automatically captured by sensors or generated through IT systems.

Beręsewicz, M., Lehtonen, R., Reis, F., Di Consiglio, L., & Karlberg, M. (2018). *An overview of methods for treating selectivity in big data sources*. Eurostat Statistical Working Paper. Doi: https://doi. org/10.2785/312232.

Abstraction

- Data are generated through a process of abstraction, so any data are the result of human decisions and choices.
  - There are choices for things abstracted and what categories or measurements to use in the abstracted representation.
- Data are never an objective description of the world, they are always partial and biased.
- Data are not a perfect representation of the real-world entities and processes we are trying to understand.
- Nevertheless, if we pay attention to how we design and gather the data, the results of our analyses will provide useful insights into real-world problems.

Data Set

- A data set consists of the data related to a collection of entities which are described in terms of a set of attributes.
- In practice, the majority of the time and effort in data science projects is spent on creating, cleaning and transforming data sets.

Data Set

Structured vs. Unstructured Data

- **Structured data** refers to data with a high level of organization, such as in relational databases and spreadsheets.
  - Depends on a data model - a model of the data types and how they will be stored, processed and accessed.
  - Easily entered, stored, queried and analyzed.
  - SQL is, for instance, is used for management of structured data.
- Every entity has the same set of attributes
  - Ex: *demographic data for a population, classical readings*
  - Can be represented by an n x m matrix:
    - n rows: a row for each entity,
    - m columns: a column for each attribute.

Title	Author	Year
Shibumi	Trevenian	1979
Perfume	P. Suskind	1985
Foundation	I. Asimov	1951

Data Set

Structured vs. Unstructured Data

- **Unstructured data** means all things that cannot be classified and fit into one simple model.
  - Photos, images, videos, web pages, e-mails, web page, blog entries, PDF files,…
- Each instance may have its own internal structure

Data Set

Structured vs. Unstructured Data

- It is relatively easy to apply data science to structured data.
- Structured can be easily stored, organized, searched, reordered, and merged with other structured data.
- Unstructured data are much more common than structured data.
- It is difficult to analyze unstructured data in its raw form.
- It may be possible to extract structured data from unstructured data using techniques such as natural-language processing, digital signal processing, and computer vision.

Data Set

Structured vs. Unstructured Data

- Structured data can be used for number-driven (*quantitative*) purposes.
  - Most frequently retweeted social media account
  - Most popular tweet, facebook page..
- Unstructured data can be used for *qualitative* purposes.
  - What is the sentiment of news today?
  - Why is a certain hashtag commonly used today?
  - How people react to messages comprising hatred?
  - What misinformation is shared among users regarding Covid?

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

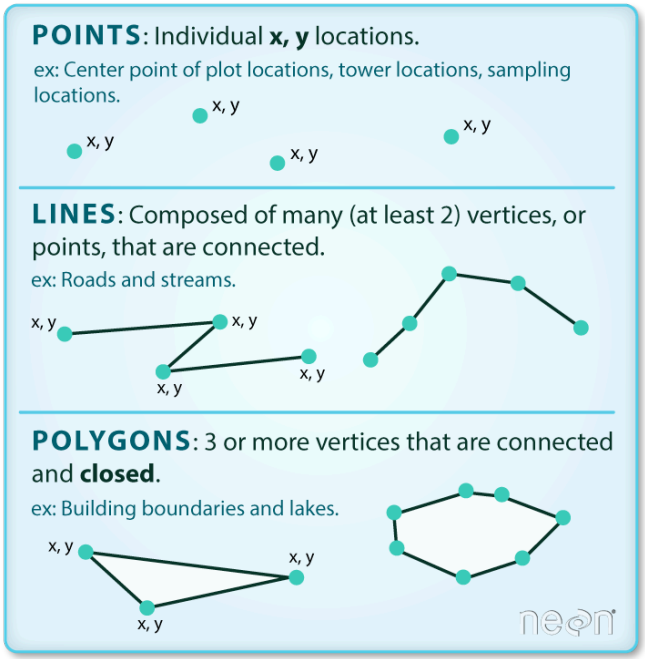
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Types of Data Sets

- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data
  - Video data
- Panel (longitudinal) data

person	year	income	age	sex
1	2001	1300	27	1
1	2002	1600	28	1
1	2003	2000	29	1
2	2001	2000	38	2
2	2002	2300	39	2
2	2003	2400	40	2

- Cross sectional data:
  - refers to data collected by observing many subjects (such as individuals, firms or countries/regions) at the same point of time, or without regard to differences in time.



Ref: https://nceas.github.io/oss-lessons/spatial-data-gis-law/3-mon-intro-gis-in-r.html

Difference between time series, cross-sectional and panel data

- 1. **Time series data** - It is a collection of observations (behaviour) for a single subject(entity) at different time intervals (generally equally spaced)
  - Example - Max Temperature, Humidity and Wind(all three behaviours) in New York City(single entity) collected on First day of every year(multiple intervals of time)

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2012	35	56%	3 mph
NYC	1/1/2013	47	65%	21 mph
NYC	1/1/2014	30	39%	16 mph
NYC	1/1/2015	55	45%	4 mph

- 2. **Cross-Sectional data** - It is a collection of observations (behaviour) for multiple subjects(entities) at single point in time.
  - Example - Max Temperature, Humidity and Wind(all three behaviours) in New York City, SFO, Boston, Chicago(multiple entities) on 1/1/2015(single instance)

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2015	55	45%	4 mph
SFO	1/1/2015	70	35%	21 mph
Boston	1/1/2015	34	39%	16 mph
Chicago	1/1/2015	29	15%	54 mph

- 3. **Panel Data (Longitudinal Data)** - It is usually called as Cross-sectional Time-series data as it a combination of above mentioned types, i.e., collection of observations for multiple subjects at different time points.
  - Example - Max Temperature, Humidity and Wind(all three behaviours) in New York City, SFO, Boston, Chicago(multiple entities) on First day of every year(multiple intervals of time)

Difference between time series, cross-sectional and panel data

- 1. **Time series data** - It is a collection of observations (behaviour) for a single subject(entity) at different time intervals(generally equally spaced)
  - Example - Max Temperature, Humidity and Wind(all three behaviours) in New York City(single entity) collected on First day of every year(multiple intervals of time)

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2012	35	56%	3 mph
NYC	1/1/2013	47	65%	21 mph
NYC	1/1/2014	30	39%	16 mph
NYC	1/1/2015	55	45%	4 mph

- 2. **Cross-Sectional data** - It is a collection of observations (behaviour) for multiple subjects(entities) at single point in time.
  - Example - Max Temperature, Humidity and Wind(all three behaviours) in New York City, SFO, Boston, Chicago(multiple entities) on 1/1/2015(single instance)

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2015	55	45%	4 mph
SFO	1/1/2015	70	35%	21 mph
Boston	1/1/2015	34	39%	16 mph
Chicago	1/1/2015	29	15%	54 mph

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2015	55	45%	4 mph
NYC	1/1/2014	30	39%	16 mph
NYC	1/1/2013	47	65%	21 mph
SFO	1/1/2015	70	35%	21 mph
SFO	1/1/2014	75	23%	2 mph
SFO	1/1/2013	71	39%	13 mph
Boston	1/1/2015	34	39%	16 mph
Boston	1/1/2014	26	17%	27 mph
Boston	1/1/2013	45	46%	18 mph

- s Cross-sectional Time-series data as it a combination of for multiple subjects at different time points.
  all three behaviours) in New York City, SFO, Boston, ar(multiple intervals of time)

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database tables: rows -> data objects; columns ->attributes.

Attributes

- An attribute is a raw abstraction for an event or object.
  - Ex: *a person's height, the number of words in a tweet, the temperature in a room, the time or location of a purchase.*
- Data can also be derived from some other data by applying a function to raw data.
  - Ex: *the average person height, the variance in the temperature of a room across a period of time.*
- It is frequently very beneficial to identify derived attributes that provide insight into a problem.
  - Ex: *The Body Mass Index (=mass/height^2) provides more information about obesity than mass or height*.

Attributes

- Choosing the correct set of attributes is a challenge faced by many data science projects.
  - Too many attributes require too much time and effort to collect, integrate and check quality.
  - Irrelevant and redundant attributes may degrade the algorithm performance and may result in finding irrelevant and spurious patterns.

Data Types of Attributes

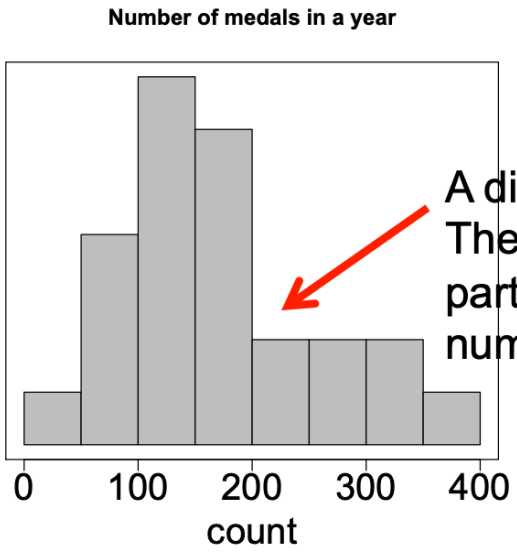
- The data type of an attribute (numeric, ordinal, nominal) affects the methods we can use to analyze and understand the data.
- Quantitative data deals with numbers and things you can measure objectively:
  - *E.g.,* dimensions such as height, width, and length. Temperature and humidity. Prices. Area and volume.
- Qualitative data deals with characteristics and descriptors that can't be easily measured, but can be observed subjectively
  - *E.g.,* as smells, tastes, textures, attractiveness, and color.
  - Descriptors may be represented by strings or numbers

Attributes

- **Quantitative**
  - Number of medals won by U.S. in a given year.
  - Can be shown with a distribution, or summarized with an average, etc.
- **Qualitative**
  - Medal Type: Gold/Silver/Bronze
  - Summarized with a table or chart.

With some reformatting of the earlier data, we can get a count of medals for each year.

Year	Count
1896	20
1900	55
1904	394
1908	63
1912	101
1920	193



A distribution. The x-axis shows part of the real-number line.

Attributes

Quantitative Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g.,
      - Zip codes, profession, or the set of words in a collection of documents
      - Number of children in a household
      - Number of languages a person speaks
      - Number of people sleeping in stats class
  - Sometimes, represented as integer variables

Attributes

Quantitative Attributes

- **Continuous Attribute**
  - Has infinite number of states
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Attribute	Value
Height	5.4, 6.2 …etc
weight	50.33 ……etc

Attributes

Quantitative Attributes

- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - *E.g.,* Time interval on a 12 hour clock (6am, 6pm)
  - No true zero-point
  - Interval data cannot be multiplied or divided, however, it can be added or subtracted.
  - Interval data is measured on an interval scale.

Does it make sense if we say that 20°C is half as hot as 40°C or that 40°C is twice as hot as 20°C?

Other examples of interval data include: IQ scores, dates on a calendar, and longitudes on a map. The key distinction is that the zero point on an interval scale is arbitrarily chosen; it doesn't represent a natural minimum quantity of the thing being measured.

Attributes

Quantitative Attributes

- **Ratio**
  - is interval data with a natural zero point.
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚is twice as high as 5 K˚).
    - *E.g.,* Age (from 0 years to 100+)
    - Distance (measured with a ruler or other such measuring device)
    - Time interval (measured with a stop-watch or similar)

Attributes

Qualitative Attributes

- **Nominal:** represents some category or state and that's why nominal attribute also referred as **categorical attributes (categoric discrete data)**
  - Might not have a logical order
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes

Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

Attributes

Qualitative Attributes

- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

Attributes

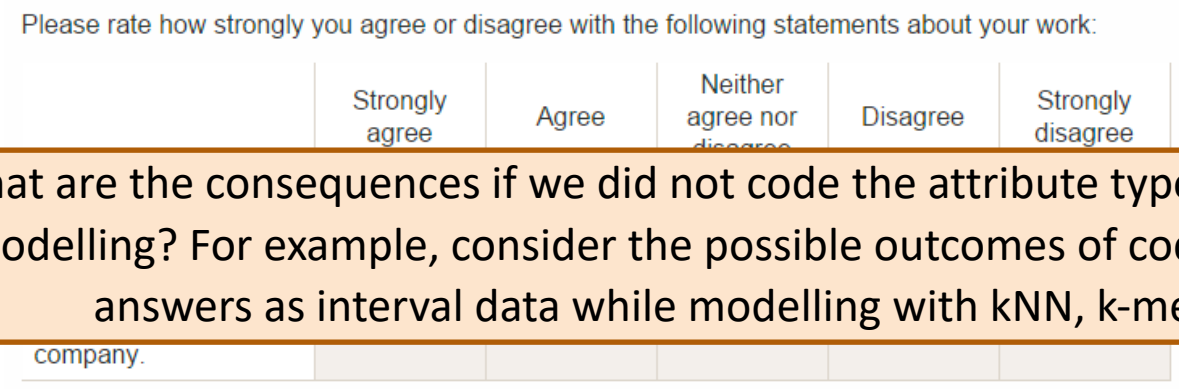
Qualitative Attributes

- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size =* {*small, medium, large*}, grades, army rankings
  - Opinion (agree, mostly agree, neutral, mostly disagree, disagree)
  - Time of day (morning, noon, night)
  - Tumour Stage (I, IIA, IIB, IIIA, IIIB, etc.)

Attributes

Qualitative Attributes

- **Likert style questions are ordinal data**
  - To be considered interval data, a scale or level of measurement must possess distance (i.e., the magnitude of the differences between successive numbers must be equal).
  - Also, with interval data, you must be able to perform addition and subtraction with the values (and the resulting numbers must make sense).
  - These definitely do not hold for Likert scale data.


Please rate how strongly you agree or disagree with the following statements about your work:

	Strongly agree	Agree	Neither agree nor	Disagree	Strongly disagree
company.					

What are the consequences if we did not code the attribute type correctly when modelling? For example, consider the possible outcomes of coding Likert type answers as interval data while modelling with kNN, k-means etc.

Attributes

Summary

Qualitative: entities are divided into distinct categories (each category may be represented by a unique string or number)

- Binary: There are only two categories e.g. dead or alive.
- Nominal: There are more than two categories e.g. whether someone is an omnivore, vegetarian, vegan, or fruitarian.
- Ordinal: The same as a nominal variable but the categories have a logical order e.g. whether people got a fail, a pass, a merit or a distinction in their exam.

Quantitative: entities get numerical scores (discrete or continuous)

- Interval : Equal intervals on the variable represent equal differences in the property being measured e.g. the difference between 6 and 8 is equivalent to the difference between 13 and 15.
- Ratio: The same as an interval variable, but the ratios of scores on the scale must also make sense e.g. a score of 16 on an anxiety scale means that the person is, in reality, twice as anxious as someone scoring 8).

Attributes

Summary

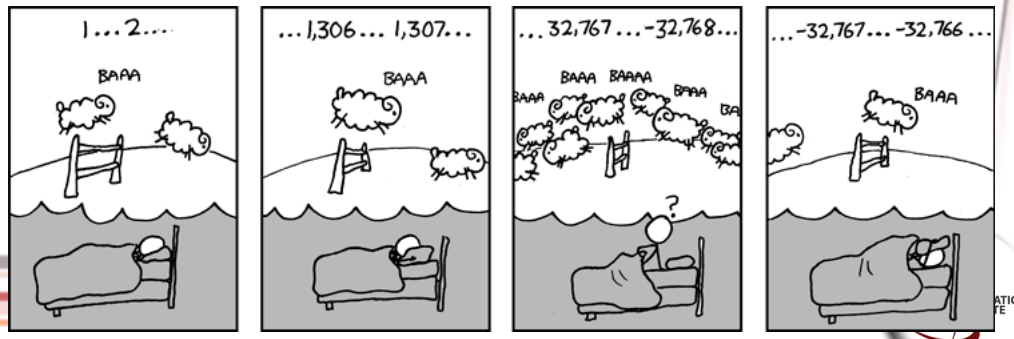
Operation	Nominal	Ordinal	Interval	Ratio
Equality	✔	✔	✔	✔
Order		✔	✔	✔
Add / subtract			✔	✔
Multiply / divide				✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Arithmetic mean			✔	✔
Geometric mean				✔

Attributes

- Context is important! The context of the study and the relevant questions of interest are important in specifying what kind of variable we will analyze.
- For example,
  - Did you get a flu? (Yes or No) -- is a binary nominal categorical variable
  - What was the severity of your flu? (Low, Medium, or High) -- is an ordinal categorical variable

Exercise

- Classify the following as either continuous or discrete data.
  - Number of road accidents in a month in Chicago
  - Customer satisfaction survey results (measured on a 1-5 scale)
  - Time taken to deliver a product to the customer in days
  - Money



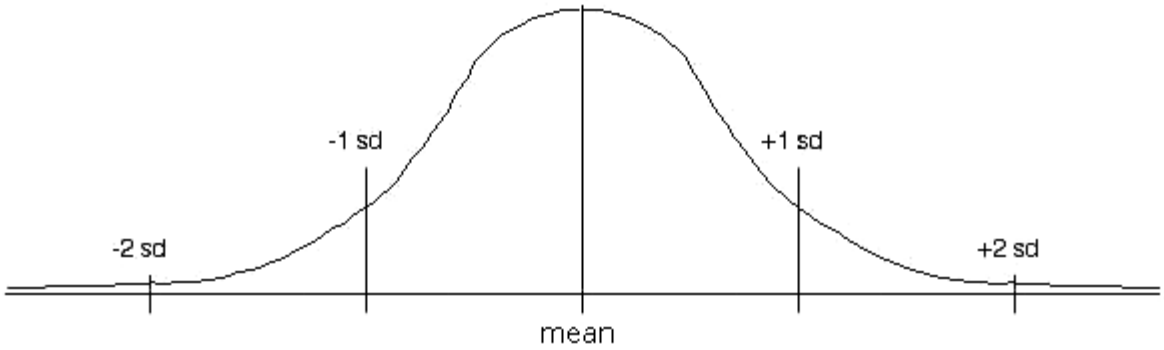
IS 709 Introduction to Data Science

Lecture 3 – Understanding Data- Part2

Statistical Descriptions of Data

Basic Statistical Descriptions of Data

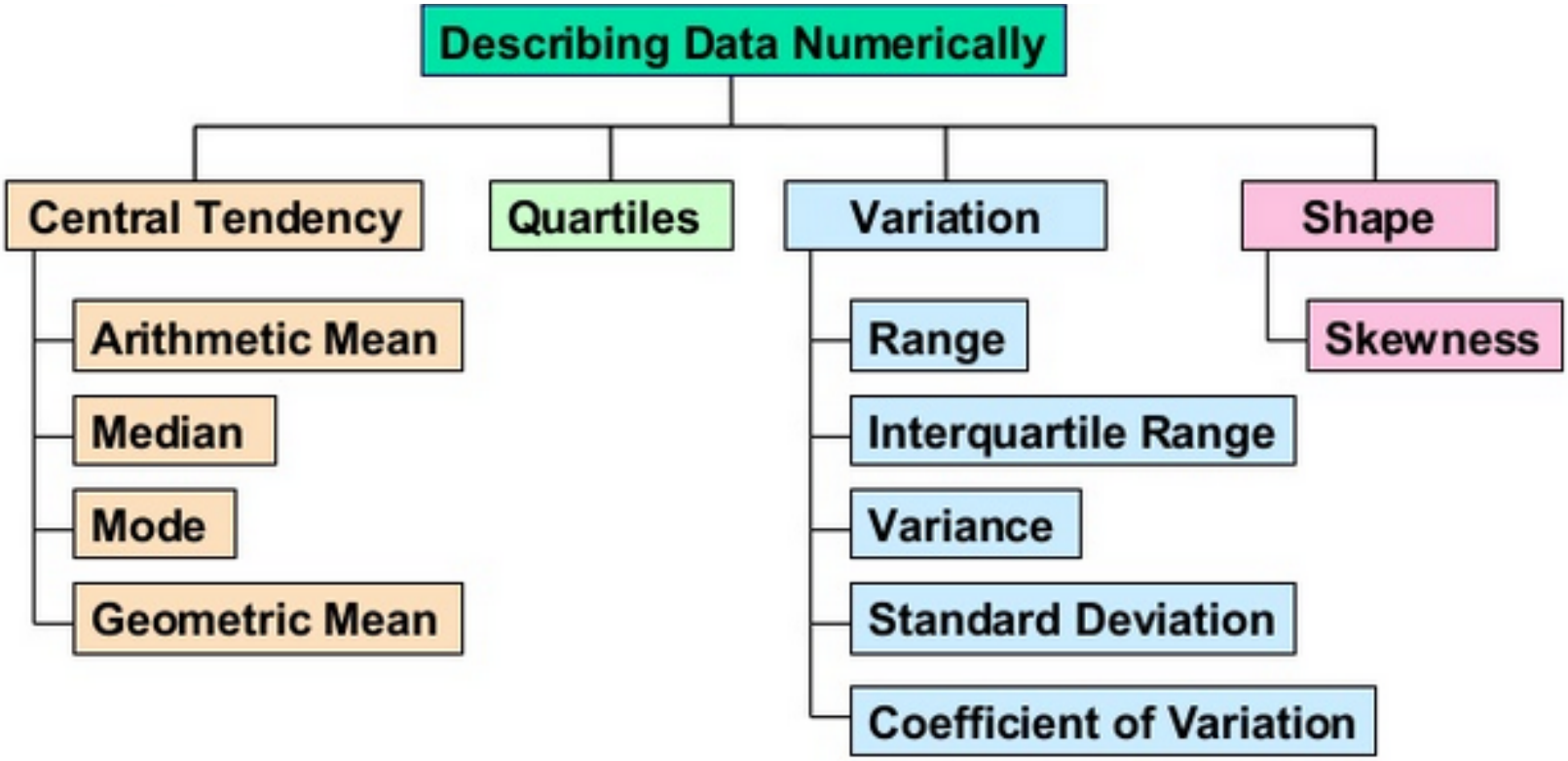
- Motivation
  - To better understand the data: central tendency, variation and spread
- Centrality and dispersion



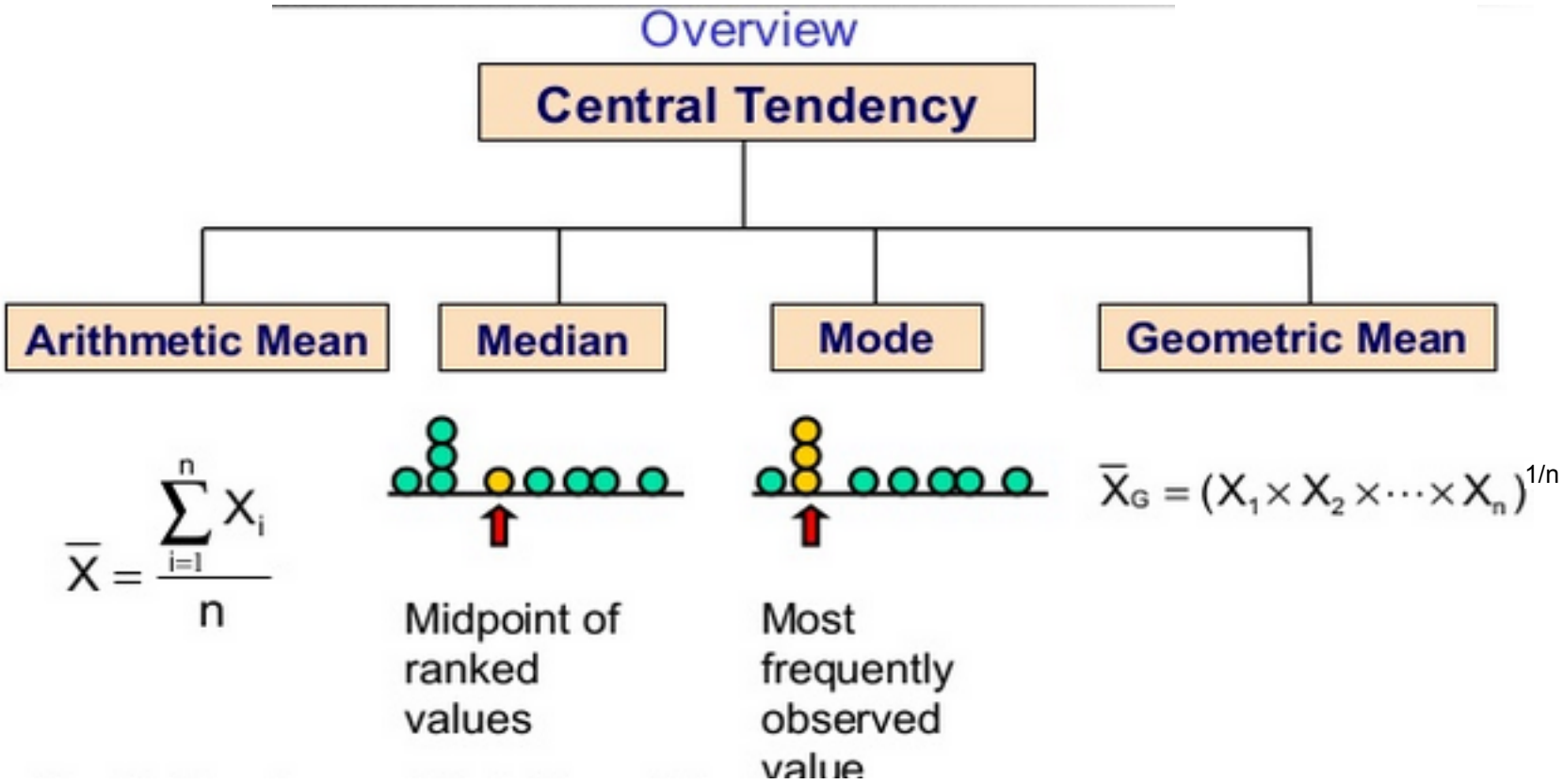
Central tendency refers to the idea that there is one number that best summarizes the entire set of measurements, a number that is in some way "central" to the set.

Dispersion refers to the idea that there is a second number which tells us how "spread out" all the measurements are from that central number.

Summary Measures



Measures of Central Tendency



Measuring the Central Tendency

Mean

- Mean (algebraic measure) (sample vs. population):

Note: *n* is sample size and *N* is population size.

(A) $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ (B) $$\mu = \frac{\sum x}{N}$$

The "mean" of a sample is the summary statistic computed with the first formula (A).

An "average" is one of many summary statistics you might choose to describe the typical value or the central tendency of a sample.

- Weighted arithmetic mean:

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

When to use weighted arithmetic mean?

Measuring the Central Tendency
Median

- Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for *grouped data*):

$$median = L_1 + (\frac{n/2 - (\sum freq)_l}{freq_{median}})width$$

	age	*frequency*
	1–5	200
	6–15	450
	16–20	300
Median interval →	21–50	1500
	51–80	700
	81–110	44

L_1 is the lower boundary of the median interval, n is the number of values in the entire set, $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and width is the width of the median interval.

n=3194
L1=20
$(\sum freq)_l$ = 950 (200+450+300)
$freq_{median}$ = 1500 (21-50 bracket frequency)
Width = 30 (50-21+1)
*Median = 20+(((3194/2)-950)/1500)*30 = 32.94*

ge of the

- Estimated by interpolation (for *grouped data*):

$$median = L_1 + (\frac{n/2 - (\sum freq)_l}{freq_{median}})width$$

	age	*frequency*
	1–5	200
	6–15	450
	16–20	300
Median interval →	21–50	1500
	51–80	700
	81–110	44

L_1 is the lower boundary of the median interval, n is the number of values in the entire set, $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and width is the width of the median interval.

Measuring the Central Tendency
Interpolated Median

- Interpolated median values are generally the most adequate measure of central tendency when there is a limited number of response categories, such as Likert scales or the level of education.
- In survey data, actual distributions frequently are non-continuous and non-normal.
  - The mean may thus be inappropriate to summarize the central tendency, and the median too rough because it is constrained to the actual categories in the data.
  - On a five-point scale, the median can only fall on one of these categories, and thus does not reflect smaller changes in the distribution. The interpolated median adjusts the median position to do just that.

Measuring the Central Tendency
Interpolated Median

- Interpolated median values are generally the most adequate measure of central tendency when there is a limited number of response categories, such as Likert scales or the level of education.
- For example:

Response	Question 1	Question 2
5 = Strongly agree	9	1
4 = Agree	10	10
3 = Neither agree nor disagree	0	6
2 = Disagree	1	1
1 = Strongly disagree	0	2
Number of points	20	20
Mean	4.35	3.35
Median	4	4
Intepolated median	4.4	3.6

Define variables as follows:

M = the standard median of the responses
nl = number of responses strictly less than *M*
ne = number of responses equal to *M*
ng = number of responses strictly greater than *M*

The interpolated median *IM* is then computed as follows:

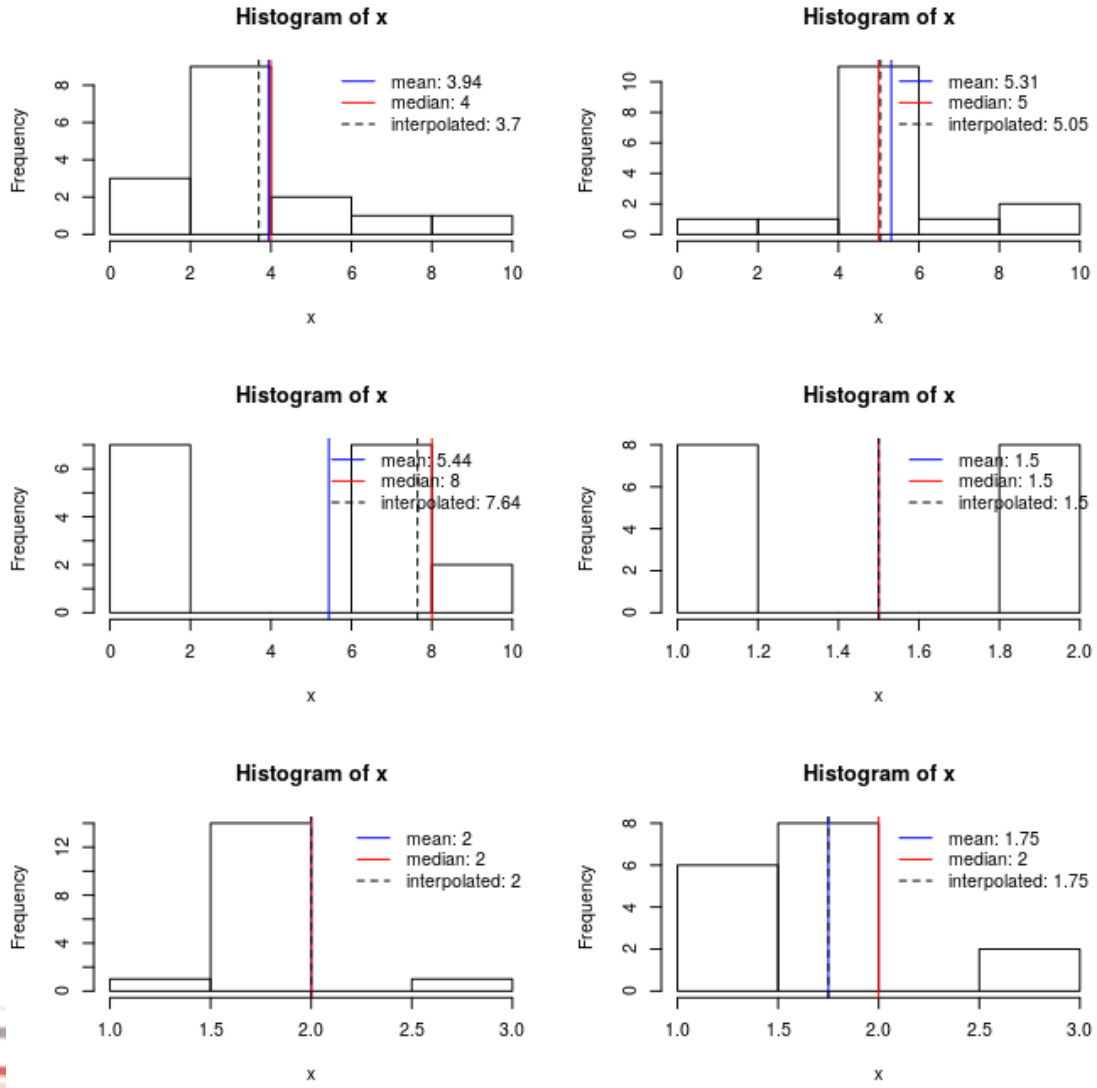
If *ne* is nonzero:

$IM = M + (ng - nl) / (2ne)$

If *ne* is zero:

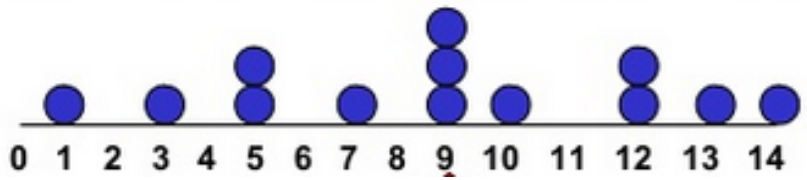
$IM = M$

Measuring the Central Tendency
Interpolated Median

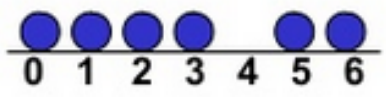


Measuring the Central Tendency
Mode

- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical (nominal) data
- There may be no mode or several modes



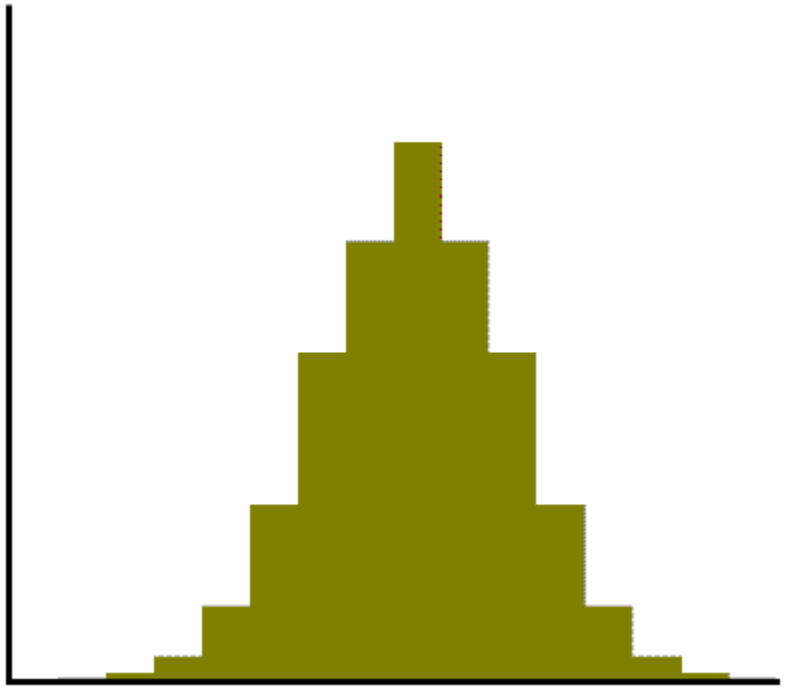
Mode = 9



No mode

Measuring the Central Tendency

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - **Can you point the mean, median and mode values in both distributions?**

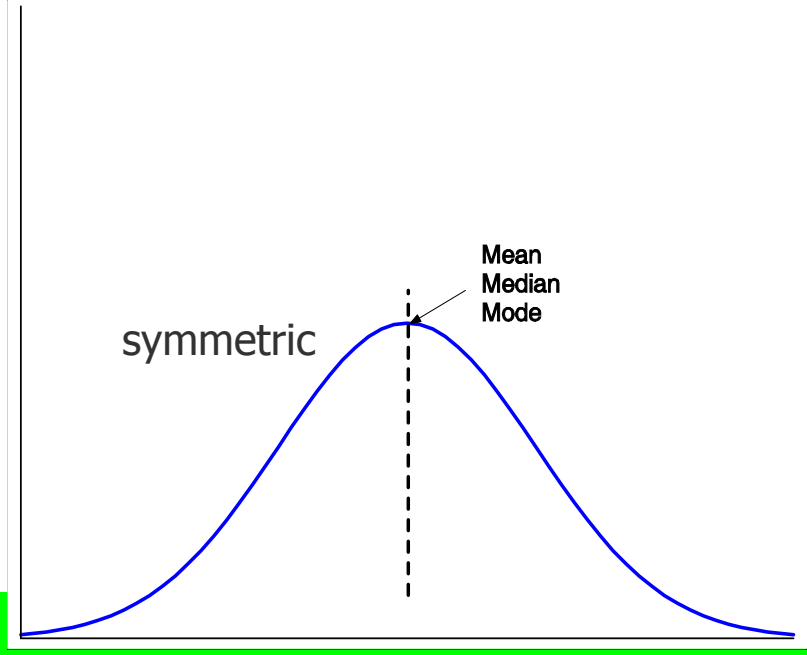


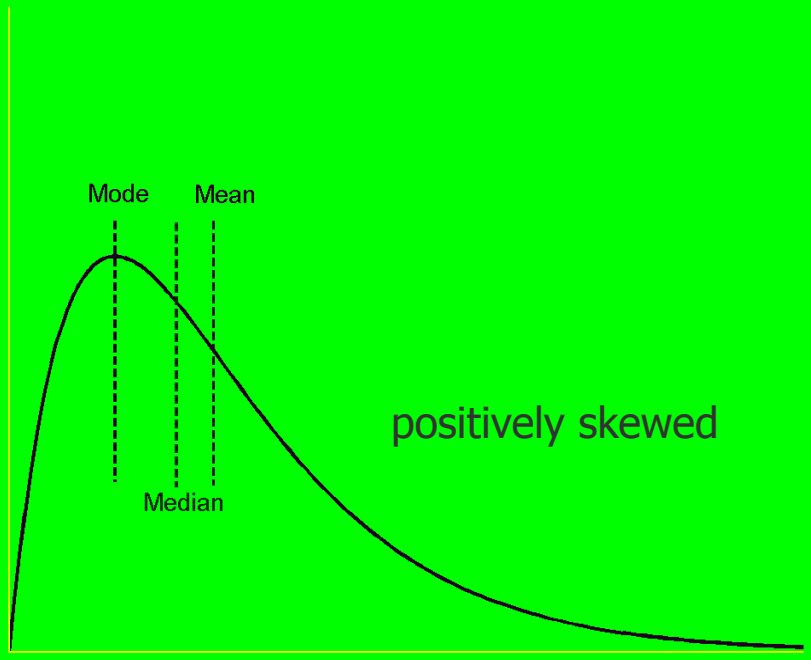
Unimodal dist.

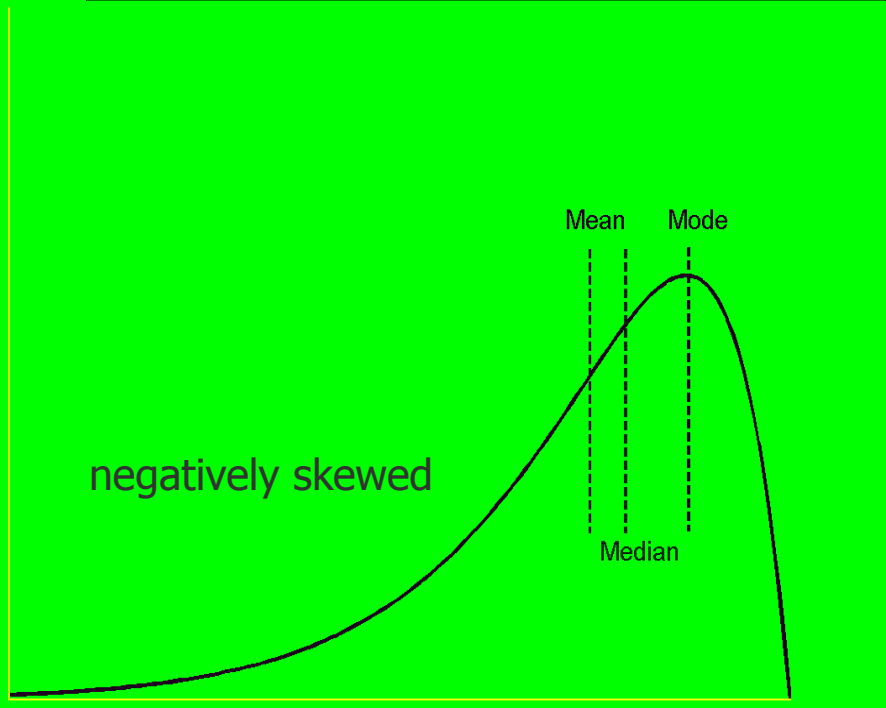
Bimodal dist.

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data





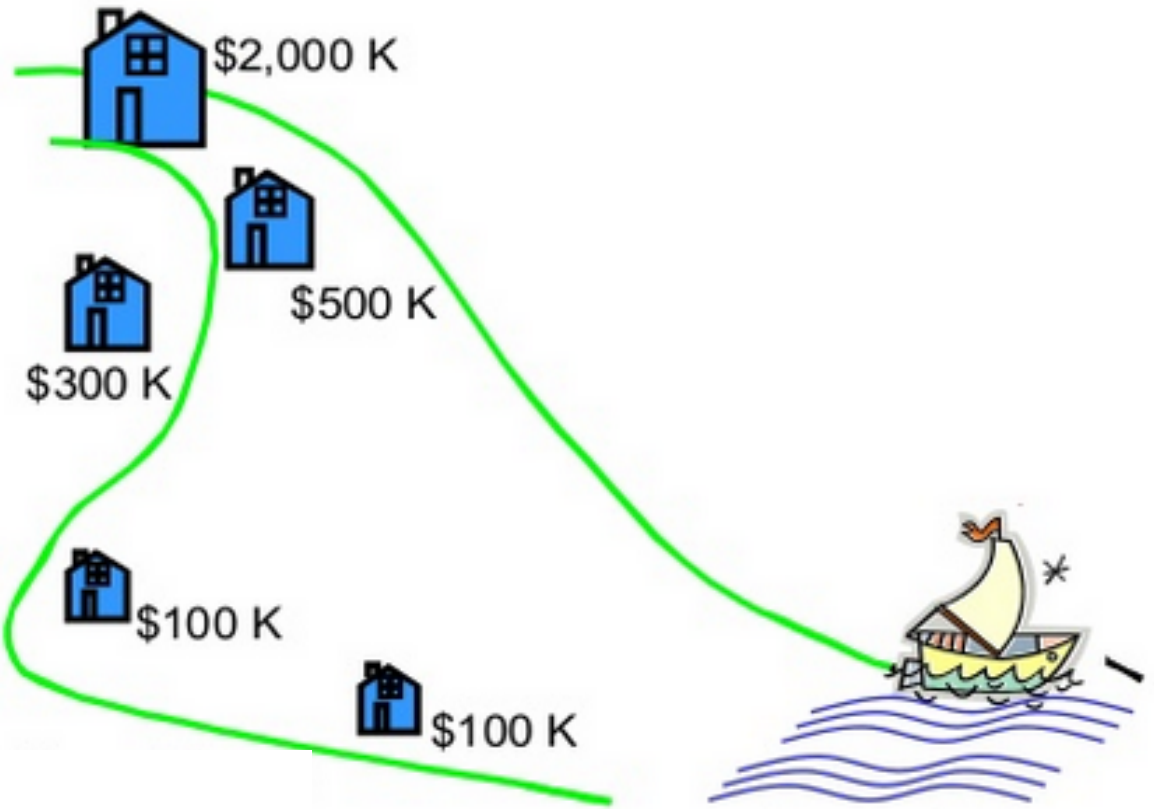


Review Example

- Five houses on a hill by the beach

House Prices:

$2,000,000
500,000
300,000
100,000
100,000



Find the mean, median and mode of the house prices.

Review Example

Summary Statistics

- Five houses on a hill by the beach

House Prices:
$2,000,000
500,000
300,000
100,000
100,000

Sum= $3,000,000
Mean = ($3,000,000 /5) =$600,000
Median = $300,000
Mode = $100,000

Discuss which measure of location is the most preferable.

Exercises

- Consider the following two data sets:
  - Ds1: 108 112 116 120 124
  - Ds2: 108 112 116 120 205
- What do the mean and median values tell us?

Measuring the Central Tendency

Geometric Mean

- The *geometric mean* of a set of *n* numbers is the n^{th} root of their product.
- Used to measure the rate of change of a variable over time.

$$\overline{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

- Geometric mean rate of return
  - Measures the status of an investment over time

$$\overline{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$

  - where r is the rate of return in time period i

Measuring the Central Tendency

Geometric Mean

- Example:An investor has annual return of 5%, 10%, 20%, -50%, and 20%.
- Using the arithmetic mean, the investor's total return is (5%+10%+20%-50% +20%)/5 = 1%

Year	Starting Equity	Return %	Return $	Closing equity
1	$1,000	5%	$50	$1,050
2	$1,050	10%	$105	$1,155
3	$1,155	20%	$231	$1,386
4	$1,386	-50%	-$693	$693
5	$693	20%	$138.6	$831.6

The actual 5 year return on the account is ($831.6 – $1,000)/$1,000 = -16.84%
The geometric mean is used to tackle continuous data series which the arithmetic mean is unable to calculate.

5th Square Root of ((1 + 0.05)(1 + 0.1)(1 + 0.2)(1 – 0.5)(1 + 0.2)) – 1 = -0.03621
Multiply the result by 100 to calculate the percentage. This results in a -3.62% annual return.

Measuring the Central Tendency

Geometric Mean

- Example:An investor has annual return of 5%, 10%, 20%, -50%, and 20%.
- Using the arithmetic mean: (5%+10%+20%-50%+20%)/5 = 1%

=1000 x -0.03621= -36.21

Year	Starting Equity	Return %	Return $	Closing equity	Annual loss	Net
1	$1,000	5%	$50	$1,050	-36.21	964
2	$1,050	10%	$105	$1,155	-34.90	929
3	$1,155	20%	$231	$1,386	-33.64	895
4	$1,386	-50%	-$693	$693	-32.42	863
5	$693	20%	$138.6	$831.6	-31.24	~832

The actual 5 year return on the account is ($831.6 – $1,000)/$1,000 = -16.84%
The geometric mean is used to tackle continuous data series which the arithmetic mean is unable to calculate.

5th Square Root of ((1 + 0.05)(1 + 0.1)(1 + 0.2)(1 – 0.5)(1 + 0.2)) – 1 = -0.03621
Multiply the result by 100 to calculate the percentage. This results in a -3.62% annual return.

Measuring the Central Tendency

Geometric Mean

- Also used when the ranges are different.
  - For example: Suppose college applicants are rated on SAT score (0 to 800), grade point average in history (0 to 4) and extracurricular activities (1 to 10).
  - Student 1: SAT=800, GPA=3, EA=8
    - $\sqrt[3]{800 \text{ x } 3 \text{ x } 8} = 26.77$
  - Student 2: SAT=400, GPA=2, EA=10
    - $\sqrt[3]{400 \text{ x } 2 \text{ x } 10} = 20$

Measuring the Central Tendency

Geometric Mean

- The Geometric Mean is useful when we want to compare things with very different properties.

Example: you want to buy a new camera.

- One camera has a zoom of 200 and gets an 8 in reviews,
- The other has a zoom of 250 and gets a 6 in reviews.

Comparing using the usual arithmetic mean gives (200+8)/2 = **104** vs (250+6)/2 = **128**. The zoom is such a big number that the user rating gets lost.

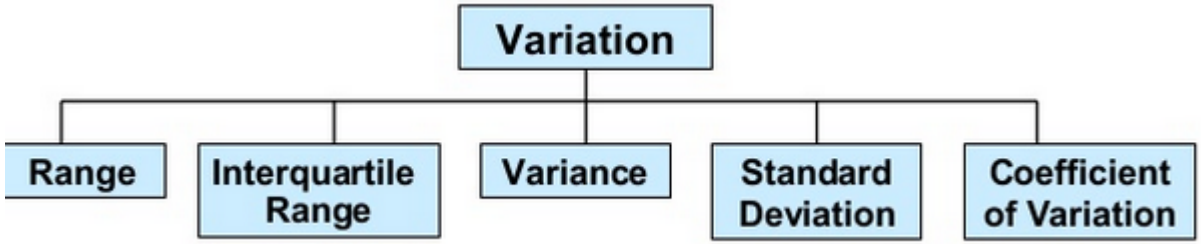
But the geometric means of the two cameras are:

- $\sqrt{(200 \times 8)} = \mathbf{40}$
- $\sqrt{(250 \times 6)} = \mathbf{38.7...}$

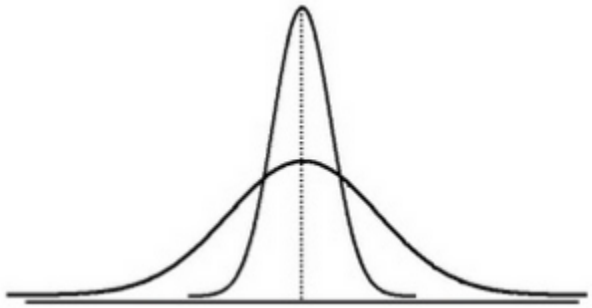
So, even though the zoom is 50 bigger, the lower user rating of 6 is still important.

Ref: https://www.mathsisfun.com/numbers/geometric-mean.html

Measuring the Dispersion of Data



- Measures of variation give information on the spread or variability of the data values.



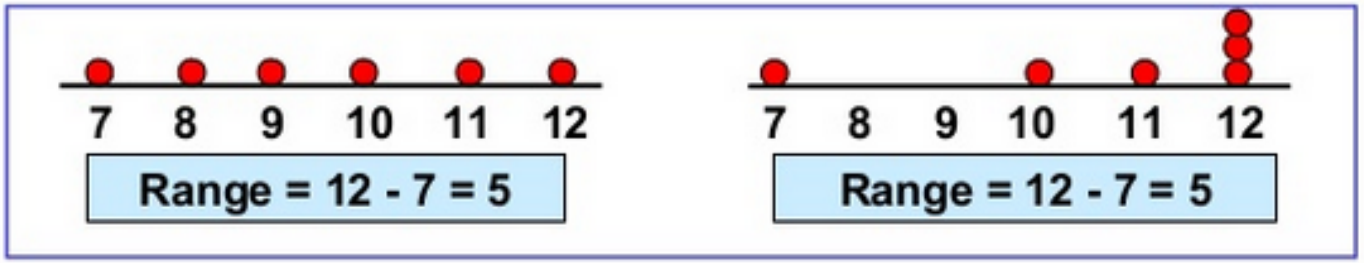
Same center
Different Variation

Measuring the Dispersion of Data
Range

- Simplest measure of variation
- Difference between the largest and the smallest values in a set of data
- Range= $X_{largest}$-$X_{smallest}$

Measuring the Dispersion of Data
Range

- Disadvantages of the range:
  - Ignores the way in which data are distributed.



- Sensitive to outliers

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

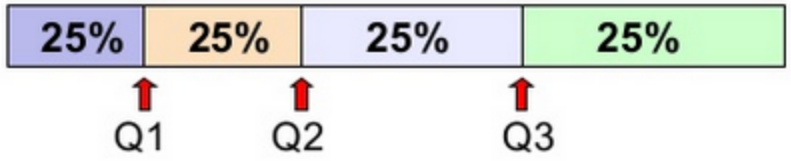
Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

Measuring the Dispersion of Data
Interquartile Range

- Quartiles split the ranked data into 4 segments with an equal number of values per segment.

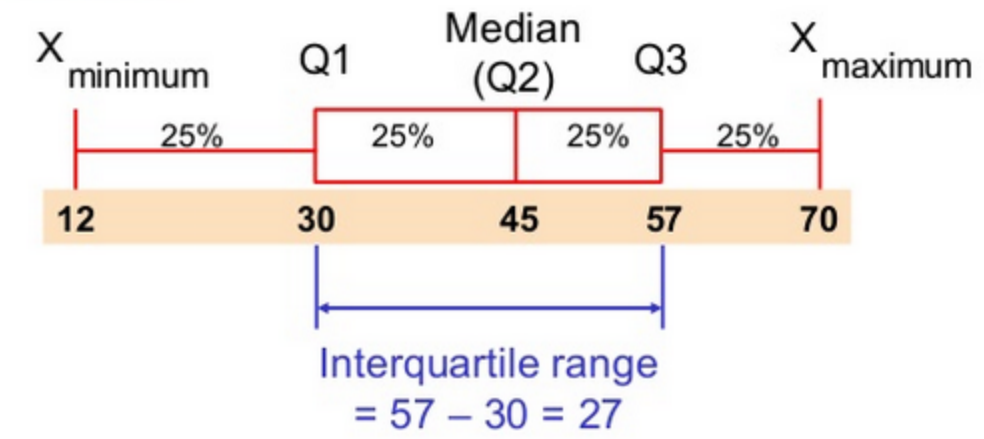


- The first quartile, Q1, is the value for which 25% of the observations are smaller and 75% are larger.
- Q2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile.

Measuring the Dispersion of Data
Interquartile Range

- Find a quartile by determining the value in the appropriate position in the ranked data, where
- First quartile position: Q1 = (n+1)/4
- Second quartile position: Q2 = (n+1)/2 (the median position)
- Third quartile position: Q3 = 3(n+1)/4
  - where n is the number of observed values.

Example:

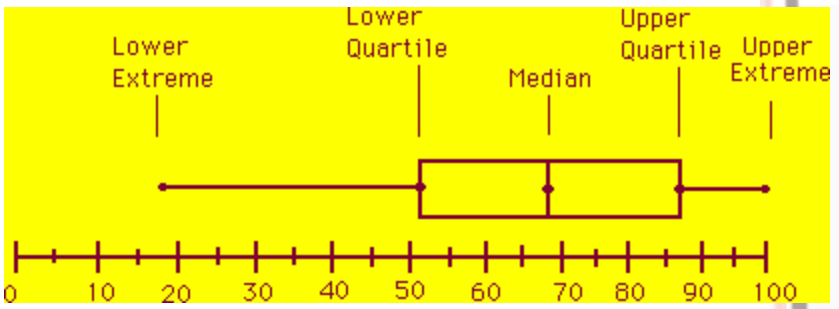


Measuring the Dispersion of Data
Interquartile Range

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: $IQR = Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value lower/higher than 1.5 x IQR from Q_1/Q_3

What is the advantage of using quartiles?

Boxplot Analysis

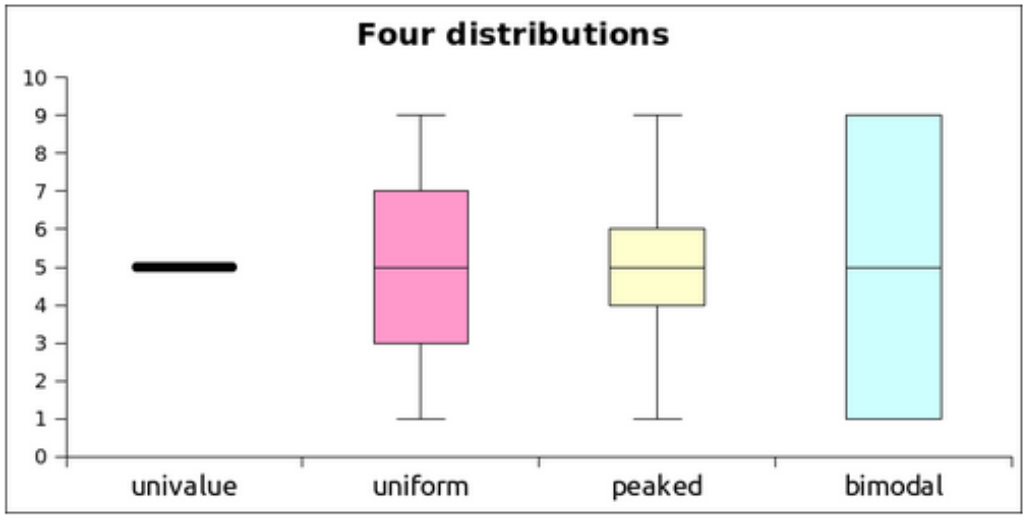


- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually

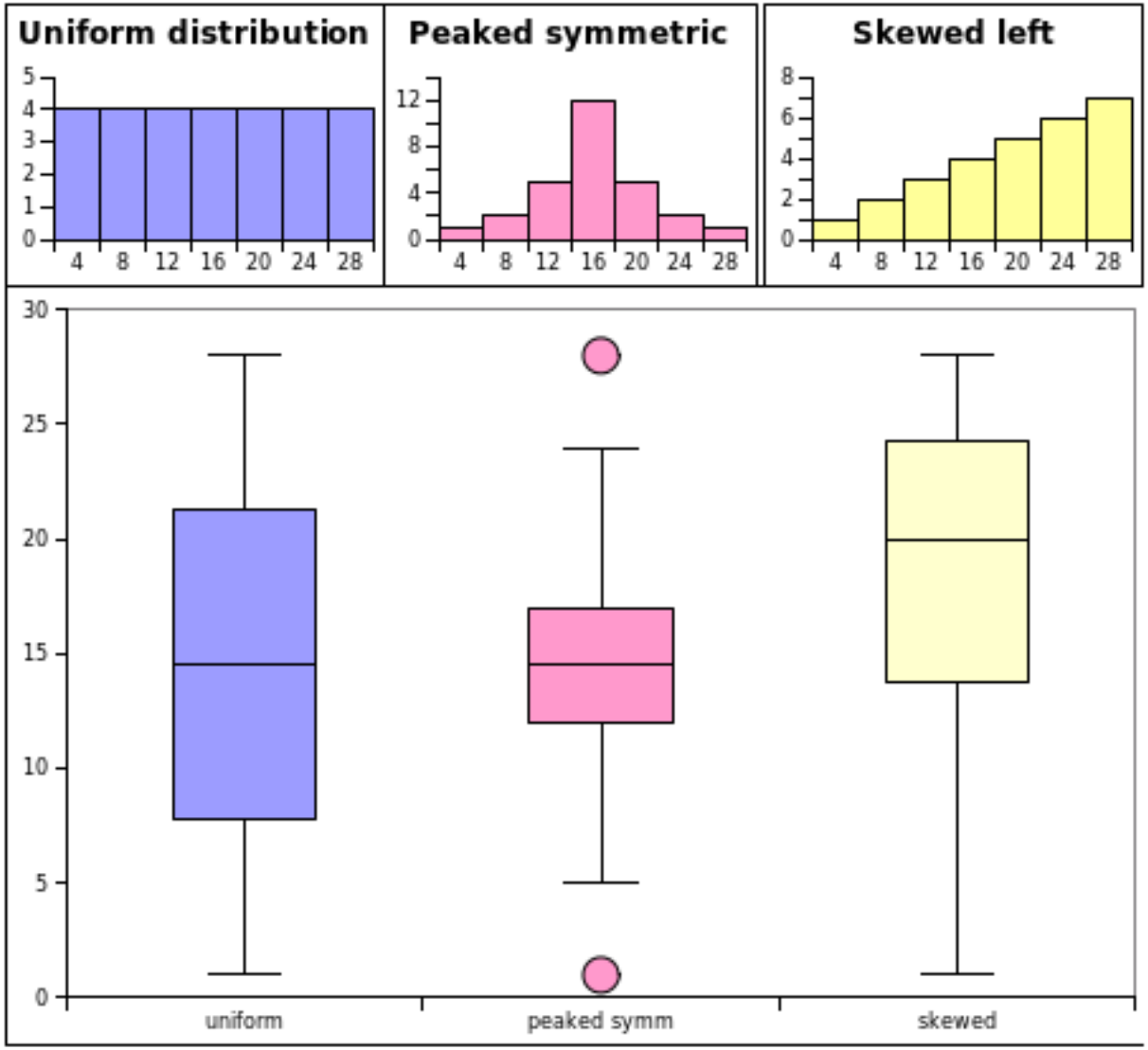
Boxplot Analysis

univalue	uniform	peaked	symmetric	bimodal
5	1	1	1	
5	2	4	1	
5	3	4	1	
5	4	5	1	
5	5	5	5	← Median
5	6	5	9	
5	7	6	9	
5	8	6	9	
5	9	9	9	

Box plots display how the data is spread across the range based on the quartile information above.

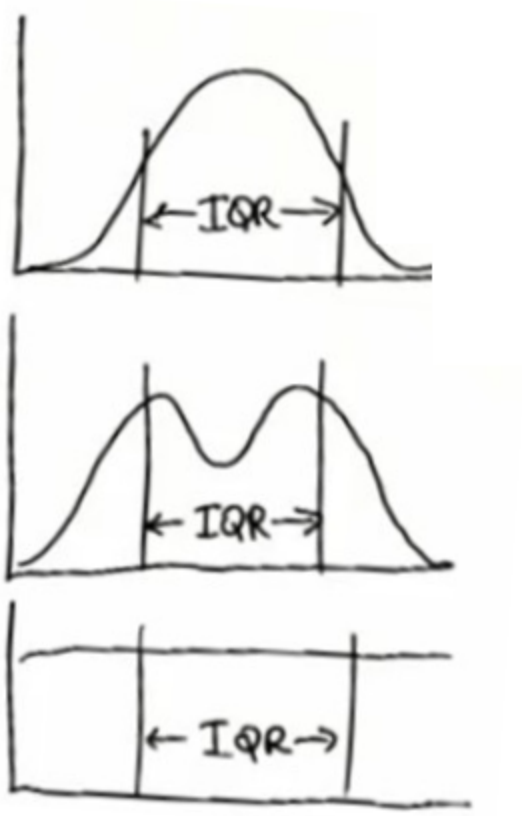


Boxplot Analysis



Boxplot Analysis

Problem with interquartile range



Measuring the Dispersion of Data

- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*
    - Has the same units as the original data

Scale=statistical dispersion
Example: The normal distribution has two parameters: a location parameter μ and a scale parameter σ.

Measuring the Dispersion of Data

Example:

Sample Data (X_i) : 10 12 14 15 17 18 18 24

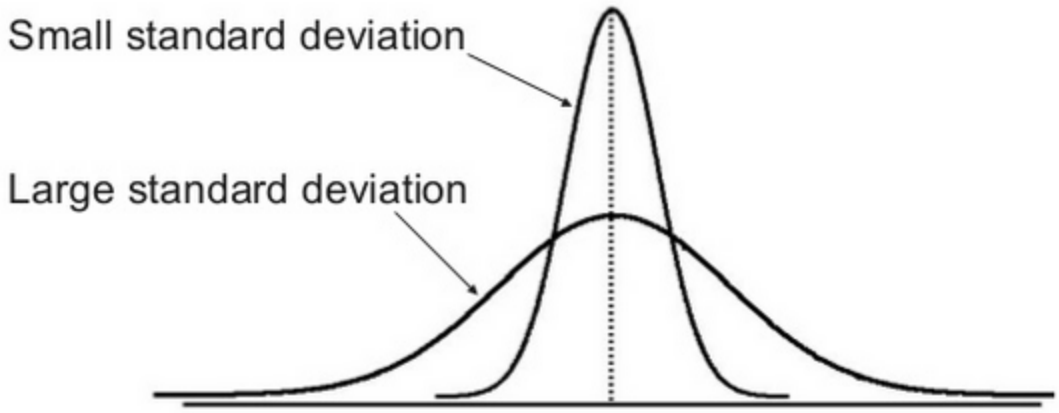
n = 8 Mean = $\overline{X}$ = 16

$$S = \sqrt{\frac{(10-\overline{X})^2 + (12-\overline{X})^2 + (14-\overline{X})^2 + \cdots + (24-\overline{X})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

=4.3095 <- A measure of the «average» scatter around the mean.

Measuring the Dispersion of Data



Advantages:

1. Each value in the data set is used in the calculation.
2. Values far from the mean are given extra weight.
   (because deviations from the mean are squared)

Measuring the Dispersion of Data
Coefficient of Variation

- Measures relative variation
  - AKA relative standard deviation (RSD)
- Always in percentage (%)
- Shows variation relative to the mean
- Can be used to compare two or more sets of data measured in different units

$$CV = \left(\frac{S}{\bar{X}}\right).100\%$$

Measuring the Dispersion of Data
Coefficient of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5
- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_A = \left(\frac{S}{\bar{X}}\right).100\% = \frac{\$5}{\$50}.100\% = 10\%$$

$$CV_B = \left(\frac{S}{\bar{X}}\right).100\% = \frac{\$5}{\$100}.100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price.

Measuring the Dispersion of Data
Coefficient of Variation

A data set of [100, 100, 100] has constant values. Its standard deviation is 0 and average is 100, giving the coefficient of variation as

0 / 100 = 0

A data set of [90, 100, 110] has more variability. Its standard deviation is 8.16 and its average is 100, giving the coefficient of variation as

8.6 / 100 = 0.086

A data set of [1, 5, 6, 8, 10, 40, 65, 88] has still more variability. Its standard deviation is 30.779 and its average is 27.875, giving a coefficient of variation of

30.779/ 27.875 = 1.104

Measuring the Dispersion of Data
Coefficient of Variation

We want to compare the homogenity of two districts (ilce) in Turkey in terms of the level of income of people.

District A= In some quarters (mahalle), high status people are living (rental prices are high), whereas in some quarters low status people are living (rental prices are low).

Monthly rental fees are: 3000, 2500, 3100, 400, 500, 200

Mean =1616
Standard deviation= 1388

COV(district A)= (1388/1616)x100=85.83

District B: Only high status people are living.

Monthly rental fees of quarters are: 3000, 2500, 3100, 2900,3200, 3050

Mean =2958
Standard deviation= 246

COV(district B)= (246/2958)x100=8.31

Conclusion: District B is more homogenous than District A.

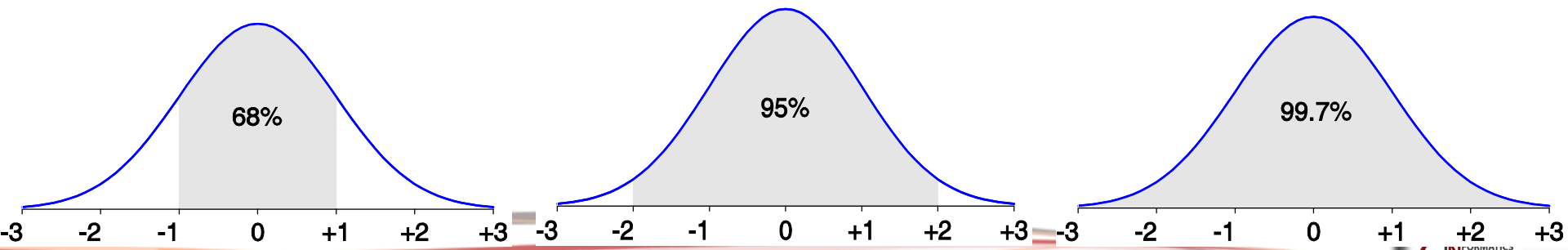
Z-scores

- A measure of distance from the mean (for example, a Z-score of 2.0 means that a value is 2.0 standard deviations from the mean)
- The difference between a value and the mean, divided by the standard deviation
- A Z-score above 3 or below -3 is considered an outlier

$$Z = \left(\frac{X - \bar{X}}{S} \right)$$

Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it



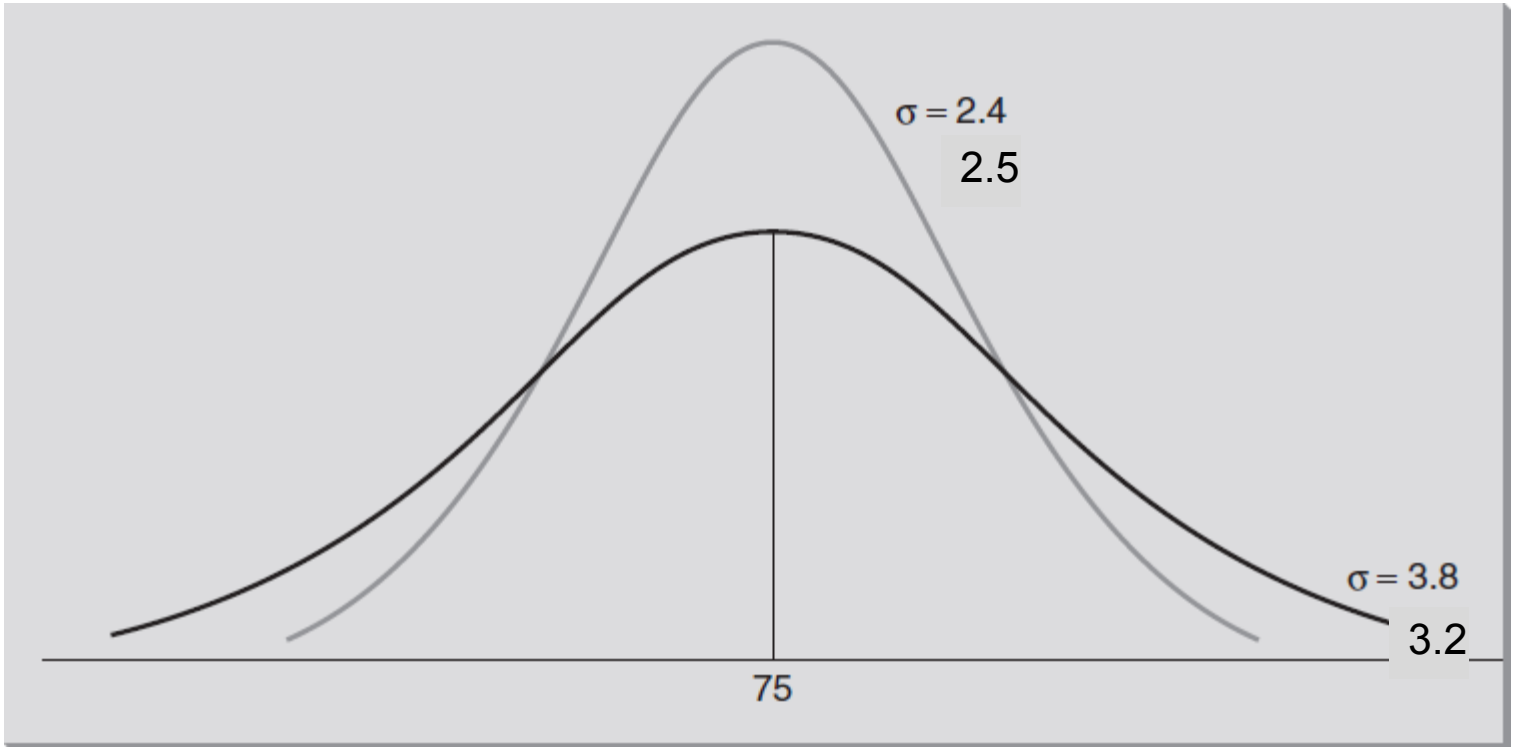
Properties of Normal Distribution Curve: Example

- A student scored 79 out of 100 on the final exam in statistics course and 60 out of 100 on the final exam in the research course. Can the student conclude that her performance was better in statistics because of the higher score in the statistics course than the research course?

Let us assume that the final exam in statistics had a mean of 75 with a standard deviation of 3 and the final exam in research had a mean of 40 with a standard deviation of 2.5.

Ref: http://samples.jbpub.com/9781449686697/45670_CH06_101_132.pdf

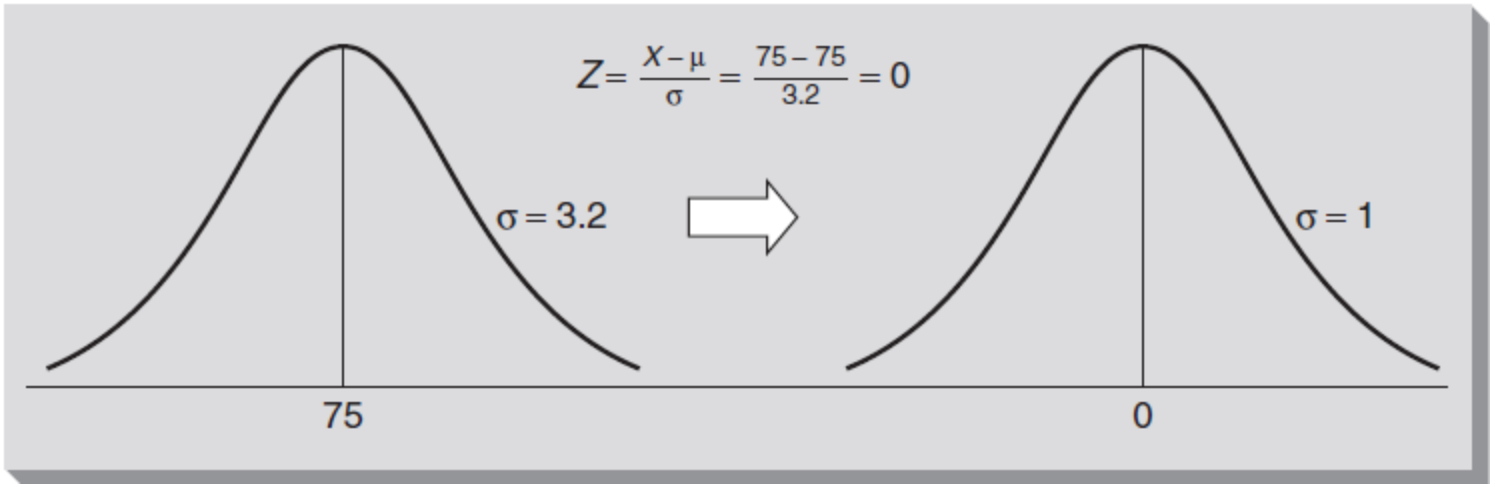
Properties of Normal Distribution Curve: Example



- Normal distributions with different standard deviations
- Final exam in statistics had a mean of 75 with a standard deviation of 3.2 and the final exam in research had a mean of 40 with a standard deviation of 2.5.

Ref: http://samples.jbpub.com/9781449686697/45670_CH06_101_132.pdf

Properties of Normal Distribution Curve



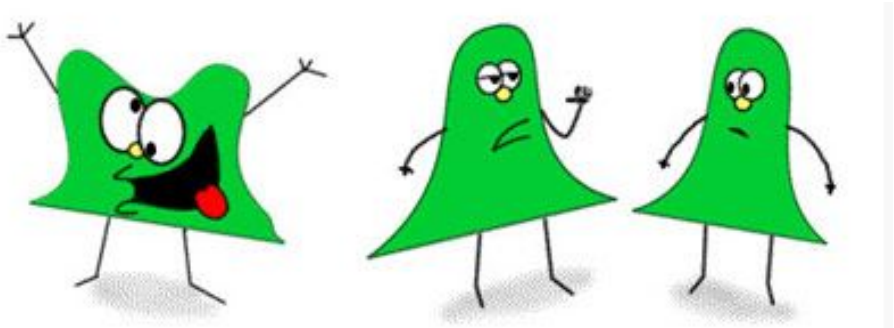
- z_score(79)=1.25 in statistics
- z_score(60)=8 in research

Her z-score of 8 tells us that her score is eight standard deviations above the average score of 40 since a standard normal distribution has a standard deviation of 1.

Ref: http://samples.jbpub.com/9781449686697/45670_CH06_101_132.pdf

Properties of Normal Distribution Curve

- Why do we care normality of a data sample?
- Because many models in data mining and statistical modelling techniques assume normality in the data.



"KEEP YOUR EYE ON THAT GUY, TOM. HES NOT, YOU KNOW...NORMAL!"

IS 709 Introduction to Data Science

Lecture 3 – Understanding Data- Part3

Graphic Displays of Basic Statistical Descriptions

Graphic Displays of Basic Statistical Descriptions

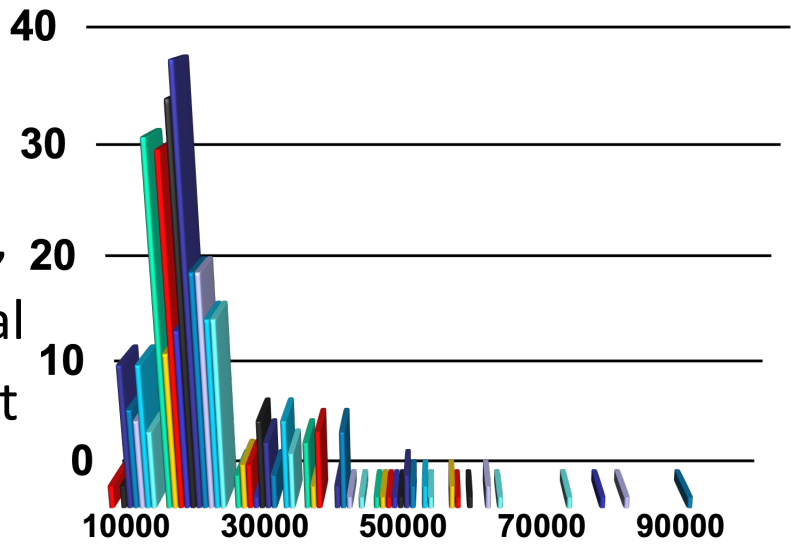
- Mean, median, variance etc. are important and they are part of the ***summary statistics***:
  - **Definition**: The result of a computation that reduces a dataset to a single number (or at least a smaller set of numbers) that captures some characteristic of the data.
- Summary statistics are concise, but dangerous, because they obscure the data. An alternative is to look at the distribution of the data, which describes how often each value appears
  - Example: histogram, box plot, bean plots

Graphic Displays of Basic Statistical Descriptions

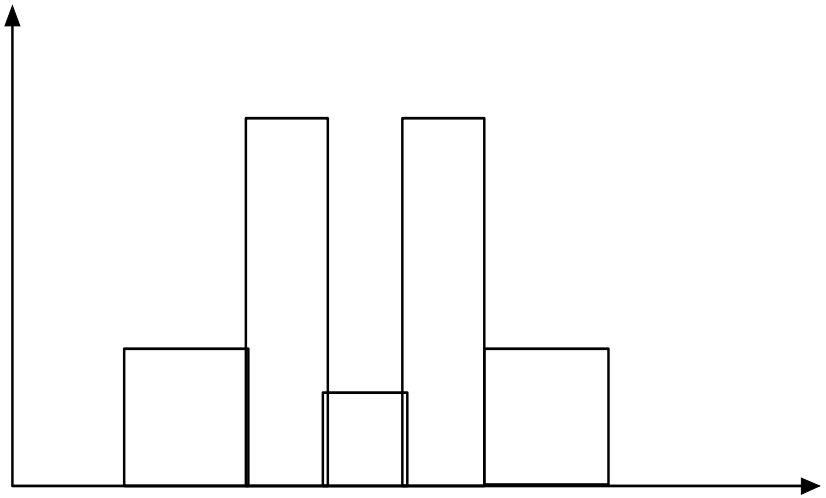
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis are frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

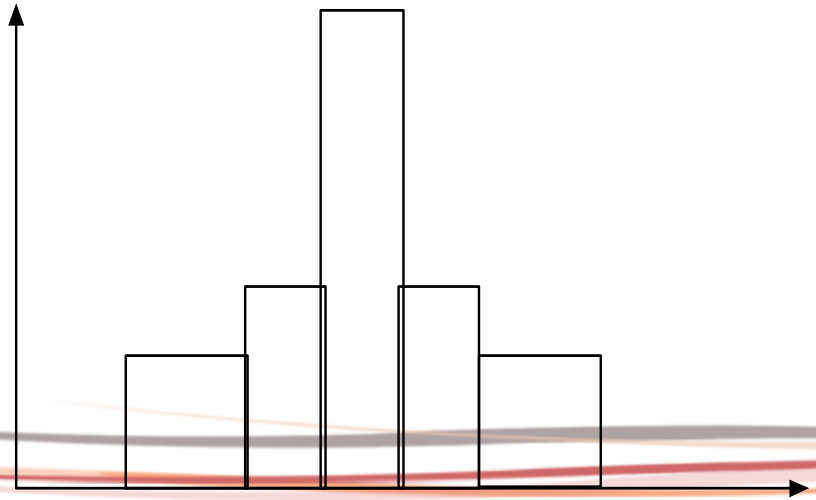
Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



Histograms Often Tell More than Boxplots

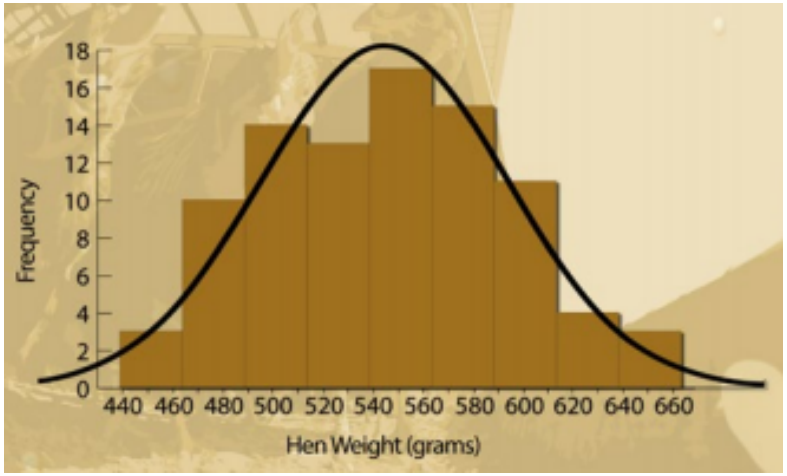


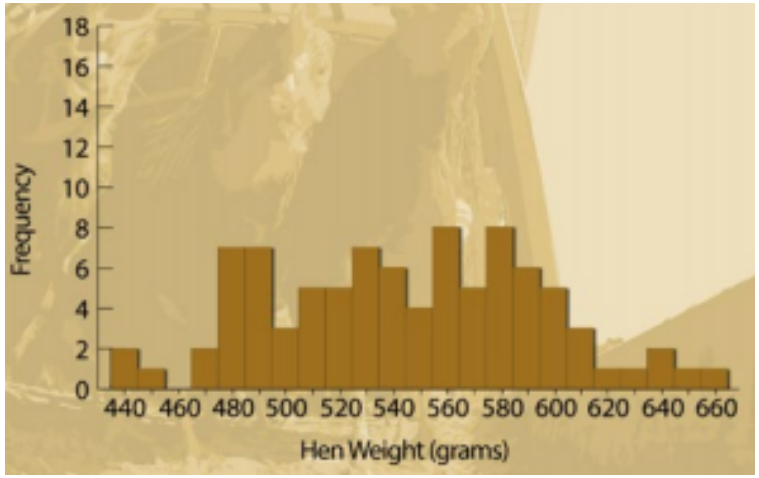


- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Histograms Often Tell More than Boxplots

- It's important to consider the class size when we are eyeballing a histogram to see if the data are normal. Sometimes, changing the class size can really change the way the histogram looks and what once appeared perfectly bell-shaped now looks quite different. The histogram of hen weights in both figures differ:

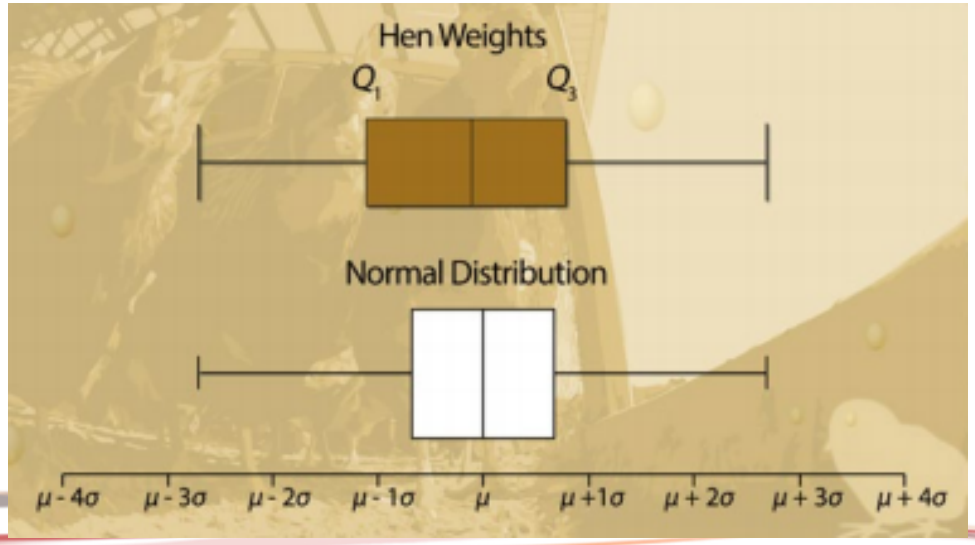




Changing the class size

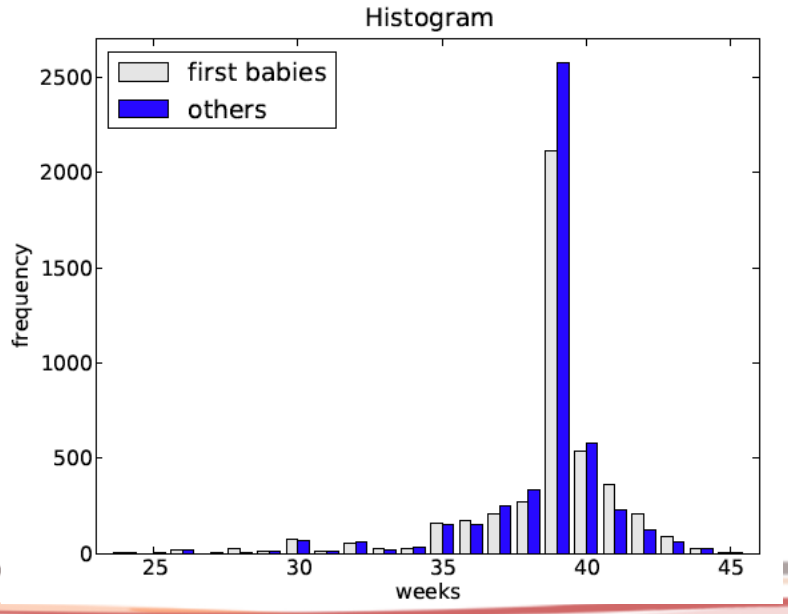
Histograms Often Tell More than Boxplots

- To assess normality, we can use the same hen weights to construct a boxplot.
- Boxplots can act as another graphical display test to see if our data are normally distributed.
  - If a distribution is normal, we would expect to see the box containing the middle 50% of the data to be pretty tightly grouped in the center of the distribution, with longer whiskers indicating the increased spread of the upper and lower quarters of the data.

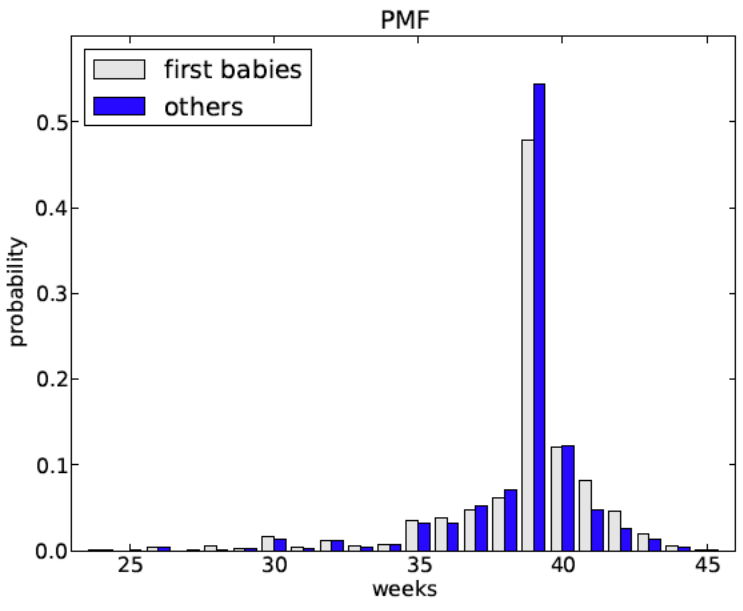


Histograms vs Probability Mass Function (PMF)

- In histograms, we compute frequencies with a dictionary.
- A probability is a frequency expressed as a fraction of the sample size, *n*.
- To get from frequencies to probabilities, we divide through by *n*, which is called normalization.
- The normalized histogram is called a PMF, which stands for "probability mass function"; that is, it's a function that maps from values to probabilities.



Histogram of pregnancy lengths.



PMF of pregnancy lengths

Other Summarization Methods

- **Risk ratio:** The relative risk (or risk ratio) is an intuitive way to compare the risks for the two groups. Simply divide the cumulative incidence in **exposed** group by the cumulative incidence in the **unexposed** group:

$$\text{Risk Ratio} = \frac{CI_e}{CI_u}$$

where CI_e is the cumulative incidence in the 'exposed' group and CI_u is the cumulative incidence in the 'unexposed' group.

A risk ratio <1: suggests that the exposure being considered is associated with a reduction in risk.

% decrease = (1 - RR) x 100

A risk ratio > 1 suggests an increased risk of that outcome in the exposed group.

% increase = (RR - 1) x 100

If the risk ratio is 1 (or close to 1), it suggests no difference or little difference in risk (incidence in each group is the same).

Other Summarization Methods

- **Risk ratio**, which is a ratio of two probabilities.
  - For example, the probability that a first baby is born early is 18.2%. For other babies it is 16.8%, so the relative risk is 1.08. That means that first babies are about 8% more likely to be early.
    - (18.2%/16.8%=1.08 then *% in increase* is computed as: (1.08-1)x100=0.08)
  - If we hypothetically find that 17% of smokers develop lung cancer and 1% of non-smokers develop lung cancer, then we can calculate the relative risk of lung cancer in smokers versus non-smokers as:
    - Relative Risk = 17% / 1% = 17
    - Thus, smokers are 17 times more likely to develop lung cancer than non-smokers.

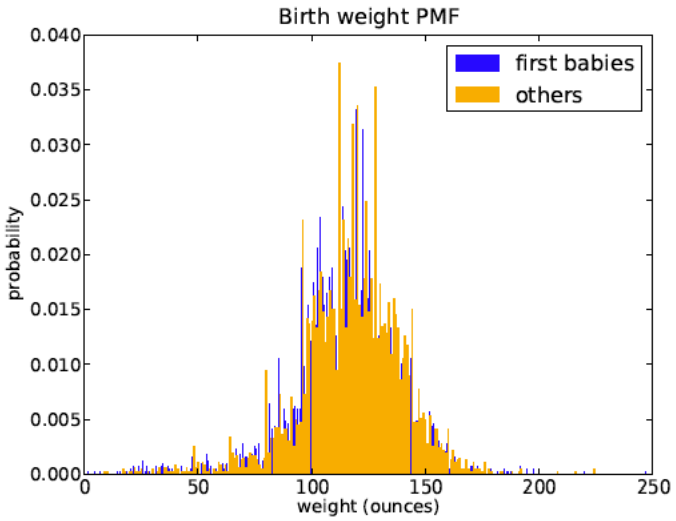
Limitations of PMF

- PMFs work well if the number of values is small. But as the number of values increases, the probability associated with each value gets smaller and the effect of random noise increases.
- For example, we might be interested in the distribution of birth weights. The below figure shows the PMF of these values for first babies and others.
- But parts of this figure are hard to interpret. For example which distribution do you think has the higher mean?

These problems can be mitigated by binning the data.

Binning can be useful, but it is tricky to get the size of the bins right. If they are big enough to smooth out noise, they might also smooth out useful information.

An alternative that avoids these problems is the cumulative distribution function, or CDF.



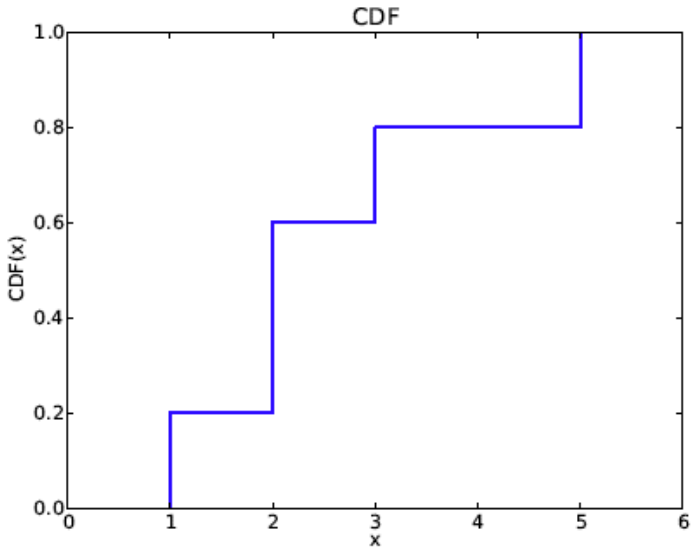
Cumulative Distribution Functions (CDF)

- The CDF is the function that maps values to their percentile rank in a distribution.
- The CDF is a function of *x*, where *x* is any value that might appear in the distribution. To evaluate CDF(x) for a particular value of *x*, we compute the fraction of the values in the sample less than (or equal to) *x*.
- As an example, suppose a sample has the values {1, 2, 2, 3, 5}. Here are some values from its CDF:

CDF(0) = 0	= 0	
CDF(1) = 0.2	= (1/5)=0.2	{1}
CDF(2) = 0.6	= (3/5)=0.6	{1,2,2}
CDF(3) = 0.8	= (4/5)=0.8	{1,2,2,3}
CDF(4) = 0.8	= (4/5)=0.8	{1,2,2,3} - no 4
CDF(5) = 1	= (5/5)=1	{1,2,2,3,5}

Cumulative Distribution Functions (CDF)

- We can evaluate the CDF for any value of *x*, not just values that appear in the sample.
  - If *x* is less than the smallest value in the sample, CDF(x) is 0.
  - If *x* is greater than the largest value, CDF(x) is 1.
  - The below figure is a graphical representation of this CDF. The CDF of a sample is a step function.



Bean Plots

- A bean plot is a plot in which (one or) multiple batches ("beans") are shown.
- Each bean consists of a density trace, which is mirrored to form a polygon shape.
- Next to that, a one-dimensional scatter plot shows all the individual measurements, like in a strip chart.
- The scatter plot is drawn using one small line for each observation in a batch. If a small line is drawn outside of the density shape, a different colour is used to draw the line.

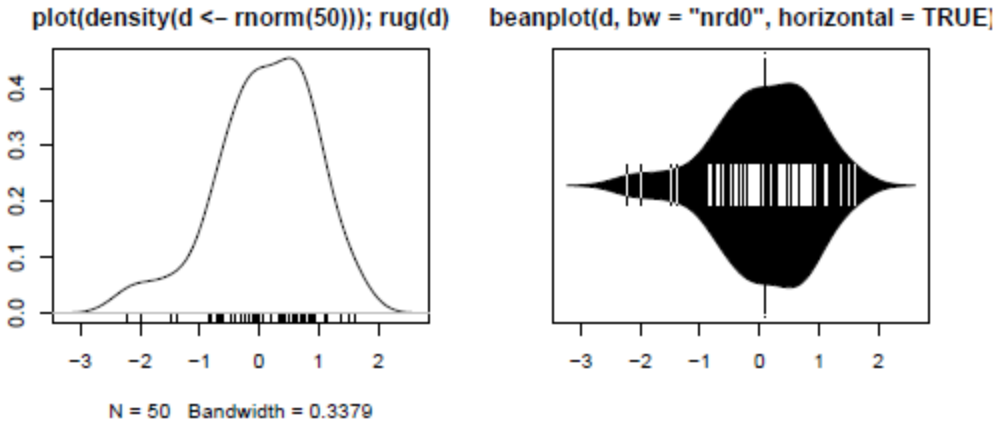


Figure 1: A density trace of a normal distribution with a `rug` (1d-scatter plot) and its corresponding beanplot. The small lines represent individual data points.

Bean Plots

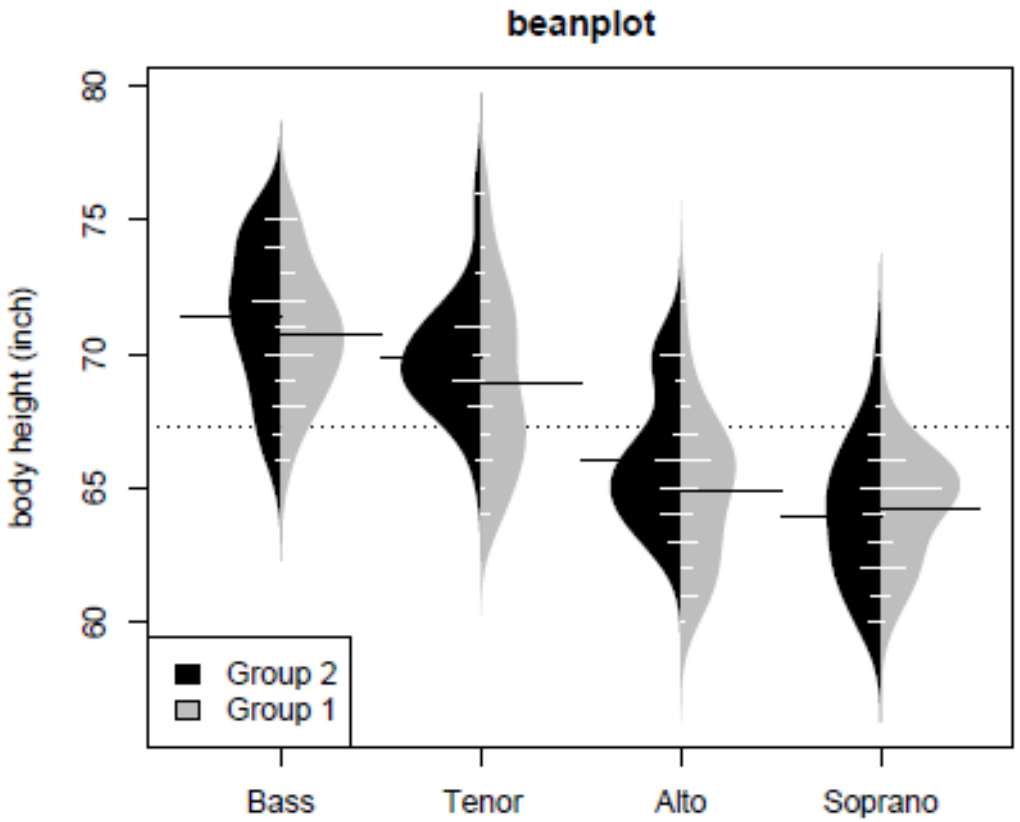
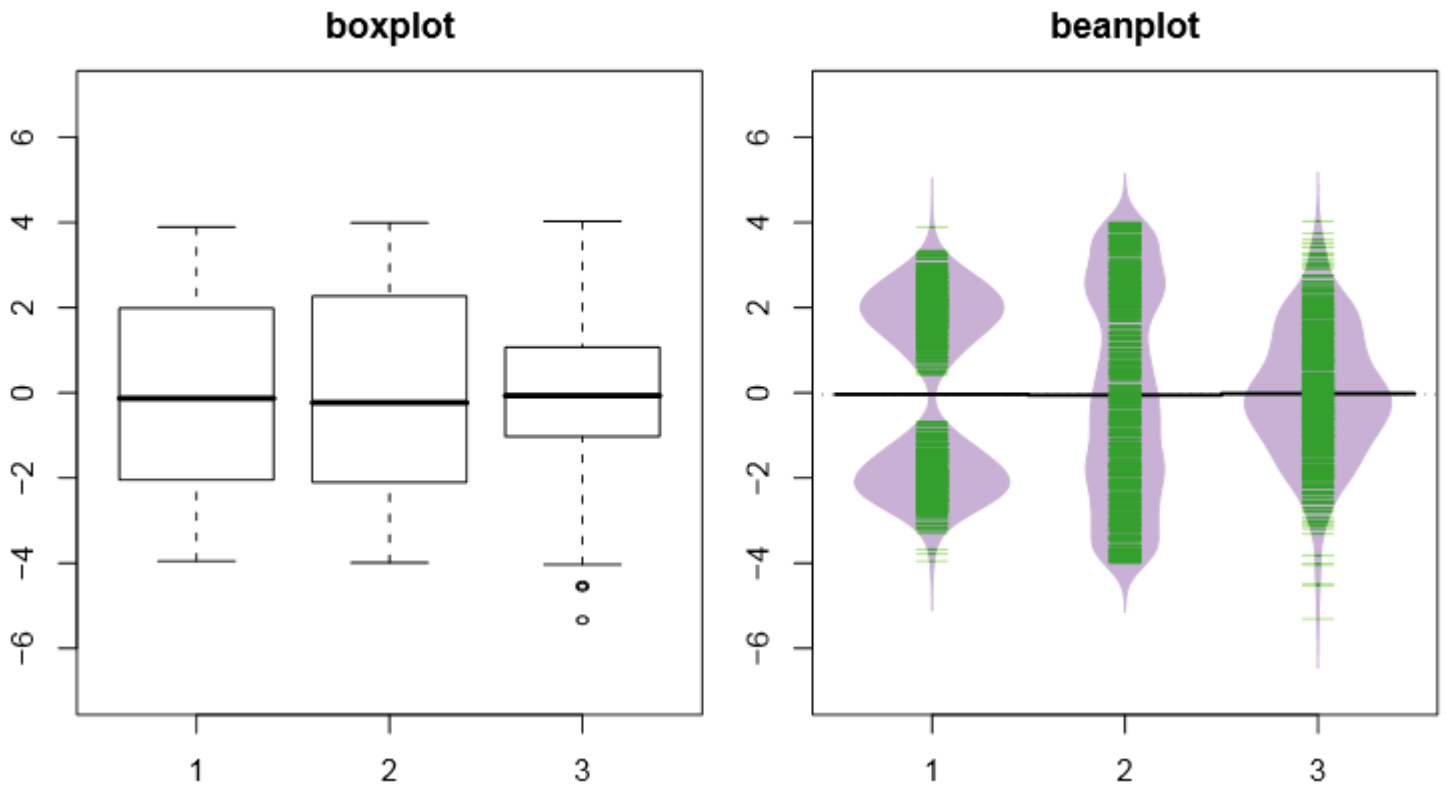


Figure shows (as produced by vioplot, Adler 2005) a beanplot for the body heights of different singers. In the beanplot it is visible that the measurements are in whole inches, and that there were many singers with a height of 65 inches in group Soprano 1. Also, an indication of the number of measurements is visible.

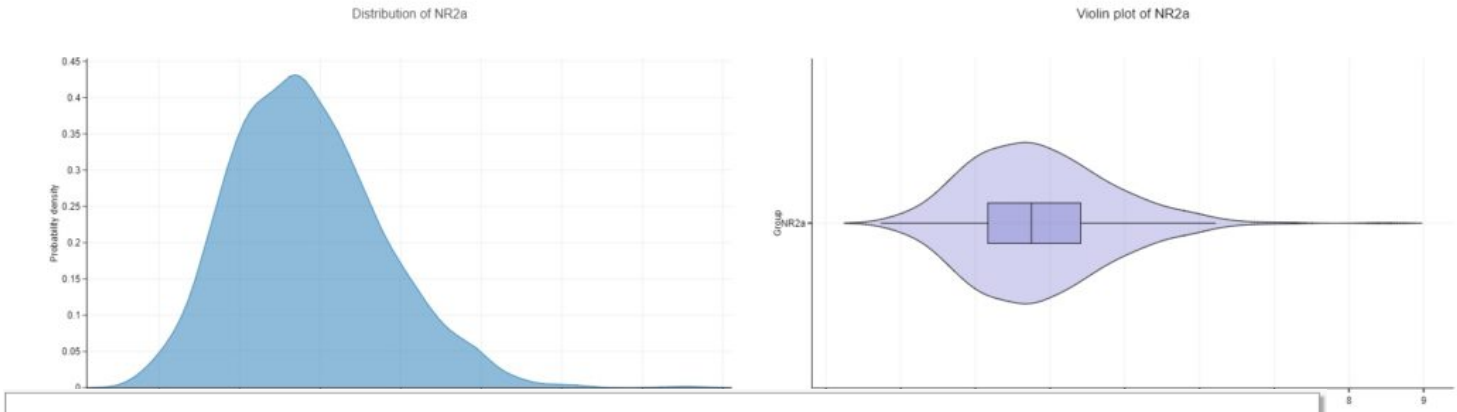
Box plots vs. Bean plots

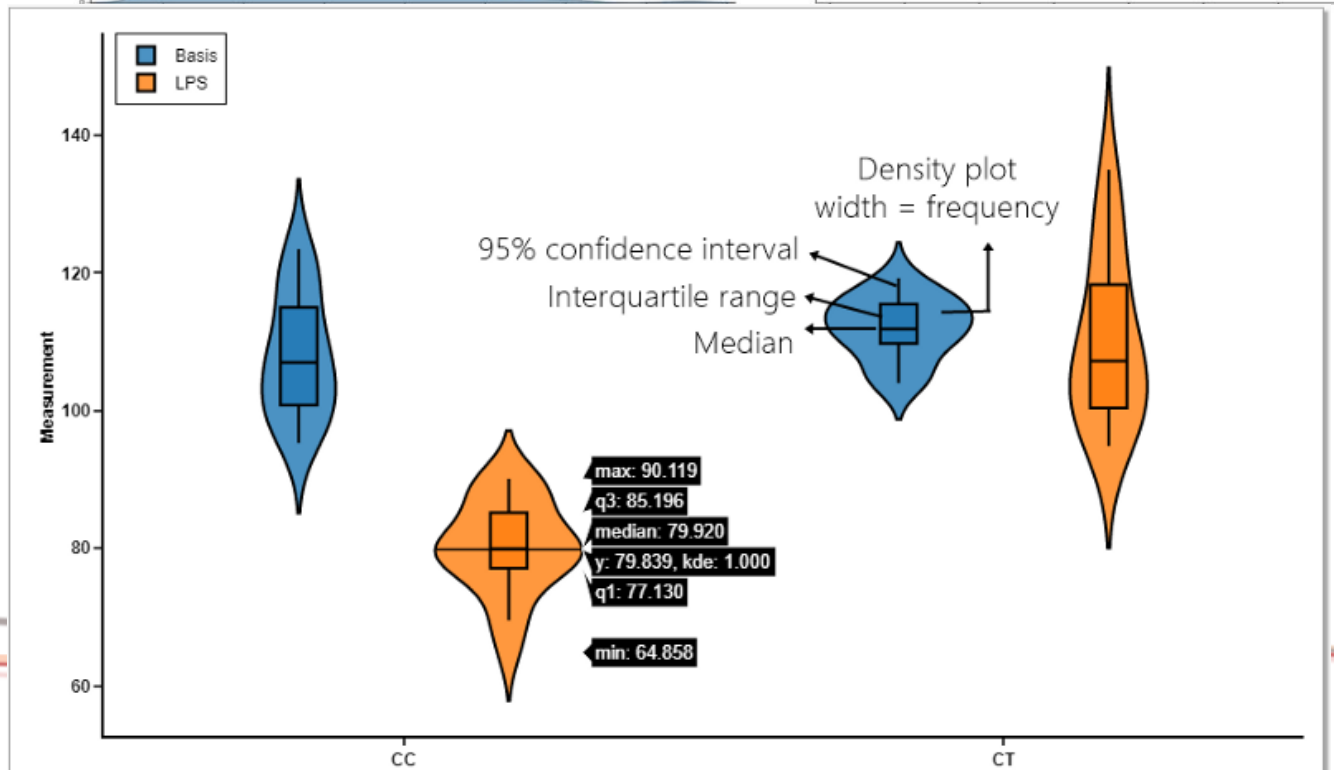


Plots for a bimodal, a uniform and a normal distribution. In the beanplot the green lines show individual observations, while the purple area shows the distribution

Violin Plots

The “violin” shape of a violin plot comes from the data’s density plot. You just turn that density plot sideway and put it on both sides of the box plot, mirroring each other.





Quantile Plot

- Quantile: each of any set of values of a variate that divide a frequency distribution into equal groups, each containing the same fraction of the total population.
- Sample quantiles are based on Order Statistics: Let $X_1, X_2, ..., X_n$ be a sample of size n. The order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are just the observations sorted in ascending order.
- Data: 76 92 83 105 102 109 106 91 110 89
- Order statistics: 76 83 89 91 92 102 105 106 109 110

Quantile Plot

- Specialized quantiles:
  - The 2-quantile is called the **median**
  - The 3-quantiles are called tertiles or terciles
  - The 4-quantiles are called **quartiles**
  - The 5-quantiles are called quintiles
  - ....

Quantile Plot

- Our data: 76 92 83 105 102 109 106 91 110 89 28 71
- Order statistics: 28 71 76 83 89 91 92 102 105 106 109 110

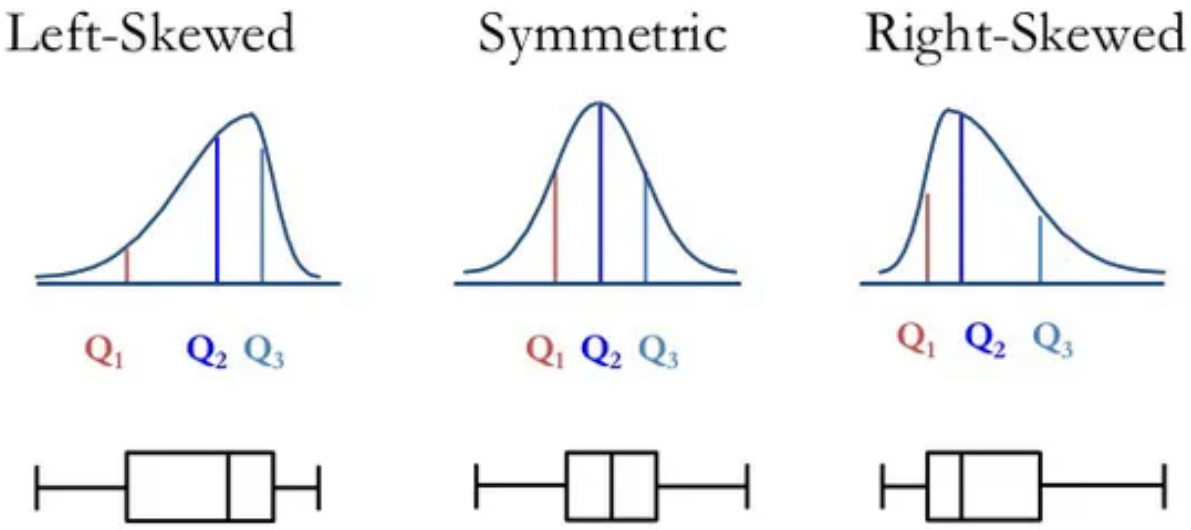
```
>>> data = pd.DataFrame({'x': [76,92,83,105,102,109,106,91,110,89,28,71]})
>>> data.quantile(q=[0,0.25,0.5,0.75,1], interpolation='lower')
        x
0.00   28
0.25   76
0.50   91
0.75  105
1.00  110
>>> data.x.quantile(q=[0.05, 0.95], interpolation='lower')
0.05     28
0.95    109
```

Quantiles are points in a distribution that relate to the rank order of values in that distribution.

Quantile Plot

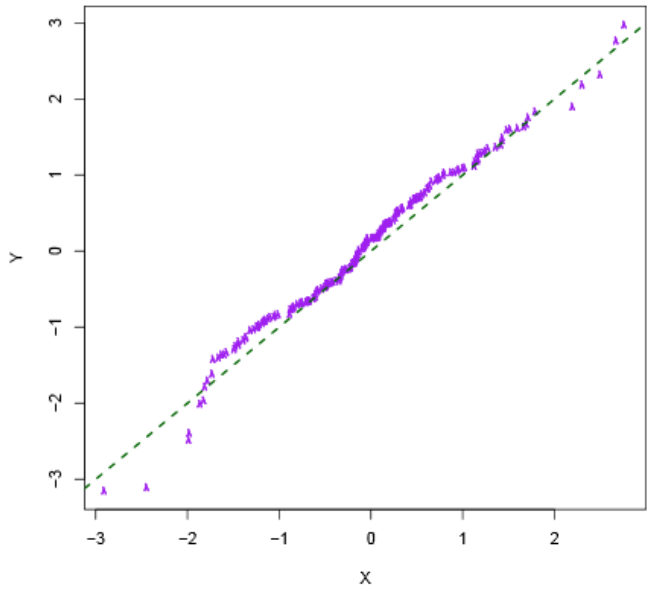
- Centiles/percentiles are descriptions of quantiles relative to 100;
  - The 75th percentile (upper quartile) is 75% or three quarters of the way up an ascending list of sorted values of a sample.
  - The 25th percentile (lower quartile) is one quarter of the way up this rank order.
- Percentile rank is the proportion of values in a distribution that a particular value is greater than or equal to.
  - For example, if a pupil is taller than or as tall as 79% of his classmates then the percentile rank of his height is 79, i.e. he is in the 79th percentile of heights in his class.

Quantile Plot



Quantile Plot

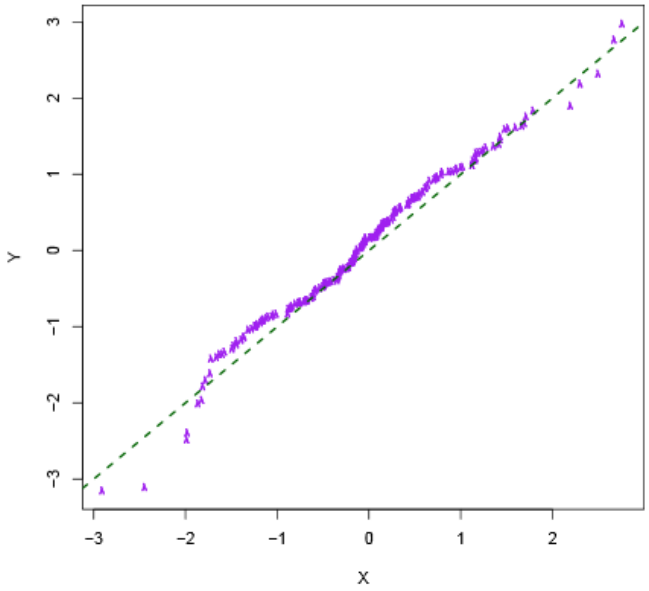
- Comparing two samples: Suppose we have two samples of size n, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$. If they were samples from the same distribution, then the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ and $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ would be estimates of the same quantiles.
- The **quantile-quantile plot, or QQplot**, is a simple graphical method for comparing two sets of sample quantiles.
- If two datasets come from the same distribution, the points should lie roughly on a line through the origin with slope 1.

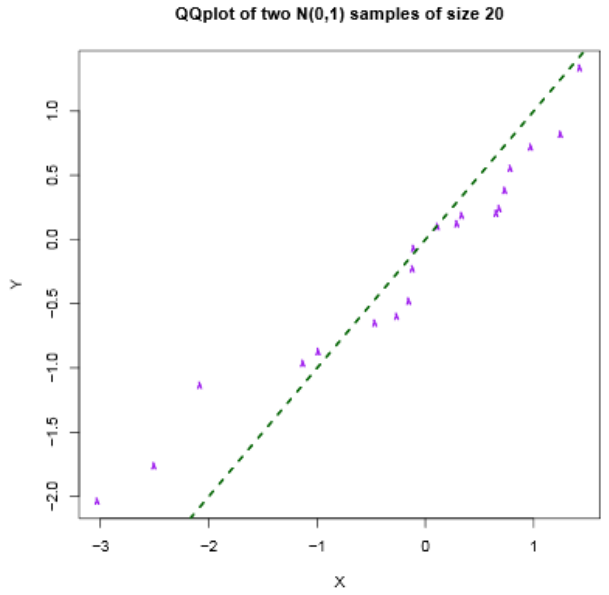


x coordinate here shows the theoretical quantiles.

Quantile Plot

- Comparing two samples: Suppose we have two samples of size n, X1, X2, ..., Xn and Y1, Y2, ..., Yn. If they were samples from the same distribution, then the order statistics X(1), X(2), ..., X(n) and Y(1),Y(2), ..., Y(n) would be estimates of the same quantiles.
- The quantile-quantile plot, or QQplot, is a simple graphical method for comparing two sets of sample quantiles.
- With small samples, variations may happen. nts should lie roughly on a line through the origin with slope 1.





Quantile Plot

- Recall our data: 76 92 83 105 102 109 106 91 110 89 28 71
- Order statistics: 28 71 76 83 89 91 92 102 105 106 109 110

```
>>> data = pd.DataFrame({'x': [76,92,83,105,102,109,106,91,110,89,28,71]})
>>> data.quantile(q=[0,0.25,0.5,0.75,1], interpolation='lower')
      x
0.00   28
0.25   76
0.50   91
0.75  105
1.00  110
>>> data.x.quantile(q=[0.05, 0.95], interpolation='lower')
0.05    28
0.95   109
```

Does my data come from normal distribution?

Quantile Plot

- Recall our data: 76 92 83 105 102 109 106 91 110 89 28 71
- Order statistics: 28 71 76 83 89 91 92 102 105 106 109 110

Mean = 88.5

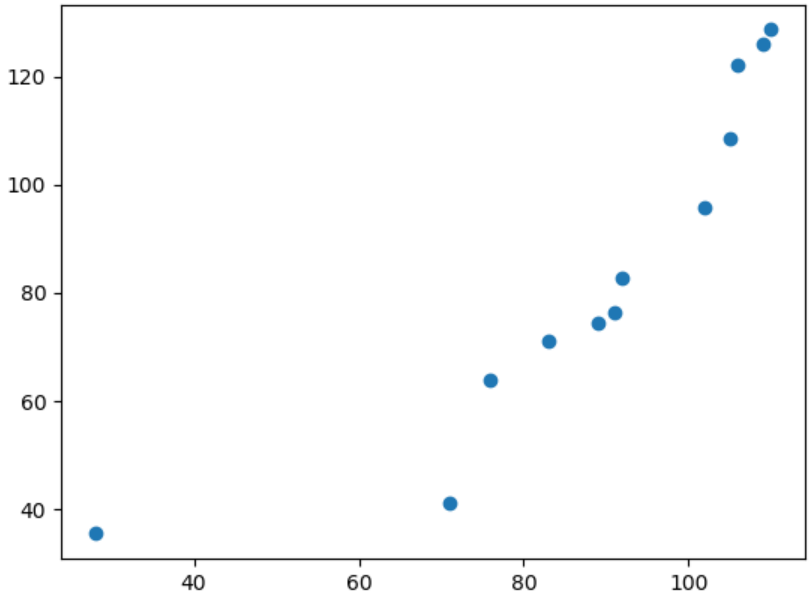
Standard deviation = 22.99209

Let's generate 12 random numbers hav

```
>>> y = numpy.random.normal(loc=data.x.mean(), scale=c
>>> y
array([ 125.84710048,  74.43443934,  76.35622569,  63.
        108.39753313,  35.58280708, 128.61687523,  70.9
         95.83537673,  82.76645301, 122.1169224 ,  41.13
```

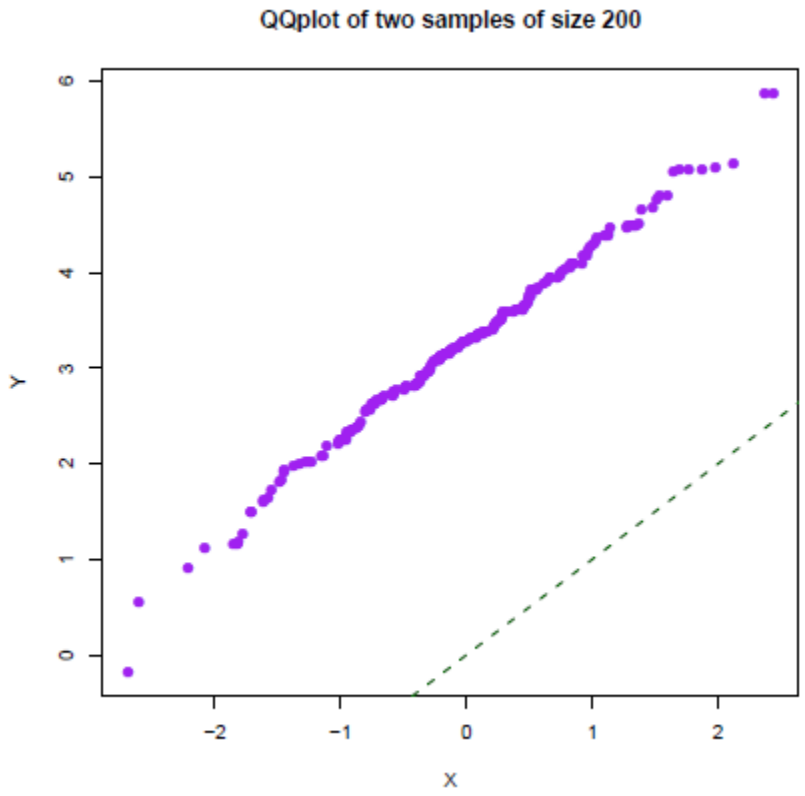
Plot quantile quantile plot (shortcut!)

```
>>> plt.scatter(x=sorted(data.x), y=sorted(y))
>>> plt.show()
```



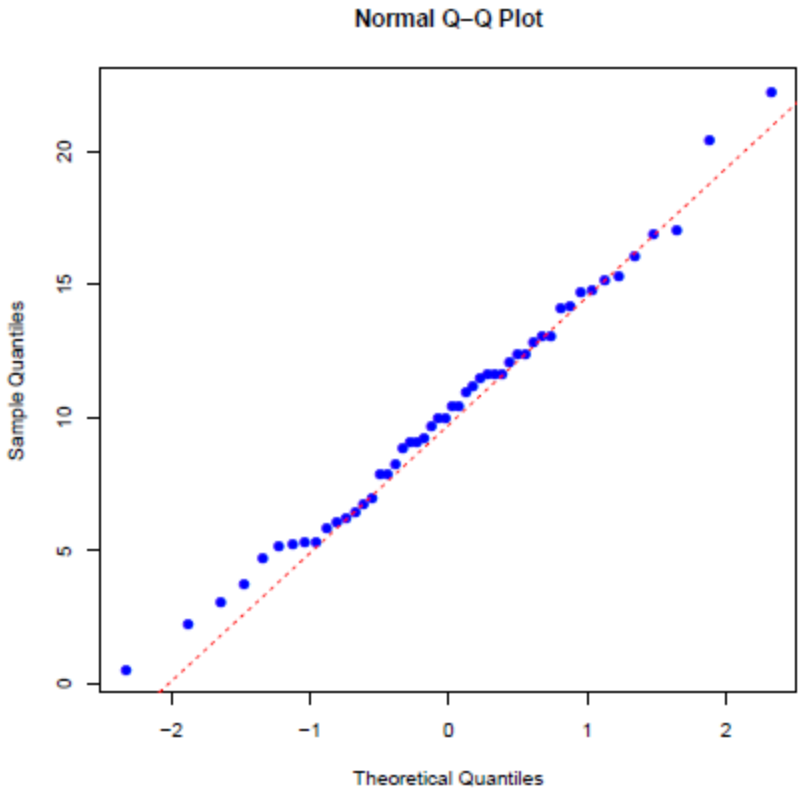
Quantile Plot

- Two samples from similar distributions which differ only in location: the green reference line is y = x



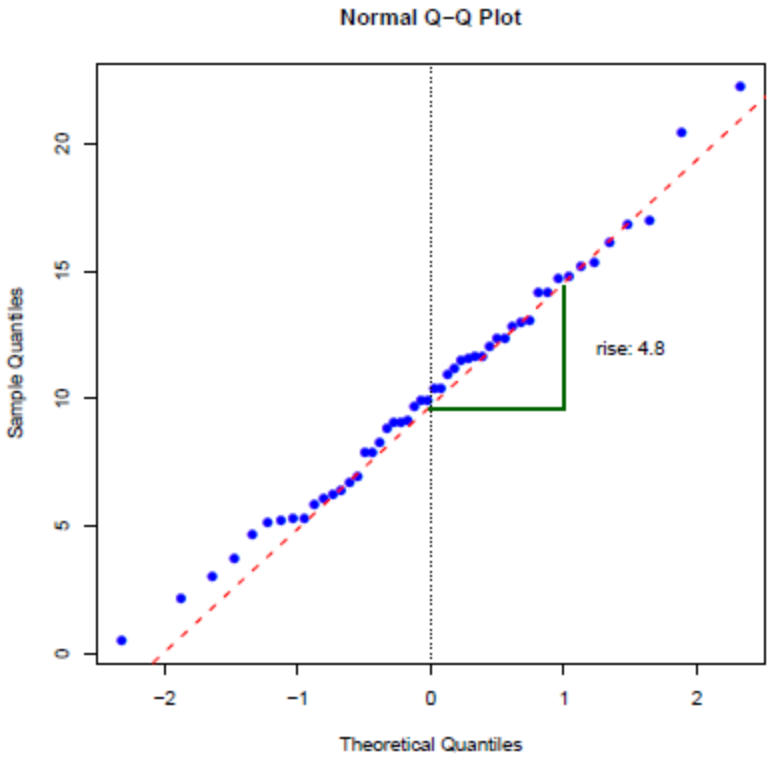
Quantile Plot

- Comparing to a normal distribution (Normal quantile plot).



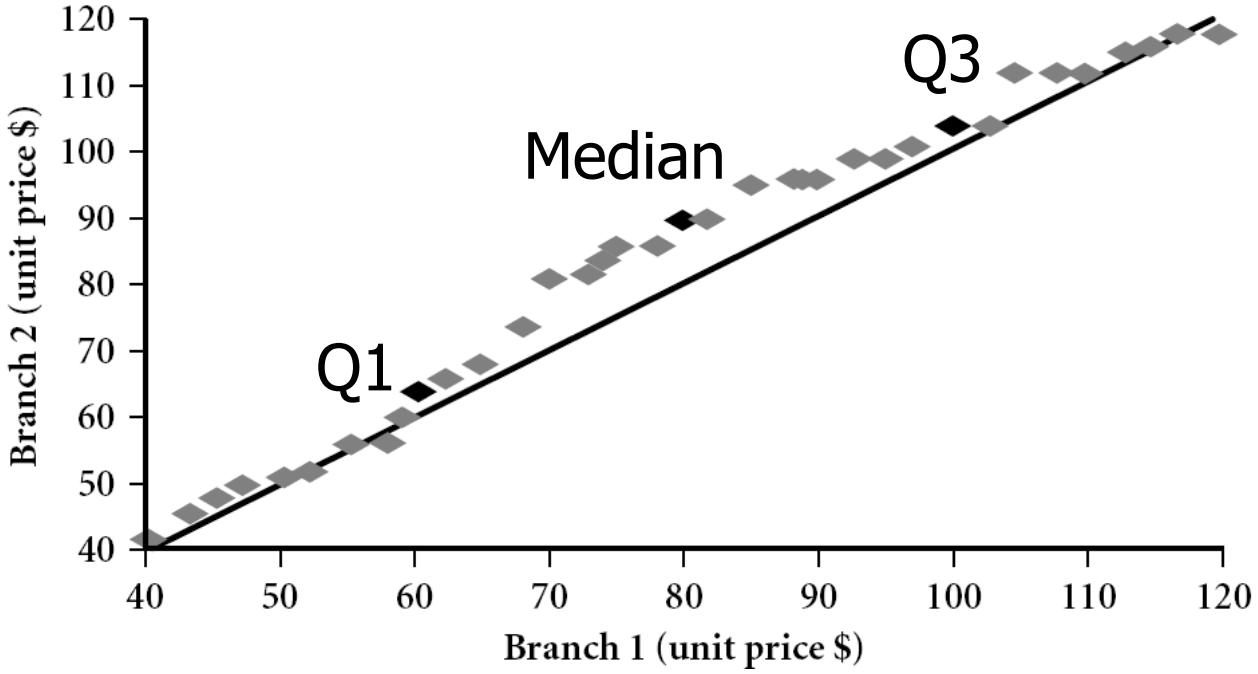
Quantile Plot

- We can estimate the mean and SD from a Normal quantile plot: the mean is roughly equal to the median (plotted above 0), and the slope is roughly the SD.



Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

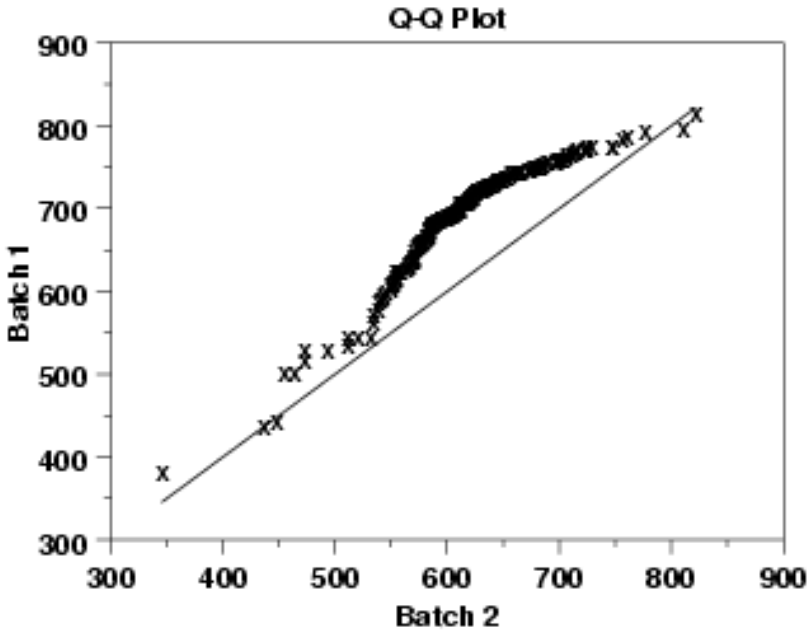


Quantile-Quantile (Q-Q) Plot

- The advantages of the q-q plot are:
  - The sample sizes do not need to be equal.
  - Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

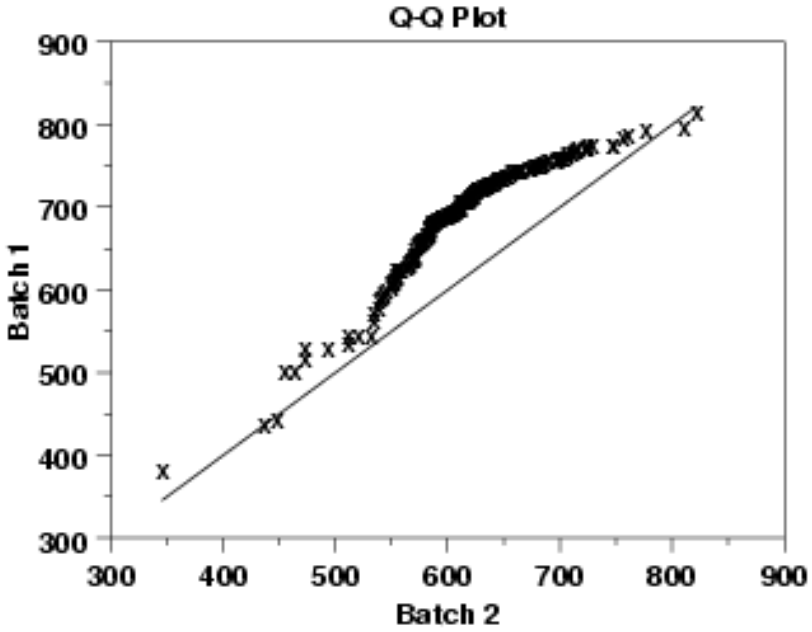
Quantile-Quantile (Q-Q) Plot

- Interpret the following plot.



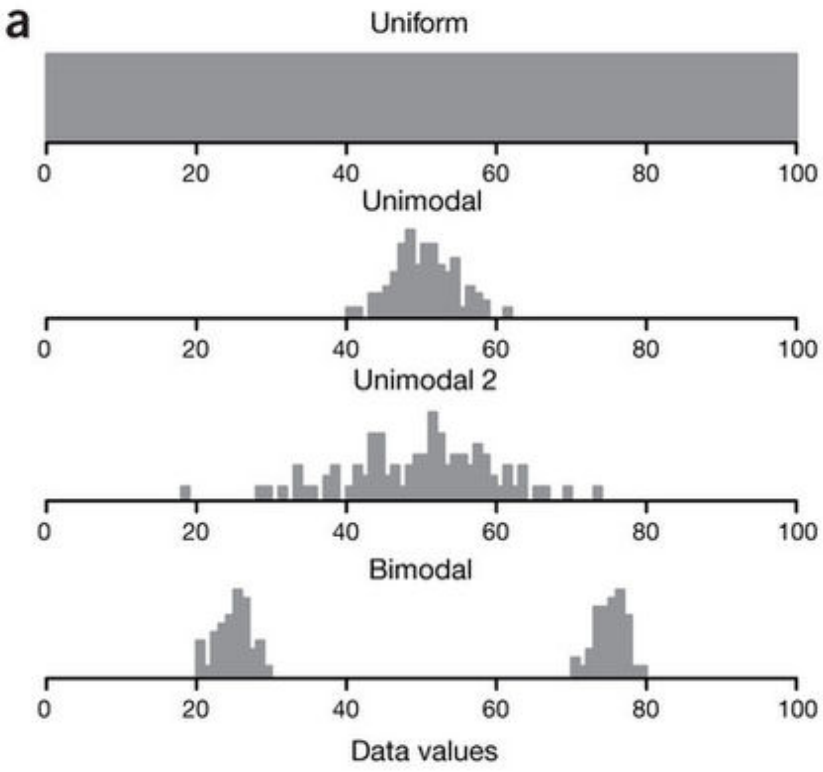
Quantile-Quantile (Q-Q) Plot

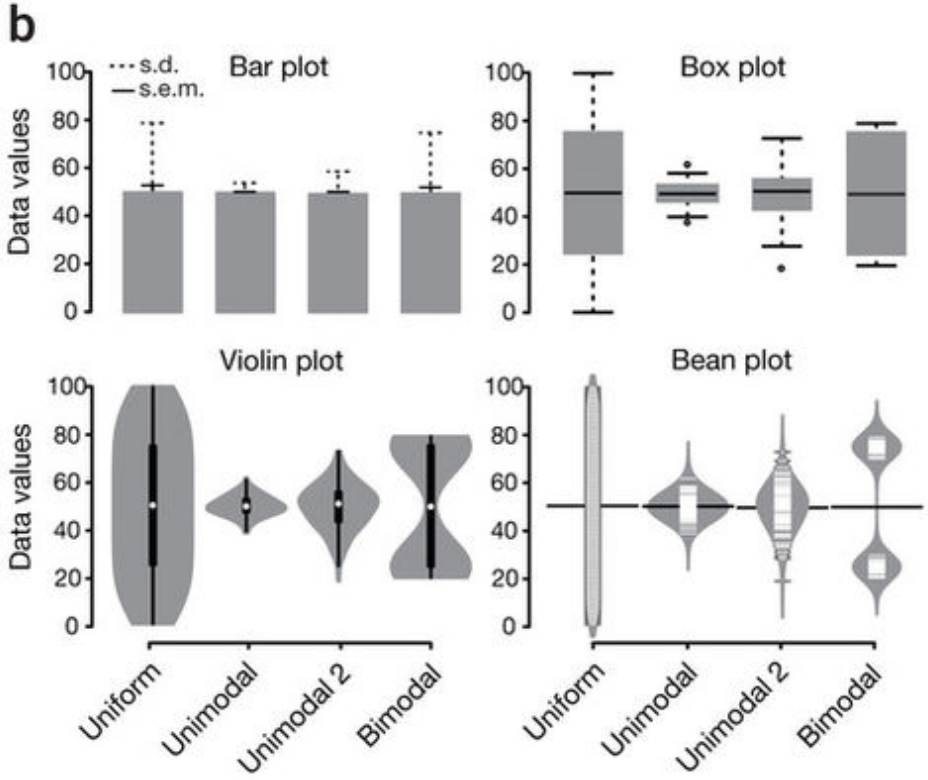
- Interpret the following plot.



This q-q plot shows that
1.These 2 batches do not appear to have come from populations with a common distribution.
2.The batch 1 values are significantly higher than the corresponding batch 2 values.
3.The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.

Summary Plots



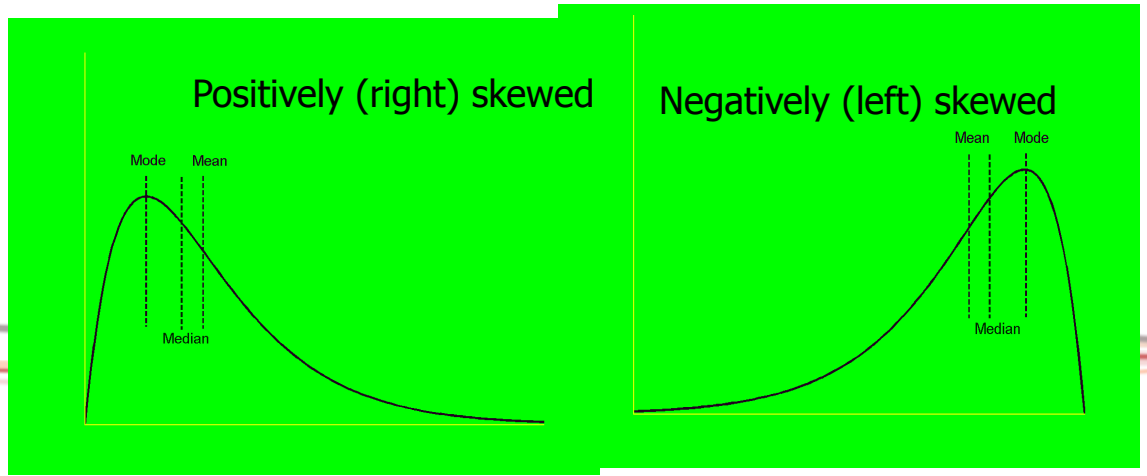


Example

- The table shows the household income of individuals based on their education level. What can you say about them based on the descriptive statistics?

Descriptive Statistics

	High school degree	Post-undergraduate degree	Did not complete high school	Some college	College degree
Mean	52.00	99.71	51.48	56.90	70.94
Std. Deviation	56.370	147.769	51.855	53.836	67.940
N	527	84	246	333	310
Median	35.00	59.50	36.00	39.00	49.00
Minimum	12	16	15	13	15
Maximum	533	1,079	497	403	512



Recommendation for Reporting Descriptive Statistics of Your Data

Adding median to the table will also be valuable because many datasets are non-normal.

Continuous Features	**Case 1**			**Case 2**		
Cont_Var_1	N (Valid)			N (Valid)		
	N (Missing)			N (Missing)		
	Mean			Mean		
	Std. Deviation			Std. Deviation		
Cont_Var_2	N (Valid)			N (Valid)		
	N (Missing)			N (Missing)		
	Mean			Mean		
	Std. Deviation			Std. Deviation		
Categorical Features	**Case 1**			**Case 2**		
Cate_Var_1	Values	Frequency	Percent	Values	Frequency	Percent
	A			A		
	B			B		
	C			C		
Cate_Var_2	Values	Frequency	Percent	Values	Frequency	Percent
	X			X		
	Y			Y		
	Z			Z		

Recommendation for Reporting Descriptive Statistics of Your Data

Continuous Features	Turkey		USA	
Body Surface Area (m^2)	N (Valid)	4523	N (Valid)	15256
	N (Missing)	354	N (Missing)	1478
	Mean	1.86	Mean	1.94
	Std. Deviation	0.24	Std. Deviation	0.3
Height (cm)	N (Valid)	4595	N (Valid)	16204
	N (Missing)	282	N (Missing)	530
	Mean	168.4	Mean	177.3
	Std. Deviation	12.5	Std. Deviation	16.1

Categorical Features	Turkey			USA		
Age	Values	Frequency	Percent	Values	Frequency	Percent
	10-19	357	7.32%	10-19	586	3.51%
	20-50	3845	78.84%	20-50	12965	77.47%
	50+	675	13.84%	50+	3183	19.02%
Gender	Values	Frequency	Percent	Values	Frequency	Percent
	Male	2702	55.4%	Male	9738	58.19%
	Female	2175	44.6%	Female	6996	41.81%

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
- Above steps are the beginning of data preprocessing
- Many methods have been developed but still an active area of research