# INM430
## Principles of Data Science

Week 03
# Data Processing & Summarization

Aidan Slingsby

CITY
UNIVERSITY OF LONDON
EST 1894

# Module Schedule

- Week 01: Introduction & Basic Concepts
- Week 02: Data Characteristics & Wrangling
- **Week 03: Data Processing & Summarization**
- Week 04: Inferential Statistics
- Week 05: Relationships and Supporting Analysis using Models and Prediction
- Week 06: Reading week (no lectures)
- Week 07: Finding structure in data
- Week 08: Analysing text
- Week 09: Networks and Knowledge Representation
- Week 10: Processing data from images
- Week 11: Wrap-up (coding in the Real World)

# Today

- Review of last week's practical
- Data processing
  - Missing values (from last week)
  - Binning
  - Transforming
- Descriptive statistics
  - Probability distributions
  - Statistical assumptions
  - Outliers
- (Inferential statistics next week)

# DATA PROCESSING

# Missing values (covered last week)

- We often have missing values in “real” data
  - Collated from sources/years with different columns
  - People didn’t answer a question
  - The measuring device didn’t measure a value (low battery, lack of network, conditions wrong)
  - Data lost
- Data processing point of view
  - Remove record, remove variable, replace, impute?
- Data analysis (later today)
  - How representative are the means?

# Missing values (covered last week)

- Check the **nature** of the missing values
  - How much (or what proportion)?
  - Do they relate to other missing values?
  - Is it random or might it introduce bias?
- What you do depends, so **think**
  - **Remove:** Whole row or whole column, or values? How will this affect other analyses?
  - **Replace:** Zero, mean/median, local mean/median, or sample from existing values
  - **Impute:** Interpolate or otherwise model the value
- Try some alternatives and see effect

# Missing values: case for removing variable

- Some variables with **many missing values** may be too **unreliable** to use and sufficiently **uninteresting** to consider

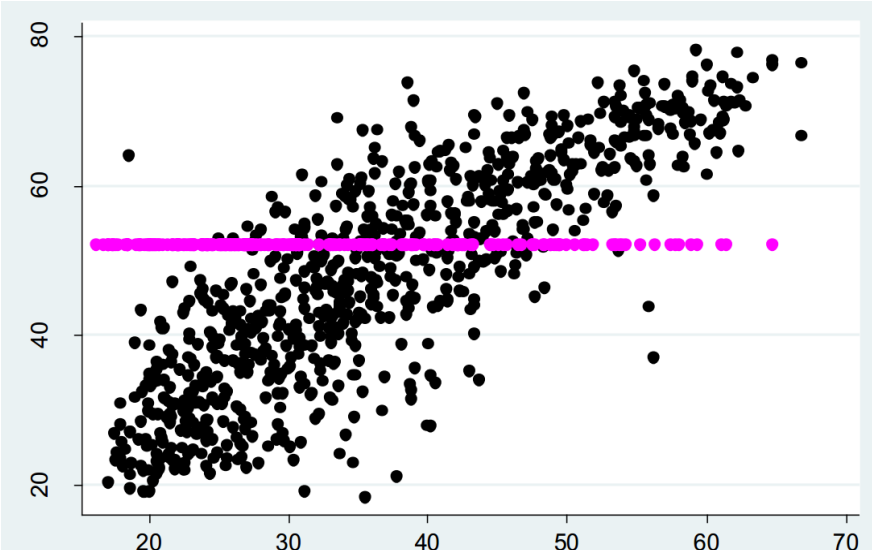
# Missing values: case for removing records

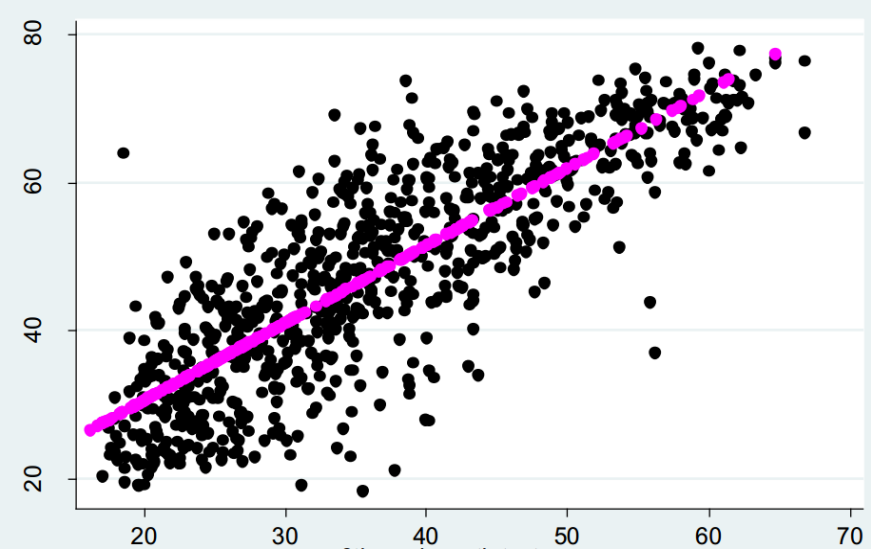
- If you have **enough data** and it seems to be **MCAR** (missing completely at random – unlikely!), consider **removing** records with missing data
  - **Check** if they seem to be MCAR
    - Use knowledge of the collection process (if you know it) – is it likely?
    - Try computing summary statistics of columns with and without records with missing values
  - **Keep copy** of original data
    - Recommend you add a **Boolean exclude column** to you can easily filter or not

# Missing values: case for imputing

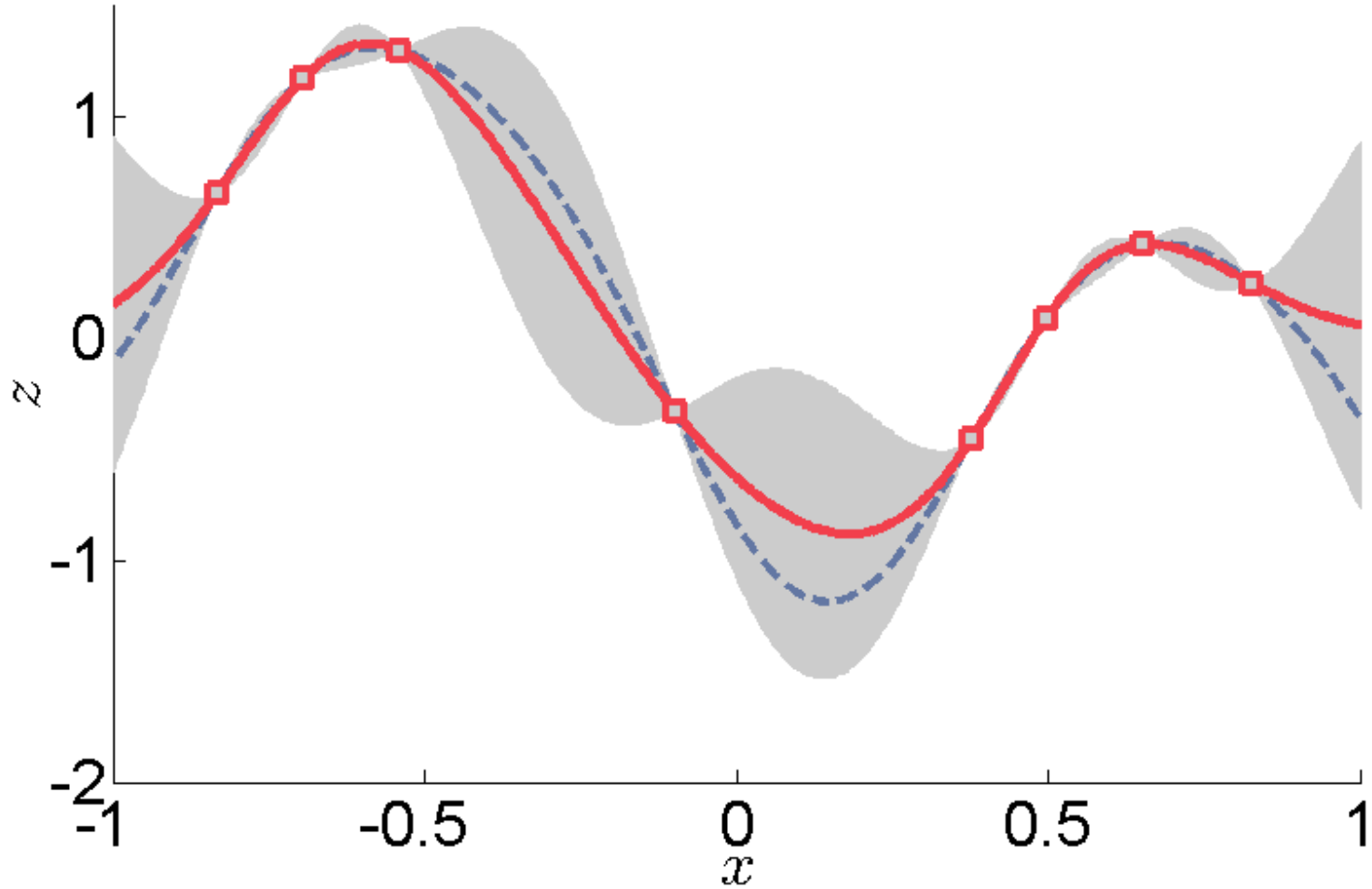
- **Keep a record** of what you have done
  - Boolean “replaced” column
- **Don’t replace with zero** by default
- **Replacing with mean/median** is better, but also problematic
- Replacing with a **local** mean/median is better, if you spot such a relationship; e.g. based on the mean/median of a category
- Replacing with globally or locally **sampled values** may retain existing variation

# Missing values: replace with zero vs regression



# Missing values: use stastical models




https://en.wikipedia.org/wiki/Kriging

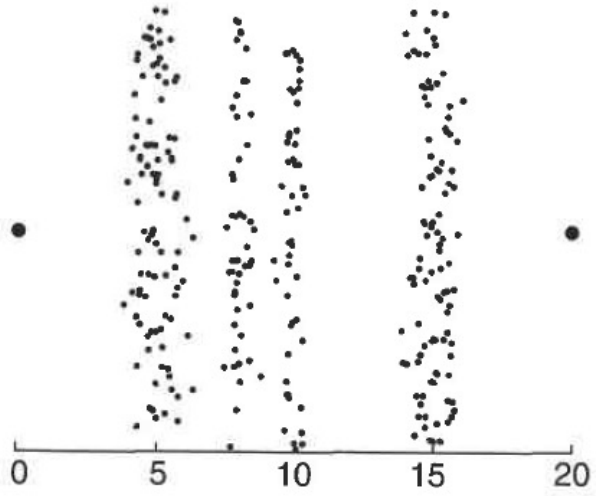
# Missing values: advice

- Be careful about “making-up/guessing” data!
- **Check nature** of missing values
  - Proportions, correlations)
- **Think** about what makes sense
  - Why might they be missing?
  - Do even you need the variables?
  - Do they seem to be MCAR?
- Try some **alternatives**
  - if you remove rows, stored as Boolean column
  - compute in separate columns
  - study the effects (are summary statistics the same?)
- Keep a **record** of what’s removed/imputed

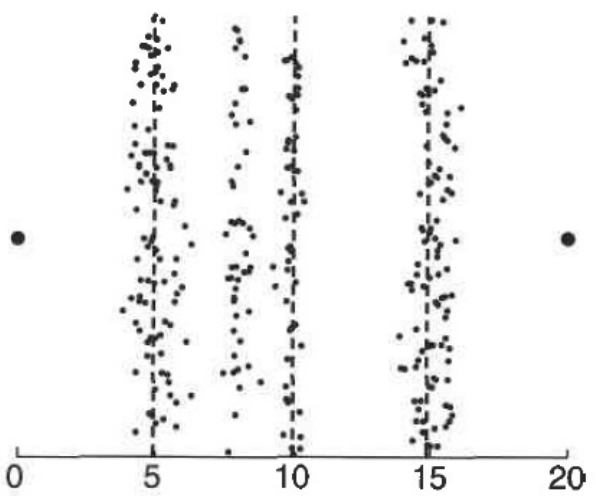
# Discretisation (binning)

- Transform a continuous attribute into a categorical attribute
- Partition the data into
  - Equal size (histograms, time-series, gridded)
  - Quantiles (equal number of data points in each partition)
    - e.g. quartiles
  - Based on gaps in the data
  - Based on predetermined thresholds
  - Based on analytical goals

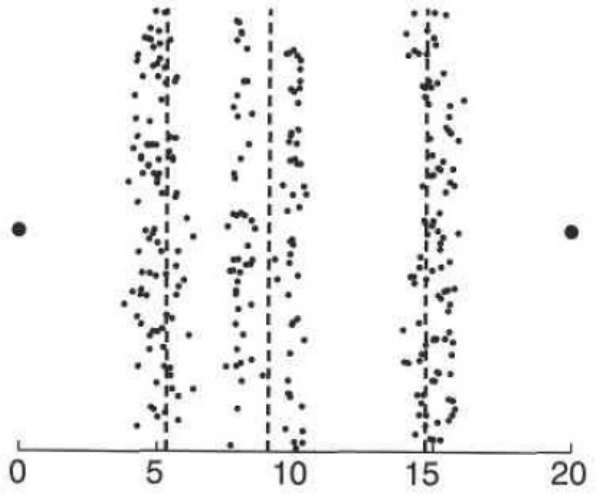
# Binning strategies



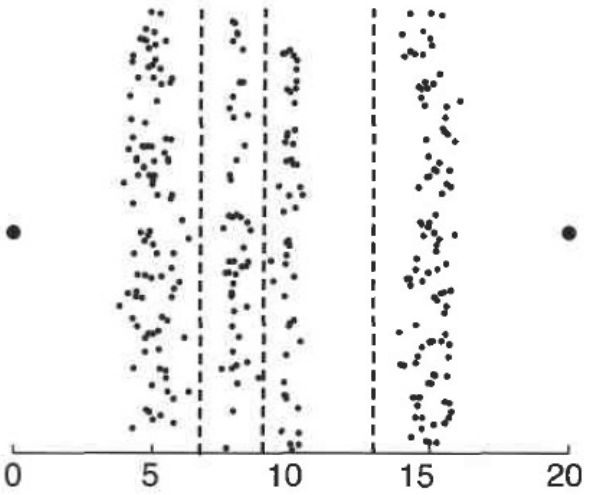

(a) Original data.




(b) Equal width discretization.

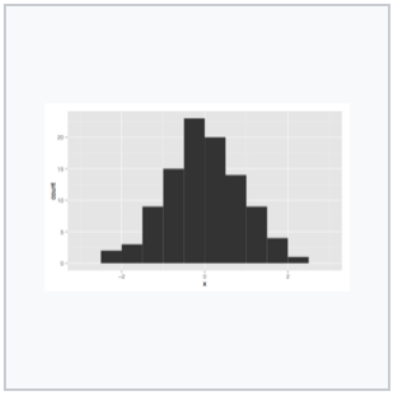



(c) Equal frequency discretization.

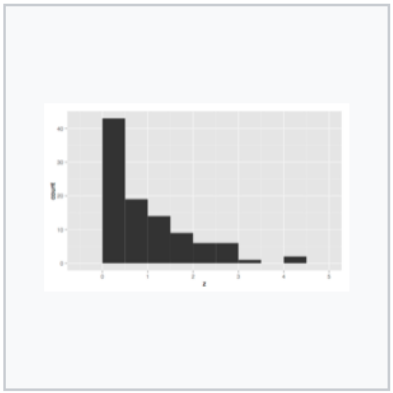



(d) K-means discretization.

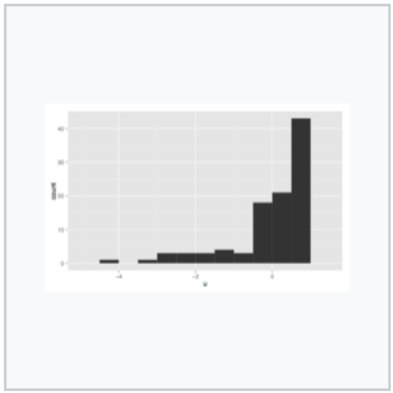
# Discretisation: equally-sized bins



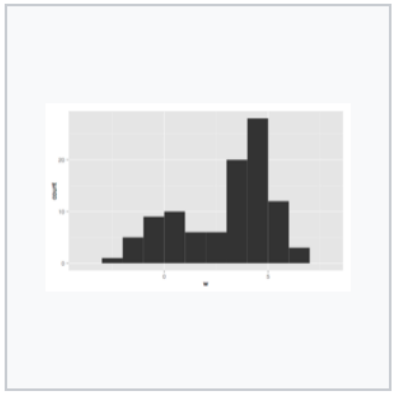
Symmetric, unimodal



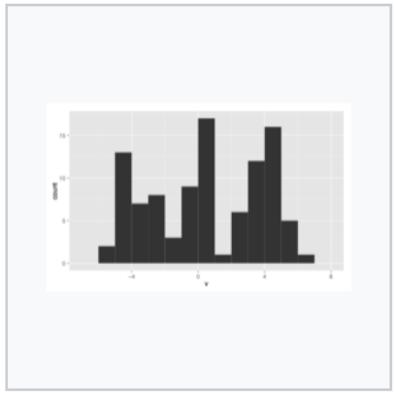
Skewed right



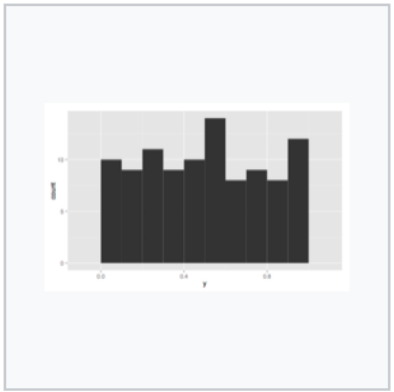
Skewed left



Bimodal



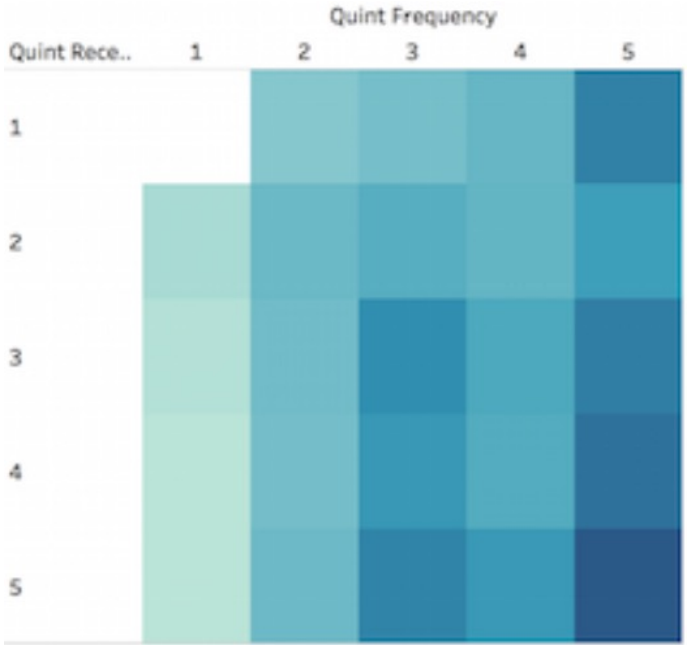
Multimodal



https://en.wikipedia.org/wiki/Histogram

# Discretisation: Quantiles (quintiles)

- Example: RFM (recency/frequency/monetary)
  - Sort customers by R/F/M and allocate them to bins with a fifth in each
  - The matrix combine two sets of bins.
  - Distribution of customers based on their RFM
    - (“best” customers in bottom right)



# Discretisation: Daps in the data

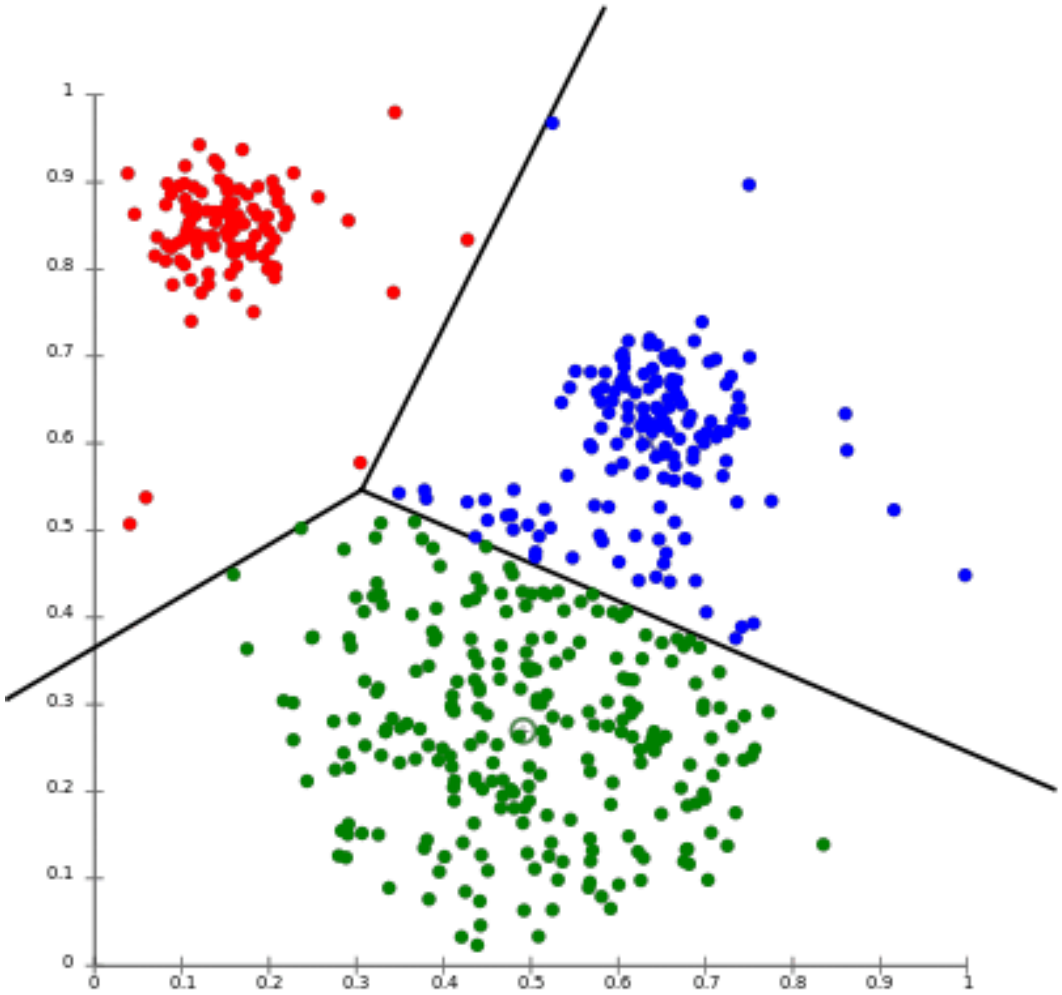


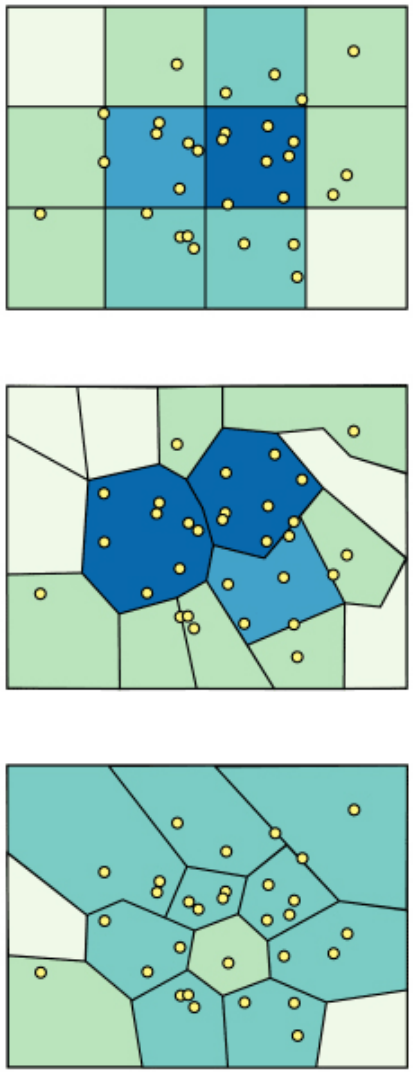

Image nicked from https://aws.amazon.com/blogs/machine-learning/k-means-clustering-with-amazon-sagemaker/

# Discretisation

- Predetermined thresholds
  - E.g. Tax bands
  - E.g. Internationally defined population/land use classes
- Analytical goals
  - E.g. Before/after a policy came into effect (2 categories)

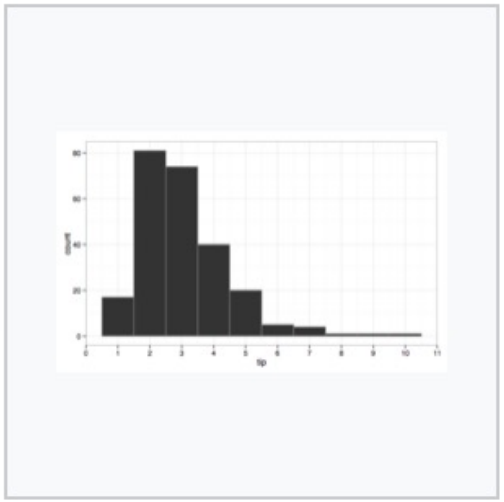
# Problems with discretisation

- In geographical analysis: “Modifiable Areal Unit Problem (MAUP)

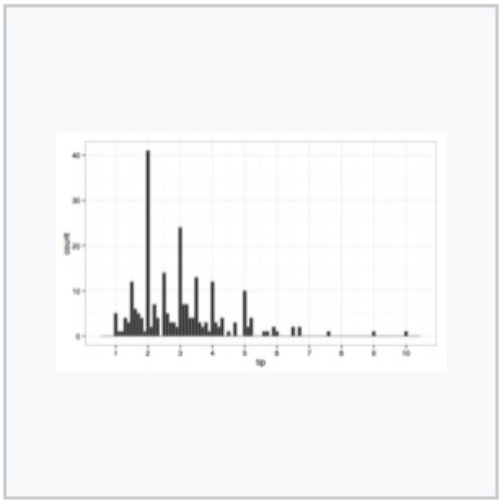


http://gispopsci.org/maup/

# Discretisation: Bin sizes



Tips using a $1 bin width, skewed right, unimodal



Tips using a 10c bin width, still skewed right, multimodal with modes at $ and 50c amounts, indicates rounding, also some outliers

https://en.wikipedia.org/wiki/Histogram

# Discretisation: Bin sizes

- Try different bin-sizes to see what works
- Examples of some algorithms:
  - Sturges' formula: $k = \lceil \log_2 n + 1 \rceil$

    depends on (log of) sample size
  - Scott's rule $h = \frac{3.5\hat{\sigma}}{n^{1/3}}$

    depends on standard deviation as well as sample size

More complex methods exist (for the interested):

- Histogram optimization (http://toyoizumilab.brain.riken.jp/hideaki/res/histogram.html )
- Variable sized bins (adaptive bandwidths) (https://jakevdp.github.io/blog/2012/09/12/dynamic-programming-in-python/ )

# Discretisation: advice

- For studying **distributions** (counts of values), consider using fixed-size bins
  - E.g. for histogram or gridded data
- For **partitioning** your data into top, bottom, middle, etc, for comparing summary statistics, consider using quantiles (this example: “terciles”)
  - Compare the means of the bottom, middle and top terciles.
- Other options for comparing summary statistics:
  - Natural breaks so groups are defined by the data rather than being artificially split
  - Domain-specific breaks (like tax-bands)

# Data transformation

- Transforming data variables so they are suitable for different analysis
  - Depends on what analysis you are doing, many make assumptions about data
- Types of transformation
  - Linear: rescale (or “normalisation”, “scaling” or “standardisation”)
  - Non-linear: Change the distribution (e.g. log)

# Data transformation: Rescaling/normalising

- Remove the effects of scale (magnitude)
- To make variables comparable by putting things on a common scale
  - Percentage
  - Range normalise (0-1 rescale) or use a different range (e.g. 5-95$^{th}$ percentile)

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

  - Z-score standardisation (subtract mean and divide by standard deviation)

$$\frac{X - \mu}{\sigma}$$

  - Difference from a reference value

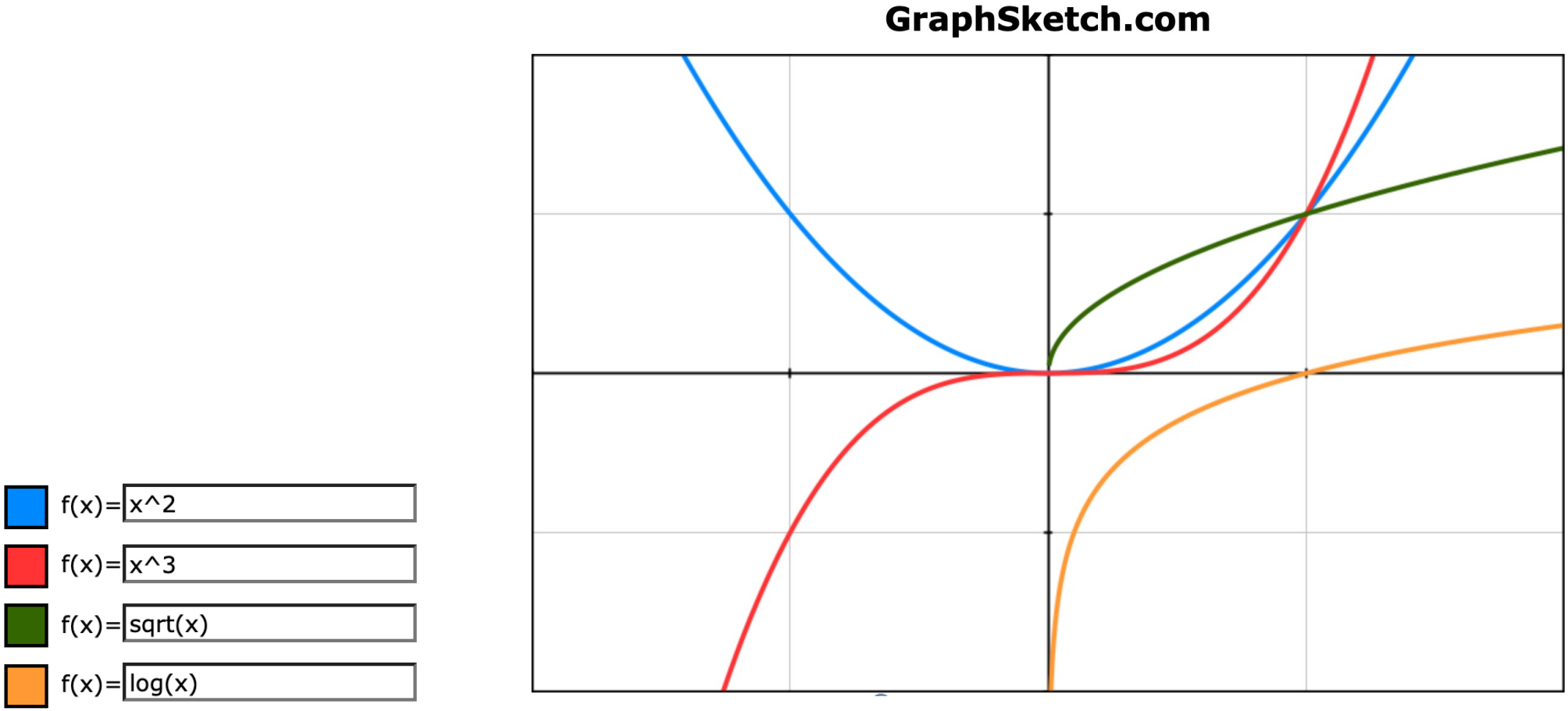
https://en.wikipedia.org/wiki/Feature_scaling

# Rescaling/normalizing: advice

- Use proportions when denominators are different
  - Eg. Population-related variable between regions
- Use z-scores to compare values with different units
- Beware of range-normalizing: affected by outliers
- To reduce the effect of outliers
  - Consider removing the1$^{st}$ and 99$^{th}$ percentiles
  - Consider log-transforming (de-emphasizes **but distorts**).

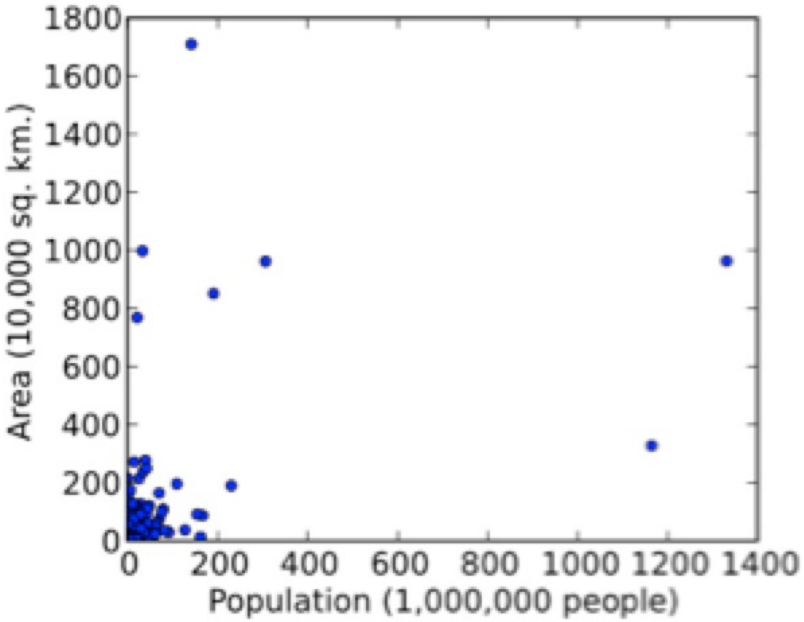
# Data transformation: Change distribution

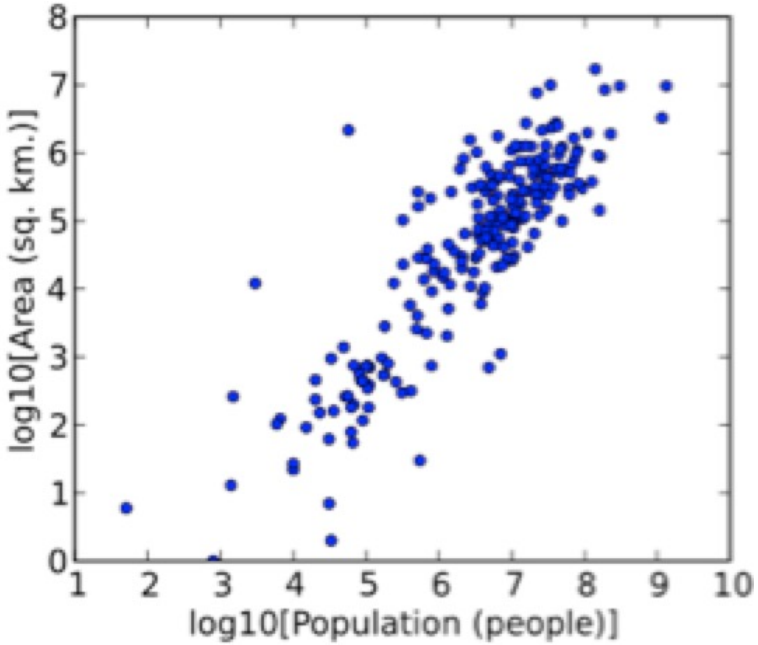
- Apply a mathematical operator to all values to change the distribution of the data
  - E.g. Give more weight to low/high values

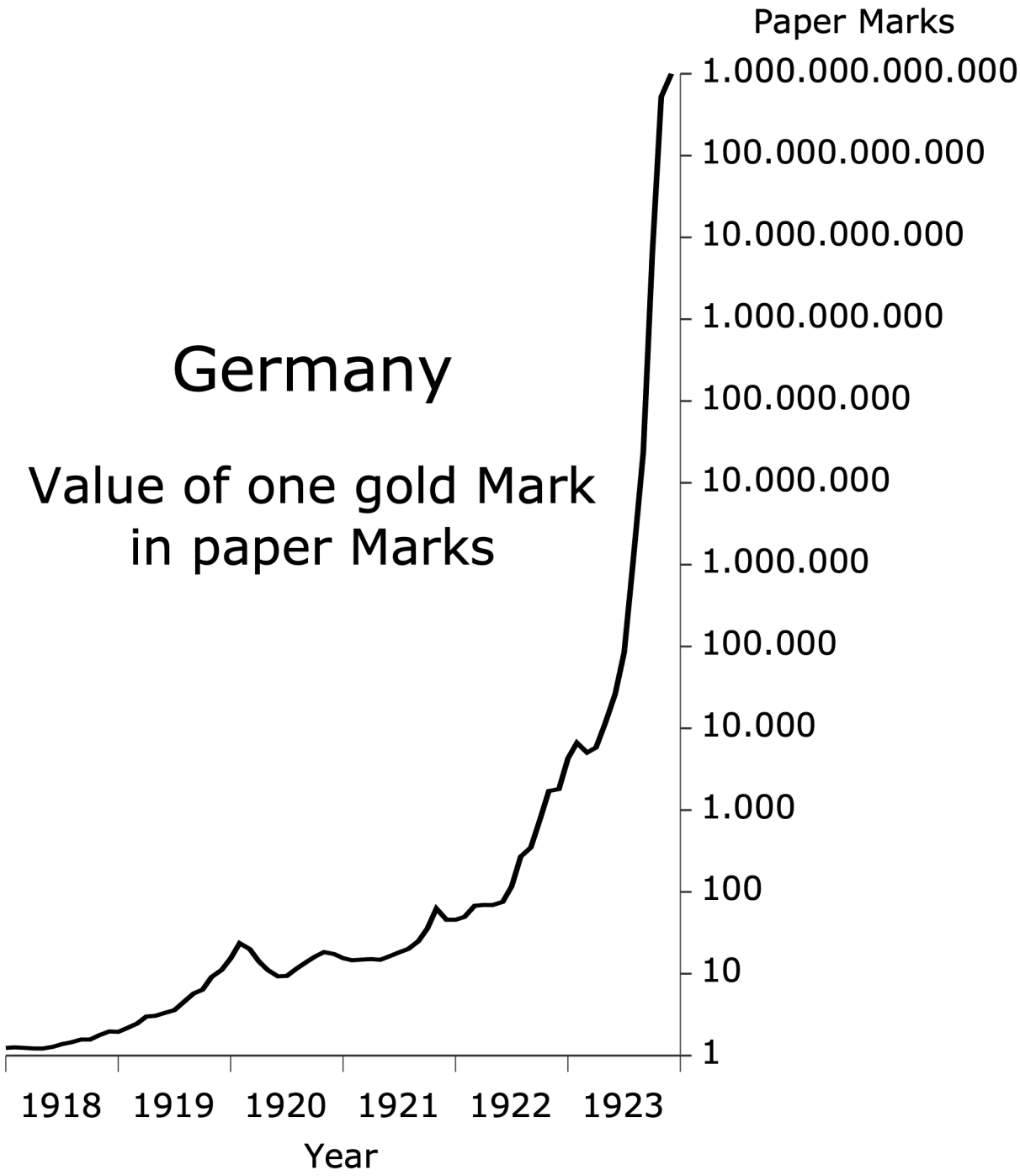


# Data transformation: Change distribution: **log**

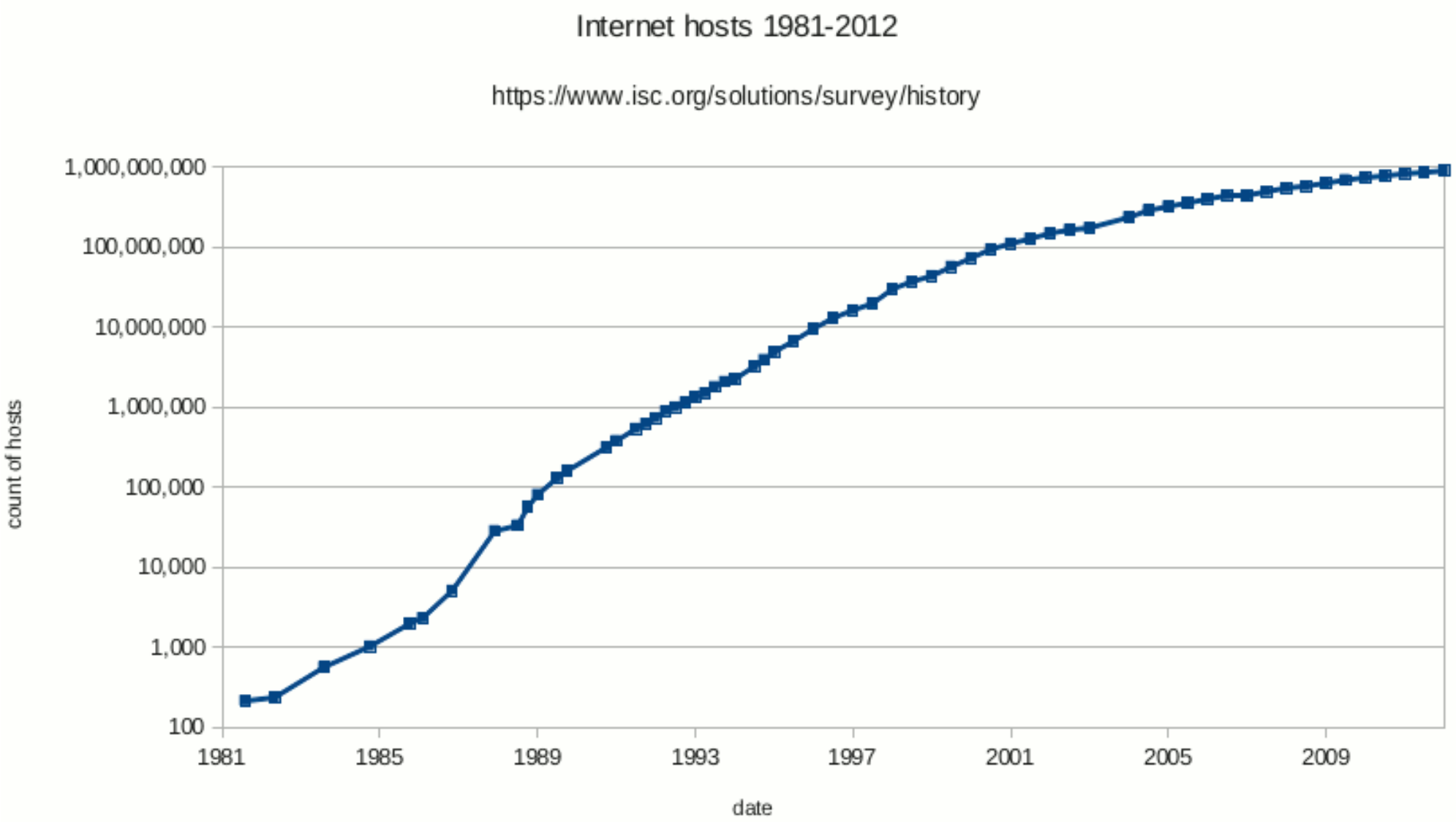
- A useful way to consider phenomena that change exponentially
  - https://en.wikipedia.org/wiki/Logarithmic_scale
- **Take the *log* of each observation**
- Spreads out small values and compresses high values



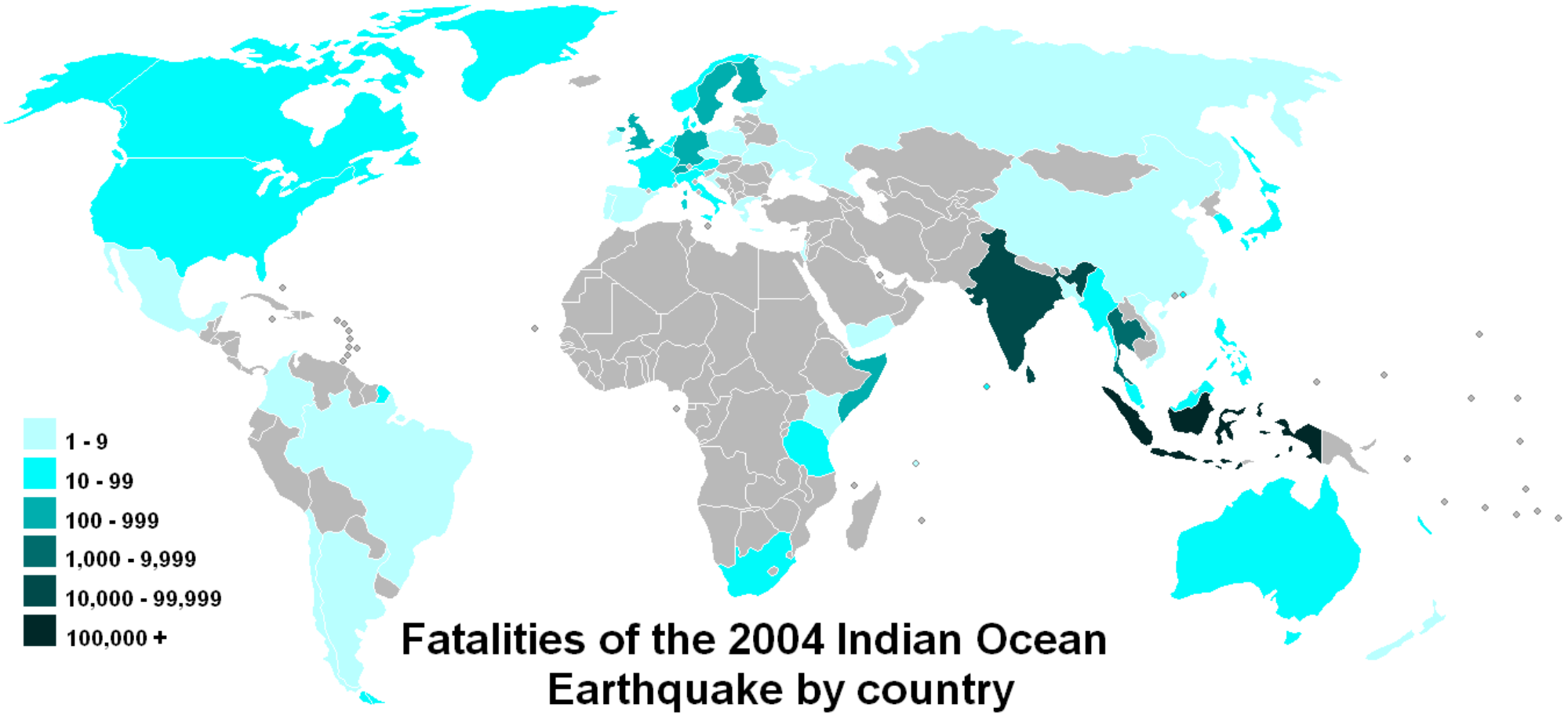






https://en.wikipedia.org/wiki/Logarithm#/media/File:Germany_Hyperinflation.svg




https://en.wikipedia.org/wiki/Logarithmic_scale#/media/File:Internet_host_count_1988-2012_log_scale.png




https://en.wikipedia.org/wiki/Logarithmic_scale#/media/File:COB_data_Tsunami_deaths.PNG

# Change distribution: advice

- Sometimes used to reduce effect of outliers
  - I don't think this use: would prefer to remove outliers (assuming you know how to define them)
- Usually for turning your data into a normal distribution
  - Most parametric models assume this (e.g. regression)
  - …but remember the transformations if you need to interpret – more later!

# Data transformation: Reasons

- Rescaling:
  - Put data on a **common scale** (e.g. percentage)
  - Only consider **part** of the distribution (e.g. remove the outliers and/or tails)
- Transforming:
  - Emphasise low values over high values (or visa versa)
  - Allow logarithmic relationships to be modelled with linear models
- Applications:
  - Computational: building models, statistical comparisons
  - Visual: e.g. colour scales and axis
  - Statistical: making distributions more normal (when parametric statistics; can make interpretation more difficult)

# DS Process

- Understand domain needs
- Collect & make data available
- Get the data ready for analysis
- Exploratively (and visually) analyse the data
- Model the phenomena (if needed)
- Evaluate findings
- ITERATE (from any stage to any other stage)!
- Communicate findings

# DESCRIPTIVE STATISTICS

# Let’s collect some data

https://forms.office.com/e/RJyZD2jhcr

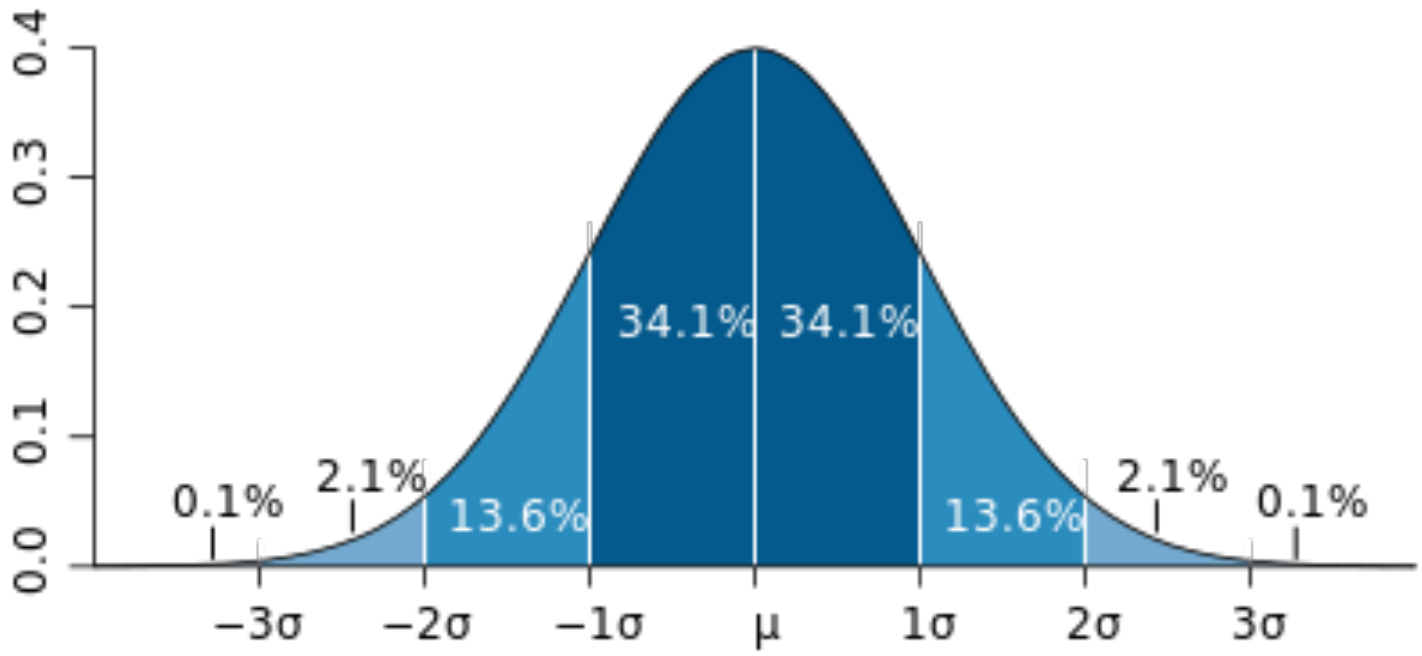
# Descriptive statistics

- **Descriptive statistics**
  - quantitatively describe the main features of data
  - distinguished from inferential statistics
    - **descriptive statistics:** summarize a sample (today)
    - **inferential statistics:** learn about the population that the sample of data is thought to represent (next week)

# Descriptive statistics

- Statistical characteristics of 1D distributions
  - Centrality
  - Spread
  - Skewness
  - Kurtosis
- Summarises the information
- Provides an overview on the distribution of the data
- Often also used with inferential methods

# Probability distribution

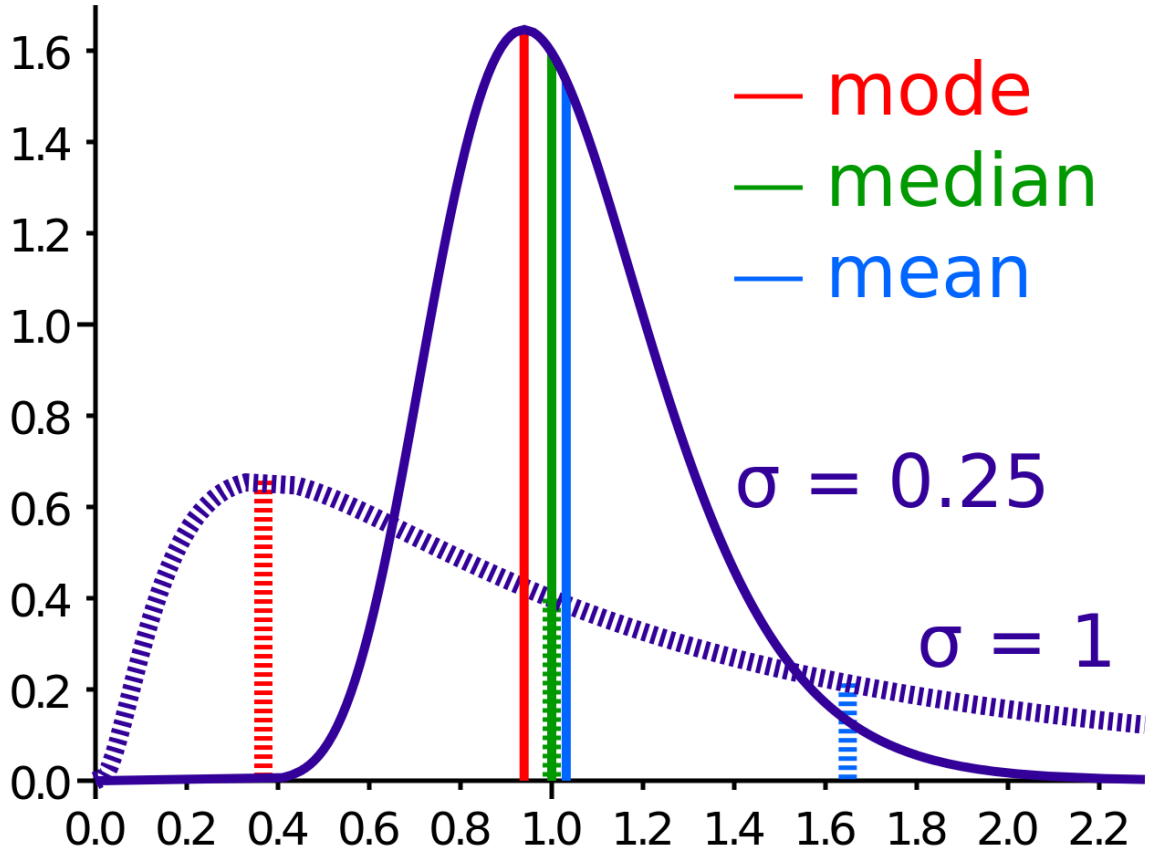



**Probability distribution** for normal distribution

(links each outcome of a statistical experiment with its **probability of occurrence**.)

# Centrality: mean/average, median, mode

- Mean
  - centre of the distribution
- Median
  - middle value
  - "robust"
- Mode
  - most frequent value
  - Often more use for **categorical data**

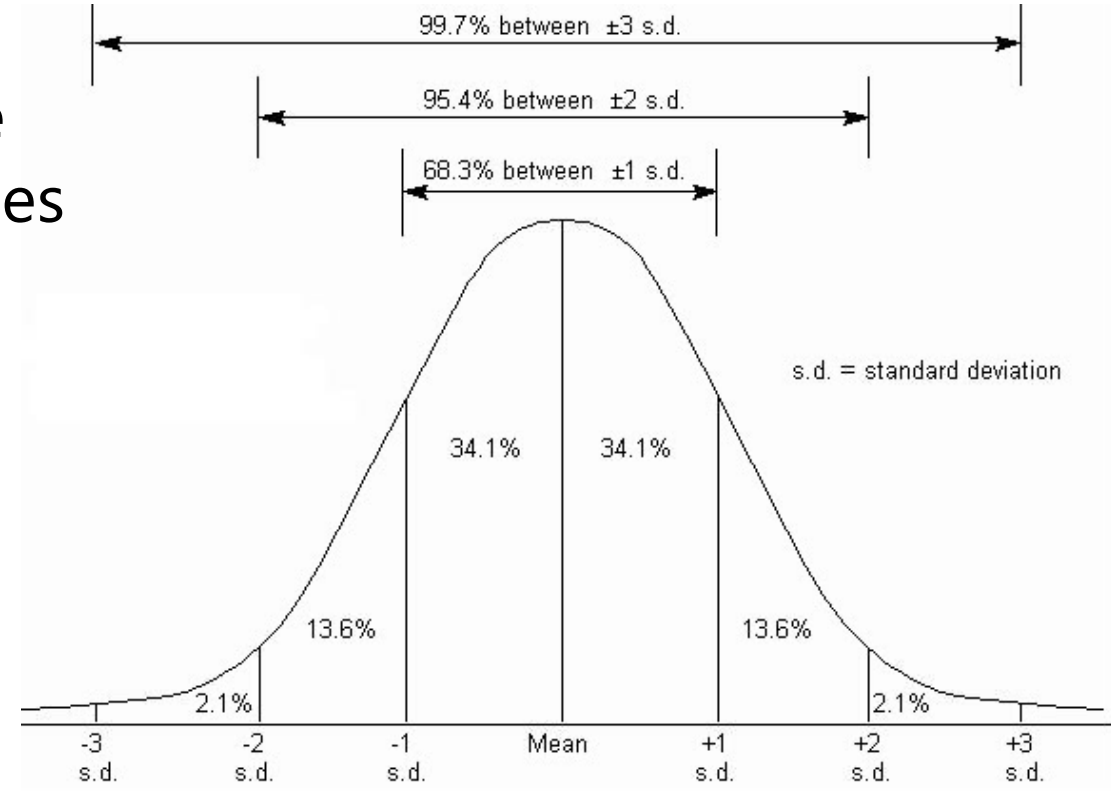


# Spread: variance, standard deviation, IQR

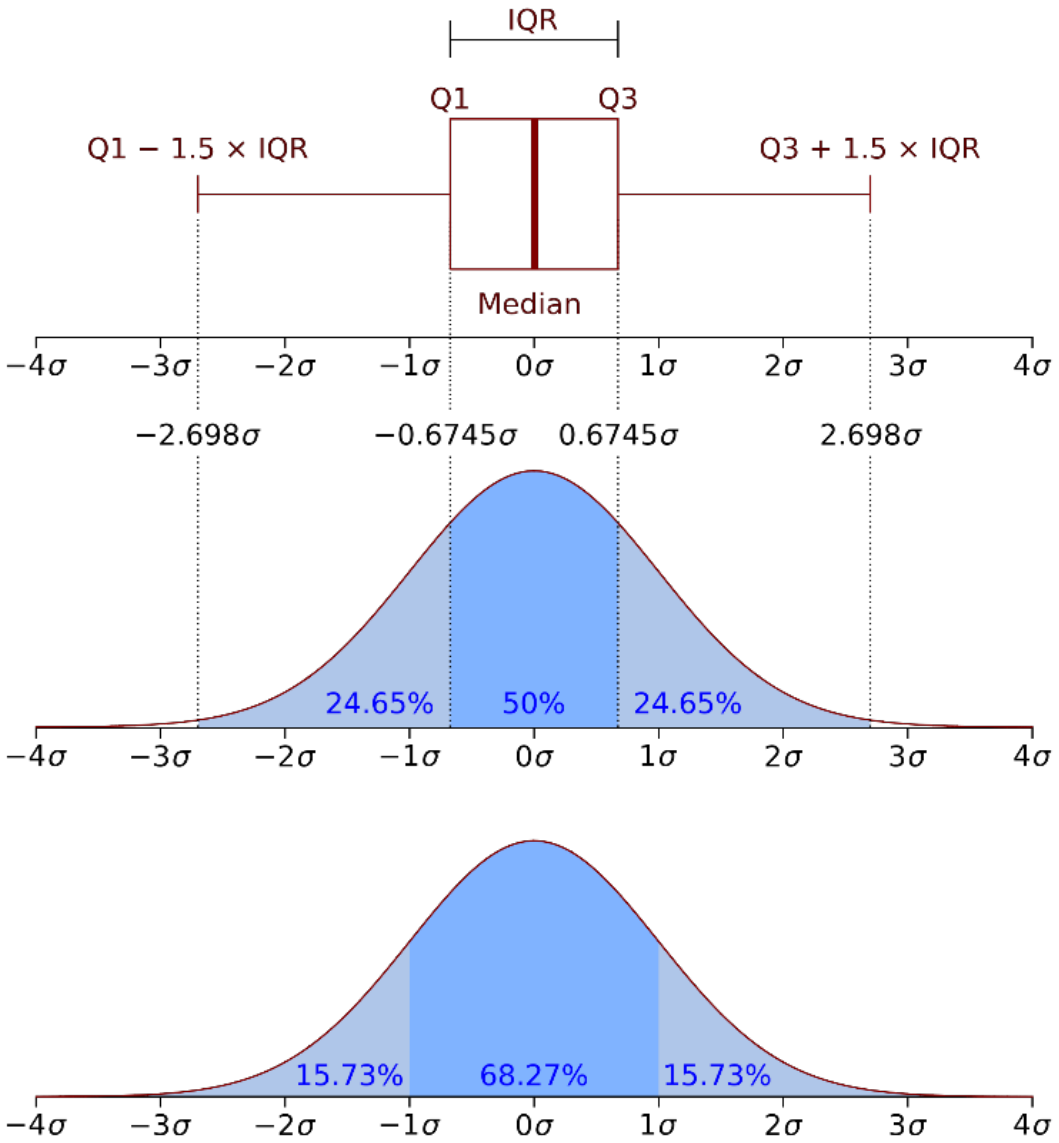
- Variance
  - average of the squared deviations from the mean
- Standard deviation
  - Square root of the variance

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

- Interquartile range
  - The range between the 25% and 75% percentiles
  - “robust”

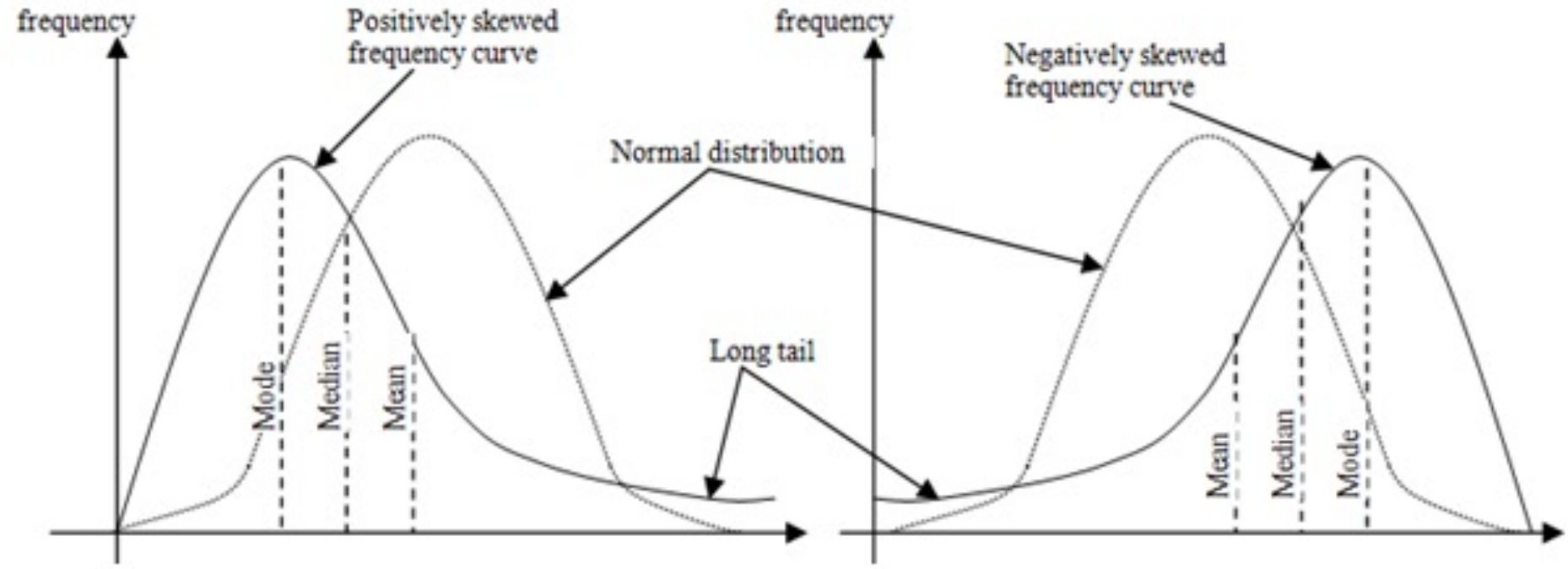


# Spread: variance, standard deviation, IQR



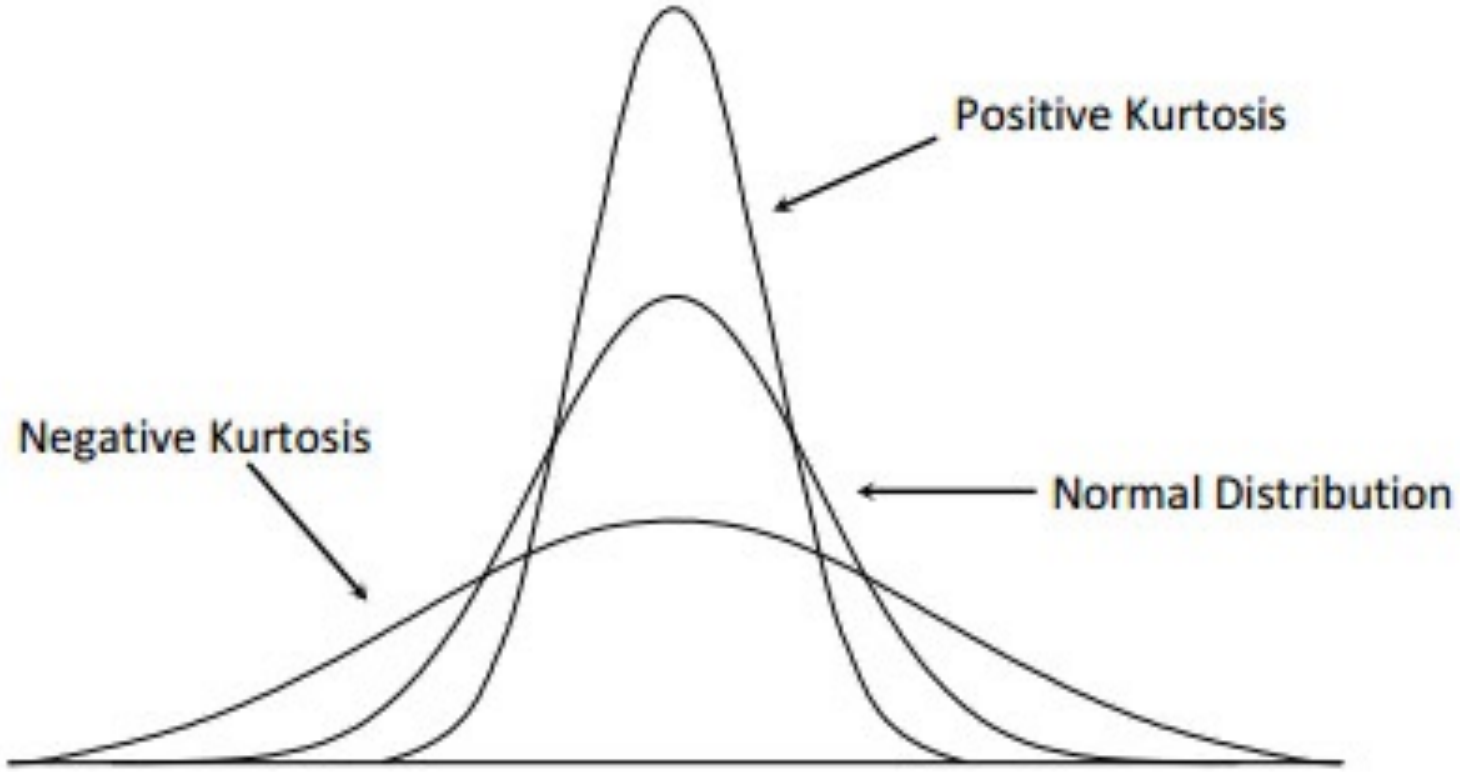
# Skewness

- ***negative skew*:** The left tail is longer: *right-skewed*, *right-tailed*, or *skewed to the right*
- ***positive skew*:** The right tail is longer: *left-skewed*, *left-tailed*, or *skewed to the left*



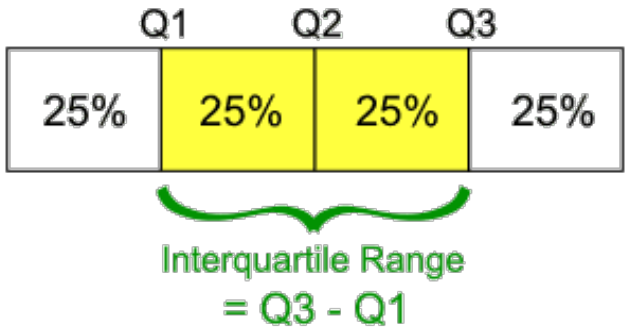
# Kurtosis

- Measure of the "peakedness"
- Unstable statistics, need ~ 500 samples at least



# Quantiles

- points taken at **regular intervals** from the **cumulative distribution function** (CDF) of a random variable
  - The 2-quantile is called the **median**
  - The 4-quantiles are called **quartiles** (Q1, Q2, Q3)
  - The 10-quantiles are called **deciles**
- **IQR** : Inter-quartile range
  - Q3 – Q1
  - Can be a basis to define a threshold to identify outliers



# Missing data and relationship to categories

- Missing data?
- Other categories

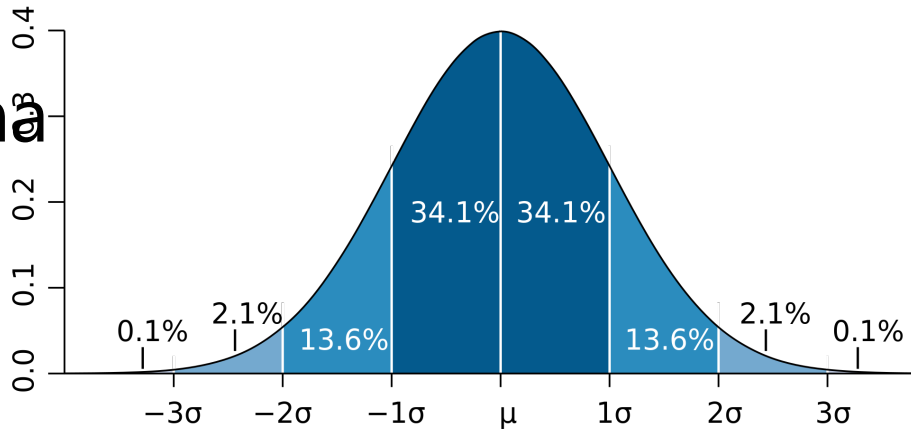
# Categorical data (& other qualitative data)

- Frequency/counts
  - how frequent is one category
- Proportions
  - Normalised by total
- Mode (most common category)
- Aggregate categories
- Qualitative data - make into categorical data
  - E.g. extracting themes/verbs from text, sentiment, NLP

| | Undergrad | Graduate | Staff |
|---|---|---|---|
| Counts | 17 | 1 | 2 |
| Proportions | 0.85 | 0.05 | 0.1 |

# Common Distributions: **Normal/Gaussian**

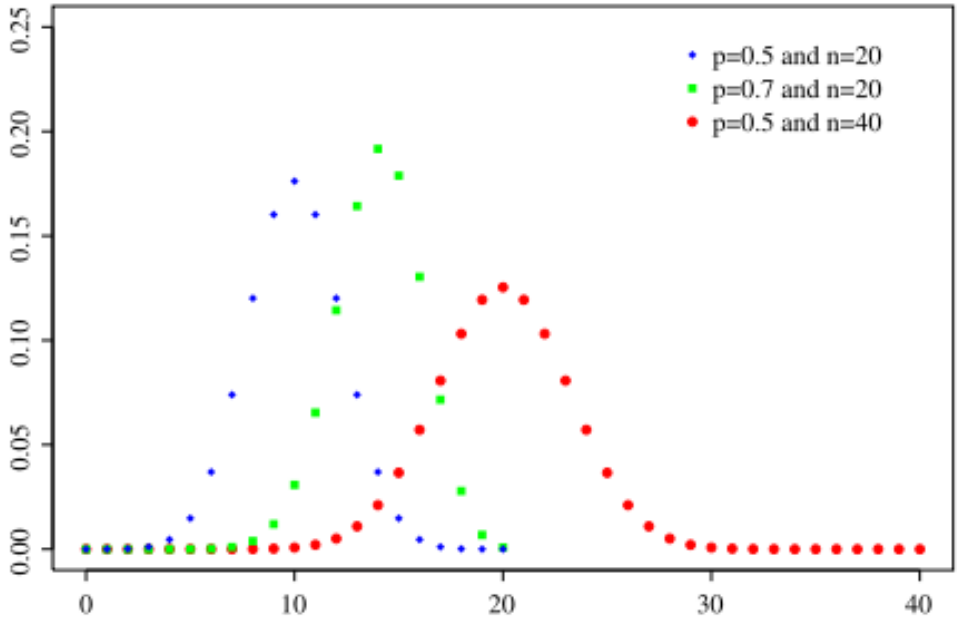
- **Continuous** distribution followed by many phenomena
  - including sample means (Central Limit Theory)




- Most statistical models assume the underlying data is normally distributed
- Examples
  - Heights of people, marks in a class
- Can be described with two parameters:
  - mean & standard deviation

# Common Distributions: **Binomial**

- A discrete probability distribution
  - Number of successes in sequence of independent yes/no experiments
  - Looks ‘normal’ if there are enough samples
  - Can be used to predict how many success outcomes




- Examples
  - number of times that a component fails
- Can be described with two parameters:
  - Number of samples (*n*) & probability of success (*p*)

# Examples for DS Lecture, Week 3

*Phong Nguyen, 10 Oct 2018*

## Binomial distribution

### Task: Draw samples from a Binomial distribution and visualise them

- a straightforward way: use a library like numpy
- a fun way: simulate data by hand

[1]:
```
n = 100
p = 0.5
```

First, simulate a biased coin with probability of Head as $p$.

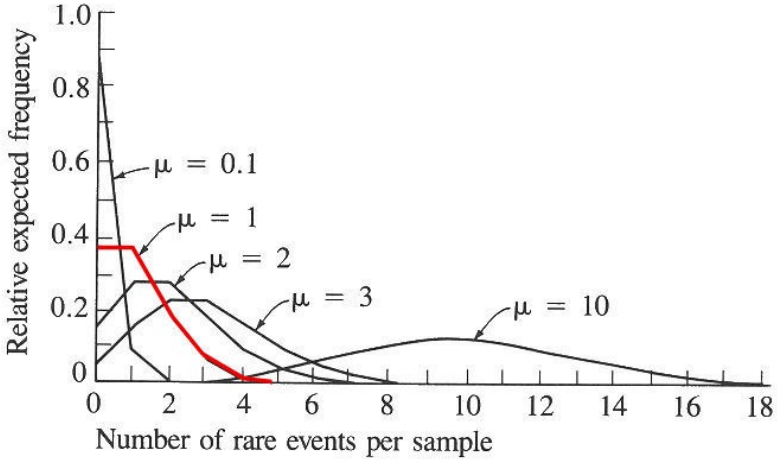
[2]:
```
from random import random

def flip_coin(p):
    'Return H or T with probability of H as p.'
    return 'H' if random() <= p else 'T'
```

[3]:
```
flip_coin(p)
```

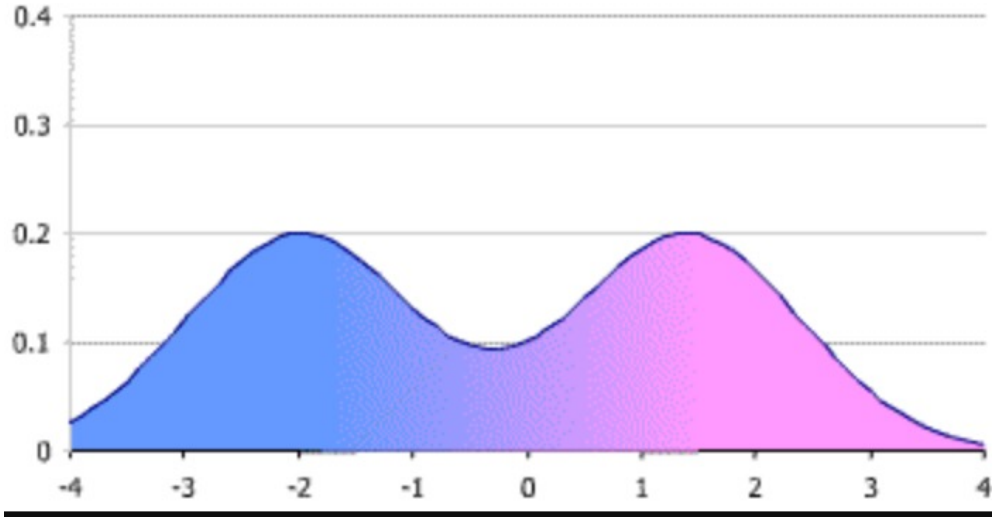
# Common Distributions: **Poisson**

- Another discrete probability distribution
  - Counts (of random "events" that occur at a certain rate)
- Examples
  - Customer sales in on a particular days; number of hurricanes in a year; number of times software fails within a time period
- Can be described with one parameter (lamba):
  - Number of occurrences within time frame



# Common Distributions: **Bimodal** (or multimodal)

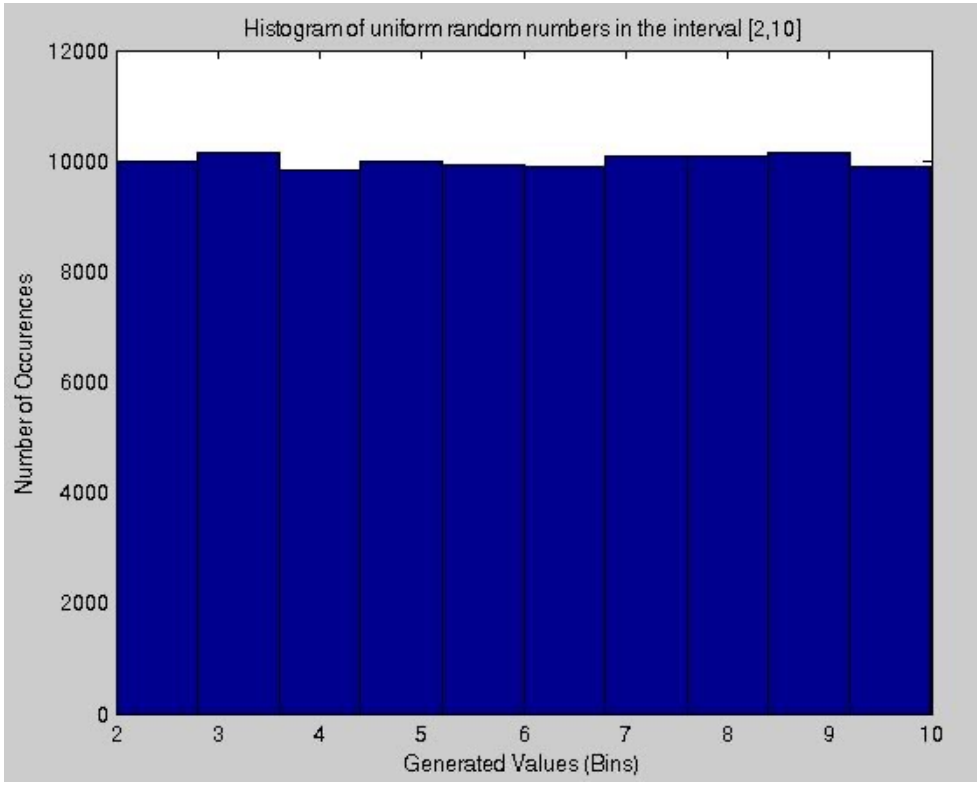
- Two modes
- May indicate that there are two cohorts with different properties
  - …so you might consider finding out how to distinguish them, and treat differently




- No (single) central tendency, so mean/median etc may be misleading
- Example
  - Male and female heights
    - Apparently, this is a bit of myth because the modes similar, so it approximates normal.

# Common Distributions: **Uniform**

- All values are equally probable
  - Random numbers
  - Some categorial variables (suits in a pack of cards)



# Why do we need to know about distributions?

- Tells us how data columns are distributed
  - for validating and error/bias checking
  - for telling us if the phenomenon is behaving as **expected**
  - interpolation (a form of modelling)
- Multimodal may indicate a dependency on other characteristics. You may partition your data.
- Helps us know whether the data meets assumptions of the statistical techniques we want to use
  - mean & standard deviation assumes a normal(-ish) distribution
  - linear regression assumes residuals are normal
- We can transform data to
  - emphasise values depending on range (for visualisation/modelling)
  - prepare variables as “feature” for modelling (later in the DS process)

# Parametric vs non-parametric models

- **Parametric models** make assumptions about the population from which the sample is drawn
  - assumptions about how the data are distributed and this parameters of the distributions
  - based on statistical assumptions
  - data irregularities or deviations from this can be problematic
- **Non-parametric** models
  - do not have such assumptions
  - can be more complex to compute, i.e., costly
  - Examples: Kernel density estimation, Spearman correlation

# Some common assumptions in statistics

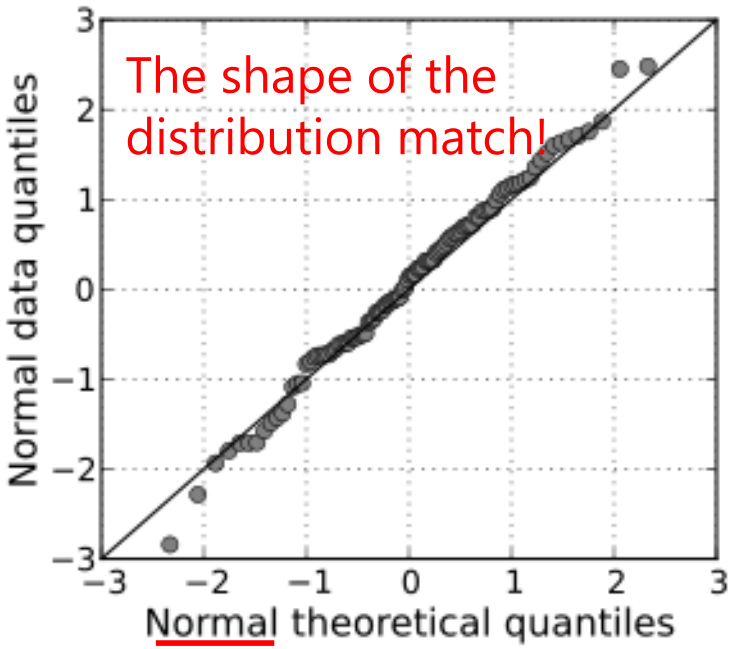
- Depends on the parametric statistics/methods
- Normality
  - Data/residuals have a normal distribution (or at least is symmetrical)
- Homogeneity of variances
  - Data from multiple groups have the same variance (Levene's test to check)
- Linearity
  - Data have a linear relationship
- Independence
  - Data are independent

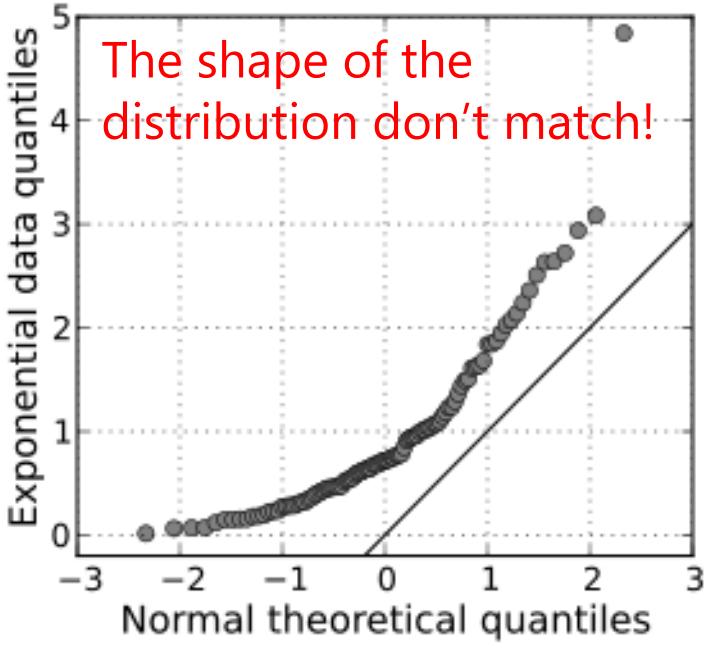
# Data are noisy

- Most real data is noisy and often only approximates conform theoretical/mathematical distributions
- Use to assess the reliability of statistics

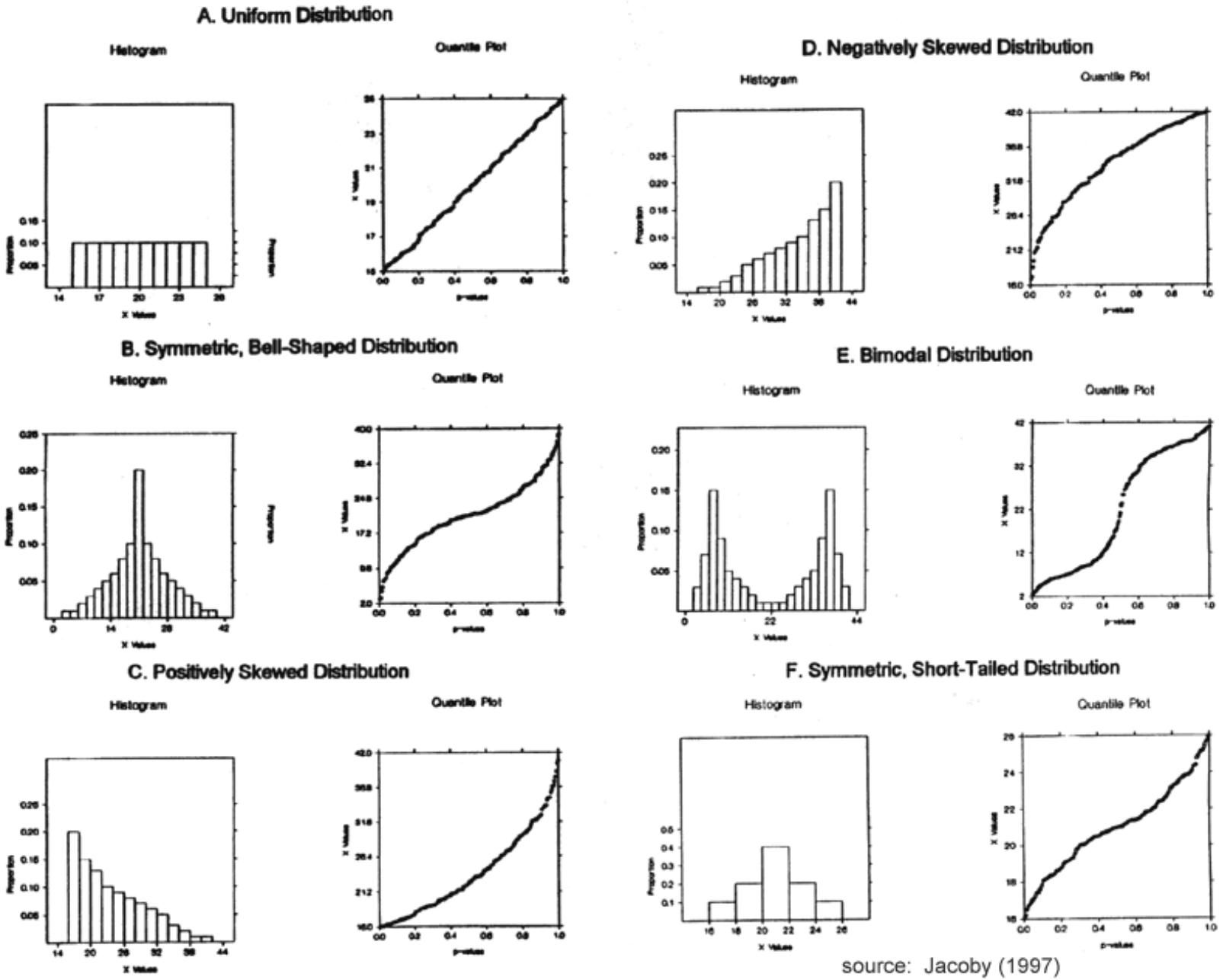
# Checking for normality: Normal Q-Q plots

- A **quantile-quantile (Q-Q) plot** is a **visual means** of comparing two distributions
  - Usually compare **data** with a **theoretical distribution**
  - **Normal Q-Q plots** are where the theoretical distribution is normal
  - We plot the corresponding quantiles



# Uniform Q-Q plots for various distributions



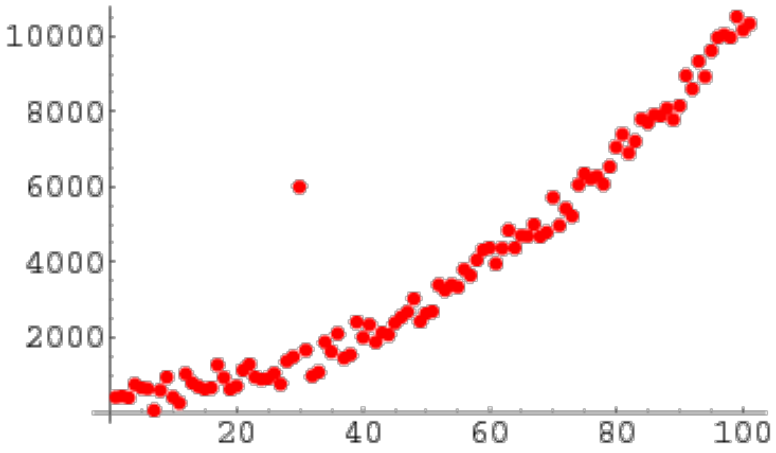

source: Jacoby (1997)

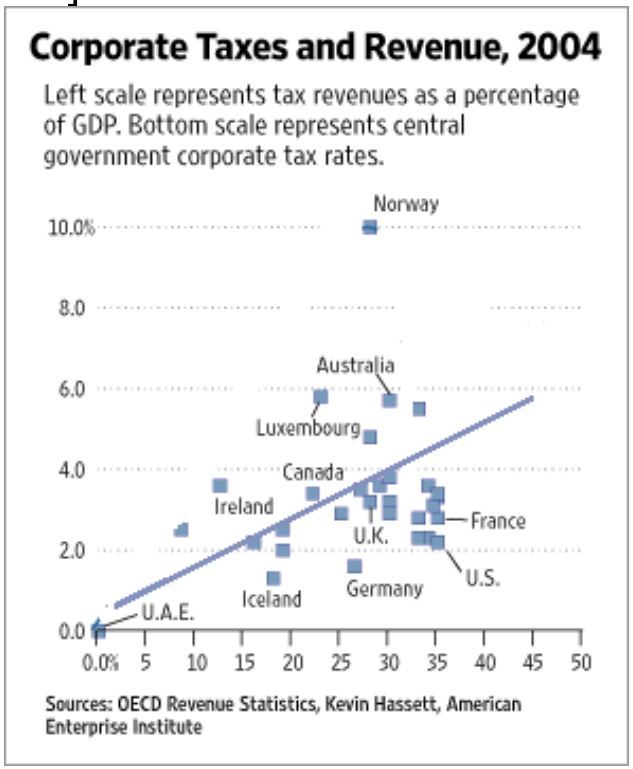
# Histograms, with superimposed theoretical distr

- Histograms with superimposed theoretical distributions
  - Also works, but may be more difficult to judge

# Outliers

- "An outlier is an observation which **deviates so much** from the other observations (**a.k.a. trends**) as to arouse suspicions that it was generated by a **different mechanism**" [Hawkins 1980]



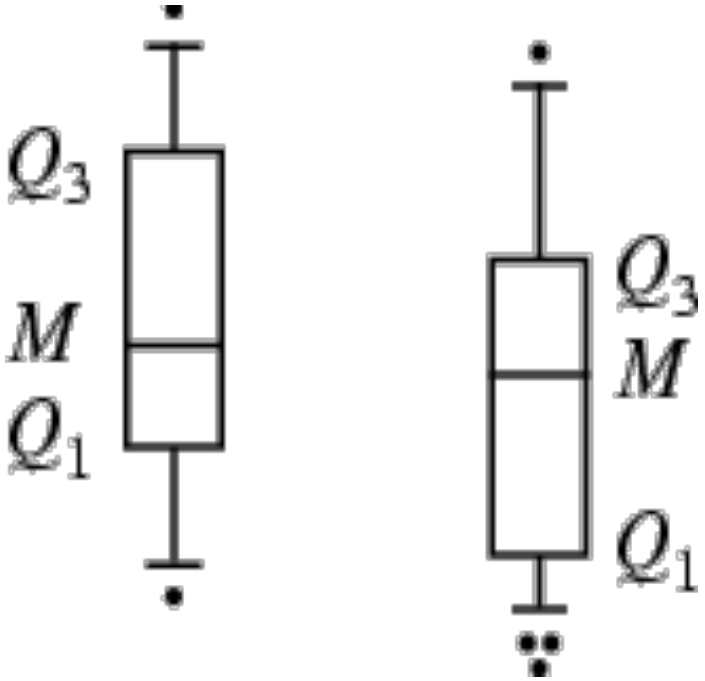



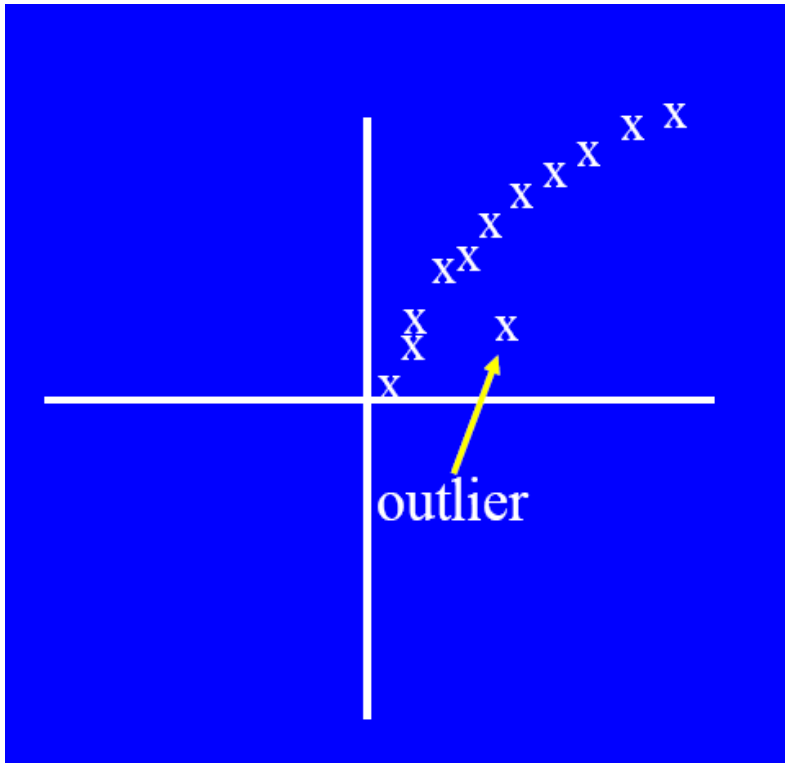
# Outliers – friend or foe?

- Often **tiny** proportions of our data (with the potential to ruin our analysis!)
- Might be **problematic values**
  - faulty readings, measurement errors, missing data, …
- Might be **what you are after**
  - fraud detection, network intrusion detection, ….
- Might be **something unexpected**
  - valuable analytical finding
  - might be filtered by automated methods

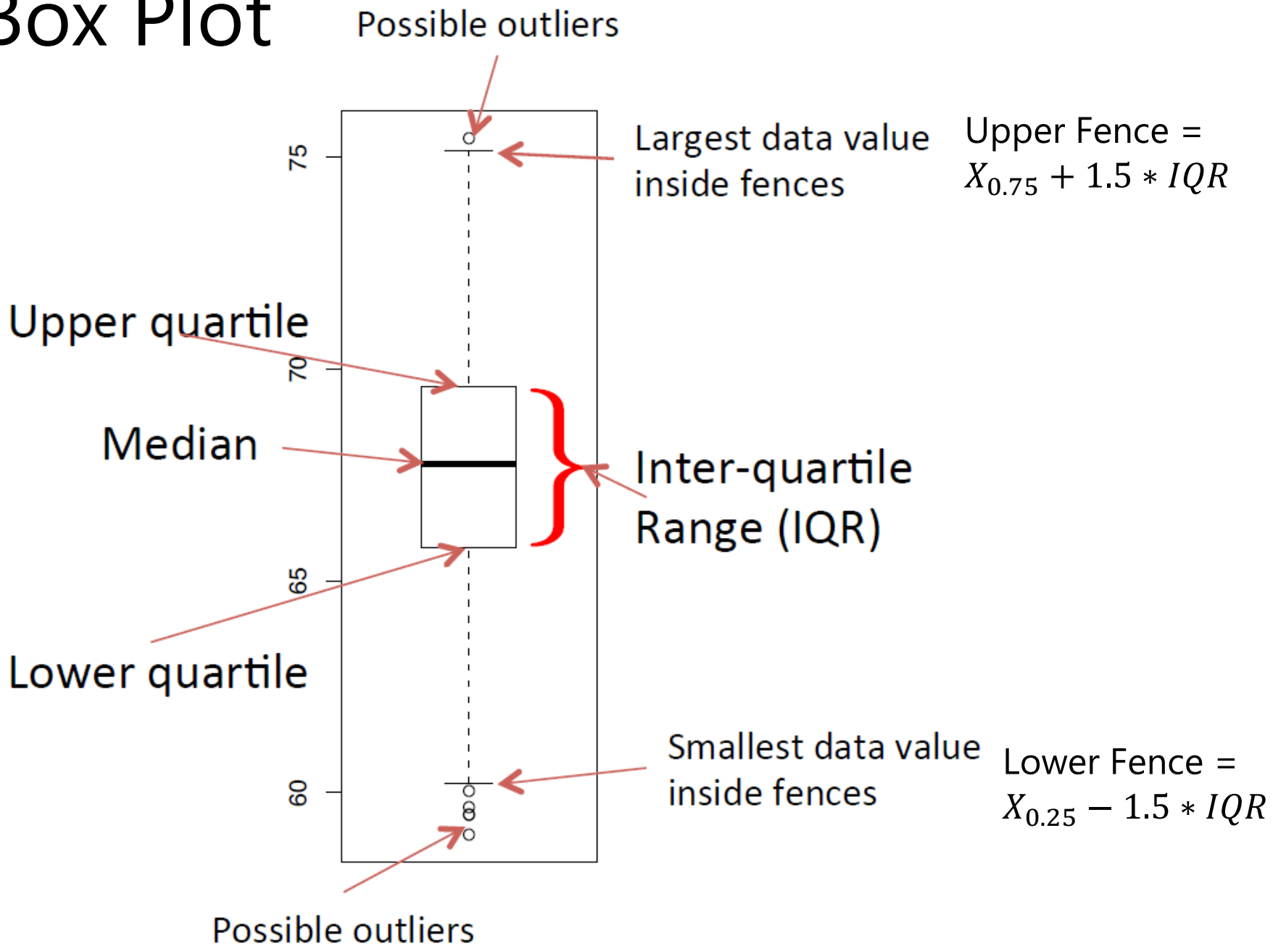
# Outlier detection – Graphical approach

- Boxplot (1-D), Scatter plot (2-D)



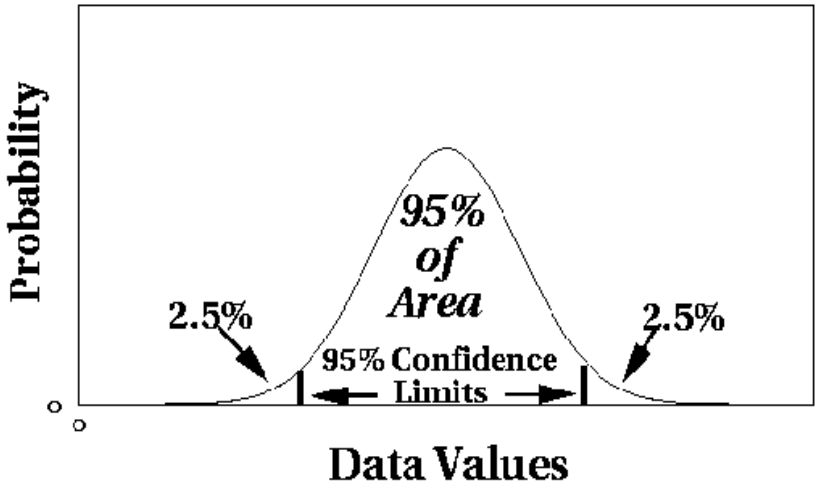



# Box Plot


Possible outliers
Largest data value inside fences
Upper Fence = $X_{0.75} + 1.5 * IQR$
Upper quartile
Median
Inter-quartile Range (IQR)
Lower quartile
Smallest data value inside fences
Lower Fence = $X_{0.25} - 1.5 * IQR$
Possible outliers
75
70
65
60

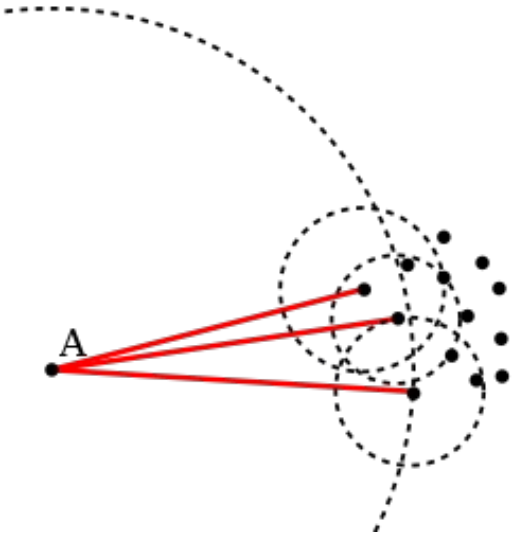
# Outlier detection – Statistical Approach

- Fit a parametric statistical model that defines the "norm", i.e., trend
- Anything outside a determined limit
  - e.g., values that are more than $3\sigma$ (it depends) away from the mean
  - Distinction between *soft & hard* outliers
- Still based on assumptions, e.g., normality

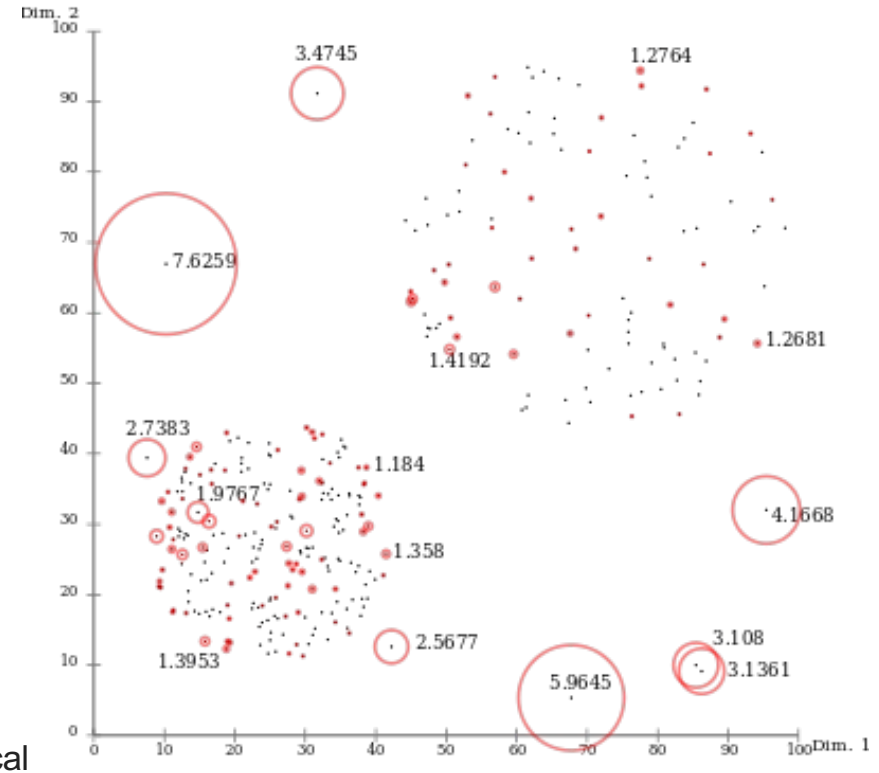


# Outlier Detection – Density Based Approach

- Local outlier factor (Breunig et al. 2000)
- Find outliers by measuring the **local** deviation of a given data point **with respect to its neighbours**




***A*** has lower density compared to neighbours




LOF scores

Breunig, Markus M., et al. "LOF: identifying density-based local outliers."*ACM sigmod record*. Vol. 29. No. 2. ACM, 2000.

# High Dimensional Outliers -- Mahalanobis distance

- Outlier resistant distance function
- Suitable for nD data
- use as an *"outlyingness score"*

$$MD_i = \sqrt{(x_i - \mu)^T C^{-1} (x_i - \mu)}$$

Where we have i rows, $\mu$ is a high-dimensional mean vector, and $C$ is the covariance matrix (quantifying how variables vary together)

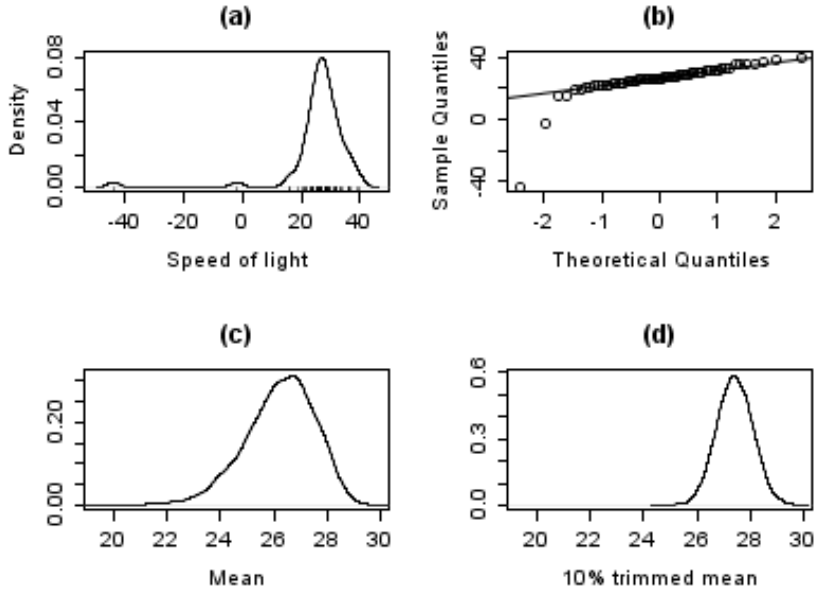
- Values above/below a threshold are considered outliers
  - any i that has an MD value above the 0.975 quartile of a chi-square distribution

# “Living with outliers”: Robust statistics

- Statistics / methods that are resistant against outliers
- No need to remove outliers, in theory
- Focus on finding better statistical estimates
- Can use robust statistics in parametric methods to “robustify” them, e.g., fit a regression line using robust $\mu$ and $\sigma$

# Robust versions of centrality, i.e, mean

- **Median**
- **Midhinge** is the average of the first and third quartiles
- **Trimmed mean**: calculation of the mean after discarding parts of data, e.g., 5 to 25 percent of the ends are discarded (a.k.a. Winsorized mean)



# Robust version of dispersion, i.e., $\sigma$

- Inter-quartile range (*IQR*), Q3 – Q1
- Median absolute deviation

$$\mathrm{MAD} = \frac{1}{N}\sum_{i=1}^{n} |X_i - \mathrm{median}\,|$$

$$\sigma \approx 1.4826\ \mathrm{MAD}$$

**NOTE:** There are also robust versions of analysis methods such as: robust regression, robust covariance estimation, etc. which we don't cover in this module

# Conclusions

- Understanding data distributions is important…
  - Tells us how data columns are distributed
  - Multimodal may indicate a dependency on other characteristics. You may partition your data.
  - Helps us know whether the data means the assumptions of the statistical techniques we want to use
- …for both processing and summarising data
  - for validating and error/bias checking
  - for telling us if the phenomenon is behaving as **expected**
  - interpolation (a form of modelling)

# Conclusions

- Transformations help prepare data for analysis
  - but **note** that is distorts when interpreting
- Effective summarisation is key to analysis (obviously…)
- Outliers
  - Often **tiny** proportions of our data (with the potential to ruin our analysis!)
  - Checking: may indicate problems in data
  - Phenomenon: may be important to know about
  - Analysis: may be problematic
  - Consider removing outlies and using robust statistics