Lecture 6: Expectation is a positive linear operator

Relevant textbook passages:

Pitman [12]: Chapters 3 and 5; Section 6.4–6.5

Larsen–Marx [10]: 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 with cumulative distribution function F is **absolutely continuous** if there is some function f is called a **probability density**, or more simply a **density**, such that

$$f(x) \geqslant 0 \quad \text{for all } x,$$

and F is an indefinite integral of f, that is, for every interval $[a, b]$ with $a \leqslant b$,

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

The density may be referred to as either the density of X or the density of F.

Pitman [12]: § 4.1

Larsen–Marx [10]: § 3.4

6.2.1 Remark The First Fundamental Theorem of Calculus, e.g., Apostol [3, Theorem 5.1, p. 202], asserts that if a cumulative distribution function F is absolutely continuous, then it will have a derivative at each point of continuity of f, and at such a point x, we have $F'(x) = f(x)$. The Second Fundamental Theorem of Calculus, e.g., Apostol [3, Theorem 5.3, p. 205], tells us that if F is continuously differentiable, then F is absolutely continuous and its derivative $f(x) = F'(x)$ is its density. The nonnegativity of f follows from the that fact that F is nondecreasing. (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]

[1] For math hawks: For an example of a cumulative distribution function that is continuous, but not absolutely continuous, consider the **Cantor ternary function** c, described in Appendix 6.14⋆. 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

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

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

If X has cumulative distribution function F, then

$$P(X \in \boldsymbol{R}) = 1 = \int_{-\infty}^{\infty} f(x)\,dx.$$

Any nonnegative 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.

The following simple lemma shows a simple way to generate new densities from old.

This is redundant with Theorem 9.1.1 and Example 9.1.2.

6.2.2 Lemma *If f is a probability density function and $a > 0$, then g defined by*

$$g(x) = af(ax)$$

is also a probability density function.

Proof: Since $a > 0$ and $f \geqslant 0$, we have $g \geqslant 0$, so all that we need to show is that $\int_{-\infty}^{\infty} g(x)\,dx = 1$. But

$$\int_{-\infty}^{\infty} g(x)\,dx = \int_{-\infty}^{\infty} af(ax)\,dx = \int_{-\infty}^{\infty} f(y)\,dy = 1,$$

where the penultimate equality comes from the change of variable $y = ax$. ∎

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 more recently than I, 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 [12, Section 4.4, pp. 302–310] and Larsen–Marx [10, Section 3.8, pp. 176–183].

up an appendix) and it remains a density for the distribution. This means that I lied when I said that if f is an indefinite integral, and is nondecreasing, then f must be nonnegative—it could be negative on a set of measure zero, such as a single point. Nevertheless, I'll continue to require that densities be nonnegative. These different densities are called **versions** of each other. They all give rise to the same cumulative distribution function and the same support. The densities that we discuss here are all piecewise continuous, and we make the pieces as large as possible.

6.3 Probabilities and densities

As mentioned above, if X has density f, then

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

That is, the probability that X takes on a value in $[a,b]$ is just the are under the density between a and b. For instance, in Figure 6.1 the shaded shows the probability X takes on a value in the interval $[-0.5, 0.8725]$ when the density is $e^{-x^2/2}/\sqrt{2\pi}$. The area is approximately 0.500, while the area under the entire for the interval $[-3,3]$ shown is approximately 0.997. (It doesn't look like it to me. What do you think?)

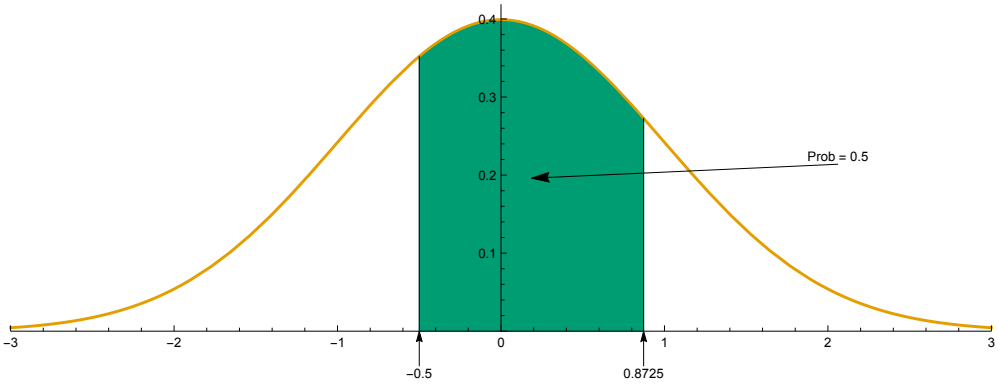


Figure 6.1. $P(-0.5 \leqslant X \leqslant 0.8725)$.

Another way to say the same thing is that if the density f is continuous at a point x, then it is the instantaneous rate of probability per unit length that the random variable X falls in an interval. That is, $P(X \in [x-\varepsilon/2, x+\varepsilon/2]) \approx \varepsilon f(x)$, which is the area of a rectangle of height $f(x)$ and base ε. See Figure 6.2. In this way, the density is analogous to it is analogous to the probability mass function for a discrete random variable.

6.3.1 Remark If X is a random variable with a density f and a is any real number, then

$$P(X = a) = \int_a^a f(x)\,dx = 0.$$

That is, the probability of any particular value must always be zero.

Yet the random experiment must have some outcome, so how can they all have probability zero? How can a line have positive length if it is made up of points each of which has length zero? This, I think, is one of those things that von Neumann was referring to when he said "In mathematics you don't understand things. You just get used to them." (See page 1–12.) If you think you understand this, I recommend David Foster Wallace's *Everything and More: A Compact History of* ∞ [14].

This is another reason probabilists invented the phrase "almost surely."

6.4 Expectation and densities

The usefulness of knowing the density of a random variable lies in the fact that we can use the density as a computational tool to calculate expectations. Proposition 5.10.3 has the following

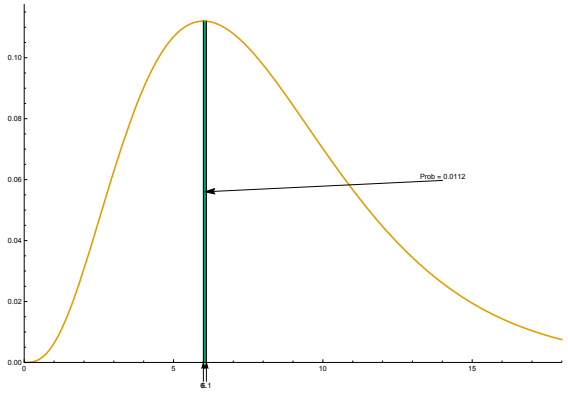


Figure 6.2. The density as an instantaneous rate of probability.

analog for random variables with a density. The proof is beyond the scope of this course.

6.4.1 Theorem *If X has a density f, we can calculate 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.

Note that some authors, as well as your textbooks, take the above theorem as a definition of the expectation, at least for random variables that have densities. For discrete random variables they define the expectation along the lines of Proposition 5.10.3. But there are random variables that are neither discrete nor have a density, so it would seem that more definitions are called for. It seems to me that our approach to defining expectations is both straightforward and consistent, and covers all the cases. I'm glad to say that there are others, for instance, Ross and Peköz [13, Section 1.6, pp. 23–25] who agree.

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

Finally, here is a "trick" for calculating the the expectation, when there is a density.

6.4.2 Proposition (Tail probabilities II) *Let X be a nonnegative random variable with a density f on the interval $[0, b]$, and cumulative distribution function F, so $f = F'$. Then*

$$\boldsymbol{E}\, X = \int_0^b 1 - F(x)\, dx.$$

Proof: The expectation of X is given by

$$\boldsymbol{E}\, X = \int_0^b x f(x)\, dx.$$

Integrating by parts gives

$$\int_0^b xf(x)\,dx = bF(b) - 0F(0) - \int_0^b F(x)\,dx.$$

But $bF(b) = b$, and $\int_0^b 1\,dx = b$, so

$$\boldsymbol{E}\,X = \int_0^b 1 - F(x)\,dx.$$

∎

There is a useful generalization of this result that is beyond the scope of this course. (It uses Fubini's Theorem.)

6.4.3 Proposition *Let X be a nonnegative random variable (not necessarily absolutely continuous or discrete). Then for any $p > 0$,*

$$\boldsymbol{E}\,X^p = \int_0^\infty px^{p-1}P(X > x)\,dx.$$

For a proof, see my on-line notes (Corollary 9) or Durrett [7, Lemma 2.2.13, pp. 54–55].

We can also use the density to calculate the expectation of a function of a random variable.

Larsen–Marx [10]: § 3.5

6.4.4 Theorem *For a random variable X with a density f, we have that $g \circ X$ is also a random variable,*[a] *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.

[a] 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.

6.4.5 Remark One way to think about this is as a kind of change-of-variables theorem. For instance, I'll bet you have transformed an integral of a function g from (x, y)-coordinates to (r, θ)-coordinates (polar coordinates). There is a multiplicative factor in the transformed integral, namely the Jacobian determinant of the transformation. In our case we start with a function $g \circ X$ on a set Ω and transform it from a function of ω to the function $g(x)$ of x defined on $\boldsymbol{R}$, and the density $f(x)$ is the multiplicative factor.

If you did not find this remark helpful, just pretend you never read it.

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

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

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

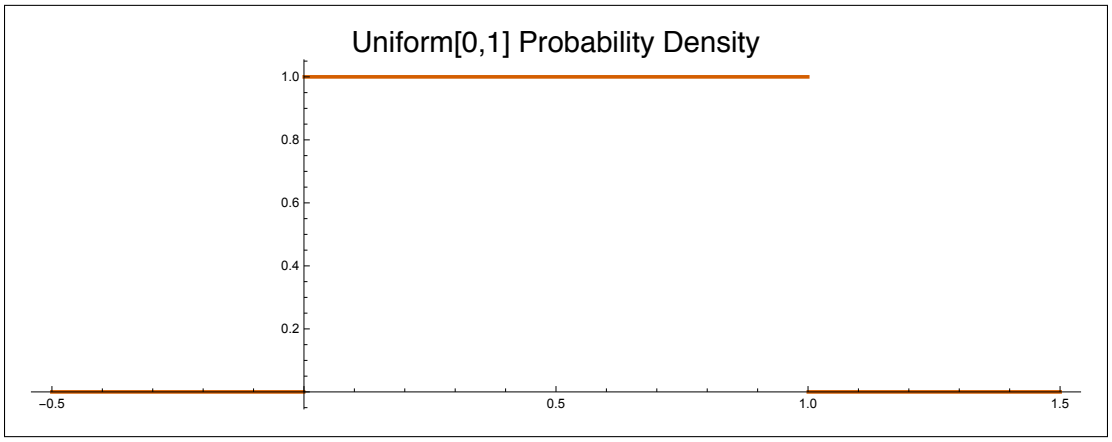


Figure 6.3. The Uniform$[0, 1]$ pdf.

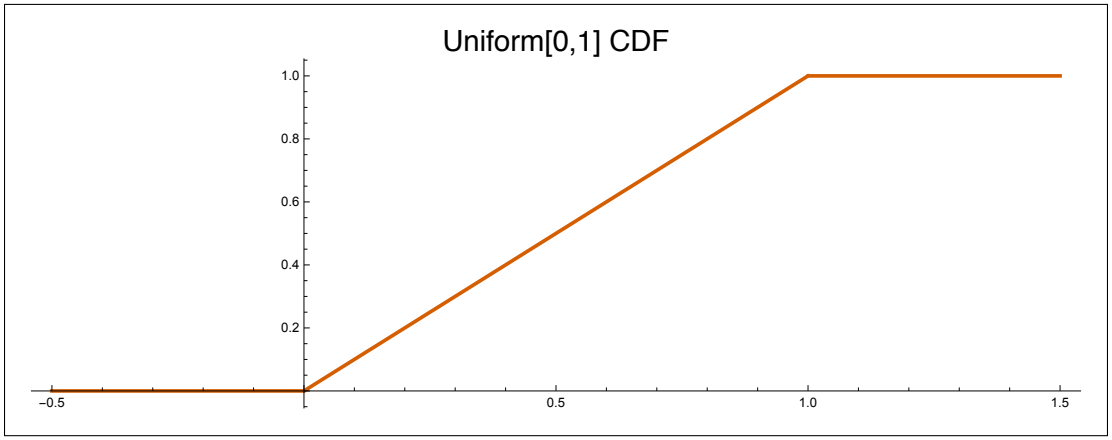


Figure 6.4. The Uniform$[0, 1]$ cdf.

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 The modes and median of a distribution

The **median** of the distribution of a random variable X with cumulative distribution function F is a number x such that

$$P(X \leqslant x) = 1/2 = P(X \geqslant x).$$

It may not be unique. For instance, if X is the number of success in three Bernoulli trials with probability of success $p = 1/2$, the both 1 and 2 (and any number in between are medians). In such a case, we usually pick the midpoint of the set of medians to be *the* median.

The **mode** of a distribution is a less formal term and it refers to a point of locally greatest probability. For distributions with densities, we look for local maximizers of the density. A distribution with a unique local maximizer is called **unimodal**. We moreover will say that a distribution is unimodal if there are several local maximizers of the density if they are all grouped together. For instance, the uniform distribution on $[0, 1]$ is usually considered to be unimodal. Same for the a Binomial distribution with parameter $p = 1/2$ and an odd number n of trials. Its probability mass function has maximizers at $(n-1)/2$ and $(n+1)/2$, but it is usually considered to be unimodal.

An example of a bimodal distribution is heights of adult Americans. Males and females both have unimodal distributions, but together there are two modes.

6.7 Expectation is a positive linear operator!!

Since random variables are just real-valued functions on a sample space Ω, 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

$$(X + Y)(\omega) = X(\omega) + Y(\omega).$$

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 [12]: pp. 181 ff.

- Expectation is a **linear operator** on $L_1(P)$, This means that

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

We have already proven this for simple random variables in Proposition 5.11.1. The proof for general random variables follows in a straight forward fashion. Until I get around to writing it out, see for instance, Ross and Peköz [13, Proposition 1.29, pp. 26–28].

- Expectation is a **positive operator**. That is, if $X \geqslant 0$, i.e., $X(\omega) \geqslant 0$ for each $\omega \in \Omega$, 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 $P(X = c) = 1$ (or $X = x$a.s.), 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:
$$\boldsymbol{E}\, \mathbf{1}_A = \sum_{\omega \in \Omega} \mathbf{1}_A(\omega) P(\omega) = \sum_{\omega \in A} P(\omega) = P(A).$$

∎

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

6.7.1 Proposition (Summary of positive linear operator properties)
In summary, for random variables with finite expectation (those in $L_1(P)$):

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

6.8 Expectation of an independent product

6.8.1 Theorem *Let X and Y be random variables on the probability space $(\Omega, \mathcal{F}, P)$, with finite expectations. If X and Y are* independent, *then*

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

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

$$\begin{aligned}
\boldsymbol{E}(XY) &= \sum_{(x,y)} xyP(X=x \text{ and } Y=y) && \text{Proposition 5.10.3}\\
&= \sum_{(x,y)} xyP(X=x)P(Y=y) && \text{by independence}\\
&= \sum_{x} \left(xp_X(x) \left(\sum_{y} yp_Y(y) \right) \right) && \text{Distributive Law}\\
&= \sum_{x} xp_X(x)(\boldsymbol{E}\,Y) && \text{definition of expectation for } Y\\
&= (\boldsymbol{E}\,Y) \left(\sum_{x} xp_X(x) \right) && \text{Distributive Law}\\
&= (\boldsymbol{E}\,Y)(\boldsymbol{E}\,X) && \text{Proposition 5.10.3 again.}
\end{aligned}$$

∎

6.9 ⋆ Jensen's Inequality

This section is not covered in Pitman [12]!

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 $0 < t < 1$.

*A function f is **(strictly) concave** if $-f$ is (strictly) convex.*

Another way to say this is that the line segment joining any two points on the graph of a convex function f lies above the graph. Note that a linear function is both convex and concave, but not strictly.

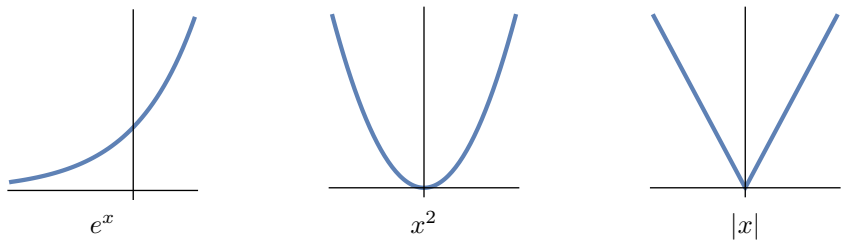


Figure 6.5. Some familiar convex functions. The exponential and the square function are strictly convex, but the absolute value function is not strictly convex.

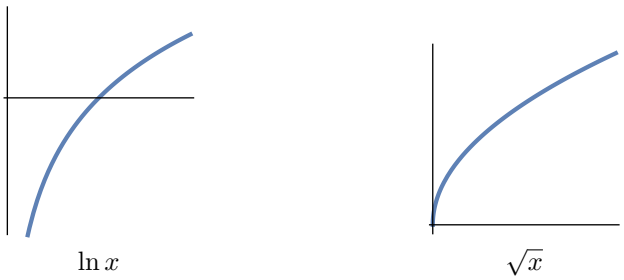


Figure 6.6. Some strictly concave functions.

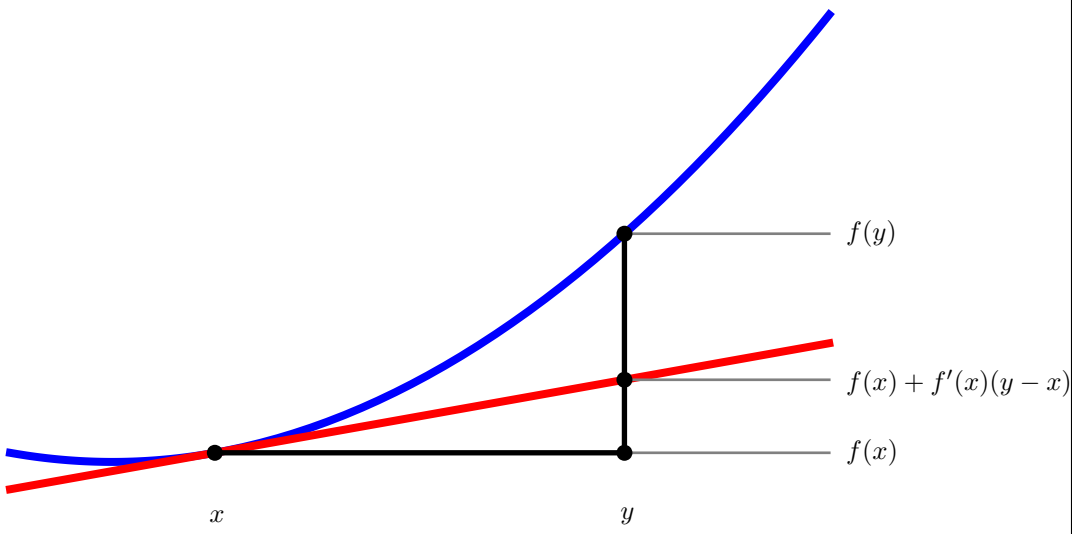


Figure 6.7. The Subgradient Inequality for a convex function.

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) and continuous 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 $x \in [a, b]$, 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 $x \in (a, b)$, then we have the* ***Subgradient Inequality****:*

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

For concave functions the inequality is reversed.

- *The geometric interpretation of this is that if f is convex, then its graph lies above the tangent line to the graph. See Figure 6.7.*

Even if f is not differentiable at $x \in (a, b)$, f will have both a left- and right-hand derivative at x and the left-hand derivative will be less than the right-hand derivative. Any number m between the left and right derivatives is called a ***subderivative****, and for all $y \in [a, b]$, we have*

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

If the left and right hand derivatives are equal at x, then f is actually differentiable at x and $f'(x)$ is the only subderivative.

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

- *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 $f\colon I \to \boldsymbol{R}$ be a convex function whose domain I is an interval that 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.

For concave functions the inequality is reversed.

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 [9], 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.

6.9.5 Example Let X be the result of rolling a single die. Then

$$\begin{aligned}
\boldsymbol{E}\,X &= \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = 3\frac{1}{2}\\
\boldsymbol{E}(X^2) &= \frac{1}{6}(1+4+9+16+25+36) = \frac{91}{6} = 15\frac{1}{6}\\
\boldsymbol{E}(1/X) &= \frac{1}{6}(1+\frac{1}{2}+\frac{2}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}) = \frac{49}{120} \approx 0.408333
\end{aligned}$$

It is straightforward to see that

$$\begin{aligned}
\boldsymbol{E}\,X^2 = 15\frac{1}{6} > 12\frac{1}{4} &= (\boldsymbol{E}\,X)^2\\
\boldsymbol{E}\left(\frac{1}{X}\right) \approx 0.408 > 0.0659341 &\approx \frac{1}{\boldsymbol{E}\,X}
\end{aligned}$$

Now Let $Y = X^2$. Then $Y/X = X$, so $\boldsymbol{E}(Y/X) = 3\frac{1}{2}$, but $(\boldsymbol{E}\,Y)/(\boldsymbol{E}\,X) = \frac{91}{6}/\frac{7}{2} = \frac{13}{3} = 4\frac{1}{3}$. □

6.10 Variance

Pitman [12]: § 3.3
Larsen–Marx [10]: § 3.6

6.10.1 Definition *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. (Some authors might also say that a random variable has infinite variance.)

The ***standard deviation*** *of X, denoted* $\operatorname{SD} X$*, 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 haven't 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.

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

Proof: To simplify things, let $\mu = \boldsymbol{E}\, X$. Then since **expectation is a positive 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. (True story.) 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 [12, 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. (Hint: Use Jensen's Inequality.)

One reason for the popularity of variance is that it is easier to work with than the mean absolute deviation. For instance, in a moment (no pun intended) we shall prove Theorem 6.11.1, which asserts that the variance of their 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. For instance, if X and Y are independent, and for simplicity's sake we'll assume that $\boldsymbol{E}\, X = \boldsymbol{E}\, Y = 0$, all I can say is that $\boldsymbol{E}\,|X + Y| \leqslant \boldsymbol{E}\,|X| + \boldsymbol{E}\,|Y|$.

The next result is referred to as the **Bienaymé Equality** by Loève [11, p. 246]. (See the 1853 paper by Jules Bienaymé [4].)

6.11.1 Theorem (Bienaymé Equality) *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,

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

so since **expectation is a positive linear operator**,

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

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

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

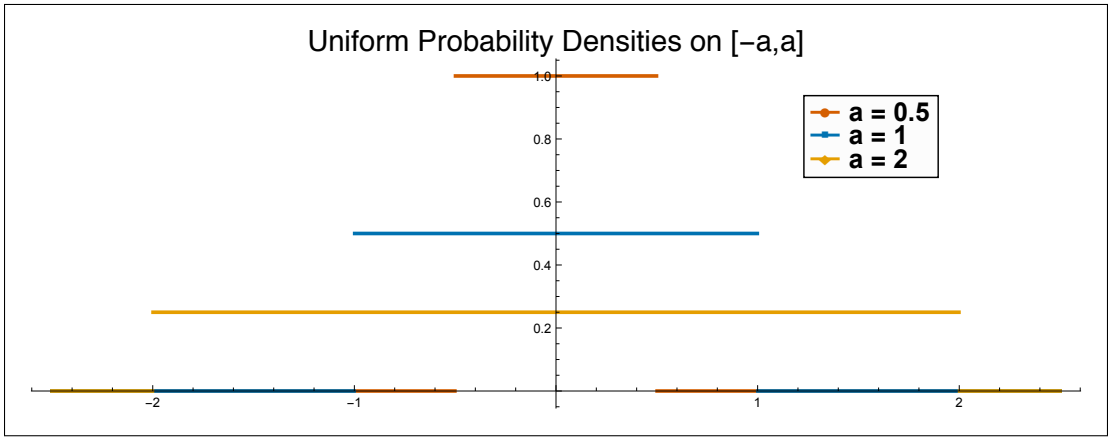
- For $a > 0$, the variance of a Uniform$[-a, a]$ random variable (which has density $1/2a$) on $[-a, a]$ and expectation 0) is

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

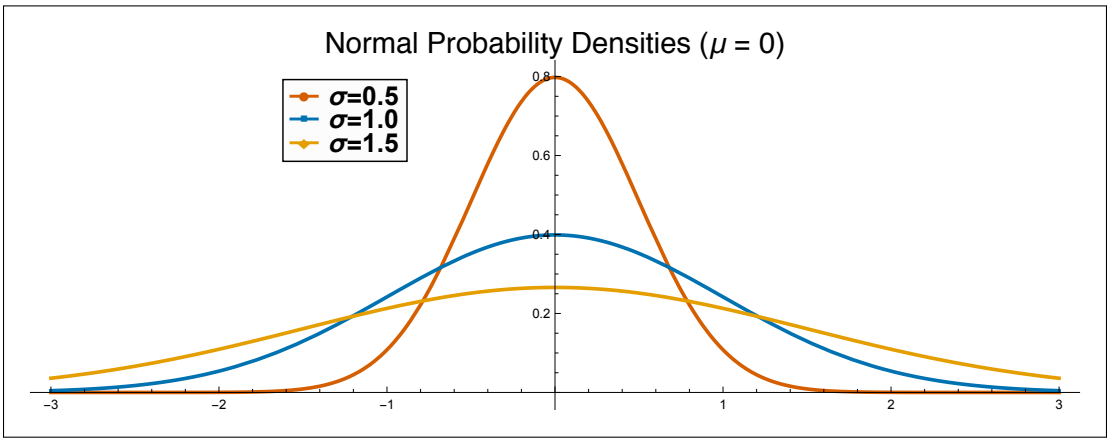
□

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 ⋆ Limits and expectations

Move to section on expectation!

Suppose we have a sequence of random variables $X_1, X_2, \dots$, which converge pointwise (on the probability space $(\Omega, \mathcal{F}, P)$) to a random variable X. When is it the case that

$$\lim_{n\to\infty} \boldsymbol{E}\, X_n = \boldsymbol{E}\, X?$$

The answer is clearly, not always.

6.12.1 Example Let $\Omega = \{1, 2, 3, \dots\}$ and define the probability P by

$$P(n) = 1/2^n.$$

Let the random variable X_n be defined by

$$X_n(\omega) = \begin{cases} 2^n & \omega = n, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\boldsymbol{E}\, X_n = 1$ for all n, but $X_n \to 0$ pointwise. In this case,

$$\lim_{n\to\infty} \boldsymbol{E}\, X_n = 1 \neq 0 = \boldsymbol{E}(\lim_n X_n).$$

□

There are two cases where we do have an affirmative answer, the Monotone Convergence Theorem 5.13.2 and the following.

6.12.2 Theorem (Dominated Convergence Theorem) *Let*

$$X_n \to X$$

pointwise on Ω*, and assume there is a random variable* Y *with*

$$\boldsymbol{E}\, Y < \infty,$$

such that Y *dominates each* X_n*, that is,*

$$|X_n| \leqslant |Y| \text{ for all } n.$$

Then

$$\lim_n \boldsymbol{E}\, X_n = \boldsymbol{E}\, X.$$

This is sometimes called the **Bounded Convergence Theorem**. It too is a standard result, and may be found in, e.g., Aliprantis and Border [1, Theorem 11.21, p. 415]), in Breiman [6, A.28, p. 398], Halmos [8, Theorem D, p. 110].

We will have use for these results when we study the long run properties of Markov chains and martingales in Lectures 16, 17, and Supplement 7.

6.13 The L_p spaces

For historical reasons,[4] we shall make the following definition.

6.13.1 Definition *Given a probability space $(\Omega, \mathcal{F}, P)$, for $p \geqslant 1 < \infty$ define*

$$L_p(P) = \{\text{random variables } X : \boldsymbol{E}\,|X|^p < \infty\}.$$

Thus $L_1(P)$, or more simply by L_1, is the set of random variables with finite expectation, and L_2 is the set of random variables with finite variance.

6.13.2 Definition *For $1 \leqslant p < \infty$, if $\boldsymbol{E}(|X|^p)$ is finite, then*

$$\|X\|_p = \Big(\boldsymbol{E}(|X|^p)\Big)^{\frac{1}{p}}$$

*is called the **L_p norm**, or simply the **p-norm** of X.*

6.13.3 Proposition *The set L_p is closed under addition and scalar multiplication, that is, it is a vector subspace of the space of random variables.*

The proof is taken from Aliprantis and Burkinshaw [2, p. 255].

Proof: Closure under scalar multiplication is straightforward.

To see that it is closed under addition, by observing that for any real x, y, we have $|x+y| \leqslant |x|+|y|$. Now note that $|x| = (|x|^p)^{1/p} \leqslant (|x|^p + |y|^p)^{1/p}$, where the inequality follows from $|x|^p \leqslant |x|^p + |y|^p$. Now interchange x and y and the two inequalities to get

$$|x+y| \leqslant |x|+|y| \leqslant 2\,(|x|^p + |y|^p)^{1/p}$$

which implies

$$|x+y|^p \leqslant 2^p\,(|x|^p + |y|^p)$$

Since **expectation is a positive linear operator**,

$$|X+Y|^p \leqslant 2^p\,(\boldsymbol{E}\,|X|^p + \boldsymbol{E}\,|Y|^p) < \infty,$$

so $L_p(P)$ is closed under addition. ∎

Recall the following definition.

6.13.4 Definition *A **norm** on a vector space is a function $\|x\|$ satisfying, for all vectors x and y, and all scalars a,*

1. $\|x\| \geqslant 0$ *and* $\|x\| = 0 \iff x = 0$.
2. $\|ax\| = |a|\ \|x\|$.
3. $\|x+y\| \leqslant \|x\| + \|y\|$.

[4] The L is for Henri Lebesgue (1875–1941) a French mathematician, who developed the approach to integration that is embodied in our notion of expectation.

The next result is well-known as **Minkowski's Inequality**, and may be found, for instance, in Aliprantis and Burkinshaw [2, Theorem 31.4, pp. 256–257].

6.13.5 Minkowski's Inequality *For $1 \leqslant 1 < \infty$, and for any $X, Y \in L_p(P)$,*

$$\|X + Y\|_p \leqslant \|X\|_p + \|Y\|_p .$$

6.13.6 Corollary *For $p \geqslant 1$, the space L_p is a vector space, and $\|\cdot\|_p$ is a norm on L_p.*

6.14⋆ Appendix: The Cantor ternary function

This section is completely optional. It is here only for the math hawks. It describes a cdf that is a continuous function, but has no density.

Given any number x with $0 \leqslant x \leqslant 1$ there is an infinite sequence $a_1, a_2, \ldots$, where each a_n belongs to $\{0, 1, 2\}$ such that $x = \sum_{n=1}^{\infty} \frac{a_n}{3^n}$. This sequence is called the **ternary representation** of x. If x is of the form $\frac{N}{3^m}$ (in lowest terms), then it has two ternary representations: $x = \sum_{n=1}^{\infty} \frac{a_n}{3^n}$, where $a_m > 0$ and $a_n = 0$ for $n > m$, and another representation of the form $x = \sum_{n=1}^{m-1} \frac{a_n}{3^n} + \frac{a_m - 1}{3^m} + \sum_{n=m+1}^{\infty} \frac{2}{3^n}$. But these are the only cases of a nonunique ternary representation, and there are only countably many such numbers. (See, e.g., Boyd [5, Theorem 1.23, p. 20].)

Given $x \in [0, 1]$, let $N(x)$ be the first n such that $a_n = 1$ in the ternary representation. If x has two ternary representations use the one that gives the larger value of $N(x)$. If x has a ternary representation with no $a_n = 1$, then $N(x) = \infty$. The **Cantor set** $\mathcal{C}$ consists of all numbers x in $[0, 1]$ for which $N(x) = \infty$. That is, those that have a ternary representation where no $a_n = 1$. That is, all numbers x of the form $x = \sum_{n=1}^{\infty} \frac{2b_n}{3^n}$, where each b_n belongs to $\{0, 1\}$. Each distinct sequence of 0s and 1s gives rise to a distinct element of $\mathcal{C}$. Indeed some authors identify the Cantor set with $\{0, 1\}^{\mathbb{N}}$ endowed with its product topology, since the mapping $(b_1, b_2, \ldots) \mapsto \sum_{n=1}^{\infty} \frac{2b_n}{3^n}$ is a homeomorphism. Also note that a sequence $(b_1, b_2, \ldots)$ of 0s and 1s also corresponds to a unique subset of $\mathbb{N}$, namely $\{n \in \mathbb{N} : b_n = 1\}$. Thus there are as many elements $\mathcal{C}$ as there are subset of $\mathbb{N}$, so the Cantor set is uncountable. (This follows from the Cantor diagonal procedure.) Yet the Cantor set includes no interval.

It is perhaps easier to visualize the complement of the Cantor set. Let

$$\mathcal{A}_n = \{x \in [0, 1] : N(x) = n\}.$$

The complement of the Cantor set is $\bigcup_{n=1}^{\infty} \mathcal{A}_n$. Define

$$\mathcal{C}_n = [0, 1] \setminus \bigcup_{k=1}^{n} \mathcal{A}_k,$$

so that $\mathcal{C} = \bigcap_{n=0}^{\infty} \mathcal{C}_n$. Now $\mathcal{A}_1$ consists of those x for which $a_1 = 1$ in its ternary expansion. This means that

$$\mathcal{A}_1 = \left(\tfrac{1}{3}, \tfrac{2}{3}\right) \qquad \text{and} \qquad \mathcal{C}_1 = \left[0, \tfrac{1}{3}\right] \cup \left[\tfrac{2}{3}, 1\right].$$

Note that $N(\frac{1}{3}) = \infty$ since $\frac{1}{3}$ can also be written as $\sum_{n=2}^{\infty} \frac{2}{3^n}$, so $a_1 = 0$, $a_n = 2$ for $n > 1$. Now $\mathcal{A}_2$ consists of those x for which $a_1 = 0$ or $a_1 = 2$ and $a_2 = 1$ in its ternary expansion. Thus

$$\mathcal{A}_2 = \left(\tfrac{1}{9}, \tfrac{2}{9}\right) \cup \left(\tfrac{7}{9}, \tfrac{8}{9}\right) \qquad \text{and} \qquad C_2 = \left[0, \tfrac{1}{9}\right] \cup \left[\tfrac{2}{9}, \tfrac{1}{3}\right] \cup \left[\tfrac{2}{3}, \tfrac{7}{9}\right] \cup \left[\tfrac{8}{9}, 1\right].$$

Each $\mathcal{C}_n$ is the union of 2^n closed intervals, each of length $\frac{1}{3^{n-1}}$, and $\mathcal{A}_{n+1}$ consists of the open middle third of each of the intervals in $\mathcal{C}_n$. The total length of the removed open segments is

$$\frac{1}{3} + 2 \cdot \frac{1}{9} + 4 \cdot \frac{1}{27} + \cdots = \sum_{n=0}^{\infty} \frac{2^n}{3^{n+1}} = \frac{1}{3} \sum_{n=0}^{\infty} \left(\frac{2}{3}\right)^n = \frac{1}{3} \cdot \frac{1}{1 - \frac{2}{3}} = 1.$$

Thus the total length of the Cantor set is $1 - 1 = 0$.

The Cantor ternary function f is defined as follows. On the open middle third $(\frac{1}{3}, \frac{2}{3})$ its value is $\frac{1}{2}$. On the open interval $(\frac{1}{9}, \frac{2}{9})$ its value is $\frac{1}{4}$ and on $(\frac{7}{9}, \frac{8}{9})$ its value is $\frac{3}{4}$. Continuing in this fashion, the function is defined on the complement of the Cantor set. The definition is extended to the entire interval by continuity. See Figure 6.8. A more precise but more opaque

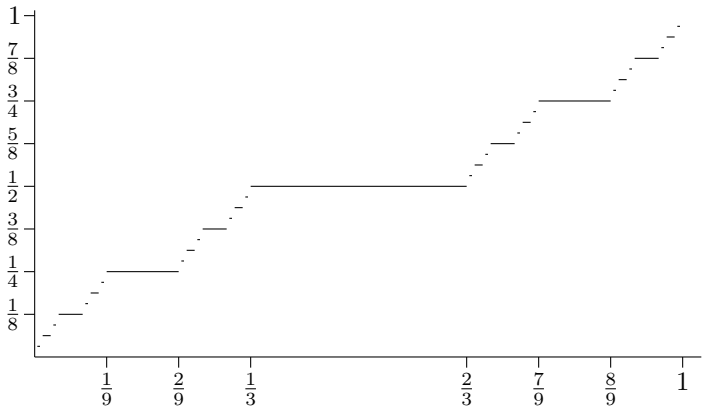


Figure 6.8. Partial graph of the Cantor ternary function.

definition is this:

$$f(x) = \begin{cases} \sum_{n=1}^{N(x)-1} \frac{\frac{1}{2}a_n}{2^n} + \frac{a_{N(x)}}{2^{N(x)}} & \text{if } N(x) < \infty, \\ \sum_{n=1}^{\infty} \frac{\frac{1}{2}a_n}{2^n} & \text{if } N(x) = \infty. \end{cases}$$

In any event notice that f is constant on each open interval in some $\mathcal{A}_n$, so it is differentiable there and $f' = 0$. Thus f is differentiable almost everywhere, and $f' = 0$ wherever it exists, but

$$f(1) - f(0) = 1 \quad \text{and} \quad \int_0^1 f'(x)\,dx = 0.$$

Bibliography

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

[2] C. D. Aliprantis and O. Burkinshaw. 1998. *Principles of real analysis*, 3d. ed. San Diego: Academic Press.

[3] T. M. Apostol. 1967. *Calculus, Volume I: One-variable calculus with an introduction to linear algebra*, 2d. ed. New York: John Wiley & Sons.

[4] J. Bienaymé. 1853. Considerations à l'appui de la découverte de laplace su la loi de probabilité dans la méthode des moindres carrés. *Comptes Rendus des Séances de l'Académie des Sciences (Paris)* 37(9):309–324.
https://gallica.bnf.fr/ark:/12148/bpt6k29948/f313.image

[5] D. Boyd. 1972. Classical analysis, volume 1. Notes prepared for Ma 108 abc. Published occasionally since at least 1972 by the Department of Mathematics, California Institute of Technology, 253-37, Pasadena CA 91125.

[6] L. Breiman. 1968. *Probability*. Reading, Massachusetts: Addison Wesley.

[7] R. Durret. 2019. *Probability: Theory and examples*, 5th. ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. DOI: 10.1017/9781108591034

[8] P. R. Halmos. 1974. *Measure theory*. Graduate Texts in Mathematics. New York: Springer–Verlag. Reprint of the edition published by Van Nostrand, 1950.

[9] 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

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

[11] M. Loève. 1977. *Probability theory*, 4th. ed. Number 1 in Graduate Texts in Mathematics. Berlin: Springer–Verlag.

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

[13] S. M. Ross and E. A. Peköz. 2007. *A second course in probability*. Boston: Probability-Bookstore.com.

[14] D. F. Wallace. 2003. *Everything and more: A compact history of* ∞. New York: Norton.