Probability for Statistics: The Bare Minimum Prerequisites

This note seeks to cover the very bare minimum of topics that students should be familiar with before starting the Probability for Statistics module. Most of you will have studied classical and/or axiomatic probability before, but given the various backgrounds it's worth being concrete on what students should know going in.

Real Analysis

Students should be familiar with the main ideas from real analysis, and *at the very least* be familiar with notions of

- Limits of sequences of real numbers;
- Limit Inferior ($\liminf$) and Limit Superior ($\limsup$) and their connection to limits;
- Convergence of sequences of real-valued vectors;
- Functions; injectivity and surjectivity; continuous and discontinuous functions; continuity from the left and right; Convex functions.
- The notion of open and closed subsets of real numbers.

For those wanting for a good self-study textbook for any of these topics I would recommend: *Mathematical Analysis*, by Malik and Arora.

Complex Analysis

We will need some tools from complex analysis when we study characteristic functions and subsequently the central limit theorem. Students should be familiar with the following concepts:

- Imaginary numbers, complex numbers, complex conjugates and properties.
- De Moivre's theorem.
- Limits and convergence of sequences of complex numbers.
- Fourier transforms and inverse Fourier transforms

For those wanting a good self-study textbook for these topics I recommend: *Introduction to Complex Analysis* by H. Priestley.

Classical Probability and Combinatorics

Many students will have already attended a course in classical probability theory beforehand, and will be very familiar with these basic concepts. Nonetheless, they are presented to consolidate terminology and notation.

- The set of all possible outcomes of an experiment is known as the **sample space** of the experiment, denoted by S.
- An **event** is a set consisting of possible outcomes of an experiment. Informally, any subset E of the sample space is known as an event. If the outcome of the experiment is contained in E, then we say that E has occurred.
- A probability P is a function which assigns real values to events in a way that satisfies the following **axioms of probability**:

1. Every probability is between 0 and 1, i.e.
$$0 \leq P(E) \leq 1.$$
2. The probability that at least one of the outcomes in S will occur is 1, i.e.
$$P(S) = 1.$$
3. For any sequence of mutually exclusive events $E_1, \ldots, E_n$, we have
$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i).$$

- A **permutation** is an arrangement of r objects from a pool of n, where *the order matters*. The number of such arrangements is given by $P(n, r)$, defined as
$$P(n, r) = \frac{n!}{(n-r)!}.$$
- A **combination** is an arrangement of r objects from a pool of n objects, where the *order does not matter*. The number of such arrangements is given by $C(n, r)$, defined as
$$C(n, r) = \frac{P(n)}{r!} = \frac{n!}{r!(n-r)!}$$

Conditional Probability

- **Bayes' rule** – For events A and B such that $P(B) > 0$, we have:
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$
Note that $P(A \cap B) = P(A)P(B \mid A) = P(A \mid B)P(B)$.
- Let $\{A_i \; i = 1, \ldots, n\}$ be a *partition of* S, i.e.
$$A_i \cap A_j = \emptyset, \quad i \neq j \quad \text{and} \quad \bigcup_{i=1}^{n} A_i = S.$$
Then we have $P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$.
- More generally,
$$P(A_k \mid B) = \frac{P(B \mid A_k)P(A_k)}{\sum_{i=1}^{n} P(B \mid A_i)P(A_i)}.$$
- Two events A and B are **independent** if and only if:
$$P(A \cap B) = P(A)P(B).$$

Random Variables

- A **random variable**, X, is a function which maps every element in a sample space to the real line. If X takes discrete values (e.g. outcomes of coin flips), then we say that X is **discrete**, otherwise, if X takes continuous values, such as the temperature in the room, then X is **continuous**.
- The **cumulative distribution function** F, is defined by
$$F(x) = P(X \leq x).$$
This is a monotonically non-decreasing function such that $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$. Clearly, $P(a < X \leq b) = F(b) - F(a)$.
- Given a continuous random variable X with CDF F, a function f is a **probability density function** if $F(x) = \int_{-\infty}^{x} f(y)\, dy$. Equivalently, $f(x) = \frac{dF(x)}{dx}$. Clearly, $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
- Given a discrete random variable X, a function f is a **probability mass function** if $P(X = x) = f(x)$. Clearly $f(x_j) \leq 1$ and $\sum_j f(x_j) = 1$. If F is the CDF of X then $F(x) = \sum_{x_i \leq x} f(x_i)$.

Expectations and Moments

The **expected value** of a random variable, also known as the mean value or the first moment, is often denoted by $\mathbb{E}[X]$ or μ. It can be considered the value that we would obtain by averaging the results of the experiment infinitely many times. If f is the probability density function or probability mass function of X, then it is computed as

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i f(x_i) \quad \text{and} \quad \mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx,$$

for discrete and continuous random variables, respectively. The expected value of a function of a random variable $g(X)$ is computed as follows:

$$\mathbb{E}[g(X)] = \sum_{i=1}^{n} g(x_i) f(x_i) \quad \text{and} \quad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx,$$

for discrete and continuous random variables, respectively. More generally, the k^{th} (non-centered) moment is the value of X^k that we expect to observe on average on infinitely many trials. It is computed as follows:

$$\mathbb{E}[X^k] = \sum_{i=1}^{n} x_i^k f(x_i) \quad \text{and} \quad \mathbb{E}[X^k] = \int_{-\infty}^{\infty} x^k f(x)\, dx,$$

for discrete and continuous random variables, respectively.

Let $\mathbf{1}_A$ be an indicator function for the set A, i.e.

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0, & \text{otherwise.} \end{cases}$$

then $\mathbb{E}[\mathbf{1}_A(X)] = \mathbb{P}[A]$.

Linearity of Expectations – If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then

$$\mathbb{E}\left[\sum_i a_i X_i\right] = \sum_i a_i \mathbb{E}[X_i].$$

In particular, if $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $\mathbb{E}[\overline{X}] = \mathbb{E}[X_1] = \mu$.

Expectations of Independent Random Variables – Let $X_1, \ldots, X_n$ be independent random variables. Then

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_i \mathbb{E}[X_i].$$

The variance of a random variable, often denoted by $\mathrm{Var}[X]$ or σ^2 is a measure of the spread of its probability distribution. It is determined as follows:

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2\,.$$

If a and b are constants, then $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$. The standard deviation of a random variable, often denoted by σ, is a measure of the spread of its distribution function which is compatible with the units of the actual random variable. It is determined as follows:

$$\sigma = \sqrt{\mathrm{Var}[X]}.$$

Linearity of Variance – If $X_1, \ldots, X_n$ are independent, and $a_1, \ldots, a_n$ are constants, then

$$\mathrm{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_i a_i^2 \mathrm{Var}[X_i].$$

In particular, if $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ where $X_1, \ldots, X_n$ are IID, then $\text{Var}[\overline{X}] = \frac{1}{n}\text{Var}[X_1] = \frac{\sigma^2}{n}$.

Characteristic Functions – A characteristic function $\psi(\omega)$ is derived from a probability density function $f(x)$ and is defined as:

$$\psi(\omega) = \sum_{i=1}^{n} e^{i\omega x_i} f(x_i) \quad \text{and} \quad \psi(\omega) = \int_{-\infty}^{\infty} e^{i\omega x} f(x)\, dx,$$

for a discrete and continuous random variable, respectively, and where i is the imaginary number. It is clear that $\psi(0) = 1$.

Euler's formula is the name given to the identity

$$e^{i\theta} = \cos(\theta) + i\sin(\theta).$$

Using Euler's formula we obtain

$$\psi(\omega) = \mathbb{E}[\cos(\omega X)] + i\mathbb{E}[\sin(\omega X)].$$

The k^{th} moment can also be computed with the characteristic function as follows:

$$\mathbb{E}[X^k] = (-i)^k \left[\frac{\partial^k \psi}{\partial \omega^k}\right]_{\omega=0}.$$

Distribution of a sum of independent random variables – Let $Y = X_1 + \ldots + X_n$ with $X_1, \ldots, X_n$ independent. We have:

$$\psi_Y(\omega) = \prod_{k=1}^{n} \psi_{X_k}(\omega).$$

Transformation of random variables – Now suppose that X and Y are continuous random variables which are linked by some one-to-one differentiable function r, e.g. Y = r(X). Denoting by f_X and f_Y the probability density function of X and Y respectively, then

$$f_Y(y) = f_X(x(y)) \left|\frac{dx}{dy}\right|.$$

Leibniz integral rule – Let g be a function and a parameter θ and boundaries $a = a(\theta)$ and $b = b(\theta)$, then

$$\frac{\partial}{\partial \theta}\left(\int_{a(\theta)}^{b(\theta)} g(x,\theta)\, dx\right) = \frac{\partial b(\theta)}{\partial \theta} \cdot g(b) - \frac{\partial a(\theta)}{\partial \theta} \cdot g(a) + \int_{a(\theta)}^{b(\theta)} \frac{\partial g}{\partial \theta}(x,\theta)\, dx.$$

Probability Distributions

Discrete distributions — The main discrete distributions to be familiar with:

Distribution	$\mathbb{P}[X = x]$	$\psi(\omega)$	$\mathbb{E}[X]$	$\text{Var}[X]$
$X \sim B(n,p)$	$\binom{n}{x} p^x (1-p)^{n-x}$	$(pe^{i\omega} + (1-p))^n$	np	$np(1-p)$
$X \sim \text{Poisson}(\lambda)$	$\frac{\lambda^x}{x!} e^{-\lambda}$	$e^{\lambda(e^{i\omega}-1)}$	λ	λ

Continuous distributions — The main continuous distributions to be familiar with:

Distribution	$f(x)$	$\psi(\omega)$	$\mathbb{E}[X]$	$\text{Var}[X]$
$X \sim \mathcal{U}(a,b)$	$\frac{1}{b-a}$	$\frac{e^{i\omega b} - e^{i\omega a}}{(b-a)i\omega}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$X \sim \mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$e^{i\omega\mu - \frac{1}{2}\omega^2\sigma^2}$	μ	σ^2
$X \sim \text{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$\frac{1}{1 - \frac{i\omega}{\lambda}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Jointly Distributed Random Variables

The joint probability mass function of two discrete random variables X and Y, denoted f_{XY} defined as

$$f_{XY}(x_i, y_j) = \mathbb{P}[X = x_i \text{ and } Y = y_j].$$

The joint probability density of two continuous random variables X and Y is defined by

$$f_{XY}(x, y)\Delta x \Delta y = \mathbb{P}[x \leq X \leq x + \Delta x \text{ and } y \leq Y \leq y + \Delta y].$$

Marginal density– We define the marginal distribution of the random variable X as follows:

$$f_X(x_i) = \sum_j f_{XY}(x_i, y_j) \text{ or } f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y)\, dy,$$

in the discrete and continuous case, respectively.

The **cumulative distribution function** for (X, Y) is defined as follows:

$$F_{XY}(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f_{XY}(x_i, y_j)$$

in the discrete case and

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x', y')\, dx'\, dy',$$

in the continuous case.

Conditional density – The conditional density of X with respect to Y, often denoted by $f_{X|Y}$, is defined as follows:

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

Independence – Two random variables X and Y are said to be independent if we have:

$$f_{XY}(x, y) = f_X(x) f_Y(y).$$

Moments of joint distributions – We define the moments of joint distributions of random variables X and Y as follows

$$\mathbb{E}[X^p Y^q] = \sum_i \sum_j x_i^p y_j^q f_{XY}(x_i, y_i) \quad \text{and} \quad \mathbb{E}[X^p Y^q] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f_{XY}(x, y)\, dx\, dy,$$

in the discrete and continuous case, respectively.

Covariance – We define the covariance of two random variables X and Y, that we note σ_{XY}^2 or $Cov(X, Y)$, as follows:

$$\text{Cov}(X, Y) = \sigma_{XY}^2 = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mu_Y,$$

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

Correlation – The correlation between the random variables X and Y, denoted ρ_{XY} is given by

$$\rho_{XY} = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y},$$

where σ_X, σ_Y are the standard deviations of X and Y. Clearly $-1 \leq \rho_{XY} \leq 1$ and $\rho_{XY} = 0$ if X and Y are independent.

Conditional Expectation – The conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X \,|Y = y] = \begin{cases} \sum_i x_i f_{X|Y}(x_i|y) & \text{discrete case} \\ \int x f_{X|Y}(x|y)\, dx & \text{continuous case.} \end{cases}$$

If $r(x, y)$ is a function of x and y:

$$\mathbb{E}[r(X,Y)\,|Y=y] = \begin{cases} \sum_i r(x_i, y) f_{X|Y}(x_i|y) & \text{discrete case} \\ \int r(x,y) f_{X|Y}(x|y)\,dx & \text{continuous case.} \end{cases}$$

For random variables X and Y, assuming the expectations exist, we have that

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y] \text{ and } \mathbb{E}\left[\mathbb{E}[X|Y]\right] = \mathbb{E}[X].$$

1 Important Inequalities

Concentration inequalities provide probability bounds on how a random variable deviates from its expectation.

Markov Inequality – Let X be any nonnegative integrable random variable, and $a > 0$,

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

Chebyshev's inequality – Let X be a random variable with expected value μ. For $a > 0$, we have the following inequality;

$$\mathbb{P}\left(|X - \mu| \geq a\right) \leq \frac{\mathrm{Var}[X]}{a^2}.$$

Hoeffding's inequality– Let $Y_1, \ldots, Y_n$ be independent observations such that $\mathbb{E}[Y_i] = 0$ and $a_i \leq Y_i \leq b_i$. Let $\epsilon > 0$. Then, for any $t > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n} Y_i \geq \epsilon\right] \leq e^{-t\epsilon} \prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8.}$$

Cauchy-Schwartz inequality – If X and Y have finite variances then

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

Jensen's inequality – Recall that a function g is convex if, for each x, y and $\alpha \in [0, 1]$,

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

Then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$. In particular $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$, and if X is positive then $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$.