# 10

# Hypothesis Testing

### Overview

Hypothesis testing is a fundamental tool in data science. It enables us to determine whether the available data provide sufficient evidence to support a certain hypothesis, which must be selected beforehand, as explained in Section 10.1. Section 10.2 describes the main idea underlying the hypothesis-testing framework: we play devil's advocate and try to interpret the data assuming our hypothesis does not hold (an assumption known as the null hypothesis). In Sections 10.3 and 10.4, we explain how to use parametric modeling to perform hypothesis testing and introduce a key concept: the p value. A small p value indicates that the data cannot be explained by the null hypothesis, which is evidence in favor of the original hypothesis of interest. Section 10.5 shows that this procedure is guaranteed to control the probability of endorsing a false finding. Section 10.6 explains how to compute the power of a test, which is the probability of correctly rejecting the null hypothesis when it does not hold. In Section 10.7 we explain how to perform hypothesis testing without parametric models. Section 10.8 describes multiple testing, a challenging setting of great practical interest where many tests are performed simultaneously. In Section 10.9 we discuss the interplay between hypothesis testing and causal inference. Finally, Section 10.10 warns against relying solely on p values to establish the importance of a scientific result.

## 10.1 Selecting A Hypothesis

The goal of hypothesis testing is to evaluate whether a hypothesis is supported by data. This is achieved by trying to interpret the data assuming that the hypothesis does not hold. If this interpretation is implausible, then we conclude that the data supports our hypothesis. For this logic to be sound, it is crucial to choose the hypothesis *beforehand*. It may be tempting to use the same data to select the hypothesis and also back it up, but this is a serious mistake! To see why, imagine that you are strolling down 5th avenue in New York with a friend and they tell you:

*Look, a car with license plate number EMC6055! About one million cars drive through Manhattan every day. Isn't it amazing that we saw this particular one?*

*It's a one-in-a-million chance!*

No, it is not amazing. The probability of seeing that particular number is indeed one in a million. However, your friend did not specify the number beforehand, so they could have said the same about any other number! Consequently, their hypothesis is not really *seeing the license plate number EMC6055*, but rather *seeing a car with any license plate number*, which is much less interesting. If they had told you the number *before seeing the car*, then the data would indeed support their claimed hypothesis, and you would be very impressed. Similarly, when we analyze data, there are often many possible hypotheses to choose from. If we select one that is consistent with the data, then we will obviously find that it is supported by the data (that is exactly why we chose it!). In order to avoid this circular logic, it is crucial to *first* define a hypothesis and *then* look at the data.

**Example 10.1** (Unfair die: Hypothesis)**.** In Example 1.23 we study a toy die that belongs to my daughter. Rolling it 60 times yielded substantially more threes (18) than any other number. This suggests the following hypothesis:

*The probability of rolling a three is greater than 1/6.*

We might be tempted to use the data in Example 1.23 as evidence to support this hypothesis, but we should not because we have already used it to come up with the hypothesis. We need fresh data, which I gathered by rolling the toy die 100 more times. We will see what happened in the following sections.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

**Example 10.2** (Free throws under pressure)**.** Giannis Antetokounmpo is an NBA superstar who plays for the Milwaukee Bucks. One of the few weaknesses in his game is free-throw shooting. Before each free throw, Antetokounmpo performs a long routine, which often exceeds the official 10-second limit to shoot. In the 2020/2021 season, fans from opposing teams started counting loudly during the routine in order to request a 10-second violation from the referees. While watching the 2021 NBA playoffs I wondered whether this affected Antetokounmpo's free-throw percentage; it seemed to me that he was shooting worse at away games than at home games, when the fans were mostly silent during his free throws. My hypothesis was:

*Antetokounmpo's free-throw percentage is better at home than away.*

I could not test this hypothesis based on the games I had already watched, since they had influenced my choice of hypothesis. Instead, I decided to use the games from the NBA finals, which had not yet occurred. We analyze these data in the following sections.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

## 10.2 The Null Hypothesis And The Test Statistic

When we perform hypothesis testing, we take a skeptical approach and try to explain the data under the assumption that our hypothesis of interest does not hold. This assumption is known as the *null hypothesis*. In contrast, the hypothesis of interest is called the *alternative hypothesis*. If the data are inconsistent with the null hypothesis, then this is interpreted as evidence supporting the alternative hypothesis.

**Example 10.3** (Unfair die: Null and alternative hypotheses)**.** In Example 10.1 we conjecture that the die is not fair. Therefore, a natural choice for the null hypothesis is:

*The probability of rolling a three is equal to 1/6.*

The corresponding alternative hypothesis is that the probability of rolling a three is different from 1/6.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

**Example 10.4** (Free throws under pressure: Null and alternative hypotheses)**.** In Example 10.2, we suspect that Antetokounmpo does not shoot free throws with equal accuracy at home and away games. Therefore a reasonable null hypothesis is:

*Antetokounmpo's free throw percentage is the same at home and away.*

The corresponding alternative hypothesis is that the percentages are different.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

In order to evaluate to what extent the data are consistent with the null hypothesis, we define a *test statistic*. This is a function of the data designed to be larger under the alternative hypothesis than under the null hypothesis. If the test statistic is very large, then this suggests that the null hypothesis is unlikely to hold.

**Example 10.5** (Unfair die: Test statistic)**.** To test the hypothesis in Example 10.1, we need a test statistic $t_{\text{data}}$ that is small when the probability of rolling a three is 1/6, and large when it is not. Since we suspect that the probability of rolling a three is actually greater than 1/6, we set the test statistic to be the number of rolled threes, which should be larger if our suspicion is correct. The value of this test statistic computed from the data (100 additional rolls of the die) is $t_{\text{data}} = 21$ (21 of the additional rolls were threes).

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

**Example 10.6** (Free throws under pressure: Test statistic)**.** As explained in Example 10.2, we would like to test whether Antetokounmpo's free throw percentage is the same at home and away. Since we expect him to shoot worse in away games, we build a test statistic $t_{\text{data}}$ by subtracting the fraction of made free throws at away games from the fraction of made free throws at home games,

$$t_{\text{data}} := \frac{\text{Made free throws at home}}{\text{Attempted free throws at home}} - \frac{\text{Made free throws away}}{\text{Attempted free throws away}}. \tag{10.1}$$

During the finals, Antetokounmpo made 34 out of 44 free throws in home games, and 22 out of 41 free throws in away games, so $t_{\text{data}} = 0.236$.

## 10.3 Parametric Testing And The P Value

As explained in Section 10.2, our goal in hypothesis testing is to determine whether the data can be explained in terms of the null hypothesis. More specifically, we compute a test statistic from the data, which is designed to be larger under the alternative hypothesis than under the null hypothesis. If the test statistic is very large, then this is evidence against the null hypothesis. A key challenge is how to quantify precisely what *very large* means. For this, we need to characterize the probabilistic behavior of the test statistic under the null hypothesis. In this section, we describe how to do this using parametric models (introduced in Sections 2.3 and 3.6). Section 10.7 describes an alternative nonparametric approach.

In parametric hypothesis testing, we interpret the test statistic as a random variable with a predefined distribution that depends on a small number of deterministic parameters. These parameters enable us to precisely define the null and alternative hypotheses. A *simple* null hypothesis states that each model parameter has a single fixed value. A *composite* null hypothesis states that each parameter is in a predefined set.

**Example 10.7** (Unfair die: Parametric model)**.** In Example 10.5 we choose the number of rolled threes as the test statistic for our die example. To derive a parametric model for the test statistic, we assume that the rolls are independent and that the probability of rolling a three is equal to a fixed parameter $\theta$. In that case, the distribution of the test statistic is binomial with parameters $n := 100$ (the number of rolls) and $\theta$ (see Section 2.3.2). We denote by $\tilde{t}_\theta$ the random variable representing the test statistic in order to emphasize its dependence on $\theta$. The pmf of $\tilde{t}_\theta$ is

$$p_{\tilde{t}_\theta}(t) = \binom{100}{t} \theta^t (1-\theta)^{(100-t)}, \quad t = 0, 1, \ldots, 100. \tag{10.2}$$

The null hypothesis defined in Example 10.3 is a simple null hypothesis, which states that $\theta = 1/6$. Alternatively, we could have chosen the composite null hypothesis $\theta \in [0, 1/6]$, stating that the probability of rolling a three is less than or equal to $1/6$.

Armed with a parametric model for the test statistic, we can finally evaluate the consistency of the data and the null hypothesis. To achieve this, we compute the probability that the test statistic is greater than or equal to the observed value, assuming that the null hypothesis holds. In our unfair die example, the observed test statistic equals 21 rolled threes. We therefore compute the probability of 21 or more rolled threes occurring when $\theta = \theta_{\text{null}}$. If this probability, which we call the p value, is very low, then this is convincing evidence against the null hypothesis.

**Definition 10.8** (P-value function and p value)**.** *The p value is the probability that the test statistic under the null hypothesis is greater than or equal to the test statistic computed from the available data. We define the p value in terms of the p-value function, which is a function mapping every possible value of the test statistic to the corresponding p value.*

*Let $\tilde{t}_\theta$ be a random variable representing a test statistic with a parametric distribution that depends on a parameter (or vector of parameters) $\theta$. For any $\theta_{\text{null}}$, consider the simple null hypothesis $\theta = \theta_{\text{null}}$. The p-value function* pv *associated to this null hypothesis maps $t$ to the probability that $\tilde{t}_{\theta_{\text{null}}}$ is greater than or equal to $t$,*

$$\text{pv}(t) := \mathrm{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq t\right). \tag{10.3}$$

*Consider a composite null hypothesis of the form $\theta \in \Theta_{\text{null}}$ for a certain predefined set $\Theta_{\text{null}}$. The corresponding p-value function* pv *maps each $t$ to the supremum over $\Theta_{\text{null}}$ of the probability that $\tilde{t}_{\theta_{\text{null}}}$ is greater than or equal to $t$,*

$$\text{pv}(t) := \sup_{\theta \in \Theta_{\text{null}}} \mathrm{P}\left(\tilde{t}_{\theta} \geq t\right). \tag{10.4}$$

*The p value of the available data is equal to* $\text{pv}(t_{\text{data}})$*, where $t_{\text{data}}$ denotes the observed test statistic.*

**Example 10.9** (Unfair die: P value)**.** Following Definition 10.8, we compute the p-value function corresponding to the simple null hypothesis $\theta = \theta_{\text{null}} := 1/6$ using the pmf of $\tilde{t}_{\theta_{\text{null}}}$ derived in Example 10.7,

$$\begin{aligned}\text{pv}(t) &:= \mathrm{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq t\right) && (10.5)\\ &= \sum_{i=t}^{100} \binom{100}{i} \left(\frac{1}{6}\right)^{i} \left(\frac{5}{6}\right)^{100-i}, \quad t = 0, 1, \ldots, 100. && (10.6)\end{aligned}$$

The p-value function under the composite null hypothesis $\theta \in \Theta_{\text{null}} := [0, 1/6]$ is the same as under the simple null hypothesis, because the supremum is attained

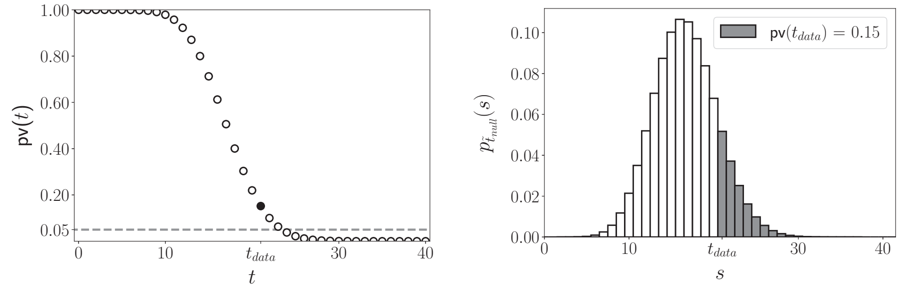


**Figure 10.1 P value for a (possibly) unfair die.** The left plot shows the p-value function pv derived in Example 10.9. The p value corresponding to the observed test statistic $t_{\text{data}} := 21$ is equal to 0.15. This is above 0.05, a popular threshold for statistical significance (see Section 10.5) represented by the dashed gray line. The right plot depicts the pmf of the test statistic (the number of rolled threes) under the null hypothesis ($\theta = \theta_{\text{null}} = 1/6$). The p value at $t_{\text{data}}$ is the sum of the pmf over all values greater than or equal to $t_{\text{data}}$.

exactly at $\theta := 1/6$:

$$\mathrm{pv}(t) := \sup_{\theta \in \Theta_{\text{null}}} \mathrm{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq t\right) \tag{10.7}$$

$$= \sup_{\theta \in [0,1/6]} \sum_{i=t}^{100} \binom{100}{i} \theta^{i} \left(1-\theta\right)^{100-i} \tag{10.8}$$

$$= \sum_{i=t}^{100} \binom{100}{i} \left(\frac{1}{6}\right)^{i} \left(\frac{5}{6}\right)^{100-i}, \quad t = 0, 1, \ldots, 100. \tag{10.9}$$

Figure 10.1 shows the p-value function. It is close to one for small values of the test statistic, because we are almost certain to observe a larger test statistic under the null hypothesis. Then it decreases until it reaches the p value for the observed test statistic, which equals

$$\mathrm{pv}(t_{\text{data}}) = \sum_{i=21}^{100} \binom{100}{i} \left(\frac{1}{6}\right)^{i} \left(\frac{5}{6}\right)^{100-i} = 0.15. \tag{10.10}$$

A common misconception is that the p value represents the probability that the null hypothesis holds. In Example 10.9, this would translate to the statement:

*The probability that the die is fair (or equivalently that $\theta = 1/6$) is* 0.15.

However, this is incorrect, because we are assuming that the model parameters are deterministic. Under this frequentist perspective, the null hypothesis is not random: it either holds or not. In the case of the die, $\theta$ is either equal to $1/6$ or it is not. In order to make probabilistic statements about the null hypothesis, we would need to take a Bayesian perspective (see Section 6.7). However, as discussed in Example 9.47, these statements would then depend on the prior distribution assigned to the parameters.

From our frequentist viewpoint, the correct statement associated to the p value computed in Example 10.9 is:

*If the die is fair ($\theta = 1/6$), the probability of observing 21 threes or more is* 0.15*.*

If this probability is very small, then we interpret it as evidence against the null hypothesis. To determine whether the p value is small enough to be convincing, we compare it to a predetermined threshold, which is often set equal to 0.05 (but can vary depending on the domain). Section 10.5 shows that this allows us to control the probability of endorsing a false finding. The p value in Example 10.9 is not very small, so the observed data are not very unlikely given the null hypothesis. If the die is indeed fair and we repeat the experiment over and over, then we will observe 21 threes or more about 15% of the time. Note that this does *not* mean that we think the null hypothesis is correct; it just means that we do not have enough evidence to rule it out. In fact, I personally still think that the die is unfair.

## 10.4 Two-Sample Tests

In Example 10.9 the data are modeled as samples from a single distribution. The corresponding test is consequently called a *one-sample* test. In contrast, in Example 10.4 the data are divided into two groups, corresponding to home and away games. The alternative hypothesis is that the groups have different distributions, whereas the null hypothesis is that they are sampled from the same distribution. This is known as a *two-sample* test. The following definition describes a parametric two-sample test called a z test, based on the Gaussian approximation to the sample mean provided by the central limit theorem (Definition 9.38).

**Definition 10.10** (Two-sample z test)**.** *Given two binary datasets of sizes $n_A$ and $n_B$, the null hypothesis of the two-sample z test is that all the data are generated as i.i.d. samples from the same Bernoulli distribution. If the test statistic is the difference between the sample proportions*

$$t_{\text{data}} := \frac{k_A}{n_A} - \frac{k_B}{n_B}, \tag{10.11}$$

*where $k_A$ and $k_B$ are the number of instances equal to one in the first and second set, respectively, then the test is said to be one-tailed. The corresponding p value equals*

$$\text{pv}(t_{\text{data}}) = 1 - F_{\tilde{z}}\left(\frac{t_{\text{data}}}{\sigma_{\text{null}}}\right). \tag{10.12}$$

*$F_{\tilde{z}}$ is the cdf of a standard Gaussian random variable with mean zero and unit variance, and*

$$\sigma_{\text{null}}^2 := \frac{k(n-k)}{n^2}\left(\frac{1}{n_A}+\frac{1}{n_B}\right), \tag{10.13}$$

*where $n := n_A + n_B$ and $k := k_A + k_B$. If the test statistic is the absolute value of the difference between the sample proportions*

$$t_{\text{data}} := \left|\frac{k_A}{n_A} - \frac{k_B}{n_B}\right|, \tag{10.14}$$

*then the test is said to be two-tailed and the p value equals*

$$\text{pv}(t_{\text{data}}) = 2\left(1 - F_{\tilde{z}}\left(\frac{t_{\text{data}}}{\sigma_{\text{null}}}\right)\right). \tag{10.15}$$

*Derivation* Let the random variables $\tilde{x}_1, \ldots, \tilde{x}_n$ represent the data. We denote by $\mathcal{A}$ and $\mathcal{B}$ the indices of the data points in the first and second group respectively. For the one-tailed test, the test statistic equals

$$\tilde{t}_{\text{1-tail}} = \frac{1}{n_A}\sum_{i\in\mathcal{A}}\tilde{x}_i - \frac{1}{n_B}\sum_{i\in\mathcal{B}}\tilde{x}_i. \tag{10.16}$$

Under the null hypothesis, the random variables $\tilde{x}_i$, $1 \leq i \leq n$, are i.i.d. Bernoulli with the same parameter, which we denote by $\theta_{\text{null}}$. Therefore, $\sum_{i\in\mathcal{A}}\tilde{x}_i$ is binomial with parameters $\theta_{\text{null}}$ and $n_A$, and $\sum_{i\in\mathcal{B}}\tilde{x}_i$ is binomial with parameters $\theta_{\text{null}}$ and $n_B$. By the Gaussian approximation to the binomial distribution (Definition 9.39), $\sum_{i\in\mathcal{A}}\tilde{x}_i$ is approximately Gaussian with mean $n_A\theta_{\text{null}}$ and variance $n_A\theta_{\text{null}}(1-\theta_{\text{null}})$. Equivalently, by Theorem 3.32, $\frac{1}{n_A}\sum_{i\in\mathcal{A}}\tilde{x}_i$ is approximately Gaussian with mean $\theta_{\text{null}}$ and variance $\theta_{\text{null}}(1-\theta_{\text{null}})/n_A$. Similarly, $\frac{1}{n_B}\sum_{i\in\mathcal{B}}\tilde{x}_i$ is approximately Gaussian with mean $\theta_{\text{null}}$ and variance $\theta_{\text{null}}(1-\theta_{\text{null}})/n_B$. By Theorem 9.36 this implies that $\tilde{t}_{\text{1-tail}}$ is approximately Gaussian with mean zero and variance

$$\sigma_{\text{null}}^2 := \theta_{\text{null}}(1-\theta_{\text{null}})\left(\frac{1}{n_A}+\frac{1}{n_B}\right) \tag{10.17}$$

$$\approx \frac{k(n-k)}{n^2}\left(\frac{1}{n_A}+\frac{1}{n_B}\right), \tag{10.18}$$

where we approximate $\theta_{\text{null}}$ as the fraction of total positive instances $k/n$. This is the maximum-likelihood estimate of the parameter under the null hypothesis that all data points are samples from i.i.d. Bernoulli random variables (see Example 2.26). The p-value function equals

$$\text{pv}(t) := \text{P}\left(\tilde{t}_{\text{1-tail}} \geq t\right) \tag{10.19}$$

$$= \text{P}\left(\tilde{z} \geq \frac{t}{\sigma_{\text{null}}}\right), \tag{10.20}$$

where we use the fact that $\tilde{z} := \tilde{t}_{\text{1-tail}}/\sigma_{\text{null}}$ is a standard Gaussian random variable with mean zero and unit variance by Theorem 3.32.

The test statistic for the two-tailed test is equal to the absolute value of the test statistic for the one-tailed test,

$$\tilde{t}_{\text{2-tails}} := \left|\tilde{t}_{\text{1-tail}}\right|, \tag{10.21}$$

so the corresponding p-value function equals

$$\begin{aligned} \text{pv}(t) &:= \text{P}\left(\tilde{t}_{\text{2-tails}} \geq t\right) && (10.22)\\ &= \text{P}\left(\tilde{t}_{\text{1-tail}} \geq t\right) + \text{P}\left(\tilde{t}_{\text{1-tail}} \leq -t\right) && (10.23)\\ &= \text{P}\left(\tilde{z} \geq \frac{t}{\sigma_{\text{null}}}\right) + \text{P}\left(\tilde{z} \leq -\frac{t}{\sigma_{\text{null}}}\right) = 2\text{P}\left(\tilde{z} \geq \frac{t}{\sigma_{\text{null}}}\right), && (10.24) \end{aligned}$$

by symmetry of the zero-mean Gaussian pdf. ■

The p value for the *one-tailed* test in Definition 10.10 is computed only from the right tail (i.e. the extreme positive values) of the distribution of the difference between the sample proportions under the null hypothesis. Therefore, the test does not reject the null hypothesis if the difference is negative, even if it is very extreme. In contrast, the *two-tailed* test takes into account both tails of the distribution. The bottom two plots in Figure 10.2 illustrate the difference between the two tests.

**Example 10.11** (Free throws under pressure: P value)**.** We apply the two-sample z test in Definition 10.10 to obtain a p value for the test statistic in Example 10.6, assuming that the free throws are independent. We begin by applying a one-tailed test, which is only sensitive to positive differences and oblivious to situations where the away percentage is higher. Setting $n_A := 44$, $k_A := 34$, $n_B := 41$, $k_B := 22$, we obtain $\sigma_{\text{null}} = 0.103$, which yields the p-value function

$$\text{pv}(t) = 1 - F_{\tilde{z}}\left(\frac{t}{0.103}\right), \tag{10.25}$$

depicted in the top left plot of Figure 10.2. The bottom left plot shows the Gaussian approximation to the difference between home and away percentages. The p value for the observed test statistic $t_{\text{data}} = 0.236$ is equal to 0.011. The data therefore seem to contradict the null hypothesis, in the sense that if it holds (and our independence assumptions are correct) then we would observe results like this only around 1% of the time.

The one-tailed test derived in Example 10.11 ignores the possibility that Antetokounmpo might actually shoot better at away games (perhaps motivated by the fans taunting him). If we instead apply the two-tailed test in Definition 10.10 then the p-value function equals

$$\text{pv}(t) = 2\left(1 - F_{\tilde{z}}\left(\frac{t}{0.103}\right)\right), \tag{10.26}$$

as depicted in the top right plot of Figure 10.2. The bottom right plot shows that the p value for the two-tailed test is obtained by integrating both tails of the Gaussian pdf approximating the difference between home and away percentages. The p value for the two-tailed test is 0.022, double that of the one-tailed test,

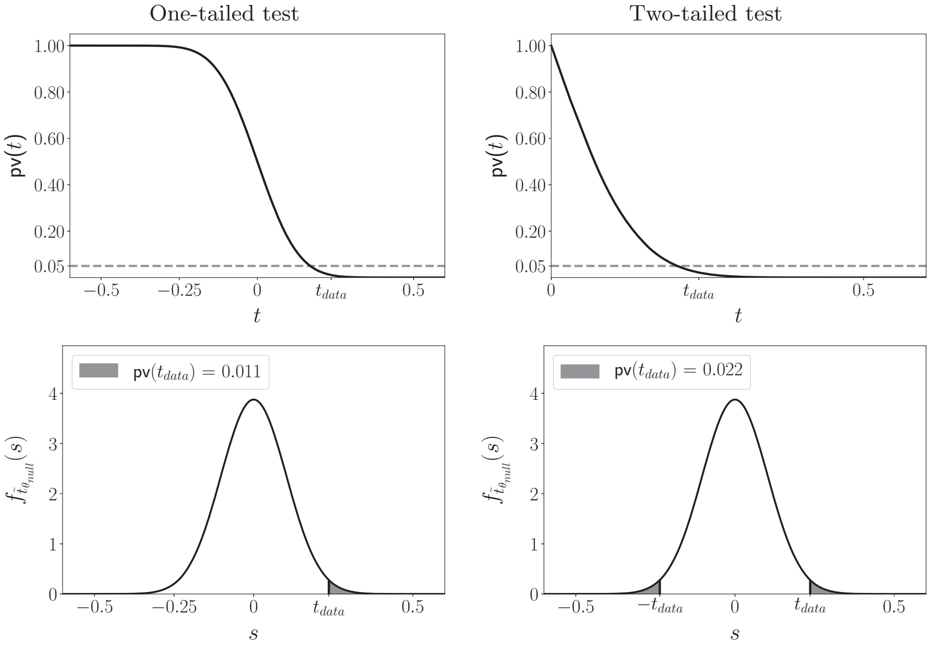


**Figure 10.2 P value for free throws under pressure.** The top row shows the p-value functions for the one-tailed (left column) and two-tailed (right column) tests in Example 10.11. The p value corresponding to the observed test statistic is equal to 0.011 and 0.022, respectively. This is below 0.05, a popular threshold for statistical significance (see Section 10.5) represented by the dashed gray line. The bottom row depicts the Gaussian approximation to the difference between home and away percentages derived in Definition 10.10. The p value for the one-tailed test is obtained by integrating the right tail of the pdf (left). The p value for the two-tailed test is obtained by integrating both tails of the pdf (right).

but still very small. Example 10.17 compares the properties of the one-tailed and two-tailed tests in more detail.

## 10.5 Statistical Significance

As described in Section 10.3, the p value quantifies the consistency between the null hypothesis and the available data. In the scientific literature, p values are typically compared to a predetermined threshold called the *significance level*. If the p value is below the threshold, then we *reject* the null hypothesis, meaning that it is inconsistent with the data, and declare the result to be *statistically significant*. If the p value is above the threshold, then we do not reject the null hypothesis; we conclude that the evidence against it is insufficient. The following definition summarizes the different steps in a hypothesis test.

**Definition 10.12** (Hypothesis test)**.** *To perform a hypothesis test we:*

*1 Choose a conjecture.*
*2 Determine the corresponding null hypothesis.*
*3 Design a test statistic and model its distribution under the null hypothesis.*
*4 Decide on a significance level $\alpha$.*
*5 Gather the data and compute the test statistic.*
*6 Compute the p value.*
*7 Reject the null hypothesis if the p value is less than or equal to $\alpha$.*

A decision based on a hypothesis test can be wrong in two different ways. A false positive or *type 1 error* occurs if the null hypothesis holds, but we reject it. We call this a false positive because rejecting the null hypothesis supports our original conjecture. Conversely, a false negative or *type 2 error* occurs if we fail to reject the null hypothesis when it does not hold. In hypothesis testing the main priority is to avoid the first type of error. In fact, the procedure described in Definition 10.12 is designed to control the probability of a false positive. To understand why, we need to understand the behavior of the test statistic under the null hypothesis.

Let us assume that the null hypothesis is simple and states that $\theta = \theta_{\text{null}}$ for a certain constant $\theta_{\text{null}}$. We denote by $\tilde{t}_{\theta_{\text{null}}}$ the random variable that represents the test statistic under this null hypothesis. A false positive happens if (1) the null hypothesis holds and (2) the p value is below the significance level $\alpha$. The probability of this event is

$$\mathrm{P}\left(\text{False positive}\right) = \mathrm{P}\left(\mathrm{pv}(\tilde{t}_{\theta_{\text{null}}}) \leq \alpha\right). \tag{10.27}$$

To make sense of this expression, recall that the p-value function is a deterministic function, which maps any possible value of the test statistic to the corresponding p value. Therefore, plugging the random variable $\tilde{t}_{\theta_{\text{null}}}$ into it yields another random variable, which we call

$$\tilde{u} := \mathrm{pv}(\tilde{t}_{\theta_{\text{null}}}). \tag{10.28}$$

A false positive occurs when $\tilde{u} \leq \alpha$, so $\mathrm{P}\left(\text{False positive}\right)$ is equal to the cdf of $\tilde{u}$ at $\alpha$. Figure 10.3 shows the cdf $F_{\tilde{u}}$ of $\tilde{u}$ for the p-value functions derived in Examples 10.9 and 10.11. Notice that in both cases $F_{\tilde{u}}\left(\alpha\right) \leq \alpha$ for all $\alpha$. This is no coincidence. Although it is not obvious at first sight, the p-value function is defined precisely so that

$$\mathrm{P}\left(\text{False positive}\right) = F_{\tilde{u}}\left(\alpha\right) \leq \alpha, \tag{10.29}$$

which means that the probability of a false positive is indeed bounded by the significance level! We dedicate the rest of the section to proving this rigorously.

The following theorem characterizes the distribution of the p value under the null hypothesis, showing that Figure 10.3 is not a fluke: the cdf of any p value looks like that under the null hypothesis.

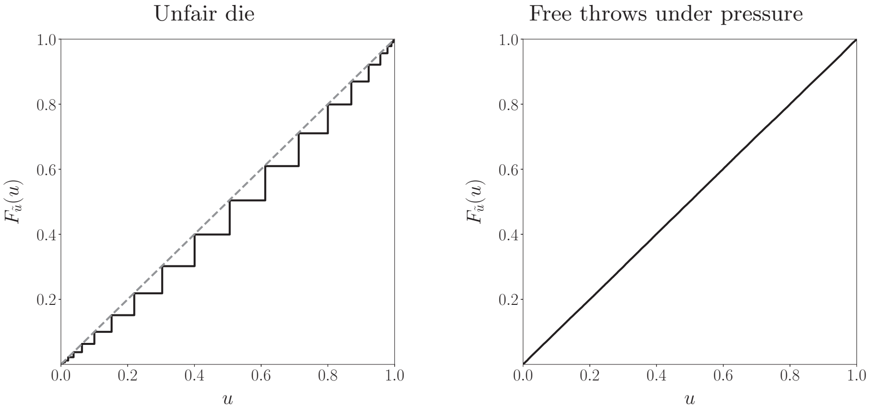


**Figure 10.3 P value under the null hypothesis.** The plots show the cdf of the random variable $\tilde{u} := \text{pv}(\tilde{t}_{\theta_{\text{null}}})$ for the p-value functions in Examples 10.9 (left) and 10.11 (right). The cdf on the left lies below the diagonal line where $F_{\tilde{u}}(u) = u$. The cdf on the right is on top of the line, indicating that the p value is uniformly distributed under the null hypothesis.

**Theorem 10.13** (Distribution of the p value under the null hypothesis)**.** *Let $\tilde{t}_\theta$ be a random variable representing a test statistic with a parametric distribution that depends on a parameter (or vector of parameters) $\theta$. For any $\theta_{\text{null}}$, let* pv *be the p-value function associated to the simple null hypothesis $\theta = \theta_{\text{null}}$. We define the random variable $\tilde{u} := \text{pv}(\tilde{t}_{\theta_{\text{null}}})$ to represent the p value under the null hypothesis. If $\tilde{t}_{\theta_{\text{null}}}$ is continuous, $\tilde{u}$ is uniformly distributed in $[0, 1]$. If $\tilde{t}_{\theta_{\text{null}}}$ is discrete, for any $u \in [0, 1]$*

$$F_{\tilde{u}}(u) \leq u, \tag{10.30}$$

*with equality if there is a $t$ such that $\text{pv}(t) = u$ and $\text{P}\left(\tilde{t}_{\theta_{\text{null}}} = t\right) \neq 0$.*

*Proof* If $\tilde{t}_{\theta_{\text{null}}}$ is continuous, the distribution of $F_{\tilde{t}_{\theta_{\text{null}}}}(\tilde{t}_{\theta_{\text{null}}})$ is uniformly distributed in $[0, 1]$ by Theorem 3.23. By definition, the p-value function equals

$$\text{pv}(t) := \text{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq t\right) = 1 - F_{\tilde{t}_{\theta_{\text{null}}}}(t), \tag{10.31}$$

so $1 - \tilde{u} = 1 - \text{pv}(\tilde{t}_{\theta_{\text{null}}}) = F_{\tilde{t}_{\theta_{\text{null}}}}(\tilde{t}_{\theta_{\text{null}}})$ is uniformly distributed in $[0, 1]$, and therefore so is $\tilde{u}$ (see Exercise 3.8). This is exactly what we see on the right graph of Figure 10.3.

If $\tilde{t}_{\theta_{\text{null}}}$ is discrete, then so is $\tilde{u}$. Let $t_1 < t_2 < \ldots$ denote the (countable) points for which $\text{P}\left(\tilde{t}_{\theta_{\text{null}}} = t_i\right) \neq 0$, $i = 1, 2, \ldots$ The corresponding values $u_i := \text{pv}(t_i)$ that $\tilde{u}$ takes with nonzero probability are decreasing: $u_1 > u_2 > \ldots$, because for

$t_i < t_j$

$$\begin{aligned} u_i := \text{pv}(t_i) &= \text{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq t_i\right) && (10.32)\\ &= \text{P}\left(\tilde{t}_{\theta_{\text{null}}} = t_i\right) + \text{P}\left(t_i < \tilde{t}_{\theta_{\text{null}}} < t_j\right) + \text{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq t_j\right) && (10.33)\\ &> \text{pv}(t_j) := u_j. && (10.34) \end{aligned}$$

Consequently, for any $i = 1, 2, ...$ the events $\tilde{u} \leq u_i$ and $\tilde{t}_{\theta_{\text{null}}} \geq t_i$ are equivalent, which implies

$$\begin{aligned} F_{\tilde{u}}(u_i) = \text{P}\left(\tilde{u} \leq u_i\right) &= \text{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq t_i\right) && (10.35)\\ &= \text{pv}(t_i) = u_i. && (10.36) \end{aligned}$$

Finally, for $u_i < u < u_{i+1}$, $\text{P}\left(u_i < \tilde{u} \leq u\right) = 0$, so

$$\begin{aligned} F_{\tilde{u}}(u) &= \text{P}\left(\tilde{u} \leq u\right) && (10.37)\\ &= \text{P}\left(\tilde{u} \leq u_i\right) && (10.38)\\ &= F_{\tilde{u}}(u_i) && (10.39)\\ &= u_i \leq u. && (10.40) \end{aligned}$$

The cdf therefore looks like the one depicted on the left of Figure 10.3. ■

For simple null hypotheses, Theorem 10.13 immediately implies that the significance level bounds the probability of a false positive. The following theorem provides a proof, and establishes that the same is true for composite null hypotheses.

**Theorem 10.14** (The significance level bounds the probability of a false positive)**.** *Let $\tilde{t}_\theta$ be a random variable representing a test statistic with a parametric distribution that depends on a parameter (or vector of parameters) $\theta$. For any $0 < \alpha < 1$, if the null hypothesis is rejected when the p value is smaller than or equal to the significance level $\alpha$, then the probability of a false positive is bounded by $\alpha$.*

*Proof* If the null hypothesis is simple, and can be expressed as $\theta = \theta_{\text{null}}$ for some $\theta_{\text{null}}$, then by Theorem 10.13,

$$\text{P}\left(\text{False positive}\right) = \text{P}\left(\text{pv}(\tilde{t}_{\theta_{\text{null}}}) \leq \alpha\right) \leq \alpha. \qquad (10.41)$$

Now, let us assume that the null hypothesis is composite, and can be expressed as $\theta \in \Theta_{\text{null}}$ for some set $\Theta_{\text{null}}$. To ease notation, we define $\text{pv}_{\theta_0}(t) := \text{P}\left(\tilde{t}_{\theta_0} \geq t\right)$, which is the p-value function under the null hypothesis $\theta = \theta_0$. By definition of the p-value function for composite hypotheses, $\text{pv}(t) = \sup_{\theta \in \Theta_{\text{null}}} \text{pv}_\theta(t)$. For a false positive to occur, the null hypothesis must hold, so $\theta = \theta_0$ for some $\theta_0 \in \Theta_{\text{null}}$.

For any such $\theta_0$,

$$\begin{aligned} \mathrm{P}\left(\text{False positive}\right) &= \mathrm{P}\left(\mathrm{pv}(\tilde{t}_{\theta_0}) \leq \alpha\right) && (10.42)\\ &= \mathrm{P}\left(\sup_{\theta \in \Theta_{\text{null}}} \mathrm{pv}_{\theta}(\tilde{t}_{\theta_0}) \leq \alpha\right) && (10.43)\\ &\leq \mathrm{P}\left(\mathrm{pv}_{\theta_0}(\tilde{t}_{\theta_0}) \leq \alpha\right) && (10.44)\\ &\leq \alpha, && (10.45) \end{aligned}$$

where (10.45) follows directly from Theorem 10.13. The inequality in (10.44) holds because $\sup_{\theta \in \Theta_{\text{null}}} \mathrm{pv}_{\theta}(\tilde{t}_{\theta_0}) \geq \mathrm{pv}_{\theta_0}(\tilde{t}_{\theta_0})$ if $\theta_0 \in \Theta_{\text{null}}$. ■

## 10.6 The Power

As explained in Section 10.5, the significance level of a hypothesis test is a bound on the probability of a false positive. In contrast, the *power* of a test is the probability of a *true positive*, i.e. of rejecting the null hypothesis when it does *not* hold. To be useful, a hypothesis test must have sufficient power. If we never reject the null hypothesis, we definitely won't incur any false positives, but won't find any true positives either!

To formally define the power for a parametric test, we introduce the power function, which maps each possible value of the parameter to the corresponding probability of rejecting the null hypothesis.

**Definition 10.15** (Power function)**.** *Let $\tilde{t}_{\theta}$ be a random variable representing a test statistic with a parametric distribution that depends on a parameter (or vector of parameters) $\theta$, and let* pv *be the p-value function associated to a certain null hypothesis. The power function maps $\theta$ to the corresponding probability of rejecting the null hypothesis,*

$$\mathrm{pow}(\theta) := \mathrm{P}\left(\mathrm{pv}\left(\tilde{t}_{\theta}\right) \leq \alpha\right), \tag{10.46}$$

*where $\alpha \in [0, 1]$ is the significance level of the test.*

The term *power function* is somewhat misleading, because the function is also defined for values of the parameters associated to the null hypothesis. Let $\Theta_{\text{null}}$ and $\Theta_{\text{alt}}$ be the set of values of $\theta$ associated to the null and alternative hypotheses, respectively. For $\theta \in \Theta_{\text{alt}}$, $\mathrm{pow}(\theta)$ is equal to the power of the test, defined as the probability of a true positive. In contrast, for $\theta \in \Theta_{\text{null}}$, $\mathrm{pow}(\theta)$ is the probability of a false positive. Therefore, we would like $\mathrm{pow}(\theta)$ to be as close to zero as possible for $\theta \in \Theta_{\text{null}}$ and as close to one as possible for $\theta \in \Theta_{\text{alt}}$.

**Example 10.16** (Unfair die: Power function)**.** In this example we derive the power function for the hypothesis test described in Example 10.9. We begin by defining the set of values of the test statistic for which we reject the null hypothesis. This set is known as the *rejection region*:

$$\mathcal{R} := \left\{\tau : \mathrm{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq \tau\right) \leq \alpha\right\}, \tag{10.47}$$

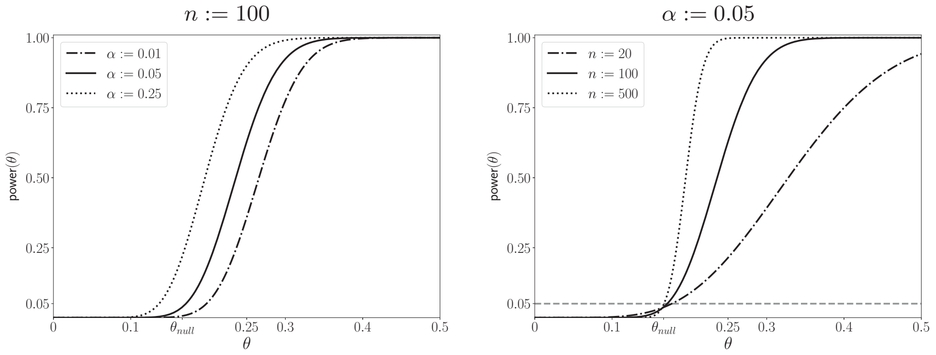


**Figure 10.4 Power function for a (possibly) unfair die.** The graphs show the power function derived in Example 10.16 for different values of the significance level $\alpha$ and number of data $n$. The power function stays below $\alpha$ for $\theta \leq \theta_{\text{null}}$. Increasing $\alpha$ shifts the whole curve upwards (left plot): the probability of true positives is larger, but so is the probability of false positives. In contrast, increasing the number of data while keeping $\alpha$ constant (right plot) increases the power for the alternative hypothesis, but not the probability of a false positive.

where $\theta_{\text{null}} := 1/6$. Let $\tau_{\text{thresh}}$ denote the minimum value in the rejection region $\mathcal{R}$. It equals

$$\tau_{\text{thresh}} := \min_{0 \leq \tau \leq n} \left\{ \tau : \mathrm{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq \tau\right) \leq \alpha \right\} \tag{10.48}$$

$$= \min_{0 \leq \tau \leq n} \left\{ \tau : \sum_{i=\tau}^{n} \binom{n}{i} \theta_{\text{null}}^{i} \left(1 - \theta_{\text{null}}\right)^{n-i} \leq \alpha \right\}. \tag{10.49}$$

If $\tau \geq \tau_{\text{thresh}}$, then $\tau \in \mathcal{R}$, since

$$\mathrm{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq \tau\right) \leq \mathrm{P}\left(\tilde{t}_{\theta_{\text{null}}} \geq \tau_{\text{thresh}}\right) \leq \alpha. \tag{10.50}$$

Conversely, if $\tau < \tau_{\text{thresh}}$, then $\tau \notin \mathcal{R}$ because $\tau_{\text{thresh}}$ is the minimum of $\mathcal{R}$. Therefore, we reject the null hypothesis if and only if the test statistic is greater or equal to the threshold $\tau_{\text{thresh}}$. If $n := 100$, the threshold equals 27 for $\alpha := 0.01$, 24 for $\alpha := 0.05$, and 20 for $\alpha := 0.25$. The observed value of the test statistic is 21, so we reject the null hypothesis if $\alpha := 0.25$, but not if $\alpha$ equals 0.01 or 0.05.

Since the events $\mathrm{pv}\left(\tilde{t}_{\theta}\right) \leq \alpha$ and $\tilde{t}_{\theta} \geq \tau_{\text{thresh}}$ are equivalent, we can express the power function in terms of $\tau_{\text{thresh}}$,

$$\mathrm{pow}(\theta) := \mathrm{P}\left(\mathrm{pv}\left(\tilde{t}_{\theta}\right) \leq \alpha\right) \tag{10.51}$$

$$= \mathrm{P}\left(\tilde{t}_{\theta} \geq \tau_{\text{thresh}}\right) \tag{10.52}$$

$$= \sum_{i=\tau_{\text{thresh}}}^{n} \binom{n}{i} \theta^{i} \left(1 - \theta\right)^{n-i}. \tag{10.53}$$

The graph on the left of Figure 10.4 shows the power function for different values of $\alpha$ when $n := 100$. In all cases, the power function is bounded by $\alpha$ for $\theta \leq \theta_{\text{null}}$, as established in Theorem 10.14.*

If we increase $\theta$, then the power also increases. This makes sense: if the probability of rolling a three is higher, then we are more likely to reject the null hypothesis. Increasing the significance level $\alpha$, also increases the power, because for any $\alpha_1 \leq \alpha_2$,

$$\mathrm{P}\left(\mathrm{pv}\left(\tilde{t}_\theta\right) \leq \alpha_1\right) \leq \mathrm{P}\left(\mathrm{pv}\left(\tilde{t}_\theta\right) \leq \alpha_2\right). \tag{10.54}$$

Therefore higher $\alpha$ results in more power. For instance, if $\theta := 1/4$, the power equals 0.358 for $\alpha := 0.01$, 0.629 for $\alpha := 0.05$ and 0.9 for $\alpha := 0.25$. If the probability of rolling a three is 1/4, the test correctly rejects the null hypothesis one third of the time if $\alpha := 0.01$, a bit less than two thirds of the time if $\alpha := 0.05$, and 90% of the time if $\alpha := 0.25$. Of course, increasing $\alpha$ does not solve all our problems. It comes at the cost of raising the probability of a false positive. If $\theta := \theta_{\text{null}} := 1/6$, then $\text{pow}(\theta_{\text{null}})$ is very close to $\alpha$, so for $\alpha := 0.25$ one out of four tests result in a false positive!

Often, the only way to increase power without also increasing the probability of false positives is to use more data. The graph on the right of Figure 10.4 shows the power function for $\alpha := 0.05$ and $n \in \{20, 100, 500\}$. The power function remains bounded by $\alpha$ for $\theta \leq \theta_{\text{null}}$, as guaranteed by Theorem 10.14, but the power clearly increases with $n$ over the region $\theta \geq \theta_{\text{null}}$. For $\theta := 1/4$, the power equals 0.214 for $n := 20$, 0.629 for $n := 100$, and 0.998 for $n := 500$. This means that if the probability of rolling a three is 1/4, then for $n := 20$ we fail to reject the null hypothesis 78.6% of the time, so a false negative is very likely. In such situations, a hypothesis test is said to be *underpowered*. For $n := 100$, we fail to reject more than one third of the time. In contrast, for $n := 500$ we are almost guaranteed to reject; the probability of a false negative is just 0.2%.

.....................................................................................

The power function makes it possible to compare different hypothesis tests designed for the same problem, such as the one-tailed and two-tailed tests in Example 10.11.

**Example 10.17** (Free throws under pressure: Power function)**.** In order to compare the one-tailed and two-tailed tests in Example 10.11, we fix the home free-throw percentage to $\theta_{\text{home}} := 0.685$, which was Antetokounmpo's free-throw percentage during the 2020/2021 regular season. We then derive the power function as a function of the away free-throw percentage $\theta_{\text{away}}$. We follow a similar strategy to Example 10.16. First, we compute the smallest value of the test statistic for which we reject the null hypothesis. Then, we compute the probability that the test statistic exceeds that value as a function of $\theta_{\text{away}}$.

*Recall that in Example 10.9 we show that this p-value function can be interpreted as being associated to the composite null hypothesis $\theta \in \Theta_{\text{null}} := [0, 1/6]$.

The p-value function for the one-tailed test in Example 10.11 is

$$\mathrm{pv}(t) = 1 - F_{\tilde{z}}\left(\frac{t}{\sigma_{\text{null}}}\right), \tag{10.55}$$

where $\tilde{z}$ is a standard Gaussian random variable, and $\sigma_{\text{null}}^2$ is the variance of the test statistic $\tilde{t}_{\text{1-tail}}$ under the null hypothesis that $\theta_{\text{away}} = \theta_{\text{home}}$. As explained in the derivation of Definition 10.10, in that case $\tilde{t}_{\text{1-tail}}$ is approximately Gaussian with mean zero and variance

$$\theta_{\text{home}}(1 - \theta_{\text{home}})\left(\frac{1}{n_{\text{home}}} + \frac{1}{n_{\text{away}}}\right), \tag{10.56}$$

where $n_{\text{home}}$ and $n_{\text{away}}$ are the number of home and away games, respectively. In order to derive the power function, we assume that the p value would be computed using the true value of $\theta_{\text{home}}$ in (10.56). Note that this is not completely accurate. In practice, we don't know $\theta_{\text{home}}$, and must approximate it from the available data.

The cdf of a Gaussian is invertible, so from (10.55) we can find the threshold $\tau_{\text{1-tail}}$ at which we reject the null hypothesis for a significance level $\alpha$ by solving the equation

$$1 - F_{\tilde{z}}\left(\frac{\tau_{\text{1-tail}}}{\sigma_{\text{null}}}\right) = \alpha, \tag{10.57}$$

which yields

$$\tau_{\text{1-tail}} = \sigma_{\text{null}} F_{\tilde{z}}^{-1}(1 - \alpha). \tag{10.58}$$

For $\alpha = 0.05$, $\tau_{\text{1-tail}} = 0.166$. Recall that the observed test statistic is $0.236 > \tau_{\text{1-tail}}$, so we reject the null hypothesis, as discussed in Example 10.11.

By the same reasoning, the threshold for the test statistic in the two-tailed test from Example 10.11 is obtained by solving the equation

$$2\left(1 - F_{\tilde{z}}\left(\frac{\tau_{\text{2-tails}}}{\sigma_{\text{null}}}\right)\right) = \alpha, \tag{10.59}$$

which yields

$$\tau_{\text{2-tails}} = \sigma_{\text{null}} F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right). \tag{10.60}$$

For $\alpha = 0.05$, $\tau_{\text{2-tails}} = 0.198$. We also reject the null hypothesis for this test, since the observed test statistic is $0.236 > \tau_{\text{2-tails}}$.

In order to derive the power function, we model the test statistic for the one-tailed test as the random variable

$$\tilde{t}_{\theta} := \frac{1}{n_{\text{home}}} \sum_{i=1}^{n_{\text{home}}} \tilde{h}_i - \frac{1}{n_{\text{away}}} \sum_{i=1}^{n_{\text{away}}} \tilde{a}_i, \tag{10.61}$$

where $\tilde{h}_i$ and $\tilde{a}_i$ are independent Bernoulli random variables representing the $i$th free throw attempted at home and away respectively. We assume that all home games share the same Bernoulli parameter $\theta_{\text{home}}$, and all away games share a different Bernoulli parameter $\theta_{\text{away}}$, so the distribution only depends on the parameter vector

$$\theta := \begin{bmatrix} \theta_{\text{home}} \\ \theta_{\text{away}} \end{bmatrix}. \tag{10.62}$$

By the Gaussian approximation to the binomial distribution (Definition 9.39) and Theorem 3.32, $\tilde{t}_\theta$ is approximately Gaussian with mean $\theta_{\text{home}} - \theta_{\text{away}}$ and variance

$$\text{Var}\left[\tilde{t}_\theta\right] = \frac{\theta_{\text{home}}(1-\theta_{\text{home}})}{n_{\text{home}}} + \frac{\theta_{\text{away}}(1-\theta_{\text{away}})}{n_{\text{away}}} := \sigma_{\text{pow}}^2. \tag{10.63}$$

The power function for the one-tailed test under this parametric model is

$$\begin{align}
\text{pow}(\theta) &:= \text{P}\left(\text{pv}\left(\tilde{t}_\theta\right) \leq \alpha\right) \tag{10.64}\\
&= \text{P}\left(\tilde{t}_\theta \geq \tau_{\text{1-tail}}\right) \tag{10.65}\\
&= \text{P}\left(\sigma_{\text{pow}}\tilde{z} + \theta_{\text{home}} - \theta_{\text{away}} \geq \tau_{\text{1-tail}}\right) \tag{10.66}\\
&= \text{P}\left(\tilde{z} \geq \frac{\tau_{\text{1-tail}} - (\theta_{\text{home}} - \theta_{\text{away}})}{\sigma_{\text{pow}}}\right), \tag{10.67}
\end{align}$$

where $\tilde{z}$ is a standard Gaussian random variable with mean zero and unit variance. The power function for the two-tailed test is

$$\begin{align}
&\text{pow}(\theta) = \text{P}\left(\left|\tilde{t}_\theta\right| \geq \tau_{\text{2-tails}}\right) \tag{10.68}\\
&= \text{P}\left(\sigma_{\text{pow}}\tilde{z} + \theta_{\text{home}} - \theta_{\text{away}} \geq \tau_{\text{2-tails}}\right) + \text{P}\left(\sigma_{\text{pow}}\tilde{z} + \theta_{\text{home}} - \theta_{\text{away}} \leq -\tau_{\text{2-tails}}\right)\\
&= \text{P}\left(\tilde{z} \geq \frac{\tau_{\text{2-tails}} - (\theta_{\text{home}} - \theta_{\text{away}})}{\sigma_{\text{pow}}}\right) + \text{P}\left(\tilde{z} \leq \frac{-\tau_{\text{2-tails}} - (\theta_{\text{home}} - \theta_{\text{away}})}{\sigma_{\text{pow}}}\right).
\end{align}$$

Figure 10.5 shows the two power functions for $\alpha = 0.05$ as a function of $\theta_{\text{away}}$. We observe both power functions are bounded by $\alpha$ when the null hypothesis holds ($\theta_{\text{away}} = \theta_{\text{home}}$), as guaranteed by Theorem 10.14. For $\theta_{\text{away}} < \theta_{\text{home}}$, the power of both tests increases smoothly as we move away from $\theta_{\text{home}}$, but it is clearly higher for the one-tailed test, which therefore has a higher probability of correctly rejecting the null hypothesis than the two-tailed test. In contrast, for $\theta_{\text{away}} > \theta_{\text{home}}$, the power of the two-tailed test again increases smoothly away from $\theta_{\text{home}}$, but the one-tailed test has no power. The main difference is therefore that the two-tailed test is able to detect situations where $\theta_{\text{away}} > \theta_{\text{home}}$ (as it was designed to do), which results in a loss of power for $\theta_{\text{away}} < \theta_{\text{home}}$.

.........................................................................................

Computing the power function of a hypothesis test analytically is often very complicated or even intractable. As an alternative, we can leverage the Monte Carlo method from Section 1.7 to estimate the power function in such situations. The key insight is that the power is a probability of a certain event (rejection of the null hypothesis), so we can approximate it via the Monte Carlo method as long as we can simulate the event.

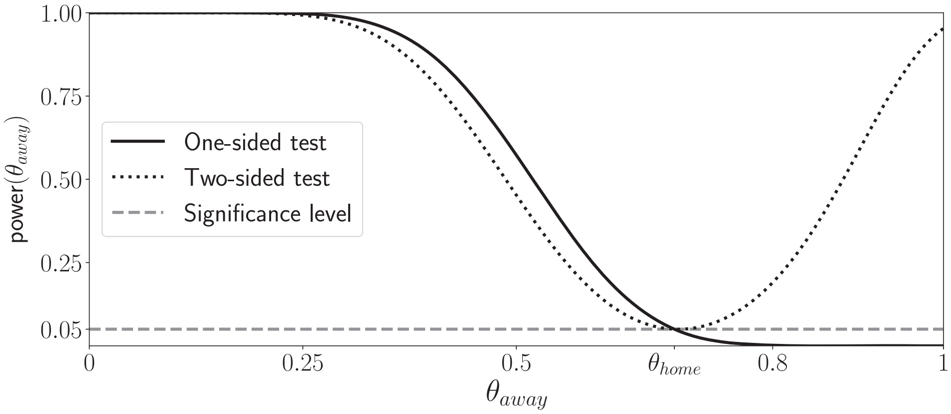


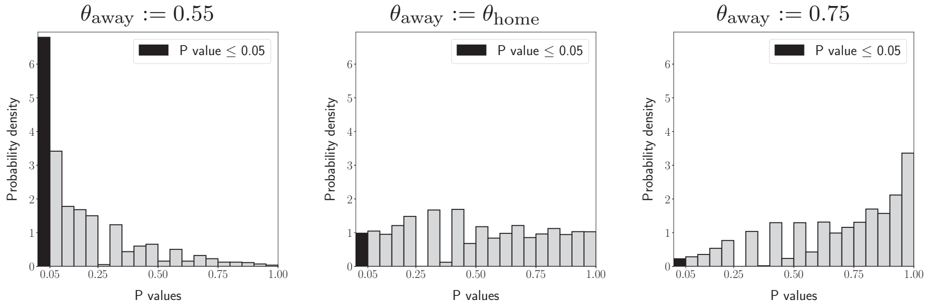
**Figure 10.5 Power function for free throws under pressure.** The graph shows the power functions of the one-tailed and two-tailed tests in Example 10.11 as a function of the free-throw percentage in away games $\theta_{\text{away}}$ when the percentage at home games is $\theta_{\text{home}} := 0.685$. For $\theta_{\text{away}} < \theta_{\text{home}}$, the power of both tests increases away from $\theta_{\text{home}}$, but it is higher for the one-tailed test. In contrast, for $\theta_{\text{away}} > \theta_{\text{home}}$, the power of the two-tailed test again increases away from $\theta_{\text{home}}$, but the one-tailed test has no power.

**Definition 10.18** (Monte Carlo power estimation)**.** *Let $\tilde{t}_\theta$ denote a test statistic following a known parametric distribution with parameters $\theta$. To estimate the power function at $\theta$, we first simulate $k$ independent samples of $\tilde{t}_\theta$: $t_1, \ldots, t_k$. Then we apply the p-value function of the hypothesis test to each of the samples. By Definition 10.15, the power function at $\theta$ is the probability of rejecting the null hypothesis, or equivalently the probability that the p value is not greater than the significance level $\alpha$. We approximate this probability by the corresponding empirical probability, which equals the fraction of p values smaller than or equal to $\alpha$,*

$$\operatorname{pow}(\theta) := \mathrm{P}\left(\operatorname{pv}\left(\tilde{t}_\theta\right) \leq \alpha\right) \approx \frac{1}{k}\sum_{i=1}^{k} 1(\operatorname{pv}\left(t_i\right) \leq \alpha), \tag{10.69}$$

*where $1(\operatorname{pv}(t_i) \leq \alpha)$ is an indicator function that is equal to one if $\operatorname{pv}(t_i) \leq \alpha$ and to zero otherwise.*

In Figure 10.6 we apply the Monte Carlo method to approximate the power function of the one-tailed test in Example 10.17 for several values of $\theta_{\text{away}}$. Interestingly, this is *more accurate* than our theoretical derivation in Example 10.17 (but also much more computationally expensive). The reason is that in the theoretical derivation we plug the true parameter $\theta_{\text{home}}$ into (10.56) in order to compute the variance of the test statistic under the null hypothesis. However, in practice the parameter is unknown and must be estimated from the data, as explained in Definition 10.10. When performing the Monte Carlo simulations, we



Derivation = 0.384
Estimate = 0.339

Derivation = 0.050
Estimate = 0.049

Derivation = 0.009
Estimate = 0.011

**Figure 10.6 Monte Carlo power-function estimation.** The plots depict a Monte Carlo estimate of the power function of the one-tailed test in Example 10.17 for different values of the parameter $\theta_{\text{away}}$. Each plot shows the histogram of one million p values computed from independent samples of the test statistic. The fraction of p values below the significance threshold (black bar) approximate the probability of rejecting the null hypothesis for that value of $\theta$, and therefore the power function of the test. Below each graph, we compare the Monte Carlo estimate and the power function derived analytically in Example 10.17.

take this into account by using the estimated value of $\theta_{\text{home}}$ to compute the p value.

## 10.7 Nonparametric Testing: The Permutation Test

In the previous sections, we focus on parametric hypothesis tests, which rely on a parametric model to describe the distribution of the test statistic. In practice, it can be challenging to design such parametric models, as illustrated for instance by Example 10.26 below. Fortunately, it is possible to perform hypothesis testing without parametric models. In this section, we study one of the most popular nonparametric tests: the permutation test. Section 10.7.1 provides an intuitive description based on a toy example. In Section 10.7.2, we define the p value of the permutation test formally, and explain how to compute it in practice. Finally, Section 10.7.3 shows that permutation tests allow us to control the probability of a false positive, just like parametric tests.

### *10.7.1 Intuition*

In this section we use a toy example to explain the logic underlying permutation tests. My friends from Spain often complain about prices in New York, claiming that it is more expensive than Madrid. Our goal is to verify this via hypothesis testing, focusing on hamburger prices.

Following the steps in Definition 10.12, we begin by defining the null hypothesis.

Our original hypothesis is that prices are different in both cities, so we choose the null hypothesis that *the distribution of burger prices is the same in both cities.* Then, we select a test statistic that should be small if the null hypothesis holds, and large if it does not. A natural choice is the difference between the mean burger price in New York and in Madrid. Our data are the price of a cheeseburger at four restaurants:

$$\text{NY: \$16}, \quad \text{NY: \$18}, \quad \text{Madrid: \$13}, \quad \text{Madrid: \$12}. \tag{10.70}$$

The observed value of this test statistic equals

$$t_{\text{data}} = m(\text{NY}) - m(\text{Madrid}) = \frac{16+18}{2} - \frac{13+12}{2} = 4.5, \tag{10.71}$$

where $m(\text{NY})$ and $m(\text{Madrid})$ denote the sample mean of burger prices in New York and Madrid respectively. This value seems pretty high, but we don't have a lot of data, so it could just be due to random fluctuations. To settle the question, we need to model the test statistic under the null hypothesis and compute the p value. Our goal is to do this without assuming that we have a parametric model for the distribution of the test statistic.

In order to bypass the use of a parametric model, we exploit a key implication of the null hypothesis. If the distribution of burger prices is the same in New York and Madrid, then the label indicating the city associated to each data point is *meaningless*. Therefore, permuting the data while fixing the labels *should not change the behavior of the test statistic*. If the null hypothesis holds, we would have been equally likely to observe

$$\text{NY: \$13}, \quad \text{NY: \$18}, \quad \text{Madrid: \$12}, \quad \text{Madrid: \$16}, \tag{10.72}$$

or any of the 4! = 24 possible permutations of the labels listed in Figure 10.7. Permuting the entry changes the value of the test statistic. For example, the test statistic for (10.72) equals

$$t = m(\text{NY}) - m(\text{Madrid}) = \frac{18+13}{2} - \frac{16+12}{2} = 1.5. \tag{10.73}$$

If the labels are meaningless, the test statistic associated to each permutation should occur with the same probability. In that case, it is unlikely for $t_{\text{data}}$ to be larger than most of them. Consequently, a large $t_{\text{data}}$ suggests that the data contradict the null hypothesis. To determine whether $t_{\text{data}}$ is abnormally large, we compute the fraction of permutations with test statistics greater than or equal to $t_{\text{data}}$. In our example, there are four such permutations (see Figure 10.7). This means that under the null hypothesis, we would expect to see this value of the test statistic $4/24 = 16.7\%$ of the time. The probability is small, but not small enough to be overwhelming evidence against the null hypothesis.

### 10.7.2 The P Value Of The Permutation Test

The approach described in Section 10.7.1 yields a probability quantifying the consistency between the null hypothesis and the observed test statistic. In other

| NY | NY | M | M | t |
|---|---|---|---|---|
| 12 | 13 | 16 | 18 | -4.5 |
| 12 | 13 | 18 | 16 | -4.5 |
| 12 | 16 | 13 | 18 | -1.5 |
| 12 | 16 | 18 | 13 | -1.5 |
| 12 | 18 | 13 | 16 | 0.5 |
| 12 | 18 | 16 | 13 | 0.5 |
| 13 | 12 | 16 | 18 | -4.5 |
| 13 | 12 | 18 | 16 | -4.5 |
| 13 | 16 | 12 | 18 | -0.5 |
| 13 | 16 | 18 | 12 | -0.5 |
| 13 | 18 | 12 | 16 | 1.5 |
| 13 | 18 | 16 | 12 | 1.5 |
| 16 | 12 | 13 | 18 | -1.5 |
| 16 | 12 | 18 | 13 | -1.5 |
| 16 | 13 | 12 | 18 | -0.5 |
| 16 | 13 | 18 | 12 | -0.5 |
| **16** | **18** | **12** | **13** | **4.5** |
| **16** | **18** | **13** | **12** | **4.5** |
| 18 | 12 | 16 | 13 | 0.5 |
| 18 | 12 | 13 | 16 | 0.5 |
| **18** | **16** | **12** | **13** | **4.5** |
| **18** | **16** | **13** | **12** | **4.5** |
| 18 | 13 | 12 | 16 | 1.5 |
| 18 | 13 | 16 | 12 | 1.5 |

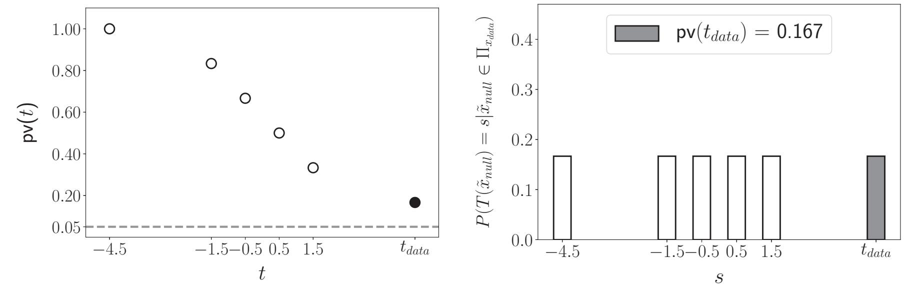


**Figure 10.7 P value for difference in burger prices.** The table shows all the permutations of the data $x_{\text{data}}$ in Section 10.7.1 and the corresponding test statistic. Below we show the p-value function (left) and the conditional pmf of the test statistic given that the data are a permutation of $x_{\text{data}}$ (right). The p value is computed by summing the pmf over all values greater or equal to the observed test statistic $t_{\text{data}}$. This corresponds to the four permutations highlighted in bold in the table. The resulting p value equals 4/24, which is above the significance level of 0.05 indicated by the dashed line.

words, it allows us to compute a p value without a parametric model! However, in contrast to the p value in Definition 10.8, this p value is a conditional probability, because it only accounts for situations where the data are equal to a permutation of the available measurements.

**Definition 10.19** (P-value function and p value of a permutation test)**.** *We define $\Pi_x$ to be the set of vectors that can be obtained by permuting the entries of the vector $x$. For example, if*

$$x := \begin{bmatrix} a \\ b \\ c \end{bmatrix} \qquad \textit{then} \qquad \Pi_x = \left\{ \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \begin{bmatrix} a \\ c \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \\ c \end{bmatrix}, \begin{bmatrix} b \\ c \\ a \end{bmatrix}, \begin{bmatrix} c \\ a \\ b \end{bmatrix}, \begin{bmatrix} c \\ b \\ a \end{bmatrix} \right\}, \tag{10.74}$$

*as long as $a \neq b \neq c$, so that all the elements in $\Pi_x$ are distinct (if this is not the case, the repeated elements should be removed).*

*Let $x_{\text{data}} \in \mathbb{R}^n$ be a vector of data, and let $T : \mathbb{R}^n \to \mathbb{R}$ be a function that maps any $n$-dimensional data vector $x$ to the test statistic $T(x)$ associated to a permutation test. We model the data under the null hypothesis as an $n$-dimensional random vector $\tilde{x}_{\text{null}}$. The p-value function* pv *of the permutation test maps every possible value $t$ of the test statistic to the conditional probability that $\tilde{t}_{\text{null}} := T(\tilde{x}_{\text{null}})$ is greater than or equal to $t$, given that $\tilde{x}_{\text{null}}$ is in $\Pi_{x_{\text{data}}}$:*

$$\text{pv}(t) := \text{P}\left(\tilde{t}_{\text{null}} \geq t \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right). \tag{10.75}$$

*The p value of the observed data $x_{\text{data}}$ is equal to* $\text{pv}(t_{\text{data}})$*, where $t_{\text{data}} := T(x_{\text{data}})$.*

Our reasoning in Section 10.7.1 hinges on the assumption that the data are *exchangeable* under the null hypothesis, which means that if we permute their order (with respect to the labels indicating the city), their joint distribution remains unchanged.

**Definition 10.20** (Exchangeability)**.** *The entries of a random vector $\tilde{x}$ are exchangeable if permuting them does not change the joint distribution of $\tilde{x}$. Let $\Pi_x$ be the set of vectors obtained by permuting the entries of the vector $x$. The entries of a discrete random vector $\tilde{x}$ are exchangeable if for all $x$ in the range of $\tilde{x}$,*

$$p_{\tilde{x}}(x) = p_{\tilde{x}}(v) \qquad \textit{for all } v \in \Pi_x, \tag{10.76}$$

*where $p_{\tilde{x}}$ is the joint pmf of $\tilde{x}$. Similarly, the entries of a $d$-dimensional continuous random vector $\tilde{x}$ are exchangeable if for all $x \in \mathbb{R}^d$,*

$$f_{\tilde{x}}(x) = f_{\tilde{x}}(v) \qquad \textit{for all } v \in \Pi_x, \tag{10.77}$$

*where $f_{\tilde{x}}$ is the joint pdf of $\tilde{x}$.*

The following lemma shows that i.i.d. random variables are exchangeable.

**Lemma 10.21** (Exchangeability of i.i.d. random variables)**.** *The elements of any sequence of $d$ i.i.d. random variables $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_d$ are exchangeable.*

*Proof* Let us assume that the random vector is continuous. The same argument can be applied to discrete random vectors replacing the joint pdf with the joint pmf. By the i.i.d. assumption,

$$f_{\tilde{x}}(x) = \prod_{i=1}^{d} f_{\tilde{x}_i}(x_i) = \prod_{i=1}^{d} f_{\text{marg}}(x_i), \tag{10.78}$$

where $f_{\text{marg}}$ denotes the marginal pdf of the entries. Therefore, for any $v \in \Pi_x$,

$$f_{\tilde{x}}(v) = \prod_{i=1}^{d} f_{\text{marg}}(v_i) = \prod_{i=1}^{d} f_{\text{marg}}(x_i) = f_{\tilde{x}}(x), \tag{10.79}$$

because the entries of $v$ are the same as the entries of $x$ (in a different order). ■

Under exchangeability, the conditional distribution of the test statistic in a permutation test is always the same, regardless of the marginal distribution. This key insight is what allows us to compute the p value without a parametric model.

**Theorem 10.22** (P-value function for permutation test)**.** *Let $x_{\text{data}} \in \mathbb{R}^d$ be a vector of data, $\tilde{x}_{\text{null}}$ an $n$-dimensional random vector representing the data under the null hypothesis, and $T : \mathbb{R}^n \to \mathbb{R}$ a test statistic associated to a permutation test. If the entries of $\tilde{x}_{\text{null}}$ are exchangeable, then the p-value function of the permutation test is*

$$\operatorname{pv}(t) = \frac{\sum_{v \in \Pi_{x_{\text{data}}}} 1\left(T(v) \geq t\right)}{|\Pi_{x_{\text{data}}}|}, \tag{10.80}$$

*where $\Pi_{x_{\text{data}}}$ is the set of vectors obtained by permuting the entries of $x$, and $|\Pi_{x_{\text{data}}}|$ is its cardinality (i.e. the number of different permutations, which is $n!$ if the entries of $x$ are distinct). $1\left(T(v) \geq t\right)$ is an indicator function that is equal to one if $T(v) \geq t$ and to zero otherwise.*

*Proof* If $\tilde{x}_{\text{null}}$ is discrete, then exchangeability implies that for any $v_1, v_2 \in \Pi_{x_{\text{data}}}$ $p_{\tilde{x}_{\text{null}}}(v_1) = p_{\tilde{x}_{\text{null}}}(v_2)$, so

$$\begin{aligned} P\left(\tilde{x}_{\text{null}} = v_1 \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) &= \frac{P\left(\tilde{x}_{\text{null}} = v_1, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)}{P\left(\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)} && (10.81) \\ &= \frac{p_{\tilde{x}_{\text{null}}}(v_1)}{P\left(\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)} && (10.82) \\ &= \frac{p_{\tilde{x}_{\text{null}}}(v_2)}{P\left(\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)} && (10.83) \\ &= P\left(\tilde{x}_{\text{null}} = v_2 \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right). && (10.84) \end{aligned}$$

Consequently, $P\left(\tilde{x}_{\text{null}} = v \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)$ is the same for all $v \in \Pi_{x_{\text{data}}}$. These $|\Pi_{x_{\text{data}}}|$ probabilities need to add up to one, because

$$\begin{aligned} \sum_{v \in \Pi_{x_{\text{data}}}} P\left(\tilde{x}_{\text{null}} = v \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) &= P\left(\cup_{v \in \Pi_{x_{\text{data}}}} \tilde{x}_{\text{null}} = v \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) \\ &= P\left(\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}} \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) && (10.85) \\ &= 1, && (10.86) \end{aligned}$$

so

$$P\left(\tilde{x}_{\text{null}} = v \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) = \frac{1}{|\Pi_{x_{\text{data}}}|}. \tag{10.87}$$

The same holds if $\tilde{x}_{\text{null}}$ is continuous. The proof is similar but requires taking limits, as in the proof of Theorem 6.6, since the probability that $\tilde{x}_{\text{null}}$ equals any specific value is zero (see Section 3.1).

The p-value function can be derived directly from (10.87), as it is the conditional probability of the union of the disjoint events $\tilde{x}_{\text{null}} = v$ for every $v$ such that $T(v) \geq t$,

$$\begin{aligned}
\operatorname{pv}(t) &:= \operatorname{P}\left(T\left(\tilde{x}_{\text{null}}\right) \geq t \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) && (10.88)\\
&= \operatorname{P}\left(\cup_{\left\{v \in \Pi_{x_{\text{data}}} : T(v) \geq t\right\}} \left\{\tilde{x}_{\text{null}} = v\right\} \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) && (10.89)\\
&= \sum_{\left\{v \in \Pi_{x_{\text{data}}} : T(v) \geq t\right\}} P\left(\tilde{x}_{\text{null}} = v \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) && (10.90)\\
&= \frac{\sum_{v \in \Pi_{x_{\text{data}}}} 1\left(T(v) \geq t\right)}{\left|\Pi_{x_{\text{data}}}\right|}. && (10.91)
\end{aligned}$$

■

Theorem 10.22 establishes that our intuitive calculation in Section 10.7.1 is correct: the p value of the permutation test equals the fraction of permutations with a test statistic greater than or equal to the observed test statistic $t_{\text{data}}$, as illustrated in Figure 10.7. A key question is whether thresholding this p value enables us to control the probability of a false positive, as in parametric testing (see Section 10.5). Section 10.7.3 establishes that this is indeed the case: rejecting the null hypothesis when the p value of a permutation test is below the significance level $\alpha$ guarantees that the probability of a false positive is bounded by $\alpha$.

It is usually impossible to compute the exact p value associated to a permutation test, because the number of possible permutations is too large. For instance, if there are $n := 40$ data points, the number of permutations is $40! > 10^{47}$. In practice, the p value must be approximated computationally. This can be achieved by sampling many independent permutations, and determining what fraction of them result in a test statistic greater than or equal to $t_{\text{data}}$. The resulting fraction is a Monte Carlo estimate of the conditional probability of the event $T\left(\tilde{x}_{\text{null}}\right) \geq t_{\text{data}}$ given $\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}$ (see Section 1.7).

**Definition 10.23** (Permutation test via the Monte Carlo method)**.** *Let $x_{\text{data}} \in \mathbb{R}^n$ be a vector of data and $T : \mathbb{R}^n \to \mathbb{R}$ a test statistic. To perform a permutation test for a null hypothesis that implies exchangeability of the data points we:*

*1 Generate $k$ independent permutations $v_1, \ldots, v_k \in \Pi_{x_{\text{data}}}$ of the entries of $x_{\text{data}}$.*
*2 Compute the test statistic corresponding to each permutation, $t_i = T(v_i)$, $1 \leq i \leq k$.*
*3 Compute the p-value estimate*

$$\operatorname{pv}(t_{\text{data}}) = \frac{\sum_{i=1}^{k} 1\left(T(v_i) \geq t_{\text{data}}\right)}{k}, \qquad (10.92)$$

*where $t_{\text{data}} := T(x)$ is the observed test statistic and $1\left(T(v) \geq t\right)$ is an indicator function that is equal to one if $T(v) \geq t$ and to zero otherwise.*

Permutation tests based on the Monte Carlo method are close in spirit to the bootstrap technique described in Section 9.9. When performing bootstrapping, we sample the data with replacement to approximate the distribution of a statistic of interest. When applying a permutation test, we sample the data without

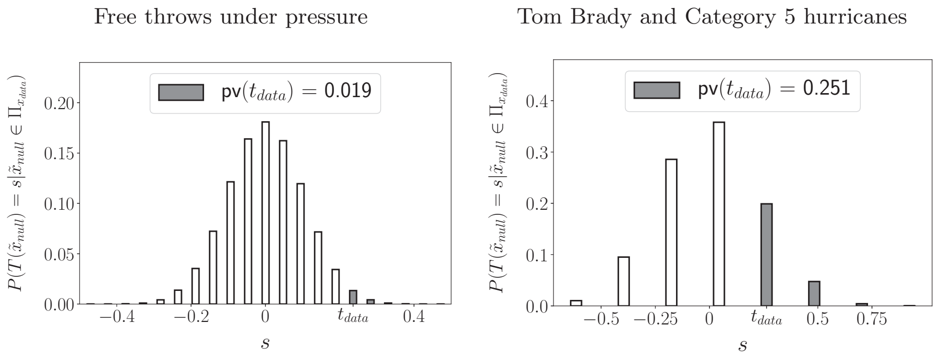


**Figure 10.8 Two-sample permutation tests.** P-value calculation for the two-sample permutation test in Definition 10.23 applied to the data in Examples 10.2 and 1.30 (see Example 10.25). The graphs show the conditional pmf of the test statistic, computed using $k := 10^6$ independent permutations of the data. The p value is obtained by summing over the values that are greater than or equal to the observed test statistic $t_{\text{data}}$ (gray bars). On the left, the p value is small enough to reject the null hypothesis based on a significance level of 0.05. On the right, the p value is very large, so we do not reject the null hypothesis.

replacement to generate the random permutations that are used to approximate the p value.

**Example 10.24** (Free throws under pressure: Permutation test)**.** In this example, we apply a permutation test to the data in Example 10.2. Under the null hypothesis that the free-throw percentage is the same at home and away (see Example 10.4), the data can be modeled as exchangeable by Lemma 10.21, as long as we assume that the free throws are independent. The number of data is 85, so we cannot consider all possible permutations. We instead resort to the Monte-Carlo approach in Definition 10.23. We set $k := 10^6$ and use the same test statistic defined in Example 10.6: the difference between the free-throw percentage at home and away.

The left graph in Figure 10.8 shows the resulting conditional pmf of the test statistic. The conditional pmf is quite different from the continuous pdf of the test statistic in the one-tailed parametric test of Example 10.11 (bottom left graph of Figure 10.2). However, the resulting p values are quite close: 0.019 for the permutation test and 0.011 for the parametric test. The p-value functions of both tests, depicted in Figure 10.9, are also similar.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

**Example 10.25** (Tom Brady and Category 5 hurricanes: Permutation test)**.** The data in Example 1.30 seem to indicate that when Tom Brady wins the Superbowl, Category 5 hurricanes in the North Atlantic Ocean are more likely. No matter how you feel about Tom Brady, this doesn't make a lot of sense. In order to check

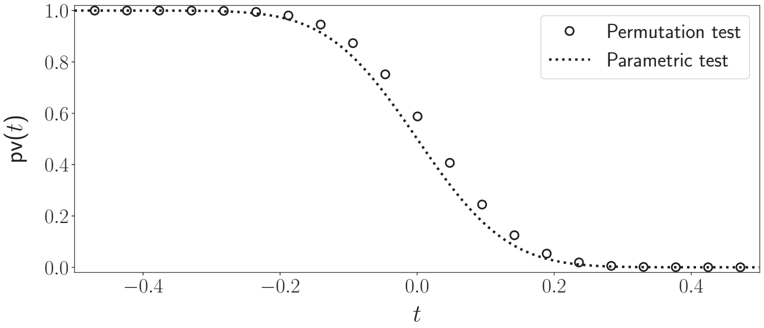


**Figure 10.9 Comparison of p-value function for permutation and parametric tests.** The graph shows the p-value functions corresponding to the permutation test in Figure 10.8 (left plot) and the parametric test derived in Example 10.11. The permutation test produces a similar p-value function, without requiring a parametric model.

whether this pattern is due to random chance, we apply a permutation test. Our null hypothesis is that the distribution of hurricanes has nothing to do with Tom Brady. Ideally, we should apply the test to new data. Otherwise, a small p value does not necessarily imply that something interesting is going on, because the data have been cherry-picked, as explained in Section 10.1. However, the result is instructive, so let us take a look anyway (but don't do this at home!).

We encode the data in Table 1.4 as an $n$-dimensional vector $x$, where $x[i] = 1$ if there was a hurricane in year $i$, and $x[i] = 0$ otherwise. A reasonable test statistic is the difference between the fraction of years with hurricanes when Tom Brady wins the Superbowl (denoted by $m(x_W)$) and when he does not (denoted by $m(x_L)$):

$$T(x) := m(x_W) - m(x_L) = 0.264 := t_{data}, \tag{10.93}$$

The graph on the right of Figure 10.8 depicts the distribution of the test statistic under the null hypothesis that the data are exchangeable, computed following Definition 10.23 with $k := 10^6$. The p value equals 0.251: one fourth of the random permutations result in a test statistic that is greater or equal to the observed test statistic. This shows that even though the test statistic looks quite large, it is not inconsistent with the null hypothesis. As we suspected, Brady's success is probably not associated to Category 5 hurricanes; we are just seeing a random fluctuation due to the limited number of data.

**Example 10.26** (School grades)**.** Permutation tests are particularly useful when we do not have a parametric model for our data. The graph on the left of Figure 10.10 shows the distribution of grades for a course on Portuguese language in two Portuguese schools, extracted from Dataset 15. Our goal is to determine whether the grades in the two schools are systematically different. Our null hy-

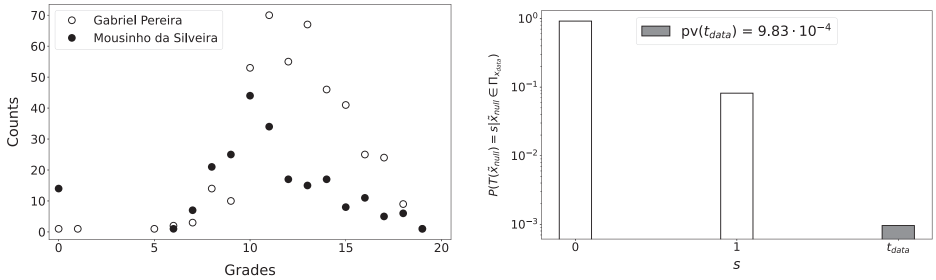


**Figure 10.10 Permutation test for school grades.** The graph on the left shows the grades for a course on Portuguese language in two different schools. The graph on the right depicts the conditional pmf of the test statistic, computed using $k := 10^6$ independent permutations of the data. The p value corresponding to the $t_{\text{data}}$ (gray bar) is smaller than $10^{-3}$, indicating that the observed data are inconsistent with the null hypothesis.

pothesis is that the grade distribution in both schools is the same. We choose the absolute difference between medians as our test statistic.

Designing a parametric test in this setting is challenging for two reasons. First, it is unclear what parametric model can provide a reasonable fit to these data, because of their unique structure. For example, there is a large jump at 10, presumably because this is the grade needed to pass the course. Second, we need to derive the distribution of the test statistic in terms of the model parameters, which may be difficult for some choices of test statistic.

In contrast, applying the permutation test is very straightforward. The observed test statistic equals

$$T(x) := |\text{median}(\text{x}_{\text{GP}}) - \text{median}(\text{x}_{\text{MS}})| = 2 := t_{data}, \tag{10.94}$$

where $x$ represents the data, $x_{\text{GP}}$ are the grades from the school Gabriel Pereira and $x_{\text{MS}}$ are the grades from the school Mousinho da Silveira. The graph on the right of Figure 10.10 depicts the distribution of the test statistic under the null hypothesis that the data are exchangeable, computed following Definition 10.23 with $k := 10^6$. The p value is smaller than $10^{-3}$. Consequently, the observed data are extremely unlikely given the null hypothesis; we conclude that the difference between the grade distributions of both schools is statistically significant.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

### *10.7.3 Statistical Significance*

In Definition 10.23, we compare the p value of a permutation test with a predefined significance level $\alpha$ in order to decide whether to reject the null hypothesis. This allows us to control the probability of a false positive, as in the case of parametric testing (see Section 10.5). In order to prove this, we first show that the conditional probability of a false positive is bounded by $\alpha$.

**Theorem 10.27** (The significance level bounds the conditional probability of a false positive)**.** *If we reject the null hypothesis of a permutation test when the p value is smaller or equal to a significance level* $\alpha \in [0,1]$*, then the conditional probability of a false positive, given that we observe a permutation of the data, is bounded by* $\alpha$*.*

*Proof* Let $x_{\text{data}}$ represent the available data, and let $\tilde{x}_{\text{null}}$ be a random vector used to model the data under the null hypothesis. We define the random variable $\tilde{t}_{\text{null}} := T(\tilde{x}_{\text{null}})$. The p-value function is the deterministic function

$$\text{pv}(t) := P\left(\tilde{t}_{\text{null}} \geq t \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right). \tag{10.95}$$

The random variable

$$\begin{aligned} \tilde{u} &:= \text{pv}(\tilde{t}_{\text{null}}) & (10.96)\\ &= 1 - F_{\tilde{t}_{\text{null}}}\left(\tilde{t}_{\text{null}} \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) & (10.97)\end{aligned}$$

captures the conditional behavior of the p value under the null hypothesis, where

$$F_{\tilde{t}_{\text{null}}}\left(t \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) := P\left(\tilde{t}_{\text{null}} \leq t \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) \tag{10.98}$$

is the conditional cdf of $\tilde{t}_{\text{null}}$ given the event $\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}$. By the same argument as in the proof of Theorem 10.13 (replacing the cdfs by conditional cdfs),

$$F_{\tilde{u}}(u \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}) \leq u. \tag{10.99}$$

Consequently,

$$\begin{aligned} \text{P}\left(\text{False positive} \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) &= \text{P}\left(\text{pv}(\tilde{t}_{\text{null}}) \leq \alpha \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) & (10.100)\\ &= F_{\tilde{u}}(\alpha \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}) \leq \alpha. & (10.101)\end{aligned}$$

■

Our bound on the conditional probability of a false positive allows us to control the *marginal* probability of a false positive.

**Corollary 10.28** (The significance level bounds the probability of a false positive)**.** *If we reject the null hypothesis of a permutation test when the p value is smaller or equal to a significance level* $\alpha \in [0,1]$*, then the probability of a false positive is bounded by* $\alpha$*.*

*Proof* Let $\tilde{x}_{\text{null}}$ be an $n$-dimensional random vector representing the data under the null hypothesis. We define a random vector $\tilde{w} := \text{sort}(\tilde{x}_{\text{null}})$, obtained by sorting the entries of $\tilde{x}_{\text{null}}$. For example,

$$\text{sort}\left(\begin{bmatrix}1\\0\\3\end{bmatrix}\right) = \begin{bmatrix}0\\1\\3\end{bmatrix}. \tag{10.102}$$

The key idea is that $\tilde{w}$ encodes the entries of $\tilde{x}_{\text{null}}$ but not their order. Consequently the event $\tilde{x}_{\text{null}} \in \Pi_w$ is equivalent to $\tilde{w} = w$. Let $\widetilde{\text{fp}}$ be a Bernoulli random variable that is equal to one if a false positive occurs, and to zero otherwise. If $\tilde{x}_{\text{null}}$ is discrete, by Theorems 4.10 and 4.14

$$\begin{aligned}
\mathrm{P}\left(\text{False positive}\right) &= \sum_{w \in \mathcal{W}} p_{\widetilde{\text{fp}} \,|\, \tilde{w}}(1 \,|\, w) p_{\tilde{w}}(w) && (10.103)\\
&= \sum_{w \in \mathcal{W}} \mathrm{P}\left(\text{False positive} \,|\, \tilde{w} = w\right) p_{\tilde{w}}(w) && (10.104)\\
&= \sum_{w \in \mathcal{W}} \mathrm{P}\left(\text{False positive} \,|\, \tilde{x}_{\text{null}} \in \Pi_w\right) p_{\tilde{w}}(w) && (10.105)\\
&\leq \alpha \sum_{w \in \mathcal{W}} p_{\tilde{w}}(w) = \alpha, && (10.106)
\end{aligned}$$

where $\mathcal{W}$ denotes the range of $\tilde{w}$ (the set of all possible sorted entries of $\tilde{x}_{\text{null}}$) and the inequality follows from Theorem 10.28. If $\tilde{x}_{\text{null}}$ is continuous, the proof is very similar. By Theorem 6.5

$$\mathrm{P}\left(\text{False positive}\right) = \int_{w \in \mathbb{R}^n} p_{\widetilde{\text{fp}} \,|\, \tilde{w}}(1 \,|\, w) f_{\tilde{w}}(w) \, \mathrm{d}w \leq \alpha \int_{w \in \mathbb{R}^n} f_{\tilde{w}}(w) \, \mathrm{d}w = \alpha.$$

■

## 10.8 Multiple Testing

In some domains, it is commonplace to perform many hypothesis tests at the same time. This is known as *multiple testing*. An important example is computational genomics, where thousands of genetic markers may be evaluated to determine whether they are associated to a disease. Sections 10.5 and 10.7.3 establish that when we perform hypothesis testing, the probability of a false positive is controlled by the significance level $\alpha$. However, this only holds *for a single hypothesis test*. If we perform many such tests simultaneously, then the probability of a false positive can be dramatically higher.

Imagine that we carry out $k$ independent hypothesis tests with significance level $\alpha$, and that the probability of a false positive in each test equals $\alpha$. Then the probability of incurring at least one false positive is

$$1 - \mathrm{P}\left(\text{No false positives}\right) = 1 - (1 - \alpha)^k. \qquad (10.107)$$

For $\alpha := 0.05$ and $k := 100$, the probability is 0.99. False positives are essentially guaranteed to occur!

The following example illustrates how easy it is to find apparent evidence for false findings in multiple testing scenarios.

**Example 10.29** (3-point shooting in the clutch)**.** Basketball analysts often praise or criticize players based on their performance *in the clutch*, which refers to the final moments that decide a game. Players that are able to play better at that crucial time are said to be *clutch*. Here we study clutch 3-point shooting during the 2014/2015 NBA season, using data extracted from Dataset 10. We define the clutch as the fourth quarter of games where the final point differential is below 10.

To determine whether a player shoots better in the clutch, we apply a parametric hypothesis test where the parameter of interest is the player's clutch 3-point percentage. Our null hypothesis is that the parameter is equal to the player's season 3-point percentage $\theta_{\text{season}}$, i.e. that the player's accuracy is the same in the clutch as at any other time. Our test statistic of interest $t_{\text{data}}$ is the number of made shots in the clutch, which should be large under the alternative hypothesis that the player shoots better in the clutch, but not under the null hypothesis.

If we assume the shots are independent, the distribution of the made shots under the null hypothesis, represented by the random variable $\tilde{t}_{\theta_{\text{season}}}$, is binomial with parameters $n$ and $\theta_{\text{season}}$, where $n$ denotes the number of clutch shots taken by the player. Therefore the p value equals

$$\operatorname{pv}(t_{\text{data}}) := \mathrm{P}\left(\tilde{t}_{\theta_{\text{season}}} \geq t_{\text{data}}\right) \tag{10.108}$$

$$= \sum_{i=t_{\text{data}}}^{n} \binom{n}{i} \theta_{\text{season}}^{i} \left(1-\theta_{\text{season}}\right)^{n-i}. \tag{10.109}$$

The resulting parametric test is the same as in Example 10.9.

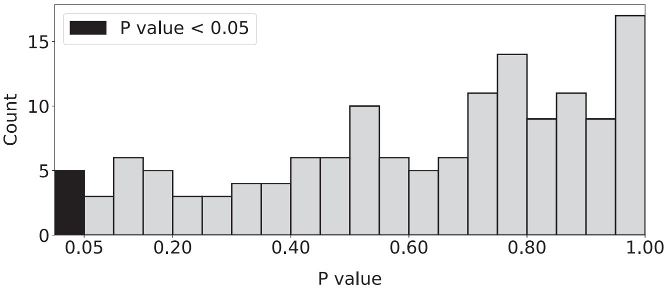
We apply the hypothesis test to the 146 players who shot at least 100 3-pointers during the season, computing the p value of their clutch 3-point shots during the first half of the season. For example, Robert Covington's season percentage was 38.2% and he made 11 out of 15 3-point shots in the clutch, so we set $\theta_{\text{season}} := 0.382$, $n := 15$ and $t_{\text{data}} := 11$ in (10.109) to obtain a p value of 0.006. A histogram of all the p values is shown at the top of Figure 10.11. Only five players (including Covington), have p values below 0.05. They are listed at the bottom of Figure 10.11.

If we were sports journalists, we would go ahead and write an article about how Covington is clutch because he used to tirelessly shoot 3-pointers in the snow as a kid. Instead, let us use some held-out data to double check whether what we are seeing is real. We compute the clutch 3-point percentages for our five clutch players during the second half of the season. Covington and Butler actually shot worse in the clutch than their season percentages! The remaining players shot better, but all the p values are much higher than 0.05. From the second-half data, we cannot conclude that any of the players are actually clutch.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

Example 10.29 illustrates the key challenge of multiple testing: the probability that a single player overperforms by chance in the clutch is low, but the probability that *one out of the 146* players overperforms by chance is very high. Because of this, a p value below the significance level is not necessarily compelling evidence against the null hypothesis. This is reflected in the distribution of the p values, shown in the histogram at the top of Figure 10.11. Every interval of length 0.05 between 0 and 1 contains at least three p values. It is therefore not surprising that five of them end up in $[0, 0.05]$ by sheer luck.

In Theorem 10.13 we show that under a simple null hypothesis, the p value of



| | | First half | | Second half | |
|---|---|---|---|---|---|
| | Season % | Clutch % | P value | Clutch % | P value |
| Rob. Covington | 38.2 | 73.3 (11/15) | 0.006 | 31.8 (7/22) | 0.796 |
| Nikola Mirotic | 34.1 | 62.5 (10/16) | 0.019 | 37.5 (6/16) | 0.478 |
| Caron Butler | 32.1 | 61.5 (8/13) | 0.027 | 25.0 (2/8) | 0.783 |
| Mike Conley | 39.2 | 60.9 (14/23) | 0.029 | 50.0 (8/16) | 0.262 |
| Kirk Hinrich | 31.7 | 52.4 (11/21) | 0.039 | 37.5 (3/8) | 0.491 |

**Figure 10.11 3-point shooting in the clutch.** The graph at the top shows a histogram of the p values corresponding to the 3-point clutch shooting of 146 NBA players in the first half of the 2014/2015 season, computed as described in Example 10.29. The five players with a p value below 0.05 are listed in the table below. The table shows their season 3-point percentage, their clutch 3-point shooting in the first and second half of the season, and the associated p values.

a continuous test statistic has a uniform distribution. If we perform a single test, that is good news: the p value falls in the interval $[0, 0.05]$ (assuming a significance level of 0.05), resulting in a false positive, only 5% of the time. However, if we perform a large number of tests, then approximately 5% end up in $[0, 0.05]$ and result in false positives.

In multiple testing a small p value is not convincing evidence, but a *very small* p value can be. It just needs to be so small that it is unlikely to be observed *even if we account for the number of simultaneous tests.* Bonferroni's correction provides a threshold to decide when this is the case.

**Definition 10.30** (Bonferroni's correction)**.** *When performing $k$ simultaneous hypothesis tests, Bonferroni's correction consists of rejecting the null hypothesis for each test when the p value is smaller or equal to $\alpha/k$, where $\alpha$ is the desired significance level.*

Imagine again that we perform $k$ independent hypothesis tests with significance level $\alpha$, where the probability of a false positive for each test is $\alpha$. If we apply Bonferroni's correction, the probability of incurring at least one false positive is

$$1 - \mathrm{P}\left(\text{No false positives}\right) = 1 - \left(1 - \frac{\alpha}{k}\right)^{k}. \tag{10.110}$$

For $\alpha := 0.05$ and $k := 100$, the probability equals $0.049 \leq \alpha$, so the correction allows us to keep the false-positive rate below the significance level. In fact, the correction allows us to control the false-positive rate for any values of $\alpha$ and $k$. This follows from the union bound in Theorem 9.19.

**Theorem 10.31** (Bonferroni's correction works)**.** *If we perform Bonferroni's correction when performing $k$ hypothesis tests, then the probability of at least one false positive occurring is bounded by the significance level $\alpha$.*

*Proof* By the union bound in Theorem 9.19,

$$\begin{aligned} \mathrm{P}\left(\text{At least one false positive}\right) &= \mathrm{P}\left(\cup_{i=1}^{k} \text{False positive in test } i\right) && (10.111) \\ &\leq \sum_{i=1}^{k} \mathrm{P}\left(\text{False positive in test } i\right) && (10.112) \\ &\leq k \cdot \frac{\alpha}{k} = \alpha. && (10.113) \end{aligned}$$

■

If we apply Bonferroni's correction to the $k := 146$ hypothesis tests in Example 10.29, then the threshold corresponding to a significance level of 0.05 equals $3.42 \cdot 10^{-4}$. All observed p values are larger than the threshold, so we don't reject the null hypothesis for any of them. The correction saves us from the false positives.

The problem with Bonferroni's correction is that it can result in a very small threshold for the p value, which can dramatically reduce our power to detect true positives. This is the main challenge of multiple testing: we can adapt our testing strategy to control the probability of false positives, but this comes at the expense of increasing the false negatives.

**Example 10.32** (Evaluating NBA players)**.** Measuring the impact of individual players is challenging in any team sport. In this example, we consider the problem of quantifying the *added value* that a specific player provides to a team without using traditional statistics such as points, rebounds, assists, etc. Our strategy is to compare the performance of the team with and without the player. To this end, we use Dataset 17, which consists of box scores of NBA regular-season games between 2012 and 2018.

For each player, we compute the point differential of their team in the games with and without the player. We then apply hypothesis testing to evaluate whether the difference is statistically significant. Our null hypothesis is that the point differentials from all games belong to the same distribution. Our alternative hypothesis is that the point differential with the player is higher on average than the point differential without the player. Therefore, we choose the difference of means as the test statistic

$$t_{\text{data}} := m_{\text{with}} - m_{\text{without}}, \tag{10.114}$$

where $m_{\text{with}}$ and $m_{\text{without}}$ denote the sample mean of point differentials with and without the player, respectively. We apply this procedure separately for each team that the player played for in their career. For example, during Lebron James's second stint with the Cleveland Cavaliers, the Cavs played 301 games with James and 27 games without him. The difference between the mean point differentials in those games is $t_{\text{data}} = 16.6$, which suggests that he was very valuable to the Cavs. As a comparison, the test statistic for his time with the Miami Heat is just $t_{\text{data}} = 3.7$.

It is not obvious how to model the distribution of the point differential under the null hypothesis, which motivates the use of a nonparametric test. We apply the permutation test based on the Monte Carlo method from Definition 10.23 to compute the p value. For James's time with the Cavs, the p value is less than $10^{-7}$. In contrast, the p value for his time with the Heat is 0.180. This does not mean that he was not valuable to the Heat (we are pretty sure that he was!), just that we do not have enough data to contradict the null hypothesis (he just missed 9 games).

In total, we apply this procedure to 1,397 player/team pairs. The left column of Figure 10.12 shows the top 20 players according to our test statistic. Surprisingly, they are mostly role players who play very few minutes. It seems quite plausible that most of them have a large test statistic by random chance. If we perform Bonferroni's correction, the threshold to reject the null hypothesis in order to ensure a significance level of 0.05 is $3.58 \cdot 10^{-5}$. Most of the players in the top 20 have p values above the threshold, but below the uncorrected threshold 0.05. This illustrates the effectiveness of Bonferroni's correction in avoiding false positives.

The right column of Figure 10.12 shows the 20 player/team pairs with the smallest p values. The list looks much more meaningful than the top 20 on its left; it includes many well known stars. However, there are only eight that are below the Bonferroni threshold. This illustrates the problem with Bonferroni's correction. The threshold ensures that we avoid false positives with high probability, but it also hinders the detection of true positives. Here we incur seven clear false negatives: Kevin Durant, Chris Paul, Stephen Curry, Anthony Davis, Marc Gasol, Kawhi Leonard and Klay Thompson. In case you don't follow the NBA, there is no doubt that these players are crucial for their respective teams, but their p values are just above the Bonferroni threshold.

The histogram at the top of Figure 10.13 shows the p values of all 1,397 player/team pairs. We can interpret the distribution as a mixture of *alternative-hypothesis players*, who are truly important to their team, and *null-hypothesis players*, who are not. We expect the null-hypothesis players to be evenly spaced out in the unit interval by Theorem 10.13. Indeed, in the histogram we observe a *floor* of around 10 players in every interval of length 0.01, which probably corresponds to null-hypothesis players. In contrast, we expect the alternative-hypothesis players to be concentrated close to zero, at least if there are enough data points. If there are not, as in the case of LeBron James in the Miami Heat, then they will be lost in the sea of null-hypothesis players. Bonferroni's hypothesis tells us how much we should zoom in close to zero in order to leave behind the null-hypothesis players with high probability.

Top 20 mean point differential

| | Mean point diff. | P value | Mins per game |
|---|---|---|---|
| Marcus Paige (CHA) | 28.5 | $2 \cdot 10^{-4}$ | 5.4 |
| N. Mohammed (OKC) | 18.5 | $3 \cdot 10^{-3}$ | 4.0 |
| Georges Niang (UTA) | 17.1 | $2 \cdot 10^{-4}$ | 3.7 |
| **L. James (CLE)** | 16.7 | $< 10^{-7}$ | 36.6 |
| A. Goudelock (HOU) | 16.5 | $3 \cdot 10^{-2}$ | 6.4 |
| **B. Caboclo (TOR)** | 16.4 | $< 10^{-7}$ | 4.6 |
| Roy Hibbert (DEN) | 16.1 | $3 \cdot 10^{-3}$ | 2.0 |
| Brandon Knight (DET) | 16.1 | $2 \cdot 10^{-3}$ | 31.5 |
| Michael Gbinije (DET) | 15.8 | $5 \cdot 10^{-3}$ | 3.4 |
| DeMarre Carroll (BKN) | 15.7 | $2 \cdot 10^{-3}$ | 29.9 |
| Enes Kanter (NY) | 15.5 | $4 \cdot 10^{-3}$ | 25.8 |
| MarShon Brooks (GS) | 15.4 | $5 \cdot 10^{-2}$ | 2.4 |
| Victor Oladipo (IND) | 15.0 | $2 \cdot 10^{-3}$ | 34.1 |
| Ronnie Brewer (OKC) | 13.7 | $2 \cdot 10^{-3}$ | 10.1 |
| J. Cunningham (ATL) | 12.5 | $3 \cdot 10^{-2}$ | 4.6 |
| Steve Novak (MIL) | 11.9 | $4 \cdot 10^{-3}$ | 3.9 |
| Ronnie Brewer (NY) | 10.7 | $5 \cdot 10^{-2}$ | 15.4 |
| Jonas Jerebko (UTA) | 10.5 | $9 \cdot 10^{-2}$ | 15.3 |
| Eric Maynor (OKC) | 10.5 | $2 \cdot 10^{-3}$ | 10.6 |
| A. McKinnie (TOR) | 10.3 | $2 \cdot 10^{-3}$ | 3.9 |

Bottom 20 p values

| | Mean point diff. | P value | Mins per game |
|---|---|---|---|
| **L. James (CLE)** | 16.7 | $< 10^{-7}$ | 36.6 |
| **B. Caboclo (TOR)** | 16.4 | $< 10^{-7}$ | 4.6 |
| **N. Mirotic (CHI)** | 10.3 | $3 \cdot 10^{-7}$ | 23.1 |
| **C. Anthony (NY)** | 8.1 | $5 \cdot 10^{-7}$ | 36.3 |
| **Ricky Rubio (MIN)** | 7.6 | $7 \cdot 10^{-7}$ | 31.4 |
| **James Jones (MIA)** | 8.2 | $6 \cdot 10^{-6}$ | 7.8 |
| **Brandon Rush (GS)** | 6.7 | $6 \cdot 10^{-6}$ | 12.6 |
| **Joel Embiid (PHI)** | 8.7 | $2 \cdot 10^{-5}$ | 28.7 |
| Kevin Durant (OKC) | 6.9 | $1 \cdot 10^{-4}$ | 37.3 |
| Kevin Garnett (MIN) | 9.2 | $2 \cdot 10^{-4}$ | 15.3 |
| Marcus Paige (CHA) | 28.5 | $2 \cdot 10^{-4}$ | 5.4 |
| Georges Niang (UTA) | 17.1 | $2 \cdot 10^{-4}$ | 3.7 |
| Chris Paul (LAC) | 6.8 | $2 \cdot 10^{-4}$ | 33.6 |
| Stephen Curry (GS) | 8.2 | $3 \cdot 10^{-4}$ | 34.6 |
| Anthony Davis (NO) | 5.1 | $4 \cdot 10^{-4}$ | 34.8 |
| Marc Gasol (MEM) | 5.5 | $5 \cdot 10^{-4}$ | 33.9 |
| DeMarre Carroll (ATL) | 10.1 | $5 \cdot 10^{-4}$ | 31.5 |
| Kawhi Leonard (SA) | 4.7 | $6 \cdot 10^{-4}$ | 31.6 |
| Nikola Pekovic (MIN) | 5.0 | $8 \cdot 10^{-4}$ | 28.7 |
| Klay Thompson (GS) | 10.0 | $9 \cdot 10^{-4}$ | 34.1 |

**Figure 10.12 Evaluating NBA players.** The left list shows the top 20 out of 1,397 NBA players between 2012 and 2018 according to the mean point differential in games where their team played with and without them. Most of the players played very few minutes per game, which suggests that they are not actually important to their teams. The right list shows the bottom 20 p values computed via a permutation test. Most of the players are well-known stars who play a lot of minutes. Players with p values below the Bonferroni threshold are highlighted in bold.

The histogram at the bottom of Figure 10.13 shows p values in the interval $[0, 0.01]$. We observe that the floor of evenly-spaced p values is quite sparse and there is a clear concentration of p values close to zero, which probably correspond mostly to alternative-hypothesis players. The Bonferroni threshold captures only a subset of them.

Example 10.32 showcases the difficulty of multiple testing: a small threshold misses some alternative-hypothesis players, but a higher threshold would increase the probability of including null-hypothesis players. In practice, it is often advisable to allow for some false positives, in order to detect more true positives. This can be achieved by thresholding the p values using rules that are less strict than Bonferroni's correction (e.g. (Benjamini and Hochberg, 1995)).

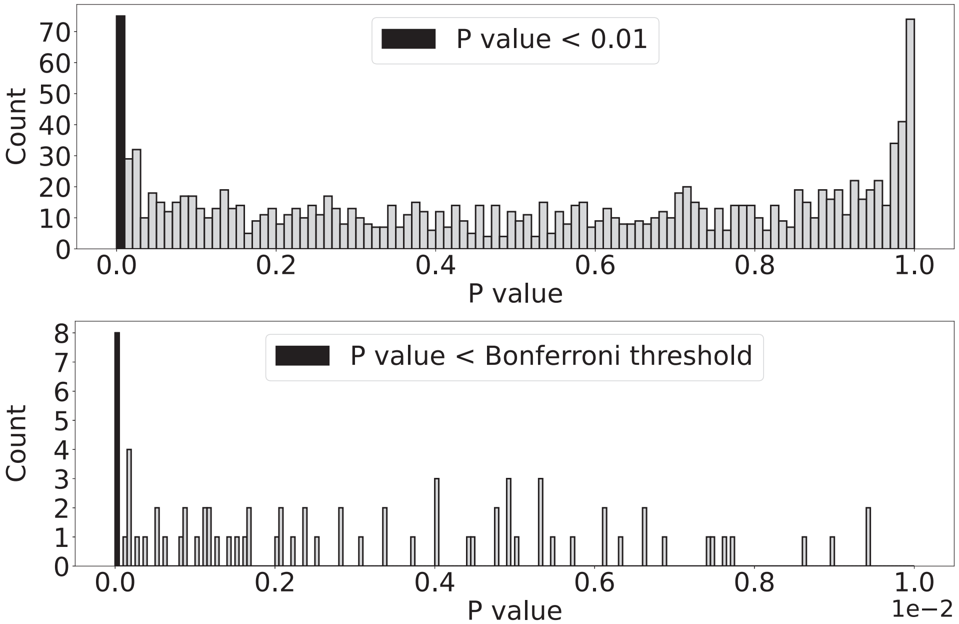


**Figure 10.13 P value distribution in multiple testing.** The histogram at the top shows the p values of all 1,397 player/team pairs in Example 10.32. The histogram has a *floor* of p values corresponding to players who are either not important to their teams, or for whom not enough data is available. It seems clear that a fraction of the smallest p values are due to such players. The histogram at the bottom shows the p values smaller than 0.01, corresponding to the black bar in the histogram above. At this higher resolution, we see a clearer separation between a group of players with very small p values and the rest, but the separation is not completely clear cut. Bonferroni's correction seems to err on the conservative side, only detecting a subset of the smallest p values (represented by the black bar).

## 10.9 Hypothesis Testing And Causal Inference

An important consideration when performing hypothesis testing is whether the results can be interpreted in terms of causal effects. In Example 10.2, our motivation to analyze Antetokounmpo's free-throw shooting is that fans were loudly chanting while he shot free throws in away games. The hypothesis test in Example 10.11 establishes that there is indeed a statistically significant difference in free-throw percentage between home and away games; the data strongly support that

$$p_{\tilde{y}\,|\,\tilde{t}}(1\,|\,0) > p_{\tilde{y}\,|\,\tilde{t}}(1\,|\,1), \tag{10.115}$$

where $\tilde{y}$ and $\tilde{t}$ are random variables that represent a free throw ($\tilde{y} = 1$ if the free throw is made) and whether fans were taunting Antetokounmpo ($\tilde{t} = 1$) or not ($\tilde{t} = 0$). Does this imply that the taunts caused the difference in free-throw percentage? No! Statistical significance does *not* imply the existence of a causal effect. As explained in Section 4.6, the difference could be due by confounding factors. For instance, he could have been more tired at away games due to traveling.

To further emphasize the distinction between statistical significance and causal effects, let us delve deeper into a surprising result in Example 10.32, where we evaluate NBA players by comparing the games played by their teams with and without them. Bruno Caboclo from the Toronto Raptors has an impressive difference between mean point differentials (16.4, sixth overall). However, he played less than five minutes per game, which makes it unlikely that his presence could have been that impactful. Surprisingly, the difference is highly statistically significant, even if we account for multiple testing: the corresponding p value is the second smallest overall! There is overwhelming evidence that

$$\mu_{\tilde{y}\,|\,\tilde{t}}(1) > \mu_{\tilde{y}\,|\,\tilde{t}}(0), \tag{10.116}$$

where $\mu_{\tilde{y}\,|\,\tilde{t}}$ is the conditional mean of the point differential represented by the random variable $\tilde{y}$ when Caboclo plays ($\tilde{t} = 1$) or not ($\tilde{t} = 0$). However, this does *not* imply that his presence *causes* an increase in point differential.

In fact, Caboclo played only 24 games for the Raptors over four years (missing more than 200). Since the Raptors won these games by a very large margin on average and Caboclo played very few minutes, it seems plausible that he was being included in games, only when they were clearly decided in favor of the Raptors. In other words, the Raptors were not winning by a lot because Caboclo played, but rather *Caboclo played because the Raptors were winning by a lot.*

As explained in Sections 4.6.3 and 7.9, randomized data acquisition guarantee that differences in conditional probabilities or conditional means reflect causal effects. In such situations, the results of hypothesis tests *can* be interpreted causally. In Example 4.25 we describe the randomized controlled trial used to evaluate Pfizer's COVID-19 vaccine. Randomization and hypothesis testing play complementary roles in the analysis of the data. Hypothesis testing allows us to determine that there is a (statistically significant) difference between the control and vaccine groups. Randomization allows us to conclude that the difference reflects a causal effect.

**Example 10.33** (COVID-19 vaccine: P value)**.** We apply the two-sample z test to the data in Example 4.25, choosing the one-tailed version of the test because we are not interested in detecting a situation where the vaccine results in more COVID-19 cases. The null hypothesis is that the probability of contracting COVID-19, represented by the parameter $\theta_{\text{null}}$ is the same in the control and the vaccine group. The test statistic is the fraction of positives for the vaccine group minus the fraction of positives in the control group,

$$t_{\text{data}} = \frac{162}{21728} - \frac{8}{21720} = 7.09 \cdot 10^{-3}. \tag{10.117}$$

Setting $n_A := 21728$, $k_A := 162$, $n_B := 21720$, $k_B := 8$ in Definition 10.10, we obtain $\sigma_{\text{null}} = 5.99 \cdot 10^{-4}$, which yields a p value equal to

$$\text{pv}(t_{\text{data}}) = 1 - F_{\tilde{z}}\left(\frac{7.09 \cdot 10^{-3}}{5.99 \cdot 10^{-4}}\right) < 10^{-23}. \qquad (10.118)$$

The p value is extremely small, so the data are completely inconsistent with the null hypothesis. The vaccine works.

In non-medical applications, the combination of randomized experiments and hypothesis testing is called A/B testing. It is often applied to evaluate user reactions.

**Definition 10.34** (A/B testing)**.** *The goal of A/B testing is to evaluate the causal effect of different options in the design of a product. Users are randomly assigned to a control group (A) or a treatment group (B), where they are exposed to two versions of the product, with and without a specific modification. Hypothesis testing is then applied to determine whether the difference in user response, measured by an appropriate metric, is statistically significant.*

In the Internet era, A/B testing has become a fundamental tool in the tech industry and beyond. Famously, it played a key role in Barack Obama's presidential campaigns.

**Example 10.35** (Obama's presidential campaign)**.** Data scientists working for the Obama campaign used A/B testing to design the campaign website. Specifically, they ran A/B testing to determine what images or videos would be more effective in getting viewers to sign up for the campaign mailing list (Siroker, 2010). For simplicity, we focus on the comparison between images and videos. The sign-up rate for the images was 14,016 out of 155,280 users (9.03%). The sign-up rate for the videos was 10,337 out of 155,102 (6.66%). Given that the users were randomly assigned to the two groups, we can safely assume that any difference in the sign-up rate is due to the choice between images and videos. We just need to determine that the difference is not due to random chance. To this end, we apply the two-tailed version of the two-sample z test in Definition 10.10.

The null hypothesis is that the sign-up rate $\theta_{\text{null}}$ is the same for all users. The test statistic is the absolute value of the difference between sign-up rates,

$$t_{\text{data}} = \left|\frac{14016}{155280} - \frac{10337}{155102}\right| = 2.36 \cdot 10^{-2}. \qquad (10.119)$$

Setting $n_A := 155280$, $k_A := 14016$, $n_B := 155102$, $k_B := 10337$ in Definition 10.10, we obtain $\sigma_{\text{null}} = 9.65 \cdot 10^{-4}$, which yields a p value equal to

$$\text{pv}(t_{\text{data}}) = 2\left(1 - F_{\tilde{z}}\left(\frac{2.36 \cdot 10^{-2}}{9.65 \cdot 10^{-4}}\right)\right) < 10^{-80}. \qquad (10.120)$$

The p value is minuscule, which is typical for A/B testing in online settings, where the number of subjects is often order of magnitudes higher than in other domains. As we discuss in Section 10.10.1, this can be a mixed blessing, because tiny differences that are not practically meaningful can be statistically significant.

## 10.10 P-Value Abuse

In many scientific domains, reporting a p value below a predefined significance level (usually 0.05) is a requisite for publication. As a result, statistical significance is often interpreted as a *stamp of approval* for scientific results. This is problematic for several reasons. First, as described in Section 10.9, statistical significance does not provide insight into causal effects in observational studies where data acquisition is not randomized. In such cases, any claim implying causality must be justified through additional analysis. Second, statistical significance does not necessarily imply practical significance, as we explain in Section 10.10.1. Third, there is a strong incentive to achieve artificially small p values by cherry-picking results, as described in Section 10.10.2. Consequently, statistical significance should not be the end goal of a scientific study, but rather a sanity check complementing a more comprehensive analysis.

### *10.10.1 Practical Significance*

Statistical significance does not automatically imply *practical significance*. A result is practically significant if it is meaningful or useful for the application of interest. A vaccine that works 1% of the time is not very useful. A change in a website that increases the click-through rate by 0.001% is probably not worth it. Practical significance cannot be directly assessed using p values, because they only quantify to what extent the observed data is inconsistent with the null hypothesis.

**Example 10.36** (Statistical vs. practical significance)**.** We consider a fictional scenario where two drugs for a certain disease are being evaluated via randomized controlled trials. Both drugs produce side effects and are expensive, so medical experts have determined that they should produce an increase of at least 5% in the cure rate (the probability that a patient recovers) in order to be approved.

In the first trial, 52 out of 100 patients in the treatment group and 30 out of 100 patients in the control group are cured. We apply the one-tailed two-sample z test in Definition 10.10 to determine whether the result is statistically significant. The null hypothesis is that the control and treatment groups have the same distribution. The test statistic is the difference in the cure rate in the two groups, which is the fraction of cured subjects. Setting $n_A := 100$, $k_A := 52$, $n_B := 100$, $k_B := 30$, we obtain $t_{\text{data}} = 0.22$ and $\sigma_{\text{null}} = 6.96 \cdot 10^{-2}$, so the p value equals

$$\text{pv}(t_{\text{data}}) = 1 - F_{\tilde{z}}\left(\frac{0.22}{6.96 \cdot 10^{-2}}\right) = 7.8 \cdot 10^{-4}. \tag{10.121}$$

In the second trial, 30,650 out of 100,000 patients in the treatment group and

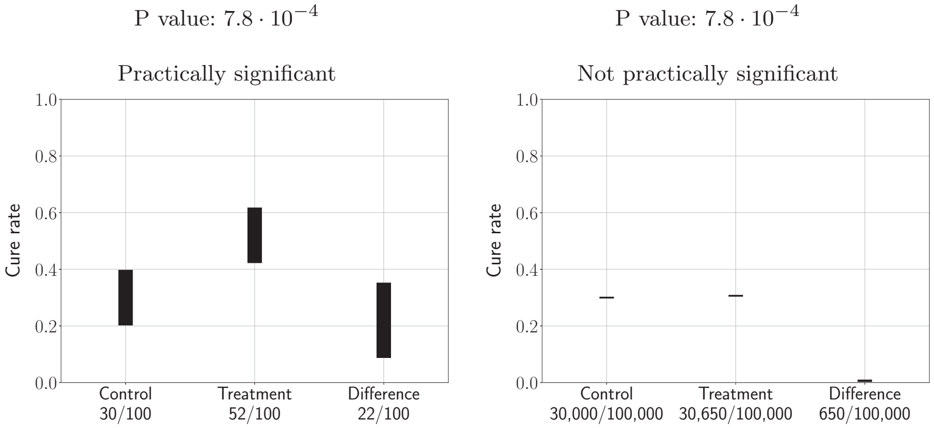


**Figure 10.14 Statistical vs. practical significance.** The two fictional datasets in Example 10.36 have the same p value. They are both statistically significant, but only the trial on the right is practically significant. The difference is obvious in the confidence intervals for the cure rate in the control and treatment groups. The improvement in the cure rate is much larger for the trial on the left. The trial on the right has the same p value because of the large number of subjects, which renders the tiny improvement statistically significant. The confidence intervals are built following are built following Definition 9.44 and Exercise 9.9.

30,000 out of 100,000 patients in the control group are cured. In this case, the difference in the cure rate is minuscule: $t_{\text{data}} = 6.5 \cdot 10^{-3}$. However, the trial has so many subjects that the standard deviation of the test statistic under the null hypothesis is also very small: $\sigma_{\text{null}} = 2.06 \cdot 10^{-3}$. As a result, the p value is the same as in the first study

$$\text{pv}(t_{\text{data}}) = 1 - F_{\tilde{z}}\left(\frac{6.5 \cdot 10^{-3}}{2.06 \cdot 10^{-3}}\right) = 7.8 \cdot 10^{-4}. \tag{10.122}$$

The p value is very small for both studies, so we can be very certain that both drugs increase the cure rate with respect to the control group. However, the p value tells us *nothing* about the actual difference in the cure rate, which is what determines the practical significance of the results.

To quantify the practical significance of each trial, we compute 0.95 confidence intervals for the difference in cure rate between the control and treatment groups, as explained in Exercise 9.9. The confidence intervals, shown in Figure 10.14, equal $[8.71\%, 35.3\%]$ for the first drug and $[0.25\%, 1.05\%]$ for the second drug. The difference is positive in both trials, but it is only practically significant (above 5%) for the first drug. Despite having exactly the same p value, the practical significance of the two results is very different.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

Example 10.36 shows that tiny differences between two groups can be sta-

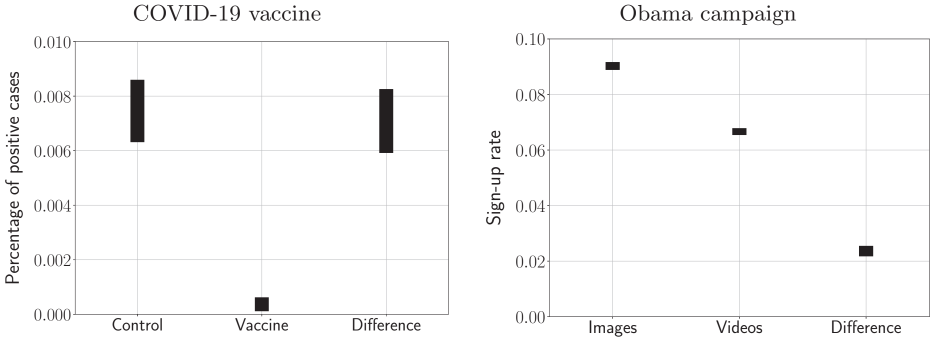


**Figure 10.15 Confidence intervals establish practical significance.** 0.95 confidence intervals for the data in Examples 10.33 and 10.35. In both cases, there is clearly a substantial difference between the two groups, so the results are practically significant as well as statistically significant. The confidence intervals are built following Definition 9.44 and Exercise 9.9.

tistically significant when data is very plentiful. This can easily occur in large randomized controlled trials and in online A/B tests. In Example 10.33, imagine there were 120 positive cases in the vaccine group (instead of 8). The corresponding p value would be 0.006, so the difference would still be statistically significant at a significance level of 0.05 (or even 0.01), but the ratio between the positive cases in the vaccine and in the control group would only be 3/4 (in the real data it is 1/20). Such a modest reduction in the infection rate might not be worth the cost of producing and administering the vaccine.

Similarly, if in Example 10.35 the users that sign up after viewing the videos were 13,650 (instead of 10,337), then the p value for the hypothesis test would be 0.027. This difference is statistically significant for the typical significance level of 0.05. However, the observed difference in sign-up rate would just be 0.2%, which is probably negligible in practice.

These examples hopefully make it clear that small p values are not sufficient evidence to establish practical significance. Figure 10.15 uses confidence intervals to show that the results in Examples 10.33 and 10.35 are indeed practically significant.

### *10.10.2 Publication Bias And P-Hacking*

P values are often interpreted as a guarantee that a published result is meaningful. Unfortunately, this is not necessarily warranted. Imagine that 100 studies are performed by different research groups around the world to test whether eating pizza cures COVID-19. This is a multiple testing scenario, where many hypothesis tests are carried out simultaneously. As explained in Section 10.8, if the null hypothesis holds for the 100 tests, then around 5 of them will have p values below 0.05 by random chance. If all of these results were made public, this would not be

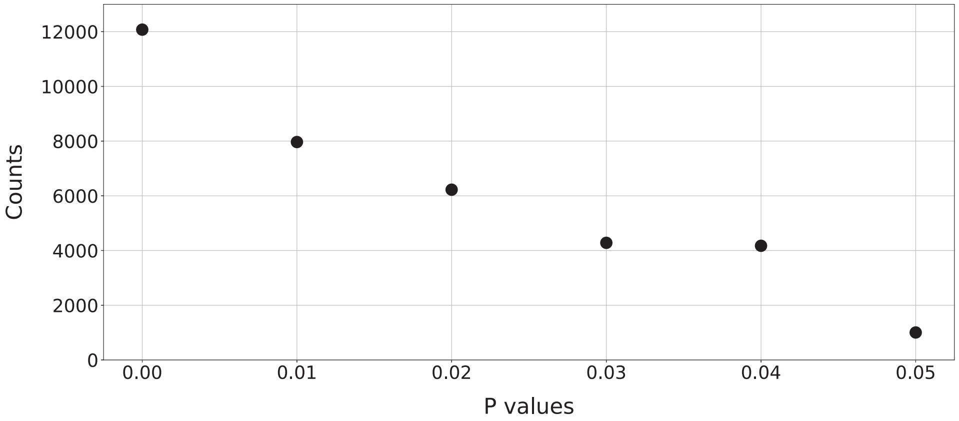


**Figure 10.16 Evidence of p-hacking.** Histogram of p values reported in the abstracts of open-access articles in PubMed, a popular biomedical database, collected by the authors of (Head et al., 2015). The p values are rounded to two decimal places, in order to account for the different precisions at which they are reported. There is a smooth decreasing trend in the p values, which is to be expected if most results are truly significant. However, at 0.04 the trend is broken: the number of p values is almost the same as for 0.03. It seems likely that p-hacking is artificially inflating the number of p values just below the standard significance threshold of 0.05.

a big deal. Unfortunately, positive findings are much more likely to be published than negative results. This is known as *publication bias* and it is somewhat understandable: in the null hypothesis framework, a negative result just indicates that there is not sufficient evidence against the null hypothesis, *not that it holds*. Therefore it may be difficult to communicate effectively. The headline *Pizza cures COVID-19!* is much more attractive than *There is no conclusive evidence that pizza cures COVID-19...*

Beyond publication bias, the acceptance of statistical significance as a sufficient condition for publication is problematic because it may motivate researchers to cherry-pick their results. For instance, consider a researcher trying to determine whether different food additives are harmful. They test thousands of additives by feeding them to mice and recording any adverse effects. One of the additives results in a small p value, but not small enough to be statistically significant after applying Bonferroni's correction. The researcher has two choices: (1) Gather additional data and perform another hypothesis test focused on the detected additive. (2) Publish the result without mentioning that thousands of additives were tested. This known as *p-hacking*, because it is a fraudulent manipulation of the hypothesis-testing framework. Unfortunately, the researcher may have strong incentives to prefer option 2 (e.g. publishing a high-impact paper, as opposed to delaying their PhD graduation by another year).

In order to study the prevalence of p-hacking in practice, we examine the distribution of p values reported in the abstracts of open-access articles in PubMed, a popular biomedical database, extracted from Dataset 18. We round the p values to two decimal places, in order to account for the different precisions at which they are reported. We observe a clear decreasing trend: 12,074 abstracts report p values in $[0, 0.005)$, 7,966 in $[0.005, 0.015)$, 6,224 in $[0.015, 0.025)$, and 4,281 in $[0.025, 0.035)$. This is expected: for truly significant results, the p values should be heavily skewed towards very small values (see the p value distribution for $\theta_{\text{away}} := 0.55$ in Figure 10.6 or the histograms in Figure 10.13). However, just below the standard threshold for statistical significance the trend is broken: the p values reported in $[0.035, 0.045)$ are 4,173, almost the same as in $[0.025, 0.035)$. This is very suspicious: 0.05 is an arbitrary value, except for the fact that it is used as a significance threshold. P-hacking is a likely explanation for the inflated number of p values just below the threshold.