Math 37500 -LM (25988) - Lectures 01

Ethan Akin
Office: NAC 6/287
Phone: 650-5136
Email: ethanakin@earthlink.net

Fall, 2019

Contents

Introduction

- The book is *Introduction to Probability* by Joseph K. Blitzstein and Jessica Hwang (hereafter BH). Supplementary material, including solutions to marked exercises, is posted by the author on the site:

$$http://projects.iq.harvard.edu/stat110.net$$

- READ THE BOOK. Keep up with the homework.
- Ask questions.
- The course information sheet with the term's homework assignments is posted on my site:

$$http://math.sci.ccny.cuny.edu/peoplename = EthanAkin$$

- I will be posting there a pdf of the slides I am using here. The first half of the course already up. I will post the other half as we get to it.
- Office: NAC 6/287. Hours: Tuesday 12-1:50pm. Phone: 650-5136, Email: ethanakin@earthlink.net

Sample Space and Set Language, , Section 1.2

In probability we consider the probability that this happens as opposed to that. There are two aspects to consider.

- The "this" and the "that" are events, subsets of a sample space of all possible outcomes.
- The probability is an assignment numbers to the events.

The first uses set theory language following the chart in BH page 6. The sample space is the set of outcomes of an experiment or possible states of the world. The simplest case is the BH's Pebble World with the sample space finite. However, it can be infinite, like the set of integers or the set of reals (countable or uncountable).

Naive Probability and Counting

The simplest way to assign probabilities is the *naive definition*:

- The sample space S is finite.
- All of the outcomes are equally likely.

Then we define for an event $A \subset S$, $P_{naive}(A) = \frac{|A|}{|S|}$.

This requires us to count the number of outcomes in A and this gets us to the tricky problem of counting.

We start by considering why multiplication is commutative.
Why is $5 \cdot 3 = 3 \cdot 5$?
This says

$$5 \textit{ threes } (\text{added}) \;=\; 3 \textit{ fives } (\text{added}).$$

That is,

$$3 + 3 + 3 + 3 + 3 \;=\; 5 + 5 + 5.$$

To see the problem, notice that it doesn't work for repeated multiplication. That is,

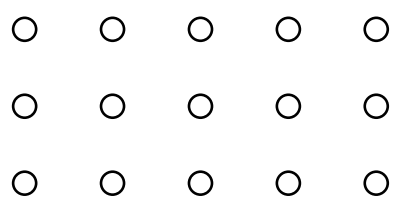
$$5 \textit{ threes} \text{ (multiplied)} \neq 3 \textit{ fives} \text{ (multiplied)}.$$

That is,

$$3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \neq 5 \cdot 5 \cdot 5,$$

because $3^5 = 243$ and $5^3 = 125$.

The answer is a picture:



The total of fifteen dots consists of 3 rows of five dots each or, equally, 5 columns of three dots each.

Our fundamental tool is the *Multiplication Rule*. Think of filling in k labeled blanks. For the first we have n_1 choices,

$$\underline{n_1} \quad \underline{n_2} \quad \dots \quad \underline{n_k}.$$

for the second n_2 choices and so on.
The total number of outcome possibilities resulting from these choices is the product $n_1 \cdot n_2 \cdot \dots \cdot n_k$.

We will say that there are k *selections* (those are the blanks), in which there are n_1 *choices* for the first, n_2 choices for the second and so on, leading to $n_1 \cdot n_2 \cdot \dots \cdot n_k$ alternative *possibilities* or *outcomes*.

While this is fundamental, the tricky part is to interpret your question so that the rule applies. It is simpler to think about when the separate selections are *independent*. This applies to the ice-cream cone examples of BH Example [1.4.3, page 8]1.4.5, page 10. The two selections are independent. The flavor choices are the same no matter what cone you pick and vice-versa.

There are three fundamental examples.

(BH Example [1.4.4]1.4.6: **Subsets**)Suppose you want to count how many subsets there are for a set A of size k. That is, $|A| = k$. For each of the k elements there is a selection with two choices, namely, whether or not to include the element. For example, if you choose $0 =$ "No" for every element, then you get the empty set. If you choose $1 =$ "Yes" for every element, then you get the whole set A. The k selections with 2 choices for each lead to 2^k possible subsets.

(BH Theorem [1.4.5]1.4.7: **Sampling WITH Replacement**) You make k selections - in order - taken from a set with n elements. After each choice is recorded, the object is replaced. So for each of the k selections there are n choices. Thus, there are n^k possible samples.

(**Functions from a set of size k to one of size n**)The key is to see that this is the same as sampling with replacement. For each of the k elements of the domain, we have n choices for the value of the function. Because different points of the domain can map to the same point of the range, the selections of the domain elements are independent. With k selections and n choices for each, again there are n^k possible functions.

Notice that we can think of the subset problem as counting functions from the set A to the two point set $\{0, 1\}$.

(BH Theorem [1.4.6]1.4.8: **Sampling WITHOUT Replacement**) You make k selections - in order - taken from a set with n elements. After each choice is recorded, the object is not replaced. Thus, we have lost independence. After the first selection with n choices is made, there are now only $n - 1$ choices for the second selection. Furthermore, which one are available depends upon the result of the first selection.

What is important is that the number of choices for the second selection is the same regardless of the result of the first selection. So we have

$$\underline{n} \quad \underline{n-1} \quad \ldots \quad \underline{n-k+1}.$$

We write the product $n \cdot n-1 \cdot n-2 \cdot \ldots n-k+1$ as n_k by analogy with n^k. Notice that if $k > n$, then this is 0 as it should be.

(**One-to-One Functions from a set of size k to one of size n**)This is the same as sampling without replacement. For the k elements of the domain, we have n choices for the value of the function at the first point, $n-1$ at the second point and so on. With k selections, there are n_k possible one-to-one functions.

This is what is happening with the Birthday problem, BH Example [1.4.8, page 11]1.4.10, page 12.

When $k = n$, $n_n = n!$, *n factorial*. Notice that $n_k = \frac{n!}{(n-k)!}$.

In this case we are counting the possible orderings of a set of size n. These are called *permutations* because in the case when the set is $\{1, 2, \ldots, n\}$ we think of them as re-arrangements of the standard order.

Adjusting for Overcounting

Start with a set of size n. We want to find the number of subsets of size k. Think of a collection of n people and we want to know how many ways we can choose a committee of size k. We know the number of lists of size k. These are just the samples of size k without replacement. There are n_k of these. For every subset of size k we have counted the $k!$ permutations of it separately. So the number of subsets is

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

Again for $k > n$ we have $\binom{n}{k} = 0$. Notice that $\binom{n}{n} = \binom{n}{0} = 1$. These are called the *binomial coefficients* because of:

(BH Example [1.4.17]1.4.19: **Binomial Theorem**)

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

Among the $1, 2, \ldots, n$ factors of $(x + y)$ we obtain a term $x^k y^{n-k}$ by choosing k x's and using y's from the remaining $n - k$ factors.

If we multiply out $(x_1 + x_2 + \cdots + x_\ell)^n$ then we obtain a term of the form $x_1^{k_1} \ldots x_\ell^{k_\ell}$ with $k_1 + k_2 + \cdots + k_\ell = n$. We call the coefficient of these term $\binom{n}{k_1 k_2 \ldots k_\ell}$. These are the *multinomial coefficients*.

Think of this as putting $\{1, \ldots, n\}$ into ℓ boxes. There are two methods we can use to count the number of ways that the first box contains k_1, the second k_2 and so forth.
For any ordering of $1, \ldots n$ of which there are $n!$, we put the first k_1 in box 1, the next k_2 in box 2, and so forth. Now we have to correct for overcounting by dividing by the number of rearrangements in each box.

Instead we can first choose a set of size k_1 for the first box. Then from the remaining $n - k_1$ choose a set of size k_2 for second. Then choose k_3 from $n - k_1 - k_2$.

Thus, we have

$$\binom{n}{k_1 k_2 \dots k_\ell} = \frac{n!}{k_1! \cdot \dots k_\ell!}$$
$$= \binom{n}{k_1} \cdot \binom{n-k_1}{k_2} \cdot \dots \binom{n-k_1-k_2 \cdots - k_{\ell-1}}{k_\ell}.$$

The identity given by the two alternatives is an example of what BH call a *story proof*.

This applies to the rearrangement of the letters in a word (BH Example [1.4.16]1.4.18).

(BH Example [1.4.20]1.4.22: **Bose-Einstein Problem**) Put n indistinguishable particles in k boxes.

First Case (At least one particle in every box): This is the number of solutions $(x_1, \ldots, x_k)$ of $x_1 + x_2 + \cdots + x_k = n$ with each x_i a positive integer. Think of the n dots in a row. We choose $k-1$ from among the $n-1$ spaces between the particles and place a wall in each. Thus, there are $\binom{n-1}{k-1} = \binom{n-1}{n-k}$ solutions.

Second Case (Some boxes may be empty): This is the number of solutions $(x_1, \ldots, x_k)$ of $x_1 + x_2 + \cdots + x_k = n$ with each x_i a non-negative integer. Following Ross, we see that if $(y_1, \ldots, y_k)$ is a positive solution of $y_1 + \ldots y_k = n + k$ then $(x_1, \ldots, x_k) = (y_1 - 1, \ldots, y_k - 1)$ is a non-negative solution and vice-versa. So from the First Case, the number of non-negative solutions is $\binom{n+k-1}{k-1} = \binom{n+k-1}{n}$.

Story Proofs, Section 1.5

On BH page [19]20 are three important examples of what the authors call *story proofs*.

Another is a version of BH Example 1.5.4 which generalizes it and which is used in problem [37]39.

Suppose we have $2n$ people consisting of n pairs, e.g. married couples. With $k \leq n$, how many ways can we choose a committee of size k with no pair on the committee?

The first way is to look at all possible lists of length k with no pair on the list. For the first selection there are $2n$ choices, but now for the second there are $2n-2$ instead of $2n-1$ possible choices. Continuing we obtain $2n \cdot (2n-2) \cdot \ldots (2n-2(k-1))$ lists. We correct for overcounting by dividing by $k!$.

Instead, from among the n couples we can choose k of them to get members of the committee. There are $\binom{n}{k}$ possible choices so far. Then, from each of the k selected couples, we choose one member of each. That is k selections with 2 choices each.

Comparing the two methods we see that

$$\frac{2n \cdot (2n-2) \cdot \ldots (2n-2(k-1))}{k!} = \binom{n}{k} \cdot 2^k.$$

But suppose with $2j \leq k$ you want exactly j couples on the committee. How do you count now?

First, choose j couples from n to be on the committee. From the remaining $n-j$ couples you want to fill the remaining $k-2j$ slots on the committee. Choose $k-2j$ couples from the $n-j$ and then choose one of the two members of each of these couples. What is the formula for the result?

General Definition of Probability, Section 1.6

The general definition of probability is an assignment to every event, a subspace of the sample space S, a non-negative number so that

- $P(\emptyset) = 0$, and $P(S) = 1$.
- If $A_1, A_2, \ldots$ is a finite or infinite sequence of disjoint events, then $P(\bigcup_i A_i) = \sum_i P(A_i)$.

Think of the probability of an event as the weight of the subset, or, actually the weight compared to the total weight of the sample space.

BH Theorem 1.6.2, page [21]23, lists properties of probability which follow from these axioms.

Conditional Probability, Sections 2.1, 2.2

The *conditional probability of A given B*, written $P(A|B)$ is defined by

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

Notice that this only makes sense if $P(B) > 0$.
As described on BH page [42]46, we now know, or have evidence, that B is true and so we can update our estimate of the probability of A.

We restrict to the portion B of the sample space and so A occurs if and only if $A \cap B$ occurs. On [pages 43 and 44] page 48, BH present the geometric and frequency interpretations.

(BH Example 2.2.5) Assuming one child is a girl (B) , what is the (conditional) probability that both are girls (A)? We can think of this two ways.

Since $A \subset B$, $P(A \cap B) = P(A)$ which is $\frac{1}{4}$, the probability of GG. $P(B) = 1 - P(B^c)$. The probability $P(B^c)$ that neither is a girl is also $\frac{1}{4}$, since this is the probability of BB. So

$$P(A|B) = \frac{1}{4} \div (1 - \frac{1}{4}) = \frac{1}{3}.$$

Alternatively, we can think of B as our new sample space. Instead of four equally likely possibilities BB, BG, GB, GG, B consists of three equally likely possibilities BG, GB, GG of which $A = GG$ is one.

Assuming that the elder is a girl (C), we still have $A \subset C$ and so $P(A \cap C) = P(A) = \frac{1}{4}$.

Now $P(C) = \frac{1}{2}$ and so

$$P(A|C) = \frac{1}{4} \div \frac{1}{2} = \frac{1}{2}.$$

When our sample space is C, it consists of two equally likely possibilities GB, GG of which $A = GG$ is one.

Properties of Conditional Probability, Section 2.3

The definition

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

is clearly equivalent to saying

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

While obvious, this is important enough to be labeled a theorem (BH Theorem 2.3.1)

(BH Theorem 2.3.6: **The Law of Total Probability (LOTP)**)If $A_1, \ldots, A_n$ is a parition of the sample space S and $P(A_i) > 0$ for every i, then

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

Bayes' Rule, Section 2.3

Of special importance is (BH Theorem 2.3.3: **Bayes' Rule**) For A, B with $P(A), P(B) > 0$,

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)}.$$

We consider (BH Examples 2.3.7, 2.39) on pages [50, 51]55, 56.

Conditional Probabilities as Probabilities, Section 2.4

It is convenient to follows the notation on BH page [55]61, and define $\tilde{P}(A) = P(A|E)$ with A varying and E fixed. As observed there, this allows us to think of successive conditional probabilities in two ways.

$$\tilde{P}(A|B) = \tilde{P}(A \cap B) \div \tilde{P}(B)$$

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

$$= P(A \cap B \cap E) \div P(B \cap E) \;=\; P(A|B \cap E).$$

Independence, Section 2.5

Two events A and B are *independent* when $P(A \cap B) = P(A) \cdot P(B)$. Notice that independence is a symmetric notion.

When $P(B) > 0$ we can divide by it and see that independence says $P(A|B) = P(A)$. That is, knowledge of B gives us no information about A. In this form, the symmetry is less obvious but independence is also equivalent to $P(B|A) = P(B)$ (if $P(A) > 0$).

With $P(E) > 0$, A and B are *conditionally independent with respect to* E when $\tilde{P}(A \cap B) = \tilde{P}(A) \cdot \tilde{P}(B)$. That is,

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

This in turn is equivalent to

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

To compare independence with conditional independence, we look at (BH Examples 2.5.9, 10, 11) on pages [58, 59]65, 66.

When independence fails, we prove the following, which comes up in Exercises 14 and 16.

Theorem Assume that $0 < P(A), P(B) < 1$. The following are equivalent.

- (1) $P(A \cap B) > P(A) \cdot P(B)$.
- (2) $P(A|B) > P(A)$.
- (3) $P(A|B) > P(A|B^c)$.
- (4) $P(A) > P(A|B^c)$.

Proof (1) $\Rightarrow$ (2): Divide by $P(B)$.

(2) $\Rightarrow$ (3): $P(A|B) > P(A|B)P(B) + P(A|B^c)P(B^c)$ and so $P(A|B)P(B^c) > P(A|B^c)P(B^c)$. Divide by $P(B^c)$.

(4) $\Rightarrow$ (3): $P(A|B)P(B) + P(A|B^c)P(B^c) > P(A|B^c)$ and so $P(A|B)P(B) > P(A|B^c)P(B)$. Divide by $P(B)$.

Each of these arguments is reversible.

Conditioning as a Tool, Section 2.7

(BH Example 2.7.1: **Monty Hall Problem**) pages [50, 51]69, 70.

(BH Example 2.7.2: **Branching Process**) pages [63, 64]71, 72. There is a subtlety here that the authors don't mention.

We obtain the equation

$$P(D) = 1 \cdot \frac{1}{3} + P(D) \cdot \frac{1}{3} + P(D)^2 \cdot \frac{1}{3},$$

and they note that $P(D) = 1$ solves this equation. It is important that it is the only solution.

Suppose that with probability $\frac{1}{3}$ the amoeba splits into a triplet instead of a pair. Now the equation is

$$P(D) = 1 \cdot \frac{1}{3} + P(D) \cdot \frac{1}{3} + P(D)^3 \cdot \frac{1}{3},$$

and again $P(D) = 1$ is a solution, but now there is another solution between 0 and 1 and this is in fact the probability that the line dies out.

(BH Example 2.7.3: **Gambler's Ruin**) pages [64, 65]72, 73. It is easier to think of gambler A playing against the house

with one dollar bets. Starting with $i > 0$ dollars, he plays until he reaches his goal of N dollars or until he is busted with 0 dollars. If p_i is the probability that he wins, we obtain the equation for $0 < i < N$:

$$p \cdot p_{i+1} - p_i + q \cdot p_{i-1} = 0.$$

Regarded as an equation for p_i with i unrestricted, we look for solutions of the form $p_i = x^i$ with $x \neq 0$. Pulling out the common factor of x^{i-1} and dividing it away we obtain the quadratic equation

$$p \cdot x^2 - x + q = 0.$$

This has roots 1 and q/p (because $q^2 - q + pq = 0$. So if $p \neq \frac{1}{2}$ we obtain two solutions $p_i = 1, p_i = (q/p)^i$. If $p = \frac{1}{2}$, then $p_i = i$ is a second solution.

Discrete Random Variables and their PMFs, Sections 3.1, 3.2

A *random variable* X is just a real-valued function defined on the sample space. What is "random" is the location of the point s of the sample space S. For each location s $X(s)$ is the value of the function X at s.

For the next two chapters we will only consider *discrete* random variables with either a finite list of values $a_1, a_2, \ldots, a_n$ or an infinite list of values $a_1, a_2, \ldots$. BH calls the set of numbers x such that $P(X = x) > 0$ the *support of* X . Observe the notation $\{X = x\}$ for the event which is the set $\{s \in S : X(s) = x\}$.

If the sample space is finite, then any random variable on it is discrete, but discrete random variables are important for all sample spaces. If A is an event, then the indicator r.v. of A, denoted I_A, is $= 1$ on A and $= 0$ elsewhere so that $A = \{I_A = 1\}$.

The *Probability Mass Function* (PMF) of a random variable X is a function $p_X : \mathbb{R} \to [0, 1]$ given by $p_X(x) = P(X = x)$ so that $p_X(x) = 0$ except at the values in the support of X.

(BH Example 3.2.5: **Sum of Two Dice**) on page [97]109 and (BH Example 3.2.6: **Chidren in U.S. Households**) on page [98]110 illustrate this.

In these case, we have an explicitly given sample space S and we think of the random variable as a measurement on S.

Bernoulli, Binomial Distributions, Section 3.3

As our study proceeds, the sample space fades into the background and we focus on r.v.'s defined on some unspecified sample space. Our important -named- r.v.'s are described by their PMF's.

A *Bernoulli* r.v. X takes on just the values 0 and 1. So with $p = P(X = 1)$ and $q = P(X = 0)$, $p + q = 1$. A Bernoulli r.v. is exactly the indicator r.v. of some subset of the sample space. We regard a Bernoulli r.v. as describing the outcome of a *Bernoulli trial* with success when $X = 1$. We write $X \sim Bern(p)$ when X is a Bernoulli r.v. with success probability p.

The r.v.'s $X_1, \ldots, X_n$ are independent when

$$P(X_1 = x_1, \ldots, X_n = x_n) \;=\; P(X_1 = x_1) \ldots P(X_n = x_n).$$

This says, for example, that knowledge of the values of $X_2, \ldots, X_n$ tells us nothing about X_1:

$$P(X_1 = x_1 | X_2 = x_2, \ldots X_n = x_n) \;=\; P(X_1 = x_1).$$

A *Binomial* r.v. $X \sim B(n, p)$ is a sum of n independent Bernoulli r.v.'s $Bern(p)$. That is, $X = I_1 + \cdots + I_n$. The list of r.v.'s $\{I_1, \ldots, I_n\}$ consists of *independent, identically distributed random variables*, i.i.d. for short.

The value $X = k$ occurs when there are k 1's and $n - k$ 0's among the I_i values. Such a list has probability $p^k q^{n-k}$ by independence. There are $\binom{n}{k}$ choices for placement of the k 1's among the n values. So the PMF for a $B(n, p)$ r.v. is given by

$$P(X = k) \;=\; \binom{n}{k} p^k q^{n-k}.$$

Hypergeometric Distribution, Section 3.4

If an urn contains w white balls and b black balls, then drawing a white ball is an event with indicator $I \sim Bern(w/(w+b))$. If we draw n balls -with replacement- then the number of white balls drawn is an r.v. $X \sim B(n, w/(w+b))$. With I_j the indicator of the j^{th} draw, X is the sum of the i.i.d.'s $I_1, \ldots, I_n$.

If the balls are drawn without replacement then $X = I_1 + \cdots + I_n$ and each I_j $Bern(w/(w+b))$. But this time the r.v.'s are not independent. Among the $w+b$ balls there are $\binom{w+b}{n}$ equally likely ways of choosing n balls. There are $\binom{w}{k}$ ways of choosing k white balls and $\binom{b}{n-k}$ ways of choosing the remaining black balls, assuming $0 \leq k \leq w, 0 \leq n-k \leq b$. So

$$P(X = k) = \frac{\binom{w}{k} \cdot \binom{b}{n-k}}{\binom{w+b}{n}}.$$

In that case we write $X \sim HGeom(w, b, n)$

(BH Example 3.4.3: **Elk Recapture**) on page [104]117.

(BH Theorem 3.4.5) is an interesting symmetry result.

Discrete Uniform Distribution, Section 3.5

This is just BH's pebble space with a finite set of numbers, each equally likely. If C is the support, the set of values, then for $x \in C$, $P(X = x) = \frac{1}{|C|}$.

(BH Example 3.5.2: **Random Slips of Paper**) on pages [106,107]119, 120 provides a good summary of the different distributions.

Cumulative Distribution Function, Section 3.6

Recall that for a r.v. X the PMF is given by $p_X(x) = P(X = x)$. . The *Cumulative Distribution Function* CDF, F_X is defined by $F_X(x) = P(X \leq x)$.
These become more important for continuous r.v's later, but even for discrete r.v's it is sometimes easier to compute the CDF.
Let $X_1, \ldots, X_n$ be independent r.v.'s and defined $Y = max(X_1, \ldots, X_n)$. Observe that $\{Y \leq x\} = \{X_1 \leq x\} \cap \cdots \cap \{X_n \leq x\}$ So it follows that

$$F_Y(x) \;=\; F_{X_1}(x) \cdot \ldots F_{X_n}(x).$$

Expectation, Sections 4.1, 4.2

The *expectation* (or *expected value* or *mean*) of a r.v. X is its weighted average:

$$E(X) = \sum_x x \cdot P(X = x) = \sum_x x \cdot p_X(x).$$

Observe that we only require the PMF of X to compute its expectation. The expectation is undefined when the series $\sum_x |x| \cdot P(X = x)$ diverges

On the other hand, in the background is the sample space S. If S is finite (or countably infinite), then $P(X = x) = \sum\{P(s) : X(s) = x\}$. So we have

$$E(X) \;=\; \sum_x x\cdot P(X=x) \;=\; \sum_x x\cdot(\sum\{P(s) : X(s)=x\})$$

$$=\; \sum_x\sum\{X(s)P(s) : X(s)=x\} \;=\; \sum\{X(s)P(s) : s\in S\}.$$

That is, we have shown that

$$E(X) \;=\; \sum_s X(s)\cdot P(s).$$

The most important, and convenient, property of expectation is *linearity* :

$$E(X+Y) \;=\; E(X)+E(Y), \qquad E(cX) \;=\; cE(X).$$

These properties are easy to see when we use the sample space formula. The r.v.'s X and Y are defined on some common sample space S. We have

$$E(X) = \sum_s X(s) \cdot P(s), \qquad E(Y) = \sum_s Y(s) \cdot P(s).$$

and so

$$E(X) + E(Y) = \sum_s X(s) + Y(s) \cdot P(s) = E(X + Y).$$

From linearity we obtain (BH Proposition 4.2.4: **Monotonicity**) page [144]157: If $X \geq Y$ then $X - Y \geq 0$ and so $E(X) - E(Y) = E(X - Y) \geq 0$. So $E(X) \geq E(Y)$.

If X and Y are independent and have expectations then $E(X \cdot Y) = E(X) \cdot E(Y)$. Proof:

$$E(XY) = \sum X(s)Y(s)P(s) =$$

$$\sum_{x,y} xy \sum \{P(s) : X(s) = x, Y(s) = y\}$$

$$= \sum_{x,y} xyP(X = x, Y = y) = \sum_{x,y} xyP(X = x)P(Y = y)$$

$$\sum_{x} xP(X = x) \cdot \sum_{y} yP(Y = y) = E(X) \cdot E(Y).$$

If $I \sim Bern(p)$ so that $P(I = 1) = p, P(I = 0) = q = 1 - p$, then $E(I) = p$.

If $X \sim B(n, p)$ then $X = I_1 + \ldots I_n$ is the sum of n independent Bernoulli $Bern(p)$'s and so by linearity $E(X) = np$.

If $X \sim HGeom(w, b, n)$ then again $X = I_1 + \ldots I_n$ is the sum of n Bernoulli $Bern(w/(w + b))$'s. This time the terms are not independent, but independence is not required for linearity. So $E(X) = nw/(w + b)$.

Geometric and Negative Binomial Distributions, Section 4.3

For a sequence of independent Bernoulli trials with success probability p, the *Geometric Distribution*, $Geom(p)$ is the number of failures up to, but not including, the first success. Thus, for $k = 0, 1, \ldots$ $p_X(k) = q^k p$ with $q = 1 - p$. The *First Success distribution* $FS(p)$ is the distribution of $Y = X + 1$.

The calculus argument of (BH Example [4.3.5]4.3.6) on page [146]159 shows that $E(X) = \frac{q}{p}$ and so $E(Y) = \frac{q}{p} + 1 = \frac{1}{p}$.

The *Negative Binomial distribution* $X \sim NBin(r, p)$ counts the number of failures up to the r^{th} success. So X is the sum $K_1 + \cdots + K_r$ of r independent r.v.'s with each $K_j \sim Geom(p)$. So $E(X) = r \cdot \frac{q}{p}$.

(BH Theorem [4.3.9]4.3.10) on page [147]161 computes the PMF for the negative binomial $p_X(n) = \binom{n+r-1}{r-1} p^r q^n$.

Indicator Functions and the Fundamental Bridge, Section 4.4

The indicator function I_A of an event A is $= 1$ when A occurs, i.e. $s \in A$ and $= 0$ otherwise. Since $I_A \sim Bern(P(A))$ the expectation satisfies $E(I_A) = P(A)$.

(BH Theorem 4.4.1) on page [151]164 lists the properties which relate the events to their indicators.

(BH Example 4.4.3) on page [152]165 shows to see the *Inclusion-Exclusion Identity* without counting.

Functions of Random Variables and LOTUS, Section 4.5

If g is a real-valued function of a real variable and X is an r.v. then we can define a new r.v. by $Y = g(X)$. For its PMF p_Y we see

$$P(Y = y) = \sum\{P(s) : g(X(s)) = y\}$$

$$= \sum\{P(s) : X(s) = x \ \& \ g(x)) = y\}$$

$$= \sum_{g(x)=y} \sum_{X(s)=x} P(s) = \sum\{P(X = x) : g(x) = y\}.$$

For the expected value, we do a similar computation, following BH page 157 to get what BH call the *Law of the Unconscious Statistician* LOTUS:

$$E(g(X)) = \sum_x g(x)P(X = x).$$

If X, Y are independent r.v.'s and g, h are real-valued functions, then $g(X), h(Y)$ are independent.

$$P(g(X) = u, h(Y) = v) = \sum\{P(X = x, Y = y) : g(x) = u, h(y)$$

$$= \sum\{P(X = x)P(Y = y) : g(x) = u, h(y) = v\} = P(g(X) = u)$$

In particular, if X and Y are independent, then $X - \mu_X$ and $Y - \mu_Y$ are independent as well, where $\mu_X = E(X), \mu_Y = E(Y)$.

Variance, Section 4.6

A measure of the spread of a r.v. X is the *variance* $Var(X) = E((X - E(X))^2)$. Equivalently, $Var(X) = E(X^2) - E(X)^2$. If $\mu = E(X)$ then $(X - \mu)^2 = X^2 - 2\mu X + \mu^2$ and the result follows from linearity of expectation.

$Var(X) \geq 0$ and is $= 0$ only if X is constant.
$Var(X + c) = Var(X)$ and $Var(cX) = c^2 Var(X)$.

With the function $g(x) = x^2$, the fact that the variance is positive except when X is a constant is an example of the fact that $E(g(X))$ is not in general equal to $g(E(X))$.

Given X and Y with means μ_X, μ_Y, if X and Y are independent, then $X - \mu_X, Y - \mu_Y$ are independent and $E(X + Y) = \mu_X + \mu_Y$. So

$$Var(X + Y) = E((X + Y - \mu_X - \mu_Y)^2) =$$

$$E((X-\mu_X)^2+(Y-\mu_Y)^2+2(X-\mu_X)(Y-\mu_Y) = Var(X)+Var(Y),$$

because $E((X - \mu_X)(Y - \mu_Y)) = E(X - \mu_X)E(Y - \mu_Y) = 0$ by independence.

For $I \sim Bern(p)$, $E(I^2) = E(I) = p$ and so $Var(I) = p - p^2 = pq$.

If $X \sim Bin(n, p)$ then it is the sum of n independent $Ber(p)$'s. So $Var(X) = npq$.

(BH Example 4.6.4) on pages [159,160]173, 174 use a calculus argument again to compute for $X \sim Geom(p)$, $Var(X) = \frac{q}{p^2}$.

As it is the sum of r independent $Geom(p)$'s we have that for $X \sim NBin(r, p)$, $Var(X) = r\frac{q}{p^2}$.

Poisson, Section 4.7

An r.v. X with non-negative integer values is *Poisson* $\sim Pois(\lambda)$ when $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ (Recall that $0! = 1$ by convention).

(BH Example 4.7.2) on pages [161, 162]175, 176 use a calculus argument again to compute for $X \sim Pois(\lambda)$ that

$$E(X) = \lambda, \qquad Var(X) = \lambda.$$

(BH Theorems 4.8.1, 4.8.2, 4.8.3) on pages [166, 167]181, 182 describe the crucial properties of the Poisson distribution.

Probability Density Functions, Section 5.1

For continuous r.v.'s we let the sample space slide completely into the background and concentrate upon the distributions.

An r.v. X has a *continuous distribution* when its CDF F_X is piecewise differentiable. The function F_X is continuous (no jumps) and the line $\mathbb{R}$ is cut into a finite number of closed intervals on each of which F_X has a continuous derivative f_X which we call the *probability density function* PDF of X. At the cut points, f_X may have a jump, but since we compute probabilities by integrating the value at the jump does not matter. In our examples, the cut points will be at the ends of the support interval for X.

Think of the density as the weight per unit length so that $f_X(x)dx$ is the weight, or probability, of a little interval containing x of length dx. We will write $[x + dx]$ for an interval containing x and of length dx, e.g. $[x, x + dx]$.

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

In particular, the probability $P(X = t) = 0$ for any point t and so $P(a < X \leq b) = P(a \leq X \leq b) = P(a < X < b)$.

A PDF is non-negative and has integral $\int_{-\infty}^{+\infty} f_X(x)dx = 1$.

Thus, if we start with the CDF F_X we obtain the PDF by differentiating:

$$f_X(x) \;=\; F_X'(x),$$

and if we start with the PDF we get the CDF by integrating

$$F_X(x) \;=\; \int_{-\infty}^{x} f_X(t)dt.$$

(Location-Scale Transformation) If $Y = a + mX$ then

$$F_Y(y) = P(Y \leq y) = P(X \leq \frac{y-a}{m}) = F_X(\frac{y-a}{m})$$

and so, differentiating, we have $f_Y(y) = \frac{1}{m} f_X(\frac{y-a}{m})$.

For a continuous r.v. X with density $f_X(x)dx$ the expectation is given by

$$E(X) \;=\; \mu_X \;=\; \int_{-\infty}^{\infty} x f_X(x) dx.$$

In general, for a function $g(X)$ of X, the LOTUS result from discrete r.v.'s becomes

$$E(g(X)) \;=\; \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

In particular,

$$E(X^2) \;=\; \int_{-\infty}^{\infty} x^2 f_X(x)dx,$$

with the variance given by

$$Var(X) \;=\; E((X-\mu_X)^2) \;=\; E(X^2) - E(X)^2.$$

Uniform Distributions, Section 5.2

With $a < b$ the uniform distribution $U \sim U(a, b)$ has constant density with a linear CDF on (a, b) BH page [201]220.

Most important is $U(0, 1)$ with $f_U(x) = 1$ for $x \in (0, 1)$ and so

$$P(U \leq u) \;=\; F_U(u) = u, \quad \text{for } 0 < u < 1.$$

$E(U) = \int_0^1 x dx = \frac{1}{2}, E(U^2) = \int_0^1 x^2 dx = \frac{1}{3}$ and so $Var(U) = \frac{1}{6}$.

(Location-Scale Transformation) With $U \sim U(0, 1)$, $X = a + (b - a)U$ has $X \sim U(a, b)$, So

$$E(X) = \frac{b - a}{2}, Var(X) = \frac{(b - a)^2}{6}.$$

Universality of the Uniform Distribution, Section 5.3

Let X be a continuous r.v. support on (a, b). That is, X has values in (a, b) (with probability 1) and $F = F_X$ is a strictly increasing function from (a, b) to $(0, 1)$. So it has an inverse function F^{-1} from $(0, 1)$ to (a, b). Let $U \sim U(0, 1)$. Define $\hat{X} = F^{-1}(U)$ and $\hat{U} = F(X)$.

$$F_{\hat{X}}(x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

$$F_{\hat{U}}(u) = P(F(X) \leq u) = P(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u.$$

Thus, $\hat{X}$ and X have the same distribution and $\hat{U}$ and U have the same distribution.

So we can use U to obtain an r.v. with CDF F provided F on (a, b) is strictly increasing. In fact, it will work even if F levels off and so the support of X has gaps.

This will work as well even if X is a discrete r.v. Now F^{-1} is not a function but has a finite or a sequence of vertical lines, but the probability that U lands exactly on one of these points is 0.

Normal Distributions, Sections 5.4

The standard *normal distribution* r.v. $Z \sim \mathcal{N}(0,1)$ has PDF on $\mathbb{R}$ given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

The important properties of the normal are given on BH pages [212-216]232-236.

Notice that if $X \sim \mathcal{N}(\mu, \sigma^2)$ we convert to the standard normal by $Z = \frac{X-\mu}{\sigma}$ or $X = \sigma Z + \mu$ so that

$$f_X(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{x-\mu}{\sigma})^2/2}.$$

Exponential Distributions, Section 5.5

The *exponential distribution* $X \sim Expo(\lambda)$ with parameter λ has distribution $f_X(x) = \lambda e^{-\lambda x}$ and so with CDF $1 - e^{-\lambda x}$.

As is shown on BH page [218]230, if $X \sim Expo(1)$ then $Y = \frac{X}{\lambda} \sim Expo(\lambda)$ and so $E(Y) = Var(Y) = \frac{1}{\lambda}$. Why $\frac{1}{\lambda}$ instead of λ? Think of Y as the time to failure of some appliance. The parameter λ is the rate that it is being used up. The faster it is used up, i.e. the greater the rate λ, the shorter is the average time to failure $E(Y)$. In fact, it is exactly inverse to the rate.

Memoryless Property of the Exponential Distribution

Call $G_X(t) = P(X \geq t) = 1 - F_X(t)$ the survival function. Notice that $\{X \geq s + t\} \cap \{X \geq s\} = \{X \geq s + t\}$ when s and t are positive. Hence,
$P(X \geq s + t | X \geq s) = G(s + t)/G(s)$.

If $X \sim Expo(\lambda)$, then $G(t) = e^{-\lambda t}$. So in that case,

$$P(X \geq s+t|X \geq s) = e^{-\lambda(t+s)}/e^{-\lambda s} = e^{-\lambda t} = G(t) = P(X \geq t).$$

Conversely, if $G(s + t) = G(s)G(t)$, then, following BH page 220, $G'(t) = G'(0)G(t)$ and so with $\lambda = G'(0)$, $G(t) = e^{-\lambda t}$. Thus, a continuous r.v. which is memoryless is exponential.

Memoryless Property of the Geometric Distribution

Recall that for a sequence of Bernoulli trials with success probability p, $X \sim Geom(p)$ is the number of failures before the first success. So

$$G(k) = P(X \geq k) = q^k$$

Because $X \geq k$ when the first k trials are failures.

So

$$P(X \geq k + j | X \geq j) = q^{k+j}/q^j = q^k = G(k) = P(X \geq k).$$

Conversely, if $G(k + j) = G(k)G(j)$, then $G(k) = G(1)^k$ and so with $q = G(1)$, $G(k) = q^k$. Thus, a discrete r.v. with support the non-negative integers which is memoryless is geometric.

Poisson Process, Section 5.6

For the sequence of Bernoulli trials the number of successes in the first n trials is binomial $\sim Bin(n, p)$. The time up to the first success, that is, the number of failures before the first success is $\sim Geom(p)$. The number of failures before the r^{th} success is the negative binomial $NBin(r, p)$ and it is the sum of r independent geometric r.v.'s $\sim Geom(p)$.

The Poisson process is the continuous time analogue of the sequence of Bernoulli trials. With rate parameter λ, in the interval $(0, t)$ we assume that the number N_t of successes is $\sim Pois(\lambda t)$. If T_r is the time to the r^{th} success, then

$$T_r > t \quad \Longleftrightarrow \quad N_t < r.$$

In particular, $T_1 > 1$ is equivalent to $N_t = 0$ which has probability $e^{-\lambda t}$. Thus, $T_1 \sim Expo(\lambda)$.

T_r is the sum of r independent exponential r.v.'s$\sim$ $Expo(\lambda)$. This is a Gamma distribution which we will study in Chapter 8. It plays the role here that the negative binomial did in the discrete case.

(BH Example 5.6.3) on page [224]246 shows that the minimum of a list of independent exponentials is an exponential.