Appendix B

Probability

B.1 Foundations

The set S of all possible outcomes of an experiment is called the **sample space** for the experiment. Take the simple example of tossing a coin. There are two outcomes, heads and tails, so we can write $S = \{H, T\}$. If two coins are tossed in sequence, we can write the four outcomes as $S = \{HH, HT, TH, TT\}$.

An **event** A is any collection of possible outcomes of an experiment. An event is a subset of S, including S itself and the null set $\emptyset$. Continuing the two coin example, one event is $A = \{HH, HT\}$, the event that the first coin is heads. We say that A and B are **disjoint** or **mutually exclusive** if $A \cap B = \emptyset$. For example, the sets $\{HH, HT\}$ and $\{TH\}$ are disjoint. Furthermore, if the sets $A_1, A_2, ...$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, ...$ is called a **partition** of S.

The following are elementary set operations:

Union: $A \cup B = \{x : x \in A \text{ or } x \in B\}$.

Intersection: $A \cap B = \{x : x \in A \text{ and } x \in B\}$.

Complement: $A^c = \{x : x \notin A\}$.

The following are useful properties of set operations.

Communtatitivity: $A \cup B = B \cup A$; $\quad A \cap B = B \cap A$.

Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$; $\quad A \cap (B \cap C) = (A \cap B) \cap C$.

Distributive Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $\quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

DeMorgan's Laws: $(A \cup B)^c = A^c \cap B^c$; $\quad (A \cap B)^c = A^c \cup B^c$.

A **probability function** assigns probabilities (numbers between 0 and 1) to events A in S. This is straightforward when S is countable; when S is uncountable we must be somewhat more careful. A set $\mathcal{B}$ is called a **sigma algebra** (or Borel field) if $\emptyset \in \mathcal{B}$, $A \in \mathcal{B}$ implies $A^c \in \mathcal{B}$, and $A_1, A_2, ... \in \mathcal{B}$ implies $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$. A simple example is $\{\emptyset, S\}$ which is known as the trivial sigma algebra. For any sample space S, let $\mathcal{B}$ be the smallest sigma algebra which contains all of the open sets in S. When S is countable, $\mathcal{B}$ is simply the collection of all subsets of S, including $\emptyset$ and S. When S is the real line, then $\mathcal{B}$ is the collection of all open and closed intervals. We call $\mathcal{B}$ the sigma algebra associated with S. We only define probabilities for events contained in $\mathcal{B}$.

We now can give the axiomatic definition of probability. Given S and $\mathcal{B}$, a probability function $\mathbb{P}$ satisfies $\mathbb{P}(S) = 1$, $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{B}$, and if $A_1, A_2, ... \in \mathcal{B}$ are pairwise disjoint, then $\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Some important properties of the probability function include the following

- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(A) \leq 1$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

- $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- If $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- Bonferroni's Inequality: $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$
- Boole's Inequality: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

For some elementary probability models, it is useful to have simple rules to count the number of objects in a set. These counting rules are facilitated by using the binomial coefficients which are defined for nonnegative integers n and r, $n \geq r$, as

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}.$$

When counting the number of objects in a set, there are two important distinctions. Counting may be **with replacement** or **without replacement**. Counting may be **ordered** or **unordered**. For example, consider a lottery where you pick six numbers from the set 1, 2, ..., 49. This selection is without replacement if you are not allowed to select the same number twice, and is with replacement if this is allowed. Counting is ordered or not depending on whether the sequential order of the numbers is relevant to winning the lottery. Depending on these two distinctions, we have four expressions for the number of objects (possible arrangements) of size r from n objects.

	Without Replacement	With Replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

In the lottery example, if counting is unordered and without replacement, the number of potential combinations is $\binom{49}{6} = 13,983,816$.

If $\mathbb{P}(B) > 0$ the **conditional probability** of the event A given the event B is

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

For any B, the conditional probability function is a valid probability function where S has been replaced by B. Rearranging the definition, we can write

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

which is often quite useful. We can say that the occurrence of B has no information about the likelihood of event A when $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, in which case we find

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \tag{B.1}$$

We say that the events A and B are **statistically independent** when (B.1) holds. Furthermore, we say that the collection of events $A_1, ..., A_k$ are **mutually independent** when for any subset $\{A_i : i \in I\}$,

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

Theorem 1 *(Bayes' Rule). For any set B and any partition $A_1, A_2, ...$ of the sample space, then for each $i = 1, 2, ...$*

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B \mid A_j)\,\mathbb{P}(A_j)}$$

B.2 Random Variables

A **random variable** X is a function from a sample space S into the real line. This induces a new sample space – the real line – and a new probability function on the real line. Typically, we denote random variables by uppercase letters such as X, and use lower case letters such as x for potential values and realized values. (This is in contrast to the notation adopted for most of the textbook.) For a random variable X we define its **cumulative distribution function** (CDF) as

$$F(x) = \mathbb{P}(X \leq x). \tag{B.2}$$

Sometimes we write this as $F_X(x)$ to denote that it is the CDF of X. A function $F(x)$ is a CDF if and only if the following three properties hold:

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$
2. $F(x)$ is nondecreasing in x
3. $F(x)$ is right-continuous

We say that the random variable X is **discrete** if $F(x)$ is a step function. In the latter case, the range of X consists of a countable set of real numbers $\tau_1, ..., \tau_r$. The probability function for X takes the form

$$\mathbb{P}(X = \tau_j) = \pi_j, \qquad j = 1, ..., r \tag{B.3}$$

where $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{r} \pi_j = 1$.

We say that the random variable X is **continuous** if $F(x)$ is continuous in x. In this case $\mathbb{P}(X = \tau) = 0$ for all $\tau \in R$ so the representation (B.3) is unavailable. Instead, we represent the relative probabilities by the **probability density function** (PDF)

$$f(x) = \frac{d}{dx}F(x)$$

so that

$$F(x) = \int_{-\infty}^{x} f(u)du$$

and

$$\mathbb{P}(a \leq X \leq b) = \int_{a}^{b} f(u)du.$$

These expressions only make sense if $F(x)$ is differentiable. While there are examples of continuous random variables which do not possess a PDF, these cases are unusual and are typically ignored.

A function $f(x)$ is a PDF if and only if $f(x) \geq 0$ for all $x \in R$ and $\int_{-\infty}^{\infty} f(x)dx$.

B.3 Expectation

For any measurable real function g, we define the **mean** or **expectation** $\mathbb{E}g(X)$ as follows. If X is discrete,

$$\mathbb{E}g(X) = \sum_{j=1}^{r} g(\tau_j)\pi_j,$$

and if X is continuous

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

The latter is well defined and finite if

$$\int_{-\infty}^{\infty} |g(x)|\, f(x)dx < \infty. \tag{B.4}$$

If (B.4) does not hold, evaluate

$$\begin{aligned} I_1 &= \int_{g(x)>0} g(x)f(x)dx \\ I_2 &= -\int_{g(x)<0} g(x)f(x)dx \end{aligned}$$

If $I_1 = \infty$ and $I_2 < \infty$ then we define $\mathbb{E}g(X) = \infty$. If $I_1 < \infty$ and $I_2 = \infty$ then we define $\mathbb{E}g(X) = -\infty$. If both $I_1 = \infty$ and $I_2 = \infty$ then $\mathbb{E}g(X)$ is undefined.

Since $\mathbb{E}(a + bX) = a + b\mathbb{E}X$, we say that expectation is a linear operator.

For $m > 0$, we define the $m'th$ **moment** of X as $\mathbb{E}X^m$ and the $m'th$ **central moment** as $\mathbb{E}(X - \mathbb{E}X)^m$.

Two special moments are the **mean** $\mu = \mathbb{E}X$ and **variance** $\sigma^2 = \mathbb{E}(X-\mu)^2 = \mathbb{E}X^2 - \mu^2$. We call $\sigma = \sqrt{\sigma^2}$ the **standard deviation** of X. We can also write $\sigma^2 = \text{var}(X)$. For example, this allows the convenient expression $\text{var}(a + bX) = b^2 \text{var}(X)$.

The **moment generating function** (MGF) of X is

$$M(\lambda) = \mathbb{E}\exp(\lambda X).$$

The MGF does not necessarily exist. However, when it does and $\mathbb{E}|X|^m < \infty$ then

$$\left.\frac{d^m}{d\lambda^m} M(\lambda)\right|_{\lambda=0} = \mathbb{E}(X^m)$$

which is why it is called the moment generating function.

More generally, the **characteristic function** (CF) of X is

$$C(\lambda) = \mathbb{E}\exp(\mathrm{i}\lambda X)$$

where $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. The CF always exists, and when $\mathbb{E}|X|^m < \infty$

$$\left.\frac{d^m}{d\lambda^m} C(\lambda)\right|_{\lambda=0} = \mathrm{i}^m \mathbb{E}(X^m).$$

The L^p **norm**, $p \geq 1$, of the random variable X is

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p}.$$

B.4 Gamma Function

The gamma function is defined for $\alpha > 0$ as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x).$$

It satisfies the property

$$\Gamma(1+\alpha) = \Gamma(\alpha)\alpha$$

so for positive integers n,

$$\Gamma(n) = (n-1)!$$

Special values include

$$\Gamma(1) = 1$$

and

$$\Gamma\left(\frac{1}{2}\right) = \pi^{1/2}.$$

Sterling's formula is an expansion for the its logarithm

$$\log\Gamma(\alpha) = \frac{1}{2}\log(2\pi) + \left(\alpha - \frac{1}{2}\right)\log\alpha - z + \frac{1}{12\alpha} - \frac{1}{360\alpha^3} + \frac{1}{1260\alpha^5} + \cdots$$

B.5 Common Distributions

For reference, we now list some important discrete distribution function.

Bernoulli

$$\begin{aligned} \mathbb{P}(X=x) &= p^x(1-p)^{1-x}, \qquad x=0,1; \qquad 0\le p\le 1 \\ \mathbb{E}X &= p \\ \operatorname{var}(X) &= p(1-p) \end{aligned}$$

Binomial

$$\begin{aligned} \mathbb{P}(X=x) &= \binom{n}{x} p^x(1-p)^{n-x}, \qquad x=0,1,...,n; \qquad 0\le p\le 1 \\ \mathbb{E}X &= np \\ \operatorname{var}(X) &= np(1-p) \end{aligned}$$

Geometric

$$\begin{aligned} \mathbb{P}(X=x) &= p(1-p)^{x-1}, \qquad x=1,2,...; \qquad 0\le p\le 1 \\ \mathbb{E}X &= \frac{1}{p} \\ \operatorname{var}(X) &= \frac{1-p}{p^2} \end{aligned}$$

Multinomial

$$\begin{aligned} \mathbb{P}(X_1=x_1, X_2=x_2, ..., X_m=x_m) &= \frac{n!}{x_1!x_2!\cdots x_m!} p_1^{x_1}p_2^{x_2}\cdots p_m^{x_m}, \\ x_1+\cdots+x_m &= n; \\ p_1+\cdots+p_m &= 1 \\ \mathbb{E}X_i &= p_i \\ \operatorname{var}(X_i) &= np_i(1-p_i) \\ \operatorname{cov}(X_i,X_j) &= -np_ip_j \end{aligned}$$

Negative Binomial

$$\begin{aligned} \mathbb{P}(X=x) &= \frac{\Gamma(r+x)}{x!\Gamma(r)} p^r(1-p)^{x-1}, \qquad x=0,1,2,...; \qquad 0\le p\le 1 \\ \mathbb{E}X &= \frac{r(1-p)}{p} \\ \operatorname{var}(X) &= \frac{r(1-p)}{p^2} \end{aligned}$$

Poisson

$$\begin{aligned} \mathbb{P}(X=x) &= \frac{\exp(-\lambda)\lambda^x}{x!}, \qquad x=0,1,2,..., \qquad \lambda>0 \\ \mathbb{E}X &= \lambda \\ \operatorname{var}(X) &= \lambda \end{aligned}$$

We now list some important continuous distributions.

Beta

$$\begin{aligned}
f(x) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \leq x \leq 1; \qquad \alpha > 0,\ \beta > 0 \\
\mu &= \frac{\alpha}{\alpha+\beta} \\
\operatorname{var}(X) &= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}
\end{aligned}$$

Cauchy

$$\begin{aligned}
f(x) &= \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty \\
\mathbb{E}X &= \infty \\
\operatorname{var}(X) &= \infty
\end{aligned}$$

Exponential

$$\begin{aligned}
f(x) &= \frac{1}{\theta} \exp\left(\frac{x}{\theta}\right), \qquad 0 \leq x < \infty; \qquad \theta > 0 \\
\mathbb{E}X &= \theta \\
\operatorname{var}(X) &= \theta^2
\end{aligned}$$

Logistic

$$\begin{aligned}
f(x) &= \frac{\exp(-x)}{(1+\exp(-x))^2}, \qquad -\infty < x < \infty; \\
\mathbb{E}X &= 0 \\
\operatorname{var}(X) &= \frac{\pi^2}{3}
\end{aligned}$$

Lognormal

$$\begin{aligned}
f(x) &= \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right), \qquad 0 \leq x < \infty; \qquad \sigma > 0 \\
\mathbb{E}X &= \exp\left(\mu + \sigma^2/2\right) \\
\operatorname{var}(X) &= \exp\left(2\mu + 2\sigma^2\right) - \exp\left(2\mu + \sigma^2\right)
\end{aligned}$$

Pareto

$$\begin{aligned}
f(x) &= \frac{\beta\alpha^\beta}{x^{\beta+1}}, \qquad \alpha \leq x < \infty, \qquad \alpha > 0, \qquad \beta > 0 \\
\mathbb{E}X &= \frac{\beta\alpha}{\beta-1}, \qquad \beta > 1 \\
\operatorname{var}(X) &= \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \qquad \beta > 2
\end{aligned}$$

Uniform

$$\begin{aligned}
f(x) &= \frac{1}{b-a}, \qquad a \leq x \leq b \\
\mathbb{E}X &= \frac{a+b}{2} \\
\operatorname{var}(X) &= \frac{(b-a)^2}{12}
\end{aligned}$$

Weibull

$$\begin{aligned} f(x) &= \frac{\gamma}{\beta}x^{\gamma-1}\exp\left(-\frac{x^{\gamma}}{\beta}\right), \qquad 0 \le x < \infty; \qquad \gamma > 0,\ \beta > 0 \\ \mathbb{E}X &= \beta^{1/\gamma}\Gamma\left(1+\frac{1}{\gamma}\right) \\ \mathrm{var}(X) &= \beta^{2/\gamma}\left(\Gamma\left(1+\frac{2}{\gamma}\right)-\Gamma^2\left(1+\frac{1}{\gamma}\right)\right) \end{aligned}$$

Gamma

$$\begin{aligned} f(x) &= \frac{1}{\Gamma(\alpha)\theta^{\alpha}}x^{\alpha-1}\exp\left(-\frac{x}{\theta}\right), \qquad 0 \le x < \infty; \qquad \alpha > 0,\ \theta > 0 \\ \mathbb{E}X &= \alpha\theta \\ \mathrm{var}(X) &= \alpha\theta^2 \end{aligned}$$

Chi-Square

$$\begin{aligned} f(x) &= \frac{1}{\Gamma(r/2)2^{r/2}}x^{r/2-1}\exp\left(-\frac{x}{2}\right), \qquad 0 \le x < \infty; \qquad r > 0 \\ \mathbb{E}X &= r \\ \mathrm{var}(X) &= 2r \end{aligned}$$

Normal

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty; \qquad -\infty < \mu < \infty,\ \sigma^2 > 0 \\ \mathbb{E}X &= \mu \\ \mathrm{var}(X) &= \sigma^2 \end{aligned}$$

Student t

$$\begin{aligned} f(x) &= \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}, \qquad -\infty < x < \infty; \qquad r > 0 \\ \mathbb{E}X &= 0 \text{ if } r > 1 \\ \mathrm{var}(X) &= \frac{r}{r-2} \text{ if } r > 2 \end{aligned}$$

B.6 Multivariate Random Variables

A pair of bivariate random variables (X, Y) is a function from the sample space into $\mathbb{R}^2$. The joint CDF of (X, Y) is

$$F(x, y) = \mathbb{P}(X \le x, Y \le y).$$

If F is continuous, the joint probability density function is

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

For a Borel measurable set $A \in R^2$,

$$\mathbb{P}((X < Y) \in A) = \int\int_A f(x, y)dxdy$$

For any measurable function $g(x,y)$,

$$\mathbb{E}g(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy.$$

The **marginal distribution** of X is

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \le x) \\ &= \lim_{y\to\infty} F(x,y) \\ &= \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(x,y)dydx \end{aligned}$$

so the **marginal density** of X is

$$f_X(x) = \frac{d}{dx}F_X(x) = \int_{-\infty}^{\infty} f(x,y)dy.$$

Similarly, the marginal density of Y is

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

The random variables X and Y are defined to be **independent** if $f(x,y) = f_X(x)f_Y(y)$. Furthermore, X and Y are independent if and only if there exist functions $g(x)$ and $h(y)$ such that $f(x,y) = g(x)h(y)$.

If X and Y are independent, then

$$\begin{aligned} \mathbb{E}\left(g(X)h(Y)\right) &= \int\int g(x)h(y)f(y,x)dydx \\ &= \int\int g(x)h(y)f_Y(y)f_X(x)dydx \\ &= \int g(x)f_X(x)dx \int h(y)f_Y(y)dy \\ &= \mathbb{E}g\left(X\right)\mathbb{E}h\left(Y\right). \end{aligned} \tag{B.5}$$

if the expectations exist. For example, if X and Y are independent then

$$\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y.$$

Another implication of (B.5) is that if X and Y are independent and $Z = X + Y$, then

$$\begin{aligned} M_Z(\lambda) &= \mathbb{E}\exp\left(\lambda\left(X+Y\right)\right) \\ &= \mathbb{E}\left(\exp\left(\lambda X\right)\exp\left(\lambda Y\right)\right) \\ &= \mathbb{E}\exp\left(\lambda' X\right)\mathbb{E}\exp\left(\lambda' Y\right) \\ &= M_X(\lambda)M_Y(\lambda). \end{aligned} \tag{B.6}$$

The covariance between X and Y is

$$\mathrm{cov}(X,Y) = \sigma_{XY} = \mathbb{E}\left(\left(X - \mathbb{E}X\right)\left(Y - \mathbb{E}Y\right)\right) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y.$$

The correlation between X and Y is

$$\mathrm{corr}\left(X,Y\right) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_x\sigma_Y}.$$

The Cauchy-Schwarz Inequality implies that

$$|\rho_{XY}| \leq 1. \tag{B.7}$$

The correlation is a measure of linear dependence, free of units of measurement.

If X and Y are independent, then $\sigma_{XY} = 0$ and $\rho_{XY} = 0$. The reverse, however, is not true. For example, if $\mathbb{E}X = 0$ and $\mathbb{E}X^3 = 0$, then $\operatorname{cov}(X, X^2) = 0$.

A useful fact is that

$$\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X,Y).$$

An implication is that if X and Y are independent, then

$$\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y),$$

the variance of the sum is the sum of the variances.

A $k \times 1$ random vector $\boldsymbol{X} = (X_1, ..., X_k)'$ is a function from S to $\mathbb{R}^k$. Let $\boldsymbol{x} = (x_1, ..., x_k)'$ denote a vector in $\mathbb{R}^k$. (In this Appendix, we use bold to denote vectors. Bold capitals $\boldsymbol{X}$ are random vectors and bold lower case $\boldsymbol{x}$ are nonrandom vectors. Again, this is in distinction to the notation used in the bulk of the text) The vector $\boldsymbol{X}$ has the distribution and density functions

$$\begin{aligned} F(\boldsymbol{x}) &= \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) \\ f(\boldsymbol{x}) &= \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F(\boldsymbol{x}). \end{aligned}$$

For a measurable function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^s$, we define the expectation

$$\mathbb{E}\boldsymbol{g}(\boldsymbol{X}) = \int_{\mathbb{R}^k} g(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

where the symbol $d\boldsymbol{x}$ denotes $dx_1 \cdots dx_k$. In particular, we have the $k \times 1$ multivariate mean

$$\boldsymbol{\mu} = \mathbb{E}\boldsymbol{X}$$

and $k \times k$ covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathbb{E}\left((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\right) \\ &= \mathbb{E}\boldsymbol{X}\boldsymbol{X}' - \boldsymbol{\mu}\boldsymbol{\mu}' \end{aligned}$$

If the elements of $\boldsymbol{X}$ are mutually independent, then $\boldsymbol{\Sigma}$ is a diagonal matrix and

$$\operatorname{var}\left(\sum_{i=1}^{k} \boldsymbol{X}_i\right) = \sum_{i=1}^{k} \operatorname{var}(\boldsymbol{X}_i)$$

B.7 Conditional Distributions and Expectation

The **conditional density** of Y given $\boldsymbol{X} = \boldsymbol{x}$ is defined as

$$f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

if $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$. One way to derive this expression from the definition of conditional probability is

$$\begin{aligned}
f_{Y|X}(y \mid \boldsymbol{x}) &= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \mathbb{P}(Y \leq y \mid \boldsymbol{x} \leq \boldsymbol{X} \leq \boldsymbol{x} + \varepsilon) \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{\mathbb{P}(\{Y \leq y\} \cap \{\boldsymbol{x} \leq \boldsymbol{X} \leq \boldsymbol{x} + \varepsilon\})}{\mathbb{P}(\boldsymbol{x} \leq X \leq \boldsymbol{x} + \varepsilon)} \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{F(\boldsymbol{x} + \varepsilon, y) - F(\boldsymbol{x}, y)}{F_{\boldsymbol{X}}(\boldsymbol{x} + \varepsilon) - F_X(\boldsymbol{x})} \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{\frac{\partial}{\partial x} F(\boldsymbol{x} + \varepsilon, y)}{f_{\boldsymbol{X}}(\boldsymbol{x} + \varepsilon)} \\
&= \frac{\frac{\partial^2}{\partial x \partial y} F(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\
&= \frac{f(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}.
\end{aligned}$$

The **conditional mean** or **conditional expectation** is the function

$$m(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x})\, dy.$$

The conditional mean $m(\boldsymbol{x})$ is a function, meaning that when $\boldsymbol{X}$ equals $\boldsymbol{x}$, then the expected value of Y is $m(\boldsymbol{x})$.

Similarly, we define the conditional variance of Y given $\boldsymbol{X} = \boldsymbol{x}$ as

$$\begin{aligned}
\sigma^2(\boldsymbol{x}) &= \operatorname{var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= \mathbb{E}\left((Y - m(\boldsymbol{x}))^2 \mid \boldsymbol{X} = \boldsymbol{x}\right) \\
&= \mathbb{E}\left(Y^2 \mid X = \boldsymbol{x}\right) - m(\boldsymbol{x})^2.
\end{aligned}$$

Evaluated at $\boldsymbol{x} = \boldsymbol{X}$, the conditional mean $m(\boldsymbol{X})$ and conditional variance $\sigma^2(\boldsymbol{X})$ are random variables, functions of $\boldsymbol{X}$. We write this as $\mathbb{E}(Y \mid \boldsymbol{X}) = m(\boldsymbol{X})$ and $\operatorname{var}(Y \mid \boldsymbol{X}) = \sigma^2(\boldsymbol{X})$. For example, if $\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \alpha + \beta'\boldsymbol{x}$, then $\mathbb{E}(Y \mid \boldsymbol{X}) = \alpha + \beta'\boldsymbol{X}$, a transformation of $\boldsymbol{X}$.

The following are important facts about conditional expectations.

Simple Law of Iterated Expectations:

$$\mathbb{E}(\mathbb{E}(Y \mid \boldsymbol{X})) = \mathbb{E}(Y) \tag{B.8}$$

Proof:

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(Y \mid \boldsymbol{X})) &= \mathbb{E}(m(\boldsymbol{X})) \\
&= \int_{-\infty}^{\infty} m(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) dy d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(y, \boldsymbol{x})\, dy d\boldsymbol{x} \\
&= \mathbb{E}(Y).
\end{aligned}$$

Law of Iterated Expectations:

$$\mathbb{E}(\mathbb{E}(Y \mid \boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X}) = \mathbb{E}(Y \mid \boldsymbol{X}) \tag{B.9}$$

Conditioning Theorem. For any function $g(\boldsymbol{x})$,

$$\mathbb{E}\left(g(\mathbf{X})Y \mid \mathbf{X}\right) = g\left(\mathbf{X}\right)\mathbb{E}\left(Y \mid \mathbf{X}\right) \tag{B.10}$$

Proof: Let

$$\begin{aligned} h(\boldsymbol{x}) &= \mathbb{E}\left(g(\mathbf{X})Y \mid \mathbf{X} = \boldsymbol{x}\right) \\ &= \int_{-\infty}^{\infty} g(\boldsymbol{x}) y f_{Y|\mathbf{X}}\left(y \mid \boldsymbol{x}\right) dy \\ &= g(\boldsymbol{x}) \int_{-\infty}^{\infty} y f_{Y|\mathbf{X}}\left(y \mid \boldsymbol{x}\right) dy \\ &= g(\boldsymbol{x}) m(\boldsymbol{x}) \end{aligned}$$

where $m(\boldsymbol{x}) = \mathbb{E}\left(Y \mid \mathbf{X} = \boldsymbol{x}\right)$. Thus $h(\mathbf{X}) = g(\mathbf{X})m(\mathbf{X})$, which is the same as $\mathbb{E}\left(g(\mathbf{X})Y \mid \mathbf{X}\right) = g\left(\mathbf{X}\right)\mathbb{E}\left(Y \mid \mathbf{X}\right)$.

B.8 Transformations

Suppose that $\mathbf{X} \in \mathbb{R}^k$ with continuous distribution function $F_{\mathbf{X}}(\boldsymbol{x})$ and density $f_{\mathbf{X}}(\boldsymbol{x})$. Let $\mathbf{Y} = \boldsymbol{g}(\mathbf{X})$ where $\boldsymbol{g}(\boldsymbol{x}) : \mathbb{R}^k \to \mathbb{R}^k$ is one-to-one, differentiable, and invertible. Let $\boldsymbol{h}(\boldsymbol{y})$ denote the inverse of $\boldsymbol{g}(\boldsymbol{x})$. The **Jacobian** is

$$J(\boldsymbol{y}) = \det\left(\frac{\partial}{\partial \boldsymbol{y}'}\boldsymbol{h}(\boldsymbol{y})\right).$$

Consider the univariate case $k = 1$. If $g(x)$ is an increasing function, then $g(X) \leq Y$ if and only if $X \leq h(Y)$, so the distribution function of Y is

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(g(X) \leq y\right) \\ &= \mathbb{P}\left(X \leq h(Y)\right) \\ &= F_X\left(h(Y)\right). \end{aligned}$$

Taking the derivative, the density of Y is

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X\left(h(Y)\right)\frac{d}{dy}h(y).$$

If $g(x)$ is a decreasing function, then $g(X) \leq Y$ if and only if $X \geq h(Y)$, so

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(g(X) \leq y\right) \\ &= 1 - \mathbb{P}\left(X \geq h(Y)\right) \\ &= 1 - F_X\left(h(Y)\right) \end{aligned}$$

and the density of Y is

$$f_Y(y) = -f_X\left(h(Y)\right)\frac{d}{dy}h(y).$$

We can write these two cases jointly as

$$f_Y(y) = f_X\left(h(Y)\right)\left|J(y)\right|. \tag{B.11}$$

This is known as the **change-of-variables** formula. This same formula (B.11) holds for $k > 1$, but its justification requires deeper results from analysis.

As one example, take the case $X \sim U[0,1]$ and $Y = -\log(X)$. Here, $g(x) = -\log(x)$ and $h(y) = \exp(-y)$ so the Jacobian is $J(y) = -\exp(y)$. As the range of X is $[0,1]$, that for Y is $[0,\infty)$. Since $f_X(x) = 1$ for $0 \leq x \leq 1$ (B.11) shows that

$$f_Y(y) = \exp(-y), \qquad 0 \leq y \leq \infty,$$

an exponential density.

B.9 Normal and Related Distributions

The **standard normal** density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad -\infty < x < \infty.$$

It is conventional to write $X \sim \mathrm{N}(0,1)$, and to denote the standard normal density function by $\phi(x)$ and its distribution function by $\Phi(x)$. The latter has no closed-form solution. The normal density has all moments finite. Since it is symmetric about zero all odd moments are zero. By iterated integration by parts, we can also show that $\mathbb{E}X^2 = 1$ and $\mathbb{E}X^4 = 3$. In fact, for any positive integer m, $\mathbb{E}X^{2m} = (2m-1)!! = (2m-1)\cdot(2m-3)\cdots 1$. Thus $\mathbb{E}X^4 = 3$, $\mathbb{E}X^6 = 15$, $\mathbb{E}X^8 = 105$, and $\mathbb{E}X^{10} = 945$.

If Z is standard normal and $X = \mu + \sigma Z$, then using the change-of-variables formula, X has density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

which is the **univariate normal density**. The mean and variance of the distribution are μ and σ^2, and it is conventional to write $X \sim \mathrm{N}(\mu, \sigma^2)$.

For $\boldsymbol{x} \in \mathbb{R}^k$, the **multivariate normal density** is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\right), \qquad \boldsymbol{x} \in \mathbb{R}^k.$$

The mean and covariance matrix of the distribution are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and it is conventional to write $X \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The MGF and CF of the multivariate normal are $\exp(\boldsymbol{\lambda}'\boldsymbol{\mu} + \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}/2)$ and $\exp(\mathrm{i}\boldsymbol{\lambda}'\boldsymbol{\mu} - \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}/2)$, respectively.

If $\boldsymbol{X} \in \mathbb{R}^k$ is multivariate normal and the elements of $\boldsymbol{X}$ are mutually uncorrelated, then $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_j^2\}$ is a diagonal matrix. In this case the density function can be written as

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{(2\pi)^{k/2} \sigma_1 \cdots \sigma_k} \exp\left(-\left(\frac{(x_1-\mu_1)^2/\sigma_1^2 + \cdots + (x_k-\mu_k)^2/\sigma_k^2}{2}\right)\right) \\ &= \prod_{j=1}^{k} \frac{1}{(2\pi)^{1/2}\sigma_j} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right) \end{aligned}$$

which is the product of marginal univariate normal densities. This shows that if $\boldsymbol{X}$ is multivariate normal with uncorrelated elements, then they are mutually independent.

Theorem B.9.1 *If* $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}$ *with* $\boldsymbol{B}$ *an invertible matrix, then* $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}')$.

Theorem B.9.2 *Let* $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_r)$. *Then* $Q = \boldsymbol{X}'\boldsymbol{X}$ *is distributed chi-square with* r *degrees of freedom, written* χ_r^2.

Theorem B.9.3 *If* $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{A})$ *with* $\mathbf{A} > 0$, $q \times q$, *then* $\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z} \sim \chi_q^2$.

Theorem B.9.4 *Let* $Z \sim \mathrm{N}(0,1)$ *and* $Q \sim \chi_r^2$ *be independent. Then* $T_r = Z/\sqrt{Q/r}$ *is distributed as student's t with* r *degrees of freedom.*

Proof of Theorem B.9.1. By the change-of-variables formula, the density of $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{BX}$ is

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{k/2} \det(\boldsymbol{\Sigma}_Y)^{1/2}} \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{\mu}_Y)' \boldsymbol{\Sigma}_Y^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_Y)}{2} \right), \qquad \boldsymbol{y} \in \mathbb{R}^k.$$

where $\boldsymbol{\mu}_Y = \boldsymbol{a}+\boldsymbol{B\mu}$ and $\boldsymbol{\Sigma}_Y = \boldsymbol{B\Sigma B}'$, where we used the fact that $\det(\boldsymbol{B\Sigma B}')^{1/2} = \det(\boldsymbol{\Sigma})^{1/2} \det(\boldsymbol{B})$. ■

Proof of Theorem B.9.2. First, suppose a random variable Q is distributed chi-square with r degrees of freedom. It has the MGF

$$\mathbb{E}\exp(tQ) = \int_0^\infty \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{r/2}} x^{r/2-1} \exp(tx) \exp(-x/2)\, dy = (1-2t)^{-r/2}$$

where the second equality uses the fact that $\int_0^\infty y^{a-1} \exp(-by)\, dy = b^{-a}\Gamma(a)$, which can be found by applying change-of-variables to the gamma function. Our goal is to calculate the MGF of $Q = \boldsymbol{X}'\boldsymbol{X}$ and show that it equals $(1-2t)^{-r/2}$, which will establish that $Q \sim \chi^2_r$.

Note that we can write $Q = \boldsymbol{X}'\boldsymbol{X} = \sum_{j=1}^r Z_j^2$ where the Z_j are independent $\mathrm{N}(0,1)$. The distribution of each of the Z_j^2 is

$$\begin{aligned} \mathbb{P}\left(Z_j^2 \leq y\right) &= 2\mathbb{P}\left(0 \leq Z_j \leq \sqrt{y}\right) \\ &= 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \\ &= \int_0^y \frac{1}{\Gamma\left(\frac{1}{2}\right) 2^{1/2}} s^{-1/2} \exp\left(-\frac{s}{2}\right) ds \end{aligned}$$

using the change–of-variables $s = x^2$ and the fact $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Thus the density of Z_j^2 is

$$f_1(x) = \frac{1}{\Gamma\left(\frac{1}{2}\right) 2^{1/2}} x^{-1/2} \exp\left(-\frac{x}{2}\right)$$

which is the χ_1^2 and by our above calculation has the MGF of $\mathbb{E}\exp\left(tZ_j^2\right) = (1-2t)^{-1/2}$.

Since the Z_j^2 are mutually independent, (B.6) implies that the MGF of $Q = \sum_{j=1}^r Z_j^2$ is $\left[(1-2t)^{-1/2}\right]^r = (1-2t)^{-r/2}$, which is the MGF of the χ_r^2 density as desired. ■

Proof of Theorem B.9.3. The fact that $\boldsymbol{A} > 0$ means that we can write $\boldsymbol{A} = \boldsymbol{CC}'$ where $\boldsymbol{C}$ is non-singular. Then $\boldsymbol{A}^{-1} = \boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}$ and

$$\boldsymbol{C}^{-1}\boldsymbol{Z} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{C}^{-1\prime}\right) = \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{C}^{-1}\boldsymbol{CC}'\boldsymbol{C}^{-1\prime}\right) = \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_q\right).$$

Thus

$$\boldsymbol{Z}'\boldsymbol{A}^{-1}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{Z} = \left(\boldsymbol{C}^{-1}\boldsymbol{Z}\right)'\left(\boldsymbol{C}^{-1}\boldsymbol{Z}\right) \sim \chi_q^2.$$

■

Proof of Theorem B.9.4. Using the simple law of iterated expectations, T_r has distribution function

$$\begin{aligned} F(x) &= \mathbb{P}\left(\frac{Z}{\sqrt{Q/r}} \leq x\right) \\ &= \mathbb{E}\left\{Z \leq x\sqrt{\frac{Q}{r}}\right\} \\ &= \mathbb{E}\left[\mathbb{P}\left(Z \leq x\sqrt{\frac{Q}{r}} \mid Q\right)\right] \\ &= \mathbb{E}\Phi\left(x\sqrt{\frac{Q}{r}}\right) \end{aligned}$$

Thus its density is

$$\begin{aligned} f(x) &= \mathbb{E}\frac{d}{dx}\Phi\left(x\sqrt{\frac{Q}{r}}\right) \\ &= \mathbb{E}\left(\phi\left(x\sqrt{\frac{Q}{r}}\right)\sqrt{\frac{Q}{r}}\right) \\ &= \int_0^\infty \left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{qx^2}{2r}\right)\right)\sqrt{\frac{q}{r}}\left(\frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}}q^{r/2-1}\exp\left(-q/2\right)\right)dq \\ &= \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)} \end{aligned}$$

which is that of the student t with r degrees of freedom. ■