STAT 512: Statistical Inference **Original notes credit: Yen-Chi Chen**

Lecture 1: Introduction to Probability and Statistics

Instructor: Emilija Perković

These notes are partially based on Perlman (2019).

Useful additional reading: Chapters 1 and 4 of Casella and Berger (2021). Please work through the examples and results in Chapter 1 of Casella and Berger (2021).

Useful recall: Basic set theory, circle and ellipse functional representations, polar coordinates, computing the area of a sector of a circle, change of variables (see also Lecture notes 8).

1.1 Sample Space and Probability Measure

The *sample space* Ω is the collection of all possible outcomes of a random experiment, e.g. toss of a coin, $\Omega = \{H, T\}$. Elements $\omega \in \Omega$ are called *outcomes*, *realizations* or *elements*. Subsets $A \subseteq \Omega$ are called *events*. You should able to express events of interest using the standard set operations. For instance:

- "Not A" corresponds to the *complement* $A^c = \Omega \setminus A$;
- "A or B" corresponds to the *union* $A \cup B$;
- "A and B" corresponds to the *intersection* $A \cap B$.

We said that $A_1, A_2, ...$ are *pairwise disjoint/mutually exclusive* if $A_i \cap A_j = \emptyset$ for all $i \neq j$. A *partition* of Ω is a sequence of pairwise disjoint sets $A_1, A_2, ...$ such that $\cup_{i=1}^{\infty} A_i = \Omega$. We use $|A|$ to denote the number of elements in A.

The sample space defines basic elements and operations of events. But it is still too simple to be useful in describing our senses of 'probability'. Now we introduce the concept of σ-algebra.

Definition 1.1 *A* σ-algebra $\mathcal{F}$ *is a collection of subsets of* Ω *satisfying:*

(A1) (full and null set) $\Omega \in \mathcal{F}$, $\emptyset \in \mathcal{F}$ *(*$\emptyset$ *= empty set).*

(A2) (complement)$A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.

(A3) (countable union) $A_1, A_2, ... \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The σ-algebra defines a collection of events. The sets in $\mathcal{F}$ are said to be *measurable* and $(\Omega, \mathcal{F})$ is a *measurable space*. The intuition of a set being measurable is that we can find a function that takes the elements of $\mathcal{F}$ and output a real number; this number represents the 'size' of the input element.

Borel σ-algebra. Generally, we will not choose $\mathcal{F}$ to be just any σ-algebra. For $\Omega \equiv \mathbb{R}$, we will consider the smallest σ-algebra that contains all intervals of the form $(-\infty, a]$, for $a \in \mathbb{R}$ (smallest in terms of the union operation). This algebra is known as the Borel σ-algebra.

We use $\mathcal{O}$ to denote the collection of all open set on $\mathbb{R}$, that is all sets that can be formed by countable union, intersection and complement of open sets of the form (a, b), $a, b \in \mathbb{R}$. The Borel σ- algebra, denoted by $\mathcal{B}$ is a σ-algebra generated by open sets $\mathcal{O}$. We also write $\mathcal{B} = \sigma(\mathcal{O})$. By definition of a σ-algebra, a Borel σ-algebra is a collection of all open or closed sets on $\mathbb{R} \cup \{-\infty, +\infty\}$ and all of their intersections, unions, and complements. Additionally, we will call an object a Borel set if it can be formed by countable union, intersection and complement of open sets of the form (a, b), $a, b \in \mathbb{R}$.

Now we introduce the concept of probability. Intuitively, probability should be associated with an event – when we say a probability of something, this 'something' is an event. Using the fact that the σ-algebra $\mathcal{F}$ is a collection of events and the property that $\mathcal{F}$ is measurable, we then introduce a measure called *probability measure* $\mathbb{P}(\cdot)$ that assigns a number between 0 and 1 to every element of $\mathcal{F}$. Namely, this function $\mathbb{P}$ maps an event to a number, describing the likelihood of the event.

Definition 1.2 *Let Ω be a sample space and $\mathcal{F}$ a σ-algebra. A probability measure is a mapping $\mathbb{P} : \mathcal{F} \mapsto \mathbb{R}$ satisfying the following three axioms*

(P1) $\mathbb{P}(\Omega) = 1$.

(P2) $\mathbb{P}(A) \geq 0$ *for all* $A \in \mathcal{F}$.

(P3) (countably additive) $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *for mutually exclusive events* $A_1, A_2, ... \in \mathcal{F}$.

The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ *is called a* probability space.

Theorem 1.3 *The three axioms (P1)-(P3) imply the following properties (among others):*

(1) $\mathbb{P}(\emptyset) = 0$,

(2) $0 \leq \mathbb{P}(A) \leq 1$,

(3) $A \subset B \Longrightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$,

(4) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,

(5) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

The countable additive property (P3) also implies that if a sequence of sets $A_1, A_2, \ldots$ in $\mathcal{F}$ satisfying $A_n \subseteq A_{n+1}$ for all n, then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

If $A_n \supseteq A_{n+1}$ for all n, then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

How do we interpret the concept probability? There are two major views in statistics. The first view is called the frequentist view – where a probability is interpreted as the limiting frequency observed over repetitions in identical situations. The other view is called the Bayesian view or the subjective view where the probability quantifies personal belief.

1.2 Random Variables

So far, we have built a mathematical model describing the probability and events. However, in reality, we are dealing with numbers, which may not be directly link to events. We need another mathematical notion that bridges the events and numbers and this is why we need to introduce random variables.

Informally, a *random variable* is a mapping $X : \Omega \mapsto \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$. For example, we toss a coin 2 times and let X represents the number of heads. The sample space is $\Omega = \{HH, HT, TH, TT\}$. Then for each $\omega \in \Omega$, $X(\omega)$ outputs a real number: $X(\{HH\}) = 2$, $X(\{HT\}) = X(\{TH\}) = 1$, and $X(\{TT\}) = 0$.

Definition 1.4 *Let Ω be a sample space and $\mathcal{F}$ a σ-algebra. A mapping $X : \Omega \to \mathbb{R}$ is a random variable (R.V.) if $X(\omega)$ is measurable with respect to $\mathcal{F}$, i.e.,*

$$X^{-1}((-\infty, c]) := \{\omega \in \Omega : X(\omega) \leq c\} \in \mathcal{F}, \quad \textit{for all } c \in \mathbb{R}.$$

Note that the condition in Definition 1.4 is equivalent to saying that $X^{-1}(B) \in \mathcal{F}$ for every Borel set B. This means that the set $X^{-1}(B)$ is indeed an event so that it makes sense to talk about $\mathbb{P}(X \in B)$, the probability that X lies in B, for any Borel set B. The function $B \mapsto \mathbb{P}(X \in B)$ is a probability measure and is called the *(probability) distribution* of X.

A very important characteristic of a random variable is its *cumulative distribution function (CDF)*, which is defined as follows

Definition 1.5 *The cumulative distribution function (CDF) of a random variable X, denoted by, $F_X(x)$ or $F(x)$, is defined by*

$$F(x) = P(X \leq x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}).$$

The inverse of the CDF $F^{-1}(x)$ is called the *quantile* function. The distribution of X is completely determined by the CDF $F(x)$, regardless of X being a discrete random variable or a continuous random variable (or a mix of them).

Theorem 1.6 *The function $F(x)$ is a CDF if and only if the following three conditions hold:*

(1) $\lim_{x\to-\infty} F(x) = 0$, *and* $\lim_{x\to+\infty} F(x) = 1$.

(2) $F(x)$ *is a nondecreasing function of* x.

(3) $F(x)$ *is right-continuous; that is, for every number* x_0, $\lim_{x\to x_0+} F(x) = F(x_0)$.

When X takes discrete values, we may characterize its distribution using the probability mass function (PMF):

$$p(x) = P(X = x) = F(x) - F(x^-),$$

where $F(x^-) = \lim_{\epsilon\to 0} F(x-\epsilon)$. In this case, one can recover the CDF from PMF using $F(x) = \sum_{x' \leq x} p(x')$.

If X is an absolutely continuous random variable, we may describe its distribution using the probability density function (PDF):

$$p(x) = F'(x) = \frac{d}{dx}F(x).$$

In this case, the CDF can be written as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(x')dx'.$$

However, the PMF and PDF are not always well-defined. There are situations where X does not have a PMF or a PDF. The formal definition of PMF and PDF requires the notion of the Radon-Nikodym derivative, which is beyond the scope of this course.

Example 1 (discrete). Suppose X takes only three possible values: $1, 2, 3$, with equal probabilities. Then the PMF $p(x) = \frac{1}{3}$ for $x = 1, 2, 3$ and $p(x) = 0$ otherwise. The CDF will be

$$F(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{3}, & 1 \leq x < 2 \\ \frac{2}{3}, & 2 \leq x < 3 \\ 1, & x \geq 3 \end{cases} \tag{1.1}$$

Example 2 (continuous). Consider a random variable X that has a uniform PDF over the interval $[1, 10]$. Namely, there is a constant c such that $p(x) = c$ for $x \in [1, 10]$ and $p(x) = 0$ otherwise. What will c be? Using the fact that $1 = \int p(x)dx$, you can easily see that $c = \frac{1}{9}$. What will the CDF be in this case? By definition,

$$F(x) = \int_{-\infty}^{x} p(u)du = \begin{cases} 0, & x \leq 1 \\ \frac{x-1}{9}, & 1 < x \leq 10 \\ 1, & x > 10. \end{cases}$$

Example 3 (no PDF or PMF). Consider a random variable X such that with a probability of 0.5, it always takes a fixed value 2 and with a probability of 0.5, it is from a uniform PDF over $[0, 1]$. In this case, can we define PDF or PMF? It turns out that this random variable X does not have a PDF (since it has a point mass at $x = 2$) but it has a strange PMF (that takes a value of 0 except at $x = 2$). So we cannot characterize its distribution well using the PDF or PMF. However, using the definition of CDF $F(x) = P(X \leq x)$, you can easily see that it has a well-defined CDF:

$$F(x) = P(X \leq x) = \begin{cases} 0, & x \leq 0 \\ \frac{x}{2}, & 0 < x \leq 1 \\ \frac{1}{2}, & 1 < x < 2 \\ 1, & x \geq 2. \end{cases}$$

If you take a more advanced probability theory course, you will find that the CDF is the formal definition of a distribution function–it is always well-defined unlike the PMF or PDF.

1.3 Common Distributions

1.3.1 Discrete Random Variables

Bernoulli. If X is a Bernoulli random variable with parameter p, then $X = 0$ or, 1 such that

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

In this case, we write $X \sim \mathsf{Ber}(p)$.

Binomial. If X is a binomial random variable with parameter (n, p), then $X = 0, 1, \cdots, n$ such that

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

In this case, we write $X \sim \mathsf{Bin}(n, p)$. Note that if $X_1, \cdots, X_n \sim \mathsf{Ber}(p)$, then the sum $S_n = X_1 + X_2 + \cdots + X_n$ is a binomial random variable with parameter (n, p).

Geometric. If X is a geometric random variable with parameter p, then

$$P(X = n) = (1-p)^{n-1} p$$

for $n = 1, 2, \cdots$. Geometric random variable can be constructed using 'the number of trials of the first success occurs'. Consider the case we are flipping coin with a probability p that we gets a head (this is a Bernoulli (p) random variable). Then the number of trials we made to see the first head is a geometric random variable with parameter p.

Poisson. If X is a Poisson random variable with parameter λ, then $X = 0, 1, 2, 3, \cdots$ and

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

In this case, we write $X \sim \mathsf{Poi}(\lambda)$. Poisson is often used to model a counting process. For instance, the intensity of an image is commonly modeled as a Poisson random variable.

1.3.2 Continuous Random Variables

Uniform. If X is a uniform random variable over the interval $[a, b]$, then

$$p(x) = \frac{1}{b-a} I(a \le x \le b),$$

where $I(\mathsf{statement})$ is the indicator function such that if the $\mathsf{statement}$ is true, then it outputs 1 otherwise 0. Namely, $p(x)$ takes value $\frac{1}{b-a}$ when $x \in [a, b]$ and $p(x) = 0$ in other regions. In this case, we write $X \sim \mathsf{Uni}[a, b]$.

If X is a uniform random variable over some surface or more generally object C, then the PDF of X is,

$$p(x) = \frac{1}{\text{volume}(C)} I(x \in C).$$

For any $A \subseteq C$, it holds that $P(X \in A) = \frac{\text{volume}(A)}{\text{volume}(C)}$.

Normal. If X is a normal random variable with parameter (μ, σ^2), then

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

In this case, we write $X \sim N(\mu, \sigma^2)$.

Exponential. If X is an exponential random variable with parameter λ, then X takes values in $[0, \infty)$ and

$$p(x) = \lambda e^{-\lambda x}.$$

In this case, we write $X \sim \mathsf{Exp}(\lambda)$. Note that we can also write

$$p(x) = \lambda e^{-\lambda x} I(x \geq 0).$$

Laplace, or double exponential. A *Laplace* or *double exponential* random variable X with parameter λ has the following PDF

$$p(x) = \frac{\lambda}{2} e^{-\lambda |x|}.$$

Cauchy. If $X \in \mathbb{R}$ is a Cauchy random variable with parameter μ, σ^2, then it has a PDF

$$p(x) = \frac{1}{\pi\sigma} \frac{1}{1 + (x-\mu)^2/\sigma^2}.$$

Interesting fact: the Cauchy distribution has *no* mean (average); the parameter μ is the median.

Gamma. A Gamma random variable $X \geq 0$ has two parameters $\alpha, \lambda > 0$ and has a PDF

$$p(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I(x \geq 0).$$

The function $\Gamma(\alpha) = \int x^{\alpha-1} e^{-x} dx$ is known as the Gamma function.

Beta. The Beta distribution is a continuous distribution on $[0, 1]$. So it is often used to model a ratio or a probability. If X is a Beta random variable with parameter α, β, then

$$p(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} I(0 \leq x \leq 1),$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

Logistic. The logistic distribution is a random variable whose CDF follows from the logit function. It has two parameter $\alpha \in \mathbb{R}, \beta > 0$ and has a CDF

$$F(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} = \frac{1}{1 + e^{-\alpha-\beta x}}.$$

The PDF is

$$p(x) = \frac{\beta e^{-\alpha-\beta x}}{(1 + e^{-\alpha-\beta x})^2} = \frac{\beta e^{\alpha+\beta x}}{(1 + e^{\alpha+\beta x})^2}$$

1.4 Random Vectors

We first need to define joint distribution functions, see Ch.4 of Casella and Berger (2021).

A multivariate random variable $(X_1, X_2, \ldots, X_n)$ is called a *random vector* (denoted rvtr in Perlman (2019)). The individual random variables in a random vector must arise in the same experiment (e.g., all features of a subject, height and weight), so they may or may not be *correlated* (later lecture). For simplicity, we consider random vectors of length two in the results below.

Definition 1.7 *The joint cumulative distribution function (joint CDF) of two random variables X and Y is defined as*

$$F_{XY}(x, y) = P(X \leq x, Y \leq y), \text{ for all } x, y.$$

Theorem 1.8 *For random variables X and Y with (marginal) CDFs $F_X(x)$ and $F_Y(y)$, the joint CDF $F_{XY}(x,y)$ satisfies the following*

(1) $F_X(x) = \lim_{y\to+\infty} F_{XY}(x,y)$, *for any* x.

(2) $F_Y(y) = \lim_{x\to+\infty} F_{XY}(x,y)$, *for any* y.

(3) $\lim_{x,y\to+\infty} F_{XY}(x,y) = 1$.

(4) $\lim_{x\to-\infty} F_{XY}(x,y) = \lim_{y\to-\infty} F_{XY}(x,y) = 0$.

(5) $P(x_1 < X \le x_2, y_1 < Y \le y_2) = F_{XY}(x_2,y_2) - F_{XY}(x_1,y_2) - F_{XY}(x_2,y_1) + F_{XY}(x_1,y_1)$.

When the vector is multivariate continuous (both variables are continuous), the corresponding joint PDF is

$$p_{XY}(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}.$$

A marginal PDF of X can be obtained from $p_{XY}(x,y)$ by integrating over y:

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x,y)dy.$$

When both X and Y are discrete (the random vector (X,Y) is disrete), the joint PMF is given by

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

A marginal PDF of X can be obtained from $p_{XY}(x,y)$ by summing over y where $p_X(x) = P(X = x) = \sum_y P(X = x, Y = y)$.

The joint distribution contains information about X and Y beyond their marginal distributions, i.e., information about their dependence. Thus, the joint distribution determines all marginal distributions but not conversely.

1.5 Conditional Probability

Now we have a basic mathematical model for probability. This model also defines an interesting quantity called conditional probability. For two events $A, B \in \mathcal{F}$, such that $P(B) \neq 0$, the conditional probability of A given B is

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

Note that when B is fixed, the function $\mathbb{P}(\cdot|B) : \mathcal{F} \mapsto \mathbb{R}$ is another probability measure.

In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$. This is sometimes called as the prosecutor's fallacy:

$$\mathbb{P}(\text{evidence}|\text{guilty}) \neq \mathbb{P}(\text{guilty}|\text{evidence}).$$

Example (Exponential). Let X be an exponential random variable with parameter $\lambda > 0$ and consider two positive numbers $x, y > 0$. What is the probability $P(X > x + y|X > y)$? In this case the two events

$A = \{X > x + y\}$ (formally, $A = \{\omega : X(\omega) > x + y\}$) and $B = \{X > y\}$. It is easy to see that $A \subset B$ so $A \cap B = A$. Thus,

$$P(X > x + y | X > y) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{P(X > x + y)}{P(X > y)}.$$

It is easy to see that for an exponential RV X, $P(X > y) = e^{-\lambda y}$, which implies

$$P(X > x + y | X > y) = \frac{P(X > x + y)}{P(X > y)} = e^{-\lambda x} = P(X > x).$$

Thus, the probability only depends on the *increment* x, not y. This is known as the *memoryless* property.

1.6 Conditional Distribution

When both variables in a random vector are continuous, the *conditional PDF* of Y given $X = x$ is

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)},$$

where $p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y)dy$ is the marginal density function of X.

When both X and Y are discrete, the conditional PMF of Y given $X = x$ is

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)},$$

where $p_X(x) = P(X = x) = \sum_y P(X = x, Y = y)$.

Example (triangle uniform). Consider two random variables (X, Y) that have a uniform PDF over the region $D = \{(x, y) : x \geq 0, y \geq 0, x + y \leq 1\}$. It is easy to see that $p(x, y) = 2$ when $(x, y) \in D$ and 0 otherwise. What is the conditional PDF of $p_{Y|X}(y|x)$? Because the joint PDF is a constant, one can easily see that $p_{Y|X}(y|x)$ will also be a constant. The key is to identify what is the feasible range of y when $X = x$. We have two constraints $y \geq 0$ and $y \leq 1 - x$ so the feasible range of y is $[0, 1 - x]$. Thus,

$$p_{Y|X}(y|x) = \frac{1}{1 - x} I(0 \leq y \leq 1 - x).$$

Example (Beta-Bernoulli). Consider two random variables $X \in \{0, 1\}$ and $Y \in [0, 1]$ such that given Y, the random variable X is a Bernoulli random variable with parameter $p = Y$. Namely,

$$P(X = 1|Y) = Y, \quad P(X = 0|Y) = 1 - Y.$$

Also, assume that Y follows a Beta distribution with parameters α, β. We are interested in the conditional distribution of Y given $X = x$.

The conditional PDF/PMF

$$p(x|y) = y^x (1 - y)^{1-x}.$$

Thus, the joint PDF/PMF

$$p(x, y) = p(x|y)p(y) = y^x (1 - y)^{1-x} \cdot \frac{1}{B(\alpha, \beta)} y^{\alpha-1} (1 - y)^{\beta-1}.$$

There are two methods we can now employ to compute $p(y|x)$:

- **Method 1**: Compute $p_X(x)$ from $p_{XY}(x, y)$ and then use conditional distribution formula to compute $p(y|x)$.
- **Method 2**: Use the "proportional" trick below.

Here is the trick: because $p(y|x) = \frac{p(x,y)}{p(x)} \propto p(x, y)$, we only need to focus on the part of $p(x, y)$ that involves y. The above product shows that

$$p(y|x) \propto p(x, y) \propto y^{\alpha+x-1}(1 - y)^{\beta-x}.$$

Therefore,

$$p(y|x) = C \cdot y^{\alpha+x-1}(1 - y)^{\beta-x},$$

where C is some constant in $\mathbb{R}$. Since we know that $p(y|x)$ is a density function, integrating it over all possible values of y must be equal to 1, we can use this information to compute C. After some computations, we can conclude that the distribution of Y conditional on $X = x$ is going to be Beta distribution with parameters $(\alpha' = \alpha + x, \beta' = \beta + (1 - x))$.

The Beta-Bernoulli example illustrates the fact that the conditioning operation can be viewed as an information flow. Suppose that Y is a variable of interest that is unobserved but it implicitly determines the distribution of X. And X is something that we can measure/observe (think of it as the data). Before seeing X, we place a model that Y is from a Beta distribution with parameter α, β. After observing X, this information should improve our knowledge about Y. A simple mathematical model to describe the improvement from the information is the conditional distribution. As is shown in the above example, the conditional distribution of Y given X is a Beta distribution with parameter $\alpha+X, \beta+(1-X)$. The change of parameter is an example of how the observed data X improves our understanding of an unobserved quantity Y.

1.7 Independence

Intuitively, when we say that two events are independent, we refer to the case that the two events will not interfere each other. Two events A and B are independent if

- $\mathbb{P}(A|B) = \mathbb{P}(A)$, or equivalently,
- $\mathbb{P}(B|A) = \mathbb{P}(B)$, or equivalently,
- $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

For three events A, B, C, we say events A and B are *conditional independent* given C if

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

Two random variables X and Y are independent under the model (Ω, P), denoted $X \perp Y$, or $X \perp\!\!\!\perp Y$, if $\{X \in A\}$ and $\{Y \in B\}$ are independent for each pair of events $A \subseteq \Omega_X$ and $B \subseteq \Omega_Y$.

Random variables X and Y are *independent* if the joint CDF can be factorized as

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

One can show that for a jointly discrete or jointly continuous random vector (X, Y), $X \perp\!\!\!\perp Y$ if and only if one of the following equivalent conditions hold

- $p_{XY}(x,y) = p_X(x)p_Y(y)$, for all $(x,y) \in \Omega_{XY}$;
- $p_{Y|X}(y|x) = p_Y(y)$, for all $(x,y) \in \Omega_{XY}$;
- $p_{X|Y}(x|y) = p_X(x)$, for all $(x,y) \in \Omega_{XY}$.

Independence **requires** that the *joint range* of (X,Y) is the Cartesian product of the marginal ranges:

$$\Omega_{XY} = \Omega_X \times \Omega_Y.$$

This is a necessary condition, but not a sufficient condition for independence!

Now we use the information interpretation (in the Beta-Bernoulli example) to think of the independence. The independence implies $p_{Y|X}(y|x) = p(y)$, which can be interpreted that *knowing X does NOT change the distribution of Y*. This is essentially what an intuitive meaning of independence should be–knowing the outcome of one variable does not provide any information about another variable.

When we have many random variables $X_1, \cdots, X_n$, they are (mutually) independent if the joint CDF

$$F(x_1, x_2, \cdots, x_n) = F(x_1)F(x_2)\cdots F(x_2),$$

which implies

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2)\cdots p(x_n).$$

Example (Uniform on a disk). Consider two random variables X, Y such that they jointly follow from a distribution that is uniform over the unit disk $S_0 = \{(x,y) : x^2 + y^2 \leq 1\}$. Clearly, X and Y are not independent because when $X = 0$, the feasible range of Y is $[-1,1]$ while when $X = 1$, the only possible value of Y is 0.

Now suppose we reparametrize the two random variables using polar coordinates $(R, \Theta) \in [0,1] \times [0, 2\pi]$. Then $F(r,\theta) = P(R \leq r, \Theta \leq \theta)$, and note that since we know the distribution is uniform on a disc we have that

$$\begin{aligned}
F(r,\theta) &= P(R \leq r, \Theta \leq \theta) \\
&= \frac{\text{area of the sector defined by } R \leq r \text{ and } \Theta \leq \theta}{\text{total area of the disc}} \\
&= \frac{1}{\pi} \cdot \pi r^2 \cdot \frac{\theta}{2\pi} \\
&= r^2 \cdot \frac{\theta}{2\pi} \\
&= F_R(r)F_\Theta(\theta), \\
F_R(r) &= r^2, \quad 0 \leq r \leq 1 \\
F_\Theta(\theta) &= \frac{\theta}{2\pi}, \quad 0 \leq \theta \leq 2\pi.
\end{aligned}$$

So $R \perp\!\!\!\perp \Theta$, i.e., they are independent (see also example 1.12 of Perlman (2019).)

Example (Independence and information). In the Beta-Bernoulli example, we have seen a probabilistic approach to infer an unobserved variable Y using the information from another random variable X. That idea is something related to *Bayesian inference*. Here we will introduce another approach to infer an unobserved quantity θ without assuming that θ is random. Suppose we observe $X_1, \cdots, X_n \in \{0,1\}$ that are independent. We assume that they are all from the same Bernoulli distribution with an unknown parameter $\theta_0 = P(X_i =$

1). In this case, we say $X_1, \cdots, X_n$ are IID (independently and identically distributed). Given $X_1, \cdots, X_n$, how do we infer θ_0? Under a probabilistic model, any parameter θ would imply a joint PMF

$$p(x_1, \cdots, x_n; \theta) = p(x_1; \theta)p(x_2; \theta) \cdots p(x_n; \theta)$$

due to the independence. Since it is a product term, we take a logarithm, which leads to

$$\log p(x_1, \cdots, x_n; \theta) = \log p(x_1; \theta) + \log p(x_2; \theta) + \cdots + \log p(x_n; \theta).$$

Since $X_1, \cdots, X_n$ are observed, we can view the above function as a function of θ, and this function is known as the log-likelihood function

$$\underbrace{\ell(\theta|X_1, \cdots, X_n)}_{\text{Total information}} = \log p(X_1, \cdots, X_n; \theta) = \sum_{i=1}^{n} \log p(X_i; \theta) = \sum_{i=1}^{n} \underbrace{\ell(\theta|X_i)}_{\text{Information of the } i\text{-th obs.}}.$$

Informally, we can call $\ell(\theta|X_1, \cdots, X_n)$ as the total information from $X_1, \cdots, X_n$ on θ. The independence assumption implies the above equality, which means that *under independence, the total information is the addition of all individual information.* In the *likelihood framework*, information about θ is determined by the log-likelihood function (Total information term). Note that unlike the Beta-Bernoulli example, here we did not specify any distribution of θ–it is just an unknown quantity and we use the likelihood function to infer plausible value of it. The famous *maximal likelihood estimator* (MLE) finds an estimated value of θ by maximizing the log-likelihood value.

1.8 Total probability and the Bayes theorem

Probability measure also has a useful property called *law of total probability.* If $B_1, B_2, ..., B_k$ forms a partition of Ω, then

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

In particular, $\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)$. And this further implies the famous *Bayes rule*: Let $A_1, ..., A_k$ be a partition of Ω. If $\mathbb{P}(B) > 0$ then, for $i = 1, ..., k$:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{k} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

For random variables, we also have the Bayes theorem:

$$\begin{aligned} p_{X|Y}(x|y) &= \frac{p_{XY}(x, y)}{p_Y(y)} \\ &= \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} \\ &= \begin{cases} \frac{p_{Y|X}(y|x)p_X(x)}{\int p_{Y|X}(y|x')p_X(x')dx'}, & \text{if } X, Y \text{ are absolutely continuous.} \\ \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')}, & \text{if } X, Y \text{ are discrete.} \end{cases} \end{aligned}$$

Example (Poisson-Binomial). Consider two random variables X and Y such that $X \sim \mathsf{Poisson}(\lambda)$ and $Y|X = x$ is from a Binomial distribution with parameters (X, p). What will the marginal distribution of Y

be? To study this, we attempt to compute the probability $P(Y = y)$.

$$\begin{aligned} P(Y = y) &= \sum_x P(Y = y, X = x) \\ &= \sum_{x \geq y} P(Y = y | X = x) P(X = x) \quad \text{(Think about why } x \geq y\text{)} \\ &= \sum_{x \geq y} \binom{x}{y} p^y (1-p)^{x-y} \frac{\lambda^x e^{-\lambda}}{x!}. \end{aligned}$$

Using the fact that $\binom{x}{y} = \frac{x!}{(x-y)!y!}$ and set $k = x - y$, we can rewrite the above as

$$\begin{aligned} P(Y = y) &= \sum_{x \geq y} \binom{x}{y} p^y (1-p)^{x-y} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= p^y e^{-\lambda} \sum_{x \geq y} \frac{x!}{(x-y)!y!} (1-p)^{x-y} \lambda^x \frac{1}{x!} \\ &= \frac{p^y e^{-\lambda}}{y!} \sum_{k=0}^{\infty} \frac{1}{k!} (1-p)^k \lambda^{y+k} \\ &= \frac{(\lambda p)^y e^{-\lambda p}}{y!} \underbrace{\sum_{k=0}^{\infty} \frac{1}{k!} (1-p)^k \lambda^k e^{-\lambda(1-p)}}_{=1}, \end{aligned}$$

which is the PMF of Poisson(λp). Thus, Y follows from a Poisson distribution with parameter λp.

1.9 Conditional independence

For three RVs X, Y, and Z, we say X, Y are conditional independent given Z if

$$P(X \leq x, Y \leq y | Z = z) = P(X \leq x | Z = z) P(Y \leq y | Z = z)$$

for every x and y and P_Z-almost everywhere of z. P_Z-almost everywhere of z means that the above equality holds for all z except for a set of values that has 0 probability. It is a slightly weaker notion than 'for every z'. We use the notation

$$X \perp\!\!\!\perp Y | Z$$

for denote the case where X, Y are conditional independent given Z.

Note that $X \perp\!\!\!\perp Y | Z$ also implies

$$P(X \leq x | Y = y, Z = z) = P(X \leq x | Z = z)$$

for every x and $P_{Y,Z}$-almost everywhere of (y, z).

Beware! Independence is not the same as conditional independence, i.e., $X \perp\!\!\!\perp Y \not\Rightarrow X \perp\!\!\!\perp Y | Z$.

Example (conditional independence $\not\Rightarrow$ indepedence). Assume $X \perp\!\!\!\perp Y | Z$ and $Z \in \{0, 1\}$ such that when $Z = 0$, X and Y are both from a uniform distribution over $[0, 1]$ and when $Z = 1$, X and Y are from a uniform distribution over $[2, 3]$. Assume that Z has an equal probability of being 0 or 1. Marginally, both X and Y are from a uniform distribution over the set $[0, 1] \cup [2, 3]$. However, if we observe $X = 2.5$, we know that Y has to be from a uniform distribution over $[2, 3]$ so $P(Y \in [0, 1] | X = 2.5) = 0 \neq P(Y \in [0, 1]) = 0.5$.

The following is a theorem about different ways of saying *conditional independence*.

Theorem 1.9 *Let p_{XYZ} be the joint PDF/PMF of X, Y, and Z. Then the followings are equivalent:*

(i) $X \perp\!\!\!\perp Y|Z$.

(ii) $p_{XY|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$ *almost everywhere (a.e.).*

(iii) $p_{X|YZ}(x|y,z) = p_{X|Z}(x|z)$ *a.e.*

(iv) $p_{XYZ}(x,y,z) = \frac{p_{XZ}(x,z)p_{YZ}(y,z)}{p_Z(z)}$ *a.e.*

(v) $p_{XYZ}(x,y,z) = g(x,z)h(y,z)$, *where g and h are some (measurable) functions.*

(vi) $p_{X|YZ}(x|y,z) = w(x,z)$, *where w is some (measurable) function.*

Proof: The equivalence between (i), (ii), (iii), and (iv) are trivial so we focus on case (v) and (vi).

(ii) $\Rightarrow$ (v):
Because

$$p_{XY|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z),$$

we have

$$\frac{p_{XYZ}(x,y,z)}{p_Z(z)} = \frac{p_{XZ}(x,z)}{p_Z(z)}\frac{p_{YZ}(y,z)}{p_Z(z)}$$

so

$$p_{XYZ}(x,y,z) = \frac{p_{XZ}(x,z)p_{YZ}(y,z)}{p_Z(z)} = g(x,z)h(y,z),$$

for $g(x,z) = \frac{p_{XZ}(x,z)}{p_Z(z)}$ and $h(y,z) = p_{YZ}(y,z)$, for instance. Hence, (v) holdes..

(v) $\Rightarrow$ (vi):
Based on (v), we have

$$p_{YZ}(y,z) = \int p_{XYZ}(x,y,z)dx = h(y,z)\int g(x,z)dx = h(y,z)q(z).$$

Thus,

$$p_{X|YZ}(x|y,z) = \frac{p_{XYZ}(x,y,z)}{p_{YZ}(y,z)} = \frac{g(x,z)h(y,z)}{h(y,z)q(z)} = \frac{g(x,z)}{q(z)} = w(x,z).$$

Finally, we show that (vi) $\Rightarrow$ (iii):

$$\begin{aligned} p_{X|Z}(x|z) = \int p_{XY|Z}(x,y|z)dy &= \int p_{X|YZ}(x|y,z)p_{Y|Z}(y|z)dy \\ &= w(x,z)\int p_{Y|Z}(y|z)dy = w(x,z) = p_{X|YZ}(x|y,z). \end{aligned}$$

■

Here are five important properties of conditional independence. For more details see Dawid (1979), or Chapter 3.1 of Lauritzen (1996).

Theorem 1.10 *Let X, Y, Z, W be RVs, the following properties hold:*

(C1) (symmetry) $X \perp\!\!\!\perp Y|Z \iff Y \perp\!\!\!\perp X|Z$.

(C2) (decomposition) $X \perp\!\!\!\perp Y|Z \implies h(X) \perp\!\!\!\perp Y|Z$ *for any (measurable) function* h.
A special case is: $(X, W) \perp\!\!\!\perp Y|Z \implies X \perp\!\!\!\perp Y|Z$.

(C3) (weak union) $X \perp\!\!\!\perp Y|Z \implies X \perp\!\!\!\perp Y|Z, h(X)$ *for any (measurable) function* h.
A special case is: $(X, W) \perp\!\!\!\perp Y|Z \implies X \perp\!\!\!\perp Y|(Z, W)$

(C4) (contraction)

$$X \perp\!\!\!\perp Y|Z \text{ and } X \perp\!\!\!\perp W|(Y, Z) \iff X \perp\!\!\!\perp (W, Y)|Z.$$

(C5) (intersection) If the joint PDF $p_{XYZW}(x, y, z, w)$ *is positive almost everywhere, then*

$$X \perp\!\!\!\perp Y|(W, Z) \text{ and } X \perp\!\!\!\perp W|(Y, Z) \iff X \perp\!\!\!\perp (W, Y)|Z.$$

(C5) can be generalized beyond positive densities, but it gets rather technical.

References

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Perlman, M. D. (2019). Probability and mathematical statistics i. https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf.

STAT 512: Statistical Inference **Original notes credit: Yen-Chi Chen**

Lecture 2: Transforming continuous random variables

Instructor: Emilija Perković

Useful additional reading: Chapter 2.1 of Casella and Berger (2021), Chapter 2 of Perlman (2019).

Useful recall: derivative rules (product, quotient, chain rule), using chain rule to compute the derivative of the inverse function. Taylor and Maclaurin series. Various representations of e.

In the previous lecture, we have seen a couple of distributions that have nice properties. When working with data, we may perform some transformation of random variables. Suppose we know the distribution of a random variable before the transformation, does this give us any hint on the distribution of the transformed variable?

2.1 One function of one random variable

Let X be a continuous random variable whose PDF $p_X(x)$ is known. Consider a given function f and another random variable $Y = f(X)$. Since the input X is random, the output Y is often random as well. What will the distribution of Y be?

When f is differentiable, we have the following useful theorem. While the below result is useful, I would suggest remembering the proof rather than the result itself.

Lemma 2.1 *Let X be a continuous r.v. on $[a, b]$, $a, b \in \mathbb{R}$ with PDF p_X. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous, strictly increasing and differentiable function, then the PDF of $Y = g(X)$ is*

$$p_Y(y) = \begin{cases} \frac{p_X(g^{-1}(y))}{g'(g^{-1}(y))}, & g(a) \le y \le g(b) \\ 0, & \text{otherwise.} \end{cases}$$

Proof:

To start with, we consider the CDF of Y:

$$\begin{aligned} P(Y \le y) &= P(g(X) \le y) \\ &= P(X \le g^{-1}(y)). \end{aligned}$$

Note that since $g(x)$ is strictly increasing and continuous, $g^{-1}(y)$ exists. The PDF will be the derivative of the CDF, leading to

$$\begin{aligned} p_Y(y) &= \frac{d}{dy} P(Y \le y) \\ &= \frac{d}{dy} P(X \le g^{-1}(y)) \\ &= p_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) \\ &= \frac{p_X(g^{-1}(y))}{g'(g^{-1}(y))}, \end{aligned}$$

which completes the proof. Note that we can obtain that $(g^{-1}(y))' = \frac{1}{g'(g^{-1}(y))}$, by applying the derivative chain rule to $g \circ g^{-1}(y) = y$ and solving for $(g^{-1}(y))'$.

■

Example. $X \sim Exp(\lambda)$, and $Y = \sqrt{X}$. What is the density of Y? **Hint:** Use the above lemma.

When $g(x)$ is not strictly increasing, but is instead strictly decreasing a similar result to above can be derived. Let's consider the general case when $g(x)$ is strictly monotonic (can be either increasing or decreasing).

Theorem 2.2 *Let X be a continuous r.v. with PDF p_X. Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable and strictly monotonic function with inverse denoted $\gamma = g^{-1}$, then the p.d.f of $Y = g(X)$ is*

$$p_Y(y) = \begin{cases} |\gamma'(y)| p_X(g^{-1}(y)) & y \in g(\mathbb{R}), \\ 0 & \text{otherwise} \end{cases}$$

where $\gamma'(y) = \frac{1}{g'(g^{-1}(y))}$.

Proof: Denote $a = \inf_x(g(x))$, $b = \sup_x(g(x))$, (possibly $a = -\infty$, or $b = +\infty$) If $t < a$, $F_Y(t) = P(g(X) \leq a) = 0$ so $f_Y(t) = 0$. If $t > b$, $F_Y(t) = P(g(X) \leq t) = 1$ so $f_Y(t) = 0$.

If g is strictly increasing, for $t \in (a, b)$, then $\gamma(t) := g^{-1}(t)$ is defined,

$$\begin{aligned} & F_Y(t) = P(Y \leq t) = P(g(X) \leq t) = P(X \leq g^{-1}(t)) = F_X(\gamma(t)) \\ \text{so} \quad & p_Y(t) = \gamma'(t) p_X(\gamma(t)). \end{aligned}$$

If g is strictly decreasing, for $t \in (a, b)$, then $g^{-1}(t)$ is defined,

$$\begin{aligned} & F_Y(t) = P(Y \leq t) = P(g(X) \leq t) = P(X \geq g^{-1}(t)) = 1 - F_X(\gamma(t)) \\ \text{so} \quad & p_Y(t) = -\gamma'(t) p_X(\gamma(t)) \end{aligned}$$

For $t \in (a, b)$, $g \circ g^{-1}(t) = t$ so $\gamma'(t) = \frac{1}{g'(g^{-1}(t))}$ so $\gamma'(t) < 0$ for g decreasing.

■

Note that if g is not defined on $\mathbb{R}$, but rather, $g : B \to \mathbb{R}$, for some $B \subseteq \mathbb{R}$, we can still use the above theorem for $y \in g(B)$ as long as $P(X \in B) = 1$. We will have that $p_Y(y) = 0$, for $y \notin g(B)$.

Example. Assume $X \sim \mathsf{Uniform}[1, e]$ and consider that we are interested in the PDF of $Y = -2 \log X$. Here, $g(x) = -2 \log(x)$, where $x \in [1, e]$. In this case, $g'(x) = -\frac{2}{X}$, so $g'(x) < 0$, for $x \in [1, e]$, hence, g is strictly decreasing.

Moreover, for $x = 1$, $-2 \log(x) = 0$, and for $x = e$, $-2 \log(x) = -2$. So the range of possible values for Y is $[-2, 0]$.

Note also that $p_X(x) = \frac{1}{e-1} I(1 \leq x \leq e)$ and $g^{-1}(y) = e^{-\frac{1}{2}y}$. We now apply Theorem 2.2. Then the PDF of Y will be

$$p_Y(y) = \frac{1}{2(e-1)} e^{-\frac{1}{2}y} I(-2 \leq y \leq 0).$$

How do we approach the case when $g(\cdot)$ is not a strictly monotonic function? → **Split the domain of X into segments on which $g(X)$ is monotonic. Use law of total probability!**

Example. Consider $X \sim N(0,1)$ and $Y = X^2$. What is the distribution of Y? Note that the underlying transformation $g : x \to x^2$ is not monotonic (and also not invertible) as a function from $\mathbb{R}$ to $\mathbb{R}^+$.

However, g is monotonic (and invertible) when restricted to $\mathbb{R}^+$ or $\mathbb{R}^- \setminus \{0\}$, i.e.,

- if $x \geq 0$ and $x^2 = t$ then $x = \sqrt{t}$,
- if $x < 0$ and $x^2 = t$ then $x = -\sqrt{t}$.

First, clearly $P(Y \leq t) = 0$ for $t \leq 0$, so we'll consider $t > 0$. Let us partition $\mathbb{R}$ as $\mathbb{R} = \mathbb{R}^+ \cup (\mathbb{R}^- \setminus \{0\})$. Then by law of total probability

$$\begin{aligned} P(Y \leq t) &= P(X^2 \leq t \text{ and } X \geq 0) + P(X^2 \leq t \text{ and } X < 0) \\ &\overset{(*)}{=} P(X \leq \sqrt{t} \text{ and } X \geq 0) + P(-\sqrt{t} \leq X \text{ and } X < 0) \\ &= P(0 \leq X \leq \sqrt{t}) + P(-\sqrt{t} \leq X < 0) \\ &= F_X(\sqrt{t}) - F_X(0) + F_X(0) - F_X(-\sqrt{t}) \end{aligned}$$

where in $(*)$ we carefully used that g is strictly decreasing on $\mathbb{R}^-$ for the second term.

We then get that

$$p_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} p_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} p_X(-\sqrt{y}) & \text{if } y \geq 0 \\ 0 & \text{if } y \leq 0 \end{cases}$$

Replacing $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ into the above equation, we obtain

$$p_Y(y) = \begin{cases} \frac{1}{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-y/2} = \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{1}{2}y} & \text{if } y \geq 0 \\ 0 & \text{if } y \leq 0 \end{cases}$$

which implies $Y \sim \mathsf{Gamma}(\frac{1}{2}, \frac{1}{2})$. Note: Gamma $(\frac{1}{2}, \frac{1}{2})$ is the same as χ^2_1, the chi-squared distribution with degree of freedom 1.

Example. (Work it out on your own.) Suppose that Y is a continuous random variable with CDF F_Y and X is a uniform random variable within $[0,1]$. Then you can show that $Z = F_Y^{-1}(X)$ has a CDF $F_Z(z) = F_Y(z)$.

2.2 One function of two or more random variables

In practice, we may encounter problems involving a function of two or more random variables. Namely, suppose we have access to a random vector (X, Y) whose joint CDF $F_{XY}(x, y)$ is known and we are interested in the distribution of another random variable $U = g(X, Y)$ for some given function g. In this case, a general strategy again is to investigate the underlying CDF and take the derivative to obtain the corresponding PDF. Note that first you need to determine the range of possible values for U (drawing is helpful). We will illustrate the idea via a few examples.

Example. Consider random vector (X, Y) that is uniform distributed over $[0,1] \times [0,1]$ and that random variables X and Y both have $\mathsf{Uniform}([0,1])$ distributions. We will use that X and Y are independent random variables in this case, but proving that is left as an exercise. We are interested in deriving the distribution of random variable $U = g(X, Y)$.

- **Case 1:** $U = X + Y$**.**

  To find the possible values of U, i.e., $\{U = X + Y \leq u\}$ we need to find intersect regions $[0,1] \times [0,1]$ and $x + y \leq u$. This intersection is empty for $u \leq 0$ and is equal to the region $[0,1] \times [0,1]$, when $u \geq 2$. Hence, $P(U \leq u) = 0$, for $u < 0$, and $P(U \leq u) = 1$, for $u \geq 2$.

  When $u \in [0,2]$, we can easily work out $P(U \leq u)$ as the area of the intersecting region. (Need to know how to compute an area of a rectangle, and a triangle.)

$$F_U(u) = P(U \leq u)$$
$$= \begin{cases} 0, & u < 0 \\ u^2/2, & 0 \leq u \leq 1 \\ 1 - (2-u)^2/2, & 1 \leq u \leq 2 \\ 1, & u > 2 \end{cases}.$$

  The PDF $p_U(u)$ will be

$$p_U(u) = \begin{cases} 0, & u < 0 \\ u, & 0 \leq u \leq 1 \\ 2 - u, & 1 \leq u \leq 2 \\ 0, & u > 2 \end{cases}.$$

- **Case 2:** $U = \max\{X, Y\}$**.** A common trick to compute the distribution of a maximum of two or more independent random variables is based on the following insight:

$$\{\max\{X, Y\} \leq u\} \equiv \{X \leq u, Y \leq u\}.$$

  Therefore,

$$F_U(u) = P(U \leq u) = P(\max\{X, Y\} \leq u) = P(X \leq u, Y \leq u) = P(X \leq u)P(Y \leq u),$$

  which implies $F_U(u) = u^2$ and $p_U(u) = 2u$ when $u \in [0,1]$.

- **Case 3:** $U = \min\{X, Y\}$**.** The case of minimum is similar to the case of maximal but we will use analogous reasoning to Case 2, by relying on the following insight:

$$\{\min\{X, Y\} > u\} \equiv \{X > u, Y > u\}.$$

  Therefore,

$$1 - F_U(u) = P(U > u) = P(\min\{X, Y\} > u) = P(X > u, Y > u) = P(X > u)P(Y > u) = (1-u)^2,$$

  Thus, $F_U(u) = 1 - (1-u)^2$ so $p_U(u) = 2 - 2u$ for $u \in [0,1]$.

Example (minimum of many uniforms). Now consider $X_1, \cdots, X_n$ that are IID $\mathsf{Uniform}([0,1])$. Define $U = n \min\{X_1, \cdots, X_n\}$. What will the distribution of U be when n is large? Using the trick that we have discussed,

$$\{\min\{X_1, \cdots, X_n\} > \frac{u}{n}\} \equiv \{X_1 > \frac{u}{n}, \cdots, X_n > \frac{u}{n}\},$$

so

$$1 - F_U(u) = P\left(\min\{X_1, \cdots, X_n\} > \frac{u}{n}\right) = \prod_{i=1}^{n} P\left(X_i > \frac{u}{n}\right) = \left(1 - \frac{u}{n}\right)^n \to e^{-u}.$$

As a result, $F_U(u) \to 1 - e^{-u}$ and $p_U(u) \to e^{-u}$ so when n is large, U behaves like an exponentially distributed random variable.

Example (exponential distributions). Consider $X, Y \overset{IID}{\sim} \mathsf{Exponential}(1)$.

- **Sum of two exponentials.** What is the distribution of $U = X + Y$? A simple trick is to fixed one variable at a time and make good use of integration. Specifically, for a given $u > 0$,

$$\begin{aligned} F_U(u) &= P(U \le u) \\ &= P(X + Y \le u) \\ &= \int_{x+y \le u} e^{-x-y} dx dy \\ &= \int_{x=0}^{u} \int_{y=0}^{u-x} e^{-x-y} dy dx \\ &= \int_{x=0}^{u} e^{-x}(1 - e^{x-u}) dx \\ &= 1 - e^{-u} - u e^{-u}. \end{aligned}$$

  Thus, $p_U(u) = u e^{-u}$.

- **Minimum of two exponentials.** Now we consider $V = \min\{X, Y\}$. Using the same trick as the minimum of many uniforms, i.e.,

$$\{\min\{X, Y\} > v\} \equiv \{X > v, Y > v\}$$

  so

$$1 - F_V(v) = P(X > v)P(Y > v) = e^{-2v},$$

  which implies that $V \sim \mathsf{Exp}(2)$. In fact, you can easily generalize it to showing that if $X_1, \cdots, X_n \sim \mathsf{Exp}(\lambda)$, then $\min\{X_1, \cdots, X_n\} \sim \mathsf{Exp}(n\lambda)$.

- **Difference.** Consider

$$Z = \max\{X, Y\} - \min\{X, Y\} = |X - Y|.$$

  What will the distribution of Z be?

  Using a direct computation, we see that

$$\begin{aligned} F_Z(z) &= P(Z \le z) \\ &= P(|X - Y| \le z) \\ &= P(-z \le X - Y \le z) \\ &= P(X - Y \le z) - P(X - Y < -z) \\ &= P(X \le Y + z) - 1 + P(X - Y \ge -z) \\ &= -1 + P(X \le Y + z) + P(Y \le X + z) \\ &= -1 + 2P(X \le Y + z) \qquad X, Y \text{ are identically distributed.} \end{aligned}$$

  Moreover,

$$\begin{aligned} P(X \le Y + z) &= \int_{y=0}^{\infty} \int_{x=0}^{y+z} e^{-x} dx e^{-y} dy \\ &= \int_{y=0}^{\infty} (1 - e^{-y-z}) e^{-y} dy \\ &= 1 - e^{-z} \int_{0}^{\infty} e^{-2y} dy \\ &= 1 - \frac{1}{2} e^{-z}. \end{aligned}$$

As a result,

$$F_Z(z) = -1 + 2P(X \le Y + z) = 1 - e^{-z},$$

which is the CDF of $\mathsf{Exp}(1)$! This is another *memoryless property*.

- **Ratio.** Lastly, we consider $W = \frac{X}{X+Y}$ and studies its distribution. Clearly, $0 \le w \le 1$ so we will focus on the range $[0, 1]$.

$$\begin{aligned}
F_W(w) &= P\left(\frac{X}{X+Y} \le w\right) \\
&= P(X \le w(X+Y)) \\
&= P((1-w)X \le wY) \\
&= P\left(X \le \frac{w}{1-w} Y\right) \\
&= \int_{y=0}^{\infty} \int_{x=0}^{\frac{w}{1-w} y} e^{-x} dx e^{-y} dy \\
&= \int_{y=0}^{\infty} (1 - e^{\frac{-w}{1-w} y}) e^{-y} dy \\
&= 1 - \int_{0}^{\infty} e^{-\frac{1}{1-w} y} dy \\
&= 1 - 1 + w = w.
\end{aligned}$$

Thus $W \sim \mathsf{Unif}[0, 1]$.

Useful properties about normal (please verify them).

- Let $X \sim N(\mu_1, \sigma^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent. Then

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Also, for any real number a,

$$aX \sim N(a\mu_1, a^2\sigma_1^2).$$

- Let $X_1, \cdots, X_n$ be IID normal random variables from $N(\mu, \sigma^2)$. Then the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n).$$

- Let $X_1, \cdots, X_n$ be IID $N(0, 1)$. Then $Z_1 = X_1^2$ follows the χ^2 distribution with a degree of freedom 1. And $Z_n = \sum_{i=1}^{n} X_i^2$ follows the χ^2 distribution with a degree of freedom n.

References

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Perlman, M. D. (2019). Probability and mathematical statistics i. https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf.

STAT 512: Statistical Inference **Original notes credit: Yen-Chi Chen**

Lecture 3: Expectation and basic asymptotic theories

Instructor: Emilija Perković *Compiled on: 2023-11-01, 11:14:26*

Useful Additional Reading: Chapters 2 and 3 of Casella and Berger (2021). Chapter 3 of Perlman (2019). For the super motivated Section 25 of Billingsley (2017).

Useful recall: Taylor expansions, representations of e, rates of convergence, little o notation, computing the minimum of $f(x)$.

3.1 Expectation

For a function $g(x)$, the expectation of $g(X)$ is

$$\mathbb{E}(g(X)) = \int g(x)dF(x) = \begin{cases} \int_{-\infty}^{\infty} g(x)p(x)dx, & \text{if } X \text{ is continuous} \\ \sum_x g(x)p(x), & \text{if } X \text{ is discrete} \end{cases}.$$

In the simplest case $g(x) = x$,

$$\mathbb{E}(X) = \int xdF(x) = \begin{cases} \int_{-\infty}^{\infty} xp(x)dx, & \text{if } X \text{ is continuous} \\ \sum_x xp(x), & \text{if } X \text{ is discrete} \end{cases}.$$

is known as the the mean (expectation) of a R.V. X. Let $\mu = \mathbb{E}(X)$, the variance of X is $\mathsf{Var}(X) = \mathbb{E}((X-\mu)^2) = \mathbb{E}((X - \mathbb{E}(X))^2)$. The mean is a common measure of the "center" of a distribution and the variance is a common measure of the spread of a distribution.

The m-th moment of a random variable X is

$$\mathbb{E}(X^m).$$

Let $\mu = \mathbb{E}(X)$ be the mean/first moment of X, the m-th *centered* moment of X is

$$\mathbb{E}((X-\mu)^m).$$

Thus, the variance is the second centered moment.

Example.

- $X \sim \mathsf{Binomial}(n, p)$. Then $\mathbb{E}(X) = np$ and $\mathsf{Var}(X) = np(1-p)$.
- $X \sim \mathsf{Geometric}(p)$. Then $\mathbb{E}(X) = 1/p$ and $\mathsf{Var}(X) = (1-p)/p^2$.
- $X \sim \mathsf{Poisson}(\lambda)$. Then $\mathbb{E}(X) = \lambda$ and $\mathsf{Var}(X) = \lambda$.
- $X \sim \mathsf{Normal}(\mu, \sigma^2)$. Then $\mathbb{E}(X) = \mu$ and $\mathsf{Var}(X) = \sigma^2$.
- $X \sim \mathsf{Exponential}(\lambda)$. Then $\mathbb{E}(X) = 1/\lambda$ and $\mathsf{Var}(X) = 1/\lambda^2$.
- $X \sim \mathsf{Gamma}(\alpha, \lambda)$. Then $\mathbb{E}(X) = \alpha/\lambda$ and $\mathsf{Var}(X) = \alpha/\lambda^2$.
- $X \sim \mathsf{Beta}(\alpha, \beta)$. Then $\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$ and $\mathsf{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

- $X \sim \mathsf{Uniform}(a, b)$. Then $\mathbb{E}(X) = (a+b)/2$ and $\mathsf{Var}(X) = (b-a)^2/12$.

Linearity and decomposability of the expectation:

$$\mathbb{E}\left(\sum_{j=1}^{k}(a_j g_j(X) + b_j)\right) = \sum_{j=1}^{k}(a_j \cdot \mathbb{E}(g_j(X)) + b_j),$$

where a_j, b_j, $j \in \{1, \dots k\}$ are constants. Note that the above equality holds even if $g_{j_1}(X)$ and $g_{j_2}(X)$ are dependent.

When a set of mutually random variables $X_1, \cdots, X_n$ are independent, then

$$\mathbb{E}\left(X_1 \cdot X_2 \cdots X_n\right) = \mathbb{E}(X_1) \cdot \mathbb{E}(X_2) \cdots \mathbb{E}(X_n).$$

In fact, for a set of mutually independent random variables, $X_1, \cdots, X_n$

$$\mathbb{E}\left(g_1(X_1) \cdot g_2(X_2) \cdots g_n(X_n)\right) = \mathbb{E}(g_1(X_1)) \cdot \mathbb{E}(g_2(X_2)) \cdots \mathbb{E}(g_3(X_n)).$$

Aside: For independent random variables X_1 and X_2, $X_1 \perp X_2$ and functions $f, g : \mathbb{R} \to \mathbb{R}$ it holds: $f(X_1) \perp g(X_2)$. Similarly, for random variables X_1 and X_2 that are conditionally independent given rv X_3, $X_1 \perp X_2 | X_3$ and functions $f, g : \mathbb{R} \to \mathbb{R}$ it holds: $f(X_1) \perp g(X_2) | X_3$.

For two random variables X and Y with means being $\mathbb{E}(X) = \mu_x$ and $\mathbb{E}(Y) = \mu_y$ and variance being σ_x^2 and σ_y^2. The *covariance*

$$\mathsf{Cov}(X, Y) = \mathbb{E}((X - \mu_x)(Y - \mu_y)) = \mathbb{E}(XY) - \mu_x \mu_y$$

and the (Pearson's) correlation

$$\rho(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sigma_x \sigma_y}.$$

When two R.V. are not independent, we have

$$\mathsf{Var}(X \pm Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) \pm 2\mathsf{Cov}(X, Y).$$

The independence implies the covariance (and correlation) is 0, i.e.,

$$X \perp Y \Rightarrow \mathsf{Cov}(X, Y) = 0.$$

As a result, if $X \perp Y$,

$$\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y).$$

A more general result is that for independent random variables $X_1, \cdots, X_n$, we have

$$\mathsf{Var}\left(\sum_{i=1}^{n}(a_i X_i + b_i)\right) = \sum_{i=1}^{n} a_i^2 \cdot \mathsf{Var}(X_i),$$

for constants a_i, b_i, $i \in \{1, \dots, n\}$.

Example (Binomial). Here we illustrate how the above properties can be useful in computing the variance of some distributions. Consider $X \sim \mathsf{Binomial}(n, p)$. By the definition of a Binomial distribution, we can rewrite $X = Y_1 + Y_2 + \cdots + Y_n$, where each Y_i is an independent Bernoulli random variable with parameter p. Thus,

$$\mathsf{Var}(X) = \mathsf{Var}(Y_1 + Y_2 + \cdots + Y_n) = \sum_{i=1}^{n} \mathsf{Var}(Y_i) = np(1-p).$$

3.2 Moment generating function (MGF)

Moment generating function (MGF) is a powerful function that uniquely describes the underlying features of a random variable. The MGF of a RV X is

$$M_X(t) = \mathbb{E}(e^{tX}).$$

Note that M_X may not exist. However, when it exists, it admits the following interpretation in terms of the moments of RV X. First note that, for $g(t) = e^{tX}$,

$$g(t) = e^{tX} = \sum_{n=0}^{\infty} \frac{g^{(n)}(0)}{n!} t^n = 1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots .$$

Thus,

$$M_X(t) = 1 + t\mu_1 + \frac{t^2\mu_2}{2!} + \frac{t^3\mu_3}{3!} + \cdots ,$$

where $\mu_j = \mathbb{E}(X^j)$ is the j-th moment of X. Therefore,

$$\mathbb{E}(X^j) = M^{(j)}(0) = \left. \frac{d^j M_X(t)}{dt^j} \right|_{t=0},$$

where $M^{(j)}(0)$ is the j-th derivative of $M(t)$ at $t = 0$. Here you see how the moments of X is generated by the function M_X.

The MGF uniquely determines the distribution of a random variable. If two random variables X and Y, have the *same MGFs* then X and Y also have the same distribution (CDF). Side note for the curious: the MGF is related to the Laplace transform (actually, they are the same).

The MGF has some interesting properties:

- **Location-scale.** $M_{aX+b}(t) = \mathbb{E}(e^{(aX+b)t}) = e^{bt}\mathbb{E}(e^{atX}) = e^{bt}M_X(at)$.
- **Multiplicity.** $M_{X+Y}(t) = \mathbb{E}(e^{(X+Y)t}) = \mathbb{E}(e^{Xt}e^{Yt})$. Thus,

$$X \perp Y \Rightarrow M_{X+Y}(t) = \mathbb{E}(e^{Xt}e^{Yt}) = \mathbb{E}(e^{Xt})\mathbb{E}(e^{Yt}) = M_X(t)M_Y(t).$$

Example (Bernoulli and Binomial). Let $X \sim \mathsf{Ber}(p)$. Its MGF is $M_X(t) = \mathbb{E}(e^{tX}) = pe^t + (1-p)$. Let $Y \sim \mathsf{Bin}(n, p)$. Using the fact that we can express it as $Y = X_1 + \cdots + X_n$, where each X_i is independent Bernoulli R.V. with parameter p. Its MGF is

$$M_Y(t) = \prod_{i=1}^{n} M_{Z_i}(t) = (pe^t + (1-p))^n.$$

Example (Poisson). Let $X \sim \mathsf{Poisson}(\lambda)$. Then its MGF is

$$M_X(t) = \mathbb{E}(e^{tX}) = \sum_{x=0} e^{tx} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \underbrace{\sum_{x=0}^{\infty} \frac{[\lambda e^t]^x}{x!}}_{=e^{\lambda e^t}} = e^{\lambda(e^t-1)}.$$

Example (Exponential). Let $X \sim \mathsf{Exp}(\lambda)$. Then its MGF is

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_0^{+\infty} e^{tx}\lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t}$$

for $t < \lambda$.

Example (Normal). Let $X \sim N(\mu, \sigma^2)$. Then you can show that (exercise)

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$

You can use the fact that the MGF uniquely determines a distribution to show that any addition of normals is still normal.

Remark (Characteristic function). A more general function than the MGF is the characteristic function. Let i be the imaginary number. The characteristic function of a RV X is defined as

$$\phi_X(t) = \mathbb{E}(e^{itX}).$$

Similarly to the MGF, the characteristic function of an RV uniquely determines its distribution. One additional useful property is that while for a random variable an MGF may not always exist (because, $\mathbb{E}(e^{tX})$ may not always converge, see e.g. the Log-Normal distribution), the characteristic function always exists. When X is absolutely continuous, the characteristic function is the Fourier transform of the PDF.

3.2.1 Multivariate MGF

The MGF can be defined for a random vector. Consider $X = (X_1, \cdots, X_d) \in \mathbb{R}^d$ be a random vector. Then its MGF will be a function of a d-dimensional argument, $t = (t_1, \cdots, t_d) \in \mathbb{R}^d$

$$M_X(t) = \mathbb{E}(e^{t^T X}).$$

Example. Let X be a multivariate normal $MVN(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ is the mean vector and $\Sigma \in \mathbb{R}^{d\times d}$ is the covariance matrix. Namely, each component $X_i \sim N(\mu_i, \Sigma_{ii})$ and the covariance $\mathsf{Cov}(X_i, X_j) = \Sigma_{ij}$. Then its MGF will be

$$M_X(t) = e^{t^T\mu + \frac{1}{2}t^T\Sigma t}.$$

Using this, you can show that the linear transformation $Z = b + AX$ follows the $MVN(b + A\mu, A\Sigma A^T)$ distribution.

Example (Normal plus Normal). Here we show that the MGF provides a simple way to see that the addition of two normal random variables still leads to a normal random variable. Let X, Y be two normal random variable such that their joint distribution is MVN with mean (μ_1, μ_2) and covariance matrix Σ. Consider $Z = X + Y$. To see why Z is still normal, consider its MGF:

$$M_Z(t) = \mathbb{E}(e^{tZ}) = \mathbb{E}(e^{tX+tY}) = M_{X,Y}(t, t),$$

which is the MGF of the normal vector (X, Y) with the argument (t, t). Thus,

$$M_Z(t) = M_{X,Y}(t, t) = e^{t(\mu_1+\mu_2)+\frac{1}{2}t^2(\Sigma_{11}+\Sigma_{22}+2\Sigma_{12})},$$

which is the MGF of a normal random variable with mean $\mu_1 + \mu_2$ and variance $\Sigma_{11} + \Sigma_{22} + 2\Sigma_{12} = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X, Y)$.

3.3 Convergence Theory

3.3.1 Convergence in distribution.

Let $Z_1, \cdots, Z_n, \cdots$ be a sequence of random variables with CDFs $F_1, \cdots, F_n, \cdots$. For a random variable Z with CDF F, we say that Z_n **converges in distribution** (a.k.a. converge weakly or converge in law) to Z

if

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

for all x where F is continuous.

In this case, we write

$$Z_n \xrightarrow{D} Z, \quad \text{or } Z_n \xrightarrow{d} Z.$$

Namely, the CDF's of the sequence of random variables converge to a the CDF of a fixed random variable.

3.3.2 Convergence in probability.

For a sequence of random variables $Z_1, \cdots, Z_n, \cdots$, we say Z_n **converges in probability** to another random variable Z if for any $\epsilon > 0$,

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

and we will write

$$Z_n \xrightarrow{P} Z$$

In other words, Z_n converges in probability implies that the distribution is concentrating at a targeting point.

3.3.3 Convergence almost surely.

For a sequence of random variables $Z_1, \cdots, Z_n, \cdots$, we say Z_n **converges almost surely** to a random variable Z if

$$P(\lim_{n\to\infty} Z_n = Z) = 1$$

or equivalently,

$$P(\{\omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\}) = 1.$$

We use the notation

$$Z_n \xrightarrow{a.s.} Z$$

to denote convergence almost surely.

Note that almost sure convergence implies convergence in probability. Convergence in probability implies convergence in distribution. In many cases, convergence in probability or almost surely converge occurs when a sequence of RVs converging toward a fixed number. In this case, we will write (assuming that μ is the target of convergence)

$$Z_n \xrightarrow{P} \mu, \quad Z_n \xrightarrow{a.s.} \mu.$$

Examples.

- Let $\{X_1, X_2, \cdots,\}$ be a sequence of random variables such that $X_n \sim N\left(0, 1 + \frac{1}{n}\right)$. Then X_n converges in distribution to Z, where $Z \sim N(0, 1)$.
- Let $\{X_1, X_2, \cdots\}$ be a sequence of random variables such that $X_i \sim N(0, 1/n)$. Then $X_n \xrightarrow{P} 0$, i.e., it converges in probability to a random variable Z that takes value 0 with probability 1. Also, the random variable $\sqrt{n}X_n \xrightarrow{D} N(0, 1)$.

- Let $\{X_1, X_2, \cdots\}$ be a sequence of random variables such that

$$P(X_n = 0) = 1 - \frac{1}{n}, \quad P(X_n = 1) = \frac{1}{n}.$$

  Then $X_n \xrightarrow{P} 0$.

- Let $\{X_1, X_2, \cdots\}$ be a sequence of independent random variables such that

$$P(X_n = 0) = 1 - \frac{1}{n}, \quad P(X_n = 1) = \frac{1}{n}.$$

  Then $X_n \xrightarrow{P} 0$ but not almost sure convergence.

Sometimes, one may be thinking that the convergence in probability/distribution may imply convergence in *expectation*. But this is not true! Here is an example that it converges in probability to 0 but its expectation diverges.

Example (diverging expectation but convergence in probability). Consider a sequence of RVs $X_1, X_2, \cdots$, such that

$$P(X_n = 0) = 1 - \frac{1}{n}, \quad P(X_n = n^2) = \frac{1}{n}.$$

Then you can easily verify that $X_n \xrightarrow{P} 0$. However, if you compute the expectation,

$$\mathbb{E}(X_n) = n \to \infty.$$

So the expectation is in fact diverging. Later we will see that convergence in expectation implies convergence in probability (follows from Markov's inequality).

3.3.4 Weak Law of Large Numbers

We write $X_1, \cdots, X_n \sim F$ when $X_1, \cdots, X_n$ are IID (independently, identically distributed) from a CDF F. In this case, $X_1, \cdots, X_n$ is called a *random sample*.

Theorem 3.1 (Markov's inequality) *Let X be a non-negative RV with $\mathbb{E}(X) < \infty$. Then for any $\epsilon > 0$,*

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

A feature of Markov inequality is that it implies that *converges in expectation $\Rightarrow$ convergence in probability.* Also, Markov's inequality implies the following useful result, known as Chebyshev's inequality.

Theorem 3.2 (Chebyshev's inequality) *Let X be a RV with $\mathbb{E}(X) < \infty$ and $\mathsf{Var}(X) < \infty$ Then for any $\epsilon > 0$,*

$$P(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\mathsf{Var}(X)}{\epsilon^2}.$$

The proof of the Chebyshev's inequality is a direct application of the Markov's inequality. The Chebyshev's inequality shows that for a sequence of random variables with equal mean but a vanishing variance, this sequence converges in probability to the mean. Applying Chebyshev's inequality to the sample mean, we obtain the weak law of large numbers.

Theorem 3.3 (Weak Law of Large Numbers) *Let* $X_1, \cdots, X_n \overset{IID}{\sim} F$, *with* $\mu = \mathbb{E}(X_1) < \infty$ *and* $\mathsf{Var}(X_1) = \sigma^2 < \infty$. *Then the sample average*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges in probability to μ. *i.e.,*

$$\overline{X}_n \xrightarrow{P} \mu.$$

Proof: Using the properties of variance and the fact that $X_1, \ldots, X_n$ are IID, one can easily show that

$$\mathsf{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Thus, by Chebyshev's inequality

$$P(|\overline{X}_n - \mu| > t) \leq \frac{\sigma^2}{nt^2} \to 0,$$

which completes the proof.

■

The above theorem is known as the Weak Law of Large Numbers. We can in fact avoid assuming the existence of a variance but the proof will become much more complicated. Note that a strong law of large numbers exists. This result states the convergence in terms of an 'almost sure convergence'.

3.3.5 Central Limit Theorem

Theorem 3.4 (Central Limit Theorem) *Let* $X_1, \cdots, X_n$ *be IID random variables with* $\mu = \mathbb{E}(X_1)$ *and* $\sigma^2 = \mathsf{Var}(X_1) < \infty$. *Let* $\overline{X}_n$ *be the sample average. Then*

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1).$$

Note that $N(0,1)$ *is also called a* standard normal random variable.

Proof: Let $Z = \sqrt{n}(\overline{X}_n - \mu)$. Proving the theorem is equivalent to showing that $Z \to N(0, \sigma^2)$.

Note that we can rewrite Z as

$$Z = \sqrt{n}(\overline{X}_n - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i,$$

Thus, the MGF of Z is

$$M_Z(t) = \mathbb{E}(e^{tZ}) = \mathbb{E}\left(e^{\frac{t}{\sqrt{n}}\sum_{i=1}^{n} Y_i}\right) = \mathbb{E}\big(e^{\frac{t}{\sqrt{n}}Y_1}\big)^n = \big(\mathbb{E}\big(e^{\frac{t}{\sqrt{n}}Y_1}\big)\big)^n = \big(M_{Y_1}(t/\sqrt{n})\big)^n. \tag{3.1}$$

Note that above we use the fact that $Y_1, \cdots, Y_n$ are IID random variables. Additionally, note that $\mathbb{E}(Y_1) = \mathbb{E}(X_1 - \mu) = 0$, and $\mathbb{E}(Y_1^2) = \mathbb{E}([X_1 - \mu]^2) = \mathsf{Var}(X_1) = \sigma^2$. We use these facts and the Taylor expansion representation of $M_{Y_1}(t/\sqrt{n})$ with the linearity of expectation below. Note that $M_{Y_1}^{(k)}(0) = \mathbb{E}[Y_1^k]$, for all k.

$$M_{Y_1}(t/\sqrt{n}) = \sum_{k=0}^{\infty} \frac{(t/\sqrt{n})^k M_{Y_1}^{(k)}(0)}{k!} = 1 + \frac{t}{\sqrt{n}}\underbrace{\mathbb{E}(Y_1)}_{=0} + \frac{t^2}{2n}\underbrace{\mathbb{E}(Y_1^2)}_{=\sigma^2} + o\left(\frac{1}{n}\right).$$

The left out Taylor expansion terms are replaced by $o(\frac{1}{n})$, which denotes the fact that these terms converge to 0 faster than $\frac{1}{n}$, when $n \to \infty$. Using the above expansion, we can see that

$$(M_{Y_1}(t/\sqrt{n}))^n \left(1 + \frac{t^2\sigma^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \overset{n\to\infty}{\to} e^{\frac{1}{2}t^2\sigma^2}.$$

Plugging this back into Equation (3.1) we obtain

$$M_Z(t) = (M_{Y_1}(t/\sqrt{n}))^n = \left(1 + \frac{t^2\sigma^2 + o(1)}{2n}\right)^n \overset{n\to\infty}{\to} e^{\frac{1}{2}t^2\sigma^2},$$

which is the MGF of a normal random variable with mean 0 and variance σ^2. ■

Note that there are other versions of the central limit theorem that allow dependent RVs or infinite variance using the idea of 'triangular array' (also known as the Lindeberg-Feller Theorem). However, the details are beyond the scope of this course so we will not pursue them here.

3.3.6 Other useful theorems

Theorem 3.5 (*Continuous mapping theorem*) *Let g be a continuous function, let X be a random variable and $X_1, \ldots, X_n, \ldots$ be a sequence of random variables.*

1. *If $X_n \overset{D}{\to} X$, then $g(X_n) \overset{D}{\to} g(X)$.*
2. *If $X_n \overset{p}{\to} X$, then $g(X_n) \overset{p}{\to} g(X)$.*
3. *If $X_n \overset{a.s.}{\to} X$, then $g(X_n) \overset{a.s.}{\to} g(X)$.*

We will sometimes abuse our notation slightly by writing $Y_n \overset{p}{\to} c$, for $Y_1, \ldots, Y_n, \ldots$ a sequence of random variables and $c \in \mathbb{R}$ a constant. What we mean by this is that Y_n converges to a binary random variable Y that takes value c with probability 1, and takes any other value with probability 0.

Theorem 3.6 (*Slutsky's theorem.*) *Let $\{X_n : n = 1, 2, \cdots\}$ and $\{Y_n : n = 1, 2, \cdots\}$ be two sequences of RVs such that $X_n \overset{D}{\to} X$ and $Y_n \overset{p}{\to} c$, where X is a RV c is a constant. Then*

$$\begin{aligned} X_n + Y_n &\overset{D}{\to} X + c \\ X_n Y_n &\overset{D}{\to} cX \\ X_n / Y_n &\overset{D}{\to} X/c \quad \textit{(if } c \neq 0\textit{)}. \end{aligned}$$

We will use Theorems 3.5 and 3.6 in future chapters. Especially in discussions of maximum likelihood estimators. They will also be important for STAT 513.

Why do we need these notions of convergence? The convergence in probability is related to the concept of statistical consistency. An estimator is statistically consistent if it converges in probability toward its target population quantity, this property will become very important later on. The convergence in distribution is often used to construct confidence intervals or perform a hypothesis tests.

3.4 Concentration inequalities

In addition to Theorems 3.5 and 3.6, we will often use concentration inequalities to obtain convergence in probability. Let $\{X_n : n = 1, 2, \cdots\}$ be a sequence of RVs. For a given $\epsilon > 0$, a concentration inequality aims to compute the function $\phi_n(\epsilon)$ such that

$$P(|X_n - \mathbb{E}(X_n)| > \epsilon) \leq \phi_n(\epsilon)$$

and we have a concentration inequality if $\phi_n(\epsilon) \overset{n\to\infty}{\to} 0$. This automatically gives us convergence in probability for X_n. Moreover, the *convergence rate* of $\phi_n(\epsilon)$ towards 0 with respect to n is a central quantity that describes how fast X_n converges toward its mean.

Example: concentration of a Gaussian mean. Markov's inequality implies a useful bound on describing how fast the sample mean of a Gaussian converges to the population mean. For simplicity, we consider a sequence of mean 0 Gaussians: $X_1, \cdots, X_n \sim N(0, \sigma^2)$. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean. It is known that $\overline{X}_n \sim N(0, \sigma^2/n)$. Then

$$\begin{aligned} P(\overline{X}_n > \epsilon) &= P(e^{\overline{X}_n} > e^{\epsilon}) \\ &= P(e^{s\overline{X}_n} > e^{s\epsilon}), \text{ for a positive number } s \\ &\leq \frac{\mathbb{E}(e^{s\overline{X}_n})}{e^{s\epsilon}} \qquad \text{by Markov's inequality} \\ &= e^{\frac{1}{2n}\sigma^2 s^2 - s\epsilon} \qquad \text{by the MGF of a Gaussian.} \end{aligned}$$

Since we know that our probability is smaller or equal than any value of $e^{\frac{1}{2n}\sigma^2 s^2 - s\epsilon}$, for $s > 0$, it is also smaller or equal than its minimum. Since $e^{\frac{1}{2n}\sigma^2 s^2 - s\epsilon}$ is a convex function in s, we can find the minimum as the solution of

$$\frac{\partial e^{\frac{1}{2n}\sigma^2 s^2 - s\epsilon}}{\partial s} = 0$$

. The minimum is reached for $s = \frac{n\epsilon}{2\sigma^2}$ Hence,

$$P(\overline{X}_n > \epsilon) \leq e^{-\frac{n\epsilon^2}{2\sigma^2}}.$$

Since the distribution of $\overline{X}_n$ is symmetric around 0, we also have $P(\overline{X}_n < -\epsilon) \leq e^{-\frac{n\epsilon^2}{2\sigma^2}}$. So we conclude

$$P(|\overline{X}_n| > \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}}$$

or more generally,

$$P(|\overline{X}_n - \mathbb{E}(X_1)| > \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}},$$

giving us our first *concentration inequality*.

Example (concentration of a maximum). Let $X_1, \cdots, X_n$ be IID normal random variables $N(0, \sigma^2)$. Define $Z_n = \max\{|X_1|, \cdots, |X_n|\}$. Intuitively, we know that when $n \to \infty$, Z_n should diverge since we are taking the maximum of more and more values. But, it is possible to find an increasing sequence $\gamma_n \to \infty$ such that Z_n/γ_n will not diverge (in probability). How do we find such a sequence γ_n?

A simple approach is based on the concentration inequality. Using the result from previous example, we know that for a single random variable X_i (replace $n = 1$ above), we have

$$P(|X_i| > \epsilon) \leq 2e^{-\frac{\epsilon^2}{2\sigma^2}}.$$

With this, we know that

$$\begin{aligned} P(Z_n > \epsilon) &= P(\max\{|X_1|, \cdots, |X_n|\} > \epsilon) \\ &\leq \sum_{i=1}^{n} P(|X_i| > \epsilon) \qquad \text{(maximum is over } \epsilon \Rightarrow \text{one of them must hold)} \\ &\leq 2ne^{-\frac{\epsilon^2}{2\sigma^2}}. \end{aligned}$$

To find the sequence γ_n, we will replace ϵ with ϵ_n above. Our goal is to identify a sequence ϵ_n that would lead to the following, for some constant $0 < \delta < 1$,

$$2ne^{-\frac{\epsilon_n^2}{2\sigma^2}} \overset{n\to\infty}{\to} \delta$$

Treating the above $\overset{n\to\infty}{\to}$ as an equality and solving for ϵ_n gets us $\epsilon_n = \sigma\sqrt{2\log(2n) - 2\log(\delta)}$. Hence, we need a sequence γ_n to diverge at the same rate as ϵ_n. This leads to the choice of $\gamma_n = \sigma\sqrt{2\log n}$, which itself gives a characterization on how fast Z_n diverges.

3.4.1 Concentration of the mean

Let $X_1, \cdots, X_n \overset{IID}{\sim} F$ be a random sample such that $\sigma^2 = \mathsf{Var}(X_1)$. Using Chebyshev's inequality, we know that the sample average $\overline{X}_n$ has the following concentration inequality:

$$P(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

When the RVs are bounded, there is a stronger notion of convergence, which we explore in the following theorem.

Theorem 3.7 (Hoeffding's inequality) *Let $X_1, \cdots, X_n$ be IID RVs such that $a \leq X_1 \leq b$ and let $\overline{X}_n$ be the sample average. Then for any $\epsilon > 0$,*

$$P(\overline{X}_n - \mathbb{E}(\overline{X}_n) \geq \epsilon) \leq e^{\frac{-2n\epsilon^2}{(b-a)^2}}$$

and

$$P(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \geq \epsilon) \leq 2e^{\frac{-2n\epsilon^2}{(b-a)^2}}.$$

Before proving Hoeffding's inequality, we first introduce the following lemma:

Lemma 3.8 *Let X be a random variable with $\mathbb{E}(X) = 0$ and $a \leq X \leq b$. Then*

$$\mathbb{E}(e^{tX}) \leq e^{t^2(b-a)^2/8}$$

for any number $t \in \mathbb{R}$.

Proof: We will use the fact that $x \mapsto e^{tx}$ is a convex function for all positive t. Recall that a function $g(x)$ is a convex function if for any two point $a < b$ and $\alpha \in [0, 1]$,

$$g(\alpha a + (1-\alpha)b) \leq \alpha g(a) + (1-\alpha)g(b).$$

Because $X \in [a, b]$,

$$X = \alpha_X b + (1 - \alpha_X)a.$$

we have that

$$\alpha_X = \frac{X - a}{b - a},$$

Using the fact that $x \mapsto e^{tx}$ is convex,

$$e^{tX} \leq \alpha_X e^{tb} + (1 - \alpha_X)e^{ta} = \frac{X - a}{b - a}e^{tb} + \frac{b - X}{b - a}e^{ta}.$$

Now taking the expectation of both sides,

$$\mathbb{E}(e^{tX}) \leq \frac{\mathbb{E}(X) - a}{b - a}e^{tb} + \frac{b - \mathbb{E}(X)}{b - a}e^{ta} = \frac{b}{b - a}e^{ta} - \frac{a}{b - a}e^{tb} = e^{g(s)}, \tag{3.2}$$

where $s = t(b - a)$ and $g(s) = -\gamma s + \log(1 - \gamma + \gamma e^s)$ and $\gamma = -a/(b - a)$. Note that $g(0) = g'(0) = 0$ and $g''(s) \leq 1/4$ for all positive s. Using Taylor's theorem,

$$g(s) = g(0) + sg'(0) + \frac{1}{2}s^2 g''(s^*)$$

for some $s^* \in [0, s]$. Thus, we conclude $g(s) \leq \frac{1}{2} \times s^2 \times \frac{1}{4} = \frac{1}{8}s^2$.

Then equation (3.2) implies

$$\mathbb{E}(e^{tX}) \leq e^{g(s)} \leq e^{\frac{s^2}{8}} = e^{\frac{t^2(b-a)^2}{8}}.$$

■

Now, we formally prove Hoeffding's inequality.

Proof:

We first prove that $P\left(\overline{X}_n - \mu \geq \epsilon\right) \leq e^{-2n\epsilon^2/(b-a)^2}$.

Let $Y_i = X_i - \mu$. Because the exponential function is monotonically increasing, for any positive t,

$$\begin{aligned}
P\left(\overline{X}_n - \mu \geq \epsilon\right) &= P\left(\overline{Y}_n \geq \epsilon\right) \\
&= P\left(\sum_{i=1}^n Y_i \geq n\epsilon\right) \\
&= P\left(e^{\sum_{i=1}^n Y_i} \geq e^{n\epsilon}\right) \\
&= P\left(e^{t\sum_{i=1}^n Y_i} \geq e^{tn\epsilon}\right) \\
&\leq \frac{\mathbb{E}(e^{t\sum_{i=1}^n Y_i})}{e^{tn\epsilon}} \quad \text{by Markov's inequality} \\
&= e^{-tn\epsilon}\mathbb{E}(e^{tY_1} \cdot e^{tY_2} \cdots e^{tY_n}) \\
&= e^{-tn\epsilon}\mathbb{E}(e^{tY_1}) \cdot \mathbb{E}(e^{tY_2}) \cdots \mathbb{E}(e^{tY_n}) \\
&= e^{-tn\epsilon}\mathbb{E}(e^{tY_1})^n \\
&\leq e^{-tn\epsilon}e^{nt^2(b-a)^2/8} \text{ by Lemma 3.8.}
\end{aligned}$$

Because the above inequality holds for all positive t, we can choose t to optimize the bound. To get the bound as sharp as possible, we would like to make it as small as possible. Taking derivatives with respect

to t and set it to be 0, we obtain

$$t_* = \frac{4\epsilon}{(b-a)^2}$$

and

$$-t_* n\epsilon + nt_*^2(b-a)^2/8 = -2n\epsilon^2/(b-a)^2.$$

Thus, the inequality becomes

$$P\left(\overline{X}_n - \mu \geq \epsilon\right) \leq e^{-t_* n\epsilon} e^{nt_*^2(b-a)^2/8} = e^{-2n\epsilon^2/(b-a)^2}.$$

The same proof also applies to the case $P\left(\overline{X}_n - \mu \leq -\epsilon\right)$ and we will obtain the same bound. Therefore, we conclude that

$$P\left(|\overline{X}_n - \mu| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2/(b-a)^2}.$$

■

Hoeffding's inequality gives a concentration of an exponential order (actually it is often called a Gaussian rate). The convergence rate is much faster than the one given b Chebyshev's inequality. Obtaining such an exponential rate is useful for analyzing the property of an estimator. Many modern statistical topics, such as high-dimensional problems, nonparametric inference, semi-parametric inference, and empirical risk minimization all rely on a convergence rate of this form. Note that the exponential rate may also be used to obtain an almost sure convergence via the Borel-Cantelli Lemma (see Section 4 of Billingsley, 2017, as well as Theorem 22.8 in Section 22.)

Example: consistency of estimating a high-dimensional proportion. To see how the Hoeffding's inequality is useful, we consider the problem of estimating the proportion of several binary variables. Suppose that we observe IID observations

$$X_1, \cdots, X_n \in \{0,1\}^d.$$

$X_{ij} = 1$ can be interpreted as the i-th individual response 'Yes' in j-th question. We are interested in estimating the proportion vector $\pi \in [0,1]^d$ such that $\pi_j = P(X_{ij} = 1)$ is the proportion of 'Yes' response in j-th question in the population. A simple estimator is the sample proportion $\hat{\pi} = (\hat{\pi}_1, \cdots, \hat{\pi}_d)^T$ such that

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}.$$

When d is much smaller than n, it is easy to see that this is a good estimator. However, if $d = d_n \to \infty$ with $n \to \infty$, will $\hat{\pi}$ still be a good estimator of π? To define a good estimator, we mean that *every proportion* can be estimated accurately. A simple way to quantify this is the vector max norm:

$$\|\hat{\pi} - \pi\|_{\max} = \max_{j=1,\cdots,d} |\hat{\pi}_j - \pi_j|.$$

We consider the problem of estimating π_j first. It is easy to see that by the Hoeffding's inequality,

$$P(|\hat{\pi}_j - \pi_j| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Thus,

$$\begin{aligned} P(\|\hat{\pi} - \pi\|_{\max} > \epsilon) &= P\left(\max_{j=1,\cdots,d} |\hat{\pi}_j - \pi_j| > \epsilon\right) \\ &\leq P\left(|\hat{\pi}_1 - \pi_1| > \epsilon \cup |\hat{\pi}_2 - \pi_2| > \epsilon \cup \cdots \cup |\hat{\pi}_d - \pi_d| > \epsilon\right) \\ &\leq \sum_{j=1}^d P(|\hat{\pi}_j - \pi_j| > \epsilon) \\ &\leq 2de^{-2n\epsilon^2}. \end{aligned} \tag{3.3}$$

Thus, as long as $2de^{-2n\epsilon^2} \to 0$ for any fixed ϵ, we have the statistical consistency. This implies that we need

$$\frac{\log d}{n} \to 0,$$

which allows the number of questions/variables to increase a lot faster than the sample size n!

References

Billingsley, P. (2017). *Probability and measure*. John Wiley & Sons.

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Perlman, M. D. (2019). Probability and mathematical statistics i. https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf.

STAT 512: Statistical Inference **Original notes credit: Yen-Chi Chen**

Lecture 4: Conditional expectation and conditional distribution

Instructor: Emilija Perković Compiled on: 2023-10-20, 09:47:29

See also Chapter 4 of Perlman (2019) and Chapter 4 of Casella and Berger (2021).

4.1 Recall: conditional distributions

Suppose (X, Y) is a random vector with joint CDF

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

In the first lecture, we explained how to obtain the conditional distribution of $Y|X = x$ when both X and Y are continuous or when both X and Y are discrete.

X, Y continuous: When both variables are continuous, the *conditional PDF* of Y given $X = x$ is

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)},$$

where $p_{XY}(x, y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$, and $p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y)dy$.

X, Y discrete: When both X and Y are discrete, the conditional PMF of Y given $X = x$ is

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)},$$

where $p_{XY}(x, y) = P(X = x, Y = y)$, and $p_X(x) = P(X = x) = \sum_y P(X = x, Y = y)$.

We now consider the cases when the random vector (X, Y) is mixed and we still want to compute the distribution of $Y|X = x$.

X discrete, Y continuous: Note that the joint CDF $F_{XY}(x, y)$ is still well-defined. That is,

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

In this case it also makes sense to consider $P(X = x, Y \leq y)$ as

$$P(X = x, Y \leq y) = P(X \leq x, Y \leq y) - P(X < x, Y \leq y) = F_{XY}(x, y) - \lim_{\epsilon \downarrow 0} F_{XY}(x - \epsilon, y).$$

Now, we define the joint PDF/PMF $p_{XY}(x, y)$ for (X, Y) as follows

$$p_{XY}(x, y) := \frac{d}{dy} P(X = x, Y \leq y).$$

Note that $:=$ denotes a defining equality. To sum up, we have extend the joint PDF/PMF p_{XY} to a mixed case where one of the random variables is continuous while the other is discrete. From simplicity, we will from now on refer to the call p_{XY} as the joint PDF even if one or both X and Y are discrete.

We can now define conditional PMF $p_{X|Y}(x|y)$ and conditional PDF $p_{Y|X}(y|x)$ as follows.

For a positive PDF $p_Y(y) > 0$, the PMF $p_{X|Y}(x|y)$ is defined as:

$$p_{X|Y}(x|y) := \frac{p_{XY}(x,y)}{p_Y(y)}.$$

For a positive PMF $p_X(x) > 0$, the PDF $p_{Y|X}(y|x)$ is defined as:

$$p_{Y|X}(y|x) := \frac{p_{XY}(x,y)}{p_X(x)}.$$

Remark (beyond the scope of this course). Formally, both a PMF and a PDF can be called *density functions* in a general sense when using Radon-Nikodym derivatives. Roughly speaking, a density function $p(x)$ is defined as the ratio $p(x) = \frac{dP(x)}{d\mu(x)}$, where $P(x)$ is a *probability measure* and $\mu(x)$ is another *measure*. When $\mu(x)$ is the Lebesgue measure, $p(x) = \frac{dP(x)}{d\mu(x)}$ is the usual PDF. When $\mu(x)$ is the counting measure (as in the case of discrete variables), $p(x) = \frac{dP(x)}{d\mu(x)}$ reduces to the PMF. So both PDF and PMF can be referred to as *density* functions (more on this in your measure theory course).

Example (Poisson-Exponential-Gamma). Suppose that we have two R.V.s X and Y such that X is a discrete random variables, $X \in \{0, 1, 2, 3, \cdots\}$, and Y is a continuous random variables, where $Y \geq 0$, The joint PDF of X and Y is

$$p_{XY}(x,y) = \frac{\lambda y^x e^{-(\lambda+1)y}}{x!},$$

$\lambda > 0$. Compute the PDFs $p_{X|Y}$ and $p_{Y|X}$.

To compute $p_{X|Y}$, we first computing $p_Y(y)$ (Method 1, see Lecture 1). Here

$$p_Y(y) = \sum_x \frac{\lambda y^x e^{-(\lambda+1)y}}{x!} = \lambda e^{-(\lambda+1)y} \underbrace{\sum_x \frac{y^x}{x!}}_{e^y} = \lambda e^{-\lambda y}.$$

So $Y \sim \mathsf{Exp}(\lambda)$. And thus

$$p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)} = \frac{\frac{\lambda y^x e^{-(\lambda+1)y}}{x!}}{\lambda e^{-\lambda y}} = \frac{y^x e^{-y}}{x!}.$$

Therefore, $X|Y = y \sim \mathsf{Poisson}(y)$.

Recall, that we can also use Method 2 (the proportional trick from Lecture 1). We know that $p_{X|Y}(x|y)$ will be a density function, where X is random, and $Y = y$ is not random. So we only need to keep track of how the function changes w.r.t x and treat y as a constant. This method leads to

$$p_{X|Y}(x|y) \propto p_{XY}(x,y) = \frac{\lambda y^x e^{-(\lambda+1)y}}{x!} \propto \frac{y^x}{x!}.$$

From this, we can see that $X|Y = y$ follows a Poisson distribution with rate parameter y.

Let's use the same trick (Method 2) to compute the conditional distribution $p_{Y|X}$, that is, we keep track of y and treat x as a constant. Hence,

$$p_{Y|X}(y|x) \propto p_{XY}(x,y) = \frac{\lambda y^x e^{-(\lambda+1)y}}{x!} \propto y^x e^{-(\lambda+1)y},$$

which leads us to conclude that $Y|X = x$ follows a Gamma distribution with parameters $\alpha = x+1, \beta = \lambda+1$.

4.2 Conditional expectations

The **conditional expectation** of Y given X is the random variable $\mathbb{E}(Y|X) = g(X)$ such that when $X = x$, its value is

$$\mathbb{E}(Y|X=x) = \begin{cases} \int yp(y|x)dy, & \text{if } Y \text{ is continuous,} \\ \sum_y yp(y|x), & \text{if } Y \text{ is discrete,} \end{cases}$$

where $p(y|x) = p(x,y)/p(x)$ is the PDF/PMF of $Y|X = x$. Essentially, the conditional expectation is the expectation of the conditional distribution.

Note that when X and Y are independent,

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y), \quad \mathbb{E}(X|Y=y) = \mathbb{E}(X).$$

Law of total expectations / Tower rule:

$$\begin{aligned} \mathbb{E}[\mathbb{E}[Y|X]] &= \int \mathbb{E}[Y|X=x]p_X(x)dx = \int\int yp_{Y|X}(y|x)p_X(x)dxdy \\ &= \int\int yp_{XY}(x,y)dxdy = \mathbb{E}[Y]. \end{aligned}$$

A more general form of this is that for any measurable function $g(x,y)$, we have

$$\mathbb{E}[g(X,Y)] = \mathbb{E}[\mathbb{E}[g(X,Y)|X]]. \tag{4.1}$$

There are many cool applications of equation (4.1).

- Suppose $g(x,y) = q(x)h(y)$. Then equation (4.1) implies

  $$\mathbb{E}[q(X)h(Y)] = \mathbb{E}[\mathbb{E}[q(X)h(Y)|X]] = \mathbb{E}[q(X)\mathbb{E}[h(Y)|X]].$$

- Let $w(X) = \mathbb{E}[h(Y)|X]$. The covariance

  $$\begin{aligned} \mathsf{Cov}(q(X),h(Y)) &= \mathbb{E}[q(X)h(Y)] - \mathbb{E}[q(X)]\mathbb{E}[h(Y)] \\ &= \mathbb{E}[\mathbb{E}[q(X)h(Y)|X]] - \mathbb{E}[q(X)]\mathbb{E}[\mathbb{E}[h(Y)|X]] \\ &= \mathbb{E}[q(X)w(X)] - \mathbb{E}[q(X)]\mathbb{E}[w(X)] \\ &= \mathsf{Cov}(q(X),w(X)) \\ &= \mathsf{Cov}(q(X),\mathbb{E}(h(Y)|X)). \end{aligned}$$

  Namely, the covariance between $q(X)$ and $h(Y)$ is the same as the covariance between $q(X)$ and $w(X) = \mathbb{E}[h(Y)|X]$. Thus, $w(X) = \mathbb{E}[h(Y)|X]$ is sometimes viewed as the projection from $h(Y)$ onto the space of X.

Law of total variance:

$$\begin{aligned} \mathsf{Var}(Y) &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \\ &= \mathbb{E}[\mathbb{E}(Y^2|X)] - \mathbb{E}[\mathbb{E}(Y|X)]^2 \quad \text{(law of total expectation)} \\ &= \mathbb{E}[\mathsf{Var}(Y|X) + \mathbb{E}(Y|X)^2] - \mathbb{E}[\mathbb{E}(Y|X)]^2 \quad \text{(definition of variance)} \\ &= \mathbb{E}[\mathsf{Var}(Y|X)] + \left\{\mathbb{E}[\mathbb{E}(Y|X)^2] - \mathbb{E}[\mathbb{E}(Y|X)]^2\right\} \\ &= \mathbb{E}\left[\mathsf{Var}(Y|X)\right] + \mathsf{Var}\left(\mathbb{E}[Y \mid X]\right) \quad \text{(definition of variance).} \end{aligned}$$

Example (Binomial-uniform). Consider two R.V.s X, Y such that

$$X|Y \sim \mathsf{Bin}(n, Y), \qquad Y \sim \mathsf{Unif}[0,1].$$

We are interested in $E[X]$, $\mathsf{Var}(X)$. For the marginal expectation $\mathbb{E}[X]$, using the law of total expectation,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[nY] = \frac{n}{2}.$$

The variance is

$$\begin{aligned}\mathsf{Var}(X) &= \mathbb{E}\left[\mathsf{Var}(X|Y)\right] + \mathsf{Var}\left(\mathbb{E}[X \mid Y]\right) \\ &= \mathbb{E}(nY(1-Y)) + \mathsf{Var}(nY) \\ &= \frac{n}{2} - \frac{n}{3} + \frac{n^2}{12}.\end{aligned}$$

Now we examine the distribution of $Y|X$. Using the fact that

$$p_{Y|X}(y|x) \propto p_{XY}(x,y) = p_{X|Y}(x|y)p_Y(y) = \binom{n}{x} y^x (1-y)^{n-x} \propto y^x (1-y)^{n-x},$$

we can easily see that this is the PDF of a Beta distribution with parameters $\alpha = x+1$ and $\beta = n-x+1$. This is an interesting case because the uniform distribution over $[0,1]$ is equivalent to $\mathsf{Beta}(1,1)$. And $Y|X \sim \mathsf{Beta}(X+1, n-X+1)$. Thus, initially, Y behaves like $\mathsf{Beta}(1,1)$. Then after observing the data X, we update the distribution of Y to $Y|X \sim \mathsf{Beta}(X+1, n-X+1)$. This is a way of modeling how data informs our decisions and is used in Bayesian inference.

Example (missing data). Consider a social survey where for each participant we record two variables X and Y, where X is the age of a participant and Y is their income. We are interested in estimate $\mu = \mathbb{E}[Y]$. We'll further assume that all participants disclose their age, but not all participants necessarily disclose their income (we have missing information for some samples of Y). In this case, we cannot estimate $\mathbb{E}[Y]$ without making additional assumptions.

We'll use a binary variable R to denote the presence/missingness of information on Y. When $R = 1$, we measure both X and Y. When $R = 0$, we only measure X. We will now make the following assumption that will later enable us to estimate $\mathbb{E}[Y]$. We will assume that $R \perp Y|X$ (this is a special case of the so-called *missing at random* assumption). Under this assumption, the response probability $P(R = 1|X, Y) = \pi(X)$ only depends on X and we will additionally assume that $\pi(X)$ is a known function.

Now consider the quantity:

$$W = \frac{RY}{\pi(X)}.$$

Interestingly, W can always be computed–when $R = 1$, $W = \frac{Y}{\pi(X)}$ and when $R = 0$, $W = 0$. A more interesting fact is that W has the same mean as Y:

$$\begin{aligned}\mathbb{E}[W] &= \mathbb{E}\left[\frac{RY}{\pi(X)}\right] \\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}[RY|X]\right] \\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}[R|X]\mathbb{E}[Y|X]\right] \\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\pi(X)\mathbb{E}[Y|X]\right] \\ &= \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y].\end{aligned}$$

Hence, suppose we observe many IID random copies of $(X, R = 1, Y)$ or $(X, R = 0)$, we estimate $\mu = \mathbb{E}[Y]$ using the mean of the **empirical CDF** (see below)

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\pi(X_i)}.$$

The quantity W is also called the IPW (inverse probability weighed) estimator.

4.3 Empirical Distribution Function

We now briefly introduce the empirical cumulative distribution function (ECDF, or EDF), an estimator of the cumulative distribution function (CDF).

Let first look at the CDF $F(x)$ more closely. Given a value x_0,

$$F(x_0) = P(X_i \leq x_0)$$

for every $i = 1, \cdots, n$. Namely, $F(x_0)$ is the probability of the event $\{X_i \leq x_0\}$.

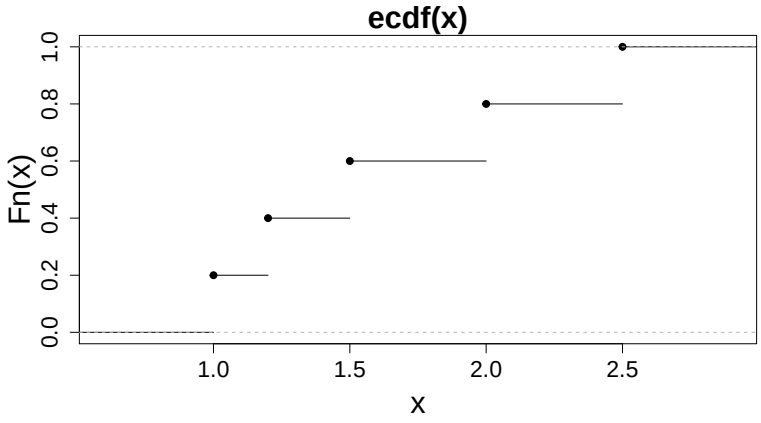
A natural estimator of a probability of an event is *the ratio of such an event in our sample*. Thus, we use

$$\hat{F}_n(x_0) = \frac{\text{number of } X_i \leq x_0}{\text{total number of observations}} = \frac{\sum_{i=1}^{n} I(X_i \leq x_0)}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x_0) \tag{4.2}$$

as the estimator of $F(x_0)$.

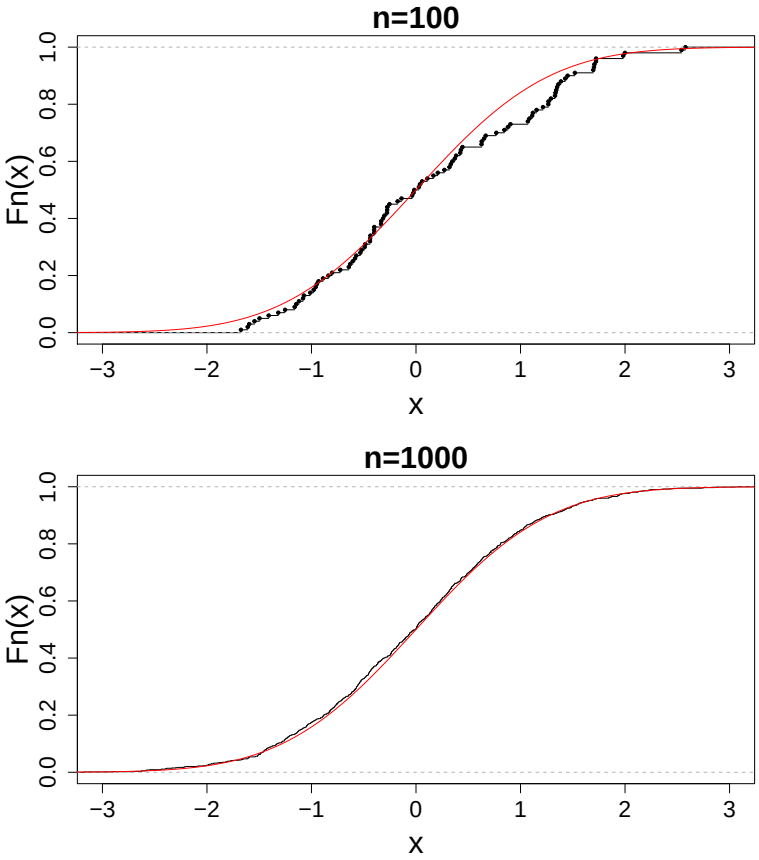
For every x_0, we can use such a quantity as an estimator, so the estimator of the CDF, $F(x)$, is $\hat{F}_n(x)$. This estimator, $\hat{F}_n(x)$, is called the *empirical distribution function (EDF)*.

Example. Here is an example of the EDF of 5 observations of $1, 1.2, 1.5, 2, 2.5$:



There are 5 jumps, each located at the position of an observation. Moreover, the height of each jump is the same: $\frac{1}{5}$.

Example. While the previous example might not be look like an idealized CDF, the following provides a case of EDF versus CDF where we generate $n = 100,1000$ random points from the standard normal $N(0, 1)$:



The red curve indicates the true CDF of the standard normal. Here you can see that when the sample size is large, the EDF is pretty close to the true CDF.

4.3.1 Properties of the EDF

Because EDF is the average of $I(X_i \leq x)$, we now study the property of $I(X_i \leq x)$ first. For simplicity, let $Y_i = I(X_i \leq x)$. What is the random variable Y_i?

Here is the breakdown of Y_i:

$$Y_i = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{if } X_i > x \end{cases}.$$

So Y_i only takes value 0 and 1–so it is actually a Bernoulli random variable! We know that a Bernoulli random variable has a parameter p that determines the probability of outputing 1. What is the parameter p for Y_i?

$$p = P(Y_i = 1) = P(X_i \leq x) = F(x).$$

Therefore, for a given x,

$$Y_i \sim \mathsf{Ber}(F(x)).$$

This implies

$$\begin{aligned} \mathbb{E}(I(X_i \leq x)) &= \mathbb{E}(Y_i) = F(x) \\ \mathsf{Var}(I(X_i \leq x)) &= \mathsf{Var}(Y_i) = F(x)(1 - F(x)) \end{aligned}$$

for a given x.

Now what about $\hat{F}_n(x)$? Recall that $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x) = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

$$\mathbb{E}\left(\hat{F}_n(x)\right) = \mathbb{E}(I(X_1 \le x)) = F(x)$$
$$\mathsf{Var}\left(\hat{F}_n(x)\right) = \frac{\sum_{i=1}^n \mathsf{Var}(Y_i)}{n^2} = \frac{F(x)(1-F(x))}{n}.$$

What does this tell us about using $\hat{F}_n(x)$ as an estimator of $F(x)$?

First, at each x, $\hat{F}_n(x)$ is an *unbiased estimator* of $F(x)$:

$$\textbf{bias}\left(\hat{F}_n(x)\right) = \mathbb{E}\left(\hat{F}_n(x)\right) - F(x) = 0.$$

Second, the *variance converges to* 0 when $n \to \infty$. This implies that for a given x,

$$\hat{F}_n(x) \xrightarrow{P} F(x).$$

i.e., $\hat{F}_n(x)$ is a *consistent* estimator of $F(x)$.

The EDF is a good approximator of the CDF in many ways. As functionals of the EDF can be computed directly from data, they will often serve as **plug-in estimators** for functionals of the CDF. For instance, the first moment of the EDF is the empirical mean:

$$\mathbb{E}_n(X) = \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i$$

This empirical mean will often be used as an estimator of $\mathbb{E}[X]$, see example above.

Example (survey sampling). Suppose a city government is planning to estimate the average income for a city resident. We assume that the city has three districts: A and B and C. 60% of population lives in district A, 30% of population lives in district B, and the remaining 10% of the population lives in district C. For each resident, we record two variables X and Y, where $X \in \{A, B, C\}$ is the indicator of the district the resident lives in and Y is their income.

The average income for a resident of our city is then

$$\mu = \mathbb{E}_Y[Y] = \mathbb{E}_X[\mathbb{E}_Y[Y|X]] = 0.6 \cdot \mathbb{E}_Y[Y|X=A] + 0.3 \cdot \mathbb{E}_Y[Y|X=B] + 0.1 \cdot \mathbb{E}_Y[Y|X=C].$$

However, when the government conducted the survey, they surveyed the same amount of individuals in each district. Our proportions are miss-aligned! So while for a random resident the probability that they live in district C is 1/10, the probability of living in district C for a person in our sample is 1/3. So for our sample, $P(X=A) = P(X=B) = P(X=C) = \frac{1}{3}$.

In this case, how should we construct a quantity $Z = g(X, Y)$ such that $\mathbb{E}[Z] = \mu$?

It turns out that we can use *importance weighting* (a similar idea to the inverse probability weighting above) to construct such $Z = g(X, Y)$. Consider

$$\begin{aligned} Z &= \frac{0.6}{1/3} I(X=A)Y + \frac{0.3}{1/3} I(X=B)Y + \frac{0.1}{1/3} I(X=C)Y \\ &= 1.8 I(X=A)Y + 0.9 I(X=B)Y + 0.3 I(X=C)Y. \end{aligned}$$

Namely, when we observe a person from district A, instead of assigning them a weight of 1, we count this person as 1.8 individuals. When we observe a person from district C, we assign this person a weight of 0.3 individuals. Then you can easily verify that

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}[\mathbb{E}[Z|X]] \\ &= 1.8\mathbb{E}[I(X=A)]\mathbb{E}[Y|X=A] + 0.9\mathbb{E}[I(X=B)]\mathbb{E}[Y|X=B] + 0.3\mathbb{E}[I(X=C)]\mathbb{E}[Y|X=C] \\ &= 0.6\mathbb{E}[Y|X=A] + 0.3\mathbb{E}[Y|X=B] + 0.1\mathbb{E}[Y|X=C] \\ &= \mu.\end{aligned}$$

References

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Perlman, M. D. (2019). Probability and mathematical statistics i. https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf.

STAT 512: Statistical Inference **Original notes credit: Yen-Chi Chen**

Lecture 5: Correlation, prediction, and regression

Instructor: Emilija Perković Compiled on: 2023-11-06, 16:56:02

Additional reading: Chapter 5 of Perlman (2019) and parts of Chapter 11 Casella and Berger (2021).

5.1 Correlation

Pearson's correlation, or simply correlation, is a common measure of association between the two random variables. Formally, for random variables X and Y, their correlation is

$$\rho_{XY} = \mathsf{Cor}(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}}.$$

It has three nice properties:

1) *(Symmetric property:)* $\mathsf{Cor}(X,Y) = \mathsf{Cor}(Y,X)$.
2) *(Location-scale property:)* $\mathsf{Cor}(aX+b, cY+d) = \mathsf{sign}(ac)\mathsf{Cor}(X,Y)$.
3) *(Bounded and colinearity property:)* $-1 \leq \mathsf{Cor}(X,Y) \leq 1$. $\mathsf{Cor}(X,Y) = \pm 1$ if and only if they are perfectly linear, i.e., $X = aY + b$ for some constant a, b.

Proof of 3): Let $U = X - \mathbb{E}[X]$, $V = Y - \mathbb{E}[Y]$, and $g(t) = \mathbb{E}[(tU+V)^2]$. Then

$$g(t) = \mathbb{E}[(tU+V)^2] = t^2\mathbb{E}[U^2] + 2t\mathbb{E}[UV] + \mathbb{E}[V^2].$$

Since, $(tU+V)^2$ will be non-negative for all values of t, it holds that $g(t) = \mathbb{E}[(tU+V)^2] \geq 0$. That is,

$$t^2\mathbb{E}[U^2] + 2t\mathbb{E}[UV] + \mathbb{E}[V^2] \geq 0 \tag{5.1}$$

If we view $g(t)$ as a quadratic function of t, then equation (5.1) implies that the discriminant of $g(t) = 0$ must be smaller or equal to zero (Recall that the discriminant for a quadratic equation $a^2x^2 + bx + c = 0$ is $b^2 - 4ac$.) Hence,

$$\begin{aligned} 4(\mathbb{E}[UV])^2 - 4\mathbb{E}[U^2]\mathbb{E}[V^2] &\leq 0 \\ (\mathbb{E}[UV])^2 &\leq \mathbb{E}[U^2]\mathbb{E}[V^2], \end{aligned} \tag{5.2}$$

where equation (5.2) is known as the *Cauchy-Schwartz inequality.* Replacing $U = X - \mathbb{E}[X]$, $V = Y - \mathbb{E}[Y]$, and $g(t) = \mathbb{E}[(tU+V)^2]$, above we obtain:

$$(\mathsf{Cov}(X,Y))^2 \leq \mathsf{Var}(X)\mathsf{Var}(Y),$$

which gives us that $\rho_{XY}^2 \leq 1$.

Note that equality in equation (5.2) holds if and only if the discriminant of $g(t) = 0$ is exactly zero, that is if there is a $t_0 \in \mathbb{R}$, such that $g(t_0) = 0$. Since $g(t_0) = \mathbb{E}[(t_0U+V)^2]$, for $g(t_0) = 0$, we need to have

$t_0U + V \equiv 0$, for all values of U, and V. So V must exactly be a linear function of U, i.e. Y needs to be exactly a linear (affine) function of X. ■

You can think of correlation as a measure of the *linear* relationship between two random variables. A large correlation implies a strong linear relationship. However, *a low correlation does not imply the two random variables are not related with each other!*

Example (0 correlation but perfectly related). Consider a random variable X take three possible values $-1, 0, 1$ with a probability $P(X = -1) = P(X = 1) = 1/4$ and $P(X = 0) = 1/2$. Let $Y = X^2$. You can see that X and Y are deterministically related. It is easy to see that $\mathbb{E}[X] = 0$ and $\mathbb{E}[XY] = 0$ and $\mathsf{Var}(X), \mathsf{Var}(Y) > 0$. However, the covariance between them will be

$$\mathsf{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 = 0,$$

which implies that $\mathsf{Cor}(X, Y) = 0$. So they are uncorrelated but perfectly related with each other.

5.2 Mean-squared error prediction

A classical problem in statistics is how to use data on random variable X to predict the value of random variable Y. This is known as a prediction problem. Thus, our goal us is to find some function of X, $g(X)$ that approximates Y well in some respect. To measure how close $g(x)$ is to some y, we use a *loss function* $L(g(x), y)$. One common loss function is a squared loss which equals:

$$L(g(x), y) = (g(x) - y)^2$$

To measure how close $g(X)$ is to Y overall we take the expectation over the loss function, $\mathbb{E}[L(g(X), Y)]$. This is known as the *risk function* $R(g) = \mathbb{E}[L(g(X), Y)]$. The risk function for a squared loss, is also known as the *mean-squared error (MSE)*:

$$R(g) = \mathbb{E}((Y - g(X))^2).$$

Namely, the MSE is the expected squared deviation of our predictor $g(X)$ to the target Y.

Ideally, we want to choose g that minimizes $R(g)$. Formally, we want to find

$$g^* = \mathsf{argmin}_g R(g).$$

We now take a deeper look at the MSE $R(g) = \mathbb{E}((Y - g(X))^2)$. Using the law of total expectation,

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

Using the fact that for any fixed constant c,

$$\mathbb{E}[(Y - c)^2] = \mathbb{E}[(Y - \mathbb{E}[Y] + \mathbb{E}[Y] - c)^2] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] + (\mathbb{E}[Y] - c)^2 = \mathsf{Var}(Y) + (\mathbb{E}[Y] - c)^2,$$

we can rewrite the MSE as

$$R(g) = \mathbb{E}[\mathbb{E}[(Y - g(X))^2|X]] = \mathbb{E}[\mathsf{Var}(Y|X) + (\mathbb{E}[Y|X] - g(X))^2] = \mathbb{E}[\mathsf{Var}(Y|X)] + \mathbb{E}[(\mathbb{E}[Y|X] - g(X))^2].$$

The first quantity is independent of g so it does not matter in the selection of g. The second quantity involves

$$(\mathbb{E}[Y|X] - g(X))^2 \geq 0.$$

The only case that the equality holds is $g(X) = \mathbb{E}[Y|X]$. As a result, to minimize the MSE, we should use the conditional expectation $\mathbb{E}[Y|X]$ as our predictor. The conditional expectation $\mathbb{E}[Y|X = x] = m(x)$ is also known as the *regression function* or the *best predictor*.

With the regression function, we can decompose Y as

$$Y = \underbrace{\mathbb{E}[Y|X]}_{\text{best predictor}} + \underbrace{(Y - \mathbb{E}[Y|X])}_{\text{residual}}. \tag{5.3}$$

That is for $g^*(X) = \mathbb{E}[Y|X]$,

$$Y = g^*(X) + Y - g^*(X).$$

Here are some interesting properties of the decomposition in equation (5.3):

- **Unbiased.** $\mathbb{E}[\text{best predictor}] = \mathbb{E}[g^*(X)] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$ and $\mathbb{E}[\text{residual}] = \mathbb{E}[Y - g^*(X)] = 0$.
- **Uncorrelated.** $\mathsf{Cov}(g^*(X), Y - g^*(X)) = \mathsf{Cov}(\mathbb{E}[Y|X], Y - \mathbb{E}[Y|X]) = 0$.
- **Residual variance.** $\mathsf{Var}(Y - g^*(X)) = \mathsf{Var}(Y - \mathbb{E}[Y|X]) = \mathbb{E}[\mathsf{Var}(Y|X)]$. To see this,

$$\begin{aligned}
\mathsf{Var}(Y - g^*(X)) &= \mathsf{Var}(Y - \mathbb{E}[Y|X]) = \mathsf{Var}(Y) - 2\mathsf{Cov}(Y, \mathbb{E}[Y|X]) + \mathsf{Var}(\mathbb{E}[Y|X]) \\
&= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 - 2(\mathbb{E}[Y\mathbb{E}[Y|X]] - \mathbb{E}[Y]\mathbb{E}[\mathbb{E}[Y|X]]) + \mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \\
&= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 - 2\mathbb{E}[\mathbb{E}[Y|X]^2] + 2\mathbb{E}[Y]^2 + \mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[Y]^2 \\
&= \mathbb{E}[Y^2] - \mathbb{E}[\mathbb{E}[Y|X]^2] \\
&= \mathbb{E}[\mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2] \\
&= \mathbb{E}[\mathsf{Var}(Y|X)].
\end{aligned}$$

  Note also that,

$$\mathsf{Var}(Y - g^*(X)) = \mathbb{E}[(Y - g^*(X))^2] - [\mathbb{E}[Y - g^*(X)]]^2 = R(g^*).$$

- **Variance decomposition.** Using the law of total variance we obtain:

$$\mathsf{Var}(Y) = \mathsf{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathsf{Var}(Y|X)].$$

  in our case this decomposition can also be interpreted as:

$$\mathsf{Var}(Y) = \mathsf{Var}(g^*(X)) + R(g^*).$$

5.3 Linear prediction (linear regression)

In the above analysis, we see that the best predictor is $\mathbb{E}[Y|X]$. However, $\mathbb{E}[Y|X]$ does not present any restrictions on the functional form of $g(x)$. This unrestricted set of options is often too rich for a straightforward statistical analysis. So generally, we may choose to restrict ourselves to some set of simple functions $g(x)$. One canonical example is the set of all linear functions. Namely, suppose we want to find constants α, β such that for $g(x) = \alpha + \beta x$, the MSE is minimized. That is, we want to minimize:

$$R(\alpha, \beta) = \mathbb{E}((Y - \alpha - \beta X)^2)$$

This way of choosing $g(x)$ is also known as the *least squares* approach. We

How do we find the best values of α and β? To solve this problem, we first expand $R(\alpha, \beta)$:

$$\begin{aligned} R(\alpha,\beta) &= \mathbb{E}((Y-\alpha-\beta X)^2) \\ &= \mathbb{E}(Y^2+\alpha^2+\beta^2X^2-2Y\alpha-2XY\beta+2\alpha\beta X) \\ &= \mathbb{E}[Y^2]+\alpha^2+\beta^2\mathbb{E}[X^2]-2\alpha\mathbb{E}[Y]-2\beta\mathbb{E}[XY]+2\alpha\beta\mathbb{E}[X], \end{aligned}$$

which is a quadratic function of α, β. Additionally, note that $R(\alpha, \beta) = \mathbb{E}((Y - \alpha - \beta X)^2) \geq 0$, must be a convex function of α, β. Thus, we want to find α^*, β^* such that

$$\alpha^*, \beta^* = \mathsf{argmin}_{\alpha,\beta} R(\alpha, \beta)$$

Since $R(\alpha, \beta)$ is convex, this amounts to solving the following system of gradient equations (known as the first order equations):

$$\begin{aligned} 0 &= \frac{\partial}{\partial\alpha} R(\alpha^*, \beta^*) \\ &= 2\alpha^* - 2\mathbb{E}[Y] + 2\beta^*\mathbb{E}[X] \\ 0 &= \frac{\partial}{\partial\beta} R(\alpha^*, \beta^*) \\ &= 2\beta^*\mathbb{E}[X^2] - 2\mathbb{E}[XY] + 2\alpha^*\mathbb{E}[X] \\ \Rightarrow\quad & \beta^*\mathsf{Var}(X) = \mathsf{Cov}(X, Y) \\ \Rightarrow\quad & \beta^* = \frac{\mathsf{Cov}(X,Y)}{\mathsf{Var}(X)} \\ \Rightarrow\quad & \alpha^* = \mathbb{E}[Y] - \mathbb{E}[X]\beta^*. \end{aligned}$$

With these, the *best linear predictor (BLP)* is

$$\begin{aligned} m^*(x) &= \alpha^* + \beta^* x \\ &= \mathbb{E}[Y] + \frac{\mathsf{Cov}(X,Y)}{\mathsf{Var}(X)}(x - \mathbb{E}[X]) \\ &= \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \end{aligned}$$

where $\mu_X = \mathbb{E}[X], \mu_Y = \mathbb{E}[Y], \sigma_X^2 = \mathsf{Var}(X), \sigma_Y^2 = \mathsf{Var}(Y)$ and ρ_{XY} is the Pearson's correlation.

Interestingly, the MSE under the best linear predictor will be

$$\begin{aligned} R(\alpha^*, \beta^*) &= \mathbb{E}((Y - \alpha^* - \beta^* X)^2) \\ &= \mathbb{E}[(Y - \mu_Y - \rho_{XY}\frac{\sigma_Y}{\sigma_X}(X - \mu_X))^2] \\ &= \sigma_Y^2 - 2\rho_{XY}\frac{\sigma_Y}{\sigma_X}\mathbb{E}[(Y-\mu_Y)(X-\mu_X)] + \rho_{XY}^2\frac{\sigma_Y^2}{\sigma_X^2}\sigma_X^2 \\ &= \sigma_Y^2(1 - \rho_{XY}^2). \end{aligned}$$

An important feature of the above analysis is that *we did NOT assume the linear model to be correct!* We can always find a best linear predictor regardless of what the true regression function looks like.

5.3.1 Multivariate linear prediction

Suppose that the covariate $X = (X_1, \cdots, X_p)$ is now multivariate. The linear prediction approach still works and the MSE will be

$$R(\alpha, \beta) = \mathbb{E}((Y - \alpha - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_p X_p)^2),$$

where $\beta = (\beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$. Let $Z = (1, X_1, \cdots, X_p)^T \in \mathbb{R}^{p+1}$ be a *data vector* and $\gamma = (\alpha, \beta_1, \cdots, \beta_p)^T \in \mathbb{R}^{p+1}$ be a coefficient vector. Then the MSE has an elegant form:

$$R(\gamma) = R(\alpha, \beta) = \mathbb{E}((Y - \gamma^T Z)^2).$$

A direct expansion shows that

$$\begin{aligned} R(\gamma) &= \mathbb{E}(Y^T Y) - 2\mathbb{E}[Y Z^T \gamma] + \mathbb{E}[\gamma^T Z Z^T \gamma] \\ &= \mathbb{E}(Y^T Y) - 2\gamma^T \mathbb{E}[ZY] + \gamma^T \mathbb{E}[ZZ^T]\gamma, \end{aligned}$$

which is a quadratic function of γ.

Differentiating this with respect to the coefficient vector γ leads to

$$0 = -2\mathbb{E}[ZY] + 2\mathbb{E}[ZZ^T]\gamma.$$

Thus, the least squares solution will be

$$\gamma^* = (\mathbb{E}[ZZ^T])^{-1}\mathbb{E}[ZY].$$

Note that $\mathbb{E}[ZZ^T]$ is a matrix and $(\mathbb{E}[ZZ^T])^{-1}$ is the matrix inverse.

With $\gamma^* = (\alpha^*, \beta^*)^T$, we can easily write down the BLP:

$$m^*(x) = \gamma^{*T} z = \alpha^* + \beta^{*T} x = \alpha^* + \sum_{j=1}^{p} \beta_j^* x_j.$$

5.3.2 Correctness of the model

In linear prediction, we did NOT assume the linear model to be correct. In the case of the linear model being correct, we have some really nice properties. We use the notation from the multivariate case for simplicity.

When the linear model is incorrect. The coefficient from the least squares approach is $\gamma^* = \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZY]$. In general, this quantity will change if the distribution of the covariates X (Z) changes. So the coefficient depends on the distribution of the covariates.

When the linear model is correct. Suppose that the linear model is correct, i.e., $Y = \bar{\gamma}^T Z + \epsilon$ for some $\bar{\gamma} \in \mathbb{R}^{p+1}$, and ϵ is a noise such that $\epsilon \perp Z$ and $\mathbb{E}[\epsilon|Z] = 0$.

Then the coefficient that minimizes the MSE will be

$$\begin{aligned} \gamma^* &= \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZY] \\ &= \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[Z(\bar{\gamma}^T Z + \epsilon)] \\ &= \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[Z(Z^T\bar{\gamma} + \mathbb{E}[\epsilon|Z])] \\ &= \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZZ^T\bar{\gamma}] \\ &= \bar{\gamma}. \end{aligned}$$

Thus, the least squares coefficient is the same as the true coefficient. This also implies that the least squares coefficient will be *invariant* to the distribution of the covariates when the linear model is correct.

5.4 Prediction of categorical outcomes: classification

All computations above were done with the understanding that we are relying on quadratic loss. We could have also chosen a different loss function. For instance, if you are interested in *robust statistics* (topic for another course) you may be interested in computing the risk under the Huber loss function, which for some $\delta > 0$ takes the following form:

$$L_\delta(y, g(x)) = \begin{cases} \frac{1}{2}(y - g(x))^2 & \text{for } |y - g(x)| \leq \delta, \\ \delta \cdot \left(|y - g(x)| - \frac{1}{2}\delta\right), & \text{otherwise.} \end{cases}$$

The Huber Loss has a quadratic form close to 0 and is otherwise linear. It's used for regression tasks that may have some outliers present in the data.

One other reason why you may consider a different loss function than the quadratic loss is if you response variable is discrete. In particular, the prediction problems where the response variable is categorical are known as *classification problems*.

Some examples of classification problems are below:

- **Email spam.** Deciding whether an email is spam or not, based on the email text and/or title.
- **Sentiment analysis.** Deciding whether a review is positive or negative, based on the text of the review.
- **Image classification.** Classifying pictures of animals by species, for instance. (Many categories for Y in this case.)

We'll consider the simple scenario – when Y is binary, known as binary classification. We will denote the two classes of Y by 0 and 1.

Our goal is to construct a classifier $c(X)$ based on our knowledge of random variable X, that does a "good job" at approximating Y. To decide how well we are doing at approximating Y, we need to define a loss function.

Since Y is binary, we may want to minimize the classification error, that is our loss could be

$$L(c(x), y) = \mathbb{1}(y \neq c(x)).$$

This loss function is known as the **0-1 loss**, because $L(c(x), y) = 1$ if our classification is wrong, and $L(c(x), y) = 0$, if our classification is correct.

How do we find the classifier c: we will minimize the risk function as before. The risk of a classifier c is

$$R(c) = \mathbb{E}(L(c(X), Y)).$$

Suppose that we know the distribution $P(y|x)$. An intuitive choice for a classifier c is then as follows:

$$c^*(x) = \mathsf{argmax}_{y=0,1} P(y|x) = \begin{cases} 0, & \text{if } P(0|x) \geq P(1|x), \\ 1, & \text{if } P(1|x) > P(0|x). \end{cases} \tag{5.4}$$

Namely, we predict the response (sometimes called label in classification) as the category with the highest conditional probability. This particular classifier is known as the **Bayes classifier.**

Is this classifier good in the sense of the classification error (risk)? In fact, yes! This is the optimal classifier (best predictor) for the $0-1$ loss, that is, $R(c^*) = \min_c R(c)$.

Derivation. Given a classifier c, the risk function $R(c) = \mathbb{E}(L(c(X), Y))$. Using tower property, we can further write it as

$$R(c) = \mathbb{E}(L(c(X), Y)) = \mathbb{E}(\underbrace{\mathbb{E}(L(c(X), Y)|X)}_{(A)}).$$

For the quantity (A), we have

$$\begin{aligned}
\mathbb{E}(L(c(X), Y)|X) &= L(c(X), 1)p(Y = 1|X) + L(c(X), 0)p(Y = 0|X) \\
&= \mathbb{1}(c(X) \neq 1)p(Y = 1|X) + \mathbb{1}(c(X) \neq 0)p(Y = 0|X) \\
&= \begin{cases} p(Y = 1|X) & \text{if } c(X) = 0 \\ p(Y = 0|X) & \text{if } c(X) = 1. \end{cases}
\end{aligned}$$

The optimal classifier should predict $c(X) = 0$ if $P(Y = 1|X) \leq P(Y = 0|X)$ and $c(X) = 1$ if $P(Y = 1|X) > P(Y = 0|X)$, which is exactly what the Bayes classifier does.

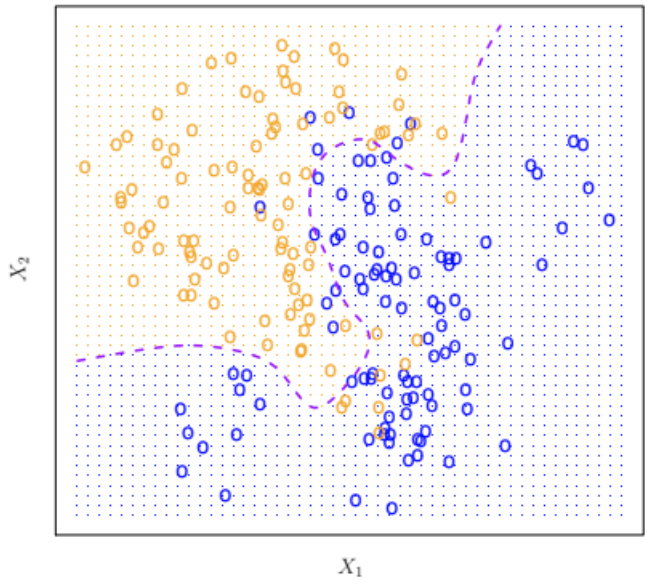


FIGURE 2.13. *A simulated data set consisting of* 100 *observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.*

Figure 5.1: Figure from ISLR2.

Note that our classifier c_*, depended on us knowing the distribution $P(y|x)$. What happens if we do not know this distribution? In these cases, we will obtain a classifier through some estimation procedure (for instance we may try to maximize $\widehat{P(y|x)}$ using maximum likelihood estimation, see next chapter). Such a classifier will necessarily have a larger (or equal) risk compared to the Bayes classifier. For such a classifier c, we define its **excess risk (regret)** as

$$\mathcal{E}(c) = R(c) - \min_c R(c).$$

The excess risk is a quantity that measures how much c compares to the optimal/Bayes classifier. If we cannot compute the optimal classifier, we will at least try to find a classifier whose excess risk is small.

References

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Perlman, M. D. (2019). Probability and mathematical statistics i. https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf.

STAT 512: Statistical Inference **Original notes credit: Yen-Chi Chen**

Lecture 6: Estimators

Instructor: Emilija Perković Compiled on: 2023-11-29, 08:34:07

Additional reading: Chapter 7.2 of Casella and Berger (2021).

In statistics, we often encounter a problem where we observe a sequence of random variables (data) $X_1, \cdots, X_n$ that represent random sample from a population and we wish to use this data to estimate some characteristics of the population.

For instance, we may assume that $X_1, \cdots, X_n$ are IID random variables from an unknown PDF p, and our goal is to estimate some parameters of p.

To make our lives easier, we may assume that p belongs to some *parametric family* of distributions (for instance the exponential family, see below). That is, $p(x) = p(x;\theta)$, where θ are some distribution parameters. In this case, we say that we are using a *parametric model*. For instance, if the population is normally distributed $\mathcal{N}(\mu, \sigma^2)$, then $\theta = (\mu, \sigma^2)$ consists of the mean and variance parameters.

An *estimator* is a statistic, that is a function of the data, $g(X_1, \cdots, X_n)$ that approximates θ in some way[1]. In this lecture, we discuss some popular approaches to finding a good estimator.

6.1 Detour: Exponential Family

One parametric family that will make a lot of appearances in STAT 513 is the exponential family of distributions.

Definition 6.1 *A parametric family of univariate distributions is said to belong to an exponential family of distributions if and only if the probability density function (or probability mass function in the case of discrete distributions) of any member of the family can be written as*

$$p_X(x;\theta) = h(x)\exp\big[\sum_{i=1}^{L}\eta_i(\theta)T_i(x) - A(\theta)\big],$$

where:

- $h : \mathbb{R} \to \mathbb{R}_+$*, is a function that only depends on* x*,*
- θ *is a* $K \times 1$ *vector of parameters;*
- $\eta = (\eta_1, \dots \eta_L)^T$*, for* $i \in \{1, \dots, L\}$*,* $L \geq K$*, and* $\eta_i : \mathbb{R}^K \to \mathbb{R}$ *is a function of the vector of parameters* θ*;*
- $T = (T_1, \dots T_L)^T$*, for* $i \in \{1, \dots, L\}$*,* $L \geq K$*, and* $T_i : \mathbb{R}^K \to \mathbb{R}$ *is a function of* x*;*
- $\eta(\theta)^T T(x)$ *is the dot product between* η *and* T*;*

[1] The concept of estimator can be generalized to other parameters of interest, not necessarily a parameter in a parametric model. For instance, we may be interested in the median of a distribution - which is a parameter of the distribution - but we may not want to assume that the the distribution is Gaussian. See lecture 10.

- $A : \mathbb{R}^K \to \mathbb{R}$ *is a function of* θ.

Example. (Normal is part of the exponential family) Note that for $X \sim \mathcal{N}(\mu, \sigma^2)$ and $\theta = (\mu, \sigma^2)$, the PDF of X is

$$\begin{aligned} p_X(x; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big[-\frac{(x-\mu)^2}{2\sigma^2}\big] \\ &= \frac{1}{\sqrt{2\pi}} \exp\big[\log\frac{1}{|\sigma|}\big] \exp\big[-\frac{(x-\mu)^2}{2\sigma^2}\big] \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[\big(-\log|\sigma| - \frac{(x-\mu)^2}{2\sigma^2}\big)\right] \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[\big(-\log|\sigma| - \frac{\mu^2}{2\sigma^2} - \frac{x^2 - 2\mu x}{2\sigma^2}\big)\right] \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[\big(-\log|\sigma| - \frac{\mu^2}{2\sigma^2} + \frac{-x^2}{2\sigma^2} + \frac{2\mu x}{2\sigma^2}\big)\right] \end{aligned}$$

We can notice above, that we can choose the following functions for h, A, T, and η.

- $h(x) = 1/\sqrt{2\pi}$,
- $A(\mu, \sigma^2) = \log|\sigma| + \frac{\mu^2}{2\sigma^2}$, and
- $\eta_1(\mu, \sigma^2)T_1(x) = \frac{-x^2}{2\sigma^2}$, that is $\eta_1(\mu, \sigma^2) = \frac{-1}{2\sigma^2}$, and $T_1(x) = x^2$.
- $\eta_2(\mu, \sigma^2)T_2(x) = \frac{2\mu x}{2\sigma^2}$, that is $\eta_2(\mu, \sigma^2) = \frac{\mu}{\sigma^2}$, and $T_2(x) = x$.

Hence, the Normal distribution belongs to the exponential family.

Aside: A distribution belonging to the exponential family is said to be *flat*, if $K = L$, and *curved* if $K < L$. Example of a *curved* distribution in the exponential family is $\mathcal{N}(\theta, \theta^2)$.

6.2 Method of Moments

The method of moments is a simple but powerful approach to finding an estimator. The idea is as follows. For a parametric model $p(x; \theta)$, its moments are determined by the underlying parameter θ. For instance, the first moment is

$$m_1(\theta) = \mathbb{E}[X] = \int x p(x; \theta) dx$$

and the second moment is

$$m_2(\theta) = \mathbb{E}[X^2] = \int x^2 p(x; \theta) dx$$

Suppose that we have k parameters in the model, i.e., $\theta = (\theta_1, \ldots, \theta_k)^T \in \mathbb{R}^k$. Then we can use the first k moments to express $\theta_1, \ldots, \theta_k$, i.e.,

$$m_j(\theta_1, \ldots, \theta_k) = \int x^j p(x; \theta_1, \ldots, \theta_k) dx,$$

for $j = 1, 2, 3, \cdots, k$.

So for instance, in the case of a normal distribution $\mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)^T$, we have that

$$\begin{aligned} m_1(\theta) &= \mathbb{E}[X] = \mu, \text{ and} \\ m_2(\theta) &= \mathbb{E}[X^2] = \mathsf{Var}(X) + (\mathbb{E}[X])^2 = \sigma^2 + \mu^2. \end{aligned}$$

How do we use this in estimation? Recall that the EDF $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$ is a good estimator of the true CDF F (Lecture notes 4). So if we want to estimate a population quantity θ that is a function of F, $\theta = T_{\mathsf{target}}(F)$, we can use $T_{\mathsf{target}}(\widehat{F}_n) = \widehat{\theta}_n$ as our estimator. Method of moments exploits exactly this connection. Thus:

$$\widehat{m_j(\theta)} = \frac{1}{n}\sum_{i=1}^n X_i^j$$

for each $j = 1, 2, 3, \cdots$. We obtain the estimator for θ, by finding $\widehat{\theta}$ that solves the following system of equations:

$$\begin{aligned} \widehat{m_1(\theta)} &= \frac{1}{n}\sum_{i=1}^n X_i \\ \widehat{m_2(\theta)} &= \frac{1}{n}\sum_{i=1}^n X_i^2 \\ \vdots \qquad & \qquad \vdots \\ \widehat{m_k(\theta)} &= \frac{1}{n}\sum_{i=1}^n X_i^k. \end{aligned}$$

The resulting quantity $\widehat{\theta}_{MoM}$ is called the **method of moments estimator**.

Example: Normal distribution. Consider $X_1, \cdots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$. We want to estimate $\theta = (\mu, \sigma^2)^T$. Since we have two parameters we want to estimate, we use the first two moment equations. That is,

$$m_1(\mu, \sigma^2) = \mu, \quad m_2(\mu, \sigma^2) = \mu^2 + \sigma^2.$$

Thus, we immediately have

$$\widehat{\mu} = \widehat{m}_1(\widehat{\mu}, \widehat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^n X_i$$

and

$$\widehat{\mu}^2 + \widehat{\sigma}^2 = \widehat{m}_2(\widehat{\mu}, \widehat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^n X_i^2,$$

which leads to

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu})^2.$$

Example: Uniform distribution. Suppose that $X_1, \cdots, X_n \overset{IID}{\sim} \mathsf{Uniform}[0, \theta]$. We want to estimate θ. By the method of moments:

$$\widehat{\theta}/2 = \widehat{m}_1(\theta) = \frac{1}{n}\sum_{i=1}^n X_i$$

so $\widehat{\theta}_{MoM} = \frac{2}{n}\sum_{i=1}^{n} X_i$.

Example: Exponential distribution. Consider the case where we use the exponential distribution to model $X_1, \cdots, X_n$. Since $p(x;\lambda) = \lambda e^{-\lambda x} I(x \geq 0)$, we have

$$m_1(\lambda) = \frac{1}{\lambda}.$$

As a result,

$$1/\widehat{\lambda} = \widehat{m}_1(\widehat{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\widehat{\lambda}_{MoM} = \frac{n}{\sum_{i=1}^{n} X_i}.$$

6.3 Maximum likelihood estimators

Another popular estimation procedure is the maximum likelihood estimation. For this procedure, we again assume that our random variables of interest belong to a certain parametric model. That is, the PDF/PMF can be written as $p(x) = p(x;\theta)$, where $\theta \in \Theta$, for some parameter space Θ.

The idea behind maximum likelihood estimation, is to treat each observation of a random variable as having a certain likelihood of being drawn that depends on θ. Then ask, given this observation X, which θ is the *most likely* parameter to have generated it? To answer this question, we can vary θ and examine the value of $p(X;\theta)$.

Because we are treating X as fixed and θ as a variable we want to optimize, we can view the problem as finding the best θ such that the **likelihood function** $L(\theta|X) = p(X;\theta)$ is maximized. The maximum likelihood estimator (MLE), will be the θ that leads to $\max_\theta L(\theta|X)$. Namely,

$$\widehat{\theta}_{MLE} = \mathsf{argmax}_\theta L(\theta|X).$$

When we assume we have observations of n random variables $X_1, \cdots, X_n$ (n samples), the likelihood function will be

$$L_n(\theta) = L(\theta|X_1, \cdots, X_n) = p(X_1, \cdots, X_n; \theta).$$

If we further assume that our variables are IID, we can further decompose the likelihood,

$$L_n(\theta) = \prod_{i=1}^{n} L(\theta|X_i) = \prod_{i=1}^{n} p(X_i;\theta).$$

For most (if not all) problems in this course, we will assume that our n random variables are IID.

As finding the maximum of a product is both numerically and analytically complicated, we consider a monotonic transformation that preserves the $argmax$ while making the problem easier to analyze. That is, instead analyzing the likelihood function, we will analyze the **log-likelihood function**

$$\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log p(X_i;\theta).$$

Under the IID assumption, each $\log p(X_i;\theta)$ is an IID random variable. So we can make use of the central limit theorem and the law of large numbers, making it possible to analyze it asymptotic behavior.

Now, to find the extremum of the log-likelihood function above, we study the first derivative of this function, also known as the gradient, or *the score function*.

$$S_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta) = \sum_{i=1}^{n} s(\theta|X_i),$$

where $s(\theta|X_i) = \frac{\partial}{\partial\theta}\ell(\theta|X_i) = \frac{\partial}{\partial\theta}\log p(X_i;\theta)$. The extrema points (either maximums or minimums) of the likelihood function, can be find as the solution to the following *score equation*:

$$S_n(\widehat{\theta}^*) = 0.$$

Note that $\theta \in \mathbb{R}^p$, the score equation will become a system of p equations. These equations are also known as the likelihood equations. Under suitable conditions, $\theta^* \equiv \widehat{\theta}_{MLE}$.

For θ^* to be the maximum of the likelihood function, we need the log-likelihood function to be log-concave. Hence, we may consider the second derivative of the log-likelihood at $\theta = \theta^*$. If this derivative is negative, we have found the maximum, that is $\theta^* \equiv \widehat{\theta}_{MLE}$. In the case, of $\theta \in \mathbb{R}^p$, we consider the Hessian of the log-likelihood for $\theta = \theta^*$ and show it's negative definite. Conveniently, most common probability distributions – in particular the exponential family – is log-concave.

The second derivative of the score function around θ^*, will also give us some information about the stability of our found maximum. This will further allow us to study the variance of our estimator $\widehat{\theta}_{MLE}$. We first define some important notation.

Definition 6.2 (Fisher information) *Fisher information for a single observation of a random variable X_1 and $\theta \in \mathbb{R}$ is defined as*

$$I_1(\theta) = \mathbb{E}\left[\left(\frac{\partial\ell_1(\theta)}{\partial\theta}\right)^2\right].$$

For n samples of IID random variables $X_1, \ldots, X_n$, the Fisher information is

$$I(\theta) = nI_1(\theta).$$

*If $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$, then $I_1(\theta)$ and $I(\theta)$ become **Fisher information matrices** with i,j-th elements equal to*

$$[I_1(\theta)]_{i,j} = \mathbb{E}\left[\left(\frac{\partial\ell_1(\theta)}{\partial\theta_i}\right)\left(\frac{\partial\ell_1(\theta)}{\partial\theta_j}\right)\right] \text{ and } [I(\theta)]_{i,j} = n[I_1(\theta)]_{i,j}.$$

Theorem 6.3 *Under some regularity conditions, for $\theta \in \mathbb{R}$, the following holds:*

$$\mathbb{E}\left(\frac{\partial\ell_1(\theta)}{\partial\theta}\right) = 0, \text{ therefore, } I_1(\theta) = \mathrm{Var}\left(\frac{\partial\ell_1(\theta)}{\partial\theta}\right).$$

By above, the Fisher information will be the variance of the score.

Theorem 6.4 *The Fisher information for a single observation of a random variable X_1 can be derived from the second derivative, that is for $\theta \in \mathbb{R}$*

$$I_1(\theta) = -\mathbb{E}\left(\frac{\partial^2\ell_1(\theta)}{\partial^2\theta}\right),$$

or for $\theta \in \mathbb{R}^p$

$$[I_1(\theta)]_{i,j} = -\mathbb{E}\left(\frac{\partial^2\ell_1(\theta)}{\partial\theta_i\partial\theta_j}\right).$$

Suppose for simplicity that $\theta \in \mathbb{R}$. By applying the CLT, one can show

$$\sqrt{n}\left(\widehat{\theta}_{MLE} - \theta_0\right) \xrightarrow{D} N(0, I_1^{-1}(\theta_0)),$$

where θ_0 is the true population value of θ. Namely, the MLE is asymptotically normally distributed around the true parameter θ_0, and the covariance is determined by the Fisher information matrix. Note that the asymptotic normality also implies that $\widehat{\theta}_{MLE} - \theta_0 \xrightarrow{P} 0$.

We now introduce a very important result, that will be explored deeper in STAT 513:

Theorem 6.5 (Cramér-Rao Lower Bound) *Let $X_1, \ldots, X_n$ be IID random variables each with PDF/PMF $p_{X_1;\theta}$ and supposed $\widehat{\theta}$ is an unbiased estimator for θ_0, that is $\mathbb{E}[\widehat{\theta}] = \theta_0$. Then*

$$\mathrm{Var}(\widehat{\theta}) \geq \frac{1}{I(\theta_0)} = \frac{1}{nI_1(\theta_0)}.$$

An estimator is said to be **efficient** if it reaches the Cramér-Rao lower bound. We can conclude from above that this will be true for an MLE (under the correct parametric model assumption). Hence, one reason for its popularity.

Example 1: Binomial Distribution. Assume that we obtain a single observation $Y \sim \mathsf{Bin}(N, p)$, and we assume that N is known. The goal is to estimate p. The log-likelihood function is

$$\ell(p) = Y \log p + (N - Y)\log(1 - p) + C_N(Y),$$

where $C_N(Y) = \log\binom{N}{Y}$ is independent of p. The score function is

$$S(p) = \frac{Y}{p} - \frac{N - Y}{1 - p}$$

so solving $S(p) = 0$ gives us $\widehat{p}_{MLE} = \frac{Y}{N}$. Moreover, the Fisher information is

$$I(p) = -\mathbb{E}\left\{\frac{\partial}{\partial p}S(p)\right\} = +\frac{\mathbb{E}(Y)}{p^2} + \frac{N - \mathbb{E}(Y)}{(1-p)^2} = \frac{N}{p(1-p)}.$$

Example 2: Poisson Distribution. Suppose we observe two integer RVs X_1, X_2. We assume that they are independently from Poisson distribution with unknown parameter λ. What will be the MLE of λ? In this case, the joint PDF is

$$p(x_1, x_2; \lambda) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\frac{\lambda^{x_2}}{x_2!}e^{-\lambda}.$$

Thus, the log-likelihood function will be

$$\ell(\lambda|X_1, X_2) = (X_1 + X_2)\log\lambda - 2\lambda - \log(X_1!) - \log(X_2!)$$

so the score function is

$$S(\lambda|X_1, X_2) = \frac{X_1 + X_2}{\lambda} - 2.$$

This leads to the MLE:

$$\widehat{\lambda} = \frac{1}{2}(X_1 + X_2).$$

Example 3: Uniform Distribution. Consider $X_1, \cdots, X_n \overset{IID}{\sim} \mathsf{Unif}[0, \theta]$. What will be the MLE of θ? Recall that the PDF will be

$$p(x_1, \cdots, x_n) = \prod_{i=1}^{n}\frac{1}{\theta}I(0 \leq x_i \leq \theta).$$

So the likelihood function is

$$L(\theta|X_1, \cdots, X_n) = \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 \leq X_i \leq \theta).$$

An interesting fact is that

$$\prod_{i=1}^{n} I(0 \leq X_i \leq \theta) = I(0 \leq X_{\min} \leq X_{\max} \leq \theta),$$

where $X_{\min} = \min\{X_1, \cdots, X_n\}$ and $X_{\max} = \max\{X_1, \cdots, X_n\}$. So the likelihood function increases when θ decreases. However, it will drop to 0 immediately when $\theta < X_{\max}$. Thus, the MLE of θ will be $\widehat{\theta} = X_{\max}$.

6.4 Bayesian estimators

Bayesian statistics is an alternative statistical paradigm that treats population parameters as random variables rather than fixed values. In Bayesian analysis, we assign a probability distribution to every parameter in our model based on some prior beliefs. Thus, for a parametric model $p(x;\theta)$, the parameter of interest θ is assumed to follow a **prior distribution** $\pi(\theta)$. The Bayesian probability model can then be written as follows:

$$\begin{aligned} X_1, \cdots, X_n|\theta &\overset{IID}{\sim} p(x|\theta) \\ \theta &\sim \pi. \end{aligned}$$

One of the goals of Bayesian inference is updating our distribution of θ after observing $X_1, \cdots, X_n$. This new distribution is known a **posterior distribution** for θ. Bayesian estimators of θ will be some functions of the posterior distribution. The posterior distribution of θ given $X_1, \ldots, X_n$, denoted by $\pi(\theta|X_1, \cdots, X_n)$ is computed using the Bayes formula. For instance, if the prior of θ is discrete:

$$\pi(\theta|X_1, \cdots, X_n) = \frac{p(X_1, \cdots, X_n, \theta)}{p(X_1, \cdots, X_n)} = \frac{p(X_1, \cdots, X_n|\theta)\pi(\theta)}{\sum_{\tilde{\theta}} p(X_1, \cdots, X_n|\tilde{\theta})\pi(\tilde{\theta}).} \propto \underbrace{p(X_1, \cdots, X_n|\theta)}_{\text{likelihood}} \times \underbrace{\pi(\theta)}_{\text{prior}}.$$

The posterior distribution reflects our belief about the parameter after seeing the data and we can use it as a measure of *uncertainty* about θ. If the posterior distribution is more spread out, then the uncertainty in our inference is larger. On the other hand, if the posterior distribution is very concentrated, then there is very little (Bayesian) uncertainty.

There are two common estimators in Bayesian inference: the posterior mean and the maximum a posteriori (MAP) estimator.

Posterior mean. The posterior mean, $\widehat{\theta}_\pi = \mathbb{E}[\theta|X_1, \ldots, X_n]$, can be seen as the estimator of θ that minimizes some mean square error (recall Lecture 5).

$$\widehat{\theta}_\pi = \mathbb{E}[\theta|X_1, \cdots, X_n] = \int \theta \cdot \pi(\theta|X_1, \cdots, X_n) d\theta.$$

Maximum a posteriori (MAP) estimation. Another common estimator of θ is the MAP estimator; it relies on a principle similar to the MLE – we choose as $\widehat{\theta}_{MAP}$ the value that is most likely a posteriori. Formally, the MAP estimator is defined as

$$\widehat{\theta}_{MAP} = \mathsf{argmax}_\theta \pi(\theta|X_1, \cdots, X_n).$$

Example: Beta-Binomial-Beta. Assume that we have an observation (random variable) $Y \sim \mathsf{Bin}(N, \theta)$ where N is known and the parameter of interest is θ:

$$P(Y = y|\theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y}.$$

We assume that the prior distribution of θ is a Beta distribution with known parameters (α, β), $\alpha, \beta > 0$. Namely,

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the Gamma function. Note that α and β are sometimes called hyperparameters. The prior mean of θ is $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$.

The posterior distribution of θ is

$$\begin{aligned}\pi(\theta|Y) &= \frac{\binom{N}{Y}\theta^Y(1-\theta)^{N-Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \binom{N}{Y}\theta^Y(1-\theta)^{N-Y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} \\ &\propto \theta^{Y+\alpha-1}(1-\theta)^{N-Y+\beta-1}.\end{aligned}$$

With a little bit more computation, we can confirm that the posterior distribution of θ is a again a Beta distribution but now with parameters $(Y+\alpha, N-Y+\beta)$. Then the posterior mean and MAP estimators are

$$\widehat{\theta}_\pi = \frac{Y+\alpha}{N+\alpha+\beta}, \qquad \widehat{\theta}_{MAP} = \frac{Y+\alpha-1}{N+\alpha+\beta-2}$$

(these are the mean and the mode of a Beta distribution).

Note that in this problem, the MLE for θ is $\widehat{\theta}_{MLE} = \frac{Y}{N}$, (computed by maximizing the likelihood $L(\theta)$ which assumes that θ is fixed.

Thus, the posterior mean has an interesting decomposition:

$$\begin{aligned}\widehat{\theta}_\pi &= \frac{Y+\alpha}{N+\alpha+\beta} \\ &= \widehat{\theta}_\pi = \frac{Y}{N+\alpha+\beta} + \frac{\alpha}{N+\alpha+\beta} \\ &= \frac{Y}{N} \times \frac{N}{N+\alpha+\beta} + \frac{\alpha}{\alpha+\beta} \times \frac{\alpha+\beta}{N+\alpha+\beta} \\ &= \widehat{\theta}_{MLE} \times W + [\text{Prior mean}] \times (1-W),\end{aligned}$$

where $W = \frac{N}{N+\alpha+\beta}$ tends to 1 when $N \to \infty$. This phenomenon that the posterior mean can be written as a weighted average of the MLE and the prior mean occurs in some cases (but does not hold for every Bayesian posterior mean estimator). Moreover, the fact that the weights $W \to 1$ as the sample size $N \to \infty$ imply that with a large enough data set, the prior distribution becomes almost irrelevant for $\widehat{\theta}_\pi$.

Example: Normal-Normal-Normal. Suppose that $X_1, \cdots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$, and that σ^2 is known, so that $\theta \equiv \mu$. Furthermore, suppose that we assume the prior distribution of μ is $\mathcal{N}(\xi, \tau^2)$, where ξ, τ^2 are pre-specified. Now, we derive the posterior distribution of μ given $X_1, \cdots, X_n$ and the specified parameters. Hence,

$$\pi(\mu|X_1, \cdots, X_n) \propto \exp(-\frac{1}{2\tau^2}(\mu-\xi)^2)\prod_{i=1}^n \exp(-\frac{1}{2\sigma^2}(X_i-\mu)^2).$$

Note that $\log \pi(\mu|X_1, \cdots, X_n)$ will be a quadratic function of μ. After some computation, we can derive that $\pi(\mu|X_1, \cdots, X_n)$ will still be a normal distribution but now with new parameters. In fact,

$$\log \pi(\mu|X_1, \cdots, X_n) = C_0 - \frac{1}{2\tau^2}(\mu - \xi)^2 - \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(X_i - \mu)^2,$$

which after some further computation

$$\begin{aligned}
\mathbb{E}[\mu|X_1, \cdots, X_n] &= \frac{\tau^2}{\tau^2 + \sigma^2/n}\overline{X}_n + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\xi \\
\mathsf{Var}(\mu|X_1, \cdots, X_n) &= \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}.
\end{aligned}$$

In this case, again, the posterior mean can be written as the weighted average of the prior mean and the MLE estimate.

$$\begin{aligned}
\widehat{\mu}_\pi &= \frac{\tau^2}{\tau^2 + \sigma^2/n}\overline{X}_n + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\xi \\
&= \widehat{\mu}_{MLE} \times W + \xi \times (1 - W) \\
&= \text{MLE} \times W + [\text{Prior mean}] \times (1 - W),
\end{aligned}$$

where $W = W_n = \frac{\tau^2}{\tau^2 + \sigma^2/n} \overset{n\to\infty}{\to} 1$.

Remark.

- *Choice of prior and conjugate prior.* Sometimes it is convenient to choose a prior distribution for θ such that given $p(X_1, \ldots, X_n|\theta)$, we know that the posterior distribution of θ will be in the same family as the prior. We observed to examples of this above, Beta $\to$ Beta, and Normal $\to$ Normal. If a prior distribution and a likelihood function lead to a posterior that belongs to the same family as the prior, we call this prior a **conjugate prior**. There are several conjugate priors know to date, see https://en.wikipedia.org/wiki/Conjugate_prior for an incomplete list of cases.
- *Uninformative priors.* If you do not have any prior belief about θ, you may want to choose a prior that is as "uninformative" as possible. This turns out to be very difficult – perhaps impossible –, see e.g. this discussion on stackexchange or this article Noninformative Priors Do Not Exist: A Discussion with José M. Bernardo (that is also a discussion). Still, a common choice for researchers looking for a prior that is as uninformative as possible is the **Jeffreys prior**[2], which equals $\pi(\theta) \propto \sqrt{\det(I_1(\theta))}$, where $I_1(\theta)$ is the Fisher information matrix.
- *Challenge in computing the posterior.* If we do not choose a conjugate prior, the posterior distribution could be difficult to compute. The challenge often comes from the normalization quantity $p(X_1, \cdots, X_n)$ in the denominator of the posterior $\pi(\theta|X_1, \cdots, X_n)$. In these cases, numerical methods such as the Monte Carlo method, are used to estimate posterior – the intuition being is that if you have a way to generate enough points from $\pi(\theta|X_1, \cdots, X_n)$ the empirical posterior should approximate the true posterior distribution well.
- *Consistency.* In a Bayesian point of view, statistical consistency (convergence in probability to the true parameter) is not an important property because there is no single true parameter. Thus, the posterior distribution is the quantity that we really need to make our inference. However, sometimes Bayesian estimators, such as the posterior mean or MAP, do have statistical consistency. Namely, $\widehat{\theta}_\pi \overset{P}{\to} \theta_0$ and $\widehat{\theta}_{MAP} \overset{P}{\to} \theta_0$, where the data $X_1, \cdots, X_n \overset{IID}{\sim} p(x; \theta_0)$. This is related to the Bernstein-von Mises theorem[3].

[2] see https://en.wikipedia.org/wiki/Jeffreys_prior for more details.

[3] https://en.wikipedia.org/wiki/Bernstein%E2%80%93von_Mises_theorem

6.5 Empirical risk minimization (ERM) and M-estimation

Recall that in Lecture 5, our goal was to find an estimator (predictor) that minimizes the risk function. Another approach for computing "good" estimators is known as **empirical risk minimziation (ERM)**. This approach is widely used in machine learning and many modern statistical procedures. In fact, this methods can be viewed as a generalization of the maximum likelihood estimation, since for certain loss functions, the estimators derived by ERM match the MLE (An example is included in Homework 4!)

6.5.1 Motivation: least squares estimates

Consider again the linear regression problem where we observe $(X_1, Y_1), \cdots, (X_n, Y_n)$, and want to find a good estimator of Y that is a linear function of X. We assume that each X_i is $p+1$ dimensional vector, that is, $X_i = (X_{i,1} = 1, \cdots, X_{i,p+1})$ (X_i is now playing the role of Z from Lecture 5). Furthermore, we may assume that

$$\mathbb{E}[Y|X] = X^T\beta,$$

where $\beta \in \mathbb{R}^{p+1}$.

Ideally, we want to compute the estimate of β by minimizing the MSE, i.e.,

$$\beta^* = \mathsf{argmin}_\beta R(\beta) = \mathsf{argmin}_\beta \mathbb{E}[(Y - X^T\beta)^2].$$

However, if we do not know the distribution of X, Y (and do not want to assume one), we cannot compute the above estimator. Instead, we can choose to approximate the MSE using *empirical mean square errors*, i.e., we approximate $R(\beta)$ by

$$\widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2.$$

The estimator we derive in this case, will be the *least squares estimate (LSE)*:

$$\widehat{\beta} = \mathsf{argmin}_\beta \widehat{R}(\beta) = \mathsf{argmin}_\beta \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2.$$

The ERM takes the above idea and runs with it, applying the same estimation process with various loss functions.

6.5.2 A general ERM approach

Recall that in prediction, a loss function $L : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ is a function that measures the quality of prediction (or estimation). Note that $\mathcal{Y}$ is the support of Y. As discussed, a popular loss function is the square loss, i.e., $L(a, b) = (a - b)^2$, $a, b \in \mathbb{R}$. But we can also consider more complex functions such as the absolute loss, $L(a, b) = |a - b|$, the Huber loss from Lecture 5, or others.

As mentioned in Lecture 5, the risk will be equal to the expected value of the loss. In the case of mean square prediction the risk is $R(\beta) = \mathbb{E}[L(Y, f_\beta(X))] = \mathbb{E}[(Y - X^T\beta)^2]$. In general, with any loss function L, the risk function is

$$R(\beta) = \mathbb{E}[L(Y, f_\beta(X))].$$

We may not be able to compute the risk function analytically if we do not know the joint distribution of (X, Y). Instead, we will approximate the risk with something computable from data, in our case, the

empirical risk:

$$\widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f_\beta(X_i)).$$

With enough data, the empirical risk, should approximate the true risk (similar intuition as with empirical CDF). We then construct $\widehat{\beta}_{ERM}$ by minimizing $\widehat{R}(\beta)$, namely,

$$\widehat{\beta}_{ERM} = \mathsf{argmin}_\beta \widehat{R}(\beta).$$

Example: least absolute deviation. Consider the loss function $L(a, b) = |a - b|$. Then the regression estimator

$$\widehat{\beta}_{LAD} = \mathsf{argmin}_\beta \widehat{R}(\beta) = \mathsf{argmin}_\beta \frac{1}{n}\sum_{i=1}^{n} |Y_i - X_i^T\beta|$$

is called the least absolute deviation (LAD) estimator. It is more robust against outliers compared to the LSE due to use of L_1 norm (absolute value) as the loss function.
Note: the LSE is approximating the conditional mean of Y given X by a linear function; the LAD will be approximating the conditional *median* of Y given X by a linear function.

6.5.3 M-estimation

The ERM is actually a special case of a more general procedure called **M-estimation**. In M-estimation, we compute an estimator by maximizing an empirical objective function, i.e.,

$$\widehat{\theta} = \mathsf{argmax}_\theta \frac{1}{n}\sum_{i=1}^{n} \omega(\theta; X_i)$$

for some function ω. To see why you can consider this process as a generalization of ERM, note that we can write:

$$\mathsf{argmin}_\beta \widehat{R}(\beta) = \mathsf{argmax}_\beta(-\widehat{R}(\beta)).$$

Note that for M-estimation, we do not need to be optimizing the risk, but can choose to optimize a different function. It is easy to see that if we choose the objective function to be the log-likelihood function

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta|X_i) = \frac{1}{n}\sum_{i=1}^{n} \log p(X_i; \theta),$$

then the M-estimator is the MLE.

References

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

STAT 512: Statistical Inference **Original notes credit: Yen-Chi Chen**

Lecture 7: Multinomial distribution

Instructor: Emilija Perković Compiled on: 2023-12-05, 08:25:28

Additional reading: Chapter 7 of Perlman (2019).

The multinomial distribution is a common distribution for characterizing categorical variables. Suppose a random variable Z has k categories. We can code each category as an integer, leading to $Z \in \{1, 2, \cdots, k\}$. Suppose that $P(Z = k) = p_k$. The parameter $\{p_1, \cdots, p_k\}$ describes the entire distribution of Z (with the constraint that $\sum_j p_j = 1$). Suppose we generate $Z_1, \cdots, Z_n$ IID from the above described distribution. and let the random vector X be such that, $X = (X_1, \cdots, X_k)$, where

$$X_j = \sum_{i=1}^{n} I(Z_i = j) = \#\text{ of observations in the category } j.$$

Then X is said to be multinomially distributed with parameter $(n, p_1, \cdots, p_k)$. We often write

$$X \sim M_k(n; p_1, \cdots, p_k)$$

to denote a multinomial distribution.

Example (pet lovers). The following is a hypothetical dataset about how many students prefer a particular animal as a pet. Each row (except the 'total') can be viewed as a random vector from a multinomial distribution. For instance, the first row $(18, 20, 6, 4, 2)$ can be viewed as a random draw from a multinomial distribution $M_5(n = 50; p_1, \cdots, p_5)$. The second and the third row can be viewed as other random draws from the same distribution.

	cat	dog	rabbit	hamster	fish	total
Class 1	18	20	6	4	2	50
Class 2	15	15	10	5	5	50
Class 3	17	18	8	4	3	50

7.1 Properties of multinomial distribution

The PMF of a multinomial distribution has a the following If $X \sim M_k(n; p_1, \cdots, p_k)$, then

$$p(X = x) = p(X_1 = x_1, \cdots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}.$$

The *multinomial coefficient* $\frac{n!}{x_1! x_2! \cdots x_k!} = \binom{n}{x_1, \cdots, x_n}$ is the number of possible ways to put n balls into k boxes.

Note that by the *multinomial theorem*

$$(a_1 + a_2 + \cdots + a_k)^n = \sum_{x_i \geq 0, \sum_i x_i = n} \frac{n!}{x_1! x_2! \cdots x_k!} a_1^{x_1} a_2^{x_2} \cdots a_k^{x_k}.$$

Hence, $\sum_{x_i \geq 0, \sum_i x_i = n} p(X = x) = (p_1 + p_2 + \cdots + p_k)^n = 1$.

By the construction of a multinomial $M_k(n; p_1, \cdots, p_k)$ above, we note that if $X \sim M_k(n; p_1, \cdots, p_k)$, then $X = \sum_{i=1}^n Z_i$, where $Z_1, \cdots, Z_n \in \{0,1\}^k$ are IID multinomial random variables with parameter $(1; p_1, \cdots, p_k)$, that is $Z_1, \cdots, Z_n \overset{IID}{\sim} M_k(1; p_1, \cdots, p_k)$.

The moment generating function of X can then be derived from the moment generating functions of Z_i's

$$M_X(s) = \mathbb{E}[e^{s^T X}] = \mathbb{E}[e^{s^T Z_1}]^n = \left(\sum_{j=1}^k p_j e^{s_j}\right)^n$$

The multinomial distribution various nice properties which we explore next.

7.1.1 Additive and marginal properties

Suppose $X \sim M_k(n; p_1, \cdots, p_k)$ and $Y \sim M_k(m; p_1, \cdots, p_k)$ and $X \perp Y$. Then we will have that,

$$X + Y \sim M_k(n + m; p_1, \cdots, p_k).$$

Suppose we focus on one particular category j, and consider the distribution of X_j. One can show that,

$$X_j \sim \mathsf{Bin}(n, p_j).$$

Note that $X_1, \cdots, X_k$ are not independent due to the constraint that $X_1 + X_2 + \cdots + X_k = n$. However, for any X_i and X_j elements of the random vector X, one can show that

$$X_i + X_j \sim \mathsf{Bin}(n, p_i + p_j).$$

An intuitive way to think of this is that $X_i + X_j$ will represent the number of observations in either category i or category j. So we are essentially pulling two categories together.

7.1.2 Conditional distributions of multinomials

Here we illustrate a property of the multinomial distributions using an example with $k = 4$, but this property applies to more general scenarios. Let $X = (X_1, X_2, X_3, X_4) \sim M_4(n; p_1, p_2, p_3, p_4)$. Suppose we combine the last two categories into a new category. Let $W = (W_1, W_2, W_3)$ be the resulting random vector. By construction, $W_3 = X_3 + X_4$ and $W_1 = X_1, W_2 = X_2$. Then

$$W \sim M_3(n, q_1, q_2, q_3), \quad q_1 = p_1, q_2 = p_2, q_3 = p_3 + p_4.$$

So pulling two or more categories together will result in a new multinomial distribution.

Let $Y = (Y_1, Y_2)$ such that $Y_1 = X_1 + X_2$ and $Y_2 = X_3 + X_4$. We know that $Y \sim M_2(n; p_1 + p_2, p_3 + p_4)$. What will the conditional distribution of $X|Y$ be?

$$\begin{aligned}
&P((X_1, X_2, X_3, X_4) = (x_1, x_2, x_3, x_4) | (Y_1, Y_2) = (y_1, y_2)) \\
&= \frac{P((X_1, X_2, X_3, X_4) = (x_1, x_2, x_3, x_4))}{P((Y_1, Y_2) = (y_1, y_2))} I(y_1 = x_1 + x_2, y_2 = x_3 + x_4) \\
&= \frac{\frac{n!}{x_1! x_2! x_3! x_4!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4}}{\frac{n!}{y_1! y_2!} (p_1 + p_2)^{y_1} (p_3 + p_4)^{y_2}} I(y_1 = x_1 + x_2, y_2 = x_3 + x_4) \\
&= \frac{(x_1 + x_2)!}{x_1! x_2!} \left(\frac{p_1}{p_1 + p_2}\right)^{x_1} \left(\frac{p_2}{p_1 + p_2}\right)^{x_2} \times \frac{(x_3 + x_4)!}{x_3! x_4!} \left(\frac{p_3}{p_3 + p_4}\right)^{x_3} \left(\frac{p_4}{p_3 + p_4}\right)^{x_4} \\
&= P((X_1, X_2) = (x_1, x_2) | Y_1 = y_1) P((X_3, X_4) = (x_3, x_4) | Y_2 = y_2)
\end{aligned}$$

which leads us to conclude that

$$X_1, X_2|X_1+X_2 \sim M_2\left(X_1+X_2; \frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right), \quad X_3, X_4|X_3+X_4 \sim M_2\left(X_3+X_4; \frac{p_3}{p_3+p_4}, \frac{p_4}{p_3+p_4}\right).$$

and that

$$(X_1, X_2) \perp (X_3, X_4)|Y,$$

i.e., they are conditionally independent.

Because $X_1 + X_2 = n - X_3 - X_4$, the above result also implies that

$$X_1, X_2|X_3, X_4 \stackrel{d}{=} X_1, X_2|n - X_3 - X_4 \sim M_2\left(n - X_3 - X_4; \frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right),$$

where $X \stackrel{d}{=} Y$ means that the two random variables have the same distribution. Thus, one can see that (X_1, X_2) and (X_3, X_4) have a negative linear relation.

General case. Suppose that we partition $X = (X_1, \cdots, X_k)$ into r blocks

$$\underbrace{(X_1, \cdots, X_{k_1})}_{B_1}, \underbrace{(X_{k_1+1}, \cdots, X_{k_2})}_{B_2}, \cdots, \underbrace{(X_{k_{r-1}+1}, \cdots X_{k_r})}_{B_r}.$$

Then we have $B_1, \cdots, B_r$ are conditionally independent given $S_1, \cdots, S_r$, where $S_1 = \sum_{i=1}^{k_1} X_i = \sum_j B_{1,j}$ and $S_r = \sum_{i=k_{r-1}+1}^{k_r} X_i = \sum_j B_{r,j}$ are the block-specific sum.

Also,

$$B_j|S_j \sim M_{k_j - k_{j-1}}\left(S_j; \frac{p_{k_{j-1}+1}}{\sum_{\ell=k_{j-1}+1}^{k_j} p_\ell}, \cdots, \frac{p_{k_j}}{\sum_{\ell=k_{j-1}+1}^{k_j} p_\ell}\right).$$

Now, we consider the case where $X \sim M_k(n; p_1, \cdots, p_k)$, and we focus on only two variables X_i and X_j $(i \neq j)$. What will the conditional distribution of $X_i|X_j$ be?

Using the above formula, we choose $r = 2$ and the first block contains everything except X_j and the second block only contains X_j. This implies that $S_1 = n - S_2 = n - X_j$. Thus,

$$(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_k)|X_j \stackrel{d}{=} (X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_k)|n - X_j \sim M_{k-1}\left(n - X_j; \frac{p_1}{1-p_j}, \cdots, \frac{p_k}{1-p_j}\right).$$

So the marginal of the above,

$$X_i|X_j \sim \mathsf{Bin}\left(n - X_j, \frac{p_i}{1-p_j}\right).$$

As a result, we see that X_i and X_j are negatively correlated. To confirm, we compute their covariance below.

$$\begin{aligned}
\mathsf{Cov}(X_i, X_j) &= \mathbb{E}[\underbrace{\mathsf{Cov}(X_i, X_j|X_j)}_{=0}] + \mathsf{Cov}(\mathbb{E}[X_i|X_j], \underbrace{\mathbb{E}[X_j|X_j]}_{=X_j}) \\
&= \mathsf{Cov}(\mathbb{E}[X_i|X_j], X_j) \\
&= \mathsf{Cov}\left((n - X_j)\frac{p_i}{1-p_j}, X_j\right) \\
&= -\frac{p_i}{1-p_j}\mathsf{Var}(X_j) \\
&= -np_ip_j.
\end{aligned}$$

7.2 Estimating multinomial parameters

Suppose we observe a random vector X from a multinomial distribution. We often know the total number of individuals n but the parameters $p_1, \cdots, p_k$ are may need to be estimated. We explain below how to use the MLE to estimate these parameters. Note that the multinomial distribution belongs to the (multivariate) exponential family.

For the parameter $\theta = (p_1, \ldots, p_k)$ of a multinomial distribution, the parameter space is $\Theta = \{(p_1, \cdots, p_k) : 0 \leq p_j, \sum_{j=1}^k p_j = 1\}$. We observe the random vector $X = (X_1, \cdots, X_k) \sim M_k(n; p_1, \cdots, p_k)$. In this case, the likelihood function is

$$L_n(p_1, \cdots, p_k|X) = \frac{n!}{X_1! \cdots X_k!} p_1^{X_1} \cdots p_k^{X_k}$$

and the log-likelihood function is

$$\ell_n(p_1, \cdots, p_k|X) = \sum_{j=1}^k X_j \log p_j + \log(\binom{n}{x_1, x_2, \ldots, x_k}),$$

where $C_n = \binom{n}{x_1, x_2, \ldots, x_k}$ is a constant is independent of p. Note that, in this case, naively computing the roots of the score function does not lead to the correct estimate for the p_js. This is due to the fact that we are ignoring the constraint of the parameter space, that is $\sum_{j=1}^k p_j = 1$. Truly we want to compute

$$\underset{(\theta \in \Theta)}{\mathsf{argmax}}\ \ell_n(\theta|X) = \underset{(p_1, \ldots, p_k); \sum_j p_j = 1}{\mathsf{argmax}} \ell_n((p_1, \ldots, p_k)|X).$$

One can show that the solution to this *constrained optimization problem* will be equivalent to the solution of the following (unconstrained) problem

$$\underset{(p_1, \ldots, p_k, \lambda)}{\mathsf{argmax}}\ F(\theta, \lambda) = \underset{(p_1, \ldots, p_k, \lambda)}{\mathsf{argmax}} \left[\sum_{j=1}^k X_j \log p_j + \lambda \left(1 - \sum_{j=1}^k p_j \right) \right].$$

The function $F(\theta, \lambda)$ above is known as the Lagrangian, or the Lagrange function, and the new parameter λ is known as the Lagrange multiplier. Please see a course on mathematical optimization for details, also discussed further in 513.

We can now treat the above problem in the classical way, that is, differentiate the Lagrangian with respect to $p_1, \cdots, p_k$, and compute the roots. Then

$$\frac{\partial F}{\partial p_j} = \frac{X_j}{p_j} - \lambda = 0 \Rightarrow X_j = \widehat{\lambda} \cdot \widehat{p}_{MLE,j}.$$

The only thing left is to compute $\hat{\lambda}$ using our constraints. Thus, $n = \sum_{j=1}^k X_j = \widehat{\lambda} \sum_{j=1}^k p_j = \widehat{\lambda}$ so $\widehat{p}_{MLE,j} = \frac{X_j}{n}$, which is just the proportion of observations that belong to category j.

7.3 Dirichlet distribution and Bayesian estimators

The Dirichlet distribution is a continuous distribution relevant to the Multinomial. Sampling from a Dirichlet distribution leads to a random vector with length k where each element of this vector is non-negative and sum over the elements is 1, meaning that the Dirichlet distribution generates a random probability vector.

The Dirichlet distribution is a multivariate distribution over the simplex $\sum_{i=1}^k x_i = 1$ and $x_i \geq 0$. Its probability density function is

$$p(x_1, \cdots, x_k; \alpha_1, \cdots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^k x_i^{\alpha_i - 1},$$

where $B(\alpha) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}$, with $\Gamma(\alpha)$ being the Gamma function and $\alpha = (\alpha_1, \cdots, \alpha_k)$ the parameters of this distribution, where $\alpha_i > 0$, for all i. The Dirichlet distribution can be viewed as a generalization of the Beta. For $Z = (Z_1, \cdots, Z_k) \sim \mathsf{Dirich}(\alpha_1, \cdots, \alpha_k)$,

$$\mathbb{E}(Z_i) = \frac{\alpha_i}{\sum_{j=1}^k \alpha_j}, \text{ and the mode of } Z_i \text{ is } \frac{\alpha_i - 1}{\sum_{j=1}^k \alpha_j - k}.$$

So each parameter α_i determines the relative importance of category (state) i. The Dirichlet distribution is very popular for use in social sciences and linguistics analysis.

The Dirichlet distribution is often used as a prior distribution for the Multinomial parameter $p_1, \cdots, p_k$ in Bayesian inference. The fact that it generates a probability vector makes it an excellent candidate for this job and in fact it is a conjugate prior for this problem. Let $p = (p_1, \cdots, p_k)$. Assume that

$$X|p = (X_1, \cdots, X_k)|p \sim M_k(n; p_1, \cdots, p_k)$$

and we place a prior

$$p \sim \mathsf{Dirich}(\alpha_1, \cdots, \alpha_k).$$

The two distributional assumptions imply that the posterior distribution of p will be

$$\begin{aligned}
\pi(p|X) &\propto \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \times \frac{1}{B(\alpha)} p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1} \\
&\propto p_1^{x_1 + \alpha_1 - 1} \cdots p_k^{x_k + \alpha_k - 1} \\
&\sim \mathsf{Dirich}(x_1 + \alpha_1, \cdots, x_k + \alpha_k).
\end{aligned}$$

If we use the posterior mean as our estimate, then

$$\widehat{p}_{\pi,i} = \frac{x_i + \alpha_i}{\sum_{j=1}^k x_j + \alpha_j},$$

which is also the MLE when we observe the counts $x' = (x'_1, \cdots, x'_k)$ such that $x'_j = x_j + \alpha_j$. However, note that above, α_j does not have to be an integer. So the prior parameter α_j can be viewed as a pseudo count of the category j before collecting the data.

References

Perlman, M. D. (2019). Probability and mathematical statistics i. https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf.

STAT 512: Statistical Inference Original notes credit: Yen-Chi Chen, Facheng Yu, Mathias Drton

Lecture 8: Linear models and multivariate normal distributions

Instructor: Emilija Perković Compiled on: 2023-12-06, 13:59:13

Additional reading: Chapter 4 of Casella and Berger (2021), Chapter 8 of Perlman (2019). Thank you to Facheng Yu for putting some of these proofs together!

8.1 Review of linear algebra

An $m \times n$ matrix $A = \{a_{ij}\}$ is an array of nm elements such that

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

In this case, we can write $A \in \mathbb{R}^{m \times n}$. The matrix applied to a vector represents a linear mapping (linear transformation) $A : \mathbb{R}^n \to \mathbb{R}^m$ ($x \mapsto Ax$), where $x \in \mathbb{R}^n$ is written as a column vector (i.e., an $n \times 1$ matrix) and

$$Ax = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \sum_j a_{1j} x_j \\ \sum_j a_{2j} x_j \\ \vdots \\ \sum_j a_{mj} x_j \end{pmatrix}$$

Clearly, the above operation implies the linear addition, i.e., for any $a, b \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$, $A(ax + by) = aAx + bAy$.

For two $m \times n$ matrices A, B, the addition $A + B$ is another $m \times n$ matrix such that $[A + B]_{ij} = a_{ij} + b_{ij}$. For an $m \times n$ matrix A and an $n \times p$ matrix B, the *matrix multiplication* AB is an $m \times p$ matrix such that

$$[AB]_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

A very important property is that $AB \neq BA$ in general even if $m = n = p$.

8.1.1 Useful characteristics of a matrix

Rank. The *rank* of a matrix A, denoted as $\mathsf{rank}(A)$, is the dimension of its column space. The column space is the vector space spanned by $a_{\cdot 1}, \cdots, a_{\cdot n}$, the column vectors of A, i.e.,

$$a_{\cdot j} = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}.$$

One can easily verify that $\mathsf{rank}(A) \leq \min\{m, n\}$. Also, $\mathsf{rank}(AB) \leq \min\{\mathsf{rank}(A), \mathsf{rank}(B)\}$.

Identity matrix. The $n \times n$ identity matrix $\mathbf{I}_n$ is a matrix that has 1's on its diagonal and 0 elsewhere. Namely, $\mathbb{I}_n = \mathsf{Diag}(1, 1, 1, \cdots, 1)$. One can easily see that for an $m \times n$ matrix A and $n \times m$ matrix B,. $A\mathbf{I}_n = A$ and $\mathbf{I}_n B = B$.

Inverse. The *inverse* of an $n \times n$ (square) matrix A, denoted as A^{-1}, is an $n \times n$ matrix with the property that $AA^{-1} = A^{-1}A = \mathbf{I}_n$. Note: the inverse may not exist. When the inverse of A exists, A is called *regular* otherwise it is called singular. The followings are equivalent of a $n \times n$ square matrix A:

- A is regular/non-singular (i.e., has an inverse matrix).
- A is full rank, i.e., $\mathsf{rank}(A) = n$.
- The determinant of A is not 0 (we will define a determinant later).

If both $n \times n$ matrices A, B are regular, then AB is also regular with inverse $(AB)^{-1} = B^{-1}A^{-1}$. For a diagonal matrix $D = \mathsf{Diag}(d_1, \cdots, d_n)$, its inverse is $D^{-1} = \mathsf{Diag}(d_1^{-1}, \cdots, d_n^{-1})$.

Transpose. For an $m \times n$ matrix A, its *transpose*, denoted as A^T, is an $n \times m$ matrix such that $[A^T]_{ij} = a_{ji}$. You can easily verify that $(A + B)^T = A^T + B^T$, $(AB)^T = B^T A^T$, and $(A^{-1})^T = (A^T)^{-1}$.

Trace. For an $n \times n$ matrix A, its *trace*, denoted as $\mathsf{Tr}(A)$, is $\mathsf{Tr}(A) = \sum_{i=1}^n a_{ii}$. One can easily verify that $\mathsf{Tr}(aA + bB) = a\mathsf{Tr}(A) + b\mathsf{Tr}(B)$ and $\mathsf{Tr}(A) = \mathsf{Tr}(A^T)$. Moreover, for an $m \times n$ matrix A and an $n \times m$ matrix B, $\mathsf{Tr}(AB) = \mathsf{Tr}(BA)$.

Triangular matrix. An $n \times n$ matrix A is upper triangular if $a_{ij} = 0$ for all $i < j$. An $n \times n$ matrix A is lower triangular if A^T is upper triangular. A matrix is called triangular if it is either upper or lower triangular.

Determinant. For s square $n \times n$ matrix A, its *determinant*, denoted as $|A|$ or $\mathsf{det}(A)$, or sometimes $\mathsf{Det}(A)$, is

$$\mathsf{det}(A) = \sum_{\sigma \in S_n} \mathsf{sgn}(\sigma) \prod_{i=1}^n a_{i\sigma(i)},$$

where S_n is the set of all possible permutations of $\{1, 2, 3, \cdots, n\}$ and σ is one of the permutations in S_n. Additionally, $\mathsf{sgn}(\sigma) = \pm 1$ is the *signature* of the permutation σ, it is $+1$ if the permutation σ can be obtained with an even number of transpositions, otherwise it is -1. Note that non-square matrices do not have determinants.

Another way to define the determinant is using the recursive Laplace expansion. Here we note that a determinant for a 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ is } \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc,$$

and for a 3×3 matrix

$$B = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \text{ is } \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh,$$

recall the Sarrus rule.

Laplace expansion. Then for a determinant of a general $n \times n$ matrix A, by Laplace expansion we have that

$$|A| = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} M_{i,j},$$

where $a_{i,j}$ is the element of A in the i-th row and j-th column, and $M_{i,j}$ is the *minor* of A, that is, the determinant of an $(n-1) \times (n-1)$ matrix that results from A by removing the i-th row and j-th column. The expression $(-1)^{i+j} M_{i,j}$ is also known as a *cofactor* of A.

Here are some useful properties of the determinant:

- $\mathsf{det}(cA) = c^n \mathsf{det}(A)$, for $A \in \mathbb{R}^{n \times n}$,
- for $A = [a_{\cdot 1}, \ldots, a_{\cdot n}]$, $\mathsf{det}(A) = \mathsf{det}([a_{\cdot 1}, \ldots, a_{\cdot j}, \ldots, a_{\cdot i} \ldots, a_{\cdot n}]) = -\mathsf{det}([a_{\cdot 1}, \ldots, a_{\cdot i}, \ldots, a_{\cdot j} \ldots, a_{\cdot n}])$. This property can be applied iteratively.
- $\mathsf{det}(AB) = \mathsf{det}(A) \cdot \mathsf{det}(B)$, when they are both square matrices
- $\mathsf{det}(A)^{-1} = \mathsf{det}(A^{-1})$,
- $\mathsf{det}(A^T) = \mathsf{det}(A)$, $\mathsf{det}(A) = \prod_{i=1}^{n} a_{ii}$ if A is triangular.
- $\mathsf{det}(AB) = \mathsf{det}(A) \cdot \mathsf{det}(B)$ if A and B are square matrices.
- $\mathsf{det}(A)^{-1} = \mathsf{det}(A^{-1})$.

Orthogonal matrix. An $n \times n$ matrix U is *orthogonal* if $U^T U = \mathbf{I}_n$. Namely, its column vectors form an orthonormal basis of $\mathbb{R}^n$. Note that one can easily see that this implies that $U^T = U^{-1}$ so $UU^T = \mathbf{I}_n$ as well.

Eigenvalues and eigenvectors. For an $n \times n$ matrix, its *eigenvalues* are the n roots $\lambda_1, \cdots, \lambda_n$ to the following polynomial equation:

$$\mathsf{det}(A - \lambda \mathbf{I}_n) = 0.$$

For each λ_j, there exists a vector u_j such that $(A - \lambda_j \mathbf{I}_n) u_j = 0$ or $A u_j = \lambda_j u_j$. Such a vector u_j is called the *eigenvector* corresponding to λ_j. Note that if λ_j is distinct from other eigenvalues, then u_j is unique. Also note that the eigenvalues and eigenvector may not be real numbers/vectors.

8.1.2 Symmetric matrices

A square matrix $A \in \mathbb{R}^{n \times n}$ is *symmetric* if $a_{ij} = a_{ji}$, i.e., $A = A^T$. In what follows, we will review some useful properties of a symmetric matrix.

For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, it has the following properties:

- Eigenvalues and eigenvectors are real numbers/vectors.
- For eigenvalues $\lambda_j \neq \lambda_k$, their corresponding eigenvectors u_j, u_k are orthogonal, i.e., $u_j^T u_k = 0$.
- **Spectral decomposition.** Let $\lambda_1, \cdots, \lambda_n$ be the eigenvalues of A and $u_1, \cdots, u_n$ be the corresponding eigenvectors. Let $\Lambda = \mathsf{Diag}(\lambda_1, \cdots, \lambda_n)$ and $U = [u_1, \cdots, u_n]$. Then

  $$A = U \Lambda U^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T.$$

  This is known as the spectral decomposition.

- **Trace.** The trace of A is $\mathsf{Tr}(A) = \sum_{i=1}^n \lambda_i$.
- **Determinant.** The determinant of A is $\mathsf{det}(A) = \prod_{i=1}^n \lambda_i$

Positive definite matrix. A particular important class of symmetric matrices is the *positive definite (PD) matrices*. A square matrix $A \in \mathbb{R}^{n\times n}$ is *positive semi-definite (PSD)* if

$$x^T Ax \geq 0$$

for all $x \in \mathbb{R}^n$. It is *positive definite* if

$$x^T Ax > 0$$

for all $x \in \mathbb{R}^n$ and $x^T x > 0$.

Here are some useful properties of PD and PSD matrices.

- The identity matrix is PD.
- A diagonal matrix D is PD if $D_{ii} > 0$ for all i and is PSD if $D_{ii} \geq 0$ for all i.
- If $S \in \mathbb{R}^{n\times n}$ is PSD and $A \in \mathbb{R}^{m\times n}$ be any matrix, then ASA^T is PSD.
- If $S \in \mathbb{R}^{n\times n}$ is PD and $A \in \mathbb{R}^{m\times n}$ be any matrix with $\mathsf{rank}(A) = m \leq n$, then ASA^T is PD.
- AA^T is PSD for any $m \times n$ matrix A.
- AA^T is PD for any $m \times n$ matrix A with $\mathsf{rank}(A) = m \leq n$.
- A is PD $\Rightarrow$ A is full rank $\Rightarrow$ A^{-1} exists $\Rightarrow$ $A^{-1} = A^{-1}AA^{-1}$ is PD.
- A symmetric matrix A is PSD (PD) if all its eigenvalues $\lambda_j \geq 0$ (> 0).
- If $A \in \mathbb{R}^{n\times n}$ is PD, then let its spectral decomposition be $A = U\Lambda U^T$. Then the square root of A, a matrix C such that $CC^T = A$, is $C = U\sqrt{\Lambda}U^T$, where $\sqrt{\Lambda} = \mathsf{Diag}(\sqrt{\Lambda_{11}}, \cdots, \sqrt{\Lambda_{nn}})$.

Partitioned PD matrix. Suppose that $A \in \mathbb{R}^{n\times n}$ is a PD matrix and we suppose that it can be decomposed into 4 submatrices

$$A = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

where $S_{ij} \in \mathbb{R}^{n_i \times n_j}$ with $i, j = 1, 2$ and $n = n_1 + n_2$. Then we have the following properties:

- S_{11} and S_{22} are both PD.
- Let $S_{11,2} = S_{11} - S_{12}S_{22}^{-1}S_{21}$. Then

$$\begin{pmatrix} \mathbf{I}_{n_1} & -S_{12}S_{22}^{-1} \\ 0 & \mathbf{I}_{n_2} \end{pmatrix} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{n_1} & 0 \\ -S_{22}^{-1}S_{21} & \mathbf{I}_{n_2} \end{pmatrix} = \begin{pmatrix} S_{11,2} & 0 \\ 0 & S_{22} \end{pmatrix}$$

  so $S_{11,2}$ is PD as well.
- Following from the above result, we have

$$\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{n_1} & S_{12}S_{22}^{-1} \\ 0 & \mathbf{I}_{n_2} \end{pmatrix} \begin{pmatrix} S_{11,2} & 0 \\ 0 & S_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{n_1} & 0 \\ S_{22}^{-1}S_{21} & \mathbf{I}_{n_2} \end{pmatrix}$$

$$\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{I}_{n_1} & 0 \\ -S_{22}^{-1}S_{21} & \mathbf{I}_{n_2} \end{pmatrix} \begin{pmatrix} S_{11,2}^{-1} & 0 \\ 0 & S_{22}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{n_1} & -S_{12}S_{22}^{-1} \\ 0 & \mathbf{I}_{n_2} \end{pmatrix}$$

- Further, the above implies that

$$A \text{ is PD } \Leftrightarrow S_{11,2}, S_{22} \text{ are PD } \Leftrightarrow S_{22,1}, S_{11} \text{ are PD .}$$

- For any vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^n$ such that $x_1 \in \mathbb{R}^{n_1}$ and $x_2 \in \mathbb{R}^{n_2}$,

$$xA^{-1}x = (x_1 - S_{12}S_{22}^{-1}x_2)S_{11,2}^{-1}(x_1 - S_{12}S_{22}^{-1}x_2) + x_2S_{22}^{-1}x_2.$$

Later we will see that the above results are very useful in analyzing the conditional normal distribution.

8.1.3 Projection matrices

An $n \times n$ matrix P is called a *projection* matrix if it is symmetric and idempotent ($P^2 = P \cdot P = P$).

P is a projection matrix if and only if there exists orthogonal matrix U such that

$$P = U \begin{pmatrix} \mathbf{I}_m & 0 \\ 0 & 0 \end{pmatrix} U^T.$$

In this case $\mathsf{rank}(P) = m$.

Suppose that we can partition $U = [U_1, U_2]$, where $U_1 \in \mathbb{R}^{n \times m}$ and $U_2 \in \mathbb{R}^{n \times (n-m)}$. Then the above result implies that $P = U_1U_1^T$ and $PU_1 = U_1$ and $PU_2 = 0$. This means that P project any vector in $\mathbb{R}^n$ into the column space of U_1 and is orthogonal to the column space of U_2. An interesting property is that $\mathsf{rank}(P) = \mathsf{Tr}(P) = m$.

Also, the matrix $\mathbf{I}_n - P$ is another projection matrix that projects any vector in $\mathbb{R}^n$ to the space orthogonal to the column space of U_1. To see this, $P(\mathbf{I}_n - P) = P - P^2 = 0$.

8.2 Transforming multiple continuous random variables

In lecture 2, we have learned techniques to deal with transforming a single continuous random variable, i.e., investigating the distribution of $U = f(X)$ when we know the distribution of X. In this section, we will study a more general problem where we are transforming two or more (continuous) random variables.

We start with a simple case where we have two random variables X, Y and we know their joint PDF. Consider two random variables $U = f(X, Y)$ and $V = g(X, Y)$, where u, v are two known functions.

We now study the joint PDF of (U, V). By definition,

$$\begin{aligned} p_{U,V}(u, v) &= \frac{\partial^2}{\partial u \partial v} P(U \le u, V \le v) \\ &= \frac{\partial^2}{\partial u \partial v} P(f(X, Y) \le u, g(X, Y) \le v) \\ &= \frac{\partial^2}{\partial u \partial v} P((X, Y) \in R(u, v)) \\ &= \frac{\partial^2}{\partial u \partial v} \int_{R(u,v)} p_{X,Y}(x, y) dx dy, \end{aligned}$$

where

$$R(u,v) = \{(x,y) : f(x,y) \le u, g(x,y) \le v\}.$$

In some scenarios, this region $R(u,v)$ has a nice form so that the probability $P((X,Y) \in R(u,v))$ has an analytical expression, so that we can take derivatives easily. However, this expression might still be hard to compute in general.

Example 1. Let $X, Y \sim \mathsf{Unif}[0,1]$. Consider $U = \max\{X,Y\}, V = \min\{X,Y\}$. Note that there is an implicit constraint on $f_{U,V}$ that $f_{U,V}(u,v) = 0$ if $v > u$. So we consider any pair $(u,v) : v \le u$. By a direct computation,

$$\begin{aligned} P(U \le u, V \le v) &= P(U \le u) - P(U \le u, V > v) \\ &= P(X \le u, Y \le u) - P(X \le u, Y \le u, X > v, Y > v) \\ &= P(X \le u)P(Y \le u) - P(v < X \le u)P(v < Y \le u) \\ &= u^2 - (u-v)^2 \end{aligned}$$

when $0 \le v \le u \le 1$. Thus,

$$p_{U,V}(u,v) = \frac{\partial^2}{\partial u \partial v} P(U \le u, V \le v) = 2I(0 \le v \le u \le 1).$$

Example 2. Consider $X, Y \sim \mathsf{Exp}(1)$ and let $U = X + Y$ and $V = \frac{X}{X+Y}$. Note that $(U,V) \in [0,\infty) \times [0,1]$. So we consider any $u \ge 0$ and $v \in [0,1]$. The joint CDF is

$$\begin{aligned} P(U \le u, V \le v) &= P(X + Y \le u, X \le v(X+Y)) \\ &= P\left(Y \le u - X, Y \ge \frac{1-v}{v} X\right) \\ &= \mathbb{E}\left[I\left(Y \le u - X, Y \ge \frac{1-v}{v} X\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[I\left(Y \le u - X, Y \ge \frac{1-v}{v} X\right) | X\right]\right] \\ &= \mathbb{E}\left[P\left(Y \le u - X, Y \ge \frac{1-v}{v} X | X\right)\right]. \end{aligned}$$

Note that $I(E)$ is the indicator function such that it returns 1 if the event E is true and 0 otherwise; one can easily see that $\mathbb{E}[I(E)] = P(E)$. Conditioning on X, the probability

$$\begin{aligned} P\left(Y \le u - X, Y \ge \frac{1-v}{v} X | X\right) &= P(\frac{1-v}{v} X \le Y \le u - X | X) \\ &= \int_{y=\frac{1-v}{v}X}^{u-X} e^{-y} dy \\ &= e^{-\frac{1-v}{v}X} - e^{X-u}. \end{aligned}$$

Thus, using the fact that $U \le u, V \le v$, $\frac{X}{X+Y} \le v$, so $X \le (X+Y)v \le uv$, we have

$$\begin{aligned} P(U \le u, V \le v) &= \mathbb{E}\left[P\left(Y \le u - X, Y \ge \frac{1-v}{v} X | X\right)\right] \\ &= \int_0^{uv} [e^{-\frac{1-v}{v}x} - e^{x-u}] e^{-x} dx \\ &= \int_0^{uv} [e^{-\frac{x}{v}} - e^{-u}] dx \\ &= v(1 - e^{-u} - ue^{-u}). \end{aligned}$$

By taking the derivative, we obtain

$$p_{U,V}(u,v) = ue^{-u}I(0 \le v \le 1) = \underbrace{ue^{-u}}_{p_U(u)} \cdot \underbrace{I(0 \le v \le 1)}_{p_V(v)}.$$

Thus, we conclude that $U \sim \mathsf{Gamma}(2,1)$ and $V \sim \mathsf{Uni}[0,1]$ and $U \perp V$.

8.2.1 Jacobian method

The Jacobian method is an elegant approach for substituting variables (change of varibales) in an integration. Consider $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ and assume that there is a 1-1 and onto mapping (a bijection) $T : \mathbb{R}^n \to \mathbb{R}^n$ for almost all x such that $y = T(x)$. We define the Jacobian matrix

$$J_T(x) = \left(\frac{\partial T(x)}{\partial x}\right) = \left(\frac{\partial y}{\partial x}\right) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

The *Jacobian* is the absolute value of the determinant of this matrix, i.e., $|\mathsf{det}(J_T(x))| = |\left(\frac{\partial y}{\partial x}\right)| = \left|\frac{\partial y}{\partial x}\right|$.

Theorem 8.1 *Assume that $y = T(x)$, where T is 1-1 and onto for almost all x and the Jacobian $\mathsf{det}(J_T(x)) \neq 0$ for all x. Let $A, B \subset \mathbb{R}^n$ be two subsets such that $B = \{T(x) : x \in A\}$. Let f be an integrable function. Then*

$$\int_A f(x)dx = \int_B f(T^{-1}(y))\,|\mathsf{det}(J_{T^{-1}}(y))|\,dy = \int_B f(T^{-1}(y))\left|\frac{\partial x}{\partial y}\right| dy.$$

Under the same condition, suppose X is a random variable with a PDF $p_X(x)$ and $Y = T(X)$. Then the PDF of Y is

$$\begin{aligned} p_Y(y) &= p_X(T^{-1}(y))\,|\mathsf{det}(J_{T^{-1}}(y)| \\ &= p_X(T^{-1}(y))\left|\frac{\partial x}{\partial y}\right|. \end{aligned}$$

The Jacobian has a nice chain rule that if $z = S(y)$ and $y = T(x)$ such that S, T are both 1-1 and onto. Then

$$\left|\frac{\partial z}{\partial x}\right| = \left|\frac{\partial z}{\partial y}\right|\left|\frac{\partial y}{\partial x}\right|.$$

Also, we have the inverse rule:

$$\left|\frac{\partial y}{\partial x}\right| = \left|\frac{\partial x}{\partial y}\right|^{-1}.$$

Example: Gamma distributions. Consider X, Y are IID Gamma (α, λ). Recall that the PDF of a Gamma (α, λ) is

$$p(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1}e^{-\lambda t}I(t \ge 0).$$

Now we consider $U = X + Y$ and $W = \frac{X}{X+Y}$. In this case, the mapping $T(x,y) = (u,w)$ such that $T = (T_1, T_2)$ with $T_1(x,y) = x + y$ and $T_2(x,y) = \frac{x}{x+y}$. Thus, the inverse mapping $T^{-1}(u,w) = (x,y)$ will

be $T_1^{-1}(u,w) = uw$ and $T_2^{-1}(u,w) = u - uw$. The Jacobian

$$\begin{aligned}\left|\frac{\partial(x,y)}{\partial(u,w)}\right| &= \left|\frac{\partial T^{-1}(u,w)}{\partial(u,w)}\right| \\ &= \left|\mathsf{det}\left(\begin{pmatrix} w & 1-w \\ u & -u \end{pmatrix}\right)\right| \\ &= u.\end{aligned}$$

We already know the joint PDF $p_{XY}(x,y)$ since they are independent Gamma. Thus,

$$\begin{aligned}p_{UW}(u,w) &= p_{XY}(T_1^{-1}(u,w), T_2^{-1}(u,w))uI(0 \le w \le 1, u \ge 0) \\ &= p_X(T_1^{-1}(u,w))p_Y(T_2^{-1}(u,w))uI(0 \le w \le 1, u \ge 0) \\ &= \frac{\lambda^{2\alpha}}{\Gamma^2(\alpha)}(uw)^{\alpha-1}e^{-\lambda uw}(u-uw)^{\alpha-1}e^{-\lambda(u-uw)}uI(0 \le w \le 1, u \ge 0) \\ &= \frac{\lambda^{2\alpha}}{\Gamma^2(\alpha)}u^{2\alpha-1}e^{-\lambda u}I(u \ge 0)w^{\alpha-1}(1-w)^{\alpha-1}I(0 \le w \le 1) \\ &= p_U(u)p_W(w)\end{aligned}$$

such that $U \sim \mathsf{Gamma}(2\alpha, \lambda)$ and $W \sim \mathsf{Beta}(\alpha, \alpha)$.

Example: Polar coordinates. A common reparametrization of two variables X, Y is through the polar coordinates R, Θ. Specifically, we choose $R = \sqrt{X^2 + Y^2}$ and $\Theta \in [0, 2\pi]$ such that

$$X = R\cos(\Theta), \quad Y = R\sin(\Theta).$$

In this case, $T(x,y) = (r,\theta)$ is 1-1 and onto for almost all points (x,y) except $(0,0)$ so we can still apply the Jacobian trick. You can easily work out that

$$\left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| = r$$

so if we know the PDF of X, Y as $p_{X,Y}(x,y)$, then

$$p_{R,\Theta}(r,\theta) = p_{X,Y}(r\cos(\theta), r\sin(\theta))r.$$

If the joint PDF of (X,Y) is over a circle or ellipse, (also called radial), i.e., $p_{X,Y}(x,y) = g(x^2+y^2)$, then $p_{R,\Theta}(r,\theta) = g(r^2)r$ so $R \perp \Theta$ and $\Theta \sim \mathsf{Uni}[0, 2\pi]$.

8.3 Random vectors and the covariance matrix

A random vector is a vector of random variables. Let $X \in \mathbb{R}^n$ be a random vector. We often express X as a column vector, i.e.,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

The expectation/expected value of X is the elementwise expectation:

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}.$$

Similar to random variables, the expectation is an linear operation of random vectors. Namely, for two random vectors $X, Y \in \mathbb{R}^n$ and two real numbers a, b,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

An important characteristic of a random vector is the *variance-covariance* matrix (often we just called it the covariance matrix):

$$\begin{aligned}\mathsf{Cov}(X) &= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T] \\ &= \begin{pmatrix} \mathsf{Var}(X_1) & \mathsf{Cov}(X_1, X_2) & \mathsf{Cov}(X_1, X_3) & \cdots & \mathsf{Cov}(X_1, X_n) \\ \mathsf{Cov}(X_2, X_1) & \mathsf{Var}(X_2) & \mathsf{Cov}(X_2, X_3) & \cdots & \mathsf{Cov}(X_2, X_n) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathsf{Cov}(X_n, X_1) & \mathsf{Cov}(X_n, X_2) & \mathsf{Cov}(X_n, X_3) & \cdots & \mathsf{Var}(X_n) \end{pmatrix}.\end{aligned}$$

Using the fact that $\mathsf{Var}(X_i) = \mathsf{Cov}(X_i, X_i)$, elements in the above matrix can be written as $\mathsf{Cov}(X)_{ij} = \mathsf{Cov}(X_i, X_j)$.

Here are some nice properties of the covariance matrices.

- $\mathsf{Cov}(X) = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T$
- For a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $b \in \mathbb{R}^m$,

$$\mathsf{Cov}(AX + b) = A\mathsf{Cov}(X)A^T.$$

- For a vector $a \in \mathbb{R}^n$, $\mathsf{Var}(a^T X) = a^T \mathsf{Cov}(X) a$.
- **The covariance matrix is positive semi-definite (PSD)**.
- The covariance matrix is PD if the only vector $a \in \mathbb{R}^n$ such that $\mathsf{Var}(a^T X) = 0$ is $a = 0$.

The covariance matrix immediately implies some useful properties of the sample mean. Suppose $X_1, \cdots, X_n$ are IID with mean u and variance σ^2. Then $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i = a^T X$, where $a_j = \frac{1}{n}$. As a result,

$$\mathsf{Var}(\overline{X}_n) = a^T \mathsf{Cov}(X) a = \frac{1}{n^2}\sum_{i=1}^n \mathsf{Var}(X_i) = \frac{\sigma^2}{n}.$$

Now, suppose that the random variables are not independent but instead, they have correlation $\mathsf{Cor}(X_i, X_j) = \rho$ when $i \neq j$. Then the variance of the sample mean will be

$$\begin{aligned}\mathsf{Var}(\overline{X}_n) &= a^T \mathsf{Cov}(X) a \\ &= \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \begin{pmatrix} \sigma^2 & \sigma^2\rho & \cdots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \cdots & \sigma^2\rho \\ \vdots & \vdots & \cdots & \vdots \\ \sigma^2\rho & \sigma^2\rho & \cdots & \sigma^2 \end{pmatrix} \begin{pmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix} \\ &= \frac{1}{n^2}(n\sigma^2 + n(n-1)\sigma^2\rho) \\ &= \frac{\sigma^2}{n}(1 + (n-1)\rho).\end{aligned}$$

8.4 The multivariate normal distribution

Recall that for a standard Normal random variable Z_1, its PDF is

$$p_{Z_1}(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

Thus, for IID standard normal random variables $Z_1, \cdots, Z_n$, we can represent them as a random vector Z and its joint PDF will be

$$p(z_1, \cdots, z_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\sum_{i=1}^n z_i^2} = \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2} z^T z}.$$

Now we consider a linear transformation that $A \in \mathbb{R}^{n\times n}$ is an invertible square matrix and $\mu \in \mathbb{R}^n$ is a vector and $X = AZ + \mu$. Since Z is a random vector, X will also be a random vector. Using the fact that $Z = A^{-1}(X - \mu)$ and the Jacobian method, you can show that the PDF of X is

$$\begin{aligned}
p_X(x) &= p(A^{-1}(x-\mu))|\mathsf{det}(A^{-1})| \\
&= \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}(x-\mu)^T[A^{-1}]^T A^{-1}(x-\mu)} \frac{1}{\sqrt{(\mathsf{det}(A))^2}} \\
&= \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}(x-\mu)^T[A^{-1}]^T A^{-1}(x-\mu)} \frac{1}{\sqrt{\mathsf{det}(AA^T)}} \\
&= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sqrt{\mathsf{det}(AA^T)}} e^{-\frac{1}{2}(x-\mu)^T[AA^T]^{-1}(x-\mu)} \\
&= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sqrt{\mathsf{det}(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)},
\end{aligned}$$

where $\Sigma = \mathsf{Cov}(X) = AA^T$ is the covariance matrix of X. Note that $\mathbb{E}[X] = \mu$ by construction. In this case, we will say that X is from a *multivariate normal distribution* with a mean (vector) μ and a covariance matrix Σ. For abbreviation, we often write $X \sim N_n(\mu, \Sigma)$.

The MGF of a multivariate normal. Using the same notation as above the MGF of Z can be derived from the univariate standard normal MGFs of $Z_1, \ldots, Z_n$ using their IID property. That is

$$\begin{aligned}
M_Z(t) &= \mathbb{E}[e^{t^T Z}] = \mathbb{E}[e^{\sum_i t_i Z_i}] = \mathbb{E}[\prod_i e^{t_i Z_i}] = \prod_i \mathbb{E}[e^{t_i Z_i}] \\
&= M_{Z_1}(t_1) M_{Z_2}(t_2) \cdots M_{Z_n}(t_n) = e^{\frac{1}{2}\sum_i t_i^2} = e^{\frac{1}{2} t^T t}.
\end{aligned}$$

Then the MGF of $X = AZ + \mu$ is

$$\begin{aligned}
M_X(t) &= \mathbb{E}[e^{t^T X}] = \mathbb{E}[e^{t^T(AZ+\mu))}] \\
&= e^{t^T\mu}\mathbb{E}[e^{t^T AZ}] = e^{t^T\mu}\mathbb{E}[e^{(A^T t)^T Z}] \\
&= e^{t^T\mu} M_Z(A^T t) \text{ (up to here true for all MGFs!)} \\
&= e^{t^T\mu + \frac{1}{2} t^T AA^T t} = e^{t^T\mu + \frac{1}{2} t^T \Sigma t}.
\end{aligned}$$

Linearity. The linear transformation of multivariate normal is still normal. Namely,

$$Y = CX + b \sim N_n(C\mu + b, C\Sigma C^T)$$

for non-singular matrix $C \in \mathbb{R}^{n\times n}$ and any vector $b \in \mathbb{R}^n$.

Proof: Using the fact $X = C^{-1}(Y-b)$ and the Jacobian method, the PDF of Y is

$$\begin{aligned}
p_Y(y) &= p_X(C^{-1}(Y-b))|\mathsf{det}(C^{-1})| \\
&= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sqrt{\mathsf{det}(\Sigma)}} e^{-\frac{1}{2}(C^{-1}(y-b)-\mu)^T\Sigma^{-1}(C^{-1}(y-b)-\mu)} \frac{1}{\sqrt{\mathsf{det}(C)^2}} \\
&= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sqrt{\mathsf{det}(\Sigma)\mathsf{det}(C)\mathsf{det}(C^T)}} e^{-\frac{1}{2}((y-b-C\mu)^T(C^{-1})^T\Sigma^{-1}C^{-1}(y-b-C\mu)} \\
&= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sqrt{\mathsf{det}(C\Sigma C^T)}} e^{-\frac{1}{2}((y-b-C\mu)^T(C\Sigma C^T)^{-1}(y-b-C\mu)}.
\end{aligned}$$

That is, $Y \sim N_n(C\mu + b, C\Sigma C^T)$. ■

For a vector $a \in \mathbb{R}^n$,

$$a^T X \sim N(a^T\mu, a^T\Sigma a).$$

Proof: Let $W = a^T X$, we will use the MGF of X for this proof. Since the MGF of X is $M_X(t) = e^{t^T\mu + \frac{1}{2}t^T\Sigma t}$, the MGF of W is

$$M_W(t) = E[e^{ta^TX}] = M_X(ta) = e^{ta^T\mu + \frac{1}{2}t^2a^T\Sigma a}.$$

That is, $W \sim N(a^T\mu, a^T\Sigma a)$. Note that in the above computation t is a scalar, $t \in \mathbb{R}$, so $t^T = t$. ■

Any marginal distribution of an MVN is Normal. Suppose $X \sim \mathcal{N}(\mu, \Sigma)$. Consider that $X_i = e_i^T \cdot X$, where $e_i \in \mathbb{R}^n$ is a basis vector that is made up of 0's except in position i, where it is equal to 1, that is, $e_i^T = (0, \ldots, 0, 1, 0, \ldots, 0)$. Then by the linearity property above, $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$.

Similarly if W is any sub-vector of X, that is, if $W = (X_{w_1}, \ldots, X_{w_k})^T$, for $w_1, \ldots, w_k \in \{1, \ldots, n\}$, $w_i \neq w_j$, for $i \neq j$. Then consider the re-ordered vector X titled $\widetilde{X}$ such that $\widetilde{X} = (\widetilde{X}_1, \widetilde{X}_2)^T$, where $\widetilde{X}_1 \equiv W$ and $\widetilde{X}_2$ is made up of all elements of X that are not in W.

We know that $\widetilde{X} \sim \mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma})$, where $\widetilde{\mu}, \widetilde{\Sigma}$ are appropriately reordered versions of μ and Σ. Then let $\widetilde{\mu}_1, \widetilde{\mu}_2$ be the mean vector that corresponds to each of the block of $\widetilde{X}_1$ and $\widetilde{X}_2$ and $\widetilde{\Sigma} = \begin{pmatrix} \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & \widetilde{\Sigma}_{22} \end{pmatrix}$.

Then $W = \begin{pmatrix} \mathbf{I}_k & 0 \\ 0 & 0 \end{pmatrix} \widetilde{X}$ and using the linearity property above it follows that

$$W \sim N_{n_1}(\widetilde{\mu}_1, \widetilde{\Sigma}_{11}).$$

so the marginals of the random vector are also multivariate normals.

Independence $\Leftrightarrow$ uncorrelation. If X is multivariate normal, then

$$X_i \perp X_j \Leftrightarrow \mathsf{Cov}(X_i, X_j) \equiv \Sigma_{ij} = 0.$$

Namely, pairwise independence is equivalent to uncorrelatedness.

Proof: ($\Rightarrow$) $X_i \perp X_j \Rightarrow \mathsf{Cov}(X_i, X_j) = \mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] = 0$.

($\Leftarrow$) From above marginal property, we know that $(X_i, X_j) \sim N_2((\mu_i, \mu_j), \begin{pmatrix} \Sigma_{ii} & 0 \\ 0 & \Sigma_{jj} \end{pmatrix})$ Then the MGF of

(X_i, X_j) is:

$$\begin{aligned} M_{X_i,X_j}(t_i,t_j) &= \exp(t_i\mu_i + t_j\mu_j + \frac{1}{2}t_i^2\Sigma_{ii} + \frac{1}{2}t_j^2\Sigma_{jj}) \\ &= \exp(t_i\mu_i + \frac{1}{2}t_i^2\Sigma_{ii})\exp(t_j\mu_j + \frac{1}{2}t_j^2\Sigma_{jj}) \\ &= M_{X_i}(t_i)M_{X_j}(t_j). \end{aligned}$$

so $X_i \perp X_j$. (We could have also chosen to look at the PDF of X_i, X_j directly.) ■

Any conditional distribution of an MVN is Normal. Suppose we partition X into two blocks

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

where $X_1 \in \mathbb{R}^{n_1}$ and $X_2 \in \mathbb{R}^{n_2}$. Let μ_1, μ_2 be the mean vector that corresponds to each of the block and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then the conditional distribution of $X_1|X_2$ is

$$X_1|X_2 \sim N_{n_1}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11,2}),$$

where $\Sigma_{11,2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. You can compare this to the partitioned of PD matrix in Section 8.1.2.

Proof: Consider the following linear transformation of X:

$$Y = \begin{pmatrix} I_{n_1} & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = CX,$$

where C is invertible (because it is upper-triangular). By the linearity,

$$Y = \begin{pmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{pmatrix} \sim N_n(C\mu, C\Sigma C^T) = N_n\left(\begin{pmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right).$$

It follows that $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \sim N_{n_1}(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ and that $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \perp X_2$. Thus,

$$X_1|X_2 \overset{d}{=} (X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 + \Sigma_{12}\Sigma_{22}^{-1}X_2)|X_2 \sim N_{n_1}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

■

Regression is Linear. Using the result above, we have that the regression function (conditional mean) is

$$\mathbb{E}[X_1|X_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2),$$

and the conditional variance

$$\mathsf{Var}(X_1|X_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

8.4.1 Chi-square distribution

Let $X = (X_1, \cdots, X_n)^T$ be a multivariate normal vector with mean 0 and identity covariance matrix. Then the random variable

$$W_n = \sum_{i=1}^{n} X_i^2 = X^T X = \|X\|^2$$

has a distribution called the χ^2 distribution with a degree of freedom n. In this case, we write $W_n \sim \chi^2_n$. The χ^2_n is the same as $\Gamma(\frac{n}{2}, \frac{1}{2})$ and $\mathbb{E}(W_n) = n$ and $\mathsf{Var}(W_n) = 2n$.

Normalizing a Gaussian vector. Suppose a random vector $Y \sim N(\mu, \Sigma)$, then

$$Z = \Sigma^{-\frac{1}{2}}(Y - \mu) \sim N(0, \mathbf{I}_n)$$

so

$$Z^T Z = (Y - \mu)^T \Sigma^{-1} (Y - \mu) \sim \chi^2_n.$$

Projection property. Here is an interesting property of a projection matrix. Lete $X \sim N(\mu, \mathbf{I}_n)$ be a multivariate normal vector in $\mathbb{R}^n$. Let $P \in \mathbb{R}^{n \times n}$ be a projection matrix with $\mathsf{rank}(P) = \mathsf{Tr}(P) = m < n$. Then

$$(X - \mu)^T P (X - \mu) \sim \chi^2_m.$$

You can prove the above result using the decomposition in Section 8.1.3.

Proof: The projection matrix P can be decomposed as

$$P = C^T \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} C,$$

where $C^T C = I_n$. Thus,

$$(X - \mu)^T P (X - \mu) = (X - \mu)^T C^T \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} C (X - \mu).$$

Using linearity,

$$Y = C(X - \mu) \sim N_n(0, C I_n C^T) = N_n(0, C C^T) = N_n(0, I_n).$$

Then it follows that

$$(X - \mu)^T P (X - \mu) = Y^T \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} Y = \sum_{i=1}^{m} Y_i^2 \sim \chi^2_m.$$

■

IID normals. Suppose $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ form an IID random sample. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ be the sample variance. Then we have the following results:

(i) $\overline{X}_n$ and S_n^2 are independent.

(ii) $\overline{X}_n \sim N(\mu, \sigma^2/n)$.

(iii) $(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}$.

Proof: Let $X = (X_1, \cdots, X_n)^T$ be a multivariate normal formed by the IID elements. Let $A \in \mathbb{R}^{\times}$ be an orthogonal matrix ($A^T A = I_n$) such that the last row of A is equal to $a_{n\cdot} = \frac{1}{\sqrt{n}}(1, 1, \cdots, 1)$. (This matrix exists and can be constructed using the Gram-Schmidt process).

Consider the vector $Z = (Z_1, \ldots, Z_n)^T$ defined as $Z = AX$. Note the following:

- $Z \sim \mathcal{N}_n(A\mu, \mathrm{Cov}(Z))$, where $\mathrm{Cov}(Z) = \mathrm{Cov}(AX) = A\,\mathrm{Cov}(X)A^T = A\sigma^2 I_n A^T = \sigma^2 A A^T = \sigma^2 I_n$,
- $Z_n = \sqrt{n}\bar{X}_n$,

- Additionally,
$$\sum_{i=1}^{n} Z_i^2 = Z^T Z = (AX)^T (AX) = X^T A^T AX = X^T X = \sum_{i=1}^{n} X_i^2.$$
- Also,
$$\sum_{i=1}^{n-1} Z_i^2 = \sum_{i=1}^{n} Z_i^2 - Z_n^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}_n = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Since $Z_n \perp Z_1, \ldots, Z_{n-1}$ by above, we have that $Z_n \perp \sum_i^{n-1} Z_n^2$.

Now, (i) follows from the fact that $\overline{X}_n = \frac{1}{\sqrt{n}} Z_n$, and $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} Z_i^2$.

Additionally, (ii) follows by linearity, since $Z_n \sim \mathcal{N}(\sqrt{n}\mu, \sigma^2)$ and $\overline{X}_n = \frac{1}{\sqrt{n}} Z_n$.

Lastly, for (iii) let $Q = I_n - a_n a_n^T$, and note that $S_n^2 = \frac{1}{n-1} \|QX\|_2^2$. We also note that Q is a projection matrix ($Q = Q^T = Q^2$) with $tr(Q) = n - 1$. Using the projection property and the fact that $\sqrt{n} a_n^T Q = \sqrt{n} Q a_n = 0$,

$$\frac{1}{\sigma^2} X^T Q X = \frac{1}{\sigma^2} (X - \mu\sqrt{n} a_n)^T Q (X - \mu\sqrt{n} a_n) \sim \chi_{n-1}^2.$$

Thus,

$$(n-1)\frac{S_n^2}{\sigma^2} = \frac{1}{\sigma^2} \|QX\|_2^2 = \frac{1}{\sigma^2} X^T Q^T Q X = \frac{1}{\sigma^2} X^T Q X \sim \chi_{n-1}^2.$$

■

References

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Perlman, M. D. (2019). Probability and mathematical statistics i. https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf.

STAT 512: Statistical Inference **Original notes credit: Yen-Chi Chen**

Lecture 9: Order Statistics: Continuous Univariate Distributions

Instructor: Emilija Perković Compiled on: 2023-11-29, 08:33:20

Additional reading: Chapter 9 of Perlman (2019).

Let $X_1, \cdots, X_n$ be IID continuous R.V.'s with a PDF $p_X(x)$ and a CDF $F_X(x)$. Since they are continuous R.V.s, we will generally assume that they all take distinct values. The order statistics $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are the ordered versions of these n random variables such that $X_{(j)}$ is the j-th smallest values among $\{X_1, \cdots, X_n\}$. Thus,

$$X_{(1)} = \min\{X_1, \cdots, X_n\}$$
$$X_{(n)} = \max\{X_1, \cdots, X_n\}.$$

In past lectures we have already considered the distributions of $X_{(1)}$ and $X_{(n)}$ as well as the joint distribution of $(X_{(1)}, X_{(n)})$, when, for instance, $X_1, \ldots, X_n \overset{IID}{\sim} \mathsf{Uniform}[0,1]$. We now revisit these computations in more generality.

Distribution of the minimum. We first compute the CDF $F_{X_{(1)}}(x)$:

$$\begin{aligned} F_{X_{(1)}}(x) &= P(X_{(1)} \le x) = 1 - P(X_{(1)} > x) \\ &= 1 - P(X_1 > x, X_2 > x, \ldots, X_n > x) \\ &\overset{IID}{=} 1 - [P(X_1 > x)]^n = 1 - [1 - F_X(x)]^n. \end{aligned}$$

We can also obtain the PDF $p_{X_{(1)}}(x)$ by differentiating the above. Hence,

$$p_{X_{(1)}}(x) = \frac{d}{dx}(1 - [1 - F_X(x)]^n) = n[1 - F_X(x)]^{n-1} p_X(x).$$

Distribution of the maximum. We first compute the CDF $F_{X_{(n)}}(x)$:

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \le x) \\ &= P(X_1 \le x, X_2 \le x, \ldots, X_n \le x) \\ &\overset{IID}{=} [P(X_1 \le x)]^n = [F_X(x)]^n. \end{aligned}$$

We can also obtain the PDF $p_{X_{(n)}}(x)$ by differentiating the above. Hence,

$$p_{X_{(n)}}(x) = \frac{d}{dx}([F_X(x)]^n) = n[F_X(x)]^{n-1} p_X(x).$$

Joint distribution of the minimum and maximum. We first compute the CDF $F_{X_{(1)},X_{(n)}}(x,y) = P(X_{(1)} \le x, X_{(n)} \le y)$. For this computation, note the following relationship that follows from the law of total probability.

$$P(X_{(n)} \le y) = P(X_{(1)} \le x, X_{(n)} \le y) + P(X_{(1)} > x, X_{(n)} \le y). \tag{9.1}$$

We know how to compute the left-hand-side of Equation (9.1), by above. So to compute the desired CDF, we need to reason about $P(X_{(1)} > x, X_{(n)} \leq y)$. In this case, we need to have $x < y$, otherwise, the above decomposition is trivial, since $P(X_{(1)} > x, X_{(n)} \leq y) = 0$. We now compute $P(X_{(1)} > x, X_{(n)} \leq y)$:

$$\begin{aligned} P(X_{(1)} > x, X_{(n)} \leq y) &= P(x < X_1 \leq y, x < X_2 \leq y, \ldots, x < X_n \leq y) \\ &\overset{IID}{=} [P(x < X_1 \leq y)]^n = [F_X(y) - F_X(x)]^n. \end{aligned}$$

Thus, the CDF $F_{X_{(1)},X_{(n)}}(x, y)$ equals

$$\begin{aligned} F_{X_{(1)},X_{(n)}}(x, y) &= P(X_{(1)} \leq x, X_{(n)} \leq y) \\ &\overset{(9.1)}{=} P(X_{(n)} \leq y) - P(X_{(1)} > x, X_{(n)} \leq y) \\ &= [F_X(y)]^n - [F_X(y) - F_X(x)]^n. \end{aligned}$$

And the PDF $p_{X_{(1)},X_{(n)}}(x, y)$ is

$$p_{X_{(1)},X_{(n)}}(x, y) = \frac{d^2}{dx\, dy}([F_X(y)]^n - [F_X(y) - F_X(x)]^n) = n(n-1)[F_X(y) - F_X(x)]^{n-2} p_X(x) p_X(y).$$

Note that above, we insist on $x < y$.

9.1 Joint Distribution of All Order Statistics

An interesting note: suppose that we have $n = 3$, and we observe x_1, x_2, x_3, such that $x_1 < x_2 < x_3$. Then $X_{(1)} = x_1, X_{(2)} = x_2$, and $X_{(3)} = x_3$. What kind of sampling of X_1, X_2, X_3 could have generated these observations? It could have been

$$\begin{aligned} &X_1 = x_1,\ X_2 = x_2,\ X_3 = x_3,\ \text{or} \\ &X_1 = x_2,\ X_2 = x_3,\ X_3 = x_1,\ \text{or} \\ &X_1 = x_2,\ X_2 = x_1,\ X_3 = x_3,\ \text{or} \ldots. \end{aligned}$$

There are $3! = 6$ permutations that could have generated the same order statistics. Hence, the mapping

$$(X_1, \cdots, X_n) \to (X_{(1)}, \cdots, X_{(n)})$$

is not 1-to-1 but $(n!)$-to-1.

We will use this information to compute the joint distribution of the order statistics. Instead of the joint CDF $P(X_{(1)} \leq x_1, \ldots, X_{(n)} \leq x_n)$ let's now consider a slightly different object that is easier to compute:

$$P(y_1 < X_{(1)} \leq x_1, \ldots, y_n < X_{(n)} \leq x_n),$$

where we insist on $y_1 < x_1 \leq y_2 < x_2 \leq \cdots \leq y_n < x_n$. The order statistics will satisfy the above for $n!$ different permutations of $X_1, \ldots, X_n$. Since these are all disjoint events, because of $y_1 < x_1 \leq y_2 < x_2 \leq \cdots \leq y_n < x_n$ and since $X_1, \ldots, X_n$ are IID, we have the following:

$$\begin{aligned} P(y_1 < X_{(1)} \leq x_1, \ldots, y_n < X_{(n)} \leq x_n) &= n! P(y_1 < X_1 \leq x_1, \ldots, y_n < X_n \leq x_n) \\ &= n! \prod_{i=1}^{n} P(y_i < X_i \leq x_i) = n! \prod_{i=1}^{n} [F_X(x_i) - F_X(y_i)]. \end{aligned}$$

Now to derive the PDF $p_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots x_n)$ of the order statistics, note that above we have obtained:

$$\int_{y_n}^{x_n}\int_{y_{n-1}}^{x_{n-1}}\cdots\int_{y_1}^{x_1} p_{X_{(1)},\ldots,X_{(n)}}(t_1,\ldots t_n)dt_1\ldots dt_n = n!\prod_{i=1}^{n}[F_X(x_i)-F_X(y_i)].$$

Differentiating both sides of the above equation with respect to $x_1,\ldots,x_n$ we obtain:

$$\frac{d}{dx_1}\cdots\frac{d}{dx_n}\left[\int_{y_n}^{x_n}\int_{y_{n-1}}^{x_{n-1}}\cdots\int_{y_1}^{x_1} p_{X_{(1)},\ldots,X_{(n)}}(t_1,\ldots t_n)dt_1\ldots dt_n\right] = \frac{d}{dx_1}\cdots\frac{d}{dx_n}\left[n!\prod_{i=1}^{n}[F_X(x_i)-F_X(y_i)]\right].$$

$$p_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots x_n) = n!p_X(x_1)p_X(x_2)\cdots p_X(x_n),$$

which gives us the joint PDF for our order statistics.

9.2 Marginal Distributions of Order Statistics

Distribution of $X_{(j)}$. We will use the fact that we have the above expression for the joint distribution $p_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots x_n)$ and integrate out all unwanted x_is.

Let's suppose that $j\neq 1$, since we know how to compute the marginal distribution of the minimum. Hence, consider integrating out x_1. What are the bounds of the integral? Note that we need to have $x_1<x_2<\cdots<x_n$ in the PDF $p_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots x_n)$. So the limits of integration are $-\infty$ and x_2. Then

$$\begin{aligned}
p_{X_{(2)},\ldots,X_{(n)}}(x_2,\ldots x_n) &= \int_{-\infty}^{x_2} p_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots x_n)dx_1\\
&= \int_{-\infty}^{x_2} n!p_X(x_1)p_X(x_2)\cdots p_X(x_n)dx_1\\
&= n!F(x_2)p_X(x_2)\cdots p_X(x_n),
\end{aligned}$$

for $x_2<\cdots<x_n$. Let's suppose that $j\neq 2$, so that we also need to integrate out x_2 above.

$$\begin{aligned}
p_{X_{(3)},\ldots,X_{(n)}}(x_3,\ldots x_n) &= \int_{-\infty}^{x_3} p_{X_{(2)},\ldots,X_{(n)}}(x_2,\ldots x_n)dx_2\\
&= \int_{-\infty}^{x_3} n!F_X(x_2)p_X(x_2)p(x_3)\cdots p_X(x_n)dx_2\\
&= n!\; p(x_3)\cdots p_X(x_n)\int_{-\infty}^{x_3} F_X(x_2)p_X(x_2)dx_2\\
&= n!\;\frac{1}{2}p_X(x_3)\cdots p_X(x_n)\big[F(x_2)\big]^2\Big|_{x_2=-\infty}^{x_2=x_3}\\
&= \frac{n!}{2}[F(x_3)]^2p_X(x_3)\cdots p_X(x_n),
\end{aligned}$$

for $x_3<\cdots<x_n$. Above, we used that if $u=F_X(x_2)$, than $du=p_X(x_2)dx_2$, and also that $F_X(-\infty)=0$.

We can keep integrating like this until we reach x_j. At that point we will have

$$p_{X_{(j)},\ldots,X_{(n)}}(x_j,\ldots x_n) = \frac{n!}{(j-1)!}[F(x_j)]^{j-1}p_X(x_j)\cdots p_X(x_n),$$

with $x_j<x_{j+1}<\cdots<x_n$.

Now, we can start integrating "from the other side". Starting with x_n (again assuming that $j \neq n$). The limits of integration will now be x_{n-1} to ∞:

$$\begin{aligned} p_{X_{(j)},\ldots,X_{(n-1)}}(x_j, \ldots x_{n-1}) &= \int_{x_{n-1}}^{+\infty} p_{X_{(j)},\ldots,X_{(n)}}(x_j, \ldots x_n) dx_n \\ &= \int_{x_{n-1}}^{+\infty} \frac{n!}{(j-1)!} [F(x_j)]^{j-1} p_X(x_j) \cdots p_X(x_n) dx_n \\ &= \frac{n!}{(j-1)!} [F(x_j)]^{j-1} p_X(x_j) \cdots p_X(x_{n-1})[1 - F_X(x_{n-1})], \end{aligned}$$

for $x_j < x_{j+1} < \cdots < x_{n-1}$.

We can keep up this process, until we reach:

$$p_{X_{(j)}}(x_j) = \frac{n!}{(j-1)!(n-j)!} [F(x_j)]^{j-1} p_X(x_j)[1 - F_X(x_j)]^{n-j},$$

for $x_j \in \mathbb{R}$.

The heuristic reasoning for the above expression is that there are $n!$ ways to arrange X_i's, and we need $(j-1)$ of X_i's to fall below x_j (each with probability $F(x_j)$) and $(n-j)$ of X_i's to fall above x_j (each with probability $1 - F(x_j)$). There are $(j-1)!$ ways of arranging the X_i's which are below x_j and $(n-j)!$ ways of arranging the X_i's above x_j. Hence, those need to be divided out of the total $n!$ arrangements.

Distribution of $X_{(i)}$ and $X_{(j)}$ for $i < j$. Using the same principles as in the section above, we can come up with all kinds of joint PDFs or conditional PDFs.

In particular, the joint PDF for $X_{(i)}$ and $X_{(j)}$ when $i < j$ can be computed as:

$$p_{X_{(i)},X_{(j)}}(x_i, x_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} [F(x_i)]^{i-1} p_X(x_i)[F(x_j) - F(x_i)]^{j-i-1} p_X(x_j)[1 - F_X(x_j)]^{n-j},$$

for $x_i < x_j$.

Aside: Order statistics and their properties and convergences are the topic of study in a branch of statistics known as Extreme Value Theory. This branch of statistics has many equivalent results to those we covered, such as a version of the CLT with different limiting distributions! Extreme value theory used to be primarily used to model insurance computations, but more recently has seen a lot of applications in climate science (see e.g. the work on Extreme Event Attribution.)

9.3 Case study: uniform distribution

Consider the case where $X_1, \cdots, X_n$ are IID from $\mathsf{Uni}[0,1]$. Then $p_X(x) = 1$ and $F_X(x) = x$ when $x \in [0,1]$. Thus,

$$p_{Y_j}(y) = \frac{n!}{(j-1)!(n-j)!} y^{j-1}(1-y)^{n-j},$$

which is the PDF of $\mathsf{Beta}(j, n-j+1)$.

Here is an interest note about the variance. The variance of Y_j is

$$\mathsf{Var}(Y_j) = \frac{j(n-j+1)}{(n+1)^2(n+2)},$$

which is maximized when $j = \frac{n+1}{2}$ assuming n is an odd number. The value $j = \frac{n+1}{2}$ corresponds to the 'median' of $\{X_1, \cdots, X_n\}$. Thus, the median has the highest variability. In this case,

$$\mathsf{Var}(Y_{\frac{n+1}{2}}) = \frac{1}{4(n+2)} = O(n^{-1}).$$

On the other hand, the maximal or minimal value has the lowest variance:

$$\mathsf{Var}(Y_1) = \mathsf{Var}(Y_n) = \frac{n}{(n+1)^2(n+2)} = O(n^{-2}).$$

Now we consider another way to look at the order statistics. Let $W_1, \cdots, W_n, W_{n+1}$ be the 'spacing' between consecutive order statistics:

$$\begin{aligned} W_1 &= Y_1 - 0 \\ W_2 &= Y_2 - Y_1 \\ W_3 &= Y_3 - Y_2 \\ &\vdots \\ W_n &= Y_n - Y_{n-1} \\ W_{n+1} &= 1 - Y_n. \end{aligned}$$

It is easy to see that $W_i \in [0, 1]$ and $W_1 + W_2 + \cdots + W_{n+1} = 1$. Also, we can reparametrize Y_j via W_i's:

$$Y_j = W_1 + W_2 + \cdots + W_j.$$

Since X_i's are uniform over $[0, 1]$, the joint PDF of $Y_1, \cdots, Y_n$ is

$$p(y_1, \cdots, y_n) = n!$$

whenever $0 < y_1 < \cdots < y_n < 1$. By the Jacobian method with the fact that $\mathsf{det}(\frac{dY}{dW}) = 1$ (think about why), we conclude that

$$p(w_1, \cdots, w_n) = n!$$

whenever $w_i \in [0, 1]$ and $w_1 + \cdots + w_n < 1$. One can easily see that $p(w_1, \cdots, w_n)$ is invariant under the permutation of $W_1, \cdots, W_n$ (i.e., they are *exchangeable*), so the marginal distribution of W_i is the same as the marginal distribution of W_j for all $i, j = 1, \cdots, n$. Because $W_1 = Y_1$ follows from $\mathsf{Beta}(1, n)$, we conclude that W_j is a $\mathsf{Beta}(1, n)$ random variable.

Note that W_i and W_j are dependent ($i \neq j$)! Due to the exchangeability property, the joint distribution (W_i, W_j) is the same as the joint distribution of W_1, W_2, so

$$\begin{aligned} \mathsf{Cov}(W_i, W_j) &= \mathsf{Cov}(W_1, W_2) \\ &= \frac{1}{2}\left(\mathsf{Var}(W_1 + W_2) - \mathsf{Var}(W_1) - \mathsf{Var}(W_2)\right) \\ &= \frac{1}{2}\left(\mathsf{Var}(Y_2) - 2\mathsf{Var}(Y_1)\right) \\ &= \frac{1}{2}\left(\frac{2(n-1)}{(n+1)^2(n+2)} - 2\frac{n}{(n+1)^2(n+2)}\right) \\ &= \frac{-1}{(n+1)^2(n+2)} < 0. \end{aligned}$$

References

Perlman, M. D. (2019). Probability and mathematical statistics i. https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf.

STAT 512: Statistical Inference **Original notes credit: Yen-Chi Chen, Patrick Breheny**

Lecture 10: Statistical functionals and the bootstrap

Instructor: Emilija Perković Compiled on: 2023-12-13, 14:49:47

Additional reading: pages 240 - 245 from Casella and Berger (2021)

10.1 The Delta Method

In this section, we will talk about a very useful technique in handling convergence for certain estimators – the *delta method*.

Example: inverse of mean. Assume we have $X_1, \ldots, X_n, \ldots \overset{iid}{\sim} F$, such that $\text{Var}(X_1) = \sigma^2$ and $\mathbb{E}[X_1] = \mu$, and that we are interested in estimating $1/\mu$. Namely, our parameter of interest θ is

$$\theta = \frac{1}{\mu} = \frac{1}{\mathbb{E}[X_i]} = \frac{1}{\int x dF(x)}.$$

The plug-in estimator for this quantity based on the EDF is

$$\widehat{\theta}_n = \frac{1}{\int x d\widehat{F}_n(x)} = \frac{1}{\bar{X}_n}.$$

What do we know about the asymptotic properties of this estimator?

Theorem 10.1 (Taylor) *If $f^{(r)}(a) = \frac{\partial^r}{\partial x^r} f(x)|_{x=a}$ exists, then for Taylor's r-order polynomial of f around a,*

$$T_r(x) = \sum_{i=0}^{r} \frac{f^{(i)}(a)}{i!}(x-a)^i$$

it holds that

$$\lim_{x \to a} \frac{f(x) - T_r(x)}{(x-a)^r} = 0.$$

Theorem 10.2 (Delta Method) *Let $Y_1, \cdots, Y_n \cdots$ be a sequence of random variables such that*

$$\sqrt{n}(Y_n - y_0) \xrightarrow{D} N(0, \sigma_Y^2), \tag{10.1}$$

for some constants y_0 and σ_Y^2. For a given differentiable function f such that $f'(y_0) \neq 0$, it then holds that

$$\sqrt{n}(f(Y_n) - f(y_0)) \xrightarrow{D} N(0, (f'(y_0))^2 \sigma_Y^2). \tag{10.2}$$

Proof: By first order Taylor expansion of $f(Y_n)$ around $Y_n = y_0$, we have that

$$f(Y_n) = f(y_0) + f'(y_0)(Y_n - y_0) + o(Y_n - y_0), \tag{10.3}$$

where by Taylor's theorem and the convergence of $Y_n \to y_0$, $o(Y_n - y_0) = o_p(Y_n - y_0) \overset{n \to \infty}{\to} 0$. Since $\sqrt{n}(Y_n - y_0)$ converges in distribution, then $\sqrt{n}(Y_n - y_0) = O_p(1)$ (this means bounded in probability, see Chapter 2.2

of Van der Vaart (2000) for details on o_p, O_p notation). Then $o(\sqrt{n}(Y_n - y_0)) = o_p(O_p(1)) = o_p(1)$, that is, $o(\sqrt{n}(Y_n - y_0)) \xrightarrow{P} 0$. Then by Equation (10.3), we have that

$$\sqrt{n}(f(Y_n) - f(y_0)) = f'(y_0)\sqrt{n}(Y_n - y_0) + o(\sqrt{n}(Y_n - y_0)).$$

Now applying the continuous mapping theorem and Slutsky's theorem (Theorem 3.6 in Lecture notes 3) to above we have that

$$\sqrt{n}(f(Y_n) - f(y_0)) \xrightarrow{D} N(0, (f'(y_0))^2\sigma_Y^2).$$

■

Now, we recall our problem and note that $\theta = f(\mu)$, where $f(x) = 1/x$ and by the CLT $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$. Then noting that $\widehat{\theta}_n = f(\bar{X}_n)$, we can apply the Delta method to obtain:

$$\sqrt{n}(\widehat{\theta}_n - \theta) = \sqrt{n}\left(\frac{1}{\bar{X}_n} - \frac{1}{\mathbb{E}(X_i)}\right) \approx -\frac{1}{\mathbb{E}^2(X_i)}\sqrt{n}\left(\bar{X}_n - \mathbb{E}(X_i)\right) \xrightarrow{D} N\left(0, \underbrace{\frac{1}{\mathbb{E}^4(X_i)}\mathsf{Var}(X_i)}_{=\mathbb{V}_{\mathsf{inv}}(F)}\right).$$

Using the fact that $\mathbb{E}(X_i) = \int x dF(x)$ and $\mathsf{Var}(X_i) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$, we obtain

$$\sqrt{n}(\widehat{\theta}_n - \theta) \approx N(0, \mathbb{V}_{\mathsf{inv}}(F)),$$

where

$$\mathbb{V}_{\mathsf{inv}}(F) = \frac{\int x^2 dF(x) - \left(\int x dF(x)\right)^2}{\left(\int x dF(x)\right)^4}.$$

Theorem 10.3 (Second Order Delta Method) *Let $Y_1, \cdots, Y_n \cdots$ be a sequence of random variables such that*

$$\sqrt{n}(Y_n - y_0) \xrightarrow{D} N(0, \sigma_Y^2), \tag{10.4}$$

for some constants y_0 and σ_Y^2. For a given twice differentiable function f such that $f'(y_0) = 0$, but $f''(y_0) \neq 0$, it then holds that

$$n(f(Y_n) - f(y_0)) \xrightarrow{D} \sigma_Y^2 \frac{f''(y_0)}{2}\chi_1^2. \tag{10.5}$$

Proof: By second order Taylor expansion of $f(Y_n)$ around $Y_n = y_0$, we have that

$$f(Y_n) = f(y_0) + f'(y_0)(Y_n - y_0) + \frac{f''(y_0)}{2}(Y_n - y_0)^2 + o((Y_n - y_0)^2), \tag{10.6}$$

where by Taylor's theorem and the convergence of Y_n, we have that $o((Y_n - y_0)^2) = o_P((Y_n - y_0)^2)$, that is $o((Y_n - y_0)^2) \xrightarrow{P} 0$. Then plugging in $f'(y_0) = 0$ into Equation (10.6), and re-arranging the terms we have that

$$n(f(Y_n) - f(y_0)) = \frac{f''(y_0)}{2}n(Y_n - y_0)^2 + o(n(Y_n - y_0)^2).$$

Now note that by definition of a Chi-squared random variable, we have that

$$\frac{n}{\sigma_Y^2}(Y_n - y_0)^2 \xrightarrow{D} \chi_1^2.$$

Now, since $n(Y_n - y_0)^2$ converges in distribution we have that $n(Y_n - y_0)^2 = O_p(1)$, so that $o(n(Y_n - y_0)^2) = o_p(O_p(1)) = o_P(1)$, which implies $o(n(Y_n - y_0)^2) \xrightarrow{P} 0$.

Hence, by applying the continuous mapping theorem and Slutsky's theorem we have the desired result

$$n(f(Y_n) - f(y_0)) \xrightarrow{D} \frac{f''(y_0)}{2}\sigma_Y^2\chi_1^2.$$

■

Theorem 10.4 (Multivariate Delta Method) *Let $\vec{X}_1, \ldots, \vec{X}_n, \ldots$ be a sequence of random vectors, where $\vec{X}_i = (X_{i1}, \ldots, X_{ip})^T, i \geq 1$ with $\mathbb{E}[X_{ij}] = \mu_j$ and $\mathrm{Cov}(X_{ij}, X_{ik}) = \sigma_{jk}$. Let $g : \mathbb{R}^p \to \mathbb{R}$ be a given function with continuous first partial derivatives and a specific value of $\vec{\mu} = (\mu_1, \ldots, \mu_p)^T$ for which*

$$\tau^2 = \sum_i \sum_j \sigma_{ij} \frac{\partial g(\vec{\mu})}{\partial \mu_i} \cdot \frac{\partial g(\vec{\mu})}{\partial \mu_j} > 0.$$

If

$$\sqrt{n}(\vec{X}_n - \vec{\mu}) \xrightarrow{D} N(\vec{0}, \Sigma), \tag{10.7}$$

where $[\Sigma]_{ij} = \sigma_{ij}$, then it also holds that

$$\sqrt{n}(g(\vec{X}_n) - g(\vec{\mu})) \xrightarrow{D} N(0, \tau^2). \tag{10.8}$$

The delta method is a very useful technique for analyzing some estimators. In the rest of this lecture, we consider various other plug-in estimators and their asymptotic properties.

10.2 Empirical Distribution Function

Let us take a look a the empirical distribution again.

Definition 10.5 *Let $X_1, \ldots, X_n$ be a (random sample) collection of iid random variables. The empirical (sample) distribution P_n is the discrete probability distribution that assigns probability $\frac{1}{n}$ to each observation X_i. Equivalently, P_n assigns probability $\frac{1}{n}$ to each order statistics $X_{(i)}$. The empirical CDF (EDF) F_n is the CDF associated with P_n, that is,*

$$\begin{aligned}\widehat{F}_n(x) &= \frac{\textit{number of } X_i \leq x}{\textit{total number of observations}} = \\ &= \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \\ &= \frac{1}{n}\sum_{i=1}^{n} I(X_{(i)} \leq x)\end{aligned}$$

Because EDF is the average of $I(X_i \leq x)$, let $Y_i = I(X_i \leq x)$.

$$Y_i = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{if } X_i > x \end{cases}.$$

So Y_i is a Bernoulli random variable What is the parameter p for Y_i?

$$p = P(Y_i = 1) = P(X_i \leq x) = F(x).$$

Therefore, for a given x,

$$Y_i \sim \mathsf{Ber}(F(x)).$$

This implies

$$\begin{aligned}\mathbb{E}(I(X_i \le x)) &= \mathbb{E}(Y_i) = F(x)\\ \mathsf{Var}(I(X_i \le x)) &= \mathsf{Var}(Y_i) = F(x)(1-F(x))\end{aligned}$$

for a given x. Recall that $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x) = \frac{1}{n}\sum_{i=1}^n Y_i$. Hence, $n\widehat{F}_n(x) = \sum_i Y_i \sim \mathsf{Bin}(n, p \equiv F(x))$. Also,

$$\begin{aligned}\mathbb{E}\left(\widehat{F}_n(x)\right) &= \mathbb{E}(I(X_1 \le x)) = F(x)\\ \mathsf{Var}\left(\widehat{F}_n(x)\right) &= \frac{\sum_{i=1}^n \mathsf{Var}(Y_i)}{n^2} = \frac{F(x)(1-F(x))}{n}.\end{aligned}$$

Hence, at each x, $\widehat{F}_n(x)$ is an *unbiased estimator* of $F(x)$:

$$\textbf{bias}\left(\widehat{F}_n(x)\right) = \mathbb{E}\left(\widehat{F}_n(x)\right) - F(x) = 0.$$

Furthermore, the *variance converges to* 0 when $n \to \infty$, which implies that for a given x,

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

i.e., $\widehat{F}_n(x)$ is a *consistent* estimator of $F(x)$.

Furthermore, by the CLT for a given x,

$$\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right) \xrightarrow{D} N(0, F(x)(1-F(x))). \tag{10.9}$$

Namely, $\widehat{F}_n(x)$ is asymptotically normally distributed around $F(x)$ with variance $F(x)(1-F(x))$.

In fact, an even stronger relationship holds:

Theorem 10.6 (Glivenko-Cantelli) *If F is a continuous CDF then (10.9) holds uniformly in x, that is, the following holds:*

$$\sup_{-\infty<x<\infty} |\widehat{F}_n(x) - F(x)| \to 0, \ \textit{as } n \to \infty.$$

The Glivenko-Cantelli Theorem implies that $\widehat{F}_n \xrightarrow{a.s.} F$.

10.3 Statistical Functionals and Nonparametric Estimation

What is a functional? A functional is just a function of a function. Namely, it is a 'function' which takes as input another function and outputs a real number. Formally speaking, a functional is a mapping $T : \mathcal{F} \mapsto \mathbb{R}$, where $\mathcal{F}$ is a collection of functions. A statistical functional is a functional T that takes as input a CDF. In this lecture we will consider estimating statistical functional, that is our parameter of interest θ is

$$\theta = T(F),$$

for some statistical functional $T(F)$, and CDF F. Examples of several statistical functionals are below.

Mean of a distribution. The mean of a distribution is a statistical functional

$$\mu = T_{\mathsf{mean}}(F) = \int x dF(x).$$

When the distribution is continuous, $dF(x) = p(x)dx$ so the mean functional reduces to:

$$\mu = T_{\mathsf{mean}}(F) = \int x dF(x) = \int x p(x) dx = \mathbb{E}[X].$$

When the distribution is discrete, we define

$$\int x dF(x) = \sum_x x P(x) \Longrightarrow \mu = T_{\mathsf{mean}}(F) = \sum_x x P(x) = \mathbb{E}[X],$$

where $P(x)$ is the PMF of the distribution F. In either case,

$$\mathbb{E}[X] = \int x dF(x) = T_{\mathsf{mean}}(F).$$

Aside: A pause here to reflect on the fact that we are now labeling the expectation as a function of F rather than X. When we write "X" in $\mathbb{E}[X]$, and talk about the mean "of X" this is actually shorthand for the "mean of the distribution of X". This is not a function of a random X (since the expectation is just a constant), though it is a functional of X, i.e. a function of the distribution of X.

Variance of a distribution. The variance of a distribution is also a statistical functional. Let X be a random variable with CDF F. Then

$$\sigma^2 = T_{\mathsf{var}}(F) = \mathsf{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \int x^2 dF(x) - \left(\int x dF(x) \right)^2.$$

Median of a distribution. Using the concept of a statistical functional, any quantile can be easily defined. The median of a distribution F is a point θ_{med} such that $F(\theta_{\mathsf{med}}) = 0.5$. Thus,

$$T_{\mathsf{med}}(F) = F^{-1}(0.5).$$

Note that when F is a CDF of a discrete random variable, F^{-1} may have multiple values. In this case, we define

$$F^{-1}(q) = \inf\{x : F(x) \geq q\}.$$

Any quantile of a distribution can be represented in a similar way. For instance, the q-quantile ($0 < q < 1$) will be

$$T_{\mathsf{q}}(F) = F^{-1}(q).$$

As a result, the interquartile range (IQR) is

$$T_{\mathsf{IQR}}(F) = F^{-1}(0.75) - F^{-1}(0.25).$$

Linear (Statistical) Functionals. More generally, recall that for any function ω,

$$\mathbb{E}[\omega(X)] = \int \omega(x) dF(x).$$

We use this property to introduce a class of statistical functionals called linear functionals. A statistical functional T_ω referred to as a *linear functional* if:

$$T_\omega(F) = \int \omega(x) dF(x).$$

10.4 Plug-in Estimators Broadly

Statistical functionals combined with the EDF provide simple plug-in estimators to various population quantities. These types of estimators are referred to as non-parametric.

Thus, if we want to estimate a population quantity $\theta = T_{\mathsf{target}}(F)$, we can use $T_{\mathsf{target}}(\widehat{F}_n) = \widehat{\theta}_n$ as our estimator. Many estimators follow this form. For instance, the estimator for the $\mu = T_{\mathsf{mean}}(F)$. If you plug-in $\widehat{F}_n$ into the statistical functional:

$$T_{\mathsf{mean}}(\widehat{F}_n) = \int x d\widehat{F}_n(x) = \sum_{i=1}^{n} X_i \frac{1}{n} = \sum_{i=1}^{n} \frac{X_i}{n} = \bar{X}_n.$$

See also earlier discussions for the method of moments.

We can apply the same principle to estimate the median:

$$T_{\mathsf{med}}(\widehat{F}_n) = \widehat{F}_n^{-1}(0.5)$$

and other quantiles of a distribution. (These estimators turn out to be equal to the sample median, or the corresponding sample quantiles.)

Are these estimators any good? Since Glivenko-Cantelli gives us that

$$\widehat{F}_n \overset{a.s.}{\to} F$$

can we also conclude that

$$T(\widehat{F}_n) \overset{a.s.}{\to} T(F)?$$

The answer is sometimes, but not always. In this lecture we discuss some cases when this convergence holds. For an example where the plug-in estimate is not consistent, note that using the empirical distribution to estimate a continuous density will not give you a consistent estimate.

When is the plug-in estimator $T(\widehat{F}_n)$ consistent? We have seen one example of consistency in the delta method above. More generally, consistency requires conditions on the smoothness (differentiability) of $T(F)$, which puts us in another predicament. How do you take a derivative with respect to a function? To do this, we expand the notion of the derivative.

Note that we leave out some details in the below two definitions please see a course on functional analysis for more details.

Definition 10.7 (Gâteaux derivative) *The Gâteaux derivative of $T(F)$ in the direction G (where G is a function in the same class as F) is defined by*

$$L_F(T; G) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon G) - T(F)}{\epsilon}. \tag{10.10}$$

Or equivalently, if $D = G - F$, than equation (10.10) becomes

$$L_F(T; D) = \lim_{\epsilon \to 0} \frac{T(F + \epsilon D) - T(F)}{\epsilon}. \tag{10.11}$$

From a statistical perspective, when F and G are CDFs, the Gâteaux derivative represents the rate of change in the statistical functional when are CDF F is "contaminated" by a small (ϵ) amount of G.

If the Gâteaux derivative of $T(F)$ exists, is that enough to have $T(\widehat{F}_n) \to T(F)$? Unfortunately, no. We need an even **stronger** condition known as *Hadamard differentiability.*

Definition 10.8 (Hadamard Differentiability) *A functional $T(F)$ is Hadamard Differentiable if, for any sequence $\epsilon_n \overset{n\to\infty}{\to} 0$ and a sequence of functions D_n satisfying $\sup_x |D_n(x) - D(x)| \overset{n\to\infty}{\to} 0$ (where $D_1, \ldots, D_n, \ldots$ and D are in the same function class as F) we have*

$$\frac{T(F + \epsilon_n D_n) - T(F)}{\epsilon_n} \to L_F(T; D) < \infty.$$

See the connection with Glivenko-Cantelli. If $T(F)$ is Hadamard differentiable, then $T(\widehat{F}_n) \overset{P}{\to} T(F)$.

10.4.1 The Influence Function

The *influence function* of a statistical functional T_{target} is a special case of the Gâteaux derivative when $G = \delta_x$, where

$$\delta_x(u) = \left\{ \begin{array}{ll} 0, & \text{if } u < x \\ 1, & \text{if } u \geq x \end{array} \right\}$$

Note that the influence function of the EDF is called the empirical influence function. We will rely on the influence function to analyze asymptotic properties of our estimators. The influence function is also related to the robustness of an estimator and plays a key role in the semi-parametric statistics (Van der Vaart, 2000).

Definition 10.9 *Let X be a random variable with CDF F and let $T(F)$ be some functional. The influence function $L(x)$ of $T(F)$ is defined as*

$$L_F(x) = \lim_{\epsilon \to 0} \frac{T_{\mathsf{target}}((1-\epsilon)F + \epsilon\delta_x) - T_{\mathsf{target}}(F)}{\epsilon}. \tag{10.12}$$

A powerful feature of the influence function is that when the statistical functional T_{target} is Hadamard differentiable then

$$\sqrt{n}\left(T_{\mathsf{target}}(\widehat{F}_n) - T_{\mathsf{target}}(F)\right) \overset{D}{\to} N\left(0, \mathbb{V}_{\mathsf{target}}(F) = \int L_F^2(x) dF(x)\right) \tag{10.13}$$

and a consistent estimator of $\mathbb{V}_{\mathsf{target}}(F)$ is $\mathbb{V}_{\mathsf{target}}(\widehat{F}_n) = \frac{1}{n}\sum_{i=1}^n L^2_{\widehat{F}_n}(X_i)$. This is related to the functional delta method see Chapters 3 and 20 of Van der Vaart (2000) for more details.

10.4.2 Linear Functionals

Many statistical functionals are of the form

$$T_\omega(F) = \int \omega(x) dF(x),$$

where ω is some function. This type of statistical functional is called a *linear* functional.

The empirical estimators of linear functionals have the following form:

$$T_\omega(\widehat{F}_n) = \int \omega(x) d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \omega(X_i).$$

Moreover, for any linear function, it's influence function is simply:

$$L_F(x) = \omega(x) - T_\omega(F). \tag{10.14}$$

A short computational explanation for the above:

$$\begin{aligned} L_F(x) &= \lim_{\epsilon \to 0} \frac{T_\omega(F_\epsilon) - T_\omega(F)}{\epsilon} = \lim_{\epsilon \to 0} \frac{\int \omega(x) dF_\epsilon(x) - T_\omega(F)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\int \omega(x) dF_\epsilon(x) - \int T_\omega(F) dF_\epsilon(x)}{\epsilon} = \omega(x) - T_\omega(F). \end{aligned}$$

Theorem 10.10 *Suppose that T_ω is a linear functional with an influence function $L_F(x)$ defined in equation (10.14) and $\int \omega^2(x) dF(x) < \infty$. Then*

$$\sqrt{n}\left(T_\omega(\widehat{F}_n) - T_\omega(F)\right) \xrightarrow{D} N\left(0, \mathbb{V}_\omega(F) = \int L_F^2(x) dF(x)\right)$$

and a consistent estimator of $\mathbb{V}_\omega(F)$ is $\mathbb{V}_\omega(\widehat{F}_n) = \frac{1}{n}\sum_{i=1}^n L^2_{\widehat{F}_n}(X_i)$.

Proof:

It is easy to see that

$$T_\omega(\widehat{F}_n) - T_\omega(F) = \int L_F(x) d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n L_F(X_i).$$

Moreover,

$$\mathbb{E}(L_F(X_i)) = \int L_F(x) dF(x) = \int \left(\omega(x) - T_\omega(F)\right) dF(x) = T_\omega(F) - T_\omega(F) = 0.$$

Note also that since by assumption $\int \omega^2(x) dF(x) < \infty$, Jensen's inequality gives us that $\int \omega(x) dF(x) < \infty$ (Existence of higher order moments implies existence of lower order moments.) Furthermore, note that

$$\int L_F^2(x) dF(x) = \int \omega^2(x) dF(x) - T_w^2(F) = \int \omega^2(x) dF(x) - (\int \omega(x) dF(x))^2 < \infty.$$

Thus, by the central limit theorem,

$$\sqrt{n}\left(T_\omega(\widehat{F}_n) - T_\omega(F)\right) \xrightarrow{D} N\left(0, \mathbb{V}_\omega(F) = \int L_F^2(x) dF(x)\right)$$

Moreover,

$$\begin{aligned} \mathbb{V}_\omega(\widehat{F}_n) &= \int L^2_{\widehat{F}_n}(x) d\widehat{F}_n(x) = \int \left(\omega^2(x) - 2\omega(x) T_\omega(\widehat{F}_n) + T_\omega^2(\widehat{F}_n)\right) d\widehat{F}_n(x) \\ &= \int \omega^2(x) d\widehat{F}_n(x) - T_\omega^2(\widehat{F}_n). \end{aligned} \tag{10.15}$$

By the Law of Large Numbers and the continuous mapping theorem,

$$T_\omega^2(\widehat{F}_n) \xrightarrow{P} T_\omega^2(F).$$

And

$$\int \omega^2(x) d\widehat{F}_n(x) = T_{\omega^2}(\widehat{F}_n) \xrightarrow{P} T_{\omega^2}(F).$$

Therefore, we conclude that when $\int \omega^2(x) dF(x) < \infty$,

$$\mathbb{V}_\omega(\widehat{F}_n) = \int \omega^2(x) d\widehat{F}_n(x) - T_\omega^2(\widehat{F}_n) \xrightarrow{P} \mathbb{V}_\omega(F) = \int L_F^2(x) dF(x).$$

■

10.4.3 Non-linear Functionals

Some statistical functionals are not linear. For instance, the median

$$T_{\mathsf{med}}(F) = F^{-1}(0.5)$$

is not a linear functional. However, it turns out that this functional is Hadamard differentiable and so, we can use Equation (10.13).

The influence function of the functional T_{med} is

$$L_F(x) = \frac{1}{2p(F^{-1}(0.5))},$$

where p is the PDF of F.

Note that $F^{-1}(0.5) = T_{\mathsf{med}}(F)$ is the median of F. So this shows not only the asymptotic normality of the sample median but also its limiting variance, which is inversely related to the PDF at the median.

10.5 The Bootstrap

Assume we are given the data points $X_1, \cdots, X_n$. Let $M_n = \mathsf{median}\{X_1, \cdots, X_n\}$. As we have seen in previous sections, our estimate for the population median will be $M_n = T(\widehat{F}_n)$. How do we estimate the variance/MSE/other uncertainty for our estimate M_n? We will use a method called the bootstrap (or empirical/non-parametric bootstrap). The bootstrap can be used in many complex scenarios such as this one.

Here is the procedure. First, we *sample with replacement* n "new" data points from our n data points, leading to a set of new observations denoted as $X_1^{*(1)}, \cdots, X_n^{*(1)}$. We then repeat the sample procedure again, generating a new sample from the original dataset $X_1, \cdots, X_n$ by sampling with replacement, leading to another new sets of observations $X_1^{*(2)}, \cdots, X_n^{*(2)}$. Now we keep repeating the same process of generating new sets of observations, after B rounds, we will obtain

$$\begin{array}{ccc} X_1^{*(1)}, & \cdots, & X_n^{*(1)} \\ X_1^{*(2)}, & \cdots, & X_n^{*(2)} \\ \vdots & \vdots & \vdots \\ X_1^{*(B)}, & \cdots, & X_n^{*(B)}. \end{array}$$

So, in total, we will have B sets of data points. Each set of the data points, say $X_1^{*(1)}, \cdots, X_n^{*(1)}$, is called a bootstrap sample. This sampling approach–sample with replacement from the original dataset–is called the *empirical bootstrap*, and was introduced by Bradley Efron (sometimes this approach is also called *Efron's bootstrap* or *nonparametric bootstrap*)[1].

Now for each "new" set of data, we compute the sample median. This leads to B sample medians, called

[1]The name is a reference to "pulling yourself by your bootstraps" quote from Surprising Adventures of Baron Munchausen.

bootstrap medians:

$$\begin{aligned} M_n^{*(1)} &= \mathsf{median}\{X_1^{*(1)}, \cdots, X_n^{*(1)}\} \\ M_n^{*(2)} &= \mathsf{median}\{X_1^{*(2)}, \cdots, X_n^{*(2)}\} \\ &\vdots \\ M_n^{*(B)} &= \mathsf{median}\{X_1^{*(B)}, \cdots, X_n^{*(B)}\}. \end{aligned}$$

Here are some interesting things you can do with all these bootstrap samples.

- **Bootstrap estimate of the variance of the sample median.** We will use the sample variance of $M_n^{*(1)}, \cdots, M_n^{*(B)}$ as an estimate of the variance of sample median M_n. Namely, we will use

$$\widehat{\mathsf{Var}}_B(M_n) = \frac{1}{B-1}\sum_{\ell=1}^{B}\left(M_n^{*(\ell)} - \bar{M}_B^*\right)^2, \quad \bar{M}_B^* = \frac{1}{B}\sum_{\ell=1}^{B} M_n^{*(\ell)},$$

  as an estimate of $\mathsf{Var}(M_n)$.

- **Bootstrap estimate of the MSE of the sample median.** Moreover, we can estimate the MSE by

$$\widehat{\mathsf{MSE}(M_n)} = \frac{1}{B}\sum_{\ell=1}^{B}\left(M_n^{*(\ell)} - M_n\right)^2.$$

- **Bootstrap confidence interval for the of the sample median.** In addition, we can construct a $1-\alpha$ confidence interval of the population median via

$$M_n \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathsf{Var}}_B(M_n)}.$$

Well... this sounds a bit weird–we generate new data points by sampling from the existing data points. However, under some conditions, this approach does work! Meaning the above estimates, lead to some sensible approximations. Here is a brief explanation on why this approach works (in some cases).

Let $X_1, \cdots, X_n \sim F$. When we sample with replacement from $X_1, \cdots, X_n$, what is the distribution we are sampling from? Let $Z \equiv X_i^{*(b)}$, $i \in \{1, \ldots, n\}, b \in \{1, \ldots, B\}$, then Z has the following probability distribution:

$$P(Z = X_i) = \frac{1}{n}, \quad \text{for each } i = 1, 2, \cdots, n.$$

Thus, each set of the bootstrap sample is an IID sample from $\widehat{F}_n$. Namely,

$$\begin{aligned} X_1^{*(1)}, \cdots, X_n^{*(1)} &\sim \widehat{F}_n \\ X_1^{*(2)}, \cdots, X_n^{*(2)} &\sim \widehat{F}_n \\ &\vdots \\ X_1^{*(B)}, \cdots, X_n^{*(B)} &\sim \widehat{F}_n. \end{aligned}$$

Because a bootstrap median, say $M_n^{*(b)}$, is the sample median of $X_1^{*(b)}, \cdots, X_n^{*(b)}$. Its CDF is $F_{M_n^{*(b)}}(x) = T(\widehat{F}_n)$, whereas the CDF of the sample median M_n is $F_{M_n}(x) = T(\widehat{F})$. We know that $\widehat{F}_n \overset{a.s.}{\to} F$. Thus,

as long as T is "smooth" with respect to F (see previous sections), $F_{M_n^{*(b)}}(x) \to F_{M_n}(x)$. This has many implications. For an example, when two CDFs are similar, their variances will be similar as well, i.e.,

$$\mathsf{Var}\left(M_n^{*(b)}|X_1,\cdots,X_n\right) \approx \mathsf{Var}(M_n).$$

The reason why we condition on $X_1,\cdots,X_n$ on the left-hand-side, is because when we compute the bootstrap estimate, the original observations $X_1,\cdots,X_n$ are fixed. Now the bootstrap variance estimate $\widehat{\mathsf{Var}}_B(M_n)$ is just a sample variance of $M_n^{*(b)}$:

$$\widehat{\mathsf{Var}}_B(M_n) = \frac{1}{B-1}\sum_{\ell=1}^{B}\left(M_n^{*(\ell)} - \bar{M}_B^*\right)^2 \approx \mathsf{Var}\left(M_n^{*(\ell)}|X_1,\cdots,X_n\right).$$

Lets take a step back and examine what is going on here. We have learned that the bootstrap sample is a new random sample from the EDF $\widehat{F}_n$. Each bootstrap sample itself forms another EDF called the bootstrap EDF, denoted as $\widehat{F}_n^*$. Namely, let $X_1^*,\cdots,X_n^*$ be a bootstrap sample. Then the bootstrap EDF is

$$\widehat{F}_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i^* \le x).$$

Note that our goal is to estimate the $\mathrm{Var}(M_n) = T_{\mathsf{target}}(F)$ and we are using the bootstrap sample, that is $\widehat{\mathsf{Var}}_B(M_n) = T_{\mathsf{target}}(\widehat{F}_n^*)$. So the estimator using the bootstrap sample is another plug-in estimator but now we are plugging in the bootstrap EDF $\widehat{F}_n^*$. As a result, we can use the results of the previous sections. That is the bootstrap estimator will be consistent for linear functionals whenever $T_{\omega^2}(F) < \infty$.

Consistency of the bootstrap variance estimator. Suppose our parameter of interest is θ, we have obtained an estimate $\widehat{\theta}_n$ based on the EDF plug-in and now we are looking to estimate the uncertainty of $\widehat{\theta}_n$ using the bootstrap. We generate bootstrap samples from the EDF $\widehat{F}_n$ and obtain several realizations of $\widehat{\theta}_n^*$'s. Namely, we generate

$$\widehat{\theta}_n^{*(1)},\cdots,\widehat{\theta}_n^{*(B)}$$

and use their sample variance, $\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*)$, as an estimator of $\mathsf{Var}(\widehat{\theta}_n)$. Note that $\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*)$ is

$$\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*) = \frac{1}{B-1}\sum_{\ell=1}^{N}\left(\widehat{\theta}_n^{*(\ell)} - \bar{\widehat{\theta}}_{n,B}^*\right), \quad \bar{\widehat{\theta}}_{n,B}^* = \frac{1}{B}\sum_{\ell=1}^{B}\widehat{\theta}_n^{*(\ell)}.$$

Then

$$\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*) \overset{B\to\infty}{\to} \mathsf{Var}(\widehat{\theta}_n^*|\widehat{F}_n). \tag{10.16}$$

Note that $\cdot|\widehat{F}_n$ means *conditioned* on $\widehat{F}_n$ being fixed. The reason why above we have a convergence to the conditioned variance is because when we generate bootstrap samples, the original EDF $\widehat{F}_n$ is fixed (and we are generating from it).

Now, to argue that the bootstrap variance $\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*)$ is a good estimate of the original variance, we need to argue

$$\mathsf{Var}(\widehat{\theta}_n^*|\widehat{F}_n) \to \mathsf{Var}(\widehat{\theta}_n).$$

Or more formally,

$$\frac{\mathsf{Var}(\widehat{\theta}_n^*|\widehat{F}_n)}{\mathsf{Var}(\widehat{\theta}_n)} \to 1 \tag{10.17}$$

(people generally use the ratio expression because both quantities often converge to 0 when the sample size $n \to \infty$). Under weak conditions (beyond the scope of this course) this convergence indeed holds.

Validity of bootstrap confidence interval. How about the validity of the bootstrap confidence interval? This will also often hold. Read more on the Berry-Essen bound if interested.

Example: Sample Mean. Our parameter of interest is now the mean of a distribution $T_{\mathsf{target}} = T_{\mathsf{mean}}$. The mean of a distribution has the form

$$\mu = T_{\mathsf{mean}}(F) = \int x dF(x).$$

The plug-in estimator is

$$\widehat{\mu}_n = T_{\mathsf{mean}}(\widehat{F}_n) = \int x d\widehat{F}_n(x) = \bar{X}_n$$

and the bootstrap estimator is

$$\widehat{\mu}_n^* = T_{\mathsf{mean}}(\widehat{F}_n^*) = \int x d\widehat{F}_n^*(x) = \bar{X}_n^*.$$

It is clear from the Central Limit Theorem that

$$\sqrt{n}(\widehat{\mu}_n - \mu) \xrightarrow{D} N(0, \mathsf{Var}(\sqrt{n} \cdot T_{\mathsf{mean}}(\widehat{F}_n)))$$

and for the bootstrap it can be shown that the following holds

$$\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \xrightarrow{D} N(0, \mathsf{Var}(\sqrt{n} \cdot T_{\mathsf{mean}}(\widehat{F}_n^*)|\widehat{F}_n)).$$

In this case, we know that

$$\mathsf{Var}(\sqrt{n} \cdot T_{\mathsf{mean}}(\widehat{F}_n)) = \mathsf{Var}(\sqrt{n}\bar{X}_n) = \mathsf{Var}(X_i) \Longrightarrow \mathbb{V}_{\mathsf{mean}}(F) = \mathsf{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}^2(X_i) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2.$$

Therefore, the bootstrap variance is

$$\mathsf{Var}(\sqrt{n} \cdot T_{\mathsf{mean}}(\widehat{F}_n^*)|\widehat{F}_n) = \mathbb{V}_{\mathsf{mean}}(\widehat{F}_n) = \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2.$$

By the Law of Large Numbers and continuous mapping theorem,

$$\int x^2 d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} \mathbb{E}(X_i^2) = \int x^2 dP(x)$$

$$\int x d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}(X_i) = \int x dP(x).$$

Thus, by Slutsky's theorem,

$$\mathbb{V}_{\mathsf{mean}}(\widehat{F}_n) \xrightarrow{P} \mathbb{V}_{\mathsf{mean}}(F).$$

Thus, the bootstrap variance estimator converges to the true variance estimator and we can conclude that

$$\frac{\mathsf{Var}(T_{\mathsf{mean}}(\widehat{F}_n^*)|\widehat{F}_n)}{\mathsf{Var}(T_{\mathsf{mean}}(\widehat{F}_n))} \xrightarrow{P} 1.$$

As a result, the bootstrap variance estimator is consistent and the bootstrap confidence interval is also valid.

Generalization to other statistics. The bootstrap can be applied to many other statistics such as sample quantiles, interquartile range, skewness (related to $\mathbb{E}(X^3)$), kurtosis (related to $\mathbb{E}(X^4)$), ...etc. The theory basically follows from the same idea.

Failure of the bootstrap. However, the bootstrap may fail for some statistics. One example is the minimum value of a distribution. Here is an illustration why the bootstrap fails. Let $X_1, \cdots, X_n \sim \mathsf{Uni}[0,1]$ and $X_{(1)} = \min\{X_1, \cdots, X_n\}$ be the minimum value of the sample. Then it is known that

$$n \cdot X_{(1)} \xrightarrow{D} \mathsf{Exp}(1).$$

♠ : Think about why it converges to exponential distribution.

Thus, $X_{(1)}$ has a continuous distribution. Assume we generate a bootstrap sample $X_1^*, \cdots, X_n^*$ from the original observations. Now let $X_{(1)}^* = \min\{X_1^*, \cdots, X_n^*\}$ be the minimum value of a bootstrap sample. Because each X_ℓ^* has an equal probability $(\frac{1}{n})$ of selecting each of $X_1, \cdots, X_n$, this implies

$$P(X_\ell^* = X_{(1)}) = \frac{1}{n}.$$

Namely, for each observation in the bootstrap sample, we have a probability of $1/n$ selecting the minimum value of the original sample. Thus, the probability that we *do not select* $X_{(1)}$ in the bootstrap sample is

$$P(\text{none of } X_1^*, \cdots, X_n^* \text{ select } X_{(1)}) = \left(1 - \frac{1}{n}\right)^n \approx e^{-1}.$$

This implies that with a probability $1 - e^{-1}$, one of the observation in the bootstrap sample will select the minimum value of the original sample $X_{(1)}$. Namely,

$$P(X_{(1)}^* = X_{(1)}) = 1 - e^{-1}.$$

Thus, $X_{(1)}^*$ has a huge probability mass at the value $X_{(1)}$, meaning that the distribution of $X_{(1)}^*$ will not be close to an exponential distribution.

References

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.