INM430
Principles of Data Science

Week 03
Data Processing & Summarization

Aidan Slingsby, giCentre
CITY
UNIVERSITY OF LONDON
EST 1894

On the menu today

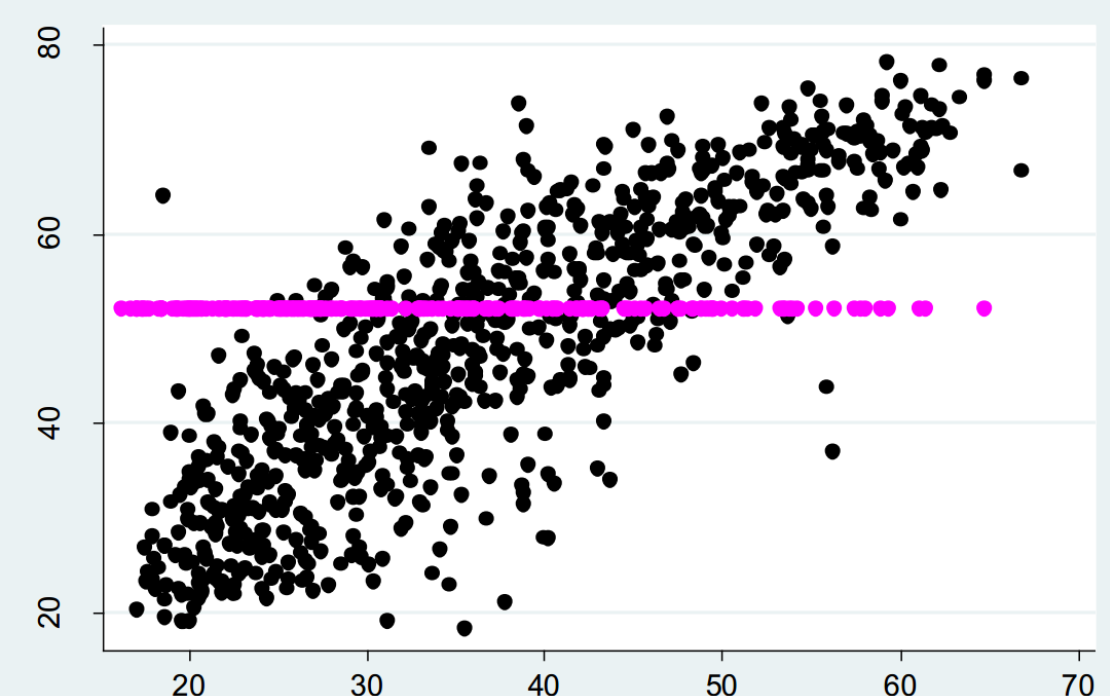
- **Missing data techniques**
- **Processing** data –data transformations
- **Summarizing** data – descriptive statistics
- **Outliers**
- Overview of **robust statistics**

Missing values (recap)

- Try to find out the **nature** of the missing values
  - How many (or what proportion)?
  - Do they relate to other missing values?
  - Is it random or might it introduce bias?
- Then decide what to do (you may try some alternatives if you're not sure). Depends on what you're doing, so **think!**
  - Don't just treat as zero!
  - Remove (whole row or just value)?
  - Impute? Can we come up with reasonable alternatives? Will this help?

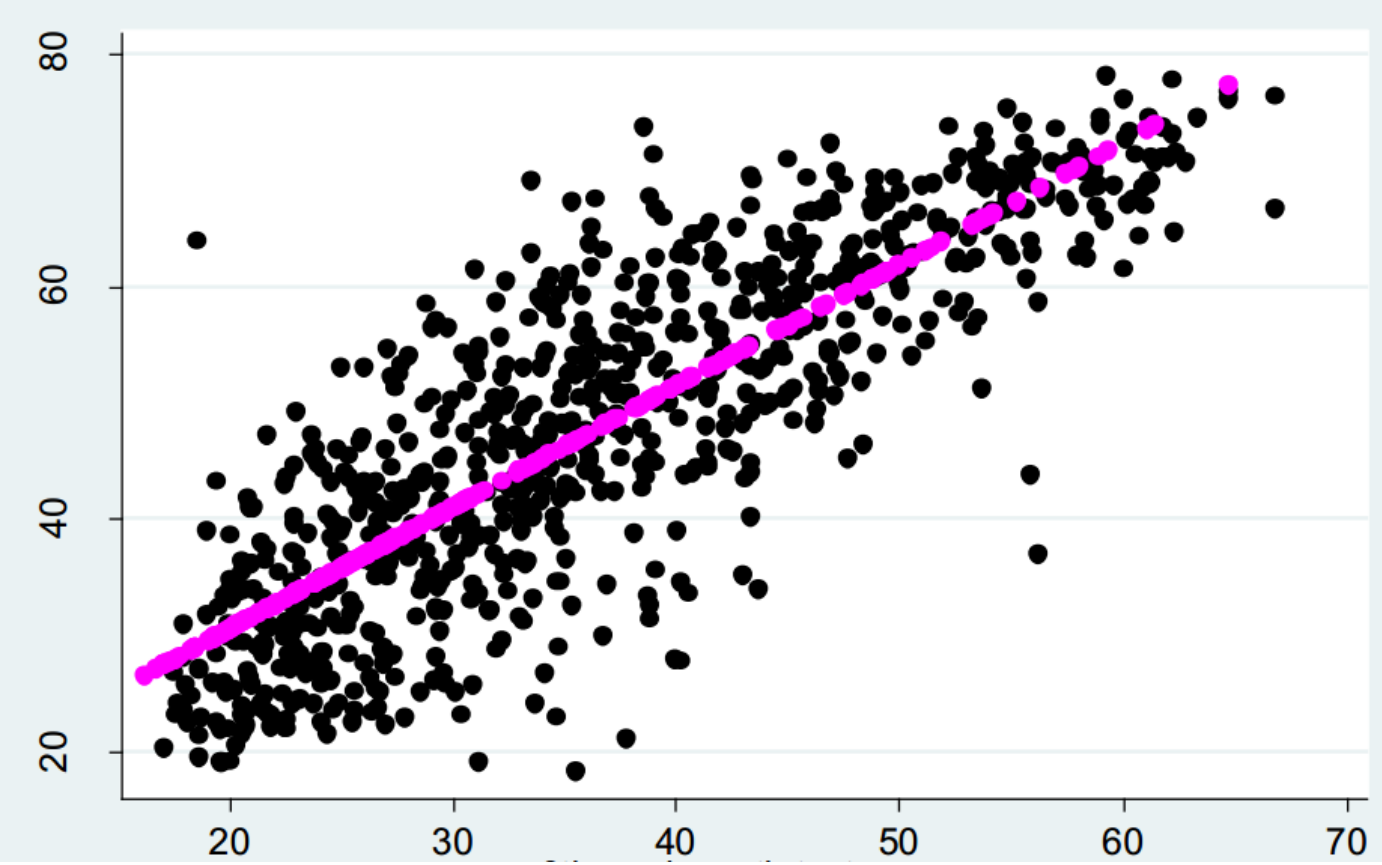
Missing value imputation

- **Mean / mode substitution**
  - Replace missing value with sample mean or mode
  - Reduces variability
  - Weakens covariance and correlation



Missing value imputation

- **Regression substitution (deterministic)**
  - replaces missing values with predictions from a regression function

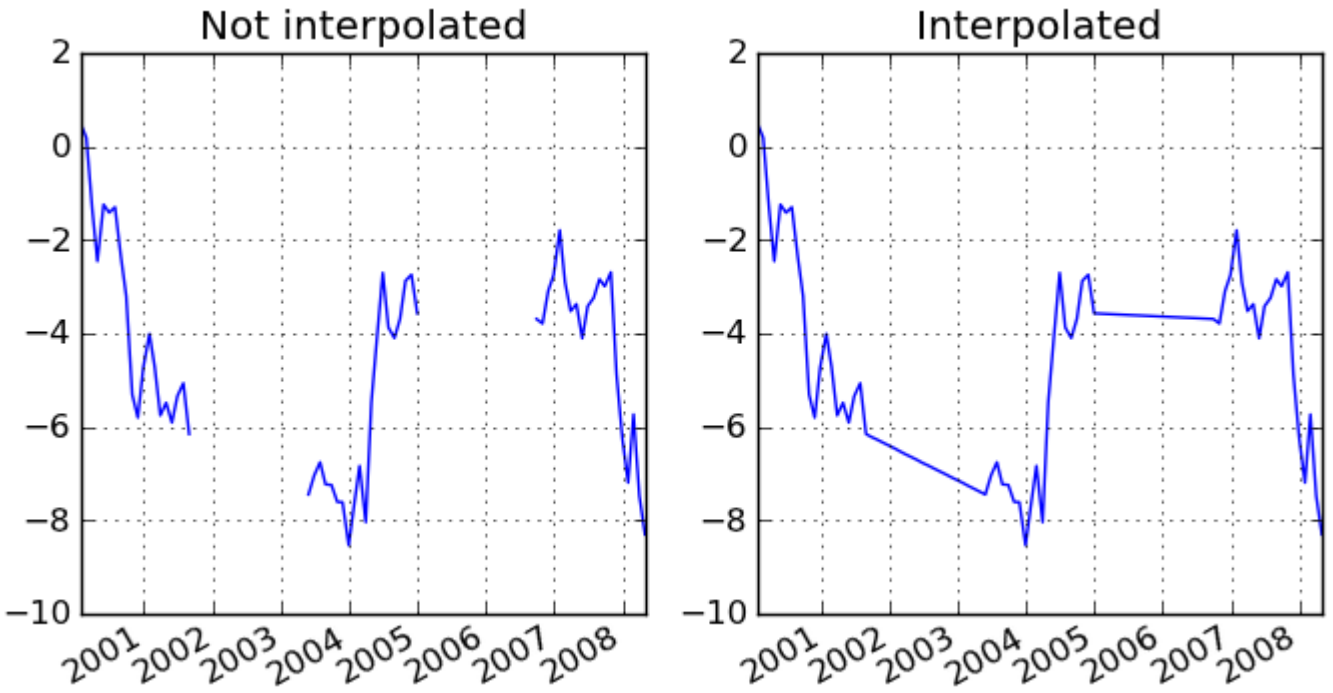


Missing value imputation

- **Random**
  - Randomly, drawn from the sample
- **Local methods**
  - Partition the data by its characteristics, and defined an imputation strategy for each
    - Male/female heights
    - Different regression models for different partitions of data

Missing value imputation

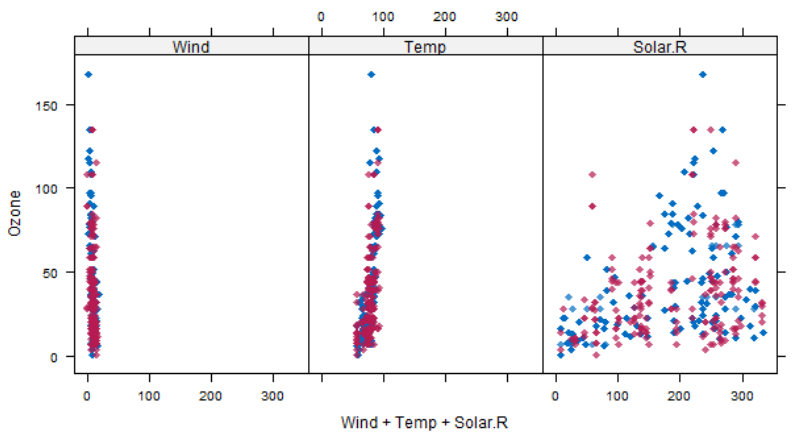
- **Interpolation**
  - Which interpolation method?
  - Use a model of expected values
  - Example, GPS location

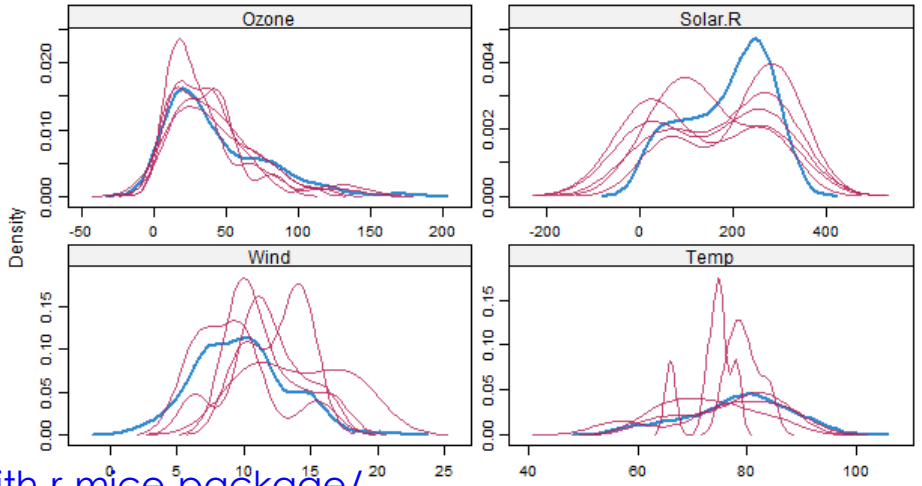


Missing value imputation

- **Multiple Imputation**
  - Rubin, Donald B. Multiple imputation for nonresponse in surveys. Vol. 81. John Wiley & Sons, 2004.
    1. Identify some good impution alternatives
    2. Apply them to come up with multiple alternative datasets and compare the results
    3. Pool: Consider using pooled results

Imputed
Original





http://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/

Missing value imputation

- In summary:
  - Decide whether you **need/want** to impute
  - **Think** about what makes sense
  - Try some **alternatives** if it's not clear and consider the **implications** on your analysis (it might be negligible…)
  - **Keep a record** of what yoy do and why. **Analytical provenance** is important (In Python, you might store imputed versions as separate columns. Computational notebook can help leave a provenance trail).

Data transformation

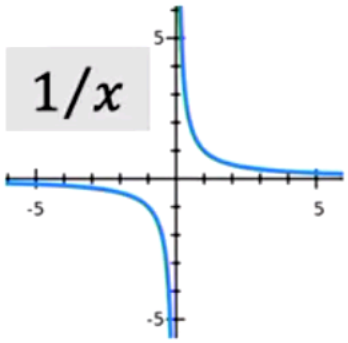
- Transforming data variables so they are suitable for analysis
  - Depends on what analysis you are doing, many make assumptions about data
- Other names:
  - Normalisation. scaling, standardisation
- Types of transformation
  - Rescale
  - Change the distribution (express in a different distribution)

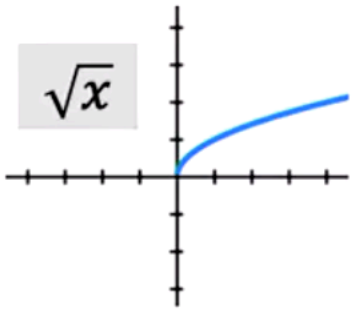
Data transformation: Rescaling/normalising

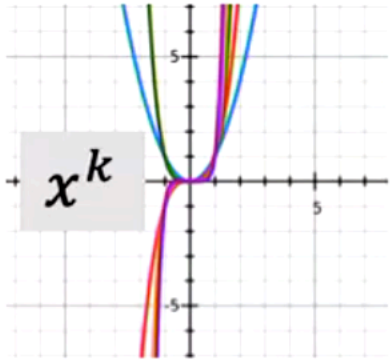
- Remove the effects of scale
- To make variables comparable – bring things into a common scale
  - Z-score standardisation: $\frac{X - \mu}{\sigma}$
  - Range normalise e.g. (0-1): $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$
  - (or use different range 5-95th percentile)
  - Scaling to unit length: $x' = \frac{x}{||x||}$

Data transformation: Change distribution

- Mathematical operators applies to single values
- Change the shape/distribution of the data
  - E.g. Give more weight to low/high values
    - x^k
    - $\log x$
    - e^x
    - $\sqrt{x}$
    - $1/x$
    - $\sin x$
    - $|x|$





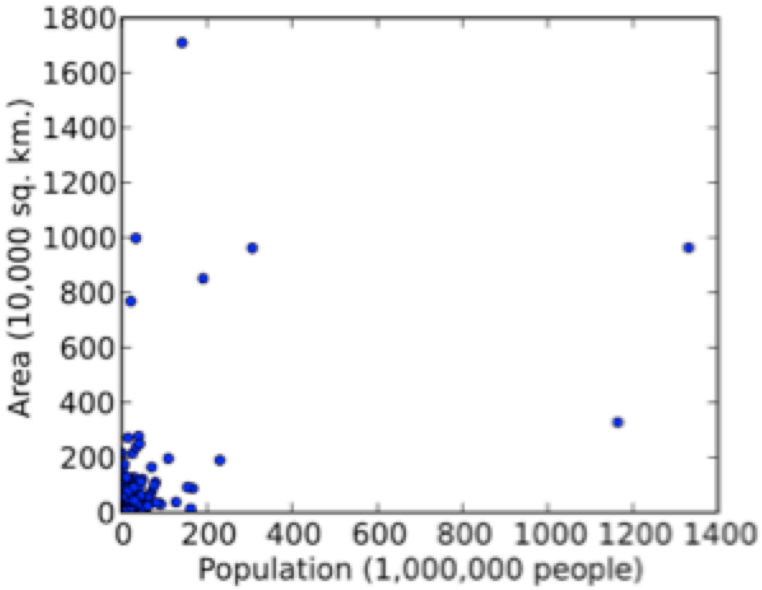


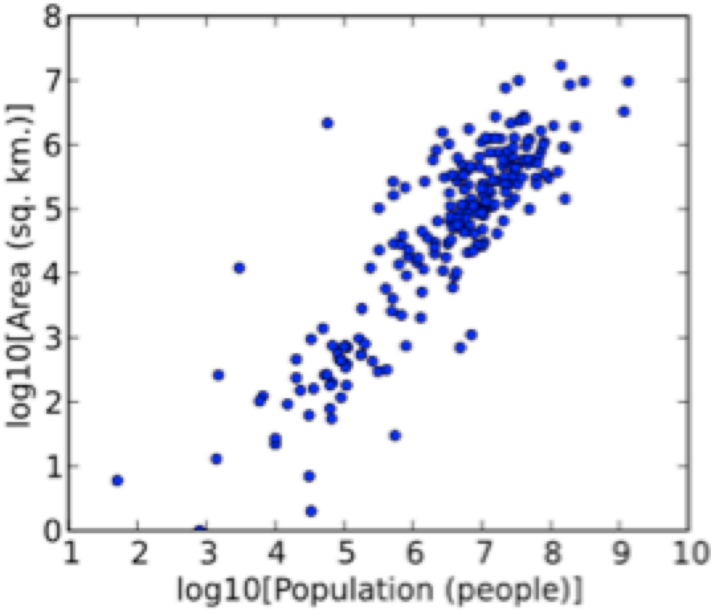
http://goo.gl/AzmUo0

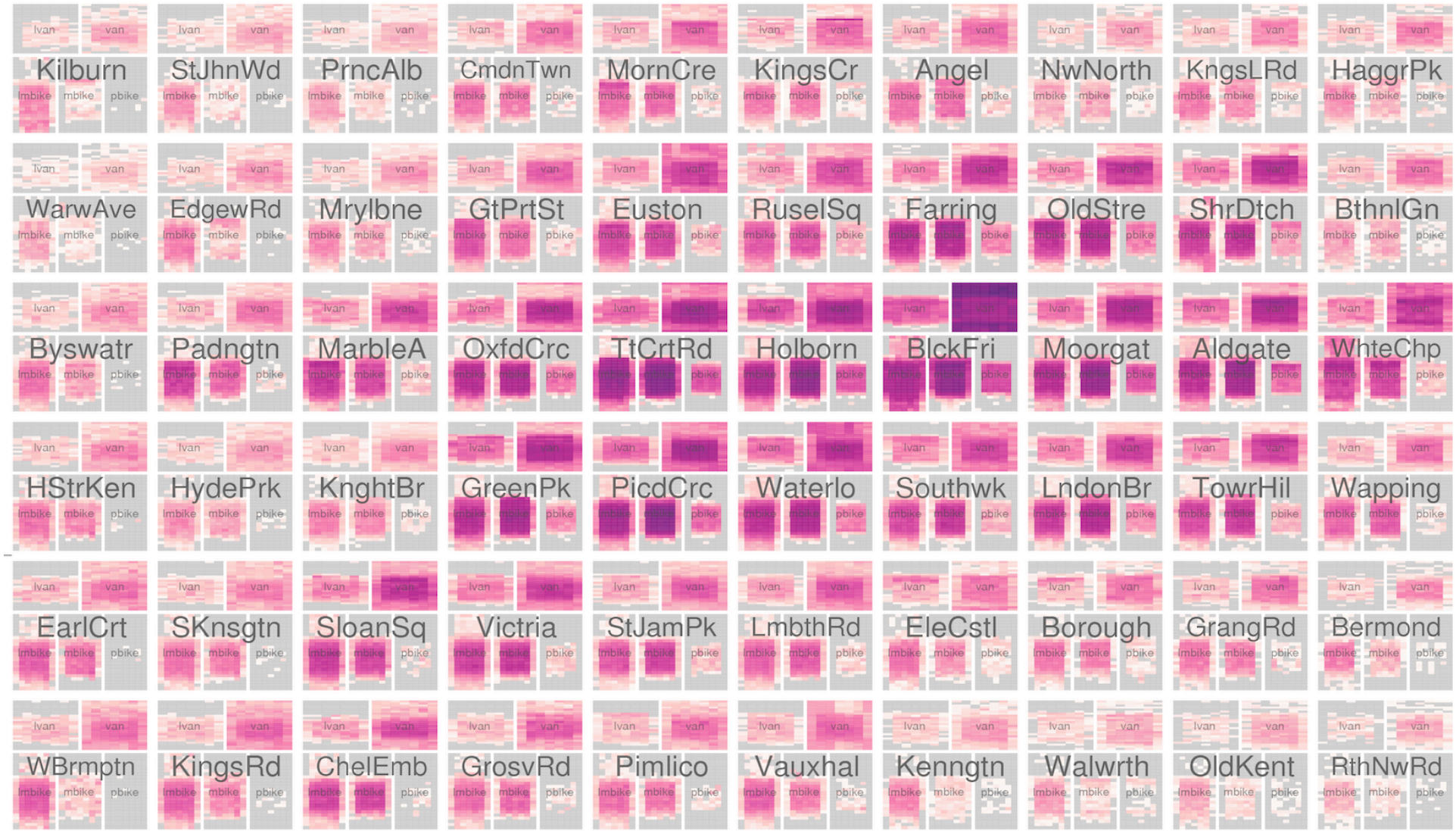
http://goo.gl/AzmUo0

Data transformation: Log transform

- A very common transformation - we find logs all over the place!
  - https://en.wikipedia.org/wiki/Logarithm
- Take the *log* of each observation
- Spreads out small values, “curing” outliers
- **Distorts** the data



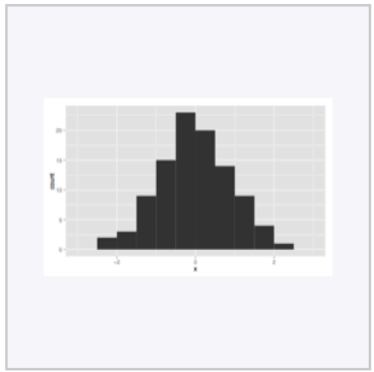



lvan van
Kilburn StJhnWd PrncAlb CmdnTwn MornCre KingsCr Angel NwNorth KngsLRd HaggrPk
WarwAve EdgewRd Mrylbne GtPrtSt Euston RuselSq Farring OldStre ShrDtch BthnlGn
Byswatr Padngtn MarbleA OxfdCrc TtCrtRd Holborn BlckFri Moorgat Aldgate WhteChp
HStrKen HydePrk KnghtBr GreenPk PicdCrc Waterlo Southwk LndonBr TowrHil Wapping
EarlCrt SKnsgtn SloanSq Victria StJamPk LmbthRd EleCstl Borough GrangRd Bermond
WBrmptn KingsRd ChelEmb GrosvRd Pimlico Vauxhal Kenngtn Walwrth OldKent RthNwRd
lmbike mbike pbike

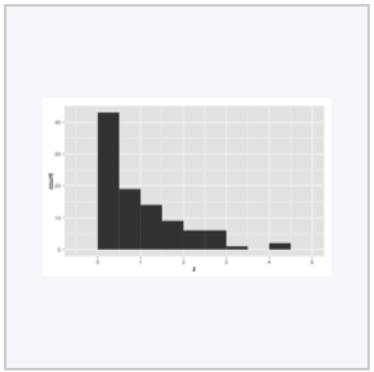
Data transformation: discretization (binning)

- Transform a continuous attribute into a categorical attribute
  - Histograms do this
- How to define bins (number/size)?
  - Histograms: usually equal width
  - May want to partition data on quantiles
  - May define bin boundaries based on gaps in data
  - May define bin boundaries based on prior/domain knowledge

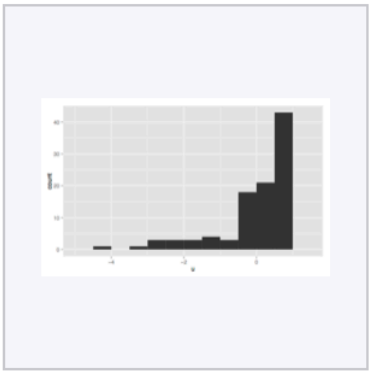
Histograms



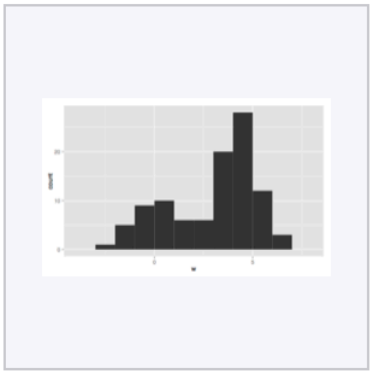
Symmetric, unimodal



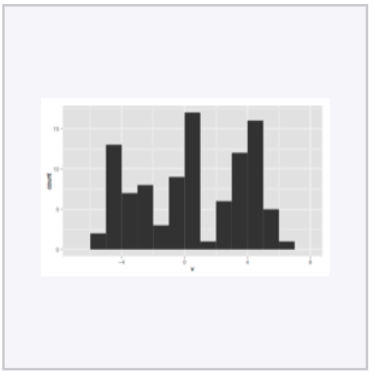
Skewed right



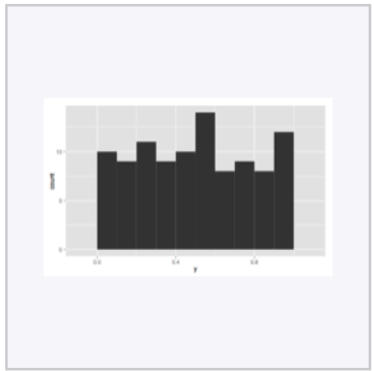
Skewed left



Bimodal

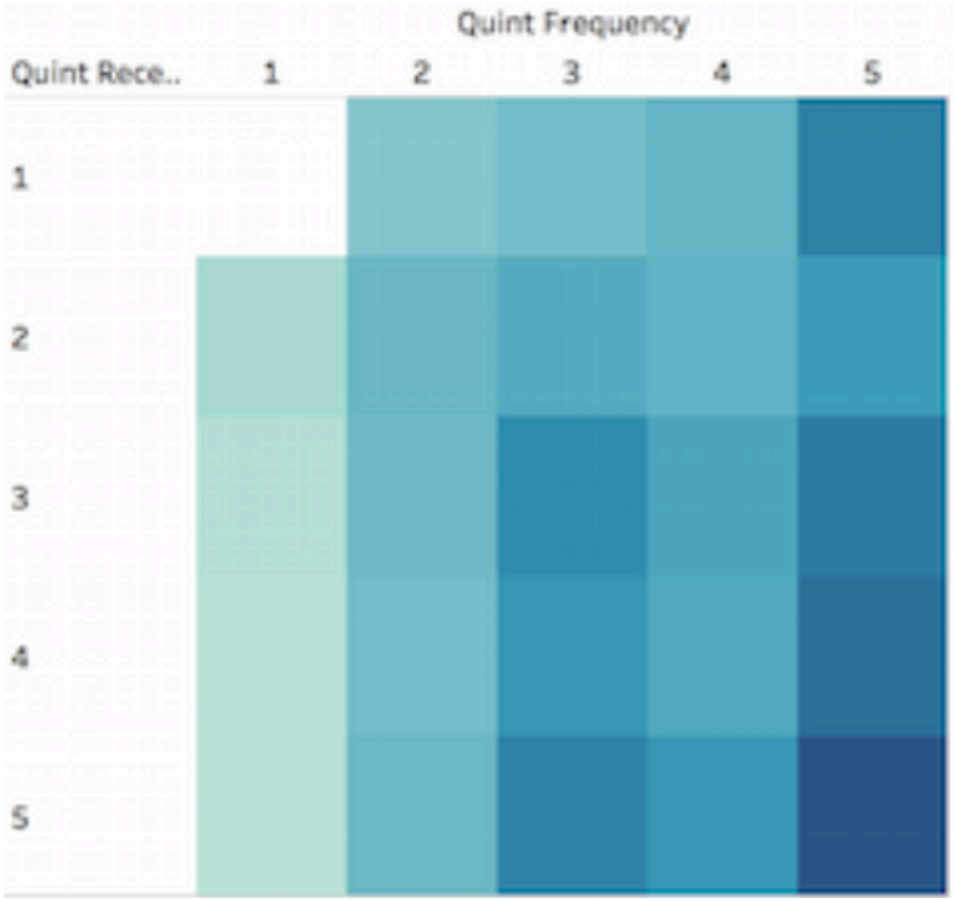


Multimodal

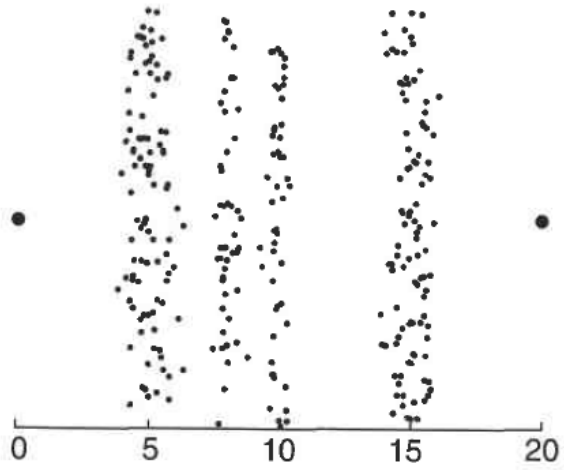


https://en.wikipedia.org/wiki/Histogram

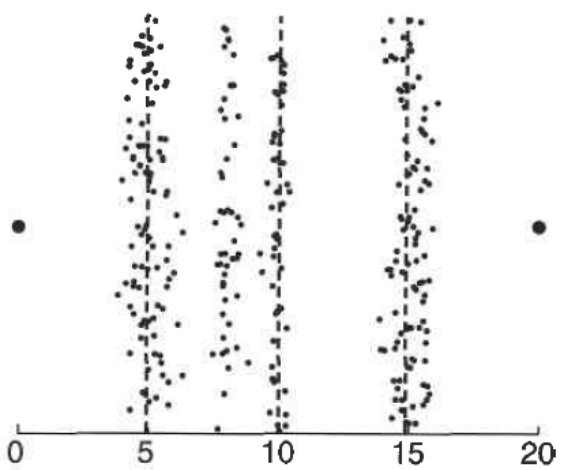
Recency-frequency



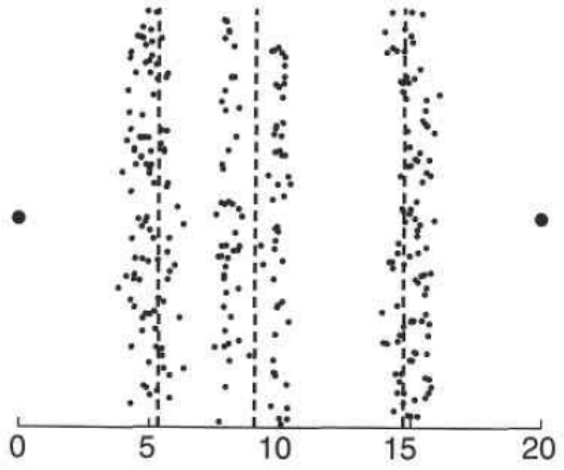
Binning strategies



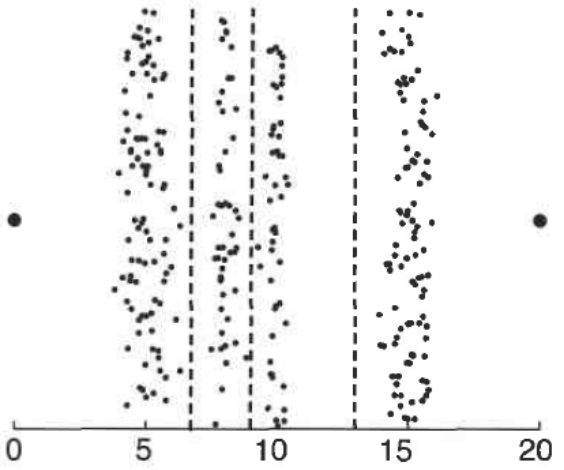
(a) Original data.



(b) Equal width discretization.

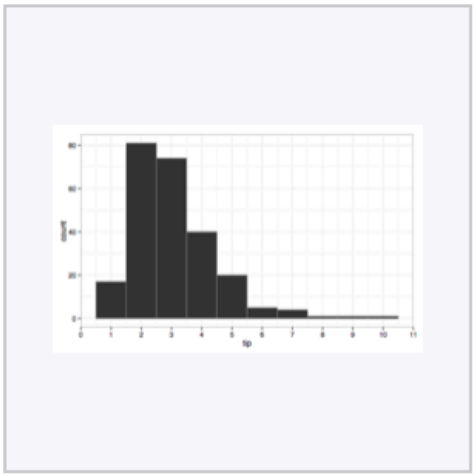


(c) Equal frequency discretization.

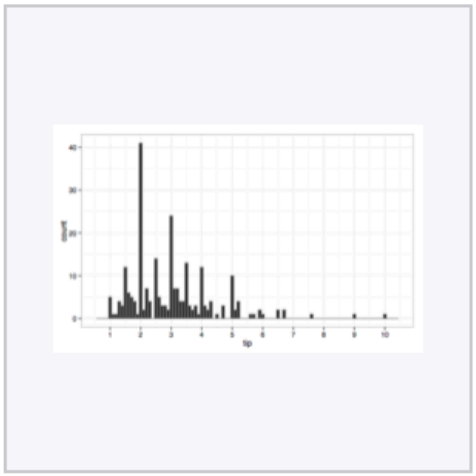


(d) K-means discretization.

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

Good binning sizes?

- Try different bin-sizes to see what works
- Some suggestions: $k = \lceil \log_2 n + 1 \rceil$

  depends on (log of) sample size

  – Sturges' formula:

  $$h = \frac{3.5\hat{\sigma}}{n^{1/3}}$$

  depends on standard deviation as well as sample size

  – Scott's rule

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/)

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

- **Descriptive statistics**
  - quantitatively describe the main features of data
  - distinguished from inferential statistics
    - **descriptive statistics:** summarize a sample
    - **inferential statistics:** learn about the population that the sample of data is thought to represent
- Later, we'll be looking at how we can infer population data from a population sample
- Population sampling
  - http://onlinestatbook.com/stat_sim/sampling_dist/index.html
- Central Limit Theory

Inferential statistics

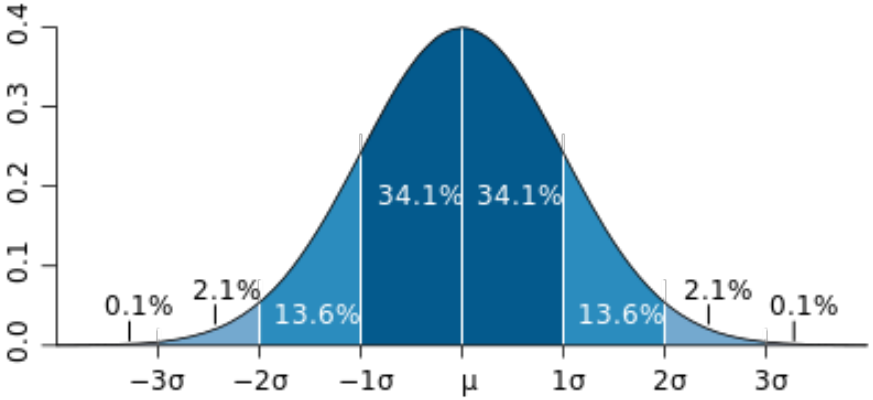
- **Statistical inference** is the process of deducing properties of an underlying distribution by analysis of data. Inferential statistical analysis infers properties about a population. The population is assumed to be larger than the observed data set; in other words, the **observed data is assumed to be sampled from a larger population**
- The idea is to compute statistics on the given sample and make inferences about the population through mathematical methods – to estimate the **unobserved** given the **observed**, so the more you observe, the better.
- The goal is to estimate parameters that define the underlying population model and test for **hypotheses** about your sample vs. population (more later)

Interactive demo on sampling distribution:
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

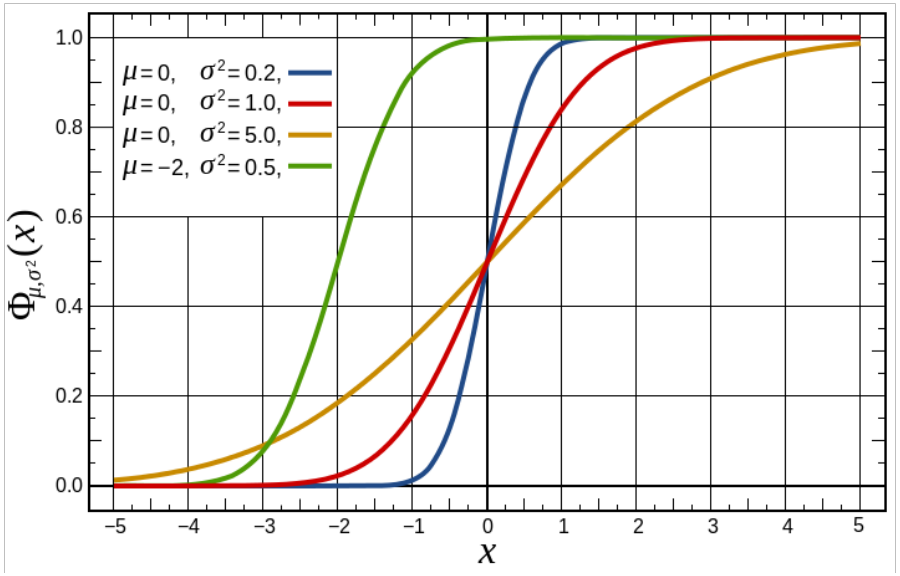
Descriptive statistics

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

Cumulative & probability distribution functions



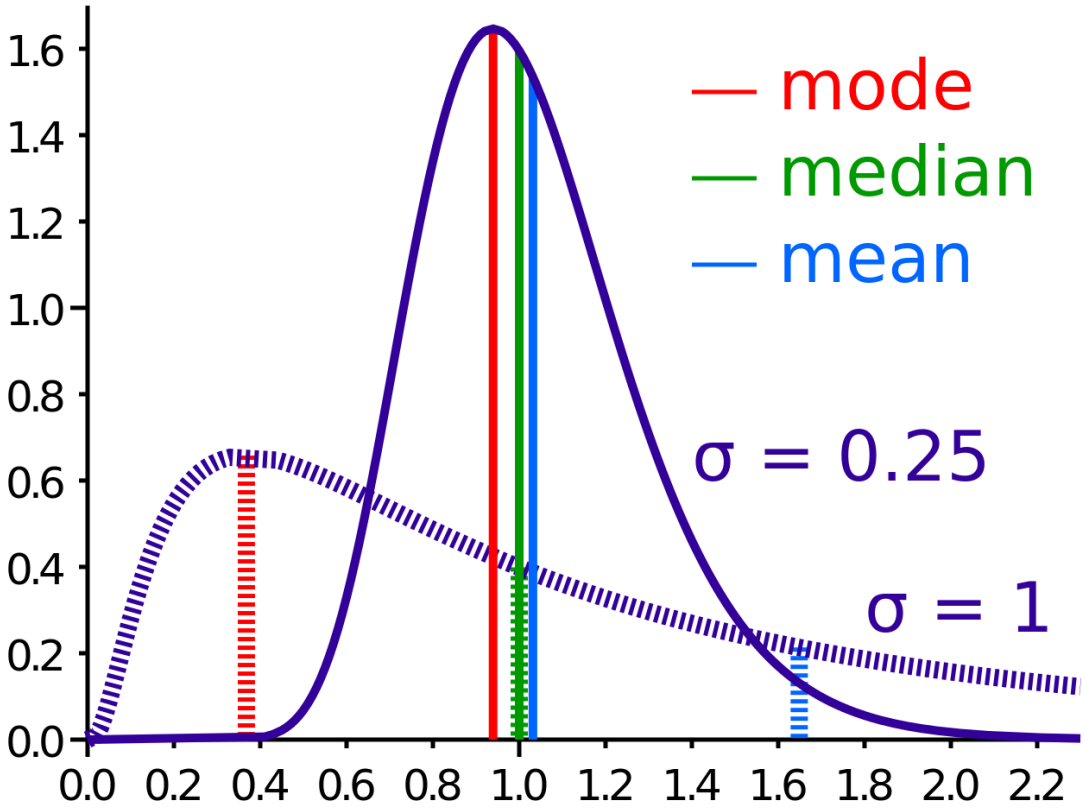
Probability distribution
for normal dist.
(links each outcome of a statistical experiment with its **probability of occurrence**.)



Cumulative distribution
for normal dist.
(..when evaluated at x, is the probability that *X* will take a value less than or equal to x)

Centrality: mean/average, median, mode

- Mean
  - centre of the distribution
- Median
  - middle value
  - “robust”
- Mode
  - most frequent value
  - More suitable for **categorical data**

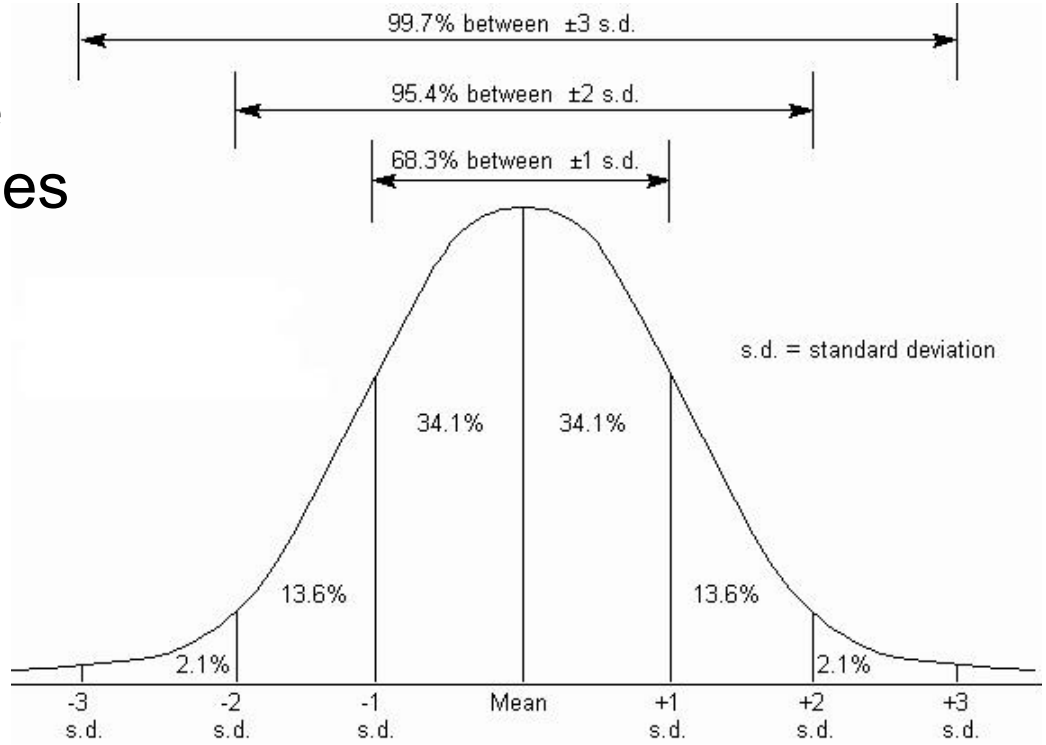


Spread: variance, standard deviation, IQR

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

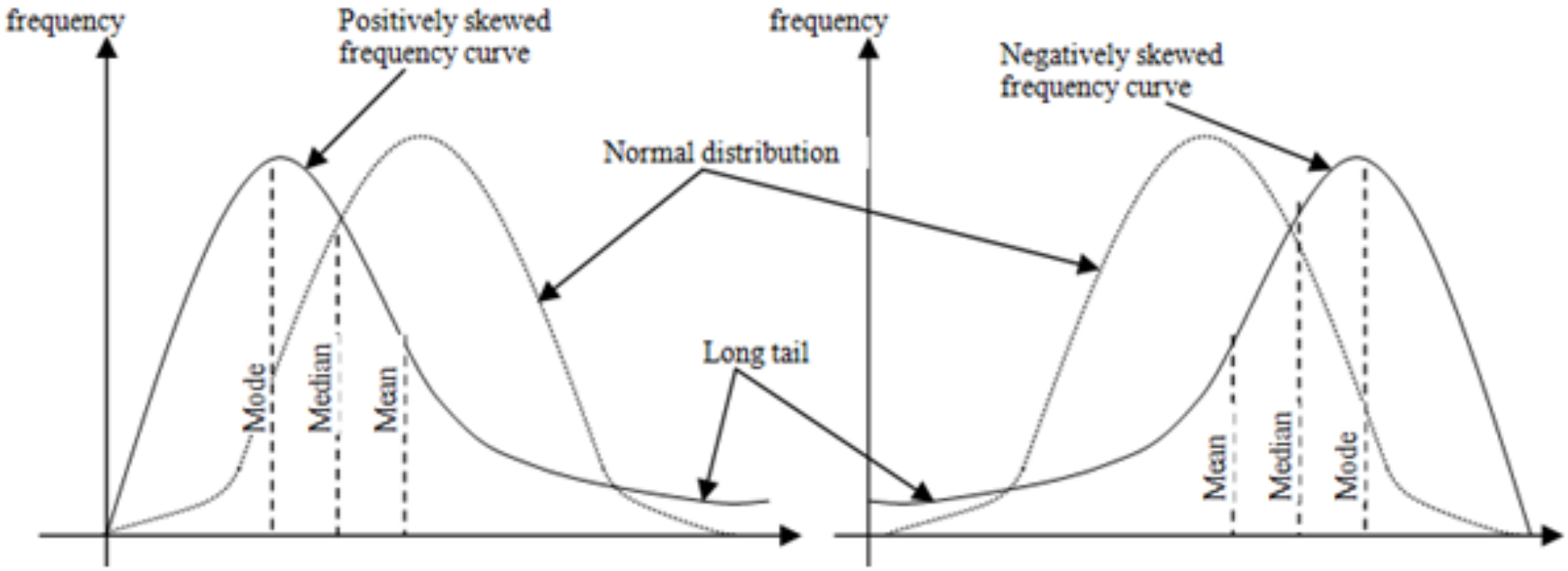
$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

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



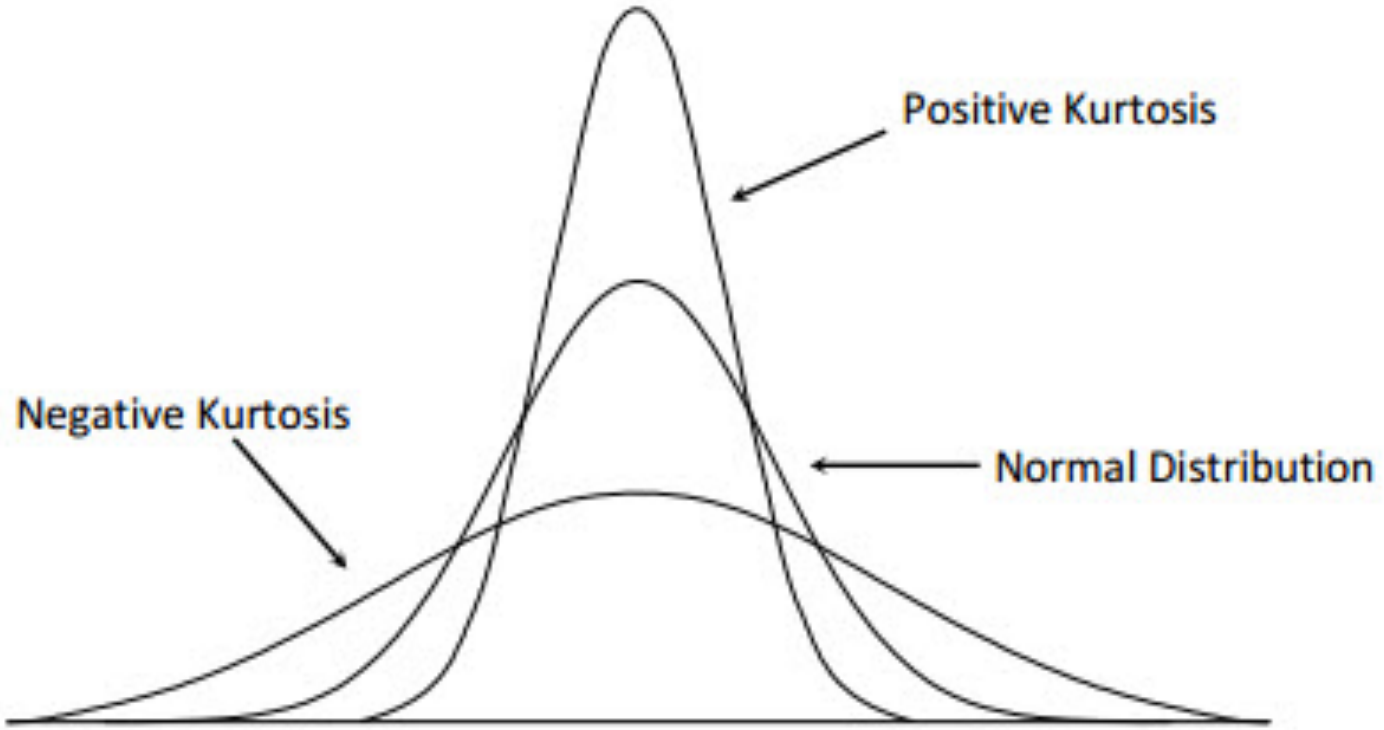
Skewness

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



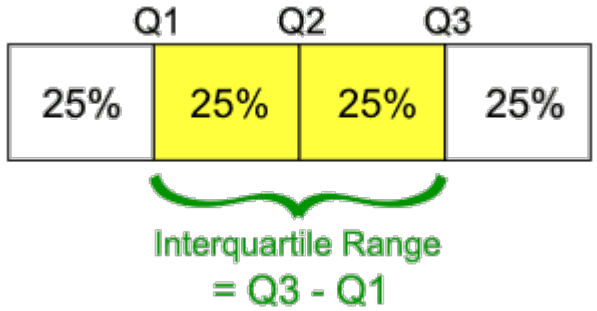
Kurtosis

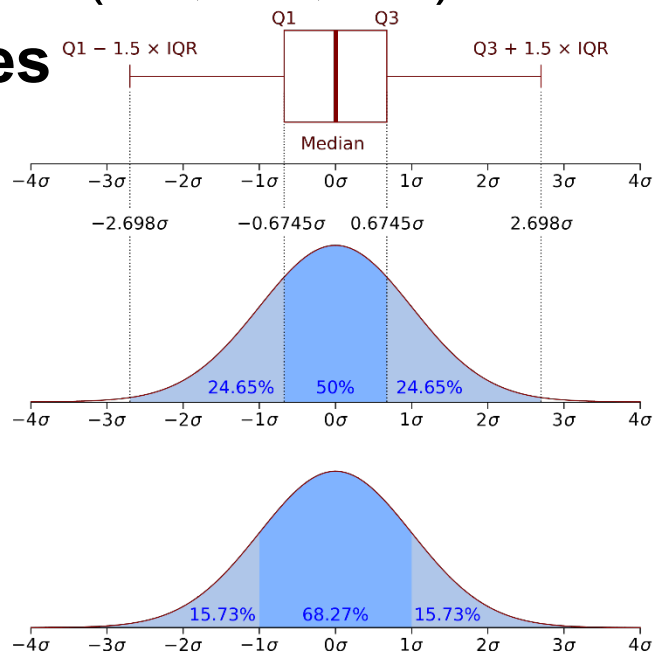
- Measure of the "peakedness“
- Unstable statistics, need ~ 500 samples at least



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
  - Useful to find outliers





Summaries for categorical data

- Frequency/counts
  - how frequent is one category
- Proportions
  - Normalised by total
- Mode
  - Use mode as the centre of the data

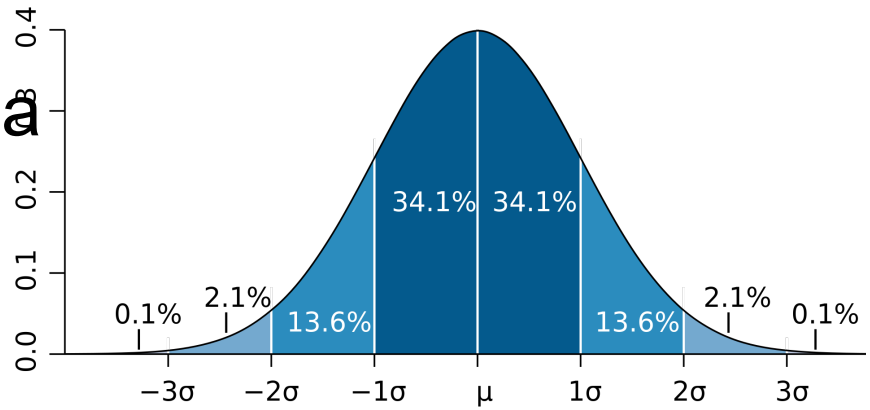
	Undergrad	Graduate	Staff
Counts	17	1	2
Proportions	0.85	0.05	0.1

Some important distributions,

There are many, see:
http://en.wikipedia.org/wiki/List_of_probability_distributions

Normal (Gaussian) distribution

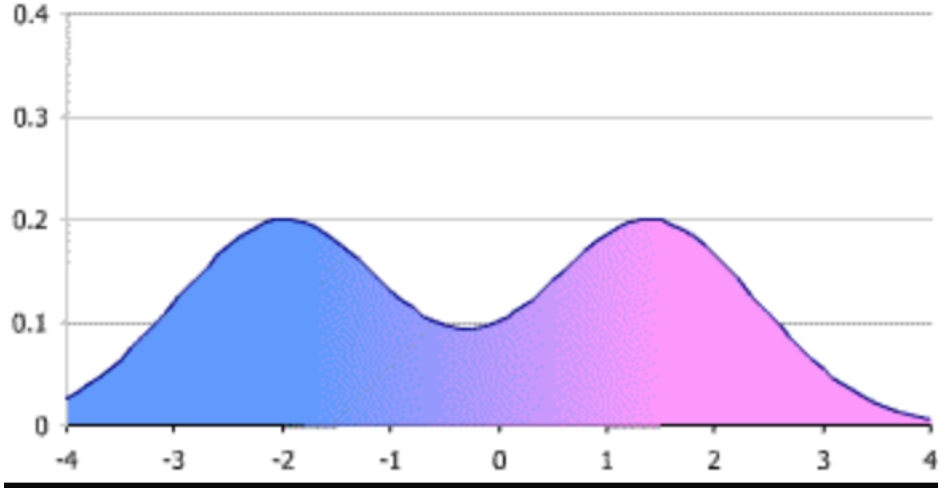
- **Continuous** distribution followed by many phenomena
  - including sample means (Central Limit Theory)



- 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

Biomodal distribution

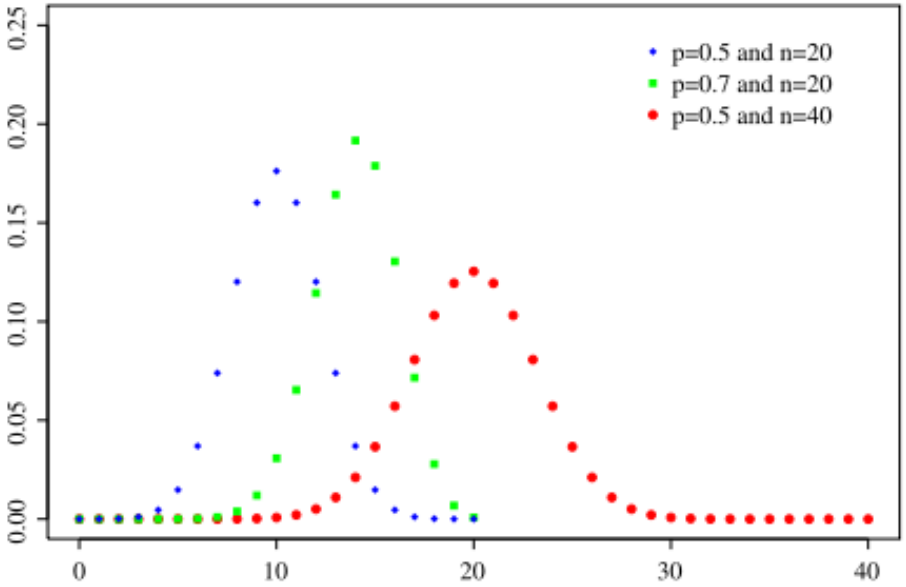
- Two modes
- May indicate that there are two cohorts with different properties
  - …so you might consider finding out how to distinguish them, and treat differently



- No (single) central tendency, so mean/median etc may be misleading
- Example
  - Male and female heights… but a myth because the modes are not so different that it approximates normal.

Binomial distribution

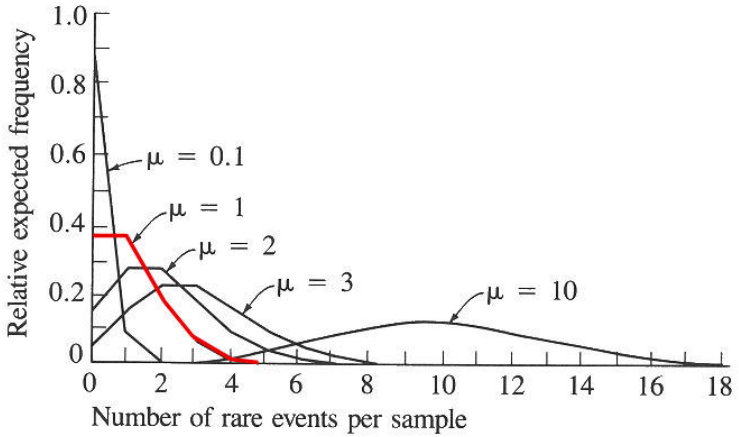
- **Discrete** probability distribution
  - Number of successes in sequence of independent yes/no experiments
  - Looks ‘normal’ if there are enough samples
  - Can be used to predict how many success outcomes



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

Poisson Distribution

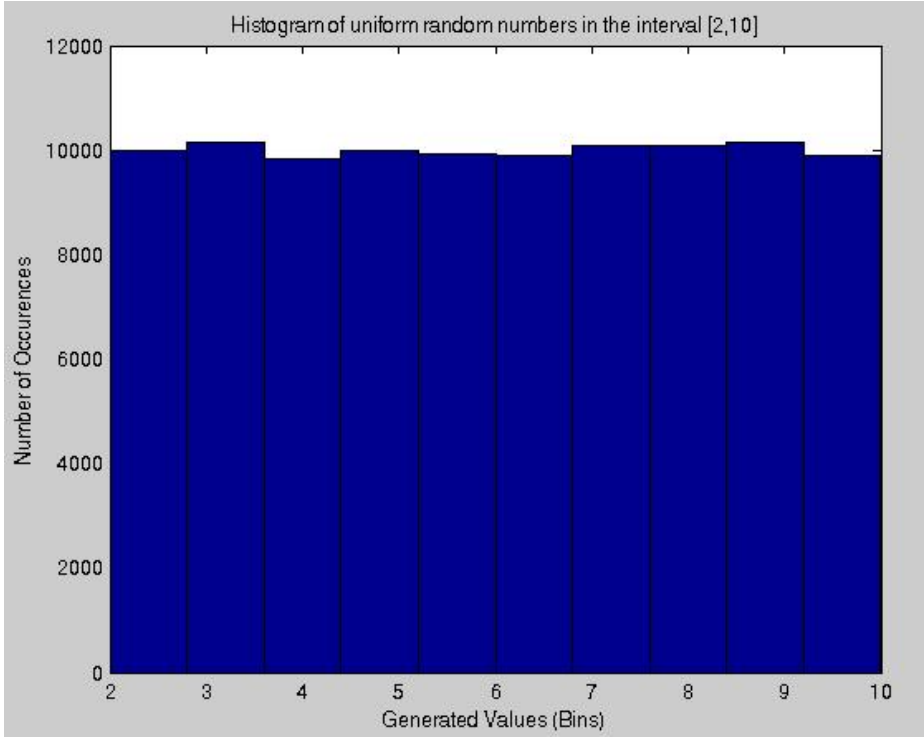
- Another **discrete** probability distribution
  - Where **random "events"** occur at a certain **rate** over a **period of time**



- 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:
  - Number of occurrences within time frame

Uniform distribution

- **Symmetric** probability distribution in which all values are equally probable



Why to know about distributions?

- Data in data columns are distributed
- The shape of a distribution can inform us about the phenomenon
- Can help us choose methods
  - e.g., linear regression assumes normality
- We can use such a distributions to build basic models of data
  - …and use it 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

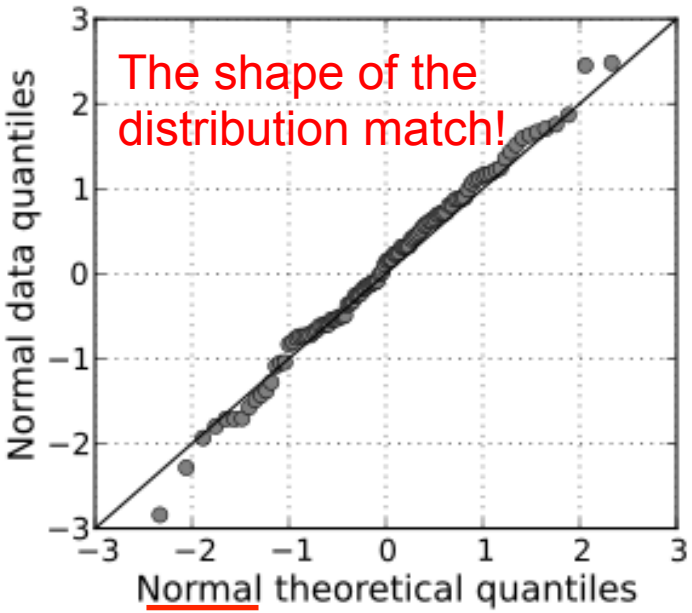
Some common assumptions in statistics

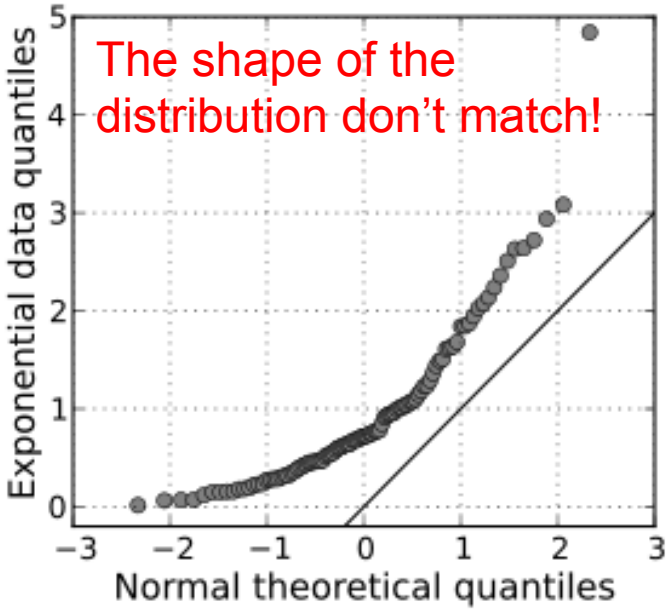
- Depends on the parametric statistics/methods
- Normality
  - Data have a normal distribution (or at least is symmetrical)
- Homogeneity of variances
  - Data from multiple groups have the same variance (Levene's test to check)
- Linearity
  - Data have a linear relationship
- Independence
  - Data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

Assumptions do not always hold,
Question **the reliability of statistics**

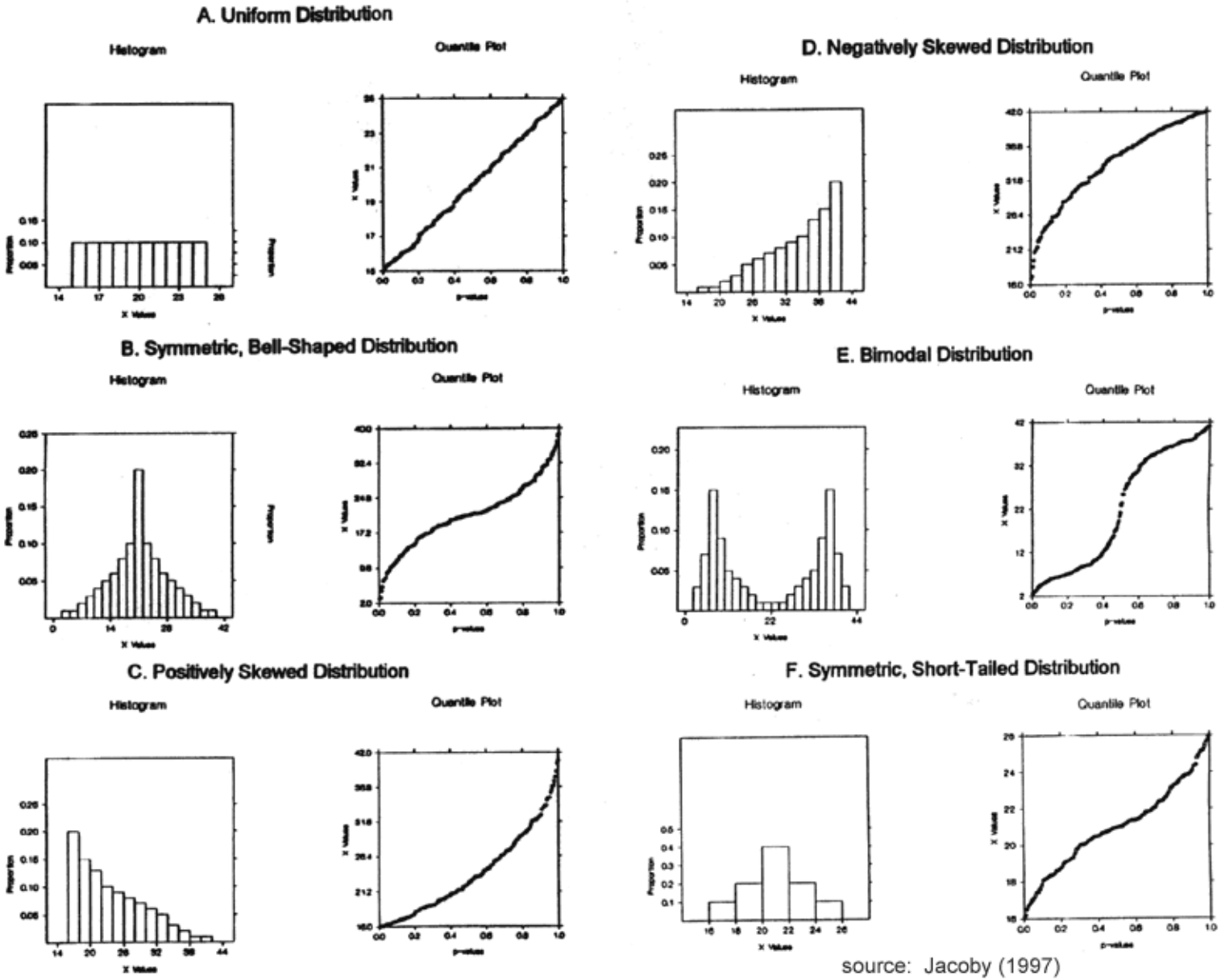
Checking for normality: Normal Q-Q plots

- 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



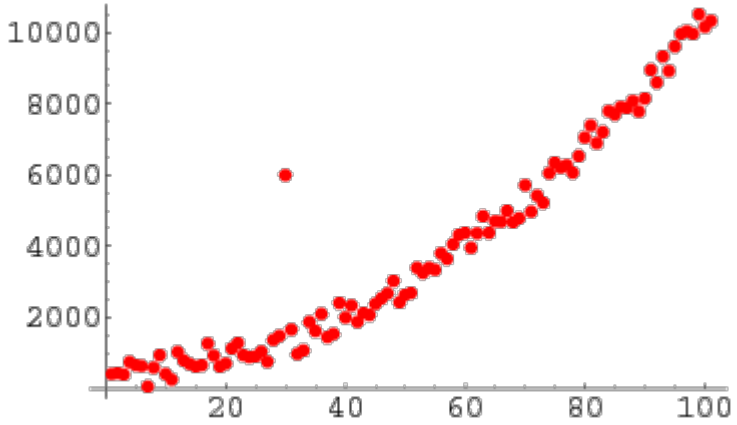
source: Jacoby (1997)

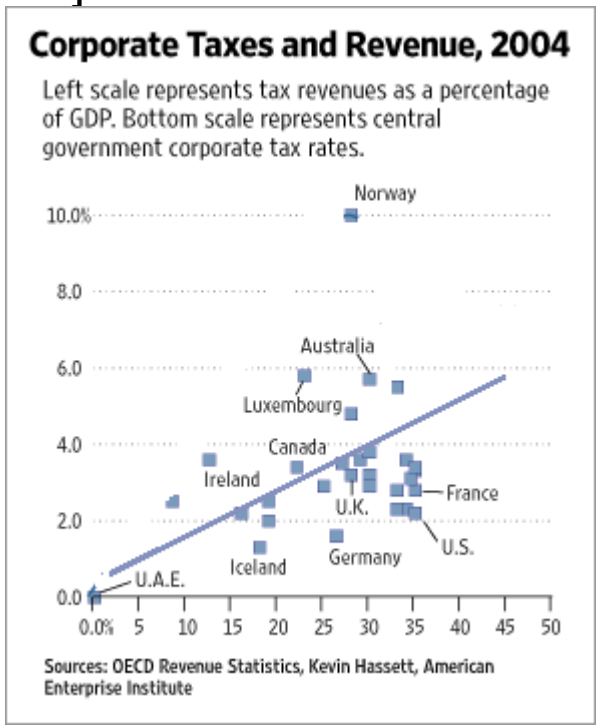
Histograms

- Personally, I prefer histograms!
  - Superimpose a theoretical distribution

Outliers (vs. trends)

- "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]



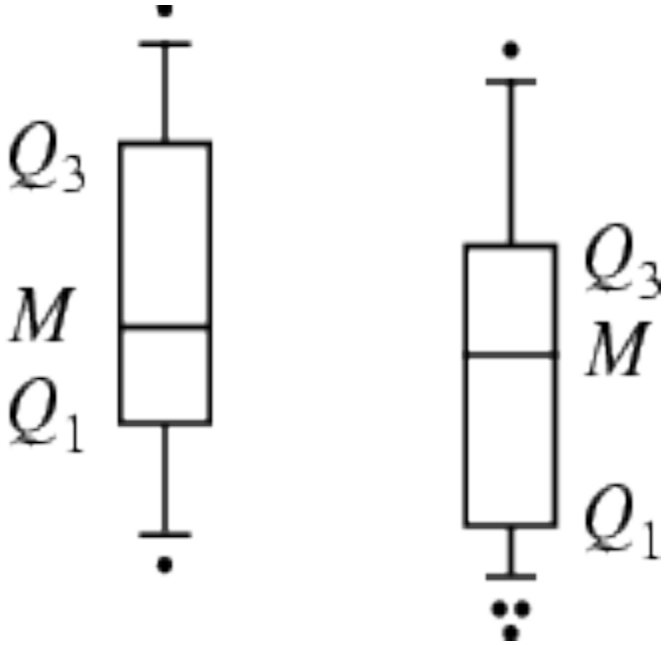


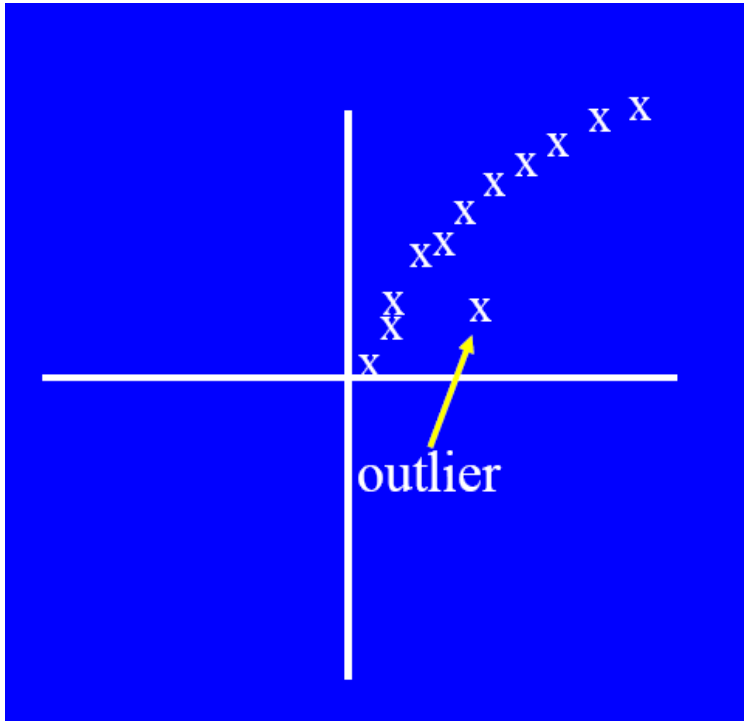
Outliers – friend or foe?

- 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

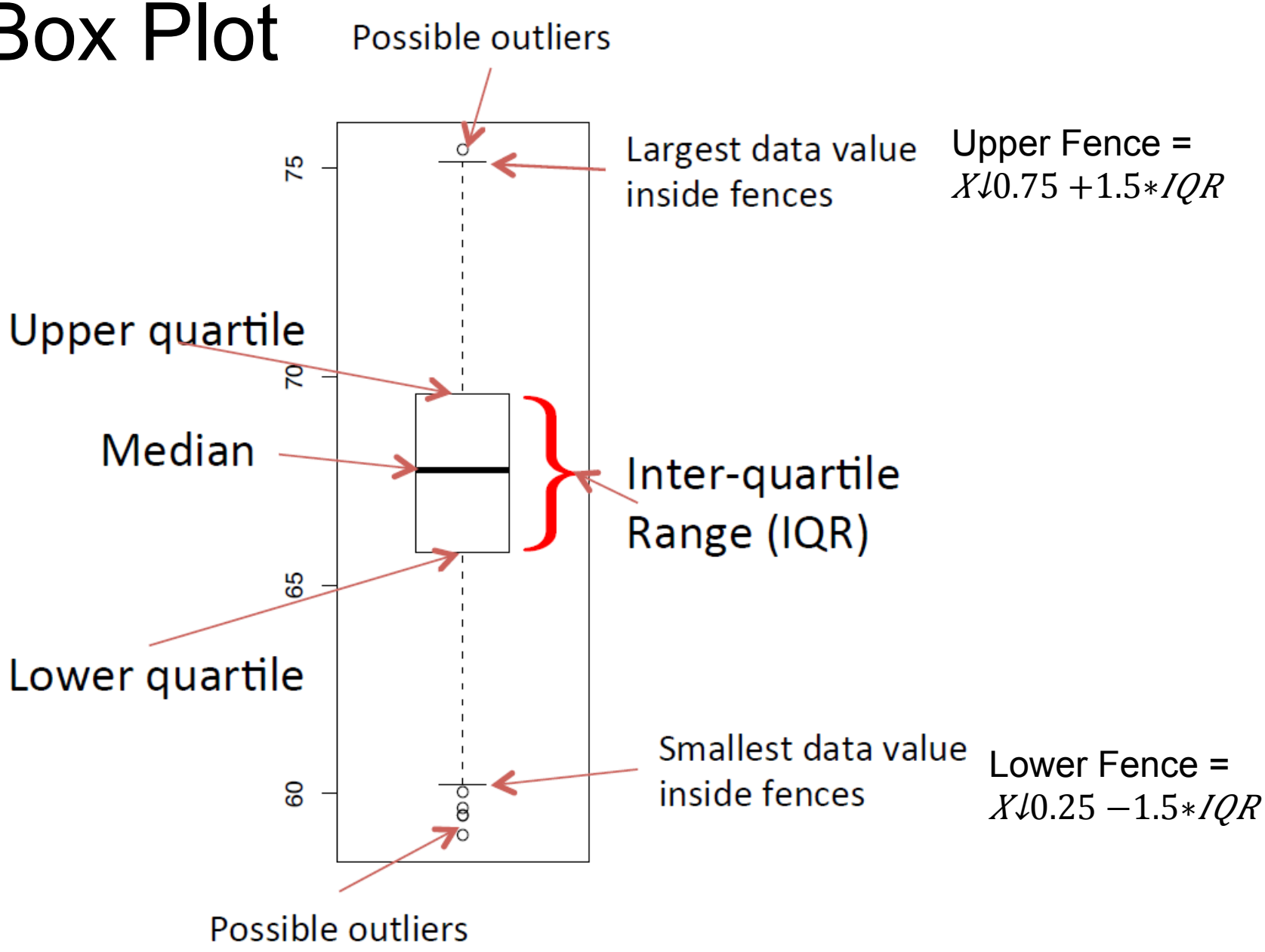
- Boxplot (1-D), Scatter plot (2-D)





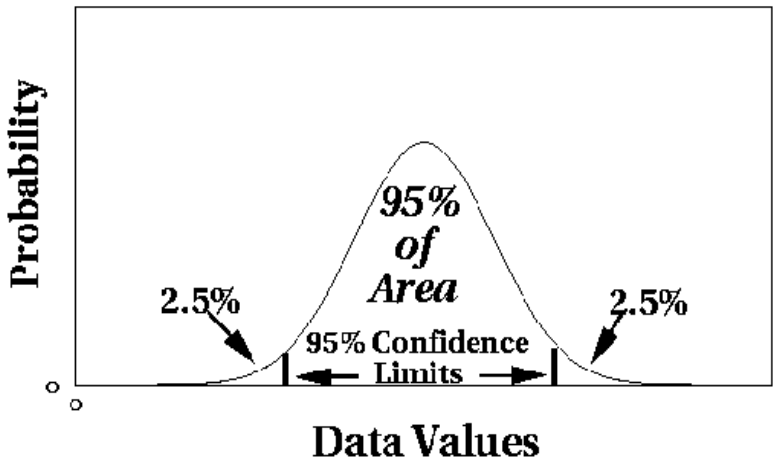
slide material by Tan et al. , 20

Box Plot


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

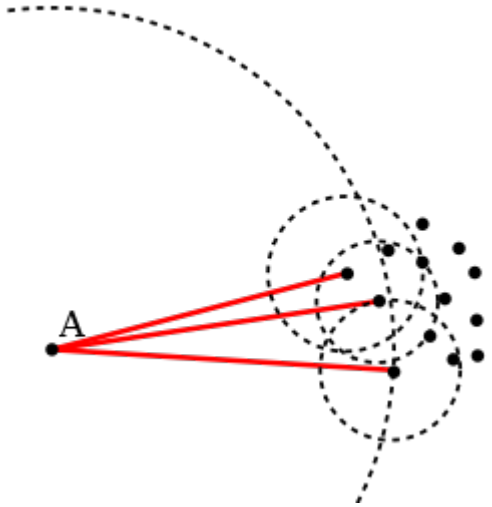
Outlier detection – Statistical Approach

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

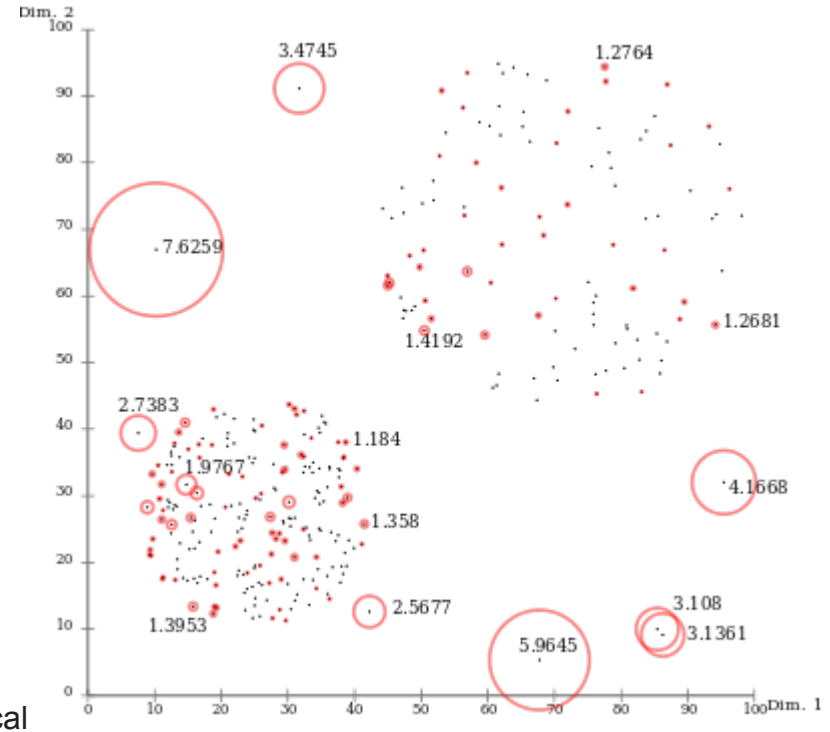


Outlier Detection – Density Based Approach

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



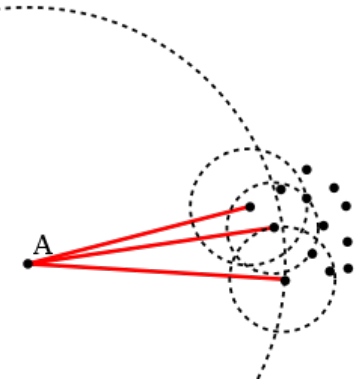
A has lower density compared to neighbours



LOF scores

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

Local outlier factor - explained



1

$$\text{k-distance}(A)$$:the distance of the object A to the k-th nearest neighbour.

2

$$\text{reachability-distance}_k(A, B) = \max\{\text{k-distance}(B), d(A, B)\}$$

the true distance of the two objects, but at least the *k-distance(B)* – for stability

3

$$\text{lrd}(A) := 1/\left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|}\right)$$

local reachability density (lrd): is the inverse of the average reachability distance of the object A from its neighbors.

4

$$\text{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\text{lrd}(B)}{\text{lrd}(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \text{lrd}(B)}{|N_k(A)|} / \text{lrd}(A)$$

LOF: *average local reachability density of the neighbours* divided by the object's own local reachability density. A value around 1 means A is similar to its neighbours

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, μ is a high-dimensional mean vector, and C is the covariance matrix (quantifying how variables vary together)

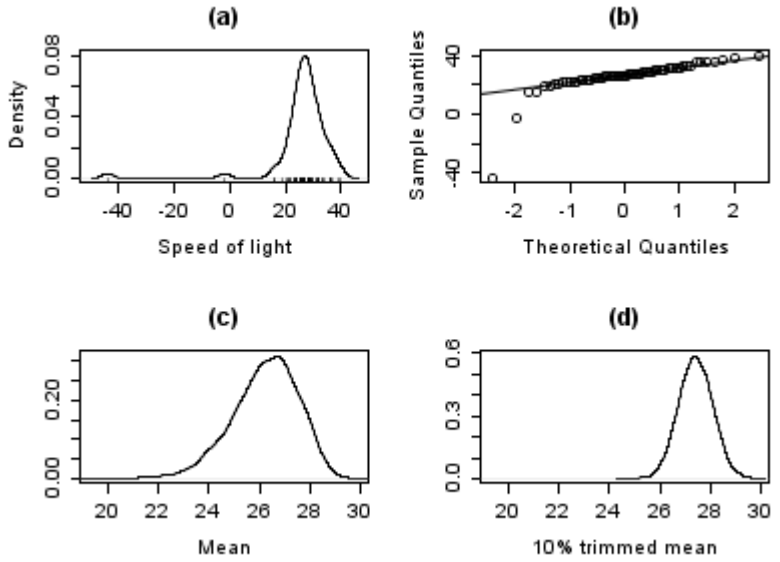
- Values above/below a threshold are considered outliers
  - any i that has an MD value above the 0.975 quartile of a chi-square distribution

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 μ and σ

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.



Robust version of dispersion, i.e., σ

- 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

A peek into next week ..

Inference & New Stats

Frequentist inference

- Make better inferences by considering **repeated sampling** of datasets similar to the one at hand
- One tries to conclude with :
  - **Point estimates** – given the data that will serve as a "best guess" of an unknown population parameter (e.g., maximum likelihood)
  - **Interval estimates** - intervals that would contain the true parameter value with the probability at the stated confidence level (confidence intervals)

Some crucial frequentist concepts

- **Central Limit Theorem**: states that given a distribution with a mean μ and variance σ^2, the **sampling distribution of the mean** approaches a normal distribution with a mean (μ) and a variance σ^2/N as N, the sample size
- The very fundamental aspect about the central limit theorem is that **no matter what the shape of the original distribution**, the **sampling distribution of the mean approaches a normal distribution**.

Interactive demo on sampling distribution:
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

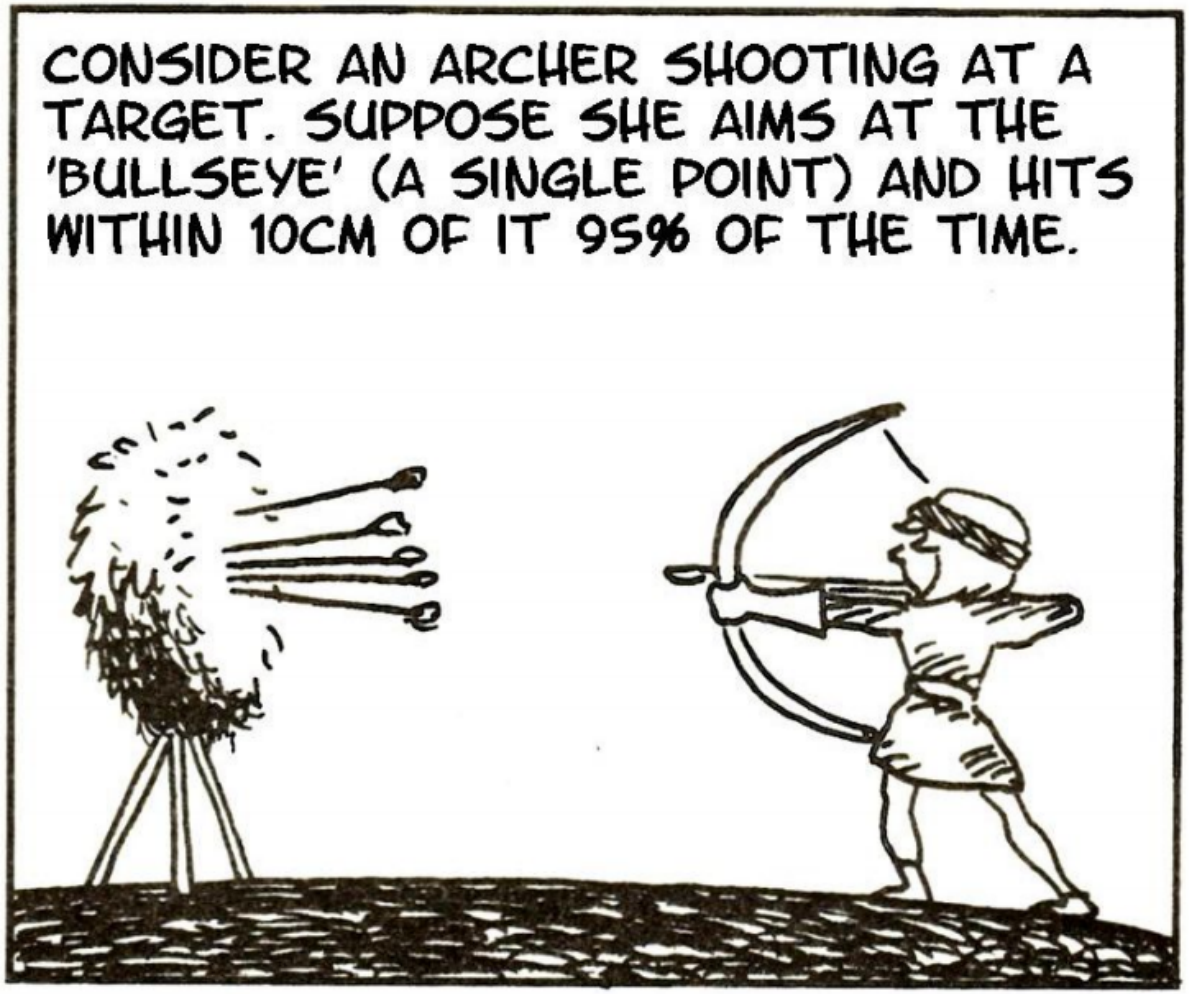
Some crucial frequentist concepts

- **Confidence Interval**: an interval estimate of a population parameter. It is an observed interval (i.e., it is calculated from the observations) that frequently includes the value of an unobservable parameter of interest if the experiment is repeated

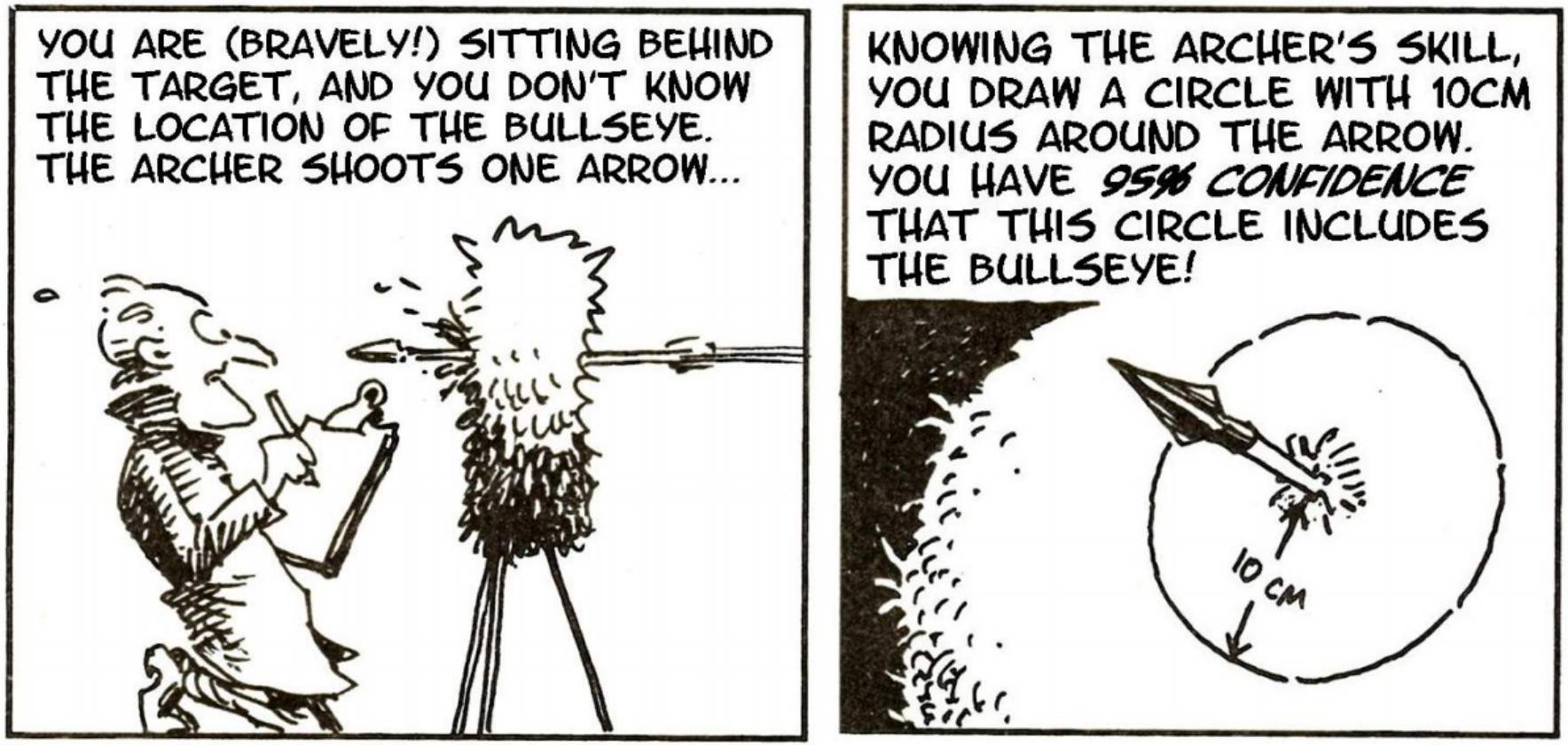
Interactive Demo : http://rpsychologist.com/d3/CI/

- **Careful** with how to interpret this: "*If the true value of the parameter lies outside the 90% (or x%) confidence interval once it has been calculated, then an event has occurred which had a probability of 10% (100-x%) (or less) of happening by chance*" [from Wikipedia]

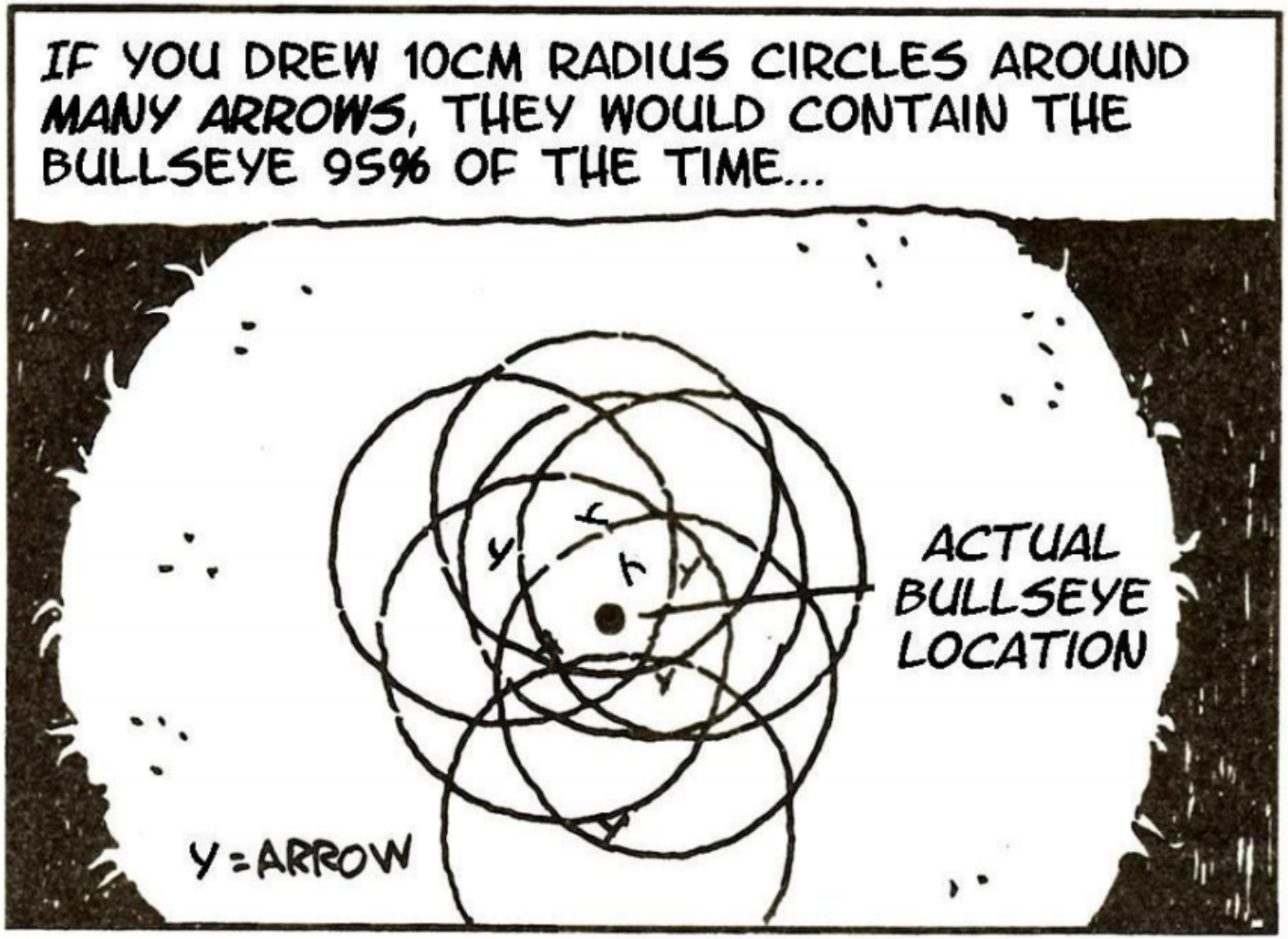
From **“The Cartoon Guide to Statistics”, by Gonick & Smith**



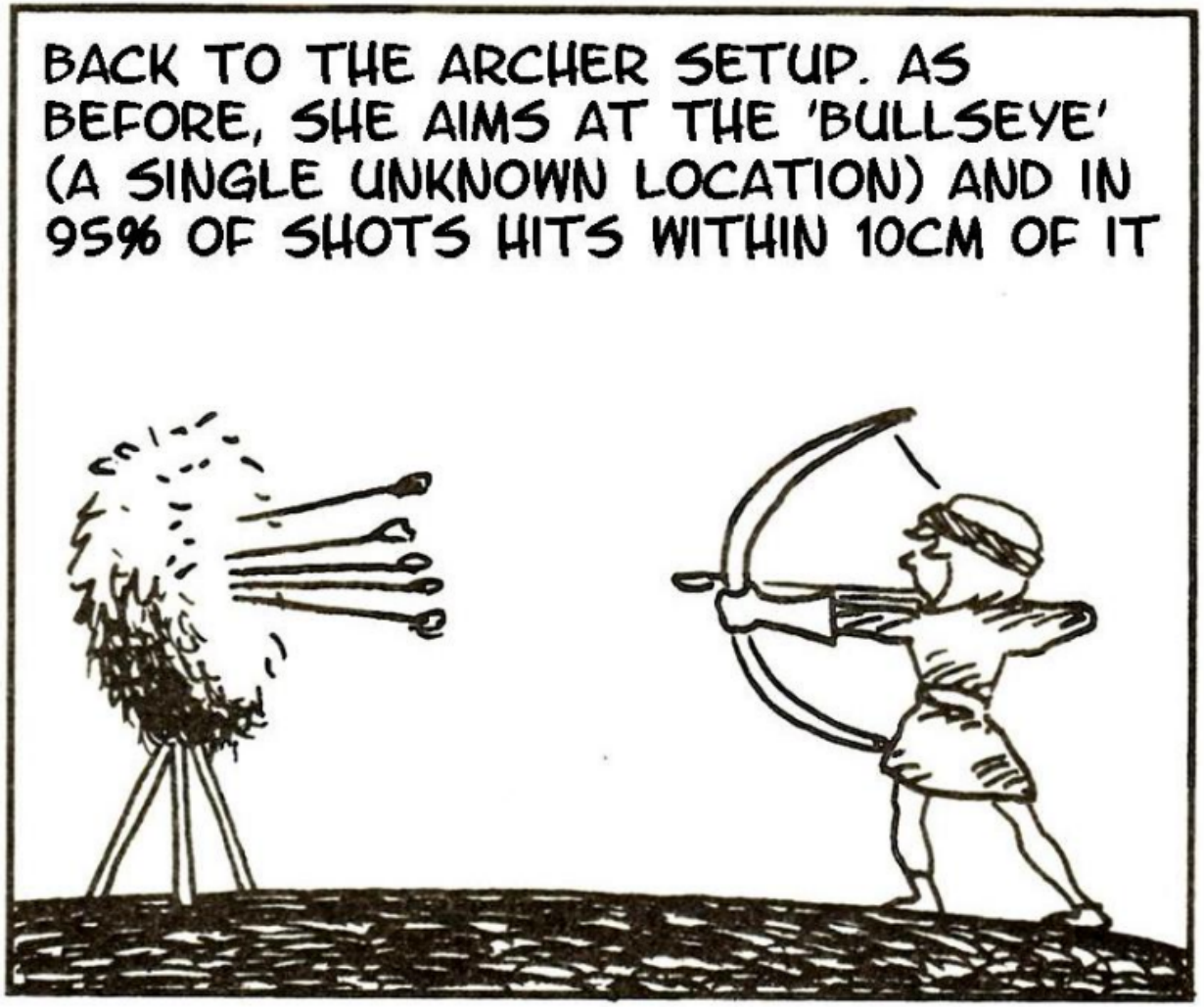
From “The Cartoon Guide to Statistics”, by Gonick & Smith



From **"The Cartoon Guide to Statistics", by Gonick & Smith**

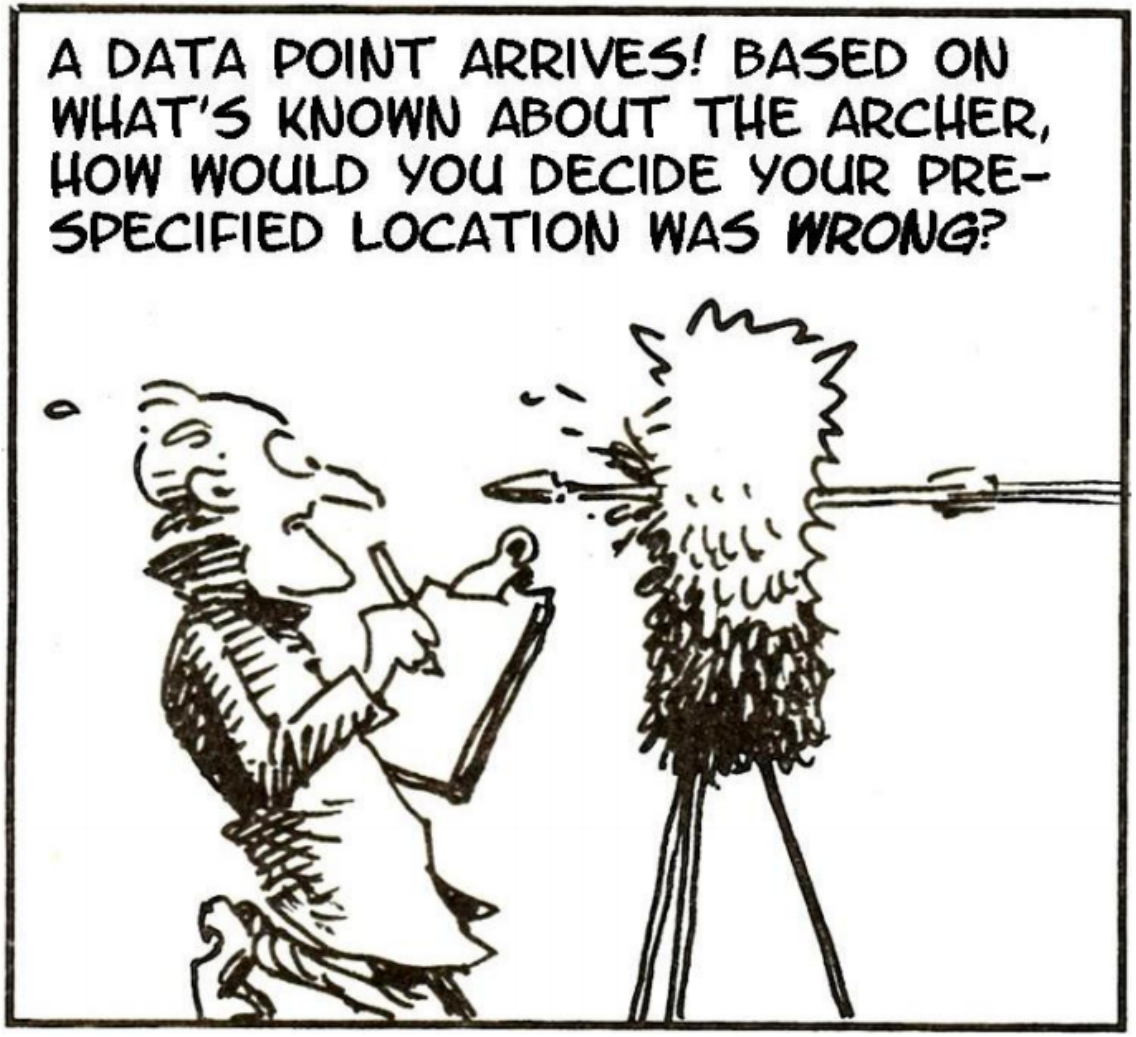


From **"The Cartoon Guide to Statistics", by Gonick & Smith**

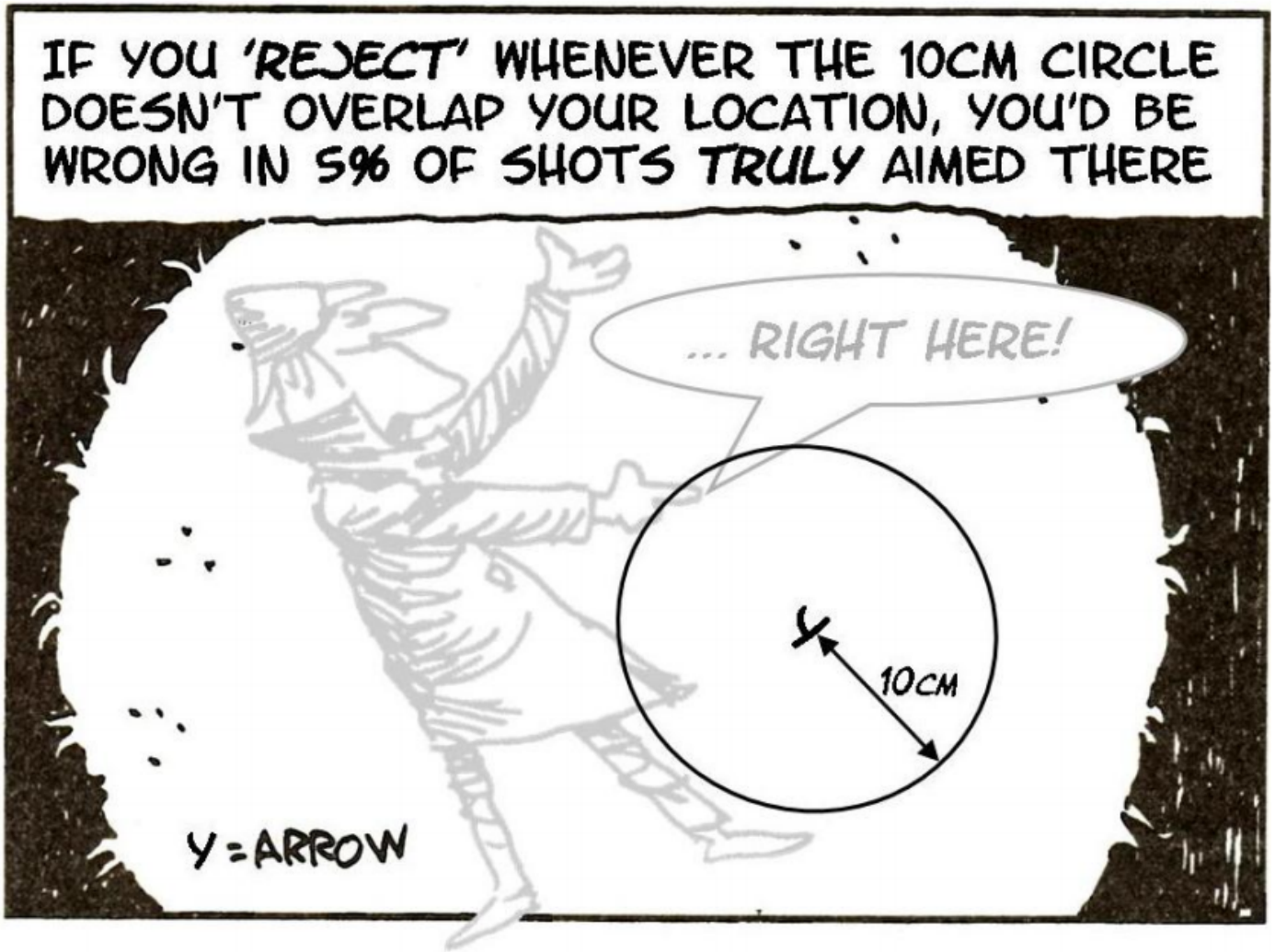


From **"The Cartoon Guide to Statistics", by Gonick & Smith**

From **"The Cartoon Guide to Statistics", by Gonick & Smith**



From **"The Cartoon Guide to Statistics", by Gonick & Smith**



Null hypothesis significance testing (NHST)

- Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses (an assumption about a population parameter)
- E.g., Want to understand "**Whether drug A has an affect on condition B, let's say insomnia**"
- So you want to check the hypothesis that "*participants who are treated with A , have improved conditions*", so this is what you are after (called the **Alternative Hypothesis, H_1**)
- Since you can never prove H_1 (according to freq. inf.), you can at least check whether it is wrong, so you come up with a **Null Hypothesis**, H_0 that says "*Drug A has no effect on insomnia*" and check if this is supported by your observations.
- You follow a formal procedure and you either ***reject the null hypothesis*** or you ***fail to reject the null hypothesis***

The infamous p-value and significance

- **Before** you do the test, you decide on a **significance level (α)** you want to observe before you make a claim
- **Significance level** is the probability of rejecting the null hypothesis given that it is true. By convention, the significance level is set to 0.05 (5%), implying that it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis
- Once the test is done, you end up with a p-value which is: "*The probability of obtaining the observed results, or results more extreme, if* $\boldsymbol{H_0}$ *is true*", so you need small p-values so that you can reject $\boldsymbol{H_0}$
- And how small is sufficient depends on the significance level you've picked, if p< α then

nature International weekly journal of science

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

ScienceNews
MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

Archive

Search Science News...

Explore

LATEST | MOST VIEWED

TELEVISION
Hollywood-made science documentary series comes to TV
BY TINA HESMAN SAEY OCTOBER 20, 2015

SCIENCE VISUALIZED
'Whalecopter' drone swoops in for a shot and a shower
BY SUSAN MILIUS OCTOBER 20, 2015

SCIENCE TICKER
Climate change could shift New

Context

SCIENCE PAST AND PRESENT
TOM SIEGFRIED

CONTEXT NUMBERS

P value ban: small step for a journal, giant leap for science

Editors reject flawed system of null hypothesis testing

BY TOM SIEGFRIED 3:18PM, MARCH 17, 2015

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

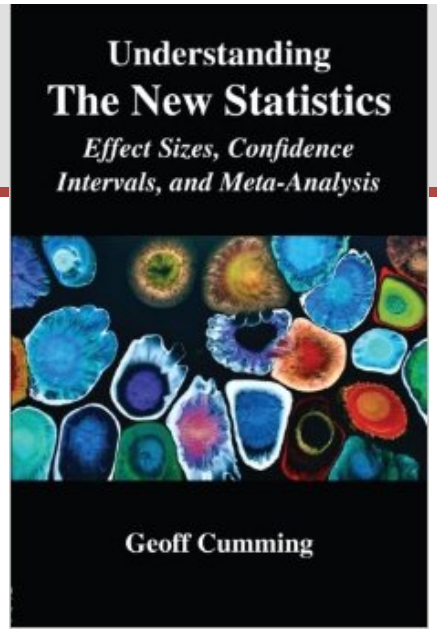
Editorial

David Trafimow and Michael Marks
New Mexico State University

The Basic and Applied Social Psychology (BASP) 2014 Editorial **emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it** (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. **From now on, BASP is banning the NHSTP.**

…. Some might view the NHSTP ban as indicating that it will be easier to publish in BASP, or that less rigorous manuscripts will be acceptable. This is not so. On the contrary, we believe that the **$p < .05$ bar is too easy to pass** and sometimes serves as an excuse for lower quality research.

The “new” statistics



- Are not really new
- ***Avoid binary decision-making*** as in NHST
- Replace ***estimation*** with testing
- Use effect sizes (ESs) and confidence intervals
- From Cumming, 2014:

“… Suppose you read in the news that “support for Proposition X is 53%, in a poll with an error margin of 2%.” Most readers immediately understand that the 53% came from a sample and, assuming that the poll was competent, conclude that 53% is a fair estimate of support in the population. The 2% suggests the largest likely error. Reporting a result in such a way, or as 53 ± 2%, or as 53% with a 95% CI of [51, 55], is natural and informative. It is more informative than stating that support is “statistically significantly greater than 50%, $p < .01$.” The 53% is our point estimate, and the CI our interval estimate, whose length indicates precision of estimation. “…

Cumming, G. (2013). The new statistics why and how. *Psychological science*.

Effect Sizes

- An ES is **simply an amount of anything of interest** (Cumming & Fidler, 2009). Means, differences between means, frequencies, correlations, and many other familiar quantities are ESs.
- Interpretation of ESs requires **informed judgment in context**. We need to trust our expertise and report our assessment of the size, importance, and theoretical or practical value of an ES, taking full account of the research situation.

Cumming, G. (2013). The new statistics why and how. *Psychological science*.

Effect Sizes – some examples

TABLE 2.1

Examples of ES Measures

Sample ES	Description	Example
Mean, M	Original units	Mean response time, $M = 462$ ms.
Difference between two means	Original units	The average price of milk increased last year by \$0.12/L, from \$1.14/L to \$1.26/L.
Median, Mdn	Original units	Median response time, $Mdn = 385$ ms.
Percentage	Units-free	35.5% of respondents were in favor. 0.7% of responses were errors.
Frequency	Units-free	39 states ran a deficit.
Correlation, r	Units-free	Income correlated with age ($r = .28$).
Cohen's d	Standardized	The average effect of psychotherapy was $d = 0.68$ (see Chapters 7 and 11).
Regression weight, b	Original units	The slope of the regression line for income against age was $b = \$1{,}350$/year.
Regression weight, β	Standardized	The β-weight for age in the regression was .23.
Proportion of variance, R^2	Units-free	Three variables of age, education, and family status in the multiple regression together gave $R^2 = .48$.
Risk	Units-free	The risk that a child has a bicycle accident in the next year is 1/45.
Relative risk	Units-free	A boy is 1.4 times as likely as a girl to have a bicycle accident in the next year.
Proportion of variance, ω^2 (Greek omega-squared)	Units-free	In the analysis of variance, the independent variable age accounted for $\omega^2 = 21.5\%$ of total variance.

Cohen's d (Jacob Cohen (1988). Statistical Power Analysis for the Behavioral Sciences)

Standardized difference between sample means

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}.$$

.. where, s is the pooled sample mean

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

.. and variance for the groups are computed as:

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n1} (x_{1,i} - \bar{x}_1)^2,$$

Very good resource:
http://rpsychologist.com/d3/cohend/

Some suggestions from Cumming G. (2013)

- **Formulate research questions in estimation terms.** To use estimation thinking, ask “How large is the effect?” or “To what extent . . . ?” Avoid dichotomous expressions such as “test the hypothesis of no difference” or “Is this treatment better?”
- **Identify the ESs that will best answer the research questions.** If, for example, the question asks about the difference between two means, then that difference is the required ES
- more in the article…

From:Cumming, G. (2013). The new statistics why and how. *Psychological science*.

But the debate will continue …

The "new statistics" are built on fundamentally flawed foundations

Michael D. Lee

University of California, Irvine

…. Methodologically, it is wrong to rely on flawed orthodox statistical theory. We should all just be Bayesian.

Next week

- A quick look into the New Statistics
- And on to more advanced computational analysis
  - Correlations
  - Regressions
  - Structures, groups, clusters