Notes on STAT-330: Mathematical Statistics

Unversity of Waterloo

David Duan

Last Updated: April 16, 2021 (V1.0)

Contents

1 Univariate Random Variables **1**
1 The Probability Model 1
2 Random Variables and CDFs 2
3 Discrete Random Variables 3
4 Continuous Random Variables 4
5 The Expectation Operator 7
6 The Variance Operator 9
7 Moments of a Random Variable 10
8 Moment Generating Functions 11

2 Multivariate Random Variables **15**
1 Joint and Marginal Cumulative Distribution Functions 15
2 Bivariate Discrete Distributions 16
3 Bivariate Continuous Distributions 18
4 Independent Random Variables 20
5 Joint Expectation 21
6 Covariance and Correlation 22
7 Conditional Distributions 23
8 Conditional Expectation 25
9 Joint Moment Generating Functions 27
10 Multinomial Distribution 28
11 Bivariate Normal Distribution 29

3 Transformations of Random Variables **30**
1 The Cumulative Distribution Function Technique 31
2 Univariate 1-1 Transformation 37
3 Bivariate 1-1 Transformation 38
4 The Moment Generating Function Technique 41
5 Important Distributions 42

4 Limiting/Asymptotic Distributions **45**
1 Convergence in Distribution 45
2 Convergence in Probability 48
3 Weak Law of Large Numbers 49
4 Central Limit Theorem 50
5 More Limit Theorems 51
6 Summary 55

5 Point Estimation **56**
1 Method of Moments 57
2 Maximum Likelihood 58

Chapter 1. Univariate Random Variables

Section 1. The Probability Model

1.1. Definition: The **probability model** consists of three components:

- **sample space**, S, the set of all distinct outcomes of a random experiment;
- **events**, $\mathcal{A}$, a collection of subset of the sample space, known as a *sigma algebra*;
- **probability measure**, $\Pr : \mathcal{A} \to \mathbb{R}$, which assigns, to each event $A \in \mathcal{A}$, a probability $\Pr(A)$. We require the following properties:
  - $\Pr(A) \geq 0$ for all $A \subseteq S$.
  - $\Pr(S) = 1$.
  - For a countable set of events $\{A_1, A_2, A_3, \ldots\} \subseteq S$ that are mutually exclusive,

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i).$$

1.2. Proposition: *Let A, B be events in a sample space S. Then*

- $\Pr(\varnothing) = 0$.
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.
- $\Pr(A \cap B^c) = \Pr(A) - \Pr(A \cap B)$.
- $\Pr(A^c) = 1 - \Pr(A)$.
- $A \subseteq B \implies \Pr(A) \leq \Pr(B)$.
- $0 \leq \Pr(A) \leq 1$.

1.3. Definition: Let A, B be events. The **conditional probability** of A given B is

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \quad \text{provided } \Pr(B) > 0.$$

1.4. Definition: Two events A and B are **independent**, denoted $A \perp B$, if

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

1.5. Intuition: It's helpful to think that A, B are independent iff $\Pr(A \mid B) = \Pr(A)$ and $\Pr(B \mid A) = \Pr(B)$. In other words, the occurrence of one event does not influence the probability of the other.

Section 2. Random Variables and CDFs

1.6. Definition: A **random variable** X is a function from S to $\mathbb{R}$ such that the set $\{X \leq x\} := \{A \in S : X(A) \leq x\}$ is defined (i.e., is a valid event) for all $x \in \mathbb{R}$.

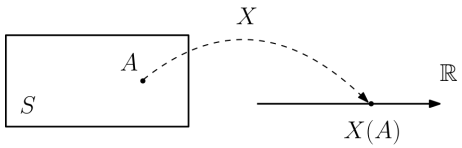

Figure 1.1: Random Variable.

1.7. Intuition: A random variable X assigns each outcome $A \in S$ a value $X(A) \in \mathbb{R}$. Its main purpose is to *quantify the outcomes of a random experiment.* For example, if we let X denote the number of heads among 3 coin flips and the outcome of a given random experiment is $A = \{\mathrm{T}, \mathrm{H}, \mathrm{H}\}$, then we have $A \overset{X}{\mapsto} 2$ (or $X(A) = 2$) as there are two heads.

1.8. Definition: The **cumulative distribution function** (**cdf**) of a random variable X, $F_X : \mathbb{R} \to [0, 1]$, is defined as $F_X(x) = \Pr(X \leq x)$ for $x \in \mathbb{R}$.

1.9. Proposition: *Let F be the cdf of some random variable X.*

(1). F is non-decreasing, i.e., $x_1 \leq x_2 \implies F(x_1) \leq F(x_2)$.
(2). $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.
(3). F is right-continuous, i.e., $\forall a \in \mathbb{R} : \lim_{x\to a^+} F(x) = F(a)$.
(4). $\forall a < b : \Pr(a < X \leq b) = \Pr(X \leq b) - \Pr(X \leq a) = F(b) - F(a)$.
(5). $\forall a \in \mathbb{R} : \Pr(X = a) = \lim_{x\to a^+} F(x) - \lim_{x\to a^-} F(x) = F(a) - \lim_{x\to a^-} F(x)$.

1.10. Example: Suppose that X is a random variable with CDF

$$F(x) = \begin{cases} 0 & x < -2 \\ \frac{x+4}{8} & -2 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

Proposition 1.9 (4) allows us to compute the probability within an interval $a < X \leq b$:

$$\Pr(-1 < X \leq 1) = F(1) - F(-1) = \frac{5}{8} - \frac{3}{8} = \frac{1}{4}$$
$$\Pr(3 < X \leq 4) = F(4) - F(3) = 1 - 1 = 0$$

Proposition 1.9 (5) allows us to compute the probability at a single value $X = a$:

$$\Pr(X = 0) = F(0) - \lim_{x\to 0^-} F(x) = \frac{1}{2} - \frac{1}{2} = 0$$
$$\Pr(X = 2) = F(2) - \lim_{x\to 2^-} F(x) = 1 - \frac{3}{4} = \frac{1}{4}$$

Section 3. Discrete Random Variables

1.11. Definition: A **discrete random variable** takes on a *finite* or *countable* number of values. The cdf of a discrete random variable looks like a *right-continuous step function.*

1.12. Example: Let X be a discrete random variable with cdf

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4} & 0 \leq x < 1 \\ \frac{3}{4} & 1 \leq x < 2 \\ 1 & 2 \leq x \end{cases}$$

Let's plot its cdf:

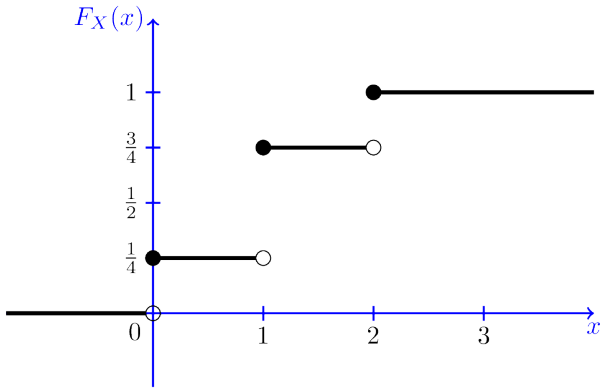

Figure 1.2: The cumulative distribution function of a discrete random variable X.

Observe that each component of the cdf of a discrete random variable will be of the form $a_i \leq x \leq a_{i+1}$, except the first component looks like $x < a_1$ and the last one looks like $a_n \leq x$. Intuitively, the set of discontinuities $A = \{a_1, a_2, \ldots, a_n\}$ are the set of values X can take. As we see below, this set A is called the *support set* of X.

1.13. Definition: The **probability function** or **probability mass function** (**pmf**) of a discrete random variable is given by

$$f(x) = \begin{cases} \Pr(X = x) & \text{if } X \text{ can take value } x \\ 0 & \text{if } X \text{ cannot take value } x \end{cases}$$

The set of values that X can take, $A_X = \{x : f(x) > 0\}$, is called its **support set** of X.

1.14. Proposition: *Let f be the pmf of some discrete random variable X.*

(1). $\forall x \in \mathbb{R} : f(x) \geq 0$.
(2). $\sum_{x \in A_X} f(x) = 1$, *where A_X denotes the support set of X.*
(3). $F(x) = \Pr(X \leq x) = \sum_{x_i \leq x} f(x_i)$, *where F denotes the cdf of X.*

Section 4. Continuous Random Variables

1.15. Definition: Suppose X is a random variable with cdf $F(x)$ such that

- $F(x)$ is *continuous* at every $x \in \mathbb{R}$, and
- F is *differentiable* everywhere except at countably many points (hint: measure zero),

then X is a **continuous random variable**.

1.16. Definition: Let X be a continuous random variable with cumulative distribution function F_X. The **probability density function** (**pdf**) of X is given by

$$f_X(x) = \begin{cases} F_X'(x) & \text{if } F_X(x) \text{ is differentiable at } x \\ 0 & \text{otherwise.} \end{cases}$$

The **support** of X is the set $A_X := \{x : f(x) > 0\}$.

1.17. Example: Let X be a random variable with CDF

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

Since F is differentiable at every $x \in \mathbb{R} \setminus \{1\}$, we get

$$f(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Since $f(x) > 0$ for every $x \in (0, 1)$, the support of X is the open interval $(0, 1)$. □

1.18. Remark: It is important to note that $f(x) \neq \Pr(X = x)$ when X is continuous! In fact, statement (4) below tells us that $\Pr(X = x)$ at a single point $x \in \mathbb{R}$ is always zero.

1.19. Proposition: *Let f be the pdf of a continuous random variable X.*

(1). $\forall x \in \mathbb{R} : f(x) \ge 0$.

(2). $\int_{-\infty}^{\infty} f(x)\,dx = \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x) = 1$.

(3). The probability over an interval is given by the integral of the pdf over that interval:

$$\Pr(a < X \le b) = \Pr(X \le b) - \Pr(X \le a) = F(b) - F(a) = \int_a^b f(x)\,dx.$$

(4). $\Pr(X = b) = 0$ *for all* $b \in \mathbb{R}$.

1.20. Note (From cdf to pdf)**:** Let X be a continuous random variable with cdf $F(x)$. We can find its pdf $f(x)$ by differentiating $F(x)$:

$$f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h\to 0} \frac{\Pr(x \le X \le x+h)}{h} = F'(x),$$

provided the limit exists.

1.21. Example (From cdf to pdf)**:** Consider the following cdf where $b > a$:

$$F(x) = \begin{cases} 0 & x \le a \\ \frac{x-b}{b-a} & a < x \le b \\ 1 & x > b \end{cases} \implies F'(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a < x < b \\ 0 & x > b \end{cases}$$

Note that $F'(x)$ does not exist at $x \in \{a, b\}$ because the one-sided derivatives at $x \in \{a, b\}$ do not match. By definition, $f(x) = 0$ at $x \in \{a, b\}$ and $f(x) = F'(x)$ otherwise, i.e.,

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

1.22. Note (From pdf to cdf)**:** Let X be a continuous random variable with pdf $f(x)$. We can find its cdf $F(x)$ by integrating $f(x)$:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

1.23. Example (From pdf to cdf)**:** Consider the following pdf.

$$f(x) = \begin{cases} \frac{1}{x^2} & x \ge 1 \\ 0 & x < 1. \end{cases}$$

To verify this is a valid pdf, observe that $f(x) \ge 0$ for all $x \in \mathbb{R}$ and

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{1}^{\infty} \frac{1}{x^2}\, dx = -\frac{1}{x}\Big|_{1}^{\infty} = 1.$$

To find its cdf, let us integrate $f(x)$:

$$x < 1 : F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{x} 0\, dt = 0$$
$$x \ge 1 : F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(t)\, dt = \int_{1}^{x} \frac{1}{t^2}\, dt = 1 - \frac{1}{x}$$

Therefore, the cdf of X is given by

$$F(x) = \begin{cases} 1 - 1/x & x \ge 1, \\ 0 & x < 1. \end{cases}$$

1.24. Example: Continuing from above, let us demonstrate two ways of computing

$$\Pr(-2 < X < 3).$$

First, by Proposition 1.9 (4), we have

$$\Pr(-2 < X < 3) = \Pr(-2 < X \leq 3) = F(3) - F(-2) = \left(1 - \frac{1}{3}\right) - 0 = \frac{2}{3}.$$

Alternatively, by 1.19 (3), we could integrate $F(x)$ over the interval $(-2, 3)$ and obtain

$$\Pr(-2 < X < 3) = \int_{-2}^{3} f(x)\,dx = \int_{1}^{3} \frac{1}{x^2}\,dx = -\frac{1}{x}\Big|_{1}^{3} = 1 - \frac{1}{3} = \frac{2}{3}.$$

1.25. Before we conclude this section, let us introduce the Gamma function. It appears in the pdf of many famous distributions and its properties often help you evaluate integrals in probability theory.

1.26. Definition: The **Gamma function**, denoted $\Gamma(\alpha)$, is defined as

$$\Gamma(\alpha) = \int_{0}^{\infty} y^{\alpha-1} e^{-y} dy \qquad \alpha > 0.$$

1.27. Proposition: *Useful properties of the Gamma function:*

(1). $\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$ *for all* $\alpha > 1$.

(2). $\Gamma(n) = (n - 1)!$ *for all* $\alpha \in \mathbb{Z}_+$.

(3). $\Gamma(1/2) = \sqrt{\pi}$.

Section 5. The Expectation Operator

1.28. Definition: Let X be a *discrete* random variable with support A and pdf $f(x)$. The **expectation** or **expected value** of X is given by

$$\mathbb{E}[X] = \sum_{x \in A} x f(x) \qquad \text{provided} \qquad \sum_{x \in A} |x| f(x) < \infty.$$

If the series does not converge absolutely, then $\mathbb{E}[X]$ does not exist.

1.29. Example: Let X be a discrete random variable with pdf

$$f(x) = \frac{1}{x(x+1)}, \qquad x = 1, 2, \ldots$$

Then $A = \{1, 2, \ldots\}$. We first verify that $f(x)$ is a valid pdf:

$$\sum_{x \in A} f(x) = \sum_{x=1}^{\infty} \left(\frac{1}{x} - \frac{1}{x+1} \right) = 1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \cdots = 1.$$

We now check if its expectation exists:

$$\sum_{x \in A} |x| f(x) = \sum_{x=1}^{\infty} x \frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \frac{1}{x+1} = \infty.$$

Thus, $\mathbb{E}[X]$ does not exist.

1.30. Definition: Let X be a *continuous* random variable with support A and pdf $f(x)$. The **expectation** or **expected value** of X is given by

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx \qquad \text{provided} \qquad \int_{-\infty}^{\infty} |x| f(x)\, dx < \infty$$

If the integral does not converge absolutely, then $\mathbb{E}[X]$ does not exist.

1.31. Example: Let X be a continuous random variable with pdf

$$f(x) = \frac{1}{\pi(x^2+1)} \qquad x \in \mathbb{R}.$$

First, let's check this is a valid pdf:

$$\int_{-\infty}^{\infty} \frac{1}{\pi(x^2+1)}\, dx = 1.$$

We again observe that the expectation does not exist for this X:

$$\int_{-\infty}^{\infty} |x| f(x)\, dx = \int_{-\infty}^{\infty} |x| \frac{1}{x^2+1}\, dx = 2 \int_0^{\infty} \frac{x}{x^2+1}\, dx = \log(x^2+1) \Big|_0^{\infty} = \infty.$$

1.32. Warning: Thus, always verify that the series/integral converges absolutely first!

1.33. Let us now look at the expectation of *functions* of random variables. The expected value operator can be viewed as a special case where $g(X) = I$, the identity function.

1.34. Definition: Let X be a discrete random variable with pdf $f(x)$ and support A. Let g be a function of X. Then

$$\mathbb{E}[g(X)] = \sum_{x \in A} g(x) f(x) \qquad \text{provided} \qquad \sum_{x \in A} |g(x)| f(x) < \infty.$$

Otherwise, $\mathbb{E}[g(x)]$ does not exist.

1.35. Definition: Let X be a continuous random variable with pdf $f(x)$ and support A. Let g be a function of X. Then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \qquad \text{provided} \qquad \int_{-\infty}^{\infty} |g(x)| f(x)\, dx < \infty.$$

Otherwise, $\mathbb{E}[g(x)]$ does not exist.

1.36. Proposition: *Let X be a random variable, $a, b, c \in \mathbb{R}$ be constants, and g, h be functions of X. Then*

$$\mathbb{E}[ag(X) + bh(X) + c] = a\mathbb{E}[g(X)] + b\mathbb{E}[h(X)] + c.$$

In other words, the expectation operator is linear.

Proof. By linearity of summation and integral. □

Section 6. The Variance Operator

1.37. Definition: Let X be a random variable. The **variance** of X is the *expected value of the squared deviation from the mean* of X, i.e.,

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

1.38. Proposition: *Let X be a random variable. Then*

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

Proof.

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

□

1.39. Proposition: *Let X be a random variable.*

- *The variance of a constant is zero (indeed, there is no deviation at all), i.e., for $a \in \mathbb{R}$,*

$$\mathrm{Var}(a) = 0.$$

- *The variance is non-negative, because the squares are non-negative:*

$$\mathrm{Var}(X) \geq 0.$$

- *The variance is invariant wrt changes in a location parameter, i.e., for $a \in \mathbb{R}$,*

$$\mathrm{Var}(X + a) = \mathrm{Var}(X).$$

- *If all values are scaled by a constant, the variance is scaled by squared of that constant:*

$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X).$$

- *The variance of a sum of two random variables is given by*

$$\begin{aligned}\mathrm{Var}(aX + bY) &= a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y), \\ \mathrm{Var}(aX - bY) &= a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) - 2ab\,\mathrm{Cov}(X, Y).\end{aligned}$$

- *Since independent random variables are uncorrelated, for $X_1, \ldots, X_n$ independent,*

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

The variance of the mean of $X_1, \ldots, X_n$ is given by

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n}.$$

Section 7. Moments of a Random Variable

1.40. The *moments* of a function are quantitative measures related to the shape of the function's graph. For a probability distribution on a bounded interval, the collection of all the moments (of all orders, from 0 to ∞) uniquely determines the distribution.[1] Expectation and variance discussed in the previous two sections are two very special moments.

1.41. Definition: Let X be a random variable.

- The kth **moment** of X is given by $\mathbb{E}[X^k]$ for $k = 1, 2, \ldots$
  - Also known as the k**th moment about the origin**.
  - The 1st moment of X is the **mean** of X:
$$\mu_X = \mathbb{E}[X].$$
- The kth **central moment** of X is given by $\mathbb{E}[(X - \mu)^k]$ for $k = 1, 2, \ldots$
  - Also known as the k**th moment about the mean**.
  - The 2nd central moment of X is the **variance** of X:
$$\mathrm{Var}(X) = \sigma^2 = \mathbb{E}[(X - \mu)^2].$$

[1] The same is not true on unbounded intervals.

Section 8. Moment Generating Functions

1.42. So far, we have seen two types of functions that uniquely determines a distribution: pdf and cdf. A third type of functions, known as *moment generating functions*, also uniquely determines a distribution. Moment generating functions provide the basis of an alternative route to analytical results compared with working directly with pdfs and cdfs. As its name implies, this function can be used to compute a distribution's moments.

1.43. Definition: Let X be a random variable. The function

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

is called the **moment generating function** (**mgf**) if $\mathbb{E}[e^{tX}]$ exists for all t in some neighbourhood around 0, i.e., for all $t \in (-h, h)$ for some $h > 0$.

1.44. Note: We now demonstrate how to derive the mgf given the pdf of a random variable. Recall that for a random variable X with pdf $f(x)$ and support A, the expectation of a function $g(X)$ is given by

$$\mathbb{E}[g(X)] = \begin{cases} \displaystyle\sum_{x \in A} g(x) f(x) & x \text{ is discrete,} \\ \displaystyle\int_{x \in A} g(x) f(x)\, dx & x \text{ is continuous.} \end{cases}$$

Let $g(X) = e^{tX}$. Then

$$\mathbb{E}[e^{tX}] = \begin{cases} \displaystyle\sum_{x \in A} e^{tx} f(x) & x \text{ is discrete,} \\ \displaystyle\int_{x \in A} e^{tx} f(x)\, dx & x \text{ is continuous.} \end{cases}$$

1.45. Example: Let $X \sim \text{Poisson}(\lambda)$ with pdf $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x \in \mathbb{Z}_{\geq 0}$. Then

$$\begin{aligned} M(t) = \mathbb{E}[e^{tX}] = \sum_{x=0}^{\infty} e^{tx} f(x) &= \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x e^{tx}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda} e^{\lambda e^t} && e^y = \sum_{n=0}^{\infty} \frac{y^n}{n!} \\ &= e^{\lambda(e^t - 1)} && \forall t \in \mathbb{R}. \end{aligned}$$

1.46. Let $Y = aX + b$. The following proposition gives us a way to directly derive $M_Y(t)$ given $M_X(t)$ without going through the computation involving expected values again.

1.47. Proposition: *Let X be a random variable with mgf $M_X(t)$ that exists for all $t \in (-h, h), h > 0$. Define $Y = aX + b$ for $a, b \in \mathbb{R}$, $a \neq 0$. Then the mgf for Y is given by*

$$M_Y(t) = e^{bt} M_X(at), \qquad t \in \left(-\frac{h}{|a|}, \frac{h}{|a|} \right).$$

Proof. Observe

$$\begin{aligned} M_Y(t) = \mathbb{E}[e^{tY}] &= \mathbb{E}[e^{t(aX+b)}] \\ &= e^{bt}\mathbb{E}[e^{taX}] && \text{exists for } |at| < h \\ &= e^{bt} M_X(at). && \text{for } |t| < \frac{h}{|a|} \end{aligned}$$

Pay attention to the third line: $M_X(t_X)$ is defined for all $t_X \in (-h, h)$, which implies that

$$\mathbb{E}[e^{taX}] = \mathbb{E}[e^{t_X X}]$$

is defined only if $t_X = ta \in (-h, h)$, or equivalently, $|at| < h$. Since a and h are fixed, we require $|t| < h/|a|$. This is the domain where $M_Y(t)$ is defined. □

1.48. Example: Let us derive the mgf of $X \sim N(\mu, \sigma^2)$. First, recall that $X = \sigma Z + \mu$ where $Z \sim N(0, 1)$. The mgf of the standard normal Z is given by

$$\begin{aligned} M_Z(t) = \mathbb{E}[e^{tZ}] &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2 + 2tx}{2} \right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-t)^2 + t^2}{2} \right) dx \\ &= \exp\left(\frac{t^2}{2} \right) \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-t)^2}{2} \right)}_{\text{pdf for } N(t,1), \text{ thus integrate to 1}} dx \\ &= \exp\left(\frac{t^2}{2} \right). \end{aligned}$$

Now use the proposition above,

$$\begin{aligned} M_X(t) &= e^{\mu t} M_Z(\sigma t) \\ &= e^{\mu t} \exp\left(\frac{(\sigma t)^2}{2} \right) \\ &= \exp\left(\mu t + \frac{\sigma^2 t^2}{2} \right). \end{aligned}$$

1.49. Note: The following proposition gives us a way of computing the kth moment about the origin. In particular, given the mgf of X, we can find its mean and variance by

(1). Calculate the first and second derivative $M'_X(t)$ and $M''_X(t)$ of $M_X(t)$.
(2). The mean of is given by $\mathbb{E}[X] = M'_X(0)$.
(3). The variance is given by $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X] = M''_X(0) - (M'_X(0))^2$.

1.50. Proposition: *Let X be a random variable with mgf $M(t)$ defined on $t \in (-h, h)$ for $h > 0$. Then $M(0) = 1$ and for $k = 1, 2, \ldots$, the kth moment about the origin is given by*

$$\mathbb{E}[X^k] = M^{(k)}(0)$$

where

$$M^{(k)}(t) := \frac{d^k}{dt^k} M(t)$$

is the kth derivative of $M(t)$.

Proof. Note that $M(0) = \mathbb{E}[X^0] = \mathbb{E}[1] = 1$ and that

$$\frac{d^k}{dt^k} e^{tx} = x^k e^{tx} \qquad \text{for } k = 1, 2, \ldots \tag{1.1}$$

Let X be a continuous r.v. with mgf $M(t)$ defined for $t \in (-h, h)$ for some $h > 0$, then

$$M^{(k)}(t) = \frac{d^k}{dt^k}\mathbb{E}[e^{tX}] = \frac{d^k}{dt^k}\int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \int_{-\infty}^{\infty} \frac{d^k}{dt^k} e^{tx} f(x)\, dx \qquad k = 1, 2, \ldots$$

Note that we are allowed to interchange differentiation and integration here (proof omitted). Using (1.1), we have

$$M^{(k)}(t) = \int_{-\infty}^{\infty} x^k e^{tx} f(x)\, dx = \mathbb{E}[X^k e^{tX}] \qquad t \in (-h, h) \text{ for some } h > 0.$$

Letting $t = 0$, we obtain $M^{(k)}(0) = \mathbb{E}[X^k], k = 1, 2 \ldots$ as required. □

1.51. Example: Let us derive the mean and variance for a random variable $X \sim \text{Poisson}(\lambda)$. Recall in Example 1.48, we derived that the mgf of X is given by

$$M_X(t) = \exp(\lambda(e^t - 1)).$$

Compute its first and second derivatives:

$$M'_X(t) = \exp(\lambda(e^t - 1))\lambda e^t = e^{\lambda(e^t - 1) + t},$$
$$M''_X(t) = \lambda e^{\lambda(e^t - 1) + t}(\lambda e^t + 1).$$

By the proposition above, we see that

$$\mathbb{E}[X] = M'_X(0) = \lambda,$$
$$\text{Var}[X] = \mathbb{E}[X^2] - \mu^2 = M''(0) - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

1.52. The following result is often known as the *uniqueness of mgf* and can be used to show that a random variable follows a certain distribution.

1.53. Proposition: *Let X, Y be random variables with mgfs $M_X(t)$, $M_Y(t)$, respectively. Then the mgfs coincide in a neighbourhood around 0 iff X and Y have the same distribution, i.e.,*

$$\begin{aligned} & \exists h > 0, \forall t \in (-h, h) : M_X(t) = M_Y(t) \\ \iff & \forall s \in \mathbb{R} : \Pr(X \leq s) = F_X(s) = F_Y(s) = \Pr(Y \leq s). \end{aligned}$$

Proof. Omitted. □

Chapter 2. Multivariate Random Variables

Section 1. Joint and Marginal Cumulative Distribution Functions

2.1. Definition: Let X and Y be random variables defined on a sample space S. The **joint cumulative distribution function** of X and Y is given by

$$F(x,y) = \Pr(X \leq x, Y \leq y), \quad \forall (x,y) \in \mathbb{R}^2.$$

2.2. Remark: This notion is well-defined as both $\{X \leq x\}$ and $\{Y \leq y\}$ are valid events, so their intersection $\{X \leq x, Y \leq y\}$ is also valid.

2.3. Proposition:

(1). Fix y, F is non-decreasing in x. Similarly, fix x, F is non-decreasing in y.

(2). $\lim_{x \to -\infty} F(x,y) = 0 = \lim_{y \to -\infty} F(x,y)$.

(3). $\lim_{(x,y) \to (-\infty,-\infty)} F(x,y) = 0$ *and* $\lim_{(x,y) \to (\infty,\infty)} F(x,y) = 1$.

2.4. Definition: The **marginal cumulative distribution function of** X is given by

$$F_X(x) = \lim_{y \to \infty} F(x,y) = \Pr(X \leq x) \qquad \forall x \in \mathbb{R}.$$

Similarly, the **marginal cumulative distribution function of** Y is given by

$$F_Y(y) = \lim_{x \to \infty} F(x,y) = \Pr(Y \leq y) \qquad \forall y \in \mathbb{R}.$$

2.5. Warning: Note that given joint cdfs, we can find marginal pdfs. But, given marginal pdfs, we cannot find the joint cdfs. In other words, it's possible to have (X_1, Y_1) and (X_2, Y_2) such that $F_{X_1}(x) = F_{X_2}(x)$ and $F_{Y_1}(y) = F_{Y_2}(y)$ but $F_{X_1,Y_1}(x,y) \neq F_{X_2,Y_2}(x,y)$.

Section 2. Bivariate Discrete Distributions

2.6. Definition: Let X and Y be random variables defined on sample space S. If there exists $A \subseteq \mathbb{R}^2$ such that A is countable and $\Pr((x,y) \in A) = 1$, then X and Y are a pair of **bivariate discrete random variables**.

2.7. Definition: The **joint pmf** of discrete random variables X and Y is given by

$$f(x,y) = \Pr(X = x, Y = y) \qquad \forall (x,y) \in \mathbb{R}^2.$$

The **joint support** of (X,Y) is given by

$$A = \{(x,y) : f(x,y) > 0\}.$$

2.8. Proposition:

(1). $f(x,y) \geq 0$ *for all* $(x,y) \in \mathbb{R}^2$.
(2). $\sum_{(x,y)\in A} f(x,y) = 1$.
(3). For $R \subseteq \mathbb{R}^2$, $\Pr((x,y) \in R) = \sum_{(x,y)\in R} f(x,y)$.

2.9. Definition: Let $f(x,y)$ be the joint pmf for X, Y. Then the **marginal pmfs** are obtained by summing out the other variable, i.e.,

$$f_X(x) = \Pr(X = x) = \sum_y f(x,y) \qquad \forall x \in \mathbb{R},$$

$$f_Y(y) = \Pr(Y = Y) = \sum_x f(x,y) \qquad \forall y \in \mathbb{R}.$$

2.10. Example: Let $p \in (0,1)$ and X, Y be discrete random variables with joint pmf

$$f(x,y) = \begin{cases} k(1-p)^2 p^{x+y} & x \geq 0, y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(1). Find the value of k.

First, $f(x,y) \geq 0$ so $k \geq 0$. Next,

$$\begin{aligned} & \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} f(x,y) = 1 \\ \implies & \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} k(1-p)^2 p^x p^y = 1 \\ \implies & k(1-p)^2 \left(\sum_{x=0}^{\infty} p^x\right)\left(\sum_{y=0}^{\infty} p^y\right) = k(1-p)^2 \left(\frac{1}{(1-p)^2}\right) = k \\ \implies & k = 1. \end{aligned}$$

(2). Find marginal pmfs.

$$f_X(x) = \sum_{y=0}^{\infty}(1-p)^2p^{x+y} = (1-p)^2p^x\sum_{y=0}^{\infty}p^y = (1-p)p^x, \qquad x = 0, 1, 2, \ldots$$
$$f_Y(x) = (1-p)p^y, \qquad x = 0, 1, 2, \ldots$$

We conclude that X and Y marginally follow a geometric distribution.

(3). Find $\Pr(X \leq Y)$.

$$\begin{aligned}
\Pr(X \leq Y) &= \sum_{x=0}^{\infty}\sum_{y=x}^{\infty}(1-p)^2p^{x+y} \\
&= (1-p)^2\sum_{x=0}^{\infty}p^x\sum_{y=x}^{\infty}p^y \\
&= (1-p)^2\sum_{x=0}^{\infty}p^x\left(p^x\sum_{y=1}^{\infty}p^y\right) \\
&= (1-p)^2\sum_{x=0}^{\infty}p^x\left(p^x\frac{1}{1-p}\right) \\
&= \frac{(1-p)^2}{1-p}\sum_{x=0}^{\infty}p^{2x} \\
&= (1-p)\frac{1}{1-p^2} \\
&= \frac{1}{1+p}.
\end{aligned}$$

Section 3. Bivariate Continuous Distributions

2.11. Definition: If $F(x,y)$ is continuous and the derivative

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

exists and is continuous except along a finite number of curves, then we say that X, Y are **bivariate continuous** and we define its **joint pdf** to be

$$f(x,y) = \begin{cases} \dfrac{\partial^2}{\partial x \partial y} F(x,y) & \text{if exists} \\ 0 & \text{otherwise} \end{cases}$$

The **joint support** of (x,y) is given by

$$A = \{(x,y) : f(x,y) > 0\}.$$

2.12. Proposition:

(1). $f(x,y) \geq 0$ *for all* $(x,y) \in \mathbb{R}^2$.
(2). $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) = 1$.
(3). For $R \subseteq \mathbb{R}^2$, $\Pr((x,y) \in R) = \iint_R f(x,y)\, dx\, dy$.

2.13. Definition: Let $f(x,y)$ be the joint pdf of X, Y. Then the **marginal pdfs** are obtained by integrating out the other variable, i.e.,

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

2.14. Note: To evaluate a double integral:

(1). Integrate over y then x:

$$\int \left[\int f(x,y)\, dy \right] dx$$

(2). Integrate over x then y:

$$\int \left[\int f(x,y)\, dx \right] dy$$

To figure out the bounds for the integrals using approach 1 (mirror for approach 2):

(1). Outer integral (over x): figure out the range of x in the region.
(2). Inner integral (over y): fix x, figure out the range of y in the region.

2.15. Example: Suppose that (X, Y) is a pair of continuous variables with joint pdf

$$f(x,y) = \begin{cases} 1, & 0 < x < 1, 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(1). Find $\Pr(X \leq Y)$.

$$\begin{aligned}
\Pr(X \leq Y) &= \Pr(X - Y \leq 0) \\
&= \iint_{x-y\leq 0} f(x,y)\,dx\,dy \\
&\overset{1}{=} \int_0^1 \int_x^1 1\,dy\,dx \\
&= \int_0^1 y\Big|_x^1\,dx \\
&= \int_0^1 (1-x)\,dx \\
&= \int_0^1 1\,dx - \int_0^1 x\,dx \\
&= x\Big|_0^1 - \frac{1}{2}x^2\Big|_0^1 \\
&= 1 - \frac{1}{2} = \frac{1}{2}.
\end{aligned}$$

Focus on $\overset{1}{=}$. The outer bound is easy as we know $x \in (0,1)$. For the inner bound, fix $x \in (0,1)$, we see that $y \in [x, 1)$ satisfies the inequality $x - y \leq 0$. This gives us the inner bound.

(2). Find the marginal pmfs.

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^1 1\,dy = y\Big|_0^1 = 1 \implies f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^1 1\,dx = x\Big|_0^1 = 1 \implies f_Y(y) = \begin{cases} 1 & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Section 4. Independent Random Variables

2.16. Recall that two events A and B are **independent** iff $\Pr(A \cap B) = \Pr(A)\Pr(B)$.

2.17. Definition: Two random variables X and Y are **independent** iff

$$\forall A, B \subseteq \mathbb{R} : \Pr(X \in A, Y \in B) = \Pr(X \in A)\Pr(Y \in B).$$

2.18. Theorem: *Let X, Y be random variables.*

- *Let $F(x, y)$ be the joint cdf and $F_X(x)$, $F_Y(y)$ be the marginal cdfs. Then*

$$X \perp Y \iff \forall (x, y) \in \mathbb{R}^2 : F(x, y) = F_X(x)F_Y(y).$$

- *Let $f(x, y)$ be the joint pdf/pmf and $f_X(x), f_Y(y)$ be the marginal pdfs/pmfs. Define the supports $A_X = \{x : f_X(x) > 0\}$ and $A_Y = \{y : f_y(y) > 0\}$. Then*

$$X \perp Y \iff \forall (x, y) \in A_X \times A_Y : f(x, y) = f_X(x)f_Y(y).$$

2.19. Theorem (Factorization Theorem for Independence)**:** *Let X, Y be random variables with joint pdf/pmf $f(x, y)$ and joint support A. Let A_X, A_Y be the support of X, Y, respectively. Then*

$$X \perp Y \iff \exists g(x) \geq 0, h(y) \geq 0 : f(x, y) = g(x)h(y)$$

for all $(x, y) \in A_1 \times A_2$.

2.20. Remark:

- If RHS holds, then $f_X(x) \propto g(x)$ and $f_Y(y) \propto h(y)$.
- If A is not rectangular, then X and Y must be dependent. Indeed, not rectangular means there exists $(x, y) \notin A$ such that $x \in A_1, y \in A_2$. This means that $f_X(x) > 0$, $f_Y(y) > 0$, but $f(x, y) = 0$. Therefore, $f(x, y) \neq f_X(x)f_Y(y)$ for this (x, y).

2.21. Theorem: *If X, Y are independent random variables and g, h are functions, then $g(X), h(Y)$ are independent.*

2.22. Remark: Note the reverse does not always hold, that is, we could have $g(X)$ and $h(Y)$ independent for some g, h but X and Y are dependent.

Section 5. Joint Expectation

2.23. Definition: Suppose X, Y are bivariate discrete and $h(X, Y)$ is a function. Then

$$\mathbb{E}[h(X,Y)] = \sum_{(x,y)\in A} h(x,y)f(x,y)$$

provided that the series converges absolutely:

$$\sum_{(x,y)\in A} |h(x,y)|f(x,y) < \infty.$$

Otherwise, we say that $\mathbb{E}[h(X,Y)]$ DNE.

2.24. Definition: Suppose X, Y are bivariate discrete and $h(X, Y)$ is a function. Then

$$\mathbb{E}[h(X,Y)] = \iint_{(x,y)\in A} h(x,y)f(x,y)\,dx\,dy$$

provided that the integral converges absolutely:

$$\iint_{(x,y)\in A} |h(x,y)|f(x,y)\,dx\,dy < \infty.$$

Otherwise, we say that $\mathbb{E}[h(X,Y)]$ DNE.

2.25. Proposition (Linearity of Expectation)**:** *For random variables $X_1, \ldots, X_n$,*

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i] \qquad a_1, \ldots, a_n \in \mathbb{R}.$$

2.26. Proposition: *If $X_1, \ldots, X_n$ are independent, then*

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

In particular, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Section 6. Covariance and Correlation

2.27. Motivation: *Covariance* is a measure of the joint probability of two random variables. The sign of the covariance shows the tendency in the linear relationship between the variables. The normalized version of the covariance, the *correlation coefficient*, gives the strength of the linear relation.

2.28. Definition: The **Covariance** of X and Y is given by

$$\begin{aligned}\sigma_{XY} = \mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

When $\mathrm{Cov}(X,Y) = 0$, we say that X and Y are **uncorrelated**.

2.29. Definition: The **correlation coefficient** of X and Y is given by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}} \in [-1,1].$$

In particular, $\rho(X,Y) = \pm 1$ indicates that X and Y have a perfect linear relationship.

2.30. Proposition:

(1). $X \perp Y \implies \mathrm{Cov}(X,Y) = 0$.
(2). $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
(3). $\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X,Y)$.
(4). $\mathrm{Var}(\sum_{i=1}^n a_iX_i) = \sum_{i=1}^n a_i^2\mathrm{Var}[X_i] + \sum_{i\neq j} a_ia_j\mathrm{Cov}(X_iX_j)$.
(5). If $X_1,\ldots,X_n$ *are independent,* $\mathrm{Var}(\sum_{i=1}^n a_iX_i) = \sum_{i=1}^n a_i^2\mathrm{Var}(X_i)$.

2.31. Remark: Focus on the first statement. If two random variables are independent, then there does not exist *any* relationship between them. In particular, no linear relationship exists. Therefore X and Y are uncorrelated.

2.32. Remark: Here's an example of the last property. Let $X_1,\ldots,X_n$ be independent with $\mathrm{Var}[X_i] = \sigma^2$ for all i. Then

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

Section 7. Conditional Distributions

2.33. Definition: Suppose X and Y are bivariate discrete random variables with joint pmf $f(x, y)$. The **conditional pmf** of X given $Y = y$ is

$$f_X(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \quad \text{provided } f_Y(y) > 0,$$

where $f_Y(y)$ is the marginal pmf of Y. We can interpret this as

$$\Pr(X = x \mid Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)}, \quad \text{provided } \Pr(Y = y) > 0.$$

The **conditional pmf** of Y given $X = x$ is

$$f_Y(y \mid x) = \frac{f(x, y)}{f_X(x)}, \quad \text{provided } f_X(x) > 0,$$

where $f_X(x)$ is the marginal pmf of X. We can interpret this as

$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}, \quad \text{provided } \Pr(X = x) > 0.$$

2.34. Proposition: *$f_X(x \mid y)$ and $f_Y(y \mid x)$ are valid probability distributions, i.e.,*

- $f_X(x \mid y) \geq 0$ *and* $\sum_x f_X(x \mid y) = 1$.
- $f_Y(y \mid x) \geq 0$ *and* $\sum_y f_Y(y \mid x) = 1$.

2.35. Definition: Suppose X and Y are bivariate continuous random variables with joint pdf $f(x, y)$. Then the **conditional pdf** of X given $Y = y$ is

$$f_X(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \quad \text{provided } f_Y(y) > 0.$$

The **conditional pmf** of Y given $X = x$ is

$$f_Y(y \mid x) = \frac{f(x, y)}{f_X(x)}, \quad \text{provided } f_X(x) > 0.$$

2.36. Remark: One can show that

$$\Pr(X \leq x \mid Y = y) = \int_{-\infty}^{x} f_X(t \mid y)\, dt$$
$$\Pr(Y \leq y \mid X = x) = \int_{-\infty}^{y} f_Y(t \mid x)\, dt$$

2.37. Proposition: *$f_X(x \mid y)$ and $f_Y(y \mid x)$ are valid probability distributions, i.e.,*

- $f_X(x \mid y) \geq 0$ *and* $\int_{-\infty}^{\infty} f_X(x \mid y) = 1$.
- $f_Y(y \mid x) \geq 0$ *and* $\int_{-\infty}^{\infty} f_Y(y \mid x) = 1$.

2.38. Proposition: *Let X, Y be random variables with marginal pdfs/pmfs $f_X(x), f_Y(y)$, marginal supports A_X, A_Y, conditional pdfs/pmfs $f_X(x \mid y)$ and $f_Y(y \mid x)$. Then*

$$X \perp Y \iff \forall x \in A_X : f_X(x \mid y) = f_X(x) \wedge \forall y \in A_Y : f_Y(y \mid x) = f_Y(y).$$

Proof. Recall that X and Y are independent iff $f(x, y) = f_X(x) f_Y(y)$. □

2.39. Theorem: $f(x, y) = f_X(x \mid y) f_Y(y) = f_Y(y \mid x) f_X(x)$.

Proof. This follows directly from

$$f_X(x \mid y) = \frac{f(x, y)}{f_Y(y)}.$$

□

Section 8. Conditional Expectation

2.40. Definition: Let Y be a random variable and $g(Y)$ be a function. The **conditional expectation** of $g(Y)$ given $X = x$ is

$$\mathbb{E}[g(Y) \mid X = x] = \begin{cases} \sum_y g(y) f_Y(y \mid x) & \text{provided that } \sum_y |g(y)| f_Y(y \mid x) < \infty \\ \int_{-\infty}^{\infty} g(y) f_Y(y \mid x) & \text{provided that } \int_{-\infty}^{\infty} |g(y)| f_Y(y \mid x) < \infty \end{cases}$$

2.41. Definition:

- For $g(y) = y$, $\mathbb{E}[Y \mid X = x]$ is called the **conditional mean**.
- For $g(y) = (y - \mathbb{E}[Y \mid X = x])^2$,

$$\begin{aligned} \operatorname{Var}[Y \mid X = x] &= \mathbb{E}[(Y - \mathbb{E}[Y \mid X = x])^2 \mid X = x] \\ &= \mathbb{E}[Y^2 \mid X = x] - [\mathbb{E}[Y \mid X = x]]^2 \end{aligned}$$

  is called the **conditional variance**.

2.42. Proposition: *If X and Y are independent, then*

$$\forall g, \forall h : \mathbb{E}[g(X) \mid Y = y] = \mathbb{E}[g(X)] \quad \wedge \quad \mathbb{E}[h(Y) \mid X = x] = \mathbb{E}[h(Y)].$$

Proof. Observe that

$$\mathbb{E}[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x) f_X(x \mid y)\, dx = \int_{-\infty}^{\infty} g(x) f_X(x) = \mathbb{E}[g(X)]$$

as X and Y are independent. The other statement is similar. □

2.43. Corollary: *If X and Y are independent, then*

$$\begin{aligned} \mathbb{E}[Y \mid X = x] &= \mathbb{E}[Y] \\ \operatorname{Var}[Y \mid X = x] &= \mathbb{E}[Y^2 \mid X = x] - \mathbb{E}^2[Y \mid X = x] = \mathbb{E}[Y^2] - \mathbb{E}^2[Y] = \operatorname{Var}[Y]. \end{aligned}$$

2.44. Theorem (Substitution Rule)**:**

$$\mathbb{E}[h(X, Y) \mid X = x] = \mathbb{E}[h(x, Y) \mid X = x].$$

2.45. Example:

$$\begin{aligned} \mathbb{E}[X + Y \mid X = x] &= \mathbb{E}[x + Y \mid X = x] = x + \mathbb{E}[Y \mid X = x] \\ \mathbb{E}[XY \mid X = x] &= \mathbb{E}[xY \mid X = x] = x\,\mathbb{E}[Y \mid X = x]. \end{aligned}$$

2.46. Remark: Note that $\mathbb{E}[g(X) \mid Y] \neq \mathbb{E}[g(X) \mid Y = y]$. LHS is a random variable (as it's a function of Y) while RHS is a scalar value.

2.47. Theorem (Double-Expectation Formula)**:**

$$\mathbb{E}[\mathbb{E}[g(X) \mid Y]] = \mathbb{E}[g(X)].$$

Proof.

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[g(X) \mid Y]] &= \mathbb{E}\left[\int_{-\infty}^{\infty} g(x) f_X(x \mid Y)\, dx\right]. \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(x) f_X(x \mid Y) dx\right] f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) \underbrace{f_X(x \mid y) f_Y(y)}_{f(x,y)}\, dx\, dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f(x,y)\, dx\, dy. \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f(x,y)\, dy\, dx. \\
&= \int_{-\infty}^{\infty} g(x) \underbrace{\left[\int_{-\infty}^{\infty} f(x,y)\, dy\right]}_{f_X(x)} dx. \\
&= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx = \mathbb{E}[g(x)].
\end{aligned}$$

□

2.48. Theorem (Law of Total Variance)**:**

$$\mathrm{Var}[Y] = \mathbb{E}[\mathrm{Var}[Y \mid X]] + \mathrm{Var}[\mathbb{E}[Y \mid X]].$$

Section 9. Joint Moment Generating Functions

2.49. Definition: Let X, Y be a pair of random variables. If there exist $h_X, h_Y \in \mathbb{R}_+$ such that $\mathbb{E}[e^{t_X X + t_Y Y}]$ exists for $t_X \in (-h_X, h_X)$ and $t_Y \in (-h_Y, h_Y)$, then

$$M(t_X, t_Y) = \mathbb{E}[e^{t_X X + t_Y Y}]$$

is called the **joint mgf** of X and Y. More generally, the joint mgf of n random variables $X_1, \ldots, X_n$ is given by

$$M(t_1, \ldots, t_n) = \mathbb{E}[e^{\sum_{i=1}^n t_i X_i}]$$

provided that $\exists h_1, \ldots, h_n > 0$ such that $\mathbb{E}[e^{\sum t_i X_i}]$ exists for all $t_i \in (-h_i, h_i)$, $i = 1, \ldots, n$.

2.50. Proposition: *Given $M(t_1, t_2)$, we can find the marginal mgfs by*

$$M_X(t_X) = M(t_X, 0) = \mathbb{E}[e^{t_X X + 0Y}] = \mathbb{E}[e^{t_X X}]$$
$$M_Y(t_Y) = M(0, t_Y) = \mathbb{E}[e^{0X + t_Y Y}] = \mathbb{E}[e^{t_Y Y}].$$

2.51. Proposition: *Let X, Y be a pair of random variables with MGF $M(t_X, t_Y)$, then*

$$X \perp Y \iff M(t_X, t_Y) = M_X(t_X) M_Y(t_Y).$$

More generally, $X_1, \ldots, X_n$ are independent iff $M(t_1, \ldots, t_n) = \prod_i M_{X_i}(t_i)$.

Section 10. Multinomial Distribution

2.52. The *multinomial distribution* is a generalization of the binomial distribution. For n independent trials each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability p_k, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.

2.53. Definition: Let $(X_1, \ldots, X_k)$ be discrete random variables with joint pmf

$$f(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

$$\text{where } x_1, \ldots, x_k \in \{0, 1, \ldots, n\} \text{ and } \sum_{i=1}^{k} x_i = n;$$

$$p_1, \ldots, p_k \in [0, 1] \text{ and } \sum_{i=1}^{k} p_i = 1.$$

Then $(X_1, \ldots, X_k)$ is said to follow a **Multinomial distribution**. We write

$$(X_1, \ldots, X_k) \sim \text{Multinomial}(n; p_1, \ldots, p_k).$$

2.54. Proposition: *If* $(X_1, \ldots, X_k) \sim \text{Multinomial}(n; p_1, \ldots, p_k)$, *then*

(1). $(X_1, \ldots, X_{k-1})$ *has joint moment generating function*

$$\begin{aligned} M(t_1, \ldots, t_{k-1}) &= \mathbb{E}\left(e^{t_1 X_1 + \cdots + t_{k-1} X_{k-1}}\right) \\ &= \left(p_1 e^{t_1} + \cdots + p_{k-1} e^{t_{k-1}} + p_k\right)^n, \quad (t_1, \ldots, t_{k-1}) \in \mathbb{R}^{k-1}. \end{aligned}$$

(2). Any subset of $\{X_1, \ldots, X_k\}$ *also has a Multinomial distribution. In particular, each* X_i *follows a binomial distribution with success probability* p_i, *i.e.,* $X_i \sim \text{Binomial}(n, p_i)$.

(3). If $T = X_i + X_j$ *with* $i \neq j$, *then* $T \sim \text{Binomial}(n, p_i + p_j)$

(4). For $i \neq j$, $\text{Cov}(X_i, X_j) = -np_ip_j$.

(5). The conditional distribution of any subset of $(X_1, \ldots, X_k)$ *given the remaining of the coordinates is a Multinomial distribution. In particular, the conditional probability function of* X_i *given* $X_j = x_j, i \neq j$, *is*

$$(X_i \mid X_j = x_j) \sim \text{Binomial}\left(n - x_j, \frac{p_i}{1 - p_j}\right).$$

(6). The conditional distribution of X_i *given* $T = X_i + X_j = t, i \neq j$, *is*

$$(X_i \mid X_i + X_j = t) \sim \text{Binomial}\left(t, \frac{p_i}{p_i + p_j}\right).$$

Section 11. Bivariate Normal Distribution

2.55. Definition: Suppose X_1 and X_2 are random variables with joint pdf

$$f(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^T\right\} \quad \text{for } (x_1, x_2) \in \mathbb{R}^2$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \quad |\Sigma| = \det(\Sigma),$$

and Σ is a non-singular positive definite matrix. Then $\mathbf{X} = (X_1, X_2)$ is said to have a **bivariate normal distribution** with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We write

$$\mathbf{X} \sim \text{BVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

2.56. Proposition: *If* $\mathbf{X} = [X_1, X_2]^T \sim \text{BVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *then*

(1). X_1, X_2 *has joint moment generating function*

$$\begin{aligned} M(t_1, t_2) &= \mathbb{E}\left(e^{t_1X_1+t_2X_2}\right) \\ &= \mathbb{E}\left[\exp\left(\mathbf{X}\mathbf{t}^T\right)\right] \\ &= \exp\left(\boldsymbol{\mu}\mathbf{t}^T + \frac{1}{2}\mathbf{t}\boldsymbol{\Sigma}\mathbf{t}^T\right) \qquad \textit{for all } \mathbf{t} = (t_1, t_2) \in \mathbb{R}^2 \end{aligned}$$

(2). $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$.

(3). $\text{Cov}(X_1, X_2) = \rho\sigma_1\sigma_2$, $\text{Cor}(X_1, X_2) = \rho$, *where* $-1 \le \rho \le 1$.

(4). $X_1 \perp X_2 \iff \rho = 0$.

(5). *For* $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^{2\times 1}$,

$$\mathbf{a}^T\mathbf{X} = a_1X_1 + a_2X_2 \sim N(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}).$$

(6). *For non-singular* $\mathbf{A} \in \mathbb{R}^{2\times 2}$ *and* $\mathbf{b} \in \mathbb{R}^{2\times 1}$,

$$\mathbf{A}\mathbf{X} + \mathbf{b} \sim \text{BVN}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T).$$

(7). *The conditional probability function of one given the other is*

$$(X_2 \mid X_1 = x_1) \sim N\left(\mu_2 + \rho(x_1 - \mu_1)\frac{\sigma_2}{\sigma_1}, \sigma_2^2(1-\rho^2)\right),$$
$$(X_1 \mid X_2 = x_2) \sim N\left(\mu_1 + \rho(x_2 - \mu_2)\frac{\sigma_1}{\sigma_2}, \sigma_1^2(1-\rho^2)\right).$$

(8). $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi_2^2$.

Chapter 3. Transformations of Random Variables

3.1. Motivation: Let $X_1, \ldots, X_n$ be random variables and suppose we want to find the distribution of a new random variable

$$Y := h(X_1, \ldots, X_n).$$

We discuss three approaches:

(1). the cdf technique;

(2). 1-1 transformation technique (for continuous random variables only);

(3). the mgf technique.

Section 1. The Cumulative Distribution Function Technique

3.2. Motivation: Given the pdf of $X_1, \ldots, X_n$, we wish to find the cdf and/or pdf/pmf of $Y = h(X_1, \ldots, X_n)$.

3.3. Note (CDF Technique, Discrete)**:** The discrete case is pretty simple. To find the probability of $Y = y$, we just need to figure out all cases of $X_1 = x_1, \ldots, X_n = x_n$, such that $h(x_1, \ldots, x_n) = y$, then add up the probabilities $\Pr(X_1 = x_1, \ldots, X_n = x_n) = f(x_1, \ldots, x_n)$ for each of these assignments to get the pdf. The cdf is obtained by summing up all values of $Y \leq y$ as usual. More specifically, for all $y \in \mathbb{R}$,

$$\begin{aligned} f_Y(y) = \Pr(Y = y) &= \Pr(h(X_1, \ldots, X_n) = y) \\ &= \Pr((X_1, \ldots, X_n) \in \{(x_1, \ldots, x_n) : h(x_1, \ldots, x_n) = y\}) \\ &= \sum_{(x_1,\ldots,x_n):h(x_1,\ldots,x_n)=y} \Pr(X_1 = x_1, \ldots, X_n = x_n) \\ &= \sum_{(x_1,\ldots,x_n):h(x_1,\ldots,x_n)=y} f(x_1, \ldots, x_n) \end{aligned}$$

$$F_Y(y) = \Pr(Y \leq y) = \sum_{t:t\leq y} f_Y(t).$$

3.4. Example: Consider

$$f_X(x) = \begin{cases} 1/4 & |x| = 1 \\ 1/2 & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

and define $Y = X^2$. Let us first write out the probability $\Pr(Y = y)$ in terms of X:

$$\Pr(Y = y) = \Pr(X^2 = y) = \begin{cases} \Pr(X = \sqrt{y}) + \Pr(X = -\sqrt{y}) & y > 0 \\ \Pr(X = 0) & y = 0 \\ 0 & y < 0 \end{cases}$$

- $y > 0$: $Y = X^2 = y \iff X = \pm\sqrt{y} \implies \Pr(Y = y) = \Pr(X = \sqrt{y}) + \Pr(X = -\sqrt{y})$.
- $y = 0$: $Y = X^2 = 0 \iff X = 0 \implies \Pr(Y = 0) = \Pr(X = 0) = 1/2$.
- $y < 0$: Since Y is a squared value, it can never be negative, so $\Pr(Y < 0) = 0$.

Now plug in the actual value, we obtain

$$\Pr(Y = y) = \begin{cases} 1/2 & y = 0 \\ 1/2 & y = 1 \\ 0 & \text{otherwise} \end{cases}$$

We conclude that $Y \sim \text{Bernoulli}(1/2)$.

3.5. Note (CDF Technique, Continuous)**:** The continuous case is a bit more complex and we need to find the cdf first. We break the procedure into three steps:

(1). For all $y \in \mathbb{R}$, find $R_y = \{(x_1, \ldots, x_n) : h(x_1, \ldots, x_n) \leq y\}$. This R can be viewed as all possible assignments of $X_1, \ldots, X_n$ that contribute densities to the cdf.

(2). Find the cdf of Y by integrating the joint pdf of $X_1, \ldots, X_n$ over R_y:

$$\begin{aligned} \forall y \in \mathbb{R} : F_Y(y) = \Pr(Y \leq y) &= \Pr(h(X_1, \ldots, X_n) \leq y) \\ &= \Pr((X_1, \ldots, X_n) \in R_y) \\ &= \int_{R_y} f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n. \end{aligned}$$

(3). Find the pdf of Y by differentiating $F_Y(y)$, i.e., $f_Y(y) = F_y'(y)$.

3.6. Example: Consider $X \sim \text{Exponential}(\theta)$ with cdf

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-x/\theta} & x > 0 \end{cases}$$

Define Y by applying F onto X:

$$Y = F(X) = \begin{cases} 0 & X \leq 0 \\ 1 - e^{-X/\theta} & X > 0. \end{cases}$$

Step 1. For all $y \in \mathbb{R}$, find $R_y = \{x : F(x) \leq y\}$.

- If $y < 0$, then $R_{y<0} = \varnothing$ because $F(x) \geq 0$ by definition.
- If $y = 0$, then we need $x \leq 0$, so $R_{y=0} = \{x : x \leq 0\}$.
- If $y \geq 1$, then any x satisfies $F(x) \leq 1 \leq y$, so $R_{y \geq 1} = \mathbb{R}$.
- Finally, we get $R_{0<y<1} = \{x : F(x) \leq y\} = \{x : x \leq 0\} \cup \{x > 0 : 1 - e^{-x/\theta} \leq y\}$. Solving the last inequality, we have $R_{0<y<1} = \{x : x \leq 0\} \cup \{x > 0 : x \leq -\theta \log(1-y)\}$.

Step 2. Find the cdf of Y.

$$F_Y(y) = \Pr(X \in R_y) = \begin{cases} \Pr(X \in \varnothing) = 0 & y < 0 \\ \Pr(X \leq 0) = \int_{x \in \mathbb{R}_{\leq 0}} 0\, dx = 0 & y = 0 \\ \Pr(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f(x)\, dx = 1 & y \geq 1 \\ \Pr(X \leq -\theta \log(1-y)) = F(-\theta \log(1-y)) = 1 - e^{\theta \log(1-y)/\theta} = y & 0 < y < 1 \end{cases}$$

Note in the last case, $\{x : x \leq 0\}$ contributes zero probability as in case 2. To summarize:

$$F_Y(y) = \begin{cases} 0 & y \leq 0 \\ y & 0 < y < 1 \\ 1 & y \geq 1 \end{cases}$$

By the uniqueness of cdfs, we conclude that $Y \sim \text{Uniform}(0, 1)$.

3.7. Remark: We have shown that when applying the cdf to a exponential random variable, the result is a Uniform$(0, 1)$ random variable. In fact, this holds for all continuous random variables, not just the ones following an exponential distribution.

3.8. Theorem: *If X is a continuous random variable with cdf F, then the random variable F defined by applying the cdf of X onto X,*

$$Y := F(X) = \int_{-\infty}^{X} f(t)\, dt,$$

has a Uniform$(0, 1)$ distribution.

Proof. Suppose the continuous random variable X has support set $A = \{x : f(x) > 0\}$. For all $x \in A$, F is an increasing by the definition of cdf. Therefore, F has an inverse on the domain A. For $0 < y < 1$, the cdf of $Y = F(X)$ is

$$\begin{aligned} G(y) &= \Pr(Y \leq y) \\ &= \Pr(F(X) \leq y) \\ &= \Pr(X \leq F^{-1}(y)) \\ &= F(F^{-1}(y)) \\ &= y, \end{aligned}$$

which is the cdf of a Uniform$(0, 1)$ random variable. It follows that $Y = F(X) \sim$ Uniform$(0, 1)$ as required. □

3.9. Remark: This provides a method for generating observations from a continuous distribution. Let u be an observation generated from a Uniform$(0, 1)$ distribution using a random number generator. Then by the corollary below, $x = F^{-1}(u)$ is an observation from the distribution with cdf F.

3.10. Corollary: *Let F be the cdf of some continuous random variable. If $U \sim$ Uniform$(0, 1)$, then the random variable $X = F^{-1}(U)$ has cdf F.*

Proof. Suppose that the support set of the random variable $X = F^{-1}(U)$ is A. For $x \in A$, the cdf of $X = F^{-1}(U)$ is

$$\begin{aligned} \Pr(X \leq x) &= \Pr(F^{-1}(U) \leq x) \\ &= \Pr(U \leq F(x)) \\ &= F(x) \end{aligned}$$

as $\Pr(U \leq u) = u$ for $0 < u < 1$ given that $U \sim$ Uniform$(0, 1)$. Therefore, $X = F^{-1}(U)$ has cdf F as claimed. □

3.11. Example: Let $X \sim N(0,1)$ and $Y = X^2$. First,

$$R_y = \{x : x^2 \leq y\} = \begin{cases} \varnothing & y < 0 \\ [-\sqrt{y}, \sqrt{y}] & y \geq 0 \end{cases}$$

where the second case follows from the fact that $x^2 \leq y \iff |x| \leq y$. Now

$$F_y(y) = \Pr(X \in R_y) = \begin{cases} 0 & y < 0 \\ \Pr(X \in [-\sqrt{y}, \sqrt{y}]) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) & y \geq 0 \end{cases}$$

For pdf,

$$\begin{aligned} f_Y(y) = F_y'(y) &= F_X'(\sqrt{y})\frac{1}{2\sqrt{y}} - F_X'(-\sqrt{y})\left(-\frac{1}{2\sqrt{y}}\right) \\ &= \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y})) \\ &= \frac{1}{2\sqrt{y}}\frac{2}{\sqrt{2\pi}}e^{-y/2} \\ &= \frac{1}{\sqrt{2\pi}}y^{-1/2}e^{-y/2}, y > 0. \end{aligned}$$

We conclude that $Y \sim \chi^2_1$.

3.12. Example: Let $X_1, X_2 \overset{iid}{\sim} \text{Uniform}(0,1)$. Then

$$f(x_1, x_2) = \begin{cases} 1 & (x_1, x_2) \in (0,1) \times (0,1) \\ 0 & \text{otherwise} \end{cases}$$

with joint support $A = (0,1) \times (0,1)$. Graphically speaking, if we view the Cartesian plane as $x_1 \times x_2$ and the "height" of the graph as $f(x_1, x_2)$, then the bars within the unit square $(0,1) \times (0,1)$ have height 1 and all other places are flat/of value zero.

Define $Y = X_1 + X_2$. Find cdf and pdf of Y.

Step 1. Consider $R_y = \{(x_1, x_2) : x_1 + x_2 \leq y\}$ for different y's. Note that

$$x_1 + x_2 \leq y \implies x_2 \leq -x_1 + y,$$

so by viewing the Cartesian plane as $x_1 \times x_2$, this line $x_2 = -x_1 + y$ has slope -1 and intercept y, and thus R_y is a halfspace (to the left/bottom of) determined by the line $x_2 = -x_1 + y$. Let us split y into cases. (The figure on the next page will be helpful.)

- Case 1. If $y < 0$, then $R_y \cap A = \varnothing$ as the unit square A is completely to the right/top of the line $x_1 + x_2 = y$. If $y = 0$, then A intersects R_y at exactly $(0,0)$. In both cases, $R_{y \leq 0}$ contributes 0 probability density.
- Case 2. If $y \geq 2$, then $R_y \cap A = A$ as the unit square A is completely contained in the halfspace R_y. In this case, $R_{y \geq 2}$ contributes 1 probability density.

- Case 3. If $0 < y \leq 1$, then the line $x_1 + x_2 = y$ intersects the "left" and the "bottom" sides of A. In this case, R_y contributes a triangle: $y^2/2$.
- Case 4. If $1 < y < 2$, then the line $x_1 + x_2 = y$ intersects the "top" and the "right" sides of A. In this case, R_y contributes a square subtract a triangle: $1 - (2-y)^2/2$.

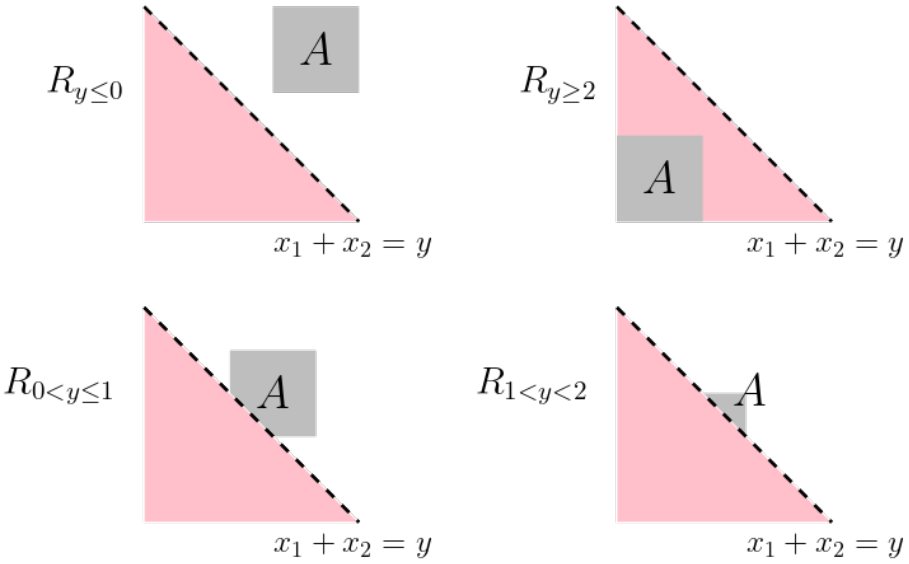

Figure 3.1: Pink: Halfplace. Gray: Support. Dashed line: $x_1 + x_2 = y$.

Time to evaluate the integral.

$$\begin{aligned} F_Y(y) &= \Pr((X_1, X_2) \in R_y) \\ &= \int_{R_y} f(x_1, x_2)\, dx_1 dx_2 \\ &= \int_{R_y \cap A} f(x_1, x_2)\, dx_1 dx_2. \end{aligned}$$

Therefore,

$$F_y(y) = \begin{cases} 0 & y \leq 0 \\ y^2/2 & 0 < y \leq 1 \\ 1 - (2-y)^2/2 & 1 \leq 2 < 2 \\ 1 & y \geq 2. \end{cases}$$

Finally,

$$f_Y(y) = \begin{cases} y & 0 < y \leq 1 \\ 2 - y & 1 < y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Lesson from this example: The first step is often the most complex one. Be patient.

3.13. Example: The next example is an extremely important one. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$. Find the distribution of $X_{(n)} = \max_i X_i$ and $X_{(1)} = \min_i X_i$.

For $X_{(n)}$,

$$\begin{aligned}
F_{X_{(n)}}(y) &= \Pr(X_{(n)} \leq y) \\
&= \Pr(\max_i X_i \leq y) \\
&= \Pr(X_1 \leq y, \ldots, X_n \leq y) \\
&= \prod_{i=1}^{n} \Pr(X_i \leq y) \\
&= [F(y)]^n \\
&= \begin{cases} 0 & y \leq 0 \\ (y/\theta)^n & 0 < y \leq \theta \\ 1 & y > \theta \end{cases} \\
f_{X_{(n)}}(y) &= \begin{cases} (n/\theta^n) y^{n-1} & 0 < y < \theta \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Now for $X_{(1)}$,

$$\begin{aligned}
F_{X_{(1)}}(y) &= \Pr(\min_i X_i \leq y) \\
&= 1 - \Pr(\min_i X_i > y) \\
&= 1 - \Pr(X_1 > y, \ldots, X_n > y) \\
&= 1 - \prod_{i=1}^{n} \Pr(X_i > y) \\
&= 1 - \prod_{i=1}^{n} (1 - \Pr(X_i \leq y)) \\
&= 1 - (1 - F(y))^n \\
&= \begin{cases} 0 & y \leq 0 \\ 1 - (1 - y/\theta)^n & 0 < y \leq \theta \\ 1 & y > \theta \end{cases} \\
f_{X_{(1)}}(y) &= \begin{cases} n/\theta (1 - y/\theta)^{n-1} & 0 < y < \theta \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Note this approach works for random variables with other distributions as well (i.e., not limited to the uniform distribution).

Section 2. Univariate 1-1 Transformation

3.14. Motivation: If the transformation h is *injective* (or *one-to-one*) on A, i.e.,

$$\forall x_1, x_2 \in A : h(x_1) = h(x_2) \implies x_1 = x_2,$$

then we can take a shortcut (which can be viewed as a special case of the cdf technique). This technique is useful when you only need the pdf of Y.

3.15. Theorem (Univariate 1-1 Transformation)**:** *If X is continuous with support A and the transformation h is 1-1 on A, then the pdf of $Y = h(X)$ is*

$$f_Y(y) = \begin{cases} f_X(h^{-1}(y)) \cdot \left| \dfrac{d}{dy} h^{-1}(y) \right| & y \in \{h(x) : x \in A\} \\ 0 & \text{otherwise} \end{cases}$$

where h^{-1} satisfies $h^{-1}(h(x)) = x$ for all $x \in A$.

Proof. Apply cdf technique. Proof omitted. □

3.16. Example: Let $\theta > 0$ and consider X with pdf

$$f(x) = \begin{cases} \theta / x^{\theta+1} & x > 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that $A_X = (1, \infty)$. Find the pdf of $Y = \log(X)$.

Proof. The log function is injective so we can use this technique. Let us find the inverse function and its derivative.

$$\begin{aligned} y = \log(x) \iff x = e^y &\implies h^{-1}(y) = e^y \\ &\implies \frac{d}{dy} h^{-1}(y) = e^y. \end{aligned}$$

Thus,

$$\begin{aligned} f_Y(y) &= \begin{cases} f_X(h^{-1}(y)) \left| \dfrac{d}{dy} h^{-1}(y) \right| & y \in \{h(x) : x \in A\} \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \theta e^{-\theta y} & y > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

as

$$f_X(e^y)|e^y| = \frac{\theta}{(e^y)^{\theta+1}} e^y = \theta e^{-\theta} y.$$

By uniqueness of pdf, we recognize that $Y \sim \text{Exponential}(\theta)$. □

Section 3. Bivariate 1-1 Transformation

3.17. Motivation: We now look at the joint distribution of a one-to-one (injective) transformation of two random variables. We begin with some notation and a theorem that gives sufficient conditions for determining whether a transformation is one-to-one in the bivariate case followed by the theorem, which gives the joint pdf for the two new random variables.

3.18. Note: Suppose the transformation S defined by

$$u = h_1(x, y)$$
$$v = h_2(x, y)$$

is a one-to-one transformation for all $(x, y) \in R_{XY}$ and that S maps the region R_{XY} into the region R_{UV} in the uv plane. Since $S : (x, y) \to (u, v)$ is a one-to-one transformation, there exists an inverse transformation T defined by

$$x = w_1(u, v)$$
$$y = w_2(u, v)$$

such that $T = S^{-1} : (u, v) \to (x, y)$ for all $(u, v) \in R_{UV}$. The Jacobian of the transformation T is

$$\frac{\partial(x, y)}{\partial(u, v)} = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix} = \left[\frac{\partial(u, v)}{\partial(x, y)}\right]^{-1},$$

where $\dfrac{\partial(u, v)}{\partial(x, y)}$ is the Jacobian of the transformation S.

3.19. We now given some sufficient but not necessary conditions for the inverse to exist.

3.20. Theorem (Inverse Mapping Theorem)**:** *Consider the transformation S defined by*

$$u = h_1(x, y)$$
$$v = h_2(x, y)$$

If $\dfrac{\partial u}{\partial x}, \dfrac{\partial u}{\partial y}, \dfrac{\partial v}{\partial x}, \dfrac{\partial v}{\partial y}$ are continuous functions and $\dfrac{\partial(u, v)}{\partial(x, y)} \neq 0$ for all $(x, y) \in R$, then S is one-to-one on R and S^{-1} exists.

3.21. We are now ready for the main result of this section.

3.22. Theorem (Bivariate Transformation Theorem)**:** *If $U = h_1(X,Y)$, $V = h_2(X,Y)$ defines an one-to-one transformation on the joint support $A = \{(x,y) : f(x,y) > 0\}$, then the joint pdf of U and V is given by*

$$g(u,v) = \begin{cases} f(w_1(u,v), w_2(u,v)) \cdot \left|\dfrac{\partial(w_1,w_2)}{\partial(u,v)}\right| & \forall (u,v) \in \{(h_1(x,y), h_2(x,y)) : (x,y) \in A\} \\ 0 & \text{otherwise} \end{cases}$$

where

$$\frac{\partial(w_1,w_2)}{\partial(u,v)} = \begin{bmatrix} \dfrac{\partial w_1}{\partial u} & \dfrac{\partial w_1}{\partial v} \\ \dfrac{\partial w_2}{\partial u} & \dfrac{\partial w_2}{\partial v} \end{bmatrix}$$

is the Jacobian matrix and $\|\cdot\|$ denotes its determinant

$$\left\|\frac{\partial(w_1,w_2)}{\partial(u,v)}\right\| = \frac{\partial w_1}{\partial u}\cdot\frac{\partial w_2}{\partial v} - \frac{\partial w_1}{\partial v}\cdot\frac{\partial w_2}{\partial v}.$$

3.23. Example: Consider

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \text{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right).$$

Define $U = X + Y$ and $V = X - Y$. The joint pdf of X and Y is given by

$$f(x,y) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}(x^2+y^2)\right).$$

with joint support $\mathbb{R}^2$. Let us first find the inverse mapping which shows that U, V is injective:

$$X = \frac{1}{2}(U+V), \quad Y = \frac{1}{2}(U-V)$$

Apply the one-to-one bivariate transformation theorem. Also, $A_{U,V} = A_{X,Y} = \mathbb{R}^2$. First,

$$\left|\frac{\partial(w_1,w_2)}{\partial(u,v)}\right| = \left|\begin{bmatrix} \dfrac{\partial w_1}{\partial u} & \dfrac{\partial w_1}{\partial v} \\ \dfrac{\partial w_2}{\partial u} & \dfrac{\partial w_2}{\partial v} \end{bmatrix}\right| = \left|\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}\right| = \frac{1}{2}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot\frac{1}{2} = -\frac{1}{2}.$$

Now the joint pdf of U, V is given by

$$\begin{aligned} g(u,v) &= f\left(\frac{1}{2}(u+v), \frac{1}{2}(u-v)\right) \cdot \left|\frac{\partial(w_1,w_2)}{\partial(u,v)}\right| \\ &= \frac{1}{2\pi}\exp\left\{-\frac{1}{2}\left(\left(\frac{u+v}{2}\right)^2 + \left(\frac{u-v}{2}\right)^2\right)\right\} \cdot \left|-\frac{1}{2}\right| = \frac{1}{4\pi}\exp\left\{-\frac{1}{4}(u^2+v^2)\right\}. \end{aligned}$$

Thus, $\begin{bmatrix} U \\ V \end{bmatrix} \sim \text{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\right).$

3.24. Example: Let X, Y be independent exponential variables with $\theta = 1$. Then

$$f(x,y) = \begin{cases} e^{-x}e^{-y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Define $U = X + Y$ and $V = X$. We wish to show that $X + Y \sim \text{Gamma}(2, 1)$.

Find the inverse mapping first:

$$\begin{aligned} X &= V \\ Y &= U - V \end{aligned}$$

Figure out the joint support of U and V:

$$\begin{aligned} \{(x+y, x) : x > 0, y > 0\} &= \{(u,v) : 0 < v < \infty, v < u < \infty\} \\ &= \{(u,v) : 0 < v < u < \infty\}. \end{aligned}$$

The determinant of the derivative:

$$\left|\frac{\partial(w_1, w_2)}{\partial(u,v)}\right| = \begin{vmatrix} \frac{\partial w_1}{\partial u} & \frac{\partial w_1}{\partial v} \\ \frac{\partial w_2}{\partial u} & \frac{\partial w_2}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1.$$

The joint pdf:

$$f(w_1(u,v), w_2(u,v)) = e^{-v}e^{-(u-v)} = e^{-u} \implies g(u,v) = \begin{cases} e^{-u} & 0 < v < u < \infty \\ 0 & \text{otherwise} \end{cases}$$

Finally,

$$g_U(u) = \begin{cases} \int_0^u e^{-u}\, dv = ue^{-u} & 0 < u < \infty \\ 0 & \text{otherwise} \end{cases}$$

This is the pdf of $U = X + Y$ which shows that $U = X + Y \sim \text{Gamma}(2, 1)$.

3.25. Remark: This concludes the sections on transformation theorems.

Section 4. The Moment Generating Function Technique

3.26. Motivation: The mgf technique is particularly useful in determining the distribution of a sum of two or more independent random variables if the mgfs of the random variables exist.

3.27. Theorem: *Suppose $X_1, \ldots, X_n$ are independent random variables and X_i has mgf $M_i(t)$ which exists for $t \in (-h, h)$ for some $h > 0$. Then $Y = X_1 + \cdots + X_n$ has mgf*

$$M_Y(t) = \prod_{i=1}^{n} M_i(t), \quad t \in (-h, h).$$

If the X_i's are iid each with mgf $M(t)$, then $Y = X_1 + \cdots + X_n$ has mgf

$$M_Y(t) = [M(t)]^n, \quad t \in (-h, h).$$

3.28. Example: We prove a useful result on transformations of normal random variables. Let $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ with $a \neq 0$. We know that

$$M_X(t) = \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}, \quad t \in \mathbb{R}.$$

Then

$$M_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}[e^{t(aX+b)}] = e^{tb}\mathbb{E}[e^{atX}] = e^{tb}M_X(at) = \exp\left\{t(a\mu + b) + \frac{(a\sigma)^2 t^2}{2}\right\}$$

By uniqueness of mgf, we see that $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$.

3.29. Example: Let $X_i \overset{\text{iid}}{\sim} N(\mu_i, \sigma_i^2)$ and $Y = \sum_{i=1}^n a_i X_i$. Then

$$M_{X_i}(t) = \exp\left\{\mu_i t + \frac{\sigma_i^2 t^2}{2}\right\}.$$

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{n} M_{x_i}(a_i t) \\ &= \prod_{i=1}^{n} \exp\left\{\mu_i a_i t + \frac{\sigma_i^2 a_i^2 t^2}{2}\right\} \\ &= \exp\left\{t\sum_{i=1}^{n} \mu_i a_i + t^2 \sum_{i=1}^{n} \sigma_i^2 a_i^2\right\} \end{aligned}$$

By uniqueness of mgf, we see that

$$Y = \sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} \mu_i a_i, \sum_{i=1}^{n} \sigma_i^2 a_i^2\right).$$

Section 5. Important Distributions

3.30. Note: The **Chi-Squared** distribution with k degrees of freedom:

$$\sum_{i=1}^{k} Z_i^2 \sim \chi_k^2, \qquad Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} N(0,1).$$

Some properties of χ_k^2:

- mgf of χ_1^2: $M(t) = (1-2t)^{-1/2}$.
- mgf of χ_k^2:

$$M_{\sum_{i=1}^k Z_i^2}(t) = \mathbb{E}[e^{t\sum_{i=1}^k Z_i^2}] = \prod_{i=1}^{k} \mathbb{E}[e^{tZ_i^2}] = \prod_{i=1}^{k}(1-2t)^{-1/2} = (1-2t)^{-k/2}.$$

- If $Y_1, \ldots, Y_m$ are independent with $Y_i \sim \chi_{ki}^2$, then

$$\sum_{i=1}^{m} Y_i \sim \chi^2_{\sum_{i=1}^k ki}.$$

as its mgf is given by

$$M_{\sum_{i=1}^m Y_i}(t) = \prod_{i=1}^{m} M_{Y_i}(t) = \prod_{i=1}^{m}(1-2t)^{-ki/2} = (1-2t)^{-\frac{1}{2}\sum_{i=1}^m k_i}.$$

3.31. Note (t-Distribution)**:** Let $X \sim N(0,1)$ and $Y \sim \chi_n^2$ be independent. Then

$$\frac{X}{\sqrt{Y/n}} \sim t_{(n)}$$

with support $\mathbb{R}$.

3.32. Note (F-Distribution)**:** Let $X \sim \chi_{(n)}^2$ and $Y \sim \chi_{(m)}^2$ be independent, then

$$\frac{X/n}{Y/n} \sim F_{(n,m)}$$

with support $(0, \infty)$.

3.33. Example: Let $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Define the mean and sample variance

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

We look at the one-sample T-statistic for $H_0 : \mu = \mu_0$:

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}.$$

Write $Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ so that $T = \dfrac{Z}{s/\sigma} = \dfrac{Z}{\sqrt{s^2/\sigma^2}} = \dfrac{Z}{\sqrt{\frac{(n-1)s^2}{(n-1)\sigma^2}}} =: \dfrac{Z}{\sqrt{Y/(n-1)}}$ where $Y = \dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$. We need to show three things:

(1) $Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. (2) $Y = \dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$. (3) Z and Y are independent.

Proof of (1). We have

$$\overline{X} \sim N\left(\sum_{i=1}^n a_i\mu, \sum_{i=1}^n a_i^2\sigma^2\right) = N\left(\mu, \frac{\sigma^2}{n}\right).$$

Define $Z = a\overline{X} + b, a = \dfrac{\sqrt{n}}{\sigma}, b = \dfrac{-\mu\sqrt{n}}{\sigma}$, we get $Z \sim N\left(a\mu + b, a^2\dfrac{\sigma^2}{n}\right) = N(0,1)$. □

Proof of (2). We first show that $\overline{X}$ is independent with each of $X_1 - \overline{X}, \ldots, X_n - \overline{X}$. For this claim, recall that

$$X_i \overset{\text{iid}}{\sim} N(\mu,\sigma^2) \iff \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}\right).$$

Now we can take a linear transformation of $(X_1, \ldots, X_n)$ to obtain

$$\begin{bmatrix} \overline{X} \\ X_1 - \overline{X} \\ \vdots \\ X_n - \overline{X} \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ 1-\frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & 1-\frac{1}{n} \end{bmatrix}}_{A} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \sim \text{MVN}\left(A\begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}, A\begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}A^T\right).$$

Skipping some steps, the covariate matrix evaluates to

$$\left[\begin{array}{c|c} \sigma^2/n & O \\ \hline O & \ddots \end{array}\right]$$

Since the first row and columns besides the $(0,0)$-th entry is zero, the claim follows. Now

$$\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2} + \frac{n(\overline{X}-\mu)^2}{\sigma^2}.$$

Now LHS $\sim \chi^2_n$ as it is a sum of n iid standard normals. The middle term can be written as $\frac{(n-1)s^2}{\sigma^2}$ and the right term can be written as $\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1$. Computing its mgf and simplify, we get the desired result. To verify your answer, note that $M_Y(t) = (1-2t)^{-(n-1)/2}$. □

Proof of (3). Y and Z are independent as Z is a function of the sample mean $\overline{X}$ and Y is a function of the sample variance. □

3.34. Corollary: *Given two independent samples*

$$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma_1^2),$$
$$Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu, \sigma_2^2),$$

then

$$T = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma_1^2}{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2/\sigma_2^2} \sim F_{n-1,m-1}.$$

Proof. We can write

$$T = \frac{\left[\frac{(n-1)s_1^2}{\sigma^2}\right]/(n-1)}{\left[\frac{(m-1)s_2^2}{\sigma^2}\right]/(m-1)}.$$

Note that

$$\frac{(n-1)s_1^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{(m-1)s_2^1}{\sigma^2} \sim \chi^2_{m-1},$$

and the numerator of T is a function of X_i's while the denominator of T is a function of the Y_i's. Now use previous results and we are done. □

Chapter 4. Limiting/Asymptotic Distributions

Section 1. Convergence in Distribution

4.1. Definition: Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables and $\{F_i(x)\}_{i=1}^{\infty}$ be the corresponding cdfs, i.e., X_i has cdf $F_i(x) = \Pr(X_i \leq x)$. Let X be the random variable with cdf $F(x) = \Pr(X \leq x)$. We say X_n **converges in distribution** to X and write $X_n \to_D X$ if $\lim_{n\to\infty} F_n(x) = F(x)$ at all points x at which $F(x)$ is continuous. We call F the **limiting** or **asymptotic distribution** of X_n.

4.2. Remark: There are a few points to note in this definition.

(1). Intuitively, $X_n \to_D X$ means that for large n, the distributions of X_n becomes close to X. Thus, all functions of X_n that uniquely determines the distribution of a random variable, such as cdf, pdf, mgf, etc., will approach that of X. Taking cdf as an example, $X_n \to_D X$ implies that for large n, $F_n(x) = \Pr(X_n \leq x) \approx \Pr(X \leq n) = F(x)$.

(2). However, convergence in distribution does not mean that the value of X_n will converge to the value of X. See Example 4.3.

(3). Note in the definition above, we only care about points where $F(x)$ is continuous. See Example 4.4.

4.3. Example: Let $W \sim \text{Bernoulli}(1/2)$, $X = 1 - W$, and $X_n = W$ for all $n = 1, 2, \ldots$ Here, $X_n \to_D X$ (indeed, both are Bernoulli(1/2) random variables) but do we have $X_n \approx X$? No! We always $X_n = W$ and $X = 1 - W$. In particular, $|X_n - X| = 1$ for all n.

4.4. Example: Let $X_n \sim \text{Uniform}(0, 1/n)$ and $X = 0$ (a constant random variable). Intuitively, the support of X_n shrinks as $n \to \infty$ and eventually approaches X. Let us prove this formally.

$$F_n(x) = \begin{cases} 0 & x \leq 0 \\ nx & 0 < x < 1/n \\ 1 & x \geq 1/n \end{cases} \implies \lim_{n\to\infty} F_n(x) = \tilde{F}(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$$

Comparing this with the cdf of X, which is

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

we see that they do not match at $x = 0$. In fact, $\tilde{F}(x)$ is not right continuous, so it is not even a cdf! Luckily, in the definition of convergence in distribution, we only care about points in $\mathbb{R} \setminus \{0\}$ as 0 is a point of discontinuity of $F(x)$. Thus, we still have $\lim_{n\to\infty} F_n(x) = F(x)$ for all x such that F is continuous at x, so $X_n \to_D X$ (which matches our intuition).

4.5. The following theorem is very useful in determining the limiting distribution of random variables.

4.6. Theorem: *Let $b, c \in \mathbb{R}$ and $\phi : \mathbb{N} \to \mathbb{R}$ such that $\lim_{n\to\infty} \phi(n) = 0$. Then*

$$\lim_{n\to\infty} \left[1 + \frac{b}{n} + \frac{\phi(n)}{n}\right]^{cn} = e^{bc}.$$

In particular, taking ϕ to be the constant zero function, we have

$$\lim_{n\to\infty} \left(1 + \frac{b}{n}\right)^{cn} = e^{bc}.$$

Proof. Omitted. □

4.7. Example: Let $X_i \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$, $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Find the asymptotic distribution of $nX_{(1)}$.

Proof. Recall the following from a previous example:

$$F_{X_{(1)}}(y) = \begin{cases} 0 & y \leq 0 \\ 1 - (1-y)^n & 0 < y \leq 1 \\ 1 & y > 1 \end{cases}$$

Now observe that

$$\begin{aligned} F_n(x) &= \Pr(nX_{(1)} \leq x) \\ &= \Pr(X_{(1)} \leq x/n) \\ &= \begin{cases} 0 & x/n \leq 0 \\ 1 - (1 - x/n)^n & 0 < x/n \leq 1 \\ 1 & x/n > 1 \end{cases} \\ &= \begin{cases} 0 & x \leq 0 \\ 1 - (1 - x/n)^n & 0 < x \leq n \\ 1 & x > n \end{cases} \end{aligned}$$

$$\lim_{n\to\infty} F_n(x) = \tilde{F}(x) = \begin{cases} 0 & x \leq n \\ 1 - e^{-x} & x > 0 \end{cases}$$

Now recall that the cdf of $X \sim \text{Exponential}(1)$ is given by

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

Since $\tilde{F}(x)$ and $F(x)$ match at all points besides 0, we see that the asymptotic distribution of $nX_{(1)}$ is Exponential(1). □

4.8. Example: Let $X_i \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$, $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Find the asymptotic distribution of $n(1 - X_{(n)})$.

Proof. Recall the following from a previous example:

$$F_{X_{(n)}}(y) = \begin{cases} 0 & y \leq 0 \\ y^n & 0 < y \leq 1 \\ 1 & y > 1 \end{cases}$$

As before, we want to find the cdf of $F_n(x)$.

$$\begin{aligned} F_n(x) &= \Pr(n(1 - X_{(n)}) \leq x) \\ &= \Pr(X_{(n)} \geq 1 - x/n) \\ &= 1 - \Pr(X_{(n)} \leq 1 - x/n) \\ &= 1 - F_{X_{(n)}}(1 - x/n) \\ &= \begin{cases} 1 & 1 - x/n \leq 0 \\ 1 - (1 - x/n)^n & 0 < 1 - x/n \leq 1 \\ 0 & 1 - x/n > 1 \end{cases} \\ &= \begin{cases} 0 & x < 0 \\ 1 - (1 - x/n)^n & 0 < x \leq n \\ 1 & x > n \end{cases} \\ \lim_{n\to\infty} F_n(x) = \tilde{F}(x) &= \begin{cases} 0 & x \leq 0 \\ 1 - e^{-x} & x > 0 \end{cases} \end{aligned}$$

It follows that $n(1 - X_{(n)}) \to_D \text{Exponential}(1)$ as well. □

Section 2. Convergence in Probability

4.9. Definition: A sequence of random variables $\{X_i\}_{i=1}^{\infty}$ **converges in probability** to a random variable X, denoted $X_n \to_p X$ if, for all $\varepsilon > 0$, $\lim_{n\to\infty} \Pr(|X_n - X| \geq \varepsilon) = 0$, or equivalently, $\lim_{n\to\infty} \Pr(|X_n - X| < \varepsilon) = 1$.

4.10. Remark: Convergence in probability is a stronger form of convergence than convergence in distribution, in the sense that $X_n \to_p X \Rightarrow X_n \to_D X$. The converse is false.

4.11. Example: Let $W \sim \text{Uniform}(0, 1)$, $X = 0$ be a constant random variable, and

$$X_n = \begin{cases} 1 & 0 < W < 1/n \\ 0 & \text{otherwise} \end{cases}$$

We proceed by definition. Let $\varepsilon > 0$, observe that

$$\begin{aligned} \Pr(|X_n - X| \geq \varepsilon) = \Pr(X_n \geq \varepsilon) = \Pr(X_n = 1) &= \Pr(0 < W < 1/n) \\ &= \int_0^{1/n} 1\, dx = \frac{1}{n} \xrightarrow{n\to\infty} 0. \end{aligned}$$

4.12. Here are two useful inequalities.

4.13. Theorem (Markov Inequality)**:** *Let X be a random variable.*

$$\Pr(|X| \geq c) \leq \frac{\mathbb{E}(|X|^k)}{c^k}, \qquad \forall k, c \geq 0.$$

4.14. Theorem (Chebyshev's Inequality)**:** *Suppose X is a random variable finite mean μ and finite variance σ^2. Then for any $k > 0$,*

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

4.15. Example: Let $\{X_i\}_{i=0}^{\infty}$ be a series of random variables with

$$\mathbb{E}[X_n] = \mu, \quad \text{Var}[X_n] = \sigma_n^2, \quad \lim_{n\to\infty} \sigma_n^2 = 0.$$

Show that $X_n \to_p \mu$.

Proof. Let $\varepsilon > 0$. Then,

$$0 \leq \Pr(|X_n - \mu| \geq \varepsilon) \leq \frac{\mathbb{E}[(X_n - \mu)^2]}{\varepsilon^2} = \frac{\text{Var}[X_n]}{\varepsilon^2} = \frac{\sigma_n^2}{\varepsilon^2} \xrightarrow{n\to\infty} 0.$$

By Squeeze Theorem,

$$\lim_{n\to\infty} \Pr(|X_n - \mu| \geq \varepsilon) = 0 \implies X_n \to_p \mu.$$

□

Section 3. Weak Law of Large Numbers

4.16. Intuition: The WLLN says that the sample mean $\overline{X}_n$ of iid random variables approaches the population mean μ as $n \to \infty$.

4.17. Theorem: *Suppose $\{X_i\}_{i=0}^{\infty}$ are iid random variables with $\mathbb{E}[X_i] = \mu$ and finite variance $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Consider the sequence of means $\{\overline{X}_i\}_{i=0}^{\infty}$ where*

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

Then $\overline{X}_n \to_p \mu$.

Proof. Fix $\varepsilon > 0$. Using Chebyshev's inequality to the random variable $\overline{X}_n$ with mean μ and variance σ^2/n, we obtain

$$\Pr\left(|\overline{X}_n - \mu| \geq \frac{k\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2}$$

for all $k > 0$. Set

$$k = \frac{\sqrt{n}\varepsilon}{\sigma}.$$

Then

$$0 \leq \Pr(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\sigma} \xrightarrow{n\to\infty} 0.$$

By Squeeze Theorem,

$$\lim_{n\to\infty} \Pr(|\overline{X}_n - \mu| \geq \varepsilon) = 0$$

as required. □

4.18. Remark: The proof of the WLLN does not actually require that the random variables are iid, only that they all have the same mean and variance. It also does not require knowing the distribution of these random variables.

4.19. Remark: A stronger result exists, but we won't talk about it in this class.

Section 4. Central Limit Theorem

4.20. Lemma: *Suppose* $\{X_i\}_{i=0}^{\infty}$ *are iid rvs with mgfs* $M_i(t)$. *Let* X *be a rv with mgf* $M(t)$. *If there exists an* $h > 0$ *such that*

$$\forall t \in (-h, h) : \lim_{n\to\infty} M_n(t) = M(t),$$

then $X_n \to_D X$.

Proof. Recall that the mgf uniquely determines the distribution of a random variable. □

4.21. Lemma: *Suppose* $\{X_i\}_{i=0}^{\infty}$ *and* X *are non-negative integer-valued random variables. If* $X_n \to_D X$, *then*

$$\lim_{n\to\infty} \Pr(X_n \le x) = \Pr(X \le x)$$

holds for all x *and in particular*

$$\lim_{n\to\infty} \Pr(X_n = x) = \Pr(X = x) \quad x = 0, 1, \ldots$$

Proof. This is like the definition of convergence in distribution but applied on non-negative integer-valued random variables. □

4.22. Theorem: *Suppose* $\{X_i\}_{i=0}^{\infty}$ *are iid rvs with* $\mathbb{E}[X_i] = \mu$ *and* $\mathrm{Var}[X_i] = \sigma^2 < \infty$. *Consider the sequence of normalized random variables* $\{Z_i\}_{i=0}^{\infty}$ *with*

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}, \qquad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then $Z_n \to_D Z \sim N(0, 1)$.

Proof. Omitted. □

4.23. Example: We show an application of CLT. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Then

$$\frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\lambda}} \to_D N(0, 1).$$

4.24. Example: Suppose we have a sequence of random variables where $X_i \sim \chi_i^2$. Since $\chi_n^2 = \sum_{i=1}^{n} \chi_1^2$, we can apply CLT. We know that $\mathbb{E}[X_n^2] = 1$ and $\mathrm{Var}[X_1^2] = 2$. Therefore,

$$\frac{\sqrt{n}\left(\frac{\sum_{i=1}^{n}\chi_i^2}{n} - 1\right)}{\sqrt{2}} \to_D N(0, 1).$$

It follows that

$$\frac{X_n - n}{\sqrt{2n}} \to_D N(0, 1).$$

Section 5. More Limit Theorems

4.25. The following two results tell us that convergence in distribution/probability behave quite nicely. In particular, we can substitute the limiting distribution when applying a continuous function and add/subtract/multiply/divide limiting distributions, as long as one sequence converges to a constant random variable. See Warning 4.28.

4.26. Theorem (Continuous Mapping)**:** *Let $g(\cdot)$ be a continuous function.*

- *If $X_n \to_p c$, then $g(X_n) \to_p g(c)$.*
- *If $X_n \to_d X$, then $g(X_n) \to_d g(X)$.*

Proof. Omitted. □

4.27. Theorem (Slutsky)**:** *If $X_n \to_d X$ and $Y_n \to_p c$, then:*

- $X_n + Y_n \to_d X + c$.
- $X_n Y_n \to_d cX$.
- $\frac{X_n}{Y_n} \to_d \frac{X}{c}$ *(when $c \neq 0$).*

Note that we can replace $\to_d$ with $\to_p$ and these results still hold.

Proof. Omitted. □

4.28. Warning: Note that $X_n \to_D X$ and $Y_n \to_D Y$ does not imply $X_n + Y_n \to_D X + Y$!

4.29. Example: Let $X_n \geq 0$ and $c \geq 0$. Then

$$X_n \to_p c \implies \sqrt{X_n} \to_p \sqrt{c}$$
$$X_n \to_p c \implies X_n^2 \to c^2$$

4.30. Example: Let $X_n \to_D X \sim N(0,1)$. Then

$$2X_n + 1 \to_D 2X + 1 \sim N(1,4)$$
$$X_n^2 \to_D X^2 \sim \chi_1^2$$

4.31. Example: If $X_n \to_D X \sim N(0,1)$ and $Y_n \to_p c \neq 0$, then

$$X_n + Y_n \to_D X + c \sim N(c,1)$$
$$X_n Y_n \to_D cX \sim N(0, c^2)$$
$$\frac{X_n}{Y_n} \to_D \frac{X}{c} \sim N\left(0, \frac{1}{c^2}\right)$$

4.32. Example: Let $\{X_i\}_{i=1}^{\infty}$ be iid Poisson(λ) random variables. We wish to find the asymptotic distribution for $U_n = \sqrt{n}(\overline{X}_n - \lambda)$ and $Z_n = U_n/\sqrt{\overline{X}_n}$.

Solution. First, by CLT, we have

$$\frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\lambda}} \to_D N(0,1).$$

Define $g(t) = \sqrt{\lambda}t$ which is continuous. Then by continuous mapping theorem,

$$U_n = g\left(\frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\lambda}}\right) \to_D g(N(0,1)) = \sqrt{\lambda} \cdot N(0,1) = N(0,\lambda).$$

Next, by WLLN, $\overline{X}_n \to_p \lambda$. Define $h(t) = \sqrt{t}$ which is continuous. Then

$$\sqrt{\overline{X}_n} = g(\overline{X}_n) \to_p g(\lambda) = \sqrt{\lambda}.$$

Finally, by Slutsky, we have

$$Z_n = \frac{U_n}{\sqrt{\overline{X}_n}} \to_D \frac{N(0,\lambda)}{\sqrt{\lambda}} = N(0,1).$$

□

4.33. Example: Let $\{X_i\}_{i=1}^{\infty}$ be iid Uniform$(0,1)$ random variables. Define $U_n = \max_{1\le i\le n} X_i$ and $V_n = e^{-n(1-U_n)}$. Find the limiting distribution of V_n.

Solution. Previously, we have shown that $n(1-U_n) \to_D$ Exponential(1). Define $g(t) = e^{-t}$, so that $V_n = g(n(1-U_n))$. Using continuous mapping theorem,

$$V_n \to_D g(\text{Exponential}(1)) = e^{-Y}$$

where $Y \sim$ Exponential(1). Let $T = e^{-Y}$. Using definition of cdf,

$$F_T(t) = \Pr(e^{-Y} \le t) = \begin{cases} \Pr(Y \ge -\log(t)) & t > 0 \\ 0 & t \le 0 \end{cases}$$

For $t > 0$,

$$\Pr(Y \ge -\log(t)) = \begin{cases} \int_{-\log t}^{\infty} e^{-y}\,dy & -\log(t) > 0 \\ 1 & -\log(t) \le 0 \end{cases} = \begin{cases} t & t < 1 \\ 1 & t \ge 1 \end{cases}$$

It follows that

$$F_T(t) = \begin{cases} 0 & t \le 0 \\ t & 0 < t < 1 \\ 1 & t \ge 1 \end{cases}$$

Observe this is a Uniform$(0,1)$ random variable. It follows that $V_n \to_D$ Uniform$(0,1)$. □

4.34. Example: Let $\{X_i\}_{i=1}^{\infty}$ be iid Uniform$(0,1)$ random variables. Define $U_n = \max_{1\le i\le n} X_i$. Find the limiting distribution of

$$W_n = \frac{n(1-U_n)}{\overline{X}_n^2}.$$

Solution. We know the numerator converges in distribution to Exponential(1). For the denominator, we use WLLN, which gives us

$$\overline{X}_n \to_p \mathbb{E}[X_i] = \frac{1}{2}.$$

Moreover, $\text{Var}[X_i] = 1/12 < \infty$. Using continuous mapping, for $g(t) = t^2$, we have

$$\overline{X}_n^2 = g(\overline{X}_n) \to_p \frac{1}{2^2} = \frac{1}{4} \neq 0.$$

Applying Slutsky, we have

$$W_n \to_D \frac{\text{Exponential}(1)}{1/4} = 4 \cdot \text{Exponential}(1) = \text{Exponential}(4).$$

□

4.35. Theorem (Delta Method)**:** *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables such that*

$$a_n(X_n - \theta) \to_D N(0, \sigma^2)$$

with $\lim_{n\to\infty} a_n = \infty$ and $g(x)$ is differentiable at $x = \theta$ with $g'(\theta) \neq 0$. Then

$$a_n[g(X_n) - g(\theta)] \to_D N(0, [g'(\theta)]^2\sigma^2).$$

4.36. Intuition: Using 1st order Taylor expansion,

$$g(X_n) \approx g(\theta) + g'(\theta)(X_n - \theta)$$
$$a_n(g(X_n) - g(\theta)) \approx a_n(g'(\theta)(X_n - \theta))$$

For large n, $a_n(X_n - \theta) \approx_D N(0, \sigma^2)$, which implies that

$$a_n g'(\theta)(X_n - \theta) \approx_D N(0, \sigma^2) - g'(\theta) = N(0, \sigma^2[g'(\theta)]^2).$$

4.37. Note: We usually use this result with $a_n = \sqrt{n}$. Recall CLT states that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \to_D N(0,1).$$

Now by continuous mapping theorem,

$$\sqrt{n}(\overline{X}_n - \mu) \to_D N(0,1) \cdot \sigma = N(0, \sigma^2).$$

4.38. Example: Let $\{X_i\}_{i=1}^{\infty}$ be iid Poisson(λ) random variables. Find the limiting distribution of

$$Z_n = \sqrt{n}(\sqrt{\overline{X}_n} - \sqrt{\lambda}).$$

Solution. Previously, we have $\sqrt{n}(\overline{X}_n - \lambda) \to_D N(0, \lambda)$. Define $g(t) = \sqrt{t}$, so that $Z_n = \sqrt{n}(g(\overline{X}_n) - g(\lambda))$. Since $\lambda > 0$ (required for Poisson), $g'(\lambda)$ exists and

$$g'(\lambda) = \frac{1}{2\sqrt{\lambda}} \neq 0.$$

By the Delta method,

$$Z_n \to_D N(0, [g'(\lambda)]^2 \cdot \lambda) \implies Z_n \to_D N(0, 1/4).$$

□

4.39. Example: Let $\{X_i\}_{i=1}^{\infty}$ be iid Exponential(θ) random variables. Find the limiting distribution of

$$Z_n = \sqrt{n}(\log(\overline{X}_n) - \log(\theta)).$$

Proof. By CLT and continuous mapping,

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{\theta} \to_D N(0, 1) \implies \sqrt{n}(\overline{X}_n - \theta) \to N(0, \theta^2).$$

Define $g(t) = \log(t)$ so that

$$Z_n = \sqrt{n}(g(\overline{X}_n) - g(\theta)).$$

We have $g'(\theta) = 1/\theta \neq 0$, so by delta method,

$$Z_n \to_D N(0, [g'(\theta)]^2 \cdot \theta^2) = N(0, 1).$$

□

4.40. Example: Let $\{X_i\}_{i=1}^{\infty}$ be iid with mean 0, variance $\sigma^2 < \infty$. Approximate the distribution of $\overline{X}_n^2$.

Solution. By CLT and continuous mapping,

$$\frac{\sqrt{n}(\overline{X}_n - 0)}{\sigma} \to_D N(0, 1) \implies \sqrt{n}\overline{X}_n \to_D N(0, \sigma^2).$$

Define $g(t) = t^2$, which exists at $t = 0$. However, $g'(0) = 0$, so we can't use the delta method. Another idea: apply the squared functionto

$$\frac{\sqrt{n}(\overline{X}_n - 0)}{\sigma} \to_D N(0, 1) \implies \frac{n\overline{X}_n^2}{\sigma^2} \to_D [N(0, 1)]^2 = \chi_1^2.$$

□

Section 6. Summary

4.41. Let us now summarize this chapter. We first list the building blocks for determining convergence in distribution/probability, then discuss the helper results.

4.42. Note: We have two ways of showing convergence in distribution:

(1). By definition, i.e., show that $F_n(x) \to F(x)$. Most likely simple limit or e limit.

(2). By CLT, when we are working with sums of iid random variables.

We have two ways of showing convergence in probability:

- By definition, i.e., $\forall \varepsilon > 0 : \Pr(|X_n - X| \geq \varepsilon) \to 0$. Simple limit, or Markov inequality.
- By WLLN, when there are sum of iid random variables.

4.43. Note: The helper results:

- Continuous mapping theorem.
- Slutsky's theorem.
- Delta method.

Chapter 5. Point Estimation

5.1. Motivation: Suppose the random variable $\mathbf{X}$ has pdf $f(\mathbf{x};\theta)$, where θ is unknown and $\theta \in \Omega$ is the parameter space. A **statistic**, $T = T(\mathbf{X})$, is a function of the data $\mathbf{X}$ which does not depend on any unknown parameters. A statistic $T = T(\mathbf{X})$ that is used to estimate $\tau(\theta)$, a function of θ, is called an **estimator** of $\tau(\theta)$ and an observed value of the statistic $t = t(\mathbf{x})$ is called an **estimate** of $\tau(\theta)$. We often denote an estimator of θ as $\hat{\theta}$.

Section 1. Method of Moments

5.2. Note: Let $X_1, \ldots, X_n$ be iid with pdf $f(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^p$. Define

$$\begin{aligned} \mu_1 &= \mathbb{E}[X_i] = g_1(\boldsymbol{\theta}) \\ \mu_2 &= \mathbb{E}[X_i^2] = g_2(\boldsymbol{\theta}) \\ &\vdots \\ \mu_p &= \mathbb{E}[X_i^p] = g_p(\boldsymbol{\theta}) \end{aligned}$$

Since we don't have true population moments in practice, we can substitute μ_i by $\hat{\mu}_i$, where

$$\hat{\mu}_j = \frac{\sum_{i=1}^n X_i^j}{n}.$$

We define the **method of moments estimates** (MME) of θ to be the solution to this system of p equations and p unknowns.

5.3. Example: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. We know the true population mean is $\mu_1 = \mathbb{E}[X_i] = \lambda$. Using MME, we derive one equation of one known:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n X_i = \lambda.$$

Since $\hat{\lambda} = \hat{\mu}_1$ solves this equation, it is the MME of λ. We look at the quality of this estimate:

- $\hat{\lambda}$ is unbiased:

$$\mathbb{E}[\hat{\lambda}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n}\cdot n\lambda = \lambda.$$

- $\hat{\lambda}$ is consistent: By WLLN, $\hat{\lambda} = \overline{X_n} \to_p \lambda$.

5.4. Example: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$. We know that $\mu_1 = \mathbb{E}[X_i] = \theta/2$. The MME estimate $\hat{\theta}$ is given by

$$\hat{\mu}_1 = \frac{\theta}{2} \implies \hat{\theta} = 2\hat{\mu}_1 = 2\frac{1}{n}\sum_{i=1}^n\sum_{i=1}^n X_i = \frac{2}{n}\sum_{i=1}^n X_i = 2\overline{X}_n.$$

Again, this estimate is unbiased (easy) and consistent (continuous mapping and WLLN).

5.5. Example: As one last example, consider $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. We know that $\mu_1 = \mathbb{E}[X_i] = \mu$ and $\mu_2 = \mathbb{E}[X_i^2] = \text{Var}[X_i] + \mathbb{E}[X_i]^2 = \sigma^2 + \mu^2$. The MMEs are obtained by solving $\hat{\mu}_1 = \mu$ and $\hat{\mu}_2 = \sigma^2 + \mu^2$. Both are consistent; $\hat{\mu}$ is unbiased while $\hat{\sigma}^2$ is biased.

$$\hat{\mu} = \overline{X}_n, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2.$$

Section 2. Maximum Likelihood

5.6. Motivation: Let $X_1, \ldots, X_n$ be iid random variables with pdf $f(\mathbf{x}; \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Omega$. Let $(x_1, \ldots, x_n)$ be observe from $(X_1, \ldots, X_n)$.

5.7. Note: The **likelihood function** $\mathcal{L} : \Omega \to [0, \infty)$ is defined by

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}).$$

It is important to recognize that

- The joint pdf is a function of data $\mathbf{x}$, indexed by parameter $\boldsymbol{\theta}$.
- The likelihood function is a function parameter $\boldsymbol{\theta}$, indexed by data $\mathbf{x}$.

The maximum likelihood method picks $\boldsymbol{\theta}$ that maximizes the likelihood function. We call this $\hat{\theta}$ the **maximum likelihood estimate** (MLE) of θ. In practice, it is often easier to maximize the **log likelihood function**:

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \log \left(\prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \right) = \sum_{i=1}^{n} \log(f(x_i; \boldsymbol{\theta})).$$

5.8. Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$. Then

$$f(x; \theta) = \begin{cases} 1/\theta & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

The likelihood function is given by

$$\mathcal{L}(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \begin{cases} 1/\theta & \theta > x_i \\ 0 & \text{otherwise} \end{cases} = \begin{cases} \theta^{-n} & \theta > \max_i x_i \\ 0 & \text{otherwise} \end{cases}$$

Observe the maximum happens at $\hat{\theta} = \max_i x_i$. This is thus the MLE.

5.9. Theorem (Invariance of MLE)**:** *If $\hat{\theta}$ is the MLE of θ, then for any function g, $g(\hat{\theta})$ is the MLE of $g(\theta)$.*

5.10. Example: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Previously, we derived that $\hat{\lambda}_{\text{MLE}} = \overline{X_n}$. How about the MLE of $\mathbb{E}[X_i^2]$? Recall that $\mathbb{E}[X_i^2] = \text{Var}[X_i] + \mathbb{E}[X_i]^2 = \lambda + \lambda^2 = \lambda(\lambda + 1)$. By the invariant property, the MLE is simply $\overline{X}_n(\overline{X}_n + 1)$. How about the MLE of

$$\Pr(X_i = 0) = \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda}?$$

Again, it's just $e^{-\overline{X}_n}$.

5.11. Note: From now on, let θ be a scalar. Define the **score function** by

$$S(\theta; \mathbf{x}) = \frac{d}{d\theta}\ell(\theta; \mathbf{x}).$$

Define the **information function** by

$$I(\theta; \mathbf{x}) = -\frac{d^2}{d\theta^2}\ell(\theta; \mathbf{x}).$$

Define the **expected information function** by

$$J(\theta) = \mathbb{E}[I(\theta; \mathbf{x})].$$

5.12. Theorem (Asymptotic Normality and Consistency of MLE)**:** *Under some regularity conditions,* $(\hat{\theta} - \theta)[J(\theta)]^{1/2} \to_D N(0, 1)$ *and* $\hat{\theta} \to_p \theta$.

5.13. Example: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Then

$$\ell(\lambda; \mathbf{x}) = \log \mathcal{L}(\lambda; \mathbf{x}) = \left(\sum_{i=1}^n x_i\right) \log \lambda - n\lambda - \sum_{i=1}^n \log(x_i!).$$

$$S(\lambda; \mathbf{x}) = \frac{d}{d\lambda}\ell(\lambda; \mathbf{x}) = \frac{\sum_{i=1}^n x_i}{\lambda} - n.$$

$$I(\lambda; \mathbf{x}) = -\frac{d^2}{d\lambda^2}\ell(\lambda; \mathbf{x}) = -\frac{-\sum_{i=1}^n x_i}{\lambda^2} = \frac{\sum_{i=1}^n x_i}{\lambda^2}$$

$$J(\lambda) = \mathbb{E}[I(\lambda; \mathbf{x})] = \mathbb{E}\left[\frac{\sum_{i=1}^n x_i}{\lambda^2}\right] = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}.$$

By this theorem, we have

$$(\hat{\lambda} - \lambda)[J(\lambda)]^{1/2} \to_D N(0, 1), \quad \hat{\lambda}\overline{X}_n \to_p \lambda.$$