Introduction to statistical inference 1

Lauri Mehtätalo
University of Eastern Finland
School of Computing

January 13, 2018

II

Contents

1 Preliminaries **1**

1.1 Introduction 1

1.2 Set theory 2

2 Random variables **7**

2.1 Random variables 7

2.2 Cumulative distribution function 9

2.3 Probability mass and density functions 11

2.4 Transformations of a random variable 13

Chapter 1

Preliminaries

1.1 Introduction

- Field of statistics builds on probability theory

> *"You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician."* - Sherlock Holmes

- The paragraph includes the important ideas of the statistical model:
  - The percentage p = the model of the process or underlying population
  - The behavior of individuals = data
- Assuming a constat probability p may be a too simplistic or naive assuption, and may be replaced by more realistic one where p is a function of the propabilities of the individual and the context where s/he is.
- We also need to specify a model, for the variability of the individuals around the p to complete the model formulation.
  - A crude summary if the variance-covariance matrix of the observations.
  - A complete definition is done by specifying the joint distribution of all individuals.
- We also may want to estimate how accurately we finally estimated p writing the available data
- The theoretical process that generates the data is called
  - Statistical model or (tilastollinen malli)

  - Stochastic process or (stokastinen prosessi)
  - Random process or (satunnaisprosessi)

- The process is random/stochastic because the "man" do not behave exactly according to model.[1]
  - Probability calculus and the theory of random variables provide tools to formulate and understand such models.
- Once model has been formulated or specified (muotoiltu), observed data can be used to[2]
  - estimate model parameters
  - evaluate the model fit (mallin sopivuus)
  - evaluate the inaccuracy related to the estimated model parameters
- When talking about models, we can talk about
  - True model (Tosi malli)
  - Estimated model (Estimoitu malli)
  - True model always stays the same, but as data used to formulate the estimated model gets larger, the estimated model gets closer to true model.
  - See example R-script `regsimu.R`

1.2 Set theory

- Consider a statistical experiment (e.g. rolling a dice, measuring the diameter of a tree, tossing a coin, measuring the photosynthetic activity in plant etc.)

Definition 1.1. All possible outcomes of a particular experiment (koe) form a set (joukko) called sample space (otosavaruus), denoted by S. For example:

A Toss of a coin; $S = \{H, T\}$

B Reaction time, Waiting time; $S = [0, \infty)$

C Exercise score of this course; $S = \{0, 1, 2, \ldots, 210\}$

D Number of points (events) within fixed area; $S = \{0, 1, 2, \ldots$

E CO_2 uptake within 0.5 hours in fixed area plot; $S = (-\infty, \infty)$

[1] This is what is done on this part of the course (ISI1).
[2] This is what is done on second part of the course (ISI2)

F Waiting time up to one hour (in minutes); $S = [0, 60)$

Sample space can be countable (numeroituva) or uncountable (ylinumeroituva). If the elements of a sample space can be put into one-to-one correspondence with a finite subset of integers, the space is countable. Otherwise, it is uncountable.

- Note: Examples A and C before are countable, the others are uncountable
- Note: If the waiting time in G are rounded to the minute / second / millisecond / microsecond, the sample space becomes countable.

Definition 1.2. An event (tapaus) is any collection of possible outcomes of an experiment, meaning it is a subset of S. Event A is said to occur, if the outcome of the experiment is in set A.

Example 1.1. Draw a card from standard deck.

$$S = \{\heartsuit, \diamondsuit, \clubsuit, \spadesuit\}$$

One possible event is $A = \{\heartsuit, \diamondsuit\}$. Another possible event is $B = \{\diamondsuit, \clubsuit, \spadesuit\}$. The union (unioni) of the two events includes all elements of both

$$A \cup B = \{\heartsuit, \diamondsuit, \clubsuit, \spadesuit\}$$

The intersection (leikkaus) includes elements that are common to both events

$$A \cap B = \{\diamondsuit\}$$

The complement (komplementti) of a set includes al elements of S that are not included in A

$$A^c = \{\clubsuit, \spadesuit\}$$

Events A and B are said to be disjoint (erillisiä), if

$$A \cap B = \emptyset$$

Number of events A_1, A_2, A_3, ... are said to be pairwise disjoint, if

$$A_i \cap A_j = \emptyset$$

for all pairs of i, j. In addition, if $\bigcup_{i=1}^{\infty} A_i = S$, then A_1, A_2, A_3, ... defines a partition of the sample space.

Example 1.2. Events $A = \{\heartsuit, \diamondsuit\}$ and $B = \{\clubsuit, \spadesuit\}$ are disjoint since

$$A \cap B = \emptyset$$

Events $A_1 = \{\heartsuit, \diamondsuit\}$, $A_2 = \{\clubsuit\}$ and $A_3 = \{\spadesuit\}$ are pairwise disjoint. Also, since

$$\bigcup_{i=1}^{3} = A_1 \cup A_2 \cup A_3 = \{\heartsuit, \diamondsuit, \clubsuit, \spadesuit\} = S$$

they are also partiotion of S.

Definition 1.3. Probability (todennäköisyys)

If a certain experiment is performed number of times (or infinite number of times), it may lead to different outcome, which is an event of the sample space. This frequency of outcome of an event is called probability.

For an event $A \subset S$ in an experiment, notation $P(A)$ (or $Pr(A)$) specifies the probability of outcome / event A.

Theorem 1.1. Axioms of probability

1. For every event A, $P(A) \neq 0$ (meaning every event is possible)
2. $P(S) = 1$ (because something will be observed)
3. For a sequence of pairwise disjoint events,

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Example 1.3. Assume we have two events A_1, $X \in A_1$ and A_2, $X \in A_2$, which have probabilities $P(A_1) = 0.2$ and $P(A_2) = 0.3$.

If the events are disjoint ($A_1 \cap A_2 = \emptyset$), probability for the union of events is

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) = 0.2 + 0.3 = 0.5$$

If events are not disjoint ($A_1 \cap A_2 \neq \emptyset$), then

$$P(A_1 \cup A_2) \neq P(A_1) + P(A_2) = 0.2 + 0.3 = 0.5$$

Example 1.4. In a fair deck, define events

$$A_1 = \{\heartsuit\}, A_2 = \{\diamondsuit\}, A_3 = \{\clubsuit\}, A_4 = \{\spadesuit\}$$

which have probabilities

$$P(A_1) = P(A_2) = P(A_3) = P(A_4) = 1/4$$

Events $A_1 \ldots A_4$ are disjoint. Therefor,

$$B = A_1 \cup A_2 = \{\heartsuit, \diamondsuit\}$$

$$P(B) = P(A_1 \cup A_2) = P(A_1) + P(A_2) = 1/4 + 1/4 = 1/2$$

Theorem 1.2. Consider events A and B

1. $P(A^c) = 1 - P(A)$
2. $P(B \cap A^c) = P(B) - P(A \cap B)$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. If $A \subset B$, then $P(A) \leq P(B)$

Note Consider case 3 of theorem 1.2

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) \leq 1$$

$$P(A) + P(B) - P(A \cap B) \leq 1$$

$$P(A \cap B) \geq P(A) + P(B) - 1$$

This equation is called the Bonferroni inequality. Idea is, that if we have intersection of two events (A and B), the probability of the intersection can be shown to be higher or equal than the right term.

$$P(A \cap B) \geq P(A) + P(B) - 1$$

Supose A and B are two events that occur with probability $P(A) = P(B) = 0.975$. Then $P(A \cap B) = 0.975 + 0.975 - 1 = 0.95\ldots$

Theorem 1.3. a) $P(A) = \sum_{i=1}^{\infty} P(A \cap C)$ for any partition $C_1, C_2, \ldots$

b) $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$, for any sets $A_1, A_2, \ldots$

Definition 1.4. IF A and B are events in a sample space S and $P(B) > 0$, then conditional probability (ehdollinen todennäkoisyys) of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Example 1.5. Clinical trial. Assume that we know the probabilities that are represented in table 1.1. Let event A be the event "the patient recovered" with sample space $S = \{\text{“}OK\text{”}, \text{“}NOT\ OK\text{”}\}$, and event B is the event "The patient was treated with placebo" with $S = \{\text{“}YES\text{”}, \text{“}NO\text{”}\}$.

Now $P(A \cap B)$ can be computed directly using the table 1.1, and it is

$$P(A \mid B) = P(\text{“}OK\text{”} \mid \text{“}Placebo\text{”}) = \frac{P(A \cap B)}{P(B)} = \frac{0.160}{0.227} \approx 0.70$$

Table 1.1: Probabilities related to clinical trial. Four different drugs and the probabilities that patient got or did not got ok after using it.

	Drug1	Drug2	Drug3	Placebo	Total
OK	0.120	0.087	0.147	0.160	0.513
NOT OK	0.147	0.167	0.107	0.067	0.487
Total	0.267	0.253	0.253	0.227	1.000

Definition 1.5. Bayes rule

Assuming we have event A and B,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \| * P(B)$$

if $P(B) > 0$

$$P(A \cap B) = P(B)P(A|B)$$

And the other way around

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} \quad \| * P(A)$$

if $P(A) > 0$

$$P(A \cap B) = P(A)P(B|A)$$

and by using the earlier $P(A \cap B) = P(B)P(A|B)$

$$P(B)P(A|B) = P(A)P(B|A) \quad \| : P(B)$$

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

This is known as Bayes rule.

Definition 1.6. Two events A and B are independent, if

$$P(A \cap B) = P(A)P(B)$$

If two events are known to be independent, we can compute $P(A \cap B)$ as $P(A)P(B)$

Example 1.6. In example 1.5 we had $P(A) = 0.513$, $P(B) = 0.227$ and $P(A \cap B) = 0.160$. By calculating

$$P(A)P(B) = 0.513 * 0.227 = 0.116 \neq 0.160 = P(A \cap B)$$

we can determine that events A and B are not independent.

Chapter 2

Random variables

2.1 Random variables

Definition 2.1. A random variable (RV for short) (satunnaismuuttuja) is a function from sample space S into the real numbers.
Note If S includes only real numbers, the definition of the RV is implicit. Some examples of this in table 2.1. Examples of events that not implicitly connected to random variables in table 2.2

The probability function of the original sample space is defined ot the RV as follows:

Consider the sample space $S = S_1, S_2, \ldots, S_n$, and has probability function P. Define RV X with sample space / range / support set

$$\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$$

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\})$$

Note About the notations: Uppercase (capital) letters are used for a random variables and lowercase letters for the realized value or range. $X = x$ means, that RV X gets value x.

Table 2.1: Examples of implicit connection between event and random variable

Experiment	**Event = Random variable**
Toss of two dice.	X = Sum of two numbers
Estimated regression from n observations from certain true model.	X = Estimate of slope coefficient $\hat{\beta}_2$
Event location in space.	X = # of events in certain fixed subset of the space.

Table 2.2: Examples of implicit connection between event and random variable

Experiment	**S**	**RV**
Toss of a coin	$\{$"H'', "$T''\}$	$X_1 = \begin{cases} 1 & \text{if Head,} \\ 0 & \text{if Tails} \end{cases}$ or $X_2 = \begin{cases} 10 & \text{if Head,} \\ 2 & \text{if Tails} \end{cases}$
Health status of tree	$\{$"$Healthy$", "$Sick$", "$Dead$"$\}$	$X = \begin{cases} 1 & \text{if Healthy,} \\ 2 & \text{if Sick,} \\ 3 & \text{if Dead} \end{cases}$ or we can define two random variables $X_1 = \begin{cases} 0 & \text{if Healthy or Dead,} \\ 1 & \text{if Sick} \end{cases}$ $X_2 = \begin{cases} 0 & \text{if Healthy or Sick,} \\ 1 & \text{if Dead} \end{cases}$

Table 2.3: Example 2.1 experiment's sample space value conversion to RV values.

s_i	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$X(s_i)$	3	2	2	1	2	1	1	0
$Y(s_i)$	0	1	1	2	2	1	2	3

Example 2.1. Tree tosses of a coin. Let RV X be the # of heads, or RV Y is the number of tails. For RV X, well get space

$$\mathcal{X} = \{0, 1, 2, 3\}$$

See table 2.3 to see how the sample space values are converted to random variable values. We can then calculate the probabilities that random variable gets certain value from its support space. For random variable X, these probabilities are

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Now, we can get the probability of event, that with three coin tosses, we get only one heads, by getting the probability that random variable X receives value 1

$$P(X = 1) = P(\{HTT,\ THT,\ TTH\}) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

Example 2.2. Waiting time in a phone service. Sample space for experiment is

$$S = [0, 60)$$

For RV X, space will be

$$\mathcal{X} = [0, 60)$$

meaning since the original experiment has sample space that is defined by real number, the experiment directly specifies the random variable

$$X(s) = s$$

2.2 Cumulative distribution function

- Starting point with any computation of probability is the distribution function.

Definition 2.2. The culmulative distribution function (kertymäfunktio) or just distribution function (cdf for short) is defined as

$$F_X(x) = P_X(X \leq x) \text{ for all } x$$

Example 2.3. Tossing of 3 coins. Random variable X = the # of heads. Cumulative distribution function for RV X can be defined as

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \quad x \in (-\infty, 0) \\ \frac{1}{8} & \text{if } 0 \leq x < 0 \quad x \in [0, 1) \\ \frac{1}{2} & \text{if } 1 \leq x < 0 \quad x \in [1, 2) \\ \frac{7}{8} & \text{if } 2 \leq x < 0 \quad x \in [2, 3) \\ 1 & \text{if } 3 \leq x < \infty \quad x \in [3, \infty) \end{cases}$$

Notes

- $F_X(x)$ is defined for all x!
- $F_X(x) = 0$ for $x < 0$ and $F_X(x) = 1$ for $x \geq 3$.
- $F_X(x)$ is right-continuous: the value of $F_X(x)$ at the jump is the one we get by approaching the jump from the right.
- $F_X(x)$ is discontinuous (in this example).

Theorem 2.1. Function $F(x)$ is a cdf if and only if the following conditions hold:

a) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

b) $F(x)$ is non-decreasing (kasvava) (meaning value of function is either increasing or constant as x increases)

c) $F(x)$ is right-continuous, i.e. for every x_0

$$\lim_{x \downarrow x_0} F(x) = F(x_0)$$

- If a function is said to be a cdf, it has to fulfill the rules a, b and c of theorem 2.1.
- If a function fulfills rules a, b and c of theorem 2.1, then it can be used as a cdf.

Theorem 2.2. A random variable X is said to be continuous, if $F_X(x)$ is a continuous function of x (meaning the function foes hot have jumps).

Note The X in example 2.1 was discrete.

Example 2.4. An example of a continuous cdf is the uniform distribution. Let RV U have equal probabilities within $\mathcal{U} \in [a, b]$. The cdf of U is

$$F_U(u) = \begin{cases} 0 & \text{if } u < a \quad u \in (-\infty, a) \\ \frac{u-a}{b-a} & \text{if } a \leq u < b \quad u \in [a, b) \\ 1 & \text{if } u \geq b \quad u \in [b, \infty) \end{cases}$$

- The cdf can be used for calculation of the probabilities. In general

$$P(x \in (l, u]) = F(u) - F(l)$$

Example 2.5. Consider the tossing of 3 coins (example 2.1). Probability that more than 1 heads is observed is

$$P(X \in (1,3]) = F_X(3) - F_X(1) = 1 - 0.5 = 0.5$$

This result is equal to

$$P(X = 2) + P(X = 3) = \frac{3}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$$

Example 2.6. Consider uniform distribution where $a = 0$ and $b = 2$.

$$P(X \in (0.5, 1]) = F(1) - F(0.5) = 0.5 - 0.25 = 0.25$$

Note For continuous RV

$$P(x \in [a,b]) = P(x \in (a,b)) = P(x \in (a,b]) = P(x \in [a,b))$$

Definition 2.3. RVs X and Y are said to be identically distributed (samoin jakautuneita), if $F_X(x) = F_Y(x)$ for every x.

Two RVs are said to be independent and identically distributed (i.i.d.) if they are both independent and identically distributed.

2.3 Probability mass and density functions

- The probability mass function (pmf) is defined only for discrete RVs as follows.

$$f_X(x) = P(X = x) \text{ for all } x$$

- The pmf is connected with cdf as follows.

$$F_X(x) = \sum_{k=1}^{X} f_X(k)$$

Example 2.7. The $Bernoulli(p)$ distribution is used for a discreate RV with two possible values (i.e. it is the distribution for so called Bernoulli trial, which has two possible outcomes: success ("S") and failure ("F")). Record the outcomes as a RV:

s_i	X(s_i)
"S"	1
"F"	0

The pmf of X is

$$F(x;p) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \\ 0 & \text{elsewhere} \end{cases}$$

The Bernoulli cdf is

$$F(x;p) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Note a and b for the uniform distribution (like in example 2.4) and the p in case of Bernoulli distribution (like in example 2.7) are called parameters.

- pmf cannot be defined for a continuous RV since it holds

$$P(X = x) \leq \lim_{\epsilon \downarrow 0} [F_X(x) - F_X(x - \epsilon)] = 0$$

- For continuous RV, the sum of pmf is replaced by integral...

$$P(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

- ... which also implies

$$\frac{d}{dx} F_X(x) = f_X(x)$$

Definition 2.4. Function $f_X(x)$ above is called probability density function (pdf for short) (tiheysfunktio).

Theorem 2.3. A function $f(x)$ is a pdf or pmf, if

a) $f(x) \geq 0$ for all x

b)

$$\sum_X f(x) = 1 \text{ (if pmf)}$$

or

$$\int_{-\infty}^{\infty} f(x)dx = 1 \text{ (if pdf)}$$

Note Notation $\sim$ means "is distributed as"; for example

- $X \sim f(x)$
- $X \sim F(x)$
- $X \sim Bernoulli(p)$

Example 2.8. Perhaps the most widely used continuous distribution is the Normal distribution, which has support $(-\infty, \infty)$, and pdf

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ -\infty < \mu < \infty, \ \sigma > 0$$

If $\mu = 0$ and $\sigma = 1$, then function is called standard normal density (standardoitu normaali jakauma). The cdf of normal distribution

$$\int_{-\infty}^{x} f_X(t)dt = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{\sigma^2}}dt = \dots$$

but it cannot be written in closed form. However, R and other software can compute the value numerically.

Note Parameters: μ is the expected value, and σ is the standard deviation of X.

Example 2.9. Let $U \sim Uniform(0, 10)$ and X be the sample mean $\frac{\sum_{i=1}^{n} u_i}{n}$ i n a sample of size n. R-script `unifmean.R` illustrates, that if n is sufficiently large, then

$$X \sim N(\frac{10}{2}, \frac{10/\sqrt{12}}{n})$$

This is a consequence of central limit theorem.

Note Often the parameters of a distribution are functions, not singe numbers. For example, the probability of success, p, in the Bernoulli case may depend on some fixed/known characteristics x of sampling unit, ie. we have $p(x)$ instad of the p. Or the mean of normally distributed RV may also depend on x, so we use $\mu(x)$ instead of μ. Examples 1.1. and 1.2. in `notes.pdf` demonstrate this

Example 2.10. $p(Age) = 0.1 + 0.0005 * Age$

If $Age = 0$, then $p(0) = 0.1$.

If $Age = 10$, then $p(10) = 0.15$

If $Age = 100$, then $p(10) = 0.1 + 0.5 = 0.6$

Since $p(Age)$ as specified above may become negative or higher than 1 with some values of the parameter Age, function $p(x)$ should be defined so, that it can only get values within $(0, 1)$. This could be achieved with binary logistic model.

2.4 Transformations of a random variable

- If X is a random variable, then any function of it, e.g. $Y = g(x) : \mathcal{X} \to \mathcal{Y}$ is also a random variable, where the function $g(x)$ is called transformation.
- The probabilistic properties of Y can be expressed using those of X:

$$P(Y \in A) = P(g(x) \in A)$$

- If $g(x)$ is monotonic function in X, we can define the inverse function $g^{-1}(y) : Y \to \mathcal{X}$ and establish the theorem 2.4.

Theorem 2.4. Let X have cdf $F_X(x)$. Let $Y = g(x)$ and $\mathcal{X} = \{x, f_X(x) > 0\}$ and $\mathcal{Y} = \{y : y = (g(x) \text{ for some } x \in \mathcal{X}\}$

a) If g is an increasing function of $\mathcal{X}$, then

$$F_Y(y) = F_X(g^{-1}(y)) \text{ for } y \in \mathcal{Y}$$

- If $g(x)$ specifies $y(x)$, we solve $y(x) = y$ for x to write x in terms of y $x(y) = g^{-1}(y)$.

b) If g is decreasing function of X and X is continuous, then

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

Example 2.11. Suppose $X \sim Uniform(0, 1)$. Now $F_X(x) = x,\ 0 < x < 1$. Consider transformation $Y = g(x) = -ln(x)$.

- $ln(x)$ is decreasing, since $\frac{d}{dx}(-ln(x)) = -\frac{1}{x} < 0$ when $x > 0$

$$\mathcal{X} \in (0, 1) \to \mathcal{Y} \in (0, \infty)$$

- Because $g(x)$ is decreasing, the minimum of Y will be obtained at the maximum value of X.

$$-ln(1) = 0, \text{ (minimum of } Y)$$

- The maximum of Y is obtained at the minimum of X.

$$-ln(x) \to \infty, \text{ as } x \to \infty$$

- We can get the g^{-1} by solving x for $y = g(x)$

$$y = -ln(x) \ \ || * (-1)$$

$$-y = ln(x)$$

$$x = e^{-y} = g^{-1}(y)$$