MAT220 Lecture 17: Testing Independent Populations

Contents

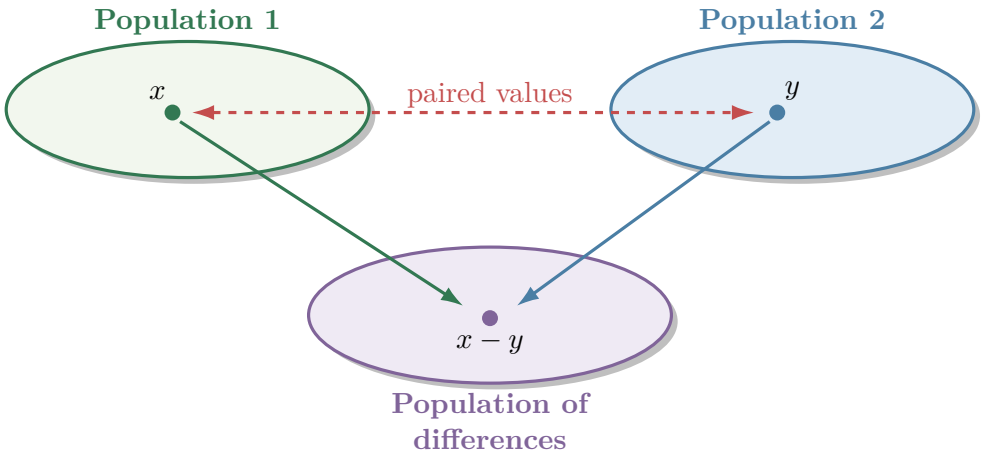
In Lecture 16, we compared two populations whose observations were naturally paired. The pairing allowed us to subtract inside each pair, form a new population of differences $D = X - Y$, and reduce the problem to a one-sample t-test. In this lecture, we will remove that assumption and instead take our populations to be *independent*. As we will see, this new assumption allows us to again perform hypothesis tests of two populations simultaneously. This time, however, we can no longer construct differences $d = x - y$. Instead, as we will see, we need to work around this loss of convenience in another way. In the end, we will use our new techniques to test the difference of population means, and also to test the difference between the success probabilities of two independent binomial experiments.

1 A Review of Paired Data

In this lecture, we will continue to discuss hypothesis tests that compare the parameters of two separate populations. Given two populations, the central question here is:

> *Are these two populations the same, on average?*

In principle, this can be answered by comparing the true means of the two populations and testing whether or not they are equal. In the process of performing a test like this, we need to create a sampling distribution that combines sample data from the two populations. In Lecture 16, we noted that, generally speaking, this is a difficult task. Therefore, we made our lives easier by assuming that our populations were naturally *paired*. Since each individual observation in one population had a unique counterpart in the other, this assumption allowed us to create a third population made up of all of the differences between these paired values:



Symbolically, we used the letter D to represent this "population of differences". Formally, D is nothing but $X - Y$, the difference of the random variables used to define the two populations.

Using some standard facts about expected values and variances, we saw that the mean and the variance of this random variable D were:

$$\mu_D = \mu_X - \mu_Y, \qquad \sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2\,\mathrm{Cov}(X, Y).$$

Although the mean takes a nice form, the variance carries an extra term known as *covariance*. Roughly speaking, this extra covariance term measures how the paired values move together. Since σ_D is generally unknown, we had to estimate the standard deviation of the differences using s_D and use a Student's t-distribution, giving the following test statistic.

Paired-data test statistic

For paired data, first compute the differences $d_i = x_i - y_i$. Then use

$$t = \frac{\bar{d} - 0}{s_D/\sqrt{n}}, \qquad df = n - 1.$$

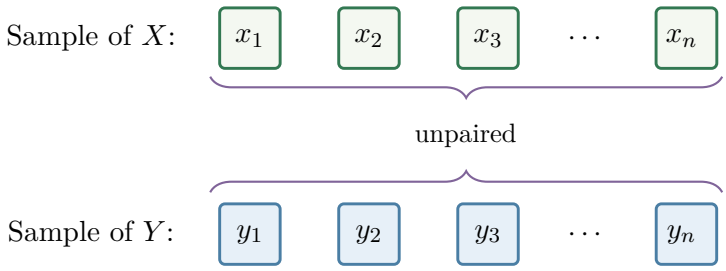
2 Independent Populations

2.1 Changing our assumptions

When performing hypothesis tests involving two populations, there is a major technical hurdle when attempting to describe the appropriate sampling distribution. In order to make things easier for ourselves, in the last lecture we assumed that the two populations were naturally paired. This allowed us to take samples from the population D and break them apart into corresponding samples of X and Y, thereby giving us a description of the sampling distribution of $\bar{D}$.

In this lecture, we will no longer assume that our populations are paired. Instead, we will assume that X and Y are *independent*, meaning that there is no probabilistic dependence between the two. In fact, we saw this term when discussing probability earlier in the course: two events were called "independent" if the occurrence of one did not change the probability of the other. In a similar spirit, we call X and Y independent whenever knowledge of one does not change the probability distribution of the other. Mathematically, assuming that X and Y are independent implies that there is *no* interaction between the two, and therefore the covariance $\text{Cov}(X, Y)$ vanishes.

Although this looks promising, we also have to admit that we have lost something in the process of relaxing the assumption of paired data. When observations of X and Y were paired, we were able to form a third population governed by $D = X - Y$, where data points were of the form $d = x - y$. This allowed us to take samples of the population D and break them apart into samples of X and Y individually. If we drop the assumption that X and Y are paired, then we can no longer perform this trick. In other words, even if we picked two simple random samples from X and Y of the same size, there is no natural way to pair them up into some kind of sample of differences:



This means that there is no natural way to create a sample of individual differences for testing the difference $\mu_X - \mu_Y$.

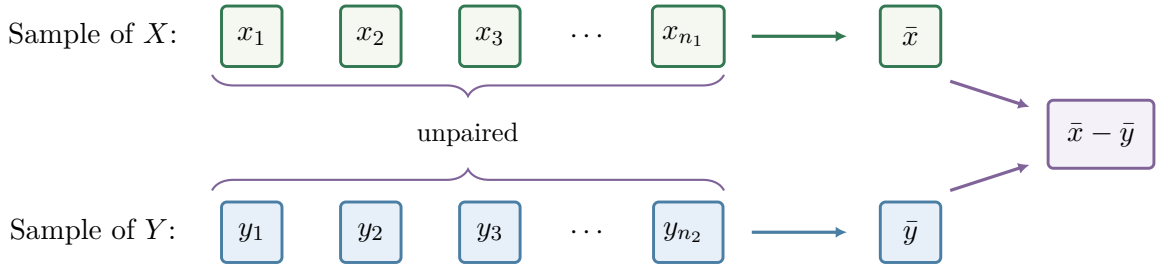
2.2 The difference of sample means

Suppose that we take two independent random samples from X and Y, respectively. Since the samples are independent, there is no need to require that these two samples are of the same size. So, let n_1 denote the size of the sample from X and let n_2 denote the size of the sample from Y. Then the sample means $\bar{x}$ and $\bar{y}$ can be calculated individually:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_{n_1}}{n_1} \qquad \text{and} \qquad \bar{y} = \frac{y_1 + y_2 + \cdots + y_{n_2}}{n_2}.$$

Ultimately, our goal is to describe a sampling distribution centred at $\mu_X - \mu_Y$, without being able

to directly appeal to some population of differences D. According to the sample data that we have available to us, we can create a difference of averages by considering the quantity $\bar{x} - \bar{y}$:



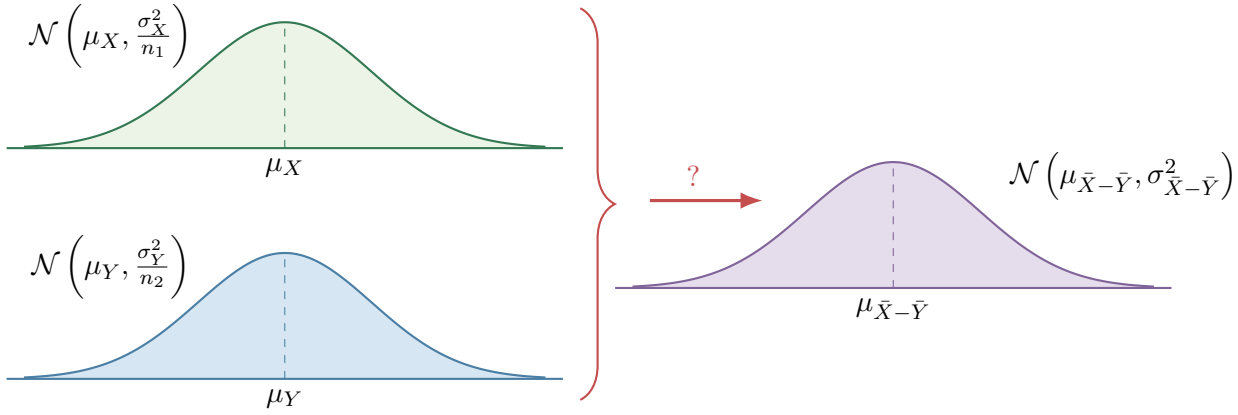
If we wanted to perform a hypothesis test by evaluating the chances of this particular observed value $\bar{x} - \bar{y}$, then we need to consider a probability distribution containing the collection of all possible randomly obtained differences $\bar{x} - \bar{y}$. Therefore, we will study the random variable $\bar{X} - \bar{Y}$.

For sufficiently large samples, the Central Limit Theorem tells us that the individual sampling distributions of $\bar{X}$ and $\bar{Y}$ are approximately normal, with parameters:

$$\bar{X} \approx \mathcal{N}\left(\mu_X, \frac{\sigma_X^2}{n_1}\right) \qquad \text{and} \qquad \bar{Y} \approx \mathcal{N}\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right).$$

We emphasize that, in contrast to the case of paired data, the two samples need not have the same size, i.e. it does not need to be the case that $n_1 = n_2$.

We are now going to use these two normal distributions to describe the overall distribution of the random variable $\bar{X} - \bar{Y}$. Schematically, we are looking for the following:



2.3 Constructing the sampling distribution for $\bar{X} - \bar{Y}$

In the previous lecture, we saw that the difference of two random variables has an expected value equal to the difference of the expected values of the individual variables. In symbols, for $\bar{X} - \bar{Y}$, we have:

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}).$$

Since $E(\bar{X})$ is simply the mean of $\bar{X}$, which is μ_X, and similarly for $\bar{Y}$, we may conclude that:

$$\mu_{\bar{X}-\bar{Y}} = E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y.$$

Conveniently, this is the same centre that appeared for paired data: the expected difference between two sample means is the difference between the two population means.

As for the variance, in fact we may again use the same fact as in Lecture 16 to conclude that

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) - 2\,\mathrm{Cov}(\bar{X}, \bar{Y}),$$

where $\mathrm{Cov}(\bar{X}, \bar{Y})$ is the covariance of $\bar{X}$ and $\bar{Y}$. This is where the magic happens: because the samples are assumed to be independent, $\bar{X}$ and $\bar{Y}$ are also independent, and therefore this extra covariance term vanishes: $\mathrm{Cov}(\bar{X}, \bar{Y}) = 0$. Therefore, the variance formula above can be conveniently simplified to:

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}).$$

The variance of $\bar{X}$ is simply the second parameter in its normal distribution, that is, the square of the standard error term for $\bar{X}$, namely $\frac{\sigma_X^2}{n_1}$. Similarly, the variance of $\bar{Y}$ is given by $\frac{\sigma_Y^2}{n_2}$. Therefore, we may conclude that for independent populations, the variance of the distribution of $\bar{X} - \bar{Y}$ will be:

$$\sigma^2_{\bar{X}-\bar{Y}} = \mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}.$$

Putting this all together, we now have a full description of the normal distribution that approximately describes $\bar{X} - \bar{Y}$.

Sampling distribution of the difference of independent means

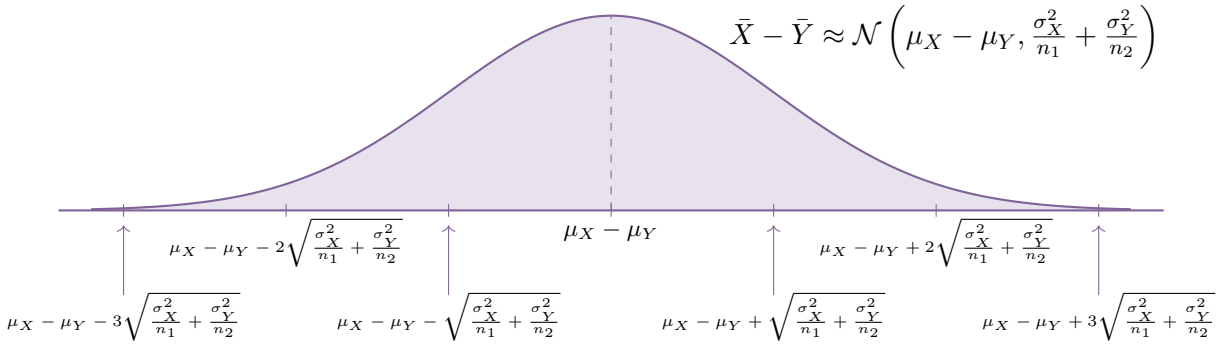
For independent samples,

$$\bar{X} - \bar{Y} \approx \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right).$$

Our normal distribution will be centred at the term $\mu_X - \mu_Y$, and it will "count out" in terms of the standard error, which is the square root of the variance term above:

$$\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}.$$

Geometrically, we obtain the following image:



Although this is complicated, the important point is that this is entirely *tractable* – there will be a formula that we can use to compute relevant probabilities.

3 Hypothesis Tests for Independent Means

3.1 The null and alternate hypotheses

As with our discussion in the previous lecture, when we want to test whether or not the two independent populations are equal on average, the null hypothesis is given by

$$H_0 : \mu_X - \mu_Y = 0.$$

We then take two independent samples from X and Y of potentially different sizes. Assuming the null hypothesis, the sampling distribution becomes

$$\bar{X} - \bar{Y} \approx \mathcal{N}\left(0, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right).$$

The two samples from X and Y have their respective sample means $\bar{x}$ and $\bar{y}$, and these have a difference $\bar{x} - \bar{y}$ that will land somewhere on the normal distribution described above. As with the other hypothesis tests that we have seen previously, the farther the observed value $\bar{x} - \bar{y}$ lies from the centre, the smaller the P-value becomes.

As with the paired populations of Lecture 16, if we wanted to test H_0 against a statement to the contrary, then we can use different statements depending on our particular suspicion:

$$H_1 : \begin{cases} \mu_X - \mu_Y < 0, & \text{says that the population for } X \text{ is smaller on average,} \\ \mu_X - \mu_Y > 0, & \text{says that the population for } X \text{ is larger on average,} \\ \mu_X - \mu_Y \neq 0, & \text{merely says that the averages are different.} \end{cases}$$

Choosing the direction

The order of subtraction must match the alternate hypothesis. If the setup uses the sample difference $\bar{x} - \bar{y}$, then:

- $\mu_X < \mu_Y$ becomes $H_1 : \mu_X - \mu_Y < 0$;
- $\mu_X > \mu_Y$ becomes $H_1 : \mu_X - \mu_Y > 0$.

3.2 Test Statistics with known and unknown σ

Suppose that both population standard deviations, σ_X and σ_Y, happen to be known. In this case, the distribution for $\bar{X} - \bar{Y}$ has standard error given by

$$\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}.$$

When we convert the observed difference into a z-score, we schematically take the quotient:

$$z = \frac{\text{observed value} - \text{mean}}{\text{standard error}}.$$

In our case, we are assuming the null hypothesis, which says that $\mu_X - \mu_Y = 0$. The standard error is given above, and can be calculated directly from the information available. Therefore, the observed difference $\bar{x} - \bar{y}$ is converted into the z-score via the formula:

$$z = \frac{(\bar{x} - \bar{y}) - 0}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}.$$

Independent means with known standard deviations

To test $H_0 : \mu_X - \mu_Y = 0$ when σ_X and σ_Y are known, we use the formula:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}.$$

Suppose now that either σ_X or σ_Y is unknown. In this case, we don't have the values to substitute into the previous formula. Therefore, the next best thing we can do is replace both population standard deviations with their sample estimates, and proceed with an approximate Student's t-distribution instead:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}$$

either σ_X or σ_Y is unknown

replace σ_X and σ_Y by s_X and s_Y

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}}.$$

Student's t, $df = \min(n_1 - 1, n_2 - 1)$

As a convention, in this course we will use the smaller of the two available degrees of freedom.

Independent means with unknown standard deviations

To test $H_0 : \mu_X - \mu_Y = 0$ when either population standard deviation is unknown, use the formula

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}},$$

where $df = \min(n_1 - 1, n_2 - 1)$.

After this, the appropriate P-value can be calculated using the corresponding z-table or t-table.

4 Testing Two Binomial Probabilities

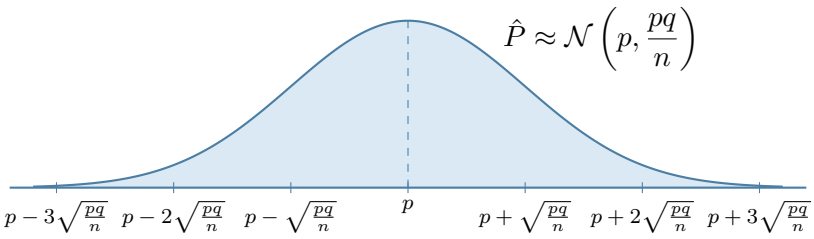
4.1 A reminder about one-sample proportions

Suppose that we have a single binomial experiment, with success probability p. If we were to take n trials and witness r successes, then we can calculate an estimate of p from this sample data. This estimate is known as the sample proportion, and it can be calculated simply as:

$$\hat{p} = \frac{r}{n}.$$

As we saw in Lectures 13 and 15, in this setting the Central Limit Theorem gives the normal approximation

$$\hat{P} \approx \mathcal{N}\left(p, \frac{pq}{n}\right), \qquad q = 1 - p.$$



When working in a situation like this, we must always check that the sample is large enough for the normal approximation to be reasonable. In our course, we use the convention of $np > 5$ and $nq > 5$.

4.2 The difference of two independent sample proportions

Suppose that we now have two binomial experiments:

- Experiment 1 has true success probability p_1, sample size n_1, and sample proportion $\hat{p}_1$;
- Experiment 2 has true success probability p_2, sample size n_2, and sample proportion $\hat{p}_2$.

Individually,

$$\hat{P}_1 \approx \mathcal{N}\left(p_1, \frac{p_1 q_1}{n_1}\right) \qquad \text{and} \qquad \hat{P}_2 \approx \mathcal{N}\left(p_2, \frac{p_2 q_2}{n_2}\right),$$

where here $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. The sample proportions play the same role as the sample means, and we can take their difference to get a new variable $\hat{P}_1 - \hat{P}_2$. If we assume that these two experiments are independent, then the covariance term will again vanish and we are left with a normal distribution of the form:

$$\hat{P}_1 - \hat{P}_2 \approx \mathcal{N}\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right).$$

Using this as motivation, we will now describe a method for testing whether or not the two experiments have the same success probability.

4.3 Hypothesis tests for differences of success probabilities

To test whether the two success probabilities are equal, we start with the null hypothesis:

$$H_0 : p_1 - p_2 = 0.$$

Under this assumption, the normal distribution mentioned above will be centred at 0:

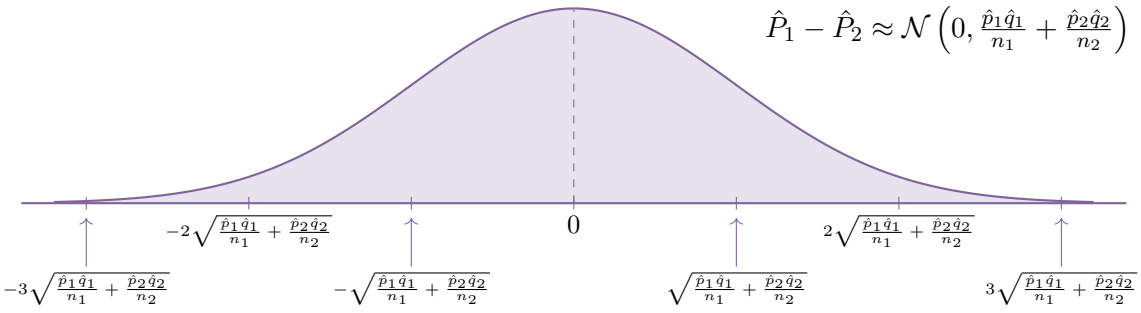
$$\mathcal{N}\left(0, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right).$$

However, we cannot directly use this distribution in testing, because we don't actually know the values of p_i and q_i. Why not? Because the null hypothesis only makes a statement about the

difference $p_1 - p_2$, not about their precise value. Therefore, we have to treat p_1 and p_2 as unknown, and we need to instead estimate their value with the sample proportions instead. This means that our normal distribution will be:

$$\hat{P}_1 - \hat{P}_2 \approx \mathcal{N}\left(0, \frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}\right),$$

where $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$. Geometrically, this is again a bell curve centred at 0, but this time we are counting out in terms of the standard error $\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$:



The corresponding z-score starts counting from the centre at 0 in terms of the standard error term above, meaning that the formula will be:

$$z = \frac{\text{observed value} - \text{mean}}{\text{standard error}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}.$$

This will be the formula to use to compute the appropriate P-value.

Testing two independent binomial probabilities

To test $H_0 : p_1 - p_2 = 0$, use

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}.$$

The P-value is calculated using $Z \sim \mathcal{N}(0, 1)$.

As with the normal distribution of Section 4.1, we should always confirm that the two binomial experiments can be meaningfully approximated by normal distributions in the first place. In our setting, we must always confirm that the four quantities

$$n_1\hat{p}_1, \qquad n_1\hat{q}_1, \qquad n_2\hat{p}_2, \qquad n_2\hat{q}_2$$

are greater than 5 before proceeding with the approximation.

5 Example: Orange Skittles

Suppose that a professor gives bags of Skittles to two classes. He would like to test whether the proportion of orange Skittles in Class 1 is less than the proportion of orange Skittles in Class 2.

The recorded data is:

- Class 1 counted 740 Skittles, of which 139 were orange;
- Class 2 counted 741 Skittles, of which 150 were orange.

We perform the test using a significance level of $\alpha = 0.05$.

Step 1: Gathering the information

The important values are

$$n_1 = 740, \qquad r_1 = 139, \qquad n_2 = 741, \qquad r_2 = 150, \qquad \alpha = 0.05.$$

Let p_1 be the true proportion of orange Skittles for Class 1, and let p_2 be the true proportion for Class 2. The null hypothesis says that the proportions are equal, while the suspicion says that the first is smaller:

$$H_0 : p_1 - p_2 = 0, \qquad H_1 : p_1 - p_2 < 0.$$

This is a left-tailed test.

Step 2: Calculate the P-value

First compute the two sample proportions:

$$\hat{p}_1 = \frac{r_1}{n_1} = \frac{139}{740} \approx 0.188, \qquad \hat{p}_2 = \frac{r_2}{n_2} = \frac{150}{741} \approx 0.202.$$

Therefore,

$$\hat{q}_1 = 1 - \hat{p}_1 \approx 0.812, \qquad \hat{q}_2 = 1 - \hat{p}_2 \approx 0.798.$$

Next check the normal approximation conditions:

$$\begin{aligned} n_1\hat{p}_1 &\approx 740(0.188) = 139 > 5, \\ n_1\hat{q}_1 &\approx 740(0.812) = 601 > 5, \\ n_2\hat{p}_2 &\approx 741(0.202) = 150 > 5, \\ n_2\hat{q}_2 &\approx 741(0.798) = 591 > 5. \end{aligned}$$

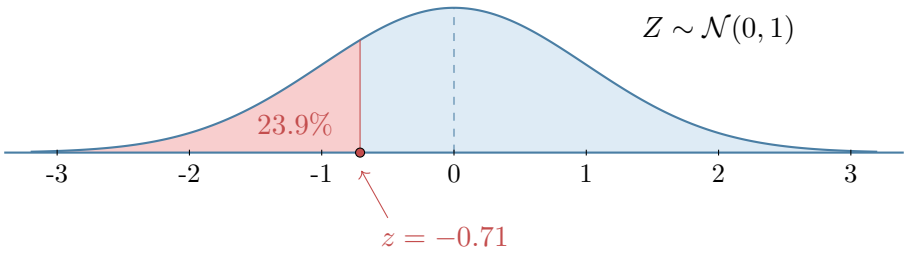
All four values are greater than 5, so the normal approximation is reasonable. Now we can use our information to calculate the z-score:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \approx \frac{0.188 - 0.202}{\sqrt{\frac{(0.188)(0.812)}{740} + \frac{(0.202)(0.798)}{741}}} \approx -0.71.$$

Since this is a left-tailed test, the P-value can be read directly from the z-table:

$$\text{P-value} = P(\hat{P}_1 - \hat{P}_2 \leq -0.0146) \approx P(Z \leq -0.71) \approx 0.2389.$$

Geometrically, we have just computed an area in the left tail that is around 23.9% of the total graph:



By construction, this area is the same as the area to the left of our sample data $\hat{p}_1 - \hat{p}_2$ in the sampling distribution built according to the null hypothesis. In words: if the two classes truly had the same proportion of orange Skittles, then there would be about a 23.9% chance of obtaining a difference at least this far in the direction $\hat{p}_1 < \hat{p}_2$.

Step 3: Compare the P-value with α

A comparison gives:

$$0.2389 > 0.05.$$

Therefore, we do not reject H_0.

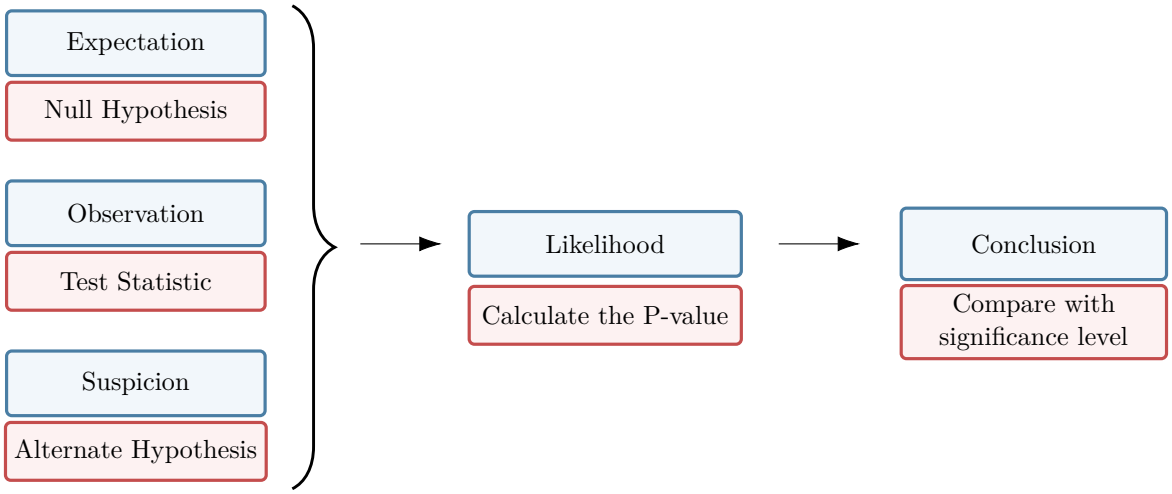
Step 4: Give an interpretation

Interpretation

At the 5% significance level, the samples do not provide convincing evidence that the proportion of orange Skittles in Class 1 is lower than the proportion in Class 2.

6 A Summary of Hypothesis Tests

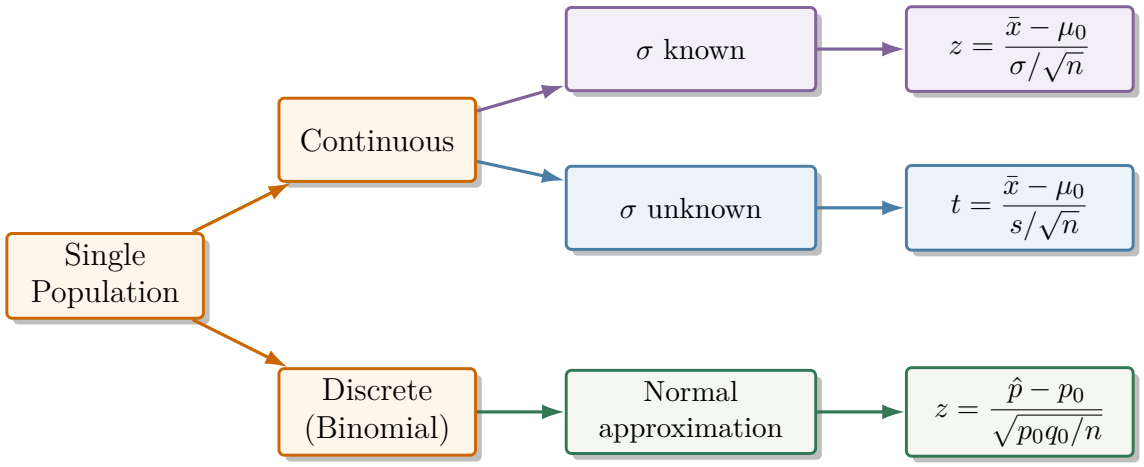
All of the hypothesis tests we have seen follow the same overall pattern:



So far in the course, we have studied seven major types of hypothesis test. These are:

1. testing a single population mean μ with known σ,
2. testing a single population mean μ with unknown σ,
3. testing a single binomial success probability p,
4. testing $\mu_X - \mu_Y$ for paired data,
5. testing $\mu_X - \mu_Y$ for independent data with known population standard deviations,
6. testing $\mu_X - \mu_Y$ for independent data with unknown population standard deviations, and
7. testing the difference $p_1 - p_2$ for two independent binomial experiments.

These can be divided into two main camps: tests of a single population and tests of two populations simultaneously. In the case of a single population, the following flow chart organises the different tests.



When testing two populations at the same time, we are met with the following flow chart: