Ask Mish

Epidemiology and Biostatistics

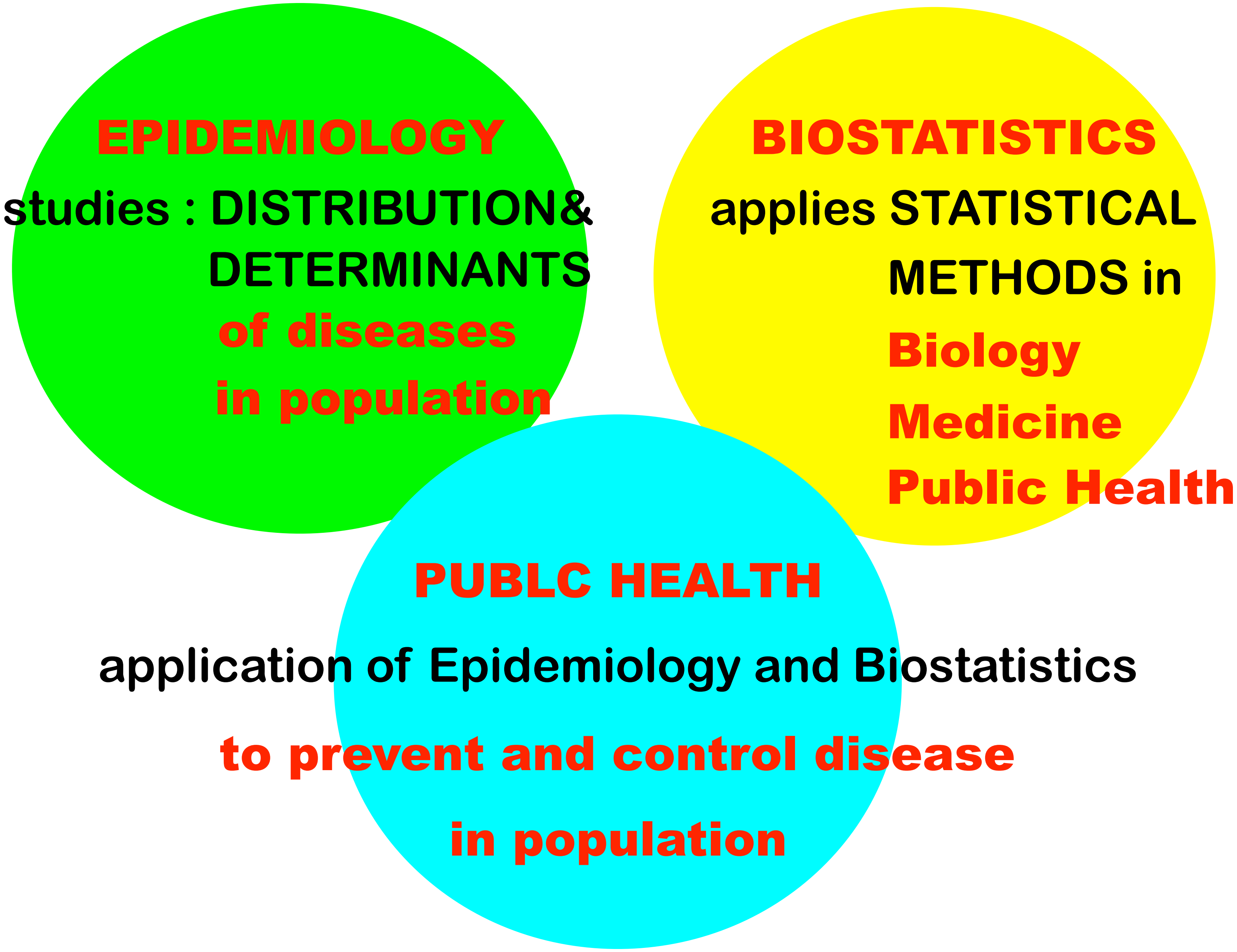
Introduction:

Ask Mish

- First 5 slides : definitions of epidemiology, biostatistics and public health and their connection
- Next slide compares two concepts : health vs disease
- Last slide from introduction part- other two concepts related to disease are compared : signs vs symptoms

why epidemiology & biostatistics?hard way:

Ask Mish



EPIDEMIOLOGY

studies : DISTRIBUTION&
DETERMINANTS
of diseases
in population

BIOSTATISTICS

applies STATISTICAL
METHODS in
Biology
Medicine
Public Health

PUBLC HEALTH

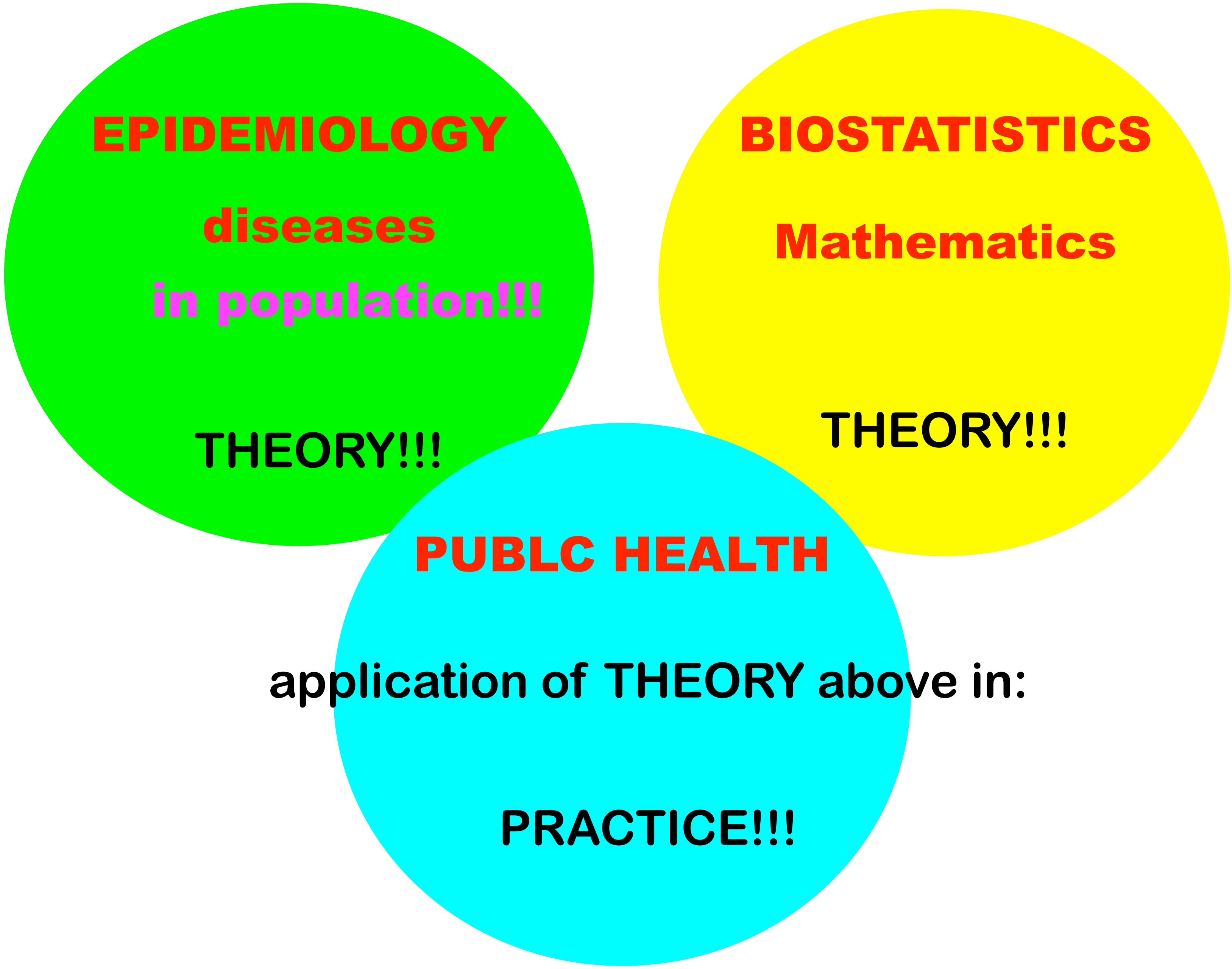
application of Epidemiology and Biostatistics

to prevent and control disease

in population

why epidemiology & biostatistics?easy way:

Ask Mish



EPIDEMIOLOGY

diseases

in population!!!

THEORY!!!

BIOSTATISTICS

Mathematics

THEORY!!!

PUBLC HEALTH

application of THEORY above in:

PRACTICE!!!

why epidemiology & biostatistics?comparison:

Ask Mish

	epidemiology	biostatistics	public health
refer to	study of DISEASES in a way that you can:	application of STATISTICS	application of theories from Epidemio & Biostat.
action	prevent and control disease (THEORY)	to exclude events in medicine that are due by chance alone	to prevent and control disease(PRACTICE)
on:	population not one person !!!!	population not one person !!!!	population not one person !!!!

health vs disease: definitions

Ask Mish

- **HEALTH**
- complete:
- physical
- mental
- social well being
- not absence of disease

- **DISEASE:**
- diagnosis using:
- signs
- symptoms
- history
- test results

signs vs symptoms : definitions

Ask Mish

- **SIGN:**
- objective evidence of disease
- can be seen/
- can be measured
- e.g.vital signs

- **SYMPTOM:**
- subjective evidence of disease
- a feeling of subject
- others cannot see/ measure
- e.g. headache

Epidemiology: history, distribution of disease and rates

Ask Mish

- next slide is about the beginning of epidemiology
- next 3 slides refer to DISTRIBUTION of disease in the world: define endemic, epidemic, pandemic (concepts applied to contagious diseases)
- next slides refer to the way the level a disease is assessed in epidemiology through rates: types of rates, rate of diseases in the US, most important rates (incidence & prevalence) and other rates used (attack rate, cumulative incidence, vital rates)

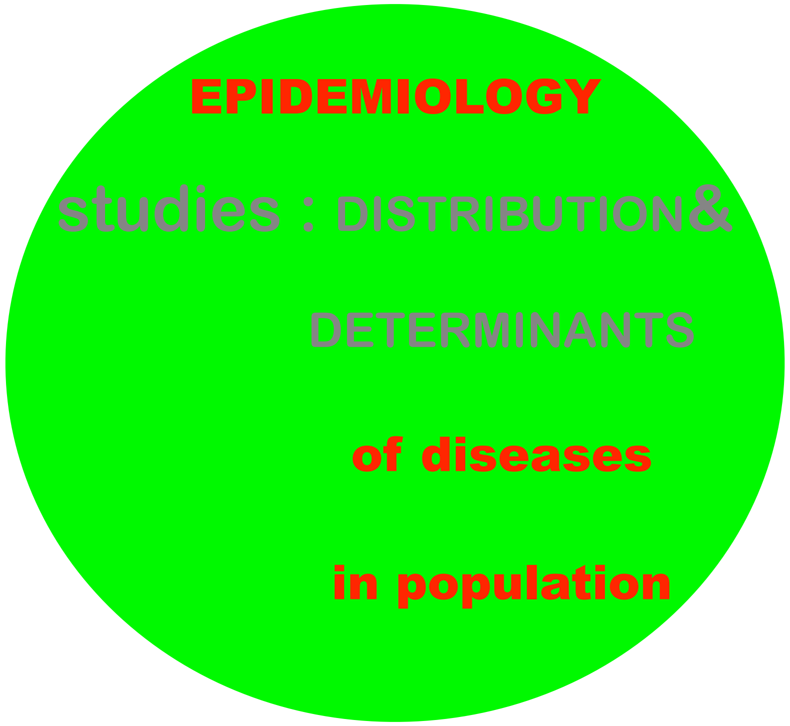
Epidemiology: the beginning

Ask Mish

- John Snow is the founder of Epidemiology
- in 1854 he investigated an outbreak of cholera in London
- he founded it was related to a water source
- he made maps and followed the addresses of dead people to find the source

epidemiology: understanding definition

Ask Mish



EPIDEMIOLOGY

studies : DISTRIBUTION&

DETERMINANTS

of diseases

in population

- Distribution = presence in the world: endemic,epidemic, pandemic
- Level of presence in any part of the world is assessed through RATES (number of diseases in population)
- Determinants refers to causes and risk factors

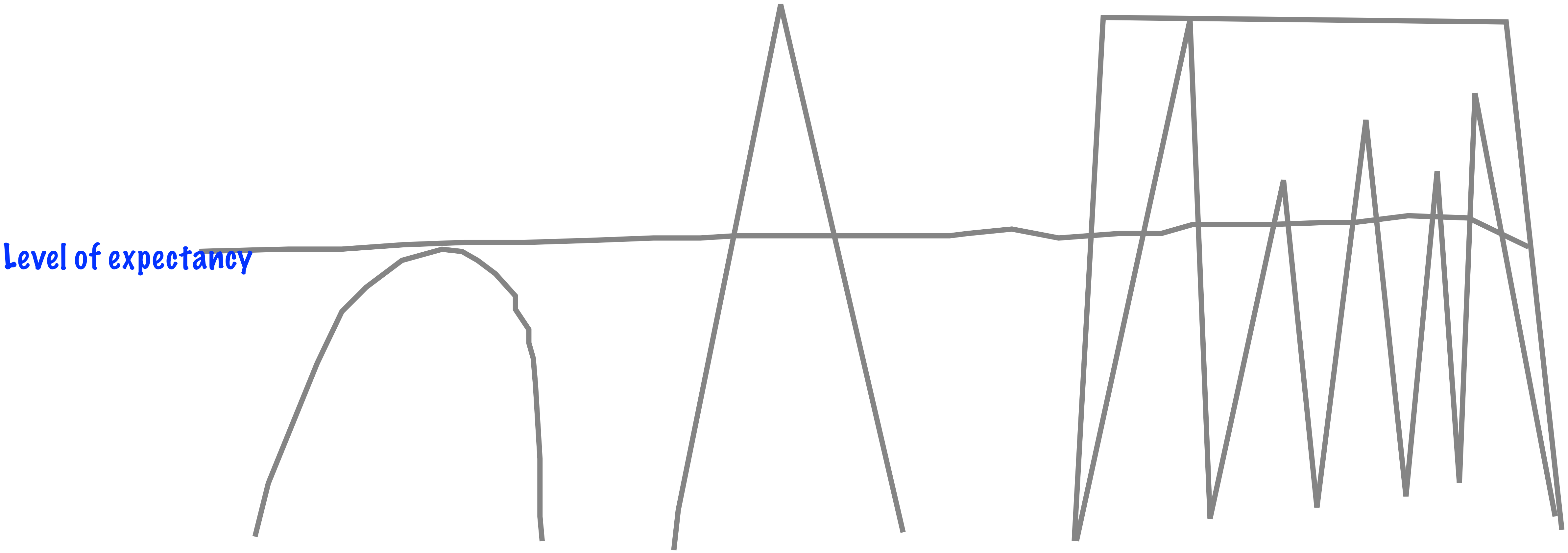
epi-demio-logy

Ask Mish

- epi=on/ upon, demos=people, logos=science, all from Gr.
- studies DISEASES among people :distribution&determinants
- **DISTRIBUTION of DISEASES**
- endemic at expectation
-
- epidemic disease above expectation in one point
- pandemic disease above expectation in many points/or spreading
- beyond national borders (worldwide)

endemic vs epidemic vs pandemic

Ask Mish



ENDEMIC :

below or @ the level of expectancy

EPIDEMIC:

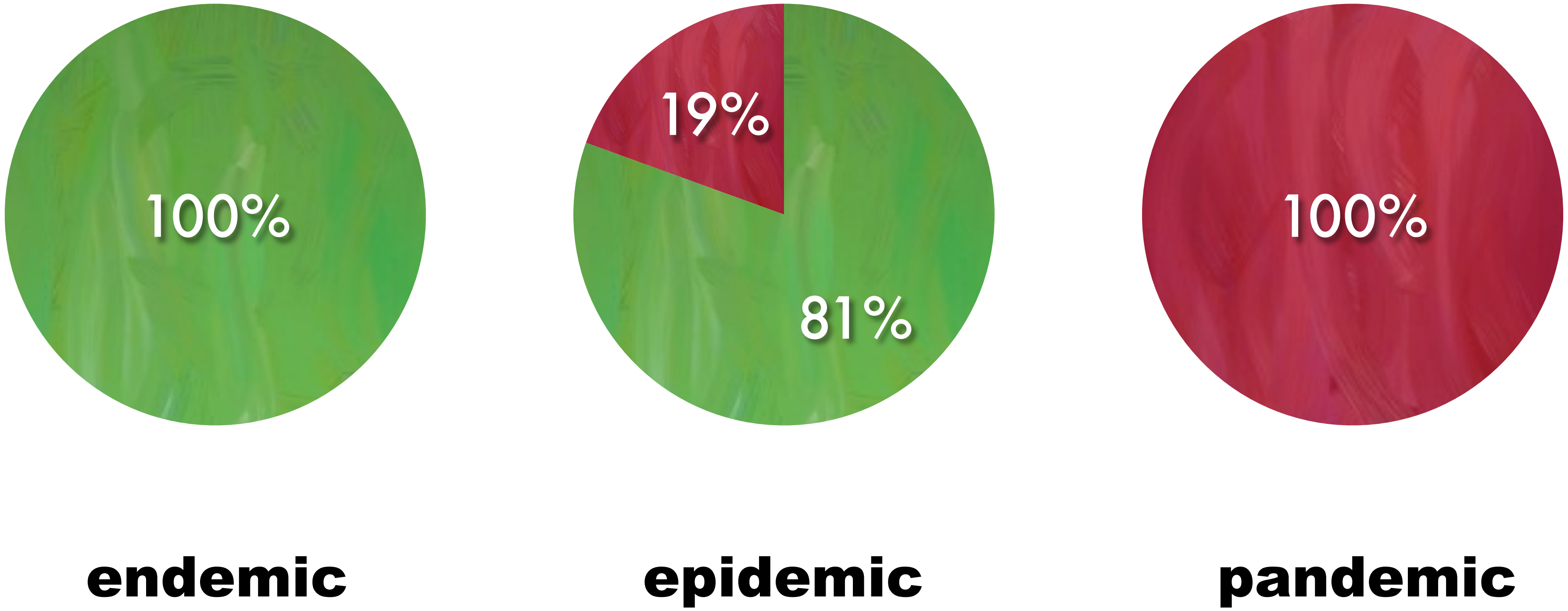
above the level of expectancy in one point meaning a limited territory

PANDEMIC:

above the level of expectancy in many points (territories) in the same time or spreading from one territory to another (beyond national borders) sometimes worldwide

another way: epidemic vs pandemic

Ask Mish



*green color: under the level of expectancy for a disease

*red color: above the level of expect.

*percentages in the middle graph are not important, just meaning a limited territory above the level vs last graph where is above the level and spreading worldwide

assessing level of disease using rates

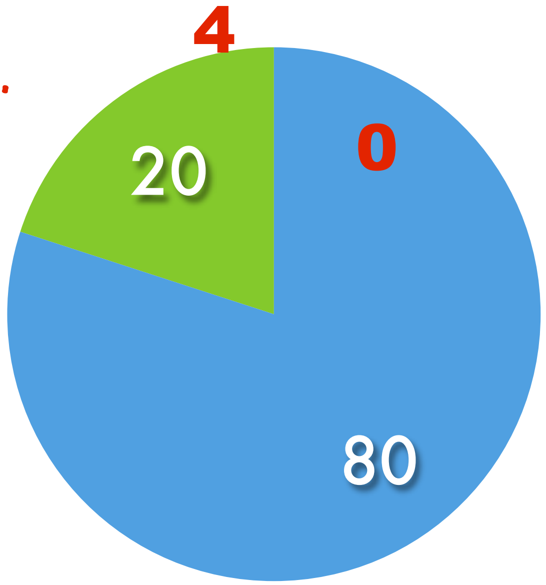
Ask Mish

- Rates are ratios (numerator/denominator)
- in epidemiology : **# diseases/#population***
- **# = number**
- ***population at risk:** *susceptible to a given disease*
- if refer to **total population** we have **crude rates**
- if refer to **group of population** we have **specific rates** (e.g.:gender, age, marital status, socioeconomic status)
- if rates are adjusted to allow comparison: **adjusted rates**(e.g comparing the same age group)

examples of crude vs specific vs adjusted

Ask Mish

POPULATION 1
100 people
4 diseased p.

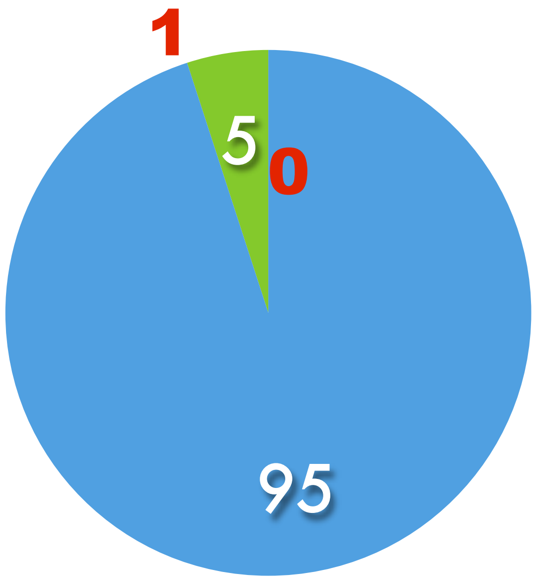


COLORS: age groups :
blue: < 60 yo
green > 60 yo

NUMBERS in white:
number of people investigated
for each age group

NUMBERS in red:
number of diseased people in each age group

POPULATION 2
100p.
1 d.



CRUDE RATE:	**4/100 (1)**	**DIFFERENT**	**1/100 (2)**
SPECIFIC RATE:	**4/20 & 0/80 (1)**	**SAME**	**1/5 & 0/95 (2)**
ADJUSTED RATE:	**4/20 (1)**	**SAME**	**1/5 (2)**
	0/80 (1)		**0/95 (2)**

question:

Ask Mish

- **What is the rate of AIDS in US** ?
- **220/100K**
- **90/100K**
- **500/100K**
- **15/100K**
- **65/100K**

answer : D

Ask Mish

- **Any disease in the US is <50/100K !!!!!!!!**

most important rates in epidemiology

Ask Mish

- **Incidence**= rate of occurrence of new cases of disease among total population in a period of time (never a point in time)
- **I = new cases/total population x 100**
- **Prevalence** = rate of all existing cases of disease among total population in either a point in time or a period in time
- **P = all cases/total population x 100**
- P = I x average time duration of disease, meaning all new cases that are not solved become cases in prevalent pot

incidence vs prevalence

Ask Mish

	Incidence	Prevalence
refer to	occurrence of new cases (rate)	occurrence of all existing cases (rate)
among	all population	all population
time	period of time	period of time or a point in time

Attack rate*

Cumulative incidence*

Attack rate:

Ask Mish

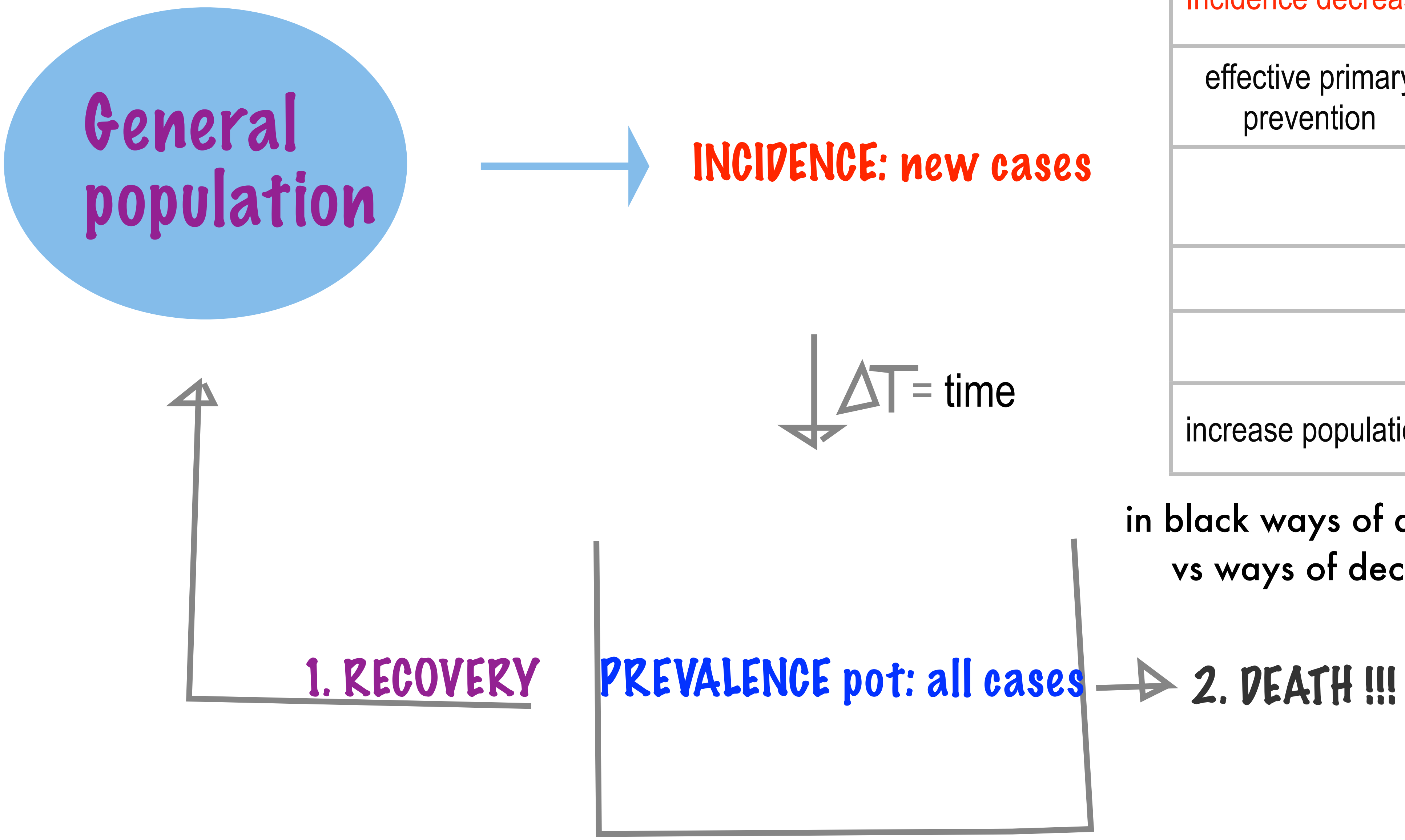
- used instead of incidence
- during a disease outbreak in a narrowly-defined population over a short period of time
- **AR= #affected/#exposed**
- e.g. :AR in case of food poisoning in a restaurant

Cumulative incidence(Proportion incidence)

Ask Mish

- is an incidence in a defined period of time
- in this case you are not interested what date exactly happened but you add all new cases in the period, that is why is called cumulative
- it is expressed as a proportion so it is also called proportion incidence

incidence and prevalence



Ask Mish

Incidence decrease	Prevalence decrease
effective primary prevention	decreased incidence
	new cases recover quickly in time
	increased recovery
	increased death
increase population	increase population

in black ways of decreasing Incidence
vs ways of decreasing Prevalence

other rates : vital rates (1)

Ask Mish

- Birth rate:number of births @1000people
- **Birth rate**=births/population x 1000
- Death rate:number of deaths @1000people
- **Death rate**=deaths/population x1000
- Case fatality rate: number of cases that end up in death;
- **CFR**= deaths from a cause/diseased x100
- **Proportionate mortality rate(PMR)**= deaths from a cause/all deaths x 100

other rates:vital rates(2)

Ask Mish

- Fertility rate=number of children/fertile woman
- **Fertility rate** = births/women of childbearing age(15-49) x1000
- **Infant mortality rate** : deaths 0-1 yo from 1000 live births; neonatal 0-28day, perinatal: 28days-1 year
- **IMR** = deaths 0-1 yo/live births x1000

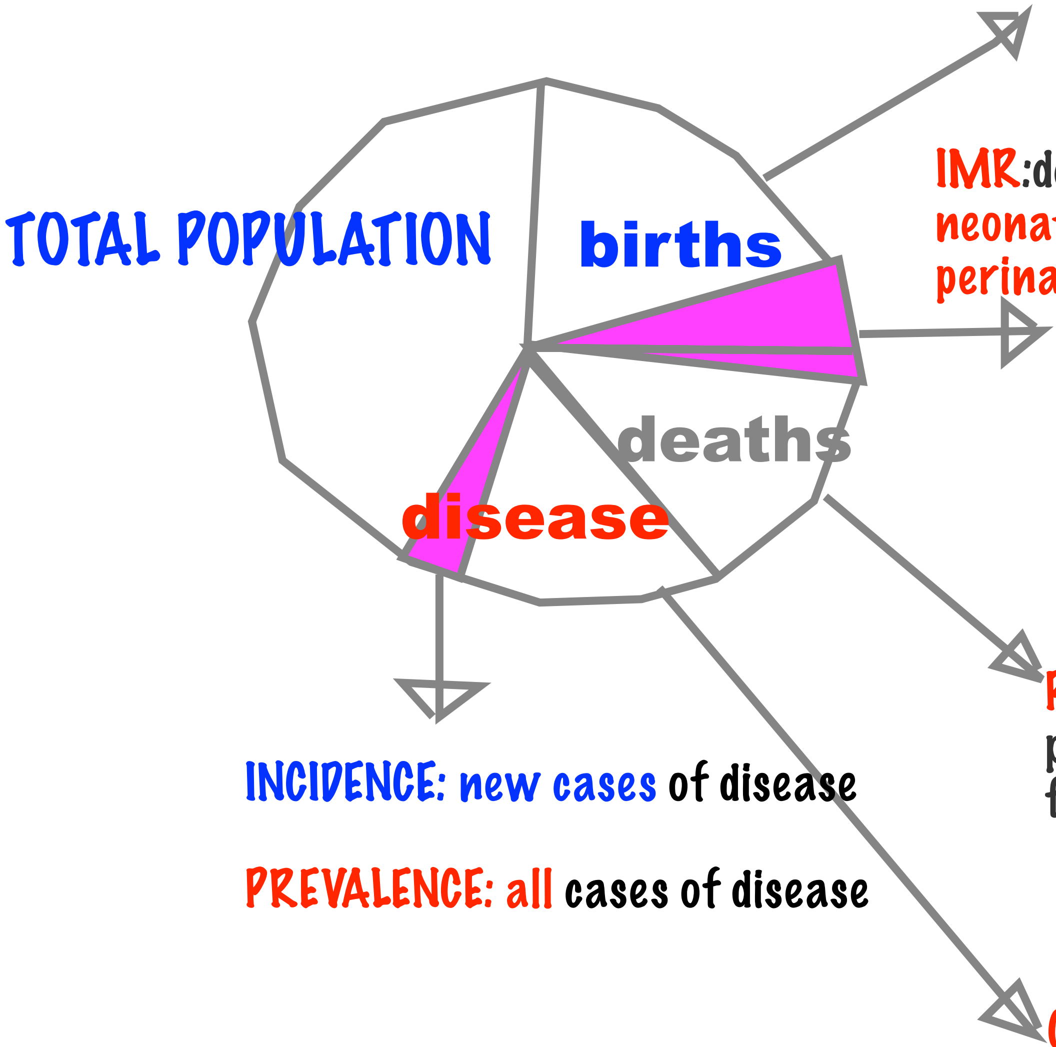
infant mortality rate in US

Ask Mish

- **IMR in US** is **7/1000**; different among ethnic groups:whites &hispanics - 6/1000, black- 13/1000
- Major causes:
- **1.genetic**
- **2.low birth weight <1500 g**
- **3.SIDS** - never let infants sleep on the belly
- Low birth weight =1st cause in blacks, **SIDS**=1st cause in native Americans

visualization of crude vs specific rates

Ask Mish



FERTILITY RATE : births/women 15-49 yo

IMR:deaths 0-1 yo from all living births:
neonatal:0-28d
perinatal:28d-1yo

PMR :
percent of deaths from a cause
from all deaths

CASE FATALITY RATE:
percent of disease from a cause
that end up in death

INCIDENCE: new cases of disease

PREVALENCE: all cases of disease

CRUDE	SPECIFIC
birth rate	fertility rate
death rate	infant mortality rate
prevalence	PMR
incidence	CFR

Vital rates in epidemiology:

Ask Mish

	BIRTHS	DEATHS	DISEASES
Total population /1000	Birth rate	Death rate	Incidence (new) Prevalence (all)
Groups of population:infants/ mothers /100,000		IMR MMR	
Groups of population: diseased/ dead /100		CFR PMR	

Epidemiology: Risk, Risk Factors and Causes

Ask Mish

- definition of risk in epidemiology
- risk vs incidence
- definition of risk factors and causes
- importance of risk factors and causes

Risk in epidemiology:

- the probability of occurrence of a new case in a time period is called **RISK.**
- if the period of time you choose is lifetime then it is a lifetime risk.

risk = probability of incidence

Ask Mish

	Incidence	Risk
refer to	occurrence of new cases (rate)	probability of occurrence of new cases (rate)
among	all population	all population
time	period of time	period of time

Epidemiology: determinants

Ask Mish

DETERMINANTS: CAUSES & RISK FACTORS

	Causes	Risk Factors
refer to	personal habits & environmental factors	personal habits & environmental factors
action	DETERMINES	INCREASE the probability of
on:	the occurrence of disease	the occurrence of disease

Determinants: analysis

Ask Mish

1. Knowing causes and risk F

-used to prevent and control disease by removing causes and risk F

2. Not knowing causes and risk F

-an epidemiological study is recommended to determine them

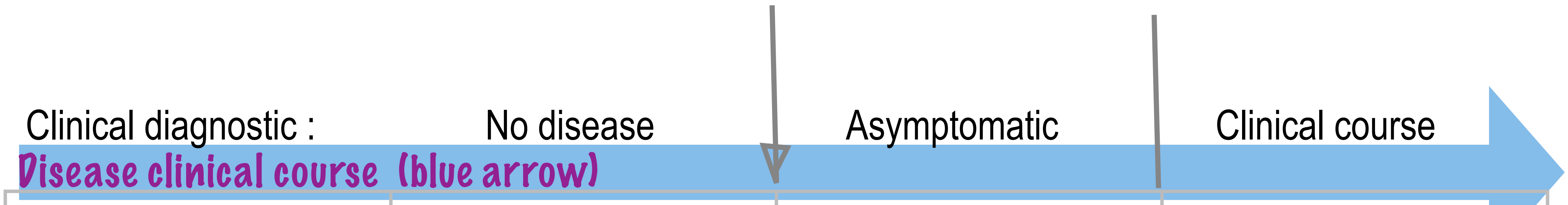
Epidemiology: prevention of disease

Ask Mish

- The next slides are about the level of prevention of a disease: primary, secondary and tertiary prevention
- Related to secondary prevention there are few slides about understanding screening tests. This includes: screening test table, concepts like: sensitivity, specificity, positive predictive value, negative predictive value, accuracy. Also includes one example of how to calculate all the values above and relationship sensitivity vs specificity for the same test.

Levels of prevention: 1

Ask Mish



ONSET disease

Disease clinical course (blue arrow)

Clinical diagnostic :	No disease	Asymptomatic	Clinical course
Levels of prevention	Primary	Secondary	Tertiary
Ways of prevention	remove risk factors	early detection & treatment	reduce complications
Examples of prevention	vaccines, folate, exercise, seat belts	screening tests	Beta-blockers post MI

Levels of prevention: 2

Ask Mish

- Preventing new cases of disease(incidence)=primary; e.g.: vaccines, spreading information about disease
- Preventing disease(prevalence) by detecting it early = secondary; e.g.: screening tests, quit smoking advice
- Preventing disease by applying recovery programs= tertiary; recovery after myocardial infarct

screening tests design

Ask Mish

	DISEASED people	HEALTHY people	
TEST POSITIVE	true positive	false positive	positive predictive value
TEST NEGATIVE	false negative	true negative	negative predictive value
	sensitivity	specificity	accuracy

screening tests concepts

Ask Mish

- **Sensitivity**= percentage of people w/disease detected by test
- **Specificity**=percentage of healthy p.detected by test
- **PPV**= if a test is positive what is the chance to be true
- **NPV**=if a test is negative what is the chance to be true
- **Accuracy :** what is the chance that a test (+ or -) is true = tp+tn/all tested(tp +tn+fp+fn); chance=percent

screening test example

Ask Mish

	disease:100	healthy:100	
test positive	80 true positive	10 false positive	**PPV:80/90**
test negative	20 false negative	90 true negative	**NPV: 90/110**
	SENSITIVITY **80/100**	**SPECIFICITY** **90/100**	**Accuracy:** **80+90/200**

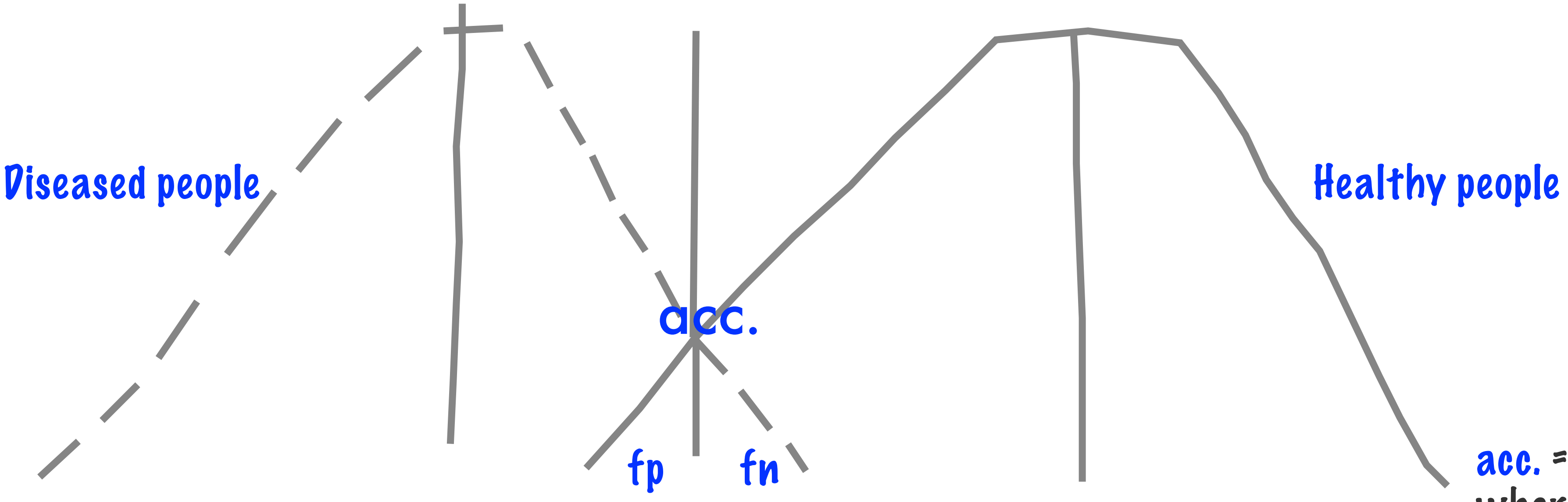
PPV vs Sensitivity vs Specificity

Ask Mish

- PPV = TP/FP low usually because TP low in comparison to FP
- SENSITIVITY = TP/FN if high, because TP high in comparison to FN
- INCREASING SENSITIVITY usually by decreasing the screening test threshold which will produce an increase in TP but also in FP
- Relationship sensitivity vs specificity: any increase in FP will decrease specificity! Remember SPECIFICITY = TN/FP. We can say any increase in sensitivity will produce a decrease in specificity!

screening tests diagrams

Ask Mish



acc. = accuracy point, where accuracy is best

fn = false negatives

fp = false positives

Moving midline to LEFT = increase SENSITIVITY will decrease fp but increase fn.
Increase fn means decrease SPECIFICITY.

CONCLUSION : Increase SENSITIVITY for a test means decrease SPECIFICITY for the same test.

Epidemiology: studies

Ask Mish

- refer to observational(non-intervention) studies vs interventional studies
- definition of each type of study
- ways to test a hypothesis in both observational studies and interventional studies.
- Most common mistakes in studies aka bias in research

Epidemiological studies:

Ask Mish

- 1.observational
- 2.experimental

observational vs experimental studies

Ask Mish

- **OBSERVATIONAL =** non interventional studies
- **EXPERIMENTAL =** interventional studies

1.Observational studies:

Ask Mish

- **1. case report**
- **2. case series**
- **3. cross - sectional**
- **4. case - control**
- **5. cohort**

2. Experimental studies:

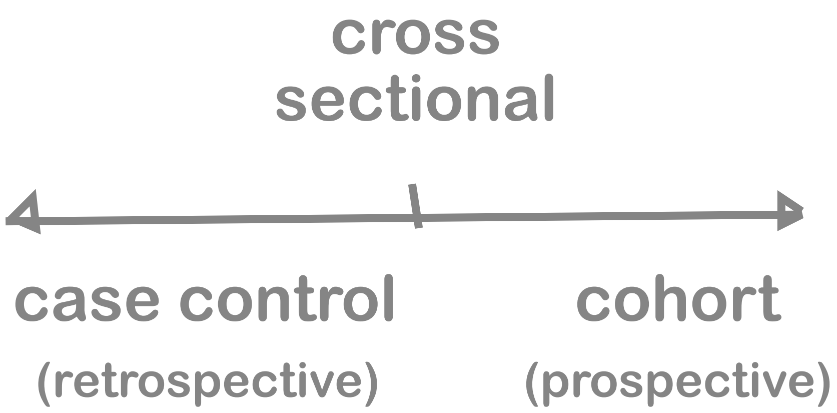
Ask Mish

- **RCT = random control trials**

what does each study mean:

Ask Mish

- **1.CASE REPORT or CASE SERIES REPORT=** report of one case or a small # of cases of a disease with low prevalence
- 2.**CROSS SECTIONAL=** disease vs non disease one point in time
- 3.**CASE CONTROL=** one disease followed back in time to find associated causes and risk F
- 4.**COHORT** = one risk F followed in the future to find associated disease(s)
- 5.**RCT** = interventional study to verify a hypothesis vs all the above which are observational (non-interventional)



cross sectional

case control

(retrospective)

cohort

(prospective)

design of an observational study

Ask Mish

	DISEASE	NO DISEASE
EXPOSED	**A**	**B**
NON EXPOSED	**C**	**D**

*2 groups of people: exposed to a risk F vs non-exposed

A,B,C,D= number of people from
the 2 groups above that have disease/not

hypothesis testing in observational studies

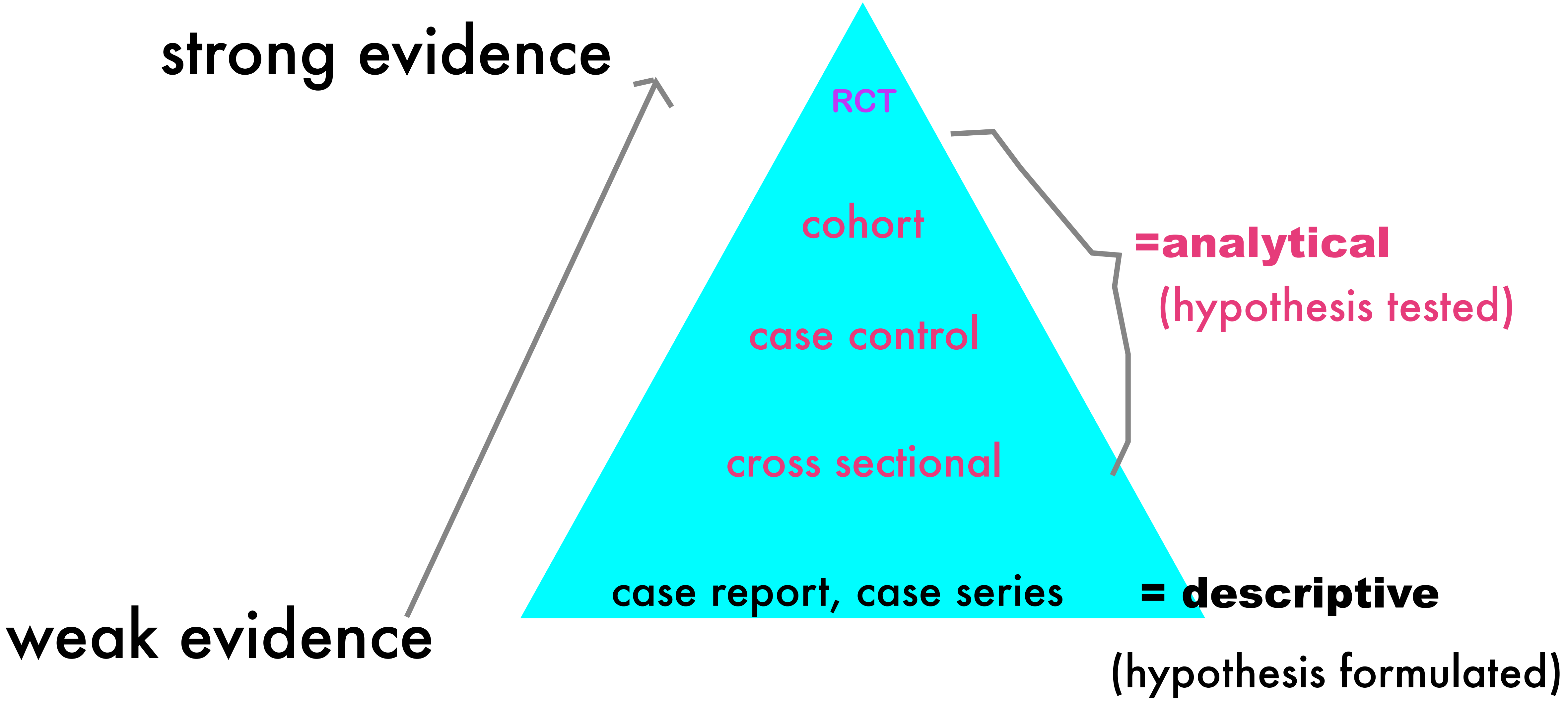
Ask Mish

- if A>>>C then A RISK FACTOR could be highly probable.
- A hypothesis is formulated but cannot be tested in case report and case series report also called DESCRIPTIVE studies
- Hypothesis could be verified in cross sectional, case control and cohort study. What we want to see is if A>C due to hazard or the value is statistically significant = ANALYTICAL studies
- Hypothesis testing : uses formulas for each study to see if the association of risk F w/ disease is due to hazard or is statistically significant. Below are the name of formulas used to test this association for each study.
- cross sectional: chi square
- case control: odds ratio
- cohort: relative/attributable risk

	DISEASE	NO DISEASE
EXPOSED	**A**	**B**
NON EXPOSED	**C**	**D**

graphical representation of studies

Ask Mish



strong evidence

RCT

cohort

case control

cross sectional

case report, case series

weak evidence

=analytical

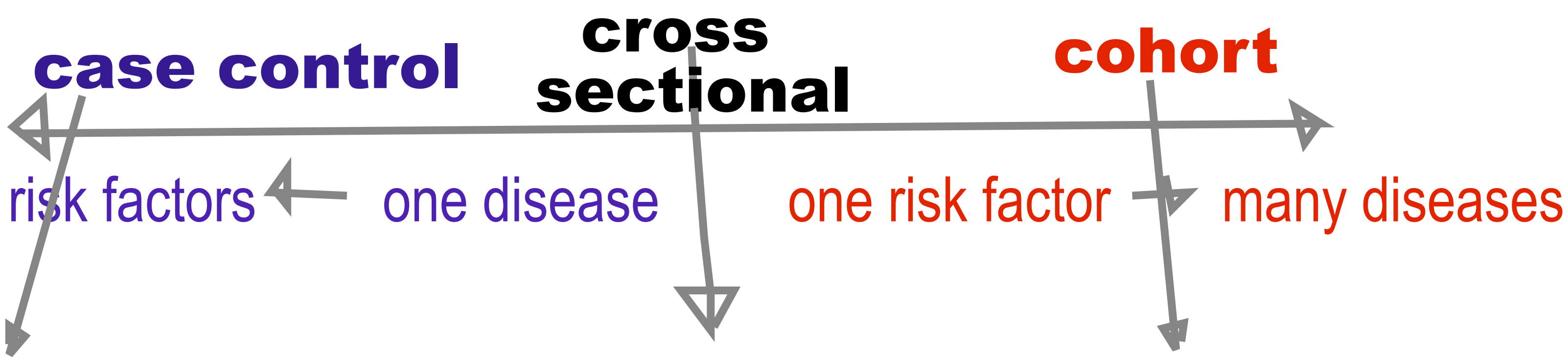
(hypothesis tested)

= descriptive

(hypothesis formulated)

Hypothesis testing in observational studies(1): concepts

Ask Mish



case control

cross sectional

cohort

risk factors ← one disease

one risk factor → many diseases

odds ratio (OR)

chi square

relative risk, attributable risk (RR,AR)

odds= interested/uninterested

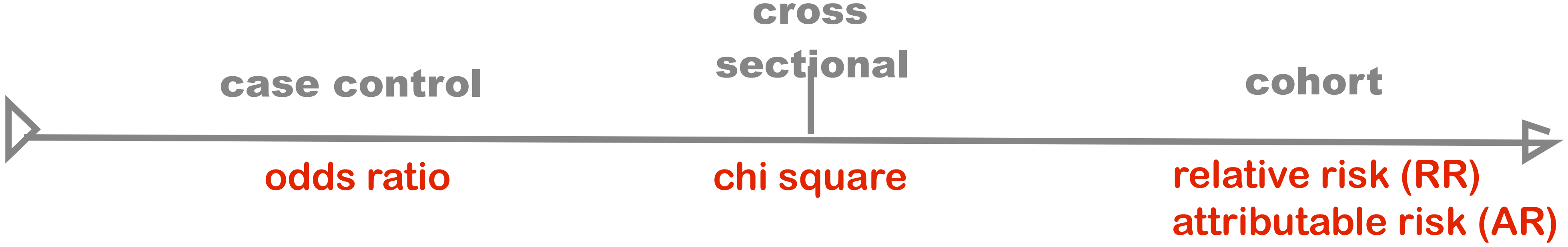
this case: Exposed/Non exposed

RRR* or RRI* relative risk reduction o relative risk increase of the exposed

	disease	no disease
exposed	a	b
non exposed	c	d

Hypothesis testing in observational studies(2) formulas

Ask Mish



case control

cross sectional

cohort

odds ratio

chi square

relative risk (RR)
attributable risk (AR)

OR = odds of exposure for cases divided by odds of exposure for controls

a/c : b/d=ad/bc

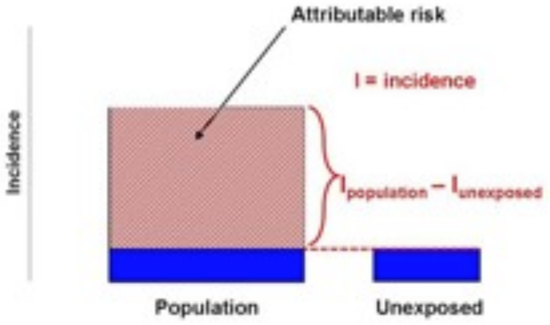
	disease	no disease
exposed	a	b
non exposed	c	d

RR = incidence among exposed vs incidence among unexposed

a/all :c/all = a/c (DIVISION)

Question for RR:
how much more likely?

RRR* or RRI* = |1- RR|



AR = also called absolute risk reduction =ARR

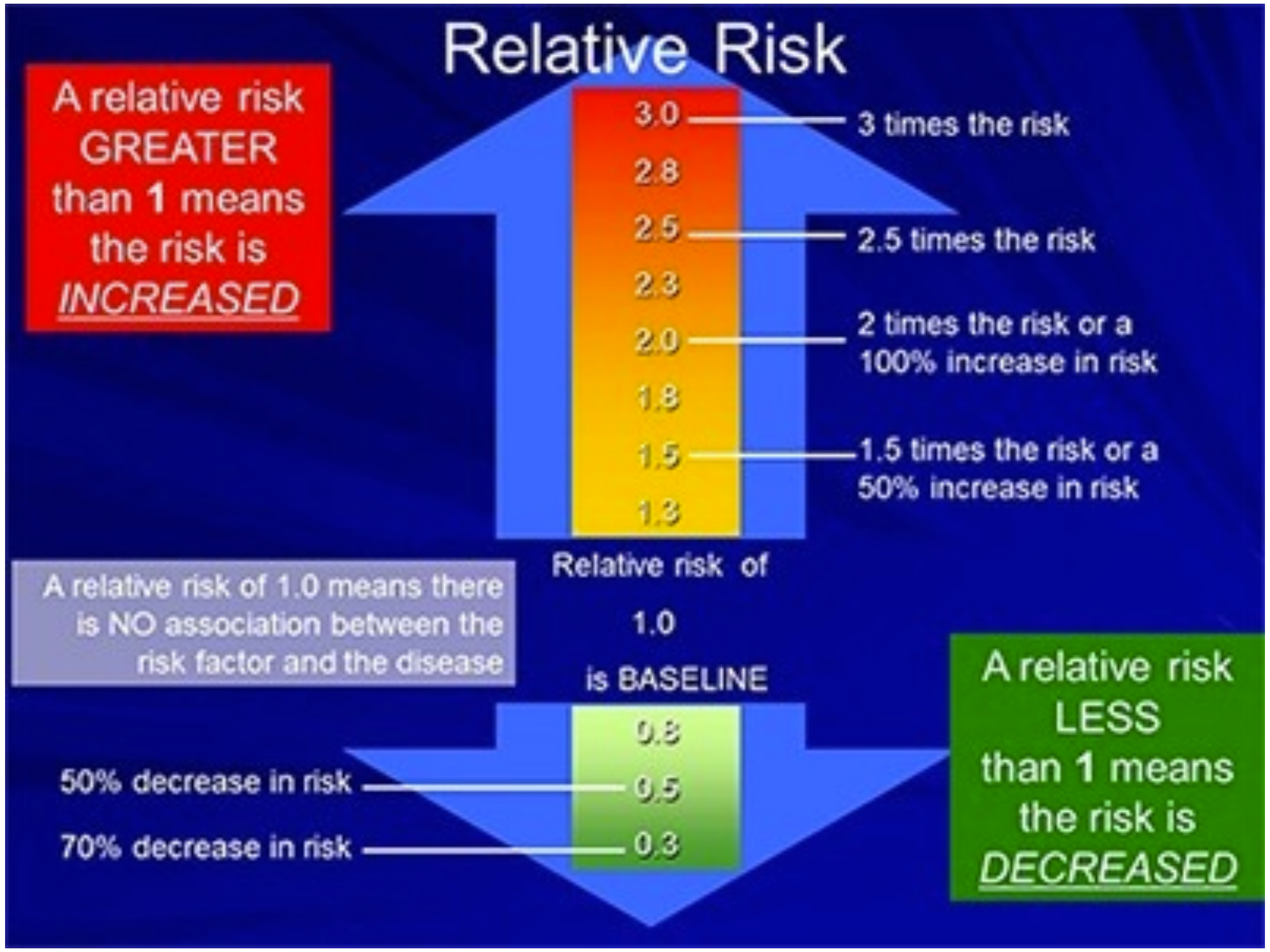
AR = incidence in the exposed -incidence in the control

Question for AR: how many more cases in E vs U?

NNT *& NNH* = 1/ ARR

How to interpret Relative Risk and Odds Ratio

Ask Mish


Relative Risk
A relative risk GREATER than 1 means the risk is INCREASED
3.0
2.8
2.5
2.3
2.0
1.8
1.5
1.3
3 times the risk
2.5 times the risk
2 times the risk or a 100% increase in risk
1.5 times the risk or a 50% increase in risk
A relative risk of 1.0 means there is NO association between the risk factor and the disease
Relative risk of 1.0 is BASELINE
0.8
0.5
0.3
50% decrease in risk
70% decrease in risk
A relative risk LESS than 1 means the risk is DECREASED

- HOW TO INTERPRET RR AND ODDS RATIO
- = 1 means no association disease-risk factor, >1 is increased risk for disease in exposed and <1 means decreased risk of disease in exposed.
- Calculation: RR=2.5 means 150% increased risk; RRI = |1-RR|x100 so RRI=|1-2.5|x100 ; RRI=150% . RR= 0.3 means 70% decreased risk; RRR= |1-0.3|x100= 70%
- Application of RR in clinical practice:
- Let's suppose we have a study in which we used Estrogen/ Progesterone to decrease the risk of CAD. Final result :RR=0.39 meaning RRR=61% equals 61% less risk of disease in the E/P group. If you have a woman with 20% Framingham CAD risk how much will be her risk of CAD if she receives E/P? Multiply 0.39 (RR)x 20%(Framingham.risk) = approximative 8% risk of CAD with E/P.
- How big should be RR or Odds ratio?
- Depends on study. RCT, least prone to bias, a small variation is enough; in COHORT study RR> 3 , in CASE CONTROL study OR>4 (Case control has a greater risk of bias)

NNT(number needed to treat) and NNH (number needed to harm)

Ask Mish

1.NNT/NNH

	NNT	NNH
refer to	treatment as cure	treatment w/ side effects
# people treated	to prevent 1 case (disease)	to prevent 1 case (disease)

2.Total:100 people

	disease	no disease
exposed:50	5	45
non exposed:50	0	50

1.NNT =NNH :# people you need to treat/harm to prevent the appearance of 1 new case of disease (definition in table 1)

2.Example : How to calculate NNT/NNH from table 2

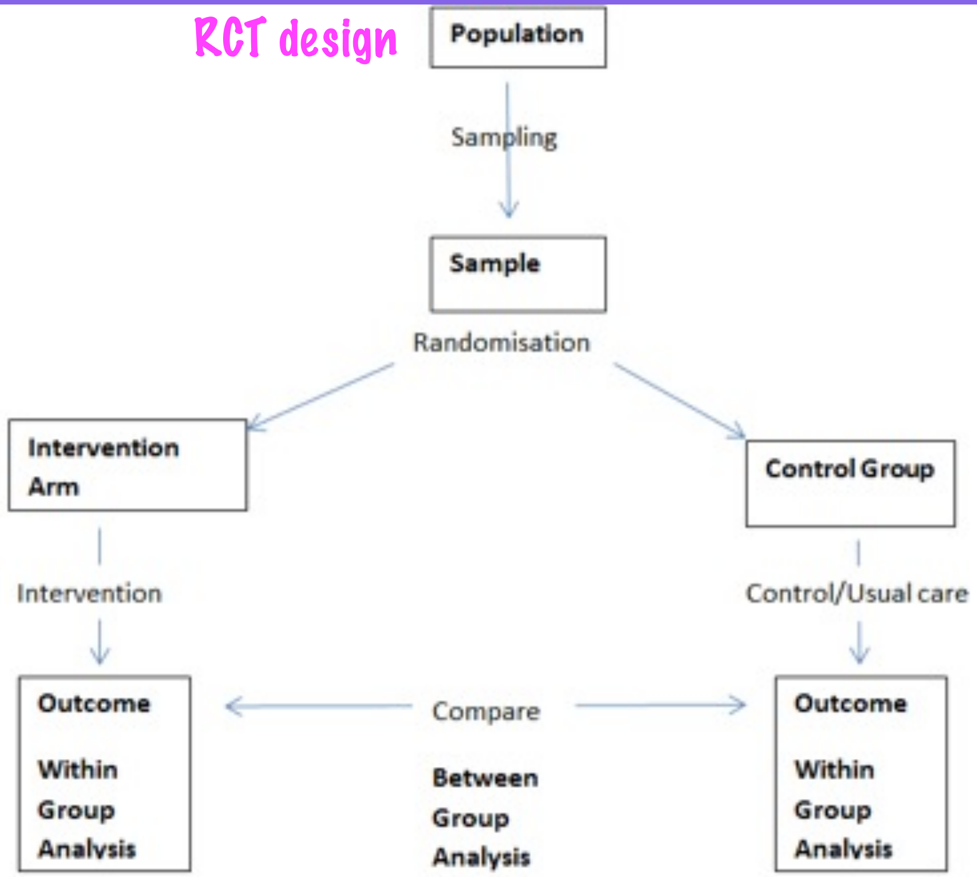
NNT/NNH : if you treat all 100 people you prevent the 5 cases of disease
so you need to treat x=NNT to prevent 1 case of disease
apply 3 simple rule: NNT= 1x100/5=20
meaning you need to treat 20 people in order to prevent 1 case of disease

also calculate : NNT/NNH = 1/ARR

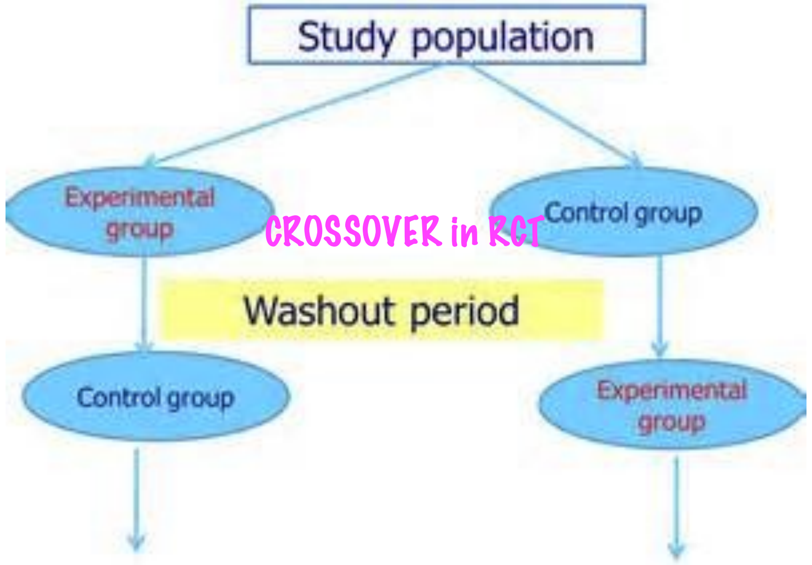
Intervention studies: Clinical Trials

Ask Mish

RCT design



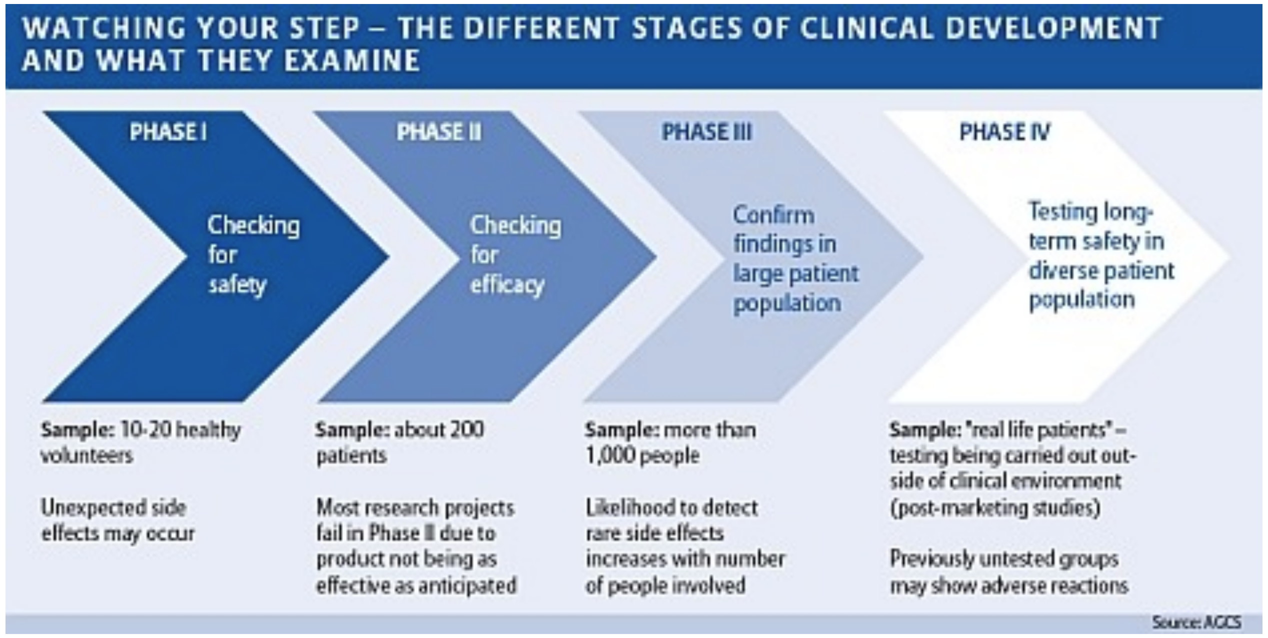
CROSSOVER in RCT



- CLINICAL TRIAL= intervention studies for the benefit of patients
- usually involves the administration of a test regimen to evaluate its safety and efficacy
- study has 2 arms: people on drug(intervention) and people on placebo (control) group
- RCT=randomized controlled clinical trial : subjects randomly allocated into one group, intervention or control
- double blind: neither subject nor researchers know which group the subject is, intervention or control
- crossover study: switch arms of the study one point in time, intervention group becomes control and control becomes intervention
- community trial: an entire community receives a regimen testing how the regimen works in the real world

FDA approval for a drug : 3 Clinical Trial Phases

Ask Mish



WATCHING YOUR STEP – THE DIFFERENT STAGES OF CLINICAL DEVELOPMENT AND WHAT THEY EXAMINE

PHASE I

Checking for safety

Sample: 10-20 healthy volunteers

Unexpected side effects may occur

PHASE II

Checking for efficacy

Sample: about 200 patients

Most research projects fail in Phase II due to product not being as effective as anticipated

PHASE III

Confirm findings in large patient population

Sample: more than 1,000 people

Likelihood to detect rare side effects increases with number of people involved

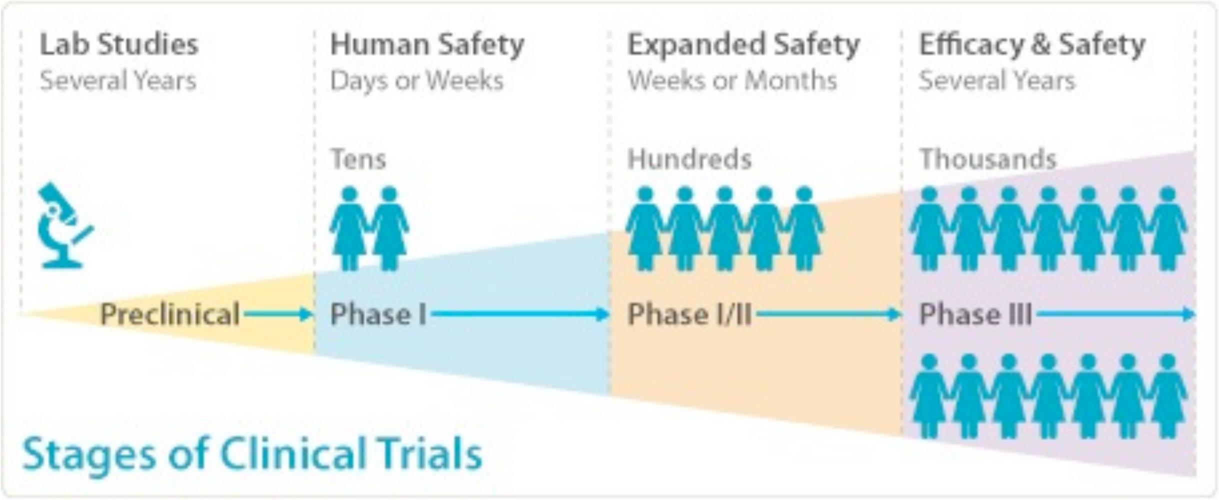
PHASE IV

Testing long-term safety in diverse patient population

Sample: "real life patients" – testing being carried out outside of clinical environment (post-marketing studies)

Previously untested groups may show adverse reactions

Source: AGCS



Stages of Clinical Trials

- For FDA approval 3 phases of the clinical trials must be passed:
- Phase 1: testing safety in healthy volunteers
- Phase 2: testing efficacy (dose levels) in small group of patient volunteers
- Phase 3: testing efficacy and safety in larger group of patient volunteers. Phase 3 is considered a definitive test for FDA.
- Phase 4: not necessary for FDA approval; is called post marketing survey and focuses on long term safety (e.g. Vioxx)

Bias in research

Ask Mish

Type of BIAS	DEFINITION	Important associations	Solutions
SELECTION	sample not representative	Berkson's bias = using hospital data nonrespondent bias= p.included in study are different than non-includ	random, independent sample
MEASUREMENT	gathering information distorts it	Hawthorne effect= people under observation behave differently	control group/placebo group
EXPERIMENTER EXPECTANCY	researcher's beliefs affect outcome	Pygmalion effect	double-blind design
LEAD-TIME	early detection confused w/ increased survival	benefits of screening	measure "back-end" survival(back-end=age of death for the disease)
RECALL	subjects cannot remember accurately	retrospective studies	confirm association w/ other sources
LATE-LOOK	severely diseased individuals are not covered	early mortality	stratify study by severity
CONFOUNDING	A 3rd factor is involved in various proportions in exposure-disease rel.	affects result	random selection, multiple studies

Biostatistics

Ask Mish

- STATISTICS means world expressed in numbers
- World includes:
- events= action and
- categories = structures that have names “this” or “that” and that’s why they are called nominal/categorical data e.g. gender (one category with 2 groups: males and females) , population in a study (also 2 groups : on drug and on placebo) or categories with no groups (most of them)

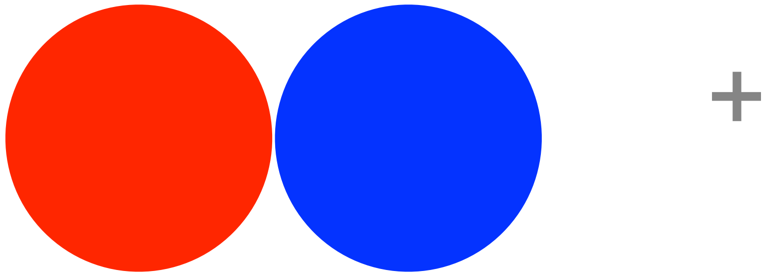
Ask Mish

2 events : probability to occur together

type of event:

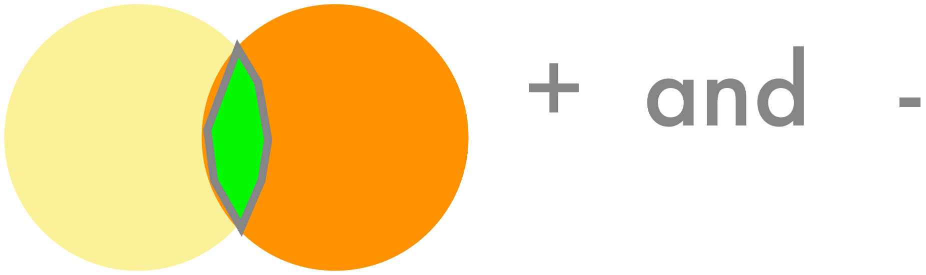
1. Independent events: no connection b/w them, e.g. blond hair and catch a cold.

Probability for a blonde to catch a cold (independent events): multiply the probability of each event expressed as hundredths.



2. Mutual exclusive events: one event excludes the possibility of the other happening in the same time, e.g. heads and tails for a coin flip

Probability to have a head or a tail when flipping a coin : ADD together the probability of each event

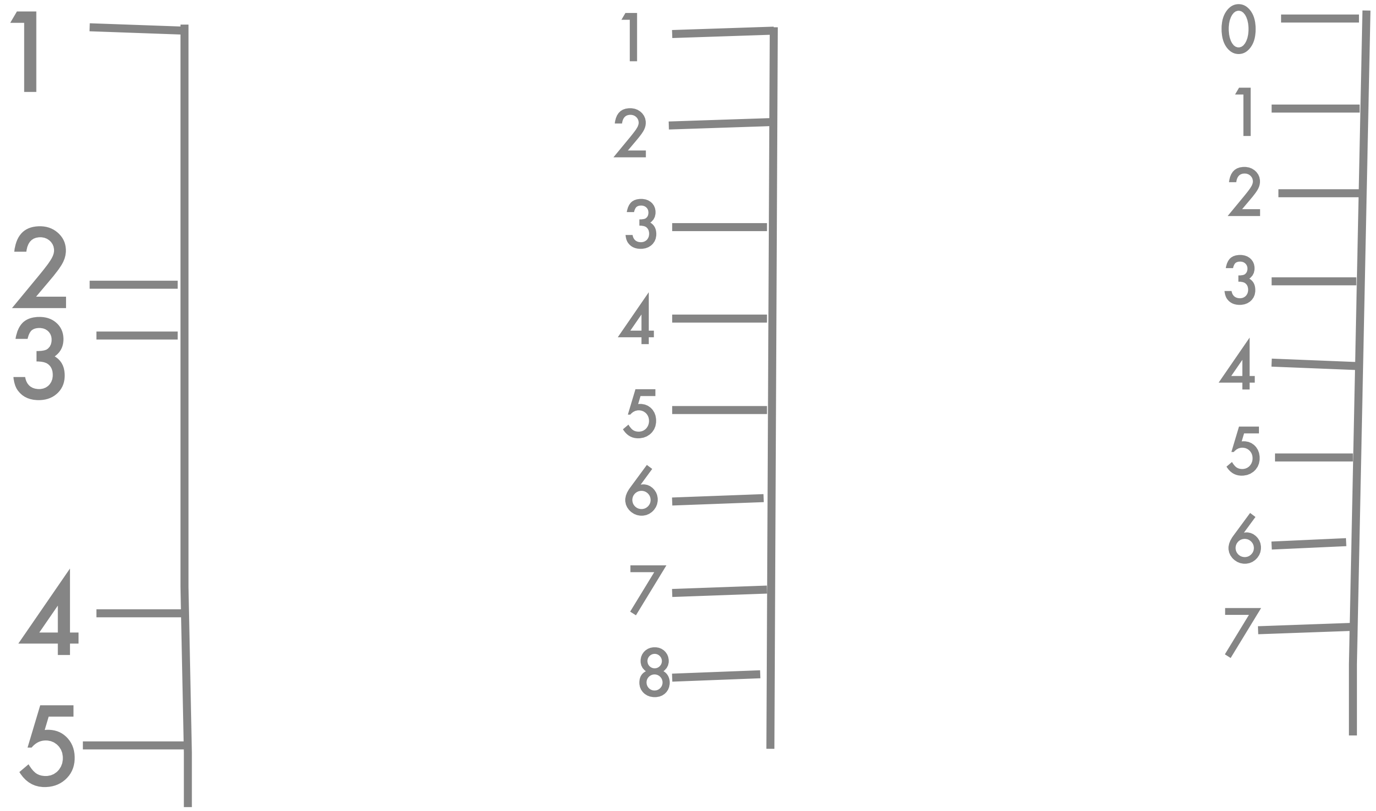


3. Non-mutual exclusive events: one event does not exclude the possibility of the other happening in the same time e.g. obese and diabetic

Probability for an obese patient to also have diabetes is add the 2 probabilities and subtract their product

categorical /nominal data measured in numbers

Ask Mish



ORDINAL data
= rank order

no similar intervals b/w rankings
any rank order

INTERVAL

similar intervals on the scale w/ **no 0**
height,weight, BP

RATIO

similar intervals on the scale
0 included
temperature in K

A nominal data can be measured using one of the three above scales
: rank order, interval or ratio

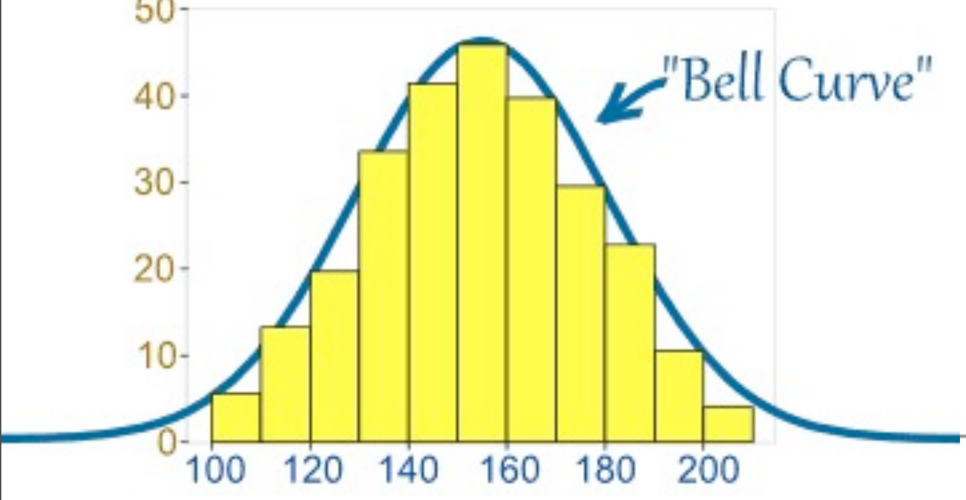
Descriptive vs Inferential Statistics

Ask Mish

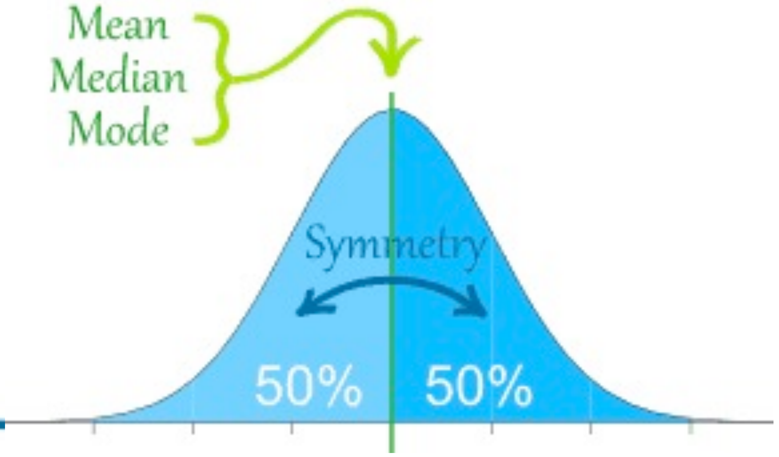
Descriptive Statistics	Inferential Statistics
measures groups/population (coz you can measure each member of the group)	takes a sample from a group and draw conclusion about the whole group (coz you cannot measure all!)
Result: distribution is a bell shape curve symmetric to a central point (mean=median=mode)	Result is expressed in confidence intervals

Descriptive statistics: normal distribution

Ask Mish



normal distribution example

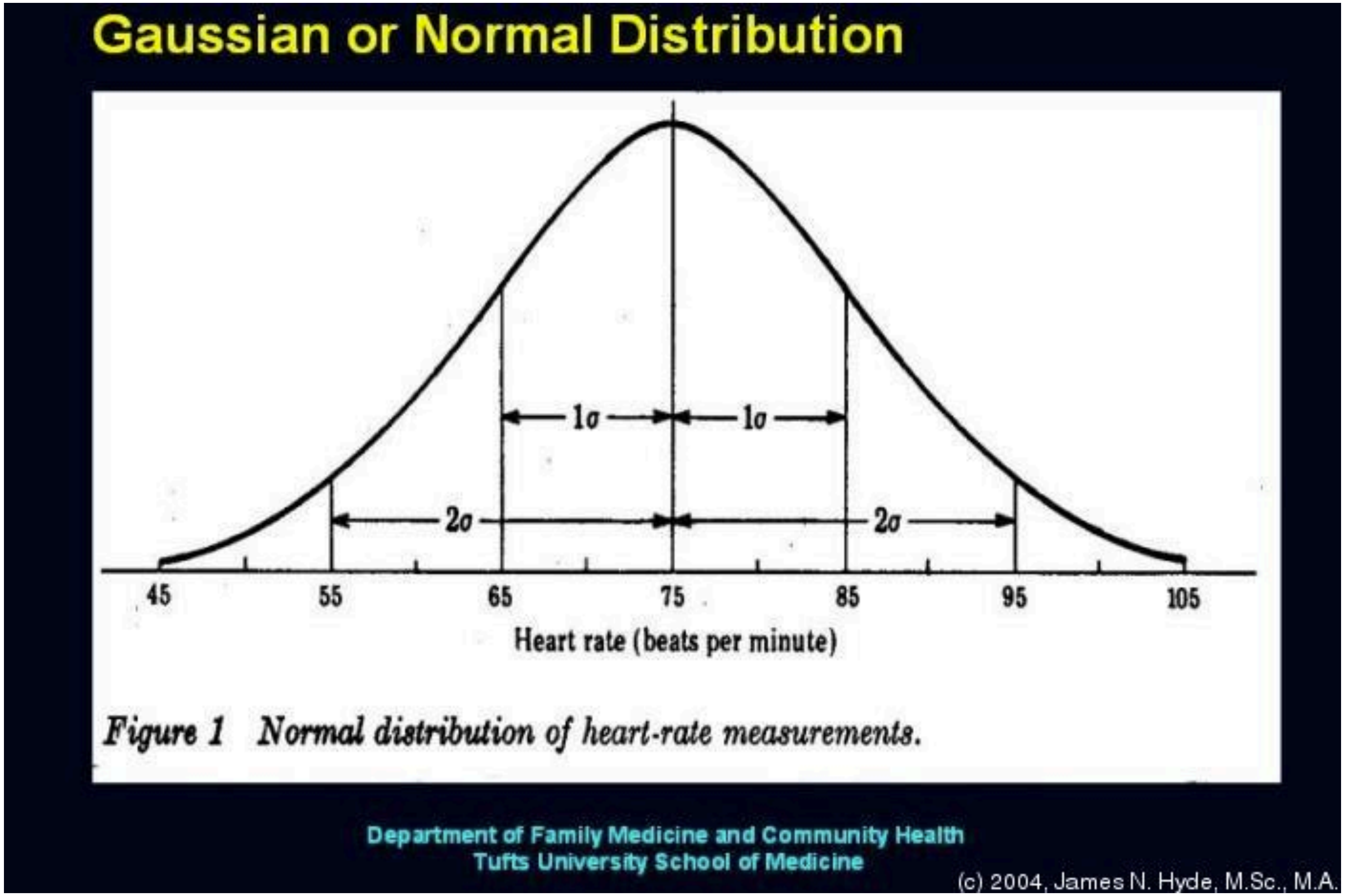


mean=median=mode

- Mean = average= add all quantities and divide by the number of quantities you added (Xo)
- Median = midpoint (Md)
- Mode = most frequent number (Mo)

Descriptive statistics:normal distribution example

Ask Mish



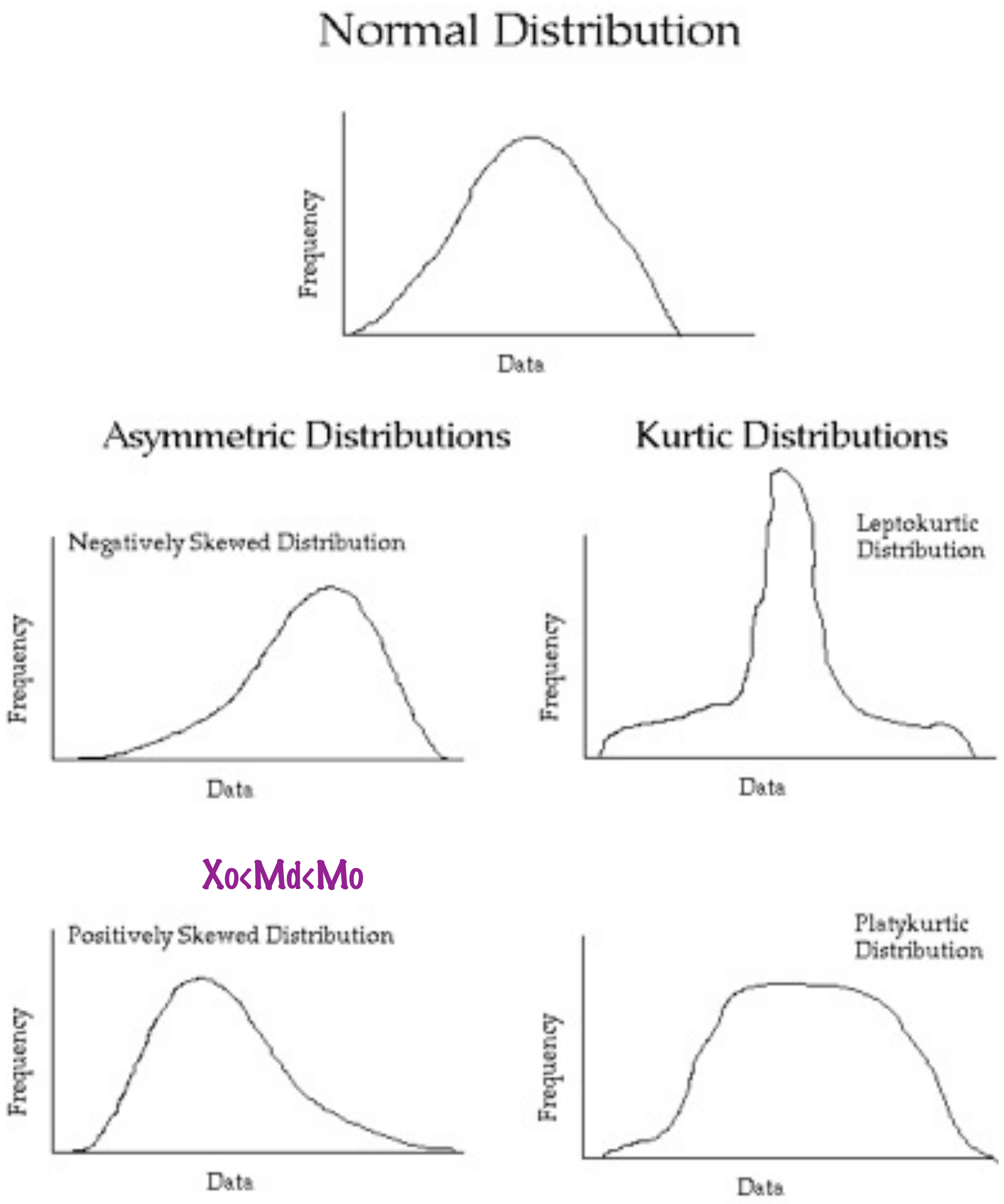
Gaussian or Normal Distribution

Figure 1 *Normal distribution of heart-rate measurements.*

Department of Family Medicine and Community Health
Tufts University School of Medicine

(c) 2004, James N. Hyde, M.Sc., M.A.

Descriptive statistics: various distributions



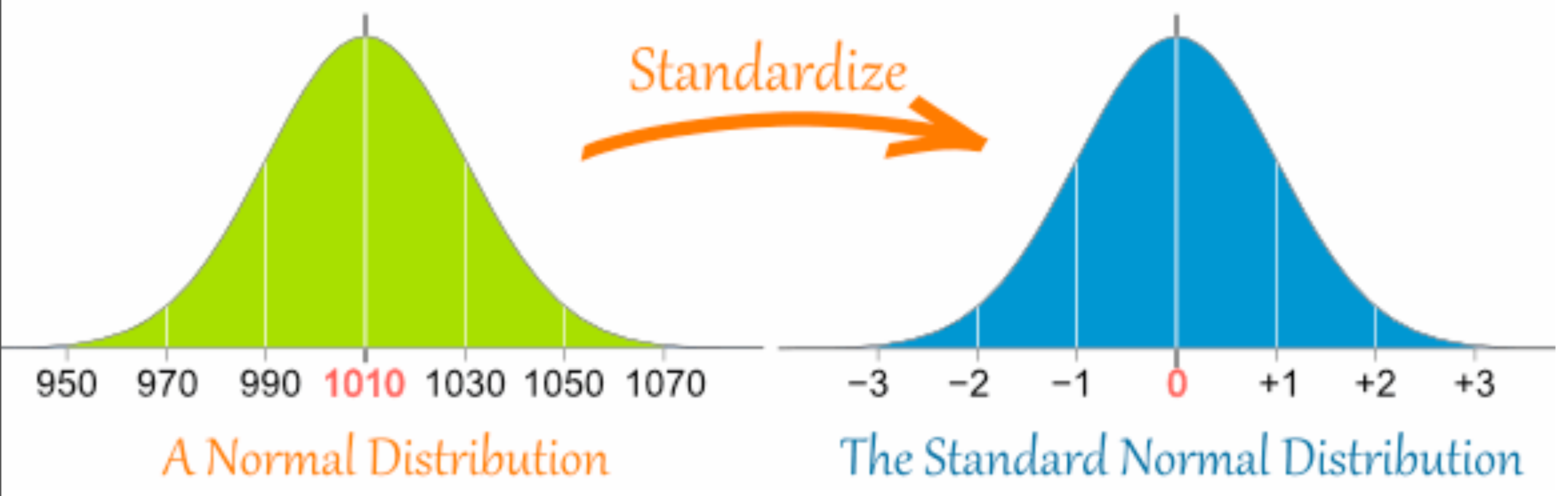
Xo<Md<Mo

Mo<Md<Xo

- ASYMMETRIC DISTRIBUTIONS:
- have a hump and a tail. If the tail is on negative side there is a negatively skewed distribution and if the tail is on positive side there is a positively skewed distribution. In both cases, mean is not equal to mode and is different from median.
- KURTIC DISTRIBUTIONS:
- LEPTOKURTIC: peaked
- PLATYKURTIC: flattened

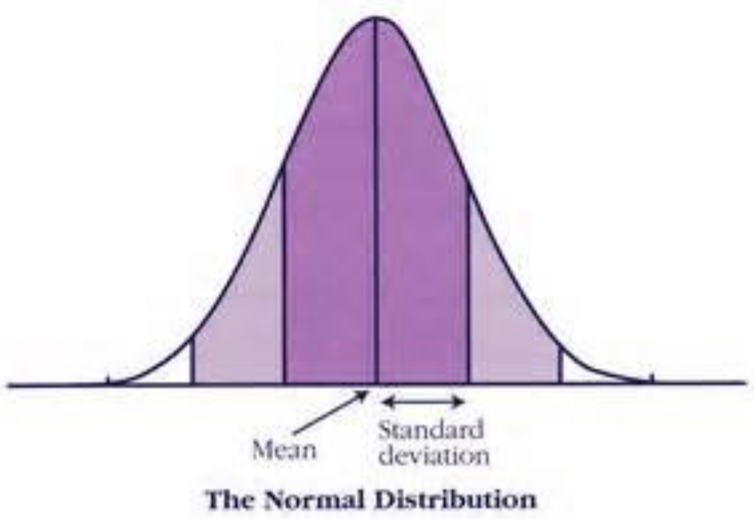
Inferential statistics : Standard Deviation

Ask Mish



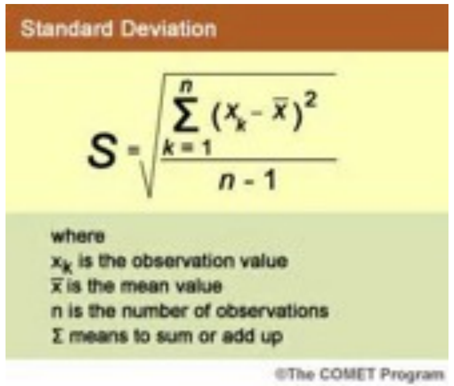
If N= sample size is too big to be measured we take a number n of observations from it and measure them.

Each measurement X is a number +error. Each time, the next measurement contains less error and is closer to the mean. This is called in statistics REGRESSION to the MEAN.



STANDARD DEVIATION (S) = average dispersion around the mean

Finally we obtain a normal distribution where observations are dispersed from min to max around a mean. One way to measure DISPERSION is using a unit called STANDARD DEVIATION (S) which is an average dispersion around the mean.



Standard Deviation

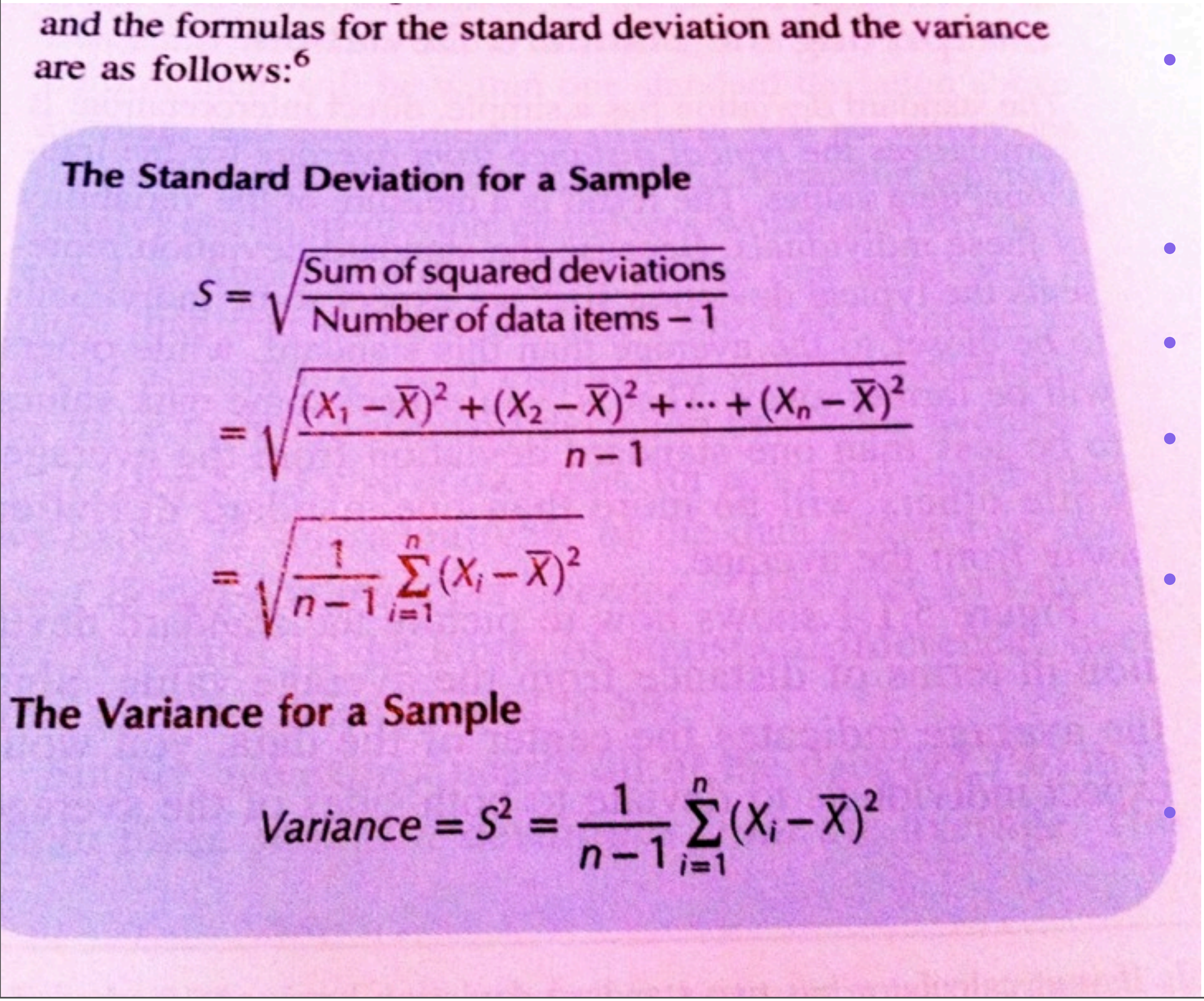
$$S = \sqrt{\frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n-1}}$$

where
x_k is the observation value
$\bar{x}$ is the mean value
n is the number of observations
Σ means to sum or add up

©The COMET Program

- n-1 = degree of freedom (observations- control)

Inferential statistics: standard deviation, variance, range



and the formulas for the standard deviation and the variance are as follows:[6]

The Standard Deviation for a Sample

$$S = \sqrt{\frac{\text{Sum of squared deviations}}{\text{Number of data items} - 1}}$$

$$= \sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}}$$

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

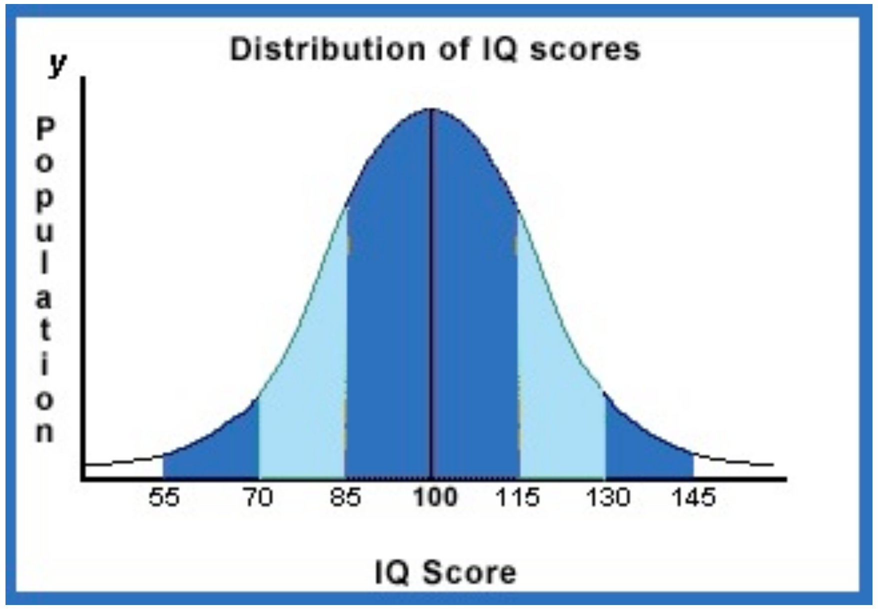
The Variance for a Sample

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

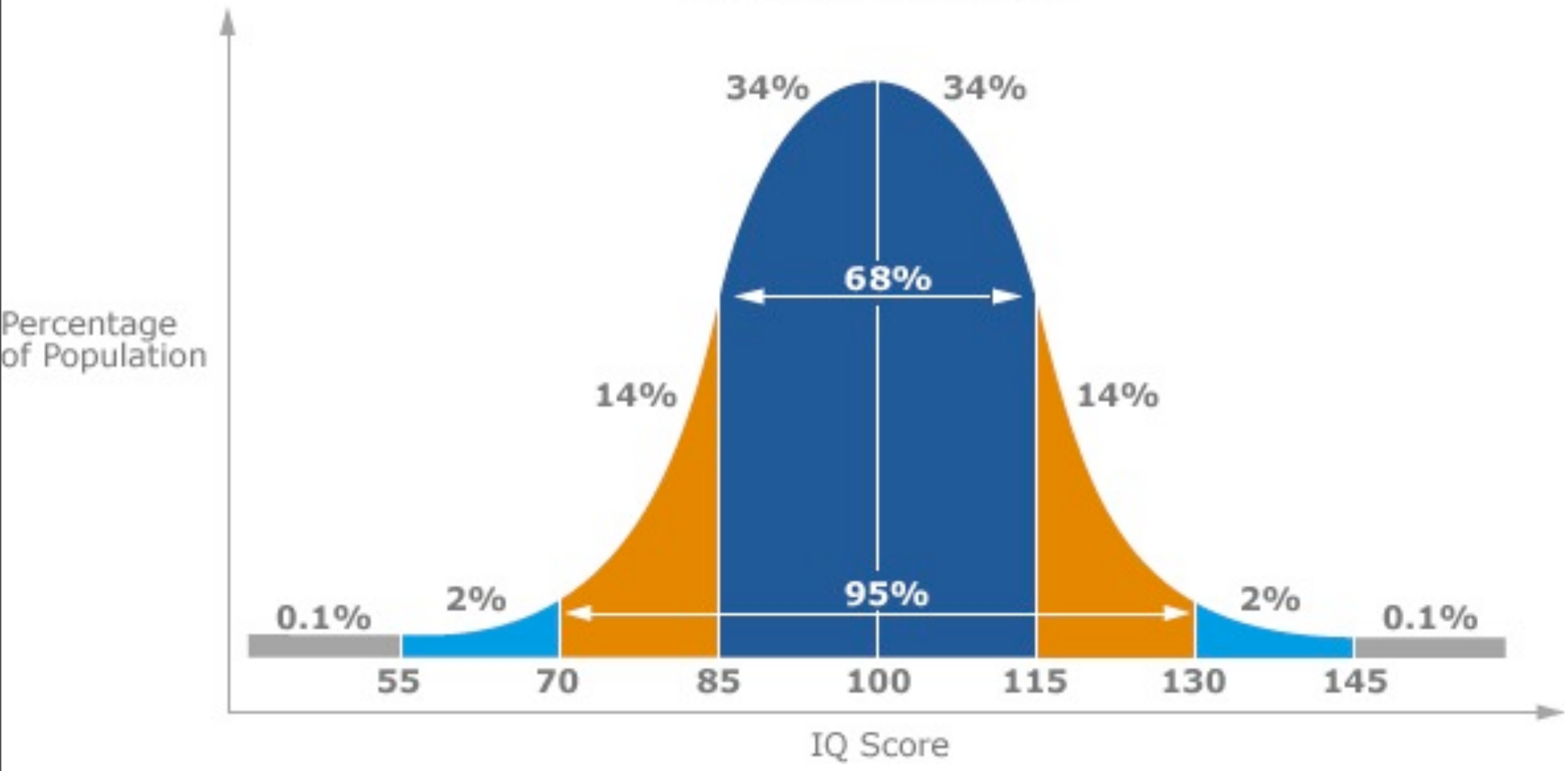
- DISPERSION IN STATISTICS can be measured not only using S, but also variance and range:
- S=standard deviation
- Variance
- Range = max. value - min. value
- in the left, the extended formulas for calculating S and variance for a sample (in case you need)
- Standard deviation is used for calculating confidence intervals

QUIZ:

Ask Mish



IQ Score Distribution



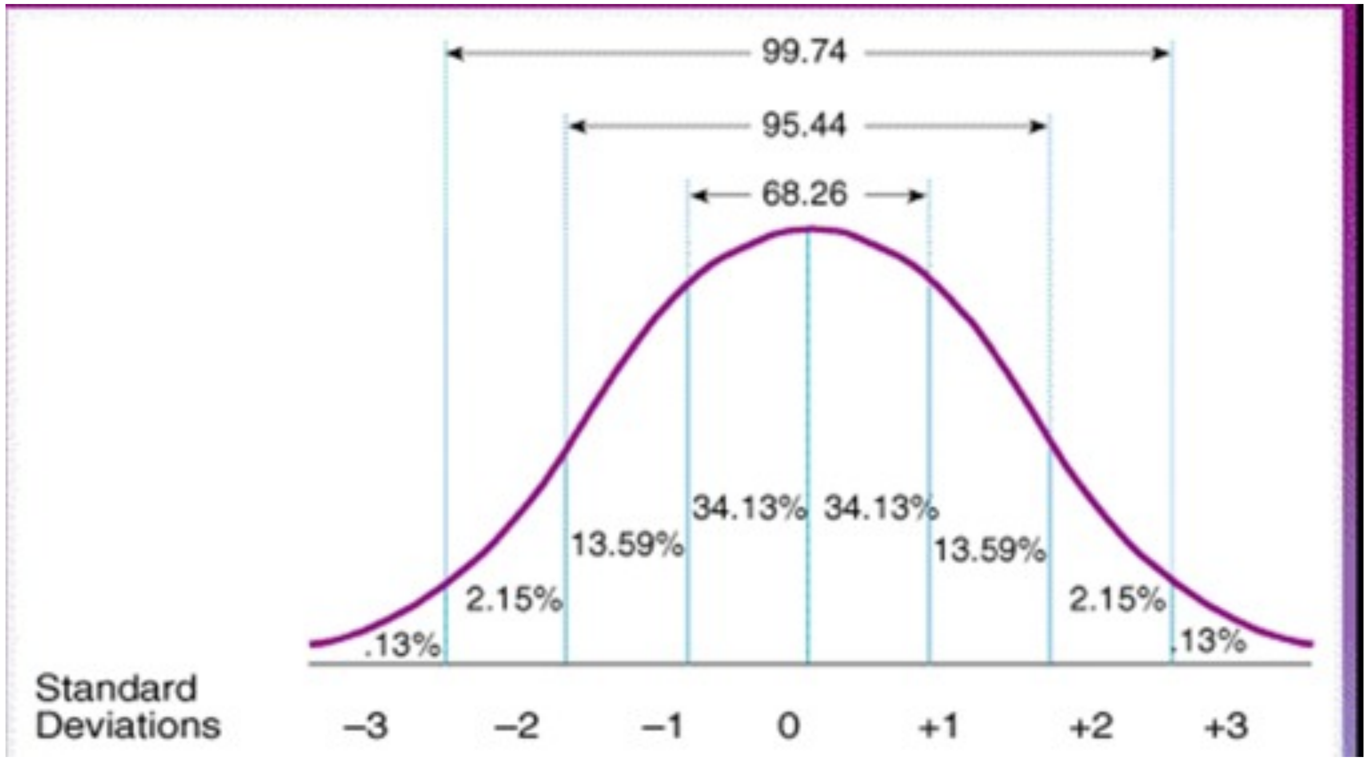
- A standardized IQ test has a mean of 100 and a standard deviation of 15. A person with IQ=115 is at what percentile of IQ?
- A.50th
- B.68th
- C.84th
- D.95th
- E.99th

Ask Mish

answer: C (84th)

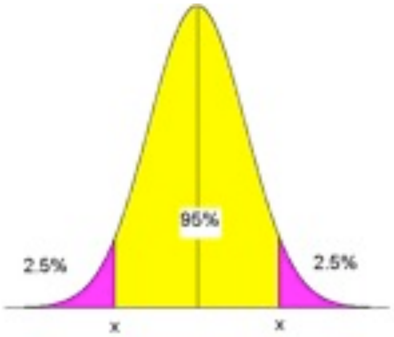
Inferential statistics - Confidence intervals: definition

Ask Mish



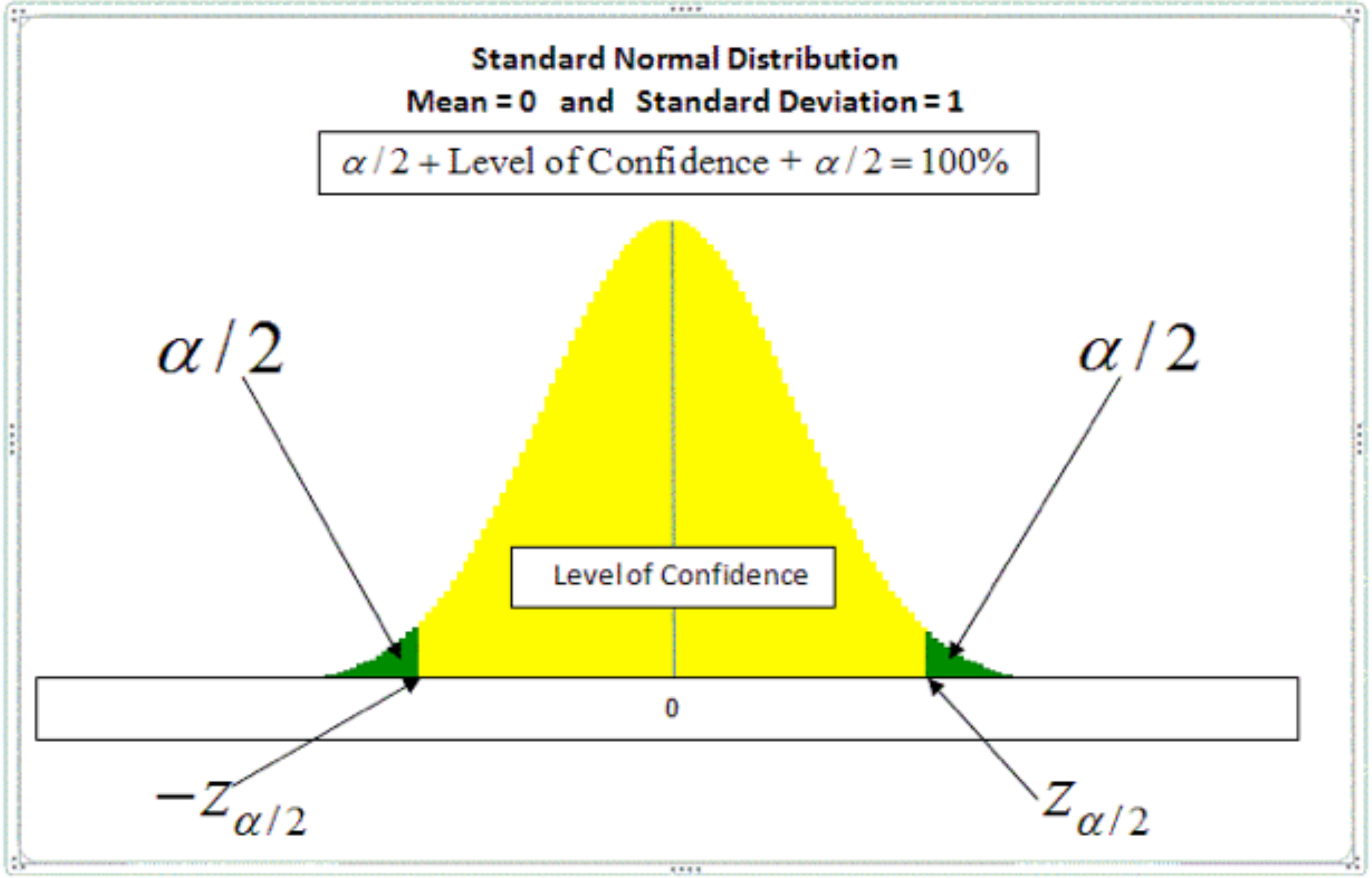
IN a NORMAL DISTRIBUTION approximative:

- 68% of the data are within one SD (-1; +1)
- 95% are within 2 SD (-2;+2)
- 99.7% are within 3 SD (-3;+3)
- 0.3% are beyond 3 SD.

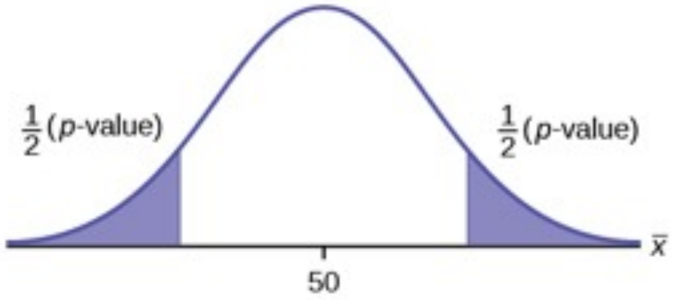
- CONFIDENCE INTERVALS:
- If you have a N=sample size -> you take and measure n outcomes -> you can calculate the mean Xo from these n outcomes. However the result is a distribution of outcomes around this mean. Confidence intervals is about how far from the mean you want to go to feel confident with the result.
- It is generally accepted that a 95% interval around the mean (meaning 2 SD above and 2 SD below the mean) would give a good estimation of the sample. This means that from 100 outcomes , based on the 95% CI formula you will be able to recognize as good 95 outcomes. You will make mistakes in 5 outcomes which you will recognize as good when in reality they are not.
- 

Inferential statistics: 95% confidence interval, p value and type I (alfa) error

Ask Mish


Standard Normal Distribution
Mean = 0 and Standard Deviation = 1
$\alpha/2 + \text{Level of Confidence} + \alpha/2 = 100\%$
$\alpha/2$
$\alpha/2$
Level of Confidence
0
$-Z_{\alpha/2}$
$Z_{\alpha/2}$

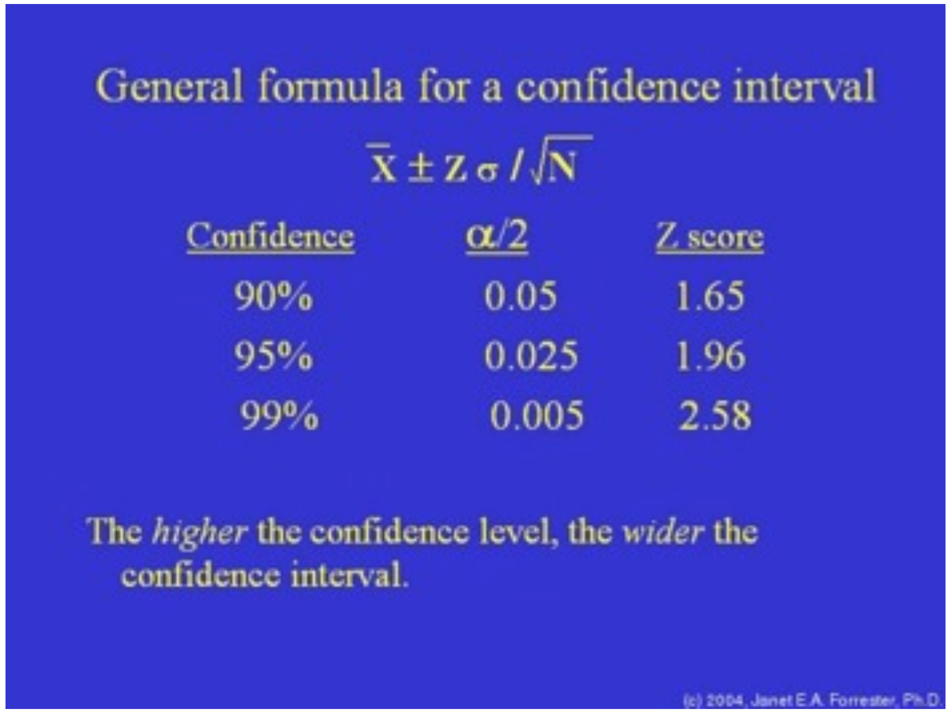
- P VALUE:
- When we chose a 95% confidence interval we accepted that we have a probability p of making a mistake in 5% cases, meaning a p=0.05 also known as p value



- TYPE I aka ALFA ERROR
- It is the error itself. When you recognize a good outcome when in reality is not you commit an error aka TYPE I or ALFA error. In statistics where 95% confidence interval is generally accepted there is a 5%cases that you can make a type I (ALFA) error.
- p=probability</=0.05 to make an error in a study while type I (ALFA) is the error you actually make in 5% of outcomes at 95% C.I.

Inferential statistics: how to calculate confidence intervals

Ask Mish



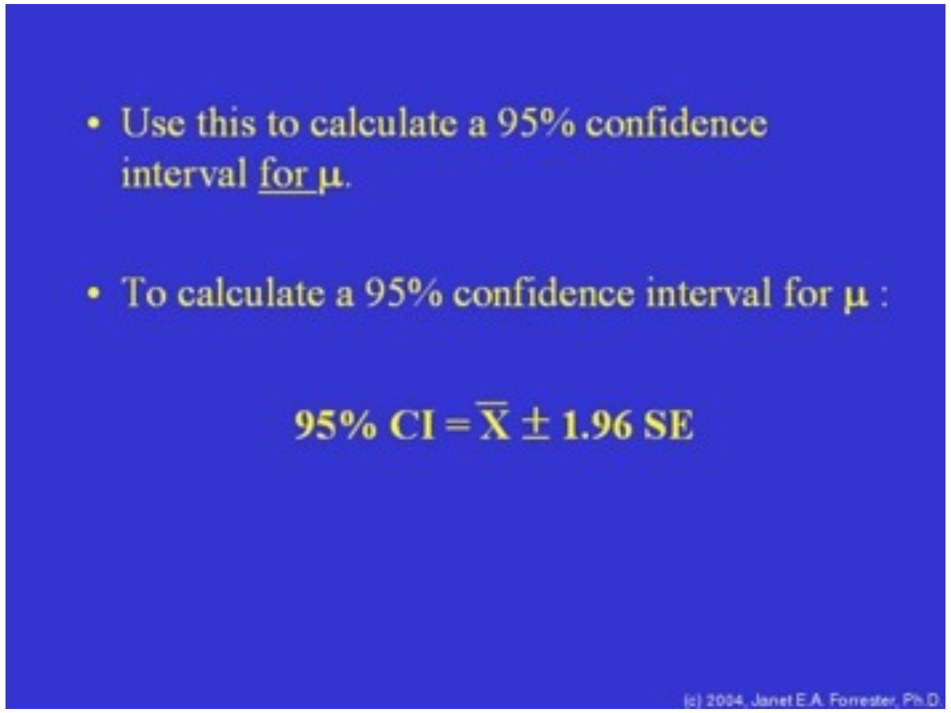
General formula for a confidence interval

$$\bar{X} \pm Z\sigma/\sqrt{N}$$

Confidence	$\alpha/2$	Z score
90%	0.05	1.65
95%	0.025	1.96
99%	0.005	2.58

The *higher* the confidence level, the *wider* the confidence interval.

(c) 2004, Janet E.A. Forrester, Ph.D.



- Use this to calculate a 95% confidence interval for μ.
- To calculate a 95% confidence interval for μ :

95% CI = $\bar{X} \pm 1.96$ SE

(c) 2004, Janet E.A. Forrester, Ph.D.

- CALCULATING CONFIDENCE INTERVAL:

$$c.i. = \bar{x} \pm \underbrace{Z\left(\frac{\sigma}{\sqrt{N}}\right)}$$

This part of the equation is called the margin of error. Your book calls this section E.

- N=sample size, you take n outcomes and calculate the X= average. Margin of error includes: standard error (SE) and Z score.

$$\text{Standard error} = \frac{\sigma_x}{\sqrt{N}}$$

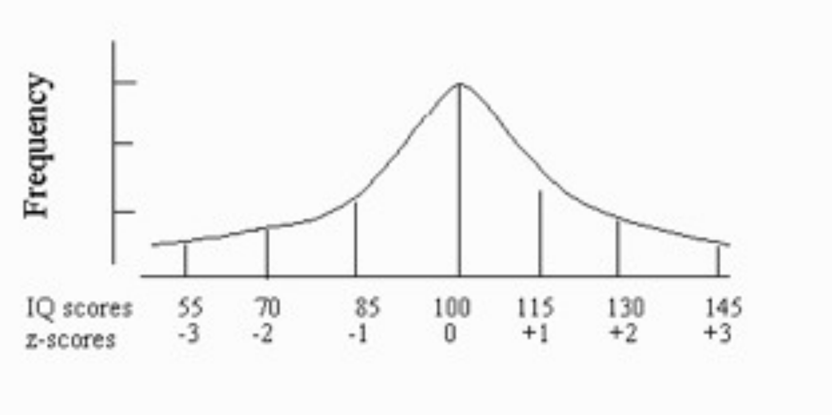
As sample size N goes up you have a better estimation from a larger N. So as N goes up, the error goes down meaning standard error (SE) is less error than standard deviation (sigma) in the formula on the left.

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean
σ = Standard Deviation

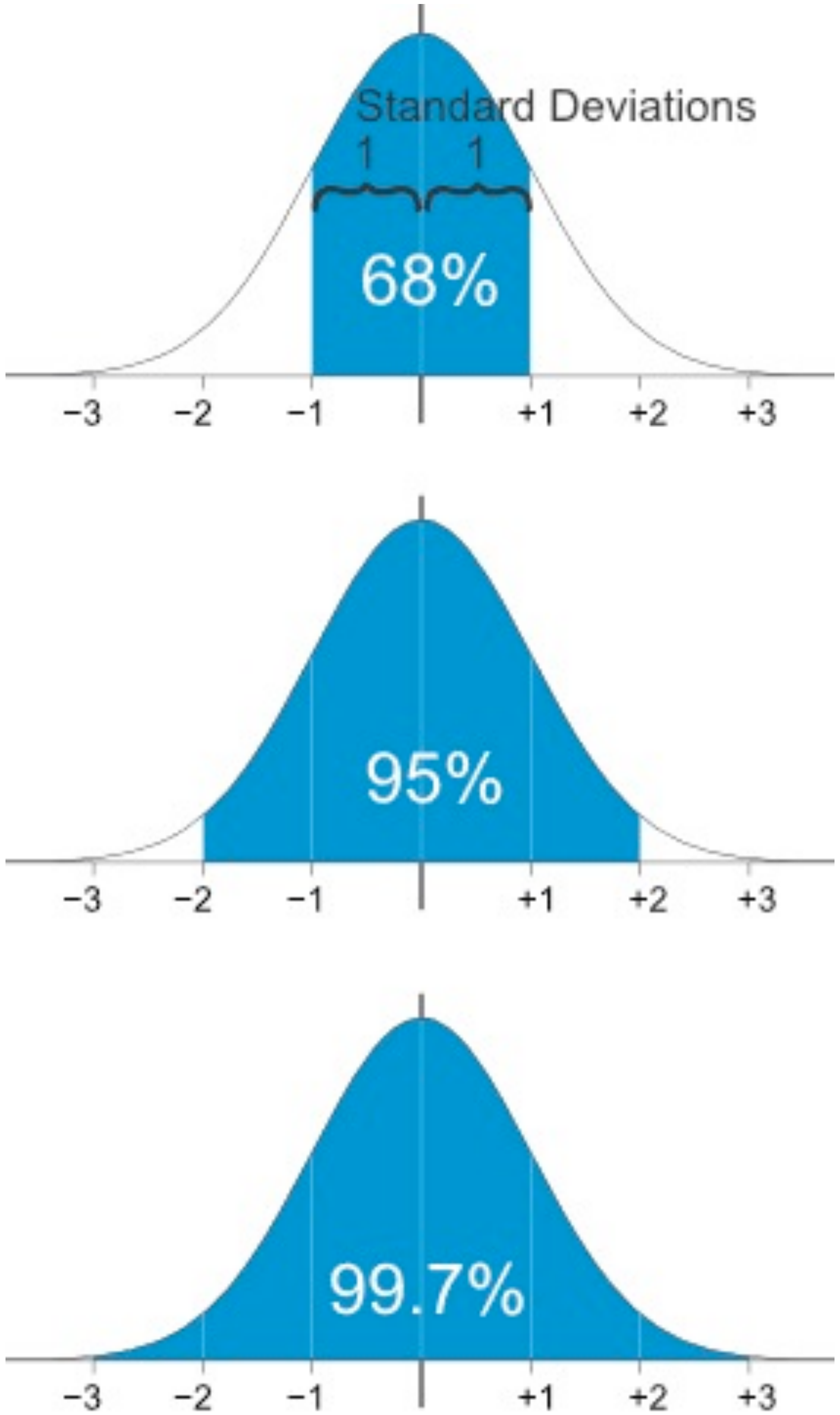
Z score or standard score tells you how far from the mean your C.I. goes and to calculate the Z score use the formula on the left where mean=0 and S= 1. Z is actually how many standard deviations far from the mean goes the C.I. you chose.

For practical purpose,
Z=2 for 95% c.i
Z=2.5 for 99% c.i.



Inferential statistics: Calculating C.I. Quiz

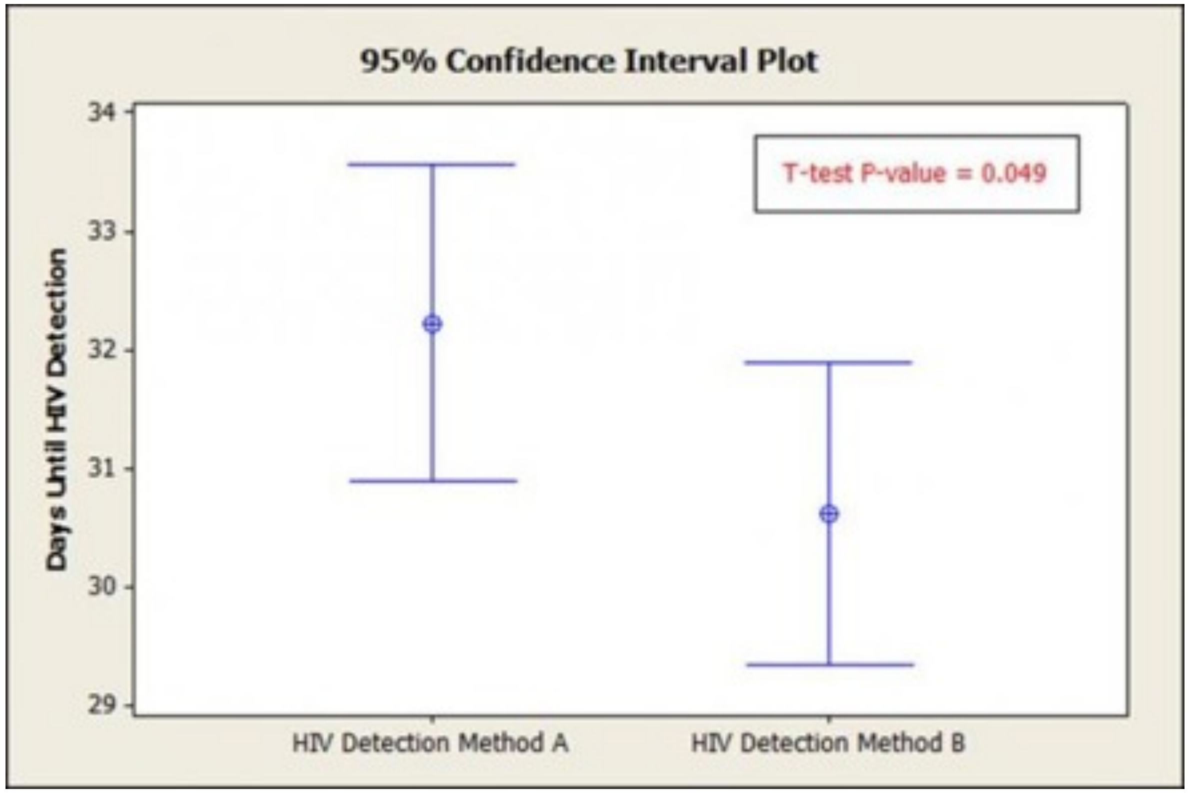
Ask Mish



- Compute a 95% C.I. knowing the following:
- mean Xo= 67
- standard deviation S =8
- sample size N= 16
- consider Z=2
- Answer: 95% CI : between 63-71 including 63 and 71.

95% Confidence Interval plots & 95% confidence intervals for relative risk and odds ratio

Ask Mish



RELATIVE RISK	95% Confidence Interval	INTERPRETATION
1.48	(1.10 - 2.20)	statistically significant
1.69	(0.80 - 2.43)	not stat. significant
0.73	(0.55 - 0.94)	statistically significant

- Q: Assuming the graph in the left presents 95%C.I. are the two HIV detection methods different from each other?
- A: When comparing 2 groups any overlap of C.I. means the groups are not statistically different. Therefore, method A and method B are no different in HIV detection.
- Q: When are the C.I for RR or odds ratio not statistically significant? (see table on left)
- A: If the given C.I contains 1.0 then there is no statistically effect for the exposure, meaning RISK is the SAME. When C.I. contains no 1.0 then there is a statistically significant INCREASED RISK.

Hypothesis testing in statistical studies (1)

Ask Mish

TEST	VARIABLES used in test
Interval/ordinal data	2 interval(I) / 2ordinal
Nominal data	2 nominal (any number of groups)
t test	1N(max. 2 groups) + 1I
ANOVA one way	1N (many groups) + 1I
ANOVA two way	2 N + 1I

STATISTIC FORMULA for each of the tests
Pearson (2I)/Spearman (2O) correlation
chi square
t statistic
F statistic
F statistic

p value

cannot tell if it is BIAS in study
cannot tell if the result is clinically significant
it can only tell you if it is statistically significant

- Now the question is: what's the link between all these we described? I refer to categories, confidence intervals, p value, alfa error, etc?
- The link is this: imagine you want to compare 2 or more categories and draw a conclusion. First you need to DESIGN A STUDY. You need to know what do you want to compare in your study: only nominal data, interval data or nominal and interval data.
- If the categories are not identical you can find either a correlation or a difference between them depending on categories.
- Let's assume you found a difference. The next question : is this difference due to hazard or it is significantly statistic? To know this you will apply for each study a specific STATISTIC FORMULA specially designed for that study. You found a number and you want to know if this number is in your 95% confidence interval, that what you found is statistically significant. You take the number you obtained and check in the tables for the p value related to your number. If the p found<0.05 then YES, your study result shows a significant statistic difference. If p found>0.05, your study result shows no statistically difference.

Hypothesis testing in statistical studies (2)

Ask Mish

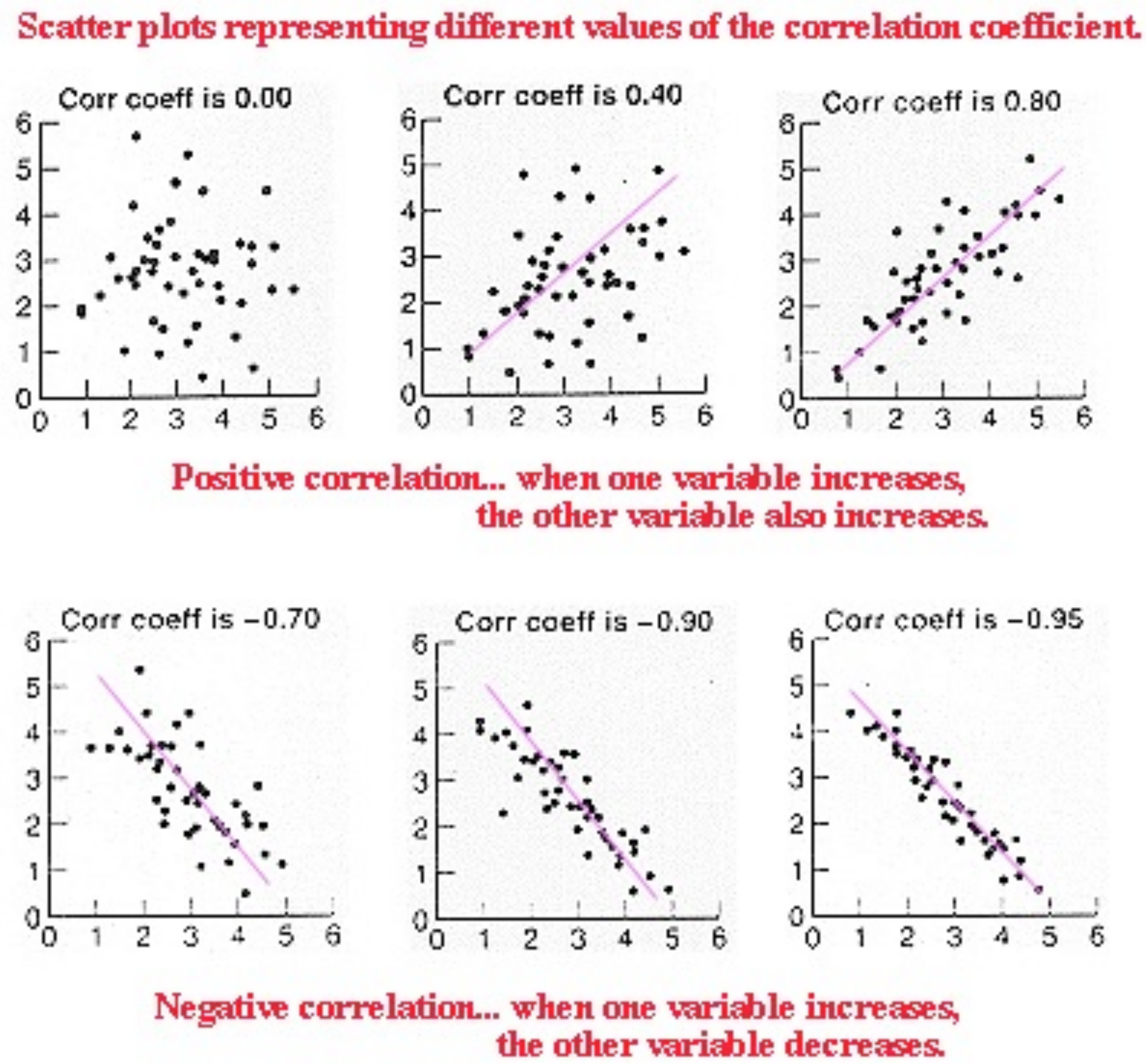
		REALITY	
		DIFFERENCE	NO DIFFERENCE
STUDY RESULT	DIFFERENCE		Type I error α error "false positive"
	NO DIFFERENCE	Type II error or β error "false negative"	

*NULL Hypothesis (Ho) = no difference found
If the study finds a difference : REJECT Ho
If the study finds no difference : FAIL to reject Ho (H1)

- HYPOTHESIS in STATISTICS (2):
- When the study finds a difference when a difference truly exists (box1) and when the study finds no difference when no difference exists (box4) everything is OK. (smiley)
- When the study finds a difference when it truly exists then this is called THE POWER of the study (to see difference)- the first box.
- In TYPE I error or alfa error the study finds a difference when no difference really exists. This is a "false positive" study. It equals p.
- In TYPE II error or beta error the study finds no difference when one truly exists. It's a "false negative" study. Usually:10-20% but no more than 20%.
- POWER= 100 - beta error(%) or 1- beta(decimal). You choose the power when you design the study. If the difference you need to find is small you need an increased power and you need to increase the SAMPLE SIZE which will also increase the costs. You have to find the optimum balance for all. Power > 80%.

Hypothesis testing in statistical studies (3): Correlation Analysis

Ask Mish


Scatter plots representing different values of the correlation coefficient.

Positive correlation... when one variable increases, the other variable also increases.

Negative correlation... when one variable increases, the other variable decreases.

- A CORRELATION:
- means two measures are related not why they are related. Does not mean one variable necessarily causes the other
- CORRELATION COEFFICIENT:
- indicates the DEGREE to which two measures are related. The further from 0 the stronger the relationship. Max. values +1 and -1 indicates a linear relationship. When coefficient = 0 means the two variables have no linear relation to one another (e.g. height and exam scores).
- POSITIVE correlation: the 2 variables go the same direction
- NEGATIVE correlation: the 2 variables go in opposite directions
- TYPES of correlation: PEARSON compares 2 interval level variables and SPEARMAN - 2 ordinal l.variables.
- SCATTER PLOT is a graphical representation of a correlation

Survival Analysis

Ask Mish

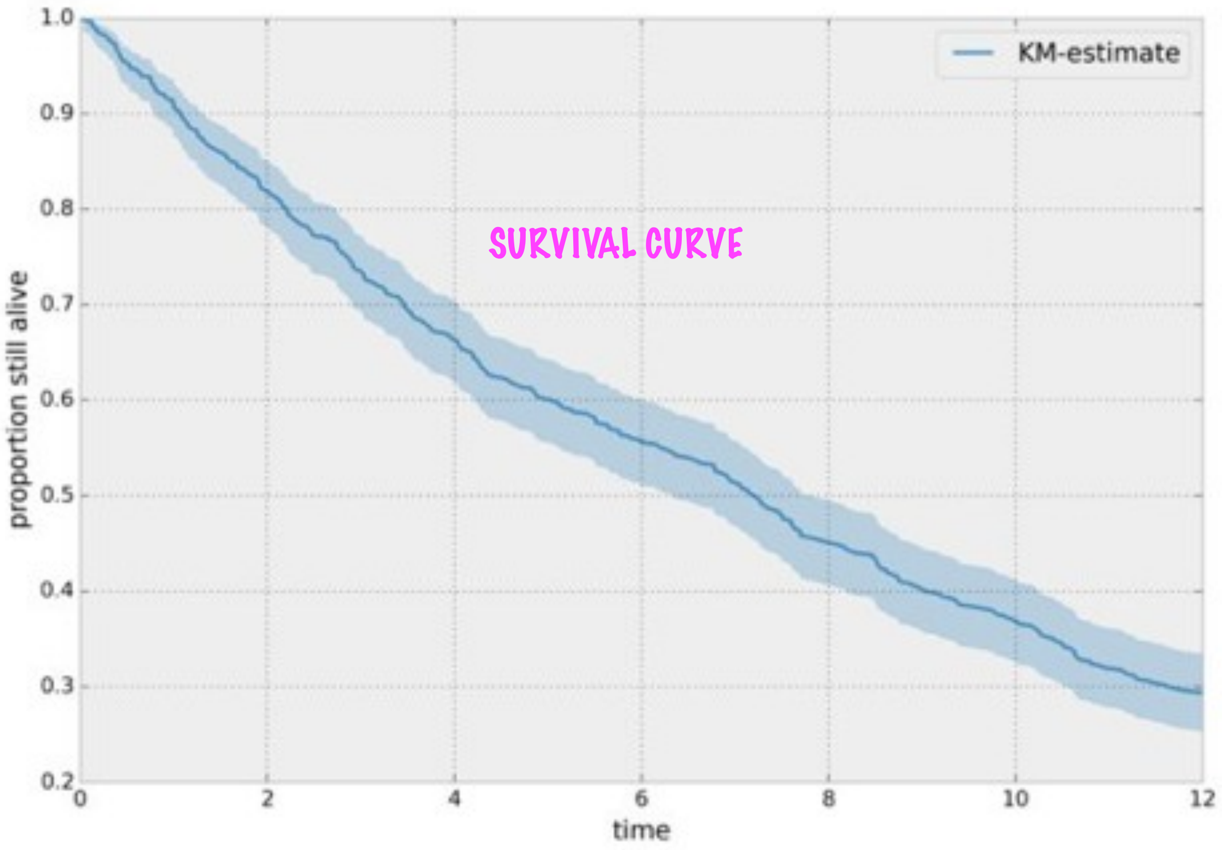


Table : SURVIVAL RATES after SURGERY

Number patients	1 YEAR	2 YEAR	3 YEAR	4 YEAR
172	90%	75%	50%	40%
take 100	90	75	50	40

- SURVIVAL ANALYSIS:
- is a class of statistical procedures for estimating the proportion of people who survive (y axis) in relation to the length survival time
- A survival curve starts with 100% (1.0 in graph) of the study population and shows the percentage of population still surviving at successive times for as long as information is available
- Median survival time is the time where 50%(0.5 in graph) are still alive.
- Median survival time is also called LIFE EXPECTANCY
- Q: What is the life expectancy after surgery?(check the table on the left)
- A: 3 years. (check 50% survival in table)
- Q: If the patient survives 2 years what is the chance for surviving for 3 years?
- A: 50/75= 67%. (At 3y: 50 survived from 75 that survived at 2 years considering 100 patients)