MAT220 Lecture 14: Introduction to Hypothesis Testing

Contents

In our recent lectures we have studied estimation, which uses sample data to guess roughly where a population parameter is. We will now continue our discussion of inferential statistics by performing *tests of claims*. Instead of estimating a parameter, we will now evaluate claims of a parameter's true value. Such a process is called a *hypothesis test*, and it is a powerful tool in inferential statistics. Throughout the next four lectures we will discuss this technique in detail. Today we will simply introduce the overall structure of a hypothesis test.

1 Introduction

Suppose that you are on a walk, and you come across two groups of people. One of these groups looks dangerous and suspicious, and the other is a group of trustworthy-looking nuns and children:

A suspicious group

A trustworthy group

You walk over to the shady-looking group first. They hand you a coin and propose the following game:

> Keep flipping the coin. Every time you land on heads, we will give you \$2. But for every tails you land, you have to give us \$1.

At first this feels generous, so you decide to play. But after 30 flips, you notice that you only landed on heads 3 times, and got 27 tails. The net total is that you have to pay out \$21, and you go about your business.

Feeling suspicious, you decide to do some calculations: since coin flips are a binomial experiment, you use the formula:

$$P(X = 3) = C_{30,3}(0.5)^3(0.5)^{27} \approx 3.78 \times 10^{-6},$$

where here you assume that the coin is fair with $p = q = 0.5$. According to the above, you notice that the probability of getting 3 heads in 30 attempts is awfully small, and therefore you conclude that probably the coin was a scam, with p much lower than 50%.

You now walk over to the trustworthy-looking group, and start talking to them. Weirdly, this second group offers you the exact same game:

> Keep flipping the coin. For every head, we will give you \$2. But for every tail, you have to give us \$1.

Against your better judgement, you decide to play again for another 30 coin flips. In this case, you manage to get 12 heads and 18 tails, and therefore you win \$6. Again assuming that the coin is fair, you calculate:

$$P(X = 12) = C_{30,12}(0.5)^{12}(0.5)^{18} = C_{30,12}(0.5)^{30} \approx 0.081.$$

In fact, this is not too unreasonable, given that the chance of getting 15 heads in 30 flips is only about 14%. From this information alone, there is no good reason to think that the second coin was unfair.

1.1 The informal structure of a statistical test

We have just seen the informal idea behind a very important statistical procedure known as *hypothesis testing*. In both situations we were presented with a system and challenged an assumption that we were working with a fair coin, based on some observations that we were able to make. In both cases we made some sort of probability calculation, and drew two different conclusions:

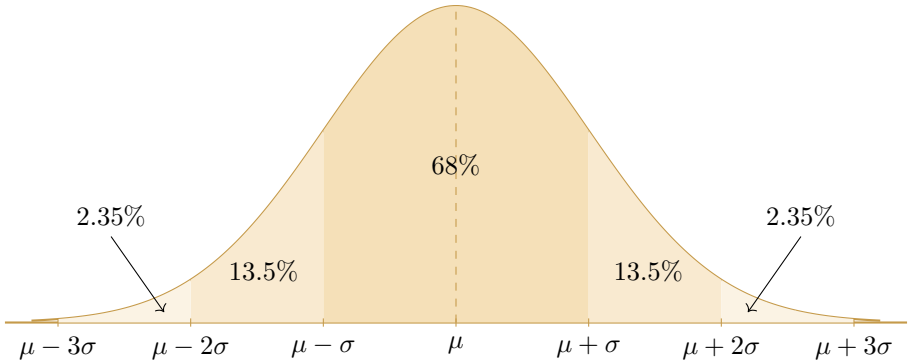
1. in the case of the suspicious coin, we concluded that our random observation was *extremely unlikely*, which cast doubt on the expectation that $p = 0.5$, causing us to reject the idea of a fair coin, whereas
2. in the case of the nun's coin, our random observation was consistent with the probability $p = 0.5$, therefore giving us no reason to reject the fairness of the coin.

Lots of hypothesis tests follow this general pattern: we have an expectation, we make an observation which drives a suspicion, and then we perform a probability calculation that informs us of how reasonable it would be to reject our original assumption. This conceptual workflow can be summarized as follows:

Throughout the next few lectures, we will explore this structure in various contexts. For now, we will simply describe the general structure. In order to do that, we will first recall some basic properties of the normal distribution.

1.2 A Brief Review of the Normal Distribution

Recall that a normal distribution is defined by two parameters: the population mean μ and the population variance σ^2. When we write $X \sim \mathcal{N}(\mu, \sigma^2)$, the x-axis of the distribution represents the possible values of the random variable X. Since this is a continuous random variable, the probability of finding X in some region is given by the area under the graph. The shape of the graph therefore encodes the likelihood of finding X nearby. Since a normal distribution will peak around its centre and then run off quickly into its tails, we would expect that randomly selecting a value of X will probably land us close to μ. In fact, based on our discussion from Lecture 10, there is about a 68% chance of landing within one standard deviation of μ:

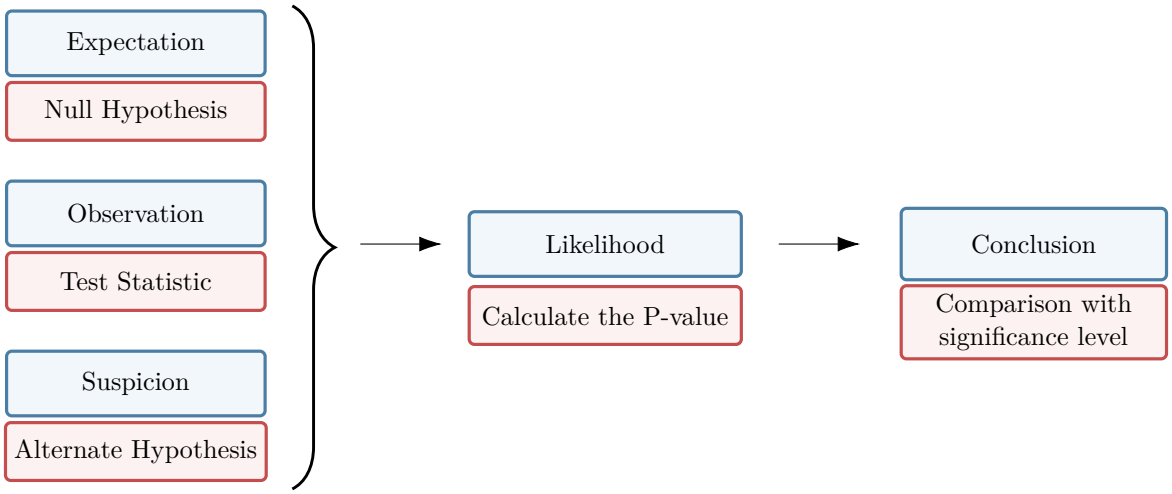


In other words, if we were to keep randomly selecting a value of X, then this value would be in the interval $[\mu - \sigma, \mu + \sigma]$ about 68% of the time. In contrast, if we were to randomly select a value of X, then it would be very unlikely that the value is far from the centre. This observation underpins the concept of hypothesis tests.

2 Hypothesis Testing

2.1 The general layout of a statistical test

In the introduction, we described a general workflow for performing a statistical test. Written below is the same workflow, decorated with the correct technical terms.



We will now explain these terms one-by-one.

Important Terminology

- **Null hypothesis (H_0):** the statement that we want to test.
- **Alternate hypothesis (H_1):** the competing statement that reflects our suspicion.
- **Test statistic:** the numerical information obtained from sample data.
- **P-value:** the probability of getting a result at least as extreme as the observation, assuming H_0 is true.
- **Significance level (α):** the cut-off probability that tells us what counts as "too unlikely".

2.2 The P-value

According to our definition above, the P-value of a hypothesis test does not simply calculate the probability of an outcome happening. Instead, the P-value calculates the chances of observing our sample data *or worse* whilst assuming the null hypothesis is true. We saw prototypes of this idea in Section 1, where we performed binomial calculations for the two coins. In fact, these probability calculations do not yet represent P-values of those hypothesis tests. Instead, we would need to make more detailed calculations.

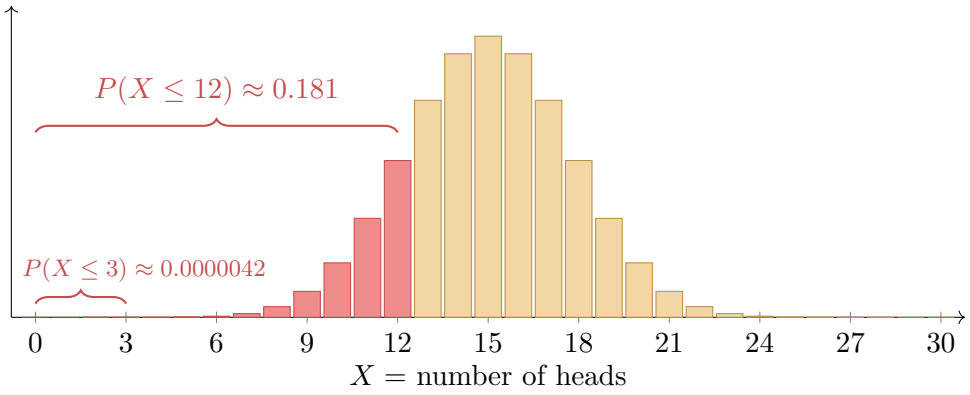
For the suspicious coin, the exact observation was $X = 3$, i.e. 3 heads. Here, the P-value of this observation assumes that the coin is fair, i.e. $p = 0.5$, and then calculates the chance of getting 3 heads *or less*:

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \approx 0.0000042.$$

For the second coin, we observed $X = 12$. Again, the P-value is calculated by computing the probability of getting 12 *or less* heads:

$$P(X \leq 12) = P(X = 0) + P(X = 1) + \ldots + P(X = 12) \approx 0.181.$$

Visually, we are looking to compute the probability found in the left tail of the binomial distribution built from $n = 30$ and $p = 0.5$:



2.3 The Significance Level

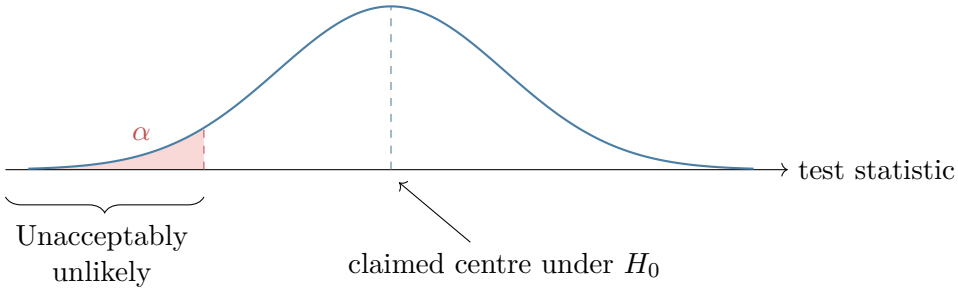
Hypothesis tests assume that the null hypothesis is actually *true*, and then they explore the consequences of this assumption. As we mentioned in Section 1.2, under most circumstances, we would

expect a randomly selected test statistic to land relatively close to the centre of its probability distribution – especially when it is normally distributed. If we were to obtain an observation very far into a tail, then this observation has a very low probability under the original model.

Of course, by the nature of randomness it is always *possible* to obtain an unusually extreme result by accident. Really, we need to ask:

> How unlucky do we need to be before we start to think that something is wrong with our null hypothesis?

The significance level α serves to answer this question by giving a "cut-off" value for luck:



Once we have sample data in hand, the P-value calculates the chance of this particular sample *or worse* happening, assuming that the true centre of the data is given by H_0. If this probability is unacceptably low, i.e. less than α, then we conclude that the value of the population parameter claimed by H_0 is probably wrong. The final judgement can be summarised as follows.

Decision Rule

- If the calculated P-value is **less than or equal to** α, then the observation is too unlikely under H_0, so we **reject** H_0.
- If the calculated P-value is **greater than** α, then the observation is not unlikely enough, so we **do not reject** H_0.

In practice, we choose the significance level at the start of the test. Common values are

$$\alpha = 0.01, \qquad \alpha = 0.02, \qquad \alpha = 0.05.$$

Importantly, "do not reject H_0" does *not* mean that we have proved H_0 is true. It only means that the sample does not provide strong enough evidence against it.

2.4 Different types of tests

The alternate hypothesis H_1 is always a statement to the contrary of the null hypothesis. As a matter of fact, there are many ways to disagree with an "is" statement. Typically, the null hypothesis will say that the true value of a population parameter equals a particular claimed value. There are three ways that this could be wrong:

- the true value could be less than the claimed value,

- the true value could be more than the claimed value, or
- the true value could be simply *different from* the claimed value.

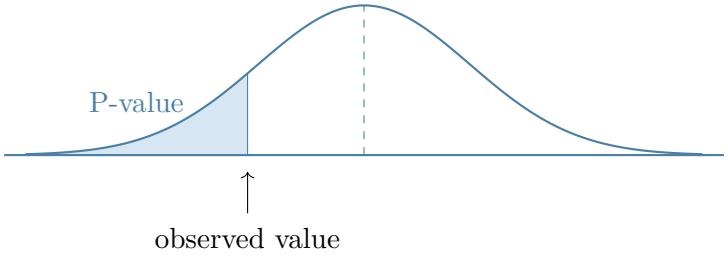
These are called left-tailed, right-tailed and two-tailed tests, respectively. We will now demonstrate how the alternate hypothesis tells us how to compute the correct P-value for a given test.

2.4.1 Left-tailed tests

A left-tailed test asks whether the true value is less than the value claimed by the null hypothesis. Formally, the test would start by writing something like:

$$H_0 : \text{true value} = \text{claimed value}, \qquad H_1 : \text{true value} < \text{claimed value}.$$

As the name may suggest, in a left-tailed test we calculate the P-value of the observed data by taking the area to the *left*:

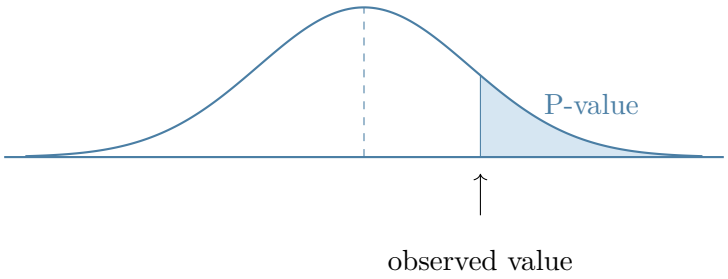


2.4.2 Right-tailed tests

In a right-tailed test, the alternate hypothesis asks whether the true value is greater than the value claimed by the null hypothesis:

$$H_0 : \text{true value} = \text{claimed value}, \qquad H_1 : \text{true value} > \text{claimed value}.$$

In this case, the P-value is the area to the *right* of the observed sample's statistic:

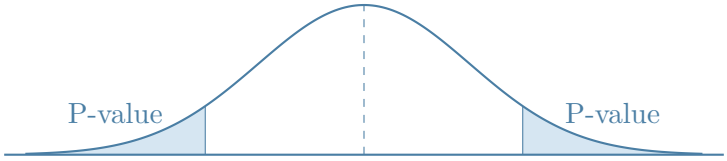


2.4.3 Two-tailed tests

For a two-tailed test, we ask whether the true value is merely different from the value claimed by the null hypothesis:

$$H_0 : \text{true value} = \text{claimed value}, \qquad H_1 : \text{true value} \neq \text{claimed value}.$$

Recall that if two numbers a and b are different from each other, then either $a < b$ or $b < a$. As such, a two-tailed test computes the P-value by combining the probabilities from both tails at the same time:



2.5 Type 1 and Type 2 errors

By their very nature, hypothesis tests always contain some amount of uncertainty – if we find a P-value lower than the significance level, then this does not *prove* that the null hypothesis is false. Instead, it only means that the sample provides evidence against the null hypothesis. Similarly, if we do not reject the null hypothesis, this does not prove that it is genuinely true. Simply, it means that there is not enough evidence to reject it.

Because of this inherent uncertainty, it is always possible to draw the wrong conclusion by accident. There are two main ways that we could make a mistake when performing a hypothesis test. These are known as Type 1 and Type 2 errors.

Type 1 and Type 2 Errors

- A **Type 1 Error** happens when we reject H_0 even though H_0 is actually true.
- A **Type 2 Error** happens when we do not reject H_0 even though H_0 is actually false.

Type 1 errors can occur if we happen to find a statistically unlikely sample on the first try. For example, technically it is not *impossible* to flip a coin 30 times and to get 27 tails and 3 heads. However, it is extremely *unlikely*, and therefore we are compelled to reject the claim that such a coin is fair. However, when doing so, there is no guarantee that we are simply not living in an unlikely world. Rejecting the null hypothesis in a situation like that could be an example of a Type 1 error.

Type 2 errors occur when we take sample data that is consistent with the null hypothesis. In such a situation, we have no good reason to reject the null hypothesis, even if it might be genuinely false in reality. This can happen if our sample observation gives a P-value above α. To illustrate this, recall that the nun's coin gave 12 heads in 30 attempts. We concluded that this observation is reasonably consistent with a fair coin and therefore there is no reason to reject the claim that the coin is fair. However, what if the coin happened to have a probability of $P(H) = 0.4$ in reality? Our sample data would not exclude the possibility of a fair coin, even though it could be false.

There are four possible outcomes that can result from a hypothesis test:

	Do not reject H_0	**Reject H_0**
H_0 is true	Correct decision	Type 1 Error
H_0 is false	Type 2 Error	Correct decision

The significance level α controls the probability of making a Type 1 Error. For example, using $\alpha = 0.05$ means that the test is designed to reject a true null hypothesis about 5% of the time in repeated use.

3 A Sample Test: The Two Coins

We will now demonstrate how to express our coin flip examples as genuine hypothesis tests. For future convenience, we will actually phrase the experiments in terms of normal approximations using the random variable $\hat{P}$, which we saw in previous lectures when creating estimates for the true binomial probability p. In what follows we will simply state the P-values associated to our sample experiments, and we will defer the actual calculations of P-values to next lecture. For now, it is enough to recall that the random variable $\hat{P}$ follows an approximate normal distribution with parameters:

$$\hat{P} \sim \mathcal{N}\left(p, \frac{pq}{n}\right).$$

Importantly, the true binomial probability p lies at the centre of this distribution, and therefore we can visualize our null hypotheses as claims about the middle of this normal distribution. Our observed sample data will then eventually manifest as points on the x-axis of this distribution. In other words, the above normal distribution tells us how to *structure our luck*.

Before getting to visualizations of our observed data, we will first demonstrate how to properly phrase our tests.

3.1 The suspicious coin

We now return to the first coin.

Step 1: state the hypotheses.

We begin by assuming that the coin is fair, but we suspect that the probability of landing on Heads is actually lower than 0.5. Therefore, our null and alternate hypotheses become:

$$H_0 : p = 0.5, \qquad \text{and} \qquad H_1 : p < 0.5,$$

respectively. Notice that according to our alternate hypothesis, this is a left-tailed test.

Step 2: record the test statistic.

According to our sample data, we observed $r = 3$ heads in $n = 30$ coin flips. This gives a sample proportion of:

$$\hat{p} = \frac{r}{n} = \frac{3}{30} = 0.1.$$

Step 3: calculate the P-value.

Ordinarily here is where we could calculate the P-value directly. However, we will simply state the calculation according to our normal approximation:

$$\hat{P} \sim \mathcal{N}\left(0.5, \frac{(0.5)(0.5)}{30}\right).$$

Therefore,

$$P(\hat{P} \leq 0.1) \approx 0.0000134.$$

Note that we compute the values to the left of the observed $\hat{p}$, because we are working with a left-tailed test.

Step 4: compare with the significance level.

Using a level of significance $\alpha = 0.05$, we have, quite obviously:

$$0.0000059 < 0.05.$$

Therefore, we are compelled to reject the null hypothesis H_0.

Interpretation

Our sample data of 3 heads in 30 coin flips provides extremely strong evidence against the hypothesis that the coin is fair. Instead, the data supports the suspicion that the probability of heads is less than 0.5.

3.2 The nun's coin

We will now perform the same test for the second coin.

Step 1: state the hypotheses.

Again, we are going to test whether the coin is biased against heads:

$$H_0 : p = 0.5, \qquad \text{and} \qquad H_1 : p < 0.5.$$

Step 2: record the test statistic.

In this case, we observed $r = 12$ heads in $n = 30$ flips, so

$$\hat{p} = \frac{r}{n} = \frac{12}{30} = 0.4.$$

Step 3: calculate the P-value.

Under the null hypothesis, we apply a continuity correction and calculate the P-value to be approximately

$$P(X \leq 12) \approx P(Z \leq -0.913) \approx 0.181.$$

Step 4: compare with the significance level.

Again using a significance level of $\alpha = 0.05$, we now see that:

$$0.181 > 0.05.$$

Therefore, we do not reject H_0.

Interpretation

Our sample of 12 heads in 30 coin flips does not provide strong enough evidence to conclude that the true probability of heads is less than 0.5. This does not prove that the coin is fair; it only means that the observation is reasonably consistent with a fair coin.

3.3 The tests visualized

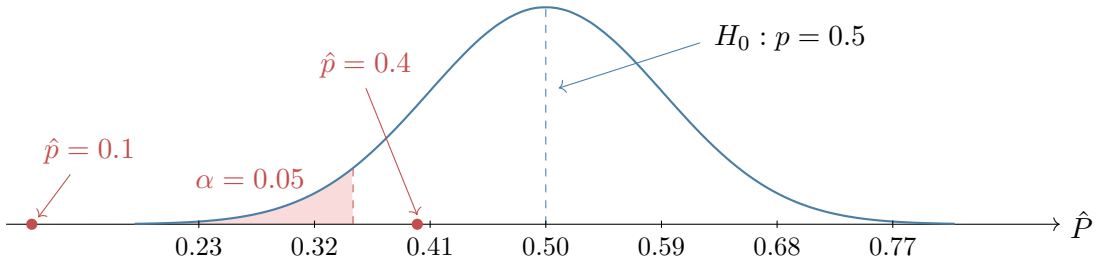
Under the null hypothesis, the sampling distribution of the observed proportion is approximately

$$\hat{P} \sim \mathcal{N}\left(0.5, \frac{(0.5)(0.5)}{30}\right).$$

The standard deviation of this distribution is

$$\sqrt{\frac{(0.5)(0.5)}{30}} \approx 0.091.$$

The suspicious coin produced $\hat{p} = 0.1$, which is extremely far into the left tail. The nun's coin produced $\hat{p} = 0.4$, which is below the centre but not especially extreme.



As you can see, the left-tailed P-value associated to the nun's coin will give an area greater than the threshold $\alpha = 0.05$, and therefore we have no good reason to reject the hypothesis that that coin is fair. However, in the case of the suspicious coin, the observed sample proportion $\hat{p} = 0.1$ is far into the left-tail of the distribution, meaning that it is extremely unlikely to occur on our first attempt at getting sample data. The area to the left of 0.1 is far smaller than 0.05, and therefore we rejected the hypothesis that the suspicious coin was fair.