Probability and Statistics Notes

By

Daniel Ruiz

November 2019

Prelude

This set of notes is a brief exposition on key concepts in Probability and Statistics. These notes are one of many that I have decided to polish and make available for anyone interested. One of the benefits of this series are having a compendium of definitions, examples and personal thoughts that I can always refer back to if I need a reminder on a particular topic / concept. In addition, I find that this medium reduces the search time for specific definitions, theorems, examples etc and thus aids in reinforcing my own knowledge when frequented.

Contents

Chapter 1

Probability Theory

1.1 Constructing a Probability Space

In developing a foundation for probability, one needs to first establish the fundamental structures that comprise a probability space. We will consider some set Ω such that its points ω are associated with possible outcomes of a measurement. We also denote $\mathcal{A}$ to be a nonempty collection of subsets of Ω which will represent collection of *events* that will be assigned probabilities.

Definition 1.1: Sample Space, Ω

A set Ω with outcomes $s_1, s_2, ..s_n$ (i.e. $\Omega = \{s_1, s_2, \ldots, s_n\}$) must meet some conditions in order to be a sample space:

1. *The outcomes must be mutually exclusive, i.e. if s_j takes place, then no other s_i will take place, $\forall i, j \in \{1, 2, \ldots, n\} \quad i \neq j$.*
2. *The outcomes must be collectively exhaustive, i.e., on every experiment (or random trial) there will always take place some outcome $s_i \in \Omega$ for $i \in \{1, 2, \ldots, n\}$.*
3. *The sample space Ω must have the right granularity depending on what we are interested in. We must remove irrelevant information from the sample space. In other words, we must choose the right abstraction (forget some irrelevant information).*

Definition 1.2: σ-algebra / σ-field $\mathcal{A}$ [Event Space]

A non-empty collection of subsets $\mathcal{A}$ of set Ω is called a σ-field of subsets of Ω provided that the following two properties hold:

1. *If A is in $\mathcal{A}$, then A^c is also in $\mathcal{A}$.*
2. *If A_n is in $\mathcal{A}$, $n = 1, 2, ...$, then $\cup_{n=1}^{\infty} A_n$ and $\cap_{n=1}^{\infty} A_n$ are both in $\mathcal{A}$.*

Definition 1.3: Event

Given a σ-field $\mathcal{A}$ that corresponds to some sample space Ω. We say that if $A \in \mathcal{A}$, then A is an event.

The statement 'the event A occurs' means that the outcome of our experiment is represented by some point $\omega \in A$. For an event A, if we let $P(A)$ denote the probability of the event, then we have $0 \leq P(A) \leq 1$.

Definition 1.4: Probability Measure

A probability measure P on a σ-field of subsets $\mathcal{A}$ of a set Ω is a real valued function having domain $\mathcal{A}$ satisfying the following properties:

(i) $P(\Omega) = 1$
(ii) $P(A) \geq 0 \quad \forall A \in \mathcal{A}$
(iii) If A_n, $n = 1, 2, 3, ...$ *are mutually disjoint sets in* $\mathcal{A}$, *then* $P\big(\cup_{n=1}^{\infty} A_n\big) = \sum_{n=1}^{\infty} P(A_n)$.

Definition 1.5: Probability Space

A probability space, denoted by $(\Omega, \mathcal{A}, P)$ *is a set* Ω, *a* σ*-field of subsets* $\mathcal{A}$, *and probability measure* P *defined on* $\mathcal{A}$.

1.2 Standard Definitions and Properties

Definition 1.6: Conditional Probability

Let A *and* B *be two events such that* $P(A) > 0$. *Then the conditional probability of* B *given* A, *written* $P(B|A)$, *is defined to be*

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{1.2.1}$$

If $P(A) = 0$, *then the conditional probability of* B *given* A *is undefined.*

Proposition 1.1

Let A *be an event and* A^c *be its complement, defined as* $A^c = \Omega - A$. *It follows from the properties of disjoint probability sets that*

$$P(A^c) = 1 - P(A) \tag{1.2.2}$$

Definition 1.7: Independent Events

Two events A *and* B *are independent if and only if*

$$P(A \cap B) = P(A)P(B) \tag{1.2.3}$$

This definition emerges as a consequence of wanting to construct a notion of an event's occurrence having no influence on the the occurrence of the other event. Through the conditional probabilistic lens, this would mean $P(B|A) = P(B)$ *(i.e. Given that* A *has occurred, this does not affect the probability that* B *will occur). Therefore, it follows that* $P(A \cap B) = P(A)P(B)$.

Definition 1.8: Mutual Exclusivity

Events A *and* B *are said to be two mutually exclusive events if both cannot occur. In essence, their intersection is disjoint* $A \cap B = \emptyset$ *so that they have the following properties:*

$$P(A \cap B) = 0 \tag{1.2.4}$$

$$P(A \cup B) = P(A) + P(B) \tag{1.2.5}$$

Definition 1.9: Discrete Random Variable

A discrete real-valued random variable X on a probability space $(\Omega, \mathcal{A}, P)$ is a function X with domain Ω and range that is a finite or countably infinite subset $\{x_1, x_2, ...\}$ of the real numbers $\mathbb{R}$ such that $\{\omega : X(\omega) = x_i\}$ is event for all i.

Hence, $\{\omega : X(\omega) = x_i\}$ is an event and we usually will write $\{X = x_i\}$ for brevity and denote the probability of this event as $P(X = x_i)$.

Definition 1.10: Discrete Density Function

The real-valued function f defined on $\mathbb{R}$ by $f(x) = P(X = x)$ is called the discrete density function of X. A number x is called a possible value of X if $f(x) > 0$.

We note that a real-valued function f defined on $\mathbb{R}$ is called a discrete density function provided that it satisfies the following properties:

(i) $f(x) \geq 0$, $x \in \mathbb{R}$.
(ii) $\{x : f(x) \neq 0\}$ is a finite or countably infinite subset of $\mathbb{R}$. Let $\{x_1, x_2, ...\}$ denote this set. Then
(iii) $\sum_i f(x_i) = 1$.

We can compute the probability of X taking on value in some set A via

$$P(X \in A) = \sum_{x \in A} f(x) \tag{1.2.6}$$

Definition 1.11: Discrete r-dimensional Random Vector

We let $\mathbb{R}^r$ denote the collection of all r-tuples of real numbers. A point $\mathbf{x} = (x_1, x_2, ..., x_r)$ of $\mathbb{R}^r$ is usually called an r-dimensional vector. Thus for each $\omega \in \Omega$, the r values $X_1(\omega), .., X_r(\omega)$ define a point

$$X(\omega) = (X_1(\omega), X_2(\omega), ..., X_r(\omega)) \tag{1.2.7}$$

of $\mathbb{R}^r$. This defines an r-dimensional vector-valued function on Ω, $\mathbf{X} : \Omega \to \mathbb{R}^r$, which is usually written as $\mathbf{X} = (X_1, X_2, ..., X_r)$.

A discrete r-dimensional random vector $\mathbf{X}$ is a function $\mathbf{X}$ from Ω tp $\mathbb{R}^r$ taking on a finite or countably infinite number of values $\mathbf{x}_1, \mathbf{x}_2, ..$ such that

$$\{\omega : \mathbf{X}(\omega) = x_0\} \tag{1.2.8}$$

is an event for all i.

Definition 1.12: Discrete Density Function for Random Vector

The discrete density function f for the random vector $\mathbf{X}$ is defined by

$$f(x_1, ..., x_r) = P(X_1 = x_1, ..., X_r = x_r) \tag{1.2.9}$$

or equivalently

$$f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^r \tag{1.2.10}$$

The probability that $\mathbf{X}$ *belongs to the subset* A *of* $\mathbb{R}^r$ *can be found by using the analog of (1.2.6), namely*

$$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x}) \tag{1.2.11}$$

Definition 1.13: Mutually Independent Random Variables

Let $X_1, X_2, ..., X_r$ *be* r *discrete random variables having densities* $f_1, f_2, ..., f_r$ *respectively. These random variables are said to be mutually independent if their joint density function* f *is given by*

$$f(x_1, x_2, ..., x_r) = f_1(x_1) f_2(x_2) \cdots f_r(x_r) \tag{1.2.12}$$

Consider two independent discrete random variables having densities f_X *and* f_Y*, respectively, Then for any two subsets* A *and* B *of* R*, we have*

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B) \tag{1.2.13}$$

Definition 1.14: Probability Generating Function

Let X *be a non-negative integer-valued random variable. The probability generating function* Φ_X *of* X *is defined as*

$$\Phi_X(t) = \sum_{x=0}^{\infty} P(X = x) t^x = \sum_{x=0}^{\infty} f_X(x) t^x, \qquad -1 \leq t \leq 1 \tag{1.2.14}$$

Definition 1.15: Random Variable

A random variable X *on a probability space* $(\Omega, \mathcal{A}, P)$ *is a real-valued function* $X(\omega)$*,* $\omega \in \Omega$*, such that for* $-\infty < x < \infty$*,* $\{\omega | X(\omega) \leq x\}$ *is an event.*

Definition 1.16: Continuous Random Variable

A random variable X *is called a continuous random variable if*

$$P(X = x) = 0, \qquad -\infty < x < \infty \tag{1.2.15}$$

We can observe that X *is a continuous random variable if and only if its distribution function* F *is continuous at every* x*, that is,* F *is a continuous function.*

Definition 1.17: Symmetric Random Variable

A random variable X *is said to be symmetric if* X *and -*X *have the same distribution function.*

Definition 1.18: Median

For any probability distribution on the real line $\mathbb{R}$ *with cumulative distribution function* F*, regardless of whether it is any kind of continuous probability distribution, in particular an absolutely continuous distri-*

bution, or a discrete probability distribution, a median is by definition any real number m that satisfies the inequalities

$$P(X \leq m) = \frac{1}{2}, \qquad P(X \geq m) = \frac{1}{2} \tag{1.2.16}$$

1.3 Distributions and Densities

Let X and Y be two discrete random variables. For any real numbers x and y, the set $\{\omega | X(\omega) = x \text{ and } Y(\omega) = y\}$ is an event that we will usually denote by $\{X = x, Y = y\}$.

Definition 1.19: Joint Density and Marginal Density

Let $\mathbf{X} = (X_1, X_2, ..., X_r)$ be an $\mathbf{r}$-dimensional random vector with density f. Then the function f is usually called the joint density of the random variables $X_1, X_2, ..., X_r$. The density function of the random variable X_i is then called the i^{th} marginal density of $\mathbf{X}$ or of f.

Definition 1.20: (Cumulative) Distribution Function [Discrete]

The function $F(t)$, $-\infty < t < \infty$, defined by

$$F(t) = P(X \leq t) = \sum_{x \leq t} f(x), \qquad -\infty < t < \infty \tag{1.3.1}$$

is called the distribution function of the random variable X or of the density f. One immediate consequence of this is that it satisfies:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \tag{1.3.2}$$

Proposition 1.2

Let X and Y be independent, non-negative integer-valued random variables. Then

$$\Phi_{X+Y}(t) = \Phi_X(t)\Phi_Y(t) \tag{1.3.3}$$

Definition 1.21: (Cumulative) Distribution Function [Continuous]

The distribution function F of a random variable X is the function

$$F(x) = P(X \leq x), \qquad -\infty < x < \infty \tag{1.3.4}$$

Proposition 1.3: Properties of Distribution Functions

Not all functions can arise as distribution functions, for the latter must satisfy certain conditions. Let X be a random variable and let F be its distribution function. Then

(i) $0 \leq F(x) \leq 1$ for all x.
(ii) F is a non-decreasing function of x.
(iii) $F(-\infty) = 0$ and $F(+\infty) = 1$.
(iv) F(x+) = F(x) for all x. (F is a right-continuous function)

We note that a distribution function is any function F satisfying properties (i)-(iv).

Definition 1.22: Probability Density Function (PDF) / Density

A density function / PDF (with respect to integration) is a non-negative function f such that

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{1.3.5}$$

Note that if f is density function, then the function F defined by

$$F(x) = \int_{-\infty}^{x} f(y)\, dy, \qquad -\infty < x < \infty \tag{1.3.6}$$

is a continuous function satisfying properties (i)-(iv) in ***Prop*** *1.3.*

Definition 1.23: Uniform Density

Let Ω be a sample space with finite measure $Vol(\Omega) < \infty$. Then, a uniform density is a constant function f, such that

$$1 = \int_{\Omega} f dV \tag{1.3.7}$$

Hence, $f = 1/Vol(\Omega)$.

Example 1.1: Uniform Density / Distribution on a Real Line Interval

Let a and b be constants with $a < b$. The uniform density on the interval (a,b) is the density f defined by

$$f(x) = \begin{cases} (b-a)^{-1} & \text{for } a < x < b, \\ 0 & \text{elsewhere} \end{cases} \tag{1.3.8}$$

The distribution function corresponding to (1.3.8) is given by

$$F(x) = \begin{cases} 0 & x < a, \\ (x-a)/(b-a), & a \le x \le b, \\ 1, & x > b. \end{cases} \tag{1.3.9}$$

Definition 1.24: Binomial Density

Let $0 < p < 1$. Then, the real valued function f defined on $\mathbb{R}$ by

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, ..., n, \\ 0, & \text{elsewhere} \end{cases} \tag{1.3.10}$$

is called the binomial density with parameters n and p.

Definition 1.25: Geometric Density

Let $0 < p < 1$. Then the real valued function f defined on $\mathbb{R}$ by

$$f(x) = \begin{cases} p(1-p)^x, & x = 0, 1, 2, ..., \\ 0, & \text{elsewhere} \end{cases} \tag{1.3.11}$$

is a discrete density function called the geometric density with parameter p.

Definition 1.26: Poisson Density

Let $0 < p < 1$ and let λ be a positive number. Then, the real valued function f defined on $\mathbb{R}$ by

$$f(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, ..., \\ 0, & \text{elsewhere.} \end{cases} \tag{1.3.12}$$

is called the Poisson density with parameter λ.

Proposition 1.4: Binomial Theorem

Let $0 < p < 1$ and $x < n \in \mathbb{Z}$. Then, we have that

$$1 = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \tag{1.3.13}$$

Which follows from the binomial theorem

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} \tag{1.3.14}$$

Proposition 1.5

Let ϕ be a differentiable strictly increasing or strictly decreasing function on an interval I, and let $\phi(I)$ denote the range of ϕ and ϕ^{-1} the inverse function to ϕ. Let X be a continuous random variable having density f such that $f(x) = 0$ for $x \notin I$. Then $Y = \phi(X)$ has density g given by $g(y) = 0$ for $y \notin \phi(I)$ and

$$g(y) = f\big(\phi^{-1}(y)\big)\left|\frac{d}{dy}\phi^{-1}(y)\right|, \qquad y \in \phi(I) \tag{1.3.15}$$

It is a bit more suggestive to wrtie this in the following form:

$$g(y) = f(x)\left|\frac{dx}{dy}\right|, \qquad y \in \phi(I) \qquad \text{and} \qquad x = \phi^{-1}(y) \tag{1.3.16}$$

Definition 1.27: Cauchy Density

The following function f, is a density known as the Cauchy Density.

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty \tag{1.3.17}$$

The corresponding distribution function is given by

$$F(x) = \frac{1}{2} + \frac{1}{\pi} arctan(x), \qquad -\infty < x < \infty \tag{1.3.18}$$

Definition 1.28: Symmetric Density

A density function f is called symmetric if $f(-x) = f(x)$ for all x. The Cauchy density and the uniform density on $(-a, a)$ are both symmetric.

Proposition 1.6

Let X be a random variable that has a density. Then f has a symmetric density if and only if X is a symmetric random variable.

Definition 1.29: Standard Normal Density

The following density, ϕ

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty \tag{1.3.19}$$

The standard normal density is clearly symmetric.

The normal density with mean μ and variance σ^2 is often denoted by $n(\mu, \sigma^2)$ or $n(y; \mu, \sigma^2)$, $-\infty < y < \infty$. Thus,

$$n(y; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)}, \qquad -\infty < y < \infty \tag{1.3.20}$$

Definition 1.30: Exponential Density

The exponential density with parameter λ is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{1.3.21}$$

The corresponding distribution function is

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{1.3.22}$$

Proposition 1.7

Let X be a random variable such that the following holds:

$$P(X > a + b) = P(X > a)P(X > b), \qquad a \geq 0 \qquad \text{and} \qquad b \geq 0 \tag{1.3.23}$$

Then either $P(X > 0) = 0$ or X is exponentially distributed.

Proposition 1.8: Sum of Random Variables

Let X, Y be continuous random variables with densities f_X and f_Y respectively. Then, the random variable $Z = X + Y$ has density f_Z, given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-k) f_Y(k)\ dk \tag{1.3.24}$$

1.4 Expectations

Notation: Let $\mathbf{X}$ be a discrete r-dimensional random vector having possible values $\mathbf{x}_1, \mathbf{x}_2, ...$ and density f, and let ϕ be an real-valued function defined on $\mathbb{R}^r$. Then $\sum_{\mathbf{x}} \phi(\mathbf{x}) f(\mathbf{x})$ is defined as

$$\sum_{\mathbf{x}} \phi(\mathbf{x}) f(\mathbf{x}) = \sum_{j} \phi(\mathbf{x}_j) f(\mathbf{x}_j) \tag{1.4.1}$$

Definition 1.31: Expectation Value

Let X be any discrete random variable that assumes a finite number of values $x_1, ..., x_r$. Then the expected value of X, denoted by EX, $E[X]$ or μ, is the number

$$E[X] = \sum_{i=1}^{r} x_i f(x_i) \tag{1.4.2}$$

The expected value $E[X]$ is also called the mean of X.

Definition 1.32: Finite / Undefined Expectation

Let X be a discrete random variable having density f. If $\sum_j |x_j| f(x_j) < \infty$, then we say that X has finite expectation and we define its expectation by (1.4.2). On the other hand if $\sum_j |x_j| f(x_j) = \infty$, then we say X does not have finite expectation and $E[X]$ is undefined.

Proposition 1.9

Let $\mathbf{X}$ be a discrete random vector having density f, and let ϕ be a real-valued function defined on $\mathbb{R}^r$. Then the random variable $Z = \phi(\mathbf{X})$ has finite expectation if and only if

$$\sum_{\mathbf{x}} |\phi(\mathbf{x})| f(\mathbf{x}) < \infty \tag{1.4.3}$$

and, when (1.4.3) holds,

$$E[Z] = \sum_{\mathbf{x}} \phi(\mathbf{x}) f(\mathbf{x}) \tag{1.4.4}$$

Proposition 1.10: Properties of Expectation Operator

Let X and Y be two random variables having finite expectation.

(i) If c is a constant and $P(X = c) = 1$, then $E[X] = c$.

(ii) ***Linearity:***
a) If c is a constant, then cX has finite expectation and $E[cX] = cE[X]$.
b) $X + Y$ has finite expectation and[a]

$$E[X + Y] = E[X] + E[Y] \tag{1.4.6}$$

(iii) Suppose that $P(X \geq Y) = 1$. Then $E[X] \geq E[Y]$; moreover, $E[X] = E[Y]$ if and only if $P(X = Y) = 1$.
(iv) $|E[X]| \leq E[|X|]$.

[a] More explicitly, we note that these are expectations w.r.t different densities:

$$E_{X+Y}[X + Y] = E_X[X] + E_Y[Y] \tag{1.4.5}$$

Proposition 1.11

Let X be a random variable such that for some constant M, $P(|X| \leq M) = 1$. Then X has finite expectation and $|E[X]| \leq M$.

Proposition 1.12

Let X and Y be two independent random variables having finite expectations. Then XY has finite expectation and

$$E[XY] = E[X]E[Y] \tag{1.4.7}$$

Proposition 1.13

Let X be a non-negative integer-valued random variable. Then X has finite expectation if and only if the series $\sum_{x=1}^{\infty} P(X \geq x)$ converges. If the series does converge, then

$$E[X] = \sum_{x=1}^{\infty} P(X \geq x) \tag{1.4.8}$$

Definition 1.33: Moments / Central Moments

Let X be a discrete random variable, and let $r \geq 0$ be an integer. We say that X has a moment of order r if X^r has finite expectation. In that case we define the r^{th} of X as $E[X^r]$.

If X has a moment of order r then the r^{th} moment of $X - \mu$, where μ is the mean of X, is called the central moment (or the r^{th} about the mean) of X.

Proposition 1.14

If the random variables X and Y have moments of order r, then $X + Y$ also has a moment of order r.

Definition 1.34: Variance

Let X be a random variable having a finite second moment. Then the variance of X, denoted by Var[X] or V[X], is defined by

$$Var[X] = E[(X - E[X])^2] \tag{1.4.9}$$

Through expanding, this works out to the following:

$$Var[X] = E[X^2] - (E[X])^2 \tag{1.4.10}$$

Definition 1.35: Standard Deviation

We often denote Var X by σ^2. The non-negative number $\sigma = \sqrt{VarX}$ is called the standard deviation of X or of f_X.

Definition 1.36: Covariance

Let X and Y be two random variables each having finite second moment. We define a quantity called the covariance of X and Y written as $Cov(X, Y)$. Thus we have the formula

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y] \tag{1.4.11}$$

We note that $X + Y$ has a finite second moment and finite variance. We therefore have an important formula:

$$Var[X + Y] = Var[X] + Var[Y] + 2\ Cov(X, Y) \tag{1.4.12}$$

Definition 1.37: Correlation Coefficient

Let X and Y be two random variables having finite nonzero variances. One measure of the degree of dependence between the two random variables is the correlation coefficient $\rho(X, Y)$ defined by

$$\rho = \rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{(Var[X])(Var[Y])}} \tag{1.4.13}$$

These random variables are said to be uncorrelated if $\rho = 0$. We can automatically see that independent random variables are uncorrelated. However, it is possible for dependent random variables to be uncorrelated as well. We observe that the correlation coefficient ρ is always between -1 and 1, and that $\rho = 1$ if and only if $P(X = aY) = 1$ for some constant a.

Definition 1.38: Cross-Correlation Matrix

Let $\boldsymbol{X}$, $\boldsymbol{Y}$ be random vectors. Then, we define the cross-correlation matrix by $E[\boldsymbol{X}\boldsymbol{Y}^T]$, where the matrix elements in the standard basis are given by $\left|E[\boldsymbol{X}\boldsymbol{Y}^T]\right|_{ij} = E[x_i y_j]$.

Theorem 1.1: The Schwartz Inequality

Let X and Y have finite second moments. Then

$$[E[XY]]^2 \leq (E[X^2])(E[Y^2]) \tag{1.4.14}$$

Furthermore, equality holds in (1.4.14) if and only if either P(Y=0) = 1 or P(X = aY) = 1 for some constant a.

Proposition 1.15: Chebyshev's Inequality

Let X be a random variable with mean μ and finite variance σ^2. Then for any real number $t > 0$

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \tag{1.4.15}$$

Theorem 1.2: Weak Law of Large Numbers

Let $X_1, X_2, ..., X_n$ be independent random variables having acommon distribution with finite mean μ and set $S_n = X_1 + ... + X_n$. Then for any $\delta > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \delta\right) = 0 \tag{1.4.16}$$

1.5 Jointly Distributed Random Variables

Definition 1.39: Joint Distribution Function

Let X and Y be two random variables defined on the same probability space. Their joint distribution function F is defined by

$$F(x, y) = P(X \leq x, Y \leq y), \qquad -\infty < x, y < \infty \tag{1.5.1}$$

Definition 1.40: Marginal Distribution Functions

The one-dimensional distribution functions F_X and F_Y defined by

$$F_X(x) = P(X \leq x) \qquad \text{and} \qquad F_Y(y) = P(Y \leq y) \tag{1.5.2}$$

are called the marginal distribution functions of X and Y. They are related to the joint distribution function F by

$$F_X(x) = F(x, \infty) = \lim_{y\to\infty} F(x, y) \tag{1.5.3}$$

$$F_Y(y) = F(\infty, y) = \lim_{x\to\infty} F(x, y) \tag{1.5.4}$$

Definition 1.41: Joint Density Function

If there is a nonnegative function f such that

$$F(x, y) = \int_{-\infty}^{x} \left(\int_{-\infty}^{y} f(u, v)dv \right) du, \qquad -\infty < x, y < \infty, \tag{1.5.5}$$

then f is called a joint density function (with respect to integration) for the distribution function F or the pair of random variables X, Y.

$$P((X, Y) \in A) = \int_A \int f(x, y)\, dx\, dy \tag{1.5.6}$$

By letting A be the entire plane we obtain

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ dx\ dy = 1 \tag{1.5.7}$$

Definition 1.42: Marginal Density

Let F be the distribution function for a pair of random variables X, Y. Then, the marginal distribution F_X has marginal density f_X given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\ dy \tag{1.5.8}$$

Similarly, F_Y has marginal density f_Y given by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\ dx \tag{1.5.9}$$

We note that it satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\ du \tag{1.5.10}$$

We can observe that

$$\frac{\partial^2}{\partial x\ \partial y}F(x,y) = f(x,y) \tag{1.5.11}$$

Definition 1.43: Independent Random Variables

The variables X and Y are called independent random variables if whenever $a \leq b$ and $c \leq d$, then

$$P(a < X \leq b, c < Y \leq d) = P(a < X \leq b)P(c < Y \leq d) \tag{1.5.12}$$

By letting $a = c = -\infty$, $b = x$, and $d = y$, it follows that if X and Y are independent, then

$$F(x,y) = F_X(x)F_Y(y), -\infty < x, y < \infty \tag{1.5.13}$$

Proposition 1.16

If X and Y are independent and A and B are unions of a finite or countably infinite number of intervals, then

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \tag{1.5.14}$$

In other words, the events

$$\{\omega | X(\omega) \in A\} \qquad \text{and} \qquad \{\omega | X(\omega) \in B\} \tag{1.5.15}$$

are independent events.

Proposition 1.17

Let X and Y be random variables having marginal densities f_X and f_Y. Then X and Y are independent if and only if the function f defined by

$$f(x,y) = f_X(x)f_Y(y), \qquad -\infty < x, y < \infty \tag{1.5.16}$$

is a joint density for X and Y.

Definition 1.44: Bivariate Density Function

A two-dimensional (or bivariate)density function f is a non-negative function on $\mathbb{R}^2$ such that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ dx\ dy = 1 \tag{1.5.17}$$

Definition 1.45: Standard Bivariate Normal Density

The density given below by f is referred to as the standard bivariate normal density.

$$f(x,y) = \frac{1}{2\pi}e^{-(x^2+y^2)/2}, \qquad -\infty < x,y < \infty \tag{1.5.18}$$

Proposition 1.18

Let X and Y be random variables having joint density f. In many contexts, we will have a random variable Z defined in terms of X and Y and we wish to calculate the density of Z. Let $Z = \phi(X,Y)$, where ϕ is a real-valued function whose domains contains the range of X and Y. For fixed z the event $\{Z \le z\}$ is equivalent to the event $\{(X,Y) \in A_z\}$ where A_z is the subset of $\mathbb{R}^2$ defined by

$$A_z = \{(x,y)|\phi(x,y) \le z\} \tag{1.5.19}$$

Thus,

$$\begin{aligned} F_Z(z) &= P(Z \le z) && (1.5.20)\\ &= P((X,Y) \in A_z) && (1.5.21)\\ &= \int_{A_z}\int f(x,y)\ dx\ dy && (1.5.22) \end{aligned}$$

If we can find a non-negative function g such that

$$\int_{A_z}\int f(x,y)\ dx\ dy = \int_{-\infty}^{z} g(v)\ dv, \qquad -\infty < z < \infty \tag{1.5.23}$$

then g is necessarily a density of Z.

Proposition 1.19

Let X and Y be independent random variables having the respective normal densities $n(\mu_1,\sigma_1^2)$ and $n(\mu_2,\sigma_2^2)$. Then X + Y has the normal density

$$n(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2) \tag{1.5.24}$$

Definition 1.46: Conditional Density (Discrete)

Let X and Y be discrete random variables having joint density f. If x is a possible value of X, then

$$P(Y=y|X=x) = \frac{P(X=x,Y=y)}{P(X=x)} = \frac{f(x,y)}{f_X(x)} \tag{1.5.25}$$

The function $f_{Y|X}$ defined by

$$f_{Y|X}(y|x) = \begin{cases} \frac{f(x,y)}{f_X(x)}, & f_X(x) \neq 0 \\ 0, & f_X(x) = 0 \end{cases} \tag{1.5.26}$$

is called the conditional density of Y given x.

Definition 1.47: Conditional Density (Continuous)

Let X and Y be continuous random variables having joint density f. The conditional density $f_{Y|X}$ is defined by

$$f_{Y|X}(y|x) = \begin{cases} \frac{f(x,y)}{f_X(x)}, & 0 < f_X(x) < \infty, \\ 0 & \text{elsewhere.} \end{cases} \tag{1.5.27}$$

If f_X is continuous and $f_X(x) \neq 0$, we have

$$P(a \leq Y \leq b|X = x) = \frac{\int_a^b f(x,y)\ dy}{f_X(x)} \tag{1.5.28}$$

Proposition 1.20: Bayes Rule

Let X and Y be random variables with marginal densities f_X and f_Y respectively and conditional densities $f_{X|Y}$ and $f_{Y|X}$. We have the continuous analog to Bayes' rule given below:

$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(x)f_{Y|X}(y|x)\ dx} \tag{1.5.29}$$

Definition 1.48: Joint Distribution Function (Multivariate)

Let $X_1, ..., X_n$ be n random variables defined on a common probability space. Their joint distribution function F is defined by

$$F(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n), \qquad -\infty < x_1, ..., x_n < \infty \tag{1.5.30}$$

Definition 1.49: Marginal Distribution Function (Multivariate)

The marginal distribution functions $F_{X_m}, m = 1, ..., n$ are defined by

$$F_{X_m}(x_m) = P(X_m \leq x_m), \qquad -\infty < x_m < \infty \tag{1.5.31}$$

The value of $F_{X_m}(x_m)$ can be obtained from F by letting $x_1, ..., x_{m-1}, x_{m+1}, ..., x_n$ all approach $+\infty$.

Definition 1.50: Joint Density Function (Multivariate)

A non-negative function f is called a joint density function (with respect to integration) for the joint distribution function F, or for the random variables $X_1, ..., X_n$ if

$$F(x_1, ..., x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(u_1, ..., u_n)\ du_1 \cdots du_n, \qquad -\infty < x_1, ..., x_n < \infty \tag{1.5.32}$$

We also note that

$$f(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(x_1, ..., x_n) \tag{1.5.33}$$

is valid at the continuity points of F.

Definition 1.51: Marginal Density Function (Multivariate)

The random variable X_m has the marginal density f_{X_m} obtained by integrating f over the remaining $n-1$ variables. For example,

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_n)\ dx_1\ dx_3 \cdots dx_n \tag{1.5.34}$$

Definition 1.52: Independent Random Variables (Multivariate)

In general, the random variables $X_1, ..., X_n$ are called independent whenever $a_m \leq b_m$ for $m = 1, ..., n$, then

$$P(a_1 < X_1 \leq b_1, ..., a_n < X_n \leq b_n) = P(a_1 < X_1 \leq b_1) \cdots P(a_n < X_n \leq b_n) \tag{1.5.35}$$

Proposition 1.21

A necessary and sufficient condition for independence is that

$$F(x_1, ..., x_n) = F_{x_1}(x_1) \cdots F_{X_n}(x_n), \qquad -\infty < x_1, ..., x_n < \infty \tag{1.5.36}$$

If F has a density f, then $X_1, ..., X_n$ are independent if and only if f can be chosen so that

$$f(x_1, ..., x_n) = f_{x_1}(x_1) \cdots f_{X_n}(x_n), \qquad -\infty < x_1, ..., x_n < \infty \tag{1.5.37}$$

If $X_1, ..., X_n$ are random variables whose joint density is given by (1.5.37) then $X_1, ..., X_n$ are independent and X_m has the marginal density f_m.

Proposition 1.22

Let $X_1, ..., X_n$ be independent random variables. Let Y be a random variable defined in terms of $X_1, ..., X_m$ and let Z be a random variable defined in terms $X_{m+1}, ..., X_n$ (where $1 < m < n$). Then Y and Z are independent.

Definition 1.53: Conditional Density (Multivariate)

If $X_1, ..., X_n$ has a joint density f, then any subcollection of these random variables has a joint density which can be found by integrating over the remaining variables. For example, if $1 \leq m < n$,

$$f_{X_1,...,X_m}(x_1, ..., x_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_n) dx_{m+1} \cdots dx_n \tag{1.5.38}$$

The conditional density of a subcollection of $X_1, ..., X_n$ given the remaining variables can also be defined in an obvious manner. Thus the conditional density of $X_{m+1}, ..., X_n$ given $X_1, ..., X_m$ is defined by

$$f_{X_{m+1},...,X_n|X_1,...,X_m}(x_{m+1}, ..., x_n|x_1, ..., x_m) = \frac{f(x_1, ..., x_n)}{f_{X_1,...,X_m}(x_1, ..., x_m)} \tag{1.5.39}$$

Where f is the joint density of $X_1, ..., X_n$.

Definition 1.54: Order Statistics

Let $U_1, ...U_n$ be independent continuous random variables, each having distribution F and density function f. Let $X_1, ..., X_n$ be random variables obtained by letting $X_1(\omega), ...X_n(\omega)$ be the set $U_1(\omega), ..., U_n(\omega)$ permuted so as to be in increasing order. In particular, we define X_1 and X_n to be the functions

$$X_1(\omega) = min\{U_1(\omega), ..., U_n(\omega)\} \tag{1.5.40}$$

and

$$X_n(\omega) = max\{U_1(\omega), ..., U_n(\omega)\} \tag{1.5.41}$$

The random variable X_k is called the k^{th} order statistic. Another related variable of interest is the range R, defined by

$$R(\omega) = X_n(\omega) - X_1(\omega) \tag{1.5.42}$$
$$= max\{U_1(\omega), ..., U_n(\omega)\} - min\{U_1(\omega), ..., U_n(\omega)\} \tag{1.5.43}$$

Proposition 1.23: Distributions for Order Statistics

Let $X_1, X_2, ..., X_n$ be identically distributed and independent random variables. Let their common CDF by denoted by F. We define $(X_{(1)}, X_{(2)}, ..., X_{(n)})$ as the vector of order statistics of $X_1, X_2, ..., X_n$. Then, the distribution for $X_{(k)}$ in a sample of size n is given by

$$F_{(k,n)}(x) = P(X_{(k)} \leq x) = \sum_{j=k}^{n} \binom{n}{k} [F(x)]^j [1 - F(x)]^{n-j} \tag{1.5.44}$$

Proof. We will break the event $(X_{(k)} \leq x)$ into disjoint sub-events given by

$$(X_{(k)} \leq x) = (X_{(n)} \leq x) \cup (X_{(n)} > x, X_{(n-1)} \leq x) \cup ... \cup (X_{(n)} > x, ..., X_{(k+1)} > x, X_{(k)} \leq x). \tag{1.5.45}$$

Recall from the property of probability measures that if $A_i's$ are all mutually disjoint sets, then $P(\cup_{i=1}^{l} A_i) = \sum_{i=1}^{l} P(A_i)$. Hence, it amounts to identify the probability of the events contained within each of these terms. Consider the event $(X_{(n)} > x, ..., X_{(j+1)} > x, X_{(j)} \leq x)$. This event tells us that the first j ordered variables have a value less than x while the rest have a value lying above x. Since the CDF, F(x) to each one tells us the probability of them occupying a value less than x, one can view this through the lens of fail / successes among n independent variables. In essence, we use the multiplicative property of independent events and combinatorics to establish the number of combinations one can arrange such an ordering of variables. We therefore have

$$P(X_{(n)} > x, ..., X_{(j+1)} > x, X_{(j)} \leq x) = \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j}. \tag{1.5.46}$$

Hence, it therefore follows that

$$F_{(k,n)}(x) = P(X_{(k)} \leq x) = \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j} \tag{1.5.47}$$

□

Theorem 1.3: Change of Variables

Let $X_1, ..., X_n$ be continuous random variables having joint density f and let random variables $Y_1, ..., Y_n$ be defined by

$$Y_i = \sum_{j=1}^{n} a_{ij} X_j, \qquad i = 1, ..., n, \tag{1.5.48}$$

where the matrix $A = [a_{ij}]$ has nonzero determinant $\det A$. Then $Y_1, ..., Y_n$ have joint density $f_{Y_1,...,Y_n}$ given by

$$f_{Y_1,...,Y_n}(y_1, ..., y_n) = \frac{1}{|det\ A|} f(x_1, ..., x_n), \tag{1.5.49}$$

where the x's are defined in terms of y's as the unique solution to the equations $y_i = \sum_{j=1}^{n} a_{ij} x_j$.

1.6 Expectations and the Central Limit Theorem

Definition 1.55: Expectation (Continuous)

Let X be a continuous random variable having density f. We say that X has finite expectation if

$$\int_{-\infty}^{\infty} |x| f(x)\ dx < \infty, \tag{1.6.1}$$

and in that case we define its expectation by

$$E[X] = \int_{-\infty}^{\infty} x f(x)\ dx \tag{1.6.2}$$

Theorem 1.4

Let $X_1, ..., X_n$ be continuous random variables having joint density f and let Z be a random variable defined in terms of $X_1, ..., X_n$ be $Z = \phi(X_1, ..., X_n)$. Then Z has finite expectation if and only if

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\phi(x_1, ..., x_n)| f(x_1, ..., x_n)\ dx_1 \cdots dx_n < \infty \tag{1.6.3}$$

in which case

$$E[Z] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(x_1, ..., x_n) f(x_1, ..., x_n)\ dx_1 \cdots dx_n < \infty \tag{1.6.4}$$

Definition 1.56: Moments (Continuous)

Let X be a continuous random variable having density f and mean μ. If X has finite m^{th} moment, then we have

$$E[X^m] = \int_{-\infty}^{\infty} x^m f(x)\ dx \tag{1.6.5}$$

If X has finite second moment, its variance σ^2 is given by

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\ dx \tag{1.6.6}$$

Definition 1.57: Conditional Expectation

Let X and Y be continuous random variables having joint density f and suppose that Y has finite expectation. Recall that we defined the conditional density of Y given $X = x$ by

$$f_{Y|X}(y|x) = \begin{cases} \frac{f(x,y)}{f_X(x)}, & 0 < f_X(x) < \infty, \\ 0 & \text{elsewhere.} \end{cases} \tag{1.6.7}$$

*For each x such that $0 < f_X(x) < \infty$ the function $f_{Y|X}(y|x)$, $-\infty < y < \infty$, is a density function with respect to **Def** 1.22. Thus we can talk about various moments of this density. Its mean is called the conditional expectation of Y given $X = x$ and is denoted by $E[Y|X = x]$ or $E[Y|x]$. Thus*

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f(y|x)\ dy \tag{1.6.8}$$

$$= \frac{\int_{-\infty}^{\infty} y f(x,y)\ dy}{f_X(x)} \tag{1.6.9}$$

when $0 < f_X(x) < \infty$. We define $E[Y|X = x] = 0$ elsewhere.

Proposition 1.24: Properties of Conditional Expectation

Let X, Y, Z be random variables and $a, b \in \mathbb{R}$, and $g : \mathbb{R} \to \mathbb{R}$. Assuming all the following expectations exist, we have that

(i) $E[a|Y] = a$
(ii) $E[aX + bY|Z] = aE[X|Y] + bE[Y|Z]$
(iii) $E[X|Y] \geq 0$ *if* $X \geq 0$.
(iv) $E[X|Y] = E[X]$ *if X and Y are independent.*
(v) $E[E[X|Y]] = E[X]$
(vi) $E[Xg(Y)|Y] = g(Y)E[X|Y]$. *In particular,* $E[g(Y)|Y] = g(Y)$.
(vii) $E[X|Y, g(Y)] = E[X|Y]$
(viii) $E[E[X|Y, Z]|Y] = E[X|Y]$

Definition 1.58: Regression Function

In statistics, the function m defined by $m(x) = E[Y|X = x]$ is called the regression function of Y on X.

Lemma 1.1

Let X be a random variable with density f_X. Then, $\frac{X-\alpha}{\beta}$ is a random variable with density

$$f_{(X-\alpha)/\beta}(z) = \beta f_X(\beta z + \alpha) \tag{1.6.10}$$

Theorem 1.5: Central Limit Theorem

Let $X_1, X_2, \ldots$ be independent, identically distributed random variables having mean μ and finite nonzero variance σ^2. Set $S_n = X_1 + \cdots + X_n$. Then

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x), \qquad -\infty < x < \infty \tag{1.6.11}$$

Where we recall that Φ is the CDF for the normal density:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\ du \tag{1.6.12}$$

Proof. Let $X_1, X_2, ...$ be independent, identically distributed random variables having mean μ and finite nonzero variance σ^2. Let their density be denoted by f. We define the random variable $S_n = X_n + S_{n-1}$, noting that its density is given by

$$f_{S_n}(x_n) = \int_{-\infty}^{\infty} \cdot\cdot \int_{-\infty}^{\infty} \mathrm{d}x_1\ \mathrm{d}x_2 \cdot\cdot \mathrm{d}x_{n-1} \Big(\prod_{i=1}^{n-1} f(x_{i+1} - x_i) \Big) f(x_1) \tag{1.6.13}$$

We define a new random variable $G_n := \frac{S_n - n\mu}{\sigma\sqrt{n}}$. Its density is given by

$$f_{G_n}(x_n) = \sigma\sqrt{n} \int_{-\infty}^{\infty} \cdot\cdot \int_{-\infty}^{\infty} \mathrm{d}x_1\ \mathrm{d}x_2 \cdot\cdot \mathrm{d}x_{n-1}\ f(\sigma\sqrt{n}x_n + n\mu - x_{n-1}) \Big(\prod_{i=1}^{n-2} f(x_{i+1} - x_i) \Big) f(x_1) \tag{1.6.14}$$

□

Chapter 2

Statistics Primer

Data is comprised of measurements (observations) $x_1, x_2, ..., x_n$ that are regarded as the realizations of random variables $X_1, ..., X_n$. Usually, we'll have $X_i \in \mathbb{R}$ (for $i = 1, ..., n$) but they can also be vector-valued $X_i \in \mathbb{R}^k$.

Definition 2.1: Observations

Let Ω denote a sample space. Then, on Ω, let $X = (X_1, .., X_n)$ denote a random vector. If $\omega \in \Omega$ represents the outcome of an experiment, then $X(\omega)$ is referred to as the observation or data. We remind ourselves that $X : \Omega \to \mathbb{R}^n$ where the notation $X(\omega)$ works element-wise:

$$X(\omega) = \big(X_1(\omega), X_2(\omega), ..., X_n(\omega)\big). \tag{2.0.1}$$

In addition, we usually refer to $X(\omega)$ as a realization of X.

Definition 2.2: Population

Let $\mathcal{T}$ denote a data set. Then, we define a population to be the set that contains all the elements of the data set. That is, if $\mathcal{P}$ denotes the population, then $\mathcal{P} = \mathcal{T}$.

Definition 2.3: Model

We note that X has an underlying probability distribution that is considered unknown to us. Instead, as we perform experiments, further computing $X(\omega)$, we gain a better sense of what this underlying distribution can be like. Generally, this distribution is assumed to be a member of a family $\mathcal{P}$ of probability distributions on $\mathbb{R}^n$. $\mathcal{P}$ is referred to as the model.

Definition 2.4: Parameterization / Parameter Space

Let $\mathcal{P}$ be a model. To describe $\mathcal{P}$, we use a parameterization which is a map $\theta \to P_\theta$ from a space of labels, called the parameter space Θ, to $\mathcal{P}$. In essence a parameterization is a map $\hat{\theta} : \Theta \to \mathcal{P}$ such that we can express the model by

$$\mathcal{P} = \{P_\theta : \ \theta \in \Theta\} \tag{2.0.2}$$

Definition 2.5: Parametric Model

Let $\mathcal{P}$ be a model and Θ be a parameter space that is a 'nice' subset of Euclidean space. Then, if the map $\theta \to P_\theta$ is smooth, we would call the model $\mathcal{P}$ parametric.

Definition 2.6: Identifiable Parameterization

Let $\mathcal{P}$ be a model. There are numerous choices in how we choose our parameterizations, but concerns arise if they are not one-to-one. Suppose that $\hat{\theta}$ is a parameterization of $\mathcal{P}$. Let $\hat{\theta}$ be one-to-one, that is $\theta_1 \neq \theta_2$ implies $P_{\theta_1} \neq P_{\theta_2}$. Then, we say that $\hat{\theta}$ is an identifiable parameterization.

A parameterization that isn't one-to-one is said to be unidentifiable.

Definition 2.7: Parameter

Let $\mathcal{P}$ be a model. We define a map ν from $\mathcal{P}$ to another space $\mathcal{N}$. Then, we call ν a parameter, which is a feature $\nu(P)$ of the distribution of X.

A function, $q : \Theta \to \mathcal{N}$ can be identified with a parameter $\nu(P)$ iff $P_{\theta_1} = P_{\theta_2}$ implies $q(\theta_1) = q(\theta_2)$ and then $\nu(P_\theta) \equiv q(\theta)$.

Definition 2.8: Statistic

Formally, a statistic T is a map from some sample space $\mathcal{X}$ to some space of values $\mathcal{T}$, which is usually a Euclidean space.

Definition 2.9: Action Space

An action space, $\mathcal{A}$ of actions or decisions or claims that we can contemplate making. Some common examples including making a prediction, hence $\mathcal{A}$ would be a set of predictions.

Definition 2.10: Decision Rule

Let $\mathcal{X}$ denote an outcome / sample space and $\mathcal{A}$ denote an action space. Then, data would be a point $X = x \in \mathcal{X}$. We define a decision rule or procedure δ to be any function from the sample space, taking its values in $\mathcal{A}$. We would express this by $\delta : \mathcal{X} \to \mathcal{A}$, so that if δ is used and $X = x$ is observed, then the statistician takes action $\delta(x)$.

Definition 2.11: Loss Function

Let $\mathcal{P}$ be a model and $\mathcal{A}$ be an action space. We define a loss function as a function $l : \mathcal{P} \times \mathcal{A} \to \mathbb{R}^+$. The interpretation of $l(P, a)$ (or $l(\theta, a)$ if $\mathcal{P}$ is parameterized) is the non-negative loss incurred by the statistician if he or she takes action a and the true 'state of Nature', that is, the probability distribution producing the data, is P.

Definition 2.12: Risk Function

Let δ be a decision rule, l a loss function and θ be the true value of a parameter. If $X = x$ is the outcome of an experiment, the loss would be given by $l\big(P, \delta(x)\big)$. However, we don't know the value of l as P us unknown to us. We therefore regard $l\big(P, \delta(x)\big)$ as a random variable and define the risk function, R by

$$R(P, \delta) = E_P\big[l\big(P, \delta(x)\big)\big] \tag{2.0.3}$$

as a measure of performance over the decision rule $\delta(x)$.

Example 2.1: Mean Squared Error

Suppose that $\nu \equiv \nu(P)$ is a real parameter we want to estimate and $\hat{\nu} \equiv \hat{\nu}(X)$ is our estimator (Decision rule). Then using quadratic loss, our risk function is called the mean squared error (MSE) of $\hat{\nu}$, given by

$$MSE(\nu) = R(P, \nu) = E_P\left[\left(\hat{\nu}(X) - \nu(P)\right)^2\right] \tag{2.0.4}$$

Definition 2.13: Bias

Let $\hat{\nu}$ represent an estimator and ν be a parameter. Then, the bias of our estimator is defined by

$$Bias(\hat{\nu}) = E[\hat{\nu}] - \nu \tag{2.0.5}$$

This quantity is also referred to as the long-run average error of $\hat{\nu}$.

Definition 2.14: Consistency

Usually, we are also concerned with the behavior of an estimator as the amount of training data grows. In particular, we usually wish that, as the number of data points m in our dataset increases, our point estimates converge to the true value of the corresponding parameters. More formally, we would like that

$$plim_{m\to\infty}\hat{\theta}_m = \theta \tag{2.0.6}$$

The symbol plim indicates convergence in probability, meaning that for any $\epsilon > 0$, $P\left(|\hat{\theta}_m - \theta| > \epsilon\right) \to 0$ as $m \to \infty$. The condition described by (2.0.6) is known as ***consistency.*** *It is also sometimes referred to as weak consistency.*

Alphabetical Index