WEO Research Workshop

Seoul, 17 November 2018

Biostatistician's role in study design

Dr Sunny H Wong
MBChB, DPhil, FRCPEd, FRCPath, FHKCP, FHKAM

Assistant Professor
Institute of Digestive Disease
The Chinese University of Hong Kong

Breath of biostatistics

- Descriptive statistics (data types, central tendency, dispersion, exploratory data analysis)
- Probability distributions & confidence intervals
- Hypothesis testing (null hypothesis, type I and II errors, sample size, power)
- Inferential statistics (t-test, chi-square, trend test, Fisher's test, log rank test, comparative data analysis)
- Correlation & regressions
- Multiple comparisons & corrections
- Survival analysis
- Meta-analysis
- Bayesian statistics
- Others (diagnosis, public health, bioinformatics)

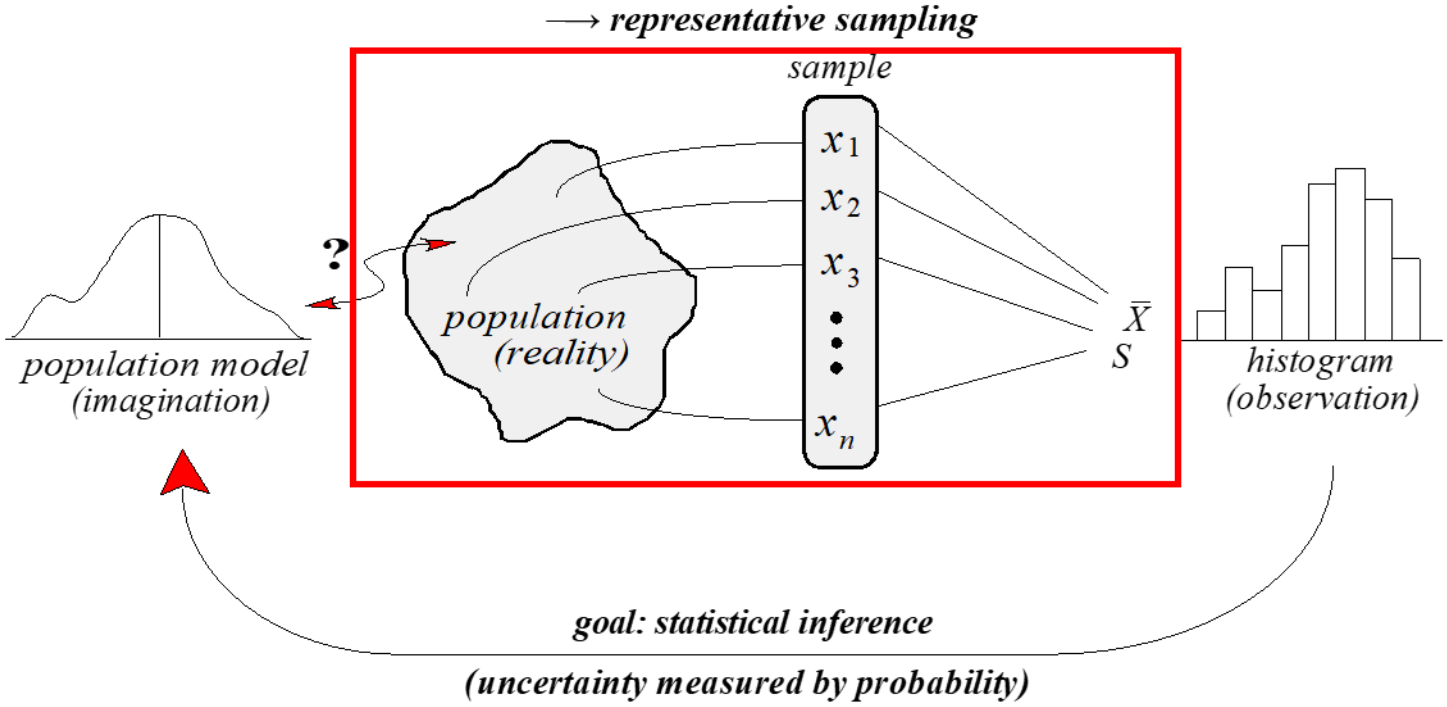
Outline

1. Study design and power
2. Descriptive statistics
3. Inferential statistics
4. Software and tips

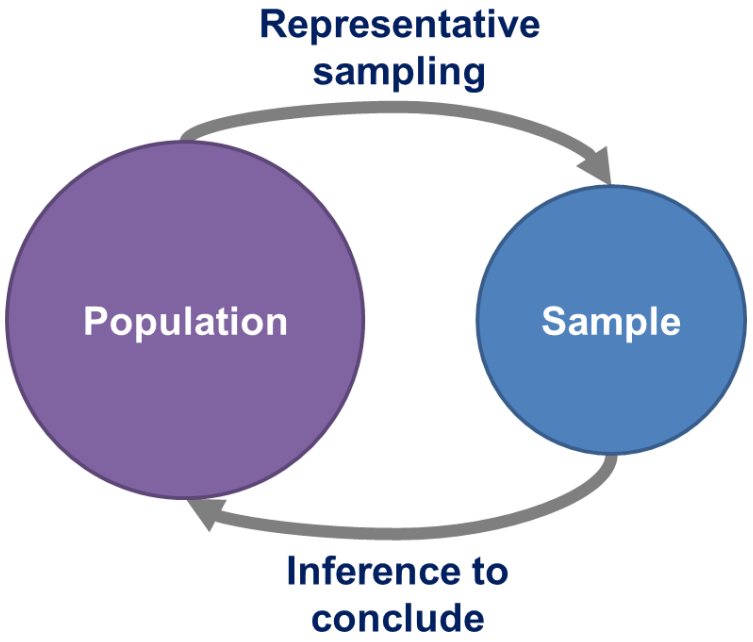
Study design and power

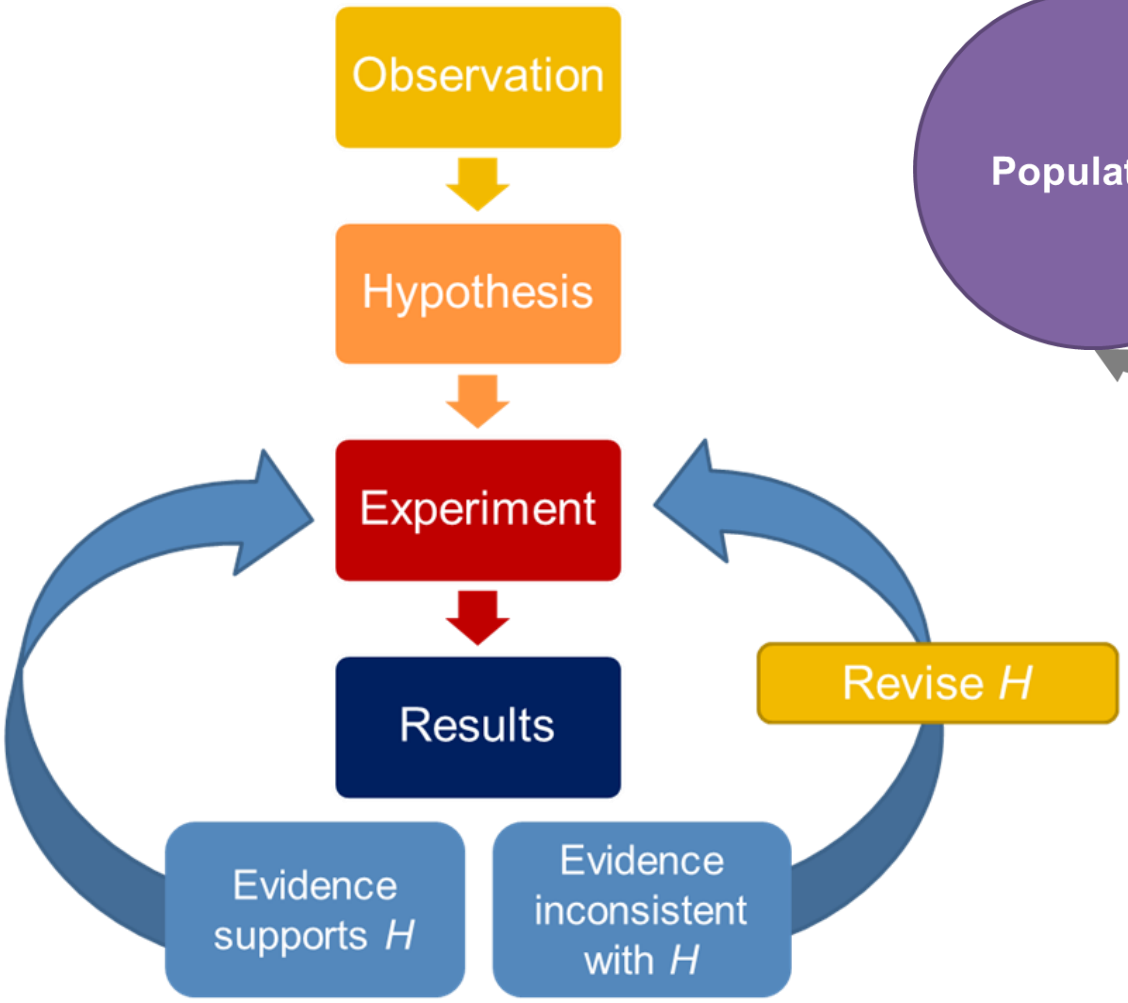
Concepts of statistics

- Science of (1) collecting and analyzing data to help make decisions in uncertainty, and (2) to make inference about a phenomenon observed

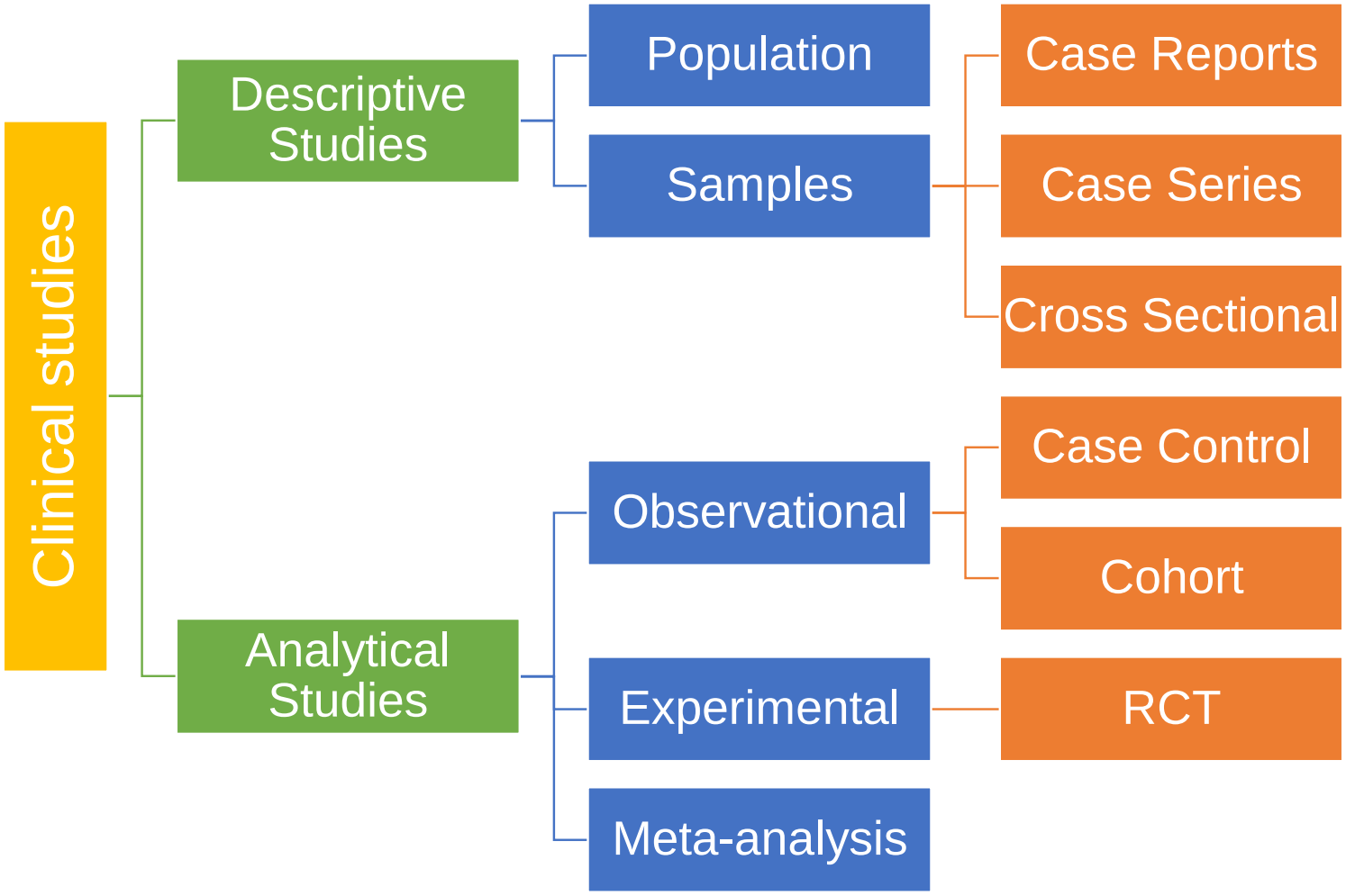


Clinical studies

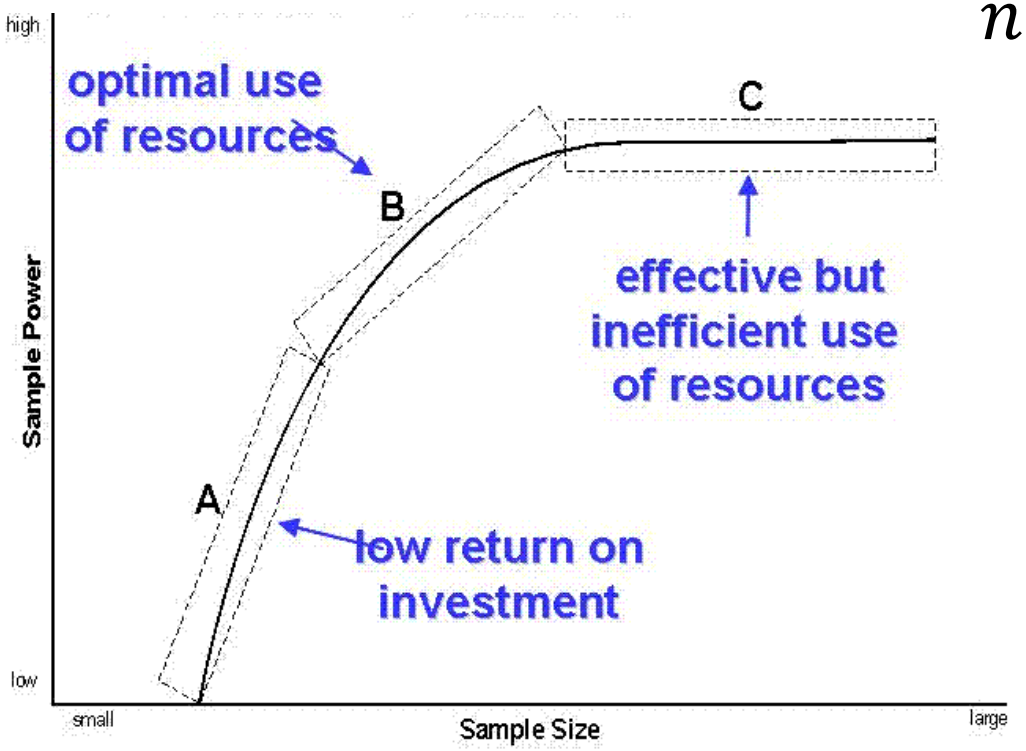




Types of clinical studies



Power calculation



$$n = \frac{2\sigma^2(Z_\beta + Z_{\alpha/2})^2}{d^2}$$

n = sample size
σ = standard deviation
β = power
α = level of significance
d = effect size

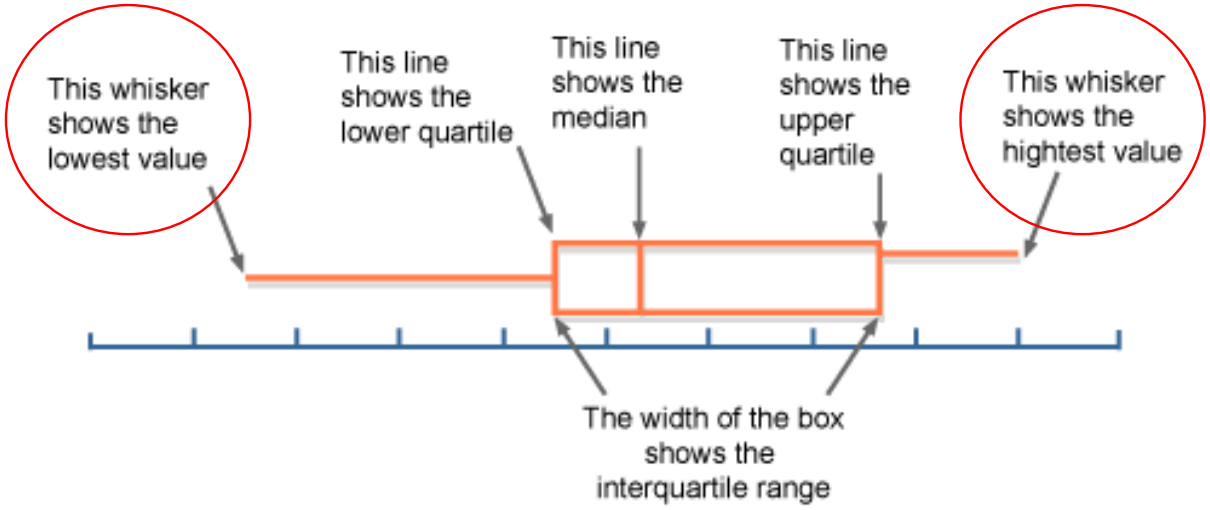
Loss to follow-up

Descriptive statistics

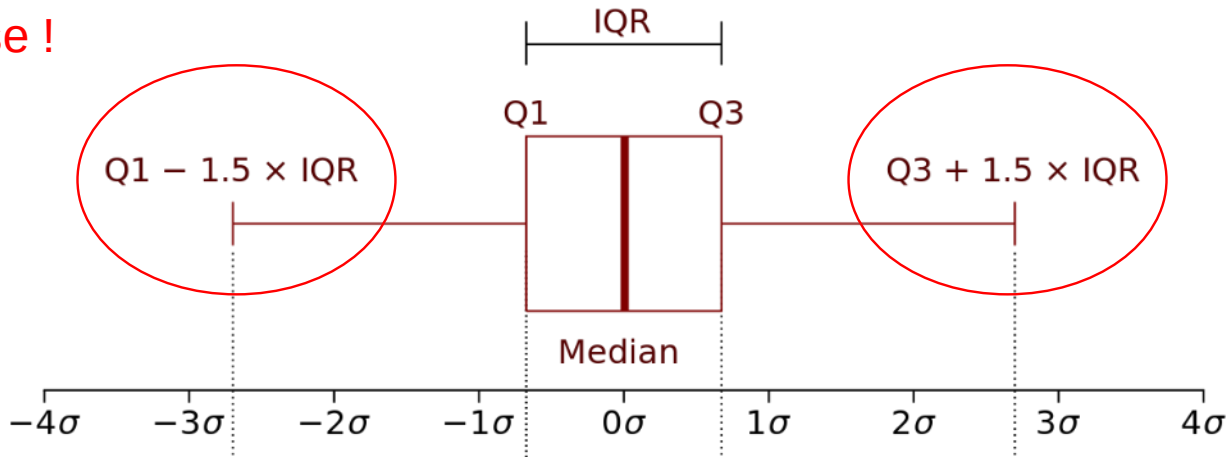
Types of variables

- Nominal - categories (e.g. gender, ethnicity)
- Ordinal - categories with a trend
  (e.g. cancer stage, grade)
- Numerical / scalar - quantitative
  - Continuous scale (e.g. age, height)
  - Discrete scale (e.g. number of polyp)

Box-and-whisker plot



Mind these !

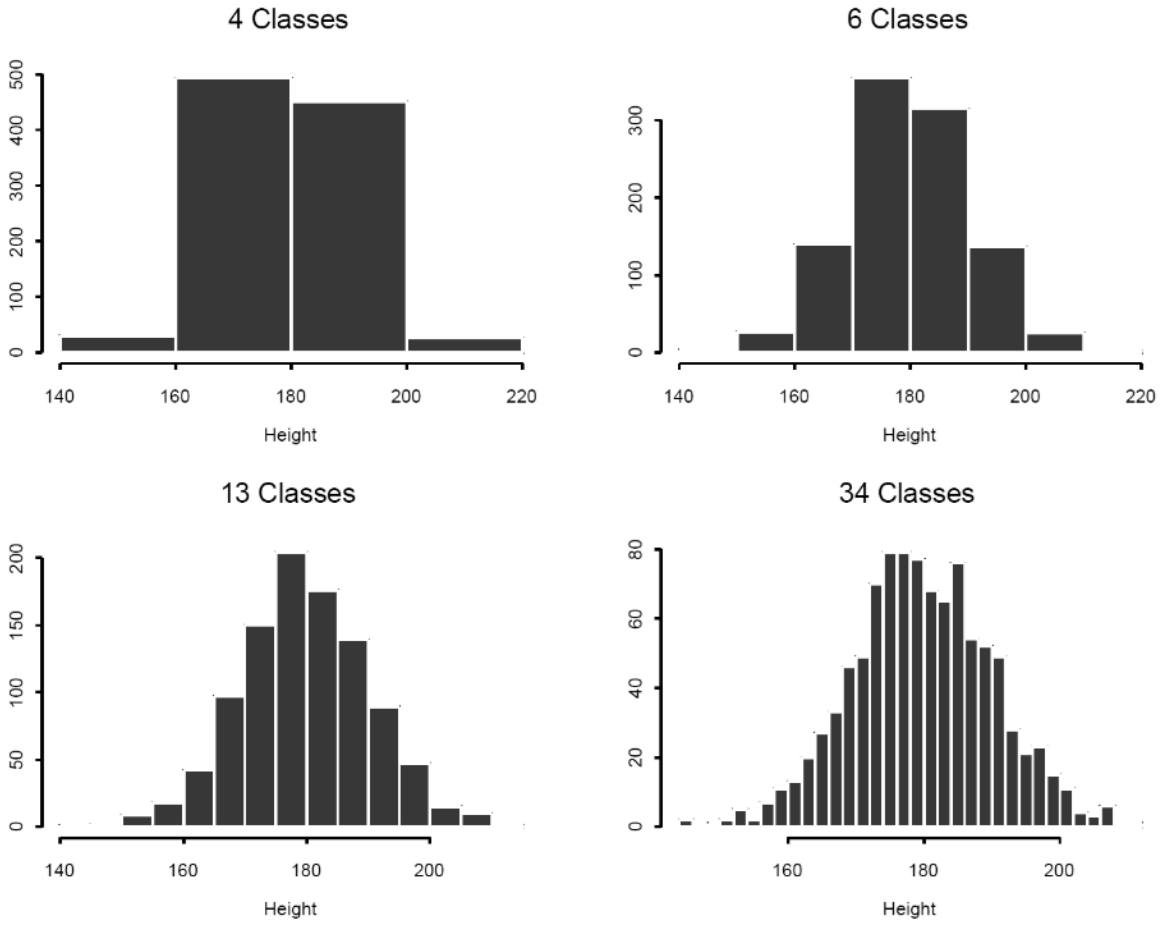


Bar chart

- Height is the mean

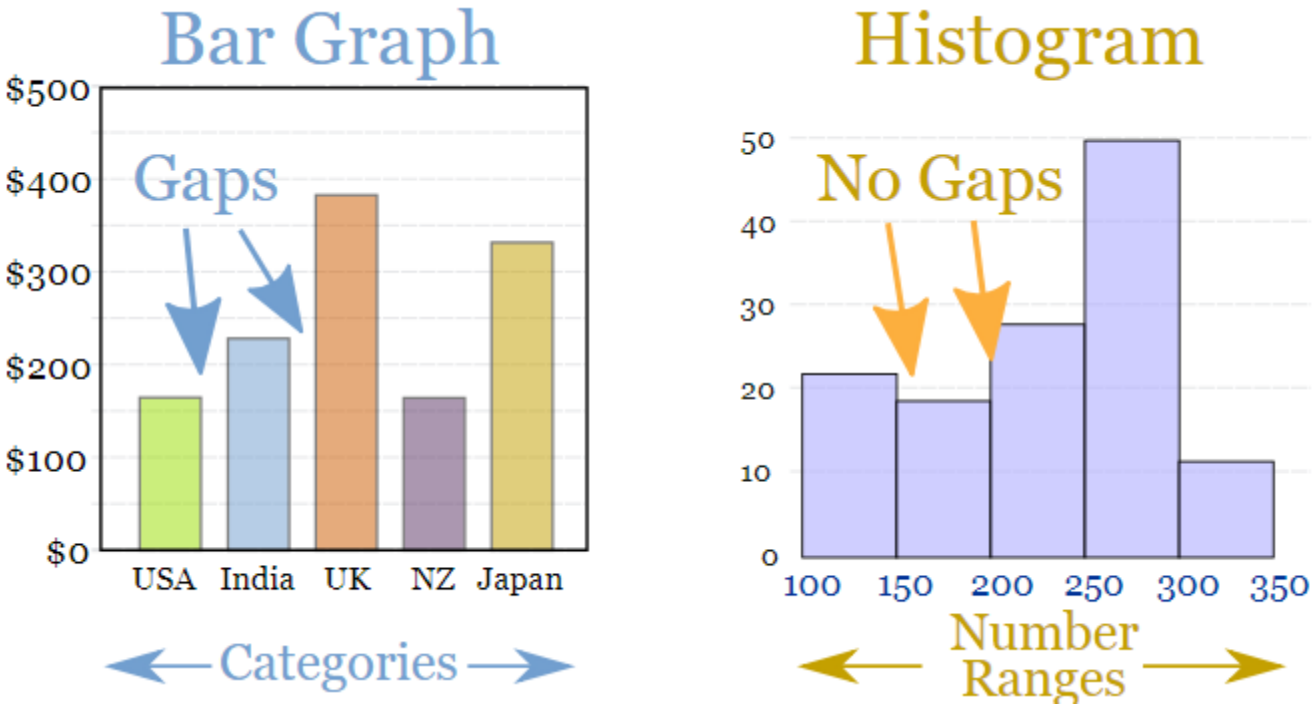
- How about error bars?
  - Standard deviation (SD)
  - Standard error of mean (SE / SEM)
  - Confidence interval (95% CI)

Histogram



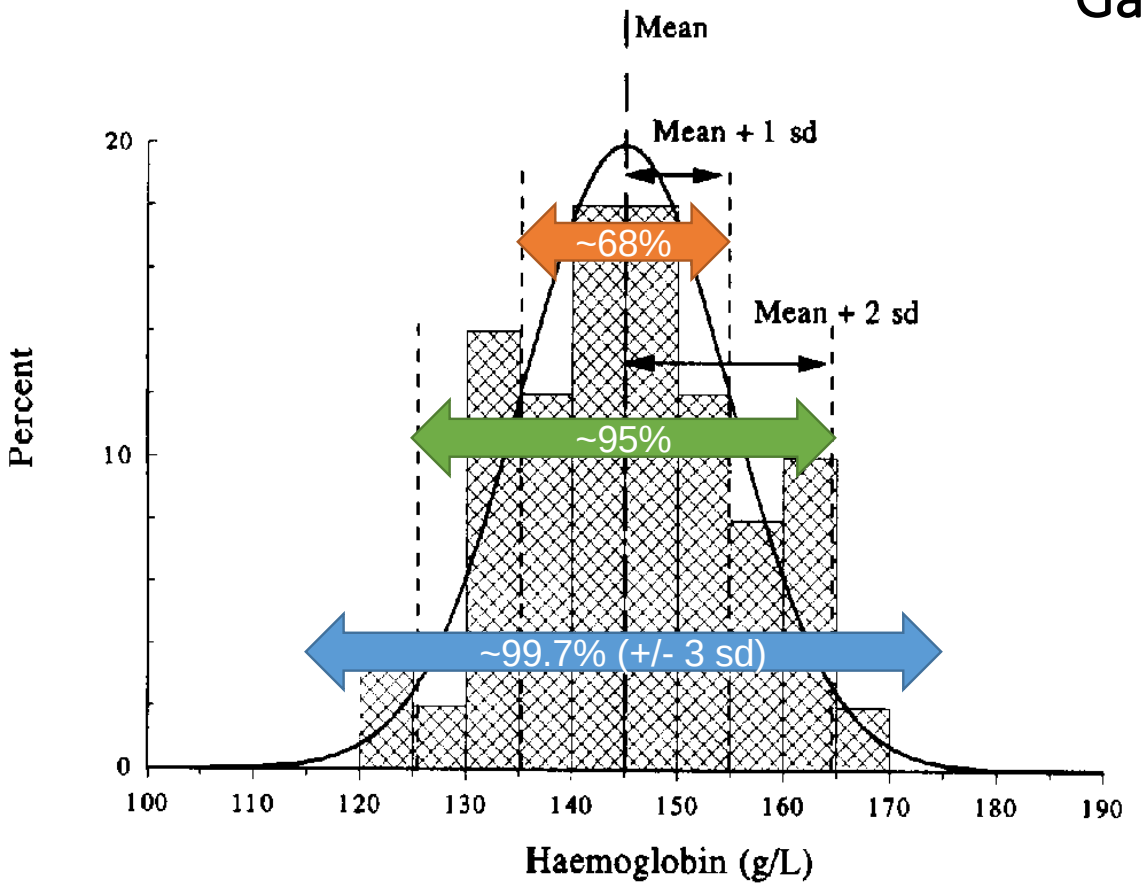
Q: What are the differences between bar chart and histogram? (P.T.O.)

Bar chart and histogram



Bonus Q: what is a Pareto chart?

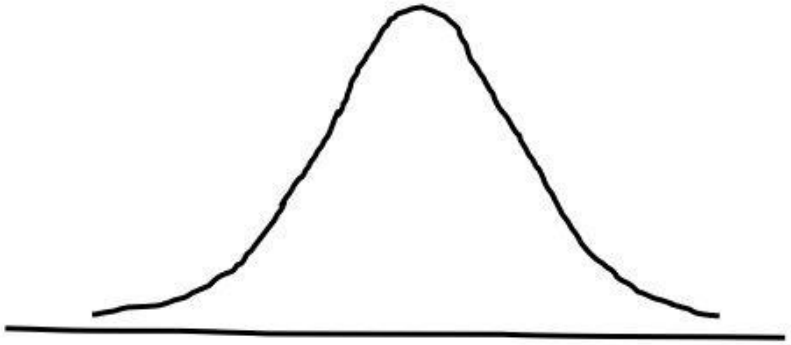
Normal distribution

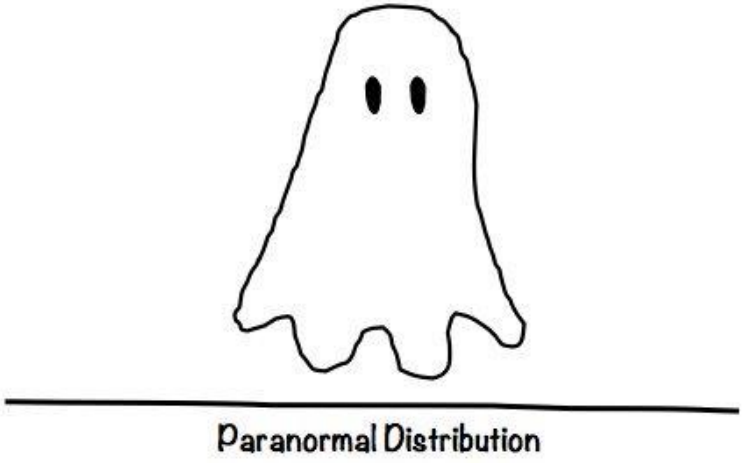


Gaussian distribution
(*z*-statistics)

de Moivre,
Gauss & Laplace

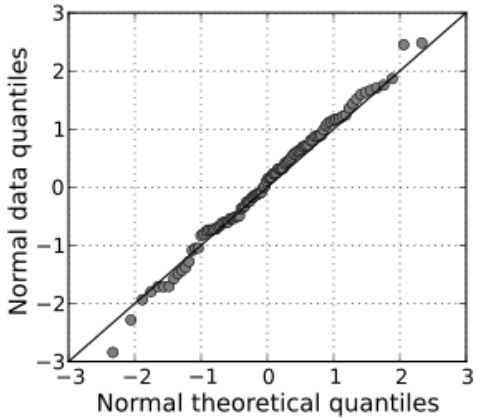
Normality

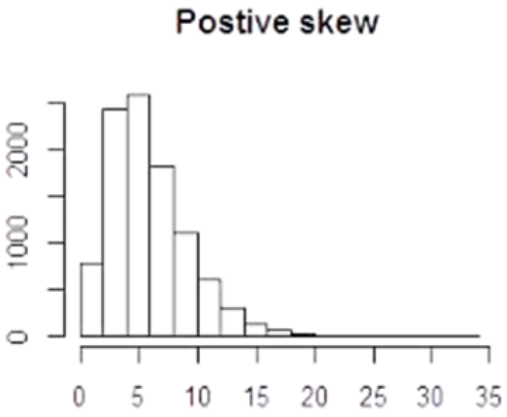


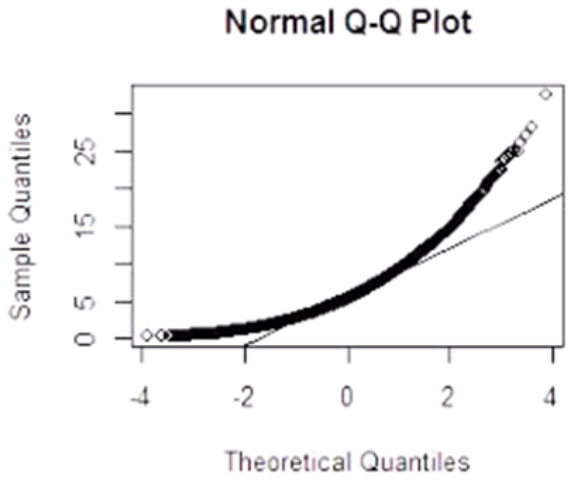


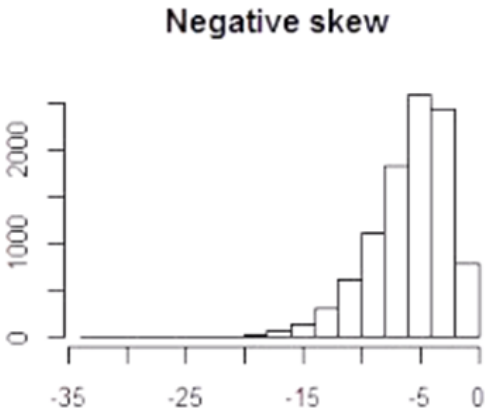
- Histogram inspection
  - n>30
  - Fitting shape
- Quantile-quantile plot
- Formal tests
  - Komolgorov-Smirnov test
  - Shapiro Wilk test

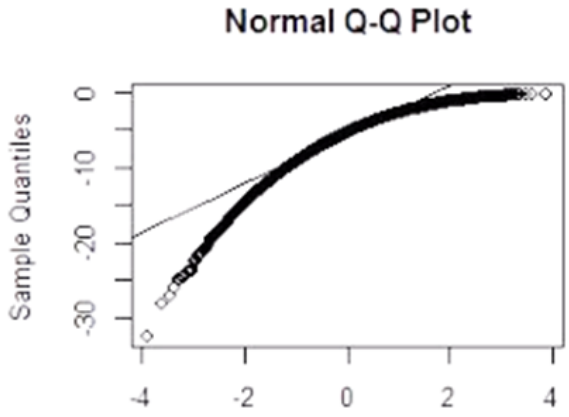
QQ plot









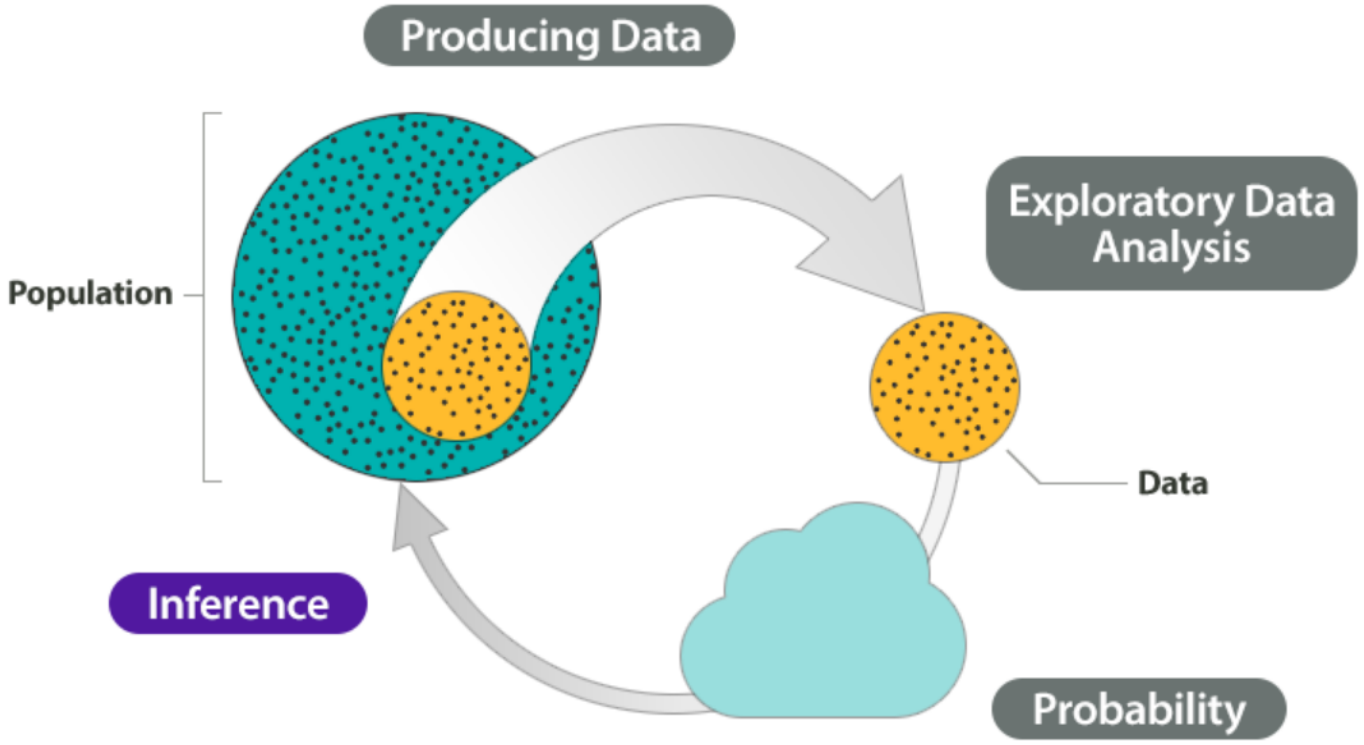


Q: What is the hospital length-of-stay distribution? (right skewed distribution)

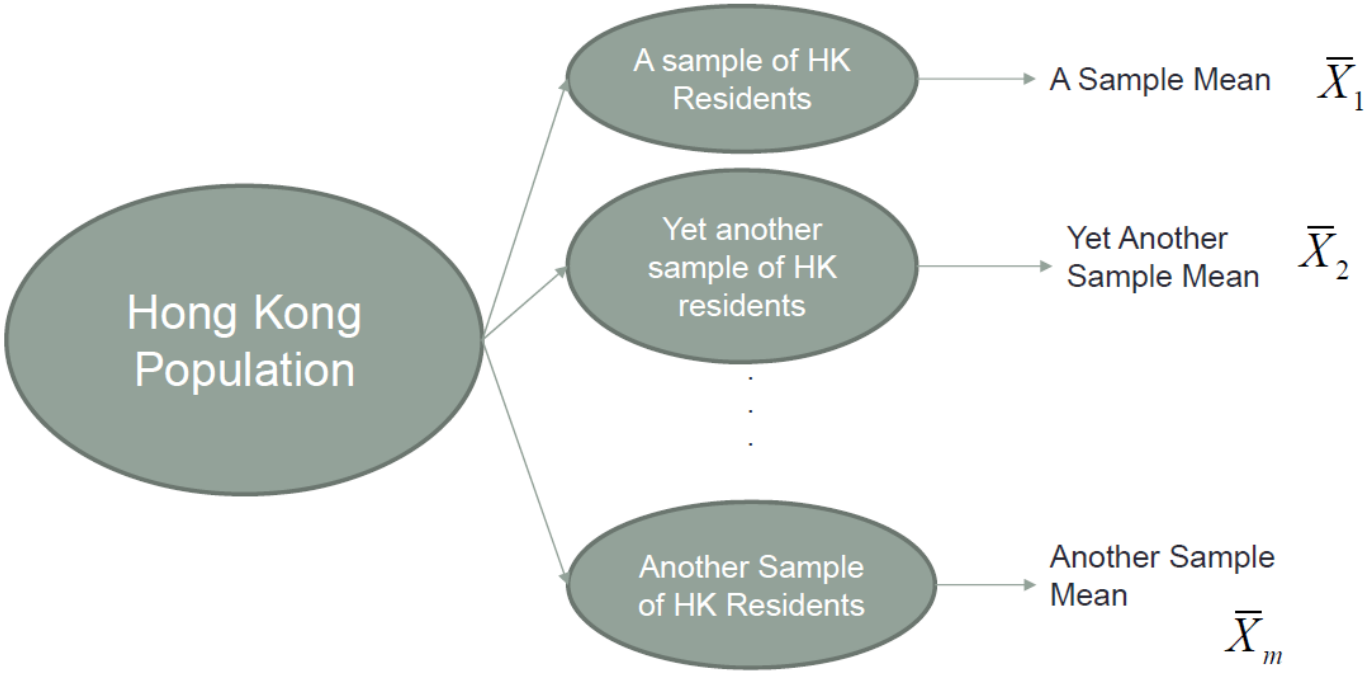
Inferential statistics

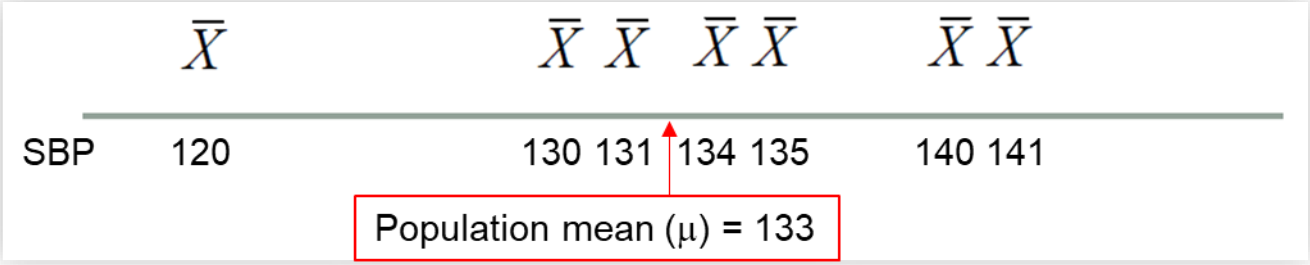
Sampling

e.g. SBP of the Hong Kong population

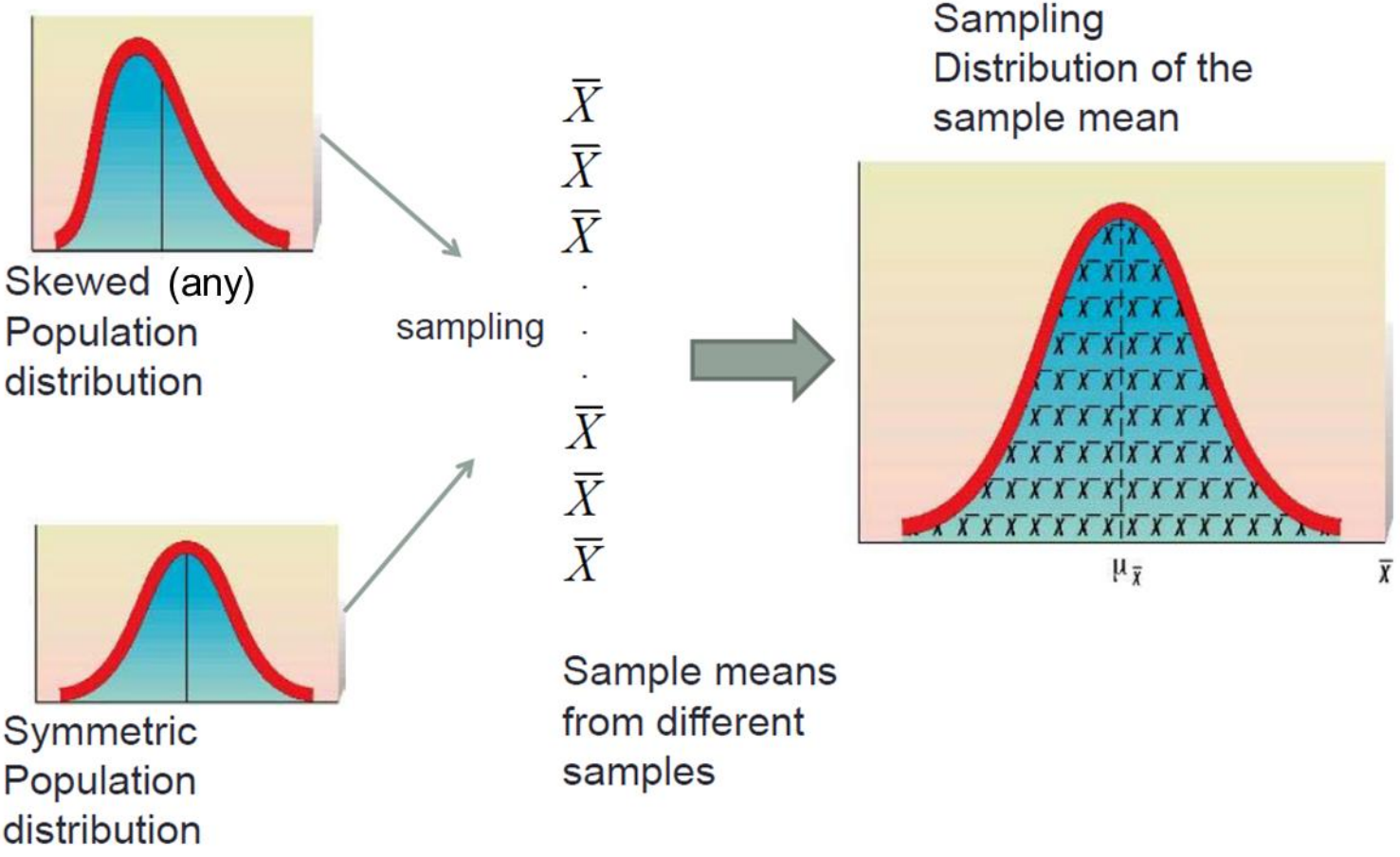


Sampling distribution





Central limit theorem

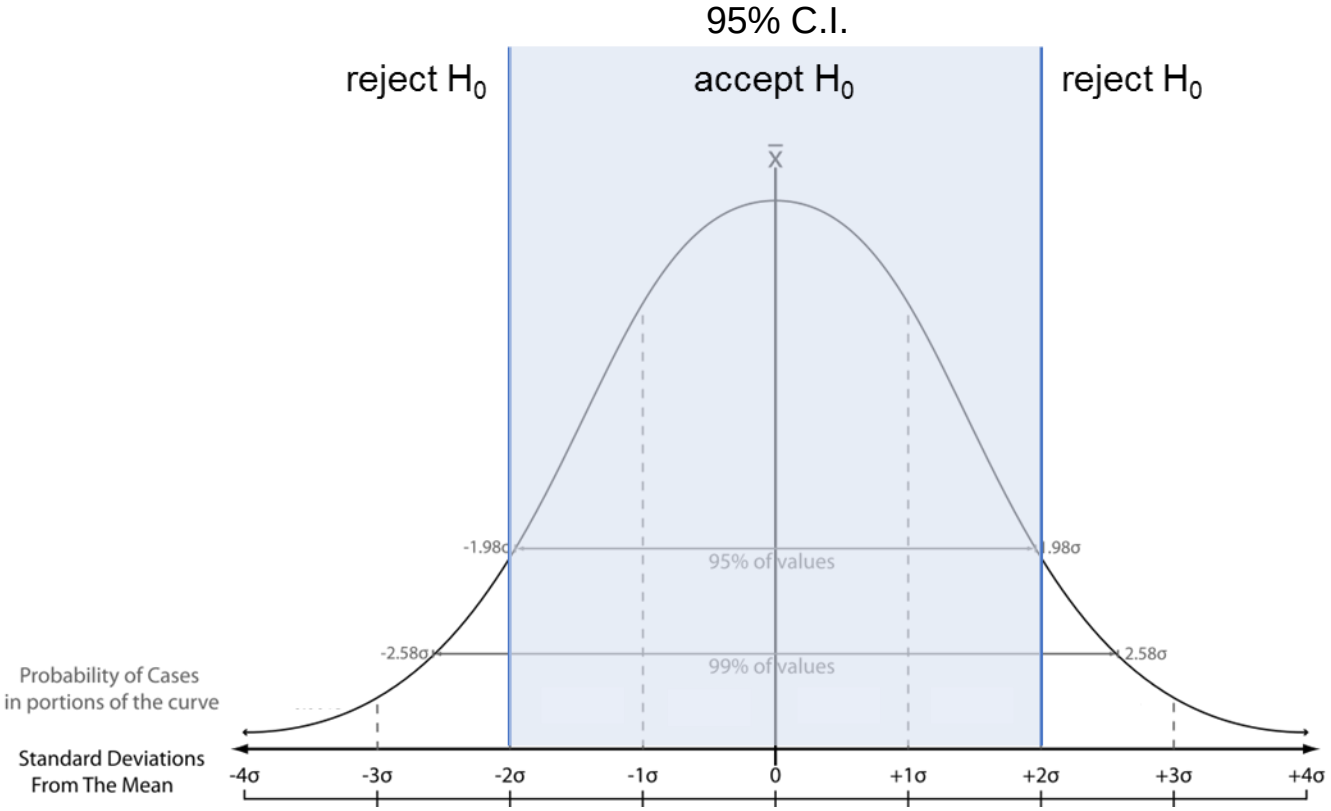


Distribution of the means

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

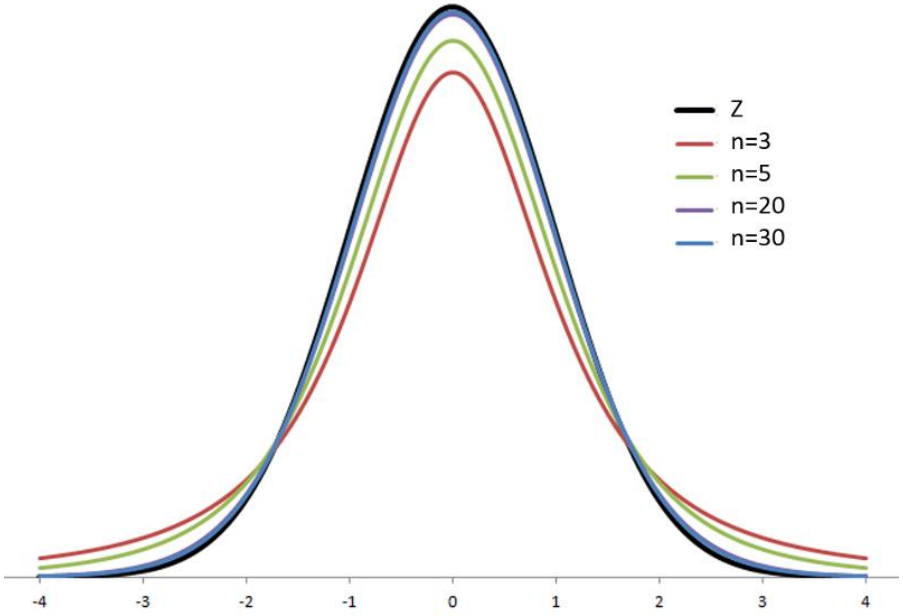
$$se = \frac{\sigma}{\sqrt{n}}$$



Student's *t* distribution

- Described by William Sealy Gosset

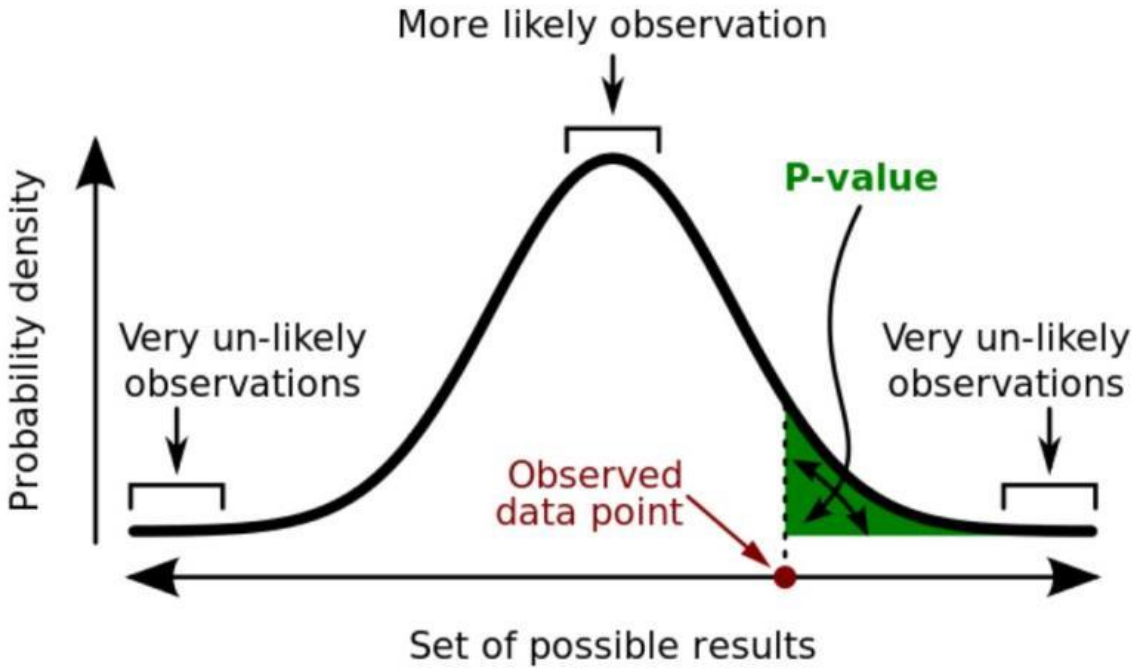
Resemble normal distribution if sample size is large (n>30)



Hypothesis testing

- Confirmatory data analysis to determine the probability that a given hypothesis is true
- Null hypothesis 'H_0': statement of no differences or association between variables
- Alternative hypothesis 'H_1': statement of differences or association between variables
- Type I (alpha) and Type II (beta) error

P-value

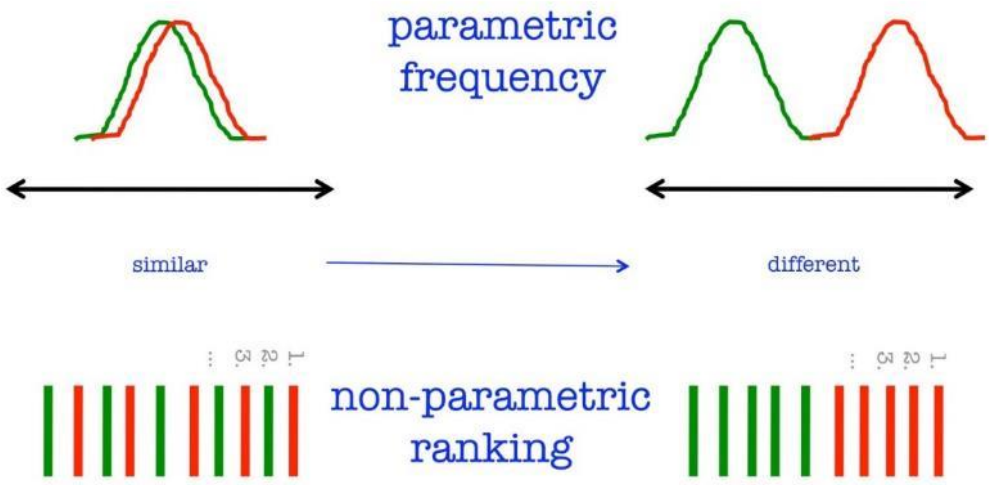


A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Probability of mistakenly rejecting the null hypothesis (α)

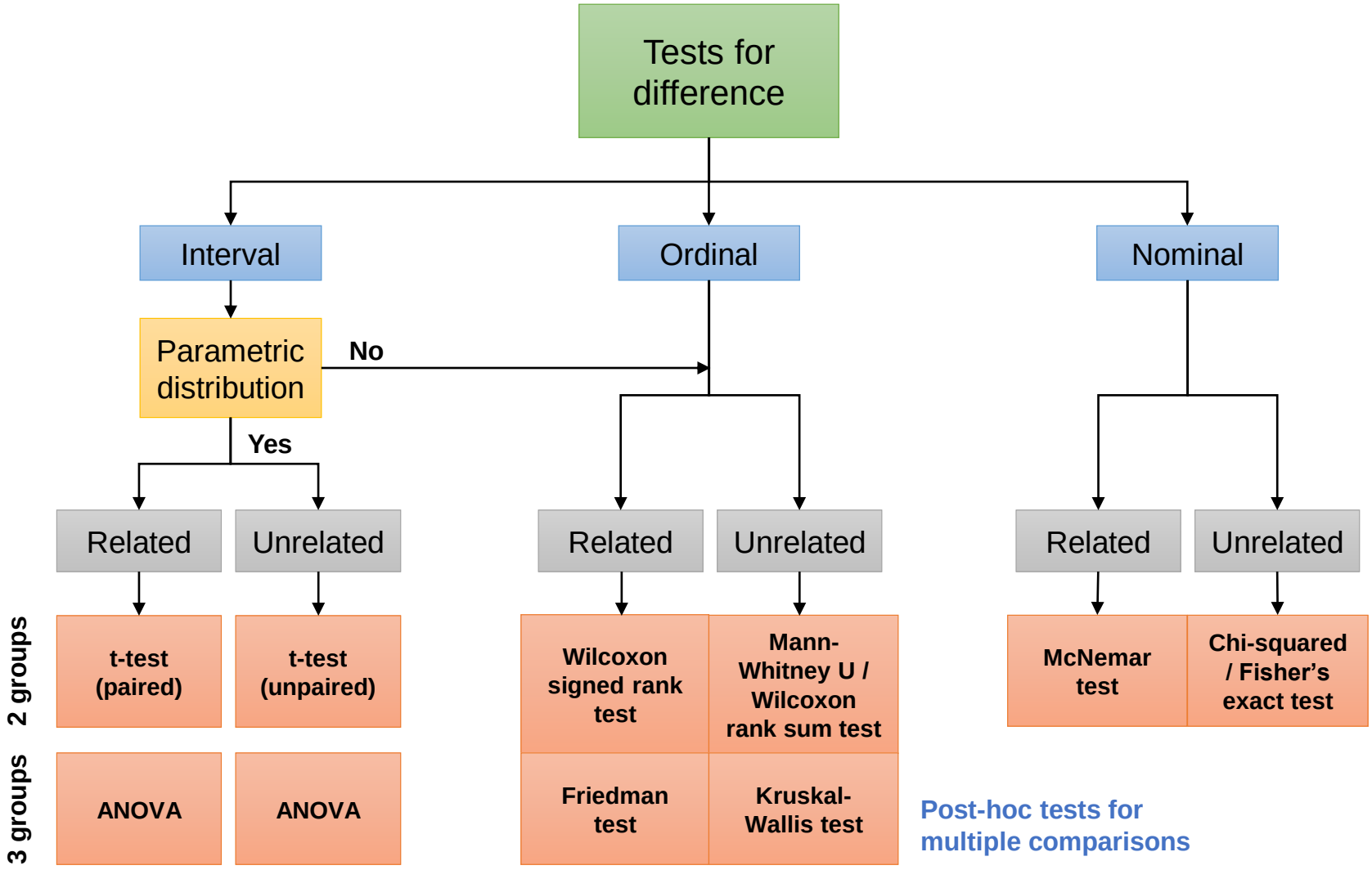
Parametric tests

- Parametric tests assume data comes from a population
  - With known probability distribution (e.g. normal)
  - Based on a fixed set of parameters (e.g. mean, SD)



Non-parametric tests usually less powerful (values discarded)

Comparative tests (interval)



A walk-through of rank sum test

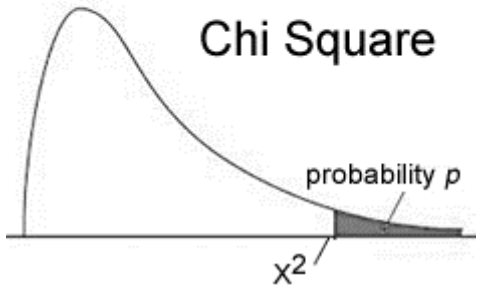
Group A	Group B	Rank A	Rank B		A	B
87	71	19	9	Total rank	127	148
72	42	10	1	Median	74	75.5
94	69	22	8	*n*	11	12
49	97	2	23			
56	78	4	14.5	**U(A)**	**71**	
88	84	20	17	**U(B)**	**61**	
74	57	12	5	**U statistic**	**61**	
61	64	6	7	***P*-value**	**0.76**	
80	78	16	14.5			
52	73	3	11			
75	85	13	18			
	91		21			

Comparative tests (categorical)

- Examine differences between ***observed*** and ***expected*** counts

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

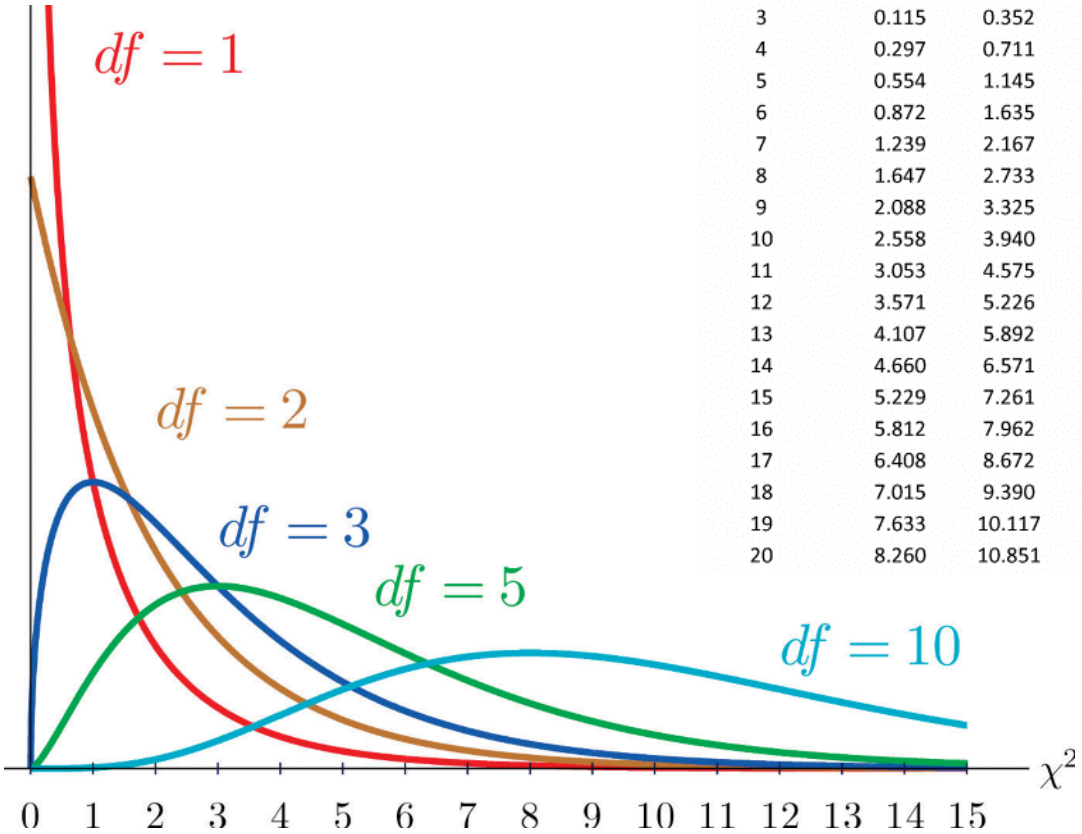
- Two assumptions
  - Independence of observations
  - Count of all cells >5



- Degree of freedom
  - *df = (R-1) x (C-1)*
  - No of free variables

	Drug A	Drug B	
Cured	20	10	**30**
Not cured	12	22	**34**
	32	**32**	**64**

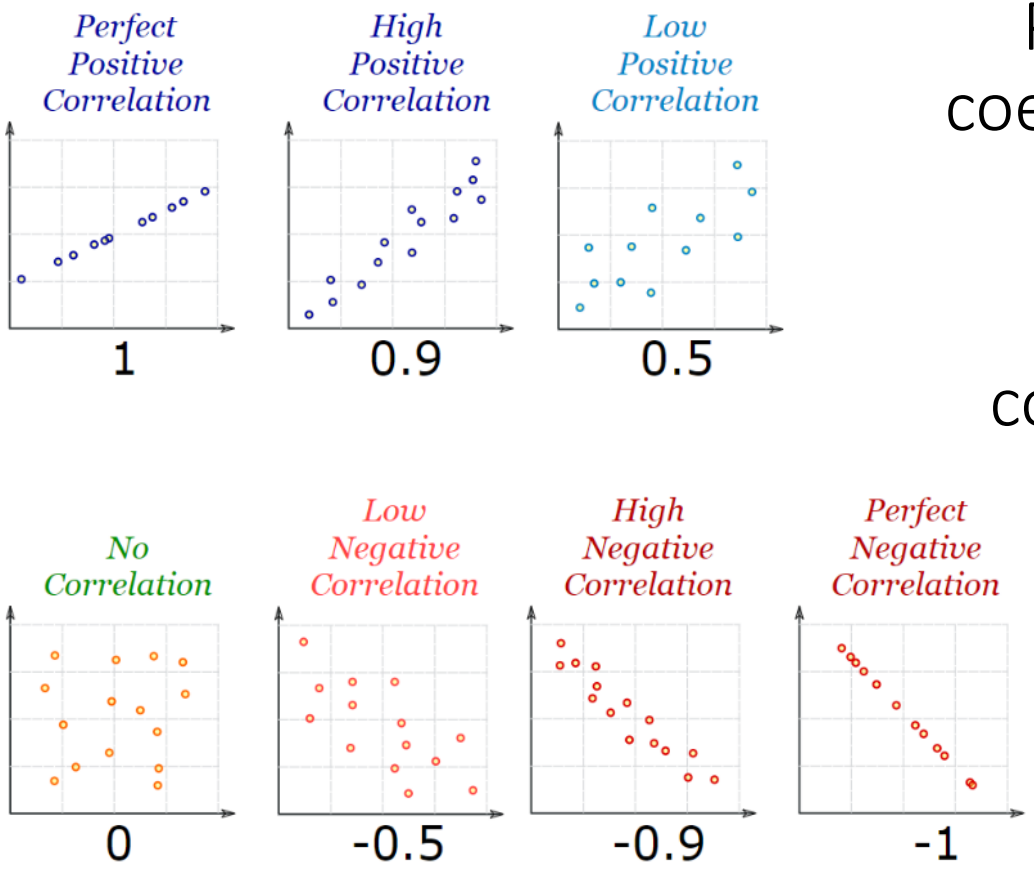
χ^2 distribution and *df*



Degrees of Freedom	Probability of a larger value of x^2							
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Correlation



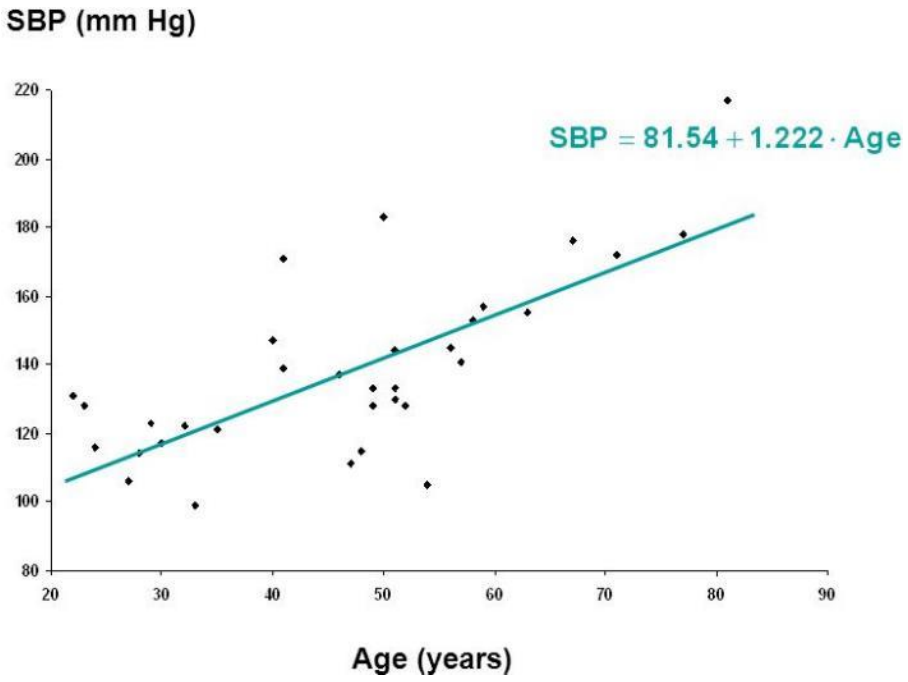
Pearson's correlation coefficient (parametric)

Spearman's rank correlation coefficient (non-parametric)

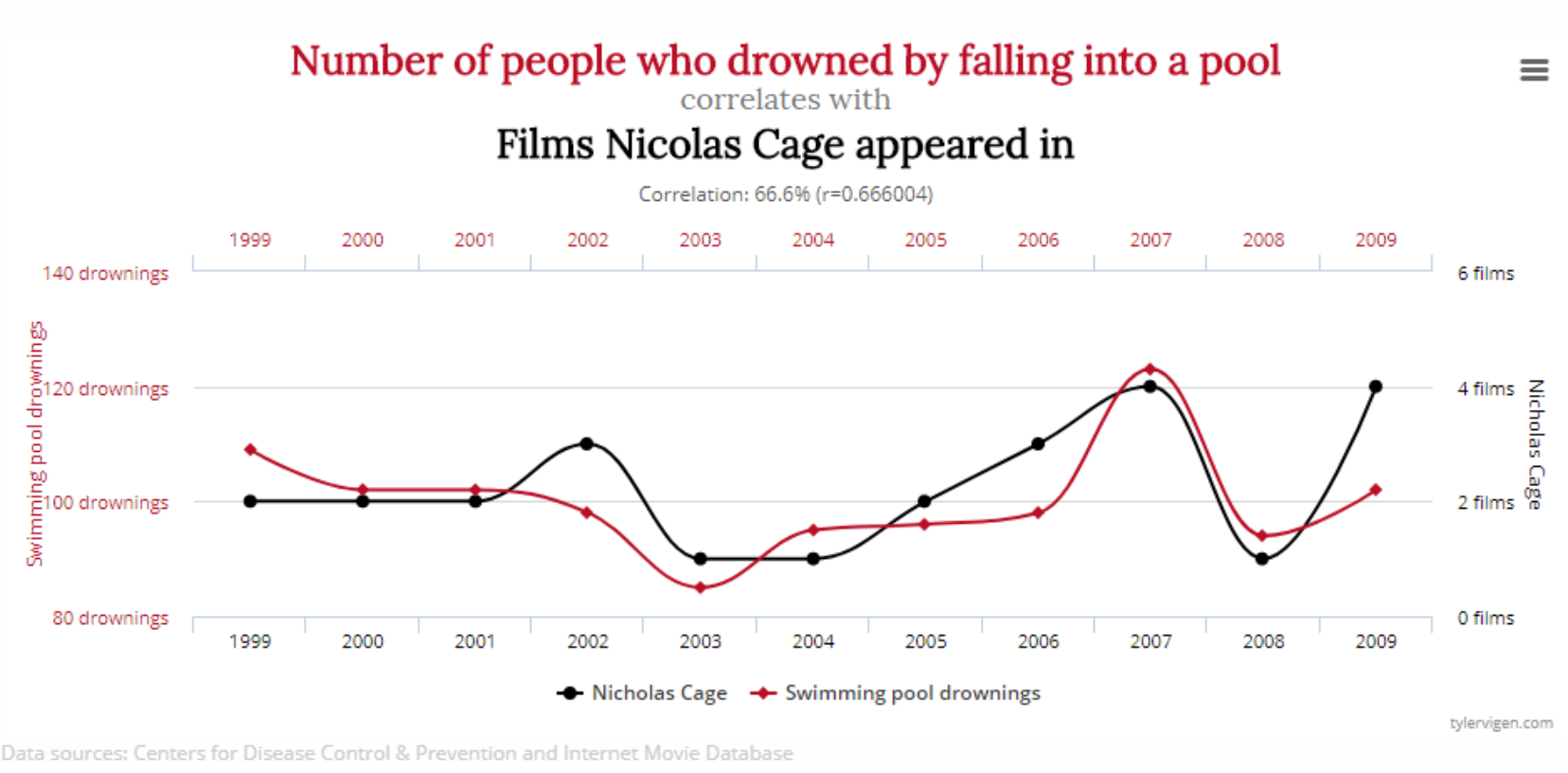
Regression

- Estimating relationships between variables (between dependent and independent variables)
- Logistic regression
- Linear regression

$$\hat{Y} = bX + a$$



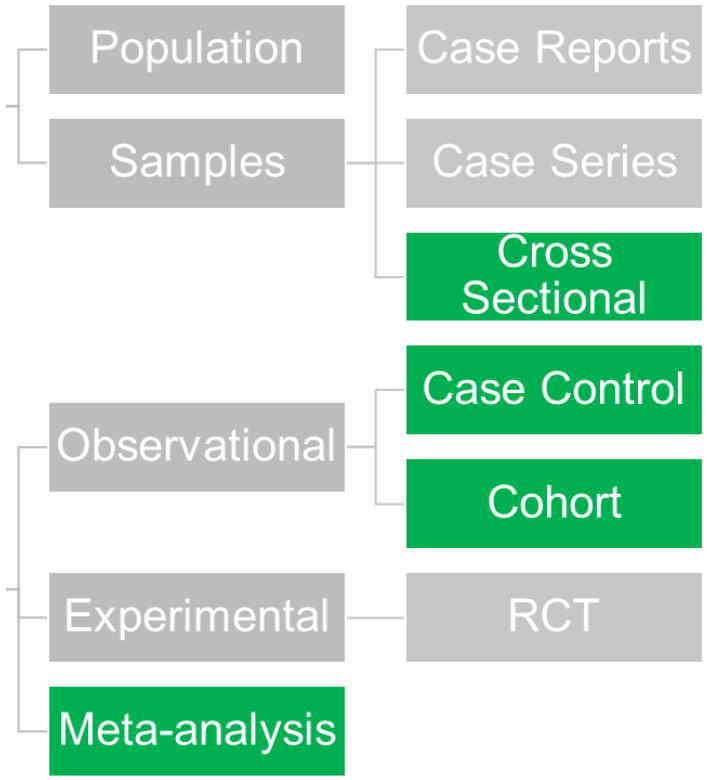
Correlation/association ≠ causality



Q: What are the differences between correlation and regression?

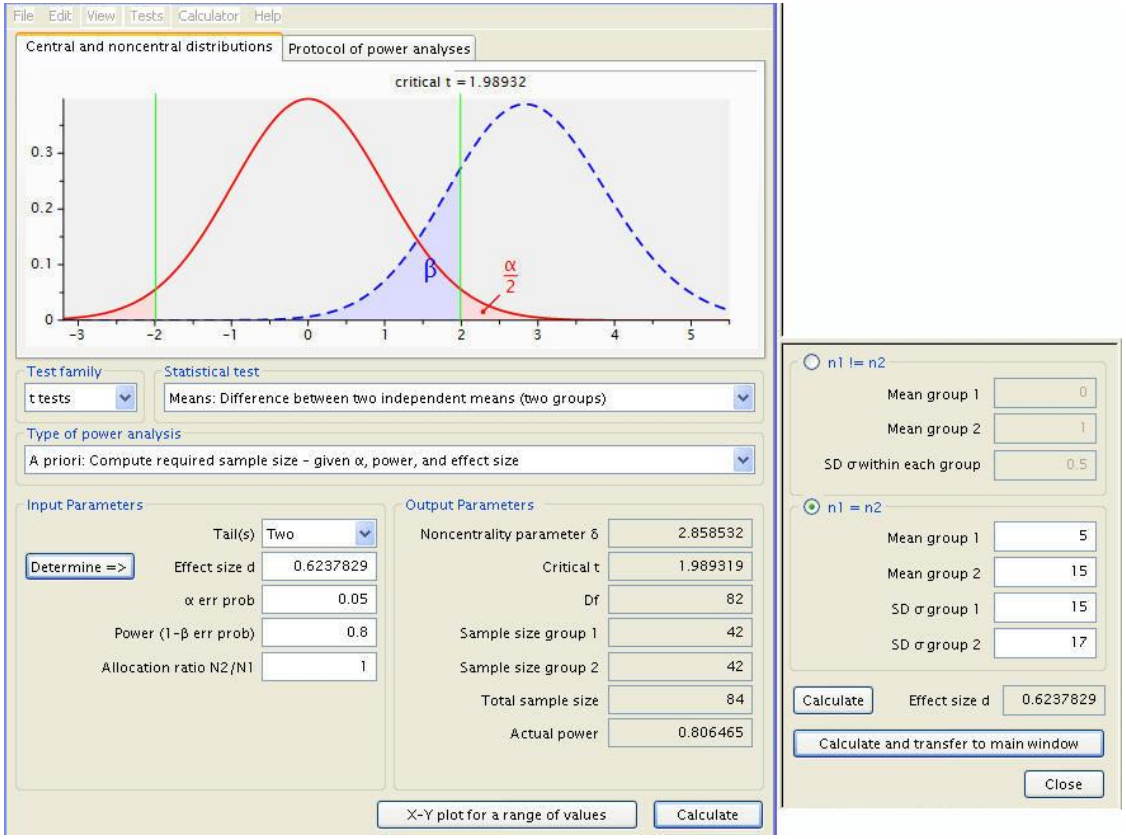
Tips and software

Study design



- Most important is your **research question**
- Consider
  - Time
  - Effort
  - Infrastructure
  - Clinical ethics, governance and compliance

Power calculators



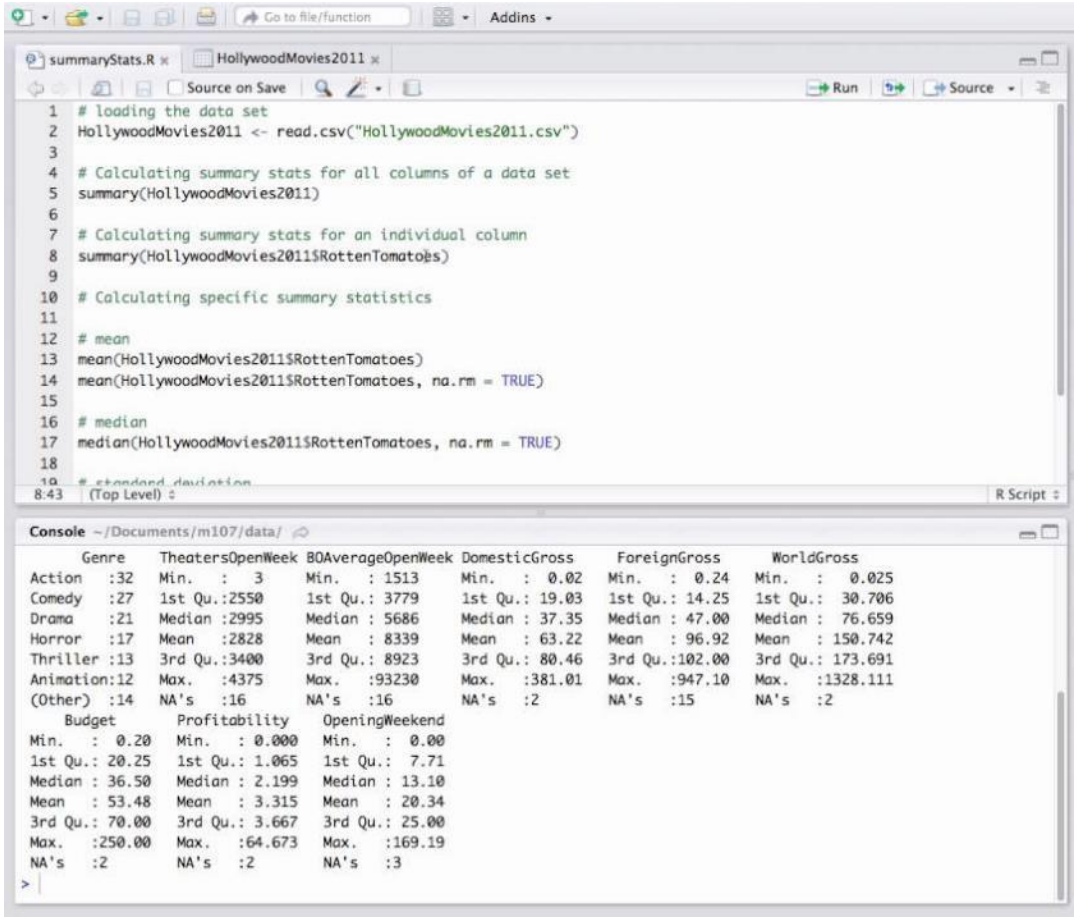
GPower: http://www.gpower.hhu.de/

CUHK CCRB: http://www2.ccrb.cuhk.edu.hk/web/

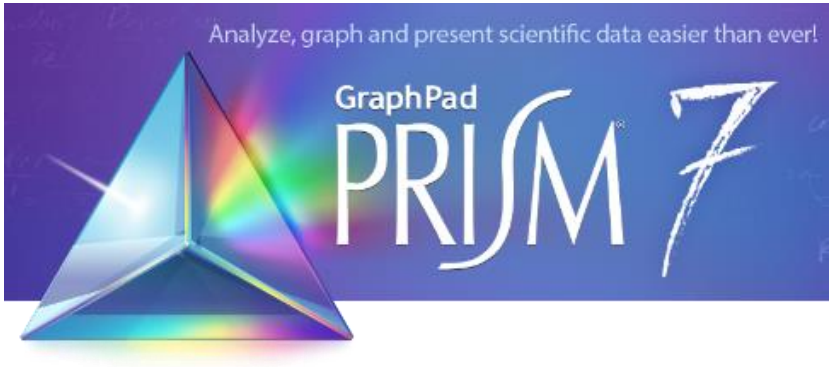
Clin Calc: http://clincalc.com/stats/samplesize.aspx

Statistical software

R Project for Statistical Computing



Statistical software



Questions?