ST202 PROBABILITY, DISTRIBUTION THEORY, AND INFERENCE
LECTURE NOTES

Matteo Barigozzi*

Michaelmas Term 2015-2016 and Lent Term 2018-2019

# Contents



*Office: COL7.11 - Statistics Department

# 1 Probability Space

The main concept of probability is a random experiment, i.e. an experiment with an uncertain outcome. Associated with random experiments is the probability space which is made of three ingredients:

1. the collection of all possible outcomes: sample space $\Omega$;
2. the collection of all possible events: $\sigma$-algebra $\mathcal{F}$;
3. the probability measure $P$;

and we write it as

$$(\Omega, \mathcal{F}, P).$$

## 1.1 Sample space

**Experiment**: a procedure which can be repeated any number of times and has a well-defined set of possible outcomes.

**Sample outcome**: a potential eventuality of the experiment. The notation $\omega$ is used for an outcome.

**Sample space**: the set of all possible outcomes. The notation $\Omega$ is used for the sample space of an experiment. An outcome $\omega$ is a member of the sample space $\Omega$, that is, $\omega \in \Omega$.

**Example**: a fair six-sided die is thrown once. The outcomes are numbers between $1$ and $6$, i.e. the sample space is given by $\Omega = \{1, \ldots, 6\}$.

**Example**: a fair six-sided die is thrown twice. The outcomes are pairs of numbers between $1$ and $6$. For example, $(3, 5)$ denotes a $3$ on the first throw and $5$ on the second. The sample space is given by $\Omega = \{(i, j) \ : \ i = 1, \ldots, 6, \ j = 1, \ldots, 6\}$. In this example the sample space is finite so can be written out in full:

| | | | | | |
|---|---|---|---|---|---|
| (1,1) | (1,2) | (1,3) | (1,4) | (1,5) | (1,6) |
| (2,1) | (2,2) | (2,3) | (2,4) | (2,5) | (2,6) |
| (3,1) | (3,2) | (3,3) | (3,4) | (3,5) | (3,6) |
| (4,1) | (4,2) | (4,3) | (4,4) | (4,5) | (4,6) |
| (5,1) | (5,2) | (5,3) | (5,4) | (5,5) | (5,6) |
| (6,1) | (6,2) | (6,3) | (6,4) | (6,5) | (6,6) |

**Example**: the measurement of people's height has the positive real numbers as sample space, if we allow for infinite precision in the measurement.

**Example**: assume to have an experiment with a given sample space $\Omega$, then the experiment corresponding to $n$ replications of the underlying experiment has sample space $\Omega^n$. Notice that, in principle we can repeat an experiment infinitely many times.

## 1.2 Elementary set theory

**Notation**: given a sample space $\Omega$, we define the following objects:

| | Set terminology | Probability terminology |
|---|---|---|
| $A$ | Subset of $\Omega$ | Event some outcome in $A$ occurs |
| $A^c$ | Complement | Event no outcome in $A$ occurs |
| $A \cap B$ | Intersection | Event outcome in both $A$ and $B$ occur |
| $A \cup B$ | Union | Event outcome in $A$ and/or $B$ occur |
| $A \backslash B$ | Difference | Event outcome in $A$ but not in $B$ occur |
| $A \subseteq B$ | Inclusion | If outcome is in $A$ it is also in $B$ occur |
| $\emptyset$ | Empty set | Impossible event |
| $\Omega$ | Whole space | Certain event |

**Properties of Intersections and Unions**

1. Commutative: $A \cap B = B \cap A$,
   $A \cup B = B \cup A$.
2. Associative: $A \cap (B \cap C) = (A \cap B) \cap C$,
   $A \cup (B \cup C) = (A \cup B) \cup C$.
3. Distributive: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,
   $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
4. With whole space: $A \cap \Omega = A$,
   $A \cup \Omega = \Omega$.
5. With empty set: $A \cap \emptyset = \emptyset$,
   $A \cup \emptyset = A$.

**Properties of the complement set**: $A^c = \Omega \backslash A$, that is, $\omega \in A^c \iff \omega \notin A$.

1. $(A^c)^c = A$.
2. $A \cap A^c = \emptyset$.
3. $A \cup A^c = \Omega$.
4. $(A \cup B)^c = A^c \cap B^c$.
5. De Morgan's theorem (a generalization of 4 above): $(\bigcup_{i=1}^n A_i)^c = \bigcap_{i=1}^n A_i^c$.

**Partition of** $\Omega$: $\{A_1, \ldots, A_n\}$ is a partition if:

1. mutually exclusive: $A_i \cap A_j = \emptyset$ for any $i \neq j$, so $A_1, \ldots, A_n$ are disjoint sets;
2. exhaustive: $\bigcup_{i=1}^n A_i = \Omega$;
3. not-empty: $A_i \neq \emptyset$ for any $i$.

Notice that $n$ can be infinite.

## 1.3 Events

For any experiment, the events form a collection of all the possible subsets of $\Omega$ which we denote $\mathcal{F}$ and has the following properties:

1. $\emptyset \in \mathcal{F}$,
2. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$,
3. if $A_1, A_2, \ldots, \in \mathcal{F}$, then $\bigcup_{i=1}^\infty A_i \in \mathcal{F}$. The union has to be infinite.

Any collection of subsets with these properties is known as a $\sigma$-algebra.

If $\Omega$ has $n$ elements, then $\mathcal{F}$ has $2^n$ elements. Indeed, the number of elements of $\mathcal{F}$ is made of the sum of all possible combinations of $n$ elements, i.e., for any $0 \leq k \leq n$, we need to compute all the possible $k$-elements subsets of an $n$-elements set:

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

The binomial coefficient is also used to find the coefficients of binomial powers, the general formula is

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$$

and by setting $x = y = 1$ we have the result above. Another useful formula for the binomial coefficient is

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

## 1.4 Probability

In an experiment the intuitive definition of probability is the ratio between the number of favorable outcomes over the total number of possible outcomes or with the above notation, the probability of an event $A \subset \Omega$, such that $A \in \mathcal{F}$, is:

$$P(A) = \frac{\#\text{elements in } A}{\#\text{elements in } \Omega}.$$

Slightly more sophisticated is the "frequentist" definition of probability which is based on the frequency $f_n$ with which a given event $A$ is realized, given a total number $n$ of repetitions of an experiment:

$$P(A) = \lim_{n\to\infty} f_n.$$

**Example**: if we toss a fair coin the sample space is $\Omega = \{H, T\}$, then the event $A = \{H\}$ has probability

$$P(\{H\}) = \frac{\#\text{elements in } A}{\#\text{elements in } \Omega} = \frac{1}{2}.$$

Alternatively, we could compute this probability by tossing the coin $n$ times, where $n$ is large, and compute the number of times we get head say $k_n$. If the coin is fair, we should get

$$P(\{H\}) = \lim_{n\to\infty} \frac{k_n}{n} = \frac{1}{2}.$$

We here adopt a more mathematical definition of probability, based on the Kolmogorov axioms.

**Probability measure**: is a function $P : \mathcal{F} \longrightarrow [0, 1]$, such that

1. $P(A) \geq 0$,
2. $P(\Omega) = 1$,
3. if $A_1, A_2, \ldots,$ is an infinite collection of mutually exclusive members of $\mathcal{F}$ then

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

This in turn implies that for any finite collection $A_1, A_2, \ldots, A_n$ of mutually exclusive members of $\mathcal{F}$ then

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

We can associate a probability space $(\Omega, \mathcal{F}, P)$ with any experiment.

**Properties of probability measures**

1. $P(A^c) = 1 - P(A)$.
2. $P(A) \leq 1$.
3. $P(\emptyset) = 0$.
4. $P(B \cap A^c) = P(B) - P(A \cap B)$.
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
6. If $A \subseteq B$ then $P(B) = P(A) + P(B \backslash A) \geq P(A)$.
7. More generally if $A_1, \ldots, A_n$ are events then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n).$$

8. For any partition $A_1, \ldots, A_n$ of $\Omega$

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i).$$

   Notice that $n$ can be infinite.

9. Boole's inequality:

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

# 2 Counting or occupancy models

**Multiplication rule for counting ordered sequences**: an operation $A_i$ can be performed in $n_i$ different ways for $i = 1, \ldots, k$. The ordered sequence (operation $A_1$, operation $A_2$, $\ldots$, operation $A_k$) can be performed in $n_1 \cdot n_2 \cdot \ldots \cdot n_k$ ways. We write this product as $\prod_{i=1}^{k} n_i$.

When the sample space $\Omega$ is finite and all the outcomes in $\Omega$ are equally likely, we calculate the probability of an event $A$ by counting the number of outcomes in the event:

$$P(A) = \frac{\#\text{elements in } A}{\#\text{elements in } \Omega} = \frac{|A|}{|\Omega|}$$

Consider the following problem: $k$ balls are distributed among $n$ distinguishable boxes in such a manner that all configurations are equally likely or analogously (from the modeling point of view) we extract $k$ balls out on $n$. We need to define the sample space and its cardinality, i.e. the number of its elements. The balls can be distinguishable or undistinguishable which is analogous to saying that the order in the extraction matters or not. Moreover, the extraction can be with or without replacement, i.e. the choice of a ball

is independent or not from the ball previously chosen. In terms of balls and boxes this means that we can put as many balls as we want in each box (with replacement) or only one ball can fit in each box (without replacement).

There are four possible cases (three of which are named after famous physicists).

**Ordered (distinct), without replacement (dependent)**: in this case we must have $k \leq n$ and the sample space is

$$\Omega = \{(\omega_1 \dots \omega_k) \ : \ 1 \leq \omega_i \leq n \ \forall i \ \omega_i \neq \omega_j \text{ for } i \neq j\},$$

where $\omega_i$ is the box where ball $i$ is located. All the possible permutations of $k$ balls that can be formed from $n$ distinct elements, i.e. not allowing for repetition, are

$$|\Omega| = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}.$$

**Ordered (distinct), with replacement (independent) - Maxwell-Boltzmann**: the sample space is

$$\Omega = \{(\omega_1 \dots \omega_k) \ : \ 1 \leq \omega_i \leq n \ \forall i\},$$

where $\omega_i$ is the box where ball $i$ is located. Each ball can be selected in $n$ ways, so the total number of outcomes is

$$|\Omega| = \underbrace{n \cdot n \cdot \ldots \cdot n}_{k \text{times}} = n^k.$$

**Unordered (not distinct), without replacement (dependent) - Fermi-Dirac**: again we need $k \leq n$ and the sample space is

$$\Omega = \left\{ (\omega_1 \dots \omega_n) \ : \ \omega_i = \{0,1\} \ \forall i \text{ and } \sum_{i=1}^{n} \omega_i = k \right\},$$

with box $i$ occupied if and only if $\omega_i = 1$. Starting from the case of distinct balls, we have to divide out the redundant outcomes and we obtain the total number of outcomes:

$$|\Omega| = \frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1)}{1 \cdot 2 \cdot \ldots \cdot k} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

**Unordered (not distinct), with replacement (independent) - Bose-Einstein**: the sample space is

$$\Omega = \left\{ (\omega_1 \dots \omega_n) \ : \ 0 \leq \omega_i \leq k \ \forall i \text{ and } \sum_{i=1}^{n} \omega_i = k \right\},$$

with $\omega_i$ the number of balls in box $i$. This is the most difficult case to count. The easiest way is to think in terms of $k$ balls and $n$ boxes. We can put as many balls as we want in each box and balls are identical. To find all the possible outcomes it is enough to keep

track of the balls and of the walls separating the boxes. Excluding the 2 external walls, we have $n+1-2=n-1$ walls and $k$ balls, hence we have $n-1+k$ objects that can be arranged in $(n-1+k)!$ ways. However, since the balls and the walls are identical we need to divide out the redundant orderings which are $k!(n-1)!$, so

$$|\Omega| = \frac{(n-1+k)!}{k!(n-1)!} = \binom{n-1+k}{k}.$$

**Example**: in a lottery 5 numbers are extracted without replacement out of $\{1,\dots,90\}$. Which is the probability of extracting the exact sequence of numbers $(1,2,3,4,5)$?
The possible outcomes of this lottery are all the 5-tuples $\boldsymbol{\omega} = (\omega_1,\dots,\omega_5)$ such that $\omega_i \in \{1,\dots,90\}$. We can extract the first number in 90 ways, the second in 89 ways and so on, so

$$|\Omega| = 90\cdot 89\cdot 88\cdot 87\cdot 86 = \frac{90!}{85!}.$$

Since all the outcomes are equally likely, the probability we are looking for is $85!/90! \simeq 1/510^9$.

**Example**: if in the previous example the order of extraction does not matter, i.e. we look for the probability of extracting the first 5 numbers independently of their ordering, then $\Omega$ contains all the combinations of 5 numbers extracted from 90 numbers:

$$|\Omega| = \binom{90}{5}.$$

Since all the outcomes are equally likely, the probability we are looking for is $1/\binom{90}{5} \simeq 1/410^7$ so as expected it is greater than before, although still very small!

**Example**: which is the probability that, out of $n$ people randomly chosen, at least two were born in the same day of the year? We can define a generic event of the sample space as $\boldsymbol{\omega} = (\omega_1,\dots,\omega_n)$ such that $\omega_i \in \{1,\dots,365\}$. Each birth date can be selected $n$ times so

$$|\Omega| = \underbrace{365\cdot 365\cdot \ldots \cdot 365}_{n\,\text{times}} = 365^n.$$

Now we have to compute the number of elements contained in the event $A = \{\boldsymbol{\omega}\in\Omega : \boldsymbol{\omega}\text{ has at least two identical entries}\}$. It is easier to compute the number of elements of the complement set $A^c = \{\boldsymbol{\omega}\in\Omega : \boldsymbol{\omega}\text{ has all entries distinct}\}$. Indeed $A^c$ is made of all $n$-tuples of numbers that are extracted out of 365 numbers without replacement, so the first entry can be selected in 365 ways, the second in 364 ways and so on, then

$$|A^c| = \frac{365!}{(365-n)!}.$$

If we assume that the outcomes of $\Omega$ are all equally likely (which is not completely correct as we now that birth rates are not equally distributed throughout the year), then

$$P(A) = 1 - \frac{365!}{365^n(365-n)!},$$

which for $n = 23$ is $0.507$, for $n = 50$ is $0.974$, and for $n = 100$ is $0.9999997$.

**Example**: an urn contains $b$ black balls and $r$ red balls, we extract without replacement $n \leq (b + r)$ balls, what is the probability of extracting $k$ red balls? We first compute all the possible ways of extracting without replacement $n$ balls out of $(b + r)$, then $|\Omega| = \binom{b+r}{n}$. Let us assume that the all the balls are numbered and that the red ones have index $\{1, \ldots, r\}$ while the black ones have index $\{r + 1, \ldots, b + r\}$ so we are interested in the event

$$A = \{\omega \ : \ \omega \text{ contains exactly } k \text{ elements with index } \leq r\},$$

then is like asking for the all possible ways of extracting $k$ balls out of $r$ and $n - k$ balls out of $b$, therefore

$$P(A) = \frac{\binom{r}{k}\binom{b}{n-k}}{\binom{b+r}{n}}.$$

# Reading

Casella and Berger, Sections 1.1, 1.2.

# 3 Conditional probability

Let $A$ and $B$ be events with $P(B) > 0$. The conditional probability of $A$ given $B$ is the probability that $A$ will occur given that $B$ has occurred;

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

It is as if we were updating the sample space to $B$, indeed $P(B|B) = 1$. Moreover, if $A$ and $B$ are disjoint, then $P(A|B) = P(B|A) = 0$, once one of the two events took place the other becomes impossible.

By noticing that

$$P(A \cap B) = P(A|B)P(B) \quad \text{and} \quad P(A \cap B) = P(B|A)P(A),$$

we have the useful formula

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

**Law of total probability**: if $A_1, \ldots, A_n$ is a partition of $\Omega$ and $B$ is any other event defined on $\Omega$, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Notice that $n$ can be infinite.

**Bayes' rule**: if $A_1, \ldots, A_n$ is a partition of $\Omega$ and $B$ is any other event defined on $\Omega$, then for any $j = 1, \ldots, n$

$$P(A_j|B) = P(B|A_j)\frac{P(A_j)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}.$$

Notice that $n$ can be infinite.

**Multiplication rule for intersections**: let $A_1, \ldots, A_n$ be a set of events defined on $\Omega$,

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{j=1}^{n} P\left(A_j|\bigcap_{i=0}^{j-1} A_i\right),$$

where we define $A_0 = \Omega$.

# 4 Independence

If the occurrence of an event $B$ has no influence on the event $A$ then

$$P(A|B) = P(A),$$

then from previous section

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

so $A$ has no influence on $B$, moreover from Bayes' rule

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

and this is the definition of statistical independence. **Two** events $A$ and $B$ are said to be **independent** ($A \perp B$) if and only if

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

1. If $P(A) > 0$ then $P(B|A) = P(B) \iff A \perp B$.
   If $P(B) > 0$ then $P(A|B) = P(A) \iff A \perp B$.
2. If $A \perp B$ then $A^c \perp B^c$, $A^c \perp B$ and $A \perp B^c$.

A **collection** of events $A_1, \ldots, A_n$ is said to be **mutually independent** if for every subset $A_{i_1}, \ldots, A_{i_k}$ we have

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

**Example** A common misconception is that an event $A$ is independent of its complement $A^c$. In fact, this is only the case when $P(A) \in \{0, 1\}$ (check this!). Otherwise, the events $A$ and $A^c$ since the never occur at the same time and hence the probability of their intersection is zero.

**Example**: another common misconception is that an event is independent of itself. If $A$ is an event that is independent of itself, then

$$P(A) = P(A \cap A) = P(A)P(A) = (P(A))^2.$$

The only finite solutions to the equation $x = x^2$ are $x = 0$ and $x = 1$, so an event is independent of itself only if it has probability 0 or 1.

**Example**: consider tossing a coin 3 times, then we have $2^3 = 8$ possible outcomes and if the coin is fair each outcome has probability $\frac{1}{8}$. If we define $H_i$ the event of having head at the the $i$-th toss for $i = 1, 2, 3$ we have only four possible outcomes contained in each event $H_i$, therefore

$$P(H_i) = \frac{4}{8} = \frac{1}{2} \text{ for } i = 1, 2, 3.$$

To verify that $H_i$s are independent we need to compute

$$P(H_1 \cap H_2 \cap H_3) = P(\{HHH\}) = \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(H_1)P(H_2)P(H_3),$$

but we also have to compute for any $i \neq j$

$$P(H_i \cap H_j) = P(H_i)P(H_j),$$

so for example when $i = 1$ and $j = 3$

$$P(H_1 \cap H_3) = P(\{HTH, HHH\}) = \frac{2}{8} = \frac{1}{2} \cdot \frac{1}{2} = P(H_1)P(H_3).$$

**Example**: consider tossing a tetrahedron (i.e. a die with just four faces) with a red, a blue, a yellow face, and a face with all three colours. Each face has equal probability $\frac{1}{4}$ to be selected.[1] We want to see if the events: red ($R$), green ($G$), blue ($B$) are independent. The probability of selecting any colour is then $P(R) = P(G) = P(B) = \frac{1}{2}$ since all colours appear twice on the tetrahedron. Consider the conditional probability

$$P(R|G) = \frac{P(RG)}{P(G)} = \frac{1/4}{1/2} = \frac{1}{2} = P(R),$$

so the event $R$ is independent of the event $G$, by repeating the same reasoning with all couples of colours we see that colours are pairwise independent. However, we do not have mutual independence indeed, for example,

$$P(R|GB) = \frac{P(RGB)}{P(GB)} = \frac{1/4}{1/4} = 1 \neq P(R) = \frac{1}{2}.$$

[1] Due to its geometry in this case the selected face is the bottom one once the tetrahedron is tossed.

**Example**. Consider the following game: your ST202 lecturer shows you three cups and tells you that under one of these there is a squashball while under the other two there is nothing. The aim of the Monty Squashball problem[2] is to win the squashball by picking the right cup. Assume you choose one of the three cups, without lifting it. At this point one of the remaining cups for sure does not contain the ball and the your lecturer lifts it showing emptiness (selecting one at random if there is a choice). With two cups still candidates to hide the squashball, you are given a second chance of choosing a cup: will you stick to the original choice or will you switch to the other cup?

We can model and solve the problem by using conditional probability and Bayes' rule.

The probability of getting the ball is identical for any cup, so

$$P(\text{ball is in } k) = \frac{1}{3}, \quad k = 1, 2, 3.$$

Once you choose a cup (say $i$), your ST202 lecturer can lift only a cup with no ball and not chosen by yourself, he will lift cup $j$ (different from $i$ and $k$) with probability

$$P(\text{ST202 lecturer lifts } j | \text{you choose } i \text{ and ball is in } k) = \begin{cases} \frac{1}{2} & \text{if } i = k, \\ 1 & \text{if } i \neq k. \end{cases}$$

Let us call the cup you pick number 1 (we can always relabel the cups). Using Bayes' rule we compute (for $j \neq k$ and $j \neq 1$)

$$P(\text{ball is in } k | \text{ ST202 lecturer lifts } j) = \frac{P(\text{ST202 lecturer lifts } j | \text{ ball is in } k) P(\text{ball is in } k)}{P(\text{ST202 lecturer lifts } j)}.$$

Since $P(\text{ball is in } k) = 1/3$, we are left to compute (for $j \neq 1$)

$$\begin{aligned} P(\text{lecturer lifts} j) &= \sum_{k=1}^{3} P(\text{lecturer lifts } j | \text{ ball is in } k) P(\text{ball is in } k) \\ &= \frac{1}{2} * \frac{1}{3} + 0 * \frac{1}{3} + 1 * \frac{1}{3} = \frac{1}{2}. \end{aligned}$$

This can also be seen by symmetry and law of total probability.

So if you choose cup 1 and the ST202 lecturer lifts cup 2, the probability that the ball is in cup 3 is

$$P(\text{ball is in } 3 | \text{ST202 lecturer lifts } 2) = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

while the probability that the ball is in cup 1, i.e. the cup you chose at the beginning

$$P(\text{ball is in } 1 | \text{ST202 lecturer lifts } 2) = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}.$$

Hence, switching gives a higher probability of winning the squashball.

[2] this is an eco-friendly version of the famous Monty Hall problem which has "doors" for "cups", "goats" for "nothing" and a "car" for "squashball"; no animals are harmed in the Monty Squashball problem. It is also closely related to Bertrand's box paradox and the Prisoners' paradox (not to be confused with the Prisoners' dilemma)

# Reading

Casella and Berger, Sections 1.3.

# 5 Random variables

We use random variables to summarize in a more convenient way the structure of experiments.

**Borel $\sigma$-algebra**: is the $\sigma$-algebra $\mathcal{B}(\mathbb{R})$ (called the Borel $\sigma$-algebra) on $\Omega = \mathbb{R}$, i.e. the $\sigma$-algebra generated by (i.e. the smalles sigma-algebra containing) the intervals $(a, b]$ where we allow for $a = -\infty$ and $b = +\infty$.
We could have equally have taken intervals $[a, b]$ (think about this for a while!).

**Random variable**: a real-valued function is defined on the sample space

$$X : \Omega \longrightarrow \mathbb{R}$$

with the property that, for every $B \in \mathcal{B}(\mathbb{R})$, $X^{-1}(B) \in \mathcal{F}$.

Define, for all $x \in \mathbb{R}$, the set of outcomes

$$A_x = \{\omega \in \Omega : X(\omega) \leq x\}$$

then $A_x \in \mathcal{F}$. Thus, $A_x$ is an event, for every real-valued $x$.

The function $X$ defines a new sample space (its range) and creates a bijective correspondence between events in the probability space $(\Omega, \mathcal{F}, P)$ with events in the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ which allows for easier mathematical computations. We need to define the probability measure on the Borel $\sigma$-algebra.

**Example**: consider the experiment of tossing a coin $n$ times, the sample space is made of all the $n$-tuples $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ such that $\omega_i = 1$ if we get head and $\omega_i = 0$ if we get tail. An example of random variable is the function: number of heads in $n$ tosses which we can define as

$$X(\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i.$$

Consider the case in which we get $m$ times head with $m < n$. Then, for every number $m$ we can define the event $A_m = \{\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n) \in \Omega : X(\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i = m\}$.

Notice that in this example the random variables have only integer values which are a subset of the real line. Notice also that the original sample space is made of $2^n$ elements,

while the new sample space is made of the integer numbers $\{0 \ldots, n\}$ which is a smaller space.

**Example**: consider the random walk, i.e. a sequence of $n$ steps $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ such that the $i$-th step can be to the left or to the right. We can introduce a random variable that represents the $i$-th step by $X_i(\boldsymbol{\omega}) = \pm 1$ where it takes the value 1 if the step is to the left and -1 if the step is to the right. We can also introduce the random variable that represents the position of the random walk after $k$ steps: $Y_k(\boldsymbol{\omega}) = \sum_{i=1}^{k} X_i(\boldsymbol{\omega})$.

## 5.1 Cumulative distribution function

We must check that the probability measure $P$ defined on the original sample space $\Omega$ is still valid as a probability measure defined on $\mathbb{R}$. If the sample space is $\Omega = \{\omega_1, \ldots, \omega_n\}$ and the range of $X$ is $\{x_1, \ldots, x_m\}$, we say that we observe $X = x_i$ if and only if the outcome of the experiment is $\omega_j$ such that $X(\omega_j) = x_i$.

**Induced probability**: we have two cases

1. finite or countable sample spaces: given a random variable $X$, the associated probability measure $P_X$ is such that, for any $x_i \in \mathbb{R}$,

$$P_X(X = x_i) = P(\{\omega_j \in \Omega : X(\omega_j) = x_i\}).$$

2. uncountable sample spaces given a random variable $X$, the associated probability measure $P_X$ is such that, for any $B \in \mathcal{B}(\mathbb{R})$,

$$P_X(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\}).$$

Hereafter, given the above equivalences, we denote $P_X$ simply as $P$.

**Cumulative distribution function (cdf)**: given a random variable $X$, it is the function

$$F : \mathbb{R} \longrightarrow [0, 1], \quad \text{s.t.} \quad F(x) = P(X \leq x), \ \forall\, x \in \mathbb{R}.$$

**Properties of cdfs**: $F$ is a cdf if and only if

1. Limits: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.
2. Non-decreasing: if $x < y$ then $F(x) \leq F(y)$.
3. Right-continuous: $\lim_{h \to 0^+} F(x + h) = F(x)$.

**Probabilities from distribution functions**

1. $P(X > x) = 1 - F(x)$;
2. $P(x < X \leq y) = F(y) - F(x)$;
3. $P(X < x) = \lim_{h \to 0^-} F(x + h) = F(x^-)$;
4. $P(X = x) = F(x) - F(x^-)$.

**Identically distributed random variables**: the random variables $X$ and $Y$ are identically distributed if, for any set $A \in \mathcal{B}(\mathbb{R})$, $P(X \in A) = P(Y \in A)$. This is equivalent to saying that $F_X(x) = F_Y(x)$, for every $x \in \mathbb{R}$.

**Example**: in the random walk the step size random variable $X_i$ is distributed as:

$$P(X_i = 1) = \frac{1}{2}, \qquad P(X_i = -1) = \frac{1}{2}.$$

while

$$F_X(-1) = \frac{1}{2}, \qquad F_X(1) = 1.$$

The random variables $X_i$ are identically distributed. Moreover, they are also independent so

$$P(\boldsymbol{\omega}) = P(X_1 = \omega_1, \ldots, X_n = \omega_n) = \prod_{i=1}^{n} P(X_i = \omega_i),$$

for any choice of $\omega_1, \ldots, \omega_n = \pm 1$. Therefore, all $n$-tuples $\boldsymbol{\omega}$ are equally probable with probability

$$P(\boldsymbol{\omega}) = P(X_1 = \omega_1, \ldots, X_n = \omega_n) = \prod_{i=1}^{n} \frac{1}{2} = \frac{1}{2^n}.$$

Consider the random variable $Z$ the counts the steps to the right, then the probability of having $k$ steps to the right and $n - k$ steps to the left is

$$\begin{aligned} P(Z = k) = F_Z(k) &= (\text{\# of ways of extracting } k \text{ 1s out of } n) \, (\text{ Prob. of a generic } \boldsymbol{\omega}) \\ &= \binom{n}{k} \frac{1}{2^n}. \end{aligned}$$

We say that $X_i$ follows a Bernoulli distribution and $Z$ follows a Binomial distribution. The previous example of a fair coin can be modeled exactly in the same way but this time by defining $X_i(\boldsymbol{\omega}) = 0$ or 1.

## 5.2 Discrete random variables

A random variable $X$ is *discrete* if it only takes values in some countable subset $\{x_1, x_2, \ldots\}$ of $\mathbb{R}$, then $F(x)$ is a step-function of $x$, but still right-continuous.

**Probability mass function (pmf)**: given a discrete random variable $X$, it is the function

$$f : \mathbb{R} \longrightarrow [0,1] \text{ s.t. } f(x) = P(X = x) \quad \forall\, x \in \mathbb{R}.$$

**Properties of pmfs**

1. $f(x) = F(x) - F(x^-)$;
2. $F(x) = \sum_{i:x_i \leq x} f(x_i)$;
3. $\sum_i f(x_i) = 1$;
4. $f(x) = 0$ if $x \notin \{x_1, x_2, \ldots\}$.

## 5.3 Continuous random variables

A random variable $X$ is *continuous* if it takes values in $\mathbb{R}$ and its distribution function $F(x)$ is an absolutely continuous function of $x$ ($F$ is differentiable "almost everywhere") .

**Probability density function (pdf)**: given a continuous random variable $X$, (a version of) its density is an integrable function $f : \mathbb{R} \longrightarrow [0, +\infty)$ such that the cdf of $X$ can be expressed as

$$F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u \qquad \forall\, x \in \mathbb{R}.$$

**Properties of continuous random variables**

1. $P(X = x) = 0$ for any $x \in \mathbb{R}$;
2. $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$;
3. $f(x) \geq 0$ for any $x \in \mathbb{R}$;
4. $\int_a^b f(u)\mathrm{d}u = P(a < X \leq b)$.

Notice that, in principle, any nonnegative function with a finite integral over its support can be turned into a pdf. So if

$$\int_{A \subset \mathbb{R}} h(x)\mathrm{d}x = K < \infty$$

for some constant $K > 0$, then $h(x)/K$ is a pdf of a random variable with values in $A$.

**Unified notation**: given a random variable $X$;

$$P(a < X \leq b) = \int_a^b \mathrm{d}F(x) = \begin{cases} \sum_{i:a<x_i\leq b} f(x_i), & \text{if } X \text{ discrete,} \\ \int_a^b f(u)du, & \text{if } X \text{ continuous.} \end{cases}$$

# Reading

Casella and Berger, Sections 1.4 - 1.5 - 1.6.

# 6 Expectations

**Mean**: given a random variable $X$, its mean is defined as

$$\mu = \mathrm{E}[X] = \int_{-\infty}^{+\infty} x \mathrm{d}F(x) = \begin{cases} \sum_i x_i f(x_i), & \text{if } X \text{ discrete,} \\ \int_{-\infty}^{+\infty} x f(x) dx, & \text{if } X \text{ continuous,} \end{cases}$$

where $f$ is either the pmf or the pdf. The definition holds provided that $\int_{-\infty}^{+\infty} |x| \mathrm{d}F(x) < \infty$.

If we interpret $\mu$ as a good guess of $X$ we may also be interested to have a measure of the uncertainty with which $X$ assumes the value $\mu$, this is known as variance of $X$.

**Variance**: given a random variable $X$, its variance is defined as

$$\sigma^2 = \mathrm{Var}[X] = \int_{-\infty}^{+\infty} (x-\mu)^2 \mathrm{d}F(x) = \begin{cases} \sum_i (x_i - \mu)^2 f(x_i), & \text{if } X \text{ discrete,} \\ \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) \mathrm{d}x, & \text{if } X \text{ continuous,} \end{cases}$$

where $f$ is either the pmf or the pdf. The standard deviation is defined as $\sigma = \sqrt{\mathrm{Var}[X]}$. Notice that $\sigma^2 = \mathrm{E}[(X-\mu)^2]$. The definition holds provided that $\int_{-\infty}^{+\infty} (x-\mu)^2 \mathrm{d}F(x) < \infty$.

**Expectations**: for an integrable function $g : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\int_{-\infty}^{\infty} |g(x)| \mathrm{d}F(x) < \infty$, the expectation of the random variable $g(X)$ as

$$\mathrm{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) \mathrm{d}F(x) = \begin{cases} \sum_i g(x_i) f(x_i), & \text{if } X \text{ discrete,} \\ \int_{-\infty}^{+\infty} g(x) f(x) dx, & \text{if } X \text{ continuous,} \end{cases}$$

Note that we have cheated a bit here, since we need to show in fact that $g(X)$ is a random variable and also that the given expression corresponds to the one given above for the random variable $g(X)$. This can be done but is beyond the scope of ST202. Feel free to ask me if you would like to hear more about this.

**Properties of expectations**: for any constant $a$, integrable functions $g_1$ and $g_2$, and random variables $X$ and $Y$:

1. $\mathrm{E}[a] = a$;

2. $\mathrm{E}[ag_1(X) + bg_2(Y)] = a\mathrm{E}[g_1(X)] + b\mathrm{E}[g_2(Y)]$;

3. if $X \geq Y$ then $\mathrm{E}[X] \geq \mathrm{E}[Y]$;

4. $\mathrm{Var}[ag_1(X) + b] = a^2\mathrm{Var}[g_1(X)]$.

The variance of $X$ can be written in a more convenient form

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

## 6.1 Moments

Moments are expectations of powers of a random variable. They characterise the distribution of a random variable. Said differently (and somewhat informally), the more moments of $X$ we can compute, the more precise is our knowledge of the distribution of $X$.

**Moment**: given a random variable $X$, for $r$ a positive integer then the $r^{th}$ moment, $\mu_r$, of $X$ is

$$\mu_r = \mathrm{E}[X^r] = \int_{-\infty}^{+\infty} x^r \mathrm{d}F(x) \begin{cases} \sum_i x_i^r f(x_i), & \text{if } X \text{ discrete,} \\ \int_{-\infty}^{+\infty} x^r f(x) dx, & \text{if } X \text{ continuous,} \end{cases}$$

where $f$ is either the pmf or the pdf. The definition holds provided that $\int_{-\infty}^{+\infty} |x|^r \mathrm{d}F(x) < \infty$.

**Central moment**: given a random variable $X$, the $r^{th}$ central moment, $m_r$ is

$$m_r = \mathrm{E}[(X - \mu_1)^r].$$

The definition holds provided that $\int_{-\infty}^{+\infty} |x|^r \mathrm{d}F(x) < \infty$. so if the $r$-th moment exists, then also the $r$-th central moment exists.

**Properties of moments**:

1. mean: $\mu_1 = \mathrm{E}[X] = \mu$ and $m_1 = 0$;

2. variance: $m_2 = \mathrm{E}[(X - \mu_1)^2] = \mathrm{Var}[X] = \sigma^2$;

3. coefficient of skewness: $\gamma = \mathrm{E}[(X - \mu_1)^3]/\sigma^3 = m_3/m_2^{\frac{3}{2}}$;

4. coefficient of kurtosis: $\kappa = (\mathrm{E}[(X - \mu_1)^4]/\sigma^4) = (m_4/m_2^2)$.

What would a distribution with postive skew and large kurtosis look like?

## 6.2 Inequalities involving expectations

**A general inequality**: let $X$ be a random variable with $X \geq 0$ and let $g$ be a positive increasing function on $\mathbb{R}^+$, then, for any $a > 0$,

$$P(g(X) \geq a) \leq \frac{\mathrm{E}[g(X)]}{a}.$$

There are two special cases.

1. **Markov's inequality**: let $X$ be a random variable with $X \geq 0$ and $\mathrm{E}[X]$ defined, then, for any $a > 0$,
$$P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}.$$
2. **Chebyshev's inequality**: let $X$ be a random variable with $\mathrm{E}[X^2] < \infty$, then, for any $a > 0$,
$$P((X - \mathrm{E}[X])^2 \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}.$$

**Jensen's inequality**: If $X$ is a random variable with $\mathrm{E}[X]$ defined, and $g$ is a convex function with $\mathrm{E}[g(X)]$ defined, then

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X]).$$

## 6.3 Moment generating functions

These are functions that help to compute moments of a distribution and are also useful to characterie the distribution. However, it can be shown that the moments do not characterise the distribution uniquely (if you would like to know more about this, check the log-normal distribution).

**Moment generating function (mgf)**: given a random variable $X$, it is a function

$$M : \mathbb{R} \longrightarrow [0, \infty) \text{ s.t. } M(t) = \mathrm{E}[e^{tX}],$$

where it is assumed $M(t) < \infty$ for $|t| < h$ and some $h > 0$, i.e. the expectation exists in a neighborhood of $0$. Therefore,

$$M(t) = \int_{-\infty}^{+\infty} e^{tx} \mathrm{d}F(x) = \begin{cases} \sum_i e^{tx_i} f(x_i), & \text{if } X \text{ discrete,} \\ \int_{-\infty}^{+\infty} e^{tx} f(x) \mathrm{d}x, & \text{if } X \text{ continuous.} \end{cases}$$

**Properties of mgfs**: if $X$ has mgf $M(t)$ then

1. Taylor expansion:

$$M(t) = 1 + t\mathrm{E}[X] + \frac{t^2}{2!}\mathrm{E}[X^2] + \ldots + \frac{t^r}{r!}\mathrm{E}[X^r] + \ldots = \sum_{j=0}^{\infty} \frac{\mathrm{E}[X^j]}{j!} t^j;$$

2. the $r^{\text{th}}$ moment is the coefficient of $t^r/r!$ in the Taylor expansion;

3. derivatives at zero:

$$\mu_r = \mathrm{E}[X^r] = M^{(r)}(0) = \left.\frac{\mathrm{d}^r}{\mathrm{d}t^r} M(t)\right|_{t=0}.$$

**Proof**: by differentiating $M(t)$ (in a neighbourhood of $0$ assuming existence)

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} M(t) &= \frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{+\infty} e^{tx} \mathrm{d}F(x) = \\
&= \int_{-\infty}^{+\infty} \frac{\mathrm{d}}{\mathrm{d}t} e^{tx} \mathrm{d}F(x) = \\
&= \int_{-\infty}^{+\infty} x e^{tx} \mathrm{d}F(x) = \\
&= \mathrm{E}[X e^{tX}],
\end{aligned}$$

and in general

$$\frac{\mathrm{d}^r}{\mathrm{d}t^r} M(t) = \mathrm{E}[X^r e^{tX}],$$

by imposing $t = 0$ we get the desired result.

**Uniqueness**: let $F_X$ and $F_Y$ be two cdfs with all moments defined, then:

1. if $X$ and $Y$ have bounded support, then $F_X(x) = F_Y(x)$ for any $x \in \mathbb{R}$ if and only if $\mathrm{E}[X^r] = \mathrm{E}[Y^r]$ for any $r \in \mathbb{N}$;

2. if the mgfs exist and $M_X(t) = M_Y(t)$ for all $|t| < h$ and some $h > 0$, then $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$.

**Cumulant generating function (cgf)**: given a random variable $X$ with moment generating function $M(t)$, it is defined as

$$K(t) = \log M(t).$$

**Cumulant**: the $r^{\text{th}}$ cumulant, $c_r$, is the coefficient of $t^r/r!$ in the Taylor expansion of the cumulant generating function $K(t)$:

$$c_r = K^{(r)}(0) = \left.\frac{\mathrm{d}^r}{\mathrm{d}t^r} K(t)\right|_{t=0}.$$

**Properties of cgfs**:

1. $c_1 = \mu_1 = \mu$ (mean, first moment);
2. $c_2 = m_2 = \sigma^2$ (variance, second central moment);
3. $c_3 = m_3$ (third central moment);
4. $c_4 + 3c_2^2 = m_4$ (fourth central moment).

# Reading

Casella and Berger, Sections 2.2 - 2.3.

# 7 Distributions

## 7.1 Discrete distributions

**Degenerate**: all probability concentrated in a single point $a$.

- $f(x) = 1 \quad$ for $x = a$.
- $M(t) = e^{at}, \quad K(t) = at$.
- $\mu = a, \quad \sigma^2 = 0$.

**Bernoulli**: trials with two, and only two, possible outcomes, here labeled $X = 0$ (failure) and $X = 1$ (success).

- $f(x) = p^x(1-p)^{1-x} \quad$ for $x = 0,\ 1$.
- $M(t) = 1 - p + pe^t, \quad K(t) = \log(1 - p + pe^t)$.
- $\mu = p, \quad \sigma^2 = p(1-p)$.

**Binomial**: we want to count the number of successes in $n$ independent Bernoulli trials, each with probability of success $p$. Consider $n$ random variables $Y_i$ with just two possible outcomes $Y_i = 0,\ 1$, their sum $X = \sum_{i=1}^{n} Y_i$ is the total number of successes in $n$ trials, so $0 \leq X \leq n$. Notation $X \sim \text{Bin}(n, p)$. We need to count all the possible ways to extract $x$ numbers out of $n$ and multiply this number for the probability of success given by the Bernoulli distribution.

- $f(x) = \frac{n!}{x!(n-x)!} p^x(1-p)^{n-x} = \binom{n}{x} p^x(1-p)^{n-x} \quad$ for $x = 0,\ 1, \ldots,\ n$.
- $M(t) = (1 - p + pe^t)^n, \quad K(t) = n\log(1 - p + pe^t)$.
- $\mu = np, \quad \sigma^2 = np(1-p)$.

The Bernoulli distribution is equivalent to a binomial distribution with $n = 1$.

Examples: tossing a coin $n$ times and counting the number of times we get head (or tail); $n$ steps in the random walk and counting the steps to the right (or to the left).
Suppose to roll a die $k$ times and we want the probability of obtaining at least one 3. So we have $k$ Bernoulli trials with success probability $p = 1/6$. Define the random variable $X$ that counts the total number of 3 in $k$ rolls, then $X \sim \text{Bin}(k, 1/6)$ and

$$P(\text{at least one 3}) = P(X > 0) = 1 - P(X = 0) = 1 - \binom{k}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^k = 1 - \left(\frac{5}{6}\right)^k$$

If, by throwing two dice, we were interested in the probability of at least double 3 we would get

$$P(\text{at least one double 3}) = 1 - \left(\frac{35}{36}\right)^k < P(\text{at least one 3}),$$

since $35/36 > 5/6$.

**Computing the moments and mgf of the binomial distribution**: just notice that a binomial random variable $X$ is the sum of $n$ Bernoulli independent random variables $Y_i$, each with mean $\text{E}[Y_i] = p$ and variance $\text{Var}[X] = p(1-p)$ hence

$$\text{E}[X] = \text{E}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \text{E}[Y_i] = np,$$

and (independence is crucial here)

$$\text{Var}[X] = \sum_{i=1}^{n} \text{Var}[Y_i] = np(1-p).$$

The mgf is computed as

$$M(t) = \sum_{x=0}^{n} e^{tx}\binom{n}{k}p^x(1-p)^{n-x} = \sum_{x=0}^{n}\binom{n}{k}(pe^t)^x(1-p)^{n-x}$$

use the binomial expansion

$$(u+v)^n = \sum_{x=0}^{n}\binom{n}{x}u^x v^{n-x}$$

and by substituting $u = pe^t$ and $v = 1-p$ we get

$$M(t) = (pe^t + 1 - p)^n.$$

**Negative Binomial**: we want to count the number of Bernoulli trials necessary to get a fixed number of successes (i.e. a waiting time). Consider a random variable $X$ denoting the trial at which the $r^{\text{th}}$ success occurs. We want the distribution of the event $\{X = x\}$ for $x = r, r+1, \ldots$. This event occurs only if we had $r-1$ successes in $x-1$ trials and a success at the $x^{\text{th}}$ trial. By multiplying these probabilities we have

- $f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$ for $x = r,\ r+1, \ldots$.
- $M(t) = \left( \frac{pe^t}{1-(1-p)e^t} \right)^r, \quad K(t) = -r \log\{(1-\frac{1}{p}) + \frac{1}{p} e^{-t}\}$ for $|t| < -\log(1-p)$.
- $\mu = \frac{r}{p}, \quad \sigma^2 = \frac{r(1-p)}{p^2}$.

It is also the defined in terms of the number of failures before the $r^{\text{th}}$ success.

**Geometric**: to count the number of Bernoulli trials before the first success occurs. Equivalent to a negative binomial with $r = 1$.

- $f(x) = (1-p)^{x-1} p$ for $x = 1,\ 2, \ldots$.
- $M(t) = \frac{pe^t}{1-(1-p)e^t}, \quad K(t) = -\log\{(1-\frac{1}{p}) + \frac{1}{p} e^{-t}\}$ for $|t| < -\log(1-p)$.
- $\mu = \frac{1}{p}, \quad \sigma^2 = \frac{1-p}{p^2}$.

This distribution is memoryless, indeed, if $X$ follows a geometric distribution, then, for integers $s > t$,

$$\begin{aligned} P(X > s | X > t) &= \frac{P(X > s \cap X > t)}{P(X > t)} = \frac{P(X > s)}{P(X > t)} \\ &= (1-p)^{s-t} = P(X > s - t), \end{aligned}$$

Given that we observed $t$ failures we observe an additional $s - t$ failures with the same probability as we observed $s - t$ failures at the beginning of the experiment. The only thing that counts is the length of the sequence of failures not its position.

**Hypergeometric**: it is usually explained with the example of the urn model. Assume to have an urn containing a total of $N$ balls made up of $N_1$ balls of type 1 and $N_2 = N - N_1$ balls of type 2, we want to count the number of type 1 balls chosen when selecting $n < N$ balls without replacement from the urn.

- $f(x) = \binom{N_1}{x} \binom{N-N_1}{n-x} / \binom{N}{n}$ for $x \in \{0, \ldots, n\} \cap \{n - (N - N_1), \ldots, N_1\}$.
- $\mu = n\frac{N_1}{N}, \quad \sigma^2 = n\frac{N_1}{N}\frac{N-N_1}{N}\frac{N-n}{N-1}$.

**Uniform**: for experiments with $N$ equally probable outcomes

- $f(x) = \frac{1}{N}$ for $x = 1,\ 2, \ldots, N$.
- $\mu = \frac{N+1}{2}, \quad \sigma^2 = \frac{N^2-1}{12}$.

**Poisson**: to count the number of events which occur in an interval of time. The assumption is that for small time intervals the probability of an occurrence is proportional to the length of the waiting time between two occurrences. We consider the random variable $X$ which counts the number of occurrences of a given event in a given unit time interval, it depends on a parameter $\lambda$ which is the intensity of the process considered. Notation $\mathrm{Pois}(\lambda)$.

- $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0,\ 1, \ldots$.
- $M(t) = e^{\lambda(e^t-1)}, \quad K(t) = \lambda(e^t - 1)$.
- $\mu = \lambda, \quad \sigma^2 = \lambda$.

The intensity is the average number of occurrences in a given unit time interval. Notice that the Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

Example: think of crossing a busy street with an average of 300 cars per hour passing. In order to cross we need to know the probability that in the next minute no car passes. In a given minute we have an average of $\lambda = 300/60 = 5$ cars passing through. If $X$ is the number of cars passing in one minute we have

$$P(X = 0) = \frac{e^{-5}5^0}{0!} = 6.7379 \cdot 10^{-3},$$

maybe is better to cross the street somewhere else. Notice that $\lambda$ has to be the intensity per unit of time. If we are interested in no cars passing in one hour then $\lambda = 300$ and clearly the probability would be even smaller. If we want to know the average number of cars passing in 5 minutes time then just define a new random variable $X$ which counts the cars passing in 5 minutes, which is distributed as Poisson with $\lambda = 300/12 = 25$ and this is also the expected value.

**The Poisson approximation**: if $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Pois}(\lambda)$ with $\lambda = np$, then for large $n$ and small $p$ we have $P(X = x) \simeq P(Y = x)$. More rigorously we have to prove that for finite $\lambda = np$

$$\lim_{n\to\infty} F_X(x; n, p) = F_Y(x; \lambda)$$

we can use mgfs and prove equivalently that

$$\lim_{n\to\infty} M_X(t; n, p) = \lim_{n\to\infty} (1 - p + pe^t)^n = e^{\lambda(e^t-1)} = M_Y(t; \lambda).$$

**Proof**: we can use the following result: given a sequence of real numbers s.t. $a_n \to a$ for $n \to \infty$, then

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

Now

$$\begin{aligned} \lim_{n\to\infty} M_X(t; n, p) &= \lim_{n\to\infty} (1 - p + pe^t)^n = \lim_{n\to\infty} \left(1 + \frac{1}{n}(e^t - 1)np\right)^n = \\ &= \lim_{n\to\infty} \left(1 + \frac{1}{n}(e^t - 1)\lambda\right)^n = e^{\lambda(e^t-1)}. \end{aligned}$$

## 7.2 Continuous distributions

**Uniform**: a random number chosen from a given closed interval $[a,\ b]$. Notation $U(a, b)$.

- $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$.
- $M(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}$ for $t \neq 0$ and $M(0) = 1$.
- $\mu = \frac{a+b}{2}, \quad \sigma^2 = \frac{(b-a)^2}{12}$.

**Normal or Gaussian**: this is the most important distribution. Notation $N(\mu, \sigma^2)$.

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$ for $-\infty < x < \infty$.
- $M(t) = e^{\mu t+\sigma^2 t^2/2}$.
- $\mathrm{E}[X] = \mu, \quad \mathrm{Var}[X] = \sigma^2$.

If $X \sim N(\mu, \sigma^2)$ then $(X - \mu)/\sigma = Z \sim N(0, 1)$, is a standard normal distribution, i.e. with zero mean and unit variance. We can use the moments of $Z$ to compute the moments of $X$, indeed

$$\mathrm{E}[X] = \mathrm{E}[\mu + \sigma Z] = \mu, \qquad \mathrm{Var}[X] = \mathrm{Var}[\mu + \sigma Z] = \sigma^2.$$

The shape of $f(x)$ is symmetric around $\mu$ with inflection points at $\mu \pm \sigma$. A statistical table (in the past) or a computer programme (nowadays) can be used to calculate the distribution function. The following values will be useful later on:

$$\begin{aligned} P(|X - \mu| \leq \sigma) &= P(|Z| \leq 1) = .6826, \\ P(|X - \mu| \leq 2\sigma) &= P(|Z| \leq 2) = .9544, \\ P(|X - \mu| \leq 3\sigma) &= P(|Z| \leq 3) = .9974. \end{aligned}$$

In particular, the so-called two-sigma rule states that (roughly) $95\%$ (in a repeated sample) of the data from a normal distribution falls within two standard deviations of its mean.

The normal distribution is characterized by just its first two moments. We can compute higher order moments by using the following relation (holding for any differentiable function $g(X)$) for $X \sim N(\mu, \sigma^2)$:

$$\mathrm{E}[g(X)(X - \mu)] = \sigma^2 \mathrm{E}[g'(X)].$$

Check this (hint: use integration by parts).

From the above relation we have that all moments of a normal distribution are computable starting from the second central moment. Moreover, for a standard normal random variable $Z$ all moments of odd order are zero, in particular

$$\begin{aligned} \text{skewness} \quad \gamma &= \frac{\mathrm{E}[Z^3]}{\mathrm{E}[Z^2]^{3/2}} = \mathrm{E}[Z^2 Z] = \mathrm{E}[2Z] = 0, \\ \text{kurtosis} \quad \kappa &= \frac{\mathrm{E}[Z^4]}{\mathrm{E}[Z^2]^2} = \mathrm{E}[Z^3 Z] = \mathrm{E}[3Z^2] = 3. \end{aligned}$$

The skewness coefficient measures the asymmetry and indeed is zero for the normal, and the kurtosis coefficient measures flatness of the tails, usually we are interested in the coefficient of excess kurtosis (with respect to the normal case), i.e. $\kappa - 3$.

**Computing moments and mgf of the standard normal distribution**: the mgf is computed as

$$\begin{aligned} M(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}+tz} \mathrm{d}z = \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2-2tz+t^2}{2}} e^{t^2/2} \mathrm{d}z = \\ &= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(z-t)^2}{2}} \mathrm{d}z = e^{\frac{t^2}{2}}. \end{aligned}$$

The Taylor expansion of $M(t)$ is

$$\begin{aligned} M(t) &= 1 + 0 + \frac{t^2}{2} + 0 + \frac{t^4}{2^2 2!} + \ldots = \sum_{j=0}^{+\infty} \frac{t^{2j}}{2^j j!} = \\ &= 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} + \ldots = \sum_{r=0}^{+\infty} \mu_r \frac{t^r}{r!}, \end{aligned}$$

hence the moments of $Z$ (which in this case are equal to the central moments) are for $r = 0, 1, 2, \ldots$

$$\begin{aligned} \mu_{2r+1} &= \mathrm{E}[Z^{2r+1}] = 0, \\ \mu_{2r} &= \mathrm{E}[Z^{2r}] = \frac{(2r)!}{2^r r!}. \end{aligned}$$

**Gamma**: is a family of distributions characterized by parameters $\alpha > 0$ and $\theta$. We need the gamma function defined by

$$\Gamma(t) = \int_0^{\infty} y^{t-1} e^{-y} \mathrm{d}y, \quad \text{for } t > 0.$$

Properties of the gamma function $\Gamma(t) = (t-1)\Gamma(t-1)$ for $t > 1$ and $\Gamma(n) = (n-1)!$ for positive integer $n$. Notation for the gamma distribution; $\mathrm{Gamma}(\alpha, \theta)$ or $G(\alpha, \theta)$.

- $f(x) = \frac{1}{\Gamma(\alpha)} \theta^{\alpha} x^{\alpha-1} e^{-\theta x}$ for $0 \leq x < \infty$.
- $M(t) = \frac{1}{(1-t/\theta)^{\alpha}}$ for $t < \theta$.
- $\mu = \alpha/\theta, \quad \sigma^2 = \alpha/\theta^2$.

$\alpha$ is the shape parameter determining if the distribution has a peak or it is monotonically decreasing, while $\theta$ is the scale parameter influencing the spread of the distribution hence its peak location.

**Chi-square**: if $Z_j$ are independent standard normal, then $X = \sum_{j=1}^{r} Z_j^2$ has a chi-square distribution with $r$ degrees of freedom. Notation $\chi_r^2$ or $\chi^2(r)$. Equivalent to a gamma distribution with $\alpha = r/2$ and $\theta = 1/2$.

- $f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2}$ for $0 \leq x < \infty$.
- $M(t) = \frac{1}{(1-2t)^{r/2}}$ for $t < 1/2$.
- $\mu = r, \quad \sigma^2 = 2r$.

**Exponential**: waiting time between events distributed as Poisson with intensity $\theta$. Notation $\mathrm{Exp}(\theta)$ (somewhat ambiguous). Equivalent to a gamma distribution with $\alpha = 1$.

- $f(x) = \theta e^{-\theta x}$ for $0 \leq x < \infty$.
- $M(t) = \frac{\theta}{\theta - t}$ for $t < \theta$.
- $\mu = 1/\theta, \quad \sigma^2 = 1/\theta^2$.

It is a memoryless distribution, indeed if $X \sim \mathrm{Exp}(\theta)$, then for integers $s > t$,

$$\begin{aligned} P(X > s | X > t) &= \frac{P(X > s \cap X > t)}{P(X > t)} = \frac{P(X > s)}{P(X > t)} \\ &= \frac{\int_s^{+\infty} \theta e^{-x\theta} \mathrm{d}x}{\int_t^{+\infty} \theta e^{-x\theta} \mathrm{d}x} = e^{-(s-t)\theta} = P(X > s - t). \end{aligned}$$

Example: it is used in modeling survival rates (see below).

**Log-normal**: it is the distribution of a random variable $X$ such that $\log X \sim N(\mu, \sigma^2)$. It is used for random variables with positive support, and it is very similar to, although less flexible, and more analytically tractable than the gamma distribution.

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} e^{-(\log x - \mu)^2/(2\sigma^2)}$ for $0 < x < \infty$.
- $\mathrm{E}[X] = e^{\mu + \sigma^2/2}, \quad \mathrm{Var}[X] = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$.

Notice that in this case $M(t)$ is not defined (see ex. 2.36 Casella & Berger). Examples are the distributions of income or consumption. This choice allows to model the logs of income and consumption by means of the normal distribution which is the distribution predicted by economic theory.

## 7.3 Survival and hazard

**Survival function**: given a continuous non-negative random variable $X$, it is the function

$$\bar{F}(x) = 1 - F(x) = P(X > x).$$

where $x$ is interpreted as a threshold and we are interested in the probability of having realizations of $X$ beyond $x$. We usually assume that $\bar{F}(0) = 1$.

In the context of survival analysis the cdf and the pdf are called lifetime distribution function and event density, respectively.

**Hazard function or hazard rate**: it is the probability of having a realization of $X$ in a small interval beyond the threshold $x$, i.e. conditional on survival of $X$ beyond $x$:

$$h(x) = \lim_{\varepsilon \to 0^+} \frac{\bar{F}(x+\varepsilon) - \bar{F}(x)}{\varepsilon \bar{F}(x)} = \lim_{\varepsilon \to 0^+} \frac{P(X \leq x + \varepsilon | X > x)}{\varepsilon},$$

it is then defined as

$$h(x) = \frac{f(x)}{\bar{F}(x)} = -\frac{\bar{F}'(x)}{\bar{F}(x)}.$$

Its relationship with cdf is:

$$h(x) = -\frac{\mathrm{d}}{\mathrm{d}x} \log(1 - F(x)), \qquad F(x) = 1 - \exp\left(-\int_0^x h(u)\mathrm{d}u\right).$$

# Reading

Casella and Berger, Sections 3.1 - 3.2 - 3.3.

# 8 Multivariate distributions

## 8.1 Bivariate joint and marginal distributions

For simplicity we first give the definitions for the bivariate case and then we generalise to the $n$-dimensional setting.

**Joint cumulative distribution function**: for two random variables $X$ and $Y$ the joint cdf is a function $F_{X,Y} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ such that

$$F(x, y) = P(X \leq x, Y \leq y).$$

**Properties of joint cdf**:

1. $F_{X,Y}(-\infty, y) = \lim_{x\to-\infty} F_{X,Y}(x, y) = 0$,
   $F_{X,Y}(x, -\infty) = \lim_{y\to-\infty} F_{X,Y}(x, y) = 0$,
   $F_{X,Y}(+\infty, +\infty) = \lim_{x\to+\infty, y\to+\infty} F_{X,Y}(x, y) = 1$;

2. Right continuous in $x$: $\lim_{h\to0^+} F_{X,Y}(x + h, y) = F_{X,Y}(x, y)$,
   Right continuous in $y$: $\lim_{h\to0^+} F_{X,Y}(x, y + h) = F_{X,Y}(x, y)$.

3. For any $y$, the function $F(x, y)$ is non-decreasing in $x$.
   For any $x$, the function $F(x, y)$ is non-decreasing in $y$.

We are interested in the probability that $X$ and $Y$ take values in a given (Borel !) subset of the plane $\mathbb{R} \times \mathbb{R} \equiv \mathbb{R}^2$. The simplest is a rectangular region $A = \{(x, y) \in \mathbb{R}^2 \text{ s.t. } x_1 < x \leq x_2 \text{ and } y_1 < x \leq y_2\}$. Then

$$\begin{aligned} P(A) &= P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \\ &= F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) - [F_{X,Y}(x_2, y_1) - F_{X,Y}(x_1, y_1)]. \end{aligned}$$

**Marginal cumulative distribution functions**: if $F_{X,Y}$ is the joint distribution function of $X$ and $Y$ then the marginal cdfs are the usual cdfs of the single random variables and are given by

$$\begin{aligned} F_X(x) &= \lim_{y\to\infty} F_{X,Y}(x, y) = F_{X,Y}(x, \infty), \\ F_Y(y) &= \lim_{x\to\infty} F_{X,Y}(x, y) = F_{X,Y}(\infty, y). \end{aligned}$$

Marginal cdfs are generated from the joint cdf, but the reverse is not true. The joint cdf contains information that is not captured in the marginals. In particular it tells us about the dependence structure among the random variables, i.e. how they are associated.

## 8.2 Bivariate joint and marginal pmf and pdf

**Joint probability mass function**: for two discrete random variables $X$ and $Y$ it is a function $f_{X,Y} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ such that

$$f_{X,Y}(x, y) = P(X = x, Y = y) \quad \forall\, x, y \in \mathbb{R}.$$

In general

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \sum_{x_1<x\leq x_2} \sum_{y_1<y\leq y_2} f_{X,Y}(x, y).$$

**Marginal probability mass functions**: for two discrete random variables $X$ and $Y$, with range $\{x_1, x_2, \ldots\}$ and $\{y_1, y_2, \ldots\}$ respectively, the marginal pmfs are

$$\begin{aligned} f_X(x) &= \sum_{y\in\{y_1,y_2,\ldots\}} f_{X,Y}(x, y) \\ f_Y(y) &= \sum_{x\in\{x_1,x_2,\ldots\}} f_{X,Y}(x, y). \end{aligned}$$

**Joint probability density function**: for two jointly continuous random variables $X$ and $Y$, it is an integrable function $f_{X,Y} : \mathbb{R} \times \mathbb{R} \to [0, +\infty)$ such that

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v) \mathrm{d}u \mathrm{d}v \quad \forall\, x, y \in \mathbb{R},$$

notice that this implies

$$f_{X,Y}(x, y) = \left. \frac{\partial^2 F_{X,Y}(u, v)}{\partial u \partial v} \right|_{u=x, v=y},$$

**Properties of joint pdf**:

1. $f_{X,Y}(x, y) \geq 0$ for any $x, y \in \mathbb{R}$;
2. normalisation:
$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y = 1;$$
3. probability of a rectangular region:
$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y;$$
4. for any (Borel) set $B \subseteq R^2$ the probability that $(X, Y)$ takes values in $B$ is
$$P(B) = \int \int_{B} f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y.$$

In the one-dimensional case events are usually intervals of $\mathbb{R}$ and their probability is proportional to their length, in two-dimensions events are regions of the plane $\mathbb{R}^2$ and their probability is proportional to their area, in three-dimensions events are regions of the space $\mathbb{R}^3$ and their probability is proportional to their volume. Lengths, areas and volumes are weighted by the frequencies of the outcomes which are part of the considered events hence they are areas, volumes and 4-d volumes under the pdfs. Probability is the measure of events with respect to the measure of the whole sample space which is 1 by definition.

**Marginal probability density functions**: for two jointly continuous random variables $X$ and $Y$, they are integrable functions $f_X : \mathbb{R} \to [0, +\infty)$ and $f_Y : \mathbb{R} \to [0, +\infty)$ such that

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f_{X,Y}(x, y) \mathrm{d}y, \quad \forall\, x \in \mathbb{R}, \\ f_Y(y) &= \int_{-\infty}^{+\infty} f_{X,Y}(x, y) \mathrm{d}x, \quad \forall\, y \in \mathbb{R}. \end{aligned}$$

Therefore, the marginal cdfs are

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} \int_{-\infty}^{+\infty} f_{X,Y}(u, y) \mathrm{d}y \mathrm{d}u, \quad \forall\, x \in \mathbb{R}, \\ F_Y(y) &= \int_{-\infty}^{y} \int_{-\infty}^{+\infty} f_{X,Y}(x, v) \mathrm{d}x \mathrm{d}v, \quad \forall\, y \in \mathbb{R}. \end{aligned}$$

## 8.3 Cdf, pmf, and pdf of $n$ random variables

**Multivariate generalization**: for $n$ random variables $X_1, \ldots, X_n$ we have analogous definitions:

1. the joint cdf is a function $F_{X_1,\ldots,X_n} : \mathbb{R}^n \to [0,1]$ such that

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots X_n \leq x_n);$$

2. the marginal cdfs are, for any $j = 1, \ldots, n$, the functions

$$F_{X_j}(x_j) = F_{X_1,\ldots,X_n}(\infty, \ldots, \infty, x_j, \infty, \ldots, \infty);$$

3. the marginal pmf or pdf are, for any $j = 1, \ldots, n$, the functions

$$f_{X_j}(x_j) = \begin{cases} \sum_{x_1} \cdots \sum_{x_{j-1}} \sum_{x_{j+1}} \cdots \sum_{x_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n), & \text{discrete case,} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_{j-1} dx_{j+1} \ldots dx_n, & \text{continuous case;} \end{cases}$$

4. if $g$ is a well-behaved function $g : \mathbb{R}^n \to \mathbb{R}$, then

$$\mathrm{E}[g(X_1, \ldots, X_n)] = \begin{cases} \sum_{x_1} \cdots \sum_{x_n} g(x_1, \ldots, x_n) f_{X_1,\ldots,X_n}(x_1, \ldots, x_n), & \text{discrete,} \\ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(x_1, \ldots, x_n) f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n, & \text{continuous.} \end{cases}$$

# Reading

Casella and Berger, Section 4.1.

# 9 Independence of random variables

## 9.1 Pairwise independence

Besides the usual univariate measures of location (mean) and scale (variance), in the multivariate case we are interested in measuring the dependence among random variables.

**Joint cdf of independent random variables**: two random variables $X$ and $Y$ are independent if and only if the events $\{X \leq x\}$, $\{Y \leq y\}$ are independent for all choices of $x$ and $y$, i.e., for all $x, y \in \mathbb{R}$,

$$\begin{aligned} P(X \leq x, Y \leq y) &= P(X \leq x) P(Y \leq y), \\ F_{X,Y}(x, y) &= F_X(x) F_Y(y). \end{aligned}$$

**Joint pmf or pdf of independent random variables**: two random variables $X$ and $Y$ are independent if and only if, for all $x, y \in \mathbb{R}$,

$$f_{X,Y} = f_X(x) f_Y(y).$$

The two above are necessary and sufficient conditions, while the following is just necessary conditions but not sufficient (see also below the distinction between independence and uncorrelation).

**Expectation and independence**: if $X$ and $Y$ are independent then

$$\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y].$$

Moreover, if $g_1$ and $g_2$ are well-behaved functions then also $g_1(X)$ and $g_2(Y)$ are independent random variables, hence

$$\mathrm{E}[g_1(X)g_2(Y)] = \mathrm{E}[g_1(X)]\mathrm{E}[g_2(Y)].$$

## 9.2 Independence of $n$ random variables

**Multivariate generalisation**: in the $n$-dimensional case we have analogous definitions:

1. the random variables $X_1,\ X_2, \ldots,\ X_n$ are mutually independent if and only if the events $\{X_1 \leq x_1\},\ \{X_2 \leq x_2\}, \ldots,\ \{X_n \leq x_n\}$ are independent for all choices of $x_1, x_2, \ldots, x_n \in \mathbb{R}$:

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \ldots F_{X_n}(x_n) = \prod_{j=1}^{n} F_{X_j}(x_j);$$

2. $X_1, X_2, \ldots, X_n$ are mutually independent if and only if $x_1, x_2, \ldots, x_n$:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n) = \prod_{j=1}^{n} f_{X_j}(x_j).$$

3. if $X_1, X_2, \ldots, X_n$ are mutually independent then

$$\mathrm{E}[X_1, X_2, \ldots, X_n] = \mathrm{E}[X_1]\mathrm{E}[X_2] \ldots \mathrm{E}[X_n] = \prod_{j=1}^{n} \mathrm{E}[X_j],$$

and if $g_1, g_2, \ldots, g_n$ are well-behaved functions then also $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ are mutually independent random variables, hence

$$\mathrm{E}[g_1(X_1)g_2(X_2) \ldots g_n(X_n)] = \mathrm{E}[g_1(X_1)]\mathrm{E}[g_2(X_2)] \ldots \mathrm{E}[g_n(X_n)] = \prod_{j=1}^{n} \mathrm{E}[g_j(X_j)].$$

## 9.3 Measures of pairwise dependence

**Covariance function**: for two random variables $X$ and $Y$ we define

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])],$$

which is equivalent to

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y].$$

**Properties of covariance**:

1. symmetry: $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$;
2. bilinearity

   $$\mathrm{Cov}(X_1+X_2, Y_1+Y_2) = \mathrm{Cov}(X_1,Y_1)+\mathrm{Cov}(X_1,Y_2)+\mathrm{Cov}(X_2,Y_1)+\mathrm{Cov}(X_2,Y_2),$$

   and for any $a, b \in \mathbb{R}$

   $$\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X,Y);$$

3. relationship with variance: $\mathrm{Var}[X] = \mathrm{Cov}(X,X)$,
   $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}(X,Y)$,
   $\mathrm{Var}[X-Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] - 2\mathrm{Cov}(X,Y)$;
4. if $X$ and $Y$ are independent, $\mathrm{Cov}(X,Y) = 0$.

**Correlation coefficient**: for random variables $X$ and $Y$,

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}.$$

Correlation is the degree of linear association between two variables. It is a scaled covariance, $|\mathrm{Corr}(X,Y)| \leq 1$. Moreover, $|\mathrm{Corr}(X,Y)| = 1$ if and only if there exist numbers $a \neq 0$ and $b$ such that $P(Y = aX + b) = 1$ ( a linear relation between variables). If $\mathrm{Corr}(X,Y) = 1$ then $a > 0$, if $\mathrm{Corr}(X,Y) = -1$ then $a < 0$.

**Uncorrelation and independence**: $\mathrm{Corr}(X,Y) = 0$, i.e. $X$ and $Y$ are uncorrelated, if and only if

$$\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y].$$

This result implies that

$$X, Y \text{ independent } \Rightarrow X, Y \text{ uncorrelated.}$$

but not the viceversa. Indeed correlation means only linear dependence.

**Example**: we know that independence implies

$$\mathrm{E}[g_1(X)g_2(Y)] = \mathrm{E}[g_1(X)]\mathrm{E}[g_2(Y)].$$

for any $g_1$, $g_2$ well-behaved functions. Consider the discrete random variables $X$ and $Y$ such that the joint pmf is

$$f_{X,Y}(x,y) = \begin{cases} 1/4 & \text{if} \quad x=0 \quad \text{and} \quad y=1 \\ 1/4 & \text{if} \quad x=0 \quad \text{and} \quad y=-1 \\ 1/4 & \text{if} \quad x=1 \quad \text{and} \quad y=0 \\ 1/4 & \text{if} \quad x=-1 \quad \text{and} \quad y=0 \\ 0 & \text{otherwise} \end{cases}$$

Now, $\mathrm{E}[XY] = 0$ and $\mathrm{E}[X] = \mathrm{E}[Y] = 0$, thus $\mathrm{Cov}(X,Y) = 0$, the variables are uncorrelated. If we now choose $g_1(X) = X^2$ and $g_2(Y) = Y^2$ we have $\mathrm{E}[g_1(X)g_2(Y)] = \mathrm{E}[X^2Y^2] = 0$, but

$$\mathrm{E}[g_1(X)]\mathrm{E}[g_2(Y)] = \frac{1}{2}\frac{1}{2} = \frac{1}{4} \neq 0.$$

So $X$ and $Y$ are not independent.

**Example**: suppose $X$ is a standard normal random variable, i.e. with $\mathrm{E}[X^k] = 0$ for $k$ odd, and let $Y = X^2$. Clearly $X$ and $Y$ are not independent: if you know $X$, you also know $Y$. And if you know $Y$, you know the absolute value of $X$. The covariance of $X$ and $Y$ is

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = \mathrm{E}[X^3] - 0 \cdot \mathrm{E}[Y] = \mathrm{E}[X^3] = 0.$$

Thus $\mathrm{Corr}(X,Y) = 0$, and we have a situation where the variables are not independent, yet they have no linear dependence. A linear correlation coefficient does not encapsulate anything about the quadratic dependence of $Y$ upon $X$.

# 10 Multivariate moments

## 10.1 Joint moments and mgfs for two random variables

**Expectation of a function of two random variables**: if $g$ is a well-behaved function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $X$ and $Y$ are random variables with joint pmf or pdf function $f_{X,Y}$ then

$$\mathrm{E}[g(X,Y)] = \begin{cases} \sum_y \sum_x g(x,y) f_{X,Y}(x,y), & \text{discrete case,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y) \mathrm{d}x\mathrm{d}y, & \text{continuous case.} \end{cases}$$

**Joint moments**: if $X$ and $Y$ are random variables with joint pmf or pdf $f_{X,Y}$ then the $(r,s)^{\text{th}}$ joint moment is

$$\mu_{r,s} = \mathrm{E}[X^r Y^s] = \begin{cases} \sum_y \sum_x x^r y^s f_{X,Y}(x,y), & \text{discrete case,} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^r y^s f_{X,Y}(x,y) \mathrm{d}x\mathrm{d}y, & \text{continuous case.} \end{cases}$$

**Joint central moments**: the $(r,s)^{\text{th}}$ joint central moment is

$$\begin{aligned} m_{r,s} &= \mathrm{E}[(X-\mathrm{E}[X])^r(Y-\mathrm{E}[Y])^s] = \\ &= \begin{cases} \sum_y \sum_x [(x-\mu_X)^r(y-\mu_Y)^s] f_{X,Y}(x,y), & \text{discrete case,} \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{\infty}[(x-\mu_X)^r(y-\mu_Y)^s] f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y, & \text{continuous case.} \end{cases} \end{aligned}$$

**Properties of joint moments**:

1. mean for $X$: $\mu_{1,0} = E[X]$;
2. $r^{\text{th}}$ moment for $X$: $\mu_{r,0} = \mathrm{E}[X^r]$;
3. variance for $X$: $m_{2,0} = E[(X-\mathrm{E}[X])^2]$;
4. $r^{\text{th}}$ central moment for $X$: $m_{r,0} = \mathrm{E}[(X-\mathrm{E}[X])^r]$;
5. covariance: $m_{1,1} = \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])] = \mathrm{Cov}(X,Y)$;
6. correlation: $m_{1,1}/\sqrt{m_{2,0}m_{0,2}} = \mathrm{Corr}(X,Y)$.

**Joint moment generating function**: given two random variables $X$ and $Y$ is a function $M_{X,Y} : \mathbb{R}\times\mathbb{R} \to [0,+\infty)$ such that

$$M_{X,Y}(t,u) = \mathrm{E}[e^{tX+uY}] = \begin{cases} \sum_y \sum_x e^{tx+uy} f_{X,Y}(x,y), & \text{discrete case,} \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{\infty} e^{tx+uy} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y, & \text{continuous case.} \end{cases}$$

**Properties of joint mgfs**:

1. Taylor expansion:
$$M_{X,Y}(t,u) = \mathrm{E}\left[\sum_{i=0}^{+\infty}\frac{(tX)^i}{i!}\sum_{j=0}^{+\infty}\frac{(uY)^j}{j!}\right] = \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty}\mathrm{E}[X^iY^j]\frac{t^iu^j}{i!j!};$$
2. the $(r,s)^{\text{th}}$ joint moment is the coefficient of $t^ru^s/(r!s!)$ in the Taylor expansion;
3. derivatives at zero:
$$\mu_{r,s} = \mathrm{E}[X^rY^s] = M_{X,Y}^{(r,s)}(0,0) = \left.\frac{\mathrm{d}^{r+s}}{\mathrm{d}t^r\mathrm{d}u^s}M_{X,Y}(t,u)\right|_{t=0,u=0};$$
4. moment generating function for marginals: $M_X(t) = \mathrm{E}[e^{tX}] = M_{X,Y}(t,0)$, $M_Y(u) = \mathrm{E}[e^{uY}] = M_{X,Y}(0,u)$;
5. if $X$ and $Y$ independent:
$$M_{X,Y}(t,u) = M_X(t)M_Y(u).$$

**Joint cumulants**: let $K_{X,Y}(t,u) = \log M_{X,Y}(t,u)$ be the joint cumulant generating function, then we define the $(r,s)^{\text{th}}$ joint cumulant $c_{r,s}$ as the coefficient of $(t^ru^s)/(r!s!)$ in the Taylor expansion of $K_{X,Y}$. Thus,

$$\mathrm{Cov}(X,Y) = c_{1,1} \text{ and } \mathrm{Corr}(X,Y) = \frac{c_{1,1}}{\sqrt{c_{2,0}c_{0,2}}}.$$

## 10.2 Joint moments and mgfs of $n$ random variables

**Multivariate generalisation**: for random variables $X_1, \ldots, X_n$ with joint pmf or pdf $f_{X_1,\ldots,X_n}$:

1. joint moments:

$$\begin{aligned}\mu_{r_1,\ldots,r_n} =& \mathrm{E}[X_1^{r_1} \ldots X_n^{r_n}] \\ =& \begin{cases} \sum_{x_1} \ldots \sum_{x_n} x_1^{r_1} \ldots x_n^{r_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n), & \text{discrete case,} \\ \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} x_1^{r_1} \ldots x_n^{r_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n, & \text{continuous case;} \end{cases}\end{aligned}$$

2. joint central moments:

$$m_{r_1,\ldots,r_n} = \mathrm{E}[(X_1 - \mathrm{E}[X_1])^{r_1} \ldots (X_n - \mathrm{E}[X_n])^{r_n}].$$

3. joint moment generating function:

$$M_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = \mathrm{E}[e^{t_1 X_1 + \ldots + t_n X_n}],$$

   and the coefficient of $t_1^{r_1} \ldots t_n^{r_n}/(r_1! \ldots r_n!)$ in the Taylor expansion of $M_{X_1,\ldots,X_n}$ is $\mathrm{E}[X_1^{r_1} \ldots X_n^{r_n}]$;

4. independence: if $X_1, \ldots, X_n$ are independent then

$$M_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = M_{X_1}(t_1) \ldots M_{X_n}(t_n) = \prod_{j=1}^{n} M_{X_j}(t_j);$$

5. joint cumulant generating function:

$$K_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = \log(M_{X_1,\ldots,X_n}(t_1, \ldots, t_n)),$$

   and the $(r_1, \ldots, r_n)^{\text{th}}$ joint cumulant is defined as the coefficient of $(t_1^{r_1} \ldots t_n^{r_n})/(r_1! \ldots r_n!)$ in the Taylor expansion of $K_{X_1,\ldots,X_n}$.

## 10.3 Inequalities

**Hölder's inequality**: let $p$ and $q$ be two integers such that $\frac{1}{p} + \frac{1}{q} = 1$, if $X$ belongs to $L^p$ and $Y$ belongs to $L^q$, then $XY$ belong to $L^1$ and

$$\mathrm{E}[|XY|] \leq \mathrm{E}[|X|^p]^{1/p} \mathrm{E}[|Y|^q]^{1/q}.$$

**Cauchy-Schwarz's inequality**: this is Hölder's inequality when $p = q = 2$; if $X$ and $Y$ belong to $L^2$, then $XY$ belongs to $L^1$ and

$$\mathrm{E}[|XY|] \leq \sqrt{\mathrm{E}[X^2]\mathrm{E}[Y^2]}.$$

As a consequence, if $X$ and $Y$ have variances $\sigma_X^2$ and $\sigma_Y^2$, then

$$|\mathrm{Cov}(X,Y)| \leq \sigma_X \sigma_Y,$$

which means $|\mathrm{Corr}(X,Y)| \leq 1$.

**Minkowski's inequality**: let $p \geq 1$, if $X$ and $Y$ belong to $L^p$, then $X+Y$ belongs to $L^p$ and

$$\mathrm{E}[|X+Y|^p]^{1/p} \leq \mathrm{E}[|X|^p]^{1/p} + \mathrm{E}[|Y|^p]^{1/p}.$$

# Reading

Casella and Berger, Sections 4.2 - 4.5 - 4.7.

# 11 Conditional distributions

When we observe more than one random variable their values may be related. By considering conditional probabilities we can improve our knowledge of a given random variable by exploiting the information we have about the other.

**Conditional cumulative distribution function**: given $X$ and $Y$ random variables with $P(X=x) > 0$, the distribution of $Y$ conditional (given) to $X=x$ is defined as

$$F_{Y|X}(y|x) = P(Y \leq y | X = x).$$

It is a possibly different distribution for every value of $X$, we have a family of distributions.

**Conditional probability mass function**: given $X$ and $Y$ discrete random variables with $P(X=x) > 0$, the conditional pmf of $Y$ given $X=x$ is

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

such that the conditional cdf is

$$F_{Y|X}(y|x) = \sum_{y_i \leq y} f_{Y|X}(y_i|x).$$

**Conditional probability density function**: given $X$ and $Y$ jointly continuous random variables with $f_X(x) > 0$, the conditional pdf of $Y$ given $X=x$ is

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

such that the conditional cdf is

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} \frac{f_{X,Y}(x,v)}{f_X(x)} \mathrm{d}v.$$

**Conditional, joint and marginal densities**: given $f_X(x) > 0$ we have:

1. conditional pmf or pdf:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \begin{cases} \frac{f_{X,Y}(x,y)}{\sum_y f_{X,Y}(x,y)}, & \text{discrete case,} \\ \\ \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy}, & \text{continuous case;} \end{cases}$$

2. joint pmf or pdf:

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x);$$

3. marginal pmf or pdf:

$$f_Y(y) = \begin{cases} \sum_x f_{Y|X}(y|x) f_X(x), & \text{discrete case,} \\ \\ \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) dx, & \text{continuous case;} \end{cases}$$

4. reverse conditioning (if also $f_Y(y) > 0$):

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

These are all direct implications of Bayes' theorem.

# 12 Conditional moments and mgfs

**Conditional expectation**: given $X$ and $Y$ random variables the expectation of $Y$ given $X = x$ is

$$\mathrm{E}[Y|X=x] = \begin{cases} \sum_y y f_{Y|X}(y|x), & \text{discrete case,} \\ \\ \int_{-\infty}^{+\infty} y f_{Y|X}(y|x) \mathrm{d}y, & \text{continuous case.} \end{cases}$$

If we consider all possible values taken by $X$ then we have a new random variable which is the conditional expectation of $Y$ given $X$ and it is written as $\mathrm{E}[Y|X]$. It is the best guess of $Y$ given the knowledge of $X$. All properties of expectations still hold.

**Law of iterated expectations**: since $\mathrm{E}[Y|X]$ is a random variable we can take its expectation:

$$\mathrm{E}[\mathrm{E}[Y|X]] = \mathrm{E}[Y].$$

Indeed, in the continuous case

$$\begin{aligned} \mathrm{E}[\mathrm{E}[Y|X]] &= \int_{-\infty}^{+\infty} \mathrm{E}[Y|X=x] f_X(x) \mathrm{d}x = \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)} f_X(x) \mathrm{d}x \mathrm{d}y = \mathrm{E}[Y]. \end{aligned}$$

A useful consequence is that we can compute $\mathrm{E}[Y]$ without having to refer to the marginal pmf or pdf of $Y$:

$$\mathrm{E}[Y] = \begin{cases} \sum_x \mathrm{E}[Y|X=x] f_X(x), & \text{discrete case,} \\ \\ \int_{-\infty}^{+\infty} \mathrm{E}[Y|X=x] f_X(x) \mathrm{d}x, & \text{continuous case.} \end{cases}$$

**Conditional expectations of function of random variables**: if $g$ is a well-behaved, real-valued function, the expectation of $g(Y)$ given $X = x$ is defined as:

$$\mathrm{E}[g(Y)|X=x] = \begin{cases} \sum_y g(y) f_{Y|X}(y|x), & \text{discrete case,} \\ \\ \int_{-\infty}^{+\infty} g(y) f_{Y|X}(y|x) \mathrm{d}y, & \text{continuous case.} \end{cases}$$

The conditional expectation of $g(Y)$ given $X$ is written as $\mathrm{E}[g(Y)|X]$ and it is also a random variable.

As a consequence any function of $X$ can be treated as constant with respect to expectations conditional on $X$. In general for well-behaved functions $g_1$ and $g_2$

$$\mathrm{E}[g_1(X) g_2(Y)|X] = g_1(X) \mathrm{E}[g_2(Y)|X].$$

Notice that also $\mathrm{E}[Y|X]$ is a function of $X$ so

$$\mathrm{E}[\mathrm{E}[Y|X] Y|X] = \mathrm{E}[Y|X] \mathrm{E}[Y|X] = (\mathrm{E}[Y|X])^2.$$

**Conditional variance**: for random variables $X$ and $Y$, it is defined as

$$\begin{aligned} \mathrm{Var}[Y|X=x] &= \mathrm{E}[(Y - \mathrm{E}[Y|X=x])^2 | X=x] = \\ &= \begin{cases} \sum_y [y - \mathrm{E}[Y|X=x]]^2 f_{Y|X}(y|x), & \text{discrete case,} \\ \\ \int_{-\infty}^{\infty} [y - \mathrm{E}[Y|X=x]]^2 f_{Y|X}(y|x) dy, & \text{continuous case.} \end{cases} \end{aligned}$$

The conditional variance of $Y$ given $X$ is written as $\mathrm{Var}[Y|X]$ and it is a random variable function of $X$. Moreover,

$$\mathrm{Var}[Y|X] = \mathrm{E}[Y^2|X] - (\mathrm{E}[Y|X])^2,$$

By using the law of iterated expectations,

$$\begin{aligned} \mathrm{Var}[Y] &= \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2 = \\ &= \mathrm{E}[\mathrm{E}[Y^2|X]] - \{\mathrm{E}[\mathrm{E}[Y|X]]\}^2 = \\ &= \mathrm{E}[\mathrm{Var}[Y|X] + (\mathrm{E}[Y|X])^2] - \{\mathrm{E}[\mathrm{E}[Y|X]]\}^2 \\ &= \mathrm{E}[\mathrm{Var}[Y|X]] + \mathrm{E}[\{\mathrm{E}[Y|X]\}^2] - \{\mathrm{E}[\mathrm{E}[Y|X]]\}^2 \\ &= \mathrm{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[\mathrm{E}[Y|X]], \end{aligned}$$

This result tells us that

$$\text{Var}[Y] \geq \text{E}[\text{Var}[Y|X]],$$

the expected value of the conditional variance is in general smaller than the unconditional variance. If $X$ contains useful information for $Y$ then conditioning on $X$ makes uncertainty about the value of $Y$ smaller. The case in which equality holds is when $\text{Var}[\text{E}[Y|X]] = 0$, i.e. when $\text{E}[Y|X]$ is no more random, which is when $X$ contains no information on $Y$, i.e. they are independent.

**Conditional distributions and independence**: if $X$ and $Y$ are independent random variables then for cdfs we have

$$\begin{aligned} F_{Y|X}(y|x) &= F_Y(y) \quad \forall\, x, y \in \mathbb{R}, \\ F_{X|Y}(x|y) &= F_X(x) \quad \forall\, x, y \in \mathbb{R}. \end{aligned}$$

and for pmfs or pdfs we have

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y) \quad \forall\, x, y \in \mathbb{R}, \\ f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x) \quad \forall\, x, y \in \mathbb{R}. \end{aligned}$$

Finally,

$$\text{E}[Y|X] = \text{E}[Y].$$

**Conditional moment generating function**: given $X = x$, it is the function defined as

$$M_{Y|X}(u|x) = \text{E}[e^{uY}|X = x] = \begin{cases} \sum_y e^{uy} f_{Y|X}(y|x), & \text{discrete case,} \\ \int_{-\infty}^{+\infty} e^{uy} f_{Y|X}(y|x)\mathrm{d}y, & \text{continuous case.} \end{cases}$$

This is a conditional expectation so it is a random variable. We can calculate the joint mgf and marginal mgfs from the conditional mgf,

$$\begin{aligned} M_{X,Y}(t,u) &= \text{E}[e^{tX+uY}] = \text{E}[e^{tX} M_{Y|X}(u|X)], \\ M_Y(u) &= M_{X,Y}(0,u) = \text{E}[M_{Y|X}(u|X)]. \end{aligned}$$

**Example**: suppose that $X$ is the number of hurricanes that form in the Atlantic basin in a given year and $Y$ is the number making landfall. We assume we know that each hurricane has a probability $p$ of making landfall independent of other hurricanes. If we know the number of hurricanes that form say $x$ we can view $Y$ as the number of success in $x$ independent Bernoulli trials, i.e. $Y|X = x \sim \text{Bin}(x, p)$. If we also know that

$X \sim \text{Pois}(\lambda)$, then we can compute the distribution of $Y$ (notice that $X \geq Y$)

$$\begin{aligned} f_Y(y) &= \sum_{x=y}^{+\infty} f_{Y|X}(y|x) f_X(x) = \\ &= \sum_{x=y}^{+\infty} \frac{x!}{y!(x-y)!} p^y (1-p)^{x-y} \frac{\lambda^x e^{-\lambda}}{x!} = \\ &= \frac{\lambda^y p^y e^{-\lambda}}{y!} \sum_{x=y}^{+\infty} \frac{[\lambda(1-p)]^{x-y}}{(x-y)!} = \\ &= \frac{\lambda^y p^y e^{-\lambda}}{y!} \sum_{j=0}^{+\infty} \frac{[\lambda(1-p)]^{j}}{j!} = \\ &= \frac{\lambda^y p^y e^{-\lambda}}{y!} e^{\lambda(1-p)} = \\ &= \frac{(\lambda p)^y e^{-\lambda p}}{y!}, \end{aligned}$$

thus $Y \sim \text{Pois}\lambda p$. So $\text{E}[Y] = \lambda p$ and $\text{Var}[Y] = \lambda p$, but we could find these results without the need of the marginal pdf. Since $Y|X = x \sim \text{Bin}(x, p)$, then

$$\text{E}[Y|X = x] = Xp \qquad \text{Var}[Y|X = x] = Xp(1-p)$$

Since $X \sim \text{Pois}(\lambda)$, by using the law of iterated expectations, we have

$$\text{E}[Y] = \text{E}[\text{E}[Y|X = x]] = \text{E}[X]p = \lambda p$$

and

$$\text{Var}[Y] = \text{E}[\text{Var}[Y|X = x]] + \text{Var}[\text{E}[Y|X = x]] = \text{E}[X]p(1-p) + \text{Var}[Xp] = \lambda p(1-p) + \lambda p^2 = \lambda p.$$

Alternatively we can use the mgfs, we have

$$M_X(t) = \exp\{\lambda(e^t - 1)\} \qquad M_{Y|X}(u|X) = (1 - p + pe^u)^X,$$

therefore

$$\begin{aligned} M_Y(u) &= \text{E}[M_{Y|X}(u|X)] = \text{E}[(1 - p + pe^u)^X] = \\ &= \text{E}[\exp\{X \log(1 - p + pe^u)\}] = \\ &= M_X(\log(1 - p + pe^u)) = \\ &= \exp\{\lambda(1 - p + pe^u - 1)\} = \\ &= \exp\{\lambda p(e^u - 1)\}, \end{aligned}$$

which is the mgf of a Poisson distribution.

# 13 An example of bivariate distribution

Consider the function

$$f_{X,Y}(x,y) = \begin{cases} x+y & \text{if } 0<x<1 \text{ and } 0<y<1, \\ 0 & \text{otherwise.} \end{cases}$$

- It is a valid density, indeed it is a positive real valued function and it is normalized

$$\begin{aligned} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y &= \int_0^1\int_0^1 (x+y)\mathrm{d}x\mathrm{d}y = \\ &= \int_0^1 \left[\frac{x^2}{2}+xy\right]_0^1 \mathrm{d}y = \int_0^1 \left[\frac{1}{2}+y\right]\mathrm{d}y = \\ &= \left[\frac{y}{2}+\frac{y^2}{2}\right]_0^1 = 1. \end{aligned}$$

- The joint cdf is

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\mathrm{d}u\mathrm{d}v = \\ &= \int_{-\infty}^{y}\int_{-\infty}^{x} (u+v)\mathrm{d}u\mathrm{d}v = \\ &= \int_0^y \left[\frac{x^2}{2}+xv\right]\mathrm{d}v = \left[\frac{x^2v}{2}+\frac{xv^2}{2}\right]_0^1 = \\ &= \frac{1}{2}xy(x+y) \quad \text{for } 0<x<1, 0<y<1. \end{aligned}$$

  More precisely we have

$$F_{X,Y}(x,y) = \begin{cases} \frac{1}{2}xy(x+y) & \text{if } 0<x<1 \text{ and } 0<y<1, \\ \frac{1}{2}x(x+1) & \text{if } 0<x<1 \text{ and } y\geq 1, \\ \frac{1}{2}y(y+1) & \text{if } x\geq 1 \text{ and } 0<y<1, \\ 1 & \text{if } x\geq 1 \text{ and } y\geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

- The marginal pdf of $X$ is

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\mathrm{d}y = \\ &= \int_0^1 (x+y)\mathrm{d}y = x+\frac{1}{2}. \end{aligned}$$

- We can compute probabilities as $P(2X < Y)$, we first define the event $B =$

$\{(x,y)$ s.t. $0<x<\frac{y}{2}, 0<y<1\}$ then

$$\begin{aligned} P(2X<Y)=P(B) &= \int\int_B f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \\ &= \int_0^1\int_0^{y/2}(x+y)\mathrm{d}x\mathrm{d}y = \int_0^1\left[\frac{y^2}{8}+\frac{y^2}{2}\right]\mathrm{d}y = \\ &= \left[\frac{y^3}{24}+\frac{y^3}{6}\right]_0^1 = \frac{5}{24}. \end{aligned}$$

Analogously we could define $C=\{(x,y)$ s.t. $0<x<\frac{1}{2}, 2x<y<1\}$ and compute $P(C)$.

- the $(r,s)^{\text{th}}$ joint moment is

$$\begin{aligned} \mathrm{E}[X^rY^s] &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x^ry^sf_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_0^1\int_0^1 x^ry^s(x+y)\mathrm{d}x\mathrm{d}y = \int_0^1\left[\frac{1}{r+2}y^s+\frac{1}{r+1}y^{s+1}\right]\mathrm{d}y = \\ &= \left[\frac{1}{(r+2)(s+1)}y^{s+1}+\frac{1}{(r+1)(s+2)}y^{s+2}\right]_0^1 = \frac{1}{(r+2)(s+1)}+\frac{1}{(r+1)(s+2)}. \end{aligned}$$

Thus, $\mathrm{E}[XY]=\frac{1}{3}$, $\mathrm{E}[X]=\mathrm{E}[Y]=\frac{7}{12}$, $\mathrm{E}[X^2]=\frac{5}{12}$ so $\mathrm{Var}[X]=\frac{11}{144}$ and finally

$$\mathrm{Cov}(X,Y)=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]=\frac{1}{3}-\frac{49}{144}=-\frac{1}{144},$$

and $\mathrm{Corr}(X,Y)=-\frac{1}{11}$, so $X$ and $Y$ are not independent.
We find this result also by noticing that given the marginals and the joint pdfs we have

$$f_X(x)f_Y(y)=xy+\frac{x+y}{2}+\frac{1}{4},$$

therefore $f_X(x)f_Y(y)\neq f_{X,Y}(x,y)$ so $X$ and $Y$ are not independent.

- The conditional pdf of $Y$ given $X=x$ is

$$f_{Y|X}(y|x)=\frac{f_{X,Y}(x,y)}{f_X(x)}=\begin{cases}\frac{x+y}{x+\frac{1}{2}} & \text{if } 0<y<1\\ 0 & \text{otherwise.}\end{cases}$$

- The conditional expectation of $Y$ given $X=x$ is

$$\begin{aligned} \mathrm{E}[Y|X=x] &= \int_{-\infty}^{+\infty}yf_{Y|X}(y|x)\mathrm{d}y = \\ &= \int_0^1 y\frac{x+y}{x+\frac{1}{2}}\mathrm{d}y = \\ &= \frac{1}{x+\frac{1}{2}}\left[\frac{xy^2}{2}+\frac{y^3}{3}\right]_0^1 = \\ &= \frac{3x+2}{6x+3}. \end{aligned}$$

- we can use the law of iterated expectations

$$\begin{aligned}\mathrm{E}[\mathrm{E}[Y|X=x]] &= \int_0^1 \frac{3x+2}{6x+3}\left(x+\frac{1}{2}\right)\mathrm{d}x = \\ &= \frac{1}{6}\int_0^1 3x+2\mathrm{d}x \\ &= \frac{1}{6}\left(\frac{3}{2}+2\right) \\ &= \frac{7}{12} = \mathrm{E}[Y].\end{aligned}$$

# Reading

Casella and Berger, Sections 4.2 - 4.4 - 4.5.

# 14 Sums of random variables

We start with the bivariate case and then we generalise it to $n$ variables.

**Moments of a sum**: if $X$ and $Y$ are random variables then:

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y], \qquad \mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}(X,Y),$$

and, by using the linearity of expectations and the binomial expansion, we have for $r \in \mathbb{N}$

$$\mathrm{E}[(X+Y)^r] = \sum_{j=0}^{r}\binom{r}{j}\mathrm{E}[X^j Y^{r-j}].$$

**Probability mass/density function of a sum**: if $X$ and $Y$ are random variables with joint density $f_{X,Y}(x,y)$ and we define $Z = X+Y$ then the pmf/pdf of $Z$ is

$$f_Z(z) = \begin{cases} \sum_u f_{X,Y}(u, z-u), & \text{discrete case,} \\ \int_{-\infty}^{+\infty} f_{X,Y}(u, z-u)\mathrm{d}u, & \text{continuous case.}\end{cases}$$

In the continuous case just change variables $X = U$ and $Y = Z - U$. In the discrete case notice that

$$\{X+Y=z\} = \bigcup_u \{X = u \cap Y = z-u\}$$

and, since this is a sum of disjoint events, for any $u$, we have

$$P(X+Y=z) = \sum_u P(X = u \cap Y = z-u).$$

**Probability mass/density function of a sum of independent random variables**: if $X$ and $Y$ are independent random variables and we define $Z = X + Y$ then the pmf/pdf of $Z$ is

$$f_Z(z) = \begin{cases} \sum_u f_X(u) f_Y(z-u), & \text{discrete case,} \\ \int_{-\infty}^{+\infty} f_X(u) f_Y(z-u) \mathrm{d}u, & \text{continuous case.} \end{cases}$$

This operation is known as convolution

$$f_Z = f_X * f_Y \Leftrightarrow \int_{-\infty}^{+\infty} f_X(u) f_Y(z-u) \mathrm{d}u.$$

Convolution is commutative so $f_X * f_Y = f_Y * f_X$.

**Moment generating function of the sum of independent random variables**: if $X$ and $Y$ are independent random variables and we define $Z = X + Y$ then the mgf of $Z$ is

$$M_Z(t) = M_X(t) M_Y(t),$$

and the cumulant generating function is

$$K_Z(t) = K_X(t) + K_Y(t).$$

**Example**: suppose the $X$ and $Y$ are independent r.v. exponentially distributed, $X \sim \mathrm{Exp}(\lambda)$ and $Y \sim \mathrm{Exp}(\theta)$, with $\lambda \neq \theta$, then the pdf of $Z = X + Y$ is

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} f_X(u) f_Y(z-u) \mathrm{d}u = \\ &= \int_0^z \lambda e^{-\lambda u} \theta e^{-\theta(z-u)} \mathrm{d}u = \\ &= \lambda\theta e^{-\theta z} \left[ \frac{-1}{\lambda - \theta} e^{-(\lambda-\theta)u} \right]_0^z = \\ &= \frac{\lambda\theta}{\lambda - \theta} (e^{-\theta z} - e^{-\lambda z}) \qquad 0 \leq z < +\infty. \end{aligned}$$

Note the domain of integration $[0, z]$. Indeed, since both $X$ and $Y$ are positive r.v., also $U$ and $Z - U$ have to be positive, thus we need $0 < U \leq Z$.
In theory, we could also use mgfs, but in this case we get a function of $t$ that does not have an expression that resembles one of a known distribution.

**Example**: suppose the $X$ and $Y$ are independent r.v. normally distributed, $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then then to compute the pdf of $Z = X + Y$ we use the cumulant generating functions

$$K_X(t) = \mu_X t + \frac{\sigma_X^2 t^2}{2}, \quad K_Y(t) = \mu_Y t + \frac{\sigma_Y^2 t^2}{2},$$

and

$$K_Z(t) = (\mu_X + \mu_Y) t + \frac{(\sigma_X^2 + \sigma_Y^2) t^2}{2}$$

by uniqueness of cumulant generating functions $Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

**Multivariate generalization**: for $n$ independent random variables $X_1, \ldots, X_n$ let $S = \sum_{j=1}^n X_j$ then

1. the pmf/pdf of $S$ is
$$f_S = f_{X_1} * \ldots * f_{X_n};$$
2. the mgf of $S$is
$$M_S(t) = M_{X_1}(t) \ldots M_{X_n}(t).$$
3. if $X_1, \ldots, X_n$ are also identically distributed they have a common mgf $M_X(t)$ thus
$$f_S = \underbrace{f * f * \ldots * f}_{n-\text{times}}, \quad M_S(t) = [M_X(t)]^n, \quad K_S(t) = nK_X(t).$$

To indicate independent and identically distributed random variables we use the notation i.i.d.

**Example**: given $n$ i.i.d. Bernoulli r.v. $X_1 \ldots X_n$ with probability $p$ and mgf
$$M_X(t) = 1 - p + pe^t,$$
the sum $S = \sum_{j=1}^n X_j$ has mgf
$$M_S(t) = (1 - p + pe^t)^n,$$
thus, by uniqueness of mgf, $S \sim \text{Bin}(n, p)$.

**Example**: given $X_1, \ldots, X_n$ independent r.v. normally distributed $X_j \sim N(\mu_j, \sigma_j^2)$ then, for fixed constants $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, we have
$$S = \sum_{j=1}^n (a_j X_j + b_j) \sim N\left(\sum_{j=1}^n (a_j \mu_j + b_j), \sum_{j=1}^n a_j^2 \sigma_j^2\right).$$
If $X_j \sim iidN(\mu, \sigma^2)$, then
$$S = \sum_{j=1}^n X_j \sim N(n\mu, n\sigma^2).$$

**Other examples of sums of independent random variables**

1. Poisson:
$$X \sim \text{Pois}(\lambda_1),\ Y \sim \text{Pois}(\lambda_2) \Rightarrow Z \sim \text{Pois}(\lambda_1 + \lambda_2)$$
$$X_j \sim iid\text{Pois}(\lambda) \Rightarrow S \sim \text{Pois}(n\lambda) \quad j = 1, \ldots n;$$

2. Gamma:

$$X \sim \text{Gamma}(r_1, \theta),\ Y \sim \text{Gamma}(r_2, \theta) \Rightarrow Z \sim \text{Gamma}(r_1 + r_2, \theta)$$

$$X_j \sim iid\text{Exp}(\lambda) \Rightarrow S \sim \text{Gamma}(n, \lambda) \quad j = 1, \dots n;$$

3. Binomial:

$$X \sim \text{Bin}(n_1, p),\ Y \sim \text{Bin}(n_2, p) \Rightarrow Z \sim \text{Bin}(n_1 + n_2, p)$$

$$X_j \sim iid\text{Bin}(k, p) \Rightarrow S \sim \text{Bin}(nk, p) \quad j = 1, \dots n.$$

## 14.1 Limit theorems for Bernoulli sums

Assume to observe $n$ independent Bernoulli trials $X_i$ with an unknown probability of success $p$. We study the behaviour of the process $S_n = \sum_{i=1}^n X_i$ which counts the number of successes in $n$ trials. If $X_i \sim iid\text{Bernoulli}(p)$, then $S_n \sim \text{Bin}(n, p)$. For any $i$ we have that $\text{E}[X_i] = p$ and $\text{Var}[X_i] = p(1-p)$ so that $\text{E}[S_n] = np$ and $\text{Var}[S_n] = np(1-p)$.

**Law of Large Numbers**: there are two forms of this law:

1. Weak Law of Large Numbers: as $n \to +\infty$, $S_n/n \overset{m.s.}{\to} p$, i.e.

$$\lim_{n \to +\infty} \text{E}\left[\left(\frac{S_n}{n} - p\right)^2\right] = 0,$$

   which implies $S_n/n \overset{P}{\to} p$, i.e.

$$\lim_{n \to \infty} P\left(\left|\frac{S_n}{n} - p\right| < \epsilon\right) = 1, \quad \forall \epsilon > 0;$$

2. Strong Law of Large Numbers: as $n \to +\infty$, $S_n/n \overset{a.s.}{\to} p$, i.e.

$$P\left(\lim_{n \to \infty} \left|\frac{S_n}{n} - p\right| = 0\right) = 1.$$

The law establishes the convergence of the empirical average (or sample mean) $S_n/n$ to the expected value of $X_i$, i.e. to $p$ (or population mean). It is uesful if we observe many Bernoulli trials and we want to determine $p$: it is a first example of inference.

**Proof of the weak law**: for each $n$ we have

$$\text{E}\left[\left(\frac{S_n}{n} - p\right)^2\right] = \frac{\text{E}\left[(S_n - np)^2\right]}{n^2} = \frac{\text{Var}[S_n]}{n^2} = \frac{p(1-p)}{n} \to 0, \quad \text{as } n \to +\infty.$$

**Example**: when tossing a coin $X_i$ is 1 if we get head or 0 if we get tail (a Bernoulli trial), $S_n$ is the number of heads we get in $n$ independent tosses. The frequency of heads will converge to $1/2$ which is the value of $p$ in this particular case.

The following result is a special case of the Central Limit Theorem which we shall see in due course.

**De Moivre-Laplace Limit Theorem**: as $n \to +\infty$, and for $Z \sim N(0,1)$,

$$\lim_{n\to\infty} P\left(\sqrt{n}\frac{S_n/n - p}{\sqrt{p(1-p)}} \leq \alpha\right) = P(Z \leq \alpha) = \int_{-\infty}^{\alpha} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\mathrm{d}z, \quad \forall\, \alpha \in \mathbb{R},$$

which implies

$$\sqrt{n}\frac{S_n/n - p}{\sqrt{p(1-p)}} \xrightarrow{d} Z.$$

We are saying that the sample mean (which is a random variable) of the Bernoulli trials converges in distribution or is asymptotically distributed as a normal random variable with mean $p$ (this we know already from the law of large numbers) and variance $p(1-p)/n$, thus the more trials we observe the smaller the uncertainty about the expected value of the sample mean, the rate of convergence being $\sqrt{n}$. This result contains useful informations not only on the point-wise estimate of the population mean but also on the uncertainty and the speed with which we have convergence.

Finally, remember that $S_n \sim \mathrm{Bin}(np, np(1-p))$, then, by rearranging the terms, we have

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z.$$

i.e. the Binomial distribution can be approximated by a normal distribution with mean $np$ and variance $np(1-p)$.

# Reading

Casella and Berger, Sections 5.2

# 15 Mixtures and random sums

**Hierarchies and mixtures**: suppose we are interested in a random variable $Y$ which has a distribution that depends on another random variables, say $X$. This is called a hierarchical model and $Y$ has a mixture distribution. In the first instance we do not know the marginal distribution of $Y$ directly, but we know the conditional distribution of $Y$ given

$X = x$ and the marginal distribution of $X$ (see the example on hurricanes of Section 26).

The key results which are necessary for characterising $Y$, are

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[\mathrm{E}[Y|X]] \\ \mathrm{Var} &= \mathrm{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[\mathrm{E}[Y|X]] \\ f_Y(y) &= \mathrm{E}[f_{Y|X}(y|X)] \quad \text{and } M_Y(t) = \mathrm{E}[M_{Y|X}(t|X)] \end{aligned}$$

**Example:** Poisson mixing. If $Y|\Lambda = \lambda \sim \mathrm{Pois}(\lambda)$, for some positive r.v. $\Lambda$, then

$$\mathrm{E}[Y|\Lambda] = \mathrm{Var}[Y|\Lambda] = \Lambda.$$

Therefore,

$$\mathrm{E}[Y] = \mathrm{E}[\Lambda], \qquad \mathrm{Var}[Y] = \mathrm{E}[\Lambda] + \mathrm{Var}[\Lambda].$$

**Random sums**: We consider the case in which $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables and $Y = \sum_{j=1}^{N} X_j$, where $N$ is also a random variable which is independent of each $X_i$. $Y$ is called random sum and can be viewed as a mixture such that $Y|N = n$ is a sum of random variables, so all results of previous section still hold.

**Conditional results for random sums**: suppose that $\{X_j\}$ is a sequence of i.i.d. random variables with mean $\mathrm{E}[X]$ and variance $\mathrm{Var}[X]$, for any $j$, and suppose that $N$ is a random variable taking only positive integer values and define $Y = \sum_{j=1}^{N} X_j$, then

$$\begin{aligned} \mathrm{E}[Y|N] &= N\mathrm{E}[X], \\ \mathrm{Var}[Y|N] &= N\mathrm{Var}[X], \\ M_{Y|N}(t|N) &= [M_X(t)]^N \quad \text{and } K_{Y|N}(t|N) = NK_X(t). \end{aligned}$$

**Marginal results for random sums**: suppose that $\{X_j\}$ is a sequence of i.i.d. random variables with mean $\mathrm{E}[X]$ and variance $\mathrm{Var}[X]$, for any $j$, and suppose that $N$ is a random variable taking only positive integer values and define $Y = \sum_{j=1}^{N} X_j$, then

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[N]\mathrm{E}[X], \\ \mathrm{Var}[Y] &= \mathrm{E}[N]\mathrm{Var}[X] + \mathrm{Var}[N]\{\mathrm{E}[X]\}^2, \\ M_Y(t) &= M_N(\log M_X(t)) \quad \text{and } K_Y(t) = K_N(K_X(t)). \end{aligned}$$

**Example**: each year the value of claims made by an owner of a health insurance policy is distributed exponentially with mean $\alpha$ independent of previous years. At the end of each year with probability $p$ the individual will cancel her policy. We want the distribution of the total cost of the health insurance policy for the insurer. The value of claims in year $j$ is $X_j$ and the number of years in which the policy is held is $N$, thus

$$X_j \sim iid\mathrm{Exp}\left(\frac{1}{\alpha}\right), \qquad N \sim \text{Geometric}(p).$$

The total cost for the insurer is $Y = \sum_{j=1}^{N} X_j$. Therefore, $\mathrm{E}[Y] = \alpha\frac{1}{p}$. To get the distribution we use the cumulant generating function

$$K_X(t) = -\log(1-\alpha t), \qquad K_N(t) = -\log\left(1 - \frac{1}{p} + \frac{1}{p}e^{-t}\right),$$

and

$$K_Y(t) = K_N(K_X(t)) = -\log\left(1 - \frac{1}{p} + \frac{1}{p}(1-\alpha t)\right) = -\log\left(1 - \frac{\alpha}{p}t\right),$$

by uniqueness we have that $Y \sim \mathrm{Exp}\left(\frac{p}{\alpha}\right)$.

**The Poisson approximation**: assume to have $X_j \sim iid\mathrm{Bernoulli}(p)$, and $N \sim \mathrm{Pois}(\lambda)$. Consider $Y = \sum_{j=1}^{N} X_j$, then $Y|N = n \sim \mathrm{Bin}(n,p)$ and

$$\begin{aligned}
\mathrm{E}[Y] &= \lambda\mathrm{E}[X], \\
\mathrm{Var}[Y] &= \lambda\mathrm{E}[X^2], \\
M_Y(t) &= M_N(\log M_X(t)) = e^{\lambda(M_X(t)-1)}, \\
K_S(t) &= \lambda(M_X(t)-1).
\end{aligned}$$

By using the mgf of a Bernoulli $M_X(t) = 1 - p + pe^t$ we get

$$M_Y(t) = e^{\lambda(M_X(t)-1)} = e^{\lambda p(e^t-1)},$$

by uniqueness of mgf, $Y \sim \mathrm{Pois}(\lambda p)$ (see the example on hurricanes of Section 26).

# Reading

Casella and Berger, Section 4.4

# 16 Random vectors

This is just a way to simplify notation when we consider $n$ random variables. Expectations are element wise and we have to remember that the variance of a vector is a matrix.

**Random vector**: an $n$-dimensional vector of random variables, i.e. a function

$$\mathbf{X} = (X_1, \ldots, X_n)^T : \Omega \to \mathbb{R}^n.$$

The cdf, pmf or pdf, and mgf of a random vector are the joint cdf, pmf or pdf, and mgf of $X_1, \ldots, X_n$ so, for any $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{t} = (t_1, \ldots, t_n) \in \mathbb{R}^n$,

$$\begin{aligned}
F_{\mathbf{X}}(\mathbf{x}) &= F_{X_1,\ldots,X_n}(x_1, \ldots, x_n), \\
f_{\mathbf{X}}(\mathbf{x}) &= f_{X_1,\ldots,X_n}(x_1, \ldots, x_n), \\
M_{\mathbf{X}}(\mathbf{t}) &= M_{X_1,\ldots,X_n}(t_1, \ldots, t_n).
\end{aligned}$$

**Expectation of a random vector**: the expectation of a random vector is a vector of the expectations, i.e. it is taken element by element

$$\mathrm{E}[\mathbf{X}] = \begin{pmatrix} \mathrm{E}[X_1] \\ \vdots \\ \mathrm{E}[X_n] \end{pmatrix}.$$

For jointly continuous random variables we have

$$\begin{aligned} \mathrm{E}[\mathbf{X}] &= \int_{\mathbb{R}^n} \mathbf{x} f_{\mathbf{X}}(\mathbf{x}) \mathrm{d}\mathbf{x} = \\ &= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} x_1 \dots x_n f_{X_1,\dots,X_n}(x_1,\dots,x_n) \mathrm{d}x_1 \dots \mathrm{d}x_n. \end{aligned}$$

**Variance-covariance matrix**: given $n$ random variables $X_1, \dots, X_n$ we know what is the variance of each of them and we know the covariance of each couple. All these informations can be summarized in just one object, defined as

$$\mathbf{\Sigma} = \mathrm{Var}[\mathbf{X}] = \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^T],$$

Where $\mathbf{X}$ is $n \times 1$ (a column vector), then $\mathbf{X}^T$ is $1 \times n$ (a row vector), and $\mathbf{\Sigma}$ is a $n \times n$ matrix. Taking element by element expectation of this matrix we get

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}[X_2] & \dots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \dots & \dots & \mathrm{Var}[X_n] \end{pmatrix}.$$

The matrix is symmetric and if the variables are uncorrelated then it is a diagonal matrix. If the variables are also identically distributed then $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_n$ where $\sigma^2$ is the variance of each random variable and $\mathbf{I}_n$ is the $n$-dimensional identity matrix. Finally, as the univaraite variance is always positive, in this case we have that $\mathbf{\Sigma}$ is a non-negative definite matrix, i.e.

$$\mathbf{b}^T \mathbf{\Sigma} \mathbf{b} \geq 0 \qquad \forall\, \mathbf{b} \in \mathbb{R}^n.$$

**Example**: if $N = 2$ and assume $\mathrm{E}[X] = \mathrm{E}[Y] = 0$ then

$$\Sigma = \mathrm{E}\left[ \begin{pmatrix} X \\ Y \end{pmatrix} (X\ Y) \right] = \mathrm{E}\begin{bmatrix} X^2 & XY \\ YX & Y^2 \end{bmatrix} = \begin{pmatrix} \mathrm{E}[X^2] & \mathrm{E}[XY] \\ \mathrm{E}[YX] & \mathrm{E}[Y^2] \end{pmatrix} = \begin{pmatrix} \mathrm{Var}[X] & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}[Y] \end{pmatrix}.$$

**Conditioning for random vectors**: if $\mathbf{X}$ and $\mathbf{Y}$ are random vectors, and if $f_{\mathbf{X}}(\mathbf{x}) > 0$, we can define the conditional pdf/pmf as

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})}.$$

or

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}).$$

**Decomposition of probability mass/density function**: given an $n$-dimensional random vector $\mathbf{X}$ and given $\mathbf{x} \in \mathbb{R}^n$, then

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{X_n|X_{n-1}\ldots X_1}(x_n|x_{n-1}\ldots x_1) f_{X_{n-1}|X_{n-2}\ldots X_1}(x_{n-1}|x_{n-2}\ldots x_1)\ldots f_{X_2|X_1}(x_2|x_1) f_{X_1}(x_1) = \\ &= \prod_{j=1}^{n} f_{X_j|\mathbf{X}_{j-1}}(x_j|\mathbf{x}_{j-1}), \end{aligned}$$

where the random vector $\mathbf{X}_{j-1}$ is the random vector $\mathbf{X}$ without its $j$-th element.

**Example**: consider 3 r.v. $X_1$, $X_2$ and $X_3$, we can group them in different ways and we get for example

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = f_{X_3|X_1,X_2}(x_3|x_1,x_2) f_{X_1,X_2}(x_1,x_2),$$

and applying again the definition above to the joint pdf/pmf of $X_1$ and $X_2$ we have

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = f_{X_3|X_1,X_2}(x_3|x_1,x_2) f_{X_2|X_1}(x_2|x_1) f_{X_1}(x_1).$$

# 17 Multivariate normal distribution

We start with the bivariate case. We want a bivariate version of the normal distribution. Given two standard normal random variables, we can build a bivariate normal that depends only on their correlation.

**Standard bivariate normal**: given $U$ and $V$ i.i.d. standard normal random variables, and for some number $|\rho| < 1$, define $X = U$ and $Y = \rho U + \sqrt{1-\rho^2}V$, then we can prove that

1. $X \sim N(0,1)$ and $Y \sim N(0,1)$;
2. $\mathrm{Corr}(X,Y) = \rho$;
3. the joint pdf is that of a standard bivariate normal random variable and depends only on the parameter $\rho$:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-(x^2 - 2\rho xy + y^2)/(2(1-\rho^2))\right].$$

The random vector $\mathbf{X} = (X,Y)^T$ is normally distributed and we write

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

or $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}_{X,Y})$ where $\mathbf{\Sigma}_{X,Y}$ is the $2 \times 2$ variance covariance matrix;

4. the joint mgf is

$$M_{X,Y}(s,t) = \exp\left[\frac{1}{2}(s^2 + 2\rho st + t^2)\right].$$

**Bivariate normal for independent random variables**: if the random variables $U$ and $V$ are independent and standard normal, the joint pdf and mgf are

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{2\pi} e^{-(u^2+v^2)/2}, \\ M_{U,V}(s,t) &= e^{(s^2+t^2)/2}. \end{aligned}$$

The random vector $(U, V)$ is normally distributed with variance covariance matrix

$$\mathbf{\Sigma}_{U,V} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

**Computing the joint pdf**: given $X = U$ and $Y = \rho U + \sqrt{1-\rho^2}V$, we have to compute $f_{X,Y}(x,y)$ given $f_{U,V}(u,v)$. Given the function $\mathbf{h} : \mathbb{R}^2 \to \mathbb{R}^2$ such that $\mathbf{h}(X,Y) = (U,V)$ and the domain of $\mathbf{h}$ is $C \subseteq \mathbb{R}^2$ and it is in one-to-one correspondence with the support of $(U, V)$, we have the rule

$$f_{X,Y}(x,y) = \begin{cases} f_{U,V}(\mathbf{h}(x,y))|J_{\mathbf{h}}(x,y)| & \text{for } (x,y) \in C \\ 0 & \text{otherwise} \end{cases}$$

where

$$J_{\mathbf{h}}(x,y) = \det \begin{pmatrix} \frac{\partial}{\partial x}h_1(x,y) & \frac{\partial}{\partial x}h_2(x,y) \\ \\ \frac{\partial}{\partial y}h_1(x,y) & \frac{\partial}{\partial y}h_2(x,y) \end{pmatrix}.$$

In this case, $C = \mathbb{R}^2$,

$$u = h_1(x,y) = x, \qquad v = h_2(x,y) = \frac{y - \rho x}{\sqrt{1-\rho^2}},$$

and $|J_{\mathbf{h}}(x,y)| = \frac{1}{\sqrt{1-\rho^2}}$, thus

$$f_{X,Y}(x,y) = f_{U,V}\left(x, \frac{y-\rho x}{\sqrt{1-\rho^2}}\right)\frac{1}{\sqrt{1-\rho^2}}.$$

**Generic bivariate normal**: if $X^* = \mu_X + \sigma_X X$ and $Y^* = \mu_Y + \sigma_Y Y$ then $X^* \sim N(\mu_X, \sigma_X^2)$ and $Y^* \sim N(\mu_Y, \sigma_Y^2)$ with $\mathrm{Corr}(X^*, Y^*) = \rho$ and the joint pdf is

$$f_{X^*,Y^*}(x,y) = \frac{1}{\sigma_X \sigma_Y} f_{X,Y}\left(\frac{x-\mu_X}{\sigma_X}, \frac{y-\mu_Y}{\sigma_Y}\right).$$

A generic jointly normal random vector is distributed as

$$\begin{pmatrix} X^* \\ Y^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \sigma_Y^2 \end{pmatrix}\right).$$

**Conditional distribution**: of $Y^*$ given $X^*$ is

$$Y^*|X^* = x \ \sim \ N\left(\mu_y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X),\ \sigma_Y^2(1-\rho^2)\right).$$

It is obtained by using the joint and the marginal pdfs.

**Multivariate case**

1. Multivariate normal density: let $X_1, \ldots, X_n$ be random variables and define the $n \times 1$ random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$. If $X_1, \ldots, X_n$ are jointly normal then $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the mean $\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}]$ is an $n \times 1$ vector and the covariance matrix $\boldsymbol{\Sigma} = \mathrm{Var}[\mathbf{X}]$ is an $n \times n$ matrix whose $(i,j)^{\text{th}}$ entry is $\mathrm{Cov}(X_i, X_j)$. The joint density functions is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2}|\det \boldsymbol{\Sigma}|^{-1/2} e^{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}.$$

2. Conditional expectation for multivariate normal: suppose that $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)^T$, for some integers $n$ and $m$, and $\mathbf{X} \sim N(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $\mathbf{Y} \sim N(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$ . If, $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}'_{YX}$, then

$$\begin{aligned}\mathrm{E}[\mathbf{Y}|\mathbf{X}] =& \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}(\mathbf{X} - \boldsymbol{\mu}_X),\\ \mathrm{Var}[\mathbf{Y}|\mathbf{X}] =& \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{XY}.\end{aligned}$$

**Joint normality and independence**:

- normally distributed and independent random variables are jointly normally distributed, however, a pair of jointly normally distributed variables need not be independent;
- while it is true that the marginals of a multivariate normal are normal too, it is not true in general that given two normal random variables their joint distribution is normal;
- in general, random variables may be uncorrelated but highly dependent, but if a random vector has a multivariate normal distribution then any two or more of its components that are uncorrelated are independent, this implies that any two or more of its components that are pairwise independent are independent;
- it is not true however that two random variables that are marginally normally distributed and uncorrelated are independent: it is possible for two random variables to be distributed jointly in such a way that each one alone is marginally normally distributed, and they are uncorrelated, but they are not independent.

**Example**: consider $X$ a standard normal random variable and define

$$Y = \begin{cases} X & \text{if } |X| > c \\ -X & \text{if } |X| < c \end{cases}$$

where $c$ is a positive number to be specified. If $c$ is very small, then$\mathrm{Corr}(X, Y) \simeq 1$; if $c$ is very large, then $\mathrm{Corr}(X, Y) \simeq -1$. Since the correlation is a continuous function of $c$, there is some particular value of $c$ that makes the correlation 0. That value is approximately $1.54$. In that case, $X$ and $Y$ are uncorrelated, but they are clearly not independent, since $X$ completely determines $Y$. Moreover, $Y$ is normally distributed. Indeed, its distribution is the same as that of $X$. We use cdfs:

$$\begin{aligned} P(Y \leq x) &= P((|X| < c \cap -X < x) \cup (|X| > c \cap X < x)) = \\ &= P((|X| < c \cap X > -x)) + P((|X| > c \cap X < x)) = \\ &= P((|X| < c \cap X < x)) + P((|X| > c \cap X < x)) \end{aligned}$$

where the last row depends on the fact that for a symmetric distribution $P(X < x) = P(X > -x)$. Thus, since the events $\{|X| < c\}$ and $\{|X| > c\}$ are a partition of the sample space which is $\mathbb{R}$, then

$$P(Y \leq x) = P(X \leq x),$$

hence $Y$ is a standard normal random variable too. Finally, notice that the sum $X + Y$ for $c = 1.54$ has a substantial probability (about 0.88) of it being equal to 0, whereas the normal distribution, being a continuous distribution, has no discrete part, i.e., does not concentrate more than zero probability at any single point. Consequently $X$ and $Y$ are not jointly normally distributed, even though they are marginally normally distributed.

# Reading

Casella and Berger, Definition 4.5.10

# 18 Bernoulli motivation for the Law of Large Numbers

This section starts off somewhat more abstract but concludes with the most important and widely-used theorem in probability, the Central Limit Theorem. Along the way we also state and prove two laws of large numbers.

To get started, as an example, consider a sequence of independent Bernoulli random variables $X_i \sim X$ with $p = 1/2$ and let $Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (2X_i - 1)$. Note that we have normalised the $X_i$ so that $E[Y_n] = 0$ and $\mathrm{Var}(Y_n) = 1$. In particular, the mean and variance of $Y_n$ does not depend on $n$. A gambler could think of $Y_n$ as their (rescaled) earnings in case they win £1 each time a fair coin ends up head and lose £1 each time the coin leads to tail. Astonishingly, even though $Y_n$ is constructed from a humble Bernoulli distribution, as $n$ gets large, the distribution of $Y_n$ approaches that of the normal distribution. Indeed,

using moment generating functions (and $M_{aX+b}(t) = e^{bt}M_X(at)$ for $a, b \in \mathbb{R}$), we get

$$\begin{aligned}
M_{Y_n}(t) &= \left(e^{-t/\sqrt{n}}M_X(2t/\sqrt{n})\right)^n \\
&= \left(e^{-t/\sqrt{n}}(1 - \frac{1}{2} + \frac{1}{2}e^{2t/\sqrt{n}})\right)^n \\
&= \left(\frac{1}{2}e^{-t/\sqrt{n}} + \frac{1}{2}e^{t/\sqrt{n}}\right)^n \\
&\to \left((\frac{1}{2} - \frac{1}{2}t/\sqrt{n} + \frac{1}{4}t^2/n) + (\frac{1}{2} + \frac{1}{2}t/\sqrt{n} + \frac{1}{4}t^2/n)\right)^n \quad \text{(Taylor)} \\
&= \left((1 + \frac{1}{2}t^2/n)\right)^n \\
&\to e^{t^2/2},
\end{aligned}$$

which we recognise as the moment generating function of a standard normal distribution. Since moment generating functions (usually, more on this below) uniquely determine the distribution, it follows that $Y_n$ "converges" to a normally distributed random variable. We shall see below that there is nothing special here about the Bernoulli distribution as hardly any distribution (though there are some) can resist the attraction of the normal distribution. But before we get to that, we first have a closer look at the various kinds of convergence of random variables and how these notions are related.

# 19 Modes of convergence

In what follows we consider a sequence of random variables $X_1, X_2, \ldots$ and we consider four (and there are more!) types of convergence.

The first notion is that of almost sure convergence. Perhaps you find the terminology surprising since in mathematical statements we are used to certainty and almost sure sounds rather vague (in fact, there is even a notion of vague convergence), but almost sure in the setting here means that convergence happens on a set with probability 1.

**Almost sure convergence**: the sequence $\{X_n\}$ converges to $X$ almost surely if

$$P\left(\omega \in \Omega : \lim_{n\to+\infty} X_n(\omega) = X(\omega)\right) = 1,$$

and we use the notation $X_n \overset{a.s.}{\to} X$.
It means that $X_n(\omega)$ converges to $X(\omega)$ for all $\omega \in \Omega$ except perhaps for some $\omega \in N$ where $P(N) = 0$.
Note that in the Casella Berger book this is stated in the equivalent form

$$P\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1, \quad \forall\, \epsilon > 0.$$

Note that whenever we write $P(A)$ we should check that $A$ is in our sigma-algebra. Indeed, with $A := \{\omega : \lim_{n\to+\infty} X_n(\omega) = X(\omega)\}$ we have that $\omega \in A$ if and only if

$$\forall k \in \mathbb{N}\ \exists N \in \mathbb{N} \text{ s.t. } \forall n \geq N \quad |X_n(\omega) - X(\omega)| < \frac{1}{k}$$

and hence

$$A = \bigcap_{k\in\mathbb{N}} \bigcup_{N\in\mathbb{N}} \bigcap_{n\geq N} \left\{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \frac{1}{k}\right\}$$

is a measurable set (being the countable intersection of a countable union of a countable intersection of measurable sets!). Useful equivalent definitions are

$$P\left(|X_n - X| > \epsilon \quad \text{for infinitely many } n\right) = 0 \quad \text{for any } \epsilon > 0$$

and

$$\lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} |X_m - X| > \epsilon\right) = 0 \quad \text{for any } \epsilon > 0.$$

To see that the latter two definitions are equivalent, first consider an increasing sequence of events $B_n$, meaning that $B_i \subset B_{i+1}$ for each $i$. Using countable additivity it follows that (with $B_0 = \emptyset$)

$$P\left(\bigcup_n B_n\right) = P\left(\bigcup_n (B_n \backslash B_{n-1})\right) = \sum_n P(B_n \backslash B_{n-1}) = \lim_{n\to\infty} P(B_n).$$

A diagram might help here to see why the above is true and the final equality is an example of a so-called telescoping series. This is called the continuity property of probability. Next, note that $\bigcup_{m=n}^{\infty}\{|X_m - X| > \epsilon\}$ is a decreasing sequence of sets and by taking complements equivalence now follows (try filling in the details).

The remaining three modes of convergence are somewhat more straightforward.

**Convergence in probability**: the sequence $\{X_n\}$ converges to $X$ in probability if

$$\lim_{n\to\infty} P\left(|X_n - X| < \epsilon\right) = 1, \quad \forall\, \epsilon > 0,$$

and we use the notation $X_n \xrightarrow{P} X$.
An obviously equivalent definition is

$$\lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0, \quad \forall\, \epsilon > 0.$$

**Mean-square convergence**: the sequence $\{X_n\}$ converges to $X$ in mean-square if

$$\lim_{n\to\infty} \mathrm{E}\left[(X_n - X)^2\right] = 0,$$

and we use the notation $X_n \xrightarrow{m.s.} X$.

**Convergence in distribution**: the sequence $\{X_n\}$ converges to $X$ in distribution if

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t),$$

for any $t$ at which $F_X$ is continuous. We use the notation $X_n \xrightarrow{d} X$.

**Relations among the modes of convergence**:

1. if $X_n \xrightarrow{a.s.} X$ then $X_n \xrightarrow{P} X$;
2. if $X_n \xrightarrow{m.s.} X$ then $X_n \xrightarrow{P} X$;
3. if $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{d} X$.

**Proof**:

1. If $X_n$ converges to $X$ almost surely, this means that for any $\epsilon > 0$

   $$\lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} |X_m - X| > \epsilon\right) = 0.$$

   Since $\{|X_n - X| > \epsilon\} \subset \bigcup_{m=n}^{\infty}\{|X_m - X| > \epsilon\}$ it follows that

   $$\lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0,$$

   so $X_n$ converges to $X$ in probability.

2. From Chebyshev's inequality we know that for any $\epsilon > 0$

   $$P\left(|X_n - X| > \epsilon\right) \leq \frac{E[(X_n - X)^2]}{\epsilon^2} \to 0$$

   as $n \to \infty$ and hence mean-square convergence indeed implies convergence in probability.

3. Suppose for simplicity that $X_n$ and $X$ are continuous random variables and assume that $X_n \xrightarrow{P} X$. From the bounds

   $$P\left(X \leq t - \epsilon\right) \leq P\left(X_n \leq t\right) + P\left(|X_n - X| \geq \epsilon\right)$$

   and

   $$P\left(X_n \leq t\right) \leq P\left(X \leq t + \epsilon\right) + P\left(|X_n - X| \geq \epsilon\right)$$

   it follows by letting $\epsilon > 0$ arbitrarily small that

   $$P\left(X_n \leq t\right) \to P\left(X \leq t\right) \quad \text{as } n \to \infty.$$

   This argument can be adapted to the case when $X_n$ or $X$ are not continuous random variables as long as $t$ is a point of continuity of $F_X$.

Note that it follows that convergence in distribution is implied by any of the other modes of convergence. None of the other implications hold in general. For some of the examples and also for the proof of (a special case of) the Strong Law of Large Numbers the so-called Borel Cantelli Lemmas are incredibly useful.

## 19.1 Borel Cantelli Lemmas

The Borel Cantelli Lemmas are two fundamental lemmas in probability theory. Let $A_n$ be a sequence of events and denote by $A := \cap_n \cup_{m=n}^{\infty} A_m$ the event that infinitely many of the $A_n$ occur. The Borel Cantelli Lemmas give sufficient conditions on the $A_n$ under which either $P(A) = 0$ or $P(A) = 1$.

**Borel Cantelli 1**: Suppose $\sum_{n=1}^{\infty} P(A_n) < \infty$. Then $P(A) = 0$.

**Proof**: Note that since by definition $A \subset \cup_{m=n}^{\infty} A_m$ for each $n$, it follows that

$$P(A) \leq P(\bigcup_{m=n}^{\infty} A_m) \leq \sum_{m=n}^{\infty} P(A_m) \to 0 \quad \text{as } n \to \infty$$

since $\sum_{n=1}^{\infty} P(A_n) < \infty$.

**Borel Cantelli 2** Suppose that $A_1, A_2, \ldots$ are independent and $\sum_{n=1}^{\infty} P(A_n) = \infty$. Then $P(A) = 1$.

**Proof**: It suffices to show that $P(A^c) = 0$. Note that

$$\begin{aligned}
P(A^c) &= P\left(\bigcup_n \bigcap_{m=n}^{\infty} A_m^c\right) \\
&= \lim_{n\to\infty} P\left(\bigcap_{m=n}^{\infty} A_m^c\right) \quad (\text{as } \textstyle\bigcap_{m=n}^{\infty} A_m^c \text{ is increasing in } n) \\
&= \lim_{n\to\infty} \prod_{m=n}^{\infty} (1 - P(A_m)) \quad (\text{independence}) \\
&\leq \lim_{n\to\infty} \prod_{m=n}^{\infty} e^{-P(A_m)} \quad (\text{since } 1 - x \leq e^{-x}) \\
&= \lim_{n\to\infty} e^{-\sum_{m=n}^{\infty} P(A_m)} \\
&= 0
\end{aligned}$$

whenever $\sum_{n=1}^{\infty} P(A_n) = \infty$.

## 19.2 Examples of various modes of convergence

**Example "in probability" does not imply "almost surely"**

Let $X_n$ be independent Bernoulli random variables with parameter $p = 1/n$. Then it obviously holds that $X_n \xrightarrow{P} 0$ since $P(|X_n - 0| > \epsilon) = P(X_n = 1) = 1/n \to 0$ as $n \to \infty$. You may find it surprising (at least upon first reading) that $X_n$ **does not** converge to $0$ almost surely. Indeed, considering $A_n := \{X_n = 1\}$ it holds that

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

as the harmonic series diverges[3]. Now, from the second Borel Cantelli Lemma it follows that $P(X_n = 1 \text{ for infinitely many n}) = 1$, so $X_n$ does not converge to $0$ almost surely.

**Example "square mean" does not imply "almost surely"**

Let $X_n$ be defined as in the previous example. Then

$$E[(X_n - 0)^2] = \frac{1}{n} \to 0$$

as $n \to \infty$, so $X_n$ converges in mean square to $0$ but not almost surely.

**Example "in probability" does not imply "square mean"**

Convergence in probability only means that the probability that $X_n$ and $X$ differ by at most $\epsilon > 0$ goes to zero as $n \to \infty$, and, in particular, it does not lead to any restriction on the values of $X_n$ when it is not close to $X$. Take for example $X_n = 0$ with $p = 1 - 1/n$ and $X_n = n$ with $p = 1/n$. Again, it holds that have that $X_n \xrightarrow{P} 0$. However, since

$$E[(X_n - 0)^2] = \frac{n^2}{n} = n$$

does not converge to zero, the random variables $X_n$ do not converge to $0$ in square mean.

**Example "almost surely" does not imply "square mean"**

If we tweak $X_n$ and define the sequence now with $P(X_n = 0) = 1 - 1/n^2$ and $P(X_n = n) = 1/n^2$ we have that for any $\epsilon > 0$

$$P(|X_n - 0| > \epsilon) = \frac{1}{n^2}.$$

Since $\sum_{n=1}^{\infty} n^{-2} < \infty$, (in fact[4], it is $\pi^2/6$), it now follows from the first Borel Cantelli Lemma that

$$P\left(|X_n - 0| > \epsilon \quad \text{for infinitely many } n\right) = 0 \quad \text{for any } \epsilon > 0,$$

or equivalently, $X_n \xrightarrow{a.s.} 0$. On the other hand,

$$E[(X_n - 0)^2] = n^2/n^2 = 1$$

and so $X_n$ does not converge to $0$ in mean square.

**Example "in distribution" does not imply anything**

Let $Z$ be a standard normal random variable and let $X_n = (-1)^n Z$. Than $X_n$ converges in distribution to $Z$ but does not converge in any of the other three modes.

**Example "almost surely" implies "in probability"**

[3]for example, this follows form the fact that the harmonic series $1 + 1/2 + (1/3 + 1/4) + (1/5 + 1/6 + 1/7 + 1/8) + \ldots$ has a lower bound $1 + 1/2 + (1/4 + 1/4) + (1/8 + 1/8 + 1/8 + 1/8) + \ldots = 1 + 1/2 + 1/2 + 1/2 + \ldots = \infty$

[4]for various proofs of this surprising result see `http://empslocal.ex.ac.uk/people/staff/rjchapma/etc/zeta2.pdf`

Consider $X_n$ and $X \sim U[0,1]$ such that $X_n(\omega) = \omega + \omega^n$ and $X(\omega) = \omega$ for any $\omega \in [0,1]$. Then if $\omega \in [0,1)$ we have $\omega^n \to 0$ and so $X_n(\omega) \to X(\omega) = \omega$. When $\omega = 1$ we have $X_n(1) = 2$ but $X(1) = 1$. However, the set in which we have problems is $A = \{\omega \text{ s.t. } \omega = 1\}$ and we have

$$P(A) = 1 - P(A^c) = 1 - P(\{\omega \text{ s.t. } \omega \in [0,1)\}) = 1 - [F_X(1) - F_X(0)] = 1 - [1-0] = 0.$$

We have also convergence in probability. We can write $X_n = X + X^n$, then

$$\begin{aligned} P(|X_n - X| > \epsilon) &= P(|X^n| > \epsilon) = \\ &= P(X^n < -\epsilon \cup X^n > \epsilon) = \\ &= P(X < -\epsilon^{1/n} \cup X > \epsilon^{1/n}) = \\ &\to P(X < -1 \cup X > 1) = 0. \end{aligned}$$

**Example "in probability" does not imply "almost surely"**

Consider $X_n$ and $X \sim U[0,1]$ such that

$$\begin{aligned} X_1(\omega) &= \omega + I_{[0,1]}(\omega), \\ X_2(\omega) &= \omega + I_{[0,1/2]}(\omega), \\ X_3(\omega) &= \omega + I_{[1/2,1]}(\omega), \\ X_4(\omega) &= \omega + I_{[0,1/3]}(\omega), \\ X_5(\omega) &= \omega + I_{[1/3,2/3]}(\omega), \\ X_6(\omega) &= \omega + I_{[2/3,1]}(\omega). \end{aligned}$$

Define also $X(\omega) = \omega$. Let's compute the probability limit

$$P(|X_n - X| > \epsilon) = P(X + I_{\delta_n} - X > \epsilon) \to 0,$$

since $\delta_n$ is an interval that becomes smaller and smaller as $n \to \infty$. Then $X_n \to X$ in probability. However, for any $\omega$ we have an $n$ such that $X_n(\omega) = \omega$, $X_{n+1}(\omega) = \omega + 1$, and $X_{n+2}(\omega) = \omega$. Therefore the set of outcomes such that $X_n$ does not converge to $X$ is the whole sample space $[0,1]$ which implies that no almost sure convergence exists.

**Example "in probability" implies "in distribution"**

Convergence in probability implies convergence in distribution. Assume that $X_n \sim U[0,1]$ and are i.i.d. such that $X_n = \max_{1 \le i \le n} X_i$. We prove that $X_n$ converges in probability to the random variable $X = 1$.

$$\begin{aligned} P(|X_n - 1| > \epsilon) &= P(X_n - 1 > \epsilon \cup X_n - 1 < -\epsilon) = \\ &= P(X_n > \epsilon + 1) + P(X_n < 1 - \epsilon) = \\ &= 0 + P(\cap_{i=1}^n X_i < 1 - \epsilon) = \\ &= \prod_{i=1}^n P(X_i < 1 - \epsilon) = \\ &= (1 - \epsilon)^n \to 0. \end{aligned}$$

Then consider $\epsilon = t/n$

$$P(X_n \leq 1 - t/n) = (1 - t/n)^n \to e^{-t},$$

therefore

$$P(X_n \geq 1 - t/n) = P(n(1 - X_n) \leq t) \to 1 - e^{-t},$$

which is a cdf of an Exponential r.v. thus, $n(1 - X_n) \sim \text{Exp}(1)$.

**Example "in distribution" and continuity of cdf**

Define $X_n \sim U[1/2 - 1/n, 1/2 + 1/n]$ then as $n \to \infty$, $X_n \to X = 1/2$ in distribution, where the limiting r.v. is a degenerate r.v. with all its mass in $1/2$. We have

$$F_{X_n}(t) = \begin{cases} 0 & t \leq 1/2 - 1/n \text{ or } t \geq 1/2 + 1/n \\ \frac{t-(1/2-1/n)}{2/n} & t \in [1/2 - 1/n, 1/2 + 1/n]. \end{cases}$$

As $n \to \infty$ the cdf converges to $F_{X_n}(1/2) = 1/2$, however the limiting r.v. has cdf $F_X(1/2) = 1$ as all the mass of $X$ is in $t = 1/2$. So in $t = 1/2$ the cdf of $X_n$ does not converge to the cdf of $X$. However, $t = 1/2$ is a point where $F_X$ is not continuous, thus we still have convergence in distribution.

# 20 Two Laws of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with moments $\text{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$, such that $\sigma^2 < \infty$, for all $i$. We define $S_n = \sum_{i=1}^n X_i$ and $S_n/n$ is the sample mean (a random variable). Then we have two results.

**Weak Law of Large Numbers**:

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) = 1, \qquad \forall \epsilon > 0$$

that is $S_n/n \xrightarrow{P} \mu$.

**Proof**: for every $\epsilon > 0$, we use Chebychev's inequality

$$P(|S_n/n - \mu| > \epsilon) \leq \frac{\text{E}[(S_n/n - \mu)^2]}{\epsilon^2} = \frac{\text{Var}[S_n/n]}{\epsilon^2} = \frac{\text{Var}[S_n]}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

which converges to $0$ as $n$ goes to $\infty$.

Whereas the weak law of large numbers numbers is straightforward to prove, perhaps not surprisingly the strong law of large numbers requires some more effort.

**Strong Law of Large Numbers**:

$$P\left(\lim_{n\to\infty}\left|\frac{S_n}{n} - \mu\right| = 0\right) = 1,$$

that is $S_n/n \stackrel{a.s.}{\to} \mu$.

**Proof**: Here we give the proof in the case that we have the additional assumption that $E[X_i^4] < \infty$. In that case, note that

$$E[(S_n/n - \mu)^4] = \frac{1}{n^4} E\left[\left(\sum_{i=1}^{n}(X_i - \mu)\right)^4\right].$$

Note that this is a rather humongous sum. Justify (exercise) that it is equal to

$$\frac{1}{n^4}\left\{nE[(X-\mu)^4] + 3n(n-1)\left(E\left[(X-\mu)^2\right]\right)^2\right\}.$$

Note that this expression can be bounded by $Cn^{-2}$ for some $C > 0$ which does not depend on $n$. Using Chebyshev's inequality with $g(x) = x^4$ we have that for $\epsilon > 0$

$$P(|S_n/n - \mu| \geq \epsilon) \leq \frac{E[(S_n/n-\mu)^4]}{\epsilon^4} \leq \frac{C}{\epsilon^4 n^2}.$$

Since $1/n^2$ is summable we deduce from Borel Cantelli 1 that $S_n/n \stackrel{a.s.}{\to} \mu$. (to see why, reconsider the example above of almost sure convergence but not convergence in mean square).

**On the assumptions**: for the proof of weak and strong law above we have used the assumption of finite second and fourth moment, respectively. This is in fact stronger than what is needed. A sufficient condition is the weaker assumption $\mathrm{E}[|X|] < \infty$ ; the proof is much more demanding though.

The Strong Law of Large Numbers implies the Weak Law of Large Numbers and also convergence in distribution $S_n/n \stackrel{d}{\to} \mu$ which can be interpreted as convergence to the degenerate distribution with all of the mass concentrated at the single value $\mu$. We shall soon see that, just as in the case of the sum of Bernoulli random variables at the beginning, we can say a lot more about the limiting distribution of $S_n$ by proper rescaling. To be more specific, since $S_n/n - \mu$ converges to zero and since $\mathrm{Var}(S_n/n - \mu) = 1/n$, a scaling with factor $\sqrt{n}$, i.e. $\sqrt{n}(S_n/n - \mu)$ seems promising. This is the subject of the next section.

# 21 Central Limit Theorem

In this section we state and prove the fundamental result in probability and statistics, namely that the normalised sample mean from an i.i.d. sample (with finite variance) converges to a standard normal distribution. We shall make use of moment generating functions and the following result from the theory of so-called Laplace transforms.

**Convergence of mgfs (Theorem 2.3.12 in Casella Berger)** If $X_n$ is a sequence of random varaibles with a moment generating functions satisfying

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$$

for all $t$ in a neighbourhood of $0$ and if $M_X(t)$ is a moment generating function of a random variable $X$, then $X_n \xrightarrow{d} X$.

**Assumptions**: given an i.i.d. sequence of random variables $X_1, X_2, \ldots$ with finite variance $\sigma^2 > 0$, define $S_n = \sum_{i=1}^n X_i$.

**Central Limit Theorem**: if the mgf $M_X(t)$ of $X_i$ exists in some neighborhood of $0$, then, as $n \to +\infty$, and for $Z \sim N(0,1)$,

$$\sqrt{n}\frac{S_n/n - \mu}{\sigma} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} Z.$$

We can state the convergence in distribution as

$$P\left(\sqrt{n}\frac{S_n/n - \mu}{\sigma} \leq \alpha\right) \to \int_{-\infty}^{\alpha} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\mathrm{d}z, \quad \forall\, \alpha \in \mathbb{R}.$$

Notice that both $\mu$ and $\sigma^2$ exist and are finite since the mgf exists in a neighbourhood of $0$.

**Proof**: Define $Y_i = (X_i - \mu)/\sigma$, then

$$\sqrt{n}\frac{S_n/n - \mu}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$$

therefore the mgf of $Y_i$ exists for $t$ in some neighbourhood of $0$ (and we shall take $t$ sufficiently small from now on), given that $Y_i$ are i.i.d.

$$\begin{aligned} M_{\sqrt{n}\sigma^{-1}(S_n/n-\mu)}(t) &= M_{n^{-1/2}\sum_{i=1}^n Y_i}(t) = \\ &= \left(\mathrm{E}\left[\exp\left(tY_i/\sqrt{n}\right)\right]\right)^n = \\ &= \left(M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n. \end{aligned}$$

By expanding in Taylor series around $t = 0$, we have

$$M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=0}^{\infty} \mathrm{E}[Y_i^k]\frac{(t/\sqrt{n})^k}{k!}.$$

Now notice that $\mathrm{E}[Y_i] = 0$ and $\mathrm{Var}[Y_i] = 1$ for any $i$, thus

$$M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{(t/\sqrt{n})^2}{2} + o\left[\left(\frac{t}{\sqrt{n}}\right)^2\right],$$

where the last term is the remainder term in the Taylor expansion such that

$$\lim_{n\to\infty} \frac{o[(t/\sqrt{n})^2]}{(t/\sqrt{n})^2} = 0.$$

Since $t$ is fixed we also have

$$\lim_{n\to\infty} \frac{o[(t/\sqrt{n})^2]}{(1/\sqrt{n})^2} = \lim_{n\to\infty} n\, o\left[\left(\frac{t}{\sqrt{n}}\right)^2\right] = 0,$$

thus

$$\begin{aligned}\lim_{n\to\infty} M_{\sqrt{n}\sigma^{-1}(S_n/n-\mu)}(t) &= \lim_{n\to\infty}\left[M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right]^n = \\ &= \lim_{n\to\infty}\left\{1+\frac{1}{n}\left(\frac{t^2}{2}+n\, o\left[\left(\frac{t}{\sqrt{n}}\right)^2\right]\right)\right\}^n = e^{t^2/2},\end{aligned}$$

which is the mgf of a standard normal random variable. Therefore, by uniqueness of the moment generating function, $\sqrt{n}(S_n/n-\mu)/\sigma$ converges in distribution to a standard normal random variable.

**On the assumptions**:

1. we can relax the assumption of finite variances, it is enough to have $X_i$ that are small with respect to $S_n$; this can be assured by imposing two conditions by Lyapunov and Lindeberg of asymptotic negligibility;

2. Independence can also be relaxed by asking for asymptotic independence.

3. The assumption on the existence on moment generating functions can be dropped and a similar proof can be given in terms of the so-called characteristic function. This is defined similarly to the moment generating function by

   $$\Phi(t) := E[e^{\mathrm{i}tX}] \quad \text{for } t \in \mathbb{R}.$$

   Here $\mathrm{i} = \sqrt{-1}$ and $e^{ix} = \cos(x) + \mathrm{i}\sin(x)$ for $x \in \mathbb{R}$. The advantage of the characteristic function over the moment generating function is that the former always exists. This is due to the property that $|e^{ix}| = \cos^2(x) + \sin^2(x) = 1$ and hence $\Phi(t) \leq 1$. Characteristic functions also uniquely determine distributions and there is a convergence result equivalent to the one above for moment generating functions. Once you have calculated the moment generating functions, it is usually straightforward to find the characteristic function. For example, if $X$ is standard normal, then

   $$\Phi_X(t) = E[e^{\mathrm{i}tX}] = e^{(\mathrm{i}t)^2/2} = e^{-t^2/2}.$$

# Reading

Casella and Berger, Sections 5.5

# 22 Properties of a Random Sample

A description of a statistical analysis may be described as follows:

1. Consider a real-world phenomenon/problem/population with uncertainty and open questions.
2. Probability model: Identify the random variable(s) associated with the problem and assign a suitable probability model. The model is described by some parameter(s) $\theta$.
3. Draw a sample from the population.
4. Use the information contained in the sample to draw inference, that is gain knowledge for the population parameters $\theta$ and provide answers to the questions.

In this course we are concerned with parts 3 and 4.

## 22.1 Random Sample

A **sample** is a collection of random variables $X = (X_1, X_2, \ldots, X_n)$. An **observed sample** is a collection of observations $(x_1, x_2, \ldots, x_n)$ on $(X_1, X_2, \ldots, X_n)$.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables each with pdf or pmf $f(x_i|\theta)$. Then $X_1, X_2, \ldots, X_n$ are called **a random sample** of size $n$ from population $f(x_i|\theta)$.

- A random sample of size $n$ implies a particular probability model described by the population $f(x_i|\theta)$, that is by the marginal pdf or pmf of each $X_i$. Notice that it depends on some parameter $\theta$, and if we know $\theta$ then the model would be completely specified. However, $\theta$ is in general unknown and it is the object we are interested in estimating. For this reason we highlight the dependence on $\theta$ when indicating the pdf or pmf.
- The random sampling model describes an experiment where the variable of interest has a probability distribution described by $f(x_i|\theta)$.
- Each $X_i$ is an observation of the same variable.
- Each $X_i$ has a marginal distribution given by $f(x_i|\theta)$.
- The joint pdf or pmf is given by

$$f_X(x|\theta) \equiv f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta)f(x_2|\theta)\ldots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

**Example**

- Poisson:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

- Exponential:

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n|\lambda) = \prod_{i=1}^{n} \frac{1}{\lambda} e^{-x_i/\lambda} = \frac{1}{\lambda^n} e^{-\sum_{i=1}^{n} x_i/\lambda}$$

## 22.2 Statistics

Let $T(x_1, x_2, \ldots, x_n)$ be a real or vector valued function whose domain includes the sample space of $X_1, X_2, \ldots, X_n$. Then the random variable $Y = T(X_1, X_2, \ldots, X_n)$ is called a **statistic**.

- Inferential questions are hard to answer just by looking at the raw data.
- Statistics provide summaries of the information in the random sample.
- They could be arbitrary functions but they cannot be functions of parameters.

The following are three statistics often used

- The sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- The sample variance

$$S^2 = \frac{1}{n-1} \sum_{i-1}^{n} \left(X_i - \bar{X}\right)^2.$$

- The sample standard deviation $S = \sqrt{S^2}$.

Notice that these are random variables and we denote their observed values as $\bar{x}, s^2, s$.

**Lemma** Let $X_1, X_2, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $\mathrm{E}[g(X_1)]$ and $\mathrm{Var}(g(X_1))$ exist. Then,

$$\mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathrm{E}\left[g(X_1)\right]$$

and

$$\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathrm{Var}\left[g(X_1)\right].$$

**Theorem** Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean $\mu$ and variance $\sigma^2 < \infty$. Then

1. $\mathrm{E}(\bar{X}) = \mu$,
2. $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$,
3. $\mathrm{E}(S^2) = \sigma^2$.

As we will see later in detail we say that $\bar{X}$ and $S^2$ are unbiased estimators of $\mu$ and $\sigma^2$ respectively.

## 22.3 Sampling Distribution

The probability distribution of a statistic $T = T(X)$ is called the **sampling distribution** of $T$.

**Theorem** Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mgf $M_{X_i}(t)$. Then the mgf of the sample mean is

$$M_{\bar{X}}(t) = \left[M_{X_i}(t/n)\right]^n.$$

When applicable, the theorem above provides a very convenient way for deriving the sampling distribution.

**Example**

- $X_1, \ldots, X_n$ i.i.d. from $N(\mu, \sigma^2)$, then

$$M_{\bar{X}}(t) = \left[\exp\left(\mu\frac{t}{n} + \frac{\sigma^2(t/n)^2}{2}\right)\right]^n = \exp\left(\mu t + \frac{(\sigma^2/n)t^2}{2}\right)$$

  that is $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

- $X_1, \ldots, X_n$ i.i.d. from Gamma$(\alpha, \beta)$, then $\bar{X} \sim$ Gamma$(n\alpha, \beta/n)$.

If we cannot use the above theorem we can derive the distribution of the transformation of random variables by working directly with pdfs.

## 22.4 Transformation of Random Variables

### 22.4.1 Transformation of Scalar Random Variables

**Theorem:** Let $X$,$Y$ be random variables with pdfs $f_X(x)$, $f_Y(y)$ and defined for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, respectively. Suppose that $g(\cdot)$ is a monotone function such that $g : \mathcal{X} \to \mathcal{Y}$ and $g^{-1}(\cdot)$ has a continuous derivative on $\mathcal{Y}$. The pdf of $Y$ is then

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

**Proof:** Let $X$ be a random variable with density $f_X(x)$ and $Y = g(X)$ or $X = g^{-1}(Y)$. If $g^{-1}(\cdot)$ is increasing

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g^{-1}(Y) \leq g^{-1}(y)), \\ &= P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)), \end{aligned}$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}$$

If $g^{-1}(\cdot)$ is decreasing

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g^{-1}(Y) \geq g^{-1}(y)), \\ &= P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)), \end{aligned}$$

$$f_Y(y) = -\frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \left[-\frac{dg^{-1}(y)}{dy}\right]$$

(The derivative of a decreasing function is negative)
Putting both cases together if $g^{-1}(.)$ is monotone

$$f_Y(y) = f_X(g^{-1}(y)) \left|\frac{dg^{-1}(y)}{dy}\right|$$

**Example:** Inverse Gamma Distribution. Let $X \sim$ Gamma$(\alpha, \beta)$,

$$f_X(x|\alpha, \beta) = \frac{x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right)}{\Gamma(\alpha)\beta^\alpha}, \quad 0 < x < \infty,$$

We want the distribution of $Y = 1/X$, therefore $g(x) = 1/x$ and $g^{-1}(y) = 1/y$. Then, $\frac{d}{dy} g^{-1}(y) = -1/y^2$. We can therefore write

$$\begin{aligned} f_Y(y) = f_X(g^{-1}(y)) \left|\frac{dg^{-1}(y)}{dy}\right| &= \frac{\left(\frac{1}{y}\right)^{\alpha-1} \exp\left(-\frac{1}{y\beta}\right)}{\Gamma(\alpha)\beta^\alpha} \frac{1}{y^2}, \\ = \frac{\left(\frac{1}{y}\right)^{\alpha+1} \exp\left(-\frac{1}{y\beta}\right)}{\Gamma(\alpha)\beta^\alpha}, &\qquad 0 < y < \infty. \end{aligned}$$

**Square Transformations:** What if $g(\cdot)$ is not monotone? For example consider $Y = X^2$, then $g^{-1}(y) = \sqrt{y}$ and clearly $F_Y(y) = \mathrm{P}(X \leq \sqrt{y})$ is not defined if $y < 0$. For $y \geq 0$

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}), \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy}[F_X(\sqrt{y}) - F_X(-\sqrt{y})], \\ &= \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}), \text{ if } y \geq 0. \end{aligned}$$

**Example:** $\chi^2$ distribution from standard Normal.

Let $X \sim N(0,1)$, and consider $Y = X^2$. Using the previous result we get

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}} \frac{1}{\sqrt{2\pi}} \exp(-\sqrt{y}^2/2) + \frac{1}{2\sqrt{y}} \frac{1}{\sqrt{2\pi}} \exp(-(-\sqrt{y})^2/2), \\ &= \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right) \end{aligned}$$

Note that

$$f_Y(y) = \frac{1}{\Gamma(\frac{1}{2}) 2^{\frac{1}{2}}} y^{\frac{1}{2}-1} \exp\left(-\frac{y}{2}\right),$$

which is the pdf of a Gamma$\left(\frac{1}{2}, 2\right)$ distribution, or else a $\chi^2$ distribution with one degree of freedom.

### 22.4.2 Transformations of Multivariate Random Variables

Let $\mathbf{X} = (X_1, \ldots, X_d)'$ be d-dimensional random variable and $\mathbf{Y} = (Y_1, \ldots, Y_d)' = \mathbf{g}(\mathbf{X})$ so that $\mathbf{X} = \mathbf{g}^{-1}(\mathbf{Y})$, or else

$$\begin{aligned} X_1 &= g_1^{-1}(Y_1, \ldots, Y_d), \\ X_2 &= g_2^{-1}(Y_1, \ldots, Y_d), \\ &\vdots \\ X_d &= g_d^{-1}(Y_1, \ldots, Y_d). \end{aligned}$$

The transformation $\mathbf{g} : \mathcal{X} \subset \mathbb{R}^d \to \mathcal{Y} \subset \mathbb{R}^d$ from $X$ to $Y$ has to be **one-to-one**.
Consider the matrix with the partial derivatives

$$\frac{\partial g^{-1}(Y_1, \ldots, Y_n)}{\partial (Y_1, \ldots, Y_n)} = \begin{pmatrix} \frac{\partial g_1^{-1}(Y)}{\partial Y_1} & \frac{\partial g_1^{-1}(Y)}{\partial Y_2} & \cdots & \frac{\partial g_1^{-1}(Y)}{\partial Y_d} \\ \frac{\partial g_2^{-1}(Y)}{\partial Y_1} & \frac{\partial g_2^{-1}(Y)}{\partial Y_2} & \cdots & \frac{\partial g_2^{-1}(Y)}{\partial Y_d} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial g_d^{-1}(Y)}{\partial Y_1} & \frac{\partial g_d^{-1}(Y)}{\partial Y_2} & \cdots & \frac{\partial g_d^{-1}(Y)}{\partial Y_d} \end{pmatrix}$$

The **Jacobian**, **J** of the transformation $\mathbf{g}(\cdot)$ is the determinant of the matrix of derivatives above. It provides a scaling factor for the change of volume under the transformation.

**Formula for multivariate transformations:**

Using standard change of variables results from multivariate calculus, we get

$$f_{Y_1,\ldots,Y_d}(y_1,\ldots,y_d) = f_{X_1,\ldots,X_d}\left(g_1^{-1}(\mathbf{Y}),\ldots,g_d^{-1}(\mathbf{Y})\right)\ |\mathbf{J}|$$

**Theorem:** If $X, Y$ are independent random variables with pdfs $f_X(x)$ and $f_Y(y)$, the pdf of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{-\infty} f_X(w) f_Y(z-w) dw$$

This formula for $f_Z(z)$ is called the **convolution** of $f_X(x)$ and $f_Y(y)$.

**Proof (of the convolution expression):** We introduce an extra random variable $W = X$ so that

$$Z = X + Y, \text{ and } W = X \text{ or}$$
$$X = W, \text{ and } Y = Z - W.$$

The Jacobian is equal to 1. Since $X, Y$ are independent their joint pdf is $f_{XY}(x,y) = f_X(x) f_Y(y)$. We can now write

$$f_{ZW}(z,w) = f_{XY}(w, z-w) \times 1 = f_X(w) f_Y(z-w)$$

Finally,

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{ZW}(z,w) dw = \int_{-\infty}^{+\infty} f_X(w) f_Y(z-w) dw$$

**Example:** If $X$ and $Y$ are independent and identically distributed exponential random variables, find the joint density function of $U = X/Y$ and $V = X + Y$.
For $U = X/Y, V = X + Y$, the inverse transformation is $X = UV/(1+U), Y = V/(1+U)$. We have

$$\frac{\partial X}{\partial U} = V/(1+U)^2, \quad \frac{\partial X}{\partial V} = U/(1+U), \quad \frac{\partial Y}{\partial U} = -V/(1+U)^2, \quad \frac{\partial Y}{\partial V} = 1/(1+U).$$

The Jacobian is

$$\begin{vmatrix} V/(1+U)^2 & U/(1+U) \\ -V/(1+U)^2 & 1/(1+U) \end{vmatrix} = V(1+U)/(1+U)^3 = V/(1+U)^2$$

which is non-negative for $U, V \geq 0$.

For $U, V > 0$ the joint density is therefore

$$f_{U,V}(u,v) = f_{X,Y}(x,y)v/(1+u)^2 = \lambda^2 e^{-\lambda(x+y)}v/(1+u)^2 = \lambda^2 v e^{-\lambda v}/(1+u)^2.$$

The joint density factorises into a marginal density for $V$, which is Gamma with a scale parameter $\lambda$ and a shape parameter 2, and a Pareto density $1/(1+u)^2$ for $U$. So $U$ and $V$ are independent.

## 22.5 Sampling from the Normal distribution

**Theorem** Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution.

1. $\bar{X}$ and $S^2$ are independent random variables.
2. $\bar{X}$ has a $N\left(\mu, \frac{\sigma^2}{n}\right)$ distribution.
3. $(n-1)S^2/\sigma^2$ has chi squared distribution with $n-1$ degrees of freedom ($\chi^2_{n-1}$).

### 22.5.1 Chi-squared distribution

1. A $\chi^2_p$ distribution is a Gamma$(p/2, 2)$ distribution. Its pdf is

$$f(y) = \frac{1}{\Gamma(p/2)2^{p/2}} y^{(p/2)-1} e^{-y/2}, \ x > 0.$$

2. If $Z$ is a $N(0,1)$ random variable, then $Z^2 \sim \chi^2_1$.
3. If $X_1, X_2, \ldots, X_n$ are independent and $X_i \sim \chi^2_{p_i}$, then

$$X_1 + X_2 + \cdots + X_n \ \sim \ \chi^2_{p_1+\cdots+p_n}$$

4. Let $X$ be distributed according to a $\chi^2_p$. Then $\mathrm{E}(X) = p$ and, for $p > 2$, $\mathrm{E}(\frac{1}{X}) = 1/(p-2)$.

### 22.5.2 Student's $t$ distribution

Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. We know

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \ \sim \ N(0,1).$$

But $\sigma^2$ is usually unknown. It would be most useful if we knew the distribution of the statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

The distribution of $T$ is known as the $t$ distribution with $n-1$ degrees of freedom. Note that

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}},$$

where $Z \sim N(0,1)$, $V \sim \chi^2_{n-1}$, and $Z$, $V$ are independent because of the above theorem.

**Example** Let $Y$ be a random variable distributed according to Student's $t$ distribution with $p$ degrees of freedom. Show that

1. The pdf of $Y$ is

$$f_Y(y) = \frac{(2\pi)^{-1/2} 2^{-p/2} p^{-1/2}}{\Gamma(p/2)} \Gamma\left(\frac{p+1}{2}\right) \left[\frac{2}{1 + y^2/p}\right]^{(p+1)/2}$$

2. $\mathrm{E}(Y) = 0$, if $p > 1$.

3. $\mathrm{Var}(Y) = p/(p-2)$, if $p > 2$.

### 22.5.3 Snedecor's $F$ distribution

Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N(\mu_x, \sigma_x^2)$ population, and $Y_1, Y_2, \ldots, Y_m$ be a random sample from a $N(\mu_y, \sigma_y^2)$ population. Suppose that we want to compare the variability between the two populations. This can be done through the variance ratio $\sigma_x^2/\sigma_y^2$ but since these are unknown we can use $S_x^2/S_y^2$. The $F$ distribution compares these quantities by giving us the distribution of

$$F = \frac{S_x^2/S_y^2}{\sigma_x^2/\sigma_y^2}$$

Note that $F$ may be written as

$$F = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} = \frac{U/(n-1)}{V/(m-1)}$$

where $U \sim \chi^2_{n-1}$, $V \sim \chi^2_{m-1}$ and $U$, $V$ are independent.

**Example** Let $Y$ be a random variable distributed according to Snedecor's $F$ distribution with $p$ and $q$ degrees of freedom.

1. Find the pdf of $Y$.

2. Show that $\mathrm{E}(Y) = q/(q-2)$, if $q > 2$.

3. Show that $1/Y$ is again an $F$ distribution with $q$ and $p$ degrees of freedom.

4. Show that if $T \sim t_q$ then $T^2 \sim F_{1,q}$

## 22.6 Order Statistics

Let $X_1, X_2, \ldots, X_n$ be a random sample from population with distribution function $F(x)$ and density $f(x)$. Define $X_{(i)}$ to be the $i-$th smallest of the $\{X_i\}$ $(i = 1, \ldots, n)$, namely

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

We want to find the density function of $X_{(i)}$. Notice that while $X_i$ is one element of the random sample, $X_{(i)}$ is a statistic which is function of the whole random sample. Informally, we may write for small $\Delta x$,

$$f_{X_{(i)}}(x) \approx \frac{F_{X_{(i)}}(x + \Delta x) - F_{X_{(i)}}(x)}{\Delta x} = \frac{P[X_{(i)} \in (x, x + \Delta x)]}{\Delta x}.$$

The probability $X_{(i)}$ is in $(x, x + \Delta x)$ is roughly equal to the probability of $(i - 1)$ observations in $(-\infty, x)$, one in $(x, x + \Delta x)$ and the remaining $(n - i)$ in $(x + \Delta x, +\infty)$. This is a trinomial probability

$$\frac{P[X_{(i)} \in (x, x + \Delta x)]}{\Delta x} = \frac{n!}{(i-1)!1!(n-i)!} F(x)^{i-1}[1-F(x+\Delta x)]^{n-i} \frac{P(\text{one observation} \in (x, x + \Delta x))}{\Delta x}$$

As $\Delta x \to 0$, rigorous calculations provide

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1} f(x)[1 - F(x)]^{n-i}.$$

Notice that this formula is function of the population cdf and pdf, i.e. of a generic $X_i$.

**Example:** Let $X_1, X_2, \ldots, X_n$ be a random sample from $U(0, 1)$. Find the density of $X_{(i)}$.

The density of $X_{(i)}$ becomes (for $0 < y < 1$)

$$\begin{aligned} f_{X_{(i)}}(y) &= \frac{n!}{(i-1)!(n-i)!} x^{i-1} 1 (1 - x)^{n-i} \\ &= \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n+1-i)} x^{i-1} (1 - x)^{(n-i+1)-1} \end{aligned}$$

This is a Beta$(i, n - i + 1)$ distribution. We get

$$\text{E}[Y_{(i)}] = \frac{i}{n+1}, \qquad \text{Var}(Y_{(i)}) = \frac{j(n-j+1)}{(n+1)^2(n+2)}.$$

# Reading

G. Casella & R. L. Berger 2.1, 4.3, 4.6, 5.1, 5.2, 5.3, 5.4

# 23 The Sufficiency Principle

- Each statistic reduces the observed sample into a single number; **data reduction**.
- Inherently there is some loss of information.
- A good inferential procedure is based on statistics that do not throw too much information.

## 23.1 Sufficient Statistics

A **sufficient statistic** for a parameter $\theta$ captures, in a certain sense, all the relevant information in the sample about $\theta$.

**Sufficiency Principle:** If $T(Y)$ is a sufficient statistic for $\theta$ then any inference for $\theta$ should be based on the sample $Y$ only through $T(Y)$. That is if $x$ and $y$ are two observed samples (that is $x = (x_1 \dots x_n)$ and $y = (y_1 \dots y_n)$) such that $T(x) = T(y)$ then the inference about $\theta$ should be the same regardless if $Y = y$ or $Y = x$ was observed.

**Definition:** Let $Y = (Y_1, Y_2, \dots, Y_n)$ be a random sample. A statistic $U = T(Y)$ is a sufficient statistic for a parameter $\theta$ if the conditional distribution of $Y|U = u$ is independent of $\theta$. In other words if $f_Y(y|\theta)$ is the joint pdf or pmf of the sample $Y$, and $f_U(u|\theta)$ is the pdf or pmf of $U$, $U$ is a sufficient statistic if

$$f_{Y|U}(y|U = u; \theta) = \frac{f_{Y,U}(y, u|\theta)}{f_U(u|\theta)} = \frac{f_Y(y|\theta)}{f_U(u|\theta)}$$

is a constant as a function of $\theta$ for all $y$ (that is it does not depend on $\theta$). Notice that since $U = T(Y)$, we have $P(Y = y, U = u) = P(Y = y)$, indeed the event $\{Y = y\}$ is a subset of the event $\{T(Y) = T(y)\}$ (consider for example the case $T(Y) = \sum_{i=1}^{n} Y_i$).

**Notes**

- If Y is discrete, the ratio above is a conditional probability mass function.

$$P(Y = y|T(Y) = T(y)) = \frac{P(Y = y,\ T(Y) = T(y))}{P(T(Y) = T(y))}$$

- If it is continuous it is just a conditional pdf.
- The definition refers to the conditional distribution. A statistic is sometimes defined as being sufficient for a family of distributions, $F_Y(y|\theta)$, $\theta \in \Theta$.

**Example:** Let $Y = (Y_1, \dots, Y_n)$ be a random sample from a Poisson($\lambda$) population, and let $U = T(Y) = \sum_{i=1}^{n} Y_i$. It can be shown that $U \sim$Poisson($n\lambda$). We can also write

$$P(Y = y|U = u) = \frac{P(Y = y,\ U = u)}{P(U = u)},$$

and note that

$$P(Y = y,\ U = u) = \begin{cases} 0 & \text{if } U \neq u, \\ P(Y = y) & \text{if } U = u. \end{cases}$$

so we can then write

$$\begin{aligned} P(Y = y|U = u) &= \frac{P(Y = y)}{P(U = u)} = \frac{\prod_{i=1}^{n} \exp(-\lambda)\lambda^{y_i}/(y_i!)}{\exp(-n\lambda)(n\lambda)^u/(u!)} \\ &= \frac{u!}{n^u \prod_{i=1}^{n} y_i!} \end{aligned}$$

U is sufficient since there is no $\lambda$ in the above.

**Theorem (Factorization Theorem):** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample with joint pdf or pmf $f_Y(y|\theta)$. The statistic $T(Y)$ is sufficient for the parameter $\theta$ if and only if we can find functions $g(.)$ and $h(.)$ such that

$$f_Y(y|\theta) = g\left(T(y)|\theta\right) h(y)$$

for all $y \in \mathcal{R}^n$ and $\theta \in \Theta$. (Note that both $T$ and $\theta$ can be vectors)

We give the proof for a discrete valued $Y$. The proof for the continuous case is quite technical and beyond the scope of this course.

**Proof of Factorization Theorem:**

**Preliminaries:** Since $T(Y)$ is a function of $Y$ we can write

$$P(Y = y) = P(Y = y, T(Y) = T(y)) \tag{1}$$

but NOT

$$P(T(Y) = T(y)) = P(Y = y, T(Y) = T(y))$$

Indeed,

$$\begin{aligned} P(T(Y) = T(y)) &= \sum_{y_k : T(y_k) = T(y)} P(Y_k = y_k, T(Y_k) = T(y)) \\ &= \sum_{y_k : T(y_k) = T(y)} P(Y_k = y_k) \end{aligned} \tag{2}$$

That is the event $\{Y = y\}$ is a subset of the event $\{T(Y) = T(y)\}$ but not the viceversa.

**If $T$ is sufficient:** Suppose $T$ is sufficient for $\theta$. That is $P(Y = y|T(Y) = T(y))$ is independent of $\theta$. We can write

$$\begin{aligned} P_\theta(Y = y) &\overset{(1)}{=} P_\theta(Y = y, T(Y) = T(y)) \\ &= P_\theta(T(Y) = T(y))P(Y = y|T(Y) = T(y)) \\ &= g(T(Y), \theta)h(Y). \end{aligned}$$

since the pmf $P_\theta(T(Y) = T(y))$ is typically a function of $T(Y)$ and $\theta$ whereas $P(Y = y|T(Y) = T(y))$ is independent of $\theta$ due to the sufficiency of $T$. Note that the function $g(T(Y), \theta)$ is the pmf of $T(Y)$.

**Converse:** Suppose $P_\theta(Y = y) = g(T(y), \theta)h(y)$. The conditional pmf

$$\begin{aligned} P_\theta(Y = y|T(Y) = T(y)) &= \frac{P_\theta(Y = y, T(Y) = T(y))}{P_\theta(T(Y) = T(y))} \\ &\overset{(1)(2)}{=} \frac{P_\theta(Y = y)}{\sum_{y_k:T(y_k)=T(y)} P_\theta(Y_k = y_k)} \\ &= \frac{g(T(y), \theta)h(y)}{\sum_{y_k:T(y_k)=T(y)} g(T(y), \theta)h(y_k)} \\ &= \frac{g(T(y), \theta)h(y)}{g(T(y), \theta)\sum_{y_k:T(y_k)=T(y)} h(y_k)} = \frac{h(y)}{\sum_{y_k:T(y_k)=T(y)} h(y_k)}, \end{aligned}$$

which is independent of $\theta$. Hence $T(Y)$ is a sufficient statistic.

**Example:** Let $Y = (Y_1, \ldots, Y_n)$ be a random sample from the following distributions find a sufficient statistic for each case.

1. Sufficient statistic for $\mu$ from a $N(\mu, \sigma^2)$ population with $\sigma^2$ known.

   The joint density may be written as

   $$\begin{aligned} f_Y(Y|\mu) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n(\bar{Y} - \mu)^2 + (n-1)S^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(\frac{(n-1)S^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{Y} - \mu)^2}{2\sigma^2}\right) \end{aligned}$$

   The statistic $T(Y) = \bar{Y}$ is sufficient for $\mu$ since if we set $g(T(Y), \mu) = \exp\left(-\frac{n(\bar{Y}-\mu)^2}{2\sigma^2}\right)$ and $h(Y) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{(n-1)S^2}{2\sigma^2}\right)$, we have $f_Y(Y|\mu) = g(T(Y), \mu)h(Y)$.

2. Sufficient statistic for $(\mu, \sigma^2)$ from a $N(\mu, \sigma^2)$ population.

   The joint density may be written as

$$f_Y(Y|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n(\bar{Y} - \mu)^2 + (n-1)S^2}{2\sigma^2}\right)$$

   The statistic $T(Y) = (\bar{Y}, S^2)$ is sufficient for $(\mu, \sigma^2)$ since if we set $g(T(Y), \mu, \sigma^2) = f_Y(Y|\mu, \sigma^2)$ and $h(Y) = 1$, we have $f_Y(Y|\mu, \sigma^2) = g(T(Y), \mu, \sigma^2)h(Y)$.

   **Note:** The statistic $\bar{Y}$ is sufficient for $\mu$ but not for $(\mu, \sigma^2)$.

3. Sufficient statistic for $\theta$ from Unif$(0, \theta)$.

   Let $I(Y \in A)$ denote an indicator function, that is a function that equals 1 if $Y \in A$ and 0 otherwise.

   The joint density may be written as

$$f_Y(y|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I(y_i > 0) I(y_i < \theta) = \frac{1}{\theta^n} I(\max_i y_i < \theta) I(\min_i y_i > 0)$$

   The statistic $T(Y) = \max_i y_i$ is sufficient for $\theta$ since if we set $g(T(Y), \theta) = \frac{1}{\theta^n} I(\max_i y_i < \theta)$ and $h(Y) = I(\min_i y_i > 0)$, we have $f_Y(Y|\theta) = g(T(Y), \theta)h(Y)$.

## 23.2 Minimal Sufficiency

**Example (Sufficiency of the sample):** Let $Y = (Y_1, \ldots, Y_n)$ is a sample from a population with $f_{Y_i}(y_i|\theta)$. Denote the joint density of the sample $Y$ by $f_Y(y|\theta)$.

Note that

$$f_Y(y|\theta) = f_Y(y|\theta) \times 1 = g(T(y)|\theta) \times h(y),$$

where

$$T(Y) = Y, \quad g(T(y)|\theta) = f_Y(y|\theta), \quad h(y) = 1.$$

Every sample is itself a sufficient statistic. Also every statistic that is a one-to-one function of a sufficient statistic is itself a sufficient statistic.

- There exist many sufficient statistics.
- Ideally, we would like the simplest possible sufficient statistic.
- If a statistic $T$ is a function of a statistic $S$, then it contains no more information than $S$.
- We look at those sufficient statistics which are still sufficient even though they are functions of other statistics.

**Definition:** A sufficient statistic $T(Y)$ is a minimal sufficient statistic if for any other sufficient statistic $T'(Y)$, $T(Y)$ is a function of $T'(Y)$.

**Some facts about minimal sufficient statistics**

- If a sufficient statistic has dimension 1, it must be a minimal sufficient statistic.
- Minimal sufficient statistics are not unique. However if two statistics are minimally sufficient they must have the same dimension.
- A one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic
- The dimension of a minimal sufficient statistic is not always the same as the dimension of the parameter of interest.

# Reading

G. Casella & R. L. Berger 3.4, 6.1, 6.2.1, 6.2.2

# 24 Point Estimation

**Problem:**

- Suppose that a real world phenomenon may be described by a probability model defined through the random variable $Y$ with $F_Y(y|\theta)$.
- Suppose also that a sample $Y = (Y_1, Y_2, \ldots, Y_n)$ is drawn from that distribution.
- We want use the information in the random sample $Y$ to get a best guess for $\theta$. In other words we want a **point estimate** for $\theta$.
- A function of $Y$ that gives a point estimate $\theta$ is an **estimator** of $\theta$. If we use the observed random sample $Y = y$, we get the **estimate** of $\theta$ which is a particular value of the **estimator**.

Next, we look at two methods for finding point estimators, the method of moments and the maximum likelihood estimators. Then we present evaluation methods for estimators.

## 24.1 Method of Moments

**Description:** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from population with pdf or pmf $f(y|\theta_1, \ldots, \theta_k)$. Let a sample moment be defined as

$$m_r = \frac{1}{n}\sum_{i=1}^{n} Y_i^r$$

Remember that the $r$-th (population) moment is

$$\mu_r = \mu_r(\theta_1, \ldots, \theta_k) = E_\theta(Y_i^r)$$

Method of moments estimators are found by equating the **first** $k$ **sample moments** to the corresponding $k$ **population moments** and solving the resulting system of simultaneous equations.

**Example:** $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from $N(\mu, \sigma^2)$ population. Find estimators for $\mu$ and $\sigma^2$ using the method of moments.

We want estimators for 2 parameters. Hence, we first write down the system of 2 equations

$$\bar{X} = E(X) = \mu,$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = E(X^2) = \mu^2 + \sigma^2,$$

and after solving it we get the following estimators

$$\hat{\mu} = \bar{X},$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$$

## 24.2 The Likelihood Function

**Definition:** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a sample from population with pdf (or pmf) $f(y_i|\theta)$. Then, given $Y = y$ is observed, the function of $\theta$ defined by the joint pdf (or pmf) of $Y = y$

$$L(\theta|Y = y) = f_Y(y|\theta)$$

is called the **likelihood function**.

**Notes**

- In most cases the pdf of $Y$ is thought as a function of $Y$ whereas the likelihood function is thought as a function of $\theta$ for a given observed sample.
- If for $\theta_1$, $\theta_2$ we have $L(\theta_1|y) > L(\theta_2|y)$ then the sample is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$. In other words $\theta_1$ is a more plausible value than $\theta_2$.
- The likelihood, as a function of $\theta$ is not always a pdf.
- Sometimes it is more convenient to work with the **log-likelihood**, $l(\theta|y)$ which is just the log of the likelihood.
- If $Y = (Y_1, Y_2, \ldots, Y_n)$ is a random sample from $f(y_i|\theta)$, then

$$L(\theta|Y = y) = f_Y(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

$$l(\theta|Y = y) = \log f_Y(y|\theta) = \sum_{i=1}^{n} \log f(y_i|\theta)$$

**Example:** Consider $X$ continuous random variable with pdf $f_X(x|\theta)$, then for small $\epsilon$ we have

$$\frac{P_\theta(x - \epsilon < X < x + \epsilon)}{2\epsilon} \simeq f_X(x|\theta) = L(\theta|x)$$

therefore if we compare the probabilities for different values of $\theta$ we have

$$\frac{P_{\theta_1}(x - \epsilon < X < x + \epsilon)}{P_{\theta_2}(x - \epsilon < X < x + \epsilon)} \simeq \frac{L(\theta_1|x)}{L(\theta_2|x)}$$

and the value of $\theta$ which gives higher likelihood is more likely to be associated to the observed sample since gives a higher probability.

**Example:** Likelihood and log-likelihood for exponential($\lambda$):

$$L(\lambda|Y = y) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right), \quad l(\lambda|Y = y) = n \log \lambda - \lambda \sum_{i=1}^{n} y_i$$

**Example:** Likelihood and log-likelihood for $N(\mu, \sigma^2)$:

$$L(\mu, \sigma^2|Y = y) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right)$$

$$l(\mu, \sigma^2|Y = y) = -\frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

## 24.3 Score Function and Fisher's Information

**Definition:** The **score function** associated with the log-likelihood $l(\theta|y)$ is

$$s(\theta|y) = \frac{\partial l(\theta|y)}{\partial \theta} = \frac{1}{L(\theta|y)} \frac{\partial L(\theta|y)}{\partial \theta},$$

**Proposition:** $E(s(\theta|Y)) = 0$.

**Proof:** (for the continuous case)

$$E[s(\theta|Y)] = \int_{\mathbb{R}^n} s(\theta|y) f(y|\theta) dy = \int_{\mathbb{R}^n} \frac{\frac{\partial L(\theta|y)}{\partial \theta}}{L(\theta|y)} f(y|\theta) dy = \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} f(y|\theta) dy = \frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} f(y|\theta) dy = 0,$$

because $L(\theta|y) = f(y|\theta)$ and the last integral is equal to one, since the pdf is normalised. Here it has to be intended that $\theta$ is the true value of the unknown parameters.

**Notes:**

1. For the discrete case replace the integrals with sums.
2. Although the score function is usually viewed as a function of $\theta$, the expectation is taken with respect to $Y$, actually with respect to the distribution of $Y$ which depends on $\theta$. This may be interpreted as follows. If the experiment was repeated many times the score function would on average equal 0. That is, if we start at the true value of the parameters, on average over many experiments the likelihood does not change if we make an infinitesimal change of the parameter.

In most cases, if we plot the likelihood function against $\theta$, we get a curve with a peak in the maximum. The sharper the peak is, the more information about $\theta$ exists in the sample. This is captured by the **Fisher's information**:

$$\mathcal{I}(\theta|y) \equiv \mathcal{I}(\theta|Y = y) = E\left[s(\theta|Y)^2\right] = E\left[\left(\frac{\partial}{\partial \theta} l(\theta|Y)\right)^2\right]$$

This is the variance of the score (when computed at the true value of $\theta$), so the larger it is the more the score in the true value is affected by minimal changes in the parameters, the sharper is the peak, the more precise is our information about $\theta$.

**Proposition:** Show that under regularity conditions (e.g. exponential family)

$$\mathcal{I}(\theta|y) = E\left[s(\theta|Y)^2\right] = -E\left[\frac{\partial^2}{\partial \theta^2} l(\theta|Y)\right] = -E\left[\frac{\partial}{\partial \theta} s(\theta|Y)\right]$$

again here it has to be intended that $\theta$ is the true value of the unknown parameters.

In this case at the true value of $\theta$ the Fisher info is also the negative Hessian of the log-likelihood so it measures the concavity of the log-likelihood. In particular since the Fisher information must be always positive (it is a variance), then the Hessian must be negative which jointly with a zero expectation of the score (first derivative of the log-likelihood) tells us that the true value of $\theta$ is a maximum of the log-likelihood for any realisation $Y = y$.

**Proof:**

$$\begin{aligned}
0 &= \frac{d}{d\theta} E\left[s(\theta|y)\right] = \frac{d}{d\theta}\int_{\mathbb{R}^n} s(\theta|y) f(y|\theta) dy = \int_{\mathbb{R}^n} \frac{d}{d\theta}\left[s(\theta|y) f(y|\theta)\right] dy \\
&= \int_{\mathbb{R}^n} \left[\left(\frac{d}{d\theta} s(\theta|y)\right) f(y|\theta) + s(\theta|y)\frac{d}{d\theta} f(y|\theta)\right] dy \\
&= \int_{\mathbb{R}^n} \left[\frac{d}{d\theta} s(\theta|y) + s(\theta|y)^2\right] f(y|\theta) dy, \quad \left(\frac{d}{d\theta} f(y|\theta) = s(\theta|y) f(y|\theta)\right) \\
&= E\left[\frac{d}{d\theta} s(\theta|y) + s(\theta|y)^2\right] = E\left[\frac{d}{d\theta} s(\theta|y)\right] + E\left[s(\theta|y)^2\right]
\end{aligned}$$

Let $Y = (Y_1, \ldots, Y_n)$ be a random sample from a pdf with $f_{Y_i}(y_i|\theta)$. Denote with $s(\theta|y_i)$ and $\mathcal{I}(\theta|y_i)$ the Score function and Fisher information for $Y_i = y_i$ respectively. Then, for a realisation of the random sample we have

$$s(\theta|y) = \sum_{i=1}^{n} s(\theta|y_i), \qquad \mathcal{I}(\theta|y) = n\mathcal{I}(\theta|y_i).$$

**Proof:** The log-likelihood function is

$$\ell(\theta|Y) = \log\left(\prod_{i=1}^{n} f(Y_i|\theta)\right) = \sum_{i=1}^{n} \ell(\theta|Y_i)$$

Hence

$$s(\theta|Y) = \frac{\partial \ell(\theta|Y)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ell(\theta|Y_i)}{\partial \theta} = \sum_{i=1}^{n} s(\theta|Y_i)$$

For the Fisher information, using the fact that $(Y_1, \ldots, Y_n)$ are i.i.d., we have

$$\mathcal{I}(\theta|y) = E\left[\left(\sum_{i=1}^{n} s(\theta|Y_i)\right)^2\right] = \sum_{i=1}^{n} E\left[(s(\theta|Y_i))^2\right] = n\mathcal{I}(\theta|y_i),$$

Or alternatively, using the Hessian,

$$\begin{aligned}
\mathcal{I}(\theta|Y) &= -E\left(\frac{\partial s(\theta|Y)}{\partial \theta}\right) = -E\left(\sum_{i=1}^{n} \frac{\partial s(\theta|Y_i)}{\partial \theta}\right) \\
&= \sum_{i=1}^{n} -E\left(\frac{\partial s(\theta|Y_i)}{\partial \theta}\right) = \sum_{i=1}^{n} \mathcal{I}(\theta|y_i) = n\mathcal{I}(\theta|y_i).
\end{aligned}$$

**Example:** Let $Y = (Y_1, \ldots, Y_n)$ be a random sample from an Exp($\lambda$). Show that the score function is

$$s(\theta|y) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i,$$

and the Fisher's Information matrix is

$$\mathcal{I}(\theta|y) = \frac{n}{\lambda^2}$$

**Vector parameter case** If $\theta = (\theta_1, \ldots, \theta_p)'$, then the score function is the vector

$$s(\theta|Y) = \nabla_\theta \ell(\theta|Y) = \left( \frac{\partial}{\partial \theta_1} \ell(\theta|Y), \ldots, \frac{\partial}{\partial \theta_p} \ell(\theta|Y) \right)'$$

and Fisher's information is the matrix

$$E\left[ s(\theta|Y) s(\theta|Y)' \right].$$

The $(i,j)^{th}$ element of the Fisher's information matrix is

$$\left[ \mathcal{I}(\theta|y) \right]_{ij} = E\left[ \frac{\partial}{\partial \theta_i} l(\theta|Y) \frac{\partial}{\partial \theta_j} l(\theta|Y) \right]$$

It also holds that

1. $E\left[ s(\theta|Y) \right] = 0_p$
2. $\mathcal{I}(\theta|y) = V\left[ s(\theta|Y) \right], \quad V[.]$ denotes a covariance matrix
3. $\mathcal{I}(\theta|y) = -E\left[ \nabla_\theta \nabla'_\theta \ell(\theta|Y) \right] = -E\left[ \nabla_\theta s(\theta|Y) \right]$

   or else

$$\left[ \mathcal{I}(\theta|y) \right]_{i,j} = -E\left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta|Y) \right]$$

**Example:** Let $Y = (Y_1, \ldots, Y_n)$ be a random sample from a $N(\mu, \sigma^2)$. Show that the score function is

$$s(\theta|y) = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu), \\ \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 \end{pmatrix}$$

and the Fisher's Information matrix is

$$\mathcal{I}(\theta|y) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

## 24.4 Maximum Likelihood Estimators

We have seen that the true value of the parameter $\theta$ must be such that the log-likelihood attains its maximum. This motivates mathematically the definition of maximum likelihood estimator.

**Definition:** For each sample point $Y = y$ let $\hat{\theta}(y)$ be the parameter value at which the likelihood $L(\theta|y)$ attains its maximum as a function of $\theta$. A **maximum likelihood estimator** of the parameter $\theta$ based on $Y$ is the function $\hat{\theta}(Y)$.

**Maximization:** In general the likelihood function can be maximized using numerical methods. However if the function is differentiable in $\theta$, calculus may be used. The values of $\theta$ such that

$$s(\theta|y) = \frac{\partial \ell(\theta|y)}{\partial \theta} = 0,$$

are **possible candidates**. These points may not correspond to the maximum because

1. They may correspond to the minimum. The second derivative must also be checked.
2. The zeros of the first derivative locate only local maxima, we want a global maximum.
3. The maximum may be at the boundary where the first derivative may not be 0.
4. These points may be outside the parameter range.

Notice that, an application of the Weak Law of Large Numbers tells us that, as $n \to \infty$, we must have

$$\frac{1}{n}s(\theta|Y) = \frac{1}{n}\sum_{i=1}^{n} s(\theta|Y_i) \xrightarrow{p} E\left[s(\theta|Y_i)\right] = 0$$

which justifies our necessary condition.

**Example:** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from $N(\mu, 1)$, $-\infty < \mu < +\infty$. Find the MLE for $\mu$. The log likelihood function is equal to

$$\ell(\mu|y) = \text{const.} - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2 = -\frac{1}{2}\sum_{i=1}^{n} y_i^2 + \mu\sum_{i=1}^{n} y_i - \frac{1}{2}n\mu^2$$

Setting the score function equal to 0 yields a candidate for the global maximum:

$$\frac{\partial}{\partial\mu}\ell(\hat{\mu}|y) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} y_i - n\hat{\mu} = 0 \quad \Rightarrow \quad \hat{\mu} = \bar{Y}.$$

We could check whether it corresponds to a maximum (and not a minimum) if the second derivative of the log-likelihood is negative

$$\frac{\partial^2}{\partial\mu^2}\ell(\hat{\mu}|y) = -n < 0$$

The MLE for $\mu$ is $\hat{\mu} = \bar{Y}$ (In fact more checking is required but it is omitted for simplicity).

**Example:** We cannot always use the above calculus recipe. For example let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from $U(0, \theta)$. Assume to observe $Y = y$ and rank the realisations as $y_{(1)} \leq \ldots \leq y_{(n)}$. These are then realisations of the order statistics $Y_{(i)}$. The likelihood for $\theta$ given $Y = y$ is

$$L(\theta|y) = \theta^{-n} I(y_{(1)} \geq 0) I(y_{(n)} \leq \theta)$$

and the log-likelihood for $Y_{(1)} \geq 0$ is (notice that by construction all realisations are such that $y_{(1)} \geq 0$)

$$\ell(\theta|y) = -n\log(\theta) \quad \text{if } \theta \geq y_{(n)},$$

The function $\ell(\theta|y)$ is maximized at $\hat{\theta} = y_{(n)}$ which is our estimate. Hence $\hat{\theta} = Y_{(n)}$ is the MLE.

**Induced likelihood:** Let $Y$ be a sample with likelihood $L(\theta|y)$ and let $\lambda = g(\theta)$. The induced likelihood for $\lambda$ given $Y = y$ is

$$L^*(\lambda|Y = y) = \sup_{\theta: g(\theta) = \lambda} L(\theta|Y = y)$$

**Theorem (Invariance property of the MLE's):** If $\hat{\theta}$ is the MLE for $\theta$, then for any function $g(.)$ the MLE of $g(\theta)$ is $g(\hat{\theta})$.

**Example:** MLE for $\mu^2$ in N$(\mu, 1)$ case is $\hat{Y}^2$, MLE for $p/(1-p)$ in Binomial$(n, p)$ is $\hat{p}/(1-\hat{p})$ etc.

## 24.5 Evaluating Estimators

Being a function of the sample, an estimator is itself a random variable. Hence it has a mean and a variance. Let $\hat{\theta}$ be an estimator of $\theta$. The quantity below

$$E(\hat{\theta} - \theta),$$

is termed as the **bias** of the estimator $\hat{\theta}$. If $E(\hat{\theta}) = \theta$ the estimator is **unbiased**.

Estimators are usually evaluated based on their bias and variance.

The **mean squared error (MSE)** of an estimator $\hat{\theta}$ is the function of $\theta$ defined by

$$MSE(\theta) = E(\hat{\theta} - \theta)^2.$$

Note that

$$\begin{aligned} E(\hat{\theta} - \theta)^2 &= E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E\{[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 + 2[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta]\} \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + E[(Bias)^2] + 2[E(\hat{\theta}) - E(\hat{\theta})][E(\hat{\theta}) - E(\theta)] \\ &= Variance + (Bias)^2 \end{aligned}$$

An estimator $\hat{\theta}_1$ is uniformly better than $\hat{\theta}_2$ if it has smaller MSE for all $\theta$.

**Example:** Compare the MSE's of $S^2$, $\frac{n-1}{n}S^2$, and $\hat{\sigma}^2 = 1$ as estimators for $\sigma^2$ in the presence of a random sample $Y$ of size $n$ from $N(\mu, \sigma^2)$

$$E\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2,$$

Therefore $\frac{n-1}{n}S^2$ is biased.

$$\text{MSE}(S^2) = ... = \frac{2}{n-1}\sigma^4 > \frac{2n-1}{n^2}\sigma^4 = MSE\left(\frac{n-1}{n}S^2\right), \forall\sigma^2, \quad n = 2, 3, \ldots$$

Thus $\frac{n-1}{n}S^2$ is uniformly better than $S^2$. But it is not uniformly better than $\hat{\sigma}^2 = 1$ which has zero MSE when $\sigma^2 = 1$.

## 24.6 Best Unbiased Estimators

As seen from the previous example, we cannot find a 'uniformly best' estimator. Hence we restrict attention to unbiased estimators. The MSE of an unbiased estimator is equal to its variance. A **best unbiased estimator** is also termed as a **minimum variance unbiased estimator**.

**Theorem (Cramér - Rao inequality):** Let $Y = (Y_1, \ldots, Y_n)$ be a sample and $U = h(Y)$ be an unbiased estimator of $g(\theta)$. Under regularity conditions the following holds for all $\theta$

$$V(U) \geq \frac{\left[\frac{\partial}{\partial\theta} g(\theta)\right]^2}{\mathcal{I}(\theta|y)}.$$

Note that if $g(\theta) = \theta$ we get

$$V(U) \geq \frac{1}{\mathcal{I}(\theta|y)},$$

and if also $Y = (Y_1, Y_2, \ldots, Y_n)$ is a random sample

$$V(U) \geq \frac{1}{n\mathcal{I}(\theta|y_i)}.$$

**Proof:** We know that when $\theta$ is the true unknown value then

$$|\text{Corr(X,Y)}| \leq 1 \;\Rightarrow\; \text{Cov}(X,Y)^2 \leq V(X)V(Y), \tag{3}$$

$$s(\theta|y) = \frac{\frac{\partial}{\partial\theta} f(y|\theta)}{f(y|\theta)} \;\Rightarrow\; \frac{\partial}{\partial\theta} f(y|\theta) = s(\theta|y) f(y|\theta), \tag{4}$$

$$E[s(\theta|Y)] = 0, \tag{5}$$

$$V[s(\theta|Y)] = \mathcal{I}(\theta|y), \tag{6}$$

and by assumption

$$E(U) = E[h(Y)] = g(\theta). \tag{7}$$

$$\begin{aligned}
\text{Cov}[h(Y), s(\theta|Y)] &= E[h(Y)s(\theta|Y)] - E[h(Y)]E[s(\theta|Y)] \\
&\overset{(5)}{=} E[h(Y)s(\theta|Y)] = \int_{\mathbb{R}^n} h(y)s(\theta|y)f(y|\theta)dy \\
&\overset{(4)}{=} \int_{\mathbb{R}^n} h(y)\frac{\partial}{\partial\theta} f(y|\theta)dy = \frac{\partial}{\partial\theta} E[h(Y)] \overset{(7)}{=} \frac{\partial}{\partial\theta} g(\theta)
\end{aligned}$$

If we replace into inequality (3), we will get

$$\text{Cov}[h(Y), s(\theta|Y)]^2 = \left[\frac{\partial}{\partial\theta} g(\theta)\right]^2 \leq V[h(Y)]V[s(\theta|Y)] \Rightarrow$$
$$\overset{(6)}{\Rightarrow} \left[\frac{\partial}{\partial\theta} g(\theta)\right]^2 \leq V(U)\mathcal{I}(\theta|y).$$

**Example:** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from $N(\mu, 1)$.

$$\ell(\mu|y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

$$\frac{\partial}{\partial\mu}\ell(\mu|y) = \sum_{i=1}^{n}(y_i - \mu) = \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \mu)$$
$$= \sum_{i=1}^{n}(y_i - \bar{y}) + \sum_{i=1}^{n}(\bar{y} - \mu) = n(\bar{y} - \mu).$$

$$\mathcal{I}(\mu|y) = -E\left(\frac{\partial}{\partial\mu} n(\bar{Y} - \mu)\right) = -E(-n) = n.$$

Hence the Cramér - Rao lower bound for $\mu$ is $1/n$.

Consider $\hat{\mu} = \bar{Y}$ as an estimator for $\mu$.

$$E(\bar{Y}) = \mu,$$
$$V(\bar{Y}) = \frac{1}{n}.$$

Since $\hat{\mu} = \bar{Y}$ is unbiased and attains the Cramér - Rao lower bound for $\mu$, it is also a MVUE for $\mu$.

**Example:** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from Poisson($\lambda$). It is not hard to check that $\mathcal{I}(\lambda|y) = n/\lambda$. Both the mean and the variance of a Poisson distribution are equal to $\lambda$. Hence

$$E(\bar{Y}) = E(Y_i) = \lambda,$$
$$E(S^2) = V(Y_i) = \lambda.$$

Consider the estimators $\hat{\lambda}_1 = \bar{Y}$ and $\hat{\lambda}_2 = S^2$. They are both unbiased. Which one to choose?

$$V(\bar{Y}) = \frac{V(Y_i)}{n} = \frac{\lambda}{n}.$$

Since $\hat{\lambda}_1$ is unbiased and attains the Cramér - Rao lower bound for $\lambda$, it is also a MVUE for $\lambda$.

**Theorem (Cramér - Rao attainment):** If $U = h(Y)$ is an unbiased estimator of $g(\theta)$, then $U$ attains the Cramér - Rao lower bound if and only if

$$s(\theta|y) = b(\theta)\left[h(y) - g(\theta)\right].$$

**Example:** Random sample $Y$ from $\mathrm{N}(\mu, 1)$.

$$s(\mu|y) = n(\bar{y} - \mu): \quad b(\mu) = n,\ h(y) = \bar{y},\ g(\mu) = \mu.$$

**Example:** Random sample $Y$ from Poisson$(\lambda)$.

$$s(\lambda|y) = -n + \frac{\sum_{i=1}^{n} y_i}{\lambda} = \frac{n}{\lambda}\left(\bar{y} - \lambda\right),$$

$$b(\lambda) = \frac{n}{\lambda},\ h(y) = \bar{y},\ g(\lambda) = \lambda.$$

**Proof of Cramér - Rao attainment theorem:** The Cramér - Rao lower bound comes from the inequality

$$\mathrm{Cov}[h(Y), s(\theta|Y)]^2 \leq V[h(Y)]V[s(\theta|Y)].$$

The lower bound is attained if and only if the equality holds in the above which is the case if and only if $s(\theta|Y)$ and $h(Y)$ are linearly related:

$$s(\theta|Y) = a(\theta) + b(\theta)h(Y). \tag{8}$$

Taking expectations on both sides we get

$$E[s(\theta|Y)] = a(\theta) + b(\theta)E[h(Y)] \overset{(5),(7)}{\Rightarrow}$$
$$\overset{(5),(7)}{\Rightarrow} 0 = a(\theta) + b(\theta)g(\theta) \Rightarrow a(\theta) = -b(\theta)g(\theta).$$

Substituting into (8), we get

$$s(\theta|Y) = -b(\theta)g(\theta) + b(\theta)h(Y) = b(\theta)[h(Y) - g(\theta)].$$

**Theorem (Uniqueness of MVUE's):** If $U$ is a best (minimum variance) unbiased estimator of $g(\theta)$, then $U$ is unique.

**Proof:** We will use again the Cauchy Schwarz inequality

$$\mathrm{Cov}(X, Y) \leq [V(X)V(Y)]^{1/2}, \tag{9}$$

and the fact that when the equality holds in the above, we can write

$$Y = a(\theta) + b(\theta)X$$

Let $U'$ be another minimum variance unbiased estimator ($V(U) = V(U')$), and consider the estimator $U^* = \frac{1}{2}U + \frac{1}{2}U'$.

Note that $U^*$ is also unbiased

$$E(U^*) = E\left(\frac{1}{2}U + \frac{1}{2}U'\right) = \frac{1}{2}E(U) + \frac{1}{2}E(U') = g(\theta),$$

and

$$\begin{aligned}
V(U^*) &= V\left(\frac{1}{2}U + \frac{1}{2}U'\right) \\
&= V\left(\frac{1}{2}U\right) + V\left(\frac{1}{2}U'\right) + 2\text{Cov}\left(\frac{1}{2}U, \frac{1}{2}U'\right) \\
&= \frac{1}{4}V(U) + \frac{1}{4}V(U') + \frac{1}{2}\text{Cov}(U, U') \\
&\overset{(9)}{\leq} \frac{1}{4}V(U) + \frac{1}{4}V(U') + \frac{1}{2}[V(U)V(U')]^{1/2} \\
&= V(U). \qquad (V(U) = V(U'))
\end{aligned}$$

We must have equality in the previous expression because $U$ is a MVUE. This implies

$$\text{Cov}(U, U') = [V(U)V(U')]^{1/2} = V(U), \quad \text{and} \tag{10}$$

$$U' = a(\theta) + b(\theta)U. \tag{11}$$

We can write

$$\begin{aligned}
V(U) = V(U') &\overset{(10)}{=} \text{Cov}(U, U') \overset{(11)}{=} \text{Cov}[U, a(\theta) + b(\theta)U] \\
&= \text{Cov}[U, b(\theta)U] = b(\theta)V(U).
\end{aligned}$$

Hence $b(\theta) = 1$. Also

$$E(U') = E(U) \overset{(11)}{\Rightarrow} E(U) + a(\theta) = E(U) \rightarrow a(\theta) = 0.$$

Since $a(\theta) = 0$ and $b(\theta) = 1$, $U$ is unique.

## 24.7 Sufficiency and Minimum Variance Unbiased Estimators

We can use the concept of sufficiency for searching for minimum variance unbiased estimators.

**Theorem (Rao-Blackwell):** Let $U(Y)$ be an unbiased estimator of $g(\theta)$ and $T(Y)$ be a sufficient statistic for $\theta$. Define $W(Y) = E(U(Y)|T(Y))$. Then for all $\theta$

1. $E(W) = g(\theta)$,
2. $V(W) \leq V(U)$,
3. $W$ is a uniformly better (unbiased) estimator than $U$.

**Proof:** The proof of the Rao-Blackwell theorem is based on the following conditional expectation properties

$$\begin{aligned} E(X) &= E[E(X|Y)] \quad (12) \\ V(X) &= V[E(X|Y)] + E[V(X|Y)] \quad (13) \end{aligned}$$

We can write

$$g(\theta) = E(U) \overset{(12)}{=} E[E(U|T)] = E[W(Y)],$$

$$\begin{aligned} V(U) &\overset{(13)}{=} V[E(U|T)] + E[V(U|T)] \\ &= V[W(Y)] + E[V(U|T)] \geq V[W(Y)]. \end{aligned}$$

It remains to prove that $W(Y)$ is indeed an estimator, i.e. independent of parameters. If $U$ is a function only of $Y$, then the distribution of $U|T$ is independent of parameters (definition of sufficiency). Hence so is $W(Y)$.

**Note:** conditioning an unbiased estimator on a sufficient statistics results in a uniform improvement, actually conditioning on anything always givens an improvement but the result might depend on $\theta$ hence it will not be an estimator. Sufficiency is crucial then.

**Example:** Let $(Y_1, \ldots, Y_n)$ be a random sample from a distribution with mean $\mu$ and variance $\sigma^2$ and suppose that $T = \sum_{i=1}^{n} Y_i$ is sufficient for $\mu$. Consider the estimator $\hat{\mu}_1 = Y_1$ for $\mu$ and find a better one.

For the estimator $\hat{\mu}_1$, we have

$$E(\hat{\mu}_1) = E(Y_1) = \mu, \qquad V(\hat{\mu}_1) = V(Y_1) = \sigma^2$$

Since $Y_1, Y_2, \ldots, Y_n$ are identically distributed

$$E(Y_i|T) = E(Y_1|T) \tag{14}$$

The Rao-Blackwell theorem states that the following estimator is better

$$\begin{aligned}\hat{\mu}_2 = E(\hat{\mu}_1|T) = E(Y_1|T) = \frac{1}{n}\sum_{i=1}^{n} E(Y_1|T) &\overset{(14)}{=} \frac{1}{n}\sum_{i=1}^{n} E(Y_i|T) \\ = \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i|T\right) &= \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y},\end{aligned}$$

Indeed

$$E(\hat{\mu}_2) = E(\bar{Y}) = \mu, \quad V(\hat{\mu}_2) = V(\bar{Y}) = \frac{\sigma^2}{n} \leq V(\hat{\mu}_1)$$

# Reading

G. Casella & R. L. Berger 6.3.1, 7.1, 7.2.1, 7.2.2, 7.3.1, 7.3.2, 7.3.3

# 25 Interval Estimation

## 25.1 Interval Estimators and Confidence Sets

- Point estimates provide a single value as a best guess for the parameter(s) of interest.
- Interval estimates provide an interval which we believe contains the true value of the parameter(s).
- More generally we may look for a confidence sets (not necessarily an interval), for example when when we are unsure whether the result of the procedure is an interval and or in cases of more than one parameters.

**Definition of interval estimator/estimate:** Let $Y = (Y_1 \dots Y_n)$ be a sample with density $f_Y(y|\theta)$ and $U_1(Y)$, $U_2(Y)$ be statistics such that $U_1(x) \leq U_2(x)$ for any $x$. The random interval $[U_1(Y), U_2(Y)]$ is an **interval estimator** for $\theta$.

If the observed sample is $y$, then interval $[U_1(y), U_2(y)]$ is an **interval estimate** for $\theta$.

**Definition of coverage probability:** The probability that the random interval contains the true parameter $\theta$ is termed as **coverage probability** and denoted with

$$P\left[U_1(Y) \leq \theta \leq U_2(Y)\right]$$

**Definition of confidence level:** The infimum of all the coverage probabilities (for each $\theta$) is termed as **confidence level (coefficient)** of the interval.

$$\inf_{\theta} P\left[U_1(Y) \leq \theta \leq U_2(Y)\right]$$

**Notes:**

- The random variables in the coverage probability are $U_1(Y)$ and $U_2(Y)$. The interval may be interpreted as the probability that $U_1(Y)$ and $U_2(Y)$ contain $\theta$.
- If an interval has confidence level $1-\alpha$ the interpretation is: 'If the experiment was repeated many times $100 \times (1-\alpha)\%$ percent of the corresponding intervals would contain the true parameter $\theta$.'
- The random variables in $P\left[U_1(Y) \leq \theta \leq U_2(Y)\right]$ are $U_1(Y)$ and $U_2(Y)$. Thus

$$\begin{aligned} P\left[U_1(Y) \leq \theta \leq U_2(Y)\right] &= P\left[U_1(Y) \leq \theta \cap U_2(Y) \geq \theta\right] \\ &= 1 - P[U_1(Y) > \theta \cup U_2(Y) < \theta] \\ &= 1 - \left\{P[U_1(Y) > \theta] + P[U_2(Y) < \theta] - P[U_1(Y) > \theta \cap U_2(Y) < \theta]\right\} \\ &= 1 - P[U_1(Y) > \theta] - P[U_2(Y) < \theta] \end{aligned}$$

since $P[U_1(Y) > \theta \cap U_2(Y) < \theta] = 0$.

**Example:** given an random sample $X = (X_1 \ldots X_4)$ from $N(\mu, 1)$, compare the sample mean $\bar{X}$ which is a point estimator with the interval estimator $[\bar{X}-1, \bar{X}+1]$. At first sight with the interval estimator we just loose precision, but we actually gained in confidence. Indeed, while $P(\bar{X} = \mu) = 0$, we have

$$P(\bar{X} - 1 \leq \mu \leq \bar{X} + 1) = P(-1 \leq \bar{X} - \mu \leq 1) = P\left(-2 \leq \frac{\bar{X} - \mu}{\sqrt{1/4}} \leq 2\right) = .9544$$

because $\bar{X} \sim N(\mu, 1/4)$. Therefore, we loose in precision but we now have over 95% chances of covering the unknown parameter with this interval estimator.

**Definition of expected length:** Consider an interval estimator $[U_1,\ U_2]$. The length $U_2 - U_1$ is a random variable. One possible measure is its expected length

$$E(U_2 - U_1)$$

A good interval estimator should minimise the expected length while maximising the confidence level.

**Some notation:** Suppose that the random variable X follows a distribution $\mathcal{X}$. We will denote with $\mathcal{X}_\alpha$ the number for which

$$P(X \leq \mathcal{X}_\alpha) = \alpha$$

Naturally

$$P(X > \mathcal{X}_\alpha) = 1 - \alpha$$

We use such notation for various distributions. In particular we use the letter $Z$ and $\mathcal{Z}_\alpha$ for the standard normal distribution where we also write

$$P(Z \leq \mathcal{Z}_\alpha) = \Phi(\mathcal{Z}_\alpha) = \alpha.$$

**Same length different confidence levels:**

Suppose that we have a random sample from a $N(\mu, 1)$ and we want an interval estimator for $\mu$. Let $k_1$, $k_2$ be positive constants. All the intervals below have the same length.

1. $[-k_1,\ k_2]$
2. $[Y_1 - k_1,\ Y_1 + k_2]$
3. $[\bar{Y} - k_1, \bar{Y} + k_2]$

Let's evaluate their confidence levels. Note that $Y_i - \mu \sim N(0, 1)$

1. This interval does not depend on the sample. If $\mu \in [k_1,\ k_2]$, the coverage probability is 1, otherwise it is 0. Thus the confidence level is 0.
2. The coverage probability is

$$\begin{aligned} &P(Y_1 - k_1 \leq \mu \leq Y_1 + k_2) = 1 - P(Y_1 - k_1 > \mu) \\ &- P(Y_1 + k_2 < \mu) = 1 - P(Y_1 - \mu > k_1) - P(Y_1 - \mu < -k_2) \\ &= \Phi(k_1) - \Phi(-k_2) = \Phi(k_1) + \Phi(k_2) - 1 \end{aligned}$$

   which is equal with the confidence level.

3. Using the fact that $\sqrt{n}(\bar{Y} - \mu) \sim N(0, 1)$ and similar calculations we get a confidence level of

$$\Phi(\sqrt{n}k_1) + \Phi(\sqrt{n}k_2) - 1 \geq \Phi(k_1) + \Phi(k_2) - 1$$

**Same confidence level different lengths:**

Suppose that we have a random sample from a $N(\mu, 1)$ and we want an interval estimator for $\mu$. We know that $Z = \sqrt{n}(\bar{Y} - \mu) \sim N(0, 1)$. If $\alpha_1$, $\alpha_2$ are positive numbers such that $\alpha = \alpha_1 + \alpha_2$, we can write

$$\begin{aligned} P(\mathcal{Z}_{\alpha_1} \leq \sqrt{n}(\bar{Y} - \mu) \leq \mathcal{Z}_{1-\alpha_2}) &= 1 - P(Z < \mathcal{Z}_{\alpha_1}) - P(Z > \mathcal{Z}_{1-\alpha_2}) \\ &= 1 - \alpha_1 - [1 - (1 - \alpha_2)] \\ &= 1 - (\alpha_1 + \alpha_2) \\ &= 1 - \alpha, \end{aligned}$$

By rearrangement we get the interval estimator for $\mu$

$$\left[\bar{Y} - \frac{1}{\sqrt{n}}\mathcal{Z}_{1-\alpha_2},\ \ \bar{Y} + \frac{1}{\sqrt{n}}\mathcal{Z}_{1-\alpha_1}\right]$$

The (expected) length of the interval above is

$$E\left[\bar{Y} + \frac{1}{\sqrt{n}}\mathcal{Z}_{1-\alpha_1} - \bar{Y} - \frac{1}{\sqrt{n}}\mathcal{Z}_{1-\alpha_2}\right] = \frac{1}{\sqrt{n}}(\mathcal{Z}_{1-\alpha_1} + \mathcal{Z}_{1-\alpha_2})$$

Using statistical tables we can construct the following table, where we fix $\alpha_1 + \alpha_1 = 0.05$. Hence we have the length of $95\%$ confidence intervals for the mean of a normal distribution with unit variance for various lower and upper endpoints.

| $\alpha_1$ | $\alpha_2$ | $\mathcal{Z}_{1-\alpha_1}$ | $\mathcal{Z}_{1-\alpha_2}$ | $\sqrt{n}$ length |
|---|---|---|---|---|
| 0 | 0.05 | $+\infty$ | 1.645 | $\infty$ |
| 0.01 | 0.04 | 2.326 | 1.751 | 4.077 |
| 0.02 | 0.03 | 2.054 | 1.881 | 3.935 |
| **0.025** | **0.025** | **1.96** | **1.96** | **3.920** |
| 0.03 | 0.02 | 1.881 | 2.054 | 3.935 |
| 0.04 | 0.01 | 1.751 | 2.326 | 4.077 |
| 0.05 | 0 | 1.645 | $+\infty$ | $\infty$ |

**Why 95%:**

Let us consider symmetric intervals with confidence levels 0.8, 0.9, 0.95, and 0.99. Using the previous procedure and statistical tables we can construct the following table where we have the length of intervals for the mean of a normal distribution with unit variance for various confidence levels.

| $\alpha_1$ | $\alpha_2$ | $\mathcal{Z}_{1-\alpha_1}$ | $\mathcal{Z}_{1-\alpha_2}$ | $\sqrt{n}$ length |
|---|---|---|---|---|
| 0.1 | 0.1 | 1.2816 | 1.2816 | 2.563 |
| 0.05 | 0.05 | 1.645 | 1.645 | 3.290 |
| **0.025** | **0.025** | **1.96** | **1.96** | **3.920** |
| 0.005 | 0.005 | 2.576 | 2.576 | 5.152 |

The level 95% is chosen as a compromise between length and confidence.

## 25.2 Finding Interval Estimators from Pivotal Functions

A way to construct a $1-\alpha$ confidence set for $\theta$ is by using a pivotal function.

**Definition of a pivotal function:** Consider a sample $Y$ with density $f_Y(y|\theta)$ and suppose that we are interested in constructing an interval estimator for $\theta$. A function $G = G(Y, \theta)$ of $Y$ and $\theta$ is a **pivotal function** for $\theta$ if its distribution is known and does not depend on $\theta$.

**Example:** Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a $N(\mu, \sigma^2)$ with $\mu$ unknown and $\sigma^2$ known. We know that

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

and we can use the above to get the following pivotal function

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Notice that $Z$ depends on $\mu$ but its distribution does not change regardless of the value of $\mu$.

**Example:** Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a $N(\mu, \sigma^2)$ with $\mu$ known and $\sigma^2$ unknown. We know that

$$Z_i = \frac{Y_i - \mu}{\sigma} \sim N\left(0, 1\right),$$

and that $Z_i$'s are independent. Getting $\sum_i Z_i^2$ gives us the following pivotal function

$$\sum_{i=1}^{n} Z_i^2 = \frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{\sigma^2} \sim \chi^2_n.$$

**Example:** Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a $N(\mu, \sigma^2)$ with both $\mu$, $\sigma^2$ unknown. Now we cannot use

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

as a pivotal function for $\mu$ because its distribution also depends on the unknown parameter $\sigma$. Instead we use

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

In the same way we cannot use $\sum_i Z_i^2$ for $\sigma^2$ since its distribution depends on $\mu$ which is unknown, instead we can use

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

**Constructing an interval from a pivotal function:**

Suppose that we have a sample $Y$. To construct an interval estimator with confidence level $1-\alpha$ for the parameter $\theta$ using a pivotal function one can use the following procedure:

**Step 1:** Find a pivotal function $G = G(Y, \theta)$ based on a reasonable point estimator for $\theta$.

**Step 2:** Use the distribution of the pivotal function to find values $g_1$ and $g_2$ such that

$$P(g_1 \leq G(Y, \theta) \leq g_2) = 1 - \alpha$$

**Step 3:** Manipulate the quantities $G \geq g_1$ and $G \leq g_2$ to make $\theta$ the reference point. This yields inequalities of the form

$$\theta \geq U_1(Y, g_1, g_2) \quad \text{and} \quad \theta \leq U_2(Y, g_1, g_2),$$

for some functions $U_1(.)$ and $U_2(.)$ independent of parameters.

**Step 4:** Give the following interval

$$[U_1(Y, g_1, g_2), \quad U_2(Y, g_1, g_2)].$$

**Note:** The endpoints $U_1$, $U_2$ are usually functions of one of the $g_1$ or $g_2$ but not the other.

**Example: Interval for $\mu$, $N(\mu, \sigma^2)$ with known $\sigma^2$**

Suppose that we have a random sample $Y$ from a $N(\mu, \sigma^2)$ (with $\sigma^2$ known) and we want an interval estimator for $\mu$ with confidence level $1-\alpha$.

**Step 1:** We know that

$$Z = Z(Y, \mu) = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Thus $Z$ is a pivotal function.

**Step 2:** We can write

$$\begin{aligned} &P(\mathcal{Z}_{\alpha/2} \leq Z \leq \mathcal{Z}_{1-\alpha/2}) = \\ &= 1 - P(Z < \mathcal{Z}_{\alpha/2}) - P(Z > \mathcal{Z}_{1-\alpha/2}) \\ &= 1 - \alpha/2 - [1 - (1 - \alpha/2)] \\ &= 1 - (\alpha/2 + \alpha/2) = 1 - \alpha. \end{aligned}$$

**Step 3:** Rearranging the inequalities we get

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \geq \mathcal{Z}_{\alpha/2} \quad \text{and} \quad \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq \mathcal{Z}_{1-\alpha/2}$$

which we can rewrite as

$$\mu \leq \bar{Y} - \frac{\sigma}{\sqrt{n}}\mathcal{Z}_{\alpha/2}, \text{ and } \mu \geq \bar{Y} - \frac{\sigma}{\sqrt{n}}\mathcal{Z}_{1-\alpha/2}$$

**Step 4:** Note that $\mathcal{Z}_{\alpha/2} = -\mathcal{Z}_{1-\alpha/2}$. We get

$$[\bar{Y} - \frac{\sigma}{\sqrt{n}}\mathcal{Z}_{1-\alpha/2}, \quad \bar{Y} + \frac{\sigma}{\sqrt{n}}\mathcal{Z}_{1-\alpha/2}]$$

**Numerical Example:** Suppose that we had $n = 10$, $\bar{Y} = 5.2$, $\sigma^2 = 2.4$ and $\alpha = 0.05$. From suitable tables or statistical software we get $\mathcal{Z}_{.975} = 1.96$, so an interval estimator for $\mu$ with confidence level $1 - \alpha$ is

$$[5.2 - 1.96\sqrt{2.4/10}, \quad 5.2 + 1.96\sqrt{2.4/10}]$$

or else $[4.24, \quad 6.16]$.

**Example: Interval for $\mu$, $N(\mu, \sigma^2)$, with unknown $\sigma^2$**

Suppose that we have a random sample $Y$ from a $N(\mu, \sigma^2)$ (with also $\sigma^2$ unknown) and we want an interval estimator for $\mu$ with confidence level $1 - \alpha$.

**Step 1:** We know that

$$T = T(Y, \mu) = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Thus $T$ is a pivotal function.

**Step 2:** We can write

$$\begin{aligned}
&P(t_{n-1,\alpha/2} \leq T \leq t_{n-1,1-\alpha/2}) = \\
&= 1 - P(T < t_{n-1,\alpha/2}) - P(T > T_{n-1,1-\alpha/2}) \\
&= 1 - \alpha/2 - [1 - (1 - \alpha/2)] \\
&= 1 - (\alpha/2 + \alpha/2) = 1 - \alpha.
\end{aligned}$$

**Step 3:** Rearranging the inequalities we get

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \geq t_{n-1,\alpha/2} \quad \text{and} \quad \frac{\bar{Y} - \mu}{S/\sqrt{n}} \leq t_{n-1,1-\alpha/2}$$

which we can rewrite as

$$\mu \leq \bar{Y} - \frac{S}{\sqrt{n}}t_{n-1,\alpha/2}, \text{ and } \mu \geq \bar{Y} - \frac{S}{\sqrt{n}}t_{n-1,1-\alpha/2}$$

**Step 4:** Note that $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$. We get

$$\left[\bar{Y} - \frac{S}{\sqrt{n}}t_{n-1,1-\alpha/2}, \quad \bar{Y} + \frac{S}{\sqrt{n}}t_{n-1,1-\alpha/2}\right]$$

**Numerical Example:** Suppose that we had $n = 10$, $\bar{Y} = 5.2$, $S^2 = 2.4$ and $\alpha = 0.05$. From suitable tables or statistical software we get $t_{9,.975} = 2.262$, so an interval estimator for $\mu$ with confidence level $1 - \alpha$ is

$$[5.2 - 2.262\sqrt{2.4/10}, \quad 5.2 + 2.262\sqrt{2.4/10}]$$

or else $[4.09, \quad 6.31]$.

**Note:** Compared with the known $\sigma^2$ case the interval is now larger despite the fact that $S = \sigma$. The $t$ distribution has fatter tails than the standard Normal. On the other hand as $n$ grows the $t$ distribution gets closer to the Normal.

# Reading

G. Casella & R. L. Berger 9.1, 9.2.1, 9.2.2, 9.3.1

# 26 Asymptotic Evaluations

So far we considered evaluation criteria based on samples of finite size $n$. But as mentioned above there may be cases where a satisfactory solution does not exist. An alternative route is to approach this problems with letting $n \to \infty$, in other words study the **asymptotic behaviour** of the problem. We will look mainly into asymptotic properties of maximum likelihood procedures.

## 26.1 Summary of the Point/Interval Estimation Issues

- In point estimation we use the information from the sample $Y$ to provide a best guess for the parameters $\theta$.
- For this we use statistics termed as **estimators**,

  $$\hat{\theta} = h(Y),$$

  that are functions of the sample $Y$. The realization of the sample provides a **point estimate** which reflects our belief for the parameter $\theta$.
- There are many ways to find estimator functions. For example one can use the **method of moments** or **maximum likelihood estimators**.
- We look for estimators with **small mean squared error**, defined as

  $$E[(\hat{\theta} - \theta)^2].$$

- But it is very hard to compare estimators based solely on MSE. Even irrational estimators like
$$\hat{\theta} = 1,$$
are not worse than reasonable ones for all $\theta$. For this reason we restrict attention to **unbiased estimators**
$$E(\hat{\theta}) = \theta.$$

- An optimal solution to the problem is given by a **minimum variance unbiased estimators**. Note that the variance of an unbiased estimator is equal to its MSE. If such an estimator exists it is **unique**.

- The Cramér-Rao theorem provides a lower bound for the variance of an unbiased estimator. Therefore if the variance of an unbiased estimator attains that bound, it provides an optimal solution to the problem.

- Alternatively if an unbiased estimator is based on a **complete sufficient statistic** it is also of minimum variance (see Rao-Blackwell theorem).

- **Problem:** Even an unbiased estimator may not be available or may not exist.

- In interval estimation we want to use the information from the sample $Y$ to provide an **interval** which we believe **contains the true value** of the parameter(s).

- The probability that the random interval contains the true parameter $\theta$ is termed as **coverage probability**.

- The infimum of all the coverage probabilities is termed as **confidence coefficient (level)** of the interval.

- A way to construct an interval is by using a **pivotal function**, that is a function of $Y$ and $\theta$ with distribution **independent** of $\theta$.

- Alternatively one may invert an $\alpha$ level test of $H_0 : \theta = \theta_0$. The **parameter points** $\theta_0$ that provide an acceptance region $A$ that **contains the observed sample**, provide a $1 - \alpha$ confidence level interval. Conversely we can find the acceptance region of an $\alpha$ level $H_0 : \theta = \theta_0$ test by taking the **sample points** $Y$ for which the resulting $1 - \alpha$ confidence level interval **contains** $\theta_0$ (see definitions in Chapter 6).

- There may exist more than one intervals with the same level. One way to choose between them is through their **expected length**.

- **Problem:** Sometimes it may be even hard to find any 'reasonable' interval estimator.

## 26.2 Asymptotic Evaluations

**Definition:** A sequence of estimators $U_n = U(X_1, \ldots, X_n)$ is a **consistent sequence of estimators** for a parameter $\theta$ if, for every $\epsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta(|U_n - \theta| < \epsilon) = 1.$$

In other words a consistent estimator converges in probability to the parameter $\theta$ it is estimating. Notice that for any $\theta$ the property must hold, that is if we change $\theta$ the probability $P_\theta$ changes but the limit will still hold.

**Theorem:** If $U_n$ is a sequence of estimators for a parameter $\theta$ satisfying

1. $\lim_{n\to\infty} V(U_n) = 0$,
2. $\lim_{n\to\infty} \text{Bias}(U_n) = 0$,

for every $\theta \in \Theta$, then $U_n$ is a consistent sequence of estimators.

**Proof:** we use Chebychev inequality, as $n \to \infty$,

$$P_\theta(|U_n - \theta| > \epsilon) \leq \frac{E[(U_n - \theta)^2]}{\epsilon^2} = \frac{\text{Bias}(U_n)^2 + V(U_n)}{\epsilon^2} \to 0.$$

An example is provided by the sample mean which has zero bias and variance $V(X_i)/n$.

**Definition:** An estimator is **asymptotically unbiased** for $\theta$ if its bias goes to 0 as $n \to \infty$ for any $\theta \in \Theta$.

**Definition:** The ratio of the Cramér-Rao lower bound over the variance of an estimator is termed as **efficiency**. An **efficient** estimator has efficiency 1. We can compare estimators in terms of their **asymptotic efficiency**, that is their efficiencies as $n \to \infty$. An estimator is **asymptotically efficient** if its asymptotic efficiency is 1.

**Theorem (Asymptotic normality of MLEs):** Under weak regularity conditions the maximum likelihood estimator $g(\hat{\theta})$ satisfies

$$\sqrt{n}\left[g(\hat{\theta}) - g(\theta)\right] \xrightarrow{d} N\left(0, \frac{g'(\theta)^2}{\mathcal{I}(\theta|y_i)}\right), \quad n \to \infty,$$

where $g$ is a continuous function. We may also write

$$g(\hat{\theta}) \overset{\text{approx}}{\sim} N(g(\theta), v(\theta)),$$

where $v(\theta) = \frac{g'(\theta)^2}{n\mathcal{I}(\theta|y_i)}$ which is the Cramér-Rao lower bound, since $\mathcal{I}(\theta|y) = n\mathcal{I}(\theta|y_i)$. Therefore, $\text{Var}(g(\hat{\theta})) = v(\theta)$.

The estimator $\hat{\theta}$ is computed using a sample of size $n$ hence it depends on $n$ and the theorem tells us its behaviour when we consider larger and larger samples.

**Corollary:** Under weak regularity conditions the MLE $\hat{\theta}$, or a function of it, is consistent, asymptotically unbiased and efficient for the parameter it is estimating.

**Slutsky's theorem** If, as $n \to \infty$, $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} a$ with $a$ constant then

a. $Y_n X_n \xrightarrow{d} aX$;

b. $X_n + Y_n \xrightarrow{d} X + a$.

Actually asymptotic normality implies always consistency. Indeed, we have

$$\sqrt{n\mathcal{I}(\theta|y_i)}(\hat{\theta} - \theta) \xrightarrow{d} Z \sim N(0,1), \quad n \to \infty,$$

then

$$(\hat{\theta} - \theta) = \left(\frac{1}{\sqrt{n\mathcal{I}(\theta|y_i)}}\right)\left(\sqrt{n\mathcal{I}(\theta|y_i)}(\hat{\theta} - \theta)\right) \xrightarrow{d} \lim_{n\to\infty}\left(\frac{1}{\sqrt{n\mathcal{I}(\theta|y_i)}}\right) Z = 0, \quad n \to \infty,$$

and convergence in distribution to a point is equivalent to convergence in probability. So $\hat{\theta}$ is consistent estimator of $\theta$.

**Asymptotic distribution of MLE's - Sketch of proof:** Assume $g(\theta) = \theta$ and let $s'(\theta|Y)$ denote $\frac{\partial}{\partial\theta}s(\theta|Y)$. Let $\hat{\theta}$ be the MLE of the true value which we denote as $\theta_0$.

Consider a Taylor series expansion around the true value $\theta_0$

$$s(\theta|Y) = s(\theta_0|Y) + s'(\theta_0|Y)(\theta - \theta_0) + \dots$$

Ignore the higher order terms and substitute $\theta$ with $\hat{\theta}$

$$\begin{aligned} s(\hat{\theta}|Y) &= s(\theta_0|Y) + s'(\theta_0|Y)(\hat{\theta} - \theta_0) \Rightarrow \\ \Rightarrow \hat{\theta} - \theta_0 &= -\frac{s(\theta_0|Y)}{s'(\theta_0|Y)} \ \big(\text{since } s(\hat{\theta}|Y) = 0\big) \Rightarrow \\ \Rightarrow \sqrt{n}(\hat{\theta} - \theta_0) &= \frac{\frac{\sqrt{n}}{n}s(\theta_0|Y)}{-\frac{1}{n}s'(\theta_0|Y)}. \end{aligned} \tag{15}$$

Then, recall that

$$s(\theta|Y) = \sum_{i=1}^{n} s(\theta|Y_i). \tag{16}$$

Using (16), the numerator of (15) becomes

$$\frac{\sqrt{n}}{n}s(\theta_0|Y) = \sqrt{n}\left(\frac{\sum_{i=1}^{n} s(\theta_0|Y_i)}{n} - 0\right) \xrightarrow{d} N\big(0, \mathcal{I}(\theta_0|y_i)\big), \qquad n \to \infty,$$

from the Central Limit Theorem for i.i.d. random variables and since $E[s(\theta_0|Y_i)] = 0$, $\text{Var}[s(\theta_0|Y_i)] = \mathcal{I}(\theta_0|Y_i)$.

For the denominator of (15), using the Weak Law of Large Numbers for i.i.d. random variables, we get

$$-\frac{1}{n}s'(\theta_0|Y) = -\frac{1}{n}\sum_{i=1}^{n} s'(\theta_0|Y_i) \xrightarrow{p} E\left[-s'(\theta_0|Y_i)\right] = \mathcal{I}(\theta_0|y_i), \qquad n \to \infty.$$

Combining these two results and using Slutsky's theorem we get that, as $n \to \infty$,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{\frac{\sqrt{n}}{n} s(\theta_0|Y)}{-\frac{1}{n} s'(\theta_0|Y)} \xrightarrow{d} N\left(0, \frac{1}{\mathcal{I}(\theta_0|y_i)}\right) \quad \Rightarrow \quad \hat{\theta} - \theta_0 \xrightarrow{d} N\left(0, \frac{1}{\mathcal{I}(\theta_0|y)}\right),$$

since $\mathcal{I}(\theta_0|y) = n\mathcal{I}(\theta_0|y_i)$.

**Asymptotic pivotal function from MLEs:** Note that

$$\sqrt{\mathcal{I}(\theta|y)}(\hat{\theta} - \theta) \overset{\text{approx}}{\sim} N(0, 1).$$

Also, since $\hat{\theta}$ is consistent for $\theta$, the quantity $\mathcal{I}(\hat{\theta}|y)$ converges in probability to $\mathcal{I}(\theta|y)$. Hence, in a second level of approximation, we can write for large sample sizes $n$

$$\sqrt{\mathcal{I}(\hat{\theta}|y)}(\hat{\theta} - \theta) \overset{\text{approx}}{\sim} N(0, 1).$$

We say that this function is asymptotically pivotal for $\theta$.

**Example: Asymptotic estimation of Bernoulli** Let $(Y_1, \ldots, Y_n)$ be a random sample from a Bernoulli($p$). We know that

- $T = \sum_{i=1}^{n} Y_i \sim$ Binomial$(n, p)$.
- The MLE for $p$ is $\hat{p} = \bar{Y}$.
- The Fisher's information is $\mathcal{I}(p) = \frac{n}{p(1-p)}$.

The MLE for $p$, $\hat{p} = \bar{Y}$ is consistent, (asymptotically) unbiased and efficient. The asymptotic distribution of $\hat{p} = \bar{Y}$ is

$$\hat{p} \overset{\text{approx}}{\sim} N\left(p, \frac{p(1-p)}{n}\right).$$

In an extra level of approximation we may use

$$\hat{p} \overset{\text{approx}}{\sim} N\left(p, \frac{\hat{p}(1-\hat{p})}{n}\right).$$

Let $S_p = \frac{\hat{p}(1-\hat{p})}{n}$. Then

$$\hat{p} \overset{\text{approx}}{\sim} N\left(p, S_p\right) \quad \Rightarrow \quad \frac{\hat{p} - p}{\sqrt{S_p}} \overset{\text{approx}}{\sim} N\left(0, 1\right).$$

We can use the above to construct the following asymptotic $1 - \alpha$ confidence interval

$$\left[\hat{p} - \mathcal{Z}_{1-\alpha/2}\sqrt{S_p}, \quad \hat{p} + \mathcal{Z}_{1-\alpha/2}\sqrt{S_p}\right]$$

**Note:** The above interval may take values outside [0,1].

**Example: Asymptotic estimators, intervals for functions of parameters** Let $Y = (Y_1, \ldots, Y_n)$ be a random sample from a $N(0, \sigma^2)$. We want a point and interval estimator for $\sigma$. The MLE for $\sigma^2$ is $\hat{\sigma}^2 = \frac{n-1}{n} S^2$, where $S^2$ is the sample variance. Hence, a consistent, asymptotically unbiased and efficient estimator is the MLE for $\sigma$, that is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = S\sqrt{\frac{n-1}{n}}$$

Note that both $\hat{\sigma}^2$ and $\hat{\sigma}$ are biased for small samples. But their bias goes to 0 as $n \to \infty$.

We are interested in $\sigma = g(\sigma^2) = (\sigma^2)^{1/2}$. The Cramér - Rao lower bound is equal to

$$v(\sigma) = \frac{\left(\frac{\partial}{\partial \sigma^2} g(\sigma^2)\right)^2}{n\mathcal{I}(\sigma^2|y_i)} = \frac{\frac{1}{4\sigma^2}}{\frac{n}{2\sigma^4}} = \frac{\sigma^2}{2n}$$

If we further substitute $v(\hat{\sigma}) = \hat{\sigma}^2/2n$ for $v(\sigma)$ we get

$$\hat{\sigma} \overset{\text{approx}}{\sim} N\left(\sigma, v(\hat{\sigma})\right) \quad \Rightarrow \quad \frac{\hat{\sigma} - \sigma}{\sqrt{v(\hat{\sigma})}} \overset{\text{approx}}{\sim} N(0,1)$$

which leads to the following asymptotic $1 - \alpha$ confidence level interval for $\sigma$

$$\left[\hat{\sigma} - \mathcal{Z}_{1-\alpha/2}\sqrt{v(\hat{\sigma})}, \quad \hat{\sigma} + \mathcal{Z}_{1-\alpha/2}\sqrt{v(\hat{\sigma})}\right].$$

# Reading

G. Casella & R. L. Berger 10.1.1, 10.1.2, 10.3.1, 10.3.2, 10.4.1

# 27 Hypothesis Testing

**Problem:**

- Suppose that a real world phenomenon/population may be described by a probability model defined through the random variable $Y$ with $F_Y(y|\theta)$.
- Suppose also that a sample $Y = (Y_1, Y_2, \ldots, Y_n)$ is drawn from that distribution/population.
- We want to use the information in the random sample $Y$ to answer statements about the population parameters $\theta$.

## 27.1 Statistical tests

### 27.1.1 Definitions

- A **hypothesis** is a statement about a population parameter.
- The two complementary hypotheses in a hypothesis testing problem are often called the **null** and **alternative**. They are denoted by $H_0$ and $H_1$ respectively.
- A **simple** hypothesis takes the form

$$H_0 : \theta = c,$$

  where $c$ is a constant. A hypothesis that is not simple is called **composite**, e.g.

$$H_0 : \theta \leq c,$$

- A **hypothesis test** is a rule that specifies
  1. For which sample values the decision is made to accept $H_0$ as true.
  2. For which sample values $H_0$ is rejected and $H_1$ is accepted as true.
- The subset of the sample space for which $H_0$ will be rejected is termed as **rejection region** or **critical region**. The complement of the rejection region is termed as a **acceptance region**.
- The test rule is based on a statistic, termed as the **test statistic**.

#### Test Statistics

- The crucial part is to identify an appropriate test statistic $T$.
- One of the desired features of $T$ is to have an interpretation such that large (or small) values of it provide evidence against $H_0$.
- We also want to know the distribution of $T$ under $H_0$, that is when $H_0$ holds.
- We focus on test statistics that have tabulated distributions under $H_0$ (Normal, $t$, $\chi^2$, $F$). But this is only for convenience and it is not a strict requirement.

**Example:** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be random sample $Y$ of size $n$ from a $N(\mu, \sigma^2)$ population (with $\sigma^2$ is known). Is $\mu$ equal to $\mu_0$ or larger for a given value $\mu_0$?

The hypotheses of the test are

$$H_0 : \ \mu = \mu_0 \quad \text{versus} \quad H_1 : \ \mu > \mu_0$$

One test may be based on the test statistic $\bar{Y}$ with rejection region

$$R = \left[\mu_0 + 1.96\frac{\sigma}{\sqrt{n}},\ \ \infty\right]$$

The rule is then to reject $H_0$ if $\bar{Y} > \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}$.

### 27.1.2 Types of errors in tests, power function and p-value

Consider a test with $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$.

**Type I error:** If $\theta \in \Theta_0$ ($H_0$ is true) but the test rejects $H_0$.

**Type II error:** If $\theta \in \Theta_1$ ($H_1$ is true) but the test does not reject $H_0$.

| | Accept $H_0$ | Reject $H_0$ |
|---|---|---|
| $H_0$ is true | Correct Decision | Type I error |
| $H_1$ is true | Type II error | Correct Decision |

The Type I error is associated with the **significance level** and the **size** of the test.

**Definition:** The test has significance level $\alpha$ if

$$\sup_{\theta\in\Theta_0} P_\theta(\text{Reject } H_0) \leq \alpha$$

The test has size $\alpha$ if

$$\sup_{\theta\in\Theta_0} P_\theta(\text{Reject } H_0) = \alpha$$

**Note:** If the null hypothesis is simple then the size of the test is the probability of a type I error.

**Power function:** Let $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_0^c$. The **power** function is defined as

$$\beta(\theta) = P_\theta(\text{Reject } H_0),$$

that is the probability that the null hypothesis is rejected if the true parameter value is $\theta$.

**Note:**

$$\beta(\theta) = P_\theta(\text{Reject } H_0) = \begin{cases} \text{probability of Type I error} & \text{if } \theta \in \Theta_0, \\ 1 - \text{probability of Type II error}, & \text{if } \theta \in \Theta_0^c. \end{cases}$$

Also we can define the level and the size of the test through the power function

$$\text{Level: } \sup_{\theta\in\Theta_0} \beta(\theta) \leq \alpha, \text{ and Size: } \sup_{\theta\in\Theta_0} \beta(\theta) = \alpha.$$

Ideally we would like the power function $\beta(\theta)$ to be 0 when $\theta \in \Theta_0$ and 1 when $\theta \in \Theta_1$, but this is not possible. In practice we fix the size $\alpha$ to a small value (usually 0.05) and for a given size we try to maximize the power. Hence

- Failure to reject the null hypothesis does not imply that it holds and we say that we do not reject $H_0$ rather than saying we accept $H_0$.
- We usually set the alternative hypothesis to contain the statement that we are interested in proving.

**Example of a power function (previous example continued):** The power function of the test is

$$\begin{aligned}
\beta(\mu) &= P(Y \in R) = P\left(\bar{Y} > \mu_0 + \mathcal{Z}_{1-\alpha}\sigma/\sqrt{n}\right) \\
&= P\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} > \mathcal{Z}_{1-\alpha}\right) \\
&= P\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} > \mathcal{Z}_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(\mathcal{Z}_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

**$p$-value:** From the definitions so far, we either reject or not reject the null hypothesis. The following quantity is also informative regarding the weight of evidence against $H_0$.

**Definition:** Let $T(Y)$ be a test statistic such that large values of $T$ give evidence against $H_0$. For an observed sample point $y$ the corresponding $p$-value is

$$p(y) = \sup_{\theta\in\Theta_0} P(T(Y) \geq T(y))$$

**Notes:**

1. Clearly $0 \leq p(y) \leq 1$. The closer to 0 the more likely to reject.
2. In words, a $p$-value is the probability that we got the result of the sample or a more extreme result. Extreme in the sense of evidence against $H_0$.

3. If we have a fixed significance level $\alpha$, then we can describe the rejection region as

$$R = \{y : p(y) \leq \alpha\}$$

We reject $H_0$ if the probability of observing a more extreme result than that of the sample, is small (less than $\alpha$).

4. A similar definition can be made if small values of $T$ give evidence against $H_0$.

**Example of a $p$-value (previous example continued):** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be random sample $Y$ of size $n$ from a $N(\mu, \sigma^2)$ population (with $\sigma^2$ is known). We want to test $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$.

Note $Y$ denotes the sample whereas $y$ the observed sample. Also we use $\bar{Y}$ as $T(Y)$ and large values of $\bar{Y}$ provide evidence against $H_0$,

$$\begin{aligned} p(y) &= \sup_{\mu \leq \mu_0} P(\bar{Y} \geq \bar{y}) = \sup_{\mu \leq \mu_0} P\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \geq \frac{\bar{y} - \mu}{\sigma/\sqrt{n}}\right) \\ &= \sup_{\mu \leq \mu_0} \left\{1 - P\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{\bar{y} - \mu}{\sigma/\sqrt{n}}\right)\right\} \\ &= \sup_{\mu \leq \mu_0} \left\{1 - \Phi\left(\frac{\bar{y} - \mu}{\sigma/\sqrt{n}}\right)\right\} = 1 - \Phi\left(\frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

### 27.1.3 Constructing Statistical Tests

The procedure for constructing a test can be given by the following general directions:

**Step 1:** Find an appropriate test statistic $T$. Figure out whether large or small values of $T$ provide evidence against $H_0$. Also find its distribution under $H_0$.

**Step 2:** Use the definition of the level/size $\alpha$ and write ($R$ is unknown)

$$P_{\theta_0}(T \in R) \leq \alpha, \quad \text{or} \quad P_{\theta_0}(T \in R) = \alpha$$

**Step 3:** Solve the equation to get $R$. The test rule is then 'Reject $H_0$' if the sample $Y$ is in $\{Y : T(Y) \in R\}$.

**Example of a Statistical test (cont'd)** Let's come back to the previous example with $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$.

**Step 1:** Test statistic $\bar{Y}$. Large values are against $H_0$. Under $H_0$, $\bar{Y} \sim N(\mu_0, \sigma^2)$ or we could use

$$\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

**Step 2:** We know that under $H_0$

$$P_{\mu_0}\left(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} > \mathcal{Z}_{1-\alpha}\right) = P\left(\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} > \mathcal{Z}_{1-\alpha}\right) = \alpha$$

**Note:** It can be shown that the above also holds for $H_0 : \mu \leq \mu_0$.

**Step 3:** From

$$P\left(\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} > \mathcal{Z}_{1-\alpha}\right) = \alpha,$$

we can get to

$$P\left(\bar{Y} > \mu_0 + \mathcal{Z}_{1-\alpha}\sigma/\sqrt{n}\right) = \alpha,$$

hence

$$R = \{Y : \bar{Y} > \mu_0 + \mathcal{Z}_{1-\alpha}\sigma/\sqrt{n}\}$$

## 27.2 Most Powerful Tests

The tests we are interested in control by construction the probability of a type I error (it is at most $\alpha$). A good test should also have a small probability of type II error. In other words it should also be a powerful test.

**Definition:** Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test in class $\mathcal{C}$, with power function $\beta(\theta)$, is a **Uniformly Most Powerful** (UMP) class $\mathcal{C}$ test if

$$\beta(\theta) \geq \beta'(\theta)$$

for every $\theta \in \Theta_0^c$ and every $\beta'(\theta)$ that is a power function of a test in class $\mathcal{C}$.

**Theorem (Neyman-Pearson Lemma):** Consider a test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ and let $f_Y(y|\theta_0)$, $f_Y(y|\theta_1)$ denote the pdf (pmf) of the sample $Y$. Suppose that a test with rejection region $R$ satisfies

$$y \in R, \text{ if } \frac{f_Y(y|\theta_1)}{f_Y(y|\theta_0)} > k,$$

for some $k > 0$ and

$$\alpha = P_{\theta_0}(Y \in R).$$

A test that satisfies the above is a uniformly most powerful test of size $\alpha$.

**Notes:**

1. The value $k$ may be chosen to satisfy

$$P_{\theta_0,k}(Y \in R) = P_{\theta_0}\left(\frac{f_Y(y|\theta_1)}{f_Y(y|\theta_0)} > k\right) = \alpha$$

2. The above ratio of pdf's (or pmf's) is the ratio of the likelihood functions.

**Proof of Neyman-Pearson Lemma: Preliminaries:** We give the proof for continuous random variables. For discrete random variables just replace integrals with sums.

Let $\phi_S(Y)$ denote the rule of a test $S$ with rejection region $R_S$. Note that $\phi_S(Y) = I(Y \in R_S)$ where $I(\cdot)$ is the indicator function. Hence for all $\theta$

$$\int_{\mathbb{R}^n} \phi_S(Y) f_Y(y|\theta) dy = \int_{R_S} f_Y(y|\theta) dy \tag{17}$$

$$\begin{aligned} E[\phi_S(Y)] &= \int_{\mathbb{R}^n} \phi_S(y) f_Y(y|\theta) dy = \int_{R_S} f_Y(y|\theta) dy \\ &= P_\theta(\text{Reject } H_0) = \beta_S(\theta) \end{aligned} \tag{18}$$

where $\beta_S(\theta)$ is the power function of $S$.

Let $T$ be the Neyman-Pearson lemma test, that is

$$R_T = \{y \in \mathbb{R}^n : f_Y(y|\theta_1) - k f_Y(y|\theta_0) > 0\},$$

with rule $\phi_T(Y) = I(Y \in R_T)$. Let $S$ be another test with $\phi_S(Y)$. Then

$$\phi_T(y) \geq \phi_S(y), \quad \text{for } \ y \in R_T, \tag{19}$$

because in $R_T$ we have $\phi_T(y) = 1$ while in general $0 \leq \phi_S(y) \leq 1$.

Consider the quantity $B = \phi_S(y)[f_Y(y|\theta_1) - k f_Y(y|\theta_0)]$. If $y \in R_T$, $B \geq 0$. If $y \notin R_T$, $B \leq 0$. Hence

$$\int_{\mathbb{R}^n} \phi_S(y)[f_Y(y|\theta_1) - k f_Y(y|\theta_0)] dy \leq \int_{R_T} \phi_S(y)[f_Y(y|\theta_1) - k f_Y(y|\theta_0)] dy \tag{20}$$

**Main Proof:** Let $T$ be the Neyman-Pearson lemma test and $S$ be another test of size $\alpha$.

$$\begin{aligned} \beta_S(\theta_1) - k\beta_S(\theta_0) &\overset{(18)}{=} \int_{\mathcal{R}^n} \phi_S(y)[f_Y(y|\theta_1) - k f_Y(y|\theta_0)] dy \\ &\overset{(20)}{\leq} \int_{R_T} \phi_S(y)[f_Y(y|\theta_1) - k f_Y(y|\theta_0)] dy \\ &\overset{(19)}{\leq} \int_{R_T} \phi_T(y)[f_Y(y|\theta_1) - k f_Y(y|\theta_0)] dy \\ &\overset{(17)}{=} \int_{\mathcal{R}^n} \phi_T(y)[f_Y(y|\theta_1) - k f_Y(y|\theta_0)] dy \\ &\overset{(18)}{=} \beta_T(\theta_1) - k\beta_T(\theta_0) \end{aligned}$$

Since both $T$ and $S$ are size $\alpha$ tests,

$$\beta_T(\theta_0) = \beta_S(\theta_0) = \alpha.$$

Therefore we can write

$$\beta_S(\theta_1) \leq \beta_T(\theta_1),$$

which implies that $T$ is a uniformly most powerful test of size $\alpha$.

**Example (Neyman-Pearson Lemma):** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be random sample $Y$ of size $n$ from a $N(\mu, \sigma^2)$ population (with $\sigma^2$ is known). We want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$.

**Step 1:** The likelihood ratio from the Neyman-Pearson Lemma is

$$\begin{aligned} LR = \frac{L(\mu_1|Y)}{L(\mu_0|Y)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}[n(\bar{Y}-\mu_1)^2 + (n-1)S^2]\}}{(2\pi\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}[n(\bar{Y}-\mu_0)^2 + (n-1)S^2]\}} \\ &= \exp\left(\frac{n}{2\sigma^2}(-\bar{Y}^2 + 2\mu_1\bar{Y} - \mu_1^2 + \bar{Y}^2 - 2\mu_0\bar{Y} + \mu_1^2)\right) \\ &= \exp\left(\frac{n}{2\sigma^2}[(\mu_0^2 - \mu_1^2) - 2\bar{Y}(\mu_0 - \mu_1)]\right) \end{aligned}$$

If $\mu_0 < \mu_1$ the above is large when $\bar{Y}$ is large. If $\mu_0 > \mu_1$ the above is large when $\bar{Y}$ is small. So $\bar{Y}$ can be a test statistic. Its distribution is known to be

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

**Step 2:** We want

$$P_{\mu_0}\left\{\exp\left(\frac{n}{2\sigma^2}[(\mu_0^2 - \mu_1^2) - 2\bar{Y}(\mu_0 - \mu_1)]\right) > k\right\} = \alpha.$$

**Step 3:** If $\mu_0 < \mu_1$ the above is equivalent with

$$P_{\mu_0}\left\{\bar{Y} > \frac{(\mu_0^2 - \mu_1^2) - \frac{2\sigma^2}{n}\log(k)}{2(\mu_0 - \mu_1)}\right\} = \alpha$$

But we also know that $\bar{Y} \sim N(\mu_0, \frac{\sigma^2}{n})$ under $H_0$, then

$$P_{\mu_0}\left(\bar{Y} > \mu_0 + \mathcal{Z}_{1-\alpha}\sigma/\sqrt{n}\right) = \alpha,$$

is an equivalent test being based on the same statistic. This will give us a most powerful test for this testing problem.

**Example (Neyman-Pearson Lemma):** Let a random sample $Y = (Y_1, \ldots, Y_n)$ from a Poisson($\lambda$) and $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda = \lambda_1$. The Neyman-Pearson lemma likelihood ratio is

$$LR = \frac{e^{-n\lambda_1}\lambda_1^{\sum_i Y_i}/\prod_i Y_i!}{e^{-n\lambda_0}\lambda_0^{\sum_i Y_i}/\prod_i Y_i!} = e^{n(\lambda_0-\lambda_1)}\left(\frac{\lambda_1}{\lambda_0}\right)^{\sum_i Y_i}$$

A test with rejection region from $LR > k$ is such that

$$P\left(\sum_{i=1}^{n} Y_i > \frac{\log k - n(\lambda_0 - \lambda_1)}{\log \lambda_1 - \log \lambda_0} = k_1\right) = \alpha$$

but we also know that $\sum_i Y_i \sim$ Poisson$(n\lambda_0)$ under $H_0$ so we can find $k$.

Let $n = 8$, $\lambda_0 = 2$ ($\sum_i Y_i \sim$ Poisson$(16)$) and $\lambda_1 = 6$. The size is 0.058 with $k_1 = 22$ and 0.037 with $k_1 = 23$. A test with significance level 0.05 corresponds to a $k_1 = 23$. What if $\lambda_1 = 150$?

**Neyman-Pearson lemma for 1-sided composite hypotheses:**

Neyman-Pearson lemma refers to tests with simple hypotheses

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

but sometimes may be used for composite hypotheses as well.

Assume a rejection region independent of $\theta_1$. If $\theta_0 < \theta_1$, the test is most powerful for all $\theta_1 > \theta_0$. Hence it is most powerful for

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

Similarly, if $\theta_0 > \theta_1$ the Neyman-Pearson lemma test is most powerful for

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0.$$

What about

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0, \quad ?$$

The power is not affected, but is the size of the test still $\alpha$? we would have to show that

$$\sup_{\theta \leq \theta_0} P(\text{Reject } H_0) = \alpha.$$

**Example (Neyman-Pearson Lemma):** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from $N(\mu, \sigma^2)$ population (with $\sigma^2$ is known).

We showed that for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, the UMP size $\alpha$ test is constructed by

$$LR = \exp\left(\frac{n}{2\sigma^2}[(\mu_0^2 - \mu_1^2) - 2\bar{Y}(\mu_0 - \mu_1)]\right) > k$$

which for $\mu_0 < \mu_1$ is equivalent to

$$\bar{Y} > \mu_0 + \mathcal{Z}_{1-\alpha}\sigma/\sqrt{n}.$$

Note that the rejection region is independent of $\mu_1$, thus the test is applicable for all $\mu_1 > \mu_0$. Hence it is also the UMP test for

$$H_0 : \mu = \mu_0, \quad \text{versus} \quad H_1 : \mu > \mu_0$$

What about testing problems of the following form?

$$H_0 : \mu \leq \mu_0, \quad \text{versus} \quad H_1 : \mu > \mu_0.$$

The UMP test above would be a size $\alpha$ test if

$$\sup_{\mu \leq \mu_0} P(\text{Reject } H_0) = \sup_{\mu \leq \mu_0} \beta(\mu) = \alpha.$$

where $\beta(\mu)$ is the power function (derived on the notes of previous lectures). We can write

$$\sup_{\mu \leq \mu_0} \beta(\mu) = \sup_{\mu \leq \mu_0} \left[1 - \Phi\left(\mathcal{Z}_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)\right].$$

The function inside the supremum is increasing in $\mu$ and equal to $\alpha$ if $\mu = \mu_0$. Therefore the above supremum is equal to $\alpha$.

**Note:** The UMP test usually does not exist for 2-sided (composite) alternative hypotheses.

**Corollary:** Consider the previous testing problem, let $T(Y)$ be a sufficient statistic for $\theta$, and $g(t|\theta_0)$, $g(t|\theta_1)$ be its corresponding pdf's (or pmf's). Then any test with rejection region $S$ (a subset of the sample space of $T$) is a UMP level $\alpha$ test if it satisfies

$$t \in S, \text{ if } \frac{g(t|\theta_1)}{g(t|\theta_0)} > k,$$

for some $k > 0$ and $\alpha = P_{\theta_0}(T \in S)$.

**Proof:** Since $T$ is a sufficient statistic we can write

$$\frac{f(y|\theta_1)}{f(y|\theta_0)} = \frac{g(t|\theta_1)h(y)}{g(t|\theta_0)h(y)} = \frac{g(t|\theta_1)}{g(t|\theta_0)}.$$

## 27.3 Likelihood ratio test

Let $\Theta = \Theta_0 \cup \Theta_0^c$ and consider hypothesis testing problems with

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_0^c.$$

**Definition of Likelihood Ratio test:** Let $Y = y$ be an observed sample and define the likelihood by $L(\theta|y)$. The likelihood ratio test statistic is

$$\lambda(y) = \frac{\sup_{\theta\in\Theta} L(\theta|y)}{\sup_{\theta\in\Theta_0} L(\theta|y)}.$$

A likelihood ratio test is a test with rejection region $y : \lambda(y) \geq c$.

The constant $c$ may be determined by the size, i.e.

$$\sup_{\theta\in\Theta_0} P_\theta(\lambda(Y) > c) = \alpha.$$

**Notes:**

- The numerator is evaluated at the value of $\theta$ corresponding to the MLE, that is the maximum of the likelihood over the entire parameter range.
- The denominator contains a maximum over a restricted parameter range.
- Hence the numerator is larger or equal to the denumerator and the statistic of the likelihood ratio test is always greater than 1.
- Its distribution is usually unknown.

**Example (Likelihood Ratio test):** Let a random sample $Y = (Y_1, \ldots, Y_n)$ from a $N(\mu, \sigma^2)$, (with $\sigma^2$ known). Consider the test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. The MLE is $\hat{\mu} = \bar{Y}$, hence the likelihood ratio test statistic is

$$\begin{aligned}\lambda(Y) &= \frac{L(\hat{\mu}|Y)}{L(\mu_0|Y)} = \frac{(2\pi\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}[n(\bar{Y}-\hat{\mu})^2 + (n-1)S^2]\}}{(2\pi\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}[n(\bar{Y}-\mu_0)^2 + (n-1)S^2]\}} \\ &= \exp\left(\frac{n(\bar{Y}-\mu_0)^2}{2\sigma^2}\right).\end{aligned}$$

The test $\lambda(Y) > k$ is equivalent with the test $\left|\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}\right| \geq k_1$.

We can write the previous rejection region as

$$R = \left\{ y : \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \leq -k_1 \right\} \cup \left\{ y : \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \geq k_1 \right\} \quad \text{or}$$

$$R = \left\{ y : \bar{y} \leq \mu_0 - k_1\sigma/\sqrt{n} \right\} \cup \left\{ y : \bar{y} \geq \mu_0 + k_1\sigma/\sqrt{n} \right\}$$

Since $\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$, if we set $k_1 = \mathcal{Z}_{1-\alpha/2}$ the size will be $\alpha$:

$$\begin{aligned} P_{\mu_0}(Y \in R) &= P_{\mu_0}\left( \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \leq -\mathcal{Z}_{1-\alpha/2} \right) + P_{\mu_0}\left( \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \geq \mathcal{Z}_{1-\alpha/2} \right) \\ &= \Phi(\mathcal{Z}_{\alpha/2}) + 1 - \Phi(\mathcal{Z}_{1-\alpha/2}) = \alpha/2 + 1 - (1 - \alpha/2) \\ &= \alpha \end{aligned}$$

**Note:** Equivalently one can use the fact that $2\log\lambda(Y) = \frac{(\bar{Y}-\mu_0)^2}{\sigma^2/n} \sim \chi^2_1$.

**Theorem (Likelihood ratio test and sufficiency):** Let $Y$ be a sample parametrised by $\theta$ and $T(Y)$ be a sufficient statistic for $\theta$. Also let $\lambda(.)$, $\lambda^*(.)$ be the likelihood ratio tests for $Y$ and $T$ respectively. Then for every $y$ in the sample space

$$\lambda(y) = \lambda^*(T(y)).$$

**Proof:** because of sufficiency we have

$$\begin{aligned} \lambda(y) &= \frac{\sup_{\theta\in\Theta} L(\theta|y)}{\sup_{\theta\in\Theta_0} L(\theta|y)} = \frac{\sup_{\theta\in\Theta} g(\theta|T(y))h(y)}{\sup_{\theta\in\Theta_0} g(\theta|T(y))h(y)} \\ &= \frac{\sup_{\theta\in\Theta} L^*(\theta|T(y))}{\sup_{\theta\in\Theta_0} L^*(\theta|T(y))} = \lambda^*(T(y)). \end{aligned}$$

**Likelihood Ratio test for nuisance parameters**

Suppose that $\theta$ can be split in two groups: $\psi$ the main parameters and $\nu$ the parameters that are of little interest. We are interested in testing the hypothesis that $\psi$ takes a particular value

$$H_0 : \psi = \psi_0, \quad \nu \in N$$

$$H_1 : \psi \neq \psi_0, \quad \nu \in N.$$

The likelihood ratio test is

$$\lambda(y) = \frac{\sup_{\psi,\nu} L(\psi, \nu|y)}{\sup_{\nu} L(\nu|\psi_0, y)}.$$

The test may be viewed as a comparison between two models:

- The constrained model under $H_0$ with parameters $\nu$.

- The unconstrained model under $H_1$ with parameters $\nu, \psi$.

Generally in statistics, models with many parameters have better fit but do not always give better predictions. **Parsimonious** models achieve a good fit with not too many parameters. They usually perform better in terms of prediction. The likelihood ratio tests provides a useful tool for finding parsimonious models.

**Example (Likelihood Ratio test):** Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are two independent random samples from two exponential distributions with mean $\lambda_1$ and $\lambda_2$ respectively. We want to test

$$H_0 : \lambda_1 = \lambda_2 = \lambda, \quad \text{versus} \quad H_1 : \lambda_1 \neq \lambda_2.$$

The likelihood function is

$$L(\lambda_1, \lambda_2 | x, y) = \lambda_1^{-n} \exp\left( -\sum_{i=1}^{n} x_i/\lambda_1 \right) \lambda_2^{-n} \exp\left( -\sum_{i=1}^{n} y_i/\lambda_2 \right).$$

Under the unconstrained model of $H_1$ we have $\hat{\lambda}_1^{MLE} = \bar{X}$ and $\hat{\lambda}_2^{MLE} = \bar{Y}$. Under the constrained model of $H_0$ we get

$$\hat{\lambda}^{MLE} = (\bar{X} + \bar{Y})/2.$$

Hence, the likelihood ratio test statistic is

$$\begin{aligned} LR &= \frac{L(\hat{\lambda}_1^{MLE}, \hat{\lambda}_2^{MLE} | x, y)}{L(\hat{\lambda}^{MLE}, \hat{\lambda}^{MLE} | x, y)} = \frac{(\bar{X} + \bar{Y})^{2n}/2^{2n}}{\bar{X}^n \bar{Y}^n} \\ &= 2^{-2n} \left\{ \sqrt{\bar{X}/\bar{Y}} + \sqrt{\bar{Y}/\bar{X}} \right\}^{2n}. \end{aligned}$$

We do not know the distribution of the LR test statistic. We may attempt to isolate $T = \bar{X}/\bar{Y}$, but since LR is not monotone in $T$ we cannot construct a test.

**Note:** We will see in the next sections how deal with such cases by constructing asymptotic tests.

## 27.4 Other tests based on the likelihood

**The Wald test:** Suitable for testing simple null hypotheses $H_0 : \theta = \theta_0$ versus $H_0 : \theta \neq \theta_0$. The statistic of the test is

$$Z = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$$

The estimator $\hat{\theta}$ is the MLE and a reasonable estimate for its standard error $se(\hat{\theta}) = \sqrt{V(\hat{\theta})}$ is given by Fisher's information.

**The Score test:** Similar to the Wald test but it takes the form

$$Z = \frac{S(\theta_0)}{\sqrt{I(\theta_0)}}$$

where $S(\cdot)$ is the Score function and $I(\cdot)$ is the Fisher information.

Multivariate versions of the above tests exist. These tests are similar to the likelihood ratio test but not identical. As with the likelihood ratio test, their distribution is generally unknown. For 'large' sample sizes the likelihood ratio, score and Wald tests are equivalent.

# 28 Asymptotic Evaluations for Hypothesis Testing

## 28.1 Summary for Hypothesis Testing Issues

- In hypothesis testing we want to use the information from the sample $Y$ to choose between two hypotheses about the parameter $\theta$: the **null hypothesis** $H_0$ and the **alternative** $H_1$. The sample values for which the $H_0$ is rejected (accepted) is called **rejection (acceptance) region**.
- There are two possible types of errors. Type I error is if we **falsely reject** $H_0$ whereas type II is if we **don't reject** $H_0$ **when we should.**
- The **level** and **size** $\alpha$ of a test provide an upper bound for the type I error.
- The rejection region $R$, and hence the test itself, is specified using the probability that the sample $Y$ belongs to $R$ under $H_0$ is bounded by $\alpha$. If $H_0 : \theta \in \Theta_0$ we use

$$\sup_{\theta \in \Theta_0} P_\theta(Y \in R)$$

- A famous test is the **likelihood ratio test.**
- The type II error determines the **power** of a test. In practice we fix $\alpha$ and try to minimize (maximize) the type II error (power).
- To find a most powerful test we can use the **Neyman-Pearson Lemma**. It refers to tests where both $H_0$ and $H_1$ are simple hypotheses but can be extended in some cases to composite hypotheses. A version based on **sufficient statistics**, rather than the whole sample $Y$, is available.
- **Problem:** A uniformly most powerful test may not be available or may not exist. Sometimes it may be even hard to find any 'reasonable' test.

## 28.2 Asymptotic Evaluations

**Theorem (Asymptotic distribution of scalar LRTs):** For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, assume a random sample $Y$ parametrised by a scalar parameter $\theta$ and let $\lambda(Y)$ be the likelihood ratio test. Under $H_0$ as $n \to \infty$

$$2 \log \lambda(Y) \xrightarrow{d} \chi^2_1$$

provided certain regularity conditions hold for the likelihood.

Suppose that $\theta$ can be split in two groups: $\theta = (\psi, \nu)$ where $\psi$ are the main parameters of interest of dimension $k$. Consider the test

$$H_0 : \psi = \psi_0, \quad \nu \in N$$

$$H_1 : \psi \neq \psi_0, \quad \nu \in N.$$

Equivalently, suppose that we want to compare the constrained model of $H_0$ with the unconstrained model of $H_1$.

**Theorem (Asymptotic distribution of multi-parameter LRTs):** Provided certain regularity conditions hold, under $H_0$

$$2 \log \lambda(Y) \xrightarrow{d} \chi^2_k \quad n \to \infty.$$

Note that $k$ is the number of restrictions we are testing for.

**Example:** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from an Exponential($\lambda$) distribution. We want to test

$$H_0 : \lambda = \lambda_0 \quad \text{versus} \quad H_1 : \lambda \neq \lambda_0.$$

The MLE of $\lambda$ is $\hat{\lambda} = \bar{Y}$. The likelihood ratio test statistic is

$$LR(Y) = \frac{\sup_{\lambda > 0} L(\lambda|Y)}{L(\lambda_0|Y)} = \frac{\hat{\lambda}^{-n} \exp(-n\bar{Y}/\hat{\lambda})}{\lambda_0^{-n} \exp(-n\bar{Y}/\lambda_0)} = \frac{\bar{Y}^{-n} \exp(-n)}{\lambda_0^{-n} \exp(-n\bar{Y}/\lambda_0)}.$$

We cannot construct an exact test since

- The distribution of $LR(Y)$ is unknown,
- The distribution of $\bar{Y}$ is known but $LR(Y)$ is non-monotone in $\bar{Y}$.

Consider the quantity

$$2\log LR(Y) = 2n\left[\log(\lambda_0) - \log(\bar{Y}) - (1 - \bar{Y}/\lambda_0)\right]$$

The previous theorem establishes that, under $H_0$ and for $n \to \infty$,

$$2\log LR(Y) \sim \chi^2_1$$

Hence, the asymptotic likelihood ratio test of size $\alpha$ rejects if $2\log LR(Y) > \chi^2_{1,1-\alpha}$, where $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$th percentile of a $\chi^2_1$ distribution.

**Example:** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from an Poisson($\lambda$) distribution. We want to test

$$H_0 : \lambda = \lambda_0 \quad \text{versus} \quad H_1 : \lambda \neq \lambda_0.$$

The MLE of $\lambda$ is $\hat{\lambda} = \bar{Y}$. The likelihood ratio test statistic is

$$\begin{aligned} LR(Y) &= \frac{\sup_{\lambda>0} L(\lambda|Y)}{L(\lambda_0|Y)} = \frac{\exp(-n\hat{\lambda})\hat{\lambda}^{n\bar{Y}}}{\exp(-n\lambda_0)\lambda_0^{n\bar{Y}}} \\ &= \exp[n(\lambda_0 - \bar{Y})]\left(\frac{\lambda_0}{\bar{Y}}\right)^{-n\bar{Y}}. \end{aligned}$$

We cannot construct an exact test since

- The distribution of $LR(Y)$ is unknown,
- The distribution of $\bar{Y}$ is known but $LR(Y)$ is non-monotone in $\bar{Y}$.

Consider the quantity

$$2\log LR(Y) = 2n\left[(\lambda_0 - \bar{Y}) - \bar{Y}\log(\lambda_0/\bar{Y})\right]$$

The previous theorem establishes that, under $H_0$ and for $n \to \infty$,

$$2\log LR(Y) \sim \chi^2_1$$

Hence, the asymptotic likelihood ratio test of size $\alpha$ rejects if $2\log LR(Y) > \chi^2_{1,1-\alpha}$.

# Reading

G. Casella & R. L. Berger 8.1, 8.2.1, 8.3.1, 8.3.2, 8.3.4, 10.3, 10.4.1