

I find a penny on the sidewalk. I am curious about the probability that this penny will come up heads when it is tossed. (You can tell I am a mathematician!) I believe that it is the case that this penny will come up heads with probability $1/2$. How do I proceed to test this belief?

1. A MODEL FOR TOSSING A PENNY

First, I adopt the model for tossing a penny in which the outcomes on each toss of then penny (heads or tails) are independent, and the probability of a head on any one toss is always the same.

In this model, if p denotes the probability of a head on any one toss, and X represents the number of heads in N tosses, then

$$\Pr(X = k) = \binom{N}{k} p^k (1-p)^{N-k},$$

the familiar binomial distribution.

2. NULL AND ALTERNATIVE HYPOTHESES

My hypothesis is that the coin will come up heads with probability $1/2$. The cornerstone of hypothesis testing is that I am looking for evidence that will cause me to discard this working hypothesis in favor of some other hypothesis, and that this evidence must be pretty strong. The term **null hypothesis** refers to my working hypothesis. In this case, we would say that, in terms of our binomial model, the null hypothesis is that $p = 1/2$. There are many other hypotheses which might take its place. Any of these is called an **alternative hypothesis**. Let us take the most obvious alternative hypothesis, that $p \neq 1/2$. (Of course, in any event, $p \in [0, 1]$.)

3. FOUR POSSIBILITIES

No matter what experiment we carry out, and no matter what conclusion we draw, there will be four possibilities, two good and two bad:

- We may conclude that the null hypothesis is true when it is true.
- We may conclude that the null hypothesis is false when it is false.
- We may conclude that the null hypothesis is false when it is true. This is called a type I error.
- We may conclude that the null hypothesis is true when it is false. This is called a type II error.

The object is to have a procedure in which the probabilities of type I and type II errors are as small as possible.

4. AN EXAMPLE

We toss the coin 10 times, and count the number of heads. If it is 4, 5 or 6 we conclude that the null hypothesis, $p = 1/2$ is true. Else we conclude that the alternative hypothesis is true.

The probability of type I error is

$$1 - \left(\binom{10}{4} \left(\frac{1}{2}\right)^{10} + \binom{10}{5} \left(\frac{1}{2}\right)^{10} + \binom{10}{6} \left(\frac{1}{2}\right)^{10} \right) = .34375$$

while the probability of type II error depends on p :

$$\sum_{k=4}^6 \binom{10}{k} p^k (1-p)^{10-k} = 42p^4(4p^2 - 4p + 5)(p-1)^4$$

which reaches a maximum value of 0.65635 when $p = 1/2$.

5. THE TYPICAL PROCEDURE, AND ITS INHERENT BIAS

Tests are always constructed to first minimize the probability of type I error. We choose a set of outcomes, A , called the **acceptance region**, and if we observe outcomes in this region, we accept the null hypothesis. The complement of the acceptance region is called the **rejection region**. Typically, a threshold is set and the probability of type I error must fall below this threshold. For example, it might be set at 0.05. The test in the previous section does not meet this criterion, so we must reduce the rejection region (so that it has lower probability). For example, if we accept the null hypothesis when the number of heads is among $\{3, 4, 5, 6, 7\}$ then the probability of type I error is

$$1 - \sum_{k=3}^7 \binom{10}{k} \left(\frac{1}{2}\right)^{10} = .109375,$$

which is still not small enough, but taking the acceptance region to be $\{2, 3, 4, 5, 6, 7, 8\}$ gives a probability of type I error of

$$1 - \sum_{k=3}^7 \binom{10}{k} \left(\frac{1}{2}\right)^{10} = .021484375$$

which is small enough. In other words, we will only reject the null hypothesis if 0, 1, 9 or 10 heads are obtained. We should then be concerned about the probability of type II error. Indeed, it is given by

$$\sum_{k=2}^8 \binom{10}{k} p^k (1-p)^{10-k} = 3p^2(6p^6 - 18p^5 + 63p^4 - 96p^3 + 95p^2 - 50p + 15)(p-1)^2$$

which reaches a maximum value of .978515625 when $p = 1/2$. What is more, it is only less than 0.05 if $p < 0.0365$ or if $p > 0.9635$. In other words, if p is not $1/2$ but is pretty close, we are almost sure to make a type II error. What to do?

6. MORE TRIALS!!!

Our experience with confidence intervals tells us the more trial should help. Indeed, if we toss the coin 40 times we anticipate a 50% improvement in some way or another. So we experiment. If the null hypothesis is true ($p = 1/2$), then

$$1 - \Pr(15 \leq X \leq 25) = 1 - \sum_{k=15}^{25} \binom{40}{k} \left(\frac{1}{2}\right)^{40} \approx .0806904678,$$

which is close, but not quite what we want. After a little fiddling around:

$$1 - \Pr(14 \leq X \leq 26) = 1 - \sum_{k=14}^{26} \binom{40}{k} \left(\frac{1}{2}\right)^{40} \approx .0384773083$$

, so our acceptance region is $\{14, 15, \dots, 25, 26\}$, which is as good as we can get if we want a symmetric interval. Since our alternative hypothesis is symmetric, this would seem to be a good idea.

What then of the type II error? The probability of type II error still depends on $p \neq 1/2$ and is

$$\Pr(14 \leq X \leq 26) = \sum_{k=14}^{26} \binom{10}{4} p^k (1-p)^{10-k}$$

which is less than 0.05 if $p < 0.225$ or if $p > 0.775$. This is clearly better. In fact, if we up the number of trials to 160, and we base the test on the acceptance region $\{68, 69, \dots, 91, 92\}$ then the probability of making a type I error is about .0477669960, and the probability of a type II error is less than 0.05 if $p < 0.359$ or $p > 0.641$.

When we tossed the coin 10 times, the acceptance region contained 64% of the possible outcomes. When we increased the number of tosses by a factor of 4, we reduced the number of outcomes in the acceptance region to 32% of the total, and when we increased by another factor of 4, we reduced to 16% of the total. Now we see what is halved!