7 Probability Theory

7.1 Probability Spaces

Let Ω be a set of **outcomes**, also referred to as a **sample space**, e.g. the outcome of a coin toss or the roll of two dice, etc. A set of subsets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is called a $\sigma-$**algebra** if and only if (i) $\Omega \in \mathcal{F}$, (ii) if $A \in \mathcal{F}$ then $\Omega \backslash A \in \mathcal{F}$ and (iii) if $A_1, A_2, A_3, ... \in \mathcal{F}$ then $A_1 \cup A_2 \cup A_3 \cup ... \in \mathcal{F}$ for any countable sequence $\{A_1, A_2, A_3, ...\}$ of subsets of Ω. The elements of $\mathcal{F}$ are called **events**. Given a sample space Ω and a σ-algebra $\mathcal{F}$, a **probability measure** is a function $P : \mathcal{F} \to \mathbb{R}$ such that (i) $P(A) \geq 0$ for all $A \in \mathcal{F}$, (ii) $P(\Omega) = 1$, and (iii) for any countable sequence of events $A_1, A_2, A_3, ...$ that are **disjoint** (i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$) then $P(A_1 \cup A_2 \cup ...) = \sum_i P(A_i)$. The triple $(\Omega, \mathcal{F}, P)$ is called a **probability space**. Whenever we use the notation $P(\cdot)$ below it will represent the probability measure over a sample space Ω with $\sigma-$algebra $\mathcal{F}$. It will be useful to review the set theory that we did at the start before proceeding.

Exercise 67. Sometimes we will use the notation A^c to denote the complement $\Omega \backslash A$ of A in Ω. First prove **De Morgan's laws**: (i) $A \cap B = (A^c \cup B^c)^c$ and (ii) $A \cup B = (A^c \cap B^c)^c$. Then use these to prove the following four properties of the probability function: (i) $A \in \mathcal{F}$ implies $P(A^c) = 1 - P(A)$, (ii) $A, B \in \mathcal{F}$ with $A \subseteq B$ implies $P(A) \leq P(B)$, (iii) $A, B \in \mathcal{F}$ implies $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, and (iv) $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ with $A_1 \subseteq A_2 \subseteq A_3 \subseteq ...$ implies $P(\bigcup_{i=1}^{\infty} A_i) = \lim_{n \to \infty} P(A_n)$. A consequence of some of these properties, you will notice, is that $P(A) \in [0, 1]$ for all $A \in \mathcal{F}$.

Consider an event $B \in \mathcal{F}$ with $P(B) > 0$. Then the **conditional probability** given event B, which we denote by the function $P(\cdot|B) : \mathcal{F} \to \mathbb{R}$ is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \qquad \forall A \in \mathcal{F} \tag{86}$$

It is easy to verify that $P(\cdot|B)$ is also a probability measure, and so it satisfies all of the probability of the probability measure that you proved in the above exercise. In addition, note that

$$A, A' \subseteq B \text{ with } P(A') > 0 \text{ implies } \frac{P(A|B)}{P(A'|B)} = \frac{P(A)}{P(A')}$$

so that the ratio of conditional probabilities of two events equals the ratio of their unconditional probabilities when they are both sub-events of the conditioning event. The definition of a conditional probability gives rise to a new probability space, which we call the **conditional probability space**.

Next, we say that two events A and B are **independent** if $P(A \cap B) = P(A)P(B)$. This means that if $P(A) = 0$ or $P(B) = 0$ then A and B must be independent, and if $P(B) > 0$ then the two events are independent if $P(A|B) = P(A)$. As well, if A and B are independent then A^c and B are independent, so are A and B^c and A^c and B^c. Similarly, two events A and A' are **conditionally independent**, conditional on a third event B, if $P(A \cap A'|B) = P(A|B)P(A'|B)$. The other properties also carry over.

Exercise 68. Prove that if A and B are two independent events then A^c and B^c are two independent events (where A^c and B^c denote the complements of A and B respectively).

Finally, we derive an important formula known as **Bayes rule**:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

This follows from noting that the numerator in (86) is simply $P(A \cap B)$ while the denominator equals $P(A \cap B) + P(A^c \cap B) = P\left((A \cap B) \cup (A^c \cap B)\right) = P(B)$. Since conditional probabilities require the conditioning event to have positive probability, the formula holds only when these necessary conditions hold. The formula is useful in computing $P(A|B)$ when $P(A)$, $P(B|A)$ and $P(B|A^c)$ are known but $P(A|B)$ is not. The following exercises provide you with examples of this.

Exercise 69. DHS has just developed a new program that is able to check foreigners entering the US for whether they are terrorists. Suppose the probability that a person trying to enter the US is a terrorist is 10^{-5} and the program correctly classifies a terrorist as being one 99.8% of the time, and correctly classifies an innocent visitor as being innocent 99.99% of the time. Suppose a person is classified as a terrorist. What is the probability that s/he is one?

Exercise 70. Prove that

$$P(X \cap Y|Z) = \frac{P(Y \cap Z|X)P(X)}{P(Z)}.$$

Exercise 71. Consider a gameshow in which there are three opaque boxes, only one of which has a cash prize inside; the others are empty, and the prize is put randomly in one box. The gameshow host asks a contestant to choose a box. Whichever box he chooses, one of the other two will be empty so she will open up one box (among the two that were not chosen, and at random if both are empty) and reveal that it is empty. Then she will ask the contestant whether he wants to stay with the box he picked, or switch to the other closed box. After that, the contestant's decision is final, and wins the prize

if it is inside the box he decided to go with. Should the contestant stay with the box he originally picked, or switch to the other closed box after an empty box was revealed to him? Explain your answer in detail using probability theory.

7.2 Random Variables

Given a probability space $(\Omega, \mathcal{F}, P)$, a (one-dimensional) **random variable** X is a function $X : \Omega \to \mathbb{R}$ for which $\{\omega : X(\omega) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$. Note that by this definition, it is possible to generate new random variables from functions of random variables; e.g., for a function $h : X(\Omega) \to \mathbb{R}$, the composition $h \circ X$ is a random variable on the same probability space if $\{\omega : h(X(\omega)) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$, etc. Given the probability space and random variable defined on it, we say that a set $A \subseteq \mathbb{R}$ is **measurable** if $\{\omega : X(\omega) \in A\} \in \mathcal{F}$.

When the probability space and random variable that we are considering are clear, we will often abuse notation and write $P(\{\omega : X(\omega) \in A\})$ as $P(X \in A)$ or $P(A)$ or $P(\text{“some description of } A\text{”})$ for any measurable set A.

The cumulative distribution function, or **cdf**, of a random variable X is a function $F : \mathbb{R} \to \mathbb{R}$ defined by $F(x) = P(\{\omega : X(\omega) \leq x\})$. It is easy to verify that F has the following three properties

(i) it is nondecreasing,

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

(iii) it is “**right continuous**;” i.e., for all x, we have $\lim_{\tilde{x} \to x^+} F(\tilde{x}) = F(x)$.

In fact, we will refer to any function F that has the above properties as a cumulative distribution function even when there is no underlying probability space and random variable associated with it. For example we might say that the “ideology is distributed according to F.” When we say this we mean that $F(x)$ gives the fraction of individuals whose ideology is at or to the left of the ideology x.

If for all but a countable number of values of $x \in \mathbb{R}$, there exists $\epsilon > 0$ such that F is constant on $[x - \epsilon, x + \epsilon]$ then X is said to be a **discrete** random variable. When this is the case, at each of the countable number of points at which this condition fails, F must have a **jump**; i.e., if x is one of these points, then

$$f(x) = F(x) - \lim_{\tilde{x} \to x^-} F(\tilde{x}) > 0$$

F is then said to be a **step function**. The size of the jump $f(x)$ is the "probability of x," i.e., $f(x) = P(\{x\}) = P(\{\omega : X(\omega) = x\})$. The function $f(\cdot)$ mapping the jump points to $\mathbb{R}$ is called the probability mass function, or **pmf**.

If F is differentiable at every point $x \in \mathbb{R}$ then X is said to be a **continuous** random variable. The derivative of F, which we denote f will be called the probability distribution function, or **pdf**. Note then that $F(x) = \int_{-\infty}^{x} f(y)dy$ by the fundamental theorem of calculus. For $x_1 < x_2$, we have

$$P(\{x : x_1 < x \leq x_2\}) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx,$$

and note that $P(\{x\}) = \int_{x}^{x} f(y)dy = 0$ if X is continuous.

If X is not discrete and its cdf is differentiable at all but a countable number of points at which there are jumps, then X is said to be **mixed**. In this case, a jump point is also often called a **mass point** or **atom**. While much (but not all) of what we say in the sequel will also apply to mixed random variables, we will implicitly assume from here on that a given random variable is either discrete or continuous. If $h : \mathbb{R} \to \mathbb{R}$ is a function and A is a measurable set, we sometimes use the notation

$$\int_A h(x)dF(x) := \begin{cases} \sum_{x \in A} h(x)f(x) & \text{if } X \text{ is discrete} \\ \int_A h(x)f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

This notation also has meaning in the case where F is mixed; to learn more about it, look up the "Riemann-Stieltjes integral." Note that $P(A) = \int_A dF(x)$ in both cases.

Finally, the **support** of a discrete/continuous random variable X can be defined as the set of points at which there are jumps/the set of points at which the derivative of F takes positive value. We will denote the support of X by $\mathsf{supp}\, X$ or $\mathsf{supp}\, F$ or $\mathsf{supp}\, f$. *Note:* The definition of support is sometimes different: the "set of points at which the derivative of F takes a positive value" is often replaced by "the smallest closed set containing the set of points at which the derivative f of F takes a positive value." Even though this would change the notion of support (by including some extra points) this typically does not pose any problems.

Exercise 72. Suppose that $f(x)$ equals 0 for negative values of x and equals $Ke^{-\alpha x}(1 - e^{-\alpha x})$ for nonnegative values of x, for some $\alpha > 0$. (i) Find K such that $f(x)$ is the pdf of a continuous random variable. (ii) Find the corresponding cdf. (iii) Find the probability that the random variable takes value strictly larger than 1.

7.3 Transformations of Random Variables

Let X be a random variable on a probability space and consider an injective function $h : \mathbb{R} \to \mathbb{R}$. Redefining its range to be the image of $\mathbb{R}$, this function is bijective so it has inverse h^{-1}. In the discrete case, the distribution of the random variable $Y = h \circ X$ is

$$f_Y(y) = P(Y = y) = P(X = h^{-1}(y)) = f_X(h^{-1}(y))$$

where f_X is the pmf of X and f_Y the pmf of Y. From this we can generate the cdf of Y, which is

$$F_Y(y) = \sum_{\tilde{y} \leq y} f_X(h^{-1}(\tilde{y})).$$

In the continuous case, suppose that h is increasing and has a differentiable inverse. Then for $y \in h(\mathbb{R})$, we have

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(h^{-1}(y))}{dy} = f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}$$

and $f_Y(y) = 0$ for $y \notin h(\mathbb{R})$. If, on the other hand, h is decreasing (but also still has a differentiable inverse) then we would derive the same expression as above but with a negative sign in front of it. This follows after noting that

$$F_Y(y) = P(Y \leq y) = P(X \geq h^{-1}(y)) = 1 - F_X(h^{-1}(y)),$$

and then taking the derivative of the left and right most sides with respect to y.

Exercise 73. Consider a continuous random variable X with pdf f_X. Let $h(x) = x^2$, and consider the random variable $Y = h \circ X$. Find an expression for the pdf of Y that depends only on y and f_X.

Exercise 74. If X is a continuous random variable with pdf $f(x) = 1/[\pi(1 + x^2)]$, what is the pdf of $1/X$?

7.4 Joint, Marginal and Conditional Distributions

Given a probability space, consider a pair of random variables (X, Y) each defined on this space. Let F_X denote the cdf of X and F_Y the cdf of Y; f_X and f_Y will denote their pmf/pdf respectively. The **joint cdf** of X and Y is the function $F_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ defined for $x = (x, y) \in \mathbb{R}^2$ by

$$F_{X,Y}(x, y) = P(\{\omega : X(\omega) \leq x \text{ and } Y(\omega) \leq y\})$$

If X and Y are discrete then

$$F_{X,Y}(x,y) = \sum_{\tilde{x}\leq x}\sum_{\tilde{y}\leq y} f(\tilde{x},\tilde{y})$$

where

$$f_{X,Y}(\tilde{x},\tilde{y}) = P(\{\omega : X(\omega) = \tilde{x} \text{ and } Y(\omega) = \tilde{y}\})$$

is the **joint pmf**, and if they are continuous then

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}$$

where

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

is the **joint pdf**. These definitions generalize the the natural way to larger collections of random variables than just pairs; e.g., a collection of n random variables $(X_1, X_2, ..., X_n)$ defined on the same probability space. What we say below also generalizes.

Note that the cdf of X can be recovered from the joint cdf of (X,Y) as

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y)$$

and the joint pdf or pmf can be recovered as

$$f_X(x) = \sum_y f_{X,Y}(x,y) \qquad \text{or} \qquad f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

in the discrete and continuous cases, respectively. Note the summation is over all jump points. These derived functions are called the **marginal** cdf, pmf, or pdf of X given the joint cdf, pmf or pdf of X, Y. The concept of "support" defined above carries over to these distribution functions.

X and Y are said to be **independent** if their joint pmf in the discrete case, or joint pdf in the continuous case, can be written as the product of their pmfs or pdfs respectively; that is, the joint pmf or pdf is the product of the marginal pmfs or pdfs. This implies that the joint cdf is the product of the marginal cdfs as well. Note that X and Y can be independent only if the support of one does not depend on the other. Also, if X and Y are independent, then for function g and h, if $g(X)$ and $h(Y)$ are also random variables on the same probability space, this pair is also independent.

If X and Y are discrete then $f_{X,Y}$ represents their joint pmf while if they are continuous then it represents their joint pdf. With this in mind, we may be able to construct

a new collection of random variables which we will denote $\{X \mid Y = y\}_y$ each member to be read as X "given" (or "conditional on") $Y = y$. If the marginal distribution $f_Y(y) > 0$, then the joint pmf or pdf of this random variable is given by

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

which we say is the **conditional distribution** of X given $Y = y$. The conditional joint cdf of this collection can be generated from the joint pmf or pdf by summing or integrating. Please note the connection between these distributions and conditional probability: $X|Y = y$ is a random variable on the conditional probability space. For example, in the discrete case, observe that

$$f_{X|Y=y}(x|y) = P\left(\{\omega : X(\omega) = x\} \mid \{\omega : Y = y\}\right).$$

Note that if X and Y are independent, then

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \text{ so } f_{X|Y=y}(x|y) = f_X(x) \text{ for all } y;$$

i.e., the conditional distribution is always equal to the marginal distribution. Also, define two random variables X and Y to be **conditionally independent** conditional on a third random variable Z if their joint conditional pmf or pdf $f_{X,Y|Z}$ is the product of their marginal conditional pdfs or pmfs $f_{X|Z}$ and $f_{Y|Z}$.

Exercise 75. Let $f_{X,Y}(x,y) = 15x^2y$ for $0 < x < y < 1$ and zero elsewhere be the joint pdf of two continuous random variables X and Y. (i) What are the marginal pdfs and cdf of X and Y? (ii) Are X and Y independent? (iii) What is the conditional distribution (both pdf and cdf) of Y given X?

Exercise 76. Consider a sample of size 2 drawn without replacement from an urn containing three balls numbered 1, 2, 3. Let X be the number on the first ball drawn and Y the larger of the two numbers drawn. (i) Find the joint pdf of X and Y. (ii) Find the distribution of $X \mid Y = 3$.

Exercise 77. Let $F_{X,Y}$ be a joint cdf of the discrete random variables X and Y with marginal cdfs F_X and F_Y. Prove that $F_X(x) + F_Y(y) - 1 \leq F_{X,Y}(x,y) \leq \sqrt{F_X(x)F_Y(y)}$.

7.5 Expectations and Other Moments

The expectations operator, denoted $E[\cdot]$, is a mapping that works on random variables defined on a probability space. If X is a random variable, then the **expectation** of X,

if it exists, is given by

$$E[X] = \int_{-\infty}^{\infty} x dF(x)$$

The expectation exists if $\int_{-\infty}^{\infty} |x| dF(x)$ converges to a finite number. Recall that integrals with bounds of ∞ or $-\infty$ on either end represent limits of sequences, which means that the sequences must converge. The expectation of a random variable is also sometimes called its **mean**.

In what follows we will work with expectations and typically drop the qualifier that some property about expectations is true only if the expectations exist. The following are four properties of the expectations operator. You should verify these.

(i) It inherits linearity from the linearity of summation and integration; i.e., for the random variable constructed by linearly combining two random variables X and Y with weights $a, b \in \mathbb{R}$, we have

$$E[aX + bY] = aE[X] + bE[Y].$$

(ii) The expectation of a function g of a random variable is the expectation of the random variable it generates; i.e., for $Y = g(X)$, we have

$$E[Y] = E[g(X)] = \int_{-\infty}^{\infty} g(x) dF(x)$$

where F is the cdf of X.

(iii) The expectation of a random variable created as a function of several random variables is computed by integrating with respect to the joint distribution; e.g., for $Z = h(X, Y)$, we have

$$E[Z] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) dF_{X,Y}(x, y).$$

(iv) If X and Y are independent, then

$$E[XY] = E[X]E[Y].$$

The conditional expectation of X given $Y = y$ is simply the expectation of the random variable $X \mid Y = y$ defined above. Thus,

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} x dF_{X|Y=y}(x|y),$$

which is a function that depends on y. We will abbreviate it by referring to it as $\mu_X(y)$; sometimes this is called a **regression** function. Note that it is a random variable; therefore, it has an expectation. Conveniently, its expectation is simply the (unconditional) expectation of X; for example, in the continuous case, we have

$$\begin{aligned} E[\mu_X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx f_Y(y)dy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx \\ &= \int_{-\infty}^{\infty} x f_X(x)dx = E[X] \end{aligned}$$

The first equality is by definition, the second follows because f_Y is independent of x, the third by definition of the conditional distribution, the fourth by Fubini's theorem, the last two by the definitions of marginal distribution and expectation of X. Nevertheless,

$$E[E[X \mid Y]] = E[X]$$

holds in the discrete and mixed cases as well, so long as the expectations exist, and is known as the **law of iterated expectations**.

Exercise 78. Let X and Y be two random variables. Consider the random variable $Y - h(x)$ where the function h of x is called the **forecast** of Y, and $Y - h(x)$ the forecast error. Suppose we wanted to minimize the expectation of the square of the forecast error, $E[(Y - h(x))^2 \mid X = x]$. Show that the optimal forecast (i.e., the function h that solves this minimization problem) is $h(x) = E[Y \mid X = x]$.

Exercise 79. Let X be a random variable that is nonnegative with probability 1, with cdf F. Show that if the expectation of X exists, then $E[X] = \int_0^{\infty}(1 - F(x))dx$ if X is continuous. (You can show that this is also true if X is discrete.)

Exercise 80. Let X be a continuous random variable with cdf F. The **median** m of X is defined as the value of x such that $F(x) = 1/2$. Suppose that m is unique. Show that $E[|X - a|]$ is minimized by choosing $a = m$.

The expectation of a random variable X is sometimes called its first moment, and often denoted $\mu = E[X]$. In general, the quantity $E[X^k]$, if it exists, is called the kth **moment** of X. The k'th centered moment of X is $E[(X - \mu)^k]$, where μ is the first

moment. The second centered moment has a special name: it is called the **variance** of X and is often denoted $Var[X]$ or σ^2. Its positive square root, σ, is called the **standard deviation**. Observe that

$$E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - \mu^2.$$

For a vector of random variables $X = (X_1, ..., X_n)$ each defined on the same probability space, the mean is defined as the vector $E[X] = (E[X_1], ..., E[X_n])$. The $n \times n$ matrix $\Sigma = E[(X - E[X])(X - E[X])]$ is called the **covariance matrix** of X. The (i,j)th element of Σ is $\sigma_{ij} = E[(X_i - E[X_i])(X_j - E[X_j])]$ is the covariance between X_i and X_j. We also write $Cov[X_i, X_j] = \sigma_{ij}$. Since $\sigma_{ij} = \sigma_{ji}$, Σ is a symmetric matrix.

Let α and β be vectors of length n. Then $E[\alpha' X] = \alpha' E[X]$, where α' denotes the transpose of α. As well, $Var[\alpha' X] = \alpha' \Sigma \alpha \geq 0$ so that Σ is positive semi-definite. The **covariance** between $\alpha' X$ and $\beta' X$ is $E[(\alpha' X - \alpha' E[X])(\beta' X - \beta' E[X]) = \alpha' \Sigma \beta$. For an $n \times k$ matrix of scalars A, the expectation of AX is $E[AX] = A' E[X]$ and the covariance matrix is $A' \Sigma A$. Finally, the **correlation** between X_i and X_j is $\rho_{ij} = \sigma_{ij} / \sqrt{\sigma_{ii}\sigma_{jj}}$.

Exercise 81. Prove that $-1 \leq \rho_{ij} \leq 1$. (*Hint:* Σ is positive semi-definite.) Then show that if X_i and X_j are independent then $\rho_{ij} = 0$ (but note that the converse is not true).

Exercise 82. Let X and Y be random variables. Assuming that all necessary moments exist, prove that the values of a and b that minimize $E[(Y - a - bX)^2]$ are

$$a = E[Y] - \frac{Cov[X,Y]}{Var[X]} E[X] \qquad \text{and} \qquad b = \frac{Cov[X,Y]}{Var[X]}.$$

Exercise 83. Let X and Y be two (jointly) continuous random variables. Define the "conditional variance of Y given $X = x$" as

$$\sigma^2_{Y|X=x}(x) = Var[Y|X=x] = E[(Y - E[Y|X=x])^2 | X = x]$$

and let $Var[Y|X] = \sigma^2_{Y|X}(X)$. Show that $Var[Y] = E[Var[Y|X]] + Var[E[Y|X]]$.

Theorem 41. *Suppose that X is a random variable defined on some probability space.*

1. *(Jensen) If $h : \mathbb{R} \to \mathbb{R}$ is a convex function then $E[h(X)] \geq h(E[X])$.*
2. *(Markov) If $P(X \geq 0) = 1$, then $P(X \geq K) \leq \frac{E[X]}{K}$ for all $K > 0$.*
3. *(Chebychev) If $Var[X] < \infty$ then $P\big(|X - E[X]| \geq K\big) \leq \frac{Var[X]}{K^2}$ for all $K > 0$.*

Proof. (i) If h is convex then by Exercise 31 there is a line $y = h(E[X]) + m(x - E[X])$ such that

$$h(x) \geq m(x - E[X]) + h(E[X]), \qquad \forall x$$

Then taking expectations, we have

$$E[h(X)] \geq E[m(x - E[X])] + E[h(E[X])] = h(E[X]).$$

(ii) I prove this in the continuous case, leaving the discrete case to you. Note that

$$E[X] = \int_0^\infty xf(x)dx \geq \int_K^\infty xf(x)dx \geq K\int_K^\infty f(x)dx = KP(X \geq K).$$

(iii) Let $Y = (X - E[X])^2$ so $P(Y \geq 0) = 1$ and $E[Y] = Var[X]$. Then we have

$$P(|X - E[X]| \geq K) = P(Y \geq K^2) \leq \frac{E[Y]}{K^2} = \frac{Var[X]}{K^2}.$$

where the inequality follows from Markov's inequality. □

7.6 Application: Expected Utility Theory

Consider a finite set of outcomes X. (For this application, we will use X rather than Ω to denote the set of outcomes.) A **lottery** on this space is a mapping $p : X \to [0, 1]$ such that $\sum_x p(x) = 1$. The idea is that $p(x)$ can be interpreted as the probability of the event $\{x\}$ under some probability measure P defined on a σ-algebra that includes every possible singleton event generated from X; i.e. $p(x) = P(\{x\})$. Let $\Delta(X)$ denote the set of all lotteries on X. For any two lotteries, p and q, the lottery $\alpha p + (1 - \alpha)q$ is the lottery defined by $[\alpha p + (1 - \alpha)q](x) = \alpha p(x) + (1 - \alpha)q(x)$.

Consider a preference relation $\succeq$ on $\Delta(X)$. Say that $\succeq$ satisfies **continuity** if $p \succ q \succ r$ implies that there are numbers $\alpha, \beta \in (0, 1)$ such that

$$\alpha p + (1 - \alpha)r \succ q \succ \beta p + (1 - \beta)r.$$

We say that $\succeq$ satisfies **independence** if for all $\alpha \in (0, 1)$, $p \succ q$ implies

$$\alpha p + (1 - \alpha)r \succ \alpha q + (1 - \alpha)r, \qquad \forall r \in \Delta(X).$$

As before, $\succeq$ is rational if it satisfies completeness and transitivity.

We say that a function U on $\Delta(X)$ has the **expected utility form** if there exists a real valued function u on X such that

$$U(p) = \sum_{x \in X} u(x)p(x).$$

Theorem 42. (von-Neumann & Morgenstern 1947) *Let $\Delta(X)$ be the set of lotteries defined on a finite probability space, and let $\succeq$ be a preference relation over $\Delta(X)$. Then, there exists a function U with expected utility form that represents $\succeq$ on $\Delta(X)$ if and only if $\succeq$ is a rational preference relation that satisfies continuity and independence.*

The function u is called a **utility index** to distinguish it from the utility function, which is U. The utility function is over lotteries while the utility index is over outcomes. The idea here is that a decision maker may be confronted with choices over lotteries that produce outcomes, rather than the actual outcomes themselves—the case we studied in Section 1.3. If a decision maker's preference relation is rational and satisfies continuity and independence, then the decision maker makes choices over lotteries as if she is maximizing an expected utility, which is the probability weighted sum of the utility index over outcomes. The converse is also true.

Exercise 84. Prove the following: If U is a function with the expected utility form that represents some preference relation $\succeq$ over the lotteries on X, and V is some other function that also takes the expected utility form, then V represents $\succeq$ if and only if there exists a positive number α and a number β such that $V = \alpha U + \beta$.

Exercise 85. Lottery p pays \$55,000 with 33% chance, \$48,000 with 66% chance and \$0 with 1% chance. Lottery q pays \$48,000 for sure. Lottery r pays \$55,000 with 33% chance, and \$0 with 67% chance. Lottery s pays \$48,000 with 34% chance and \$0 with 66% chance. Between p and q, which lottery do you prefer? Which do you prefer between r and s? Now, suppose a decision maker strictly prefers lottery q to lottery p and strictly prefers lottery r to lottery s. Show that this decision maker's preferences violate independence, and therefore cannot be represented in expected utility form.

Monetary Outcomes Under some technical conditions, the representation of preference relations over lotteries with expected utility form generalizes to the case where the set of outcomes X is an interval $X = [w, b]$, and therefore can be interpreted as a set of monetary outcomes. Here, a lottery is redefined to be a cumulative distribution function F with support contained in $[w, b]$, and the expected utility form is taken to be

$$U(F) = \int_w^b u(x)dF(x).$$

where u is again called the utility index. We denote by δ_x the degenerate lottery that gives outcome x for certain. For any lottery F, the **expected value** of the lottery is

$$E[F] = \int xdF(x)$$

which is a number in $[w, b]$. A decision maker with preference $\succeq$ is (strictly) **risk averse** if for any non-degenerate lottery F with expected value $E[F]$,

$$\delta_{E[F]} \succ F;$$

that is, the decision maker would rather take the expected value of the lottery for sure than to expose himself to the risk associated with the lottery.

Theorem 43. *A decision maker with preference relation $\succeq$ over monetary outcomes that is represented in expected utility form with utility index u is strictly risk averse if and only if u is strictly concave.*

Proof. u is concave if and only if $\int_w^b u(x)dF(x) \leq u(\int_w^b xdF(x))$, whose argument follows the argument that we used to prove Jensen's inequality. But this is equivalent to $U(F) \leq U(\delta_{E[F]})$, or $\delta_{E[x]} \succ F$, which defines risk aversion of $\succeq$. □

A decision-maker is **risk neutral** if $\delta_{E[F]} \sim F$ for all non-degenerate lotteries F and is **risk-seeking** if $F \succ \delta_{E[F]}$ for all such lotteries. When her preferences can be represented in expected utility form, these are respectively equivalent to saying that the utility index is linear, and convex.

For any lottery F and utility index u, the **certainty equivalent** is the dollar amount $c(F, u) \in \mathbb{R}$ such that

$$u(c(F, u)) = \int_w^b u(x)dF(x).$$

If the utility index u is twice differentiable, the **Arrow-Pratt measure** of risk aversion is $A_u(x) = -u''(x)/u'(x)$. If for two preference relations $\succeq$ and $\succeq'$ over monetary outcomes on $[w, b]$, it is the case that for all non-degenerate lotteries δ_x, $F \succeq' \delta_x$ implies $F \succeq \delta_x$, then we say that $\succeq$ is more risk averse than $\succeq'$.

Theorem 44. *Let $\succeq$ and $\succeq'$ be two preference relations over lotteries over monetary outcomes that can be represented in expected utility form with twice differentiable utility indices u and v respectively. Then the following are equivalent:*

(i) $\succeq$ is more risk averse than $\succeq'$,

(ii) for all lotteries F, $c(F, u) \leq c(F, v)$,

(iii) there exists an increasing concave function φ such that $u = \varphi \circ v$, and

(iv) for all $x \in X$, $A_u(x) \geq A_v(x)$.

Exercise 86. Prove that (i), (iii) and (iv) are equivalent in the theorem above.

Suppose F and G are two lotteries over monetary outcomes X. We that F **first order stochastically dominates** (fosd) G when

$$F(x) \leq G(x), \qquad \forall x \in X.$$

That is, F "pays more" than G. If F and G have the same expected value, then F **second order stochastically dominates** (sosd) G when

$$\int_{-\infty}^{x} F(\tilde{x})d\tilde{x} \leq \int_{-\infty}^{x} G(\tilde{x})d\tilde{x}, \qquad \forall x \in X.$$

That is, F is "less risky" than G.

Theorem 45. *F fosd G if and only if for every nondecreasing function $u : X \to \mathbb{R}$,*

$$\int u(x)dF(x) \geq \int u(x)dG(x).$$

For two lotteries F and G with the same expected value, F sosd G if and only if for every concave function $u : X \to \mathbb{R}$,

$$\int u(x)dF(x) \geq \int u(x)dG(x).$$

Exercise 87. Consider the lottery that works as follows. We repeatedly flip a fair coin until the first time it shows heads. If heads appears on the nth flip, then the lottery pays $\$2^n$. What is the expected value of this lottery? What is the certainty equivalent of this lottery for a decision maker with risk neutral preferences? What is the certainty equivalent of this lottery for a risk averse decision maker with utility index $u(x) = \log x$? What is the maximum that you would pay to play this lottery?

Exercise 88. A rational decision maker has preferences over monetary outcomes that can be represented in expected utility form with utility index $u(x) = \log x$. She holds an asset worth Y but faces a probability p of incurring a loss L on it. An insurance company is selling a policy that indemnifies her in the event of a loss. If she covers I amount of the loss, then the insurance premium is qI. How much of the loss should she cover? Interpret the meaning of this result by examining how coverage varies with the difference between p and q. (When $p = q$, insurance is said to be **actuarially fair**.)

7.7 The Moment Generating Function & Select Distributions

The moment generating function, or **mgf**, for a random variable X with cdf $F(x)$ is

$$M(t) = E\left[e^{tX}\right]$$

Since $M(t) = \int_{-\infty}^{\infty} e^{tx} dF(x)$, we have

$$M^{(k)}(t) = \int_{-\infty}^{\infty} x^k e^{tx} dF(x), \qquad \text{so } M^{(k)}(0) = E\left[X^k\right]$$

where $M^{(k)}$ denotes the kth derivative of $M(t)$. The moment generating function may not exist for all random variables, but if it does exist on some interval $(-\xi, \xi)$ centered at 0, then it uniquely characterizes the cdf of X. More formally,

Theorem 46. *Suppose X and Y are random variables with cdfs F_X and F_Y and mgfs $M_X(t)$ and $M_Y(t)$ that exist for t sufficiently close to 0. For all $\epsilon > 0$ suppose there exists $\xi > 0$ such that*

$$|M_X(t) - M_Y(t)| < \epsilon \qquad \forall t \in (-\xi, \xi).$$

Then there is a function $\delta(\epsilon)$ such that at all continuity points $a \in \mathbb{R}$ of F_X and F_Y,

$$|F_X(a) - F_Y(a)| < \delta(\epsilon) \qquad \text{and} \qquad \lim_{\epsilon \to 0} \delta(\epsilon) = 0.$$

Thus, when the mgfs of two random variables are close, their cdfs are close as well. The proof of this theorem is outside the scope of this class. However, we will use it below to prove the "Central Limit Theorem."

Exercise 89. Suppose that the random variables $X_1, ..., X_K$ are all (mutually) independent with mgfs $M_1(t), ..., M_K(t)$ that exist for t sufficiently close to 0. Show that the mgf of the random variable created by taking the sum of these variables $Y = \sum_k X_k$ is the product of the mgfs, denoted $\prod_k M_k(t)$.

Exercise 90. Let X be a discrete random variable with pmf $f(x) = (1/2)^x$, $x = 1, 2, 3, ...$ Find $E[X]$ and the mgf $M(t)$ of X for $|t| < \log 2$.

Exercise 91. Suppose the mgf of X and Y are $M_X(t) = M_Y(t) = e^{t^2+3t}$ and X and Y are independent. What is the mgf of $Z = 2X - 3Y + 4$?

I now derive the mgf (and some properties) of several well-known discrete and continuous random variables/distributions. This is also an opportunity for you to learn these distributions and their properties, which will come up in applications.

Bernoulli distribution. The pmf of X is $f(1) = p$, $f(0) = 1 - p$ and $f(x) = 0$ for all $x \notin \{0, 1\}$. p is said to be the **parameter** of the Bernoulli distribution. The mgf is

$$M(t) = E\left[e^{tX}\right] = pe^{1t} + (1-p)e^{0t} = 1 - p + pe^t.$$

so the mean of the distribution is $E[X] = p$ and the variance is $Var[X] = p(1-p)$.

Binomial distribution. Suppose that $X_1, ..., X_n$ are independent and identically distributed (**iid**) Bernoulli random variables each with parameter p. Then $Y = \sum_i X_i$ has the Binomial distribution with support $\mathsf{supp}\, Y = \{0, 1, ..., n\}$ and for $y \in \mathsf{supp}\, Y$, the pmf of Y is

$$f_Y(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

By the result in Exercise 89, the mgf of Y is

$$M(t) = (1 - p + pe^t)^n.$$

The mean of the distribution is therefore np and the variance is $np(1-p)$.

Poisson distribution. X takes values on $\mathsf{supp}\, X = \{0, 1, 2, ...,\}$ and for $x \in \mathsf{supp}\, X$, the pmf is

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

The mgf is

$$M(t) = \sum_{n=0}^{\infty} e^{tn} \frac{\lambda^n e^{-\lambda}}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(e^t\lambda)^n}{n!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}.$$

The mean and the variance are therefore both equal λ, which is the parameter of this distribution. Note also that the Poisson distribution is the limit as $n \to \infty$ of a sequence of Binomial distributions with parameters $p_n = \lambda/n$; to see this, note that the limit as $n \to \infty$ of the sequence of corresponding mgfs is

$$\lim_{n\to\infty} M_n(t) = \lim_{n\to\infty} \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n} e^t\right)^n = e^{\lambda(e^t - 1)}$$

where the second inequality follows from Exercise 40. You can therefore think of the Poisson distribution as modeling the probability λ of one success over a period of unit time. The number of successes in non-overlapping periods are independent.

Exercise 92. Let X and Y be independent random variables both distributed Poisson with $E[X] = \lambda_X$ and $E[Y] = \lambda_Y$. Find the distribution of $X + Y$.

Uniform distribution. X takes values on an interval $\mathsf{supp}\, X = [a, b] \subset \mathbb{R}$ with every value in this interval equally "likely." That is, the pmf is $f(x) = 1/(b-a)$ on $[a, b]$ and 0 everywhere else. The mgf is therefore

$$M(t) = \int_a^b e^{tx} \frac{1}{b-a} dx = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

if $t \neq 0$ and 1 if $t = 0$. The mean of the distribution is the midpoint of the interval $E[X] = \frac{1}{2}(a+b)$ and the variance is $Var[X] = \frac{1}{12}(b-a)^2$.

Exercise 93. Consider random variable X with cdf $F(x)$. Let $h(x) = F(x)$ and let the inverse $h^{-1}(y)$ be defined as the smallest x such that $F(x) = y$ (which exists because F_X is right continuous). Show that $Y = h(X)$ has uniform distribution with support $[0, 1]$.

Exercise 94. If X has a uniform distribution on support $[0, 1]$, find the distribution of $1/X$. Does $E[1/X]$ exist? If so, find it.

Normal distribution. X takes on values on $\mathsf{supp}\, X = \mathbb{R}$ and for parameters (μ, σ) where $\sigma > 0$ the pdf is

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

Note that the moment generating function is

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2 + xt} dx \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-(\mu+\sigma^2 t))^2} dx \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \end{aligned}$$

since the second integral is the integral of the pdf of the normal distribution with parameters $((\mu + t\sigma^2), \sigma^2)$ and so must integrate to 1. The distribution is often denoted $\mathcal{N}(\mu, \sigma^2)$ since $E[X] = \mu$ is the mean of the distribution and $Var[X] = \sigma^2$ the variance. The **standard normal** is the distribution $\mathcal{N}(0, 1)$ with zero mean and unit variance.

Exercise 95. Suppose X is a random variable with standard normal distribution. Show that the random variable $Y = cX$ where c is a constant has distribution $\mathcal{N}(0, c^2)$.

Exercise 96. Suppose that $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables each having Poisson distribution. Suppose the parameter of the distribution of X_n is n for all n. Consider the corresponding sequence of **standardized** Poisson random variables $\{(X_n - n)/\sqrt{n}\}_{n=1}^{\infty}$. Show that the corresponding sequence of distributions of the standardized sequence converges to the standard normal distribution.

Exercise 97. Suppose that X is distributed $\mathcal{N}(\mu, \sigma^2)$. Show that $E[|X - \mu|] = \sigma\sqrt{2/\pi}$.

Chi-squared distribution. Let $X_1, ..., X_n$ be a sequence of iid random variables each with a standard normal distribution. Then $Y = \sum_i (X_i)^2$ is said to have a chi-squared distribution with n degrees of freedom. The distribution is denoted χ^2_n.

Exercise 98. Show that the mean of the chi-squared distribution with k degrees of freedom is $E[Y] = k$ and the variance is $Var[Y] = 2k$.

$F-$distribution. Let Y_1 and Y_2 be two independent random variables each where Y_1 has distribution χ^2_k and Y_2 has distribution χ^2_l. Then $Q = (Y_1/k)/(Y_2/l)$ is said to be have an F_{kl} distribution with k degrees in the numerator and l in the denominator.

$t-$distribution. Let Z be a random variable with standard normal distribution and Y be a random variable with a χ^2_k distribution. If Y and Z are independent, then $T = Z/\sqrt{Y/k}$ is said to have the **student's** t_k-distribution, with k degrees of freedom.

The Multivariate Normal Distribution. Consider the random variables X_1 and X_2, construct the pair (X_1, X_2) and treat it as a vector. $X = (X_1, X_2)$ has the bivariate normal distribution if and only if for all vectors $\alpha \in \mathbb{R}^2$, the random variable defined $\alpha' X$ is normally distributed. If X_1 and X_1 are jointly bivariate normal, then they are each normally distributed. The mean and covariance matrix of (X_1, X_2) both exist and we denote them by μ (a vector of size 2) and Σ (a square matrix of order 4). Let α be a vector of size two. Then $Y = \alpha' X$ is normal with mean and variance $E[\alpha' X] = \mu'\alpha$ and $Var[\alpha' X] = \alpha'\Sigma\alpha$. Therefore,

$$M_X(\alpha) = E[e^{\alpha' X}] = E[e^Y] = M_Y(1) = e^{\alpha'\mu + \frac{1}{2}\alpha'\Sigma\alpha}$$

where M_X and M_Y are the mgfs of X and Y, respectively. By the uniqueness of the mgf, this implies that the distribution of X is completely characterized by μ and Σ, and we write $X = (X_1, X_2) \sim \mathcal{N}_2(\mu, \Sigma)$. These things generalize to the case where $X = (X_1, ..., X_n)$, in which case X has the multivariate normal distribution.

7.8 Convergence Concepts & Results

Consider a probability space $(\Omega, \mathcal{F}, P)$, a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ each defined on the space, and another random variable X also defined on the same space. Then we have the following notions of convergence.

1. Let
$$A = \left\{ \omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega) \right\}$$
be the outcomes on which $\{X_n\}$ converges to X. (Perhaps A is empty.) Then the sequence $\{X_n\}$ is said to **converge almost surely** to X (denoted $X_n \xrightarrow{a.s.} X$) if
$$P(A) = 1.$$

2. Alternatively, for any $\varepsilon > 0$ let
$$p_n(\varepsilon) = P(|X_n - X| > \varepsilon)$$
be the probability that X_n differs by more than ϵ from X. Then if for all ε, the sequence $\{p_n(\varepsilon)\}$ converges to 0, i.e. if
$$\lim_{n\to\infty} p_n(\varepsilon) = 0, \qquad \forall \varepsilon > 0,$$
then $\{X_n\}$ is said to **converge in probability** to X (denoted $X_n \xrightarrow{p} X$ or sometimes as plim $X_n = X$).

3. Suppose that
$$\lim_{n\to\infty} E\left[|X_n - X|^r\right] = 0$$
for $r = 1$. Then $\{X_n\}$ is said to **converge in mean** to X (denoted $X_n \xrightarrow{m} X$). If the same limit holds for $r = 2$ then $\{X_n\}$ is said to **converge in mean square** (denoted $X_n \xrightarrow{m.s.} X$).

4. Let $\{F_n\}$ denote the corresponding sequence of cdfs for the sequence $\{X_n\}$ of random variables and let F be the cdf of X. If for all points at which F is continuous,
$$\lim_{n\to\infty} F_n(x) = F$$
then $\{X_n\}$ is said to **converge in distribution** to X (denoted $X_n \xrightarrow{d} X$). Sometimes we also say that $\{X_n\}$ **converges weakly** to X.

Exercise 99. (i) Show that convergence in mean square implies convergence in mean. (*Hint:* Use Jensen's inequality.) (ii) Show that convergence in mean implies convergence in probability (*Hint:* Use Markov's inequality.) *Extra credit:* (iii) Show that almost sure convergence implies convergence in probability. (iv) Finally, show that convergence in probability implies convergence in distribution.

Theorem 47. (**weak law of large numbers**) *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables with means denoted μ_i and variances denoted σ_i^2. Suppose that they are uncorrelated meaning the covariances are $\sigma_{ij} = 0$ for all $i \neq j$. Let*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i, \qquad \bar{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2$$

and suppose that

$$\lim_{n\to\infty} \bar{\sigma}_n^2/n = 0.$$

Then we have

$$\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0.$$

Proof. Note that by Chebychev's inequality we have for all $\varepsilon > 0$,

$$P(|\bar{X}_n - \bar{\mu}_n| > \varepsilon) \leq \frac{E\left[(\bar{X}_n - \bar{\mu}_n)^2\right]}{\varepsilon^2} = \frac{\frac{1}{n^2}E\left[(\sum_{i=1}^{n}(X_i - \mu_i))^2\right]}{\varepsilon^2} = \frac{\frac{1}{n}\bar{\sigma}_n^2}{\varepsilon^2}$$

Therefore $P(|\bar{X}_n - \bar{\mu}_n| > \varepsilon)$ converges to 0 for all ε since the right side of this expression converges to 0 in the limit as $n \to \infty$. □

Exercise 100. This exercise asks you to prove the **Condorcet Jury Theorem** with sincere voting. A jury of $2n+1$ jurors must vote to convict or acquit the suspect who is equally likely to be guilty or innocent of stealing cookies. If the suspect is guilty, then each juror sees crumbs with probability $p > 1/2$ while if the suspect is innocent, each juror does *not* see crumbs with probability p. The event that one juror sees crumbs, for this exercise, is assumed to be independent of the event that any other juror sees crumbs. Each juror that sees crumbs votes to convict and each that does not see crumbs votes to acquit. The suspect is convicted if and only if a simple majority of jurors votes to convict. Show that as $n \to \infty$, the probability that the jury makes the right decision converges to 1. Then use this result to defend democracy.

Theorem 48. (**central limit theorem**) *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid random variables each with mean μ and variance σ^2 and mgf $M(t)$ that exists for all t in some interval $(-\xi, \xi)$. Suppose that $M''(t)$ is continuous at 0 and let*

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

Then,

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0,1).$$

Proof. Let $Z_n = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$ and $Y_i = (X_i - \mu)/\sigma$ so that $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$. Let

$$\psi(t) := E\left[e^{Y_i t}\right] = E\left[e^{(X_i/\sigma - \mu/\sigma)t}\right] = e^{-\mu t/\sigma} M(t/\sigma)$$

be the mgf of Y_i which exists on the interval $(-\xi\sigma, \xi\sigma)$. Note that $\psi(0) = 1$, $\psi'(0) = 0$ and $\psi''(0) = 1$. By the version of the Taylor's theorem presented in Exercise 61, for $n = 1$, and taking $a = 0$ in that exercise, we have

$$\begin{aligned}\psi(t) &= \psi(0) + \psi'(0)t + \frac{1}{2}\psi''(0)t^2 + \frac{1}{2}(\psi''(\tau(t)) - \psi''(0))t^2 \\ &= 1 + \frac{1}{2}t^2 + \frac{1}{2}\big(\psi''(\tau(t)) - 1\big)t^2\end{aligned}$$

where $\tau(t)$ is a number between 0 and t such that $\lim_{t\to 0}\tau(t) = 0$. Since ϕ'' is continuous at zero (because it is differentiable), we also have $\lim_{t\to 0}\psi''(\tau(t)) = \psi''(0) = 1$. Then, the mgf of Z_n on the interval $(-\xi\sigma, \xi, \sigma)$ is

$$\begin{aligned}M_{Z_n}(t) &= (\phi(t/\sqrt{n}))^n \\ &= \left(1 + \frac{1}{2}(t/\sqrt{n})^2 + \frac{1}{2}\big(\psi''(\tau(t/\sqrt{n})\big) - 1)(t/\sqrt{n})^2\right)^n \\ &= \left(1 + \frac{\frac{1}{2}t^2 + \frac{1}{2}(\psi''(\tau(t/\sqrt{n})) - 1)t^2}{n}\right)^n \qquad (87)\end{aligned}$$

Now as n goes to $+\infty$ the right hand side of this converges to $e^{t^2/2}$, which is the mgf of the standard normal distribution. Then by invoking Theorem 46, the proof is complete. □

Exercise 101. This assertion in the proof above is the same as saying that if $\{z_n\}$ is a real sequence converging to z then $\lim_{n\to\infty}(1 + z_n/n)^n = e^z$. Look back at how you proved exercise 40 and convince yourself that this assertion holds.

Theorem 49. (**the Delta Method**) *Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables such that $\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable at μ. Then,*

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, (g'(\mu))^2\sigma^2).$$

Proof. By the mean value theorem

$$g(X_n) - g(\mu) = (X_n - \mu)g'(\tilde{X}_n)$$

for some $\tilde{X}_n(\omega)$ between μ and $X_n(\omega)$. Since X_n converges in probability to μ, $\tilde{X}_n$ converges in probability to μ as well. Then, there is a theorem called the "continuous mapping theorem" (look it up) that says that $g'(\tilde{X}_n)$ converges in probability to $g'(\mu)$ since g' is continuous. Thus, $\sqrt{n}(g(X_n) - g(\mu))$ converges in distribution to $\mathcal{N}(0, (g'(\mu))^2\sigma^2)$. □

The Delta method generalizes to vectors of random variables in the following way. Let $\{X_n\}$ denote a sequence of vectors of random variables each of size k such that $\sqrt{n}(X_n - \mu)$ converges in distribution to $\mathcal{N}_k(0, \Sigma)$, the k-variate normal distribution, and let $g : \mathbb{R}^k \to \mathbb{R}^l$ be continuously differentiable at the vector μ of size k. Let A denote the $l \times k$ Jacobian matrix of first derivatives of g at μ. Then $\sqrt{n}(g(X_n) - g(\mu))$ converges in distribution to $\mathcal{N}_l(0, A\Sigma A')$.