Statistics for Data Science

Ricardo Hortelano

2019

Release:
v0.3.1

Todo list

Contents

Preface

Introduction

Write all this

Who am I?
Why this book?
What is maths?
How to read?
How to reference it?
How to collaborate?

Chapter 1

Linear Algebra

This chapter will focus on the very basis of mathematical knowledge and linear algebra, but builded from the very beginning.

1.1 Number Sets

- Natural $\mathbb{N} = \{1, 2, 3, \dots\}$
- Integer $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$
- Rational $\mathbb{Q} = \{r | r = \frac{m}{n}, \text{ where } m, n \in \mathbb{Z}, n \neq 0\}$
- Irrational $\mathbb{P} = \mathbb{R} \setminus \mathbb{Q} = \{\pi, \sqrt{2}, e, \dots\}$
- Real $\mathbb{R} = \mathbb{P} \cup \mathbb{Q}$
- Complex $\mathbb{C} = \{z = a + bi, \text{ where } a, b \in \mathbb{R}, i^2 = -1\}$

1.2 Fundamental theorem of Algebra

Any nth degree polynomial, such as,

$$r(x) = x^n + \alpha_1 x^{n-1} + \alpha_2 x^{n-2} + \cdots + \alpha_{n-1} x + \alpha_n$$

has n roots in $\mathbb{C}$ allowing multiplicity. Where

- A root is a value such as $r(x) = 0$

- Multiplicity means that a root can be repeat. ex: A root repeated two times is called a root with multiplicity of two.

We can use those roots to factorize a polynomial. For example, for a polynomial of grade two we can use the formula $\frac{-b\pm\sqrt{b^2-4ac}}{2a}$

$$r(x) = ax^2 + bx + c \Leftrightarrow r(x) = a(x - x_+)(x - x_-)$$

There is also a formula for cubic polynomials (grade three)

For polynomials of greather grade, it can be used the Ruffini's Rule

insert example of ruffini

1.3 Linear Equations

Chapter 2

Probability

This chapter will focus on the main probabilistic knowledge necessary to have a strong mathematical base. One of the key thing to know about probabilities is that all the theory is builded from the set theory. Because of that the very first points to take into account will be the fundamental set aspects. After that the chapter will go inside into the probabilistic theory.

2.1 Set Theory

Informally, a set can be defined as a collection of objects. The formally definition of what it's a set vary depending of the axiomatic definition of choice. Because set theory is not into the aim of this book and because sets can be studied intuitively, we are going to refer to it in his more fundamental form. Venn diagrams can be used to understand graphically most of the properties of sets.

2.1.1 Notation

If an object o is an element of a set A, then we can write $o \in A$. A set and his elements can be denoted by

$$A = \{o_1, o_2, \ldots, o_n\} \text{ where } n \text{ can be } \textit{finite} \text{ or } \textit{infinite}$$

If all elements of a set B are also member of a set A then we can say that B is a subset of A and it's denoted by $B \subset A$ or $B \subseteq A$ if A and B has exactly the same elements.

Sometimes a especial set U is used to refer the set that contains all possible objects.

2.1.2 Binary Operations

Set theory defines a group of operations that can be performed between two sets. Some of them are:

- **Union** of A and B, denoted $A \cup B$, is the set of all objects that are a member of A, or B, or both.
- **Intersection** of A and B, denoted $A \cap B$, is the set of all objects that are members of both A and B.
- **Complement** of A refers to elements not contained in A. Denoted $\overline{A}$ or A^c
- **Difference** of A and B, denoted $A \setminus B$, is the set of all members of A that are not members of B
- **Power** of A, denoted $\mathcal{P}(A)$, is the set whose members are all of the possible subsets of A.

2.1.3 Properties

Derived from above there is some general properties called the **fundamental properties of set algebra**. These properties are essential to understand the set theory. In fact, if the set theory is completely understand, these properties are automatically assimilated. Thus it's not necessary to memorize it.

- Commutative:

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

- Associative:

$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$

- Distributive:

Union with Intersection: $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

Intersection with Union: $(A \cup B) \cap C = (A \cap C) \cup (B \cup C)$

- Neutral Element: $A \cup \emptyset = A = A \cap S$
- Complementation: $A \cup \overline{A} = S$ and $A \cap \overline{A} = \emptyset$
- Idempotence: $A \cup A = A$ and $A \cap A = A$
- Abosortion: $A \cup S = S$ and $A \cap \emptyset = \emptyset$
- Simplication: $A \cap (A \cup B) = A = A \cup (A \cap B)$

De Morgan's laws

- $\overline{A \cup B} = \overline{A} \cap \overline{B}$
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$

2.2 Combinatory

Also know as expertise in counting

2.2.1 Rule of Product

If there are α ways of doing something and β ways of doing another thing, then there are $\alpha\beta$ ways of performing both actions.

2.2.2 k-permutations

k-permutations of n elements:
The number of **ordered sequences** with $1 \leq k \leq n$ that can be formed of n elements is,

$$n(n-1)(n-2)\dots(n-k+1) = \frac{n!}{(n-k)!}$$

2.2.3 k-combinations

k-combinations of n objects:
The number of **unordered sequences** within a set with $1 \leq k \leq n$ that can be formed of n elements is,

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

2.2.4 Partitions

r groups containing n_i objects each,

$$\binom{n}{n_1, n_2 \ldots, n_r} = \frac{n!}{n_1! n_2! \ldots n_r!}$$

2.3 Random Experiments

A experiment is deterministic when knowing the state of the set of variables involved in the experiment the outcome is always the same. Conversely a experiment is random when knowing the state (or when we ignore some of these states) of the set of variables involved in the experiment the outcome differs.

In other terms, **an experiment is random** if although it is repeated in the same manner every time, can result in different outcomes. Thus, the fixed outcome of a random experiment is impossible to predict in advance although the number of individual possibles outcomes is known in advance. The probability theory study the random experiments.

The notation of some elements involved in a random experiment are the followings:

- **Sample Space**, is the set of all possible outcomes of the random experiment. It's denoted by S or Ω
- **Individual Outcomes**, are the type of possible outcomes of a random experiment. It's denoted by ω
- **Event**, is a subset of S. Usually denoted by capital letters starting by A, B, C, etc. Sometimes denoted by $\mathcal{F}$ for his general form.

- **Null Event**, is a special event that never occurs. Denoted by $\emptyset$
- **Frequency** of ω is the number of times the individual outcome ω occurs in a random experiment.
- **Relative Frequency** of ω is the ratio between the frequency of ω and the total number of outcomes of a random experiment.

Once the experiment has been performed, it is said that A "happened" if the outcome of the experiment (ω) belongs to A.

2.3.1 Events Operations

If the logical meaning from the set operations is taken and restricted to a probabilistic perspective, then a new meaning for these operations can be defined in this new scope. Thus:

- **Union** $A \cup B$ (Grammatically A or B), occurs when either of the two events (or both of them simultaneously) do occur.
- **Intersection** $A \cap B$ (Grammatically A and B), occurs when both of them do simultaneously occur.
- **Complementary** $\overline{A}$ (Grammatically not A), occurs when the event does not occur.
- **Difference** $A \setminus B$ (Grammatically A and not B), occurs when the first event does occurs, but the second does not. Note that $A \setminus B = A \cap \overline{B}$

2.3.2 Formal Definitions

From an formal perspective, the concept of probability has been defined multiples times. Here they are collected the fundamental ones that tries to formalize the abstract concepts of what it's usually understand as probability and his parts.

σ-algebra

A σ-algebra (sigma-algebra) $\mathcal{A}$ over a set Ω is a family (collection) of subsets (with elements $E_1, E_2, \dots$) of Ω that satisfies:

- $\emptyset \in \mathcal{A}$
- If $E \in \mathcal{A}$ then $\overline{E} \in \mathcal{A}$
- If $E_1, E_2, \cdots \in \mathcal{A}$ then $\cup_{n=1}^{\infty} E_i \in \mathcal{A}$

σ-algebra discrete

The discrete σ-algebra of Ω is the power set of Ω ($\mathcal{P}(\Omega) = \{E : E \subset \Omega\}$), that is, the collection of all subsets of Ω.

For example, given a random experiment that toss one coin,

H: The coin shows head.

T: The coin shows tail.

$S = \{H, T\}, \quad \mathcal{P}(S) = \{\emptyset, \{H\}, \{T\}, S\}$

Measurable Space

The pair $(\Omega, \mathcal{A})$ is a measurable space if $\mathcal{A}$ is a σ-algebra over Ω

Frequentist Definition

The probability of an event A is the limit of the relative frequency of that event when the number of repetitions of the experiment tends to infinity. If the experiment is repeated n times, and n_A is the number of repetitions in which A happens, then the probability of A is

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

LaPlace's Definition

This definition can be used for random experiments that have a finite number of outcomes and all of them are equally likely.
The probability of an event A is the ratio between the favorable outcomes to A and the total outcomes of the experiment, thus,

$$P(A) = \frac{\#A}{\#\Omega}$$

This implies that, given,

$$S = \{s_1, s_2, \dots s_n\}, \quad P(s_n) = \frac{1}{n}$$

Kolmogorov Definition

The Kolmogorov definition is the only one that does not define what is a probability function. In fact, this definition stablish three axioms that must be satisfied by any probability function. This axioms are a fundamental part of probability theory.
Let the pair $(\Omega, \mathcal{A})$ be a *measurable space*. In it, the probability function P of some event E, denoted $P(E)$ is an application over the real numbers $(P : \mathcal{A} \to \mathbb{R})$ that satisfies:

- **Non Negativity**: $P(E) \geq 0, \forall E \in \mathcal{A}$
- **Unitarity**: $P(\Omega) = 1$
- **Additivity**: Any countable sequence $E_1, E_2, \dots$ of disjoint events of $\mathcal{A}$, satisfies
$$P(\bigcup_{n=1} E_i) = \sum_{n=1} P(E_i)$$

Cox's Theorem

Used by bayesians. write explanation

2.3.3 Properties

These are the properties that apply independency of the probability definition of choice.

- $P(\overline{A}) = 1 - P(A)$
- $P(\emptyset) = 0$
- If $A \subseteq B$ then $P(A) \leq P(B)$
- $P(A \setminus B) = P(A) - P(A \cap B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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

2.3.4 Conditional Probability

The probability of the event A occurs knowing before hand that the event B has occurred is,

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

Note that when we have a conditional probability, it can be say that B becomes the new sample space of the experiment.

Properties

The conditional probability holds all the properties of the regular probability.

2.3.5 Bayes Theorem

Given a set of events $A_1, A_2, \ldots, A_n$, the probability of all of them occurring simultaneously is called **probability chain rule**, and it is

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)\ldots P(A_n|A_1 \cap \cdots \cap A_{n-1}) = \prod_{k=1}^{n} P(A_k|\bigcap_{j=1}^{k-1} A_j)$$

or alternatively, in his two events form,

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

Total Probability Rule

Given $A_1, \ldots, A_n$ mutually exclusive (disjoint) events whose union is the whole of the sample space (partition) and assume $P(A_i) > 0$ for every i .For every event B we have

$$P(B) = P(A_1 \cap B) + \cdots + P(A_n \cap B) = \prod_{k=1}^{n} P(A_k|\bigcap_{j=1}^{k-1} A_j)$$

Bayes Theorem

Knowing of above, then we can write,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{P(A_1 \cap B) + \cdots + P(A_n \cap B)}$$

Note: The different components of the bayes formula have names:

1. $P(A|B)$ is called the **posterior probability**.
2. $P(A)$ is called the **prior probability**.
3. $P(B)$ is called **marginal distribution**.

2.3.6 Independency of two events

Two events are independence if and only if,

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

If A and B are independence, then the occurrence of A does no provide any information about B. Formally,

If A, B independ. and $P(B) > 0$ then $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\cancel{P(B)}}{\cancel{P(B)}} = P(A)$

Conditional Independece

Given two events A and B, they are conditionally independence if given C and $P(C) > 0$,

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

and

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

explanatory Venn diagram

2.4 Discrete Random Variables

When we are dealing with random experiments, it's very common to use a number to represent a event of the sample space. It's called random variable.

A **random variable** (r.v.) is the numerical outcome of a random experiment. Formally, a random variable X is a mapping $X : \Omega \to \mathbb{R}$

Write formal definition using borel sigma-algebra

It associates each outcome of a random experiment with a real number and is measurable because the inverse image of every borelian set does belong to the σ-algebra of events.

Support of Random Variables

We can define the support of the random variable depending of the size of the mapped set. A r.v. is:

- **Discrete** if $X(\Omega)$ is a finite or numerable set.
- **Continuous** if $X(\Omega)$ contains an interval of $\mathbb{R}$

2.4.1 Probability Mass Function

A probability mass function (p.m.f.) match a real number x with the probability that an event exactly occurs.

$$p(x) = P(X = x) = P_X(x), \quad \forall x \in \mathbb{R}$$

Properties

write this

2.4.2 Cumulative Distribution Function

The cumulative distribution function (c.d.f.) measures the probability that an event is not greater or equal than a real number.

$$F(x) = P(X \leq x) = P_X((-\infty, x])), \quad \forall x \in \mathbb{R}$$

Properties

- $\lim_{x \to -\infty} F(X) = 0$
- $\lim_{x \to \infty} F(X) = 1$
- F is non-decreasing
- F is right-continuous

2.4.3 Mean, Variance and Standard Deviation

Mean or Expectation

Let X be a r.v. with the probability function $p(x)$, then, the *expected value* of X is,

$$\mu_X \approx \mathbb{E}[X] = \sum_x x\, P(X = x)$$

if $p(x)$ is an accurate characterization of the *population* frequency distribution, then,

$$\mu_X = \mathbb{E}[X]$$

Properties

- $\mathbb{E}[c] = c$, if c is a constant
- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- $\mathbb{E}[g(x)] = \sum_x g(x)P(X = x)$
- $\mathbb{E}[X^k] = \sum_x x^k P(X = x)$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[X - Y] = \mathbb{E}[X] - \mathbb{E}[Y]$
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, if X and Y are independent

Variance

Let X be a r.v. with mean $\mathbb{E}[X] = \mu$, probability function $p(x)$, then, the variance of X is defined to be the expected value of $(X - \mu)^2$, thus,

$$\sigma_X^2 \approx V[X] = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mu^2 = \sum_x (x - \mathbb{E}[X])^2 P(X = x)$$

if $p(x)$ is an accurate characterization of the *population* frequency distribution, then,

$$\sigma_X^2 = V[X]$$

Properties

- $V[X] \geq 0$
- $V[aX + b] = a^2 V[X], \quad \forall a, b \in \mathbb{R}$
- $V[X \pm Y] = V[X] + V[Y]$, if X and Y are independent.
- $V[X + Y] = V[X] + V[Y] + 2\text{Cov}(X, Y)$

Standard Deviation

The standard deviation of X, is the positive square root of $V[X]$

$$\sigma = \sqrt{V[X]}$$

2.4.4 Median and Quantiles

Median

The median is the most central value of a distribution of a r.v. X, thus, the median m of X satisfies,

$$P(X \leq m) \geq \frac{1}{2}$$

and

$$P(X \leq m) \leq \frac{1}{2}$$

explanatory image

α-quantiles

For $0 < \alpha < 1$ the α-quantile of a r.v. X is a number q_α that satisfies,

$$P(X \leq q_\alpha) \geq \alpha$$

and

$$P(X \geq q_\alpha) \geq 1 - \alpha$$

explanatory image

2.4.5 Quantile Function

For $0 < \alpha < 1$ the *quantile function* of a r.v. is defined as,

$$F_X^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}$$

explanatory image

Properties

write this

2.5 Discrete Random Distributions

2.5.1 Bernoulli Process

A **Bernoulli process** is random experiment in which outcome can only result in two different values. This outcomes are commonly refered as *success* and *failure*. The probability of success is $0 \leq p \leq 1$ and all experiment trials are *independent*.

$$X \sim Bernoulli(p)$$

2.5.2 Binomial Distribution

Consider a Bernoulli process with probability of success p and that is carried out independently n times, a Binomial random variable X with parameters n and p represents the number of trials that result in success.

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

with

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, 2, \ldots, n\}$$

and have expectation and variance

$$\mathbb{E}[X] = np, \quad V[X] = np(1-p)$$

Note: Be careful with the definition. It's not the same 'number of trials' and 'number of success'.

Using this base experiment we cant try to fit other rand. exp. and see if this distribution is valid.

Add base example that serves as base experiment.

Properties

If X and Y are independent, and $X \sim B(n_1, p), Y \sim B(n_2, p)$ then

$$X + Y \sim B(n_1 + n_2, p)$$

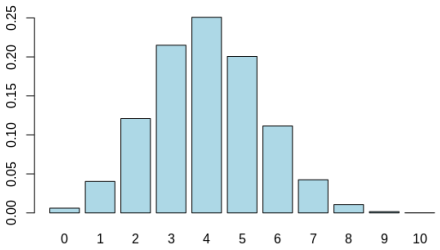


Figure 2.1: Binomial p.m.f.

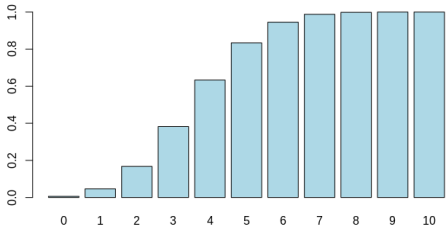


Figure 2.2: Binomial c.d.f.

2.5.3 Geometric Distribution

Consider a Bernoulli trial with probability of success p, the number of independent trials that result in *failure* obtained before the first success follows a Geometric distribution with parameter p.

$$X \sim \mathcal{G}(p)$$

with

$$P(X = x) = p(1 - p)^x, \quad x \in \{0, 1, 2, \ldots, n\}$$

and have expectation and variance

$$\mathbb{E}[X] = \frac{1 - p}{p}, \quad V[X] = \frac{1 - p}{p^2}$$

Ussing this base experiment we cant try to fit other rand. exp. and see if this distribution is valid.

Add base example that serves as base experiment.

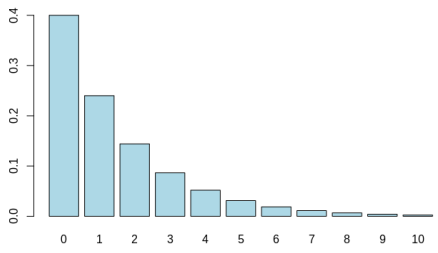


Figure 2.3: Geometric p.m.f.

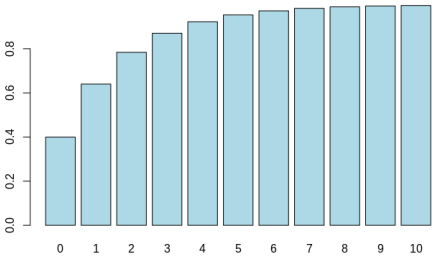


Figure 2.4: Geometric c.d.f.

2.5.4 Negative Binomial Distribution

Consider a Bernoulli trial with probability of success p, the number of failures (independent trials that result in failure) before the k-th success (that is, **number of trials** until k-success) follows a Negative Binomial distribution with parameters k and p.

$$X \sim NB(k, p)$$

with

$$P(X = x) = \binom{x + k - 1}{x} p^k (1 - p)^x, \quad x \in \{0, 1, 2, \ldots, n\}$$

and have expectation and variance

$$\mathbb{E}[X] = \frac{k(1-p)}{p}, \quad V[X] = \frac{k(1-p)}{p^2}$$

Note: Be careful with the definition. This distribution counts the **number of fails** not the number of fails + success

Add base example that serves as base experiment.

Ussing this base experiment we cant try to fit other rand. exp. and see if this distribution is valid.

2.5.5 Hypergeometric Distribution

Consider a **finite population** with $N_1 + N_2$ objects, such that N_1 are of type 1 and N_2 are of type 2. A total number of k objects are selected from the population **without replacement**. The number of objects of type N_1 in the selection follows a Hypergeometric distribution with parameters N_1, N_2, and k.

$$X \sim H(N_1, N_2, k)$$

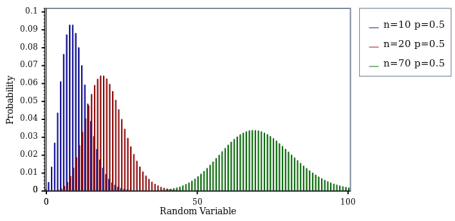


Figure 2.5: Negative Bin. p.m.f.

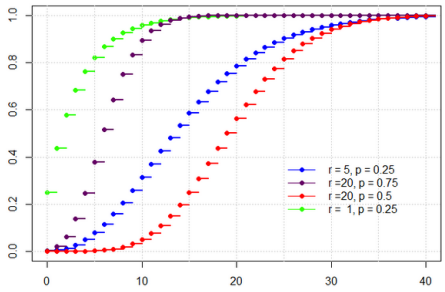


Figure 2.6: Negative Bin. c.d.f.

with

$$P(X = x) = \frac{\binom{N_1}{x}\binom{N_2}{k-x}}{\binom{N_1+N_2}{k}}, \quad x \in \{\max\{0, k - N_2\}, \ldots, \min\{k, N_1\}\}$$

and have expectation and variance

$$\mathbb{E}[X] = \frac{kN_1}{N_1 + N_2}, \quad V[X] = k \cdot \frac{N_1 N_2}{(N_1 + N_2)^2} \cdot \frac{N_1 + N_2 - k}{N_1 + N_2 - 1}$$

Add base example that serves as base experiment.

Ussing this base experiment we cant try to fit other rand. exp. and see if this distribution is valid.

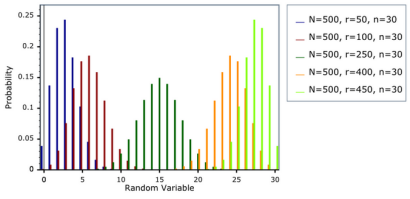


Figure 2.7: Hypergeometric p.m.f.

2.5.6 Poisson Distribution

The number of events that occur in a region of space (or time) independently one from the others and at a constant rate $\lambda > 0$ follows a Poisson distribution with parameter λ.

$$X \sim \mathcal{P}(\lambda)$$

with

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x \in \{0, 1, 2, \ldots, n\}$$

and have expectation and variance

$$\mathbb{E}[X] = \lambda, \quad V[X] = \lambda$$

Add base example that serves as base experiment.

Ussing this base experiment we cant try to fit other rand. exp. and see if this distribution is valid.

Properties

If X and Y are independent, and $X \sim \mathcal{P}(\lambda_1), Y \sim \mathcal{P}(\lambda_2)$ then

$$X + Y \sim \mathcal{P}(\lambda_1 + \lambda_2)$$

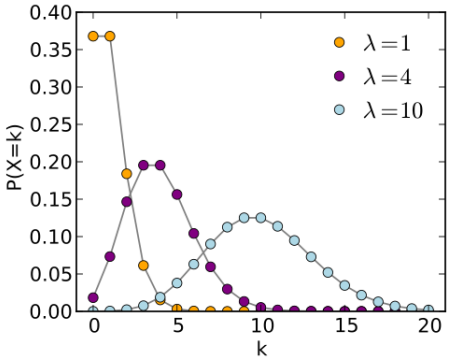


Figure 2.8: Poisson p.m.f.

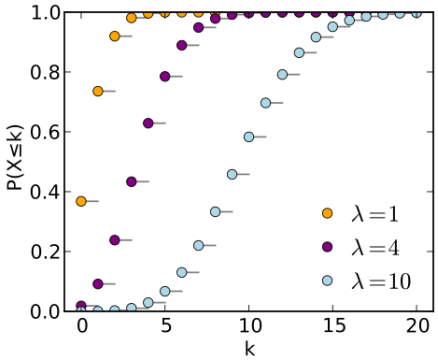


Figure 2.9: Poisson c.d.f.

2.5.7 Multi Bernoulli Distribution

write this

2.5.8 Zero Inflated Poisson Distribution

write this

2.6 Continuous Random Variables

add moment generating function section

A continuous r.v. X is the one that satisfies the following conditions:

- The distribution function F_x is always continuous.
- The distribution function F_x is differentiable and its derivative is continuous except for a countable set of points.

2.6.1 Probability density function

The p.d.f. (also referred as density mass function) of X is a function such as $f_X : \mathbb{R} \to \mathbb{R}$ and it is

$$f_X(x) = \begin{cases} 0 & \text{if } x \in S \\ F'_X(x) & \text{if } x \notin S \end{cases}$$

Note that the above definition is similar to:

$$P(X \in A) = \int_A f_X(x)dx$$

For $A \subset \mathbb{R}$ a borelian set.

explanatory p.d.f.

2.6.2 Cumulative distribution function

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

Properties

- $\lim_{x\to-\infty} F(x) = 0$
- $\lim_{x\to\infty} F(x) = 1$
- F is non decreasing
- F is continuous

In order to compute the probability that X lies in an interval $[a, b]$ just

$$P(a \leq X \leq b) = F_X(b) - F_X(a)$$

Relationship between p.d.f and c.d.f.

It's important to note that the c.d.f. is just the primitive of the p.d.f. and thus, the p.d.f. is the derivative of the c.d.f.

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \qquad \text{and} \qquad f_X(x) = F'_X(x)$$

2.6.3 Mean, variance and quantiles

Mean

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx$$

The mean holds the same properties as it is does in his discrete form

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- $\mathbb{E}[g(X)] = \int g(x)f_X(x)dx$
- $\mathbb{E}[(X - \mathbb{E}[X])^2] = \min_{x\in\mathbb{R}} \mathbb{E}[(X - x)^2]$

Variance

$$V[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int (x - \mathbb{E}[X])^2 f_X(x) dx$$

Properties

- $V[X] \geq 0$
- $V[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- $V[aX + b] = a^2 V[X] \quad \forall a, b \in \mathbb{R}$

Standard deviation

$$\sigma_X = \sqrt{V[X]}$$

complete section with, median, quantiles

2.7 Continuous Random Distributions

2.7.1 Uniform Distribution

write description

$$X \sim U(a, b)$$

with

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathbb{E}[X] = \frac{a+b}{2}, \qquad V[X] = \frac{(b-a)^2}{12}$$

pdf and cdf

2.7.2 Exponential distribution

Used to measure the time between the occurrence of two events in a poisson distribution $X_t \sim \mathcal{P}(t\lambda)$ where t is the number of events in $[0, t]$.

write description

$$T \sim Exp(\lambda)$$

with c.d.f.

$$F_T(t) = \begin{cases} 1 - e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

with p.d.f

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

and

$$\mathbb{E}[T] = \lambda^{-1}, \quad V[T] = \lambda^{-2}$$

Lack of memory property

$$P(T > t_1 + t_2 | \, T > t_1) = P(T > t_2)$$

pdf and cdf

2.7.3 Normal Distribution

write description

$$X \sim \mathcal{N}(\mu, \sigma)$$

with

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}$$

if $\mu = 0$ and $\sigma = 1$ then it is said that $\mathcal{N}$ is in standard normal and

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad x \in \mathbb{R}$$

Note that a linear transformation of a normal is also a normal

$$aX + b \sim \mathcal{N}(a\mu + b, \quad |a|\sigma)$$

Probably the most important transformation is the one called **standarization**

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

pdf and cdf

Normal aproximation to the binomial distribution

complete

COMPLETE WITH gamma, beta distributions

2.8 Random Vectors

Until now the only distributions that we have seen are in $\mathbb{R}$, or in other words in one single dimension. Now we are going to study random variables in more dimensions.

2.8.1 Marginal distributions

The distributions of each fo the components of a r.v. alone is referred to as Marginal distribution.

Random Vector

A random vector is a measurable mapping from a sample space S into $\mathbb{R}^d$. For example, a bivariate random vector maps S into $\mathbb{R}^2$

$$(X, Y) : S \to \mathbb{R}^2$$

Joint distribution of a r.v.

The joint distribution is the one that describes the behavior of all variables that compose the random vector. So, since now, we are going to talk about joint distributions of two or more random variables.

2.8.2 Discrete random vectors

Let X, Y be two discrete random variables on the same probability space.

Joint probability mass function

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

with $p_{X,Y}(x, y) > 0$ and $\sum_x \sum_y p_{X,Y}(x, y) = 1$

Joint cumulative distribution function

$$F_{X,Y}(x_0, y_0) = P(X \leq x_0, Y \leq y_0) = \sum_{x \leq x_0} \sum_{y \leq y_0} p_{X,Y}(x, y)$$

Marginal discrete distributions

Given X, Y with $p_{X,Y}(x, y)$ his joint probability mass function.

- **Marginal p.m.f of X**: $p_X(x) = P(X = x) = \sum_y P(X = x, Y = Y) = \sum_y p_{X,Y}(x, y)$
- **Marginal p.m.f of Y**: $p_Y(y) = P(Y = y) = \sum_x P(X = x, Y = Y) = \sum_x p_{X,Y}(x, y)$

2.8.3 Continuous random vectors

Let X, Y be two continuous random variables on the same probability space.

Joint density mass function

$$f_{X,Y}(x, y) = P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y) dy dx$$

with $f_{X,Y}(x, y) > 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy dx = 1$

Joint cumulative distribution function

$$F_{X,Y}(x_0, y_0) = P(X \leq x_0, Y \leq y_0) = \int_{-\infty}^{x_0} \int_{-\infty}^{y_0} f_{X,Y}(x, y) dy dx$$

Marginal continuous distributions

Given X, Y with $f_{X,Y}(x, y)$ his joint density mass function.

- **Marginal p.d.f of X**:
  $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$
- **Marginal p.d.f of Y**:
  $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx$

example with image of the diamond distribution

2.8.4 Independence of random vectors

X and Y are independent if any of the followings conditions

Discrete Variables	**Continuous Variables**
$pY\|X(y\|x) = p_Y(y)$	$fY\|X(y\|x) = f_Y(y)$
$p_{X\|Y}(x\|y) = p_X(x)$	$f_{X\|Y}(x\|y) = f_X(x)$
$p_{X,Y}(x, y) = p_X(x)p_Y(y)$	$f_{X,Y}(x, y) = f_X(x)f_Y(y)$

square random vector showing the only true indepence variables

2.8.5 Transformations of random vectors

Consider a d-variate r.v. $\mathbf{X} = (X_1, \ldots, X_d)^t$ (note the bold X) and a function $g : \mathbb{R}^d \to \mathbb{R}^k$ then $\mathbf{Y} = g(\mathbf{X})$ is a k-variate random vector. Note that if $k = 1$ then $Y = g(\mathbf{X})$ is a random variable (note that the bold Y disappears here).

Mean of a univariate transformation

- **X** discrete: $\mathbb{E}[Y] = \mathbb{E}[g(\mathbf{X})] = \sum g(\mathbf{x}) p_{\mathbf{X}}(\mathbf{x})$
- **X** continuous: $\mathbb{E}[Y] = \mathbb{E}[g(\mathbf{X})] = \int_{\mathbb{R}^d} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$

Join density mass function of a transformation

Is the jacobian matrix

complete, or maybe remove from here

Convolutions or sum of independent r.v.

complete

2.8.6 Mean Vector

The mean vector of a random vector **X** is a column vector with components equals to the mean of the components

$$\mu = \mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}$$

2.8.7 Covariance

The covariance is a value that measures how one value vary respect the other. Is one of the principal values used to determine if two variables are independent.

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

2.8.8 Correlation

The correlation is a value that measures the dependency between two variables.

Pearson's product-moment coefficient

The most popular measure of dependency. It's a value in $[-1, 1]$

$$\rho_{X,Y} = \frac{\mathrm{Cov}[X, Y]}{\sqrt{V[X]V[Y]}}$$

picture showing the 4 zones of a correlation

Note that, if X and Y are independent then $\mathrm{Cov}[X, Y] = 0$ and the reverse does not necessary hold. Correlation is not causality.

add more types of correlations, linear and not linear, rank-corr etc

2.8.9 Covariance matrix

Denoted by Σ is a square matrix $d \times d$ symmetric positive semidefinite matrix, such that the element in position (i, j) is $\mathrm{Cov}[X_i, X_j]$

complete

Linear transformations

complete

2.8.10 Linear combinations of components of a random vector

complete with definition and properties

2.8.11 Multivariate Normal Distribution

Is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions. For $k = 1$ is the normal distribution, for $k = 2$ is the bivariate normal, for $k > 2$ is the multivariate.

$$\mathbf{X} \sim \mathcal{N}_d(\mu, \Sigma)$$

In R, library(MASS), function mvrnorm

complete

descriptive figure with diferent parameters

2.8.12 Bivariate Normal distribution

$$(X_1, X_2)^t \sim \mathcal{N}_2(\mu, \Sigma)$$

with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{\sigma_1,\sigma_2} \\ \rho_{\sigma_2,\sigma_1} & \sigma_2^2 \end{bmatrix}$$

complete

descriptive figure with diferent parameters

2.8.13 Multinomial Distribution

The multinomial distribution is the generalation of the binomial distribution. It allows to model the Bernoulli process with more than a binary result. For example, it models the probability of counts of each side for rolling a k-sided die n times.

$$\mathbf{X} \sim M(n, p_1, \ldots, p_k) \quad \text{with} \quad \sum_{i=1}^{k} p_i = 1$$

with

$$P(X = x) = \frac{n!}{x_1! \ldots X_k!} p_1^{x_1} \cdots p_k^{x_k}$$

and have expectation and variance

$$\mathbb{E}[X_i] = np_i \quad V[X_i] = np_i(1 - p_i)$$

complete

descriptive figure with diferent parameters

2.9 Mixtures

Let $F_1, F_2, \ldots, F_k$ be c.d.f. of different distributions and $p_1, p_2, \ldots, p_k > 0$ with $\sum_{i=1}^{k} p_i = 1$, then a new c.d.f. of a mixture distribution is defined as

$$G(x) = p_1 F_1(x) + \cdots + p_k F_k(x)$$

This new distribution mixes the later distributions according to the probability distribution given by $p_1, \ldots, p_k$. The cdf, density (or probability) mass function, or the random number generation can be done directly with the original distributions, BUT the quantile function is not straightforwardly computed from the ones of the original distributions.

remove this last paragraph?

mixture of two normals

2.9.1 Mean and variance

If $G = \sum^k p_i F_i$ with mean μ and variance σ^2

$$\mu = \sum_{i=1}^{k} p_i \mu_i \qquad \sigma^2 = \sum_{i=1}^{k} p_i(\mu_i^2 + \sigma_i^2) - \mu^2$$

2.9.2 Uncountable or continuous Mixtures

Let ω be a weight function

$$G(x) = \int_A \omega(a) F_a(x) da$$

complete this

2.10 Some statistics?

probably move this to inference chapter

2.10.1 Lorenz curve

If $X \geq 0$

$$L_X(x) = \frac{1}{\mathbb{E}[X]} \int_0^x F_X^{-1}(t) dt$$

The Lorenz curve represents the proportion of a given characteristic (wealth) earned by the fraction x of individuals with the smallest value in the characteristic (poorest individuals).

review this and write about his generalized form

lorenz curve

2.10.2 Gini coefficient

The Gini index or Gini coefficient equals twice the area between the Lorenz curve and the line segment from $(0, 0)$ to $(1, 1)$

$$G(X) = 1 - 2\int_0^1 L_X(t)dt = \frac{\text{GMD}(X)}{\mathbb{E}[X]}$$

Gini mean difference

Let X, Y be independent and follow the same distribution

$$\text{GMD}(X) = \frac{1}{2}\mathbb{E}[|X - Y|]$$

Chapter 3

Statistical Inference

This chapter will focus in the procedures to analyze a phenomena of which some information is unknown. Until now, we had analyzed problems and experiments in which the whole structure of the problem is known before. But, what happen when we couldn't know the probabilities related with all events? There is any approach to overcome this? Someone could just think, "Well, lets perform the experiment n times and let's see what happen."

This approach is the one we are going to follow in this chapter, and in it, it's analyzed the problems and benefits of performing such approach. The only prior we have to take in to account is that **all random phenomena follows an underlying probability function**. With this in mind. Let's start defining in a formal way "perform the experiment n times"

3.0.1 Parametric Family of Distributions

Let F be a p.d.f. with parameter θ

Usually, for some random experiments, the trials could give us some information about the shape of the underlying probability function F of the experiment. In practice, F is not exactly know, but we know that his parameter θ model his shape. So, we can know that F belongs to a set or *parametric family of distributions* $\{F(\,\cdot\,;\theta) \mid \theta \in \Theta\}$.

Because our objective is to know more about the real F we need to perform some procedures in order to *infer* the real value of θ. If we achieve to

infer the real value, or at least, an approximately one, then we could find, a good approximation of F inside his family of distributions. In order to extract this information about F we need to perform independent realizations of the experiment related with F. With this extracted knowledge we could perform inference over F or his parameter θ.

The fact that we perform independent repetitions of the experiment means that, for each repetition of the experiment, we have an associated r.v. with the same distribution F. This is called, *simple random sample*.

3.0.2 Types of estimation (Inference)

Review the taxonomy and refactor to be the same in the whole book

- Model Based (exact, approximated)
- Sampling Based
- Bayesian Based

3.0.3 Random Sample

A *simple random sample* of a r.v. X with distribution F is a collection of r.v. $(X_1, X_2, \ldots, X_n)$ that are **independent and identically distributed (i.i.d.)** and all of them follows the same distribution F

Because the random variables are independent, the c.d.f. of the sample is,

$$F_X(x_1, x_2, \ldots, x_n) = F(X_1)F(x_2)\ldots F(x_n)$$

3.0.4 Statistic

A statistic T is any measurable function $T : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}^k, \mathcal{B}^k)$, where k is the dimension of the statistic. For example:

- $T_1(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \triangleq \bar{X}_n$, this is called **Sample Mean**
- $T_1(X_1, X_2, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \triangleq S_n'^2$, this is called **Sample Variance**

- $T_1(X_1, X_2, \ldots, X_n) = \min\{X_1, \ldots, X_n\} \triangleq X_{(1)}$
- $T_1(X_1, X_2, \ldots, X_n) = \max\{X_1, \ldots, X_n\} \triangleq X_{(n)}$
- $T_1(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} \log X_i$
- $T_1(X_1, X_2, \ldots, X_n) = (X_{(1)}, X_{(n)})$

All these statistics have dimension $k = 1$ except the last with $k = 2$. The distribution induced by the statistic T is called the *sampling distribution* of T, since it depends on the distribution of the sample.

3.0.5 Estimators or Point Estimators

An estimator is an statistic that tries to estimate the value of a parameter θ that belongs to the distribution of the variable X. Thus,

Let $X \sim F(\cdot\,;\,\theta)$ where θ is a parameter vector with possibles values in the *parameter space* Θ. An estimator $\hat{\theta}_n$ of θ is a statistic $\hat{\theta}_n(X_1, \ldots, X_n)$ with values also in Θ

To clarify with a example, in the real world many r.v. follows a normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}^+$ that are unknown. Because of this, it's usually assumed that the distribution of a r.v. belong to a normal family of distributions $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$. Knowing this, it is said that the **sample mean** $\bar{X}$ and the **sample variance** S^2 are good estimators of the distribution mean μ and distribution variance σ^2.

Population Parameter	Sample Estimator
$p = P(X = 1)$	$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
$\mu = \mathbb{E}[X]$	$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
$\sigma^2 = V[X]$	$S_n'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

3.1 Exact inference under Normal Distributions

Sampling distributions in Normal Populations

The sample mean $\bar{X}$ and sample variance S^2 estimators play an important

role in statistical inference, since both are "good" estimators of μ and σ^2, respectively. As a consequence, it is important to obtain their sampling distributions in order to know their random behaviors. We will do so under the assumption of normal populations.

3.1.1 Expected value and Variance of the sample mean

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then, it holds,

$$\mathbb{E}[\bar{X}_n] = \mu, \quad V[\bar{X}_n] = \frac{\sigma^2}{n}$$

3.1.2 Expected value of the sample variance

Let $S_n'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ be the sample variance. Then, it holds,

$$\mathbb{E}[S_n'^2] = \sigma^2$$

3.1.3 Logarithmic Distribution Transformations

Review all this

If we take a constant k and perform a logarithmic transformation of the sample distribution $\bar{X}_n$ with this constant. For a fixed k it holds,

$$\log(\bar{X}_n + k) \sim \mathcal{N}(\mu, \sigma^2)$$

3.1.4 Sample Mean of Normal Distribution

Let $(X_1, \ldots, X_n)$ a s.r.s. of size n of a r.v. $\mathcal{N}(\mu, \sigma^2)$. Then, the **sample mean** $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ satisfies,

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

3.1.5 Z statistic or Standardization

The Z statistic is obtained when the sample mean is standardized,

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

3.1.6 Chi-Square Distribution

A Chi-square distribution $\mathcal{X}_n^2$ with n degrees of freedom, is the distribution that gives the sum of n independent r.v. $Z_1, \ldots, Z_n$ all of them with $\mathcal{N}(0,1)$ distributions. Thus,

$$\sum_{i=1}^{n} Z_i^2 \sim \mathcal{X}_n^2$$

and it holds that $\mathcal{X}_n^2$ =Gamma$(n/2, 2)$

rewrite this to show that is the square of the Z statistic

3.1.7 Fisher's Theorem

Let $(X_1, \ldots, X_n)$ be a s.r.s of r.v. that follows $\mathcal{N}(\mu, \sigma^2)$. Then,

$$\frac{(n-1)S_n'^2}{\sigma^2} = \sum_{i=1}^{n} (\frac{X_i - \hat{X}_n}{\sigma})^2 \sim \mathcal{X}_{n-1}^2$$

and

$$\hat{X}_n \text{ and } \mathcal{X}_n^2 \text{ are independent}$$

Dregrees of freedom remark

diapo 2.21

3.1.8 Student's Distribution

Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{X}_v^2$ be independent. The distribution of the r.v.

$$T = \frac{X}{\sqrt{Y/v}}$$

is a Student's t_v with v degrees of freedom

3.1.9 T statistic

Let $(X_1, \ldots, X_n)$ a s.r.s. of a r.v. that follows a $\mathcal{N}(\mu, \sigma^2)$. Then, a T statistic is,

$$T = \frac{\bar{X}_n - \mu}{S_n'^2/\sqrt{n}} \sim t_{n-1}$$

3.1.10 Snedecor's Distribution

Let $X_1 \sim \mathcal{X}^2_{v_1}$ and $X_2 \sim \mathcal{X}^2_{v_2}$ be independent. The distribution of the r.v.

$$F = \frac{X_1/v_1}{X_2/v_2}$$

is a Snedecor's F distribution with v_1 degrees of freedom in the numerator, and v_2 degrees of freedom in the denominator. It's denoted $\mathcal{F}_{v_1,v_2}$

add notes explaining that the snedecor's is ussefull when we have to deal with ratios

Note:

$$F_{1,v} = t^2_v$$

3.1.11 F statistic

Let $(X_1, \ldots, X_{n_1})$ be a s.r.s. of a $\mathcal{N}(\mu_1, \sigma^2_1)$ with sample variance S'^2_n. Let $(Y_1, \ldots, Y_{n_2})$ be a different s.r.s. of a $\mathcal{N}(\mu_2, \sigma^2_2)$ with sample variance S'^2_2. Let the first and the second distributions be independent. Then,

$$F = \frac{S'^2_1/\sigma^2_1}{S'^2_2/\sigma^2_2} \sim \mathcal{F}_{n_1-1,n_2-1}$$

3.2 Large Sample Inference

What happens when we can't ensure that the underlying distribution is a normal?

3.2.1 Convergence in distribution

The sequence of r.v. converges in distribution to the r.v. X if,

$$\lim_{x\to\infty} F_{X_n}(x) = F_X(x)$$

for all the points x where $F_X(x)$ is continuous. This is denoted as,

$$X_n \xrightarrow{d} X$$

3.2.2 Central Limit Theorem

explain much better this theorem

Let $X_1, \ldots, X_n$ i.i.d. r.v. with expectation $\mathbb{E}[X_i] = \mu$ and variance $V[X_i] = \sigma^2 < \infty$. Then, the c.d.f. of the r.v. converges to the c.d.f of a $\mathcal{N}(0,1)$ as long as $n \to \infty$. This is,

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Summary

$$(X_1, \ldots, X_n) \begin{cases} \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n) \\ \nsim \mathcal{N}(\mu, \sigma^2) \begin{cases} n < 30 \Rightarrow ? \\ n \geq 30 \Rightarrow \bar{X}_n \cong \mathcal{N}(\mu, \sigma^2/n) \end{cases} \end{cases}$$

In general, when inference, if the original distribution follows a normal and the sample size is small, we need to use the Student's t distribution. Otherwise, use the normal.

explain better

3.3 Properties of estimators

3.3.1 Biased and Unbiased

Given an estimator $\hat{\theta}$ the quantity

$$\mathbf{B}[\hat{\theta}] = \mathbb{E}[\hat{\theta}] - \theta$$

is the bias of the estimator. It is said that a estimator with zero bias is unbiased. $\mathbb{E}[\hat{\theta}] = \theta$

two explanatory dartboard

3.3.2 Estimation Error

Once observed the value of $\hat{\theta}$ of a sample the estimation error is the quantity $\hat{\theta}_n - \theta$. Note that the estimation error is usually different from zero. The bias is the mean estimation error.

$$\mathbf{B}[\hat{\theta}] = \mathbb{E}[\hat{\theta}_n] - \theta = \mathbb{E}[\hat{\theta}_n - \theta]$$

Error measures

There are lots of error measurements and each one is used for different implications. The most commons are:

- Mean Absolute Error: $\text{MSE} = \mathbb{E}[|\hat{\theta}_n - \theta|]$

  Note that
  $$\text{MSE} = \mathbf{B}^2(\hat{\theta}_n) + V(\hat{\theta}_n)$$

- Mean Square Error: $\text{MAE} = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$

Add more error measurements and his implications

Estimator selection

When we have to choose between two estimators $\hat{\theta}_{n,1}$ or $\hat{\theta}_{n,2}$

- If both are unbiased, choose the one with smallest variance.
- If at leas is biased, then choose the one with smallest error (e.g. MSE).

3.3.3 Relative Error

One problem that appears when measuring errors is that in general one error cant be directly compared with another one. One way to overcome this is to use a measurement that holds the same scale.

Coefficient of Variation

Given a estimator $\hat{\theta}$ with standard deviation $\sigma(\hat{\theta})$. The coefficient of variation or relative standard deviation (RSD) is defined as

$$\text{C.V.}(\hat{\theta}) = \frac{\sigma(\hat{\theta})}{\hat{\theta}}\text{x100} \qquad \text{if} \hat{\theta} \text{ is unbiased}$$

$$\text{RRMSE}(\hat{\theta}) = \text{C.V.}(\hat{\theta}) = \frac{\sqrt{\text{MSE}(\hat{\theta})}}{\hat{\theta}}\text{x100} \qquad \text{if } \hat{\theta} \text{ is biased}$$

3.3.4 Invariant

add more invariant properties

Translation Invariant

It is said that a estimator is translation invariant when it satisfies

$$\hat{\theta}(X_1 + c, \ldots, X_n + c) = \hat{\theta}(X_1, \ldots, X_n) + c, \quad \forall c \in \mathbb{R}$$

example

Scale Invariant

It is said that a estimator is scale invariant when it satisfies

$$\hat{\theta}(cX_1, \ldots, cX_n) = c\hat{\theta}(X_1, \ldots, X_n), \quad \forall c > 0$$

example

3.3.5 Consistency

The idea of consistency is related with the size of the sample. This property apply when the probability that the estimator $\hat{\theta}$ decay from the real θ when $n \to \infty$. More formally

add explanatory plot

$$P(|\bar{X}_n - \mu|) > 1 \quad \text{when} \quad n \to \infty$$

Weak Consistency

Let X be a r.v. with induced probability $P(\cdot) = P(\cdot\,;\theta)$. Let $(X_1, \ldots, X_n)$ be a s.r.s of X, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be an estimator of θ. The sequence $\{\hat{\theta}_n,\ n \in \mathbb{N}\}$ is consistent (or weak consistent, or consistent in probability) for θ, denoted $\hat{\theta} \xrightarrow{P} \theta$, if and only if,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0, \quad \forall \epsilon > 0$$

Consistency in squared mean

The following definition is stronger than the previous one,
A sequence of estimators $\{\hat{\theta}_n, n \in \mathbb{N}\}$ is consistent in squared mean for θ, denoted $\hat{\theta} \xrightarrow{sq.m.} \theta$, if it verifies

$$\lim_{n\to\infty} \mathrm{MSE}(\hat{\theta}_n) = 0$$

or equivalently,

$$\lim_{n\to\infty} \mathbf{B}(\hat{\theta}_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} V(\hat{\theta}_n) = 0$$

This theorem also states that,

$$\hat{\theta}_n \xrightarrow{sq.m.} \theta \Rightarrow \hat{\theta}_n \xrightarrow{P} \theta$$

3.4 Law of Large Numbers

The law of large numbers (LLN) is along with the central limit theorem the two key laws in probability. It states that when the estimator is consistent and the sample size is enough big, the sample mean equals the population mean. More formally,

Definition of LLN

Let $(X_1, \ldots, X_n)$ be a s.r.s. of a r.. X with mean μ and variance $\sigma^2 < \infty$. Then,

$$\bar{X}_n \xrightarrow{P} \mu$$

3.5 Algebra of Consistency

It is stated that any continuous transformation of a consistent estimator is consistent for the same transformation of the parameter. We can define formally in two ways,

Single estimator definition

Let $\hat{\theta} \xrightarrow{P} \theta$ and let $g(x)$ be a continuous at $x = \theta$. Then $g(\hat{\theta}_n) \xrightarrow{P} g(\theta)$

Multiple estimator definition

Let $\hat{\theta} \xrightarrow{P} \theta$ and $\hat{\theta}' \xrightarrow{P} \theta'$. Let $g(x, y)$ be a continuous at $(x, y) = (\theta, \theta')$. Then $g(\hat{\theta}_n, \hat{\theta}'_n) \xrightarrow{P} g(\theta, \theta')$

Corollary

From this two definitions we can summarize multiples algebras between estimators.

- $\hat{\theta}_n + \hat{\theta}'_n \xrightarrow{P} \theta + \theta'$
- $\hat{\theta}_n \hat{\theta}'_n \xrightarrow{P} \theta\theta'$
- $\hat{\theta}/\hat{\theta}'_n \xrightarrow{P} \theta/\theta' \quad \text{if } \theta' \neq 0$
- $\sqrt{\hat{\theta}_n} \xrightarrow{P} \sqrt{\theta} \quad \text{if } P(\hat{\theta}_n \geq 0) = 1$
- $a_n \hat{\theta}_n \xrightarrow{P} a\theta \quad \text{with} a_n \text{ being a sequence of constants.}$

3.5.1 Slutsky's Theorem

Let U_n and W_n be random sequences satisfying

$$U_n \xrightarrow{d} \mathcal{N}(0, 1), \qquad W_n \xrightarrow{P} 1$$

Then, it holds,

$$\frac{U_n}{W_n} \xrightarrow{d} \mathcal{N}(0, 1)$$

Add asymptotic normality of T stat example

3.6 Fisher's Information

Let X be a continuous r.v. with distribution that depends of $\theta \in \Theta \subset \mathbb{R}$. The Fisher's information of X over θ is defined as

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right]$$

for continuous distributions. For discrete distributions the definition is the same but replacing the p.d.f for the p.m.f (replacing $f(X;\theta)$ by $p(X;\theta)$)

Note that

$$\frac{\partial \log f(X;\theta)}{\partial \theta} = \frac{\frac{\partial f(X;\theta)}{\partial \theta}}{f(X;\theta)}$$

This quantity is the relative variation rate of f when varying θ, for a realized value x of X. It represents the information of x to discriminate θ from a near value $\theta + h$. Fisher's information is the mean information of the r.v. X about θ.

add better and conceptual explanation. Maybe an image too

3.6.1 Fisher's Information of a sample

The Fisher's information of a s.r.s $(X_q, \ldots, X_n)$ of a continuous r.v. X over θ is defined as

$$\mathcal{I}_n(\theta) = \mathbb{E}\left[\left(\frac{\partial \log f(X_1, \ldots, X_n;\theta)}{\partial \theta}\right)^2\right]$$

For X discrete just replace $f(\cdot;\theta)$ by the joint p.m.f. $p(\cdot;\theta)$

Fisher's equality

search and add information about this equality

Under certain regularity assumptions,

$$\mathcal{I}_n(\theta) = nI(\theta)$$

Frechet-Crámer-Rao lower bound

Check and write this assumptions

Under certain regularity assumptions, any unbiased estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ of θ satisfies

$$V(\hat{\theta}_n) \geq 1/\mathcal{I}_n(\theta)$$

3.6.2 Efficient estimator

An unbiased estimator $\hat{\theta}_n$ of θ that satisfies $V(\hat{\theta}_n) = 1/\mathcal{I}_n(\theta)$ is called efficient. If an estimator is efficient it means that it is the best possible estimator.

add example

3.7 Robust Estimators

search more info about robust estimators and complete this section

In real problems it is common that the data contains some contamination in form of measurements errors or other problems. In statistics it is said that an estimator $\hat{\theta}$ is robust if it preserves good properties (small bias and variance) even if the sample is contaminated.

The theory of statistical robustness is deep. For this, we are going to use this widely-used contamination model for $f(\cdot\,; \theta)$.

$$(1-\epsilon)f(x;\theta) + \epsilon g(x), \quad x \in \mathbb{R}, \text{ with } 0 < \epsilon < 0.5 \text{ and arbitrary p.d.f } g$$

3.7.1 Outliers

rewrite this whole part, because tis copy-paste from another book

The concept of outlier is intimately related with robustness. Outliers are "abnormal" observations in the sample that seem very unlikely for the assumed distribution model or are remarkably different from the rest of sample observations. Outliers can be originated by measurement errors, exceptional circumstances, changes in the data generating process, etc.

There are two main approaches for preventing outliers or contamination to undermine the estimation of θ:

1. Detect the outliers through a diagnosis of the model fit and re-estimate the model once the outliers have been removed.

2. Employ a robust estimator

The first approach is the traditional one and is still popular due to its simplicity. Besides, it allows us to employ non-robust efficient estimators that tend to be simpler to compute, provided the data has been cleared adequately. However, this procedure may quickly run into problems, since, for example, detecting outliers in higher dimensions is usually complicated and this detection may require manual inspection of the data.

In addition, robust estimators may be needed even when performing the first approach, as the following example illustrates. A simple rule to detect outliers in a normal population is to flag as outliers the observations that lie further away than 3σ from the mean μ, since those observations are highly extreme. Since their probability is 0.0027, we expect to flag as an outlier 1 out of 371 observations if the data comes from a perfectly valid normal population. However, applying this procedure entails estimating first μ and σ from the data. But the conventional estimators, sample mean and variance, are also very sensitive to outliers, and therefore their resulting values may hide the existence of outliers. Therefore, it is better to rely on a robust estimator, which brings us back to the second approach. As a consequence, it is sometimes preferred to employ robust estimators from the beginning.

The next definition introduces a simple measure of the robustness of an estimator.

3.7.2 Finite-sample breakdown point

The breakdown point of an estimator can be interpreted as the maximum fraction of the sample that can be changed without modifying the value of $\hat{\theta}$ to an arbitrary large value.

add better explanation, examples of robust estimators, trimmed mean, etc

3.8 Estimation methods

add introductory text

3.8.1 Method of moments

Population moments

Let's consider a population X with a unknown distribution that depends on K unknown parameters $\theta_1, \ldots, \theta_K$. The populations moments are functions of the unknown parameters, if they exists. Formally,

$$\alpha_r = \alpha_r(\theta_1, \ldots, \theta_K) = \mathbb{E}[X^r], \quad r = 1, 2, 3, \ldots$$

Sample moment

Given a s.r.s. of X, it's denoted by a_r to the sample moment of order r that estimate α_r

$$a_r = \bar{X}^r = \frac{1}{n}\sum_{i=1}^{n} X_i^r \qquad r = 1, 2, 3, \ldots$$

It's important to note that the sample moments do not depend of $\alpha_1, \ldots, \alpha_K$ but the population moment does.

Method of Moments

Let X be a r.v. that follow a unknown distribution of unknown parameters $\theta_1, \ldots, \theta_K$. The method of moments estimates this parameters by resolving a system of equations of shape

$$\alpha_r(\theta_1, \ldots, \theta_K) = a_r \quad r = 1, \ldots, R$$

where $R \geq K$ is the lowest integer such that the system admits a unique solution. This estimator of θ is denoted by $\hat{\theta}_{MM}$

add example

Consistency of moments estimators

In the moment estimator method, if $\mathbb{E}[(X^{r_k} - \alpha_{r_k})^2] < \infty$ and $\theta_k = g_k(\alpha_{r_1}, \ldots, \alpha_{r_K})$, with g_k continuous and $k = 1, \ldots, K$ then

$$\hat{\theta}_k = g_k(a_{r_1}, \ldots, a_{r_K}) \xrightarrow{P} \theta_k, \quad i = 1, \ldots, K$$

3.8.2 Maximum Likelihood Method

introductory example with unfair toss coin 0.8,0.2

This method relies in the believe that whatever happens in the reality is what is more probably to happen.

The Maximum Likelihood estimator (MLE) denoted by $\hat{\theta_{\text{MLE}}}$ of θ given the realized sample $(X_1 = x_1, \ldots, X_n = x_n)$ is the value $\hat{\theta}$ that maximizes the likelihood

$$L(\hat{\theta}; x_1, \ldots, x_n) = \max_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n)$$

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n)$$

Note that the maximum $\hat{\theta}$ of the likelihood is also the maximum of the loglikelihood

$$\ell(\hat{\theta}; x_1, \ldots, x_n) = \log L(\theta; x_1, \ldots, x_n)$$

This is because the logarithm is monotonously increasing.

It's important to remark that when calculating the MLE sometimes is more easy to just compute the logMLE.

Computation of MLE

In order to find that maximum likelihood, there are some remarks to take into account. If Θ is finite, the maximum can be found evaluating all possible values of $L(\theta; x_1, \ldots, x_n)$. If Θ is infinite and $L(\theta; x_1, \ldots, x_n)$ is differentiable respect θ, then the solutions of the likelihood equations can be calculated

$$\frac{\partial}{\partial \theta_k} L(\theta; x_1, \ldots, x_n) = 0, \quad k = 1, \ldots, K$$

Later, and because the equations can have multiples solutions (local minimas, and global minimas) we need to compare the solutions among themselves and with the boundaries values of Θ to find the global maximum.

example MLE under Normal

Properties

- The MLE of θ is not necessarily unbiased.
- The MLE of θ is not necessarily unique.
- If an unbiased and efficient estimator of θ exists, then that estimator is the unique MLE of θ
- The MLE is invariant with respect to one-to-one transformations of the parameter, that is, if $\omega = h(\theta)$ where h is one-to-one and $\hat{\theta}$ is the MLE of θ, then $\hat{\omega} = h(\hat{\theta})$ is the MLE of ω

Asymptotic efficency of MLEs

Let $f(x;\theta)$ be a p.d.f. (or p.m.f.) of a r.v. X, where $\theta \in \Theta$. Let Θ be a open interval from $\mathbb{R}$. Under certain regularity conditions, it holds that any sequence $\hat{\theta}_n$ of solutions of the likelihood equations that is consistent for θ verifies

check this conditions

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$$

This definition can be used to calculate the MLE when a parameter θ is unknown.

example

Chapter 4

Multivariate Analysis

4.1 Multidimensional Datasets

4.1.1 Graphs

For single quantitative variable

1. Barplot
2. Boxplot
3. Histogram
4. Kernel Density

add plots and explanation of how kernel denstity works (gaussian kernel)

For multiple quantitative variables:

1. Scatterplots (single 2D, single 3D or matrix)
2. Parallel Coordinate
3. Andrews plot

add example plots and clarification of (how works, when are usefull) for all of them

4.1.2 Descriptive Measurements

Some plots have lack of information when concerning the whole data. To overcome this, some measurements can be done.

For single variables:

- Sample Mean
- Sample Variance
- Sample Standard deviation
- Sample Covariance (with standardized variables or not)
- Sample Correlation (with standardized variables or not)

examples of correlation between two variables

For data matrices:

- Sample Mean Vector

$$\hat{x} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_p \end{bmatrix} = \frac{1}{n} X' 1_n \tag{4.1}$$

investigate this

- Sample Covariance Matrix (sample Covariance matrix of y?)
- Shrinkage sample covariance matrix

Eigenvalues of the covariance matrix analysis

write this

Shrinkage sample covariance matrix

$$\widetilde{S}_x = (1 - \lambda)S_x + \lambda I_p \tag{4.2}$$

write this

Chapter 5

Statistical Learning

5.1 Introduction - Supervised Learning Framework

rewrite all this using all the content of the slides but without being a scheme

The usual framework used in machine learning is DGP (data generating process?)

research about this

$$\text{Data} = \text{Model} + \text{Noise}$$

or in a statistical point of view, we assume $g(x)$ as the true model of the data

$$y = g(x_1, \ldots, x_k) + \text{Noise}$$

If we use this equation for a linear approximation

$$\text{Statistical approach: } y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$
$$\text{Machine Learning approach: } y = \text{map}(x_1, \ldots, x_p) + \epsilon$$

The two main categories of problems are named as, **classification and regression.**

5.1.1 Dimensionality Curse

write dimensionality curse problem

5.1.2 Prediction Error

In predictive models there are three sources of uncertainty

1. Estimation error: The error in the coefficients when the estimation is true.

2. Model Bias: The error in the linear approximation when the true model is non-linear, or contains other variables.

3. Irreducible Error: The noise in the data.

$$(\text{Prediction Error}^2) = \sigma^2 + \text{Bias}^2 + \text{Var}$$

Statistics

It focus on minimizing bias (by assuming knowledge of the population). By doing so, it is able to get formulas for the Var that provides explanation. The Var can be large in practice.

Machine Learning

It focus on minimizing $\text{Bias}^2 + \text{Var}$. No assumptions needed, hence no formulas and no direct explanation. Best predictive performance.

bias vs variance image

Overfitting

write this

example plot of a training showing the diverge between train and test loss

5.1.3 Classification Problems

write extensively about this

Some well-know statistical classification models are:

1. Logistic regression
2. Bayes classifiers
   (a) LDA
   (b) QDA
   (c) Naive Bayes
   (d) Shrinkage classification

Some well-know machine learning classification models are:

1. Nearest Neighbors
2. Neural Networks
3. Support Vector Machines
4. Decision Trees, Random Forest, Gradient Boosting

5.2 Probabilistic Learning

This approach aims to first predict the probabilities of a observation to belong to each of the categories. Later, select the most probable as the correct category.

They are two main families:

1. Bayes Classifiers
2. Logistic Regression

5.2.1 Bayes Classifiers

Given the predictors $y|x_1, \ldots, x_p$ focus on the conditional probability $p(y|x_1, \ldots, x_p)$.

In the bayes approach, this is modeled in an indirect way:

1. First model the predictors X separately for each given class y: $p(x|y)$
2. Second, apply the bayes formula to get $p(y|x)$

This approach is more stable than logistic regression when the classes are well separated. It performs well when the variables are gaussian distributed.

Classifier

Let $g = \{1, 2, \ldots, G\}$ a set of labels, let $\pi_g = P(y \in g)$ denote the prior probability of y to belong to g. Let $g : f_g(x)$ the multivariate distribution of predictors to model. Thus,

$$p_g(x) = P(y \in g|X = x) = \frac{f_g(x)\pi_g}{\sum_k f_k(x)\pi_k} \tag{5.1}$$

Finally, the maximum probability will be the ones to assign, $\max_g \quad f_g(x)\pi_g$.

> Note: The posterior probabilities are the same as the logistic regression ones.

The bayes classifiers are optimal in the sense they minimize the classification error rate. In practice π_g and $f_g(x)$ needs to be estimated, in order to do that using the proportion of training observations to belongs to class $g : \hat{\pi}_g \frac{n_g}{n}$. Whiout previous knoledge, $\hat{\pi_g} = 1/G$ is usually used.

To estimate $f_g(x)$ it is usually assumed to follow a multivariate normal, $f_g \sim \mathcal{N}(\mu_g, \Sigma_g)$. Using bayes rule we have,

$$\max_g \quad f_g(x)\pi_g = \min_g \quad (x - \mu_g)^T \Sigma_g^{-1}(x - \mu_g) + \log\det(\Sigma_g) - 2\log\pi_g \tag{5.2}$$

investigate about mahalanobis distance

This is called **quadratic discriminant analysis (QDA)**.

5.2.2 Quadratic discriminant analysis (QDA)

Chapter 6

Stochastic Processes

6.1 Introduction

A stochastic process $\boldsymbol{X} = \{X_t, t \in T\}$is a collection of random variables in the same probabilistic space $(\Omega, \boldsymbol{\mathcal{A}}, P)$. Thus, for each t in the index set T, X_t is random variable $X_t : \Omega \to S$, where S is the state space.

In this enviroment, an individual realization of the process is called a trajectory, $\boldsymbol{X}(\omega) = \{X_t(\omega), t \in T\}$.

- If T is countable, it is said to be a discrete-time stochastic process.
- If T is uncountable, it is said to be a continuous-time stochastic process.
- If S is countable, it is said to be a stochastic process with discrete state space.
- If S is uncountable, it is said to be a stochastic process with continuous state space.

6.1.1 Elements

- The mean function is $\mathbb{E}[X_t]$.
- The covariance function is $\gamma(s,t) = \mathrm{Cov}(X_s, X_t), \quad s,t \in T$

6.1.2 Stationary Stochastic Process

A s.p. is **weakly stationary** if:

- The expectation is constant, $\mathbb{E}[X_t] = \mathbb{E}[X_s], \forall s, t \in T$
- The covariance only depends on the lag, $\gamma(s,t) = \widetilde{\gamma}(t-s), \forall s, t \in T$

A s.p is **strong stationary** if $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ and $(X_{t_1+s}, X_{t_2+s}, \ldots, X_{t_n+s})$ are identically distributed.

A strong stationary process is also a weakly stationary one. The opposite is (not necessarily?) true.

6.2 Discrete-time Markov Chains

With S being a countable set, a discrete-time Markov chain is a sequence of random variables $\boldsymbol{X} = \{X_1, X_2, \ldots\}$ that takes values in S with the property

$$P(X_n = j | X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = i) = P(X_n = j | X_{n-1} = i), \quad \forall i, j, x_0, x_{n-2} \in S \tag{6.1}$$

This property is know as the **markov property**. The index set $T = 0, 1, 2, \ldots$.

Chapter 7

Regression Models

7.1 Introduction

Regression modeling is a set of statistical tools that aims to model associations rules between variables. This models can be later use to predict new observations, system explanation, variable screening or parameter estimation.

It's important to note the difference between correlation and regression. In correlation the relationship is not directional, so, it's interest is only on how they are mutually associated. In regression the interest come from how one variable respond to others.

7.2 Linear Regression

7.2.1 Linear Regression Models

Problems where there is only one independent variable (regressor or covariate) is called **simple linear regression model**. For these problems, the approach is to try to estimate the mean values of a variable respect to other $\mathbb{E}[Y|x]$, more formally

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{7.1}$$

where ϵ is an error term.

simple linear regression problem, ex: birth rate/poverty index

A regression model that contains more than one regressor variable is called **multiple linear regression model**. The model is similar to the simple one, they have in common a **response variable** Y, an **intercept** $beta_0$ and an **error term** ϵ. Contrary to the simple one, the multiple linear regression model has **more than one covariate or regressor** x_{ij}.

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \epsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \epsilon_n \end{aligned}$$

Matrix Formulation

The above model can be expressed in matrix equation (note the bold font):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{7.2}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ & & \vdots & \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{7.3}$$

Each column of $\boldsymbol{X}$ contains a particular covariate, is assumed to be know. $\boldsymbol{\beta}$ is the vector of unknown parameters to be estimated from the model and $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$ are random vectors whose elements are random variables.

7.2.2 Assumptions

These are the assumptions to be aware of when using linear regression models.

- Linearity: The mean of the response is a linear function of the predictors:

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_k = x_k] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- Independence: The errors ϵ_i are independent,

$$\text{Cov}(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$$

- Homocedasticity: The variance of the errors ϵ_i at each value of x_i is constant.

$$\text{Var}[\epsilon_i|\boldsymbol{X}_1] = x_1, \ldots, \boldsymbol{X}_k = x_k] = \sigma^2$$

- Normality: The errors of ϵ are normal distributed

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In summary

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}\sigma^2) \Rightarrow \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}\sigma^2)$$

7.2.3 Least Squares

This approach tries to minimize the residual sum of squares (RSS)

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

This approach, penalizes more the points that are further from the regression line. Also, it's computational less expensive than other approaches.

For the simple linear regression model, the solution for the least squares is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

where $\bar{X}$ and $\bar{Y}$ are the sample mean of X and Y. S_x^2 is the sample variance of X. S_{xy} is the sample covariate between X and Y.

The Least Squares method does not make any assumptions about the distribution of the response. Although this might be a good thing, it has the drawback that does not allow us to make any inference on the estimated parameters, and therefore, on the predictions.

Matrix Form

The least squares for multiple linear regression can be denoted

$$\text{RSS}(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{ik} + \cdots + \beta_k x_{ik}))^2 \tag{7.4}$$

or in matrix form

$$\text{RSS}(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{7.5}$$

Solution: In order to calculate the vector of β_i we calculate the derivate of the matrix

$$\frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X'Y} + 2\boldsymbol{X'X}\boldsymbol{\beta}$$

setting the equation to zero, we obtain

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X}^{-1}\boldsymbol{X'Y})$$

Once the parameters are estimated we obtain

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} = \boldsymbol{H}\boldsymbol{Y} \tag{7.6}$$

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \tag{7.7}$$

where $\boldsymbol{H}$ is called the *Hat Matrix* and what it does is to project $\boldsymbol{Y}$ into the regression hyperplane.

Properties

- The sum of residuals is zero.

$$\sum_{i=1}^{n} \hat{\epsilon}_i = \mathbf{1}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = 0$$

- The sum of observed data is equal to the sum of fitted values,

$$\sum_{i=0}^{n} Y_i = \sum_{i=0}^{n} \hat{Y}_i = \mathbf{1}'\hat{\boldsymbol{Y}}$$

- The residuals are orthogonal to the predictors

$$\sum_{i=0}^{n} x_i \hat{\epsilon}_i = \boldsymbol{X}'\hat{\boldsymbol{\epsilon}} = 0$$

- The residuals are orthogonal to the fitted values

$$\sum_{i=0}^{n} \hat{y}_i \hat{\epsilon}_i = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{\epsilon}} = 0$$

add complete subsection of QR decomposition least squares

7.2.4 Maximum Likelihood

In maximum likelihood estimation, we search over all possible sets of parameter values for a specified model to find the set of values for which the observed sample was most likely. That is, we find the set of parameter values that, given a model, were most likely to have given us the data that we have in hand.

$$\ln \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma^2}|\boldsymbol{X}) = \cdots \propto -\frac{n}{2}\ln(\sigma^2) - \frac{(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})}{2\sigma^2} \tag{7.8}$$

Minimizing $\ln \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma^2}|\boldsymbol{X})$ is equivalent to minimizing least squares because the only element of the equation that depends on $\boldsymbol{\beta}$ is $(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})$.

Estimation of σ^2

Using ML we can estimate σ^2. The unbiased estimator is,

$$\hat{\sigma}^2 = \frac{(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})}{n - (k+1)} \tag{7.9}$$

where $k + 1$ is the number of estimated parameters.

Degrees of freedom

The degrees of freedom is the number of independent pieces of information in a sample. Here, since we use $k+1$ parameters to estimate, the number of remaining parameters we can use to estimate σ^2 will be $n-(k+1)$.

Interpretation of the coefficients

The interpretation of a multivariate regression coefficient is the expected change in the response per unit change in the regressor, holding all of the other regressors fixed. The latter part of the phrase is important,by holding the other regressors constant, we are investigating an adjusted effect.

7.2.5 Inference of model parameters

Sampling distribution of estimated coefficients

$$\mathcal{E}[\hat{\beta}] = \boldsymbol{\beta} \tag{7.10}$$

$$\text{Var}[\hat{\beta}] = \sigma^2(\boldsymbol{X'X})^{-1} \tag{7.11}$$

We can conclude that, the smaller the variance of the error, the larger the sample size and the variability of the predictor, the more precise the estimates are.

Confidence intervals

It is given by,

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_{jj}}} \sim t_{n-k-1} \tag{7.12}$$

thus,

$$CI_{(1-\alpha)}\left(\beta_j\right) = \left[\hat{\beta}_j - t_{1-\alpha/2}; n-k-1\hat{\sigma}\sqrt{v_{jj}}, \quad \hat{\beta}_j + t_{1-\alpha/2;n-k-1}\hat{\sigma}\sqrt{v_{jj}}\right] \tag{7.13}$$

where v_{ii} is the i-th element of the diagonal of $\boldsymbol{(X'X)^{-1}}$.

Hypothesis Test

$$H_0 : \beta_j = 0 \qquad H_1 : \beta_j \neq 0 \tag{7.14}$$

In fact, what we are testing is whether the relationship between $\boldsymbol{X}_i$ and $\boldsymbol{Y}$ is linear or not.

7.2.6 ANOVA

Anova test the variance of the estimators and the data. A way to think about regression is in the decomposition of variability of our response.

The total variability in our response is the variability around an intercept. This is also the variance estimate from a model with only an intercept:

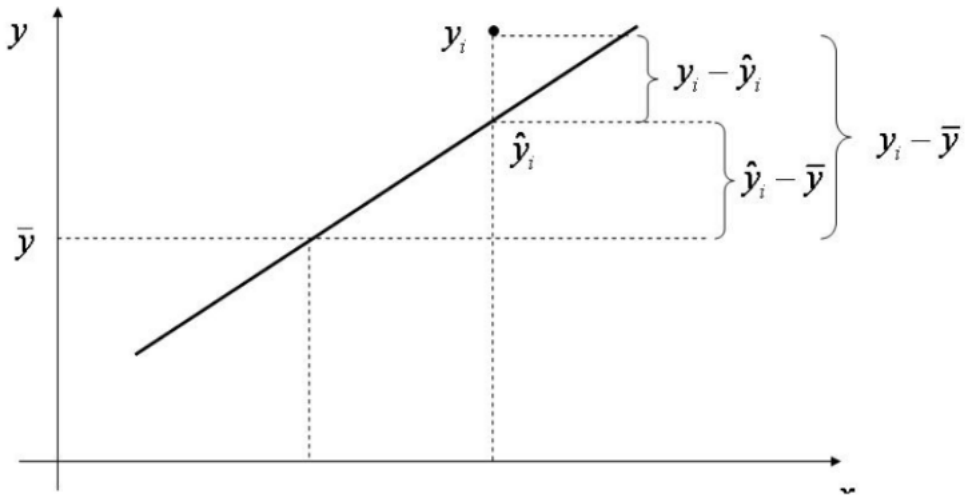
$$\text{Total variability} = \text{SST} = \sum_{i=1}^{n} (Y_i - \bar{\boldsymbol{Y}})^2 \tag{7.15}$$

The regression variability is the variability that is explained by adding the predictors. The regression variability will be small if the predicted line is almost plane, and large if the slope of the estimator is large.

$$\text{Regression variability} = \text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{\boldsymbol{Y}})^2 \tag{7.16}$$

The residual variability is what is leftover around the regression line. This value will be large if the points are far from the predicted line and small if they are close.

$$\text{Residual variability} = \text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{\boldsymbol{Y}}_{\boldsymbol{i}})^2 \tag{7.17}$$



$$\underbrace{\sum_{i=1}^{n} \left(Y_i - \overline{\mathbf{Y}}\right)^2}_{SST} = \underbrace{\sum_{i=1}^{n} \left(\hat{Y}_i - \overline{\mathbf{Y}}\right)^2}_{SSR} + \underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}_{SSE}$$

complete this

Chapter 8

Numerical Methods

This chapter will focus on the methods and processes needed to choose the optimal solution given a mathematical model. This chapter also include the processes needed to build these models.

This field of study is also know as **Prescriptive Analysis** or **Operations Research**.

The schematic way to proceed in this problems is the following:

Problem Description → Model Formulation → Analysis & Algorithms → Computer Solution → Interpretation

8.1 Notation

8.1.1 Elements

The elements involved in all mathematical decision models are the following:

- Decision Variables: Are the ones we want to know his optimal value. (e.g. number of products to build in a period of time)

$$x = (x_1, x_2, \ldots, x_n)$$

- Objective Function: Is the one that model the problem. It will be always needed to *maximize* or *minimize* it.

$$\text{maximize } f(x) \text{ or minimze } f(x)$$

- Functional Constraints (or structural constrains): Are the limitations to the objective function and is a function of all the variables.

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \leq b_1$$

- Non-negativity Constraints: Are the limitations to the objective function.

$$g_1(x) \leq b_1, \;\; g_2(x) \geq b_2, \;\; b_3 = b_3$$

The main goal to a correct model building is to identify these elements and define it in a proper way. Contrary that the logic intuition could make us think, simplest models are better than complex one with lot of constraints and elaborated objective functions.

All models are wrong, but some are useful (? citation)

Why use this models? Because most of the times the time and memory computational complexity to resolve some problems in a brute force or in a non analytical way is way greater than the models developed using this approach.

8.1.2 Standard Form of the Model

A problem is described in the standard form if it is described in the following way:

$$\text{maximize } c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

Subject to the restrictions

$$\begin{gathered} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \leq b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \leq b_2 \\ \cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \leq b_m \end{gathered}$$

and

$$x_1 \geq 0, \quad x_2 \geq 0, \quad \ldots \quad x_n \geq 0$$

8.1.3 Other Forms

- Minimizing rather than maximizing the objective function.

$$\text{minimize } c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

- Functional constrains with a greather-or-equal inequality

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \geq b_i$$

- Functional constrains in equation form

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i$$

- Deleting the non-negativity constrains

$$x_j \text{is unrestricted in sign for some values of } j$$

8.1.4 Terminology for solutions of the model

- **Feasible solution** is a solution that satisfies all the constrains. It is possible for a problem to have no feasible solutions.

- **Infeasible solution** is a solution with at least one constrains unsatisfied.

- **Feasible region** is the set of possibles solutions of the model. It's always a convex polytope. This polytope is defined by the intersection of the problem constraints

- **Optimal solution** of the problem. Is whose that *maximize* or *minimize* the objective function inside the feasible region. Usually, this optimal solution is find computationally. It is possible for a problem to have multiples optimal solutions, also is possible to do not have them.

- A **Corner-point feasible (CPF)** solution is a solution that lies at a corner of the feasible region.

Realtionship between optimal solutions and CPF solutions

Consider any linear pro-gramming problem with feasible solutions and a bounded feasible region. The problem must possess CPF solutions and at least one optimal solution. Furthermore, the best CPF solution must be an optimal solution. Thus, if a problem has exactly one optimal solution, it must be a CPF solution. If the problem has multiple optimal solutions, at least two must be CPF solutions.

8.2 Linear Optimization (LO) Models

Most important type of decision optimization models. Is the foundation to understand the complex ones. Also is one of the most widely applied models because of his simplicity.

Limitations

- Decision variables needs to be continuous.

$$x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}$$

- Objective function needs to be linear in x. Called .

$$f(x_1, x_2, \ldots, x_n) = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

- Constraints needs to be linear in x

$$\begin{gathered} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \leq b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \leq b_2 \\ \ldots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \leq b_m \end{gathered}$$

- Decision Variables needs to be non-negative

$$x_1 \geq 0, \quad x_2 \geq 0, \quad \ldots \quad x_n \geq 0$$

An equation in a form called **Slope-intercept form** is the one in the form $x_n = ax_i + bZ, \ \ \forall a, b \in \mathbb{R}$ and demonstrate that the slope of the line is a. This means that an increase of one value in x_n implies an increment of a in x_i. Whereas, the intercept of the line with the x_n axis is bZ.

Assumptions of a LO problem

- Proportionally, realted with linearity of the objective functions and constrains.
- Additivity, Every function is the sum of the individual contributions of the respectives activities.
- Divisibility, decision variables are in $\mathbb{R}$
- Certainty, the value assigned to each parameter is assumed to be a known constant. This allow us to perform sensitivity analysis.

8.2.1 Duality

Every linear programming problem has associated with it another linear programming problem called the **dual**.

Wikipedia
The duality principle is the principle that optimization problems may be viewed from either of two perspectives, the **primal problem** or the **dual problem**. The solution to the dual problem provides a lower bound to the solution of the primal (minimization) problem. However in general the optimal values of the primal and dual problems need not be equal. Their difference is called the duality gap. For convex optimization problems, the duality gap is zero under a constraint qualification condition.

These conditions are:

- Non-negative variables

$$x = (x_1^+, x_2^+, \ldots, x_n^+) \in \mathbb{R}$$

- The objective functions need to be maximize

$$\text{maximize } c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

- All constrains must be equality (non inecuations)

$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m$$

- The right-hand side values must be non-negatives

$$b_m > 0, \quad \forall m \in \{1, 2, \ldots, m\}$$

If some of this conditions are not satisfied, we can convert the equations to make it satisfied.

For constrains that are non-equality

Given, $x_1 \leq 4$ we can convert it to a new variable. Just,

$$x_1 \leq 4 \equiv x_3 = 4 - x_1$$

This new variable x_3 is called **slack or surplus variable** depending of what is doing. So for a problem,

Original Problem	Augmented Problem
$\text{maximize } Z = 3x_1 + 5x_2$	$\text{maximize } Z = 3x_1 + 5x_2$
subject to	subject to
$x_1 \leq 4$	$x_1 + x_3 = 4$
$2x_2 \leq 12$	$2x_2 + x_4 = 12$

For variables with unrestricted sign

write example and explanation

Formulation

Primal Problem	Dual Problem
$$\text{maximize } Z = \sum_{j=1}^{n} c_j x_j$$	$$\text{minimize } W = \sum_{i=1}^{m} b_i y_i$$
subject to	subject to
$$\sum_{j=1}^{n} a_{ij} x_j \leq b_i, \quad \forall i \in \{1, 2, \ldots, m\}$$	$$\sum_{i=1}^{m} a_{ij} y_i \leq c_j, \quad \forall j \in \{1, 2, \ldots, m\}$$
and	and
$$x_j \geq 0, \quad \forall j \in \{1, 2, \ldots, n\}$$	$$y_i \geq 0, \quad \forall i \in \{1, 2, \ldots, n\}$$

Properties

- **Weak duality**: If x is a feasible solution for the (P) and y is a feasible solution for (D), then
$$cx \leq yb$$
- **Strong duality**: If x^* is an optimal solution for (P) and y^* is an optimal solution for (D), then,
$$cx^* = y^* b$$

Thus, these two properties imply that $cx < yb$ for feasible solutions if one or both of them are not optimal for their respective problems, whereas equality holds when both are optimal.

- **Symmetry property**: For any primal problem and its dual problem, all relationships between them must be symmetric because the dual of this dual problem is this primal problem.

Relation between primal and dual problem

The following are the only possible relationships between the two problems.

1. If one problem has feasible solutions and a bounded objective function (and so has an optimal solution), then so does the other problem, so both the weak and strong duality properties are applicable.

2. If one problem has feasible solutions and an unbounded objective function (and so no optimal solution), then the other problem has no feasible solutions.

3. If one problem has no feasible solutions, then the other problem has either no feasible solutions or an unbounded objective function.

Complementary Basic Solutions

Because the problem (D) is also a LO problem, it also has a corner point solution. Ussing the augmented form of the problem we can express these corner-point solution as a basic solution. Because the functional constrains have $\geq$, this solution is obtained by subtracting the surplus (rather than adding the slack) from the Left-HS of each constraint j. Thus,

$$z_j - c_j = \sum_{i=1}^{m} a_{ij} y_i - c_j, \quad \forall j \in \{1, 2, \ldots, n\}$$

So $z_j - c_j$ is the surplus variable for constraint j (or its slack variable if the constraint is multiplied through by -1).

One of the important relationships between the primal and dual problems is a direct cor- respondence between their basic solutions.

- **Complementary basic solutions property**: Each basic solution in the primal problem has a **complementary basic solution** in the dual problem, where their respective objective function values (Z and W) are equal.

- **Complementary slackness property**: CHECK THIS and addapt to the defition given here

$$(\sum_{i=1}^{m} a_{ij} \pi_i - r_j) \bar{x}_j = 0, \quad \forall j \in \{1, 2, \ldots, n\}$$

8.2.2 Sensitivity Analysis

Once we already have the optimal solution for a LO problem, it would be important to analyze if slight changes in the constraints or the objective

function affects the optimal solution or it remains the same. Specifically, we would want to analyse and find the **parameter bounds** of the optimal solution. We are going to analyze two types of changes:

1. Changes in objective coefficients (changes in c_j values)
2. Changes in the right-hand side of the constraints (changes in b_i values)

Changes in the coefficients of nonbasic variables

Because the variable involved is nonbasic (value of zero), changing its coefficients cannot affect the feasibility of the solution. Therefore, the open question in this case is whether it is still optimal.Since these changes affect the dual problem by changing only one constraint, this question can be answered simply by checking whether this complementary basic solution still satisfies this revised constraint.

General Procedure

Model Revision → Final tableau revision → Gaussian elimination → Feasibility text → Optimality tes

Changes in objective coefficients

Simplified example.
Given an objective function (o.f.),

$$\text{maximize } z(x_1, x_2) = 5x_1 + 4x_2$$

Subject to

$$\text{M1: } 6x_1 + 4x_2 \leq 24$$

$$\text{M2: } x_1 + 2x_2 \leq 6$$

with optimal solution in $x^* = (3, 3/2)$, we can perform the sensitivity analysis if the changes in the o.f. are expressed as,

$$\text{maximize } \hat{z}(x_1, x_2) = (5 + \Delta r_1)x_1 + (4 + \Delta r_2)x_2$$

Solution:
We know that $x^* = (3, 3/2)$ for $(\Delta r_1, \Delta r2) = 0$, but, for which increments of $(\Delta r_1, \Delta r2)$ is optimal x^*? x^* is optimal only and only if

$$\text{slope of M1 constraint} \leq \text{slope of } \hat{z} \leq \text{slope of M2 constraint}$$

$$-\frac{3}{2} \leq \frac{5+\Delta r_1}{4+\Delta r_2} \leq -\frac{1}{2}$$

Note: We can draw the region of solution if we represent the increments as axis (include image)

Changes in the RHS of the constraints

Simplified example.
Given constraints,
M1: $6x_1 + 4x_2 = 24$
M2: $x_1 + 2x_2 = 6$
We can repeat the same approach
M1: $6 + x_1 + 4x_2 = 24 + \Delta b_1$
M2: $x_1 + 2x_2 = 6 + \Delta b_2$

Shadow prize and range of validity Shadow price is change of the objective function per unit in the RHS
Range of validity is the range of values of the RHS in which the solution remains optimal

Appendix A

Appendix