PROBABILITY AND STATISTICS – MATH 241

LECTURE NOTES BY STEFAN WANER
BASED ON G.G. ROUSSAS *A COURSE IN MATHEMATICAL STATISTICS*

Contents

Date: July 2004.

1. Sets and Set Operations

A **set** is a collection of (distinct) **elements**, and we write $s \in S$ to indicate that s is an element of S. If S and T are sets, then S is a **subset** of T (we write $S \subseteq T$) if every element of S is an element of T. S is a **strict subset** of T (we write $S \subset T$) if every element of $S \subseteq T$ and $S \neq T$.

An axiom of set theory called **extensionality** says that a set is completely characterized by its elements; that is two sets are equal if and only if they have the same elements. The **empty set** $\emptyset$ is the set containing no elements. Two sets are **disjoint** if they contain no elements in common.

We think of the set S as a **universal set** if every set we consider is a subset of S. For example, if we are talking about real numbers, then we can take $S = \mathbb{R}$, the set of all real numbers as our universal set.

1.1. Set Operations Let S be a universal set.

(a) The **complement** of A is the set

$$A^c = \{\, s \in S \mid s \notin A \,\}$$

(b) The **union** of the sets $A_i;\ i = 1 \ldots n$ is the set

$$A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{i=1}^{n} A_i = \{\, s \in S \mid s \in A_i \text{ for at least one } i \,\}$$

Similarly, the union of the infinite sequence of sets $A_i;\ i = 1 \ldots$ is the set

$$A_1 \cup A_2 \cup \cdots = \bigcup_{i=1}^{\infty} A_i = \{\, s \in S \mid s \in A_i \text{ for at least one } i \,\}$$

(c) The **intersection** of the sets $A_i;\ i = 1 \ldots n$ is the set

$$A_1 \cap A_2 \cap \cdots \cap A_n = \bigcap_{i=1}^{n} A_i = \{\, s \in S \mid s \in A_i \text{ for every } i \,\}$$

Similarly, the intersection of the infinite sequence of sets $A_i;\ i = 1 \ldots$ is the set

$$A_1 \cap A_2 \cap \cdots = \bigcap_{i=1}^{\infty} A_i = \{\, s \in S \mid s \in A_i \text{ for every } i \,\}$$

Notice that two sets A and B are disjoint if and only if $A \cap B = \emptyset$.

(d) The difference $A - B$ of the sets A and B is defined by

$$A - B = \{\, s \in S \mid s \in A,\ s \notin B \,\}$$

Notation 1.2. The collection $A_1, A_2, \ldots$ of sets is **mutually** (or **pairwise**) **disjoint** if $A_i \cap A_j = \emptyset$ for every $i \neq j$. When this is so, we sometimes write $\bigcup_{i=1}^{n} A_i$ as $\sum_{i=1}^{n} A_i$, and similarly for infinite sums.

Lemma 1.3. *The following hold for any three sets A, B and C and any indexed collection of sets B_α ($\alpha \in \Omega$):*

(1) Associativity:
$$A \cup (B \cup C) = (A \cup B) \cup C \qquad A \cap (B \cap C) = (A \cap B) \cap C$$
(2) Commutativity:
$$A \cup B = B \cup A \qquad A \cap B = B \cap A$$
(3) Identity and Idempotent
$$A \cup A = A, \qquad A \cap A = A$$
$$A \cup \emptyset = A, \qquad A \cap \emptyset = \emptyset$$
(4) De Morgan's Laws:
$$D - (A \cup B) = (D - A) \cap (D - B), \quad D - (A \cap B) = (D - A) \cup (D - B)$$
$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c$$
Fancy Form:
$$A - \bigcup_i B_i = \bigcap_i (A - B_i), \quad A - \bigcap_i B_i = \bigcup_i (A - B_\alpha)$$
(5) Distributive Laws:
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C), \quad A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
Fancy Form:
$$A \cup (\bigcap_i B_i) = \bigcap_i (A \cup B_i), \quad A \cap (\bigcup_i B_i) = \bigcup_i (A \cap B_i)$$

Proof We prove (1), (2), (3) and a bit of (4) in class. The rest you will prove in the exercise set. □

Definitions 1.4.

a. The sequence (A_i) of sets is **monotone** if either
$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots, \qquad \text{written } A_n \uparrow \qquad \text{or}$$
$$A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots, \qquad \text{written } A_n \downarrow$$

b. Given a monotone sequence of sets, $(A_n) \uparrow$ (resp. $(A_n) \downarrow$), we define their **limit** by
$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n \qquad (\text{ resp. } \lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n).$$

c. More generally, if (A_n) is any sequence of sets, we define two associated limits, $\bar{A}$ and $\underline{A}$, the **limit superior and the limit inferior,** by
$$\bar{A} = \limsup_{n\to\infty} A_n = \lim_{n\to\infty} \bigcup_{i=n}^{\infty} A_i \qquad \text{(The limit of a decreasing sequence)}$$
and
$$\underline{A} = \liminf_{n\to\infty} A_n = \lim_{n\to\infty} \bigcap_{i=n}^{\infty} A_i \qquad \text{(The limit of an increasing sequence)}$$
The sequence (A_n) is **convergent** if the limit inferior and limit superior agree.

Notes 1.5. The following are shown in the exercises:

a. In general

$$\liminf_{n\to\infty} A_n \subseteq \limsup_{n\to\infty} A_n.$$

b. If $(A_n) \uparrow$ or $(A_n) \downarrow$ then (A_n) is convergent.

Exercise Set 1.

1. Prove Lemma 1.3 (4) and (5).

2. By consulting the book or otherwise, show that any union $\bigcup_i A_i$ of sets A_i is equal to a disjoint union, $\sum_i B_i$ for suitable sets $B_i \subseteq A_i$. [Draw a picture to see what is going on.]

3. Prove the assertions in Notes 1.5.

4. Give an example of a convergent sequence of sets that is not (even eventually) monotone. [Hint: Try a moving interval . . .]

5. For each of the following, determine whether (i) the given sequence is monotone, and (ii) whether it converges. Justify your assertions by showing the computations.

a. $\left[1-\frac{1}{n}, 3+\frac{1}{n}\right]$

b. $\left[1+\frac{1}{n}, 3-\frac{1}{n}\right]$

c. $\left[1-\frac{1}{n}, 3-\frac{1}{n}\right]$

6. Let

$$A_n = \left\{ (x,y) \in \mathbb{R}^2 \mid x^2+y^2 \le \frac{1}{n} \right\},$$

Show that A_n is monotone and find its limit.

7. Compute lim inf and lim sup for the following:

a. $\left[1+\frac{1}{n}, 3-\frac{1}{n}\right]$

b. $\left\{ (x,y) \in \mathbb{R}^2 \mid x^2+y^2 = 1-\frac{1}{n} \right\}$

c. $\left\{ (x,y) \in \mathbb{R}^2 \mid 1-\frac{1}{n} \le x^2+y^2 \le 1 \right\}$

d. $\left\{ (x,y) \in \mathbb{R}^2 \mid x^2+y^2 \le n^2-10n \right\}$

8. (From the book) Show that

$$\underline{A} = \{\, x \in S \mid x \text{ belongs to all but finitely many of the } A\text{s} \,\}$$
$$\bar{A} = \{\, x \in S \mid x \text{ belongs to infinitely many of the } A\text{s} \,\}$$

9. Let $\mathcal{A}$ be any collection of subsets of the set S, and define a possibly larger collection $\mathcal{F}$ as follows: the sets in $\mathcal{F}$ are defined to be all subsets of the form

$$\Delta_{n_1} \dots \Delta_{n_r} A_{n_1,n_2,\dots,n_r},$$

where the n_i are in countable sets, each symbol Δ represents either $\bigcup$ or $\bigcap$, and each $A_{n_1,n_2,\ldots,n_r}$ is either a set in $\mathcal{A}$ or the complement of set in $\mathcal{A}$. (In words, we are taking all sets in $\mathcal{A}$, throwing in all complements, and then including everything that can be obtained in a finite number of steps by taking countable unions and countable intersections.) Show that $\mathcal{F}$ is closed under countable unions and complements; that is, show that countable unions of sets in $\mathcal{F}$ are in $\mathcal{F}$, and that complements of sets in $\mathcal{F}$ are in $\mathcal{F}$.

2. Probability Functions

Intuitively, we think of a **non-deterministic experiment** as an experiment whose outcome is uncertain. The **sample space** associated with an experiment is the set S of all conceivable outcomes. However, for our discussion of abstract probability, S will be any set whatsoever.

Definition 2.1. Let S be a set. A **σ-field** $\mathcal{F}$ in S is a collection of subsets of S such that:

(a) $S \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$;
(b) $A \in \mathcal{F}$ implies $A' \in \mathcal{F}$;
(c) $A_n \in \mathcal{F}$ implies $\bigcup_n A_n \in \mathcal{F}$.

We call the sets in $\mathcal{F}$ **events.** Mutually disjoint sets in $\mathcal{F}$ are called **mutually exclusive events.**

It follows that σ-fields are closed under intersections and complements, as well as any other operations we might be interested in. We think of a σ-field as the **collection of events** associated with an experiment whose sample space is S.

Definition 2.2. (Kolmogorov) Let S be a set and let $\mathcal{F}$ be a σ-field on S. A **probability function** (or **probability measure**) **on** S is a function $P : \mathcal{F} \to \mathbb{R}$ such that

(P1) P is non-negative: $P(A) \geq 0$ for each $A \in \mathcal{F}$.
(P2) P is normed: $P(S) = 1$.
(P3) P is σ-additive: $P(\sum_i A_i) = \sum_i P(A_i)$ for every collection $\{ A_i \}$ of mutually disjoint sets in $\mathcal{F}$.

Proposition 2.3. *[Further Properties of Probability Measures]*

(C1) $P(\emptyset) = 0$
(C2) *P is finitely additive: For any finite sequence $\{ A_i \}_{i=1}^n$ of mutually disjoint events, one has $P(\sum_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$.*
(C3) *(Monotonicity) For every pair of events A and B, one has $A \subseteq B \Rightarrow P(A) \leq P(B)$.*
(C4) *For every event A, $0 \leq P(A) \leq 1$.*
(C5) *(Complements) For every event A, one has $P(A') = 1 - P(A)$.*
(C6) *For every pair of events A and B, one has*

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

(C7) *(Subadditivity) For every collection $\{ A_i \}$ of events,*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} P(A_i)$$

$$\text{and } P\Big(\bigcup_{i=1}^{n} A_i\Big) \leq \sum_{i=1}^{n} P(A_i)$$

To prove (C1), write S as $S+\emptyset+\emptyset+\dots$ and apply σ-additivity. For (C2) just add on infinitely many empty sets as well. The rest of the properties follow formally, except for 7, which is proved int he exercises.

Examples 2.4.

A. Let$S = \{ s_1, s_2, \dots, s_n \}$ be any finite set, then take the associated σ-field $\mathcal{F}$ to be the collection of all subsets of S. A probability function on S is then any function $P : S \to \mathbb{R}$ with
   $0 \leq P(s_i) \leq 1$
   $P(s_1) + \dots + P(s_n) = 1$
   This is the definition given in "baby probability" courses.

B. Take $S = [0, 1]$, the closed unit interval, and let $\mathcal{F}$ be defined to be the smallest σ-field[1] containing the closed sub-intervals of $[0, 1]$. (Sets in $\mathcal{F}$ are referred to as **Borel sets.**) We now define $P : \mathcal{F} \to \mathbb{R}$ as follows: If $[a_i, b_i]$ is a non-overlapping sequence[2] of subsets of $[0, 1]$, take

$$P(\sum_i [a_i, b_i]) = \sum_i (b_i - a_i)$$

   Next, observe that every union of closed intervals can be expressed as the union of non-overlapping closed intervals, so this allows us to define P for arbitrary unions of closed intervals. It can be verified that this is well-defined (that is, independent of decomposition into non-overlapping sets). This particular probability function is referred to as **Lesbégue measure.**

C. S'pose $f : D \to \mathbb{R}$ is any integrable function whose domain D is an interval (of any type — so that D is a Borel set — see the exercises). Assume further that f satisfies the following properties:
   (a) $f(x) \geq 0$ for every $x \in D$.
   (b) $\int_D f(x)dx = 1$
   Then f is called a **probability density function,** and we can define

$$P([a, b]) = \int_a^b f(x)\ dx$$

   for every interval $[a, b] \subseteq D$. As in item B above, this permits us to define P on the σ-field of Borel sets in D.

Proposition 2.5 (Another Property of Probability Functions)**.**
If $\{ A_i \}$ *is any monotone sequence of events, then*

$$P(\lim_{i\to\infty} A_i) = \lim_{i\to\infty} P(A_i)$$

[1]Formally, this is the intersection of all σ-fields that contain the closed intervals in $[0, 1]$. In practice, we can get it by following the procedure in Exercise Set 1 #9.

[2]Two closed intervals are non-overlapping if their intersection is at most a single point. A sequence of closed intervals is non-overlapping if they are pairwise non-overlapping.

Proof. Assume first that $\{ A_i \}$ is increasing, so that its limit is given by $\bigcup_i A_i$. Then we can write $\bigcup_i A_i$ as

$$\bigcup_i A_i = A_1 + (A_2 - A_1) + (A_3 - A_2) + \dots$$

By σ-additivity, one has

$$\begin{aligned} P(\bigcup_i A_i) &= P(A_1) + P(A_2 - A_1) + P(A_3 - A_2) + \dots \\ &= \lim_{i\to\infty} P(A_1) + P(A_2 - A_1) + \dots + P(A_i - A_{i-1}) \\ &= \lim_{i\to\infty} P(A_1) + P(A_2) - P(A_1) + \dots + P(A_i) - P(A_{i-1}) \qquad \text{(Exer. 3)} \\ &= \lim_{i\to\infty} P(A_i). \end{aligned}$$

The proof for decreasing sequences is left to the exercise set.

□

Exercise Set 2.

1. Prove Proposition 2.3(C7)

2. Show that the Borel sets (Example 2.4 B) include all open intervals, and single point sets.

3. Prove that probability functions satisfy

$$P(A - B) = P(A) - P(B)$$

for all events A and $B \subseteq A$.

4. Finish the proof of Proposition 2.5.

5. Let $S = \{ 1, 2, \dots, n\}$ and let $P(k)$ be the probability of getting k heads in n tosses of a fair coin.

- **a.** Give a formula for $P(k)$
- **b.** Obtain a formula for $P(A)$, where A is any subset of S so that P is a probability function.
- **c.** Verify directly from your formulas that $P(S) = 1$.

6. Let $S = \{ 1, 2, \dots \}$ and let $P(k)$ be the probability of getting heads for the first time on the k^{th} toss of a fair coin.

- **a.** Give a formula for $P(k)$ [Hint: Draw a tree.]
- **b.** Obtain a formula for $P(A)$, where A is any subset of S so that P is a probability function.
- **c.** Verify directly from your formulas that $P(S) = 1$.

7. Take $S = \mathbb{R}$, let $f(x) = \dfrac{1}{\sqrt{\pi}}\, e^{-x^2}$, and define

$$P(B) = \int_B f(x)\, dx.$$

Verify that P is a probability function on $\mathbb{R}$, at least as defined on the set of all intervals.

8. If (a_n) is a sequence of real numbers, define

$$\liminf_{n\to\infty} a_n = \lim_{k\to\infty} \left(\inf_{i\geq k} a_i \right)$$

and

$$\limsup_{n\to\infty} a_n = \lim_{k\to\infty} \left(\sup_{i\geq k} a_i \right)$$

Note that these limits always exist, the first is $\leq$ the second, and they are equal to each other iff the actual limit exists (in which case they are also equal to the limit) and prove that, for an arbitrary sequence $\{ A_i \}$ of events, one has

$$P(\underline{\mathrm{A}}) \leq \liminf_{n\to\infty} P(A_n) \leq \limsup_{n\to\infty} P(A_n) \leq P(\bar{A})$$

Deduce that, if $\{ A_i \}$ is a convergent set, one has

$$P(\lim_{i\to\infty} A_i) = \lim_{i\to\infty} P(A_i)$$

3. Conditional Probability

Definition 3.1. Let A be an event, and suppose that $P(A) > 0$. Then the **conditional probability given** A, $P(-, A)$, is the probability function defined by

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

Notice that this can also be written as

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

If it happens that $P(B|A) = P(B)$, we say that A and B are **independent.** Looking at the second formulation, this amounts to saying that two events with non-zero probability are independent iff

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

We check that it is indeed a probability function in class.

Examples 3.2.

A. $P(-|S) = P(-)$.
B. Find the probability that throwing two fair dice leads to a sum of 5, given that the first throw is odd.
C. Find the probability that throwing two fair dice leads to a sum of 5, given that at least one of the throws is odd.

Theorem 3.3 (Multiplicative Theorem)**.**
S'pose $\{ A_i \}_{i=1}^n$ *is a sequence of events with the property that* $P(A_1 \cap \ldots A_{n-1}) \neq 0$. *Then*

$$P(\bigcap_{i=0}^{n} A_i) = P(A_n \mid A_1 \cap \cdots \cap A_{n-1}) \times P(A_{n-1} \mid A_1 \cap \cdots \cap A_{n-2}) \ldots P(A_2 \mid A_1)P(A_1).$$

The theorem is proved in the exercises. In class, we illustrate this theorem by means of a tree diagram.

Examples 3.4 (from the book)**.**

A. If $\{ A_i \}_{i=1}^n$ consists of two sets $A_1,\ A_2$.
B. A fair coin is tossed 3 times in a row, and then a fair die is rolled. Assume that the successive outcomes are independent of the history. Find the probability of the outcome HTH4.
C. A bag contains five black marbles, three red ones, and two white ones. Four are drawn randomly and in sequence (without replacement). Find the probability that the first one is black, the second is red, the third white, and the fourth black.

Definition 3.5. A collection $\{ A_i \}$ of events is called a **partition of** S if they are mutually disjoint, and their union is S.

If $\{ A_i \}$ is a partition of S and B is any event, then

$$B = \sum_i (B \cap A_i), \quad \text{whence}$$
$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B \mid A_i) P(A_i), \tag{1}$$

by Definition 3.1. This formula leads to the following:

Theorem 3.6 (Bayes' Formula)**.**
If $\{ A_i \}$ is a partition of S and B is any event, then

$$P(A_j \mid B) = \frac{P(B \mid A_i) P(A_j)}{\sum_j P(B \mid A_i) P(A_i)}.$$

Proof. We have

$$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j \cap B)}{\sum_i P(B \mid A_i) P(A_i)},$$

by equation (1). □

Corollary 3.7 (Bayes' Formula for Babies)**.**
If A and B are any two events with non-zero probability, then

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B \mid A) P(A) + P(B \mid A') P(A')}.$$

Proof. Apply Bayes' formula to the partition $A,\ A'$ of S. □

Exercise Set 3.

1. Compute the following conditional probabilities, and determine whether the given pairs of events are independent.
   - **a.** Two fair dice (one red and one green) are rolled. Find the probability that the sum is 5 given that the dice have opposite parity.
   - **b.** A bag contains 3 red marbles, 2 green ones, 1 fluorescent pink one, 2 yellows and 2 orange ones. An alien grabs 4 at random. Find the probability that it gets none of the red ones, given that it gets the fluorescent pink one
2. A box contains two chocolate truffles, three chocolate-covered almonds, and one Turkish delight.
   - **a.** While watching exciting events on CSpan, you eat three of them, selecting one at random each time. What is the probability that your choices are: chocolate truffle, almond, and almond, in that order?
   - **b.** Your friend is watching CSpan with you, and greedily grabs three of your candies at once (before you have even had a chance to have one). Find the probability that she has a chocolate truffle and two almonds.

3. A company wishes to enhance productivity by running a one-week training course for its employees. Let T be the event that an employee participated in the course, and let I be the event that an employee's productivity improved the week after the course was run.
   a. Assuming that the course has a positive effect on productivity, how are $P(I \mid T)$ and $P(I)$ related?
   b. If T and I are independent, what can one conclude about the training course?

4. A man was arrested for attempting to smuggle a bomb on board an airplane. During the subsequent trial, his lawyer claimed that, by means of a simple argument, she would prove beyond a shadow of a doubt that her client was not only innocent of any crime, but was in fact contributing to the safety of the other passengers on the flight. This was her eloquent argument: "Your Honor, first of all, my client had absolutely no intention of setting off the bomb. As the record clearly shows, the detonator was unarmed when he was apprehended. In addition-and your Honor is certainly aware of this-there is a small but definite possibility that there will be a bomb on any given flight. On the other hand, the chances of there being two bombs on a flight are so remote as to be negligible. There is in fact no record of this having ever occurred. Thus, since my client had already brought one bomb on board (with no intention of setting it off) and since we have seen that the chances of there being a second bomb on board were vanishingly remote, it follows that the flight was far safer as a result of his action! I rest my case." This argument was so elegant in its simplicity that the judge acquitted the defendant. Where is the flaw in the argument?

5. Prove Theorem 3.3.

6. Any athlete who fails the Enormous State University's women' soccer fitness test is automatically dropped from the team. (The fitness test is traditionally given at 5 AM on a Sunday morning.) Last year, Mona Header failed the test, but claimed that this was due to the early hour. In fact, a study by the ESU Physical Education Department suggested that 50% of athletes fit enough to play on the team would fail the soccer test, although no unfit athlete could possibly pass the test. It also estimated that 45% of the athletes who take the test are fit enough to play soccer. Assuming these estimates are correct, what is the probability that Mona was justifiably dropped?

7. Two of the mathematics professors at Enormous State are Professor A (known for easy grading) and Professor F (known for tough grading). Last semester roughly three quarters of Professor F's class consisted of former students of Professor A; these students apparently felt encouraged by their (utterly undeserved) high grades. (Professor F's own former students had fled in droves to Professor A's class to try to shore up their grade point averages.) At the end of the

semester, as might have been predicted, all of Professor A's former students wound up with a C- or lower. The rest of the students in the class-former students of Professor F who had decided to "stick it out"—fared better, and two thirds of them earned higher than a C-. After discovering what had befallen them, all the students who earned C- or lower got together and decided to send a delegation to the Department Chair to complain that their grade point averages had been ruined by this callous and heartless beast! The contingent was to consist of ten representatives selected at random from among them. How many of the ten would you estimate to have been former students of Professor A?

8. The following letter appeared in The New York Times (January 16, 1996, p. A16).

> To the Editor:
> It stretches credulity when William Safire contends (column, Jan. 11) that 90 percent of those who agreed with his Jan. 8 column, in which he called the First Lady, Hillary Rodham Clinton, "a congenital liar," were men and 90 percent of those who disagreed were women. Assuming these percentages hold for Democrats as well as Republicans, only 10 percent of Democratic men disagreed with him. Is Mr. Safire suggesting that 90 percent of Democratic men supported him? How naive does he take his readers to be?
> A. D.
> New York, Jan. 12, 1996

Comment on the letter writer's reasoning.

9. (From the book #2.2.13 *et seq.*) Consider two urns U_1 and U_2, where U_i contains m_i white balls and n_i black balls.

- **a.** A ball is drawn at random from each urn and placed into a third urn. Then a ball is drawn at random from the third urn. Compute the probability that the ball is black.
- **b.** A fair die is rolled, and if an even number appears, a ball, chosen at random from U_1, is placed in U_2. If an odd number appears, a ball, chosen at random from U_2, is placed in U_1. What is the probability that, after the above experiment is performed twice, the number of white balls in urn U_2 remains the same?
- **c.** Consider 6 urns $U_1, \ldots U_6$ such that urn U_i contains m_i (≥ 2) white balls and n_i (≥ 2) black balls. A balanced die is tossed once, and if the number j appears on the die, then two balls are selected at random from Urn U_j. What is the probability that one ball is white and the other black?

4. Independence

Definitions 4.1. The events A and B are **independent** if one or both of $P(A)$ and $P(B)$ are zero, or

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

Equivalently, A and B are **independent** if

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

More generally, the events $A_i,\ i = 1, \ldots n$ are **mutually independent** if

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \ldots P(A_{i_k}).$$

whenever $1 \leq i_1 \leq \cdots \leq i_k \leq n$. Also, the events $A_i,\ i = 1, \ldots n$ are **pairwise independent** if

$$P(A_i \cap A_j) = P(A_i)P(A_j)$$

whenever $i \neq j$.

Examples 4.2.

A. Cast two fair dice;
   A : The first is even.
   B : the sum is 7.
B. Cast two fair dice;
   A : The first is even.
   B : the sum is 6.
C. (Pairwise independent but not mutually independent) Toss two fair coins.
   A : The first coin comes up heads.
   B : The second coin comes up heads.
   C : Both coins come up the same.
D. (Mutually independent—and hence pairwise independent) Toss three fair coins.
   A : The first coin comes up heads.
   B : The second coin comes up heads.
   C : The third coin comes up heads.

Theorem 4.3. *If the events $A_i,\ i = 1, \ldots n$ are mutually independent (resp. pairwise independent), then so are their complements.*

The proof for mutual independence is in the textbook. The proof for pairwise independence is in the exercise set. □

Corollary 4.4 (Probability of the Union of Independent Events)**.**

(1) If $A_i,\ i = 1, \ldots n$ are mutually independent, then

$$P(A_{i_1} \cup \cdots \cup A_{i_k}) = 1 - P(A'_{i_1}) \ldots P(A'_{i_k}).$$

whenever $1 \leq i_1 \leq \cdots \leq i_k \leq n$.

(2) If $A_i,\ i = 1, \ldots n$ are pairwise independent, then

$$P(A_i \cup A_j) = 1 - P(A_i')P(A_j')$$

whenever $i \neq j$.

Exercise Set 4.

1. **a.** Show that the events $S,\ \emptyset,\ A$ are mutually independent for any event A not equal to S or $\emptyset$.
   b. Show that an event A is independent of itself iff $P(A) = 0$ or 1.
2. Consider the experiment in which you toss two fair coins. Find four mutually independent events.
3. Continuing with the two coin experiment, is it possible to find three mutually independent events excluding S and $\emptyset$? Justify your assertion.
4. Prove Theorem 4.3 in the case that the given events are mutually independent.
5. Prove that two events are both mutually exclusive and independent iff at least one of them has zero probability. Generalize this result to n events.
6. Sosa SoSo Stores sell gold chains, bracelets and fake tan lotion at various branches. Annual sales at the Washington and Texas branches are shown in the following table:

	Gold Chains	Bracelets	Tan Lotion	Total
Washington	10,000	10,000	60,000	80,000
Texas	15,000	20,000	85,000	120,000
Total	25,000	30,000	145,000	200,000

   Let C be the event that a gold chain was sold, let B be the event that a bracelet was sold, and let W be the event that an item was sold in Washington.
   a. Evaluate $P(C \mid W)$ and $P(W \mid C)$.
   b. Are C and W independent? Why?
   c. Are B and W independent? Why?
   d. Are C and B independent? Why?
7. According to the weather service, there is a 50% chance of rain in New York and a 30% chance of rain in Honolulu. Assuming that New York's weather is independent of Honolulu's, find the probability that it will rain in at least one of these cities.
8. A company wishes to enhance productivity by running a one-week training course for its employees. Let T be the event that an employee participated in the course, and let I be the event that an employee's productivity improved the week after the course was run.
   a. Assuming that the course has a positive effect on productivity, how are $P(I \mid T)$ and $P(I)$ related?

b. If T and I are independent, what can one conclude about the training course?

9. (From the text) Two people A and B play a game by repeatedly (and independently) rolling two fair dice. If the total rolled in any turn is 6, player A is the winner and the game is terminated. If the total rolled in any turn is 10, player B is the winner and the game is terminated.

a. Find the probability that the game terminates after n steps with A the winner.

b. Find the probability that the game terminates after n turns.

c. Find the probability that player A eventually wins.

d. Find the probability that player B eventually wins.

e. Find the probability that the game never terminates.

10. Let $S = [0, 1]$, and take the set of events to be the Borel sets in $[0, 1]$. If $0 \leq a \leq b \leq c \leq d \leq 1$, Prove that the events $[a, c]$ and $[b, d]$ are independent iff $c = a$ or $d = b + \frac{(c-b)}{(c-a)}$.

5. Random Variables

Definition 5.1. Let P be a probability function on $(S, \mathcal{F})$. A **random variable** is a function $X : S \to \mathbb{R}$ such that the preimage of every Borel set is in $\mathcal{F}$. That is, the preimage of every Borel set is an event.

We denote by $X(S)$ the image of X; that is, the set of possible values of X.

Notes 5.2.

(1) Since the single point sets $\{ b \}$ are Borel, their preimages under X are events, and hence have probabilities. In other words, we can define

$$P(X = b) = P(X^{-1}(b)) = P(\{ s \in S \mid X(s) = b \}),$$

the probability of the event that $X = b$.

(2) More generally, we would like to consider the probability that $X \in B$, for some Borel set B. And so we define

$$P(X \in B) = P(X^{-1}(B)) = P(\{ s \in S \mid X(s) \in B \}),$$

noting that $X^{-1}(B)$ is an honest-to-goodness event.

(3) Notice that, if we define $P_X : \mathcal{B} \to \mathbb{R}$ by

$$P_X(B) = P(X \in B)$$

we get a new probability function P_X. Put another way, the collection of sets $X^{-1}(B)$ gives us a σ-subfield of $\mathcal{F}$, and the restriction of P to this new σ-field is P_X. In other words,

P_X is just P with a new set of events.

Examples 5.3.

A. Toss three coins, and assign to each outcome the number of heads. Compute the probability function P_X we get.
B. Cast two fair dice, and assign to each outcome the sum of the numbers facing up. Compute the resulting probability function.
C. In general, if $S \subseteq \mathbb{R}$, then we can just take X to be the identity function, $X(s) = s$. In that even, one has $P(X \in B) = P(B)$, so we just have the original probability function.
D. Roll a die repeatedly until you get a 6. (So S is the set of all finite sequences of numbers 1–5 in which 6 occurs exactly once and at the end.) Let X be the number of times you rolled the die. Compute the resulting probability function.
E. Take $S = \mathbb{R}$, and let $X(s) =$ Temperature in Central Park at time s.

Definition 5.4. The random variable X is called **discrete** if there is a countable set of real numbers $D = \{ x_1, x_2, \dots \}$ such that $X(S) \subseteq D$ for each i, so that

$$\sum_i P(X = x_i) = 1.$$

If we define $f: D \to \mathbb{R}$ by

$$f(x_i) = P(X = x_i)$$

then we can compute $P(X \in B)$ for some subset B of D by just summing the $f(x_i)$s for $x_i \in B$.

Definition 5.5. If X is a discrete random variable with associated discrete values in the discrete set $D \subset \mathbb{R}$, then the associated **(discrete) probability density function** is given by $f: D \to \mathbb{R}$ as defined above.

Examples 5.6 (Of Discrete Random Variables)**.**
We look to see which of the above examples is discrete.

Definition 5.7. The random variable X is called **continuous** if $P(X = x) = 0$ for every $x \in X$.

Typical examples of continuous random variables arise from *(continuous) probability density functions.*

Definition 5.8. Let I be an interval (possibly infinite) of the real line. A **(continuous) probability density function on** I is a non-negative integrable function $f : \mathbb{R} \to \mathbb{R}$ with the property that $\int_I f(x)\ dx = 1$.

Any probability density function leads to a probability function P defined on the Borel subsets of I by taking

$$P(B) = \int_B f(x)\ dx,$$

as we did in Exercise Set 2. Note that the sample space in this context is just the real line: $S = \mathbb{R}$. We can now define a continuous random variable $X : S \to \mathbb{R}$ by taking X to be the identity function on $\mathbb{R}$. Thus, in particular

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

5.9. Very Important Remark
The Riemann integral is not defined over arbitrary Borel sets — in fact, it is only defined for disjoint unions of closed intervals. To integrate over an arbitrary Borel set, we need a more powerful form of integral: the *Lesbégue integral*, which extends the Riemann integral to permit integration over arbitrary Borel sets, and also enlarges the collection of functions that we can integrate. The study of these integrals is beyond the scope of this course, but we shall assume everything we need about such integrals with gay abandon. The squeamish among you can think "disjoint union of closed intervals" whenever these notes mention "Borel set."

Q Why is this a continuous random variable?
A Because $P(X = a) = \int_a^a f(x)\ dx = 0$ for every $a \in X$.

Examples 5.10.

A. **Uniform Random Variable:** Define $f : \mathbb{R} \to \mathbb{R}$ by taking

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b]; \\ 0 & \text{otherwise.} \end{cases}$$

B. **Normal Distribution with St. Deviation $\frac{1}{\sqrt{2}}$:**

$$f(x) = \frac{1}{\sqrt{\pi}}\, e^{-x^2}$$

C. **General Normal Distribution:**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ and σ are non-zero constants.

Exercise Set 5.

1. In each of the following, give the probability distribution P_X associated with the indicated random variable X. In each case, verify that $\sum_i P_X(x_i) = 1$.

a. A red and a green die are rolled, and $X = 0$ if the numbers are the same, and equal to 1 if the numbers are different.

b. Three fair tetrahedral dice (four faces; numbered 1, 2, 3, 4) are rolled. X is the sum of the numbers facing up.

c. Consider any experiment performed repeatedly and independently in which there are two outcomes: success, with probability a and failure, with probability b. The experiment ends when success is accomplished. Take X to be the number of times the experiment is performed.

2. For which value of the constant a is the given function a probability density function?

(1) $f : [0, +\infty) \to \mathbb{R}$; $f(x) = ae^{-rx}$; $(r > 0)$

(2) $f : [0, 1] \to \mathbb{R}$; $f(x) = ax^{\beta}$ (β a fixed positive constant)

(3) $f : [0, 1] \to \mathbb{R}$; $f(x) = axe^{-x^2}$

3. Verify that the general normal distribution (Example 5.10(C)) is indeed a probability density function. [To integrate it, use a suitable change of variables to reduce it to the simpler integral you have already considered.]

6. Some Discrete Random Variables

Binomial Random Variables

Definition 6.1. A **Bernoulli trial** is an experiment that has two possible outcomes, called **success** and **failure**. (Thus, if the probability of success is p then the probability of failure is $q = 1-p$.) A **binomial random variable** is one that counts the number of successes in a sequence of independent Bernoulli trials.

S'pose that we have a sequence of n independent Bernoulli trials. Then the image of X is $X(S) = \{ 0, 1, 2, \ldots, n \}$, and

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

In class we check, using the binomial theorem to expand $1 = (p + (1-p))^n$ that $\sum_{i=0}^{n} P(X = i) = 1$. Also, the distribution is basically bell-shaped.

Examples 6.2.

A. Flip a fair coin n times; $X =$ the number of successes.
B. Choose 50 people at random in a large population in which 20% of the population uses Wishy Washy detergent. $X =$ the number in the sample that use it. Compute $P(48 \leq X \leq 50)$.

Poisson Random Variables

S'pose we have a very large sequence of Bernoulli trials with a very small probability, so that the product $\lambda = np$ is reasonable large.

For instance, if it is known that an average of 150 people arrive at a car wash each lunch hour. Consider this as a sequence of Bernoulli trials in which $n =$ entire population of the town and $p =$ the probability that a randomly selected person will just happen to go to he carwash on that particular lunch hour. Then $np =$ expected number of successes $= 150$. We will see that, under these circumstances, we can approximate the (horrendus) binomial distribution computation using the following approximation:

Definition 6.3. Let λ be a positive real number. Then the associated **Poisson random variable** X has image $X(S) = \{ 1, 2, \ldots \}$ and is given by

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

Notes 6.4.

(1) Notice that summing $P(X = x)$ from $x = 0$ to ∞ gives $e^{-\lambda} e^{\lambda} = 1$.
(2) The more conventional interpretation of the carwash problem: Think of the lunch hours being broken up into a large number n of small time intervals Δt, and take p to be the probability that a car will arrive in that interval. Then again $np =$ expected number of cars arriving in the lunch hour $= 150$.

(3) The parameter λ is numerically equal to the expected number of successes.
(4) Note that there is no upper limit to the value of X (unlike the case for the binomial distribution).
(5) To generate the entire distribution in Excel with, say, $\lambda = 20$, use the formula `=EXP(-20)*20^x\FACT(x)`.

Examples 6.5.

A. A radioactive source is emitting an average of 10 particles per millisecond. What is the probability that, in a given millisecond, the number of particles emitted will be exactly 9.
B. In a bank, an average of 3 people arrive each minute. Find the probability that, in a given minute, more than 2 people will arrive.

Hypergeometric Random Variables

This is similar to the binomial random variable, except that, instead of of performing trials with replacement (eg. select a lightbulb, determine whether it is defective, then replace it and repeat the experiment) we do not replace it. This makes the success more likely after a string of failures.

For example, we know that 30 of the 100 workers at the Petit Mall visit your diner for lunch. You choose 10 workers at random; X is the number of workers who visit your diner. (Note that the problem approaches the binomial distribution for a large population, where we can ignore the issue of replacement). If X is hypergeometric, we compute $P(X = x)$ as follows: If

$$\begin{aligned} N &= \text{ population size} \\ n &= \text{ number of trials} \\ r &= \text{ number of possible successes} \end{aligned}$$

then

$$P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

Example 6.6. The Gods of Chaos have promised you that you will win on exactly 40 of the next 100 bets at the Happy Hour Casino. However, your luck has not been too good up to this point: you have bet 50 times and have lost 46 times. What are your chances of winning both of the next two bets?
Solution Here N = number of bets left = $100 - 50 = 50$, n= number of trials = 2 and r = number of successes possible = $40 - 4 = 36$ (you have used up 4 of your guaranteed 40 wins). So we can now compute $P(X = 2)$ using the formula.

Multinomial Random Variables

This is like the binomial distribution, except that we have an trial with k possible outcomes (rather than two) performed repeatedly and independently n times. Think of the set of outcomes of each trial as $\{1, 2, \ldots, k\}$. At the end of the experiment, let us say we got 1 x_1 times, 2 x_2 times, and so on. Then, since we performed n trials, we must have

$$x_1 + x_2 + \cdots + x_k = n.$$

Thus, for the sample space, let us use the collection S of all n-tuples of outcomes chosen from $\{1, 2, \ldots, k\}$ (for example, a typical outcome in the experiment with $n = 3$ might be $(2, 2, 5)$. For our "random variable," we actually have many separate random variables:

$$\begin{aligned} X_1 &= \text{ Number of } 1'\text{s} \\ X_2 &= \text{ Number of } 2'\text{s} \\ &\ldots \end{aligned}$$

which we can think of as being combined into a single **vector random variable**:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$$

which assigns to each outcome the vector whose i^{th} coordinate is the number of times outcome i ha occurred. To obtain the probability of each of these vectors, we need to know the probabilities of the individual outcomes, which we can call p_i. Since the experiments are done independently, we can now use the binomial theorem to obtain the desired probability:

$$P(\boldsymbol{X} = (x_1, x_2, \ldots, x_n)) = \frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k},$$

which directly generalizes the binomial distribution formula. The fact that they equal 1 follows from the multinomial theorem;

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

where the sum is taken over all possible outcomes (sequences $(x_1, x_2, \ldots, x_k)$ of nonnegative integers that add up to n).

Exercise Set 6.

1. Use the binomial tables at the back of the book to compute the following probabilities for $p = 0.125$ and $n = 25$:
   - **a.** $P(X \geq 1)$
   - **b.** $P(X \leq 20)$
   - **c.** $P(5 \leq X \leq 20)$
2. According to a July 1999 article in The New York Times,[3] venture capitalists had this "rule of thumb": The probability that an Internet start-up company will be a "stock market success," resulting in

[3]Not All Hit It Rich in the Internet Gold Rush, *New York Times,* July 20, 1999, p. A1.

"spectacular profits for early investors" is .2. If you were a venture capitalist who invested in 10 Internet start-up companies, what was the probability that at least one of them would be a stock market success? (Round your answer to four decimal places.)

3. In March 2004 the US Agriculture Department announced plans to test approximately 243,000 slaughtered cows per year for mad cow disease (bovine spongiform encephalopathy).[4] When announcing the plan, the Agriculture Department stated that "by the laws of probability, that many tests should detect mad cow disease even if it is present in only 5 cows out of the 45 million in the nation." Test the Department's claim by computing the probability that, if only 5 out of 45 million cows had mad cow disease, at least one cow would test positive in a year (assuming the testing was done randomly). [Use the binomial distribution.]

4. On average, 5% of all hits on SlimShady.com by Macintosh users result in an order for beauty products, while 10% of the hits by Windows users result in orders. Due to on-line promotional efforts, your site traffic is approximately 10 hits per hour by Macintosh users, and 20 hits per hour by Windows users.
   - **a.** What is the probability that exactly 3 Windows users will order beauty products in the next hour?
   - **b.** What is the probability that at most 3 Windows users will order beauty products in the next hour?
   - **c.** What is the probability that exactly 1 Macintosh user and 3 Windows users will order beauty products in the next hour?

5. *(From the text)* Let X be a Poisson distributed random variable with parameter λ. Given that $P(X = 0) = 0.1$, compute the probability that $X > 5$.

6. The number of neutrons emitted from a sample of Uranium that breach a containment vessel in a millisecond is distributed according to a a Poisson distribution with parameter λ.
   - **a.** It is found that, on average, 100 particles breach the vessel each millisecond. Compute λ and the following probabilities:
     - **(i)** That exactly 120 particles breach the vessel in a given millisecond
     - **(ii)** That no particles breach it in a given millisecond.
     - **(iii)** That at least 50 particles breach the vessel in a given millisecond. [Use the tables on p. 520 – you might need to rescale the value of λ.]

7. *(From the text)* There are four blood types: A, O, B, and AB.Assume they occur with the following relative frequencies: 0.45, 0.40, 0.10, 0.05 respectively, and that 20 people are chosen independently and at random. Find the following probabilities:

[4]Source: *New York Times*, March 17, 2004, p. A19.

a. All 20 have the same blood type [No fancy formulas needed here]

b. Nine people have type O, eight type A, two type B, and one type AB.

7. Some Continuous Random Variables

Recall Definition 5.8: S is the real line, $X\colon S \to \mathbb{R}$ is the identity, and one has a nonnegative $f\colon I \to \mathbb{R}$ such that $P(B) = \int_B f(x)\,dx$ for Borel $B \subseteq I$.

Uniform
For a random variable to be uniform, the image of X is $X(S) = [\alpha, \beta]$, and

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{if } \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases}$$

Example 7.1. A freely rotating spinner is spun. Find the probability that
a. It lands on 15°
b. It lands somewhere between 15° and 18°.

Normal
Here, $X(S) = \mathbb{R}$ and

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big[-\frac{(x-\mu)^2}{2\sigma^2}\Big]$$

for a given $\mu \in \mathbb{R}$ and $\sigma > 0$. You have already proved (Exercise Set 5 # 3) that this is indeed a probability density function function. When $\mu = 0$ and $\sigma = 1$ we denote the resulting random variable by Z and refer to is as the **standard normal random variable.**

Example 7.2. SAT test scores are normally distributed with a mean of 500 and a standard deviation of 100. Find the probability that a randomly chosen test-taker will score 650 or higher.

Gamma
(This one will include two important special cases: The Chi-square distribution and the exponential distribution.) The gamma distributions are defined for $X(S) = (0, \infty)$ and have the form

$$f(x) = Ax^B e^{-Cx}$$

for suitable constants $A,\ B,\ C$. If you set $B = \alpha - 1$ and $C = 1/\beta$ (which you are perfectly entitled to do) and impose normalcy (see Exercise 4 in the homework set), you get A in terms of α and β as follows, giving you the following formula:

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha, \beta > 0$ are parameters of the distribution.

Question: What are these things $\Gamma(\alpha)$?
Answer: $\Gamma\colon (0,+\infty) \to \mathcal{R}$ is the **gamma function,** defined by

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

If you stare at both formulas, you will find it apparent that f is indeed a probability density function.

In order to derive some special cases (which are the ones in applications) we look at some properties of the gamma function:

Lemma 7.3. *Properties of the Gamma Function*

(1) $\Gamma(x) = (x-1)\Gamma(x-1)$ *for every* $x > 1$.
(2) If $n > 0$ *is an integer, then* $\Gamma(n) = (n-1)!$.
(3) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.
(4) More generally, $\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\sqrt{\pi}$, $\Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\frac{1}{2}\sqrt{\pi}$, *etc.*

You will prove this lemma in the homework.

Chi-square Distributions
These are just the gamma distributions with β set equal to 2 and $\alpha = r/2$ for a positive integer r. Thus,

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(r/2)2^{r/2}}x^{(r/2)-1}e^{-x/2} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

for $r > 0$ an integer, called the *number of degrees of freedom.* These distributions occur all over the place in statistics. When $r = 1$,

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}}x^{-1/2}e^{-x/2} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Exponential Distributions
These are just the gamma distributions with $\alpha = 1$ and $\beta = 1/\lambda$:

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

We shall see in the exercises that this distribution is intimately related to the discrete Poisson distribution. Exponential distributions also arise in scenarios like the following:

Question You are an investment analyst, and recent surveys show that 5% of all saving and loan (S&L) institutions fail each year. What is the probability that a randomly selected S&L will fail sometime within the next x years?

Answer To answer the question, suppose that you started with 100 S&Ls. Since they are failing (continuously) at a rate of 5% per year, the number left after x years is given by the decay equation:

$$\text{Number left } = 100e^{-0.05x},$$

Thus, the percentage that will have failed by that time—and hence the probability that we are asking for—is given by

$$P = \frac{100(1 - e^{-0.05x})}{100} = 1 - e^{-0.05x}.$$

Now let X be the number of years a randomly chosen S&L will take to fail. We have just calculated the probability that X is between 0 and x. In other words,

$$P(0 \leq X \leq x) = 1 - e^{-0.05x}.$$

But we also know that

$$P(0 \leq X \leq x) = \int_0^x f(t)\ dt$$

for a suitable probability density function f. The Fundamental Theorem of Calculus tells us that the derivative of this left side is $f(x)$. Thus,

$$f(x) = \frac{d}{dx}\int_0^x f(t)\ dt = 0.05e^{-0.05x},$$

an exponential probability density function.

Exercise Set 7.

1. Use the normal distribution tables to compute the following:
   - **a.** $P(-i \leq Z \leq i)$; $\quad i = 1, 2, 3$
   - **b.** $P(Z \geq 2.4)$
   - **c.** $P(Z \leq -2.4)$
   - **d.** $P(|Z| \geq 1.22)$
2. Prove that, if X is normally distributed with mean μ and standard deviation σ, then
   $$Z = \frac{X - \mu}{\sigma}$$
   is the standard normal random variable. [Hint: Compute $P(a \leq Z \leq b)$ directly from the formula for Z making an appropriate change of variables when doing the integral.]
3. The following information was gathered from student testing of a statistical software package called MODSTAT. Students were asked to complete certain tasks using the software, without any instructions. The results were as follows. (Assume that the time for each task is normally distributed.)

Task	Mean Time (min)	St. Dev. (min)
Task 1: Descriptive analysis	11.4	5.0
Task 2: Standardizing scores	11.9	9.0

a. Find the probability that a student will take at least 10 minutes to complete Task 1, and the probability that a student will take at least 10 minutes to complete Task 2.

b. Assuming that the time it takes a student to complete each task is independent of the others, find the probability that a student will take at least 10 minutes to complete each of Tasks 1 and 2.

c. It can be shown (see later) that, if X and Y are independent normal random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y respectively, then their sum $X+Y$ is also normally distributed and has mean $\mu = \mu_X + \mu_Y$ and standard deviation $\sigma = \sqrt{\sigma_X^2 + \sigma_Y^2}$. Assuming that the time it takes a student to complete each task is independent of the others, find the probability that a student will take at least 20 minutes to complete both Tasks 1 and 2.

4. (Derivation of the formula for the Gamma Distributions) Let $f\colon (0, +\infty) \to \mathbb{R}$ be defined by $f(x) = Ax^B e^{-Cx}$ for positive constants A, B, C. Define α and β by the equations $B = \alpha - 1$ and $C = 1/\beta$. Prove that, if f is a probability density function, then f must be the gamma distribution with parameters α and β.

5. Your media company's new television series "Avocado Comedy Hour" has been a complete flop, with viewership continuously declining at a rate of 30% per month. Use a suitable density function to calculate the probability that a randomly chosen viewer will be lost sometime in the next three months.

6. The half-life of Carbon-14 is 5,730 years. What is the probability that a randomly selected Carbon-14 atom will not yet have decayed in 4,000 years' time?

7. The probability that a "doomsday meteor" will hit the earth in any given year and release a billion megatons or more of energy is on the order of $0.000\,000\,01$.[5]

a. What is the probability that the earth will be hit by a doomsday meteor at least once during the 21st Century? (Use an exponential distribution with $a = 0.000\,000\,01$. Give the answer correct to 2 significant digits.)

b. What is the probability that the earth has been hit by a doomsday meteor at least once since the appearance of life (about 4 billion years ago)?

[5]Source: NASA International Near-Earth-Object Detection Workshop *New York Times*, January 25, 1994, p. C1.

8. Prove Lemma 7.3

9. Refer back to the scenario we discussed in the Poisson random variable: 150 people arrive at a car wash every hour, so that the number of people X arriving at your carwash is distributed according to the Poisson distribution: $P(X = x) = e^{-\lambda}\lambda^x/x!$ where $\lambda = 150$.

- **a.** What is the average number of cars arriving in the first y hours?
- **b.** Let Y be the time you need to wait for the first car to arrive. Use the answer to part (a) and the Poisson distribution to show that
$$P(Y > y) = e^{-\lambda y}$$
[Hint: The event that $Y > y$ is precisely the event that no cars arrive in the first y hours.]
- **c.** Hence compute $P(X \leq y)$ and use the Fundamental Theorem of Calculus to conclude that y an exponential distribution with probability density function $f(y) = \lambda e^{-\lambda y}$.

8. Distribution Functions

Definition 8.1. If X is a random variable, then define the associated **cumulative distribution function** $F : \mathbb{R} \to \mathbb{R}$ by

$$F(x) = P(X \leq x) = P_X((-\infty, x])$$

Proposition 8.2 (Properties of Cumulative Distributions)**.** *The cumulative distribution F of X satisfies:*

(a) $0 \leq F(x) \leq 1$
(b) F is monotone increasing.
(c) $F(x) \to 1$ *as* $x \to \infty$, *and* $F(x) \to 0$ *as* $x \to -\infty$.
(d) If $a \leq b$ *then* $P(a < x \leq b) = F(b) - F(a)$
(e) F is continuous from the right: If $x \searrow a$ *then* $F(x) \to a$.
(f) For every a, $F(a) - \lim_{x \to a^-} F(x) = P(X = a)$.

Proof.

(a) This follows from the definition of probability.
(b) This follows from monotonicity of probability measures (Proposition 2.3(3)).
(c) If (x_n) is any sequence of numbers increasing to ∞, then we can apply Proposition 2.5 to the monotone sequence $(-\infty, x_n]$ of events, which converges to $\mathbb{R}$, showing that $F(x_n) \to 1$.
(d) $F(b) - F(a) = P_X(-\infty, b] - P_X(-\infty, a] = P_X(a, b]$.
(e) If (x_n) is any sequence of numbers decreasing to x, we can again apply Proposition 2.5 to the monotone sequence $(-\infty, x_n]$, which has limit $(-\infty, x]$.
(f) One has

$$\begin{aligned} P(X = a) &= P_X(-\infty, a] - P_X(-\infty, a) \\ &= F(a) - \lim_{n \to \infty} P_X(-\infty, x_n] \text{ where } x_n \nearrow a \\ &= F(a) - \lim_{x \to a^-} F(x) = P(X = a). \end{aligned}$$

□

Examples 8.3.

A. *Continuous Probability Distributions* If X is continuous with associated probability density function f, then F is computed as

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

Notes 8.4.

(a) It follows from the fundamental theorem of calculus that we can recover the density function f as the derivative of the distribution function F. Put another way, if we know that the distribution function is continuously differentiable (everywhere except possibly at a discrete set of points), then there exists a density function

(continuous everywhere except possibly at a discrete set of points) for X given by the above formula.

(b) It follows further that, to prove that two random variables have the same density function at all but finitely many points, it suffices to prove that they have the same distribution function.

B. *Discrete Probability Distributions* S'pose that X is a discrete random variable with $X(S) = \{0, 1, 2, \dots\}$. In this case,

$$F(x) = P(X \le x) = \sum_{k=0}^{x} f(k),$$

where $f(k) = P(X = k)$ is the associated discrete density function. and it can be seen that F is a step function with

$$P(X = 0) = F(0)$$
$$P(X = k) = F(k) - \lim_{x \to k^-} F(x) = \text{ Jump of } F \text{ at } x = k$$

Here is a nice little application of distribution functions to obtain a theoretical result:

Lemma 8.5. *If X is the standard normal distribution (we say that X is $N(0,1)$-distributed) then X^2 is distributed as χ^2_1.*

Proof. By Remarks 8.4, it suffices to prove that X^2 and χ^2 have the same distribution function. The distribution for $Y = X^2$ is

$$\begin{aligned} F(y) = P(Y \le y) &= P(-\sqrt{y} \le X \le \sqrt{y}) && (y \ge 0) \\ &= 2 \cdot \frac{1}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-x^2/2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_0^{y} \frac{1}{\sqrt{t}} e^{-t/2}\, dt && (\text{Using } t = x^2) \end{aligned}$$

Therefore, the density function is

$$\frac{dF}{dy} = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2},$$

which agrees with the formula for the density function of χ^2_r with $r = 1$. When $y < 0$, $F(y) = P(X^2 \le y) = 0$, and so its derivative is also zero, giving us agreement with χ^2_1 once again. □

Vector Distributions

Often we will be interested in relating two or more random variables. The way to do that mathematically is through the use of vector random variables, such as $\boldsymbol{X} = (X_1, X_2)$.

Definition 8.6. If $\boldsymbol{X} = (X_1, X_2)$ a vector random variable, define the associated joint **distribution function** $F : \mathbb{R}^2 \to \mathbb{R}$ by

$$F(x_1, x_2) = P(X_1 \le x_1,\ X_2 \le x_2).$$

Notes 8.7.

(1) We can think of $P(X_1 \leq x_1,\ X_2 \leq x_2)$ as specifying a probability measure $P_{\boldsymbol{X}}$ on Borel sets in $\mathbb{R}^2$ by defining $P_{\boldsymbol{X}}([a,b] \times [c,d]) = P(X_1 \in [a,b],\ X_2 \in [c,d])$.

(2) If the X_i are continuous random variables, then we would like some analogue of the density function. The solution is inspired by the preceding point. A **joint probability density function** is a non-negative function $f:\ \mathbb{R}^2 \to \mathbb{R}$ satisfying

$$P(X_1 \in [a,b],\ X_2 \in [c,d]) = \int_{[a,b]} \int_{[c,d]} f(x_1, x_2) dx_2\, dx_1.$$

Thus,

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(t_1, t_2) dt_2\, dt_1.$$

The functions F and f are then related by

$$\frac{\partial^2}{\partial x_1\, \partial x_2} F(x_1, x_2) = f(x_1, x_2).$$

(3) In the discrete case—where, to simplify the notation, we assume that the discrete values of all random variables are integers; $D = \mathbb{Z}$—one has instead a (joint) probability function f with the property that

$$F(x_1, x_2) = \sum_{t_1=-\infty}^{x_1} \sum_{t_2=-\infty}^{x_2} f(t_1, t_2)$$

where $f(t_1, t_2) = P(X_1 = t_1,\ X_2 = t_2)$ is the associated discrete density function.

Question How are the density functions of the individual X_i related to the joint density function?
Answer In the continuous case, one has

$$P(X_1 \in [a,b]) = P(X_1 \in [a,b],\ X_2 \in \mathbb{R}) = \int_a^b \int_{-\infty}^{\infty} f(x_1, x_2)\ dx_2\, dx_1.$$

In other words, the inner integral,

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\ dx_2,$$

serves as a density function for X_1, so that it must agree with that density function, assuming we were given one to begin with. We call f_1 as defined above, and the similar one f_2, the **marginal probability density functions.**

In the discrete case,

$$P(X_1 = x_1) = P(X_1 = x_1,\ X_2 \in \mathbb{Z}) = \sum_{x_2=-\infty}^{\infty} f(x_1, x_2).$$

So, the associated discrete density functions are given by the above formula, meaning that

$$f_1(x_1) = \sum_{x_2=-\infty}^{\infty} f(x_1, x_2)$$

becomes the associated joint discrete density function.

Conditional Density Functions

Definition 8.8. S'pose $f(x_1, x_2)$ is a joint density function of some type, and assume that $a \in \mathbb{R}$ is such that $f_1(a) \neq 0$. Define the associated **conditional density function** by

$$f(x_2 \mid a) = \frac{f(a, x_2)}{f_1(a)}.$$

The conditional density function $f(x_1 \mid b)$ is defined similarly. That these are indeed density functions is shown in the exercises.
Question What do these have to do with ordinary conditional probability?
Answer Look at the discrete case first:

$$P(X_1 = x_1 \mid X_2 = x_2) = \frac{P(X_1 = x_1,\ X_2 = x_2)}{P(X_1 = x_1)} = \frac{f(x_1, x_2)}{f_1(x_1)},$$

so the conditional density function is exactly the density function for the conditional probability as shown. In the continuous case, we first need to interpret ordinary density functions in terms of probability itself: If h is small and positive, then $hf(x)$ is approximately $P(x \leq X \leq x + h)$. Likewise, $hf(x_1 \mid x_2)$ can be seen to be the approximately

$$P(x_1 \leq X_1 \leq x_1 + h \ \mid \ X_2 = x_2),$$

or, perhaps less disturbingly,

$$hf(x_1 \mid x_2) \approx P(x_1 \leq X_1 \leq x_1 + h \ \mid \ x_2 \leq X_2 \leq x_2 + k)$$

for any small k.

Exercise Set 8.

1. Suppose that X is a random variable with associated distribution function F. Determine the distribution functions of the following random variables in terms of F:
   - **a.** $-X$
   - **b.** X^2
   - **c.** $aX + b$ $(a, b$ constants$)$
2. A **logistic random variable** has associated distribution function $F : \mathbb{R} \to \mathbb{R}$ given by

$$F(x) = \frac{1}{1 + e^{-(at+b)}} \qquad (a > 0, b \text{ constants })$$

a. Show that F is indeed a distribution function and find the associated density function f.

b. Show that f and F are related by $f(x) = aF(x)[1 - F(x)]$.

3. Looking at the definitions of the marginal density functions in the continuous and discrete cases, define and obtain formulas for the associated marginal distribution functions.

4. Check that the conditional density functions are indeed density functions in both the discrete and continuous cases.

5. Suppose that $f_1(x),\ f_2(x), \ldots, f_n(x)$ are probability density functions. Show that their product is a joint probability density function.

9. Moments of a Random Variable

We first look at single (rather than vector-valued) random variables.

Definitions 9.1. If X is a discrete or continuous random variable with associated density function f and $n \geq 0$, then:

(1) The nth **moment of** X is given by

$$E[X]^n = \int x^n f(x)\ dx \qquad \text{(Continuous)}$$

$$E[X]^n = \sum x^n f(x) \qquad \text{(Discrete)}$$

where the sum or integral is taken over the target of X as usual. The corresponding **absolute moment** is defined as the moment of $|X|$ denoted by $E|X|$. Thus,

$$E|X|^n = \int |x|^n f(x)\ dx \qquad \text{(Continuous)}$$

$$E|X|^n = \sum |x|^n f(x) \qquad \text{(Discrete)}$$

In each of the special cases below, we have a corresponding absolute moment, but we won't mention it each time.

(2) The **mathematical expectation or mean of** X is defined as the first moment of X:

$$\mu(X) = EX = \int x f(x)\ dx \qquad \text{(Continuous)}$$

$$\mu(X) = EX = \sum x f(x) \qquad \text{(Discrete)}$$

(3) For an arbitrary constant c, the nth **moment of** X **about** c is given by

$$E(X-c)^n = \int (x-c)^n f(x)\ dx \qquad \text{(Continuous)}$$

$$E(X-c)^n = \sum (x-c)^n f(x) \qquad \text{(Discrete)}$$

(4) The nth **central moment of** X is given by its central moment about EX:

$$E(X-EX)^n = \int (X-EX)^n f(x)\ dx \qquad \text{(Continuous)}$$

$$E(X-EX)^n = \sum (X-EX)^n f(x) \qquad \text{(Discrete)}$$

(5) The second central moment is called the **variance** of X, and is denoted by $\sigma^2(X)$. Its square root, $\sigma(X)$, is called the **standard deviation of** X**.**

(6) In general, if $g\colon \mathbb{R} \to \mathbb{R}$ is any measurable function, then define the **expected value of** $g(X)$ by

$$E[g(X)] = \int g(x)f(x)\,dx \qquad \text{(Continuous)}$$

$$E[g(X)] = \sum g(x)f(x) \qquad \text{(Discrete)}$$

In mechanics, we think of the probability density function as the mass density function. Then, then mean and standard deviation of each coordinate function give, respectively, the coordinates of the center of mass and the moment of inertia.

The definition of EX gives the following general properties (which we will not bother to prove):

Proposition 9.2 (Properties of the Expected Value)**.**
Let X be a random variable and c constant. Then:

(1) $E(c) = c$
(2) $E(cX) = cE(X)$
(3) $E(X - EX) = 0$

The definition of $\sigma^2(X)$ gives the following general properties (which you will prove in the exercise set):

Proposition 9.3 (Properties of the Variance)**.**
Let X be a random variable and c constant. Then:

(1) $\sigma^2(c) = 0$
(2) $\sigma^2(cX) = c^2\sigma^2(X)$
(3) $\sigma^2(X + c) = \sigma^2(X)$
(4) $\sigma^2(X) = E(X^2) - (EX)^2$

Using Joint Probability Density Functions

S'pose that X and Y are random variables. What, one might ask, is the expected value of things like $X + Y$, XY, XY^2, and so on? The way to answer these questions is to think first of the vector random variable (X, Y), and its associated joint probability density function $f(x, y)$. For then we can use the following definition:

Definition 9.4. If X and Y are random variables, if (X, Y) with associated density function $f(x, y)$, and if $Z = K(X, Y)$, we define

$$E(Z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(x, y)f(x, y)dx\,dy$$

and

$$\sigma(Z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [K(x, y) - E(Z)]^2 f(x, y)dx\,dy,$$

and similarly for other moments.

Note that there is nothing special about simply using two random variables X, Y. We can equally well use n of them, and we shall do so later.

Examples 9.5.

A. S'pose that (X, Y) has joint density function $f(x, y) = \dfrac{3}{2\pi(x^2 + y^2)^{5/2}}$ defined on the region $x^2+y^2 \geq 1$. Compute the expected value of x^2+y^2.

B. S'pose that (X, Y) has joint density function $\dfrac{1}{\pi}e^{-(x^2+y^2)}$. Find the expected value of X^2Y^2. [Hint for doing the integral: Use Integration by parts.]

C. S'pose that (X, Y) has joint density function $f(x)g(y)$, where f and g are the marginal density functions. Then E(X+Y) = E(X)E(Y)

Exercise Set 9.

1. Prove Proposition 9.3.

2. **a.** Show that, for every random variable X and every constant c

$$E(X - c)^2 = E(X^2) - 2cE(X) + c^2.$$

Deduce that the second central moment about c is a minimum when $c = EX$.

b. Give an example to show that the first central absolute moment about c may be a minimum at values of c other than EX (which is a reason that we don't use this moment in the definition of the variance or standard deviation).

3. Obtain formulas for the expected value and variance of each of the following:[6]

a. A uniform random variable with range of values $[\alpha, \beta]$.

b. An exponential distribution with parameter λ. (Do this from scratch rather than as a special case of a gamma distribution.)

c. A normal distribution

d. A Gamma distribution

e. A Chi-square distribution

f. A binomial distribution

4. The function $f\colon \mathbb{R} \to \mathbb{R}$ is symmetric about c if $f(c - x) = f(c + x)$ for every x. Prove that, if X has a density function that is symmetric about c, then $EX = c$ and that all odd moments vanish (assuming they exist).

5. The following table shows the number of U.S. households at various income levels in 2000, based on a (fictitious) population of 100,000 households:[7]

[6]Most of the derivations appear in Section 5.2 of the textbook.

[7]Incomes are rounded mean income per quintile. Source: U.S. Census Bureau, Current Population Survey, selected March Supplements as collected in Money Income in the United States: 2000 (p. 60-213) September, 2001.

2000 Income (thousands)	$10	$25	$42	$66	$142
Households (thousands)	120	20	20	20	20

Compute the expected value and the standard deviation of the associated random variable X. If we define a "low income" family as one whose income is more than one standard deviation below the mean, and a "high income" family as one whose income is at least one standard deviation above the mean, what is the income gap between high-and low-income families in the U.S.? (Round your answers to the nearest $1000.)

6. A roulette wheel (of the kind used in the U.S.) has the numbers 1 through 36, 0 and 00.

a. A bet on a single number pays 35 to 1. This means that if you place a $1 bet on a single number and win (your number comes up), you get your $1 back plus $35 (that is, you gain $35). If you lose, you lose the $1 you bet. What is the expected gain from a $1 bet on a single number? What is the variance?

b. The numbers 0 and 00 are green, half the remaining 36 are red, and the other half black. A bet on red or black pays 2 to 1. What is the expected gain from a $M bet on red? What is the variance?

7. Suppose that X and Y are continuous random variables with probability density functions f and g respectively, and suppose that the joint density function of (X, Y) is $f(x)g(y)$.

a. How are $E(XY)$, $E(X)$, and $E(Y)$ related?

b. Obtain a formula for the *distribution* function of $X + Y$.

8. Give an example of two discrete random variables X and Y such that $E(X + Y) \neq E(X) + E(Y)$.

10. Some Inequalities

Theorem 10.1. *Let X be a random variable, let $g\colon \mathbb{R} \to \mathbb{R}$ be measurable, and let $c > 0$ be constant. Then*

$$P(g(X) \geq c) \leq \frac{E(g(X))}{c}.$$

Proof.

$$\begin{aligned} E[g(X)] &= \int_{-\infty}^{\infty} g(x)f(x)\ dx \\ &\geq \int_{\{\, x|g(x)\geq c\,\}} g(x)f(x)\ dx \\ &\geq c\int_{\{\, x|g(x)\geq c\,\}} f(x)\ dx \\ &= cP(g(X)\geq c) \quad (!) \end{aligned}$$

Justification of the last step: The probability of any Borel set is defined as the integral of the density function over that set. □

Corollary 10.2 (Markov's Inequality)**.** *Let X be a random variable with mean μ, and let $c > 0$. Then*

$$P(|X-\mu| \geq c) \leq \frac{E\,|X-\mu|^r}{c^r}$$

for every $r > 0$.

Proof. In the theorem, take $g(X) = (X-\mu)^r$ and observe that

$$P(|X-\mu| \geq c) = P(|X-\mu|^r \geq c^r).$$

□

Corollary 10.3 (Tchebichev's Inequality)**.** *Let X be a random variable with mean μ and standard deviation σ, and let $k > 0$. Then*

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Equivalently,

$$P(|X-\mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}.$$

Proof. In Markov's inequality, take $c = k\sigma$ and $r = 2$:

$$P(|X-\mu| \geq k\sigma) \leq \frac{E\,|X-\mu|^2}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

□

Theorem 10.4. *Cauchy-Schwartz Inequality*
Let X and Y be random variables with means μ_X, μ_Y and standard deviations σ_X, σ_Y. Then

$$\Big(E[(X-\mu_X)(Y-\mu_Y)]\Big)^2 \leq E(X-\mu_X)^2 \cdot E(Y-\mu_Y)^2 = \sigma_X^2\sigma_Y^2$$

with

$$E[(X-\mu_X)(Y-\mu_Y)] = \sigma_X\sigma_Y \Leftrightarrow P\Big[Y = \mu_Y + \frac{\sigma_Y}{\sigma_X}(X-\mu_X)\Big] = 1$$

and

$$E[(X-\mu_X)(Y-\mu_Y)] = -\sigma_X\sigma_Y \Leftrightarrow P\Big[Y = \mu_Y - \frac{\sigma_Y}{\sigma_X}(X-\mu_X)\Big] = 1$$

Proof. We start with a special case where $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$. In this special case,

$$\begin{aligned} 0 \le E[(X-Y)^2] &= E(X^2 - 2XY + Y^2) \\ &= EX^2 + EY^2 - 2E(XY) = 2 - 2E(XY), \end{aligned}$$

showing that $E(XY) \le 1$. Similarly, the fact that $E[(X+Y)^2] \ge 0$ gives us $E(XY) \ge -1$.

Now look at the extreme cases. If $E(XY) = 1$, then the first inequality shows that $E[(X-Y)^2] = 0$. Now, the only way that a non-negative random variable can have expectation zero is that it is zero with probability 1. Thus also $X - Y = 0$ with probability 1 (since that is true of its square.) In other words, $P(X = Y) = 1$. Similarly, if $E(XY) = -1$, then $P(X = -Y) = 1$. This proves the theorem for the special case.

For the general case, let $X' = (X-\mu_X)/\sigma_X$ and $Y' = (Y-\mu_Y)/\sigma_Y$. Then X' and Y' are in the special case, and so $E^2(X'Y') \le 1$. Substituting what they are and a little algebra now gives the general case. □

Exercise Set 10.

1. (The first step in computing the variance of the sum of two independent random variables) Use the equation in Proposition 9.3 to show that, if X and Y are random variables, then

$$\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) - 2[E(XY) - E(X)E(Y)].$$

2. Your newly launched company, CyberPromo, Inc., sells computer games on the Internet. Statistical research indicates that the lifespan of an Internet marketing company such as yours has an expected value of 30 months and standard deviation of 4 months.

(1) Complete the following sentence: There is (Multiple Choice: at least/at most/approximately) a _____% chance that CyberPromo will still be around for more than 38 months.

(2) How does your answer change if you know that the lifespans of Internet marketing companies are symmetrically distributed about the mean?

3. Let X be a χ^2_r be the Chi-square random variable with parameter r. You proved in an earlier exercise set that $E(X) = r$ and $\sigma^2(X) = 2r$. Using that information, take $r = 40$ and use Tchebichev's inequality to find a lower bound for $P(|X - 40| \le 20)$ and compare it to the actual value (Table 3 in Appendix III).

11. Independent Random Variables

S'pose you want to measure the average income of a household in the US. One way of approximating it is to repeatedly measure the income of randomly selected households, and then take the average. Let X be the random variable that assigns to each household the income. If we are sampling n households, then strictly speaking, we have n identical random variables: X_1 assigns to each household its income on at the time you survey the first household. X_2 gives the incomes at the time of second houehold is surveyed, and so on. Since the households are selected at random, the events $(X_1 \in [a,b])$, $(X_2 \in [a,b])$, etc. are independent, so we say that the *random variables* $X_1,\ X_2, \ldots, X_n$ are independent, leading to the following definition:

Definition 11.1. The random variables $X_1,\ X_2, \ldots, X_k$ are **independent** if for any Borel sets B_i, one has

$$P(X_i \in B_i;\ i = 1, \ldots, k) = \prod_{i=1}^{k} P(X_i \in B).$$

The following result tells us how the joint density function is related to the individual density functions in the case of independence.

Lemma 11.2. *The following statements are equivalent:*

(a) $X_1,\ X_2, \ldots, X_k$ *are independent*
(b) $F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = F_{X_1}(x_1) \ldots F_{X_k}(x_k)$
(c) $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = f_{X_1}(x_1) \ldots f_{X_k}(x_k)$

Proof.
(a) $\Rightarrow$ (b): Just take $B_i = (-\infty, x_i]$ in the definition of independence.
(b) $\Rightarrow$ (c): Start with (a) and take $\partial/\partial x_1 \ldots \partial x_k$ of both sides.
(c) $\Rightarrow$ (a): Integrating both sides repeatedly gives you (b), which gives independence of a certain class $B_i = (-\infty, x_i]$, of Borel sets. We can now use the Baby Rules of probability to get it true for arbitrary intervals, and thence to Borel sets with some manipulation. □

Corollary 11.3.

(1) If X and Y are independent, then $E(XY) = E(X)E(Y)$.
(2) More generally, if $X_1,\ X_2, \ldots, X_k$ are independent, then $E(X_1 X_2 \ldots X_n) = E(X_1)E(X_2) \ldots E(X_n)$.

A nice consequence of the definition of independence is the following:

Proposition 11.4. *If X and Y are independent, then so are $f(X)$ and $g(Y)$, where f and g are any continuous functions of a single variable.*

Proof. Just test for independence, noting that $P(f(X) \in B) = P(X \in f^{-1}(B)$ for any Borel set B, and similarly for $P(g(Y) \in B)$. □

Complex-Valued Random Variables

To talk properly about complex-valued random variables, we need to talk about Borel sets in $\mathbb{C}$. For this, refer back to Example 2.4(B), and we take as our Borel sets the σ-field generated by the collection of all closed rectangles.

A **complex-valued random variable** is a function $X\colon S \to \mathbb{C}$ such that the preimage of every Borel set in $\mathbb{C}$ is an event. A **probability density function for a complex-valued random variable** is then a function $f : \mathbb{C} \to \mathbb{R}$ whose integral over any region (= Borel set) R in $\mathbb{C}$ equals $P(X \in R)$. Think of a typical complex-valued random variable Z as a sum $X + iY$, where X and Y are ordinary real-valued random variables. In this way, we treat it in just the same way as a vector-valued random variable (X, Y). In fact: A **probability density function** for $Z = X+iY$ is defined as just a joint probability density function for (X, Y), and similarly for a **distribution function**.

Thus, the distribution function of a complex-valued random variable is the function $F\colon \mathbb{C} \to \mathbb{R}$ given by

$$F(z) = F(x + iy) = P(X \leq x, Y \leq y).$$

Definition 11.5. The complex-valued random variables $Z_1,\ Z_2, \ldots, Z_k$ are **independent** if for any Borel sets $B_i \subseteq \mathcal{C}$, one has

$$P(Z_i \in B_i;\ i = 1, \ldots, k) = \prod_{i=1}^{k} P(Z_i \in B).$$

The definition of independence, as well as the above results — see the exercises — all make perfect sense for complex random variables.

Exercise Set 11.

1. (From the textbook) Let X, Y, Z be continuous random variables with joint distribution function $f(x, y, z) = 8xyz\chi(I)$, where I is the unit cube $[0, 1]^3 \subset \mathbb{R}^3$, and $\chi(I)$ is its characteristic function. Prove that X, Y, and Z are independent, and compute $P(X_1 < X_2 < X_3)$. [Hint: this probability can be computed without actually performing an integral. Consider instead how many regions of the type $x < y < z$ are possible by using a simple counting argument.]
2. S'pose that X and Y are continuous random variables and that (X, Y) has density function the form $h(x, y) = f(x)g(y)$.
   - **a.** Show that the density functions of X and Y are, respectively, f and g. [Hint: Take a brief look at the material on joint distributions in Section 8.]
   - **b.** Deduce that X and Y are independent by just quoting a certain result.
   - **c.** S'pose that $f = g$ above. Prove that $P(X > y) = \frac{1}{2}$. [Hint: in evaluating the integral, consider the derivative of $[F(x)]^2$.]

3. Continuing with Exercise 1 in Section 10. Prove that, if X and Y are independent, then $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y)$. Note the significance of this result to #3(c) In Exercise Set 7.

4. If X and Y are independent, what can be said about the quantity $E[(X-\mu_X)(Y-\mu_Y)]$?

5. **a.** State and prove the complex versions of Lemma 11.2 and Corollary 11.3.

b. Hence prove the following generalization of Proposition 11.4: If Z_1 and Z_2 are independent complex-valued functions, then so are $f(Z_1)$ and $g(Z_2)$, where f and g are any continuous functions of a single complex variable.

6. Let X be a random variable, and consider the associated complex random variable $T(X) = e^{iX} = \cos X + i \sin X$. Show that, if X and Y are independent and $U = X+Y$, then $E[T(U)] = E[T(X)]E[T(Y)]$.

12. Covariance, Correlation, and Characteristic Functions

Covariance and Correlation

Definitions 12.1.

(1) If X and Y are two random variables with means μ_X and μ_Y respectively, then their **covariance** is

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)].$$

Notice that Cauchy-Schwartz tells us that $|\mathrm{Cov}(X,Y)| \leq \sigma_X\sigma_Y$. Think of this as a kind of "dot product". In fact, it is formally an inner product. (See the exercises.)

(2) If X and Y are two random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y respectively, then their **correlation coefficient** is

$$\rho(X,Y) = E\Big[\Big(\frac{X-\mu_X}{\sigma_X}\Big)\Big(\frac{Y-\mu_Y}{\sigma_Y}\Big)\Big] = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

This is, formally, a *normalized* inner product. (See the exercises.)

Cauchy-Schwartz now tells us that $|\rho(X,Y)| \leq 1$, and is equal to 1 iff X and Y are linearly related with probability 1. Thus, $\rho(X,Y)$ is a measure of the linear dependence between X and Y. They are **completely positively (resp. negatively) correlated** if $\rho(X,Y) = 1$ (resp. -1).

Characteristic Functions

Definition 12.2. Let X be a random variable (continuous or discrete. Then its **characteristic function** of **Fourier transform** $\phi_X : \mathbb{R} \to \mathbb{C}$ is defined by

$$\phi_X(t) = E[\cos(tX) + i\sin(tX)] = E[e^{itX}].$$

Equivalently, we can define it in terms of its density function:

$$\text{Continuous: } \phi_X(t) = \int_{-\infty}^{+\infty} e^{itx} f(x)\, dx = \int_{-\infty}^{+\infty} [\cos(tx) + i\sin(tx)] f(x)\, dx$$

$$\text{Discrete: } \phi_X(t) = \sum_{x\in D} e^{ix} f(x) = \sum_{x\in D} [\cos(tx) + i\sin(tx)] f(x)$$

Theorem 12.3 (Properties of the Characteristic Function)**.**

a. $\phi_X(0) = 1$
b. $|\phi_X(t)| \leq 1$
c. $\phi_{cX+d}(t) = e^{itd}\phi_X(ct)$ *if* c *and* d *are constants.*
d. $\dfrac{d^n}{dt^n}\phi_X(t)\,|_{t=0} = i^n E(X^n)$ $(n = 1, 2, \dots)$

e. *Inversion Formulas: If Φ_X is absolutely integrable on $[-\infty, +\infty]$, then*

$$\textit{Continuous: } f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{itx} \phi_X(t)\ dt$$

$$\textit{Discrete: } f(x_j) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} e^{itx_j} \phi_X(t)\ dt$$

Exercise Set 12.

1. If X is a random variable with mean ν_X, show that $\mathrm{Cov}(X, X) = \sigma_X^2$, and that $\rho(X, X) = 1$.

2. Show that, if X and Y are independent, then $\mathrm{Cov}(X, Y) = \rho(X, Y) = 0$. In words, *Independent random variables are completely uncorrelated.*

3. (A slight generalization of Exercise 3 in Section 11.) Prove that, if X and Y are completely uncorrelated (that is, $\rho(X, Y) = 0$), then $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$. (Again, note the significance of this result to #3(c) In Exercise Set 7.)

4. If $X_1, X_2, \ldots, X_n$ are independent random variables with characteristic functions ϕ_{X_i}, show that

$$\phi_{X_1 + \cdots + X_n}(t) = \prod_{i=1}^{n} \phi_{X_i}(t).$$

5. (Refer to Section 6.3 in the book.)

a. If Z is the standard normal distribution, show that $\phi_Z(t) = e^{-t^2/2}$.

b. If X is an arbitrary normal distribution, obtain a formula for X in terms of Z, and hence use Theorem 12.3c to obtain a formula for ϕ_X.

c. Obtain a formula for the characteristic function of the Chi-square distribution.

13. Applications of Characteristic Functions

Theorem 13.1 (The Sum of Independent Binomial Random Variables is Binomial)**.**
If $X_1, \ldots, X_n$ are independent discrete binomial random variables of type (n_i, p), then their sum X is also binomial, and of type (N, p), where $N = \sum_i n_i$.

Proof. By Exercise 4 in Section 12, the characteristic function of X is the product of the ϕ_{X_i}. So, the question is, what is the characteristic function of a binomial random variable of type (n_i, p)? Well,

$$\begin{aligned}
\phi_{X_i}(t) &= E(e^{itX_i}) \\
&= \sum_{k=0}^{n_i} P(X_i = k) e^{itk} && \text{Defn of Expected Value} \\
&= \sum_{k=0}^{n_i} \binom{n_i}{k} p^k q^{n_i - k} e^{itk} && \text{Here, } q = 1 - p \\
&= \sum_{k=0}^{n_i} \binom{n_i}{k} (pe^{it})^k q^{n_i - k} \\
&= (pe^{it} + q)^{n_i}
\end{aligned}$$

□

Now, to get the characteristic function of the sum, just multiply all these characteristic functions together to get $\phi_X(t) = (pe^{it} + q)^N$, as required.

Theorem 13.2 (Linear Combinations of Independent Normal Distributions)**.**
Let $X_1, \ldots, X_k$ be independent normal random variables where X_i has mean μ_i and standard deviation σ_i, and let $c_1, \ldots, c_k$ be constants. Then $X = \sum_{j=1}^{k} c_j X_j$ is also normal, with mean $\mu = \sum_j c_j \mu_i$ and variance $\sigma^2 = \sum_j c_j^2 \sigma_j^2$.

Proof.

$$\begin{aligned}
\phi_X(t) &= \phi_{\sum_j c_j X_j}(t) \\
&= \prod_j \phi_{c_j X_j}(t) \\
&= \prod_j \phi_{X_j}(c_j t) \\
&= \prod_j \exp\Big[ic_j t\mu_j - \frac{\sigma_j^2 c_j^2 t^2}{2}\Big] \\
&= \exp\Big[it\mu - \frac{\sigma^2 t^2}{2}\Big]
\end{aligned}$$

where σ and μ are as stated in the hypothesis. The result follows, since the last term is the characteristic function of the stated normal random variable. □

Corollary 13.3 (The Sampling Distribution of the Mean)**.**
If $X_1, \ldots, X_k$ are identical independent random variables with mean μ and standard deviation σ, then their mean, $\overline{X} = \frac{\sum_j X_j}{k}$ is normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Theorem 13.4 (Sum of Chi-Square Distributions)**.**
If $X_1, \ldots, X_k$ are independent random variables with X_j a $\chi^2_{r_j}$ random variable, then their sum X is of type χ^2_r, where $r = \sum_j r_j$.

Proof. From the exercises in the preceding section, we know that

$$\phi_{X_j}(t) = (1 - 2it)^{-r_j/2}.$$

Therefore,

$$\begin{aligned}\phi_X(t) &= \phi_{\sum_j X_j}(t) \\ &= \prod_j \phi_{X_j}(t) \\ &= \prod_j \Big[(1-2it)^{-r_j/2}\Big] \\ &= (1-2it)^{-r/2}\end{aligned}$$

where r is as stated in the hypothesis. □

In view of Lemma 8.5, we now have:

Corollary 13.5. *Let X_j be normal with mean μ_j and standard deviation σ_j. Then*

$$X = \sum_{j=1}^{k} \Big(\frac{X_j - \mu_j}{\sigma_j}\Big)^2$$

is a Chi-square random variable χ^2_k (that is, Chi-square with k degrees of freedom.

13.6. Application S'pose we need to ascertain whether the average capacity of our 10 oz sodas are really $\mu_1 = 10$ oz, and the same for the $\mu_2 = 5$ oz and $\mu_3 = 2$ oz. One can get a single statistic that measures the discrepancy by taking

$$\chi^2 = \Big(\frac{\overline{x}_1 - \mu_1}{\sigma_1/\sqrt{n_1}}\Big)^2 + \text{ Similar terms for the other two}$$

By the above results, this really is Chi-square, provided the distribution of soda bottle fills X_i are normal to begin with, which is tacitly assumed in the real world. (Actually, the Central Limit Theorem will justify this for us, when n is large.)

14. Central Limit Theorem

Definition 14.1. If $X_1, \ldots, X_n$ are random variables, then their **mean** is the random variable $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$.

Theorem 14.2 (Central Limit Theorem)**.**
S'pose that $X_1, \ldots, X_n$ are independent identical random variables with expected value μ and standard deviation σ. Let

$$G_n(x) = P\Big[\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le x\Big] = \text{ Distribution function of } \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt = \text{ Distribution function of } Z.$$

Then $G_n(x) \to N(x)$ as $n \to \infty$.

Note Convergence of the distribution functions does not necessarily imply convergence of the associated density functions (the associated derivatives). But the theorem is sufficient for all probability computations, since it is the distribution functions that are used in computing probability. Thus, for purposes of probability, the mean approaches the standard normal distribution.

Proof of the Central Limit Theorem. We prove instead that the characteristic function of $Y = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ approaches that of Z. The result we want will then follow by a theorem of P. Lèvy. Now, the characteristic function of Z was shown in the exercises to be $\phi(t) = e^{-t^2/2}$. For the given mean, write

$$Y = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Z_j,$$

where $Z_j = \dfrac{X_j - \mu}{\sigma}$. Since the Z_i are identical independent random variables, one has

$$\phi_{\sum Z_i}(t) = \phi_{Z_1}(t)^n.$$

and so

$$\phi_Y(t) = \phi_{\sum Z_i/\sqrt{n}}(t) = \phi_{\sum Z_i}(t/\sqrt{n}) = \phi_{Z_1}(t/\sqrt{n})^n.$$

Now expand ϕ_{Z_1} as a Taylor series about $t = 0$:

$$\phi_{Z_1}(x) = \phi_{Z_1}(0) + x\phi_{Z_1}{}'(0) + \frac{x^2}{2!}\phi_{Z_1}{}''(0) + o(x^2).$$

One has $x = t/\sqrt{n}$, and $\phi_{Z_1}(0) = 1$ (by Theorem 12.3, which also tells us that $\phi_{Z_1}{}'(0) = iE(Z_1) = i(0) = 0$, and $\phi_{Z_1}{}''(0) = i^2E(Z_1^2) = -1$ (since $\sigma = 1$ for the Z_i. Substituting in the Taylor series gives:

$$\phi_{Z_1}(t/\sqrt{n}) = 1 - \frac{t^2}{2n} + o\Big(\frac{t^2}{n}\Big),$$

whence

$$\phi_Y(t) = \phi_{Z_1}(t/\sqrt{n})^n = \Big[1 - \frac{t^2}{2n} + o\Big(\frac{t^2}{n}\Big)\Big]^n \to e^{-t^2/2}$$

as $n \to \infty$ (see the exercise set). Since this is the characteristic function of Z, we are done. □

Exercise Set 13.

1. The monthly revenue from the sale of skin creams at SlimShady.com is normally distributed with a mean of \$38,000 with a standard deviation of \$21,000. If an audit agency selects 64 months' figures at random, what is the probability that it finds a mean monthly revenue of more than \$21,050?

2. Prove that

$$\Big[1 - \frac{t^2}{2n} + o\Big(\frac{t^2}{n}\Big)\Big]^n \to e^{-t^2/2} \text{ as } n \to \infty.$$

[Hint: Put $h = 1/n$ and use L'Hospital.]