Topic 7: Expected Values

October 1, 2009

1 Discrete Random Variables

Recall for a data set $x_1, x_2, \ldots, x_n$, we can compute the sample average of a function of the data

$$h(x) = \frac{1}{n} \sum_{x} h(x)p(x).$$

where $p(x)$ is the proportion of observations taking the value $x$

Analogously, for a finite sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$, we can define the expectation or the expected value of a random variable $X$ by

$$EX = \sum_{j=1}^{N} X(\omega_j)P(\omega_j).$$

(1)

In this case, two properties of expectation are immediate:

1. If $X(\omega) \geq 0$ for every $\omega \in \Omega$, then $EX \geq 0$.

2. Let $X_1$ and $X_2$ be two random variables and $c_1, c_2$ be two real numbers, then

$$E[c_1X_1 + c_2X_2] = c_1EX_1 + c_2EX_2.$$

Taking these two properties, we say that expectation is a positive linear functional. Another example of a positive linear functional is the integral

$$f \mapsto \int_{a}^{b} f(x) \, dx$$

that takes a positive function and gives the area between the graph of $f$ and the $x$-axis between the vertical lines $x = a$ and $x = b$.

Example 1. Roll one die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let $X$ be the value on the die. So, $X(\omega) = \omega$. If the die is fair, $P(\omega) = 1/6$ and

$$EX = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = \frac{7}{2}.$$

If $X_1$ and $X_2$ are the values on two rolls of a die, then the expected value of the sum

$$E[X_1 + X_2] = EX_1 + EX_2 = \frac{7}{2} + \frac{7}{2} = 7.$$

We can generalize the identity in (1) to

$$Eg(X) = \sum_{j=1}^{N} g(X(\omega_j))P(\omega_j).$$
As before, we can simplify

\[
Eg(X) = \sum_x \sum_{\omega : X(\omega) = x} g(X(\omega))P(\omega) = \sum_x \sum_{\omega : X(\omega) = x} g(x)P(\omega)
\]

\[
= \sum_x g(x) \sum_{\omega : X(\omega) = x} P(\omega) = \sum_x g(x)P(X = x) = \sum_x g(x)f_X(x)
\]

where \(f_X\) is the probability mass function for \(X\).

A similar formula holds if we have a vector of random variables \(X = (X_1, X_2, \ldots, X_n)\), \(f_X\), the joint probability mass function and \(g\) a real-valued function of \(x = (x_1, x_2, \ldots, x_n)\).

**Example 2.** Flip a biased coin twice and let \(X\) be the number of heads. Then,

<table>
<thead>
<tr>
<th>(x)</th>
<th>(f_X(x))</th>
<th>(xf_X(x))</th>
<th>(x^2f_X(x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>((1-p)^2)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>(2p(1-p))</td>
<td>2p</td>
<td>2p</td>
</tr>
<tr>
<td>2</td>
<td>(p^2)</td>
<td>2p</td>
<td>4p</td>
</tr>
</tbody>
</table>

Thus, \(EX = 2p\) and \(EX^2 = 2p + 2p^2\).

### 2 Counting

Suppose that two experiments are to be performed.

- Experiment 1 can have \(n_1\) possible outcomes and
- for each outcome of experiment 1, experiment 2 has \(n_2\) possible outcomes.

Then together there are \(n_1 \times n_2\) possible outcomes.

**Exercise 3.** Generalize this basic principle of counting to \(k\) experiments.

#### 2.1 Permutations

Assume that we have a collection of \(n\) objects and we wish to make an ordered arrangement of \(k\) of these objects. Using the generalized principle of counting, the number of possible outcomes is

\[
n \times (n-1) \times \cdots \times (n-k+1).
\]

We will write this as \(n)_k\) and say \(n\ falling\ \(k\).

**Example 4** (birthday problem). In a list the birthday of \(k\) people, there are \(365^k\) possible lists (ignoring leap year births) and

\[
(365)_k
\]

possible lists with no date written twice. Thus, the probability, under equally likely outcomes, that no two people on the list have the same birthday is

\[
\frac{(365)_k}{365^k}
\]

and, under equally likely outcomes,

\[
P(\text{at least one pair of individuals share a birthday}) = 1 - \frac{(365)_k}{365^k}
\]

For example
### Introduction to Statistical Methodology

#### Expected Values

<table>
<thead>
<tr>
<th>$k$</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>18</th>
<th>20</th>
<th>22</th>
<th>23</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>0.027</td>
<td>0.117</td>
<td>0.253</td>
<td>0.347</td>
<td>0.411</td>
<td>0.476</td>
<td>0.507</td>
<td>0.569</td>
<td>0.706</td>
<td>0.891</td>
<td>0.970</td>
<td>0.994</td>
</tr>
</tbody>
</table>

**The R code and output**

```r
> prob=rep(1,30)
> for (n in 2:30) {prob[n]=prob[n-1]*(365-n+1)/365}
> data.frame(1-prob)
```

<table>
<thead>
<tr>
<th>X1...prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000000000</td>
</tr>
<tr>
<td>0.002739726</td>
</tr>
<tr>
<td>0.008204166</td>
</tr>
<tr>
<td>0.016355912</td>
</tr>
<tr>
<td>0.027135574</td>
</tr>
<tr>
<td>0.040462484</td>
</tr>
<tr>
<td>0.056235703</td>
</tr>
<tr>
<td>0.074335292</td>
</tr>
<tr>
<td>0.094623834</td>
</tr>
<tr>
<td>0.116948178</td>
</tr>
<tr>
<td>0.141141378</td>
</tr>
<tr>
<td>0.167024789</td>
</tr>
<tr>
<td>0.194410275</td>
</tr>
<tr>
<td>0.223102512</td>
</tr>
<tr>
<td>0.252901320</td>
</tr>
<tr>
<td>0.283604005</td>
</tr>
<tr>
<td>0.315007665</td>
</tr>
<tr>
<td>0.346911418</td>
</tr>
<tr>
<td>0.379118526</td>
</tr>
<tr>
<td>0.411438384</td>
</tr>
<tr>
<td>0.443688335</td>
</tr>
<tr>
<td>0.475695308</td>
</tr>
<tr>
<td>0.507297234</td>
</tr>
<tr>
<td>0.538344258</td>
</tr>
<tr>
<td>0.568699704</td>
</tr>
<tr>
<td>0.598240820</td>
</tr>
<tr>
<td>0.626859282</td>
</tr>
<tr>
<td>0.654461472</td>
</tr>
<tr>
<td>0.680968537</td>
</tr>
<tr>
<td>0.706316243</td>
</tr>
</tbody>
</table>

The ordered arrangement of all $n$ objects is

\[(n)_n = n \times (n - 1) \times \cdots \times 1 = n!,\]

**n factorial.** We take $0! = 1$.

**Exercise 5.**

\[(n)_k = \frac{n!}{(n-k)!}.

### 2.2 Combinations

Write

\[
\binom{n}{k}
\]
The number of different groups of \( k \) objects that can be chosen from a collection of \( n \).

**Theorem 6.**

\[
\binom{n}{k} = \frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!}.
\]

Here is an example of a combinatorial proof.

We will form an ordered arrangement of \( k \) objects from a collection of \( n \) by:

1. First choosing a group of \( k \) objects.
   The number of possible outcomes for this experiment is \( \binom{n}{k} \).
2. Then, arranging this \( k \) objects in order.
   The number of possible outcomes for this experiment is \( k! \).

So, by the basic principle of counting,

\[
(n)_k = \binom{n}{k} \times k!.
\]

Now complete the proof by dividing both sides by \( k! \).

**Exercise 7** (binomial theorem).

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

**Exercise 8.** \( \binom{n}{1} = \binom{n-1}{n-1} = n. \) \( \binom{n}{k} = \binom{n}{n-k}. \) Thus, we set \( \binom{n}{n} = \binom{n}{0} = 1 \)

The number of combinations is computed in \( \mathbb{R} \) using \texttt{choose}. For example, \( \binom{8}{5} \)

> choose(8, 5)

[1] 56

**Theorem 9** (Pascal’s triangle).

\[
\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.
\]

To establish this identity, distinguish one of the \( n \) objects in the collection.

1. If the distinguished object is the group, then we must choose \( k - 1 \) from the remaining \( n - 1 \) objects. Thus, \( \binom{n-1}{k-1} \) groups have the distinguished object.
2. If the distinguished object is not the group, then we must choose \( k \) from the remaining \( n - 1 \) objects. Thus, \( \binom{n-1}{k} \) groups do not have the distinguished object.
3. These choices of groups of no overlap.

**Example 10** (Bernoulli trials). Random variables \( X_1, X_2, \ldots, X_n \) are called a sequence of Bernoulli trials provided that:

1. Each \( X_i \) takes on two values 0 and 1. We call the value 1 a success and the value 0 a failure.
2. \( P\{X_i = 1\} = p \) for each \( i \).
3. The outcomes on each of the trials is independent.
For each $i$, 
\[ EX_i = 0 \cdot P\{X_i = 0\} + 1 \cdot P\{X_i = 1\} = 0 \cdot (1 - p) + 1 \cdot p = p. \]

Let $S = X_1 + X_2 + \cdots + X_n$ be the total number of successes. A sequence having $x$ successes has probability 
\[ p^x (1 - p)^{n-x}. \]

In addition, we have \( \binom{n}{x} \) mutually exclusive sequences that have $x$ successes. Thus, we have the mass function 
\[ f_S(x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \ldots \]

The fact that $\sum_x f_S(x) = 1$ follows from the binomial theorem. Consequently, $S$ is called a binomial random variable.

Using the linearity of expectation 
\[ ES = E[X_1 + X_2 + \cdots + X_n] = p + p + \cdots + p = np. \]

## 3 Continuous Random Variables

For $X$ a continuous random variable with density $f_X$, consider the discrete random variable $\hat{X}$ obtained from $X$ by rounding down to the nearest multiple of $\Delta x$. Denoting the mass function of $\hat{X}$ by $f_{\hat{X}}(\hat{x}) = P\{\hat{x} \leq X < \hat{x} + \Delta x\}$, we have 
\[ Eg(\hat{X}) = \sum_{\hat{x}} g(\hat{x}) f_{\hat{X}}(\hat{x}) = \sum_{\hat{x}} g(\hat{x}) P\{\hat{x} \leq X < \hat{x} + \Delta x\} \]
\[ \approx \sum_{\hat{x}} g(\hat{x}) f_x(\hat{x}) \Delta x \approx \int_{-\infty}^{\infty} g(x) f_X(x) \, dx. \]

Taking limits as $\Delta x \to 0$ yields the identity 
\[ Eg(X) = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx. \]

As in the case of discrete random variables, a similar formula holds if we have a vector of random variables $X = (X_1, X_2, \ldots, X_n)$, $f_X$, the joint probability density function and $g$ a real-valued function of $x = (x_1, x_2, \ldots, x_n)$. The expectation in this case is an $n$-fold Riemann integral.

Integration by parts give an alternative to computing expectation. Let $X$ be a positive random variable and $g$ an increasing function.
\[ u(x) = g(x) \quad v(x) = -(1 - F_X(x)) \]
\[ u'(x) = g'(x) \quad v(x) = f_X(x) = F_X'(x). \]

Then, 
\[ \int_0^b g(x) f_X(x) \, dx = -g(b)(1 - F_X(b)) \bigg|_0^b + \int_0^b g'(x)(1 - F_X(x)) \, dx \]

Now, substitute $F_X(0) = 0$, then the first term,
\[ g(x)(1 - F_X(x)) \bigg|_0^b = g(b)(1 - F_X(b)) = \int_b^\infty g(b) f_X(x) \, dx \leq \int_b^\infty g(x) f_X(x) \, dx \]
Because, \( \int_{0}^{\infty} g(x)f_X(x) \, dx < \infty, \int_{b}^{\infty} g(x)f_X(x) \, dx \rightarrow 0 \) as \( b \rightarrow \infty \). Thus,

\[
Eg(X) = \int_{0}^{\infty} g'(x)P\{X > x\} \, dx.
\]

For the case \( g(x) = x \), we obtain

\[
EX = \int_{0}^{\infty} P\{X > x\} \, dx.
\]

In words, the expected value is the area between the cumulative distribution function and the line \( y = 1 \) or the area under the survival function. For the case of the dart board, we see that the area under the distribution function between \( y = 0 \) and \( y = 1 \) is \( \int_{0}^{1} x^2 \, dx = 1/3 \), so the area below the survival function \( EX = 2/3 \).

Example 11. Let \( T \) be an exponential random variable, then for some \( \lambda \), \( P\{T > t\} = \exp(-\lambda t) \). Then

\[
ET = \int_{0}^{\infty} P\{T > t\} \, dt = \int_{0}^{\infty} \exp(-\lambda t) \, dt = -\frac{1}{\lambda} \exp(-\lambda t) \bigg|_{0}^{\infty} = 0 - \left(-\frac{1}{\lambda}\right) = \frac{1}{\lambda}.
\]

Example 12. For a standard normal random variable, the probability density function

\[
\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \quad z \in \mathbb{R}.
\]

The expectation

\[
EZ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \exp\left(-\frac{z^2}{2}\right) \, dz = 0
\]

because the integrand is an odd function.
\[ EZ^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \exp\left(-\frac{z^2}{2}\right) dz \]

To evaluate this integral, integrate by parts

\[ u(z) = z \quad v(z) = -\exp\left(-\frac{z^2}{2}\right) \]
\[ u'(z) = 1 \quad v'(z) = z \exp\left(-\frac{z^2}{2}\right) \]

Thus,

\[ EZ^2 = \frac{1}{\sqrt{2\pi}} \left( -z \exp\left(-\frac{z^2}{2}\right) \right|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz \].

Use l'Hôpital’s rule to see that the first term is 0 and the fact that the integral of a probability density function is 1 to see that the second term is 1.

Several choices for \( g \) have special names.

1. If \( g(x) = x \), then \( \mu = EX \) is called variously the mean, and the first moment.
2. If \( g(x) = x^k \), then \( EX^k \) is called the \( k \)-th moment.
3. If \( g(x) = (x-x_1) \cdots (x-x_k) \), then \( E(X-x_1)^k \) is called the \( k \)-th factorial moment.
4. If \( g(x) = (x-\mu)^k \), then \( E(X-\mu)^k \) is called the \( k \)-th central moment.
5. The second central moment \( \sigma^2 = E(X-\mu)^2 \) is called the variance.
6. The third moment of the standardized random variable is called the skewness.
7. The fourth moment of the standardized is called the kurtosis.
8. If \( X \) is \( R^d \)-valued and \( g(x) = e^{i\langle \theta, x \rangle} \), where \( \langle \cdot, \cdot \rangle \) is the standard inner product, then \( \phi(\theta) = E e^{i\langle \theta, X \rangle} \) is called the Fourier transform or the characteristic function. The characteristic function receives its name from the fact that the mapping from the distribution to this function is one-to-one.
9. Similarly, if \( X \) is \( R^d \)-valued and \( g(x) = e^{i\langle \theta, x \rangle} \), then \( m(\theta) = E e^{i\langle \theta, X \rangle} \) is called the Laplace transform or the moment generating function. The moment generating function also gives a one-to-one mapping. However, not every distribution has a moment generating function. To justify the name, consider the one-dimensional case \( m(\theta) = E e^{\theta X} \). Then,

\[ m'(\theta) = EX e^{\theta X}, \quad m'(0) = EX \]
\[ m''(\theta) = EX^2 e^{\theta X}, \quad m''(0) = EX \]
\[ \vdots \]
\[ m^{(k)}(\theta) = EX^k e^{\theta X}, \quad m^{(k)}(0) = EX^k. \]
10. If \( X \) is \( Z^+ \)-valued and \( g(x) = z^x \), then \( \rho(z) = E z^X = \sum_{x=0}^{\infty} P(X = x) z^x \) is called the (probability) generating function. For \( N \)-valued random variable, the probability generating function is used. It allows us to use ideas from complex variable and power series to perform computations.

**Exercise 13.** \( \text{Var}(aX + b) = a^2 \text{Var}(X) \).
4 Independence

If $X_1$ and $X_2$ are independent discrete random variables and $g_1$ and $g_2$ are real valued functions, then

$$E[g_1(X_1)g_2(X_2)] = \sum_{x_1} \sum_{x_2} g_1(x_1)g_2(x_2)f_{X_1,X_2}(x_1,x_2) = \sum_{x_1} \sum_{x_2} g_1(x_1)g_2(x_2)f_{X_1}(x_1)f_{X_2}(x_2)$$

$$= \left( \sum_{x_1} g_1(x_1)f_{X_1}(x_1) \right) \left( \sum_{x_2} g_2(x_2)f_{X_2}(x_2) \right) = E[g_1(X_1)] \cdot E[g_2(X_2)]$$

A similar identity that the expectation of the product of two independent random variables equals to the product of the expectation holds for continuous random variables.

For example, if $X_1$ and $X_2$ are random variables with respective means $\mu_1$ and $\mu_2$, then

$$\text{Var}(X_1 + X_2) = E[((X_1 + X_2) - (\mu_1 + \mu_2))^2] = E[((X_1 - \mu_1) + (X_2 - \mu_2))^2]$$

$$= E[(X_1 - \mu_1)^2] + 2E[(X_1 - \mu_1)(X_2 - \mu_2)] + E[(X_2 - \mu_2)^2]$$

$$= \text{Var}(X_1) + 2\text{Cov}(X,Y) + \text{Var}(X_2).$$

where the covariance $\text{Cov}(X,Y) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$.

If $X_1$ and $X_2$ are independent, then $\text{Cov}(X,Y) = E[(X_1 - \mu_1)] \cdot E[(X_2 - \mu_2)]$ and the variance of the sum is the sum of the variances.