# Chapter 33

# Null Hypothesis Significance Testing

The Null Hypothesis Significance Testing procedure, or NHST for short, is a recipe-like data analysis technique that you can use to support a scientific claim or new theory. The term *significance* carries a very specific meaning in this context. We say a scientific claim is *statistically significant* according to some threshold value $\alpha$ (usually $\alpha = 0.05$) if the probability of the observed data occurring "by chance" is smaller than $\alpha$. The notion of what data might occur "by chance" is defined by some baseline probability model that assumes the claim is not true, which we call the *null hypothesis*. The NHST "recipe" was introduced in the last century for the purpose of defining a minimum standard for statistical analysis that scientists must perform before reporting their research findings. Before any new theory or new scientific claim is published, the scientist must follow the NHST procedure to show that the data observed is not simply due to chance (the null hypothesis).

The idea of a standardized procedure for analyzing data agreed upon by all scientists in a field has greatly advanced scientific research. Indeed we could say that for many decades the notion of "doing Science" was synonymous with using the NHST recipe. The NHST procedure has been passed through generations of scientists without too much modifications and it is actively used to this day for data analysis. Modern statistics has developed many more advanced, detailed, and nuanced ways of doing statistical analysis, but it's important that we learn about NHST because it is the most common type of statistical analysis you're likely to encounter.

## 33.1 Definitions

The main purpose of the NHST procedure is to test the plausibility of some scientific claim, which is usually expressed as a mathematical statement about a population parameter. We begin this chapter by introducing all the necessary terminology and definitions needed to understand and apply the NHST procedure.

### 33.1.1 Hypotheses

We formalize the scientific claim we want to test using two precise precise mathematical statements called *hypotheses*.

- The *alternative hypothesis*, denoted $H_A$, is a statement about the value of a population parameter that corresponds to the new scientific claim. The alternative hypothesis describes the new theory that the scientists suspect is true.
- The *null hypothesis*, denoted $H_0$, is a skeptical claim about the value of the population parameter that is contrary to the alternative hypothesis. The null hypothesis is assumed to be true unless we find evidence that shows otherwise.

The null hypothesis and the alternative hypothesis should be *mutually exclusive* and *collectively exhaustive*, meaning they cannot both be true at the same time and together they cover all possible cases.

The two competing hypotheses are the starting point of the NHST procedure, which involves designing a scientific experiment, performing the experiment, collecting sample data, and doing the statistical analysis based on the sample data to reach one of two possible conclusions:

- We "reject the null hypothesis" whenever we find the observed data to be very unlikely to occur by chance under $H_0$. This is the conclusion that scientists hope to reach at the end of their statistical analysis, because it means the observations cannot be explained by the baseline model.
- On the contrary, we "fail to reject the null hypothesis" when the observed sample data can be explained by the null hypothesis. In these cases, there is no need for an alternative theory beyond the baseline model we already have.

The NHST procedure is shown in simplified form in Figure 33.1.

When we reach the conclusion "reject $H_0$" we have not actually shown $H_A$ to be true. Rejecting the null hypothesis is just a necessary prerequisite for further study of alternative hypotheses.

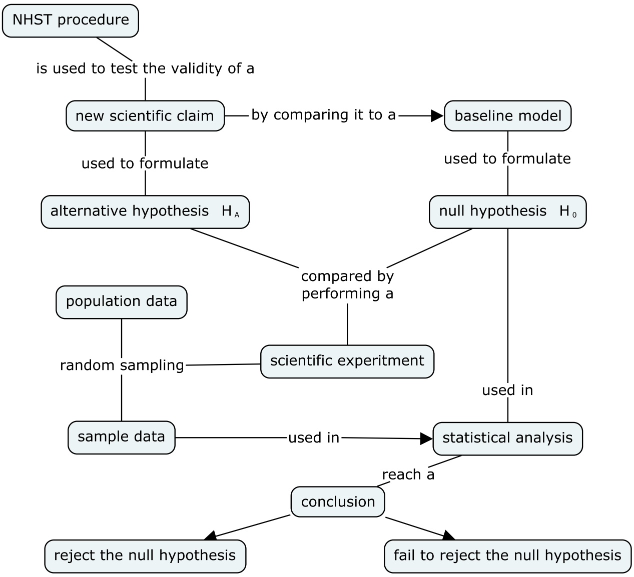

**Figure 33.1:** High-level overview of the NHST procedure, which starts with two competing hypotheses $H_A$ and $H_0$ and reaches one of two possible conclusions. The rest of this chapter is dedicated to understanding the details of each step.

### 33.1.2 Statistical modelling and assumptions

So how exactly do we perform the statistical analysis necessary to reach the correct conclusion? The answer is that we'll statistical modelling techniques to describe the distribution of hypothetical samples taken from the population *under the assumption that $H_0$ is true*. We can then compare the characteristics of the real sample we obtained from the population and make a judgement about how likely or unlikely it is to occur under $H_0$.

Statistical models embody what we know or *assume* to be true about the population and the data sample. We often describe the population in an idealized way simple probability distributions described by a few parameters. For example, a common assumption is that the population data is normally distributed according to $X \sim \mathcal{N}(\mu, \sigma^2)$ with known variance $\sigma^2$ and an unknown mean $\mu$.

In reality the population is probably not exactly symmetrical or perfectly gaussian, but making making these simplifying assumptions allows to work with simple formulas for estimators, sample statistics, and easily compute probabilities. We can't chose assumptions *just* for the sake of convenience though; models must be reasonable approximations of populations, otherwise our conclusions might be faulty.

### 33.1.3 Test statistic

Starting with a clear picture of the hypotheses we want to compare and the statistical modelling assumptions we're making allows us to "do the math" for the statistical analysis procedure in advance of performing the experiment and collecting the data. Indeed, over the years statisticians have "done the math" for numerous statistical analysis scenarios in order to simplify the task of working scientists to just computing a single number, called the *test statistic*.

We'll start by with the definitions:

- *Estimator*: a generic term applied to functions computed from data samples like $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$. For example, the sample mean is $\overline{x} = g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- *Test statistic $t, z, d, \ldots$*: an estimator with applications to the NHST analysis procedure. The term *test static* usually refers to the value of the estimator computed from a particular sample $\mathbf{x}$. For example the value of the $z$ test statistic is computed using $z = h(\mathbf{x}, \mu, \sigma)$, where $\mu$ and $\sigma$ are the population parameters.
- *Sampling distribution of the test statistic $T, Z, D, \ldots$*: a random variable that describes the test statistic computed on random samples like $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, where each $X_i$ is a random sample taken from the population. The sampling distribution of the $z$ statistic is described by the random variable $Z = h(\mathbf{X}, \mu, \sigma)$.

Note the distinction between the lowercase $z = h(x_1, x_2, \ldots, x_n, \mu, \sigma)$, which is a particular value of the test statistic computed from a given sample $\{x_1, x_2, \ldots, x_n\}$, and the uppercase $Z = h(X_1, X_2, \ldots, X_n, \mu, \sigma)$, which is a random variable since it is computed from a random sample $\{X_1, X_2, \ldots, X_n\}$. The actual computation we perform in each case is described by the same function $h(\cdot, \mu, \sigma)$, but the inputs are completely different.

It's important that you understand the notion of a sampling distribution well before you proceed with this chapter. If you're not 100% on top of the idea of random variable defined in terms of hypothetical random samples of size $n$ taken from a population, it's recommended you go back to the previous chapter and an review the figures and examples given. It might also be a good idea to solve some practice problems to get some hands on experience with the concept. The relationships between the characteristics of the general sampling distribution $Z$ and particular values of the test statistic $z$

is at the heart of the NHST logic. This is what will ultimately allow us to judge how likely it is to observe a particular value of the test statistic $z$ under the null hypothesis, and thus make a decision about the two alternatives under consideration.

The formulas for test statistics commonly have the following structure:

$$\text{test statistic} = \frac{\text{estimator} - \text{mean}}{\text{standard error}},$$

where some base estimator quantity is "standardized" by subtracting the mean and dividing by the standard error. For example, the formula for the $z$-score is given by $z = \frac{\overline{x}-\mu}{\sigma}$. This transformation allows us to compute the standardized $z$-score from the sample mean $\overline{x}$ for *any* normally distributed population.

The use of standard test statistics is the main mathematical tool that makes it possible to perform NHST. There are many types of test statistics used for different type of statistical analysis and replying on different sets of assumptions, but all of them boil down to the same idea: a test statistic measures how far data deviates from what is expected under the null hypothesis.

### 33.1.4 Critical value and decision rule

The purpose of reducing the data analysis task to the computation of single number, the test statistic, is to allow scientists to use a very simple decision rule. The decision rule used in NHST requires only a simple numerical comparison to a predefined threshold, called the *critical value*. Let's take the time to formally define the notion of a critical values and all the related concepts within the statistical testing procedure.

- *critical value* CV: a specific value for the test statistic that we use to decide whether to reject or retain $H_0$. The critical value can be determined in advance of the experiment, before we have collected data or computed any test statistics.
- *decision rule*: a simple algorithm that determines which of the two possible NHST conclusions we declare. The decision rule performs a comparison of the test statistic $z$ obtained from given sample and the pre-determined critical value. If the value of test statistic computed from the sample is greater than the critical value, we will reject the null hypothesis. If the observed value of the test statistic is smaller is smaller than the critical value then we fail to reject the null hypothesis.
- *critical region*: the set of values for the test statistic that will lead us to reject $H_0$. A test statistic value that falls within the critical

region tells us that $H_0$ a very unlikely explanation for the data observed.

- *region of acceptance*: the set of values for test statistic that will lead us to retain $H_0$ as the most likely explanation for the observed data. Another way to describe this outcome is to say "fail to reject $H_0$," meaning observed value of the test statistic is not very unlikely under $H_0$ so there is no need to consider any alternative hypotheses.

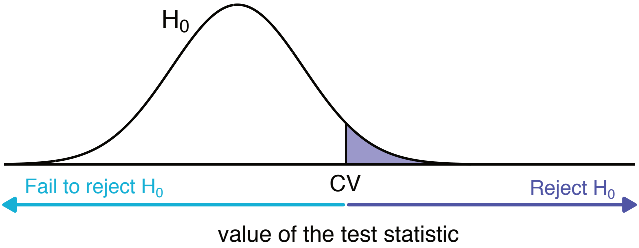

**Figure 33.2:** Illustration of the *region of acceptance* and the *critical region* used to draw conclusion as part of the NHST procedure. If the value of the test statistic computed falls above the critical value, our conclusion will be to reject $H_0$. If the value of the test statistic is below the critical value, we fail to reject $H_0$.

The region of acceptance is the complement of the critical region. The critical values are the boundaries between the critical region and the region of acceptance, as illustrated in Figure 33.2. The shape of the critical region is determined by the type of comparison encoded in the two hypotheses. Figure 33.2 shows an example of an *upper-tailed* rejection region, but *lower-tailed* and *two-tailed* rejection regions also exist (see discussion on page 60).

The notion of critical value and critical region apply generally to instances of the NHST procedure. To keep things simple, in this chapter we focus exclusively the $z$-test. The $z$-test is a general-purpose statistical analysis procedure based on the $z$ test statistic, which is nothing other than the standard normal distribution which you're familiar with:

- $z_q$: A value such that $F_Z(z_q) = q$, where $F$ is the CDF for the standard normal distribution $Z \sim \mathcal{N}(0,1)$. The normal distribution plays a central role and must often use compute quantities like $F_Z^{-1}(q)$ and $F_Z^{-1}(1-q)$, to the shorthand notation $z_q$ and $z_{1-q}$ is very convenient.

- The critical values of the $z$-test are specified in terms of the normal distribution $\text{CV}_z = z_{1-\alpha}$ where $\alpha$ is the Type I error parameter, which we'll formally define in the next section.

### 33.1.5 Errors

There are two types of mistakes you could make when following the NHST procedure:

- *Type I error* occurs when $H_0$ is true, but you reject $H_0$. This is also called a *false positive*.
- *Type II error* occurs when $H_A$ is true but you fail to reject $H_0$. This is also called a false negative.

| | | $H_0$ is true | $H_A$ is true |
|---|---|---|---|
| decision reached | Reject $H_0$ | Type I error<br>false positive<br>Probability = $\alpha$ | Type II success<br>true positive<br>Probability = 1 - $\beta$ |
| | Fail to reject $H_0$ | Type I success<br>true negative<br>Probability = 1-$\alpha$ | Type II error<br>false negative<br>Probability = $\beta$ |

This table shows the four possible outcomes of following the NHST procedure. There are two types of success cases and two types of errors cases, depending on the decision you reach and which hypothesis is actually true.

We decide how tolerant we are to making Type I and Type II errors in advance of doing the statistical analysis by choosing the error rate parameters:

- *Type I error rate* $\alpha$: The probability of making a Type I error.

$$\alpha = \Pr\big(\text{reject } H_0 \mid H_0 \text{ is true}\big).$$

  The number $\alpha$ is also called the *significance level* or the *rejection level*. For example, we could choose $\alpha = 0.05$. This means that if $H_0$ is actually true, we have only a 5% chance of rejecting a true null hypothesis.
- *Type II error rate* $\beta$: The probability of making a Type II error:

$$\beta = \Pr\big(\text{fail to reject } H_0 \mid H_A \text{ is true}\big).$$

In words, the coefficient $\beta$ describes the probability of false-negatives—when $H_A$ is true, but following the NHST procedure leads us to retain $H_0$.

Statisticians more commonly talk about the Type II error rate in the form of it's inverse, *Statistical Power*.

- *Statistical Power* $(1-\beta)$: The ability to detect a pattern in the sample if a pattern exists in the population.

$$\text{power} = (1-\beta) = \Pr\big(\text{reject } H_0 \mid H_A \text{ is true}\big).$$

For example, if we choose $\beta = 0.2$, this means we'll have $1 - 0.2 = 0.8 = 80\%$ chance to correctly reject the null hypothesis.

The parameters $\alpha$ and $\beta$ are specific to the particular statistical question you are studying, and can vary from experiment to experiment.

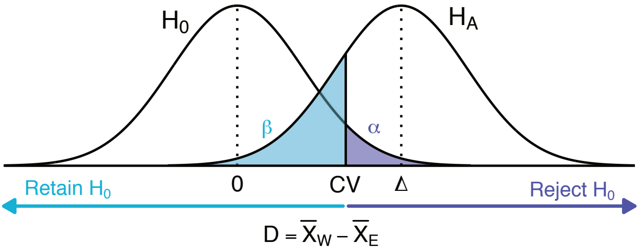

**Figure 33.3:** This diagram a sketch of the distribution of the $z$ scores under the null hypothesis and the alternative hypothesis. The tails of the distributions that correspond to Type I and Type II errors are highlited. The critical value $\text{CV}_z$ represents the boundary critical region.

As you see in Figure 33.3, choosing the critical value $\text{CV}_z$ is of central importance for the statistical test, since it affects both Type I error rate $\alpha$ and the Type II error rate $\beta$. The choice of sample size $n$ is equally important. Using larger sample sizes will reduce the variance of the sampling distributions and thus make the two probability distributions under the two hypotheses easier to distinguish.

### 33.1.6 Calculating the sample size

In order to satisfy the target false-positive error rate $\alpha$ and target power of $1-\beta$, we require a minimum sample size $n$ for our statistical analyses. For the purpose of better flow of explanations, we'll delay the detailed discussion about the find-$n$-from-$\alpha$-and-$\beta$ calculation until the end of the chapter, where where we'll show three different ways of doing the calculations.

### 33.1.7 Reporting results

Let's now talk about all the juicy information we can draw from our statistical tests.

- *p-value*: the probability you would get a value of the test statistic *at least as extreme* as the one you calculated from your sample purely by chance, assuming that $H_0$ is true. For an upper-tailed $z$-test, the $p$-value is
$$p = \Pr(\, Z \geqslant z \mid H_0) \,, \text{ where } \; z = g(x_1, x_2, \ldots, x_n).$$
- *Effect size*: a statistic that estimates the size or magnitude of the parameter predicted by the alternative hypothesis. For example, say your $H_A$ claimed that mean of one population $\mu_1$ differed from another $\mu_2$. You could describe the effect size as the difference between two sample means $\overline{x}_1 - \overline{x}_2$.
- *Confidence interval*: a range of numbers indicating the precision of some estimate. For example the $(1-\alpha)$-confidence interval for the population parameters $\theta$ is
$$\text{CI}_{1-\alpha} = [\ell, u], \quad \text{ where } \Pr(\ell \leqslant \Theta \leqslant u) = 1 - \alpha,$$
where $\ell$ and $u$ depend on the estimator value $\hat{\theta}$ obtained from the observed sample and the sampling distribution of the estimator $\Theta$. The $(1-\alpha)$ is the proportion of intervals that would contain the true value of $\theta$ if we resampled and reran analyses many times. A confidence interval tells us about the precision of our estimate.

The $p$-value is a measure of the the strength of evidence against $H_0$. The smaller the $p$-value, the less probable it is to obtain a sample at least as extreme as your observed sample under $H_0$.

Note that a scientific result can be statistically significant (small $p$-value), but of no *practical significance* if the effect size is very small. To determine the practical significance of your results, you must judge how important your estimated effect size is in the context of the system or situation you're looking at.

### 33.1.8 NHST procedure

Now that you have some understanding of the "ingredients" required for NHST, let's take a look at the steps of the NHST "recipe."

Step 1: Set up two competing claims that define the property you want to test ($H_0$ vs. $H_A$). Consider assumptions behind each hypothesis and choose the appropriate null and alternative models and data collection method.

Step 2: Decide how conservative or risky you want to be in your decision-making by choosing a statistical significance level ($\alpha$) and statistical power ($1 - \beta$).

Step 3: Calculate the required sample size $n$ you'll need and the critical value $\text{CV}_z$, then collect the necessary data.

Step 4: Using your data, check your assumptions.

Step 5: Decide whether to *reject* or *fail to reject* the hypothesis $H_0$ based on the comparison of the value of the test statistic $z = g(x_1, x_2, \ldots, x_n)$ and the critical value $\text{CV}_z$.

Step 6: Measure the strength of your evidence. Calculate the $p$-value associated with your test statistic $z$, report the effect size and its associated confidence interval. Plot your results.

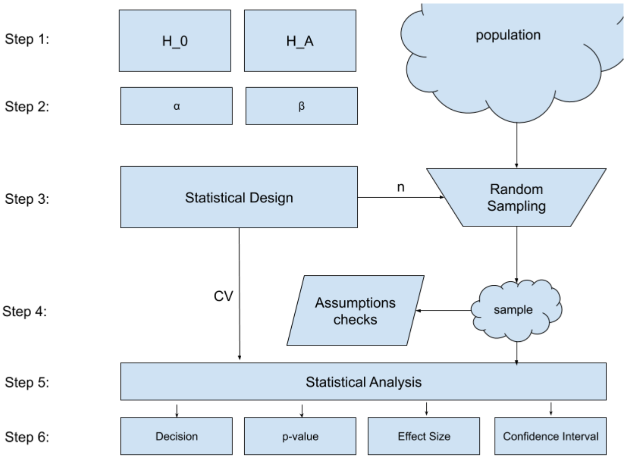

**Figure 33.4:** A visual overview of the data flow between the different steps of the NHST procedure. On the left we see theoretical calculations and design considerations, while on the right the actual data procedures.

Note the key step happens in Step 5 where we apply the decision rule based on comparing the value of the test statistic to the predetermined threshold value $\text{CV}_z$. In other words, we decide based on the value of $z$ statistic if the evidence is strong enough to reject $H_0$. The main "feature" of the NHST procedure is that it gives you the ability to make precise, quantitative claims backed by well-established statistical analysis methods.

## 33.2 Statistical model for beer prices

The best way to understand NHST is to apply it to a real-world situation. Let's focus our attention on a particular example of a decision making process. We'll go through the six steps of the NHST procedure for a hypothetical scenario where we perform statistical analysis on data about beer prices. First we'll start by applying the procedure "mechanically" by plugging numbers into equations, then later on in the chapter we'll look "under the hood" to show all the gory math details of the probabilistic reasoning behind each equation. Don't worry you can handle this.

Before we get started, we should warn you that we're about to make some pretty wild assumptions: we'll assume that beer prices are normally distributed and that we know the true population variance for the distribution. On top of that, we'll use a dubious data collection protocol that violates a fundamental stats assumption (that observations are randomly selected and independent). These unrealistic assumptions and reckless sampling procedures serves a noble purpose—to make your first contact with the world of NHST a little bit gentler. Thanks to the simplifying assumptions about normality and the known variance, we'll be able to use the *z-test for difference of sample means* for the data analysis, which the simple enough for you to understand all the formulas. Don't worry, we'll have plenty of time to discuss more realistic statistical testing procedures in Chapter ??.

### 33.2.1 Hipster problems

Imagine you feel like heading out for a drink after a long day of studying statistics. You live in the West End of the city. Because of rampant gentrification, all your favourite local bars have either closed or upped the price of draught IPAs, charging an outrageous eight or nine dollars per pint. You've only got six dollars and change to spare, but don't worry, there is hope. Your savvy friend Kayla says she went out in the East End last week and got a super hoppy small-batch craft beer for only $5.25. That sounds way too good to be true. Is there a rigorous, numerical procedure that could help you test whether Kayla's claim "beer is cheaper in the East End" is true? The Null Hypothesis Significance Testing recipe is what you need. Let's collect some data and follow the six steps of NHST to decide whether it's worth biking to the East End in search of cheap beer.

### 33.2.2 Step 1

During the first step we want to clearly write down the two hypotheses under consideration and state the assumptions we're making about our population and sampling procedure.

The alternative hypothesis $H_A$ is that East End beer is cheaper on average than West End beer. This is the "scientific" claim we want to investigate. The null hypothesis $H_0$ describes the case where mean beer price is the same on both ends of the city, or worse, that East End beer is actually more expensive. This is the statement we aim to reject. If we call the mean price of beer on the West End $\mu_W$, and the mean price of beer on the East End $\mu_E$, we can write the two hypotheses as follows:

$$H_0 : \mu_W \leqslant \mu_E, \qquad\qquad H_A : \mu_W > \mu_E.$$

Note that these two hypotheses are mutually exclusive—both can't be true at the same time—East End beer cannot be less expensive *and* simultaneously more expensive than West End beer. They are also collectively exhaustive—together they cover every possible relationships between $\mu_W$ and $\mu_E$. We can state these same hypotheses in terms of the difference between the means. Our alternative hypothesis ($H_A$) is that the difference in beer prices is positive, while the null hypothesis ($H_0$) is that the difference in beer prices is zero or negative:

$$H_0 : \mu_W - \mu_E \leqslant 0, \qquad\qquad H_A : \mu_W - \mu_E > 0.$$

Both hypotheses are stated in terms of the difference in beer prices for the *population parameters* $\mu_W$ and $\mu_E$, which are unknown. We'll estimate the value of $\mu_W - \mu_E$ by collecting two samples of beer prices: one from the East End $x_E = [x_{E1}, x_{E2}, \ldots, x_{En}]$ and one from the West End $x_W = [x_{W1}, x_{W2}, \ldots, x_{Wn}]$. We'll compute the mean of each sample and calculate the difference between the sample means:

$$\overline{x}_W - \overline{x}_E = \tfrac{1}{n}\sum_{i}^{n} x_{Wi} - \tfrac{1}{n}\sum_{j}^{n} x_{Ej}.$$

We also assume that:

- The sampling method for each population is simple random sampling.
- The observations in each sample are independent.
- Beer prices in the East and West End are normally distributed with variance $\sigma^2 = 5$.

These assumptions allow us to use the $z$-test for a difference between means. We will calculate a $z$-score for our sample difference $\overline{x}_W - \overline{x}_E$ under the assumption that $H_0$ is true. If the $z$-score is high, meaning the sample difference is quite large and therefore unusual under $H_0$, we can reject the idea that $\mu_W - \mu_E \leqslant 0$. We will bike East and get beer! If the $z$-score is close to zero, it means the sample difference is consistent with $H_0$. In that case, we'll retain $H_0$ and stay home and drink water.

### 33.2.3 Step 2

In order to quantify and control the likelihood of coming to a wrong conclusion we must choose appropriate values for the parameters $\alpha$ and $\beta$. Let's go over the two ways you could mess up.

**Type I error** If the null hypothesis is true but we reject the null hypothesis, we would be making a Type I error. You would be making a Type I error if you went the East End because you believed there was cheap beer there, but you actually only found expensive beer. Let's decide we're willing to make this mistake in one out of every 20 cases. We choose the value $\alpha = \frac{1}{20} = 0.05$ as the upper bound on this type of error:

$$\Pr\big(\text{Type I error}\big) = \Pr\big(\text{reject } H_0 \mid H_0 \text{ is true}\big) \leqslant \alpha = 0.05.$$

**Type II error** If the alternative hypothesis is actually true but we fail to reject the null hypothesis we would be making a Type II error. This second type of error would be the situation where you stay home because you believe that beer on the East is expensive, but you're wrong and you miss out on actual cheap beer. Suppose we decide this type of error is not quite as bad as a Type I error. We're willing to tolerate at a probability of $\beta = 0.2$ at most. This choice of $\beta$ means our statistical test will have *statistical power* at least $(1-\beta) = 0.8$:

$$\text{Statistical Power} = \Pr\big(\text{reject } H_0 \mid H_A \text{ is true}\big) \geqslant (1-\beta) = 0.8.$$

| | | Beer prices are the same | East End beer is cheaper |
|---|---|---|---|
| decision reached | Bike to East End | Biking for nothing<br>(Type I error)<br>Probability = 5% | Yay, cheap beer!<br>(Type II success)<br>Probability = 80% |
| | Stay home | No time wasted<br>(Type I success)<br>Probability = 95% | Miss out on cheap beer<br>(Type II error)<br>Probability = 20% |

This table shows the possible outcomes when you follow the NHST procedure to make a decision about your beer plans. There are two types of errors that can occur and two success scenarios where you make the right decision.

### 33.2.4 Step 3

Having chosen the parameters $\alpha = 0.05$ and $\beta = 0.2$ for our statistical experiment, we can now compute the sample size $n$ and critical value $z_{1-\alpha}$ to use for our statistical test.

**Finding the critical value** The critical value $z_{1-\alpha}$ is the value such that

$$\Pr\big(Z \geqslant z_{1-\alpha} \mid H_0 \text{ is true}\big) \leqslant \alpha = 0.05.$$

Since we're talking about standard $z$-score, you can find the value $z_{1-\alpha}$ in a lookup table, or you use Python module `norm` as follows:

```
>>> from scipy.stats.distributions import norm
>>> Z = norm(0, 1)
>>> Z.ppf(0.95)
1.6448536269514722
```

The decision rule for the statistical experiment is the following:

$$\begin{cases} \text{if } z > 1.645 & \Rightarrow \quad \text{reject } H_0 \\ \text{if } z \leqslant 1.645 & \Rightarrow \quad \text{fail to reject } H_0 \end{cases}$$

In other words, if the difference in beer prices has a $z$-score of 1.645 or more, you'll reject the idea that the means of beer prices between East and West are the same (or less). Following this approach ensures your Type I error rate stays below 5%, since $\Pr(Z \geqslant 1.645) = 0.05$. If in reality $\mu_W - \mu_E \leqslant 0$, we'll only be unlucky enough to get a sample that gives us a value of $z > 1.645$ in one out of 20 cases.

**Calculating the sample size** In order to keep your target Type I error rate at $\alpha = 0.05$ *and* also have power $1 - \beta = 0.80$, you'll need to have a sufficiently large sample size $n$. Power is influenced by the sample size $n$ and by the effect size (the amount by which average beer prices differ). The ability to detect a difference in beer prices depends on how large those differences are. To carry out the power calculation, we must decide on the minimum difference in beer prices $\mu_W - \mu_E$ that we want to be able to detect, if it exists.

You count your pocket change and decide it's worth biking to the East End if beer is at least \$2.62 cheaper there on average. This is a somewhat arbitrary decision and it is motivated by the specifics of this scenario and not some some absolute mathematical truth. The range between \$0 and \$2.62 is called your "zone of indifference." You might not have the power to detect differences this small, but you don't really care since you're not going to bike to the East End for a difference less than \$2.62 anyways.

To find out the sample size $n$ you need to satisfy these conditions, you can use this formula:

$$n = \frac{(|z_{1-\alpha}| + |z_\beta|)^2(\sigma_E^2 + \sigma_W^2)}{(\mu_W - \mu_E)^2},$$

where $\sigma_W^2$ and $\sigma_W^2$ are the assumed variances of west and East End beer prices (both 5), $\mu_W - \mu_E$ is your minimum detectable difference (2.62), $z_{1-\alpha}$ is your critical value (1.645), and $z_\beta$ is the critical value of the normal distribution at $\beta$ ($z_{0.20} = -0.84$). We'll explain where this formula comes from later on; for now let's just just plug in the values:

$$n = \frac{(1.64 + 0.84)^2(5+5)}{2.62^2} = 8.959851 \approx 9.$$

Great! We now know that with nine sample beer prices from the East End and nine samples from the West End, we'll be able to correctly reject the null hypothesis in at least 8 out of 10 tests, when the difference in beer prices is 2.62 or more:

$$\Pr\big(Z \geqslant 1.645 \mid \mu_W - \mu_E \geqslant 2.62\big) \geqslant 1 - \beta = 0.80.$$

**Collecting data** Suppose you text nine friends that live in the East End, and 9 friends in the West End and ask them to check the price of beer in the pub closest to their house. While you wait for texts to come back, you wonder whether this data collection strategy violates two of our three assumptions: that samples are random and that observations are independent. If you're only getting beer prices from pubs that happen to be located near your friends, does every beer price really have an equally opportunity of being selected in your sample? Since your friends have similar tastes and hang out at similar places, won't your sample be biased? Before you can finish that thought, your phone starts pinging with sample beer prices:

$$x_E = [7.7, 5.9, 7.0, 4.8, 6.3, 6.3, 5.5, 5.4, 6.5],$$
$$x_W = [11.8, 10.0, 11.0, 8.6, 8.3, 9.4, 8, 6.8, 8.5].$$

Finally you've got data!

### 33.2.5 Step 4

Now that you have your samples, it's time to check our assumptions. Recall that we assumed the populations of beer prices where these samples are taken from are normally distributed with variance $\sigma_E^2 = \sigma_W^2 = 5$.

Let's start by making two histograms to check that beer prices are normally distributed, as shown in Figure 33.5.

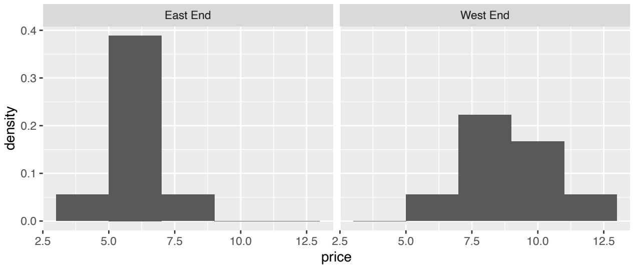

**Figure 33.5:** Histogram plots showing the distribution of beer prices from the two samples.

These histograms look reasonably bell-shaped, but it's hard to tell given that you've only got 9 samples per city end. You decide the it's good enough.

Next you calculate the variance of each sample of beer prices. The sample variance for the East End is $s_E^2 = 0.7702778$ and the variance for the West End is $s_W^2 = 2.440278$. The two sample variances are very different from each other much smaller than the assumed population variance 5. This should normally make us reconsider our assumptions and the use of the $z$-test, but let's say you're very thirsty and decide ignore these discrepancies. You're willing to go ahead with the $z$-test even if the assumptions are not satisfied because there is beer on the line!

### 33.2.6 Step 5

With your data you can compute the sample means $\overline{x}_E$ and $\overline{x}_W$:

$$\overline{x}_E = \tfrac{1}{9}\sum_{i=1}^{9} x_{Ei} = 6.155, \qquad \overline{x}_W = \tfrac{1}{9}\sum_{i=1}^{9} x_{Wi} = 9.155,$$

and compute your $z$ test statistic:

$$z = \frac{(\overline{x}_W - \overline{x}_E) - (\mu_W - \mu_E)}{\sqrt{\frac{\sigma_W^2}{n} + \frac{\sigma_E^2}{n}}}$$

Under $H_0$, $\mu_W - \mu_E$ could be any value less than or equal to zero. Because zero is the most extreme value that's still possible under $H_0$, this is the number we'll use for the hypothesized difference in beer prices.

$$\begin{aligned} z &= \frac{(9.155 - 6.155) - (0)}{\sqrt{\frac{5}{9} + \frac{5}{9}}} \\ &= 2.84605. \end{aligned}$$

Having obtained the value of the test statistic $z = 2.84605$, we can finally make our decision. Since the value of the test statistic is greater than the critical value $z_{1-\alpha} = 1.645$, our conclusion is to reject the null hypothesis. In other words, observing a difference of $z = 2.84605$ standard errors above the hypothesized difference of 0 is very unlikely if the null hypothesis is true. This means it's plausible that beer is cheaper in the East End. Hop on that bike and go get you some cheap beer—the scientific method demands it!

### 33.2.7 Step 6

But wait, before you get excited about this difference in beer prices and start looking for your bike helmet, there is some really important information you need to consider. How strong is your evidence of cheap East End beer? More importantly, *how much* cheaper is that East End beer? Let's calculate the $p$-value, effect size, and a confidence interval to find out.

You can think of the $p$-value as a rough measure of the strength of evidence against $H_0$. It's the probability of observing a value of the $z$ statistic equal to or more extreme than our result $z = 2.84605$, assuming that the null hypothesis of $\mu_W - \mu_E \leqslant 0$ is true, and that all our statistical assumptions are met. The $p$-value corresponds to the area under the curve of the null distribution,

$$p = \int_{z}^{\infty} f_{Z_0}(x)\, dx = 1 - F_{Z_0}(z),$$

where $F_{Z_0}$ is the cumulative density distribution the test statistic under the null hypothesis $Z_0 \sim \mathcal{N}(0, 1)$. We'll use the trusty `norm` function from `scipy.stats.distributions` to compute the $p$-value:

```
>>> from scipy.stats.distributions import norm
>>> z = 2.84605
>>> 1 - norm.cdf(z)
0.0022132629289599204
```

The $p$-value 0.0022 tells us that if the price of beer on the East and West End were actually equal, we would end up with a sample that gives us a $z$-score of at least 2.84605 in only about 22 out of 10000 tests.

A low $p$-value only tells you that you're likely to find cheaper beer in the East End, but it doesn't say anything about the magnitude of the difference. The *effect size* is the estimated difference in average price, which will certainly impact your choice. For instance, if the estimated difference in beer prices was only 5 cents, you may question whether it's worth risking your biking on the streets for 45 minutes just to save a nickel per pint. In other words, you want to know if the difference in beer prices is of *practical significance*.

We obtain the estimated difference in beer prices by computing the difference between the sample means

$$\overline{x}_W - \overline{x}_E = 9.155 - 6.155 = 3.$$

Based on the two samples we collected, we estimate that East End beer is about $3 cheaper on average than West End beer. This is called a *point estimate*, meaning it corresponds to a single value that is our "best guess" of the true difference in beer prices.

We can also calculate a *confidence interval* for this estimate to get an idea of how precise it is. Recall that a $100(1-\alpha)\%$-confidence interval (CI) describes a range of values that contain the true value with probability $(1-\alpha)$. We will use a one-sided confidence interval to go along with our one-sided hypothesis test. We only care about the lower bound of the CI because we only care about the lower end of the precision of our estimate. If our precision estimate tells us that East End beer might not actually be as cheap as we estimated, that could influence our decision. On the other hand, we will not change our decision if East End beer is likely to be even cheaper than expected. The formula for a one-sided confidence interval is given by

$$\mathrm{CI}_{1-\alpha} = \left( (\overline{x}_W - \overline{x}_E) - z_\alpha \sqrt{\tfrac{\sigma_E^2}{n} + \tfrac{\sigma_W^2}{n}},\ \infty \right),$$

Putting all the values we know into the confidence interval formula we obtain

$$\begin{aligned} \mathrm{CI}_{0.95} &= \left( 3 - 1.645 \cdot \sqrt{\frac{5}{9} + \frac{5}{9}},\ \infty \right) \\ &= (3 - 1.733982,\ \infty) \\ &= (1.266018,\ \infty). \end{aligned}$$

This interval $(1.266018, \infty)$ is likely to contain the true difference in average beer prices with 95% confidence. In other words, we're 95% confident that that the true mean difference in beer prices is $\geqslant$ \$1.27. That lower bound is rather low. In fact, it dips into our "zone of indifference." Recall that we're interested in biking to the East End only if the expected beer prices are \$2.62 cheaper than in the West End. This is good to keep in mind so you won't be disappointed.

Remember that the interval is a random variable and the true mean difference in beer prices is fixed. That means if we repeated this test with new data many times, the CI would capture the true mean difference in beer price in about 95% of cases. In other other 5% of cases, the lower bound of the confidence interval would end up higher than the true mean.

To test our beer-pothesis, we used the rule:

$$\text{reject } H_0 \text{ if } z > z_{1-\alpha}.$$

It would be equivalent for us to use a rule involving $p$-values:

$$\text{reject } H_0 \text{ if } p < \alpha$$

or to use confidence intervals:

$$\text{reject } H_0 \text{ if CI does not contain } \mu_W - \mu_E = 0$$

In all of these cases, our result is statistically significant when we reject $H_0$.

### 33.2.8 Reporting your results

No matter whether you choose your test statistic, $p$-value, or CI to decide on statistical significance, you should include all these values when reporting your results, along with your choice fo the values of $\alpha$ and $\beta$, your assumptions, and your data collection protocol. We don't just want to reach a conclusion about the existence of a pattern, after all, we want to be able to explain how certain we are of our conclusion and quantify the size of the pattern we have observed.

Say you want to convince a very skeptic friend to join you on a beer run to the East End. You tell your friend you used the formal hypothesis testing procedure with a false positive rate of $\alpha = 0.05$ and concluded that it's unlikely that the beer prices are the same. You tell them the $z$-score you calculated and explain that if beer prices *were* the same, you'd only get a $z$-score this high in 22 out of 10000 tests. Your friend obviously asks about how much cheaper the prices are, at which point you pull out the 3 dollars estimate for the effect size, which gets their attention. Being skeptical and statistics savvy, you friend claims that your estimate 3 is obtained from small samples and might not be accurate. They say "Sure, in the particular samples you obtained, you found an effect size of 3, but if you had obtained different samples you would have computed a different effect size." This is where you pull out the 95% confidence interval $(1.266018, \infty)$ and interpret it for your friend saying that you're 95% confident that beer prices are at least $1.26 cheaper in the East End than in the West End. To top it all off, you pull out some plots you just made to illustrate your point:

*plot or plots go here. I'm thinking paired dot plot (small sample size)...*

Now that's a convincing proposal even for the biggest skeptic.

### 33.2.9 Statistical reality check

Finally you convince your friend to join you on the bike ride. It's cheap beer time! You bike 45 minutes eastward, stop at the first patio you find and stare in horror at the prices you see posted. Eight dollars per pint! You pedal furiously from bar to bar, only to discover that prices are pretty much the same as in your neighbourhood. Sweaty and beerless, you ask yourself what went wrong? You followed all the steps of the NHST procedure, but still somehow reached a wrong conclusions: beer prices in the East End are not actually cheaper than in the West End. In other words, you made a Type I error.

The first thing to remember is that NHST doesn't guarantee a correct outcome, it just gives you an bounded estimate of the a Type I error that you can expect if you follow the NHST procedure. Remember that $\alpha = 0.05$ corresponds to 5% false-positive rate when we observe extreme values in the sample and falsely reject $H_0$, even when $H_0$ is true. You samples of beer prices could just randomly have come out extra different. The East End sample could just have happened to be well below the true population average. Although this should only happen in 5% of cases, it's still possible that it happens.

More importantly, recall that all of these calculations were made under some very specific assumptions. The probabilities of making Type I and Type II errors only hold if your assumptions are valid. Let's start with how we collected data. Recall from [chapter about collecting data], that for a sample to be truly random, each price of beer in the city should have an equal chance of being included in your sample. A better strategy in this situation would be to get a list of all the bars in your city, then have a computer randomly select a portion of them to use in your sample. You could then contact the bars directly to ask for their beer prices, instead of texting your friends.

Next, it was pretty unreasonable to assume beer prices are normally distributed. There are lots of variables in the natural world that tend to follow a normal distribution like human heights, but not all things are normally distributed. We can't assume that beer prices follow a normal distribution without some background knowledge.

If our sample size $n$ had been larger ($n \geqslant 30$), we could have gotten away with the population following some non-normal distribution. That's because the central limit theorem tells us that the distribution of the the estimators $\overline{X}_E$ and $\overline{X}_W$ will be normally distributed regardless of the distribution of the population.

Perhaps the most unreasonable assumption was assuming we *knew* the *true* population variance of beer prices on the East and West End. Actually, it's pretty hard to think of a real-life situation where we would know the population variance. This unrealistic assumption means the $z$-test wasn't the right statistical test for this data analysis scenario. In the next chapter, we'll learn about the $t$-test which uses estimates the variance computed from sample the data samples.

To understand why these assumptions are so important, we need to look under the hood of the NHST procedure we used, and see exactly how we modelled our population and test statistic.

* * *

Don't worry if you were not able to follow all the numeric calculations presented in each step. The essential thing is that you understand the general logic behind NHST and become familiar with the design, data collection, calculations, and reporting steps needed to perform statistical tests.

We intentionally skipped the details of the numeric calculations for the $p$-value and the confidence interval in the above text. These are straightforward to perform either using lookup tables or computer software. We'll talk more about that in discussion section. If you're impatient and interested in seeing how the calculations done, you can check this spreadsheet.

The key point to remember I want you to remember about the NHST procedure is that you must choose appropriate values for $\alpha$ and $\beta$ that are appropriate for your application. Don't just choose some values just because you've seen them in other papers. Once you choose the values for $\alpha$ and $\beta$, you'll be able to calculate the sample size $n$ you need, and only then you can start collecting data. Review Figure 33.4 on page 38 to remember the dependencies between the six steps of the NHST procedure.

## 33.3 Explanations

If you're seeing the NHST procedure for the first time it's natural if it seems very complicated and involved. Don't worry about that for now! You'll get a chance to become more familiar with the NHST procedure in the next chapter where we'll apply the NHST procedure to building many other scenarios that depend on different test statistics. By the end of this book you'll have seen the steps of NHST so often that you'll be able to do them in your sleep! For now I want you to focus on understanding the general idea behind NHST, and why each of the steps is required. The following extra explanations will hopefully help with that.

### 33.3.1 Hypothesis Testing as a Trial

Statisticians like to use double-negatives. When the $p$-value in some statistical test is high, they say things like, "We failed to reject the null hypothesis." Why don't they just say that they proved their null hypothesis true, and their alternative hypothesis false? Do they need to read the No Bullshit Guide to Clear and Grammatically Acceptable Sentence Structure? This unconventional sentence structure actually serves a purpose. In this case it reflects the philosophy underlying NHST that is also used throughout science—the null hypothesis testing recipe is similar to a trial in court.

Suppose Bob is undergoing a trial. Jurors must presume Bob is innocent ($H_0$) until there is extremely persuasive evidence that he's guilty ($H_A$). The burden of proof is on the prosecution: they have to show compelling evidence for the alternative hypothesis. If the prosecution can't present strong enough evidence to convict Bob, then the jurors' verdict will be "not guilty." Jurors will only reject $H_0$ in light of compelling evidence that Bob is guilty. Otherwise they declare Bob is not guilty. Note that they're not saying that Bob is innocent, they're just saying they haven't seen enough evidence for $H_A$. This is just like in "statistics duty," where scientists must go in assuming the null hypothesis is true unless otherwise convinced by their data. When a statistical test fails to reject the null hypothesis, this doesn't mean that we've shown the null hypothesis is true, just that the alternative hypothesis is not convincing.

Even with convincing evidence, is it really fair to say that Bob is "guilty"? Unlike in court, in statistics, we don't really think so. Statisticians are a little more conservative (or at least they should be). Rejecting the null hypothesis doesn't necessarily mean the alternative one is true either. It means the alternative hypothesis survives to be tested again. It means we should call a re-trial with new evidence. Even if you reject your null hypothesis, you should maintain some skepticism of your alternative, at least until more tests (with new data) reach a positive conclusion. Good science relies on replication—the idea that multiple, independent studies need to be carried and reach the same conclusion.

**Understanding the $p$-value** Okay, let's continue with the trial analogy. Suppose the prosecution presents a lot of very persuasive evidence making Bob look pretty guilty. An NHST statistician would think, "If Bob were innocent, this would be a super bizarre and unfortunate combination of coincidences." That's because they judge the probability of data, given the null hypothesis (innocence). This is exactly what the $p$-value tells you, but with numbers. It quantifies the probability of finding evidence at least as suspicious as what's been found, in a world where Bob is innocent ($H_0$ is true). Note that the $p$-value doesn't tell you the probability that Bob is innocent ($P(H_0|data)$), nor does it tell you the chance that you will falsely imprison Bob ($P(reject H_0|H_0)$).

**Choosing a significance level** At what point is the evidence compelling enough to reject the idea that Bob is innocent? In law, the threshold is "a reasonable doubt." In stats, it's $\alpha$. Remember that $\alpha$ is the risk of a Type I error, so to decide its value, you have to think about the cost of finding a pattern that isn't there. If the worst-case-scenario is false hope of cheap beer, then go wild and choose a recklessly high significance level ($\alpha = 0.1$ maybe). Just be careful in situations like Bob's trial. If you choose the significance level $\alpha = 0.1$, this means you're cool with trials that send one out of every 10 innocent people to jail. This is probably not how you want to run things, so remember to always think about the real-life consequences and choose a low $\alpha$ when Type I errors are critical.

**Choosing a level of statistical power** Also consider the consequence of the other possible error: not finding a pattern that is there. Are you failing to uncover cheap beer, are you setting a criminal free, or are you labelling a life-saving drug as worthless? When Type II errors are serious, you should choose lower $\beta$ values, which leads to tests with higher statistical power.

**The right balance between significance and power** The astute reader may still be wondering why exactly we don't choose $\alpha = 0$ and $1 - \beta = 1$ and avoid all errors. Dear perfectionist reader, it's okay to make mistakes. If you think about it, when you try to perfect one thing, there's always some other thing that you end up compromising (sleep, happiness, time spent learning about NHST, etc). It's the same with stats. Decreasing the chance of a Type I error increases your chance of a Type II error. In other words, statistical power decreases when significance levels are set lower. In both stats and life, the trick is to find a healthy balance.

To increase the power of a test without compromising $\alpha$, you need to use larger sample sizes or study phenomena with bigger effect sizes. That's because it's easier to detect patterns when you have a lot of data or if the the patterns are large. You can only control the potential effect size when you're running a manipulative experiment and you're able to increase the magnitude of the intervention (for example, increase the dosage in a medical trial). In many situations, increasing the sample size can be infeasible or prohibitively expensive. So what is a resource-limited statistician to do? You should do your best to actually quantify the cost of a Type I error compared to a Type II error. Is convicting innocent Bob ten times worse than setting guilty Bob free? Then your $\alpha$ should be ten times lower than your $\beta$.

**The alternative hypothesis is not on trial** Note that the details of the alternative hypothesis $H_A$ do not come into play in the NHST procedure, except in the consideration of statistical power. Indeed, all we have shown is that intervention $X$ causes "something different from the baseline model" so NHST doesn't really test any specific aspects of the alternative hypothesis. This is a known limitation of NHST procedure, which can be remedied by reporting estimates of the effect size observed and confidence intervals.

### 33.3.2 Looking under the hood

In this chapter we applied the general NHST procedure to a particular data analysis scenario where we compared the difference between two population means using the $z$-test. In this section we'll look in a little bit more detail at the probability models that underpin the formulas we used, in order to understand where the formulas come from.

The fundamental question we investigated concerns difference in beer prices $\mu_W - \mu_E$, which is an expression involving the difference between two *population parameters*. To known the true value of the expression $\mu_W - \mu_E$, we'd have to call *every* bar in the city and asked them for their beer list, then took the mean of the West End beers prices and subtracted the mean of the East End beer prices. As you can imagine, this is impractical to do since there are lots of bars and pubs in the city.

We can estimate the value of $\mu_W - \mu_E$ by collecting two samples of beer prices: one from the East End $x_E = [x_{E1}, x_{E2}, \ldots, x_{En}]$ and one from the west $x_W = [x_{W1}, x_{W2}, \ldots, x_{Wn}]$, finding the mean of each sample, then computing the difference between sample means:

$$d \equiv \overline{x}_W - \overline{x}_E = \tfrac{1}{n}\sum_{i}^{n} x_{Wi} - \tfrac{1}{n}\sum_{j}^{n} x_{Ej}.$$

We define the single-letter variable $d$ to represent the difference between sample means, in order to avoid writing $\overline{x}_W - \overline{x}_E$ all the time.

The statistic $d = \overline{x}_W - \overline{x}_E$ is an instance of the random variable $D \equiv \overline{X}_W - \overline{X}_E$, which is a function of the random samples $X_E = [X_{E1}, X_{E2}, \ldots, X_{En}]$ and $X_W = [X_{W1}, X_{W2}, \ldots, X_{Wn}]$. By modelling the distribution of individual sample values $X_{Wi}$ and $X_{Ei}$, we can get an idea of which values of $d$ are probable—this is called the sampling distribution of the random variable $D$.

Modelling $X_W$ and $X_E$ requires making assumptions about the two populations of beer prices. We assumed the price of beer on both ends of the city were normally distributed with known variance $\sigma^2 = 5$. These assumptions allowed us to model beer prices in the East End ($X_E$) and the West End ($X_W$) of the city as normal distributions ($\mathcal{N}$) with unknown means $\mu_E$ and $\mu_W$, and known variance $\sigma^2 = 5$:

$$X_E \sim \mathcal{N}(\mu_E, 5), \qquad\qquad X_W \sim \mathcal{N}(\mu_W, 5).$$

We can treat the sample beer prices we obtained as independent draws from $\mathcal{N}(\mu_E, 5)$ and $\mathcal{N}(\mu_W, 5)$, assuming we collected beer prices through a random sampling process and ensured that each observation was independent. The central limit theorem tells us the sample means for independent samples of size $n$ from a population with variance $\sigma^2$ are normally distributed with variance $\sigma^2_{\overline{x}} = \frac{\sigma^2}{n} = \frac{5}{n}$:

$$\overline{X}_E \sim \mathcal{N}\left(\mu_E, \tfrac{5}{n}\right), \qquad \overline{X}_W \sim \mathcal{N}\left(\mu_W, \tfrac{5}{n}\right).$$

The difference between the random variables $\overline{X}_W - \overline{X}_E$ is also normal with mean equal to the difference of means and variance equal to the sum of the variances for $\overline{X}_E$ and $\overline{X}_W$:

$$D \equiv \overline{X}_W - \overline{X}_E \sim \mathcal{N}\left(\mu_W - \mu_E, \tfrac{5}{n} + \tfrac{5}{n}\right).$$

The variable $D$ describes difference between random sample means $\overline{X}_E$ and $\overline{X}_W$.

Instead of working with the random variable $D$ to compute probabilities, we can "standardize" $D$ by subtract its mean $(\mu_W - \mu_E)$ (the difference under a given hypothesis), then divide by the its standard deviation:

$$Z = \frac{D - (\mu_W - \mu_E)}{\sqrt{\frac{10}{n}}}.$$

The resulting random variable $Z$ has the standard normal distribution $\mathcal{N}(0, 1)$, with mean zero and standard deviation one.

For every probability calculation that we might want to perform using the random variable $D$ there is an equivalent probability calculation we can carry out using the random variable $Z$. One of the nice properties of gaussian random variables is that they they can be transformed to the standard normal distribution, and we want to take advantage of this property to simplify simplify the following calculations:

- Whenever we need to compute the value of the cumulative distribution $F_D(d) = \int_{-\infty}^{d} f_D(x)\ dx$, we can instead compute the cumulative distribution $F_Z\left(\frac{d-\mu_D}{\sigma_D}\right)$, where $F_Z$ is the CDF of the standard normal $Z \sim \mathcal{N}(0, 1)$.
- Whenever you want to compute a value of the inverse cumulative distribution $F_D^{-1}(q)$, you can instead find the equivalent $z$-score, $z_q = F_Z^{-1}(q)$, then compute $\mu_D + z_q\sigma_D$ which is equal to $F_D^{-1}(q)$.

It's important to note that doing the change-of-variables transformation $D$ to $Z$ is not a required step, but simply a computational trick. If you look at Table ZZ in Appendix YY you'll see the table that contains all the values of the CDF of the standard normal distribution $Z$, which you can use to lookup any value of $z_q$ that you're interested in. The main benefit of standardization is that you can do probability calculations simply by "looking up" the appropriate values in Table ZZ. This means you can do statistical analysis without needing a computer—all the probability calculations have been done for you for the standard normal and recorded for you so don't need a computer. As you can imagine, in the early days of statistics when computers were not available, such computational "hacks" were all the rage since it allowed people to do statistical analysis using only basic algebra followed by table lookups.

In the modern day when computing power is plentiful, the need for tables of pre-computer probabilities for standard test statistics has decreased. Today we can easily do probability calculations with random variable $D \sim \mathcal{N}\left(\mu_D, \frac{10}{n}\right)$ just as easily as with the standard normal $Z \sim \mathcal{N}(0,1)$. Even if we no longer need the standardization procedure for computational purposes, it's still worth learning about the $z$-score for the procedural standardization it provides. By converting all possible normally distributed test statistics to standard test statistic $z$, we just need to learn about one set of formula for the NHST statistical analysis. In this chapter we learned about the $z$-test for comparing the difference between beer prices, but the exact same steps apply to differences between sample means from any two normal populations with known variance.

We could carry out the entire NHST procedure using the $d$-test statistic whose sampling distribution is described by the random variable $D \equiv \overline{X}_W - \overline{X}_E \sim \mathcal{N}\left(0, \frac{10}{9}\right)$. If we want to use the $d$-test as part of the NHST procedure, in Step 3 we'll need to find the critical value $\text{CV}_d$ we need to build the decision rule:

$$\begin{cases} \text{if } d > \text{CV}_d & \Rightarrow \quad \text{reject } H_0 \\ \text{if } d \leqslant \text{CV}_d & \Rightarrow \quad \text{fail to reject } H_0 \end{cases}$$

Recall that the critical value for the test is computed based on the formula $\text{CV}_d = F_D^{-1}(0.95) = 1.734$, which involves computing the inverse of the CDF of the random variable $D$. The value 0.95 corresponds to $1 - \alpha$, where $\alpha = 0.05$ is the maximum allowed probability of Type I error we chose in Step 2. The value of $F_D^{-1}(q)$ can obtained using the formula `=NORM.INV(q,0,SQRT(10/9))` in Excel, by calling `qnorm(q,0,sqrt(10/9))` in R, or by calling `norm.ppf(q,0,sqrt(10/9))` in Python after importing `norm` from `scipy.stats.distributions`.

### 33.3.3 Sampling distributions under the two hypotheses

The two hypotheses we consider as part of the NHST procedure, correspond to two "alternative realities" and thus two different probability models for the sampling distribution of the test statistic. Let's take a moment to derive the explicit formulas the probability distributions under each hypothesis. We'll show the probability distributions of both the non-standardized real difference between means $D$ (measured in dollars), and the standardized $z$-scores.

- The null hypothesis is that average beer prices are the same everywhere in the city or more expensive in the East:

$$H_0 : \quad \mu_W - \mu_E \leqslant 0.$$

- Under specific case of the null hypothesis ($H_0 : \mu_W - \mu_E = 0$), the sampling distributions of the test statistics $d$ and $z$ are

$$D_0 \sim \mathcal{N}(0, \sigma_D) \qquad \Leftrightarrow \qquad Z_0 \sim \mathcal{N}(0,1),$$

where $Z_0 = \frac{D_0 - 0}{\sigma_D}$ and $\sigma_D = \frac{5}{n} + \frac{5}{n} = \frac{10}{n}$.

- The alternative hypothesis is that beer prices are on average cheaper in the East End than in the West End of the city. We could make a very broad statement like "there is some difference," stated mathematically as:

$$H_A : \quad \mu_W - \mu_E > 0,$$

which states that average beer price is cheaper in the East End than in the West End, but doesn't say how much cheaper. A more precise way to state an alternative hypothesis is

$$H_A : \quad \mu_W - \mu_E = \mu_{D_A},$$

where $\mu_{D_A}$ is some unknown constant computed as the difference between the average beer prices of the populations $\mu_{D_A} = \mu_E - \mu_W$. For example, the value $\mu_{D_A} = 2.62$ represents the alternative hypothesis that beer prices are \$2.62 cheaper in the East End than in the West End.

- Under the alternative hypothesis ($H_A : \mu_W - \mu_E = \mu_{D_A}$), the sampling distributions of the test statistics $d$ and $z$ are

$$D_A \sim \mathcal{N}(\mu_{D_A}, \sigma_D) \qquad \Leftrightarrow \qquad Z_A \sim \mathcal{N}(0,1),$$

where $Z_A = \frac{D_A - \mu_{D_A}}{\sigma_D}$.

### 33.3.4 Understanding the sample size formula

The process of calculating the correct choice of sample size $n$ and and critical value $\mathrm{CV}_z$ requires solving the following two equations simultaneously:

$$\Pr\big(Z_0 > \mathrm{CV}_z \mid H_0 \text{ is true}\big) = \alpha, \qquad \Pr\big(Z_A \leqslant \mathrm{CV}_z \mid H_A \text{ is true}\big) \leqslant \beta,$$

where $Z_0 \equiv \frac{(\overline{X}_W - \overline{X}_E) - 0}{\sigma_D} \sim \mathcal{N}(0,1)$ is the distribution of the test statistic $z$ under the null hypothesis, and $Z_A \equiv \frac{(\overline{X}_W - \overline{X}_E) - \mu_{D_A}}{\sigma_D} \sim \mathcal{N}(0,1)$ is the distribution of the test statistic $d$ under the alternative hypothesis.

After some algebraic calculations, we can convert the two probability inequalities to two simple algebraic inequalities:

$$\mathrm{CV}_z = z_{1-\alpha}\sigma_D, \qquad \mathrm{CV}_z \geqslant \mu_{D_A} + z_\beta \sigma_D.$$

The first equations follows from the fact that we choose the cutoff value $\mathrm{CV}_z$ in order to satisfy the Type I error level. The second inequality tells us the minimum value of $\mathrm{CV}_z$ for the Type II errors to be as intended. After combining these expressions and looking for the equality condition, we obtain $z_\alpha \sigma_D = \mu_{D_A} + z_{1-\beta}\sigma_D$, which after some algebra steps leads us the formula $n = \frac{(z_\alpha - z_{1-\beta})^2(\sigma_W^2 + \sigma_W^2)}{\mu_{D_A}^2}$.

It might be helpful to see the above calculations with the particular values $z_{0.95} = 1.64$ and $z_{0.2} = -0.84$, for the values $\alpha = 0.05$ and $\beta = 0.2$ we have chosen. Recall that we also assumed the value $\mu_{D_A} = 2.62$ for the real difference between mean beer prices. Plugging these values into the combined equations for the value of $\mathrm{CV}_z$ gives us the rather simple equation

$$\mathrm{CV}_z = 1.64 \cdot \sigma_D = (2.62) - 0.84 \cdot \sigma_D.$$

Since we know $\sigma_D = \sqrt{\frac{10}{n}}$ we can solve for $n$ to find $n \approx 9$. You can visualize the two sides of this equation in Figure 33.6.

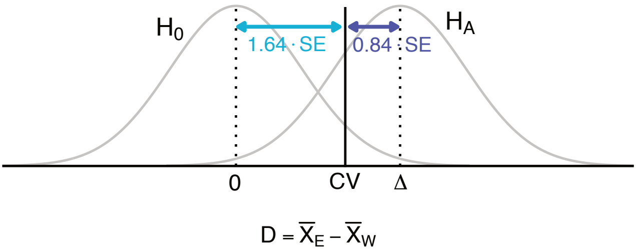

**Figure 33.6:** The dual constraint of the critical value $\text{CV}_z$. In order to have Type I errors at most $\alpha$ percent of the time we must choose $\text{CV}_z$ to be 1.64 standard deviations away from the centre of the distribution under $H_0$. In order to keep Type II errors small, we must choose $\text{CV}_z$ to be more than 0.84 standard deviations away from the mean of the sampling distribution under $H_A$.

## 33.4 Discussion

Our understanding of statistics and the NHST procedure is always evolving. Statisticians have been engaging in cordial scientific debate on the pros and cons of NHST (peppered with occasional creative name-calling like "bone-headedly misguided"). Like most statistical techniques, NHST has some valid criticisms which have led to recommendations for improvement. We've baked some of these suggestions into the NHST recipe described above: statistical power considerations, assumption checks, estimating effect sizes, and calculating confidence intervals.

### History of statistical testing

The NHST procedure we presented in this chapter has an interesting history, which will now examine briefly. The idea of using a *null hypothesis* was originally proposed by Ronald Fisher in his book *Statistical Methods for Research Workers*. Fisher starts from the assumption that scientists have observed some pattern in data, and possibly have some new theory that can explain the data. Before they can make any claims about their new theory, they must first show that the data observed cannot be explained by some baseline model, which he called the *null hypothesis*. The name *null hypothesis* comes from the fact that this is the hypothesis that stands to be nullified.

Fisher introduced $p$-values as the statistical tool to judge the *statistical significance* of a pattern in some observed data. Observing a lower $p$-value is stronger evidence for the pattern than observing a large $p$-value. For example, a small $p$-value like $p = 0.01$ means the observations are very unlikely to be due to chance (one in a hundred), versus a large $p$-value like $p = 0.33$ (one chance in three). Note that Fisher initially proposes the threshold value of $p = 0.05$ as a convenient rule of thumb, not as some universal standard of significance:

> "The value for which P=.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not."
>
> —Ronald Fisher, *Statistical Methods for Research Workers*

Somehow this recommendation for the significance level of 0.05 stuck, and became the "gold standard" for research findings that determines which results get published in many science journals. This was certainly not Fisher's intent, as he wrote that different types of research may require different standards of significance. For Fisher, the best thing to do for science authors is not to make any significance judgments but to report the $p$-values they obtained, and let readers decide for themselves whether the results are significant in the context.

Later, Jerzy Neyman and Egon Pearson introduced the notion of the *alternative hypothesis* and formulated statistical analysis as a binary decision rule used to reach one of two possible conclusions: "reject null hypothesis" or "fail to reject the null hypothesis." Instead of asking readers of scientific papers to judge the significance of the evidence provided for themselves, we can asks paper's authors to make that determination and report a hard conclusion, based on two tolerance-to-error parameters $\alpha$ and $\beta$. Note that for Neyman and Pearson the actual $p$-value is not important; it's just some number that we know will be smaller than predetermined value $\alpha$ to control Type I errors. The goal of introducing the hypothesis testing approach was to take away the arbitrary and situation-dependent notion of "significance" and replace it with a well-defined inferential procedure, governed by the tradeoff between the parameters $\alpha$ (tolerance for false positives) and $\beta$ (tolerance for false negatives).

Over time, scientific research "best practices" evolved to give us NHST which is a hybrid that mixes concepts from both Fisher's significance testing, and the Neyman–Pearson hypothesis testing methodologies. Instead of carefully choosing the $\alpha$ and $\beta$ values specific to each problem they want to study, many scientists blindly choose the $\alpha = 0.05$ rule of thumb to get a cutoff value, and only think about experiment power after having collected the data. For optimal statistical results, we recommend that you follow the six steps as described in this chapter: and perform power analysis before collecting your data samples. Additionally, we recommend that you always report effect size and confidence intervals for any estimated parameters, instead of just focussing on $p$-values.

## One-tailed and two-tailed tests

The definitions and worked example we presented in this chapter were using *upper-tailed* tests, which correspond to the case when the alternative hypothesis tests for a positive model parameter $\theta$. There are two other hypothesis formulation scenarios that you need to be aware of to complete your knowledge NHST, so we'll briefly describe them in this section.

In all the scenarios we'll use a critical value of the form $\text{CV}_z = z_q$ defined through the equation $F_Z(z_q) = q$, where $F$ is the CDF for the standard normal distribution $Z \sim \mathcal{N}(0,1)$. The critical values of the $z$-test are specified in terms of the normal distribution, and are of the form $\text{CV}_z = z_{1-\alpha}$, $\text{CV}_z = z_\alpha$, or $\text{CV}_z = \pm z_{1-\alpha/2}$, depending on comparison encoded in the hypotheses $H_0$ and $H_A$. In all cases the decision rule corresponds to a comparison of the value of a test statistic $z$ and the critical value, but type of comparison will according to one of the three possible cases:

- *upper-tailed*: When you want to test for the possibility of a positive parameter
$$H_0 : \theta \leqslant 0, \qquad\qquad H_A : \theta > 0.$$
Decision rule: Reject $H_0$ if $z > z_{1-\alpha}$, otherwise retain $H_0$.
- *lower-tailed*: When you want to test for the possibility of a negative parameter
$$H_0 : \theta \geqslant 0, \qquad\qquad H_A : \theta < 0.$$
Decision rule: Reject $H_0$ if $z < z_\alpha$, otherwise retain $H_0$.
- *two-tailed test*: When you want to test for the possibility of a parameter in two directions
$$H_0 : \theta = 0, \qquad\qquad H_A : \theta \neq 0.$$
Decision rule: Reject $H_0$ if $z < z_{\alpha/2}$ or if $z > z_{1-\alpha/2}$, otherwise retain $H_0$.

The *acceptance region* is defined by a different type of inequality in each of the above cases: it is $\{z \in \mathbb{R} \mid z \leqslant \text{CV}_z\}$ for upper tailed tests, $\{z \in \mathbb{R} \mid z \geqslant \text{CV}_z\}$ for lower-tailed tests, and $\{z \in \mathbb{R} \mid |z| \leqslant \text{CV}_z\}$ for two-tailed tests. The *critical region* is the complement of the region of acceptance for the statistical test. Recall that the critical region is the set of values for the test statistic that will lead us to reject $H_0$, as illustrated in Figure 33.7. The values of the test statistic that fall somewhere within the tails of the distribution tell us that $H_0$ a very unlikely explanation for the data observed.

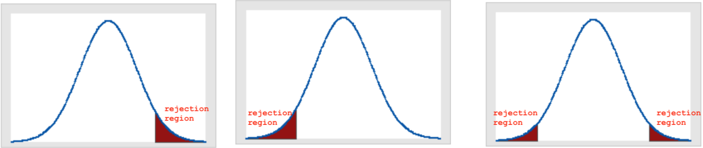

**Figure 33.7:** Illustration of the three types of "tails" that describe the probability weights of the false-positive conclusions. In each figure, the shaded region under the curve has total area $\alpha$ and corresponds to the Type I error for the test. When we observe a value of the test that falls within this range by change, we'll erroneously reject the null hypothesis even when $H_0$ is true.

The parts of the sampling distribution under the null hypothesis that fall in the *rejection region* of the statistical test correspond to the Type I error, as illustrated in Figure 33.7. You need to keep this picture in mind when choosing the critical value for a given test. For example, assuming we've chosen $\alpha = 0.05$, the critical value for an upper-tailed test will be $z_{1-\alpha} = z_{0.95} = 1.645$, while the critical value for an lower-tailed test is $z_\alpha = z_{0.05} = -1.645$. For a two-tailed test, the critical values will be at $z_{\alpha/2} = z_{0.025} = -1.96$ and $z_{1-\alpha/2} = z_{0.975} = 1.96$.

The type of test also affect the calculation of the $p$-value. Recall that the $p$-value is the probability of observing a value of the test statistic *at least as extreme* as the one you calculated from your sample purely by chance, assuming that $H_0$ is true. For an upper-tailed $z$-test, the $p$-value is given by $p = \Pr(Z \geqslant z \mid H_0)$. For a lower-tailed $z$ test the $p$-values is calculated using $\Pr(Z \leqslant z \mid H_0)$, and for two-tailed tests the $p$-values is $\Pr(|Z| \geqslant z \mid H_0)$.

## On being a self-respecting scientist

You need to know about NHST if you want to publish science papers or make better business decisions. It's possible to follow the NHST procedure by blindly carrying out rote calculations without intuition or understanding. This approach often leads to misleading interpretations, wrong statements, and false scientific conclusions. If you choose this approach, you'll have to memorize all kinds of formulas and "rules" to follow on a case-by-case basis. Dear readers, you'll have to trust me on this one, you can do better than that. Having proven yourself by surviving all the prerequisites topics we covered in the data and probability chapters, you have no excuses for skipping the under-the-hood story for the NHST procedure, which is one of the core pillars of the modern scientific establishment. It would be a shame if you were to miss out on the beautiful mathematics construct that builds on top of thousands of years of analytical thought. Think about it, what other math topics do you know that requires using Greek symbols like $\alpha$ and $\beta$, Roman numerals for the type errors, the best of seventeenth century math theory, and modern computer-aided calculations.

In particular, you need to know probability theory to correctly interpret $p$-values. The $p$-value is the probability of observing a value of the test statistic equally or more extreme than the value of the test statistic obtained form the data. Without this under-the-hood understanding of probabilistic modelling assumptions, it's easy to make a wrong decision in Step 5 and give a incorrect interpretation of the $p$-value in Step 6. A significant proportion of scientists currently publishing papers, including scientists with prestigious awards and distinctions, commit statistical errors when using the NHST procedure mechanically without thinking much about the probability assumptions they're making, which means some significant portion (think 20%+) of published papers are wrong. I'm counting on you to do better than that in your research.

In the next chapter we'll learn about the various other statistical tests that can be used with the NHST procedure. The formulas and calculation procedures will differ in each case, but we'll always follow the six-step NHST procedure as described in this chapter. If you ever start to get confused about what is going on in later chapters, you can always come back to this chapter to review and re-acquaint yourself with the meaning of $\alpha$, $\beta$, critical values, $p$-values, effect sizes, and confidence intervals.

In this chapter we looked in detail at one example based on the $z$-test, but the $z$-test applies for many other scenarios (see Exercises). Indeed $z$-test is the go-to tool whenever we perform statistical analysis on a populations with known variance. But what about cases where variance is not known? And what about other scenarios when we're not comparing means but proportions, or two sided tests, or any of the myriad other data analysis scenarios that can come up. In the next chapter we'll discuss a number of other tests that can be used as part of the NHST procedure. The steps of the NHST procedure are the same as in this chapter, but have to use different formulas depending on the assumptions and the probability models for the sampling distributions that arise in each case.

## 33.5 Exercises

You should be able to solve all the exercises in this chapter without the need for a computer, by looking up probability values in Table ZZ. If you do have access to a computer, you can skip the lookup table and instead compute $z_q \equiv F_Z^{-1}(q)$ using the formula $z_q$ `=NORM.INV(q,0,1)` in Excel, by calling $z_q$ `=qnorm(q,0,1)` in R, or $z_q$ `=norm.ppf(q,0,1)` in Python.

**E33.1** Repeat the numerical analysis techniques for the beer example using R. ..

**E33.2** Suppose you receive two new samples for beer prices for which $\overline{x}_W = 10$ and $\overline{x}_E = 9.3$. Assuming $n = 9$ and we continue with the assumption of a known population variance $\sigma_E^2 = \sigma_W^2 = 5$, what conclusion will you reach? Is Kayla's claim that beer prices are cheaper in the East End supported by the second sample? Fail to reject $H_0$.

**E33.3** Another question but this time ask to do a two-tailed $z$-test. ??

## 33.6 Links

[ An interactive visualization about statistical power and significance testing]
`https://rpsychologist.com/d3/NHST`

[ Historical context about the foundations of statistics ]
`https://en.wikipedia.org/wiki/Foundations_of_statistics`

[ The Wikipedia bios of the inventors of NHST ]
`https://en.wikipedia.org/wiki/Ronald_Fisher`
`https://en.wikipedia.org/wiki/Jerzy_Neyman`
`https://en.wikipedia.org/wiki/Egon_Pearson`