Introduction to statistical inference 1

Lauri Mehtätalo
University of Eastern Finland
School of Computing

February 14, 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

2.5 Expected value 16

3 Bivariate random variables **21**

3.1 Conditional distributions 24

3.2 Independence 27

3.3 Covariance and correlation 29

3.4 Bivariate normal distribution 33

3.5 Properties of the bivariate normal distribution 35

4 Matrix and vector notations in statistics **37**

4.1 Random vector 39

5 Multivariate distributions **43**

5.1 Multivariate normal distribution 44

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 using 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

$$A_i \cap A_j = \emptyset, \text{ for all pairs of } i, j$$

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 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, $P(A) \neq 0$ (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. Therefore,

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

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

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) $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ 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, and if $P(B) > 0$

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

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

And the other way around, if $P(A) > 0$

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

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

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 for 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!
- $F_X(x)$ has jumps at $x_i \in \mathcal{X}$. The jumps equal to $P(X = x_i)$.
- $F_X(x) = 0$ 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 does hot have jumps).

Note The X in example 2.1 is 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 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

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

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

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

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

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

- ...which also implies

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

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) $f(x) \ge 0$ for all x

b)

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

or

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

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 the 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 $\frac{\sum_{i=1}^{n} u_i}{n}$ in 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 the 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. $p(Age) = 0.1 + 0.0005 * Age$

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

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

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

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 $g^{-1}(y) : \mathcal{Y} \to \mathcal{X}$ 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 $\mathcal{X}$, 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 $\frac{\mathrm{d}}{\mathrm{d}x}(-ln(x)) = -\frac{1}{x} < 0$ for $x > 0$

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

- 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 $g^{-1}(y)$ by solving x for $y = g(x)$

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

$$-y = ln(x)$$

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

- 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

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{\mathrm{d}(1-e^{-y})}{\mathrm{d}y} = -e^{-y}(-1) = e^{-y}$$

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

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

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

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 $q(x) = q(0.95) = \ldots = x_*$

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 $U \sim Uniform(0,1)$ 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

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

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

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

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

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

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

$$f_X(x) = \tfrac{1}{\lambda} e^{-\frac{x}{\lambda}} \quad \text{where } 0 < x < \infty, \quad \lambda > 0$$

The expected value is

$$\begin{aligned} E(X) &= \textstyle\int_0^{\infty} x f_X(x)\,\mathrm{d}x \\ &= -\textstyle\int_0^{\infty} x \left(-\frac{1}{\lambda} e^{-\frac{x}{\lambda}}\right) \mathrm{d}x \\ &= \Big|_0^{\infty} - xe^{-\frac{x}{\lambda}} + \textstyle\int_0^{\infty} 1e^{-\frac{x}{\lambda}}\,\mathrm{d}x \\ &= -0 - (-0) + \left(-\lambda \textstyle\int_0^{\infty} -\frac{1}{\lambda} e^{-\frac{x}{\lambda}}\,\mathrm{d}x\right) \\ &= -\lambda \Big|_0^{\infty} e^{-\frac{x}{\lambda}} \\ &= -\lambda(0 - 1) = \lambda \end{aligned}$$

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 the distributions 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 in x, 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) $E(g(x)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x$

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

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 (or sd for short)

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

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

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

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

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

or

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

Example 2.16. Assume $X \sim N(10, 2^2)$. Then

$$\begin{aligned} E(X^2) &= 104 \\ (E(X))^2 &= 100 \\ Var(X) &= E(X^2) - (E(X))^2 = 104 - 100 = 4 \\ sd(X) &= \sqrt{Var(X)} = \sqrt{4} = 2 \end{aligned}$$

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

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

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

Expected value of X is the 1$^{\text{st}}$ moment and the variance of X is the 2$^{\text{nd}}$ cental moment of X. These two are called also the 1$^{\text{st}}$ and 2$^{\text{nd}}$ 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 kurtosis.

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

$$f(x,y) = P(X = x, Y = y) \qquad \text{for all } x, y.$$
$$\text{or} \quad f_{X,Y}(x,y) = P(X = x, Y = y) \quad f : \mathbb{R}^2 \to \mathbb{R}$$

Example 3.1. Two dice.

$$X = \text{The sum of two dice}$$
$$Y = |\text{The difference of two dice}|$$

The sample space includes $\{(1,1),(1,2),(1,3),\ldots,(6,6)\}$. There are total of 36 different possibilities, and they are 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
	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
Y	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

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

- 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 $f(x,y)$ ($\mathbb{R}^2 \to \mathbb{R}$) is called the joint pdf of the continuous random variables (X,Y), if for every $A \in \mathbb{R}^2$ we have

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

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

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

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

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

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

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

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

Example 3.2. Define a joint pdf by

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

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?

$$\begin{aligned} \textstyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \textstyle\int_0^1\int_0^1 f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \textstyle\int_0^1\int_0^1 6xy^2\,\mathrm{d}x\,\mathrm{d}y \\ &= \textstyle\int_0^1 \big|_0^1 3x^2y^2\,\mathrm{d}y \\ &= \textstyle\int_0^1 3y^2\,\mathrm{d}y \\ &= \textstyle\big|_0^1 3 * \frac{1}{3}y^3 \\ &= \textstyle\big|_0^1 y^3 \\ &= 1^3 - 0^3 = 1 \end{aligned}$$

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

Lets compute some probabilities:

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

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

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

Example probability for marginal pdf

$$\begin{aligned} P(\tfrac{1}{2} < x < \tfrac{3}{4}) &= \int_{\frac{1}{2}}^{\frac{3}{4}} f_X(x) \,\mathrm{d}x \\ &= \int_{\frac{1}{2}}^{\frac{3}{4}} 2x \,\mathrm{d}x \\ &= \Big|_{\frac{1}{2}}^{\frac{3}{4}} x^2 \\ &= (\tfrac{3}{4})^2 - (\tfrac{1}{2})^2 \\ &= \tfrac{9}{16} - \tfrac{4}{16} = \tfrac{5}{16} \end{aligned}$$

- The joint cdf is defined by

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

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

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s,t) \,\mathrm{d}s \,\mathrm{d}t$$

$$\frac{\mathrm{d}F(x,y)}{\mathrm{d}x \,\mathrm{d}y} = f(x,y)$$

3.1 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 $Y > 80$kg, if $X = 180$cm, than if $X = 150$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 3.3. 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 $f(y \mid x)$ and defined by

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

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

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

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

See example 1.24. in `notes.pdf`.

- The extension to continuous random variables is given next.

Definition 3.4. Let (X, Y) be a continuous random variables 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 $f(y \mid x)$ and defined by

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

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

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

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

Example 3.3. Using the joint pdf form the excercise 3.2

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

Also, the marginal pdf for X is

$$f_X(x) = 2x$$

The conditional pdf $f(y \mid X = 0)$ 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

$$\int_{-\infty}^{\infty} f(y \mid X = 1)\,\mathrm{d}y = \int_0^1 3y^2\,\mathrm{d}y = \Big|_0^1 \frac{1}{3}3y^3 = 1 - 0 = 1$$

Corresponding cdf is

$$\begin{aligned} F_{Y|X=1}(y \mid x = 1) &= \textstyle\int_{-\infty}^{y} f_{Y|X=1}(t \mid x = 1)\,\mathrm{d}t \\ &= \textstyle\int_0^y 3t^2\,\mathrm{d}t \\ &= \textstyle\Big|_0^y t^3 \\ &= y^3 - 0^3 = y^3 \end{aligned}$$

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

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

$$\begin{array}{ll} E(Y \mid X = x) = E(Y \mid x) = \sum_{y \in \mathcal{Y}} y f(y \mid x) & \text{if } y \text{ is discrete} \\ E(Y \mid X = x) = E(Y \mid x) = \int_{-\infty}^{\infty} y f(y \mid x)\,\mathrm{d}y & \text{if } y \text{ is continuous} \end{array}$$

- For $g(Y)$ we have correspondingly

$$\begin{array}{ll} E(g(Y) \mid x) = \sum_{y \in \mathcal{Y}} g(y) f(y \mid x) & \text{if } y \text{ is discrete} \\ E(g(Y) \mid x) = \int_{-\infty}^{\infty} g(y) f(y \mid x)\,\mathrm{d}y & \text{if } y \text{ is continuous} \end{array}$$

- $E(Y \mid X)$ is "The best guess of Y we could make given the knowledge on X"

- The continuous variance of Y given that $X = x$ is

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

Example 3.4. Using the conditional pdf from example 3.3, well get following conditional expected value and variance for Y on condition that $X = 1$

$$\begin{aligned} E(Y \mid X = 1) &= \int_{-\infty}^{\infty} y f(y \mid X = 1)\,\mathrm{d}y \\ &= \int_0^1 y 3y^2\,\mathrm{d}y \\ &= \int_0^1 3y^3\,\mathrm{d}y = \tfrac{3}{4}\big|_0^1 y^4 = \tfrac{3}{4}(1^4 - 0^4) = \tfrac{3}{4} \end{aligned}$$

$$\begin{aligned} E(Y^2 \mid X = 1) &= \int_{-\infty}^{\infty} y^2 f(y \mid X = 1)\,\mathrm{d}y \\ &= \int_0^1 y^2 3y^2\,\mathrm{d}y \\ &= \int_0^1 3y^4\,\mathrm{d}y = \tfrac{3}{5}\big|_0^1 y^5 = \tfrac{3}{5}(1^5 - 0^5) = \tfrac{3}{5} \end{aligned}$$

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

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

Theorem 3.2. Let X and Y be any two random variables. Then

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

and

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

Illustration See 1.27 in `notes.pdf`.

3.2 Independence

Definition 3.5. Let (X, Y) be a bivariate random variable with joint pdf or pmf $f(x, y)$, and marginal pmf's/pdf's $f_X(x)$ and $f_Y(y)$.

X and Y are said to be independent (riippumattomia) random variables if for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$

$$f(x, y) = f_X(x) f_Y(y)$$

If X and Y are independent, it follows that

$$\begin{aligned} f(y \mid x) &= \tfrac{f(x,y)}{f_X(x)} && \text{(by definition of conditional pmf/pdf)} \\ &= \tfrac{f_X(x) f_Y(y)}{f_X(x)} && \text{(by definition of independence)} \\ &= f_Y(y) \end{aligned}$$

regardless of the value of x.

- For independent random variables, the marginal distribution of Y is equal to the conditional distribution of $Y \mid x$ for all values of X.
- Same holds naturally for the marginal distribution of X and all conditional distributions of $X \mid Y$.
- For $X \mid Y = y$, if for some choice of the y the pdf differs, then X and Y are not independent
- For independent random variables, the knowledge of the realized value of X does not provide any information on Y compared to the situation where the value of X is not known.

Example 3.5. Checking for independence

Assume that $f(x, y)$ is defined as follows:

$$f(10,1) = f(20,1) = f(20,2) = \tfrac{1}{10}$$
$$f(10,2) = f(10,3) = \tfrac{1}{5}$$
$$f(20,3) = \tfrac{3}{10}$$

Joint and marginal probabilities can be seen on table 3.3. Consider events $A: \ X = 10$ and $B: \ Y = 1$.

$$P(A \cap B) = \frac{1}{10}$$
$$P(A)P(B) = \frac{5}{10} * \frac{2}{10} = \frac{1}{10}$$

Since $P(A \cap B) = P(A)P(B)$, events A and B are independent. Lets then consider events $C: \ X = 10$ and $D: \ Y = 2$.

$$P(C \cap D) = \frac{1}{5} = \frac{4}{20}$$
$$P(C)P(D) = \frac{1}{2} * \frac{3}{10} = \frac{3}{20}$$

Since $P(C \cap D) \neq P(C)P(D)$, events C and D are not independent, which also means that random variables X and Y are not independent.

Table 3.3: Probabilities for bivariate pmf $f(x, y)$ of example 3.5

		Y			
		1	2	3	
X	10	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{5}{10} = \frac{1}{2}$
	20	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{5}{10} = \frac{1}{2}$
		$\frac{2}{10}$	$\frac{3}{10}$	$\frac{5}{10}$	

Theorem 3.3. Let X and Y be independent random variables.

a) For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

that is, events $\{X \in A\}$ and $\{Y \in B\}$ are independent.

b) Let $g(x)$ be a function of x only and $h(y)$ is a function of y only, then

$$E(g(X)h(Y)) = E(g(X))E(h(Y)) \ (*)$$

and naturally also

$$E(XY) = E(X)E(Y)$$

which is a special case of $(*)$.

3.3 Covariance and correlation

- Measures the strength of the relationship between X and Y. Especially, they measure the linear association between the two random variables.

Notations

$$\begin{aligned} E(X) &= \mu_X \qquad E(Y) = \mu_Y \\ Var(X) &= \sigma_X^2 \quad \rightarrow sd(X) = \sigma_X \\ Var(Y) &= \sigma_Y^2 \quad \rightarrow sd(Y) = \sigma_Y \end{aligned}$$

We always assume that $0 < \sigma_X^2 < \infty$ and $0 < \sigma_Y^2 < \infty$.

Definition 3.6. The covariance of X and Y is the number defined by

$$cov(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

The correlation of X and Y is the number defined by

$$\rho_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

It can also be called correlation coefficient.

- If large values of X tend to be observed together with large values of Y and small values of X with small values of Y, then the product $(X - \mu_X)(Y - \mu_Y)$ tends to get high values, meaning $cov(X, Y) = E((X - \mu_X)(Y - \mu_Y))$ is high and positive.

- If large values of X are associated with low values of Y, then the product $(X - \mu_X)(Y - \mu_Y)$ tends to be negative with high absolute values, meaning covariance is negative and has high absolute value.
- Same holds for correlation, because $\sigma_X \sigma_Y$ is always positive and constant.
- The covariance can get any values within $(-\infty, \infty)$, but the correlation is restricted to interval $[-1, 1]$.
- Values $\rho_{XY} = -1$ and $\rho_{XY} = 1$ of correlation indicate a perfect linear relationship between X and Y, while value $\rho_{XY} = 0$ of correlation indicates no linear relationship between X and Y

Theorem 3.4. For any random variables X and Y

$$cov(X, Y) = E(XY) - \mu_X \mu_Y$$

Proof

$$\begin{aligned}
&E((X - \mu_X)(Y - \mu_Y)) \\
&= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\
&= E(XY) - E(\mu_X Y) - E(\mu_Y X) + E(\mu_X \mu_Y) \quad \mu_X \text{ and } \mu_Y \text{ are fixed} \\
&= E(XY) - \mu_X \underbrace{\mathrm{E}(Y)}_{\mu_Y} - \mu_Y \underbrace{E(X)}_{\mu_X} + \mu_X \mu_Y \\
&= E(XY) - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\
&= E(XY) - \mu_X \mu_Y
\end{aligned}$$

Note If $\mu_X = 0$ or $\mu_Y = 0$, then $cov(X, Y) = E(XY)$

Note The covariance of random variable with itself is

$$\begin{aligned}
cov(X, X) &= E(XX) - E(X)E(X) \\
&= E(X^2) - (E(X))^2 = Var(X)
\end{aligned}$$

In that case

$$cor(X, X) = \frac{cov(X, X)}{\sigma_X \sigma_X} = \frac{Var(X)}{Var(X)} = 1$$

Example 3.6. Let the joint pdf of (X, Y) be

$$f(x, y) = \begin{cases} 1 & 0 < x < 1,\ x < y < x + 1 \\ 0 & \text{otherwise} \end{cases}$$

The volume under curve $f(x, y)$ is $1 \times 1 \times 1 = 1$ The marginal distribution of X is $Uniform(0, 1)$. The marginal distribution of Y is

$$f_Y(y) = \begin{cases} y & \text{if } 0 < y \leq 1 \\ 2 - y & \text{if } 1 < y < 2 \end{cases}$$

$$E(Y) = \int_0^2 f_Y(y)dy = \ldots = 1$$

$$E(X) = \frac{0+1}{2} = \frac{1}{2} \ (X \sim Unif(0,1))$$

$$\begin{aligned} E(XY) = \int_0^1 \int_x^{x+1} xy1 \,\mathrm{d}y \,\mathrm{d}x &= \int_0^1 \Big|_x^{x+1} \tfrac{1}{2}xy^2 \,\mathrm{d}x \\ &= \int_0^1 \tfrac{1}{2}x(x+1)^2 - \tfrac{1}{2}xx^3 \,\mathrm{d}x \\ &= \ldots = \int_0^1 x^2 + \tfrac{1}{2}x \,\mathrm{d}x \\ &= \ldots = \tfrac{7}{12} \end{aligned}$$

$$\begin{aligned} cov(X,Y) &= E(XY) - E(X)E(Y) \\ &= \tfrac{7}{12} - \tfrac{1}{2} * 1 \\ &= \tfrac{7}{12} - \tfrac{6}{12} = \tfrac{1}{12} \end{aligned}$$

The correlation is

$$cor(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{1/12}{\sqrt{1/12}\sqrt{1/2}} = \frac{1}{\sqrt{2}}$$

Theorem 3.5. If X and Y are independent, then $cov(X,Y) = 0$ and $\rho_{XY} = 0$.

Proof Under independence, theorem 3.3 holds, meaning $E(XY) = E(X)E(Y)$,

$$\begin{aligned} cov(X,Y) &= E(XY) - E(X)E(Y) \\ &= E(X)E(Y) - E(X)E(Y) = 0 \end{aligned}$$

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{0}{\sigma_X \sigma_Y} = 0$$

Theorem 3.6. Calculating using covariance

If X, Y and U are random variables, and a and b are constants,

$$\begin{aligned} Var(X+Y) &= Var(X) + Var(Y) + 2 * cov(X,Y) \\ Var(X-Y) &= Var(X) + Var(Y) - 2 * cov(X,Y) \\ Var(aX+bY) &= a^2 Var(X) + b^2 Var(Y) + 2ab * cov(X,Y) \\ cov(X+Y,U) &= cov(X,U) + cov(Y,U) \end{aligned}$$

If X and Y are independent, then

$$Var(X+Y) = Var(X) + Var(Y)$$

Example 3.7. The variance of sample mean

Let $Y = \frac{1}{n}\sum_{i=1}^n X_i$. We already know from exercises 4, that $E(Y) = \frac{1}{n}\sum_{i=1}^n \mu_i$, and if $\mu_i = \mu$ for all $i = 1, \ldots, n$, $E(Y) = \mu$

If $X_1, \ldots, X_n$ are independent or at least uncorrelated,

$$\begin{aligned} Var(Y) &= Var(\tfrac{1}{n}\sum X_i) \\ &= \tfrac{1}{n^2} Var(\sum X_i) \quad \text{(uncorrelated random variables)} \\ &= \tfrac{1}{n^2}\sum Var(X_i) \\ &= \tfrac{1}{n^2}\sum \sigma_i^2 \end{aligned}$$

If $\sigma_i^2 = \sigma^2$ for all $i = 1, \ldots, n$ (the random variables have common, constant variance)

$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$$sd(Y) = \sqrt{Var(Y)} = \frac{\sigma}{\sqrt{n}}$$

The well known standard error of the mean of an independent sample.

Note If X_i are independent and identically distributed (i.i.d) and $X_i \sim Uniform(l, u)$, we have $Var(X_i) = \frac{1}{12}(u-l)^2$ and

$$Var(Y) = \frac{1}{n} Var(X_i) = \frac{1}{n}\frac{1}{12}(u-l)^2 = \frac{1}{12n}(u-l)^2$$

Recall R example `unifmean.R`, where $X_1 \sim Unif(0, 10)$

$$\begin{aligned} Var(X_1) &= \tfrac{1}{12}(u-l)^2 = \tfrac{1}{12}10^2 = \tfrac{100}{12} \\ Var(Y) &= \tfrac{100/12}{n} \\ sd(Y) &= \tfrac{10}{\sqrt{12n}} \end{aligned}$$

Example 3.8. Spatial data

Let random variables X_1, X_2 and X_3 be observed at fixed locations u_1, u_2 and u_3:

$$u_1 = (0, 1)$$
$$u_2 = (0, 0)$$
$$u_3 = (2, 0)$$

Assume that the correlation among X_i is specified by the exponential spatial correlation

$$\rho_{X_i X_j} = e^{\frac{-s_{ij}}{d}}$$

where s_{ij} is the spatial distance between locations u_i and u_j, and d specifies the strength of the spatial correlation. Assume that d is known to be $d = 2$. Then, the correlations for all the pairs can be seen in the table 3.4.

Table 3.4: Correlations between the random variable pairs of 3.8

Pair	Distance	Correlation
1,2	1	0.61
1,3	$\sqrt{5}$	0.3
2,3	2	0.37
		$\sum \approx 1.30$

Let $Y = \frac{1}{3}(X_1 + X_2 + X_3)$, $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ and $E(Y) = \mu$. Then, the variance of Y can be computed as

$$\begin{aligned}
Var(Y) &= \tfrac{1}{n^2} Var(X_1 + X_2 + X_3) \\
&= \tfrac{1}{n^2} \left[Var(X_1 + X_2) + Var(X_3) + 2cov(X_1 + X_2, X_3) \right] \\
&= \tfrac{1}{n^2} \left[Var(X_1) + Var(X_2) + 2cov(X_1, X_2) + Var(X_3) + 2cov(X_1, X_3) + 2cov(X_2, X_3) \right] \\
&= \tfrac{1}{n^2} \left[\underbrace{\sum Var(X_i))}_{3\sigma^2} + 2 \sum_{i<j} \underbrace{cov(X_i, X_j)}_{\sigma\sigma\rho_{X_i X_j}} \right] \\
&= \tfrac{1}{3^2} * 3\sigma^2 + \tfrac{2\sigma^2}{3^2} (cor(X_1, X_2) + cor(X_1, X_3) + cor(X_2, X_3)) \\
&= \sigma^2 (\tfrac{1}{3} + \tfrac{2}{9} * 1.30) = \underbrace{0.62}_{<1} \sigma^2
\end{aligned}$$

If the random variables would be uncorrelated and $n = 3$, we would have had

$$Var(Y) = \frac{1}{3}\sigma^2 \approx 0.33\sigma^2$$

If n were 2, we would have had

$$Var(Y) = \frac{1}{2}\sigma^2 = 0.5\sigma^2$$

3.4 Bivariate normal distribution

- If we have dependent data of size n, we cannot think it as n (independent) replicates of a random variable X, but as a single replicate of a n-variate random vector.
  - We need to define n-variate distribution
  - We need to extend the ideas we have presented for bivariate case to n-variate case.
- The most commonly used n-variate distribution is multivariate normal distribution.

- To understand its properties, we first take a look into bivariate normal distribution.

Definition 3.7. Let $-\infty < \mu_X < \infty, -\infty < \mu_Y < \infty, \sigma_X > 0, \sigma_Y > 0, -1 < \rho < 1$ be real numbers.

The bivariate normal pdf with (marginal) means μ_X and μ_Y, (marginal) variances σ_X^2 and σ_Y^2 and correlation ρ is the bivariate pdf given by

$$f(x,y) = \left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right)^{-1} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$

Example 3.9. $\mu_X = 10$, $\mu_Y = 5$, $\sigma_X = 5$, $\sigma_Y = 2$ and $\rho = 0$.

Because in this case $\rho = 0$, the marginal mean of $Y \mid x$ is equal to $E(Y)$, and the conditional variance of $Y \mid x$ is equal to $Var(Y \mid x)$.

- Because both marginal and conditional distributions are univariate normals, the marginal distributions are therefore the same as the conditional ones, meaning the two random variables are independent.
- The equation looks complex, but it can be presented nicely in matrix form that covers implicitly the extension to multivariate case.
- Also called bivariate Gaussian density.
- Basis for the wide class of statistical models and limiting distribution of parameter estimates in many cases as the sample size increases.
  - Normality of e_i in the linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ leads to bivariate normality of $\hat{\beta}_0$ and $\hat{\beta}_1$ in small samples.
  - Asymptotic distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ in the linear regression regardless of the distribution of e_i, as the sample size $\to \infty$. The convergence is quite fast, if the distribution of e_i is not too different from the normal distribution, and the variance $Var(e_i)$ is constant over the range of x.
  - It is a limit distribution (as $n \to \infty$) of parameter estimates in a wide class of statistical models: multiple linear regression, generalized linear models, mixed-effect models ect. In general, all models where parameters are estimated using maximum likelihood.
  - Starting assumption in factor analysis, and model based geostatistics, Gaussian mixture models, model-based time-series analysis ect.
- The bivariate normal density is an intermediate step to the more general multivariate normal distribution!

3.5 Properties of the bivariate normal distribution

- The marginal distribution of X is $N(\mu_X, \sigma_X^2)$.
- The marginal distribution of Y is $N(\mu_Y, \sigma_Y^2)$.
- The correlation between X and Y is $\rho_{XY} = \rho$.
- For any constant a and b, the distribution of $aX + bY$ is

$$N\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y\right)$$

  - Recall:

$$E(aX + bY) = aE(X) + bE(Y)$$
$$Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abcov(X, Y)$$

    for X and Y regardless of their joint distribution, meaning it also holds for the normal case.

- The conditional distributions of $X \mid y$ and $Y \mid x$ are also normal. The expected value and variance

$$E(Y \mid x) = \mu_Y + \rho\tfrac{\sigma_Y}{\sigma_X}(x - \mu_X)$$
$$Var(Y \mid x) = \sigma_Y^2(1 - \rho^2)$$

  - Recall the example 3.9 where $\rho = 0$. In that case

$$E(Y \mid x) = \mu_Y + 0 * \tfrac{\sigma_Y}{\sigma_X}(x - \mu_X) = \mu_Y$$
$$Var(Y \mid x) = \sigma_Y^2(1 - 0^2) = \sigma_Y^2$$

Note The conditional expected value $E(Y \mid x)$ is related to x linearly

$$\begin{aligned} E(Y \mid x) &= \mu_Y + \tfrac{\rho\sigma_Y}{\sigma_X}x - \tfrac{\rho\sigma_Y}{\sigma_X}\mu_X \\ &= \underbrace{\mu_Y - \frac{\rho\sigma_Y}{\sigma_X}\mu_X}_{\text{Constant a}} + \underbrace{\frac{\rho\sigma_Y}{\sigma_X}}_{\text{Constant b}} x \\ &= a + bx \end{aligned}$$

Note Conditional variance $Var(Y \mid x) = \sigma_Y^2(1-\rho^2)$ does not include x, meaning it's constant over the range of x

- Dependency of Y on x occurs only through the mean $E(Y \mid x)$, which is linear in x.
- Correlation, which means linear association between X and Y, completely describes the dependency of X and Y.
- Correlation 0 implies also independence under bivariate normality.

Chapter 4

Matrix and vector notations in statistics

Example 4.1. Single predictor regression

Consider model

$$y_i = \beta_0 + \beta_1 x_i + e_i \text{ , where } i = 1, \ldots, n$$

Lets define the following matrices and vectors.

$$\boldsymbol{y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X}_{n\times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta}_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{e}_{n\times 1} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

The model for whole data can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n\times 2} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_{2\times 1} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 * \beta_0 + x_1\beta_1 \\ 1 * \beta_0 + x_2\beta_1 \\ 1 * \beta_0 + x_3\beta_1 \\ \vdots \\ 1 * \beta_0 + x_n\beta_1 \end{bmatrix}_{n\times 1} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} \beta_0 + x_1\beta_1 + e_1 \\ \beta_0 + x_2\beta_1 + e_2 \\ \beta_0 + x_3\beta_1 + e_3 \\ \vdots \\ \beta_0 + x_n\beta_1 + e_n \end{bmatrix}_{n\times 1}$$

Note This extends easily to multiple regression, e.g. for the model with two predictors x_i and z_i

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \text{ , where } i = 1, \ldots, n$$

Then,

$$\boldsymbol{X}_{n\times 3} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{bmatrix}, \quad \boldsymbol{\beta}_{3\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix},$$

4.1 Random vector

Consider random variables $X_1, X_2, \ldots, X_n$. A random vector is the vector of univariate random variables.

$$\boldsymbol{x} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \quad E(\boldsymbol{x}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix}$$

The variance of a random vector is $n \times n$ symmetric matrix called variance-covariance matrix.

$$Var(\boldsymbol{x})_{n\times n} = \begin{bmatrix} Var(X_1) & cov(X_1, X_2) & cov(X_1, X_3) & \ldots & cov(X_1, X_n) \\ cov(X_2, X_1) & Var(X_2) & cov(X_2, X_3) & \ldots & cov(X_2, X_n) \\ cov(X_3, X_1) & cov(X_3, X_2) & Var(X_3) & \ldots & cov(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(X_n, X_1) & cov(X_n, X_2) & cov(X_n, X_3) & \ldots & Var(X_n) \end{bmatrix}$$

Note Because $cov(X_1, X_2) = cov(X_2, X_1)$, the variance-covariance matrix is symmetric, meaning $[Var(\boldsymbol{x})]' = Var(\boldsymbol{x})$

Note Because $Var(X) = cov(X, X)$,

$$Var(\boldsymbol{x})_{n\times n} = \begin{bmatrix} cov(X_1, X_1) & cov(X_1, X_2) & cov(X_1, X_3) & \ldots & cov(X_1, X_n) \\ cov(X_2, X_1) & cov(X_2, X_2) & cov(X_2, X_3) & \ldots & cov(X_2, X_n) \\ cov(X_3, X_1) & cov(X_3, X_2) & cov(X_3, X_3) & \ldots & cov(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(X_n, X_1) & cov(X_n, X_2) & cov(X_n, X_3) & \ldots & cov(X_n, X_n) \end{bmatrix}$$

Consider two random vectors $\boldsymbol{x}_{1\times p}$ and $\boldsymbol{y}_{1\times q}$. The covariance between them is $p \times q$ matrix (not a square matrix unless $p = q$)

$$cov(\boldsymbol{x}, \boldsymbol{y}')_{p\times q} = \begin{bmatrix} cov(X_1, Y_1) & cov(X_1, Y_2) & \ldots & cov(X_1, Y_q) \\ cov(X_2, Y_1) & cov(X_2, Y_2) & \ldots & cov(X_2, Y_q) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_p, X_1) & cov(X_p, X_2) & \ldots & cov(X_n, X_q) \end{bmatrix}$$

Example 4.2. Recall example 3.8 on spatial data at 3 locations: $(1,0)$, $(0,0)$ and $(0,2)$, and $\rho_{X_iX_j} = e^{\frac{-s_{ij}}{2}}$

On table 3.4, we calculated the correlations $\rho_{X_iX_j}$ between the pairs (i,j). We also assumed, that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for all i.

If we place the random variables into random vector $\boldsymbol{x}$

$$\boldsymbol{x} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

then, the expected value and variance of random vector $\boldsymbol{x}$ are

$$E(\boldsymbol{x}) = \begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix} \quad Var(\boldsymbol{x})_{3\times 3} = \sigma^2 \begin{bmatrix} 1 & 0.61 & 0.3 \\ 0.61 & 1 & 0.37 \\ 0.3 & 0.37 & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & \sigma^2 * 0.61 & \sigma^2 * 0.3 \\ \sigma^2 * 0.61 & \sigma^2 & \sigma^2 * 0.37 \\ \sigma^2 * 0.3 & \sigma^2 * 0.37 & \sigma^2 \end{bmatrix}$$

Covariance terms inside the variance-covariance matrix are calculated by using following

$$cov(X_i, X_j) = \rho_{X_iX_j} \underbrace{sd(X_i)}_{\sigma} \underbrace{sd(X_j)}_{\sigma} = \rho_{X_iX_j}\sigma^2$$

Example 4.3. Ordinary least squares (OLS) estimators

Recall the example 4.1 about single predictor regression.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

If we assume that

$$Var(\boldsymbol{e}) = Var(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_{n\times n} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

meaning all components of $\boldsymbol{e}$ have a common variance σ^2 and $cov(e_i, e_j) = 0$ for all i, j where $i \neq j$. Given the realized values of $\boldsymbol{y}$, and the known values of $\boldsymbol{X}$, the unknown values of parameters $\boldsymbol{\beta}$ can estimated by minimizing the sum of squared residual errors $\boldsymbol{e}'\boldsymbol{e} = \sum e_i^2$ with respect to $\boldsymbol{\beta}$ to get ordinary least squares (OLS) estimator (PNS, pienimmän neliösumman estimaattori).

$$\widehat{\boldsymbol{\beta}_{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Estimator has the variance $Var\left(\widehat{\boldsymbol{\beta}_{OLS}}\right) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

Example 4.4. See example about OLS in R-script `ols.R`

Observations	$\boldsymbol{x}' = \begin{bmatrix} 1 & 2 & \dots & 10 \end{bmatrix}$
and true model is	$y_i = \beta_0 + \beta_1 x_i + e_i = 5 + 10 * x_i + e_i$ where $e_i \sim N(0, 10^2)$

Chapter 5

Multivariate distributions

- We extend the bivariate random variables to the n-dimensional case.

  Note In this section, we do not make the difference between random variable and its realizations when speaking about random vectors. That is, $\boldsymbol{y}$ may be a random variable even though it is not capitalized.

- The joint pdf for a continuous random variable $\boldsymbol{x}_{n\times 1}$ is defined by $f(\boldsymbol{x})$ in the integral

$$\begin{aligned} P(\boldsymbol{x} \in A) &= \int \dots \int_A f(\boldsymbol{x}\, \mathrm{d}\boldsymbol{x}) \\ &= \int \dots \int_A f(x_1, x_2, \dots, x_n)\, \mathrm{d}x_1 \dots\, \mathrm{d}x_n \end{aligned}$$

  where A is n-dimensional set.

- For a discrete random variables, this is defined correspondingly by replacing integrals with sums.

- The expected value of $g(\boldsymbol{x})$

$$E\left[g(\boldsymbol{x})\right] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(\boldsymbol{x}) f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

  and

$$E\left[g(\boldsymbol{x})\right] = \sum_{\boldsymbol{x}\in\, \mathbb{R}^n} g(\boldsymbol{x}) f(\boldsymbol{x})$$

- The marginal pmf/pdf for any set of elements of $\boldsymbol{x}$ is computed by summing or integrating $f(\boldsymbol{x})$ over all other elements of $\boldsymbol{x}$.

$$f(x_1, \dots, x_k) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n)\, \mathrm{d}x_{k+1} \dots\, \mathrm{d}x_n$$

  or

$$f(x_1, \dots, x_k) = \sum_{x_{k+1}, \dots, x_n \in\, \mathbb{R}^{n-k}} \underbrace{f(x_1, x_2, \dots, x_n)}_{f(\boldsymbol{x})}$$

- The conditional pdf/pmf of $x_{k+1}, \ldots, x_n$ conditional on $x_1, \ldots, x_k$ is

$$f(x_{k+1}, \ldots, x_n \mid x_1, \ldots, x_k) = \frac{f(x_1, \ldots, x_n)}{f(x_1, \ldots, x_k)}$$

5.1 Multivariate normal distribution

- Consider a random variable $\boldsymbol{x}$ of length n. It is said to follow the multivariate normal distribution, if the density is

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

  where $\boldsymbol{\mu} = [E(\boldsymbol{x})]_{n \times 1}$ and $\boldsymbol{\Sigma} = [Var(\boldsymbol{x})]_{n \times n}$

- If the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ were written using scalar notations, they would include
  - n expected values, one for each element of $\boldsymbol{x}$ (in $\boldsymbol{\mu}$).
  - n variances (from the diagonal of $\boldsymbol{\Sigma}$), one for each element of $\boldsymbol{x}$.
  - $(n-1)+(n-2)+\ldots+1 = \frac{n(n-1)}{2}$ covariances, one for each pair (x_i, x_j)

- In practice, we make some restrictions to the parameters

E.g.	$E(X_i) = \mu$	for all i
	$Var(X_i) = \sigma^2$	for all i
or	$Var(X_i) = \sigma^2 [E(X)]^{\delta}$	where δ is some power parameter
and	$cov(X_i, X_j) = \sigma^2 e^{\frac{-S_{ij}}{\delta}}$	in spatial data
or	$cov(X_i, X_j) = \rho$	if X_i and X_j belong to a same group in grouped data
or	$cov(X_i, X_j) = \sigma^2 \phi^{\lvert i-j \rvert}$	in time series data ($AR(1)$)

- Lets consider the following partition of the normally distributed random vector $\boldsymbol{x}$:

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \\ X_{k+1} \\ \vdots \\ X_n \end{bmatrix} \quad \text{where } \boldsymbol{x}_1 = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \text{ and } \boldsymbol{x}_2 = \begin{bmatrix} X_{k+1} \\ \vdots \\ X_n \end{bmatrix}$$

Now

$$E(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \begin{bmatrix} E(\boldsymbol{x}_1) \\ E(\boldsymbol{x}_2) \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} Var(\boldsymbol{x}_1)_{k\times k} & cov(\boldsymbol{x}_1, \boldsymbol{x}_2')_{k\times(n-k)} \\ cov(\boldsymbol{x}_2, \boldsymbol{x}_1')_{(n-k)\times k} & Var(\boldsymbol{x}_2)_{(n-k)\times(n-k)} \end{bmatrix}_{n\times n} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_2 \end{bmatrix}$$

The conditional expected value and variance of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2$ are

$$E(\boldsymbol{x}_1 \mid \boldsymbol{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$$
$$Var(\boldsymbol{x}_1 \mid \boldsymbol{x}_2) = \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_{21}$$

Example 5.1. Recall example 4.2 on spatial data at 3 locations: $u_1 = (1, 0)$, $u_2 = (0, 0)$ and $u_3 = (0, 2)$, and $\rho_{X_iX_j} = e^{\frac{-s_{ij}}{2}}$.

Consider additional random variables X_4 and X_5 at locations $u_4 = (1, 2)$ and $u_5 = (3, 1)$.

$$\boldsymbol{x} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix}$$

Lets assume that

$E(X_i) = 10$ for all i

$Var(X_i) = \sigma^2 = 5$ for all i

then

$Cov(X_i, X_j) = \sigma^2 e^{\frac{-s_{ij}}{2}} = 5 * e^{\frac{-s_{ij}}{2}}$

If we have following partitions

$$\boldsymbol{y} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \quad \boldsymbol{z} = \begin{bmatrix} X_4 \\ X_5 \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix}$$

and $\boldsymbol{z}$ has been observed to be

$$\boldsymbol{z} = \begin{bmatrix} X_4 \\ X_5 \end{bmatrix} = \begin{bmatrix} 9 \\ 13 \end{bmatrix}$$

Variance-covariance matrix of $\boldsymbol{x}$ would be formed as

$$Var(\boldsymbol{x}) = \begin{bmatrix} Var(\boldsymbol{y}) & \boldsymbol{c}ov(\boldsymbol{y}, \boldsymbol{z}') \\ cov(\boldsymbol{z}, \boldsymbol{y}') & Var(\boldsymbol{z}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_2 \end{bmatrix}$$

Then, we could calculate the conditional expected value and variance for $\boldsymbol{y}$ with condition $\boldsymbol{z}$ as (see R-script `MVNormalMean.R` to calculate the values)

$$\begin{aligned} E(\boldsymbol{y} \mid \boldsymbol{z}) &= E(\boldsymbol{y}) + cov(\boldsymbol{y}, \boldsymbol{z}') \left[Var(\boldsymbol{z})\right]^{-1} \left[\boldsymbol{z} - E(\boldsymbol{z})\right] \\ &= \ldots = \begin{bmatrix} 10.4 \\ 10.0 \\ 9.4 \end{bmatrix} \\ Var(\boldsymbol{y} \mid \boldsymbol{z}) &= Var(\boldsymbol{y}) + cov(\boldsymbol{y}, \boldsymbol{z}') \left[Var(\boldsymbol{z})\right]^{-1} cov(\boldsymbol{z}, \boldsymbol{y}') \\ &= \ldots = \begin{bmatrix} 4.1 & 2.3 & 0.5 \\ 2.3 & 4.4 & 0.8 \\ 0.5 & 0.8 & 3.16 \end{bmatrix} \end{aligned}$$