1 Probability measure and random variables

1.1 Probability spaces and measures

We will use the term *experiment* in a very general way to refer to some process that produces a random outcome.

Definition 1. *The set of possible outcomes is called the* **sample space**. *We will typically denote an individual outcome by* ω *and the sample space by* Ω.

Set notation: $A \subset B$, A is a subset of B, means that every element of A is also in B. The union $A \cup B$ of A and B is the of all elements that are in A or B, including those that are in both. The intersection $A \cap B$ of A and B is the set of all elements that are in both of A and B.

$\cup_{j=1}^{n} A_j$ is the set of elements that are in at least one of the A_j.

$\cap_{j=1}^{n} A_j$ is the set of elements that are in all of the A_j.

$\cap_{j=1}^{\infty} A_j, \quad \cup_{j=1}^{\infty} A_j$ are ...

Two sets A and B are **disjoint** if $A \cap B = \emptyset$. $\emptyset$ denotes the empty set, the set with no elements.

Complements: The **complement** of an event A, denoted A^c, is the set of outcomes (in Ω) which are not in A. Note that the book writes it as $\Omega \setminus A$.

De Morgan's laws:

$$\begin{aligned} (A \cup B)^c &= A^c \cap B^c \\ (A \cap B)^c &= A^c \cup B^c \\ (\bigcup_j A_j)^c &= \bigcap_j A_j^c \\ (\bigcap_j A_j)^c &= \bigcup_j A_j^c \end{aligned} \tag{1}$$

Definition 2. *Let* Ω *be a sample space. A collection* $\mathcal{F}$ *of subsets of* Ω *is a* σ*-field if*

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_n \in \mathcal{F}$ *for* $n = 1, 2, 3, \cdots \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

Definition 3. *Let $\mathcal{F}$ be a σ-field of events in Ω. A probability measure on $\mathcal{F}$ is a real-valued function $\mathbf{P}$ on $\mathcal{F}$ with the following properties.*

1. $\mathbf{P}(A) \geq 0$, *for* $A \in \mathcal{F}$.
2. $\mathbf{P}(\Omega) = 1$, $\mathbf{P}(\emptyset) = 0$.
3. *If $A_n \in \mathcal{F}$ is a disjoint sequence of events, i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$, then*

$$\mathbf{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbf{P}(A_n) \tag{2}$$

We refer to the triple $(\Omega, \mathcal{F}, \mathbf{P})$ as a *probability space.*

Theorem 1. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space.*

1. $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$ *for* $A \in \mathcal{F}$.
2. $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$ *for* $A, B \in \mathcal{F}$.
3. $\mathbf{P}(A \setminus B) = \mathbf{P}(A) - \mathbf{P}(A \cap B)$. *for* $A, B \in \mathcal{F}$.
4. *If $A \subset B$, then* $\mathbf{P}(A) \leq \mathbf{P}(B)$. *for* $A, B \in \mathcal{F}$.
5. *If $A_1, A_2, \cdots, A_n \in \mathcal{F}$ are disjoint, then*

$$\mathbf{P}(\bigcup_{j=1}^{n} A_j) = \sum_{j=1}^{n} \mathbf{P}(A_j) \tag{3}$$

1.2 Conditional probability and independent events

Definition 4. *If A and B are events and $\mathbf{P}(B) > 0$, then the (conditional) probability of A given B is denoted $\mathbf{P}(A|B)$ and is given by*

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \tag{4}$$

Theorem 2. *Let $\mathbf{P}$ be a probability measure, B an event ($B \in \mathcal{F}$) with $\mathbf{P}(B) > 0$. For events A ($A \in \mathcal{F}$), define*

$$Q(A) = \mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$$

Then Q is a probability measure on $(\Omega, \mathcal{F})$.

We can rewrite the definition of conditional probability as

$$\mathbf{P}(A \cap B) = \mathbf{P}(A|B)\mathbf{P}(B) \tag{5}$$

In some experiments the nature of the experiment means that we know certain conditional probabilities. If we know $\mathbf{P}(A|B)$ and $\mathbf{P}(B)$, then we can use the above to compute $\mathbf{P}(A \cap B)$.

In general $\mathbf{P}(A|B) \neq \mathbf{P}(A)$. Knowing that B happens changes the probability that A happens. But sometimes it does not. Using the definition of $\mathbf{P}(A|B)$, if $\mathbf{P}(A) = \mathbf{P}(A|B)$ then

$$\mathbf{P}(A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \tag{6}$$

i.e., $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. This motivate the following definition.

Definition 5. *Two events are independent if*

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) \tag{7}$$

Theorem 3. *If A and B are independent events, then*

1. *A and B^c are independent*
2. *A^c and B are independent*
3. *A^c and B^c are independent*

In general A and A^c are not independent.

The notion of independence can be extended to more than two events.

Definition 6. *Let $A_1, A_2, \cdots, A_n$ be events. They are independent if for all subsets I of $\{1, 2, \cdots, n\}$ we have*

$$\mathbf{P}(\cap_{i \in I} A_i) = \prod_{i \in I} \mathbf{P}(A_i) \tag{8}$$

They are just pairwise independent if $\mathbf{P}(A_i \cap A_j) = \mathbf{P}(A_i)\mathbf{P}(A_j)$ for $1 \leq i < j \leq n$.

1.3 The partition theorem and Bayes theorem

Definition 7. *A partition is a finite or countable collection of events B_j such that $\Omega = \cup_j B_j$ and the B_j are disjoint, i.e., $B_i \cap B_j = \emptyset$ for $i \neq j$.*

Theorem 4. *(Partition theorem) Let $\{B_j\}$ be a partition of Ω. Then for any event A,*

$$\mathbf{P}(A) = \sum_j \mathbf{P}(A|B_j)\,\mathbf{P}(B_j) \tag{9}$$

Bayes theorem deals with the situation where we know all the $\mathbf{P}(A|B_j)$ and want to compute $\mathbf{P}(B_i|A)$.

Theorem 5. *(Bayes theorem) Let $\{B_j\}$ be a partition of Ω. Then for any event A and any k,*

$$\mathbf{P}(B_k|A) = \frac{\mathbf{P}(A|B_k)\,\mathbf{P}(B_k)}{\sum_j \mathbf{P}(A|B_j)\,\mathbf{P}(B_j)} \tag{10}$$

1.4 Random Variables and their distribution

A random variable is a function X from Ω to the real numbers. It must be measurable, meaning that for all Borel subsets B of the real line, $X^{-1}(B)$ must belong to $\mathcal{F}$. In this course RV's will come in two flavors - discrete and continuous. We will not worry about measurability.

We can consider functions from Ω into other spaces. A function that maps to $\mathbb{R}^n$ is called a random vector. More general range spaces are possible, for example, X could map into a metric space S. S could be a set of graphs, in which case we would call X a graph valued random variable.

Important idea: The sample space Ω may be quite large and complicated. But we may only be interested in one or a few RV's. We would like to be able to extract all the information in the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ that is relevant to our random variable(s), and forget about the rest of the information contained in the probability space. We do this by defining the distribution of a random variable. The distribution measure of X is the Borel measure μ_X on the real line given by $\mu_X(B) = \mathbf{P}(X \in B)$. We can also specify the distribution by the cumulative distribution function (CDF). This is the function on the real line defined by $F(x) = \mathbf{P}(X \leq x)$. If we want to make it clear which RV we are talking about, we write it as $F_X(x)$.

Theorem 6. *For any random variable the CDF satisfies*

1. *$F(x)$ is non-decreasing, $0 \leq F(x) \leq 1$.*
2. *$\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$.*
3. *$F(x)$ is continuous from the right.*

Theorem 7. *Let $F(x)$ be a function from $\mathbb{R}$ to $[0,1]$ such that*

1. *$F(x)$ is non-decreasing.*
2. *$\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$.*
3. *$F(x)$ is continuous from the right.*

Then $F(x)$ is the CDF of some random variable, i.e., there is a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a random variable X on it such that $F(x) = \mathbf{P}(X \leq x)$.

Another important idea: Suppose we have two completely different probability spaces $(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbf{P}_2)$, and RV's X_1 on the first and X_2 on the second. Then it is possible that X_1 and X_2 have the same distribution measure, i.e., for any Borel set B $\mathbf{P}(X_1 \in B) = \mathbf{P}(X_2 \in B)$. If we only look at X_1 and X_2 when we do the two experiments, then we won't be able to tell the experiments apart. When the two random variables have the same distribution measure we say that X_1 and X_2 are *identically distributed*.

1.5 Expected value

Given a random variable X, the expected value or mean of X is

$$E[X] = \int X \, dP \tag{11}$$

The integral is over ω and the definition requires abstract integration theory. But as we will see, in the case of discrete and continuous random variables this abstract integral is equal to either a sum or a calculus type integral on the real line. Here are some trivial properties of the expected value.

Theorem 8. *Let X, Y be discrete RV's with finite mean. Let $a, b \in \mathbb{R}$.*

1. $\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y]$
2. *If $X = b$, b a constant, then* $\mathbf{E}[X] = b$.
3. *If* $\mathbf{P}(a \leq X \leq b) = 1$, *then* $a \leq \mathbf{E}[X] \leq b$.
4. *If $g(X)$ and $h(X)$ have finite mean, then* $\mathbf{E}[g(X) + h(X)] = \mathbf{E}[g(X)] + \mathbf{E}[h(X)]$

The next theorem is not trivial; it is extremely useful if you want to actually compute an expected value.

Theorem 9. *(change of variables or transformation theorem) Let X be a random variable, μ_X its distribution measure. Let $g(x)$ be a Borel measurable function from $\mathbb{R}$ to $\mathbb{R}$. Then*

$$E[g(X)] = \int_{\mathbb{R}} g(x)\, d\mu_X \tag{12}$$

provided

$$\int_{\mathbb{R}} |g(x)|\, d\mu_X < \infty \tag{13}$$

In particular

$$E[X] = \int x\, d\mu_X(x) \tag{14}$$

Definition 8. *The variance of X is*

$$var(X) = \mathbf{E}[(X - \mu)^2]$$

where $\mu = \mathbf{E}X$. The standard deviation of X is $\sqrt{var(X)}$. The variance is often denoted σ^2 and the standard deviation by σ. The mean of X, i.e., $E[X]$ is also called the first moment of X. The kth moment of X is $\mathbf{E}[X^k]$.

Proposition 1. *If X has finite first and second moments, then*

$$var(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

and

$$var(aX + b) = a^2 var(X), \quad a, b \in \mathbb{R}$$

1.6 Joint distributions

If $X_1, X_2, \cdots, X_n$ are RV's, then their joint distribution measure is the Borel measure on $\mathbb{R}^n$ given by

$$\mu_{X_1,X_2,\cdots,X_n}(B) = \mathbf{P}((X_1, X_2, \cdots, X_n) \in B) \tag{15}$$

where B is a Borel subset of $\mathbb{R}^n$.

Definition 9. *Random variables $X_1, X_2, \cdots, X_n$ are independent if their joint distribution measure is the product of their individual distribution measures, i.e.,*

$$\mu_{X_1,X_2,\cdots,X_n} = \mu_{X_1} \times \mu_{X_2} \times \cdots \times \mu_{X_n} \tag{16}$$

This is equivalent to saying that

$$\mathbf{P}(X_1 \in B_1, X_2 \in B_2, \cdots, X_n \in B_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \in B_i) \tag{17}$$

for all Borel sets B_i in $\mathbb{R}$. An infinite set of random variables is independent if every finite subset of them is independent.

Theorem 10. *Let $X_1, X_2, \cdots, X_n$ be independent. Then*

$$E[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} E[X_i] \tag{18}$$

Furthermore, if $g_i : \mathbb{R} \to \mathbb{R}$ are measurable, then $g_1(X_1), g_2(X_2), \cdots, g_n(X_n)$ are independent and so

$$E[\prod_{i=1}^{n} g_i(X_i)] = \prod_{i=1}^{n} E[g_i(X_i)] \tag{19}$$

(We are omitting some hypothesis that insure things are finite.)

Corollary 1. *If $X_1, X_2, \cdots X_n$ are independent and have finite variances, then*

$$var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} var(X_i) \tag{20}$$

There is a generalization of the change of variables theorem:

Theorem 11. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be measurable. Let $X_1, \cdots, X_n$ be random variables. Then*

$$E[g(X_1, \cdots, X_n)] = \int_{\mathbb{R}^n} g(x_1, x_2, \cdots, x_n)\, d\mu_{X_1,\cdots,X_n} \tag{21}$$

(We are omitting a hypothesis that insures the integrals involved are defined.)

2 Discrete Random Variables

2.1 Probability mass function

A random variable is said to be discrete if its range is finite or countably infinite.

Definition 10. *The probability mass function (pmf) $f(x)$ of a discrete RV X is the function on $\mathbb{R}$ given by*

$$f(x) = \mathbf{P}(X = x)$$

For discrete RV's the distribution measure μ_X is just a sum of point masses

$$\mu_X = \sum_x f(x)\delta_x \tag{22}$$

Here δ_x denotes the measure such that $\delta_x(B)$ is 1 if $x \in B$ and is 0 if $x \notin B$.

Notation/terminology: If we have more than one RV, then we have more than one pmf. To distinguish them we use $f_X(x)$ for the pmf for X, $f_Y(x)$ for the pmf for Y, etc. Sometimes the pmf is called the "density function" and sometimes the "distribution of X." The latter can be confusing as the term "distribution function" usually means the cumulative distribution function. For a discrete RV the change of variables theorem becomes

Theorem 12. *Let X be a discrete RV with probability mass function $f_X(x)$. Then the expected value of X is given by*

$$\mathbf{E}[X] = \sum_x x f_X(x)$$

If $g : \mathbb{R} \to \mathbb{R}$, then

$$\mathbf{E}[g(X)] = \sum_x g(x) f_X(x)$$

(We are omitting some hypothesis that insure things are finite.)

2.2 Discrete RV's - catalog

Bernoulli RV (one parameter $p \in [0, 1]$) This is about as simple as they get. The RV X only takes on the values 0 and 1.

$$p = \mathbf{P}(X = 1), \quad 1 - p = \mathbf{P}(X = 0)$$

We can think of this as coming from a coin with probability p of heads. We flip it only once, and $X = 1$ corresponds to heads, $X = 0$ to tails. It is standard to refer to the outcome with $X = 1$ as success and the outcome with $X = 0$ as failure.

Binomial RV (two parameters: $p \in [0, 1]$, positive integer n) The range of the random variable X is $0, 1, 2, \cdots, n$.

$$\mathbf{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Think of flipping an unfair coin n times. p is the probability of heads on a single flip and X is the number of head we get out of the n flips. The parameter n is often called the "number of trials."

Poisson RV (one parameter: $\lambda > 0$) The range of the random variable X is $0, 1, 2, \cdots$.

$$\mathbf{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

There is no simple experiment that produces a Poisson random variable. But it is a limiting case of the binomial distribution and it occurs frequently in applications.

Geometric (one parameter: $p \in [0, 1]$) The range of the random variable X is $1, 2, \cdots$.

$$\mathbf{P}(X = k) = p(1 - p)^{k-1} \tag{23}$$

Think of flipping an unfair coin with p being the probability of heads until we gets heads for the first time. Then X is the number of flips (including the flip that gave heads.)

Caution: Some books use a different convention and take X to be the number of tails we get before the first heads. In that case $X = 0, 1, 2, ...$ and the pmf is different.

2.3 Conditional expectation

Fix an event B. If we define a function Q on events by $Q(A) = \mathbf{P}(A|B)$, then this defines a new probability measure. So if we have a RV X, then we can consider its probability mass function with respect to the probability measure Q. And so we can compute its expected value with repect to this new pmf. This is called the conditional expectation of X given B. The formal definition follows.

Definition 11. *Let X be a discrete RV. Let B be an event with $\mathbf{P}(B) > 0$. The conditional probability mass function of X given B is*

$$f(x|B) = \mathbf{P}(X = x|B)$$

The conditional expectation of X given B is

$$\mathbf{E}[X|B] = \sum_x x\, f(x|B)$$

(provided $\sum_x |x|\, f(x|B) < \infty$).

The above is a bit of a cheat. There is a general definition of the conditional expectation that applies to any RV. If we used this definition, then the above definition would be a theorem that says for discrete RV's the general definition reduces to the above. The general definition is pretty abstract, so I am skipping it.

Recall that the partition theorem gave a formula for the probability of an event A in terms of conditional probabilities of A given the events in a partition. There is a similar partition theorem for the expected value of a RV. It is useful when it is hard to compute the expected value of X directly, but it is relatively easy if we know something about the outcome of the experiment.

Theorem 13. *Let $B_1, B_2, B_3, \cdots$ be a finite or countable partition of Ω. (So $\cup_k B_k = \Omega$ and $B_k \cap B_l = \emptyset$ for $k \neq l$.) We assume also that $\mathbf{P}(B_k) > 0$ for all k. Let X be a discrete random variable. Then*

$$\mathbf{E}[X] = \sum_k \mathbf{E}[X|B_k]\,\mathbf{P}(B_k)$$

provided that all the expected values are defined.

3 Multiple Discrete Random Variables

3.1 Joint densities

Definition 12. *If $X_1, X_2, \cdots, X_n$ are discrete RV's, then their joint probability mass function is*

$$f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = \mathbf{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

The joint density for n RV's is a function on $\mathbb{R}^n$. Obviously, it is a non-negative function. It is non-zero only on a finite or countable set of points in $\mathbb{R}^n$. If we sum it over these points we get 1:

$$\sum_{x_1,\cdots,x_n} f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = 1$$

In the discrete case the joint distribution measure is a sum of point masses.

$$\mu_{X_1,X_2,\cdots,X_n} = \sum_{(x_1,x_2,\cdots,x_n)} f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n)\delta_{(x_1,x_2,\cdots,x_n)} \tag{24}$$

When we have the joint pmf of X, Y, we can use it to find the pmf's of X and Y by themselves, i.e., $f_X(x)$ and $f_Y(y)$. These are called "marginal pmf's." The formula for computing them is :

Corollary 2. *Let X, Y be two discrete RV's. Then*

$$\begin{aligned} f_X(x) &= \sum_y f_{X,Y}(x, y) \\ f_Y(y) &= \sum_x f_{X,Y}(x, y) \end{aligned}$$

This generalizes to n discrete RV's in the obvious way.

For discrete RV's, the multivariate change of variables theorem becomes

Theorem 14. *Let $X_1, X_2, \cdots, X_n$ be discrete RV's, and $g : \mathbb{R}^n \to \mathbb{R}$. Then*

$$E[g(X_1, X_2, \cdots, X_n)] = \sum_{x_1, x_2, \cdots, x_n} g(x_1, x_2, \cdots, x_n)\, f_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n)$$

3.2 Independence of discrete RV's

Theorem 15. *Discrete RV's $X_1, X_2, \cdots, X_n$ are independent if and only if*

$$f_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i),$$

Remark: In general, knowing the individual pmf's of X and Y, i.e., $f_X(x)$ and $f_Y(y)$, is not enough to determine the joint pmf of X and Y. But if we also know that the two RV's are independent, then $f_X(x)$ and $f_Y(y)$ completely determine the joint pmf.

Why do we care about independence? The following paradigm occurs often, expecially in statistics and Monte Carlo.

Sampling paradigm: We have an experiment with a random variable X. We do the experiment n times. We will refer to this n-fold repetition of the experiment as the *super-experiment*. Let $X_1, X_2, \cdots, X_n$ be the resulting values of X. These are random variables for the super-experiment. We assume the repetitions of the experiment do not change the experiment and they are independent. So the distribution of each X_j is the same as that of X and $X_1, X_2, \cdots, X_n$ are independent. So the joint pmf is

$$f_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = \prod_{j=1}^{n} f_X(x_j)$$

The RV's $X_1, X_2, \cdots, X_n$ are called i.i.d. (independent, identically distributed). We are often interested in the sample mean

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

It is a RV for the super-experiment. We can find its mean and variance.

The mean of $\bar{X}$ is

$$\mathbf{E}[\bar{X}] = \mathbf{E}[\frac{1}{n}\sum_{j=1}^{n} X_j] = \frac{1}{n}\sum_{j=1}^{n} \mathbf{E}[X_j] = \frac{1}{n}\sum_{j=1}^{n} \mathbf{E}[X] = \mathbf{E}[X]$$

Since the X_j are independent, the variance of $X_1 + X_2 + \cdots X_n$ is the sum of their variances. Since they are identically distributed they have the same variance. In fact, the variance of X_j is the variance of X. So the variance of $X_1 + X_2 + \cdots X_n$ is $n\ var(X)$. Thus

$$var(\bar{X}) = \frac{1}{n^2} n\ var(X) = \frac{1}{n} var(X)$$

So if n is large, the variance of the sample average is much smaller than that of the random variable X.

4 Absolutely continuous random variables

4.1 Densities

Definition 13. *A random variable X is absolutely continuous if there is a non-negative function $f_X(x)$, called the probability density function (pdf) or just density, such that*

$$\mathbf{P}(X \leq t) = \int_{-\infty}^{t} f_X(x)\, dx$$

In undergraduate courses such RV's are often said to be just continuous rather than absolutely continuous. A more abstract way to state the definition is that a random variable X is absolutely continuous if its distribution measure is absolutely continuous with respect to Lebesgue measure on the real line. The Radon-Nikodym derivation of μ_X with respect to Lebesgue measure is then the density $f_X(x)$.

Proposition 2. *If X is an absolutely continuous random variable with density $f(x)$, then*

1. $\mathbf{P}(X = x) = 0$ *for any* $x \in \mathbb{R}$.
2. $\mathbf{P}(a \leq X \leq b) = \int_a^b f(x)\, dx$

3. For any Borel subset C of $\mathbb{R}$, $\mathbf{P}(X \in C) = \int_C f(x)\,dx$

4. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

For absolutely continuous RV's, $d\mu_X(x)$ is $f_X(x)dx$, so the change of variables theorem becomes

Theorem 16. *Let X be an absolutely continuous RV with density $f_X(x)$. Then the expected value of X is given by*

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

If g is a measurable function from $\mathbb{R}$ to $\mathbb{R}$, then

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

(We are omitting some hypothesis that insure things are finite.)

4.2 Catalog

Uniform: (two parameters $a, b \in \mathbb{R}$ with $a < b$) The uniform density on $[a, b]$ is

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

Exponential: (one real parameter $\lambda > 0$)

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \ge 0 \\ 0, & \text{if } x < 0 \end{cases}$$

Normal: (two real parameters $\mu, \sigma > 0$)

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

The range of a normal RV is the entire real line.

4.3 Function of a random variable

Theorem 17. *Let X be a continuous RV with pdf $f(x)$ and CDF $F(x)$. Then they are related by*

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)\,dt, \\ f(x) &= F'(x) \end{aligned}$$

Let X be a continuous random variable and $g : \mathbb{R} \to \mathbb{R}$. Then $Y = g(X)$ is a new random variable. We want to find its density. One way to do this is to first compute the CDF of Y and then differentiate it to get the pdf of Y.

4.4 Histograms and the meaning of the pdf

For a discrete RV the pmf $f(x)$ has a direct interpretation. It is the probability that $X = x$. For a continuous RV, the pdf $f(x)$ is not the probability that $X = x$ (which is zero), nor is it the probability of anything. If $\delta > 0$ is small, then

$$\int_{x-\delta}^{x+\delta} f(u)\,du \approx 2\delta f(x)$$

This is $\mathbf{P}(x - \delta \leq X \leq x + \delta)$. So the probability X is in the small interval $[x - \delta, x + \delta]$ is $f(x)$ times the length of the interval. So $f(x)$ is a *probability density.*

Histogram are closely related to the pdf and can be thought of as "experimental pdf's." Suppose we generate N independent random samples of X where N is large. We divide the range of X into intervals of width Δx (usually called "bins"). The probability X lands in a particular bin is $\mathbf{P}(x \leq X \leq x + \Delta x) \approx f(x)\Delta x$. So we expect approximately $Nf(x)\Delta x$ of our N samples to fall in this bin.

To construct a histrogram of our N samples we first count how many fall in each bin. We can represent this graphically by drawing a rectangle for each bin whose base is the bin and whose height is the number of samples in the bin. This is usually called a frequency plot. To make it look like our pdf we should rescale the heights so that the area of a rectangle is equal to the fraction of the samples in that bin. So the height of a rectangle should be

$$\frac{number\,of\,samples\,in\,bin}{N\,\Delta x}$$

With these heights the rectagles give the histogram. As we observed above, the number of our N samples in the bin will be approximately $Nf(x)\Delta x$, so the above is approximately $f(x)$. So if N is large and Δx is small, the histogram will approximate the pdf.

5 Jointly continuous random variables

5.1 Joint density functions

Definition 14. *Two random variables X and Y are jointly absolutely continuous if there is a function $f_{X,Y}(x,y)$ on $\mathbb{R}^2$, called the joint probability density function, such that*

$$\mathbf{P}(X \le s, Y \le t) = \int \int_{x \le s, y \le t} f_{X,Y}(x,y)\, dxdy$$

The integral is over $\{(x,y) : x \le s, y \le t\}$. We can also write the integral as

$$\begin{aligned} \mathbf{P}(X \le s, Y \le t) &= \int_{-\infty}^{s} \left(\int_{-\infty}^{t} f_{X,Y}(x,y)\, dy \right) dx \\ &= \int_{-\infty}^{t} \left(\int_{-\infty}^{s} f_{X,Y}(x,y)\, dx \right) dy \end{aligned}$$

In this case the distribution measure is just $f_{X,Y}(x,y)$ times Lebesgue measure on the plane, i.e.,

$$d\mu_{(X,Y)}(x,y) = f_{X,Y}(x,y)dxdy \tag{25}$$

In order for a function $f(x,y)$ to be a joint density it must satisfy

$$\begin{aligned} f(x,y) &\ge 0 \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy &= 1 \end{aligned}$$

The definition and formula generalize to n RV's in an obvious way.

The multivariate change of variables theorem becomes (with $n = 2$)

Theorem 18. *Let X, Y be jointly continuous random variables with joint density $f(x, y)$. Let $g(x, y) : \mathbb{R}^2 \to \mathbb{R}$.*

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

(Usual caveat about a missing hypothesis.)

What does the pdf mean? In the case of a single discrete RV, the pmf has a very concrete meaning. $f(x)$ is the probability that $X = x$. If X is a single continuous random variable, then

$$\mathbf{P}(x \leq X \leq x + \delta) = \int_{x}^{x+\delta} f(u)\, du \approx \delta f(x)$$

If X, Y are jointly continuous, than

$$\mathbf{P}(x \leq X \leq x + \delta, y \leq Y \leq y + \delta) \approx \delta^2 f(x, y)$$

5.2 Independence and marginal distributions

Proposition 3. *If X and Y are jointly continuous with joint density $f_{X,Y}(x, y)$, then the marginal densities are given by*

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx \end{aligned}$$

This generalizes to n RV's in the obvious way.

Theorem 19. *Let $X_1, X_2, \cdots, X_n$ be jointly continuous random variables with joint density $f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n)$ and marginal densities $f_{X_i}(x_i)$. They are independent if and only if*

$$f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

5.3 Change of variables

Suppose we have two random variables X and Y and we know their joint density. We have two functions $g : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$, and we define two new random variables by $U = g(X, Y)$, $W = h(X, Y)$. Can we find the joint density of U and W? In principle we can do this by computing their joint CDF and then taking partial derivatives. In practice this can be a mess. There is a another way involving Jacobians which we will study in this section. This all generalizes to the situation where we have n RV's and from n new RV's by taking n different functions of the original RV's.

First we return to the case of a function of a single random variable. Support that X is a continuous random variable and we know it's density. g is a function from $\mathbb{R}$ to $\mathbb{R}$ and we define a new random variable $Y = g(Z)$. We want to find the density of Y. Our previous approach was to compute the CDF first. Now suppose that g is strictly increasing on the range of X. Then we have the following formula.

Proposition 4. *If X is a continuous random variable whose range is D and $f : D \to \mathbb{R}$ is strictly increasing and differentiable, then*

$$f_Y(y) = f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y)$$

We review some multivariate calculus. Let D and S be open subsets of $\mathbb{R}^2$. Let $T(x, y)$ be a map from D to S that is 1-1 and onto. (So it has an inverse.) We also assume it is differentiable. For each point in D, $T(x, y)$ is in $\mathbb{R}^2$. So we can write T as $T(x, y) = (u(x, y), w(x, y))$ We have an integral

$$\int \int_D f(x, y)\, dxdy$$

that we want to rewrite as an integral over S with respect to u and w. This is like doing a substitution in a one-dimensional integral. In that case you have $dx = \frac{dx}{du} du$ The analog of dx/du here is the Jacobian

$$J(u, w) = det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial w} \end{pmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial w} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial w}$$

We then have

$$\int \int_D f(x, y)\, dxdy = \int \int_S f(T^{-1}(u, w))\, |J(u, w)|\, dudw$$

Often $f(T^{-1}(u,w))$ is simply written as $f(u,w)$. In practice you write f, which is originally a function of x and y as a function of u and w.

If A is a subset of D, then we have

$$\int\int_A f(x,y)\,dxdy = \int\int_{T(A)} f(T^{-1}(u,w))\,|J(u,w)|\,dudw$$

We now state what this results says about joint pdf's.

Proposition 5. *Let $T(x,y)$ be a 1-1, onto map from D to S. Let X,Y be random variables such that range of (X,Y) is D, and let $f_{X,Y}(x,y)$ be their joint density. Define two new random variables by $(U,W) = T(X,Y)$. Then the range of (U,W) is S and their joint pdf on this range is*

$$f_{U,W}(u,w) = f(T^{-1}(u,w))\,|J(u,w)|$$

where the Jacobian $J(u,w)$ is defined above.

5.4 Conditional density and expectation

Now suppose X and Y are jointly absolutely continuous random variables. We want to condition on $Y = y$. We cannot do this since $\mathbf{P}(Y = y) = 0$. How can we make sense of something like $\mathbf{P}(a \le X \le b | Y = y)$? We can define it by a limiting process:

$$\lim_{\epsilon \to 0} \mathbf{P}(a \le X \le b | y - \epsilon \le Y \le y + \epsilon)$$

Now let $f(x,y)$ be the joint pdf of X and Y.

$$\mathbf{P}(a \le X \le b | y - \epsilon \le Y \le y + \epsilon) = \frac{\int_a^b \left(\int_{y-\epsilon}^{y+\epsilon} f(u,w)\,dw \right) du}{\int_{-\infty}^{\infty} \left(\int_{y-\epsilon}^{y+\epsilon} f(u,w)\,dw \right) du}$$

Assuming f is continuous and ϵ is small,

$$\int_{y-\epsilon}^{y+\epsilon} f(u,w)\,dw \approx 2\epsilon f(u,y)$$

So the above just becomes

$$\frac{\int_a^b 2\epsilon f(u,y) du}{\int_{-\infty}^{\infty} 2\epsilon f(u,y) du} = \int_a^b \frac{f(u,y)}{f_Y(y)}\,du$$

This motivates the following definition:

Definition 15. *Let X, Y be jointly continuous RV's with pdf $f_{X,Y}(x,y)$. The conditional density of X given $Y = y$ is*

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, if\ f_Y(y) > 0$$

When $f_Y(y) = 0$ we can just define it to be 0. We also define

$$\mathbf{P}(a \le X \le b|Y = y) = \int_a^b f_{X|Y}(x|y)\,dx$$

We have made the above definitions. We could have defined $f_{X|Y}$ and $\mathbf{P}(a \le X \le b|Y = y)$ as limits and then proved the above as theorems.

What happens if X and Y are independent? Then $f(x,y) = f_X(x)f_Y(y)$. So $f_{X|Y}(x|y) = f_X(x)$ as we would expect.

The conditional expectation is defined in the obvious way

Definition 16.

$$\mathbf{E}[X|Y = y] = \int x\, f_{X|Y}(x|y)\,dx$$

As in the discrete case, we have cheated a bit. If we use the general abstract definition of condition expectation, then the above definition would be a theorem.

For continuous random variables we have the following "partition theorems."

Theorem 20. *Let X, Y be jointly absolutely continuous random variables. Then*

$$\mathbf{P}(a \le X \le b) = \int \mathbf{P}(a \le X \le b|Y = y)\, f_Y(y)\,dy$$

where

$$\mathbf{P}(a \le X \le b|Y = y) = \int_a^b f_{X|Y}(x|y)dx$$

and

$$\mathbf{E}[Y] = \int \mathbf{E}[Y|X = x]\, f_X(x)\,dx$$

6 Laws of large numbers, central limit theorem

6.1 Introduction

Let X_n be a sequence of independent, indentically distributed RV's, an i.i.d. sequence. The "sample mean" is defined to be

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

Note that $\overline{X}_n$ is itself a random variable. Intuitively we expect that as $n \to \infty$, $\overline{X}_n$ will converge to $\mathbf{E}[X]$. What exactly do we mean by "convergence" of a sequence of random variables? And what can we say about the rate of convergence and the error? We already saw that the mean of $\overline{X}_n$ is $\mu = \mathbf{E}[X]$. And its variance is σ^2/n where σ^2 is the common variance of the X_j.

6.2 Laws of large numbers

Definition 17. *Let Y_n be a sequence of random variables, and Y a random variable, all defined on the same probability space. We say Y_n converges to Y in probability if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbf{P}(|Y_n - Y| > \epsilon) = 0$$

Theorem 21. *(***Weak law of large numbers***) Let X_j be an i.i.d. sequence with finite mean. Let $\mu = \mathbf{E}[X_j]$. Then*

$$\overline{X}_n \to \mu \quad \textit{in probability}$$

Definition 18. *Let Y_n be a sequence of random variables and Y a random variable. We say X_n converges to X "almost surely" or "with probability one" if*

$$\mathbf{P}(\{\omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1$$

More succintly,

$$\mathbf{P}(Y_n \to Y) = 1$$

This is a stronger form of convergence than convergence in probability. (This is not at all obvious.)

Theorem 22. *If Y_n converges to Y with probability one, then Y_n converges to Y in probability.*

Theorem 23. *(***Strong law of large numbers***) Let X_j be an i.i.d. sequence with finite mean. Let $\mu = \mathbf{E}[X_j]$. Then*

$$\overline{X}_n \to \mu \quad a.s.$$

i.e.,

$$\mathbf{P}(\lim_{n\to\infty} \overline{X}_n = \mu) = 1$$

6.3 Central limit theorem

Let X_n be an i.i.d. sequence with finite variance. Let μ the their common mean and σ^2 their common variance. Define

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{j=1}^n X_j - n\mu}{\sqrt{n}\sigma}$$

Note that $\mathbf{E}[Z_n] = 0, var(Z_n) = 1$.

Theorem 24. *(***Central limit theorem***) Let X_n be an i.i.d. sequence of random variables with finite mean μ and variance σ^2. Define Z_n as above. Then for all $a < b$*

$$\lim_{n\to\infty} \mathbf{P}(a \le Z_n \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

If we take $a = -\infty$, then the theorem says that the CDF of Z_n converges pointwise to the CDF of the standard normal. This is an example of what is called "convergence in distribution" in probability. However, we caution the reader that the general definition of convergence in distribution involves some technicalities.

Confidence intervals: The following is an important problem in statistics. We have a random variable X (usually called the population). We know its variance σ^2, but we don't know its mean μ. We have a "random sample," i.e.,

random variables $X_1, X_2, \cdots, X_n$ which are independent random variables which all have the same distribution as X. We want to use our one sample $X_1, \cdots, X_n$ to estimate μ. The natural estimate for μ is the sample mean

$$\overline{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j$$

How close is $\overline{X}_n$ to the true value of μ ? This is a vague question. We make it precise as follows. For what $\epsilon > 0$ will $\mathbf{P}(|\overline{X}_n - \mu| \leq \epsilon) = 0.95$? We say the $[\overline{X}_n - \epsilon, \overline{X}_n + \epsilon]$ is a 95% confidence interval for μ. (The choice of 95% is somewhat arbitrary. We can use 98% for example.

If n is large we can use the CLT to figure out what ϵ should be. As before we let

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

So $|\overline{X}_n - \mu| \leq \epsilon$ is equivalent to $|Z_n| \leq \epsilon\sqrt{n}/\sigma$ So we want

$$\mathbf{P}(|Z_n| \leq \epsilon\sqrt{n}/\sigma) = 0.95$$

The CLT says that the distribution for Z_n is approximately that of a standard normal. If Z is a standard normal, then $\mathbf{P}(|Z| \leq 1.96) = 0.95$. So $\epsilon\sqrt{n}/\sigma = 1.96$. So we have found that the 95% confidence interval for μ is $[\mu - \epsilon, \mu + \epsilon]$ where

$$\epsilon = 1.96 * \sigma/\sqrt{n}$$