Principles of Biostatistics and Informatics

3rd Lecture: Elements of Probability Calculus

25th September 2017

Dániel VERES

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...)

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?

Population and Sample

Population

RANDOMNESS!

Sample

The size of the **population** usually does not allow the examination of all of its elements.

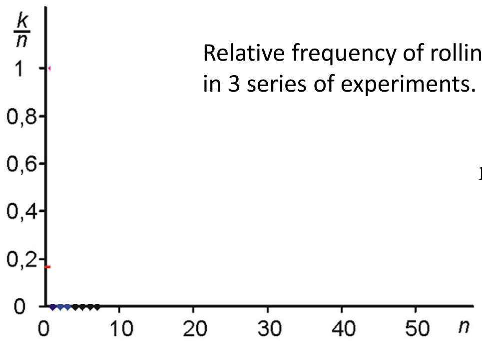
Therefore, only a subset of the population is examined. That is what we call a **sample**.

We carry out measurements on the sample elements, then this data set (which is also called **sample**) will be characterized by graphs and numbers

Characteristics of the sample can be used to draw conclusions on the population.

UNCERTAINTY!

An Other Experiment...

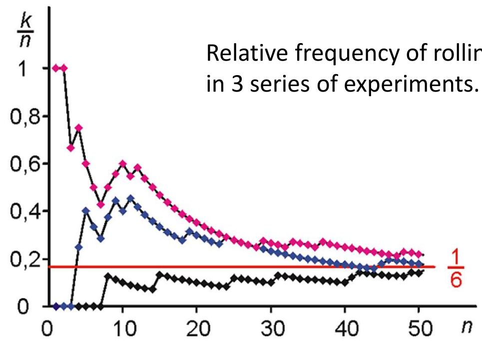


Relative frequency of rolling 6 with a die in 3 series of experiments.

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

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**.

An Other Experiment...

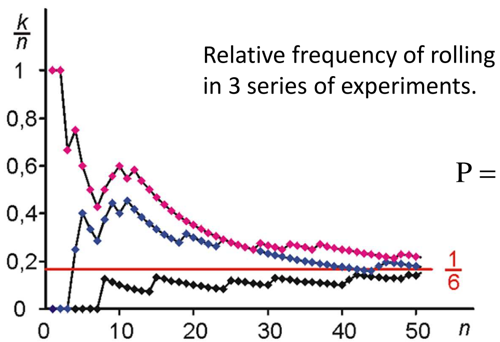


Relative frequency of rolling 6 with a die in 3 series of experiments.

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

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**.

Probability as a Quantity



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

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

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

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.

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)

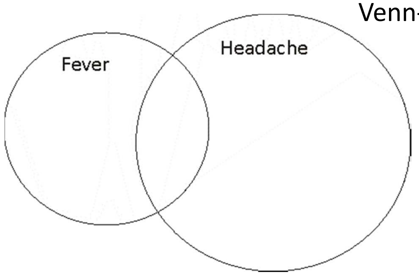
Complementary (complement) event: **Ā**
(the patient has NO fever)
Probability that event A NOT occurs**: P(Ā) or P(notA)**
(the probability that the patient has NO fever)

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)

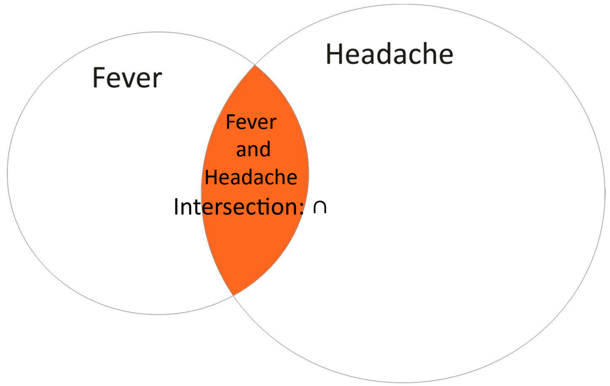
Probability that event A **or** event B occur: **P(AorB)**, **P(A+B)**, **P(AUB)**
(the probability that the patient has fever or headache)

Venn-diagrams



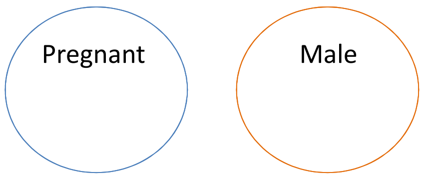
Probability of Events II.

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)



Probability of Events III.

Mutually exclusive events: A and B cannot occur at the same time.
(the patient is both pregnant and male) *(A∩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)

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)

Probability of Events V.

Axioms on probability of events (Kolmogorov):

1. **0 ≤ P(A) ≤ 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)

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*

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%.

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*:

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

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

Relative Risk or Risk Ratio (RR):
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.

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

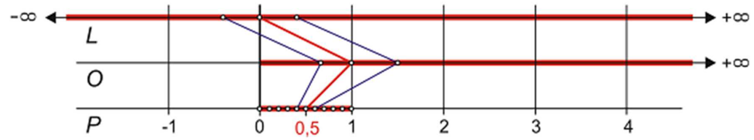
Odds

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)

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

Logit (L): natural logarithm of odds

Logit, Odds, Probability



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*:

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

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.

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

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

$$a << b \quad c << d \qquad OR \Rightarrow RR$$

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

Probability Calculus

Permutations,
Variations,
Combinations

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?

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

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.

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.

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