# Principles of Biostatistics and Informatics

3rd Lecture: Elements of Probability Calculus

26th September 2018

Dániel VERES

1

In this lecture we are discussing on probability.
First we „define" the probability as a quantity based on the law of large numbers.
Thereafter we discuss on the probability of events (notation, and/or relation between events, mutually exclusive and independent events, Kolmogorov axioms and conditional probability). After that we discuss other probability terms as risk and odds.
After that we discuss on how to estimate/calculate probabilities using probability calculus.
At the end I give you two example for how our mind works and how it have to be....

# An Experiment...

We have a quick test for a disease:
blue: healthy
green: ill
We want to figure out whether there is an epidemic in a certain area based on the proportion of ill people. What we know is:
- In non-affected („healthy") areas:
1-2 are green out of 10 people
- In affected areas:
7-9 are green out of 10 people

Is there an epidemic in the unkown area in question?
(??Actions hard consequenies...)

In this experiment we are modelling an epidemic investigation. We have a quick test for a disease. Blue colored result indicate healthy, green one indicate ill people.
We want to figure out whether there is an epidemic in a certain area based on the proportion of ill people.
What we know is:

in a non-affected (blue) area there are 1-2 ill out of 10 and
in the affected area there is 7-9 ill out of 10 people.

Now we are to test the people in an unknown area where we would like to know whether there is a disease – and may be an epidemic.

We have limited time and resources so we couldn't test every people, but we have to make a decision.

# An Experiment...

We have a quick test for a disease:

blue: healthy

green: ill

We want to figure out whether there is an epidemic in a certain area based on the proportion of ill people. What we know is:

- In non-affected („healthy") areas:

  1-2 are green out of 10 people

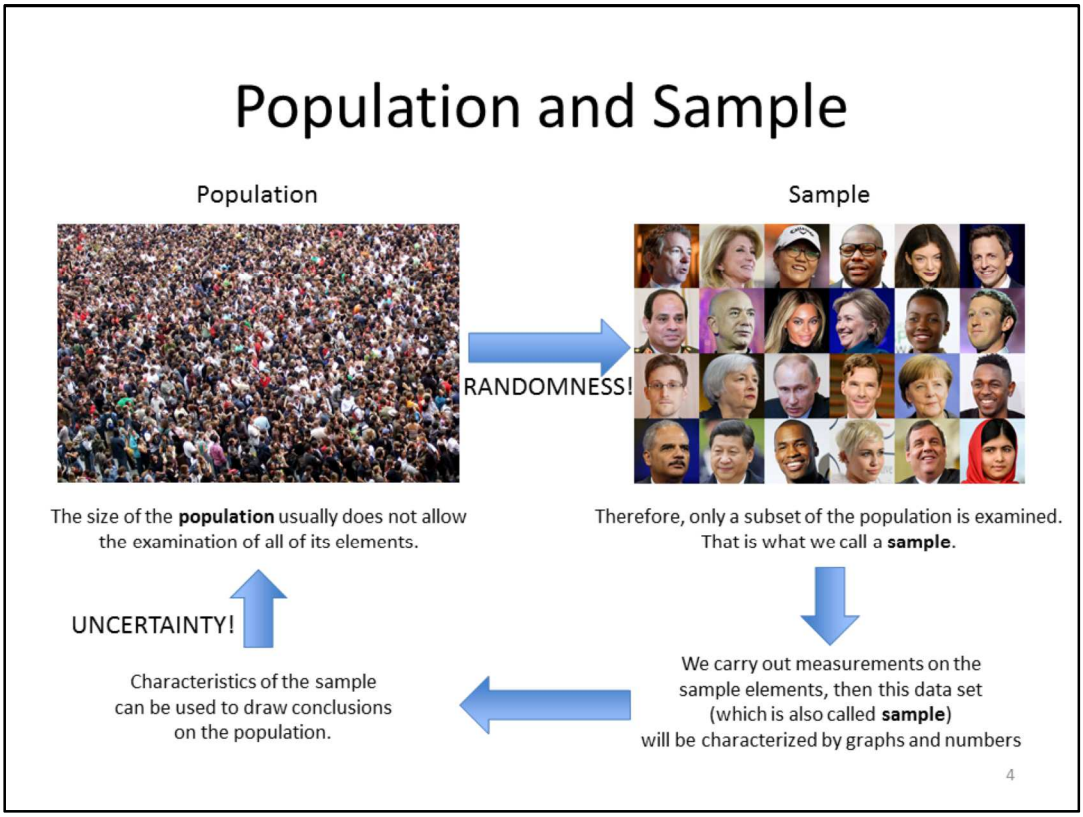
- In affected areas:

  7-9 are green out of 10 people

Is there an epidemic in the unkown area in question?

***Increasing the number of measurements increase the „certainty".***
***How many measurements are required?***
***But a small uncertainty still remain... – How much is that?***

We have limited time and resources so we couldn't test every people, but we have to make a decision.
We found *that increasing the number of measurements increase the „certainty".* But there is always an *uncertainty* if we don't measure everybody. How much is that uncertainty?
But there is always an *uncertainty* if we don't measure everybody. How much is that uncertainty?




To answer the question about the number of measurements are required let us clarify the meaning of population and sample.

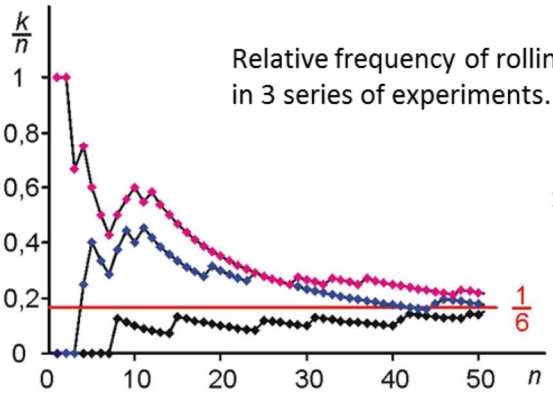
As we mentioned before, statistics examines random mass phenomena. This means that during examination of a phenomenon many, if not infinitely many measurements would be possible. The set containing the outcomes of all these theoretically possible measurements is called **population**. Theoretically, the complete understanding of a variable would require the execution of all the possible measurement, but of course it is not possible.

Consequently, we only observe a subset of the population, which is called **sample**. The most evident way of generating this subset is **random selection**.

We carry out measurements on the sample, the set of measurement results is also called **sample**. (That is: in less precise way the sample may be a group of students [individuals, objects] of the university as a population. In more precise way, the population is the height of all people at the university, the sample is the set of height values for a group that was actually measured.) The sample might be characterized graphically or numerically as we learned in the last lecture, then the properties learned that way may be extrapolated to the population. E.g. if 25% of people in a group have blood type “A”, we may expect the same from the whole population. Since the sample is chosen randomly, it will not necessary represent the population, the frequency of occurrence of different values within the population perfectly. As a result, every conclusion drawn from a sample carries a burden of **uncertainty**.

We carry out measurements on the sample, the set of measurement results is also called **sample**. (That is: in less precise way the sample may be a group of students [individuals, objects] of the university as a population. In more precise way, the population is the height of all people at the university, the sample is the set of height values for a group that was actually measured.)
The sample might be characterized graphically or numerically as we learned in the last lecture, then the properties learned that way may be extrapolated to the population. E.g. if 25% of people in a group have blood type “A”, we may expect the same from the whole population. Since the sample is chosen randomly, it will not necessary represent the population, the frequency of occurrence of different values within the population perfectly. As a result, every conclusion drawn from a sample carries a burden of **uncertainty**.
What is the quantity of the uncertainty? How to define?

# An Other Experiment...




Relative frequency of rolling 6 with a die in 3 series of experiments.

$$\text{rel. freq.} = \frac{N_{favorable}}{N_{all}}$$

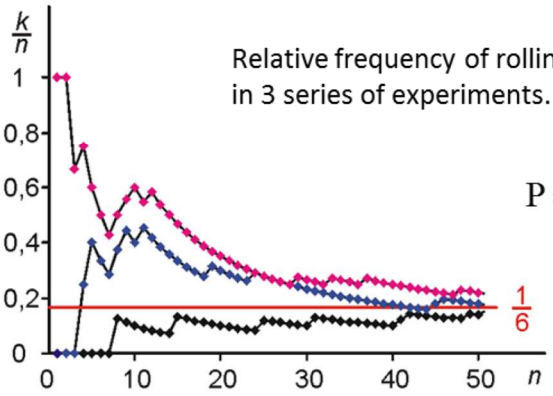
We experience that **relative frequencies** – although with fluctuations – **tend to a certain value** independetly from the actual series of experiments if we **increase the number of the experiments**.

In an other experiment we roll a die 50 times and we count the relative frequency of rolling 6. We repeat this experiment 3 times.

We experience that **relative frequencies** – although with fluctuations – **tend to a certain value** independently from the actual series of experiments if we **increase the number of the experiments**.

In an other experiment we roll a die 50 times and we count the relative frequency of rolling 6. We repeat this experiment 3 times. We experience that the *relative frequencies* (frequency of favorable/all measurement) *tend to a certain value independently from the actual series of experiments if we increase the number of the rolls.*

# Probability as a Quantity




Relative frequency of rolling 6 with a die in 3 series of experiments.

$$\mathrm{P} = \text{rel. freq.} = \frac{N_{favorable}}{N_{all}}$$

$$N_{all} \to \infty$$

**Law of large numbers** (on relative frequencies): the relative frequency in an infinite sequence tends to a certain value.
We assign that **certain value** to an **event**: ***1/6*** to ***rolling 6*** with a die.
This value is called the ***probability of an event***.
This is an *empirical law – cannot be proven* by logical sequence.

*Law of large numbers (on relative frequencies): the relative frequency in an infinite sequence tends to a certain value.*
We assign that certain value to an event: 1/6 to rolling 6 with a die.
This value is called *probability of an event*. (The probability of an outcome of a variable in a given situation.)
The relative frequency is equal to the probability if the sequence is infinite.
This is an empirical law – can not be proven by logical sequence.

# Probability of Events I.

**Notation:**
Event: **A**
*(the patient has fever)*
Probability that event A occurs**: P(A)**
*(the probability that the patient has fever)*

Complementary (complement) event: **$\bar{A}$**
*(the patient has NO fever)*
Probability that event A NOT occurs**: P($\bar{A}$) or P(notA)**
*(the probability that the patient has NO fever)*

Now we describe some properties of events' probability. First about the notation. (Examples are given in parentheses with italian format)
An event is notated with a capital letter – e.g. A (*the patient has fever*). Its probability symbolized as P(A). Example P(A) = the probability that the patient has fever.
We called complementary or complement event if we negate a given event: the event NOT occur. It is notated as $\bar{A}$. e.g. $\bar{A}$ (*the patient has no fever*). Its probability symbolized as P($\bar{A}$). Example P($\bar{A}$) = the probability that the patient has NO fever.

# Probability of Events I.

**Notation:**

Event: **A**

*(the patient has fever)*

Probability that event A occurs**: P(A)**

*(the probability that the patient has fever)*

Probability that event A **or** event B occur: **P(AorB)**, **P(A+B)**, **P(AUB)**

*(the probability that the patient has fever or headache)*

Probability that event A *or* event B occur (*the probability that the patient has a fever or a headache*) could be notated in three way: P(AorB), P(A+B), P(AUB). The last one refer to a set definition: the *Union* – abbreviated U – is all value that belongs to at least one of the sets. On Venn-diagram we can plot it as it shown in the slide.

# Probability of Events I.

**Notation:**

Event: **A**

*(the patient has fever)*

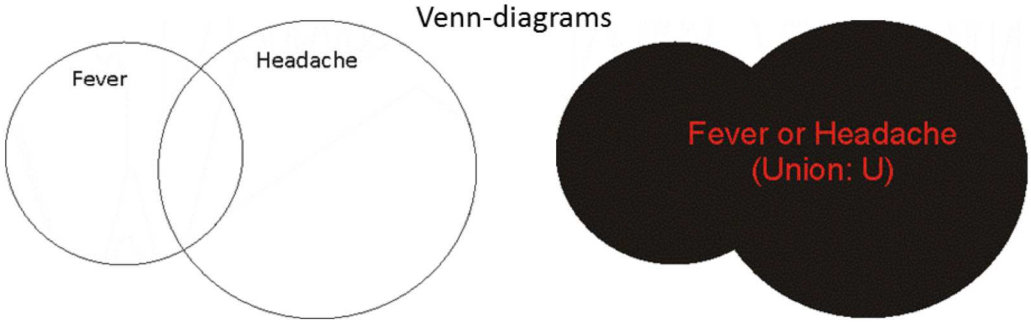
Probability that event A occurs**: P(A)**

*(the probability that the patient has fever)*

Probability that event A **or** event B occur: **P(AorB)**, **P(A+B)**, **P(AUB)**

*(the probability that the patient has fever or headache)*

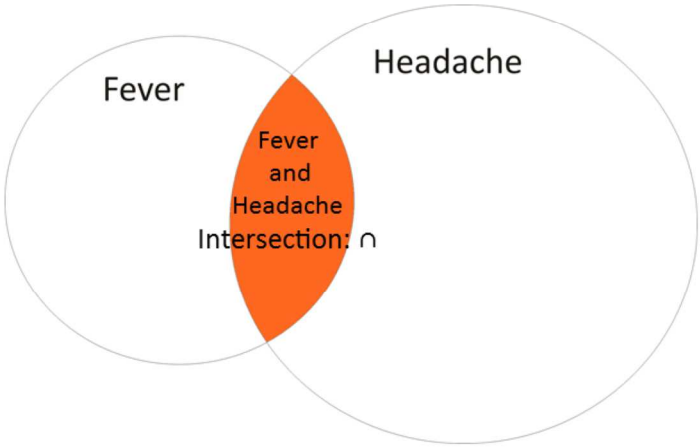
Venn-diagrams




Probability that event A *or* event B occur (*the probability that the patient has a fever or a headache*) could be notated in three way: P(AorB), P(A+B), P(AUB). The last one refer to a set definition: the *Union* – abbreviated U – is all value that belongs to at least one of the sets. On Venn-diagram we can plot it as it shown in the slide.

# Probability of Events II.

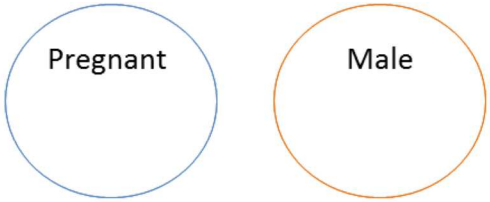
Prob. that both events A **and** B occur: **P(AandB)**, **P(A*B)**, **P(AB)**, **P(A∩B)**
*(the probability that the patient has both fever and headache)*




The probability that both events A and B occur could be notated as P(AandB), P(A*B), P(AB), P(A∩B). (*The probability that the patient has fever and headache.*) The ∩ is the symbol in set theory for the intersection. The Venn-diagram for intersection is shown in the slide.

# Probability of Events III.

**Mutually exclusive events:** A and B cannot occur at the same time.
*(the patient is both pregnant and male)* $(A \cap B)=0$




**Independent events**: occurrence of A does not affect the occurrence of B
*(our first patient is male and the second one is female)*

We have to highlight two event relations.
First the mutually exclusive events that mean two events (A,B) cannot occur at the same time. (*The patient is both pregnant and male.*) It means the intersection of this two event is an empty set. On Venn-diagram we see that the two set has no intersection.

In the case of independent events the occurrence of an event doesn't affect the occurrence of the second one. (*Our first patient is male and the second one is female.*)

# Probability of Events IV.

**Conditional** probability
Probability of A **given that** B has occurred: **P(A|B).**
*(the probability that a patient suffering from a viral infection has actually flu – and not some other type of viral infection)*

Before we go on we have to define the conditional probability. Conditional probability is the probability of an event given an other event has occurred. (*The probability that a patient suffering from a viral infection has actually flu – and not some other type of viral infection.*) The notation for the conditional probability is P(A|B).

# Probability of Events V.

**Axioms on probability of events (Kolmogorov):**

*1.* **$0 \leq P(A) \leq 1$**

*2.* **P(*sure*) = 1** *(The patient will die sooner or later)*
**P(*impossible*) = 0** *(I'm 310 cm tall)*

*3. Mutually exclusive events (i.e.* P(AandB)=0)
***P(AorB)=P(A)+P(B)***
*(probability of being pregnant or male)*

And a theorem:
+4. *Independent events*: **P(AandB)=P(A)*P(B)**
*(probability that our first patient is male and the second one is female)*

To describe the probability of events we have axioms. Now we show the Kolmogorov axioms. (In a simplified way.)

1. The probability of an event is between 0 and 1.
2. The probability of a sure event (*the patient will die sooner or later* – we know that life is a sexually transmitted lethal disease☺) is 1. The probability of an impossible event is 0 (*I'm 310 cm tall*).
3. The probability of A or B events occur if A and B are mutually exclusive events is the sum of the probability of A and the probability of B events. (*The probability that being pregnant or male is the probability that being pregnant + the probability that being male.*)

A theorem based on the axioms:

+4. The probability of A and B events occur if A and B are independent events is the multiplication of the probability of A and the probability of B.
(*Probability that our first patient is male and the second one is female is the probability that our first patient is male * the probability that our second patient is female.*)

These mentioned statements are true from other way round. For example if P(A)*(P(B)=P(AandB) then A and B are independent events.

# Probability of Events VI.

*Conditional events calculation:*

*general form:* **P(A|B)=P(AandB)/P(B)**

**Special cases:**

*I. Independent events*:

*Probability that our second patient is male*
***if*** *the first one is female*

**P(A|B)=P(AandB)/P(B)**
**P(A|B)=P(A)*P(B)/P(B)**
**P(A|B)=P(A)**

*Probability that our second patient is male*
***if*** *the first one is female = Probability that our second patient is male*

There is an other important calculation that you have to know on conditional events. I showed here a simplified form of Bayes' law. The general form of the „multiplication role" is P(A|B)=P(AandB)/P(B). First examine 2 special case.

Case1: check the conditional probability in case of independent event. E.g. *probability that our first patient is male* ***if*** *the second one is female.* As we see the result the independent condition has no effect on the probability of the event; e.g. *probability that our second patient is male* ***if*** *the first one is female = probability that our second patient is male.*

We have the same with tossing a coin, rolling a die, etc.: the previous results has no effect to the next one.

# Probability of Events VI.

II. event A is a subset of event B

*Probability that a patient has a flu if suffering from a viral infection*

**P(A|B)=P(AandB)/P(B)**
**P(A|B)=P(A)/P(B)**

*Calculation:*

*The probability that a patient coming to our office has viral infection is 8% =P(B)*

*The probability of occurrence of flu infections at our office is 2% = P(A)*

*The probability that a patient suffering from a viral infection has actually flu is: P(A|B) = 2% / 8 % = 25%.*

Case2: event A is a subset of event B (all A is a B, but not all B is A).
An example:
*The probability that a patient coming to our office has viral infection is 8% =P(B)* – that is the probability of the condition.
*The probability of occurrence of flu infections at our office* – that is the probability of our event in the given sample. *P(A)* = 2%
*The probability that a patient suffering from a viral infection has actually flu - P(A|B)* – is 25%.

## Risk

| | | Illness | | |
|---|---|---|---|---|
| | | Yes | No | Sum |
| Risk factor | Yes | a | b | a+b |
| | No | c | d | c+d |
| | Sum | a+c | b+d | a+b+c+d |

Risk (probability) of the illness if the risk factor is *present*:

$$P(Ill_y \mid Risk_y) = \frac{P(Ill_y \cap Risk_y)}{P(Risk_y)} = \frac{\frac{a}{a+b+c+d}}{\frac{a+b}{a+b+c+d}} = \frac{a}{a+b}$$

Risk (probability) of the illness if the risk factor is *NOT present*:

$$P(Ill_y \mid Risk_n) = \frac{P(Ill_y \cap Risk_n)}{P(Risk_n)} = \frac{\frac{c}{a+b+c+d}}{\frac{c+d}{a+b+c+d}} = \frac{c}{c+d}$$

In the medical practice instead of conditional events sometimes we use the term *risk*. It is widely used in the studies with bivariate nominal variables where one of variable is called *risk factor (or exposure)* and the other one is the examined *illness*. In the table we can give the frequencies according to the headings.
E.g.: risk factor is the smoking habit (smoker/non-smoker), illness is the presence of lung cancer.

First look the risk of the illness if the risk factor is present. That is a conditional probability that we can calculate. Here I used the symbol for intersection instead of „and”)
We can conclude that this probability is the relative frequency of the illness in the risk factor group. (What is the ratio of ill people in the risk factor group.)

Calculate the risk of the illness if the risk factor is not present too.

# Risk Ratio

| | | Illness | | |
|---|---|---|---|---|
| | | Yes | No | Sum |
| Risk factor | Yes | a | b | a+b |
| | No | c | d | c+d |
| | Sum | a+c | b+d | a+b+c+d |

**Risk Ratio (RR) (or Relative Risk)**:
ratio of the probability of an **event occurring** if a risk factor is **present** to the probability of an **event occurring** if a risk factor does **not prese**nt.

$$\frac{P(Ill_y \mid Risk_y)}{P(Ill_y \mid Risk_n)} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{a*(c+d)}{c*(a+b)}$$

In these studies we are usually interested in the ratio of risks: how many times higher the probability of an illness if the risk factor is present comparison if the risk factor not present. E.g.: how many times higher is the probability to have a lung cancer if we smoke than we not smoke. This ratio is called the risk ratio (or Relative Risk), abbreviated by RR. (Note: in the US nomenclature relative risk could mean different ratios, therefore it is better to call it risk ratio)

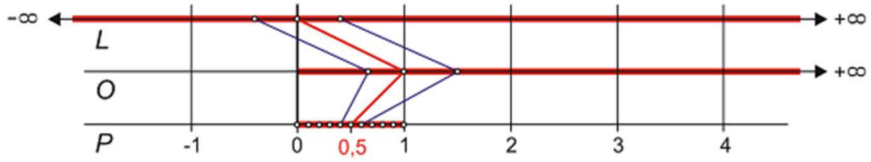
# Odds

***Odds (O)***: the ratio of the probability that a given event occurs and the probability that it does not occur. (how many times is the probability of an event occurring greater than not occurring)

$$O = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)}$$

***Logit (L)***: natural logarithm of odds

Logit, Odds, Probability




There is an other „probability like" parameter in probability calculus (and in gambling) that used very often. This parameter is the odds. Calculated as a ratio of the probability that a given event occurs and the probability that it does not occur (complementary event). (The meaning is how many times is the probability of an event occurring greater than not occurring.) [It is the equivalent scientific term for chance in the everyday speech.]

There is an other „probability like" parameter in probability calculus that used very often. This parameter is the odds. Calculated as a ratio of the probability that a given event occurs and the probability that it does not occur. (The meaning is how much larger is the probability of an evet occure than of not occure.)

The logit is a rarely used parameter. It is the natural logarithm of odds.
The connection between the value of the logit, odds and probability is shown in the slide.
We can notice for example that the probability is between 0 and 1. The odds between 0 and infinite. The odds is over 1 if the probability is over 0.5. The logit is negative if the probability is smaller than 0.5.

# Odds Ratio

| | | Illness | | |
|---|---|---|---|---|
| | | Yes | No | Sum |
| Risk factor | Yes | a | b | a+b |
| | No | c | d | c+d |
| | Sum | a+c | b+d | a+b+c+d |

Odds of the illness if the risk factor is *present*:

$$\frac{P(Ill_y \mid Risk_y)}{P(Ill_n \mid Risk_y)} = \frac{\frac{P(Ill_y \cap Risk_y)}{P(Risk_y)}}{\frac{P(Ill_n \cap Risk_y)}{P(Risk_y)}} = \frac{P(Ill_y \cap Risk_y)}{P(Ill_n \cap Risk_y)} = \frac{\frac{a}{a+b+c+d}}{\frac{b}{a+b+c+d}} = \frac{a}{b}$$

Odds of the illness if the risk factor is *NOT present*:

$$\frac{P(Ill_y \mid Risk_n)}{P(Ill_n \mid Risk_n)} = \frac{c}{d}$$

In the medical practice in the risk factor studies we use odds too. Here you can find how to calculate.

# Odds Ratio

| | | Illness | | |
|---|---|---|---|---|
| | | Yes | No | Sum |
| Risk factor | Yes | a | b | a+b |
| | No | c | d | c+d |
| | Sum | a+c | b+d | a+b+c+d |

**Odds Ratio (OR)**:
ratio of the odds of an **event occurring** if a risk factor is **present** to the odds of an **event occurring** if a risk factor does **not prese**nt.

$$\frac{\left(\frac{P(Ill_y \mid Risk_y)}{P(Ill_n \mid Risk_y)}\right)}{\left(\frac{P(Ill_y \mid Risk_n)}{P(Ill_n \mid Risk_n)}\right)} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a*d}{c*b}$$

E.g.: how many times higher is the *odds* to have a lung cancer if we smoke than we not smoke. This ratio is called the *odds ratio*, abbreviated by OR.

# Risk Ratio and Odds Ratio

| | | Illness | | |
|---|---|---|---|---|
| | | Yes | No | Sum |
| Risk factor | Yes | a | b | a+b |
| | No | c | d | c+d |
| | Sum | a+c | b+d | a+b+c+d |

**OR** **RR**

$$\frac{a*d}{c*b} \neq \frac{a*(c+d)}{c*(a+b)}$$

**Illness is rare**

$$\begin{matrix} a << b \\ c << d \end{matrix} \quad OR \Rightarrow RR$$

Let's compare the OR and the RR. As we see, OR not equal with RR – but OR tends to RR if a<<b and c<<d – so the illness is rare.

# Risk Ratio and Odds Ratio - calc

| | | Lung cancer | | |
|---|---|---|---|---|
| | | Cancer | No cancer | Sum |
| Smoking habit | Smoker | 79 | 71 | 150 |
| | Non-smoker | 9 | 18 | 27 |
| | Sum | 88 | 89 | 177 |

**OR**

$$\frac{a*d}{c*b}$$

$$\frac{79*18}{9*71} = 2{,}23$$

**RR**

$$\frac{a*(c+d)}{c*(a+b)}$$

$$\frac{79*27}{9*150} = 1{,}58$$

**Meaning? (R: Ratios)**

R=1 – „no risk effect"

R>1 – increased risk/odds with factor

R<1 – decresed risk with factor

**May be, may be NOT**

Let's make a calculation in an example. Meaning – really? Are you sure? Uncertainty? MAY BE (see later in the chi-square tests...)

# Probability Calculus

Permutations,
Variations,
Combinations

Probability calculation and statistics are based on the permutations, variations and combinations. But in this course we won't go into the details of math.

## *Probability Calculus Example*

During last year's flu epidemic 402 out of the total 2989 patients who turned up at a doctor's office required vaccination. Based on last year's data what is the probability that 4 vaccines will be sufficient (exactly, i.e. no vaccines will be left), if we are expecting a total number of 25 patients?

$$P = \binom{n}{k} \cdot (p)^k \cdot (1-p)^{(n-k)} = \binom{25}{4} \cdot \left(\frac{402}{2989}\right)^4 \cdot \left(1-\frac{402}{2989}\right)^{(25-4)} \approx 0{,}2$$

How to calculate (in excel)? How to read out from a graph, table? Which equation, table, excel function should we use?

An example why and how we use the probability calculus.
During last year's flu epidemic 402 out of the total 2989 patients who turned up at a doctor's office required vaccination. Based on last year's data what is the probability that 4 vaccines will be sufficient (exactly, i.e. no vaccines left) in a certain day, if we are expecting a total number of 25 patients?

For answering the question we use the Bernoulli distribution's (see later) equation. I show this equation to highlight that Bernoulli distribution is based on probability calculus.

To answer question like that, our main questions will be: How to calculate (in excel)? How do we know the equations? Which equation, table, excel function we should use?

## Human thinking and probability...

Tom is a quiet, shy, modest, hard-working guy who is happy to help others. Which is more probable?

a) Tom is a librarian
b) Tom is a blue-collar worker

Perhaps you think based on the description that Tom is a librarian more likely but if you think it over the frequency of male librarians and male blue-collar worker you should realize that Tomi is probably a blue-collar worker and not a librarian.

# Human thinking and probability...

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

a) Linda is a teacher in a secondary school
b) Linda works in bookstore and participates in yoga courses
c) Linda is a member of the league of women voters
d) Linda is a bank teller.
e) Linda is an insurance agent
f) Linda is a bank teller and is active in the feminist movement.

I'd like to highlight two statement: *d* and *f*. I hope everybody found out that co-occurrence is less probable than occurrence of a given event. The intersection of sets is always equal or smaller than the sets. So it is less probable that Linda is a bank teller and active feminist than she is a bank teller.

# Test Questions #1

- Give the definition of fprobability based on relative frequences.
- What is the law of large numbers?
- How tends the relative frequencies to the probability? [fluctuations, infinite seuence]
- How we can prove the law of large numbers?.
- What is the union of two sets?
- How we can notate the probability that events A or B occur?
- How we can notate the probability that both event A and B occur at the same time?
- What is the intersection of two event?
- What does it mean mutually exclusive events?
- Give an example for mutually excusive events.
- What is the the value of intersection of two mutually exclusive events?
- What does independent events mean?
- Give an example for independent events.
- What is the conditional probability?.
- Give an example for conditional probability.
- How we could notate conditional probability?
- How to calculate P(A) if P(A|B) and P(B) is given?
- What are the Kolmogorov's axioms?
- What is the relation betweenA and B events, if P(AorB)=P(A)+P(B) is true?
- What is the relation between A and B events, if P(AB)=P(A)*P(B) is true?
- What is the probability of sure event?
- What is the probabilty of an impossible event?
- Give an example for sure and impossible events.
- What could the value of an event's probability be?
- Define the odds.
- Define the logit.
- Calculate the logit if the probability of an event is 0,12.
- Calculate the odds if the probability is 0,4.
- Calculate the probability if the odds is 3.
- Calculate the probability if the logit is – 32.

The following questions may be answered using lecture material, consultation with practice teacher, or your own investigation (on the library or the internet). These test questions are examples for multiple choice items that may occur in the midterm and exam tests.

# Test Questions #2

Calculate the risk ratio and the odds ratio of the cancer among smokers comparison to non-smokers.

| | | Lung cancer | | |
|---|---|---|---|---|
| | | Cancer | No cancer | Sum |
| Smoking habit | Smoker | 79 | 71 | 150 |
| | Non-smoker | 9 | 18 | 27 |
| | Sum | 88 | 89 | 177 |