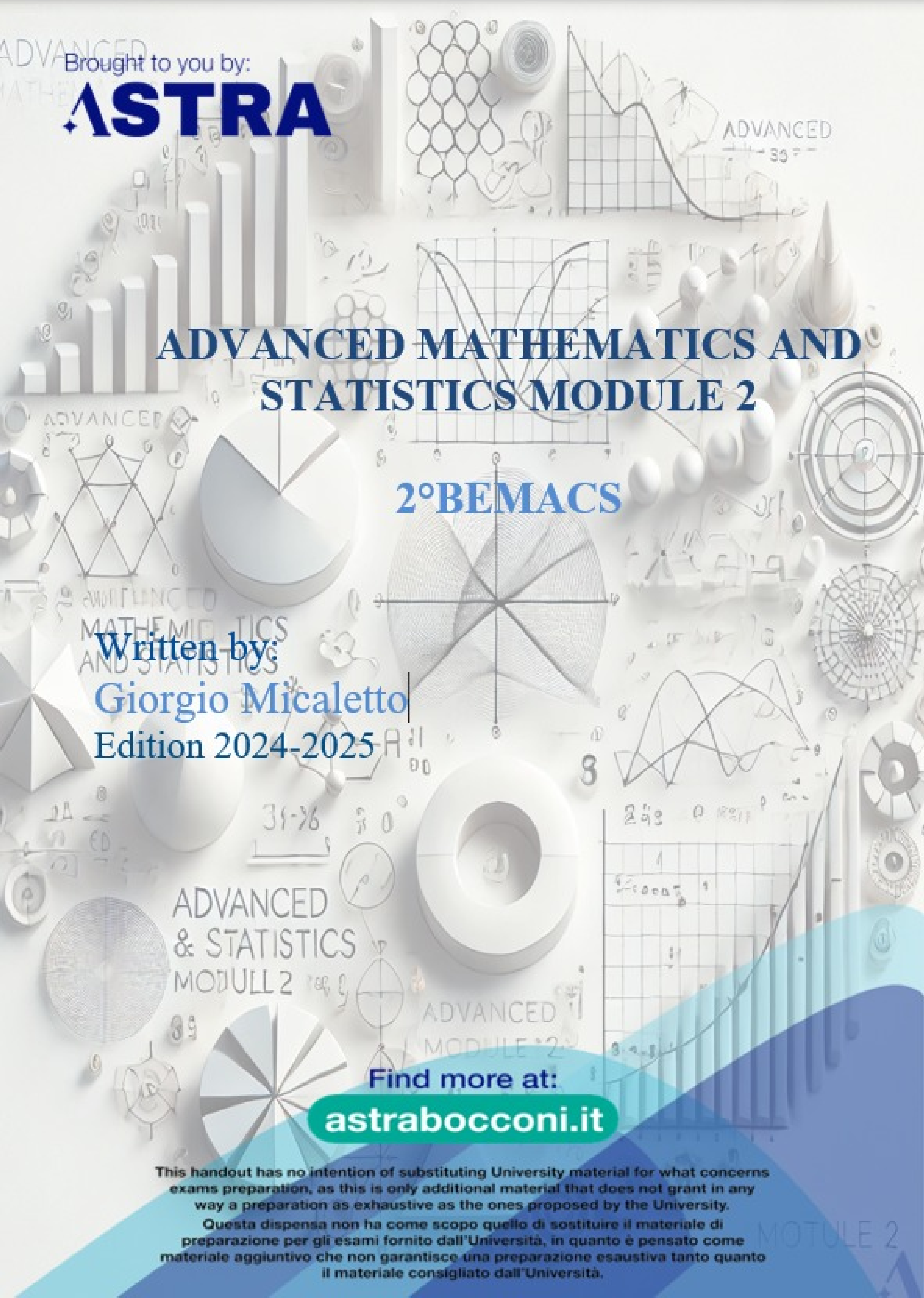


Brought to you by:
ASTRA

ADVANCED MATHEMATICS AND STATISTICS MODULE 2

2°BEMACS

Written by:
Giorgio Micaletto
Edition 2024-2025

Find more at:
astrabocconi.it

This handout has no intention of substituting University material for what concerns exams preparation, as this is only additional material that does not grant in any way a preparation as exhaustive as the ones proposed by the University.

Questa dispensa non ha come scopo quello di sostituire il materiale di preparazione per gli esami fornito dall'Università, in quanto è pensato come materiale aggiuntivo che non garantisce una preparazione esaustiva tanto quanto il materiale consigliato dall'Università.

Contents

Lecture 1

Introduction

We define a random variable X as a function that maps outcomes from a sample space $\mathcal{S}$ to real numbers, formally denoted as $X : \mathcal{S} \to \mathbb{R}$. This concept is essential for quantifying uncertainty and making predictions based on probabilistic models. **Examples**:

- **Dice Roll:** The experiment of throwing two dice has a sample space $\mathcal{S}$ with cardinality 36. A random variable X, representing the sum of the dice, reduces the problem's dimensionality to 11 possible outcomes (2 through 12). For example, the probability of rolling a sum of 2 (both dice showing 1) is given by:

$$\mathbb{P}[X = 2] = \mathbb{P}[\mathcal{S} = (1, 1)] = \frac{1}{36} \tag{1}$$

- **Poker Hands:** In poker, the sample space $\mathcal{S}$ for all possible 5-card hands from a 52-card deck is $\text{card}(\mathcal{S}) = \binom{52}{5}$. If we are interested in the number of spades in a hand, this random variable X can take 6 values (0 through 5).

Before delving into random variables, an useful function that is worth defining is the indicator function. Formally, if A is a subset of $\mathbb{R}$, we define the indicator function of A as

$$\mathbb{1}_A = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

.

Classification of Random Variables

Random variables can be broadly divided in discrete and continuous.

Discrete Random Variables

What we have described above are discrete random variables.

To give a more formal definition, a discrete random variable has a countable (see refresher 0.1) set of possible values. Its probability distribution is characterized by a probability mass function (PMF) $p_X : \mathbb{R} \to \mathbb{R}_+$, satisfying:

- $\sum_{x \in D} p_X(x) = 1$ for some countable set $D \subset \mathbb{R}$.
- $p_X(x) = 0$ for all $x \notin D$.

- $D = X(\mathcal{S})$ and $p_X(x) = \mathbb{P}[X = x]$.

Discrete distributions that you should know:

1. **Binomial**: $X \sim \text{Binomial}(n, p)$ with $n \geq 1$ and $p \in [0, 1]$ defined as

$$p_X(x) = \binom{n}{x} p^x (1-p)^{(n-x)} \mathbb{1}_{\{\mathbb{N}\}}(x) \tag{3}$$

   It describes is the success out of n independent trial each with probability of success p.

2. **Bernoulli**: A special case of the Binomial is when $n = 1$, in which case we have that $X \sim \text{Bernoulli}(p)$.

$$p_X(x) = p^x (1-p)^{n-x} \mathbb{1}_{[0,1]}(x) \tag{4}$$

3. **Poisson**: $X \sim \text{Poisson}(\lambda)$, with $\lambda > 0$ and $D = \{0, 1 \ldots\}$, described as

$$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!} \mathbb{1}_{\{D\}}(x) \tag{5}$$

   It is the counting number of events in a unit of time, with λ being the average number of events

4. **Negative Binomial** $X \sim \text{Neg-binomial}(r, p)$ with $r \geq 1$, $p \in [0, 1]$ and $D = \{0, 1, \ldots\}$, defined as

$$p_X(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \mathbb{1}_{\{D\}}(x) \tag{6}$$

   It describes the number of trials you must experience before gaining the r-th success.

5. **Geometric** A particular case of the negative binomial is when $r = 1$, in which case we have that $X \sim \text{Geom}(p)$, with $D = \{0, 1, \ldots\}$:

$$p_X(x) = p \cdot (1-p)^{x-1} \mathbb{1}_{\{D\}}(x) \tag{7}$$

   And it describes the number of values before gaining the first success.

Continuous Random Variables

A continuous random variable, X is characterized by a probability density function, $f_X : \mathbb{R} \to \mathbb{R}_+$ such that

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1 \tag{8}$$

Popular continuous distributions are:

1. **Negative Exponential** $X \sim \text{Exponential}(\lambda)$, which is defined as:

   $$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{\{\mathbb{R}_+\}}(x) \tag{9}$$

   This probability is used to find the time to the occurrence of an event, and is generally used in survival analysis.

2. **Gamma** The negative exponential is a special case of the gamma distribution. If $X \sim \text{Gamma}(a, b)$ then it's defined as:

   $$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \mathbb{1}_{\{\mathbb{R}_+\}}(x) \tag{10}$$

3. **Normal** $X \sim \mathcal{N}(\mu, \sigma^2)$, defined as:

   $$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathbb{1}_{\{\mathbb{R}\}}(x) \tag{11}$$

4. **Uniform** $X \sim \text{Uniform}(a, b)$ with $D = [a, b]$, where $-\infty < a < b < \infty$

   $$f_X(x) = \frac{1}{b-a} \mathbb{1}_{\{[a,b]\}}(x) \tag{12}$$

   A characteristic of this distribution is that any interval of the same size has the same probability.

Cumulative Distribution Functions

Cumulative distribution functions (CDFs) provide a comprehensive method for characterizing the behavior of both discrete and continuous random variables.

For a given random variable X, its CDF, denoted as F_X, is a function that maps real numbers to the non-negative real numbers. Formally, it is defined as:

$$F_X : \mathbb{R} \to \mathbb{R}_+, \tag{13}$$

where for any real number x, the value of $F_X(x)$ is given by the probability that X assumes a value less than or equal to x:

$$F_X(x) = \mathbb{P}[X \leq x].$$

Two random variables, X and Y, are considered to be equal in distribution, denoted by $X \overset{d}{=} Y$, if their CDFs are identical, i.e., $F_X = F_Y$. This indicates that X and Y share the same distribution across their respective domains. In the context of discrete random variables, equality in distribution implies that their probability mass functions (PMFs) are also equivalent:

$$p_X = p_Y \Rightarrow X \overset{d}{=} Y.$$

This means that for every possible value, the probability of X and Y assuming that value is the same. For continuous random variables, equality in distribution is indicated by the equivalence of their probability density functions (PDFs):

$$f_X = f_Y \Rightarrow X \overset{d}{=} Y.$$

Thus, the likelihood of X and Y assuming values within any given interval is identical. The principle of identification leverages CDFs (as well as PMFs and PDFs for discrete and continuous variables, respectively) to ascertain whether two random variables have the same distribution. This principle is instrumental in the study of probability and statistics, facilitating the comparison and analysis of random variables' behaviors.

Introduction to Moment Generating Functions

The moment generating function (MGF) of a random variable X is defined as the expected value of e^{tX}, where t is a real number. Formally, the MGF, $m_X(t)$, is given by:

$$m_X(t) = \mathbb{E}[e^{tX}] = \int_{\inf(D)}^{\sup(D)} e^{tx} f_X(x)\, dx \tag{14}$$

for continuous random variables, and

$$m_X(t) = \mathbb{E}[e^{tX}] = \sum_{x \in D} e^{tx} p_X(x) \tag{15}$$

for discrete random variables, where $f_X(x)$ is the probability density function (PDF) of X, $p_X(x)$ is the probability mass function (PMF) of X and $\inf(D)$ and $\sup(D)$ denote the lower and upper bounds of D.

The significance of the MGF lies in its ability to characterize the distribution of a random variable. If the MGF of a random variable exists for t in some neighborhood of 0, it uniquely determines the probability distribution of the random variable. Moreover, moments of the random variable (such as mean, variance) can be derived by taking derivatives of the MGF with respect to t and evaluating at $t = 0$:

$$\mu'_n = \left.\frac{d^n m_X(t)}{dt^n}\right|_{t=0} \tag{16}$$

where μ'_n denotes the nth moment about the origin of the random variable X.

Lecture 2

Moment Generating Functions

The moment generating function (MGF) of a random variable X is $\mathbb{E}[e^t X] = m_X(t)$. If such an expectation is finite neighbourhood of the origin, then the moment generating function can be written as

$$m_X(t) = \mathbb{E}[e^{tX}] = \begin{cases} \displaystyle\sum_{x \in D} e^{tx} p_X(x) & \text{for discrete random variables,} \\ \displaystyle\int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx & \text{for continuous random variables,} \end{cases} \tag{17}$$

As you can see, for the continuous random variables, the MGF is an integral transformation. An integral transformation is a mathematical process that converts one function into another through the integration of the product of the original function and a kernel function (Refresher 0.2) over a specified domain.

In our case:

- e^{tx} represents the kernel function, dependent on the parameter t,
- $f_X(x)$ is the original function,
- D is the domain on which we integrate.

This definition places the MGF in the category of integral transformations, as it transforms the PDF of a random variable X into a new function $m_X(t)$ through integration over D. You can rely on m_X to evaluate moments of X, $\mathbb{E}X^k$. Here the MGFs for the most common distributions, with their proofs:

1. **Binomial** If $X \sim \text{Binom}(n, p)$, with $D = \mathbb{N}$:

$$\begin{aligned} m_X(t) &= \mathbb{E}[e^{tX}] \\ &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\ &= (pe^t + 1 - p)^n \end{aligned}$$

   We are able to solve the summation because it is a binomial sum, which we recall is given by $\sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} = (a+b)^n$, with $a = (pe^t)^x$ and $b = (1-p)$

2. **Bernoulli**, as the Bernoulli distribution is a binomial with $n = 1$, the MGF of $X \sim \text{Bernoulli}(p)$ will be

$$m_X(t) = (pe^t + 1 - p)$$

3. **Geometric** If $X \sim \text{Geom}(p)$, with $D = \mathbb{N} \backslash 0$, then

$$\begin{aligned} m_X(t) &= \mathbb{E}[e^{tX}] \\ &= \sum_{x=1}^{\infty} e^{tx} p (1-p)^{x-1} \\ &= \frac{p}{1-p} \sum_{x=1}^{\infty} (e^t(1-p))^x \\ &= \frac{p}{1-p} \frac{e^t(1-p)}{1-e^t(1-p)} \\ &= \frac{pe^t}{1-e^t(1-p)} \end{aligned}$$

   We are able to solve the summation because it is a geometric series,

$$\sum_{x=k}^{\infty} r^x = \frac{r^k}{1-r} \tag{18}$$

   with $r = e^t(1-p)$. As the geometric series requires $|r| < 1$ (refresher 0.3), we have to set that $|e^t(1-p)| < 1$ which means that $e^t < \frac{1}{1-p}$, which further means that $t < -\log(1-p)$, which implies that $\log(1-p) < 0$. This further means that the MGF exists $\forall t < -\log(1-p)$

4. **Poisson** If $X \sim \text{Pois}(\lambda)$, with $D = \mathbb{N}$

$$\begin{aligned} m_X(t) &= \mathbb{E}[e^{tX}] \\ &= \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda} e^{e^t \lambda} \\ &= e^{\lambda(e^t - 1)} \end{aligned}$$

   We were able to solve the summation because it is an exponential series with $b = (\lambda e^t)$ and $k = x$

5. **Negative Binomial** $X \sim \text{Neg-binomial}(r, p)$ with $r \geq 1$, $p \in [0, 1]$ and $D = \{0, 1, \dots\}$

$$\begin{aligned} m_X(t) &= \mathbb{E}[e^{tX}] \\ &= \sum_{x=0}^{\infty} e^{tx} \binom{x + r - 1}{x} p^x (1-p)^r \\ &= \left(\frac{1-p}{1-pe^t} \right)^r, \quad \text{for } t < -\ln(p) \end{aligned}$$

6. **Exponential** If $X \sim \text{Exp}(\lambda)$ with $D = \mathbb{R}_+$

$$\begin{aligned} m_X(t) &= \mathbb{E}[e^{tX}] \\ &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^{\infty} e^{-(\lambda - t)x} dx \\ &= \lambda \frac{1}{\lambda - t}, \forall t < \lambda \end{aligned}$$

7. **Uniform** If $X \sim \text{Unif}(0, 1)$, then

$$\begin{aligned} m_X(t) &= \int_0^1 e^{tx} dx \\ &= \frac{1}{t}(e^t - 1) \end{aligned}$$

Note: there's no discontinuity at $t = 0$, as taking the limit of the first order taylor expansion (refresher 0.3) tells us that

$$\lim_{t \to 0} \frac{e^t - 1}{t} = 1 \tag{19}$$

8. **Normal** If $X \sim \mathcal{N}(0,1)$

$$\begin{aligned}
m_X(t) &= \mathbb{E}[e^{tX}] \\
&= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2} + tx} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2 - 2tx + t^2 - t^2)} dx \\
&= \frac{e^{\frac{t^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2} dx \\
&= \frac{e^{\frac{t^2}{2}}}{\sqrt{2\pi}} \sqrt{2\pi} \\
&= e^{\frac{t^2}{2}}
\end{aligned}$$

For $Z \sim \mathcal{N}(\mu, \sigma^2)$, we have that $Z = \sigma X + \mu$

$$\begin{aligned}
m_Z(t) &= \mathbb{E}[e^{t\sigma X + t\mu}] \\
&= e^{t\mu} \mathbb{E}[e^{t\sigma X}] \\
&= e^{t\mu} m_X(t\sigma) \\
&= e^{t\mu} e^{\frac{(t\sigma)^2}{2}}
\end{aligned}$$

9. **Gamma** If $X \sim \text{Gamma}(a, b)$,

$$\begin{aligned}
m_X(t) &= \mathbb{E}[e^{tX}] \\
&= \int_0^{\infty} e^{tx} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \, dx \\
&= \frac{b^a}{\Gamma(a)} \int_0^{\infty} e^{tx} x^{a-1} e^{-bx} \, dx \\
&= \frac{b^a}{\Gamma(a)} \frac{\Gamma(a)}{(b-t)^a} \\
&= (\frac{b}{b-t})^a
\end{aligned}$$

We can solve the integral because, given the property of any PDF f_X, $\int_{-\infty}^{\infty} f_X(x)dx = 1$, in this case, given our PDF, taking out $\frac{b^a}{\Gamma(a)}$, it follows that $\int_0^{\infty} x^{a-1}e^{-bx}\, dx$ must be equal to the inverse of $\frac{b^a}{\Gamma(a)}$, which is $\frac{\Gamma(a)}{b^a}$

Continuity theorem

The Continuity Theorem is a key result in probability theory that gives conditions under which a sequence of probability distributions converges to a limiting distribution. One common version of the theorem pertains to characteristic functions, as follows:

Let $\{\phi_n(t)\}$ be a sequence of characteristic functions corresponding to a sequence of probability distributions $\{F_n\}$. Suppose there exists a function $\phi(t)$ such that for every $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \phi_n(t) = \phi(t), \tag{20}$$

and $\phi(t)$ is continuous at $t = 0$. Then, $\phi(t)$ is the characteristic function of some probability distribution F, and the sequence of distributions $\{F_n\}$ converges weakly to F.

The theorem implies that if we can show the pointwise convergence of characteristic functions of a sequence of random variables to a limit, and this limit is continuous at the origin, then the sequence of random variables converges in distribution to a random variable with the characteristic function being the limit.

Lecture 3

More on MGFs

Properties of Moment Generating Functions:

- We can find the moment generating function of a linear transformation. Let $Y = a+bX$, if m_X is defined, then we can define $m_Y(t) = \mathbb{E}e^{t(a+bX)} = e^{at}\mathbb{E}e^{bX} = e^{at}m_X(bt)$. We have thus found m_Y (assuming that m_X is defined on bt)

- It is possible to recover a moment, $\mathbb{E}X^k$ from m_X, by evaluating its k-th derivative at $t = 0$:

$$\mathbb{E}X^k = \frac{d^k}{dt^k} m_X(t)\bigg|_{t=0} \tag{21}$$

A relevant moment generating function is $Y = X^2$, where $X \sim \mathcal{N}(0,1)$. In this case we say that $Y = \chi^2_{(1)}$, where $\chi^2_{(1)}$ is a chi-squared distribution with 1 degree of freedom. The chi-squared distribution with k degrees of freedom is a special case of the gamma distribution where the shape parameter α is $k/2$ and the scale parameter β is 2. Thus, its probability density function can be written as:

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} \quad \text{for } x > 0. \tag{22}$$

Here, k represents the degrees of freedom, which in this context, relates to the number of independent variables squared and summed to form the chi-squared statistic (refreshers 0.4). The chi-squared distribution is particularly useful in statistical tests that compare observed and expected frequencies to determine goodness-of-fit, test for independence, and estimate variances.

Introduction to Random Vectors

A random vector is a vector of random variables that represent multi-dimensional random phenomena, and which are defined on the same $\mathcal{S}$.

More rigorously:

Let $\mathcal{S}$ be a sample space. Consider p real-valued random variables $X_1, X_2, \ldots, X_p$, each defined on $\mathcal{S}$ and mapping into $\mathbb{R}$. Together, these random variables can be viewed as a vector-valued random variable $(X_1, X_2, \ldots, X_p)$ that maps from $\mathcal{S}$ into $\mathbb{R}^p$, formally written as:

$$(X_1, X_2, \ldots, X_p) : \mathcal{S} \to \mathbb{R}^p. \tag{23}$$

This vector-valued random variable associates each outcome in $\mathcal{S}$ with a p-dimensional vector in $\mathbb{R}^p$, where each component of the vector is the value of one of the p random variables at that outcome.

For a 2D random vector (X, Y), the components X and Y are individual random variables.

The CDF for a 2D random vector (X, Y), denoted $F_{X,Y}(x, y)$, is defined as:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) \tag{24}$$

Random vectors hold the same properties of random variables

1. Limits at Infinity: $F_{X,Y}(x, y)$ approaches the marginal CDFs $F_X(x)$ and $F_Y(y)$ as X and Y approach $+\infty$, respectively, and 1 if $X \to +\infty$ and $Y \to +\infty$.

2. Limits at Negative Infinity: $F_{X,Y}(x, y)$ approaches 0 as x or y approaches $-\infty$.

3. Continuity: $F_{X,Y}(x, y)$ is right-continuous, with $F_{X,Y}(x+h, y+k) \to F_{X,Y}(x, y)$ as $h, k \to 0$.

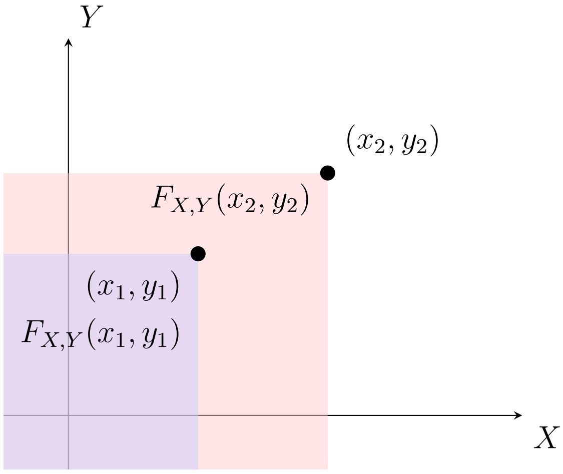


Figure 1: Visualization of Point 4, the shaded area in blue is $F_{X,Y}$ at point (x_1, y_1) and the area in red s $F_{X,Y}$ at point (x_2, y_2), which as you can see, is smaller

4. Ordering: If $x_1 \leq x_2$ and $y_1 \leq y_2$, then $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$.

   Note: we can also find the area that sits in the rectangle with vertices $(x_1, y_1), (x_1, y_2), (x_2, y_2), (x_2, y_1)$ with the following formula:

   $F_{X,Y}(x_2, y_2) - F_{X,Y}(x_2, y_1) - F_{X,Y}(x_1, y_2) + F_{X,Y}(x_1, y_1)$

Lecture 4

Random vectors can be classified as discrete or continuous.

Note: To streamline notation, we will restrict ourselves to 2d (see refresher 0.5 for the n dimensional)

Discrete random vectors

A discrete random vector is described by a PMF $p_{X,Y}$ where $p_{X,Y}(x, y) = \mathbb{P}[X = x, Y = y]$. From this, we can derive:

- The **marginal** probabilities:

$$p_X(x) = \sum_{\mathbf{y}} p_{X,Y}(x, \mathbf{y}) \tag{25}$$

$$p_Y(y) = \sum_{\mathbf{x}} p_{X,Y}(\mathbf{x}, y) \tag{26}$$

Note: the joint distribution is the multiplication of the two marginals if and only if X and Y are independent. For this reason, in general, from the marginals we usually can't go back to the joint

- The **conditional** probabilities:

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \quad \text{if } p_Y(y) > 0 \tag{27}$$

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \quad \text{if } p_X(x) > 0 \tag{28}$$

- **Moments of a function**: given a function $g : \mathbb{R}^2 \to \mathbb{R}$, we can find the expected value and that is

$$\mathbb{E}g(x,y) = \sum_{x,y} g(x,y) p_{X,Y}(x,y) \tag{29}$$

From the conditional probabilities we can find the expected value of X given Y and Y given X, and the definitions are:

$$\mathbb{E}[X \mid Y = y] = \sum_{x} x p_{X|Y}(x,y) \tag{30}$$

$$\mathbb{E}[Y \mid X = x] = \sum_{y} y p_{Y|X}(x,y) \tag{31}$$

The expected value of Y given X, denoted as $\mathbb{E}[Y \mid X]$, represents the optimal predictor of Y as a function of X. This optimality is defined in terms of minimizing the mean squared error (MSE) between the actual values of Y and the predictions from any function $g(X)$. Mathematically, the criterion for optimality is expressed as minimizing $\mathbb{E}\left[(Y - g(X))^2 \mid X\right]$, where $g(X)$ is the function that achieves this minimum.

Note: $\mathbb{E}[Y \mid X]$ is the regressor function

Continuous random vectors

A continuous random vector is described by a PDF $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ with the following properties

- $f_{X,Y}(x,y) \geq 0, \forall x, y$
- $\iint_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\, dx = 1$

- $\displaystyle\int_a^b \int_c^d f_{X,Y}(x,y)\,dy\,dx = \mathbb{P}[a < X \leq b, c < Y \leq d]$
- $f_{X,Y}(x,y) = \mathbb{P}[X \in [x+dx], Y \in [y+dy]]$, from the mean value theorem (refresher 0.6)

From this function we can derive

- The **CMF** which is given by

  $$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)\,dt\,ds \tag{32}$$

  At the same time, we can go back from the CMF to the PDF by taking the second partial derivative, which means that

  $$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) \tag{33}$$

- The **conditional** probability is given by

  $$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad \text{if } f_Y(y) > 0 \tag{34}$$

  $$f_{Y|X}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_X(x)} \qquad \text{if } f_X(x) > 0 \tag{35}$$

- The **moments** of a function: given a function $g : \mathbb{R}^2 \to \mathbb{R}$, we can find the expected value and that is

  $$\mathbb{E}g(x,y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dy\,dx \tag{36}$$

From the conditional probabilities we can find the expected value of X given Y and Y given X, and the definitions are:

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x,y) dx \tag{37}$$

$$\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(x,y) dy \tag{38}$$

Lecture 5

Joint and Marginal Distributions: Although we can recover from the joint the marginals, the same can't be said for the other way around, unless X and Y are independent. However, if f_X and $f_{Y|X}$ are known, then

$$f_{X,Y}(x,y) = f_{Y|X}(y \mid x) f_X(x). \tag{39}$$

Covariance and Correlation

Covariance in a Joint Distribution: Covariance in a joint distribution is given by:

$$\mathrm{Cov}(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \mathbb{E}XY - (\mathbb{E}X)(\mathbb{E}Y). \tag{40}$$

With $\mathrm{Cov}(X,Y) > 0$ then an increase in X will lead to an increase in Y, and with $\mathrm{Cov}(X,Y) < 0$ an increase in X will lead to a decrease Y.

Properties of Covariance:

1. $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X,Y)$
2. $\mathrm{Cov}(X,Y) \leq \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}$
3. $\mathrm{Var}(aX, bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X,Y)$

From 2 we also get that

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \tag{41}$$

With $-1 \leq \rho_{X,Y} \leq 1$. The two extreme cases happen if and only if $\mathbb{P}[Y = a + bX] = 1$ for some $b > 0$ for $\rho_{X,Y} = 1$ and $b < 0$ for $\rho_{X,Y} = -1$.

If $\rho_{X,Y} = 0$, then we can only say they are uncorrelated, but this doesn't mean they are independent. The only case in which this happens is if X and Y are normally distributed.

$$\rho_{aX,bY} = \frac{ab\mathrm{Cov}(X,Y)}{\sqrt{a^2\mathrm{Var}(X)}\sqrt{b^2\mathrm{Var}(Y)}} = \frac{ab\mathrm{Cov}(X,Y)}{|a| \cdot |b|\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \mathrm{sign}(ab)\rho_{X,Y}. \tag{42}$$

Properties Related to Conditioning

Expectation:

$$\mathbb{E}X = \mathbb{E}\mathbb{E}(X \mid Y). \tag{43}$$

Proof:

Assuming X and Y linear,

$$\mathbb{E}\mathbb{E}(X \mid Y) = \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y)\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,dx\, f_Y(y)\,dy \tag{44}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) f_Y(y)\,dy\,dx = \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\right) dx \tag{45}$$

$$= \int_{-\infty}^{\infty} x f_X(x)\,dx = \mathbb{E}X. \tag{46}$$

We are able to switch the order in which we integrate by using Fubini's Theorem that states $\int_X \int_Y g(x,y)\,dy\,dx = \int_Y \int_X g(x,y)\,dx\,dy$ if $g(x,y) \geq 0\,\forall x,y$.

Variance of X: The variance of X is given by:

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X \mid Y)] + \text{Var}(\mathbb{E}[X \mid Y]). \tag{47}$$

Starting with the definition of variance, we have:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]. \tag{48}$$

Using the law of iterated expectations, we expand this as:

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X \mid Y] + \mathbb{E}[X \mid Y] - \mathbb{E}[X])^2] && (49)\\ &= \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2] + 2\mathbb{E}[(X - \mathbb{E}[X \mid Y])(\mathbb{E}[X \mid Y] - \mathbb{E}[X])] + \mathbb{E}[(\mathbb{E}[X \mid Y] - \mathbb{E}[X])^2]. && (50)\end{aligned}$$

The second term becomes zero, and we are left with:

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X \mid Y)] + \text{Var}(\mathbb{E}[X \mid Y]). \tag{51}$$

Lecture 6

Independence

Definition and Properties

Independence of Random Variables: Random variables X and Y are said to be independent, denoted as $X \perp Y$, if for any $A, B \subset \mathbb{R}$,

$$\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[X \in A]\mathbb{P}[Y \in B]. \tag{52}$$

This implies that the occurrence of events in X does not affect the probability of events in Y and vice versa.

Characteristic Forms of Independence:

- For continuous random variables:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \forall x, y \in \mathbb{R}. \tag{53}$$

- For discrete random variables:

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \quad \forall x, y \in \mathbb{R}. \tag{54}$$

- In general, X and Y are independent if and only if their joint distribution function factors into the product of their marginal distribution functions:

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \forall x, y \in \mathbb{R}. \tag{55}$$

Independence in transformations of Random Variables

Note:This section is highly technical and I think not necessary to know to this depth

Given two independent random variables X and Y with values in $\mathbb{R}$, and functions $g, h : \mathbb{R} \to \mathbb{R}$, we define $S = g(X)$ and $T = h(Y)$. We aim to prove that S and T are independent, which, by definition, means showing that for any Borel sets C and D in $\mathbb{R}$,

$$P(S \in C, T \in D) = P(S \in C)P(T \in D). \tag{56}$$

Two random variables X and Y are independent if and only if for any Borel sets (see refresher 0.7) A and B,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B). \tag{57}$$

Given $S = g(X)$ and $T = h(Y)$, we express the events $S \in C$ and $T \in D$ in terms of X and Y:

$$\begin{aligned} S \in C &\Leftrightarrow X \in g^{-1}(C), \\ T \in D &\Leftrightarrow Y \in h^{-1}(D), \end{aligned}$$

where $g^{-1}(C)$ and $h^{-1}(D)$ are the pre-images (see refresher 0.8) of C and D under g and h, respectively, which are also Borel sets due to the measurability of g and h.

Since X and Y are independent,

$$P(X \in g^{-1}(C), Y \in h^{-1}(D)) = P(X \in g^{-1}(C))P(Y \in h^{-1}(D)). \tag{58}$$

Substituting the equivalences into the probability expression, we get

$$P(S \in C, T \in D) = P(X \in g^{-1}(C), Y \in h^{-1}(D)) = P(S \in C)P(T \in D). \tag{59}$$

Special Cases A random variable X degenerate at a point $c \in \mathbb{R}$ (i.e., $\mathbb{P}[X = c] = 1$) is independent of any other random variable.

Applications in Statistics

- **Independently and Identically Distributed (i.i.d.) Variables:** Assuming $X_1, \cdots, X_n \sim \mathcal{P}$ implies that they are independent and identically distributed. For such variables:

  - In the discrete case:

  $$p_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i). \tag{60}$$

  - In the continuous case:

  $$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i). \tag{61}$$

- **Variance of a Linear Combination:** If $X_1, \cdots, X_n$ are independent with constants $a_1, \cdots, a_n \in \mathbb{R}$, then:

  $$\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i). \tag{62}$$

The second point is necessary for the Central Limit Theorem (CLT). The CLT states that as the number of i.i.d. random variables increases, the distribution of their sum (or average) tends towards a normal distribution, irrespective of the variables' original distribution.

Consider $X_1, X_2, \ldots, X_n$ as a sequence of i.i.d. random variables with a common mean μ and variance σ^2. The sample mean, denoted $\bar{X}_n$, is given by:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{63}$$

To understand the distribution of $\bar{X}_n$, we examine its variance. Setting $a_1 = \cdots = a_n = \frac{1}{n}$, we apply the formula for the variance of a linear combination of independent random variables:

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\frac{1}{n^2}\text{Var}(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}. \tag{64}$$

Moment Generating Functions and Independence

If $X \perp Y$ with moment generating functions $m_X(t)$ and $m_Y(t)$ respectively, then for $X+Y$, the moment generating function is:

$$m_{X+Y}(t) = m_X(t)m_Y(t). \tag{65}$$

For i.i.d. variables $X_1, \ldots, X_n$, the MGF of their sum is the product of their individual MGFs, which simplifies to $(m_X(t))^n$.

Multivariate Normal Distribution

Consider a random vector $\mathbf{X} = (X_1, X_2)^\top$ that follows a multivariate (bi-variate in the 2-dimensional case) normal distribution. The joint density function of $\mathbf{X}$, given mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ and covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, is defined as:

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\} \tag{66}$$

where ρ is the correlation coefficient between X_1 and X_2, and σ_1, σ_2 are the standard deviations of X_1 and X_2, respectively.

Properties:

1. **Linear Transformations:** A linear transformation of $\mathbf{X}$, say $Y = aX_1 + bX_2$, where a and b are constants, will also follow a normal distribution. The mean and variance of Y are given by $\mathbb{E}[Y] = a\mu_{X_1} + b\mu_{X_2}$ and $\text{Var}(Y) = a^2\sigma_{X_1}^2 + b^2\sigma_{X_2}^2 + 2ab\rho\sigma_{X_1}\sigma_{X_2}$, respectively.

2. **Marginal Distributions:** The marginal distribution of X_1, obtained by integrating the joint density over X_2, is:

   $$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_{X_1}} \exp\left(-\frac{(x_1 - \mu_{X_1})^2}{2\sigma_{X_1}^2}\right). \tag{67}$$

   Similarly, the marginal distribution of X_2 is:

   $$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}\sigma_{X_2}} \exp\left(-\frac{(x_2 - \mu_{X_2})^2}{2\sigma_{X_2}^2}\right). \tag{68}$$

   The expected values are $\mathbb{E}[X_1] = \mu_{X_1}$ and $\mathbb{E}[X_2] = \mu_{X_2}$, respectively.

3. **Conditional Distributions:** The conditional distribution of X_1 given $X_2 = x_2$ is normally distributed with:

   $$\mathbb{E}[X_1 \mid X_2 = x_2] = \mu_{X_1} + \rho\frac{\sigma_{X_1}}{\sigma_{X_2}}(x_2 - \mu_{X_2}), \tag{69}$$

   and variance $\text{Var}(X_1 \mid X_2) = \sigma_{X_1}^2(1 - \rho^2)$. The formula demonstrates how knowledge of X_2's value adjusts the expected value of X_1.

4. **Independence and Covariance:** For X_1 and X_2 to be independent, the covariance (and thus the correlation ρ) must be zero. In this case, the joint density function simplifies, indicating that knowledge of one variable does not affect the distribution of the other. Independence implies that the joint density function factors into the product of the marginal densities.

Lecture 7

Multinomial Distribution

The multinomial distribution generalizes the binomial distribution to scenarios with more than two possible outcomes. Specifically, while the binomial distribution concerns binary outcomes (success or failure), the multinomial distribution applies to experiments with k possible outcomes, where $k \geq 2$.

For an experiment with n trials, let X_i denote the number of occurrences of the i-th outcome, for $i = 1, 2, \ldots, k$. The total number of occurrences across all outcomes is $\displaystyle\sum_{i=1}^{k} X_i = n$, implying $X_i = n - \displaystyle\sum_{j \neq i} X_j$ for each i.

The probability of observing a specific outcome pattern is defined by a probability vector $\mathbf{p} = (p_1, p_2, \ldots, p_k)$, where p_i represents the probability of the i-th outcome, subject to the conditions $p_i \geq 0$ for all i, and $\displaystyle\sum_{i=1}^{k} p_i = 1$.

A random vector $\mathbf{X} \sim M_k(n; p_1, \ldots, p_k)$, with $\mathbf{X} = (X_1, \ldots, X_k)$, is said to follow a multinomial distribution if its probability mass function (PMF) is given by:

$$p_{X_1, \cdots, X_k}(x_1, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \tag{70}$$

The marginal distributions of each X_i are binomial, with $X_i \sim \text{Binom}(n; p_i)$. This implies that the expected value of X_i is $\mathbb{E}[X_i] = np_i$, and the covariance between any two different components X_i and X_j (for $i \neq j$) is $\text{Cov}(X_i, X_j) = -np_i p_j$.

Transformations of Random Variables and Random Vectors

We detail three primary methods for determining the distribution of $Y = g(X)$, where X is a random variable, and g is a transformation function.

Method 1: Moment Generating Function (MGF) Method

The MGF of a random variable Y, when $Y = g(X)$, is given by:

$$m_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}[e^{tg(X)}] = \begin{cases} \displaystyle\sum_{x} e^{tg(x)} p_X(x) & \text{if } X \text{ is discrete} \\ \displaystyle\int_{-\infty}^{\infty} e^{tg(x)} f_X(x)\, dx & \text{if } X \text{ is continuous} \end{cases} \tag{71}$$

Method 2: Cumulative Distribution Function (CDF) Method

The CDF method relates directly to the transformation of probabilities:

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[g^{-1}(X) \leq y] \tag{72}$$

For continuous X with a continuous and strictly increasing CDF F_X, the transformation $Y = F_X(x)$ yields $Y \sim \text{Unif}(0, 1)$. The proof is based on calculating $P(Y \leq y)$

and demonstrating it equals y for $y \in (0,1)$, which matches the CDF of a uniform distribution.

Remark: This result extends beyond strictly increasing F_X. When F_X is not strictly increasing, one uses the generalized inverse $F_X^{-1}(y) = \inf\{x \in \mathbb{R} : F_X(x) \geq y\}$ (more on refresher 0.10).

Method 3: Change of Variable Technique

Univariate Case: For $X \sim f_X$ and a differentiable, bijective function g, with $Y = g(X)$, the density of Y is:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \tag{73}$$

Multivariate Case: The extension to multivariate transformations involves the Jacobian determinant of the transformation. For a random vector $X \sim f_X$ and a transformation $g : \mathbb{R}^n \to \mathbb{R}^n$ that is differentiable and bijective, the density of $Y = g(X)$ is given by:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \det J_{g^{-1}}(y) \right| \tag{74}$$

where $J_{g^{-1}}(y)$ is the Jacobian matrix of the inverse transformation.

Procedure:

1. Determine the inverse transformation g^{-1}.
2. Compute the Jacobian matrix $J_{g^{-1}}(y)$, which is

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{\partial g_1^{-1}}{\partial y_1} & \cdots & \frac{\partial g_1^{-1}}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n^{-1}}{\partial y_1} & \cdots & \frac{\partial g_n^{-1}}{\partial y_n} \end{bmatrix} \tag{75}$$

3. Calculate the determinant $|\det J_{g^{-1}}(y)|$.
4. Apply these to find $f_Y(y) = f_X(g^{-1}(y)) \left| \det J_{g^{-1}}(y) \right|$.

Lecture 8

We did only exercises, suggested revising the distributions found at the start of the refreshers

Lecture 9

Box-Müller Transformation

The Box-Muller transformation is a method for generating independent, standard, normally distributed random variables given two independent uniformly distributed random variables. Let's denote these uniformly distributed variables as U_1 and U_2, where $U_1, U_2 \sim \text{Uniform}(0, 1)$.

The transformation consists of the following steps to produce two independent standard normal variables, Z_0 and Z_1:

1. Compute two intermediate values based on U_1 and U_2:

$$R^2 = -2\log(U_1) \tag{76}$$

$$\Theta = 2\pi U_2 \tag{77}$$

2. Apply the Box-Muller transformation to obtain Z_0 and Z_1:

$$Z_0 = R\cos(\Theta) = \sqrt{-2\log(U_1)}\cos(2\pi U_2) \tag{78}$$

$$Z_1 = R\sin(\Theta) = \sqrt{-2\log(U_1)}\sin(2\pi U_2) \tag{79}$$

The Box-Muller transformation's use of R^2 and Θ directly relates to the concept of polar coordinates, which offer a different method for representing points in a plane. In polar coordinates, a point's position is determined by its distance from the origin and the angle formed with a reference direction.

In the context of the Box-Muller transformation:

- R represents the radius (distance from the origin) in polar coordinates. It is defined as $R = \sqrt{-2\log(U_1)}$, where U_1 is a uniformly distributed random variable. The square of the radius, $R^2 = -2\log(U_1)$, is used to ensure that the distribution of R is such that when converted back to Cartesian coordinates (using Z_0 and Z_1), it produces a normal distribution.

- Θ represents the angle in polar coordinates, defined as $\Theta = 2\pi U_2$, where U_2 is another independent uniformly distributed random variable. This angle is uniformly distributed between 0 and 2π, which ensures that the direction of the generated point is random and uniformly distributed around the circle.

This method exploits the fact that if (Z_0, Z_1) are independent standard normal random variables, their squared distance from the origin, $Z_0^2 + Z_1^2$, follows the exponential distribution when $R^2 = -2\log(U_1)$.

Properties:

- The resulting variables Z_0 and Z_1 are independent.

- Both Z_0 and Z_1 follow a standard normal distribution, i.e., $Z_0, Z_1 \sim N(0, 1)$.

Order Statistics

Order statistics provide a way to arrange random variables in ascending or descending order. Given a sample of n iid continuous random variables $X_1, X_2, \ldots, X_n$ from a distribution $F(x)$, the kth order statistic, denoted $X_{(k)}$, is the kth smallest value in the sample.

Distribution of a Single Order Statistic

The distribution function of the kth order statistic, $X_{(k)}$, can be derived using the properties of continuous iid random variables. Let's denote the probability density function (pdf) of X_i as $f(x)$ and the cumulative distribution function (cdf) as $F(x)$.

The probability that $X_{(k)}$ is less than or equal to x is equivalent to the probability that at least k out of the n observations fall at or below x. This can be expressed using the binomial distribution:

$$F_{X_{(k)}}(x) = P(X_{(k)} \leq x) = \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j} \tag{80}$$

The corresponding pdf is found by differentiating the cdf with respect to x:

$$f_{X_{(k)}}(x) = \frac{d}{dx} F_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x) \tag{81}$$

For the smallest order statistic $X_{(1)}$, and the largest order statistic $X_{(n)}$, the density functions are given by:

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - [1 - F(x)]^n. \\ F_{X_{(n)}}(x) &= [F(x)]^n. \\ f_{X_{(1)}}(x) &= n[1 - F(x)]^{n-1} f(x), \\ f_{X_{(n)}}(x) &= n[F(x)]^{n-1} f(x). \end{aligned}$$

Joint Distribution of Two Order Statistics

The joint distribution of two order statistics, $X_{(j)}$ and $X_{(k)}$ for $j < k$, can also be derived. It accounts for the probability that $X_{(j)}$ falls in one interval and $X_{(k)}$ falls in another, without any other values in between.

$$f_{X_{(j)}, X_{(k)})}(x, y) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!} [F(x)]^{j-1} [F(y) - F(x)]^{k-j-1} [1 - F(y)]^{n-k} f(x) f(y) \tag{82}$$

for $x < y$.

The joint density function for the smallest and largest order statistics, $X_{(1)}$ and $X_{(n)}$, is particularly interesting as it captures the spread of the entire sample. It is given by:

$$f_{X_{(1)},X_{(n)}}(x_1, x_n) = n(n-1)[F(x_n) - F(x_1)]^{n-2} f(x_1) f(x_n),$$

for $x_1 < x_n$.

Properties and Applications

- **Minimum and Maximum**: The first and last order statistics, $X_{(1)}$ and $X_{(n)}$, represent the minimum and maximum values in the sample, respectively.
- **Medians and Percentiles**: Middle order statistics can serve as empirical medians or other percentiles, depending on their position.
- **Range and Spacing**: Differences between successive order statistics provide information about the sample's dispersion.

Lecture 10

Markov Inequality

Markov's Inequality provides a way to bound the probability that a non-negative random variable X is at least some positive value a. Formally, for any $a > 0$, the inequality is given by:

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \tag{83}$$

Proof

Assume X is a continuous random variable with a probability density function $f(x)$. The expected value of X is defined as:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx \tag{84}$$

Considering X is non-negative, we modify the lower bound of integration:

$$\mathbb{E}[X] = \int_{0}^{\infty} x f(x) dx \tag{85}$$

To establish Markov's inequality, we consider the integral from a to infinity:

$$\mathbb{E}[X] \geq \int_a^\infty xf(x)dx \geq \int_a^\infty af(x)dx = a\int_a^\infty f(x)dx \tag{86}$$

Here, $\int_a^\infty f(x)dx$ represents the probability $P(X \geq a)$. Thus, we derive:

$$\mathbb{E}[X] \geq aP(X \geq a) \tag{87}$$

Rearranging gives Markov's inequality:

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \tag{88}$$

Special Cases

1. **Chebyshev's Inequality**: Let X be a random variable with mean μ and variance σ^2. For any $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \tag{89}$$

2. **Jensen's Inequality**: Let X be a random variable and ϕ a convex function. Then,

$$\phi(E[X]) \leq E[\phi(X)]. \tag{90}$$

3. **The Chernoff Bounds**: Let $X_1, X_2, \ldots, X_n$ be independent random variables. For any $t > 0$,

$$P\left(X \geq a\right) \leq \inf_{t>0} \frac{\mathbb{E}e^{tX}}{e^{ta}}. \tag{91}$$

For $t < 0$, then

$$P\left(X \leq a\right) \leq \inf_{t<0} \frac{\mathbb{E}e^{tX}}{e^{ta}}. \tag{92}$$

Lecture 11

Convergence of Random Variables

Convergence in Probability

A sequence of random variables $\{X_n\}$ converges in probability towards the random variable X if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0. \tag{93}$$

This type of convergence is denoted as $X_n \xrightarrow{p} X$ and signifies that the probability of X_n deviating from X by more than ϵ becomes arbitrarily small as n increases.

Almost Sure Convergence

A sequence $\{X_n\}$ converges almost surely (a.s.) to X if

$$P(\lim_{n\to\infty} X_n = X) = 1. \tag{94}$$

Denoted by $X_n \xrightarrow{a.s.} X$, it means that the events where X_n does not converge to X occur with zero probability. This is a stronger form of convergence than convergence in probability, implying that almost every sequence realization converges to X.

Convergence in Distribution

A sequence $\{X_n\}$ converges in distribution to X if for all t at which the cumulative distribution function (CDF) of X, $F_X(t)$, is continuous,

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t). \tag{95}$$

This is denoted as $X_n \xrightarrow{d} X$ and concerns the convergence of the distribution functions rather than the random variables themselves.

Convergence in Quadratic Mean

A sequence $\{X_n\}$ converges in quadratic mean to X if

$$\lim_{n\to\infty} \mathbb{E}[(X_n - X)^2] = 0. \tag{96}$$

This type of convergence, denoted as $X_n \xrightarrow{qm} X$, requires that the mean squared difference between X_n and X goes to zero as n approaches infinity. It implies convergence in probability and is particularly useful for analyzing the properties of estimators in statistics.

Convergence in Distribution

A sequence of random variables X_n converges in distribution to a random variable X if for every continuity point x of $F_X(x)$, the CDF of X,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x). \tag{97}$$

Comparison and Examples

- **Convergence in probability** is useful for large-sample properties of estimators but does not guarantee that a particular sequence realization will converge.
- **Almost sure convergence** is a stronger guarantee than convergence in probability, ensuring that almost every realization of the sequence converges to the limit.
- **Convergence in distribution** is essential for understanding the limiting behavior of distributions, particularly in the context of the Central Limit Theorem.
- **Convergence in quadratic mean** is powerful for statistical inference, ensuring not just convergence in probability but also that the average squared discrepancies converge to zero.

Laws of Large Numbers

Weak Law of Large Numbers (WLLN)

The WLLN states that for a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n$ with expected value $\mathbb{E}[X_i] = \mu$, the sample average converges in probability towards the population mean. Formally, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0. \tag{98}$$

Strong Law of Large Numbers (SLLN)

The SLLN asserts that the sample averages converge almost surely (with probability 1) to the expected value, that is,

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1. \tag{99}$$

Lecture 12

The Central Limit Theorem (CLT)

Statement

Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Define the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, as $n \to \infty$, the standardized sample mean $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converges in distribution to a standard normal distribution, that is,

$$Z_n \xrightarrow{d} Z \sim N(0,1). \tag{100}$$

Proof

For i.i.d. random variables $X_1, \ldots, X_n$ with MGF $m_X(t)$, the MGF of their sum $S_n = \sum_{i=1}^{n} X_i$ is given by $m_{S_n}(t) = [m_X(t)]^n$.

The MGF of the standardized variable Z_n is derived by considering the transformation applied to S_n, resulting in:

$$m_{Z_n}(t) = e^{-\mu t\sqrt{n}/\sigma} \left[m_X\left(\frac{t}{\sigma\sqrt{n}} \right) \right]^n. \tag{101}$$

Given X_i with mean μ and variance σ^2, the MGF $M_X(t)$ near $t = 0$ can be expanded as:

$$m_X(t) = 1 + \mu t + \frac{\sigma^2 t^2}{2} + o(t^2), \tag{102}$$

where $o(t^2)$ represents terms of higher order that become negligible as t approaches 0.

Using the expansion of $m_X(t)$ and considering the limit as n approaches infinity, we find that $m_{Z_n}(t)$ converges to the MGF of a standard normal distribution:

$$m_{Z_n}(t) \to e^{t^2/2}, \tag{103}$$

demonstrating that Z_n converges in distribution to $N(0,1)$.

Lecture 13

Simulating a Random Variable with a known CDF

Given a random variable X with a known CDF $F_X(x)$, the Inverse Transform Sampling method involves the following steps:

1. Find the CDF $F_X(x)$ of the random variable X.
2. Compute the inverse of the CDF, denoted as F_X^{-1}.
3. Generate a uniform random variable $u \sim Unif(0,1)$.
4. Return the value $X = F_X^{-1}(u)$.

This method is based on the principle that if u is a uniform random variable on the interval $(0, 1)$, then the variable $X = F_X^{-1}(u)$ will have the distribution F_X.

Below is a pseudocode representation of the Inverse Transform Sampling method:

```
# Pseudocode for simulating a random variable X
# using Inverse Transform Sampling

def inverse_transform_sampling(F_inv):
    # Generate a uniform random number u from 0 to 1
    u = random.uniform(0, 1)

    # Compute the inverse CDF value
    X = F_inv(u)

    return X

# Example usage
# Define the inverse CDF function F_inv for the target distribution
# Call inverse_transform_sampling(F_inv) to simulate a random variable
```

Listing 1: Inverse Transform Sampling Pseudocode

Simulating a random variable without a CDF

While some distributions are straightforward to sample from, others require more sophisticated techniques. Two widely used methods for generating random variables are the Box-Muller method for normal distributions and the Acceptance-Rejection method for sampling from more complex distributions.

Box-Muller Method

The Box-Muller method is a procedure for generating pairs of independent, standard, normally distributed (zero mean, unit variance) random variables from two independent uniform random variables.

Given two independent random variables U_1 and U_2 uniformly distributed over $(0, 1)$, two independent standard normally distributed random variables Z_0 and Z_1 can be generated as follows:

$$Z_0 = \sqrt{-2 \ln U_1} \cos(2\pi U_2) \tag{104}$$

$$Z_1 = \sqrt{-2 \ln U_1} \sin(2\pi U_2) \tag{105}$$

This method utilizes the polar coordinates system to transform uniform random variables into normally distributed variables.

Below is a pseudocode representation of the Box-Muller Method

```
import math
import random

def box_muller():
    U1 = random.random()  # Generate U1, U2 ~ Unif(0, 1)
    U2 = random.random()
    Z0 = math.sqrt(-2 * math.log(U1)) * math.cos(2 * math.pi * U2)  # Standard normal Z0
    Z1 = math.sqrt(-2 * math.log(U1)) * math.sin(2 * math.pi * U2)  # Standard normal Z1
    return Z0, Z1
```

Listing 2: Box-Muller Method Pseudocode

Acceptance-Rejection Method

The Acceptance-Rejection method allows sampling from a distribution $f(x)$ by utilizing a simpler proposal distribution $g(x)$ from which we can readily sample. The method is based on finding an envelope of $g(x)$ that covers $f(x)$.

To simulate a random variable with density $f(x)$:

1. Choose a proposal distribution $g(x)$ such that there exists a constant c where $cf(x) \geq g(x)$ for all x.

2. Generate a candidate X from $g(x)$.

3. Generate a uniform random variable U on $(0, 1)$.

4. Accept X as a sample from $f(x)$ if $U \leq \frac{f(X)}{cg(X)}$; otherwise, reject X and return to step 2.

This method is efficient if the proposal distribution $g(x)$ closely resembles the target distribution $f(x)$, and the constant c is close to 1. The efficiency of the method depends on the choice of $g(x)$ and the acceptance rate, which ideally should be high to minimize computational waste.

Below is a pseudocode representation of the Acceptance-Rejection Method

```
import random

def acceptance_rejection(f, g, c):
    while True:
        X = sample_from_g()  # Generate candidate X from proposal distribution g
        U = random.random()  # Uniform random number U

        # Acceptance condition
        if U <= f(X) / (c * g(X)):
            return X  # Accept X as sample from f

# f: Target density function
# g: Proposal density function (from which we can sample directly)
# c: Constant such that c*g(x) >= f(x) for all x
# sample_from_g: Function to sample from g
```

Listing 3: Acceptance-Rejection Method Pseudocode

Lecture 14 - 15

We did only exercises, suggested revising everything done up until now

Lecture 16

Parameter Space and Statistical Models

Assuming data $X_1, \dots, X_n$ are realizations of a random variable under identical conditions and are independent from one another, we represent this as:

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} F_\theta,$$

where F_θ is the distribution function parameterized by θ within a parameter space Θ. The parameter space Θ defines all possible values of the parameter θ, and the statistical model is a collection of probability density functions (pdfs) or probability mass functions (pmfs) defined as:

$$\{f(\cdot;\theta) : \theta \in \Theta\},$$

where the joint density (or mass) function for i.i.d. data is given by the product of individual densities (or masses):

$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Point Estimators

Point estimators are used to guess the value of a parameter θ from the data.

A point estimator $\hat{\theta}$ is a function mapping from the sample space to the parameter space:

$$\hat{\theta} : \mathbb{R}^n \to \Theta,$$

where $\hat{\theta}(x_1, \ldots, x_n)$ becomes a deterministic value after observing data.

An estimator $\hat{\theta}$ is unbiased for estimating θ if its expected value equals θ for all $\theta \in \Theta$, mathematically represented as:

$$\mathbb{E}_\theta[\hat{\theta}] = \theta, \quad \forall \theta \in \Theta.$$

Evaluating and comparing estimators is essential for selecting the most appropriate method for parameter estimation. MSE is a common measure used to evaluate the accuracy of an estimator (see Refresher 0.11 for more), and is defined as:

$$\text{MSE} = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

For an unbiased estimator, MSE simplifies to the variance of the estimator, as shown below:

$$\text{MSE} = \mathbb{E}(\hat{\theta} - \theta)^2 = \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 = \text{Var}(\hat{\theta})$$

Using MSE we can compare the efficiency of multiple estimators. An estimator $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2) \quad \forall \theta \in \Theta;$$

and for some $\theta \in \Theta$,

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2).$$

This may not always be applicable however, as it may be possible that the variance of an estimator is lower for some values and higher for others, as shown in the image below (location of the image may vary).

In these cases, we compare the relative efficiency of $\hat{\theta}_1$ and $\hat{\theta}_2$, defined as

$$\text{RE}(\hat{\theta}_1, \hat{\theta}_2) = \frac{\text{Var}(\hat{\theta}_1)}{\text{Var}(\hat{\theta}_2)} \tag{106}$$

and we say that $\hat{\theta}_1$ is at least as efficient as $\hat{\theta}_2$ at estimating θ if $\text{RE}(\hat{\theta}_1, \hat{\theta}_2) \leq 1 \ \forall \theta$.

Lecture 17

Unbiased Estimators and Their Existence

Finding an unbiased estimator for a parameter of interest is a fundamental task. An estimator $\hat{\theta}$ of a parameter θ is considered *unbiased* if its expected value equals the parameter itself, i.e., $\mathbb{E}[\hat{\theta}] = \theta$. This concept is crucial in ensuring that the estimation process does not systematically overestimate or underestimate the true parameter value.

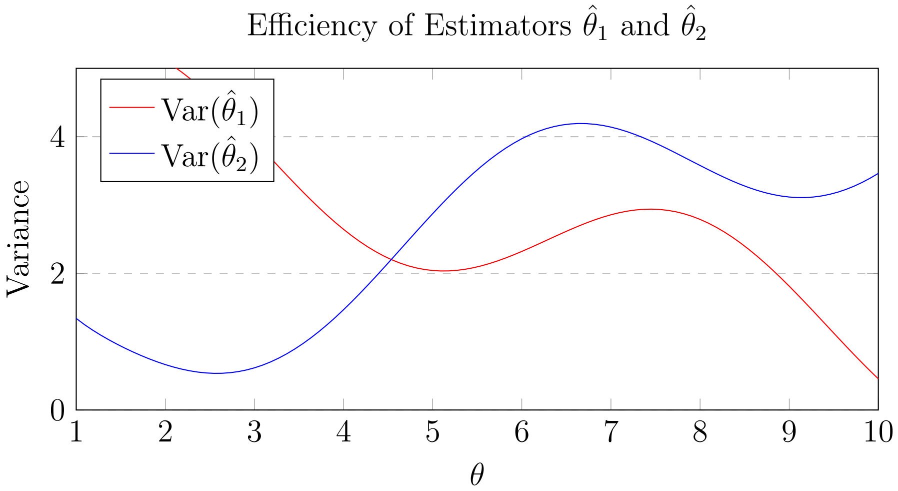
Example of an Unbiased Estimator

Consider a scenario where we are interested in estimating the probability $P_\theta[X \in A] = g(\theta)$ for a given set A and distribution P_θ. If we sample a single observation $X_1 \sim P_\theta$, the indicator function $\mathbb{1}_A(X_1)$ serves as an unbiased estimator for $P_\theta[X \in A]$. This is because:

$$\mathbb{E}[\mathbb{1}_A(X_1)] = \mathbb{P}_\theta[\mathbb{1}_A(X_1) = 1] = P_\theta[X_1 \in A]. \tag{107}$$

Non-Existence of an Unbiased Estimator

However, an unbiased estimator does not always exist for every statistical measure. An example is attempting to estimate p^2 for a Bernoulli distribution with parameter p. Suppose $X_1 \sim \text{Bern}(p)$; we wish to find a function $g(X_1)$ such that $\mathbb{E}[g(X_1)] = p^2$. It can be shown that:



$$\mathbb{E}[g(X_1)] = pg(1) + (1-p)g(0) \neq p^2 \quad \forall p \in [0,1]. \tag{108}$$

This equation cannot hold for all p within the interval $[0,1]$, implying that no function g can serve as an unbiased estimator for p^2. The contradiction arises from the inability of g's outputs for discrete inputs to satisfy a quadratic relationship across all p values, illustrating the limitations in finding unbiased estimators for certain parameters.

Fisher Information

Fisher information quantifies the amount of information that an observable random variable X, or a sample $(X_1, X_2, \ldots, X_n)$, carries about an unknown parameter θ upon which the probability of X or the joint probability of the sample depends.

There are two equivalent ways of defining Fisher information:

1. The Fisher information for a single observation X_1 from a probability distribution with parameter θ is defined as the expected value of the squared score function:

$$\mathcal{I}_{X_1}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X_1;\theta)\right)^2\right]$$

   where $f(X_1;\theta)$ is the probability density function (for continuous distributions) or probability mass function (for discrete distributions) of X_1, parameterized by θ.

2. Alternatively, Fisher information can also be defined using the double derivative of the log-likelihood function with respect to θ (for the proof see refresher 0.13):

$$\mathcal{I}_{X_1}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_1;\theta)\right]$$

We will primarily use the second definition involving the double derivative, as it often simplifies the computation and theoretical analysis.

Fisher Information for a Sample of Observations

For a sample of n independent and identically distributed (i.i.d.) observations, the Fisher information is the sum of the individual Fisher informations because the log-likelihood of the sample is the sum of the log-likelihoods of the individual observations:

$$\mathcal{I}_{X_n}(\theta) = n\mathcal{I}(\theta)$$

Conditions for Applicability

The application of Fisher information and the second definition, in particular, require certain conditions: (Refresher 0.12)

- **Differentiability**: The log-likelihood function must be sufficiently smooth, allowing for the existence and continuity of the first and second derivatives with respect to θ.
- **Regularity Conditions**: Several regularity conditions must be met to ensure that operations such as differentiation under the integral sign are valid and to guarantee the existence of the expected values involved in the definitions.

Lecture 18

The Cauchy-Schwarz Inequality

The Cauchy-Schwarz inequality finds significant applications in statistics, particularly in deriving properties of estimators and in proving the optimality of certain estimators under specific conditions.

The Cauchy-Schwarz inequality can be stated as follows:

Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be functions such that:

$$\int_{\mathbb{R}^n} f^2(x)\, dx < \infty \qquad \text{and} \qquad \int_{\mathbb{R}^n} g^2(x)\, dx < \infty, \tag{109}$$

then,

$$\left(\int_{\mathbb{R}^n} g(x) f(x)\, dx \right)^2 \leq \left(\int_{\mathbb{R}^n} f^2(x)\, dx \right) \left(\int_{\mathbb{R}^n} g^2(x)\, dx \right). \tag{110}$$

This inequality is significant for the derivation of the Cramér-Rao Lower Bound, which sets a lower bound on the variance of unbiased estimators.

Regular Statistical Models

Definition: A statistical model, $\{f(\cdot;\theta) : \theta \in \Theta\}$, is considered regular if it satisfies the following conditions:

- The support of $f(\cdot;\theta)$, denoted as $A = \{x : f(x,\theta) > 0\}$, is independent of θ.
- The function $f(x;\theta)$ is twice differentiable for any $x \in \mathbb{R}^n$.

- For any sample statistic $T : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{E}|T(X_1, \ldots, X_n)| < \infty$, one has that:

$$\frac{\partial}{\partial\theta} \int_{\mathbb{R}^n} T(x) f(x;\theta) dx = \int_{\mathbb{R}^n} T(x) \frac{\partial}{\partial\theta} f(x;\theta) dx, \tag{111}$$

$$\frac{\partial^2}{\partial\theta^2} \int_{\mathbb{R}^n} T(x) f(x;\theta) dx = \int_{\mathbb{R}^n} T(x) \frac{\partial^2}{\partial\theta^2} f(x;\theta) dx. \tag{112}$$

The Cramér-Rao Lower Bound

Theorem: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(\cdot;\theta)$ within a regular statistical model $\{f(\cdot;\theta) : \theta \in \Theta\}$. If T is an unbiased estimator of θ, that is, $\mathbb{E}T = \theta$ for any $\theta \in \Theta$, then:

$$\operatorname{Var}(T) \geq \frac{1}{n\mathcal{I}_{X_1}(\theta)}, \tag{113}$$

where $\mathcal{I}_{X_1}(\theta)$ denotes the Fisher information in a single observation. This inequality is known as the Cramér-Rao Lower Bound (CRLB).

If there exists a statistic T that achieves the CRLB, it is considered optimal and is known as a Minimum Variance Unbiased Estimator (MVUE).

Sufficiency and the Fisher-Neyman Theorem

Definition: A sample statistic $T = T(X_1, \ldots, X_n)$ is sufficient for estimating θ if the conditional distribution of $(X_1, \ldots, X_n)$ given $T = t$ does not depend on θ; in other words, for all $\theta \in \Theta$:

$$f_\theta(x_1, \ldots, x_n \mid t) = f(x_1, \ldots, x_n \mid t). \tag{114}$$

Theorem(Fisher-Neyman), also known as Factorization Theorem:

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(\cdot;\theta)$ for $\theta \in \Theta \subset \mathbb{R}^d$. A statistic T is sufficient for θ if and only if there exist functions $\nu : \mathbb{R}^d \times \Theta \to \mathbb{R}^+$ and $W : \mathbb{R}^n \to \mathbb{R}^+$ such that for all $x_1, \ldots, x_n$ in the sample space:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i;\theta) = \nu(t, \theta) W(x_1, \ldots, x_n), \tag{115}$$

where $t = T(x_1, \ldots, x_n)$, ν depends on the observations only through t, and W does not depend on θ.

Lecture 19

One-parameter Exponential Families

A probability density or mass function, $f(\cdot;\theta)$, belongs to a one-parameter exponential family of distributions if the set $\Omega = \{x : f(x;\theta) > 0\}$ does not depend on θ and if there exist functions:

- $\mathcal{A} : \Theta \to \mathbb{R}$;
- $\mathcal{B} : \mathbb{R} \to \mathbb{R}$;
- $\mathcal{C} : \mathbb{R} \to \mathbb{R}$;
- $\mathcal{D} : \Theta \to \mathbb{R}$;

such that the density or mass function can be expressed as:

$$f(x;\theta) = \exp\left\{\mathcal{A}(\theta)\mathcal{B}(x) + \mathcal{C}(x) + \mathcal{D}(\theta)\right\} \mathbb{1}_{\Omega}(x), \tag{116}$$

Writing the function in this form simplifies the identification of ν and W for determining a sufficient statistic.

The Rao-Blackwell Theorem

Theorem: Given an estimator T of a parameter θ that is unbiased, i.e., $\mathbb{E}[T] = \theta$, and another statistic S that is sufficient for θ, the Rao-Blackwell theorem states that the conditional expectation of T given S, denoted by $\mathbb{E}[T \mid S]$, is also an unbiased estimator of θ. Moreover, $\mathbb{E}[T \mid S]$ is at least as good as T in terms of the mean squared error (MSE), meaning:

$$\mathbb{E}\left[(\mathbb{E}[T \mid S] - \theta)^2\right] \le \mathbb{E}\left[(T - \theta)^2\right], \tag{117}$$

with equality if and only if T is a function of S.

Proof: Given the unbiasedness of T and applying the law of total expectation and total variance, we establish that $\mathbb{E}[T \mid S]$ maintains unbiasedness and achieves a lower or equal MSE compared to T.

Completeness of a Statistic

Definition of Completeness: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(\cdot;\theta)$ for $\theta \in \Theta$. A sample statistic $T = T(X_1, \ldots, X_n)$ is complete for θ if, for any function $g : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[g(T)] = 0$, it holds that $\mathbb{P}[g(T) = 0] = 1$.

Lecture 20

Lehmann and Scheffè Theorem

The Rao-Blackwell Theorem offers a methodology to improve an unbiased estimator if a sufficient statistic exists. However, sufficiency alone doesn't guarantee the optimality of an estimator. The Lehmann-Scheffé Theorem extends this by introducing the concept of a complete statistic, allowing the identification of the Minimum Variance Unbiased Estimator (MVUE).

Theorem (Lehmann and Scheffè) Let:

1. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{P}_\theta$, where $\theta \in \Theta$.
2. $T = T(X_1, \ldots, X_n)$ is sufficient and complete for estimating θ.
3. $\hat{\theta} = g(T)$ is an unbiased estimator of θ.

under the given conditions, $\hat{\theta}$ is the Minimum Variance Unbiased Estimator (MVUE) of θ.

Proof:

The Rao-Blackwell theorem states that if $\hat{\theta}$ is an unbiased estimator of θ, and T is a sufficient statistic for θ, then the estimator $\hat{\theta}^* = \mathbb{E}[\hat{\theta} \mid T]$ has a variance that is less than or equal to the variance of $\hat{\theta}$, i.e., $\text{Var}(\hat{\theta}^*) \leq \text{Var}(\hat{\theta})$, and $\hat{\theta}^*$ is also unbiased for θ.

1. By assumption, $\hat{\theta} = g(T)$ is an unbiased estimator for θ. Thus, $\mathbb{E}[\hat{\theta}] = \theta$.
2. Since T is sufficient for θ, by the Rao-Blackwell theorem, we consider $\hat{\theta}^* = \mathbb{E}[\hat{\theta} \mid T] = \mathbb{E}[g(T) \mid T]$. However, since $\hat{\theta} = g(T)$ is a function of T alone, we have $\hat{\theta}^* = g(T)$, which means $\hat{\theta}^*$ is the same as our original estimator $\hat{\theta}$.
3. The completeness of T for θ implies that if $\mathbb{E}[h(T)] = 0$ for all $\theta \in \Theta$, then $h(T) = 0$ almost surely. Since $\hat{\theta}$ is unbiased and a function of the complete sufficient statistic T, it is the unique function (up to almost everywhere equality) satisfying its own expected value equation, making it the MVUE by the definition of completeness and the Rao-Blackwell theorem.

Conclusion: $\hat{\theta} = g(T)$ is the MVUE for θ, leveraging the properties of sufficiency, unbiasedness, completeness, and the Rao-Blackwell theorem.

Determination of a complete statistic

The determination of a complete statistic may be challenging, unless we have a one-parameter exponential family. If

$$f(x;\theta) = \exp\left\{\mathcal{A}(\theta)\mathcal{B}(x) + \mathcal{C}(x) + \mathcal{D}(\theta)\right\} \tag{118}$$

Then $T = \sum_{i=1}^{n} \mathcal{B}(X_i)$ is sufficient and complete

Lecture 21

We did only exercises, suggested revising the last 4 lectures.

Lecture 22

Maximum Likelihood Estimator

Given a parameter vector θ and a data set $X = \{x_1, x_2, \ldots, x_n\}$, the likelihood function is defined as:

$$L(\theta; X) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{119}$$

where $f(x_i \mid \theta)$ is the probability mass function (for discrete data) or the probability density function (for continuous data). The log-likelihood function (see refresher 0.14) is defined as:

$$\ell(\theta; X) = \sum_{i=1}^{n} \log f(x_i \mid \theta) \tag{120}$$

The goal of Maximum Likelihood Estimation (MLE) is to find the parameter $\hat{\theta}$ that maximizes $\ell(\theta; X)$. This involves taking the derivative of $\ell(\theta; X)$ with respect to θ, setting it equal to zero, and solving for θ. This derivative, known as the score function, is given by:

$$\frac{d}{d\theta}\ell(\theta; X) = \sum_{i=1}^{n} \frac{d}{d\theta} \log f(x_i \mid \theta) \tag{121}$$

Setting the score function equal to zero gives the equations necessary to solve for $\hat{\theta}$. To verify that the point $\hat{\theta}$ found is indeed a maximum, we examine the second derivative of $\ell(\theta; X)$ with respect to θ, which is:

$$\frac{d^2}{d\theta^2}\ell(\theta; X) = \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f(x_i \mid \theta) \tag{122}$$

For a maximum, this second derivative evaluated at $\hat{\theta}$ should be negative. In the case of multiple parameters, the Hessian matrix, which is a matrix of second-order partial derivatives, is used to determine concavity:

$$H(\theta; X) = \begin{bmatrix} \frac{\partial^2 \ell(\theta;X)}{\partial \theta_1^2} & \cdots & \frac{\partial^2 \ell(\theta;X)}{\partial \theta_1 \partial \theta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell(\theta;X)}{\partial \theta_n \partial \theta_1} & \cdots & \frac{\partial^2 \ell(\theta;X)}{\partial \theta_n^2} \end{bmatrix} \tag{123}$$

The criterions for a maximum in this multivariate case are:

1. the Hessian matrix evaluated at $\hat{\theta}$ is **negative definite**
2. $\det\Big(H(\hat{\theta}, X)\Big) > 0$

Lecture 23

Sufficient Statistic and Maximum Likelihood Estimation

We can use the Fisher-Neyman Theorem to find the MLE by following these steps:

1. Write the likelihood function.
2. Apply the Factorization Theorem:
   - Find a way to express the likelihood function in the form required by the Fisher-Neyman factorization theorem.
   - Identify $T(X_1, \ldots, X_n)$, the function ν, and the function W.
3. **Maximize the Likelihood:**
   - Focus on maximizing $\nu(t, \theta)$ with respect to θ since $W(X_1, \ldots, X_n)$ does not affect the maximization.
   - Differentiate $\nu(t, \theta)$ with respect to θ and set the derivative to zero to find the MLE. Ensure that the second derivative is negative, confirming that the solution corresponds to a maximum.

We can just maximise the function $\nu(t,\theta)$ because, as shown in the image below, the point of maximum for $\nu(t,\theta)$ will also be the point of maximum of $\nu(t,\theta)W(X_1,\ldots,X_n)$

From the Fisher-Neyman theorem we have the following theorem.

Theorem: If $T(X_1,\ldots,X_n)$ is a sufficient statistic for a parameter θ, and if there exists a unique maximum likelihood estimator (MLE) of θ, then this unique MLE can be expressed as a function of the sufficient statistic $T(X_1,\ldots,X_n)$. (see refresher 0.15)

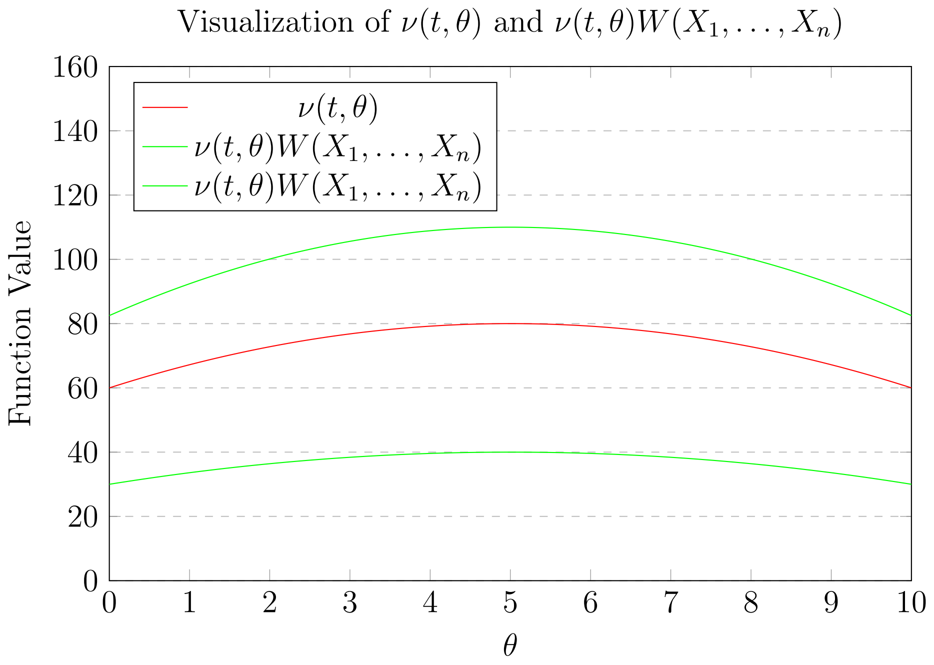
Lecture 24

Estimation of Transformed Parameters

Assume $X_1,\ldots,X_n \overset{\text{iid}}{\sim} F_\theta$ and that we have a maximum likelihood estimator (MLE), $\hat{\theta}$. We aim to find an estimator for η.

Theorem: Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} F_\theta$ with $\theta \in \Theta$, and suppose that $\hat{\theta}$ is a maximum likelihood estimator (MLE). Then for every bijective function $g : \Theta \to \Omega$, the maximum likelihood estimator of η (where $\eta \in \Omega$), $\hat{\eta}$, is given by $\hat{\eta} = g(\hat{\theta})$.

Proof: For every $\theta \in \Theta$, we have $L(\theta) = L\left(g^{-1}\left(g(\theta)\right)\right) = L\left(g^{-1}(\eta)\right)$. Assuming



$\max_{\theta \in \Theta} L(\theta) = L(\hat{\theta})$, then $L(\hat{\theta}) = L\left(g(\hat{\theta})\right)$. From here, we deduce that $\hat{\theta} = g^{-1}(\hat{\eta})$ which is equivalent to $\hat{\eta} = g(\hat{\theta})$. (see refresher 0.16 for what logit/probit means from the example)

Asymptotic Properties of Maximum Likelihood Estimators

MLEs are said to have some asymptotic properties:

- The MLE $\hat{\theta}$ is said to be consistent if $\hat{\theta}$ converges in probability to the true parameter θ as the sample size n tends to infinity. Formally,

  $$\hat{\theta} \xrightarrow{p} \theta \quad \text{as } n \to \infty.$$

  Consistency ensures that the MLE estimates are reliable with large samples.

- Under certain regularity conditions, the MLE $\hat{\theta}$ is not only consistent but also asymptotically normally distributed. As $n \to \infty$, the scaled difference between the MLE and the true parameter approximates a normal distribution:

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

  where $\mathcal{I}(\theta)$ is the Fisher information.

  For the asymptotic normality of MLEs, the following regularity conditions must be satisfied:

  1. The parameter space Θ should be an open subset of $\mathbb{R}^k$.
  2. The true parameter value θ must lie in the interior of Θ.
  3. The likelihood function must be three times continuously differentiable with respect to θ.
  4. The Fisher information $\mathcal{I}(\theta)$ is positive and is continuous as a function of θ. (In the n dimensional case, we have that the Fisher information matrix $\mathcal{I}_n(\theta)$ must be positive definite and the function of each entry is continuous for any θ)
  5. The likelihood equations have a unique solution, and the log-likelihood function should satisfy uniform convergence properties. (see refresher 0.17)

- Asymptotically the MLE is the MVUE

- $\hat{\theta}_n \xrightarrow{\mathrm{d}} Y \sim N_d(\theta, \frac{\mathcal{I}^{-1}(\theta)}{n})$

Multidimensional MLE Asymptotics

For a vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^k$, the MLE $\hat{\boldsymbol{\theta}}$ extends the unidimensional asymptotic properties. The distribution of the estimator vector can be expressed as:

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

where $\mathcal{I}_k(\boldsymbol{\theta})$ is the Fisher information matrix. This result is fundamental for multivariate statistical inference.

The Newton-Raphson algorithm is a root-finding method that uses the first and second derivatives of a function to rapidly converge on a solution that makes the function zero. This method is particularly useful in the context of statistical estimation and numerical optimization.

Newton-Raphson Algorithm

Unidimensional Case

In the unidimensional case, the algorithm seeks to find the roots of a function $f(x)$ by iteratively moving closer to the solution starting from an initial guess x_0. The update rule is given by:

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$$

where $f'(x)$ is the derivative of $f(x)$.

This method is commonly used to find the maximum likelihood estimates of parameters in statistical models by setting $f(x)$ as the derivative of the log-likelihood function, thus solving $f'(x) = 0$.

Multidimensional Case

In the multidimensional case, the algorithm extends to finding roots of a vector-valued function $\mathbf{F}(\mathbf{x})$. The update formula is:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - [\mathbf{J}(\mathbf{x}_t)]^{-1}\mathbf{F}(\mathbf{x}_t)$$

where $\mathbf{J}(\mathbf{x}_t)$ is the Jacobian matrix of partial derivatives of $\mathbf{F}$ at $\mathbf{x}_t$.

Convergence

The convergence of the Newton-Raphson algorithm depends critically on the choice of the initial guess and the nature of the function:

- **Rapid Convergence**: Near the root, the algorithm converges quadratically, making it very efficient for practical use.
- **Sensitivity**: The method can be sensitive to the initial guess, especially if $f'(x)$ is near zero or the function is not well-behaved.

Lecture 25

Introduction to Bayesian Statistics

In frequentist statistics, given a sample $X_1, \ldots, X_n$, we estimate θ, a parameter or a vector of parameters, which then enables us to predict the outcome of a new data point using the probability $\mathbb{P}(X_{n+1} \mid \hat{\theta})$.

In Bayesian statistics, however, we treat θ itself as a random variable. This approach allows us to use all available information, encapsulated in $X_1, \ldots, X_n$, to update our beliefs about the distribution of θ.

Bayes' Theorem

The foundation of Bayesian statistics is Bayes' Theorem, which is used to update our probability estimate for a hypothesis as more evidence or information becomes available.

Bayes' Theorem (Simple Form)

Bayes' Theorem in its simple form is expressed as:

$$\mathbb{P}(C_i \mid B) = \frac{\mathbb{P}(B \mid C_i)P(C_i)}{\mathbb{P}(B)}$$

where $\mathbb{P}(B)$ is the probability of the evidence, and $\mathbb{P}(B \mid C_i)$ is the probability of the evidence given that hypothesis C_i is true. (proof refresher 0.18)

Bayes' Theorem (General Form)

Assuming $C_1, \ldots, C_k$ are partitions of the sample space and each $\mathbb{P}(C_i) > 0$ for $i = 1, \ldots, k$, and let B be an event with $\mathbb{P}(B) > 0$. Bayes' Theorem can then be generalized as:

$$\mathbb{P}(C_i \mid B) = \frac{\mathbb{P}(B \mid C_i)\mathbb{P}(C_i)}{\sum_{j=1}^{k} \mathbb{P}(B \mid C_j)\mathbb{P}(C_j)} \tag{124}$$

Proof of General Bayes' Theorem

First, recognize that $\mathbb{P}(C_i \mid B) = \frac{\mathbb{P}(C_i,B)}{\mathbb{P}(B)}$. Also, $\mathbb{P}(B) = \sum_{j=1}^{k} \mathbb{P}(B, C_j)$ can be rewritten using the simple form of Bayes' Theorem as $\sum_{j=1}^{k} \mathbb{P}(B \mid C_j)\mathbb{P}(C_j)$. Combining these, we arrive at the general form of Bayes' Theorem.

Bayesian Model

To define a Bayesian model, we begin by specifying a statistical model for the data conditional on a parameter $\theta \in \Theta$. The model is expressed as:

$$\mathcal{M} = \{f(\cdot \mid \theta),\, \theta \in \Theta\} \tag{125}$$

where $f(\cdot \mid \theta)$ represents the probability distribution of the data given the parameter θ.

Next, we assume a prior distribution for θ, denoted by $\theta \sim G$, with a probability density function $g(\theta)$. This prior should reflect prior knowledge about the phenomenon being modeled:

$$\theta \sim G(\theta) \tag{126}$$

Note: Specifying the model $\mathcal{M}$ and the prior G implies that observations X_i are not independent and identically distributed (iid), even if $X_i \perp X_j \mid \theta$. This is because the prior introduces a dependency through θ. Under a Bayesian model, the data are exchangeable rather than independent. Exchangeability means that any permutation of the data is equally likely, which is a weaker condition than independence and is more appropriate when a common prior affects all observations.

Under the assumption of exchangeability, we define the likelihood function as:

$$\mathcal{L}_x(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{127}$$

where we assume that $X_1, \ldots, X_n$ are conditionally i.i.d. from $f(\cdot \mid \theta)$ given θ.

With this, we can now define the posterior distribution:

$$g\left(\theta \mid X_1, \ldots, X_n\right) \propto \mathcal{L}_x(\theta) g(\theta) \tag{128}$$

where $\propto$ indicates that it is equal up to a proportional constant, defined as:

$$C = \frac{1}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} f(x_i \mid \theta) g(\theta)\, d\theta} \tag{129}$$

Bayesian Inference using Bayes' Rule

The core of Bayesian inference is updating our belief about θ after observing data $X_1, \ldots, X_n$. This is done by applying Bayes' rule to find the posterior distribution of θ given the data.

Theorem (Bayesian Updating): The posterior density function of θ given $X_1, \ldots, X_n$ is:

$$g(\theta \mid X_1, \ldots, X_n) = \frac{g(\theta) \prod_{i=1}^{n} f(x_i \mid \theta)}{m(X_1, \ldots, X_n)} \tag{130}$$

where the marginal likelihood $m(X_1, \ldots, X_n)$ is given by:

$$m(X_1, \ldots, X_n) = \int \prod_{i=1}^{n} f(x_i \mid \theta) g(\theta) \, d\theta \tag{131}$$

Making Predictions and Determining Credible Intervals

With the posterior distribution $g(\theta \mid X_1, \ldots, X_n)$, we can make predictions for future observations and perform inference on θ. One common inferential goal is to compute credible intervals for θ, which are intervals $[\theta_1, \theta_2]$ within Θ such that:

$$\mathbb{P}(\theta_1 < \theta < \theta_2 \mid X_1, \ldots, X_n) \geq 1 - \alpha \tag{132}$$

This is equivalent to:

$$\int_{\theta_1}^{\theta_2} g(\theta \mid X_1, \ldots, X_n) \, d\theta \geq 1 - \alpha$$

Lecture 26

Definition: a prior is conjugate for a statistical model $\mathcal{M} = \{f(\cdot, \theta), \theta \in \Theta\}$ if the posterior distribution has the same analytical form of the prior with the updated parameters

Lecture 27

Bayesian Statistics and Estimation

In frequentist statistics, we often define optimality criteria based on properties such as unbiasedness, efficiency, and consistency, with methods like maximum likelihood estimation being prevalent. In contrast, Bayesian statistics introduces a probabilistic approach to estimation, incorporating prior knowledge through a prior distribution.

Another way of defining optimal estimators is through the `minimax` criterion, where $\hat{\theta}$ is defined as minimax if

$$R(\hat{\theta}) = \min_{\hat{\theta}\in A} \max_{\theta\in\Theta} R(\theta, \hat{\theta}) \tag{133}$$

Where A is the family of all estimators We begin by defining a set of possible actions $\mathcal{A}$, each representing a potential estimator of an unknown parameter θ. We associate with each action $a \in \mathcal{A}$ a loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}^+$, where Θ is the parameter space, defined by:

$$L(\theta, a) = \text{the loss incurred from estimating } \theta \text{ with } a. \tag{134}$$

We consider a loss function that quantifies the "cost" of the difference between the true parameter value and the estimator. Common choices for L include:

1. **Quadratic Loss Function:** $L(\theta, a) = (\theta - a)^2$. This loss penalizes squared deviations from the true value, emphasizing larger errors more heavily.

2. **Absolute Error Loss Function:** $L(\theta, a) = |\theta - a|$. This function penalizes the absolute value of the error, providing a linear response to the estimation error.

3. **0-1 Loss Function:**
$$L(\theta, a) = \begin{cases} 1 & \text{if } \theta \neq a, \\ 0 & \text{if } \theta = a. \end{cases}$$
This loss function is useful for categorical decisions, penalizing any incorrect estimation without considering the magnitude of the error.

The risk function, which represents the expected loss for an estimator, is defined as:

$$R(\theta, \hat{\theta}) = \mathbb{E}[L(\theta, \hat{\theta})] = \int L(\theta, \hat{\theta}(x_1, \ldots, x_n)) \prod_{i=1}^{n} f(x_i|\theta)\, dx_1 \cdots dx_n, \tag{135}$$

where $f(x_i|\theta)$ denotes the likelihood of observing x_i given θ.

In Bayesian statistics, we integrate over all possible values of θ weighted by a prior distribution $g(\theta)$, leading to the Bayesian risk function:

$$r(g, \hat{\theta}) = \int R(\theta, \hat{\theta}) g(\theta)\, d\theta = \int \left(\int L(\theta, \hat{\theta}(x_1, \ldots, x_n)) \prod_{i=1}^{n} f(x_i|\theta)\, dx_i \right) g(\theta)\, d\theta. \tag{136}$$

The optimal Bayesian estimator $\hat{\theta}_g$ is defined as the action (or estimator) that minimizes the Bayesian risk:

$$\hat{\theta}_g = \arg\min_{\hat{\theta}\in\mathcal{A}} r(g, \hat{\theta}), \tag{137}$$

and equivalently minimizes the posterior risk:

$$r(g, \hat{\theta}_g|x_1, \ldots, x_n) = \min_{\hat{\theta}\in\mathcal{A}} \int L(\theta, \hat{\theta}) g(\theta|x_1, \ldots, x_n)\, d\theta. \tag{138}$$

Definition: A Bayesian estimator for θ, with respect to the loss function L and prior g, is the statistic $\hat{\theta}_g$ that minimizes the Bayesian risk function, incorporating both the likelihood of the data given the parameter and the prior distribution of the parameter.

Note: Depending on the loss, we have different optimal estimators:

- If quadratic loss is used, then the optimal bayesian point is the posterior mean
- If absolute error loss is used, then the optimal bayesian point estimator is the median
- If 0-1 loss is used, the optimal bayesian point estimator is the Maximum A Posteriori (MAP) - ie. the value that maximizes the posterior (you do the same procedure as MLE but instead of taking the derivative of $\log(f(x))$ and setting it to zero, you take $g(\theta \mid X_1, \ldots, X_n))$

Lecture 28

Introduction to Statistical Hypothesis Testing

Given a sample $X_1, \ldots, X_n$ that are independently and identically distributed (*iid*) from a distribution F_θ, where the parameter θ belongs to a parameter space Θ, hypothesis testing involves comparing two hypotheses. These are the null hypothesis $H_0 : \theta \in \Theta_0$ and the alternative hypothesis $H_1 : \theta \in \Theta_1$, with $\Theta_0, \Theta_1 \subset \Theta$.

Definitions and Concepts

- **Simple hypothesis**: Specifies a fixed numerical value for θ. For example, $H_0 : \theta = \theta_0$.
- **Composite hypothesis**: Specifies a set of possible values for θ. For example, $H_0 : \theta \in [\theta_0, \theta_1]$.

- **One-sided hypothesis**: Considers a one-directional alternative. For example, $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

- **Two-sided hypothesis**: Considers both directions as alternatives. For example, $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

- **Hypothesis test**: A statistical procedure used to test H_0 against H_1.

Critical Region and Types of Errors

Critical Region Definition: For a given test statistic T, the critical region is the set of values of T for which the null hypothesis H_0 is rejected.

In hypothesis testing, two types of errors can occur:

- **Type I Error (False Positive)**: Rejecting H_0 when it is true. The probability of making a Type I error is denoted by α.

- **Type II Error (False Negative)**: Not rejecting H_0 when it is false. The probability of making a Type II error is denoted by β.

The following table summarizes these errors along with correct decisions:

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN) - Type II Error
Actual Negative	False Positive (FP) - Type I Error	True Negative (TN)

Probabilities of Errors

The error probabilities α and β are defined as:

$$\alpha = \mathbb{P}(\text{Reject } H_0 \mid H_0), \tag{139}$$

$$\beta = \mathbb{P}(\text{Not reject } H_0 \mid H_1). \tag{140}$$

Typically, α and β are chosen to balance the risks of these errors, often by selecting a small value for α.

Definition of β for Composite H_0

When H_0 is composite, α is defined as the supremum of the rejection probabilities under all θ in Θ_0:

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}(\text{Reject } H_0 \mid \theta). \tag{141}$$

β is similarly defined as:

$$\beta = \inf_{\theta \in \Theta_1} \mathbb{P}(\text{Not reject } H_0 \mid \theta), \tag{142}$$

reflecting the smallest probability of a Type II error over all parameter values that make H_1 true.

Lecture 29

Power Function

Power Function **Definition**:
Let $X_1, \ldots, X_n \overset{iid}{\sim} F_\theta$. Let $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, with $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. Let T be the test statistic of a test with critical region CR. The power function is defined as

$$p(\theta) : \theta \rightarrow \mathbb{P}(\text{reject } H_0 \mid \theta) = \mathbb{P}\left((X_1, \ldots, X_n) \in CR \mid \theta\right) \tag{143}$$

If $\theta \in \Theta_0$ then $p(\theta)$ is the probability of type I error and if $\theta \in \Theta_1$, it's $1-$probability of type II error.

Neyman Pearson Theorem

Neyman Pearson **Theorem**: Let X be a random variable with

$$H_0 : x \sim f(\cdot \mid \theta_0) \qquad H_1 : X \sim f(\cdot \mid \theta_1) \tag{144}$$

If $k > 0$ is such that

$$\mathbb{P}\left(\frac{f(x \mid \theta_1)}{f(x \mid \theta_0}) > k\right) = \alpha \tag{145}$$

then the test with critical region $CR = \{x : f(x \mid \theta_1) > kf(x \mid \theta_0)\}$ has the minimum probability of type II error among the tests with size α.
Proof: Consider a test defined by the critical region:

$$CR = \left\{x : \frac{f(x \mid \theta_1)}{f(x \mid \theta_0)} > k\right\}$$

where k is chosen such that:

$$\mathbb{P}\left(\frac{f(X \mid \theta_1)}{f(X \mid \theta_0)} > k \mid \theta_0\right) = \alpha$$

This ensures that the test has size α.

Suppose there is an alternative test defined by another critical region CR', which also has size α. We need to show that for all $\theta_1 \in \Theta_1$:

$$\mathbb{P}(X \in CR \mid \theta_1) \geq \mathbb{P}(X \in CR' \mid \theta_1)$$

Define:

$$A = \left\{ x : \frac{f(x \mid \theta_1)}{f(x \mid \theta_0)} > k \right\} \quad \text{and} \quad B = CR'$$

Then,

$$\mathbb{P}(X \in A \mid \theta_1) = \int_A f(x \mid \theta_1)\, dx$$

$$\mathbb{P}(X \in B \mid \theta_1) = \int_B f(x \mid \theta_1)\, dx$$

We can write the difference in power between the two tests as:

$$\int_{A \cap B} f(x \mid \theta_1)\, dx + \int_{A \setminus B} f(x \mid \theta_1)\, dx - \int_{B \setminus A} f(x \mid \theta_1)\, dx$$

Given that $\frac{f(x|\theta_1)}{f(x|\theta_0)} > k$ for $x \in A$ and $\leq k$ for $x \notin A$, it follows that:

$$\int_{A \setminus B} f(x \mid \theta_1)\, dx \geq k \int_{A \setminus B} f(x \mid \theta_0)\, dx$$

$$\int_{B \setminus A} f(x \mid \theta_1)\, dx \leq k \int_{B \setminus A} f(x \mid \theta_0)\, dx$$

Since the size of both tests is α, we have:

$$\int_{A \setminus B} f(x \mid \theta_0)\, dx = \int_{B \setminus A} f(x \mid \theta_0)\, dx$$

Thus,

$$\int_{A \setminus B} f(x \mid \theta_1)\, dx \geq \int_{B \setminus A} f(x \mid \theta_1)\, dx$$

This implies:

$$\mathbb{P}(X \in CR \mid \theta_1) \geq \mathbb{P}(X \in CR' \mid \theta_1)$$

which shows that the Neyman-Pearson test is the most powerful test.

Note: if there exists a sufficient statistics T for the model, the likelihood ratio can be expressed in terms of functions involving T, which are $\nu(t, \theta)$ and $w(x_1, \ldots, x_n)$.

The expression simplifies because $w(x_1, \ldots, x_n)$ cancels out in the numerator and the denominator, resulting in:

$$\frac{\nu(t, \theta_1)}{\nu(t, \theta_0)}$$

The simplified form of the likelihood ratio allows establishing a rejection rule based on the statistic T rather than computing the likelihood ratio for every data point. This is particularly useful when T has a simple relationship with the parameter θ:

- **Monotonely Increasing Function of X**: If T increases as X increases, then you reject H_0 if $T > k$, where k is a threshold determined by the desired significance level α.

- **Monotonely Decreasing Function of X**: If T decreases as X increases, then you reject H_0 if $T < k$.

Lecture 30

The problem of the Neyman Pearson theorem is that it doesn't account for composite hypothesis, so to extend it, we use the following definition.

Definition Let $X_1, \ldots, X_n \overset{iid}{\sim} F_\theta$ and consider a testing problem involving two composite hypothesis, with $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, a test $\mathcal{A}$ based on $X_1, \ldots, X_n$ is said to be the Uniformly Most Powerful (UMP) test of size α if $\mathcal{A}$ has size α and for any alternative test $\mathcal{B}$ with $\leq \alpha$, $\mathcal{A}$ has the same or greater power than $\mathcal{B}$ for every $\theta \in \Theta_1$

Lecture 31

Leibniz Rule:

$$\frac{\partial}{\partial s} \int_{a(s)}^{b(s)} f(s, y)\, dy = f\left(s, b(s)\right) \cdot \frac{\partial}{\partial s} b(s) - f\left(s, a(s)\right) \cdot \frac{\partial}{\partial s} a(s) + \int_{a(s)}^{b(s)} \frac{\partial}{\partial s} f(s, y) dy \quad (146)$$

Refresher

Refresher of last year

Distributions

- **Normal Distribution** ($\mathcal{N}(\mu, \sigma^2)$)
  - PDF: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - Parameters: Mean $\mu \in \mathbb{R}$, Variance $\sigma^2 > 0$
- **Exponential Distribution** ($\text{Exp}(\lambda)$)
  - PDF: $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$
  - Parameter: Rate $\lambda > 0$
- **Uniform Distribution** ($\mathcal{U}(a, b)$)
  - PDF: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
  - Parameters: Interval endpoints $a < b$
- **Gamma Distribution** ($\text{Gamma}(\alpha, \beta)$)
  - PDF: $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x > 0$
  - Parameters: Shape $\alpha > 0$, Rate $\beta > 0$
- **Beta Distribution** ($\text{Beta}(\alpha, \beta)$)
  - PDF: $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ for $0 < x < 1$
  - Parameters: Shape $\alpha > 0$, $\beta > 0$
- **T Distribution (Student's T)**
  - PDF: $f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}$ for $-\infty < t < \infty$
  - Parameters: Degrees of freedom $k > 0$
  - Notes: Used for estimating the mean of a normally distributed population when the sample size is small and population standard deviation is unknown. It is given by $\frac{X}{\sqrt{\frac{Y}{k}}}$ where $X \sim N(0, 1)$ and $Y \sim \chi^2(k)$
- **Cauchy Distribution**

  - PDF: $f(x) = \frac{1}{\pi\theta\left(1+\left(\frac{x-m}{\theta}\right)^2\right)}$ for $-\infty < x < \infty$
  - Parameters: Location parameter $m \in \mathbb{R}$, scale parameter $\theta > 0$
  - Notes: The mean and variance are undefined. Known for its "heavy tails." It's given by $\frac{X_1}{X_2}$ where $X_1, X_2 \overset{i.i.d}{\sim} N(0,1)$

- **F Distribution**
  - PDF: $f(x) = \frac{\Gamma\left(\frac{k+r}{2}\right)\left(\frac{k}{r}\right)^{k/2} x^{k/2-1}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{r}{2}\right)\left(1+\frac{k}{r}x\right)^{\frac{k+r}{2}}}$ for $x > 0$
  - Parameters: Degrees of freedom $k > 0$ and $r > 0$
  - Notes: Used in the analysis of variance, comparing the variances of different samples.

- **Logistic Distribution**
  - PDF: $f(x) = \frac{e^{-\frac{x-\mu}{s}}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)^2}$ for $-\infty < x < \infty$
  - Parameters: Mean $\mu \in \mathbb{R}$, scale $s > 0$
  - Notes: Used for modeling growth, and in logistic regression.

- **Log-Normal Distribution**
  - PDF: $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)$ for $x > 0$
  - Parameters: Mean $\mu \in \mathbb{R}$, standard deviation $\sigma > 0$
  - Notes: If X is log-normally distributed, then $\ln(X)$ is normally distributed. Often used to model stock prices.

- **Pareto Distribution**
  - PDF: $f(x) = \frac{\alpha\cdot k}{x^{\alpha+1}} \cdot \mathbb{1}_{(k,+\infty)}$
  - Parameters: Shape $\alpha > 0$, scale $k > 0$
  - Notes: Used to model the distribution of wealth, sizes of cities, etc.

Discrete Distributions

- **Binomial Distribution** (Bin(n,p))
  - PMF: $f(k) = \binom{n}{k}p^k(1-p)^{n-k}$ for $k = 0, 1, \ldots, n$

  - Parameters: Number of trials $n \geq 0$, Success probability $p \in [0, 1]$

- **Poisson Distribution** ($\text{Pois}(\lambda)$)
  - PMF: $f(k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for $k = 0, 1, 2, \ldots$
  - Parameter: Rate $\lambda > 0$

- **Geometric Distribution** ($\text{Geom}(p)$)
  - PMF: $f(k) = p(1-p)^{k-1}$ for $k = 1, 2, \ldots$
  - Parameter: Success probability $p \in [0, 1]$

- **Negative Binomial Distribution** ($\text{NegBin}(r, p)$)
  - PMF: $f(k) = \binom{k+r-1}{k} p^r (1-p)^k$ for $k = 0, 1, \ldots$
  - Parameters: Number of successes $r > 0$, Success probability $p \in [0, 1]$

- **Hypergeometric Distribution** ($\text{Hypergeom}(N, K, n)$)
  - PMF: $f(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$ for $k = 0, 1, \ldots, \min(K, n)$
  - Parameters: Population size N, Number of successes in population K, Sample size n

0.1 Countable and uncountable sets

In mathematics, the concepts of countable and uncountable sets are crucial in understanding the different sizes of infinity and the structure of the real number line.

A set is **countable** if it is finite, or if it has the same cardinality as the set of natural numbers $\mathbb{N}$. In other words, a set S is countable if there exists a bijection (a one-to-one correspondence) between S and $\mathbb{N}$, or if S can be listed in a sequence $(s_1, s_2, s_3, \ldots)$ where every element of S appears exactly once in the sequence. This implies that the elements of a countable set can be enumerated without omission.
Examples:

- The set of natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$ is countable.
- The set of integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ is countable.
- The set of rational numbers $\mathbb{Q}$, which can be written as fractions of integers, is countable.

A set is **uncountable** if it is not countable, meaning there is no bijection between the set and the set of natural numbers $\mathbb{N}$. Uncountable sets have a higher cardinality (size of infinity) than countable sets. The most well-known example of an uncountable set is the set of real numbers $\mathbb{R}$.

0.2 Kernel Functions

Kernel functions serve as the foundation for a wide range of applications, from solving differential equations to data classification and pattern analysis.

In the field of integral transforms, a kernel function plays a crucial role in transforming one function into another via integration. This function, when integrated in conjunction with another function, results in a newly transformed function. Examples include:

- The exponential function e^{-st} used in the Laplace transform.
- The sine and cosine functions in the Fourier transform, represented as $e^{i\omega t}$.
- General kernel functions $K(x, y)$ in convolution integrals and other transforms.

0.3 Series

Geometric Series

A geometric series is a sequence of terms where each subsequent term is the product of the previous term and a constant known as the common ratio. If we let the series start from an arbitrary term indexed at k, the series can be expressed as:

$$S = ar^k + ar^{k+1} + ar^{k+2} + \cdots \tag{147}$$

Here, a represents the initial term of the series (when $k = 0$), r is the common ratio, and k is the starting index of the summation, which can be any integer, allowing the series to begin at any term.

- For a finite geometric series starting at an arbitrary term k and ending at term n, the sum can be represented using the summation notation as:

  $$S_{k,n} = \sum_{i=k}^{n} ar^i \tag{148}$$

  The formula for the sum of the first n terms from the start index k (inclusive) is given by:

$$S_{k,n} = a\left(\frac{r^k - r^{n+1}}{1-r}\right), \quad \text{for } r \neq 1. \tag{149}$$

- For an infinite geometric series starting from an arbitrary term k, the sum, assuming $|r| < 1$, is:

$$S_{k,\infty} = \sum_{i=k}^{\infty} ar^i = a\frac{r^k}{1-r}. \tag{150}$$

Arithmetic Series

An arithmetic series is the sum of the terms of an arithmetic sequence, where each term is derived by adding a constant difference to the previous term. When considering the series starting from an arbitrary term indexed at k, the sum of the series up to the n-th term can be defined as:

$$S_{k,n} = \sum_{i=k}^{n} \left(a + d(i-1)\right), \tag{151}$$

where a is the first term of the sequence, d is the common difference, and i represents the index of the term within the sequence.

The formula for the sum of terms from the arbitrary start index k to the n-th term is given by:

$$S_{k,n} = \frac{(n-k+1)}{2}\left(2a + (n-k)d\right), \tag{152}$$

This formula accounts for the sum of an arithmetic series starting from an arbitrary index k to the n-th term, adjusting the classical formula to accommodate the arbitrary starting point.

Exponential Series

The exponential series is the expansion of the exponential function e^x as an infinite series:

$$e^b = 1 + b + \frac{b^2}{2!} + \frac{b^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{b^k}{k!}. \tag{153}$$

0.3.1 Taylor Series

A Taylor series expands a function into an infinite sum of terms calculated from the values of its derivatives at a single point:

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n. \tag{154}$$

Popular expansions are

- e^x
$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots \tag{155}$$

- $\sin(x)$
$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \cdots \tag{156}$$

- $\cos(x)$
$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + \cdots \tag{157}$$

- $\ln(1+x)$
$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots + (-1)^{n+1}\frac{x^n}{n} + \cdots \tag{158}$$

0.4 Chi-squared statistic

The chi-squared statistic is a measure of the discrepancy between observed and expected frequencies under a specific hypothesis. It is defined as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{159}$$

where:

- O_i represents the observed frequency for the i^{th} category,
- E_i represents the expected frequency for the i^{th} category, as predicted by the hypothesis,
- The summation runs over all categories involved in the analysis.

The chi-squared statistic is widely used in various statistical tests, including:

1. **Goodness-of-Fit Test:** To determine how well the observed data fit a specified distribution.
2. **Test of Independence:** To assess whether there is an association between two categorical variables in a contingency table.
3. **Homogeneity Test:** To check if different populations have the same distribution of a categorical variable.

Characteristics

- The chi-squared statistic is always non-negative.
- A higher chi-squared value indicates a greater discrepancy between observed and expected frequencies, which may lead to rejecting the null hypothesis.
- The distribution of the chi-squared statistic under the null hypothesis follows a chi-squared distribution, with the degrees of freedom depending on the specifics of the test being performed.

Calculation of Degrees of Freedom

- In a goodness-of-fit test, the degrees of freedom are typically $n-1-p$, where n is the number of categories, and p is the number of parameters estimated from the data.
- In a test of independence within a contingency table of size $r \times c$, the degrees of freedom are $(r-1) \times (c-1)$, where r and c represent the number of rows and columns, respectively.

0.5 N-Dimensional random vectors

Let $(X_1, \cdots, X_n$ be a random vector, then, if $X_1, \cdots, X_n$ are PMFs, the PMF of the vector is described by

$$p_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \mathbb{P}[X_1 = x_1, \cdots, X_n = x_n] \tag{160}$$

With

$$\sum_{x_1} \cdots \sum_{x_n} p_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = 1 \tag{161}$$

$$p_{X_1,\cdots,X_n} \geq 0 \tag{162}$$

If $X_1, \cdots, X_n$ are PDFs, the PDF of the vector is described by

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) \tag{163}$$

with

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) dx_1, \cdots, dx_n = 1 \tag{164}$$

The CMF would be described by

$$\begin{aligned} F_{X_1,\cdots,X_n}(x_1, \cdots, x_n) &= \mathbb{P}[X_1 \leq x_1, \cdots, X_n \leq x_n] & (165) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,\cdots,X_n}(s_1, \cdots, s_n) ds_1, \cdots, ds_n & (166) \end{aligned}$$

And $f_{x_1,\ldots,x_n} = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F_{X_1,\cdots,X_n}$

0.6 Mean Value Theorem

The Mean Value Theorem (MVT) is formally stated as follows: Let f be a function that satisfies both of the following conditions:

1. f is continuous on the closed interval $[a, b]$.
2. f is differentiable on the open interval (a, b).

Then, there exists at least one point c in (a, b) such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \tag{167}$$

This theorem essentially states that there is at least one point on the graph of f where the tangent is parallel to the secant line connecting $(a, f(a))$ and $(b, f(b))$.

0.7 Borel σ-Algebra

The Borel σ-algebra serves as the bridge between topological and measurable spaces. It is defined on any topological space but is most commonly associated with $\mathbb{R}^n$, the n-dimensional Euclidean space. For simplicity, we focus on $\mathbb{R}$.

What is σ Algebra

A σ-algebra is a collection of subsets of a given set, satisfying certain properties that make it suitable for the development of measure theory and probability theory. Let X be a set, which could represent anything from a set of real numbers $\mathbb{R}$ to any abstract space. A σ-algebra $\mathcal{F}$ on X is a collection of subsets of X that includes X itself and is closed under complementation and countable unions and intersections.

Properties: For $\mathcal{F}$ to be considered a σ-algebra on X, it must satisfy the following properties:

1. **Non-emptiness:** The set X is in $\mathcal{F}$, and consequently, the empty set $\emptyset$ is also in $\mathcal{F}$, since the empty set is the complement of X in X.

$$\emptyset \in \mathcal{F} \quad \text{and} \quad X \in \mathcal{F}. \tag{168}$$

2. **Closure under complementation:** If a set A is in $\mathcal{F}$, then so is its complement $A^c = X \setminus A$.

$$A \in \mathcal{F} \implies A^c \in \mathcal{F}. \tag{169}$$

3. **Closure under countable unions:** If a countable collection of sets $A_1, A_2, \ldots$ are in $\mathcal{F}$, then their union is also in $\mathcal{F}$.

$$\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}. \tag{170}$$

   By De Morgan's laws (refresher 0.9), closure under countable unions and complementation implies closure under countable intersections:

$$\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}. \tag{171}$$

Examples and Intuition

- **Power Set:** The power set of X, denoted by $\mathcal{P}(X)$, which contains all possible subsets of X, is a σ-algebra. It is the largest possible σ-algebra on X.

- **Trivial σ-Algebra:** The smallest σ-algebra on X contains only the empty set and the set X itself, $\{\emptyset, X\}$. It represents the minimal structure that satisfies the σ-algebra properties.

Significance in Mathematics

The concept of a σ-algebra is foundational in measure theory, where it is used to define measurable spaces $(X, \mathcal{F})$ as a precursor to introducing measures on X. In probability theory, σ-algebras underlie the formal definition of probability spaces, enabling the rigorous treatment of events and their probabilities. By specifying which subsets of X are measurable (i.e., those in $\mathcal{F}$), it becomes possible to assign sizes or probabilities in a consistent manner.

Construction of Borel σ-Algebra

To construct the Borel σ-algebra $\mathcal{B}(\mathbb{R})$, we start with a base, which is a collection of open intervals (a, b) where $a, b \in \mathbb{R}$ and $a < b$. The Borel σ-algebra is the smallest σ-algebra containing this base. Formally, it is the intersection of all σ-algebras that contain the open intervals.

Generators of $\mathcal{B}(\mathbb{R})$ Besides open intervals, $\mathcal{B}(\mathbb{R})$ can also be generated by:

- Closed intervals $[a, b]$.
- Open sets in $\mathbb{R}$.
- Closed sets in $\mathbb{R}$.
- Half-open intervals $(a, b]$ and $[a, b)$.

Properties and Operations

$\mathcal{B}(\mathbb{R})$ possesses several key properties inherent to σ-algebras:

1. **Closure under complementation:** If $A \in \mathcal{B}(\mathbb{R})$, then its complement $A^c \in \mathcal{B}(\mathbb{R})$.
2. **Closure under countable unions and intersections:** If $\{A_i\}_{i=1}^{\infty}$ are in $\mathcal{B}(\mathbb{R})$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}(\mathbb{R})$ and $\bigcap_{i=1}^{\infty} A_i \in \mathcal{B}(\mathbb{R})$.

Role in Measure Theory

In measure theory, the Borel σ-algebra is essential for defining Borel measures, which are measures defined on $\mathcal{B}(\mathbb{R})$. The Lebesgue measure, a fundamental example, assigns "length" to sets in $\mathcal{B}(\mathbb{R})$.

Lebesgue Measure on $\mathcal{B}(\mathbb{R})$

The Lebesgue measure m is defined such that for any interval $[a, b]$, $m([a, b]) = b - a$. This measure is then extended to all sets in $\mathcal{B}(\mathbb{R})$ through the process of completion.

Significance in Probability Theory

In probability theory, the Borel σ-algebra underlies the formal definition of continuous random variables. A continuous random variable $X : \Omega \to \mathbb{R}$ is such that for every Borel set B, the preimage $X^{-1}(B)$ is in the σ-algebra of the sample space Ω, ensuring that probabilities can be assigned to events involving X.

Measurability of Functions and Borel Sets

The concept of measurability plays an important role in determining the properties of functions within the realm of probability theory. Specifically, functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are considered measurable if, for any Borel set B in $\mathbb{R}$, their pre-images $g^{-1}(B)$ and $h^{-1}(B)$ are also Borel sets. This property is crucial because it ensures that operations involving these functions, such as the transformation of random variables, preserve the structure necessary for applying probabilistic models.

Implications for Continuous Random Variables

When g and h are measurable, this implies that transformations of continuous random variables through these functions also result in sets that are measurable within the Borel σ-algebra. For example, if X is a continuous random variable and $g(X)$ represents a transformation of X, the measurability of g ensures that the event $g(X) \in C$, where C is a Borel set, corresponds to a measurable event in the original probability space. This compatibility is fundamental for defining the distribution and properties of transformed random variables in a rigorous manner.

Rigorous Treatment of Probability

This framework allows for the rigorous definition and analysis of events, enabling the calculation of probabilities and expectations for a wide range of scenarios involving continuous random variables and their transformations. The measurability of functions g and h thus extends the reach of probabilistic analysis by ensuring that even complex transformations of random variables can be accommodated within the established probabilistic framework.

Expectation and Integration

The expectation of a continuous random variable, as well as probabilities of events defined by it, are computed using integrals with respect to the Lebesgue measure on $\mathcal{B}(\mathbb{R})$, allowing for a rigorous treatment of probability. The measurability of functions g and h, ensuring their pre-images under Borel sets remain Borel sets, underpins the mathematical foundation necessary for these integrations, thereby reinforcing the coherence and integrity of probabilistic analysis.

0.8 Pre-Images

The concept of a pre-image is fundamental in the study of functions, topology, measure theory, and probability theory. It provides a way to analyze and understand how functions map elements from one set to another, especially in contexts where the structure of sets and their elements' relationships are crucial.

Definition of Pre-images

Given a function $f : X \to Y$, where X and Y are sets, the pre-image (or inverse image) of a subset $B \subseteq Y$ under f is defined as the set of all elements in X that f maps into B. Formally, the pre-image of B under f is denoted and defined as:

$$f^{-1}(B) = \{x \in X \mid f(x) \in B\}. \tag{172}$$

Key Points:

- The pre-image $f^{-1}(B)$ is a subset of X.
- The notation f^{-1} does not necessarily imply that f is invertible. Even non-bijective (non-invertible) functions have pre-images for sets.

Properties of Pre-images

Pre-images exhibit several important properties that are pivotal in various branches of mathematics:

1. **Pre-images and Set Operations:** For any function $f : X \to Y$, and subsets $A, B \subseteq Y$, the following properties hold:

   - $f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$
   - $f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B)$
   - $f^{-1}(A^c) = (f^{-1}(A))^c$

   These properties illustrate how pre-images interact with union, intersection, and complement operations, reflecting the structural preservation of set operations through the function f.

2. **Pre-images and Empty Set:** The pre-image of an empty set under any function is always the empty set in X:

$$f^{-1}(\emptyset) = \emptyset. \tag{173}$$

Significance in Measure Theory and Probability

In measure theory and probability theory, the concept of pre-images is crucial for defining measurable functions and setting up the foundational framework for these disciplines:

- A function $f : X \to Y$ between measurable spaces is **measurable** if the pre-image of every measurable set in Y is a measurable set in X. This property ensures that measure and probability can be appropriately applied to the function's outcomes.
- In probability theory, the measurability of a random variable, viewed as a function from a sample space to the real numbers, is essential for defining events and computing probabilities, expectations, and variances.

0.9 De Morgan's Laws

De Morgan's laws provide a way to simplify complex logical statements (not covered) or set expressions. They are presented in two parts, each illustrating the relationship between union and intersection through complementation.

De Morgan's Laws in Set Theory

Given two sets A and B, De Morgan's laws state:

1. The complement of the union of A and B is equal to the intersection of their complements:
$$(A \cup B)^c = A^c \cap B^c. \tag{174}$$
2. The complement of the intersection of A and B is equal to the union of their complements:
$$(A \cap B)^c = A^c \cup B^c. \tag{175}$$

These laws can be extended to any finite or countable number of sets.

0.10 Generalized Inverse

Given a random variable X with its CDF denoted as $F_X(x)$, the usual inverse, $F_X^{-1}(y)$, is defined for values of y where $F_X(x)$ is strictly increasing and continuous. The generalized inverse caters to situations where $F_X(x)$ may not have these properties.

Definition

The generalized inverse of F_X, denoted as $F_X^{-1}(y)$, is defined for any $y \in [0,1]$ as:

$$F_X^{-1}(y) = \inf\{x \in \mathbb{R} : F_X(x) \geq y\} \tag{176}$$

where inf represents the infimum, or the greatest lower bound of the set. This definition ensures that the generalized inverse exists even when $F_X(x)$ is not strictly increasing or when there are jumps in the CDF due to discontinuities.

Transformation to Uniform

A fundamental result in probability theory states that if X is a random variable with CDF F_X, then $Y = F_X(X)$ is uniformly distributed on the interval $[0,1]$. This holds true even when using the generalized inverse for non-strictly increasing CDFs.

Simulation of Random Variables

The generalized inverse is instrumental in the simulation of random variables. Given a random variable $U \sim \mathrm{Unif}(0,1)$, a random variable X with CDF F_X can be simulated as $X = F_X^{-1}(U)$.

0.11 Beyond Mean Squared Error (MSE)

Mean Absolute Deviation (MAD)

The Mean Absolute Deviation (MAD) of an estimator is defined as the expected value of the absolute differences between the estimator and the parameter it estimates. For an estimator $\hat{\theta}$ estimating the parameter θ, the MAD is given by:

$$\mathrm{MAD}(\hat{\theta}) = \mathbb{E}[|\hat{\theta} - \theta|].$$

Unlike MSE, MAD is not as sensitive to large errors because it does not square the differences. This can make MAD a more robust measure of accuracy in the presence of outliers.

Root Mean Squared Error (RMSE)

Although closely related to MSE, the RMSE is often considered separately due to its interpretability, as it is in the same units as the data. It is defined as the square root of MSE:

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathbb{E}[(\hat{\theta} - \theta)^2]}.$$

Bias

The bias of an estimator is a measure of systematic error, defined as the difference between the estimator's expected value and the true value of the parameter:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

Median Unbiased Estimators

An estimator is median unbiased if, for all parameter values, the median of the distribution of the estimator is equal to the true parameter value:

$$\text{P}(\hat{\theta} \leq \theta) = \text{P}(\hat{\theta} \geq \theta) = 0.5.$$

Robustness

An estimator is said to be robust if its performance (in terms of bias and variance) is not significantly affected by deviations from model assumptions. Although not quantifiable in a single metric, robustness is a desirable property in estimators used in real-world data analysis.

0.12 Fisher Information Conditions

To utilize Fisher Information $\mathcal{I}(\theta)$ effectively, the following conditions must be met:

1. **Differentiability:** $f(\theta; x)$ must be differentiable with respect to θ,.

2. **Existence and Finiteness of the Second Derivative:** The second derivative of the log-likelihood function with respect to θ must exist and be finite to define Fisher Information accurately.

3. **Regularity Conditions:** A series of regularity conditions ensure the validity of operations involving the likelihood function and its derivatives:

   (a) **Invariant Support:** The support of the probability distribution, or the set of values that the random variable can assume, should not depend on the parameter θ.

   (b) **Exchangeability of Integration and Differentiation:** It must be permissible to interchange the order of integration (or summation for discrete variables) and differentiation when computing the expected values of the likelihood function's derivatives. This ensures that operations involving the expectation of the score and its square are valid under the integral sign.

(c) **Boundedness of Higher-Order Derivatives:** Higher-order derivatives of the log-likelihood function (beyond the second order) should be bounded in expectation. This condition aids in the expansion and approximation processes, ensuring that the Fisher Information can be accurately defined and computed.

0.13 Proof of Fisher Information

The Fisher Information for a single observation X_1 and parameter θ can be represented in two forms.

To prove these representations are equivalent, consider the following steps:

Step 1: The derivative of the log-likelihood function with respect to θ is:

$$\frac{\partial}{\partial\theta}\ln f(X_1;\theta) = \frac{1}{f(X_1;\theta)}\frac{\partial}{\partial\theta}f(X_1;\theta) \tag{177}$$

Step 2: The expectation of this derivative, under regularity conditions, is zero:

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f(X_1;\theta)\right] = 0 \tag{178}$$

Step 3: By leveraging regularity conditions that allow differentiation under the integral sign and knowing that the total derivative of a probability density function with respect to its parameter integrates to zero, we can show:

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X_1;\theta)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_1;\theta)\right] \tag{179}$$

This establishes the equivalence between the two forms of the Fisher Information, highlighting its role in quantifying the sensitivity of the likelihood function to changes in the parameter θ.

0.14 Usefulness of the log-likelihood function

The use of the logarithm in MLE offers significant advantages:

1. **Simplification**: Transforming the product into a sum through the logarithm makes differentiation with respect to θ simpler. This transformation facilitates the process of finding the maximum likelihood estimate $\hat{\theta}$.

2. **Concavity**: The log-likelihood function often turns out to be concave in θ, making the maximization process straightforward. Optimization techniques perform more reliably when applied to concave functions, ensuring a single global maximum.

3. **Interpretability**: The log transformation can enhance the interpretability, especially when comparing the influence of different data points or understanding the role of the parameter θ.

0.15 Proof of MLE Theorem

According to the Fisher-Neyman factorization theorem, the likelihood function $L(\theta; x)$ for a parameter θ based on data $X_1, \ldots, X_n$ can be factorized as:

$$L(\theta; x) = \nu(t, \theta) \cdot W(X_1, \ldots, X_n)$$

where:

- ν is a function that depends on the sample only through the statistic $t = T(X_1, \ldots, X_n)$ and the parameter θ,
- W is a function of the data that does not depend on θ.

Since $T(X_1, \ldots, X_n)$ is a sufficient statistic, it captures all the information in the sample about θ that is available from the likelihood function. When maximizing the likelihood function for estimation, the function $W(X_1, \ldots, X_n)$ does not influence the estimation because it does not depend on θ. Therefore, the maximization problem reduces to:

$$\hat{\theta} = \arg\max_{\theta} \nu(t, \theta)$$

This shows that the estimation of θ depends only on $t = T(X_1, \ldots, X_n)$.

If the MLE $\hat{\theta}$ is unique, it implies that the maximization of $\nu(t, \theta)$ with respect to θ leads to a single solution for each value of $T(x)$, denoted as $f(T(x))$. Therefore, the unique MLE $\hat{\theta}$ can be expressed as:

$$\hat{\theta} = f(T(X_1, \ldots, X_n))$$

where f is a function that maps the sufficient statistic to the parameter space.

Thus, the unique MLE of θ, if it exists, is a function of the sufficient statistic $T(X_1, \ldots, X_n)$, encapsulating all necessary information for parameter estimation within $T(X_1, \ldots, X_n)$.

0.16 Logit and Probit

Logit and probit models are statistical approaches used to model binary outcome variables in the framework of generalized linear models (GLM). These models are particularly suited for categorical outcomes that are binary, typically represented as 0 and 1.

Logit Model

The logit model, also known as logistic regression, models the probability p that $Y = 1$ given predictors $X_1, X_2, \ldots, X_k$ as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where:

- p is the probability of the dependent variable being a specific case (often coded as 1).
- $\frac{p}{1-p}$ is the odds ratio, representing the odds that an outcome occurs, given the predictors, relative to it not occurring.
- $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients that the model aims to estimate.

Probit Model

The probit model is similar to the logit model but uses the cumulative distribution function (CDF) of the standard normal distribution to link the predictors to the outcome:

$$\Phi^{-1}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where Φ^{-1} is the quantile function (inverse of the CDF) of the standard normal distribution.

Differences and Similarities

- **Link Function:** The main difference between the two models lies in the link function used. The logit model uses the logistic function, while the probit model uses the normal CDF.
- **Interpretation:** In both models, the coefficients β represent the change in the log-odds (logit) or z-score (probit) of the outcome for a one-unit change in the predictor. The specific scale and interpretation of these changes differ due to the different link functions used.
- **Applications:** Logit models are commonly used across various fields due to their straightforward interpretation (odds ratios). Probit models may be preferred when assumptions about the underlying distribution of the error terms are normative.

0.17 Uniform Convergence of the Log-Likelihood Function

Uniform convergence of the log-likelihood function, $\log L(\theta; X)$, towards its expectation means that the convergence

$$\sup_{\theta \in \Theta} |\log L_n(\theta) - E[\log L_n(\theta)]| \to 0 \quad \text{as } n \to \infty$$

holds for all θ in the parameter space Θ. Here, $\log L_n(\theta)$ represents the log-likelihood function based on a sample of size n, and $E[\log L_n(\theta)]$ is its expected value under the true parameter.

0.18 Proof of the Simple Form of Bayes' Theorem

To derive the simple form of Bayes' Theorem, we start with the definition of conditional probability and use the law of total probability. Bayes' Theorem allows us to update our prior beliefs based on new evidence.

Definitions and Setup

- Let A and B be two events within a probability space.
- $\mathbb{P}(A \mid B)$ is the probability of event A given that B has occurred.
- $\mathbb{P}(B \mid A)$ is the probability of event B given that A has occurred.
- $\mathbb{P}(A)$ and $\mathbb{P}(B)$ are the probabilities of A and B occurring independently.

Proof

1. **Start with the Definition of Conditional Probability:**

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

   This formula states that the probability of A given B is the probability of both A and B occurring divided by the probability of B occurring.

2. **Express $\mathbb{P}(A \cap B)$ in Terms of $\mathbb{P}(B \mid A)$:**

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

   Here, we use the definition of conditional probability again, but in reverse. The probability of A and B occurring together is the probability of B occurring given A times the probability of A occurring.

3. **Substitute $\mathbb{P}(A \cap B)$ Back into the Conditional Probability:**

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

   This step substitutes the expression from step 2 into the conditional probability formula from step 1, yielding the simple form of Bayes' Theorem.

This theorem is fundamentally important in Bayesian statistics as it provides a mathematical basis for updating beliefs in light of new evidence.

ASTRA

BOCCONI