Hypothesis Testing and the Binomial Distribution

Prerequisites

You should be familiar with the binomial distribution.

Example (1)

A gardener buys a packet of seeds. It says on the packet that 60% of the seeds are expected to germinate. He plants 25 seeds and finds that only 8 seeds actually germinate.

(*a*) Assuming that the true proportion of seeds germinating from this packet is 0.60, find the probability that when 25 seeds are selected at random only 9 or less seeds actually germinate.

(*b*) Discuss whether your answer would give the gardener sufficient reason to conclude that the claim on the seed packet was false.

Solution

(*a*) Let X denote the number of seeds that germinate.

$X \sim B(25, 0.6)$

$P(X \leq 9) = 0.0132$ [from tables]

(*b*) The probability that 9 or less seeds germinate is 1.32%, which is low. Assuming that the gardener himself is not to blame for the poor rate of germination, this is evidence that there is something wrong with the packet and that it is false to claim that 60% of the seeds in the packet will germinate.

If the proportion of seeds germinating in all packets whatsoever really is 60% then when 25 seeds are taken at random from a packet it is expected that $0.6 \times 25 = 15$ of them on average will germinate. That is the expected mean of the *population*. When 25 seeds are taken at random we can call this a *random sample*. The statement that the expected mean of the sample is 15 is a *hypothesis*. The actual number germinating is a *test result*. On the basis of the *test result* we will decide whether or not the *hypothesis* is true. We have to establish a rule that will act as the criterion for when we will accept that a hypothesis is true or not. The actual result of the test, which provides the test result, is an event. Suppose we establish the criterion that we will reject

the hypothesis if the probability of the event represented by the test result is less than 5%. Then 5% is our *significance level*.

Hypothesis testing in the binomial distribution

In the play *Rosencrantz and Guildenstern are Dead* (by Tom Stoppard), the opening scene depicts two characters, Rosencrantz and Guildenstern, spinning a coin. The coin keeps coming up heads. This raises the possibility that the coin is biased. Suppose we form two hypotheses, which we shall call the *null hypothesis* and the *alternative hypothesis*.

(1) Null hypothesis: the coin is not biased

(2) Alternative hypothesis: The coin is biased.

At this stage the hypotheses are formulated in words. We will need to turn each into precise mathematical statements in order to test them. We need also to adopt a procedure whereby we will decide which hypothesis is true.

Whatever procedure we adopt we will never be able to determine *for certain* whether the hypothesis is true or not. The reason is that even if the coin is a fair coin, there is always a possibility that a fair coin will turn up heads eight times in a row, or fifty times in a row. For example, the probability of a head turning up eight times in a row with a fair coin is $\frac{1}{2^8} = \frac{1}{256} \approx 0.0039$. This is not very likely, but it is possible nonetheless. Thus, whatever procedure we adopt there is always the possibility that we will reject as false a hypothesis that is in fact true.

This reasoning reveals a number of key features about the process of testing this hypothesis.

(1) The hypothesis test is based on an assumption about the general nature of a probability distribution. For example, here, we are assuming that the spinning of any fair coin whatsoever results in a binomial distribution. The entire underlying set from which a sample is drawn is called a *population*, so we are making an assumption about the background distribution of the population. We then take a *sample* drawn from this population. For example, here we might spin the coin 8 times; then the population is the set of spins of any fair coin whatsoever, and the sample is the result of spinning the coin 8 times. Let X stand for the number of successes (here heads) in the sample. We are assuming that the population follows a binomial distribution and this assumption about the background population is *not* tested. On the basis of this assumption we conclude that sample is $X \sim B(n,p)$, where n is the sample size and p is the proportion of outcomes in the entire population that will result in a "success". The symbol p stands for a *population parameter*, which here is the probability of a success; the symbol n stands for the number of trials in the sample.

(2) The hypothesis is specifically a test about the *parameter* of the distribution. Here, we are testing whether the probability of a coin coming up heads is $\frac{1}{2}$. This assumption is called the *null hypothesis.* This is represented by $H_0 : p = \frac{1}{2}$. The symbol H_0 is used to introduce the null hypothesis.

(3) The *null hypothesis* is to be tested against an *alternative hypothesis,* which is denoted by H_1. We shall see below that the alternative hypothesis can be precisely formulated in one of *two* different ways.

(4) In order to test the hypothesis we conduct an experiment in which there are a number of trials. The experiment constitutes a *sample* drawn from the *population* as a whole. The number of trials in the sample is called the *sample size*, and this is what *n* denotes in the expression $X \sim B(n,p)$. In the context of a binomially distributed population the symbol *X* is the variable denoting the number of successes in the sample. The number of successes in the sample is called the *test result*.

(5) In advance of the test we establish a *significance level* or *critical value* that establishes whether we will accept or reject the null hypothesis in favour of the alternative hypothesis. We usually denote this significance level by α. For example, a significance level $\alpha = 0.05$ means that if the probability of the test result occurring is less than 0.05 (5%) then we will reject the null hypothesis. This would also mean that there is a 5% possibility that we will reject as false something that is actually true. This is the inevitable consequence of the need to establish some form of decision criterion.

One-tailed versus two-tailed tests

The alternative hypothesis can take one of two forms – it can be *one-tailed* or *two-tailed*.

(1) A one-tailed test

$$H_0 : \quad p = \frac{1}{2} \qquad \text{or} \qquad H_0 : \quad p = \frac{1}{2}$$
$$H_1 : \quad p > \frac{1}{2} \qquad\qquad H_1 : \quad p < \frac{1}{2}$$

As the above shows a one-tailed test uses a single inequality $(< \text{ or } >)$, and so can take two forms. For example, we could question whether a coin was biased either towards heads **or** tails. Letting *X* denote the number of heads (successes) then if the coin is biased towards heads then the proportion of heads in the sample will be greater than $\frac{1}{2}$. Hence

the alternative hypothesis is $H_1: p > \frac{1}{2}$. On the other hand, we might believe that the coin was biased towards tails, in which case we would be testing whether the probability of a head coming up was $H_1: p < \frac{1}{2}$ and only that. Either case would constitute a one-tailed test.

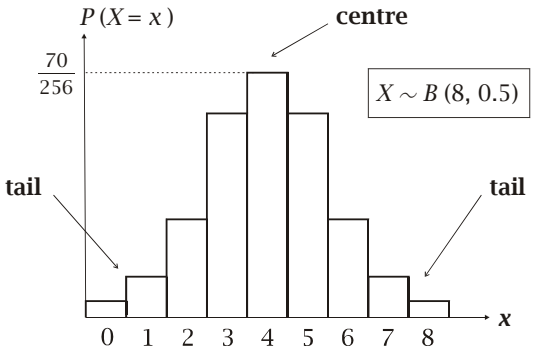
(2) A two-tailed test

$H_0: \quad p = \frac{1}{2}$

$H_1: \quad p \neq \frac{1}{2} \quad \left(\text{That is } \textbf{either } p > \frac{1}{2} \textbf{ or } p < \frac{1}{2}\right)$

In this case the null hypothesis is that the coin is fair $\left(p = \frac{1}{2}\right)$ and the alternative hypothesis is that the coin is biased $\left(p \neq \frac{1}{2}\right)$ but we do not say in advance whether it is biased towards heads or tails. Hence if we reject the null hypothesis this will be because **either** the test result indicates that the coin is biased towards heads $p > \frac{1}{2}$ **or** the test result indicates that the coin is biased towards tails $p < \frac{1}{2}$.

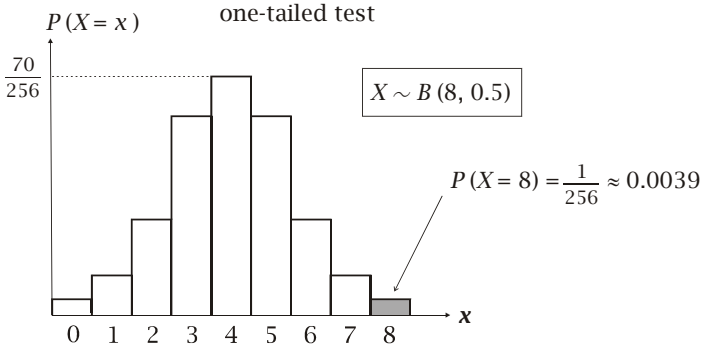
The terms one-tailed and two-tailed derive from the graphical representation of the probability distribution for the sample. The following diagram illustrates how we may think of any binomial distribution $X \sim B(n,p)$ as comprising a central portion and two tails. We show this for the specific case where $n = 8$ and $p = \frac{1}{2}$.



When we are making a one-tailed test our critical region corresponding to the significance level is derived from one only of these two tails. For example in the test

$H_0: p = \frac{1}{2}$ $\qquad$ $H_1: p > \frac{1}{2}$

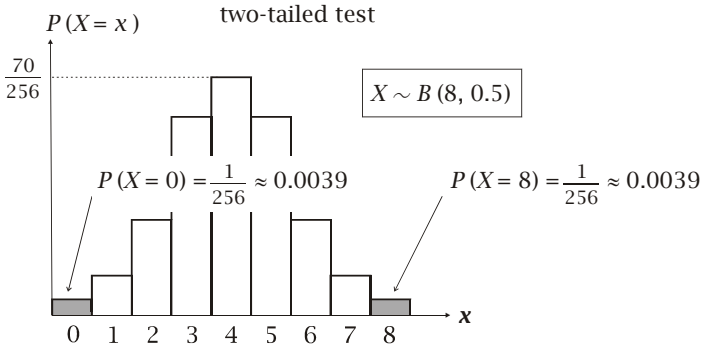
suppose we say that we will reject the null hypothesis $\left(H_0 : p = \frac{1}{2}\right)$ if heads comes up 8 times (8 successes). Then the critical region corresponds to the rectangle representing the probability $P(X = 8)$. This is shown in the next diagram.



Here we have set a very tough test. We will only say that the coin is biased in favour of heads if exactly 8 heads in a row come up. The probability of this event occurring, if the coin is actually fair is $P(X = 8) = \frac{1}{256} = 0.00391\ (3\text{ s.f.})$, which is less than 1%. On the other hand, in a two-tailed test we are looking at *both* tails of the distribution simultaneously. For example, in the test

$H_0 : p = \frac{1}{2}$ $\qquad$ $H_1 : p \neq \frac{1}{2}$ $\left(\textbf{either } p > \frac{1}{2} \textbf{ or } p < \frac{1}{2}\right)$

suppose we say that we will reject the null hypothesis if heads comes up **either** 8 times (8 successes) or 0 times (0 successes). Then the critical region corresponds to the combined areas of both rectangles representing $P(X = 8)$ **and** $P(X = 0)$. The test is two-tailed.



Here we have still set a very tough test. We will say that the coin is biased if either exactly 8 heads in a row come up or 8 tails in a row come up. The probability of this event occurring, if the coin is actually fair is $P(X = 8) + P(X = 0) = 2 \times \frac{1}{256} = 0.00781\ (3\text{ s.f.})$, which is still less than 1%.

The need for both kinds of test (one-tailed and two-tailed) derives from the practical uses of hypothesis testing, for example in science and business. Sometimes in advance of an experiment we have reason to suspect that the outcome will differ from the population mean in a specific direction. This, then, will be a one-tailed test. In our example, we may have reason to suspect in advance of the test that the coin is biased towards heads. Then we shall choose the one-tailed test

$H_0 : p = \frac{1}{2}$ $\qquad$ $H_1 : p > \frac{1}{2}$

A consequence of this choice is that if by chance the coin came up always tails $(X = 0)$ then we would **not** say that the coin was biased because we specifically were testing the alternative hypothesis that the coin was biased in favour of heads (and only heads). On the other hand, in advance of an experiment, we may have no reason to suppose that the outcome is biased one way or the other. So then we choose the two-tailed test

$H_0 : p = \frac{1}{2}$ $\qquad$ $H_1 : p \neq \frac{1}{2}$ $\left(\textbf{either } p > \frac{1}{2} \textbf{ or } p < \frac{1}{2}\right)$

In our example, we will then conclude that the coin is biased if either it comes up all heads $(X = 8)$ or it comes up all tails $(X = 0)$.

Example (2)

Guildenstern suspects that the coin Rosencrantz is spinning is biased in favour of heads. Rosencrantz will only agree if the probability of the test result is less than 1%. In an experiment they agree to spin the coin 8 times and when they do so it comes up heads on all 8 occasions. Should they conclude that the coin is biased?

Solution

(1) We assume that the background distribution is binomial. In other words, when a coin is tossed, the probability of it coming up heads n times is binomially distributed; that is, if X stands for the number of times heads will come up, n for the number of trials, and p for the probability of a success, then $X \sim B(n,p)$ and here $X \sim B(8,p)$

(2) The test result was $X = 8$; this means, that when we spun the coin 8 times, on all 8 occasions there was a "success" (a head).

(3) We are testing whether the coin is biased; so we assume that it is unbiased, which means that the probability of a head is $\frac{1}{2}$. This means that the null hypothesis is

H_0 $\qquad$ $p = \frac{1}{2}$

This entails that the assumption about the background distribution is $X \sim B\left(8, \frac{1}{2}\right)$. The alternative hypothesis is that the coin is biased in favour of heads. This means that we are expecting the coin to turn up more heads than tails (more "successes" than "failures"). So this is a *one-tailed* test; and the alternative hypothesis is

$H_1 \qquad p > \frac{1}{2}$

(4) The significance level in the question is $\alpha = 0.01 = 1\%$

(5) We now ask ourselves the question, **if** the distribution were $X \sim B\left(8, \frac{1}{2}\right)$ what would be the probability of obtaining 8 successes? This asks for $P(X = 8)$. This is given by the binomial distribution as

$$P(X = 8) = \binom{8}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^0 = \frac{1}{256} = 0.0039 \text{ (2 s.f.)}$$

This is the test result. Since this test statistic is less than the significance level, we reject the null hypothesis and accept the alternative hypothesis.

$0.0039 < 0.01$

test result < α

Reject H_0, accept H_1

The coin is biased

In a solution to such a problem it is not normal to explain every step so fully. This is how it could be written.

Solution

$X \sim B(8, p)$

$H_0 \qquad p = \frac{1}{2}$

$H_1 \qquad p > \frac{1}{2}$ one-tailed test

Under H_0 : $X \sim B\left(8, \frac{1}{2}\right)$

$\alpha = 0.01$

Test result $X = 8$

$$P(X = 8) = \binom{8}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^0 = \frac{1}{256} = 0.0039 \text{ (2 s.f.)}$$

$0.0039 < 0.01$

test statistic $< \alpha$

Reject H_0, accept H_1

The coin is biased

Extending the tests to include cumulative probabilities

In the above examples we set a significance level of $\alpha = 0.01$. In practice, because the binomial distribution is a discrete distribution, we had **actual** significance levels corresponding to the test result $X = 8$ in the one-tailed test of $P(X = 8) = 0.0039$ (2 s.f.) and to

$$P(X = 8) + P(X = 0) = 2 \times \frac{1}{256} = 0.0078 \text{ (2 s.f.)}$$

in the two-tailed test. Both were actually less than $\alpha = 0.01$. We may believe that such tests are two stringent and that we would be prepared to conclude that the coin was biased if in fact the outcome was unlikely, but not quite so unlikely. Scientists will very often conclude that a result is significant at a significance level of $\alpha = 0.05$ (5%).

Example (3)

Given $X \sim B\left(8, \frac{1}{2}\right)$ find

(a) $P(X = 8)$

(b) $P(X \leq 7)$

(c) $P(X \leq 6)$

Solution

$$P(X = 8) = \binom{8}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^0 = \frac{1}{256} = 0.0039 \text{(4 d.p.)}$$

$$P(X = 7) = \binom{8}{7}\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^1 = 8 \times \frac{1}{256} = 0.0314 \text{ (4 d.p.)}$$

$$P(X = 6) = \binom{8}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^2 = 28 \times \frac{1}{256} = 0.1094 \text{ (4 d.p.)}$$

(a) $P(X = 8) = 0.0039$ (4 d.p.)

(b) $P(X \leq 7) = P(X = 7) + P(X = 8) = 0.0313 + 0.0039 = 0.0352$ (4 d.p.)

(c) $P(X \leq 6) = P(X = 7) + P(X = 8) + P(X = 6) = 0.0313 + 0.0039 + 0.1094 = 0.1446$ (4 d.p.)

This means that the probability of obtaining either 7 or 8 heads is 0.0352 and the probability of obtaining 6, 7 or 8 heads is 0.1446. These are examples of cumulative probabilities.

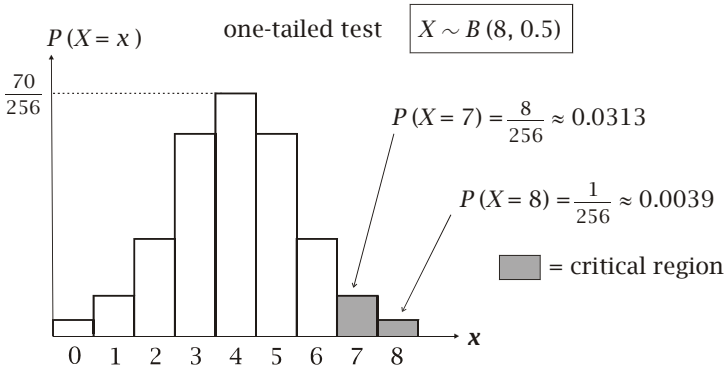
Suppose we suspect the coin of bias in favour of heads, so we are conducting a one tailed test. suppose also that the significance level is $\alpha = 0.05\ (5\%)$.

$X \sim B(8,p)$

$H_0 : p = \frac{1}{2}$ $\quad$ $H_1 : p > \frac{1}{2}$ $\quad$ one-tailed test

Under $H_0 : X \sim B\left(8,\frac{1}{2}\right)$

The cumulative probability of obtaining either a 7 or an 8 is 0.0352, and this is less than the significance level of 0.05. So if we had either a 7 or an 8 as the test result we would reject the null hypothesis and conclude that the coin was biased. Notice this is **either** a 7 or an 8. We do not know in advance of the experiment that the coin will come up 7 times. We would reject the null hypothesis if either result came true. On the other hand, the cumulative probability of a 6, 7 or 8 is 0.1446, which is greater than the significance level of 0.05. Thus, if the coin comes up heads 6 times out of 8, this will not be sufficient reason to believe that it is biased at a significance level of 0.05, and we will continue to accept the null hypothesis.



The actual significance level is less than $\alpha = 0.05\ (5\%)$. It is

$P(X \le 7) = P(X = 7) + P(X = 8) = 0.0313 + 0.0039 = 0.0352\ (4\text{ d.p.})$.

We cannot test the hypothesis to precisely the level $\alpha = 0.05$ because of the discrete nature of the binomial distribution.

Example (4)

A restaurant claims that 70% of its customers agree that their new waffle is "simply delicious". A doubting Thomas, who believed that the proportion was less, asked 12 clients who had eaten the new waffle whether they agreed. Only 4 said that the waffle was "simply delicious". Formulate an appropriate null and alternative hypothesis and test at the 5% significance level whether the restaurant's claim was valid. What is the actual critical value used in the test?

Solution

$H_0 : p = 0.7$

$H_1 : p < 0.7$

$\alpha = 0.05$

This is a one-tailed test.

If H_0 is true then $X \sim B(12, 0.7)$

The probability associated with this level is

p-value $= P(X \le 4) = 0.0095 \quad (2 \text{ d.p.})$ $\quad [\text{From tables}]$

Since p-value $= 0.0095 \ll 0.05 = \alpha$ the result is significant at the 5% level. Therefore

Reject H_0

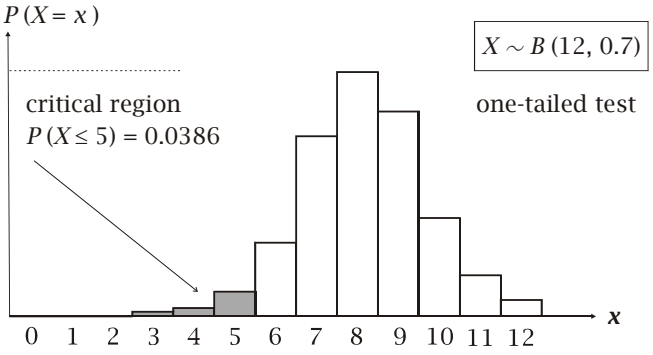
Accept H_1

The claim of the restaurant is false.

From tables of cumulative binomial probabilities we see that

$P(X \le 5) = 0.0386 \quad (4 \text{ d.p.})$

$P(X \le 6) = 0.1178 \quad (4 \text{ d.p.})$



Thus we would have also rejected the claims of the restaurant if 5 or less people had said the waffle was "simply delicious". We would have upheld the restaurant's claim if 6 or

more had said it was "simply delicious". The actual critical region corresponds to a significance level of 0.0386.

In the above example (5) we have introduced the symbol *p*-value to stand for the *probability* associated with the outcome of the experiment. In that example the event $X = 4$ corresponded to the *p*-value $= P(X \le 4) = 0.0095$ (4 d.p.). It is this probability that we compare with the previously determined significance level, here $\alpha = 0.05$. The decision criterion is that we will reject the null hypothesis is the *p*-value is less than the significance level.

Summary

Sample distribution

Suppose $X \sim B(n, p_0)$ where *n* is the sample size and p_0 is the population parameter

Null hypothesis

$H_0: \quad p = p_0$ The sample proportion is the same as that of the population

Alternative hypothesis

One-tailed test $H_1: \quad p > p_0$ **or** $p < p_0$ (That is, one or the other, **not both**)

Two-tailed test $H_1: \quad p \ne p_0$ (That is, **both** $p > p_0$ **or** $p < p_0$)

Test result

$X_{\text{test}} = x$ The number of successes in the sample

p-value The probability of obtaining that number of successes

One-tailed test *p*-value $= P(X \le x)$ **or** *p*-value $= P(X \ge x)$

Two-tailed test *p*-value $= P(X \le x$ **or** $X \ge x)$

Decision process

Significance level α The probability of rejecting H_0 given that H_0 is true.

$$\alpha = P(\text{Reject } H_0 \,|\, H_0 \text{ is true}) = P(\text{Accept } H_1 \,|\, H_0 \text{ is true})$$

Reject H_0 if *p*-value $< a$ otherwise accept H_0. To reject H_0 is to accept H_1 and vice-versa.

Critical region The actual region(s) corresponding to those value(s) of X that would result in the null hypothesis H_0 being rejected.

Critical values The boundary values of the critical region(s).

Example (6)

A gardener buys a packet of seeds. On the packet the producer claims that 65% of the seeds are expected to germinate. The gardener suspects that less than 65% of the seeds germinate. He plants 25 seeds and finds that only 9 seeds actually germinate. For a suitable hypothesis and test at the 5% significance level whether the claim on the packet should be rejected.

Solution

General	**This example**
Sample distribution	
Suppose $X \sim B(n, p_0)$	$X \sim B(25, 0.65)$
Null hypothesis	
$H_0: \quad p = p_0$	$H_0: \quad p = 0.65$
Alternative hypothesis	
One-tailed test	$H_1: p < 0.65$
Test result	
$X_{\text{test}} = x$	$X_{\text{test}} = 9$
p-value	
One-tailed test	p-value $= P(X \le x$ **or** $X \ge x) = P(X \le 9) = 0.0029$
Decision process	
Significance level	$a = 0.05$
	$r_{test} < a$
	Reject H_0. Accept H_1.
	$p \neq 0.65$
	The producer's claim is false.
Critical region	$X \le 11$ with probability $P(X \le 11) = 0.0255$.
Critical value	$X = 11$

Further note about significance level

In practical applications of hypothesis testing is usual to set the significance level in advance. It has already been mentioned that scientists often take $a = 0.05$. Because of the nature of the binomial distribution as a discrete probability distribution it is usually not possible to define the critical region in such a way that the actual significance level is equal to this predetermined value. In that case the **true** significance level is the one defined by the critical region and not the predetermined value. This follows from the definition of the significance level as

$\alpha = P(\text{Reject } H_0 \,|\, H_0 \text{ is true}) = P(\text{Accept } H_1 \,|\, H_0 \text{ is true})$.

This shows that the significance level is the probability of a test result falling in the critical region. However, it will be clear from context what the significance level is. In the last example although we were told to test the result at the 5% significance level, the actual significance level was $\alpha = 0.0255 = 2.55\%$. No practical difference arises from this. The null hypothesis is rejected at both the actual level of $\alpha = 0.0255$ and the predetermined level of $\alpha = 0.05$. When testing in the context of the binomial distribution the instruction, for example, "test at the 5% significance level" means "choose the largest critical region corresponding to a probability less than 5%".

Example (7)

A manufacturer of cereals claims that 35% of their packets of *Breakfast Wheat* contain a toy bear. A consumer protection agency decides to investigate this claim. They purchase 30 packets of *Breakfast Wheat*. They assume that a proportion p of the packets contain a toy bear and establish the hypotheses

$H_0 : p = 0.35$

$H_1 : p \neq 0.35$

They define the critical region to be $X \leq 5$ or $X \geq 16$

(*i*) Calculate the significance level of this procedure.

(*ii*) Calculate the probability of drawing the correct conclusion if the value of p is actually 0.30.

Solution

(*a*) Under $H_0 : X \sim B(30, 0.35)$

$$\begin{aligned}\alpha &= P(X \leq 5 \text{ or } X \geq 16) \\ &= P(X \leq 5) + P(X \geq 16) \\ &= 0.0233 + (1 - P(X \leq 15)) \\ &= 0.0233 + 0.0301 \\ &= 0.0534\end{aligned}$$

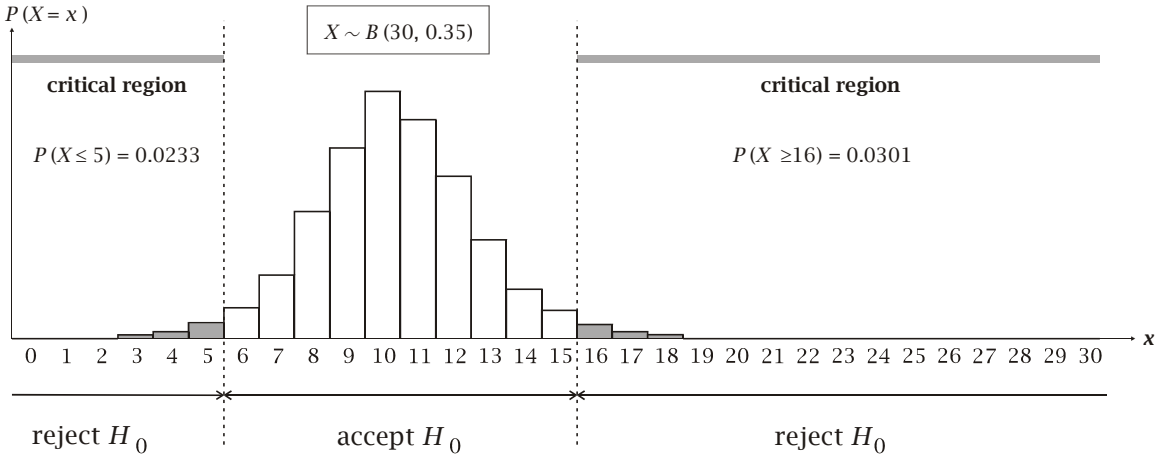
(*b*) Suppose in fact $p = 0.30$. The critical region means that we reject $H_0 : p = 0.35$ if $X \leq 5$ or $X \geq 16$.

Under $X \sim B(30, 0.30)$

$$\begin{aligned}P(X \leq 5 \text{ or } X \geq 16) &= P(X \leq 5) + (1 - P(X \leq 15)) \\ &= 0.0766 + 0.0064 \\ &= 0.083\end{aligned}$$

Thus if in fact $p = 0.30$ the probability of reaching the conclusion that $p \neq 0.35$ is only 0.083 = 8.3%. This is rather small.

The solution to this example (7) illustrates two further points about hypothesis testing under the binomial distribution. Firstly, when we employ a two-tailed test it is usual to try to match the probabilities of events falling into the two halves of the critical region corresponding to the two tails of the distribution.



As the above diagram shows the probabilities of the two critical regions are not in fact equal. Secondly, the answer to part (*b*) illustrates the fact that the test is not very sensitive to the case when the probability differs only slightly from the assumed value of $p = 0.35$. In fact, the true population parameter must differ quite considerably from $p = 0.35$ for this test to stand much chance of rejecting the then false hypothesis that $p = 0.35$. This poses a serious limitation on the use of this test.

Two suggestions for overcoming the second limitation are

(1) Take a larger sample size. As tables tend to give values of binomial distributions up to a sample size of $n = 30$ this will require using a normal approximation to the binomial distribution and extending our concepts of hypothesis testing to cover the normal distribution. The effect of the larger sample size will be to bunch the probabilities tighter about the expected mean, making the test more sensitive to divergences from the mean in the case when the true probability is not equal to the assumed one. However, in real-life applications taking a larger sample size may have cost implications – that is, it is usually more expensive to collect a large sample than a small one.

(2) Increase the size of the critical region. This is equivalent to taking a larger significance level. However, this has the side effect that it increases the likelihood as well of rejecting the null hypothesis even when the null hypothesis is true, for this is what the significance level means.

So choosing a significance level and a sample size is a matter of delicate balancing in order to minimise the possibility of two different types of error.

Type 1 error	The error of rejecting the null hypothesis when it is in fact true. Equal to the significance level.
Type 2 error	The error of accepting the null hypothesis when it is in fact false.

This topic is pursued at a higher level under the heading of *operating characteristics*.