

Lectures in Probability

David M. McClendon

Department of Mathematics
Ferris State University

2020 edition
©2020 David M. McClendon

Contents

Contents	2
1 Probability spaces	5
1.1 The big picture	5
1.2 Probability spaces	7
1.3 Properties of probability spaces	18
1.4 Conditional probability and independence	26
1.5 The Law of Total Probability and Bayes' Law	31
2 Discrete random variables	36
2.1 Introducing random variables	36
2.2 Density functions of discrete random variables	38
2.3 Counting principles	42
2.4 Bernoulli processes	56
2.5 Summary of Chapter 2	66
3 Continuous random variables	67
3.1 Density functions of continuous random variables	67
3.2 Distribution functions	72
3.3 Transformations of random variables	76
3.4 Poisson processes	83
3.5 More on gamma distributions	94
3.6 Normal random variables	97
3.7 Stirling's formula	103
3.8 Summary of Chapter 3	105
4 Discrete joint distributions	107
4.1 Basic examples	107
4.2 Multinomial and hypergeometric distributions	112

4.3	Independence of discrete random variables	114
4.4	Transformation problems with joint discrete distributions	115
5	Continuous joint distributions	118
5.1	Definitions and elementary properties	118
5.2	Examples	125
5.3	Conditional densities	129
5.4	Transformations of continuous joint distributions	133
6	Expected value	140
6.1	Definition and properties	140
6.2	Variance and covariance	151
6.3	Conditional expectation and conditional variance	159
7	Moment generating functions	164
7.1	Probability generating functions	164
7.2	Moments and moment generating functions	168
7.3	Joint normal densities	174
8	Limit theorems	184
8.1	Markov and Chebyshev inequalities	184
8.2	Laws of large numbers	186
8.3	Central Limit Theorem	190
9	Applications to insurance	195
9.1	Deductibles	195
9.2	Benefit limits	198
9.3	Proportional coverage	201
10	Homework exercises	203
10.1	Exercises from Chapter 1	203
10.2	Exercises from Chapter 2	207
10.3	Exercises from Chapter 3	212
10.4	Exercises from Chapter 4	217
10.5	Exercises from Chapter 5	220
10.6	Exercises from Chapter 6	225
10.7	Exercises from Chapter 7	230
10.8	Exercises from Chapter 8	233
10.9	Exercises from Chapter 9	236
A	Tables	238
A.1	Charts of properties of common random variables	238
A.2	Useful sum and integral formulas	241

A.3	Table of values for the cdf of the standard normal	242
A.4	Road map of standard computations with joint distributions	243
Index		244

Chapter 1

Probability spaces

1.1 The big picture

First question: What is “probability”?

Some history of probability

Pascal & Fermat (1654): correspondence regarding fair odds in games of chance

Bernoulli (1713), de Moivre (1718): basic laws of discrete probability

Boltzmann (1896), Gibbs (1902): statistical mechanics of gases expressed in terms of random motion of large numbers of particles

Kolmogorov (1933): mathematical foundation of the entire subject

Black-Scholes (1973): application of probability to pricing of derivatives

General setup of probability

1. You intend to perform an “experiment” which has different possible “outcomes”.
2. Use *mathematical language* to predict frequencies of these outcomes under repetitions of the experiment.

MOTIVATING EXAMPLES

1. Roll a die repeatedly, and record the number you roll (the number is the outcome).

In this setting, you might be interested in knowing things like:

- What is the likelihood (a.k.a. probability) you will roll a 4 on the third roll?
 - What is the probability you will roll between 9 and 12 fours if you roll the die 60 times?
 - How many rolls on the average will it take you until you roll a 4 for the eighth time?
 - What is the probability you eventually roll nineteen 5s in a row?
 - What is the probability that the sum of the first 200 numbers you roll is less than 650?
2. A driver will be involved in a random number of accidents over the course of a year, and each of these accidents will cause a random amount of damage to his/her car.
 - How long will it take (on the average) for the driver to be involved in three accidents?
 - What is the probability the driver can be accident-free for at least six years?
 - What is the probability the driver will cause more than \$3000 worth of damage over the course of two years?
 - What is the smallest number A such that you can be 99% sure that the driver will cause less than \$ A worth of damage over the next three years?
 - What amount of damage should the driver expect (on the average) to cause over the course of a year?

Probability is the branch of mathematics which solves these types of questions. To solve them, and questions like them, we will

1. learn about a bunch of commonly used models in probability, and
2. learn the *general theory* of models that can be used in probability.

The general theory involves mastery of three intertwining mathematical concepts: *probability spaces*, *random variables* and *stochastic processes*. Loosely speaking:

1. a “probability space” is a structure on which one can formulate a legal method of computing probability;
2. a “random variable” is a measured quantity arising randomly as the result of some experiment (like the number you roll or the amount of damage done in an accident);
3. a “stochastic process” is a collection of random variables indexed by time (like the running total of the numbers you roll or the running amount of total damage done by the driver or the price of a stock).

1.2 Probability spaces

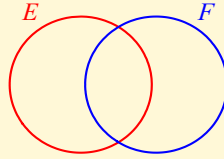
Recall that we seek mathematical language to describe probabilistic experiments.

Definition 1.1 (Outcomes, sample spaces and events)

1. Any possible result of a probabilistic experiment is called an **outcome**.
2. The set of all possible outcomes is called a **sample space** (and is usually denoted Ω).
3. Any “observable” (more on what observable means later) subset of the sample space is called an **event** (events are usually denoted by capital letters like A , B , E , F , ...)

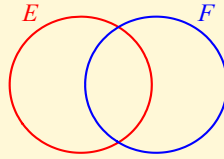
Definition 1.2 (Unions, intersections and complements)

1. Given two events E and F , the event “ E or F or both happen” is the **union** of E and F and is denoted $E \cup F$.



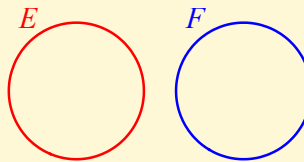
2. Given a bunch of events E_α indexed by α , the **union** of these events, denoted $\bigcup_\alpha E_\alpha$, is the event that at least one of the E_α occur.

3. Given two events E and F , the event “ E and F both happen” is the **intersection** of E and F and is denoted $E \cap F$.

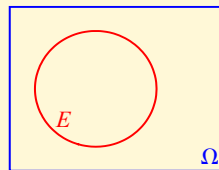


4. Given a bunch of events E_α indexed by α , the **intersection** of these events, denoted $\bigcap_\alpha E_\alpha$, is the event that all of the E_α occur.

5. Two events are called **mutually exclusive** or **disjoint** if they cannot both occur, i.e. if $E \cap F = \emptyset$.



6. Given an event E , the event “ E does not occur” is the **complement** of E and is denoted E^c , E^C , E' , $\Omega - E$, \tilde{E} (and other ways as well).



Examples of this vocabulary can be found on the next page:

EXPERIMENT	SAMPLE SPACE	OUTCOMES	EXAMPLES OF EVENTS
Toss a coin			
Roll a die			
Flip a coin over and over until you flip a heads; record the # of flips			
Record the amount of time (starting now) until your phone rings			

Observability and the definition of a probability space

Start with a sample space Ω . Regard Ω as a mathematical set. We want to describe “observable” subsets of Ω , that is, subsets which we can distinguish.

Some philosophy: I.

II.

Given this philosophical constraint, nothing mathematical “forces” an event to be observable. We are allowed (in the most general sense) to choose a collection \mathcal{A} of subsets of Ω which obey I and II above and decree the subsets belonging to \mathcal{A} to be observable. The idea is that our choice of \mathcal{A} should be a list of observable sets which appropriately models the problem at hand.

(It turns out that there are only two reasonable choices of \mathcal{A} in Math 414.)

Definition 1.3 Let Ω be a set. A nonempty collection \mathcal{A} of subsets of Ω is called a σ –**algebra** (a.k.a. σ –**field**) if

1. \mathcal{A} is “closed under complements”, i.e. whenever $E \in \mathcal{A}$, $E^C \in \mathcal{A}$.
2. \mathcal{A} is “closed under finite and countable unions and intersections”, i.e. whenever $E_1, E_2, E_3, \dots \in \mathcal{A}$, both $\bigcup_j A_j$ and $\bigcap_j A_j$ belong to \mathcal{A} as well.

A subset E of Ω is called \mathcal{A} –**measurable** (or just **measurable**) if $E \in \mathcal{A}$.

The phrases “event”, “measurable set” and “observable set” are synonyms.

Theorem 1.4 Let \mathcal{A} be a σ –algebra of subsets of Ω . Then $\emptyset \in \mathcal{A}$ and $\Omega \in \mathcal{A}$.

PROOF By definition, \mathcal{A} is nonempty, so there is some set E which belongs to \mathcal{A} . Since \mathcal{A} is closed under complements, E^C is also in \mathcal{A} . Now:

- \mathcal{A} closed under finite intersections $\Rightarrow E \cap E^C = \emptyset \in \mathcal{A}$.
- \mathcal{A} closed under finite unions $\Rightarrow E \cup E^C = \Omega \in \mathcal{A}$. \square

EXAMPLES OF σ -ALGEBRAS

Suppose you have a six-sided die where the sides are labeled with a red 1, a red 2, a red 3, a green 1, a green 2, and a green 3. Roll the die once and let Ω be the set of outcomes, i.e.

$$\Omega = \{R1, R2, R3, G1, G2, G3\}.$$

Let's look at some σ -algebras on Ω .

1. Suppose a blind man rolls the die. He can tell whether the die has been rolled (by the sound), but has no idea what number is rolled. Thus the only sets he can observe are \emptyset (the die hasn't been rolled) and Ω (the die has been rolled). He cannot observe the set $\{R1, R2\}$ or $\{R1, G3\}$, because to determine whether or not the outcome lies in that set, he would have to see the die.

The σ -algebra representing the subsets a blind person can see is $\mathcal{A} = \{\emptyset, \Omega\}$.

(Notice that this collection of sets is a σ -algebra, i.e. it is closed under complements, countable unions and countable intersections.)

2. Suppose a red-green colorblind person rolls the die. She can tell what number is rolled, but not what color. So she can observe sets like $\{R1, G1\}$, because to determine whether the outcome is in that set she only needs to know whether or not the number rolled was 1. But she can't observe sets like $\{R1\}$, because she can't tell the difference between $R1$ and $G1$.

The σ -algebra representing the subsets a colorblind person can see can't be easily listed, but can be described as follows: the σ -algebra \mathcal{A} is the set of sets E satisfying

$$Rj \in E \text{ if and only if } Gj \in E, \text{ for all } j \in \{1, 2, 3\}.$$

(Notice that this collection of sets is also a σ -algebra, i.e. it is closed under complements, countable unions and countable intersections.)

3. Suppose a person with 20/20 vision rolls the die. She can distinguish any outcome. Thus the σ -algebra representing the subsets she can see is the collection of all subsets of Ω . (This is clearly closed under complements, countable unions and countable intersections).

Examples 1 and 3 above generalize:

- Let Ω be any set. Then $\mathcal{A} = \{\emptyset, \Omega\}$ is a σ -algebra called the **trivial** σ -algebra on Ω .
- Let Ω be any set. Then the collection of all subsets of Ω , called the **power set** of Ω and denoted 2^Ω , is a σ -algebra on Ω .

Fact If the sample space Ω is finite or countable (including all cases where $\Omega \subseteq \mathbb{Z}$), then we can decree \mathcal{A} to be the power set of Ω and never have a problem. Thus **every subset of a finite or countable sample space can be thought of as observable/measurable.**

Next, we want to calculate the probability of measurable sets:

More philosophy: Given set Ω and σ -algebra \mathcal{A} :

I.

II.

III.

Given these philosophical constraints, nothing else mathematical is forced on us. We are free to choose any assignments of probabilities to events which satisfies these rules. Our choice should appropriately model the context of the original problem.

Definition 1.5 Given set Ω and a σ -algebra \mathcal{A} of subsets of Ω , a **probability measure** on (Ω, \mathcal{A}) is a function $P : \mathcal{A} \rightarrow \mathbb{R}$ satisfying

1. P is **normalized**, i.e. $P(\Omega) = 1$;
2. P is **positive**, i.e. $P(E) \geq 0$ for all $E \in \mathcal{A}$;
3. P is **countably additive on disjoint sets**, i.e. if $E_1, E_2, \dots \in \mathcal{A}$ are all mutually disjoint, then $P\left(\bigcup_j E_j\right) = \sum_j P(E_j)$.

Note: Statement (3) above necessarily implies that if there are infinitely many j with $P(E_j) > 0$, then the infinite series $\sum_j P(E_j)$ must converge.

Definition 1.6 A **probability space** is a triple (Ω, \mathcal{A}, P) where Ω is a set (called the **sample space**), \mathcal{A} is a σ -algebra on Ω (members of \mathcal{A} are called **events**) and P is a probability measure on (Ω, \mathcal{A}) .

EXAMPLE 1

Describe a probability space which represents the result when a fair coin is tossed.

Remark If the sample space Ω is finite or countable (including all cases where $\Omega \subseteq \mathbb{Z}$), then we can define $P : \mathcal{A} \rightarrow \mathbb{R}$ by writing down $P(\omega)$ for each $\omega \in \Omega$, because then for any event E , we can set $P(E) = \sum_{\omega \in E} P(\omega)$.

EXAMPLE 2

Suppose you roll a weighted die where 3 and 4 are three times as likely to appear as any of the other four numbers (3 and 4 are equally likely to occur). Describe a probability space which represents this experiment.

EXAMPLE 3

Flip a fair coin repeatedly until a tail lands for the first time. Describe a probability space which records the number of flips, and verify that you have constructed a probability space.

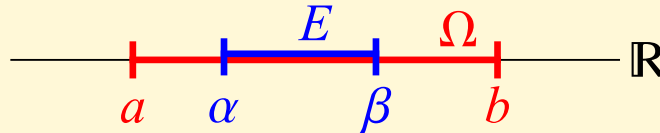
Observability in uncountable sample spaces

EXAMPLE 4

Choose a real number from the interval $[0, 1]$ with all numbers “relatively equally likely”. What is a probability space that models this problem?

Even more philosophy:

Definition 1.7 *Given any interval Ω of finite length (Ω does not have to be closed), there is a σ -algebra of subsets of Ω called the **Lebesgue σ -algebra** (denoted $L(\Omega)$) containing all intervals, all single points, and all countable unions of intervals. Furthermore, there is a probability measure P on $(\Omega, L(\Omega))$ which assigns the probability of any interval to be its normalized length:*



$$P(E) = \frac{\text{length}(E)}{\text{length}(\Omega)} = \frac{\beta - \alpha}{b - a}.$$

*This $(\Omega, L(\Omega), P)$ is called the **uniform distribution** or **normalized Lebesgue measure** on Ω .*

Fact You cannot take \mathcal{A} to be the power set of Ω and obtain a probability measure on $(\Omega, 2^\Omega)$ which assigns the probability of any interval to be its normalized length.

Thus if the sample space of some experiment is an interval of real numbers, and if we are going to compute probabilities in a reasonable way, we must assume that there are some sets which are not observable. It is beyond the scope of MATH

Pick a number X from $[-2, 6)$ with the uniform distribution (i.e. pick the number “uniformly”).

1. What is the probability that $X < 0$?
2. What is the probability that $X = 1$?
3. What is the probability that $X < 0$ or $X > 5$?

All of this stuff generalizes to higher dimensions:

Definition 1.8 Given any set $\Omega \subseteq \mathbb{R}^d$ whose volume is finite (recall that one calculates volumes using multiple integrals), there is a σ -algebra $L(\Omega)$ on Ω and a probability measure P on $(\Omega, L(\Omega))$ such that

1. $L(\Omega)$ contains all subsets of Ω whose volume is calculable using methods of calculus;
2. $(\Omega, L(\Omega), P)$ is a probability space;
3. If $E \in L(\Omega)$, then $P(E) = \frac{\text{volume}(E)}{\text{volume}(\Omega)}$.

This $(\Omega, L(\Omega), P)$ is called the **uniform distribution** or **normalized Lebesgue measure** on Ω , and $L(\Omega)$ is called the **Lebesgue σ -algebra** on Ω .

Note: The word “volume” in Definition 1.8 is a catch-all; in dimension 1, “volume” means length; in dimension 2, “volume” means area; in dimension 3, “volume” means volume, etc.

EXAMPLE 6

Pick a point (X, Y) from the square with vertices $(0, 0)$, $(2, 0)$, $(0, 2)$ and $(2, 2)$ uniformly.

1. What is the probability that $Y \geq X$?

2. What is the probability that $Y = 2X$?

3. What is the probability that $Y < X^2$?

Summary so far

- A probability space is a triple, consisting of
 - a sample space Ω (the set of all outcomes);
 - a σ -algebra \mathcal{A} (the collection of measurable sets, closed under complements, countable unions and countable intersections);
 - and a probability measure P on (Ω, \mathcal{A}) (P is a function which measures the probability of a measurable sets; P must be normalized, positive, and countably additive on disjoint sets).
- If the sample space Ω is finite or countable, we can always decree every subset of Ω to be measurable and can define P as a function on outcomes, rather than a function on events.
- If the sample space Ω is a subset of \mathbb{R}^d , we generally set $\mathcal{A} = L(\Omega)$, the Lebesgue σ -algebra on Ω . This σ -algebra contains all reasonable subsets of Ω , but not all subsets of Ω .
- To calculate probabilities associated to uniform choices of numbers or points, we compute lengths/areas/volumes as appropriate.

1.3 Properties of probability spaces

Recall that a probability space is a triple (Ω, \mathcal{A}, P) where \mathcal{A} is a σ -algebra of subsets of Ω and P is a function from \mathcal{A} to \mathbb{R} such that

- 1.
- 2.
- 3.

We are now going to derive a long list of properties which hold in any probability space. They are called *elementary* properties of probability spaces, because they follow from the definition of a probability space without introducing other deep mathematical ideas.

Theorem 1.9 (Complement Rule) *Let (Ω, \mathcal{A}, P) be a probability space. Then for any event E , $P(E^C) = 1 - P(E)$.*

PROOF E and E^C are disjoint and $E \cup E^C = \Omega$, so by the additivity of P , we have

$$1 = P(\Omega) = P(E \cup E^C) = P(E) + P(E^C).$$

Subtract $P(E)$ from both sides to get the result. \square

Theorem 1.10 (Probabilities are always between 0 and 1) *Let (Ω, \mathcal{A}, P) be a probability space. Then for any event E , $P(E) \in [0, 1]$.*

PROOF By definition, $P(E) \geq 0$. By the complement rule,

$$P(E) = 1 - P(E^C)$$

and since $P(E^C) \geq 0$, $P(E) \leq 1$. \square

Theorem 1.11 *Let (Ω, \mathcal{A}, P) be a probability space. Then $P(\emptyset) = 0$.*

PROOF Apply the Complement Rule to $E = \Omega$. \square

Theorem 1.12 (Monotonicity) *Let (Ω, \mathcal{A}, P) be a probability space, and let E and F be events. If $E \subseteq F$, then $P(E) \leq P(F)$.*

PROOF HW (as a hint, start by writing F as $F = E \cup (F \cap E^C)$.) \square

Theorem 1.13 (De Morgan Law) *Let (Ω, \mathcal{A}, P) be a probability space, and let E_j be an event for all j . Then $P\left(\bigcup_j E_j\right) = 1 - P\left(\bigcap_j E_j^C\right)$.*

PROOF We will first show

$$\left(\bigcup_j E_j\right)^C = \bigcap_j E_j^C$$

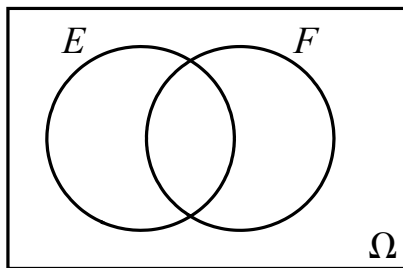
and then apply the Complement Rule.

Recall that we wanted to show

$$\left(\bigcup_j E_j\right)^C = \bigcap_j E_j^C.$$

Theorem 1.14 (Inclusion-Exclusion) *Let (Ω, \mathcal{A}, P) be a probability space, and let E and F be events. Then $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.*

PROOF Start with a Venn diagram:



Theorem 1.15 (Bonferonni Inequality) *Let (Ω, \mathcal{A}, P) be a probability space, and let E and F be events. Then $P(E \cap F) \geq P(E) + P(F) - 1$.*

PROOF HW (as a hint, use Theorems 1.9 and 1.13). \square

Theorem 1.16 (General subadditivity) Let (Ω, \mathcal{A}, P) be a probability space, and let E and F be events. Then $P(E \cup F) \leq P(E) + P(F)$.

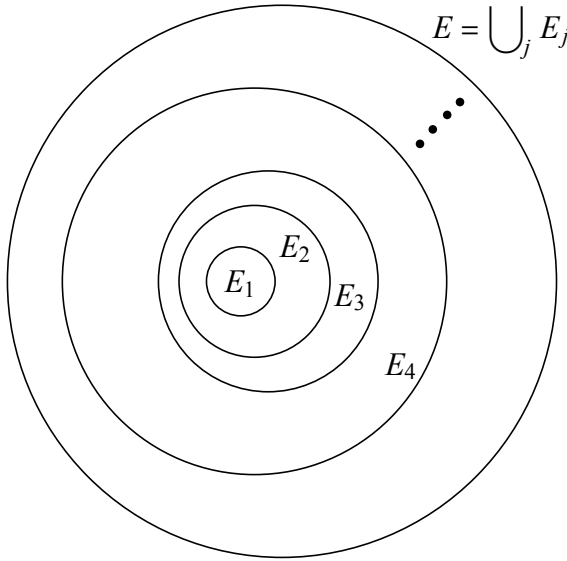
PROOF Follows from Inclusion-Exclusion, together with fact that $P(E \cap F) \geq 0$. \square

Theorem 1.17 (General subadditivity) Let (Ω, \mathcal{A}, P) be a probability space, and let E_1, E_2, E_3, \dots be events. Then $P\left(\bigcup_j E_j\right) \leq \sum_j P(E_j)$.

PROOF Follows from Theorem 1.16 and induction on j . \square

Theorem 1.18 (Continuity of probability measures I) Let (Ω, \mathcal{A}, P) be a probability space, and let E_1, E_2, E_3, \dots be events with $E_1 \subseteq E_2 \subseteq E_3 \subseteq \dots$. Let $E = \bigcup_j E_j$. Then $P(E) = \lim_{j \rightarrow \infty} P(E_j)$.

PROOF The first step of this proof is to “disjointify” the E_j by creating a related sequence of sets F_1, F_2, F_3, \dots



Theorem 1.19 (Continuity of probability measures II) *Let (Ω, \mathcal{A}, P) be a probability space, and let E_1, E_2, E_3, \dots be events with $E_1 \supseteq E_2 \supseteq E_3 \supseteq \dots$. Let $E = \bigcap_j E_j$. Then $P(E) = \lim_{j \rightarrow \infty} P(E_j)$.*

PROOF From the hypothesis, $E_1^C \subseteq E_2^C \subseteq E_3^C \subseteq \dots$. Now by Theorem 1.18,

$$P\left(\bigcup_j (E_j^C)\right) = \lim_{j \rightarrow \infty} P(E_j^C) = \lim_{j \rightarrow \infty} 1 - P(E_j) = 1 - \lim_{j \rightarrow \infty} P(E_j).$$

Applications

EXAMPLE 7

Assume $A \cup B = \Omega$, $P(A \cap B) = \frac{1}{2}$ and $P(A^C) = \frac{1}{3}$. Find $P(B)$.

EXAMPLE 8

Toss a fair coin six times. Find the probability that the first toss or the second toss is heads.

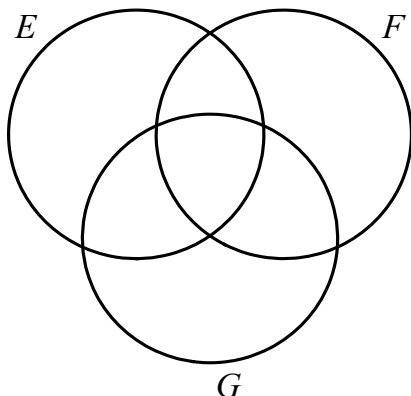
EXAMPLE 9

The chance you lose your umbrella is at least 80%. The chance you lose your glasses is at least 75%. The chance you lose your keys is at least 60%. What is the minimum chance you lose all three items?

Generalized Inclusion-Exclusion

Recall Theorem 1.14 (Inclusion-Exclusion) above, which says that $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

Can you say something similar about $P(E \cup F \cup G)$? Yes:



Theorem 1.20 (3-way Inclusion-Exclusion) *Let (Ω, \mathcal{A}, P) be a probability space, and let E_1, E_2 and E_3 be events. Then*

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3) &= P(E_1) + P(E_2) + P(E_3) \\ &\quad - P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_2 \cap E_3) \\ &\quad + P(E_1 \cap E_2 \cap E_3). \end{aligned}$$

What about $P(E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n)$?

Theorem 1.21 (General Inclusion-Exclusion) *Let (Ω, \mathcal{A}, P) be a probability space, and let $E_1, E_2, E_3, \dots, E_n$ be events. Then*

$$P\left(\bigcup_{j=1}^n E_j\right) = S_1 - S_2 + S_3 - S_4 \dots \pm S_n = \sum_{r=1}^n (-1)^{r+1} S_r$$

where

$$S_r = \sum_{1 \leq i_1 < i_2 < \dots < i_r \leq n} P(E_{i_1} \cap E_{i_2} \cap \dots \cap E_{i_r}).$$

EXAMPLE 10

Suppose that there are three risk factors which affect the chance one will contract a certain disease. Suppose that for any one risk factor, the probability that a randomly chosen person has any one particular risk factor is .8. Suppose that for any two risk factors, the probability that a randomly chosen person has those two risk factors is .7, and suppose that the probability that a person has all three risk factors is .55. What is the probability that a person has none of the three risk factors?

1.4 Conditional probability and independence

Suppose you roll two fair dice. What is the probability that you roll two numbers which sum to 10?

	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

Now let's change the problem a bit. Again, roll two fair dice. What is the probability that you roll two numbers that sum to 10, *given that at least one die roll is a 6*?

When you are asked to compute the probability of one event *given* that another one occurs, the quantity you compute is called a *conditional probability*:

Definition 1.22 Let (Ω, \mathcal{A}, P) be a probability space, and let E and F be events with $P(F) > 0$. The **conditional probability of E given F** , denoted $P(E | F)$, is defined as

$$P(E | F) = \frac{P(E \cap F)}{P(F)}.$$

1.4. Conditional probability and independence

The definition of conditional probability can be rearranged by multiplying through by $P(F)$ to obtain

$$P(E \cap F) = P(F) \cdot P(E | F).$$

This law, called the **Multiplication Principle**, is useful for computing probabilities like these:

EXAMPLE 11

Choose a number from 1 to 4 (each with probability $1/4$) and then roll a die that number of times. What is the probability that you roll a sum of 24?

Definition 1.23 Let (Ω, \mathcal{A}, P) be a probability space, and let E and F be events. E and F are said to be **independent** (denoted $E \perp F$) if $P(E \cap F) = P(E) \cdot P(F)$.

Notes on the definition of independence of events:

1. If $P(E) = 0$ or $P(E) = 1$, then E is independent of any event. (So $\emptyset \perp F$ and $\Omega \perp F$ for any event F .)
2. If $P(E) > 0$ and $P(F) > 0$, then the following are equivalent:

$$E \perp F \Leftrightarrow P(E | F) = P(E) \Leftrightarrow P(F | E) = P(F)$$

Thus to say that two events are independent means heuristically **that the probability that either event occurs is not affected by knowing whether or not the other event occurs.**

3. The following four statements are equivalent (HW):

$$E \perp F \iff E \perp F^C \iff E^C \perp F \iff E^C \perp F^C$$

4. $E \perp E$ if and only if _____ (HW).

EXAMPLE 12

Roll two fair dice. Let E be the event that you roll at least one 6, and let F be the event that you roll a total of at least 10. Are E and F independent? Give a heuristic justification of your answer, and then justify your answer algebraically.

Definition 1.24 Let (Ω, \mathcal{A}, P) be a probability space, and let E_1, \dots, E_n be events. The events E_1, \dots, E_n are called **pairwise independent** if $E_i \perp E_j$ for any $i \neq j$.

Heuristic interpretation: To say some events are “pairwise independent” means that knowing whether or not any **one** of the events occurring does not, by itself, affect the likelihood of any **one** other event.

Definition 1.25 Let (Ω, \mathcal{A}, P) be a probability space, and let E_1, \dots, E_n be events. The events E_1, \dots, E_n are called **mutually independent** (or just **independent**) if for any subset $J \subseteq \{1, \dots, n\}$,

$$P\left(\bigcap_{j \in J} E_j\right) = \prod_{j \in J} P(E_j).$$

Heuristic interpretation: To say that a collection of events is “independent” means that knowing whether or not any **subcollection** of events occur does not affect the likelihood of any other event (or collection of events) occurring.

1.4. Conditional probability and independence

(Mutual) independence implies pairwise independence, but not the other way around, as we see in this example:

EXAMPLE 13

Let $\Omega = \{1, 2, 3, 4\}$ have the uniform distribution. Let $E = \{1, 2\}$, $F = \{1, 3\}$ and $G = \{2, 3\}$. Are E, F, G pairwise independent? Are E, F, G independent?

EXAMPLE 14

Let $\Omega = [0, 1] \times [0, 1]$ have the uniform distribution. Let $E = [0, \frac{1}{2}] \times [0, 1]$, $F = [0, 1] \times [0, \frac{1}{2}]$ and $G = ([0, 1] \times [0, \frac{1}{4}]) \cup ([0, 1] \times [\frac{1}{2}, \frac{3}{4}])$. Are E, F, G pairwise independent? Are E, F, G independent?

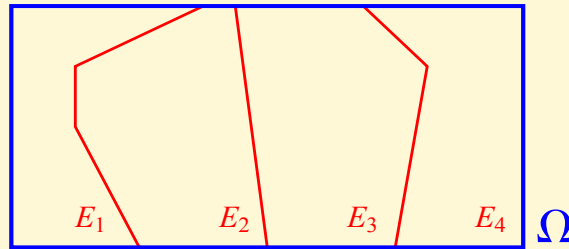
The Monty Hall (or Wayne Brady) Problem

There are three doors on a game show “Let’s Make a Deal”. One door has a car behind it; two doors have piles of manure behind them. You pick a door. Then the game show host shows you that behind a door you did not pick, there is a pile of manure. Then he gives you the option of keeping your door, or switching to the other door you haven’t seen yet. Should you switch?

1.5 The Law of Total Probability and Bayes' Law

Definition 1.26 A **partition** of a probability space (Ω, \mathcal{A}, P) is a collection of events E_1, \dots, E_n such that

1. $P(E_i \cap E_j) = 0$ for all $i \neq j$ (i.e. the E_j are essentially disjoint); and
2. $P(E_1 \cup E_2 \cup \dots \cup E_n) = 1$ (i.e. the union of the E_j s is essentially Ω).



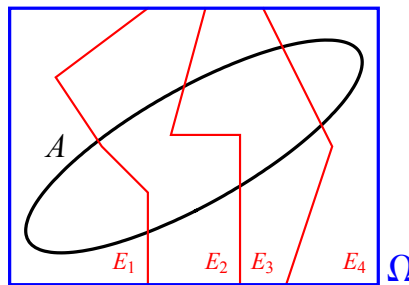
EXAMPLE

$E_1 = [0, \frac{1}{2}]$ and $E_2 = [\frac{1}{2}, 1]$ form a partition of $[0, 1]$ (with Lebesgue measure).

Theorem 1.27 (Law of Total Probability) Let (Ω, \mathcal{A}, P) be a probability space, and let $E_1, E_2, E_3, \dots, E_n$ be a partition. Then for any event A ,

$$P(A) = \sum_{j=1}^n P(A | E_j)P(E_j).$$

PROOF Start by splitting A into its intersections with each of the E_j :



1.5. The Law of Total Probability and Bayes' Law

As a special case of the Law of Total Probability, note that for any event E , E and E^C form a partition of Ω . Thus for any two events A and E , the Law of Total Probability gives us

$$P(A) = P(A | E)P(E) + P(A | E^C)P(E^C).$$

EXAMPLE 15

A fair coin is flipped. If the coin lands heads, a die is rolled once. If the coin lands tails, a die is rolled twice independently. Find the probability that the number(s) rolled sum to 5.

EXAMPLE 16

A survey shows 54% of men believe in aliens, and 33% of women believe in aliens. If 48% of people are men, what percent of people believe in aliens?

Implementing the Law of Total Probability in more complicated situations often involves drawing a diagram called a *tree diagram*, rather than formally describing the events with capital letters:

EXAMPLE 17

A vase contains 3 red and 5 blue marbles. One marble is drawn from the jar and its color recorded, after which it is returned to the jar along with 2 marbles of the opposite color. Then another marble is drawn and its color recorded, after which it is returned to the jar with 2 marbles of the same color. Finally a third marble is drawn. What is the probability that of the three marbles drawn, exactly two of them are blue?

EXAMPLE 18

(from Nate Silver's book *The Signal and the Noise*) Studies show that the chance that a woman in her forties will develop breast cancer is 1.4%. Studies also show that if a woman in her forties does not have cancer, a mammogram will incorrectly claim that she does 10% of the time, and if a woman in her forties does have breast cancer, a mammogram will detect it 75% of the time. Suppose a woman in her forties has a mammogram which indicates she has breast cancer. Given this, what is the probability she actually has breast cancer?

Without reading ahead, guess the answer to this question:

Theorem 1.28 (Bayes' Law) Let (Ω, \mathcal{A}, P) be a probability space, and let E_1, \dots, E_n be a partition. Then for any event A and any $k \in \{1, \dots, n\}$,

$$P(E_k | A) = \frac{P(A | E_k)P(E_k)}{\sum_{j=1}^n P(A | E_j)P(E_j)}.$$

PROOF By direct calculation:

$$\begin{aligned} P(E_k | A) &= \frac{P(E_k \cap A)}{P(A)} \text{ (by definition of cond'l probability)} \\ &= \frac{P(A | E_k)P(E_k)}{P(A)} \text{ (by Multiplication Principle)} \\ &= \frac{P(A | E_k)P(E_k)}{\sum_{j=1}^n P(A | E_j)P(E_j)} \text{ (by Law Total Prob.) } \square \end{aligned}$$

Importance: Bayes' Law tells you how to get $P(E_k | A)$ given all the $P(A | E_j)$.

Application: Think of the E_j as hypotheses and think of the A as some bit of evidence. Theoretically, you should have an idea as to the likelihood that each hypothesis is true (i.e. you know the *prior probabilities* $P(E_k)$). Suppose you actually witness evidence A ; what is the likelihood that hypothesis E_k is the correct hypothesis? This *posterior probability* $P(E_k | A)$ can be computed from the prior probability using Bayes' Law.

Again, note that for any event E , E and E^C form a partition of Ω . Thus for any two events A and E , Bayes' Law gives us

$$P(E | A) = \frac{P(A | E)P(E)}{P(A | E)P(E) + P(A | E^C)P(E^C)}.$$

EXAMPLE 18, REPEATED

Studies show that the chance that a woman in her forties will develop breast cancer is 1.4%. Studies also show that if a woman in her forties does not have cancer, a mammogram will incorrectly claim that she does 10% of the time, and if a woman in her forties does have breast cancer, a mammogram will detect it 75% of the time. Suppose a woman in her forties has a mammogram which indicates she has breast cancer. Given this, what is the probability she actually has breast cancer?

Chapter 2

Discrete random variables

2.1 Introducing random variables

Definition 2.1 A **random variable (r.v.)** X is a (measurable) function $X : \Omega \rightarrow \mathbb{R}^d$, where (Ω, \mathcal{A}, P) is a probability space. The **range** of X is the set of values taken by X .

Definition 2.2 A r.v. is called **real-valued** if its range is a subset of \mathbb{R} . It is called **vector-valued** (or **d -dimensional** or a **joint distribution**) if its range is a subset of \mathbb{R}^d for $d > 1$.

Technical remark: In Math 414/416, the adjective “measurable” can be ignored without a problem. To be technically precise, a function $X : \Omega \rightarrow \mathbb{R}^d$ is measurable if given any subset S of the codomain \mathbb{R}^d whose volume you can compute with calculus, the inverse image of S under X is an event. You would never need to worry about this technicality unless you go to graduate school in mathematics, however.

EXAMPLES OF RANDOM VARIABLES

Example A: Roll a fair die and let X be the number rolled.

Example B: Flip a fair coin 3 times and let X be the number of times you flip heads.

Example C: Roll a die repeatedly; let X be the number of rolls it takes for the running total of your rolls to be even.

Example D: Let X be the smallest amount of time between successive text messages you receive in the next 48 hours.

Example E: You and your friend plan to meet at The Rock between 6 and 7 PM. Let X record both your arrival time and your friend's arrival time, in terms of the number of minutes after 6 that you each arrive.

On the face of things, it seems (based on the definition) that you need a lot of information to describe a random variable: the Ω , the \mathcal{A} , the P and the rule for X . In practice, you don't actually use any of this to characterize a random variable.

First concept: Random variables can be partitioned into three types:

- 1.
- 2.
- 3.

The way you think about a r.v. (and the way you perform calculations related to the r.v.) depend heavily on which type of r.v. you are dealing with. So the first thing you must do when dealing with any r.v. is to determine which of these three types it is.

2.2 Density functions of discrete random variables

For now, we study discrete r.v.s; we'll deal with the others in Chapter 3.

Definition 2.3 A subset S of \mathbb{R}^d is called **discrete** if given any $x \in S$, you can draw a circle (or sphere) of positive radius around x such that the only point inside that circle belonging to S is x itself.

EXAMPLES

\mathbb{N} , \mathbb{Z} , and \mathbb{Z}^d are discrete; any finite set is discrete; any subset of a discrete set is discrete.

NONEXAMPLES

\mathbb{Q} , \mathbb{R} , \mathbb{Q}^d are not discrete; any set containing an interval or a curve is not discrete.

Remark Knowing the examples and nonexamples above is enough in MATH 414 and MATH 416.

Some enrichment: Discreteness is not really a concept of probability theory. It comes from a branch of mathematics called *topology*. In fact, a better definition of discreteness comes from topology - a subset of a metric space is discrete if and only if it has no cluster points (if and only if all its points are isolated).

Definition 2.4 A random variable X is called **discrete** if its range is a discrete set.

QUESTION

Which one or ones of Examples A,B,C,D,E given above are discrete r.v.s?

Return to Example B: (Flip a fair coin 3 times and record the number of heads)

Definition 2.5 Let $X : \Omega \rightarrow \mathbb{R}^d$ be a discrete random variable. A **density function** for X is a function

$$f_X : \text{Range}(X) \rightarrow \mathbb{R}$$

which satisfies

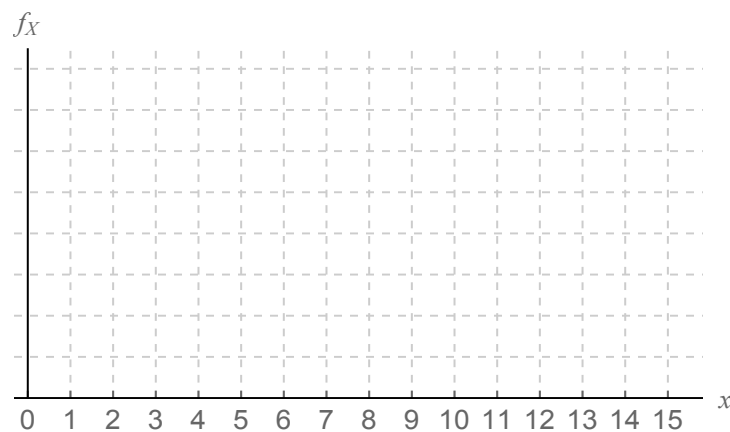
$$f_X(x) = P(X = x)$$

for all $x \in \mathbb{R}^d$.

We express density functions either by giving a formula for them, or in a chart:

EXAMPLE 1

Find density functions for the r.v.s described in Examples A and B above.

[illegible]

Key idea: If you want to do any probabilistic calculations related to a discrete r.v., all you need to be given (or all you need to figure out) is the density function of that r.v. This is because if you are given any set $E \subseteq \mathbb{R}^d$,

$$P(X \in E) = \sum_{x \in E} P(X = x) = \sum_{x \in E} f_X(x)$$

so long as X is discrete.

Properties of density functions of discrete r.v.s

Theorem 2.6 (Properties of density functions) *A function f is the density function of a discrete r.v. if and only if:*

1. $f(x) \geq 0$ for all x ;
2. $\{x : f(x) > 0\}$ is a discrete set; and
3. $\sum_{x \in \{x : f(x) > 0\}} f(x) = 1$.

EXAMPLE 3

Suppose a r.v. X takes only the values 2, 3 and 4 and has a density function that is proportional to $\frac{1}{x^2}$. What is the probability that $X = 2$?

2.3 Counting principles

The first situation we want to model using random variables is when we select a number (or vector or some other kind of object) from a finite set, with all numbers (vectors/objects) equally likely. The random variable that describes this is called a uniform r.v.:

Definition 2.7 Let $\Omega \subseteq \mathbb{R}^d$ be a finite set. A **uniform** random variable on Ω is a r.v. X whose density function is

$$f_X(x) = \begin{cases} \frac{1}{\#(\Omega)} & \text{if } x \in \Omega \\ 0 & \text{else} \end{cases}$$

If X is uniform on $\{a, a+1, a+2, \dots, b\}$, we write $X \sim \text{Unif}(\{a, a+1, a+2, \dots, b\})$.

EXAMPLE 4

Let X be the number rolled if you roll one fair die. Describe X , by giving its density function and characterizing X with appropriate language using the \sim symbol.

If X is uniform on Ω , then given any subset E of Ω , we can compute the probability that $X \in E$ by **counting**:

$$P(E) = P(X \in E) = \frac{\#(E)}{\#(\Omega)}.$$

EXAMPLE 5

Deal 2 cards from a 52 card deck. What is the probability that you get two aces?

To solve problems like this, it behooves us to learn how to count certain sets of objects quickly. The study of counting complicated sets of objects is called **combinatorics**.

(In what follows, $\#(E)$ refers to the number of elements in set E ; **all sets in this section should be assumed finite.**)

Basic counting principles

The first principle of counting is very simple: if you can divide the things you are counting into two disjoint groups, you can count the groups separately and add the answers. For example, if you have 5 red apples and 3 green apples, how many apples do you have?

Theorem 2.8 (Addition Principle of Counting) *Let E and F be finite sets. If $E \cap F = \emptyset$, then*

$$\#(E \cup F) = \#(E) + \#(F).$$

If you divide the things you are counting into two groups which overlap, you can use Inclusion-Exclusion to count them. The proof of this principle is virtually identical to the probabilistic version given in the previous chapter:

Theorem 2.9 (Inclusion-Exclusion Principle (Counting Version)) *Let E and F be finite sets. Then:*

$$\#(E \cup F) = \#(E) + \#(F) - \#(E \cap F).$$

EXAMPLE 6

Suppose that 17 students surveyed like pepperoni on their pizza, 13 students surveyed like mushroom on their pizza and 20 students like pepperoni or mushroom on their pizza. How many students like pepperoni and mushroom on their pizza?

Theorem 2.10 (Multiplication Principle of Counting) *If E is a finite set of objects, each of which can be described as the result of a sequence of “choices”, where:*

- *there are m_1 options for the first choice;*
- *each of the first choices allows m_2 options for the second choice;*
- *each of the first two choices allows m_3 options for the third choice; etc.*

then

$$\#(E) = m_1 m_2 m_3 \cdots m_n.$$

EXAMPLE 7

How many different license plates can a state make if each plate has 4 letters followed by 3 nonzero digits?

Orderings and factorials

EXAMPLE 8

How many different orderings of the letters in the English alphabet are there?

The result of the previous example generalizes:

Definition 2.11 Let $n \in \mathbb{N}$. Then $n!$, read n **factorial**, is

$$n! = n(n-1)(n-2)(n-3) \cdots 3 \cdot 2 \cdot 1.$$

As a special definition, we let $0! = 1$.

Observe: Let $n \in \mathbb{N}$. Then $n \cdot (n-1)! = n!$ (this explains why $0!$ should be 1).

The significance of factorials is as follows:

Theorem 2.12 (Orderings) The number of distinct ways to order n different objects is $n!$.

Permutations

EXAMPLE 9

There are 10 people in a club. How many different sets of officers (president, VP, secretary and treasurer) can be selected from this club?

In Example 9, we are selecting an ordered subset of 4 from a set of 10. These ordered subsets have names:

Definition 2.13 *An ordered subset taken from a larger finite set is called a **permutation**.*

Theorem 2.14 (Permutations) *The number of ordered sets of size k , taken from a set of size n is*

$$\frac{n!}{(n-k)!} = n(n-1)(n-2) \cdots (n-k+1).$$

Combinations

EXAMPLE 10

If there are 10 people in a club, how many different 4–person committees can be formed? (In other words, how many unordered groups of 4 from the group of 10 are there?)

Definition 2.15 *An unordered subset (equivalently, just a subset) taken from a larger finite set is called a **combination**.*

Theorem 2.16 (Combinations) *The number of unordered sets of size k taken from a set of size n is denoted $\binom{n}{k}$ (read “ n choose k ”) or $C(n, k)$ or ${}_nC_k$ and is given by the formula*

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

EXAMPLE 11

$$\binom{7}{3} = \frac{7!}{3!(7-3)!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1} =$$

The numbers $\binom{n}{k}$ are called *binomial coefficients*:

Theorem 2.17 (Properties of binomial coefficients) *Let $n, k \in \mathbb{N}$. Then:*

1. $\binom{n}{k} = \binom{n}{n-k}$;
2. $\binom{n}{k} = 0$ if $n < k$ (this is a definition that makes sense, since if $k > n$, there are no subsets of size k that can be taken from a set of size n);
3. $\binom{n}{0} = \binom{n}{n} = 1$;
4. $\binom{n}{n} = \binom{n}{n-1} = n$;
5. $\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}$;

PROOF Statements (1), (3) and (4) follow from Theorem 2.16 directly. We will prove Statement (5) two different ways: first, a straight algebra proof, where we add fractions by finding common denominators:

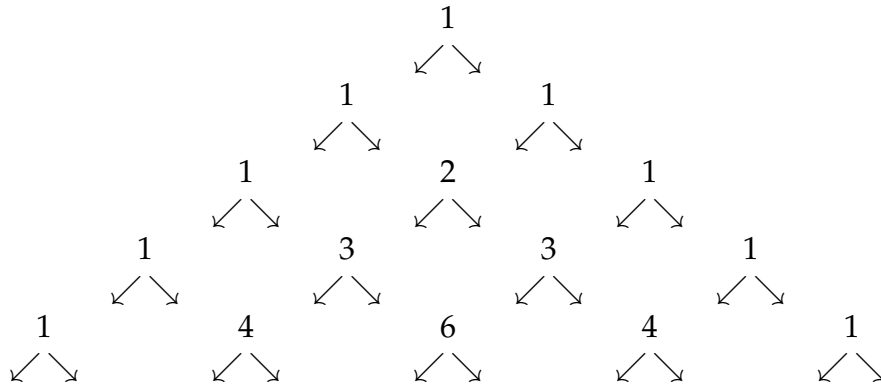
$$\begin{aligned}
 \binom{n}{k-1} + \binom{n}{k} &= \frac{n!}{(k-1)!(n-k+1)!} + \frac{n!}{k!(n-k)!} \\
 &= \frac{n!k}{k!(n-k+1)!} + \frac{n!(n-k+1)}{k!(n-k+1)!} \\
 &= \frac{n![k + (n-k+1)]}{k!(n-k+1)!} \\
 &= \frac{(n+1)!}{k!(n+1-k)!} \\
 &= \binom{n+1}{k}.
 \end{aligned}$$

Next, we give a combinatorial proof of Statement (5). Let E be a set with $n+1$

objects, and let a be one of the elements of E . Then

$$\begin{aligned}
 \binom{n+1}{k} &= \#(\text{subsets of } E \text{ of size } k) \\
 &= \# \left(\begin{array}{c} \text{subsets of } E \\ \text{of size } k \\ \text{not containing } a \end{array} \right) + \# \left(\begin{array}{c} \text{subsets of } E \\ \text{of size } k \\ \text{containing } a \end{array} \right) \\
 &= \# \left(\begin{array}{c} \text{subsets of } E - \{a\} \\ \text{of size } k \end{array} \right) + \# \left(\begin{array}{c} \text{subsets of } E - \{a\} \\ \text{of size } k-1 \end{array} \right) \\
 &=
 \end{aligned}$$

Pascal's Triangle



Based on Property (5) of Theorem 2.17 above, the entries of Pascal's Triangle must be the binomial coefficients (because they have the same entries down the side and they satisfy the same addition law). So Pascal's Triangle is really an array of the binomial coefficients:

$$\begin{array}{ccccccc}
 \binom{0}{0} = 1 & & & & & & \\
 & \binom{1}{0} = 1 & & \binom{1}{1} = 1 & & & \\
 & & \binom{2}{0} = 1 & & \binom{2}{1} = 2 & & \binom{2}{2} = 1 \\
 & & & \binom{3}{0} = 1 & & \binom{3}{1} = 3 & & \binom{3}{2} = 3 & & \binom{3}{3} = 1 \\
 & & & & \binom{4}{0} = 1 & & \binom{4}{1} = 4 & & \binom{4}{2} = 6 & & \binom{4}{3} = 4 & & \binom{4}{4} = 1
 \end{array}$$

EXAMPLE 12

A restaurant has 12 appetizers, 20 entrees and 5 desserts. If your table splits 3 appetizers, 5 entrees and 2 desserts, how many different meals are possible?

EXAMPLE 13

Deal 5 cards from a standard deck. What is the probability of being dealt a full house?

EXAMPLE 14

Deal 5 cards from a standard deck. What is the probability of being dealt two pair (but not a full house)?

Pascal's Triangle is often used to expand expressions, for example

$$\begin{aligned}
 (x + y)^4 &= (x + y)^2(x + y)^2 \\
 &= (x^2 + 2xy + y^2)(x^2 + 2xy + y^2) \\
 &= x^4 + 2x^3y + x^2y^2 + 2x^3y + 2x^2y^2 + 2xy^3 + x^2y^2 + 2xy^3 + y^4 \\
 &= x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4
 \end{aligned}$$

More generally, we have:

Theorem 2.18 (Binomial Theorem) *Let $x, y \in \mathbb{R}$ and let $n \in \mathbb{N}$. Then*

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}.$$

PROOF Expand out the left-hand side and combine like terms:

$$\begin{aligned}
 (x + y)^n &= (x + y)(x + y)(x + y) \cdots (x + y) \\
 &= xxx \cdots x + xx \cdots xy + xx \cdots yx + x \cdots xyxx + \dots \\
 &= x^n + x^{n-1}y + x^{n-2}yx + x^{n-3}yx^2 + \dots + x^{n-2}y^2 + x^{n-3}y^2x + \dots
 \end{aligned}$$

Note: In probability theory, the Binomial Theorem is most often used as a mechanism to compute the sum of some series obtained as the answer to some probability computation:

EXAMPLE 15

$$\sum_{x=0}^{14} \binom{14}{x} 2^x y^{14-x} =$$

Corollary 2.19 *Let $n \in \mathbb{N}$. Then*

$$\sum_{k=0}^n \binom{n}{k} = 2^n.$$

PROOF

$$\sum_{k=0}^n \binom{n}{k} =$$

Distinguishable arrangements and partition problems

EXAMPLE 16

How many different arrangements of the letters in the word MISSISSIPPI are there?

Theorem 2.20 (Distinguishable arrangements) *Suppose you have $n = n_1 + n_2 + \dots + n_r$ objects of r different types:*

- n_1 objects of type 1;
- n_2 objects of type 2;
- \vdots
- n_r objects of type r .

Then the number of distinguishable ways to order these objects is

$$\binom{n}{n_1, n_2, n_3, \dots, n_r} = \frac{n!}{n_1! n_2! n_3! \dots n_r!}.$$

I often call the formula given above the “Mississippi rule”.

Note: Distinguishable arrangements can be thought of as an extension of the idea of a combination. Suppose you have n objects of two types; where k objects are of the first type and $n - k$ objects are of the second type. The number of distinguishable arrangements of these objects is therefore

the same as the number of k combinations from a set of n . This is because arranging the objects is the same as choosing an unordered set of k “slots” in which to place the objects of the first type.

EXAMPLE 17

A box contains 30 red marbles and 20 blue marbles. If you draw 9 marbles from the box all at once, what is the probability that of those 9 marbles, 7 are red?

Theorem 2.21 (Partition problems (sampling without replacement)) *Suppose you have $n = n_1 + n_2 + \dots + n_r$ total objects of r different types:*

- n_1 objects of type 1;
- n_2 objects of type 2;
- \vdots
- n_r objects of type r .

Suppose you draw $k = k_1 + k_2 + \dots + k_r$ objects simultaneously. Then, the probability that you draw k_j objects of type j (for each j) is

$$\frac{\binom{n_1}{k_1} \binom{n_2}{k_2} \dots \binom{n_r}{k_r}}{\binom{n}{k}}.$$

Note: in this setting, drawing objects simultaneously is the same (mathematically) as drawing the objects one at a time without replacement (i.e. without putting back each object you draw before drawing the next object).

What if you draw the objects with replacement (i.e. put each draw back before drawing the next one)? We'll discuss that later.

Suppose that there were only two types of objects: r of type 1 and $n - r$ of type 2. Then, if you draw k objects all at once, you can let X be the number of objects of type 1 you draw.

We summarize this in the following definition:

Definition 2.22 Let $n > 0$, $k \leq n$ and $r \leq n$ be whole numbers. A **hypergeometric** random variable with parameters n, r and k is a discrete r.v. X with range $\{0, 1, 2, \dots, k\}$ whose density function is

$$f_X(x) = \frac{\binom{r}{x} \binom{n-r}{k-x}}{\binom{n}{k}}.$$

If X is hypergeometric with parameters n, r and k , we write $X \sim \text{Hyp}(n, r, k)$.

A $\text{Hyp}(n, r, k)$ r.v. **counts the number of special objects drawn when k objects are drawn at once from a set of n objects, r of which are special.**

Just to make sure the notation is clear, to say

" X is $\text{Hyp}(8, 5, 4)$ " or " $X \sim \text{Hyp}(8, 5, 4)$ "

means X is a hypergeometric r.v. whose density function is

$$f_X(x) = \frac{\binom{5}{x} \binom{3}{4-x}}{\binom{8}{4}}.$$

Theorem 2.23 (Vandermonde's Identity) *Let $r, n, k \in \mathbb{N}$. Then*

$$\sum_{x=0}^k \binom{r}{x} \binom{n-r}{k-x} = \binom{n}{k}.$$

PROOF By the Binomial Theorem,

$$(1+t)^n = \sum_{k=0}^n \binom{n}{k} t^k. \quad (2.1)$$

At the same time,

$$\begin{aligned} (1+t)^n &= (1+t)^r (1+t)^{n-r} = \left[\sum_{x=0}^r \binom{r}{x} t^x \right] \cdot \left[\sum_{y=0}^{n-r} \binom{n-r}{y} t^y \right] \\ &= \sum_{x=0}^r \sum_{y=0}^{n-r} \binom{r}{x} \binom{n-r}{y} t^{x+y} \end{aligned}$$

Next, we do what is called an *index change* in the second summation, which is analogous to a u -substitution in an integral.

In particular, we let $k = x + y$ so that $y = k - x$, and observe that $k = x + y$ goes from $0 + 0 = 0$ to $r + (n - r) = n$. Then the double sum above becomes

$$(1+t)^n = \sum_{x=0}^r \sum_{k=x}^n \binom{r}{x} \binom{n-r}{k-x} t^k = \sum_{k=0}^n \left[\sum_{x=0}^k \binom{r}{x} \binom{n-r}{k-x} \right] t^k.$$

To match equation (2.1) above, the term inside the bracket must be equal to $\binom{n}{k}$. This is Vandermonde's identity. \square

Corollary 2.24 *The density function of a hypergeometric r.v. is in fact a density function (its values sum to 1).*

PROOF Take Vandermonde's identity and divide through by $\binom{n}{k}$. \square

More examples with combinatorics

EXAMPLE 18

Pick a random number with 5 digits (ex: 00312, 15923, etc.) What is the probability that any two digits are the same? What is the probability that exactly two digits are the same?

EXAMPLE 19

Roll seven dice. What is the probability you roll 4 sixes, 2 threes and a one?

EXAMPLE 20 (THE COAT CHECK PROBLEM)

Suppose N people leave their coat at a coat check. The coats get jumbled randomly, so when the people leave, they each get a coat at random (everyone gets a different coat).

1. What is the probability a specified person gets their coat back?
2. What is the probability n specified people get their coat back?
3. What is the probability at least one person gets their coat back?
4. Suppose there are an infinite number of people (i.e. let $N \rightarrow \infty$). What is the probability that no one gets their coat back?

2.4 Bernoulli processes

Definition 2.25 A **stochastic process** $\{X_t : t \in \mathcal{I}\}$ is a collection of random variables indexed by t . The set \mathcal{I} of values of t is called the **index set** of the stochastic process.

Almost always, the index set is $\{0, 1, 2, 3, \dots\}$ or \mathbb{Z} (in which case we call the stochastic process a **discrete-time** process and usually use n instead of t for the index), or the index set is $[0, \infty)$ or \mathbb{R} (in which case we call the stochastic process a **continuous-time** process).

In MATH 414, we will focus on two stochastic processes which are of fundamental importance (we will learn a lot more about stochastic processes in MATH 416). The first one, called the Bernoulli process, is discussed in this section.

Definition 2.26 Let $p \in [0, 1]$. A **Bernoulli experiment** is a probabilistic experiment consisting of a “subexperiment” called a **trial** which is repeated over and over again, where the trials have the following properties:

1. Each trial has two outcomes, **success** and **failure**.
2. On any one trial, the probability of success is p (so the probability of failure is $1 - p$).
3. The result of any one trial is independent of the results of any other trials.

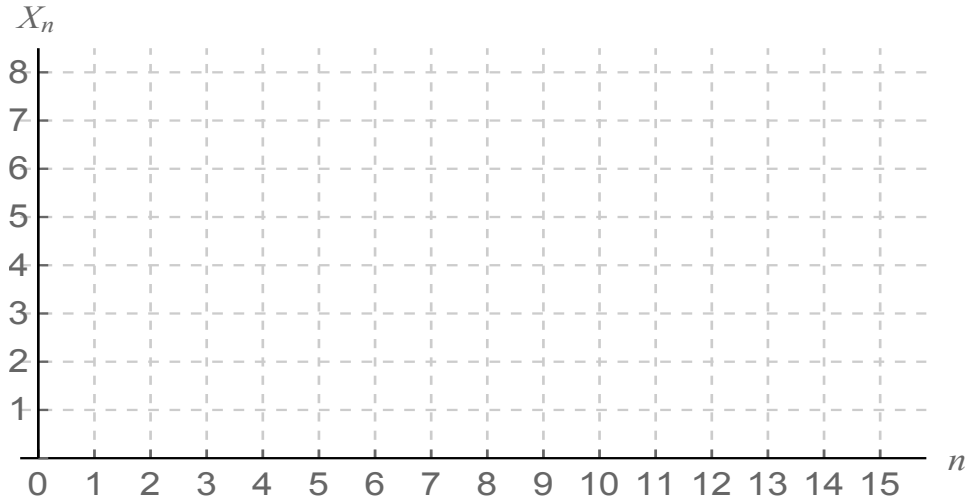
If we let, for $n \in \{0, 1, 2, 3, \dots\}$, X_n be the number of successes in the first n trials, $\{X_n : n \in \{0, 1, 2, \dots\}\}$ is a stochastic process called a **Bernoulli process** and p is called the **success probability**.

To picture a Bernoulli process in your mind, think of flipping a coin repeatedly (which flips heads with probability p) and writing down the sequence of heads and tails you get. X_n is the number of heads you flip in the first n flips.

Suppose you flip this coin repeatedly and get the following results:

T H T H T T T H H T T H T T H ...

You can represent the result of this process by the following picture:



Note: In this setting (these generalize to arbitrary Bernoulli processes):

1. $X_0 = 0$ (since there is no way to flip any positive number of heads in zero flips);
2. every time you flip heads, the value of X_n goes up by 1;
3. every time you flip tails, the value of X_n stays the same;
4. X_n never decreases nor jumps by more than 1 unit at a time.

The definition of a Bernoulli process alone is enough to figure out some basic conditional probability questions:

EXAMPLE 21

Let $\{X_n\}$ be a Bernoulli process with success probability p .

1. Find the probability that $X_8 = 5$, given that $X_6 = 3$.
2. Find the probability that $X_7 = 3$, given that $X_3 = 3$.

3. Let X_n be a Bernoulli process with success probability p . Find the probability that $X_8 = 2$, given that $X_5 = 1$.

4. In Question 3 of this example, what is really relevant? For example, if I asked you to find the probability that $X_t = b$ given that $X_s = a$, what matters about s, t, a and b ?

Binomial random variables

At this point, we want to define a random variable which counts the number of successes in n trials coming from a Bernoulli experiment:

Definition 2.27 A **binomial random variable** with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ is a discrete r.v. taking values in $\{0, 1, 2, \dots, n\}$ whose density function is

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

If X is binomial with parameters n and p , we write $X \sim b(n, p)$ or $X \sim \text{binomial}(n, p)$.

A binomial r.v. with parameters n and p **counts the number of successes in n trials of a Bernoulli process with success probability p .**

The numbers which occur as values of the density function of binomial r.v.s are commonly encountered in probability. We denote by $b(n, p, k)$ the number

$$b(n, p, k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Theorem 2.28 *The density function of a binomial(n, p) r.v. is a density function (i.e. its values sum to 1).*

PROOF Use the binomial theorem:

$$\sum_{x=0}^n f_X(x) = \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} =$$

Remark: Let $\{X_n\}$ be a Bernoulli process with success probability p . Then:

1. For any fixed m and n with $m < n$, the r.v. $X_n - X_m$ is binomial($n - m, p$);
2. For any fixed n , the r.v. X_n is binomial(n, p);
3. If $m < n$, $P(X_n = y \mid X_m = x)$ is the number $b(n - m, p, y - x)$.

Sampling with/without replacement: Suppose you have a bag containing 40 marbles. Of the 40 marbles, 8 are orange. If you draw 20 marbles from the bag, what is the probability that you draw exactly 5 orange marbles?

The answer to this question depends on whether you draw the marbles without replacement (including if they are all drawn at once) or with replacement (i.e. you put each marble back before you draw again).

If the sampling is without replacement:

If the sampling is with replacement:

EXAMPLE 22

Suppose you guess at every question on a 10-question multiple choice test (four choices per question). What is the probability you get exactly 7 questions correct?

EXAMPLE 23 (CHALLENGE)

Suppose you know 75% of the questions that might be asked on a 10-question exam. If you guess at the other 25% of the questions, what is the probability you get all ten questions correct?

Remark: This sum can be computed with *Mathematica* to get $\frac{137858491849}{1099511627776} \approx .1253$.

EXAMPLE 24

A machine produces parts which are defective 1% of the time. Out of 2000 parts produced, what is the probability that exactly 30 parts are defective?

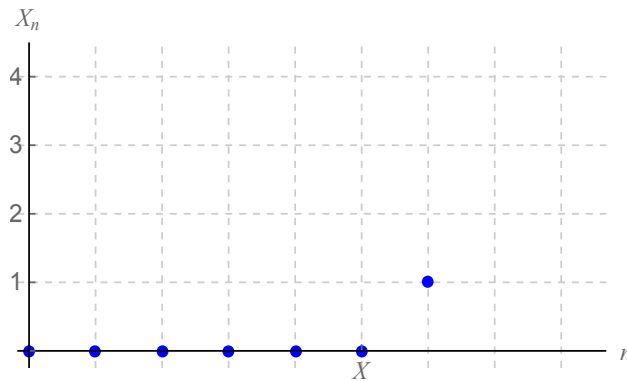
EXAMPLE 25

A fair coin is tossed 11 times (equivalently, 11 fair coins are tossed at once).

1. What is the probability of flipping exactly 7 heads?
2. What is the probability of at least 8 heads?
3. What is the probability of at least one head?

Geometric and negative binomial random variables

We earlier discussed binomial random variables, which describe the height of the graph coming from a Bernoulli process at time n . Now we introduce random variables which describe horizontal measurements on the graph. For example, suppose $\{X_n\}$ is a Bernoulli process with success probability p . Let X be a r.v. which measures the amount of time that passes before the first time the graph of $\{X_n\}$ hits height 1. X is called a **geometric** random variable.



Question: What is the density function of X ?

Definition 2.29 A **geometric** random variable with parameter $p \in [0, 1]$ is a discrete r.v. taking values in $\{0, 1, 2, 3, \dots\}$ whose density function is

$$f_X(x) = p(1 - p)^x.$$

If X is geometric with parameter p , we write $X \sim \text{Geom}(p)$.

A $\text{Geom}(p)$ r.v. counts the number of failures before the first success in a Bernoulli process with success probability p .

Theorem 2.30 The density function of a $\text{Geom}(p)$ r.v. is a density function (i.e. its values sum to 1).

PROOF

$$\sum_{x=0}^{\infty} f_X(x) = \sum_{x=0}^{\infty} p(1 - p)^x = \quad \square$$

The next proposition is a fact about geometric r.v.s worth remembering:

Theorem 2.31 (Hazard law for geometric r.v.s) Let $X \sim \text{Geom}(p)$. Then for any $n \in \mathbb{N}$,

$$P(X \geq n) =$$

PROOF

$$P(X \geq n) = \sum_{x=n}^{\infty} f_X(x) = \sum_{x=n}^{\infty} p(1-p)^x = \quad \square$$

Geometric random variables are exactly the random variables which have an important property called *memorylessness*:

Definition 2.32 A random variable X is called **memoryless** if for all $m, n \geq 0$,

$$P(X \geq m + n \mid X \geq m) = P(X \geq n).$$

To say that a r.v. is memoryless means that if you think of the r.v. as the time it takes for something to happen, if you know you have been waiting for m units, the probability you will wait another n units is the same as the probability you would wait n units from the get go (in other words, you “forget” that you have waited the m units).

Theorem 2.33 *A random variable X taking values in $\{0, 1, 2, \dots\}$ is memoryless if and only if it is geometric.*

PROOF (\Leftarrow) Assume $X \sim \text{Geom}(p)$. Then:

$$\begin{aligned} P(X \geq m+n \mid X \geq m) &= \frac{P(X \geq m+n)}{P(X \geq m)} \\ &= \frac{(1-p)^{m+n}}{(1-p)^m} \\ &= (1-p)^n \\ &= P(X \geq n) \end{aligned}$$

so X is memoryless by definition.

(\Rightarrow) Assume X is memoryless and let $p = P(X = 0)$. By the definition of memorylessness, for all m ,

$$P(X \geq m+1 \mid X \geq m) = P(X \geq 1) = 1 - P(X = 0) = 1 - p.$$

Therefore for all $m \geq 0$, we have

$$P(X \geq m+1) = (1-p)P(X \geq m),$$

and by replacing m with $m-1$ we see that for all $m \geq 1$,

$$P(X \geq m) = (1-p)P(X \geq m-1).$$

This means

Let's now generalize the idea of a geometric random variable. Suppose we wanted to count the number of failures before the r^{th} success in a Bernoulli process, where $r \in \mathbb{N}$. Let X be such a r.v.; what is the density function of X ?

Definition 2.34 A **negative binomial** random variable with parameters $r \in \mathbb{N}$ and $p \in [0, 1]$ is a discrete r.v. taking values in $\{0, 1, 2, 3, \dots\}$ whose density function is

$$f_X(x) = \binom{x+r-1}{r-1} p^r (1-p)^x.$$

If X is negative binomial with parameters r and p , we write $X \sim NB(r, p)$.

That this function is in fact a density function will not be proven here. It uses the power series expansion of the function $(1-p)^{-x}$. (The “ $-$ ” sign here is why we call this the “negative” binomial r.v.)

A $NB(r, p)$ r.v. counts the number of failures before the r^{th} success in a Bernoulli process with success probability p .

Note that a negative binomial r.v. with parameters 1 and p is the same thing as a geometric r.v. with parameter p . (We shorthand this fact by writing “ $NB(1, p) \sim \text{Geom}(p)$ ”.)

Examples with geometric and negative binomial r.v.s**EXAMPLE 26**

Let X be a geometric r.v. so that $P(X \geq 5) = .3$. What is $P(X = 1)$?

EXAMPLE 27

The number of hurricanes that hit Florida in a given year is assumed to be geometric with parameter .85. What is the probability that either 3 or 4 hurricanes will hit Florida this year?

EXAMPLE 28

An urn contains 30 red, 20 green and 50 blue marbles. Marbles are drawn from the urn, one at a time with replacement. What is the probability that the fifth time a green marble is drawn is on the 18th draw?

2.5 Summary of Chapter 2

- A discrete random variable is a function $X : \Omega \rightarrow \mathbb{R}^d$ taking values in a discrete set (like \mathbb{N} or \mathbb{Z} or \mathbb{Z}^d).
- We can completely describe a discrete r.v. X by giving its density function $f_X : \text{Range}(X) \rightarrow [0, 1]$, which is defined by

$$f_X(x) = P(X = x).$$

Such a function must take only values between 0 and 1, and its values must sum to 1. The density function of a discrete r.v. is used to compute probabilities by adding its values: if E is any subset of the range of X ,

$$P(X \in E) = \sum_{x \in E} f_X(x).$$

- Classes of commonly encountered discrete random variables include the following:
 1. uniform r.v.s, which assign equal likelihood to all values in the range of X ;
 2. hypergeometric r.v.s, which count the number of special objects drawn when a sample is drawn without replacement;
 3. binomial r.v.s, which count the number of successes in n trials of a Bernoulli process (and also describe sampling with replacement);
 4. geometric r.v.s, which count the number of failures before the first success in a Bernoulli process (and are the only memoryless discrete r.v.s);
 5. negative binomial r.v.s, which count the number of failures before the r^{th} success in a Bernoulli process.

You must know (or be able to refer to on your cheat sheet) the range and density function of these common r.v.s.

- We solve probability questions associated to uniform r.v.s by counting. Techniques used to count sets include inclusion-exclusion, the multiplication principle, permutations, combinations, distinguishable arrangements, and partition problems.

Chapter 3

Continuous random variables

3.1 Density functions of continuous random variables

Recall that a r.v. X is a function $X : \Omega \rightarrow \mathbb{R}^d$, where (Ω, \mathcal{A}, P) is a probability space.

In the previous chapter, we studied r.v.s which are discrete, including those whose range is finite or countable. Now, we will study r.v.s which are not discrete. First, a definition:

Definition 3.1 A r.v. $X : \Omega \rightarrow \mathbb{R}^d$ is called **continuous (cts)** if, for every $x \in \mathbb{R}^d$, we have

$$P(X = x) = 0.$$

Definition 3.2 A r.v. $X : \Omega \rightarrow \mathbb{R}^d$ is called **mixed** if it is neither discrete nor continuous.

EXAMPLE 1

Pick a number uniformly from $[0, 3]$ and let X be the result.

3.1. Density functions of continuous random variables

Recall: To describe a discrete r.v., we write down a _____
for that r.v. This object tells us two things:

- 1.
- 2.

Question: What's the analogue of this for a cts r.v.? Unfortunately, we can't accomplish both (1) and (2) above when X is cts:

Definition 3.3 Let $X : \Omega \rightarrow \mathbb{R}$ be a cts r.v. We say X has a **density function** f_X (equivalently, f_X is a **density function** for X) if $f_X : \mathbb{R} \rightarrow [0, \infty)$ is such that for any real numbers $a \leq b$,

$$P(X \in [a, b]) = \int_a^b f_X(x) dx.$$

EXAMPLE 1, CONTINUED

What is the density function of the uniform r.v. on $[0, 3]$?

Theorem 3.4 A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is the density function of a cts r.v. $X : \Omega \rightarrow \mathbb{R}$ if and only if all of the following hold:

1. f is measurable (i.e. you can compute $\int_a^b f(x) dx$ for every a and b);
2. $f(x) \geq 0$ for all x ;
3. $\int_{-\infty}^{\infty} f(x) dx = 1$.

As before, the word “measurable” is a technical condition you need not worry about in undergraduate mathematics. There are functions which are not measurable, but they are not motivated by real-world problems and you would never encounter them unless you go to graduate school.

EXAMPLE 2

Suppose X is a continuous r.v. whose density function is

$$f_X(x) = \begin{cases} cx & \text{if } 0 \leq x \leq 3 \\ 0 & \text{else} \end{cases}$$

for some constant c .

1. What is the range of X ?
2. What is the value of c ?
3. Find $P(X \leq 1)$.
4. Find $P(X \geq 2)$.
5. Find $P(X > 2)$.
6. Which is more likely, that $X = 1$ or $X = 2$?
7. Which is more likely, that X is close to 1 or X is close to 2?

3.1. Density functions of continuous random variables

Main idea: If you want to do any probabilistic calculations related to a continuous r.v., all you need to be given (or all you need to figure out) is the density function of that r.v. This is because if you are given any set $E \subseteq \mathbb{R}^d$,

$$P(X \in E) = \int_E f_X(x) dx.$$

Contrast this with how you compute probabilities for discrete r.v.s:

	DISCRETE R.V.S	CONTINUOUS R.V.S
How the density function is defined		
How probabilities are computed using the density		

Bad news: There are continuous r.v.s which do not have a density function. You would not encounter these in any normal situation, however.

The most common type of continuous r.v. is where you choose from a set with all numbers/points equally likely. This is called a *uniform* r.v.:

Definition 3.5 Let $\Omega \subseteq \mathbb{R}$ be a union of intervals whose total length is finite. A **uniform** random variable on Ω is the cts r.v. X with density function

$$f_X(x) = \begin{cases} \frac{1}{\text{total length}(\Omega)} & \text{if } x \in \Omega \\ 0 & \text{else} \end{cases}$$

If X is uniform on a single interval $[a, b] \subseteq \mathbb{R}$, we write $X \sim \text{Unif}([a, b])$.

Example 1 describes a uniform r.v. on $[0, 3]$, for instance.

3.1. Density functions of continuous random variables

EXAMPLE 3

Describe a density function for the uniform r.v. on $[0, 4] \cup [10, 11) \cup \{13\}$.

Remark	The density function of a cts r.v. is never unique – it can be altered on any finite or countable set without affecting any probability computations.
---------------	---

EXAMPLE 4

Find a density function for X , if $X \sim Unif([0, \frac{1}{2}])$.

Remark	Unlike density functions for discrete r.v.s, density functions for cts r.v.s can take values greater than 1.
---------------	--

3.2 Distribution functions

In this section, we address two questions:

1. How do we represent a r.v. which is mixed (neither discrete nor cts)?
2. Is there an object which describes r.v.s, which unifies the theory of discrete, cts and mixed r.v.s?

The answer to these questions is given in the following definition. For now, we'll stick to real-valued r.v.s (and discuss vector-valued r.v.s later).

Definition 3.6 Let $X : \Omega \rightarrow \mathbb{R}$ be a r.v. The **cumulative distribution function** (a.k.a. **distribution function** a.k.a. **cdf**) of X is the function $F_X : \mathbb{R} \rightarrow \mathbb{R}$ defined by

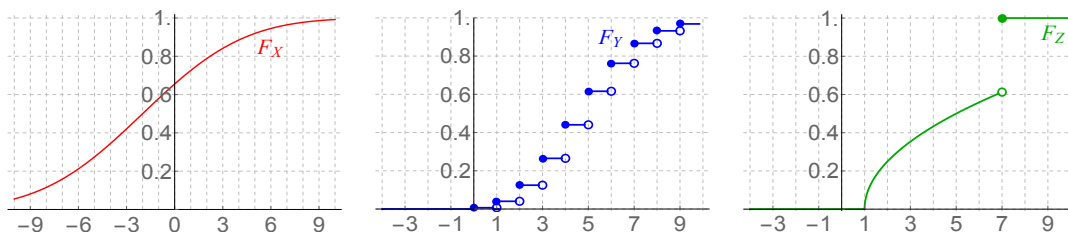
$$F_X(x) = P(X \leq x).$$

EXAMPLE 5

What is the cdf for the uniform r.v. on $[0, 4]$?

EXAMPLE 6

Shown below are graphs of the cdfs for three r.v.s X , Y and Z . What can you tell about X , Y and Z from these graphs? What are the commonalities across these three graphs?



Theorem 3.7 (Properties of distribution functions) Let $X : \Omega \rightarrow \mathbb{R}$ be a r.v. whose cdf is F_X . Then:

1. F_X is the only cdf of X ;
2. F_X is nondecreasing;
3. $\lim_{x \rightarrow -\infty} F_X(x) = 0$;
4. $\lim_{x \rightarrow \infty} F_X(x) = 1$;
5. If $\text{Range}(X) \subseteq (a, b)$, then $F_X(x) = 0$ for all $x \leq a$;
6. If $\text{Range}(X) \subseteq (a, b)$, then $F_X(x) = 1$ for all $x \geq b$;
7. F_X is right-continuous everywhere
(meaning $\lim_{x \rightarrow c^+} F_X(x) = F_X(c)$ for all c).

Theorem 3.8 (Calculating probabilities from density functions) Let $X : \Omega \rightarrow \mathbb{R}$ be a r.v. whose cdf is F_X . Then:

1. $P(X \in (a, b]) = F_X(b) - F_X(a)$ for all $a < b$;
2. $P(X = c) = F_X(c) - \lim_{x \rightarrow c^-} F_X(x)$
(this is the size of the jump in F_X at c);
3. $P(X = c) = 0$ if and only if F_X is continuous at c .

The next theorem generalizes what we observed in the first two graphs in Example 6:

Theorem 3.9 Let $X : \Omega \rightarrow \mathbb{R}$ be a r.v. whose cdf is F_X . Then:

1. X is a cts r.v. if and only if F_X is a continuous function;
2. X is a discrete r.v. if and only if F_X is piecewise constant.

EXAMPLE 7

Suppose X has distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{1}{4}\sqrt{x} & \text{if } x \in (0, 1) \\ \frac{1}{2}x & \text{if } x \in [1, 2) \\ 1 & \text{if } x \geq 2 \end{cases}$$

1. Find $P(X = x)$ for any single real number x .
2. Find $P(X \leq 1)$.
3. Find $P(X < 1)$.
4. Find $P(X \geq 1)$.
5. Find $P(X > 1)$.
6. Find $P(\frac{1}{2} \leq X < \frac{3}{2})$.

Theorem 3.10 (Relationship between density and dist. functions) Let $X : \Omega \rightarrow \mathbb{R}$ be a cts r.v. with density function f_X . Then:

1. $\frac{d}{dx}(F_X(x)) = f_X(x)$; and
2. $\int_{-\infty}^x f_X(t) dt = F_X(x)$.

PROOF Statement (2) follows from definitions of density and distribution functions:

$$\int_{-\infty}^x f_X(t) dt = P(X \in (-\infty, x]) = F_X(x).$$

Statement (1) follows from (2) and the Fundamental Theorem of Calculus:

$$\frac{d}{dx}F_X(x) = \frac{d}{dx} \left[\int_{-\infty}^x f_X(t) dt \right] = f_X(x). \quad \square$$

EXAMPLE 8

Suppose X is a cts. r.v. whose distribution function is

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ \sin x & 0 < x \leq \frac{\pi}{2} \\ 1 & x > \frac{\pi}{2} \end{cases}.$$

1. Find a density function of X .
2. Compute $P(X < \frac{\pi}{6})$ using the cdf of X .
3. Compute $P(X < \frac{\pi}{6})$ using a density function of X .

Definition 3.11 Let X be a real-valued r.v. The **survival function** of X is the function $H_X(x) = P(X > x) = 1 - F_X(x)$.

Note: if X is cts, then $H_X(x) = P(X \geq x)$ as well.

EXAMPLE 8

Compute the survival function of X , if $X \sim \text{Unif}([0, 8])$.

3.3 Transformations of random variables

Let $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ be a function and let X be a real-valued r.v. (By the way, φ is “phi”.) Then $Y = \varphi(X)$ is a r.v. which is called a **transformation** of X . The object of this section is to compute the density function of a transformation, given a density function of the original r.v.

When X is discrete

In this situation, $Y = \varphi(X)$ must also be discrete. To compute the density function of Y , first determine the range of Y . Then, for y belonging to the range of Y , start with the definitions as follows:

$$f_Y(y) = P(Y = y) = P(\varphi(X) = y)$$

and then solve the equation inside the parentheses for X . Then use the density function of X to compute probabilities.

EXAMPLE 9

Suppose $X \sim \text{Unif}(\{-2, -1, 0, 1, 2\})$. Let $Y = X^4$. Find a density function of Y .

Remark	Once you have a density function of Y , you can compute any probability associated to Y as described earlier.
---------------	---

When X is continuous

In this situation, $Y = \varphi(X)$ could be discrete, continuous or neither. Since you don't even know that Y has a density function, the best way to proceed is to find the distribution function of Y first. First, determine the range of Y . If this range is $[a, b]$ or (a, b) , you know that

and

Now, let y be in the range of Y . By the definition of distribution function and the definition of φ , we get

$$F_Y(y) = P(Y \leq y) = P(\varphi(X) \leq y).$$

Next, solve the inequality $\varphi(X) \leq y$ for X (this may involve multiple cases) and use either the density or distribution function of X to obtain the cdf of Y . Finally, differentiate F_Y to obtain f_Y . Let's do some examples to see how this works:

EXAMPLE 10

Let X be uniform on $[0, 2]$ and let $Y = X^3$. Find a density function of Y .

EXAMPLE 11

Suppose that an insurance company has to make two kinds of annual payments, “direct” and “indirect”. If X is the size of the direct payment and Y is the size of the indirect payment the company has to make, assume that (X, Y) is modeled by a uniform r.v. on the unit square (this is the square whose vertices are $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$). Find the density function of the total annual payment the insurance company has to make.

EXAMPLE 12

Choose a point (X, Y) uniformly from the rectangle whose vertices are the four points $(1, 0)$, $(1, 1)$, $(4, 0)$ and $(4, 1)$. Let $Z = Y/X$. Find f_Z .

EXAMPLE 13

You and your friend decide to meet at the library to study math. Each of you choose a random time (uniformly and independently) to arrive at the library between 6 and 7 PM. What is the density function of the length of time the first person to arrive has to wait for the second person to arrive?

Definition 3.12 A continuous, real-valued r.v. Y is called **Cauchy** if $Y = \tan X$, for X uniform on $[-\frac{\pi}{2}, \frac{\pi}{2}]$.

The Cauchy r.v. **measures the slope of an angle which is uniformly chosen**, because $\tan \theta$ is the slope of a line at angle θ to the horizontal.

EXAMPLE 14

Find a density function of the Cauchy r.v.

Solution: First, notice that since $X \sim \text{Unif}([-\frac{\pi}{2}, \frac{\pi}{2}])$,

$$f_X(x) = \frac{1}{\frac{\pi}{2} - (-\frac{\pi}{2})} = \frac{1}{\pi}$$

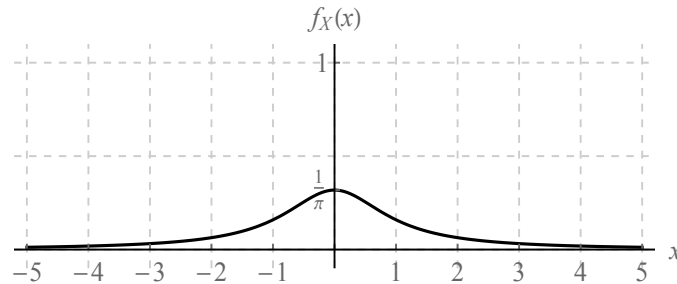
for $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ (and $f_X(x) = 0$ otherwise).

Now, let $Y = \tan X$; the range of Y is \mathbb{R} . For any $y \in \mathbb{R}$,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(\tan X \leq y) \\ &= P(X \leq \arctan y) \\ &= \int_{-\pi/2}^{\arctan y} f_X(x) dx \\ &= \int_{-\pi/2}^{\arctan y} \frac{1}{\pi} dx \\ &= \frac{1}{\pi} \left(\arctan y - \left(-\frac{\pi}{2}\right) \right) \\ &= \frac{1}{\pi} \arctan y + \frac{1}{2}. \end{aligned}$$

Therefore a density function for Y is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} \left[\frac{1}{\pi} \arctan y + \frac{1}{2} \right] =$$



EXAMPLE 15

Let $I \subseteq \mathbb{R}$ be an interval and let $\varphi : I \rightarrow J$ be a strictly increasing function (this means $\varphi(I)$ is an interval and $\varphi^{-1} : \varphi(I) \rightarrow I$ exists). Let X be a cts r.v. taking values only in I which has density function f_X . Find a density function of $Y = \varphi(X)$.

The result of Example 15 is really the proof of this theorem:

Theorem 3.13 (Transformation theorem for increasing φ) *Let $X : \Omega \rightarrow \mathbb{R}$ be a cts r.v. with density f_X , where X takes values only in interval I . Let $Y = \varphi(X)$ where $\varphi : I \rightarrow \mathbb{R}$ is a strictly increasing function. Then*

$$f_Y(y) = \begin{cases} f_X(\varphi^{-1}(y)) \cdot \frac{d}{dy}\varphi^{-1}(y) & \text{if } y \in \varphi(I) \\ 0 & \text{else} \end{cases}$$

A similar theorem holds for decreasing φ :

Theorem 3.14 (Transformation theorem for decreasing φ) *Let $X : \Omega \rightarrow \mathbb{R}$ be a cts r.v. with density f_X , where X takes values only in interval I . Let $Y = \varphi(X)$ where $\varphi : I \rightarrow \mathbb{R}$ is a strictly decreasing function. Then*

$$f_Y(y) = \begin{cases} -f_X(\varphi^{-1}(y)) \cdot \frac{d}{dy}\varphi^{-1}(y) & \text{if } y \in \varphi(I) \\ 0 & \text{else} \end{cases}$$

3.4 Poisson processes

In the last chapter, we discussed Bernoulli processes, which count the number of “successes” occurring when time is kept track of discretely (i.e. in terms of the number of trials that have been performed). In this section we describe a second important type of process, which can be thought of as keeping track of the number of “successes” occurring when time is kept track of continuously (i.e. in terms of elapsed physical time). Such a process is called a Poisson process:

Definition 3.15 Suppose “successes” are occurring at random times in $[0, \infty)$ according to the following rules:

1. the probability of two successes happening simultaneously is zero;
2. the number of successes happening in any interval of time depends only on the length of that interval (and not on the starting point or endpoint of that interval); and
3. the number of successes occurring on any collection of disjoint intervals are mutually independent of one another.

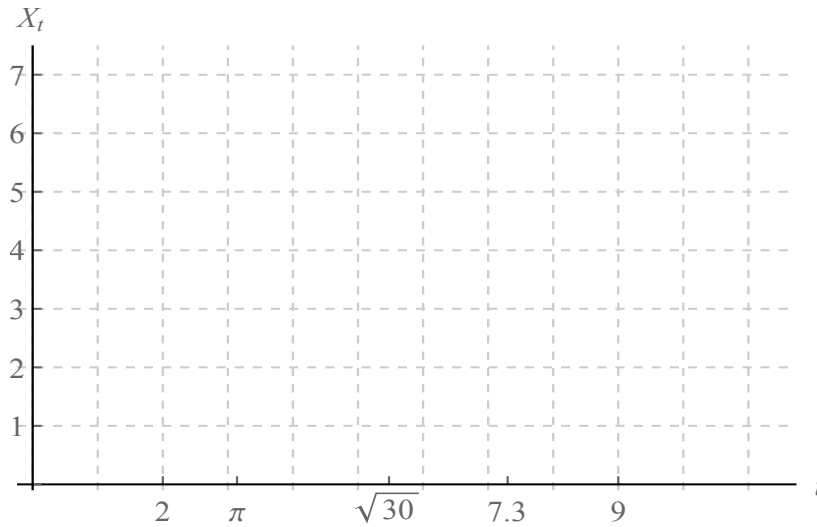
In this setting, if we define X_t to be the number of successes in time interval $[0, t]$, we obtain a continuous-time stochastic process $\{X_t : t \in [0, \infty)\}$ called a **Poisson process**.

Things from the real world that are modeled by Poisson processes include:

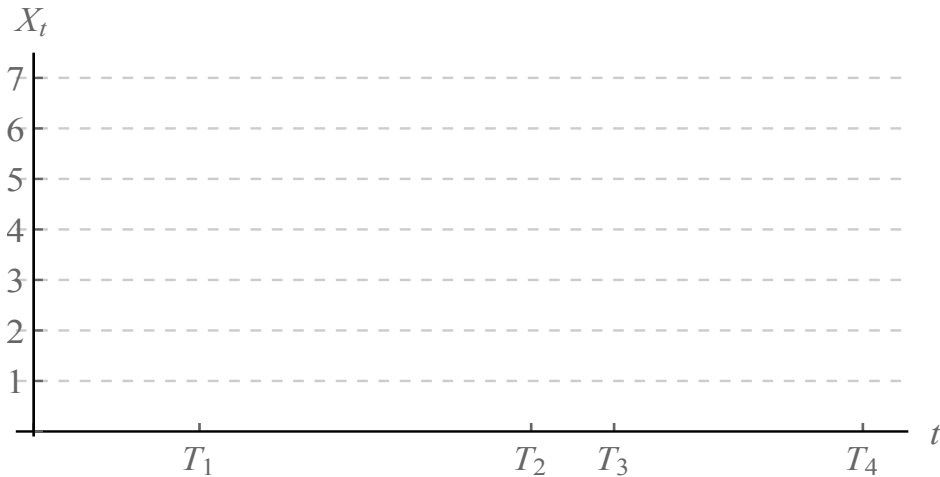
- times of arrivals of customers to a service center;
- times of radioactive emissions;
- times when a cell phone receives a text message;
- times when an earthquake hits the San Andreas Fault;
- times when an error occurs during a transmission.

In all these situations, each time when one of these things occurs is the time of a “success”.

To get a picture in your mind of what a Poisson process “looks like”, suppose successes happen at times $2, \pi, \sqrt{30}, 7.3$. Pictorially, if we graph X_t against t , we get this picture:



More generally, suppose the times that successes occur are (in increasing order) T_1, T_2, T_3, \dots . This produces the following picture, from which we can define random variables associated to the Poisson process:



Definition 3.16 Let $\{X_t\}$ be a Poisson process. For $j = 1, 2, 3, \dots$, define the following r.v.s:

$T_j =$ the j^{th} smallest time at which a success occurs (set $T_0 = 0$)

$W_j = T_j - T_{j-1} =$ the j^{th} **waiting time**
(the time between the $j - 1^{\text{st}}$ and j^{th} successes)

Notice the parallels between these r.v.s and the r.v.s arising from a Bernoulli process:

	Bernoulli process	Poisson process
time measurement	discrete ($t \in \mathbb{N}$)	continuous ($t \in [0, \infty)$)
parameter	success probability p	
distribution of X_t	binomial(t, p)	
time to first success	$\text{Geom}(p)$ (memoryless)	
time to r^{th} success	$NB(r, p)$	

Our goal is to determine the density function for each of the r.v.s associated to a Poisson process. We start with the distribution of the waiting times W_j :

Observations about waiting times:

1. $T_j = W_1 + W_2 + W_3 + \dots + W_j$.
2. If $i \neq j$, the values of W_i and W_j are independent (follows from part 3 of the definition of Poisson process).
3. For any j , the density function of W_j is the same as the density function of any other W_i , hence the same as the density function of W_1 (follows from part 2 of the definition of Poisson process). So we can call each of the waiting times W .
4. W is memoryless (follows from part 2 of the definition of Poisson process).

Recall that a r.v. X is memoryless if for all $m, n \geq 0$, $P(X \geq m + n | X \geq m) = P(X \geq n)$. If X is discrete, we showed that X would have to be geometric. The waiting time W in a Poisson process is memoryless, but is continuous. To classify it, we use the following theorem:

Theorem 3.17 *Let X be a continuous r.v. taking values in $[0, \infty)$ which is memoryless. Then X has density function*

$$f_X(x) =$$

PROOF First, let F_X be the cdf of X and consider the survival function

$$H_X(x) = 1 - F_X(x) = P(X > x) = P(X \geq x).$$

Note that $H_X(x) \in (0, 1)$ so $-\ln H_X(1) > 0$. Let $\lambda = -\ln H_X(1)$ so that $H_X(1) = e^{-\lambda}$.

Since X is memoryless,

$$\begin{aligned} \frac{P(X \geq m + n)}{P(X \geq m)} &= P(X \geq n) \quad \Rightarrow \quad P(X \geq m + n) = P(X \geq m)P(X \geq n) \\ &\Rightarrow \end{aligned}$$

Therefore for any positive integer m ,

$$H_X(m) = H_X(1 + 1 + \dots + 1) = H_X(1)H_X(1) \cdots H_X(1) = [H_X(1)]^m = e^{-\lambda m}.$$

Now for any positive rational number $\frac{m}{n}$,

$$H_X(m) = H_X\left(\frac{m}{n} + \frac{m}{n} + \dots + \frac{m}{n}\right) = H_X\left(\frac{m}{n}\right) H_X\left(\frac{m}{n}\right) \cdots H_X\left(\frac{m}{n}\right) = \left[H_X\left(\frac{m}{n}\right)\right]^n$$

so by taking n^{th} roots of both sides of the above equation we get

$$H_X\left(\frac{m}{n}\right) = \sqrt[n]{H_X(m)} = \sqrt[n]{H_X(1)^m} = [H_X(1)]^{m/n} = e^{-\lambda(m/n)}.$$

Since $H_X(\frac{m}{n}) = e^{-\lambda(m/n)}$ for all rational numbers m/n and since H_X is continuous (because X is cts by hypothesis), it must be that for all **real** numbers x , $H_X(x) = e^{-\lambda x}$. Thus

$$F_X(x) = 1 - H_X(x) = 1 - e^{-\lambda x}$$

and

$$f_X(x) =$$

Definition 3.18 An **exponential** r.v. X with parameter $\lambda \in (0, \infty)$ is a continuous r.v. whose density function is

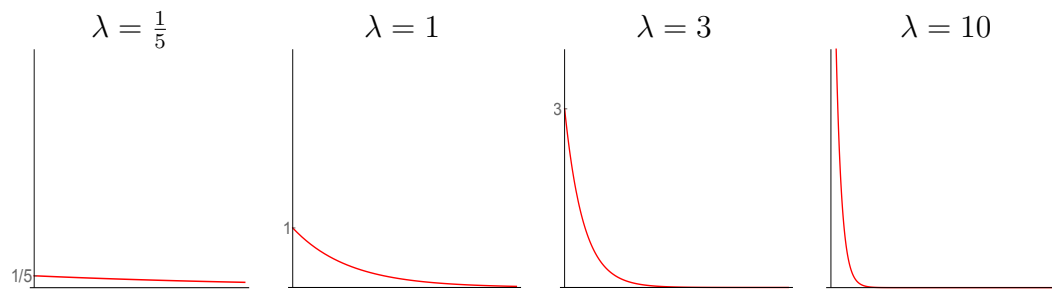
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{else} \end{cases}.$$

An exponential r.v. with parameter λ is denoted $Exp(\lambda)$.

If X is $Exp(\lambda)$, then its cdf is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}.$$

Here are some plots of density functions of $Exp(\lambda)$ r.v.s for various λ :



Thus if X is exponential, we are more likely to get smaller values for X if λ is large, and more likely to get larger values for X if λ is small.

Theorem 3.19 If X is a memoryless, continuous r.v. taking values in $[0, \infty)$, then X is exponential with some parameter $\lambda > 0$.

In the context of our previous reasoning we showed the following useful fact:

Theorem 3.20 If X is $Exp(\lambda)$, then its survival function is $H_X(x) = e^{-\lambda x}$.

Corollary 3.21 (Waiting times are exponential) Let $\{X_t\}$ be a Poisson process. Then there is a number $\lambda > 0$, called the **rate** of the process, such that the waiting times of the process are each $Exp(\lambda)$.

An exponential r.v. with parameter λ **gives the waiting time to the first event in a Poisson process with rate λ .**

Now, we turn our attention to figuring out the density of X_t (for a fixed t). Notice first that each X_t is discrete because it **counts** the number of successes in $[0, t]$.

Recall that as the rate λ increases, the waiting times are likely to get smaller (since they are exponential r.v.s with larger parameters). In fact, something more precise can be said (that we will prove later): in a Poisson process with rate λ , the *average* or *expected* waiting time between successes is $\frac{1}{\lambda}$.

This means that would expect (on the average) to have

successes in the time interval $[0, t]$, so we expect the value of X_t to be λt on the average.

Now, divide $[0, t]$ into n equal-length subintervals. For each subinterval, estimate the probability that there is a success in that subinterval. This probability should be

- the same for each subinterval (by property 2 in the definition of Poisson process),
- and should sum to λt , the expected number of successes in $[0, t]$.

Thus the probability of a success in each subinterval is

Furthermore, the probability of having a success in any one interval is independent of having a success in any other interval (by property 3 in the definition of Poisson process). Last, if n is big enough (see below), there will be at most one success in each subinterval (by property 1 in the definition of Poisson process). So if n is big enough, we can estimate X_t by a binomial r.v. $b(n, \frac{\lambda t}{n})$, i.e.

How big does n have to be? Essentially, ∞ . Thus, for $x \in \{0, 1, 2, \dots\}$,

$$\begin{aligned}
 P(X_t = x) &= \lim_{n \rightarrow \infty} b(n, \frac{\lambda t}{n}, x) \\
 &= \lim_{n \rightarrow \infty} \binom{n}{x} \left(\frac{\lambda t}{n}\right)^x \left(1 - \frac{\lambda t}{n}\right)^{n-x} \\
 &= \frac{(\lambda t)^x}{x!} \lim_{n \rightarrow \infty} \frac{n!}{(n-x)! n^x} \left(1 - \frac{\lambda t}{n}\right)^{-x} \left(1 - \frac{\lambda t}{n}\right)^n \\
 &= \frac{(\lambda t)^x}{x!} \lim_{n \rightarrow \infty} \frac{n(n-1)(n-2) \cdots (n-x+1)}{n^x} \left(1 - \frac{\lambda t}{n}\right)^{-x} \exp \left[n \ln \left(1 - \frac{\lambda t}{n}\right) \right] \\
 &= \frac{(\lambda t)^x}{x!} \lim_{n \rightarrow \infty} \frac{n^x + \text{smaller powers of } n}{n^x} \cdot \left(1 - \frac{\lambda t}{n}\right)^{-x} \cdot \exp \left[\frac{\ln \left(1 - \frac{\lambda t}{n}\right)}{n^{-1}} \right] \\
 &= \frac{(\lambda t)^x}{x!} \lim_{n \rightarrow \infty} \left[\left(1 + n^{\text{negative powers}}\right) \left(1 - \frac{\lambda t}{n}\right)^{-x} \exp \left(\frac{\ln \left(1 - \frac{\lambda t}{n}\right)}{n^{-1}} \right) \right] \\
 &\stackrel{L}{=} \frac{(\lambda t)^x}{x!} (1+0)(1-0)^{-x} \exp \left[\lim_{n \rightarrow \infty} \frac{\frac{1}{1-\frac{\lambda t}{n}} \cdot \frac{\lambda t}{n^2}}{-n^{-2}} \right] \\
 &= \frac{(\lambda t)^x}{x!} \exp \left[\lim_{n \rightarrow \infty} \frac{-\lambda t}{1 - \frac{\lambda t}{n}} \right] \\
 &= \frac{(\lambda t)^x}{x!} \exp \left[\lim_{n \rightarrow \infty} \frac{-\lambda t}{1 - 0} \right] \\
 &= \frac{(\lambda t)^x}{x!} e^{-\lambda t}.
 \end{aligned}$$

Definition 3.22 Let $\lambda \in (0, \infty)$. A **Poisson r.v.**, denoted $Pois(\lambda)$, is a discrete r.v. taking values $\{0, 1, 2, 3, \dots\}$ whose density function is

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

λ is called the **parameter** of the Poisson r.v.

Theorem 3.23 The density function of a Poisson r.v. is in fact a density function (its values sum to 1).

PROOF Apply the formula for the Taylor series of e^λ :

$$\sum_{x=0}^{\infty} f_X(x) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^\lambda = 1. \quad \square$$

Theorem 3.24 Let $\{X_t : t \in [0, \infty)\}$ be a Poisson process with rate λ . Then for each t , X_t is $Pois(\lambda t)$.

A Poisson r.v. with parameter λ **counts the number of events taking place in a Poisson process with rate λ over any one unit of time**. A Poisson r.v. with parameter λt **counts the number of events taking place in a Poisson process with rate λ over any time period of length t** .

In the context of our work, we have established the following relationship between binomial and Poisson r.v.s:

Theorem 3.25 (Law of Small Numbers) $\lim_{n \rightarrow \infty} b(n, \frac{\lambda}{n}) = Pois(\lambda)$. Restated, this means that for any $x \in \{0, 1, 2, 3, \dots\}$,

$$\lim_{n \rightarrow \infty} b\left(n, \frac{\lambda}{n}, x\right) = \frac{e^{-\lambda} \lambda^x}{x!}.$$

This law says that if you perform more and more trials in a Bernoulli experiment, but simultaneously lower the probability of success on each trial so that the expected number of successes is constant, you achieve a Poisson r.v. in the limit.

The last r.v. associated to a Poisson process whose density we need to find is the time T_r to the r^{th} success in a Poisson process. We start by noting that the range of $T = T_r$ is $[0, \infty)$; next we compute its distribution function. Let $t \in [0, \infty)$. Then

$$\begin{aligned} F_T(t) &= P(T \leq t) = P(X_t \geq r) \\ &= 1 - P(X_t < r) \\ &= 1 - P(\text{Pois}(\lambda t) < r) \\ &= 1 - \sum_{x=0}^{r-1} \frac{e^{-\lambda t} (\lambda t)^x}{x!}. \end{aligned}$$

$$\begin{aligned} \Rightarrow f_T(t) &= \frac{d}{dt} F_T(t) = \frac{d}{dt} \left[1 - \sum_{x=0}^{r-1} \frac{e^{-\lambda t} (\lambda t)^x}{x!} \right] \\ &= - \sum_{x=0}^{r-1} \frac{\lambda^x}{x!} \frac{d}{dt} [e^{-\lambda t} t^x] \\ &= - \sum_{x=0}^{r-1} \frac{\lambda^x}{x!} \frac{d}{dt} [e^{-\lambda t} e^{x \ln t}] \\ &= - \sum_{x=0}^{r-1} \frac{\lambda^x}{x!} \frac{d}{dt} [e^{x \ln t - \lambda t}] \\ &= - \sum_{x=0}^{r-1} \frac{\lambda^x}{x!} [e^{x \ln t - \lambda t}] \left(\frac{x}{t} - \lambda \right) \\ &= \sum_{x=0}^{r-1} \frac{\lambda^x}{x!} [e^{x \ln t - \lambda t}] \left(\lambda - \frac{x}{t} \right) \\ &= \sum_{x=0}^{r-1} \frac{\lambda^{x+1}}{x!} t^x e^{-\lambda t} - \sum_{x=0}^{r-1} \frac{\lambda^x}{(x-1)!} t^{x-1} e^{-\lambda t} \\ &= \sum_{x=1}^r \frac{\lambda^x}{(x-1)!} t^{x-1} e^{-\lambda t} - \sum_{x=0}^{r-1} \frac{\lambda^x}{(x-1)!} t^{x-1} e^{-\lambda t} \\ &= \frac{\lambda^r}{(r-1)!} t^{r-1} e^{-\lambda t}. \end{aligned}$$

Definition 3.26 Let $\lambda \in (0, \infty)$ and let $r \in \{1, 2, 3, \dots\}$. A **gamma** r.v., denoted $\Gamma(r, \lambda)$, is a cts r.v. X taking values in $[0, \infty)$ whose density function is

$$f_X(x) = \begin{cases} \frac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}.$$

r and λ are called the **parameters** of the gamma r.v.

We will prove that f_X is actually a density function later.

Theorem 3.27 *Let $\{X_t : t \in [0, \infty)\}$ be a Poisson process with rate λ . Then for each $r \in \{1, 2, 3, \dots\}$, T_r (the time to the r^{th} success) is $\Gamma(r, \lambda)$.*

A $\Gamma(r, \lambda)$ r.v. **measures the time until the r^{th} event in a Poisson process with rate λ .**

Problems with r.v.s related to Poisson processes

EXAMPLE 16

The number of people in a community who live to 100 years of age is a Poisson r.v. with parameter 6. Find the probability that exactly 4 people live to 100, and the probability that at least 2 people live to 100.

EXAMPLE 17

The time (in hours) it takes to repair a machine is an exponential r.v. with parameter $\frac{1}{2}$. Find the probability that the repair time is at least 2 hours.

EXAMPLE 18

Suppose X is exponential with parameter 4. Let $Y = X^2$; find the density function of Y .

EXAMPLE 19

Suppose that hits to a certain website follows a Poisson process with rate 200.

1. What is the probability there are exactly 630 hits in the first 3 units of time?
2. Suppose there is a hit at time 10. What is the probability that there are no hits between times 10 and 11?
3. Write the density function of the r.v. measuring the time to the fifth hit.

3.5 More on gamma distributions

The gamma function

We begin this question by trying to determine the value of $n!$ when n is not a whole number. For example, what is $\frac{1}{2}!$? What is $\pi!$?

More precisely, we seek a function $f : \mathbb{R} \rightarrow \mathbb{R}$ (or at least $f : [0, \infty) \rightarrow \mathbb{R}$) with the following properties:

1. $f(n) = n!$ for all $n \in \{0, 1, 2, \dots\}$;
2. f is continuous;
3. $xf(x-1) = f(x)$ for all x

Such an f would be a “continuous version of factorial”.

To do this, we will start by trying to incorporate (3) above via some creative integration by parts. Our attempt will be slightly off, but “close enough”.

Definition 3.28 *The gamma function is the function $\Gamma : (0, \infty) \rightarrow \mathbb{R}$ defined by*

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx.$$

It turns out that $\Gamma(r) = \frac{1}{r} \prod_{n=1}^{\infty} \frac{(1+\frac{1}{n})^r}{1+\frac{r}{n}}$ (this isn’t relevant to MATH 414 or 416).

Theorem 3.29 (Properties of the gamma function) *Let Γ be the gamma function. Then:*

1. $\Gamma : (0, \infty) \rightarrow \mathbb{R}$ is continuous.
2. $\Gamma(1) = 1$.
3. For every $r > 0$, $\Gamma(r+1) = r\Gamma(r)$.
4. For every $r > 1$, $\Gamma(r) = (r-1)\Gamma(r-1)$.
5. For $n \in \{1, 2, 3, \dots\}$, $\Gamma(n) = (n-1)!$.
6. For every $n \in \mathbb{N}$, $n! = \Gamma(n+1)$.

PROOF (1) All functions which are defined as integrals are cts by the Fund. Thm. of Calculus.

(2) $\Gamma(1) = \int_0^\infty e^{-x} dx = -e^{-x} \Big|_0^\infty = 0 - (-1) = 1.$

(4) follows from (3) by replacing all the r s with $r - 1$;

(5) follows from (2) and repeated application of (4);

(6) follows from (5) by replacing the n s with $n + 1$. That leaves (3).

To establish (3) use integration by parts with $u = x^r$ and $dv = e^{-x} dx$:

We can now extend the definition of gamma random variables to the situation where r is not necessarily a whole number:

Definition 3.30 Let $\lambda \in (0, \infty)$ and let $r \in (0, \infty)$. A **gamma r.v.**, denoted $\Gamma(r, \lambda)$, is a cts r.v. X taking values in $[0, \infty)$ whose density function is

$$f_X(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}.$$

r and λ are called the **parameters** of the gamma r.v.

Note: A $\Gamma(1, \lambda)$ r.v. is the same thing as an $\text{Exp}(\lambda)$ r.v.

Theorem 3.31 *The density function of a $\Gamma(r, \lambda)$ r.v. is in fact a density function.*

PROOF Perform the u -substitution $u = \lambda x$; $du = \lambda dx$ inside the integral:

$$\int_0^\infty f_X(x) dx = \int_0^\infty \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} dx =$$

Corollary 3.32 (Gamma Integral Formula) *Let $r, \lambda > 0$. Then:*

$$\int_0^\infty x^{r-1} e^{-\lambda x} dx = \frac{\Gamma(r)}{\lambda^r}$$

3.6 Normal random variables

There is one last common class of continuous random variable to discuss. This class arises when studying the average of a large amount of data. More precisely:

Definition 3.33 A discrete-time stochastic process $\{X_n : n \in \{1, 2, 3, \dots\}\}$ is called an **i.i.d. process** (or an **i.i.d. sequence of r.v.s**) if

1. the r.v.s are independent, i.e. $X_i \perp X_j$ whenever $i \neq j$; and
2. the r.v.s are identically distributed, i.e. the density function of X_j is the same for all j .

Definition 3.34 Given an i.i.d. process $\{X_n\}$, we define the **sequence of averages** of the process by defining for each $n \in \{1, 2, 3, \dots\}$

$$A_n = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

(Each A_n is a r.v., so $\{A_n\}$ is a discrete-time stochastic process.)

For example, suppose we have an i.i.d. process where each X_n is uniform on $\{1, 2, 3, 4, 5, 6\}$. This process models repeated rolling of a fair die. The r.v.s A_n keep track of the average of the first n die rolls; for example if the sequence of die rolls is

3, 5, 1, 3, 2, 6, 4, 1, 1, ...

then the corresponding sequence of averages is

A theorem which will be proven in Chapter 8 (the Central Limit Theorem) says that if $\{X_n\}$ is an i.i.d. process, then (almost) no matter what the distribution of the X_n is, then for large n the distribution of the average A_n always becomes well approximated by one of a class of r.v.s called “normal” r.v.s.

To define this class, start by letting $g(x) = e^{-x^2/2} = \exp\left(\frac{-x^2}{2}\right)$.

QUESTION

Is $g(x) = e^{-x^2/2}$ a density function? (In other words, is $\int_{-\infty}^{\infty} g(x) dx = 1$?)

$$\left[\int_{-\infty}^{\infty} g(x) dx \right]^2 = \left[\int_{-\infty}^{\infty} \exp\left(\frac{-x^2}{2}\right) dx \right] \left[\int_{-\infty}^{\infty} \exp\left(\frac{-y^2}{2}\right) dy \right]$$

Definition 3.35 The **standard normal** r.v., abbreviated $n(0, 1)$ or $\mathcal{N}(0, 1)$ or Z , is the continuous r.v. whose density function is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right).$$

The cumulative distribution function of the standard normal r.v. is denoted Φ :

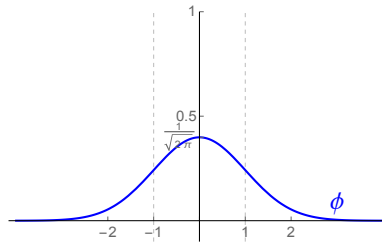
$$\Phi(x) = P(n(0, 1) \leq x) = \int_{-\infty}^x \phi(t) dt.$$

There is no better formula for Φ ; values of Φ are estimated using a calculator or tables (one such table can be found in Appendix A.3 of these notes). On exams in Math 414, we will leave answers to questions in terms of Φ .

Theorem 3.36 (Properties of the standard normal density) Let ϕ be the density of the standard normal r.v.. Then

1. $\phi(-x) = \phi(x)$.
2. $\phi(0) = \frac{1}{\sqrt{2\pi}}$.
3. $\int_{-\infty}^{\infty} \phi(x) = 1$.
4. ϕ is increasing on $(-\infty, 0)$ and decreasing on $(0, \infty)$.
5. ϕ is concave down on $(-1, 1)$ and concave up on $(-\infty, -1) \cup (1, \infty)$.
6. $\lim_{x \rightarrow \infty} \phi(x) = \lim_{x \rightarrow -\infty} \phi(x) = 0$.

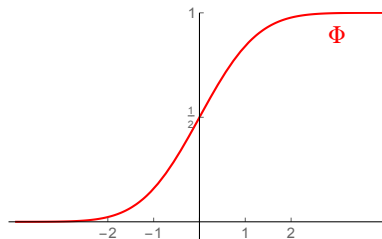
These properties imply that the graph of ϕ looks like a "bell curve":



Theorem 3.37 (Properties of Φ) Let Φ be the cdf of the standard normal. Then:

1. Φ is a cdf (so it has properties common to all cdfs);
2. Φ is continuous;
3. Φ is differentiable and $\Phi' = \phi$;
4. $\Phi(0) = \frac{1}{2}$;
5. For all x , $\Phi(-x) = 1 - \Phi(x)$.
6. For all x , $P(|N(0, 1)| \leq x) = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1$.

The properties above indicate that the graph of Φ looks like



Linear transformations of the standard normal

Let $Z \sim n(0, 1)$ and let $\mu \in \mathbb{R}$, $\sigma \in (0, \infty)$. Define $X = \sigma Z + \mu$. Then

$$F_X(x) =$$

and therefore

$$f_X(x) =$$

Definition 3.38 A random variable X is called **normal** with parameters μ and σ^2 (denoted $n(\mu, \sigma^2)$ or $\mathcal{N}(\mu, \sigma^2)$) if it is a continuous r.v. with density function

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

In this setting μ is called the **mean** and σ^2 is called the **variance**.

We will see in Chapter 8 that normal r.v.s **approximate the average of a bunch of i.i.d. r.v.s, no matter what distribution the individual r.v.s have.**

Theorem 3.39 (Characterization of normal r.v.s) Let X be $n(\mu, \sigma^2)$. Then:

1. $X = \mu + \sigma Z$ where $Z \sim n(0, 1)$.
2. $F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$.

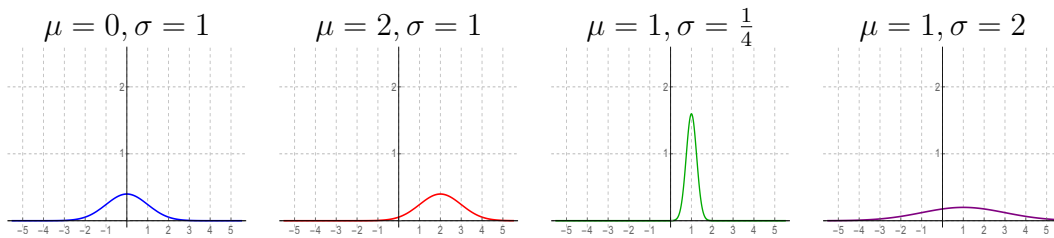
Theorem 3.40 (Linear transformations on normal r.v.s) Suppose $X \sim n(\mu, \sigma^2)$. Let $Y = a + bX$ where $a, b \in \mathbb{R}$. Then $Y \sim n(a + b\mu, b^2\sigma^2)$.

PROOF Let Z be $n(0, 1)$. Since X is $n(\mu, \sigma^2)$, we see by Theorem 3.39 that $X = \mu + \sigma Z$ so $Y = a + bX = a + b(\mu + \sigma Z) = (a + b\mu) + (b\sigma)Z$. Thus by Theorem 3.39, Y is $n(a + b\mu, b^2\sigma^2)$. \square

EXAMPLE 20

Suppose X is normal with mean 20 and variance 36. Find, in terms of Φ , the probability that $12 < X \leq 20$.

Here are some plots of density functions for various values of μ and σ :



In general, the graph of the density function of any normal r.v. is a “bell curve” which has its peak at μ and inflection points at $\mu \pm \sigma$ (HW). This means that if σ is small, then the function has a tall, skinny peak (meaning that X takes values close to μ with very high probability) and if σ is large, the function has a short, wide peak (meaning that the values of X are more spread out).

Normal random variables arise naturally as averages of i.i.d. processes; examples of data which can be assumed to be normally distributed include:

1. Heights of people;
2. Exam grades;
3. Velocities of gas particles (Maxwell’s Law);
4. Measurement errors in lab experiments;
5. The change in the price of a stock over a fixed period of time.

Normal r.v.s have this interesting connection with gamma r.v.s:

Theorem 3.41 *Let $X \sim n(0, \sigma^2)$ and let $Y = X^2$. Then Y is $\Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$.*

PROOF Y has range $[0, \infty)$; let $y \geq 0$. Then

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= \Phi\left(\frac{\sqrt{y}}{\sigma}\right) - \Phi\left(\frac{-\sqrt{y}}{\sigma}\right) \\ &= \Phi\left(\frac{\sqrt{y}}{\sigma}\right) - \left[1 - \Phi\left(\frac{\sqrt{y}}{\sigma}\right)\right] \\ &= 2\Phi\left(\frac{\sqrt{y}}{\sigma}\right) - 1. \end{aligned}$$

Therefore

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} \left[2\Phi\left(\frac{\sqrt{y}}{\sigma}\right) - 1 \right] = 2\phi\left(\frac{\sqrt{y}}{\sigma}\right) \cdot \frac{1}{\sigma 2\sqrt{y}} \\ &= \frac{1}{\sigma\sqrt{y}} \cdot \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-y}{2\sigma^2}\right] \\ &= \frac{\left(\frac{1}{2\sigma^2}\right)^{1/2}}{\sqrt{\pi}} y^{\frac{1}{2}-1} e^{-(1/2\sigma^2)y} \end{aligned}$$

Now the density of a $\Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$ r.v. is

$$f(y) = \frac{\left(\frac{1}{2\sigma^2}\right)^{1/2}}{\Gamma\left(\frac{1}{2}\right)} y^{\frac{1}{2}-1} e^{-(1/2\sigma^2)y}$$

Corollary 3.42 $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

3.7 Stirling's formula

Theorem 3.43 (Stirling's Formula)

$$\lim_{n \rightarrow \infty} \frac{n!}{n^n e^{-n} \sqrt{2\pi n}} = 1$$

(More generally, $\lim_{n \rightarrow \infty} \frac{\Gamma(n+1)}{n^n e^{-n} \sqrt{2\pi n}} = 1$).

Consequence: For large n , $n! = \Gamma(n+1)$ is approximately equal to $n^n e^{-n} \sqrt{2\pi n}$ and in many instances (i.e. proofs), $n!$ can be replaced with $n^n e^{-n} \sqrt{2\pi n}$ without a problem.

PROOF (OF STIRLING'S FORMULA) Define $\psi : \mathbb{R} \rightarrow \mathbb{R}$ by

$$\psi(x) = \begin{cases} \frac{2}{x^2}(e^x - 1 - x) & \text{if } x \neq 0 \\ 1 & \text{if } x = 0 \end{cases}$$

Observe $\lim_{x \rightarrow 0} \psi(x) = 1$ so ψ is everywhere continuous.

Next, define $f : [0, 1] \rightarrow \mathbb{R}$ by

$$f(t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp[-x^2 \psi(xt)] dx.$$

(this integral does in fact converge, and f is continuous by the FTC).

We will prove Stirling's formula by computing $f(0)$ two different ways. First, we compute it directly, by recognizing its value as the integral of a density function:

$$\begin{aligned} f(0) &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp[-x^2 \psi(0)] dx \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp[-x^2(1)] dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\frac{1}{2}} \sqrt{2\pi}} \exp\left[\frac{-x^2}{2\left(\sqrt{\frac{1}{2}}\right)^2}\right] dx \\ &= \int_{-\infty}^{\infty} f_{n(0, \sqrt{\frac{1}{2}})}(x) dx \\ &= 1. \end{aligned}$$

Now we compute $f(0)$ a different way. This involves a lot of symbol-crunching:

$$\begin{aligned}
f(t) &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp \left[-x^2 \psi(xt) \right] dx \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp \left[-x^2 \left(\frac{2}{x^2 t^2} \right) (e^{xt} - 1 - xt) \right] dx \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp \left(\frac{-2}{t^2} e^{xt} \right) \exp \left(\frac{2}{t^2} \right) \exp \left(\frac{2x}{t} \right) dx
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sqrt{\pi}} \exp \left(\frac{2}{t^2} \right) \int_0^{\infty} e^{-u} \left(\frac{t^2 u}{2} \right)^{2/t^2} \frac{1}{tu} du \\
&= \frac{1}{t\sqrt{\pi}} \exp \left(\frac{2}{t^2} \right) \left(\frac{t^2}{2} \right)^{2/t^2} \int_0^{\infty} e^{-u} u^{(\frac{2}{t^2}-1)} du \\
&= \frac{1}{t\sqrt{\pi}} \exp \left(\frac{2}{t^2} \right) \left(\frac{t^2}{2} \right)^{2/t^2} \Gamma \left(\frac{2}{t^2} \right)
\end{aligned}$$

$$f(t) = f \left(\sqrt{\frac{2}{n}} \right) = \frac{1}{\sqrt{\frac{2\pi}{n}}} e^n \left(\frac{1}{n} \right)^n \Gamma(n) = \frac{1}{\sqrt{\frac{2\pi}{n}}} e^n n^{-n} \frac{\Gamma(n+1)}{n} = \frac{\Gamma(n+1)}{n^n e^{-n} \sqrt{2\pi n}}$$

3.8 Summary of Chapter 3

- A continuous random variable is a function $X : \Omega \rightarrow \mathbb{R}^d$ such that the probability of any individual value of X is zero.
- We usually describe continuous r.v.s by specifying a density function $f_X : \mathbb{R}^d \rightarrow [0, \infty)$, which satisfies

$$P(X \in E) = \int_E f_X(x) dx$$

for any set E . Such a function must be everywhere nonnegative and must integrate to 1. We compute probabilities associated to a cts r.v. by integrating the density function as above.

- All real-valued r.v.s can be described by giving a distribution function $F_X : \mathbb{R} \rightarrow [0, 1]$ defined by

$$F_X(x) = P(X \leq x).$$

Distribution functions have many properties; notably X is cts if and only if F_X is cts; and if X is cts with density f_X ,

$$f_X(x) = \frac{d}{dx} F_X(x) \quad \text{and} \quad F_X(x) = \int_{-\infty}^x f_X(t) dt.$$

- To find the density function of a continuous transformation Y of a continuous r.v. X , first find the range of Y , then compute F_Y by back-substitution. Last, differentiate F_Y to get f_Y .
- Classes of commonly encountered continuous random variables include the following:
 1. uniform r.v.s, which assign relatively equal likelihood to all values in the range of X ;
 2. exponential r.v.s, which measure the amount of time until a success happens in a Poisson process (and are the only memoryless cts r.v.s);
 3. gamma r.v.s, which measure the amount of time until the r^{th} success of a Poisson process;
 4. normal r.v.s, which approximate averages of i.i.d. sequences of random variables.
 5. the Cauchy r.v., which gives the tangent of a uniformly chosen angle.

You should know the range, distribution and density function of each of these common r.v.s, and additional facts relevant to each class.

- One additional class of discrete r.v.s not previously encountered are Poisson r.v.s, which count the number of successes over a fixed length of time in a Poisson process.

- The gamma function Γ , defined by

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx,$$

extends the idea of factorial to positive real numbers (for $n \in \mathbb{N}$, $n! = \Gamma(n+1)$).

- Stirling's Formula says that $n! = \Gamma(n+1) \approx n^n e^{-n} \sqrt{2\pi n}$ for large n .

Chapter 4

Discrete joint distributions

4.1 Basic examples

Suppose that in a probabilistic experiment you are taking more than one measurement, say d distinct (real-valued) random variables. Often, the right way to think of these d quantities is as a single random variable which takes values in \mathbb{R}^d .

EXAMPLE 1

Pick a sample of 6 marbles (simultaneously) from an urn with 10 red, 12 blue, 18 black and 20 green marbles in it. Let

$X_1 = \#$ of red marbles drawn

$X_2 = \#$ of blue marbles drawn

$X_3 = \#$ of black marbles drawn

$X_4 = \#$ of green marbles drawn

Obtain $\mathbf{X} = \vec{X} = X = (X_1, X_2, X_3, X_4) : \Omega \rightarrow \mathbb{R}^4$ (discrete, 4-diml r.v.)

EXAMPLE 2

Pick a point uniformly from the unit square. Let

$X = x$ – coordinate of the chosen point

$Y = y$ – coordinate of the chosen point

Obtain $\mathbf{X} = \vec{X} = (X, Y) : \Omega \rightarrow \mathbb{R}^2$ (cts, 2-diml r.v.)

EXAMPLE 3

Pick a point uniformly from the triangle whose vertices are $(0, 0)$, $(0, 2)$ and $(4, 0)$.
Let

$X = x$ – coordinate of the chosen point

$Y = y$ – coordinate of the chosen point

Obtain $\mathbf{X} = \vec{X} = (X, Y) : \Omega \rightarrow \mathbb{R}^2$ (cts, 2–dim'l r.v.)

Notice: In Example 2, you obtain no information about either X or Y when you are told the value of the other coordinate. This is not the case in Examples 1 and 3; as you learn information about one or more coordinates, your belief about the values of the remaining coordinates changes.

Definition 4.1 A d –**dimensional random variable** (a.k.a. d –**dimensional random vector**) is a random variable whose range is a subset of \mathbb{R}^d . We denote such a r.v. by \mathbf{X} or X or \vec{X} .

As with real-valued discrete r.v.s, a discrete d –dim'l r.v. is determined by a density function

$$f_{\mathbf{X}} = f_X = f_{\vec{X}} : \mathbb{R}^d \rightarrow [0, \infty)$$

satisfying

$$f_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) \text{ for all } \mathbf{x} \in \mathbb{R}^d$$

and

$$P(\mathbf{X} \in E) = \sum_{\mathbf{x} \in E} f_{\mathbf{X}}(\mathbf{x})$$

for any event E .

Definition 4.2 Let $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ be a discrete d –dim'l r.v. with density function $f_{\mathbf{X}}$. The coordinates X_1, X_2, \dots, X_d of X are called its **marginals**, and any such \mathbf{X} is called a **joint distribution** of its marginals. $f_{\mathbf{X}}$ is called the **joint density (function)** of \mathbf{X} .

Note: Given a bunch of marginals X_1, \dots, X_d , one can construct lots of different joint distributions \mathbf{X} of those marginals (see the next two examples).

Theorem 4.3 (Density function of marginals, discrete case) *Let $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ be a discrete d -dim'l r.v. with density function $f_{\mathbf{X}}$. Then the density function of the j^{th} marginal X_j is*

$$f_{X_j}(x) = P(X_j = x) = \sum_{\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}_j = x\}} f_{\mathbf{X}}(\mathbf{x}).$$

In other words, this theorem says that to find the density function of a marginal, you **add up the values of the joint density over all the coordinates other than the marginal you want**. As a special case, given a two-dimensional joint density $f_{X,Y}$,

$$f_X(x) = \sum_y f_{X,Y}(x, y) \quad \text{and} \quad f_Y(y) = \sum_x f_{X,Y}(x, y).$$

EXAMPLE 4

Independently roll a fair die and flip a coin. Let X record the number on the die and let Y record 0 for tails and 1 for heads. Describe the joint density of X and Y , and the marginals.

EXAMPLE 5

Roll a fair die and flip a coin, with the assumption that the coin “knows” what number is rolled, i.e. if you roll an even number then the coin flips heads with probability $2/3$ and if you roll an odd number then the coin flips heads with probability $1/3$. Let X record the number on the die and let Y record 0 for tails and 1 for heads. Describe the joint density of X and Y , and the marginals.

EXAMPLE 6

Draw 4 balls without replacement from an urn with 15 green and 5 black balls in it. Let X and Y be the number of green and black balls drawn, respectively. Describe the joint density of X and Y , and the marginals.

EXAMPLE 7

1000 people are surveyed, and the results are summarized in the following table:

	SMOKERS	NON-SMOKERS
MEN	10%	38%
WOMEN	18%	34%

For each question, give the correct notation for what the question is asking, and answer the question.

1. What % of those surveyed are men?
2. What is the probability that a surveyed female smokes?
3. What is the probability that a given non-smoker is a male?

4.2 Multinomial and hypergeometric distributions

Consider a Bernoulli experiment consisting of n repetitions of a trial with d different outcomes. On each trial, the probability of outcome j is p_j (so $p_1 + \dots + p_d = 1$). Assume that the results of the trials are independent of one another, and that the probabilities p_1, \dots, p_d do not change from trial to trial.

Let $\mathbf{X} = (X_1, \dots, X_d)$ be the joint distribution which satisfies

$$X_j = \#(\text{trials resulting in outcome } j)$$

QUESTION 1

What is the j^{th} marginal of \mathbf{X} ?

QUESTION 2

What is the joint density $f_{\mathbf{X}}$?

4.2. Multinomial and hypergeometric distributions

Definition 4.4 Let $n \in \mathbb{N}$ and let $p_1, \dots, p_d \geq 0$ be such that $\sum_j p_j = 1$. A discrete joint distribution $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ is said to be **multinomial** with parameters n and (p_1, \dots, p_d) if it has joint density

$$f_{\mathbf{X}}(x_1, \dots, x_d) = \binom{n}{x_1, x_2, \dots, x_d} \prod_{j=1}^d p_j^{x_j} = \frac{n!}{x_1! x_2! \cdots x_d!} p_1^{x_1} p_2^{x_2} \cdots p_d^{x_d}$$

(for nonnegative integers x_1, \dots, x_d satisfying $\sum_{j=1}^d x_j = n$).

The multinomial distribution describes **sampling with replacement**, as in problems like Example 8 below.

EXAMPLE 8

Suppose a jar has 10 red marbles, 30 blue marbles and 40 green marbles. Draw 13 marbles from the jar, one at a time with replacement. What is the probability that you draw 2 red, 4 blue and 7 green marbles?

Note: The j^{th} marginal of a multinomial($n, (p_1, \dots, p_d)$) distribution is binomial(n, p_j).

What if you sample without replacement? This was discussed in Chapter 2; we rephrase the information there in the language of joint distributions:

Definition 4.5 Let $n \in \mathbb{N}$ and let $n_1, \dots, n_d \in \mathbb{N}$ be such that $\sum_j n_j = n$. Let $k \leq n$. A discrete joint distribution $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ is called **hypergeometric** (or d -**dim'l hypergeometric**) if it has density function

$$f_{\mathbf{X}}(x_1, \dots, x_d) = \begin{cases} \frac{\binom{n_1}{x_1} \binom{n_2}{x_2} \cdots \binom{n_d}{x_d}}{\binom{n}{k}} & \text{for } (x_1, \dots, x_d) \in \mathbb{N}^d \text{ satisfying } \sum_{j=1}^d x_j = k \\ 0 & \text{else} \end{cases}$$

d -dimensional hypergeometric r.v.s model the situation where you have n_j objects of type j in a jar (for a total of n objects) and you draw k objects **without replacement**. If you let X_j be the number of objects of type j you draw, then $\mathbf{X} = (X_1, \dots, X_d)$ is hypergeometric with the above density.

4.3 Independence of discrete random variables

Definition 4.6 Let X_1, \dots, X_d be discrete, real-valued r.v.s with joint distribution \mathbf{X} . The r.v.s (just as well, the distribution) are (is) called **(mutually) independent** if

$$P(X_1 = x_1, X_2 = x_2, \dots, X_d = x_d) = P(X_1 = x_1) P(X_2 = x_2) \cdots P(X_d = x_d)$$

i.e.

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^d f_{X_j}(x_j)$$

for all $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$.

EXAMPLES

In example 4 (page 108),

In example 5 (page 109),

Heuristic idea: To say a collection of random variables are independent means that knowledge about any one or ones of the r.v.s does not affect any probabilities associated to the others.

4.4 Transformation problems with joint discrete distributions

EXAMPLE 9

Suppose X and Y are integer-valued r.v.s with joint density

$$f_{X,Y}(x,y) = \begin{cases} \frac{c}{x^2 y^2} & \text{if } 1 \leq x, 1 \leq y \\ 0 & \text{else} \end{cases}$$

where c is a constant. Find (in terms of c) the density function of $W = X + Y$.

EXAMPLE 10

Suppose X and Y are independent geometric r.v.s (where X has parameter p and Y has parameter q). Find the density of $Z = \min(X, Y)$.

4.4. Transformation problems with joint discrete distributions

EXAMPLE 11

Suppose $X \sim \text{Geom}(p)$. Find the density of $X + X$.

Chapter 5

Continuous joint distributions

5.1 Definitions and elementary properties

In this section we take the usual language associated to non-discrete, real-valued r.v.s and extend it to joint distributions.

As usual, given 2 real-valued r.v.s X and Y , we think of $\mathbf{X} = (X, Y) : \Omega \rightarrow \mathbb{R}^2$.

(Similarly, write $\mathbf{X} = (X_1, \dots, X_d) : \Omega \rightarrow \mathbb{R}^d$.)

Joint distribution functions

<u>DIMENSION</u>	<u>DEFINITION OF DIST. FUNCTION</u>	<u>APPLICATION TO PROBABILITIES</u>
$d = 1$ $(X : \Omega \rightarrow \mathbb{R})$	$F_X : \mathbb{R} \rightarrow [0, 1]$ $F_X(x) = P(X \leq x)$	$P(a < X \leq b) =$ $F_X(b) - F_X(a)$
$d = 2$ $(\mathbf{X} : \Omega \rightarrow \mathbb{R}^2)$ $(\mathbf{X} = (X, Y))$		
general d $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$		

Moral

Distribution functions are not as useful for joint distributions as they are for real-valued r.v.s.

Theorem 5.1 (Properties of joint distribution functions) Let $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ be a joint distribution with joint cdf $F_{\mathbf{X}} : \mathbb{R}^d \rightarrow [0, 1]$. Then:

1. $\lim_{x_j \rightarrow \infty \forall j} F_{\mathbf{X}}(\mathbf{x}) = 1.$
2. $\lim_{x_j \rightarrow -\infty \forall j} F_{\mathbf{X}}(\mathbf{x}) = 0.$
3. If all but one coordinate is fixed, $F_{\mathbf{X}}$ is increasing with respect to that coordinate.

Marginals

As with the discrete case, the coordinates of a joint non-discrete r.v. are called its **marginals**. The marginals are described by giving their distribution functions:

Theorem 5.2 (Distribution functions of marginals) *Let $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ be a joint distribution with joint cdf $F_{\mathbf{X}} : \mathbb{R}^d \rightarrow [0, 1]$. Then the cdf F_{X_j} of the j^{th} marginal X_j is*

$$F_{X_j}(x_j) = P(X_j \leq x_j) = \lim_{x_i \rightarrow \infty \forall i \neq j} F_{\mathbf{X}}(x_1, \dots, x_d).$$

PROOF

As a special case, given joint distribution (X, Y) with joint cdf $F_{X,Y}$, we have

$$F_X(x) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y) \quad \text{and} \quad F_Y(y) = \lim_{x \rightarrow \infty} F_{X,Y}(x, y).$$

Joint density functions for cts r.v.s

Recall: A r.v. $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ is called **continuous** if $P(\mathbf{X} = \mathbf{x}) = 0$ for every $\mathbf{x} \in \mathbb{R}^d$.

When $d = 1$, most cts r.v.s have a density function which is used to compute probabilities: if $X : \Omega \rightarrow \mathbb{R}$ with density function f_X , then

Definition 5.3 *Let $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ be a r.v. We say a function $f_{\mathbf{X}} : \mathbb{R}^d \rightarrow [0, \infty)$ is a **(joint) density function** for \mathbf{X} if for every subset $E \subseteq \mathbb{R}^d$ whose size (i.e. length/area/volume/etc.) can be computed using calculus,*

$$P(\mathbf{X} \in E) = \int_E f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

Note: The integral in the above definition is really a multiple integral (a double integral if $d = 2$, a triple integral if $d = 3$, etc.).

Note: Density functions for a specific cts joint distribution X are not unique (they can be changed at single points, etc. without affecting probability computations).

Theorem 5.4 (Properties of joint density functions) *Let $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ be a d -dimensional r.v.*

1. *If \mathbf{X} is not cts but has a density, then \mathbf{X} must be cts.*
2. *A (measurable) function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is the density function of a cts joint distribution \mathbf{X} if and only if*
 - (i) *$f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$ and*
 - (ii) *$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) d\mathbf{x} = 1$.*
3. *Suppose continuous $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ has joint distribution function $F_{\mathbf{X}}$ and joint density function $f_{\mathbf{X}}$. Then for all $\mathbf{x} \in \mathbb{R}^d$,*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^d}{\partial x_1 \partial x_2 \cdots \partial x_d} F_{\mathbf{X}}(\mathbf{x}).$$

Note: There are cts joint distributions which do not have densities, but we don't have to worry about those in MATH 414 or 416.

As a special case of (3), note that if (X, Y) is a cts joint distribution with joint cdf $F_{X,Y}(x, y)$ and joint density $f_{X,Y}(x, y)$, then

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

PROOF

Joint density functions are used to compute probabilities, applying their definition:

EXAMPLE 1

Suppose X and Y are cts random variables with joint density function

$$f_{X,Y}(x,y) = \begin{cases} 6xy^2 & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{else} \end{cases}.$$

Find $P(X + Y \leq \frac{1}{2})$.

Density functions for marginals (cts case)

Theorem 5.5 (Density functions of marginals, continuous case) *Let $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ be a cts joint distribution with joint density function $f_{\mathbf{X}} : \mathbb{R}^d \rightarrow [0, \infty)$. Then:*

1. *Each marginal X_j is continuous and has a density function;*
2. *For each j ,*

$$f_{X_j}(x_j) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) dx_1 dx_2 \cdots dx_{j-1} dx_{j+1} \cdots dx_d.$$

This theorem tells us that to find the density function of the marginal of a continuous joint distribution, you **integrate the joint density with respect to all the other coordinates**.

As a special case, if X and Y are cts r.v.s with joint density function $f_{X,Y}(x, y)$, then

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx.$$

PROOF

Independence of cts r.v.s

Definition 5.6 Let X_1, \dots, X_d be real-valued r.v.s with joint distribution \mathbf{X} . The r.v.s (just as well, the distribution) are (is) called **(mutually) independent** if

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^d F_{X_j}(x_j)$$

for all $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$, where $F_{\mathbf{X}}$ is the joint cdf and the F_{X_j} are the cdfs of the marginals.

Again, whether r.v.s are independent depends on the joint distribution, and not just on the marginals.

Theorem 5.7 Let X_1, \dots, X_d be continuous real-valued r.v.s with joint density $f_{\mathbf{X}} : \mathbb{R}^d \rightarrow [0, \infty)$. Then the X_j are independent if and only if

$$f_{\mathbf{X}}(x_1, \dots, x_d) = \prod_{j=1}^d f_{X_j}(x_j) \text{ for all } (x_1, \dots, x_d) \in \mathbb{R}^d.$$

As a special case, we see that $X \perp Y$ iff $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all x, y .

PROOF

5.2 Examples

EXAMPLE 2

Pick a point (X, Y) uniformly from the unit square $[0, 1]^2$.

EXAMPLE 3

Pick a point (X, Y) uniformly from the line segment connecting $(0, 0)$ and $(1, 1)$.

EXAMPLE 4

Suppose X and Y are continuous r.v.s whose joint density

$$f_{X,Y}(x,y) = \begin{cases} \frac{C}{(x+y)^4} & \text{if } x \geq 1, y \geq 1 \\ 0 & \text{else} \end{cases}$$

1. Find C .
2. Find $P(Y \leq 2X)$.
3. Find the densities of the marginals.
4. Determine if $X \perp Y$.

3. We compute the density of X first, by integrating with respect to y :

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy = \int_1^{\infty} \frac{24}{(x+y)^4} dy \\ &= -8(x+y)^{-3} \Big|_1^{\infty} \\ &= 0 - (-8(x+1)^{-3}) \\ &= 8(x+1)^{-3}. \end{aligned}$$

This holds when $x \geq 1$; otherwise $f_X(x) = 0$. So formally, the density is

$$f_X(x) = \begin{cases} 8(x+1)^{-3} & \text{if } x \geq 1 \\ 0 & \text{else} \end{cases}$$

Next, we compute the density of Y by integrating with respect to x :

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx = \int_1^{\infty} \frac{24}{(x+y)^4} dx \\ &= -8(x+y)^{-3} \Big|_1^{\infty} \\ &= 0 - (-8(1+y)^{-3}) \\ &= 8(1+y)^{-3}. \end{aligned}$$

Formally, the answer is

$$f_Y(y) = \begin{cases} 8(1+y)^{-3} & \text{if } y \geq 1 \\ 0 & \text{else} \end{cases}$$

4. To determine whether or not $X \perp Y$, we test as follows:

5.3 Conditional densities

Recall: Given probability space (Ω, \mathcal{A}, P) and given an event E with $P(E) > 0$, we define the **conditional probability of F given E** by

$$P(F | E) = \frac{P(E \cap F)}{P(E)}.$$

Let's try to create something that is analogous on the level of random variables:

QUESTION

Let X, Y be real-valued r.v.s. (either cts or discrete). What is the “probability” of X given a particular value of Y ? e.g.

$$“P(X = x | Y = y)” =$$

Definition 5.8 Let X and Y be real-valued r.v.s with joint density function $f_{X,Y}$. The **conditional density of X given Y** is the function $f_{X|Y} : \mathbb{R}^2 \rightarrow [0, \infty)$ defined by

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

where f_Y is the density of the marginal Y (if $f_Y(y) = 0$, we say $f_{X|Y}(x|y) = 0$).

Theorem 5.9 (Properties of conditional densities) Let X and Y be real-valued r.v.s. Then:

1. For every y such that $f_Y(y) > 0$, $f_{X|Y}(x|y)$ is a density function for the random variable $X|Y$ (whose value is x), i.e.

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y) dx = 1.$$

2. (Multiplicative property)

$$f_{X|Y}(x|y) \cdot f_Y(y) = f_{X,Y}(x, y).$$

3. (Probability calculations)

$$P(X \in E | Y = y) = \int_E f_{X|Y}(x|y) dx.$$

EXAMPLE 5

Suppose X and Y have joint density

$$f_{X,Y}(x, y) = \begin{cases} cy(2 - x - y) & \text{if } (x, y) \in [0, 1]^2 \\ 0 & \text{else} \end{cases}.$$

1. Find the conditional density of Y given X .
2. Find the conditional density of Y given $X = \frac{1}{3}$.
3. Find the probability that $Y \in [\frac{1}{4}, \frac{3}{4}]$ given that $X = \frac{1}{3}$.

Solution: 1.

2. Having computed $f_{Y|X}$ in # 1, we compute this simply by plugging in $x = \frac{1}{3}$:

$$f_{Y|X}\left(y \middle| \frac{1}{3}\right) = \frac{6y(2 - \frac{1}{3} - y)}{4 - 3(\frac{1}{3})} = 2y\left(\frac{5}{3} - y\right).$$

3. Integrate the conditional density found in # 2:

$$\begin{aligned} P\left(Y \in \left[\frac{1}{4}, \frac{3}{4}\right] \mid X = \frac{1}{3}\right) &= \int_{1/4}^{3/4} f_{Y|X}\left(y \middle| \frac{1}{3}\right) dy \\ &= \int_{1/4}^{3/4} 2y\left(\frac{5}{3} - y\right) dy \\ &= \int_{1/4}^{3/4} \left(\frac{10}{3}y - 2y^2\right) dy \\ &= \left[\frac{5}{3}y^2 - \frac{2}{3}y^3\right]_{1/4}^{3/4} = \frac{9}{16}. \end{aligned}$$

EXAMPLE 6

Suppose that $X \sim \text{Exp}(\lambda)$, and that $Y|X \sim \text{Exp}(x)$.

1. Find the joint density of X and Y .
2. Find the density of Y .

EXAMPLE 7

Suppose that X is $n(0, 2)$ and that given $X = x$, Y is $n(x, 4)$. Describe Y (i.e. identify it as a common type of r.v.), giving its parameters.

$$\begin{aligned}
 &= \frac{1}{4\pi\sqrt{2}} \int_{-\infty}^{\infty} \exp \left[\frac{-2x^2y^2 + 2xy - x^2}{8} \right] dx \\
 &= \frac{1}{4\pi\sqrt{2}} \int_{-\infty}^{\infty} \exp \left[\frac{-3}{8} \left(x^2 - \frac{2}{3}xy + \frac{1}{3}y^2 \right) \right] dx \\
 &= \frac{1}{4\pi\sqrt{2}} \int_{-\infty}^{\infty} \exp \left[\frac{-3}{8} \left(\left(x - \frac{1}{3}y \right)^2 + \frac{2}{9}y^2 \right) \right] dx \\
 &= \frac{1}{4\pi\sqrt{2}} \int_{-\infty}^{\infty} \exp \left[\frac{-3}{8} \left(\left(x - \frac{1}{3}y \right)^2 + \frac{2}{9}y^2 \right) \right] dx \\
 &= \frac{1}{4\pi\sqrt{2}} \int_{-\infty}^{\infty} \exp \left(\frac{-y^2}{12} \right) \exp \left[\frac{-\left(x - \frac{1}{3}y \right)^2}{\frac{8}{3}} \right] dx
 \end{aligned}$$

Note: There will be an easier method of solving this problem given later in the course.

5.4 Transformations of continuous joint distributions

We want to consider two types of problems dealing with transformations of a joint distribution:

Class 1: Compute the density of a real-valued r.v. U obtained as a function of several r.v.s X_1, \dots, X_d which have some given joint distribution.

Ex: Given joint density of X and Y , find the density of $Z = X + 2Y$.

Class 2: Compute the joint density of some r.v.s U_1, \dots, U_d obtained as functions of several r.v.s X_1, \dots, X_d which have some given joint distribution.

Ex: Given joint density of X and Y , find the joint density of U and V if $U = X + Y$ and $V = \frac{X}{X+Y}$.

We handle problems in each of these classes separately.

Class 1 Examples

Setup: $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ is some function; $U = \varphi(X_1, \dots, X_d) = \varphi(\vec{X})$ is real-valued.

Method of solution:

1. Classify U as discrete or continuous.
2. Find the range of U .
3. If U is discrete, compute the density by back-substitution:

$$\begin{aligned} f_U(u) &= P(U = u) = P(\varphi(\mathbf{X}) = u) = P(\mathbf{X} \in \varphi^{-1}(u)) \\ &= \begin{cases} \int_{\varphi^{-1}(u)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} & \text{if } \mathbf{X} \text{ cts} \\ \sum_{\mathbf{x} \in \varphi^{-1}(u)} f_{\mathbf{X}}(\mathbf{x}) & \text{if } \mathbf{X} \text{ discrete} \end{cases} . \end{aligned}$$

4. If U is continuous, first compute the cdf of Y by back-substitution:

$$\begin{aligned} F_U(u) &= P(U \leq u) = P(\varphi(\mathbf{X}) \leq u) = P(\mathbf{X} \in \varphi^{-1}(-\infty, u]) \\ &= \begin{cases} \int_{\varphi^{-1}(-\infty, u]} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} & \text{if } \mathbf{X} \text{ cts} \\ \sum_{\mathbf{x} \in \varphi^{-1}(-\infty, u]} f_{\mathbf{X}}(\mathbf{x}) & \text{if } \mathbf{X} \text{ discrete} \end{cases} . \end{aligned}$$

Then differentiate F_U with respect to u to obtain f_U .

EXAMPLE 8

Let (X, Y) be independent, exponential r.v.s, both with parameter λ . Find the density of $X + Y$.

EXAMPLE 9

Suppose that the amount X an insurance company pays in claims and the amount Y it collects in premiums are modeled by a joint density

$$f_{X,Y}(x, y) = \begin{cases} \frac{3}{500}x & \text{if } 0 \leq x \leq y \leq 10 \\ 0 & \text{else} \end{cases}$$

Let R be the ratio of premiums to claims; find the distribution function of R .

Class 2 Examples

Setup: $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is some function (we will assume that φ is invertible, otherwise the problem is much harder); $\mathbf{U} = (U_1, \dots, U_d) = \varphi(X_1, \dots, X_d) = \varphi(\mathbf{X})$ is a joint distribution. $f_{\mathbf{U}}(\mathbf{u}) = ?$

Let's write $\varphi(x_1, \dots, x_d) = (u_1, \dots, u_d)$ for convenience.

This problem has a theoretical solution: suppose for now that $d = 2$. Then, the joint density of \mathbf{U} should satisfy, for every (measurable) set $E \subseteq \mathbb{R}^2$,

Since this holds for every $E \subseteq \mathbb{R}^2$, we have

$$\begin{aligned} f_{\mathbf{U}}(u_1, u_2) \cdot |J(\varphi)| &= f_{\mathbf{X}}(x_1, x_2) \\ \Rightarrow f_{\mathbf{U}}(u_1, u_2) &= \frac{1}{|J(\varphi)|} f_{\mathbf{X}}(x_1, x_2) \end{aligned}$$

where $J(\varphi)$ is the **Jacobian** of φ :

$$J(\varphi) = \det \begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} \end{pmatrix}$$

This generalizes:

Theorem 5.10 (Transformation theorem, higher-dimensions) Suppose $\mathbf{X} = (X_1, \dots, X_d)$ has joint density $f_{\mathbf{X}} : \mathbb{R}^d \rightarrow [0, \infty)$. Suppose that $\mathbf{U} = (U_1, \dots, U_d) = \varphi(X_1, \dots, X_d)$ where $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is some function whose partial derivatives $\frac{\partial u_i}{\partial x_j}$ exist everywhere and are continuous (such a φ is called a “ C^1 ” function). If the Jacobian determinant

$$J(\varphi) = \det \begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \dots & \frac{\partial u_1}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial u_d}{\partial x_1} & \frac{\partial u_d}{\partial x_2} & \dots & \frac{\partial u_d}{\partial x_d} \end{pmatrix}$$

is everywhere nonzero, then the U_j are all continuous and have joint density given by

$$f_{\mathbf{U}}(u_1, \dots, u_d) = \frac{1}{|J(\varphi)|} f_{\mathbf{X}}(x_1, \dots, x_d),$$

i.e.

$$f_{\mathbf{U}}(\mathbf{u}) = \frac{1}{|J(\varphi)|} f_{\mathbf{X}}(\varphi^{-1}(\mathbf{u})).$$

As a special case, let’s see what happens if $d = 1$ (so that X and U are real-valued). Then $\varphi : \mathbb{R} \rightarrow \mathbb{R}$;

$$J(\varphi) =$$

so from this theorem we can conclude that

$$f_U(u) =$$

(Compare this result to the two theorems on page 81.)

EXAMPLE 10

Let (X_1, X_2) be uniform on $[0, 1]^2$. Find the joint density of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$.

EXAMPLE 11

Suppose $X_1 \sim \Gamma(\alpha, \lambda)$, $X_2 \sim \Gamma(\beta, \lambda)$ and $X_1 \perp X_2$. Find the joint density of $Y_1 = X_1 + X_2$ and $Y_2 = \frac{X_1}{X_1 + X_2}$.

Chapter 6

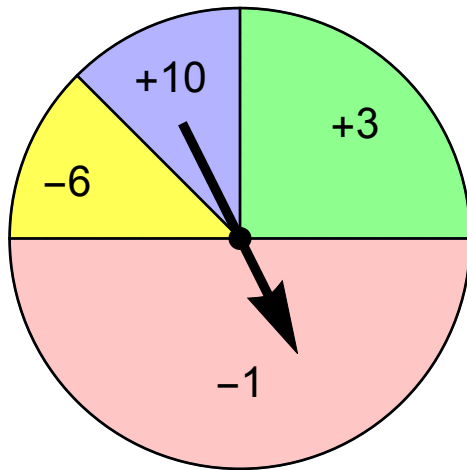
Expected value

6.1 Definition and properties

Motivating question: What is the “average” value of a random variable?

EXAMPLE 1

You and your friend play a game with a spinner. You spin the spinner and then exchange money depending on where the spinner lands:



Flawed definition: The **expected value** of a discrete, real-valued r.v. X , denoted EX , is

$$EX = \sum_{x \in \text{Range}(X)} x f_X(x).$$

Technical point: The range of X might be an infinite set (i.e. it might be \mathbb{Z}). Then there are potential issues with the convergence of the infinite series

$$\sum_{x \in \text{Range}(X)} x f_X(x)$$

if we try to rearrange terms. To get around any problems, we require that this series *converge absolutely*.

Definition 6.1 Let $X : \Omega \rightarrow \mathbb{R}$ be a discrete r.v., with density f_X . We say X has **finite expectation** (and write $EX < \infty$) if

$$\sum_{x \in \text{Range}(X)} |x| f_X(x) < \infty;$$

in which case we say the **expected value** (a.k.a. **mean** a.k.a. **expectation**) of X is the real number

$$EX = \sum_{x \in \text{Range}(X)} x f_X(x).$$

If $\sum_{x \in \text{Range}(X)} |x| f_X(x) = \infty$, we say X **does not have finite expectation** and we write $EX = \infty$.

A similar definition works for continuous, real-valued r.v.s:

Definition 6.2 Let $X : \Omega \rightarrow \mathbb{R}$ be a continuous r.v., with density f_X . We say X has **finite expectation** (and write $EX < \infty$) if

$$\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$$

in which case we say the **expected value** (a.k.a. **mean** a.k.a. **expectation**) of X is the real number

$$EX = \int_{-\infty}^{\infty} x f_X(x) dx.$$

If $\int_{-\infty}^{\infty} |x| f_X(x) dx$ diverges, we say X **does not have finite expectation** and we write $EX = \infty$.

The expected value of a r.v. X is also denoted μ , μ_X , $E[X]$, $E(X)$ and $\mathbb{E}(X)$.

Note If $X : \Omega \rightarrow \mathbb{R}$ is neither discrete nor cts, then it makes no sense to talk about EX . Also, if X is not real valued (including when $X : \Omega \rightarrow \mathbb{R}^d$ is a joint distribution), it makes no sense to talk about EX .

EXAMPLE 2

Suppose X has density function $f_X(x) = \frac{3}{28}x^2$ for $-1 \leq x \leq 3$ (and $f_X(x) = 0$ otherwise). Find the expected value of X .

Note If the range of X is either bounded above or bounded below, then one can simultaneously check that X has finite expectation and compute EX by computing $\sum_x xf_X(x)$ (if X is discrete) or $\int_{-\infty}^{\infty} xf_X(x) dx$ (if X is continuous).

LOTUS (Expected values of transformations)

Suppose you know the density of r.v. X . To get the expected value of X , you compute

$$EX = \sum_x xf_X(x) \quad \text{or} \quad EX = \int xf_X(x) dx.$$

How would you compute the expected value of a transformation of X , i.e. what is EY if $Y = \varphi(X)$?

Long way:

Seemingly dumb way:

Actually, this seemingly dumb way works! It's called "LOTUS", which is an acronym for the Law of the Unconscious Statistician:

Theorem 6.3 (LOTUS) Suppose $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ is a r.v. (either discrete or cts) with density $f_{\mathbf{X}}$. Let $U = \varphi(\mathbf{X})$ where $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a real-valued function of d variables. (This makes U a real-valued r.v.) Then:

(a) U has finite expectation if and only if

$$\begin{cases} \sum_{\mathbf{x}} |\varphi(\mathbf{x})| f_{\mathbf{X}}(\mathbf{x}) < \infty & \text{if } \mathbf{X} \text{ discrete} \\ \int_{-\infty}^{\infty} |\varphi(x)| f_X(x) dx < \infty & \text{if } \mathbf{X} = X \text{ cts, real-valued} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\varphi(\mathbf{x})| f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} < \infty & \text{if } \mathbf{X} \text{ cts, vector-valued} \end{cases}$$

(b) if $EU < \infty$, then

$$EU = \begin{cases} \sum_{\mathbf{x}} \varphi(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) & \text{if } \mathbf{X} \text{ discrete} \\ \int_{-\infty}^{\infty} \varphi(x) f_X(x) dx & \text{if } \mathbf{X} = X \text{ cts, real-valued} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \varphi(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} & \text{if } \mathbf{X} \text{ cts, vector-valued} \end{cases}$$

PROOF (when \mathbf{X} is discrete): In this case, U is also discrete; let u_1, u_2, \dots be the values in the range of U .

For each j , let $A_j = \varphi^{-1}(u_j) = \{\mathbf{x} \in \text{Range}(\mathbf{X}) : \varphi(\mathbf{x}) = u_j\}$. The A_j form a partition of the range of X . Now

$$(\mathbf{X} \in A_j \iff U = u_j) \Rightarrow f_U(u_j) = P(U = u_j) = P(\mathbf{X} \in A_j) = \sum_{\mathbf{x} \in A_j} f_{\mathbf{X}}(\mathbf{x}).$$

So

$$\begin{aligned} E|U| &= \sum_j |u_j| f_U(u_j) = \sum_j |u_j| \sum_{\mathbf{x} \in A_j} f_{\mathbf{X}}(\mathbf{x}) \\ &= \sum_j \sum_{\mathbf{x} \in A_j} |u_j| f_{\mathbf{X}}(\mathbf{x}) \\ &= \sum_{\mathbf{x} \in \text{Range}(\mathbf{X})} |\varphi(\mathbf{x})| f_{\mathbf{X}}(\mathbf{x}). \end{aligned}$$

Thus $EU < \infty$ if and only if $\sum_{\mathbf{x}} |\varphi(\mathbf{x})| f_{\mathbf{X}}(\mathbf{x}) < \infty$, proving statement (a).

For the second statement, repeat the argument of the previous paragraph with no absolute values around the u_j .

The proof when X is continuous is beyond the scope of this course (uses measure theory). \square

EXAMPLE 3

Suppose X has density function $f_X(x) = x + \frac{1}{2}$ for $0 < x < 1$ (and $f_X(x) = 0$ otherwise). Let $Y = 3X^2 + 6X + 7$. Find EY .

Note: In this solution, by using LOTUS we don't have to bother with finding f_Y explicitly.

EXAMPLE 4

Let $X \sim \text{Pois}(\lambda)$. Find $E[e^X]$.

Expected values and survival functions

A useful, alternate method to compute expected values is by means of the survival function. Recall that for a real-valued r.v. X , $H_X(x) = P(X > x) = 1 - F_X(x)$.

Theorem 6.4 (Expected value from survival function) Suppose X is a random variable taking values in $[0, \infty)$. Then:

1. if X is discrete, then $EX = \sum_{x=0}^{\infty} H_X(x)$.
2. if X is continuous, then $EX = \int_0^{\infty} H_X(x) dx$.

PROOF If X is discrete, then

$$\begin{aligned}
 EX &= \sum_{x=0}^{\infty} x f_X(x) = 1f_X(1) + 2f_X(2) + 3f_X(3) + 4f_X(4) + \dots \\
 &= [f_X(1) + f_X(2) + f_X(3) + \dots] + [f_X(2) + f_X(3) + \dots] + [f_X(3) + \dots] \\
 &= P(X > 0) + P(X > 1) + P(X > 2) + \dots \\
 &= \sum_{x=0}^{\infty} P(X > x) = \sum_{x=0}^{\infty} H_X(x).
 \end{aligned}$$

If X is continuous, then

$$\begin{aligned}
 EX &= \lim_{b \rightarrow \infty} \int_0^b x f_X(x) dx = \lim_{b \rightarrow \infty} \left[x F_X(x) \Big|_0^b - \int_0^b F_X(x) dx \right] \\
 &= \lim_{b \rightarrow \infty} \left[b F_X(b) - \int_0^b F_X(x) dx \right] \\
 &= \lim_{b \rightarrow \infty} \left[F_X(b) \int_0^b 1 dx \right] - \lim_{b \rightarrow \infty} F_X(b) \int_0^b F_X(x) dx \\
 &\quad \text{(since } \lim_{b \rightarrow \infty} F_X(b) = 1) \\
 &= \lim_{b \rightarrow \infty} \left[F_X(b) \int_0^b 1 dx \right] - \lim_{b \rightarrow \infty} \left[F_X(b) \int_0^b F_X(x) dx \right] \\
 &= \lim_{b \rightarrow \infty} \left[F_X(b) \int_0^b [1 - F_X(x)] dx \right] \\
 &= 1 \cdot \lim_{b \rightarrow \infty} \int_0^b [1 - F_X(x)] dx \\
 &= \int_0^\infty H_X(x) dx. \quad \square
 \end{aligned}$$

Theorem 6.5 (Properties of Expected Value) Suppose $EX < \infty$ and $EY < \infty$. Then:

1. (Preservation of constants) If $P(X = c) = 1$, then $EX = c$.
2. (Preservation of bounds) If $P(|X| \leq M) = 1$, then $|EX| \leq M$.
3. (Linearity I) cX has finite expectation and $E[cX] = cEX$.
4. (Linearity II) $X + Y$ has finite expectation and $E[X + Y] = EX + EY$.
5. (Monotonicity) If $P(X \geq Y) = 1$ then $EX \geq EY$.
6. (Definiteness) If $P(X \geq Y) = 1$ and $EX = EY$ then $P(X = Y) = 1$.
7. (Triangle inequality) $|EX| \leq E|X|$.

PROOF We start with property (1): if $P(X = c) = 1$, then X is discrete and $f_X(c) = 1$. Then

$$EX = \sum_x x f_X(x) = c \cdot 1 = c.$$

Now for property (3): suppose X has finite expectation. If X is discrete, then so is cX and

$$\sum_x |cx| f_X(x) = |c| \sum_x |x| f_X(x) < \infty$$

so cX has finite expectation by Theorem 6.3. Then, also by Theorem 6.3,

$$E[cX] = \sum_x cx f_X(x) = c \sum_x x f_X(x) = cEX.$$

If X is continuous, the same proof works using integrals instead of sums.

Next, property (4): suppose X and Y have finite expectation and let $Z = X + Y = \varphi(X, Y)$. Then if X and Y are discrete, then

$$\begin{aligned} \sum_{x,y} |x+y| f_{X,Y}(x,y) &\leq \sum_{x,y} (|x| + |y|) f_{X,Y}(x,y) \\ &= \sum_{x,y} |x| f_{X,Y}(x,y) + \sum_{x,y} |y| f_{X,Y}(x,y). \end{aligned}$$

Since $EX < \infty$, the first sum is finite, and since $EY < \infty$, the second sum is finite. Thus the sum is finite so by Theorem 6.3, $E[X + Y] < \infty$. Now, again using Theorem 6.3,

$$E[X + Y] = \sum_{x,y} (x + y) f_{X,Y}(x,y) = \sum_{x,y} x f_{X,Y}(x,y) + \sum_{x,y} y f_{X,Y}(x,y) = EX + EY.$$

The proof if X and/or Y is continuous is similar and uses integrals where necessary instead of sums.

Now for property (5): Let $Z = X - Y$; by (3) and (4) $EZ = EX - EY$. If $P(X \geq Y) = 1$, then $P(Z \geq 0) = 1$. So if Z is discrete, then

$$EZ = \sum_z z f_Z(z) \geq 0$$

since all the numbers in the sum are nonnegative. Thus $EX - EY \geq 0$ so $EX \geq EY$. (The proof if Z is continuous is similar.)

To prove (7): $-|X| \leq X \leq |X|$ implies $-E|X| \leq EX \leq E|X|$ by (5). Thus $|EX| \leq E|X|$.

(2) follows from (5) and (7).

For (6), again let $Z = X - Y$, since $EX = EY$ we have $EZ = 0$. At the same time, repeating the argument in (5) we have

$$EZ = \sum_z z f_Z(z) = 0$$

and since all the z s in the sum are ≥ 0 and all the $f_Z(z)$ s are ≥ 0 , the only way this can be consistent with $\sum_z f_Z(z) = 1$ is if $f_Z(0) = 1$ (otherwise there would be a positive term that could not have any negative term that cancels it). Thus $P(Z = 0) = 1$ so $P(X - Y = 0) = 1$ so $P(X = Y) = 1$. Again, the proof is similar if X and/or Y are continuous, replacing the sums with integrals as necessary. \square

Theorem 6.6 (Independence Properties of Expected Value) Suppose $EX < \infty$ and $EY < \infty$. If $X \perp Y$, then:

1. if $\varphi(X)$ and $\psi(Y)$ both have finite expectation, then so does $\varphi(X)\psi(Y)$, and

$$E[\varphi(X)\psi(Y)] = E[\varphi(X)] \cdot E[\psi(Y)].$$

2. $E[XY] = EX \cdot EY$.

WARNING: The converse of this is false, i.e. $E[XY] = EX \cdot EY$ does not imply $X \perp Y$.

PROOF To prove (1), note that $f_{X,Y}(x, y) = f_X(x)f_Y(y)$. Then

$$\begin{aligned} \sum_{x,y} |\varphi(x)\psi(y)| f_{X,Y}(x, y) &= \sum_x \sum_y |\varphi(x)| |\psi(y)| f_X(x) f_Y(y) \\ &= \left(\sum_x |\varphi(x)| f_X(x) \right) \left(\sum_y |\psi(y)| f_Y(y) \right). \end{aligned}$$

Since $E[\varphi(X)] < \infty$, the first sum is finite, and since $E[\psi(Y)] < \infty$, the second sum is finite. Thus the entire expression is finite so by Theorem 6.3, we have $E[\varphi(X)\psi(Y)] < \infty$. Then

$$\begin{aligned} E[\varphi(X)\psi(Y)] &= \sum_{x,y} \varphi(x)\psi(y) f_{X,Y}(x, y) = \sum_x \sum_y \varphi(x)\psi(y) f_X(x) f_Y(y) \\ &= \left(\sum_x \varphi(x) f_X(x) \right) \left(\sum_y \psi(y) f_Y(y) \right) \\ &= E[\varphi(X)] \cdot E[\psi(Y)]. \end{aligned}$$

Again, the proof is similar if X and/or Y are continuous, replacing the sums with integrals as necessary.

(2) follows from (1) by setting $\varphi(x) = x$ and $\psi(y) = y$. \square

Expected values of common random variables

Theorem 6.7 *The expected values of the common classes of r.v.s encountered in Chapters 2 and 3 are as follows:*

X	EX
$Unif(\{1, 2, \dots, n\})$	$\frac{n+1}{2}$
$Geom(p)$	$\frac{1-p}{p}$
$NB(r, p)$	$r \left(\frac{1-p}{p} \right)$
$\text{binomial}(n, p)$	np
$Pois(\lambda)$	λ
$Hyp(n, r, k)$	$\frac{kr}{n}$
$Unif([a, b])$	$\frac{a+b}{2}$
$Exp(\lambda)$	$\frac{1}{\lambda}$
$\Gamma(r, \lambda)$	$\frac{r}{\lambda}$
std. normal $n(0, 1)$	0
normal $n(\mu, \sigma^2)$	μ
Cauchy	∞

PROOF (OF SOME OF THESE) In the homework, you will prove the results when X is hypergeometric, exponential, gamma, and normal (not standard normal).

6.2 Variance and covariance

Definition 6.8 Let $X : \Omega \rightarrow \mathbb{R}$ be a r.v. such that $EX < \infty$ and $E[(X - EX)^2] < \infty$. The **variance** of X , denoted $Var(X)$ (or $V(X)$ or σ^2 or σ_X^2), is

$$Var(X) = E[(X - EX)^2].$$

The **standard deviation** of X , denoted σ or σ_X , is $\sigma = \sqrt{Var(X)}$.

Observations:

1. $Var(X) \geq 0$.
2. The more spread out X is, the further from zero $X - EX$ is, so the greater $Var(X)$ is. Thus variance is a measure of spread of a random variable.

Theorem 6.9 (Variance Formula) If $Var(X)$ exists, then $Var(X) = EX^2 - (EX)^2$.

PROOF

Theorem 6.10 $Var(X) = 0$ if and only if X is constant (i.e. $\exists c$ s.t. $P(X = c) = 1$).

PROOF

Variances of common random variables

Theorem 6.11 *The variances of the common classes of r.v.s encountered in Chapters 2 and 3 are as follows:*

X	$Var(X)$
$Unif(\{1, 2, \dots, n\})$	$\frac{n^2-1}{12}$
$Geom(p)$	$\frac{1-p}{p^2}$
$NB(r, p)$	$r \left(\frac{1-p}{p^2} \right)$
$\text{binomial}(n, p)$	$np(1-p)$
$Pois(\lambda)$	λ
$Hyp(n, r, k)$	$\frac{kr}{n} \left(\frac{n-r}{n} \right) \frac{n-k}{n-1}$
$Unif([a, b])$	$\frac{(b-a)^2}{12}$
$Exp(\lambda)$	$\frac{1}{\lambda^2}$
$\Gamma(r, \lambda)$	$\frac{r}{\lambda^2}$
std. normal $n(0, 1)$	1
normal $n(\mu, \sigma^2)$	σ^2
Cauchy	DNE

PROOF (OF SOME OF THESE) The proofs when X is continuous uniform, exponential and gamma are left as HW exercises.

Theorem 6.12 (Properties of Variance) *Let X be a r.v. having finite variance. Then:*

1. *For any constant b , $\text{Var}(X + b) = \text{Var}(X)$;*
2. *For any constant a , $\text{Var}(aX) = a^2\text{Var}(X)$.*

PROOF HW (as a hint, these follow from the variance formula)

New Question: What is a formula for $\text{Var}(X + Y)$ in terms of $\text{Var}(X)$ and $\text{Var}(Y)$?

Answer:

Definition 6.13 *Given two r.v.s X and Y , each having finite variance, the **covariance** between X and Y , denoted $\text{Cov}(X, Y)$ (or $C(X, Y)$ or $\sigma_{X,Y}$ or σ_{XY}) is*

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)].$$

The covariance between two random variables measures the “tendency of the r.v.s to change together”. In other words:

- If $\text{Cov}(X, Y) > 0$, then as X increases, we expect Y to increase and as X decreases, we expect Y to decrease.
- If $\text{Cov}(X, Y) < 0$, then as X increases, we expect X to decrease and as Y decreases, we expect Y to increase.
- If $\text{Cov}(X, Y) = 0$, then changes in X should not lead to any expected change in Y .

Theorem 6.14 (Properties of Covariance) *Let X and Y be r.v.s having finite variance. Then:*

1. $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y);$
2. $Cov(X, Y) = Cov(Y, X);$
3. (Covariance Formula) $Cov(X, Y) = E[XY] - EX \cdot EY;$
4. $Cov(X, X) = Var(X);$
5. *If $X \perp Y$, then $Cov(X, Y) = 0$ and $Var(X + Y) = Var(X) + Var(Y);$
(Note: the converse of (5) is false.)*
6. $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y);$
7. $Cov(X, Y_1 + Y_2) = Cov(X, Y_1) + Cov(X, Y_2);$
8. *For any constant a , $Cov(aX, Y) = aCov(X, Y) = Cov(X, aY).$*

PROOF Property (1) was established before the definition of covariance.

Property (2) is obvious.

For Property (3), notice

$$\begin{aligned}
 Cov(X, Y) &= E[(X - EX)(Y - EY)] \\
 &= E[XY - EX \cdot Y - EY \cdot X + EX \cdot EY] \\
 &= E[XY] - E[EX \cdot Y] - E[EY \cdot X] + E[EX \cdot EY] \\
 &= E[XY] - EX \cdot EY - EY \cdot EX + EX \cdot EY \\
 &= E[XY] - EX \cdot EY.
 \end{aligned}$$

Property (4) is immediate from the definitions of covariance and variance.

Property (5) follows from (3) together with the independence property of expected value.

Properties (6), (7) and (8) are homework exercises. \square

A problem with covariance: Suppose X and Y , both measured in hours, have covariance 2. Then if we let X_M and Y_M be the same quantities as X and Y , but measured in minutes rather than hours, we have

$$\text{Cov}(X_M, Y_M) =$$

Thus the covariance between two quantities *depends greatly on the units the quantities are measured in*. We don't really want this, because the covariance is "supposed" to measure how correlated the random variables are. To fix this, we invent a new quantity called "correlation":

Definition 6.15 Given two r.v.s X and Y , each having finite variance, the **correlation** between X and Y , denoted $\rho(X, Y)$ (or ρ_{XY} or $\rho_{X,Y}$) is

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}.$$

X and Y are **uncorrelated** if $\rho(X, Y) = 0$ (equivalently, if $\text{Cov}(X, Y) = 0$).

Note that independent r.v.s are uncorrelated, but uncorrelated r.v.s may not be independent (HW).

Theorem 6.16 (Schwarz Inequality) Let X and Y be real-valued r.v.s with finite variances. Then

$$(E[XY])^2 \leq EX^2 \cdot EY^2.$$

PROOF The proof of the Schwarz inequality has two cases:

Case 1: If $P(Y = 0) = 1$, then $E([XY])^2 = 0 \leq 0 = EX^2 \cdot 0 = EX^2 \cdot EY^2$ as desired.

Case 2: $P(Y = 0) < 1$. This implies $P(Y^2 = 0) < 1$ so $E[Y^2] > 0$; this will allow us to divide through by EY^2 later on. Now, define a function $f : \mathbb{R} \rightarrow \mathbb{R}$ by setting

$$f(t) = E[(X - tY)^2].$$

Notice that $f(t) \geq 0$ for all t since f is the expected value of a nonnegative r.v. Expanding f , we get

$$f(t) = E[(X - tY)(X - tY)] = E[X^2 - 2tXY + t^2Y^2] = EX^2 - 2tE[XY] + t^2EY^2.$$

Thus f is a quadratic formula of t ; from the above, the vertex of this parabola must lie above the t -axis since $f(t) \geq 0$ for all t . Let's find the y -coordinate of this vertex:

$$f'(t) = 2tEY^2 - 2E[XY]$$

Therefore $f'(t) = 0$ when $t = \frac{E[XY]}{EY^2}$, and the corresponding y -coordinate is

$$\begin{aligned} f\left(\frac{E[XY]}{EY^2}\right) &= EX^2 - 2\left(\frac{E[XY]}{EY^2}\right)E[XY] + \left(\frac{E[XY]}{EY^2}\right)^2 EY^2 \\ &= EX^2 - 2\frac{(E[XY])^2}{EY^2} + \frac{(E[XY])^2}{EY^2} \\ &= EX^2 - \frac{(E[XY])^2}{EY^2}. \end{aligned}$$

Therefore

$$\begin{aligned} 0 &\leq EX^2 - \frac{(E[XY])^2}{EY^2} \\ \Rightarrow \frac{(E[XY])^2}{EY^2} &\leq EX^2 \\ \Rightarrow (E[XY])^2 &\leq EX^2 \cdot EY^2 \end{aligned}$$

as desired. \square

You may recall from linear algebra another inequality called the Cauchy-Schwarz Inequality (important in the context of computing projections of one vector onto another, angles between vectors, etc.). That inequality is basically the same as this one; it says that for two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\| \|\mathbf{y}\|$$

where $\|\cdot\|$ denotes the norm or length of a vector (recall that $\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$). Denoting the “dot product” of two random variables as “ $X \cdot Y$ ” = $E[XY]$, the Schwarz inequality here is exactly the same thing as the C-S inequality from linear algebra.

Theorem 6.17 (Properties of Correlation) *Let X and Y be r.v.s having finite variance. Then:*

1. $\rho(X, Y) = \rho(Y, X)$;
2. $\rho(X, X) = 1$;
3. $|\rho(X, Y)| \leq 1$;
4. If $X \perp Y$, then $\rho(X, Y) = 0$ (the converse of this is false);
5. For any positive constants a and b , and any constants c and d ,

$$\rho(aX + c, bY + d) = \rho(X, Y);$$

6. $\rho(X, Y) = \pm 1$ if and only if $Y = aX + b$ for some constants a and b with $a \neq 0$.

PROOF (1) is obvious from the definition, since $Cov(X, Y) = Cov(Y, X)$.

(2) is a direct calculation:

$$\rho(X, X) = \frac{Cov(X, X)}{\sqrt{Var(X) \cdot Var(X)}} = \frac{Var(X)}{\sqrt{(Var(X))^2}} = \frac{Var(X)}{Var(X)} = 1.$$

To prove (3), apply the Schwarz Inequality to $X - EX$ and $Y - EY$ to get

$$E[(X - EX)(Y - EY)]^2 \leq E[(X - EX)^2] \cdot E[(Y - EY)^2]$$

i.e.

$$Cov(X, Y)^2 \leq Var(X) \cdot Var(Y).$$

Take the square root of both sides to get

$$|Cov(X, Y)| \leq \sqrt{Var(X) \cdot Var(Y)}$$

i.e.

$$|\rho(X, Y)| = \frac{|Cov(X, Y)|}{\sqrt{Var(X) \cdot Var(Y)}} \leq 1.$$

(4) follows from the fact that $X \perp Y$ implies $Cov(X, Y) = 0$.

(5) and (6) are HW problems. \square

6.3 Conditional expectation and conditional variance

Definition 6.18 Let X and Y be real-valued r.v.s. The **conditional expectation of Y given X** , also called the **regression of Y on X** , is the function

$$E(Y|X) = \begin{cases} \sum_y y f_{Y|X}(y|x) & \text{if } Y|X \text{ is discrete} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy & \text{if } Y|X \text{ is cts} \end{cases}$$

In this setting, there is also a conditional expectation of X given Y , defined by

$$E(X|Y) = \begin{cases} \sum_x x f_{X|Y}(x|y) & \text{if } X|Y \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx & \text{if } X|Y \text{ is cts} \end{cases}$$

Theorem 6.19 (Properties of conditional expectation) Let X and Y be real-valued r.v.s. Then:

1. $E(Y|X)$ is a function of x , not a number.
(We can think of $E(Y|X)$ as a r.v. by thinking of it as a function of X .)
2. (Law of Total Expectation) $E[E(Y|X)] = EY$.
(In other words, $\int_{-\infty}^{\infty} E(Y|X)(x) f_X(x) dx = EY$ or $\sum_x E(Y|X)(x) f_X(x) dx$.)
3. (Linearity)
 - (i) $E(Y_1 + Y_2 | X) = E(Y_1|X) + E(Y_2|X)$; and
 - (ii) $E(cY|X) = c E(Y|X)$ for any constant c .
4. (Independence) The following are equivalent:
 - (i) $X \perp Y$
 - (ii) $E(Y|X)$ is a constant function.
 - (iii) $E(Y|X) = EY$ for all x .

What does this all mean? As an example, suppose X and Y are chosen from this set E with some joint density function:

Useful integral formulas when computing conditional expectations

Gamma integral formula: $\int_0^\infty x^{r-1} e^{-\lambda x} dx = \frac{\Gamma(r)}{\lambda^r}$

Beta integral formula: $\int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

Normal integral formula: $\int_{-\infty}^\infty e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \sigma\sqrt{2\pi}$

Note If you recognize the conditional density $f_{Y|X}$ as a common density, then you can immediately conclude the value of $E(Y|X)$ from the facts known about expected values of common r.v.s.

EXAMPLE 5

Suppose the conditional density is (for $x, y > 0$)

$$f_{Y|X}(y|x) = xe^{-xy}.$$

Then we know

EXAMPLE 6

Let X and Y have joint density

$$f_{X,Y}(x,y) = \begin{cases} \frac{12}{5}y(2-x-y) & \text{if } (x,y) \in [0,1]^2 \\ 0 & \text{else} \end{cases}.$$

Find the conditional expectation of Y given X and the conditional expectation of Y given $X = \frac{1}{3}$.

EXAMPLE 7

Three contestants on a game show are given the same question, and each person answers the question correctly with probability $1 - x$ (their answers are independent). The difficulty x of the question is itself a r.v. chosen from $(0, 1)$ with density function $6x(1 - x)$. Find the expected difficulty level of the question, given that all three contestants answer incorrectly.

Conditional variance

Definition 6.20 Let X and Y be real-valued r.v.s. The **conditional variance of Y given X** , is the function

$$\begin{aligned} \text{Var}(Y|X) &= E[(Y - E[Y|X])^2 | X] \\ &= E(Y^2 | X) - E(Y|X)^2. \end{aligned}$$

That the two formulas given in the box above are the same is a HW problem.

As with conditional expectation, the conditional variance is a function of x (and can be thought of as a random variable).

Theorem 6.21 (Law of Total Variance) Let X and Y be real-valued r.v.s. Then

$$\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}[E(Y|X)].$$

PROOF HW (use the definitions and crunch the symbols appropriately)

This theorem is extremely useful for computing the variance of Y , when X and $Y|X$ are given as common random variables:

EXAMPLE 8

The number of accidents on a stretch of highway is uniform on $\{1, 2, 3, \dots, 9\}$. Given N accidents on the stretch of highway, the total amount of damage caused by the accidents is exponential with mean $2N$. Find the variance of the total amount of damage caused by accidents on this stretch of highway.

Chapter 7

Moment generating functions

Recall from Calculus 2: You can think of the exponential function as e^x , or as

$$\sum_{j=0}^{\infty} \frac{x^j}{j!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^j}{j!} + \dots$$

Normally, we think of r.v.s as being described by density functions.

Question: Is there an alternate (“power-series” like) representation of a r.v.?

Answer:

7.1 Probability generating functions

Definition 7.1 Let $X : \Omega \rightarrow \mathbb{N}$ be a discrete r.v., taking values only in $\{0, 1, 2, 3, \dots\}$. The **probability generating function** of X (a.k.a. **pgf** or **generating function**), denoted G_X or Φ_X , is the function $G_X : [-1, 1] \rightarrow \mathbb{R}$ defined by

$$G_X(t) = E[t^X] = \sum_{x=0}^{\infty} f_X(x)t^x.$$

Note: The t in this definition is just a dummy variable.

Theorem 7.2 (Properties of probability generating functions) Let $X : \Omega \rightarrow \mathbb{N}$ be a r.v. with probability generating function G_X . Then:

1. G_X is a continuous and differentiable function of t on $[-1, 1]$.
2. $|G_X(t)| \leq 1$ for all t .
3. $G_X(1) = 1$.
4. $G'_X(1) = EX$.
5. $G_X(0) = f_X(0) = P(X = 0)$ (the constant term on G_X).
6. (Inversion formula for PGFs) $f_X(n) = \frac{G_X^{(n)}(0)}{n!}$ for all $n \in \{0, 1, 2, 3, \dots\}$
7. The equation $G_X(t) = t$ has a solution in $(0, 1)$ if and only if $EX > 1$.
8. $G''_X(1) = E[X(X - 1)] = EX^2 - EX$.
9. $Var(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$.

PROOF (1) follows from the fact that G_X is a power series (from Calculus 2, all power series are cts and diffble).

(2): $|G_X(t)| = |E[t^X]| \leq E|t^X| \leq E[1] = 1$.

(3): $G_X(1) = E[1^X] = E[1] = 1$.

(4): $G_X(t) = \sum_{x=0}^{\infty} t^x f_X(x)$ so $G'_X(t) = \sum_{x=1}^{\infty} x t^{x-1} f_X(x)$. Therefore

$$G'_X(1) = \sum_{x=1}^{\infty} x 1^{x-1} f_X(x) = \sum_{x=1}^{\infty} x f_X(x) = \sum_{x=0}^{\infty} x f_X(x) = EX.$$

(5): $G_X(0) = f_X(0) = P(X = 0)$, the constant term on G_X .

(6) follows from uniqueness of power series (Calculus 2).

(7): From (3) and (4), $G_X(t)$ passes through $(1, 1)$ with slope EX . From (5), $G_X(t)$ passes through $(0, f_X(0))$. From (1), G_X is cts. So the graph of G_X looks like

(8): From above, $G'_X(t) = \sum_{x=1}^{\infty} x t^{x-1} f_X(x)$. Therefore $G''_X(t) = \sum_{x=2}^{\infty} x(x-1) t^{x-2} f_X(x)$

so

$$G''_X(1) = \sum_{x=2}^{\infty} x(x-1) 1^{x-2} f_X(x) = \sum_{x=2}^{\infty} x(x-1) f_X(x) = \sum_{x=0}^{\infty} x(x-1) f_X(x) = E[X(X-1)].$$

(9) follows from (4) and (8) and the variance formula. \square

Theorem 7.3 (PGFs of common random variables) *For the discrete random variables encountered in Chapter 2, their probability generating functions are as follows:*

X	$G_X(t)$
$Unif(\{1, 2, \dots, n\})$	$\frac{t(t^n-1)}{n(t-1)}$
$Geom(p)$	$\frac{p}{1-t(1-p)}$
$NB(r, p)$	$\left[\frac{p}{1-t(1-p)} \right]^r$
$binomial(n, p)$	$(pt + 1 - p)^n$
$Pois(\lambda)$	$e^{\lambda(t-1)}$

PROOFS (OF SOME OF THESE)

Theorem 7.4 (Uniqueness of probability generating functions) *Let $X : \Omega \rightarrow \mathbb{N}$ and $Y : \Omega \rightarrow \mathbb{N}$ be a r.v. with the same pgf (i.e. $G_X(t) = G_Y(t)$ for all $t \in [-1, 1]$). Then $X = Y$ (i.e. $f_X(x) = f_Y(x)$ for all $x \in \mathbb{N}$).*

PROOF By property 6 of Theorem 7.2, the density f_X is determined by G_X . Thus if $G_X = G_Y$, $f_X = f_Y$, i.e. $X = Y$. \square

Theorem 7.5 (Independence property of pgfs) Let $X : \Omega \rightarrow \mathbb{N}$ and $Y : \Omega \rightarrow \mathbb{N}$ be independent r.v.s with respective pgfs G_X and G_Y . Then

$$G_{X+Y}(t) = G_X(t) G_Y(t).$$

PROOF

Corollary 7.6 Let X_1, \dots, X_d be independent r.v.s with respective pgfs G_{X_1}, \dots, G_{X_d} . Then:

$$G_{\sum_{j=1}^d X_j}(t) = \prod_{j=1}^d G_{X_j}(t).$$

A main application of probability generating functions is to derive facts about the sums of independent random variables:

Theorem 7.7 Suppose X_1, \dots, X_d are independent r.v.s, and let $S = X_1 + \dots + X_d$. Then:

1. If each X_j is $\text{Pois}(\lambda_j)$, then S is $\text{Pois}(\lambda_1 + \dots + \lambda_d)$.
2. If each X_j is $\text{binomial}(n_j, p)$ (same p), then S is $\text{binomial}(n_1 + \dots + n_d, p)$.
3. If each X_j is $\text{Geom}(p)$ (same p), then S is $\text{NB}(d, p)$.
4. If each X_j is $\text{NB}(r_j, p)$ (same p), then S is $\text{NB}(r_1 + \dots + r_d, p)$.

PROOF OF (1)

7.2 Moments and moment generating functions

Definition 7.8 Let $X : \Omega \rightarrow \mathbb{R}$ and let $r \in \{0, 1, 2, 3, \dots\}$. If X^r has finite expectation, then we define the r^{th} **moment** of X , denoted μ_r , to be $E[X^r]$. Otherwise, we say X **does not have a moment of order r** .

Heuristic analogy:

Definition 7.9 Given real-valued r.v. X (X can be cts or discrete), the **moment generating function (mgf)** of X , denoted M_X or g_X , is defined by

$$M_X(t) = E[e^{tX}].$$

The domain of M_X is the set of all $t \in \mathbb{R}$ such that e^{tX} has finite expectation.

Why is M_X called a “moment generating function”?

Theorem 7.10 (Properties of mgfs) Let $X : \Omega \rightarrow \mathbb{R}$ be a r.v. with mgf $M_X(t)$. Then:

1. $M_X(0) = 1$.
2. $M'_X(0) = EX = \mu_1$.
3. $M''_X(0) = EX^2 = \mu_2$.
4. $\text{Var}(X) = M''_X(0) - [M'_X(0)]^2$.
5. For all $r \in \{1, 2, 3, \dots\}$, $M_X^{(r)}(0) = \mu_r = E[X^r]$.
6. For any two constants a and b , $M_{aX+b}(t) = e^{bt} M_X(at)$.

Theorem 7.11 (Independence property of mgfs) Let $X : \Omega \rightarrow \mathbb{R}$ and $Y : \Omega \rightarrow \mathbb{R}$ be independent r.v.s with respective mgfs M_X and M_Y . Then

$$M_{X+Y}(t) = M_X(t) M_Y(t).$$

Similarly, if X_1, \dots, X_d are independent r.v.s with respective mgfs $M_{X_1}, M_{X_2}, \dots, M_{X_d}$, then:

$$M_{\sum_{j=1}^d X_j}(t) = \prod_{j=1}^d M_{X_j}(t).$$

PROOF HW (similar to proof for pgfs)

Moment generating functions of common random variables

Theorem 7.12 (MGFs of common r.v.s) For the common classes of random variables encountered in Chapters 2 and 3, their moment generating functions are as follows:

X	$M_X(t)$
$Unif(\{1, 2, \dots, n\})$	$\frac{e^t(e^{nt}-1)}{n(e^t-1)}$
$Geom(p)$	$\frac{p}{1-(1-p)e^t}$
$NB(r, p)$	$\left[\frac{p}{1-(1-p)e^t} \right]^r$
$binomial(n, p)$	$(1-p+pe^t)^n$
$Pois(\lambda)$	$e^{\lambda(e^t-1)}$
$Unif([a, b])$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
$Exp(\lambda)$	$\frac{\lambda}{\lambda-t}$ for $t < \lambda$
$\Gamma(r, \lambda)$	$\left(\frac{\lambda}{\lambda-t} \right)^r$ for $t < \lambda$
$std. normal \ n(0, 1)$	$e^{t^2/2}$
$normal \ n(\mu, \sigma^2)$	$\exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$

7.2. Moments and moment generating functions

PROOFS (OF SOME OF THESE) The Poisson and gamma r.v.s are left as HW.

It turns out that you can explicitly recover the density function of a real-valued r.v. from its moment generating function. The proof of the next theorem is beyond the scope of Math 414, as it uses complex analysis heavily:

Theorem 7.13 (Inversion formula) *Let $X : \Omega \rightarrow \mathbb{R}$ have mgf M_X . Then:*

1. *If X is discrete and integer-valued, then for every $x \in \mathbb{Z}$,*

$$f_X(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ixt} M_X(it) dt.$$

2. *If X is continuous, then X has density*

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} M_X(it) dt.$$

WARNING: If you are ever using these formulas to do a MATH 414 or 416 problem, you are doing the problem wrong. The significance of these formulas is that they explain the following principle:

Corollary 7.14 (Uniqueness of mgfs) *Let $X : \Omega \rightarrow \mathbb{R}$ and $Y : \Omega \rightarrow \mathbb{R}$ be any two real-valued r.v.s (not necessarily discrete or cts) such that $M_X(t) = M_Y(t)$. Then $X = Y$ (i.e. $F_X(x) = F_Y(x)$ for all x , so X and Y have the same density if they have a density).*

As with pgfs, an important application of mgfs is to establish results about the sum of independent random variables:

Theorem 7.15 (Sums of \perp r.v.s) *Suppose X_1, \dots, X_d are independent r.v.s, and let $S = X_1 + \dots + X_d$. Then:*

1. *If each X_j is $\text{Pois}(\lambda_j)$, then S is $\text{Pois}(\lambda_1 + \dots + \lambda_d)$.*
2. *If each X_j is $\text{binomial}(n_j, p)$ (same p), then S is $\text{binomial}(n_1 + \dots + n_d, p)$.*
3. *If each X_j is $\text{Geom}(p)$ (same p), then S is $\text{NB}(d, p)$.*
4. *If each X_j is $\text{NB}(r_j, p)$ (same p), then S is $\text{NB}(r_1 + \dots + r_d, p)$.*
5. *If each X_j is $\text{Exp}(\lambda)$ (same λ), then S is $\Gamma(d, \lambda)$.*
6. *If each X_j is $\Gamma(r_j, \lambda)$ (same λ), then S is $\Gamma(r_1 + \dots + r_d, \lambda)$.*
7. *If each X_j is $n(\mu_j, \sigma_j^2)$, then S is $n(\mu_1 + \dots + \mu_d, \sigma_1^2 + \dots + \sigma_d^2)$.*

PROOF (OF SOME OF THESE)

Joint moment generating functions

One can also define a joint moment generating function for a joint distribution:

Definition 7.16 Let X_1, \dots, X_d be real-valued r.v.s with some joint distribution \mathbf{X} . The **joint moment generating function** of \mathbf{X} , denoted $M_{\mathbf{X}}$ or $g_{\mathbf{X}}$, is the function $M_{\mathbf{X}} : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by

$$M_{\mathbf{X}}(\mathbf{t}) = E[e^{\mathbf{t} \cdot \mathbf{X}}].$$

The domain of $M_{\mathbf{X}}$ is the set of all $\mathbf{t} \in \mathbb{R}^d$ such that $e^{\mathbf{t} \cdot \mathbf{X}}$ has finite expectation.

Many of the same properties of moment generating functions carry over to the joint case:

Theorem 7.17 (Properties of joint mgfs) Let X_1, \dots, X_d be real-valued r.v.s with some joint mgf $M = M_{\mathbf{X}}$. Then:

1. $M(\mathbf{0}) = 1$.
2. For any real numbers a_1, \dots, a_d , $M_{a_1 X_1 + \dots + a_d X_d}(t) = M_{\mathbf{X}}(a_1 t, \dots, a_d t)$.

3. For each $j \in \{1, \dots, d\}$,

$$M_{X_j}(t) = M_{\mathbf{X}}(0, 0, \dots, 0, t, 0, \dots, 0) \text{ (the } t \text{ is in the } j^{\text{th}} \text{ position)}.$$

4. For any nonnegative integers r_1, \dots, r_d ,

$$E[X_1^{r_1} X_2^{r_2} \dots X_d^{r_d}] = \frac{\partial^{r_1 + \dots + r_d} M_{\mathbf{X}}}{\partial t_1^{r_1} \partial t_2^{r_2} \dots \partial t_d^{r_d}} \Big|_{\mathbf{t}=\mathbf{0}}.$$

5. For each $j \in \{1, \dots, d\}$,

$$E[X_j] = \frac{\partial M_{\mathbf{X}}}{\partial t_j} \Big|_{\mathbf{t}=\mathbf{0}} \text{ and } E[X_j^r] = \frac{\partial^r M_{\mathbf{X}}}{\partial t_j^r} \Big|_{\mathbf{t}=\mathbf{0}}.$$

6. For any $a \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^d$, $M_{a\mathbf{X}+\mathbf{b}}(\mathbf{t}) = e^{\mathbf{b} \cdot \mathbf{t}} M_{\mathbf{X}}(a \mathbf{t})$.

7. (Inversion formula) If \mathbf{X} is continuous, then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\mathbf{x} \cdot \mathbf{t}} M_{\mathbf{X}}(\mathbf{t}) d\mathbf{t}.$$

8. (Uniqueness) If $M_{\mathbf{X}} = M_{\mathbf{Y}}$, then \mathbf{X} and \mathbf{Y} have the same joint distribution.

Theorem 7.18 (Independence test using joint mgfs) Let X_1, \dots, X_d be real-valued r.v.s. Then X_1, \dots, X_d are independent if and only if

$$M_{\mathbf{X}}(\mathbf{t}) = \prod_{j=1}^d M_{X_j}(t_j)$$

for all $\mathbf{t} = (t_1, \dots, t_d) \in \mathbb{R}^d$.

PROOF

7.3 Joint normal densities

In this section, we generalize the idea of a normal random variable to higher dimensions. The following class of joint distributions arise in statistics, econometrics, signal processing and other fields:

Definition 7.19 A collection $\mathbf{X} = (X_1, \dots, X_d)$ of real-valued r.v.s with joint density $f_{\mathbf{X}} : \mathbb{R}^d \rightarrow \mathbb{R}$ is called **joint(ly) normal** or **joint(ly) Gaussian** if every finite linear combination

$$\sum_{j=1}^d b_j X_j$$

(where $b_1, \dots, b_d \in \mathbb{R}$) is normal.

Observe: \mathbf{X} joint normal \Rightarrow each X_j is normal

(The converse of this is false: just because the X_j are normal does not mean they have a joint normal distribution.)

Write $\mu_j = EX_j$, $\sigma_j = \sqrt{\text{Var}(X_j)}$ so that each X_j is $n(\mu_j, \sigma_j^2)$. Now let

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}_{d \times 1} ; \quad \sigma = \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_d \end{pmatrix}_{d \times 1} ; \quad \sigma^2 = \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \vdots \\ \sigma_d^2 \end{pmatrix}_{d \times 1} .$$

We are thinking of these vectors as $d \times 1$ matrices, so their transposes are row matrices:

$$\mu^T = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_d \end{pmatrix}_{1 \times d}, \text{ etc.}$$

Definition 7.20 Let \mathbf{X} be the joint distribution of real-valued r.v.s X_1, \dots, X_d . The **covariance matrix** Σ of \mathbf{X} is

$$\Sigma = \begin{pmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_d) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \cdots & \text{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_d, X_1) & \text{Cov}(X_d, X_2) & \cdots & \text{Cov}(X_d, X_d) \end{pmatrix}_{d \times d} .$$

Theorem 7.21 (Properties of covariance matrices) Let Σ be the covariance matrix of any joint distribution $\mathbf{X} = (X_1, \dots, X_d)$. Then:

1. Σ is $d \times d$.
2. Σ is symmetric (i.e. $\Sigma^T = \Sigma$).
3. The diagonal entries of Σ are the variances of the X_j .
4. For any vector $\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_d \end{pmatrix}$, $Var(\mathbf{b} \cdot \mathbf{X}) = Var\left(\sum_{j=1}^d b_j X_j\right) = \mathbf{b}^T \Sigma \mathbf{b}$
5. Σ is nonnegative definite (for any vector $\mathbf{b} \in \mathbb{R}^d$, $\mathbf{b}^T \Sigma \mathbf{b} \geq 0$).
6. If \mathbf{X} is joint normal, then Σ is invertible.

PROOF

Corollary 7.22 Suppose \mathbf{X} has a joint normal density with mean vector μ and covariance matrix Σ . Let $Y = b_1 X_1 + \dots + b_d X_d = \mathbf{b} \cdot \mathbf{X}$. Then Y is normal with parameters

$$EY = \mathbf{b} \cdot \mu; \quad Var(Y) = \mathbf{b}^T \Sigma \mathbf{b}.$$

The main fact in this section is that the joint density function of a joint normal distribution can be written down in terms of the mean vector and covariance matrix:

Theorem 7.23 (Characterization of joint normal densities) Suppose X_1, \dots, X_d are jointly normal r.v.s with mean vector $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix}$ and covariance matrix Σ . Then the joint density function is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right].$$

Corollary 7.24 (Uniqueness of joint normal densities) If two jointly normal distributions have the same means and same covariances between the marginals, then they are the same distribution.

Corollary 7.25 If \mathbf{X} is a joint normal density such that $\text{Cov}(X_i, X_j) = 0$ for all $i \neq j$, then $X_i \perp X_j$ for all $i \neq j$.

PROOF (of the theorem):

This proof proceeds in three parts: first, we will assume \mathbf{X} is joint normal, and compute the joint moment generating function of \mathbf{X} . Second, we will assume \mathbf{X} has the joint density of the theorem, and compute the joint mgf of \mathbf{X} from that. These mgfs will be the same, so by uniqueness of mgfs the theorem will follow.

First, suppose \mathbf{X} is joint normal with mean vector μ and covariance matrix Σ .

Let $V = \mathbf{t} \cdot \mathbf{X} = t_1 X_1 + \dots + t_d X_d$. That means that by the definition of joint normal, V is normal and from Corollary 7.20, we see that

$$EV = E[\mathbf{t} \cdot \mathbf{X}] = \mathbf{t} \cdot \mu$$

$$\text{Var}(V) = \mathbf{t}^T \Sigma \mathbf{t}$$

Therefore V is $n(\mathbf{t} \cdot \mu, \mathbf{t}^T \Sigma \mathbf{t})$ so

$$M_{\mathbf{X}}(\mathbf{t}) = E[e^{\mathbf{t} \cdot \mathbf{X}}] = E[e^V] = E[e^{1V}] = M_V(1) = \exp \left(\mathbf{t} \cdot \mu + \frac{1}{2} \mathbf{t}^T \Sigma \mathbf{t} \right).$$

Second, suppose \mathbf{X} has the density given in the theorem:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left[\frac{-1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

Then

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= E \left[e^{\mathbf{t} \cdot \mathbf{x}} \right] \\ &= \int \int \cdots \int_{\mathbb{R}^d} e^{\mathbf{t} \cdot \mathbf{x}} \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left[\frac{-1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right] d\mathbf{x} \end{aligned}$$

FACT: If Σ is symmetric, invertible and positive definite, then there is a positive definite, symmetric, invertible matrix $\sqrt{\Sigma}$ such that $(\sqrt{\Sigma})^2 = \Sigma$.

Let $\mathbf{y} = g(\mathbf{x}) = (\sqrt{\Sigma})^{-1} (\mathbf{x} - \mu)$

so that $J(g) = \det[(\sqrt{\Sigma})^{-1}] = \frac{1}{\sqrt{\det \Sigma}}$ and $\mathbf{x} = \sqrt{\Sigma} \mathbf{y} + \mu$.

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= \int \int \cdots \int_{\mathbb{R}^d} e^{\mathbf{t} \cdot (\sqrt{\Sigma} \mathbf{y} + \mu)} \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp \left[\frac{-1}{2} (\sqrt{\Sigma} \mathbf{y})^T \Sigma^{-1} (\sqrt{\Sigma} \mathbf{y}) \right] \frac{1}{J(g)} d\mathbf{y} \\ &= \frac{e^{\mathbf{t} \cdot \mu}}{(2\pi)^{d/2}} \int \int \cdots \int_{\mathbb{R}^d} e^{\mathbf{t} \cdot \sqrt{\Sigma} \mathbf{y}} \exp \left[\frac{-1}{2} \mathbf{y}^T \mathbf{y} \right] d\mathbf{y} \end{aligned}$$

Now let $\mathbf{w} = \sqrt{\Sigma} \mathbf{t}$. Then $\mathbf{t} \cdot \sqrt{\Sigma} \mathbf{y} = \mathbf{w} \cdot \mathbf{y}$ and

$$\sum_{j=1}^d w_j^2 = \mathbf{w} \cdot \mathbf{w} = \mathbf{w}^T \mathbf{w} = \mathbf{t}^T \Sigma \mathbf{t}.$$

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= \frac{e^{\mathbf{t} \cdot \mu}}{(2\pi)^{d/2}} \int \int \cdots \int_{\mathbb{R}^d} e^{\mathbf{w} \cdot \mathbf{y}} \exp \left[\frac{-1}{2} \mathbf{y}^T \mathbf{y} \right] d\mathbf{y} \\ &= \frac{e^{\mathbf{t} \cdot \mu}}{(2\pi)^{d/2}} \int \int \cdots \int_{\mathbb{R}^d} \exp \left[\sum_{j=1}^d w_j y_j \right] \exp \left[\frac{-1}{2} \sum_{j=1}^d y_j^2 \right] dy_j \\ &= \frac{e^{\mathbf{t} \cdot \mu}}{(2\pi)^{d/2}} \prod_{j=1}^d \int_{-\infty}^{\infty} \exp [w_j y_j] \exp \left[\frac{-1}{2} y_j^2 \right] dy_j \\ &= \frac{e^{\mathbf{t} \cdot \mu}}{(2\pi)^{d/2}} \prod_{j=1}^d \int_{-\infty}^{\infty} e^{-i(w_j i) y_j} M_{n(0,1)}(i y_j) dy_j \quad \text{where } i = \sqrt{-1} \\ &= \frac{e^{\mathbf{t} \cdot \mu}}{(2\pi)^{d/2}} \prod_{j=1}^d 2\pi f_{n(0,1)}(w_j i) \quad \text{by the Inversion Formula} \end{aligned}$$

From the previous page, we have

$$\begin{aligned}
 M_{\mathbf{X}}(\mathbf{t}) &= \frac{e^{\mathbf{t} \cdot \boldsymbol{\mu}}}{(2\pi)^{d/2}} \prod_{j=1}^d 2\pi f_{n(0,1)}(w_j i) \\
 &= \frac{e^{\mathbf{t} \cdot \boldsymbol{\mu}}}{(2\pi)^{d/2}} (2\pi)^d \prod_{j=1}^d \frac{1}{\sqrt{2\pi}} \exp \left[\frac{-(i w_j)^2}{2} \right] \\
 &= e^{\mathbf{t} \cdot \boldsymbol{\mu}} \exp \left(- \sum_{j=1}^d \frac{(i w_j)^2}{2} \right) \\
 &= \exp \left(\mathbf{t} \cdot \boldsymbol{\mu} + \sum_{j=1}^d \frac{w_j^2}{2} \right) \\
 &= \exp \left(\mathbf{t} \cdot \boldsymbol{\mu} + \frac{1}{2} \mathbf{w} \cdot \mathbf{w} \right) \\
 &= \exp \left(\mathbf{t} \cdot \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}^T \Sigma \mathbf{t} \right).
 \end{aligned}$$

This is the same joint MGF as originally computed when we assumed the density was joint normal; by uniqueness of joint MGFs the joint normal density must have this form. \square

Bivariate normal densities

Now we take this general theory and restrict it to the case where $d = 2$ (i.e. we have a jointly normal distribution of two variables X and Y). In this setting we say X and Y have a **bivariate normal distribution**.

Let X and Y have a bivariate normal distribution. Thus X and Y are each normal (as is any linear combination of X and Y). Let

$$\mu_X = EX; \sigma_X^2 = Var(X); \mu_Y = EY; \sigma_Y^2 = Var(Y); \sigma_{XY} = Cov(X, Y).$$

Thus

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$$

so

$$\Sigma^{-1} = \frac{1}{\det \Sigma} \begin{pmatrix} \sigma_Y^2 & -\sigma_{XY} \\ -\sigma_{XY} & \sigma_X^2 \end{pmatrix}.$$

Write

$$\Sigma^{-1} = \begin{pmatrix} a & b \\ b & d \end{pmatrix}.$$

Now

$$f_{X,Y}(x, y) = \frac{1}{(2\pi)^{2/2} \sqrt{\det \Sigma}} \exp \left[\frac{-1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

From the previous page, we have

$$f_{X,Y}(x,y) = \frac{1}{(2\pi)^{2/2} \sqrt{\det \Sigma}} \exp \left[\frac{-1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

Now

$$\begin{aligned} (\star) &= \frac{-1}{2} \begin{pmatrix} x - \mu_X & y - \mu_Y \end{pmatrix} \begin{pmatrix} a & b \\ b & d \end{pmatrix} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix} \\ &= \frac{-1}{2} \begin{pmatrix} x - \mu_X & y - \mu_Y \end{pmatrix} \begin{pmatrix} ax - a\mu_X & by - b\mu_Y \\ bx - b\mu_X & dy - d\mu_Y \end{pmatrix} \\ &= \frac{-1}{2} \left[ax^2 - a\mu_X x + bxy - b\mu_Y x - a\mu_X x + a\mu_X^2 - b\mu_X y + b\mu_X \mu_Y \right. \\ &\quad \left. + bxy - b\mu_X y + dy^2 - d\mu_Y y - b\mu_Y x + b\mu_X \mu_Y - d\mu_X \mu_Y + d\mu_Y^2 \right] \\ &= \frac{-1}{2} \left[ax^2 + 2bxy + dy^2 + (-2a\mu_X - 2b\mu_Y)x + (-2b\mu_X - 2d\mu_Y)y \right. \\ &\quad \left. + (a\mu_X^2 + 2b\mu_X \mu_Y + d\mu_Y^2) \right] \\ &= \frac{-1}{2} ax^2 - bxy - \frac{1}{2} dy^2 + (a\mu_X + b\mu_Y)x + (b\mu_X + d\mu_Y)y + (\text{constant}). \end{aligned}$$

Punchline: Given $f_{X,Y}$ for a bivariate normal distribution (X, Y) :

1.

2.

3.

EXAMPLE 1

Suppose X and Y have bivariate normal density

$$f_{X,Y}(x, y) = K \exp \left[\frac{-1}{2} (5x^2 - 6xy + 2y^2 - 40x + 24y + 80) \right].$$

Find the expected values and variances of X and Y ; find the marginal densities; find the covariance between X and Y ; compute Σ and K . Last, find the expected value and variance of $Z = 7X - 3Y$.

The last order of business is to compute the conditional density of Y given X . First, this density must be normal. To see this, note that X is normal so (here the $(*)$ s represent arbitrary constants)

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{(*) \exp [(x) x^2 + (x) xy + (x) y^2 + (x) x + (x) y + (x)]}{(*) \exp [(x) x^2 + (x) x + (x)]} \\ &= (*) \exp [(x) x^2 + (x) xy + (x) y^2 + (x) x + (x) y + (x)] \\ &= \frac{1}{\sqrt{\text{Var}(Y|X)}\sqrt{2\pi}} \exp \left[\frac{-(y - E(Y|X))^2}{2\text{Var}(Y|X)} \right]. \end{aligned}$$

so $Y|X$ is $n(E(Y|X), \text{Var}(Y|X))$.

Now we need to compute the conditional expectation and conditional variance of Y given X :

Theorem 7.26 Suppose (X, Y) have a bivariate normal density. Then $Y|X$ is normal with parameters

$$E(Y|X) = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X); \quad \text{Var}(Y|X) = \sigma_Y^2(1 - \rho^2).$$

(Here ρ is the correlation between X and Y .)

(Note that the X s and Y s can be reversed in the above formula to give $E(X|Y)$ and $\text{Var}(X|Y)$.)

EXAMPLE 2

Suppose (X, Y) have a bivariate normal density where $E(X|Y) = 3.7 - .15y$; $E(Y|X) = .4 - .6x$ and $\text{Var}(Y|X) = 3.64$.

1. Compute EX .
2. Compute $\text{Var}(X)$.
3. Compute EY .
4. Compute $\text{Var}(Y)$.
5. Compute $\rho(X, Y)$.
6. Compute $\text{Cov}(X, Y)$.

Chapter 8

Limit theorems

8.1 Markov and Chebyshev inequalities

In this section we discuss inequalities which give us quick bounds on certain probabilities related to the mean and variance of a random variable.

Theorem 8.1 (Markov inequality) *Let $X : \Omega \rightarrow [0, \infty)$ be a nonnegative r.v. with finite expected value. Then for all $a > 0$,*

$$P(X \geq a) \leq \frac{EX}{a}.$$

PROOF Let $I : \Omega \rightarrow \{0, a\}$ be defined by

$$I(\omega) = \begin{cases} a & \text{if } X \geq a \\ 0 & \text{else} \end{cases}$$

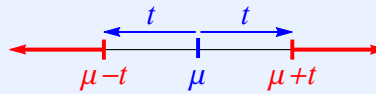
Notice that $X \geq I$, so

$$\begin{aligned} EX &\geq EI = a \cdot P(I = a) + 0 \cdot P(I = 0) \\ &= aP(X \geq a). \end{aligned}$$

Divide both sides by a to get the result. \square

Application: Suppose the time it takes for a radioactive element to decay is a random variable whose mean is 23. Use the Markov inequality to find an upper bound on the probability that it will take at least 230 units of time for the element to decay.

Theorem 8.2 (Chebyshev inequality) Let $X : \Omega \rightarrow \mathbb{R}$ be a r.v. with finite expected value μ and finite variance σ^2 . Then for all $t > 0$,



The diagram shows a horizontal number line with a central point labeled μ . Two points, $\mu - t$ and $\mu + t$, are marked on the line. Red arrows point outwards from $\mu - t$ and $\mu + t$, indicating the tails of the distribution. Blue double-headed arrows above the line indicate the distance t from μ to $\mu - t$ and from μ to $\mu + t$.

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

PROOF Apply the Markov inequality to the r.v. $(X - \mu)^2$ with $a = t^2$ to get

$$P((X - \mu)^2 \geq t^2) \leq \frac{E[(X - \mu)^2]}{t^2} = \frac{Var(X)}{t^2}.$$

But $P(|X - \mu| \geq t) = P((X - \mu)^2 \geq t^2)$. This proves the result. \square

Application: Suppose the number of items produced in a factory is a random variable with mean 100 and variance 40. Use the Chebyshev inequality to find a lower bound on the probability that between 90 and 110 items will be produced by the factory.

8.2 Laws of large numbers

We are interested in studying the results of an experiment which is repeated over and over.

Definition 8.3 A discrete-time stochastic process $\{X_t : t \in \mathbb{N}\}$ is called an **i.i.d. process** if the process is “independent and identically distributed”, i.e. $X_j \perp X_k$ for all $j \neq k$ and the X_j have the same distribution for all j .

The prototype example of an i.i.d. process is coin flipping: if you flip the same coin over and over again and let

$$X_j = \begin{cases} 1 & \text{if } j^{\text{th}} \text{ flip is heads} \\ 0 & \text{if } j^{\text{th}} \text{ flip is tails} \end{cases}$$

What we are most interested in is the **average** behavior of such a process.

Definition 8.4 Given an i.i.d. process $\{X_t\}$, define the following processes:

1. $\{S_n\}_{n \in \mathbb{N}}$, the **sequence of sums**, is $S_n = X_1 + X_2 + \dots + X_n$;
2. $\{A_n\}_{n \in \mathbb{N}}$ the **sequence of averages**, is

$$A_n = \frac{1}{n} S_n = \frac{X_1 + \dots + X_n}{n};$$

3. $\{A_n^*\}_{n \in \mathbb{N}}$, the **sequence of normalized averages**, is

$$A_n^* = \frac{A_n - E[A_n]}{\sqrt{\text{Var}(A_n)}} = \frac{A_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{S_n - \mu n}{\sigma \sqrt{n}}.$$

Notice that if each X_t has mean μ and variance σ^2 , then

$$E[A_n] =$$

$$\text{Var}(A_n) =$$

$$E[A_n^*] =$$

$$\text{Var}(A_n^*) =$$

Applying Chebyshev's inequality to the random variable A_n , we obtain:

Theorem 8.5 (Quantitative Weak Law of Large Numbers (QWLLN)) *Let $\{X_t\}$ be an i.i.d. process, where each X_j is a r.v. with finite expected value μ and finite variance σ^2 . For each $n \in \mathbb{N}$, set $A_n = \frac{1}{n}(X_1 + \dots + X_n)$. Then for all $\delta > 0$,*

$$P(|A_n - \mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2}.$$

The above theorem says that if you fix an “error tolerance” δ , if you take enough measurements (n), then the probability that the average of your measurements A_n is within δ of the theoretical average μ is high.

Application: Marbles are drawn from a jar containing 3 red and 5 marbles, one at a time with replacement. What is the smallest number n such that you can be 99% assured that between 37% and 38% of the first n marbles drawn are red?

By taking a limit of each side of the inequality in the QWLLN as $n \rightarrow \infty$, we obtain the following important theoretical result:

Theorem 8.6 (Weak Law of Large Numbers (WLLN)) *Let $\{X_t\}$ be an i.i.d. process, where each X_j is a r.v. with finite expected value μ and finite variance σ^2 . For each $n \in \mathbb{N}$, set $A_n = \frac{1}{n}(X_1 + \dots + X_n)$. Then for all $\delta > 0$,*

$$\lim_{n \rightarrow \infty} P(|A_n - \mu| \geq \delta) = 0.$$

Remark: One can derive the WLLN without the assumption that the X_j have finite variance.

Loosely speaking, the WLLN says that if you take a large sample, the probability that the average of your sample is within δ of the “theoretical average” of each measurement (i.e. the expected value μ) is large, and that as the size of the sample increases, this probability goes to 1. Let’s think about what this means in terms of flipping a fair coin repeatedly:

Let $\{X_t\}$ be an i.i.d sequence of r.v.s, each uniform on $\{0, 1\}$ (think of $X_j = 1$ as corresponding to the j^{th} flip being heads and $X_j = 0$ meaning the j^{th} flip being tails). In this setting, $\mu = EX_j = \frac{1}{2}$.

Under these assumptions, what is A_n ?

Let $\delta = \frac{1}{10}$. Let’s say that a sequence of flips is “ n -good” (or “ n, δ -good”) if $|A_n - \mu| < \delta$, i.e. the proportion of heads in the first n flips is between $\frac{4}{10}$ and $\frac{6}{10}$.

Ex: H,H,H,H,T,T,T,... is not $4, \frac{1}{10}$ -good, but is $8, \frac{1}{10}$ -good.

The WLLN says that if you fix a δ , then if you choose a large enough n , most sequences are n, δ -good.

HOWEVER: what the WLLN doesn’t tell you (and why it is called the “Weak” LLN) is any relationship between sequences that are good at different values of n . For example, it does not guarantee that most sequences are “eventually good”, i.e. are n -good for all sufficiently large n (for example, it might be the case that typical sequences of heads and tails are n -bad for infinitely many, very sparsely spaced n).

This weakness is fixed with the following stronger result, which says (among other things) that with probability 1, a randomly chosen sequence of heads and tails is eventually good (i.e. the proportion of heads in the sequence becomes close to $\frac{1}{2}$ and stays close to $\frac{1}{2}$ forever:

Theorem 8.7 (Strong Law of Large Numbers (SLLN)) *Let $\{X_t\}$ be an i.i.d. process, where each X_j is a r.v. with finite expected value μ . Then*

$$P\left(\lim_{n \rightarrow \infty} \frac{S_n}{n} = \mu\right) = 1$$

PROOF (with the extra assumption that $EX^4 < \infty$) Suppose first that $EX_j = 0$. That means $Var(X_j) = EX_j^2 - 0^2 = EX_j^2$. Now,

$$\begin{aligned} E[S_n^4] &= E[(X_1 + \dots + X_n)^4] \\ &= E[(X_1 + \dots + X_n)(X_1 + \dots + X_n)(X_1 + \dots + X_n)(X_1 + \dots + X_n)] \end{aligned}$$

$$\begin{aligned} \Rightarrow E[S_n^4] &= nEX_j^4 + \binom{n}{2} \binom{4}{2} [Var(X_j)]^2 \\ &= nEX_j^4 + 3n(n-1)[Var(X_j)]^2 \end{aligned}$$

$$\leq nEX_j^4 + 3n(n-1)EX_j^4.$$

Therefore

$$E\left[\frac{S_n^4}{n^4}\right] \leq \frac{nEX_j^4 + 3n(n-1)EX_j^4}{n^4} \leq \frac{1}{n^3}EX_j^4 + \frac{3}{n^2}EX_j^4$$

so

$$\lim_{n \rightarrow \infty} E[A_n^4] = \lim_{n \rightarrow \infty} E\left[\left(\frac{S_n}{n}\right)^4\right] = \lim_{n \rightarrow \infty} E\left[\frac{S_n^4}{n^4}\right] = 0.$$

By definiteness, $\lim_{n \rightarrow \infty} A_n^4 = 0$ with probability 1, so $\lim_{n \rightarrow \infty} A_n = 0$ with probability 1 as wanted.

If $EX_j = \mu \neq 0$, then apply the above to $X_j - \mu$ to see that $\lim_{n \rightarrow \infty} (A_n - \mu) = 0$ with probability 1, i.e. $\lim_{n \rightarrow \infty} A_n = \mu$ with probability 1 as wanted. \square

8.3 Central Limit Theorem

Question: Let $\{X_t\}$ be an i.i.d. process with normalized averages A_n^* . What happens to the distribution of A_n^* as $n \rightarrow \infty$?

To answer this question, let's use moment generating functions. Suppose that each X_j has moment generating function $M_X(t)$. Then

$$M_{S_n}(t) = M_{X_1+\dots+X_n}(t) = \prod_{j=1}^n M_{X_j}(t) = [M_X(t)]^n$$

and therefore

$$\begin{aligned} M_{A_n^*}(t) &= E[e^{tA_n^*}] = E\left[\exp\left(t\frac{S_n - n\mu}{\sigma\sqrt{n}}\right)\right] \\ &= E\left[\exp\left(\frac{t}{\sigma\sqrt{n}}S_n - \frac{n\mu t}{\sigma\sqrt{n}}\right)\right] \\ &= E\left[\exp\left(\frac{t}{\sigma\sqrt{n}}S_n\right)\right] \cdot \exp\left(\frac{-\mu tn}{\sqrt{n}\sigma}\right) \\ &= M_{S_n}\left(\frac{t}{\sigma\sqrt{n}}\right) \cdot \exp\left(\frac{-\mu tn}{\sqrt{n}\sigma}\right) \\ &= \left[M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \cdot \exp\left(\frac{-\mu tn}{\sqrt{n}\sigma}\right) \\ &= \exp\left[n\left(\ln M_X\left(\frac{t}{\sigma\sqrt{n}}\right) - \frac{\mu t}{\sigma\sqrt{n}}\right)\right] \\ &= \exp[\Delta]. \end{aligned}$$

Thus

$$\lim_{n \rightarrow \infty} M_{A_n^*}(t) = \lim_{n \rightarrow \infty} \exp[\Delta] = \exp\left[\lim_{n \rightarrow \infty} \Delta\right].$$

Now if $t = 0$,

$$\lim_{n \rightarrow \infty} \Delta = \lim_{n \rightarrow \infty} n(\ln M_X(0) - 0) = \lim_{n \rightarrow \infty} n(\ln 1 - 0) = n \cdot 0 = 0 = \frac{0^2}{2}$$

and if $t \neq 0$,

$$\begin{aligned}
 \lim_{n \rightarrow \infty} \Delta &= \lim_{n \rightarrow \infty} n \left(\ln M_X \left(\frac{t}{\sigma\sqrt{n}} \right) - \frac{\mu t}{\sigma\sqrt{n}} \right) \\
 &= \lim_{n \rightarrow \infty} \frac{\frac{t^2}{\sigma^2} \cdot \left(\ln M_X \left(\frac{t}{\sigma\sqrt{n}} \right) - \frac{\mu t}{\sigma\sqrt{n}} \right)}{\frac{t^2}{\sigma^2} \cdot \frac{1}{n}} \\
 &= \frac{t^2}{\sigma^2} \lim_{n \rightarrow \infty} \frac{\ln M_X \left(\frac{t}{\sigma\sqrt{n}} \right) - \frac{\mu t}{\sigma\sqrt{n}}}{\left(\frac{t}{\sigma\sqrt{n}} \right)^2} \\
 &\quad (\text{let } s = \frac{t}{\sigma\sqrt{n}} \text{ so that as } n \rightarrow \infty, s \rightarrow 0) \\
 &= \frac{t^2}{\sigma^2} \lim_{s \rightarrow 0} \frac{\ln M_X(s) - \mu s}{s^2} = \frac{0}{0} \\
 &\stackrel{L}{=} \frac{t^2}{\sigma^2} \lim_{s \rightarrow 0} \frac{\frac{M'_X(s)}{M_X(s)} - \mu}{2s} = \frac{0}{0} \\
 &\stackrel{L}{=} \frac{t^2}{\sigma^2} \lim_{s \rightarrow 0} \frac{\frac{M''_X(s)M_X(s) - (M'_X(s))^2}{[M_X(s)]^2}}{2} \\
 &= \frac{t^2}{\sigma^2} \frac{\frac{EX^2 \cdot 1 - (EX)^2}{1^2}}{2} = \frac{t^2}{\sigma^2} \cdot \frac{\text{Var}(X)}{2} = \frac{1}{2}t^2.
 \end{aligned}$$

We have proven that for any t ,

$$\lim_{n \rightarrow \infty} M_{A_n^*}(t) = \exp \left(\frac{t^2}{2} \right) = M_{n(0,1)}(t).$$

Therefore as $n \rightarrow \infty$, by uniqueness of moment generating functions, A_n^* must approach that of the standard normal **no matter what the original density of X was**, i.e. for large n ,

$$P(A_n^* \leq x) \approx P(n(0,1) \leq x) = \Phi(x).$$

The content of this argument is what is called the Central Limit Theorem:

Theorem 8.8 (Central Limit Theorem (CLT)) *Let $\{X_n\}$ be an i.i.d. process such that each of the X_j have finite mean μ and finite variance σ^2 . Then if A_n^* is the normalized average of the first n X_j as defined above,*

$$\lim_{n \rightarrow \infty} P(A_n^* \leq x) = \Phi(x)$$

where Φ is the cdf of the standard normal.

WARNING: If the X_j do not have finite mean or do not have finite variance (for example if X_j is Cauchy), then the CLT does not apply.

Corollary 8.9 *Let $\{X_n\}$ be an i.i.d. process such that each of the X_j have finite mean μ and finite variance σ^2 . Then:*

1. *if A_n is the (non-normalized) average of the first n X_j , A_n is approximated by a normal r.v. with parameters μ and $\frac{\sigma^2}{n}$.*
2. *if S_n is the sum of the first n X_j , S_n is approximated by a normal r.v. with parameters μn and $n\sigma^2$.*

This last corollary is usually shorthanded as follows:

$$\text{"If } \{X_t\} \text{ is i.i.d., then } \boxed{A_n \approx n \left(\mu, \frac{\sigma^2}{n} \right)} \text{ and } \boxed{S_n \approx n \left(\mu n, n\sigma^2 \right)}."$$

EXAMPLE 1

A basketball player expects to make 80% of his free throws (assume the result of each free throw is independent of any of the others). Use the Central Limit Theorem to estimate the probability that he makes at least 261 of 300 attempts.

EXAMPLE 2

Two candidates (A and B) run against each other in an election. Suppose candidate A is favored by p of the population ($0 < p < 1$).

- (a) Suppose 1600 people are polled as to the candidate they support. What is the probability that the poll result differs from the actual situation by more than 2.5%?
- (b) If 1600 people are polled, we can be 99% sure that the poll results are accurate to what percent?
- (c) How many people need to be polled so that the pollster is 99% sure the poll is accurate to within 1%?

Chapter 9

Applications to insurance

9.1 Deductibles

A more interesting application of the formula for the expected value of a transformation occurs in the context of insurance. Most of the time, when you buy an insurance policy, the policy includes a **deductible** of some fixed amount d . That means that when you incur a loss covered by the insurance policy, you must pay the first d of the loss; the insurance company only covers anything that is left after that.

Why do insurance companies like deductibles?

1. Less to pay out to policyholders
2. Reduced overhead costs (no processing of small claims)
3. Creates some risk for policyholder, which gives incentive for policyholder to be risk-averse

Suppose a policyholder holding a policy with a deductible d incurs loss X (where X is some r.v.). If Y is the claim payment, what is Y as a function of X ?

On the previous page we saw that

$$Y = \varphi(X) = \begin{cases} 0 & \text{if } X \leq d \\ X - d & \text{if } X > d \end{cases}$$

This means the cdf of Y can be computed as follows:

Therefore, the density function of Y is:

Now, let's use LOTUS to compute EY :

We have proven:

Theorem 9.1 (Expected value for insurance policy with deductible) *Suppose the loss incurred by a policyholder with deductible d is a cts r.v. X with finite expectation. If Y is the claim payment associated to this loss, then*

$$EY = \int_d^\infty (x - d) f_X(x) dx.$$

EXAMPLE 1

Suppose that a loss random variable is uniform on $[0, 10]$.

1. Find the expected amount paid by the insurer if the policy has a deductible of 3.
2. Find the variance of the amount paid by the insurer if the policy has a deductible of 3.
3. If a deductible of size d is applied before any insurance payment, and the expected payment of the insurer is 1.5, find the size of the deductible.

EXAMPLE 2

Suppose that a loss random variable is exponential with mean 10. If a deductible of size 5 is applied, find the expected payment of the insurer.

Solution: X exp. w/ mean 10 means $X \sim \text{Exp}(\quad)$, i.e. $f_X(x) =$

$$\begin{aligned} EY &= \int_d^\infty (x - d)f_X(x) dx \\ &= \int_5^\infty (x - 5)\frac{1}{10}e^{-(1/10)x} dx \end{aligned}$$

9.2 Benefit limits

A second way that insurance companies mitigate their risk is by selling policies that have **benefit limits** (a.k.a. **coverage limits**). Suppose a policy has a benefit limit of l (and no deductible). This means that the maximum amount the insurance company will pay its policyholder for a loss is l . Then if loss X is incurred by the policyholder, the corresponding claim payment Y is

If there is both a benefit limit l and a deductible d , then if loss X is incurred by the policyholder, the corresponding claim payment Y is

Using the LOTUS formula, we can derive these formulas:

Theorem 9.2 (Expected value for insurance policy with benefit limit) *Suppose the loss X incurred by policyholder with a benefit limit l is a continuous r.v. with finite expectation. Then if Y is the claim payment associated to this loss,*

$$EY = \int_0^l x f_X(x) dx + l \cdot P(X \geq l).$$

PROOF This will follow from the next theorem by setting $d = 0$. \square

Theorem 9.3 (Exp. value for policy with deductible and benefit limit) *Suppose the loss X incurred by policyholder with deductible d and benefit limit l is a continuous r.v. with finite expectation. Then if Y is the claim payment associated to this loss,*

$$EY = \int_d^{d+l} (x - d) f_X(x) dx + l \cdot P(X \geq d + l).$$

PROOF From the previous page, we know

$$Y = \varphi(X) = \begin{cases} 0 & \text{if } X \leq d \\ X - d & \text{if } d < X < d + l \\ l & \text{if } X \geq d + l \end{cases}$$

So by LOTUS, we have

$$\begin{aligned} EY &= \int_0^\infty \varphi(x) f_X(x) dx \\ &= \int_0^d 0 f_X(x) dx + \int_d^{d+l} (x - d) f_X(x) dx + \int_{d+l}^\infty l f_X(x) dx \\ &= 0 + \int_d^{d+l} (x - d) f_X(x) dx + l \int_{d+l}^\infty f_X(x) dx \\ &= \int_d^{d+l} (x - d) f_X(x) dx + l \cdot P(X \geq d + l). \quad \square \end{aligned}$$

EXAMPLE 3

Suppose that a loss random variable is uniform on $[0, 10]$. Find the expected amount paid by the insurer, if the policy has a deductible of 1 and a coverage limit of 6.

EXAMPLE 4

Suppose the loss from an accident is a continuous random variable with density $f(x) = \frac{24}{7}x^{-4}$ when $1 < x < 2$. Suppose that the insurance policy has a coverage limit of 1.5. What is the standard deviation of the loss to the insurance company?

9.3 Proportional coverage

A third way insurance companies limit their exposure is by offering **proportional coverage**. This means that they only cover a fraction of the loss, as opposed to the entire loss. To compute quantities associated to proportional coverage, think of the original loss as X and the claim payment as Y . Write Y as a piecewise-defined function φ of X and answer the question asked (if the question asks for an expected value, variance or standard deviation, you have to do several integrals separately according to each piece of the function φ).

EXAMPLE 5

Suppose that the damage (in thousands of dollars) caused when a piece of equipment breaks is given by a continuous random variable with density $f(x) = \frac{2}{x^3}$ when $x > 1$. Suppose that the piece of equipment breaks 25% of the time. If an insurance company agrees to cover 100% of the first \$3000 in damage and 50% of the next \$3000 in damage, what is the expected value of the amount the insurance company will have to pay?

EXAMPLE 6

The cumulative distribution function for health care costs experienced by a policyholder is

$$F(x) = \begin{cases} 1 - e^{-x/100} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The policy has a deductible of 20. An insurer reimburses the policyholder for 100% of health care costs between 20 and 120 (less the deductible); health care costs from 120 to 420 are reimbursed at 50%; health care costs above 420 are not reimbursed. Find the cdf of the reimbursements.

Chapter 10

Homework exercises

10.1 Exercises from Chapter 1

1. Suppose you flip a fair coin three times, and record the outcomes with H s and T s. Describe the following events in words (your description should be as efficient as possible):
 - a) $E = \{HHH, TTT\}$
 - b) $E = \{HHT, HTH, THH\}$
 - c) $E = \{HHH, HHT, HTH, HTT\}$
 - d) $E = \{HHT, HTH, HTT, THH, THT, TTH, TTT\}$
2. A box contains 4 marbles: 2 red, 1 green, and 1 blue.
 - a) Consider an experiment that consists of taking 1 marble from the box, putting it back and drawing a second marble from the box (recording both choices in order). Describe the sample space for this experiment (your sample space should be constructed so that all the outcomes are equally likely).
 - b) Suppose you didn't put the first marble back before you drew the second marble. Describe the sample space in this context (again, your sample space should be constructed so that all the outcomes are equally likely).
3. Suppose you perform an experiment where there are eight possible outcomes. Assuming that every subset of outcomes constitutes an event, how many distinct events are there?

Hint: You may want to try this problem in the situation where there are two, three and/or four outcomes, and look for a pattern.

4. Suppose you roll a fair die repeatedly until a 4 turns up. You record the number of rolls it takes to roll a 4. Describe a probability space for this experiment. Verify that you have constructed a probability space.
5. Verify the following set identities, called **De Morgan's Laws**, using Venn diagrams or an argument written in English:
 - a) $(A \cap B)^C = A^C \cup B^C$
 - b) $(A \cup B)^C = A^C \cap B^C$
6. Suppose a point (x, y) is picked at random (with the uniform distribution) from the triangle in the xy -plane with vertices at $(0, 0)$, $(4, 0)$, and $(4, 4)$.
 - a) What is the probability that $x \geq 2$?
 - b) What is the probability that $x < y^2$?
7. Prove that there is no such thing as a uniform distribution on $\mathbb{N} = \{1, 2, 3, \dots\}$.
Hint: Prove this by contradiction: suppose there is a uniform distribution on \mathbb{N} . This means that $P(m) = P(n)$ for every $m, n \in \mathbb{N}$. There are two possibilities: either $P(1) = 0$ or $P(1) > 0$. Explain why both of these cases are impossible thinking about the value of $P(\Omega)$.
8. Let (Ω, \mathcal{A}, P) be a probability space. Prove (using the definition of probability space) that if E and F are events with $E \subseteq F$, then $P(E) \leq P(F)$.
Hint: Write F as the union of the two sets $E \cap F$ and $E' \cap F$.
9. Prove Bonferonni's Inequality, which says that given any two events E and F , $P(E \cap F) \geq P(E) + P(F) - 1$.
10. Suppose two fair dice are rolled and that the 36 possible outcomes are equally likely. Find the probability that the sum of the numbers on the two faces is even.
11. (AE) (The "(AE)" means this is, or closely resembles, an old actuarial exam problem.) The probability that a small fire in a kitchen destroys a microwave oven is 70%. The probability that a small fire in a kitchen destroys a refrigerator is 50%. If the probability that a small fire destroys both is 45%, find the probability that the fire destroys neither the microwave nor the refrigerator.
12. A survey reveals that 20% of the population is afraid of ghosts, 35% of the population is afraid of vampires, and 40% of the population is afraid of zombies. 15% of the population fears ghosts and vampires; 12% of the population fears ghosts and zombies, and 20% of the population fears vampires and zombies. If 8% of the population fears ghosts, vampires and zombies, what

percent of the population isn't afraid of any of the three mythical creatures discussed in the survey?

13. Suppose a point is picked uniformly from the square whose vertices are $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$. Let E be the event that the selected point is in the triangle bounded by the lines $y = 0$, $x = 1$ and $x = y$, and let F be the event that it is in the rectangle with vertices $(0, 0)$, $(1, 0)$, $(1, \frac{1}{2})$, and $(0, \frac{1}{2})$.
 - a) Compute $P(E)$.
 - b) Compute $P(F)$.
 - c) Compute $P(E \cup F)$.
 - d) Compute $P(E \cap F)$.
 - e) Compute $P(E|F)$.
 - f) Compute $P(F|E)$.
 - g) Are E and F independent? Why or why not? (You need an algebraic proof.)
14.
 - a) (AE) Suppose events A and B are such that $P(A) = \frac{2}{5}$ and $P(B) = \frac{2}{5}$. If you also know $P(A \cup B) = \frac{1}{2}$, compute $P(A \cap B)$.
 - b) (AE) If $P(A) = .7$, $P(A \cap B^C) = .6$ and A and B are independent, what is $P(B)$?
15. A coin is tossed three times. Consider the following events:
 - A = flipping heads on the first toss
 - B = flipping tails on the second toss
 - C = flipping heads on the third toss
 - D = flipping the same side of the coin all three times
 - E = flipping heads exactly once in the three tosses
 - a) Which one or ones of the following pairs of these events are independent? A and B , A and D , A and E , D and E (No proof is required here, if you want to use the heuristic idea of independence.)
 - b) Which one or ones of the following triples of these events are independent? A, B and C ; A, B and D ; C, D and E (No proof is required here, if you want to use the heuristic idea of independence.)
16.
 - a) Suppose E and F are independent. Prove that E^C and F^C are independent.
 - b) Suppose E and F are independent. Prove that E and F^C are independent.

- c) If an event E is pairwise independent with itself, what must be true about E ? Prove your statement.
17. A point is chosen uniformly from the unit square $[0, 1] \times [0, 1]$. Find a positive number c so that the events $E = \{(x, y) : y + cx \leq 1\}$ and $F = \{(x, y) : y \leq 2x/3\}$ are independent.
Hint: There are two values of c which solve this problem; you need to find one or the other, not both.
18. Choose one of (a) or (b):
- a) Three players, Al, Bal, and Cal, take turns flipping a fair coin (Al goes first followed by Bal, then Cal, then Al again, then Bal, etc.). The first player to flip a head wins. What is the probability of each player winning?
- b) (This is a famous problem in probability called *The Triangle Problem*.) Suppose you take a stick of length 1 and break it into three pieces, choosing the break points uniformly and independently. What is the probability that the three pieces can be used to form a triangle?
Hint: In a triangle, the sum of the lengths of any two sides must be at least the length of the third side.
19. There are three boxes, labeled I, II and III. Box I contains 2 white balls and 2 black balls; box II contains 2 white balls and 1 black ball; and box III contains 1 white ball and 3 black balls.
- a) One ball is selected from each box (the draws are independent of one another). Calculate the probability of drawing all white balls.
- b) Suppose you have five slips of paper, two labeled "I", two labeled "II" and one labeled "III". One of these five slips is drawn uniformly and then a ball is drawn from the box indicated by the slip of paper chosen. Calculate the probability that the drawn ball is white.
20. An urn contains 3 red and 2 blue marbles. One marble is drawn from the jar and its color noted. That marble, along with 2 marbles of the opposite color. A second marble is drawn from the jar.
- a) What is the probability that the two marbles drawn are of the same color?
- b) What is the probability that the second marble drawn is red?
21. Suppose a student takes a multiple choice exam where each question has 5 possible answers, exactly one of which is correct. If the student knows the answer to the question, she selects the correct answer. Otherwise, she guesses

uniformly from the 5 possible answers. Assume that the student knows the answer to 70% of the questions.

- a) What is the probability that on any single given question, the student gets the correct answer?
 - b) What is the probability that the student knows the answer to a question, given that she got the question correct?
22. (AE) Suppose a factory has two machines A and B which make 64% and 36% of the total production, respectively. Of their output, machine A produces 2% defective items and machine B produces 5% defective items. Find the probability that a given defective part was produced by machine B .
23. (AE) The probability that a randomly chosen male has a blood circulation problem is .325. Males who have a circulation problem are twice as likely to be smokers as those who do not have a blood circulation problem. What is the conditional probability that a male has a blood circulation problem, given that he is a smoker?

10.2 Exercises from Chapter 2

24. Suppose X is a discrete r.v. with density function f given by

x	-3	-1	0	1	2	3	5	8
$f_X(x)$.1	.2	.15	.2	.1	.15	.05	.05

- a) Compute the probability that X is negative.
 - b) Compute the probability that X is nonpositive.
 - c) Compute the probability that X is even.
 - d) Compute $P(X \in [1, 8])$.
 - e) Compute $P(X = -3 | X \leq 0)$.
 - f) Compute $P(X \geq 3 | X > 0)$.
25. Choose two of (a), (b), (c):
- a) Suppose a box has 12 balls numbered 1 to 12. Two balls are selected from the box independently, with replacement. Let X denote the larger of the two numbers on the selected balls. Compute the density of X .

- b) Suppose you choose a zip code (i.e. a five-digit sequence of numbers) uniformly from all possible zip codes and let X be the number of nonzero digits in the zip code. Calculate the density function of X .
- c) Suppose you uniformly and independently choose three whole numbers from 0 to 9. Let X be the first digit of the number you get when you add these whole numbers together. Calculate the density function of X .
26. (AE) Among a group of 20000 people, 7200 are below age 40, 8200 are childless and 12300 are male. In the same group, there are 5400 males below age 40, 4700 childless persons below age 40 and 6000 childless males. Finally, there are 3100 childless males below age 40. How many people are females above 40 who have children?
27. A 7-person committee, consisting of 3 Democrats, 3 Republicans and 1 Independent, is to be chosen from a group of 20 Democrats, 15 Republicans and 10 Independents. How many different committees are possible?
28. A bus starts with 6 people and stops at 10 different stops. Assuming that each passenger is equally likely to depart at any stop, calculate the probability that the 6 people get off at 6 different stops.
29. My niece's iPhone has 100 songs on it, of which 10 are performed by Taylor Swift. If she sets her iPod to shuffle mode, which will play all 100 songs in a random order (without repeating any songs until they are all played once), what is the probability that the first Taylor Swift song my niece hears is the eighth song played?
30. A **domino** is a rectangular block divided into two equal subrectangles as below, where each subrectangle has a number on it:

x	y
-----	-----

(The numbers x and y might be the same or different.) Since dominos are symmetric, the domino (x, y) is the same as (y, x) . How many different domino blocks can be made if the x and y are to be chosen from n different numbers?

Hint: Count the dominos where $x = y$ separately from the dominos where $x \neq y$. Then add these two separate counts.

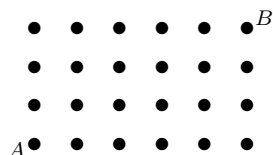
31. How many distinct arrangements of the letters in each of the following words are possible?

a) COFFEE

b) ASSESS

c) BOOKKEEPER

32. a) Consider the grid of points shown below. Suppose that starting at the point A you move from point to point, moving only one unit to the right or one unit up at a time, ending at the point B . How many different paths from A to B are possible?



- b) The above picture gives a 6×4 grid of dots. Answer the same question that was posed in part (a), if the grid is $m \times n$ (i.e. it has n horizontal rows, each containing m dots).
33. How many distinct, non-negative integer-valued vectors (x_1, x_2, \dots, x_5) satisfy $x_1 + x_2 + x_3 + x_4 + x_5 = 12$?
- Hint:* This has something to do with distinguishable arrangements, and might have something to do with Problem 32, depending on how you think about it.

In Problems 34-38, you are to give both a formula for the answer in terms of standard combinatorial notation, and a decimal approximation of your answer.

34. Suppose you deal a five-card hand from a standard deck of cards. Find the probability of being dealt each of the following hands:
- A royal flush (the A,K,Q,J and 10 of the same suit)
 - A flush (any five cards of the same suit)
 - Three-of-a-kind, but not a full house or four-of-a-kind
 - A straight (five cards in a sequence, regardless of suit)
- Note:* An ace may be the highest card (10-J-Q-K-A) or lowest card (A-2-3-4-5) in a straight, but a sequence like K-A-2-3-4 is not a straight because the ace is in the middle.
- A hand which contains no pair (nor three- nor four-of-a-kind)
35. In Texas Hold'Em, each player is dealt 2 cards from a standard deck.
- What is the probability that a Texas Hold'Em player is dealt a pair?
 - What is the probability that a Texas Hold'Em player's hand is a "Broadway" hand (i.e. both cards are 10 or higher)?

- c) What is the probability that a Texas Hold'Em player is dealt "suited connectors", meaning that the cards are of the same suit and adjoining rank (like (A-2) or (8-9) or (10-J) or (K-A))?
36. In the card game Bridge, each player is dealt 13 cards from a standard deck.
- a) A **Yarborough** is a (terrible) Bridge hand that contains no card higher than a 10 (i.e. no jacks through aces). Find the probability that a Bridge hand is a Yarborough.
 - b) A Bridge hand is said to have a **void** if there is at least one suit for which the hand has no cards in that suit. Find the probability that a Bridge hand has exactly one void.
37. In the card game Shanghai Rummy, two 54-card decks (each including the standard 52 cards and 2 jokers) are shuffled together. Then, each player is dealt a 12-card hand. What is the probability that a Shanghai Rummy hand contains at least one joker?
38. **Set** is a card game played with a deck of 81 different cards. Unlike normal playing cards, which have two attributes (a suit and a rank), each card in a Set deck has four attributes: a color (one of red, green, or purple), a shape (one of diamonds, ovals or waves), a number (1, 2 or 3), and a pattern (solid, striped, or open).
- a) If you choose five cards randomly from a Set deck, what is the probability that your hand is a color flush (meaning all five cards are of the same color)?
 - b) If you choose five cards randomly from a Set deck, what is the probability that your hand is a color and shape flush (meaning all five cards are of the same color and shape)?
 - c) If you choose five cards randomly from a Set deck, what is the probability that your hand is a color and shape and pattern flush (meaning all five cards are of the same color, shape and pattern)?
 - d) If you choose five cards randomly from a Set deck, what is the probability that your hand is a flush with respect to any two attributes?
 - e) If you choose five cards randomly from a Set deck, what is the probability that your hand is a flush with respect to at least one attribute?
39. A fair die is rolled 12 times (independently). Find the probability of rolling exactly 2 sixes, and the probability of rolling at most 2 sixes.
40. Experience shows that 20% of the people reserving tables at a certain restaurant never show up. If the restaurant has 50 tables and takes 52 reservations,

- what is the probability that it will be able to accomodate everyone who shows up?
41. A circular target of radius 1 is divided into four annular zones (an “annular” shape is like a ring) of outer radii $1/4$, $1/2$, $3/4$ and 1, respectively. Suppose 10 shots are fired at the target independently, and that each shot hits a random point in the target chosen uniformly.
- a) Find the probability that exactly four shots land in the region of radius $1/4$.
 - b) What is the probability that at most three shots land in the zone bounded on the inside by the circle of radius $1/2$ and on the outside by the circle of radius $3/4$?
 - c) If exactly 5 shots land inside the circle of radius $1/2$, find the probability that at least one shot lands inside the circle of radius $1/4$.
42. (AE) You own a business that gets bolts from two bolt manufacturers: A and B (you get 70% of your bolts from A and 30% from B). Suppose that 5% of all bolts from manufacturer A are defective, and that 20% of all bolts from manufacturer B are defective. You get a shipment of 12 bolts from one of the two manufacturers. If exactly 3 of the 12 bolts are defective, what is the probability that the shipment came from manufacturer B?
43. There are 40 gumballs in a bag, of which 20 are red, 10 are orange, 8 are green, and 2 are purple.
- a) If you draw 10 gumballs from the bag without replacement, what is the probability that you draw 5 red, 3 orange, and 2 purple gumballs?
 - b) If you draw 7 gumballs from the bag without replacement, what is the probability that you draw exactly 4 green gumballs?
 - c) If you draw 7 gumballs from the bag with replacement, what is the probability that you draw exactly 4 green gumballs?
 - d) If you draw 6 gumballs from the bag without replacement, what is the probability you draw at least 5 orange gumballs?
 - e) If you draw 10 gumballs from the bag with replacement, what is the probability that you draw 3 orange gumballs?
44. Continuing with the same bag of gumballs as in the previous problem:
- a) If you draw 15 gumballs from the bag without replacement and take a bite out of them, then put them back in the bag, and if you subsequently draw 5 gumballs from the bag with replacement, what is the probability that you drew 3 gumballs that you bit?

- b) Suppose you draw gumballs from the bag repeatedly, with replacement. What is the probability that the first time you draw a purple gumball is on the 9th draw?
 - c) Suppose you draw gumballs from the bag repeatedly, with replacement. What is the probability that the fifth time you draw a red gumball is on the 14th draw?
 - d) Suppose you draw gumballs from the bag two at a time, putting each group back after you draw it. What is the probability that the first time you draw 2 red gumballs (on a single draw) is the 4th time you draw 2 gumballs from the bag?
 - e) Divide the 40 gumballs randomly into four disjoint groups of 10. What is the probability that the first and second groups have the same number of green gumballs?
45. Suppose a box has 6 red balls and 4 black balls in it. A random sample of size n is selected; let X denote the number of red balls in the sample.
- a) Calculate the density function of X , if the sampling is without replacement.
 - b) Calculate the density function of X , if the sampling is with replacement.
46. Suppose $X \sim \text{Geom}(.8)$. Compute the following:
- a) $P(X > 3)$
 - b) $P(4 \leq X \leq 7 \text{ or } X > 9)$
 - c) $P(X \leq 2 \mid X \leq 3)$
 - d) $P(X \geq 85 \mid X \geq 80)$

10.3 Exercises from Chapter 3

47. Suppose you choose a real number X from the interval $[2, 10]$ with a density function of the form $f_X(x) = Cx$, where C is some constant.
- a) What is the value of C ?
 - b) Compute $P(X > 5)$.
 - c) Compute $P(X \leq 7)$.

48. Let X be a r.v. whose distribution function is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x}{4} & \text{if } 0 \leq x < 1 \\ \frac{x}{2} & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

Compute each quantity:

- | | |
|--|--|
| a) $P\left(\frac{1}{4} \leq X \leq \frac{5}{4}\right)$ | d) $P\left(1 \leq X \leq \frac{7}{4}\right)$ |
| b) $P\left(\frac{1}{4} \leq X \leq 1\right)$ | e) $P(1 < X < 2)$. |
| c) $P\left(\frac{1}{4} \leq X < 1\right)$ | f) $P(X \text{ is an integer})$ |

49. Suppose X is a r.v. whose distribution function is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{10} & \text{if } 1 \leq x < 3 \\ \frac{x}{10} & \text{if } 3 \leq x < 4 \\ K - \frac{2}{x} & \text{if } x \geq 4 \end{cases}$$

where K is a constant. Compute each quantity:

- | | |
|-------------------|----------------------------|
| a) K | e) $P(X > 1)$ |
| b) $P(X = 3)$ | f) $P(X > 4 X \geq 4)$ |
| c) $P(2 < X < 3)$ | g) $P(X < 3.5 X \leq 4)$ |
| d) $P(3 < X < 4)$ | h) $P(X > 2 X > 3)$ |

50. Choose (a) or (b):

- Let p be a point chosen uniformly from the interior of a circular disk of radius r . Let X denote the distance from p to the center of the disc; compute the distribution and density functions of X .
- Let a point be chosen uniformly from the interior of a triangle having a base of length l and a height h (measured perpendicular to the base). Let X be defined as the distance from the point chosen to the base of the triangle (this means the length of a perpendicular drawn from the point to the base); compute the distribution and density functions of X .

51. Let X be a discrete r.v. with density function f_X defined as follows:

$$f_X(-1) = \frac{1}{5}, \quad f_X(0) = \frac{1}{5}, \quad f_X(1) = \frac{2}{5}, \quad f_X(2) = \frac{1}{5}.$$

- a) Calculate a density function of the r.v. $Y = 2X + 1$.
- b) Calculate a density function of $Z = X^2$.
52. A point is chosen uniformly from the interval $(-10, 10)$. Let X be the r.v. defined so that X denotes the coordinate of the chosen point if the point is in $[-5, 5]$, $X = -5$ if the point is in the interval $(-10, -5)$, and $X = 5$ if the point is in the interval $(5, 10)$. Compute the distribution function of X . Does X have a density function? Why or why not?
53. Let X be a continuous, real-valued r.v. with distribution function F_X and density function f_X .
- a) Compute (in terms of F_X) the distribution function of $Y = e^X$.
- b) Compute (in terms of f_X) the density function of $Y = e^X$.
54. Let X be a continuous real-valued r.v. with density function f_X . Compute (in terms of f_X) the density function of $Y = |X|$.
55. Suppose a point (X, Y) is chosen uniformly from the rectangle whose vertices are $(1, 0)$, $(5, 0)$, $(1, 2)$ and $(5, 2)$. Compute the density function of $Z = XY$.
56. a) The number of bad checks that a bank receives during a 5-hour business day is a Poisson r.v. with $\lambda = 2$. What is the probability that the bank will receive no more than 2 bad checks in its business day?
- b) The mileage (in thousands of miles) that car owners get with a certain kind of radial tire is a r.v. whose distribution is exponential with parameter $\frac{1}{40}$. Compute the probability that one of these tires will last at least 20,000 miles.
57. (AE) The loss due to a fire in a commercial building is modeled by a continuous r.v. X with density function $f(x) = k(20 - x)$ for $0 < x < 20$ ($f(x) = 0$ otherwise). Given that a fire loss exceeds 8, what is the probability that it exceeds 16?
58. (AE) You are given the following information about N , the annual number of claims for a randomly selected insured person:

$$P(N = 0) = \frac{1}{2}; \quad P(N = 1) = \frac{3}{8}; \quad P(N = 2) = \frac{1}{8}.$$

Let S denote the total annual claim amount for an insured. When $N = 1$, S is exponentially distributed with parameter $\frac{1}{6}$. When $N > 1$, S is exponentially distributed with parameter $\frac{1}{10}$. Compute $P(4 < S < 8)$.

Hint: Use the Law of Total Probability.

59. Suppose that events occur according to a Poisson process with hourly rate $\lambda = 3$.
- Let p be the probability that no events occur between 8 AM and 10 AM.
 - Compute p , using the density function of an appropriate discrete r.v.
 - Compute p , using the density function of an appropriate continuous r.v.
 - Suppose you are given that v events occur between times 0 and t . Let $s < t$; compute the probability that exactly x of the v events occur between times 0 and s .
 - Suppose you are given that v events occur between times 0 and t . Let $s < t$. If X is the number of events occurring between times 0 and s , what kind of r.v. is X ? Include its parameters.
Hint: You computed the density function of X in part (b). Simplify this density function and identify it as the density function of a common r.v.
 - Suppose eight events occur between 8 AM and noon. What is the probability that (exactly) three of those events occurred after 11 AM?
60. As in the preceding problem, suppose that events occur according to a Poisson process with hourly rate $\lambda = 3$.
- What is the conditional probability that at least one event takes place between 8 AM and noon, given that no events take place between 8 AM and 10 AM?
 - What is the probability that exactly one event occurs between 8 and 9 AM and exactly one event occurs between 2 and 4 PM?
 - What is the probability that exactly one event occurs between 8 and 10 AM and exactly one event occurs between 9 and 11 AM?
61. Suppose X is exponential with parameter λ , where λ is such that $P(X \geq .02) = .35$. Find the number t such that $P(X \geq t) = .85$.
62. (AE) Suppose the number of claims filed by an insurance policyholder is a Poisson r.v. If the filing of (exactly) one claim is four times as likely as the filing of (exactly) two claims, find the probability the policyholder files exactly five claims.
63. Choose (a) or (b):
- Let X have an exponential density with parameter λ . Compute the density of $Y = cX$, where $c > 0$ is a positive constant.
 - Let X have the Cauchy density. Compute the density of $Y = a + bX$, where a and b are constants such that $b > 0$.

64. a) Evaluate $\Gamma(7)$.
 b) Evaluate $\Gamma(3.5)$.
 c) Simplify $\frac{\Gamma(3.2)}{\Gamma(5.2)}$.
 d) A useful and amazing fact to know about the gamma function is the following:

$$\Gamma(r)\Gamma(1-r) = \frac{\pi}{\sin(\pi r)}.$$

Use this fact to evaluate $\Gamma(1/3)\Gamma(2/3)$.

- e) Evaluate $\Gamma(7/6)\Gamma(5/6)$.
65. Suppose Z has the standard normal distribution. Compute decimal approximations to the following probabilities (trust me, there are no typos in these inequalities):
- | | |
|------------------------|--------------------------------------|
| a) $P(Z < 1.33)$ | e) $P(-1.90 \geq Z \geq .44)$ |
| b) $P(Z < -.425)$ | f) $P(Z > -.2)$ |
| c) $P(Z \geq .79)$ | g) $P(-.63 \leq Z < .3)$ |
| d) $P(.55 < Z < 1.22)$ | h) $P(Z < -1.3 \text{ or } Z > .58)$ |

66. Suppose X is normal with parameters $\mu = 20$ and $\sigma^2 = 100$. Compute decimal approximations to the following probabilities:

- | | |
|-------------------------------|---|
| a) $P(X \geq 17)$ | e) $P(X \geq 21 \mid X < 24)$ |
| b) $P(X < 24.5)$ | f) $P(X < 15 \text{ or } X \geq 24)$ |
| c) $P(X = 18)$ | g) $P(Y \geq 54), \text{ assuming } Y = 3X + 9$ |
| d) $P(X < 17 \mid X \leq 23)$ | |

67. Suppose X is normal $n(\mu, \sigma^2)$ and let $c > 0$. Compute, in terms of Φ , μ and σ , $P(|X - \mu| < c)$.

68. Let f be the density function of the normal r.v. with parameters μ and σ^2 .

- a) Show that f has its maximum when $x = \mu$.
 b) Show that the x -coordinates of the inflection points of f are $x = \mu \pm \sigma$.

69. Suppose that during periods of transcendental meditation the reduction of a person's oxygen consumption is a normal $n(37.6 \text{ cc/min}, 4.6^2 \text{ cc/min})$.

- a) Calculate (a decimal approximation to) the probability that during a period of transcendental meditation a person's oxygen consumption will be reduced by at least 42.5 cc/min.

- b) Calculate (a decimal approximation to) the probability that during a period of transcendental meditation a person's oxygen consumption will be reduced by anywhere from 30 to 40 cc/min.
70. A study shows that an experimental drug causes patient's blood pressure to lower by an amount that is normally distributed with parameters $\mu = 32$ mmHg and $\sigma^2 = 60$ mmHg. What is the probability that by taking this drug, a patient's blood pressure will be lowered by at least 25 mmHg but by no more than 35 mmHg? (Give both an answer in terms of Φ , and a decimal approximation.)
71. Use Stirling's formula to show that for large n , $\binom{2n}{n} \approx \frac{4^n}{\sqrt{\pi n}}$.

Remark: We will use this fact in MATH 416.

10.4 Exercises from Chapter 4

72. Suppose X and Y are discrete, integer-valued r.v.s with joint density function

$$f_{X,Y}(x, y) = \begin{cases} \frac{2}{9} \frac{2^x}{3^{x+y}} & x \geq 0, y \geq 0 \\ 0 & x < 0 \text{ or } y < 0 \end{cases}$$

- Verify that this $f_{X,Y}$ is in fact a density function.
 - Compute the probability that $X = 3$ and $Y = 4$.
Note: This is one question, asking for the probability that ($X = 3$ and $Y = 4$).
 - Compute the probability that $X = 2$.
 - Calculate a density function of the marginal Y .
 - Based on the computation you did in part (d), how would you describe Y as a common r.v.? (Include any appropriate parameters.)
73. Suppose you have two dice numbered 1 to 6 that you can load however you want (i.e. you can assign whatever probabilities you want to each number on each die). Is it possible to load the dice in such a manner that makes every sum from 2 to 12 equally likely when the dice are rolled independently? If so, explain how. If not, explain why not.

Hint: It is useful to look at the probabilities of rolling 1 and 6 with each die.

74. Let X and Y be r.v.s having joint density function given by the following table:

$Y \backslash X$	-1	0	2	6
-2	$\frac{1}{27}$	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{1}{9}$
1	$\frac{1}{9}$	0	$\frac{1}{9}$	$\frac{2}{9}$
3	0	$\frac{2}{27}$	$\frac{1}{9}$	$\frac{2}{27}$

- Compute the probability that X is even.
 - Compute the probability that XY is odd.
 - Compute the probability that $X > 0$ and $Y \geq 0$.
 - Compute the density function of X .
 - Compute the density function of Y .
75. Suppose X and Y are discrete r.v.s, **each taking values on the nonnegative integers**, with joint density function $f_{X,Y}$. For each given probability, write an expression, involving one or more sums, which gives the probability. As an example, if asked to compute $P(0 \leq X \leq 5, 2 \leq Y \leq 4)$, one possible correct answer is

$$P(0 \leq X \leq 5, 2 \leq Y \leq 4) = \sum_{x=0}^5 \sum_{y=2}^4 f_{X,Y}(x, y).$$

- $P(5 \leq X < 10, 0 < Y < 4)$
 - $P(X = 6, 9 \leq Y)$ (here (and always in this type of statement) the comma means “and”)
 - $P(X = 5 \text{ or } Y \geq 4)$
 - $P(X = 1)$
 - $P(3 \leq X, 12 \leq Y \leq 20)$
 - $P(X + Y = 11)$
 - $P(X - Y = 9)$
76. Same directions as the previous question:
- $P(0 \leq X \leq Y \leq 10)$
 - $P(0 \leq X \leq Y)$
 - $P(0 \leq Y \leq -X)$
 - $P(0 \leq X \leq Y^2)$

- e) $P(X + Y \leq 15)$
 f) $P(X + Y = z)$, where z is a constant
 g) $P(Y - X = z)$, where z is a nonnegative constant
 h) $P(X \geq 0)$
 i) $P(X \leq 20, Y \leq 20, X + Y \geq 20)$
77. Suppose X and Y are independent r.v.s, each being uniform on the discrete set $\{1, 2, \dots, N\}$. Compute the density of $X + Y$.
78. Let X and Y be independent r.v.s, where $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$. Prove that $X + Y$ is Poisson; what is its parameter? (The way you do this for now is to explicitly compute the density function of $X + Y$.) **The fact you are proving in this problem should be memorized** (and will be generalized later).
79. Let X and Y be independent r.v.s, where $X \sim \text{Geom}(p)$ and $Y \sim \text{Geom}(q)$ (do not assume any relationship between p and q in this problem).
- a) Compute $P(X = Y)$. b) Compute $P(X \geq Y)$.
80. Let $X \sim \text{Unif}(\{0, 1\})$ and $Y \sim \text{Unif}(\{0, 1\})$. Characterize all possible joint distributions of X and Y . For each of these joint distributions, compute the density of $X + Y$.
81. Suppose X and Y are discrete, integer-valued r.v.s with joint density function
- $$f_{X,Y}(x, y) = \begin{cases} \frac{2}{9} \frac{2^x}{3^{x+y}} & x \geq 0, y \geq 0 \\ 0 & \text{else} \end{cases}$$
- a) Compute the probability that $X + Y = 8$.
Hint: I want an answer with no “ Σ ”s in it. To evaluate your sum, you will need the following formula for a finite geometric sum given on the pink sheet.
- b) Compute the probability that $X + Y \geq 12$ (again, no “ Σ ”s in your answer are allowed).
82. Suppose X and Y have the density function of the previous problem (# 81). Compute the density of $Z = X - Y$.
83. There are 40 gumballs in a bag, of which 20 are red, 10 are orange, 8 are green, and 2 are purple.

- a) Suppose you randomly draw 15 gumballs from the bag, one at a time, with replacement. What is the probability you draw 5 red, 5 orange, and 5 green gumballs?
- b) Suppose you randomly draw 15 gumballs from the bag simultaneously. What is the probability you draw 5 red, 5 orange, and 5 green gumballs?

10.5 Exercises from Chapter 5

84. Suppose X and Y are continuous r.v.s such that $X \geq 0$ and $Y \geq 0$, with joint density function $f_{X,Y}$. For each given probability, write an expression involving integrals which gives the probability. As an example, if asked to compute $P(0 \leq X \leq 5, 2 \leq Y \leq 4)$, one possible correct answer is

$$P(0 \leq X \leq 5, 2 \leq Y \leq 4) = \int_0^5 \int_2^4 f_{X,Y}(x, y) dy dx.$$

- | | |
|---------------------------------|---------------------|
| a) $P(3 \leq X < 8, 0 < Y < 5)$ | e) $P(X \leq Y)$ |
| b) $P(X \geq 4)$ | f) $P(Y/X < 5)$ |
| c) $P(X + Y \leq 8)$ | g) $P(X - 2Y > 5)$ |
| d) $P(\min(X, Y) \leq 6)$ | h) $P(Y \leq 2X^2)$ |
85. Repeat Problem 84, but under the extra assumptions that X and Y take values only in the square whose vertices are $(0, 0)$, $(10, 0)$, $(0, 10)$ and $(10, 10)$.
86. Suppose X and Y are continuous r.v.s such that $0 < Y < X$, with joint density function $f_{X,Y}$. For each given probability, write an expression involving integrals which gives the probability.

- | | |
|---|-------------------------|
| a) $P(5 \leq X \leq 8, 3 \leq Y \leq 10)$ | e) $P(X \geq 14)$ |
| b) $P(3 \leq X \leq 10, 5 \leq Y \leq 8)$ | f) $P(Y \leq 2)$ |
| c) $P(X + Y \leq 8)$ | g) $P(X < 8, Y \leq 6)$ |
| d) $P(Y \geq \frac{1}{4}X)$ | h) $P(X - Y > 7)$ |

87. Suppose X and Y are two real-valued r.v.s with joint density function

$$f_{X,Y}(x, y) = \begin{cases} C(x^2 + \frac{xy}{2}) & \text{if } 0 < x < 1, 0 < y < 2 \\ 0 & \text{else} \end{cases}$$

where C is some constant. Compute each quantity:

- 221

92. Suppose X and Y are real-valued r.v.s with joint density

$$f_{X,Y}(x, y) = \begin{cases} \lambda^2 e^{-\lambda y} & 0 \leq x \leq y \\ 0 & \text{else} \end{cases}.$$

- a) Compute the marginal densities of X and Y .
 - b) Compute the probability that $Y \leq 4$.
93. Let X and Y denote the coordinates of a point chosen uniformly from the unit square. Let $Z_1 = X^2$, let $Z_2 = Y^2$ and let $Z_3 = X + Y$.
- a) Are Z_1 and Z_2 independent? Why or why not? (Give a heuristic argument only.)
 - b) Are Z_1 and Z_3 independent? Why or why not? (Give a heuristic argument only.)
94. Suppose X and Y are continuous r.v.s with joint density

$$f_{X,Y}(x, y) = \begin{cases} x e^{-x(y+1)} & \text{if } x > 0, y > 0 \\ 0 & \text{else.} \end{cases}$$

Compute the conditional density of X given Y .

95. Suppose X and Y are discrete r.v.s, taking values in the integers, whose joint density is

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{x!y!} \lambda^x e^{-\lambda-x-1} (x+1)^y & \text{if } 0 \leq x, 0 \leq y \\ 0 & \text{else.} \end{cases}$$

Compute the conditional density of Y given $X = 3$.

96. (AE) An insurance company supposes that each person has an accident parameter a and that the yearly number of accidents of someone who has accident parameter a is a Poisson r.v. X with parameter a . The company also supposes that the parameter of a newly insured person is itself a $\Gamma(r, \lambda)$ r.v. If a newly insured person has n accidents in his first year,
- a) Compute the conditional density of his accident parameter.
 - b) Identify the conditional density you found in part (a) as the density of a common r.v. (including appropriate parameters).
97. Let $Y \sim \text{Exp}(\lambda)$, where λ is itself a r.v. $\Lambda \sim \Gamma(r, \beta)$.
- a) Compute a density of Y .
 - b) Compute the conditional density of Λ given $Y = y$.

98. (AE) The distribution of Y , given X , is uniform on $[0, X]$. The marginal density of X is $f_X(x) = 2x$ for $0 < x < 1$ ($f_X(x) = 0$ otherwise). Find the conditional density of X given $Y = y$ (where this conditional density is positive).
99. Compute the conditional density $f_{Y|X}$, for the joint density given in Problem 92.
100. (AE) An auto insurance policy will pay for damage to both the policyholder's car and the other driver's car in the event that the policyholder is responsible for an accident. Assume that the size X of the payment for damage to the policyholder's car is uniform on $(0, 1)$, and that given $X = x$, the size Y of the payment to the other driver's car is uniform on $(x, x + 1)$. If the policyholder is responsible for an accident, what is the probability that the payment for damage to the other driver's car is greater than $\frac{1}{2}$?
101. (AE) Let X and Y be continuous r.v.s with joint density function

$$f(x, y) = \begin{cases} 24xy & \text{if } 0 < x < 1 \text{ and } 0 < y < 1 - x \\ 0 & \text{else} \end{cases}$$

Calculate $P[Y < X \mid X = \frac{1}{3}]$.

102. Suppose X and Y are discrete r.v.s whose joint density is given in the chart in Problem 74.
- a) Calculate $P(X < 4 \mid Y = 1)$.
- b) Calculate $P(Y < 3 \mid X = 6)$.
103. Suppose (X, Y) have joint density

$$f_{X,Y}(x, y) = \begin{cases} 4xy & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{else} \end{cases}$$

Compute the density of $W = X + Y$.

Hint: The computation requires separate cases, depending on whether $W \geq 1$ or $W < 1$.

104. (AE) A company offers earthquake insurance. Annual premiums are modeled by an exponential random variable with parameter 1. Annual claims are modeled by an exponential random variable with parameter 2. Assume that the annual premiums and claims are independent; let X denote the ratio of claims to premiums. What is the density function of X ?
105. If $X \sim \Gamma(r, \lambda)$, what is the density of $Y = \sqrt{X}$?

106. (AE) The time T that a computer is not working is a random variable whose cumulative distribution function is $F(t) = 1 - \frac{1}{4}t^{-2}$ for $t > 2$. The resulting cost X to the business as a result of the computer malfunctioning is $X = T^2$. Find the density function of X (when $X > 4$).
107. Let X and Y be independent standard normal r.v.s. What common r.v. is Y/X ?
Hint: First, compute the joint density of X and Y/X .
108. Let X and Y be continuous r.v.s with some unknown joint density function f . Compute (in terms of f) the joint density of X and $Z = X + Y$.
109. Let X and Y be continuous r.v.s with some unknown joint density function f . Compute the joint density (in terms of f) of W and Z , where $W = Y/X$ and $Z = X + Y$.
110. Let X and Y be independent Poisson r.v.s, with respective parameters λ and μ . Let $Z = X + Y$.
- Compute the joint density of X and Z .
Hint: In terms of X and Z , the joint density of X and Z is $f_{X,Z}(x, z) = P(X = x, Z = z)$. Back-substitute to see what this is in terms of X and Y .
 - Compute the conditional density of X given Z .
Hint: You should know what the density of Z is without computing its marginal again (since you studied this situation in a previous homework problem).
111. Suppose X and Y are continuous, real-valued r.v.s having some unknown joint density function $f_{X,Y}$. Let $W = a + bX$ and $Z = c + dY$ where a, b, c and d are constants, such that $b > 0$ and $d > 0$.
- Calculate a joint density function of W and Z .
 - Prove or disprove: if $X \perp Y$, then $W \perp Z$.
112. Suppose X_1, \dots, X_d are independent r.v.s.
- Let $MIN = \min(X_1, \dots, X_d)$. Derive a formula for F_{MIN} in terms of the F_{X_j} .
 - Prove that if X_1, \dots, X_d are independent exponential r.v.s with respective parameters $\lambda_1, \dots, \lambda_d$, then $\min(X_1, \dots, X_d)$ is exponential with parameter $\lambda_1 + \dots + \lambda_d$.
 - Let $MAX = \max(X_1, \dots, X_d)$ be the maximum of the X_j s. Derive a formula for F_{MAX} in terms of the F_{X_j} .

- d) (AE) A company decides to accept the highest of five sealed bids on a property. The sealed bids are regarded as five independent r.v.s, each with common cumulative distribution function

$$F(x) = \frac{(x-3)^2}{4} \text{ for } 3 \leq x \leq 5.$$

Find the density function of the accepted bid.

Note: The results of parts (a)-(c) of this problem are good to memorize for the actuarial exam. The maximum and minimum of the r.v.s are part of what are called the **order statistics** of the X_j .

10.6 Exercises from Chapter 6

113. Compute the expected value of X in each of these cases:

- X is cts and has density function $f(x)$ defined by $f(x) = \frac{3}{4}(1-x^2)$ for $x \in (-1, 1)$ and $f(x) = 0$ otherwise.
- X has cdf $F_X(x)$ defined by $F_X(x) = 1 - \frac{5}{x}$ if $x \geq 5$ and $F_X(x) = 0$ otherwise.
- X is the marginal of the joint distribution obtained when one selects a point (X, Y) uniformly from the triangle with vertices $(0, 0)$, $(4, 0)$ and $(0, 4)$.
- X takes values in $\{0, 1, 2, \dots\}$ and has survival function $H_X(x) = \frac{1}{x!}$.

NOTE: in all homework problems from this point forward, you may assume without proof that all r.v.s under consideration have finite expectation.

114. Choose three of (a),(b),(c),(d):

- Prove that the expected value of the uniform distribution on the interval $[a, b]$ is $\frac{a+b}{2}$.
- Prove that the expected value of an $Exp(\lambda)$ r.v. is $\frac{1}{\lambda}$.
- Prove that the expected value of a $n(\mu, \sigma^2)$ r.v. is μ .

Hint: A hard (but doable) way to do this is to directly compute

$$\int_{-\infty}^{\infty} x f_X(x) dx.$$

There is a much more clever approach, however.

- d) Verify that the expected value of a $Hyp(n, r, k)$ r.v. is $\frac{kr}{n}$.
Hint: Vandermonde's Identity may be useful.
115. a) Suppose $W \sim binomial(4, \frac{1}{3})$. Compute $E\left[\sin\left(\frac{\pi W}{2}\right)\right]$, simplifying your answer.
 b) Suppose $X \sim Pois(5)$. Calculate the mean of $(1 + X)^{-1}$.
 c) Let Y be the sine of an angle chosen uniformly from $(-\pi/2, \pi/3)$. Compute the expected value of Y .
116. If X has expected value 3 and Y has expected value -1 , what is the expected value of $3X - 5Y$?
117. Suppose that the density function f_X of X is:

$$f_X(x) = \begin{cases} a + bx^2 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases}.$$

If $EX = \frac{3}{5}$, find the values of a and b .

118. Suppose you play a carnival game that works like this: there are two bags, each with discs numbered 1 to 5 in them. You draw one disc uniformly from each bag. Whatever disc is the smaller number you draw, you win that amount of money (for example, if you draw a 2 and a 4, you would win 2).
- a) How much would you expect to win if you played this game 100 times?
 b) How much should the person running the game charge you if she expects to make a profit of .30 per game?
 c) Suppose that there were n discs in each bag, numbered 1 to n . How much would you now expect to win if you played the same game 100 times?

Hint: The summation formulas

$$\sum_{i=1}^n i = \frac{n(n+1)}{2} \text{ and/or } \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$$

may be useful.

119. (AE) A new plasma TV costs \$650. The lifetime of the TV is exponentially distributed with mean 4 years. Best Buy sells a warranty where they give a full refund to a buyer if the TV fails within the first two years, they give a half refund to a buyer if the TV fails during the third or fourth year, and they give no refund otherwise. How much should Best Buy expect to pay in refunds, if they sell 1000 plasma TVs?

120. (AE) Let T_1 be the time between a car accident and the reporting of a claim to an insurance company; let T_2 be the time between the reporting of this claim and the payment of this claim. Assume that (T_1, T_2) is uniform on the region of points (t_1, t_2) satisfying $0 < t_1 < 16$; $0 < t_2 < 16$; $0 < t_1 + t_2 < 20$. Find the expected amount of time between the accident and the payment of the claim.
121. A pond contains equal numbers of four different types of fish. You go fishing, and each time you cast, you catch one of the four types of fish (each type is equally likely). What is the expected number of casts it will take you to have caught at least one of all four types of fish?
122. a) Let $X \sim \text{Exp}(\lambda)$. Compute $E(X^2)$ directly (using the change of variables formula together with the Gamma integral formula) and use your answer to verify that the variance of X is $\frac{1}{\lambda^2}$.
 b) Prove that the variance of the uniform distribution on the interval (a, b) is $\frac{(b-a)^2}{12}$.
 c) Prove that the mean of a $\Gamma(r, \lambda)$ r.v. is $\frac{r}{\lambda}$ and that the variance of a $\Gamma(r, \lambda)$ r.v. is $\frac{r}{\lambda^2}$.
 d) Suppose X is a cts r.v. with density f given by $f(x) = cx^4$ for $0 < x < 2$ and $f(x) = 0$ otherwise. Calculate the variance of X .
123. Let X be a r.v. with finite expectation and finite variance. Prove:
 a) For any constant a , $\text{Var}(aX) = a^2 \text{Var}(X)$.
 b) For any constant b , $\text{Var}(X + b) = \text{Var}(X)$.
124. a) Suppose X and Y are two independent r.v.s such that $EX^4 = 2$, $EY^2 = 1$, $EX^2 = 1$ and $EY = 0$. Compute the variance of X^2Y .
 b) Let S and T be two independent r.v.s with finite expectation and finite variance. Let $W = 2S + 3T$; compute the mean and variance of W in terms of the means and variances of S and T .
125. Choose two of (a),(b),(c):
 a) (AE) An actuary has discovered that policyholders are six times as likely to file three claims as they are to file four claims. If the number of claims filed has a Poisson distribution, what is the variance of the number of claims filed?
 b) (AE) A company has two electric generators. The time until failure for each generator is exponential with mean 13. The company will begin using the second generator immediately after the first one fails. What is the variance of the total time the generators produce electricity?

- c) (AE) The profit for a new product is given by $Z = 5X - 4Y + 8$, where $X \perp Y$, $Var(X) = 3$ and $Var(Y) = 2$. What is the variance of the profit for the new product?

126. Let (X, Y) be a point chosen uniformly from the finite set of four points

$$\{(0, 1), (1, 0), (0, -1), (-1, 0)\}.$$

Prove that X and Y are uncorrelated, but not independent.

127. a) Suppose a box contains three balls numbered 1 to 3. Two balls are selected without replacement from the box. Let U be the number on the first ball selected, and let V be the number on the second ball selected. Compute $Cov(U, V)$ and $\rho(U, V)$.
b) Compute the covariance of X and Y , if they have joint density

$$f_{X,Y}(x, y) = \begin{cases} 2 & \text{if } x > 0, y > 0, \text{ and } x + y < 1 \\ 0 & \text{else} \end{cases}.$$

128. (AE) Let X and Y denote the price of two stocks at the end of a five-year period. Suppose X is uniform on $[0, 6]$ and that given $X = x$, Y is uniform on $[0, x]$. Determine $Cov(X, Y)$.
129. (AE) Let X denote the size of a surgical claim, and let Y denote the size of the associated hospital claim. An actuary is using a model in which $EX = 6$, $EX^2 = 47.4$, $EY = 3$, $EY^2 = 21.4$ and $Var(X + Y) = 13.5$. Let $C_1 = X + Y$ be the size of the combined claims before the application of a 20% surcharge on the hospital portion of the claim, and let C_2 denote the size of the combined claims after the surcharge. Calculate $Cov(C_1, C_2)$.
130. Prove any two of the following three statements:
- $Cov(X, Y) = Cov(Y, X)$
 - $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$
 - $Cov(aX, Y) = a Cov(X, Y)$

Note: These three statements generalize to the following important property of covariance called *bilinearity*:

$$Cov\left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j Cov(X_i, Y_j)$$

131. Suppose a and b are positive constants, and that c and d are constants (not necessarily positive). Show that $\rho(aX + c, bY + d) = \rho(X, Y)$, i.e. that the correlation between two quantities does not depend on the units used to measure those quantities.

132. a) Prove that if $Y = aX + b$ for constants a and b (with $a \neq 0$), then $\rho(X, Y) = \pm 1$. Under what conditions is $\rho(X, Y) = 1$ (as opposed to -1)?
- b) In this problem, we will prove that if $\rho(X, Y) = \pm 1$, then $Y = aX + b$ where a and b are constants. To start, define

$$\widehat{X} = \frac{1}{\sqrt{\text{Var}(X)}}(X - EX) \text{ and } \widehat{Y} = \frac{1}{\sqrt{\text{Var}(Y)}}(Y - EY).$$

- i. Compute $E[\widehat{X}]$, $E[\widehat{Y}]$, $E[\widehat{X}^2]$ and $E[\widehat{Y}^2]$.
 - ii. Prove that $\rho(X, Y) = \text{Cov}(\widehat{X}, \widehat{Y})$.
 - iii. Prove that $\text{Cov}(\widehat{X}, \widehat{Y}) = E[\widehat{X}\widehat{Y}]$.
 - iv. Use parts (i)-(iii) to prove that $E[(\widehat{Y} - \rho(X, Y)\widehat{X})^2] = 1 - \rho(X, Y)^2$.
 - v. Use part (iv) to prove that if $\rho(X, Y) = \pm 1$, then $Y = aX + b$ where a and b are constants.
133. Suppose X and Y are discrete r.v.s whose joint density is given in the chart in Problem 74.
- a) Calculate $E(Y | X)$.
 - b) Calculate $E(X^3 | Y = 1)$.
 - c) Calculate $\text{Var}(X | Y = 3)$.
134. Let (X, Y) be chosen uniformly from the triangle whose vertices are $(0, 0)$, $(2, 0)$ and $(1, 2)$. Compute the conditional expectation of Y given X .
135. (AE) A fair die is rolled repeatedly. Let X be the number of rolls needed to obtain a 5 and let Y be the number of rolls needed to obtain a 6. Calculate $E[X | Y = 2]$.
136. Let X and Y be independent, where X is $\Gamma(r, \lambda)$ and Y is $\Gamma(s, \lambda)$. Compute $E[X | X + Y]$.
- Hint:* First calculate the joint density of X and $X + Y$.
137. (AE) Let N_1 and N_2 represent the numbers of claims submitted to a life insurance company in January and February, respectively. The joint density function of N_1 and N_2 is

$$f_{N_1, N_2}(n_1, n_2) = \begin{cases} \frac{2}{3} \left(\frac{1}{3}\right)^{n_1} e^{-n_1-1} (1 - e^{-n_1-1})^{n_2} & \text{for } n_1, n_2 \in \{0, 1, 2, 3, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

Calculate the expected number of claims that will be submitted to the company in February if exactly 2 claims were submitted in January.

138. (AE) A driver and a passenger are in a car accident. Each of them independently has a probability .3 of being hospitalized. If they are hospitalized, the loss is uniform on $[0, 1]$. When two hospitalizations occur, the losses are independent. Calculate the expected number of people who are hospitalized, given that the total loss due to hospitalizations from the accident is less than 1.
139. The time it takes an insurance company to process a claim of size S is uniform on $[S, S + 1]$. If S is itself exponentially distributed with parameter $\frac{1}{2}$, what is the expected time to process a claim?
140. a) Prove that the two formulas given in the notes as definitions of conditional variance are the same.
b) Prove the Law of Total Variance, which says:

$$E[\text{Var}(X|Y)] + \text{Var}[E(X|Y)] = \text{Var}(X).$$

141. a) (AE) The number of workplace injuries, N , occurring in a factory on any given day is Poisson with mean λ . The parameter λ is itself a r.v. depending on the level of activity in the factory, and is assumed to be uniformly distributed on the interval $[0, 6]$. Find $\text{Var}(N)$.
b) (AE) The stock prices of two companies at the end of any given year are modeled with r.v.s X and Y whose joint density function is

$$f(x, y) = \begin{cases} 2x & \text{for } 0 < x < 1, x < y < x + 1 \\ 0 & \text{otherwise} \end{cases}.$$

What is the conditional variance of Y given $X = x$?

10.7 Exercises from Chapter 7

142. Let $X \sim \text{binomial}(n, p)$. Use the probability generating function of X to compute the expected value and variance of X .
143. a) Compute the moment generating function of a $\text{Pois}(\lambda)$ r.v.
b) Compute the moment generating function of a $\Gamma(r, \lambda)$ r.v.
144. (AE) Let X, Y and Z be i.i.d. r.v.s, each taking the value 0 with probability p and the value 1 with probability $(1 - p)$. Find the moment generating function of $W = XYZ$.

145. Let X be a continuous r.v. having the density $f_X(x) = \frac{1}{2}e^{-|x|}$ for all x . Compute the moment generating function of X .
146. a) Compute the first and second moments of X , if its moment generating function is $M_X(t) = \frac{1}{\sqrt{1-4t}}$ for $t < \frac{1}{4}$.
 b) Suppose X and Y are exponential r.v.s with respective means 3 and 7. If $X \perp Y$, what is the moment generating function of $4X + Y$?
 c) (AE) Assume that the number of claims related to traffic accidents on a certain road is a r.v. X whose moment generating function is $M_X(t) = (1 - 2500t)^{-4}$. Find the standard deviation of the claim size for this class of accidents.
147. Explain why each of the following functions *cannot* be the moment generating function of a real-valued r.v. X :
- a) $h(t) = \frac{e^{-t}}{2-t}$ for $t < 2$;
 b) $j(t) = \frac{1+t}{1-t}$ for $t < 1$;
 c) $k(t) = \exp(\frac{-t^2}{2})$ for $-\infty < t < \infty$.
148. (AE) Let X represent the number of customers arriving during the morning hours, and let Y be the number of customers arriving during the evening hours to a restaurant. Assuming that X and Y are both Poisson, and that the first moment of X is 8 less than the first moment of Y , and that the second moment of X is 60% of the second moment of Y , what is the variance of Y ?
149. (AE) The number N of babies born in a hospital during any one week is a r.v. satisfying $P(N = n) = \frac{1}{2^{n+1}}$, for $n \in \{0, 1, 2, \dots\}$. Suppose that the number of babies born in any one week is independent of the number of babies born in any other week. Determine the probability that exactly seventeen babies are born in a given four-week period.
150. In this problem we will derive the **Beta integral formula** (which we have used before to solve certain expected value and conditional expectation problems):

$$\int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

- a) Let X be $\Gamma(\alpha, \lambda)$ and let Y be $\Gamma(\beta, \lambda)$. Suppose that $X \perp Y$. Determine the density function of $Z = X + Y$ using moment generating functions.
 b) Given X and Y as above, compute the joint density function of X and $Z = X + Y$ by the transformation method of Chapter 5.
 c) Use your answer to part (b) to compute the marginal density of Z (write your answer as an integral with respect to x).

- d) Derive the Beta integral formula by equating the answers to part (a) and (c) of this problem, and solving the resulting equation for the Beta integral above.

Hint: in the integral you obtain from part (c), use the u -substitution $u = \frac{x}{z}$.

151. A **Beta random variable** with parameters $r_1 > 0$ and $r_2 > 0$ (denoted $B(r_1, r_2)$) is a continuous r.v. whose density is

$$f(x) = \begin{cases} \frac{\Gamma(r_1+r_2)}{\Gamma(r_1)\Gamma(r_2)} x^{r_1-1} (1-x)^{r_2-1} & \text{if } 0 < x < 1 \\ 0 & \text{else} \end{cases}.$$

- a) Prove that the function above is in fact a density function.
 b) Determine the expected value of a Beta $B(r_1, r_2)$ r.v.
152. (AE) Suppose X and Y are independent r.v.s which have the same moment generating function: $M_X(t) = M_Y(t) = e^{t^2}$. Determine the joint moment generating function of $W = X + Y$ and $Z = Y - X$.
153. Suppose Y is a discrete r.v. taking the values 0, 1, 4 and 10 with respective probabilities $\frac{3}{8}, \frac{1}{8}, \frac{1}{3}$ and $\frac{1}{6}$. Compute $M_Y(t)$.
154. Suppose X is a r.v. with $EX = \frac{1}{2}$ whose moment generating function is

$$M_X(t) = \frac{1}{7} + \frac{2}{7}e^t + Ce^{-t} + De^{2t},$$

where C and D are constants.

- a) Find C and D .
 b) Find a density function of X .
Hint: Look at the moment generating function you computed in the preceding problem, and use that to make an educated guess as to the density of X . (Uniqueness of MGFs can be used to show that your guess is correct.)
 c) Find $P(X \geq 0)$.
 d) Find the variance of X .
155. (AE) Let X and Y be i.i.d. r.v.s such that the moment generating function of $X + Y$ is
- $$M_{X+Y}(t) = .09e^{-2t} + .24e^{-t} + .34 + .24e^t + .09e^{2t}$$
- for all t . Calculate $P(X \leq 0)$.

156. Suppose (X, Y) have the following bivariate normal density:

$$f_{X,Y}(x, y) = C \exp \left[\frac{-1}{54} (x^2 + 4y^2 + 2xy + 2x + 8y + 4) \right]$$

Compute each quantity:

- | | |
|-------------|-----------------------------------|
| a) EX | e) $Cov(X, Y)$ |
| b) EY | f) $\rho(X, Y)$ |
| c) $Var(X)$ | g) C |
| d) $Var(Y)$ | h) The covariance matrix Σ |

157. Let X and Y have the density given in the preceding problem (# 156).

- Compute the conditional density of Y given $X = x$.
- Compute the conditional density of X given $Y = -1$.

158. Let X and Y have the density given in Problem 156.

- Compute the conditional variance of Y given $X = x$.
- Compute the density of $W = 8X + 5Y$.

10.8 Exercises from Chapter 8

159. Let X be a discrete r.v. whose density is

x	1	2	3
$f_X(x)$	$\frac{1}{18}$	$\frac{16}{18}$	$\frac{1}{18}$

Show that when $\delta = 1$, $P(|X - \mu| \geq \delta) = \frac{Var(X)}{\delta^2}$ (the point of this problem is to show that in general, the \leq sign in Chebyshev's inequality cannot be replaced by a $<$).

160. A bolt manufacturer knows that 5% of his production is defective. He gives a guarantee on his shipment of 10000 parts by promising that no more than a bolts are defective. Use Chebyshev's inequality to find the smallest number a can be, so that the manufacturer is assured of not paying a refund more than 1% of the time.

161. Let X be Poisson with mean λ . Use Chebyshev's inequality to verify that $P(X \leq \frac{\lambda}{2}) \leq \frac{4}{\lambda}$.
162. Let X be Poisson with mean λ . Use Chebyshev's inequality to verify that $P(X \geq 2\lambda) \leq \frac{1}{\lambda}$.
163. Suppose X is a r.v. with mean and variance both equal to 20. What can be said about $P(0 < X < 40)$? (In particular, what is the maximum or minimum value of this expression?)
164. Suppose X and Y are two real-valued r.v.s with

$$EX = 75, EY = 75, Var(X) = 10, Var(Y) = 12, Cov(X, Y) = -3.$$

What can be said about $P(|X - Y| \geq 15)$?

165. For each $\lambda > 0$, let X_λ be Poisson with parameter λ and let $Y_\lambda = \frac{X_\lambda - \lambda}{\sqrt{\lambda}}$.
- a) Show that for all t ,
- $$\lim_{\lambda \rightarrow \infty} M_{Y_\lambda}(t) = \exp(t^2/2).$$
- b) Fix $c > 0$ and use part (a) to estimate, for large λ , the value of $P(X_\lambda \leq c\lambda)$. Your answer should be in terms of Φ , the cumulative distribution function of the standard normal r.v.
- c) Compute $\lim_{\lambda \rightarrow \infty} P(X_\lambda \leq c\lambda)$.
Hint: There may be different answers depending on the value of c .
- d) Let Q be a Poisson r.v. with mean 5000. Use your answers to the previous parts of this question to estimate the following, in terms of Φ :
- $P(Q \leq 5100)$
 - $P(Q < 4920)$
 - $P(Q \geq 5050)$

166. Fix $\lambda > 0$ and for each $r > 0$ let X_r be $\Gamma(r, \lambda)$ and define $Y_r = \frac{X_r - (\frac{r}{\lambda})}{(\frac{\sqrt{r}}{\lambda})}$.

- a) Find the expected value and variance of Y_r .
- b) Show that for all t ,

$$\lim_{r \rightarrow \infty} M_{Y_r}(t) = \exp(t^2/2).$$

- c) Find $\lim_{r \rightarrow \infty} P(Y_r \leq x)$ in terms of Φ , the cumulative distribution function of the standard normal r.v.
- d) Find $\lim_{r \rightarrow \infty} P(X_r \leq r/\lambda)$.

e) Suppose Q is a gamma r.v. with parameters $r = 2000$ and $\lambda = 2$. Estimate the following probabilities in terms of Φ :

- i. $P(Q \leq 1800)$
- ii. $P(Q > 1850)$
- iii. $P(Q < 2150)$

167. Suppose X_1, X_2, \dots are i.i.d. r.v.s (taking only positive real values), each having finite mean μ . Show that with probability 1, the geometric averages of the X_j converge, where the *geometric average* of X_1, \dots, X_n is

$$G_n = \sqrt[n]{\prod_{j=1}^n X_j}.$$

Determine $\lim_{n \rightarrow \infty} G_n$.

Hint: Apply the SLLN to $\log G_n$ (here \log means natural logarithm).

168. The amount of liquid a student puts in their drink at the Rock is a r.v. with mean 350 mL and variance 1500 mL. Use the Central Limit Theorem to estimate the probability that a randomly selected group of 12 students put an average of 320 mL or more in their drinks. Give both the exact answer in terms of Φ and a decimal approximation to the answer.
169. 1000 fair dice are rolled independently. Use the Central Limit Theorem to estimate the probability that the sum of these 1000 rolls is at least 3450 and no greater than 3650. Give both the exact answer in terms of Φ and a decimal approximation to the answer.
170. You play a game where you lose \$1 with probability .7, you lose \$2 with probability .2, and win \$10 with probability .1. If you play this game 10000 times, what is the probability that you will be ahead (that is, you have won more money than you have lost)? (You are to approximate this answer using the Central Limit Theorem; give both the exact answer in terms of Φ and a decimal approximation to the answer.)
171. Let $X \sim \text{Pois}(40)$. Let $p = P(X \geq 48)$. Approximate p using the Central Limit Theorem, by approximating X as the sum of 40 i.i.d. r.v.s. Give both the exact answer in terms of Φ and a decimal approximation to the answer.
172. A tobacco company claims that the amount of nicotine in one of its cigarettes is a r.v. with mean 2.2 mg and standard deviation .8 mg. Use the Central Limit Theorem to estimate the probability that 100 randomly chosen cigarettes would have an average nicotine content of at most 2.09 mg. Give both the exact answer in terms of Φ and a decimal approximation to the answer.

173. (AE) In an analysis of healthcare data, ages are rounded to the nearest multiple of 5 years. The difference between the true age and the rounded age is assumed to be uniformly distributed on the interval from -2.5 years to 2.5 years. The healthcare data is based on a random sample of 80 people. What is the approximate probability (as estimated using the CLT) that the mean of the rounded ages is within 0.125 years of the mean of the true ages?
174. (AE) A charity receives 3100 contributions, each of which are assumed to be independent and identically distributed with mean 150 and standard deviation 40. Use the CLT to approximate the number b so that it is 90% likely that the total contributions to the charity are less than or equal to b .
175. (AE) The total claim amount for a property insurance policy follows a distribution that has density function

$$f(x) = \frac{1}{2000} e^{-x/2000} \quad \text{for } x > 0.$$

The premium for the policy is set at 250 over the expected total claim amount. If the insurance company sells 300 policies, what is the approximate probability (as estimated using the CLT) that the insurance company will have claims exceeding the premiums collected?

10.9 Exercises from Chapter 9

176. Suppose that a loss random variable is uniform on $[0, 1000]$. Determine the expected amount paid by the insurer in each of the following cases:
- a) The policy has a deductible of 300.
 - b) The policy has a coverage limit of 500.
 - c) The policy has a deductible of 100 and a coverage limit of 600.
177. Calculate the variance of the amount paid by the insurer in situation (a) of the previous problem.
178. (AE) Suppose that a loss random variable is uniform on $[0, 1000]$. A deductible of size d is applied before any insurance payment. If the expected payment of the insurer is 150, find d .
179. (AE) Suppose that a loss random variable is exponential with mean 10. If a deductible of size 5 is applied, find the expected payment of the insurer.

180. (AE) An insurance policy pays for a random loss X subject to a deductible C , where $0 < C < 1$. The loss amount is modeled as a continuous r.v. whose density is $f(x) = 4x^3$ on $[0, 1]$. If the probability that the insurance payment is less than .5 is .2401, what is C ?
181. Suppose the loss from an accident is a continuous r.v. with density $f(x) = \frac{24}{7}x^{-4}$ when $1 < x < 2$. Suppose that the insurance policy has a coverage limit of 1.5. What is the expected value of the loss to the insurance company? What is the standard deviation of the loss to the insurance company?
182. (AE) Suppose that the damage (in thousands of dollars) caused when a piece of equipment breaks is given by a continuous r.v. with density $f(x) = \frac{2}{x^3}$ when $1 < x$. Suppose that the piece of equipment breaks 25% of the time. If an insurance company agrees to cover 100% of the first \$3000 in damage and 50% of the next \$3000 in damage, what is the expected value of the amount the insurance company will have to pay?
183. The cumulative distribution function for health care costs experienced by a policyholder is

$$F(x) = \begin{cases} 1 - e^{-x/100} & \text{for } x > 0 \\ 0 & \text{else} \end{cases}$$

- The policy has a deductible of 20. An insurer reimburses the policyholder for 100% of health care costs between 20 and 120 (less the deductible); health care costs from 120 to 420 are reimbursed at 50%; health care costs above 420 are not reimbursed. Compute the cdf of the reimbursements.
184. (AE) The lifetime of a printer costing 100 is exponential with mean 2 years. The manufacturer agrees to pay a full refund to a buyer if the printer fails during the first year following its purchase, and a half refund if it fails during its second year. If the manufacturer sells 100 printers, how much should it expect to pay in refunds?
185. An insurance policy reimburses a loss up to a benefit limit of 15. The policyholder's loss follows a distribution with density function $f(x) = 2/x^3$ for $x > 1$.
- What is the probability that the benefit paid is less than 10?
 - What is the expected value of the benefit paid under this policy?

Appendix A

Tables

A.1 Charts of properties of common random variables

The next page has a chart listing relevant properties of the common discrete random variables.

The following page has a chart listing relevant properties of the common continuous random variables.

A.1. Charts of properties of common random variables

DISCRETE DISTRIBUTION X	DENSITY FUNCTION $f_X(x)$	$E(X)$	$\text{Var}(X)$	PGF $G_X(t)$ MGF $M_X(t)$
uniform on $\{1, \dots, n\}$	$f(x) = \frac{1}{n}$ for $x = 1, 2, \dots, n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$G_X(t) = \frac{t(t^{n-1}-1)}{n(t-1)}$ $M_X(t) = \frac{e^t(e^{nt}-1)}{n(e^t-1)}$
binomial(n, p)	$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, \dots, n$	np	$np(1-p)$	$G_X(t) = (1-p+pe^t)^n$ $M_X(t) = (1-p+pe^t)^n$
$Geom(p)$ $0 < p < 1$	$f(x) = p(1-p)^x$ for $x = 0, 1, 2, \dots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$G_X(t) = \frac{p}{1-(1-p)t}$ $M_X(t) = \frac{p}{1-(1-p)e^t}$
negative binomial $NB(r, p)$	$f(x) = \binom{r+x-1}{x} p^r (1-p)^x$ for $x = 0, 1, 2, \dots$	$r \frac{1-p}{p}$	$r \frac{1-p}{p^2}$	$G_X(t) = \left(\frac{p}{1-(1-p)t} \right)^r$ $M_X(t) = \left(\frac{p}{1-(1-p)e^t} \right)^r$
$Pois(\lambda)$	$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0, 1, 2, \dots$	λ	λ	$G_X(t) = e^{\lambda(t-1)}$ $M_X(t) = e^{\lambda(e^t-1)}$
hypergeometric $Hyp(n, r, k)$	$f(x) = \frac{\binom{r}{x} \binom{n-r}{k-x}}{\binom{n}{k}}$ for $x = 0, 1, \dots, k$	$\frac{kr}{n}$	$\frac{kr}{n} \binom{n-r}{n} \frac{n-k}{n-1}$	not given here
d -dimensional hypergeometric with parameters $n, (n_1, \dots, n_d), k$	$f(x_1, \dots, x_d) = \frac{\binom{n_1}{x_1} \binom{n_2}{x_2} \dots \binom{n_d}{x_d}}{\binom{n}{k}}$ for $x_1 + x_2 + \dots + x_d = k$	N/A	N/A	N/A
multinomial $n, (p_1, \dots, p_d)$	$f(x_1, \dots, x_d) = \frac{n!}{x_1! x_2! \dots x_d!} p_1^{x_1} p_2^{x_2} \dots p_d^{x_d}$ for $x_1 + x_2 + \dots + x_d = n$	N/A	N/A	N/A

CONTINUOUS DISTRIBUTION X	DENSITY FUNCTION $f_X(x)$ DISTRIBUTION FUNCTION $F_X(x)$	EXPECTED VALUE EX VARIANCE $Var(X)$ MGF $M_X(t)$
uniform on $[a, b]$	$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{else} \end{cases}$ $F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x > b \end{cases}$	$EX = \frac{a+b}{2}$ $Var(X) = \frac{(b-a)^2}{12}$ $M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$
exponential with parameter $\lambda > 0$	$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$ $F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$	$EX = \frac{1}{\lambda}$ $Var(X) = \frac{1}{\lambda^2}$ $M_X(t) = \frac{\lambda}{\lambda - t} \text{ for } t < \lambda$
Cauchy	$f(x) = \frac{1}{\pi(1+x^2)}$ $F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$	$EX = \infty$ $Var(X) \text{ DNE}$ $M_X(t) \text{ DNE}$
std. normal $n(0, 1)$	$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ $F(x) = \Phi(x)$	$EX = 0$ $Var(X) = 1$ $M_X(t) = e^{t^2/2}$
normal $n(\mu, \sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$	$EX = \mu$ $Var(X) = \sigma^2$ $M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$
gamma $\Gamma(r, \lambda)$	$f(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$ $F_X \text{ not given here}$	$EX = \frac{r}{\lambda}$ $Var(X) = \frac{r}{\lambda^2}$ $M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^r \text{ for } t < \lambda$
joint normal with mean vector μ ; covariance matrix Σ	$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} \sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right]$	$EX \text{ and } Var(X) \text{ DNE}$ $M_X(\mathbf{t}) = \exp\left(\mathbf{t} \cdot \mu + \frac{1}{2} \mathbf{t}^T \Sigma \mathbf{t}\right)$

A.2 Useful sum and integral formulas

Triangular Number Formula: For all $n \in \{1, 2, 3, \dots\}$,

$$1 + 2 + 3 + \dots + n = \sum_{j=0}^n j = \frac{n(n+1)}{2}.$$

Finite Geometric Series Formula: for all $r \in \mathbb{R}$,

$$\sum_{n=0}^N r^n = \frac{1 - r^{N+1}}{1 - r}.$$

Infinite Geometric Series Formulas: for all $r \in \mathbb{R}$ such that $|r| < 1$,

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1 - r} \qquad \sum_{n=N}^{\infty} r^n = \frac{r^N}{1 - r}.$$

Derivative of the Geometric Series Formula: for all $r \in \mathbb{R}$ such that $|r| < 1$,

$$\sum_{n=0}^{\infty} n r^n = \frac{r}{(1 - r)^2}.$$

Exponential Series Formula: for all $r \in \mathbb{R}$,

$$\sum_{n=0}^{\infty} \frac{r^n}{n!} = e^r.$$

Binomial Theorem: for all $n \in \mathbb{N}$, and all $x, y \in \mathbb{R}$,

$$\sum_{k=0}^n \binom{n}{k} x^k y^{n-k} = (x + y)^n.$$

Vandermonde Identity: for all $n, k, r \in \mathbb{N}$,

$$\sum_{x=0}^n \binom{r}{x} \binom{n-r}{k-x} = \binom{n}{k}.$$

Gamma Integral Formula: for all $r > 0, \lambda > 0$,

$$\int_0^{\infty} x^{r-1} e^{-\lambda x} dx = \frac{\Gamma(r)}{\lambda^r}.$$

Normal Integral Formula: for all $\mu \in \mathbb{R}$ and all $\sigma > 0$,

$$\int_{-\infty}^{\infty} \exp\left(\frac{-(x - \mu)^2}{2\sigma^2}\right) dx = \sigma\sqrt{2\pi}.$$

Beta Integral Formula: for all $r > 0, \lambda > 0$,

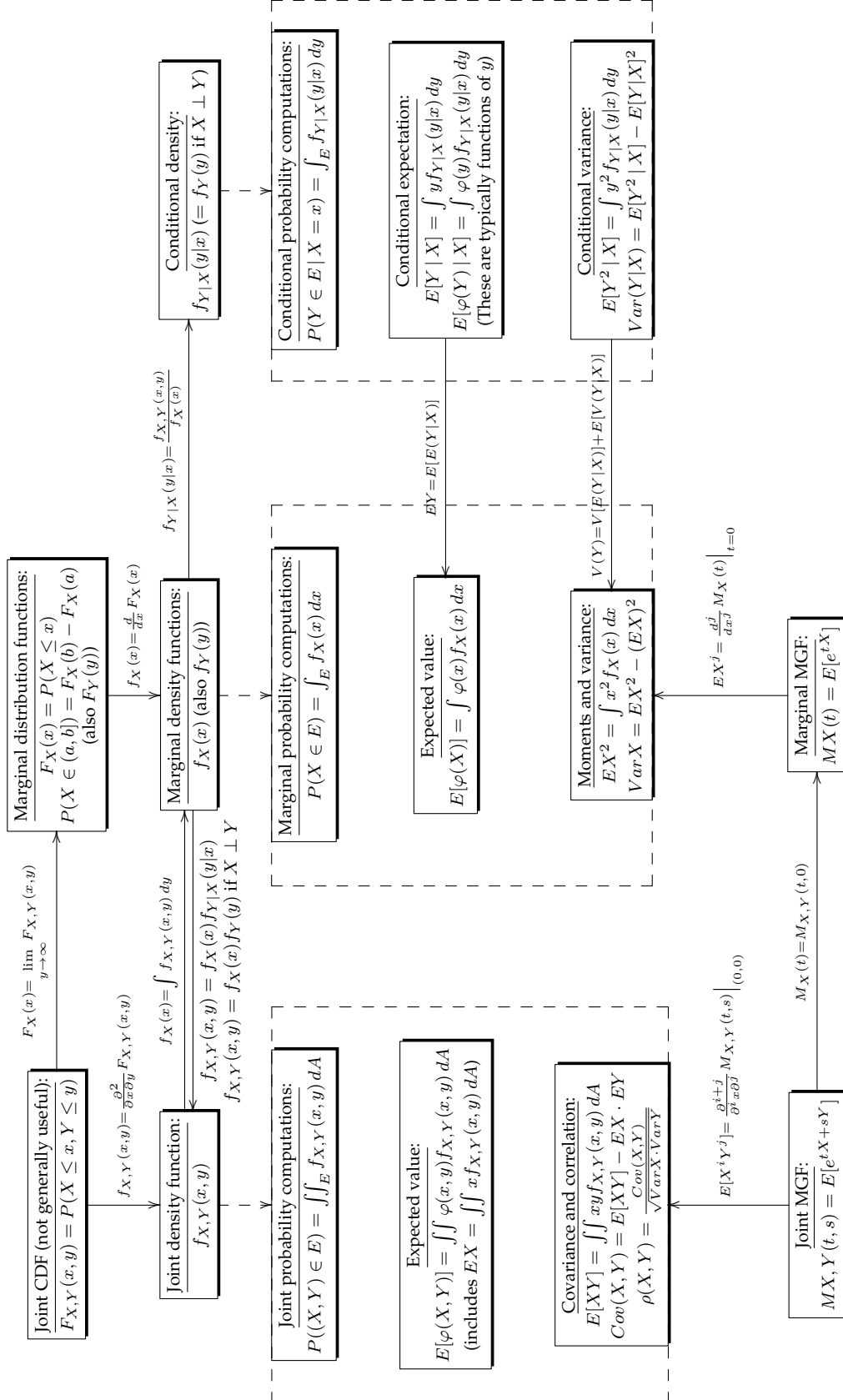
$$\int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

A.3 Table of values for the cdf of the standard normal

Entries represent $\Phi(z) = P(n(0, 1) \leq z)$. The value of z to the first decimal is in the left column. The second decimal place is given in the top row.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8436	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

A.4 Road map of standard computations with joint distributions



Index

- Φ , 98
- σ -algebra, 10
- σ -field, 10
- d -dim'l hypergeometric r.v., 113
- d -dimensional r.v., 108
- r^{th} moment, 168
- Addition Principle of Counting, 43
- Bayes' Law, 34
- benefit limit, 198
- Bernoulli process, 56
- Beta integral formula, 160
- binomial coefficients, 46
- binomial coefficients, properties of, 46
- binomial r.v., 58
- Binomial Theorem, 49
- bivariate normal distribution, 178
- Bonferonni inequality, 20
- Cauchy r.v., 81
- Cauchy-Schwarz inequality, 157
- cdf, 72
- Central Limit Theorem, 191
- characterization of joint normal r.v.s, 176
- characterization of normal r.v.s, 100
- Chebyshev inequality, 185
- Class 1 transformations, 133
- Class 2 transformations, 136
- Coat Check problem, 55
- combination, 45
- combinatorics, 42
- complement, 8
- Complement Rule, 19
- conditional density, 129
- conditional expectation, 159
- conditional expectation properties, 159
- conditional probability, 26
- conditional variance, 163
- continuity (of prob. measures), 22
- continuous (r.v.), 67
- continuous-time stochastic process, 56
- correlation, 156
- covariance, 154
- covariance matrix, 174
- covariance matrix properties, 175
- covariance properties, 155, 158
- coverage limit, 198
- coverage, proportional, 201
- cumulative distribution function, 72
- De Morgan Law, 19
- deductible, 195
- definiteness (of expected value), 146
- density function (of cts r.v.), 68
- density function (of discrete r.v.), 39
- diagram, tree, 33
- discrete (r.v.), 38
- discrete (set), 38
- discrete-time stochastic process, 56

-
- disjoint, 8
 - distinguishable arrangements, 50
 - distribution function, 72
 - event, 7
 - expected value, 141
 - expected value of a transformation, 143
 - expected value of common r.v.s, 149
 - expected value properties, 146
 - expected value, computing from survival function, 145
 - exponential r.v., 87
 - factorial, 44
 - failure (in a Bernoulli experiment), 56
 - finite expectation, 141
 - gamma function, 94
 - gamma function, properties, 94
 - Gamma Integral Formula, 96, 160
 - gamma r.v., 91, 95
 - generating function, 164
 - geometric r.v., 62
 - Hazard law for geometric r.v.s, 62
 - hazard law, exponential r.v., 87
 - hypergeometric r.v., 52
 - hypergeometric r.v. (d -dim'l), 113
 - i.i.d. process, 186
 - i.i.d. sequence of r.v.s, 97
 - identically distributed, 97, 186
 - Inclusion-Exclusion, 20
 - Inclusion-Exclusion (counting version), 43
 - Inclusion-Exclusion, 3-way, 24
 - Inclusion-Exclusion, general, 24
 - independence (of discrete r.v.s), 114
 - independence of cts r.v.s, 124
 - independence property of expected value, 148
 - independence property of mgfs, 169
 - independence property of pgfs, 167
 - independence test using mgfs, 173
 - independent (events), 27
 - independent events, 28
 - index set (of a stochastic process), 56
 - intersection, 8
 - Inversion formula, 171
 - Jacobian, 136
 - joint density function (of cts r.v.), 120
 - joint density function (of discrete r.v.), 108
 - joint distribution, 108
 - joint distribution function, 119
 - joint Gaussian r.v., 174
 - joint mgf, 172
 - joint moment generating function, 172
 - joint normal r.v., 174
 - Law of Small Numbers, 90
 - Law of the Unconscious Statistician, 143
 - Law of Total Expectation, 159
 - Law of Total Probability, 31
 - Law of Total Variance, 163
 - Lebesgue σ -algebra, 15, 16
 - Lebesgue measure, 15, 16
 - limit, benefit, 198
 - linearity (of expected value), 146
 - LOTUS, 143
 - marginals (cts case), 120
 - marginals (discrete case), 108
 - Markov inequality, 184
 - mean, 141
 - measurable (set), 10
 - memoryless, 86
 - memoryless (r.v.), 62
 - mgf, 168
 - mgf properties, 168
 - mgfs of common r.v.s, 169
 - Mississippi rule, 50
 - mixed (r.v.), 67
 - moment, 168
 - moment generating function, 168

- moment of order r , 168
- monotonicity (of expected value), 146
- Monotonicity Rule, 19
- Monty Hall problem, 30
- multinomial r.v., 113
- Multiplication Principle (of counting), 43
- Multiplication Principle (of probability), 27
- mutual independence (of events), 28
- mutually exclusive, 8

- negative binomial r.v., 64
- Normal integral formula, 160
- normal r.v., 100
- normalized average, 186

- orderings, 44
- outcome, 7

- pairwise independence (of events), 28
- partition, 31
- partition problems (counting), 51
- Pascal's Triangle, 47
- permutation, 45
- pgf, 164
- pgfs of common r.v.s., 166
- Poisson process, 83
- Poisson r.v., 90
- power set, 11
- preservation of bounds (expected value), 146
- preservation of constants (expected value), 146
- probability generating function, 164
- probability measure, 12
- probability space, 13
- properties of Φ , 99
- properties of binomial coefficients, 46
- properties of conditional densities, 130
- properties of conditional expectation, 159
- properties of covariance, 155, 158
- properties of covariance matrices, 175
- properties of cts joint densities, 121
- properties of discrete density functions, 41
- properties of distribution functions, 73
- properties of expected value, 146
- properties of joint mgfs, 172
- properties of mgfs, 168
- properties of pgfs, 165
- properties of the standard normal density, 99
- properties of variance, 154
- proportional coverage, 201

- Quantitative Weak Law of Large Numbers, 187

- random variable, 36
- random vector, 108
- rate (of a Poisson process), 87
- regression, 159

- sample space, 7
- sampling with replacement, 59, 113
- sampling without replacement, 51, 113
- Schwarz Inequality, 156
- sequence of averages, 97, 186
- sequence of normalized averages, 186
- sequence of sums, 186
- standard deviation, 151
- standard normal r.v., 98
- Stirling's Formula, 103
- stochastic process, 56
- Strong Law of Large Numbers, 189
- subadditivity (of prob. measures), 21
- success (in a Bernoulli experiment), 56
- success probability (of a Bernoulli process), 56
- sums of independent discrete r.v.s, 167
- sums of independent r.v.s, 171
- survival function, 75
- survival function, computing expected value from, 145

survival function, exponential r.v., 87
survival function, geometric r.v.s, 62

transformation (of a r.v.), 76
transformation (of cts joint distribution),
133
Transformation Theorem (1-dim, de-
creasing φ), 82
Transformation Theorem (1-dim, increas-
ing φ), 82
Transformation Theorem, higher dimen-
sions, 137
transformations, Class 1, 133
transformations, Class 2, 136
tree diagram, 33
trial (of a Bernoulli experiment), 56
Triangle inequality, 146
trivial σ -algebra, 11

uncorrelated, 156
uniform r.v. (cts), 70
uniform r.v. (discrete), 42
union, 8
uniqueness of joint normal densities,
176
uniqueness of mgfs, 171
uniqueness of pgfs, 166

Vandermonde's identity, 53
variance, 151
Variance Formula, 151
variance of common r.v.s, 152
variance properties, 154
variance, conditional, 163
vector-valued r.v., 108
Vitali set, 16

waiting time (in Poisson process), 85
Weak Law of Large Numbers, 187