INM3061/INM430
Principles of Data Science

Week 04
Inferential Statistics

Aidan Slingsby,
Constantino Carlos Reyes-Aldasoro

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 (and writing code in the Real World)

On the menu today

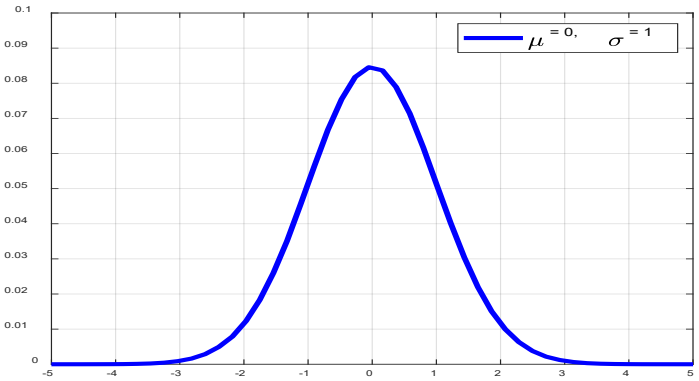
- (Some more) descriptive statistics
  - Recap of last week
  - Outliers and robust statistics (from last week)
  - Correlation
- Inferential statistics
  - Frequentist and Bayesian approaches
  - NHST
  - P-values and statistical significance
- Caution on relying on statistical significance
  - P-value controversy

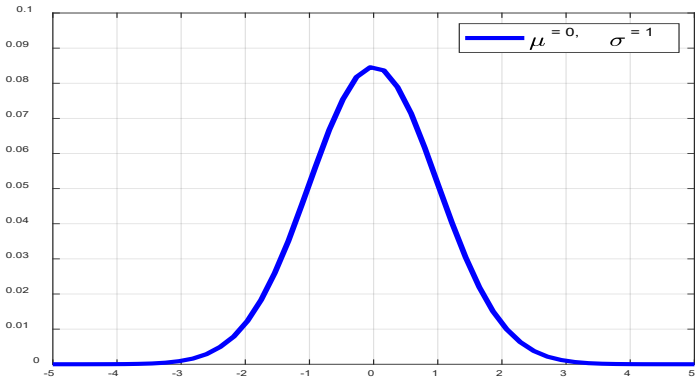
RECAP OF LAST WEEK

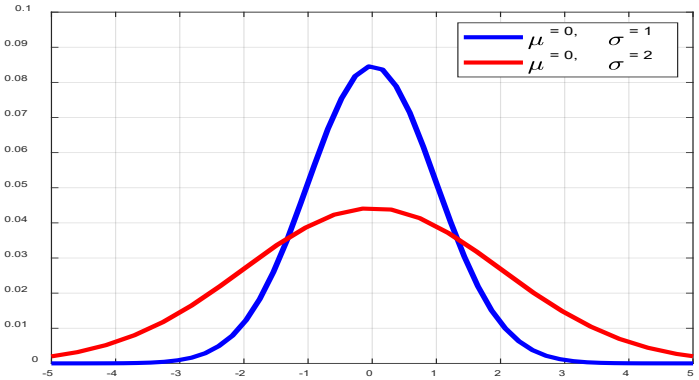
Last week's lab

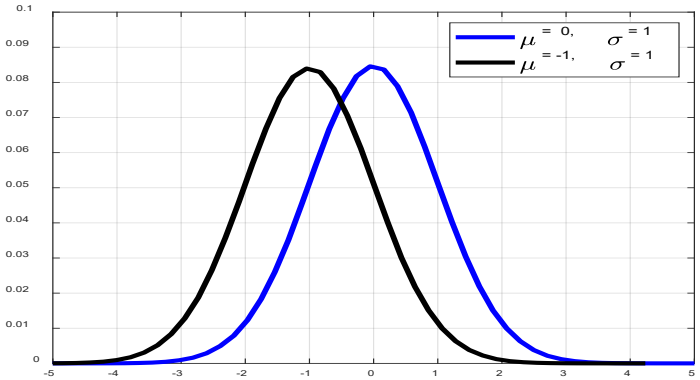
- Identifying outliers
  - Using the standard deviation
- Q-Q plots
  - for comparing distributions with theoretical distributions
- Sampling
  - sampling size
  - robust statistics

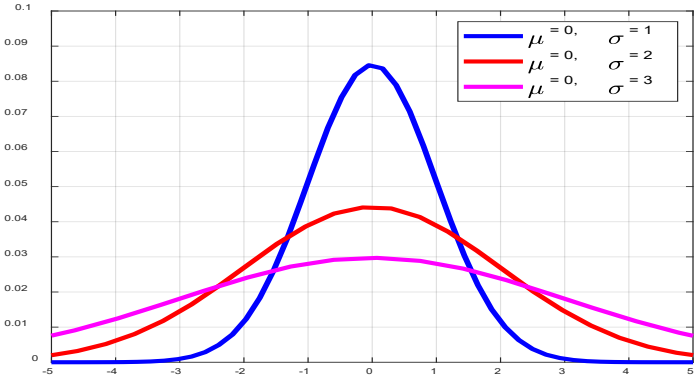
Normal distribution

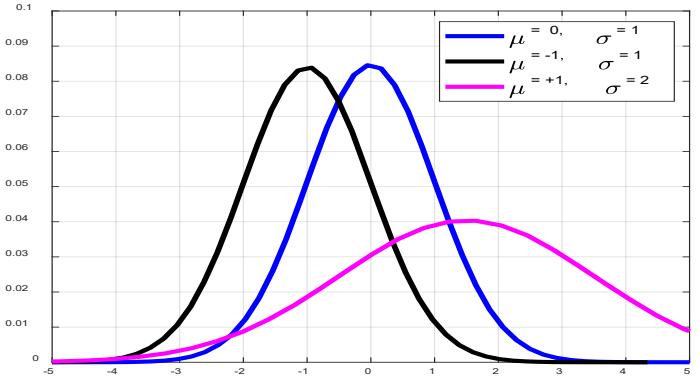




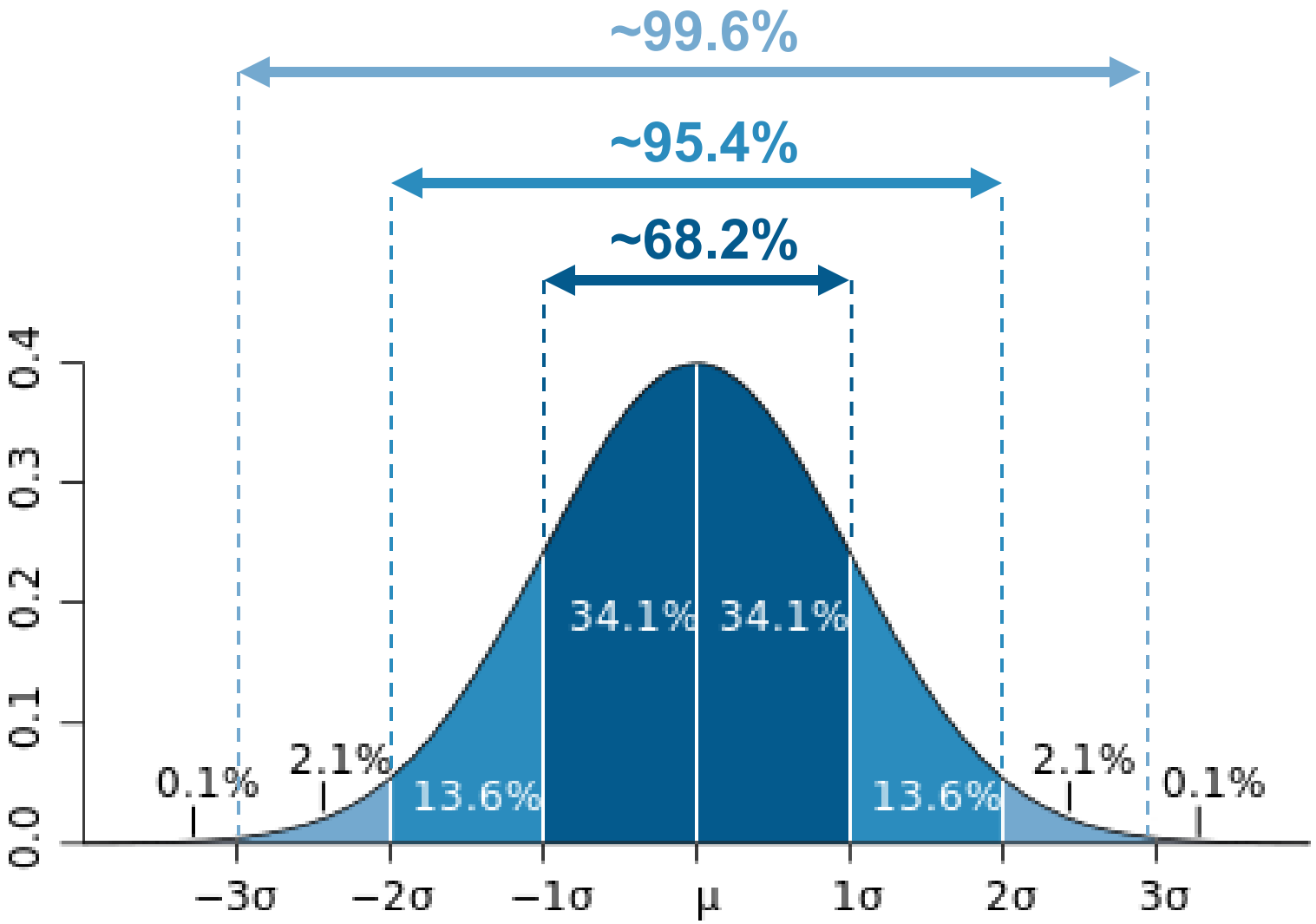








Normal distribution



Tests for normality

- Q-Q plots
- Are the right proportions of data within the right number of standard deviations?
- Shapiro–Wilk test

MORE DESCRIPTIVE STATISTICS

Correlation

- Determines the **degree** to which two variables are **related**
- To find the relationship between two quantitative variables (without being able to infer causal relationships)

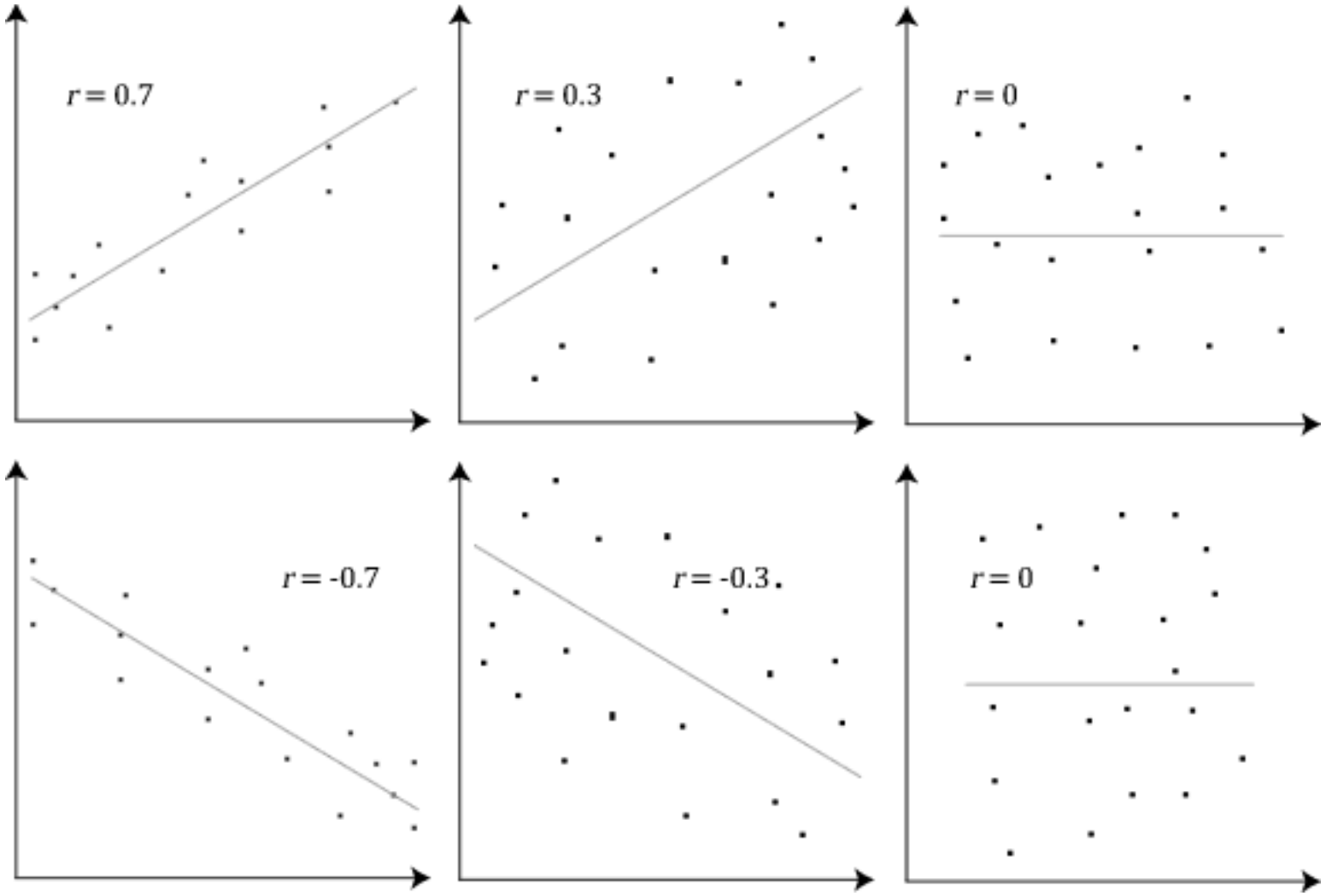
Pearson's Correlation

- Correlation coefficient $\boldsymbol{\rho}$ (or r) in [-1, 1], where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation

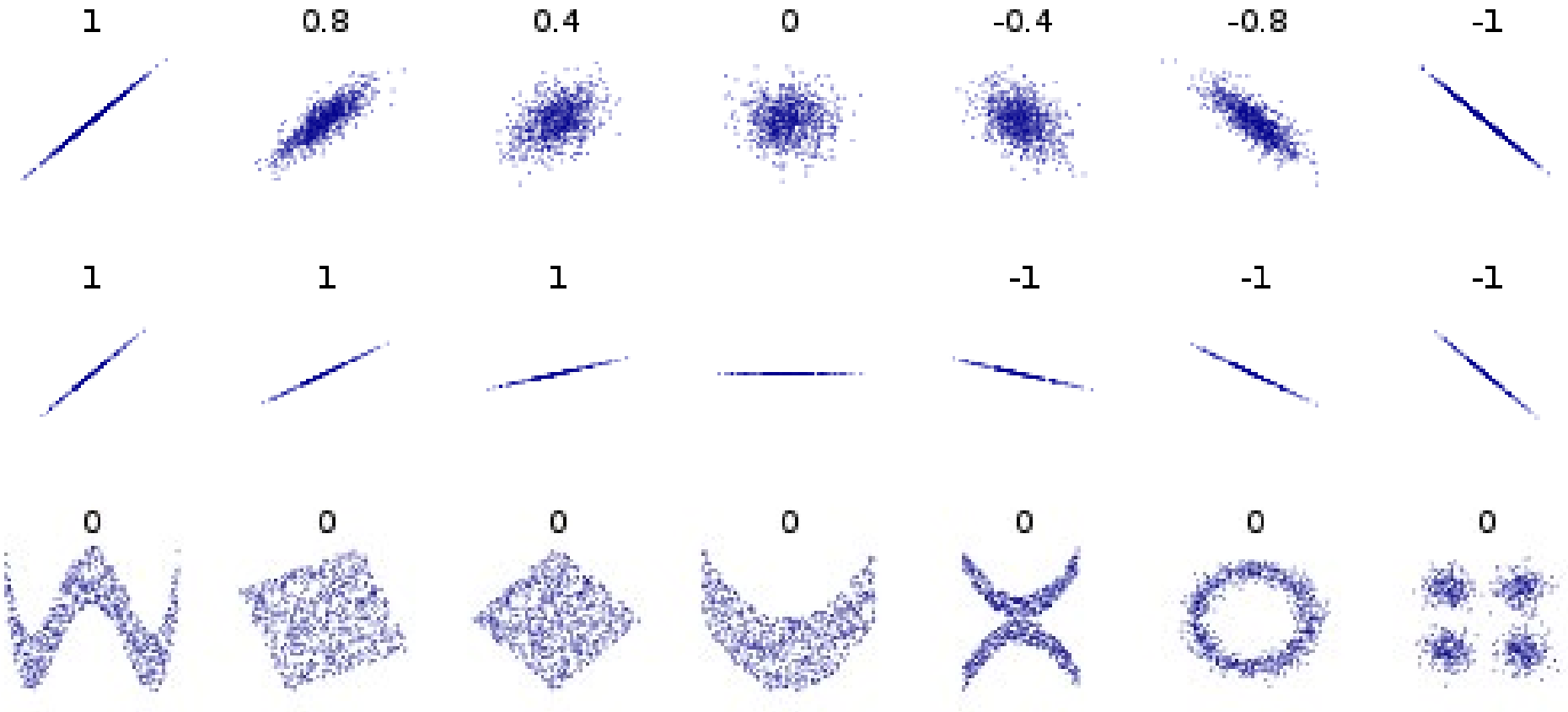
Sample mean

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

Some levels of correlation



Some levels of correlation



Pearson's Correlation

- Suitable for linear relations (assumption)
- Outliers is an issue
- Might need to test for significance, use a t-test
- Rules of thumb, r =

+.70 or higher	Very strong positive relationship
+.40 to +.69	Strong positive relationship
+.30 to +.39	Moderate positive relationship
+.20 to +.29	Weak positive relationship
+.01 to +.19	No or negligible relationship
-.01 to -.19	No or negligible relationship
-.20 to -.29	Weak negative relationship
-.30 to -.39	Moderate negative relationship
-.40 to -.69	Strong negative relationship
-.70 or higher	Very strong negative relationship

Spearman's rank correlation

- Nonparametric measure (more resistant to outliers)
- Rank based, captures **monotonic** relations
- Linearity not assumed
- Works with categorical data

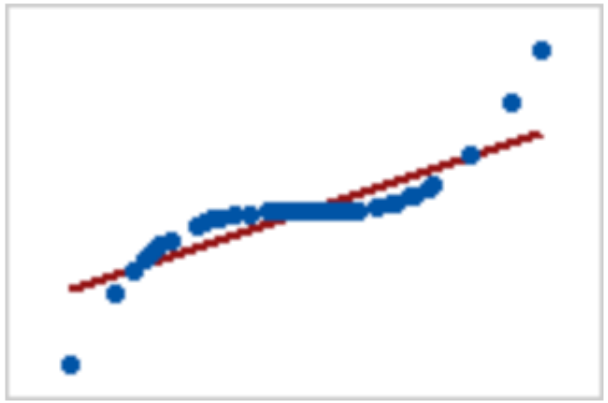
Distance in ranks

$$d_i = x_i - y_i$$

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

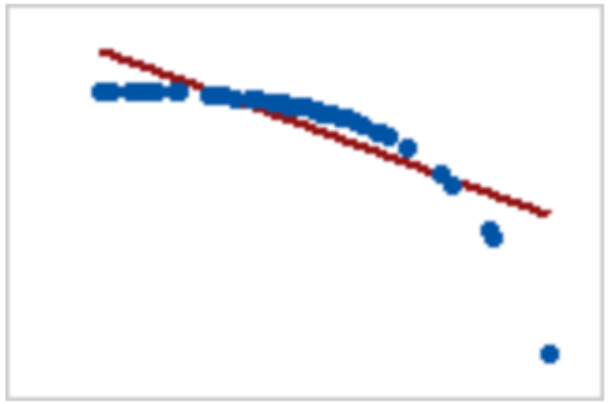
Pearson vs Spearman

Pearson = +1, Spearman = +1



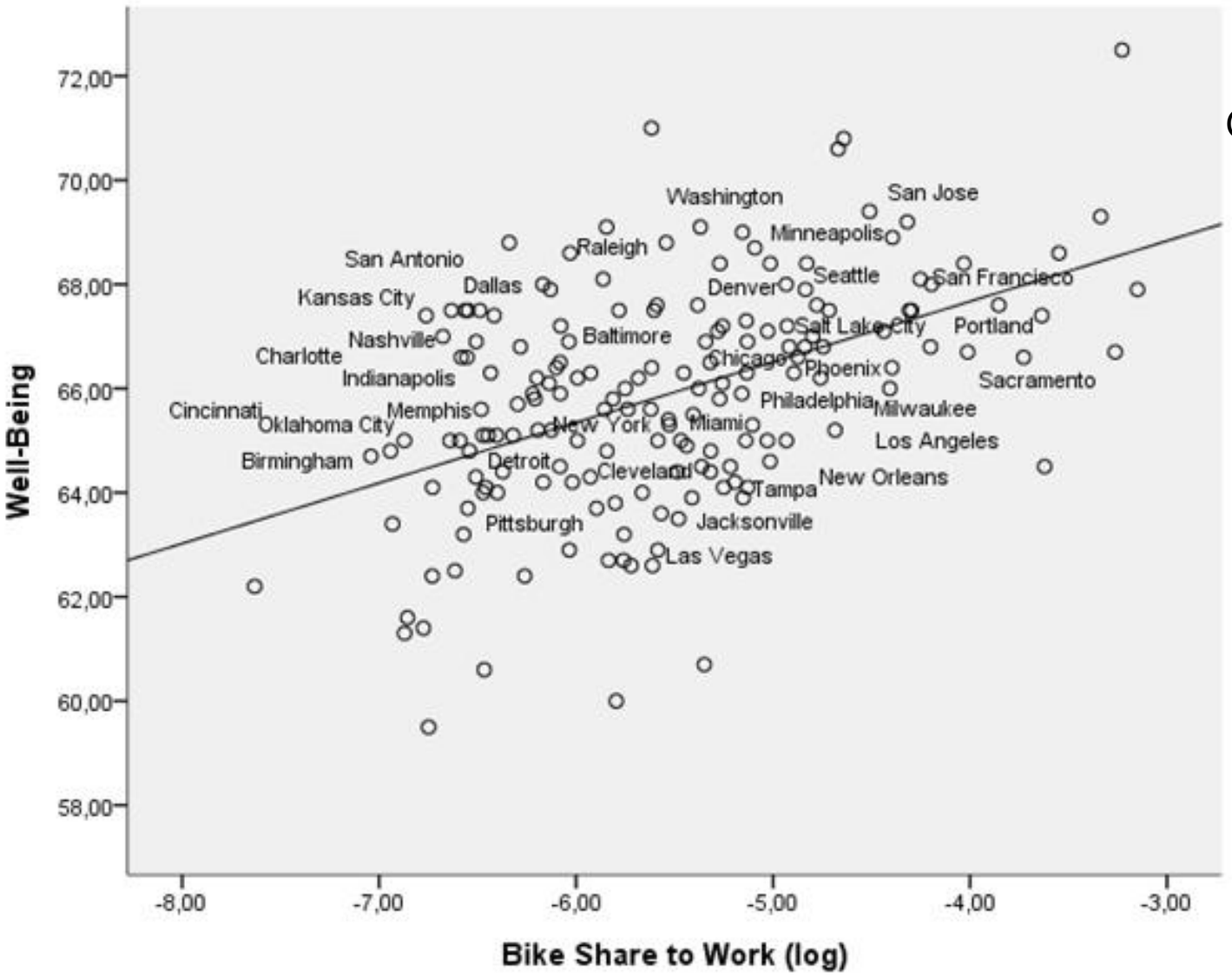
Pearson = +0.851, Spearman = +1

Pearson = −1, Spearman = −1



Pearson = −0.799, Spearman = −1

Example – Well-being vs. cycling



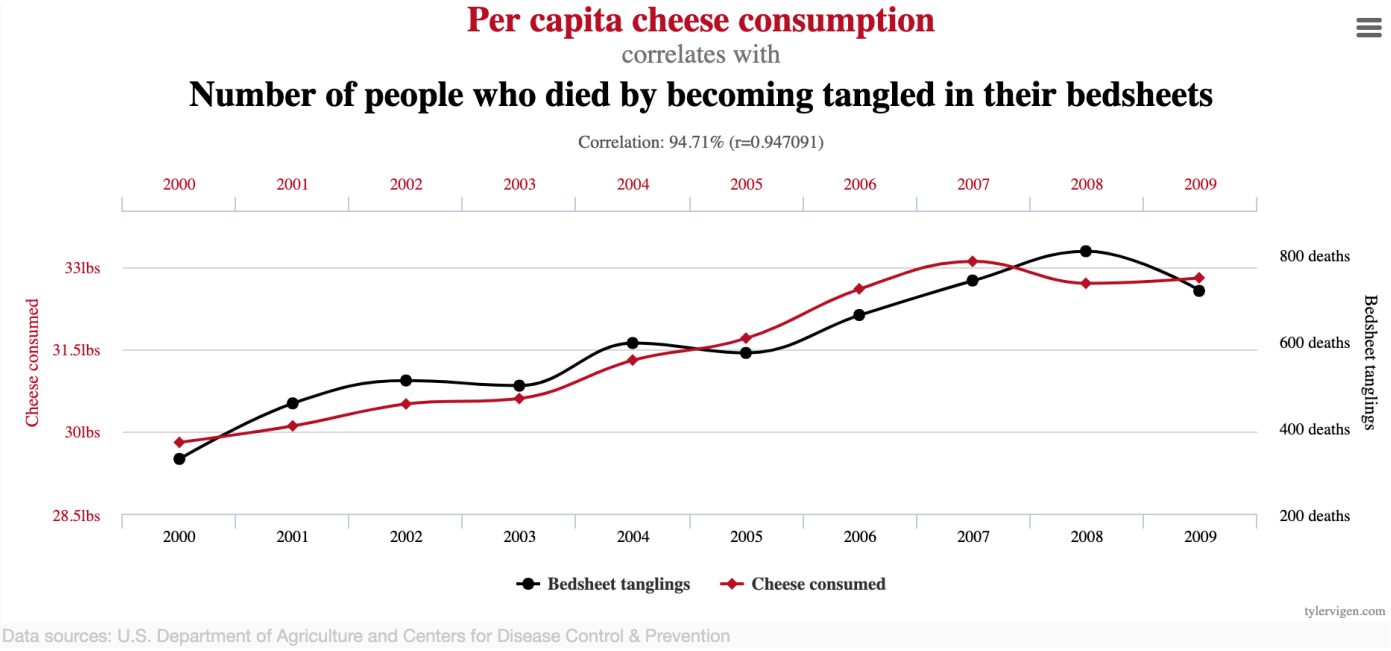
Correlation of 0.5

http://www.theatlantic.com/national/archive/2011/06/americas-top-cities-for-bike-commuting-happier-too/240265/

Correlation does not imply causation

- "Correlation coefficients are simply summaries of association and cannot be used to conclude that there is an underlying relationship between variables, let alone why one might exist"

David Spiegelhalter, The Art of Statistics, Pelican



https://www.tylervigen.com/spurious-correlations

Correlation does not imply causation

Journal of Epidemiology & Community Health

Latest content Current issue

Home / Archive / Volume 70, Issue 12

Socioeconomic position and the risk of brain tumour: a Swedish national population-based cohort study

PDF

Amal R Khanolkar [1, 2], Rickard Ljung [2], Mats Talbäck [2], Hannah L Brooke [2], Sofia Carlsson [2], Tiit Mathiesen [3], Maria Feychting [2]

Correspondence to Dr Amal R Khanolkar, Institute of Child Health, University College London, 30, Guildford Street, London WC1N 1E
a.khanolkar@ucl.ac.uk

Abstract

Background The aim was to investigate associations between different measures of socioeconomic position (SEP) an of brain tumours (glioma, meningioma and acoustic neuroma) in a nationwide population-based cohort.

Methods We included 4 305 265 individuals born in Sweden during 1911–1961, and residing in Sweden in 1991. Coho were followed from 1993 to 2010 for a first primary diagnosis of brain tumour identified from the National Cancer Regis regression was used to compute incidence rate ratios (IRR) by highest education achieved, family income, occupationa marital status, with adjustment for age, healthcare region of residence, and time period.

Results We identified 5735 brain tumours among men and 7101 among women during the study period. Highly educated (≥3 years university education) had increased risk of glioma (IRR 1.22, 95% CI 1.08 to 1.37) compared to men with primary education. High income was associated with higher incidence of glioma in men (1.14, 1.01 to 1.27). Women with ≥3 years university education had increased risk of glioma (1.23, 1.08 to 1.40) and meningioma (1.16, 1.04 to 1.29) compared to tho primary education. Men and women in intermediate and higher non-manual occupations had increased risk of glioma comp low manual groups. Compared to t decreased risk of glioma in men. M in low manual occupations.

INDEPENDENT

Rugby World Cup › Visit The USA › ENHANC

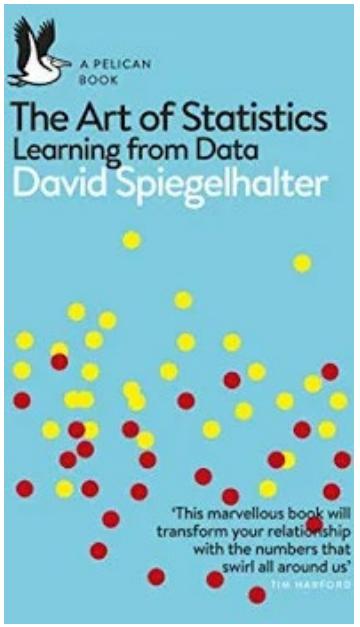
NEWS SPORT VOICES CULTURE LIFESTYLE TRAVEL PREMIUM MORE

Lifestyle > Health & Families > Health News

University education linked to increased brain tumour diagnosis

People in professional and managerial roles are also more likely to be diagnosed

Emma Henderson • Tuesday 21 June 2016 13:54 • Comments



INFERENTIAL STATISTICS

Descriptive vs Inferential statistics

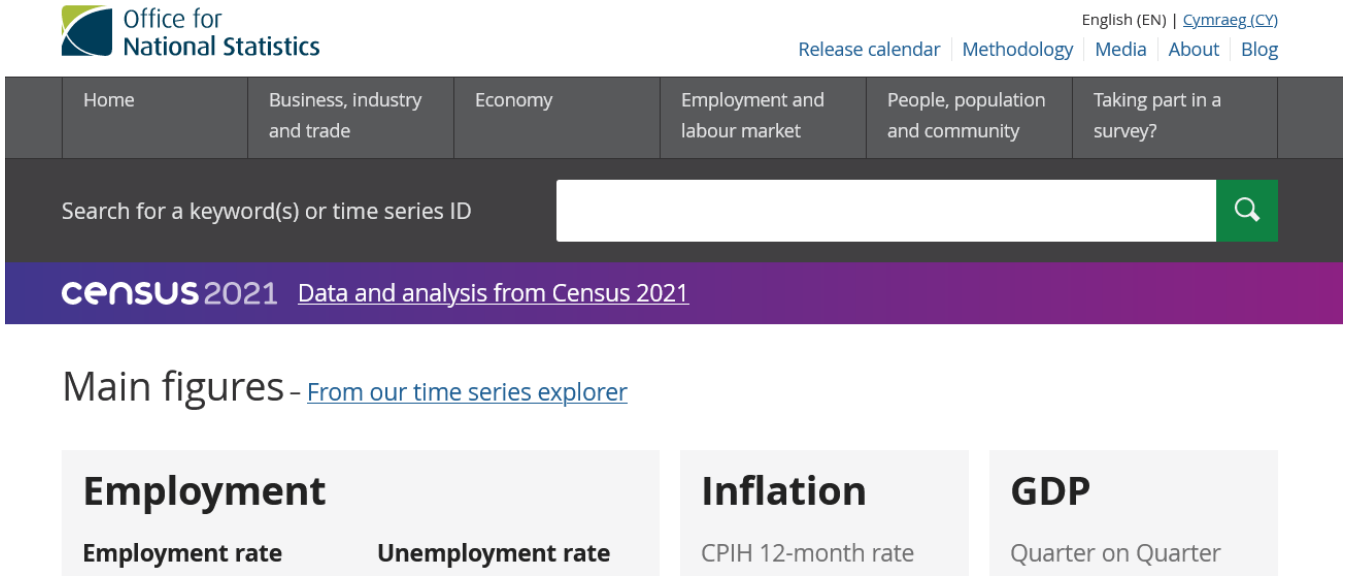
- **Descriptive:** Simply describing data that you have sampled
  - Checking for problems
  - Determining their characteristics (distribution)
  - Drawing conclusions about our **sample**
  - Informing data preparation and feature engineering
  - Helping determine whether the data behave as expected
- **Inferential:** investigating an unknown population from sampled data
  - Is the effect that you see in your sample likely of the unknown population?
  - Is it "statistically significant"?

Summarise Concisely

Learn something

Descriptive vs Inferential statistics

- **Descriptive:**
  - Observe general situation, derive a conclusion



- **Inferential:**
  - Take particular instances and derive general conclusions

Inferential Statistics

- Works by making inferences about the **population** from your **sample** based on statistical models and assumptions
  - assume that the **population** follows a **certain distribution**
  - infer how representative our sample is of our population
- **Question**: How many trees exist in the world? (Or in a country)
  - Mmmh … But what is a tree?

Tree

210 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

For other uses, see Tree (disambiguation).

In botany, a **tree** is a perennial plant with an elongated stem, or trunk, usually supporting branches and leaves. In some usages, the definition of a tree may be narrower, including only woody plants with secondary growth, plants that are usable as lumber or plants above a specified height. In wider definitions, the taller palms, tree ferns, bananas, and bamboos are also trees.

Trees are not a monophyletic taxonomic group but consist of a wide variety of plant species that have independently evolved a trunk and branches as a way to tower above other plants to compete for sunlight. The majority of tree species are angiosperms or hardwoods; of the rest, many are gymnosperms or softwoods. Trees tend to be long-lived, some reaching several thousand years old. Trees have been in existence for 370 million years. It is estimated

Shrub

106 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

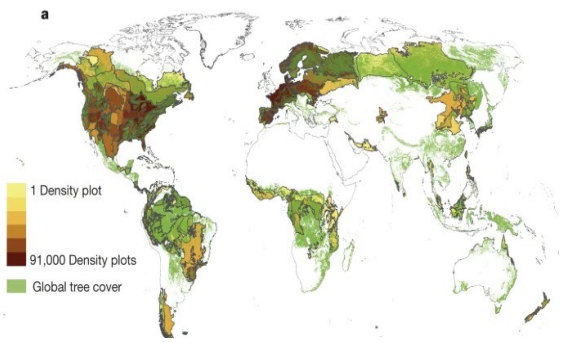
A **shrub** (often also called a **bush**) is a small-to-medium-sized perennial woody plant. Unlike herbaceous plants, shrubs have persistent woody stems above the ground. Shrubs can be either deciduous or evergreen. They are distinguished from trees by their multiple stems and shorter height, less than 6–10 m (20–33 ft) tall.[1][2] Small shrubs, less than 2 m (6.6 ft) tall are sometimes termed as subshrubs. Many botanical groups have species that are shrubs, and others that are trees and herbaceous plants instead.

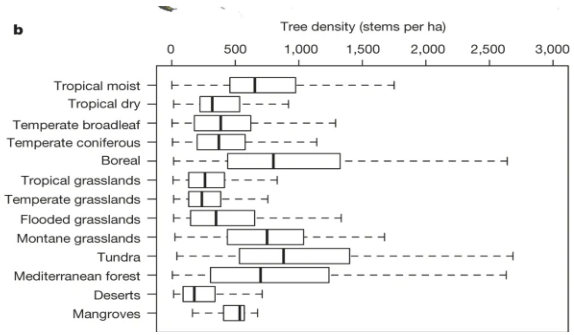
Some define a shrub as less than 6 m (20 ft) and a tree as over 6 m. Others use 10 m (33 ft) as the cutoff point for classification.[2] Many trees do not reach this mature height because of hostile less than ideal growing conditions, and resemble shrub-sized plants. Others in such species have the potential to grow taller in ideal conditions. For longevity,

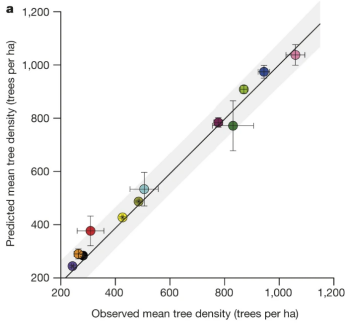
A broom shrub in flower

How many trees exist in the world?

- Unlike in a census, it is not possible to go around the world counting all trees (once properly defined).
- Example methodology:
  1. Define type area: *Forest, Rain Forest, Desert, Tundra, City, Town, Farm, etc. (actually biomes)*
  2. Count number of trees per area above (km^2)
  3. Find distribution of areas (satellite images are handy)
  4. Extrapolate with care = 3.04 trillion







T. W. Crowther, Mapping tree density at a global scale, Nature 525, (2015)

Inductive Inference Methodology

- From Data to Sample
  - Measurement, Reliable, Repeatable, Valid (no systematic bias)

92% Of Ryanair Customers Satisfied With Flight Experience

Ryanair, Europe's No.1 airline, today (5 Apr) released its quarterly 'Rate My Flight' statistics, which show that 92% of surveyed customers were happy with their o[illegible] flight experience in January, February and March 2017.

Category	Excellent/Very Good/ Good	Excellent	Very Good	Good	Fair	Ok
Overall Experience	92%	43%	35%	14%	4%	4%

- From Sample to Study Population
  - Internal validity, does the sample reflects the group?
  - Randomness
- From Study Population to Target Population
  - External validity, generalization, replication

Data

Sample

Study Population

Target Population

Inferential Statistics

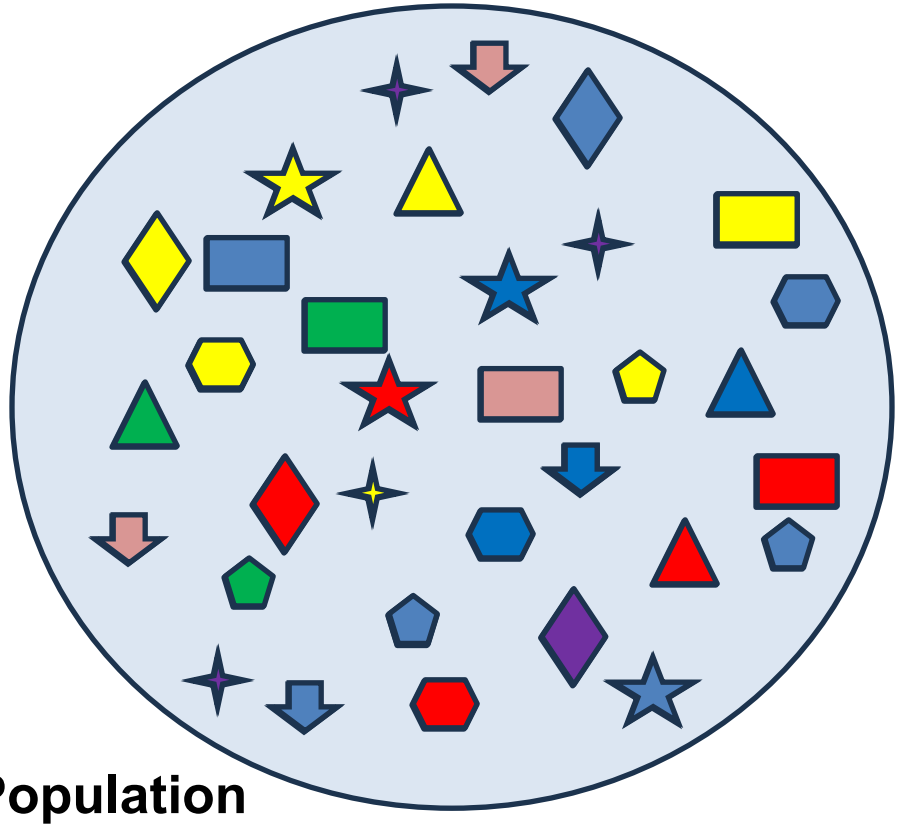
- Works by making inferences about the **population** from your **sample** based on statistical models and assumptions
  - assume that the **population** follows a **certain distribution** or **different distributions** (like the trees)
  - infer how representative our sample is of our population
- Underpinned by the **Central Limit Theorem**
  - if you have a population with mean μ and standard deviation σ and take **sufficiently large random samples** from the population with replacement, then the **distribution of the sample means** will be approximately **normally distributed**

Inferential Statistics

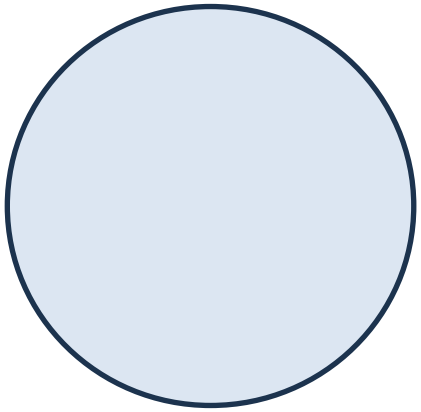
- Only necessary you are inferring from a sample
- If you're just describing your data, you do not need inferential statistics.
  - Describing the differences in heights in a class
  - Inferring population-level differences in heights
- If you collect a whole population, you do not need inferential statistics (unless you want to forecast).

Sampling

- Sampling is the selection of a subset or a statistical sample of individuals from within a statistical population.



Population



Sample

Sampling

- How do we guarantee that the sample is representative? How big does a sample need to be to be representative?
- *If you have cooked a large pan of soup, you do not need to eat it all to find out if it needs more seasoning. You can just taste a spoonful, provided you have given it a good stir"*, George Gallup.

George Gallup

Born	George Horace Gallup November 18, 1901 Jefferson, Iowa, U.S.
Died	July 26, 1984 (aged 82) Tschingel ob Gunten, Bernese Oberland, Switzerland
Alma mater	University of Iowa
Occupation	Statistician
Known for	Gallup poll

Cream of asparagus soup

11 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Cream of asparagus soup is a soup prepared with asparagus, stock and milk or cream as primary ingredients.

Ingredient variations exist. Cream of asparagus finished with various garnishes such as chives,

Ingredients and preparation me

Asparagus, a light chicken or vegetable stock ar asparagus may be puréed or pulped in its prepa and solid forms of asparagus, such as cooked a sieve to remove stray stringy asparagus matter. flavor of puréed versions of cream of asparagus season (during spring).[5][6]

Cream of asparagus soup

Stew

Article Talk

From Wikipedia, the free encyclopedia

This article is about the food. For NBA pl

A **stew** is a combination of solid food ingredi resultant gravy. Ingredients can include any c especially tougher meats suitable for slow-co sausages, and seafood. While water can be small amount of red wine or other alcohol is may also be added. Stews are typically cook allowing flavours to mingle.

Stewing is suitable for the least tender cuts o heat method. This makes it popular for low-c gelatinous connective tissue give moist, juicy

Stews are thickened by reduction or with flou

Hot and sour soup

Article Talk

From Wikipedia, the free encyclopedia

This article **needs additi** to reliable sources. Unso

Find sources: "Hot and sour s

remove this template messag

Hot and sour soup is a popular example of Sichuan cuisine. Similar versions are found in Henan province, near Beijing, and in Henan cuisine itself, where it may also be known as *hulatang* or "pepper hot soup" (胡辣湯). Also popular in Southeast Asia, India, Pakistan and the United States, it is a flexible soup which allows ingredients to be substituted or added depending on availability. For example, the American-Chinese version can be thicker as it commonly includes corn starch, whilst in Japan, sake is often added.

North America [edit]

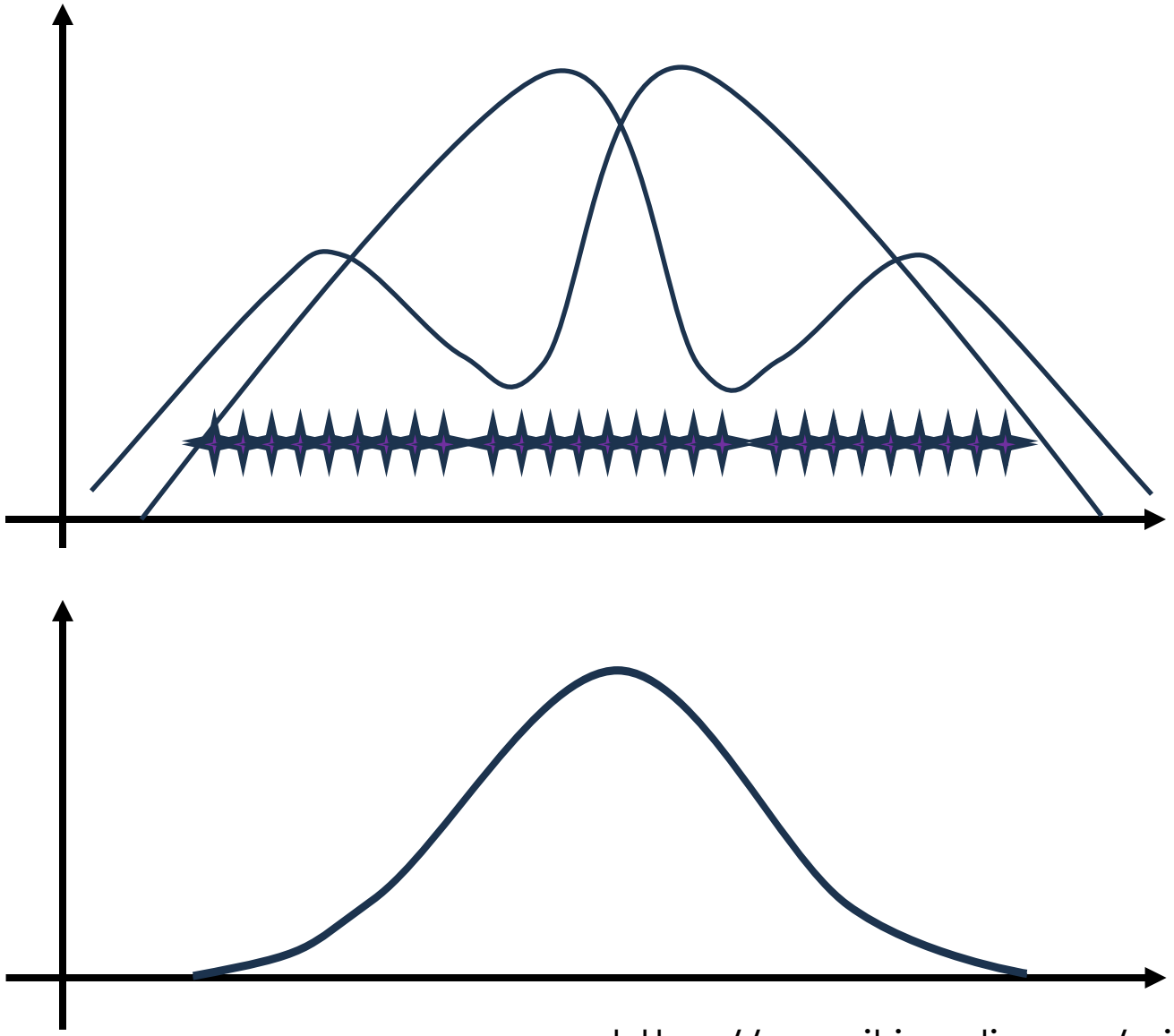
United States [edit]

Soup preparation may use chicken or pork broth, or may be meat-free. Common basic ingredients in the American Chinese version include bamboo shoots, toasted sesame oil, wood ear, cloud ear fungus, day lily buds, vinegar, egg, corn starch, and white pepper.[1] Other ingredients include button mushrooms, shiitake

Hot and sour soup

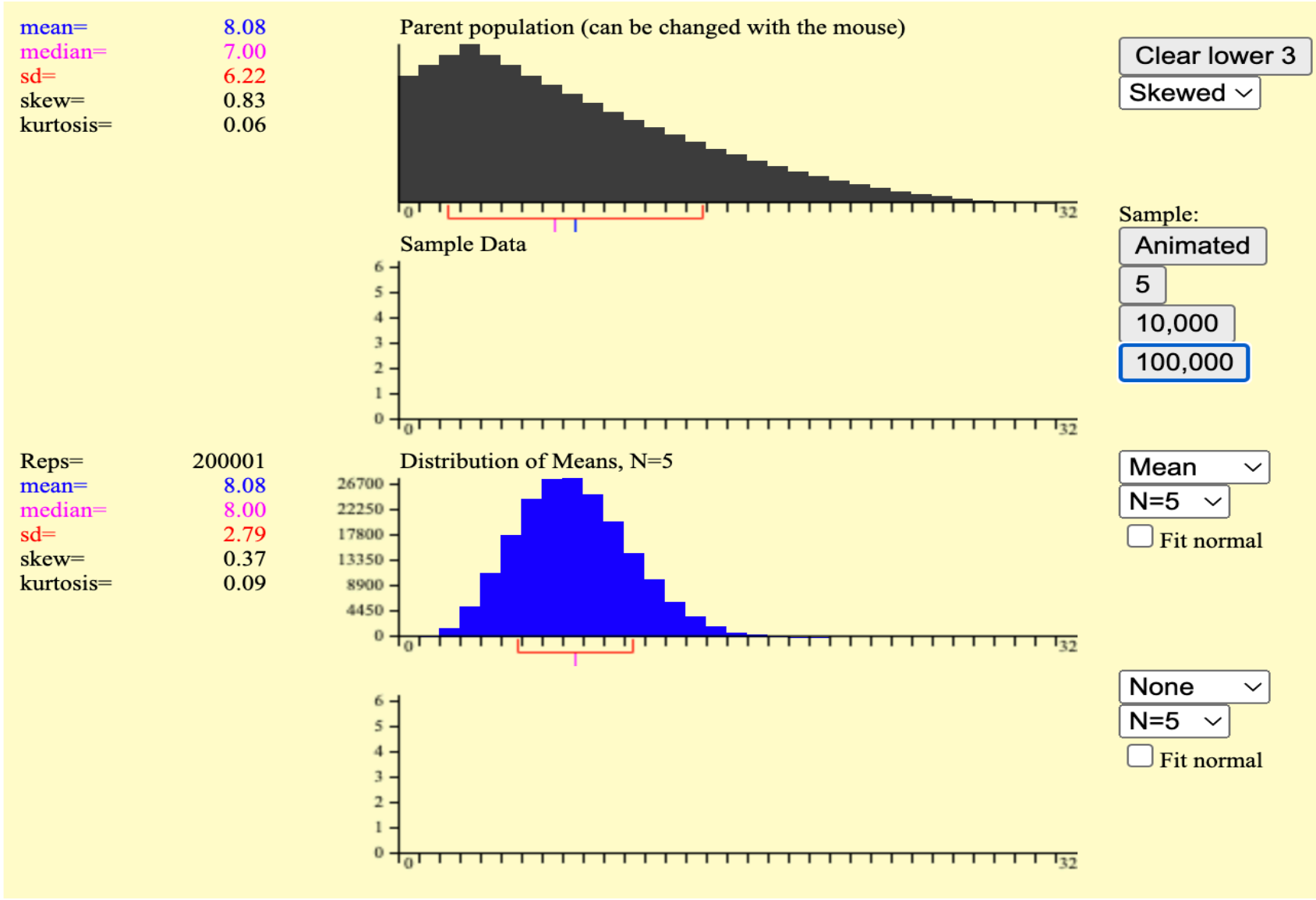
Type	Soup
Place of origin	China

Sampling: central limit theorem



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

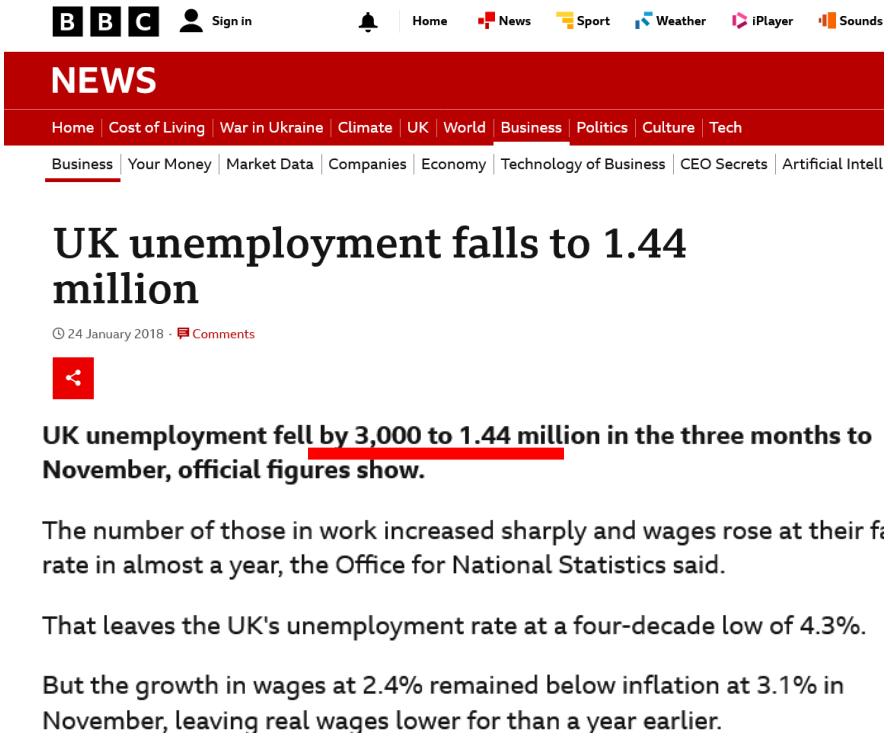
Sampling: central limit theorem

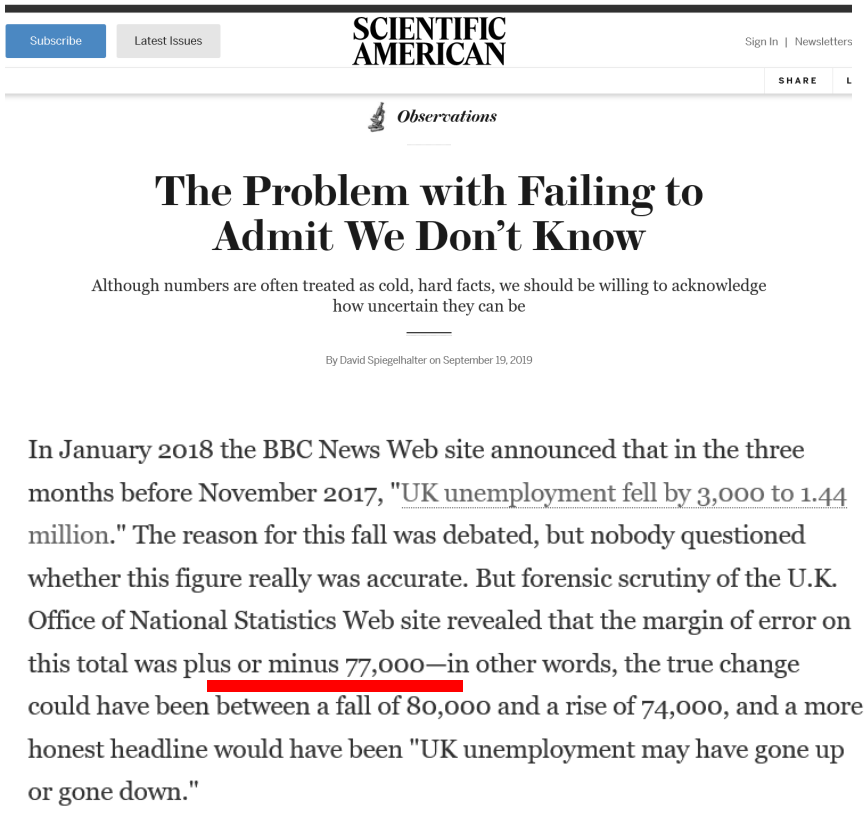


http://onlinestatbook.com/stat_sim/sampling_dist/index.html

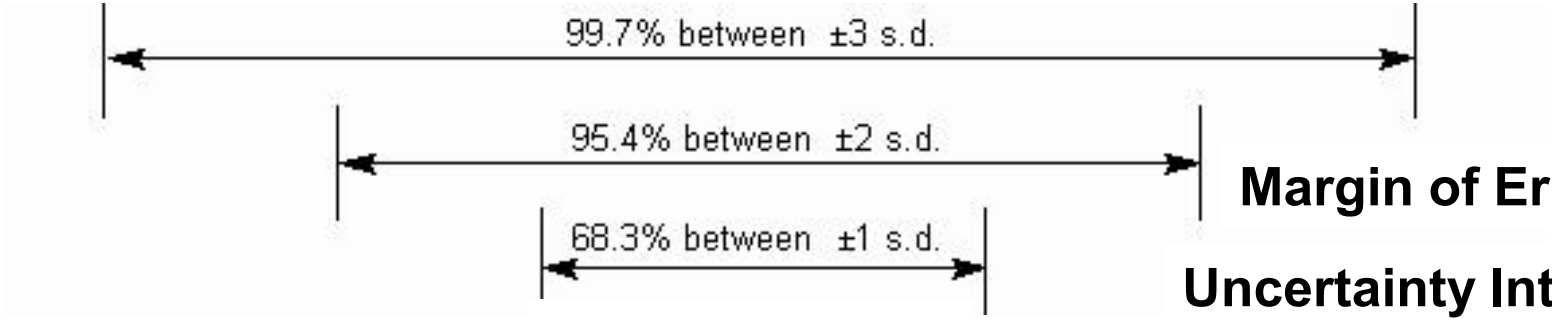
Sampling: margin of error

- The margin of error is a statistic expressing the amount of random sampling error in the results of a survey.
- The larger the margin of error, the less confidence one should have that a poll result would reflect the result of a census of the entire population.





Sampling: margin of error



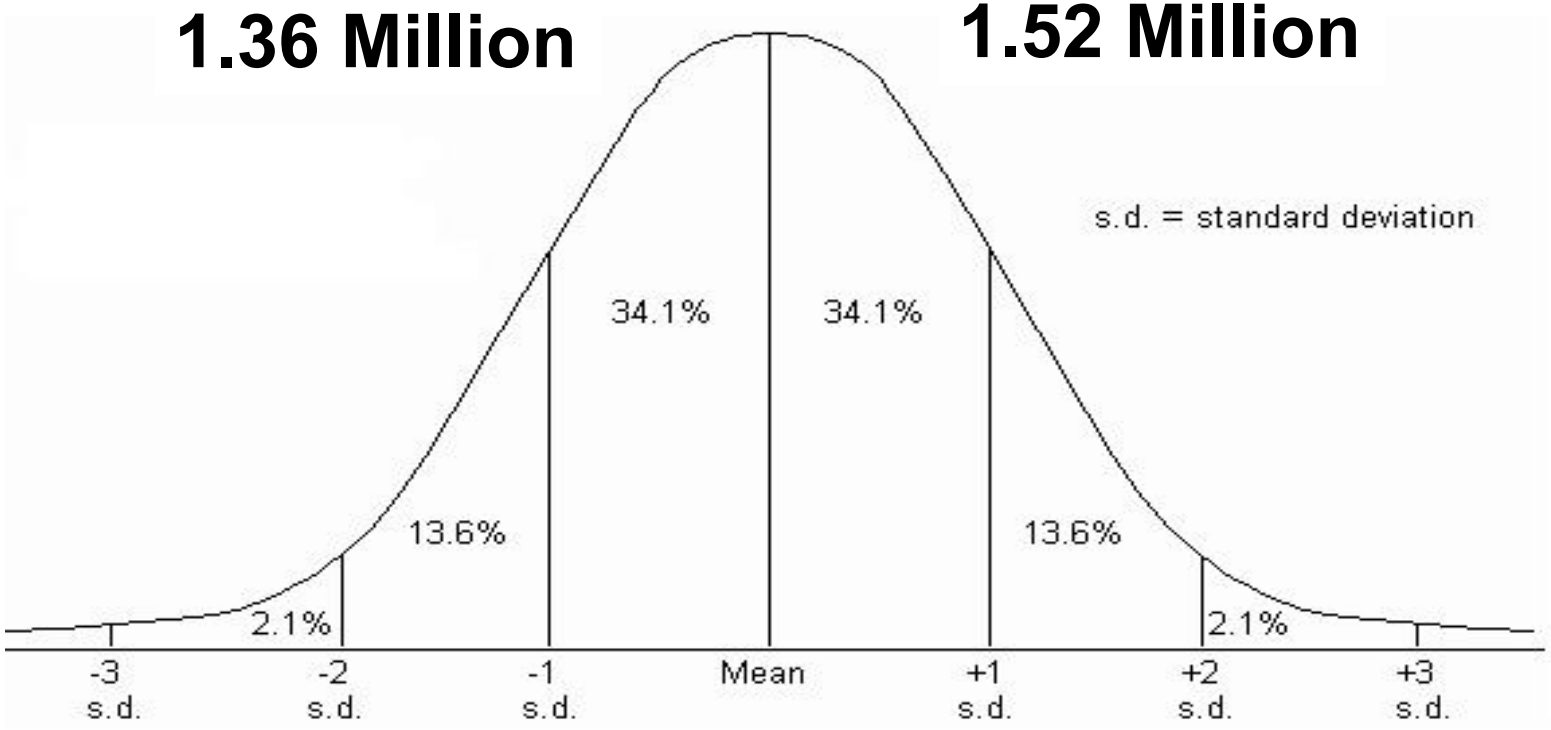
Margin of Error

Uncertainty Interval

1.44 Million

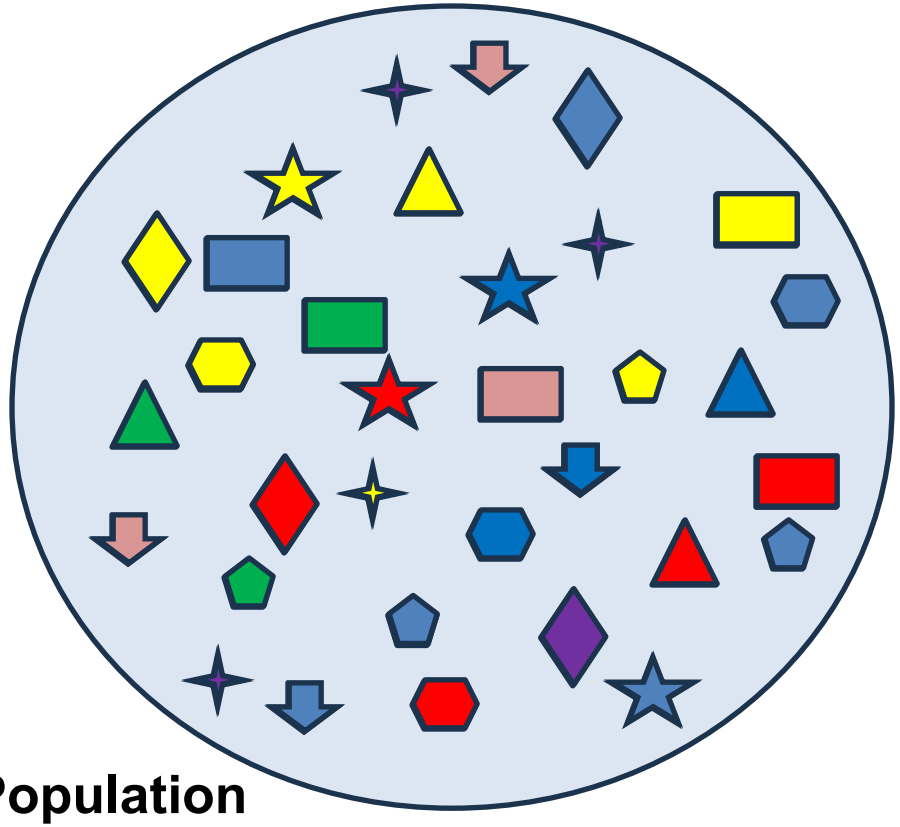
1.36 Million

1.52 Million

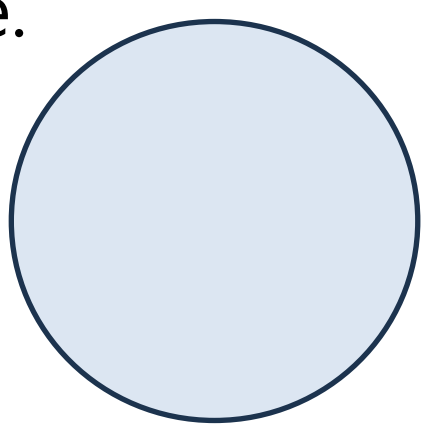


Confidence Interval, at, say 95%

- **Confidence interval**: range of estimates for a value.
- **Confidence level**: proportion of CIs that theoretically contain the true value.



Population

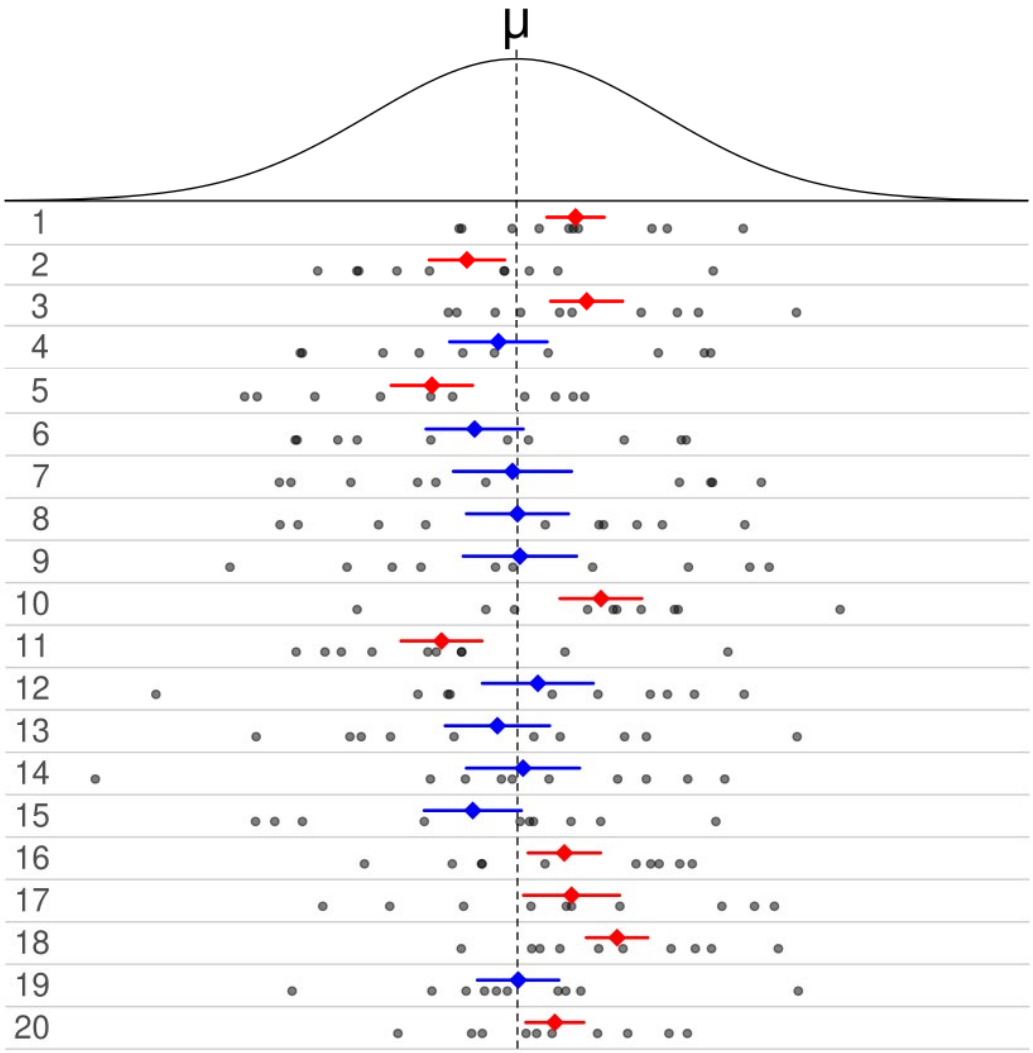


Sample

At 95% (= 19/20) If we sample 20 times, at least 19 will contain the true value that is estimated. 1 will not contain the true value.

Confidence Interval, at, say 50%

- 50% implies half of the samples will not contain the value (mean).
- Red = C.I. of the samples that do not include the correct mean
- Blue = C.I. of the samples that do include the mean



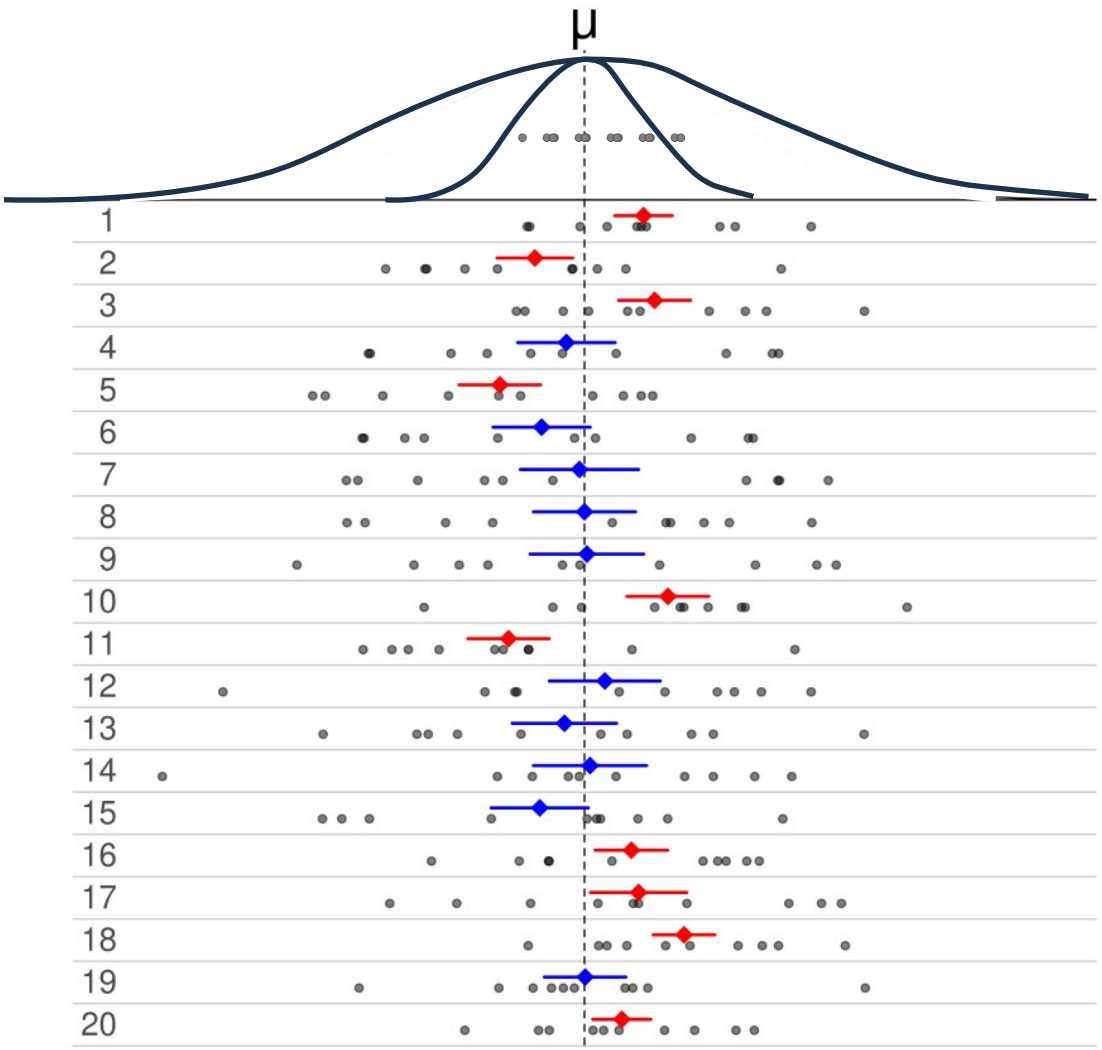
https://en.wikipedia.org/wiki/Confidence_interval

Confidence Interval, Important Factors

- Variability

- Sample Size

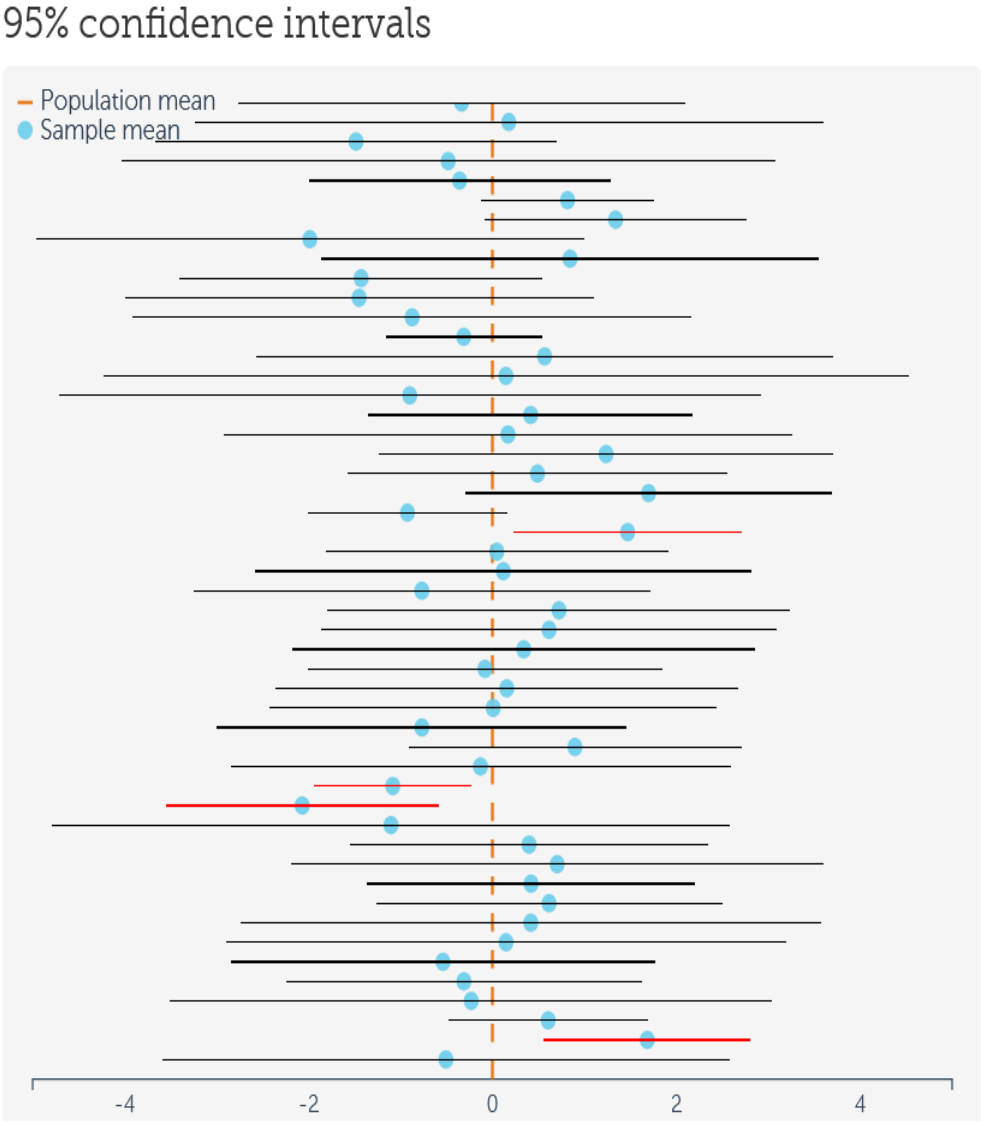
- Confidence Level



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

Confidence interval (CI)

- An interval estimate of a population parameter.
  - observed interval (calculated from observations)
  - For the 95% case, CIs from repeated sampling *should include the parameter, 95% of the time*
  - This is different from: *there is a 95% chance of including the parameter*



https://rpsychologist.com/d3/CI/

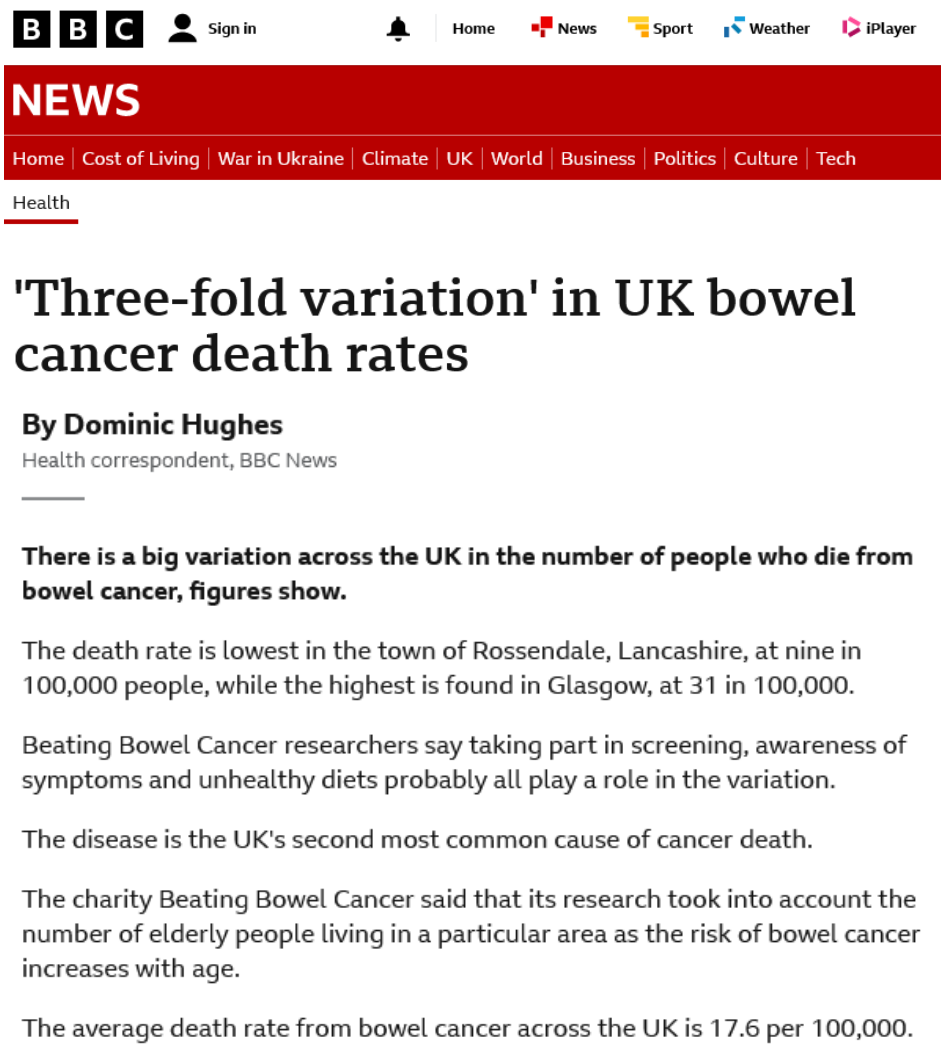
Reliance on models of distributions

- Inferential statistics assume a statistical distribution
- They estimate the **parameters of the distribution** from the sample data and make inferences within certain probability bounds.
  - How likely is the difference due to sampling error?
  - How likely is the difference a real effect?
- We want to use the samples (known) to infer about population (unknown).
- Important not to confuse:
  - Sample parameters (like the sample mean)
  - Population parameters (like the population mean)
  - **And many other factors!**

Example: Bowel Cancer in the UK

- Bowel cancer reported to have three-fold variation in the UK

https://www.bbc.co.uk/news/health-14854019



BBC Sign in Home News Sport Weather iPlayer

NEWS

Home | Cost of Living | War in Ukraine | Climate | UK | World | Business | Politics | Culture | Tech

Health

'Three-fold variation' in UK bowel cancer death rates

By Dominic Hughes

Health correspondent, BBC News

There is a big variation across the UK in the number of people who die from bowel cancer, figures show.

The death rate is lowest in the town of Rossendale, Lancashire, at nine in 100,000 people, while the highest is found in Glasgow, at 31 in 100,000.

Beating Bowel Cancer researchers say taking part in screening, awareness of symptoms and unhealthy diets probably all play a role in the variation.

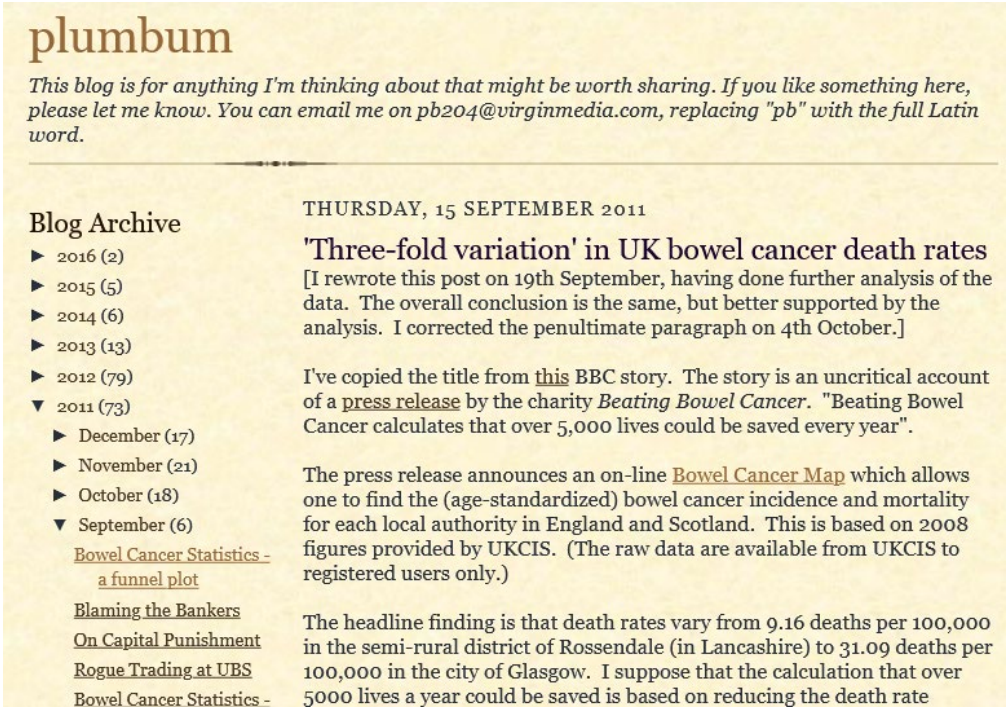
The disease is the UK's second most common cause of cancer death.

The charity Beating Bowel Cancer said that its research took into account the number of elderly people living in a particular area as the risk of bowel cancer increases with age.

The average death rate from bowel cancer across the UK is 17.6 per 100,000.

Example: Bowel Cancer in the UK

- Bowel cancer reported to have three-fold variation in the UK
- Paul Barden investigates the case …



plumbum

This blog is for anything I'm thinking about that might be worth sharing. If you like something here, please let me know. You can email me on pb204@virginmedia.com, replacing "pb" with the full Latin word.

Blog Archive

- ► 2016 (2)
- ► 2015 (5)
- ► 2014 (6)
- ► 2013 (13)
- ► 2012 (79)
- ▼ 2011 (73)
  - ► December (17)
  - ► November (21)
  - ► October (18)
  - ▼ September (6)
    - Bowel Cancer Statistics - a funnel plot
    - Blaming the Bankers
    - On Capital Punishment
    - Rogue Trading at UBS
    - Bowel Cancer Statistics -

THURSDAY, 15 SEPTEMBER 2011

'Three-fold variation' in UK bowel cancer death rates

[I rewrote this post on 19th September, having done further analysis of the data. The overall conclusion is the same, but better supported by the analysis. I corrected the penultimate paragraph on 4th October.]

I've copied the title from this BBC story. The story is an uncritical account of a press release by the charity *Beating Bowel Cancer*. "Beating Bowel Cancer calculates that over 5,000 lives could be saved every year".

The press release announces an on-line Bowel Cancer Map which allows one to find the (age-standardized) bowel cancer incidence and mortality for each local authority in England and Scotland. This is based on 2008 figures provided by UKCIS. (The raw data are available from UKCIS to registered users only.)

The headline finding is that death rates vary from 9.16 deaths per 100,000 in the semi-rural district of Rossendale (in Lancashire) to 31.09 deaths per 100,000 in the city of Glasgow. I suppose that the calculation that over 5000 lives a year could be saved is based on reducing the death rate

https://pb204.blogspot.com/2011/09/three-fold-variation-in-uk-bowel-cancer.html

Example: Bowel Cancer in the UK

- Bowel cancer reported to have three-fold variation in the UK
- Paul Barden investigates the case …
- Data available (a posteriori) from the Guardian

theguardian

Home | UK | World | Sport | Football | Opinion | Culture | Economy | Lifestyle | Fashion

Comment is free

Series: Bad science Previous | Next | Inde

DIY statistical analysis: experience the thrill of touching real data

The story of one man's efforts to re-analyse the stats behind a BBC report on bowel cancer is a heartwarmingly nerdy one

Ben Goldacre
The Guardian, Friday 28 October 2011 22.31 BST

Bowel cancer mortality

By UK local authority, deaths per 100,000

Bowel cancer mortality by region

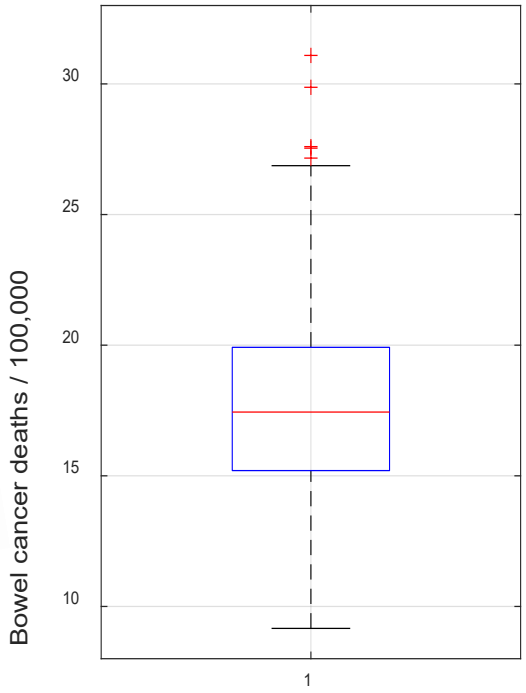
Area	Rate	Population	Number
Shetland Islands	19.15	31332	6
Limavady	21.49	32573	7
Ballymoney	17.05	35191	6
Orkney Islands	29.87	36826	11
Larne	27.54	39942	11
Magherafelt	15.26	45872	7
Carrickfergus	25.94	46261	12
Strabane	11.32	53004	6
Dungannon	22.17	54127	12
Antrim	21.07	56953	12
Banbridge	27.16	58910	16
Richmondshire	20.41	68594	14
Armagh	18.84	69002	13

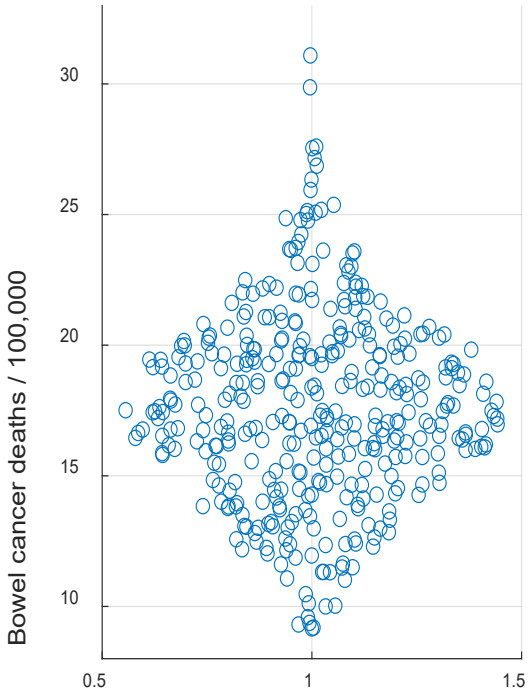
https://www.theguardian.com/commentisfree/2011/oct/28/bad-science-diy-data-analysis

Example: Bowel Cancer in the UK

- Bowel cancer reported to have three-fold variation in the UK
- Paul Barden investigates the case …
- Data available (a posteriori) from the Guardian
- Visualise the data

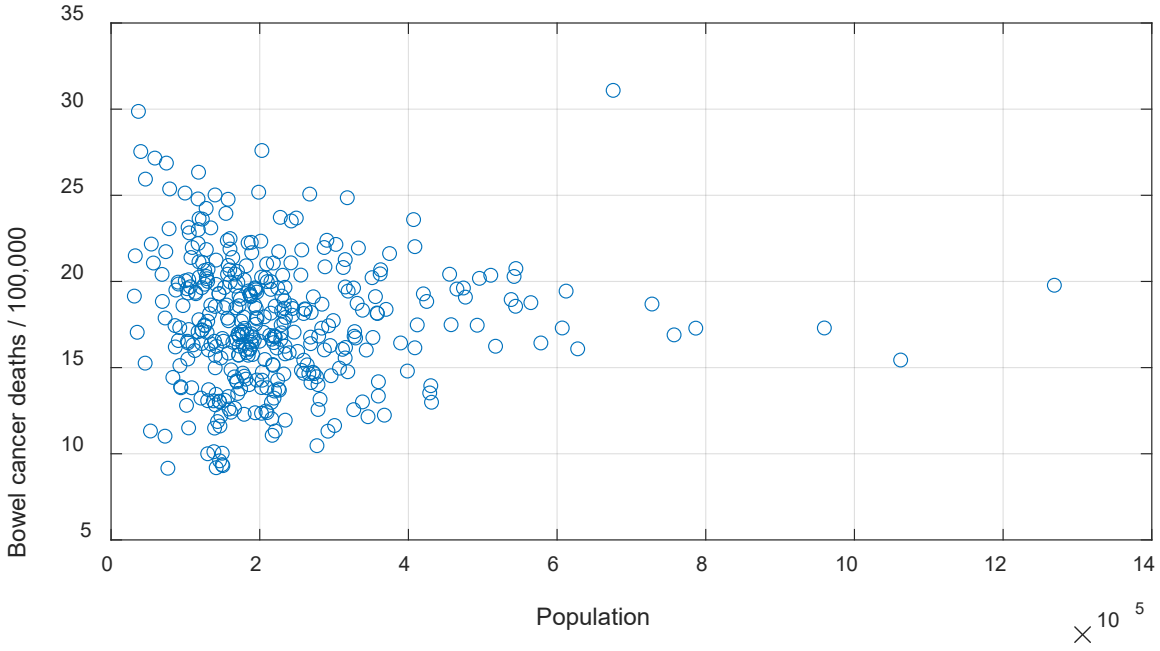
Bowel cancer mortality by region			
Area	Rate	Population	Number
Shetland Islands	19.15	31332	6
Limavady	21.49	32573	7
Ballymoney	17.05	35191	6
Orkney Islands	29.87	36826	11
Larne	27.54	39942	11
Magherafelt	15.26	45872	7
Carrickfergus	25.94	46261	12
Strabane	11.32	53004	6
Dungannon	22.17	54127	12
Antrim	21.07	56953	12
Banbridge	27.16	58910	16
Richmondshire	20.41	68594	14
Armagh	18.84	69002	13



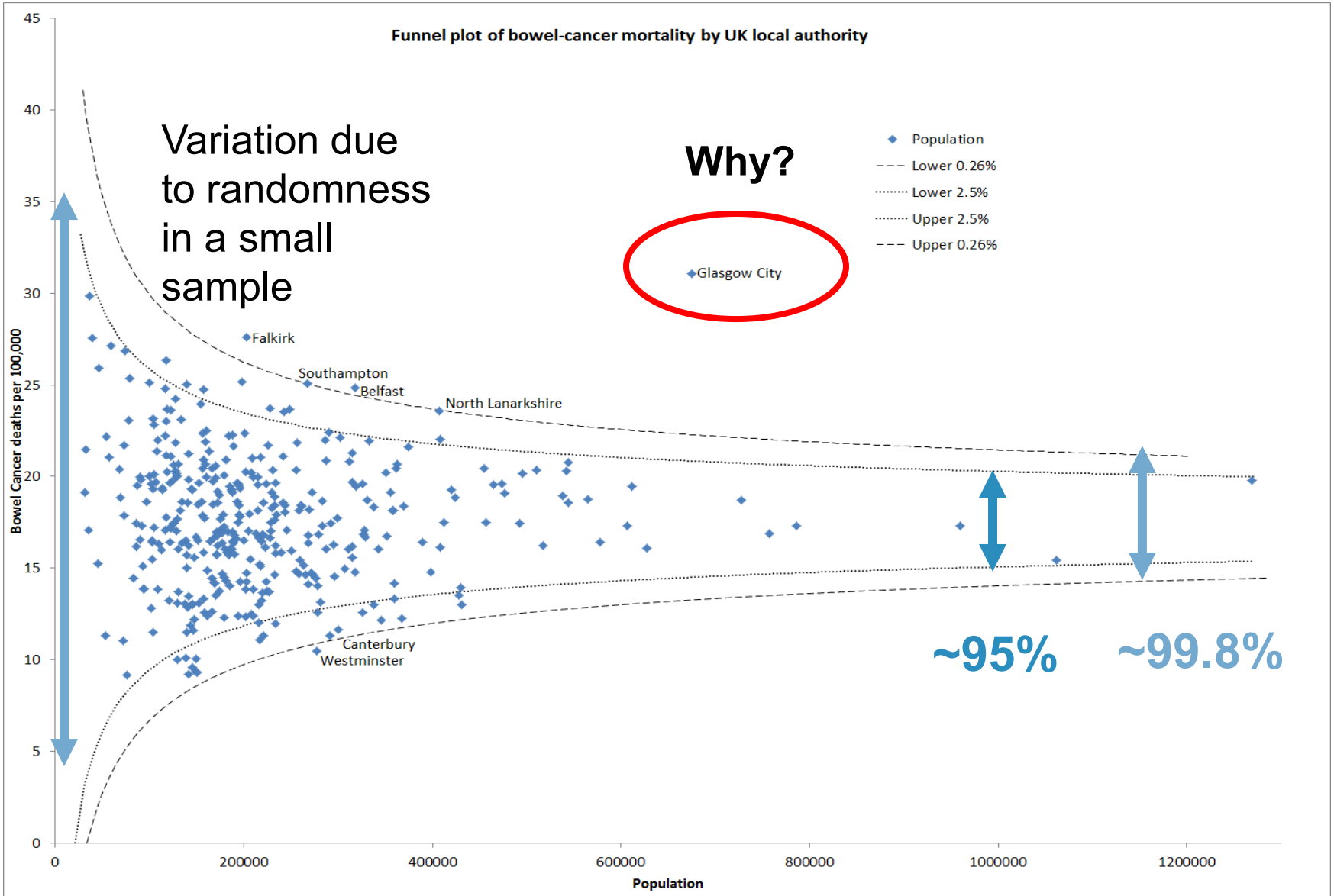


Example: Bowel Cancer in the UK

- Bowel cancer reported to have three-fold variation in the UK
- Paul Barden investigates the case …
- Data available (a posteriori) from the Guardian
- Visualise the data



Example: Bowel Cancer in the UK



https://pb204.blogspot.com/2011/09/im-grateful-to-david-spiegelhalter-of.html

Hypothesis

- So far, we had been checking if the data is correct, if it fits a distribution, how can it be described or if we can generalize from sample to the population.
- We now face a question: *why?*
- To answer we need a hypothesis.
- It is not absolute truth, but rather a working assumption.

Hypothesis

96 languages

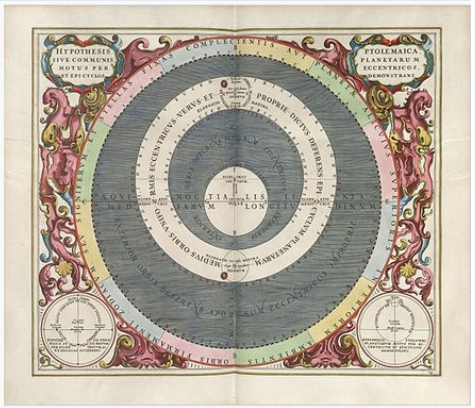
Article Talk | Read View source View history Tools

From Wikipedia, the free encyclopedia

For other uses, see Hypothesis (disambiguation) and Hypothetical (disambiguation).

A **hypothesis** (PL: **hypotheses**) is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it. Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories. Even though the words "hypothesis" and "theory" are often used interchangeably, a scientific hypothesis is not the same as a scientific theory. A working hypothesis is a provisionally accepted hypothesis proposed for further research[1] in a process beginning with an educated guess or thought.[2]

A different meaning of the term *hypothesis* is used in formal logic, to denote the antecedent of a proposition; thus in the proposition "If *P*, then *Q*", *P* denotes the hypothesis (or antecedent); *Q* can be called a consequent. *P* is the assumption in a (possibly counterfactual) *What If* question. The adjective *hypothetical*, meaning "having the nature of a hypothesis", or "being assumed to exist as an immediate consequence of a hypothesis", can refer to any of these meanings of the term "hypothesis".



The hypothesis of Andreas Cellarius, showing the planetary motions in eccentric and epicyclical orbits.

Hypothesis

- So far, we had been checking if the data is correct, if it fits a distribution, how can it be described or if we can generalize from sample to the population.
- We now face a question: *why?*
- To answer we need a hypothesis.
- It is not absolute truth, but rather a working assumption.
- The hypothesis will be based on a statistical model that will consider the model + error (or chance)

Null Hypothesis H_0

- The null hypothesis is a simplified form a model that will be used until there is enough evidence against it.
- No relationship exists between two sets of data or variables being analysed.

› Psychiatry Res. 2003 Nov 30;124(3):177-89. doi: 10.1016/s0925-4927(03)00070-2.

Detection of structural differences between the brains of schizophrenic patients and controls

Vassili A Kovalev [1], Maria Petrou, John Suckling

Affiliations + expand

PMID: 14623069 DOI: 10.1016/s0925-4927(03)00070-2

Abstract

This paper investigates the validity of the null hypothesis: there are no structural differences between the brains of schizophrenic and normal control subjects that manifest themselves in MRI-T(2) data and distinguish the two populations in a statistically significant way. The data used refer to 21 schizophrenic patients and 19 normal controls, matched for age, sex and social background. The methodology used is based on three-dimensional texture analysis, which is used to quantify anisotropy in the data at scales of the order of a few millimetres. These data reject the null hypothesis.

Null Hypothesis H_0

- Experiments will be developed to reject the null hypothesis. If we cannot reject, we must accept it. It does not mean it is true, just that we could not reject it.
- Criminal trial analogy
  - Defendants are assumed to be innocent unless they are proven guilty. No one is found innocent, just not proven guilty.
- There can be an Alternative Hypothesis H_1.

NHST: Null Hypothesis Significance Testing

- Commonly-used process for determining whether an effect measured in a **sample** is likely to exist in the **population**, within certain bounds:
  - Does a drug have an effect on the symptom severity.
- Helps determine whether there is statistical support for the effect observed in the sample to also be in the population.
  - The bigger the sample size and effect, the more likely this is.

NHST Strengths

- It is a **standard methodology** that **many** people use, based on **rigorous statistical theory.**
- It accounts for the fact that we are **sampling from a population** and helps us **judge whether there is enough evidence for an effect**.
- Widely used in psychology, medicine, biology, …

Null Hypothesis H_0

- Possible cases of the H_0 and H_1: TP, TN, FP, FN.
- Level at which the hypothesis will be supported.
- Level of risk (α) to conclude H_0 is not correct when it is. Most common α=0.05 but can be higher or lower.
- Level of risk (β) to conclude H_1 is not correct when it is. (1- β) is called Power. Common $\beta = 0.15, 0.2$

α,β: Levels of risk we are wiling to accept

		Reality	
		H_0 is true	H_1 is true
Decision	H_0 is true	True Positive $1-\alpha$	False Negative **Type II Error**
	H_1 is true	False Positive **Type I Error**	True Negative $1-\beta$

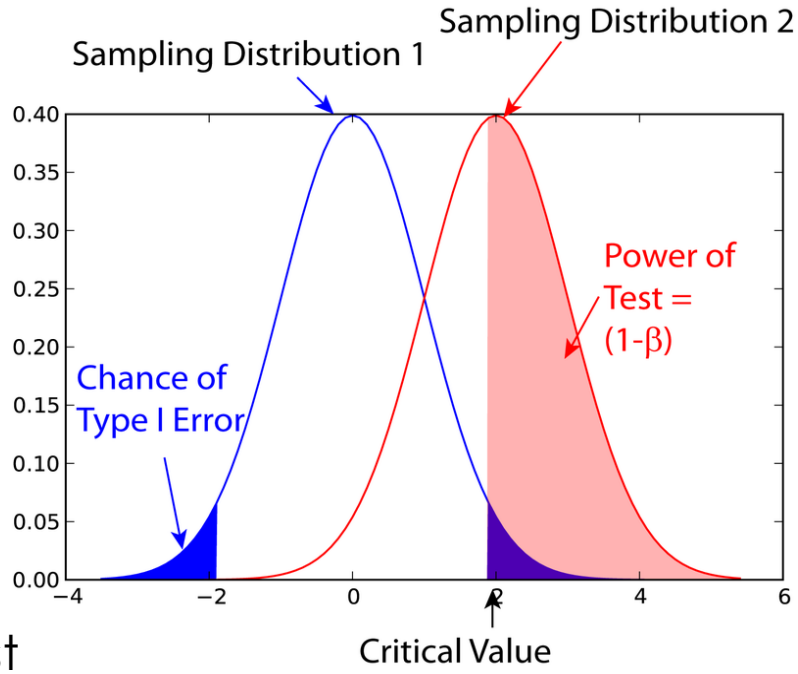
Null Hypothesis H_0

- Criminal trial analogy
  - Defendants are assumed to be innocent unless they are proven guilty. No one is found innocent, just not proven guilty.
- Type I error:
  - falsely convict an innocent
- Type II error:
  - find a criminal as not guilty
- Note that rejecting the null hypothesis does not necessarily mean the Alternative Hypothesis is true, it's just an indication of the amount of evidence.

Lady Justice statue on the top of the court building

NHST: Null Hypothesis Significance Testing

- **Hypothesis** (**H_1**, or the "Alternative Hypothesis"): *there is an effect*
  - treatment with drug **affects** the severity of the symptoms
- **Null hypothesis (H_0):** *there is not an effect*
  - treatment with drug reduces **has no effect** on the severity of the symptoms
- Test the *difference* in means between two samples



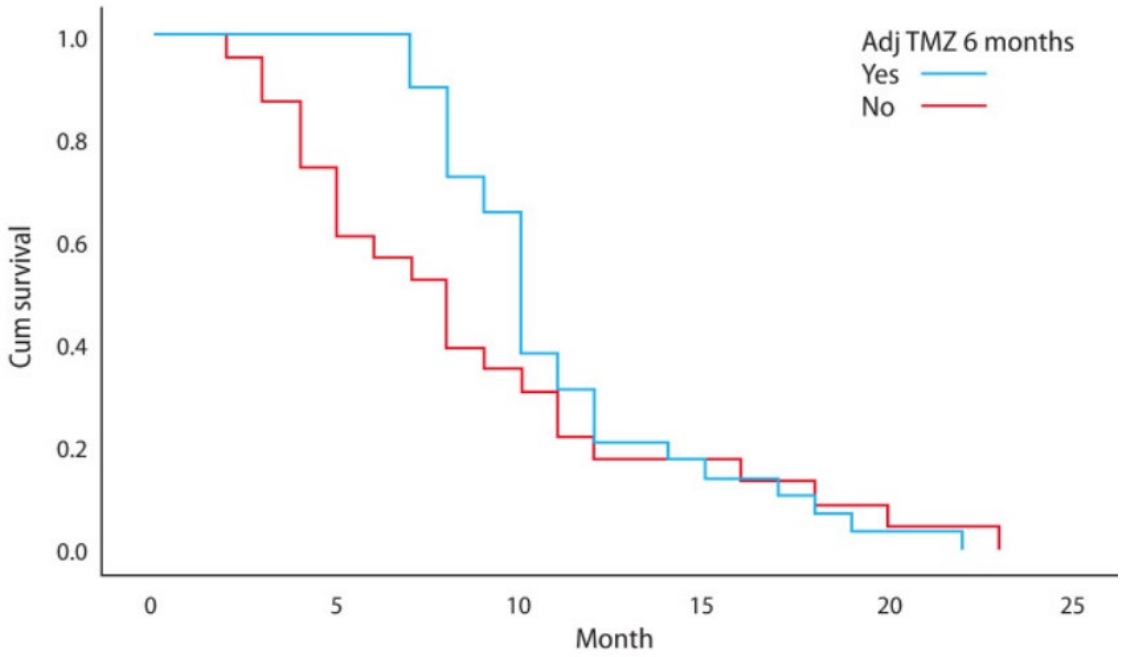
https://en.wikipedia.org/wiki/Power_of_a_test

p-values and significance

- Significance level is the probability that rejection of the null hypothesis is true
  - Alpha: Conventionally 5% (0.05) is considered "significant"
- The p-value is the probability that obtaining the observed results (or results more extreme) if H_0 is true
  - So need small values to reject the null hypothesis
- If a mean difference is statistically significant, then the effect is likely to be present in the population
  - (Given the assumptions)

Effect of intervention

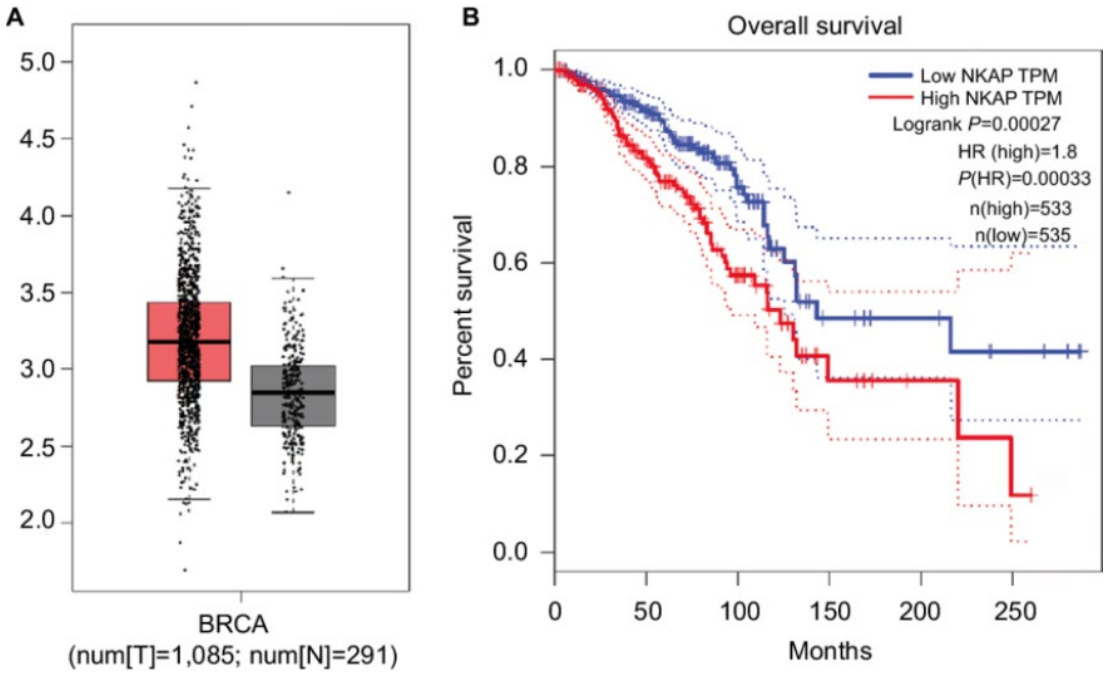
- The null hypothesis is tested to compare two populations, many times of the influence of an intervention (treatment) against nothing (or a placebo), or between variants.



Soniya Mohammed, Survival and quality of life analysis in glioblastoma multiforme with adjuvant chemoradiotherapy: a retrospective study Rep Pract Oncol Radiother. 2022; 27(6): 1026–1036.

Effect of intervention

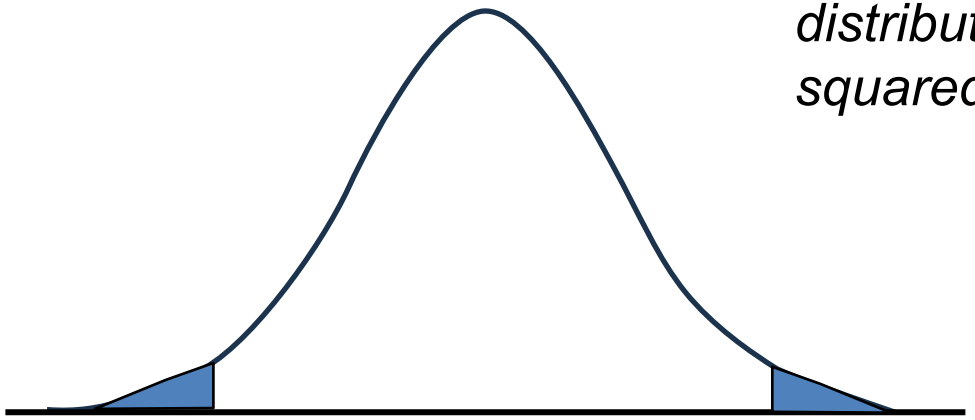
- The null hypothesis is tested to compare two populations, many times of the influence of an intervention (treatment) against nothing (or a placebo), or between variants.



Jiangtao Liu, NKAP functions as an oncogene and its expression is induced by CoCl2 treatment in breast cancer via AKT/mTOR signaling pathway, Cancer Manag Res. 2018; 10: 5091–5100.

p-value

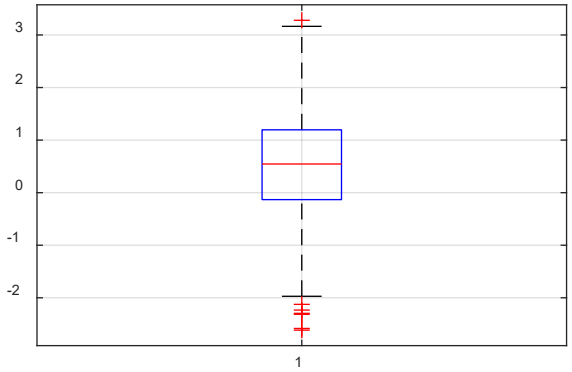
- *p*-value: evidence for a hypothesis
  - The p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.
  - In practice, a measure to summarise the probability is the tail area of a distribution.

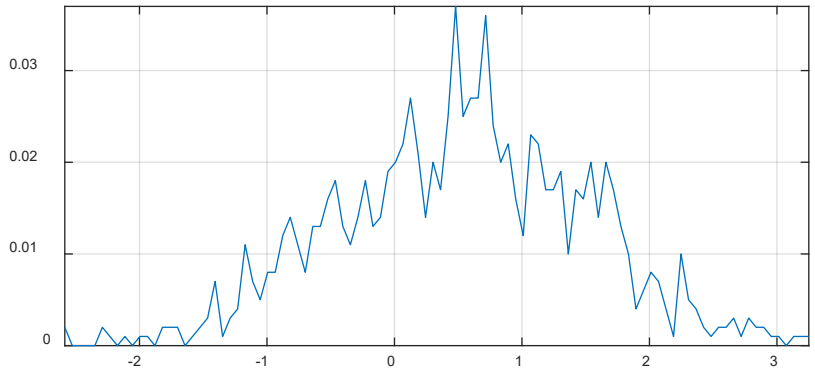


Distribution*: Normal, t-distribution, Binomial, Chi-squared, …*

p-value: example

- Does the following sample come from a Normal distribution $\mu = 0$? **H_0: Data is normal with $\mu = 0$.** Apply a t-test to check difference in means.





William Sealy Gosset, who developed the "*t*-statistic" and published it under the pseudonym of "Student"


```
a = 0.5+randn(1000,1);

[h,p,c]    = ttest(a)
h          = 1
p          = 3.5359e-56
c          = 0.4597    0.5809
```

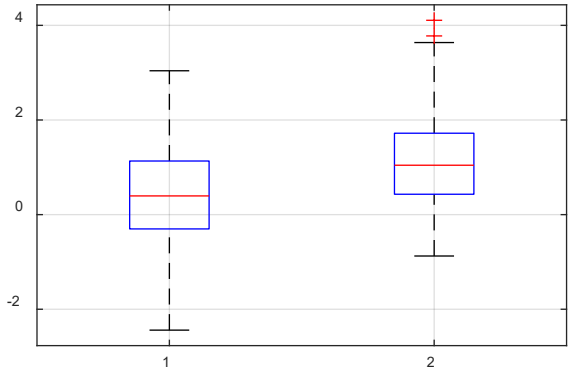
We can reject H_0

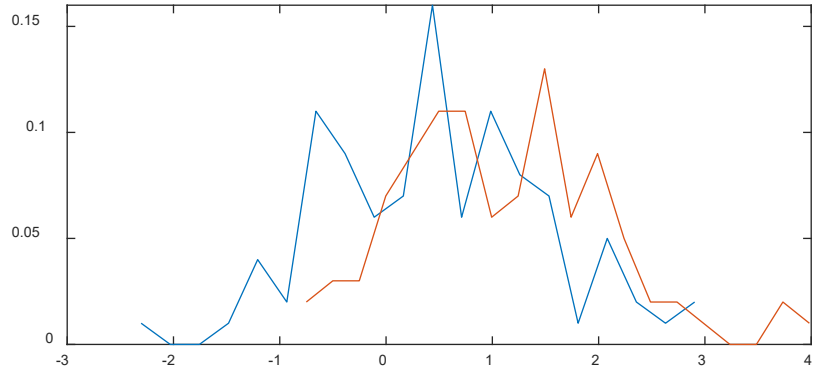
$p<0.05$

Confidence Interval

p-value: example

- **H_0: Data in vectors (a,b) come from independent random samples from normal distributions with equal means.**

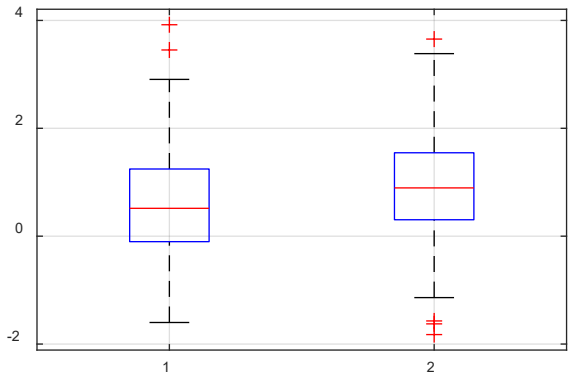


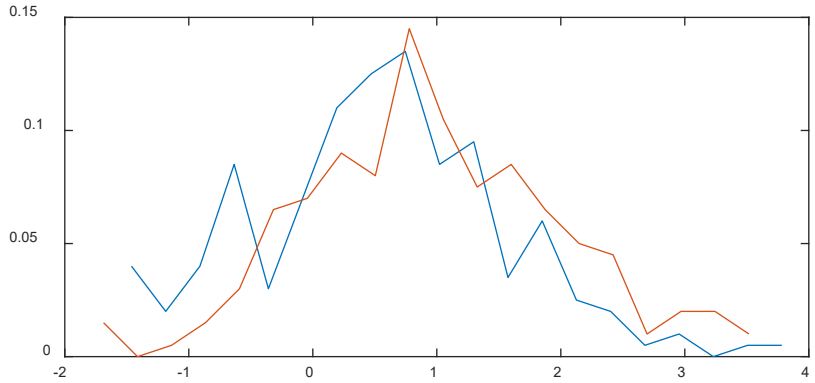



```
a=0.5+randn(100,1);
b=1+randn(100,1);
[h,p,c]    = ttest2(a,b)
h          = 0                                  We cannot reject H0
p          = 0.0538                             p>0.05
c          = -0.6044     0.0050                 Confidence Interval
```

p-value: example

- **H_0: Data in vectors (a,b) come from independent random samples from normal distributions with equal means.**






```
a=0.5+randn(200,1);
b=1+randn(200,1);
[h,p,c]    = ttest2(a,b)
h          = 1
p          = 7.3033e-05
c          = -0.6039    -0.2065
```

We can reject H_0

p<0.05

Confidence Interval

p-value

- Sample size can be a huge influence, how big should our samples be?
- Statistical Software
  - G*Power
  - R
  - SPSS
  - Stata
  - …

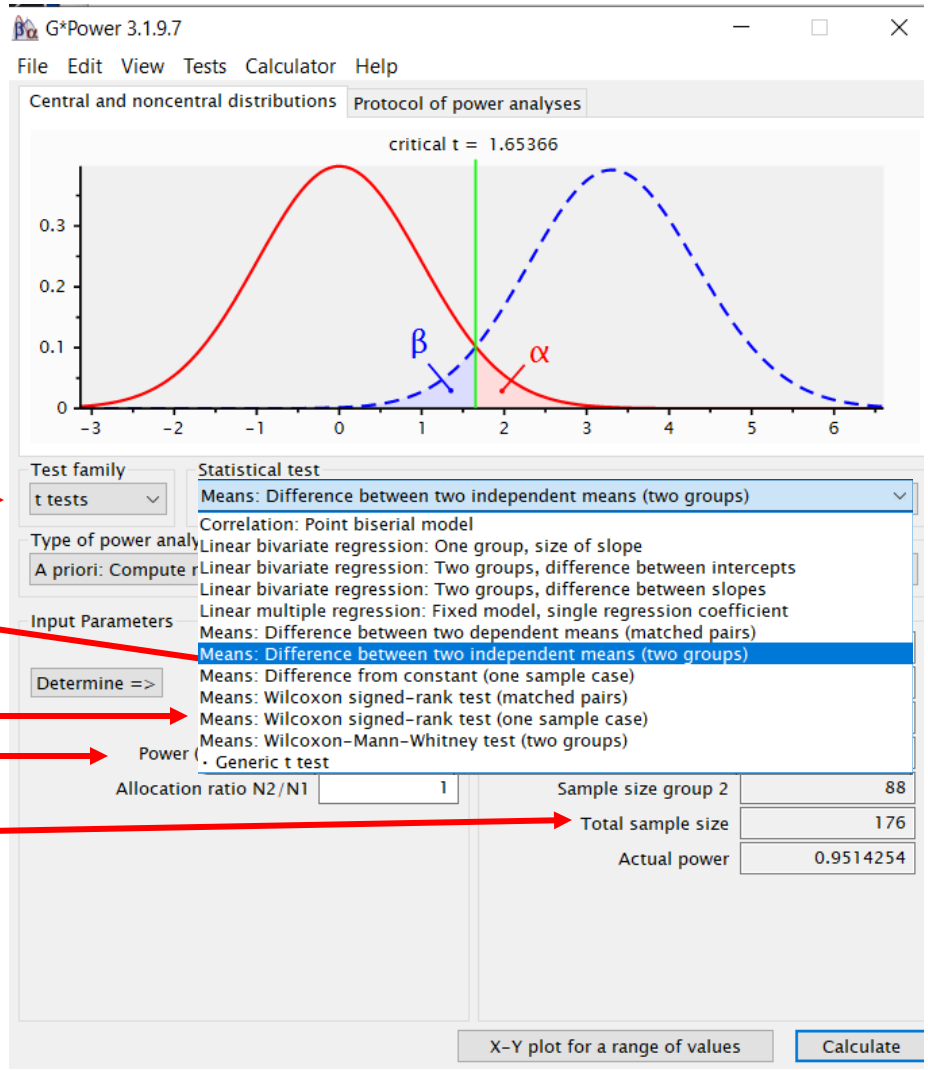
Type of test

??

α

β

Sample size



Effect Sizes

- Measure of the strength of an effect
  - mean
  - differences between means
  - correlations
  - (p-values don't indicate effect size)
- Their interpretation requires informed judgment in context
  - Is the size of the effect important?

Cohen's *d*

- A standardised effect size
- Standardized difference between sample means

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

sample means

standard deviation

- .. where, s is the pooled standard deviation

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

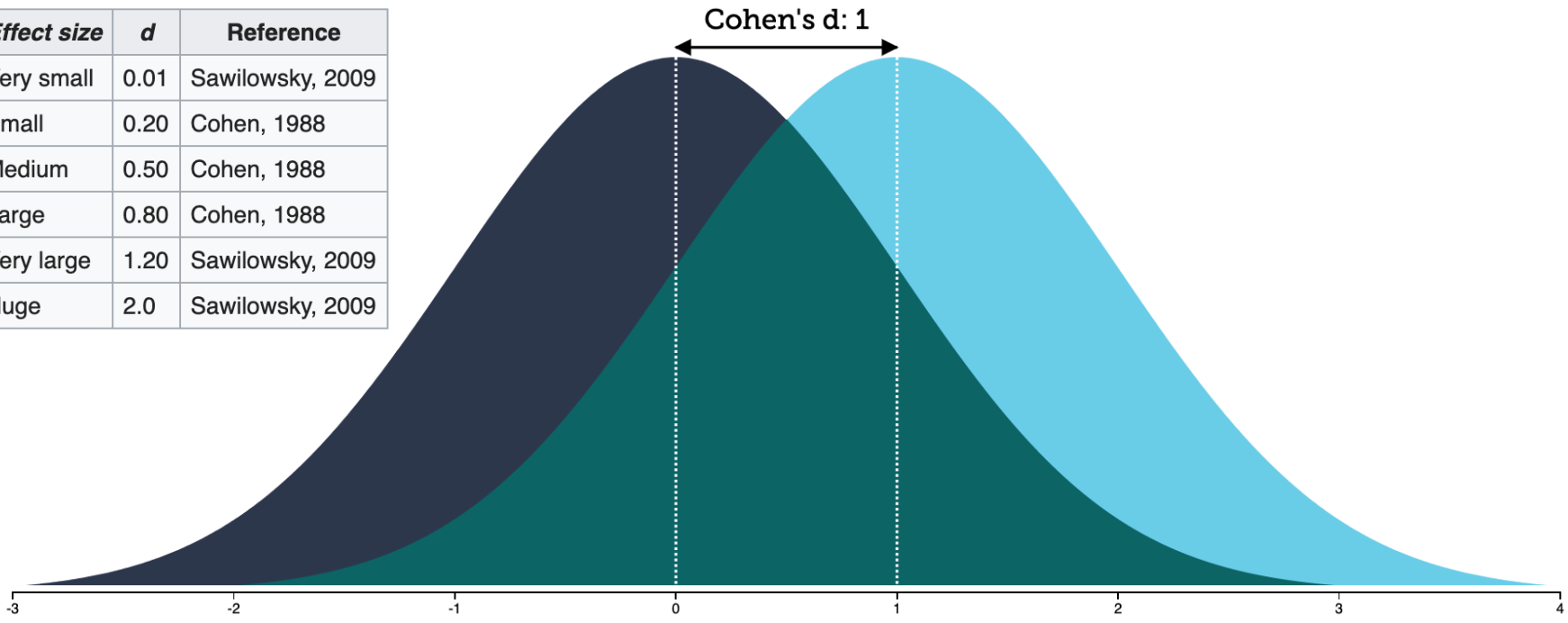
- .. and variance for the groups can be computed as:

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

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

Cohen's *d*

Effect size	*d*	Reference
Very small	0.01	Sawilowsky, 2009
Small	0.20	Cohen, 1988
Medium	0.50	Cohen, 1988
Large	0.80	Cohen, 1988
Very large	1.20	Sawilowsky, 2009
Huge	2.0	Sawilowsky, 2009



https://en.wikipedia.org/wiki/Effect_size#Cohen's_d

https://rpsychologist.com/d3/cohend/

Interpretation

Cohen's U_3

% Overlap

Probability of Superiority

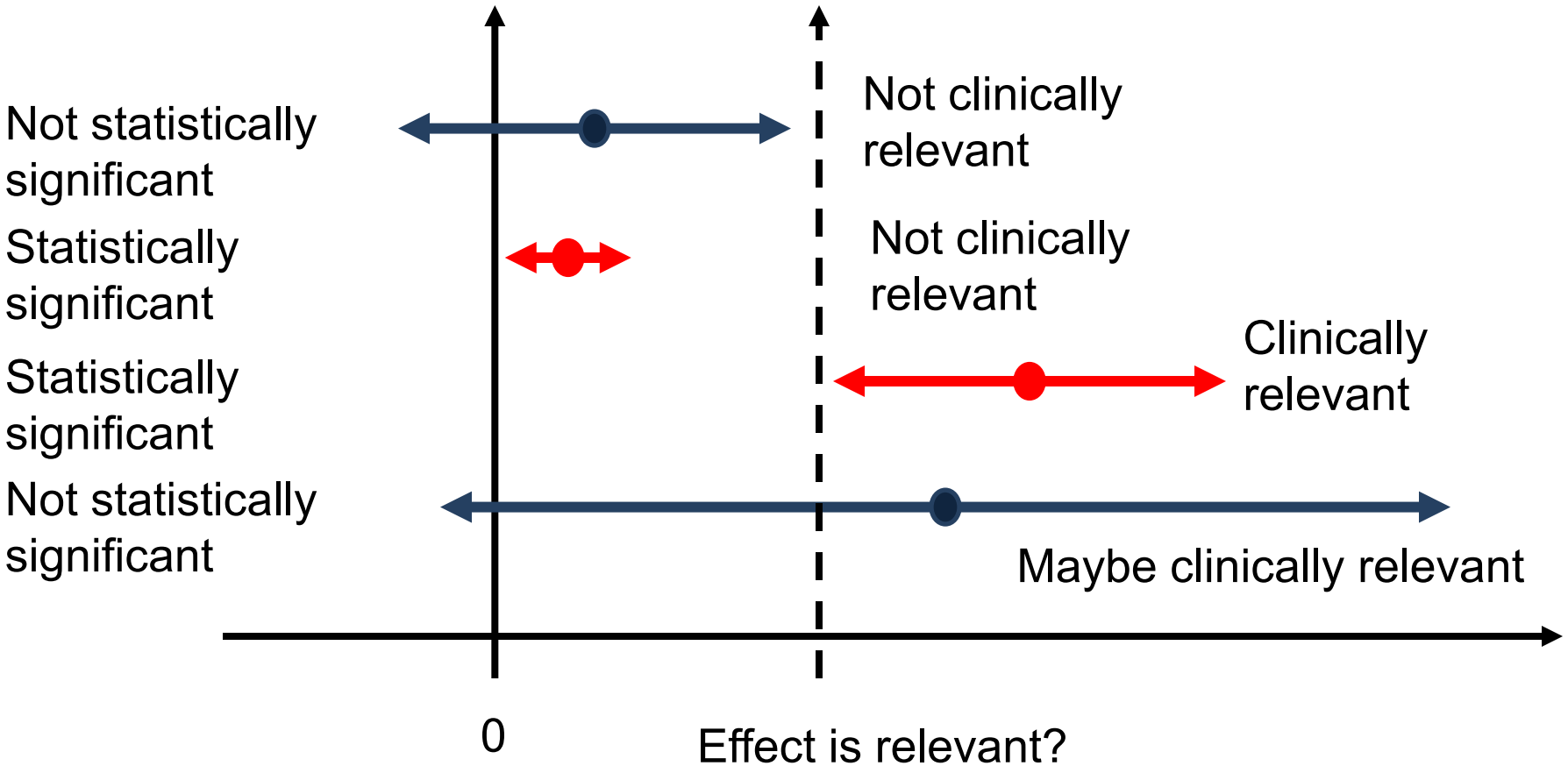
Number Needed to Treat[1]

A Common Language Explanation

With a Cohen's *d* of 1, 84 % of the treatment group will be above the mean of the control group (Cohen's U_3), 62 % of the two groups will overlap, and there is a 76 % chance that a person picked at random from the treatment group will have a higher score than a person picked at random from the control group (probability of superiority). Moreover, in order to have one more favorable outcome in the treatment group compared to the control group we need to treat 2.8 people. This means that if 100 people go through the treatment, 36.3 more people will have a favorable outcome compared to if they had received the control treatment[1].

Effect Size: why it matters

- Imagine a series of clinical tests (drug, intervention,…)



Ceyhan Ceran Serdar, Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies, Biochem Med (Zagreb). 2021 Feb 15; 31(1): 010502.

CAUTION ON RELYING ON STATISTICAL SIGNIFICANCE

Multiple Testing

- Many significance tests carry a ***significant*** danger!
  - Assume there is a drug that does not work
  - Carry experiments with $\alpha = 0.05$

H_0 is true		
True Positive	$1-\alpha$	
False Positive		

5% chance

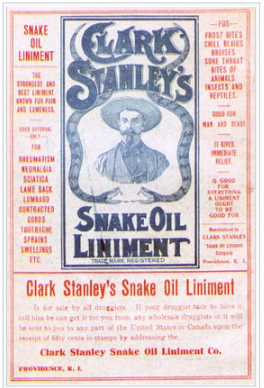
H_0 is true		
True Positive	$1-\alpha$	
False Positive		

5% chance

…

H_0 is true		
True Positive	$1-\alpha$	
False Positive		

5% chance

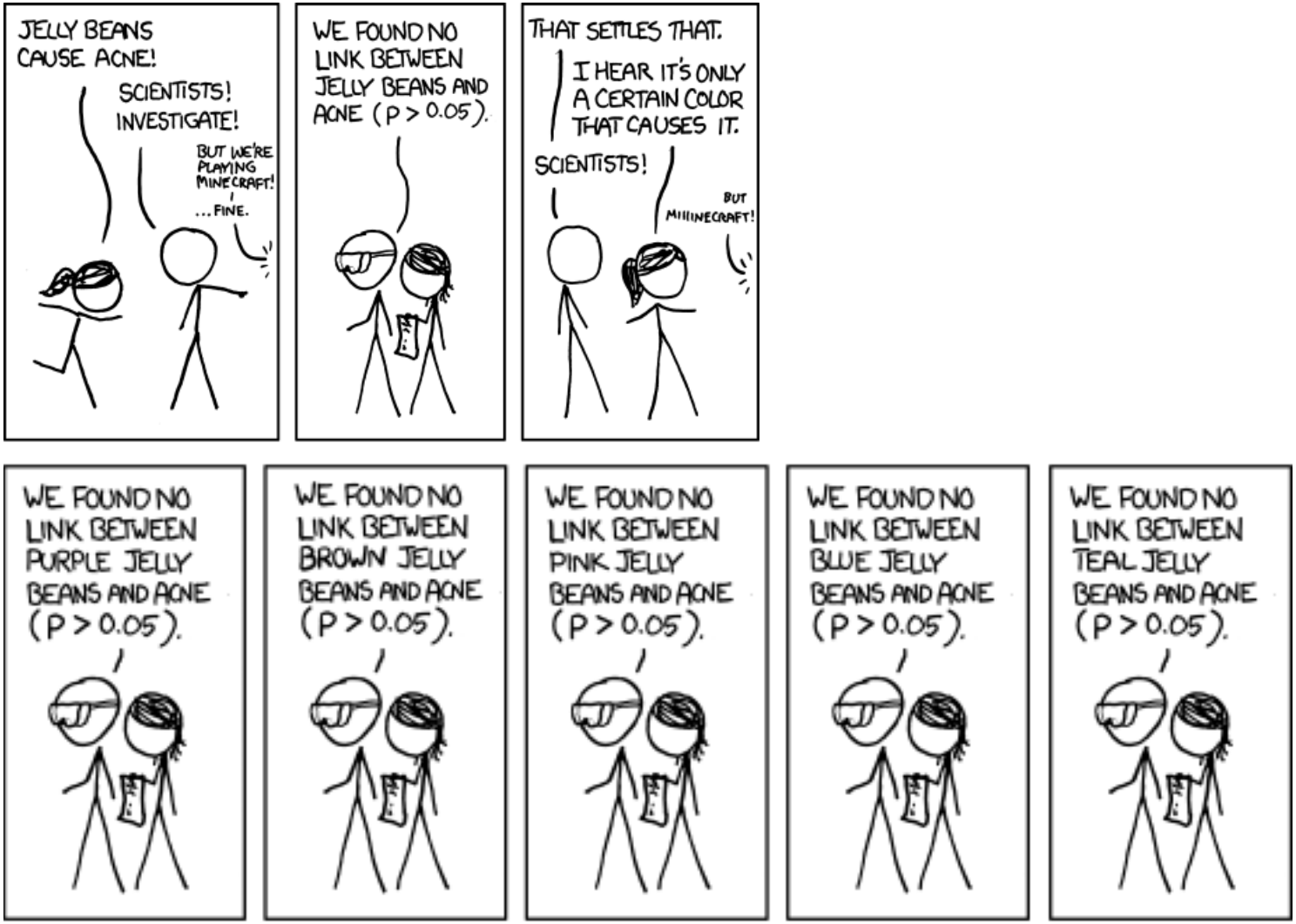


Clark Stanley's Snake Oil


Dalby's Carminative, Daffy's Elixir and Turlington's Balsam of Life bottles dating to the late 18th and early 19th centuries. These "typical" patent or quack medicines were marketed in very different, and highly distinctive, bottles. Each brand retained the same basic appearance for more than 100 years.

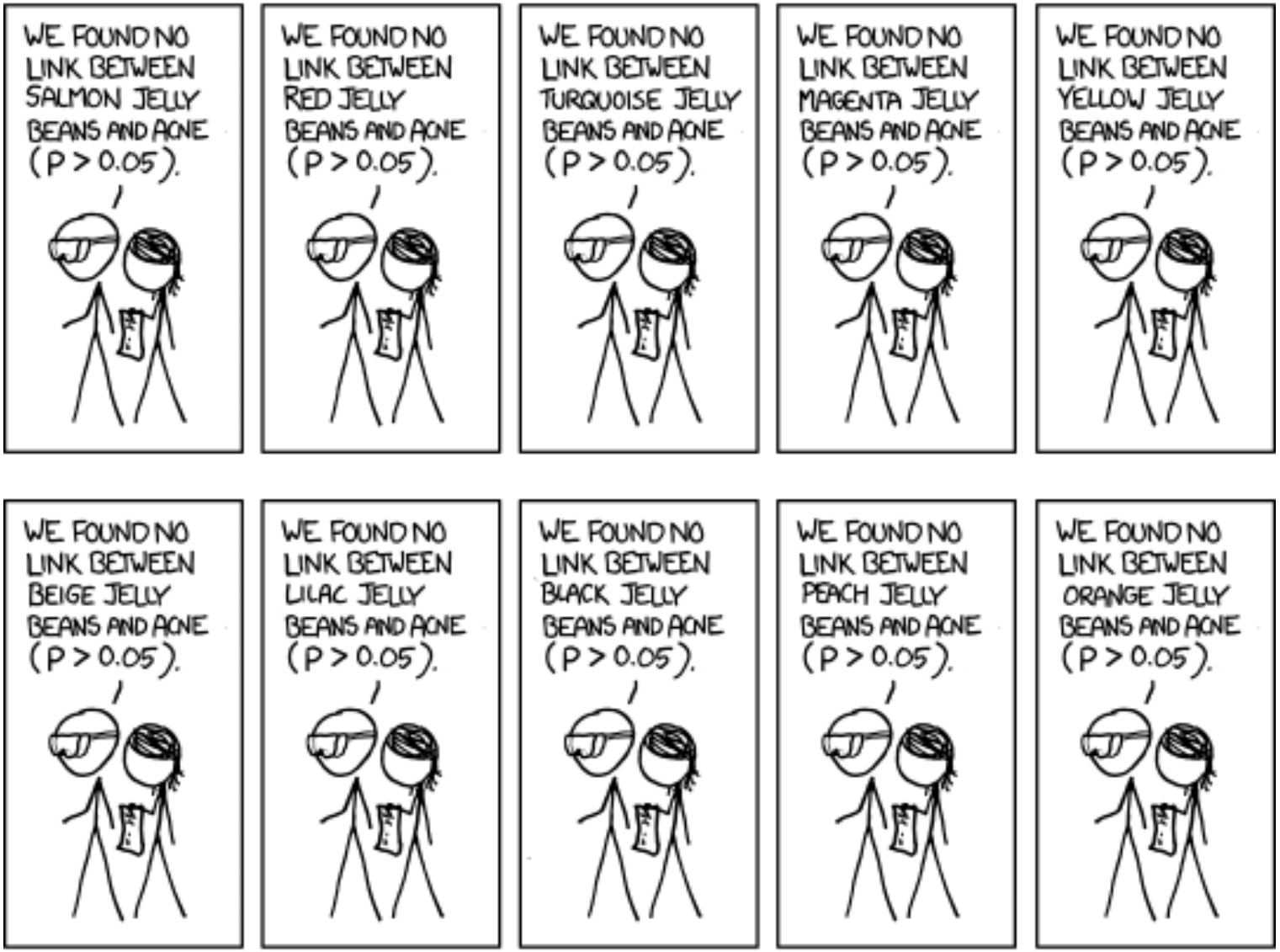
https://en.wikipedia.org/wiki/Quackery

Multiple Testing: cartoon version



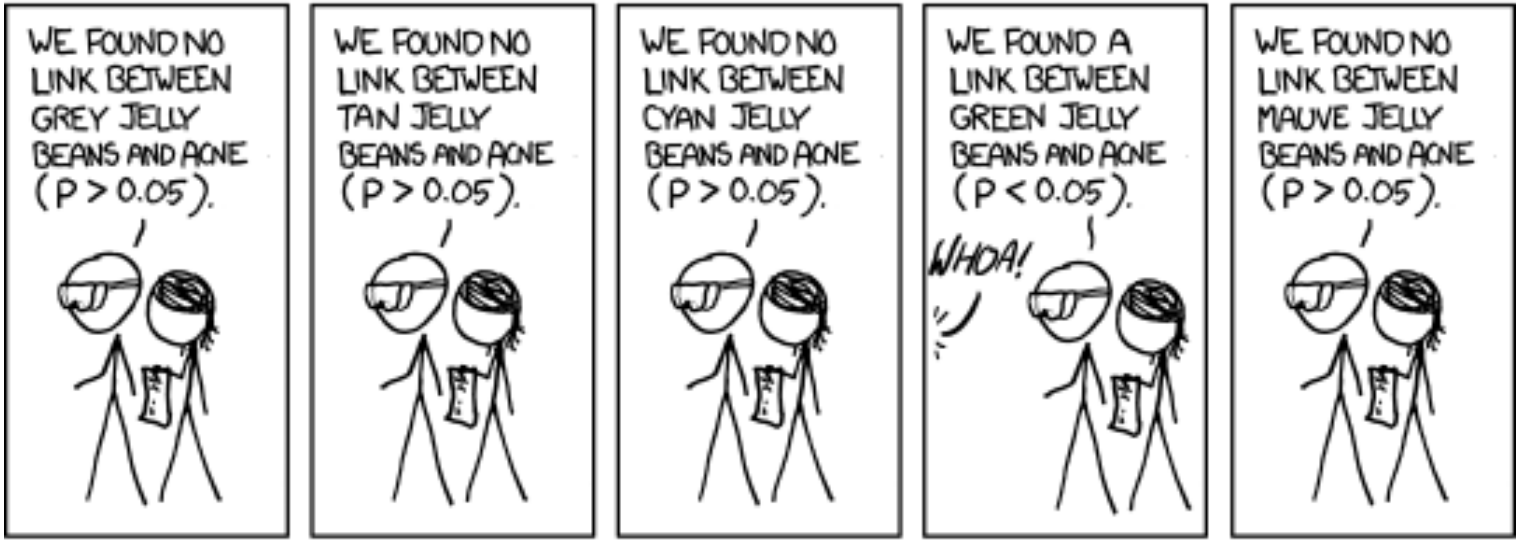
https://xkcd.com/882/

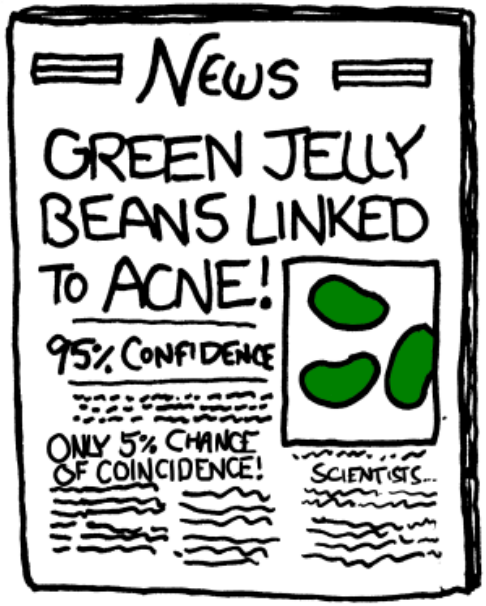
Multiple Testing: cartoon version



https://xkcd.com/882/

Multiple Testing: cartoon version





https://xkcd.com/882/

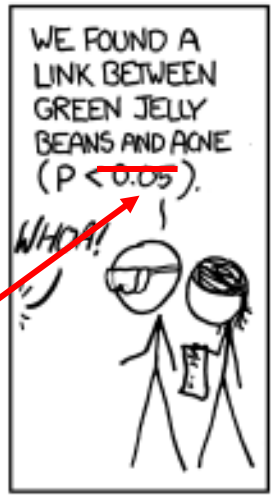
Multiple Testing: Bonferroni Correction

- We rejecting the null hypothesis if the likelihood of the observed data under the null hypotheses is lower than a pre-defined α (0.05, 0.01, 0.001).
- The Bonferroni correction compensates for the increase of multiple testing, if n tests are conducted:

$\alpha = 0.05$ $\alpha = 0.05 / n$

$\alpha = 0.05 / 20$

$\alpha = 0.0025$



Controversy

Controversy

Comment > Psychol Sci. 2014 Jun;25(6):1289-90. doi: 10.1177/0956797614525969. Epub 2014 Mar 6.

Why hypothesis tests are essential for psychological science: a comment on Cumming (2014)

Richard D Morey [1], Jeffrey N Rouder [2], Josine Verhagen [3], Eric-Jan Wagenmakers [3]

Affiliations + expand

PMID: 24604147 DOI: 10.1177/0956797614525969

No abstract available

FULL TEXT

Sage Jc

ACTIONS

Col

SHARE

Comment in

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

The ASA Board was also stimulated by highly visible discussions over the last few years. For example, ScienceNews (Siegfried 2010) wrote: "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation." A November 2013, article in Phys.org Science News Wire (2013) cited "numerous deep flaws" in null hypothesis significance testing. A ScienceNews article (Siegfried 2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on *p*-values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provides the community a service. The intended audience would be researchers, practitioners, and science writers who are not primarily statisticians. Thus, this statement would be quite different from anything previously attempted.

The Board tasked Wasserstein with assembling a group of

aps
ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2021, Vol. 16(3) 639–648
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1745691620958012
www.psychologicalscience.org/PPS
SAGE

The Practical Alternative to the *p* Value Is the Correctly Used *p* Value

Daniël Lakens
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology

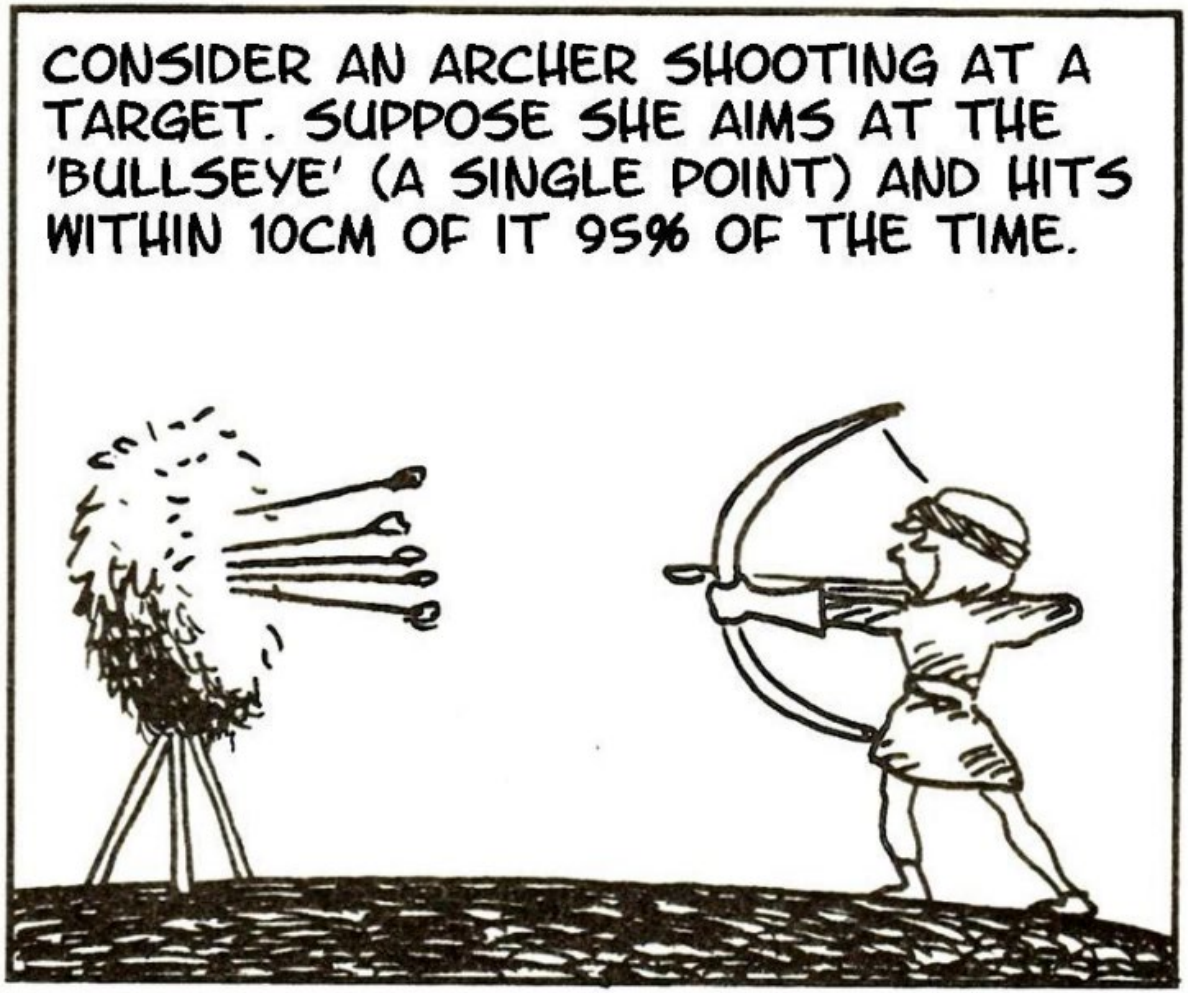
Abstract

Because of the strong overreliance on *p* values in the scientific literature, some researchers have argued that we need ...race practical alternatives. When proposing alternatives to *p* values statisticians often ... whereby they declare which statistic researchers really "want to know." Instead of ...t to know, statisticians should teach researchers which questions they can ask. In ... question they are most interested in will be the *p* value. As long as null-hypothesis ...chers have suggested including minimum-effect tests and equivalence tests in our ...s have the potential to greatly improve the questions researchers ask. If anyone

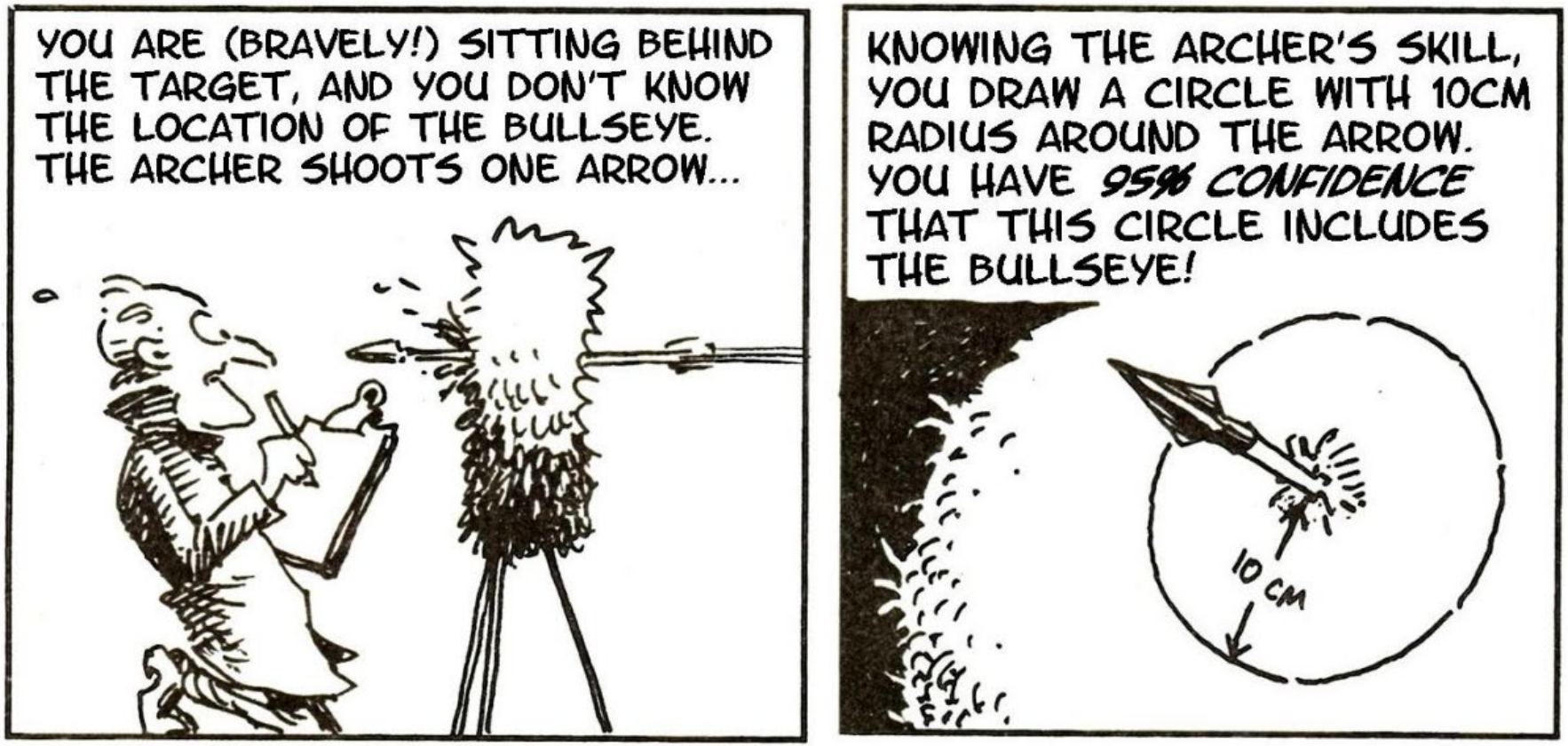
Two main approaches

- Frequentist
  - Make inferences by repeated sampling
  - Report whether the inference in the sample is reflective of the inference for the population with a given probability (p-values)
  - Usually yes/no significant or not
    - E.g. Null hypothesis testing and statistical significance
- Bayesian
  - We have a prior belief about population
  - We run experiments to update our belief
  - "The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective" by John K. Kruschke & Torrin M. Liddell (https://doi.org/10.3758/s13423-016-1221-4)

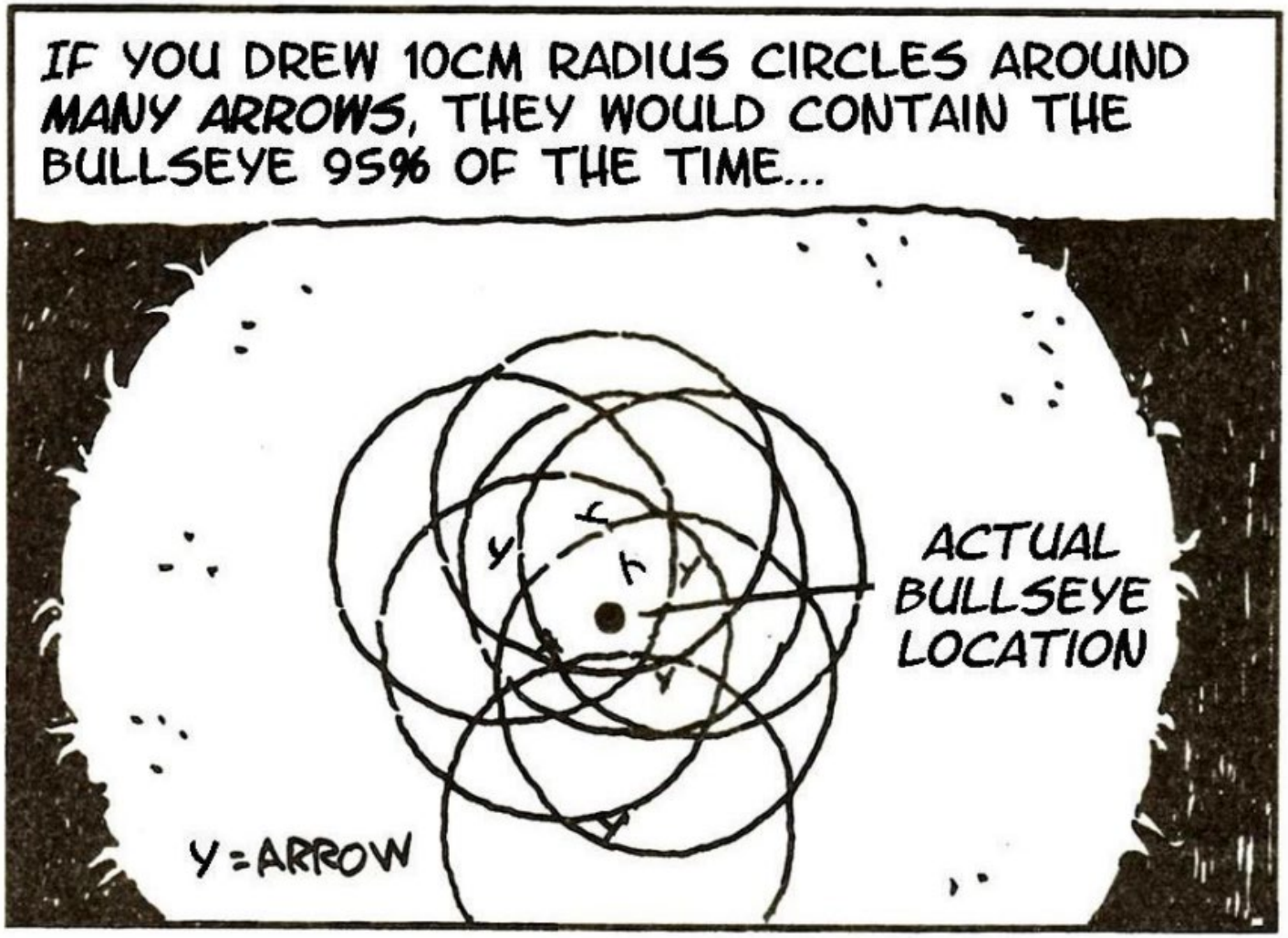
Summary: "The Cartoon Guide to Statistics", by Gonick & Smith



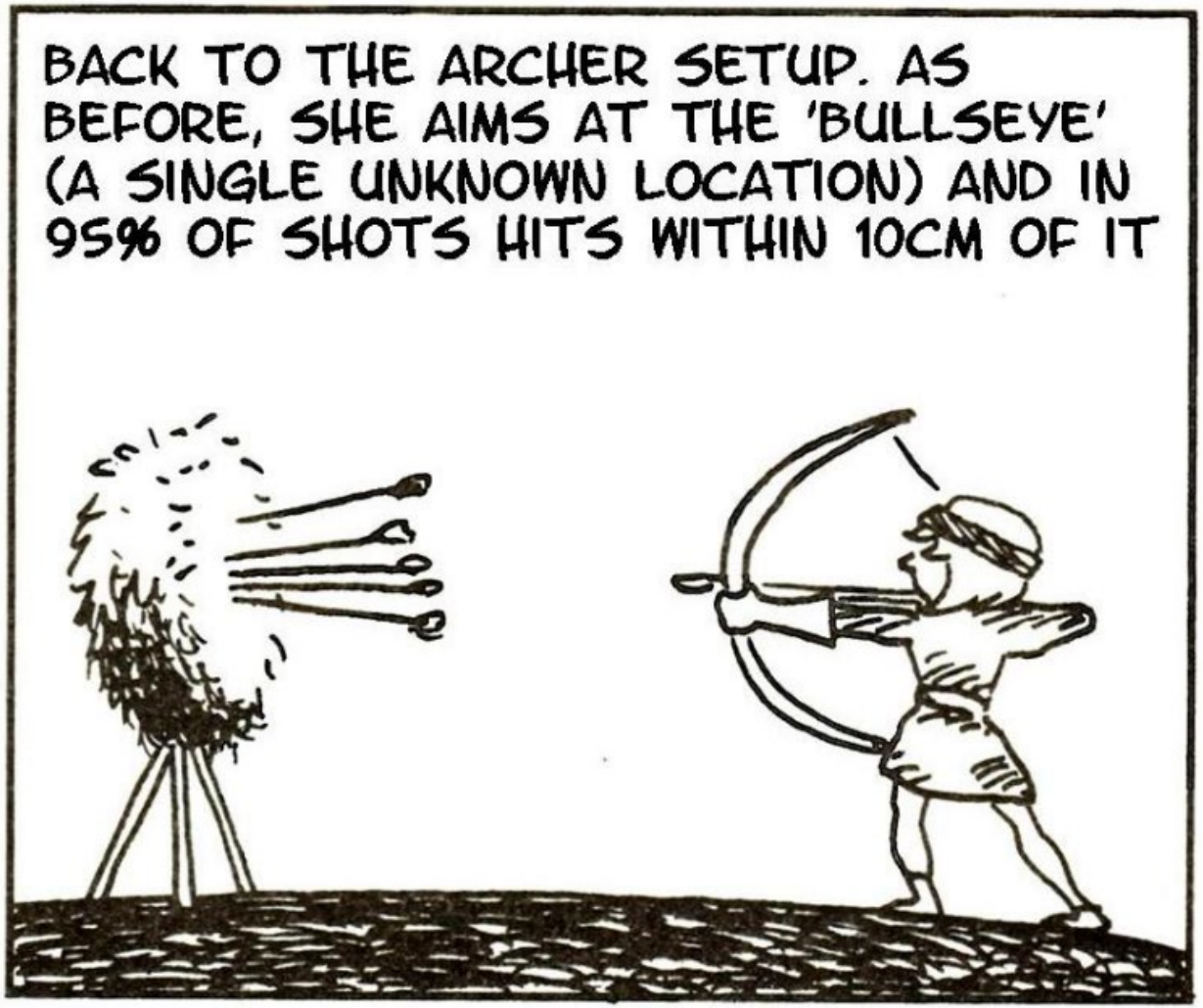
Summary: "The Cartoon Guide to Statistics", by Gonick & Smith



Summary: "The Cartoon Guide to Statistics", by Gonick & Smith

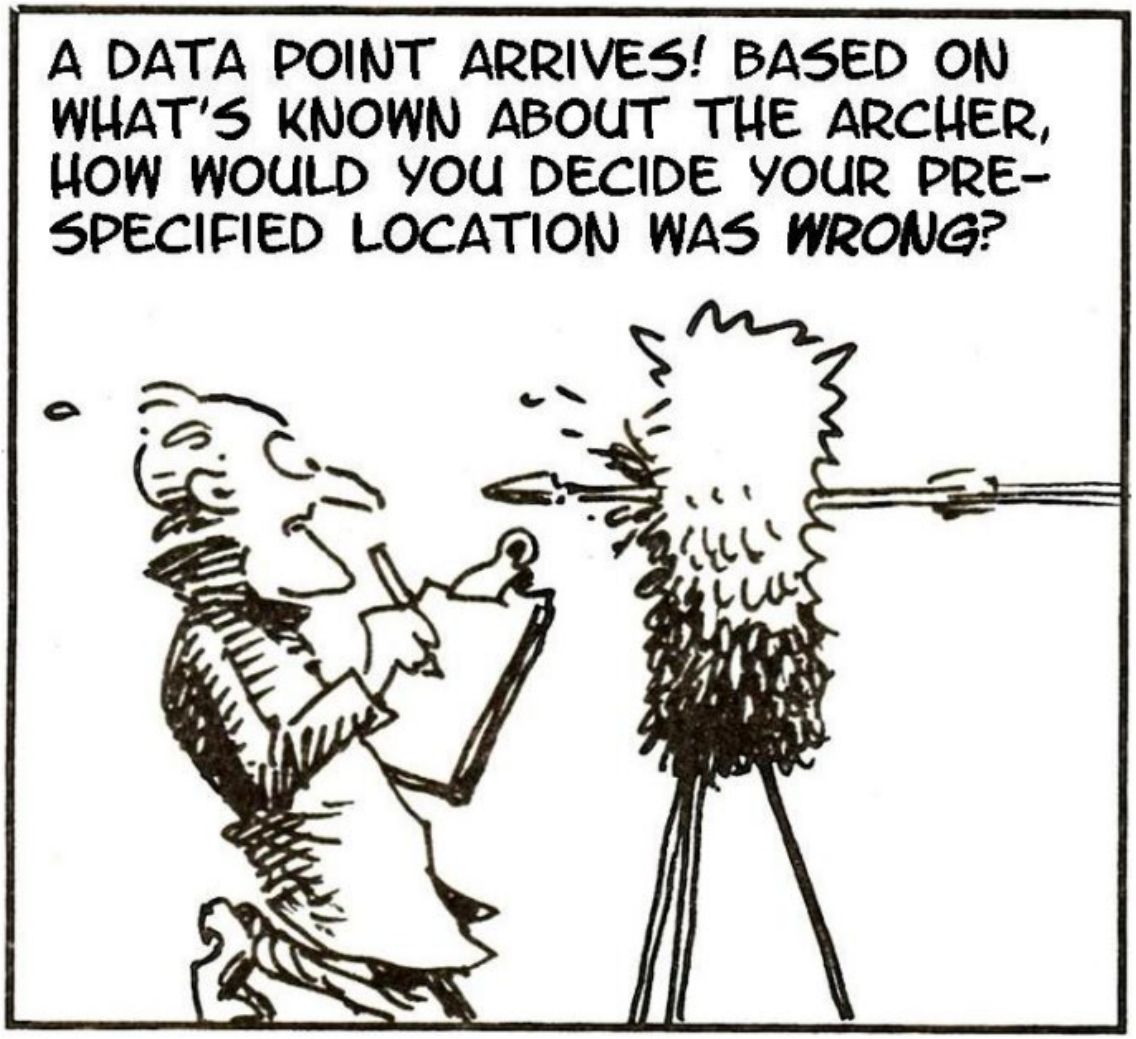


Summary: "The Cartoon Guide to Statistics", by Gonick & Smith

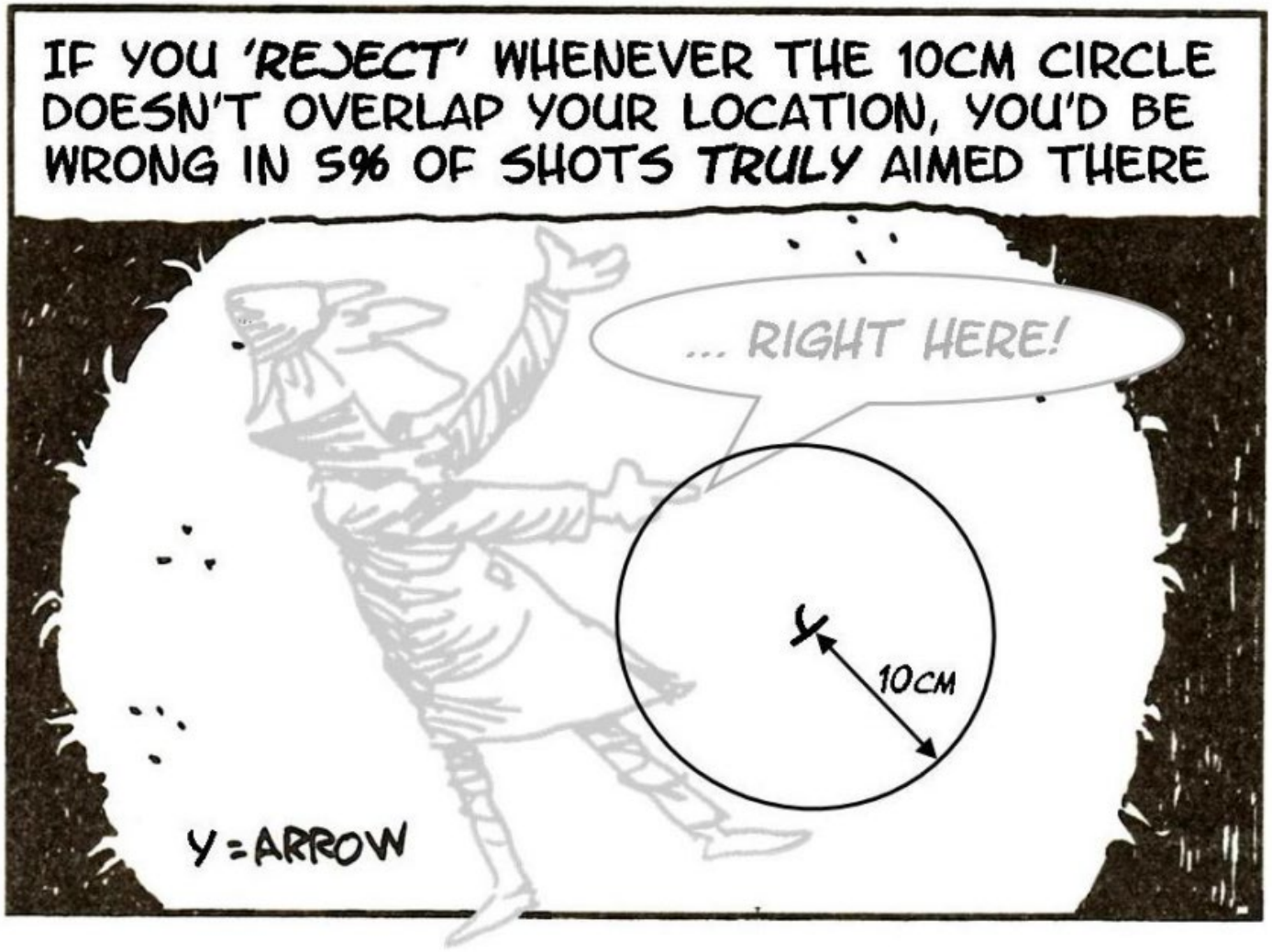


Summary: "The Cartoon Guide to Statistics", by Gonick & Smith

Summary: "The Cartoon Guide to Statistics", by Gonick & Smith



Summary: "The Cartoon Guide to Statistics", by Gonick & Smith



Conclusions from Spiegelhalter

- Statistical methods should enable data to answer scientific questions; *Why am I doing this?*
- Signals always come with noise;
- Plan ahead, really ahead;
- Worry about data quality;
- Statistical analysis is more than a set of computations;
- Keep it simple; *Good visualisations*
- Provide assessments of variability;
- Check assumptions; *Is this sensible?*
- Replicate when possible;
- Make analyses reproducible, allowing others to access data and code; *Github*

Conclusions

- Descriptive statistics are for describing data.
- Inferential statistics are for making inferences about populations.
  - they make assumptions about the population distribution.
- Statistical significance only indicates whether an effect is likely to exist in the population.