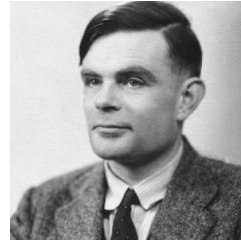


Decision Theory



Thomas Bayes, Pierre Simon de Laplace,, Bruno de Finetti, Alan Turing, Irving Good, Leonard Jimmie Savage, Dennis Lindley, Arnold Zellner, Kathryn Chaloner, Susie Bayarri , Daniel Kahneman

Who am I?

Anders Nordgaard

Reader and Forensic specialist in statistics
Swedish Police Authority – National Forensic Centre.

Former senior lecturer and director of studies at the Division of
Statistics (and Machine Learning), LiU.

Nowadays, adjunct lecturer at this division (up to 20 % of full
time)

Teaching this course
Supervision of Master's thesis work

Easiest way of contact: anders.nordgaard@liu.se

A course on decision making under uncertainty – Reasoning with probabilities

- Course responsible and tutor:

Anders Nordgaard (Anders.Nordgaard@liu.se)

- Course web page:

www.ida.liu.se/~732A66

- Teaching:

Lectures on theory

Practical exercises

Discussion of assignments

- Course book:

Winkler R.L. *An Introduction to Bayesian Inference and Decision* 2nd ed.

Probabilistic Publishing, 2003 ISBN 0-9647938-4-9

- Additional literature:
 - Taroni F., Bozza S., Biedermann A., Garbolino P., Aitken C. : Data analysis in forensic science – A Bayesian decision perspective, Chichester: Wiley, 2010
 - Gittelsohn S. (2013). Evolving from Inferences to Decisions in the Interpretation of Scientific Evidence. Thèse de Doctorat, Série criminalistique LVI, Université de Lausanne. ISBN 2-940098-60-3. Available at http://www.unil.ch/webdav/site/esc/shared/These_Gittelsohn.pdf.
 - Scientific papers (will be announced later)
- Examination:
 - Assignments (compulsory to pass)
 - Final oral exam (compulsory, decides the grade)

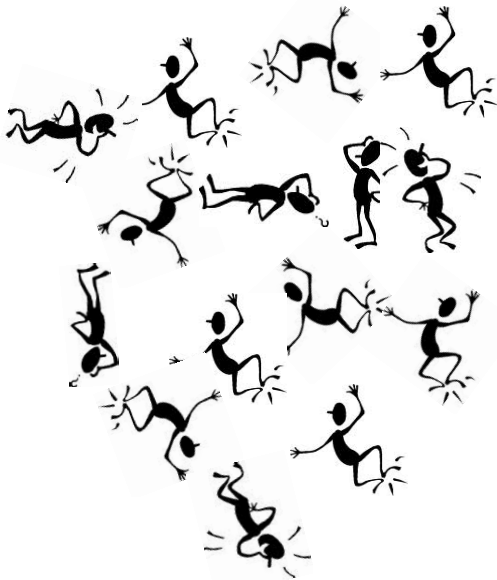
Changes of the schedule




Week	Day	Date	Time	Teacher	Type	Room
36	Wed	2019-09-04	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann
36	Fri	2019-09-06	13-15	Anders Nordgaard	Lecture/Tutorial	John von Neumann
37	Mon	2019-09-09	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann
37	Wed	2019-09-11	08-10	Anders Nordgaard	Lecture/Tutorial	John von Neumann
38	Fri	2019-09-20	13-15	Anders Nordgaard	Lecture/Tutorial	John von Neumann
39	Fri	2019-09-27	13-15	Anders Nordgaard	Lecture/Tutorial	John von Neumann
40	Wed	2019-10-02	08-10	Anders Nordgaard	Lecture/Tutorial	John von Neumann
41	Mon	2019-10-07	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann
41	Wed	2019-10-09	08-10	Anders Nordgaard	Lecture/Tutorial	John von Neumann
42	Mon	2019-10-14	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann
45	Mon	2019-11-04	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann
45	Wed	2019-11-06	08-10	Anders Nordgaard	Lecture/Tutorial	John von Neumann
46	Mon	2019-11-11	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann
47	Mon	2019-11-18	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann
48	Mon	2019-11-25	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann
48	Wed	2019-11-27	08-10	Anders Nordgaard	Lecture/Tutorial	John von Neumann
49	Mon	2019-12-02	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann
50	Mon	2019-12-09	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann
51	Mon	2019-12-16	15-17	Anders Nordgaard	Lecture/Tutorial	John von Neumann

Lecture 1 (Meeting 1 & 2): Repeat and extend...

Probability, random variables and likelihood

The concept of probability



<i>Category</i>	<i>Frequency</i>	<i>Probability</i> ?
	9	0.6
	3	0.2
	3	0.2

The general definition of probability

A random event:

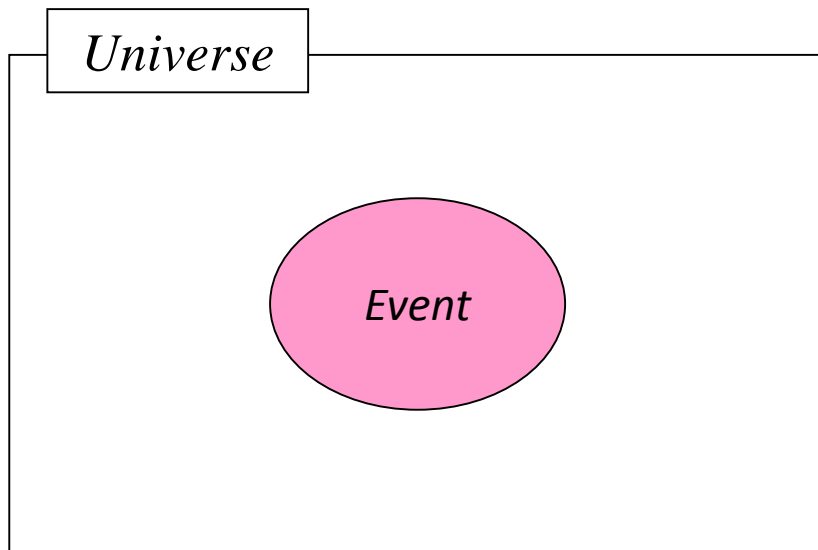
- A well-defined outcome or a collection of outcomes from an experiment
- The attained value or the collection of attained values of a quantity of interest
- The state of a variable

The universe (sample space):

- All possible outcomes from an experiment
- All possible values of a quantity of interest
- All possible states of a variable

The *probability* of an event is...

- the degree of belief in the event (that the event has happened)
- a measure of the size of the event relative to the size of the universe



Probability of event = $\Pr (Event)$

- $0 \leq \Pr (Event) \leq 1$
- $\Pr (Universe) = 1$
- If two events, A and B are mutually exclusive then

$$\Pr (A \text{ or } B) = \Pr (A) + \Pr (B)$$

The universe, all events in it and the probabilities assigned to each event constitute the *probability space*.

“Kolmogorov axioms”

This does not mean that...

“probabilities and stable relative frequencies are equal” (*Frequentist definition of probability*)

merely...

If any event is assigned a probability, that probability must satisfy the axioms.

Example: Coin tossing

Suppose you toss a coin. One possible event is “heads”, another is “tails”

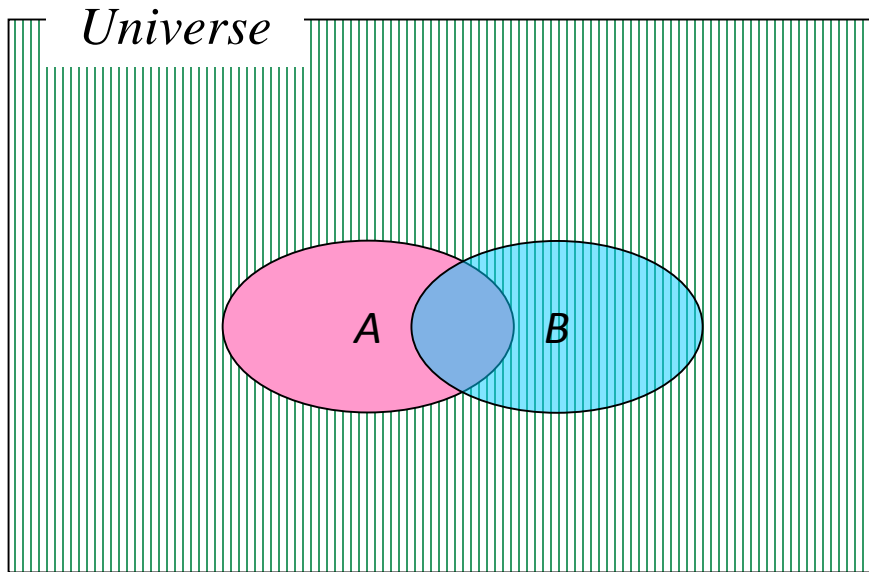
If you assign a probability p to “heads” and a probability q to “tails” they both must be between 0 and 1.

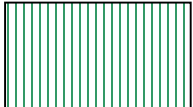
As “heads” cannot occur simultaneously with “tails”, the probability of “heads or tails” is $p + q$.

If no other event is possible then “heads or tails” = Universe →
 $p + q = 1$




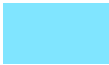

Calculation rules



Complementary event: \overline{A} ($\neg A, A^*, A^C$)  $\Pr(\overline{A}) = 1 - \Pr(A)$

Intersection: $A \cap B = (A, B)$  $\Pr(A, B)$

Union:

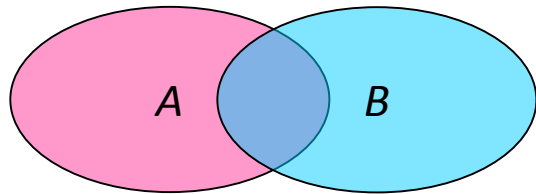
$A \cup B = A \text{ or } B$  +  -  $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A, B)$

Relevance, Conditional probabilities

An event B is said to be *relevant* for another event A if the probability (degree of belief) that A is true depends on the state of B .

The *conditional* probability of A given that B is true is

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$



If B is true then \bar{B} is *irrelevant* to consider.

If A is to be true under these conditions, only the part of A inside B should be considered.

This part coincides with (A, B)

The measure of the size of this event must be relative to the size of B

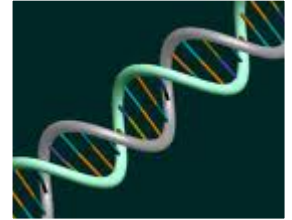
Example:

Assume you believe that approx. 1% of all human beings carry both a gene for developing disease *A* and a gene for developing disease *B*.

Further you believe that 10% of all human beings carry the gene for developing disease *B*.

Then as a consequence your degree of belief that a person who has developed disease *B* also carries the gene for developing disease *A* should be 10% ($0.01/0.10$)

Carrying the gene for *B* is relevant for carrying the gene for *A*.



What about the opposite conditioning?

Reversing the definition of conditional probability:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)} \Rightarrow \Pr(A, B) = \Pr(A|B) \cdot \Pr(B)$$

“The multiplication law of probability”

but also... $\Pr(A, B) = \Pr(B|A) \cdot \Pr(A)$

$$\Rightarrow \Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)} \quad \text{and} \quad \Pr(B|A) = \frac{\Pr(A|B) \cdot \Pr(B)}{\Pr(A)}$$

➔ For sorting out conditional probabilities it is not necessary to assign the probabilities of intersections

“All probabilities are conditional...”

How a probability is assigned depends on background knowledge.

E.g. if you assign the probability 0.5 for the event “heads” in a coin toss, you have assumed that

- the coin is fair
- the coin cannot land endways



...but it may be the case that you cannot assign any probability to the background knowledge

Let I denote all background knowledge *relevant* for A

$$\Rightarrow \Pr(A) = \Pr(A|I)$$

Extensions:

$$\Pr(A, B|I) = \Pr(A|B, I) \cdot \Pr(B|I)$$

$$\begin{aligned} \Pr(A_1, A_2, \dots, A_n|I) &= \\ &= \Pr(A_1|I) \cdot \Pr(A_2|A_1, I) \cdot \dots \cdot \Pr(A_n|A_1, A_2, \dots, A_{n-1}, I) \end{aligned}$$

Example: Suppose you randomly pick 3 cards from a well-shuffled deck of cards. What is the probability you will in order get a spade, a hearts and a spade?

I = The deck of cards is well-shuffled \Rightarrow It does not matter how you pick your cards.

Let A_1 = First card is a spade; A_2 = Second card is a hearts; A_3 = Third card is a spade

$$\begin{aligned}\Rightarrow \Pr(A_1, A_2, A_3|I) &= \Pr(A_1|I) \cdot \Pr(A_2|A_1, I) \cdot \Pr(A_3|A_1, A_2, I) = \\ &= \frac{13}{52} \cdot \frac{13}{51} \cdot \frac{12}{50} \approx 0.015\end{aligned}$$

Relevance and (conditional) independence

If B is relevant for A then $\Pr(A|B, I) \neq \Pr(A|I)$

If B is *irrelevant* for A then $\Pr(A|B, I) = \Pr(A|I)$

which in turn gives

$$\Pr(A, B|I) = \Pr(A|I) \cdot \Pr(B|I)$$

In this case A and B is said to be conditionally independent events. (In common statistical literature only *independent* is used as term.)

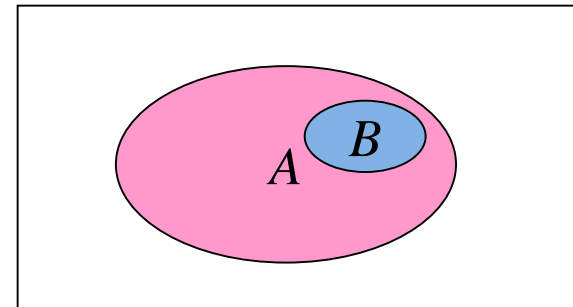
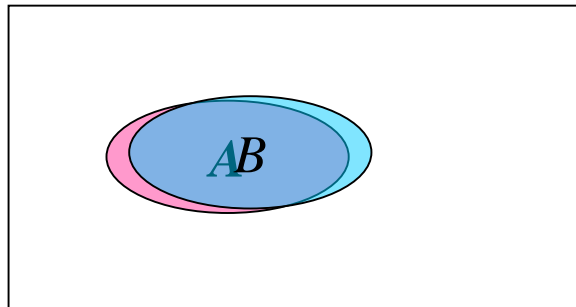
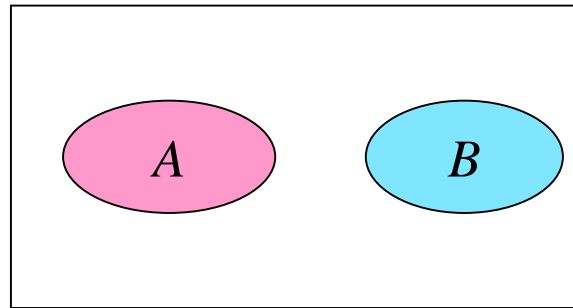
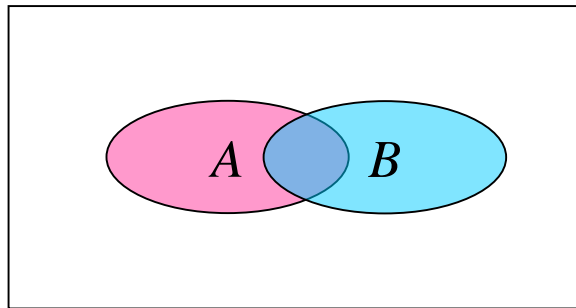
Note that it is the background knowledge I that determines whether this holds or not.

Note also that if $\Pr(A|B, I) = \Pr(A|I)$ then $\Pr(B|A, I) = \Pr(B|I)$

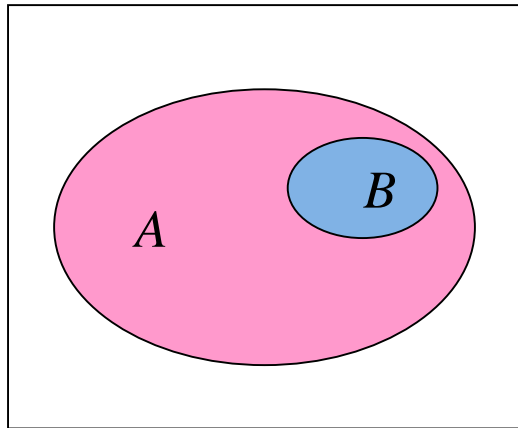
Irrelevance is reversible!

Assume that the sets below are drawn according to scale (the sizes of the sets are proportional to the probabilities of the events).

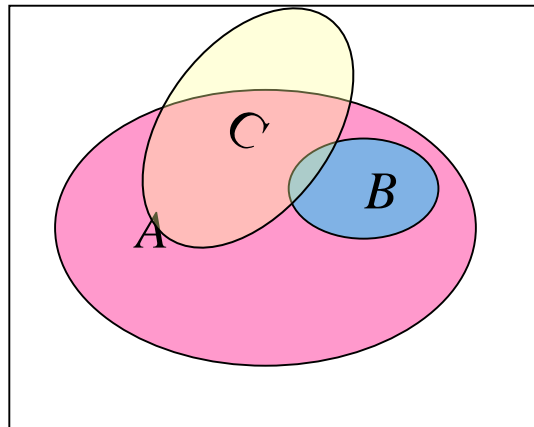
In which of the cases below may A and B be conditionally independent (given I)?



Further conditioning...



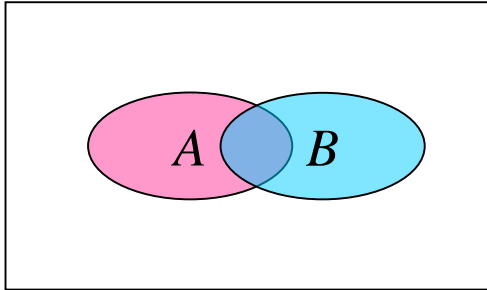
$$\Pr(A, B|I) \neq \Pr(A|I) \cdot \Pr(B|I)$$



$$\Pr(A, B|C, I) = \Pr(A|C, I) \cdot \Pr(B|C, I)$$

Two events that are conditionally dependent under one set of assumptions may be conditionally *independent* under another set of assumptions

The law of total probability and Bayes' theorem



The law of total probability:

$$\begin{aligned}\Pr(A|I) &= \Pr(A, B|I) + \Pr(A, \bar{B}|I) = \\ &= \Pr(A|B, I) \cdot \Pr(B|I) + \Pr(A|\bar{B}, I) \cdot \Pr(\bar{B}|I)\end{aligned}$$

➔ Bayes' theorem:

$$\Pr(A|B, I) = \frac{\Pr(B|A, I) \cdot \Pr(A|I)}{\Pr(B|A, I) \cdot \Pr(A|I) + \Pr(B|\bar{A}, I) \cdot \Pr(\bar{A}|I)}$$

➔ We don't need $\Pr(B|I)$ to compute $\Pr(A|B, I)$

Example:

Assume a method for detecting a certain kind of dye on banknotes is such that

- it gives a positive result (detection) in 99 % of the cases when the dye is present, i.e. the proportion of false negatives is 1%
- it gives a negative result in 98 % of the cases when the dye is absent, i.e. the proportion of false positives is 2%

The presence of dye is rare: prevalence is about 0.1 %



Assume the method has given positive result for a particular banknote.

What is the conditional probability that the dye is present?

Solution:

Let A = “Dye is present” and B = “Method gives positive result”

What about I ?

- We must assume that the particular banknote is as equally likely to be exposed to dye detection as any banknote in the population of banknotes.
- **Is that a realistic assumption?**

Now, $\Pr(A) = 0.001$; $\Pr(B|A) = 0.99$; $\Pr(B|\bar{A}) = 0.02$

Applying Bayes' theorem gives

$$\begin{aligned}\Pr(A|B) &= \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B|A) \cdot \Pr(A) + \Pr(B|\bar{A}) \cdot \Pr(\bar{A})} = \\ &= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.02 \cdot 0.999} = \underline{\hspace{2cm}}\end{aligned}$$

Odds and Bayes' theorem on odds form

The *odds* for an event A “is” a quantity equal to the probability:

$$Odds(A) = \frac{\Pr(A)}{\Pr(\overline{A})} = \frac{\Pr(A)}{1 - \Pr(A)} \Rightarrow \Pr(A) = \frac{Odds(A)}{Odds(A) + 1}$$

Why two quantities for the same thing?

- Sometimes practical (easier to talk in terms of “this many against that many”)
- The odds may take any value between 0 and infinity (∞), while the probability is restricted to values between 0 and 1

Example: An “epidemiological” model

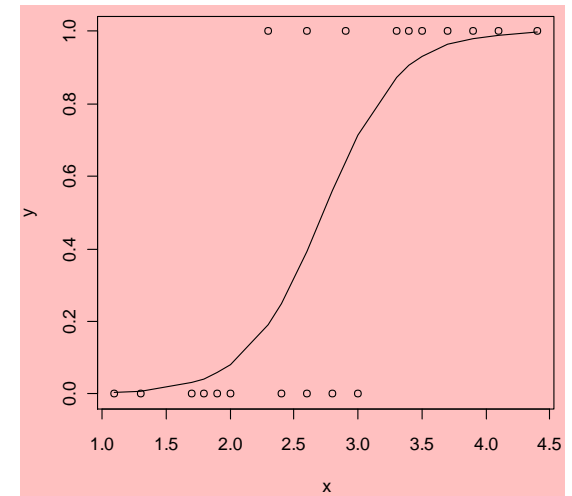
Assume we are trying to model the probability p of an event (i.e. the prevalence of some disease).

The *logit link* between p and a set of k explanatory variables x_1, x_2, \dots, x_k is

$$\text{logit}(p) = \ln \frac{p}{1-p} = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_k \cdot x_k$$

This link function is common in *logistic regression analysis*.

Note that we are modelling the natural logarithm of the odds instead of modelling p .



As the odds can take any value between 0 and ∞ the logarithm of the odds can take any value between $-\infty$ and ∞ ➔ Makes the model practical.

Conditional odds

$$Odds(A|B) = \frac{\Pr(A|B)}{\Pr(\bar{A}|B)}$$

expresses the *updated* belief that A holds when we take into account that B holds

Like probabilities, all odds are conditional if we include background knowledge I as our basis for the calculations.

$$Odds(A|I) = \frac{\Pr(A|I)}{\Pr(\bar{A}|I)}; \quad Odds(A|B, I) = \frac{\Pr(A|B, I)}{\Pr(\bar{A}|B, I)}$$

The odds ratio:

$$OR = \frac{Odds(A|B, I)}{Odds(A|I)} = \frac{\frac{Pr(A|B, I)}{Pr(\bar{A}|B, I)}}{\frac{Pr(A|I)}{Pr(\bar{A}|I)}}$$

expresses *how* the belief that A holds updates when we take into account that B holds

Example: In the epidemiological study we may want to assess how the odds for having a disease changes when an explanatory variable (like age) increases with one unit.

$$\text{logit}(p) = \ln \frac{p}{1-p} = \beta_0 + \beta_1 \cdot \text{age}$$

\Rightarrow

$$\text{odds}(\text{age}) = \frac{p}{1-p} = e^{\beta_0 + \beta_1 \cdot \text{age}}$$

$$\Rightarrow OR = \frac{\text{odds}(\text{age}+1)}{\text{odds}(\text{age})} = \frac{e^{\beta_0 + \beta_1 \cdot (\text{age}+1)}}{e^{\beta_0 + \beta_1 \cdot \text{age}}} = e^{\beta_1}$$

➔ The estimated value of β_1 is directly related to the *OR*

Now

$$\begin{aligned}\underline{Odds(A|B, I)} &= \frac{\Pr(A|B, I)}{\Pr(\bar{A}|B, I)} = \frac{\frac{\Pr(B|A, I) \cdot \Pr(A|I)}{\Pr(B|I)}}{\frac{\Pr(B|\bar{A}, I) \cdot \Pr(\bar{A}|I)}{\Pr(B|I)}} = \\ &= \frac{\Pr(B|A, I)}{\Pr(B|\bar{A}, I)} \cdot \frac{\Pr(A|I)}{\Pr(\bar{A}|I)} = \frac{\Pr(B|A, I)}{\Pr(B|\bar{A}, I)} \cdot \underline{Odds(A|I)}\end{aligned}$$

“Bayes’ theorem on odds form”

➔ The odds ratio is

$$\frac{\Pr(B|A, I)}{\Pr(B|\bar{A}, I)}$$

...and we notice that we do not need $\Pr(B | I)$ at all in the calculations.

The ratio

$$\frac{\Pr(B|A, I)}{\Pr(B|\bar{A}, I)}$$

is a special case of what is called a *likelihood ratio* (the concept of “likelihood” will follow)

$$LR = \frac{\Pr(B|A, I)}{\Pr(B|C, I)}$$

where we have substituted C for \bar{A} and we no longer require A and C to be complementary events (not even mutually exclusive).

$$\frac{\Pr(A|B, I)}{\Pr(C|B, I)} = \frac{\Pr(B|A, I)}{\Pr(B|C, I)} \cdot \frac{\Pr(A|I)}{\Pr(C|I)}$$

always holds, but the ratios involved are not always odds

“The updating of probability ratios when a new event is observed goes through the likelihood ratio based on that event.”

Example, cont.

Return to the example with detection of dye on bank notes.

(A = “Dye is present” and B = “Method gives positive result”)

$$\frac{\Pr(A|I)}{\Pr(\bar{A}|I)} = \frac{0.001}{0.999} = \frac{1}{999}$$

$$\frac{\Pr(A|B, I)}{\Pr(\bar{A}|B, I)} = \frac{\Pr(B|A, I)}{\Pr(B|\bar{A}, I)} \cdot \frac{\Pr(A|I)}{\Pr(\bar{A}|I)} = \frac{0.99}{0.02} \cdot \frac{1}{999} \approx 0.0495$$

$$\Rightarrow \Pr(A|B, I) \approx \frac{0.0495}{0.0495 + 1} \approx 0.047$$

Note! With Bayes’ theorem (original or on odds form) we can calculate

$\Pr(A | B, I)$ without explicit knowledge of $\Pr(B | I)$

Random variables and parameters

For physical measurements but also for observations it is most often convenient to formalise an event as a (set of) outcome(s) of a variable.

A random variable

- in strict mathematical definition is a mapping from the probability space to the Euclidean space \mathbf{R}^n ($n \geq 1$) – to comply with the definition of a cumulative distribution function (or to \mathbf{C}^n to account for complex-valued variables)
- is a variable, the value of which is not known in advance and cannot be exactly forecasted

➔ All variables of interest in a measurement situation would be random variables.

A parameter is another kind of variable, that is assumed to have a *fixed* value throughout the experiment (scenario, survey).

A parameter can often be controlled and its value is then known in advance (i.e. can be exactly forecasted).

The value attained by a random variable is often called *state*.

Examples:

1) Inventory of a grassland.

One random variable can be the area percentage covered by a specific weed.

The states of this variable constitute the range from zero to the total area of the grassland.

One parameter can be the distance from the grassland to the nearest water course.

2) DNA profiling

One random variable can be the genotype in a *locus* of the DNA double helix (*genotype* = the combination of two *alleles*, one inherited from the mother and one from the father; *DNA double helix* = the “entire DNA”).

The states of this variable are all possible combinations of alleles in that locus.

One parameter can be the locus itself.

Scales and probability measures

The states of a random variable can be given on different *scales*.

1) **Nominal scale**

A scale where the states have no numerical interrelationships.

Example: The colour of a sampled pill from a seizure of suspected illicit drug pills.

Each state can be assigned a probability > 0 .

Note that this does not go together with a random variable being a mapping to the Euclidean space.

2) Numerical scale

a) *Discrete states*

(i) **Ordinal scale**

A scale where the states can be put in ascending order.

Example: Classification of a dental cavity as small, medium-sized or large.

Each state can be assigned a probability > 0 .

Once probabilities have been assigned it is also meaningful to interpret statements as “at most”, “at least”, “smaller than”, “larger than”.....

➔ If we denote the random variable by X assigned state probabilities would be written $\Pr(X = a)$ and we can also interpret

$\Pr(X \leq a)$, $\Pr(X \geq a)$, $\Pr(X < a)$ and $\Pr(X > a)$.

(ii) Interval scale

An ordinal scale where the distance between two consecutive states is the same no matter where in the scale we are.

Example: The number of gun shot residues found on the hands of a person suspected to have fired a gun. The distance between 5 and 4 is the same as the distance between 125 and 124.

Probabilities are assigned and interpreted the same way as for an ordinal scale.

Interval scale discrete random variables very often fit into a family of *discrete probability distributions* where the assignment consists of choosing one or several parameters.

➔ Probabilities can be written on parametric form using a *probability mass function*, e.g. if X denotes the random variable and θ the parameter:

$$p(x|\theta) = \Pr(X = x|\theta)$$

Examples:

Binomial distribution:

Typical application: The number of “successes” out of n independent trials where for each trial the assigned probability of success is π .

$$p(x|n, \pi) = \frac{n!}{x!(n-x)!} \cdot \pi^x \cdot (1-\pi)^{n-x} ; x = 0, 1, \dots, n$$

Poisson distribution:

Typical application: Count data, e.g. the number of times an event occur in a fixed time period where λ is the expected number of counts in that period.

$$p(x|\lambda) = \frac{\lambda^x}{x!} \cdot e^{-\lambda} ; x = 0, 1, \dots$$

b) *Continuous states*

(i) **Interval scale**

This scale is of the same kind as for discrete states.

Example: Daily temperature in Celsius degrees

However, a probability > 0 cannot be assigned to a particular state.

Instead probabilities can be assigned to intervals of states.

The whole range of states has probability one.

The probability of an interval of states depends on the assigned *probability density function* for the range of states.

$$f(x|\theta)$$

Denote the random variable by X . It is thus only meaningful to assign probabilities like $\Pr (a < X < b)$ [which is equal to $\Pr (a \leq X \leq b)$].

Such probabilities are obtained by *integrating* the assigned density function (see further below).

(ii) **Ratio scale**

An interval scale with a well-defined zero state.

Example: Measurements of weight and length

The probability measure is the same as for continuous interval scale random variables.

More on continuous variables will come in future lectures

Probability and Likelihood

Two synonyms?

An event can be *likely* or *probable*, which for most people would be the same.

Yet, the definitions of probability and likelihood are different.

In a simplified form:

- The probability of an event measures the degree of belief that this event is true and is used for reasoning about not yet observed events
- The likelihood of an event is a measure of how likely that event is in light of another *observed* event
- Both use probability calculus

More formally...

Consider the *unobserved* event A and the *observed* event B .

There are probabilities for both representing the degrees of belief for these events in general:

$$\Pr(A | I), \quad \Pr(B | I)$$

However, as B is observed we might be interested in

$$\Pr(A | B, I)$$

which measures the *updated* degree of belief that A is true once we know that B holds. Still a probability, though.

How interesting is $\Pr(B | A, I)$?

$\Pr(B | A, I)$ might look meaningless to consider as we have actually observed B .

However, it says something about A .

We have observed B and if A is relevant for B we may compare $\Pr(B | A, I)$ with $\Pr(B | \bar{A}, I)$.

Now, even if we have not observed A or \bar{A} , one of them must be true (as a consequence of A and B being relevant for each other).

If $\Pr(B | A, I) > \Pr(B | \bar{A}, I)$ we may conclude that A is more *likely* to have occurred than is \bar{A} , or better phrased:

“ A is a better *explanation* to why B has occurred than is \bar{A} ”.

$\Pr(B | A, I)$ is called the *likelihood* of A given the observed B (and $\Pr(B | \bar{A}, I)$ is the likelihood of \bar{A}).

Note! This is different from the conditional probability of A given B : $\Pr(A | B, I)$.

Potential danger in mixing things up:

When we say that an event is the more likely one in light of data we do not say that this event has the highest probability.

Using the likelihood as a measure of how likely is an event is a matter of *inference to the best explanation*.

Logics: Implication:

$$A \rightarrow B$$

- If A is true then B is true, i.e. $\Pr(B \mid A, I) \equiv 1$
- If B is false then A is false, i.e. $\Pr(A \mid \bar{B}, I) \equiv 0$
- If B is true we cannot say anything about whether A is true or not (implication is different from equivalence)

“Probabilistic implication”:

$$A \xrightarrow{\text{Pr}} B$$

- If A is true then B *may* be true, i.e. $\Pr(B \mid A, I) > 0$
- If B is false the A may still be true, i.e. $\Pr(A \mid \bar{B}, I) > 0$
- If B is true then we may decide which of A and \bar{A} that is the best explanation

Inference to the best explanation:

- B is observed
- A_1, A_2, \dots, A_m are potential alternative explanations to B
- If for each $j \neq k$ $\Pr(B \mid A_k, I) > \Pr(B \mid A_j, I)$ then A_k is considered the best explanation for B and is provisionally accepted