Lecture 4

The Binomial Distribution

4.1 Introduction

In Lecture 2 we saw that we need to study probability so that we can calculate the 'chance' that our sample leads us to the wrong conclusion about the population. To do this in practice we need to 'model' the process of taking the sample from the population. By 'model' we mean describe the process of taking the sample in terms of the probability of obtaining each possible sample. Since there are many different types of data and many different ways we might collect a sample of data we need lots of different probability models. The Binomial distribution is one such model that turns out to be very useful in many experimental settings.

4.2 An example of the Binomial distribution

Suppose we have a box with a very large number[1] of balls in it: $\frac{2}{3}$ of the balls are black and the rest are red. We draw 5 balls from the box. How many black balls do we get? We can write

$$\text{X} = \text{No. of black balls in 5 draws.}$$

X can take on any of the values 0, 1, 2, 3, 4 and 5.

X is a **discrete random variable**

[1] We say "a very large number" when we want to ignore the change in probability that comes from drawing without replacement. Alternatively, we could have a small number of balls — 2 black and 1 red, for instance — but replace the ball (and mix well!) after each draw.

Some values of X will be more likely to occur than others. Each value of X will have a probability of occurring. What are these probabilities? Consider the probability of obtaining just one black ball, i.e. X = 1.

One possible way of obtaining one black ball is if we observe the pattern BRRRR. The probability of obtaining this pattern is

$$\text{P(BRRRR)} = \tfrac{2}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3}$$

There are 32 possible patterns of black and red balls we might observe. 5 of the patterns contain just one black ball

BBBBB	RBBBB	BRBBB	BBRBB	BBBRB	BBBBR	RRBBB	RBRBB
RBBRB	RBBBR	BRRBB	BRBRB	BRBBR	BBRRB	BBRBR	BBBRR
RRRBB	RRBRB	RRBBR	RBRRB	RBRBR	RBBRR	BRRRB	BRRBR
BRBRR	BBRRR	BRRRR	RBRRR	RRBRR	RRRBR	RRRRB	RRRRR

The other 5 possible combinations all have the same probability so the probability of obtaining one head in 5 coin tosses is

$$\text{P(X = 1)} = 5 \times \left(\tfrac{2}{3} \times (\tfrac{1}{3})^4\right) = 0.0412 \text{ (to 4 decimal places)}$$

What about P(X = 2)? This probability can be written as

$$\begin{aligned} P(X = 2) &= \text{No. of patterns} \times \text{Probability of pattern} \\ &= {}^5C_2 \times \left(\frac{2}{3}\right)^2 \times \left(\frac{1}{3}\right)^3 \\ &= 10 \times \frac{4}{243} \\ &= 0.165 \end{aligned}$$

It's now just a small step to write down a formula for this situation specific situation in which we toss a coin 5 times

$$P(X = x) = {}^5C_x \times \left(\frac{2}{3}\right)^x \times \left(\frac{1}{3}\right)^{(5-x)}$$

We can use this formula to tabulate the probabilities of each possible value of X.

These probabilities are plotted in Figure 4.1 against the values of X. This shows the **distribution** of probabilities across the possible values of X. This

$$
\begin{aligned}
P(X = 0) &= {}^5C_0 \times \left(\tfrac{2}{3}\right)^0 \times \left(\tfrac{1}{3}\right)^5 &= 0.0041 \\
P(X = 1) &= {}^5C_1 \times \left(\tfrac{2}{3}\right)^1 \times \left(\tfrac{1}{3}\right)^4 &= 0.0412 \\
P(X = 2) &= {}^5C_2 \times \left(\tfrac{2}{3}\right)^2 \times \left(\tfrac{1}{3}\right)^3 &= 0.1646 \\
P(X = 3) &= {}^5C_3 \times \left(\tfrac{2}{3}\right)^3 \times \left(\tfrac{1}{3}\right)^2 &= 0.3292 \\
P(X = 4) &= {}^5C_4 \times \left(\tfrac{2}{3}\right)^4 \times \left(\tfrac{1}{3}\right)^1 &= 0.3292 \\
P(X = 5) &= {}^5C_5 \times \left(\tfrac{2}{3}\right)^5 \times \left(\tfrac{1}{3}\right)^0 &= 0.1317
\end{aligned}
$$

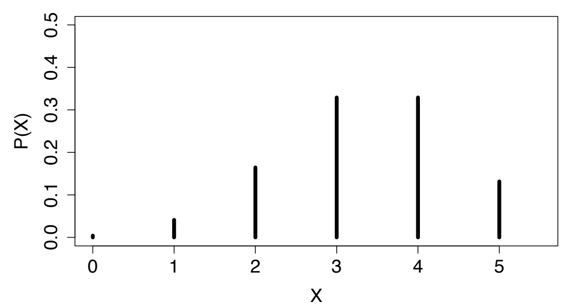

Figure 4.1: A plot of the Binomial(5, 2/3) probabilities.

situation is a specific example of a Binomial distribution.

Note It is important to make a distinction between the probability distribution shown in Figure 4.1 and the histograms of specific datasets seen in Lecture 2. A probability distribution represents the distribution of values we 'expect' to see in a sample. A histogram is used to represent the distribution of values that actually occur in a given sample.

4.3 The Binomial distribution

The key components of a Binomial distribution

In general a Binomial distribution arises when we have the following 4 conditions

- n identical trials, e.g. 5 coin tosses

- 2 possible outcomes for each trial "success" and "failure", e.g. Heads or Tails
- Trials are independent, e.g. each coin toss doesn't affect the others
- P("success") = p is the same for each trial, e.g. P(Black) = 2/3 is the same for each trial

Binomial distribution probabilities

If we have the above 4 conditions then if we let

$$\text{X} = \text{No. of "successes"}$$

then the probability of observing x successes out of n trials is given by

$$\boxed{\text{P(X = x)} \quad = \quad {}^{n}C_{x}p^{x}(1-p)^{(n-x)} \qquad x = 0, 1, \ldots, n}$$

If the probabilities of X are distributed in this way, we write

$$\boxed{\text{X}\sim\text{Bin}(n,\, p)}$$

n and p are called the **parameters** of the distribution. We *say* X follows a binomial distribution with parameters n and p.

Examples

Armed with this general formula we can calculate many different probabilities.

(i). Suppose X ~ Bin(10, 0.4), what is P(X = 7)?

$$\begin{aligned} \text{P(X = 7)} &= {}^{10}C_{7}(0.4)^{7}(1-0.4)^{(10-7)} \\ &= (120)(0.4)^{7}(0.6)^{3} \\ &= 0.0425 \end{aligned}$$

(ii). Suppose Y ~ Bin(8, 0.15), what is P(Y < 3)?

$$\begin{aligned} \text{P(Y < 3)} &= \text{P(Y = 0) + P(Y = 1) + P(Y = 2)} \\ &= {}^{8}C_{0}(0.15)^{0}(0.85)^{8} + {}^{8}C_{1}(0.15)^{1}(0.85)^{7} + {}^{8}C_{2}(0.15)^{2}(0.85)^{6} \\ &= 0.2725 + 0.3847 + 0.2376 \\ &= 0.8948 \end{aligned}$$

(iii). Suppose W $\sim$ Bin(50, 0.12), what is P(W > 2)?

$$\begin{aligned}
P(W > 2) &= P(W = 3) + P(W = 4) + \ldots + P(W = 50) \\
&= 1 - P(W \leq 2) \\
&= 1 - \Big(P(W = 0) + P(W = 1) + P(W = 2)\Big) \\
&= 1 - \Big({}^{50}C_0(0.12)^0(0.88)^{50} + {}^{50}C_1(0.12)^1(0.88)^{49} + {}^{50}C_2(0.12)^2(0.88)^{48}\Big) \\
&= 1 - \Big(0.00168 + 0.01142 + 0.03817\Big) \\
&= 0.94874
\end{aligned}$$

4.4 The mean and variance of the Binomial distribution

Different values of n and p lead to different distributions with different shapes (see Figure 4.2). In Lecture 1 we saw that the mean and standard

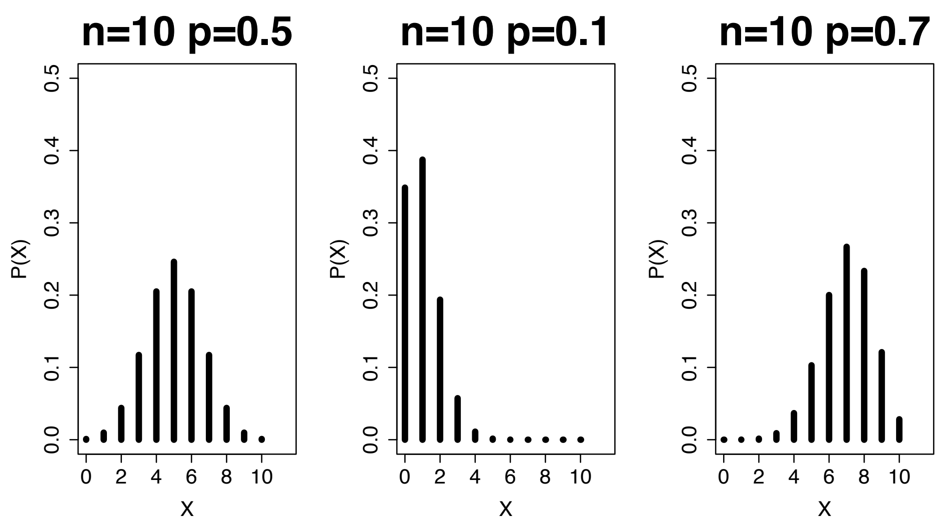

Figure 4.2: 3 different Binomial distributions.

deviation can be used to summarize the shape of a dataset. In the case of a probability distribution we have no data as such so we must use the probabilities to calculate the *expected* mean and standard deviation. In other words, the mean and standard deviation of a random variable is the mean and standard deviation that a collection of data would have if the numbers appeared in exactly the proportions given by the distribution. The mean of a distribution is also called the **expectation** or **expected value** of the distribution.

Consider the example of the Binomial distribution we saw above

x	0	1	2	3	4	5
P(X = x)	0.004	0.041	0.165	0.329	0.329	0.132

The expected mean value of the distribution, denoted μ can be calculated as

$$\begin{aligned}\mu &= 0 \times (0.004) + 1 \times (0.041) + 2 \times (0.165) + 3 \times (0.329) + 4 \times (0.329) + 5 \times (0.132) \\ &= 3.333\end{aligned}$$

In general, there is a formula for the mean of a Binomial distribution. There is also a formula for the standard deviation, σ.

If X $\sim$ Bin(n, p) then

$$\begin{aligned}\mu &= np \\ \sigma &= \sqrt{npq} \quad \text{where } q = 1 - p\end{aligned}$$

In the example above, X $\sim$ Bin(5, 2/3) and so the mean and standard deviation are given by

$$\mu = np = 5 \times (2/3) = 3.333$$

and

$$\sigma = \sqrt{npq} = 5 \times (2/3) \times (1/3) = 1.111$$

Shapes of Binomial distributions

The skewness of a Binomial distribution will also depend upon the values of n and p (see Figure 4.2). In general,

- if $p < 0.5$ the distribution will exhibit POSITIVE SKEW
- if $p = 0.5$ the distribution will be SYMMETRIC
- if $p > 0.5$ the distribution will exhibit NEGATIVE SKEW

However, for a given value of p, the skewness goes down as n increases. All binomial distributions eventually become approximately symmetric for large n. This will be discussed further in Lecture 6.

4.5 Testing a hypothesis using the Binomial distribution

Consider the following simple situation: You have a six-sided die, and you have the impression that it's somehow been weighted so that the number 1 comes up more frequently than it should. How would you decide whether this impression is correct? You could do a careful experiment, where you roll the die 60 times, and count how often the 1 comes up.

Suppose you do the experiment, and the 1 comes up 30 times (and other numbers come up 30 times all together). You might expect the 1 to come up one time in six, so 10 times, so 30 times seems high. But is it too high? There are two possible hypotheses:

(i). The die is biased.

(ii). Just by chance we got more 1's than expected.

How do we decide between these hypotheses? Of course, we can never prove that any sequence of throws *couldn't* have come from a fair die. But we can find that the results we got are extremely unlikely to have arisen from a fair die, so that we should seriously consider whether the alternative might be true.

Since the probability of a 1 on each throw is $1/6$, so we apply the formula for the binomial distribution with $n = 60$ and $p = 1/6$. Then we have

Now we summarise the general approach:

- posit a **hypothesis**
- design and carry out an **experiment** to collect a **sample** of data
- **test** to see if the sample is consistent with the hypothesis

Hypothesis The die is **fair**. All 6 outcomes have the same probability.

Experiment We roll the die.

Sample We obtain 60 outcomes of a die roll.

Testing the hypothesis Assuming our hypothesis is true what is the probability that we would have observed such a sample or a sample more extreme, i.e. is our sample quite unlikely to have occurred under the assumptions of our hypothesis?

Assuming our hypothesis is true the experiment we carried out satisfies the conditions of the Binomial distribution

- n identical trials, i.e. 60 die rolls.
- 2 possible outcomes for each trial: "1" and "not 1".

- Trials are independent.
- P("success") = p is the same for each trial, i.e. P(1 comes up) = 1/6 is the same for each trial

We define X = No. of 1's that come up

We observed X = 30. Which samples are more extreme than this?

Under our hypothesis we would expect X = 10

$X \geq 30$ are the samples as or more extreme than X = 30.

We can calculate each of these probabilities using the Binomial probability formula

$$P(\#\ 1\text{'s is exactly } 30) = ^{60}C_{30} - \left(\frac{1}{6}\right)^{18}\left(\frac{5}{6}\right)^{60-30} = 2.25 \times 10^{-9}.$$

$$P(\#\ 1\text{'s is at least } 30) = \sum_{x=30}^{60} {}^{60}C_x - \left(\frac{1}{6}\right)^{x}\left(\frac{5}{6}\right)^{60-x} = 2.78 \times 10^{-9}.$$

Which is the appropriate probability? The "strange event" from the perspective of the fair die was not that 1 came up exactly 30 times, but that it came up *so many* times. So the relevant number is the second one, which is a little bigger. Still, the probability is less than 3 in a billion. In other words, if you were to perform one of these experiments once a second, continuously, you might expect to see a result this extreme once in 10 years. So you either have to believe that you just happened to get that one in 10 years outcome the one time you tried it, or you have to believe that there really is something biased about the die. In the language of hypothesis testing we say we would 'reject the hypothesis'.

Example 4.1: Analysing the Anturane Reinfarction Trial

From Table 2.1, we know that there were 163 patients who died, out of a total of 1629. Now, suppose the study works as follows: Patients come in the door, we flip a coin, and allocate them to the treatment group if heads comes up, or to the control group if tails comes up. (This isn't exactly how it was done, but close enough. Next term, we'll talk about other ways of getting the same results.)

We had a total of 813 heads out of 1629, which is pretty close to half, which seems reasonable. On the other hand, if we look at the 163 coin flips for the patients who died, we only had 74 heads, which seems pretty far from half (which would be 81.5). It seems there are two plausible hypotheses:

(i). Anturane works. In that case, it's perfectly reasonable to expect that fewer patients who died got the anturane treatment.

(ii). Purely by chance we had fewer heads than expected.

One way of thinking about this formally is to use Bayes' rule:

$$\begin{aligned} P(\text{coin heads} \,|\, \text{patient died}) &= P(\text{coin heads}) \frac{P(\text{patient died} \,|\, \text{coin heads})}{P(\text{patient died})} \\ &= \frac{1}{2} \times \frac{P(\text{patient died} \,|\, \text{anturane treatment})}{P(\text{patient died})}. \end{aligned}$$

If the conditional probability of dying is lowered by anturane, then retrospectively the coin flips for the patients who died have less than probability 1/2 of coming up heads; but if anturane has no effect, then these are fair coin flips, and should have come out 50% heads.

So how do we figure out which possibility is true? Of course, we can never conclusively rule out the possibility of hypothesis 2. Any number of heads is **possible**. But we can say that some numbers are extremely unlikely, indicating that it would be advisable to accept hypothesis 1 — anturane works — rather than believe that we had gotten such an exceptional result from the coin flips.

So... 74 heads in 163 flips is not the most likely outcome. But how unlikely is it? Let's consider three different probabilities:

$$P(\#\text{ heads is exactly }74)$$
$$P(\#\text{ heads is at most }74)$$
$$P(\#\text{ heads is at most }74\text{ or at least }89).$$

Which one is the probability we want? It's pretty clearly not the first one. After all, any particular number of heads is pretty unlikely (and getting exactly 50% heads is impossible, since the number of tosses was odd). And if we had gotten 73 heads, that would have been considered even better evidence for hypothesis 1.

Choosing between the other two probabilities isn't so clear, though. After all, if we want to answer the question "How likely would it be to get such a strange outcome purely by chance?" we probably should consider all outcomes that would have seemed equally "strange", and 89 is as far away from the "expected" number of heads as 74. There isn't a definitive answer to choosing between these "one-tailed" and "two-tailed" tests, but we will have more to say about this later in the course.

Now we compute the probabilities:

$$P(\#\text{ heads is exactly }74) = {}^{163}C_{74}\left(\frac{1}{2}\right)^{74}\left(\frac{1}{2}\right)^{163-74} = 0.0314,$$

$$\begin{aligned}
P(\#\text{ heads is at most }74) &= \sum_{i=0}^{74} {}^{163}C_i\left(\frac{1}{2}\right)^{i}\left(\frac{1}{2}\right)^{163-i}\\
&= {}^{163}C_0\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{163-0} + {}^{163}C_1\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{163-1}\\
&\qquad + \cdots + {}^{163}C_{74}\left(\frac{1}{2}\right)^{74}\left(\frac{1}{2}\right)^{163-74}\\
&= 0.136,
\end{aligned}$$

$$\begin{aligned}
P(\#\text{ heads is at most }74\text{ or at least }89) &= \sum_{i=0}^{74} {}^{163}C_i\left(\frac{1}{2}\right)^{i}\left(\frac{1}{2}\right)^{163-i}\\
&\qquad + \sum_{i=89}^{163} {}^{163}C_i\left(\frac{1}{2}\right)^{i}\left(\frac{1}{2}\right)^{163-i}\\
&= 0.272.
\end{aligned}$$

Note that the "two-tailed" probability is exactly twice the "one-tailed". We show these probabilities on the probability histogram of Figure 4.3.

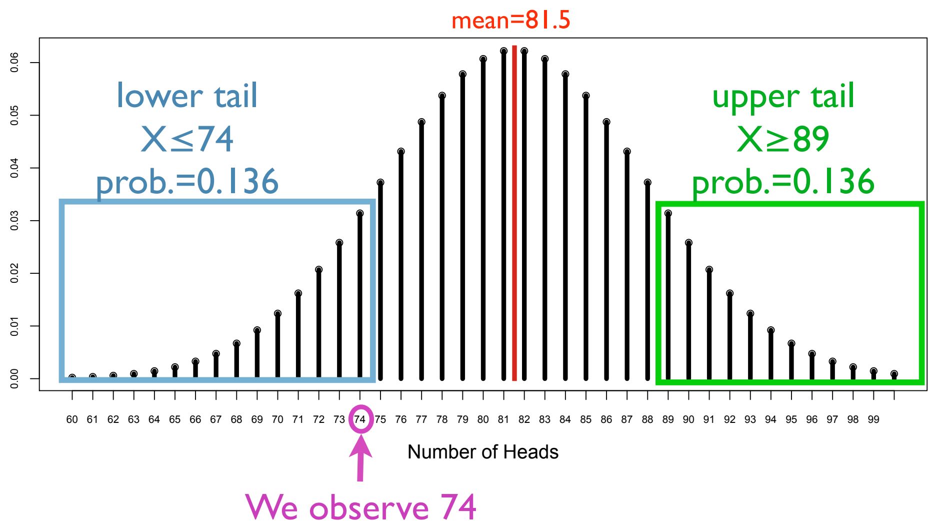

Figure 4.3: The tail probabilities for testing the hypothesis that anturane has no effect.

What is the conclusion? Even 0.136 doesn't seem like such a small probability. This says that we would expect, even if anturane does nothing at all, that we would see such a strong apparent effect about one time out of seven. Our conclusion is that this experiment does not provide significant evidence that anturane is effective. In the language of hypotheses testing we say that we *do not reject* the hypothesis that anturane is ineffective. ■