STAT 830
Expectation and Moments

I begin by reviewing the usual undergraduate definitions of expected value. For absolutely continuous random variables X we usually say:

Definition: If X has density f then

$$\mathrm{E}\{g(X)\} = \int g(x)f(x)\,dx\,.$$

For discrete random variables we say:

Definition: If X has discrete density f then

$$\mathrm{E}\{g(X)\} = \sum_x g(x)f(x)\,.$$

There is something of a problem with these two definitions. They seem to define, for instance, $\mathrm{E}(X^2)$, in two different ways. If X has density f_X then we would have

$$\mathrm{E}(X^2) = \int x^2 f_X(x)\,dx.$$

But we could also define $Y = X^2$ and try to figure out a density f_Y for Y. Then we would have

$$\mathrm{E}(Y) = \int y f_Y(y) dy.$$

Are these two formulas the same? The answer is yes.

Fact: If $Y = g(X)$ for some one-to-one smooth function g (by which I mean say g is continuously differentiable) then

$$\begin{aligned}\mathrm{E}(Y) &= \int y f_Y(y)\,dy = \int g(x) f_Y(g(x))g'(x)\,dx \\ &= \mathrm{E}\{g(X)\}\end{aligned}$$

by change of variables formula for integration so we must have

$$f_X(x) = f_Y(g(x))g'(x).$$

For the moment I won't prove this but let me consider the case where, for instance $Y = e^{2X}$. Then $g(x) = e^{2x}$ and $g'(x) = 2e^{2x}$. Moreover

$$\begin{aligned} f_X(x) &= \frac{d}{dx}F_X(x) \\ &= \frac{d}{dx}P(X \le x) \\ &= \frac{d}{dx}P(e^{2X} \le e^{2x}) \\ &= \frac{d}{dx}P(Y \le e^{2x}) \\ &= \frac{d}{dx}F_Y(e^{2x}) \\ &= f_Y(e^{2x})\frac{d}{dx}e^{2x} \end{aligned}$$

as advertised.

General Definition of E

There are random variables which are neither absolutely continuous nor discrete. I now give a definition of expected value which covers such cases and includes both discrete and continuous random variables.

Definition: We say that a random variable X is simple if we can write

$$X(\omega) = \sum_1^n a_i 1(\omega \in A_i)$$

for some constants $a_1, \ldots, a_n$ and events A_i.

Definition: For a simple random variable X we define

$$\mathrm{E}(X) = \sum a_i P(A_i)\,.$$

I remark that logically it might be possible to write X in two ways, say

$$\sum_{i=1}^n a_i 1(\omega \in A_i) = \sum_{i=1}^m b_i 1(\omega \in B_i)$$

some constants $a_1, \ldots, a_n$, $b_1, \ldots, b_m$ and events $A_1, \ldots, A_n$ and $B_1, \ldots, B_m$. I claim that if this happens then we must have

$$\sum_{i=1}^{n} a_i P(A_i) = \sum_{i=1}^{m} b_i P(B_i).$$

I won't prove the claim!

For positive random variables which are not simple we extend our definition by approximation from below:

Definition: If $X \geq 0$ then

$$\mathrm{E}(X) = \sup\{\mathrm{E}(Y) : 0 \leq Y \leq X, Y \text{ simple}\}.$$

This notation hides the fact that for positive, simple, random variables X we appear to have given 2 definitions for $\mathrm{E}(X)$. It is possible to prove they are the same.

Finally we extend the definition to general random variables:

Definition: A random variable X is **integrable** if

$$\mathrm{E}(|X|) < \infty.$$

In this case we define

$$\mathrm{E}(X) = \mathrm{E}\{\max(X, 0)\} - \mathrm{E}\{\max(-X, 0)\}.$$

Again it might seem we have another definition for simple random variable or for non-negative random variables but it is possible to prove all the definitions agree.

Fact: : E is a linear, monotone, positive operator. This means:

1. **Linear**: $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$ provided X and Y are integrable.

2. **Positive**: $P(X \geq 0) = 1$ implies $\mathrm{E}(X) \geq 0$.

3. **Monotone**: $P(X \geq Y) = 1$ and X, Y integrable implies $\mathrm{E}(X) \geq \mathrm{E}(Y)$.

Jargon: An *operator* is a function whose domain is itself a set of functions. That makes E an operator because random variables are functions. Sometimes we call operators whose range is in real or complex numbers a *functional*.

Convergence Theorems

There are some important theorems about interchanging limits with integrals and our definition of E is really the definition of an integral. In fact you will often see a variety of notations:

$$\begin{aligned}
\mathrm{E}(g(X)) &= \int g(x)F(dx) \\
&= \int g(x)dF(x) \\
&= \int g dF
\end{aligned}$$

Sometimes the integral notations make it easier to see how a calculation works out. The notation $dF(x)$ has the advantage that if F has a density $f = F'$ we can write

$$dF(x) = f(x)dx.$$

In calculus courses there is quite a bit of attention paid to such questions as when

$$\frac{d}{dy}\int g(x,y)dx = \int \frac{\partial}{\partial y} g(x,y)dx.$$

The issue is that the definition of a derivative involves a limit. The left hand side is

$$\lim_{h\to 0}\int \frac{g(x,y+h)-g(x,y)}{h}dx$$

while the right hand side is

$$\int \lim_{h\to 0}\frac{g(x,y+h)-g(x,y)}{h}dx$$

and the issue is whether or not you can pull limits in and out of integrals. You often can; the next two theorems give precise conditions for this to work.

Theorem 1 (Monotone Convergence) *If* $0 \le X_1 \le X_2 \le \cdots$ *and* $X = \lim X_n$ *(the limit* X *automatically exists) then*

$$\mathrm{E}(X) = \lim_{n\to\infty} \mathrm{E}(X_n)\,.$$

Remark: In the hypotheses we need $P(X_{n+1} \ge X_n) = 1$ and $P(X_1 \ge 0) = 1$.

Theorem 2 (Dominated Convergence) *If* $|X_n| \le Y_n$ *and* $\exists$ *a random variable* X *such that* $X_n \to X$ *(technical details of this convergence come later in the course) and a random variable* Y *such that* $Y_n \to Y$ *with* $\lim_{n\to\infty} \mathrm{E}(Y_n) = \mathrm{E}(Y) < \infty$ *then*

$$\lim_{n\to\infty} \mathrm{E}(X_n) = \mathrm{E}(X)\,.$$

Remark: The dominated convergence theorem is often used with all Y_n the same random variable Y. In this case the hypothesis that $\lim_{n\to\infty} \mathrm{E}(Y_n) = \mathrm{E}(Y) < \infty$ is just the hypothesis that $\mathrm{E}(Y) < \infty$.

Remark: These theorems are used in *approximation*. We compute the limit of the expected values of a sequence of random variables X_n and then approximate $\mathrm{E}(X_{225})$ (or whatever n we actually have instead of 225) by $\mathrm{E}(X)$.

Connection to ordinary integrals

Theorem 3 *With this definition of E:*

1. *if* X *has density* $f(x)$ *(even in* R^p *say) and* $Y = g(X)$ *then*

$$\mathrm{E}(Y) = \int g(x) f(x) dx\,.$$

   (This could be a multiple integral.)

2. *If* X *has probability mass function* f *then*

$$\mathrm{E}(Y) = \sum_x g(x) f(x)\,.$$

3. *The first conclusion works, e.g., even if* X *has a density but* Y *doesn't.*

Moments

- **Definition**: The r^{th} moment (about the origin) of a real random variable X is $\mu'_r = \mathrm{E}(X^r)$ (provided it exists).

- We generally use μ for $\mathrm{E}(X)$.

- **Definition**: The r^{th} central moment is

$$\mu_r = \mathrm{E}[(X-\mu)^r]$$

- We call $\sigma^2 = \mu_2$ the variance.

- **Definition**: For an R^p valued random vector X

$$\mu_X = \mathrm{E}(X)$$

  is the vector whose i^{th} entry is $\mathrm{E}(X_i)$ (provided all entries exist).

- **Definition**: The $(p \times p)$ variance covariance matrix of X is

$$\mathrm{Var}(X) = \mathrm{E}\left[(X-\mu)(X-\mu)^t\right]$$

  which exists provided each component X_i has a finite second moment.

Moments and independence

Theorem 4 *If $X_1, \ldots, X_p$ are independent and each X_i is integrable then $X = X_1 \cdots X_p$ is integrable and*

$$\mathrm{E}(X_1 \cdots X_p) = \mathrm{E}(X_1) \cdots \mathrm{E}(X_p)\,.$$

Proof: Suppose each X_i is simple:

$$X_i = \sum_j x_{ij} 1(X_i = x_{ij})$$

where the x_{ij} are the possible values of X_i. Then

$$\begin{aligned}
\mathrm{E}(X_1 \cdots X_p) &= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} \mathrm{E}(1(X_1 = x_{1j_1}) \cdots 1(X_p = x_{pj_p})) \\
&= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1} \cdots X_p = x_{pj_p}) \\
&= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1}) \cdots P(X_p = x_{pj_p}) \\
&= \sum_{j_1} x_{1j_1} P(X_1 = x_{1j_1}) \cdots \sum_{j_p} x_{pj_p} P(X_p = x_{pj_p}) \\
&= \prod \mathrm{E}(X_i)\,.
\end{aligned}$$

Non-negative Case: Now consider non-negative random variables X_i, Let X_{in} be X_i rounded down to the nearest multiple of 2^{-n} to a maximum of n. That is: if

$$\frac{k}{2^n} \le X_i < \frac{k+1}{2^n}$$

then $X_{in} = k/2^n$ for $k = 0, \ldots, n2^n$. For $X_i > n$ put$X_{in} = n$. Now apply the case we have just done:

$$\mathrm{E}(\prod X_{in}) = \prod \mathrm{E}(X_{in}) .$$

Monotone convergence applies to both sides to prove the result for non-negative X_i.

General case: now consider general X_i and write each X_i as the difference of positive and negative parts:

$$X_i = \max(X_i, 0) - \max(-X_i, 0) .$$

Write out $\prod_i |X_i|$ as a sum of products and apply the positive case to see that if all the X_i are integrable then so is $\prod_i X_i$.

Conditional Expectations

- Abstract definition of conditional expectation is:

- **Definition**: $\mathrm{E}(Y|X)$ is any function of X such that

  $$\mathrm{E}\left[R(X)\mathrm{E}(Y|X)\right] = \mathrm{E}\left[R(X)Y\right]$$

  for any bounded function $R(X)$.

- **Definition**: $\mathrm{E}(Y|X = x)$ is a function $g(x)$ such that

  $$g(X) = \mathrm{E}(Y|X)$$

- **Fact**: If X, Y has joint density $f_{X,Y}(x, y)$ and conditional density $f(y|x)$ then

  $$g(x) = \int y f(y|x) dy$$

  satisfies these definitions.

Proof:

$$
\begin{aligned}
\mathrm{E}(R(X)g(X)) &= \int R(x)g(x)f_X(x)dx \\
&= \int R(x)\int yf(y|x)dyf_X(x)dx \\
&= \int\int R(x)yf_X(x)f(y|x)dydx \\
&= \int\int R(x)yf_{X,Y}(x,y)dydx \\
&= \mathrm{E}(R(X)Y)
\end{aligned}
$$

Interpretation of conditional expectation

- **Intuition**: Think of $\mathrm{E}(Y|X)$ as average Y holding X fixed.
- Behaves like ordinary expected value but functions of X only are like constants:
$$\mathrm{E}(\sum A_i(X)Y_i|X) = \sum A_i(X)\mathrm{E}(Y_i|X)$$
- Statement called Adam's law by Jerzy Neyman – he used to say it comes before all the others:
$$\mathrm{E}[\mathrm{E}(Y|X)] = \mathrm{E}(Y)$$
which is just the definition of $\mathrm{E}(Y|X)$ with $R(X) \equiv 1$.
- In regression courses we say that the total sum of squares is the sum of the regression sum of squares plus the residual sum of squares:
$$\mathrm{Var(Y)} = \mathrm{Var}(\mathrm{E}(Y|X)) + \mathrm{E}[\mathrm{Var}(Y|X)]$$
- The conditional variance means
$$\mathrm{Var}(Y|X) = \mathrm{E}[(Y - \mathrm{E}(Y|X))^2|X].$$

Moments

Moment is an old word from physics used in such terms as moments of inertia. There is actually a good analogy between the physics use of the term and our use. If you made a block of wood shaped like the density of a random variable X and you tried to balance the block (it will be thin, long, flat on the bottom and curved into the shape of the density on the top) on a pencil the pencil would have to be located under the mean of the density. The *moment of force* about this pencil would be 0. Warning: go elsewhere to learn physics.

Definition: The r^{th} moment (about the origin) of a real random variable X is $\mu'_r = \text{E}(X^r)$ (provided it exists – that is, provided X^r is integrable).

Notation: We generally use μ for $\text{E}(X)$.

Definition: The r^{th} central moment is

$$\mu_r = \text{E}[(X-\mu)^r]$$

Notation: We call $\sigma^2 = \mu_2$ the variance.

Definition: For an R^p valued random vector X

$$\mu_X = \text{E}(X)$$

is the vector whose i^{th} entry is $\text{E}(X_i)$ (provided all entries exist). Similarly for matrices we take expected values entry by entry.

Definition: The $(p \times p)$ variance covariance matrix of X is

$$\text{Var}(X) = \text{E}\left[(X-\mu)(X-\mu)^t\right]$$

which exists provided each component X_i has a finite second moment.

The ijth entry in $(X-\mu)(X-\mu)^t$ is $(X_i-\mu_i)(X_j-\mu_j)$. As a result this matrix has diagonal entries which are the usual variances of the individual X_i and off diagonal entries which are covariances.

Moments and independence

Theorem 5 *If $X_1, \ldots, X_p$ are independent and each X_i is integrable then $X = X_1 \cdots X_p$ is integrable and*

$$\text{E}(X_1 \cdots X_p) = \text{E}(X_1) \cdots \text{E}(X_p)\,.$$

Proof: First suppose each X_i is simple:

$$X_i = \sum_j x_{ij} 1(X_i = x_{ij})$$

where the x_{ij} are the possible values of X_i. Then

$$\begin{aligned} E(X_1 \cdots X_p) &= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} E(1(X_1 = x_{1j_1}) \cdots 1(X_p = x_{pj_p})) \\ &= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1} \cdots X_p = x_{pj_p}) \\ &= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1}) \cdots P(X_p = x_{pj_p}) \\ &= \sum_{j_1} x_{1j_1} P(X_1 = x_{1j_1}) \cdots \sum_{j_p} x_{pj_p} P(X_p = x_{pj_p}) \\ &= \prod E(X_i) \,. \end{aligned}$$

Now we consider the case of general $X_i \geq 0$. Let X_{in} be X_i rounded down to nearest multiple of 2^{-n} (to maximum of n). That is, if

$$\frac{k}{2^n} \leq X_i < \frac{k+1}{2^n}$$

then we define $X_{in} = k/2^n$ for $k = 0, \ldots, n2^n$ and for $X_i > n$ we put $X_{in} = n$.

Now we apply the case we have just done:

$$E(\prod X_{in}) = \prod E(X_{in}) \,.$$

Finally we apply the monotone convergence theorem to both sides.

It remains to consider X_i which might not be positive. Use the previous case to prove that

$$|\prod X_i| = \prod |X_i|$$

is integrable. Then expend the product of positive minus negative parts,

$$X_i = \max(X_i, 0) - \max(-X_i, 0) \,.$$

Next check that all of the 2^p terms you get, after expanding out, are integrable and apply the previous case. The details are algebraically messy and not very informative in my view. An alternative theory is that I am too lazy to write them out.

Conditional Expectations

I am going to give here the abstract "definition" of a conditional expectation. The definition is indirect – it is a thing which has a certain property. That means that I ought to prove there is a thing with that property and that the thing with the property is unique. As usual – I won't be doing that here.

The abstract definition of conditional expectation is:

Definition: $\mathrm{E}(Y|X)$ is any function of X such that

$$\mathrm{E}\left[R(X)\mathrm{E}(Y|X)\right] = \mathrm{E}\left[R(X)Y\right]$$

for any bounded function $R(X)$.

Definition: $\mathrm{E}(Y|X = x)$ is a function $g(x)$ such that

$$g(X) = E(Y|X)$$

that is, such that $g(X)$ satisfies the previous definition.

Fact: If X, Y has joint density $f_{X,Y}(x, y)$ and conditional density $f(y|x)$ then

$$g(x) = \int yf(y|x)dy$$

satisfies these definitions.

Proof:

$$\begin{aligned}
E(R(X)g(X)) &= \int R(x)g(x)f_X(x)dx \\
&= \int R(x) \int yf(y|x)dyf_X(x)dx \\
&= \int \int R(x)yf_X(x)f(y|x)dydx \\
&= \int \int R(x)yf_{X,Y}(x, y)dydx \\
&= E(R(X)Y)
\end{aligned}$$

Interpretation and properties of conditional expectation

- **Intuition**: Think of $E(Y|X)$ as average Y holding X fixed.
- Behaves like ordinary expected value but functions of X only are like constants:
$$E(\sum A_i(X)Y_i|X) = \sum A_i(X)E(Y_i|X)$$
- Statement called Adam's law by Jerzy Neyman – he used to say it comes before all the others:
$$E[E(Y|X)] = E(Y)$$
which is just the definition of $E(Y|X)$ with $R(X) \equiv 1$.
- In regression courses we say that the total sum of squares is the sum of the regression sum of squares plus the residual sum of squares:
$$\text{Var(Y)} = \text{Var}(E(Y|X)) + E[\text{Var}(Y|X)]$$
- The conditional variance means
$$\text{Var}(Y|X) = E[(Y - E(Y|X))^2|X].$$