COS 341, October 11, 2000
Handout Number 4

Some Probability Theory

1. Expected value and variance

Let X be a random variable on a probability space $\Omega = (U, p)$. We define $E(X) = \sum_{u \in U} p(u)X(u)$, and $Var(X) = \sum_{u \in U} p(u)(X(u) - t)^2$, where $t = E(X)$. Clearly $Var(X) \geq 0$. The *standard deviation* $\sigma(X)$ is defined to be $\sqrt{Var(X)}$. Intuitively, the value $E(X)$ tells you what the average value of X you can expect, if you perform many many experiments; the variance gives you some idea on how close to $E(X)$ you expect to see the typical value of X fall in these experiments. (This last statement is made more precise in the Chebyshev's Inequality to be discussed later.)

Note that one can write $Var(X) = E(Z)$, where $Z = (X - E(X))^2$. (Recall that $E(X)$ is a real number t, and for each $u \in U$, the random variable $X - E(X)$ has value $X(u) - t$, and the random variable Z has value $Z(u) = (X(u) - t)^2$.)

Example 1 Let $\Omega = (U, p)$, where $U = \{a, b, c, d, e, f, g, h\}$ and $p(a) = 0.2, p(b) = 0.13, p(c) = 0.17, p(d) = 0.1, p(e) = 0.1, p(f) = 0.2, p(g) = 0.04, p(h) = 0.06$. Let X be the random variable on Ω with $X(a) = 1, X(b) = X(c) = 3, X(d) = X(e) = X(f) = 4, X(g) = X(h) = 9$. Then by defintion

$$E(X) = p(a)X(a) + p(b)X(b) + \cdots + p(h)X(h).$$

Let $t = E(X)$ denote the above value. Then by definition

$$Var(X) = p(a)(X(a) - t)^2 + p(b)(X(b) - t)^2 + \cdots + p(h)(X(h) - t)^2.$$

From these formulas, we obtain $E(X) = 3.6$, $Var(X) = 4.44$, and hence $\sigma(X) = \sqrt{4.44} = 2.1....$

Example 2 (Fair n coin tosses)
Let $\Omega = (U, p)$, where $U = \{0, 1\}^n$ and $p(u) = 2^{-n}$. Let X be the random variable, where for each $u \in U$, $X(u)$ is defined as the number of 1's in u.

Example 3 (Magician's n coin tosses)
Let $\Omega = (U, p)$, where $U = \{0^n, 1^n\}$ and $p(0^n) = p(1^n) = 1/2$. Let X be the random variable corresponding to the number of 1's as in Example 2. Thus $X(0^n) = 0$ and $X(1^n) = n$.

In Example 3, by definition $E(X) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot n = n/2$. Also, $Var(X) = \frac{1}{2} \cdot (0 - n/2)^2 + \frac{1}{2} \cdot (n - n/2)^2 = n^2/4$. Hence $\sigma(X) = \sqrt{Var(X)} = n/2$.

The calculation of the various quantities in Example 2 is a little more involved. In class we showed $E(X) = n/2$, $Var(X) = n/4$, and hence $\sigma(X) = \sqrt{n}/2$. The derivation involved the evaluation of $\sum_k k\binom{n}{k}$ and $\sum_k k^2\binom{n}{k}$, which we already knew how to do. We won't duplicate it here.

It is of interest to compare Examples 2 and 3. The two random variables X have the same $E(X)$ but very different $\sigma(X)$. This is consistent with our intuition about $\sigma(X)$. In one case (Example 2) the typical value of X is within $O(\sqrt{n})$ of $E(X)$, while the other case, the value of X is always quite far away from $E(X)$.

2. Three important theorems

Recall that, if $X_i, 1 \leq i \leq m$ are random variables in Ω and c_i are real numbers, then the random variable $W = \sum_{1 \leq i \leq m} c_i X_i$ is defined by $W(u) = \sum_{1 \leq i \leq m} c_i X_i(u)$. The following important formula was derived in class.

Theorem 1 (Linearity of Expected Value)

$$E(W) = \sum_{1 \leq i \leq m} c_i E(X_i).$$

Theorem 1 is quite a remarkable statement. Quite often it allows us to compute $E(X)$ for a complicated X from the calculation of the expected values of less complicated random variables X_i. The remarkable fact is that it assumes nothing about the relationships among the X_i's; they can be independent of each other, or correlated in any way.

As an illustration, consider either Example 2 or Example 3. Let X_i be the random variable such that $X_i(u)$ is the i-th bit of u. Then clearly $X = \sum_{1 \leq i \leq n} X_i$. Notice that $E(X_i) = 1/2$, and Theorem 1 gives us immediately $E(X) = n/2$. This argument is true for both Examples 2 and 3, even though X_i's are highly correlated in one case, and quite independent in the other case.

Now for the next theorem. It essentially says that, perform an experiment according to Ω, then the probability of seeing the value of X deviating from $E(X)$ by c (or more) standard deviations is less than $1/c^2$.

Theorem 2 (Chebyshev's Inequality) Let $c > 0$,

$$\Pr\{|X - E(X)| > c\sigma(X)\} \leq c^{-2}.$$

Proof If $Var(X) = 0$, then $X(u) = E(X)$ for all $u \in U$ (why?), and the Inequality is clearly true.

Assume $Var(X) > 0$. Let $Y = X - E(X)$. Then

$$\begin{aligned} Var(X) &= E(Y^2) \\ &= \sum_{u \in U} p(u)(Y(u))^2 \\ &\geq \sum_{u \in U, |Y(u)| > c\sigma(X)} p(u)(Y(u))^2 \\ &\geq (c\sigma(X))^2 \sum_{u \in U, |Y(u)| > c\sigma(X)} p(u) \\ &= c^2 Var(X) \Pr\{|Y(u)| > c\sigma(X)\}. \end{aligned}$$

This implies the Inequality. □

For example, take $c = 10$. The probability for X to deviate from $E(X)$ by $10\sigma(X)$ (or more) is less than 1 percent.

Finally, we consider a formula which comes handy for calculating the variance and for analytic studies.

Theorem 3 (A useful formula)

$$Var(X) = E(X^2) - (E(X))^2. \tag{1}$$

Proof Let $t = E(X)$ and $Z = (X - t)^2$. Note that $Z(u) = (X - t)^2(u) = (X(u))^2 - 2tX(u) + t^2$ for all $u \in U$. Thus $(X - t)^2 = X^2 - 2tX + t^2$ expresses the random variable Z as a linear combination of the random variables X^2, X. By the linearity of the expected value, we have $E(Z) = E(X^2) - 2tE(X) + t^2 = E(X^2) - t^2$, which proves (1). □

3. Generating Functions

We have seen earlier that generating functions are useful for evaluating sums such as $\sum_{k=0,2,4,\cdots} \binom{n}{k}$. We now demonstrate that the generating functions are also useful for evaluating $E(X)$ and $Var(X)$, when X take on only nonnegative integer values. Take any such X, let $a_k = \Pr\{X = k\}$. Define

$$F(x) = \sum_{k \geq 0} a_k x^k.$$

Then

$$F'(x) = \sum_{k \geq 0} k a_k x^{k-1},$$

and

$$F''(x) = \sum_{k \geq 0} k(k-1)a_k x^{k-2}.$$

Thus,

$$F'(1) = \sum_{k \geq 0} k a_k,$$

and

$$F''(1) = \sum_{k \geq 0} k(k-1)a_k.$$

It follows that

$$E(X) = \sum_{k \geq 0} k a_k = F'(1). \tag{2}$$

And

$$\begin{aligned} E(X^2) &= \sum_{k \geq 0} k^2 a_k \\ &= \sum_{k \geq 0} (k(k-1)+k)a_k \\ &= \sum_{k \geq 0} k(k-1)a_k + \sum_{k \geq 0} k a_k \\ &= F''(1) + F'(1). \end{aligned}$$

By (2) this gives

$$Var(X) = F''(1) + F'(1) - (F'(1))^2. \tag{3}$$

As an application, consider Example 2. Then $a_k = \binom{n}{k}/2^n$ and $F(x) = \sum_k a_k x^k = (1+x)^n/2^n$. Clearly, $F'(x) = n(1+x)^{n-1}/2^n$ and $F''(x) = n(n-1)(1+x)^{n-2}/2^n$. Thus $F'(1) = n/2$ and $F''(1) = n(n-1)/4$. From (3) we have $Var(X) = n(n-1)/4 + (n/2) - (n^2/4) = n/4$.

4. About polling

Consider an upcoming election in a large metropolitan area, with a large voting population. Two candidates A and B are running. If you are a pollster, taking a random sample of size n (n is a moderate number compared with the population). How confident are you about the result you get? If your result has m votes for A, how do you decide on ϵ so that you can say that "the result is m/n for A, $(n-m)/n$ for B, with a margin of error of ϵ"?

You can fairly accurately model your polling as follows. Let $0 < r < 1$ be the true fraction of voters for A. Throw a biased (with bias r for $HEAD$ and $1-r$ for $TAIL$) coin n times. The HEADs are for A and TAILs are for B. Let X denote the random variable corresponding to the total number of HEADs.

You polling corresponds to making an observation of the value of X, say with result m. By Chebyshev's Inequality, m is likely to be within a few standard deviations of the expected value of X. That means in turn that rn (which is $E(X)$) is within a few standard deviation of m. That is, rn is within $\sigma(x)$ of m. Hence r is within $\sigma(X)/n$ of your polling result m/n.

It is easy to calculate $\sigma(X)$. Clearly, $a_k = \Pr\{X = k\} = \binom{n}{k} r^k (1-r)^{n-k}$. The generating function $F(x) = \sum_{0 \le k \le n} a_k x^k$, which is equal to $((1-r)+rx)^n$ by the Binomial Theorem. It follows that $F'(x) = nr((1-r)+rx)^{n-1}$ and $F''(x) = n(n-1)r^2((1-r)+rx)^{n-2}$. From this, we obtain $Var(X) = F''(1) + F'(1) - F'(1)^2 = nr(1-r)$. This leads to $\sigma(X) = (Var(X))^{1/2} = (nr(1-r))^{1/2}$. Note that $(r(1-r))^{1/2} \le 1/2$ (Can you prove it?), implying $\sigma(X) \le \sqrt{n}/2$. Thus, from the discussions in the last paragraph, it is reasonable to say that your polling result has a margin of error about $1/(2\sqrt{n})$.