

Introduction to statistical inference 1

Lauri Mehtätalo
University of Eastern Finland
School of Computing

January 25, 2018

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 |
| 2.5 | Expected value | 16 |
| 3 | Bivariate random variables | 19 |
| 4 | Conditional distributions | 23 |

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 constant probability p may be a too simplistic or naive assumption, and may be replaced by more realistic one where p is a function of the probabilities 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 to²
 - 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; $S = \{0, 1, 2, \dots, 210\}$
- D Number of points (events) within fixed area; $S = \{0, 1, 2, \dots\}$
- E CO₂ uptake within 0.5 hours in fixed area plot; $S = (-\infty, \infty)$

¹This is what is done on this part of the course (ISI1).

²This 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 all 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, \dots 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, \dots 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

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

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

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

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 $A_1 \dots A_4$ 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.

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

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

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

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

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

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

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 $S = \{“OK”, “NOT OK”\}$, and event B is the event “The patient was treated with placebo” with $S = \{“YES”, “NO”\}$.

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

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

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 ,

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

if $P(B) > 0$

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

And the other way around

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

if $P(A) > 0$

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

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

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

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

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

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

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

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 | $\{“H”, “T”\}$ | $X_1 = \begin{cases} 1 & \text{if Head,} \\ 0 & \text{if Tails} \end{cases}$ <p>or</p> $X_2 = \begin{cases} 10 & \text{if Head,} \\ 2 & \text{if Tails} \end{cases}$ |
| Health status of tree | $\{“Healthy”, “Sick”, “Dead”\}$ | $X = \begin{cases} 1 & \text{if Healthy,} \\ 2 & \text{if Sick,} \\ 3 & \text{if Dead} \end{cases}$ <p>or we can define two random variables</p> $X_1 = \begin{cases} 0 & \text{if Healthy or Dead,} \\ 1 & \text{if Sick} \end{cases}$ $X_2 = \begin{cases} 0 & \text{if Healthy or Sick,} \\ 1 & \text{if Dead} \end{cases}$ |

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

$$P(X = x) \begin{array}{c|cccc} x & 0 & 1 & 2 & 3 \\ \hline & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array}$$

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

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

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) $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$.
- 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 does not 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

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

- 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 discrete 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

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

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

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

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

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

This is a consequence of central limit theorem.

Note Often the parameters of a distribution are functions, not single 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 $p(x)$ instead 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. $p(\text{Age}) = 0.1 + 0.0005 * \text{Age}$

If $\text{Age} = 0$, then $p(0) = 0.1$.

If $\text{Age} = 10$, then $p(10) = 0.15$

If $\text{Age} = 100$, then $p(100) = 0.1 + 0.5 = 0.6$

Since $p(\text{Age})$ 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. $Y = g(x) : \mathcal{X} \rightarrow \mathcal{Y}$ 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 $g^{-1}(y) : \mathcal{Y} \rightarrow \mathcal{X}$ 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 \mathcal{X} and X is continuous, then

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

Example 2.11. Suppose $X \sim \text{Uniform}(0, 1)$. 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 $x > 0$

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

- 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 .

$$-\ln(x) \rightarrow \infty, \text{ as } x \rightarrow 0$$

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

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

$$-y = \ln(x)$$

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

- The cdf of Y is

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

- The pdf of Y is

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d(1 - e^{-y})}{dy} = -e^{-y}(-1) = e^{-y}$$

- Note: $F_Y(y)$ is called the exponential cdf with the rate parameter 1. $f_Y(y)$ is the corresponding density.

Theorem 2.5. Let $X \sim f_X(x)$ and $Y = g(X)$ be a monotone transformation, and supports $\mathcal{X} = \{x, f_X(x) > 0\}$ and $\mathcal{Y} = \{y : y = (g(x) \text{ for some } x \in \mathcal{X})\}$ (like in theorem 2.4). Suppose that $f_X(x)$ is continuous on \mathcal{X} and $g^{-1}(y)$ has continuous derivative on \mathcal{Y} .

The pdf of Y is

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases}$$

Definition 2.5. Let X have cdf $F(x)$. Quartile function of X is defined as the inverse of cdf:

$$q(u) = F^{-1}(u)$$

Example 2.12. We want to find a value of X such that $P(X \leq x) = 0.95$. The solution is given by $q(x) = q(0.95) = \dots = x_*$

Example 2.13. Random number generator. Let $q_x(u)$ be the quartile function of X . If we are able to generate a sample from $U \sim \text{Uniform}(0, 1)$ distribution, transformation $g(u) = q_X(u)$ provides a random sample from the distribution of X .

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 0.1x & \text{when } 0 < x \leq 5 \quad (*) \\ 0.25 + 0.05x & \text{when } 5 < x \leq 15 \quad (**) \\ 1 & x > 15 \end{cases}$$

Then, the quartile functions for parts (*) and (**) are solved from $F(x) = u$ by solving x.

Part (*)

$$0.1x = u$$

$$x = 10u$$

$$\text{for } F(0) < u \leq F(5) \rightarrow 0 < u \leq 0.5$$

Part (**)

$$0.25 + 0.05x = u$$

$$x = 20u - 5$$

$$\text{for } F(5) < u \leq F(15) \rightarrow 0.5 < u \leq 1$$

Which provide us with the quartile function for X

$$q(u) = \begin{cases} 10u & \text{if } u \in [0, 0.5) \\ 20u - 5 + 0.05x & \text{if } u \in [0.5, 1] \end{cases}$$

Quartile function $q(u)$ implemented in R-script `probIntTrans.R`.

2.5 Expected value

Definition 2.6. Expected value or mean (odotusarvo) of a random variable $g(X)$ is

$$E(g(X)) = \begin{cases} \int_{-\infty}^{\infty} g(x)f(x) dx & \text{if } x \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x)f(x) & \text{if } x \text{ is discrete} \end{cases}$$

provided that sum/integral exists. If $E(g(X)) = \infty$, we say that the expected value does not exist.

Note If $g(x) = x$, the above rule simplifies to

$$E(g(X)) = \begin{cases} \int_{-\infty}^{\infty} xf(x) dx & \text{if } x \text{ is continuous} \\ \sum_{x \in \mathcal{X}} xf(x) & \text{if } x \text{ is discrete} \end{cases}$$

Example 2.14. Let $X \sim \text{Exponential}(\lambda)$ which has the pdf

$$f_X(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \quad 0 < x < \infty$$

And the expected value is

$$\begin{aligned} E(X) &= \int_0^{\infty} xf_X(x) dx \\ &= - \int_0^{\infty} x \left(-\frac{1}{\lambda} e^{-\frac{x}{\lambda}} \right) dx \\ &= \left|_0^{\infty} -xe^{-\frac{x}{\lambda}} + \int_0^{\infty} 1e^{-\frac{x}{\lambda}} dx \right. \\ &= -0 - (-0) + \left(-\lambda \int_0^{\infty} -\frac{1}{\lambda} e^{-\frac{x}{\lambda}} dx \right) \\ &= -\lambda \left|_0^{\infty} e^{-\frac{x}{\lambda}} \right. \\ &= -\lambda(0 - 1) = \lambda \end{aligned}$$

Computation of the expected values can get quite hard. Examples 1.10. and 1.11. in `notes.pdf` demonstrate numerical approximations.

Theorem 2.6. Let X and Y be random variables and a , b and c fixed constants. The following rules hold:

$$E(c) = c$$

$$E(cX) = cE(X)$$

$$E(X + Y) = E(X) + E(Y)$$

$$E(X + c) = E(X) + c$$

$$E(XY) \neq E(X)E(Y)$$

– $E(XY)$ depends on distribution of X and Y .

$$E(g(X)) \neq g(E(X))$$

- If g is nonlinear function of X .
- Also depends on the distribution of X .
- If g is linear, then we get rule (*).

Also

$$E(a * g_1(X) + b * g_2(Y) + c) = a * E(g_1(X)) + b * E(g_2(Y)) + c$$

Note $E(Y) = E(g(X))$ can be computed in two ways:

- 1) $E(g(x)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$
- 2) $E(g(x)) = E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} y f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| dy$

Definition 2.7. The variance is defined as

$$Var(X) = E \left[(X - E(X))^2 \right], \text{ and } var(X) \geq 0 \text{ always.}$$

The square root of $Var(X)$ is called the standard deviation sd

$$sd(x) = \sqrt{Var(X)}$$

Example 2.15. Let $X \sim Exponential(\lambda)$, meaning the X has pdf $f_X(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}$.

The variance of X is

$$\begin{aligned} Var(X) &= E \left[(x - \lambda)^2 \right] \\ &= E(g(X)) \\ &= \int_{-\infty}^{\infty} (x - \lambda)^2 \frac{1}{\lambda} e^{-\frac{x}{\lambda}} dx \\ &= \int_{-\infty}^{\infty} (x^2 - 2x\lambda + \lambda^2) \frac{1}{\lambda} e^{-\frac{x}{\lambda}} dx \\ &= \int_{-\infty}^{\infty} x^2 \frac{1}{\lambda} e^{-\frac{x}{\lambda}} dx - \int_{-\infty}^{\infty} 2x\lambda \frac{1}{\lambda} e^{-\frac{x}{\lambda}} dx + \int_{-\infty}^{\infty} \lambda^2 \frac{1}{\lambda} e^{-\frac{x}{\lambda}} dx \\ &= (\text{integrate each term by parts}) = \lambda^2 \end{aligned}$$

Note A computationally convenient form of $Var(x)$ is obtained as follows:

$$\begin{aligned} Var(X) &= E \left[(X - E(X))^2 \right] \\ &= E \left[X^2 - 2XE(X) + [E(X)]^2 \right] \\ &= E(X^2) - E[2XE(X)] + E[(E(X))^2] \\ &= E(X^2) - 2[E(X)]^2 + [E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

So $Var(X) = E \left[(X - E(X))^2 \right] = E(X^2) - [E(X)]^2$. For continuous random variables, the variance can be approximated numerically using integrate. See example 1.4. in notes.pdf. We can either use

$$Var(X) = \int_{-\infty}^{\infty} [X - E(X)]^2 f(x) dx$$

or

$$Var(X) = E(X^2) - [E(X)]^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \left(\int_{-\infty}^{\infty} x f(x) dx \right)^2$$

Theorem 2.7. Let X be a random variable with a finite variance and a and b are finite constants. Then

$$\begin{aligned} Var(aX + b) &= a^2 Var(X) \\ sd(aX + b) &= \sqrt{Var(aX + b)} \\ &= \sqrt{a^2 Var(X)} = a \sqrt{Var(X)} \\ &= a * sd(x) \end{aligned}$$

Definition 2.8. In general, the expected value $\mu'_n = E(X^n)$ is called n^{th} moment of random variable X . The expected value $\mu_n = E(X - \mu)^n$ is called the n^{th} central moment.

Expected value of X is the 1st moment and the variance of X is the 2nd moment of X . These two are called also the 1st and 2nd order properties of the (univariate) random variable X .

$$E(X) = \mu_1, \quad Var(X) = \mu_2 = \mu_2 - (\mu_1)^2$$

1st moment specifies the location of the distribution. The 2nd moment is related to the width of the distribution. The higher moments specify the shape of the distribution, eg. 3rd moment is related to the skewness of the distribution, and the 4th on the centroids.

The moments are “in the order of importance”: expected value is often the most important single number we can specify for the random variable X ; the variance is the second most important etc.

Chapter 3

Bivariate random variables

Definition 3.1. Let X and Y be two discrete random variables. The joint pmf is defined as

$$f(x, y) = P(X = x, Y = y) \quad \text{for all } x, y.$$

$$\text{or } f_{X,Y}(x, y) = P(X = x, Y = y) \quad f : \mathbb{R}^2 \rightarrow \mathbb{R}$$

Example 3.1. Two dice.

X = The sum of two dice

Y = |The difference of two dice|

The sample space includes $\{(1, 1), (1, 2), (1, 3), \dots, (6, 6)\}$. There are total of 36 different possibilities, and they are equally likely with probability $\frac{1}{36}$.

The joint pmf of (X, Y) can be expressed in a 2x2 table 3.1

Table 3.1: Probabilities of joint pmf $f_{X,Y}(x, y)$ (or just $f(x, y)$)

| | | X | | | | | | | | | | |
|-----|---|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Y | 0 | $\frac{1}{36}$ | 0 | $\frac{1}{36}$ | 0 | $\frac{1}{36}$ | 0 | $\frac{1}{36}$ | 0 | $\frac{1}{36}$ | 0 | $\frac{1}{36}$ |
| | 1 | 0 | $\frac{2}{36}$ | 0 | $\frac{2}{36}$ | 0 | $\frac{2}{36}$ | 0 | $\frac{2}{36}$ | 0 | $\frac{2}{36}$ | 0 |
| | 2 | 0 | 0 | $\frac{2}{36}$ | 0 | $\frac{2}{36}$ | 0 | $\frac{2}{36}$ | 0 | $\frac{2}{36}$ | 0 | 0 |
| | 3 | 0 | 0 | 0 | $\frac{2}{36}$ | 0 | $\frac{2}{36}$ | 0 | $\frac{2}{36}$ | 0 | 0 | 0 |
| | 4 | 0 | 0 | 0 | 0 | $\frac{2}{36}$ | 0 | $\frac{2}{36}$ | 0 | 0 | 0 | 0 |
| | 5 | 0 | 0 | 0 | 0 | 0 | $\frac{2}{36}$ | 0 | 0 | 0 | 0 | 0 |

- The joint pmf completely defines the probabilities of all combinations of X and Y .

$$P((X, Y) \in A) = \sum_{(x, y) \in A} f(x, y)$$

- From joint pmf of example 3.1, a probability that sum of two dices is in range $(2, 5]$, and that at same time, the difference of two dices is in range $[0, 1]$.

$$P(2 < X \leq 5, 0 \leq Y \leq 1) = \frac{2}{36} + 0 + \frac{2}{36} + 0 + \frac{1}{36} + 0 = \frac{5}{36}$$

- The expected value is defined as

$$E(g(x, y)) = \sum_{x, y \in \mathbb{R}^2} g(x, y) f(x, y)$$

- Expected value from joint pmf of example 3.1

$$\begin{aligned} E(XY) &= \sum_{(x, y) \in \mathbb{R}^2} xy f(x, y) \\ &= 2 * 0 * \frac{1}{36} + 4 * 0 * \frac{1}{36} + \dots + 12 * 0 * \frac{1}{36} \\ &\quad + 3 * 1 * \frac{2}{36} + 5 * 1 * \frac{2}{36} + \dots + 7 * 5 * \frac{2}{36} \\ &= 13 \frac{11}{18} \end{aligned}$$

- Them marginal pmf specifies the univariate distribution of one component of (X, Y) over all possible values of the other component.

Theorem 3.1. Let (X, Y) be discrete bivariate vector with joint pmf $f_{X,Y}(x, y)$. The marginal pmf's of it are given by

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y)$$

$$f_Y(y) = \sum_{x \in \mathbb{R}} f_{X,Y}(x, y)$$

- From example 3.1, we get marginal pdf seen in table 3.2 for random variable Y of joint pmf.

| Table 3.2: Probabilities of marginal pmf $f_Y(y)$ | | | | | | | |
|---|----------------|-----------------|----------------|----------------|----------------|----------------|---------------------|
| Y | 0 | 1 | 2 | 3 | 4 | 5 | \sum |
| $P(Y = y)$ | $\frac{6}{36}$ | $\frac{10}{36}$ | $\frac{8}{36}$ | $\frac{6}{36}$ | $\frac{4}{36}$ | $\frac{2}{36}$ | $\frac{36}{36} = 1$ |

- Note: Marginal pmf's do not completely specify the probabilities for all combinations of (X, Y)
- For continuous bivariate random variables, the joint pdf is defined correspondingly

Definition 3.2. A function $f(x, y)$ ($\mathbb{R}^2 \rightarrow \mathbb{R}$) is called the joint pdf of the continuous random variable (X, Y) , if for every $A \in \mathbb{R}^2$ we have

$$P((X, Y) \in A) = \iint_A f(x, y) \, dx \, dy$$

The difference to univariate case is that the integration is done over 2-dimensional set A . The expected value is defined as

$$E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) \, dx \, dy$$

The marginal pdf's are defined as before, but the sums are replaced by integrals

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y) \, dy \quad -\infty < x < \infty \\ f_Y(y) &= \int_{-\infty}^{\infty} f(x, y) \, dx \quad -\infty < y < \infty \end{aligned}$$

Any function satisfying rules a) and b) below is a joint pdf for some random variable (X, Y) .

a) $f(x, y) \geq 0$ for all $x, y \in \mathbb{R}^2$

b) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1$

Example 3.2. Define a joint pdf by

$$\begin{aligned} f(x, y) &= 6xy^2 \quad 0 < x < 1 \text{ and } 0 < y < 1 \\ \text{and } f(x, y) &= 0 \quad \text{otherwise} \end{aligned}$$

Is this a valid pdf?

a) $f(x, y) = 6xy^2 \geq 0$ for all $x, y \in \mathbb{R}^2$

b) Is the volume under the curve 1?

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy &= \int_0^1 \int_0^1 f(x, y) \, dx \, dy \\ &= \int_0^1 \int_0^1 6xy^2 \, dx \, dy \\ &= \int_0^1 3y^2 \, dy \\ &= \left|_0^1 3 * \frac{1}{3} y^3 \right. \\ &= \left|_0^1 y^3 \right. \\ &= 1^3 - 0^3 = 1 \end{aligned}$$

The volume under curve is 1, meaning it is a proper pdf.

Lets compute some probabilities:

$$\begin{aligned}
 P(X < 0.5, Y < 0.5) &= \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}} 6xy^2 \, dx \, dy \\
 &= \int_0^{\frac{1}{2}} \left|_0^{\frac{1}{2}} 3x^2 y^2 \, dy \right. \\
 &= \int_0^{\frac{1}{2}} 3 \frac{1}{4} y^2 \, dy \\
 &= \int_0^{\frac{1}{2}} \frac{3}{4} y^2 \, dy \\
 &= \left|_0^{\frac{1}{2}} \frac{3}{12} y^3 \, dy \right. \\
 &= \frac{1}{4} \left(\frac{3}{12} \right)^3 - 0 = \frac{1}{4 \cdot 8} = \frac{1}{32}
 \end{aligned}$$

The marginal pdf of X is for $0 < x < 1$

$$\begin{aligned}
 f_X(x) &= \int_{-\infty}^{\infty} f(x, y) \, dy \\
 &= \int_0^1 6xy^2 \, dy \\
 &= \left|_0^1 2xy^3 \right. \\
 &= 2x * 1^3 - 2x * 0^3 = 2x
 \end{aligned}$$

Example probability for marginal pdf

$$\begin{aligned}
 P\left(\frac{1}{2} < x < \frac{3}{4}\right) &= \int_{\frac{1}{2}}^{\frac{3}{4}} f_X(x) \, dx \\
 &= \int_{\frac{1}{2}}^{\frac{3}{4}} 2x \, dx \\
 &= \left|_{\frac{1}{2}}^{\frac{3}{4}} x^2 \right. \\
 &= \left(\frac{3}{4}\right)^2 - \left(\frac{1}{2}\right)^2 \\
 &= \frac{9}{16} - \frac{4}{16} = \frac{5}{16}
 \end{aligned}$$

- The joint cdf is defined by

$$F(x, y) = P(X \leq x, Y \leq y) \text{ for all } (x, y) \in \mathbb{R}^2$$

- Not convenient for discrete random variables, but for continuous case, we have

$$\begin{aligned}
 F(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f(s, t) \, ds \, dt \\
 \frac{dF(x, y)}{dx \, dy} &= f(x, y)
 \end{aligned}$$

Chapter 4

Conditional distributions

- Often two random variables are related.
- E.g. X = height and Y is the weight of a person.
- It is more likely, that $Y > 80\text{kg}$, if $X = 180\text{cm}$, than if $X = 150\text{cm}$.
- X gives some information on Y , but it does not tell Y exactly.
- **Note** If the relationship between X and Y would be exact, then we would have $Y = g(X)$ and we could analyze the data by considering Y as a transformation of X .
- Sometimes Y does not tell anything about X .
- We analyze such cases using the conditional distributions, which apply the previously presented idea about conditional probability to probability distribution.

Definition 4.1. Let (X, Y) be discrete bivariate random vector with joint pmf $f(x, y)$ and marginal pmf's $f_X(x)$ and $f_Y(y)$.

For any x such that $P(X = x) = f_X(x) > 0$, the conditional pmf of Y given that $X = x$ is the function of y denoted by $f(y | x)$ and defined by

$$f(y | x) = P(Y = y | X = x) = \frac{f(x, y)}{f_X(x)}$$

Correspondingly for any y such that $P(Y = y) > 0$, the conditional pmf of x given that $Y = y$ is

$$f(x | y) = P(X = x | Y = y) = \frac{f(x, y)}{f_Y(y)}$$

Note $f(y|x)$ is a pmf since

- 1) $f(y|x) \geq 0$ for all y because $f(x, y) \geq 0$ for all x, y .

$$2) \sum_{y \in Y} f(y | x) = \frac{\sum_y f(x, y)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1$$

See example 1.24. in notes .pdf.

- The extension to continuous random variables is given next.

Definition 4.2. Let (X, Y) be a continuous random variable with joint pdf $f(x, y)$ and marginal pdf's $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the conditional pdf of Y given that $X = x$ is the function of y denoted by $f(y | x)$ and defined by

$$f(y | x) = \frac{f(x, y)}{f_X(x)}$$

For any y such that $f_Y(y) > 0$, the conditional pdf that $Y = y$ is

$$f(x | y) = \frac{f(x, y)}{f_Y(y)}$$

Computing $\int_{-\infty}^{\infty} f(y | x) dy$, which will be 1, justifies that $f(y | x)$ is a pdf.

Example 4.1. Using the joint pdf from the exercise 3.2

$$\begin{aligned} f(x, y) &= 6xy^2 & 0 < x < 1 \text{ and } 0 < y < 1 \\ \text{and } f(x, y) &= 0 & \text{otherwise} \end{aligned}$$

Also, the marginal pdf for X is

$$f_X(x) = 2x$$

The conditional pdf $f(y | X = 0)$ is undefined, since then, $f_X(x) = 2 * 0 = 0$.

$$f(y | X = 1) = \frac{f(1, y)}{f_X(1)} = \frac{6 * 1 * y^2}{2 * 1} = 3y^2$$

This is proper pdf since

$$\int_{-\infty}^{\infty} f(y | X = 1) dy = \int_0^1 3y^2 dy = \left| \frac{1}{3} 3y^3 \right|_0^1 = 1 - 0 = 1$$

Corresponding cdf is

$$\begin{aligned} F_{Y|X=1}(y | x = 1) &= \int_{-\infty}^y f_{Y|X=1}(t | x = 1) dt \\ &= \int_0^y 3t^2 dt \\ &= \left| t^3 \right|_0^y \\ &= y^3 - 0^3 = y^3 \end{aligned}$$

$$\begin{aligned} P(Y > 0.5 | X = 1) &= 1 - P(Y \leq 0.5 | x = 1) \\ &= 1 - F_{Y|X=1}(0.5 | x = 1) \\ &= 1 - 0.5^3 = \frac{7}{8} \end{aligned}$$

- The conditional expected value of Y given that $X = x$ is the expected value based on the conditional distribution:

$$E(Y | X = x) = E(Y | x) = \sum_{y \in \mathcal{Y}} y f(y | x) \quad \text{if } y \text{ is discrete}$$

$$E(Y | X = x) = E(Y | x) = \int_{-\infty}^{\infty} y f(y | x) dy \quad \text{if } y \text{ is continuous}$$

- For $g(Y)$ we have correspondingly

$$E(g(Y) | x) = \sum_{y \in \mathcal{Y}} g(y) f(y | x) \quad \text{if } y \text{ is discrete}$$

$$E(g(Y) | x) = \int_{-\infty}^{\infty} g(y) f(y | x) dy \quad \text{if } y \text{ is continuous}$$

- $E(Y | X)$ is “The best guess of Y we could make given the knowledge on X ”
- The continuous variance of Y given that $X = x$ is

$$\begin{aligned} Var(Y | x) &= E\left((Y - E(Y | x))^2\right) \\ &= E(Y^2 | x) - [E(Y | x)]^2 \end{aligned}$$

Example 4.2. Using the conditional pdf from example 4.1, we'll get following conditional expected value and variance for Y on condition that $X = 1$

$$\begin{aligned} E(Y | X = 1) &= \int_{-\infty}^{\infty} y f(y | X = 1) dy \\ &= \int_0^1 y 3y^2 dy \\ &= \int_0^1 3y^3 dy = \frac{3}{4} \Big|_0^1 y^4 = \frac{3}{4} (1^4 - 0^4) = \frac{3}{4} \end{aligned}$$

$$\begin{aligned} E(Y^2 | X = 1) &= \int_{-\infty}^{\infty} y^2 f(y | X = 1) dy \\ &= \int_0^1 y^2 3y^2 dy \\ &= \int_0^1 3y^4 dy = \frac{3}{5} \Big|_0^1 y^5 = \frac{3}{5} (1^5 - 0^5) = \frac{3}{5} \end{aligned}$$

$$\begin{aligned} Var(Y | X = 1) &= E(Y^2 | X = 1) - [E(Y | X = 1)]^2 \\ &= \frac{3}{5} - \left(\frac{3}{4}\right)^2 \approx 0.0375 \end{aligned}$$

$$sd(Y | x = 1) = \sqrt{0.0375} = 0.194$$

Theorem 4.1. Let X and Y be any two random variables. Then

$$E(X) = E_Y\left(\underbrace{E_{X|Y}(X | Y)}_{g(y)}\right)$$

and

$$Var(X) = E_Y\left(\underbrace{Var_{X|Y}(X | Y)}_{\text{function of } Y, g(y)}\right) + Var_Y\left(\underbrace{E_{X|Y}(X | Y)}_{\text{function of } Y, g(y)}\right)$$

Illustration See 1.27 in notes.pdf.