Notes on STAT-330: Mathematical Statistics

Unversity of Waterloo

David Duan

Last Updated: February 26, 2021 (Draft)

Contents

Chapter 1. Univariate Random Variables

Section 1. Probability

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 function**, $\Pr : \mathcal{A} \to \mathbb{R}$, a function that satisfies three axioms:
  - $\Pr(A) \geq 0$ for all $A \subseteq S$.
  - $\Pr(S) = 1$.
  - If we have $\{A_1, A_2, A_3, \dots\} \subseteq S$ that are mutually exclusive, then

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

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

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

Proof. Trivial. □

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 does not influence the probability of the other.

Section 2. Random Variables

1.6. Definition: A **random variable** X is a function from S to $\mathbb{R}$ such that

$$\{X \leq x\} := \{A \in S : X(A) \leq x\}$$

is defined (i.e., is a valid event) for all $x \in \mathbb{R}$.

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 $A = \{\text{Tail}, \text{Head}, \text{Head}\}$, then the random variable X has the following effect:

$$\{\text{Tail}, \text{Head}, \text{Head}\} \overset{X}{\mapsto} 2 \qquad \text{(as there are two heads).}$$

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

1.9. Proposition: *Let F be a 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:

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

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

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

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

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

Let's plot its cdf:

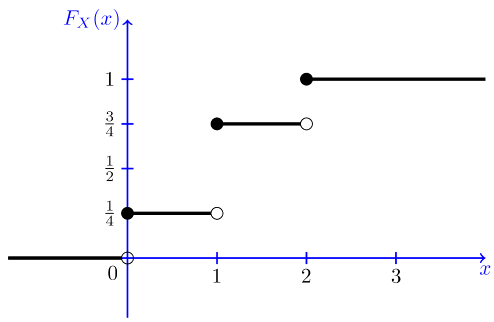

Figure 1.1: The cumulative distribution function of a discrete random variable X.

Observe that each component of the cdf of a discrete random variable will be of the form $a_i \le x \le a_{i+1}$, except the first component looks like $x < a_1$ and the last one looks like $a_n \le x$. 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, $\{x : f(x) > 0\}$, is called its **support set**.

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

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

Proof. Trivial. ∎

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: The **probability density function** (pdf) of a continuous random variable X is given by

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

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

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 $x \in \mathbb{R} \setminus \{1\}$, so we get

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

The support of X is the set $\{x : f(x) > 0\}$, i.e., 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* $b \in \mathbb{R}$.

Proof. Trivial. □

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 $b > a$:

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

By taking the derivative of $F(x)$, we obtain

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

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

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

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

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

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

$$\Pr(-2 < X < 3) = \int_{-2}^{3} f(x)\,dx = \int_{1}^{3} \frac{1}{x^2}\,dx = -\frac{1}{x}\bigg|_1^3 = 1 - \frac{1}{3} = \frac{2}{3}.$$

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). $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ *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** 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 diverges, 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** 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 diverges, 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. 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. $\square$

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] Expected value 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 **kth moment about the origin**.
  - The 1st moment of X is known as 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 **kth moment about the mean**.
  - The 2nd central moment of X is known as 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 functions that uniquely determines a distribution: the pdf and cdf. A third type of function, known as the *moement generating function*, also uniquely determines a distribution. It provides the basis of an alternative route to analytical results compared with working directly with pdfs and cdfs. As its name implies, the mgf 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. Example: 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

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

For example, let $X \sim \text{Poisson}(\lambda)$ with pdf $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x \in \mathbb{Z}_{\geq 0}$. Then its mgf is given by

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

1.45. 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.

1.46. 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 $t_X \in (-h, h)$. Thus,

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

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. □

1.47. 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

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

Now use the proposition above,

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

1.48. 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 $M''_X(t)$.
(2). The mean is given by $\mathbb{E}[X] = M'_X(0)$.
(3). The variance is given by $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X] = M''_X(0) - (M'_X(0))^2$.

1.49. 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*

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

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 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. □

1.50. Example: Let us derive the mean and variance for $X \sim \text{Poisson}(\lambda)$. Recall in Example 1.47, we derived that the mgf of X is given by

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

Compute its first and second derivatives:

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

By the proposition above, we see that

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

1.51. 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.,*

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

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 the sample space S. The **joint cumulative distribution function** of X and Y is given by

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

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

2.2. 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* $\lim_{(x,y)\to(\infty,\infty)} F(x,y) = 1$.

2.3. 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

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

2.4. 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

2.5. 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**.

2.6. 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

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

2.7. 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 $R \subseteq \mathbb{R}^2$, $\Pr((x,y) \in R) = \sum_{(x,y)\in R} f(x,y)$.

2.8. 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.,

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

2.9. Example: Let $p \in (0,1)$ and X, Y be discrete random variable 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,

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

(2). Find marginal pmfs.

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

We conclude that X and Y marginally follow geometric distribution.

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

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

Section 3. Bivariate Continuous Distributions

2.10. Definition: If $F(x, y)$ is continuous and

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

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

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

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

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

2.11. 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 $R \subseteq \mathbb{R}^2$, $\Pr((x, y) \in R) = \iint_R f(x, y)\, dx\, dy$.

2.12. 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.,

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

2.13. 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.

2.14. Example: Suppose that (X, Y) are 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 $\Pr(X \leq Y)$*.*

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

(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

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

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

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

2.17. 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*

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

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

2.19. 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).

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

2.21. 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

2.22. 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.

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

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

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.

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

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

2.25. Proposition: *If $X_1, \dots, X_n$ 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, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

2.26. *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.

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

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

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

2.28. 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.

2.29. 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). $\mathrm{Var}(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 \mathrm{Var}[X_i] + \sum_{i\neq j} a_i a_j \mathrm{Cov}(X_i X_j)$.
(5). If $X_1, \ldots, X_n$ *are independent,* $\mathrm{Var}(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 \mathrm{Var}(X_i)$.

2.30. 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 6. Conditional Distributions

2.31. Definition: Suppose X and Y are bivariate discrete r.v. 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

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

2.32. 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 $\sum_y f_Y(y \mid x) = 1$.*

2.33. Definition: Suppose X and Y are bivariate continuous r.v. 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

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

2.34. Remark: One can show that

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

2.35. 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 $\int_{-\infty}^{\infty} f_Y(y \mid x) = 1$.*

2.36. 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*

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

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

2.37. 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 $f_X(x \mid y) = f(x, y)/f_Y(y)$. □

Section 7. Conditional Expectation

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

$$\mathbb{E}[g(Y) \mid X = x] = \begin{cases} \sum_y g(y) f_Y(y \mid x) & \text{provided that } \sum_y |g(y)| f_Y(y \mid x) < \infty \\ \int_{-\infty}^{\infty} g(y) f_2(y \mid x) & \text{provided that } \int_{-\infty}^{\infty} |g(y)| f_2(y \mid x) < \infty \end{cases}$$

2.39. 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**.

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

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

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. □

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

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

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

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

2.43. Example:

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

2.44. 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.

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

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

Proof.

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

□

2.46. Theorem:

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

Section 8. Joint Moment Generating Functions

2.47. Definition: Let X, Y be a pair of random variables. If $\mathbb{E}[e^{t_X X + t_Y Y}]$ exists for $t_X \in (-h_X, h_X)$ and $t_Y \in (-h_Y, h_Y)$, $h_X, h_Y > 0$, 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 MFG of 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 $t_i \in (-h_i, h_i)$, $i = 1, \ldots, n$.

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

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

2.49. 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 $M(t_1, \ldots, t_n) = \prod M_{X_i}(t_i)$.

Section 9. Multinomial Distribution

2.50. 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.

2.51. Definition: Let $(X_1, X_2, \ldots, X_k)$ be discrete random variables with joint probability function

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

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

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

Then $(X_1, X_2, \ldots, X_k)$ is said to have a **Multinomial distribution**. We write

$$(X_1, X_2, \ldots, X_k) \sim \text{Multinomial}\,(n; p_1, p_2, \ldots, p_k).$$

2.52. Proposition: *If* $(X_1, X_2, \ldots, X_k) \sim \text{Multinomial}\,(n; p_1, p_2, \ldots, p_k)$, *then*

(1). $(X_1, X_2, \ldots, X_{k-1})$ *has joint moment generating function*

$$\begin{aligned} M(t_1, t_2, \ldots, t_{k-1}) &= \mathbb{E}\left(e^{t_1 X_1 + t_2 X_2 + \cdots + t_{k-1} X_{k-1}}\right) \\ &= \left(p_1 e^{t_1} + p_2 e^{t_2} \cdots + p_{k-1} e^{t_{k-1}} + p_k\right)^n, \quad (t_1, t_2, \ldots, t_{k-1}) \in \mathbb{R}^{k-1}. \end{aligned}$$

(2). Any subset of $\{X_1, X_2, \ldots, X_k\}$ *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 $i \neq j$, $\text{Cov}\,(X_i, X_j) = -np_i p_j$.

(5). The conditional distribution of any subset of $(X_1, X_2, \ldots, X_k)$ *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 10. Bivariate Normal Distribution

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

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

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

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

2.54. Proposition: *If* $\mathbf{X} = [X_1, X_2]^T \sim \mathrm{BVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *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). $\mathrm{Cov}(X_1, X_2) = \rho\sigma_1\sigma_2$, $\mathrm{Cor}(X_1, X_2) = \rho$, *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* $\mathbf{b} \in \mathbb{R}^{2\times 1}$,

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

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

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

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