MTH 224: Probability and Statistics **Fall 2017**
Course Notes 2 **Drew Armstrong**

Oct 17

We have finished covering the basics of probability. The next section of the course is all about "random variables." What is a random variable?

To motivate the definition let's discuss a hypothetical scenario. Suppose a student's grade in a certain class is based on their letter grades on three in-class exams, and suppose that this student received the following grades:

	Exam 1	Exam 2	Exam 3
Grade	A	B-	A-

Question: If the three exams are weighted equally, what letter grade does this student receive in the course?

Joke Answer: Since each exam is worth 1/3 of the final grade we obtain

$$\text{Final Grade} = \frac{1}{3}\cdot \text{“A”} + \frac{1}{3}\cdot \text{“B-”} + \frac{1}{3}\cdot \text{“A-”} = \frac{\text{“A”} + \text{“B-”} + \text{“A-”}}{3}.$$

///

Of course you see that this is nonsense. It is meaningless to take the average of the three symbols "A," "B-," and "A-" because these three symbols are **not numbers**. In order to compute the final grade with this method we would need to have some recipe for converting letter grades into numbers and then converting numbers back into letters. The following table shows one possible way to do this (called the Grade Point Average):

Letter Grade	GPA
A	4.00
A-	3.67
B+	3.33
B	3.00
B-	2.67
etc.	etc.

Under this scheme our student's final GPA is

$$\frac{4.00 + 2.67 + 3.67}{3} \approx 3.45,$$

which I guess translates to a high B+.[1] Thus we see that for some purposes it is necessary to convert the possible outcomes of an experiment into numbers.

Definition of Random Variable. Consider an experiment with a sample space S of possible outcomes. These outcomes can take any form, such as strings of Hs and Ts or brands of cat food. A *random variable* is any function X that turns outcomes into real numbers:

$$X : S \to \mathbb{R}$$

///

For example, suppose that we flip a coin three times and record the number of Hs that we get. The sample space of this experiment is

$$S = \{TTT, HTT, THT, TTH, HHT, HTH, THH, HHH\}.$$

The random variable under consideration is $X =$ "number of Hs," which we can think of as a function $X : S \to \mathbb{R}$ that takes in a string $s \in S$ and spits out the number $X(s)$ of Hs it contains. Here is a table:

outcome $s \in S$	TTT	HTT	THT	TTH	HHT	HTH	THH	HHH
$X(s)$	0	1	1	1	2	2	2	3

We will use the notation $S_X \subseteq \mathbb{R}$ for the set of all possible values that the random variable can take X. The textbook calls this the *space* of the random variable X. For example, the space of our random variable $X =$ "number of heads" is

$$S_X = \{0, 1, 2, 3\}.$$

Warning: The textbook often writes S instead of S_X which I find very confusing. The space S of all possible outcomes of the experiment and the space S_X of all possible values of the random variable X are **not the same set**, except in very rare cases.

For each possible value $k \in S_X$ we consider the event "$X = k$"$\subseteq S$ which is the set of all outcomes $s \in S$ that satisfy $X(s) = k$. In our example, we have four possible values:

$$\begin{aligned}
&\text{"}X = 0\text{"} = \{TTT\},\\
&\text{"}X = 1\text{"} = \{HTT, THT, TTH\},\\
&\text{"}X = 2\text{"} = \{HHT, HTH, THH\},\\
&\text{"}X = 3\text{"} = \{HHH\}.
\end{aligned}$$

[1] Let me emphasize that I do not use any such scheme in my teaching. Instead, I keep all scores in numerical form through the semester and only convert to letter grades at the very end. I sometimes estimate grade ranges for individual exams, but these can only be approximations.

If the coin is fair then each of the 8 possible outcomes is equally likely, and we obtain the probabilities:

$$P(X=0)=\frac{1}{8},\quad P(X=1)=\frac{3}{8},\quad P(X=2)=\frac{3}{8},\quad P(X=3)=\frac{1}{8}.$$

Since the events "$X=k$" partition the sample space (i.e., these events are exhaustive and mutually exclusive), we observe that their probabilities must add up to 1:

$$\begin{aligned} S &= \text{“}X=0\text{”} \sqcup \text{“}X=1\text{”} \sqcup \text{“}X=2\text{”} \sqcup \text{“}X=3\text{”} \\ P(S) &= P(\text{“}X=0\text{”} \sqcup \text{“}X=1\text{”} \sqcup \text{“}X=2\text{”} \sqcup \text{“}X=3\text{”}) \\ 1 &= P(X=0)+P(X=1)+P(X=2)+P(X=3). \end{aligned}$$

In general, the function that takes in a possible value $k \in S_X$ and spits out the probability $P(X=k)$ is called the "frequency function" or the "probability mass function" (pmf) of the random variable X.

Definition of Probability Mass Function. Let S be the sample space of an experiment and consider a random variable $X : S \to \mathbb{R}$. Let $S_X \subseteq \mathbb{R}$ be the space of all possible values of X. The *probability mass function (pmf)* of X is a real-valued function $f_X : S_X \to \mathbb{R}$ that takes in a possible value $k \in S_X$ and spits out the probability $P(X=k)$ of getting this value. Kolomogorov's three rules of probability imply that the pmf satisfies:

- For any $k \in S_X$ we have $f_X(k) = P(X=k) \geqslant 0$.
- For any set of values $A \subseteq S_X$ we have

$$P(X \in A) = \sum_{k \in A} f_X(k) = \sum_{k \in A} P(X=k).$$

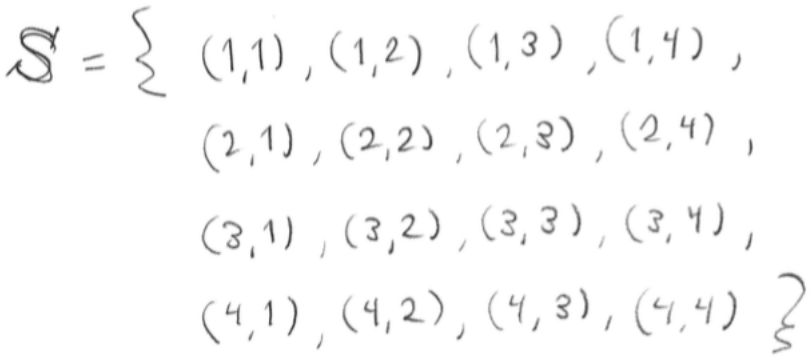
- The sum over all possible values is

$$\sum_{k \in S_X} f_X(k) = \sum_{k \in S_X} P(X=k) = 1.$$

///

To show you that this is not as difficult as it sounds, let's do an example from the textbook.

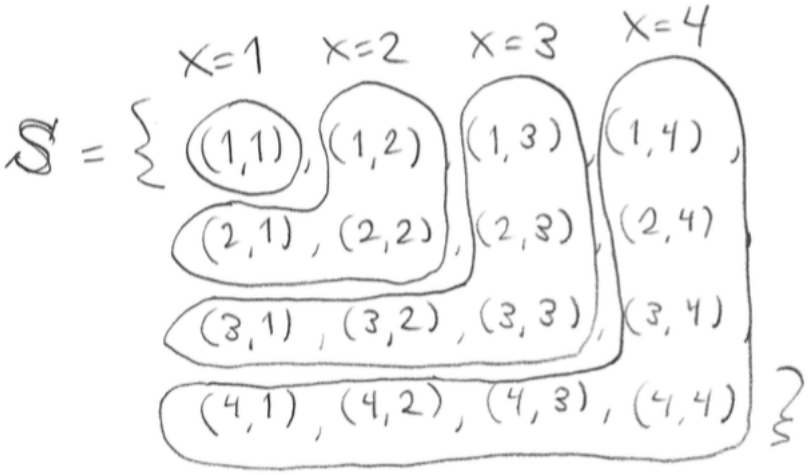
Example 2.1-3. Roll a fair 4-sided die twice and let X be the maximum of the two outcomes. The sample space for this experiment is



[Remark: I have treated the two dice as ordered because this leads to a sample space with equally likely outcomes.] If X is the maximum of the two outcomes then we observe that the space of possible values is

$$S_X = \{1, 2, 3, 4\}.$$

To compute the pmf we need to find the probability $P(X = k)$ for each value $k \in S_X$, and since the elements of S are equally likely, we need only to count how many outcomes correspond to each value of k. After a few moments of thought we find the following picture:



And we obtain the following table of probabilities:

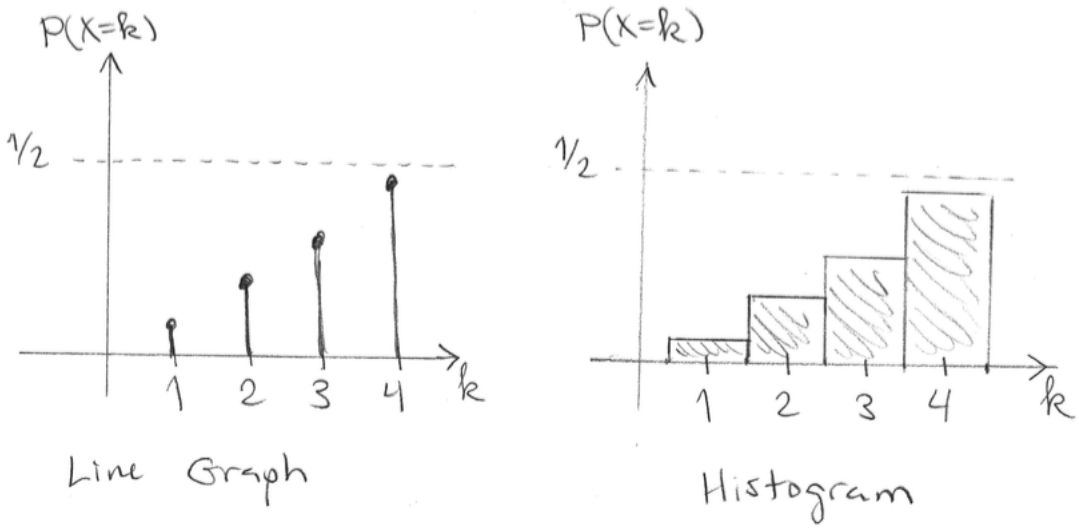
k	1	2	3	4
$P(X = k)$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{5}{16}$	$\frac{7}{16}$

Observe that we have

$$\sum_{k \in S_X} P(X = k) = \sum_{k=1}^{4} P(X = k) = \frac{1}{16} + \frac{3}{16} + \frac{5}{16} + \frac{7}{16} = \frac{1 + 3 + 5 + 7}{16} = \frac{16}{16} = 1,$$

as expected. The table above is one way to display the pmf, but there are others.

1. The pmf can be drawn. Since the input and output of the pmf are both real numbers, we can graph the function $f_X : \mathbb{R} \to \mathbb{R}$ just as we would in Calculus class. Two popular ways to do this are the *line graph* and the *histogram*:



In the line graph, the probability of $P(X = k)$ is represented as the height a vertical line drawn above k. In the histogram we draw instead a rectangle of width 1 and height $P(X = k)$ centered at k. Therefore the probability in the histogram is represented by area. This will be very important later when we consider **continuous** random variables; then the area of rectangles in the histogram will be replaced by the area under a smooth curve.

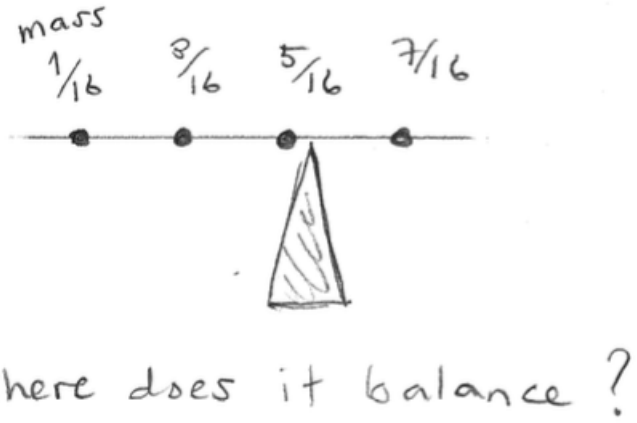
2. The pmf might have a formula. In our case, one can see by trial and error that

$$f_X(k) = P(X = k) = \frac{2k - 1}{16}.$$

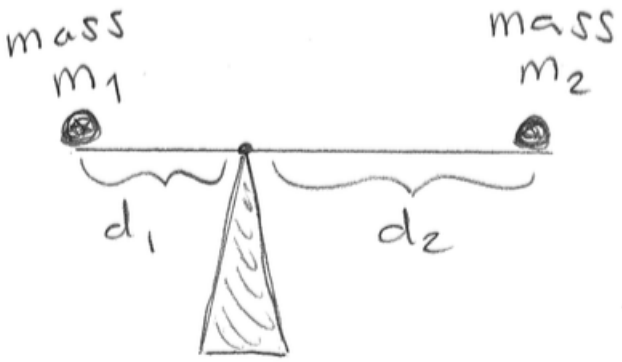
This is the most succinct way of encoding the distribution of this random variable. ///

We have seen that probability can be visualized as a **length** (in the line graph) or as an **area** (in the histogram). So why do we call it the probability **mass** function?

To understand this, we should think of the probabilities $f_X(k)$ as *point masses* distributed along a line. Is it natural to consider the question of where this distribution of masses will balance:



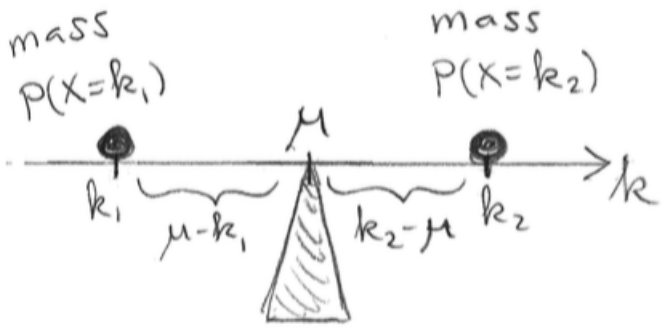
The only way to solve this is by using principles from physics. In particular, we will use *Archimedes' Law of the Lever*, which says the following:



Suppose that we have two point masses m_1 and m_2 on a balance board, at distances d_1 and d_2 from the fulcrum, respectively. Archimedes says that this system is in balance precisely when

$$\boxed{m_1 d_1 = m_2 d_2.}$$

To apply this principle to probability, let us consider a simple random variable $X : S \to \mathbb{R}$ with only two possible values, $S_X = \{k_1, k_2\}$. We are looking for a real number $\mu \in \mathbb{R}$ ("μ" is for "mean") such that the following system is in balance:



I will assume that $k_1 < \mu < k_2$, but this isn't really important. (The math will work out in any case.) Observe from the diagram that the point masses $P(X = k_1)$ and $P(X = k_2)$ have distance $\mu - k_1$ and $k_2 - \mu$ from the fulcrum, respectively. According to Archimedes, the system is in balance precisely when

$$\begin{aligned}(\mu - k_1)P(X = k_1) &= (\mu - k_2)P(X = k_2)\\ \mu P(X = k_1) + \mu P(X = k_2) &= k_1 P(X = k_1) + k_2 P(X = k_2)\\ \mu\,[P(X = k_1) + P(X = k_2)] &= k_1 P(X = k_1) + k_2 P(X = k_2),\end{aligned}$$

and since $P(X = k_1) + P(X = k_2) = 1$ this simplifies to

$$\boxed{\mu = k_1 P(X = k_1) + k_2 P(X = k_2).}$$

The same computation can be carried out for random variables with more than two possible values. This motivates the following definition.

Definition of Expected Value. Consider a discrete[2] random variable $X : S \to \mathbb{R}$ and let $S_X \subseteq \mathbb{R}$ be its space of possible values. Let $f_X : S_X \to \mathbb{R}$ denote the probability mass function. Then we define the *mean* or the *expected value* of X by the following formula:

$$\mu = E[X] = \sum_{k \in S_X} kP(X = k) = \sum_{k \in S_X} k f_X(k).$$

The intuition is that μ is the *center of mass* for the probability mass function. ///

Getting back to our example, let X be the **maximum** value in two rolls of a fair 4-sided die and recall that the pmf satisfies

k	1	2	3	4
$P(X = k)$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{5}{16}$	$\frac{7}{16}$

[2]I'll explain when this means when we discuss continuous random variables below.

Then the expected value of X is

$$\begin{aligned}
E[X] &= \sum_{k \in S_X} kP(X=k) \\
&= \sum_{k=1}^{4} kP(X=k) \\
&= 1P(X=1) + 2P(X=2) + 3P(X=3) + 4P(X=4) \\
&= 1 \cdot \frac{1}{16} + 2 \cdot \frac{3}{16} + 3 \cdot \frac{5}{16} + 4 \cdot \frac{7}{16} \\
&= \frac{1 \cdot 2 + 2 \cdot 3 + 3 \cdot 5 + 4 \cdot 7}{16} \\
&= \frac{1 + 6 + 15 + 28}{16} = \frac{50}{16} = 3.125.
\end{aligned}$$

Interpretation: If we perform this experiment many times and take the average of the resulting values of X (i.e., the maximum number that showed up each time), then we expect the average to be approximately 3.125.[3] It is remarkable that Archimedes' law of the lever helps us to solve this problem.

If you open the front cover of the textbook you will see a catalogue of famous probability distributions. We have already met a few of these:

Binomial Distribution. Consider a coin with $P(H) = p$. Flip the coin n times and X be the number of H's that we get. As we know by now, the pmf of this random variable is

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

Any random variable with this pmf is called a *binomial random variable.* ///

Let's compute the expection of a binomial random variable in the case $n = 3$. The pmf is given by the following table:

k	0	1	2	3
$P(X=k)$	$(1-p)^3$	$3p(1-p)^2$	$3p^2(1-p)$	p^3

Hence the expected value is

$$E[X] = 0 \cdot P(X=0) + 1 \cdot P(X=1) + 2 \cdot P(X=2) + 3 \cdot P(X=3)$$

[3] Don't just take my word for it. Try it yourself.

$$
\begin{aligned}
&= \cancel{0 \cdot (1-p)^3} + 1 \cdot 3p(1-p)^2 + 2 \cdot 3p^2(1-p) + 3 \cdot p^3 \\
&= 3p(1-p)^2 + 6p^2(1-p) + 3p^3 \\
&= 3p\left[(1-p)^2 + 2p(1-p) + p^2\right] \\
&= 3p\left[(1 - \cancel{2p} + \cancel{p^2}) + (\cancel{2p} - \cancel{2p^2}) + \cancel{p^2}\right] \\
&= 3p\,[1] \\
&= 3p.
\end{aligned}
$$

[Remark: There was a lot of lucky cancellation there.] Does the answer make sense? If our coin has $P(H) = p$ then we can think of p as a **ratio**, telling us how often the coin shows heads on average. If we flip the coin 3 times, then it is reasonable to expect that we will get $3p$ heads.

If we flip the coin n times, then it is reasonable to expect that we will see np heads because

$$np = (\text{total \# flips})(\text{ratio of flips that show heads}).$$

In other words, we believe that the following equation must be true:

$$E[X] = \sum_{k=0}^{n} kP(X=k) = \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \stackrel{?}{=} np.$$

One of the problems on HW3 will guide you through a tricky proof of this. In the next class I will show you another way to prove it that is easier but more abstract.

Hypergeometric Random Variable. Suppose that a bowl contains N_1 red balls and N_2 blue balls. You reach into the bowl and pull out n balls at random. Let X be the number of red balls that you get. As we know by now, the pmf of this random variable is

$$P(X=k) = \frac{\binom{N_1}{k}\binom{N_2}{n-k}}{\binom{N_1+N_2}{n}}.$$

Any random variable with this pmf is called a *hypergeometric random variable.*[4] ///

Example 2.2-5 in the textbook uses some tricks to compute the mean of the hypergeometric distribution as follows:

$$E[X] = \sum_{k=0}^{n} kP(X=k) = \sum_{k=0}^{n} k\frac{\binom{N_1}{k}\binom{N_2}{n-k}}{\binom{N_1+N_2}{n}} = n\left(\frac{N_1}{N_1+N_2}\right).$$

Without worrying about the details, does the final answer make sense? Yes: The ratio of red balls in the bowl is $N_1/(N_1+N_2)$. Therefore if we reach in and grab n it makes sense that the number of red balls we expect to get is

$$n\left(\frac{N_1}{N_1+N_2}\right) = (\text{\# of balls we grab})(\text{ratio of red balls in the bowl}).$$

[4] In my opinion this name is way too fancy. Sadly, it is too late to change it.

The final example for today is a bit less familiar, but we've seen it once or twice.

Geometric Random Variable. Consider a coin with $P(H)$. Begin to flip the coin and continue until you get H, then stop. Let X be the number of flips that you did. I claim that the pmf of this random variable is

$$P(X = k) = (1-p)^{k-1}p.$$

Any random variable with this pmf is called a *geometric random variable.* ///

Proof: This experiment is a bit tricky to analyze because it has an **infinite sample space**:

$$S = \{H, TH, TTH, TTTH, TTTTH, TTTTTH, \ldots\}.$$

The random variable X is just the length of the string, so that

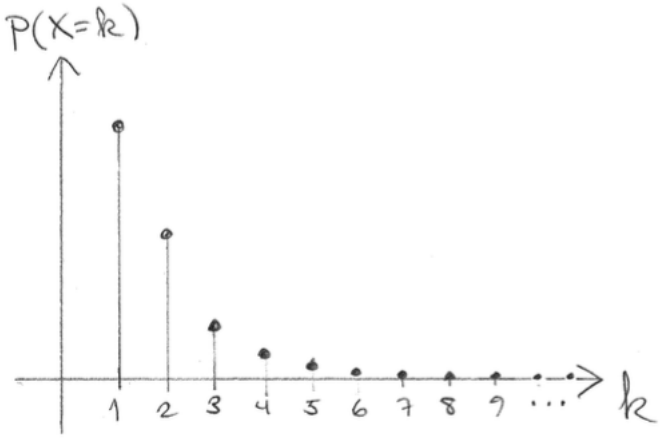
$$\begin{aligned} X(H) &= 1, \\ X(TH) &= 2, \\ X(TTH) &= 3, \\ &\vdots \end{aligned}$$

Thus the space of possible values of X is $S_X = \{1, 2, 3, \ldots\}$. Since there is only one way to get the value $k \in S_X$ we conclude that

$$P(X = k) = P(\underbrace{TT\cdots T}_{k-1 \text{ times}} H) = \underbrace{P(T)P(T)\cdots P(T)}_{k-1 \text{ times}} P(H) = P(T)^{k-1}P(H) = (1-p)^{k-1}p^k.$$

///

The line graph of a geometric random variable looks like this:



Since the set of possible values is infinite, the picture goes on forever to the right. However, since we are dealing with a probability distribution, we know that the infinite sum of probabilites must equal 1. That is, we must have a convergent series:

$$\begin{aligned}
1 &= \sum_{k \in S_X} P(X = k) \\
1 &= \sum_{k=1}^{\infty} (1-p)^{k-1} p^k \\
1 &= p + (1-p)p + (1-p)^2 p + (1-p)^3 p + \cdots \\
1 &= p\left[1 + (1-p) + (1-p)^2 + (1-p)^3 + \cdots\right] \\
\frac{1}{p} &= 1 + (1-p) + (1-p)^2 + (1-p)^3 + \cdots .
\end{aligned}$$

Is this true? Well, you may remember from Calculus II that if q is any number with $|q| < 1$ then the so-called *geometric series* is convergent:

$$1 + q + q^2 + q^3 + \cdots = \frac{1}{1-q}.$$

Now plugging in $q = 1 - p$ gives the desired result. This also explains why we call it a **geometric** random variable.[5]

Thinking Homework:[6] How many flips do you **expect** to make before you see the first H?

Oct 19

Consider an experiment with sample space S and let $X : S \to \mathbb{R}$ be any random variable. Let $S_X \subseteq \mathbb{R}$ be the set of possible values that X can take. Last time we defined the *expected value* of X by the formula

$$E[X] = \sum_{k \in S_X} kP(X = k).$$

The intuition behind this is that it represents the "center of mass" for the "probability mass function" $f_X(k) = P(X = k)$. However, there is another important formula for the expected value that we will need today. I claim that

$$\boxed{\sum_{k \in S_X} kP(X = k) = \sum_{s \in S} X(s)P(s),}$$

where the sum on the left is over all possible **values** of X and the sum on the right is over all possible **outcomes** of the experiment. To see why this equation is true, the key idea is to consider the event "$X = k$" $\subseteq S$, which is defined as the set of outcomes with value k:

$$\text{"}X = k\text{"} = \{s \in S : X(s) = k\} \subseteq S.$$

[5] Although we still might wonder why the infinite series is called "geometric."

[6] Not to be handed in; just think about it. We'll compute the answer later.

Then $P(X = k)$ is the sum of the probabilites $P(s)$ for all $s \in S$ such that $X(s) = k$. I've asked you to grapple with this on HW3. For now let's just look at an example.

Example. Consider a coin with $P(H) = p$. Flip the coin 2 times and let X be the number of heads. On the one hand, we have the following table of outcomes and probabilities:

s	TT	HT	TH	HH
$X(s)$	0	1	1	2
$P(s)$	$(1-p)^2$	$p(1-p)$	$p(1-p)$	p^2

Using the $s \in S$ formula for expectation gives:

$$\begin{aligned} E[X] &= \sum_{s \in S} X(s)P(s) \\ &= X(TT)P(TT) + X(HT)P(HT) + X(TH)P(TH) + X(HH)P(HH) \\ &= 0 \cdot (1-p)^2 + 1 \cdot p(1-p) + 1 \cdot p(1-p) + 2p^2. \end{aligned}$$

On the other hand, we have the following pmf for the binomial distribution:

k	0	1	2
$P(X = k)$	$(1-p)^2$	$2p(1-p)$	p^2

Using the $k \in S_X$ formula for expectation gives

$$\begin{aligned} E[X] &= \sum_{k \in S_X} kP(X = k) \\ &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) + 2 \cdot P(X = 2) \\ &= 0 \cdot (1-p)^2 + 1 \cdot 2p(1-p) + 2 \cdot p^2. \end{aligned}$$

Do you see why these formulas give the same answer? It always works like this.

Today we will talk about creating new random variables from old. For example, suppose that we have two (possibly different) random variables on the same experiment:

$$X, Y : S \to \mathbb{R}.$$

You may remember from Calculus that real-valued functions can be "added pointwise." In this case we can define the new random variable $X + Y : S \to \mathbb{R}$ by setting

$$(X + Y)(s) = X(s) + Y(s) \quad \text{for all } s \in S.$$

We can also multiply random variables by defining

$$(XY)(s) = X(s) \cdot Y(s) \quad \text{for all } s \in S$$

and for any constant $\alpha \in \mathbb{R}$ we can "scale" the random variable X by defining

$$(\alpha X)(s) = \alpha \cdot X(s) \quad \text{for all } s \in S.$$

In summary, we have three ways to combine random variables on the same experiment:

- addition,
- multiplication,
- scaling by constants.

Example. Continuing from the previous example, suppose we flip a coin twice, so that

$$S = \{TT, HT, TH, HH\}.$$

We define two new random variables:

$$X_1 = \begin{cases} 1 & \text{if 1st flip is } H \\ 0 & \text{if 1st flip is } T \end{cases} \qquad X_2 = \begin{cases} 1 & \text{if 2nd flip is } H \\ 0 & \text{if 2nd flip is } T \end{cases}$$

The following table shows the distribution of these random variables, and of their sum:

s	TT	HT	TH	HH
$P(s)$	$(1-p)^2$	$p(1-p)$	$p(1-p)$	p^2
$X_1(s)$	0	1	0	1
$X_2(s)$	0	0	1	1
$X_1(s) + X_2(s)$	0	1	1	2

Observe that the sum $X = X_1 + X_2$ is just a fancy way to describe the "number of heads," which we know has a binomial distribution. We saw in the previous example that

$$E[X] = 0 \cdot (1-p)^2 + 1 \cdot p(1-p) + 1 \cdot p(1-p) + 2 \cdot p^2 = 2p\left[(1-p) + p\right] = 2p.$$

Now let's compute the expectations of X_1 and X_2 using the same $s \in S$ formula. We have

$$\begin{aligned} E[X_1] &= X_1(TT)P(TT) + X_1(HT)P(HT) + X_1(TH)P(TH) + X_1(HH)P(HH) \\ &= 0 \cdot (1-p)^2 + 1 \cdot p(1-p) + 0 \cdot p(1-p) + 1 \cdot p^2 \\ &= p\left[(1-p) + p\right] \\ &= p \end{aligned}$$

and

$$\begin{aligned} E[X_2] &= X_2(TT)P(TT) + X_2(HT)P(HT) + X_2(TH)P(TH) + X_2(HH)P(HH) \\ &= 0 \cdot (1-p)^2 + 0 \cdot p(1-p) + 1 \cdot p(1-p) + 1 \cdot p^2 \\ &= p\left[(1-p) + p\right] \\ &= p. \end{aligned}$$

Is it a coincidence that

$$2p = E[X] = E[X_1 + X_2] = E[X_1] + E[X_2] = p + p?$$

No, in fact this is a very general phenomenon, called "linearity of expectation."

Theorem (Linearity of Expectation). Consider an experiment with sample space S. Let $X, Y : S \to \mathbb{R}$ be any random variables and let $\alpha, \beta \in \mathbb{R}$ be any constants. Then we have

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y].$$

Remark: The study of expected value brings us very close to the subject called "linear algebra." What we are really proving here is that expected value is a "linear function." ///

This is really hard to prove if we try to use the $k \in S_X$ formula for expected value, but it's really easy if we use the $s \in S$ formula.

Proof: We just apply the definitions:

$$\begin{aligned} E[\alpha X + \beta Y] &= \sum_{s \in S} (\alpha X + \beta Y)(s) P(s) \\ &= \sum_{s \in S} \left[\alpha X(s) + \beta Y(s)\right] \cdot P(s) \\ &= \sum_{s \in S} \left[\alpha X(s) P(s) + \beta Y(s) P(s)\right] \\ &= \sum_{s \in S} \alpha X(s) P(s) + \sum_{s \in S} \beta Y(s) P(s) \\ &= \alpha \sum_{s \in S} X(s) P(s) + \beta \sum_{s \in S} Y(s) P(s) \\ &= \alpha E[X] + \beta E[Y]. \end{aligned}$$

///

WARNING: This theorem says that expected value "preserves" addition and scaling of random variables. I want to emphasize, however, that expected value (usually) does **not** preserve multiplication:

$$E[XY] \neq E[X] \cdot E[Y]\,!$$

In particular (when $Y = X$) this tells us that

$$E[X^2] \neq E[X]^2.$$

This will be important below. ///

Linearity of expectation is an abstract concept, but it has powerful applications.

Application (Expectation of a Binomial). Consider a coin with $P(H) = p$. Flip the coin n times and let X be the number of heads that you get. We know that

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

On the other hand, we can express the binomial random variable X as a sum of simpler random variables. For each $i = 1, 2, \ldots, n$ let us define

$$X_i = \begin{cases} 1 & \text{if the } i\text{th flip is } H \\ 0 & \text{if the } i\text{th flip is } T \end{cases}$$

Observe that each X_i has two possible values $S_{X_i} = \{0, 1\}$ with pmf given by

$$P(X_i = 0) = 1 - p \qquad \text{and} \qquad P(X_i = 1) = p.$$

Hence we can compute the expected value of X_i using the $k \in S_{X_i}$ formula:

$$E[X_i] = 0 \cdot P(X_i = 0) + 1 \cdot P(X_i = 1) = 0 \cdot (1-p) + 1 \cdot p = p.$$

What does this mean? Another way to phrase the definition of X_i is "the number of heads that we get on the ith flip." Thus the formula $E[X_i] = p$ says that (on average) we expect to get p heads on the ith flip. That sounds reasonable, I guess. Then since X_i is the number of heads on the ith flip, the sum of the X_i gives the total number of heads:

$$\begin{aligned} (\text{total \# heads}) &= \sum_{i=1}^{n} (\text{\# heads on the } i\text{th flip}) \\ X &= X_1 + X_2 + \cdots + X_n. \end{aligned}$$

Finally, we can use the linearity of expectation to compute the expected number of heads:

$$\begin{aligned} E[X] &= E[X_1 + X_2 + \cdots + X_n] \\ &= E[X_1] + E[X_2] + \cdots + E[X_n] \\ &= \underbrace{p + p + \cdots + p}_{n \text{ times}} \\ &= np. \end{aligned}$$

That was pretty slick, right? On HW3 you will give a much uglier proof that $E[X] = np$. ///

And we're not done yet. Pretty soon we'll develop a Calculus trick that will allow us to compute all of the so-called *moments of* X, i.e., the expected values $E[X^k]$ for $k = 1, 2, 3, \ldots$. Taken together, the infinite sequence of moments

$$E[X], E[X^2], E[X^3], \ldots$$

tells us everything we want to know about the random variable X. In the next section we'll discuss the meaning of the *second moment* $E[X^2]$.

We have seen that the expected value $E[X]$ gives us useful information about the random variable X. But it doesn't tell us everything.

Example: Consider the following two random variables.

- Roll a fair 6-sided die with sides labeled $1, 2, 3, 4, 5, 6$ and let X be the number that shows up.
- Roll a fair 6-sided die with sides labeled $3, 3, 3, 4, 4, 4$ and let Y be the number that shows up.

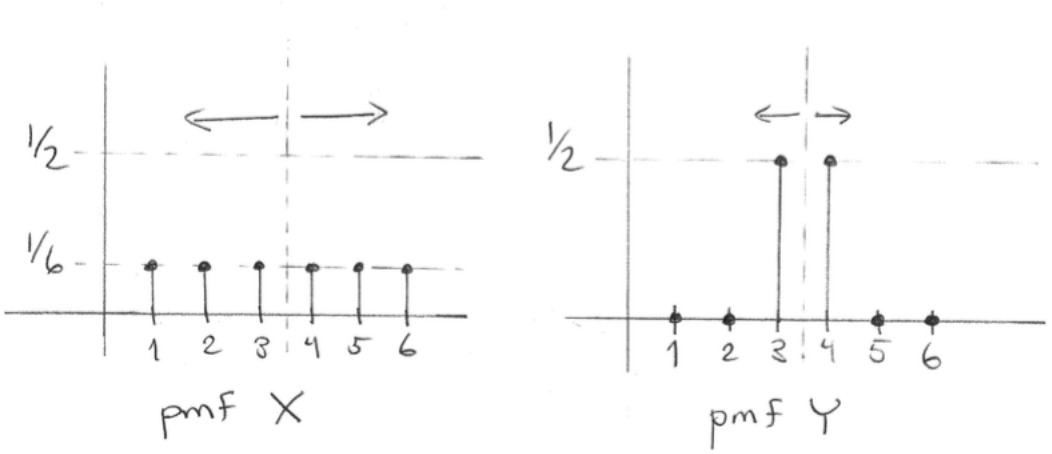
To compute the expected value of X we note that $S_X = \{1, 2, 3, 4, 5, 6\}$ with $P(X = k) = 1/6$ for all $k \in S_X$. Hence

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5.$$

To compute the expected value of Y we note that $S_Y = \{3, 4\}$ with $P(X = 3) = P(X = 4) = 3/6 = 1/2$. Hence

$$E[Y] = 3 \cdot P(Y = 3) + 4 \cdot P(Y = 4) = 3 \cdot \frac{1}{2} + 4 \cdot \frac{1}{2} = \frac{7}{2} = 3.5.$$

We conclude that X and Y have the same expected value. But they certainly do **not** have the same distribution, as we can see in the following line graphs:



We see that both distributions are centered at 3.5, but the distribution of X is more "spread out" than the distribution of Y. We would like to attach some **number** to each distribution to give a measure of its "spread," and to verify quantitatively that

$$\text{spread}(X) > \text{spread}(Y).$$

How can we do this?

Idea of "Spread." Let X be a random variable with expected value $\mu := E[X]$ (also known as the *mean* of X). We want to answer the question:

On average, how far away is X from its mean μ?

///

The most obvious way to do this is to compute the expected value of the difference $X - \mu$. We use the linearity of expectation and the fact that $E[\mu] = \mu$ (because μ is a **constant**, i.e., it's not random) to get

$$E[X - \mu] = E[X] - E[\mu] = E[X] - \mu = \mu - \mu = 0.$$

Oops. Maybe we should have seen that coming. Since X spends about half of its time on the right of μ and half of its time on the left of μ, it seems that the differences canel out.

We can fix this by taking the **distance** between X and μ, which is the absolute value of the difference: $|X - \mu|$. We will call the expected value of this distance the *spread*[7] of X:

$$\text{spread}(X) = E\left[|X - \mu|\right].$$

[7]Warning: This is not standard terminology.

To see if this idea is reasonable, let's compute the spread of the random variables X and Y from above. Unfortunately, the function $|X - \mu|$ is a bit complicated so we will have to go back to the explicit formula:

$$E\left[|X - \mu|\right] = \sum_{s \in S} |X(s) - \mu| \cdot P(s).$$

To compute the spread of X, we form the following table:

s	face 1	face 2	face 3	face 4	face 5	face 6
X	1	2	3	4	5	6
μ	3.5	3.5	3.5	3.5	3.5	3.5
$\lvert X - \mu \rvert$	2.5	1.5	0.5	0.5	1.5	2.5
P	1/6	1/6	1/6	1/6	1/6	1/6

And then we apply the formula

$$E\left[|X - \mu|\right] = (2.5)\frac{1}{6} + (1.5)\frac{1}{6} + (0.5)\frac{1}{6} + (0.5)\frac{1}{6} + (1.5)\frac{1}{6} + (2.5)\frac{1}{6} = \frac{9}{6} = 1.5.$$

We conclude that, on average, the random variable X has a distance of 1.5 from its mean. To compute the spread of Y, we form the following table:

s	face 1	face 2	face 3	face 4	face 5	face 6
Y	3	3	3	4	4	4
μ	3.5	3.5	3.5	3.5	3.5	3.5
$\lvert Y - \mu \rvert$	0.5	0.5	0.5	0.5	0.5	0.5
P	1/6	1/6	1/6	1/6	1/6	1/6

And then we apply the formula

$$E\left[|Y - \mu|\right] = (0.5)\frac{1}{6} + (0.5)\frac{1}{6} + (0.5)\frac{1}{6} + (0.5)\frac{1}{6} + (0.5)\frac{1}{6} + (0.5)\frac{1}{6} = \frac{3}{6} = 0.5.$$

We conclude that, on average, the random variable Y has a distance of 0.5 from its mean. This confirms our earlier intuition that

$$1.5 = \text{spread}(X) > \text{spread}(Y) = 0.5.$$

///

Now the bad news. Even though our definition of "spread" is very reasonable, this definition is not commonly used in probability and statistics. The main reason we don't use it is because

the absolute value function is not very algebraic. To make the algebra work out smoothly we prefer to work with the **square of the distance** between X and μ:

$$(\text{distance between } X \text{ and } \mu)^2 = |X-\mu|^2 = (X-\mu)^2.$$

Notice that when we do this the absolute value signs disappear.

Definition of Variance and Standard Deviation. Let X be a random variable with mean $\mu = E[X]$. We define the *variance* as the expected value of the squared distance between X and μ:

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

Then since we feel remorse about squaring the distance, we try to correct the situation by defining the *standard deviation* σ as the square root of the variance:

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{E\left[(X-\mu)^2\right]}.$$

///

Remark: The standard deviation of X is **not the same** as the spread that we defined earlier. In general we have the inequality

$$E\left[|X-\mu|\right] = \text{spread}(X) \leqslant \sigma.$$

However, the standard deviation has certain theoretical advantages that we will see later.

To finish this section, let's compute the standard deviations of X and Y. For now we'll use the explicit formula

$$E\left[(X-\mu)^2\right] = \sum_{s\in S}(X(s)-\mu)^2 P(s),$$

but later we'll see that there are tricks. Here's the table for X:

s	face 1	face 2	face 3	face 4	face 5	face 6
X	1	2	3	4	5	6
μ	3.5	3.5	3.5	3.5	3.5	3.5
$(X-\mu)^2$	6.25	2.25	0.25	0.25	2.25	6.25
P	1/6	1/6	1/6	1/6	1/6	1/6

Thus we have

$$\sigma_X^2 = E\left[(X-\mu)^2\right] = (6.25)\frac{1}{6}+(2.25)\frac{1}{6}+(0.25)\frac{1}{6}+(0.25)\frac{1}{6}+(2.25)\frac{1}{6}+(6.25)\frac{1}{6} = \frac{17.5}{6} \approx 2.92.$$

and $\sigma_X = \sqrt{\sigma_X^2} \approx 1.71$. And here's the table for Y:

s	face 1	face 2	face 3	face 4	face 5	face 6
Y	3	3	3	4	4	4
μ	3.5	3.5	3.5	3.5	3.5	3.5
$(Y-\mu)^2$	0.25	0.25	0.25	0.25	0.25	0.25
P	1/6	1/6	1/6	1/6	1/6	1/6

Thus we have

$$\sigma_Y^2 = E\left[(Y-\mu)^2\right] = (0.25)\frac{1}{6} + (0.25)\frac{1}{6} + (0.25)\frac{1}{6} + (0.25)\frac{1}{6} + (0.25)\frac{1}{6} + (0.25)\frac{1.5}{6} = \frac{1.5}{6} = 0.25.$$

and $\sigma_X = \sqrt{\sigma_X^2} = 0.5$. You should observe that

$$\begin{aligned} 1.5 = \text{spread}(X) &\leqslant \sigma_X \approx 1.71, \\ 0.5 = \text{spread}(Y) &\leqslant \sigma_Y = 0.5. \end{aligned}$$

But we still have

$$1.71 \approx \sigma_X > \sigma_Y = 0.5,$$

so it seems that standard deviation is a reasonable measure of the spread of X and Y.

Oct 26

Let X be a random variable and recall our definition of the *variance*:

$$\sigma^2 = \text{Var}(X) = E[(X-\mu)^2] = \sum_k (k-\mu)^2 P(X=k).$$

In other words, the variance is the expected value of the **squared distance** between X and its mean μ. This is some measure of "how spread out" X is around its mean. The fact that we use the squared distance instead of the plain distance is somewhat arbitrary, but it leads to a nice mathematical theory. For example, we have the following nice trick.

Trick for Computing Variance: $\text{Var}(X) = E[X^2] - E[X]^2$.

Proof: Let $\mu = E[X]$ be the mean. Then we use the **linearity of expectation** and the fact that μ **is constant** to compute

$$\begin{aligned} \text{Var}(X) &= E[(X-\mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \end{aligned}$$

$$
\begin{aligned}
&= E[X^2] - 2\mu E[X] + E[\mu^2] \\
&= E[X^2] - 2\mu \cdot \mu + \mu^2 \\
&= E[X^2] - \mu^2 \\
&= E[X^2] - E[X]^2.
\end{aligned}
$$

///

Remark: Since the squared distance $(X - \mu)^2$ is always non-negative, its expected value $\mathrm{Var}(X) = E[(X - \mu)^2]$ is also non-negative, and it follows from this that

$$
\begin{aligned}
\mathrm{Var}(X) &\geqslant 0 \\
E[X^2] - E[X]^2 &\geqslant 0 \\
E[X^2] &\geqslant E[X]^2.
\end{aligned}
$$

That might be useful later.

To illustrate these concepts, let X be a binomial random variable with $n = 2$ and $p = 1/3$. That, is suppose we have a coin with $P(H) = 1/3$ where we flip the coin twice and let X be the number of heads that we get. As we already know, the expected number of heads is

$$
E[X] = np = 2 \cdot \frac{1}{3} = \frac{2}{3}.
$$

Now let's compute the variance in two different ways. Here is a table with all the information we will need:

k	0	1	2
k^2	0	1	4
μ	2/3	2/3	2/3
$(k-\mu)^2$	4/9	1/9	16/9
$P(X=k)$	4/9	4/9	1/9

Using the definition of variance gives

$$
\begin{aligned}
\mathrm{Var}(X) &= \sum_k (k-\mu)^2 P(X=k) \\
&= (0-\mu)^2 P(X=0) + (1-\mu)^2 P(X=1) + (2-\mu)^2 P(X=2) \\
&= \frac{4}{9} \cdot \frac{4}{9} + \frac{1}{9} \cdot \frac{4}{9} + \frac{16}{9} \cdot \frac{1}{9} \\
&= \frac{16 + 6 + 16}{81} \\
&= \frac{36}{81} \\
&= \frac{4}{9}.
\end{aligned}
$$

On the other hand, using the trick gives

$$\begin{aligned}\mathrm{Var}(X) &= E[X^2] - E[X]^2 \\ &= \left[\sum_k k^2 P(X=k)\right] - \mu^2 \\ &= \left[0\cdot\frac{4}{9} + 1\cdot\frac{4}{9} + 4\cdot\frac{1}{9}\right] - \left(\frac{2}{3}\right)^2 \\ &= \frac{8}{9} - \frac{4}{9} \\ &= \frac{4}{9}.\end{aligned}$$

Which method do you prefer? In either case, since the variance is $\sigma^2 = 4/9$ we conclude that the standard deviation is $\sigma = \sqrt{4/9} = 2/3$.

Now let's consider a general binomial random variable with pmf

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

We saw on HW3 and in class that the mean is $\mu = E[X] = np$. It turns out that the variance has a similarly nice formula:

$$\sigma^2 = \mathrm{Var}(X) = np(1-p).$$

But why? If we try to compute the variance from the definition we get a terrible formula

$$\mathrm{Var}(X) = \sum_k (k-np)^2 \binom{n}{k} p^k (1-p)^{n-k}.$$

It's possible to simplify this monster using algebraic manipulations, but from your experience on HW3 you know that this will not be fun.

Here's a better explanation: Suppose we flip a coin n times and let X_i be the number of heads that we get on the ith flip, so that

$$X_i = \begin{cases} 1 & \text{if the } i\text{th flip is } H, \\ 0 & \text{if the } i\text{th flip is } T. \end{cases}$$

(This is our friend the Bernoulli random variable.) It is easy to compute the variance of X_i. To do this, we recall that $E[X_i] = p$ and we compute that

$$E[X_i^2] = \sum_k k^2 P(X_i = k) = 0^2 P(X_i = 0) + 1^2 P(X_i = 1) = 0(1-p) + 1p = p.$$

Thus we conclude that

$$\mathrm{Var}(X_i) = E[X_i^2] - E[X_i]^2 = p - p^2 = p(1-p).$$

Finally, we use the fact that a binomial random variable is a sum of Bernoullis:

$$\begin{aligned} X &= X_1 + X_2 + \cdots + X_n \\ \mathrm{Var}(X) &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \qquad (?) \\ &= p(1-p) + p(1-p) + \cdots + p(1-p) \\ &= np(1-p). \end{aligned}$$

This computation is correct, but I still haven't explained why the step (?) true.

Question: Why is it okay to replace the variance of the sum $\mathrm{Var}(X_1 + \cdots + X_n)$ with the sum of the variances $\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$?

Answer: This only works because our random variables X_i and X_j are **independent** for all $i \neq j$.[8] In general, it is **not** okay to replace the variance of a sum by the sum of the variances. That is, unlike the expected value $E[-]$, the variance function $\mathrm{Var}(-)$ is **not linear**.

For general (i.e., non-independent) random variables X and Y we will have

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + \text{some junk}.$$

Let's investigate what kind of junk this is. To keep track of the means we will use the notation

$$\mu_X = E[X] \qquad \text{and} \qquad \mu_Y = E[Y].$$

Since the expected value **is** linear we have

$$E[X+Y] = E[X] + E[Y] = \mu_X + \mu_Y.$$

Now we compute the variance of $X+Y$ directly from the definition:

$$\begin{aligned} \mathrm{Var}(X+Y) &= E\left[\left[(X+Y) - (\mu_X + \mu_Y)\right]^2\right] \\ &= E\left[\left[(X-\mu_X) + (Y-\mu_Y)\right]^2\right] \\ &= E\left[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)\right] \\ &= E\left[(X-\mu_X)^2\right] + E\left[(Y-\mu_Y)^2\right] + 2E\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2E\left[(X-\mu_X)(Y-\mu_Y)\right] \end{aligned}$$

[8] If $i \neq j$, then the number of heads you get on the ith flip has no relation to the number of heads you get on the jth flip. This is what we mean when we say that a coin has "no memory."

We don't like to write this out every time so we give the junk a special name.

Definition of Covariance. Let X and Y be random variables with means $\mu_X = E[X]$ and $\mu_Y = E[Y]$. We define their *covariance* as

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

Then the variance of the sum $X+Y$ is given by

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\cdot\mathrm{Cov}(X,Y).$$

///

Remark: Observe that $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$.

The covariance $\mathrm{Cov}(X,Y)$ is some measure of how "non-independent" or "entangled" the random variables are. In particular, when X and Y are **independent** we will have

$$\mathrm{Cov}(X,Y) = 0 \qquad \text{and hence} \qquad \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

To be specific about this I must finally define the word "independent."

Definition of Joint pmf and Independence. Let X and Y be random variables with supports $S_X \subseteq \mathbb{R}$ and $S_Y \subseteq \mathbb{R}$. Then for all possible values $k \in S_X$ and $\ell \in S_Y$ we define

$$\begin{aligned} f_{XY}(k,\ell) &= P(X=k, Y=\ell) \\ &= P(X=k \text{ and } Y=\ell) \\ &= P(\text{“}X=k\text{”} \cap \text{“}Y=\ell\text{”}). \end{aligned}$$

The function $f_{XY} : S_X \times S_Y \to \mathbb{R}$, which takes a pair of possible values (k,ℓ) for X and Y and spits out the probability of "$X=k$ and $Y=\ell$" is called the *joint pmf* of X and Y.

The random variables X and Y are called *independent* if for all $k \in S_X$ and $\ell \in S_Y$ we have

$$\begin{aligned} f_{XY}(k,\ell) &= f_X(k) f_Y(\ell) \\ P(X=k, Y=\ell) &= P(X=k)P(Y=\ell). \end{aligned}$$

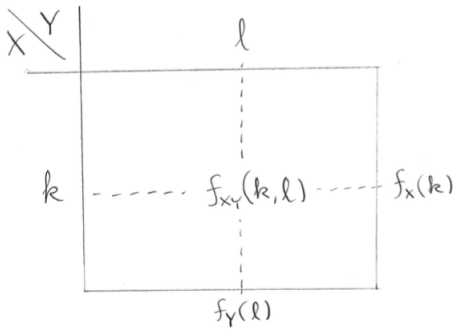
In other words, X and Y are independent if their joint pmf f_{XY} is the product of their *marginal pmfs* f_X and f_Y. ///

There are a lot of letters in that definition, so let's see an example with some numbers.

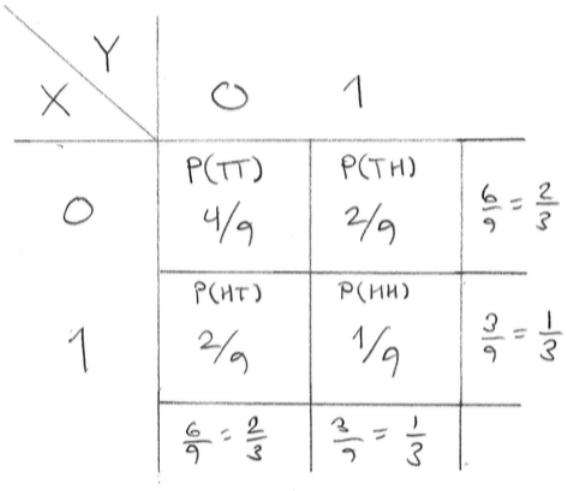
Example of Independence and Depencence. Consider a coin with $P(H) = 1/3$. Flip the coin twice and let

$$X = \text{“\# heads on 1st flip,”}$$
$$Y = \text{“\# heads on 2nd flip,”}$$
$$W = \text{“total \# heads”} = X + Y.$$

Our intuition is that X and Y are independent, but X and W are probably not. To see this we will compute the joint pmfs. In general, we display the joint pmf f_{XY} in a rectangular table with the marginal pmfs f_X and f_Y in the margins,[9] as follows:



Here is the joint pmf for our specific X and Y:

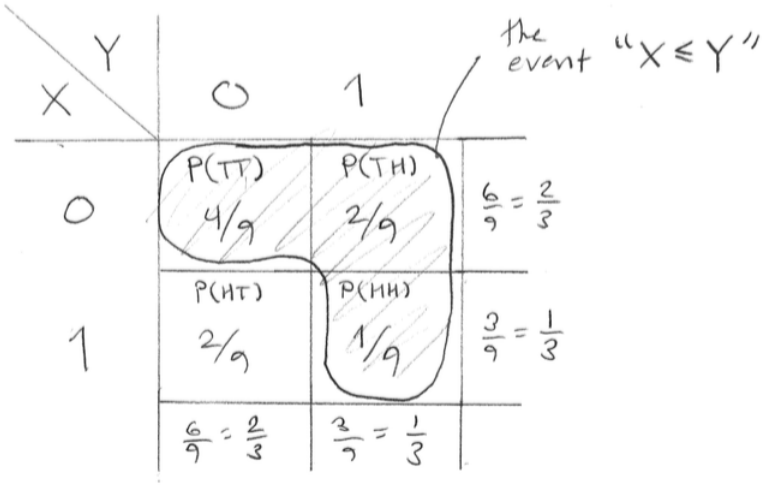


Note that the marginal pmfs f_X and f_Y are obtained by summing the entries in each row and column, respectively. We observe that each of these is a Bernoulli pmf with $p = 1/3$, as expected. Observe also that each entry of the table $f_{XY}(k, \ell)$ is equal to the product of the

[9] This is the reason for the name “marginal.”

marginal below, i.e., $f_X(k)$, and the marginal to the right, i.e., $f_Y(\ell)$. This confirms that the random variables X and Y are independent.

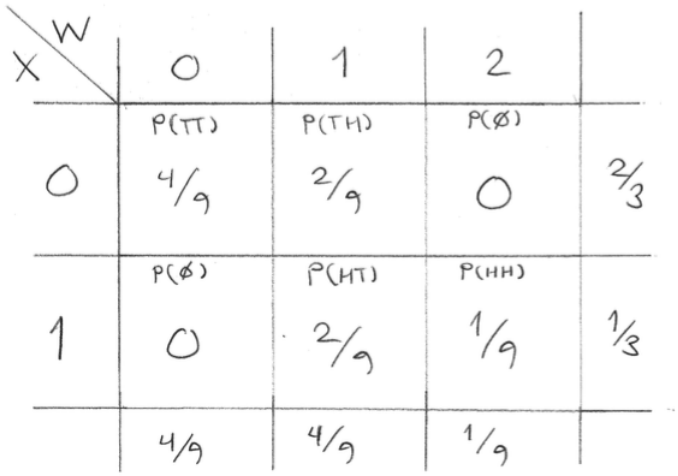
The joint pmf table tells us everything we could ever want to know about the joint distribution of X and Y. For example, to compute $P(X \leqslant Y)$ we simply add the probabilies from the cells corresponding to the event "$X \leqslant Y$" as in the following picture:



We conclude that

$$P(X \leqslant Y) = \frac{4}{9} + \frac{2}{9} + \frac{1}{9} = \frac{7}{9}.$$

Now let's move on to the joint distribution of X and W. Here is the pmf table:



Again, the marginal pmfs f_X and f_W are familiar to us, being a Bernoulli and a binomial,

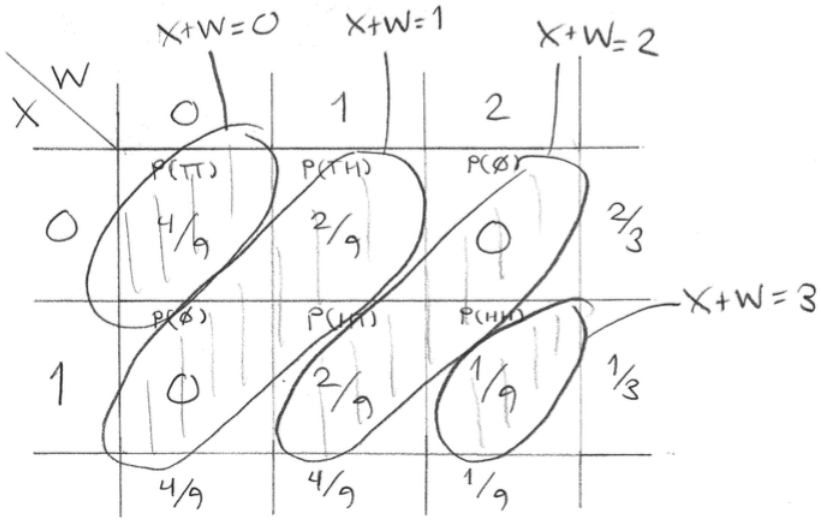
respectively.[10] Now it is immediately clear that the random variables X and W are not independent. To see this we only need to note, for example, that the entry $P(X = 0, W = 2) = 0$ is **not** equal to the product of the marginals $P(X = 0)P(W = 2) = (2/3)(1/9)$. Because of this, we expect that the covariance is not zero:

$$\mathrm{Cov}(X, W) \neq 0.^{11}$$

Let's go ahead and compute $\mathrm{Cov}(X, W)$ to be sure. There are many ways to compute covariance. For extra practice with joint distributions I'll use the formula

$$\mathrm{Var}(X + W) = \mathrm{Var}(X) + \mathrm{Var}(W) + 2 \cdot \mathrm{Cov}(X, W).$$

We already know that $\mathrm{Var}(X) = p(1-p) = (1/3)(2/3) = 2/9$ and $\mathrm{Var}(W) = np(1-p) = 2(1/3)(2/3) = 4/9$, so it only remains to compute $\mathrm{Var}(X + W)$. To do this, we note that $X + W$ can take on values $S_{X+W} = \{0, 1, 2, 3\}$. I have circled the events "$X + W = k$" for each $k \in S_{X+W}$ in the following picture:



By adding up the probabilities in each blob, we obtain the pmf of $X + W$:

k	0	1	2	3
$P(X + W = k)$	4/9	2/9	2/9	1/9

This allows us to compute

$$E[X + W] = 0 \cdot (4/9) + 1 \cdot (2/9) + 2 \cdot (2/9) + 3 \cdot (1/9) = 9/9 = 1,$$
$$E[(X + W)^2] = 0^2 \cdot (4/9) + 1^2 \cdot (2/9) + 2^2 \cdot (2/9) + 3^2 \cdot (1/9) = 19/9,$$

[10] In fact, we already computed the distribution f_W in today's lecture.

[11] It can sometimes happen by accident that $\mathrm{Cov}(X, Y) = 0$ for non-independent random variables X and Y, but this is quite rare. On ther other hand, if X and Y are independent, it is guaranteed that $\mathrm{Cov}(X, Y) = 0$.

and hence

$$\mathrm{Var}(X+W) = E[(X+W)^2] - E[X+W]^2 = \frac{19}{9} - 1^2 = \frac{10}{9}.$$

That certainly was good practice. Now let me mention an easier way to compute $\mathrm{Var}(X+W)$. If X is any random variable and $\alpha \in \mathbb{R}$ is any constant then it is easy to check that

$$\boxed{\mathrm{Var}(\alpha X) = \alpha^2 \mathrm{Var}(X).}$$

Proof: We use the linearity of expectation:

$$\begin{aligned}
\mathrm{Var}(\alpha X) &= E[(\alpha X)^2] - E[\alpha X]^2 \\
&= E[\alpha^2 X^2] - (\alpha E[X])^2 \\
&= \alpha^2 E[X^2] - \alpha^2 E[X]^2 \\
&= \alpha^2 \left(E[X^2] - E[X]^2 \right) \\
&= \alpha^2 \mathrm{Var}(X).
\end{aligned}$$

///

Applying this to our problem, we can use the fact that $W = X + Y$ and that X and Y are independent to get

$$\begin{aligned}
\mathrm{Var}(X+W) &= \mathrm{Var}(X+X+Y) \\
&= \mathrm{Var}(2X+Y) \\
&= \mathrm{Var}(2X) + \mathrm{Var}(Y) \\
&= 2^2 \cdot \mathrm{Var}(X) + \mathrm{Var}(Y) \\
&= 4 \cdot \frac{2}{9} + \frac{2}{9} \\
&= \frac{10}{9}.
\end{aligned}$$

Finally, we conclude that

$$\begin{aligned}
\mathrm{Var}(X+W) &= \mathrm{Var}(X) + \mathrm{Var}(W) + 2 \cdot \mathrm{Cov}(X,W) \\
10/9 &= 2/9 + 4/9 + 2 \cdot \mathrm{Cov}(X,W) \\
4/9 &= 2 \cdot \mathrm{Cov}(X,W) \\
2/9 &= \mathrm{Cov}(X,W).
\end{aligned}$$

This confirms once again that X and W are not independent. In fact, this is quite reasonable since the total number of heads (i.e., W) depends in some way on the number of heads that we get on the 1st flip (i.e., X). The covariance $\mathrm{Cov}(X,W) = 2/9$ just attaches a number to our intuition.

Oct 31 (I wore a skull shirt under my other shirt)

Let's recap.

Consider a fixed experiment with sample space S. Then given any two random variables $X, Y : S \to \mathbb{R}$ and any two constants $\alpha, \beta \in \mathbb{R}$ we have

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y].$$

In short, we say that "expectation is a linear function." If $\mu_X = E[X]$ is the expected value of X then we defined the variance and showed that

$$\sigma_X^2 = \mathrm{Var}(X) = E[(X - \mu_X)^2] = E[X^2] - \mu_X^2 = E[X^2] - E[X]^2.$$

Then we made the following observation:

Unlike the expected value $E[-]$, the variance $\mathrm{Var}(-)$ is **not a linear function.**

So what is it? For any random variables $X, Y : S \to \mathbb{R}$ we did a computation to show that

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X, Y),$$

where the *covariance* is defined by

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

This is a number that is supposed to measure the "non-independence" or the "entanglement" of X and Y. If X and Y are *independent* (technically: their joint pmf is the product of marginal pmfs) then we will have $\mathrm{Cov}(X, Y) = 0$ and hence $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Last time we computed an example of covariance using brute force. Today we'll use a nice trick.

Trick for Computing Covariance. Consider any random variables X, Y with expected values $\mu_X = E[X]$ and $\mu_Y = E[Y]$. Then we have

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

///

Remark: Before seeing the proof, you should observe that this is just a generalization of our earlier trick for computing the variance. Indeed, by setting $Y = X$ we obtain

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = E[X \cdot X] - E[X] \cdot E[X] = E[X^2] - E[X]^2.$$

The proof will be pretty much the same.

Proof: We will use the linearity of expectation and the fact that μ_X and μ_Y are constants. From the definition of covariance we have

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\
&= E[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] \\
&= E[XY] - E[\mu_X Y] - E[\mu_Y X] + E[\mu_X\mu_Y] \\
&= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X\mu_Y \\
&= E[XY] - \mu_X\mu_Y - \cancel{\mu_Y\mu_X} + \cancel{\mu_X\mu_Y} \\
&= E[XY] - \mu_X\mu_Y \\
&= E[XY] - E[X]\cdot E[Y].
\end{aligned}$$

///

Let's revisit the example from last time, using our new trick formula.

Example Continued from Oct 26. Consider a coin with $P(H) = 1/3$. Flip the coin twice and consider the following random variables:

$$\begin{aligned}
X &= \text{“\# heads on 1st flip,”} \\
Y &= \text{“\# heads on 2nd flip,”} \\
W &= \text{“total \# heads”} = X + Y.
\end{aligned}$$

To examine the relationship between X and W we only need to compute the expectations

$$E[X], E[X^2], E[W], E[W^2], E[XW].$$

And these are easy to compute from the following table:

s	TT	TH	HT	HH		
P	4/9	2/9	2/9	1/9		
X	0	0	1	1	$\longrightarrow$	$E[X] = 1(2/9) + 1(1/9) = 3/9,$
X^2	0	0	1	1	$\longrightarrow$	$E[X^2] = 1(2/9) + 1(1/9) = 3/9,$
W	0	1	1	2	$\longrightarrow$	$E[W] = 1(2/9) + 1(2/9) + 2(1/9) = 6/9,$
W^2	0	1	1	4	$\longrightarrow$	$E[W^2] = 1(2/9) + 1(2/9) + 4(1/9) = 8/9,$
XW	0	0	1	2	$\longrightarrow$	$E[XW] = 1(2/9) + 2(1/9) = 4/9.$

We use the expected values to compute the variances

$$\begin{aligned}
\mathrm{Var}(X) &= E[X^2] - E[X]^2 = (3/9) - (3/9)^2 = 2/9, \\
\mathrm{Var}(W) &= E[W^2] - E[W]^2 = (8/9) - (6/9)^2 = 4/9,
\end{aligned}$$

and the covariance

$$\mathrm{Cov}(X,W) = E[XW] - E[X]\cdot E[W] = (4/9) - (3/9)(6/9) = 2/9.$$

That was easier than last time, right? ///

For posterity, let me record a few last properties of variance.

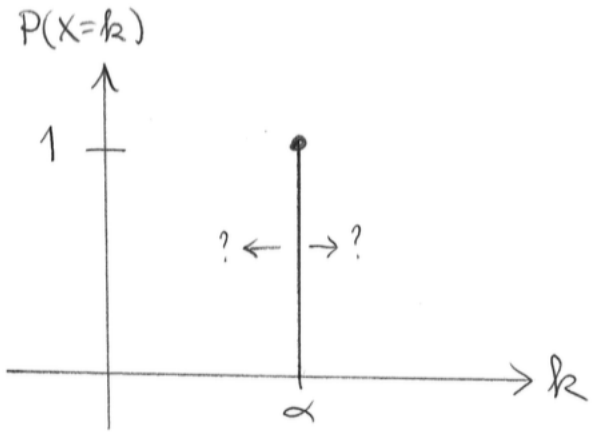
Properties of Variance.

- Let $\alpha \in \mathbb{R}$ be any constant. Then we have $\mathrm{Var}(\alpha) = 0$.

  Proof: We use the fact that $E[\alpha] = \alpha$ to obtain

  $$\mathrm{Var}(\alpha) = E[\alpha^2] - E[\alpha]^2 = \alpha^2 - (\alpha)^2 = 0.$$

  The idea behind this is that a constant random variable has **zero spread**, i.e., it is totally concentrated at its center of mass:

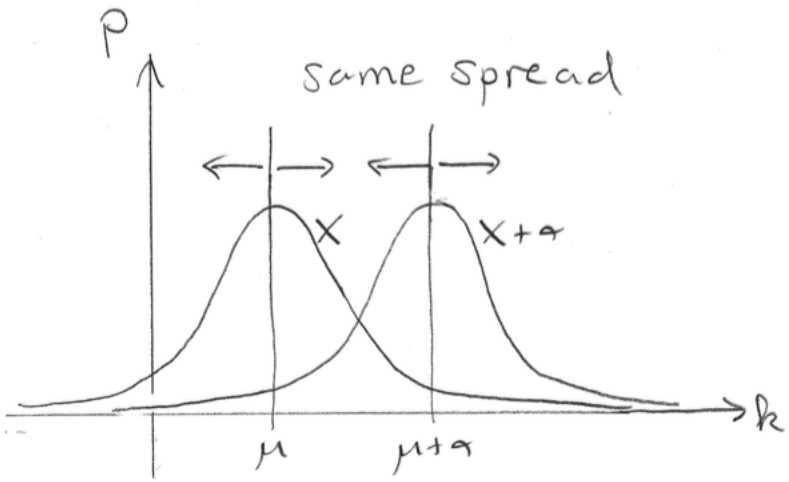


- Let $\alpha \in \mathbb{R}$ be constant and let X be any random variable. Then $\mathrm{Var}(X + \alpha) = \mathrm{Var}(X)$.

  Proof: We use $E[\alpha] = \alpha$ and the linearity of E to obtain

  $$\begin{aligned}
  \mathrm{Var}(X+\alpha) &= E[(X+\alpha)^2] - E[X+\alpha]^2 \\
  &= E[X^2 + 2\alpha X + \alpha^2] - (E[X] + \alpha)^2 \\
  &= \left(E[X^2] + \cancel{2\alpha E[X]} + \cancel{\alpha^2}\right) - \left(E[X]^2 + \cancel{2\alpha E[X]} + \cancel{\alpha^2}\right) \\
  &= E[X^2] - E[X]^2 \\
  &= \mathrm{Var}(X).
  \end{aligned}$$

  The idea behind this is that shifting a random variable to the left or right doesn't change the spread:



- For any constant $\alpha \in \mathbb{R}$ and random variable X we have $\mathrm{Var}(\alpha X) = \alpha^2 \mathrm{Var}(X)$.

  Proof: We use the fact that $E[\alpha X] = \alpha E[X]$ to get

$$\begin{aligned}
\mathrm{Var}(\alpha X) &= E[(\alpha X)^2] - E[\alpha X]^2 \\
&= E[\alpha^2 X^2] - (\alpha E[X])^2 \\
&= \alpha^2 E[X^2] - \alpha^2 E[X]^2 \\
&= \alpha^2 \left(E[X^2] - E[X]^2 \right) \\
&= \alpha^2 \mathrm{Var}(X).
\end{aligned}$$

  I don't really know what this means, but there's no arguing with the math.

///

The notion of "covariance" is constantly used in applied statistics, but you are more likely to see it in a modified form called "correlation." Recall that the variance of a random variable (i.e., the covariance of a random variable with itself) is always non-negative:

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X) = \sigma_X^2 \geqslant 0.$$

However, the covariance of two random variables can sometimes be negative. The most we can say about the covariance in general is the following inequality.

Cauchy-Schwarz Inequality. For all random variables X and Y we have

$$\mathrm{Cov}(X, Y)^2 \leqslant \mathrm{Var}(X)\mathrm{Var}(Y) = \sigma_X^2 \sigma_Y^2.$$

///

I won't bother to prove this, but I do want to note an important consequence. By dividing both sides of the inequality by the positive number $\sigma_X \sigma_Y$ we obtain

$$\begin{aligned} \mathrm{Cov}(X,Y)^2 &\leqslant \sigma_X^2 \sigma_Y^2 \\ \frac{\mathrm{Cov}(X,Y)^2}{\sigma_X^2 \sigma_Y^2} &\leqslant 1 \\ \left(\frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \right)^2 &\leqslant 1 \end{aligned}$$

This implies that the number $\mathrm{Cov}(X,Y)/(\sigma_X \sigma_Y)$, which could possibly be negative, satisfies

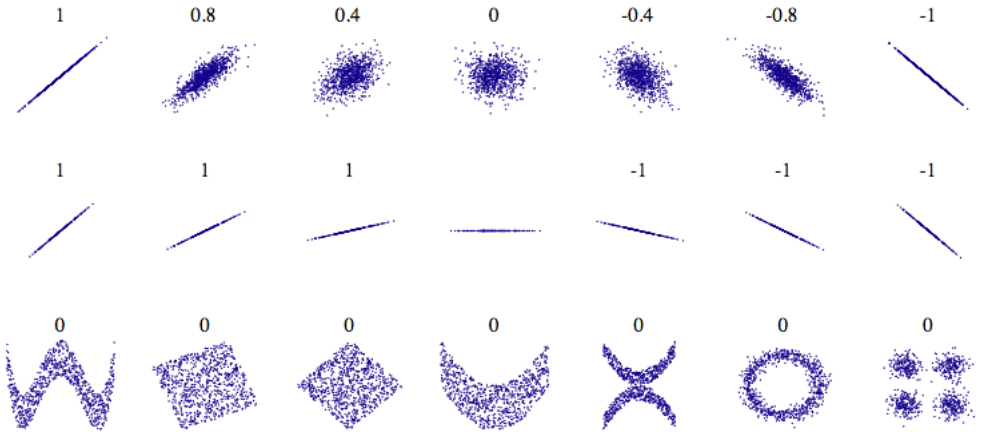
$$-1 \leqslant \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \leqslant +1.$$

Definition of Correlation. For any random variables X and Y we define the *correlation* (also called the *correlation coefficient*) by the formula

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

///

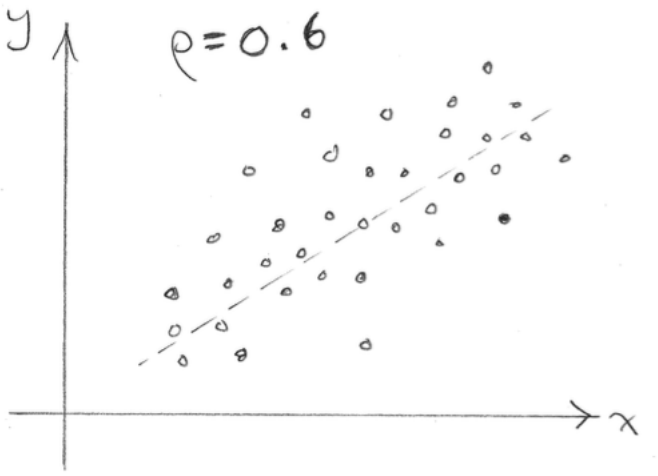
The following picture from Wikipedia will help us to interpret the number ρ_{XY}:



Consider a fixed experiment with sample space S and let $X, Y : S \to \mathbb{R}$ be any two random variables. Suppose that we run the experiment many times and for each run we record the values of X and Y. This will give us a sequence of ordered pairs

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots$$

If we plot these as points in the (x, y)-plane then we will obtain some kind of cloud of dust. The correlation ρ_{XY} is supposed to measure how close our cloud is to being a straight line. For example, if $\rho_{XY} = 0.6$ then we might obtain a picture like this:



A value of ρ_{XY} close to 1 means that our cloud of dust is close to a line of positive slope, and a value of ρ_{XY} close to -1 means that our cloud of dust is close to a line of negative slope.[12] This is demonstrated by the first two rows in the Wikipedia picture.

The third row of the picture is a warning. It says that there are certain interesting kinds of relationships between X and Y that cannot be detected by the number ρ_{XY}. This is closely related to the fact that non-independent random variables might "accidentally" have zero covariance. That is, we have

$$X, Y \text{ independent} \qquad \Longrightarrow \qquad \mathrm{Cov}(X, Y) = 0$$

but

$$\mathrm{Cov}(X, Y) = 0 \qquad \not\Longrightarrow \qquad X, Y \text{ independent.}$$

To detect more subtle relationships between and X and Y we will need more subtle tools than just the covariance and correlation.

We will end this section of the course with a Calculus trick. We might not use it much in this course but it will become very important if you go on to MTH 524/525.

For most random variables X, we can learn anything that we want about the distribution of X by studying the sequence of moments

$$E[X], E[X^2], E[X^3], \ldots.$$

[12] Vertical and horizontal lines correspond to the cases $\sigma_X = 0$ and $\sigma_Y = 0$, respectively. In either case we see that the correlation ρ_{XY} is **not defined** because we can't divide by zero.

For example, we have seen that the first and second moments $E[X]$ and $E[X^2]$ tell us about the center of mass and the spread of X. There is a convenient way to organize this infinite sequence by using Calculus.

Definition of Moment Generating Function (mgf). Let X be a random variable and let $t \in \mathbb{R}$ be a constant. Then we define the *moment generating function* by the formula

$$M_X(t) = E[e^{tX}].$$

///

Why would we do this? Well, you may remember from Calculus that the exponential function has the following infinite series expansion:

$$e^{tX} = 1 + tX + \frac{t^2}{2!}X^2 + \frac{t^3}{3!}X^3 + \cdots$$

Let's assume that this infinite series converges. Then by applying $E[-]$ to both sides and using the fact that t is constant we obtain

$$E[e^{tX}] = 1 + tE[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \cdots .$$

In other words, the moment generating function $M_X(t)$ is a power series in t whose coefficients are basically the moments of the random variable X. Supposing that we have a nice formula for the mgf, it becomes easy[13] to compute the moments.

Theorem (Moment Generating Function). Let X be a random variable with moment generating function $M_X(t)$. Then for any integer $r \geqslant 1$, the rth moment of X is given by the rth derivative of $M_X(t)$ evaluated at $t = 0$:

$$\frac{d^r}{dt^r} M_X(t)\bigg|_{t=0} = E[X^r].$$

Proof: Let's assume that we can bring derivatives inside the expected value. Then we have

$$\frac{d^r}{dt^r} M_X(t) = \frac{d^r}{dt^r} E[e^{tX}] = E\left[\frac{d^r}{dt^r} e^{tX}\right] = E[X^r e^{tX}]$$

and hence

$$\frac{d^r}{dt^r} M_X(t)\bigg|_{t=0} = E[X^r e^{tX}]\big|_{t=0} = E[X^r e^0] = E[X^r].$$

///

[13] for a computer

You should just think of this as another trick for computing moments. Sometimes it works, sometimes it doesn't. It works really well for the binomial distribution.

Application (mgf of a Binomial Random Variable). Let X be a binomial random variable with pmf

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

Then from the binomial theorem and the definition of mgf we have

$$\begin{aligned} M_X(t) = E[e^{tX}] &= \sum_k e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_k \binom{n}{k} \left(e^t p\right)^k (1-p)^{n-k} \\ &= (e^t p + 1 - p)^n. \end{aligned}$$

To compute the first moment we first differentiate:

$$\frac{d}{dt} M_X(t) = \frac{d}{dt}(e^t p + 1 - p)^n = n(e^t p + 1 - p)^{n-1} e^t p$$

And then we evaluate at $t = 0$ to obtain

$$E[X] = \left.\frac{d}{dt} M_X(t)\right|_{t=0} = n(e^0 p + 1 - p)^{n-1} e^0 p = n(1)^{n-1} 1p = np.$$

To compute the second moment we differentiate again (using the product rule):

$$\begin{aligned} \frac{d^2}{dt^2} M_X(t) &= \frac{d}{dt}\left[n(e^t p + 1 - p)^{n-1} e^t p\right] \\ &= e^t p \cdot \frac{d}{dt}\left[n(e^t p + 1 - p)^{n-1}\right] + n(e^t p + 1 - p)^{n-1} \cdot \frac{d}{dt}\left[e^t p\right] \\ &= e^t p n(n-1)(e^t p + 1 - p)^{n-2} e^t p + n(e^t p + 1 - p)^{n-1} e^t p \end{aligned}$$

And then we evaluate at $t = 0$ to obtain

$$E[X^2] = \left.\frac{d^2}{dt^2} M_X(t)\right|_{t=0} = 1pn(n-1)(1)^{n-2} 1p + n(1)^{n-1} 1p = n(n-1)p^2 + np.$$

Finally, we conclude that

$$\begin{aligned} \mathrm{Var}(X) = E[X^2] - E[X]^2 &= \left[n(n-1)p^2 + np\right] - (np)^2 \\ &= n(n-1)p^2 + np - n^2 p^2 \\ &= \cancel{n^2 p^2} - np^2 + np - \cancel{n^2 p^2} \\ &= np - np^2 \\ &= np(1-p). \end{aligned}$$

I don't consider that fun, nor do I think it's the best way to compute the variance of a binomial. But it works. The main advantage of the mgf technique is that it allows us "easy" access to the higher moments. For example, my computer spit out the following information in a milisecond:

$$E[X^3] = np(1 - 3p + 2p^2 + 3np - 3np^2 + n^2p^2).$$

I'm not really sure what that's good for. ///

On second thought, I probably didn't need to introduce the mgf at this moment in this course. (I was following the advice of a fairly bad textbook.) Oh well. There are one or two problems involving the mgf on HW4 and then you can forget about the concept until MTH 524/525. Yes, this means that there will be no problems about moment generating functions on Exam2.

Nov 2

We discussed the solutions to HW4 and then I promised that I would create a list of topics/formulas that you can use to study for Exam2. Here it is:

- **Random Variable.** Consider a fixed experiment with sample space S. A *random variable* is any function

  $$X : S \to \mathbb{R}$$

  that assigns a real number to each outcome of the experiment. Example: Flip a coin 3 times. The sample space is

  $$S = \{TTT, HTT, THT, TTH, HHT, HTH, THH, HHH\}.$$

  Let X be "number of heads squared minus number of tails." Then we have

$s \in S$	TTT	HTT	THT	TTH	HHT	HTH	THH	HHH
$X(s)$	-3	-1	-1	-1	3	3	3	9

  The *support* of X is the set of possible values $S_X \subseteq \mathbb{R}$ that X can take. In our example, $S_X = \{-3, -1, 3, 9\}$.

- **PMF.** The *probability mass function* of a random variable X is the function $f_X : S_X \to \mathbb{R}$ defined by

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

  We can express it with a table or (sometimes) with a formula. If the coin in our example is **fair** them the pmf of our random variable X is

k	-3	-1	3	9
$P(X = k)$	$1/8$	$3/8$	$3/8$	$1/8$

More generally, if the coin satisfies $P(H) = p$ then our pmf becomes

k	-3	-1	3	9
$P(X=k)$	$(1-p)^3$	$3p(1-p)^2$	$3p^2(1-p)$	p^3

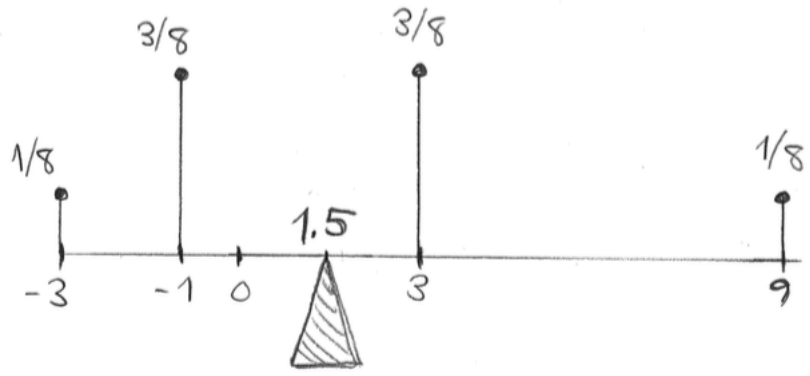
- **Expected Value.** We define the *expected value* or *mean* of a random variable X by the following two equivalent[14] formulas:

$$\mu_X = E[X] = \sum_{s \in S} X(s)P(s) = \sum_{k \in S_X} kP(X=k).$$

For our example random variable X with a **fair** coin we compute

$$E[X] = \sum_k kP(X=k) = -3 \cdot \frac{1}{8} - 1 \cdot \frac{3}{8} + 3 \cdot \frac{3}{8} + 9 \cdot \frac{1}{8} = \frac{12}{8} = 1.5.$$

Meaning 1: If we view the pmf as a distribution of "point masses" on a line, then the expected value is the "center of mass":



Meaning 2: If we perform the experiment many times and compute the average value of X, we expect to get 1.5.

- **Linearity of Expectation.** Let $X, Y : S \to \mathbb{R}$ be random variables and let $\alpha, \beta \in \mathbb{R}$ be constants. Then

$$\begin{aligned} E[\alpha] &= \alpha, \\ E[\alpha X] &= \alpha E[X], \\ E[X+Y] &= E[X] + E[Y], \\ E[\alpha X + \beta Y] &= \alpha E[X] + \beta E[Y]. \end{aligned}$$

We say that the function $E[-]$ is *linear*.

[14] You do not need to prove that they are equivalent.

- **Variance and Standard Deviation.** We define the *variance* as the expected value of $(X - \mu_X)^2$, i.e., the square of the distance between X and its mean:

$$\mathrm{Var}(X) = E[(X - \mu_X)^2].$$

Using the linearity of $E[-]$ we can also show that

$$\mathrm{Var}(X) = E[X^2] - \mu_X^2 = E[X^2] - E[X]^2.$$

We define the *standard deviation* as the non-negative square root of the variance:

$$\sigma_X = \sqrt{\mathrm{Var}(X)}.$$

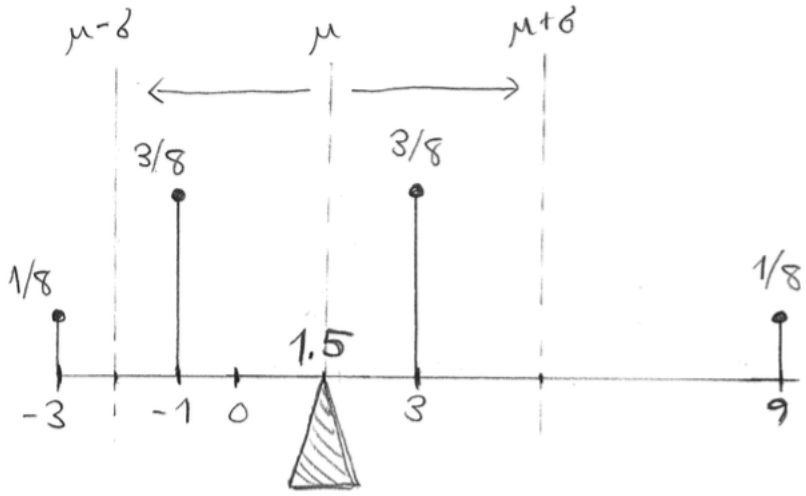
This makes sense because the variance is always non-negative. In our example we have

k	-3	-1	3	9
k^2	9	1	9	81
$P(X = k)$	$1/8$	$3/8$	$3/8$	$1/8$

and hence

$$E[X^2] = 9 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 9 \cdot \frac{3}{8} + 81 \cdot \frac{1}{8} = \frac{12}{8} = 15,$$
$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = 15 - (1.5)^2 = 12.75,$$
$$\sigma_X = \sqrt{12.75} \approx 3.57.$$

I added this information to the picture:



- **Basic Properties of Variance.** For $X : S \to \mathbb{R}$ and $\alpha \in X$ we have

$$\mathrm{Var}(\alpha) = 0,$$
$$\mathrm{Var}(X + \alpha) = \mathrm{Var}(X),$$
$$\mathrm{Var}(\alpha X) = \alpha^2 \mathrm{Var}(X).$$

- **Covariance and Correlation.** The function of $\mathrm{Var}(-)$ is **not linear**. Let $X, Y : S \to \mathbb{R}$ be random variables. Then in general we have

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X,Y)$$

where the *covariance* is defined by

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

We observe that $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$. Unlike the variance, however, **the covariance may be negative**. The most we can say in general is that

$$\mathrm{Cov}(X,Y)^2 \leqslant \sigma_X^2 \sigma_Y^2.$$

This implies that the *correlation* $\rho_{XY} = \mathrm{Cov}(X,Y)/(\sigma_X\sigma_Y)$ is between -1 and $+1$:

$$-1 \leqslant \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} \leqslant +1.$$

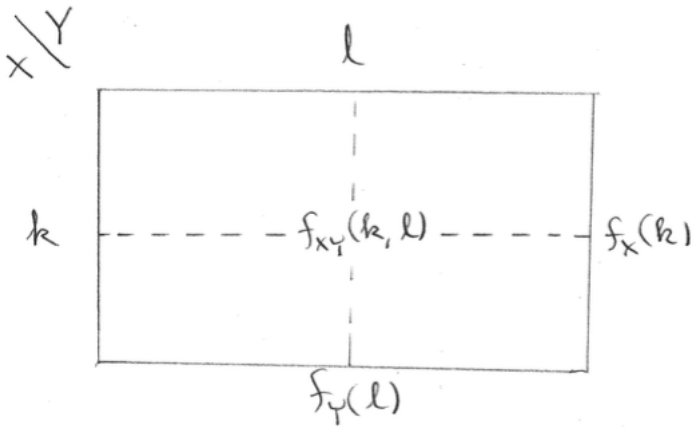
- **Joint PMF and Independence.** Let $X, Y : S \to \mathbb{R}$ be random variables. For all $k \in S_X$ and $\ell \in S_Y$ we define the function

$$f_{XY}(k,\ell) = P(X=k, Y=\ell) = P(X=k \text{ and } Y=\ell),$$

which is called the *joint pmf* of X and Y. We can recover the *marginal pmfs* of X and Y by summing:

$$\begin{array}{rcl} P(X=k) & = & \sum_\ell P(X=k, Y=\ell) \\ f_X(k) & = & \sum_\ell f_{XY}(k,\ell) \end{array} \quad \text{and} \quad \begin{array}{rcl} P(Y=\ell) & = & \sum_k P(X=k, Y=\ell) \\ f_Y(\ell) & = & \sum_k f_{XY}(k,\ell) \end{array}$$

Sometimes the pmfs have formulas, sometimes not. We like to display the joint and marginal pmfs with a table:



We say that X and Y are *independent* if for all values of k and ℓ we have

$$f_{XY}(k,\ell) = f_X(k) f_Y(\ell)$$
$$P(X = k, Y = \ell) = P(X = k)P(Y = \ell).$$

If there is even one cell of the table where this doesn't happen, we say that X and Y are *dependent*. If X and Y are **independent** then we have $\mathrm{Cov}(X,Y) = 0$, which implies

$$E[XY] = E[X] \cdot E[Y] \qquad \text{and} \qquad \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

For dependent X and Y these formulas may be false.

Statisticians have a small collection of random variables that they like to use. (Just look inside the cover of the textbook.) Here are the ones that we know.

- **Bernoulli RV.** Flip a coin and let $X = 1$ if you get heads and $X = 0$ if you get tails. Assuming that $P(H) = p$ we get the following pmf:

k	0	1
$P(X = k)$	$1-p$	p

Some people express this pmf as a formula: $P(X = k) = \binom{1}{k} p^k (1-p)^{1-k}$. But I think that's kind of silly. We can easily compute the mean and variance:

$$E[X] = 0(1-p) + 1p = p,$$
$$E[X^2] = 0^2(1-p) + 1^2 p = p,$$
$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = p^2 - p = p(1-p).$$

- **Binomial RV.** Flip a coin n times and let X be the number of heads that you get. This pmf has a famous formula:

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

There are various ways to compute the mean and variance of X, some of which are quite ugly. Here's a nice way: Let X_i be the number of heads that you get on the i-th flip, which is a Bernoulli random variable. Then we have

$$\begin{aligned} X &= X_1 + X_2 + \cdots + X_n \\ E[X] &= E[X_1] + E[X_2] + \cdots + E[X_n] \\ &= p + p + \cdots + p \\ &= np, \end{aligned}$$

and since the coin flips are **independent** we have

$$X = X_1 + X_2 + \cdots + X_n$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\\ &= p(1-p) + p(1-p) + \cdots + p(1-p)\\ &= np(1-p).\end{aligned}$$

- **Geometric RV.** Consider a coin with $P(H) = p$ and $P(T) = q$. Start flipping the coin and let X be the number of flips until you see H. If $X = k$, i.e., if H occurs for the first time on the k-th flip, then our sequence of flips must be $TT\cdots TH$. Thus,

$$\begin{aligned}P(X = k) &= P(TT\cdots TH)\\ &= P(T)P(T)\cdots P(T)P(H)\\ &= qq\cdots qp\\ &= q^{k-1}p.\end{aligned}$$

  The mean and variance are tricky to compute,[15] but here they are:

$$E[X] = \frac{1}{p} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}.$$

  To do anything with the geometric random variable you must remember the *geometric series* from Calculus:

$$\boxed{1 + q + q^2 + \cdots = \frac{1}{1-q} = \frac{1}{p}.}$$

  Then, for example, we can show for any non-negative integer ℓ that

$$\begin{aligned}P(X > \ell) = \sum_{k=\ell+1}^{\infty} P(X = k) &= \sum_{k=\ell+1}^{\infty} q^{k-1}p\\ &= q^\ell p + q^{\ell+1}p + q^{\ell+2}p + \cdots\\ &= q^\ell p \cdot \left[1 + q + q^2 + \cdots\right]\\ &= q^\ell p \cdot \frac{1}{p}\\ &= q^\ell.\end{aligned}$$

  This implies that $P(X > 0) = q^0 = 1$, as it should be.

- **Hypergeometric RV.** Everything so far has been based on a sequence of independent "coin flips." The hypergeometric random variable, on the other hand, is based on a sequence of **non-independent trials**.

  Consider a bowl containing B blue balls and R red balls. Suppose you reach in and grab n ordered balls. Define $X_i = 1$ if the i-th ball is blue and $X_i = 0$ if the i-th ball is red. Let $X = X_1 + X_2 + \cdots + X_n$ be the total number of blue balls that you get.

[15] One could use the moment generating function, but moment generating functions are not on Exam2.

Each X_i is a Bernoulli random variable with $P(X_i = 1) = B/(R + B)$ so we obtain

$$\begin{aligned} X &= X_1 + X_2 + \cdots + X_n \\ E[X] &= E[X_1] + E[X_2] + \cdots + E[X_n] \\ &= \frac{B}{B+R} + \frac{B}{B+R} + \cdots + \frac{B}{B+R} \\ &= \frac{nB}{B+R}. \end{aligned}$$

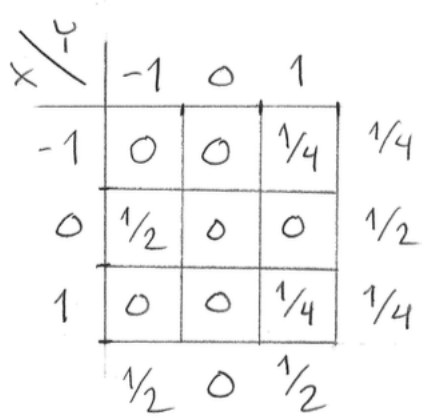
However, these X_i are **not independent**. (For example, if the 1st ball is blue, this *decreases* the chance that the 2nd ball is blue.) To compute the pmf it is easiest to consider the n balls as unordered.[16] Then every collection of n unordered balls is equally likely, so we have

$$P(X = k) = \frac{(\#\text{ collections with } k \text{ blue balls})}{(\text{total } \#\text{ collections})} = \frac{\binom{B}{k}\binom{R}{n-k}}{\binom{B+R}{n}}.$$

The variance $\mathrm{Var}(X)$ and covariances $\mathrm{Cov}(X_i, X_j)$ are tricky so we won't discuss them.

- **Worked Example of Covariance.** If X and Y are independent then we always have $\mathrm{Cov}(X, Y) = 0$, but it doesn't work the other way around. Here is a strange example of **dependent** random variables X, Y such that $\mathrm{Cov}(X, Y) = 0$.

  Consider the following joint pmf table:



  We observe that X and Y are **not independent** because, for example, the joint probability $P(X = -1, Y = -1) = 0$ is not equal to the product of the marginal probabilities:$P(X = -1)P(Y = -1) = (1/4)(1/2) = 1/8 \neq 0$. We will show, however, that X and Y are "uncorrelated," i.e., that $\mathrm{Cov}(X, Y) = \rho_{XY} = 0$.

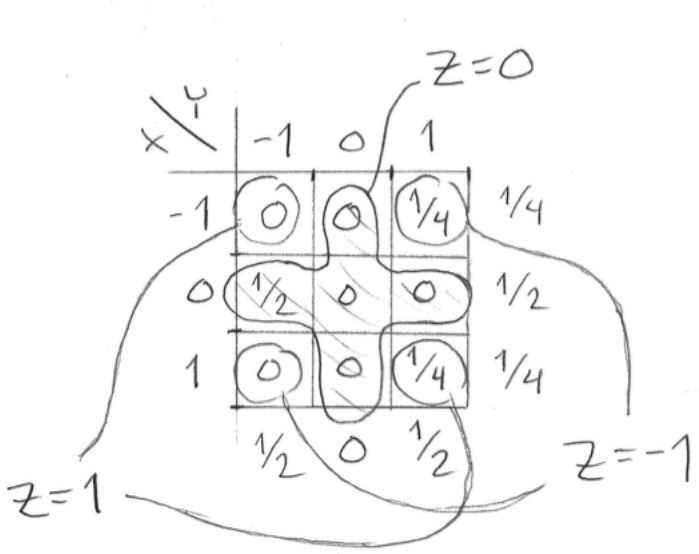
  To do this we will use the formula $\mathrm{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y]$. First, note that

$$E[X] = (-1)(1/4) + (0)(1/2) + (1)(1/4) = 0,$$

[16] We can leave them ordered, but this makes the counting problem harder.

$$E[Y] = (-1)(1/2) + (0)(0) + (1)(1/2) = 0.$$

Now define the random variable $Z = XY$ and observe that its support is $S_Z = \{-1, 0, 1\}$. We have circled the events "$Z = -1$," "$Z = 0$" and "$Z = 1$" in the following picture:



By adding the probabilities inside each blob we obtain the pmf of Z:

k	-1	0	1
$P(Z = k)$	$1/4$	$1/2$	$1/4$

We conclude that

$$E[XY] = E[Z] = (-1)(1/4) + (0)(1/2) + (1)(1/4) = 0$$

and hence

$$\mathrm{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y] = 0 - 0 \cdot 0 = 0.$$

Warning: This was a carefully chosen example. It's actually pretty difficult to find uncorrelated random variables that are not independent.

///