MAT120: Lecture 15 Handout

Hypothesis Testing

---

1 Introduction **2**

2 A Brief Review of the Normal Distribution **3**

2.1 Probability as area under the curve 4

2.2 The standard normal distribution 4

2.3 Converting other normal distributions into the standard normal 5

3 The Central Limit Theorem **6**

3.1 The sampling distribution of the sample mean 6

3.2 The theorem itself 6

3.3 Why this matters 7

4 Hypothesis Testing **8**

4.1 The general layout of a statistical test 8

4.2 Summary of Terminology 9

4.3 Examples 9

4.4 The philosophy of testing the population mean 10

4.5 Type 1 and Type 2 Errors 11

5 Testing Binomial Experiments **12**

5.1 From binomial experiments to the Central Limit Theorem 12

5.2 The sampling distribution for a sample proportion 13

5.3 A full test: the suspicious coin 13

5.4 An important condition before using the normal approximation 14

6 Examples and Exercises **15**

6.1 The Haribo data 15

6.2 A picture of the Haribo test 16

6.3 Worked exercise 17

Last lecture we studied the normal distribution, the standard normal distribution, and how z-scores let us move between them. In this lecture we use that normal-distribution machinery, together with the Central Limit Theorem, to make decisions using data. The big idea is *hypothesis testing*, which is a formal way to decide whether an observation is so unlikely that we should begin to doubt our

original assumption. We will focus on hypothesis tests for binomial experiments, such as coin flips and candy draws.

1 Introduction

Suppose that a shady-looking group of people hand you a coin and propose the following game:

> Keep flipping the coin. For every head, we will give you \$2. But for every tail, 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 got heads 3 times, and got 27 tails. The net total is that you have to pay the shady group \$21, and you go about your business.

After getting home, you decide to do a calculation and figure out the chances of this unlikely situation occurring. Since coin flips are a binomial experiment, you perform the calculation:

$$P(\text{3 heads in 30 flips}) = C_{30,3}(0.5)^3(0.5)^{27} \approx 3.78 \times 10^{-6}.$$

where here $n = 30$ and you have assumed that the coin was fair, giving $p = 0.5$. This is the probability of this exact outcome, not yet the P-value of a full hypothesis test. The probability is extraordinarily small, which gives us very strong reason to doubt that the coin was fair, and to suspect that the probability of landing a heads is much smaller than 0.5.

Now suppose that the next day, you go for a walk and meet another group of people. This group does not seem as shady as the ones from the day before, so you decide to talk to them. Weirdly, this second group offers you the exact same game to play. They say:

> 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 are surprised to find that you get heads 12 times out of the 30 attempts, and therefore you win \$6.

Again assuming that the coin is fair, you decide to calculate the chances of this occurring:

$$P(\text{12 heads in 30 flips}) = C_{30,12}(0.5)^{12}(0.5)^{18} = C_{30,12}(0.5)^{30} \approx 0.081.$$

So the exact probability of this outcome is about 8%. As a matter of fact, this is not an unreasonable number, given that the expected number of 15 heads in 30 attempts only has a probability of 14% anyways. From this information, there is no good reason to think that you were playing with an unfair coin.

Hypothesis Testing

We have just seen the informal idea behind a very important statistical procedure: hypothesis testing. In both cases, we were presented with a system (in this case, a Binomial system involving coin flips), and we *challenged* an assumption that we were working with a fair coin, based on the small observations that we were able to make. In the first case, we were able to make an *inference* about the coin based on the wildly unlikely scenario we found ourselves in, and in the second case, we found no such evidence to reject our initial assumptions. As a matter of fact, lots of hypothesis tests follow this 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:

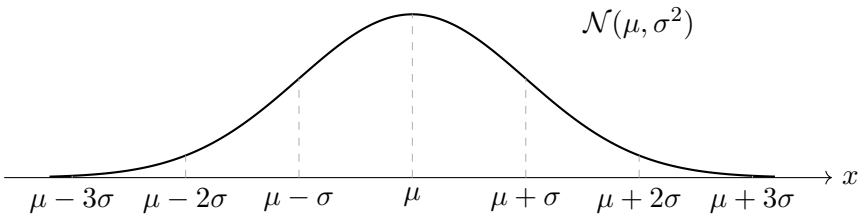
In this lecture we will explore this process in detail by building up the machinery needed to test binomial experiments. In reality, we need a little bit more mathematics than simply the binomial distribution, because the probabilities associated with large systems are very difficult to calculate. In fact, the binomial number $C_{30,12}$ from before involves the computation of numbers up to 30!, which is remarkably large. In practice, it is easier to upgrade our system to a continuous random variable, so that normal distributions can be used freely. The associated probability calculation also changes: instead of trying to compute the probability of an individual outcome occurring, we compute the probability of a given outcome or something *more extreme*. Therefore, to begin with, we will review some features of the normal distribution.

2 A Brief Review of the Normal Distribution

Lots of data in the real world follows the bell-shaped pattern of a normal distribution. This distribution is particularly nice because it only revolves around two key pieces of information:

- the **population mean** μ, which tells us where the center is;
- the **population standard deviation** σ, which tells us how spread out the data is.

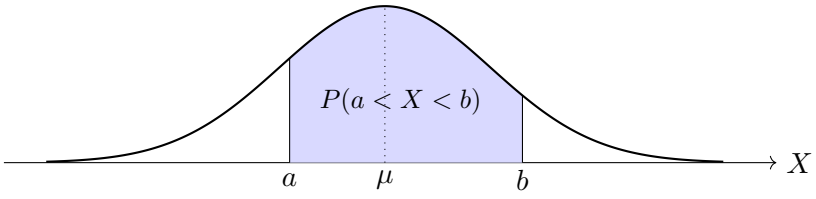
Conveniently, these two numbers are enough to uniquely characterise a normal distribution. For this reason, we write a normal distribution with the notation $\mathcal{N}(\mu, \sigma^2)$. This notation should be read something like "the normal distribution with mean μ and variance σ^2". Since the standard deviation has the same physical units as the mean, we can describe the x-axis of a normal distribution by starting in the center at μ, and walking outwards by steps of σ in both directions:



2.1 Probability as area under the curve

The x-axis of a probability distribution represents the possible values that a random variable can take. In the case of a binomial distribution this random variable is a discrete counting number, and for a normal distribution this random variable is a continuous scale. In fact, for any probability distribution built from a continuous random variable, probability is represented by *area under the curve*, instead of the height of the curve at a point.

Given a continuous random variable X, we write $P(a < X < b)$ to mean the probability that X randomly takes a value somewhere between a and b. This probability is given by the area under the curve between the points a and b. For a normal distribution, this may look like:



2.2 The standard normal distribution

Mathematically speaking, it is often difficult to compute these "areas under curves" precisely. Therefore, getting an exact description of the probability associated to a region of the x-axis can be tricky. In practice, we tend to restrict our attention to the simplest possible setting: the so-called *standard normal distribution*. We use the special letter Z to denote the random variable attached to this distribution, so we write the standard normal distribution as:

$$Z \sim \mathcal{N}(0, 1).$$

Observe that the standard normal distribution has mean equal to 0 and a standard deviation of 1.

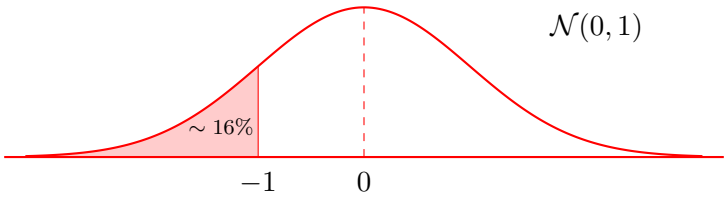
Because the relevant probabilities are difficult to calculate by hand, we often use a table of common values for

$$P(Z < z),$$

which means the probability of the random variable taking a value to the *left* of some fixed input z. This also corresponds to the area to the *left* of z. In this course, we will rely on the following table:

z	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
$P(Z < z)$	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98

Selecting a value of z gives us the associated area to the left of that value. For example, picking a value of $z = -1$ tells us that approximately 16% of the total area under the standard normal distribution exists to the left of -1:

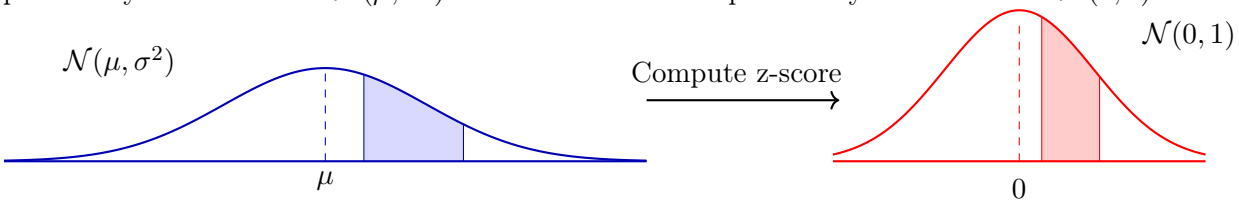


2.3 Converting other normal distributions into the standard normal

If we have any other normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$, then we can always calculate probabilities in this distribution by first converting everything to the standard normal distribution. To convert a value x into the standard normal world, we compute the *z-score*:

$$z = \frac{x - \mu}{\sigma}.$$

This formula simply tells us how far away a value x is from μ, where here distance is measured in terms of σ. This z-score formula has the useful feature of *preserving areas*, meaning that the probability calculations in $\mathcal{N}(\mu, \sigma^2)$ can be converted into probability calculations in $\mathcal{N}(0, 1)$ instead.



2.3.1 Example: heights

Suppose male heights are approximately normally distributed with mean 171 cm and standard deviation 5.6 cm. What is the probability that we meet somebody taller than 180 cm?

First compute the z-score:

$$z = \frac{180 - 171}{5.6} \approx 1.61.$$

So

$$P(X > 180) = P(Z > 1.61).$$

A standard normal table or calculator can then be used to estimate the right-tail area.

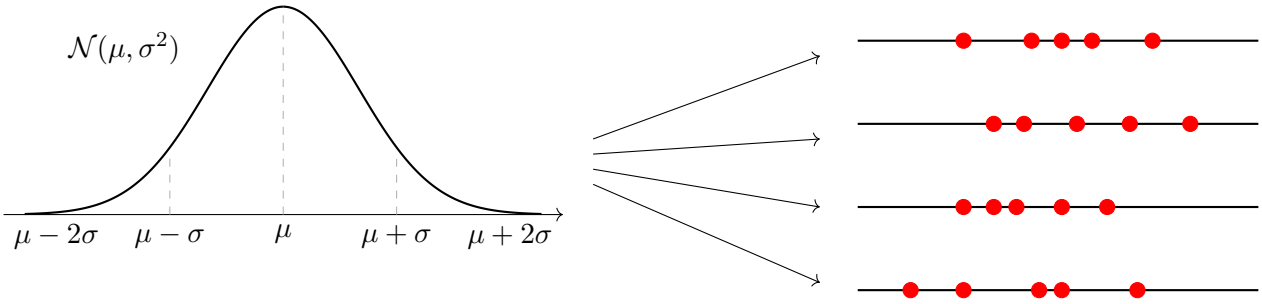
The key point is not the exact number. The key point is that we can translate a question about *any* normal distribution into a question about the single universal distribution $\mathcal{N}(0, 1)$.

3 The Central Limit Theorem

3.1 The sampling distribution of the sample mean

Suppose that we have some sort of normal distribution, i.e. $\mathcal{N}(\mu, \sigma^2)$. According to our previous discussion, if we were to randomly select a value for the variable, then this is like randomly drawing a point somewhere on the x-axis. As we just saw, the probability of finding our point within any given region is given by the *area* of that region. This means that those regions with larger areas have a higher chance of containing our randomly-chosen point.

Suppose now that we randomly pick say 5 points on the x-axis. According to the terminology of Lecture 13, this is also called a *simple random sample* of size 5. The figure below depicts several randomly chosen samples of size 5:



For each of these samples, we can compute the sample mean $\bar{x}$. In particular, we can also take the collection of *all* sample means corresponding to *all* simple random samples of size 5. Since we took our samples randomly, we can therefore also view $\bar{x}$ as a random variable in its own right. The collection of all of the $\bar{x}$'s then form their own probability distribution, where now the x-axis is written in terms of $\bar{x}$. This new probability distribution is called the *sampling distribution of the sample mean* for $n = 5$.

3.2 The theorem itself

Given that we can create a sampling distribution of $\bar{x}$'s for any continuous probability distribution, a natural question is:

> What do these sampling distributions actually look like in general?

The answer is one of the most important facts in all of elementary statistics: The Central Limit Theorem.

The Central Limit Theorem

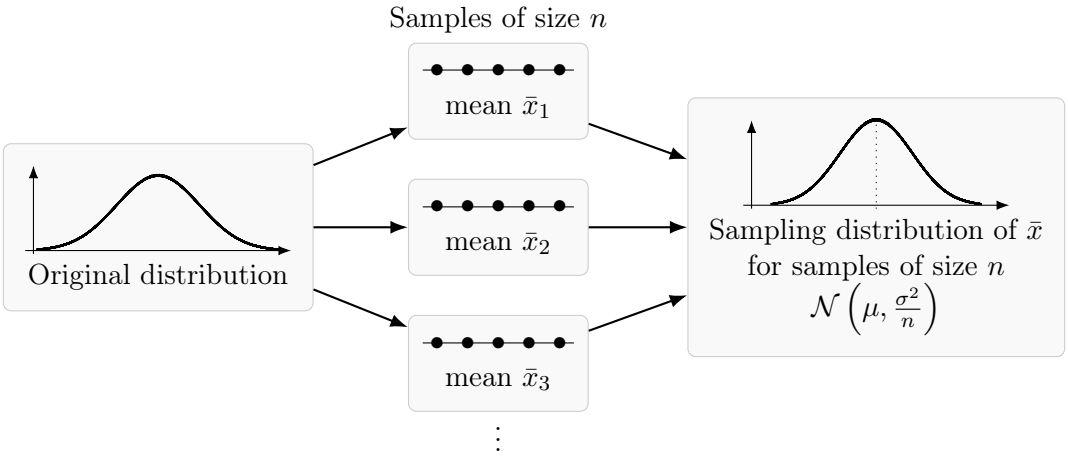
If a population has mean μ and variance σ^2, then for sufficiently large sample size n, the sampling distribution of the sample mean $\bar{x}$ is approximately normal, with:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

So:

- the centre of the sampling distribution is still the population mean μ, and
- the spread becomes *smaller*, because the standard deviation is now $\sigma/\sqrt{n}$ instead of σ.

One way to visualize the Central Limit Theorem is as follows:



3.3 Why this matters

Notice that in the statement of the central limit theorem, we did not assume that the original probability distribution was normally-distributed. In fact, this is the great power of the Central Limit Theorem: it works for a huge range of distributions. In other words, the Central Limit Theorem says that the sampling distribution of the sample means is often approximately normal for large enough samples, *even when the original population is not normal.*

This means that normal distributions arise not merely because the world is always normal, but because randomly selected *averages* naturally arrange themselves within an approximate normal distribution.

Inferential Statistics

Under some basic assumptions, the Central Limit Theorem works even if we don't know what the population mean μ actually is. Whenever we can apply this theorem, the sampling distribution of $\bar{x}$ is centred at the *true* population mean μ. Therefore, we can create a bridge from sample data to population data and say something about a larger population that we don't completely know.

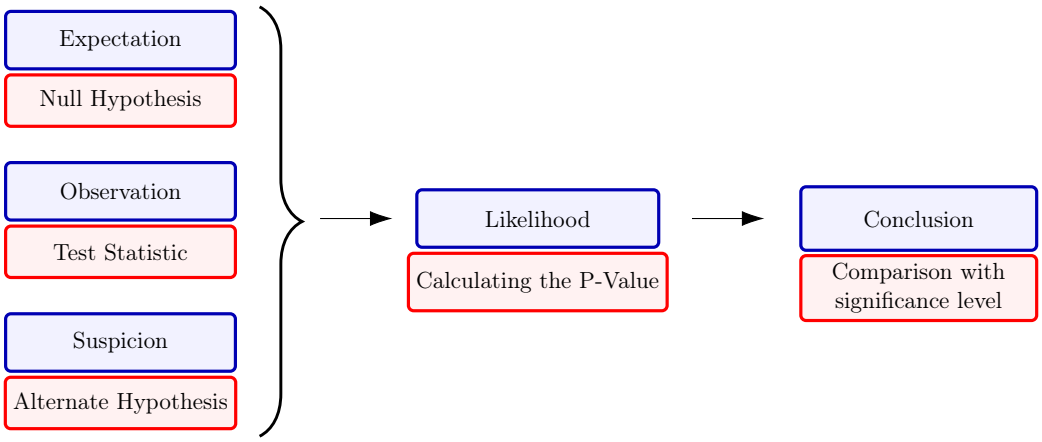
In reality, we rarely know the true population mean μ or population standard deviation σ. Moreover,

we often *cannot* know these parameters by recording the information, because we are constrained by resources. Instead, we only have access to sample data and their associated sample statistics. The Central Limit Theorem tells us that sample statistics have a predictable normal structure, and this builds a bridge between sample data and population parameters. We will now learn how to use this structure to test claims about the population parameter μ.

4 Hypothesis Testing

4.1 The general layout of a statistical test

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



We will now explain these one-by-one:

- At the start of any hypothesis test is the claim that we are trying to test. This is called the *null hypothesis*, and it is denoted by H_0 in mathematical notation. The null hypothesis can take many different forms depending on the data that we are working with. In this lecture we will stick to an introductory level and we will assume that the null hypothesis is making a claim about the value of the population mean μ.
- Next is the *alternate hypothesis* – the claim that you would like to test the null hypothesis against. In practice, this will often be a suspected value that differs from the statement of the null hypothesis. If the null hypothesis is making the claim that the mean μ equals a particular value c, then the alternate hypothesis will be a claim that μ is *not* equal to c in some way. This comes in three forms: $\mu < c$, or $\mu > c$, or $\mu \neq c$. These three types of alternate hypothesis turn the test into a *left-tailed*, *right-tailed,* or *two-tailed* test, respectively.
- Within any hypothesis test, there is the crucial sample data. This allows us to compute a sample mean $\bar{x}$, and then to plot that mean onto the sampling distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, where here n is the sample size. This sample mean $\bar{x}$ will exist *somewhere* in the sampling distribution, and therefore it has a corresponding z score. The z-score is often the *test statistic* that we will use to calculate a probability.
- The previous three points are the setup that is required to perform the hypothesis test. Next, we must determine the *P-value* of our sample data. This is the probability of finding this sample

mean, or something more extreme, *assuming that the null hypothesis were true*. Mathematically, for a left-tailed test this amounts to using the z-score to determine the probability $P(Z < z)$.

- The P-value provides the approximate probability of finding our sample data *or something more extreme* assuming that the null hypothesis were true. If this probability is *very small*, then the observation is very hard to explain under the model. To make this precise, we impose a cut-off of "acceptable probabilities", which is called the *level of significance*. We denote this with the letter α – typically α is either 5% or 2%. But, its exact value is ultimately an input of the test.

The final judgement of a hypothesis test is made as follows.

Decision rule

- If the calculated **P-value** $< \alpha$, then the observation is deemed too unlikely under H_0, so we **reject** H_0.
- If **P-value** $\geq \alpha$, then the observation is not unlikely enough, so we **do not reject** H_0.

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

4.2 Summary of Terminology

The important terminology from this section can be summarised as follows.

Important Terminology

- **Null hypothesis (H_0):** the statement that we begin by testing.
- **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".

At this level, an intuitive translation is also useful:

- H_0: the thing we are testing;
- H_1: the suspicion we have;
- test statistic: the observation we actually made;
- P-value: how lucky you would need to be to see that observation if H_0 were true;
- significance level: the level of luck we are willing to tolerate.

4.3 Examples

In the introduction we gave two informal prototype examples: we flipped coins 30 times and then calculated their probabilities.

Example 1: the suspicious coin. In this case, our statistical test was:

- **Null Hypothesis:** we expected that we were working with a fair coin, i.e. that the probability of

landing a heads was the same as the probability of landing a tails. Therefore, our null hypothesis would state $H_0 : p = 0.5$.

- **Test Statistic:** we observed 3 heads in 30 flips. Although we did not explicitly state it at the time, this was our test statistic: the sample data suggests a success probability of 10%.
- **Alternate Hypothesis:** after getting tails 27 times out of 30, we suspected that the probability of getting a heads may actually be smaller than 0.5. Therefore, our alternate hypothesis would state $H_1 : p < 0.5$.
- **Probability Calculation:** In this case, we directly calculated the probability of this exact outcome: we used the formula $C_{30,3}(0.5)^{30} \approx 0$. This is not yet the full P-value of a formal test, but it already shows that this observation is extraordinarily unlikely if the coin were fair.
- **Conclusion:** we concluded that this particular outcome is extraordinarily unlikely, and therefore we had very strong reason to reject the fair-coin assumption. Our conclusion was that the data gave strong evidence that we may be working with an unfair coin in which the probability of getting a heads is smaller.

Example 2: the more ordinary coin

- **Null Hypothesis:** again we tested the claim that the coin was fair, so our null hypothesis was $H_0 : p = 0.5$.
- **Test Statistic:** we observed 12 heads in 30 flips. Again without stating it explicitly, our sample data suggested a probability of $\frac{12}{30} = 0.4$.
- **Alternate Hypothesis:** nonetheless, to test that the coin might not be fair, we took an alternate hypothesis that the probability of landing a heads was less than 0.5. In symbols: $H_1 : p < 0.5$.
- **Probability Calculation:** We again considered the probability of this particular outcome, based on the assumption that the coin was fair. The calculation was $C_{30,12}(0.5)^30 \approx 8\%$. From this point of view, the outcome does not seem especially suspicious.
- **Conclusion:** we concluded that a probability of 8% for this exact outcome is actually quite reasonable, and therefore we had no strong reason to think that the coin was unfair.

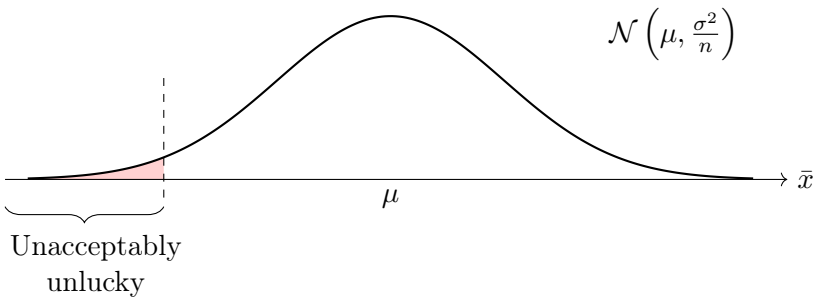
4.4 The philosophy of testing the population mean

The idea of hypothesis testing is to assume that the null hypothesis is true, and then to explore the logical consequences of this assumption. When working with a hypothesis that makes a definite claim about the mean of some population, we can invoke the Central Limit Theorem and calculate the probability that a given sample would occur randomly. Under most circumstances we would expect any simple random sample to have a sample mean that is relatively close to the center of the normal distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, where here μ is the value of the population mean claimed in the null hypothesis. In fact, according to our discussion in Lecture 14, we would expect there to be a $\sim 68\%$ chance for the sample mean to have a z-score between -1 and $+1$, that is, to fall within the interval $\left[\mu - \frac{\sigma}{\sqrt{n}},\ \mu + \frac{\sigma}{\sqrt{n}}\right]$ in the distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

If, for example, we were to take a simple random sample and were to obtain a sample mean that is far away from the claimed population mean μ, then this would correspond to a very low chance of happening in reality. Of course, by the nature of randomness it is always possible to draw a sample that is particularly unlucky and gives a sample mean quite far from the population mean μ. However, how unlucky do we need to be before we start to think that something is wrong with

our model? For example, it is statistically possible to obtain a random sample that is 5 standard deviations away from the center of a normal distribution. However, if we were to do this on our *first attempt at taking a sample*, then it is much more reasonable to suggest that our model is simply wrong.

In order to distinguish between "reasonably unlucky" and "unreasonably unlucky", we impose the significance level α. This value acts as a cut-off point at which we can say that any observed value landing past this cut-off point is simply too unlucky to reasonably accept. Instead of saying that the original model has been proved wrong, we treat such an observation as strong evidence against the original model. That is why we reject the null hypothesis whenever our sample statistic has a P-value lower than the level of significance.



This is why a statistical test has a built-in cut-off: we decide in advance how much "luck" we are willing to tolerate before we stop trusting the original model.

4.5 Type 1 and Type 2 Errors

Ultimately, hypothesis tests always have some uncertainty in them. If, for example, we perform a hypothesis test and find a P-value that is lower than the significance level, then that does not mean that we have *proved* that the null hypothesis is false. It simply means that the sample provides evidence against the null hypothesis. Moreover, in such a situation we also have not, therefore, *proved* the alternate hypothesis either – instead, we have found that the data is hard to explain under the null hypothesis.

Whenever we perform a statistical test there is always an element of uncertainty, and therefore it is always possible to draw the wrong conclusion from a test by accident. There are two main errors, called a *Type 1 error* and a *Type 2 error*. They are defined as follows.

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.

In simpler language:

- a **Type 1 Error** means that we became suspicious of the null hypothesis and rejected, even though it was true in reality, and

- a **Type 2 Error** means that our sample failed to detect that something was wrong with the null hypothesis, and therefore we did not reject it even though it is false in reality.

5 Testing Binomial Experiments

In our coin flipping examples from before, the probabilities were calculated naively with the binomial probability formulas $C_{n,r}p^r q^{n-r}$. However, in the case of large systems like samples of $n = 30$ coin flips, the binomial coefficient becomes:

$$C_{30,r} = \frac{30!}{r!(30-r)!}.$$

For low values of n, we have been calculating the binomial coefficient using Pascal's triangle, which was very useful up until now. However, when dealing with *large* values of n, the triangle becomes unreasonable. In fact, other computation methods also become unreasonable, since the factorial $n!$ becomes very large very fast. In practice, it is much easier to simply *approximate* this probability calculation by using a normal distribution that roughly fits the binomial distribution. That way, we can use z-scores to calculate P-values and the Central Limit Theorem to help us make a judgement based on these values. This approximation will produce a different probability: now, since we are working with a continuous random variable, the probability calculation gives the area of a *region*, rather than the height of an individual bar in a binomial distribution. In practice, we will consider the probability to be the likelihood of receiving this outcome *or anything more extreme*. This is how we will proceed with our hypothesis tests.

5.1 From binomial experiments to the Central Limit Theorem

Suppose we run a binomial experiment, such as repeatedly flipping a coin. Each trial has two outcomes, success or failure, where p is the probability of success and $q = 1 - p$ is the probability of failure. We can view a binomial experiment through the eyes of statistics, by interpreting:

- each trial is an individual data point, and
- doing n trials is like taking a sample of size n.

Any fixed outcome of the experiment is like a sample of size n.

If we choose to interpret the outcomes of a binomial experiment as data, we can then perform statistics on it. One way to do this is to make the data numerical, i.e. assigning a successful outcome as 1 and a failure outcome as 0. Then the sample mean is exactly the sample proportion, because

$$\bar{x} = \frac{\text{number of successes}}{n} = \hat{p}.$$

So in a binomial experiment, talking about the sample mean and talking about the sample proportion amount to the same thing.

For this 0-1 version of the binomial experiment, the mean is equal to p, and the variance is equal to pq.

5.2 The sampling distribution for a sample proportion

For a binomial experiment, it can be shown that

$$\text{population mean} = p, \qquad \text{population variance} = pq.$$

When we average n such trials, we obtain the sample proportion $\hat{p}$. So this is the same sampling-mean story as before, just written in binomial language.

By the Central Limit Theorem, for sufficiently large n, the collection of all sampling proportions $\hat{p}$ is approximately normal. Specifically, we have the following formula.

The key formula for binomial hypothesis testing

If the true success probability is p, then the sampling distribution of the sample proportions $\hat{p}$ is approximately

$$\mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

This is an extremely important fact. It tells us that binomial luck has an approximately normal structure when the sample is large enough.

In a binomial setup, the unknown population parameter is the true success probability p.

For example, in a coin-flipping experiment:

- the **population parameter** is the true probability of heads;
- the **sample statistic** is the observed proportion of heads.

So we compare:

$$\text{true probability } p \qquad \text{versus} \qquad \text{sample probability } \hat{p}.$$

5.3 A full test: the suspicious coin

Let's return to the earlier scam story. We will now sketch out a hypothesis test in the case of the first coin.

Step 1: state the hypotheses.
If we suspect that the coin is biased *against* heads, then the natural hypotheses are:

$$H_0 : p = 0.5, \qquad H_1 : p < 0.5.$$

Step 2: compute the observed sample proportion.
We observed 3 heads in 30 flips, so

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

Step 3: compute the z-score.
Instead of using a binomial calculation, here we will approximate with a normal curve and use the associated z-score as our test statistic. Here we have:

$$p_0 = 0.5, \qquad q_0 = 0.5, \qquad n = 30.$$

So the associated z-score is:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0.1 - 0.5}{\sqrt{(0.5)(0.5)/30}} \approx -4.38.$$

Step 4: compute the P-value.
A z-score of -4.38 means that our observed outcome is over 4 standard deviations away from the centre of the sampling distribution. Therefore, we would expect that the probability associated to this region is very low. In fact, this area is so far to the left of the centre that standard z-scores tables do not write these probabilities down. Using a z-scores calculator, we can find that the P-value is

$$P(Z < -4.38) \approx 0.00001.$$

Step 5: compare with a significance level.
If we use a level of significance of $\alpha = 0.05$, then clearly

$$0.00001 < 0.05.$$

So, we reject H_0.

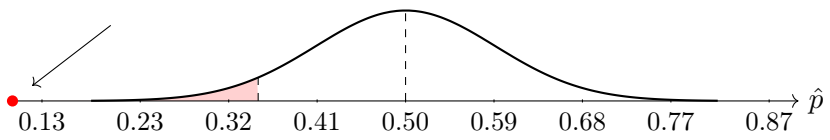
Conclusion

There is extremely strong evidence against the hypothesis that the coin is fair.

In this test, the null hypothesis predicts a sampling distribution for $\hat{p}$ that is centred at 0.5:

$$\hat{p} \approx N\left(0.5, \frac{(0.5)^2}{30}\right).$$

The standard deviation of this sampling distribution is therefore equal to $\sqrt{\frac{(0.5)^2}{30}} \approx 0.09$. Below is the graph of this distribution. The observed value of $\hat{p} = 0.1$ is extremely far away from what the null hypothesis is claiming the centre is.



The fair-coin model expects most observed proportions to lie near 0.5. An observed value like $\hat{p} = 0.1$ is very far into the unlikely left tail. So, we reject the null hypothesis.

5.4 An important condition before using the normal approximation

It should be noted that the normal approximation does *not* automatically apply to every binomial test. Instead, we need the sample to be large enough under the null hypothesis.

Condition for the normal approximation

Before using the z-test for a proportion, we must check:

$$np_0 > 5 \qquad \text{and} \qquad nq_0 > 5.$$

If both conditions hold, the normal approximation is usually reasonable at this level.

For the suspicious coin example,

$$np_0 = 30(0.5) = 15 > 5, \qquad nq_0 = 30(0.5) = 15 > 5,$$

so the approximation is allowed.

6 Examples and Exercises

6.1 The Haribo data

In class we collected data from 19 bags of Haribo sweets, counting the number of each candy type in each bag.

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Rings	3	7	4	2	6	6	7	3	7	4	2	6	6	4	7	5	5	2	6
Hearts	5	4	6	4	5	2	5	7	1	3	6	3	4	6	4	3	2	4	5
Bottles	4	1	4	2	4	2	3	3	4	6	4	0	2	4	4	4	8	1	5
Eggs	4	5	3	6	0	6	1	2	6	2	4	4	1	2	1	4	3	7	1
Gummy bears	7	8	6	12	11	10	10	9	7	11	6	14	15	8	10	9	6	12	7
Total per bag	23	25	23	26	26	26	26	24	25	26	22	27	28	24	26	25	24	26	24

The total number of sweets counted is 476, and the proportions are:

Type	**Total**	**Proportion**
Rings	92	$92/476 \approx 19.3\%$
Hearts	79	$79/476 \approx 16.6\%$
Bottles	65	$65/476 \approx 13.7\%$
Eggs	62	$62/476 \approx 13.0\%$
Gummy bears	178	$178/476 \approx 37.4\%$

So gummy bears occur much more frequently than the others.

6.1.1 Turning the Haribo question into a hypothesis test

In order to test the hypothesis that the proportion of Gummy Bears is greater than 20%, we will interpret our system as a binomial experiment. This can be achieved by first defining:

- **success** = "getting a gummy bear";
- **failure** = "not getting a gummy bear".

At this introductory level, we will treat the 476 sweets as 476 separate observations, so that $n = 476$. We are also ignoring possible bag-to-bag effects.

We will now test the claim that the true probability of drawing a gummy bear is 20%.

Step 1: hypotheses.

Let p be the true probability that a randomly selected sweet is a gummy bear. Based on our sample data, we suspect that the real proportion of gummy bears is greater than 20%. So, our null hypothesis and alternate hypothesis are:

$$H_0 : p = 0.20, \qquad \text{and} \quad H_1 : p > 0.20.$$

Step 2: observed sample proportion.

In the sample, we observed 178 gummy bears in a total of $n = 476$ sweets. The sample proportion of gummy bears is then:

$$\hat{p} = \frac{178}{476} \approx 0.374,$$

in agreement with the table above.

Step 3: checking the approximation conditions. In order to approximate our binomial experiment with a normal distribution, we first need to check the two conditions mentioned in Section 5. Under H_0, we have:

$$p_0 = 0.20, \qquad \text{and} \quad q_0 = 0.80.$$

Calculating the two conditions:

$$np_0 = 476(0.20) = 95.2 > 5, \qquad nq_0 = 476(0.80) = 380.8 > 5.$$

We confirm that both of these numbers are greater than 5, and therefore we may proceed with the hypothesis test.

Step 4: compute the test statistic.

Using the values of $p_0 = 0.2$, $q_0 = 0.8$, $\hat{p} = 0.374$ and $n = 476$, we have:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0.374 - 0.20}{\sqrt{(0.20)(0.80)/476}} \approx 9.49.$$

Therefore, according to the null hypothesis our sample observation of $\hat{p}$ is over 9 standard deviations away from the supposed centre of the sampling distribution. **Step 5:** interpret the P-value.

Our alternate hypothesis is claiming that the proportion of gummy bears is *greater than* 0.2. Therefore, we are looking in the right tail of the sampling distribution. The associated P-value is

$$P(Z > 9.49),$$

which is so small that it is essentially 0 for all practical purposes.

Conclusion

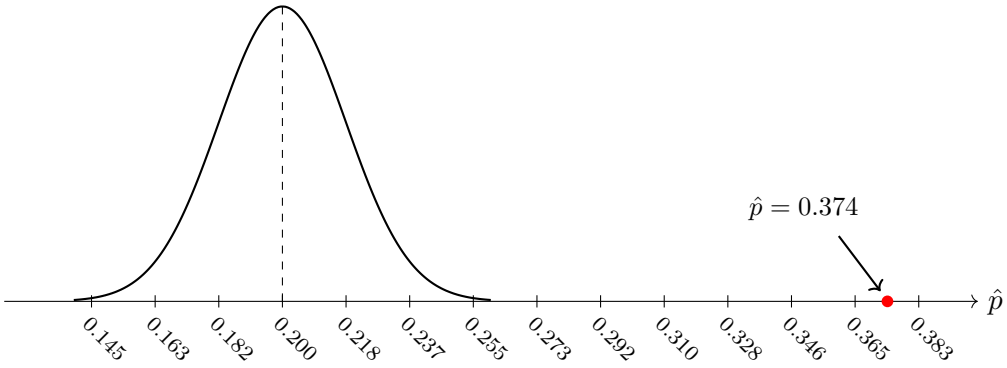
The sample gives overwhelming evidence against the claim that gummy bears occur only 20% of the time. At least in this data set, gummy bears appear to be heavily overrepresented.

6.2 A picture of the Haribo test

Under the null hypothesis, the sampling distribution is approximately

$$\hat{p} \approx N\left(0.20, \frac{(0.20)(0.80)}{476}\right).$$

Its standard deviation is only about 0.018. That means an observed value like 0.374 is extremely far out in the right tail.



6.3 Worked exercise

Exercise

A news website claims that 60% of visitors click on at least one article. In a simple random sample of $n = 96$ visitors, only $x = 48$ clicked on an article.
Let p be the true click-through rate.

(a) State hypotheses to test whether the true click-through rate is lower than 60%.
(b) Check the normal approximation conditions under H_0.
(c) Compute $\hat{p}$ and the z-score.
(d) Find the P-value using the standard normal table below.
(e) Make a decision at significance level $\alpha = 0.05$.
(f) Interpret the conclusion in context.

z	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
$P(Z < z)$	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98

Solution

(a) Since we want to test whether the true click-through rate is lower than 60%,

$$H_0 : p = 0.60, \qquad H_1 : p < 0.60.$$

(b) Under H_0,

$$p_0 = 0.60, \qquad q_0 = 0.40.$$

So

$$np_0 = 96(0.60) = 57.6 > 5, \qquad nq_0 = 96(0.40) = 38.4 > 5.$$

Therefore the normal approximation is valid.

(c) The sample proportion is

$$\hat{p} = \frac{48}{96} = 0.50.$$

Hence

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0.50 - 0.60}{\sqrt{(0.60)(0.40)/96}} = -2.0.$$

(d) Because this is a left-tailed test,

$$\text{P-value} = P(Z < -2.0).$$

From the table,

$$P(Z < -2.0) \approx 0.02.$$

(e) Since

$$0.02 < 0.05,$$

we reject H_0.

(f) There is evidence that the true click-through rate is lower than 60%.