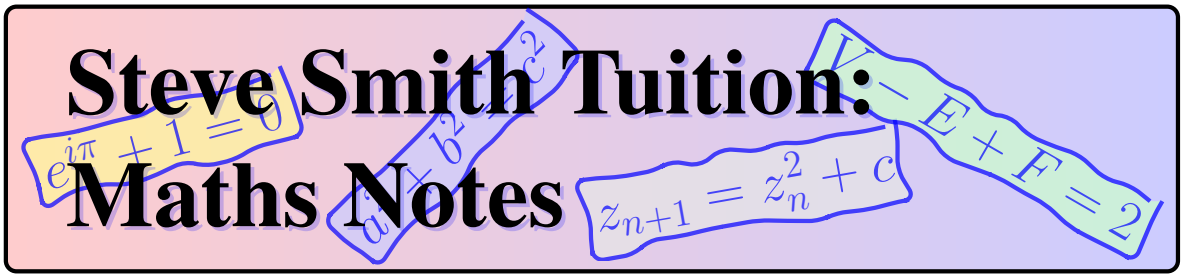


Hypothesis Testing 1 : Using the Binomial Distribution

Contents

Prerequisites

None.

Notes

None.

Document History

Date	Version	Comments
13th December 2012	0.1	Initial creation of the document.

1 Recap

1.1 Calculating Binomial Probabilities

I'm hoping you remember how to calculate binomial probabilities. If not, check out Smith (2013).

Anyway, here's how it goes. Let's say that you throw a dart at a dartboard. Each time you throw, you are trying to hit the 20. Let's say that you throw the dart 10 times, and that your probability of hitting the 20 with any given dart is 0.3. The binomial distribution enabes you to find the probability of hitting no 20s; of hitting one 20; of hitting two 20s; ...; of hitting ten 20s.

Here's how you calculate the probabilities:

Number of 20s	Probability of getting this many 20s	Value (to 4 dp)
0	${}^{10}C_0 \times (0.3)^0 \times (0.7)^{10}$	0.0282
1	${}^{10}C_1 \times (0.3)^1 \times (0.7)^9$	0.1211
2	${}^{10}C_2 \times (0.3)^2 \times (0.7)^8$	0.2335
3	${}^{10}C_3 \times (0.3)^3 \times (0.7)^7$	0.2668
4	${}^{10}C_4 \times (0.3)^4 \times (0.7)^6$	0.2001
5	${}^{10}C_5 \times (0.3)^5 \times (0.7)^5$	0.1029
6	${}^{10}C_6 \times (0.3)^6 \times (0.7)^4$	0.0368
7	${}^{10}C_7 \times (0.3)^7 \times (0.7)^3$	0.0090
8	${}^{10}C_8 \times (0.3)^8 \times (0.7)^2$	0.0014
9	${}^{10}C_9 \times (0.3)^9 \times (0.7)^1$	0.0001
10	${}^{10}C_{10} \times (0.3)^{10} \times (0.7)^0$	0.0000

Figure 1: Calculating Binomial Probabilities

The ${}^nC_r$ values you can get from your calculator. There's a button especially for this.

Here's a picture of the probability distribution for this example:

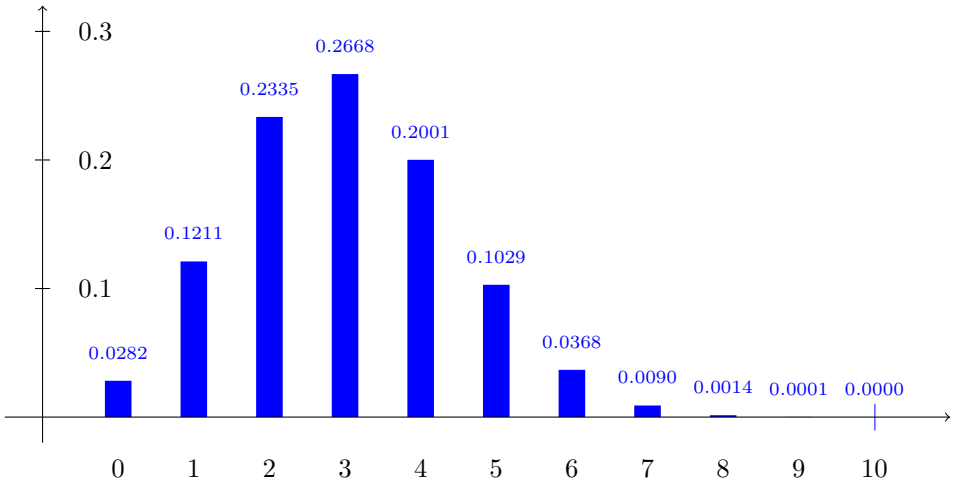


Figure 2: A Picture of Binomial Probabilities

1.2 Calculating Cumulative Binomial Probabilities

The "cumulative" in cumulative probability just means "keep a running total of the probabilities so far". For the example in Section 1.1, here is the corresponding table of cumulative probabilities:

Number of 20s	Probability of getting this many 20s	Cumulative Probability
0	0.0282	0.0282
1	0.1211	0.1493
2	0.2335	0.3828
3	0.2668	0.6496
4	0.2001	0.8497
5	0.1029	0.9527
6	0.0368	0.9894
7	0.0090	0.9984
8	0.0014	0.9999
9	0.0001	1.0000
10	0.0000	1.0000

Figure 3: Calculating Cumulative Probabilities

The idea being that, for example, the probability of getting 0, 1 or 2 20s in 10 throws would be 0.3828. It is a table of cumulative probabilities that you are given in your formula book in the exam.

Using cumulative probabilities, we can answer questions such as: "What is the probability of getting two or fewer 20s in 10 throws?" Or: "What is the probability of getting eight or more 20s in 10 throws?"

The first can be answered simply by looking at the cumulative probability table. The probability of getting two 20s or fewer will just be the entry for "2 20s" in the cumulative probability column. In other words, the probability of getting two or fewer 20s will be 0.3828 (or about 38%).

But how do you find the probability of getting eight or more 20s? Well, we can still get this from the cumulative probability table, but we have to do a bit of work to get it. Since all the probabilities added together for no 20s, one 20, two 20s, etc, will be 1, then the probability of getting eight or more will be one minus the probability of getting seven or less! So the probability of getting eight or more 20s in 10 throws will be $1 - 0.9984 = 0.0016$ (or 0.16%).

These last two questions are completely equaivalent to: "How likely is it for me to get two 20s or fewer in ten throws? (About 38%.) Or: "How likely is it for me to get eight or more 20s in ten throws? (0.16%.)

This leads us on to...

1.3 Recognising When a Result is Unlikely

You are now about to throw 10 darts at the dartboard. If I were now to ask you if it was likely that you would get two or fewer 20s in your 10 throws, you might answer: "Pretty likely. Because the probability of getting two or fewer 20s is about 38%, it will happen more than once every 3 times I throw 10 darts". Absolutely right.

On the other hand, if I were now to ask you if it was likely that you would get eight or more 20s in your 10 throws, you might answer: "Pretty unlikely. Because the probability of getting eight or more 20s is 0.16%, it will only actually happen roughly once every 63 times I throw 10 darts". Absolutely right again.

So getting two or fewer 20s in 10 throws is pretty common; eight or more will be very rare. This is the basis of hypothesis testing...

2 Hypothesis Testing : The Basic Idea

Hypothesis testing is used to answer questions like "If each dart hits the 20 with a probability of 0.3, would it be likely for me to get eight or more 20s in 10 throws of the dart?" So the result of a hypothesis test is either: "Yes, it is likely to get this outcome by chance" or "No, it is not likely to get this outcome by chance".

And if the answer is "No: it is not likely to get this outcome by chance", then what can you conclude? Well, the main idea behind hypothesis testing is that if a particular result occurs, but that result is unlikely to happen by chance, then *we have the wrong probabiliity in our calculation*!

Let's have a look at a couple of examples.

2.1 Example 1 : Rolling a Die

Let's say that you were rolling a die 8 times. Let's assume for a moment that it's a fair die so the probability of getting any given number when you roll it will be $\frac{1}{6}$. Let's say we were interested in the number of 6s that come up in the eight rolls. Here's the cumulative probability table:

Number of 6s	Probability of getting this many 6s	Cumulative Probability
0	0.2326	0.2326
1	0.3721	0.6047
2	0.2605	0.8652
3	0.1042	0.9694
4	0.0260	0.9954
5	0.0042	0.9996
6	0.0004	1.0000
7	0.0000	1.0000
8	0.0000	1.0000

Figure 4: Cumulative Probability Table : N = 8; p = $\frac{1}{6}$

What would be the probability of getting six or more 6s in eight rolls of this die? Well using our cumulative probability table it would be

$$p(Six\ or\ more\ 6s) = 1 - 0.9996 = 0.0004 = 0.04\%$$

which is very small. Let's have a look at the probability distribution picture:

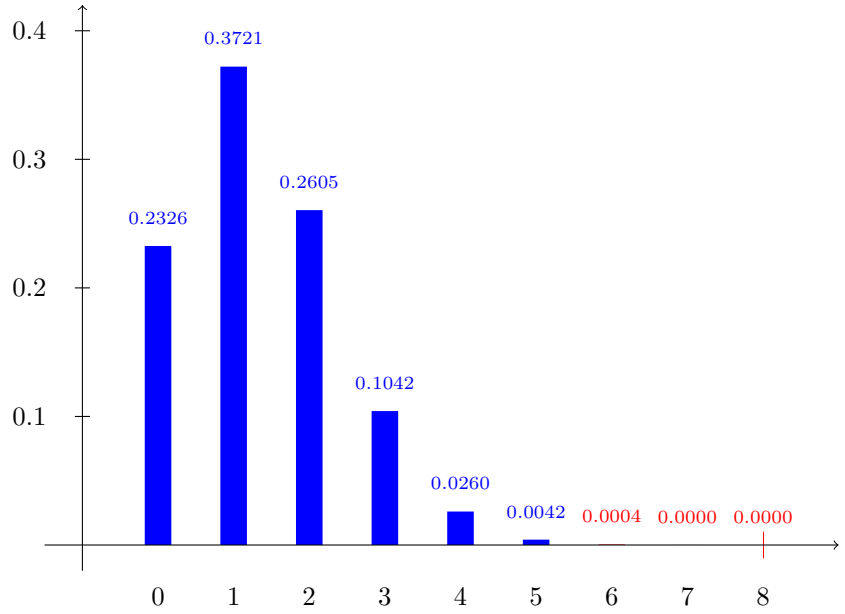


Figure 5: Binomial Probabilities : N = 8; p = $\frac{1}{6}$

So, what if we rolled this die eight times and got six 6s? Now do you remember our assumption? We assumed that the die was fair. What if it isn't fair? What if the probabililty of getting a six isn't $\frac{1}{6}$? What if it was higher than that? Say 0.5? What then? Well, if the probability of getting a six was 0.5 each time we roll the die, the cumulative probability table would be:

Number of 6s	Probability of getting this many 6s	Cumulative Probability
0	0.0039	0.0039
1	0.0313	0.0352
2	0.1094	0.1445
3	0.2188	0.3633
4	0.2734	0.6367
5	0.2188	0.8555
6	0.1094	0.9648
7	0.0313	0.9961
8	0.0039	1.0000

Figure 6: Cumulative Probability Table : N = 8; p = 0.5

And let's have a look at the picture for this distribution:

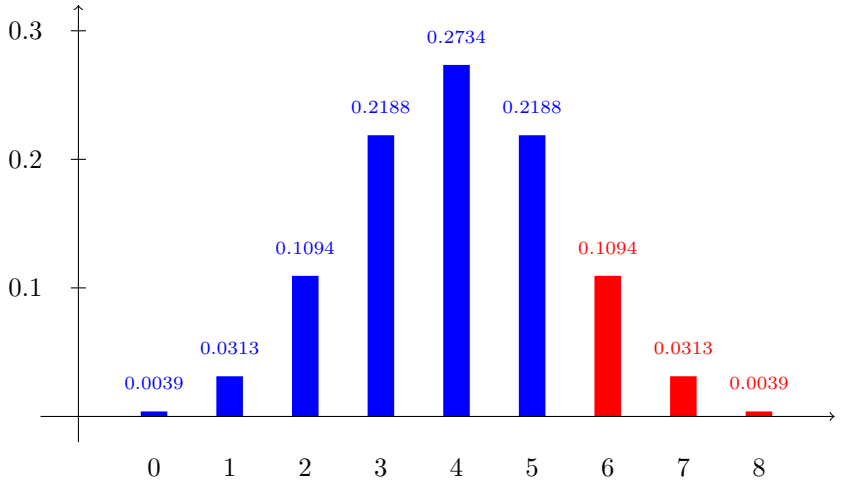


Figure 7: Binomial Probabilities : N = 8; p = 0.5

So this time,

$$p(Six\ or\ more\ 6s) = 1 - 0.8555 = 0.1445 \approx 14\%$$

and wow - 14% isn't that unlikely: if the probability of getting a six on each throw was 0.5, we would get six or more 6s roughly once every seven times we did this experiment. So it's not an unlikely outcome now, as it would definitely have been if the probability of getting a 6 on each throw was $\frac{1}{6}$.

So it looks to me as though our original assumption that the die was fair was wrong. It's much more likely that the proability of getting a six when you roll this die is greater than $\frac{1}{6}$.

Here's a flow chart of the thought process we've just gone through:

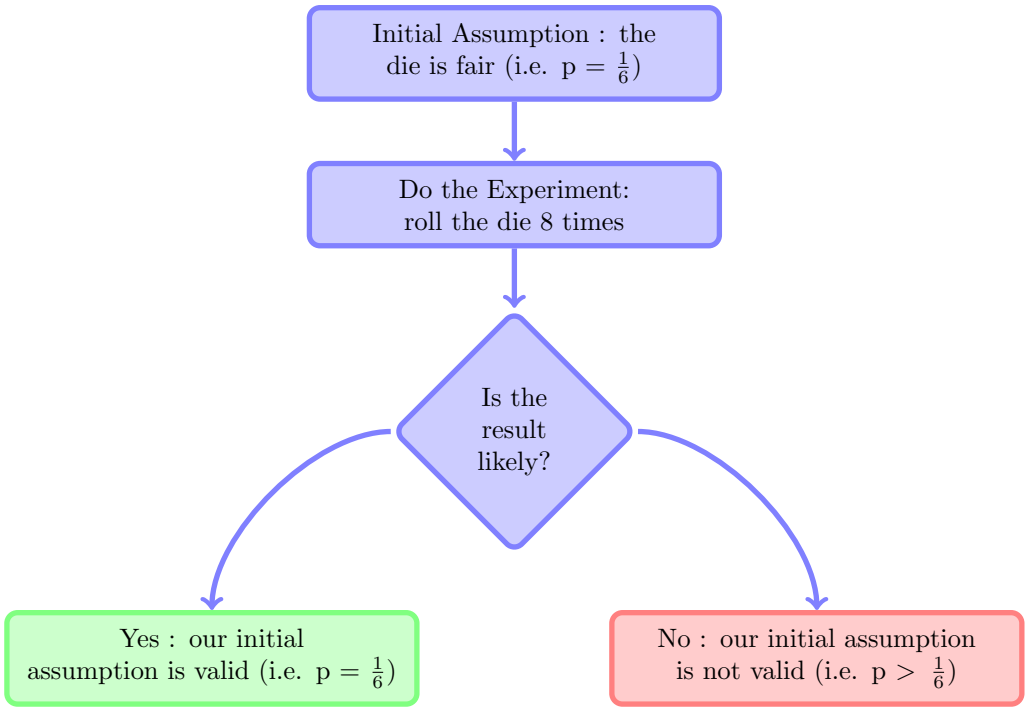


Figure 8: Hypothesis Test Flow Chart

And if we were to do this experiment (roll the die eight times) and get six 6s, we would conclude that the probability of getting a six on each throw was more than $\frac{1}{6}$.

2.2 Example 2 : Tossing a Coin

Let's say that you were tossing a coin ten times. Let's assume for a moment that it's a fair coin so the probability of getting a head when you toss it it will be 0.5. Let's say we were interested in the number of

heads that come up in the ten tosses. Here's the cumulative probability table:

Number of Heads	Probability of getting this many Heads	Cumulative Probability
0	0.0010	0.0010
1	0.0098	0.0107
2	0.0439	0.0547
3	0.1172	0.1719
4	0.2051	0.3770
5	0.2461	0.6230
6	0.2051	0.8281
7	0.1172	0.9453
8	0.0439	0.9893
9	0.0098	0.9990
10	0.0010	1.0000

Figure 9: Cumulative Probability Table : N = 10; p = 0.5

And here's a picture of this distribution:

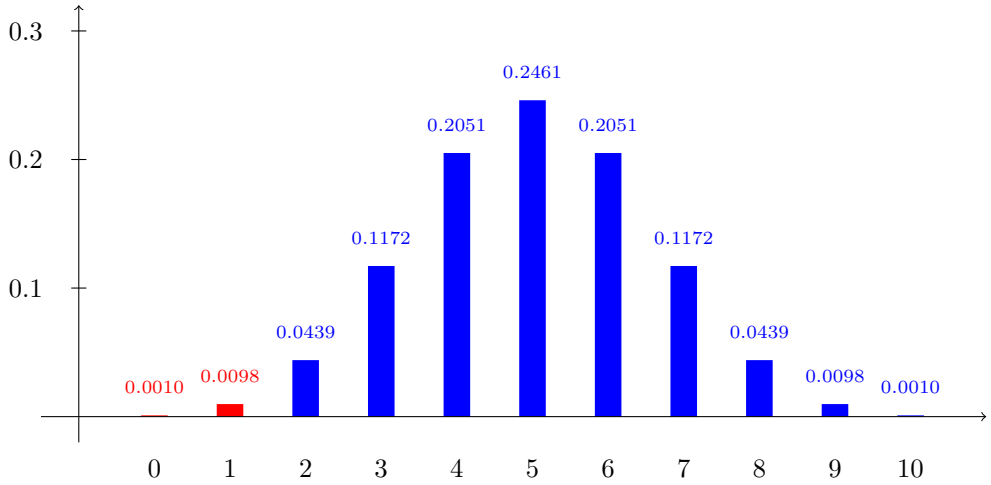


Figure 10: Binomial Probabilities : N = 10; p = 0.5

What would be the probability of getting one or fewer heads in ten tosses of this coin? Well using our cumulative probability table it would be

$$p(One\ or\ fewer\ heads) = 0.0107 \approx 1\%$$

which is very small. So, what if we tossed this coin ten times and got one head? Now do you remember our assumption? We assumed that the coin was fair. What if it isn't fair? What if the probabililty of getting a head isn't 0.5? What if it was lower than that? Say 0.2? What then? Well, if the probability of getting a head was 0.2 each time we toss the coin, the cumulative probability table would be:

Number of Heads	Probability of getting this many Heads	Cumulative Probability
0	0.1074	0.1074
1	0.2684	0.3758
2	0.3020	0.6778
3	0.2013	0.8791
4	0.0881	0.9672
5	0.0264	0.9936
6	0.0055	0.9991
7	0.0008	0.9999
8	0.0001	1.0000
9	0.0000	1.0000
10	0.0000	1.0000

Figure 11: Cumulative Probability Table : N = 10; p = 0.2

And the picture of this distribution is:

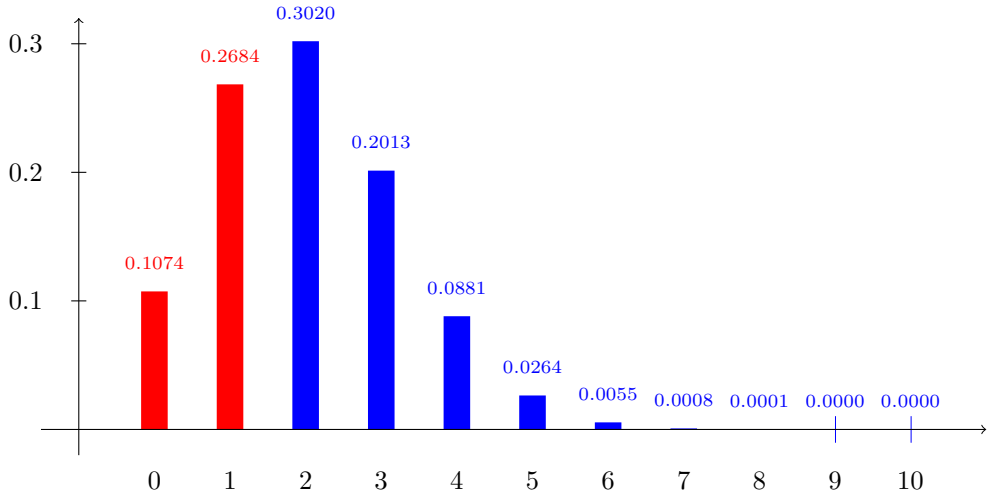


Figure 12: Binomial Probabilities : N = 10; p = 0.2

So this time,

$$p(One\ or\ fewer\ heads) = 0.3758 \approx 38\%$$

and wow - 38% isn't unlikely: if the probability of getting a head on each throw was 0.2, we would get one or fewer heads roughly once every three times we did this experiment. So it's not an unlikely outcome now, as it would definitely have been if the probability of getting a head on each toss was 0.5.

So it looks to me as though our original assumption that the coin was fair was wrong. It's much more likely that the probability of getting a head when you toss this coin is less than 0.5. Again, here's a flow chart of our thought process:

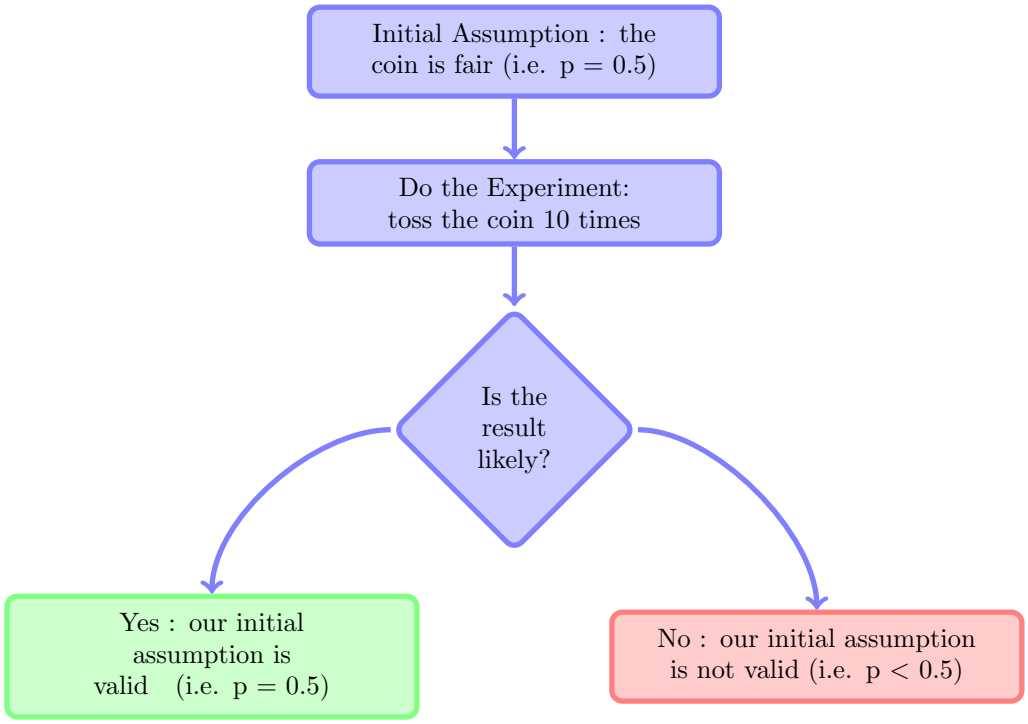


Figure 13: Hypothesis Test Flow Chart

2.3 The Story So Far...

So, here's the idea behind hypothesis testing.

- We are interested in the probability of something happening (eg getting a 6 when you roll a die, or getting a head when you toss a coin).
- We have a first stab at what we think this probability is (e.g. for the die it was $\frac{1}{6}$, and for the coin it was 0.5). This is our "p".
- We do an experiment by doing the activity a number of times. The number of times we do the activity is our "N".
- We count how many times out of the N times we do the activity that our outcome (e.g. the 6 or the head) comes up.
- We compare this with what we expect. We do this by looking at cumulative probability tables or pictures of the probability distribution for our situation.
- If the result is likely, then we conclude that our initial guess for the p was correct; but if the result is unlikely, we conclude that our initial guess for p was wrong.

There is another interesting issue that I have glossed over so far. It's this business about "six *or more* 6s in eight rolls", or "one *or less* heads in ten tosses". This is discussed in Appendix A.

If you look back at the pictures of binomial probabilities for the two examples so far, the parts of the distribution that we've been interested in have been marked in red. And the red regions are the ends of the distributions. This is because these ends of the distributions represent small probabilities (and consequently unlikely outcomes). When you are doing hypothesis test calculations, it's the red ends of the distributions that you are interested in.

2.4 But...Who Decides What "Likely" Means?

Now that is a very good question indeed.

Going back to our rolling the die experiment, we concluded that if we have our guess for p as $\frac{1}{6}$, there was only a 0.04% probability of getting six or more 6s when we rolled the die eight times. That's not very likely.

But when our guess for p was 0.5, there was a 14% probability of getting six or more 6s when we rolled the die eight times. We decided that that could easily happen.

But what if we had a value of p between $\frac{1}{6}$ and 0.5? Presumably our probability of getting six or more 6s then would be somewhere between 0.04% and 14%. What if our result was 6%? Is that likely, or not? Where do we draw the line? Is there some magic probability that we consider to be "likely"?

Yes and no. It's obvious that we need a value for our probability that we can compare our result with. So we know which way to go when we hit the "Is the result likely" bit in the flow chart. But it turns out that essentially, you can pick any value you like[1].

To get around this problem, statisticians use particular values for this likelihood threshold. Notice - value*s*. And which one you pick is way beyond A-Level. In A-Level questions, the likihood value will be given to you. Statisticians call this value *the significance level*.

Typical significance levels are 1%, 5% and 10%. And as I say, you will be told in the question which one to use.

So, for example, if we were using the 5% significance level, when you get to the "Is the result likely" bit in the flow chart, you compare your result with 5%. If the probability of getting your result is less than 5%, then it is considered unlikely; if the probability of getting your result is more than 5%, then it is considered to be likely.

So, let's take this "significance level" idea on board, and have a look at another couple of examples.

2.5 Example 3 : Toast

2.5.1 The Burning Question

A group of 18 students decides to investigate the truth of the saying that if you drop a piece of toast it is more likely to land butter-side down.

They each take one piece of toast, butter it on one side, and throw it in the air. Eleven land butter-side down (BSD), the rest butter side-up (BSU). Use their results to carry out a hypothesis test at the 5% significance level.

2.5.2 The Burning Answer

OK: the first thing we need to do is to make an initial assumption about p, the probability of the toast landing butter-side down. Let's assume for a moment that the toast lands on either side with equal chance. So, in this case, $p = 0.5$.

Now, we are doing the activity 18 times, so $N = 18$. Here's the cumulative probability table:

[1] I just *love* statistics, don't you?

Number of BSDs	Probability of getting this many BSDs	Cumulative Probability
0	0.0000	0.0000
1	0.0001	0.0001
2	0.0006	0.0007
3	0.0031	0.0038
4	0.0117	0.0154
5	0.0327	0.0481
6	0.0708	0.1189
7	0.1214	0.2403
8	0.1669	0.4073
9	0.1855	0.5927
10	0.1669	0.7597
11	0.1214	0.8811
12	0.0708	0.9519
13	0.0327	0.9846
14	0.0117	0.9962
15	0.0031	0.9993
16	0.0006	0.9999
17	0.0001	1.0000
18	0.0000	1.0000

Figure 14: Cumulative Probability Table : N = 18; p = 0.5

And here's a picture of this distribution:

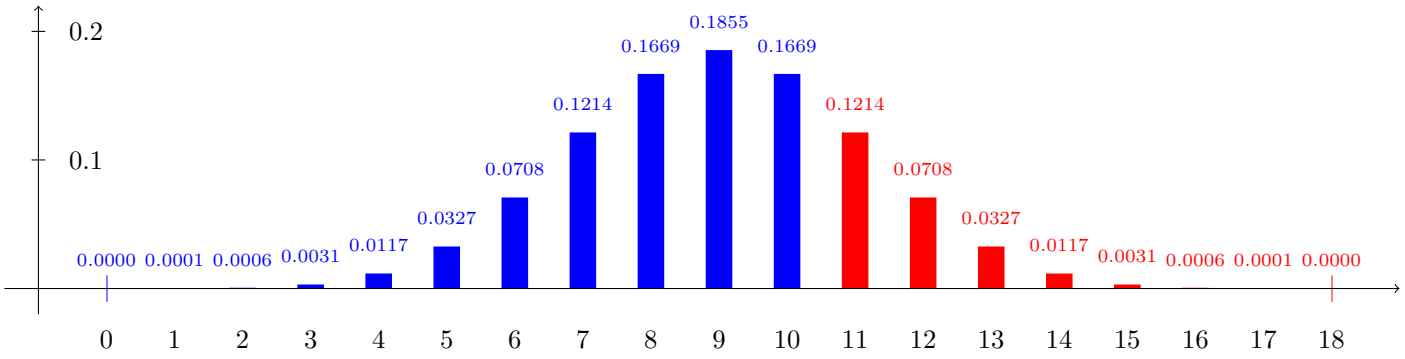


Figure 15: Binomial Probabilities : N = 18; p = 0.5

Now in the experiment there were eleven pieces BSD. What is the probability of this? Now the reason why we're doing this experiment is because of the old wive's tale that the toast will always land BSD when you drop it. So the inference is that $p > 0.5$. If that was the case, then we would expect more BSD pieces than BSU pieces. So we are looking at the right-hand end of the distribution.

So using our cumulative probability table the probability of getting *eleven or more* BSD pieces would be

$$p(\textit{Eleven or more pieces land BSD}) = 1 - 0.7597 = 0.2403 \approx 24\%$$

Remember that we always have to find the entire red region: not just the value of getting eleven BSDs, but eleven *or more* BSDs.

Now we compare this result with our significance level. Since 24% is larger than 5% (that is, the chance of eleven or more pieces of toast landing BSD is *likely*), then we say that this could easily happen by chance, so there's no reason to suspect that p is greater than 0.5.

Here's a flow chart of our thought process:

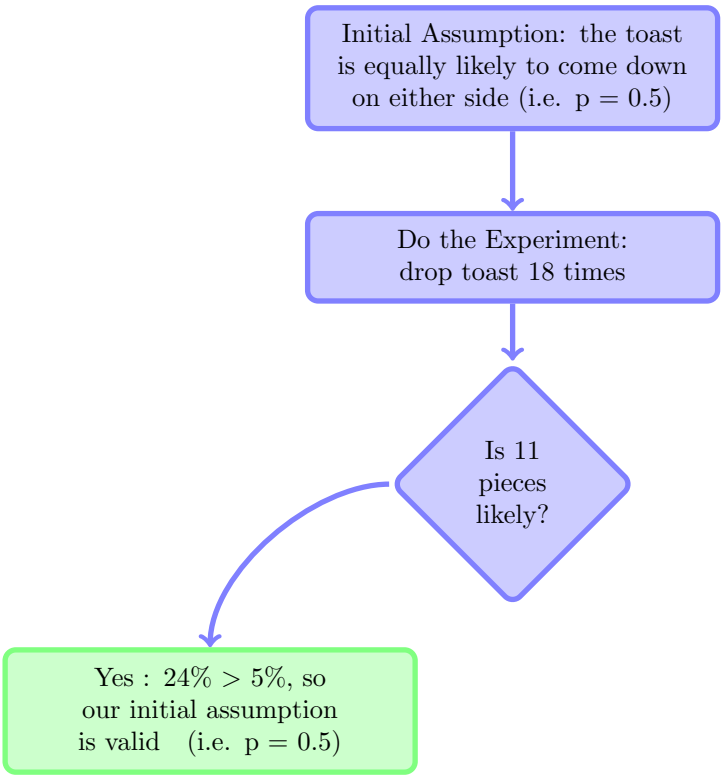


Figure 16: Hypothesis Test Flow Chart

2.6 Example 4 : Driving Tests

2.6.1 The Testing Question

On average, 45% of people pass their driving test first time.

There are complaints that Mr Smith is too harsh, and so, unknown to himself, his work is monitored. It is found that he passes 4 out of a batch of 20 candidates.

Does this provide evidence, at the 5% significance level, that the complaints are justified?

2.6.2 The Testing Answer

OK: the first thing we need to do is to make an initial assumption about p, the probability a given candidate passes first time. In this case, that's given to us in the question. It is $p = 0.45$ (0.45 is 45% as a decimal).

Now, we are doing the activity 20 times, so $N = 20$. Here's the cumulative probability table:

Number of Passes	Probability of getting this many Passes	Cumulative Probability
0	0.0000	0.0000
1	0.0001	0.0001
2	0.0008	0.0009
3	0.0040	0.0049
4	0.0139	0.0189
5	0.0365	0.0553
6	0.0746	0.1299
7	0.1221	0.2520
8	0.1623	0.4143
9	0.1771	0.5914
10	0.1593	0.7507
11	0.1185	0.8692
12	0.0727	0.9420
13	0.0366	0.9786
14	0.0150	0.9936
15	0.0049	0.9985
16	0.0013	0.9997
17	0.0002	1.0000
18	0.0000	1.0000
19	0.0000	1.0000
20	0.0000	1.0000

Figure 17: Cumulative Probability Table : N = 20; p = 0.45

And here's a picture of this distribution:

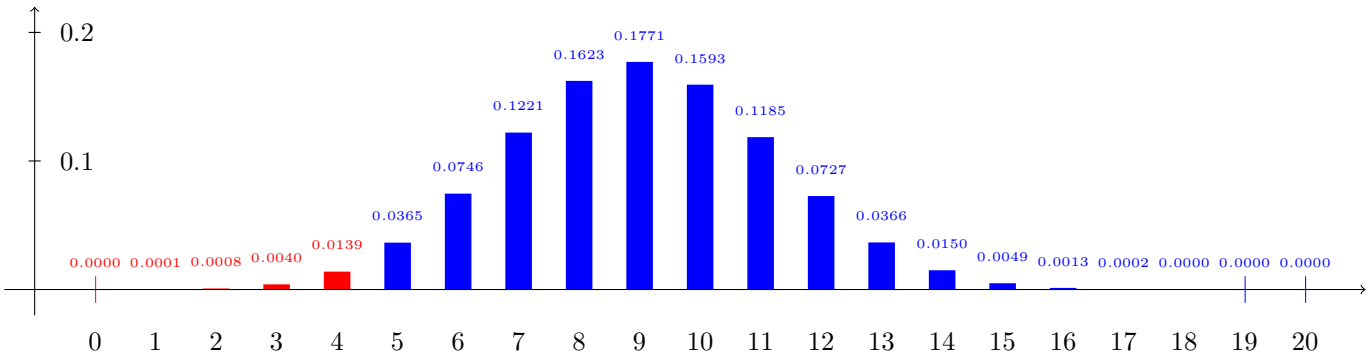


Figure 18: Binomial Probabilities : N = 20; p = 0.45

Now we've four first-time passes out of 20. What is the probability of this? In this case, the inference is that $p < 0.45$ since the claim is made that Mr Smith is too harsh when testing candidates. That means we expect the number of passes will be too low. So this time, we are looking at the left-hand end of the distribution.

So using our cumulative probability table the probability of getting four passes or less would be

$$p(Four\ or\ fewer\ passes) = 0.0189 \approx 1.9\%$$

Now we compare this result with our significance level. Since 1.9% is smaller than 5% (that is, the chance of four or fewer candidates passing first time from a batch of 20 is *not likely*), then we say that this could *not* easily happen by chance, so there *is* reason to suspect that p is smaller than 0.45 in the case of Mr Smith!

Here's a flow chart of our thought process:

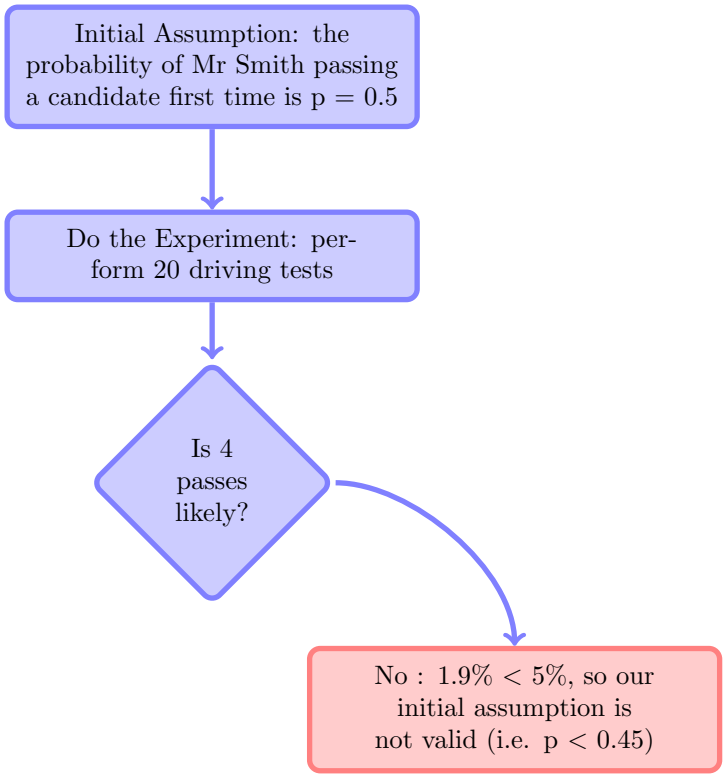


Figure 19: Hypothesis Test Flow Chart

2.7 Critical Regions

Instead of working out probabilities of outcomes and comparing them with the significance level, as we did in Examples 1 and 2, you could look at the problem the other way round.

2.7.1 Example 1 Revisited

In the example of rolling the die, we could look at the problem this way: how many 6s would we need to get in eight rolls for the result to be unlikely?

The first thing is: what do we mean by unlikely? This means we have to set a significance level. Let's set our significance level at 5%.

If you think about it, this means that we need to find the largest red region at the right-hand end of the distribution where the red probabilities add up to less than 5%.

Here's my picture of what I mean:

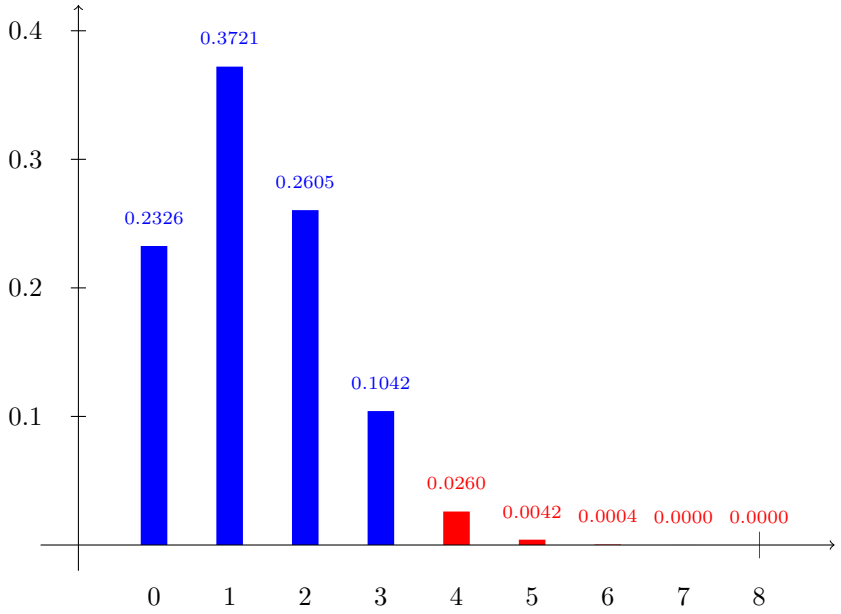


Figure 20: Binomial Probabilities : N = 8; p = $\frac{1}{6}$

Taking my red region from four 6s up, then the probability of four or more 6s is (from the cumulative probability table in Section 2.1) 1 - 0.9694 = 0.0306 (or about 3%). This is less than my significance level (5%), so is unlikely.

However, if I had picked my red region to be three 6s or more, then the probability would be 1 - 0.8652 = 0.1348 (which is about 13%). This is greater than my significance level (5%) so it is likely:

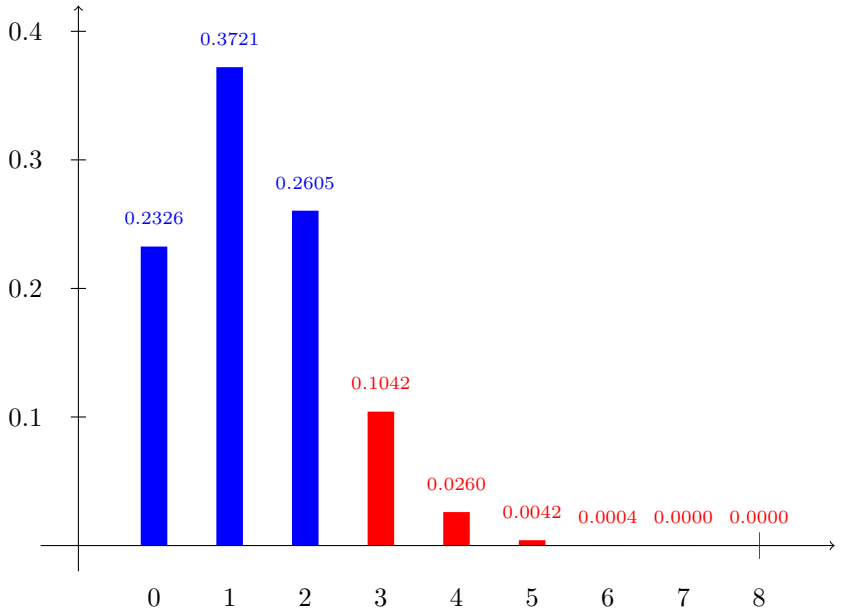


Figure 21: Binomial Probabilities : N = 8; p = $\frac{1}{6}$

So the unlikely region is the region for which I get four or more 6s when I roll my die eight times. This region (four or more 6s) is what they call the critical region.

2.7.2 The Use of Critical Regions

The use critical regions is often more convenient calculating probabilities. If you had a machine that made widgets, for example, and you wanted to check that it was working satisfactorily, then the best thing to tell the operator is to take a sample of 20 widgets every hour, and if four or more are faulty, then shut the machine down, and call maintenance.

Since the operator might not know anything about hypothesis tests, he might not be able to use the cumulative binomial probability tables to calculate whether it is likely or not to get four faulty widgets in a batch of 20. So for him, it is easier to tell him the critcal region than it is to get him to calculate probabilities and compare them with the significance level.

Here's a flow chart of a hypothesis test when you are using critical regions:

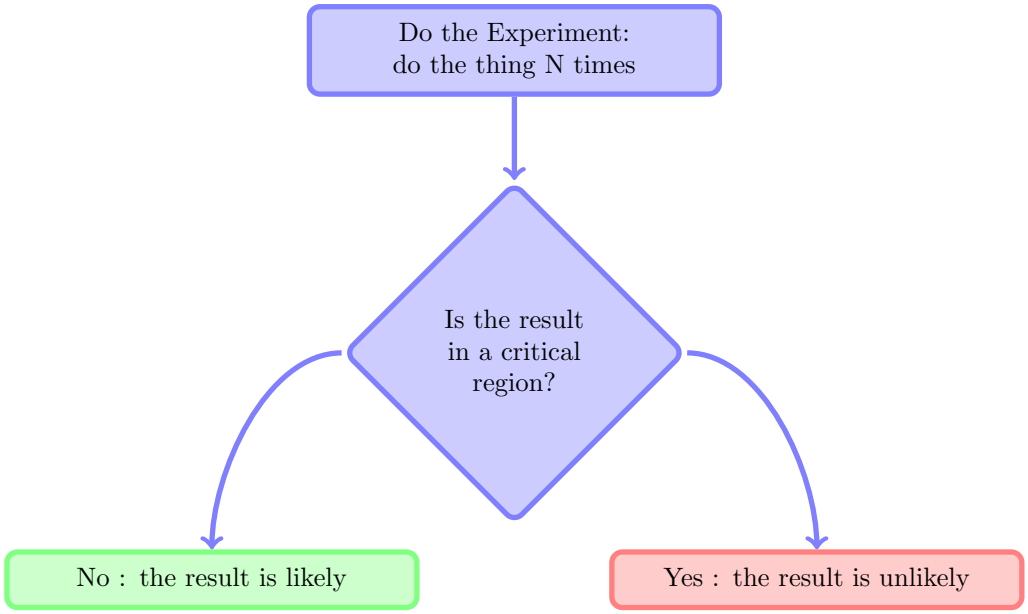


Figure 22: Hypothesis Test Flow Chart With Critical Regions

2.7.3 Example 2 Revisited

So what would be the critical region in Example 2? Well, if you remember, in that example we were looking at getting a small number of heads when you toss a coin ten times. Don't forget that we need a significance level, so let's make it 5%.

In that case, using the cumulative probability table in Section 2.2, we find our critical region to be 0 or 1 head, since the probability of getting one head or less is 0.0107 ($\approx 1\%$), whereas the probability of getting two heads or less is 0.0547 ($= 5.47\%$), which is greater than 5%, so is considered to be likely.

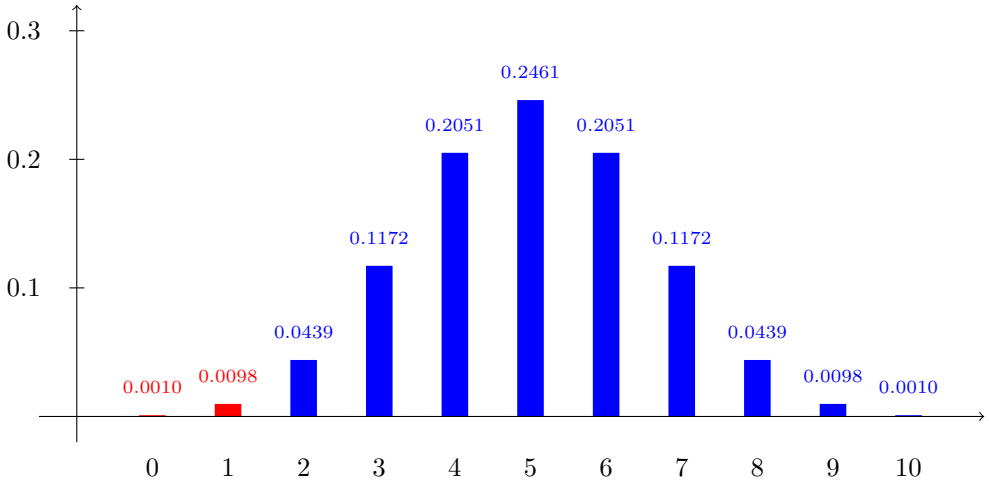


Figure 23: Binomial Probabilities : N = 10; p = 0.5

2.8 Example 5 : Meg's Mugs

Meg is an entrepeneur. She has her own business, making mugs. Unfortunately, she has a quality control problem.

Meg finds that 30% of the mugs she makes are faulty, and cannot be sold to the public. Aghast, she finds a company on the internet that claims that their mug-making machine is much better than that. They claim that their fault rate is much lower than 30%.

Meg arranges a test of their machine. She wants to see for herself whether the new machine has a fault rate lower than 30%. But how is she to do this? Well, her reasoning goes like this: I'm going to use their machine to make a batch of 16 mugs. If I can calculate the critical region for this situation, I can determine whether it is likely or not that the fault rate of the new machine was 30%.

First of all, let's have a look at the cumulative probability table for this situation:

Number of Faulty Mugs	Probability of getting this many faulty mugs	Cumulative Probability
0	0.0033	0.0033
1	0.0228	0.0261
2	0.0732	0.0994
3	0.1465	0.2459
4	0.2040	0.4499
5	0.2099	0.6598
6	0.1649	0.8247
7	0.1010	0.9257
8	0.0487	0.9743
9	0.0185	0.9929
10	0.0056	0.9984
11	0.0013	0.9997
12	0.0002	1.0000
13	0.0000	1.0000
14	0.0000	1.0000
15	0.0000	1.0000
16	0.0000	1.0000

Figure 24: Cumulative Probability Table : N = 16; p = 0.3

So, assuming that the fault rate of the new machine was 30%, then if Meg did this trial of the new machine, the probability of getting no faulty mugs would be 0.0033 (0.33%); the probability of getting 1 or fewer faulty mugs would be 0.0261 (about 2.6%); the probability of getting 2 or fewer faulty mugs would be 0.0994 (about 10%).

So here, the critical region will be one or less faulty mugs:

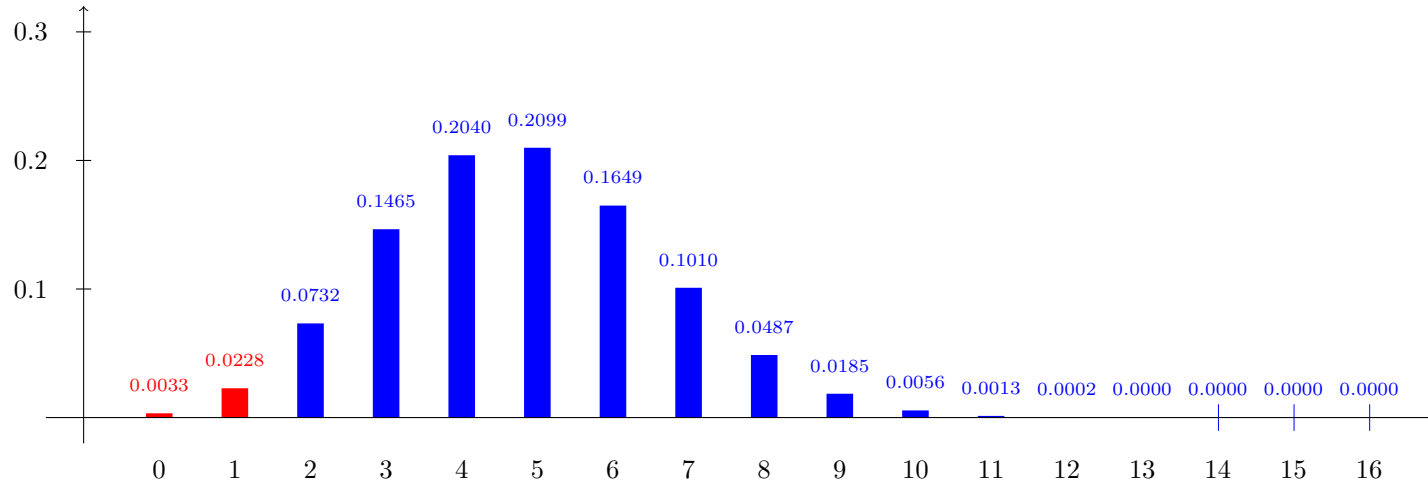


Figure 25: Binomial Probabilities : N = 10; p = 0.5

because that's the largest region at the left-hand end of the distribution to be less than 5% of the total probability.

So the new machine would need to make at most one faulty mug in a batch of 16 to pass the test.

2.9 One- and Two- Tailed Tests

So far, we have only been looking at situations where the critical region was at one end of the distribution. These are called *One-Tailed Tests*. These situations arise when we are only interested in whether the p is either *greater than* our initial assumption (as in Example 1), or *less than* our initial assumption (as in Example 2).

But some questions are phrased differently, leading us to the idea of *Two-Tailed Tests*...

2.10 Example 6 : Nesting Birds

2.10.1 The Sexy Question

A biologist discovers a colony of a previously unknown type of bird nesting in a cave. Out of the 16 chicks which hatch during his period of study, 13 are female. Test, at the 5% significance level, whether this supports the view that the sex ratio for the chicks differs from 1[2].

2.10.2 The Sexy Answer

Now in this case, the question is NOT "Test the view that there are more females born than males". If our initial p was the probability of having a female chick born (and our initial assumption would be $p = 0.5$), then this test would be to see whether it was likely that $p > 0.5$. This would be a one-tailed test, using the right-hand end of the distribution.

Neither is it: "Test the view that there are more males born than females". If our initial p was the probability of having a female chick born (and our initial assumption would be $p = 0.5$), then this test would be to see whether it was likely that $p < 0.5$. This would be a one-tailed test, using the left-hand end of the distribution.

The question actually is: "Test the view that there are different numbers of females and males born." This means that if our initial p was the probability of having a female chick born (and our initial assumption would be $p = 0.5$), then this test would be to see whether it was likely that $p \neq 0.5$.

Instead, this would actually be a *two-tailed test*, using the *both* ends of the distribution.

So: how do we tackle this? First, let's have a look at our cumulative probability table for this situation:

[2]If the sex ratio is 1, you'd expect the same numbers of each to be born.

Number of Female Birds	Probability of getting this many Female Birds	Cumulative Probability
0	0.0000	0.0000
1	0.0002	0.0003
2	0.0018	0.0021
3	0.0085	0.0106
4	0.0278	0.0384
5	0.0667	0.1051
6	0.1222	0.2273
7	0.1746	0.4018
8	0.1964	0.5982
9	0.1746	0.7728
10	0.1222	0.8949
11	0.0667	0.9616
12	0.0278	0.9894
13	0.0085	0.9979
14	0.0018	0.9997
15	0.0002	1.0000
16	0.0000	1.0000

Figure 26: Cumulative Probability Table : N = 16; p = 0.5

Now because we have a two-tailed test, what we have to do this time is to analyse both ends of the distribution, splitting the 5% significance level between the two ends, so that there is 2.5% at each end:

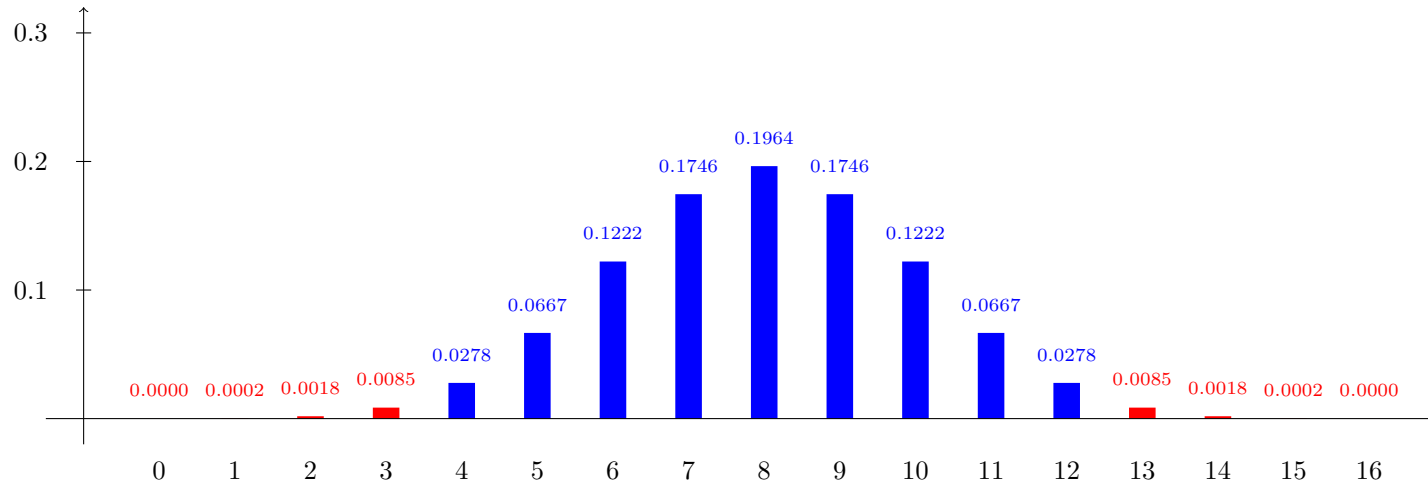


Figure 27: Binomial Probabilities : N = 16; p = 0.5

So assuming that the probability of any chick being female is 0.5, then:

the probability of getting three or fewer females in a batch of 16 will be 0.0106 (≈ 1%), which is less than 2.5%;

the probability of getting four or fewer females in a batch of 16 will be 0.0384 (≈ 4%), which is more than 2.5%;

the probability of getting thirteen or more females in a batch of 16 will be 1 - 0.9894 = 0.0106 (≈ 1%), which is less than 2.5%;

the probability of getting twelve or more females in a batch of 16 will be 1 - 0.9616 = 0.0384 (≈ 4%), which

is more than 2.5%.

So the critical regions in this example are three females or fewer AND thirteen females or more. And the result of this experiment indicates that the sex ratio of these birds is not 1.

2.11 Example 7 : Quiz Kids

2.11.1 The Quizzy Question

A multiple-choice test has 12 questions, with the answer for each allowing four options, A, B, C, and D. A student in a class taking the test tells her teacher that she guessed all 12 answers. The teacher does not believe her. What test (at the 10% significance level) can the teacher apply to see if it was likely that the student selected her answers randonly?

2.11.2 The Quizzy Answer

Well, in the case, if our p is the probability of getting an answer right by guessing, then it is going to be 0.25, since there is a 1 in 4 chance of getting an answer right by guessing.

The N in this problem is 12, so the first thing to do is to draw up the cumulative probability table:

Number of Correct Questions	Probability of getting this many Correct Questions	Cumulative Probability
0	0.0317	0.0317
1	0.1267	0.1584
2	0.2323	0.3907
3	0.2581	0.6488
4	0.1936	0.8424
5	0.1032	0.9456
6	0.0401	0.9857
7	0.0115	0.9972
8	0.0024	0.9996
9	0.0004	1.0000
10	0.0000	1.0000
11	0.0000	1.0000
12	0.0000	1.0000

Figure 28: Cumulative Probability Table : N = 12 p = 0.25

And since the significance level is 10%, then we have to split that between the two ends of the distribution, so there will be 5% at each end. And we want to find the largest regions that have probabilities less than 5% at each end. Here's my picture:

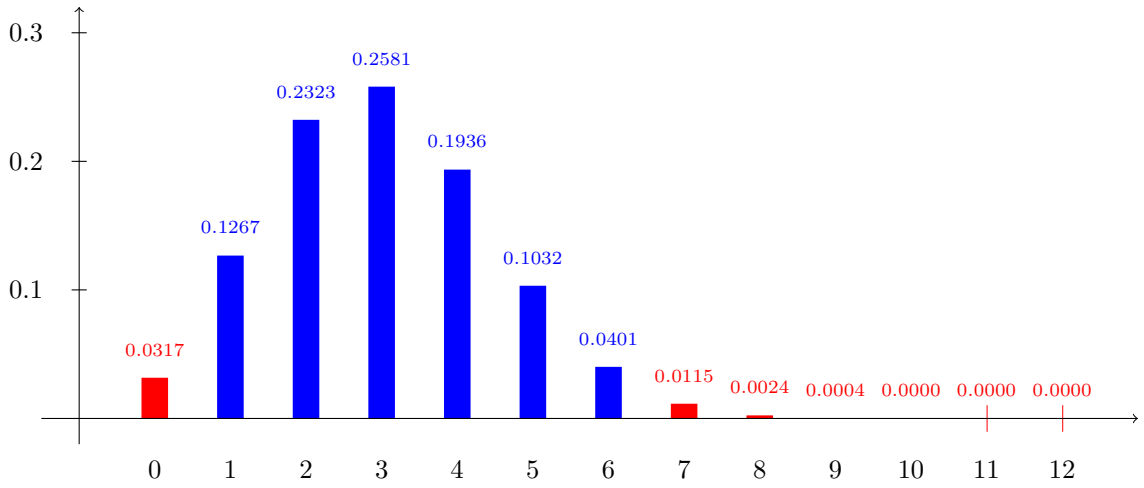


Figure 29: Binomial Probabilities : N = 12; p = 0.25

This time the critical regions are not symmetrical: that's because the distribution is not symmetrical, and that's because $p \neq 0.5$.

At the left-hand end, the critical region is 0 questions correct. Why??

And at the right-hand end the critical region is 7 or more questions correct. Why??

So, if a student gets 7 or more questions correct, it is unlikely that she is guessing the answers. It is more likely that she is trying to get them right by picking the correct answers each time.

And if a student gets no questions correct, it is unlikely that she is guessing the answers. It is more likely that she is trying to get them wrong by picking the wrong answers each time!!

3 Hypothesis Test Terminology

You may have come across the phrases "the null hypothesis" and "the alternative hypothesis". This kind of stuff abounds in statistics. It's annoying to have terminology that confuses people.

Here are the definitions of the two statements:

The *null hypothesis*: this is your initial assumption about the probability of an individual event. If the null hypothesis is supported, then it simply means that the result of your experiment is likely.

The *alternative hypothesis*: this is the opposite of your initial assumption about the probability of an individual event. If the alternative hypothesis is supported, then it simply means that the result of your experiment is unlikely.

3.1 Null and Alternative Hypothesis Examples

3.1.1 Example 1 Re-revisited

In Example 1, where we rolled the die eight times,

the null hypothesis was "The probability of getting a 6 when you roll the die is $\frac{1}{6}$";

the alternative hypothesis was "The probability of getting a 6 when you roll the die is greater than $\frac{1}{6}$".

If we performed the experiment and got six 6s in eight rolls, which is unlikely, then the alternative hypothesis was supported, and we reject the null hypothesis. This is just code for "it was unlikely to get six 6s in eight rolls, so that the probability of rolling a 6 with this die is probably greater than $\frac{1}{6}$".

3.1.2 Example 2 Re-revisited

In Example 2, where we tossed a coin ten times:

the null hypothesis was "The probability of getting a head when you toss the coin is 0.5";

the alternative hypothesis was "The probability of getting a head when you toss the coin is less than 0.5".

If we performed the experiment and got three heads in ten tosses, which is likely (check out the cumulative probability table in Section 2.2 to convince yourself of this!), then the null hypothesis was supported, and we reject the alternative hypothesis. This is just code for "it was likely to get three heads in ten tosses, so that there is nothing to indicate that the probability of tossing a head with this coin is anything other than 0.5".

A Another Way of Looking at Hypothesis Tests...

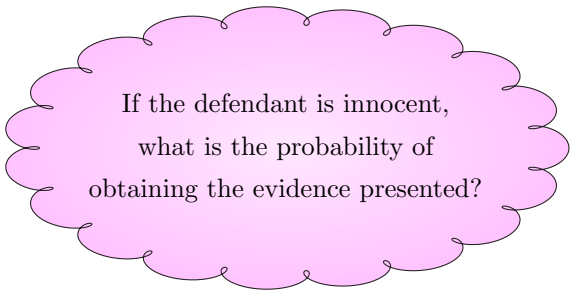
The majority of this introduction comes directly from the Edexcel-endorsed S2 book, Ref (Attwood et al., 2000).

A.1 The Concept and Interpretation of a Hypothesis Test

The scene is a courtroom. The defendant is in the dock and is accused of committing murder. The prosecution and defence counsels will both present evidence and the judge and jury have to reach a verdict. Two important principles operate under the system of British law and they are:

- the defendant is "innocent until proved guilty", ***and***
- the proof must be "beyond all reasonable doubt".

Clearly the defendant either did or did not commit the murder. But during the course of the trial the assumption the judge and jury must make is that he did not (i.e. that he is innocent). They must then examine the evidence and essentially answer the following question:



If this probability is very small then they will conclude that the assumption of innocence is not sustainable and declare the defendant guilty. For example, if the defendant's fingerprints were found on the gun that shot the victim and the defendant was seen leaving the scene shortly after the time of death then you might think that the probability of these events happening and the defendant being innocent is quite small. Followers of TV detective dramas will know that it may well not be sufficiently small to secure a conviction though!

The above situation is very similar to the processes involved in carrying out a ***hypothesis test***. The role of the defendant is played by the ***hypotheses***. We start with a basic assumption called the ***null hypothesis*** (this is equivalent to the defendant being innocent), which is assumed to be true. We also specify the ***alternative hypothesis*** which describes the situation if the null hypothesis is not true (in the above situation it would be that the defendant was guilty and did commit the murder). In a statistical hypothesis test the evidence comes from a ***sample***. This sample is summarised in the form of a statistic called a ***test statistic*** and by assuming the null hypothesis to be true it should be possible to calculate probabilities relating to this test statistic.

At this point another feature of the courtroom scenario is worth mentioning. Suppose that at the end of the trial the evidence presented is such that the judge and jury could decide that the defendant is guilty. Further evidence detrimental to the defendant could still be produced but a certain ***threshold*** has already been crossed. The probability of obtaining evidence as bad as this or worse is sufficiently small to cause the judge and jury to reject the defendant's innocence. The phrase ***"as bad or worse"*** is sometimes helpful. We calculate the probability of obtaining evidence as bad or worse as that which we have been presented with to make our judgement.

We said that the judge and jury must assess the evidence and attempt to estimate the probability of obtaining evidence "as bad or worse" as that presented if the defendant is innocent. If that probability is very small they would reject the assumption of innocence, but how small is "very small"? Clearly there is a threshold probability and it may vary, depending on the nature of the problem. In the context of a hypothesis test we call this threshold probability the ***significance level***.

The significance level is the level of probability that we call unlikely. If your test gives a probability as unlikely as the significance level then you reject the null hypothesis. The usual significance level is 5% or 0.05 but other levels such as 1% (0.01) and 10% (0.10) are often used.

If the probability of a value of the test statistic
"as bad or worse" as that obtained is p,
then we reject the null hypothesis if p is less than
or equal to the significance level α.

A.2 One and Two-Tailed Tests

One rainy day during the summer holidays a family of four were playing a simple game of cards. The game was one of chance so the probability of any particular person winning should be $\frac{1}{4}$. After playing a number of games Steve complained that his younger sister Sarah must have been cheating as she kept winning. Their parents quickly intervened and decided to carry out a proper investigation.

Now Sarah may be cheating, but if she is not then the probability of her winning should be $\frac{1}{4}$, or the proportion, p, of games that Sarah wins is $\frac{1}{4}$. The test that the parents wish to carry out is about this proportion, p, of games that Sarah wins. It is a general feature of hypothesis tests that they are about the value of unknown population parameters. The defendant in this case is Sarah and her claim that $p = \frac{1}{4}$. We assume that Sarah is innocent and wish to formulate a null hypothesis to express this idea in terms of the parameter p. We usually write H_0 for the null hypothesis and in this case you have:

$$H_0 \ : \ p \ = \frac{1}{4}$$

lf Sarah is guilty then the proportion of games that she wins must be more than $\frac{1}{4}$, (the complaint would probably not have arisen if she had not appeared to be winning more than her fair share of the games) and we write the alternative hypothesis, H_1, as

$$H_1 \ : \ p \ > \frac{1}{4}$$

This specification gives you a ***one-tailed test***, since you are only considering deviations of p in one direction, namely $p > \frac{1}{4}$. Sarah's parents may have been interested in checking the hypothesis $p = \frac{1}{4}$ against the alternative hypothesis $p \neq \frac{1}{4}$. They could be concerned if she won very few games: maybe she had not understood the rules and this left her at a disadvantage, or she could be cheating and thus winning more games than expected. In this situation you would specify the alternative hypothesis as

$$H_1 \ : \ p \ \neq \frac{1}{4}$$

and this is called a ***two-tailed test*** since you are considering deviations of p in two directions.

Once the null and alternative hypotheses have been specified we need some procedure to decide between these two opposing hypotheses and the one that we use is called a ***hypothesis test***.

A hypothesis test about a population parameter θ
tests a null hypothesis H_0, specifying a particular value for θ,
against an alternative hypothesis H_1, which will
indicate whether the test is one-tailed or two-tailed.

The parents needed some evidence upon which to base their judgement and they examined a random sample of 10 games and discovered that Sarah had won 5 times. They then had to calculate the probability of obtaining evidence "as bad or worse" than this, assuming that the null hypothesis is true. What sort of evidence would be "as bad or worse" than that which their sample gave? The alternative hypothesis is that

Sarah is cheating and that $p > \frac{1}{4}$; the sample saw Sarah winning 5 out of 10 games, a proportion of $\frac{1}{4}$, so if Sarah won 5 or more games that would constitute evidence "as bad or worse".

If we let the random variable X represent the number of times that Sarah wins in a sample of 10 games then, if the null hypothesis is true the probability that $X \geq 5 = 0.0781$ [This is obtained from a calculation using the Binomial Distribution. Trust me!]

This probability (about 7.8%) is reasonably large (it is certainly more than 5%, which is the usual significance level) so there is no reason to suspect the validity of H_0, and Sarah remains innocent. We usually say that the sample is ***not significant*** and that there is insufficient evidence to reject the null hypothesis that $p = \frac{1}{4}$.

Notice that the test was based simply on the statistic X and this is the ***test statistic*** in this case. Its sampling distribution is the binomial distribution.

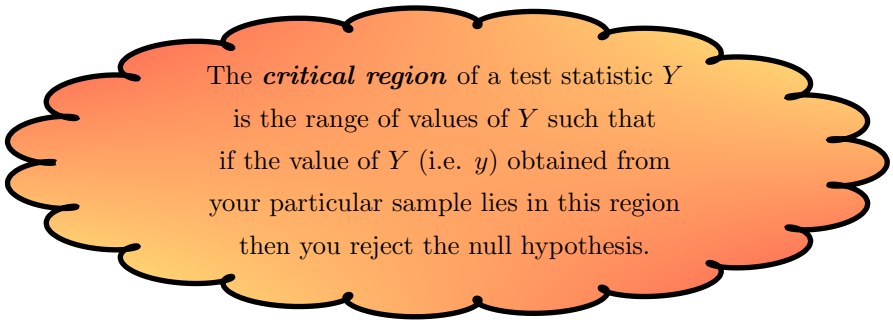
In this case it is certainly clear that Sarah has been quite lucky; 5 wins out of 10 is twice the number you might expect (out of 10 games she should win a quarter of them, i.e. 2.5 games) but it is not so unusual that you would suspect foul play.

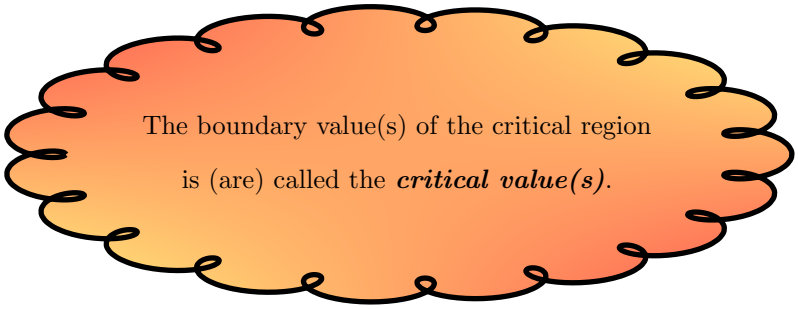
A.3 The Critical Region

It is sometimes helpful to consider what value x of the test statistic you would have needed before the parents might consider that Sarah was cheating. In other words, what value of x would provide sufficient evidence to reject the claim that $p = \frac{1}{4}$? If you use a 5% significance level then you require the value of c such that: $\mathrm{p}(\mathrm{X} \geq \mathrm{c}) \leq 0.05$ In this case (X is binomially distributed with 10 trial and $\mathrm{p} = \frac{1}{4}$), it turns out that $\mathrm{p}(\mathrm{X} \geq 6) = 0.0197$

So if Sarah had won 6 or more games out of the sample of 10 we would have had a ***significant*** result and rejected H_0. So any value $X \geq 6$ would mean that the probability of obtaining a sample "as bad or worse" is less than or equal to 5%, which is unlikely. This means that the assumption that H_0 is true is called into question and ***we reject*** H_0 ***at the 5% level of significance***. Notice that although we used a 5% level of significance, the probability of rejecting H_0 is only 0.0197, i.e. 1.97%, which is quite a bit less than the 5% we aimed for. This will often happen when our test statistic follows a discrete distribution, but when hypothesis tests are based on a continuous variable such as the normal distribution this situation will not arise.

We call the region $\mathrm{X} \geq 6$ the ***critical region*** of the statistic X and the value 6 is called the ***critical value***.





References

Attwood, G., Dyer, G. and Skipworth, G. (2000). *Statistics 2*. Heinemann, first edition.

Smith, S. (2013). Maths Notes : The Binomial Distribution. Calculating probabilities of outcomes of binomial events.