

STRAND: STATISTICS

Unit 6 *Hypothesis Testing*

TEXT

Contents

Section

- | | |
|-----|----------------------|
| 6.1 | Forming a Hypothesis |
| 6.2 | The Sign Test |

6 Hypothesis Testing

6.1 Forming a Hypothesis

We start this section with an experiment to determine whether people really can tell the difference between two similar foods or drinks.

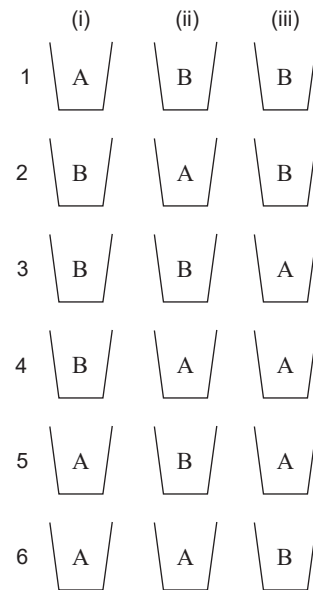
Each person taking part in the test is given 3 samples: two of one product and one of another (so that they may have two cups containing genuine Coca Cola (A) and one cup containing a supermarket brand (B) or vice versa).

Ask each person to identify the sample which is different from the other two.

Note that there are six possible groups of samples and a die can be used to decide which grouping to give to each individual person taking part.

Plan the experiment carefully before you start. Write out a list showing the samples and order of presentation for all your subjects (about 10, say).

Ensure that your subjects take the test individually in quiet surroundings, free from odours. All 3 samples must be of the same size and temperature. If there are any differences in colour you can blindfold your subject.



Record each person's answer. Count the number of subjects giving the correct answer. Subjects who are unable to detect any difference at all in the 3 samples must be left out of the analysis.

In any experiment you usually have your own hypothesis as to how the results will turn out.

However it is usual to set up a **null hypothesis** that states the opposite of what you want to prove. This can only then be abandoned in the face of overwhelming evidence, thus placing the onus of proof on you.

The null hypothesis H_0

For the activity above your null hypothesis is that subjects cannot tell the difference between the 3 samples and that they are guessing.

The alternative hypothesis H_1

This is your experimental hypothesis (or what you really wish to prove). For the activity above your alternative hypothesis is that subjects really can distinguish between the samples (or some of them can at least).

These hypotheses can be written in mathematical terms as :

$$H_0 : p = \frac{1}{3}$$

$$H_1 : p > \frac{1}{3}$$

Here p is the probability of success assuming that H_0 is true; that is, subjects cannot tell the difference and are randomly guessing.

In order to reject H_0 and adopt H_1 , your **experimental** results will have to be ones which are very difficult to explain under the null hypothesis.

So if we assume that H_0 is correct, then the numbers of successful guesses is given by the distribution

$$X \sim B\left(10, \frac{1}{3}\right)$$

You can use this distribution to calculate the probabilities of people achieving various results by guesswork. Here are the probabilities for all the possible results for 10 subjects.

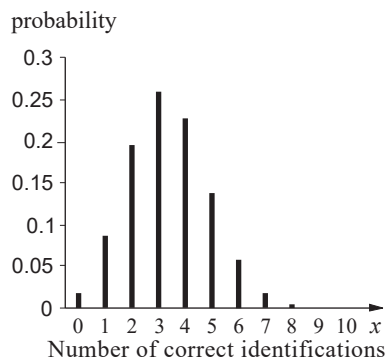
<i>Number of people giving correct answer</i>	<i>Binomial probability</i>
0	$\binom{10}{0} \left(\frac{2}{3}\right)^{10} = 0.0173$
1	$\binom{10}{1} \left(\frac{2}{3}\right)^9 \left(\frac{1}{3}\right) = 0.0867$
2	$\binom{10}{2} \left(\frac{2}{3}\right)^8 \left(\frac{1}{3}\right)^2 = 0.1951$
3	$\binom{10}{3} \left(\frac{2}{3}\right)^7 \left(\frac{1}{3}\right)^3 = 0.2601$
4	$\binom{10}{4} \left(\frac{2}{3}\right)^6 \left(\frac{1}{3}\right)^4 = 0.2276$
5	$\binom{10}{5} \left(\frac{2}{3}\right)^5 \left(\frac{1}{3}\right)^5 = 0.1366$
6	$\binom{10}{6} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^6 = 0.0569$
7	$\binom{10}{7} \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^7 = 0.0163$
8	$\binom{10}{8} \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^8 = 0.0030$
9	$\binom{10}{9} \left(\frac{2}{3}\right) \left(\frac{1}{3}\right)^9 = 0.0003$
10	$\binom{10}{10} \left(\frac{1}{3}\right)^{10} = 0.00002$

As the probability of 10 people guessing correctly is so small, if this actually happened you would be much more inclined to believe that they can actually tell the difference between the samples. So in this case it would be more rational to reject H_0 , because the explanation offered by H_1 is more plausible.

But how many of the other possible results are not easily explained by H_0 (and so better explained by H_1)?

Under H_0 the probability of :

10 correct guesses is	0.00002
9 or 10 correct guesses	0.00032
8, 9 or 10 correct guesses	0.00332
7, 8, 9 or 10 correct guesses	0.01962
6, 7, 8, 9 or 10 correct guesses	0.07652



Note that if you adopt 9 correct as a 'significant' result you must include the probability for 10 as well (because 10 is actually a 'better' result than 9). Similarly with 8 you must include the probabilities for 9 or 10 correct and so on.

In scientific experiments, it is usual to take results with probabilities of 0.05 (5%) or less as convincing evidence for rejecting the null hypothesis. This is called the **significance level** for an experiment.

If 10 subjects take the taste test then you will conclude that they *can* tell the difference between the samples if 7 or more of them make correct identifications.



Worked Example 1

Using the same experiment as described above, if your sample tested was 8, what is the smallest number for which you will reject H_0 (they are guessing) and accept H_1 (they really can distinguish between the samples)?



Solution

Again we use the binomial distribution with $n = 8$ and $p = \frac{1}{3}$ giving the following results.

<i>Number of people giving correct answer</i>	<i>Binomial probability</i>
0	$\binom{8}{0} \left(\frac{2}{3}\right)^8 \approx 0.03902$
1	$\binom{8}{1} \left(\frac{2}{3}\right)^7 \left(\frac{1}{3}\right) \approx 0.03902$
2	$\binom{8}{2} \left(\frac{2}{3}\right)^6 \left(\frac{1}{3}\right)^2 \approx 0.27312$
3	$\binom{8}{2} \left(\frac{2}{3}\right)^5 \left(\frac{1}{3}\right)^3 \approx 0.20485$
4	$\binom{8}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^4 \approx 0.17071$
5	$\binom{8}{5} \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^5 \approx 0.05121$
6	$\binom{8}{6} \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^6 \approx 0.01707$
7	$\binom{8}{7} \left(\frac{2}{3}\right) \left(\frac{1}{3}\right)^7 \approx 0.00244$
8	$\binom{8}{8} \left(\frac{1}{3}\right)^8 \approx 0.00015$

So under the assumption that H_0 is true, the probability of

8 correct guesses is	0.00015	(< 0.05)
7 or 8 correct guesses is	0.00259	(< 0.05)
6, 7 or 8 correct guesses	0.01966	(< 0.05)
5, 6, 7 or 8 correct guesses	0.07087	(> 0.05)

So 6, 7 or 8 are significant, that is

$$p(X \geq 6) = 0.01966 < 0.05$$

whereas

$$p(X \geq 5) = 0.07087 > 0.05$$

Here the smallest significant number is 6. This means that if 6 or 7 or 8 people make a correct identification, we have evidence to reject H_0 and accept H_1 ; that is, that people can distinguish between the brands.



Worked Example 2

An advert stated that "8 out of 10 cats prefer Whiskers cat food",
Is this significant, using a 5% significance level?



Solution

Here the hypotheses are

$$H_0 : \text{no preference} \quad \left(p = \frac{1}{2} \right)$$

$$H_1 : \text{preferred Whiskers} \quad \left(p > \frac{1}{2} \right)$$

assuming that Whiskers is being compared with Brand X .

If we assume that H_0 is true, that is the cats have no preference, then the distribution of the numbers preferring Whiskers is given by

$$X \sim B\left(10, \frac{1}{2}\right)$$

To test if 8 is significant, we need to find

$$\begin{aligned} p(x \geq 8) &= \binom{10}{8} \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2 + \binom{10}{9} \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right) + \binom{10}{10} \left(\frac{1}{2}\right)^{10} \\ &= \left(\frac{1}{2}\right)^{10} \left(\binom{10}{8} + \binom{10}{9} + 1 \right) \\ &= \left(\frac{1}{2}\right)^{10} (45 + 9 + 1) \\ &= \frac{56}{2^{10}} \\ &\approx 0.0547 \end{aligned}$$

Hence $p(x \geq 8) > 0.05$. So 8 out of 10 is *not* significant and we do not have sufficient evidence to reject H_0 ; that is, the cats show no preference.



Note

Although '8 out of 10' feels to be significant, at a 5% significance level, which is saying that you might be wrong 5% of the time (i.e. 1 in 20), it is not significant.

If you were using a 10% significance level, then it would be significant as

$$p(X \geq 8) = 0.0537 < 0.10$$

It should also be noted that there are tables for the Binomial distribution, for example at

<http://filestore.aqa.org.uk/subjects/FORMULAE.PDF> (page 15)

and many modern calculators have these values stored.



Worked Example 3

A coin is tossed 4 times and each time it lands HEADS. Using a 5% significance level, would you conclude that this coin is biased?

The coin is tossed again and it lands HEADS. What do you now conclude?



Solution

Here the hypotheses are

$$H_0 : \text{not biased; } p(\text{HEADS}) = \frac{1}{2}$$

$$H_1 : \text{biased for HEADS; } p(\text{HEADS}) > \frac{1}{2}$$

If we assume H_0 is true, that is, the coin is not biased, then if X is the random variable, the number of HEADS, we expect $X \sim B\left(4, \frac{1}{2}\right)$

$$\begin{aligned} \text{and } p(X=4) &= \left(\frac{1}{2}\right)^4 \\ &= \frac{1}{16} \\ &\approx 0.0625 \end{aligned}$$

So $p(\text{H H H H}) > 0.05$ and we conclude that we do not have enough evidence to reject H_0 ; that is, at 5% significance level, we do not have sufficient evidence to reject H_0 , the coin is not biased.

If we obtain 5 Heads out of 5, then

$$\begin{aligned} p(X=5) &= \left(\frac{1}{2}\right)^5 \\ &= \frac{1}{32} \\ &= 0.03125 \end{aligned}$$

So $p(\text{H H H H H}) < 0.05$ and we have evidence, using 5% significance level, to reject H_0 and accept H_1 ; that is, the coin is biased.



Exercises

1. A woman who claims to be able to tell margarine from butter correctly picks the 'odd' sample out of the 3 presented, for 5 out of 7 trials. Is this sufficient evidence to back up her claim?
2. A company has 40% women employees, yet of the 10 section heads, only 2 are women. Is this evidence of discrimination against women?
3. A subject takes a test for ESP (extra-sensory perception) in which he has to identify the suit of a playing card held by the experimenter. (The experimenter can see the card, but the subject cannot.) For 10 cards he makes 7 correct identifications. Is this evidence of ESP?

4. (a) In Worked Example 2, we showed that '8 out of 10' is not significant. Show that '16 out of 20' is significant, and explain the different results between '8 out of 10' and '16 out of 20'.
- (b) At 5% significance level, what values are significant in the context of Worked Example 2.

6.2 The Sign Test

Another important use of hypothesis testing is to find out if, for a particular situation, you improve with practice.

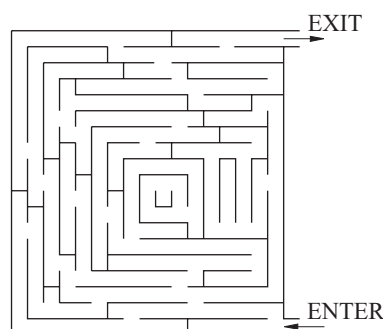
For example, time yourself finding your way through the maze shown opposite. Then have another try.

Are you faster at the second attempt?

Suppose that 7 students took the maze test.

Here are their times in seconds.

<i>First try</i>	<i>Second try</i>	<i>Improvement?</i>
9.0	3.5	✓
6.7	4.0	✓
5.8	2.6	✓
8.3	4.6	✓
5.1	5.4	X
4.9	3.7	✓
9.2	5.7	✓



Out of 7 subjects, 6 have improved, but could this result have occurred by chance? You can set up a hypothesis test in a similar way to the method used in the previous section.

The null hypothesis H_0

The null hypothesis is that any improvement or deterioration in times is quite random and that both are equally likely.

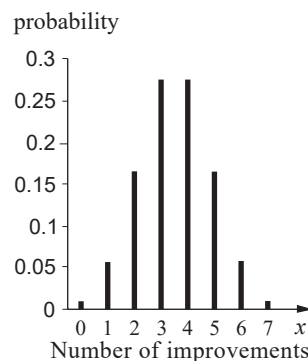
$$p(\text{improvement}) = \frac{1}{2}$$

Alternative hypothesis H_1

In this situation you expect people to improve with practice so the alternative hypothesis is that

$$p(\text{improvement}) > \frac{1}{2}$$

In order to analyse the experimental results, ignore any students who manage to achieve identical times in both trials. Their results actually do not affect your belief in the null hypothesis either way. If there are n students with non-zero differences in times, the binomial distribution can be used as a model to generate probabilities under the null hypothesis. If X is the random variable 'number of positive differences', then $X \sim B(7, \frac{1}{2})$.



This gives the table of probabilities shown below

<i>Number of positive differences</i>	<i>Probability</i>
0	$\binom{7}{0} \left(\frac{1}{2}\right)^7 = 0.0078$
1	$\binom{7}{1} \left(\frac{1}{2}\right)^7 = 0.0547$
2	$\binom{7}{2} \left(\frac{1}{2}\right)^7 = 0.1641$
3	$\binom{7}{3} \left(\frac{1}{2}\right)^7 = 0.2734$
4	$\binom{7}{4} \left(\frac{1}{2}\right)^7 = 0.2734$
5	$\binom{7}{5} \left(\frac{1}{2}\right)^7 = 0.1641$
6	$\binom{7}{6} \left(\frac{1}{2}\right)^7 = 0.0547$
7	$\binom{7}{7} \left(\frac{1}{2}\right)^7 = 0.0078$

The binomial probabilities have been calculated according to $H_0 : X \sim B(7, \frac{1}{2})$. Under H_0 , the probability of :

7 improvements is 0.0078

6 or 7 improvements is 0.0625

If you adhere to a 5% level of significance (0.05 probability of rejecting H_0 when it may be true), the result of 6 improvements is actually not sufficient grounds for rejecting H_0 . This method of hypothesis testing is called the **sign test**.



Worked Example 1

10 'Keep Fit' adults were timed running 100 m.

They were then given exercises to complete over the following 7 days and then timed again running 100 m.

The data is given below.

ADULT	1	2	3	4	5	6	7	8	9	10
1st Run	16.7	20.5	18.9	17.4	22.5	17.8	15.4	20.7	20.6	19.7
2nd Run	15.8	19.2	19.3	17.6	20.1	17.6	15.7	18.3	19.3	18.5

Is there significance evidence to show that the exercises have resulted in improved running times?



Solution

We assume that

$$p\left(\text{improvement} = \frac{1}{2}\right)$$

This is the null hypothesis, H_0 .

The alternative hypothesis, H_1 , is that

$$p\left(\text{improvement} > \frac{1}{2}\right)$$

So assuming H_0 is true, if X is the random variable 'number of positive differences', then

$$X \sim B\left(10, \frac{1}{2}\right)$$

From the data, we see that adults 1, 2, 5, 6, 8, 9 and 10 have improved times; that is, $x = 7$.

So we need to find $p(x \geq 7) = p(x = 10) + p(x = 9) + p(x = 8) + p(x = 7)$

$$= \left(\frac{1}{2}\right)^{10} + \binom{10}{9}\left(\frac{1}{2}\right)^{10} + \binom{10}{8}\left(\frac{1}{2}\right)^{10} + \binom{10}{7}\left(\frac{1}{2}\right)^{10}$$

$$= \left(1 + 10 + \frac{10 \times 9}{1 \times 2} + \frac{10 \times 9 \times 8}{1 \times 2 \times 3}\right) \left(\frac{1}{2}\right)^{10}$$

$$= (1 + 10 + 45 + 120) \left(\frac{1}{2}\right)^{10}$$

$$= \frac{176}{1024}$$

$$= 0.171875$$

But

$$p(x \geq 7) = 0.171875 > 0.05$$

so that using a 5% significance level, there is no evidence of improvement. Even at 10%, there is still no evidence that running times have improved!

You could also calculate $p(x \geq 7)$ by noting that

$$\begin{aligned} p(x \geq 7) &= 1 - p(x \leq 6) \\ &= 1 - 0.8281 \quad (\text{from tables}) \\ &= 0.1719 \end{aligned}$$



Note

Although the sign test is used extensively, it is based only on the sign of the differences rather than on the magnitude of the increases or decreases in values. For example, many of the increases are quite small whereas some of the decreases are considerably larger. So this is an easy significance test to use, but be aware that it does not use the magnitude of the changes.



Exercises

In all questions, assume a 5% level of significance.

1. A group of students undertook an intensive six week training programme, with a view to improving their times for swimming 25 metres breaststroke. Here are their times measured before and after the training programme. Have they improved significantly?

	<i>25 m breast stroke times in seconds</i>							
<i>Student</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>
<i>Before programme</i>	26.7	22.7	18.4	27.3	19.8	20.2	25.2	29.8
<i>After programme</i>	22.5	20.1	18.9	24.8	19.5	20.9	24.0	24.0

2. Twelve young children (6-year-olds) were given a simple jigsaw puzzle to complete. The times they took were measured on their first and second attempts. Did they improve significantly?

	<i>Jigsaw puzzle times in seconds</i>					
<i>Child</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
<i>First attempt</i>	143	43	271	63	232	51
<i>Second attempt</i>	58	45	190	49	178	58

<i>Child</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>	<i>11</i>	<i>12</i>
<i>First attempt</i>	109	156	304	198	83	115
<i>Second attempt</i>	73	127	351	170	74	97

3. A group of 9 children wanted to see whether the amount of air in their bicycle tyres made a difference in how easy it was to pedal their bikes. They decided to ride a particular route under two different conditions: once with a tyre pressure of 40 pounds per square inch (psi) and once with 65 psi. (The order in which they did this was to be decided by tossing a coin.) The time it took (in minutes) for each circuit was:

<i>40 psi</i>	34	54	23	67	46	35	49	51	27
<i>65 psi</i>	32	45	21	63	37	40	51	39	23

Are the children significantly faster with the higher pressure tyres?

4. A group of engineering students run a test to see whether cars will travel as many mpg on 'Standard' fuel as on 'Super Grade' fuel.

	<i>Car</i>									
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>
<i>'Standard' fuel</i>	15	23	21	35	42	28	19	32	31	24
<i>'Super Grade' fuel</i>	18	21	25	34	47	30	19	27	34	20

Does 'Super Grade' fuel give significantly better results?

5. An office manager would like to know whether it will take less time to key data for a standard monthly report using existing software or with a new version. A random sample of 7 data entry clerks was selected and the amount of time taken to key the data was recorded in hours.

	<i>Data Entry Clerk</i>						
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>
<i>Existing software</i>	7.0	7.4	7.8	6.7	6.1	8.1	7.5
<i>New software</i>	6.3	7.5	6.8	6.0	5.3	7.4	7.2

Is the task completed significantly faster using the new software?