CALIFORNIA INSTITUTE OF TECHNOLOGY

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2015

Lecture 6: Higher expectations; Densities

Relevant textbook passages:

Pitman [4]: Chapter 3

Larsen–Marx [3]: Chapter 3

6.1 Bonus Question

Since a long weekend is coming up, I am giving you a bonus question. The first two answers that I receive from students in the class and that I judge worthy will be my guests for lunch at the Athenaeum.

> Alex tosses a fair coin n independent times and Blair tosses a fair coin m independent times. Find an *elegant* or clever argument to compute the probability that they have equal numbers of Tails. (I will be the judge of whether the argument is elegant, but it had better not involve any lengthy sums.)

6.2 Independent random variables

Pitman [4]: pp. 151–154

> **6.2.1 Definition** *X and Y are* ***independent random variables*** *if for every $B_1, B_2 \subset \boldsymbol{R}$,* [a]
>
> $$P(X \in B_1 \text{ and } Y \in B_2) = P(X \in B_1) \cdot P(Y \in B_2)$$
>
> *More generally, a set $\mathcal{X}$ of random variables is* ***stochastically independent*** *if for every finite subset of random variables $X_1, \dots, X_n$ of $\mathcal{X}$ and every collection $B_1, \dots, B_n$ of subsets*[1] *of $\boldsymbol{R}$,*
>
> $$P(X_1 \in B_1, \dots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n).$$
>
> [a] Caveat: B_i must be a Borel set.

6.2.2 Example (Pairwise independence does not imply independence) Let X and Y be independent Bernoulli(1/2) random variables (coin tosses), and let Z

be the parity of $X+Y$. Then X and Y are stochastically independent, Y and Z are stochastically independent, and X and Z are stochastically independent; but the set X, Y, Z is *not* stochastically independent.

You will be asked to prove this in the homework. □

6.2.3 Definition *A sequence $X_1, X_2, \dots$ (finite or infinite) is* ***independent and identically distributed****, abbreviated* ***i.i.d.****, if they have a common distribution function and are stochastically independent.*

6.3 Continuous distributions, densities, and expectation

If X is not discrete, the cumulative distribution function still makes sense as we dfined it earlier, and there is an analog of the probability mass function.

Pitman [4]: § 4.1
Larsen–Marx [3]: § 3.4

If the cumulative distribution function F of X is differentiable everywhere,[1] its derivative is called the **probability density function** or simply the **density** of X.

Note that since the cumulative distribution function F is nondecreasing, if the density exists, it is nonnegative.

If a random variable X has cumulative distribution function F and density f, then

$$P\left(X \in [a,b]\right) = F(b) - F(a) = \int_a^b f(x)\,dx,$$

and

$$P\left(X \in [a,b]\right) = P\left(X \in (a,b)\right) = P\left(X \in (a,b]\right) = P\left(X \in [a,b)\right).$$

The definition of expectation that I gave for discrete random variables has the following analog for random variables with a density.

6.3.1 Definition *If X has a density f, we can define its expectation using the density:*

$$\boldsymbol{E}\, X = \int_{\boldsymbol{R}} x f(x)\,dx,$$

provided $\int_{\boldsymbol{R}} |x| f(x)\,dx$ is finite.

[1] Aside for math majors:

Actually, all we need is that F is **absolutely continuous**. Then it is differentiable almost everywhere and is the indefinite integral of its derivative. For an example of a cumulative distribution function that is not absolutely continuous, look up the **Cantor ternary function** c. It has the property that $c'(x)$ exists almost everywhere and $c'(x) = 0$ everywhere it exists, but nevertheless c is continuous and $c(0) = 0$ and $c(1) = 1$. It is the cumulative distribution function of a distribution supported on the Cantor set.

Aside: If a random variable has a continuous distribution, its underlying sample space S must be uncountably infinite. This means the that the set $\mathcal{E}$ of events will not consist of all subsets of S. I will largely ignore the difficulties that imposes, but in case you're interested, the "real" definition of the expectation of X is as the abstract Lebesgue integral of X with respect to the probability P on S, written $\boldsymbol{E}\,X = \int_S X\,dP$ or $\int_S X(s)\,dP(s)$. Summation is just a special case of abstract Lebesgue integration when the probability measure is discrete.

6.4 Expectation of a function of a random variable with a density

Larsen–Marx [3]: § 3.5

For a random variable X with a density f, we have that $g \circ X$ is also a random variable,[2] and

$$\boldsymbol{E}\, g \circ X = \int_{\boldsymbol{R}} g(x) f(x)\,dx,$$

provided $\int_{\boldsymbol{R}} |g(x)| f(x)\,dx$ is finite.

6.5 An example: Uniform$[a, b]$

A random variable U with the **Uniform$[a, b]$ distribution**, where $a < b$, has the cumulative distribution function F defined by

$$F(t) = \begin{cases} 0 & t < a, \\ 1 & t > b, \\ \dfrac{t-a}{b-a} & a \leqslant t \leqslant b \end{cases}$$

(that is, $F(a) = 0$, $F(b) = 1$, and F is linear in between) and density f defined by

$$f(t) = \begin{cases} 0 & t < a, \\ 0 & t > b, \\ \dfrac{1}{b-a} & a \leqslant t \leqslant b. \end{cases}$$

The density is constant on $[a, b]$ and its value is chosen so that $\int_a^b f(x)\,dx = 1$.

The expectation is just

$$\boldsymbol{E}\,U = \frac{a+b}{2}.$$

We will explore more distributions as we go along.

[2] Once again there is the mysterious caveat that g must be a **Borel function**. All step functions, and all continuous functions are Borel functions, as are all linear combinations and limits of sequences of such functions.

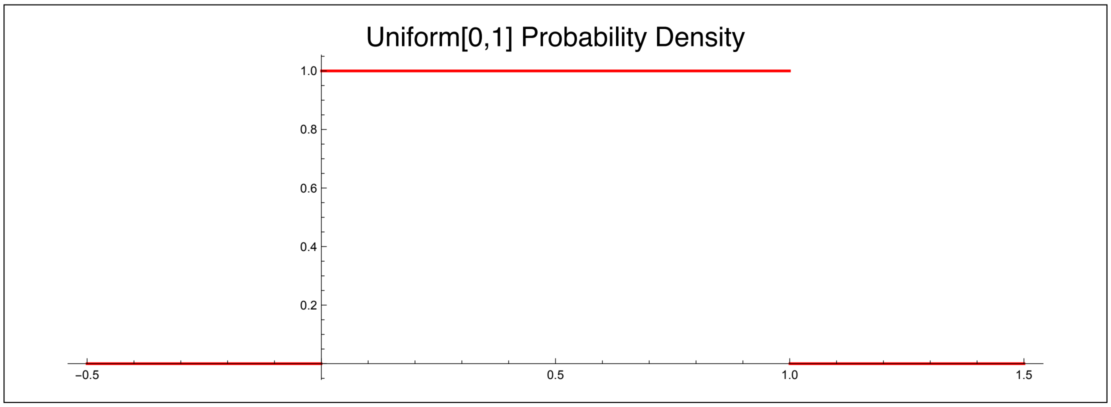

Figure 6.1. The Uniform$[0,1]$ pdf.

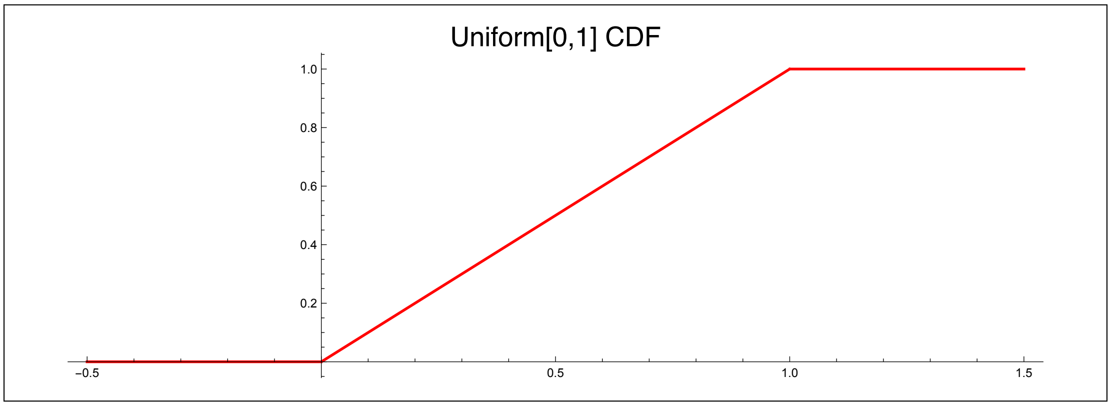

Figure 6.2. The Uniform$[0,1]$ cdf.

6.6 Expectation is a positive linear operator!!

Since random variables are just real-valued functions on a sample space S, we can add them and multiply them just like any other functions. For example, the sum of random variables X and Y is given by

$$(X+Y)(s) = X(s) + Y(s).$$

Thus the set of random variables is a vector space. In fact, the subset of random variables that have a finite expectation is also a vector subspace of the vector space of all random variables, due to the following simple results:

Pitman [4]: pp. 181 ff.

- Expectation is a **linear operator**. This means that

$$\boldsymbol{E}(aX+bY) = a\,\boldsymbol{E}\,X + b\,\boldsymbol{E}\,Y.$$

Proof: The Distributive Law. Here's the case for discrete random variables.

$$\begin{aligned}\boldsymbol{E}(aX+bY) &= \sum_{s\in S}\Big(aX(s)+bY(s)\Big)P(s)\\ &= a\sum_{s\in S}X(s)P(s) + bsum_{s\in S}Y(s)P(s)\\ &= a\,\boldsymbol{E}\,X + b\,\boldsymbol{E}\,Y.\end{aligned}$$

Save these formulas on the board!

- Expectation is a **positive operator**. That is, if $X \geqslant 0$, i.e., $X(s) \geqslant 0$ for each $s \in S$, then $\boldsymbol{E}\,X \geqslant 0$.
- If $X \geqslant Y$, then $\boldsymbol{E}\,X \geqslant \boldsymbol{E}\,Y$.

Proof: Let $X \geqslant Y$, and observe that $X - Y \geqslant 0$. Write

$$X = Y + (X - Y),$$

so since expectation is a linear operator, we have

$$\boldsymbol{E}\,X = \boldsymbol{E}\,Y + \boldsymbol{E}(X-Y).$$

Since expectation is a positive operator, $\boldsymbol{E}(X-Y) \geqslant 0$, so we conclude.

$$\boldsymbol{E}\,X \geqslant \boldsymbol{E}\,Y.$$

■

Chant!!

Special Cases:

- If X is degenerate (constant), say $P(X=c) = 1$, then $\boldsymbol{E}\,X = c$.
- So $\boldsymbol{E}(\boldsymbol{E}\,X) = \boldsymbol{E}\,X$.

- So $\boldsymbol{E}(X - \boldsymbol{E}\,X) = 0$.
- For an indicator function $\mathbf{1}_A$,

$$\boldsymbol{E}\,\mathbf{1}_A = P(A).$$

Proof:

$$\boldsymbol{E}\,\mathbf{1}_A = \sum_{s \in S} \mathbf{1}_A(s) P(s) = \sum_{s \in A} P(s) = P(A).$$

∎

- $\boldsymbol{E}(cX) = c\,\boldsymbol{E}\,X$. (This is a special case of linearity.)
- $\boldsymbol{E}(X + c) = \boldsymbol{E}\,X + c$. (This is a special case of linearity.)

6.7 Summary of positive linear operator properties

$$\begin{gathered}
\boldsymbol{E}(aX + bY) = a\,\boldsymbol{E}\,X + b\,\boldsymbol{E}\,Y \\
X \geqslant 0 \implies \boldsymbol{E}\,X \geqslant 0 \\
X \geqslant Y \implies \boldsymbol{E}\,X \geqslant \boldsymbol{E}\,Y \\
P\left(X = c\right) = 1 \implies \boldsymbol{E}\,X = c \\
\boldsymbol{E}(\boldsymbol{E}\,X) = \boldsymbol{E}\,X \\
\boldsymbol{E}(X - \boldsymbol{E}\,X) = 0 \\
\boldsymbol{E}\,\mathbf{1}_A = P(A) \\
\boldsymbol{E}(cX) = c\,\boldsymbol{E}\,X \\
\boldsymbol{E}(X + c) = \boldsymbol{E}\,X + c \\
\boldsymbol{E}(aX + c) = a\,\boldsymbol{E}\,X + c
\end{gathered}$$

See the chart in Pitman [4, p. 181].

6.8 Expectation of an independent product

6.8.1 Theorem *Let X and Y be* independent *random variables on the common probability space $(S, \mathcal{E}, P)$, with finite expectations. Then*

$$\boldsymbol{E}(XY) = (\boldsymbol{E}\,X)(\boldsymbol{E}\,Y).$$

Proof: I'll prove this for the discrete case. In what follows, the sum is over the range of X and Y.

$$\begin{aligned}
\boldsymbol{E}(XY) &= \sum_{(x,y)} xyP\left(X=x \text{ and } Y=y\right) && \text{definition of expectation}\\
&= \sum_{(x,y)} xyP\left(X=x\right)P\left(Y=y\right) && \text{by independence}\\
&= \sum_{x}\left(xp_X(x)\left(\sum_{y} yp_Y(y)\right)\right) && \text{Distributive Law}\\
&= \sum_{x} xp_X(x)\,\boldsymbol{E}\,Y && \text{definition of expectation}\\
&= (\boldsymbol{E}\,Y)\left(\sum_{x} xp_X(x)\right) && \text{linearity of expectation}\\
&= (\boldsymbol{E}\,Y)(\boldsymbol{E}\,Y) && \text{definition of expectation.}
\end{aligned}$$

■

6.9 Jensen's Inequality

This section is not covered in Pitman [4]!

6.9.1 Definition *A function $f\colon I \to \boldsymbol{R}$ on an interval I is* ***convex*** *if*

$$f\Big((1-t)x+ty\Big) \leqslant (1-t)f(x)+tf(y)$$

for all x, y in I with $x \neq y$ and all $0 < t < 1$.

A function $f\colon I \to \boldsymbol{R}$ on an interval I is ***strictly convex*** *if*

$$f\Big((1-t)x+ty\Big) < (1-t)f(x)+tf(y)$$

for all x, y in I with $x \neq y$ and all $0 < t < 1$.

Another way to say this is that the line segment joining any two points on the graph of f lies above the graph.

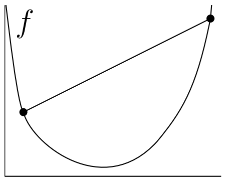

Figure 6.3. A (strictly) convex function.

6.9.2 Fact *If f is convex on an interval $[a, b]$, then f is continuous on (a, b).*

Let f be twice differentiable everywhere on (a, b). Then f is convex on (a, b) if and only if $f''(x) \geqslant 0$ for all $x \in (a, b)$. If $f''(x) > 0$ for all x, then f is strictly convex.

If f is convex on the interval $[a, b]$, then for every x and y in I, if f is differentiable at x, then

$$f(y) \geqslant f(x) + f'(x)(y - x).$$

The geometric interpretation of this is that if f is convex, then its graph lies above the tangent line to the graph.

Even if f is not differentiable at $x \in (a, b)$, say it has a "kink" at x, there is a slope m (called a ***subderivative****) such that for all $y \in [a, b]$, we have*

$$f(y) \geqslant f(x) + m(y - x).$$

Define: A random variable X is called **degenerate** if there is some x with $P(X = x) = 1$, that is, it isn't really random in the usual sense of the word. Otherwise it is **nondegenerate**.

6.9.3 Theorem (Jensen's Inequality) *Let X be a random variable with finite expectation, and let $f\colon \boldsymbol{R} \to \boldsymbol{R}$ be a convex function whose domain includes the range of X. Then*

$$\boldsymbol{E}\Big(f(X)\Big) \geqslant f(\boldsymbol{E}\, X).$$

If X is nondegenerate and f is strictly convex, then the inequality is strict.

Sketch of proof for X simple: Let X have range $\{x_1, \dots, x_n\}$ and probability mass function$p$. Then

$$\boldsymbol{E}\, f(X) = \sum_{i=1}^{n} f(x_i)p(x_i) \geqslant f\left(\sum_{i=1}^{n} x_i p(x_i)\right) = f(\boldsymbol{E}\, X).$$

A more general proof makes use of the fact for all x and y in the domain of f, $f(x) \geqslant f(y) + m(x - y)$, where m is a subderivative of f at y. Take $y = \boldsymbol{E}\, X$ and take the expectation on both sides. (Unless X is degenerate, $\boldsymbol{E}\, X$ is in the interior of the domain of f, so a subderivative exists. If X is degenerate, then we have equality.)

Jensen's Inequality is named for the Danish mathematician Johan Jensen [2].

.

Some consequences of this are:

- Let X be a positive nondegenerate random variable. Then,

$$\boldsymbol{E}\left(\frac{1}{X}\right) > \frac{1}{\boldsymbol{E}\,X}$$

since $f(x) = 1/x$ is strictly convex.

- Let X be a nondegenerate random variable. Then

$$\boldsymbol{E}(X^2) > (\boldsymbol{E}\,X)^2,$$

since $f(x) = x^2$ is strictly convex.

6.10 Variance and Higher "Moments"

Let X be a random variable with finite expectation.

Pitman [4]: § 3.3
Larsen–Marx [3]: § 3.6

The **variance** of X is defined to be

$$\begin{aligned}\boldsymbol{Var}\,X = \boldsymbol{E}(X - \boldsymbol{E}\,X)^2 &= \boldsymbol{E}\Big(X^2 - 2X \cdot \boldsymbol{E}\,X + (\boldsymbol{E}\,X)^2\Big) \\ &= \boldsymbol{E}\,X^2 - 2\,\boldsymbol{E}(X \cdot \underbrace{\boldsymbol{E}\,X}_{\text{a constant}}) + (\boldsymbol{E}\,X)^2 \\ &= \boldsymbol{E}\,X^2 - 2(\boldsymbol{E}\,X)\,\boldsymbol{E}\,X + (\boldsymbol{E}\,X)^2 \\ &= \boldsymbol{E}\,X^2 - (\boldsymbol{E}\,X)^2.\end{aligned}$$

provided the expectation is finite. (We might also say that a random variable has infinite variance.)

The **standard deviation** of X, denoted s. d. X, is just the square root of its variance.

The set of random variables with finite variance is also a vector space.

The variance is a measure of "dispersion."

6.10.1 Proposition $\boldsymbol{Var}(aX + b) = a^2\,\boldsymbol{Var}\,X$.

6.10.2 Theorem *If X and Y are independent random variables with finite variance, then*

$$\boldsymbol{Var}(X + Y) = \boldsymbol{Var}\,X + \boldsymbol{Var}\,Y$$

Proof: By definition,

$$\begin{aligned}
\boldsymbol{Var}(X+Y) &= \boldsymbol{E}\Big(X+Y-\boldsymbol{E}(X+Y)\Big)^2 \\
&= \boldsymbol{E}\Big((X-\boldsymbol{E}\,X)+(Y-\boldsymbol{E}\,Y)\Big)^2 \\
&= \boldsymbol{E}\left((X-\boldsymbol{E}\,X)^2+2(X-\boldsymbol{E}\,X)(Y-\boldsymbol{E}\,Y)+(Y-\boldsymbol{E}\,Y)^2\right) \\
&= \boldsymbol{E}(X-\boldsymbol{E}\,X)^2+2\,\boldsymbol{E}(X-\boldsymbol{E}\,X)(Y-\boldsymbol{E}\,Y)+\boldsymbol{E}(Y-\boldsymbol{E}\,Y)^2 \\
&= \ \boldsymbol{Var}\,X+2\,\boldsymbol{E}(X-\boldsymbol{E}\,X)(Y-\boldsymbol{E}\,Y)+\ \boldsymbol{Var}\,Y.
\end{aligned}$$

But by independence

$$\boldsymbol{E}(X-\boldsymbol{E}\,X)(Y-\boldsymbol{E}\,Y)=\boldsymbol{E}(X-\boldsymbol{E}\,X)\,\boldsymbol{E}(Y-\boldsymbol{E}\,Y)=0\cdot 0=0.$$

■

6.10.3 Example • The variance of Bernoulli(p): A Bernoulli(p) random variable X has expectation p, so the variance is given by

$$\sum_{x=0}^{1}(x-p)^2\times P\,(X=x)=(1-p)^2p+(0-p)^2(1-p)=p-p^2.$$

• The Binomial(n,p) distribution can be described as the distribution of the sum of n Bernoulli(p) random variables. Thus its variance is sum of the variances of n Bernoulli(p) random variables. That is,

$$n(p-p^2).$$

• The variance of a Uniform[0,1] random variable (which has density one on $[0,1]$ and expectation 1/2) is

$$\int_0^1 (x-1/2)^2\,dx=\int_0^1 x^2-x+1/4\,dx=1/3-1/2+1/4=1/12.$$

□

6.10.1 Higher moments

6.10.4 Definition *The* $\boldsymbol{n}^{\mathbf{th}}$ ***central moment of*** $\boldsymbol{X}$ *is*

$$\boldsymbol{E}(X-\boldsymbol{E}\,X)^n$$

The $\boldsymbol{n}^{\mathbf{th}}$ ***moment of*** $\boldsymbol{X}$ *is*

$$\boldsymbol{E}\,X^n.$$

6.11 Higher moments imply lower moments

There is a theorem called **Hölder's Inequality** that has as one of its consequences the following result.

6.11.1 Fact *If $1 \leqslant p < q$, if $\boldsymbol{E}|X^q|$ is finite, then $\boldsymbol{E}|X^p|$ is finite.*

This result appears in any good analysis text, but out of laziness I'll just cite [1, Corollary 13.3, p. 463].

Bibliography

[1] C. D. Aliprantis and K. C. Border. 2006. *Infinite dimensional analysis: A hitchhiker's guide*, 3d. ed. Berlin: Springer–Verlag.

[2] J. L. W. V. Jensen. 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica* 30(1):175–193. DOI: 10.1007/BF02418571

[3] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[4] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.