Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2017

Lecture 6: Expectation is a positive linear operator

Relevant textbook passages:
Pitman [3]: Chapter 3
Larsen–Marx [2]: Chapter 3

6.1 Non-discrete random variables and distributions

So far we have restricted attention to discrete random variables. And in practice any measurement you make will be a rational number. But there are times when it is actually easier to think in terms of random variables whose values might be any real number. This means we have to deal with nondenumerable sample spaces, which can lead to technical difficulties that I shall mostly ignore. The distribution of a random variable X, and its cumulative distribution function are well defined as above, but we need to replace the notion of a probability mass function with something we call a **probability density function**. We will also replace sums by integrals (which are, after all, just limits of sums).

6.2 Absolutely continuous distributions, densities, and expectation

A random variable X is **absolutely continuous** if its cumulative distribution function F is an indefinite integral, that is, if there is some function $f\colon \boldsymbol{R} \to \boldsymbol{R}$ such that for every x, $f(x) \geqslant 0$, and for every interval $[a, b]$ with $a \leqslant b$,

$$F(b) - F(a) = \int_a^b f(x)\,dx.$$

The function f is called the **density** of F (or of X). If F is absolutely continuous, then it will have a derivative almost everywhere, and it is the indefinite integral of its derivative. If the cumulative distribution function F of X is differentiable everywhere, its derivative is its density. (Sometimes we get a bit careless, and simply refer to an absolutely continuous cumulative distribution function as continuous.[1]) The **support** of a distribution with density f is the closure[2] of $\{x : f(x) > 0\}$.[3]

Pitman [3]: § 4.1
Larsen–Marx [2]: § 3.4

[1] For math majors: For an example of a cumulative distribution function that is continuous, but 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 (but not absolutely continuous) and $c(0) = 0$ and $c(1) = 1$. It is the cumulative distribution function of a distribution supported on the Cantor set. You'll learn about this in **Ma 108**.

[2] The closure of a set is the set of all its limit points.

[3] For math majors: You can change the density at single point and it remains a density (its integral doesn't change. In fact you can change it on any set of measure zero (if you don't know what that means, I might write up an appendix) and it remains a density for the distribution. These different densities are called **versions** of each other. They all give rise to the same cumulative distribution function and the same support.

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

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

and

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

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

6.2.1 Definition *If X has a density f, we 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.

Aside: If a random variable has an absolutely 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.

6.3 What makes a density?

Any function $f\colon \boldsymbol{R} \to \boldsymbol{R}_+$ such that

$$\int_{-\infty}^{\infty} f(x)\,dx < \infty$$

can be turned into a probability density function by **normalizing** it. That is, if the real number satisfies $c = \int_{-\infty}^{\infty} f(x)\,dx$, then $f(x)/c$ is a probability density. The constants c are sometimes called normalizing constants, and they account for the odd look of many densities.

For instance, $\int_{-\infty}^{\infty} e^{-z^2/2}\,dz = \sqrt{2\pi}$ is the normalizing constant for the Normal family.

Much space is devoted in introductory statistics and probability textbooks to computing various integrals. In this course, I shall not spend time in lecture on the details of evaluating integrals. My view is that the evaluation of integrals, while a necessary part of the subject, frequently offers little insight. You all have had serious calculus classes recently, so you are probably better at integration than I am these days. On the occasions where it does provide some insight, we may spend some time on it. I do recommend the exposition in Pitman [3, Section 4.4, pp. 302–310] and Larsen–Marx [2, Section 3.8, pp. 176–183].

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

Larsen–Marx [2]: § 3.5

For a random variable X with a density f, we have that $g \circ X$ is also a random variable,[4] and

[4] 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.

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

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

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

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

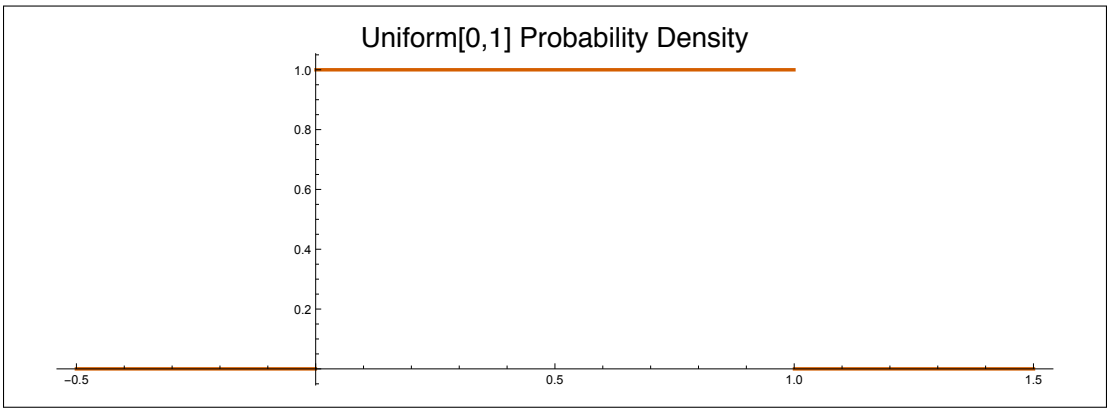


Figure 6.1. The Uniform$[0, 1]$ pdf.

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

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

The expectation is

$$\boldsymbol{E}\, U = \int_a^b x f(x)\, dx = \int_a^b \frac{x}{b-a}\, dx = \frac{1}{b-a}\frac{2}{b^2-a^2} = \frac{a+b}{2},$$

which is just the midpoint of the interval.

We will explore more distributions as we go along.

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

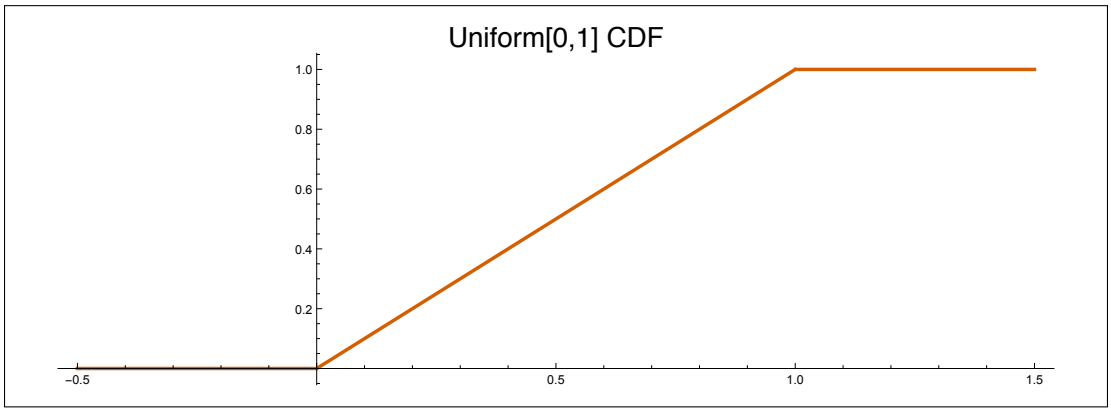


Figure 6.2. The Uniform$[0,1]$ cdf.

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 $L_1(P)$ 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 [3]: pp. 181 ff.

- Expectation is a **linear operator** on $L_1(P)$, 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.

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

- 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, $\boldsymbol{E}(X - Y) \geqslant 0$, and since it is a linear operator $\boldsymbol{E}(X - Y) = \boldsymbol{E}\,X - \boldsymbol{E}\,Y$, so

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

∎

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

6.7.1 Proposition *In summary, for random variables with finite expectation (those in $L_1(P)$):*

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

See the chart in Pitman [3, 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.

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

∎

6.9 Jensen's Inequality

This section is not covered in Pitman [3]!

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. See Figure 6.3.

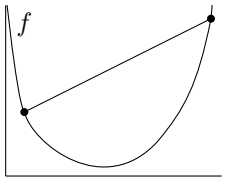


Figure 6.3. A (strictly) convex function.

6.9.2 Fact *Here are some useful properties of convex functions.*

- *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.*

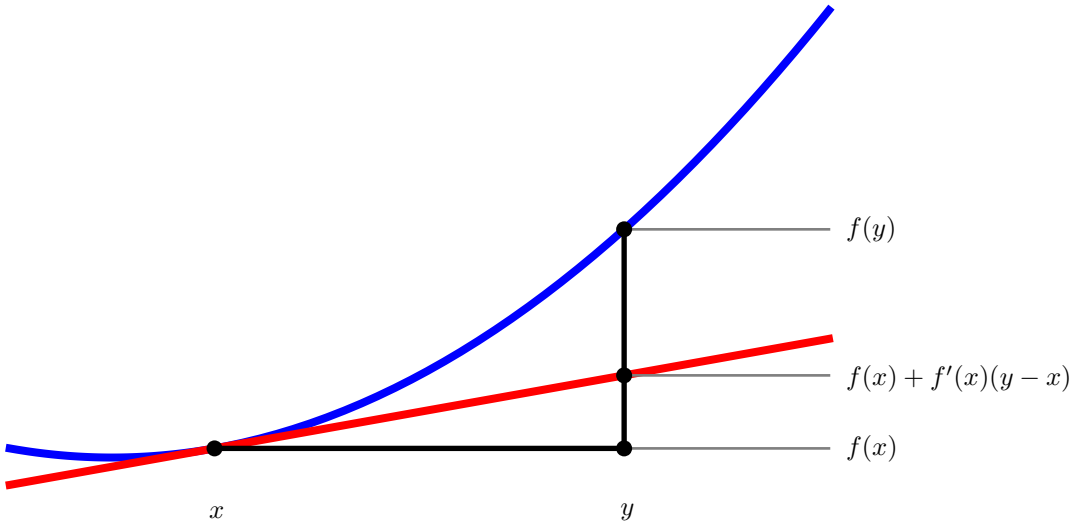


Figure 6.4. The Subgradient Inequality.

- *If f is convex on the interval $[a, b]$, then for every x and y in I, if f is differentiable at x, then we have the* ***Subgradient Inequality****:*

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

- *The geometric interpretation of this is that if f is convex, then its graph lies above the tangent line to the graph. See Figure 6.4.*

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*

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

For instance, the absolute value function has a kink at 0 and any $m \in [-1, 1]$ is a subderivative there.

In fact, f is differentiable at an interior point x if and only if it has a unique subderivative, in which case the subderivative is the derivative $f'(x)$.

- *If f is strictly convex, $x \neq y$, and if m is a subderivative of f at x, then the Subgradient Inequality is strict:*

$$f(y) > f(x) + m(y - x).$$

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

6.9.4 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 the function f is strictly convex, then the inequality holds with equality if and only if X is degenerate.

Proof: For convenience, let $\mu = \boldsymbol{E}\, X$. By the Subgradient Inequality, if f is differentiable at μ,

$$f(X) \geqslant f(\mu) + f'(\mu)(X - \mu).$$

(Even if f is not differentiable, we can replace $f'(\mu)$ by a subderivative.) Since expectation is a **positive linear operator**, we have

$$\boldsymbol{E}\, f(X) \geqslant \underbrace{\boldsymbol{E}\, f(\mu)}_{=f(\boldsymbol{E}\, X)} + f'(\mu)\, \underbrace{\boldsymbol{E}(X - \mu)}_{=0}.$$

The claim about degeneracy follows from the strictness of the Subgradient Inequality for strictly convex functions. ∎

Jensen's Inequality is named for the Danish mathematician Johan Jensen [1], so it should be pronounced Yen-sen.

Some consequences of Jensen's Inequality 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 on the interval $(x > 0)$.

- 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"

Pitman [3]: § 3.3
Larsen–Marx [2]: § 3.6

Let X be a random variable with finite expectation.

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

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

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

The **standard deviation** of X, denoted $\mathrm{SD}\, X$, is just the square root of its variance.

The variance is often denoted σ^2 and the standard deviation by σ.

- One of the virtues of the standard deviation over the variance of X is that it is in the same units as X.

- The set of random variables with finite variance is also a vector space, known as $L_2(P)$,, or more simply as L_2.

• The standard deviation is the L_2 norm of $X - \boldsymbol{E}\,X$. (Don't worry if you heard of the L_2 norm before, but it behaves like the Euclidean norm on $\boldsymbol{R}^{\mathrm{n}}$.)

The variance is a measure of the "dispersion" of the random variable's distribution about its mean.

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

Proof: To simplify things, let $\mu = \boldsymbol{E}\,X$. Then since expectation is a linear operator, $\boldsymbol{E}(aX+b) = a\mu + b$, and

$$\boldsymbol{Var}(aX + b) = \boldsymbol{E}\Big[\big((aX + b) - (a\mu + b)\big)^2\Big] = \boldsymbol{E}\Big[\big(a(X - \mu)\big)^2\Big] = a^2\,\boldsymbol{E}\big((X - \mu)^2\big) = a^2\ \boldsymbol{Var}\,X.$$

∎

6.11 Why variance?

A student stopped me at lunch one day to chat about measures of dispersion. He was curious as to who invented variance, and why it is used so much. Another sensible measure of the dispersion of X is $\boldsymbol{E}|X - \mu|$, which I'll call the **mean absolute deviation from the mean**. Pitman [3, Problem 3.3.26] leaves it as an exercise to prove the interesting fact that

$$\mathrm{SD}\,X \geqslant \boldsymbol{E}|X - \mu|.$$

You will be asked to prove this as an exercise at some point.

One reason for the popularity of variance is that it is easier to work with. For instance,in a moment we shall prove Theorem 6.11.1, which asserts that the variance of the sum of two independent random variables is the sum of the variances. To my knowledge there is no analog of this for mean absolute deviation. That is, if X and Y are independent, and for simplicity's sake we'll assume that $\boldsymbol{E}\,X = \boldsymbol{E}\,Y = 0$, then $\boldsymbol{Var}(X + Y) = \boldsymbol{Var}\,X + \boldsymbol{Var}\,Y$, but all I can say is that $\boldsymbol{E}|X + Y| \leqslant \boldsymbol{E}|X| + \boldsymbol{E}|Y|$.

The variance relation plays a central role in the Law of Large Numbers and in the Central Limit Theorem, and I don't know how to reformulate these in terms of mean absolute deviation.

6.11.1 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

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

∎

6.11.2 Example Here are the variances of some familiar distributions.

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

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

- 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

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

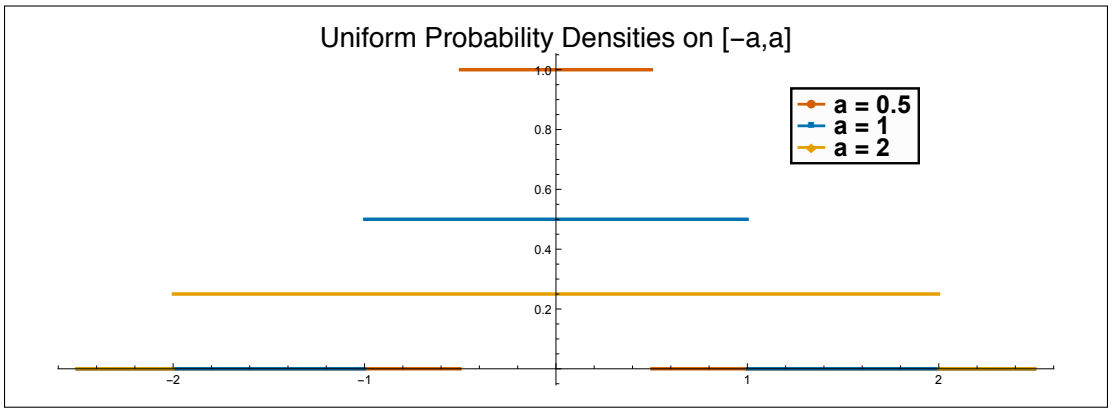
- For $a > 0$, the variance of a Uniform$[-a,a]$ random variable (which has density $1/2a$) on $[-a,a]$ and expectation 0) is

$$\int_{-a}^{a} \frac{x^2}{2a}\,dx = \frac{a^2}{3}.$$

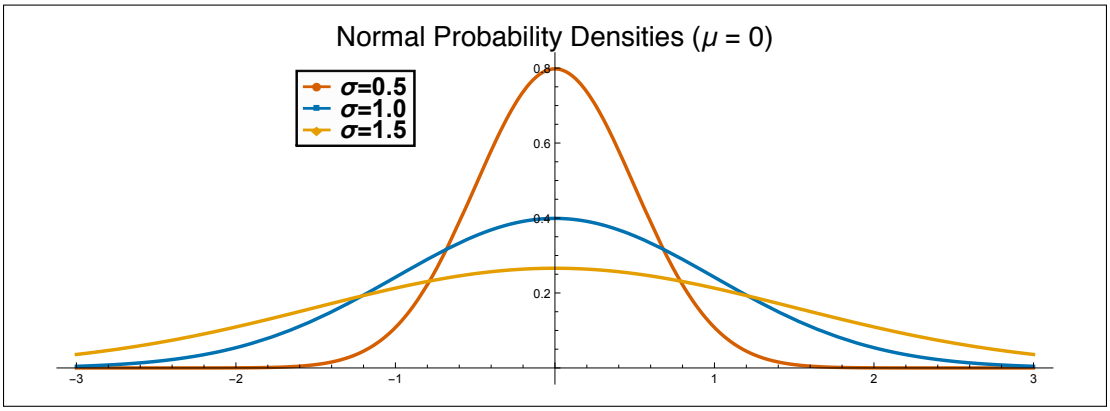
□

6.11.3 Example Here are some diagrams of densities that show the effect of increasing variance.

- Uniform densities on $[-a,a]$. The variance $a^2/3$ is increasing in a:



- "Normal" densities:



Increasing the variance spreads out and flattens the densities. □

6.12 Standardized random variables

Pitman [3]: p. 190

6.12.1 Definition *Given a random variable X with finite mean μ and variance σ^2, the* ***standardization*** *of X is the random variable X^* defined by*

$$X^* = \frac{X - \mu}{\sigma}.$$

Note that

$$\boldsymbol{E}\, X^* = 0, \quad \text{and} \quad \boldsymbol{Var}\, X^* = 1,$$

and

$$X = \sigma X^* + \mu,$$

so that X^* is just X measured in different units, called ***standard units****.*

[Note: Pitman uses both X^* and later X_* to denote the standardization of X.]

A convenient feature of standardized random variables is that they are invariant under change of scale and location.

6.12.2 Proposition *Let X be a random variable with mean μ and standard deviation σ, and let $Y = aX + b$, where $a > 0$. Then*

$$X^* = Y^*.$$

Proof: The proof follows from Propositions 6.7.1 and 6.10.1, which assert that $\boldsymbol{E}\, Y = a\mu + b$ and $\operatorname{SD} Y = a\sigma$. So

$$Y^* = \frac{Y - a\mu - b}{a\sigma} = \frac{\overbrace{aX + b}\, Y - a\mu - b}{a\sigma} = \frac{a(X - \mu)}{a\sigma} = \frac{X - \mu}{\sigma} = X^*.$$

∎

Bibliography

[1] 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

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

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