CALIFORNIA INSTITUTE OF TECHNOLOGY

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

Chapter 5: Bayes Law (cont.); Random variables and expectation

Relevant textbook passages:

Pitman [5]: Sections 3.1–3.2

Larsen–Marx [4]: Sections 3.3–3.5

5.1 Two-stage experiments

Bayes' Law is useful for dealing with two-stage experiments.

5.1.1 Example (Guessing urns) There are n urns filled with black and white balls. (Figure 5.1 shows what I picture of when hear the word "urn.") Let f_i be the fraction of white balls in urn i. In stage 1 an urn is chosen at random (each urn has probability $1/n$). In stage 2 a ball is drawn at random from the urn. Thus the sample space is $S = \{1, \dots, n\} \times \{B, W\}$. Let $\mathcal{E}$ be the set of all subsets of S.

Pitman [5]: § 1.5

Suppose a white ball is drawn from the chosen urn. What can we say about the event that Urn i was chosen. (This is the subset $E = \{(i, W), (i, B)\}$.) According to Bayes' Rules this is:

$$P(i \mid W) = \frac{P(W \mid i)P(i)}{P(W \mid 1)P(1) + \cdots + P(W \mid n)P(n)} = \frac{f_i}{f_1 + \cdots + f_n}.$$

(Note that $P(W \mid i) = f_i$.) □

Sometimes, in a multi-stage experiment, such as in the urn problem, it is easier to specify conditional probabilities than the probabilities of every point in the sample space. A **tree** diagram is then useful for describing the probability space. **Read section 1.6 in Pitman [5].**

5.1.2 Example (A numerical example)

For concreteness say there are two urns and urn 1 has 10 white and 5 black balls ($f_1 = 10/15$), and urn 2 has 3 white and 12 black balls ($f_2 = 3/15$). (It's easier to leave these over a common denominator.) Each Urn is equally likely to be selected. Figure 5.2 gives a tree diagram for this example.

Then

$$P(\text{Urn } 1 \mid W) = \frac{\frac{10}{15} \cdot \frac{1}{2}}{\frac{10}{15} \cdot \frac{1}{2} + \frac{3}{15} \cdot \frac{1}{2}} = \frac{10}{13}.$$

Figure 5.1. The archetypal urn.

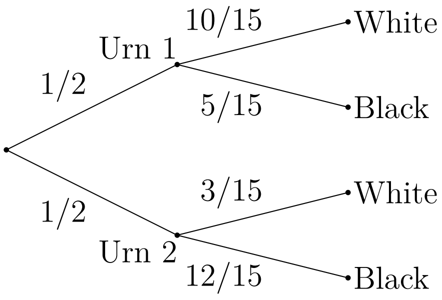

Figure 5.2. Tree diagram for urn selection

and

$$P(\text{Urn }1 \,\big|\, B) = \frac{\frac{5}{15}\cdot\frac{1}{2}}{\frac{5}{15}\cdot\frac{1}{2}+\frac{12}{15}\cdot\frac{1}{2}} = \frac{5}{17}.$$

Sometimes it easier to think in terms of posterior odds. Recall (Section 4.6) that the **posterior odds** against B given A are

$$\frac{P(B^c \,\big|\, A)}{P(B \,\big|\, A)} = \frac{P(A \,\big|\, B^c)}{P(A \,\big|\, B)}\,\frac{P(B^c)}{P(B)}.$$

Letting B be the event that Urn 1 was chosen and A be the event that a White ball was drawn we have

$$\frac{P(\text{Urn }2 \,\big|\, \text{White})}{P(\text{Urn }1 \,\big|\, \text{White})} = \frac{P(\text{White} \,\big|\, \text{Urn }2)}{P(\text{White} \,\big|\, \text{Urn }1)}\,\frac{P(\text{Urn }2)}{P(\text{Urn }1)} = \frac{\frac{3}{15}}{\frac{10}{15}}\,\frac{\frac{1}{2}}{\frac{1}{2}}.$$

The odds against Urn 1 are $3:10$, so the probability is $10/(3+10) = 10/13$.

□

5.1.1 A shortcut for a special case

In urn problems like this, if

- each Urn is equally likely to be drawn, and
- each has the same total number of balls,

then each ball is equally likely to be drawn regardless of its urn.

- Let w_i denote the number of White balls in Urn i, and

$$W = \sum_i w_i$$

be the total number of white balls.

- In this special case,

$$\boxed{P(\text{Urn i} \,\big|\, \text{White ball}) = \frac{w_i}{W}.}$$

To see this, denote the events $B_i = \{\text{Urn } i\}$ and $A = \{\text{White ball}\}$. Bayes Rule for Partitions, Theorem 4.6.3, states

$$P(B_i \,\big|\, A) = \frac{P(A \,\big|\, B_i)P(B_i)}{P(A \,\big|\, B_1)P(B_1) + \cdots + P(A \,\big|\, B_n)P(B_n)}.$$

Since each $P(B_i)$ is the same, they cancel, and each $P(B_i|A) = w_i/m$, where m is the (unspecified, but common) number of balls per urn, and it cancels out, too.

5.2 The Monty Hall Problem

When I was young there was a very popular TV show called *Let's Make A Deal*, hosted by one Monty Hall, hereinafter referred to as MH. At the end of every show, a contestant was offered the choice of a prize behind one of three numbered doors. Behind one of the doors was a very nice prize, often a new car, and behind the other two doors were booby prizes, often a goat. Once the contestant had made his or her selection, MH would often open one of the *other* doors to reveal a goat. (He always knew where the car was.) He would then try to buy back the door selected by the contestant. On one occasion, the contestant asked to trade for the unopened door. *What is the probability that the car is behind the unopened door?*

A popular, but incorrect answer to this question runs like this. Since there are two doors left, and since we know that there is always a goat to show, the opening of the door with the goat conveys no information as to the whereabouts of the car, so the probability of the car being behind either of the two remaining doors must each be one-half. This is wrong, even though intelligent people have argued at great length that it is correct. (See the Wikipedia article, which claims that even Paul Erdös believed it was fifty-fifty until he was shown simulations.)

To answer this question, we must first carefully describe the random experiment, which is a little ambiguous. This has two parts, one is to describe the sample space, and the other is to describe MH's decision rule, which governs the probability distribution. I claim that by rearranging the numbers we may safely assume that the contestant has chosen door number 1. [1] We now assume that the car has been placed at random, so it's equally likely to be behind each door. We now make the following assumption on MH's behavior (which seems to be born out by the history of the show): We assume that MH will always reveal a goat (and never the car), and that if he has a choice of doors to reveal a goat, he chooses randomly between them with equal probability.

The sample space consists of ordered pairs, the first component representing where the car is, and the second component which door MH opened. Since we have assumed that the contestant holds door 1, if the car is behind door 3, then MH *must* open door 2; if the car is behind door 2, then MH *must* open door 3; and if the car is behind door number 1, then MH is free to randomize between doors 2 and 3 with equal probability. Thus the sample space and probabilities are

$$S = \{\underbrace{(1,2)}_{\frac{1}{6}}, \underbrace{(1,3)}_{\frac{1}{6}}, \underbrace{(2,3)}_{\frac{1}{3}}, \underbrace{(3,2)}_{\frac{1}{3}}\}.$$

Suppose now that MH opens door 3. This corresponds to the event

$$\text{MH opens } 3 = \{(1,3),(2,3)\}$$

which has probability $\frac{1}{6} + \frac{1}{3} = \frac{1}{2}$. The event that the car is behind the unopened door (door 2) is the event

$$\text{car behind } 2 = \{(2,3)\},$$

which has probability $\frac{1}{3}$. Now the conditional probability that the car is behind door 2 given that MH opens door 3 is given by

$$\begin{aligned} P(\text{car behind } 2 \,\big|\, \text{MH opens } 3) &= \frac{P(\text{car behind 2 and MH opens 3})}{P(\text{MH opens 3})} \\ &= \frac{P\{(2,3)\}}{P\{(1,3),(2,3)\}} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}. \end{aligned}$$

A similar argument shows that $P(\text{car behind } 3 \,\big|\, \text{MH opens } 2) = 2/3$.

While it is true that MH is certain to reveal a goat, the number of the door that is opened to reveal a goat sheds light on where the car is. To understand this

[1] The numbers on the doors are irrelevant. The doors do not even have to have numbers, they are only used so we can refer to them in a convenient fashion. That, and the fact that they did have numbers and MH referred to them as “Door number one,” “Door number two,” and “Door number three.”

a bit more fully, the consider a different behavioral rule for MH: open the highest numbered door (of door 2 and 3) that has a goat. The difference between this and the previous rule is that if the cars is behind 1, then MH will always open door 3. The only time he opens door 2 is when the car is behind door 3. The new probability space is:

$$S = \{\underbrace{(1,2)}_{0}, \underbrace{(1,3)}_{\frac{1}{3}}, \underbrace{(2,3)}_{\frac{1}{3}}, \underbrace{(3,2)}_{\frac{1}{3}}\}.$$

In this case, $P(\text{car behind } 3 \mid \text{MH opens } 2) = 1$, and $P(\text{car behind } 2 \mid \text{MH opens } 3) = 1/2$.

5.3 Random variables

5.3.1 Definition *A* ***random variable*** *on a probability space* $(S, \mathcal{E}, P)$ *is a real-valued function on* S *which has the property that for every interval* $I \subset \boldsymbol{R}$ *the inverse image of* I *is an event.*

Note that when the collection $\mathcal{E}$ of events consists of all subsets of S, then the requirement that inverse images of intervals be events is automatically satisfied.

5.3.2 Remark An interpretation of random variables used by engineers is that they represent *measurements* on the state of a system. See, e.g., Robert Gray [3].

There is another definition of random variable that is quite common, especially in electrical engineering.

5.3.3 Definition (Another kind of random variable) *Given a set* $\mathcal{A}$ *of* symbols *or* letters, *called the* ***alphabet****, a random variable is defined to be a function from* S *into* $\mathcal{A}$*.*

While we could enumerate the symbols in the alphabet and treat the random variable as a real-valued function, the arithmetic operations have no significance: what letter is the sum of the letters A and B?

Traditionally, probabilists and statisticians use upper-case Latin letters near the end of the alphabet to denote random variables. This has confused generations of students, who have trouble thinking of random variables as functions. For the sake of tradition, and so that you get used to it, we follow suit. So a **random variable** X is a function

$$X\colon S \to \boldsymbol{R} \quad \text{such that for each interval } I, \quad \{s \in S : X(s) \in I\} \in \mathcal{E}.$$

We shall adopt the following notational convention, which I refer to as **statistician's notation**, that

$$(X \in I) \text{ means } \{s \in S : X(s) \in I\}.$$

Likewise $(X \leqslant t)$ means $\{s \in S : X(s) \leqslant t\}$, etc.

If E belongs to $\mathcal{E}$, then its **indicator function** $\mathbf{1}_E$, defined by

$$\mathbf{1}_E(s) = \begin{cases} 0 & s \notin E \\ 1 & s \in E, \end{cases}$$

is a random variable.

5.4 The distribution of a random variable

A random variable X on the probability space $(S, \mathcal{E}, P)$ induces a probability measure or distribution on the real line as follows. Given an interval I, we define

$$P_X(I) = P\left(\{s \in S : X(s) \in I\}\right).$$

This gives us probabilities for intervals. We wish to extend this to probabilities of other sets, such as complements of intervals, countable unions of intervals, countable intersections of countable unions of intervals, etc.

It turns out that the probabilities of the intervals pin down the probabilities on a whole σ-algebra of subsets of real numbers, called the **Borel σ-algebra**. This result is known as the **Carathéodory Extension Theorem**, and may be found in many places, such as [1, Chapter 10]. Sets that belong to the Borel σ-algebra are called **Borel sets**. Every interval, every open set, and every closed set belongs to this σ-algebra. In fact, you need to take an advanced analysis class to be able to describe a set that is not a Borel set. (This is beginning to sound like a broken record. Oops! Have you ever even heard a broken record?)

Pitman [5]: § 3.1

5.4.1 Definition *The* ***distribution*** *of the random variable $X : S \to \boldsymbol{R}$ on the probability space $(S, \mathcal{E}, P)$ is the probability measure P_X defined on $\boldsymbol{R}$ (and its Borel σ-algebra) by*

$$P_X(B) = P\left(X \in B\right).$$

The virtue of the distribution is that for many purposes we can ignore the probability space and only worry about the distribution. But be sure to read section 3.1 in Pitman [5], especially p. 146, on the difference between two variables being equal and having the same distribution:

A random variable is a function on a sample space, and a distribution is a probability measure on the real numbers. It is possible for two random variables to be defined on different sample spaces, but still have the same distribution. For example, let X be the indicator that is one if a coin comes up Tails, and Y be the indicator that a die is odd. Assuming both the coin and the die are "fair," X and Y will have the same distribution, namely each is equal to one with probability 1/2 and zero with probability 1/2, but they are clearly different random variables.

5.5 Discrete random variables

A random variable X is **simple** if the range of X is finite. A random variable X is **discrete** if the range X is countable (finite or denumerably infinite).

5.6 The probability mass function

For a discrete random variable, let x belong to the range of X. The **probability mass function** p_X is given by

$$p_X(x) = P\,(X = x)$$

It completely determines the distribution of X.

5.7 The cumulative distribution function

Pitman [5]: § 4.5
Larsen–Marx [4]: p. 127, p. 137

5.7.1 Definition *The **cumulative distribution function** F_X of the random variable X defined on the probability space $(S, \mathcal{E}, P)$ is the function $F_X \colon \boldsymbol{R} \to [0,1]$ defined by*

$$F_X(t) = P\,(X \leqslant t) = P_X(-\infty, t].$$

N.B. Many authors whom I respect, for instance, C. Radikrishna Rao [6], Leo Breiman [2], and most of the French define the cumulative distribution function using the strict inequality $<$ rather than $\leqslant$.

5.7.2 Fact *The cumulative distribution function F_X is a nondecreasing, right continuous function, and satisfies* $\lim_{t\to-\infty} F_X(t) = 0$ *and* $\lim_{t\to\infty} F_X(t) = 1$.

We often write

$$X \sim F$$

to mean that the random variable X has cumulative distribution function F.

5.8 Examples

5.8.1 Bernoulli random variables

The **Bernoulli distribution** is a discrete distribution that generalizes coin tossing. A random variable X with a Bernoulli(p) distribution takes on two values: 1 (“success”) and 0 (“failure”).

The probability mass function is

$$p\,(X = x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0. \end{cases}$$

Its pmf and cdf are not very interesting.

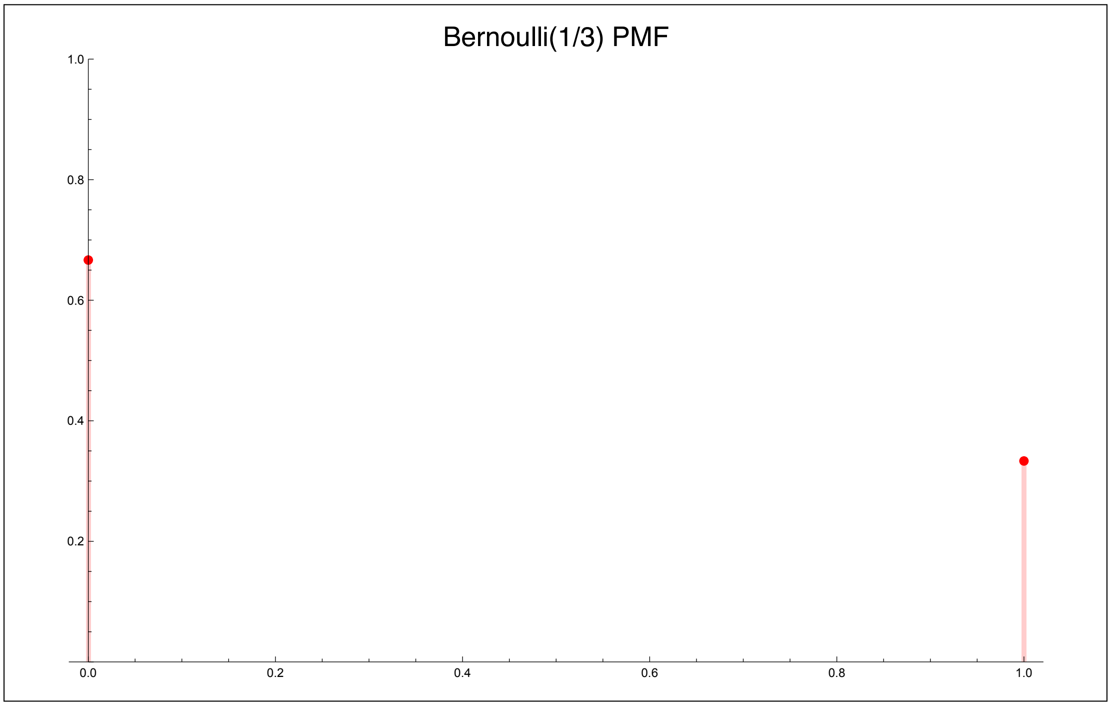

Figure 5.3. The Bernoulli pmf

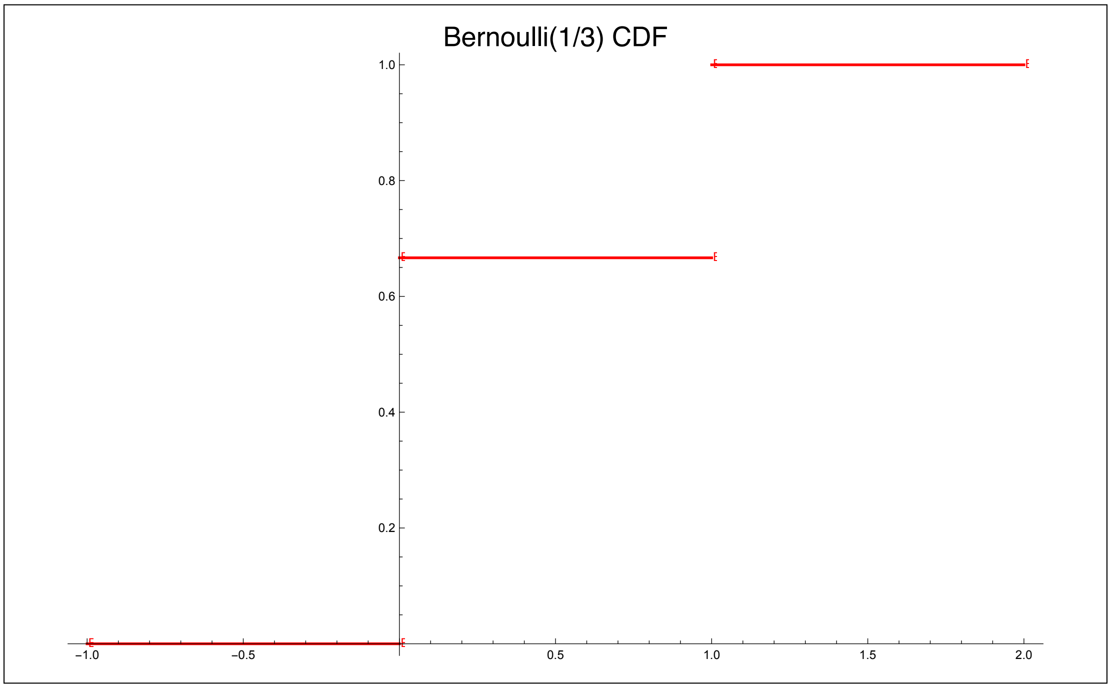

Figure 5.4. The Bernoulli cdf

5.8.2 Binomial random variables

The **Binomial$(\boldsymbol{n}, \boldsymbol{p})$ distribution** is the distribution of the number X of "successes" in n independent Bernoulli(p) trials. The probability mass function is

$$P\left(X = k\right) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, \dots, n.$$

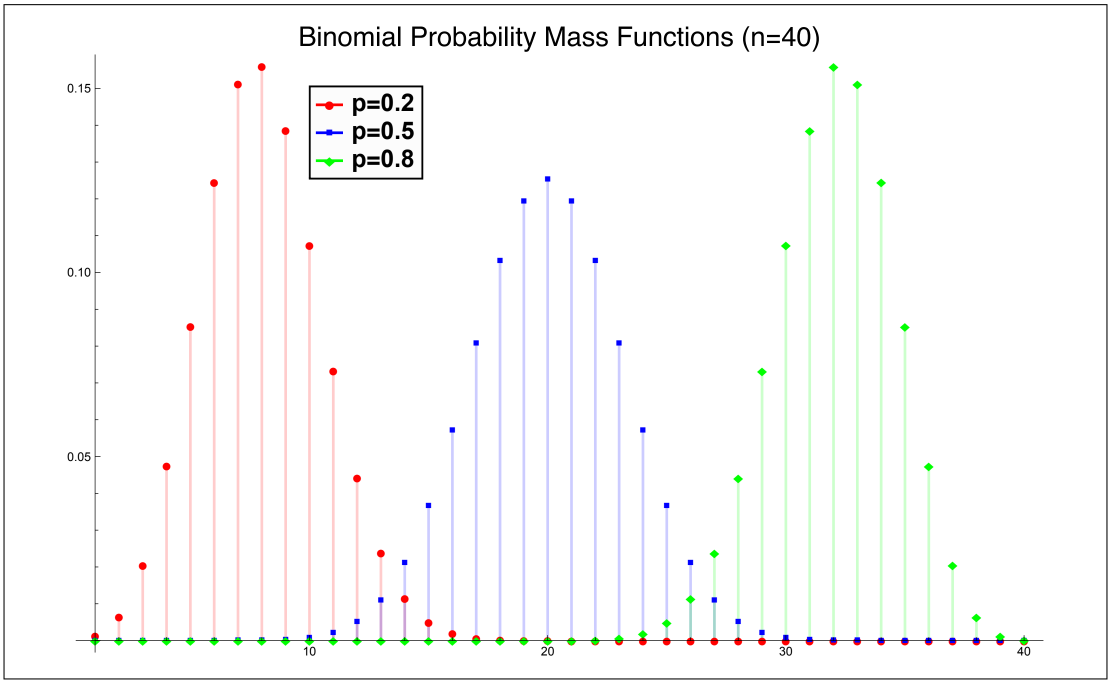

Figure 5.5. Binomial probability mass functions.

Note that the Binomial pmfs are **unimodal**. The **mode** is the value where the pmf assumes its maximum. Here this occurs at $X = pn$. When pn is not an integer, the mode(s) will be adjacent to pn. Note that the pmf for $p = 0.5$ is symmetric about pn, the height of the mode is lower, and the pmf is more "spread out." The pmfs for $p = 0.2$ and $p = 0.8$ are mirror images, which should be obvious from the formula for the pmf.

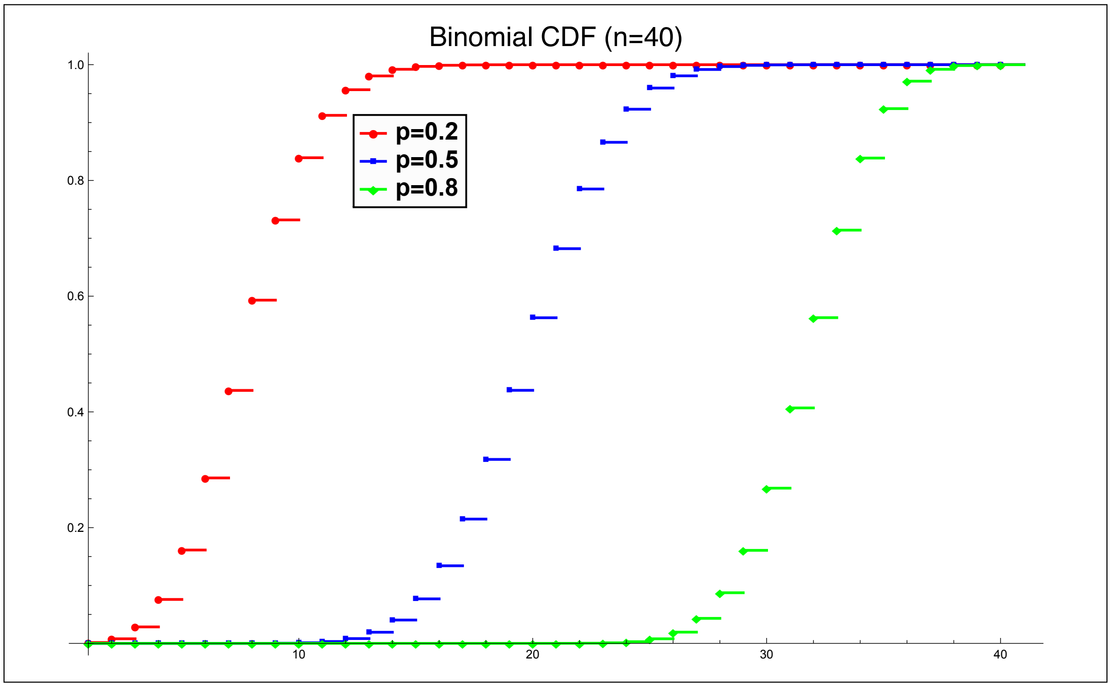

Figure 5.6. Binomial cumulative distribution functions.

5.9 ⋆ Stochastic dominance

Note: This material is in neither Pitman [5] nor Larsen–Marx [4].

Given two random variables X and Y, we say that X **stochastically dominates** Y if for every real number t

$$P(X \geqslant t) \geqslant P(Y \geqslant t),$$

and for some t this holds as a strict inequality. In other words, X stochastically dominates Y if for every t

$$F_X(t) \leqslant F_Y(t),$$

with a strict inequality for at least one t.

If X is the time to failure for one brand of hard drive, and Y is the time to failure for another, which hard drive do you want in your computer?

Note that the Binomial distributions for a fixed n are ordered so that a larger p stochastically dominates a smaller p. See Figure 5.6.

5.10 Expectation

Pitman [5]: § 3.1
Larsen–Marx [4]: § 3.5

The expectation of a random variable is a concept that grew out the study of gambling games. Suppose the sample space for a gambling game is the finite set

$$S = \{s_1, \dots, s_n\},$$

and that the probability of each outcome is given by the probability measure P on S. Suppose further that in outcome $s \in S$, you win $X(s)$. What is a fair price to pay the casino to play this game? What the early probabilists settled on is what we now call the expectation of X.

Pitman [5]: § 3.2

5.10.1 Definition *Let S be a finite or denumerably infinite sample space and let X be a random variable on S. The **expectation**, or **mathematical expectation**, or the **mean** of X is defined to be*

$$\begin{aligned} \boldsymbol{E}\, X &= \sum_{s \in S} X(s)P(s), \\ &= \sum_{x \in \mathit{range}\, X} xp(x) \end{aligned}$$

provided that in case S is infinite, the series is absolutely convergent.

In other words the expectation is a weighted average of the values of X where the weights are the probabilities attached to those values. Note that

the expectation of X is determined by its distribution on $\boldsymbol{R}$.

Why is this considered the "fair price?". For simplicity assume that each outcome is equally likely (e.g., roulette). If we play the game n times and we get each possible out s_i once we shall have won $\sum X(s)$. So the fair price per play should be $\sum X(s)/n - \boldsymbol{E}\, X$.

5.10.2 Remark Here is an interpretation of the expectation that you may find useful. At least it appears in many textbooks.

For a discrete random variable X with values $x_1, x_2, \dots$ imagine the real line as a balance beam with masses $p(x_i)$ placed at x_i for each i. Now place a fulcrum at the position μ. From what I recall of Ph 1a, the total torque on the beam is

$$\sum_k p(x_k)(x_k - \mu).$$

Which value of μ makes the total torque equal to zero? Since $\sum_k p(x_k) = 1$, it is easy to see that

$$\mu = \sum_k p(x_k) x_k$$

is the balancing point. That is, the beam is balanced at the expectation of X. In this sense, the expectation is the **location of the "center" of the distribution**.

Since the torque is also called the **moment** of the forces[2] the expectation is also known as the **first moment** of the random variable's distribution.

It follows that

$$\boxed{\boldsymbol{E}\Big(X - (\boldsymbol{E}\, X)\Big) = 0.}$$

5.10.3 Remark We shall soon see that the expectation is the long run average value of X in independent experiments. This is known as the Law of Large Numbers, or more informally as the Law of Averages.

The expectation of the indicator function $\mathbf{1}_A$ of an event A is $P(A)$.

Interpretations of $\boldsymbol{E}\, X$:

- The "fair price" of a gamble X.
- The location of the center of the distribution of X.
- Long run average value of X in independent experiments.
- The probability of the set indicates by its indicator function.

[2] According to my copy of the *OED* (yes, I have a 28-volume set in my living room), the term "moment" comes from the Latin *momentum*, meaning "movement" or "moving force."

5.11 Expectation of a function of a discrete random variable

If X is a discrete random variable on a probability space $(S, \mathcal{E}, P)$ and g is a function from $\boldsymbol{R}$ to $\boldsymbol{R}$, then the composition $g \circ X$ is also a discrete random variable, and

$$\begin{aligned} \boldsymbol{E}\, g \circ X &= \sum_{s \in S} g\Big(X(s)\Big) P(s), \\ &= \sum_{x \in \operatorname{range} X} g(x) p(x) \end{aligned}$$

provided that in case S is infinite, the series is absolutely convergent.

5.12 The St. Petersburg Paradox

There is a problem with the interpretation of expectation as a fair price.

5.12.1 Example (The St. Petersburg Paradox) (See also Larsen–Marx [4, Example 3.5.5, pp. 144–145].) Consider the following game: Toss a fair coin until the first Tails. If this happens on n^{th} toss, you win 2^n.

What is the expected value of this game?

$$\begin{aligned} \boldsymbol{E}\,\text{Value} &= \sum_{n=1}^{\infty} (\text{winnings if first Tails is on toss } n) \times \text{Prob}\,(\text{first Tails is on toss } n) \\ &= \sum_{n=1}^{\infty} 2^n \frac{1}{2^n} \\ &= \sum_{n=1}^{\infty} 1 \\ &= \infty \text{ (?!)} \end{aligned}$$

So if the expectation is a fair price, you should be willing to pay *any* price to play this game.

But wait! What is the probability that the game stops in a finite number of tosses? It is easier to look at the complementary event. What is the probability of the event A that Tails never occurs? Let A_n be the event that Tails has not occurred in the first n tosses. Then $A \subset A_n$, so $P(A) \leqslant P(A_n) = 1/2^n$. But this is true for every n, so $P(A) = 0$. Therefore with probability 1 the game will end for some finite n, and you will receive $2^n < \infty$.

We shall see later that the reason expectation is not a good measure of "fairness" in this case is that the "Law of Averages" breaks down for random variables that do not have a finite expectation. □

5.12.2 Remark The expected length of a St. Petersburg game is

$$\sum_{k=1}^{\infty} k2^{-k} = 2.$$

For a derivation of the value of the series, see the supplementary notes on series.

Bibliography

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

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

[3] R. M. Gray. 1988. *Probability, random processes, and ergodic properties*. New York: Springer–Verlag.

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

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

[6] C. R. Rao. 1973. *Linear statistical inference and its applications*, 2d. ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.