Lecture 4 : The Binomial Distribution

Jonathan Marchini

October 25, 2004

1 Introduction

In Lecture 3 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.

2 An example of the Binomial distribution

Suppose, we have an unfair coin for which the probability of getting a head is $\frac{2}{3}$ and the probability of a tail is $\frac{1}{3}$. Consider tossing the coin five times in a row and counting the number of times we observe a head. We can denote this number as

$$X = \text{No. of heads in 5 coin tosses}$$

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

X is a **discrete random variable**

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? Lets consider the probability of obtaining just one head in 5 coin tosses, i.e. X = 1.

One possible way of obtaining one head is if we observe the pattern HTTTT. The probability of obtaining this pattern is

$$P(HTTTT) = \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$$

There are 32 possible patterns of heads and tails we might observe. 5 of the patterns contain just one head

HHHHH	THHHH	HTHHH	HHTHH	HHHTH	HHHHT	TTHHH	THTHH
THHTH	THHHT	HTTHH	HTHTH	HTHHT	HHTTH	HHTHT	HHHTT
TTTHH	TTHTH	TTHHT	THTTH	THTHT	THHTT	HTTTH	HTTHT
HTHTT	HHTTT	HTTTT	THTTT	TTHTT	TTTHT	TTTTH	TTTTT

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

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

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.

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

These probabilities are plotted in Figure 1 against the values of X. This shows the **distribution** of probabilities across the possible values of X. This situation is a specific example of a Binomial distribution.

Note It is important to make a distinction between the probability distribution shown in Figure 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.

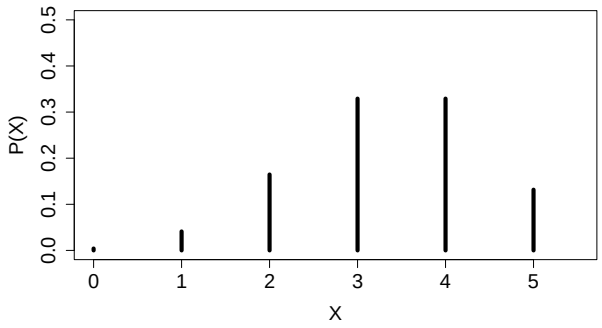


Figure 1: A plot of the Binomial(5, 2/3) probabilities.

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(Head) = 2/3 is the same for each trial

Binomial distribution probabilities

If we have the above 4 conditions then if we let

X = No. of "successes"

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

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

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.

1. Suppose X $\sim$ Bin(10, 0.4), what is P(X = 7)?

$$\begin{aligned} \text{P(X = 7)} &= {}^{10}C_7(0.4)^7(1-0.4)^{(10-7)} \\ &= (120)(0.4)^7(0.6)^3 \\ &= 0.0425 \end{aligned}$$

2. Suppose Y $\sim$ Bin(8, 0.15), what is P(Y < 3)?

$$\begin{aligned} \text{P(Y} < 3) &= \text{P(Y = 0) + P(Y = 1) + P(Y = 2)} \\ &= {}^8C_0(0.15)^0(0.85)^8 + {}^8C_1(0.15)^1(0.85)^7 + {}^8C_2(0.15)^2(0.85)^6 \\ &= 0.2725 + 0.3847 + 0.2376 \\ &= 0.8948 \end{aligned}$$

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

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

The mean and variance of the Binomial distribution

Different values of n and p lead to different distributions with different shapes (see Figure 2). In Lecture 2 we saw that the mean and standard 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. 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.

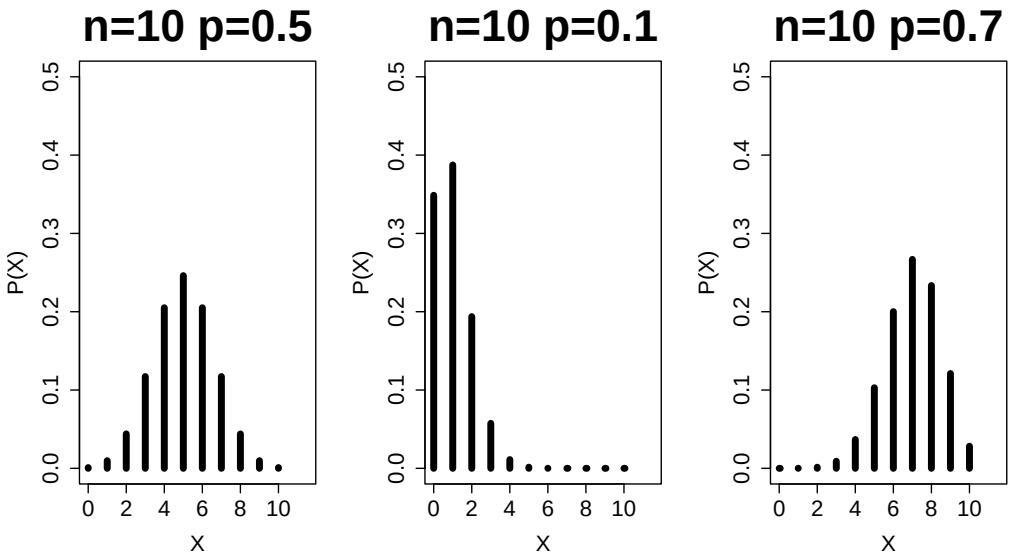


Figure 2: 3 different Binomial distributions.

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 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

4 Testing a hypothesis using the Binomial distribution

Up until know our treatment of the Binomial distribution has been mostly theoretical. To demonstrate its usefulness consider the 'Game show problem' that was included as exercise 8 on Exercise Sheet 2

You have reached the final of a game show. The host shows you 3 doors and tells you that there is a prize behind one of the doors. You pick a door. The host then opens one of the doors you didn't pick that contains no prize and asks you if you want to change from the door you chose to the other remaining door. Should you change?

Intuitively, you might think there would be no advantage to changing doors, i.e. there are two doors to choose from so the probability that one of them is correct is 1/2.

We can test this in a scientific way

The basic idea is to

- 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 probability that you win the prize if you change doors is 1/2.

Experiment To test the hypothesis we could play out the scenario many times and count the number of occasions in which changing your choice would result in you winning the prize.

Sample For example, lets suppose i carry out the experiment 100 times and observe that on 71 occasions i would have won the prize if i'd changed my choice.

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. 100 game shows
- 2 possible outcomes for each trial "success" and "failure", i.e. "Changing doors leads to a WIN" or "Changing doors leads to a LOSS"
- Trials are independent, i.e. each game show is independent

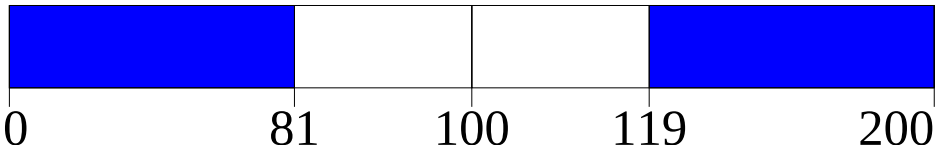
- P(“success”) = p is the same for each trial, i.e. P(Changing doors leads to a WIN) = 1/2 is the same for each trial

We define X = No. of game shows in which changing doors lead to a WIN

We observed X = 71. Which samples are more extreme than this?

Under our hypothesis we would expect X = 50

$X \geq 71$ and $X \leq 29$ are the samples as or more extreme than X = 71. Thus



we want to calculate $P(X \geq 71 \cup X \leq 29)$

We can calculate each of these probabilities using the Binomial probability formula (see the Examples above)
$\Rightarrow$

$$P(X \geq 71 \cup X \leq 29) = 0.00003216$$

This is a very small probability. This tells us that if our hypothesis is true then it is very unlikely that we would have observed 71 out of 100 experiments in which changing doors leads to a WIN. In the language of hypothesis testing ‘we say we would reject the hypothesis’.