MSc - Data Mining

Topic 01 : Module Overview

Part 04 : A Review of Statistical Concepts

Dr Bernard Butler and Dr Kieran Murphy

Department of Computing and Mathematics, WIT.
(bbutler@tssg.org and kmurphy@wit.ie)

Spring Semester, 2021

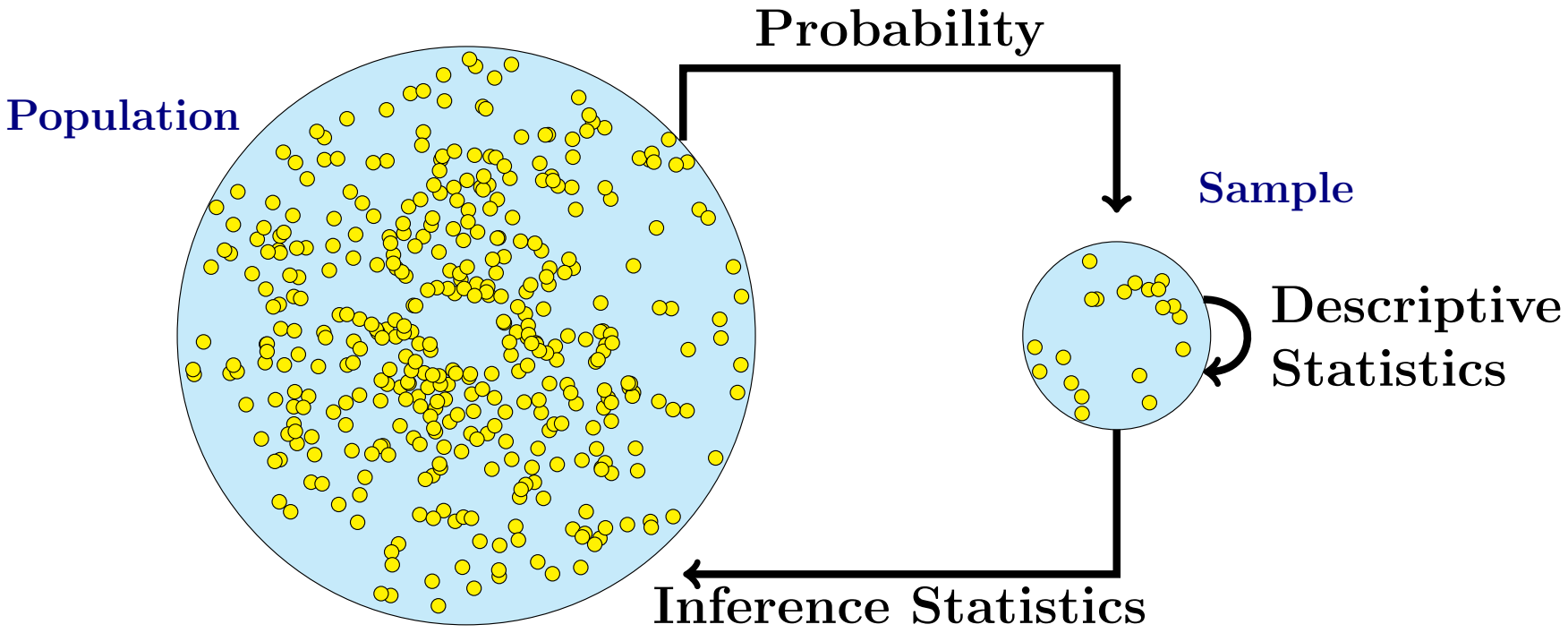
Outline

- Probability in five minutes (probably)
- Type of Data
- Descriptive statistics

Part I

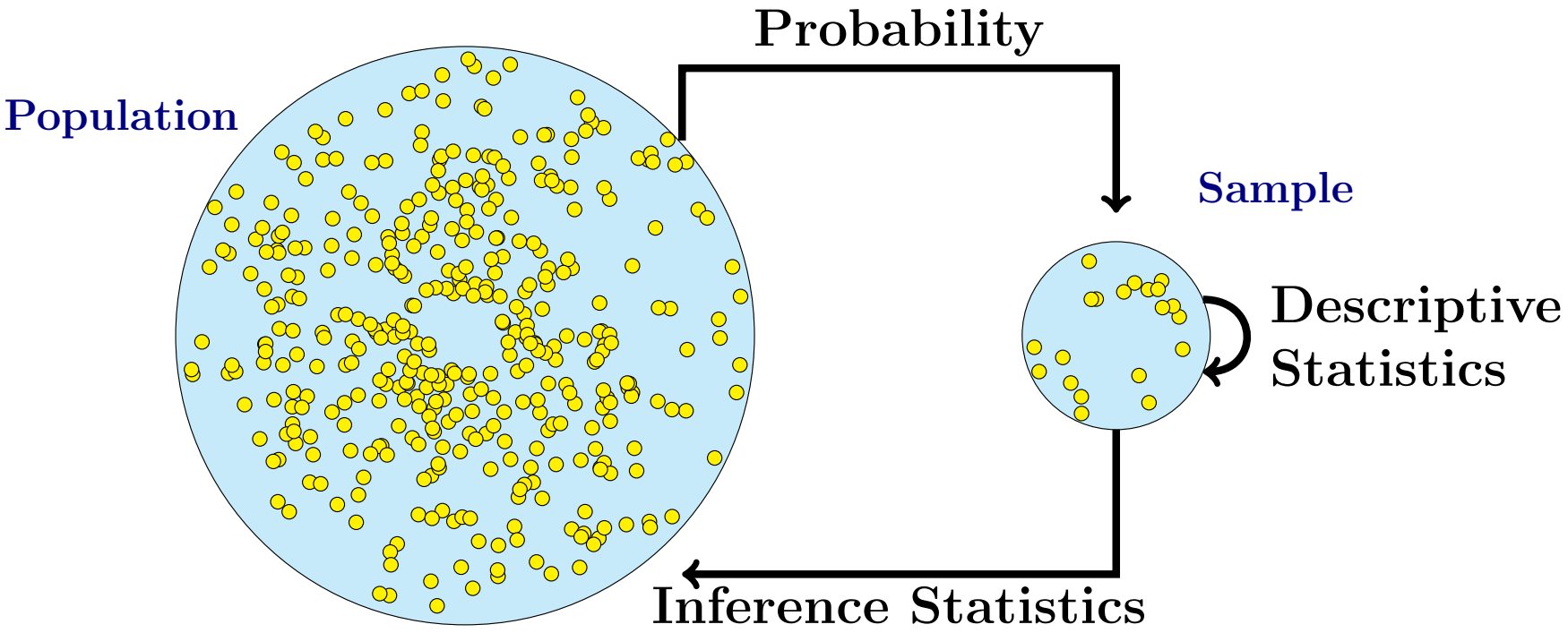
Fundamental Concepts

"Central Dogma" of Statistics



- The population is the group of all items of interest.
  - Frequently very large or even sometimes infinite.
  - Population attributes are called population parameters.
- A sample is a subset of the population drawn from the population.
  - Potentially very large, but usually of orders smaller than the population.
  - Sample attributes are called sample statistics.

"Central Dogma" of Statistics II



- Descriptive Statistics is a set of methods of organising, summarising, and presenting data in a convenient and informative way.
- Inferential statistics is a set of methods, to draw conclusions or inferences about characteristics of populations based on data from a sample.

So we use the sample statistics to make inferences about the population parameters

Descriptive vs Inference Statistics

We use sample statistics to make inferences about population parameters.

- Therefore, we can make an estimate, prediction, or decision about a (unknown) population based on (known) sample data.

Rationale

- Large populations make investigating each member impractical and expensive.
- Easier and cheaper to take a sample and make estimates about the population from the sample.

Warning

Such conclusions and estimates are not always going to be correct. For this reason, we build into the statistical inference "measures of reliability" namely confidence level and significance level.

— **even when all of the assumptions made in constructing the sample are true!.**

How good is your data? I

Validity

A valid measurement is one which actually measures what it claims to measure.

- Unemployment figures are validly measured using the Labour Force Survey not the Live Register

Reliability

A reliable measurement is one which will give approximately the same result time after time, when taken on the same individual or object.

- Most physical measurements are reliable, for example measuring your weight using a bathroom scales.
- Some measurements may be reliable but not necessarily valid.
  - Are exams reliable measuring devices?
  - Are exams results valid measurements of intelligence?

How good is your data? II

Bias

Sometimes when measurements are made a systematic error is made which underestimates or overestimates the true value. Such a measurement is called a biased measurement.

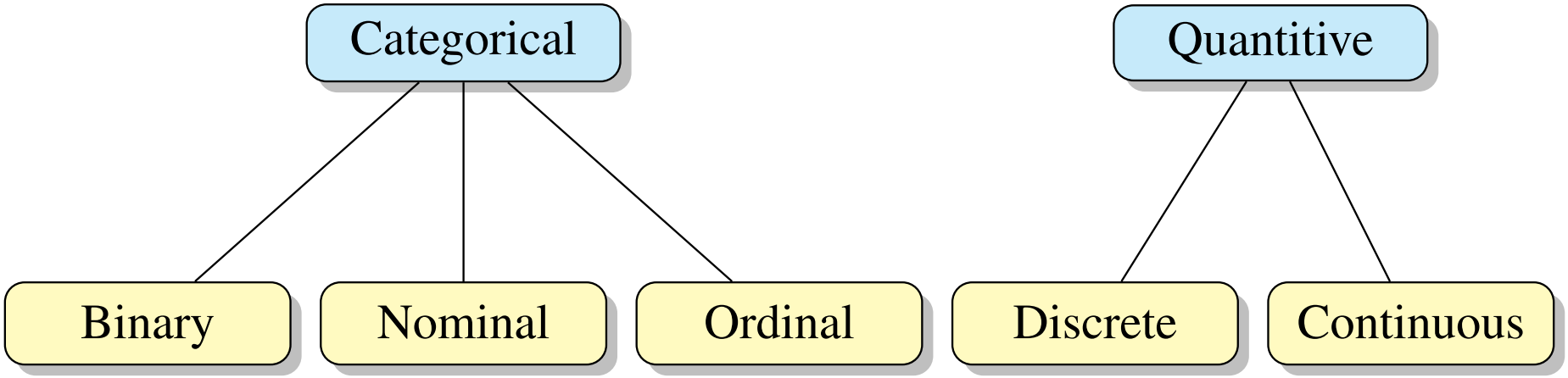
- Suppose your bathroom scales always overestimated your weight.
- Car speedometers are deliberately biased to overestimate a car's real speed.

Variability

Some datasets are more variable than others.

- A dataset consisting of the ages of 100 students in WIT will be less variable than a dataset consisting of the ages 100 randomly chosen Irish people.

Types of Data



2 Categories	More Categories	Order matters	Numerical	+ Uninterrupted
TRUE/FALSE YES/NO, etc	Gender, Religion, Favourite Team, etc		**CAN** be represented by integers on the number line	Takes values from intervals on the number line
			Number of students in this class	Duration of this class

Cannot order

Cannot add / subtract / multiple or divide

Dimensionality of Data Sets

We will frequently use a table to represent our data:

- Each row represents a observation / subject / case.
  - Univariate — single measurement per subject.
  - Bivariate — two measurements per subject.
  - Multivariate — Multiple measurements per subject.
- Each column represents a variable / attribute.
  - Each variable is then Binomial/Nominal/Ordinal/Discrete/Continuous

Tips dataset

```
df.head(10)
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
5	25.29	4.71	Male	No	Sun	Dinner	4
6	8.77	2.00	Male	No	Sun	Dinner	2
7	26.88	3.12	Male	No	Sun	Dinner	4
8	15.04	1.96	Male	No	Sun	Dinner	2
9	14.78	3.23	Male	No	Sun	Dinner	2

```
df.shape
```

```
(244, 7)
```

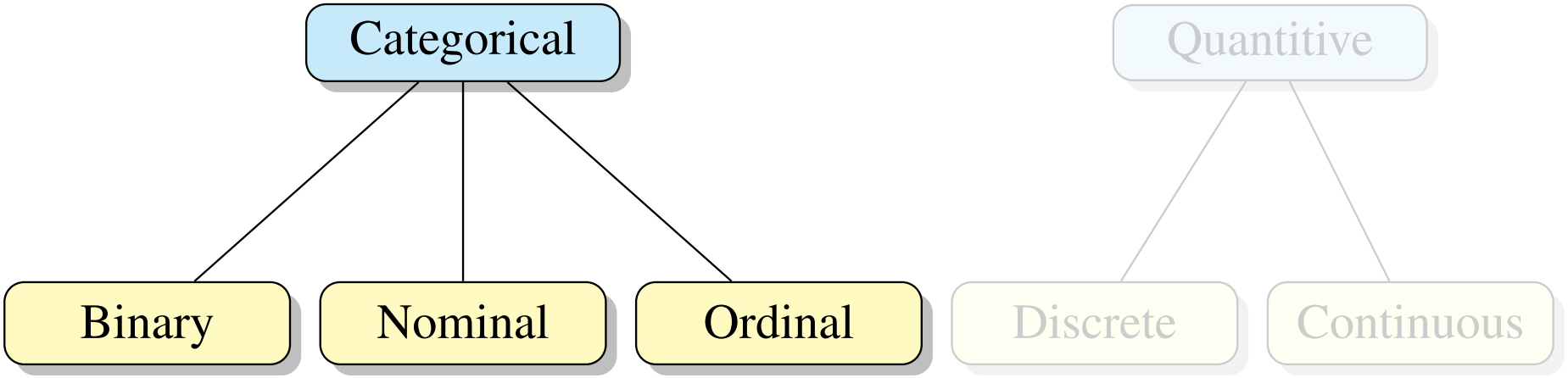
```
df.dtypes
```

```
total_bill   float64
tip          float64
sex         category
smoker      category
day         category
time        category
size           int64
dtype: object
```

Part II

Analysing Categorical Data

Types of Data — Categorical Data



2 Categories	More Categories	Order matters
TRUE/FALSE YES/NO, etc	Gender, Religion, Favourite Team, etc	

Graphical Methods

- Bar chart
- Pie chart

Numerical Methods

- Frequency counts
- Number of unique values

Cannot order

Cannot add / subtract / multiple or divide

$\Longrightarrow$ Limited to methods that only deal with counts/frequencies

Example: `tips.day`

```
df.day.nunique()
```

```
4
```

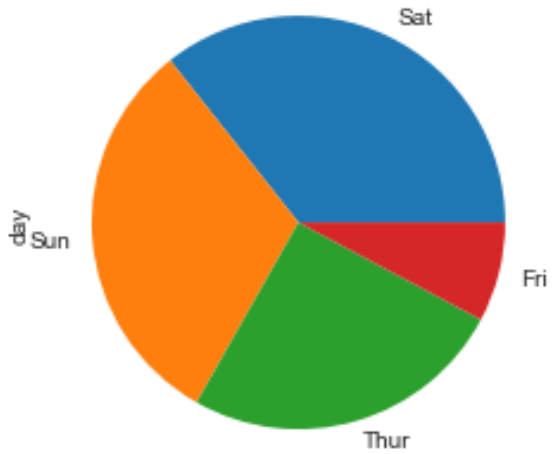
```
df.day.value_counts()
```

```
Sat     87
Sun     76
Thur    62
Fri     19
Name: day, dtype: int64
```

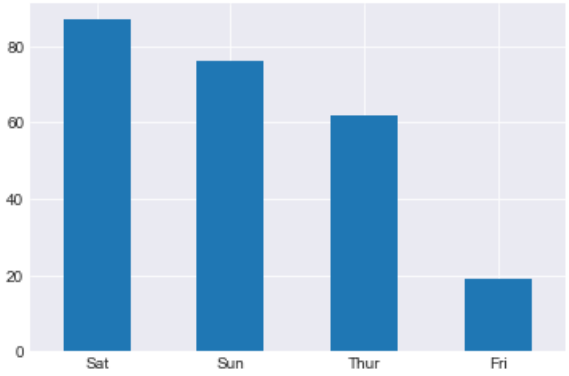
```
df.day.describe()
```

```
count    244
unique     4
top      Sat
freq      87
Name: day, dtype: object
```

```
df.day.value_counts().plot(kind='pie');
```




```
df.day.value_counts().plot(kind='bar')
plt.xticks(rotation=0);
```



These are horrible visualisations — pie chart does not show counts, and what about the category order in the bar charts

Comments

- Pie charts are often over used (especially by newspapers) but they have significant limitations:
  - Comparing two pie charts is problematic since people find it harder to compare angles to lengths.
  - If one uses relative frequencies (or percentages) then the number of observations needs to be clearly stated.
  - Should not be used if the number of categories is large.

To my mind, the best use of a pie chart is when you have one value that is overwhelmingly larger than the rest and you don't want the audience to focus on the actual values, but just bamboozle them with the overwhelming size of the leading segment. Of course, this seems to come close to embracing the old adage, "There are lies, damn lies and statistics."

— *www.juiceanalytics.com/writing/writing/the-problem-with-pie-charts*

See also

- www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=00018S
  (Edward Tufte is a statistician and author of 4 books on data visualisation.
- "Save the Pies for Dessert" article at
  www.perceptualedge.com/articles/08-21-07.pdf

Part III

Analysing Quantitative Data

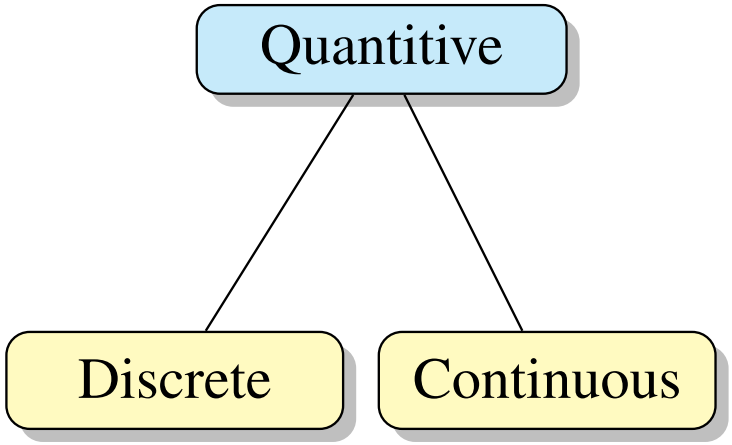
Types of Data — Quantitative Data

Graphical Methods

- Histogram — for shape, spread
- Boxplot — for shape, location, outliers
- ...

Numerical Methods

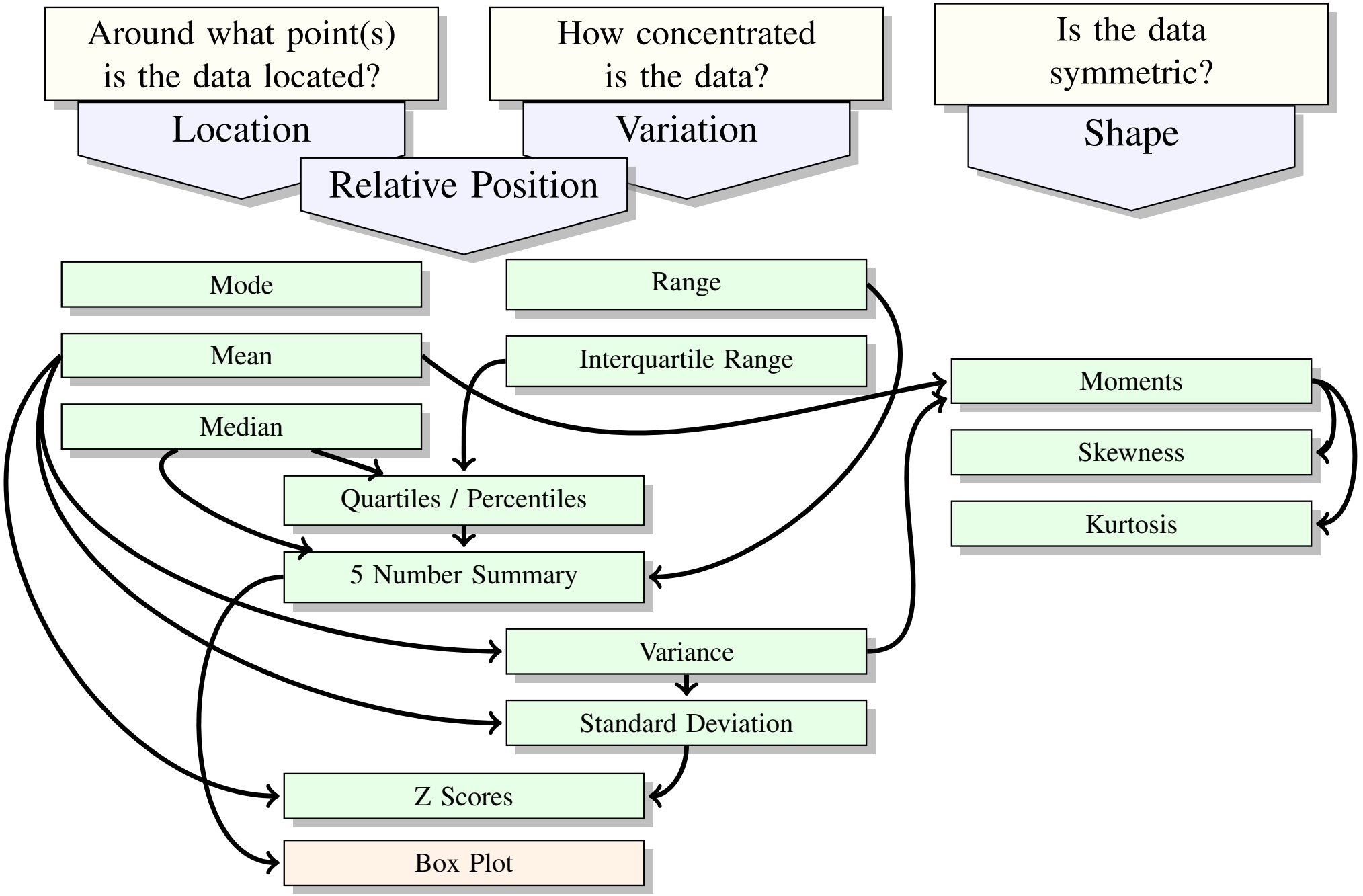
- mean, median — centre
- standard deviation, range — spread
- z-scores, quantiles — location
- ...



Numerical	Numerical+Uninterrupted
CAN be represented by integers on the number line	Takes values from intervals on the number line
Number of students in this class	Duration of this class

Can add / subtract / multiple and divide (even for discrete data)

Numerical Methods for Quantitive Data



Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.

- Often called measures of central location.
- The mode, mean, and median are common* measures of central tendency, but under different condition, some measures of central tendency become more appropriate to use than others.

*Wikipedia lists 14 measures, `https://en.wikipedia.org/wiki/Central_tendency`

Mode vs Mean vs Median

Metrics

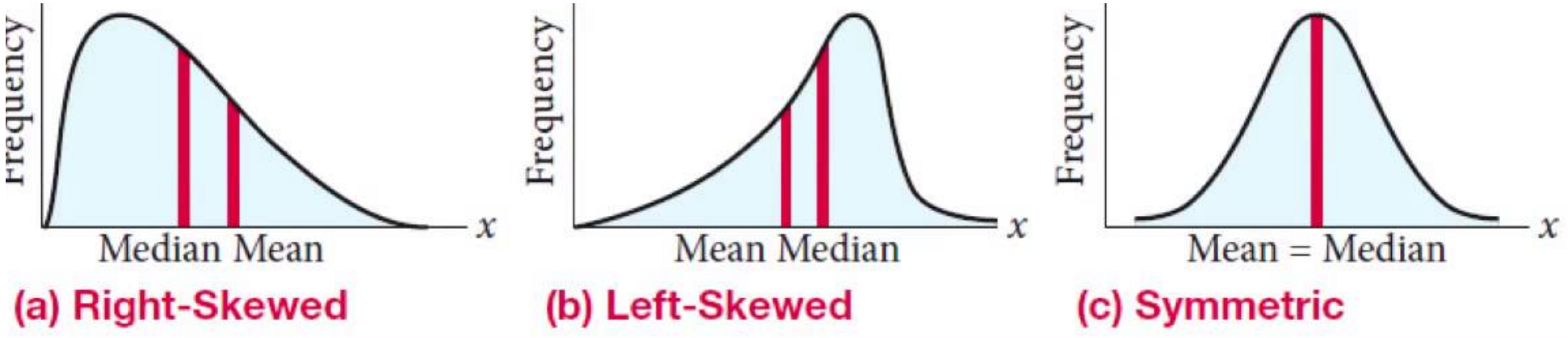
- The (arithmetic) mean is the sum of its data values divided by data count.
- The median is the point in the variable that splits the values in two equal groups.
- The mode is the most frequently occurring value in a variable.

Comparision

- If variable is nominal then mode is your only option.
- If variable is ordinal then you can use either mode or median.
  - The median only makes use of the order information in the variable (i.e., which numbers are bigger), but doesn't depend on the precise numbers involved. Hence median can be applicable while the mean is not.
- For quantitive data, either the mean or the median is generally acceptable.
  - Which one to pick depends on the data and what you're trying to achieve. The mean has the advantage that it uses all the information in the data (which is useful when you don't have a lot of data), but it's very sensitive to extreme values.

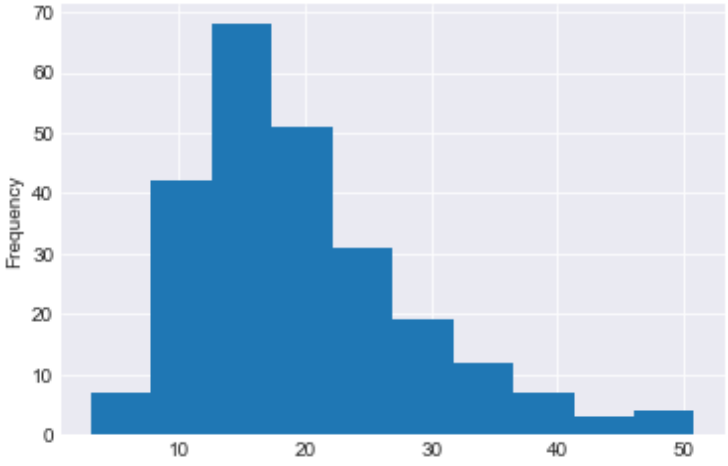
Mean vs Median and Shape

$$\text{If the distribution is } \left\{ \begin{array}{c} \text{negatively skewed} \\ \text{symmetric} \\ \text{positively skewed} \end{array} \right\} \text{ then } \left\{ \begin{array}{l} \text{mean} < \text{median} \\ \text{mean} = \text{median} \\ \text{mean} > \text{median} \end{array} \right.$$

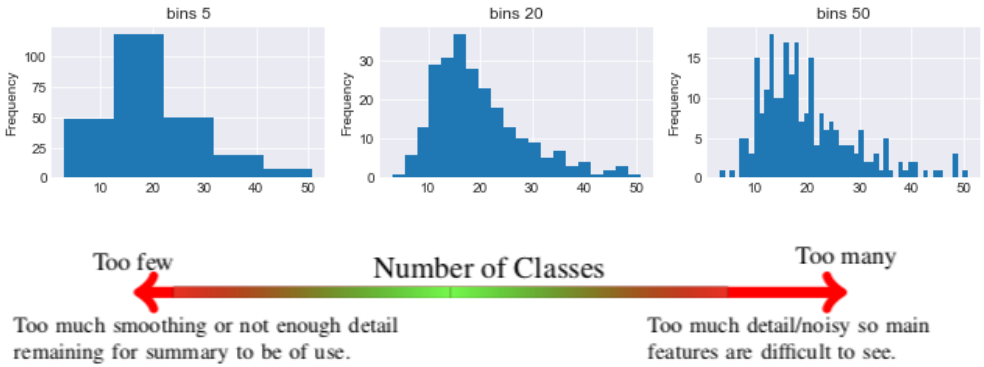


Histogram and Box-Plot

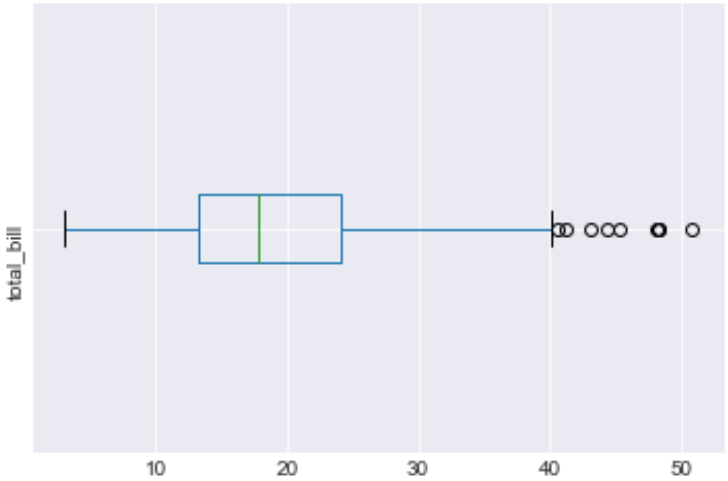
```
df.total_bill.plot(kind='hist');
```



- Parameter **bin** controls level of granularity.



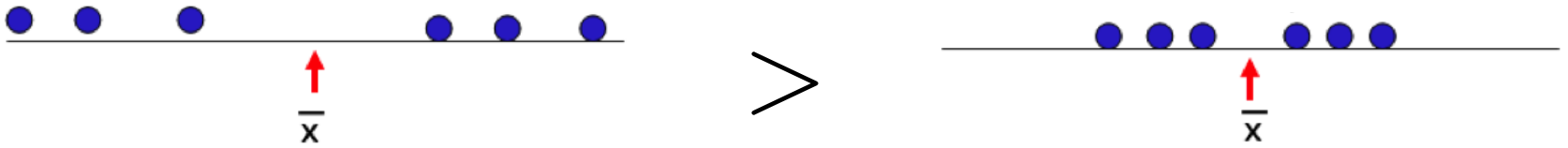
- Distribution is uni-modal
- and right skewed (longer tail to the right)
- Boxplot clearly show outliers (dots), skewness
  - Centre line is the median
  - Box edges are the first and third quartiles.




```
df.total_bill.plot(kind='box', vert=False)
plt.yticks(rotation=90);
```

Measures of Variation

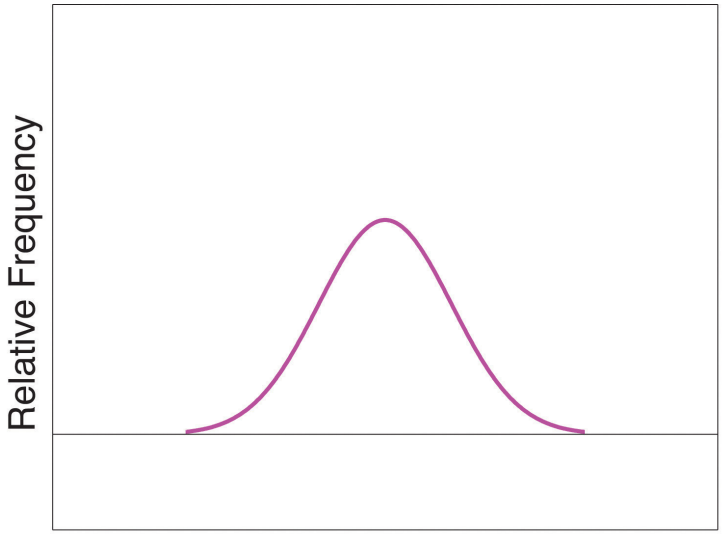
Measures of variation measures the amount that values vary or differ among themselves.



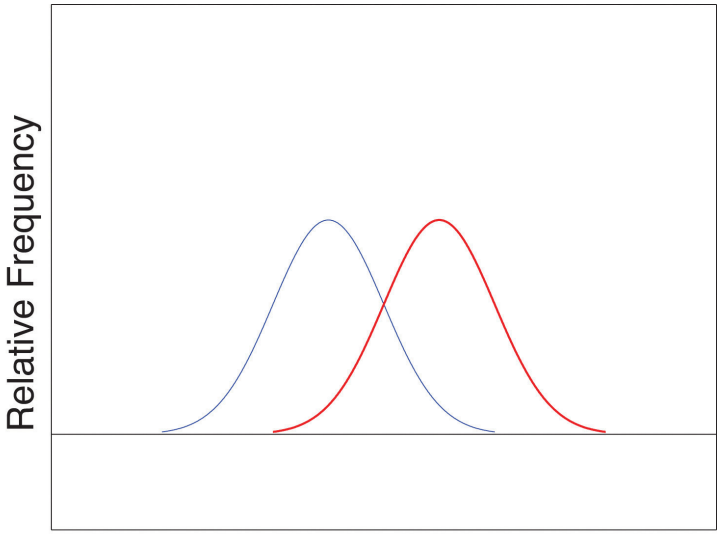
- A measure of statistical variation/dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data become more diverse.
- Dispersion is contrasted with location or central tendency, and together they are the most used properties when summarising variables.
- There are different ways of measuring variation:† range, variance and standard deviation

†Wikipedia lists over 12 measures,
https://en.wikipedia.org/wiki/Statistical_dispersion

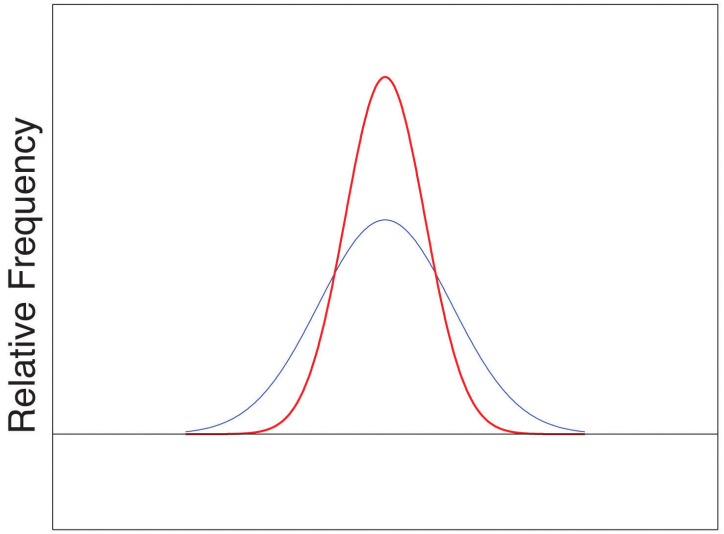
Central Tendency (Location) vs Variation



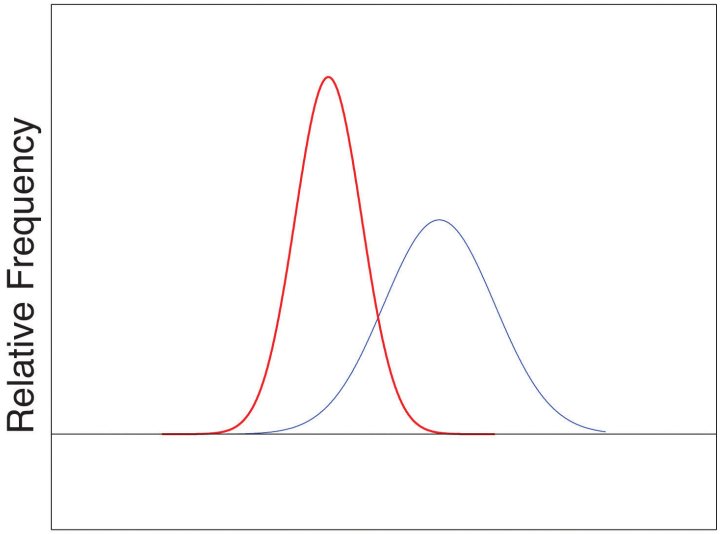
(a) Two Identical Sets



(b) Locations Differ



(c) Variabilities Differ



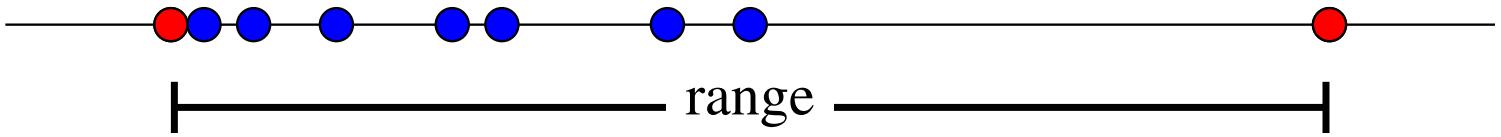
(d) Locations and Variabilities Differ

Measures of Variation Metrics

Range

The range of a data set is the difference between the maximum and minimum values in the set.

- ✔ Simplest to calculate and has easy interpretation, and only two values are used in its calculation.
- ✘ The range is highly sensitive to outliers (in fact it depends exclusively on the extreme values).



Variance

The variance, σ^2, where σ is the standard deviation and is defined as follows

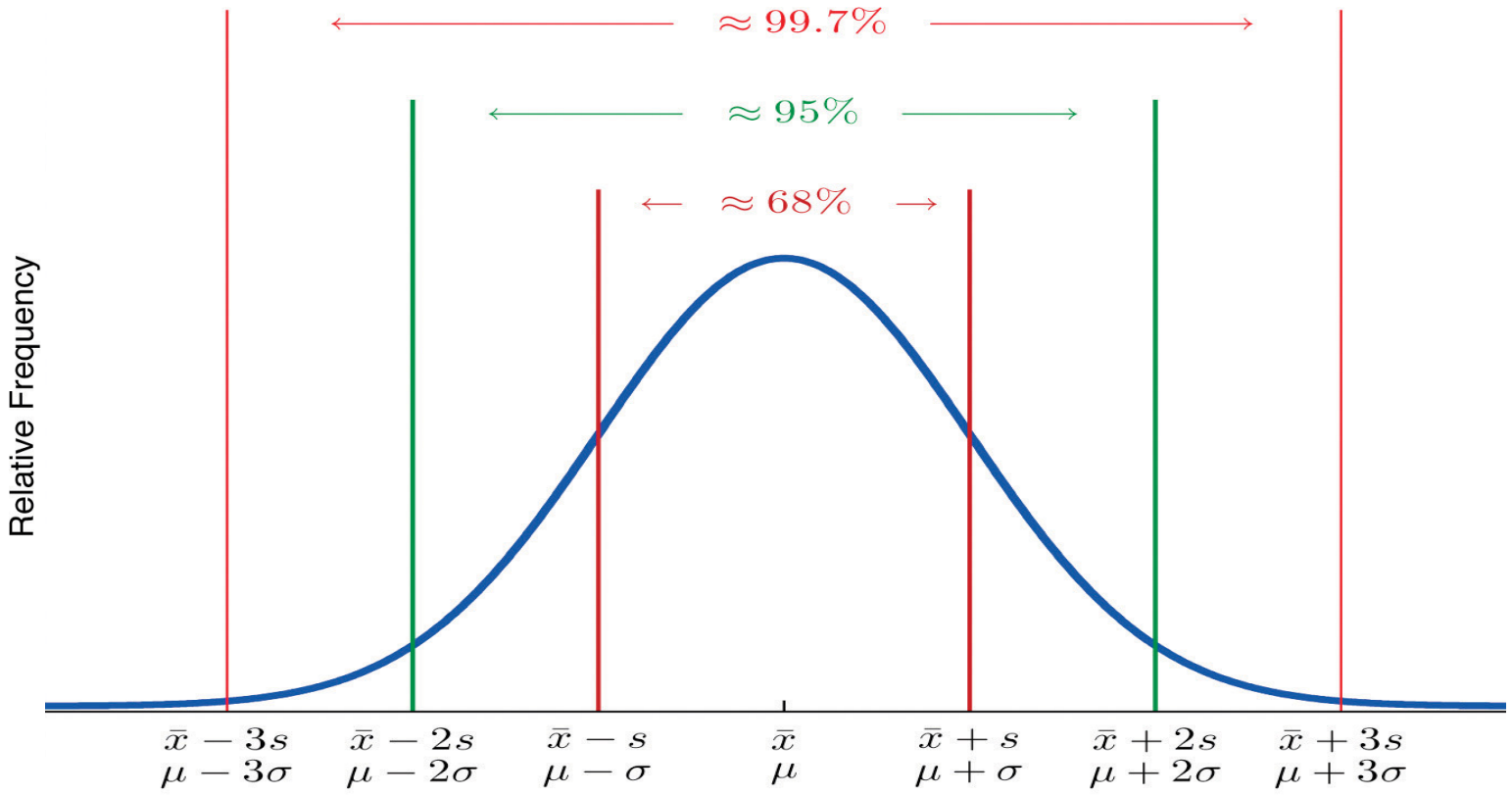
$$\sigma^2 = \sum_{k=1}^{n} \frac{(\bar{x} - x_k)^2}{n} \tag{1}$$

Interpreting and Understanding Standard Deviation

Property: Empirical rule for data with a Bell-shaped distribution

If the variable has a bell-shaped distribution, then

- About 68% of all values fall within 1 standard deviation of the mean.
- About 95% of all values fall within 2 standard deviation of the mean.
- About 99.7% of all values fall within 3 standard deviation of the mean.

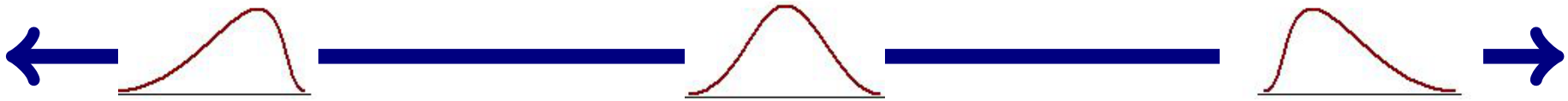


Measures of Shape

Finally we look symmetric, and how "peaked" the set of data is:

- We start with a reference distribution — the standard normal curve.
- Skewness is a measure of how symmetric the distribution is.
- Kurtosis is a measure of how peaked the distribution is.

Skewness

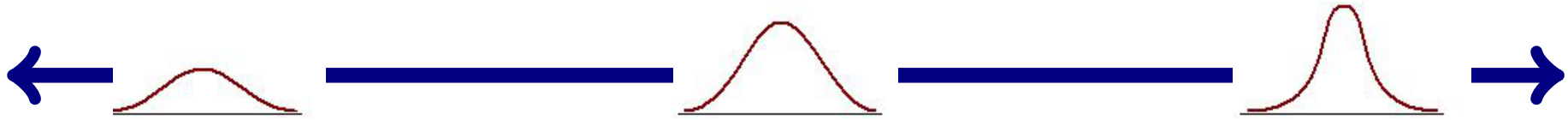


Negatively Skewed
(Skewed to the Left)
Skewness < 0

Standard Normal Curve
Symmetric
Skewness = 0

Positively Skewed
(Skewed to the Right)
Skewness > 0

Kurtosis



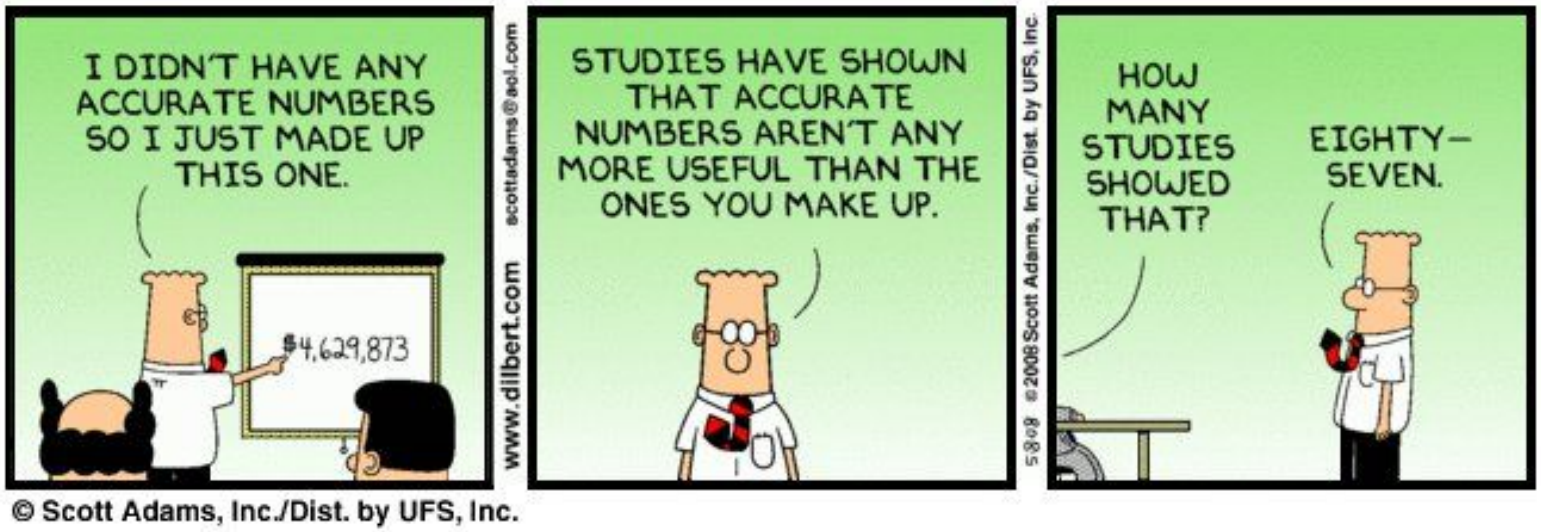
Platykurtic distribution
(Low degree of peakedness)
Kurtosis < 0

Standard Normal Curve
Mesokurtic distribution
Kurtosis = 0

Leptokurtic distribution
(High degree of peakedness)
Kurtosis > 0

Outliers

- An outlier is an observation which does not appear to belong with the other data.
- Outliers can arise because of a measurement or recording error or because of equipment failure during an experiment, etc..
- An outlier might be indicative of a sub-population, e.g. an abnormally low or high value in a medical test could indicate presence of an illness in the patient.



Detecting Potential Outliers

There are many criteria for identifying[‡] outliers, all have issues. We will adopt the criteria:

> *Points that are beyond the quartiles by one-and-a-half IQR's will be deemed* **potential** *outliers.*

Detecting Potential Outliers

STEP 1 Determine, Q_1 and Q_3, the first and third quartiles.

STEP 2 Determine IQR, the interquartile range.

STEP 3 Determine the lower and upper fences

$$\text{LIF} = Q_1 - 1.5 \times \text{IQR} \qquad \text{UIF} = Q_3 + 1.5 \times \text{IQR}$$

Points beyond these fences are potential outliers.

[‡]Automatic removal of outliers is very dangerous. Google "automatic removal of outliers ozone" and see `www.math.uni-augsburg.de/htdocs/emeriti/pukelsheim/1990c.pdf`

Part IV

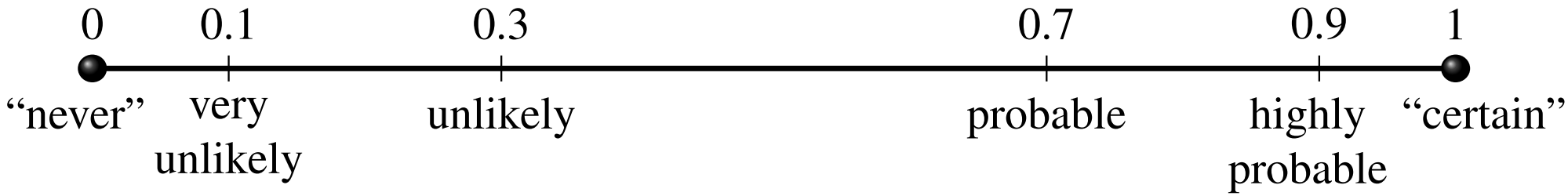
Probability

Concept: Probability

Definition 1 (Probability)

Probability is a measure of the likelihood or chance that a particular event would occur. It is given as a number in the range 0 to 1.

- Informal interpretation ...



- Other scales can be used such as
  - odds for (betting, medicine)

$$\text{odds for} = \frac{\text{probability}}{1 - \text{probability}} \iff \text{probability} = \frac{(\text{odds for})}{(\text{odds for}) + 1}$$

  - entropy (in machine learning/communication theory) uses scale $[0, \infty)$

$$\text{entropy} = (\text{probability}) \times \log_2 \left(\frac{1}{\text{probability}} \right)$$

Probability Laws I

Probability Law:

If an event, E, is a particular outcome then $\Pr(E)$ represents the probability that E will occur and

$$0 \leq \Pr(E) \leq 1 \tag{2}$$

The sum of the probabilities of all possible outcomes of an experiment equals 1, i.e.,

Probability Law: (Total Law of Probability)

If $E_1, E_2, \ldots, E_n$ are all the mutually exclusive outcomes of an experiment then

$$\Pr(E_1) + \Pr(E_2) + \cdots + \Pr(E_n) = 1 \tag{3}$$

In English — "Some outcome must happen."

Probability Law: (Independent Events)

Events A and B are independent if and only if

$$\Pr(A \text{ AND } B) = \Pr(A)\Pr(B) \tag{4}$$

Probability Laws II

If two events are mutually exclusive, then they cannot both happen ...

Probability Law: (Mutually Exclusive)

If events A and B are mutually exclusive then

$$\Pr(A \cap P) = \Pr(A \text{ AND } P) = 0 \tag{5}$$

Probability Law: ('OR'ing events)

If A and B are two events then, in general

$$\Pr(A \text{ OR } B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \tag{6}$$

If A and B are mutually exclusive then, from Equation (5),

$$\Pr(A \text{ OR } B) = \Pr(A) + \Pr(B). \tag{7}$$

i.e., OR $\longrightarrow +$

Probability Law: (Classical Probability 1)

If an experiment has n possible outcomes, with all equally likely, then the probability of any one of these outcomes occurring is $1/n$.

Example: Picking a card from a deck of 52 cards

$$\Pr(\text{Picking an ace of hearts}) = \Pr(\text{Picking a 4 of clubs}) = \cdots = \frac{1}{52}$$

Probability Law: (Classical Probability 2))

Given an experiment with each possible outcome equally likely and

S = Sample Space — Set of possible outcomes

E = Desired result — Set of outcomes that give the desired result

then

$$\Pr(E) = \frac{\#E}{\#S} = \frac{\text{Number of desired outcomes}}{\text{Number of possible outcomes}} \tag{8}$$