2. Working with Random Variables

2.1 Random Variables

Statistics deals with random data, not just random events. Like events, the variables are dynamic and depend on the experimental outcome.

DEFINITION 2.1 Let $\mathcal{S}$ be a sample space. A random variable (rv) X is a real-valued function defined on $\mathcal{S}$. If $s \in \mathcal{S}$ is the actual outcome then X takes value $X(s)$.

We will usually use upper case notation $\big($$X$ or Y or Z, etc.$\big)$ to denote *both the name and the value* of the random variable, suppressing the s.

On the other hand, we typically use lower case notation $\big($$x$ or y or z, etc.$\big)$ to represent single specific possible values. That is, $\{X = x\}$ is the event that the outcome s is such that $X(s)$ has value x.

**** The label X is the name of a function. As is always the case, different functions in the same context need different labels. It is also very important to distinguish random variables, such as X, from their possible values, such as x.*

Before we give examples, we first describe the indicator function of a set A (in any context). This is

$$1_A(x) = \begin{cases} 0 & \text{if } x \notin A, \\ 1 & \text{if } x \in A. \end{cases}$$

Some authors use $I_A(x)$ instead.

Example 2.1 Roll two distinguishable dice. $\mathcal{S}$ consists of 36 pairs (i, j). Examples of random variables include

$$X = X(i,j) = i = \text{ result of } 1^{st} \text{ die,}$$

$$Y = Y(i,j) = i + j = \text{ total of the dice,}$$

$$W = W(i,j) = 1_{\{(6,6)\}}(i,j) = \text{ indicator of the event ``two 6's''.}$$

Example 2.2 Flip a coin indefinitely. $\mathcal{S}$ consists of all infinite sequences of 0's and 1's, denoted $s = (s_1, s_2, \dots)$. Possible random variables include

$$X(s) = s_1 + \cdots + s_n = \text{ number of Heads in first } n \text{ flips,}$$

$$T(s) = \min\{n : s_n = 1\} = \text{ number of flips until a Head is first observed.}$$

Note that both are implicitly functions of the entire sequence, especially T.

Example 2.3 Observe lifetime of a single cell phone battery. Here, we can use $\mathcal{S} = [0, \infty)$ and

$$X(s) = s = \text{ the lifetime.}$$

If instead we observe the lifetimes of three batteries, we might label them X_1, X_2 and X_3. If that was all we observed (or cared about) then the sample outcome could simply be $s = (X_1, X_2, X_3)$. More generally, however, the sample outcomes would precisely include other information (explicitly observed or not), and the three rvs would be functions of those outcomes.

**** The sample space is frequently more complex than a few numerical observations. Indeed, often the sample space itself might simply be in the background and not fully identified.*

2.2 Distributions, pmfs and pdfs

Generally speaking, a distribution describes how the values of a random variable are distributed on the real line, and how likely values or sets of values are. A distribution can be expressed or formulated in several (equivalent) ways.

First, we will employ the following shorthand for an event about X: "$X \in C$" $= \{s : X(s) \in C\}$, where C is a subset of $\mathbb{R}$ (the real numbers).

DEFINITION 2.2 A random variable X has a discrete distribution (we say X *is discrete*) if there is a finite or countably infinite set C such that $\mathsf{P}(X \in C) = 1$.

In this case, X has a probability mass function (pmf) f_X defined by

$$f_X(x) = \mathsf{P}(X = x).$$

(The pmf is defined for all real x, but it is positive only for $x \in C$.)

In many practical cases, C will be a set of integers, but this is not at all necessary for the definition.

**** We subscript the pmf with the name of the random variable because often we will have more than one random variable at a time.*

Example 2.2 (cont.) Let T be the number of flips until the 1^{st} Head. Then T must be an integer, so it is discrete. If the coin is fair then its pmf is

$$f_T(t) = 2^{-t} 1_{\{1,2,\dots\}}(t).$$

Equivalently, we can express this

$$f_T(t) = 2^{-t}, \quad t = 1, 2, \dots,$$

where we interpret $f_T(t) = 0$ for values of t not explicitly indicated. Note, however, that simply saying $f_T(t) = 2^{-t}$ is *not sufficient*.

Example 2.4 The simplest random variable is the Bernoulli(p) rv. Suppose A is any event and let $X(s) = 1_A(s)$, $p = \mathsf{P}(A)$. Then

$$\mathsf{P}(X = 1) = \mathsf{P}(\{s : 1_A(s) = 1\}) = \mathsf{P}(A) = p$$

and $\mathsf{P}(X = 0) = 1 - p$. So the pmf for X is

$$f_X(x) = (1-p)1_{\{0\}}(x) + p1_{\{1\}}(x) = (1-p)^{1-x}p^x 1_{\{0,1\}}(x).$$

The last expression is called the *exponential representation* and is convenient for mathematical reasons.

Sometimes X is called a Bernoulli trial, the event "$X = 1$" $= A$ is called a success and "$X = 0$" $= A^c$ is called a failure.

Technically, there is an even simpler rv.

DEFINITION 2.3 A degenerate rv X is one such that $\mathsf{P}(X = x_0) = 1$ for some real x_0. Essentially, X is a constant on some event that has probability 1.

Example 2.5 Consider *independent* events $A_1, \ldots, A_n$, each with the *same* probability p. The corresponding indicators $1_{A_1}(s), \ldots, 1_{A_n}(s)$ are therefore a sample of n independent Bernoulli trials.

Let

$$Y = 1_{A_1}(s) + \cdots + 1_{A_n}(s),$$

which is the number of A_i's that occur (successes). We computed the probabilities for this in Ex. 1.17 and obtained (now in our new notation)

$$f_Y(y) = \mathsf{P}(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} 1_{\{0,\ldots,n\}}(y).$$

This is called the binomial(n, p) distribution and Y is a binomial rv.

Typically the events $A_1, \ldots, A_n$ described above will refer to the same response from a sample of independently selected individuals, such as the voters polled with replacement in Ex. 1.10 or the repeated flips of a coin in Ex. 2.2.

At this point it is helpful to mention an important subtlety.

It is easy to think of random variables as if they inherently have particular distributions.

But they don't. A random variable is defined as a function of the outcome – there is *no mention of probability* in that.

- For example, the number of successes in a set of n Bernoulli trials does not have to have binomial distribution if the assumptions are not met. And even if they are then *which* binomial distribution depends on the value of p.
- Similarly, the time until the first success in a sequence of Bernoulli trials can have many different distributions, depending on the assumptions (not the least of which is whether the trials are independent).
- A random variable's distribution will always depend on just what the probability model is.
- Even a random variable that is degenerate (i.e., essentially constant) for one probability model may not be degenerate for another.

Example 2.2 (cont.) Suppose we repeatedly and *independently* flip a coin with $p = \mathsf{P}(\text{Head on } n^{th} \text{ flip})$ (*same probability* for all flips). This is a sequence of Bernoulli trials with events $A_n =$ "Head on n^{th} flip". Let

$$T = \min\{n : 1_{A_n}(s) = 1\},$$

that is, the number of flips until the first Head. In Ex. 1.18 we computed

$$f_T(t) = \mathsf{P}(T = t) = p(1-p)^{t-1} 1_{\{1,2,\dots\}}(t).$$

This is called the geometric(p) distribution. Note the use of the indicator function to identify when the pmf is positive.

Distributions can be described in another way.

DEFINITION 2.4 Let X be a rv. The cumulative distribution function (cdf) for X is

$$F_X(x) = \mathsf{P}(X \le x), \quad \text{for any real } x.$$

We often write $X \sim F$ to mean F is the cdf for X, or $X \sim$ (dist. name) if the distribution has a particular name.

**** A cdf is defined all real values, even if the range of the random variable is only a subset of $\mathbb{R}$.*

Example 2.4 (cont.) $X \sim \text{Bernoulli}(p)$. Then

$$x < 0 \implies F_X(x) = \mathsf{P}(X \le x) = 0,$$
$$0 \le x < 1 \implies F_X(x) = \mathsf{P}(X = 0) = 1 - p,$$
$$1 \le x \implies F_X(x) = \mathsf{P}(X = 0) + \mathsf{P}(X = 1) = 1.$$

This can also be expressed

$$F_X(x) = (1 - p)1_{[0,\infty)}(x) + p1_{[1,\infty)}(x). \qquad \text{(exercise)}$$

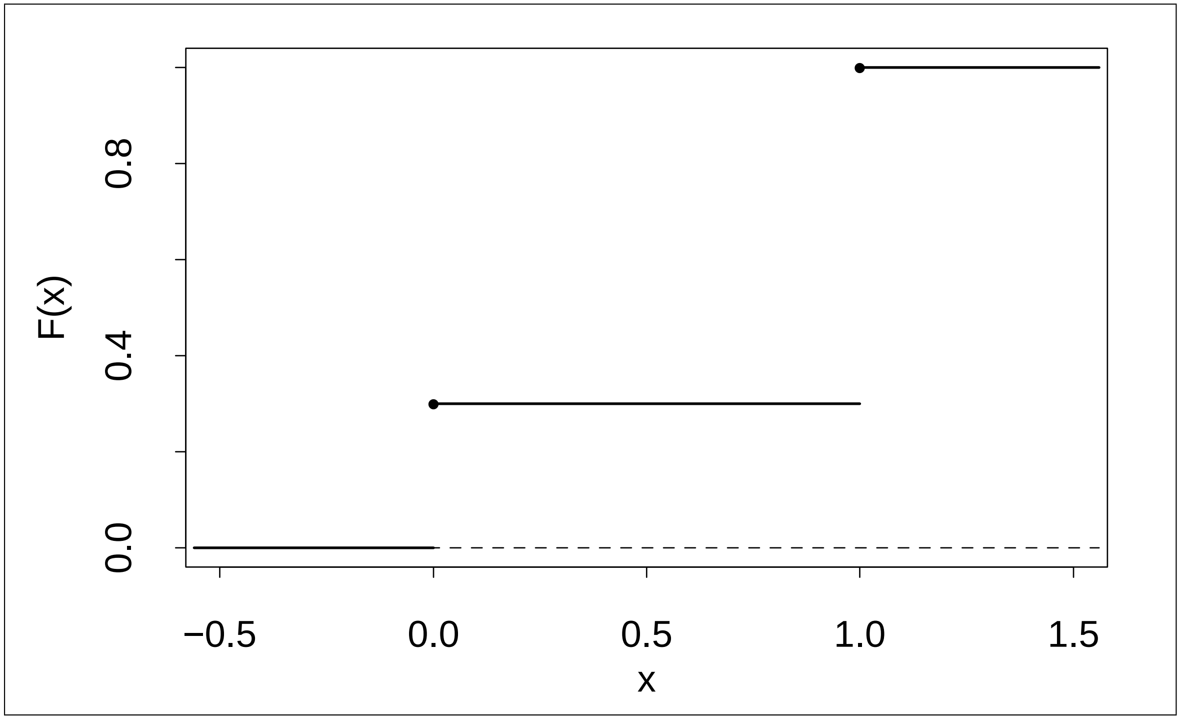

Figure 2.1 Bernoulli(.7) cumulative distribution function.

Example 2.2 (cont.) $T \sim$ geometric(p). We use another approach. If t is a positive integer, the event "$T > t$" means "no Heads in the first t flips". Hence $\mathsf{P}(T > t) = (1-p)^t$ and therefore $\mathsf{P}(T \le t) = 1 - (1-p)^t$.

For arbitrary nonnegative real t we see that $\mathsf{P}(T \le t) = \mathsf{P}(T \le \lfloor t \rfloor)$, so the general formula is

$$F_T(t) = (1 - (1-p)^{\lfloor t \rfloor})1_{[0,\infty)}(t).$$

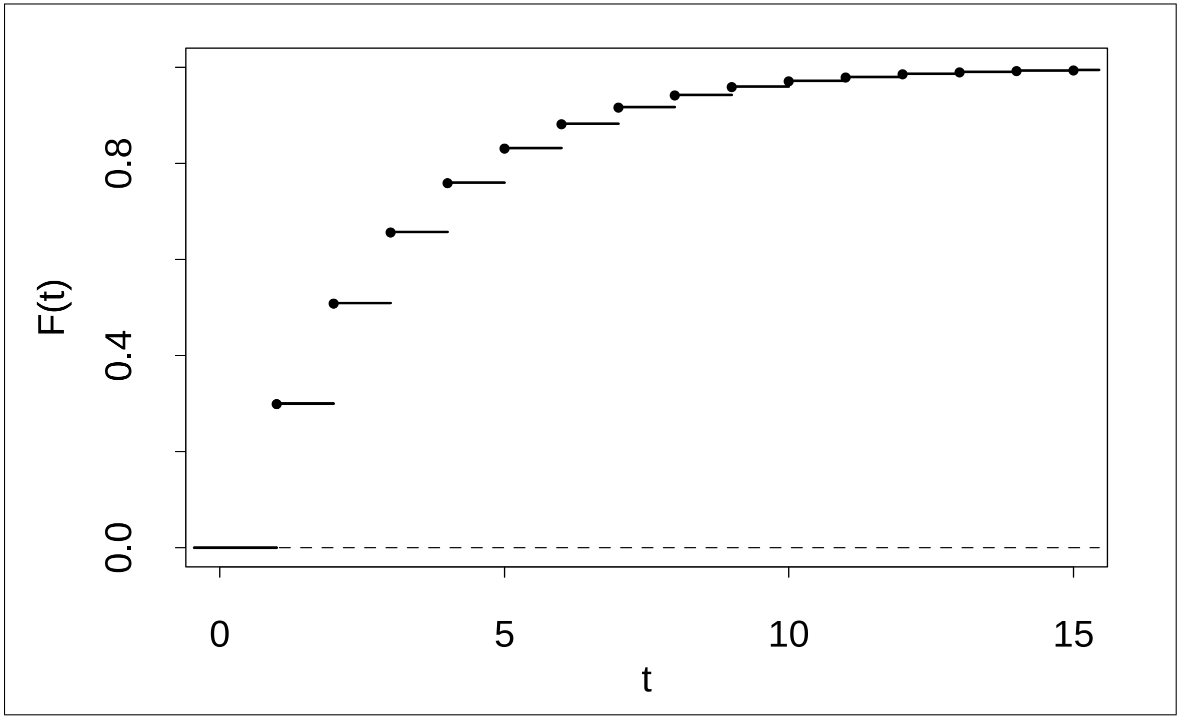

Figure 2.2 Geometric(.7) cumulative distribution function.

In both our examples, the cdf was a *step function*. This is no coincidence.

THEOREM 2.5 X is a discrete rv if and only if its cdf is a step function. In this case, if f_X is its pmf, positive at $x_1, x_2, \ldots$ then

$$F_X(x) = \sum_{j=1}^{\infty} f_X(x_j) 1_{(-\infty,x]}(x_j) = \sum_{j=1}^{\infty} f_X(x_j) 1_{[x_j,\infty)}(x).$$

PROOF ($\implies$) If X is discrete then $\mathsf{P}(X \le x)$ increases only at a countable number of points $x_1, x_2, \ldots$ and the increase is always a jump of size $\mathsf{P}(X \le x_j) - \mathsf{P}(X < x_j)$. Also, X has the cdf shown above (exercise).

($\impliedby$) Suppose F_X is a step function. Since $0 \le F_X(x) \le 1$ for all x, there at most n jumps of size $1/n$ or greater. It follows that there can only be a countable number of jumps altogether. Furthermore, $\mathsf{P}(a < X \le b) = F_X(b) - F_X(a)$ is the sum of the jumps of the points in $(a, b]$, by definition of a step function.

Letting $a \to -\infty$ and $b \to \infty$, it must also be true that

$$1 = \mathsf{P}(-\infty < X < \infty) = \sum_{j=1}^{\infty} f_X(x_j).$$

This says X is discrete by Def. 2.2. □

Contained in the previous proof, we mentioned the following. If $a < b$ then

$$\mathsf{P}(a < X \leq b) = \mathsf{P}(X \leq b) - \mathsf{P}(X \leq a) = F_X(b) - F_X(a).$$

Note the strict and non-strict inequalities. This holds because

$$\{X \leq a\} \cup \{a < X \leq b\} = \{X \leq b\}.$$

It is true for all random variables, not just discrete random variables, and it is very useful. (So *memorize* it!)

THEOREM 2.6 $F(x)$ is a cdf for some rv X if and only if

i. $x \leq y \implies F(x) \leq F(y)$ (nondecreasing),

ii. $F(y) \to F(x)$ as $y \downarrow x$ (right-continuous),

iii. $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$ (total measure 1).

PROOF ($\implies$)

i. $x \leq y$ implies "$X \leq x$" $\subset$ "$X \leq y$" and hence $\mathsf{P}(X \leq x) \leq \mathsf{P}(X \leq y)$.

ii. Let $A_n =$ "$X \leq x + 1/n$". This is a decreasing sequence of events. By continuity (Thm. 1.12), $\mathsf{P}(A_n) \to \mathsf{P}(\bigcap_n A_n) = \mathsf{P}(X \leq x)$.

(Note, however, $\mathsf{P}(X \leq x - 1/n) \to \mathsf{P}(X < x)$ which is not necessarily equal to $\mathsf{P}(X \leq x)$.)

iii. Similar arguments with decreasing events $A_n =$ "$X \leq -n$" and increasing events $A_n =$ "$X \leq n$".

($\Longleftarrow$) This requires much deeper analysis. (However, we will see later how we can put this point into practice.) □

- Let $F(x^-) = \lim_{y \uparrow x} F(y) = \mathsf{P}(X < x)$, which is the *left-hand* limit at x. By the note in the proof above, we can see that

$$F(x) - F(x^-) = \mathsf{P}(X \leq x) - \mathsf{P}(X > x) = \mathsf{P}(X = x)$$

which is the jump at x, if any.

- As we will see in examples later, for a rv X with a *continuous* cdf, $F(x) = F(x^-)$ and thus $\mathsf{P}(X = x) = 0$ for all x.

Example 2.6 Let $\mathcal{S} = [0,1]$ and define P by

$$\mathsf{P}((x,y]) = x - y \qquad \text{(length of the interval).}$$

It is easy to see why this should be additive (think of the total length of a union of disjoint intervals) and that it satisfies the other conditions of a probability measure. Now let $U(s) = s$ for $s \in \mathcal{S}$. Then

$$F_U(u) = \mathsf{P}(U \le u) = \mathsf{P}([0,u]) = u \quad \text{if } 0 \le u \le 1.$$

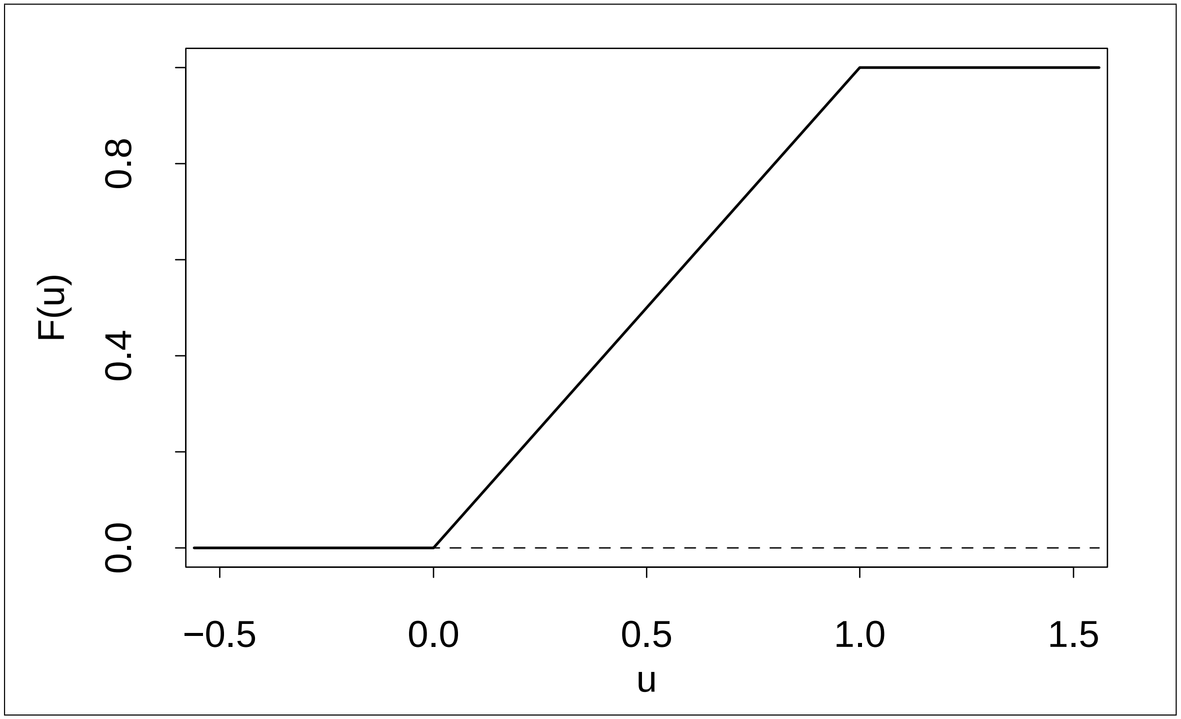

Figure 2.3 Uniform(0,1) cumulative distribution function.

Note that F_U is continuous and is not a step function.

Definition 2.7

i. X is a continuous rv if its cdf is a continuous function.

ii. X is absolutely continuous if there exists a probability density function (pdf) $f_X(x)$ such that

$$F_X(y) - F_X(x) = \int_x^y f_X(t)\,\mathrm{d}t \quad \text{for all } x < y.$$

It follows that $F_X(x) = \int_{-\infty}^x f_X(t)\,\mathrm{d}t$, and $f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_X(x)$ for those values x at which the derivative exists.

There are several important points here.

- If X is a continuous rv then $\mathsf{P}(X = x) = 0$ for every real x, unlike the discrete case.
- The derivative of an absolutely continuous cdf need not exist at every x.
- Any pdf can be changed at countably many x without changing the cdf. Of course, we generally want to use the simplest representation.
- There are continuous cdfs which are not absolutely continuous, although we generally avoid those in this course.

Example 2.6 (cont.) If $0 \le x < y \le 1$ we clearly have

$$F_U(y) - F_U(x) = \int_x^y 1du.$$

In fact,

$$F_U(y) - F_U(x) = \int_x^y 1_{[0,1]}(u)du \quad \text{for all } x < y.$$

So U is absolutely continuous with pdf $f_U(u) = 1_{[0,1]}(u)$.

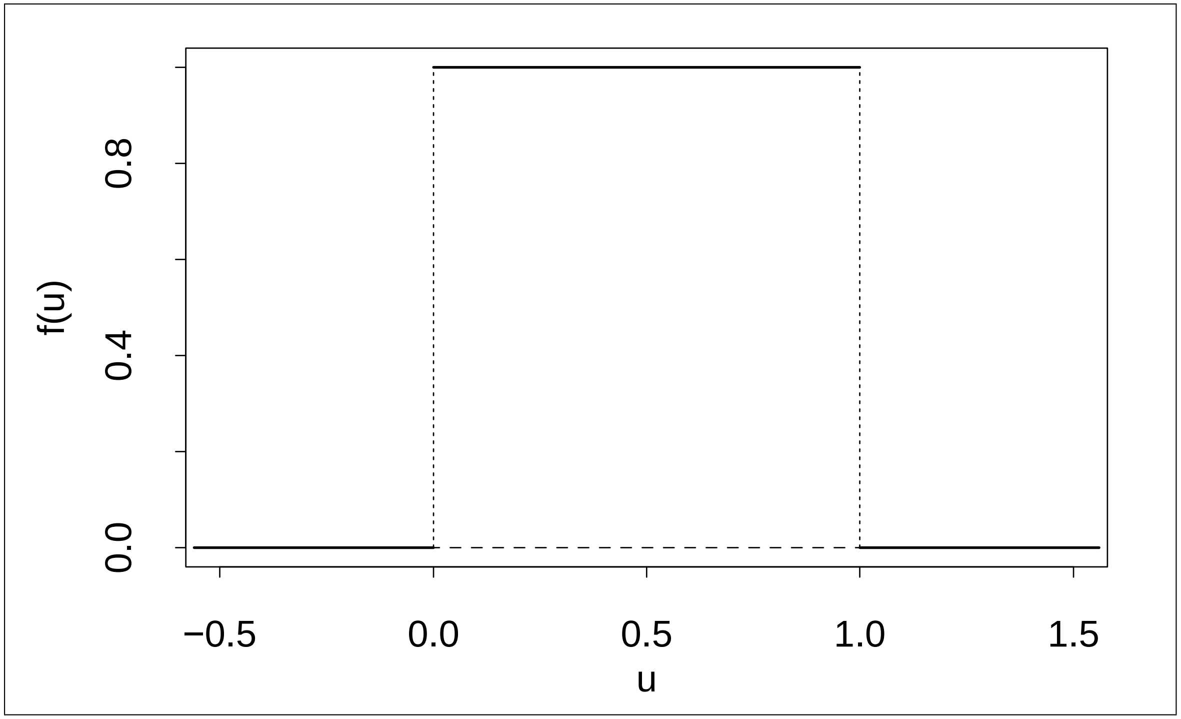

Figure 2.4 Uniform(0,1) probability density function.

This is called the uniform(0,1) distribution because the density is constant on the interval $[0, 1]$.

More generally, the uniform(a, b) distribution has constant density on the interval $[a, b]$ (and 0 elsewhere). What is the constant value of the pdf on this interval? Hint: the *cdf* ranges from 0 to 1.

**** We are using the same notation f_X for both the pmf for a discrete rv and the pdf of a continuous rv. Consequently, it is crucial to be aware which type of random variable is being considered.*

The set $\{x : f(x) > 0\}$ is called the support of f.

- The support of the uniform(a, b) distribution is the interval $[a, b]$.
- The support of a binomial(n, p) rv is $\{0, 1, \ldots, n\}$.

**** Identifying the support of a pmf or pdf is a necessary part of its definition.*

If C is the support of f and $f^*(x)$ is the "formula" for $x \in C$ then either we can write $\underline{f(x) = f^*(x)1_C(x)}$ or we can say $\underline{f(x) = f^*(x) \text{ for } x \in C}$ (with the understanding that $f(x) = 0$ for $x \notin C$).

Note: the *range* of a random variable is the set of its possible values, which technically is fixed regardless of the probability model.

Nevertheless, we will tend to conflate the terms "range" and "support".

Example 2.3 (cont.) X = lifetime of a cell phone battery. Recall the probability model suggested in Ex. 1.4: $\mathsf{P}(X > t) = \mathrm{e}^{-t/10}$ for $t > 0$. Since $\mathsf{P}(X \le t) = 1 - \mathsf{P}(X > t)$, we have

$$F_X(t) = \begin{cases} 0 & t < 0 \\ 1 - \mathrm{e}^{-t/10} & t \ge 0, \end{cases}$$

and

$$f_X(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{10}\mathrm{e}^{-t/10} & t \ge 0. \end{cases}$$

Note that F_X is indeed continuous, even at 0, but f_X is not continuous at 0. This is called the exponential(β) distribution, where $\beta = 10$ in this example.

We can either use the cdf to find, for example,

$$\mathsf{P}(10 \le X \le 20) = F_X(20) - F_X(10) = (1 - \mathrm{e}^{-20/10}) - (1 - \mathrm{e}^{-10/10}),$$

or we can use the pdf,

$$\mathsf{P}(10 \le X \le 20) = \int_{10}^{20} \frac{1}{10}\mathrm{e}^{-t/10}\,\mathrm{d}t,$$

whichever is more convenient.

**** Of course, if you already have the cdf there is no need to take a derivative and then integrate again.*

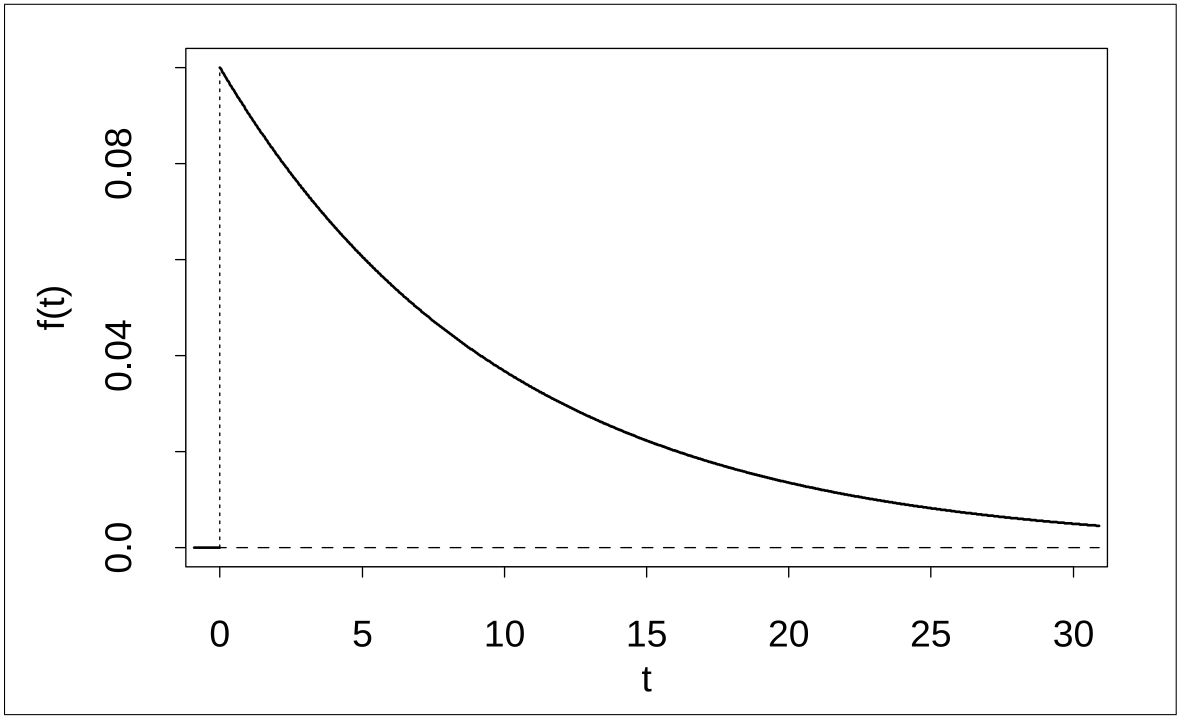

Figure 2.5 Exponential(10) probability density function.

Note that the distribution can be “parameterized” differently, for example by replacing β with $\frac{1}{\lambda}$ so that $F(t) = 1 - \mathrm{e}^{-\lambda t}$ and $f(t) = \lambda \mathrm{e}^{-\lambda t}$, for $t > 0$. (Some authors and texts use this representation.)

THEOREM 2.8

i. A function $f(x)$ is a pmf if and only if $f(x) \geq 0$ for all real x and $\sum_x f(x) = 1$. (This requires $f(x) > 0$ for only countably many x.)

ii. A function $f(x)$ is a pdf if and only if $f(x) \geq 0$ for all real x and $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

PROOF These follow from the relationships we have discussed between a cdf and its pmf or pdf. □

If $f(x) = Ch(x)$ for some positive constant C and “kernel” $h(x)$ then the distribution is determined by $h(x)$, because only one value is possible for C. Thus, we often can identify distributions *without determining the value of C*.

Example 2.7 Consider the function $h(x) = \frac{x}{(1+x^2)^2} 1_{[0,\infty)}(x)$.

Since $h(x) \geq 0$ and $\int_0^\infty h(x)\,\mathrm{d}x = 1/2$, we can conclude that $f(x) = 2h(x) = \frac{2x}{(1+x^2)^2}$, for $x > 0$, is a pdf.

Why is it obvious that $\frac{x}{(1+x^2)^2}$, for all real x, *cannot* be a pdf?

On the other hand, $\frac{|x|}{(1+x^2)^2}$, for all real x, is a pdf. (exercise.)

There are distributions which are mixtures of continuous and discrete.

Example 2.3 (cont.) Suppose the cell phone battery actually has a chance $.05$ of failing immediately, but given that it does not fail immediately it has an exponential(10) lifetime.

To confuse matters even more, we actually only observe the time until failure or 30 hours, whichever comes first (called censored data).

Let T be the observed value. Then $F_T(0) = \mathsf{P}(T \le 0) = \mathsf{P}(T = 0) = .05$ and, if $0 < t < 30$,

$$F_T(t) = \mathsf{P}(T \le t) = \mathsf{P}(T = 0) + (1 - \mathsf{P}(T = 0))\mathsf{P}(0 < T \le t | T \ne 0)$$

$$= .05 + .95(1 - \mathrm{e}^{-t/10}).$$

But $F_T(t) = 1$ for $t \ge 30$ (because T cannot be larger than 30) and $F_T(t) = 0$ for $t < 0$ (because T cannot be less than 0).

We can get $\mathsf{P}(T = 30)$ by considering the jump at 30:

$$\mathsf{P}(T = 30) = \mathsf{P}(T \le 30) - \mathsf{P}(T < 30)$$

$$= 1 - (.05 + .95(1 - \mathrm{e}^{-30/10})) = .95\mathrm{e}^{-3} \doteq .04730.$$

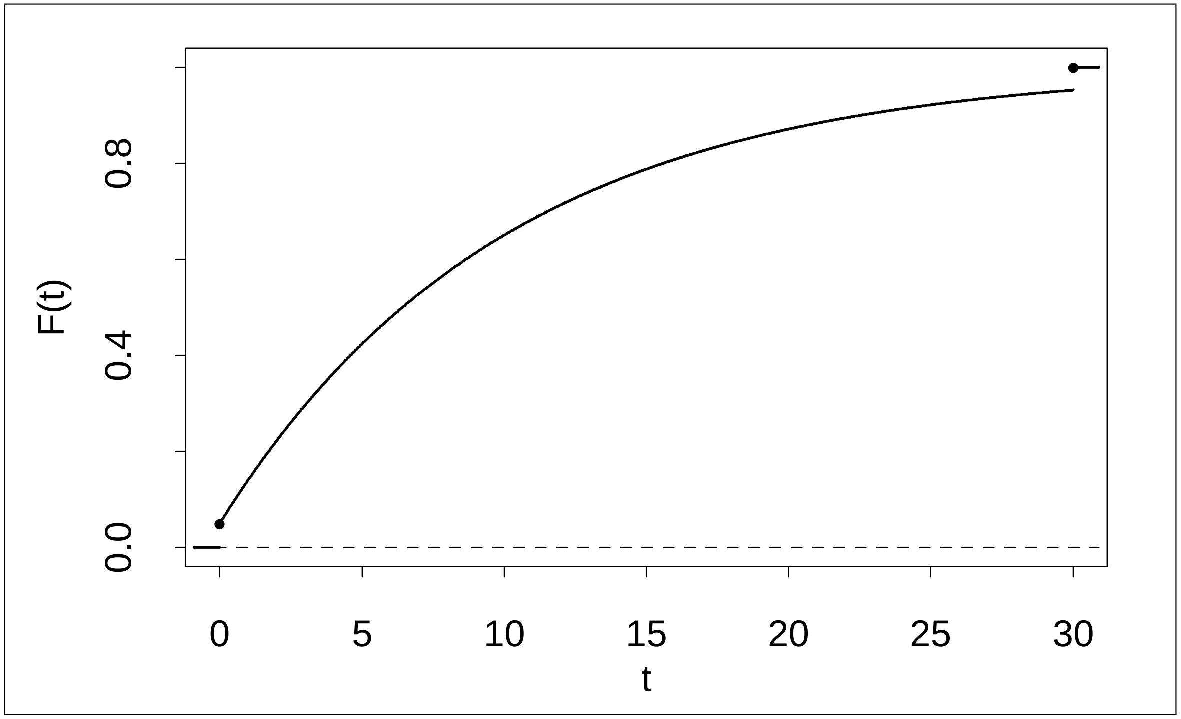

Figure 2.6 Distribution of a censored lifetime, with positive chance of immediate failure.

F_T has a derivative at all except 2 points, but the derivative does not integrate to 1. It has jumps, but they do not sum to 1.

But the sum of the jumps plus the integral is 1, and we can compute any probability by adding the relevant jumps and integral.

As we mentioned earlier, there are continuous cdfs which do not have a pdf. Here is the classic example of that.

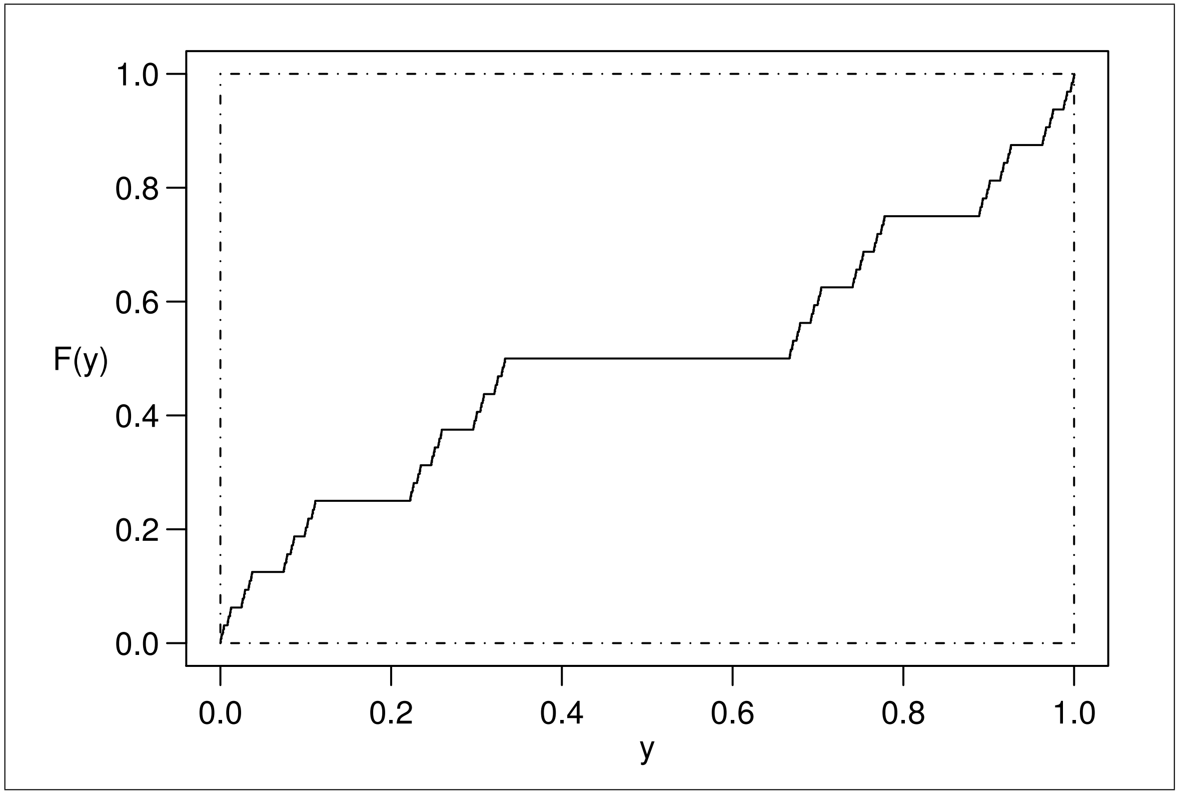

Figure 2.7 The Cantor distribution.

The derivative is 0 where it exists, but there are no jumps. All of the increase is concentrated on an *uncountable set* that has *total length equal to* 0.

2.3 Transformations

We are often interested in various transformations of a random variable and not just the variable itself. While not always necessary, it is frequently convenient to identify the distribution of the resulting new variable.

Example 2.8 Consider rolling two fair dice until a total of 7 appears, the payoff being $Y = 2^X$, where X = number of rolls.

What is the distribution of Y?

Solution Y is integer-valued, hence discrete. Actually, only certain integers (powers of 2) are possible, and $X = \log_2 Y$. The pmf of Y is

$$f_Y(y) = \mathsf{P}(Y = y) = \mathsf{P}(X = \log_2 y) = f_X(\log_2 y)$$

$$= \begin{cases} \frac{1}{6}\left(\frac{5}{6}\right)^{\log_2 y - 1} & \text{if } \log_2 y \text{ is a positive integer,} \\ 0 & \text{otherwise.} \end{cases}$$

We can also obtain the cdf:

$$F_Y(y) = 1 - \mathsf{P}(X > \log_2 y) = 1 - \left(\frac{5}{6}\right)^{\lfloor \log_2 y \rfloor} \quad \text{for } y > 0.$$

Example 2.9 Suppose $U \sim \text{uniform}(0,1)$ and $V = |U - .5|$, which is the distance from U to the middle of possibilities. We may compute, for $0 \le v \le .5$,

$$F_V(v) = \mathsf{P}(V \le v) = \mathsf{P}(.5 - v \le U \le .5 + v) = F_U(.5 + v) - F_U(.5 - v) = 2v$$

and $f_V(v) = \frac{\mathrm{d}}{\mathrm{d}v}F_V(v) = 2 \cdot 1_{[0,.5]}(v)$ (constant on $[0,.5]$). So $V \sim \text{uniform}(0,.5)$.

DEFINITION 2.9 Let $h(x)$ be a real-valued function of a real-valued argument. For any set $A \subset \mathbb{R}$, the inverse image of A by h is

$$h^{-1}A = \{x : h(x) \in A\}.$$

In particular, for $y \in \mathbb{R}$,

$$h^{-1}(-\infty, y] = \{x : h(x) \le y\}.$$

**** The notation here is for a set inverse, not the inverse function of a 1-1 function (but see below).*

Example 2.9 (cont.) Let $h(u) = |u - .5|$ for $u \in [0,1]$. Then, for $v \in [0,.5]$,

$$h^{-1}(-\infty, v] = \{u : h(u) \le v\} = \{u : |u - .5| \le v\} = [.5 - v, .5 + v].$$

Note from above that $\mathsf{P}(V \le v) = \mathsf{P}(U \in [.5 - v, .5 + v])$.

THEOREM 2.10 Let X be a random variable and let $Y = h(X)$, where $h(x)$ is real-valued. Then the cdf for Y is

$$F_Y(y) = \mathsf{P}(X \in h^{-1}(-\infty, y]).$$

PROOF $F_Y(y) = \mathsf{P}(Y \le y) = \mathsf{P}(h(X) \le y) = \mathsf{P}(X \in h^{-1}(-\infty, y])$. □

This is less mysterious than it looks. The point is simply that the distribution of Y can be computed from that of X. Just *how it is computed* depends on the nature of h and on the distribution of X, but basically one can follow the steps of the proof above to determine the cdf of the new rv.

In fact, you can avoid using set inverse by just saying $\mathsf{P}(h(X) \le y)$.

**** The first step, however, is always to ask "what is the support of Y?". That is, what is $h^{-1}R_X$ where R_X is the support of X?*

COROLLARY 2.11 Let X and Y be as in Thm. 2.10 and assume Y is discrete.

i. The pmf for Y is $f_Y(y) = \mathsf{P}(h(X) = y) = \mathsf{P}(X \in h^{-1}\{y\})$.

ii. If X is also discrete with pmf f_X then $f_Y(y) = \sum_{x \in h^{-1}\{y\}} f_X(x)$.

iii. If h is 1-1 with inverse function h^{-1} then X must also be discrete and $f_Y(y) = f_X(h^{-1}(y))$.

PROOF Same argument as before:

$$\mathsf{P}(Y = y) = \mathsf{P}(h(X) \in \{y\}) = \mathsf{P}(X \in h^{-1}\{y\}).$$

If h is 1-1 then $h^{-1}\{y\}$ is a set with just one value which typically would be denoted by $h^{-1}(y)$. □

Example 2.8 (cont.) . Recall from above that $Y = 2^X$ and

$$f_Y(y) = f_X(\log_2 y).$$

Note that strictly increasing continuous functions are 1-1, as are strictly decreasing continuous functions.

Example 2.5 (cont.) $Y =$ number of successes in n independent Bernoulli(p) trials. Recall $Y \sim$ binomial(n, p).

Now let $W = n - Y$ be the number of failures. W can take the same values as Y: $\{0, 1, \ldots, n\}$.

Then (with $y = n - w$)

$$f_W(w) = f_Y(n-w) = \binom{n}{n-w} p^{n-w}(1-p)^{n-(n-w)} 1_{\{0,\ldots,n\}}(n-w)$$

$$= \binom{n}{w}(1-p)^w p^{n-w} 1_{\{0,\ldots,n\}}(w).$$

So $W \sim$ binomial$(n, 1-p)$.

The first part of Cor. 2.11 does not require X to be discrete.

Example 2.10 Suppose X is a randomly selected insurance claim with pdf $f_X(x) = \frac{2x}{(1+x^2)^2}$. (See Ex. 2.7.)

Due to reinsurance, the cost to the company is

$$Y = 1_{[1,2)}(X) + 2 \cdot 1_{[2,10)}(X) + 10 \cdot 1_{[10,\infty)}(X),$$

which takes just four values: $0, 1, 2, 10$. So Y is discrete with pmf

$$f_Y(0) = \mathsf{P}(X < 1),\ f_Y(1) = \mathsf{P}(1 \leq X \leq 2),$$

$$f_Y(2) = \mathsf{P}(2 \leq X \leq 10),\ f_Y(10) = \mathsf{P}(10 \leq X).$$

We can use the cdf for X, $F_X(x) = \frac{x^2}{1+x^2}$ for $x > 0$, to compute these probabilities. For example,

$$f_Y(2) = F_X(10) - F_X(2) = \frac{100}{101} - \frac{4}{5}.$$

**** Again, keep in mind that we are using similar notation for pmfs and pdfs. In the example above, $f_X(x)$ is the pdf for X while $f_Y(y)$ is the pmf for Y.*

If Y is continuous, however, things look somewhat different.

COROLLARY 2.12 Let X and Y be as in Thm. 2.10.

i. If Y is known to be absolutely continuous then the pdf for Y is $f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathsf{P}(h(X) \le y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathsf{P}(X \in h^{-1}(-\infty, y])$.

ii. If h is *1-1* with inverse h^{-1} and h is *differentiable* then Y is absolutely continuous if and only if X is absolutely continuous, and

$$f_Y(y) = \frac{1}{|h'(h^{-1}(y))|} f_X(h^{-1}(y))).$$

**** A simple way to remember Cor. 2.12.ii is to note that when changing variables in the integral we have $f_X(x)\mathrm{d}x = f_Y(y)\mathrm{d}y$. So this gives the mnemonic*

$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|,$$

where we interpret both x and $\frac{\mathrm{d}x}{\mathrm{d}y}$ as functions of y.

Note also that $\frac{\mathrm{d}x}{\mathrm{d}y} = (\frac{\mathrm{d}y}{\mathrm{d}x})^{-1}$. One way or another, however, we have to get and use the inverse of h.

PROOF i. The first part is immediate from Thm. 2.10 and the definition of pdf.

ii. If h is strictly increasing and differentiable then, by using the chain rule,

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathsf{P}(X \le h^{-1}(y))$$

$$= \frac{\mathrm{d}}{\mathrm{d}y}F_X(h^{-1}(y)) = \left(\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y)\right) f_X(h^{-1}(y)).$$

But $\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y) = \frac{1}{h'(h^{-1}(y))}$.

If h is is strictly decreasing and differentiable then instead (exercise)

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}(1 - F_X(h^{-1}(y))).$$

In this case, we get a minus sign from the above and $h'(h^{-1}(y)) < 0$, which is why we need an absolute value (exercise). □

******* *Any time h is not 1-1, it is best to start by first obtaining $F_Y(y)$ and only then compute the derivative, assuming that is appropriate. See the proof of Thm. 2.10 and Ex. 2.9.*

Example 2.11 Suppose $X \sim$ exponential(10) and $Y = \log(X)$. (In this course, “log” *always* means $\log_e$, the natural logarithm.)

First note that the range of X is $(0, \infty)$, which means the range of Y is $(-\infty, \infty)$. The two rvs have different ranges.

Then the relationship is $x = \mathrm{e}^y$ so that $\frac{\mathrm{d}x}{\mathrm{d}y} = \mathrm{e}^y$. Hence

$$f_Y(y) = \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| f_X(x) = \frac{\mathrm{d}x}{\mathrm{d}y}\frac{1}{10}\mathrm{e}^{-x/10}1_{(0,\infty)}(x)$$

$$= \mathrm{e}^y\frac{1}{10}\mathrm{e}^{-\mathrm{e}^y/10}1_{(0,\infty)}(\mathrm{e}^y) = \frac{\mathrm{e}^{y-\mathrm{e}^y/10}}{10}1_{(-\infty,\infty)}(y).$$

Alternatively, we can go back to basics and first find

$$F_Y(y) = \mathsf{P}(Y \le y) = \mathsf{P}(X \le \mathrm{e}^y) = 1 - \mathrm{e}^{-\mathrm{e}^y/10}, \text{ for all real } y,$$

and then take a derivative to get the pdf $f_Y(y)$.

Example 2.12 Suppose we wish to simulate a random variable Y with continuous distribution F_Y and pdf f_Y.

A basic "random" number generator simulates $U \sim$ uniform$(0, 1)$. How can we convert U to a random variable with the distribution F_Y?

Solution We have $f_U(u) = 1_{[0,1]}(u)$ and we want $f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_Y(y)$. Ignoring the indicator for the moment, we thus want

$$\frac{\mathrm{d}u}{\mathrm{d}y} = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y}$$

which means we must have $U = F_Y(Y)$ or $Y = F_Y^{-1}(U)$. Since both U and $F_Y(Y)$ can take any value in $(0, 1)$, we are done.

For a particular case, assume we want $Y \sim$ exponential(β), $\beta > 0$. Let $u = F_Y(y) = 1 - \mathrm{e}^{-y/\beta}$ and we have $y = -\beta \log(1 - u)$. So we can simulate Y by defining $Y = -\beta \log(1 - U)$. (*Check:* this gives positive values for Y, which is necessary for the exponential distribution.)

THEOREM 2.13 Suppose F is a cdf that is continuous and 1-1.

i. If $U \sim$ uniform$(0,1)$ and $Y = F^{-1}(U)$ then $Y \sim F$.

ii. If $Y \sim F$ and $U = F(Y)$ then $U \sim$ uniform$(0,1)$.

The previous example is a case when F is absolutely continuous.

PROOF For example,

$$\mathsf{P}(Y \leq y) = \mathsf{P}(F^{-1}(U) \leq y) = \mathsf{P}(U \leq F(y)) = F(y).$$

□

Indeed, only continuity is needed for the second part and “inverse” can be generalized in a way so that the first part is valid for any distribution. See Thm. 2.37.

2.4 Expectation

Statistics is about averages, not just probabilities.

Example 2.5 (cont.) Suppose $Y \sim \text{binomial}(n, p)$. Assume first that we have sampled *with replacement* from an population of size N (and $M = pN$ is the number of successes in the population). There are N^n *equally likely samples* (outcomes).

What is the average value of $Y(s)$ among all these samples? What is the average of $h(Y(s))$?

Solution There are $\binom{n}{y} M^y (N-M)^{n-y}$ samples for which $Y(s) = y$, for each $y = 0, 1, \ldots, n$. So

$$\begin{aligned}\text{average}(Y) &= \frac{1}{N^n}\sum_{s\in\mathcal{S}} Y(s) = \frac{1}{N^n}\sum_{y=0}^{n}\sum_{s:Y(s)=y} Y(s) \\ &= \frac{1}{N^n}\sum_{y=0}^{n} y\binom{n}{y} M^y(N-M)^{n-y} = \sum_{y=0}^{n} y\binom{n}{y} p^y(1-p)^{n-y} \\ &= \sum_{y=0}^{n} y f_Y(y).\end{aligned}$$

Likewise, replacing $Y(s)$ with $h(Y(s))$,

$$\begin{aligned}\text{average}(h(Y)) &= \frac{1}{N^n}\sum_{s\in\mathcal{S}} h(Y(s)) = \sum_{y=0}^{n} h(y)\binom{n}{y}p^y(1-p)^{n-y} \\ &= \sum_{y=0}^{n} h(y)f_Y(y).\end{aligned}$$

We see that these average values depend only on the pmf for Y and not on the sample space per se. This suggests we can define average for any pmf.

DEFINITION 2.14 Let Y be a discrete rv with pmf f_Y, and let $h(y)$ be real valued.

i. If h is nonnegative then the expectation (mean, average) of $h(Y)$ is

$$\mathsf{E}(h(Y)) = \sum_y h(y)f_Y(y) \qquad \text{(can be } \infty\text{)}.$$

ii. More generally, if $\mathsf{E}(|h(Y)|) < \infty$ then we define

$$\mathsf{E}(h(Y)) = \sum_y h(y)f_Y(y).$$

Example 2.5 (cont.) $Y \sim \text{binomial}(n,p)$. Note that $y\binom{n}{y} = n\binom{n-1}{y-1}$ if $y > 0$. Evaluating the sum above, by changing variables $x = y - 1$,

$$\begin{aligned} \mathsf{E}(Y) &= \sum_{y=0}^{n} y\binom{n}{y}p^y(1-p)^{n-y} = \sum_{y=1}^{n} n\binom{n-1}{y-1}p^y(1-p)^{n-y} \\ &= np\sum_{x=0}^{n-1}\binom{n-1}{x}p^x(1-p)^{n-1-x} = np, \end{aligned}$$

because the last sum is the total of a binomial$(n-1,p)$ pmf (that is, equal to 1). We certainly "expect" Y to be near np, hence the term "expectation".

By a similar calculation (using $y(y-1)\binom{n}{y} = n(n-1)\binom{n-2}{y-2}$),

$$\begin{aligned} \mathsf{E}(Y(Y-1)) &= \sum_{y=0}^{n} y(y-1)\binom{n}{y}p^y(1-p)^{n-y} = \sum_{y=2}^{n} n(n-1)\binom{n-2}{y-2}p^y(1-p)^{n-y} \\ &= n(n-1)p^2\sum_{x=0}^{n-2}\binom{n-2}{x}p^x(1-p)^{n-2-x} = n(n-1)p^2. \end{aligned}$$

**** We often can evaluate a sum by recognizing it as a constant times the total of a pmf or as a constant times the value of some known sum.*

Example 2.4 (cont.) $X(s) = 1_A(s)$ where $\mathsf{P}(A) = p$. Here, $X \sim$ Bernoulli(p).

$$\mathsf{E}(1_A) = 0 \cdot \mathsf{P}(X = 0) + 1 \cdot \mathsf{P}(X = 1) = \mathsf{P}(A) = p.$$

This shows that probability is actually a special case of expectation!

If X is continuous with pdf f_X, we can discretize it: let $Y = \frac{\lfloor nX \rfloor}{n}$. Then $X - \frac{1}{n} < Y \le X$. If j is an integer then

$$\mathsf{P}(Y = j/n) = \int_{j/n}^{(j+1)/n} f_X(x)\mathrm{d}x \doteq \frac{1}{n} f_X(j/n).$$

Thus,

$$\mathsf{E}(X) \doteq \mathsf{E}(Y) = \sum_{j=-\infty}^{\infty} \frac{j}{n} f_X(j/n) \frac{1}{n} \doteq \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x$$

by the definition of integral (at least for “reasonable” functions).

This suggests the following.

DEFINITION 2.15 Let Y be a continuous rv with pdf f_Y, and let $h(y)$ be real valued.

i. If h is nonnegative then the expectation (mean, average) of $h(Y)$ is

$$\mathsf{E}(h(Y)) = \int_{-\infty}^{\infty} h(y) f_Y(y) \mathrm{d}y \qquad \text{(can be } \infty\text{)}.$$

ii. More generally, if $\mathsf{E}(|h(Y)|) < \infty$ then we define

$$\mathsf{E}(h(Y)) = \int_{-\infty}^{\infty} h(y) f_Y(y) \mathrm{d}y.$$

- In fact, the idea of defining expectation by first discretizing the random variable and then taking a limit of the discretized expectation is possible regardless of the type of rv. Doing this formally and showing that the result is coherently defined, however, is beyond this course.

- The reason that Def. 2.14 and Def. 2.15 have two parts is because we need to avoid the scenario where the positive part of the sum/integral and the negative part are both infinite. In that case, the expectation is *not defined*.

THEOREM 2.16 $\int_0^\infty y^m e^{-y}\mathrm{d}y = m!$ for any nonnegative integer m.

PROOF We know $\int_0^\infty \mathrm{e}^{-y}\mathrm{d}y = 1 = 0!$. We now integrate by parts. Note that a clever double integration can solve the integration by parts:

$$\begin{aligned}\int_0^\infty y^m \mathrm{e}^{-y}\mathrm{d}y &= \int_0^\infty \int_0^y m x^{m-1}\mathrm{d}x\ \mathrm{e}^{-y}\mathrm{d}y = \int_0^\infty \int_x^\infty \mathrm{e}^{-y}\mathrm{d}y\ m x^{m-1}\mathrm{d}x \\ &= m\int_0^\infty x^{m-1}\mathrm{e}^{-x}\mathrm{d}x.\end{aligned}$$

(Make sure you take care of the constants of integration! They are conveniently equal to 0 in this example.) The result then holds by induction (recursively applying the above). □

Example 2.13 Suppose $X \sim$ exponential(β). We have $F_X(x) = 1 - e^{-x/\beta}$ for $x > 0$ so $f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_X(x) = \frac{1}{\beta}e^{-x/\beta}1_{(0,\infty)}$. Let $h(x) = x^m$ for some positive integer m. Then, using a change of variables $y = x/\beta$,

$$\begin{aligned}\mathsf{E}(X^m) &= \int_0^\infty x^m f_X(x)\mathrm{d}x = \frac{1}{\beta}\int_0^\infty x^m \mathrm{e}^{-x/\beta}\mathrm{d}x \\ &= \beta^m \int_0^\infty y^m \mathrm{e}^{-y} dy = m!\beta^m.\end{aligned}$$

- Note that *a change of variables* can make the integration easier. But pay close attention to variable changes.
- There often are clues as well. For example, let $Y = X/\beta$ in Ex. 2.13 above. Then $F_Y(y) = 1 - \mathrm{e}^{-y}$ (check) which shows that the distribution of Y does not depend on β. Since $X = \beta Y$ then it is sensible that $\mathsf{E}(X^m) = \beta^m \mathsf{E}(Y^m)$, and that means the value must be proportional to β^m.
- There are three principal ways to compute an expectation *indirectly* (which is often better than computing it directly).
  i. Show that the integral (sum) is proportional to the total integral (sum) of a pdf (pmf) or other known function.
  ii. Integrate (sum) by parts.
  iii. Differentiate or integrate with respect to a parameter under the integral (sum).
- These methods tend to require we know families of functions and their integrals or sums. Which method works best or easiest depends on the problem.

Example 2.2 (cont.) T = number of trials until the 1^{st} success. So $T \sim$ geometric(p), $f_T(t) = p(1-p)^{t-1}1_{\{1,2,\dots\}}$.

Find E(T).

Solution Try each of the three methods mentioned above.

i. Recall (Ex. 1.18) we found the pmf for $V = \#$ flips until the k^{th} success. For $k = 2$ we have $f_V(v) = (v-1)p^2(1-p)^{v-2}1_{\{2,3,\dots\}}$, which must sum to 1. Now let $t = v - 1$ and compute

$$\mathsf{E}(T) = \sum_{t=1}^{\infty} tp(1-p)^{t-1} = \frac{1}{p}\sum_{v=2}^{\infty}(v-1)p^2(1-p)^{v-2} = \frac{1}{p}.$$

ii. Summation by parts can often be done by using a double sum.

$$\begin{aligned}\mathsf{E}(T) &= \sum_{t=1}^{\infty} tp(1-p)^{t-1} = \sum_{t=1}^{\infty}\sum_{j=1}^{t} p(1-p)^{t-1} = p\sum_{j=1}^{\infty}\sum_{t=j}^{\infty}(1-p)^{t-1} \\ &= p\sum_{j=1}^{\infty}\frac{(1-p)^{j-1}}{1-(1-p)} = \sum_{j=1}^{\infty}(1-p)^{j-1} = \frac{1}{p}.\end{aligned}$$

iii. Technically, this method requires a justification for exchanging summation with differentiation, okay here since $|1-p| < 1$. We will rarely need to worry about that, but it is something to keep in mind more generally.

$$\begin{aligned} \mathsf{E}(T) &= \sum_{t=1}^{\infty} tp(1-p)^{t-1} = p\sum_{t=1}^{\infty}\Big(\frac{\mathrm{d}}{\mathrm{d}x}x^t\Big)_{x=1-p} = p\Big(\frac{\mathrm{d}}{\mathrm{d}x}\sum_{t=1}^{\infty}x^t\Big)_{x=1-p} \\ &= p\Big(\frac{\mathrm{d}}{\mathrm{d}x}\frac{x}{1-x}\Big)_{x=1-p} = \frac{p}{(1-(1-p))^2} = \frac{1}{p}. \end{aligned}$$

Another option to consider is “which distribution to use”, when you have a choice.

THEOREM 2.17 (Change of Variables) Let $Y = h(X)$. The value of $\mathsf{E}(Y)$ (computed using F_Y) is the same as the value of $\mathsf{E}(h(X))$ (computed using F_X).

PROOF This is not nearly as obvious as it seems, for it holds quite generally. At least for simple cases, it follows from the change of variables formulas discussed in Sec. 2.3. □

Reminder: Although we have not done so explicitly, expectation is defined for all nonnegative random variables and for all random variables whose absolute value has finite expectation. This frees us up to discuss and utilize expectation without always having to refer back to pmfs or pdfs.

THEOREM 2.18 (Properties of Expectations)

i. $h(X) \geq 0 \implies \mathsf{E}(h(X)) \geq 0.$ (positivity)

ii. $\mathsf{E}(ah(X) + b) = a\mathsf{E}(h(X)) + b$ and $\mathsf{E}(h_1(X) + h_2(X)) = \mathsf{E}(h_1(X)) + \mathsf{E}(h_2(X))$. (linearity)

iii. $h_1(X) \geq h_2(X) \implies \mathsf{E}(h_1(X)) \geq \mathsf{E}(h_2(X))$. (monotonicity)

iv. $a \leq h(X) \leq b \implies a \leq \mathsf{E}(h(X)) \leq b.$

v. $h(X) \geq 0$ and $\mathsf{E}(h(X)) = 0 \implies \mathsf{P}(h(X) = 0) = 1.$

vi. $X_1 \sim F$ and $X_2 \sim F \implies \mathsf{E}(h(X_1)) = \mathsf{E}(h(X_2))$. (equivalence)

**** Even though we have limited definitions to discrete and absolutely continuous cases, each of the results in the theorem above are completely general.*

PROOF i.–v. all follow from the analogous properties of sums and integrals, the fact that $f_X(x)$ is nonnegative, and the fact that f_X sums or integrates to 1.

For example, if X has pdf f_X then

$$\begin{aligned}\mathsf{E}(ah(X)+b) &= \int_{-\infty}^{\infty}(ah(x)+b)f_X(x)\mathrm{d}x \\ &= a\int_{-\infty}^{\infty}h(x)f_X(x)dx + b\int_{-\infty}^{\infty}f_X(x)\mathrm{d}x = a\mathsf{E}(h(X))+b.\end{aligned}$$

vi. holds because only the distribution of the rv is used to compute any expectation. □

Example 2.14 Suppose Z has density $f_Z(z) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}$. This is the standard normal distribution.

Let $X = \mu + \sigma Z$ with $\sigma > 0$ and compute $\mathsf{E}(X)$ and $\mathsf{E}(X^2)$.

Solution Method 1. First, note that $\frac{\mathrm{d}}{\mathrm{d}z} f_Z(z) = -z f_Z(z)$ by the chain rule. We can now compute

$$\mathsf{E}(Z) = \int_{-\infty}^{\infty} z f_Z(z)\mathrm{d}z = -f_Z(z)\Big|_{-\infty}^{\infty} = 0.$$

Then, using integration by parts (with $u = z$, $v = f_Z(z)$),

$$\mathsf{E}(Z^2) = \int_{-\infty}^{\infty} z\Big(z f_Z(z)\Big)\mathrm{d}z = -z f_Z(z)\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} f_Z(z)\mathrm{d}z = 1.$$

Next,

$$\mathsf{E}(X) = \mathsf{E}(\mu + \sigma Z) = \mu + \sigma\mathsf{E}(Z) = \mu$$

and

$$\mathsf{E}(X^2) = \mathsf{E}(\mu^2 + 2\mu\sigma Z + \sigma^2 Z^2) = \mu^2 + 2\mu\sigma\mathsf{E}(Z) + \sigma^2\mathsf{E}(Z^2) = \mu^2 + \sigma^2.$$

Method 2. First, find the pdf for X. Note that $h(z) = \mu + \sigma z$ is a differentiable 1-1 function and for $z = h^{-1}(x) = \frac{x-\mu}{\sigma}$ we have $\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{1}{\sigma}$.

By Cor. 2.12ii.,

$$f_X(x) = \frac{\mathrm{d}z}{\mathrm{d}x} f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}.$$

This is the normal(μ, σ^2) density function.

We can then compute, for example,

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}\,\mathrm{d}x = \int_{-\infty}^{\infty} (\mu + \sigma z) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \mathrm{d}z = \cdots = \mu,$$

where we make an immediate change of variables $z = (x - \mu)/\sigma$, and then basically integrate as in Method 1.

Example 2.3 (cont.) Recall the cell phone battery that failed immediately with probability $.05$ and has exponential(10) lifetime otherwise. Recall also that the measured lifetime T is censored at value 30. We saw that the cdf $F_T(t)$ has jumps of size $.05$ at $t = 0$ and $.95\mathrm{e}^{-3}$ at $t = 30$. Moreover, for $0 < t < 30$, the cdf has derivative $.095\mathrm{e}^{-t/10}$.

Just as we would do for probabilities, we can compute expectations by separately handling the discrete part with a sum and the continuous part with an integral, and then *combining* the results. So, for example,

$$\begin{aligned}\mathsf{E}(T) &= 0(.05) + \int_0^{30} t(.095\mathrm{e}^{-t/10})\,\mathrm{d}t + 30(.95\mathrm{e}^{-3}) \\ &= \cdots = 9.5(1 - \mathrm{e}^{-3}).\end{aligned}$$

Similarly,

$$\begin{aligned}\mathsf{E}(\mathrm{e}^{T/5}) &= \mathrm{e}^0(.05) + \int_0^{30} \mathrm{e}^{t/5}(.095\mathrm{e}^{-t/10})\,\mathrm{d}t + \mathrm{e}^6(.95\mathrm{e}^{-3}) \\ &= \cdots = 10.45\mathrm{e}^3 - 9.45.\end{aligned}$$

**** When the cdf is a mixture of discrete and absolutely continuous parts, both probabilities and expectations are computed by adding the corresponding sum and integral.*

2.5 Moments, Mean and Variance

There are many ways to characterize random variables. The most common of these are in terms of primitive expectations called moments.

DEFINITION 2.19 Let X be a random variable and let k be a positive integer.

i. The k-th moment of X is $\mu'_k = \mathsf{E}(X^k)$, if $\mathsf{E}(|X|^k) < \infty$.

ii. The k-th central moment of X is $\mu_k = \mathsf{E}((X - \mathsf{E}(X))^k)$, if $k > 1$ and $\mathsf{E}(|X|^k) < \infty$.

iii. Important special cases are the mean of X: $\mu_X = \mu'_1 = \mathsf{E}(X)$, and the variance of X: $\mathsf{var}(X) = \mu_2 = \mathsf{E}((X - \mu_X)^2)$ (often denoted σ_X^2).
The standard deviation of X is $\sigma_X = \sqrt{\mathsf{var}(X)}$.

Note: k need not be positive or an integer in Def. 2.19.i if X is nonnegative.

THEOREM 2.20 Suppose $Y = aX + b$ with $a \neq 0$.

i. $\sigma_X^2 = \mathsf{E}(X^2) - (\mathsf{E}(X))^2 = \mathsf{E}(X(X-1)) - \mathsf{E}(X)(\mathsf{E}(X)-1)$.

ii. $\mu_Y = a\mu_X + b$.

iii. $\sigma_Y^2 = a^2\sigma_X^2$ and $\sigma_Y = |a|\sigma_X$.

PROOF

i. Use the linearity property (Thm. 2.18.ii)

$$\begin{aligned}\sigma_X^2 &= \mathsf{E}((X-\mu_X)^2) = \mathsf{E}(X^2 - 2\mu_X X + \mu_X^2) \\ &= \mathsf{E}(X^2) - 2\mu_X\mathsf{E}(X) + \mu_X^2 = \mathsf{E}(X^2) - (\mathsf{E}(X))^2,\end{aligned}$$

since $\mu_X = E(X)$. The other equality is established similarly.

ii. This is Thm. 2.18.ii.

iii. Again use linearity.

$$\begin{aligned}\sigma_Y^2 &= \mathsf{E}((Y-\mu_Y)^2) = \mathsf{E}((aX+b-(a\mu_X+b))^2) = \mathsf{E}((aX-a\mu_X)^2) \\ &= a^2\mathsf{E}((X-\mu_X)^2) = a^2\sigma_X^2.\end{aligned}$$

The formula for the σ_Y then follows, taking the *positive* square root.

A helpful consequence of i. is $\mathsf{E}(X^2) = \sigma_X^2 + \mu_X^2$. □

Example 2.14 (cont.) $Z \sim$ standard normal and $X = \mu + \sigma Z$. We have seen that $\mu_Z = \mathsf{E}(Z) = 0$ and $\mathsf{var}(Z) = \mathsf{E}((Z - \mu_Z)^2) = \mathsf{E}(Z^2) = 1$.

So we also obtain $\mu_X = \mu + \sigma\mu_Z = \mu$ and $\sigma_X^2 = \sigma^2\,\mathsf{var}(Z) = \sigma^2$. Thus the parameters for the normal(μ, σ^2) distribution are *its mean and variance*. (Note: this is not usually the case for parametric distributions.)

The standard deviation is a measure of how much a rv X differs from its mean, in an average or typical sort of way. (Since it really is the square root of the average squared deviation, it actually is a bit larger than the average absolute deviation.) As such it is a simple measurement of the "randomness" of X in the same units as X.

Recalling that one objective of statistics is to predict uncertainty in the face of randomness, we can see that the standard deviation (or the variance) is *as powerful a tool* as the average itself is.

Example 2.13 (cont.) $X \sim \text{exponential}(\beta)$. Here, $\mu_X = \mathsf{E}(X) = \beta$ and

$$\text{var}(X) = \mathsf{E}(X^2) - \mu_X^2 = 2\beta^2 - \beta^2 = \beta^2.$$

Example 2.5 (cont.) $Y \sim \text{binomial}(n, p)$. We have seen that $\mu_Y = \mathsf{E}(Y) = np$.

Since also $\mathsf{E}(Y(Y-1)) = n(n-1)p^2$ we obtain (using the second equality in Thm. 2.20.i)

$$\begin{aligned} \text{var}(Y) &= \mathsf{E}(Y(Y-1)) - \mathsf{E}(Y)(\mathsf{E}(Y)-1) \\ &= n(n-1)p^2 - np(np-1) = np(1-p). \end{aligned}$$

Recall (Ex. 1.10) that we first derived the binomial(n, p) distribution in order to have a way to predict the number of Yes's (successes) in n responses (Bernoulli trials) selected with replacement (independently).

We desired to be able to statistically predict the observed success rate $\widehat{p} = \frac{Y}{n}$, which estimates p. We now can see that

$$\mathsf{E}(\widehat{p}) = \frac{np}{n} = p$$

and

$$\text{var}(\widehat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Some pdfs and pmfs, such as the normal pdf, are symmetric in the sense that there exists a value c such that $f(c+x) = f(c-x)$ for all x. It is almost obvious, and easy to show, that if X has a mean and its pdf or pmf is symmetric about c then $\mu_X = c$.

On the other hand, if the pdf or pmf is not symmetric then there is no clear or unique definition of "center".

Nevertheless, μ_X is the "center of balance", thinking as if $f_X(x)$ is a weight function on the real line. The mean is also the best "predictor" of X, in an *average squared error* sense, as the next theorem explains.

THEOREM 2.21 Suppose X is a rv with $\mathsf{E}(X^2) < \infty$. The value c that minimizes $\mathsf{E}((X-c)^2)$ is $c = \mu_X$.

PROOF (exercise: The expression to minimize is a quadratic function in c.) □

Example 2.15 Let $f_X(x) = \frac{1}{m!}x^m \mathrm{e}^{-x}$. This is an example of the gamma distribution (with parameter $\alpha = m+1$). Using Thm. 2.16,

$$\mathsf{E}(X) = \frac{1}{m!}\int_0^\infty x^{m+1}\mathrm{e}^{-x}\mathrm{d}x = \frac{(m+1)!}{m!} = m+1.$$

Likewise, $\mathsf{E}(X^2) = \frac{(m+2)!}{m!} = (m+1)(m+2)$, so that
$\mathsf{var}(X) = (m+1)(m+2) - (m+1)^2 = m+1.$

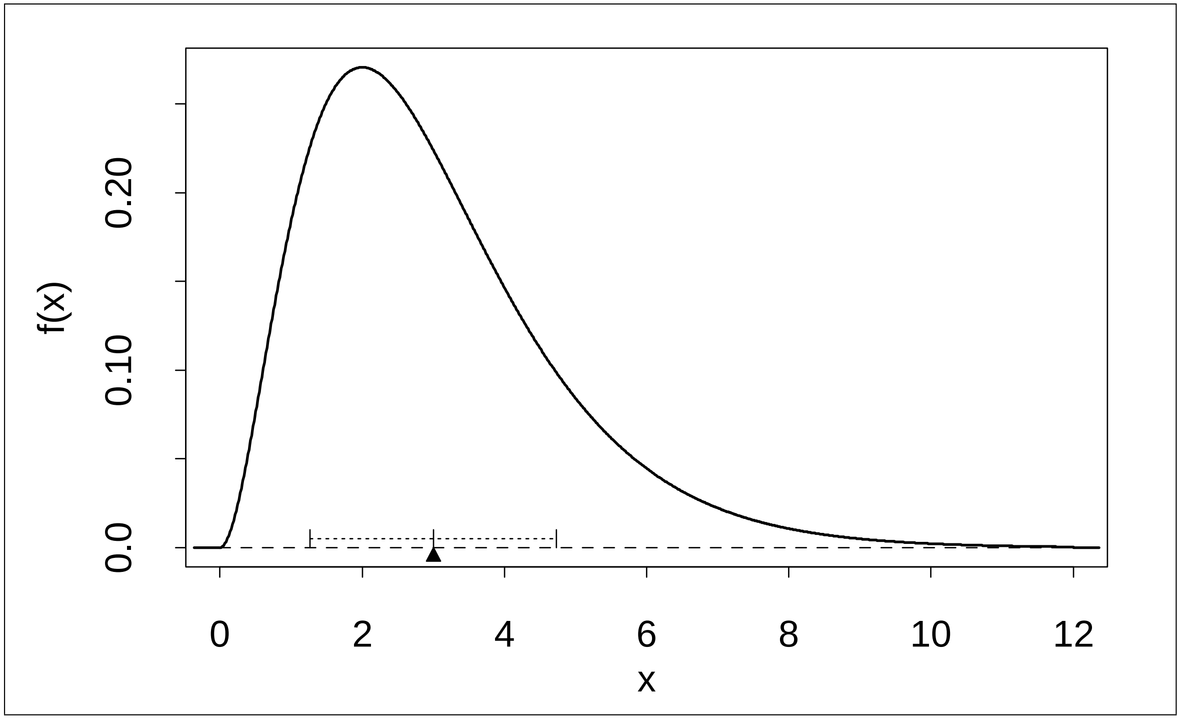

Figure 2.8 Gamma(3,1) probability density function, with mean (center of balance) and standard deviation ("typical" distance from the mean) indicated.

Many tend to view the mean and variance as the primary characteristics of a distribution, so much so that they forget they really only provide the most basic of information about the distribution.

There are other useful values that depend on the moments.

DEFINITION 2.22 (Additional Moment Quantities) Let X be a rv.

i. If X is *nonnegative*, the coefficient of variation, or relative variation, is σ_X/μ_X.

ii. If $\mathsf{E}(|X|^3) < \infty$, the skewness coefficient is $\frac{1}{\sigma_X^3}\mathsf{E}((X-\mu_X)^3)$.

iii. If $\mathsf{E}(|X|^4) < \infty$, the kurtosis coefficient is $\frac{1}{\sigma_X^4}\mathsf{E}((X-\mu_X)^4)$.

Skewness is an indicator of asymmetry, especially in the “tails” of the distribution (the nature of the distribution on the extreme left and right).

Kurtosis is an indicator of how “heavy” the tails are, that is, whether they damp quickly or slowly.

In fact moments have a lot to say about the distribution tails.

THEOREM 2.23 Let X be a rv.

i. (Markov) If $t > 0$ and $m > 0$ then $\mathsf{P}(|X| > t) \le \frac{\mathsf{E}(|X|^m)}{t^m}$.

ii. (Chebychev) If $t > 0$ then $\mathsf{P}(|X - \mu_X| > \sigma_X t) \le \frac{1}{t^2}$.

PROOF

i. We first note a simple inequality: $1_{|x|>t} \le \left(\frac{|x|}{t}\right)^m$ as long as t and m are both positive. Applying Thm. 2.18.iii.,

$$\mathsf{P}(|X| > t) = \mathsf{E}(1_{|X|>t}) \le E\left(\left(\frac{|X|}{t}\right)^m\right) = \frac{\mathsf{E}(|X|^m)}{t^m}.$$

ii. This follows from the Markov inequality applied to $\frac{X-\mu_X}{\sigma_X}$ in place of X, and with $m = 2$.

□

Probabilities such as $\mathsf{P}(X < -t)$, $\mathsf{P}(X > t)$ and $\mathsf{P}(|X| > t)$ are known as tail probabilities for (possibly large) positive t.

Additional points about moments are the following.

THEOREM 2.24 Let X be a rv and $m > 0$.

i. If $\mathsf{E}(|X|^m) < \infty$ then $\mathsf{E}(|X|^k) < \infty$ for all $k \in (0, m)$.

ii. If $x^{m+\delta}\mathsf{P}(|X| > x)$ is *bounded*, as a function of x, for some $\delta > 0$, then $\mathsf{E}(|X|^m) < \infty$.

iii. If $\mathsf{E}(|X|^m) < \infty$ then $x^m\mathsf{P}(|X| > x) \to 0$, as $x \to \infty$.

PROOF

i. Suppose $0 < k < m$. By considering the two cases $|x| \le 1$ and $|x| > 1$, we note that $|x|^k \le 1 + |x|^m$. Applying Thm. 2.18.iii.,

$$\mathsf{E}(|X|^k) \le 1 + \mathsf{E}(|X|^m) < \infty.$$

ii. and iii. These are not too difficult to show here, but the proof is somewhat involved.

□

Note the distinction between ii. and iii. They are *not quite* converses of each other.

It is helpful to know that if the pdf (or the pmf in the integer-valued case) of X decreases as fast as $|x|^{-\beta-1}$, as $x \to -\infty$ or $x \to \infty$ then $\mathsf{P}(|X| > x)$ will decrease as fast as $x^{-\beta}$, as $x \to \infty$.

The converse is true if the pdf/pmf is eventually *monotone* decreasing. But, unfortunately, it is not true in general.

Example 2.16 Let $\alpha > 0$ and $\beta > 0$. Suppose V has pdf

$$f_V(v) = K\frac{v^{\alpha-1}}{(1+v)^{\alpha+\beta}}, \quad \text{for } v > 0,$$

where K is chosen to ensure f_V integrates to 1. Then $f_V(v) \sim Kv^{-\beta-1}$ as $v \to \infty$, and

$$\mathsf{E}(V^m) = \int_0^\infty v^m K\frac{v^{\alpha-1}}{(1+v)^{\alpha+\beta}}\mathrm{d}v,$$

which is finite only if $m < \beta$.

This distribution is closely related to Snedecor's F distribution, a very important distribution used in statistical analysis.

It is very tempting to hope that moments are related to each other in a simple functional way. But, alas, this prospect is foiled by randomness. For example, as a general rule, $\mathsf{E}(h(X)) \neq h(\mathsf{E}(X))$ except when $h(x)$ is linear.

THEOREM 2.25 (Jensen's Inequality) Suppose $g(x)$ is a *convex* function but *not a straight line* on the support of X, and X is a *nondegenerate* rv such that both $\mathsf{E}(X)$ and $\mathsf{E}(g(X))$ are finite. Then $\mathsf{E}(g(X)) > g(\mathsf{E}(X))$.

In particular,

i. $|\mathsf{E}(X)| \leq \mathsf{E}(|X|)$, with equality only if $\mathsf{P}(X \geq 0) = 1$ or 0.

ii. if $0 < k < m$ and $X \geq 0$ then $\mathsf{E}(X^k) < (\mathsf{E}(X^m))^{k/m}$,

iii. $\mathrm{e}^{a\mathsf{E}(x)} < \mathsf{E}(\mathrm{e}^{aX})$, for any real $a \neq 0$,

iv. if $X > 0$ then $\mathsf{E}(1/X) > 1/\mathsf{E}(X)$.

Recall that if $g(x)$ is twice differentiable and $g''(x) \geq 0$ for all x then g is convex.

More simply sometimes, if $g(x)$ is differentiable and $g'(x)$ is nondecreasing then $g(x)$ is convex.

PROOF One definition of convex is that the graph of $y = g(x)$ lies above all its tangent lines. So choose a tangent line $g(\mathsf{E}(X)) + b(x - \mathsf{E}(X))$ passing through the point $(\mathsf{E}(X), g(\mathsf{E}(X)))$. (See Fig. 2.9.) Then

$$\mathsf{E}(g(X)) \geq \mathsf{E}(g(\mathsf{E}(X)) + b(X - \mathsf{E}(X))) = g(\mathsf{E}(X)),$$

with equality only if g coincides with the tangent line over the support of X.

Special cases include (i) $g(x) = |x|$, (ii) $g(x) = x^{m/k}$ if $k < m$ and $x \geq 0$, (iii) $g(x) = \mathrm{e}^{ax}$ and (iv) $g(x) = 1/x$ if $x > 0$. □

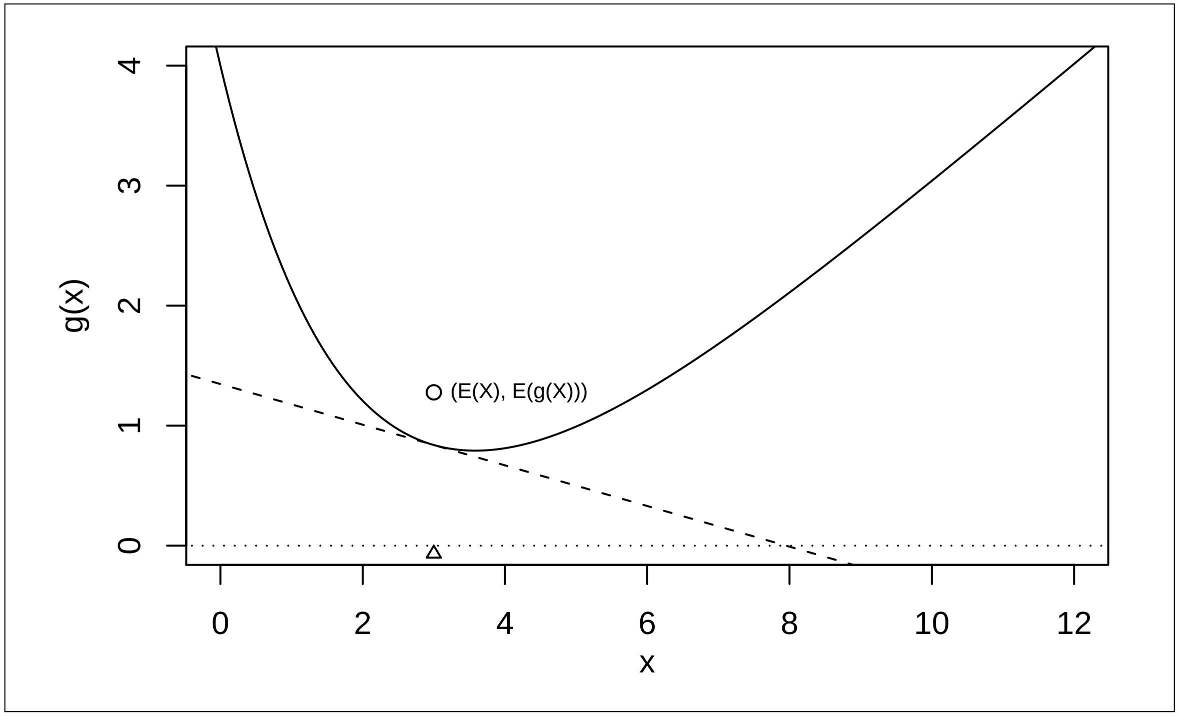

Figure 2.9 Example illustrating Jensen's inequality.

2.6 Generating Functions

If you are familiar with the idea of Laplace and Fourier transforms, you know that many functions have dual functions (transforms) that are unique to the original functions, but in a different form. This is true of distribution functions as well.

DEFINITION 2.26 Let X be a random variable.

i. The moment generating function (mgf) of X is $M_X(t) = \mathsf{E}(\mathrm{e}^{tX})$, *assuming this is finite* in an open interval containing 0.

ii. If X is *nonnegative integer-valued*, the probability generating function (pgf) is $p_X(t) = \mathsf{E}(t^X)$, $t \geq 0$.

iii. The characteristic function (cf) of X is $\phi_X(t) = \mathsf{E}(\mathrm{e}^{\mathrm{i}tX}) = \mathsf{E}(\cos(tX) + \mathrm{i}\sin(tX))$. ($\mathrm{i} = \sqrt{-1}$.)

The moment generating function does not necessarily exist. The characteristic function, however, is always finite but it usually is complex-valued. Despite the fact that mgfs do not always exist, we will focus on mgfs in this course.

The three types of transforms are related to each other.

THEOREM 2.27 Assume $M_X(t) < \infty$ for all $t \in (-\delta, \delta)$ where $\delta > 0$.

i. $\phi_X(t) = M_X(\mathrm{i}t)$.

ii. $p_X(t) = M_X(\log t)$.

PROOF These are apparent from the definitions. □

**** Observe that $M_X(t) \geq 0$ and $M_X(0) = 1$. These are useful points to check when deriving a mgf.*

Example 2.13 (cont.) $X \sim$ exponential(β), $f_X(x) = \frac{1}{\beta}e^{-x/\beta}$, $x > 0$. Find the mgf for X.

Solution

$$\begin{aligned} M_X(t) &= \int_0^\infty \mathrm{e}^{tx}\tfrac{1}{\beta}\mathrm{e}^{-x/\beta}\mathrm{d}x = \tfrac{1}{\beta}\tfrac{1}{1/\beta-t}\int_0^\infty (1/\beta - t)\mathrm{e}^{-(1/\beta-t)x}\mathrm{d}x \\ &= \begin{cases} \frac{1}{1-\beta t} & \text{if } t < 1/\beta, \\ \infty & \text{if } t \geq 1/\beta. \end{cases} \end{aligned}$$

Check that $M_X(t) \geq 0$ and $M_X(0) = 1$.

Example 2.5 (cont.) $Y \sim$ binomial(n,p), $f_Y(y) = \binom{n}{y}p^y(1-p)^{n-y}$, for $y = 0,\ldots,n$. Find the mgf for Y.

Solution First, recall the binomial formula

$$\sum_{y=0}^{n}\binom{n}{y}a^y b^{n-y} = (a+b)^n.$$

Then

$$\begin{aligned} M_Y(t) &= \sum_{y=0}^{n} \mathrm{e}^{ty}\binom{n}{y}p^y(1-p)^{n-y} = \sum_{y=0}^{n}\binom{n}{y}(p\mathrm{e}^t)^y(1-p)^{n-y} \\ &= (p\mathrm{e}^t + 1 - p)^n. \end{aligned}$$

Check that $M_X(t) \geq 0$ and $M_X(0) = 1$.

Example 2.14 (cont.) $Z \sim$ normal$(0,1)$, $X = \mu + \sigma Z \sim$ normal(μ, σ). Find the mgfs for both Z and X.

Solution For the mgf of Z, we use the trick of *completing the square* in the exponent: $tz - z^2/2 = -(z-t)^2/2 + t^2/2$, and then observing that we can integrate the normal$(t,1)$ pdf. This goes as follows.

$$M_Z(t) = \int_{-\infty}^{\infty} \mathrm{e}^{tz} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \mathrm{d}z = \mathrm{e}^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(z-t)^2/2} \mathrm{d}z = \mathrm{e}^{t^2/2}.$$

Then to get the mgf for X, we need only to relate it to the mgf for Z.

$$M_X(t) = \mathsf{E}(\mathrm{e}^{tX}) = \mathsf{E}(\mathrm{e}^{t(\mu+\sigma Z)}) = \mathrm{e}^{\mu t} \mathsf{E}(e^{\sigma t Z}) = \mathrm{e}^{\mu t} M_Z(\sigma t) = \mathrm{e}^{\mu t + \sigma^2 t^2/2}.$$

Check that $M_X(t) \geq 0$ and $M_X(0) = 1$.

In the previous example we essentially proved the following.

THEOREM 2.28 Let X have mgf M_X and let $Y = aX + b$. Then Y has mgf $M_Y(t) = \mathrm{e}^{bt} M_X(at)$.

A characterization and the foremost property of mgfs are in the following theorem, which is too much to prove here. The second part is the one to note.

Theorem 2.29

i. $M(t)$ is a mgf if and only if $M(0) = 1$ and $M(t) = M_1(t) + M_2(-t)$ where $M_1(t)$ and $M_2(t)$ are nonnegative, infinitely differentiable (wherever they are finite) and all their derivatives are nonnegative.

ii. (Invertibility) If $M(t)$ is a mgf, finite for $t \in (-\delta, \delta)$, then *there is exactly one distribution* that it is the mgf for.

Thm. 2.29.ii. says that knowing $M_X(t)$ is the mgf of X is *equivalent* to knowing the distribution of X, and that all the properties of X are (in principle) derivable from $M_X(t)$. The astounding thing is that it may be much easier to demonstrate some properties with the mgf than with the cdf.

One such problem gives the mgf its name.

THEOREM 2.30 If $M_X(t)$ exists as the mgf for X then every moment of X is finite and

$$\mathsf{E}(X^m) = \frac{\mathrm{d}^m}{\mathrm{d}t^m} M_X(t)\Big|_{t=0} \qquad \text{for each } m = 1, 2, \ldots$$

The mean and variance can also be found by

$$\mathsf{E}(X) = \frac{\mathrm{d}}{\mathrm{d}t} \log(M_X(t))\Big|_{t=0} \quad \text{and} \quad \mathsf{var}(X) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} \log(M_X(t))\Big|_{t=0}.$$

PROOF (Heuristic) The condition that M_X is finite in an open interval containing 0 ensures we can switch differentiation with expectation (integral or sum). Thus

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m} M_X(t) = \frac{\mathrm{d}^m}{\mathrm{d}t^m} \mathsf{E}(\mathrm{e}^{tX}) = \mathsf{E}\left(\frac{\mathrm{d}^m}{\mathrm{d}t^m} \mathrm{e}^{tX}\right) = \mathsf{E}(X^m \mathrm{e}^{tX}).$$

Setting $t = 0$ gives $\mathsf{E}(X^m)$ since $\mathrm{e}^0 = 1$.

The alternate calculation for mean and variance is left for exercise. □

Indeed, using the Taylor series expansion for e^{tx} we see

$$M_X(t) = \mathsf{E}(\mathrm{e}^{tX}) = \mathsf{E}\left(\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right) = \sum_{n=0}^{\infty} \frac{\mathsf{E}(X^n) t^n}{n!}.$$

Example 2.14 (cont.) $Z \sim$ normal$(0,1)$, $X = \mu + \sigma Z \sim$ normal(μ, σ^2).

$$\mathsf{E}(Z) = \frac{\mathrm{d}}{\mathrm{d}t} M_Z(t) \Big|_{t=0} = t\mathrm{e}^{t^2/2} \Big|_{t=0} = 0 \quad \text{(which we knew)},$$

$$\mathsf{E}(Z^2) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} M_Z(t) \Big|_{t=0} = (1+t^2)\mathrm{e}^{t^2/2} \Big|_{t=0} = 1 \quad \text{(which we also knew)},$$

$$\mathsf{E}(Z^3) = \frac{\mathrm{d}^3}{\mathrm{d}t^3} M_Z(t) \Big|_{t=0} = 0 \quad \text{(symmetry)},$$

$$\mathsf{E}(Z^4) = \frac{\mathrm{d}^4}{\mathrm{d}t^4} M_Z(t) \Big|_{t=0} = 3 \quad \text{(exercise)}.$$

It follows that the skewness and kurtosis of X are

$$\mathsf{E}\left(\left(\tfrac{X-\mu}{\sigma}\right)^3\right) = \mathsf{E}(Z^3) = 0 \quad \text{and} \quad \mathsf{E}\left(\left(\tfrac{X-\mu}{\sigma}\right)^4\right) = \mathsf{E}(Z^4) = 3,$$

respectively.

Example 2.5 (cont.) $Y \sim \text{binomial}(n,p)$. We have determined that $M_Y(t) = (p\mathrm{e}^t + 1 - p)^n$. Now use this to evaluate the mean and variance of Y.

Solution First,

$$\mathsf{E}(Y) = \frac{\mathrm{d}}{\mathrm{d}t}(p\mathrm{e}^t + 1 - p)^n \Big|_{t=0} = np\mathrm{e}^t(p\mathrm{e}^t + 1 - p)^{n-1} \Big|_{t=0} = np.$$

Next,

$$\begin{aligned} \mathsf{E}(Y^2) &= \frac{\mathrm{d}}{\mathrm{d}t}\Big(np\mathrm{e}^t(p\mathrm{e}^t + 1 - p)^{n-1}\Big) \Big|_{t=0} \\ &= \Big(n(n-1)p^2\mathrm{e}^{2t}(p\mathrm{e}^t + 1 - p)^{n-2} + np\mathrm{e}^t(p\mathrm{e}^t + 1 - p)^{n-1}\Big) \Big|_{t=0} \\ &= n(n-1)p^2 + np. \end{aligned}$$

Finally,

$$\mathsf{var}(Y) = \mathsf{E}(Y^2) - (\mathsf{E}(Y))^2 = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

(Try using the alternate method based on derivatives of $\log(M_Y(t))$.)

Pgfs have a property analogous to Thm. 2.30 for mgfs.

Theorem 2.31

i. Let $p_X(t) = \sum_{x=0}^{\infty} f_X(x)t^x$ be the pgf of a *nonnegative integer* rv X. Then

$$f_X(n) = \mathsf{P}(X = n) = \frac{1}{n!}\left(\frac{\mathrm{d}^n}{\mathrm{d}t^n} p_X(t)\,\Big|_{t=0}\right) \quad \text{for each } n = 0, 1, \ldots$$

($f_X(0) = p_X(0)$.)

ii. For $p(t)$ to be the pgf of a nonnegative integer rv, it suffices that $p(1) = 1$ and $\frac{d^n}{dt^n} p(t)\,\Big|_{t=0} \geq 0$ for each $n = 0, 1, \ldots$

Example 2.17 Let $p(t) = \mathrm{e}^{\lambda(t-1)}$, $\lambda > 0$. Then $p(1) = 1$ and

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} p(t) \Big|_{t=0} = \lambda^n e^{-\lambda} \geq 0.$$

Therefore, $p(t)$ is a pgf and $f(x) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$, $x = 0, 1, \ldots,$ is the pmf of some rv X. This is the Poisson(λ) pmf.

(The mgf is thus $M(t) = p(\mathrm{e}^t) = \mathrm{e}^{\lambda(\mathrm{e}^t - 1)}$.)

Recalling the Taylor series expansion of e^λ, we can check:

$$\sum_{x=0}^{\infty} f(x) = \sum_{x=0}^{\infty} \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} = \mathrm{e}^{\lambda} \mathrm{e}^{-\lambda} = 1.$$

The real mathematical power of mgfs is in the following, which would be proven in a high-level probability course or in a real analysis course. (There is a more general version for characteristic functions.)

THEOREM 2.32 Let X be a rv and let $X_1, X_2, \ldots$ be a sequence of rvs. Suppose, for some $\delta > 0$, the mgfs of each are finite on $(-\delta, \delta)$. Then, as $n \to \infty$,

$$\begin{aligned} &M_{X_n}(t) \to M_X(t) \quad \text{for all } t \in (-\delta, \delta) \\ &\quad \iff F_{X_n}(x) \to F_X(x) \text{ for all } x\text{, except possibly where } F_X \text{ has a jump.} \end{aligned}$$

In this case we say the sequence X_n converges in distribution to F_X.

In other words, *convergence of the mgfs is equivalent to convergence of the cdfs*.

The reason for the qualifier "except possibly where F_X has a jump" is because the limit may not exist at such points, and where it does the actual limiting function need not be right continuous. But in all other respects the limit looks just like F_X.

**** It is only the distribution functions that converge, not necessarily the sequence of rvs.*

Before applying the theorem, we recall a result from calculus.

Theorem 2.33 Suppose $x_n \to x$, as $n \to \infty$. Then $(1 + \frac{x_n}{n})^n \to \mathrm{e}^x$, as $n \to \infty$.

Proof (Very heuristic) For all large n,

$$n \log\left(1 + \frac{x_n}{n}\right) = n\left(\frac{x_n}{n} - \frac{x_n^2}{2n^2} + \frac{x_n^3}{3n^3} - \cdots\right) \doteq x_n \to x.$$

□

Example 2.18 (Poisson Approximation to Binomial) Binomial probabilities are clearly difficult to compute if n is large. However, if p is very small (rare successes) then the possible values of the rv are relatively small also. Suppose $Y_n \sim$ binomial(n, p_n) and $np_n \to \lambda > 0$, as $n \to \infty$. From Ex. 2.5 we have

$$M_{Y_n}(t) = (p_n \mathrm{e}^t + 1 - p_n)^n = \left(1 + \frac{np_n(\mathrm{e}^t - 1)}{n}\right)^n \to \mathrm{e}^{\lambda(\mathrm{e}^t - 1)}.$$

We recognize the limit as the mgf for a Poisson(λ) distribution (Ex. 2.17). Therefore, Y_n converges in distribution to the Poisson(λ) distribution. In particular, this means that any probabilities we desire for Y_n may be approximated by a corresponding Poisson probability.

For example, suppose the incidence of fatality from a smallpox vaccination is known to be 1 in 5000. The government starts vaccinating people and among the first $10{,}000$ there are 5 fatalities when only 2 were expected. Should you be concerned?

Solution The question essentially is, "Is seeing 5 (or more) fatalities quite unexpected if the chance is really only $.0002$?"

Let Y_n be the (binomial) rv for the number of fatalities in a random sample of $n = 10{,}000$ and let V be a Poisson rv with $\lambda = np = 2$. Recall (Ex. 2.17) that $\mathsf{P}(V = v) = \frac{\lambda^v \mathrm{e}^{-\lambda}}{v!}$. We find

$$\begin{aligned} \mathsf{P}(Y_n \geq 5) &\doteq \mathsf{P}(V \geq 5) = 1 - \mathsf{P}(V \leq 4) \\ &= 1 - \mathrm{e}^{-2}\left(1 + \frac{2}{1} + \frac{2^2}{2} + \frac{2^3}{6} + \frac{2^4}{24}\right) \doteq .05265. \end{aligned}$$

The public officials can debate whether this makes the event "quite unexpected".

**** Convergence in distribution means only that probabilities converge. The random variables themselves generally do not. In Ex. 2.18, the sequence Y_n does not converge to a Poisson rv.*

Example 2.19 (Normal Approximation to Symmetric Binomial) Suppose $Y_n \sim$ binomial$(n, 1/2)$ (such as the number of Heads when flipping a fair coin).

We know that Y_n has mean $\mu_n = np = n/2$ and standard deviation $\sigma_n = \sqrt{np(1-p)} = \sqrt{n}/2$. We also know the mgf is

$$M_{Y_n}(t) = (p\mathrm{e}^t + 1 - p)^n = \left(\frac{1+\mathrm{e}^t}{2}\right)^n.$$

Consider the standardized variable $Z_n = \frac{Y_n - \mu_n}{\sigma_n}$ which has mean $= 0$ and variance $= 1$.

By Thm. 2.28 this has mgf

$$\begin{aligned} M_{Z_n}(t) &= \mathrm{e}^{-\mu_n t/\sigma_n} M_{Y_n}(t/\sigma_n) = e^{-\sqrt{n}t}\left(\frac{1+\mathrm{e}^{2t/\sqrt{n}}}{2}\right)^n \\ &= \left(\frac{\mathrm{e}^{t/\sqrt{n}} + \mathrm{e}^{-t/\sqrt{n}}}{2}\right)^n = \left(1 + \frac{t^2}{2n} + \frac{t^4}{24n^2} + \cdots\right)^n, \end{aligned}$$

where we again use Taylor's series for e^x (twice) and find that the odd order terms drop out.

Now we apply Thm. 2.33 again to find $M_{Z_n}(t) \to \mathrm{e}^{t^2/2}$, as $n \to \infty$. That is, Z_n converges to the standard normal in distribution.

John Kerrick was a (bored) war prisoner during World War II. He flipped a coin $10{,}000$ times and observed $5{,}065$ Heads. Can we say the coin was unfair?

Solution The number of Heads was off by 65 from what was expected. What is the chance of this happening, assuming the coin was fair?

Specifically, we want

$$\mathsf{P}(|Y_n - \mu_n| \geq 65) = \mathsf{P}(|Z_n| \geq \frac{2 \times 65}{\sqrt{10,000}} = 1.30) \doteq .19360$$

(from a table of normal probabilities). Kerrick's results were not so unlikely.

2.7 Quantiles

Besides moments such as the mean and variance, quantiles (percentiles) are measures that are often used. They also have some practical advantages over moments. For example, quantiles are insensitive to the most extreme values and they may in fact be more relevant, as in identifying the value that is exceeded only 5% of the time. Quantiles are indispensable in statistical inference, as well.

DEFINITION 2.34 Let X be a rv and let $0 < p < 1$.

i. A p-th quantile ($100p$-th percentile) of X is any value x_p satisfying

$$\mathsf{P}(X < x_p) \le p \le \mathsf{P}(X \le x_p).$$

ii. The median of X ($\operatorname{med}(X)$) is its 0.50-th quantile (50-th percentile).

iii. The quantile function for X is $Q_X(p) = \min\{x : F_X(x) \ge p\}$. (That is, $Q_X(p)$ is the smallest value that is a p-th quantile.) Thus,

$$Q_X(p) \le x \iff p \le F_X(x).$$

**** Note the terminology difference between quantile and percentile.*

The slightly awkward definition of Q_X above accounts for all distributions, including those with jumps (such as discrete cdfs) and cdfs that have flat regions corresponding to zero probability. Continuous 1-1 cdfs are the easiest to handle.

THEOREM 2.35 Let X have cdf F_X and p-th quantile x_p.

i. If F_X is continuous at x_p then $F_X(x_p) = p$.

ii. If F_X is strictly increasing at x_p then x_p is the unique p-th quantile.

In particular, if F_X is continuous and strictly increasing then its inverse F_X^{-1} is well-defined and $Q_X(p) = F_X^{-1}(p)$ is the unique p-th quantile.

PROOF

i. Continuity implies $F_X(x_p) = \mathsf{P}(X \le x_p) = \mathsf{P}(X < x_p)$. So their common value must be p.

ii. If F_X is strictly increasing at x_p then for any $x < x_p$ we have $\mathsf{P}(X \le x) < \mathsf{P}(X < x_p) \le p$ and for any $x > x_p$ we have $p \le \mathsf{P}(X \le p) < \mathsf{P}(X < x)$. In either case, x cannot be a p-th quantile.

□

Note that if F_X has a pdf f_X then F_X is continuous and it is strictly increasing wherever f_X is positive.

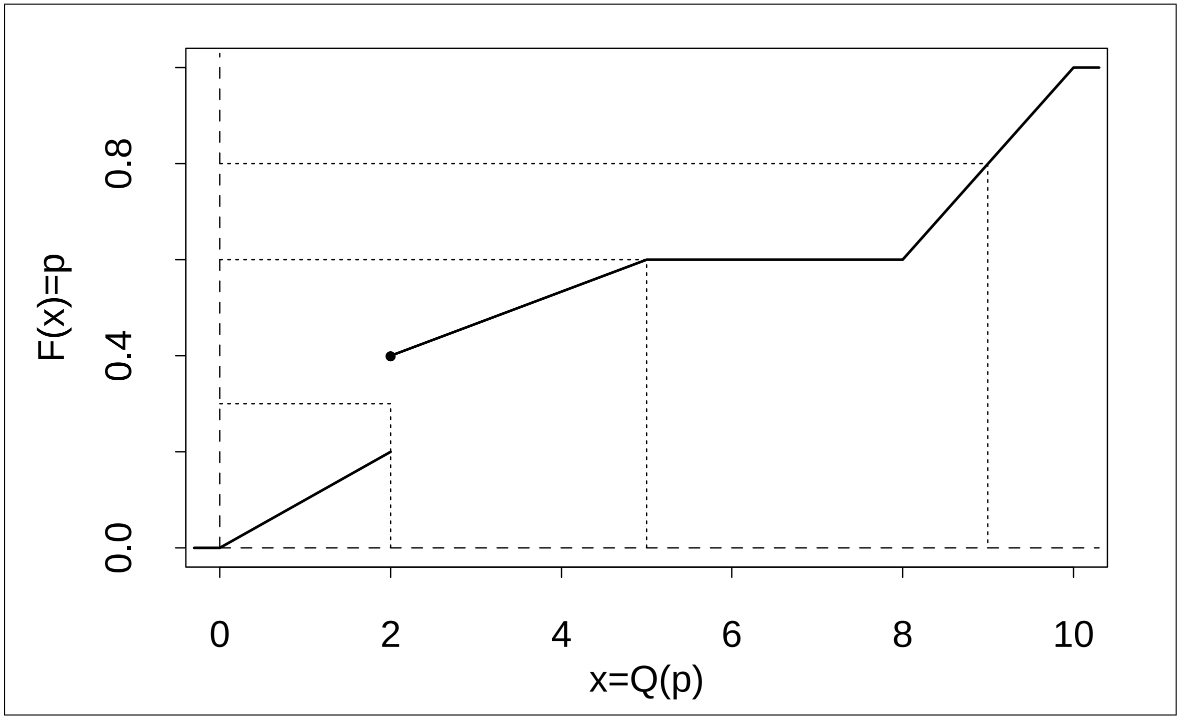

Figure 2.10 The relationship between a cumulative distribution function and its quantile function.

Example 2.20 Suppose $U \sim \text{uniform}(0,1)$ and $V = a + (b-a)U$, $b > a$. Note that $a \le V \le b$.

We can find

$$F_V(v) = \mathsf{P}(V \le v) = \mathsf{P}\Big(U \le \frac{v-a}{b-a}\Big) = F_U\Big(\frac{v-a}{b-a}\Big) = \frac{v-a}{b-a},$$

if $a \le v \le b$.

F_V is continuous and strictly increasing, so the p-th quantile of V satisfies $p = \frac{v_p - a}{b-a}$, or $v_p = a + (b-a)p$. In particular, the median is $v_{.50} = \frac{a+b}{2}$.

Example 2.21 Let $X \sim \text{exponential}(1)$ and $W = (\beta X)^{1/\gamma}$, with $\beta > 0$, $\gamma > 0$. Then W has the Weibull(γ, β) distribution and $F_W(w) = F_X(w^\gamma/\beta) = 1 - \mathrm{e}^{-w^\gamma/\beta}$ for $w > 0$. (What is the pdf for W?)

Solving $F_W(w_p) = p$, the p-th quantile is given by

$$w_p = (-\beta \log(1-p))^{1/\gamma}.$$

The median is $(\beta \log 2)^{1/\gamma}$.

Note that the p-th quantile of X is $-\log(1-p)$. (Take $\beta = \gamma = 1$.) The quantiles of X and W *have the same relationship* as do the rvs themselves.

THEOREM 2.36 (Properties of Quantiles)

i. For $a > 0$ and any $p \in (0,1)$, $Q_{aX+b}(p) = aQ_X(p) + b$. (linearity)

ii. Let h be 1-1 and *increasing*. Then, $Q_{h(X)}(p) = h(Q_X(p))$ for any $p \in (0,1)$. (monotone invariance)

PROOF Fix $0 < p < 1$ and let $x_p = Q_X(p)$.

i. This is a special case of ii, with $h(x) = ax + b$.

ii. Let $Y = h(X)$ and $y = h(x_p)$. Then, since h is 1-1 and increasing,

$$\mathsf{P}(Y < y) = \mathsf{P}(h(X) < h(x_p)) = \mathsf{P}(X < x_p) \leq p$$

and $\mathsf{P}(Y \leq y) = \mathsf{P}(X \leq x_p) \geq p$. So y must be a p-th quantile of Y. Furthermore, $y = h(x_p)$ is the smallest value that gives $\mathsf{P}(Y \leq y) \geq p$. Thus, $y = Q_{h(X)}(p)$.

□

What happens if $h(x)$ is *decreasing* instead? (exercise)

**** Invariance is not a property enjoyed by expectations.*

For example, if X is a nonnegative rv, then $Q_{X^m}(p) = (Q_X(p))^m$ but $\mathsf{E}(X^m) \neq (\mathsf{E}(X))^m$ for $m \neq 1$ (unless X is degenerate; recall Jensen's inequality, Thm. 2.25). In fact, $\mathsf{E}(X^2) - (\mathsf{E}(X))^2 = \mathsf{var}(X) > 0$ so $\mathsf{E}(X^2) > (\mathsf{E}(X))^2$. Likewise, if $X = \log(Y)$ then $\mathrm{med}(X) = \log(\mathrm{med}(Y))$ but $\mathsf{E}(X) < \log(\mathsf{E}(Y))$.

An especially useful property is the following. This is the generalization of Thm. 2.13 mentioned earlier.

THEOREM 2.37 Suppose F is a cdf with quantile function Q. If $U \sim \mathsf{uniform}(0,1)$ and $X = Q(U)$ then $X \sim F$.

PROOF By Def. 2.34, $Q(u) = \min\{x : F(x) \geq u\}$, so

$$Q(u) \leq x \iff u \leq F(x).$$

Thus, $\mathsf{P}(Q(U) \leq x) = \mathsf{P}(U \leq F(x)) = F(x)$. □

Thm. 2.37 shows that random variables can be represented in terms of "simple" rvs like the uniform. This can make proofs and other computations easier. It also provides a way to simulate rvs.

Example 2.21 (cont.) Let $W \sim$ Weibull(γ, β) with cdf $F_W(w) = 1 - \mathrm{e}^{-w^\gamma/\beta}$ for $w > 0$. We found the quantile function for W above:

$$Q_W(p) = (-\beta \log(1-p))^{1/\gamma}.$$

Therefore, if $U \sim$ uniform$(0,1)$ then $W^* = Q_W(U) = (-\beta \log(1-U))^{1/\gamma}$ has the same distribution as W.

Example 2.22 (Simulating a discrete rv) Assume X is nonnegative integer valued with cdf F_X and pmf f_X. Define $p_n = F_X(n) = \sum_{i=0}^n f_X(i)$ and $p_{-1} = 0$. Then

$$p_{n-1} < u \le p_n \iff Q_X(u) = n.$$

To simulate X, first simulate $U \sim$ uniform$(0,1)$ and then assign X the value n if $p_{n-1} < U \le p_n$. (Check: $\mathsf{P}(X = n) = \mathsf{P}(p_{n-1} < U \le p_n) = p_n - p_{n-1} = f_X(n)$.)