1. Probability and Distributions

1.1. Set Theory.

Definition 1.1. The *union* of two sets

$$A \cup B := \{x \vdots x \in A \text{ or } x \in B\}$$

consists of all points x which are elements of A or elements of B (or both). The *intersection* of two sets

$$A \cap B := \{x \vdots x \in A \text{ and } x \in B\}$$

consists of all points x which are elements of both A and B simultaneously. The intersection of A and B is the points they have in common.

Theorem 1.2. $A \cap B \subseteq A$, $A \cap B \subseteq B$, $A \subseteq A \cup B$, $B \subseteq A \cup B$, and $A \cap B \subseteq A \cup B$,

Theorem 1.3. (Distributive Laws)
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Definition 1.4. The *complement* of A

$$A^* := \{x \vdots x \notin A\}$$

consists of all points x which are not in A. The *difference* of two sets

$$A - B := \{x \in A \text{ and } x \notin B\}$$

consists of all points in A that aren't in B. Thus, $A^* = \mathcal{C} - A$, where $\mathcal{C}$ is the whole space.

Theorem 1.5. (DeMorgan's Laws)
$A - (B \cup C) = (A - B) \cap (A - C)$ and $A - (B \cap C) = (A - B) \cup (A - C)$.

Definition 1.6. When $A \cap B = \varnothing$, we say A and B are *disjoint*. In probability, this is often phrased *A and B are mutually exclusive events*. Similarly for a collection, the sets $\{C_i\}$ are *disjoint* if no two of them have a point in common, i.e.,

$$C_i \cap C_j = \varnothing \text{ whenever } i \neq j.$$

1.2. The Probability Set Function.

Definition 1.7. A *probability set function*, or *probability measure* is a function which assigns numbers to sets $A \subseteq \mathcal{C}$ in such a way that

(i) $0 \leq P(A) \leq 1$ always,
(ii) $P(\mathcal{C}) = 1$, where $\mathcal{C}$ is the whole space, and
(iii) if $\{C_i\}$ are disjoint, then

$$P\left(\bigcup C_i\right) = \sum P(C_i).$$

Intuitively, the *probability of the event A* can be thought of as the "size" or "measure" of the set A, as measured by the function P. The closer it is to 1, the more likely it is that the event A will occur.

Here, $\{C_i\}$ is a collection that can consist of any number of sets: 2, 3, n, or even infinitely many.

Theorem 1.8. $P(A^*) = 1 - P(A)$. Or equivalently, $P(A) + P(A^*) = 1$.

Theorem 1.9. $P(\varnothing) = 0$.

Theorem 1.10. (Monotonicity) If $A \subseteq B$, then $P(A) \leq P(B)$.

Theorem 1.11. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Corollary 1.12. $P(A \cup B) = P(A) + P(B)$ if and only if $P(A \cap B) = 0$.

Remark 1.13. Note that the condition $P(A \cap B) = 0$ in this corollary doesn't necessarily mean $A \cap B = \varnothing$! Example: $\mathcal{C} = [0, 1]$, the unit interval, with $P(A) = \int_A dx$. Then let $A = [0, \frac{1}{2}]$ and $B = [\frac{1}{2}, 1]$, so $A \cap B = \{\frac{1}{2}\} \neq \varnothing$.

Definition 1.14. The events $\{C_i\}$ are *exhaustive* if and only if

$$\mathcal{C} = \bigcup C_i,$$

that is, every element of $\mathcal{C}$ is in some C_i.

Definition 1.15. A *partition* of $\mathcal{C}$ is an exhaustive collection of mutually exclusive events. In other words, a partition of $\mathcal{C}$ consists of chopping up $\mathcal{C}$ into collection of subsets $\{C_i\}$ such that

(i) $C_i \cap C_j = \varnothing$ whenever $i \neq j$, and
(ii) $\mathcal{C} = \bigcup C_i$.

Definition 1.16. (**Basic notion of probability**)
Suppose there are several different possible outcomes of an experiment, and each one is equally likely. The *probability of an event* is

$$P(A) = \frac{\text{number of ways that } A \text{ can occur}}{\text{total possible outcomes}}.$$

Formally, suppose the events $\{C_i\}_{i=n}^n$ form a partition of $\mathcal{C}$, and each of the C_i is equally likely (i.e., $P(C_1) = \cdots = P(C_n) = \frac{1}{n}$). Then for

$$E = C_1 \cup \cdots \cup C_k,$$

we have

$$P(E) = P(C_1 \cup \cdots \cup C_k) = P(C_1) + \cdots + P(C_k) = k \cdot P(C_1) = \tfrac{k}{n}.$$

Here, n is the total number of outcomes, and k is the number of ways in which E can occur.

Example 1. (a) Flipping a fair coin. The probability of the event $A = \textit{head}$ is

$$P(A) = \frac{1 \text{ way to get a head}}{2 \text{ possible outcomes}} = \frac{1}{2}.$$

(b) Throwing a die. The probability of the event $B = \textit{even}$ is

$$P(B) = \frac{3 \text{ ways to get an even number}}{6 \text{ possible different outcomes}} = \frac{3}{6} = \frac{1}{2}.$$

(c) Throwing two dice. The probability of the event $C = (\textit{total} < 8)$ is

$$P(C) = \frac{21 \text{ outcomes have total 7 or less}}{36 \text{ possible different outcomes}} = \frac{21}{36} = \frac{7}{12}.$$

1.3. **Counting.**

The material in this section is not in the textbook, but is very useful for assessing probabilities by counting the number of ways in which an outcome may occur.

Rule. (The Product Rule) Suppose a procedure can be broken down into a sequence of two tasks. If there are n ways to do the first task and m ways to do the second, then there are nm ways to do the procedure.

Similarly, a procedure consisting of k tasks with n_i ways of doing each task can be done in $n_1 n_2 \dots n_k$ ways.

Example 2. How many possible outcomes are there for 4 consecutive flips of a coin? Each flip can have two outcomes, so there are

$$2 \cdot 2 \cdot 2 \cdot 2 = 2^4 = 16$$

possible outcomes.

Example 3. How many possible outcomes are there when two dice are thrown? Each die has 6 possible outcomes, so there are

$$6 \cdot 6 = 6^2 = 36$$

possible outcomes.

Example 4. How many possible license plates are there, if every plate must have a number followed by a sequence of three letters, followed by three more numbers? If all combinations are allowed, there are 10 choices for each number and 26 choices for each letter, so there are

$$10 \cdot 26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 = 10^4 \cdot 26^3 = 175,760,000$$

possible license plates.

Important Note! For these examples, the same number or letter may appear multiple times. This is an example of *replacement*. Also, note that *order matters*: in Example (2) for instance, we are counting HTTT, THTT, TTHT, and TTTH as four distinct events.

Rule. (The Sum Rule) If a first task can be done in n ways, and a second task in m ways, and if these tasks cannot be done at the same time, then there are $n + m$ ways to do one of these tasks.

Example 5. Delta Airlines has 3 flights going to Miami on a given day, and United has 4. How many ways are there to fly to Miami? There are 3+4=7. Here, flying via Delta is one task, and flying via United is another. They are mutually exclusive because you can't do both at the same time.

Example 6. Either a member of the mathematics faculty, or a student who is a mathematics major, will be picked to go on a university committee. If there are 42 faculty members and 122 students, how many different choices are there for a representative? 42+122=164.

Rule. (The Pigeonhole Principle) If $k + 1$ objects are placed into k boxes, then there is at least one box which contains 2 or more objects.

Definition 1.17. (Combinations) A *combination* of elements of a set is an unordered selection of some of them. If the set has n elements and we choose k of them, this is written

$$\binom{n}{n} = \frac{n!}{k!(n-k)!} = C(n,k)$$

and called "n choose k". Combinations are also sometimes called *binomial coefficients* because they appear as the coefficients in the expansion of $(x+y)^n$.

Important Note! Combinations are used when *order does not matter*, and when there is *no replacement*. This is quite different from the Product Rule.

Example 7. How many ways are there to select 6 players from a 10-player team, to compete in a match? Since it doesn't matter what order the players are picked in, and you can only pick a given player once (ie, no replacement), we use combinations:

$$\binom{10}{6} = \frac{10!}{4!6!} = \frac{7\cdot 8\cdot 9\cdot 10}{2\cdot 3\cdot 4} = 7\cdot 3\cdot 10 = 210$$

Example 8. Card games might be the most common source of examples of combinations, because whenever cards are drawn, you can only choose a given card once. Also, the order you get your cards doesn't matter for most simpler games.

(a) How many ways are there to get dealt a hand of 5 cards? Since there are 52 in a deck, there are

$$\binom{52}{5} = \frac{52!}{5!47!} = \langle\text{something huge}\rangle.$$

(b) How many ways are there to get exactly three spades? There are 13 spades, and 39 non-spades. This procedure corresponds to two tasks: choosing 3 spades and choosing 2 non-spades, so we use the product rule. There are $\binom{13}{3}$ ways to choose the two spades, and there are $\binom{39}{2}$ ways to choose the remaining cards, so the product rule says there are

$$\binom{13}{3}\binom{39}{2}$$

different possible hands that contain exactly three spades.

(c) How many ways are there to get at least three spades? We break this up into three disjoint tasks so that we may apply the sum rule:

(at least 3 spades) = (exactly 3 spade) **or** (exactly 4 spades) **or** (exactly 5 spades).

We just saw how to count the number of ways to get exactly 3 spades. Similarly, the number of ways to get exactly 4 spades is $\binom{13}{4}\binom{39}{1}$, and the number of ways to get exactly 5 spades is $\binom{13}{5}\binom{39}{0} = \binom{13}{5}$, using the power rule again for each of these computations. Now, we are ready to use the sum rule to put this all together: the number of ways to at least 3 spades is

$$\binom{13}{3}\binom{39}{2} + \binom{13}{4}\binom{39}{1} + \binom{13}{5}.$$

You haven't seen this yet, but it's coming up in chapter 3, so I'll include it here:

Theorem 1.18. (The Binomial Theorem)

$$(x+y)^n = \sum_{j=0}^{n} \binom{n}{j} x^{n-j} y^j$$
$$= \binom{n}{0} x^n + \binom{n}{1} x^{n-1} y + \binom{n}{2} x^{n-2} y^2 + \cdots + \binom{n}{n-2} x^2 y^{n-2} + \binom{n}{n-1} x y^{n-1} + \binom{n}{n} y^n$$

1.4. **Conditional Probability and Independence.**

Definition 1.19. The *conditional probability* of the event A to occur, given that the event B is known to have occurred is

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

This immediately gives the multiplication rule for probabilities:

$$P(A \cap B) = P(B)P(A|B).$$

This allows us to think of the probability of "two things happening simultaneously" $P(A \cap B)$ as the "probability of one thing happening" $P(B)$ times the "probability that the other happened, given that the first also happened" $P(A|B)$.

Theorem 1.20. For any fixed B, $P(A|B)$ is a probability set function defined for all sets $A \subseteq \mathcal{C}$. Check that it satisfies the three properties.

Definition 1.21. Two events A and B are *independent* if and only if

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

Note: this **does not** imply that $A \cap B = \varnothing$, it **does not** even imply that $P(A \cap B) = 0$! A nice way to think of independence is that if A and B are independent, then the probability of A shouldn't depend on whether or not B has occurred, or vice versa. In other words, if A and B are independent,

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

so the probability of A given B is the same as the probability of A not given B. Whether or not B has occurred has no influence on the likelihood of A.

Definition 1.22. A collection $\{C_i\}$ is *pairwise independent* iff each pair is independent, i.e.,

$$P(C_i \cap C_j) = P(C_i)P(C_j) \text{ whenever } i \neq j.$$

A collection $\{C_i\}$ is *mutually independent* iff every finite subcollection of $\{C_i\}$ satisfies a multiplication rule like this:

$$P(C_i \cap C_j \cap \cdots \cap C_m) = P(C_i)P(C_j)\ldots P(C_m) \text{ whenever all indices are different.}$$

Theorem 1.23. (Law of total probability)
Suppose $\{C_i\}_{i=1}^n$ is a partition of $\mathcal{C}$. Then the probability of some event A can be given in terms of this partition as

$$\begin{aligned} P(A) &= P(C_1)P(A|C_1) + P(C_2)P(A|C_2) + \cdots + P(C_n)P(A|C_n) \\ &= \sum_{i=1}^{n} P(C_i)P(A|C_i) \end{aligned}$$

Note that this allows us to compute $P(A)$ in terms of the probability of A given that C_1 occurred, plus the probability of A given that C_2 occurred, plus ... etc.

Theorem 1.24. (Bayes' theorem)
Suppose $\{C_i\}_{i=1}^n$ is a partition of $\mathcal{C}$. Then

$$P(C_j|A) = \frac{P(A \cap C_j)}{P(A)} = \frac{P(C_j)P(A|C_j)}{\sum_{i=1}^{n} P(C_i)P(A|C_i)}.$$

1.5. **Random Variables.**

Definition 1.25. A *random variable* $X : \mathcal{C} \to \mathbb{R}$ is any function from the sample space to the real numbers. Thus, it maps outcomes to numbers and events to subsets of $\mathbb{R}$.

Definition 1.26. X is a *discrete* random variable if it takes finitely many values in $\mathbb{R}$, or at most a countably infinite number of values.
X is a *continuous* random variable if it takes uncountably many values in $\mathbb{R}$, or varies continuously throughout some subset of $\mathbb{R}$, e.g., whenever X can take any value in an interval (a, b), then you have a continuous random variable.

Definition 1.27. The *probability density function* (pdf) f of a random variable X indicates the relative likelihood of X to take a given value.

(a) When X is discrete, $f(x)$ is exactly the probability that $X = x$ occurs, i.e., the definition of the pdf simplifies to
$$f(x) := Pr(X = x).$$

(b) When X is continuous, it is easier to define the pdf in terms of the distribution function. If F is the df of X, then define the pdf by
$$f(x) := \tfrac{d}{dx}F(x) = F'(x).$$

Definition 1.28. If X is a (discrete or continuous) random variable, then its *distribution function* (df) or *cumulative distribution function* F is defined by

$$F(x) := Pr(X \leq x).$$

The df can also be defined in terms of f as

$$F(x) := \sum_{y \leq x} f(y) \quad \text{or} \quad F(x) := \int_{y \leq x} f(y),$$

depending on whether X is discrete or continuous, respectively.

Remark 1.29. (Relationship of f to F)
This was just stated above, but for emphasis, *they are related by the Fundamental Theorem of Calculus*, i.e.:

$$f(x) = F'(x), \quad \text{and} \quad F(x) = \int_{-\infty}^{x} f(t)\,dt.^{1}$$

Often when X is continuous, you can use this to get f from F or vice versa. When X is discrete, you may be better off trying to use the graph to construct one from the other.

Theorem 1.30. (Properties of the pdf)

(1) $f(x) \geq 0$.
(2) If X is discrete, $f(x) \leq 1$. If X is continuous, this may not be the case.
(3) $\sum_{x\in\mathbb{R}} f(x) = 1$ or $\int_{\mathbb{R}} f(x)dx = 1$, always.

Theorem 1.31. (Properties of the df)

(1) $0 \leq F(x) \leq 1$, always.
(2) $x < y$ implies that $F(x) \leq F(y)$, i.e., $F(x)$ increases with x.
(3) (a) If X does not take any value less than a, then $F(x) = 0$ for any $x < a$.
    (b) If X does not take any value greater than b, then $F(x) = 1$ for any $x > b$.
(4) $\sum_{x\in\mathbb{R}} f(x) = 1$ or $\int_{\mathbb{R}} f(x)dx = 1$, always.
(5) If X is discrete, the $F(x)$ is a step function. Each step occurs at a point of $\mathbb{R}$ which lies in the range of X, and has a height equal to the probability $f(x) = Pr(X = x)$. $F(x)$ is constant along each step.
(6) (a) $F(\infty) := \lim_{x\to\infty} F(x) = 1$.
    (b) $F(-\infty) := \lim_{x\to-\infty} F(x) = 0$.
(7) $F(x)$ is continuous from the right, i.e., for $\varepsilon > 0$ and for any point a,

$$F(a+) := \lim_{\varepsilon\to 0^+} F(a+\varepsilon) = F(a).$$

so the limit as you approach from the right exists and is equal to the function value at that point. (On the graph, "the dot on the left end of any step/interval is filled in.")

Definition 1.32. In terms of a random variable X, the *probability of an event* A is

$$P(A) := Pr(X \in A).$$

Theorem 1.33. The probability of an event A is computed via the random variable X by

$$P(A) = Pr(X \in A) = \sum_{x\in A} f(x) \quad \text{or} \quad \int_{x\in A} f(x)\,dx.$$

Example 9. In general, whenever you are given a pdf, rv or df, and asked to find the probability of a set, this is the strategy to use:

[1] This should be clear for X continuous. It is also true for X discrete, but the derivative requires a little refinement before you can make it precise (there is a remark on p.45 about this). Still, turn to page 34-35 and compare Figures 1.3 and 1.4. Stare at it until you can see how the "weights" of f at points $\{1, 2, 3\}$ in Fig. 1.4 *do* somehow correspond to the rate of change of F at these points. A derivative is supposed to be the rate of change, right ... ? When f has a bump of height $\frac{1}{2}$, F increases by $\frac{1}{2}$, etc.

(1) Let $A = (\frac{1}{4}, \frac{9}{2}]$. Then if X is continuous,

$$P(A) = Pr\left(\tfrac{1}{4} < X \leq \tfrac{9}{2}\right) = \int_{1/4}^{9/2} f(x)\,dx = F(\tfrac{9}{2}) - F(\tfrac{1}{4}).$$

(2) Let $A = (\frac{1}{4}, \frac{9}{2}]$. Then if X is discrete and only takes values among the positive integers, its pdf f will only be positive on the positive integers (and zero elsewhere). Then

$$P(A) = \sum_{x \in A \cap \mathbb{N}} f(x) = \sum_{x=1}^{4} f(x),$$

since $A \cap \mathbb{N} = (\frac{1}{4}, \frac{9}{2}] \cap \{1, 2, 3, 4, 5, \dots\} = \{1, 2, 3, 4\}$.

Example 10. (The Uniform Distribution) When probability is uniformly distributed, the likelihood of X taking a given value is the same as the likelihood of X taking any other value. So the pdf f of X gives the same "weight" (probability density) to each point, i.e., f is a constant:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{else} \end{cases}.$$

So if we look at the df F, it must increase at a constant rate, i.e., it is linear on $[a, b]$:

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

Draw the graphs of both of these to compare, and convince yourself that the horizontal line you see in the graph of f really means "equally likely" or "completely random".

1.6. **Expectation.**

Definition 1.34. The *expectation* of a random variable is just the sum of $xf(x)$ over the whole space:

$$E(X) := \sum_{x \in \mathbb{R}} xf(x), \quad \text{or} \quad E(X) := \int_{x \in \mathbb{R}} xf(x)\,dx.$$

The first sum is written over $\mathbb{R}$ because x can take any real value, but you should keep in mind that $f(x)$ will only ever be nonzero at finitely many (or countably infinitely many) points, so you only have to count the values of $xf(x)$ for these points.

Remark 1.35. Expectation, or *expected value*, should be thought of intuitively as the *payoff* you might expect to receive. Another important way to think of it is the *average value*. This is probably easiest to see for the finite discrete case, like a simple gambling game. See the Remark on page 52 or the following example.

Example 11. Suppose you are playing a game with three possible outcomes for each "round" of the game. Each outcome may have a different probability, say $p_1 = \frac{1}{2}, p_2 = \frac{1}{3}, p_3 = \frac{1}{6}$. Suppose the outcomes also have "payoffs", e.g., the first outcome pays $x_1 = \$1$, the second

pays $x_2 = \$5$, and the third pays $x_3 = \$8$. Then if you were to play this game repeatedly, over time you might expect to win (on average)

$$x_1p_1 + x_2p_2 + x_3p_3 = \frac{1}{2} + \frac{5}{3} + \frac{8}{6} = \$3.50,$$

each round. This is your expectation. Also, it is the most you should be willing to pay the game: if it costs \$4 to play this game, you will likely lose (on average) 50 cents per round.

It may be helpful to dissect the above example in terms of the random variable, etc. In this game, X is mapping outcomes of the game to 1,5, or 8, the payoffs. The probability of each payoff is given by the pdf $f(x)$, whose graph has a bump of height $\frac{1}{2}$ over 1, $\frac{1}{3}$ over 5, and $\frac{1}{6}$ over 8.

Note: expectation can be negative! There is more than just probabilities involved here, there are payoffs, too! If the payoffs are sufficiently negative, you can expect to lose money on the game. Consider the above example again, but this time, you lose when x_1 comes up, i.e., $x_1 = -\$10$ (so you have to pay \$10 when this outcome occurs). Then the expectation is

$$x_1p_1 + x_2p_2 + x_3p_3 = -\frac{10}{2} + \frac{5}{3} + \frac{8}{6} = -\$2.00,$$

and you can expect to lose \$2 every round. Ouch!

(Re)considered in terms of random variables, X is mapping outcomes of the game to -10,5, or 8, the payoffs. The probability of each payoff is given by the pdf $f(x)$, whose graph has a bump of height $\frac{1}{2}$ over -10, $\frac{1}{3}$ over 5, and $\frac{1}{6}$ over 8. f still only takes positive values, but now it takes one of them at a negative point on the real line, and it is THIS which results in a negative expectation.

Example 12. Standard figures for a state lottery are about 135 million players, with a jackpot of \$10 million. Suppose a ticket costs \$5. Let's compute the expectation. The outcomes here are: you win, or you don't. Probability of winning is

$$p = \tfrac{1}{135{,}000{,}000},$$

so probability of losing is

$$1 - p = \tfrac{134{,}999{,}999}{135{,}000{,}000}.$$

The payoff for winning is \$10 million, the payoff for losing is \$0. So the expectation is

$$10{,}000{,}000 \cdot \tfrac{1}{135{,}000{,}000} + 0 \cdot \tfrac{134{,}999{,}999}{135{,}000{,}000} \approx 0.074.$$

So you can expect to lose, on average, \$4.92 every time you play the lotto.

Theorem 1.36. The expectation of a function $u(X)$ of a random variable is given by the formula

$$E(u(X)) := \sum_x u(x)f(x) \quad \text{or} \quad \int_{\mathbb{R}} u(x)f(x)\,dx.$$

Remark 1.37. (Expectation is Linear)

$$E(aX + bY) = aE(X) + bE(Y),$$

where $a, b \in \mathbb{R}$ are constants and X, Y are random variables. This should not come as any great surprise, because expectation is just an integral (or sum) and you've known since basic calculus that

$$\int (af + bg)dx = a \int f\,dx + b \int g\,dx$$

and

$$\sum (af + bg) = a \sum f + b \sum g.$$

In particular, for a constant k,

$$E(k) = \int kf(x)\,dx = k \int f(x)\,dx = k \cdot 1 = k.$$

Definition 1.38. The *mean* μ of an rv is just its expectation:

$$\mu := E(X).$$

The *variance* σ^2 of an rv measures how far X is from μ, on average:

$$\text{var}(X) = \sigma^2 := E\left((X - \mu)^2\right) = E\left(X^2\right) - \mu^2.$$

You should be able to prove the second equality here! It is a theorem, not part of the definition! The *standard deviation* of an rv is just the square root of the variance:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\text{var}(X)}.$$

The *moments* of X are as follows:

$$\begin{aligned}
1^{st} \text{ moment:} \qquad & E(X^1) = E(X) = \int xf(x)\,dx \quad \text{or} \quad \sum_x xf(x) \\
2^{nd} \text{ moment:} \qquad & E(X^2) = \int x^2 f(x)\,dx \quad \text{or} \quad \sum_x x^2 f(x) \\
3^{rd} \text{ moment:} \qquad & E(X^3) = \int x^3 f(x)\,dx \quad \text{or} \quad \sum_x x^3 f(x) \\
& \vdots \\
k^{\text{th}} \text{ moment :} \qquad & E(X^k) = \int x^k f(x)\,dx \quad \text{or} \quad \sum_x x^k f(x)
\end{aligned}$$

The *moment-generating function* of X is

$$M(t) := E\left(e^{tX}\right) = \int e^{tx} f(x)\,dx \quad \text{or} \quad \sum_x e^{tx} f(x).$$

Theorem 1.39. We can find the moments of a random variable X by computing the derivatives of the moment-generating function and evaluating at $t = 0$:

$$E(X^k) = M^{(k)}(t).$$

For example,

$$\begin{aligned} 1^{st} \text{ moment:} &\qquad \mu = E(X) = M'(0) \\ 2^{nd} \text{ moment:} &\qquad E(X^2) = M''(0) \\ 3^{rd} \text{ moment:} &\qquad E(X^3) = M'''(0) \end{aligned}$$

1.7. **Chebyshev's Inequality.**

Lemma 1.40. If $u(X)$ is a nonnegative function, then

$$Pr(u(X) \geq c) \leq \frac{E(u(X))}{c}.$$

Theorem 1.41. (Chebyshev's Inequality)

$$Pr\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2} \quad \text{and} \quad Pr\left(|X - \mu| < k\sigma\right) \geq 1 - \frac{1}{k^2}.$$

This allows you to compute upper and lower bounds for the probability of some sets. For example, if you wanted to find out what percentage of the class received test scores within two standard deviations of the mean, you could compute

$$Pr\left(|X - \mu| < 2\sigma\right) \geq 1 - \frac{1}{2^2} \geq 1 - \frac{1}{4} = \frac{3}{4} = 75\%.$$

Similarly, the percentage of students with test scores within 3 standard deviations of the mean would be

$$Pr\left(|X - \mu| < 3\sigma\right) \geq 1 - \frac{1}{3^2} \geq 1 - \frac{1}{9} = \frac{8}{9} \approx 89\%.$$

Throughout this review sheet, you may assume that every rule given for continuous random variables in terms of integrals has an analogous rule for discrete random variables in terms of sums. The only real exception to this is that you can't get a discrete pdf from a discrete df by differentiating, but everything else should be fine. If you are confused or doubtful about something, ask me.

2. Multivariate Distributions

2.1. Distributions of Two Random Variables.

Definition 2.1. The *joint pdf* of two random variables X, Y is a function $f(x, y)$ such that

$$P(A) = Pr((X, Y) \in A) = \iint_A f(x, y)\, dy\, dx$$

or

$$P(A) = Pr((X, Y) \in A) = \sum_{(x,y)\in A} f(x, y).$$

As before, we always have that the integral/sum over the entire space is 1; this is because the probability of the whole space is 1. In other words,

$$\iint_{\mathbb{R}^2} f(x, y)\, dy\, dx = 1 \qquad \text{and} \qquad \sum_{\text{all } (x,y)} f(x, y) = 1.$$

Definition 2.2. The *joint df* of two random variables X, Y is

$$F(x, y) = Pr(X \le x \text{ and } Y \le y).$$

It is computed via the formula

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

or

$$F(x, y) = \sum_{s\le x, t\le y} f(s, t).$$

Remark 2.3. To find the probability of a given event (i.e., set), sketch the region corresponding to that event and set up an integral over it. For example, the probability that X takes a value between a and b and Y takes a value between c and d would be

$$Pr(a < X \le b, c < Y < d) = \int_a^b \int_c^d f(x, y)\, dy\, dx.$$

Remark 2.4. Note that

$$Pr(a < X \le b, -\infty < Y < \infty)$$

means "the probability that X takes a value in $(a, b]$ and Y takes a value in $(-\infty, \infty)$, at the same time." But of course, Y always takes a value in $(-\infty, \infty)$ (there is nowhere else to go!), so saying $-\infty < Y < \infty$ gives no new information. In other words,

$$Pr(a < X \le b, -\infty < Y < \infty) = Pr(a < X \le b).$$

Earlier, we saw that for a one-variable df,

$$Pr(a < X \le b) = Pr(X \le b) - Pr(X \le a) = F(b) - F(a).$$

Now for a two-variable df, we have that

$$\begin{aligned} Pr(a < X \leq b) &= Pr(a < X \leq b, -\infty < Y < \infty) \\ &= Pr(X \leq b, -\infty < Y < \infty) - Pr(X \leq a, -\infty < Y < \infty) \end{aligned}$$

is the probability that (X, Y) takes a value in the (infinite) vertical strip between the points a and b on the x-axis.

Use this discussion to show that the probability of (X, Y) being in the rectangle $(a, b] \times (c, d]$ is

$$Pr(a < X \leq b, c < Y \leq d) = F(b, d) - F(b, c) - F(a, d) + F(a, c).$$

Hint 1: remember that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
Hint 2: draw the picture! For example, $F(a, c)$ is the probability of landing in the region left of $x = a$ and below $y = c$, so sketch this rectangle. Compare it to the others.

Definition 2.5. If $f(x, y)$ is the joint pdf of X and Y, we can recover the pdf of X by integrating with respect to Y ("integrating out the y's"):

$$f_x(x) = \int_{\mathbb{R}} f(x, y)\, dy$$

is called the marginal pdf of X. Similarly, "integrating out the x's" yields the marginal pdf of Y:

$$f_y(y) = \int_{\mathbb{R}} f(x, y)\, dx.$$

Remark 2.6. You know that for problems involving just a single random variable X, the expectation of X is

$$E(X) = \int_{\mathbb{R}} x f(x)\, dx \qquad \text{or} \qquad E(X) = \sum_{\mathbb{R}} x f(x).$$

Thus, if you are given a joint pdf $f(x, y)$ and asked to find the expectation of X, you must

(1) Find the pdf of X, i.e. the marginal pdf $f_x(x)$ of X, by integrating

$$f_x(x) = \int_{\mathbb{R}} f(x, y)\, dy.$$

(2) Use this to find the expectation of X by evaluating

$$E(X) = \int_{\mathbb{R}} x f_x(x)\, dx.$$

2.2. **Conditional Distributions and Expectations.**

Definition 2.7. The conditional pdf of X given Y is

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

Similarly, the conditional pdf of Y given X is

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

In general, $f_{y|x}(y|x)$ will be a function of both variables x, y. The intuitive meaning of this function $f_{y|x}(y|x)$ is that if you plug in some value for x (e.g. let $x = 0$), then $f_{y|x}(y|x)$ will be the function of y at that "slice of x" which corresponds to a pdf in the y direction. Have a look at the following figures.

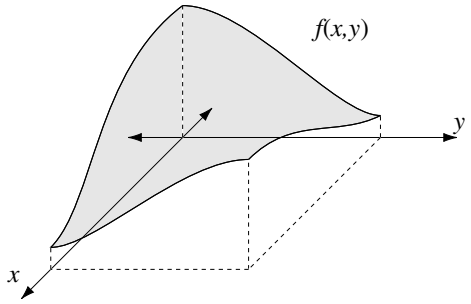

FIGURE 1. Suppose this is the joint pdf of X, Y. It is positive on the unit square $[0, 1] \times [0, 1]$ and 0 elsewhere.

If we fix $x = 0$, then the crosscut through this graph is a function of y that looks like a pdf, provided we "normalize" (i.e., divide by the appropriate marginal so that the integral of the whole thing is 1).

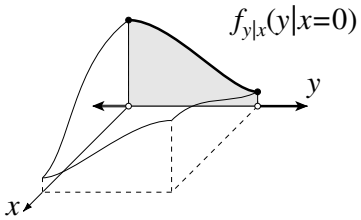

FIGURE 2. Here, the graph over the shaded slice is $f_{y|x}(y|x = 0)$.

By looking at the shape of the graph of the pdf $f_{y|x}(y|x = 0)$, you can see that when given $X = 0$, it is much more likely that Y also takes values near 0. This is simply because

$$f_{y|x}(0|x = 0) > f_{y|x}(1|x = 0).$$

Now look at Figure 3. If we fix $x = 1$, then this slice also looks like a pdf if we normalize.

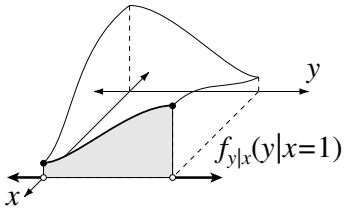

FIGURE 3. Here, the graph over the shaded slice is $f_{y|x}(y|x = 1)$.

Conclusion:

$$f_{y|x}(y|x) \text{ gives you a pdf, for any fixed value of } x.$$

By looking at the shape of the graph of the pdf $f_{y|x}(y|x = 1)$, you can see that when we already know $X = 1$, it is much more likely that Y also takes values near 1. So by comparing the graphs, you can see how when we are given information like $X = 0$ or $X = 1$, it tells us something about whether Y is likely to be large or small.

Definition 2.8. We saw in the 1-variable case that the expectation of X is given by

$$E(X) = \int_{\mathbb{R}} xf(x)\,dx,$$

where $f(x)$ is the pdf of X. By direct analogy, define the *conditional expectation* of X given $Y = y$ to be

$$E(X|y) = \int_{\mathbb{R}} xf_{x|y}(x|y)\,dx,$$

that is, the expectation calculated with the conditional pdf. Similarly, the *conditional expectation* of Y given $X = x$ is

$$E(Y|x) = \int_{\mathbb{R}} yf_{y|x}(y|x)\,dy.$$

Remark 2.9. Note that $E(Y)$ is just a number, but $E(Y|x)$, also denoted $E(Y|X = x)$, is a function. Specifically, it is the function that gives you the expectation when you tell it what slice of x you are at. For example, let's refer to the previous pictures. In Figure 2, Y takes values close to 0 more often than values close to 1. If you average these values (take the **mean** value!) of Y, you might find that on average, Y is $\frac{1}{3}$ (i.e., half of the shaded area is to the left of $y = \frac{1}{3}$ and the other half is to the right). This is another way of saying $E(Y|x = 0) = \frac{1}{3}$.

In Figure 3, Y takes values close to 1 more often than values close to 0. If you average these values (take the **mean** value!) of Y, you might find that on average, Y is $\frac{2}{3}$ (i.e., half of the shaded area is to the left of $y = \frac{2}{3}$ and the other half is to the right). This is another way of saying $E(Y|x = 1) = \frac{2}{3}$.

Conclusion:

$$E(Y|x) = E(Y|X = x) \text{ is a function of } x.$$

We've seen two values for this function:

$$E(Y|X = 0) = \tfrac{1}{3} \qquad \text{and} \qquad E(Y|X = 1) = \tfrac{2}{3}.$$

2.2.1. *Functions of a Random Variable.*

Remark 2.10. In the 1-variable case, we saw that the expectation of a function of a random variable is given by

$$E(u(X)) = \int_{\mathbb{R}} u(x)f(x)\,dx,$$

where $f(x)$ is the pdf of X. For example, let u be the polynomial $u(x) = 3x - x^2$. Then

$$E(3X - X^2) = E(u(X)) = \int_{\mathbb{R}} (3x - x^2)f(x)\,dx.$$

All this carries over immediately to the conditional case. Using the same example as above, $u(x) = 3x - x^2$,

$$E(3X - X^2|y) = E(u(X)|y) = \int_{\mathbb{R}} (3x - x^2) f_{x|y}(x|y)\, dx.$$

It's the same thing, but now using the conditional pdf.

In fact, this carries over the same way to the two variable case as well. For example,

$$E(3XY - Y^2) = E(u(X,Y)) = \iint_{\mathbb{R}^2} (3xy - y^2) f(x,y)\, dx\, dy,$$

where $u(x) = 3xy - y^2$. It's the same thing, but now using the joint pdf $f(x,y)$ of X, Y.

In fact, and this should not be surprising by now, this carries over the same way to the general multivariable case as well. For example,

$$E(3XY - Y^2Z + Z^2) = E(u(X,Y,Z)) = \iiint_{\mathbb{R}^3} (3xy - y^2z + z^2) f(x,y,z)\, dx\, dy\, dz,$$

where $u(x) = 3xy - y^2z + z^2$. The pattern should now be burned into your mind.

Remark 2.11. The comments about conditional expectation apply equally well to conditional variance. In the 1-variable case, we saw

$$\mu = E(X) = \int_{\mathbb{R}} x f(x)\, dx,$$

and

$$\text{var}(X) = \sigma^2(X) = E\left((X-\mu)^2\right) = E\left(X^2\right) - E(X)^2 = E\left(X^2\right) - \mu^2.$$

Likewise, in the conditional case, we have that the conditional variance is given by

$$\text{var}(X|y) = \sigma^2(X|y) = E\left((X - E(X|y))^2|y\right) = E\left(X^2|y\right) - E(X|y)^2.$$

This is computed using the rules outlined in the previous remark, for example,

$$E\left(X^2|y\right) = \int_{\mathbb{R}} x^2 f_{x|y}(x|y).$$

This is just $E(u(X)|y)$ for $u(x) = x^2$.

The situation for variance parallels that of expectation, in the sense that while $\text{var}(Y)$ is a number, $\text{var}(Y|x)$ is a function. In particular, it will be a function that gives the variance for a particular x-slice. For example, the variance $\text{var}(Y|X=0)$ will be a number describing the spread (dispersion) of the pdf highlighted in Figure 2, and the variance $\text{var}(Y|X=1)$ will be a number describing the spread (dispersion) of the pdf highlighted in Figure 3.

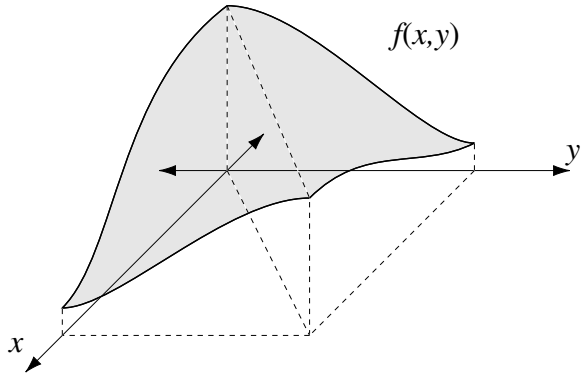

FIGURE 4. The line $y = x$ has been added to Figure 1 and drawn across the surface of the graph of $f(x, y)$, to highlight the ridge where f takes its maximum.

2.3. **The Correlation Coefficient.** In the example given in the figures above, we noticed that when $X = 0$, it seems that Y is also near 0. Additionally, when $X = 1$, it seems that Y is also near 1. In fact, Figure 4 reveals that the joint pdf $f(x, y)$ forms sort of a ridge over the line $y = x$. We describe this phenomenon by saying X and Y are related to each other, or *correlated*. See Figure 4.

If X, Y were more strongly correlated, the graph of their joint pdf might look more like Figure 5. Now, a formula to make this precise.

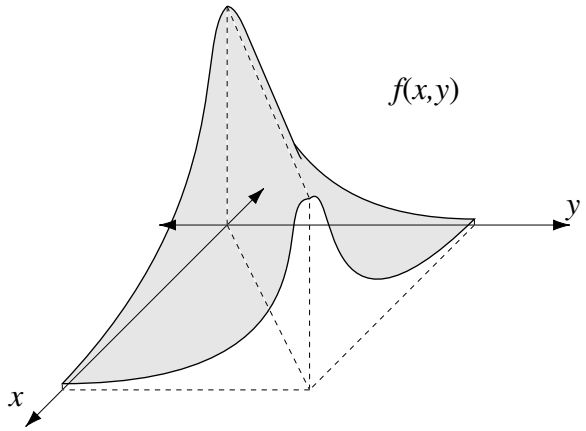

FIGURE 5. The probability is REALLY clustered around the line $y = x$.

Definition 2.12. We write the means of the random variables X, Y as

$$\mu_X = E(X) \qquad \text{and} \qquad \mu_Y = E(Y).$$

Since the variance of X alone is

$$\text{var}(X) = E\left((X - \mu_X)^2\right) = E((X - \mu_X)(X - \mu_X))$$

and the variance of Y alone is

$$\text{var}(Y) = E\left((Y - \mu_Y)^2\right) = E((Y - \mu_Y)(Y - \mu_Y)),$$

we describe how they vary jointly with the *covariance*

$$\operatorname{cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y)) = E(XY) - \mu_X\mu_Y.$$

This is computed as in Remark 2.10:

$$\operatorname{cov}(X,Y) = \iint_{\mathbb{R}^2} xyf(x,y)\,dx\,dy - \left(\int_{\mathbb{R}} xf_x(x)\,dx\right)\left(\int_{\mathbb{R}} yf_y(y)\,dy\right).$$

Remark 2.13. Covariance is an attempt to measure the degree to which X and Y tend to be large at the same time (positively correlated), or the degree to which one tends to be large when the other is small (negatively correlated). For example, suppose that $E[(X-\mu_X)(Y-\mu_Y)]$ is positive. Then

$$(X > \mu_X \text{ and } Y > \mu_Y) \qquad \text{and/or} \qquad (X < \mu_X \text{ and } Y < \mu_Y)$$

must occur together to a greater extent than

$$(X > \mu_X \text{ and } Y < \mu_Y) \qquad \text{and/or} \qquad (X < \mu_X \text{ and } Y > \mu_Y).$$

Otherwise, the mean would be negative.

If the covariance of X, Y is a large positive number, it means that when one is large, the other is very likely to be large. If the covariance of X, Y is a large negative number, it means that when one is large, the other is very likely to be small. If the covariance of X, Y is 0 (or close to it), it means that whatever value X takes is not related to whatever value Y takes.

KEY POINT: In the next section, you see that if X, Y are independent, then

$$E(XY) = E(X)E(Y).$$

So if X, Y are independent, their covariance is

$$\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y).$$

So if X, Y are independent, whatever value X takes is not related to whatever value Y takes.

The converse of the KEY POINT is not true. That is, you *can* have $\operatorname{cov}(X,Y) = 0$ when X and Y are *dependent*. Here is a simple example: let $f(x) = \frac{1}{3}, x = -1, 0, 1$. Define $Y = X^2$. Since Y is defined entirely in terms of X, it is clear that X and Y are dependent; the value of Y is entirely determined by X. However,

$$E(XY) = E(X^3) = E(X) = 0.$$

So we have $E(XY) = E(X)E(Y) = 0$ with X, Y uncorrelated but NOT independent.

However, the magnitude of $\operatorname{cov}(X,Y)$ is also influenced by the magnitudes of X and Y, and this may be misleading. For example, you can show that $\operatorname{cov}(2X,Y) = 2\operatorname{cov}(X,Y)$. Do it![2] This is motivation for the idea of correlation, a sort of normalized (i.e., max value is 1) version of covariance.

[2]This is a good exercise. In fact, it is not much harder to show that for any constants $a, b, c, d \in \mathbb{R}$,

$$\operatorname{cov}(aX+b, cY+d) = ac\operatorname{cov}(X,Y).$$

Definition 2.14. If X, Y have variances which are positive and finite, then the *correlation of X and Y* is

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

If $\rho(X,Y) = 1$, then X, Y are basically distributed the same way. If $\rho(X,Y) = 0$, then X, Y are unrelated. If $\rho(X,Y) = 1$, then X and Y are inversely related.

The next theorem shows why $-1 < \rho(X,Y) < 1$.

Theorem 2.15. (Not in the book). For any X, Y, we have $-1 < \rho(X,Y) < 1$.

Proof. We use the following form of the Cauchy-Schwartz inequality:[3]

$$[E(XY)]^2 \leq E\left(X^2\right) E\left(Y^2\right).$$

From this inequality, it follows that

$$\begin{aligned} [\operatorname{cov}(X,Y)]^2 &\leq \sigma_X^2 \sigma_Y^2 \\ |\operatorname{cov}(X,Y)| &\leq \sigma_X \sigma_Y. \end{aligned}$$

But then from the definition of correlation,

$$|\rho(X,Y)| = \frac{|\operatorname{cov}(X,Y)|}{\sigma_X \sigma_Y} \leq \frac{\sigma_X \sigma_Y}{\sigma_X \sigma_Y} = 1$$

□

2.3.1. *The Moment-Generating Function.* We saw in the 1-variable case that the expectation of a function of a random variable could be calculated by

$$E(u(X)) = \int_{\mathbb{R}} u(x) f(x)\, dx,$$

and in the multivariable case, we extended this to

$$E(u(X_1, X_2, \ldots, X_n)) = \iint \ldots \int_{\mathbb{R}^n} u(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2\, \ldots dx_n.$$

Now we apply this in the case of a particular function, the exponential.

Definition 2.16. The *moment-generating function* (mgf) of a random variable X is

$$M(t) = E\left(e^{tX}\right) = \int_{\mathbb{R}} e^{tx} f(x)\, dx.$$

The *moment-generating function* of two random variables X, Y is

$$M(t,s) = E\left(e^{tX+sY}\right) = \iint_{\mathbb{R}^2} e^{tx+sy} f(x,y)\, dx\, dy.$$

[3]If you are a math major, you should take a good look at this form of the C-S inequality and compare it to what you've seen before. Convince yourself that this is actually the same as

$$|\vec{a} \cdot \vec{b}| \leq \|\vec{a}\| \cdot \|\vec{b}\|.$$

Ask me if you have doubts ...

And in the most general case, the *moment-generating function* of n random variables $X_1, \dots X_n$ is

$$\begin{aligned} M(t_1, t_2, \dots, t_n) &= E\left(e^{t_1X_1+t_2X_2+\cdots+t_nX_n}\right) \\ &= \iint \dots \int_{\mathbb{R}^n} e^{t_1x_1+t_2X_2+\cdots+t_nX_n} f(x_1, x_2, \dots, x_n)\, dx_1\, dx_2 \dots dx_n. \end{aligned}$$

Remark 2.17. The significance of the mgf is that it contains all the same information as the df, and vice versa. Thus, they are two different forms of presenting the same random variable. The advantage of the mgf is that it allows you to find the k^{th} moment of X by taking derivatives, rather than integrating, and differentiation is often easier than integrating. Consider that the third moment of X is

$$E(X^3) = \int_{\mathbb{R}} x^3 f(x)\, dx,$$

but this can also be found by

$$E(X^3) = M'''(0),$$

the third derivative of the mgf, evaluated at 0. Of course, you have to integrate to find the mgf initially, but then you needn't integrate again. Note that there are also some discrete distributions which are basically impossible to compute by direct summation, but which are simple to compute via mgf (see homework problems, esp. §1.9). However, you can usually integrate $e^{g(x)}f(x)$ using integration by parts.

2.3.2. *How to use the mgf to answer some common questions.*

For two random variables X, Y, compute the mgf

$$M(t_1, t_2) = E\left(e^{t_1X+t_2Y}\right) = \iint_{\mathbb{R}^2} e^{t_1x+t_2y} f(x, y)\, dx\, dy.$$

Then use it to find:

$$\begin{array}{lll} \text{means} & \mu_1 = E(X) & = \dfrac{\partial M(0,0)}{\partial t_1} \\ & \mu_2 = E(Y) & = \dfrac{\partial M(0,0)}{\partial t_2} \\ \text{variances} & \sigma_1^2 = E(X^2) - \mu_1^2 & = \dfrac{\partial^2 M(0,0)}{\partial t_1^2} - \mu_1^2 \\ & \sigma_2^2 = E(Y^2) - \mu_2^2 & = \dfrac{\partial^2 M(0,0)}{\partial t_2^2} - \mu_2^2 \\ \text{covariance} & \operatorname{cov}(X,Y) = E\left[(X-\mu_1)(Y-\mu_2)\right] & = \dfrac{\partial^2 M(0,0)}{\partial t_1 \partial t_2} - \mu_1\mu_2 \end{array}$$

$$\begin{array}{lll} \text{marginals} & X \text{ has mgf } = M_1(t_1) = E\left(e^{t_1X}\right) = M(t_1, 0) & \\ & Y \text{ has mgf } = M_2(t_2) = E\left(e^{t_2Y}\right) = M(0, t_2) & \end{array}$$

$$\text{independence} \qquad X, Y \text{ are independent} \iff M(t_1, t_2) = M(t_1, 0)M(0, t_2).$$

2.4. **Independence.**

Definition 2.18. In §1.3, we saw that two sets A, B are independent iff

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

When this is true, the conditional probability of A given B becomes

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A). \tag{1}$$

Remark 2.19. In §1.5 and §1.6, we saw that the right way to think about $P(A)$ is in terms of random variables:

$$P(A) = Pr(X \in A) = Pr(\{X \in A\}).$$

(We usually drop the curly braces $\{, \}$ when it seems clear.)
So the definition of independence becomes

$$Pr(\{X \in A\} \cap \{X \in B\}) = Pr(X \in A)Pr(X \in B),$$

and we can rephrase the equality (1) as

$$Pr(X \in A|X \in B) = \frac{Pr(X \in A \cap B)}{Pr(X \in B)} = \frac{Pr(X \in A)Pr(X \in B)}{Pr(X \in B)} = Pr(X \in A).$$

In other words, if A, B are independent, then the probability that X takes a value in A, when you know that X *does* take a value in B, is the same as the probability of X taking values in A when you have no other information. In other words, knowing that X is in B gives you no information about whether or not X is in A. **This is the meaning of independence!**

This situation can only happen if A and B intersect, and if that intersection is "small". This is codified in the equality

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

When you multiply two numbers from $(0, 1)$, you get another number between $(0, 1)$ which is smaller than either of the original numbers. Since $0 \leq P(A) \leq 1$ and $0 \leq P(B) \leq 1$, independence implies that $P(A \cap B)$ must be very small.

To determine whether or not A and B are independent, ask yourself:

> Question: "Does knowing $X \in A$ tell me if $X \in B$ or $X \notin B$?" and vice versa.

Examples

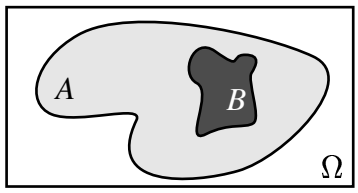

FIGURE 6. Here, if $X \in B$, then clearly X is also in A. The answer to the above question is YES, so A, B are not independent.

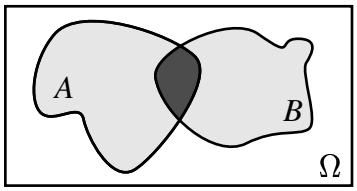

FIGURE 7. Here, if $X \in A$, then maybe X is also in A, but maybe not. You can't say either way. Thus, the answer to the above question is NO, so A, B are independent.

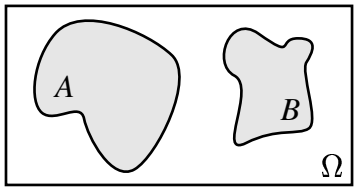

FIGURE 8. Here, if $X \in A$, then clearly X is not in B. The answer to the above question is YES, so A, B are not independent.

2.4.1. *Independent Random Variables.*

We saw the definition of independence for sets:

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

Now we look at independence for random variables.

Definition 2.20. (Independent Random Variables) Suppose X, Y have joint pdf $f(x, y)$, the marginal pdf of X is $f_x(x)$, and the marginal pdf of Y is $f_y(y)$. Then X *and* Y *are independent random variables* iff

$$f(x, y) = f_x(x)f_y(y).$$

Remark 2.21. In equation (1) we saw how conditioning doesn't do anything for independent events:

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

The notion of independence of random variable is defined to make the analogous equation hold true for conditional pdfs:

$$f_{x|y}(x|y) = \frac{f(x, y)}{f_y(y)} = \frac{f_x(x)f_y(y)}{f_y(y)} = f_x(x).$$

In words, if X, Y are independent, then the conditional pdf of X given Y is just the marginal pdf of X; it doesn't depend on Y.

We saw previously that $E(X)$ is a number, and $E(X|y)$ is a function of y; the function that spits out the expectation of X for a given y-value. So what happens when X, Y are independent? In this case, $E(X|y)$ becomes the constant function with value $E(X)$.

$$E(X|y) = \int_{\mathbb{R}} x f_{x|y}(x|y)\, dx = \int_{\mathbb{R}} x \frac{f_x(x)f_y(y)}{f_y(y)}\, dx = \int_{\mathbb{R}} x f_x(x)\, dx = E(X).$$

Theorem 2.22. X, Y are independent iff the joint pdf $f(x, y)$ can be written as a function of x alone multiplied by a function of y alone:

$$f(x, y) = g(x)h(y).$$

Remark 2.23. Point of the theorem: you don't have to calculate the marginals to determine independence. What it says, is that if you can split the pdf into a function $g(x)$ of x and a function $h(y)$ of y, then the only difference between g and h and the marginals is some constant. Example: suppose the pdf of X, Y is

$$f(x, y) = 6x^2y.$$

Then X, Y are independent because

$$\begin{aligned} f(x, y) &= 6x^2 \cdot y \\ &= 2x^2 \cdot 3y \\ &= 3x^2 \cdot 2y \\ &= \tfrac{1}{5}x^2 \cdot 30y \\ &= \dots \end{aligned}$$

The marginals of X and Y will be $f_x(x) = c_1x^2$ and $f_y(y) = c_2y$. We don't know what c_1, c_2 are (and we don't care!), but we know $c_1c_2 = 6$.

Theorem 2.24. (Thm 2, p.103) If X, Y are independent, then

$$Pr(a < X < b, c < Y < d) = Pr(a < X < b)Pr(c < Y < d),$$

or more generally,

$$Pr(X \in A, Y \in B) = Pr(X \in A)Pr(Y \in B).$$

Remark 2.25. For two random variables X and Y and two sets (events) A and B, compare:

$$\begin{aligned} A, B \text{ independent sets} \quad &\implies \quad Pr(X \in A, X \in B) = Pr(X \in A)Pr(X \in B). \\ X, Y \text{ independent rv's} \quad &\implies \quad Pr(X \in A, Y \in B) = Pr(X \in A)Pr(Y \in B). \end{aligned}$$

Theorem 2.26. (Thm 3, p.105) Suppose $u(X)$ is a function of X only, and $v(Y)$ is a function of Y only. If X, Y are independent, then

$$E(u(X)v(Y)) = E(u(X))E(v(Y)).$$

This is the most general case, but it's not the form we use most often. The following immediate corollary is more helpful.

Corollary 2.27. If X, Y are independent, then

$$E(XY) = E(X)E(Y).$$

Proof. Let $u(X) = X$ and $v(Y) = Y$ in the previous theorem. □

Corollary 2.28. If X, Y are independent, then

$$M(t_1, t_2) = M(t_1, 0)M(0, t_2).$$

Proof. Let $u(X) = e^{t_1 X}$ and $v(Y) = e^{t_2 Y}$ in the previous theorem. □

Corollary 2.29. If X, Y are independent, then they are uncorrelated.

Proof. Two corollaries back, we saw that if X, Y are independent, then

$$E(XY) = E(X)E(Y).$$

So the covariance is

$$\operatorname{cov}(X, Y) = E(XY) - E(X)E(Y) = 0,$$

and hence the correlation coefficient is

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{0}{\sigma_X \sigma_Y} = 0.$$

□

Remark 2.30. One of the best and easiest ways to tell if X, Y are independent is hidden in a comment about Example 2 on page 103. Basically, it says that if the space of positive probability is not a rectangle (with sides parallel to the axes), then X, Y must be dependent (i.e., cannot be independent). To see why, let's consider an example similar to #2.35 from the homework.

Example 13. Let X, Y have joint pdf $f(x, y) = 3x, 0 \leq y \leq x \leq 1$. Are X, Y independent?

We sketch the space of positive probability (where $f > 0$): Suppose we are given that

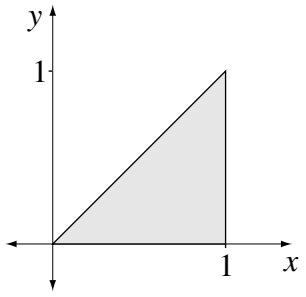

FIGURE 9. The region $0 \leq y \leq x \leq 1$.

$X = 1$. Does this tell us anything about Y? Well, all we know about Y is that for $X = 1$, Y is uniformly distributed between 0 and 1. But what if we instead fix $X = 0$? Does this tell us anything about Y? **YES!** If $X = 0$, then Y can only be 0! So Y depends on X, and they are not independent! Similarly, if we are given that $Y = 1$, it is clear that X must also be 1.

Example 14. To see how this technique can really save you time, consider a problem like #2.34: let X, Y have joint pdf $f(x, y) = \frac{1}{\pi}, (x_1 - 1)^2 + (x_2 + 2)^2 \leq 1$. Are X, Y independent?

To check this directly from the definitions via

$$f(x, y) = f_x(x) f_y(y),$$

you would need to compute a couple of integrals like

$$f_x(x) = \int_{-\sqrt{1-(x_1-1)^2}-2}^{\sqrt{1-(x_1-1)^2}-2} \frac{1}{\pi}\, dx.$$

Instead of such pain, however, you could simply graph the support of $f(x, y)$ (i.e., where $f > 0$):

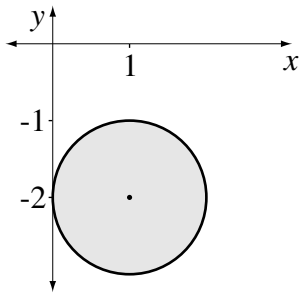

FIGURE 10. The region $(x_1 - 1)^2 + (x_2 + 2)^2 \leq 1$ is a circle of radius 1 which has been translated 1 to the right and 2 down.

As in the previous example, we can see that for $Y = -2$, X is uniformly distributed between 0 and 2, but given $Y = -1$ or given $Y = -3$, we know that X must be 1. Thus X depends on Y and they are NOT independent random variables.

2.4.2. *Summary of ways to check independence.*

(1) X, Y are independent $\iff$ $f(x, y) = g(x)h(y)$.
(2) X, Y are independent $\iff$ $f(x, y) = f_x(x)f_y(y)$.
(3) X, Y are independent $\iff$ $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$.
(4) X, Y are independent $\implies$ $Pr(X \in A, Y \in B) = Pr(X \in A)Pr(Y \in B)$.
(5) X, Y are independent $\implies$ $E(XY) = E(X)E(Y)$.
(6) X, Y are independent $\implies$ $\text{cov}(X, Y) = \rho(X, Y) = 0$.
(7) $\{f(x, y) > 0\}$ is not a rectangle $\implies$ X, Y NOT independent.

Note the direction of the arrows! Anything which points only in one direction cannot be reversed. Recall the counterexample from a previous section:

Example 15. Let $f(x) = \frac{1}{3}, x = -1, 0, 1$. Define $Y = X^2$. Since Y is defined entirely in terms of X, it is clear that X and Y are dependent; the value of Y is entirely determined by X. However,

$$E(XY) = E(X^3) = E(X) = 0.$$

So we have $E(XY) = E(X)E(Y) = 0$ with X, Y uncorrelated but NOT independent.