Epidemiology and Biostatistics
Introduction:

• 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:

EPIDEMIOLOGY
studies:
DISTRIBUTIONS
deeterminants
of diseases
in population

BIOSTATISTICS
applies
STATISTICAL
METHODS
in
Biology
Medicine
Public Health

PUBLIC HEALTH
application of Epidemiology and
Biostatistics
to prevent and control disease
in population
why epidemiology & biostatistics? easy way:

Epidemiology
- diseases
  - in population!!!
  - THEORY!!!

Biostatistics
- Mathematics
  - THEORY!!!

Public Health
- application of THEORY above
  - in: PRACTICE!!!
### Why Epidemiology & Biostatistics?

**Comparison:**

<table>
<thead>
<tr>
<th></th>
<th>Epidemiology</th>
<th>Biostatistics</th>
<th>Public Health</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Refer to</strong></td>
<td>Study of diseases in a way that you can:</td>
<td>Application of statistics</td>
<td>Application of theories from Epidemiology &amp; Biostatistics.</td>
</tr>
<tr>
<td><strong>Action</strong></td>
<td>Prevent and control disease (Theory)</td>
<td>To exclude events in medicine that are due by chance alone</td>
<td>To prevent and control disease (Practice)</td>
</tr>
<tr>
<td><strong>On:</strong></td>
<td>Population not one person !!!!</td>
<td>Population not one person !!!!</td>
<td>Population not one person !!!!</td>
</tr>
<tr>
<td>HEALTH</td>
<td>DISEASE:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>complete:</td>
<td>diagnosis using:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>physical</td>
<td>signs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mental</td>
<td>symptoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>social well being</td>
<td>history</td>
<td></td>
<td></td>
</tr>
<tr>
<td>not absence of</td>
<td>test results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>disease</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Signs vs Symptoms: Definitions

<table>
<thead>
<tr>
<th>SIGN</th>
<th>SYMPTOM:</th>
</tr>
</thead>
<tbody>
<tr>
<td>objective evidence</td>
<td>subjective evidence</td>
</tr>
<tr>
<td>of disease</td>
<td>of disease/ a feeling of subject</td>
</tr>
<tr>
<td>can be seen/</td>
<td>others cannot see</td>
</tr>
<tr>
<td>can be measured</td>
<td>others cannot measure</td>
</tr>
<tr>
<td>e.g. vital signs</td>
<td>e.g. headache</td>
</tr>
</tbody>
</table>
Epidemiology: history, distribution of disease and rates

- 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

- 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

- 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

- 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

ENDEMIC:
below or \( \leq \) 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

- **endemic**
  - green color: under the level of expectancy for a disease
  - red color: above the level of expect.

- **epidemic**
  - 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

- **pandemic**
assessing level of disease using rates

- Rates are ratios (numerator/denominator)

- in epidemiology: \( \# \ \text{diseases}/\#\text{population}^* \)

- \( \# = \text{number} \)

- *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 Rates**

**Population 1**
- 100 people
- 4 diseased people

**Crude Rate:** 4/100 (1)

**Specific Rate:** 4/20 & 0/80 (1)

**Adjusted Rate:** 4/20 (1)

**Population 2**
- 100 people
- 1 diseased person

**Crude Rate:** 1/100 (2)

**Specific Rate:** 1/5 & 0/95 (2)

**Adjusted Rate:** 1/5 (2)

Colors:
- 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

**Different**

**Same**
question: What is the rate of AIDS in US?

- 220/100K
- 90/100K
- 500/100K
- 15/100K
- 65/100K
Any disease in the US is <50/100K !!!!!!!!
most important rates in epidemiology

- **Incidence** = rate of occurrence of new cases of disease among total population in a period of time (never a point in time)

  \[ I = \frac{\text{new cases}}{\text{total population}} \times 100 \]

- **Prevalence** = rate of all existing cases of disease among total population in either a point in time or a period in time

  \[ P = \frac{\text{all cases}}{\text{total population}} \times 100 \]

  \[ P = I \times \text{average time duration of disease} \]

  meaning all new cases that are not solved become cases in prevalent pot
**incidence vs prevalence**

<table>
<thead>
<tr>
<th></th>
<th>Incidence</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>refer to</td>
<td>occurrence of new cases (rate)</td>
<td>occurrence of all existing cases (rate)</td>
</tr>
<tr>
<td>among</td>
<td>all population</td>
<td>all population</td>
</tr>
<tr>
<td>time</td>
<td>period of time</td>
<td>period of time or a point in time</td>
</tr>
</tbody>
</table>
Attack rate:

- used instead of incidence
- during a disease outbreak in a narrowly-defined population over a short period of time
- \( \text{AR} = \frac{\text{affected}}{\text{exposed}} \)
- e.g. : AR in case of food poisoning in a restaurant
Cumulative incidence (Proportion incidence)

- 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

<table>
<thead>
<tr>
<th>Incidence decrease</th>
<th>Prevalence decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>effective primary prevention</td>
<td>decreased incidence</td>
</tr>
<tr>
<td>new cases recover quickly in time</td>
<td>increased recovery</td>
</tr>
<tr>
<td>increased death</td>
<td></td>
</tr>
<tr>
<td>increase population</td>
<td>increase population</td>
</tr>
</tbody>
</table>

in black ways of decreasing Incidence vs ways of decreasing Prevalence

1. RECOVERY

2. DEATH !!!!
other rates: vital rates (1)

- Birth rate: number of births @1000 people
  
  \[ \text{Birth rate} = \frac{\text{births}}{\text{population}} \times 1000 \]

- Death rate: number of deaths @1000 people
  
  \[ \text{Death rate} = \frac{\text{deaths}}{\text{population}} \times 1000 \]

- Case fatality rate: number of cases that end up in death;
  \( \text{CFR} = \frac{\text{deaths from a cause}}{\text{diseased}} \times 100 \)

- Proportionate mortality rate (PMR): deaths from a cause/all deaths x 100
other rates: vital rates

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

- **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

**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

<table>
<thead>
<tr>
<th>CRUDE</th>
<th>SPECIFIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>birth rate</td>
<td>fertility rate</td>
</tr>
<tr>
<td>death rate</td>
<td>infant mortality rate</td>
</tr>
<tr>
<td>PREVALENCE</td>
<td>PMR</td>
</tr>
<tr>
<td>incidence</td>
<td>CFR</td>
</tr>
</tbody>
</table>
### Vital rates in epidemiology:

<table>
<thead>
<tr>
<th></th>
<th>BIRTHS</th>
<th>DEATHS</th>
<th>DISEASES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total population</td>
<td>Birth rate</td>
<td>Death rate</td>
<td>Incidence (new)</td>
</tr>
<tr>
<td>/1000</td>
<td></td>
<td></td>
<td>Prevalence (all)</td>
</tr>
<tr>
<td>Groups of</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>population:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>infants/mothers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/100,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groups of</td>
<td></td>
<td>IMR</td>
<td>CFR</td>
</tr>
<tr>
<td>population:</td>
<td></td>
<td>MMR</td>
<td>PMR</td>
</tr>
<tr>
<td>diseased/dead</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/100</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Epidemiology: Risk, Risk Factors and Causes

- 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

<table>
<thead>
<tr>
<th></th>
<th>Incidence</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>refer to</td>
<td>occurrence of new cases (rate)</td>
<td>probability of occurrence of new cases (rate)</td>
</tr>
<tr>
<td>among</td>
<td>all population</td>
<td>all population</td>
</tr>
<tr>
<td>time</td>
<td>period of time</td>
<td>period of time</td>
</tr>
</tbody>
</table>
# Epidemiology: determinants

## DETERMINANTS: CAUSES & RISK FACTORS

<table>
<thead>
<tr>
<th></th>
<th>Causes</th>
<th>Risk Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>refer to</td>
<td>personal habits &amp; environmental factors</td>
<td>personal habits &amp; environmental factors</td>
</tr>
<tr>
<td>action</td>
<td>DETERMINES</td>
<td>INCREASE the probability of</td>
</tr>
<tr>
<td>on:</td>
<td>the occurrence of disease</td>
<td>the occurrence of disease</td>
</tr>
</tbody>
</table>
1. Knowing causes and risk
   - used to prevent and control disease by removing causes and risk

2. Not knowing causes and risk
   - an epidemiological study is recommended to determine them
Epidemiology: prevention of disease

• 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

## Levels of Prevention
- **Primary**
- **Secondary**
- **Tertiary**

## Ways of Prevention
- **Primary**: remove risk factors
- **Secondary**: early detection & treatment
- **Tertiary**: reduce complications

## Examples of Prevention
- **Primary**: vaccines, folate, exercise, seat belts
- **Secondary**: screening tests
- **Tertiary**: Beta-blockers post MI
1. Prevent and control disease

- 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

<table>
<thead>
<tr>
<th></th>
<th><strong>DISEASED people</strong></th>
<th><strong>HEALTHY people</strong></th>
<th><strong>positive predictive value</strong></th>
<th><strong>negative predictive value</strong></th>
<th><strong>accuracy</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TEST POSITIVE</strong></td>
<td>true positive</td>
<td>false positive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TEST NEGATIVE</strong></td>
<td>false negative</td>
<td>true negative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sensitivity</td>
<td>specificity</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
screening tests concepts

- **Sensitivity** = percentage of people with disease detected by test
- **Specificity** = percentage of healthy people 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 = $\frac{tp+tn}{all\ tested} (tp+tn+fp+fn)$; chance = percent
## Screening Test Example

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Disease: 100</th>
<th>Healthy: 100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Positive</strong></td>
<td>80 true positive</td>
<td>10 false positive</td>
</tr>
<tr>
<td><strong>Test Negative</strong></td>
<td>20 false negative</td>
<td>90 true negative</td>
</tr>
</tbody>
</table>

### Performance Metrics
- **Sensitivity**: 80/100
- **Specificity**: 90/100
- **Accuracy**: 80 + 90/200
PPV vs Sensitivity vs Specificity

- 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!
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

- 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:

- 1. observational
- 2. experimental
observational vs experimental studies

- **OBSERVATIONAL** = non interventional studies
- **EXPERIMENTAL** = interventional studies
1. Observational studies:

- 1. case report
- 2. case series
- 3. cross-sectional
- 4. case-control
- 5. cohort
2. Experimental studies:

- RCT = random control trials
What does each study mean:

1. **CASE REPORT or CASE SERIES**
   - REPORT = report of one case or a small number 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 factors

4. **COHORT**
   - one risk factor 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)
**Design of an observational study**

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>No Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exposed</strong></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td><strong>Non-exposed</strong></td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

*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

- 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 with 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

<table>
<thead>
<tr>
<th></th>
<th>DISEASE</th>
<th>NO DISEASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXPOSED</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>NON EXPOSED</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>
graphical representation of studies

strong evidence

weak evidence

RCT

cohort

case control

cross sectional

case report, case series

= analytical

(hypothesis tested)

= descriptive

(hypothesis formulated)
Hypothesis testing in observational studies(1): concepts

**Case Control**
- Risk factors (Exposed/Non exposed)
- One disease
- Odds ratio (OR)

**Cross sectional**
- Risk factors
- One disease
- Chi square

**Cohort**
- One risk factor
- Many diseases
- Relative risk, attributable risk (RR, AR)

**Relative risk reduction or increase of the exposed**

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>No Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Non Exposed</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>
Hypothesis testing in observational studies (2) formulas

**Case Control**

- **Odds Ratio (OR)**
  
  \[
  OR = \frac{a}{b} / \frac{c}{d} = \frac{ad}{bc}
  \]

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

- **Chi-square**

  \[
  \chi^2 = \frac{(|a-d| - |b-c|)^2}{(a+b)(c+d)(a+c)(b+d)}
  \]

- **Relative Risk (RR)**
  
  \[
  RR = \frac{a / (a + c)}{b / (b + d)} = \frac{a}{c} (DIVISION)
  \]

  Question for RR: how much more likely?

- **Attributable Risk (AR)**
  
  \[
  AR = \frac{a}{a + c} - \frac{b}{b + d}
  \]

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

**Cohort**

- **Relative Risk Reduction (RRR)**
  
  \[
  RRR = |1 - \frac{RR}{RR}|
  \]

- **Number Needed to Treat (NNT)**
  
  \[
  NNT = \frac{1}{ARR}
  \]

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

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)

#### 1. NNT/NNH

<table>
<thead>
<tr>
<th></th>
<th>NNT</th>
<th>NNH</th>
</tr>
</thead>
<tbody>
<tr>
<td>refer to</td>
<td>treatment as cure</td>
<td>treatment w/ side effects</td>
</tr>
<tr>
<td># people treated</td>
<td>to prevent 1 case(disease )</td>
<td>to prevent 1 case(disease )</td>
</tr>
</tbody>
</table>

#### 2. Total: 100 people

<table>
<thead>
<tr>
<th></th>
<th>disease</th>
<th>no disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>exposed: 50</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>non exposed: 50</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

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 \( x = \text{NNT} \) to prevent 1 case of disease 
   apply 3 simple rule: \( \text{NNT} = \frac{1\times 100}{5} = 20 \) 
   meaning you need to treat 20 people in order to prevent 1 case of disease  
   also calculate : \( \text{NNT/NNH} = \frac{1}{\text{ARR}} \)
Clinical Trials

- **Intervention studies**: Clinical Trials

- *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

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

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

- 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)
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
A nominal data can be measured using one of the three above scales: rank order, interval or ratio.
### Descriptive vs Inferential Statistics

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

- **Descriptive Statistics**
  - Measures groups/population (because you can measure each member of the group).
  - Result: Distribution is a bell shape curve symmetric to a central point (mean = median = mode).

- **Inferential Statistics**
  - Takes a sample from a group and draws a conclusion about the whole group (because you cannot measure all).
  - Result is expressed in confidence intervals.
Descriptive statistics: normal distribution

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

Mean = Median = Mode
Descriptive statistics: normal distribution example

Gaussian or Normal Distribution

Figure 1 Normal distribution of heart-rate measurements.
Descriptive statistics: various distributions

- **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

- 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.

- 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.

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

where
- \( x_k \) is the observation value
- \( \bar{x} \) is the mean value
- \( n \) is the number of observations
- \( \Sigma \) means to sum or add up

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

\( S \) = standard deviation

\( \bar{x} \) = mean

\( n \) = number of observations

\( \Sigma \) = sum
Inferential statistics: standard deviation, variance, and the formulas for the standard deviation and the variance are as follows:

**The Standard Deviation for a Sample**

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

\[ = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n - 1}} \]

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

**The Variance for a Sample**

\[ \text{Variance} = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \]

**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
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
answer: C (84th)
Inferential statistics - Confidence intervals: definition

CONFIDENCE INTERVALS:

If you have a N=sample size -> you take and measure n outcomes -> you can calculate the mean \( \bar{X} \) 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.

IN a NORMAL DISTRIBUTION

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

- **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.

- $p=probability\leq0.05$ 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

General formula for a confidence interval

\[ \bar{x} \pm Z \frac{\sigma}{\sqrt{N}} \]

<table>
<thead>
<tr>
<th>Confidence</th>
<th>( \alpha/2 )</th>
<th>Z score</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.05</td>
<td>1.65</td>
</tr>
<tr>
<td>95%</td>
<td>0.025</td>
<td>1.96</td>
</tr>
<tr>
<td>99%</td>
<td>0.005</td>
<td>2.58</td>
</tr>
</tbody>
</table>

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

- Use this to calculate a 95% confidence interval for \( \mu \).
- To calculate a 95% confidence interval for \( \mu \):  
  
  \[ 95\% \text{ CI} = \bar{x} \pm 1.96 \text{ SE} \]

**Calculating confidence interval:**

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

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 \( \bar{x} \) = average. Margin of error includes: standard error (SE) and Z score.

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.

- \( Z \) 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.

\[ z = \frac{\bar{x} - \mu}{\sigma} \]

\( \mu \) = Mean  
\( \sigma \) = Standard Deviation

For practical purpose,  
\( Z=2 \) for 95% c.i  
\( Z=2.5 \) for 99% c.i.
Compute a 95% C.I. knowing the following:

- mean $X_0 = 67$
- standard deviation $S = 8$
- sample size $N = 16$
- consider $Z = 2$

Answer: 95% CI: between 63-71 including 63 and 71.
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.

<table>
<thead>
<tr>
<th>RELATIVE RISK</th>
<th>95% Confidence Interval</th>
<th>INTERPRETATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.48</td>
<td>(1.10 - 2.20)</td>
<td>statistically significant</td>
</tr>
<tr>
<td>1.69</td>
<td>(0.80 - 2.43)</td>
<td>not stat. significant</td>
</tr>
<tr>
<td>0.73</td>
<td>(0.55 - 0.94)</td>
<td>statistically significant</td>
</tr>
</tbody>
</table>
Hypothesis testing in statistical studies (1)

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.

<table>
<thead>
<tr>
<th>TEST</th>
<th>VARIABLES used in test</th>
<th>STATISTIC FORMULA for each of the tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interval/ordinal data</td>
<td>2 interval(I) /2ordinal</td>
<td>Pearson (2I)/Spearman(2O) correlation</td>
</tr>
<tr>
<td>Nominal data</td>
<td>2 nominal (any number of groups)</td>
<td>chi square</td>
</tr>
<tr>
<td>t test</td>
<td>1N(max. 2 groups) + 1I</td>
<td>t statistic</td>
</tr>
<tr>
<td>ANOVA one way</td>
<td>1N (many groups) + 1I</td>
<td>F statistic</td>
</tr>
<tr>
<td>ANOVA two way</td>
<td>2 N + 1I</td>
<td>F statistic</td>
</tr>
</tbody>
</table>

- HYPOTHESIS in STATISTICS (1)
- 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)

<table>
<thead>
<tr>
<th>STUDY RESULT</th>
<th>DIFFERENCE</th>
<th>NO DIFFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIFFERENCE</td>
<td>POWER</td>
<td>Type I error</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha$ error</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&quot;false positive&quot;</td>
</tr>
<tr>
<td>NO DIFFERENCE</td>
<td>Type II error or $\beta$ error</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&quot;false negative&quot;</td>
</tr>
</tbody>
</table>

*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%.
- $\text{POWER} = 100 - \beta \text{ error(\%)}$ or $1-\beta \text{ (decimal)}$. 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

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.

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.
Survival Analysis

Survival Analysis:
- is a class of statistical procedures for estimating the proportion of people who survive (y axis) in relation to the length of 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 = 66.67%. (At 3y: 50 survived from 75 that survived at 2 years considering 100 patients)

<table>
<thead>
<tr>
<th>Number patients</th>
<th>1 YEAR</th>
<th>2 YEAR</th>
<th>3 YEAR</th>
<th>4 YEAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>172</td>
<td>90%</td>
<td>75%</td>
<td>50%</td>
<td>40%</td>
</tr>
<tr>
<td>take 100</td>
<td>90</td>
<td>75</td>
<td>50</td>
<td>40</td>
</tr>
</tbody>
</table>