Probability and Statistics Notes

Yusif Mehdiyev

June 2024

Preface

This thing is abandoned now, I wanted it to be a notebook to prove I have actually self-studied the topic but this is really pointless (I am so naive and stupid!).

Please be warned that this PDF probably contains plagiarilism from the books I have read, this is a failed personal project anyways, do not take it seriously.

Source

- Larry Wasserman - All of Statistics - A Concise Course in Statistical Inference.
- Probability and Statistics by Morris H. DeGroot and Mark J. Schervish.
- R for Data Science.
- An Introduction To Statistical Learning with R/Python (ISLR/ISLP)
- MIT OpenCourseWare Introduction to Probability and Statistics, spring 2022.
- Countless Wikipedia Articles.

Contents

Part I

Probability

1 Probability

§1.1 Set Theory

Set Theory is well discussed and studied theory across the topics of the mathematics, a great amoount of books across the field starts with the study of the sets. Therefore, I won't really write any notes about this here other than the important definitions.

Definition 1.1.1. Event and Sample Space **Sample Space**, as name suggests, is the set of all the possible outcomes of an experiment. The events are subsets of the sample space.

§1.2 Definition of Probability

Although there are multiple interpretations of probability, the soon-to-be-defined axioms rigorously define the explain for further exploration.

Definition 1.2.1. Probability Function Let A be a event such that $A \subset \Omega$. Then, **Probability of A**, written as $P(A)$, is a function such that

1. $1 \geq P(a) \geq 0$

2. $P(\Omega) = 1$

3. If $\{A_i\}$ are disjoint events of Ω, then $P(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$

First and Second axioms are self-explainatory. Third axiom states the probabilities of two events are independent if the events themselves are independent.

Lemma 1.2.2

For events A and B,

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

Proof. We can rewrite $A \cup B$ as union of $A \setminus B$, $B \setminus A$, and $A \cap B$, since these are the slices of the thing we want to begin with. Moreover, these slices are disjoint, therefore we can apply our third axiom (P is additive):

$$\begin{aligned} P\left(A \cup B\right) &= P\big((A \setminus B) \cup (B \setminus A) \cup (A \cap B)\big) \\ &= P(A \setminus B) + P(B \setminus A) + P(A \cap B) \\ &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) + P(A \cap B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

□

§1.3 Independent Events

Definition 1.3.1. Two events A and B are **independent** if

$$P(AB) = P(A)P(B)$$

Example 1.3.2

Let $A = \{2,4,6\}$, $B = \{1,2,3,4\}$. Since $P(A)P(B) = P(AB)$, they are independent.

Example 1.3.3

Let $A = \{2,4,6\}$, $B = \{2,4,5\}$. Since $P(A)P(B) \neq P(AB)$, they are dependent.

Note that even though $AB \neq \emptyset$, in above examples, the result is not the same. The independence merely shows that another event can't change other event's probability, even though intuitively it makes no sense.

§1.4 Conditional Probability

Conditional Probability, as the name implies, is the probability of an event with a condition. More precisely, **Conditional Probability** is the probability of an event A, given that another event B is already occurred. In such probability, the sample space is reduced to B's, while we want to find probability of A from B's space (Which increases of probability of A, since sample space is also reduced). We can show this neatly in venn diagram:

Example 1.4.1

If we tossed a six sided dice one time, and we rolled an even number B, what is the probability of getting number 2, event A?

Since the first toss' result is already happened, we know that $\Omega_{reduced} = \{2,4,6\}$ and $A = \{2\}$, then $P(A)_{\Omega_{reduced}} = \frac{1}{3}$.

If there wasn't any condition, the probability of getting 2 would be $\frac{1}{6}$. Simply, in a simple probability we defined a new condition and sort of updated our measurement to $\frac{1}{3}$. This is an important idea in Probability and Statistics, which we will revisit shortly in **Bayes' Rule**

Definition 1.4.2 (Conditional Probability)**.** We can show the conditional probability of A given B as:

$$P(A|B) = \frac{P(AB)}{P(B)} \qquad \text{for } P(B) \neq 0$$

Note that $P(A|B) \neq P(B|A)$.

§1.5 Bayes' Theorem

Theorem 1.5.1 (Simplified Bayes' Theorem)

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

Proof. We apply the definition of conditional probability twice:

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

Using above properties directly gives our theorem. □

Theorem 1.5.2 (Law of Total Probability)

Let $A_1, A_2, ..., A_n$ be partition of Ω. Then for any event B,

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

Proof. Let $C_i = A_iB$. Then we know that $C_1, C_2, ..., C_n$ are the partition of B. Therefore using the partition property,

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

□

This theorem becomes very handy in practical situations. Moreover, with the help of this theorem we can generalize our Bayes' Theorem,

Theorem 1.5.3 (Bayes' Theorem)

Let $A_1, A_2, .., A_n$ be a partition of Ω such that $P(A_i) > 0$. For $P(B) \neq 0$ and for any $i = 1, 2, ..., n$,

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

2 Random Variables

“A random variable can be compared with the holy roman empire: The Holy Roman Empire was not holy, it was not roman, and it was not an empire.”

§2.1 Introduction to Random Variables

A Random Variable describes the data or the outcome ω as a real number.

Definition 2.1.1 (Random Variables). A **Random Variable X** is a function,

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

That assigns a real number $X(\omega)$ to each outcome ω.

Example 2.1.2

Flip a coin. We know that $\Omega = \{H, T\}$. A r.v X might assign $X(H) = 1$ and $X(T) = 0$. That is, heads is “coded” as 1 and tails is “coded” as 0.

Example 2.1.3

Flip a fair coin n times. Let X represent the number of heads we get. Then, X is a random variable that takes values $\{0, 1, 2, ..., n\}$.

Example 2.1.4

Toss a fair six sided dice 2 times. Let X be the sum of the two rolls we get. Then, X is a random variable that takes values $\{2, 3, 4, ..., 12\}$.

Example 2.1.5

A students wants to write a real number in intervals $[0, 1]$. Let X be the number student writes. Then, X is also a random variable that takes any real numbers in that interval.

Random Variables also have *Independence, Conditional Random Variable, a probability function* and so on. Additionally , Random Variables can be either **Discrete** or **Continous**.

Discrete Random Variable’s range is finite or countably infinite. The first two examples we gave are Discrete. Continuous Random Variables’s range is uncountably infinite like the third example

§2.2 Distribution Functions CDF, PMF, PDF, PPF

CDF and PPF

We define **Cumulative Distribution Function** as,

Definition 2.2.1. CDF The **Cumulative Distribution Function** or shortly **CDF** is a function $F_X : \mathbb{R} \to [0,1]$ such that

$$F_X(x) = P(X \leq x)$$

Example 2.2.2

We toss a fair coin two times. Let X represent the number of heads we get. Then CDF of X is,

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

Now, let's look at some properties of CDF,

Theorem 2.2.3

Let X have CDF F, and Y have CDF G. If $F(x) = G(x)$ for all x, then,

$$P(X \in A) = P(Y \in A) \quad \text{for all } A$$

Theorem 2.2.4

the function $F : \mathbb{R} \to [0,1]$ is a CDF for some r.v if and only if F satisfies three conditions:

1. F is non-decreasing
2. F is normalized i.e

$$\lim_{x \to -\infty} F(x) = 0 \quad \wedge \quad \lim_{x \to \infty} F(x) = 1$$

3. F is right continuous.

Definition 2.2.5. **Quantile Percent point function**, or shortly **PPF**, is defined as inverse of CDF i.e,

$$Q(x) = F^{-1}(x)$$

PMF and PDF

Similar to probabilities of Events, we can calculate probability of X, depending on discrete or Continuous with functions called **Probability Mass Function** and **Probability Density Function**, shortly PMF and PDF respectively.

Definition 2.2.6. If X is discrete, and it takes *countably* values $\{x_1, x_2, .., x_n\}$ we define **Probability Mass Function** of X as follows:

$$f_X(x) = P(X = x)$$

Note that $P(X = x)$ is a function, not a number. We have to specify x first to get a number.

With the properties of probability, we have $f_X \geq 0$ for all $x \in \mathbb{R}$ and $\sum_i f_X(x_i) = 1$.

Example 2.2.7

We toss a fair coin two times. Let X represent the number of heads we get. Then CDF of X is,

$$f_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Moreover, for any set of real numbers, S, we have

$$P(X \in S) = \sum_{x \in S} f_X(x)$$

Since all $\{X = x\}$ are disjoint.

We can apply similar rules to continuous r.vs,

Definition 2.2.8. If X is continuous, we can represent the probability distribution of X with,

$$P(a < X < b) = \int_a^b f_X(x)dx$$

Function f_X is called **Probability Density Function** or PDF as shortly.

Definition 2.2.9. CDF is related to PMF and PDF. For discrete r.v,

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

And for continuous r.vs,

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

And $f_X(x) = F_X^{'}(x)$ for for all differentiable points x.

§2.3 Important Random Variables and their distribution

Definition 2.3.1. If X has distribution A, we write

$$X \sim A$$

Usually A depends on some fixed numbers to define properly, we call them **parameters**. For example, the distribution **Bernoulli**(p) has parameter p. We show parameters in PMF and CDF as,

$$f(x; parameters) \quad \text{and} \quad F(x; parameters)$$

There are some specific examples of r.v. that are very useful in practical applications. We will show most important ones, and briefly discuss them. In later chapters, we will learn more about them. Note that we will write the notation with the name of the distribution.

Degenerate distribution or Point mass distribution: $X \sim \delta_a$

Consider tossing coin or dice where all the sides show the same value. The PMF is ,

$$f_X(x; \delta_a) = 1 \quad \text{for } x = a$$

Discrete Uniform distribution

This distribution is the one of the most known ones. When there are finitely many values and each of them have the same probability, then $p = \frac{1}{n}$. Simple coin tossing, dice rolling are prime example of these. The PMF is,

$$f_X(x) = \frac{1}{n}$$

Where $x \in \{1, 2, ..., n\}$. Nothing new here. for other cases, $f_X(x) = 0$.

Bernoulli distribution

Bernoulli(p) describes "Yes or No" type of experiments such as coin flipping. Therefore, $P(X = 1) = p$, $P(X = 0) = 1 - p$. We can also calculate PMF,

$$f_X(x; p) = p^x(1 - p)^{1-x} \quad \text{for } x \in \{0, 1\}$$

Binomial distribution

Binomial(n, p) is the generalized form of Bernoulli distribution. Similar to Bernoulli, this distribution describes "Yes or no" type of experiments, but for n times of tries e.g tossing a coin n times. Assuming tries are independent of each other, we can show PMF as,

$$f_X(x; n, p) = \binom{n}{x} p^x(1 - p)^{n-x} \quad \text{for } x \in \{0, 1, ..., n\}$$

Notice that Binomial$(1, p)$ = Bernoulli(p).

Geometric distribution

Geom(p) is also specIID with Bernoulli. The geometric distribution describes the probability of the first occurrence of success requires after x independent trials e.g getting the first head after x tosses. The PMF is,

$$f_X(x; p) = (1 - p)^{x-1}p \quad \text{for } x \geq 1$$

Poisson distribution

Poisson(λ) is mainly used for counts of events like photons hitting a detector in a time interval, number of car accidents, students achieving a low and high mark on exam, or number of pieces of chewing gum on a tile of a sidewalk. its PMF is,

$$f_X(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{for } x \geq 0$$

Usually, $\lambda = rt$ where r is average rate the events occur and t is the time interval. The r.v X represents the number of events.

Unfiorm distribution

The PDF of $X \sim \text{Uniform}(a, b)$ is defined as,

$$f_X(x;a,b) = \frac{1}{b-a} \quad \text{for } x \in [a,b]$$

Normal (Gaussian) distribution: $X \sim N(\mu, \sigma^2)$

PDF is defined as,

$$f_X(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } x \in \mathbb{R}$$

We will learn about μ and σ in later chapters.

We define **standart Normal distribution** as $N(0, 1)$, r.v as Z. This specific distribution is very important, so much that we show its PDF and CDF with new notation, namely $\phi(z)$ and $\Phi(z)$ respectively. There is no closed form expression for $\Phi(z)$. It can be shown that we can show any normal probabilities we want with $\Phi(z)$.

Exponential distribution

$X \sim \text{Exp}(\lambda)$ distribution is continous analogue of the geometric distribution. We define its CDF as ,

$$f_X(x;\lambda) = \lambda e^{-x\lambda}$$

Gamma distribution

$X \sim \text{Gamma}(\alpha, \beta)$, we start with,

Definition 2.3.2. For $\alpha > 0$, we define **Gamma function** as,

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$$

It is generalized form of more simple and specific version,

$$\Gamma(n) = (n-1)! \quad n \in \mathbb{N}$$

We define PDF of x as,

$$f_X(x;\alpha,\beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta} \quad x > 0$$

Notice that $\text{Gamma}(1, \beta) = \text{Exp}(\beta)$. That is, exponentional distribution is a specific case of gamma distribution.

§2.4 Multivariate Distributions

Definition 2.4.1. Let $X_1, X_2, .., X_n$ be r.vs. We call $X = \{X_1, X_2, ..., X_n\}$ **a random vector**.

Definition 2.4.2. If r.vs $X_1, X_2, ..., X_n$ are independent an have the same marginal distribution with CDF F, we define these r.vs as **independent and identically distributed**, shortly IID, with notation,

$$X_1, ..., X_n \sim F$$

similarly, we show the PDF the same way. IID property is very important in statistical field.

We can apply multivariate CDF as

Definition 2.4.3. For n r.v $\{X_1, X_2, .., X_n\}$, the multivariate CDF $F_{X_1,X_2,...,X_n}$ is given by,

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n)$$

There is nothing fancy here, actually. We simply redefine CDF in general sense for n r.vs.

Similarly, we can define multivariate PMF as,

Definition 2.4.4. For random vector X,, the multivariate PMF $f_{X_1,X_2,...,X_n}$ is given by,

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = P(X_1 = x_1, ..., \ X_n = x_n)$$

This is generalized form of PMF

Similarly, we define,

Definition 2.4.5. We know that CDF and PDF are related by derivative. Then, For random vector X, the multivariate PDF $f_{X_1,..,x_N}$ is given by,

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \frac{\partial^n F_{X_1,..,X_n}(x_1, ..., x_n)}{\partial x_1 \partial x_2 ... \partial x_n}$$

The Properties and theorems are similar, but are generalized for n r.vs.

§2.5 Marginal Distribution

If more than one variable is defined in an experiment, it is important to distinguish between the multivariate probability of $(X_1, X_2, .., X_n)$ and individual probability distributions of $X_1, X_2, .., X_n$

Formally, **Marginal distribution is the probability of a single event (or r.v) occuring, independent of other events**. Therefore implementing marginal distributions are rather easy. In multivariate distributions, we redefine the needed variable as a "constant" and work with other variables only.

Definition 2.5.1. If X is a random vector with PMF $f_{X_1,X_2,..,X_n}$, then we define marginal distribution as,

$$f_{X_1} = P(X_1 = x_1) = \sum_{x_1 \ constant} P(X_1 = x_1, .., X_n = x_n) = \sum_{x_1 \ constant} f_{X_1,X_2,..,X_n}(x_1, x_2, ..., x_n)$$

Similarly,

Definition 2.5.2. We define marginal PDF as ,

$$f_{X_i}(x_i) = \int\int\int ... \int f(x_1, x_2, .., x_n) dx_1..dx_{i-1}dx_{i+1}...dx_n$$

Similarly, CDF follows the same rule. $F_X(x) = F(x, a, b, c, ...)$.
Remark: Marginality and conditionality are not the same thing. They look similiar, but their definitions are subtly different.

§2.6 Independence

Similar to events, r.vs also can be independent,

Definition 2.6.1. Two r.vs X and Y are **independent** if, for every A and B,

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

The definition persists for multivariate distributions.

To check Independence, we need to check the above question for every subsets A, B. Additionally, we have the theorem,

Theorem 2.6.2

Let X and Y have PMF f_{X_y}. THen X and Y are independent only and only if ,

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

The definiton persists for multivariate distributions.

§2.7 Conditioning

Similar to events, r.v X can also have conditional distributions given that we have $Y = y$. We show the conditionality with,

Definition 2.7.1. We can show conditional distribution of X respect to Y with,

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

Moreover we can also define **conditional PMF** as,

Definition 2.7.2. PMF of X conditional respect to Y can be written as

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

§2.8 Transformations of a Random Variable

In some applications, we really are interested in distributions of some function of X. We call this concept **Transformation of X**.

Definition 2.8.1. Let X be r.v with PDF/PMF f_X and CDF F_X. Let $Y = r(X)$ i.e $Y = X^2$ or $Y = \ln X$. We call $Y = r(X)$ **transformation of x**.

If Y is discrete, PMF is given by,

$$f_Y(y) = P(Y = y) = P(r(X) = y) = P(\{x \ : \ r(x) = y\}) = P(X \in r^{-1}(y))$$

If Y is continuous, we first calculate CDF and find derivative of it.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(r(X) < y) \\ &= P(\{x \ : \ r(x) \leq y\}) = P(A_y) \\ &= \int_{A_y} f_X(x)dx \end{aligned}$$

And the last step, $f_Y(y) = F^{'}(y)$.

We can also generalize this concepts for Multivariate distributions, which we just increase dimensions we work with (too lazy, add this later).

3 Expectations and Invariance

§3.1 Expectation of a Random Variable

The distribution of X contains all the probabilistic data we need about X. However, we need additional tools to describe these data more cleanly.

One of these tools is **Expectation**, or **Expected Value** or **Mean** of X.

Definition 3.1.1. The **expected value** of X is defined as,

$$|E\rangle\langle E|\left(X\right) = \begin{cases} \sum xf(x) & \text{if } X \text{ is discrete} \\ \int xf(x)dx & \text{if } X \text{ is continous} \end{cases}$$

If expected value is infinite, we say that expected value of X doesn't exist.
We can also combinte both the notations into a whole generalized equation with a notation,

$$|E\rangle\langle E|\left(X\right) = \int xdF(x) = \mu = \mu_X$$

We have discussed that $dF(x) = f(X)$. in the second chapter. **Important Note**. Expectation, by nature, is a theorical mean of the variables we get. It is sometimes possible to get mean that you can't get in a practical settings.

By definition of probability, sum of all $f(x)$ is is simply 1. Then, the above equation is weighted mean of X, which is what we wanted to convey.

Example 3.1.2

Suppose that we have a discrete r.v X to describe the probability of getting heads from tossing a coin 3 times. Let c.d.f of X be f. Then,

$$X = \begin{cases} f(0) = 1/8 \\ f(1) = 3/8 \\ f(2) = 3/8 \\ f(3) = 1/8 \end{cases}$$

Let's use our above formula to calculate $\mathbb{E}(X)$,

$$\mathbb{E}(X) = \frac{1}{8} \cdot 0 + \frac{3}{8} \cdot 1 + \frac{3}{8} \cdot 2 + \frac{1}{8} \cdot 3 = 1.5$$

This number shows that if we repeat our experiment for a very long time, the mean of the heads we got woud be (or approach to) 1.5.
Observe that weighted mean is equivalent to arithmetic mean. Because gettings $X = 2$ is simply getting $X = 1$ two times.

But, what if $Y = g(X)$ and we want to compute $E(Y)$? We have a theorem for that,

Theorem 3.1.3 (Law of the Unconscious Statistician)
Let $Y = g(X)$. Then,

$$|E\rangle\langle E|\,(Y) = \mathbb{E}(g(X)) = \int g(x)dF_X(x)$$

The general proof of this theorem is out of the scope of this book. Comparing this to original expectation equation, we can see that the only thing that changes is $g(x)$ and x, which intuitively makes sense if you think about it. In transformations, probabiltiies remains unchanged, while the result of probabiltiies gets transformed by a function.
Moreover, for a special case $g(x) = I_A(x)$, where $I_A(x) \in \{0,1\}$ depending on $x \in A$, then,

$$\mathbb{E}(I_A(X)) = \int I_A(x)dF_X(x) = \int_A dF_X(x) = |P\rangle\langle P|\,(X \in A)$$

This means that probability is special case of expectation, which makes sense, considering probability itself is some average by definition.

Definition 3.1.4. We call n-th **raw moment** of X as

$$\mu_n = \mathbb{E}(X^n) = \int x^n dF_X(x)$$

If $\mathbb{E}(|X^k|)$ is infinite, then k^{th} moment do not exist.
We also define k**-th central moment** as moments about its mean μ i.e $\mathbb{E}[(X-\mu)^k]$. Additionally, k**-th standardized moments** as $\frac{\mathbb{E}[(X-\mu)^n]}{\sigma^n}$.

The 1st moment, the 2nd central moment, 3rd and 4rd standarized moments are called mean (expected value), **variance** ,**skewness** and **kurtosis** in order. We will learn more about them in later chapters.

The moments are very useful and practical. Although there are infinitely many moments, only smaller ones are important for practical purposes. We already know the first moment and its significance.

Properties of Expectation

Theorem 3.1.5 (Non-negativity)
If $X \geq 0$ is a r.v, then $\mathbb{E}(X) \geq 0$.

Proof. By definition of expectation, we have

$$\mathbb{E}(X) = \int x dF_X(x) \geq 0$$

since by definition, $dF_X(X) \geq 0$ and $x \geq 0$. □

Theorem 3.1.6 (Linearity)
For **any** random variables $X_1, X_2, ..., X_n$ and constants $a_1, a_2, ..., a_n$, we have

$$\mathbb{E}\left[\sum_i^n a_i X_i\right] = \sum_i^n a_i \mathbb{E}(X_i)$$

Proof. We will first prove the theorem for $n = 2$ with X, Y.. $n = 1$ is trivial.

$$\begin{aligned}\mathbb{E}\left[a_1 X_1 + a_2 Y\right] &= \int (a_1 x + a_2 y) dF_{X,Y}(x, y) \\ &= \int (a_1 x) dF_X(x) + \int (a_2 y) dF_Y(y) \\ &= a_1 \int x dF_X(x) + a_2 \int y dF_X(y) \\ &= a_1 \mathbb{E}(X) + a_2 \mathbb{E}(Y)\end{aligned}$$

The second line is the direct consequence of marginality. With induction, $n \geq 3$ is also true, however I will omit the solution for the sake of briefity. □

This theorem is very useful and very practical.

Theorem 3.1.7 (multiplicity)
For **independent** r.v $X_1, X_2, ..., X_n$, we have

$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbb{E}(X_i)$$

Proof. Similiar to last one, we will use induction. $n = 1$ is trivial. For $n = 2$, let r.v be X, Y. Remember that independence has property $dF_{X,Y}(x, y) = dF_X(x) \cdot dF_Y(y)$.

$$\mathbb{E}(XY) = \int (xy) dF_{X,Y}(x, y) = \int xy dF_X(x) dF_Y(y) = \int y dF_Y(y) \int x dF_X(x) = \mathbb{E}(X)\mathbb{E}(Y)$$

For the sake of briefity, I won't show the induction part. □

§3.2 Conditional Expectation

Suppose that we we want to calculate mean of X when $Y = y$. This is called conditional expectation, similar to conditional r.v and probability.

Definition 3.2.1. conditional expectation of X by $Y = y$ is given by,

$$\mathbb{E}(X|Y = y) = \int x dF_{X|Y}(x|y)$$

Note that $\mathbb{E}(X|Y)$ is a r.v itself since we don't know value of Y beforehand, or more precisely Y is a "function".

Theorem 3.2.2 (Law of total Expectations)

for all r.v X and Y,

$$\mathbb{E}[\mathbb{E}(Y|X)] = \mathbb{E}(Y) \qquad \text{and} \qquad \mathbb{E}[\mathbb{E}(X|Y)] = \mathbb{E}(X)$$

Proof. It is direct consequence of definition of conditional expectation and the fact that $dF(x, y) = dF(x)dF(y|x)$

$$\mathbb{E}[\mathbb{E}(Y|X)] = \int \mathbb{E}(Y|X = x)dF(x)$$

writelater □

§3.3 Variance

We have dicussed about the expectation, a way of showing a property of a distribution. However, the expectation alone doesn't convey much. We have another tool called '**Variance**'. Variance, in layman terms, describes how value of random variable varies are spread in the graph. Or in other terms, the distance between the expectation value.

We can define variance ass,

Definition 3.3.1. Let X be a r.v with mean $\mu = \mathbb{E}(X)$. The Variance of X, denoted as $\mathbb{V}(X)$ or $\text{Var}(X)$ or σ^2 is the 2^{nd} central moment and is defined by,

$$\sigma^2 = \mathbb{E}[(X - \mu)^2] = \int (x - \mu)^2 dF(x)$$

We also define **standart deviation** as $\text{sd}(X) = \sqrt{\sigma^2} = \sigma$.

The standart deviation and variance convey the same information. They both represent the spread of our data. The difference between them is purely mathematical. The variance is more useful in mathematical applications, where standart deviation is very intuitive and practical. mathisfun explains it very well.

Calculating variance directly can be complicated and tedious directly sometimes. We can derive a theorem from the original definition for practical purposes.

Theorem 3.3.2

Let X be a random variable. Then,

$$\sigma^2 = E([(X-\mu)^2]) = E(X^2) - \mu^2$$

Proof. It is derived directly by algebraic manipulation and basic calculus,

$$\begin{aligned}\sigma^2 &= \int (x-\mu)^2 dF(x) \\ &= \int x^2 dF(x) - 2\mu \int x dF(x) + \mu^2 \int dF(x) \\ &= \int x^2 dF(x) - \mu^2 \\ &= \mathbb{E}(X^2) - \mu^2\end{aligned}$$

□

§3.4 Conditional Variance

Definition 3.4.1. Let $\mu = \mathbb{E}(X|Y=y)$. The **conditional variance** is defined as,

$$\mathbb{V}(X|Y=y) = \int (x-\mu)^2 dF_{X|Y}(x|y)$$

The conditional variance tells us how much of spread is left after We use $Y=y$. Reminder that $\mathbb{V}(X|Y)$ is a r.v itself since Y is a sort of "function" here.

Theorem 3.4.2 (Law of Total Variance)

for any r.v X, Y, it is always true that,

$$\mathbb{V}(Y) = \mathbb{E}\big[\mathbb{V}(Y|X)\big] + \mathbb{V}\big(\mathbb{E}[Y|X]\big)$$

We have stated before that $V(Y|X)$ and $E(Y|X)$ are random variables, not numbers. Therefore We compute variance and expectation of these random variables, and add them up to get the variance $V(Y)$.

§3.5 Covariance and Corelation

Ley X and Y be r.v. **Covariance** and **Corelation** describes the linear relationship between X and Y.

Definition 3.5.1. If X and Y are r.v with mean μ_X, μ_Y and standart deviations σ_X, σ_Y, we define **covariance** as,

$$\mathrm{Cov}(X,Y) = \mathbb{E}\Big((X-\mu_X)(Y-\mu_Y)\Big)$$

and **corelation** as,

$$\rho = \rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Notice that $\text{Cov}(X, X) = |V\rangle\langle V|(X)$ and $\rho_{X,X} = 1$.

Similiar to variance, calculatig covariance can be tedious. We can derive a better formula by simple algebraic manipulations,

Theorem 3.5.2

For all random variables with non-infinite means, we have

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Proof. Similar to Variance one, we have,

$$\begin{aligned}\text{Cov(X, Y)} &= \mathbb{E}\Big((X - \mu_X)(Y - \mu_Y)\Big) \\ &= \mathbb{E}(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y) \\ &= \mathbb{E}(XY) - \mu_Y\mathbb{E}(X) - \mu_X\mathbb{E}(Y) + \mu_X\mu_Y \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

□

Theorem 3.5.3

For all random variables with non-infinite means, we have,

$$-1 \leq \rho_{X,Y} \leq 1$$

Proof. It is direct consequence of Cauchy-Schwarz inequality. □

§3.6 References

§3.7 Exercises

4 Statistical Inequalities

Statistical Inequalities provide a means of bounding measures and quantities, which is very useful practically and theorically. These inequalities are used for computations on Machine learning and AI, developing new methods and techniques for practical purposes and so on. IMO, International Mathematical Olympiads, also have a "Inequalities" section with questions that heavily use some of the methods and techniques soon be discussed here.

§4.1 Probability inequalities

Probability inequaltiies are useful for boundign quantities that is hard to compute. Moreover, they are heavily used in the **theory of convergence**. Our first theory is,

Theorem 4.1.1

Let X be non-negative r.v and assume $\mathbb{E}(X)$ exists. Then **Markov's inequality** states that for any $a > 0$,

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

Proof.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} xf(x)dx \\ &= \int_{0}^{a} xf(x)dx + \int_{a}^{\infty} xf(x)dx \\ &\geq \int_{a}^{\infty} xf(x)dx \geq a\int_{a}^{\infty} f(x)dx \\ &= aP(X \geq a) \end{aligned}$$

□

Theorem 4.1.2

Let $\mu = E(X)$ and $\sigma^2 = V(X)$. **Chebyshev's inequality** states that,

$$P(|X-\mu| \geq a) \leq \frac{\sigma^2}{a^2} \quad \text{and} \quad P(|Z| \geq k) \leq \frac{1}{k^2}$$

Where Z is **standard score** i.e $Z = \frac{x-\mu}{\sigma}$.

Proof. The theorem is the direct consequence of the markov ineqaulity

$$P(|X-\mu| \geq a) = P(|X-\mu|^2 \geq a^2) \leq \frac{\mathbb{E}([X-\mu]^2)}{a^2} = \frac{\sigma^2}{a^2}$$

The second one is just a substituion. □

§4.2 Expectation ineqaulities

Probably one of the most known ineqaulities of all time, that is also used in different field of mathematics, **Cauchy-Scwartz inequality** Simply states that,

Theorem 4.2.1

Cauchy-Scwarz ineqaulity states that,

$$\mathbb{E}^2(|XY|) \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

Another form of this theorem is,

$$|Cov\rangle\langle Cov|^2\,(X,Y) \leq \sigma_X^2\sigma_Y^2$$

This theorem is also known with its vector form. Consequently, this ineqaulity is very useful and practical, hence there are many unique and different proofs.
This ineqaulity is also popular on mathematical olympiads, with its familiar algebraic form.

Recall that a function f is **convex** if f is twice differentiable $f''(x) \geq 0$. Moreover f is **concave** if $-f$ is convex. Another very well known and popular ineqaulity,

Theorem 4.2.2

Jensen's ineqaultiy states that if g is convex,

$$\mathbb{E}g(X) \geq g(\mathbb{E}X)$$

If g is concave, the inequality symbol flips. Note that we used $\mathbb{E}X$ instead of $\mathbb{E}(X)$ for asthetical purposes.

§4.3 References

§4.4 Exercises

5 Convergence, CLT and LLN

§5.1 Introduction

We already know the caclulus definition of convergence. We say that x_n **converges** to x if, for every $\epsilon > 0$.

$$|x_n - x| < \epsilon$$

as n goes to infinity.
However, there are multiple definitions of convergence in Probability and Statistics. The core idea is simple, in layman terms, as we repeat a process for a long time, something approaches to a thing. Mathematics requires regirousity, and hence there are multiple convergence definitions for the use case. We will learn convergence in the next section

CLT, in other words, **Central Limit theorem**, states that, under correct conditions, the distribution of **normalized verison** of the sample mean **converges in distribution** to a standard normal distribution. There are multiple types of CLT.

LLN, or **Law of Large Numbers** states that under right conditions, the sample average $\overline{X}$ **convergec in probability** to the expectation $\mu = \mathbb{E}(X)$. There are also types of LLN.

§5.2 Types of Convergence

There are two main, most used types of convergence in the Statistics and Probability. The weakest one is, the **convergence in distribution**

Definition 5.2.1. Let $X_1, X_2, ...$ be a sequence of r.v with c.d.f $F_1, \ldots.$ F_n is sait to be **converging in distribution** or **converge weakly** to c.d.f F of a r.v X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

For every x that F is **continous**. Convergence in distribution may be denoted as

$$X_n \overset{d}{\to} X$$

Concept of convergence in distribution **does not require** X_n to be close to X.

More stronger covnergence, **convergence in probability** is defined as,

Definition 5.2.2. Let $X_1, X_2, \ldots$ be a sequence of r.v. X_n is said to be **converging in probability** to r.v X if for all $\epsilon > 0$,

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

Convergence in probabiltiy is denoted as,

$$X_n \overset{p}{\to} X$$

Visually, the converge in probability states that, as n grows large X_n tends to be inside the ϵ brackets of X. However, **it is still possible that X_n goes out of the bound for some time**, but very unlikely.
Almost surely converge is just stronger version of this converge, eliminating the "out of the bound´´ chance.

And other two less commonly used convergences,

Definition 5.2.3. Sequence $\{X_n\}$ converges **Almost surely** or **strongly** towards X if,

$$P\left(\lim_{n\to\infty} |X_n - X| = 0\right) = 1$$

Almost surely convergence is denoted as,

$$X_n \overset{a.s}{\to} X$$

We can also rewrite the definition as, (Ferguson 1996, p. 5)

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

Notice how this equation similar is to probability convergence.
This converge states that, there exists a large number n that X_n **will always be** in the bounds of ϵ. There is also stronger type of this version, called **sure convergence**. However, it is rarely used and there is no practical difference between this and the weaker version almost sure convergence. Therefore we don't talk about it.

and lastly, **convergence in mean**,

Definition 5.2.4. For a real number $r \geq 1$, X_n said to be **converges in r-th mean** to a r.v X if,

$$\lim_{x\to\infty} \mathbb{E}(|X_n - X|^r) = 0$$

This convergence is also called L_r convergence or L^r convergence and is denoted as,

$$X_n \overset{L^r}{\to} X$$

§5.3 Properties of Convergences

Here is a basic diagram showing the chain of implications, from Wikipedia.

$$\begin{array}{ccccc} \xrightarrow{L^s} & \underset{s>r\geq 1}{\Rightarrow} & \xrightarrow{L^r} & & \\ & & \Downarrow & & \\ \xrightarrow{a.s} & \Rightarrow & \xrightarrow{p} & \Rightarrow & \xrightarrow{d} \end{array}$$

- Almost surely convergence implies convergence in probability

$$X_n \xrightarrow{a.s} X \quad \Rightarrow \quad X_n \xrightarrow{p} X$$

- Convergence in probability implies convergence in distribution

$$X_n \xrightarrow{p} X \quad \Rightarrow \quad X_n \xrightarrow{d} X$$

- Convergence in L^s implies convergence in L^r such that $s > r \geq 1$.
$$X_n \xrightarrow{L^s} X \quad \Rightarrow \quad X_n \xrightarrow{L^r} X$$
- Convergence in L^r implies convergence in probability.
$$X_n \xrightarrow{L^r} X \quad \Rightarrow \quad X_n \xrightarrow{p} X$$

All of these are intuitively, from the definition, makes sense. No proofs shall be provided.

Theorem 5.3.1

Continous mapping theorem Let $\{X_n\}$ be r.v sequence, and let X be a r.v. Let g be a continous function. then, the below is true,

$$\begin{aligned} X_n \xrightarrow{p} X \quad &\Rightarrow \quad g(X_n) \xrightarrow{p} g(X) \\ X_n \xrightarrow{d} X \quad &\Rightarrow \quad g(X_n) \xrightarrow{d} g(X) \\ X_n \xrightarrow{a.s} X \quad &\Rightarrow \quad g(X_n) \xrightarrow{a.s} g(X) \end{aligned}$$

Note to myself: Add addivity and multiplicative property with its proof, Slutzky's theorem.

§5.4 LLN, Law of Large Numbers

There are **strong** and **weak** type of LLN, Law of large Numbers. However, both of them state the same idea, $\overline{X_n}$ converges to μ as n goes to infinity.

Theorem 5.4.1

The weak Law of Large Numbers or shortly **WLLN** states that if $X_1, \ldots X_n$ are i.i.d r.vs, then,
$$\overline{X}_n \xrightarrow{p} \mu$$

Proof. For the sake of simplicity, assume variance is finite. The theorem is the direct consequence of Chebysev's inequality.

$$P(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\mathbb{V}(\overline{X}_n)}{\epsilon^2} = \frac{\mathbb{V}(X_1)}{n\epsilon^2}$$

Which right side obviously converges to 0. □

Theorem 5.4.2

The strong Law of Large Number or shortly SLLN states that if $X_1, \ldots, X_n$ are i.i.d r.vs and $\mu < \infty$, then,
$$\overline{X}_n \xrightarrow{a.s} \mu$$

Proof is complex, so I will avoid giving it here.

Practically there is not much difference between WLLN and SLLN, therefore SLLN is preferred.

§5.5 CLT, Central Limit Theorem

Theorem 5.5.1

Central Limit Theorem, or shortly **CLT** states that for i.i.d r.vs $X_1, \dots X_n$ with mean μ and variance σ^2,

$$\frac{\overline{X}_n - \mu}{\sqrt{\mathbb{V}(\overline{X}_n)}} \xrightarrow{d} N(0, 1)$$

§5.6 References

1. https://www.stat.cmu.edu/~larry/=stat325.01/chapter5.pdf

2. https://en.wikipedia.org/wiki/Convergence_of_random_variables

3. https://imai.fas.harvard.edu/teaching/files/Convergence.pdf

4. https://stats.stackexchange.com/questions/2230/convergence-in-probability-vs-almost-

5. LafayedeMicheaux,P.,Liquet,B.(2009).UnderstandingConvergenceConcepts: AVisual-MindedandGraphicalSimulation-BasedApproach.TheAmericanStatistician, 63(2),173178.doi:10.1198/tas.2009.0032

6. https://stats.stackexchange.com/questions/3734/what-intuitive-explanation-is-there-

§5.7 Exercises

Part II

Statistical Inference

6 Statistical Models and Statistical Inference

Through the last chapter, we have talked about multiple probability distributions and their functions such as c.d.f, p.d.f, and p.m.f. However, we have assumed that we already knew the distribution and its properties. In practical world, it is not the case. We may try to find the average salary of the country, the fatality rate of a virus, and so on. Statistical inference, in shortly, is study of using the information sample we are given to deduce the characteristics of a population. Since majority of population is defined with **paramateres**, our investigation is mainly on finding, or estimating such paramteres.

§6.1 Model and Inference

Assume that $X \sim N(0, 1)$ and $Y \sim N(0, 2)$. The X and Y don't have the same probability characteristics, more cleanly, they don'thave the same distribution. However, we generalize them in a "group" called *normal distribution*. That is, there are infinitely many distribution with $N(\mu, \sigma^2)$, and to convey them in a more general way, we use **statistical models**

Definition 6.1.1. **Statistical Model** is a set of distribution such as,

$$\mathfrak{F} = \left\{ f(x; \theta) : \theta \in \Theta \right\}$$

where Θ is a **parameter space**.
There are also **nonparametric** $\mathfrak{F}$.

In a parametric model, if we are interested in only one parameter θ, we call it **target parameter** or **estimand**. The process of getting the estimand is called **Statistical Inference**. Majority of inferental problems are divided into three types: **estimation, confidence sets, hypothesis testing**.

§6.2 Point Estimation

By convention we write the estimate of θ as $\widehat{\theta}$. Since θ is constant and by definition $\widehat{\theta}$ is a function, $\widehat{\theta}$ is a r.v. Remark that funtion of r.vs is a r.v. In more mathematical way,

Definition 6.2.1. Let $X_1, ..X_n \sim F$ be i.i.d. A point estimator $\widehat{\theta}$ is defined as,

$$\widehat{\theta} = g(X_1, \ldots, X_n)$$

We also define a very useful variable **bias** as,

$$|bias\rangle\langle bias|\,(\widehat{\theta}, \theta) = |E\rangle\langle E|\,(\widehat{\theta}) - \theta$$

Here, θ is our target parameter, $\widehat{\theta}$ is the function we use to estimate our target, or the estimator. We usually write $|bias\rangle\langle bias|\,(\widehat{\theta}, \theta) = |bias\rangle\langle bias|\,(\widehat{\theta})$

Bias, in a literal sense, tells us the bias of the estimator we use. That is, the error that we may find when we estimate our parameter. We say that $\widehat{\theta}$ is **unbiased** if,

$$|bias\rangle\langle bias|\,(\widehat{\theta}) = 0 \Rightarrow \mathbb{E}(\widehat{\theta}) = \theta$$

We know that $\widehat{\theta}$ is a r.v. We call this r.v's distribution as **sampling distribution**. We also define, **standard error of** $\widehat{\theta}$ or standard deviation,

$$|se\rangle\langle se| = \sqrt{\mathbb{V}(\widehat{\theta})}$$

It is logical to think that the estimator should converge (with more samples) to its target value, we define such property as,

Definition 6.2.2. If a point estimator $\widehat{\theta}$ converges to θ, we call that $\widehat{\theta}$ is **consistent**

With bias alone, we can't characterize the quality of the estimator. Because the values of $\widehat{\theta}$ may be far away than real value θ, but still be $\mathbb{E}(\widehat{\theta}) = \theta$. Therefore, we also have to measure the variance in some way.
For such thing, we already have a tool,

Definition 6.2.3. The mean square error is defined as,

$$|MSE\rangle\langle MSE|\,(\widehat{\theta}) = \mathbb{E}([\widehat{\theta} - \theta]^2)$$

in similiar fashion to the Variance definition, we can rewrite this equation as,

$$|MSE\rangle\langle MSE|\,(\widehat{\theta}) = |bias\rangle\langle bias|^2\,(\widehat{\theta}) + \mathbb{V}(\widehat{\theta})$$

MSE is function of both its variance and bias, hence it is a better way of showing the quality of the estimator.

Example 6.2.4

Let $X_1, \ldots, X_n \sim |Bernoulli\rangle\langle Bernoulli|\,(p)$. Let $\widehat{p} = n^{-1}\sum_{i=1}^{n} X_i$. We already know that $\mathbb{E}(X) = p$ and $\mathbb{V}(X) = p(1-p)$. Our estimator is unbiased since,

$$\mathbb{E}(\widehat{p}) = \frac{1}{n}\sum \mathbb{E}(X_i) = p$$

Moreover, the estimator's variance is,

$$\mathbb{V}(\widehat{p}) = \mathbb{E}(\widehat{p}^2) - \mathbb{E}(\widehat{p})^2 = p - p^2$$

We can intutively guess some estimators that could be effective for our purposes. But for many complex problems, it is not the case. We will learn new methods to calculate estimators in later chapters.

§6.3 Properties of Point Estimation: Efficienty, Consistency, Sufficiency

Relative Efficiency

We already know that it is possible to have multiple estimators for one target parameter. We even learnt a new definiion, MSE to convey the quality of such estimators. If we have two unique and unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$, it is logical to pick the estimator that has the lowest variance, since the lower the MSE, more efficient the estimator is. To convey such idea, we use,

Definition 6.3.1. Given two unbaised estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$, then **the efficiency of $\widehat{\theta}_1$ relative to $\widehat{\theta}_2$**, denoted as $|eff\rangle\langle eff|\,(\widehat{\theta}_1, \widehat{\theta}_2)$, is defined as,

$$|eff\rangle\langle eff|\,(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{\mathbb{V}(\widehat{\theta}_2)}{\mathbb{V}(\widehat{\theta}_1)}$$

Note that if $|t_1 : w, t_2\rangle\langle t_1 : w, t_2|$ is bigger than one, then it is true that $\widehat{\theta}_1$ is relatively more efficient than $\widehat{\theta}_2$.

Consistency

We have already talked about consistency before, we say the estimator is consistent of it converged to the target parameter,

Definition 6.3.2. The estimator $\widehat{\theta}$ of θ is consistent if for any positive number ϵ,

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| \leq \epsilon) = 1$$

That is, $\widehat{\theta} \xrightarrow{p} \theta$

The graph (below) from latter exercises is also consistent, since visually it becomes a straight line where it equals to the target parameter.

Since consistent estimators converge to the target parameter, it is logical to think the variance also converges to 0. Think in a way that the graph of the estimator has to become straight from long wavy and curvy lines i.e it is direct consequence of convergence (real analysis stuff?). Indeed,

Theorem 6.3.3

The unbiased estimator $\widehat{\theta}$ of θ is a consistent estimator if,

$$\lim_{n\to\infty} \mathbb{V}(\widehat{\theta}_n) = 0$$

Sufficiency

We know that the value $\overline{X}$ (average value) is a unbiased estimator for mean μ of X. At this point, we no longer need the sample data to estimate the μ, since we can summarize the information just with the estimator $\overline{X}$. But, do the $\overline{X}$ retain all the information about X?. If it does, we call such estimator **sufficient**. That is all the sufficiency is for.

We can mathematically convery this property as conditional distribution of our sample data, given the estimator. If the distribution is dependent on our target parameter, it can't be sufficient. In more mathematical way,

Definition 6.3.4. A **statistic** is a function of data (Remark: all estimators are statistic but not all statistic are estimators). A statistic $U = t(X_1, .., X_n)$ of θ is sufficient if conditional distribution of $X_1, ..., X_n$ given U is not dependent on θ.

If conditional distribution is dependent on the target parameter, it is intuitive to think the statistic does not contain all the information.

Sufficiency is useful since it helps us to *assessing information on the entire population without the need of all the data.*
Say you get your grade on an exam and you want to know how well you did compared to your classmates. If you are given a sample mean and variance, you can do this without asking everyone's grades.

§6.4 Confidence Intervals

Let's assume we are a scientist that want to measure the mean of the specific kind of mice's weight. It is unrealistic to measure **all of the mice**, hence we catch a small amount of them, probably in hundreds, measure them and gather the data in a datasheet. We **bootstrap** (we will learn what that term is in later chapters) the sample data, and find the sample mean. Now, we repeat the bootstrapping process thousands of times, which now we have a sample mean data.
Now, let's find numbers a, b such that 95% of our sample mean data resides in interval $[a, b]$. That is what **confidence interval** basically is.

Definition 6.4.1. Let X be a random vector. The $1-\alpha$ **confidence interval** for a parameter θ is an interval $[a, b]$ and $a = a(X)$, $b = b(X)$ functions such that,

$$P(a \le \theta \le b) \ge 1 - \alpha$$

Note that θ is unknown constant value, while a and b are random variables.
Taking the above example, $\alpha = 0.05$, which is a mathematical standard number used majority of time. $1-\alpha$ is called **confidence coefficient**. We also call **lower and upper confidence limits** to a and b, sometimes also donated as θ_L and θ_U.

It is also possible to form *one sided confidence interval*, i.e,

$$P(\theta_L \le \theta) \ge 1 - \alpha \qquad \text{or} \qquad P(\theta \le \theta_U) \ge 1 - \alpha$$

The confidence intervals may be **closed or open**. For our purpose they are indifferent.

Theorem 6.4.2 (Normal-Based Confidence Intervals)
Let $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. Then, there is $1-\alpha$ chance that the below interval,

$$\widehat{\theta} \pm z_{\alpha/2} \, |\widehat{se}\rangle\langle\widehat{se}|$$

will containt the real value θ.

Proof. Let the above interval be C. Then,

$$\begin{aligned} P(\theta \in C) =& P(\widehat{\theta} - z_{\alpha/2} \, |\widehat{se}\rangle\langle\widehat{se}| < \theta < \widehat{\theta} + z_{\alpha/2} \, |\widehat{se}\rangle\langle\widehat{se}|) = \\ & P(-z_{\alpha/2} < \frac{\widehat{\theta} - \theta}{|\widehat{se}\rangle\langle\widehat{se}|} < z_{\alpha/2}) \xrightarrow{d} \\ & P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha \end{aligned}$$

□

§6.5 Hypothesis Testing

Please read the Hypothesis chapter instead.

§6.6 References

1. https://math.stackexchange.com/questions/1767877/in-laymans-terms-what-is-the-diffe

2. https://stats.stackexchange.com/questions/3911/when-are-confidence-intervals-useful

3. https://mason.gmu.edu/~alaemmer/bio214/sampling-distributions.pdf

§6.7 Exercises

7 Methods of Estimation (Parametric Inference)

In later chapter we have shortly talked about the point estimation. The estimator $\widehat{\theta}$ of a target parameter θ is a function of random variables of a sample, and therefore it itself is a random variable. The estimator has its own probability distribution, *sampling distribution*. We already know about *unbaised estimators* i.e $\mathbb{E}(\widehat{\theta}) = \theta$ and the *consistent estimator*. In this chapter, we will learn more deeply about the mathematical properties of the point estimators. Additionally, we will learn new methods to derive estimators, since until now we listened our intuition.

§7.1 Method of Moments

Until now, we have used our intuiton to find estimators. For example, it is logical to think that $\overline{X}$ would be an ideal estimator for μ of X. However, in practical world we have to generate the parametric estimators more "mathematically". First, we introduce with a new simple definition,

Definition 7.1.1. k-th sample moment $\widehat{\alpha}_k$ is moment of sample i.e

$$\widehat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

In section 3.1 we talked about **raw moments**. Raw moments convey the properties of the distribution i.e raw moments are some functions of the desired parameters. The first raw moment is the mean $\mu_1 = \mu$, the second raw moment is expression of variance $\mu_2 = \sigma^2 + \mu^2$ and so on.

The idea method of moment is we can use $\widehat{\alpha}_k$ as good estimator of μ_k, and from μ_k we can derive expressions for our target parameter.

Example 7.1.2

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. We have that $\mathbb{E}(X) = \mu$ and $\mathbb{E}(X^2) = \sigma^2 + \mu^2$.

$$\widehat{\alpha}_1 = \widehat{\mu}$$
$$\widehat{\alpha}_2 = \widehat{\mu}^2 + \widehat{\sigma^2}$$

Solving the system equation will give us estimators for μ and σ.

§7.2 Method of Maximum Likelihood

The method of moments are very simple and intuitive, but it is unefficient. We have a better and sophisticated method called **method of maximum likelihood**. There is a great video by Josh Starmer that explains the method very well.

Assume that we have a sample data, and we want to estimate parameters of the distribution that describes the sample data. The idea is that we find such estimator that maximaze the **likelihood** of getting our sample data relative to the parameter.

Definition 7.2.1. The **Likelihood function** is defined as,

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

Also we define **log-likelihood function** as,

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta)$$

At last, we define the **maximum likelihood estimator** MLE denoted by $\widehat{\theta}_n$ as the value of θ that maximizes $\mathcal{L}_n(\theta)$, or better $\ell_n(\theta)$, since working with logs are easier that multiplicative functions for maximizing.

We already know that θ is a unknown constant we want to estimate. The $\mathcal{L}_n(\theta)$ describes the likelihood of each sample data, respect to θ. Since it is intuitive to maximize the likelihood (because the sample data is already happened and should be maximized), it should also estimate our value θ.

§7.3 Properties of MLE

Will write later.

8 Estimating Statistical Functions (Non-Parametric Inference)

§8.1 Emprical Distribution Function, e.d.f

Definition 8.1.1. let $X_1, \dots, X_n \sim F$ be i.i.d sample with F as $c.d.f.$
We can estimate F with **empriical distribution function** or shortly **e.d.f** $\widehat{F}$,

$$\widehat{F}_n(x) = n^{-1} \sum_{n=1}^{n} I(X \leq x)$$

Theorem 8.1.2

e.d.f $\widehat{F_n}$ coverges almost surely to F, or is **consistent** that is,

$$\widehat{F}_n(x) \xrightarrow{a.s} F(x)$$

It also estimates F with no bias, that is,

$$\mathbb{E}(\widehat{F}_n(x)) = F(x)$$

Theorem 8.1.3

Dvoretzky–Kiefer–Wolfowitz–Massart inequality (DKW)
Let $X_1, \dots, X_n \sim F$. Then,

$$P\left(\sup_{x \in \mathbb{R}} |F(x) - \widehat{F}_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \ \forall \epsilon > 0$$

This inequality is useful for constructing **confidence intervals** such that,

$$P\left(\widehat{F}_n(x) - \epsilon \leq F(x) \leq \widehat{F}_n(x) + \epsilon\right) = 1 - \alpha \quad \forall \epsilon = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$$

§8.2 Statistical Functionals

Statistical functionals are functions of data, that is $T(F)$. Examples are mean, variance, median and so on.

Definition 8.2.1. **plug-in estimator** of $\theta = T(F)$ is defined as,

$$\widehat{\theta}_n = T(\widehat{F}_n)$$

That is, plug in $\widehat{F}_n$ to estimate our statistical function $T(F)$.

Definition 8.2.2. T is **linear functional** if $T(F) = \int r(x)dF(x)$. Linear functional T satisfishes linearity properties, that is,

$$T(aF + bG) = aT(F) = bT(G)$$

§8.3 References

1. https://en.wikipedia.org/wiki/Empirical_distribution_function
2. https://en.wikipedia.org/wiki/Dvoretzky%E2%80%93Kiefer%E2%80%93Wolfowitz_inequality

9 Hypothesis Testing and p-value

§9.1 To-learn and write

wald test, z-test, t-test,chi, various methods and distributions.

§9.2 Null and Alternate Hypothesis

The hypothesis testing is very similar to the scientific method. Scientists across the different fields use scientific method for their academical purposes. The observe, formulate a theory, experiment and test the theory. There is a similar method called **hypothesis testing** for statistical inference. First, we will introduce some notations and definitions,

Definition 9.2.1. **Null Hypothesis**, denoted by H_0, is the hypothesis to be tested. **Alternate Hypothesis**, denoted by H_1, is the hypothesis contradictory to the null hypothesis. We usually try to support, since this way we could use *proof by contradiction.* If our evidence (data) favors the alternative hypothesis, we reject the null hypothesis. Formally, we wish to test,

$$H_0 : \theta \in \Theta_0 \qquad or \qquad H_1 : \theta \in \Theta_1$$

Where Θ_0 and Θ_1 are disjoint sets of parameter space Θ.

Example 9.2.2

Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$ with known variance σ^2 and uknown mean μ. We wish to test the hypothesis,

$$H_0 : \mu = \mu_0 \qquad or \qquad H_1 : \mu \neq \mu_0$$

In order to test the hypothesis, it is logical to calculate $\overline{X}$ and compare it with μ_0. It is reasonable to reject H_0 if $\overline{X}$ is far away than μ_0. But how much far away exaclty?
Rejection or Critical Region is a set denoted by R to describe "how far away" the result can be. If the result is in the set R, we reject the hypothesis.

Definition 9.2.3. The **Critical Region** is defined as,

$$R = \left\{ x : T(x) > c \right\}$$

Here, c is called **critical value**, and T is a **test statistic** to help testing our hypothesis. The main question in hypothesis testing is find appropriate T and c.

In above example, $\overline{X}$ is our T. We will learn about finding c now.

A hypothesis in the form of $\theta = \theta_0$ is called **simple** i.e θ has only one value, while in the form of $\theta > \theta_0$ or $\theta < \theta_0$ is called **composite** i.e θ has multiple values

There are 4 different possibilities we can conclude from our test. See the table for the brief introduction.
If H_0 is true, and we reject it, we call this error **type I error**. Similarly, if H_1 is true and we keep the H_0, (or reject H_1) we call this error **type II error**. Summarize this in table,

	Retain Null	Reject Null
H_0 true	$(1-\alpha)$	type I error (α) significance level of the test
H_1 true	type II error (β)	power of the test $(1-\beta)$

Definition 9.2.4. The **power function** of a test with critical region R is a function defined as,

$$\beta(\theta) = P_\theta(X \in R)$$

that is, $\beta(\theta)$ represents the probability of rejecting H_0 if $\theta \in \Theta_0 \cup \Theta_1$ i.e H_0 or H_1 true. It is very general
The **size** of the test is defined by,

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$$

that is, for θ that H_0 is true, α is the worst case scenario probability, representing **type I error** (if the test is simple).
We also define **significance level** of the test as α if its size is less than or equal to α. That is, upper bound for incorrectly rejecting H_0. Significance level is chosen independent of data or the tests, usually as 0.05. Note in particular that both size and level don't relate to the sample.

If we are conducting pointwise test, for example $\mu = \mu_0$, then **significance level, the size of the test, and type I error** coincide, that is they are equal. However, for some special cases they may differ. Similarly, **power,1- type II error** may also differ. So, we must be careful on our definitions when our test is not simple, that is, is composite.

§9.3 p-value

The size of the test and power of the test is not that much related to our sample data. That is where **p-value** comes in.
The p-value is the probability under the null hypothesis of obtaining a real-valued test statistic at least as extreme as the one obtained, more rigirously,

Definition 9.3.1. Suppose for every $\alpha \in \{0,1\}$ we have a rejection region R_α. Then **p-value** is defined as,

$$p = \inf\left\{\alpha : T(X) \in R\right\}$$

That is, the p-value is the smallest level at which we can reject H_0. We can calculate p with,

$$p = \sup_{\theta \in \Theta_0} P_\theta(T(X) \geq T(x))$$

Notice that the expression contains sample data x.

§9.4 Learn and Write missing sections about Hypothesis Testing(*)

§9.5 References

1. https://en.wikipedia.org/wiki/Statistical_hypothesis_test#Definition_of_terms
2. https://stats.stackexchange.com/questions/183800/how-to-understand-the-size-of-hypo
rq=1
3. https://stats.stackexchange.com/questions/299873/differences-between-p-value-level-
4. https://en.wikipedia.org/w/index.php?title=P-value&oldid=554910098

10 Statistical Decision Theory

Statistical Decision Theory is widely used in machine learning fields, for example training neural networks and other statistical models via computers.
There are multiple methods of generating estimators. Decision Theory helps us decide which method is more suited for our job.
The "relationship" between θ and $\widehat{\theta}$ can be measured with a **loss function** $L(\theta, \widehat{\theta})$. The examples are,

$$\text{Squared Error Loss} \quad (\theta - \widehat{\theta})^2$$
$$\text{absolute Error Loss} \quad |\theta - \widehat{\theta}|1$$
$$L_p\text{Loss} \quad |\theta - \widehat{\theta}|^p$$

We kind of already studied these functions before.

§10.1 Risk Function

To judge the estimators, we use average of loss, or **risk**.

Definition 10.1.1. The **risk** of an estimator $\widehat{\theta}$ is,

$$R(\theta, \widehat{\theta}) = \mathbb{E}\left(L(\theta, \widehat{\theta})\right)$$

Notice that if our loss function is squared error,

$$R(\theta, \widehat{\theta}) = |MSE\rangle\langle MSE|\,(\theta, \widehat{\theta})$$

Part III

Statistical Models

11 Linear Regression

Given the data with random vectors X, Y of the form

$$(\mathbf{X}, \mathbf{Y}) \sim F_{X,Y}$$

Regression is a method finding a linear function to fit properly our data. This way, we can predict the value of Y_i, or estimate in other words.

§11.1 Simple Linear Regression Model

Definition 11.1.1. We can write the simpliest linear regression as follows,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Here, ϵ_i is called **error variable, disturbance term, or error term**. Error term represents factors other than x that affects y. It is normall assumed that $\epsilon_i \sim N(0, \sigma^2)$. That is,

$$\mathbb{E}(\epsilon_i | X_i) = 0 \quad \wedge \quad \mathbb{V}(\epsilon | X_i) = \sigma^2$$

We work with conditionalities in linear regression, since the idea of the regression itself is conditional i.e Y depends on X.

The unknown variables in our model are **intercept** β_0, **slope** β_1 and our ϵ_i. We can estimate ϵ_i with **residual** $\widehat{\epsilon_i}$. In practice we would like to minimize the values of,

Definition 11.1.2. **Residual Sums of Squares** or shortly **r.s.s or RSS**, which measures how our model works with our data, defined as,

$$|RSS\rangle\langle RSS| = \sum_{i=1}^{n} \widehat{\epsilon_i}^2$$

We also define **least square estimates**, that are values β_0 and β_1 such that minimizes RSS. We can derive an expression for least square estimates. In later chapters, we will also work with multivariate versions of the same topic.

Theorem 11.1.3

The least square estimates are given by,

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} \\ \widehat{\sigma}^2 &= \left(\frac{1}{n-2}\right)\sum_{i=1}^{n}\widehat{\epsilon}^2\end{aligned}$$

The last estimator is also called **residual standar error** or shortly RSE.
All of these estimators are unbiased. We usually denote the least square estimators as matrices i.e $\widehat{\beta}^T = (\widehat{\beta}_0, \widehat{\beta}_1)^T$, for convenience.

Proof Sketch. Notice that

$$|RSS\rangle\langle RSS| = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)$$

Therefore it is enough to minimize the right expression, which is trivial via derivatives, and solving the system equations. We will later see the multivariate proof of the theorem via matrices and such. □

§11.2 Maximum Likelihood and Least Square Estimators

Assume that Y is normally distributed. It is very practical, and holds true in practice. The likelihood function is then,

$$\prod_{i=1}^{n} f(X_i, Y_i) = \prod_{i=1}^{n} f_X(X_i) \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i) = \mathcal{L}_1 \cdot \mathcal{L}_2$$

We can use maximum likelihood method to estimate our parameters β_1, β_2. $\mathcal{L}_1$ doesn't include our target parameters, so we will treat it as a constant. We wish to maximize $\mathcal{L}_2$. Using the log-likelihood and our normal distribution assumption,

$$\ell(\beta_0, \beta_1, \sigma) = -n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$$

Note that we used the p.d.f of normal distribution and the fact that $\mu = \beta_0 + \beta_1 X_i$. To maximize the log-likelihood, we have to maximize $|MLE\rangle\langle MLE|$ of (β_0, β_1)

$$|RSS\rangle\langle RSS| = \sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$$

Which is the least square estimator. In fact, we have the following theorem,

Theorem 11.2.1

As long as we have the normality assumption, the least square estimator and likelihood estimator are equivalent.

§11.3 Properties of Least Sqaure Estimators

We begin with the expectation and variance of least square estimators. Remember that we work with conditionals on X.

Theorem 11.3.1

The least square estimator matrix $\widehat{\beta}$ is unbiased i.e

$$\mathbb{E}(\widehat{\beta}|X) = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Proof Sketch. We already derived a way of writing least square estimators in the latter chapter, **Theorem 12.1.1**. Using that fact and $\mathbb{E}(\epsilon|X) = 0$, it is matter of manupilating the expectation until we get our result. □

Theorem 11.3.2

Following properties are true:

1. Consistency: $\widehat{\beta} \xrightarrow{p} \beta$
2. Asymptotically normal: $\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0,1)$
3. $1 - \alpha$ confidence intervals:

$$\beta \pm z_{\alpha/2} \, |\widehat{se}(\beta)\rangle\langle\widehat{se}(\beta)|$$

§11.4 Multivariate Regression

Suppose that, now we have multiple input variables $X_1, \ldots X_k$ i.e

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1k} \\ 1 & X_{21} & \ldots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \ldots & X_{nk} \end{pmatrix}$$

In this notation, we have k variables with total n datas for each variable. Similarly we define,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Then we can write our simple regression model with matrix notations, i.e

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

It is very convenient to work with the matrices. The next theorem, which is the generalization of the least square estimate expressions, shows that

Theorem 11.4.1

Assume that $\mathbf{X^TX}$ is invertible, then

$$\begin{aligned} \widehat{\beta} &= (\mathbf{X^TX})^{-1}\mathbf{X^TY} \\ \mathbb{V}(\widehat{\beta}|\mathbf{X^n}) &= \sigma^2(\mathbf{X^TX})^{-1} \end{aligned}$$

Proof-Sketch. Let $S = \sum_{i=1}^{n} \widehat{\epsilon_i}^2$
Assuming the inverse of $\mathbf{X^TX}$ exists, we wish to minimize

$$S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^2$$

Opening the paranthesises, taking the derivative and equating to 0, (note that S is convex) proves the first part of the theorem.
Second theorem uses the definiton variance and matrix manipulation (simply put).

□

§11.5 References

1. https://stats.stackexchange.com/questions/221891/difference-between-residual-and-di

2. https://stats.stackexchange.com/questions/46151/how-to-derive-the-least-square-esti

3. https://igpphome.ucsd.edu/~cathy/Classes/SIO223A/sio223a.chap7.pdf

4. https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Unbiasedness_and_variance_of_%7F'%22%60UNIQ--postMath-00000037-QINU%60%22'%7F

5. https://stats.stackexchange.com/questions/124818/logistic-regression-error-term-and

6. https://stats.stackexchange.com/questions/23479/why-do-we-assume-that-the-error-is-

7. https://stats.stackexchange.com/questions/148803/how-does-linear-regression-use-the

12 Classification

In linear regresison models, it is assumed that response variable Y is quantative, that is, it issome real number. In practical world, Y can be qualitative e.g blood type, eye color and so on.
Study of predicting qualitative responess are called **classification**.

§12.1 Logistic Regression

Assume our response variable Y takes binary values i.e $Y \in \{0, 1\}$. There is a method called **logistic regression** that models the probability that Y belongs to a particular category, in this example 0 or 1, that is,

$$P(Y = 1|X) \quad \text{and} \quad P(Y = 2|X)$$

Trying linear regression model, we see that the value is not in the interval $[0, 1]$, e.g it sometimes takes negative values. To avoid this problem, we use **logistic function**.

Definition 12.1.1. **Logistic function**, sometimes called **sigmoid function**, is defined by formula,

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x} = 1 - \sigma(-x)$$

Note that σ in this formula do not represent the variance, but rather the logistic function.

Using this function, we can derive a statistical model caleld **logistic regression**,

Definition 12.1.2. **Logistic regression** is defined as,

$$p(X) \equiv P(Y_i = 1|X = x) = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\beta + \sum_{i=1}^{n} \beta_i x_i}}$$

With simple algebraic manipulation, we can find that,

$$\ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

The log expresion in the left side is usually called **logit**, that is,

$$\left|logit\right\rangle\!\left\langle logit\right| (p(X)) = \ln\left(\frac{p(X)}{1 - p(X)}\right)$$

There are subtle differences between linear regression and logistic regression. (A chapter about this?+bernoulli+error term distirbution so on)

§12.2 Estimating Logistic Coefficients

It is possible to generate the coefficient via MLE. Since Y is binary, the data is in binomial distribution, that is,

$$Y|X \sim |Bernoulli\rangle\langle Bernoulli|\,(p)$$

therefore we wish to maximize,

$$\ell(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i) \prod_{i=1}^{n} 1 - p(x_i)$$

Statistical softwares such as R and python libraries can easily compute the MLE. There are also multiple algorithms, one of them being **Gradient Descent**, a very useful technique used heavily in Deep Learning.

§12.3 Multinomial Logistic Regression

Assume our dependent variables Y get more than 2 values. It is possible to use logistic model, with minor changes, for this purpose. This model is called **Multinomial Logistic Regression**.

Let the number of classes (dependent variables) be K. The idea is choose a class as "pivot", in this example $Y = K$, and other $K-1$ outcomes are logictic regressed against the pivot. That is, we apply binary logistic regression multiple times, e.g (We wont write conditional part of the probabilities for the sake of the simplicity)

$$\ln \frac{P(Y_i = k)}{P(Y_i = K)} = \beta_k \cdot \mathbf{X_i}, \quad k < K$$

Note that here β_k is a matrix.

Afterwards, we use the fact that all the probability sum up to 1, then

$$P(Y_i = K) = 1 - \sum_{j=1}^{K-1} P(Y_i = j) = 1 - \sum_{j=1}^{K-1} P(Y_i = K) e^{\beta_j \mathbf{X_i}}$$

With basic algebraic manipulation, we derive that

$$P(Y_i = K) = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\beta_j \mathbf{X_i}}}$$

Then other non-pivot probabilities are,

$$P(Y_i = k) = \frac{e^{\beta_k \mathbf{X_i}}}{1 + \sum_{j=1}^{K-1} e^{\beta_j \mathbf{X_i}}}$$

This statistical model works, and we estimate $K-1$ coefficients to generate our model. However, this model is not symmetrical i.e Kth class is "being ignored". There is a better method, first we introduce **softmax function**, generalization of the logistic function

Definition 12.3.1. **Softmax Function** takes input matrix $\mathbf{X} = \{x_1, \ldots, x_K\} \in \mathbb{R}^K$ and converts the value to probabilities,

$$\sigma(\mathbf{X})_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$$

Intuitively, this function is also called **normalized exponential function**, because this is what basically the function does. Moreover, the sum of all the components adds up to 1.

Using the above function,

Definition 12.3.2. Softmax Multinomial Logistic Regression, for all K classes, we have

$$P(Y_i = k) = \frac{e^{\beta_k \mathbf{X_i}}}{\sum_{j=1}^{K} e^{\beta_j \mathbf{X_i}}}$$

§12.4 Estimating Multinomial Logistic coefficients

We can either use multinomial logit, or generalized maximum likelihood for joints of coefficients, which is out of the scope of my knowledge and the book.

§12.5 Linear Discriminant Analysis, l.d.a or LDA

§12.6 QDA

§12.7 KNN, write the above chapters after finish learning(*)

§12.8 References

1. https://stats.stackexchange.com/questions/124818/logistic-regression-error-term-and

2. https://en.wikipedia.org/wiki/Multinomial_logistic_regression#

§12.9 Exercises

13 Resampling Methods

§13.1 Bootstrap

Bootstrap is a technique about generating more sample data from existing data. It is used for computing confidence intervals and estimating standard errors. The idea is simple. We have sample data $X_1, \dots X_n$ from unknown distribution F. First we estimate F with $\widehat{F}_n$, then we draw random resamples from $\widehat{F}_n$ multiple times and calculate our wanted statistical functions.

§13.2 Basic Introduction

Definition 13.2.1. Bootstrapping
Let $X = \{X_1, \dots X_n\}$ be our sample data from unknown distribution F. That is,

$$X_1, X_2, \dots X_n \sim F$$

Let' s assume we are interested in $T(X)$. In practical statistical problems, we need to know about the distribution of $T(X)$. We can estimate our unknown distribution F with $\widehat{F}_n$, and draw random samples from that known distribution, that is,

$$X_1^*, X_2^*, \dots, X_n^* \sim \widehat{F}_n$$

and compute T_n^* from these samples.

§13.3 References

1. https://stats.stackexchange.com/questions/26088/explaining-to-laypeople-why-bootstra

2. https://ocw.mit.edu/courses/14-384-time-series-analysis-fall-2013/2fdf997bca65d6ed8
   MIT14_384F13_lec9.pdf

14 Deep learning

15 Unsupervised Learning