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

$$\overline{h(x)} = \sum_x h(x)p(x).$$

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

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

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 postive 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,*

x	$f_X(x)$	$xf_X(x)$	$x^2f_X(x)$
0	$(1-p)^2$	0	0
1	$2p(1-p)$	$2p(1-p)$	$2p(1-p)$
2	p^2	$2p^2$	$4p^2$
		$2p$	$2p+2p^2$

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 outdomces 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,

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

For example

k	5	10	15	18	20	22	23	25	30	40	50	100
probability	0.027	0.117	0.253	0.347	0.411	0.476	0.507	0.569	0.706	0.891	0.970	0.994

The R *code and output*

```
> prob=rep(1,30)
>  for (n in 2:30){prob[n]=prob[n-1]*(365-n+1)/365}
> data.frame(1-prob)
     X1...prob
1  0.000000000
2  0.002739726
3  0.008204166
4  0.016355912
5  0.027135574
6  0.040462484
7  0.056235703
8  0.074335292
9  0.094623834
10 0.116948178
11 0.141141378
12 0.167024789
13 0.194410275
14 0.223102512
15 0.252901320
16 0.283604005
17 0.315007665
18 0.346911418
19 0.379118526
20 0.411438384
21 0.443688335
22 0.475695308
23 0.507297234
24 0.538344258
25 0.568699704
26 0.598240820
27 0.626859282
28 0.654461472
29 0.680968537
30 0.706316243
```

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

$$\binom{n}{k}$$

for the number of 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. $\binom{n}{1} = \binom{n}{n-1} = n$. $\binom{n}{k} = \binom{n}{n-k}$. *Thus, we set* $\binom{n}{n} = \binom{n}{0} = 1$

The number of combinations is computed in R using `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 $\tilde{X}$ obtained from X by rounding down to the nearest multiple of Δx. Denoting the mass function of $\tilde{X}$ by $f_{\tilde{X}}(\tilde{x}) = P\{\tilde{x} \leq X < \tilde{x} + \Delta x\}$, we have

$$\begin{aligned} Eg(\tilde{X}) &= \sum_{\tilde{x}} g(\tilde{x}) f_{\tilde{X}}(\tilde{x}) = \sum_{\tilde{x}} g(\tilde{x}) P\{\tilde{x} \leq X < \tilde{x} + \Delta x\} \\ &\approx \sum_{\tilde{x}} g(\tilde{x}) f_x(\tilde{x}) \Delta x \approx \int_{-\infty}^{\infty} g(x) f_X(x)\,dx. \end{aligned}$$

Taking limits as $\Delta x \to 0$ yields the identity

$$Eg(X) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx. \tag{2}$$

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.

$$\begin{array}{ll} u(x) = g(x) & v(x) = -(1 - F_X(x)) \\ u'(x) = g'(x) & v(x) = f_X(x) = F_X'(x). \end{array}$$

Then,

$$\int_0^b g(x) f_X(x)\,dx = -g(x)(1 - F_X(x))\Big|_0^b + \int_0^b g'(x)(1 - F_X(x))\,dx$$

Now, substitute $F_X(0) = 0$, then the first term,

$$g(x)(1 - F_X(x))\Big|_0^b = g(b)(1 - F_X(b)) = \int_b^{\infty} g(b) f_X(x)\,dx \leq \int_b^{\infty} g(x) f_X(x)\,dx$$

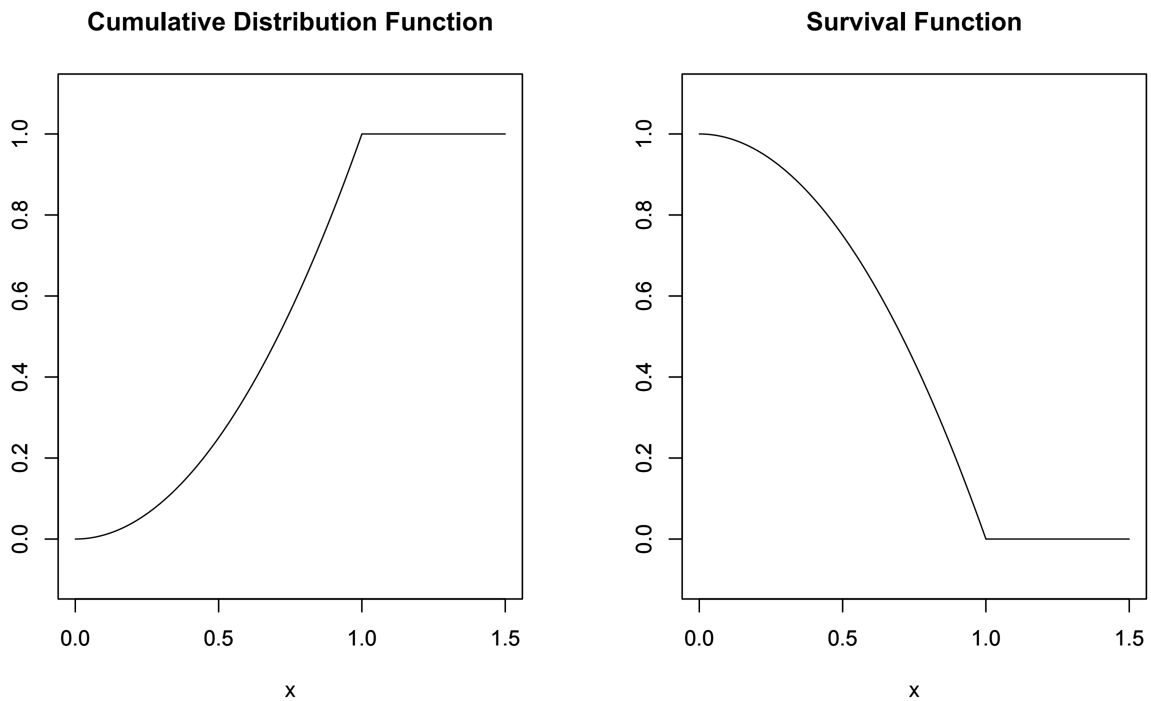
Because, $\int_0^\infty g(x)f_X(x)\,dx < \infty$, $\int_b^\infty g(x)f_X(x)\,dx \to 0$ as $b \to \infty$. Thus,

$$Eg(X) = \int_0^\infty g'(x)P\{X > x\}\,dx.$$

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 λ, $P\{T > t\} = \exp -(\lambda t)$. Then*

$$ET = \int_0^\infty P\{T > t\}\,dt = \int_0^\infty \exp -(\lambda t)\,dt = -\frac{1}{\lambda}\exp -(\lambda t)\Big|_0^\infty = 0 - (-\frac{1}{\lambda}) = \frac{1}{\lambda}.$$

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

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

The expectation

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

because the integrand is an odd function.

$$EZ^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \exp(-\frac{z^2}{2})\, dz$$

To evaluate this integral, integrate by parts

$$\begin{array}{ll} u(z) = z & v(z) = -\exp(-\frac{z^2}{2}) \\ u'(z) = 1 & v'(z) = z\exp(-\frac{z^2}{2}) \end{array}$$

Thus,

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

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 choice for g have special names.

1. If $g(x) = x$, then $\mu = EX$ is call variously the **mean**, and the **first moment.**
2. If $g(x) = x^k$, then EX^k is called the k-**th moment**.
3. If $g(x) = (x)_k$, where $(x)_k = x(x-1)\cdots(x-k+1)$, then $E(X)_k$ 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 $\sigma_X^2 = E(X-\mu)^2$ is called the **variance.** Note that

$$\text{Var}(X) = E(X-\mu)^2 = EX^2 - 2\mu EX + \mu^2 = EX^2 - 2\mu^2 + \mu^2 = EX^2 - \mu^2.$$

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 $\mathbb{R}^d$-valued and $g(x) = e^{i\langle\theta,x\rangle}$, where $\langle\cdot,\cdot\rangle$ is the standard inner product, then $\phi(\theta) = Ee^{i\langle\theta,X\rangle}$ 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 $\mathbb{R}^d$-valued and $g(x) = e^{\langle\theta,x\rangle}$, then $m(\theta) = Ee^{\langle\theta,X\rangle}$ 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 $m(\theta) = Ee^{\theta X}$. Then,

$$\begin{aligned} m'(\theta) &= EXe^{\theta X}, & m'(0) &= EX \\ m''(\theta) &= EX^2e^{\theta X}, & m''(0) &= EX \\ &\vdots & &\vdots \\ m^{(k)}(\theta) &= EX^ke^{\theta X}, & m^{(k)}(0) &= EX^k. \end{aligned}$$

10. If X is $\mathbb{Z}^+$-valued and $g(x) = z^x$, then $\rho(z) = Ez^X = \sum_{x=0}^{\infty} P\{X = x\}z^x$ is called the **(probability) generating function**. For $\mathbb{N}$-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

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

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 μ_1 and μ_2, then

$$\begin{aligned}\mathrm{Var}(X_1+X_2) &= E[((X_1+X_2)-(\mu_1+\mu_2))^2] = E[((X_1-\mu_1)+(X_2-\mu_2))^2]\\ &= E[(X_1-\mu_1)^2] + 2E[(X_1-\mu_1)(X_2-\mu_2)] + E[(X_2-\mu_2)^2]\\ &= \mathrm{Var}(X_1) + 2\mathrm{Cov}(X,Y) + \mathrm{Var}(X_2).\end{aligned}$$

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

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