Notes for Recitation 13 (5/8/98)

1 Covariance

We saw in class that the variance of a random variable is a measure of how it deviates from the mean. *Covariance* is another statistical measure. It quantifies the similarity of the distributions of two random variables.

Definition 1.1. The covariance of random variables R_1 and R_2 is

$$\mathrm{Cov}\left[R_1, R_2\right] ::= \mathrm{E}\left[R_1 R_2\right] - \mathrm{E}\left[R_1\right] \mathrm{E}\left[R_2\right].$$

If the covariance of R_1 and R_2 is positive, we will say that they are "positively correlated", and if their covariance is negative we say that they are "negatively correlated". Intuitively, positively correlated random variables tend to have close values, while negatively correlated variables tend to have values far apart. Independent random variables have zero correlation.

Examples: Let R_1 and R_2 be uniform in the range $[1, 6]$.

1. $R_1 = R_2$. In this case,

   $$\mathrm{Cov}\left[R_1, R_2\right] = \mathrm{E}\left[R_1^2\right] - \mathrm{E}^2\left[R_1\right] = \mathrm{Var}\left[R_1\right] = 2\frac{11}{12}.$$

   The variables are positively correlated. This is not a surprise — they are always equal.

2. $R_2 = 7 - R_1$. Now,

   $$\begin{aligned}
   \mathrm{Cov}\left[R_1, R_2\right] &= \mathrm{E}\left[R_1(7 - R_1)\right] - \mathrm{E}\left[R_1\right] \mathrm{E}\left[7 - R_1\right] \\
   &= 7\mathrm{E}\left[R_1\right] - \mathrm{E}\left[R_1^2\right] - 7\mathrm{E}\left[R_1\right] + \mathrm{E}^2\left[R_1\right] \\
   &= -\mathrm{E}\left[R_1^2\right] + \mathrm{E}^2\left[R_1\right] \\
   &= -\mathrm{Var}\left[R_1\right] = -2\frac{11}{12}
   \end{aligned}$$

   The variables are negatively correlated, i.e., when one is high the other is low. This is reflected by the negative covariance.

3. R_1 are R_2 are independent (e.g., consecutive coin tosses). Then

$$\mathrm{Cov}\left[R_1, R_2\right] = \mathrm{E}\left[R_1 R_2\right] - \mathrm{E}\left[R_1\right]\mathrm{E}\left[R_2\right] = \mathrm{E}\left[R_1\right]\mathrm{E}\left[R_2\right] - \mathrm{E}\left[R_1\right]\mathrm{E}\left[R_2\right] = 0.$$

   The correlation of independent random variables is zero.

So if the variables are independent then, indeed their covariance is zero. What about the converse? If the covariance is zero, are the variables independent? The answer is no. For example, let R_1, R_2 be in the range $[1,3]$ such that $\Pr\left[R_1 = i \& R_2 = j\right] = 1/8$ for all (i,j) except $(2,2)$ and $\Pr\left[R_1 = R_2 = 2\right] = 0$. We can check that this gives valid probability distributions for R_1 and R_2:

$$\Pr\left[R_1 = 1\right] = \frac{3}{8},\ \Pr\left[R_1 = 2\right] = \frac{2}{8},\ \Pr\left[R_1 = 3\right] = \frac{3}{8},$$

which add up to 1.

The distribution of R_1 is the same as the distribution of R_2 but since $\Pr\left[R_1 = 2 \mid R_2 = 2\right] = 0 \neq \Pr\left[R_1 = 2\right]$, R_1 and R_2 are not independent. However,

$$\begin{aligned}
\mathrm{Cov}\left[R_1, R_2\right] &= \mathrm{E}\left[R_1 R_2\right] - \mathrm{E}\left[R_1\right]\mathrm{E}\left[R_2\right] \\
&= \sum_{i,j\in[1,3]} ij \Pr\left[R_1 = i \& R_2 = j\right] - \Big(\sum_{i\in[1,3]} i \Pr\left[R_1 = i\right]\Big)^2 \\
&= \frac{1}{8}(1+2+3+2+6+3+6+9) - \left(\frac{3}{8}(1+3) + \frac{2}{8}2\right)^2 \\
&= 4 - 4 = 0
\end{aligned}$$

Theorem 1.2. For random variables R_1 and R_2,

$$\mathrm{Var}\left[R_1 + R_2\right] = \mathrm{Var}\left[R_1\right] + \mathrm{Var}\left[R_2\right] + 2\mathrm{Cov}\left[R_1, R_2\right].$$

Proof.

$$\begin{aligned}
\mathrm{Var}\left[R_1 + R_2\right] &= \mathrm{E}\left[(R_1 + R_2)^2\right] - \mathrm{E}^2\left[R_1 + R_2\right] \\
&= \mathrm{E}\left[R_1^2 + 2R_1R_2 + R_2^2\right] - \left(\mathrm{E}\left[R_1\right] + \mathrm{E}\left[R_2\right]\right)^2 \\
&= \mathrm{E}\left[R_1^2\right] + 2\mathrm{E}\left[R_1R_2\right] + \mathrm{E}\left[R_2^2\right] - \mathrm{E}^2\left[R_1\right] - 2\mathrm{E}\left[R_1\right]\mathrm{E}\left[R_2\right] - \mathrm{E}^2\left[R_2\right] \\
&= \left(\mathrm{E}\left[R_1^2\right] - \mathrm{E}^2\left[R_1\right]\right) + \left(\mathrm{E}\left[R_2^2\right] - \mathrm{E}^2\left[R_2\right]\right) + 2(\mathrm{E}\left[R_1R_2\right] - \mathrm{E}\left[R_1\right]\mathrm{E}\left[R_2\right]) \\
&= \mathrm{Var}\left[R_1\right] + \mathrm{Var}\left[R_2\right] + 2\mathrm{Cov}\left[R_1, R_2\right]
\end{aligned}$$

□

2 Variance of Sums

As a corollary to the previous theorem we have:

Corollary 2.1. If R_1 and R_2 are independent, then

$$\mathrm{Var}\left[R_1 + R_2\right] = \mathrm{Var}\left[R_1\right] + \mathrm{Var}\left[R_2\right]$$

Proof. If R_1, R_2 are independent, then $\mathrm{Cov}\left[R_1, R_2\right] = 0$. □

A similar result holds for the sum of many independent random variables — but only needs that the variables are *pairwise* independent.

Theorem 2.2. If $R_1, \ldots, R_n$ are pairwise independent random variables, then

$$\mathrm{Var}\left[R_1 + \cdots + R_n\right] = \mathrm{Var}\left[R_1\right] + \cdots + \mathrm{Var}\left[R_n\right].$$

The proof is simple and is left to the reader. (Calculations will be easier if one defines new random variables $D_i ::= R_i - \mathrm{E}\left[R_i\right]$. Then $\mathrm{Var}\left[\sum_i R_i\right] = \mathrm{E}\left[(\sum_i D_i)^2\right]$, $\mathrm{E}\left[D_i\right] = 0$, and $\mathrm{Var}\left[R_i\right] = \mathrm{E}\left[D_i^2\right]$, which will simplify the expressions.)

Using this result we can now derive the variance of a binomial distribution. Let R have an n, p binomial distribution, i.e.,

$$\Pr\left[R = k\right] = \binom{n}{k} p^k (1-p)^{n-k}.$$

We can think of R as the sum of n Bernoulli independent random variables. That is, the number of heads in n tosses with the probability of heads equal to p:

$$R = R_1 + R_2 + \ldots + R_n, \ \Pr\left[R_i = 1\right] = p, \ \text{and} \ \Pr\left[R_i = 0\right] = 1 - p.$$

Now

$$\mathrm{Var}\left[R_i\right] = \mathrm{E}\left[R_i^2\right] - \mathrm{E}^2\left[R_i\right] = p - p^2.$$

Hence from the theorem,

$$\mathrm{Var}\left[R\right] = \sum_{i=1}^{n} \mathrm{Var}\left[R_i\right] = n\mathrm{Var}\left[R_i\right] = p(1-p)n.$$

3 Wald's Theorem

In this section we will prove Wald's Theorem, that was stated without proof in lecture. The proof will make use of the following observation.

Lemma 3.1. Let R be a random variable with range in the non-negative integers, then

$$\mathrm{E}[R] = \sum_{k \geq 0} \Pr[R \geq k].$$

Proof. We will begin with the right-hand expression and transform it into $\mathrm{E}[R]$.

$$\begin{aligned}
\sum_{i=0}^{\infty} \Pr[R > i] &= \Pr[R > 0] + \Pr[R > 1] + \Pr[R > 2] + \ldots \\
&= \underbrace{\Pr[R = 1] + \Pr[R = 2] + \Pr[R = 3] + \ldots}_{=\Pr[R>0]} \\
&\quad + \underbrace{\Pr[R = 2] + \Pr[R = 3] + \ldots}_{=\Pr[R>1]} \\
&\quad + \underbrace{\Pr[R = 3] + \ldots}_{=\Pr[R>2]} \\
&= \Pr[R = 1] + 2 \cdot \Pr[R = 2] + 3 \cdot \Pr[R = 3] + \ldots \\
&= \sum_{i=0}^{\infty} i \cdot \Pr[R = i] \\
&= \mathrm{E}[R]
\end{aligned}$$

In the first step, the summation is rewritten with the $\ldots$ notation. In the second step, we rewrite each term $\Pr[R > i]$ as a series; if R is greater than i, then it must be one of $i + 1$, $i + 2$, etc. (This is where we use the fact that the range of R is the set of non-negative integers.) Next, we sum all the resulting series, and then revert to $\sum$ notation. The last summation is the definition of $\mathrm{E}[R]$. □

We recall the statement of Wald's Theorem[1]

[1]Here's a better version of Wald's theorem than the one presented in lecture and recitation this term:

Theorem 3.2 (Better Version of Wald's Theorem). Let Q be a non-negative integer valued random variable such that $\mathrm{E}[Q]$ is finite, and let $R_1, R_2, \ldots$ be a countable sequence of random variables. Let T be the sum $R_1 + R_2 + \cdots + R_Q$. Suppose that, conditioned on being included in the sum, every variable R_k has the same distribution as R_1, and also suppose that $\mathrm{E}[|R_1|]$ is finite.

Theorem 3.3 (Wald's Theorem). Let Q be a random variable that takes on only non-negative, integer values and such that $\mathrm{E}[Q]$ is finite. Let $R_1, R_2, \ldots$ be a sequence of random variables, all with the same distribution and with $\mathrm{E}[|R_1|]$ finite. Let T be the sum $R_1 + R_2 + \ldots + R_Q$. Suppose that every variable R_k is independent of the event that it is included in the sum. Then,

$$\mathrm{E}[T] = \mathrm{E}[Q] \cdot \mathrm{E}[R_1].$$

Technical conditions abound in the previous theorem. For example, in simple applications such as in the one presented in lecture, R_k is completely independent of Q. The more weaker, technical condition merely that R_k is "independent of the event that it is included in the sum" is important in other common applications, cf., Team Problem 2 below.

We let I_k be an indicator variable for the event that R_k "is included in the sum $R_1 + \ldots + R_Q$", that is, we define I_k as follows:

$$I_k ::= \begin{cases} 1 & \text{if } Q \geq k, \\ 0 & \text{if } Q < k. \end{cases}$$

Then the precise condition on the R_k's is that, for any k, R_k is independent of I_k.

Proof. The indicator variables I_k play an important role in this proof. First of all, we can rewrite T as shown below:

Then, $\mathrm{E}[T] = \mathrm{E}[Q] \cdot \mathrm{E}[R_1]$.

More precisely, the condition "conditioned on being included in the sum, every variable R_k has the same distribution," means

$$\Pr[R_k = r \mid Q \geq k] = \Pr[R_1 = r \mid Q \geq 1]$$

for all integers $k \geq 1$ and real numbers r. Note that $\Pr[R_1 = r \mid Q \geq 1] = \Pr[R_1 = r]$, since $Q \geq 1$ by hypothesis.

This version has the advantage that it does not put any constraints on the behavior of R_k when R_k is *not* included in the sum—which makes better sense since the behavior of R_k when it is not included in the sum is not relevant to the conclusion of the theorem. So R_k can be defined arbitrarily—say to be zero—if it is not in the sum. This is natural in many examples such as Team Problem 2 where R_k describes results of the kth in a series of trials, and Q is defined by a stopping condition on the values of the R_k's; in this case R_k isn't really defined when the experiment stops before the kth trial.

The proof is actually simpler with $R_k = 0$ when it is not included in the sum, since then

$$\sum_{k=1}^{Q} R_k = \sum_{k=1}^{\infty} R_k.$$

A further technical advantage of this version is that with R_k unconstrained when it is not in the sum, we can restrict to sample points consisting solely of the stopped experiments, thereby modeling the series of trials with a countable sample space as in Handout 54.

$$\begin{aligned} T &= R_1 + R_2 + \ldots + R_Q \\ &= \sum_{k=1}^{\infty} R_k I_k \end{aligned}$$

The effect of the indicator variables is to cancel out all but the first Q terms in the infinite sum. We can now compute the expected value of T as follows:

$$\mathrm{E}[T] = \mathrm{E}\left[\sum_{k=1}^{\infty} R_k I_k\right]$$

On the right, we have the expectation of an infinite sum, and we would like to apply linearity of expectation. Recall that linearity of expectation holds also for infinite sums, provided that the sum is absolutely convergent with probability one. But for any point of our sample space, Q takes some finite value q, and only the indicator variables $I_1, \ldots, I_q$ are non-zero; therefore, for every point of the sample space all but finitely many terms of the infinite sum are zero. So it's OK to use the infinite version of the linearity of expectation:

$$= \sum_{k=1}^{\infty} \mathrm{E}[R_k I_k]$$

The third condition of Wald's Theorem is precisely that R_k and I_k are independent. Therefore, the expectation of the product is the product of the expectations:

$$= \sum_{k=1}^{\infty} \mathrm{E}[R_k]\,\mathrm{E}[I_k]$$

The expected value of R_k is equal to the expected value of R_1 by the second condition of Wald's Theorem. Since it is independent of k, we can pull the term $\mathrm{E}[R_1]$ outside the summation:

$$= \mathrm{E}[R_1] \sum_{k=1}^{\infty} \mathrm{E}[I_k].$$

The expected value of an indicator variable like I_k is the probability that it is 1, and, by the definition of I_k, we have that $I_k = 1$ if and only if $Q \geq k$, so that

$$= \mathrm{E}[R_1] \sum_{k=1}^{\infty} \Pr[Q \geq k].$$

Since Q is a non-negative, integer-valued random variable, the remaining sum is equal to $\mathrm{E}[Q]$ by Lemma 3.1.

$$= \mathrm{E}[R_1] \cdot \mathrm{E}[Q].$$

□

As an application of Wald's Theorem, here's a simple proof that

Corollary 3.4. The mean time to failure is the reciprocal of the probability of failure.

Namely, consider flipping a weighted coin until a head shows up. Assume the flips are independent. Then

$$\mathrm{E}[\text{number of flips to first head}] = \frac{1}{\Pr[\text{heads on one flip}]}.$$

Proof. Assume the probability of a head is h. Regard flipping until a head shows up as a game in which each tail flipped pays nothing and each head pays \$1.00. We know from Handout 54 that the probability of eventually flipping a head is one, so with probability one, a game pays off \$1.00. It follows that the *expected* win obviously must also be \$1.00. But by Wald's Theorem, the expected win is the expected number of flips until a head turns up, times the expected number, h, of heads per flip. That is,

$$1 = \mathrm{E}[\text{number of flips to first head}] \cdot h.$$

□

4 Team Problems

Problem 1. The Petersburg Casino offers the following game: the gambler bets a fixed wager, and then the dealer flips a fair coin (dealers do not flip coins in US casinos, but they do in Petersburg) until it comes up heads. The gambler receives \$1 if the coin shows heads the first time, \$2 if it shows the first head at second toss, and in general $\$2^{k-1}$ if the dealer tosses the coin k times to get the first head. The fixed wager is \$100,000.

(a) What is the expected amount of money that the Petersburg Casino is going to pay to a gambler that plays this game? What about the variance? Is it reasonable for the gambler to risk \$100,000 dollars in a game where there is a 50% chance of getting only *one dollar* back (this is not really a mathematical question: just think about that.)

(b) What is the probability that the gambler does not lose money in this game, i.e. what is the probability that he gets back from the Casino more than \$100,000?

Problem 2. A Harvard student, lost in the underground corridors of MIT, finds himself in a room containing three doors. Door 1 leads him to freedom (out of MIT) after an hour of travel; door 2 returns him to the same room after four hours travel, and door 3 returns him to the room after seven hours travel. Each time he finds himself in the room, the Harvard student picks a door with equal probability. Let T denote the time until he finds the exit.

(a) Define a sequence of independent identically distributed random variables R_i and a stopping rule Q such that

$$T = \sum_{i=1}^{Q} R_i$$

(b) Argue that R_k is independent of the event that $Q \geq k$. Is R_k independent of Q?

(c) Use Wald's Theorem to find $\mathrm{E}[T]$.

(d) Now suppose that an MIT student is also lost and finds herself in a similar situation to the Harvard student. The MIT student never chooses the same door twice. What is $\mathrm{E}[T]$ for the MIT student?

Problem 3. If a random variable has infinite expectation, then the variance is undefined, however a random variable may have finite expectation and infinite variance.

(a) Show an example of a random variable whose expectation is finite but whose variance is infinite.
[It might be easier to use a random variable whose expectation is zero.]

(b) If you used a random variable which can assume negative values in answering the previous question, try to find an example of a *non-negative* random variable whose expectation is finite and whose variance is infinite.
[Hint: notice that $\sum_{n=1}^{\infty} 1/2^{2n} = 1/3$ and $\sum_{n=1}^{\infty} 1/2^n = 1$.]

Problem 4. Let X be a non-negative random variable with expectation $n/2$. Markov's inequality implies that

$$\Pr\left[X \geq \frac{2}{3}n\right] \leq \frac{3}{4}. \tag{1}$$

(a) Define a random variable X with expectation $n/2$ such that the above bound holds with equality.

(b) If $X_1, \ldots, X_n$ are a collection of *independent* unbiased Bernoulli variables over the same sample space, then $X = \sum_i X_i$ has expectation $n/2$ and is non-negative, therefore the bound (1) above applies. Unfortunately it is a much larger bound than necessary in this case. However, under the general hypotheses for Markov's inequality—where independence is not assumed—the bound (1) is still the best one can hope for.

Define unbiased Bernoulli variables $X_1, \ldots, X_n$ on a sample space such that for $X = \sum_{i=1}^{n} X_i$, the bound (1) holds with equality (assume n is a multiple of 3). *Hint*: The variables will have to be dependent. Define things so that $\Pr[X_1 = \cdots = X_n = 0] = 1/4$.

5 Solutions

Problem 1 Solution

Let V be the random variable corresponding to the amount of money that the gambler is paid by the dealer. The distribution of V is as follows: for any $n \geq 1$,

$$\Pr\left[V = 2^{n-1}\right] = 2^{-n},$$

where the probability refers to the event that the dealer tosses $n-1$ tails followed by a head. The average of V is

$$\mathrm{E}[V] = \sum_{n=1}^{\infty}[2^{-n}2^{n-1}] = \sum_{n=1}^{\infty} 1/2 = \infty.$$

The variance is not defined for a random variable having infinite expectation. It does not sound too bad to risk \$100,000 on a game where we expect to win an infinite amount of money. However the probability of recovering the money of the wager is only

$$\begin{aligned}
\Pr[X \geq 100,000] &= \Pr\left[X \geq 2^{17}\right] \\
&= \Pr[\text{ there are 17 consecutive tails }] \\
&= \frac{1}{131,072}
\end{aligned}$$

Problem 2 Solution

We view this an experiment where at each step the student makes a choice between three equally likely outcomes. We define R_i as the number of hours to return to the same room, or to escape, starting at the ith step. The experiment does not end. That is, as soon as the student escapes, he is doomed, like Sisyphus, to be stuck back in the room and start trying to escape again. This ensures that there is an ith step for all $i \geq 0$, and that the R_i are identically distributed (they are also independent, though this fact is not needed in applying Wald's Theorem). In particular, $\Pr[R_i = 1] = \Pr[R_i = 4] = \Pr[R_i = 7] = 1/3$, and $\mathrm{E}[R_i] = (1 + 4 + 7)/3 = 4$.

Define Q to be the number of steps from the first step until the escape door is selected *for the first time*. Now Q is geometrically distributed with probability $1/3$ and $\mathrm{E}[Q] = 3$.

Assuming Wald's Theorem is applicable, we conclude

$$\mathrm{E}[T] = \mathrm{E}\left[\sum_{i=1}^{Q} R_i\right] = 3 \times 4 = 12.$$

To justify the use of Wald's Theorem, we establish that R_i is independent of the event that $Q \geq i$.

It's worth noting that R_i is *not* independent of Q. For example, if $Q = i$, then it must be that the student found the escape door at the ith step, and so $\Pr[R_i = 1 | Q = i] = 1 \neq 1/3 = \Pr[R_i = 1]$.

However, suppose we are given that the student has never managed to escape before the ith step, that is, that $Q \geq i$. Then we can observe that the experiment starting at the ith step has exactly the same structure as the experiment at the first step—where we were given, implicitly, that the student had not escaped before the first step. So we can conclude that the distribution of R_i given $Q \geq i$ is the same as that of R_1 given that $Q \geq 1$, that is, the same as R_1, which we already observed is the same as R_i. In other words, R_i is independent of the event that $Q \geq i$, as required. (This informal argument can be justified rigorously as in Handout 54 on "Probabilities in a Dice Game," by explicitly describing the sample space for the problem along with a suitable bijection between the whole sample space and the event $Q \geq i$.)

Regarding the MIT student, the problem can be solved by complete enumeration; call A the door that leads out, B the door that leads back in four hours, and C the door that leads back in seven hours. The possible events are

- A with probability $1/3$, in which case $T = 1$
- BA with probability $\frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$, in which case $T = 5$;
- BCA with probability $\frac{1}{2} \cdot \frac{1}{3} \cdot 1 = \frac{1}{6}$, in which case $T = 12$;
- CA with probability $\frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$, in which case $T = 8$;
- CBA with probability $\frac{1}{2} \cdot \frac{1}{3} \cdot 1 = \frac{1}{6}$, in which case $T = 12$.

The average of T this time is $\mathrm{E}[T] = 39/6 = 6.5$.

Problem 3 Solution

There are several possible solutions. One possibility (that solves directly also the second question) is to consider a random variable V such that for any $n \geq 1$ we have $\Pr[V = 2^n] ::= 3 \cdot 2^{-2n}$; this is a valid distribution since $\sum_{n=1}^{\infty} 3 \cdot 2^{-2n} = 1$. Then

$$\mathrm{E}[V] = \sum_{n=1}^{\infty} 2^n \cdot 3 \cdot 2^{-2n} = \sum_{n=1}^{\infty} 3 \cdot 2^{-n} = 3.$$

Therefore,

$$\mathrm{Var}[V] = \mathrm{E}\left[V^2\right] - \mathrm{E}^2[V] = \mathrm{E}\left[V^2\right] - 9,$$

and we just have to compute the expectation of V^2. For every $n \geq 1$ we have that $\Pr[V^2 = 2^{2n}] = 3 \cdot 2^{-2n}$, so

$$\mathrm{E}\left[V^2\right] = \sum_{n=1}^{\infty} 2^{2n} \cdot 3 \cdot 2^{-2n} = \sum_{n=1}^{\infty} 3 = \infty.$$

Problem 4 Solution

Consider the sample space where $X_1 = \cdots = X_n = 0$ occurs with probability $1/4$, and where for any $i = 1, \ldots, n$, the event $(X_1, \ldots, X_n) = \mathrm{shift}_k(0^{n/3}1^{2n/3})$ happens with probability $3/4n$. The notation $\mathrm{shift}_k(a_1, \ldots, a_n)$, where $a_1, \ldots, a_n$ is a binary string, stands for the string $(a_{k+1} \cdots a_n a_1 \cdots a_k)$ obtained by doing a k times a cyclic shift to the left of the string $a_1, \ldots, a_n$. Notice that with probability exactly $3/4$ exactly a fraction $2/3$ of the X_i get value 1, therefore expression (1) is satisfied with equality. We have to check that for any i we have $\Pr[X_1 = 1] = 1/2$. This follows from the fact that there are exactly $2n/3$ shifts such that $X_i = 1$, each one happening with probability $3/4n$; since they are disjoint events we can sum their probability, and their sum is $1/2$.