Principles of Statistical Modelling
Part II

Adalbert Wilhelm

Fall 2015

1 Some practical distributions

By now, I hope you all got an understanding of the fundamental principles of probability, where it comes from, how we formalise and quantify it.

We will now continue with a closer look at random variables and their distributions. What you should recall is the fundamental definition of a random variable

$$X : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{F})$$

and its distribution

$$\begin{aligned} P_X : \mathcal{F} &\to [0,1] \\ P_X(A) &= P(X \in A), \qquad \forall A \in \mathcal{F}. \end{aligned}$$

1.1 Probability vectors and probability mass functions

We have already seen that a discrete distribution can be represented by its probability vector. So, for a discrete random variable $X : \Omega \to S = \{s_1, \ldots, s_n\}$ its distribution can be fully characterised by the probability vector

$$\begin{aligned} p(s_i) &= P_X(\{s_i\}) \\ &= P(\{\omega \in \Omega : X(\omega) = s_i\}) \end{aligned}$$

Reversely: For a given function $X : \Omega \to S$ with finite or countably infinite domain S, any function $p : S \to [0,1]$ fulfilling the two properties

$$\begin{aligned} p(s) &\geq 0 \\ \sum_{s \in S} p(s) &= 1 \end{aligned}$$

defines a distribution of X via

$$P_X(A) = \sum_{s \in A} P_X(\{s\}) = \sum_{s \in A} p(s)$$

.

Such functions $p : S \to [0,1]$ fulfilling the two properties

$$\begin{aligned} p(s) &\geq 0 \\ \sum_{s \in S} p(s) &= 1 \end{aligned}$$

are called **probability mass functions** (PMF).

The discussion here points to the fact that a function $X : \Omega \to S$ can have more than one distribution. However, the random variable $X : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{F}, P_X)$ with its distribution is unique since the random variable requires for its full specification also the distribution associated. In that respect a mapping $X : \Omega \to S$ only turns into a random variable by attaching the corresponding distribution to it.

Example 1 *Bernoulli Distribution: Assume we have an experiment with only two possible outcomes for each trial*

success – failure
survival – death
female – male
U.S. citizen – no U.S. citizen
retired – not retired
married – unmarried
correct – incorrect
pass – fail
0 – 1

All of these experiments can be modelled by a random variable that maps the Grundgesamtheit into the set $\{s_1, s_2\}$.

Definition 1 *The distribution of the random variable* $X : \Omega \to \{s_1, s_2\}$ *with probability mass function*

$$p(s_i) = \begin{cases} 1-p & \text{for } i = 1 \\ p & \text{for } i = 2 \end{cases}$$

is called Bernoulli Distribution with success parameter p.

Terminology gets sometimes a bit bizarre when statisticians talk a bout Bernoulli Distributions and the success parameter even in cases where one sees anything else but success.

For random variables that map the universe (Grundgesamtheit) to an ordered discrete set, e.g. a countable subset of the real numbers, there is a second function that can be used to characterise the distribution of X, the so-called cumulative distribution function (CDF). For sake of simplicity, we restrict our description to random variables mapping to a subset of the real line. For these cases, the CDF accumulates the probability of individual points and assigns probabilities to intervals and hence is defined for all real numbers..

Definition 2 *Let* $X : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{F})$, *with* $S \subseteq \mathbb{R}$. *For each real number* t, *the function* $F(t) = P(X \leq t)$ *is called the cumulative distribution function, i.e.*

$$F(t) = P(X \leq t) = \sum_{x \leq t} p(x), \qquad \forall t \in \mathbb{R}.$$

Any CDF must show the following features:

1. $0 \leq F(t) \leq 1 \qquad \forall t \in \mathbb{R}$
2. $F(-\infty) = 0$ short for: $\lim_{t \to -\infty} F(t) = 0$
3. $F(\infty) = 1$ short for: $\lim_{t \to \infty} F(t) = 1$
4. F is monotone increasing, i.e. $\forall t, u \in \mathbb{R}$ with $t < u$, we have: $F(t) \leq F(u)$
5. F is right continuous, i.e. $F(t+) = F(t) \quad \forall t \in \mathbb{R}$, i.e. $\forall \{t_n\}_{n \in \mathbb{N}} \searrow t$ we have: $\lim_{n \to \infty} F(t_n) = F(t)$

PMF and CDF carry same information, but they are organised differently. For discrete random variables it is usually more efficient to write down the PMF and not the CDF.

Example 2 *Given a function* $X : \Omega \to S = \{0, 1, 2, 3\}$ *and a function* $p : S \to [0, 1]$ *defined by*

$$p(s) = \begin{cases} \frac{1}{4} & \text{for } s = 0 \\ \frac{1}{2} & \text{for } s = 1 \\ \frac{1}{8} & \text{for } s = 2 \\ \frac{1}{8} & \text{for } s = 3 \end{cases}$$

Since $p(s)$ *is a probability mass function* ($p(s) \geq 0$ *and* $\sum_{s \in S} p(s) = 1$) *the function* $X : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{P}(S))$ *turns into a RV.*

The corresponding CDF $F_X : \mathbb{R} \to [0, 1]$ *is given by*

$$F(t) = \begin{cases} 0 & \text{for } t < 0 \\ \frac{1}{4} & \text{for } 0 \leq t < 1 \\ \frac{3}{4} & \text{for } 1 \leq t < 2 \\ \frac{7}{8} & \text{for } 2 \leq t < 3 \\ 1 & \text{for } t \geq 3 \end{cases}$$

If we repeat a series of independent Bernoulli Experiments we can look at the number of "successes" obtained in the series. The distribution of the number of successes in n trials is called the Binomial Distribution with parameters n and p.

Definition 3 *Let* n *be the number of trials of independent Bernoulli experiments with "success" probability* p *in each trial. The distribution of the number of successes is called the binomial distribution with parameters* n *and* p *and its PMF is given by*

$$p(s) = \binom{n}{s} p^s (1-p)^{n-s}, \quad s = 0, 1, 2, ..., n$$

Notation:

$$X \sim Bi(n, p)$$

The CDF for the Binomial Distribution is given by

$$F_{Bi(n,p)}(t) = P(X \leq t) = \sum_{i=0}^{t} \binom{n}{i} p^i (1-p)^{n-i}, \quad x \in \mathbb{R}$$

For a binomial distributed random variable $X \sim Bi(10, 0.25)$ the PMF and CDF are shown in Fig. 1

Remember: There is a whole family of binomial distributions and by specifying the parameters n and p we get a particular representative.

If we repeat Bernoulli experiments for an infinite number of times then under certain regularity conditions, we obtain a probability distribution on the space $S = \{0, 1, 2, 3, ...\}$. This distribution is called the Poisson Distribution.

Definition 4 *A random variable* X *with values in* $\{0, 1, 2, 3, ...\}$ *and PMF*

$$p(s) = e^{-\lambda} \frac{\lambda^s}{s!}$$

for some $\lambda > 0$ *is called Poisson random variable with parameter* λ.

Poisson random variables are often used to model the number of incidents in a given time frame.

For a random variable $X \sim Po(2.5)$ the PMF and CDF are given in Fig. 2.

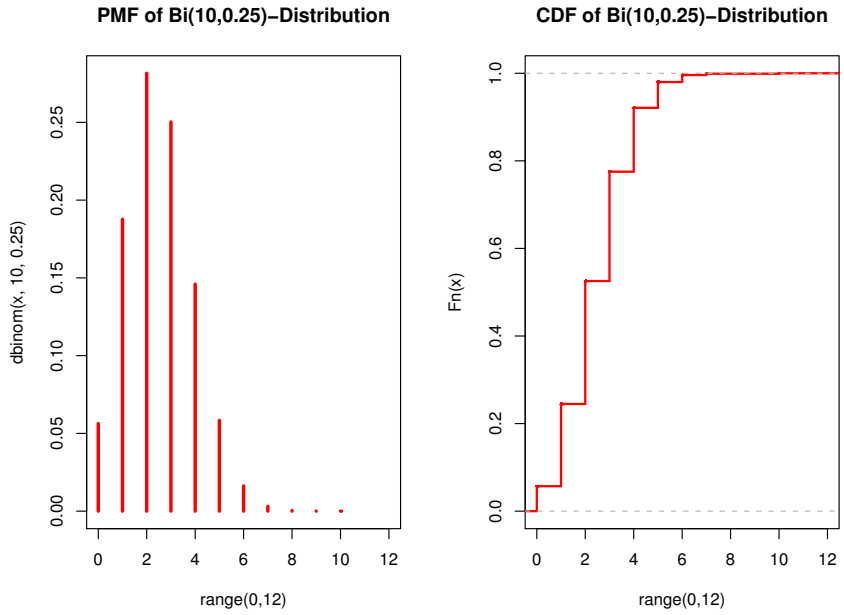


Figure 1: PMF and CDF of a random variable following a binomial distribution with parameters $n = 10$ and $p = 0.25$.

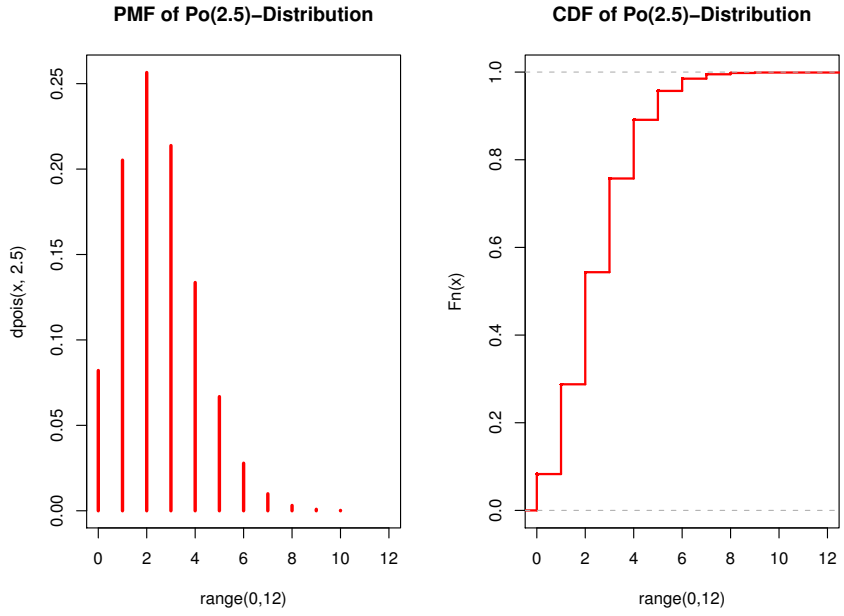


Figure 2: PMF and CDF of a random variable following a Poisson distribution with parameter $\lambda = 2.5$.

1.2 Continuous random variables

So far we have discussed random variables that take only a countable number of possible values.

Definition 5 *A random variable $X : \Omega \to \mathbb{R}$ with $X(\Omega)$ being an infinite, not count-*

able subset of $\mathbb{R}$ *is called a continuous random variable.*

It is impossible to define a continuous probability distribution by specifying a probability mass function, since we can't assign positive probability to an infinite number of points while restricting the sum of all probabilities to be equal to one.

Definition 6 *A function* $f : \mathbb{R} \to \mathbb{R}$ *with*

- $f(x) \geq 0, \forall x$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

is called a probability density function (PDF).

Lemma 1 *Any probability density function uniquely defines the probability distribution of a continuous random variable* X *via*

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

The corresponding CDF is characterized via

$$F(\alpha) = \int_{-\infty}^{\alpha} f(x)dx$$

The above lemma describes how we obtain the CDF once we know the PDF. The reverse way uses differentiation.

Lemma 2 *For a continuous random variable* X *the following properties hold:*

1. $f(x) = \frac{d}{dx}F(x) = F'(x)$
2. $P(X = a) = \int_a^a f(x)dx = 0$

So, in contrast to a discrete random variable, single points have the probability zero to occur. They are not impossible, but they are highly unlikely. There is a fine distinction between the notion of impossible and unlikely events. For those points that are impossible the probability density function will already be 0. When we observe a continuous random variable the outcome will be one of the unlikely events, but if we would repeat the observation many times we would never get exactly the same result (at least as long as our measuring device is good enough). And this is reflected in the zero probability.

Definition 7 *The distribution of the random variable* $X : \Omega \to [c, d]$ *having the following PDF is called* ***Uniform Distribution on*** $[c, d]$

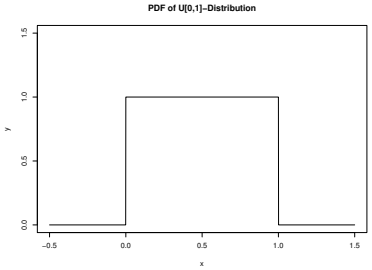
$$f(x) = \begin{cases} \frac{1}{d-c}, & \text{for } c \leq x \leq d \\ 0, & \text{otherwise.} \end{cases}$$

The CDF of a uniformly distributed RV X is given by

$$F(x) = \begin{cases} 0 & \text{for } x < c \\ \frac{x-c}{d-c}, & \text{for } c \leq x \leq d \\ 1, & \text{for } x > d. \end{cases}$$

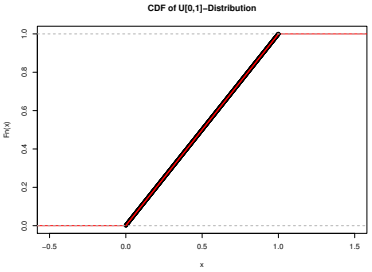
A special case is the **Uniform Distribution on** $[0,1]$ with PDF

$$f(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$



The corresponding CDF is given by

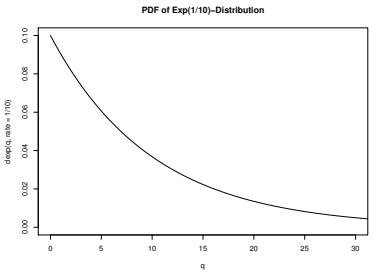
$$F(t) = \begin{cases} 0, & \text{for } t < 0 \\ t, & \text{for } 0 \leq t \leq 1 \\ 1, & \text{for } t > 1. \end{cases}$$



Definition 8 *The distribution of the random variable* $X : \Omega \to [0, \infty)$ *with PDF*

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \textit{for } x > 0 \\ 0, & \textit{otherwise} \end{cases}$$

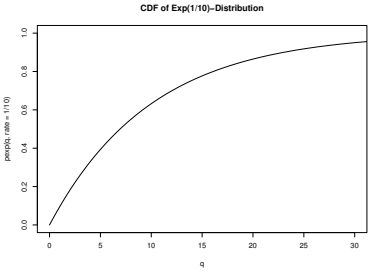
with parameter $\lambda > 0$



is called ***exponential distribution*** *with rate* λ*.*

The corresponding CDF is given by

$$F(t) = \begin{cases} 0, & \text{for } t < 0 \\ 1 - e^{-\lambda t}, & \text{for } t \geq 0. \end{cases}$$



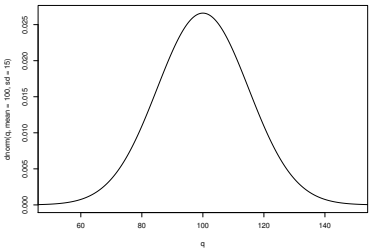
Example 3 *It is 10 minutes before your Statistics lecture starts and you are watching an important match of your favorite hockey team and the game goes overtime. Assuming that the length of overtime until sudden death follows an exponential distribution with rate* $1/7$ *minutes, what is the probability that overtime only lasts for at most 5 minutes (i.e. you'll be in time for your lecture)?*

$$\begin{aligned} X &\sim \text{Exp}(\frac{1}{7}) \\ P(X \leq 5) &= F_{Exp\frac{1}{7}})(5) \\ &= \left(1 - e^{-\frac{1}{7} \cdot 5}\right) \\ &= 0.5105 \end{aligned}$$

Definition 9 *The distribution of the random variable* $X : \Omega \to \mathbb{R}$ *with PDF*

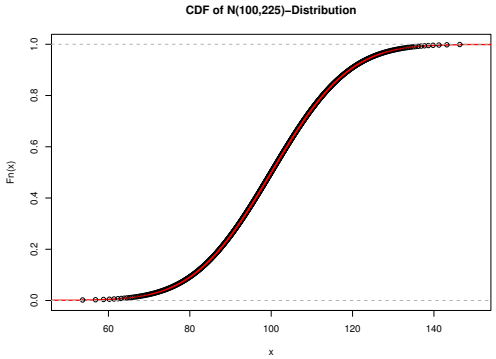
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$$

with $\mu \in \mathbb{R}$ *and* $\sigma^2 > 0$



is called ***normal distribution*** *with parameters* μ *and* σ^2.

There is no closed form expression for the CDF of a normal distributed random variable X



A special case is the normal distribution with $\mu = 0$ and $\sigma^2 = 1$ which is called **Standard Normal distribution**
Standard normal distributed RV are usually denoted by Z and the CDF of $N(0,1)$ is tabulated and usually denoted by $\Phi(z)$.

Theorem 1 *Let* $X \sim N(\mu, \sigma^2)$ *and let* Z *be the random variable defined by* $Z = \frac{X-\mu}{\sigma}$ *then* Z *is standard normally distributed, (Notation:* $Z \sim N(0,1)$*.)*

Example 4 *Let* $X \sim \mathrm{N}(10, 4)$*. Calculate* $P(X \leq 12.5)$*.*

$$\begin{aligned} P(X \leq c) &= P(\frac{X-\mu}{\sigma} \leq \frac{c-\mu}{\sigma}) \\ &= P(Z \leq \frac{c-\mu}{\sigma}) \\ &= P(Z \leq \frac{12.5-10}{2}) \\ &= P(Z \leq 1.25) \\ &= 0.8944 \end{aligned}$$

Example 5 *Points a student achieves in Statistics Quiz follows a normal distribution with* $\mu = 60$ *and* $\sigma^2 = 25$*. What is the probability of a student to earn a 'satisfactory' which means to have more than 55 but less than or equal to 70 points.*

$$\begin{aligned} X &\sim \mathrm{N}(60, 25) \\ P(56 \leq X \leq 70) &= P(X \leq 70) - P(X \leq 56) \\ &= F(70) - F(56) \\ &= \Phi(\frac{70-60}{5}) - \Phi(\frac{56-60}{5}) \\ &= \Phi(2) - \Phi(-0.8) \\ &= \Phi(2) - (1 - \Phi(0.8)) \\ &= 0.9772 - 0.2118 = 0.7654. \end{aligned}$$

1.3 Chararcteristics of random variables

1.3.1 Percentiles and Quantiles

We do not only want to know the probability of a particular outcome or for certain events, intervals or numbers, quite often we are interested in the reverse question. At which point does our distribution cover more than 10%, 20%, 50%, or 90%? Questions of this kind yield to the notion of percentiles (or more generally, quantiles).

Definition 10 *Let $X : (\Omega, \mathfrak{A}, P) \to (\mathbb{R}, \mathfrak{B})$ be a (real-valued) random variable and let $\alpha \in [0, 1]$. Every number $q \in \mathbb{R}$ for which we have*

$$\begin{aligned} P_X((-\infty, q]) &\geq \alpha \\ P_X([q, \infty)) &\geq 1 - \alpha \end{aligned}$$

is called an α-quantile (100α-percentile).

The quantiles separate the set of real numbers into two classes, an upper and a lower class. Most prominent are the following three quantiles: the first quartile ($\alpha = 0.25$), the second quartile or median ($\alpha = 0.5$), and the third quartile ($\alpha = 0.75$).

1.3.2 Expected values

Quite naturally the set of all possible values and the corresponding probabilities completely describes the behavior of a random variable. But this takes into account all instances and covers also rather unlikely events. The interest more often focusses on those values that are rather common and occur frequently or even more those values that might be called typical for a situation. This means that we are interested to know around which values a random variable is located or centered. Thus, we look at so-called location parameters. The median defined above is one indicator for that, the middle value in the sense that it divides the sample space into two equally likely classes. The most prominent location parameter, however, is the **expected value**.

Definition 11 *The expected value $E[X]$ of a (real-valued) random variable X is defined by*

$$E[X] = \begin{cases} \sum\limits_{x} xp(x) & \text{if } X \text{ is discrete} \\ \int\limits_{-\infty}^{\infty} xf(x)dx & \text{if } X \text{ is continuous.} \end{cases}$$

*Quite often the expected value is denoted by the Greek letter μ and is synonymously called **mean**.*

The expected value tells us how the value of the random variable will be on the average and describes the center of mass of the PMF or PDF, respectively.

Lemma 3 *Expected values for some of the distributions we already know:*
Bernoulli Distribution $X \sim Bi(1, p)$: $\mathrm{E}[X] = p$
Binomial Distribution $X \sim Bi(n, p)$: $\mathrm{E}[X] = np$
Poisson Distribution $X \sim Po_\lambda$: $\mathrm{E}[X] = \lambda$
Exponential Distribution $X \sim Exp(\lambda)$: $\mathrm{E}[X] = \frac{1}{\lambda}$
Uniform Distribution $X \sim U([0, 1])$: $\mathrm{E}[X] = \frac{1}{2}$
Uniform Distribution $X \sim U([c, d])$: $\mathrm{E}[X] = \frac{c+d}{2}$
Normal Distribution $X \sim N(\mu, \sigma^2)$: $\mathrm{E}[X] = \mu$

Expected values for functions of a random variable can be easily derived.

Theorem 2 *If* $Y = g(X)$ *then* $\mathrm{E}[Y] = \begin{cases} \sum_x g(x)p(x) \\ \int_{-\infty}^{\infty} g(x)f(x)dx \end{cases}$ *, i.e we don't need to calculate PMF/PDF of* Y

Theorem 3 *If* $Y = g_1(X) + g_2(X) + \ldots + g_n(X)$ *then*

$$\mathrm{E}[Y] = \mathrm{E}[g_1(X)] + \mathrm{E}[g_2(X)] + \ldots + \mathrm{E}[g_n(X)]$$

Corollary 1 *If* $Y = a + bX$ *then* $\mathrm{E}[Y] = a + b\mathrm{E}[X]$

1.3.3 Variance

The expected value is just one characteristic of a probability distribution.

Definition 12 *Let* X *be a random variable, then the expected squared deviation from the mean is called variance of* X*, i.e.* $VAR[X] = E[(X-\mu)^2]$.
The square root of the variance is called Standard Deviation , $STD[X] = \sqrt{VAR[X]}$

Lemma 4 *The following equations hold:*

- $VAR[X] = E[X^2] - \mu^2$
- $VAR[X] \geq 0, \qquad STD[X] \geq 0$
- $$\begin{aligned} Y = a + bX \Rightarrow \mathrm{VAR}[Y] &= b^2\mathrm{VAR}[X] \\ \mathrm{STD}[Y] &= |b|\mathrm{STD}[X] \end{aligned}$$

Lemma 5 *Variances for some of the distributions we already know:*
Bernoulli Distribution $X \sim Bi(1,p)$*:* $VAR[X] = p(1-p)$
Binomial Distribution $X \sim Bi(n,p)$*:* $VAR[X] = np(1-p)$
Poisson Distribution $X \sim Po_\lambda$*:* $VAR[X] = \lambda$
Exponential Distribution $X \sim Exp(\lambda)$*:* $VAR[X] = \frac{1}{\lambda^2}$
Uniform Distribution $X \sim U([0,1])$*:* $VAR[X] = \frac{1}{12}$
Uniform Distribution $X \sim U([c,d])$*:* $VAR[X] = \frac{(d-c)^2}{12}$
Normal Distribution $X \sim N(\mu, \sigma^2)$*:* $VAR[X] = \sigma^2$

1.4 Joint probability mass functions and independence

When we look at two random variables X and Y simultaneously we have the two individual probability mass functions but as well a joint probability mass function.

Example 6 *Let us pick a student's name at random from the rooster of Jacobs University students. Let* X *denote the random variable describing the gender of the student and* Y *the school he/she is enrolled. The probabilities for the joint features might be summarized in the following table:*

	School (Y)	
Gender (X)	*SHSS*	*SES*
male	*0.20*	*0.30*
female	*0.25*	*0.25*

Definition 13 *Let X and Y be two discrete random variables. The function*

$$\begin{aligned} p(x,y) &= P(X=x, Y=y) \\ &= P(\{\omega \in \Omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\}) \end{aligned}$$

is called the joint probability mass function of X and Y.

Any joint probability mass function must satisfy the following conditions:

1. $0 \leq p(x,y) \leq 1 \qquad \forall x \in X(\Omega)$ and $y \in Y(\Omega)$
2. $\sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} p(x,y) = 1$

From a given joint PMF, we can straightforwardly calculate the marginal PMF, that is, the PMF for each individual variable.

$$\begin{aligned} p_X(x) &= \sum_{y \in Y(\Omega)} p(x,y) \\ p_Y(y) &= \sum_{x \in X(\Omega)} p(x,y) \end{aligned}$$

Example 7 *For our gender/school example above we yield the following table:*

	School (Y)		
Gender (X)	*SHSS*	*SES*	$p_X(x)$
male	*0.20*	*0.30*	*0.50*
female	*0.25*	*0.25*	*0.50*
$p_Y(y)$	*0.45*	*0.55*	*1*

Definition 14 *Given a joint PMF and one marginal PMF, we define the function*

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

to be the conditional PMF of Y given $X = x$, and the function

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

to be the conditional PMF of X given $Y = y$.

Example 8 *Continuing our above example we see that the probability of a female student to be enrolled in SHSS is 0.5, since*

$$p(Y = SHSS \mid X = female) = \frac{0.25}{0.5} = 0.5.$$

Knowing the two marginals is not sufficient to calculate the joint PMF. Only, when the two random variables are independent.

Definition 15 *Two random variables X and Y are independent if the following equation holds: $p(x,y) = p_X(x)p_Y(y)$ for every x, y.*

Lemma 6 *The following conditions are equivalent:*

- $p(x,y) = p_X(x)p_Y(y)$ *for every* x, y
- $p(x \mid Y = y) = p_X(x)$ *for every* x, y
- $p(y \mid X = x) = p_Y(y)$ *for every* x, y

One particular example for a joint distribution is the **Multinomial Distribution**. This distribution is based on an experiment with k possible outcomes $a_1, \ldots, a_k$, where each outcome a_i has the probability p_i to occur, i.e.

$$P(a_i) = p_i, \quad \sum_{i=1}^{k} p_i = 1.$$

Definition 16 *Given an experiment with k possible outcomes $a_1, \ldots, a_k$, where each outcome a_i has the probability p_i to occur, that is independently repeated n times, we call the distribution of the random variable*

$$Y = (Y_1, ..., Y_k), \text{ where } Y_i \text{ counts the number of times that } a_i \text{ occurs}$$

the multinomial distribution (with parameters $p_1, p_2, \ldots, p_k$. The PMF of Y is given by

$$p(y_1, y_2, \ldots, y_k) = \frac{n!}{y_1! y_2! \ldots y_k!} p_1^{y_1} p_2^{y_2} \ldots p_k^{y_k}$$

1.5 Joint Probability Density Functions and Independence

Analogously, the joint distribution is defined for continuous random variables X and Y. Visually, the joint probability is the volume under a surface $z = f(x, y)$

Definition 17 *A function $f : \mathbb{R}^2 \to \mathbb{R}$ satisfying the conditions*

- $f(x, y) \geq 0, \qquad \forall x, y \in \mathbb{R}$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$

is called a joint probability density function and defines a probability distribution on the subsets of $\mathbb{R}^2$ by

$$P((X, Y) \in A) = \iint_{\{(x,y):(x,y)\in A\}} f(x, y) dx dy, \qquad \forall A \in \mathbb{R}^2.$$

Example 9 ***Joint Uniform Distribution on region*** B

$$f(x, y) = \begin{cases} \frac{1}{|B|}, & \text{for } (x, y) \in B \\ 0, & \text{otherwise.} \end{cases}$$

For all $A \subset B$:

$$P((X, Y) \in A) = \frac{\text{area of } A}{\text{area of } B}$$

Given a joint probability density function we can derive marginal and conditional PDF's.

marginal pdf

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

conditional pdf

for x s.t. $f_X(x) > 0$

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

for y s.t. $f_Y(y) > 0$

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

Having the conditional PDF, we can straightforwardly define the conditional expected value by

$$E[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x).$$

Definition 18 *Two continuous RV X and Y are independent if*

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

Lemma 7 *The following conditions are equivalent:*

- $f(x,y) = f_X(x) f_Y(y)$ *for every* x, y
- $f(x \mid Y = y) = f_X(x)$ *for every* x, y
- $f(y \mid X = x) = f_Y(y)$ *for every* x, y

Theorem 4 *Let X and Y be two random variables. If X and Y are independent, then*

$$E[XY] = E[X]E[Y]$$

Theorem 5 *Let X and Y be two random variables and let $W = aX + bY$. If X and Y are independent, then*

$$VAR[W] = a^2 VAR[X] + b^2 VAR[Y].$$

Definition 19 *Let X and Y be two random variables. The covariance of X and Y is defined by*

$$COV[X,Y] = E[(X - \mu_X)(Y - \mu_y)].$$

Accordingly, the correlation of X and Y is defined by

$$CORR[X,Y] = \frac{COV[X,Y]}{STD[X]STD[Y]}$$

Some remarks:

- $-1 \leq CORR[X,Y] \leq 1$
- The correlation is a measure of strength of linear relationship between the random variables X and Y.
- if $W = aX + bY$ then

$$\begin{aligned} VAR[W] &= a^2 VAR[X] + b^2 VAR[Y] \\ &\quad +2abCOV[X,Y] \end{aligned}$$

- if X and Y are independent $\Rightarrow$ X and Y are uncorrelated

1.6 Functions of Random Variables

(Almost) all functions of random variables generate again a random variable. We have seen above that some major characteristics of the distribution of this newly created random variable (e.g. expected value, variance) are easy to calculate without knowing the probability density function of the newly created RV. Using the rule of substitution provides a way to compute the pdf of the composition of a function with a random variable.

Theorem 6 *Rule of substitution:*
Let $\mathcal{X} \subseteq \mathbb{R}^k$, X be a random variable with values in $\mathcal{X}$ and PDF $f(x)$. Moreover, let $H : \mathcal{X} \to H(\mathcal{X}) \subseteq \mathbb{R}^k$ be a 1-1 mapping with differentiable inverse function H^{-1}, i.e. the differential mapping DH^{-1} does exist.
Then, the random variable $Y = H(X)$ has PDF

$$g(y) = f \circ H^{-1} |\det DH^{-1}|.$$

Lemma 8 *Convolution Theorem. Let X and Y be two independent real-valued random variables with PDF's $f(x)$ and $g(y)$.*
Then

- *$Z = X + Y$ has PDF $h(z) = \int_{\mathbb{R}} f(z-y)g(y)dy$*
- *if $Y > 0$ the RV $Z = \frac{X}{Y}$ has PDF $h(z) = \int_0^\infty f(zy)g(y)dy$.*

Theorem 7 *Let X and Y be two independent and normally distributed random variables with means μ and ν resp. and variances σ^2 and τ^2. Then the sum $X + Y$ is normally distributed with mean $\mu + \nu$ and variance $\sigma^2 + \tau^2$,*

$$i.e. X \sim N(\mu, \sigma^2), Y \sim N(\nu, \tau^2) \Rightarrow X + Y \sim N(\mu + \nu, \sigma^2 + \tau^2).$$

Definition 20 *Let X be a standard normal distributed random variable. The distribution of the random variable X^2 is called chi-square-distribution with one degree of freedom, i.e. $X^2 \sim \chi_1^2$.*
Let $X_1, X_2, \ldots, X_n$ be independent and standard normally distributed random variables, then the distribution of the sum of their squares $\sum_{i=1}^n X_i^2$ is called chi-square-distribution with n degrees of freedom, i.e. $\sum_{i=1}^n X_i^2 \sim \chi_n^2$.

Theorem 8 *Let $X_1, X_2, \ldots, X_n$ be independent and identically normal distributed random variables with means μ and variance σ^2. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.*
Then we have

1. *$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$*
2. *$\frac{n-1}{\sigma^2} S^2 \sim \chi_{n-1}^2$*
3. *$\bar{X}$ and S^2 are independent*

2 Statistical models

So, far we have been studying probability, how it arises, where in practice it might come form and how we can formalise this with probability spaces, random variables and so forth. We also have introduced ways to characterise probability distributions

on the range (co-domain, image) of a random variable by using probability mass functions, probability density functions or cumulative distribution functions. In all these instances the actual randomness lives on a probability space Ω which is unknown to us and also not needed for specifying the probability distribution of the random variable X under consideration.

In this section, we build on the notion of probability to create a framework for data analysis which is the ultimate goal in almost all kind of research in science and society.

2.1 General description

To set the scene let us start with a couple of typical examples for statistical analysis which have been determining the terminology and the standard procedures.

Example 10 *1. Mercedes Benz in their plant in Bremen are assembling a total of N GPS systems per day into their cars. Their supplier offers them a new version of the GPS system. It is known that a small fraction $N\theta$ of these systems are defective. It is too expensive to examine all of the systems prior to assembly. So to get information about the failure rate θ a sample of n systems is randomly selected for inspection. The number of defective systems found in this sample are our data gathered for further analysis.*

2. An economist wants to study the income distribution in her country. An exhaustive census is impossible to administer so the study is based on a representative sample of n individuals (randomly) drawn from the population.

3. An experimenter makes n independent determinations of the value of a physical constant μ (e.g. speed of light, gravitational constant, ...). The measurements are subject to random fluctuation and the data can be thought of as μ plus some random errors.

4. In medical research the efficacy of drugs is usually compared against the placebo effect. So, we want to compare two methods (placebo, drug) and see which of the two has the stronger effect on the individuals of a certain population. Such experiments consist of selecting $m + n$ individuals, of which m are assigned to method 1 (placebo) and n are assigned to method 2 (drug). Random variability in such an experiment would typically come from a variety of sources, mainly the differing responses among patients to the same drug, but also from error in measurements and slight variations in the composition of the drug.

Let us now provide some formal mathematical models for these examples:

Example 11 ***Sampling Inspection.*** *Our random variable X counts the number of defectives found in a sample of size n taken from a large lot of size N with failure rate θ, i.e. the total number of defectives in the full lot is $N\theta$. Then $X : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{F})$ with $S = \{0, 1, 2, \ldots, n\}$, and $\mathcal{F} = Pot[S]$ has the following distribution*

$$P(X = k) = \frac{\binom{N\theta}{k}\binom{N-N\theta}{n-k}}{\binom{N}{n}}$$

for $\max(n - N(1 - \theta), 0) \leq k \leq \min(N\theta, n)$. The distribution of X is called hypergeometric distribution $\mathcal{H}(N\theta, N, n)$.

In contrast to the probability modle studied so far, now $N\theta$ is unknown, and, in principle, can take any value between 0 and N. So, our model is actually a family of hypergeometric distributions $\mathcal{P} = \{\mathcal{H}(N\theta, N, n) : 0 \leq \theta \leq 1\}$ for X, any one of which could have generated the data that we actually observed.

Example 12 ***One-Sample Models.*** *The situation in Example 10.2 can be modelled by a set of data points $x_1, \ldots, x_n$ which are realizations of n independent and identically distributed random variables $X_1, \ldots, X_n$, with $X_i : (\Omega, \mathfrak{A}, P) \to (\mathbb{R}, \mathfrak{B})$ with common but unknown distribution P_X. It is common to require that the distribution P_X belongs to a defined set of distributions $\mathcal{P}$ (e.g. all normal distributions) on $(\mathbb{R}, \mathfrak{B})$.*

The situation in Example 10.3 can be cast into the same model. The n measurements taken can here also be seen as realizations of a a set of i.i.d. RVs X_i with

$$X_i = \mu + \epsilon_i, \qquad 1 \leq i \leq n$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is the vector of random errors. The RVs $X_i : (\Omega, \mathfrak{A}, P) \to (\mathbb{R}, \mathfrak{B})$ inherit their distribution in essence from the random errors. So, we can equivalently formulate the set of candidate distributions $\mathcal{P}_\epsilon$ in terms of the errors.

Conventional assumptions are that

- *The errors are independent (i.e. the different experimental measurements are independent, i.e. do not affect each other). So the random variables $\epsilon_1, \ldots, \epsilon_n$ are independent.*
- *The error distribution is the same, i.e. $\epsilon_1, \ldots, \epsilon_n$ are identically distributed.*
- *The distribution of ϵ is independent of μ.*

We then have $P_X(x) = P_\epsilon(x - \mu)$ and the model can be either specified by a set $\mathcal{P}_X$ that we postulate or by the set $\{(\mu, P_\epsilon :)\mu \in \mathbb{R}, P_\epsilon \in \mathcal{P}_\epsilon\}$ where $\mathcal{P}_\epsilon$ is the set of all allowable error distributions that we postulate.

The default model for error distributions is the family of normal distributions with mean 0 and unknown variance σ^2, i.e. $\mathcal{P}_\epsilon = \{N(0, \sigma^2 : \sigma > 0\}$ or alternatively, $\mathcal{P}_X = \{\Phi(\frac{x-\mu}{\sigma}) : \mu \in \mathbb{R}, \sigma > 0\}$ where Φ is the CDF of the standard normal distribution.

Example 13 ***Two-Sample Models.*** *Our drug-placebo example can be modelled in the following way: Let $x_1, \ldots x_m$ be the recorded responses (health status, reactions, biomedical measurements) for the m subjects who were given the placebo and let $y_1, \ldots y_n$ be the recorded responses (same measurements as for the placebo patients) for the subjects who were given the drug under investigation. We also have to assume that all subjects have a similar health status. We again assume that for each subject group (placebo = control, drug = treatment) our data are realizations of independent and identically distributed random variables. So, this means we assume to have i.i.d. RVs $X_1, \ldots, X_m$ with $X_i \sim P_X$ and i.i.d RVs $Y_1, \ldots Y_n$ with $Y \sim P_Y$. Hence our model consists of the set of all pairs of distributions (P_X, P_Y) with candidate sets for P_X and P_Y to be defined.*

To do so, researchers have to make additional specifications. Quite commonly, the ***constant treatment effect*** *assumption is made. This means that we assume that if a subject was given the placebo and had response x that the same subject when given the drug would yield response $y = x + \Delta$ where the parameter Δ does not depend on x. This model is often called the* ***shift*** *model since the distribution functions then have to satisfy the condition $F_Y(y) = F_X(y - \Delta)$.*

Making additional assumptions such as restricting the shift model to normal distributions with equal variance further simplifies the model. The Gaussian two-sample model with equal variance is then given by $\{(P_X, P_Y) : X \sim N(\mu, \sigma^2), Y \sim N(\mu + \Delta, \sigma^2)\}$.

In the examples above, we have twisted the description targeting at resulting in classical examples. You should not forget that in each of these specifications we took deliberate choices and decided about a set of assumption to be used. Some ingredients come from physical considerations, the nature of the experiment and so on (e.g the hypergeometric distribution in Inspection sampling). Others are derived from

experience, some might be rather arbitrary or just convenient for further computation (e.g. the independence requirements we have).

As a second note: I tried to stick with the notion of probability distributions to characterise the statistical model, i.e. the set of candidate distributions. In practice, it is often more convenient (and I also made use of it twice above) to work with the cumulative distribution functions, the probability density functions or the probability mass functions instead. In the end it does not matter since they all describe the same candidate set of probability distributions.

2.2 Parametrisations and Identifiability

In order to get a better handle on the set of statistical models it is common practice to use parametrisations of the candidate set of probability distributions. That is we introduce a map $\theta \mapsto P_\theta$ from a a space of labels, the so-called **parameter space** to $\mathcal{P}$, i.e. we write

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}.$$

Depending on the problem at hand, the parameter space might be clearly given (e.g. in our inspection sampling example the failure rate θ must be an element in $\Theta = \{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$) or it might be the consequence of a number of modelling choices. For example, in the one-sample model (see Example 12) making the three assumptions listed as bullet points (independent errors, identically distributed errors, and independence between errors and true response μ) and assuming that errors have mean 0, we can take $\Theta = \{(\mu, P_\epsilon) : \mu \in \mathbb{R}, P_\epsilon \text{with density } g \text{ such that } \int tg(t)dt = 0\}$ and $P_{(\mu,P_\epsilon)}$ has density $\prod_{i=1}^{n} g(x_i - \mu)$. If we additionally assume the errors to be normal we get for $\theta = (\mu, \sigma^2)$ that $\Theta = \{(\mu, P_\epsilon) : \mu \in \mathbb{R}, P_\epsilon \in \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}\} = \mathbb{R} \times \mathbb{R}^+$ and density function $\prod_{i=1}^{n} \frac{1}{\sigma}\varphi\left(\frac{x_i-\mu}{\sigma}\right)$ for P_θ where φ is the standard normal density.

The structure of the set Θ is taken to classify models into the following three categories:

- parametric models
- semiparametric models
- nonparametric models.

Our sampling inspection model and the classical one-sample normal error model are examples for parametric model. The one-sample model with $\Theta = \{(\mu, P_\epsilon) : \mu \in \mathbb{R}, P_\epsilon \text{ with density } g \text{ such that } \int tg(t)dt = 0\}$ and $P_{(\mu,P_\epsilon)}$ has density $\prod_{i=1}^{n} g(x_i - \mu)$ is a semiparametric model. Finally, the two-sample model with $\Theta = \{(P_X, P_Y)\}$ is a nonparametric model. The differentiation between the three classes basically depends on the **smoothness** of the map $\nu : \Theta \to \mathcal{P}$. A notion which we are going to make precise later.

Parametrisations are usually not unique. For example, we have defined the sampling inspection model in terms of the fraction of defectives θ. We could have used the number of defectives in the population $N\theta$ as well. Any one-to-one function of θ will yield a new parametrization. Which one to use might be a matter of convenience and simplicity. It is also a question of **identifiability**. We call a parametrisation **unidentifiable** when we have $P_{\theta_1} = P_{\theta_2}$ for two elements $\theta_1 \neq \theta_2$.

Example 14 *Assume the one-sample model*

$$\Theta = \{(\mu, P_\epsilon) : \mu \in \mathbb{R}, P_\epsilon \textit{ has (arbitrary) density } g\}$$

permitting the error distribution P_ϵ to be arbitrary. Let $\theta_1 = (0, N(0, 1))$ and $\theta_2 = (1, N(-1, 1))$. Then $P_{\theta_1} = N(0, 1) = P_{\theta_2}$.

The idea of parametrisation is very closely linked to the notion of a **parameter**, which is formally a map from the statistical model $\mathcal{P} \to \mathcal{N}$ some other space. A parameter is hence a feature of the distribution. For example in the shift model, we are interested in the parameter Δ which can be thought of as the difference in means of the two populations of responses. Some parameters are so-called **parameters of interest**, others are **nuisance parameters** such as the unknown variance σ^2 in the Gaussian one-sample model.

2.3 The general framework

At first glance, it might look as if all we need for a statistical model is a good choice of candidate distributions $\mathcal{P}$ or corresponding parameter space Θ or the map between them, called parametrisation. So, why has this course started with a lot of other notions, data generating environment, data recording procedure, sigma fields, you name it?

Well, they all play a crucial role in defining the set of candidate distributions $\mathcal{P}$. We might sometimes ignore the Grundgesamtheit Ω but we formally need it because our random variables X are defined on it. The universe Ω will be very much defined by the data generating environment. Quite often, the knowledge that we have will not be sufficient to completely describe Ω, so we leave many aspects vague and carry the relevant and known parts over to our random variables, the sample space S and the set of candidate distributions $\mathcal{P}$.

But what use for are the sigma fields $\mathfrak{A}$ and $\mathcal{F}$?

Let us look at an example.

Example 15 *Performance of students at Jacobs University is assessed in many courses and exams. To measure students' performance, the instructors have to record a grade for each student in their class using the Jacobs grading scheme. So our data recording procedure is assigning each student a grade. The data generating environment can be seen as the set of study conditions for each student during a particular semester. Rather difficult to describe in more detail, but in general we have some imagination of that.*

So, for a given class with n students, assuming no cheating, no group work, etc. we have n independent and identically distributed random variables $X : \Omega \to S$ with $S = \{1.00, 1.33, 1.67, 2.00, \ldots, 4.33, 4.67, 5.00\}$. As default σ-fields we take some unknown σ-field $\mathfrak{A}$ over Ω and the power set $Pot[S]$ over S.

At the end of each semester you can decide to clean your transcript by using a pass option. We can model this by a number of different ways. For example:

1. *Define the random variable $Y = h(X)$ with $h : S \to \{pass, fail\}, h(s) = \begin{cases} pass & \text{for } s \leq 4.33 \\ fail & \text{for } s \geq 4.67. \end{cases}$*
2. *Replace the data recording procedure by $\tilde{X} : \Omega \to \tilde{S}$ with $\tilde{S} = \{pass, fail\}$.*
3. *Keep the data recording procedure X and replace the power set $Pot[S]$ by the σ-field $\mathcal{S} = \{\emptyset, \{1.00, 1.33, 1.67, 2.00, \ldots, 4.33\}, \{4.67, 5.00\}, S\}$.*

The last option by changing just the σ-field reflects best the idea that the data recording was done in greater detail, but that only limited information is given to the outside. Exactly, what you want to do by using your pass option. You want to hide bad grades to people outside, in particular, grad schools and future employers.

This idea of modelling different levels of information by using different σ-fields is quite common. In particular, it is used for modeling financial data such as daily stock market prices. Here, you have essentially every day the same data generating environment as well as data recording procedure. Hence you want to keep your RVs

rather fixed. But every day you have a different amount of information, in particular, you add the information that came in after the stock market closed yesterday. So a common model are RVs $X_t : (\Omega, \mathfrak{A}_t, P) \to (\mathbb{R}, \mathfrak{B})$, for all $t = 0, 1, \ldots$.

3 The statistical problem

In the previous section, we have seen how we formulate a statistical model and got a glimpse on how we can actually put the different aspects that we know about our data collection process into the statistical model. But the description of the statistical model is only the first step. We typically collect data and want to use the data to take some decision.

In the sampling inspection example we record the number of defectives. By that we can select a particular distribution in our statistical model (i.e. the candidate set $\mathcal{P}$), but the ultimate goal is to decide whether we accept the delivery or request improved quality.[1] The collection of possible actions that the statistician can take after computing a statistic is called **decision space** and we denote it by D.

So, let us go back to a simple example, coin tossing.

Example 16 *Assume we do ten tosses with a coin of which we do not know the probability of coming up heads. So, we have*

$$X = (X_1, X_2, \ldots, X_{10}) : \Omega \to \{0, 1\}^{10}$$

and

$$\mathcal{P} = \bigotimes_{i=1}^{10} P_{X_i}.$$

It is convenient to use the parametrisation $\nu : [0, 1] \to \mathcal{P}$, with $\theta = P(X_i = 1) = 1 - P(X_i = 0)$, $0 \leq \theta \leq 1$.

Since we want to know what the probability of coming up heads is, we can describe our decision space by the set $D = \{d : 0 \leq d \leq 1\}$ where the decision $d = \cdot$ stands for the statement "my guess for the probability of this coin coming up heads is $\cdot$".

We could also have other decision spaces, e.g. $D = \{0, \frac{1}{10}, \frac{2}{10}, \ldots, 1\}$ or $D = \{fair, unfair\}$ or $D = \{fair, biased\ towards\ heads, biased\ towards\ tails\}$.

Which decision space is used depends typically on the kind of task that we want to perform. Statistical problems can be classified according to the different structure of the decision spaces into

- point estimation
- interval estimation
- hypothesis testing
- ranking
- regression
- others

Now having a set of possible actions, we actually want to take the best decision possible. In order to do so, we need an assessment of how good or bad the various possible decisions are for each possible underlying probability distribution $P_X \in \mathcal{P}$.

Definition 21 *A function $L : \mathcal{P} \times D \to \mathbb{R}^+$ is called a loss function.*

[1] There could also be other options, such as increasing the sampling inspection until you clearly know whether to accept or reject the delivery.

Often, the loss function is defined on the parameter space Θ instead of $\mathcal{P}$.

Example 17 *In our coin tossing example above with $D = \{fair, unfair\}$ a reasonable loss function is*

$$L(\theta, fair) = \begin{cases} 0 & for \mid \theta - 0.5 \mid < 0.005, \\ c & otherwise; \end{cases}$$

$$L(\theta, unfair) = \begin{cases} c & for \mid \theta - 0.5 \mid < 0.005, \\ 0 & otherwise. \end{cases}$$

There are plenty of loss functions. We typically choose them such that they qualitatively reflect what we are trying to do and to be mathematically convenient. A very common loss function for situations in which the the decision space is modelled by the real line is the quadratic loss function.

Definition 22 *The function $L : \Theta \times \mathbb{R} \to \mathbb{R}^+$ with $L(\theta, d) = (\theta - d)^2$ is called quadratic loss function.*

Other common loss functions are:

1. *$L : \Theta \times \mathbb{R} \to \mathbb{R}^+$ with $L(\theta, d) = \mid \theta - d \mid$* ***absolute value loss***
2. *$L : \Theta \times \mathbb{R} \to \mathbb{R}^+$ with $L(\theta, d) = \min\{(\theta - d)^2, a^2\}$* ***truncated quadratic loss***
3. *$L : \Theta \times \mathbb{R} \to \mathbb{R}^+$ with $L(\theta, d) = \begin{cases} 0 & for \mid \theta - d \mid \leq a \\ 1 & otherwise. \end{cases}$* ***confidence interval loss***

Specifying the statistical problem is not enough. In order to perform statistical inference, we need to select a **statistical procedure** which links the recorded data to the decisions to be taken.

Definition 23 *A function $t : S \to D$ is called a* ***statistical procedure*** *or* ***decision rule****.*

Actually, for the complete specification of a statistical procedure corresponding σ-fields on S and D are needed. For decision spaces which are finite we will typically use the power set on D. For ease of notation we leave out the σ-fields but should keep in mind that the statistical procedure as a function of a random variable (or as a function of random variables) is itself a random variable.

Example 18 *Let us consider the one-sample model in which we want to estimate the constant μ based on n realizations $x_1, \ldots, x_n$ of n i.i.d. RVs. $X_1, \ldots, X_n$ with distribution $P_X \in \mathcal{P}$. Two reasonable statistical procedures are $t_1(x_1, \ldots x_n) = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$ and $t_2 = median(x_1, \ldots, x_n) = \tilde{x}$.*

Example 19 *For the coin tossing example with 10 tosses, i.e. $X = (X_1, X_2, \ldots, X_{10}) : \Omega \to \{0, 1\}^{10}$, parameter space $\Theta = [0, 1]$ and decision space $D = \{d : 0 \leq d \leq 1\}$ we*

could use the following five decision rules:

$$
\begin{aligned}
t_1(x_1,\ldots,x_{10}) &= \frac{1}{10}\sum_{i=1}^{10} x_i \\
t_2(x_1,\ldots,x_{10}) &= \begin{cases} \frac{1}{j} & \text{if } x_1 = \ldots = x_{j-1} = 0 \text{ and } x_j = 1 \\ 0 & \text{if } x_1 = \ldots x_{10} = 0 \end{cases} \\
t_3(x_1,\ldots,x_{10}) &= \begin{cases} \frac{\pi}{5} & \text{if } \sum_{i=1}^{10} x_i \text{ is odd} \\ 0 & \text{if } \sum_{i=1}^{10} x_i \text{ is even} \end{cases} \\
t_4(x_1,\ldots,x_{10}) &= \frac{1}{2} \\
t_5(x_1,\ldots,x_{10}) &= \frac{1}{6}\sum_{i=1}^{6} x_i
\end{aligned}
$$

As you can see, not all procedures are equally good or reasonable. The first one is the sample mean which appears quite reasonable. The second procedure counts the coin flips until the first head occurs. Also a rather meaningful procedure for the given task. we would expect that the higher the probability for head, the earlier does head arise in a series of ten tosses. The third procedure appears to be rather esoteric. Why should we use only the information about the sum of heads in ten tosses being odd or even? And why should we then assign the parameter estimates given? The fourth procedure completely ignores the experimental results and firmly beliefs that the coin is fair. The fifth one only considers the first six tosses of the coin, ignoring the results of the last four ones.

Statistical procedures are very much dependent on the choice of decision space that we have made. Using a different decision space induces different statistical procedures.

Example 20 *Let us again look at the coin tossing example with 10 tosses, i.e.*

$$X = (X_1, X_2, \ldots, X_{10}) : \Omega \to \{0,1\}^{10}, \qquad X_i \sim Bi_{1,\theta},$$

parameter space $\Theta = [0,1]$ *and decision space*

$$D = \{\textit{fair}, \textit{biased towards heads}, \textit{biased towards tails}\}.$$

We could use the following decision rule:

$$
t(x_1,\ldots,x_{10}) = \begin{cases} \textit{fair} & \text{if } \sum_{i=1}^{10} = 5 \\ \textit{biased towards heads} & \text{if } \sum_{i=1}^{10} > 5 \\ \textit{biased towards tails} & \text{if } \sum_{i=1}^{10} < 5 \end{cases}
$$

Introducing the statistical procedure as a random variable which inherits its probability distribution from the data recording procedure X turns our loss function for a given parameter into a random variable. Since for the loss function we are allowing any member of our candidate set $\mathcal{P}$ as potentially correct probability distribution we aim at indicating this clearly by our notation. Hence, we will use the notation $E_P[L]$ to point out that we refer to the expected value of the loss function under the probability distribution P.

Definition 24 *Let $\mathcal{T}$ be a set of statistical procedures and $t : S \to D$ be a statistical procedure in $\mathcal{T}$, $L : \Theta \times \mathbb{R} \to \mathbb{R}^+$ a loss function, and X the outcome of the experiment (data recording procedure). Then the function*

$$R_t : \mathcal{P} \to \mathbb{R}^+ \qquad R_t(P) = E_P[L(P, t(X))],$$

providing an a-priori measure of the performance of the statistical procedure t, is called ***risk function of the procedure*** *t.*

Example 21 *(Example 19 continued) For the coin tossing example with $\Theta = [0, 1]$ and the above defined statistical procedures we obtain the following risk functions:*

$$\begin{aligned}
R_{t_1}(\theta) &= E_\theta[(t_1(X) - \theta)^2] \\
&= E_\theta[\left(\frac{1}{10}\sum_{i=1}^{10} X_i - \theta\right)^2] \\
&= \left(\frac{1}{10}\right)^2 E_\theta[(\sum_{i=1}^{10} X_i - 10\theta)^2] \\
&= \left(\frac{1}{10}\right)^2 VAR[\sum_{i=1}^{10} X_i] \\
&= \left(\frac{1}{10}\right)^2 10 \cdot \theta(1-\theta) \\
&= \frac{1}{10}\theta(1-\theta)
\end{aligned}$$

$$\begin{aligned}
R_{t_3} &= E_\theta[(t_3(X) - \theta)^2] \\
&= \frac{5}{11}(\frac{\pi}{5} - \theta)^2 + \frac{6}{11}(0 - \theta)^2 \\
&= \frac{\pi^2}{55} - \frac{2}{11}\pi\theta + \theta^2
\end{aligned}$$

$$\begin{aligned}
R_{t_4} &= E_\theta[(t_4(X) - \theta)^2] \\
&= E_\theta[(\frac{1}{2} - \theta)^2] \\
&= (\frac{1}{2} - \theta)^2.
\end{aligned}$$

$$\begin{aligned}
R_{t_5} &= E_\theta[(t_5(X) - \theta)^2] \\
&= E_\theta[(\frac{1}{6}\sum_{i=1}^{6} X_i - \theta)^2] \\
&= \frac{1}{6}\theta(1-\theta)
\end{aligned}$$

The risk functions of procedures 1, 3, 4 and 5 are visualised in Fig. 3.

In principle, we would now love to find a statistical procedure that is best in the sense that its risk function values are for any parameter value θ always smaller or at least as large as the risk function values for any competing procedure. Formalised,

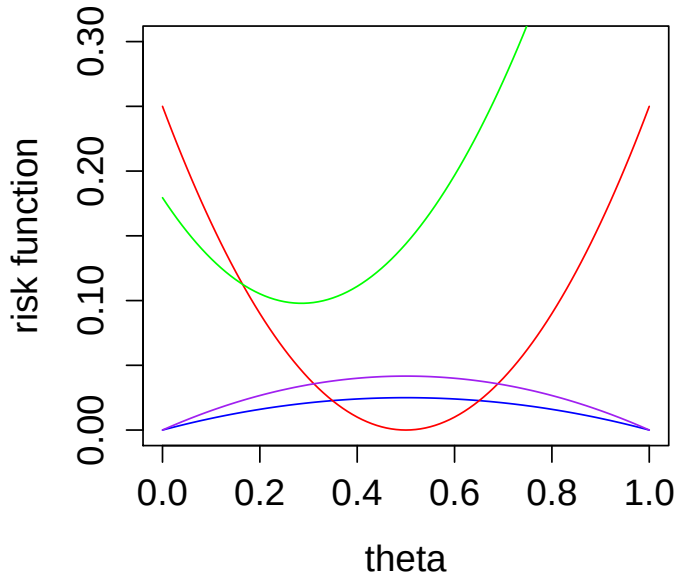


Figure 3: Some risk functions for the coin tossing example. The risk functions of procedure t_1 is given in blue, the one of t_3 in green, t_4 in red, and t_5 in purple.

that means that in a given statistical problem a procedure t' is uniformly best in the set of procedures $\mathcal{T}$, if

$$R_{t'}(P) \leq R_t(P), \qquad \forall P \in \mathcal{P} \text{ and all } t \in \mathcal{T}.$$

Now, the sad truth is that typically such uniformly best procedures do not exist (unless $\mathcal{P}$ is very peculiar). However, we can always exclude from our considerations statistical procedures which are outperformed by a competing model. That is, if t' is a decision rules such that

$$R_{t'}(P) \leq R_{t''}(P), \qquad \forall P \in \mathcal{P},$$

with

$$R_{t'}(P_1) < R_{t''}(P_1)$$

for some distribution $P_1 \in \mathcal{P}$ we say that t' dominates the procedure t''. Procedures for which a better procedure exists are called **inadmissible**. All procedures which are not dominated by another are called **admissible**.

In the example that we just have shown we can see that procedure 3 is outperformed by both procedures 1 and 5. Procedure 5 is outperformed by procedure 1. Hence procedures 3 and 5 are inadmissible, while procedures 1 and 4 are admissible (Actually, in the plot we only see that they are admissible with respect to the four procedures given. A formal proof showing admissibility in comparison to all procedures would be needed which we leave to the reader.)

4 Criteria for choosing a statistical procedure

As was pointed out before, it is usually impossible to find a statistical procedure that is uniformly best for all possible statistical models we are considering in our

statistical problem. Looking back at our example with the coin tossing: procedures such as t_4 – which have zero loss for one specific parameter value but do rather badly on many other parameters – seem rather meaningless since they do not refer to the data recorded. On the other hand, methods like t_4 do extremely well at one single parameter value and hence pose a big challenge for any other procedure whenever you look for the uniformly best procedure. So, two different ways are used to avoid such a dilemma. The first one is to restrict the number of competing procedures by imposing additional constraints or criteria. The second major approach gives up on the idea of uniformly best but compares risk functions on a more global rather than on a pointwise basis.

Imposing additional criteria and hence reducing the number of competing procedures does not per se guarantee that we will be able to find a uniformly best procedure. It can well be that also in the restricted set of procedures we are confronted with crossing risk functions such that no uniformly best method can be determined. Hence, quite regularly additional constraints will be imposed or two criteria will be mixed. We should also keep in mind that the imposition of criteria needs to be done with care since there is no good in restricting our set of potential procedures in such a way that the final set of competing procedures is too narrow.

There are a number of different approaches and we will point out some of them in more detail. We will first start with approaches that compare risk functions on global instead of a pointwise manner.

4.1 Minimax criterion

Definition 25 *A statistical procedure $t : (S, \mathcal{F}) \to (D, \mathcal{D})$ is said to be **minimax**, if*

$$\max_P R_{t^*}(P) = \min_{t \in \mathcal{T}} \left[\max_P R_t(P) \right].$$

So, the minimax strategy selects the one statistical procedure that minimizes the worst-case error. In very simple cases, we might be able to directly compute the minimax procedure by simply comparing the risk functions for all statistical procedures under consideration. This typically only works for small parameter spaces and limited classes of statistical procedures. For any reasonable statistical problem, the minimax procedure needs to be computed indirectly. There are various approaches for computing minimax procedures. One of them is closely related to the next approach.

4.2 Bayes criterion

The Bayesian approach assumes that the parameter space Θ can be turned into a probability space. So, we specify some distribution $\tilde{P}$ on the parameter space Θ and some appropriate σ-field on it. Since our parameter space is mostly taken to be a subset of the real line, the default choice is to use the corresponding Borel σ-field over this subset. In the classical Bayesian context, the probability distribution on the parameter space selected is supposed to quantify the knowledge we have about the likelihood of the different parameters **prior** to observing data.

As a measure of risk the Bayes criterion now asks to choose a statistical procedure t that minimizes the *average* error according to our prior beliefs about θ. This quantity is called *Bayes risk* of t and is defined by

$$r(t) = E_{\tilde{P}}[R_t(\theta)] = E_{\tilde{P}}[E_\theta[L(P, t(X))]] = \int_\Theta \int_S \tilde{p}(\theta) L(P, t(x)) dx d\theta$$

Here, $\tilde{p}$ is the probability density function for the a prior distribution $\tilde{P}$ for the unknown parameter θ.

Example 22 *Instead of a final exam your stats professor offers you a game. He has two coins, one of which is fair, so head and tail occur with probability* $\frac{1}{2}$*. The other coin is fabricated and comes up heads with probability* $\frac{1}{3}$*. Your professor selects one coin and you are allowed to make one single toss. You have to come up with a guess whether the coin is fair or fabricated.*

The setting can be modelled by the sample space $S = \{0, 1\}$ *where* 0 *means 'tail occurs' and* 1 *that 'head occurs'. The statistical model is described by the set* $\mathcal{P} = \{P_{\frac{1}{2}}, P_{\frac{1}{3}}\}$ *with* $P_\theta(1) = \theta$*. The default parameter space is* $\Theta = \{\frac{1}{2}, \frac{1}{3}\}$*, the decision space* $D = \{d_0, d_1\}$ *with* $d_i = \frac{1}{2+i}, i = 0, 1$*.*

Now we assume that we have zero-one loss, i.e.

$$L(\theta, t) = \begin{cases} 0 & \text{for } \theta = t \\ 1 & \text{for } \theta \neq t. \end{cases}$$

Using this loss function yields risk function

$$\begin{aligned} r_t(\theta) = E_\theta[L(\theta, t)] &= 0 \cdot P_\theta(\theta = t) + 1 \cdot P_\theta(\theta \neq t) \\ &= P_\theta(\theta \neq t) \\ &= P_\theta(\text{'wrong decision'}) \\ &= \begin{cases} P_{\frac{1}{2}}(d_1) & \text{for } \theta = \frac{1}{2} \\ P_{\frac{1}{3}}(d_0) & \text{for } \theta = \frac{1}{3} \end{cases} \end{aligned}$$

Student A now comes up with the idea the professor will always use the fair coin, Student B is convinced that he always uses the fabricated one, Student C and D make their decision dependent on the outcome of the coin toss. So their four statistical procedures are given by

$$\begin{aligned} t_A &: S \to D, \qquad x \mapsto d_0, \forall x \in S \\ t_B &: S \to D, \qquad x \mapsto d_1, \forall x \in S \\ t_C &: S \to D, \qquad x \mapsto \begin{cases} d_1, & \text{for } x = 0 \\ d_0, & \text{for } x = 1 \end{cases} \\ t_D &: S \to D, \qquad x \mapsto \begin{cases} d_0, & \text{for } x = 0 \\ d_1, & \text{for } x = 1 \end{cases} \end{aligned}$$

The corresponding risk functions are given by:

$$\begin{aligned} R_{t_A}(\theta) &= \begin{cases} 0 & \text{for } \theta = \frac{1}{2} \\ 1 & \text{for } \theta = \frac{1}{3} \end{cases} \\ R_{t_B}(\theta) &= \begin{cases} 1 & \text{for } \theta = \frac{1}{2} \\ 0 & \text{for } \theta = \frac{1}{3} \end{cases} \\ R_{t_C}(\theta) &= \begin{cases} \frac{1}{2} & \text{for } \theta = \frac{1}{2} \\ \frac{1}{3} & \text{for } \theta = \frac{1}{3} \end{cases} \\ R_{t_D}(\theta) &= \begin{cases} \frac{1}{2} & \text{for } \theta = \frac{1}{2} \\ \frac{2}{3} & \text{for } \theta = \frac{1}{3} \end{cases} \end{aligned}$$

Apparently, procedure t_D *is inadmissible since it is dominated by procedure* t_C*.*
Now, let us assume that the following a priori distribution ξ *is given by*

$$p_\xi(\theta) = \begin{cases} 0.9 & \text{for } \theta = \frac{1}{2} \\ 0.1 & \text{for } \theta = \frac{1}{3}. \end{cases}$$

The Bayes risk for the four procedures is then given by

$$\begin{aligned}
r_{t_A}(\xi) &= p_\xi(P_{\frac{1}{2}})R_{t_A}(\frac{1}{2}) + p_\xi(P_{\frac{1}{3}})R_{t_A}(\frac{1}{3}) \\
&= 0.9 \cdot 0 + 0.1 \cdot 1 \\
&= 0.1 \\
r_{t_B}(\xi) &= p_\xi(P_{\frac{1}{2}})R_{t_B}(\frac{1}{2}) + p_\xi(P_{\frac{1}{3}})R_{t_B}(\frac{1}{3}) \\
&= 0.9 \cdot 1 + 0.1 \cdot 0 \\
&= 0.9 \\
r_{t_C}(\xi) &= p_\xi(P_{\frac{1}{2}})R_{t_C}(\frac{1}{2}) + p_\xi(P_{\frac{1}{3}})R_{t_C}(\frac{1}{3}) \\
&= 0.9 \cdot \frac{1}{2} + 0.1 \cdot \frac{1}{3} \\
&= \frac{29}{60} \\
r_{t_D}(\xi) &= p_\xi(P_{\frac{1}{2}})R_{t_D}(\frac{1}{2}) + p_\xi(P_{\frac{1}{3}})R_{t_D}(\frac{1}{3}) \\
&= 0.9 \cdot \frac{1}{2} + 0.1 \cdot \frac{2}{3} \\
&= \frac{31}{60}
\end{aligned}$$

Now, let us assume that the following a priori distribution ζ is given by

$$p_\zeta(\theta) = \begin{cases} 0.4 & \text{for } \theta = \frac{1}{2} \\ 0.6 & \text{for } \theta = \frac{1}{3}. \end{cases}$$

The Bayes risk for the four procedures is then given by

$$\begin{aligned}
r_{t_A}(\zeta) &= p_\zeta(P_{\frac{1}{2}})R_{t_A}(\frac{1}{2}) + p_\zeta(P_{\frac{1}{3}})R_{t_A}(\frac{1}{3}) \\
&= 0.4 \cdot 0 + 0.6 \cdot 1 \\
&= 0.6 \\
r_{t_B}(\zeta) &= p_\zeta(P_{\frac{1}{2}})R_{t_B}(\frac{1}{2}) + p_\zeta(P_{\frac{1}{3}})R_{t_B}(\frac{1}{3}) \\
&= 0.4 \cdot 1 + 0.6 \cdot 0 \\
&= 0.4 \\
r_{t_C}(\zeta) &= p_\zeta(P_{\frac{1}{2}})R_{t_C}(\frac{1}{2}) + p_\zeta(P_{\frac{1}{3}})R_{t_C}(\frac{1}{3}) \\
&= 0.4 \cdot \frac{1}{2} + 0.6 \cdot \frac{1}{3} \\
&= 0.4 \\
r_{t_D}(\zeta) &= p_\zeta(P_{\frac{1}{2}})R_{t_D}(\frac{1}{2}) + p_\zeta(P_{\frac{1}{3}})R_{t_D}(\frac{1}{3}) \\
&= 0.4 \cdot \frac{1}{2} + 0.6 \cdot \frac{2}{3} \\
&= 0.6
\end{aligned}$$

Now, we can see that procedure t_C performs as good as procedure t_B, and both are now better than t_A. For both prior distributions, the procedure t_D is outperformed by t_C. A fact which we already knew from our considerations above on the risk function and the consequential inadmissibility of procedure t_D. How does the best Bayes procedure depend on the choice of prior distributions? Or phrased differently,

for which prior distributions is procedure t_C better than the rather naive procedures t_A or t_B?

Looking at the Bayes risks above, we see that the Bayes risk of procedure t_C depends on the a priori belief by $\frac{1}{2}\alpha + \frac{1}{3}(1-\alpha)$. The naive procedures t_A and t_B have minimum Bayes risk score $1-\alpha$ and α respectively.

Now solving

$$\begin{array}{rlrl} \frac{1}{2}\alpha + \frac{1}{3}(1-\alpha) \leq 1-\alpha & \textit{and} & \frac{1}{2}\alpha + \frac{1}{3}(1-\alpha) \leq \alpha \\ (1 - \frac{1}{3} + \frac{1}{2})\alpha \leq \frac{2}{3} & & \frac{1}{3} \leq (1 - \frac{1}{2} + \frac{1}{3})\alpha \\ \frac{7}{6}\alpha \leq \frac{2}{3} & & \frac{1}{3} \leq \frac{5}{6}\alpha \\ \alpha \leq \frac{4}{7} & & \alpha \geq \frac{2}{5} \end{array}$$

yields

$$\frac{2}{5} \leq \alpha \leq \frac{4}{7}.$$

So, for any a priori distribution that assigns a weight within $[\frac{2}{5}, \frac{4}{7}]$ to the probability of observing head, the statistical procedure t_C is the best Bayes procedure.

Comparing the risk functions we can also see that procedure t_C is the minimax procedure.

Computations for continuous distributions follow along similar lines by using the integral instead of summation. The convolution of a priori distribution and likelihood yields via Bayes rule the posterior distribution. Assuming appropriate probability densities by Bayes' rule we obtain the posterior density

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int_{-\infty}^{\infty} p(x \mid \theta)p(\theta)d\theta}.$$

For a single sample x, the conditional mean of θ can be written as

$$E[\theta \mid x] = \int_{-\infty}^{\infty} \theta p(\theta \mid x)d\theta,$$

in general.

For squared error loss we have a rather special situation. It actually turns out that in general, no matter which prior distribution is chosen for θ and no matter what the distribution of the data looks like (as long as we have i.i.d.), for squared error loss the best Bayes procedure is given by the conditional mean, i.e.

$$\hat{\theta}_{Bayes} = E_{\theta|X=x}[\theta \mid X = x].$$

So, instead of minimizing a rather complex collection of expected risks, we need to compute a conditional expectation only. This is yet another reason why squared error loss has become so popular in statistics.

Theorem 9 *Suppose a procedure $t*$ is the Bayes procedure relative to some prior distribution ξ and that the risk function $R_{t*}(P)$ is constant over $\mathcal{P}$. Then the procedure $t*$ is also minimax.*

Proof 1 *For any procedure t' we have $r_{t'}(\xi) \geq r_{t*}(\xi)$ since $t*$ is Bayes for ξ. Also $r_{t*}(\xi) = \max_P R_{t*}(P)$, since $R_{t*}(P)$ is constant over $\mathcal{P}$. Therefore,*

$$r_{t'}(\xi) \geq \max_P R_{t*}(P).$$

Now, the Bayes risk is an average of the values of the risk function, hence

$$r_{t'}(\xi) \leq \max_P R_{t'}(P).$$

So, we have

$$r_{t*}(\xi) = \max_P R_{t*}(P) \leq r_{t'}(\xi) \leq \max_P R_{t'}(P).$$

4.3 Unbiasedness

The concept of unbiasedness is widely used when our statistical problem deals with estimation. Let us assume that we want to estimate some parameter $\theta = \nu^{-1}(P)$ of the underlying distribution P.

Definition 26 *A statistical procedure $t : (S, \mathcal{F}) \to (D, \mathcal{D})$ is called an unbiased estimator of θ if*

$$E_P[t(X)] = \theta, \qquad \forall P \in \mathcal{P}.$$

The major motivation of using unbiased estimators is that the statistical procedure that we use should have a probability distribution which changes with P in such a way that it is always closely concentrated about our parameter θ. Unbiasedness now assumes that the **mean** (expected value) of the probability distribution of t is a good characteristic for the whole distribution and hence should be close to θ. This is however a rather strong assumption. There are plenty of probability distributions for which the mean is not a good indicator for the location of the probability mass under consideration.

Definition 27 *For any estimator t whose mean exists for all P, the function*

$$b_t : \mathcal{P} \to \mathbb{R}, \qquad b_t(P) = E_P[t(X) - \theta]$$

*is called the **bias function** of t for estimating $\theta = \nu^{-1}(P)$.*

Thus, a statistical procedure t is unbiased if and only if the bias function $b_t(P)$ equals zero for all candidate distributions.

In the current situation, assume further that our loss function is quadratic loss. Then our risk function turns into

$$R_{t(X)}(P) = E_P[(t(X) - \theta)^2]$$

and is called the **mean squared error**, denoted by $MSE[t(X)]$.

Straightforward calculation shows that the MSE depends on the variance of the statistical procedure and the **bias** of our statistical procedure.

Proposition 1

$$MSE[t(X)] = (Bias[t(X)])^2 + VAR[t(X)].$$

This proposition is the central piece of making the variance of a statistical procedure play the major role when comparing different estimators. The simple equality between risk function and variance of a statistical procedure is built on two modeling choices:

1. the use of squared error loss
2. the restriction to unbiased estimators.

There are plenty of statistical procedures which have zero variance but are highly biased and hence have high risk function. Consider for example the estimator t_4 in Example 19. This estimator has variance 0, but a bias that grows quadratically if you move away from the true parameter value.

Example 23 *Let $X_1, \dots X_n$ be i.i.d. RVs measuring the constant μ with $N(0, \sigma^2)$ errors (i.e. $X_i \sim N(\mu, \sigma^2)$). Using the mean $\bar{X}$ as our estimate of μ and taking quadratic loss, we obtain*

$$\begin{aligned} Bias[\bar{X}] &= E[\bar{X}] - \mu = 0 \\ VAR[\bar{X}] &= \frac{1}{n^2} \sum_{i=1}^{n} VAR[X_i] = \frac{\sigma^2}{n} \\ MSE[\bar{X}] &= R(\mu, \sigma^2, \bar{X}) = \frac{\sigma^2}{n}. \end{aligned}$$

If we know the precision σ^2 of our data recording procedure to be equal to a constant σ_0^2 or if we know at least that this precision is smaller than σ_0^2 then we can use our risk function for planning. For example, if we want to be guaranteed that $MSE[\bar{X}] \leq \epsilon^2$ we can achieve this by taking at least $n_0 = \frac{\sigma_0^2}{\epsilon^2}$ measurements.

Not knowing σ^2 restricts our planning possibility. Having taken n measurements we can estimate σ^2, for instance by $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ and obtain an ***a posteriori*** *estimate of risk $\frac{\hat{\sigma}^2}{n}$.*

The restriction to unbiased estimators is very common and widespread. One should, however, not forget that requiring unbiasedness might restrict the class unnecessarily.

4.4 Maximum likelihood

5 Hypothesis Tests

5.1 Test for binomial probabilities

Example 24 *You are gambling with one of your friends and tossing a coin. You are betting on heads, your friend on tails. After 20 games you have won only 5 times, the other 15 times your friend has been lucky. You are getting suspicious and wonder whether the coin is actually a fair one.*

How likely is it under the assumption of a fair coin that such a result occurs?

Let X be the number of heads, then $X \sim Bi(20, p)$. Under the assumption of a fair coin p is equal to 0.5. Then

$$\begin{aligned} P(X = 5) &= \binom{20}{5} 0.5^5 (1 - 0.5)^{20-5} \\ &= \binom{20}{5} 0.5^{20} \\ &= 0.01478577. \end{aligned}$$

So, under the assumption of a fair coin this result is rather unlikely.

After you claimed that your friend might have used a forged coin, he argues that the binomial distribution in this case assigns positive probability to even more extreme events and thus this is no sufficient proof.

Hence you calculate the probability of the observed event and all even more extreme events.

$$\begin{aligned} P(X \leq 5) &= \sum_{x=0}^{5} \binom{20}{5} 0.5^x (1 - 0.5)^{20-x} \\ &= F_{Bi(20,0.5)}(5) \\ &= 0.02069473. \end{aligned}$$

Now, you repeat your claim of a forged coin and underline it by the fact that the observed outcome and even more extreme outcomes are rather unlikely (probability = 2%).

Your calculations in the above example just showed that the observed outcome is unlikely under the assumed setting, however, it doesn't prove without doubt that the coin was forged. It just gives you a lot of evidence. Therefore, some conventions are needed.

Traditionally, so-called signi-ficance levels have been speci-fied a-priori (usually, one of the three choices $\alpha = 0.01, 0.05, 0.1$) to determine whether the results of such an observation evidently show a deviation of the actual situation from your assumption.

Nowadays, it is common to work with p-values, that is, to calculate as we did above the probabilities of the observed outcome and even more extreme events and the to decide according to the following conventions whether your observation gives enough evidence.

p-value > 0.1	$\Leftrightarrow$	not enough evidence
$0.05 <$ p-value < 0.1	$\Leftrightarrow$	there is some evidence, but possibly not enough
$0.01 <$ p-value < 0.05	$\Leftrightarrow$	there is evidence
p–value < 0.01	$\Leftrightarrow$	there is a lot of evidence

Formal description of binomial test:

- Let $X_1, \ldots, X_n$ be i.i.d. $Bi(1, p)$, p unknown.
- Let $p_0 \in (0, 1)$ be a hypothetical value (0.5 in above example, prob. of heads).
- Testproblem: $H_0 : p = p_0$ vs. $H_1 : p < p_0$
- Calculate test statistic $T = \sum_{i=1}^{n} x_i$.
- T follows a binomial distribution $Bi(n, p)$.
- After observing outcome x, calculate p-value $P(T < x)$.
- Decision based on $p - value$.

Step 1: Specify significance level α

Step 2: Compute test statistic

$$T = \sum_{i=1}^{n} x_i$$

Step 3: Find rejection region R

$$R = \begin{cases} [0, Bi_{n,p_0,1-\frac{\alpha}{2}}) \cup (Bi_{n,p_0;1-\frac{\alpha}{2}}, n] & \text{“two-sided”} \\ (Bi_{n,p_0;1-\alpha}, n) & \text{“one-sided”} \\ [0, Bi_{n,p_0;1-\alpha}) & \text{“one-sided”} \end{cases}$$

where $Bi_{n,p_0;1-\alpha}$ is the $(1-\alpha)$ - quantile of Bi_{n,p_0}-distribution.

Step 4: IF $T \in R$ reject null hypothesis H_0

5.2 z-test (Gauss-Test)

Example 25 *A brewery is starting to use a new filling machine for their bottles. A customer suspects that the new machine fills less than the indicated 500 ml in the bottles. From the manufacturer of the machine he receives the information that the content of a bottle is normally distributed and the machine when delivered was programmed to work with a mean of 500 ml and a standard deviation of 5 ml. The customer takes a random sample of 20 bottles and measures their content yielding an average of 497.50 ml.*

Did the brewery change the setting of the machine to fill less than the required 500 ml?

Let us calculate the probability.

$$\begin{aligned} P(\bar{X} \leq 497.50) &= P(\frac{\bar{X} - 500}{5/\sqrt{20}} \leq \frac{497.50 - 500}{5/\sqrt{20}}) \\ &= P(Z \leq \frac{-\sqrt{20}}{2}) \\ &= P(Z \leq -\sqrt{5}) \\ &= 0.01267366. \end{aligned}$$

Formal description of Gauss-test (one-sided "less than"):

- Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, μ unknown.
- Let $\mu_0 \in \mathbb{R}$ be a hypothetical value (500 in above example, average content).
- Test problem: $H_0 : \mu = \mu_0 \qquad$ vs. $\qquad H_1 : \mu < \mu_0$
- Calculate test statistic $T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$.
- T follows a standard normal distribution.
- After observing outcome $\bar{x}$, calculate p-value $P(T < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}})$.
- Decision based on $p - value$.

Step 1: Specify significance level α

Step 2: Compute test statistic

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Step 3: Find rejection region R

$$R = \begin{cases} (-\infty, \nu_{\frac{\alpha}{2}}) \cup (\nu_{1-\frac{\alpha}{2}}, \infty) & \text{"two-sided"} \\ (\nu_{1-\alpha}, \infty) & \text{"one-sided"} \\ (-\infty, \nu_\alpha) & \text{"one-sided"} \end{cases}$$

where $\nu_{1-\alpha}$ is the $(1-\alpha)$ - quantile of $N(0,1)$-distribution.

Step 4: IF $T \in R$ reject null hypothesis H_0

Example 26 *After the customer claimed at the brewery, the owner explains that he has changed the setting of the machine after he had taken a random sample of 25 bottles, measured their content, and calculated an average value of 501.50 ml.*

Did this give enough evidence to change the setting?

Let us calculate the probability.

$$\begin{aligned} P(\bar{X} \geq 501.50) &= P(\frac{\bar{X} - 500}{5/\sqrt{25}} \geq \frac{501.50 - 500}{5/\sqrt{25}}) \\ &= P(Z \geq 1.5) \\ &= 1 - P(Z \leq 1.5) \\ &= 0.0668072. \end{aligned}$$

Not, really, the probability is still small but now we know that such a result or even a more extreme one might occur in more than 6% of the cases.

Formal description of Gauss-test (one-sided "greater than"):

- *Let* $X_1, \ldots, X_n$ *be i.i.d.* $N(\mu, \sigma^2)$, μ *unknown.*
- *Let* $\mu_0 \in \mathbb{R}$ *be a hypothetical value (500 in above example, average content).*
- *Test problem:* $H_0 : \mu = \mu_0$ *vs.* $H_1 : \mu > \mu_0$
- *Calculate test statistic* $T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$.
- T *follows a standard normal distribution.*
- *After observing outcome* $\bar{x}$, *calculate p-value* $P(T < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}})$.
- *Decision based on* $p - value$.

Thus, we have seen that for the same general problem different viewpoints can be taken, resulting in different alternatives (one from the viewpoint of customer, one from the viewpoint of the brewery).

Let's take a third viewpoint and assume that an independent calibration institute wants to know whether the machine works properly.

For this institute deviations towards both sides of the target value, 500 ml, matter. Thus, they want to know how likely it is that under the assumption of the target value a sample results in an observation that deviates to such an extent from the target value. Assume that the institute takes a sample of 100 bottles yielding an average value of 499.2 ml.

Gives this enough evidence to assume that the machine is not working properly? Let us calculate the probability.

$$\begin{aligned} P(\bar{X} \leq 499.2 \text{ or } \bar{X} \geq 500.80) &= 1 - P(499.2 < \bar{X} < 500.80) \\ &= 1 - \left(P\left(\frac{\bar{X} - 500}{5/\sqrt{100}} \leq \frac{500.80 - 500}{5/\sqrt{100}} \right) \right. \\ &\quad \left. - P\left(\frac{\bar{X} - 500}{5/\sqrt{100}} \leq \frac{499.20 - 500}{5/\sqrt{100}} \right) \right) \\ &= 1 - (P(Z \leq 1.6) - P(Z \leq -1.6)) \\ &= 2 \cdot (1 - P(Z \leq 1.6)) \\ &= 0.1096. \end{aligned}$$

As we have seen there are different viewpoints to formulate the null and alternative hypothesis. The viewpoint decides what we want to show in our test and this we call the research hypothesis. It is the research hypothesis that is put in the statistical alternative. This is a consequence of inductive reasoning. We can never accept the null hypothesis, because our observation would just be an example for the validity of the null hypothesis, but if our observation is a counter-example to the null hypothesis we have proof that the null hypothesis is invalid and hence the alternative hypothesis must be true (since null hypothesis and alternative cover all possible parameter values).

In hypothesis testing there are three possible arrangements for the division of the parameter space into null and alternative hypothesis.

Two-sided alternative: "equal versus unequal"

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0$$

One-sided alternatives: "greater than or equal versus smaller"

$$H_0 : \mu \geq \mu_0 \text{ vs. } H_1 : \mu < \mu_0$$

One-sided alternatives: "less than or equal versus greater"

$$H_0 : \mu \leq \mu_0 \text{ vs. } H_1 : \mu > \mu_0$$

The following figures show the different alternatives.

Choosing a two-sided alternatives is the conservative solution and whenever a two-sided test yields a significant result you can also derive the direction of the deviation from the sign of the test statistic.

To get the p-value of the one-sided test simply divide the p-value of the two-sided test by 2.

Formal description of Gauss-test (two-sided):

- Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, μ unknown.
- Let $\mu_0 \in \mathbb{R}$ be a hypothetical value (500 in above example, average content).
- Test problem: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
- Use test statistic $T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$.
- T follows a standard normal distribution.
- After observing outcome $\bar{x}$, calculate p-value $P(\mid T \mid > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}})$.
- Decision based on $p - value$.

5.3 Student's t-Test

In the above examples we assumed that the standard deviation of the random variables is known. Usually, this is not the case. There might be quite a few reasons to justify the normal distribution, but the exact parameters of it will in practice be unknown. Thus, the above described z-test plays in practice just a minor role. But we can pursue the above outlined procedure and replace the theoretical standard deviation σ by an estimate. Taking the sample variance $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ leads to the new test statistic $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$.

The distribution of this test statistic is called t-distribution with $n - 1$ degrees of freedom.

Example 27 *If we return to our beer bottle tester, we see that in the above calculations we always assumed that the standard deviation is known (and is identical to the value given by the manufacturer). If the customer wouldn't know the standard deviation from the manufacturer, he could calculate the sample standard deviation (and let's say that his result was $s^2 = 28.4$, i.e. $s = 5.30$.). So, now he would like to know how unusual his observed average value $\bar{x} = 497.50$ was, yielding:*

$$\begin{aligned} P(\bar{X} \leq 497.50) &= P(\frac{\bar{X} - 500}{s/\sqrt{20}} \leq \frac{497.50 - 500}{5.30/\sqrt{20}}) \\ &= P(T \leq \frac{-2.5}{5.30/\sqrt{20}}) \\ &= P(T \leq \frac{-2.5}{1.185}) \\ &= 0.01745. \end{aligned}$$

Although this values is about twice as large as his result for the z-test, the p-value is still lower than 0.05, so he might reject the null hypothesis that the brewery is filling the bottles correctly.

Note, that the probability in the above calculation is obtained using a t-distribution with 19 degrees of freedom, since $n-1$ equals 19 in this example. Not knowing the standard deviation changes the probability calculation at two places, first, we get a different value ($\frac{-2.5}{1.2}$ compared with $-\sqrt{5}$) and second, we use the t_{19}-distribution instead of the standard normal distribution.

Formal description of (one-sample) t-test (two-sided):

- Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, μ and σ unknown.
- Let $\mu_0 \in \mathbb{R}$ be a hypothetical value (500 in above example, average content).
- Test problem: $H_0 : \mu = \mu_0 \qquad$ vs. $\qquad H_1 : \mu \neq \mu_0$
- Use test statistic $T = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$.
- T follows a t_{n-1}-distribution.
- After observing outcome $\bar{x}$, calculate p-value $P(\mid T \mid > \frac{\bar{x}-\mu_0}{s/\sqrt{n}})$.
- Decision based on $p - value$.

Step 1: Specify significance level α

Step 2: Compute test statistic

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Step 3: Find rejection region R

$$R = \begin{cases} \left(-\infty, t_{n-1}(\frac{\alpha}{2})\right) \cup \left(t_{n-1}(1-\frac{\alpha}{2}), \infty\right) & \text{“two-sided”} \\ (t_{n-1}(1-\alpha), \infty) & \text{“one-sided”} \\ (-\infty, t_{n-1}(\alpha)) & \text{“one-sided”} \end{cases}$$

where $t_{n-1}(\alpha)$ is the α - quantile of t_{n-1})-distribution.

Step 4: IF $T \in R$ reject null hypothesis H_0

5.4 Two-sample t-test

In many situations we do not compare a series of measurements with a theoretical value, but we would like to know whether the measurements of one series are significantly different from those of the other series. Do women have on average a higher IQ-score than men, do second year IUB students perform on average better than first year ones, do students majoring in the SHSS get better grades than students in the SES, does brewery A fill in less than brewery B, or is car model 1 consuming less gasoline than car model 2?

Let $X_{11}, \ldots, X_{1m}$ be the random variables associated with the measurements of model 1 and $X_{21}, \ldots, X_{2n}$ the ones for model 2. Assuming that all random variables follow a normal distribution, those for model 1 with mean μ_1 and variance σ_1^2, and those for model 2 with mean μ_2 and variance σ_2^2, we can derive a two-sample Gauss-test.

Let $\bar{X}_1 = \frac{1}{m}\sum_{i=1}^{m} X_{1i}$ and $\bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i}$. Under the assumption that all random variables here are independent we know that $\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{m})$ and $\bar{X}_2 \sim$

$N(\mu_2, \frac{\sigma_2^2}{n})$. Furthermore, the difference $\bar{X}_1 - \bar{X}_2$ is normally distributed with mean 0 and variance $\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$. The hypothesis $\mu_1 = \mu_2$ is equivalent to the hypothesis $\mu_1 - \mu_2 = 0$.

Our test statistic is hence $T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$ and this follows a standard normal distribution.

However, in practice the variances σ_1^2 and σ_2^2 are not known.

Assuming that we know that the variances for the two samples are the same, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we get the distribution of $\bar{X}_1 - \bar{X}_2$ to be $N(\mu_1 - \mu_2, \frac{\sigma^2}{m+n})$. Standardization yields the test statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

which when plugging in the estimator $S = \frac{1}{m+n-2}\left((m-1)S_1^2 + (n-1)S_2^2\right)$ turns under the null hypothesis to

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Here, $S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_{1i} - \bar{X}_1\right)^2$ and $S_2^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{2j} - \bar{X}_2\right)^2$ are the usual sample variances for each group individually.

Theorem 10 *The distribution of*

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

follows a t_{n+m-2}-distribution.

Proof 2 *We write*

$$T = \frac{\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}}{\frac{S}{\sigma}}.$$

From Theorem 1 we know that

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

follows a standard normal distribution and from Theorem 8 that $\frac{S^2}{\sigma^2}$ follows a scaled χ^2-distribution. Hence T is the ratio of a standard normal distribution and the square root of a chi^2-distribution. Thus, according to star problem 1 of Homework 5 the assertion is proved.

Formal summary of two-sample t-test:
Under the assumption that $x_{11}, \ldots, x_{1m}$ and $x_{21}, \ldots, x_{2m}$ are realisations of independent $N(\mu_i, \sigma_i^2)-$distributed RV with unknown means μ_i and **unknown** variances σ_i^2, $i = 1, 2$, but $\sigma_1^2 = \sigma_2^2$.

Step 1: Specify significance level α

Step 2: Compute test statistic

$$T = \sqrt{\frac{m \cdot n}{m+n}} \cdot \frac{\bar{X}_1 - \bar{X}_2}{S} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2} \cdot \frac{m+n}{n \cdot m}}}$$

with $S^2 = \frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}$, where
$S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_{1i} - \bar{X}_1)^2$ and $S_2^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{2i} - \bar{X}_2)^2$

Step 3: Find rejection region R

$$R = \begin{cases} (-\infty, -t_{m+n-2;1-\frac{\alpha}{2}}) \cup (t_{m+n-2;1-\frac{\alpha}{2}}, \infty) & \text{“two-sided”} \\ (t_{m+n-2;1-\alpha}, \infty) & \text{“one-sided”} \\ (-\infty, -t_{m+n-2;1-\alpha}) & \text{“one-sided”} \end{cases}$$

where $t_{m+n-2;1-\alpha}$ is the $(1-\alpha)$ - quantile of t_{m+n-2}-distribution.

Step 4: IF $T \in R$ reject null hypothesis H_0

Example 28 *An automobile club is testing the consumption of two different models. Measuring the consumption on a standard test drive yields two series of measurements each, one with 25 measurements for model 1 and one with 16 measurements for model 2. Let $\bar{X}_1 = 30.0$ be the average of the measurements for model 1, and $\bar{X}_2 = 25.5$ for model 2, the sample variances for model 1 and model 2 are $S_1^2 = 14.6$ and $S_2^2 = 14.8$ resp.*

Is there a significant difference in the consumption?

Let us calculate the test statistic.

$$\begin{aligned} T &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{m+n}{m \cdot n(m+n-2)}\left((m-1)S_1^2 + (n-1)S_2^2\right)}} \\ &= \frac{4.5}{\sqrt{\frac{25+16}{25 \cdot 16 \cdot (25+16-2)}(24 \cdot 14.6 + 15 \cdot 14.8)}} \\ &= 3.6689. \end{aligned}$$

Looking up the 0.975*-quantile of* t_{36} *we get 2.028 and hence reject the null hypothesis. Alternatively, we can calculate the probability*

$$P(\mid T \mid > 3.6689) = 1 - 0.999997 \leq 0.00001.$$

What happens when the two variances are not identical?

Then, under the null hypothesis our test statistic looks like

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}.$$

Once again from Theorem 8 we know that

$$\begin{aligned} S_1^2 &\sim \frac{\sigma_1^2}{m-1}\chi_{m-1}^2 \\ S_2^2 &\sim \frac{\sigma_2^2}{n-1}\chi_{n-1}^2. \end{aligned}$$

But what is the distribution for the convolution

$$\frac{1}{m}S_1^2 + \frac{1}{n}S_2^2?$$

There is no valid theoretical solution, but a practically satisfying approximation by using a t-distribution with degrees of freedoms that depend on the sample variances. In practice, we use t_f with

$$f = \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{S_1^2}{m}\right)^2 + \frac{1}{n-1}\left(\frac{S_2^2}{n}\right)^2}.$$

To check whether we can assume that the two variances are the same we can use Fisher's test for comparing two variances.

Let us look at the ratio $\frac{S_1^2}{S_2^2}$. Once again from Theorem 8 we know that

$$\frac{\sigma_1^2}{\sigma_2^2}\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim \frac{\sigma_1^2}{\sigma_2^2}\frac{\frac{1}{m-1}\chi_{m-1}^2}{\frac{1}{n-1}\chi_{n-1}^2}.$$

Under the null hypothesis that $\sigma_1^2 = \sigma_2^2$ the distribution of T is a F-distribution with $m-1$ numerator degrees of freedom and $n-1$ denominator degrees of freedom.

5.5 Most powerful tests

The previous applied perspectives on hypothesis testing shall now be integrated into the theoretical discussion of statistical models, statistical procedures and their quality. The hypothesis testing situation is characterized by the fact that our statistical model consist of two distinct subgrous $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$. Accordingly, the decision space also consists only of two elements $D = \{d_0, d_1\}$ where a decision d_i actually means that "I guess that true distribution P_X is in $\mathcal{P}_i$". To ease notation in the following I assume that we have a parametrization for our statistical model, i.e. $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ that also splits accordingly into two classes $\Theta_0 = \{\theta \in \Theta : P_\theta \in \mathcal{P}_0\}$ and $\Theta_1 = \{\theta \in \Theta : P_\theta \in \mathcal{P}_1\}$ It is also convenient and common practice to use zero-one loss

$$L(\theta, d_i) = \begin{cases} 1 & \text{if } P_\theta \notin \mathcal{P}_i \\ 0 \text{ if } P_\theta \in \mathcal{P}_i. \end{cases}$$

Under this loss, the risk function turns into the probability of making a wrong decision, i.e.

$$R_t(\theta) = P_\theta\left(\{\text{"}t \text{ makes wrong decision"}\}\right).$$

Since $\mid D \mid = 2$ it suffices to consider the behaviour of statistical procedures only for one of the two potential outcomes. So, in the following I use the short notation $t(x) = t(x, d_1)$.

Definition 28 *The function*

$$\beta_t(\theta) = P_\theta\left(\{\text{"}t \textit{ makes decision } d_1\text{"}\}\right) = E_\theta[(t(X)]$$

measuring the performance of a test t is called power function.

The risk function of the procedure t can now be written as

$$R_t(\theta) = \begin{cases} beta_t(\theta) & \text{if } \theta_{\in}\Theta_0 \\ 1 - \beta_t(\theta) & \text{if } \theta \in \Theta_1 \end{cases}.$$

Definition 29 *For a procedure t*

1. *the number $0 \leq \alpha_t \leq 1$ defined by*

$$\alpha_t = \max_{\theta \in \Theta_0} \beta_t(\theta)$$

is called significance level

2. *the subset of the sample space* $B_t = \{x : t(x) = 1\}$ *is called "critical region" or "rejection region"*

3. *the complementary subset* $S \setminus B_t$ *is called "acceptance region"*

Definition 30 *For some* $0 \leq \alpha \leq 1$ *let us denote by* $\mathcal{T}_\alpha = \{t \in \mathcal{T} : \alpha_t \leq \alpha$ *all the procedures that keep the significance level* α*. The procedure* $t* \in \mathcal{T}_\alpha$ *is called "uniformly most powerful test of level* α*" if*

$$R_{t*}(\theta) \leq R_t(\theta), \qquad forall\ t \in \mathcal{T}\alpha, \theta \in \Theta.$$

Testing between simple hypotheses: Let us now restrict to the case in which we only have two candidate distributions included in our model. Let us further assume that each of the distributions is characterized by a pdf. So we have $\mathcal{P} = \{P_0, P_1\}, \mathcal{P}_0 = \{P_0\}, \mathcal{P}_1 = \{P_1\}$ and each P_i defined by a pdf f_i.

Lemma 9 ***Neyman-Pearson Lemma***
For testing between two simple hypotheses at level $\alpha, 0 < \alpha < 1$*, every most powerful (MP) test of level* α *has the form*

$$t*(x) = \begin{cases} 1 & if\ f_1(x) > cf_0(x) \\ 0 & if\ f_1(x) < cf_0(x) \end{cases}$$

where c *is an appropriate constant (depending on* α*). Such a* c *and* $t*$ *exist for each value of* α*,* $0 < \alpha < 1$*.*

Example 29 *Let* $S = \{x : x > 0\}, f_0(x) = e^{-x}, f_1(x) = 2e^{-2x}$*, and* $0 < \alpha < 1$ *be given. Then*

$$\frac{f_1(x)}{f_0(x)} > c \Longleftrightarrow 2e^{-x} > c \Longleftrightarrow x < -\log\frac{c}{2}\,(= c')$$

So, we seek a critical region $B = \{x : x < c'\}$ *such that* $P_0(B) = \alpha$*. That is*

$$\alpha = \int_B f_0(x)dx = \int_0^{c'} e^{-x}dx = 1 - e^{-c'}$$

or $c' = \log\frac{1}{1-\alpha}$*. Then according to the NP-Lemma, the procedure* t *that rejects* H_0 *if* $x < \log\frac{1}{1-\alpha}$ *and decides for* H_0 *otherwise, i.e. the procedure*

$$t(x) = \begin{cases} 1 & if\ x < \log\frac{1}{1-\alpha} \\ 0 & else \end{cases}$$

is the most powerful test of level α*.*

Let X *above be replaced by* $X_1, \ldots, X_n$ *i.i.d. with* f_0 *of* f_1*. Then we have*

$$\begin{aligned} \frac{f_1(x_1, \ldots, x_n)}{f_0(x_1, \ldots, x_n)} &= \prod_{i=1}^n \frac{2e^{-2x_i}}{e^{-x_i}} \\ = \prod_{i=1}^n 2e^{-x_i} &> c \\ \Longleftrightarrow 2e^{-\sum x_i} > c \quad &\Longleftrightarrow \sum x_i < -\log\tfrac{c}{2} = c'' \end{aligned}$$

The density of $Y = \sum_{i=1}^n X_i$ *under* H_0 *is given by*

$$f_Y = y^{n-1}\frac{e^{-y}}{(n-1)!}, \qquad for\ y > 0.$$

This yields to

$$\alpha = \frac{1}{(n-1)!}\int_0^{c''} y^{n-1}e^{-y}dy.$$

For small n ($n = 2, 3$) you can solve for c'' by trial and error.
For $n > 10$ use the Central Limit Theorem (CLT) to conclude that Y is asymptotically normal with mean n and variance n under H_0:

$$\begin{aligned} \Longrightarrow P_0\left(\{\sum_{i=1}^{n} X_i < c''\}\right) &= P_0\left(\{\frac{\sum_{i=1}^{n} X_i - n}{\sqrt{n}} < \frac{c''-n}{\sqrt{n}}\}\right) \\ &\approx \Phi\left(\frac{c''-n}{\sqrt{n}}\right) \end{aligned}$$

where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution.

Given α, one determines k_α such that $\Phi(k_\alpha) = \alpha$ and then solves $\frac{c''-n}{\sqrt{n}} = k_\alpha$.

Composite hypotheses – UMP tests Now let us expand the Neyman-Pearson Lemma to statistical problems tat comprise more than just simple hypotheses. Let $\mathcal{P} \subseteq \mathcal{P}* = \{\theta : a < \theta < b\}$, $S \subseteq \mathbb{R}$ and the pdfs $f_\theta(x)$ such that we have for all $\theta_0, \theta_1 \in \mathcal{P}*$ that from $\theta_0 < \theta_1$ it follows that the likelihood ratio $\frac{f_{\theta_1}(x)}{f_{\theta_0}}$ is nondecreasing in $x \in S$. Such a set of distributions $\mathcal{P}*$ is called a "monotoen likelihood ratio (MLR) family".

If $\mathcal{P}$ only consists of two elements θ_0 and θ_1 with $\theta_0 < \theta_1$ for the hypotheses

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1.$$

Then the rejection region is defined by $B = \{x : x > c'\}$ with c' given by

$$\int_{c'}^{\infty} f_{\theta_0}(x)dx = \alpha.$$

The structure of the test including the threshold c' does not depend on θ_1 but only on θ_0.

Lemma 10 *Now let $\mathcal{P} = \{\theta : \theta_0 \leq \theta < b\}$*

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1$$

Then $t*(x) = \begin{cases} 1 & \text{for } x \in B \\ 0 & \text{for } x \notin B \end{cases}$ *is UMP of level α.*

Proof 3 *Assume it exists a $t \in \mathcal{T}_\alpha$ with $\beta_t(\theta_2) \leq \beta_{t*}(\theta_2)$ for some $\theta_2 > \theta_1$. Then t would be MP for the simple hypothesis θ_0 vs. θ_2. But this is a contradiction, since $t*$ is according to the Neyman-Pearson Lemma most powerful test for simple hypotheses.*

6 Linear models and unbiased linear estimators

6.1 Unbiased linear estimators

Let $X_1, X_2, \ldots, X_n$ be a series of random variables, $X_i : \Omega \to S, S \subseteq \mathbb{R}$, with statistical model $\mathcal{P}$ with parameter space Θ and some real-valued parameter of interest $\phi : \Theta \to \mathbb{R}$.

Definition 31 *1. The set of procedures* $\mathcal{L} = \{a_0 + \sum_{i=1}^{n} a_i X_i : a_0, \ldots, a_n \in \mathbb{R}\}$ *is called the class of linear estimators.*

2. its subset $\mathcal{L}(\phi) = \{t \in \mathcal{L} : \forall \theta \in \Theta : E_\theta[t] = \phi(\theta)\}$ *is called the subclass of linear estimators which are unbiased for* ϕ*..*

Our goal is to find for each $\theta \in \Theta$ a procedure that minimizes $VAR_\theta[t]$.

$$\min_{t \in \mathcal{L}(\phi)} VAR_\theta[t]$$

.

- If solution $t*$ depends on θ we call $t*$ locally minimum variance unbiased linear estimator for ϕ (LMVUL)
- If solution $t*$ is the same for all θ we call $t*$ uniformly minimum variance unbiased linear estimator for ϕ (UMVUL) or BLUE (best linear unbiased estimator)

Example 30 *There are some important cases for which*

1. *BLUE* $t*$ *exists and*
2. $t*$ *is also minimax among* ***all*** *estimators.*

Two of these cases are for the mean parameter $\phi(\theta) = E_\theta[X_1]$ *with*

1. $X_1, X_2, \ldots, X_n$ *i.i.d. with* $X_i \sim N(\mu, \sigma^2)$ *(i.e.* $\theta = (\mu, \sigma^2), \phi(\theta) = \mu$*) and a loss function* $L(\phi(\theta), d)$ *which is nondecreasing in* $\mid \phi(theta) - d \mid$
2. $X_1, X_2, \ldots, X_n$ *i.i.d. with* $X_i \sim P$*, P arbitrary distribution, but with loss function* $L(\phi(\theta), d)$ *which is strictly convex in* $\mid \phi(theta) - d \mid$

Example 31 *There are other cases in which the BLUE is not minimax and actually has fairly higher risks than the minimax procedure.*

1. *The uniform distribution on* $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$
   The sample mean $\bar{X}_n$ *is BLUE for* θ *and minimax in* $\mathcal{L}$ *with risk independent of* θ
   $$R_{\bar{X}_n}(\theta) = \frac{1}{12n}.$$
   The procedure $t_{minimax}(x_1, \ldots, x_n) = \frac{\max x_i + \min x_i}{2}$ *has risk*
   $$R_{t_{minimax}}(\theta) = \frac{1}{2(n+1)(n+2)} << \frac{1}{12n}$$
   and is minimax for all estimators of θ*.*
2. *The Cauchy-Distribution (=*t_1*-distribution).*
   Let P_θ *be given by pdf* $f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}$*. Then also the sample mean* $\bar{X}_n$ *follows a Cauchy-distribution. Since* $VAR[X_i] = \infty$*, the averaging* $VAR[\bar{X}_n] = \frac{\sigma^2}{n}$ *is without effect and the average is as good (or bad) as one single observation. However,* $t(X) = \tilde{X} = median(x_1, \ldots, x_n)$ *has variance* $VAR[\tilde{X}] = \frac{\pi^2}{4n}$*.*

6.2 The general linear model

Let $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ be an n-dimensional random vector (i.e. $Y_i : \Omega \to \mathbb{R}$)

Definition 32 *The model*

1. $E[Y] = B\phi \qquad D[Y] = \sigma^2 A$, *with*
   - *unknown mean parameter* $\phi \in \mathbb{R}^k$
   - *unknown variance parameter* $\sigma^2 \in (0, \infty)$
   - *known model (or design) matrix* $B \in \mathbb{R}^{n \times k}$
   - *known dispersion matrix* $A \in NND(n)$

   is called general linear model with moment assumptions.
2. $Y \sim N(B\phi, ; \sigma^2 A)$ *is called general linear model with normality assumption.*
3. $E[Y] = B\phi \qquad D[Y] = \sigma^2 I_n$, *with* I_n *denoting the* $(n \times n)$*-identity matrix is called classical linear model with moment assumptions.*
4. $Y \sim N(B\phi, ; \sigma^2 I_n)$ *is called classical linear model with normality assumption.*

6.2.1 Linear regression

Linear regression looks at the relationships between two or more continuous variables? The main aim is to describe the dependency of one random variable Y from knowing the values of some other variables $X_1, X_2, ..., X_k$.

Example 32 *Let us look at the salaries for a certain group of employees. The salary of a person might depend on the person's age, the education level, she or he has achieved, the months of working experience in the job, the salary they got when starting in the job, among other possible influential factors. Thus, here the current salary would be the dependent variable* Y *and all others would be explanatory variables.*

In general we use the following names to distinguish between the two kinds of variables that arise in linear regression:
$Y-$Variable: dependent variable, response variable
$X-$Variable: independent variable, explanatory variable
When using regression you suggest that there is causality, in the form that X implies Y.

Linear regression as a classical linear model We assume the linear model

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i, \quad \forall i = 1, \ldots, n,$$

where ϵ_i are independent errors with mean $E[\epsilon_i] = 0$ and variance $VAR[\epsilon_i] = \sigma^2$.

To ease notation we use a matrix notation for this model: Let

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix}, X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

then we can write the linear regression model with moment assumption as

$$E[Y] = X\beta \qquad D[Y] = \sigma^2 I_n,$$

where I_n denotes the $(n \times n)$-identity matrix.

We have $Y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times k}, x_i \in \mathbb{R}^k, \beta \in \mathbb{R}^k$ and $\sigma^2 > 0$.The matrix X is also called the "design matrix". The goal is to estimate the unknown parameter vector β to assess the magnitude of the linear relationship between the explanatory variables and the dependent variable.

with normal distribution assumption For estimating the parameter vector it is sufficient to ask for the above moment assumption (for expected value and dispersion matrix), however, to test the individual parameters in the model we have to assume normality of the errors, that is we assume

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i, \quad \forall i = 1, \ldots, n,$$

where ϵ_i are independent and identically normally distributed errors with mean $E[\epsilon_i] = 0$ and variance $VAR[\epsilon_i] = \sigma^2$ or in matrix notation:

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I),$$

or equivalently

$$Y \sim N(X\beta, \sigma^2 I).$$

There is no difference in the estimation process of the two model specifications, the normal distribution model allows in addition for testing the parameter values.

6.2.2 Least-squares-estimation

The goal in linear regression modelling is to specify the parameter vector β in such a way that the resulting errors $e_i = y_i - x_i'\beta$ are small. A suitable measurement for the size of the errors is the sum of the squared errors $\sum_{i=1}^{n}(y_i - x_i'\beta)^2$ or in matrix notation:

$$(Y - X\beta)'(Y - X\beta).$$

Our goal is hence to find a parameter vector $\hat{\beta}$ that minimizes the function $Q(\beta) = (Y - X\beta)'(Y - X\beta)$.

We find a solution for this problem by taking the derivative $\frac{\partial Q(\beta)}{\partial \beta}$ and solving the equation $\frac{\partial Q(\beta)}{\partial \beta} = 0$.

We have

$$\begin{aligned} Q(\beta) &= Y'Y - 2\beta' X'Y + \beta' X'X\beta \\ \frac{\partial Q(\beta)}{\partial \beta} &= -2X'Y + 2X'X\beta \\ \frac{\partial Q(\beta)}{\partial \beta} &= 0 \\ \iff X'Y &= X'X\beta \\ \iff \beta &= (X'X)^{-1}X'Y \end{aligned}$$

We thus define $\hat{\beta} = (X'X)^{-1}X'Y$ and call it the least squares estimator for β (short: LSE for β).

The matrix $H = X(X'X)^{-1}X'$ is called the "hat matrix".

Theorem 11 ***Gauss-Markov***
In the linear regression model $E[Y] = X\beta \qquad D[Y] = \sigma^2 I_n$, *the estimators* $\hat{\beta}$ *and* $S^2 = \frac{1}{n-k}\left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right)$ *have the following properties:*

1. $\hat{\beta}$ *is unbiased, i.e.* $E[\hat{\beta}] = \beta$
2. $D[\hat{\beta}] = \sigma^2 (X'X)^{-1}$
3. $\hat{\beta}$ *is the best, linear, and unbiased estimator (BLUE) for* β, *i.e. for all linear and unbiased estimators* $\tilde{\beta} = LY$ *with* $E[\tilde{\beta}] = \beta$ *we have* $VAR[c'\tilde{\beta}] \geq VAR[c'\hat{\beta}], forall c \in \mathbb{R}^n$.

4. $E[S^2] = \sigma^2$.

Theorem 12 *In the linear regression model with normality assumption* $Y \sim N(X\beta, \sigma^2 I)$ *the estimators* $\hat{\beta}$ *and* $S^2 = \frac{1}{n-k}\left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right)$ *have in addition the following properties:*

1. $\hat{\beta}$ *and* S^2 *are independent,*
2. $\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$,
3. $S^2 \sim \frac{\sigma^2}{n-k}\chi^2_{n-k}$,
4. $\hat{Y} \sim N(X\beta, \sigma^2 H)$,
5. $Y - \hat{Y} \sim N(0, \sigma^2(I - H))$.

From the above we can straightforwardly derive test statistics to test whether the parameters differ significantly from zero.

For the hypothesis $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ we hence use the test statistic

$$T_{\beta_j} = \frac{\hat{\beta}_j - 0}{\sqrt{S^2 (X'X)^{-1}_{jj}}}$$

and compare it with a suitable quantile of the t_{n-k}-distribution.

For the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \forall i = 1, \ldots, n,$$

we have

$$\begin{aligned} VAR[\hat{\beta}_0] &= \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sigma^2 \\ VAR[\hat{\beta}_1] &= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sigma^2 \end{aligned}$$

Design diagnostic:

Looking at the hat-matrix H we can derive the following properties:

- $H = H' = H^2$
- $h_i = H_{ii} \in \{0, 1\}$ and for $h_{ij} = H_{ij}$ we have

$$h_{ij} \leq h_i \cdot (1 - h_i).$$

- $h_i \approx 1 \Rightarrow h_{ij} \approx 0$ and $\hat{y}_i \approx y_i$ "Leverage effect"
- heuristic: if $h_i > \frac{3k}{n}$ then the ith observation is a high leverage point

7 Appendix: Multivariate Normal Distribution

Let $X_1, X_2, \ldots, X_n$ be a series of random variables.

Definition 33 *Then the vector* $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ *is called an n-dimensional random vector.*

The vector $E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}$ *is called the "vector of expected values" or the "expectation vector" of* X.

The matrix

$$\begin{aligned} D[X] &= E[(X - E[X])(X - E[X])'] \\ &= \begin{pmatrix} VAR[X_1] & COV[X_1, X_2] & \dots & COV[X_1, X_n] \\ COV[X_2, X_1] & VAR[X_2] & & COV[X_2, X_n] \\ \vdots & & \ddots & \vdots \\ COV[X_n, X_1] & COV[X_n, X_2] & \dots & VAR[X_n] \end{pmatrix} \end{aligned}$$

is called the "dispersion matrix" or the "variance-covariance" matrix.

The dispersion matrix is a symmetric and positive definite matrix, i.e.

$$z' D[X] z > 0, \forall z \in \mathbb{R}^n, \text{ with } z \neq 0.$$

Lemma 11 *Let* X *and* Y *be n-dimensional random variables. Then we have* $\forall A \in \mathbb{R}^{p \times n}, \forall b \in \mathbb{R}^p$

1. $E[AX + b] = AE[X] + b$
2. $D[AX + b] = AD[X]A'$
3. $E[X + Y] = E[X] + E[Y]$

Definition 34 *Let* $X \in \mathbb{R}^p$ *and* $Y \in \mathbb{R}^q$ *be random vectors, then*

$$COV[X, Y] = E[(X - E[X])(Y - E[Y])'] \in \mathbb{R}^{p \times q}$$

is called the "covariance" of X *and* Y. *The random vectors* X *and* Y *are uncorrelated if* $COV[X, Y] = 0$.

Lemma 12 *Let* $X \in \mathbb{R}^p$ *and* $Y \in \mathbb{R}^q$ *be random vectors. Then we have*

1. X *and* Y *are independent* $\Rightarrow$ X *and* Y *are uncorrelated*
2. X *and* Y *are uncorrelated and* $p = q \Rightarrow D[X + Y] = D[X] + D[Y]$.

Definition 35 *Let* $Z_1, \dots, Z_n$ *be independent and identically standard normal distributed random variables, i.e.* $Z_i \sim N(0, 1)$. *Then the distribution of the random vector* $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$ *is called the n-variate (or generally multivariate) standard normal distribution. We write:* $Z \sim N(0, I_n)$ *where* I_n *denotes the n-dimensional identity matrix.*

We have

- $E[Z] = \begin{pmatrix} E[Z_1] \\ \vdots \\ E[Z_n] \end{pmatrix} = 0$

- $D[Z] = \begin{pmatrix} VAR[Z_1] & COV[Z_1, Z_2] & \dots & COV[Z_1, Z_n] \\ COV[Z_2, Z_1] & VAR[Z_2] & & COV[Z_2, Z_3] \\ \vdots & & \ddots & \vdots \\ COV[Z_n, Z_1] & COV[Z_n, Z_2] & \dots & VAR[Z_n] \end{pmatrix}$
- the PDF of Z is given by $f_{0,I_n}(z) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}z'z}$

Definition 36 *Let $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$, A invertible, and Z a n-variate standard normal distributed random variable. Then we call the distribution of $Y = AZ + b$ the standard normal distribution with expectation vector b and dispersion matrix AA'.*

The PDF of Y is given by

$$\begin{aligned} f(y) &= f_{0,I_n}(A^{-1}y - A^{-1}b) \mid \det A^{-1} \mid \\ &= \frac{1}{\sqrt{(2\pi)^n} \mid \det A \mid} e^{-\frac{1}{2}(y-b)'(A^{-1})'A^{-1}(y-b)} \\ &= \frac{1}{\sqrt{(2\pi)^n \mid \det AA' \mid}} e^{-\frac{1}{2}(y-b)'(AA')^{-1}(y-b)} \end{aligned}$$