Probability Notes 18.310 (cont'd), Fall 2010

1 Variance and Covariance

Another quantity associated with a random variable is its *variance*. This is defined as

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

That is, the variance is the expectation of the square of the difference between the value of f and the expected value $\bar{f}$ of f. We can also expand this into:

$$\mathrm{Var}(f) = \sum_{x \in S} p(x)(f(x) - \bar{f})^2.$$

The variance tells us how closely the value of a random variable is clustered around its expected value. We can rewrite the definition of the variance as follows:

$$\begin{aligned} \mathrm{Var}(f) &= \mathrm{E}[(f - \bar{f})^2] \\ &= \mathrm{E}(f^2 - 2\bar{f}f + \bar{f}^2) \\ &= \mathrm{E}[f^2] - 2\bar{f}\mathrm{E}[f] + \bar{f}^2 \\ &= \mathrm{E}[f^2] - \bar{f}^2 \end{aligned}$$

so the variance of f is the expectation of the square of f minus the square of the expectation of f. We get from the second line to the third line by using linearity of expectation, and the third to the fourth by using the definition $\mathrm{E}f = \bar{f}$. Notice that the variance is always nonnegative, and that it is equal to 0 if f is constant. The *standard deviation* σ of f is defined to be the square root of $\mathrm{Var}(f)$.

Let us compute the variance and standard deviation of the roll of a die. Let the number on the die be the random variable X. We have that each of the numbers 1 through 6 are equally likely, so

$$EX = \sum_{i=1}^{6} p(i)i = \sum_{i=1}^{6} \frac{1}{6} i = \frac{21}{6}$$

and

$$E[X^2] = \sum_{i=1}^{6} p(i)i^2 = \sum_{i=1}^{6} \frac{1}{6} i^2 = \frac{91}{6}.$$

So

$$\mathrm{Var}(X) = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{35}{12}$$

and the standard deviation

$$\sigma = \sqrt{\frac{35}{12}} = 1.7078.$$

One can show that by the Cauchy-Schwartz inequality that

$$E[|f - \bar{f}|] \leq \sqrt{\mathrm{Var}(f)}.$$

Recall that the Cauchy Schwartz inequality says that for two vectors $\vec{s}$ and $\vec{t}$, the inner product $\sum_i s_i t_i$ is at most the product of their lengths. For s_i choose $\sqrt{p(x_i)}|f(x_i) - \bar{f}|$ and for t_i choose $\sqrt{p(x_i)}$. Then their inner product is the expected value of $|f - \bar{f}|$, the length of $\vec{s}$ is the standard deviation, and the length of $\vec{t}$ is 1. For the die example above, we have that the expected value of $|f - \bar{f}|$ is $3/2$, which is slightly less than the standard deviation of 1.7078.

If we a random variable f and a real c then $\mathrm{Var}(cf) = c^2\mathrm{Var}(f)$. Suppose we have two random variables f and g, and want to compute the variance of their sum. We get

$$\begin{aligned}
\mathrm{Var}(f+g) &= \mathrm{E}[(f+g)^2 - (\bar{f}+\bar{g})^2] \\
&= \mathrm{E}[f^2] + 2\mathrm{E}[fg] + \mathrm{E}[g^2] - (\bar{f}^2 + 2\bar{f}\bar{g} + \bar{g}^2) \\
&= \mathrm{E}[f^2] - \bar{f}^2 + 2\mathrm{E}[g^2] - \bar{g}^2 + 2\mathrm{E}[fg] - 2\bar{f}\bar{g} \\
&= \mathrm{Var}(f) + \mathrm{Var}(g) + 2\mathrm{E}[fg] - 2\bar{f}\bar{g}
\end{aligned}$$

This last quantity, $\mathrm{E}[fg] - \bar{f}\bar{g}$, is called the *covariance*. The covariance has the property that it is 0 if f and g are independent (but not vice versa ... as an exercise, you should be able to come up with two random variables f and g which have covariance 0, but which are not independent.) We can see this as follows:

$$\begin{aligned}
\mathrm{E}[f]\mathrm{E}[g] &= \sum_{\alpha \in \mathrm{range}(f)} P(f=\alpha)\alpha \sum_{\beta \in \mathrm{range}(g)} P(g=\beta)\beta \\
&= \sum_{\alpha,\beta} P(f=\alpha \text{ and } g=\beta)\alpha\beta \\
&= \mathrm{E}[fg].
\end{aligned}$$

Finally, suppose we have k random variables, $f_1 \ldots f_k$. What is the variance of the random variable $f = f_1 + f_2 + f_3 + \ldots + f_k$? We have, using the same reasoning as above,

$$
\begin{aligned}
\mathrm{Var}(f) &= \mathrm{E}(f_1 + \ldots + f_k)^2 - (\bar{f}_1 + \ldots + \bar{f}_k)^2 \\
&= \sum_{i=1}^{k} \mathrm{E}[f_i^2] + 2 \sum_{1 \le i < j \le k} \mathrm{E}[f_i f_j] - \sum_{i=1}^{k} \bar{f}_i^{\,2} - 2 \sum_{1 \le i < j \le k} \bar{f}_i \bar{f}_j \\
&= \sum_{i=1}^{k} \mathrm{Var}(f_i) + \sum_{1 \le i < j \le k} \mathrm{Cov}(f_i, f_j).
\end{aligned}
$$

This lets us calculate the variance of the number of heads if you flip a coin n times, say. Suppose a biased coin has probability p of coming up heads and probability $q = 1 - p$ of coming up tails. Then define f to be 1 if the coin comes up heads and 0 if the coin comes up tails. The variance is

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

If you flip the coin n times, the results of each coin flip are independent, so the covariance of any two coin flips is 0. This shows that the variance is n times the variance of a single coin flip, or $np(1-p)$.

1.1 Chebyshev's inequality

A very useful inequality, that can give bounds on probabilities, can be proved using the tools that we have developed so far. This is Chebyshev's inequality, and it is used to get an upper bound on the probability that a random variable takes on a value that is too far away from its mean.

Suppose you know the mean and variance of a random variable f? Is there some way that you can put a bound on the probability that the random variable is a long way away from its mean? This is exactly what Chebyshev's inequality does. Let's first derive Chebyshev's inequality intuitively, and then figure out how to turn this into a mathematically rigorous proof.

We will turn the probability around. Suppose we fix the probability p that the random variable is farther than $c\sigma$ away (for some given c) from the mean $\bar{f}$ (recall σ is the standard deviation $\sqrt{\mathrm{Var}(f)}$). Let's see how small the variance can be. Let's divide the sample space into two events. The first (which happens with probability say p) is that

$$f - \bar{f} < c\sigma.$$

In this case, the way to minimize their contribution to the variance is to set $f(x) = \bar{f}$, and the contribution of this case to the variance is 0. The second event is when

$$f - \bar{f} \ge c\sigma.$$

In this case, the way to minimize the variance is to set $|f(x) - \bar{f}| = c\sigma$, and the contribution of this case to the variance is $pc^2\sigma^2$. Since the variance is σ^2, we have

$$\begin{aligned} \sigma^2 &= pc^2\sigma^2, \\ p &= c^{-2}. \end{aligned}$$

Since this was the case that minimized the variance, in any other case, the variance has to be greater than this. This gives us Chebyshev's inequality, namely

$$P(\ |f - \bar{f}| \geq c\sigma\ ) \leq 1/c^2$$

or, letting $y = c\sigma$,

$$P(\ |f - \bar{f}| \geq y) \leq \sigma^2/y^2.$$

Now, let's turn this derivation into rigorous mathematical formulas. We first write down the formula for the variance

$$\sigma^2 = \sum_{x \in S} p(x)|f(x) - \bar{f}|^2.$$

Now, let's divide it into the two cases we talked about above.

$$\sigma^2 = \sum_{x:|f(x)-\bar{f}|<c\sigma} p(x)|f(x) - \bar{f}|^2 + \sum_{x:|f(x)-\bar{f}|\geq c\sigma} p(x)|f(x) - \bar{f}|^2.$$

For the first sum, we have $\sum p(x) = 1-p$ and for the second sum, we have $\sum p(x) = p$, where p is the probability that $|f - \bar{f}| \geq c\sigma$. Similarly, for the first sum, we have $|f(x) - \bar{f}|^2 \geq 0$ and for the second sum, we have $|f(x) - \bar{f}|^2 \geq c^2\sigma^2$. Putting these facts together, we have

$$\sigma^2 \geq pc^2\sigma^2,$$

which gives

$$p \leq 1/c^2,$$

and proves Chebyshev's inequality.

1.2 Weak law of large numbers

Using Chebyshev's inequality, we can now show the so-called *weak law of large numbers*. This law says that if we have a random variable f (say the value resulting from the throw of a die) and take many independent copies of it, the average value of all these copies will be very close to the expected value of f. More formally, the weak law of large numbers is the following.

theorem 1 (Weak law of large numbers) *Fix $\epsilon > 0$. Let $f_1, \cdots, f_n$ be n independent copies of a random variable f. Let*

$$g_n = \frac{1}{n}(f_1 + f_2 + \cdots f_n).$$

Then

$$P\left[|g_n - \bar{f}| \geq \epsilon\right] \to 0$$

as $n \to \infty$.

In plain English, the probability that g_n deviates from the expected value of f by at least ϵ becomes arbitrarily small as n grows arbitrarily large.

The weak law of large numbers can be proved by using Chebyshev's inequality applied to g_n. For this, we need to know E$[g_n]$ and Var(g_n). By linearity of expectations, we have

$$E[g_n] = E\left[\frac{1}{n}(f_1 + \cdots + f_n)\right] = \frac{1}{n}\sum_{i=1}^{n} E[f_i] = \frac{n}{n}E[f] = \bar{f}.$$

For the variance, we get

$$\begin{aligned} \text{Var}[g_n] &= \text{Var}\left[\frac{1}{n}(f_1 + \cdots + f_n)\right] \\ &= \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n} f_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[f_i] \\ &= \frac{1}{n}\text{Var}[f], \end{aligned}$$

the third equality being true since the f_i's are *independent*. Thus, as n tends to infinity, $E[g_n]$ remains constant while Var$[g_n]$ tends to 0. For example, we saw that the roll of a fair die gives a variance of $\frac{35}{12}$. If we were to roll the die 1000 times and average all 1000 values, we would get a random variable whose expected value is still 3.5 but whose variance is much smaller, it is only $\frac{35}{12{,}000}$.

Now that we know the expected value and variance of g_n, we can simply use Chebyshev's inequality on g_n to get:

$$P[|g_n - f| \geq \epsilon] \leq \frac{\text{Var}[g_n]}{\epsilon^2} = \frac{\text{Var}[f]}{n\epsilon^2},$$

and indeed this probability tends to 0 as n tends to infinity. This proves the weak law of large numbers.