Lecture Notes 1: Wednesday, Jan. 6, 2016
Brief Review of Basic Probability

1 Probability: Basic Set Theory

Terminology:

- The **sample space** Ω is the set of possible outcomes of an experiment. An **element** of Ω is denoted by ω. Subsets of Ω are called **Events**.
- The **empty set** $\emptyset$ is the set with no elements.
- Given an event A, the complement of $A \subset \Omega$ is

$$A^c = \{\omega \in \Omega : \omega \notin A\}.$$

- $A \cup B$ is the union of event A and B (diagram).
- $A \cap B$ is the intersection of event A and B (diagram).
- If $A \cap B = \emptyset$, then events A and B are **disjoint** or **mutually exclusive**.
- We say that $A_1, A_2, \ldots$ are **disjoint** or **mutually exclusive** if $A_i \cap A_j = \emptyset$ whenever $i \neq j$.
- A **partition** of Ω is a sequence of disjoint sets $A_1, A_2, \ldots$ such that

$$\bigcup_{i=1}^{\infty} A_i = \Omega.$$

- Given an event A, denote the **indicator function** of A by

$$I_A(\omega) = I(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise.} \end{cases}$$

- A sequence of sets $A_1, A_2, \ldots$ is **monotone increasing** if $A_1 \subset A_2 \subset \ldots$ and we define

$$\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i.$$

- A sequence of sets $A_1, A_2, \ldots$ is **monotone decreasing** if $A_1 \supset A_2 \supset \ldots$ and we define

$$\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i.$$

Example 1 *Let* $\Omega = \mathbf{R}$ *and let* $A_i = [0, 1/i)$ *for* $i = 1, 2, \ldots$. *Then*

$$\bigcup_{i=1}^{\infty} A_i = [0, 1) \quad \bigcap_{i=1}^{\infty} A_i = \{0\}.$$

If instead we define $A_i = (0, 1/i)$ *for* $i = 1, 2, \ldots$. *Then*

$$\bigcup_{i=1}^{\infty} A_i = (0, 1) \quad \bigcap_{i=1}^{\infty} A_i = \emptyset.$$

We will assign a real number $\mathbb{P}(A)$ to every event A, called the probability of A. we call $\mathbb{P}$ a probability distribution or probability measure if it satisfies the following three axioms

Definition 2 *A function* $\mathbb{P}$ *that assigns a real number* $\mathbb{P}(A)$ *to each event* A *is a* **probability distribution** *or a* **probability measure** *if it satisfies the following three axioms*

Axiom 1 $\mathbb{P}(A) \geq 0$ *for every* A.

Axiom 2 $\mathbb{P}(\Omega) = 1$.

Axiom 3 If $A_1, A_2, \ldots$ *are disjoint, then*

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

Generally, it is not feasible to assign probabilities to all subsets of a sample space Ω. Instead, one restricts attention to a set of events called σ-field.

Definition 3 *A* σ*-field (or* σ*-algebra)* $(\Omega, \mathcal{B})$ *consists of a sample space* Ω *and a collection of subsets of* Ω*, denoted as* $\mathcal{B}$*, satisfying the following conditions.*

1. $\emptyset \in \mathcal{B}$.
2. *If* $A \in \mathcal{B}$, *then* $A^c \in \mathcal{B}$.
3. *If* $A_1, A_2, \ldots$ *is a sequence of elements of* $\mathcal{B}$ *then* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$.

Hence $\mathcal{B}$ is closed under countable union, countable intersection and complements. In particular, if $A_1, A_2, \ldots$ is a sequence of elements of $\mathcal{B}$, then from rules 2 and 3, we have

$$\bigcap_{i=1}^{\infty} A_i = \left(\bigcup_{i=1}^{\infty} A_i^c\right)^c \in \mathcal{B}.$$

Definition 4 *A probability space is a triple* $(\Omega, \mathcal{B}, \mathbb{P})$ *such that* Ω *is a sample space,* $\mathcal{B}$ *is a* σ*-algebra of subsets of* Ω*, and* $\mathbb{P}$ *is a* **probability measure***; that is,* $\mathbb{P}$ *is a function with domain* $\mathcal{B}$ *and range* $[0,1]$ *such that*

1. $\mathbb{P}(A) \geq 0$ *for all* $A \in \mathcal{B}$.
2. $\mathbb{P}(\Omega) = 1$.
3. *If* $A_1, A_2, \ldots$ *is a sequence of elements of* $\mathcal{B}$ *that are disjoint, then*

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

Example 5 **σ-algebra** *Let* $\Omega = \mathbf{R}$*. Then* $\mathcal{B}$ *is chosen to contain all sets of the form*

$$[a,b], \quad (a,b], \quad (a,b), \quad [a,b)$$

for all real numbers a and b. Also, from the Properties of $\mathcal{B}$*, it follows that* $\mathcal{B}$ *contains all sets that can be formed by taking (possibly countably infinite) unions and intersections of sets of the above varieties.*

2 Conditional Probability

Definition 6 *If A and B are two events with* $\mathbb{P}(B) \neq 0$*. The conditional probability of A given B is defined to be:*

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

Conditional probability $\mathbb{P}(\cdot|B)$ is a probability measure on B, satisfying the axioms of probability for fixed event B. That is, the relevant sample space becomes B rather Ω.

Let A, B be two events and $\mathbb{P}(B) \neq 0$, then

- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$.
- If $\mathbb{P}(A) \neq 0$, then $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$.
- It follows that
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$
- For a partition $A_1, A_2, \ldots$,

$$\mathbb{P}(B) = \sum_{i=1}^{\infty} \mathbb{P}(B|A_i)\mathbb{P}(A_i). \tag{1}$$

Theorem 7 *(Bayes' Theorem) Let $A_1, A_2, \ldots, A_n$ be a partition of Ω such that $\mathbb{P}(A_i) > 0$ for all i. If $\mathbb{P}(B) > 0$, then for each $i = 1, \ldots, n$*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}.$$

PROOF.

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B \cap A_i)}{P(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}.$$

■

Example 8 *(Three prisoners) Here we use events A, B, C and $\mathcal{W}$ to illustrate the slippery nature of conditional probabilities. Three prisoners A, B, and C are on death row. The governor decides to pardon one of the three and chooses at random the prisoner to pardon. He informs the warden of his choice but requests that the name be kept secret for a few days. The next day, A tries to get the warden to tell him who has been pardoned. The warden refuses. A then asks which of B or C will be executed. The warden thinks for a while, then tells A that B is to be executed.*

Questions:

- *Does Warden give A additional information whether A will be pardoned?*
- *Does Warden give A additional information whether C will be pardoned?*
- *Why does A think his chance of being pardoned has risen to $1/2$? Why is he wrong?*
- *Where did we use the Bayes' Theorem?*

Exercise. Read through all Examples in the textbook Chapter 1.

Definition 9 *A and B are said to be independent events if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.*

Definition 10 *A collection of events $A_1, A_2, \ldots, A_n$ are* **mutually independent** *if for any subcollections $A_{i_1}, \ldots, A_{i_k}$ we have*

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}).$$

3 Distribution Functions

Definition 11 *A random variable is a function from a sample space Ω into real numbers* $\mathbf{R}$*:*

$$X : \Omega \to \mathbf{R}$$

that assigns a real number $X(\omega)$ to each outcome ω.

Definition 12 *The Cumulative Distribution Function (cdf) of a random variable X, denoted by $F_X(x)$, is defined by*

$$F_X(x) = P(X \leq x), \qquad \textit{for all} \quad -\infty < x < \infty.$$

Theorem 13 *The function $F(x)$ is a cdf if and only if the following three conditions hold:*

1. $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$.
2. *$F(x)$ is a nondecreasing function of x.*
3. *F is right-continuous: $F(x) = F(x^+)$ for all x, where $F(x) = F(x^+) = \lim_{y\to x, y>x} F(y)$.*

Definition 14 *A random variable X is continuous if $F_X(x)$ is a continuous function of x.*

Definition 15 *A random variable X is discrete if $F_X(x)$ is a step function of x.*

Exercise. Plot the step function for $F_X(x)$ for $X = \#$ of heads showing up after tossing 3 fair coins. Convince yourself it is right-continuous.

Example 16 *An example of a continuous cdf is the function*

$$F_X(x) = \frac{1}{1 + e^{-x}},$$

which satisfies the conditions of Theorem 13. For example, condition 1 is satisfied given

$$\lim_{x\to-\infty} e^{-x} = \infty \quad \textit{and} \quad \lim_{x\to\infty} e^{-x} = 0.$$

Differentiating $F_X(x)$ gives

$$\frac{d}{dx} F_X(x) = \frac{e^{-x}}{(1 + e^{-x})^2} > 0,$$

showing that $F_X(x)$ is increasing. F_X is not only right-continuous but also continuous. This is a special case of the logistic distribution.

Google this distribution to find out the plots for the CDF.

Exercise. Understand the arguments in this example.

Definition 17 *X has a point mass distribution at a, written as $Z \sim \delta_a$, if $\mathbb{P}(Z = a) = 1$ in which case*

$$F_Z(z) = \begin{cases} 0 & \text{if } z < a \\ 1 & \text{if } z \geq a. \end{cases}$$

and the probability mass function is $f(x) = 1$ for $x = a$ and 0 otherwise.

Example 18 (Sticky gauge) *If $F_X(x)$ is not a continuous function of x, it is possible for it to be a mixture of continuous pieces and jumps. For example, if we modify Example 16 to be, for some ϵ, $1 > \epsilon > 0$*

$$F_Y(y) = \begin{cases} \frac{1-\epsilon}{1+e^{-y}} & \text{if } y < 0 \\ \epsilon + \frac{1-\epsilon}{1+e^{-y}} & \text{otherwise.} \end{cases}$$

Then $F_Y(y)$ is the CDF of a random variable Y. The function $F_Y(y)$ has a jump of height ϵ at $y = 0$ and otherwise is continuous. This model might be appropriate if we were observing the reading from a gauge, which sometimes sticks at 0, where ϵ is the probability that the gauge sticks. That is, we have for $Z \sim \delta_0$ and $X \sim$ logistics distribution as in Example (16)

$$F_Y(y) = \epsilon F_Z(y) + (1 - \epsilon) F_X(y)$$

Exercise. Show that F_Y as in the above example is indeed a CDF.

4 Random Variables

Notation for Random Variables.

- Discrete RV:
  - A discrete RV is a random variable that can take on only a finite or at most a countably infinite number of values $\{x_1, x_2, \dots, \}$
  - In general, a countably infinite set is one that can be put into one-to-one correspondence with the integers.
  - Probability mass function (pmf)/frequency function $f_X(x) = P(X = x)$ for all x.
  - Let A be any subset $\mathbf{R}$, then $P(X \in A) = \sum_{x \in A} f_X(x)$. Since X is discrete, $f_X(x)$ is nonzero for at most a countable number of points x by definition.
  - Thus the sum can be interpreted as a countable sum even if A contains an uncountable number of points.

5 Continuous Distributions

In applications, we are often interested in random variables that can take on a continuum of values rather than a finite or countably infinite number.

Definition 19 *A random variable X is* **continuous** *if there exists a function f_X such that $f(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x) = 1$ and*

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx$$

The function f_X is called the **probability density function** *(PDF). We have*

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \quad \text{for all } x. \tag{2}$$

and $f_X(x) = F_X'(x)$ at all points at which F_X is differentiable.

- *If f is continuous, then $f_X(x) = F_X{}'(x)$ for all x by the Fundamental Theorem of Calculus.*
- *Otherwise, $f_X(x) = F_X'(x)$ almost everywhere – that is, except for x in a set of Lebesgue measure 0. (This is not required material)*

Warning!
The relationship (2) does not always hold because $F_X(x)$ may be continuous but not differentiable. In fact, there exists continuous random variables for which the integral relationship does not exist for any $f_X(x)$.

In summary, for a continuous RV X, for which we assume the existence of $f(x)$ that satisfies the requirement in Definition 19 with respect to $F(x)$,

- $F(x) = P(X \leq x) = \int_{-\infty}^{x} f((u)du$ is a continuous function of x.
- From the fundamental theorem of calculus, if f is continuous at x, then $F'(x) = f(x)$.
- $P(a < X < b) = \int_a^b f(x)dx$.
- As a consequence, $P(x = c) = \int_c^c f(x)dx = 0$.
- The cdf can be used to evaluate the probability that X falls into an interval:

$$P(a \leq x \leq b) = \int_a^b f(x)dx = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = F(b) - F(a).$$

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = \int_a^b f(x)dx$$

which is not true for a discrete RV.

6 Density and Mass Functions

Theorem 20 *A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if*

1. $f_X(x) \geq 0$ *for all* x.
2. $\sum_x f_X(x) = 1$ *(pmf)* *or* $\int_{-\infty}^{\infty} f_X(x) = 1$ *(pdf).*

PROOF. If $f_X(x)$ is a pdf (or pmf), then the two properties follow immediately from the definitions; in particular, for a pdf, using (2), we have that

$$1 = \lim_{x\to\infty} F_X(x) = \int_{-\infty}^{\infty} f_X(u)du.$$

For the converse: once we have $f_X(x)$ satisfying the two conditions above, we can define $F_X(x)$. Suppose X is a continuous RV and $\int_{-\infty}^{\infty} f_X(x) = 1$, then define

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

which clearly satisfies all conditions required in Theorem 13. In particular:

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
2. $F(x)$ is a nondecreasing function of x given $f_X(x) \geq 0$ for all x.
3. By the dominated convergence theorem, $F_X(x)$ is continuous.

Thus $f_X(x)$ is the pdf of a continuous random variable X by definition;

Otherwise, let $F_X(x) = \sum_{u\leq x} f_X(u)$, which is a step function and is right-continuous for X being a discrete RV. (Check it by yourself) ■

7 Expected Values

- Discrete RV $E\left(g(X)\right) = \sum_{x\in\mathcal{X}} g(x)f_X(x)$ provided that $\sum_{x\in\mathcal{X}} |g(x)|f_X(x) < \infty$. If the sum diverges, the expectation is undefined.
- Continuous RV: $E\left(g(X)\right) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$ provided that $\int_{-\infty}^{\infty} |g(x)|f_X(x)dx < \infty$.
- **Linear Combinations**
  - Even if the RVs are not independent,

$$E\left(g_1(X) + bg_2(X) + c\right) = aE\left(g_1(X)\right) + bE\left(g_2(X)\right) + c,$$

for example,

$$E\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_i a_iE\left(X_i\right).$$

- If $g_1(x) \geq g_2(x)$ for all x, then $E\left(g_1(X)\right) \geq E\left(g_2(X)\right)$.

- If $X_1, \ldots, X_n$ are *independent*, then

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_i E\left(X_i\right).$$

- **Mean and Variance**

  - $\mu = E\left(X\right)$ is the **mean**.
  - $\sigma^2 = \mathrm{Var}\left(X\right) = E\left((X-\mu)^2\right)$ is the **Variance**. Also

$$\mathrm{Var}\left(X\right) = E\left(X^2\right) - \mu^2.$$

  - If $X_1, \ldots, X_n$ are independent RVs, then

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

  - Covariance $\mathrm{Cov}(X,Y) = E(X-\mu_x)(Y-\mu_y) \quad \mathrm{Corr}(X,Y) = \mathrm{Cov}(X,Y)/\sigma_x\sigma_y$.

Example: Coupon Collection

The following description comes from Wikipedia (Google "Coupon collector's problem"): "In probability theory, the coupon collector's problem describes the "collect all coupons and win" contests. It asks the following question: Suppose that there are n coupons, from which coupons are being collected with replacement. What is the probability that more than t sample trials are needed to collect all n coupons? The mathematical analysis of the problem reveals that the expected number of trials needed grows as $\Theta(n\log(n))$. For example, when $n = 50$ it takes about 225 samples to collect all 50 coupons."

Collecting Flags of Swiss Cantons and half cantons. Coupons in this context map to the dominos with flags of Swiss Cantons and half cantons. The key for solving the problem is to understand that it takes very little time to collect the first few dominos. On the other hand, it takes a long time to collect the last few new dominos.

Let T be the time to collect all $n = 36$ dominos. Let T_i be the time to collect the i-th domino after $i-1$ dominos have been collected. Note that $T_1 = E[T_1] = 1$. Observe that the probability of collecting a new domino given that $i-1$ unique ones have been collected is $p_i = (n-i+1)/n$ and hence $E(T_i) = 1/p_i$. Note that $E(T_n) = n$. We will use the following facts:

$$E(T) = E(T_1) + \ldots + E(T_n) = nH_n = n\ln n + \gamma n + \frac{1}{2} + o(1)$$

as $n \to \infty$, where $H_n = \sum_{i=1}^{n} \frac{1}{n}$ is the harmonic number and $\gamma \approx 0.5772156649$.

Exercise. Compute $\mathrm{Var}(T)$.

8 Transformations, Chapter 2

Transformations

X is the *old* variable. Y is the *new* variable.

$$X \sim f_X(x) \qquad Y = g(X).$$

Define the support of a distribution with density $f_X(x)$:

$$\mathcal{X} = \{x : f_X(x) > 0\} \quad \text{and} \quad \mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}. \tag{3}$$

Notation. Let $A \subset \mathcal{Y}$.

$g(x) : \mathcal{X} \to \mathcal{Y}$ defines a mapping from sample space $\mathcal{X}$ to a new sample space $\mathcal{Y}$.
$g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\}$: g^{-1} defines an inverse mapping from $\mathcal{Y}$ to $\mathcal{X}$.

Now we can write for any set $A \subset \mathcal{Y}$,

$$P(Y \in A) = P(g(X) \in A) = P(\{x \in \mathcal{X} : g(x) \in A\}) = P(X \in g^{-1}(A)).$$

Discrete: If X is a discrete RV, then $\mathcal{X}$ is countable. The sample space $\mathcal{Y}$ for $Y = g(X)$ is also a countable set. Thus Y is also a discrete random variable.

- For $y \in \mathcal{Y}$,

  $$f_Y(y) = P(Y = y) = \sum_{x \in g^{-1}(y)} P(X = x) = \sum_{x \in g^{-1}(y)} f_X(x) := P(X \in g^{-1}(y)).$$

  where $\{x \in g^{-1}(y)\} = g^{-1}(\{y\}) = \{x \in \mathcal{X} : g(x) = y\}$.

- **Example:** $Y = n - X$ for $X \sim$ Binomial(n, p). Show that $Y \sim$Binomial$(n, 1 - p)$.

Continuous: CDF method: This method can be summarized as:

1. For each y, find the set $A_y = \{x \in \mathcal{X} : g(x) \leq y\}$.

2. Find the CDF by by definition of the r.v., that is,

   $$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(g(X) \leq y) = P(\{x \in \mathcal{X} : g(X) \leq y\}) \\ &= \int_{\{x \in \mathcal{X} :\ g(x) \leq y\}} f_X(x)dx = \int_{A_y} f_X(x)dx \end{aligned}$$

3. The pdf is $f_Y(y) = F_Y'(y)$.

Example.
Let $f_X(x) = e^{-x}$ for $x > 0$. Hence $F_X(x) = 1 - e^{-x}$. Let $Y = g(X) = \log X$. Then

$$\begin{aligned} F_Y(y) = P(Y \le y) &= P(\log(X) \le y) \\ &= P(X \le e^y) = F_X(e^y) = 1 - e^{-e^y}. \end{aligned}$$

Notes.
$\mathcal{X}$ is called the **support set** of the distribution. A function $g(x) : \mathcal{X} \to \mathcal{Y}$ is monotone means:

$$u > v \Rightarrow g(u) > g(v) \ \text{(increasing)} \quad \text{or} \quad u < v \Rightarrow g(u) > g(v) \ \text{(decreasing)}.$$

If the transformation is monotone, then it is *one-to-one* and *onto* from $\mathcal{X} \to \mathcal{Y}$. That is, each x goes to only one y and each y comes from at most one x (one-to-one). Also, for each $y \in \mathcal{Y}$ for $\mathcal{Y}$ as defined in (3), there is an $x \in \mathcal{X}$ such that $g(x) = y$ (onto). Thus the transformation uniquely pairs xs and ys. If g is monotone, then g^{-1} is single-valued; that is $g^{-1}(y) = x$ if and only if $y = g(x)$.

Theorem 21 *Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function:*

$$g(x) : \mathcal{X} \to \mathcal{Y} \qquad g \text{ monotone.}$$

$$X \to Y = g(X) \qquad \text{1-1 transformation.}$$

Suppose that $f_X(x)$ is continuous on $\mathcal{X}$ and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then the pdf of Y is given by

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y))}{dy} \right| \quad y \in \mathcal{Y} \quad \text{and} \quad f_Y(y) = 0 \quad \text{otherwise.}$$

Exercise. Show that the continuity of $f_X(x)$ and $\frac{d}{dy} g^{-1}(y)$ ensures that f_Y is continuous on $\mathcal{Y}$.

PROOF. Define

$$A(y) = \{x \in \mathcal{X} : \ g(x) \le y\}.$$

If g is monotonic increasing, this implies that

$$A(y) = \{x \in \mathcal{X} : \ g(x) \le y\} = \{x \in \mathcal{X} : \ x \le g^{-1}(y)\}$$

so for $h(y) = g^{-1}(y)$, we have

$$\begin{aligned} F_Y(y) &= P(X \in \mathcal{X} : g(X) \le y) = \int_{\{x \in \mathcal{X} : \ g(x) \le y\}} f_X(x)dx \\ &= \int_{\{x \in \mathcal{X} : \ x \le g^{-1}(y)\}} f_X(x)dx \ \text{ by monotonicity of } g \\ &= \int_{-\infty}^{g^{-1}(y)} f_X(x)dx = F_X(g^{-1}(y)) = F_X(h(y)) \end{aligned}$$

Now by letting $u = h(y)$, we apply the Chain Rule to obtain

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(h(y))}{dy} = \frac{dF_X(u)}{du} \cdot \frac{du}{dy} = f_X(h(y))\frac{dh(y)}{dy} = f_X(h(y))\left|\frac{dh(y)}{dy}\right|.$$

or simply

$$f_Y(y) = F'_Y(y) = F'_X(h(y))h'(y) = f_X(h(y))\frac{dh(y)}{dy} = f_X(h(y))\left|\frac{dh(y)}{dy}\right|.$$

If g is monotonic decreasing, this implies that

$$A(y) = \{x \in \mathcal{X} :\ g(x) \leq y\} = \{x \in \mathcal{X} :\ x \geq g^{-1}(y)\}$$

so for $h(y) = g^{-1}(y)$, we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = \int_{\{x \in \mathcal{X}:\ g(x) \leq y\}} f_X(x)dx \\ &= \int_{\{x \in \mathcal{X}:\ x \geq g^{-1}(y)\}} f_X(x)dx \ \text{ by monotonicity of } g \\ &= \int_{g^{-1}(y)}^{\infty} f_X(x)dx = 1 - F_X(g^{-1}(y)) = 1 - F_X(h(y)). \end{aligned}$$

The continuity of X is used to obtain the second equality on the line immediately above. Now by the chain rule we have

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

■

Recall that the transformation need not be monotone, for example, let $Y = X^2$. Then g is not monotone.

Example: Square transformation Suppose X is a continuous random variable. For $y > 0$ the cdf of $Y = X^2$ is

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= P(-\sqrt{y} < X \leq \sqrt{y}) = P(X \leq \sqrt{y}) - P(X \leq -\sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

The pdf of Y can now be obtained from the cdf by differentiation:

$$f_Y(y) = \frac{1}{2\sqrt{y}}f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}}f_X(-\sqrt{y}).$$

Theorem 22 (Theorem 2.1.10, CB Probability integral transformation) *Let X have continuous cdf $F_X(x)$ and define random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0, 1)$, that is, $P(Y \leq y) = y,\ 0 < y < 1$.*

Proposition 23 *Let X be a random variable with density f and cdf F, where F is strictly increasing on some interval I, $F = 0$ to the left of I and $F = 1$ to the right of I. Let U be uniform on $[0, 1]$, and let $X = F^{-1}(U)$. Then the cdf of X is F.*

For proof, see CB Page 54.

Example.

Recall that the transformation need not be monotone, for example, let $Y = X^2$.

$$P(Y \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y})$$

Example 2.1.6, CB (Inverted gamma pdf) Let $f_X(x)$ be gamma pdf

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta},\ 0 < x < \infty,\ \alpha, \beta > 0.$$

Suppose we want to find the pdf of $g(X) = 1/X$ for $\alpha = n$.

Now suppose $\mathcal{X}$ and $\mathcal{Y}$ are both the interval $(0, \infty)$. If we let $y = g(x)$, then $g^{-1}(y) = 1/y$ and $\frac{d}{dy} g^{-1}(y) = -1/y^2$. Applying the above theorem, for $y \in (0, \infty)$, we have

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= \frac{1}{\Gamma(n)\beta^n} (1/y)^{n-1} e^{-/(y\beta)} \frac{1}{y^2} \\ &= \frac{1}{(n-1)!\beta^n} (1/y)^{n+1} e^{-/(y\beta)}, \end{aligned}$$

a special case of a pdf known as the *inverted gamma pdf.*

9 Joint and Marginal Distributions: Chapter 4

Definition 24 *An n-dimensional random vector is a function from a sample space Ω into $\mathbf{R}^n$, n-dimensional Euclidean space.*

Definition 25 *For any two random variables X and Y, the Cumulative Distribution Function (cdf) of X and Y denoted by $F(x, y)$, is defined by*

$$F(x, y) = P(X \leq x, Y \leq y), \qquad \text{for all } -\infty < x, y < \infty$$

regardless of whether X and Y are continuous or discrete.

- The probability that (X, Y) belongs to a given rectangle is:

$$P(x_1 < x \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1).$$

- If $X_1, \ldots, X_n$ are jointly distributed random variables, their joint cdf is

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n), \qquad \text{for all } -\infty < x_1, x_2, \ldots, x_n < \infty.$$

Definition 26 *Let (X, Y) be discrete bivariate random vector. Then the function from $\mathbf{R}^2$ to $\mathbf{R}$ defined by $f(x, y) = P(X = x, Y = y)$ is called the joint probability mass function (joint pmf) of (X, Y). Notation $f_{X,Y}(x, y)$ will also be used.*

The joint *pmf* can be used to compute the probability of any event defined in terms of (X, Y).

- Let A be any subset of $\mathbf{R}^2$. Then $P((X, Y) \in A) = \sum_{(x,y)\in A} f(x, y)$.
- Since (X, Y) is discrete, $f(x, y)$ is nonzero for at most a countable number of points (x, y). Thus the sum can be interpreted as a countable sum even if A contains an uncountable number of points.

Theorem 27 *Let (X, Y) be discrete bivariate random vector with (joint pmf) $f_{X,Y}(x, y)$. Then the marginal pmfs of X and Y, $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, are given by:*

$$f_X(x) = \sum_{y\in\mathbf{R}} f_{X,Y}(x, y) \quad \textit{and} \quad f_Y(y) = \sum_{x\in\mathbf{R}} f_{X,Y}(x, y).$$

Proof. We will prove the result for $f_X(x)$. The proof for $f_Y(y)$ is similar. For any $x \in \mathbf{R}$, let $A_x = \{(x, y) : -\infty < y < \infty\}$. That is, A_x is the line in the place with the first coordinate equal to x. Then for any $x \in \mathbf{R}$,

$$\begin{aligned} f_X(x) &= P(X = x) = P(X = x, -\infty < y < \infty) \\ &= P((X, Y) \in A_x) = \sum_{(x,y)\in A_x} f_{X,Y}(x, y) \\ &= \sum_{y\in\mathbf{R}} f_{X,Y}(x, y) \end{aligned}$$

■

Example: Multinomial Distribution The multivariate version of a Binomial is called a Multinomial. Consider drawing a ball from an urn with has balls with k different colors labeled "color 1, color 2, ..., color k." Let $p = (p_1, p_2, \ldots, p_k)$ where $\sum_j p_j = 1$ and p_j is the probability of drawing color j. Draw n balls from the urn (independently and with replacement) and let $X = (X_1, X_2, \ldots, X_k)$ be the count of the number of balls of each color drawn. We say that X has a Multinomial (n, p) distribution. The pdf is

$$f(x) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \ldots p_k^{x_k}.$$

Exercise: Show the following lemma.

Lemma 28 *The marginal distribution of X_i is the Binomial(n, p_i) distribution.*

Suppose (X, Y) are continuous random variables with a joint cdf, $F(x, y)$. For the joint cdf $F(x, y)$, we have

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t) dt ds.$$

From the bivariate fundamental theorem of calculus,

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

at continuity points of $f(x, y)$. This relationship is useful in situations where an expression for $F(x, y)$ can be found. The mixed partial derivative can be computed to find the joint pdf.

Definition 29 *A function $f(x, y)$ from $\mathbf{R}^2$ into $\mathbf{R}$ is called the joint probability density function (joint pdf) of the continuous bivariate random vector (X, Y) if, for every $A \subset \mathbf{R}^2$,*

$$P((X, Y) \in A) = \int_A \int f(x, y) dx dy.$$

- In particular, if $A = \{(X, Y) | X \leq x \text{ and } Y \leq y\}$,

$$P((X, Y) \in A) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) dv du = F(x, y)$$

- The expected value of $Eg(X, Y)$ is defined by:

$$E\left(g(X, Y)\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

- The marginal pdfs of X and Y are given by:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y) dy \quad -\infty < x < \infty \\ f_Y(y) &= \int_{-\infty}^{\infty} f(x, y) dx \quad -\infty < y < \infty \end{aligned}$$

- Any function $f(x, y)$ satisfying $f(x, y) \geq 0$ for all $(x, y) \in \mathbf{R}^2$ and

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_X(x, y) dx dy$$

  is the joint pdf of some continuous bivariate random vector (X, Y).

The marginal cdfs of X, or $F_X(x)$ is given by:

$$F_X(x) = \lim_{y\to\infty} F(x,y) := F(x,\infty)$$

PROOF. Let $y_n \to \infty$ be an increasing sequence.

$$\begin{aligned} F_X(x) &= P(X \le x) = P(X \le x, -\infty < y < \infty) \\ &= P(\bigcup_n \{X \le x, Y \le y_n\}) \\ &= P\left(\lim_{y_n\to\infty} \{X \le x, Y \le y_n\}\right) = \lim_{y_n\to\infty} P(X \le x, Y \le y_n) \ (*) \\ &= \lim_{y_n\to\infty} F(x,y_n) := F(x,\infty) \end{aligned}$$

■

EXERCISE: PROVE $(*)$. In "A First Course in Probability" by Sheldon Ross, 8th Edition, this property is called **Probability as a continuous set function**. Hint: use the Axiom 3 of probability measure.

10 Conditional Distributions and Independence

Definition 30 *Let (X,Y) be discrete bivariate random vector with joint pmf $f_{X,Y}(x,y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) = P(X = x) > 0$, the conditional pmf of Y given $X = x$ is the function of y denoted by $f(y|x)$ and defined by*

$$f(y|x) = P(Y = y|X = x) = \frac{f(x,y)}{f_X(x)}.$$

For any y such that $f_Y(y) = P(Y = y) > 0$, the conditional pmf of X given $Y = y$ is the function of x denoted by $f(x|y)$ and defined by

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

Exercise: verify that conditional pmf is indeed a pmf.

Definition 31 *Let (X,Y) be a continuous bivariate random vector with joint pdf $f_{X,Y}(x,y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the conditional pdf of Y given $X = x$ is the function of y denoted by $f(y|x)$ and defined by*

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

For any y such that $f_Y(y) = P(Y = y) > 0$, the conditional pdf of X given $Y = y$ is the function of x denoted by $f(x|y)$ and defined by

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

11 Independence

Definition 32 *Let (X, Y) be a bivariate random vector with joint pdf or pmf $f_{X,Y}(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then X and Y are called independent random variables if, for every $x \in \mathbf{R}$ and every $y \in \mathbf{R}$,*

$$f(x, y) = f_X(x) f_Y(y).$$

If X and Y are independent, the conditional pdf of Y given $X = x$ is:

$$f(y|x) \quad := \quad \frac{f(x, y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y),$$

regardless of the value of x. Thus for any $A \subset \mathbf{R}$ and $x \in \mathbf{R}$,

$$P(Y \in A|x) = \int_A f(y|x)dy = \int_A f_Y(y)dy = P(Y \in A)$$

where for the first equality, we used the fact that conditional pdf is indeed a pdf. (Exercise: show this is true!)

Theorem 33 (CB Lemma 4.2.7) *Let (X, Y) be a bivariate random vector with $f_{XY}(x, y)$. X and Y are independent iif there exists functions g, h such that*

$$f_{XY}(x, y) = g(x)h(y).$$

12 Conditional Expectation

Example 34 Calculating conditional pdfs *Consider the following joint pdf for continuous random vector (X, Y) defined by*

$$f(x, y) = e^{-y} \quad 0 < x < y < \infty$$

Suppose we wish to compute the conditional pdf of Y given $X = x$. The marginal pdf of X is computed as follows: for $x > 0$, $f(x, y) > 0$ only for $\infty > y > x$, and thus

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

Thus marginally, X has an exponential distribution. Now

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = e^{-y}/e^{-x} = e^{-(y-x)} \quad \text{if } y > x > 0$$

and

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = 0/e^{-x} = 0 \quad \text{if } y \leq x \text{ and } x > 0$$

Now

$$E(Y|X=x) = \int_x^\infty yf(y|x)dy = \int_x^\infty ye^{-(y-x)}dy = 1+x$$

and

$$\mathrm{Var}(Y|X=x) = \int_x^\infty y^2 f(y|x)dy - \left(\int_x^\infty yf(y|x)dy\right)^2 = 1$$

Here are more questions:

1. *Compute* $\mathrm{Var}(Y)$.
2. *Convince yourself* $Y = X + Z$ *for some random variable* Z *independent of* X. *What is the distribution of* Z?
3. *What is the marginal distribution of* Y?

Definition 35 *The* **conditional expected value** *of* $g(Y)$ *given that* X *is denoted by*

- *Discrete RV* $E\left(g(Y)|x\right) = \sum_y g(y)f(y|x)$.
- *Continuous RV:* $E\left(g(Y)|x\right) = \int_{-\infty}^{\infty} g(y)f(y|x)dy$.

Warning! $E\left(Y|x\right)$ is a function of x, while $E(Y)$ is a constant. (Diagram).

Definition 36 *The conditional variance of the probability distribution described by* $f(x|y)$ *is:*

$$\mathrm{Var}(X|Y=y) = \int_{-\infty}^{\infty}(x-\mu(y))^2 f(x|y)dx$$

where $\mu(y) = E(X|Y=y)$. *We also use notation*

$$\mathrm{Var}(X|Y) := E\left[(X-E(X|Y))^2|Y\right].$$

Theorem 37 *(A law of total expectation)* $E(Y) = E(E(Y|X))$.

Proof. Using the definition of conditional expectation, we have for $E\left(Y|x\right) = h(x)$

$$\begin{aligned} E(E(Y|X)) = E(h(x)) &= \int_{-\infty}^{\infty} E(Y|X=x)f_X(x)dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf_{Y|X}(y|x)dyf_X(x)dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf_{Y|X}(y|x)f_X(x)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf_{X,Y}(x,y)dxdy \\ &= E(Y) \end{aligned}$$

■

Theorem 38 *(Conditional variance identity) For any two random variables X and Y,*

$$\mathrm{Var}(X) = E(\mathrm{Var}(X|Y)) + \mathrm{Var}(E(X|Y))$$

provided that the expectations exists.

PROOF. By definition, we have

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

where we have added and subtracted $E(X|Y)$ and expanded the square in the second line. Now the first term on the RHS of the last line is:

$$\begin{aligned}E(X - E(X|Y))^2 &= E(E\left\{(X - E(X|Y))^2|Y\right\}) \\ &= E(\mathrm{Var}(X|Y)),\end{aligned}$$

by the law of total expectations and by definition of conditional variance; And the second term on the RHS of the last line is:

$$\begin{aligned}E(E(X|Y) - E(X))^2 &= E\left\{E(X|Y) - E(E(X|Y))\right\}^2 \\ &= \mathrm{Var}(E(X|Y)).\end{aligned}$$

Now we show that the middle term equals 0 by using the law of total expectations,

$$E\left\{(X - E(X|Y))(E(X|Y) - E(X))\right\} = EE\left[\left\{(X - E(X|Y))(E(X|Y) - E(X))\right\}|Y\right]$$

where

$$\begin{aligned}E\left[\left\{(X - E(X|Y))(E(X|Y) - E(X))\right\}|Y\right] &= (E(X|Y) - E(X))E\left[(X - E(X|Y))|Y\right] \\ &= (E(X|Y) - E(X))(E(X|Y) - E(X|Y)) = 0\end{aligned}$$

The theorem thus holds. ■

Example 39 *Let's continue to develop Example 34 using Theorem 38. Now let's apply*

$$\begin{aligned}\mathrm{Var}(Y) &= E(\mathrm{Var}(Y|X)) + \mathrm{Var}(E(Y|X)) \\ &= 1 + \mathrm{Var}(1 + X) = 1 + \mathrm{Var}(X) = 2\end{aligned}$$

where

$$\mathrm{Var}(X) = \int_0^\infty x^2 e^{-x} dx - \left(\int_0^\infty x e^{-x} dx\right)^2 = 2 - 1 = 1$$

Given the knowledge of $X = x$, the variability in Y is considerably reduced. It is easy to verify that

$$\int_x^\infty f(y|x)dy = \int_x^\infty e^{-(y-x)} dy = 1,$$

a property which we know holds for conditional densities.

13 Covariance and Correlation

Theorem 40 *Assuming the variance is well defined. That is, we assume* $0 < \mathrm{Var}(X_i) < \infty$ *for all* i*. It has the following properties:*

1. $\mathrm{Var}(X) = EX^2 - \mu^2$ *where* $\mu = EX$*.*
2. *If* a *and* b *are constants, then* $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$*.*
3. *If random variables* $X_1, \ldots, X_n$ *are independent and* $a_1, \ldots, a_n$ *are constants, then*

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i).$$

Theorem 41 *The covariance satisfies:*

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

The correlation $\rho(X, Y) = \mathrm{Cov}(X, Y)/\sigma_x\sigma_y$ *satisfies*

$$-1 \leq \rho(X, Y) \leq 1. \tag{4}$$

If $Y = aX + b$ *for some constants* a *and* b *then* $\rho(X, Y) = 1$ *if* $a > 0$ *and* $\rho(X, Y) = -1$ *if* $a < 0$*. If* X *and* Y *are independent, then* $\mathrm{Cov}(X, Y) = \rho = 0$*. The converse is not true in general.*

Proof. We have by definition of $\mathrm{Cov}(X, Y)$,

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E(X - \mu_x)(Y - \mu_y) \\ &= E(XY) - \mu_X E(Y) - E(X)\mu_y + \mu_X\mu_Y \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

Proof for (4) is omitted for now. Now for $Y = aX + b$, we have

$$\begin{aligned} \rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} &= \frac{E(X(aX + b)) - E(X)E(aX + b)}{|a|\mathrm{Var}(X)} \\ &= \frac{aE(X^2) + bE(X) - aE(X)^2 - bE(X)}{|a|\mathrm{Var}(X)} \\ &= a/|a| \end{aligned}$$

■

Theorem 42 $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$ *and* $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y)$*. More generally, for random variables* $X_1, \ldots, X_n$

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j>i}^{n} a_i a_j \mathrm{Cov}(X_i, X_j).$$

Exercises: Prove Theorem 42.

14 Review on Independence

Lemma 43 *Let (X, Y) be a bivariate random vector with $f_{XY}(x, y)$. X and Y are independent iif there exists functions g, h such that*

$$f_{XY}(x, y) = g(x)h(y).$$

Proof.

1. If X and Y are independent then $f_{XY}(x, y) = g(x)h(y)$.
2. Assume $f_{XY}(x, y) = g(x)h(y)$. Define $\int g(x)dx = G$ and $\int h(y)dy = H$. Then

$$f_X(x) = \int f(x, y)dy = \int g(x)h(y)dy = g(x) \int h(y)dy = g(x)H.$$

Similarly

$$f_Y(y) = h(y) \int g(x)dx = h(y)G.$$

Therefore

$$f_X(x)f_Y(y) = g(x)h(y)GH.$$

But $HG = 1$ because

$$\int \int f_{XY}(x, y)dxdy = \int \int g(x)h(y)dxdy = \int g(x)dx \int h(y)dy = GH = 1.$$

Consequently, $f_{XY}(x, y) = g(x)h(y) = f_X(x)f_Y(y)$, and the RV are independent.

■

Theorem 44 (Theorem 4.2.10, CB) *Let X and Y be independent random variables.*

- *(a) For any $A \subset \mathbf{R}$ and $B \subset \mathbf{R}$, $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$; that is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.*
- *(b) Let $g(x)$ be a function of only x and $h(y)$ be a function of only y. Then*

$$E(g(X)h(Y)) = Eg(X)Eh(Y)$$

15 Functions of Jointly Distributed Random Variables

CDF method for transformations of several random variables: The cdf method can also be used for either discrete or continuous random variables that map many random variables into one, such as $Z = X + Y$ or $Z = X/Y$.

Sums, $Z = X + Y$. Show

$$\text{for discrete RVs, } f_Z(z) = \sum_{x=-\infty}^{\infty} f(x, z-x)$$

$$\text{for continuous RVs, } f_Z(z) = \int_{-\infty}^{\infty} f(x, z-x)dx \text{ if } \int_{-\infty}^{\infty} f(x, z-x)dx \text{ is continuous at } z$$

PROOF. ON BOARD

When X and Y are independent,

- $f_Z(z) = \sum_{x=-\infty}^{\infty} f_X(x) f_Y(z-x)$ for discrete random variables. This sum is called the **convolution** of the sequences f_X and f_Y.
- $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx$ for continuous random variables. This integral is called the **convolution** of functions f_X and f_Y.

Example A Let T_1 and T_2 be independent exponentials with parameter β. Let $S = T_1 + T_2$. Then $S \sim \Gamma(2, \beta)$.

Example B $Z = Y/X$ Let $X, Y \sim N(0,1)$ be independent random variables. Show that

$$f_Z(z) = \frac{1}{\pi(z^2+1)}, \quad -\infty < z < \infty.$$

This density is called **Cauchy density**. The tails of the Cauchy tend to zero very slowly compared to the tails of the normal.

Let's summarize the CDF method for transformations of several variables. Let $Z = g(X, Y)$ be the function of interest.

1. For each z, find the set $A_z = \{(x, y) : g(x, y) \leq z\}$.
2. Find the CDF

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(g(X,Y) \leq z) \\ &= P(\{(x,y) : g(x,y) \leq z\}) \\ &= \int\int_{A_z} f_{X,Y}(x,y)dxdy \end{aligned}$$

3. The PDF $f_Z(z) = F_Z'(z)$ at z where $f_Z(z)$ is continuous.

The last step applies only for continuous rvs.

16 Transformation of 2 Continuous RV's

Let (X, Y) be a bivariate random vector with a known probability distribution. Now consider

$$U = g_1(X, Y) \quad V = g_2(X, Y)$$

Let us define the following.

- $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$.
- Define $\mathcal{B} = \{(u, v) \ : \ u = g_1(x, y) \text{ and } \ v = g_2(x, y) \text{ for some } \ (x, y) \in \mathcal{A}\}$.
- Assume g_1, g_2 defines a one-to-one transformation of $\mathcal{A}$ onto $\mathcal{B}$.
- The transformation is onto because of the definition of $\mathcal{B}$.
- We are assuming that for each $(u, v) \in \mathcal{B}$, there is only one $(x, y) \in \mathcal{A}$ such that $(u, v) = (g_1(x, y), g_2(x, y))$.
- For such a one-to-to, onto transformation,we can solve equations $u = g_1(x, y)$ and $v = g_2(x, y)$ to obtain x, y in terms of u and v.

We denote the **inverse transformation** by

$$x = h_1(u, v) \text{ and } \ y = h_2(u, v)$$

Thus we have

$$\begin{aligned}(g_1, g_2) : \mathcal{A} &\rightarrow \mathcal{B} \\ (h_1, h_2) : \mathcal{B} &\rightarrow \mathcal{A}\end{aligned}$$

Now the function of (u, v) denoted by J is the determinant of a matrix of partial derivatives:

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v}$$

where

$$\frac{\partial x}{\partial u} = \frac{\partial h_1(u, v)}{\partial u}, \ \frac{\partial x}{\partial v} = \frac{\partial h_1(u, v)}{\partial v}, \ \frac{\partial y}{\partial u} = \frac{\partial h_2(u, v)}{\partial u}, \ \text{ and } \frac{\partial y}{\partial v} = \frac{\partial h_2(u, v)}{\partial v}.$$

We assume J is not identically 0 on $\mathcal{B}$. The joint pdf of (U, V) is 0 outside of the set $\mathcal{B}$ and on the set $\mathcal{B}$ is given by: for $|J|$ being the absolute value of J,

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v)) \cdot |J|.$$

It is simply an adjustment of scale when you make a change of variable to calculate an integral. Remember it as

$$\left| \frac{\partial \text{ old}}{\partial \text{ new}} \right|.$$

In general, let $Y = (Y_1, \ldots, Y_n)$ and $X = (X_1, \ldots, X_n)$ then

$$f_Y(y) = f_X(x(y)) \left| \partial x / \partial y \right|.$$

Example 45 *Sum and difference of standard normal RVs Let X, Y be independent $N(0,1)$ RVs. Consider $U = X + Y$ and $V = X - Y$. Show U and V are independent using Theorem 33.*

Notes on transformation: For example, think of changing a point (x, y) in R^2 to their polar coordinates (r, θ) where

$$\begin{aligned} r &= \sqrt{x^2 + y^2} \\ \theta &= \tan^{-1}(y/x). \end{aligned}$$

Now the inverse transformation is

$$\begin{aligned} x &= r\cos(\theta) \\ y &= r\sin(\theta). \end{aligned}$$

Example (CB 4.3.3). Distribution of the product of Beta-variables.

17 Review on Expectation

We did reviews on expectations, conditional expectations, and expectations for a function. Please work out the discrete cases. Recall the following for continuous RVs.

- $E\left(g(X)\right) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$.
- $E\left(g(X,Y)\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f(x,y) dx dy$.
- $E\left(g(Y)|x\right) = \int_{-\infty}^{\infty} g(y) f(y|x) dy$.
- $\mathrm{Var}(X|Y = y) = \int_{-\infty}^{\infty}(x - \mu(y))^2 f(x|y) dx$ where $\mu(y) = E(X|Y = y)$. We also use notation
$$\mathrm{Var}(X|Y) := E\left[(X - E(X|Y))^2|Y\right].$$

18 Moment generating function

Definition 46 *Let X be a RV with cdf F_X. The* **Moment generating function (mgf)** *of X (or F_X), denoted by $M_X(t)$ is*

$$M_X(t) = E\left(e^{tX}\right),$$

provided that the expectation exists for t in some neighborhood of 0. That is, there is some $h > 0$ such that for all t in $-h < t < h$, $E\left(e^{tX}\right)$ exists.

Theorem 47 *For any sum of independent random variables $Y = X_1 + X_2$,*

$$M_Y(t) = E\left(e^{tY}\right) = M_{X_1}(t)M_{X_2}(t).$$

PROOF. We apply Theorem 44 with $g(X_1) = e^{tX_1}$ and $h(X_2) = e^{tX_2}$ to obtain:

$$M_Y(t) = E\left(e^{tY}\right) = E\left(e^{tX_1} \cdot e^{tX_2}\right) = E\left(e^{tX_1}\right) E\left(e^{tX_2}\right).$$

Thus the theorem holds by definition. ■

19 Transformation of 2 discrete RV's

Let (X, Y) be a bivariate random vector with a known probability distribution. Now consider

$$U = g_1(X, Y) \quad V = g_2(X, Y)$$

Recall the following fact: Let (X, Y) be discrete bivariate random vector. Then $f(x, y)$ is nonzero for at most a countable set of values (x, y). Call this set $\mathcal{A}$.

- $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$.
- Define $\mathcal{B} = \{(u, v) \ : \ u = g_1(x, y) \text{ and } \ v = g_2(x, y) \text{ for some } \ (x, y) \in \mathcal{A}\}$.
- For any $(u, v) \in \mathcal{B}$, define

$$A_{u,v} = \{(x, y) \in \mathcal{A} \ : \ g_1(x, y) = u \text{ and } \ g_2(x, y) = v\}.$$

- The the joint pmf of (U, V) can be computed from the joint pmf of (X, Y) by

$$f_{U,V}(u, v) = P(U = u, V = v) = P((X, Y) \in A_{u,v}) = \sum_{(x,y)\in A_{uv}} f_{X,Y}(x, y).$$

Theorem 48 *(Theorem 4.3.2, CB) if $X \sim$ Poisson(λ) and if $Y \sim$ Poisson(μ) and X and Y are independent, then $X + Y \sim$ Poisson($\lambda + \mu$)*

PROOF. ON BOARD

Example 49 *We compute the density for $U = X + Y$ for X, Y as defined in Theorem 48 using this method by defining*

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

We use this trick a couple of times in this class.

Example 50 *Use Theorem 47, we can show that the conclusion of Theorem 48 hold by computing the mgf for $U = X_1 + X_2$ where$X_1 \sim$ Poisson (λ_1) and $X_2 \sim$ Poisson (λ_2), we have $E\left(e^{tX_1}\right) E\left(e^{tX_2}\right) = e^{(\lambda_1+\lambda_2)(e^t-1)}$, which is the mgf for the desired Poisson $(\lambda_1 + \lambda_2)$ distribution.*

20 Large deviation bounds: Chapter 3

Inequalities are useful for bounding quantities that might otherwise be hard to compute. They will also be used in the theory of convergence. Our first inequality is Markov's inequality.

Theorem 51 (Markov's inequality) *Let X be a non-negative random variable and suppose that $\mathbb{E}(X)$ exists. For any $t > 0$,*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}.$$

PROOF. Since $X > 0$,

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty x f(x)dx = \int_0^t x f(x)dx + \int_t^\infty x f(x)dx \\ &\geq \int_t^\infty x f(x)dx \geq t \int_t^\infty f(x)dx = t\,\mathbb{P}(X > t) \quad \blacksquare \end{aligned}$$

Theorem 52 (Chebyshev's inequality) *Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$. Then,*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \text{and} \quad \mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2}$$

where $Z = (X - \mu)/\sigma$. In particular, $\mathbb{P}(|Z| > 2) \leq 1/4$ and $\mathbb{P}(|Z| > 3) \leq 1/9$.

PROOF. We use Markov's inequality to conclude that

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^2 \geq t^2) \leq \frac{\mathbb{E}(X - \mu)^2}{t^2} = \frac{\sigma^2}{t^2}.$$

The second part follows by setting $t = k\sigma$. $\blacksquare$

21 Location and Scale Families

Let $f(x)$ be a pdf.

$$\text{Location family}: \ \{f(x|\mu) = f(x - \mu): \ \mu \in \mathbb{R}\}$$

$$\text{Scale family}: \ \left\{f(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right): \ \sigma > 0\right\}$$

$$\text{Location} - \text{Scale family}: \ \left\{f(x|\mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right): \ \mu \in \mathbb{R}, \sigma > 0\right\}$$

(1) Location family. Shifts the pdf.

e.g., Uniform with $f(x) = 1$ on $(0, 1)$ and $f(x - \theta) = 1$ on $(\theta, \theta + 1)$.

e.g., Normal with standard pdf the density of a $N(0, 1)$ and location family pdf $N(\theta, 1)$.

(2) Scale family. Stretches the pdf.

e.g., Normal with standard pdf the density of a $N(0, 1)$ and scale family pdf $N(0, \sigma^2)$.

(3) Location-Scale family. Stretches and shifts the pdf.

e.g., Normal with standard pdf the density of a $N(0, 1)$ and location-scale family pdf $N(\theta, \sigma^2)$, i.e., $\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$.

Example 53 *Let X_1, X_2 be independent, and both uniformly distributed on the interval $[0, \theta]$, with $\theta > 0$. Define $Y = X_1 + X_2$. Show the density of Y.*
Hint: *You can do whichever way you like, but you can think of $Y = \theta Z$, where $Z = X_3 + X_4$, where X_3, X_4 are independent, and both uniformly distributed on the interval $[0, 1]$, which you have seen in your homework. See solution to the first midterm.*

Example 54 *Let X and Z be continuous random variables. Let X have a uniform$(0, 1)$ distribution and Z have a uniform$(0, 1/10)$ distribution. Suppose X and Z are independent. Let $Y = X + Z$ and consider the random vector (X, Y). Show that the joint pdf for (X, Y) is*

$$f_{XY}(x, y) = 10, \;\; 0 < x < 1, \; x < y < x + 1/10.$$

Hint: *We used multiple methods to prove this in class. Please review these by yourself.*

Theorem 55 *(CB Theorem 3.5.6). Let $f(x)$ be any pdf. Let μ be any real number, and let $\sigma > 0$ be any positive real number. Then X is a random variable with pdf*

$$\frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right) : \; \mu \in \mathbb{R}, \; \sigma > 0$$

if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.

Exercise. Prove this theorem using the transformation theorem you have learned.

Hierarchical Models

See pp 162-168 in CB.

- A way to model a complicated process in stages.

- Based on law of total probability in Lecture 1, for a partition $B_1, B_2, \ldots$ of the sample space,

$$P(A) = \sum_k P(A|B_k)P(B_k).$$

Thus we have for X discrete

$$f_Y(y) = \sum_x P(Y = y|X = x)P(X = x) = \sum_x f_{Y|X}(y|x)P(X = x)$$

or for X continuous

$$f_Y(y) = \int f(x, y)dx = \int_x f_{Y|X}(y|x)f_X(x)dx.$$

Example 56 (Example 4.4.1. Binomial-Poisson hierarchy, CB 4.4)

1. *X = number of espressos sold per day.*
2. *N = the number of customers per day.*
3. *Let $p = P(espresso)$, that is, the probability that a customer buys an espresso.*
4. *If we assume*

$$N \sim Poisson(\lambda)$$

$$X|N = n \sim Bin(n, p)$$

Then it is easy to compute express $P(X = x)$ as a function of the 2 parameters (λ, p). We concluded that $X \sim Poisson(\lambda p)$. Now how many espressos we sell in expectation per day?

5. *Thus far in this example we have assumed we know parameters, but for Bayesian inference we would might also have a model for p, such as $p \sim beta(a, b)$.*