Chapter 11

Hypothesis testing

In a medical study we observe that 10% of the women and 12.5% of the men suffer from heart disease. If there are 20 people in the study, we would probably be hesitant to declare that women are less prone to suffer from heart disease than men; it is very possible that the results occurred by chance. However, if there are 20,000 people in the study, then it seems more likely that we are observing a real phenomenon. Hypothesis testing makes this intuition precise; it is a framework that allows us to decide whether patterns that we observe in our data are likely to be the result of random fluctuations or not.

11.1 The hypothesis-testing framework

The aim of hypothesis testing is to evaluate a predefined conjecture. In the example above, this could be that heart disease is more prevalent in men than in women. The hypothesis that our conjecture is false is called the **null hypothesis**, denoted by H_0. In our example, the null hypothesis would be that heart disease is at least as prevalent in men as in women. If the null hypothesis holds, then whatever pattern we are detecting in our data that seems to support our conjecture is just a fluke. There just happen to be a lot of men with heart disease (or women without) in the study. In contrast, the hypothesis under which our conjecture is true is known as the **alternative hypothesis**, denoted by H_1. In this chapter we take a frequentist perspective: *the hypotheses either hold or not*, they are *not* modeled probabilistically.

A **test** is a procedure to determine whether we should *reject* the null hypothesis or not based on the data. Rejecting the null hypothesis means that we consider unlikely that it happened, which is evidence in favor of the alternative hypothesis. If we fail to reject the null hypothesis, this does *not* mean that we consider it likely, we just don't have enough information to discard it. Most tests produce a decision by thresholding a **test statistic**, which is a function that maps the data (i.e. a vector in $\mathbb{R}^n$) to a single number. The test rejects the null hypothesis if the test statistic belongs to a **rejection region** $\mathcal{R}$. For example, we could have

$$\mathcal{R} := \{t \mid t \geq \eta\}, \tag{11.1}$$

where t is the test statistic computed from the data and η is a predefined threshold. In this case, we would reject the null hypothesis only if t is larger than η.

As shown in Table 11.1, there are two possible errors that we can make. A **Type I error** is a *false positive*: our conjecture is false, but we reject the null hypothesis. A **Type II error** is

	Reject H_0?	
	No	Yes
H_0 is true	☺	Type I error
H_1 is true	Type II error	☺

Table 11.1: Type I and II errors.

a *false negative*: our conjecture holds, but we do not reject the null hypothesis. In hypothesis testing, our priority is to control Type I errors. When you read in a study that a result is **statistically significant** at a level of 0.05, this means that the probability of committing a Type I error is bounded by 5%.

Definition 11.1.1 (Significance level and size)**.** *The size of a test is the probability of making a Type I error. The significance level of a test is an upper bound on the size.*

Rejecting the null hypothesis does not give a quantitative sense of the extent to which the data are incompatible with the null hypothesis. The **p value** is a function of the data that plays this role.

Definition 11.1.2 (p value)**.** *The p value is the smallest significance level at which we would reject the null hypothesis for the data we observe.*

For a fixed significance level, it is desirable to select a test that minimizes the probability of making a Type II error. Equivalently, we would like to maximize the probability of rejecting the null hypothesis when it does not hold. This probability is known as the **power** of the test.

Definition 11.1.3 (Power)**.** *The power of a test is the probability of rejecting the null hypothesis if it does not hold.*

Note that in order to characterize the power of a test we need to know the distribution of the data under the alternative hypothesis, which is often unrealistic (recall that the alternative hypothesis is just the complement of the null hypothesis and consequently encompasses many different possibilities).

The standard procedure to apply hypothesis testing in the applied sciences is the following:

1. Choose a conjecture.
2. Determine the corresponding null hypothesis.
3. Choose a test.
4. Gather the data.
5. Compute the test statistic from the data.

6. Compute the p value and reject the null hypothesis if it is below a predefined limit (typically 1% or 5%).

Example 11.1.4 (Clutch)**.** We want to test the conjecture that a certain player in the NBA is *clutch*, i.e. that he scores more points at the end of close games than during the rest of the game. The null hypothesis is that there is no difference in his performance. The test statistic t that we choose is whether he makes more or less points per minute in the last quarter than in the rest of the game

$$t\left(\vec{x}\right) = \sum_{i=1}^{n} 1_{\vec{x}_i > 0}, \tag{11.2}$$

where $\vec{x}_i$ is the difference between the points per minute he scores in the 4th quarter and in the rest of the quarters of game i for $1 \leq i \leq n$.

The rejection region of the test is of the form

$$\mathcal{R} := \left\{ t\left(\vec{x}\right) \mid t\left(\vec{x}\right) \geq \eta \right\}, \tag{11.3}$$

for a fixed threshold η. Under the null hypothesis the probability of scoring more points per minute in the 4th quarter is 1/2 (for simplicity we ignore the possibility that he scores the same number of points), so we can model the test statistic under the null hypothesis as a binomial random variable with parameters n and 1/2. If η is an integer between 0 and n, then the probability that the test statistic is in the rejection region if the null hypothesis holds is

$$P\left(T_0 \geq \eta\right) = \frac{1}{2^n} \sum_{k=\eta}^{n} \binom{n}{k}. \tag{11.4}$$

So the size of the test is $\frac{1}{2^n} \sum_{k=\eta}^{n} \binom{n}{k}$. Table 11.2 shows this value for all possible values of η. If we want a significance level of 1% or 5% then we need to set the threshold at $\eta = 16$ or $\eta = 15$ respectively.

We gather the data from 20 games $\vec{x}$ and compute the value of the test statistic $t\left(\vec{x}\right)$ (note that we use a lowercase letter because it is a specific realization), which turns out to be 14 (he scores more points per minute in the fourth quarter in 14 of the games). This is not enough to reject the null hypothesis for our predefined level of 1% or 5%. Therefore the result is not statistically significant.

In any case, we compute the p value, which is the smallest level at which the result would have been significant. From the table it is equal to 0.058. Note that under a frequentist framework we *cannot* interpret this as the probability that the null hypothesis holds (i.e. that the player is not better in the fourth quarter) because the hypothesis is not random, it either holds or it doesn't. Our result is almost significant and although we do not have enough evidence to support our conjecture, it does seem plausible that the player performs better in the fourth quarter.

△

11.2 Parametric testing

In this section we discuss hypothesis testing under the assumption that our data are sampled from a known distribution with *unknown* parameters. We again take a frequentist perspective,

η	1	2	3	4	5	6	7	8	9	10
$P(T_0 \geq \eta)$	1.000	1.000	1.000	0.999	0.994	0.979	0.942	0.868	0.748	0.588
η	11	12	13	14	15	16	17	18	19	20
$P(T_0 \geq \eta)$	0.412	0.252	0.132	0.058	0.021	0.006	0.001	0.000	0.000	0.000

Table 11.2: Probability of committing a Type I error depending on the value of the threshold in Example 11.1.4. The values are rounded to three decimal points.

as is usually done in most studies in the applied sciences. The parameter is consequently deterministic and so are the hypotheses: the null hypothesis is true or not, there is no such thing as *the probability that the null hypothesis holds*.

To simplify the exposition, we assume that the probability distribution depends only on one parameter that we denote by θ. P_θ is the probability measure of our probability space if θ is the value of the parameter. $\vec{X}$ is a random vector distributed according to P_θ. The actual data that we observe, which we denote by $\vec{x}$ is assumed to be a realization from this random vector.

Assume that the null hypothesis is $\theta = \theta_0$. In that case, the size of a test with test statistic T and rejection region $\mathcal{R}$ is equal to

$$\alpha = \mathrm{P}_{\theta_0}\left(T(\vec{X}) \in \mathcal{R}\right). \tag{11.5}$$

For a rejection region of the form (11.1) we have

$$\alpha := \mathrm{P}_{\theta_0}\left(T(\vec{X}) \geq \eta\right). \tag{11.6}$$

If the realization of the test statistic is $T(x_1, \ldots, x_n)$ then the significance level at which we would reject H_0 would be

$$p = \mathrm{P}_{\theta_0}\left(T(\vec{X}) \geq T(\vec{x})\right), \tag{11.7}$$

which is the p value if we observe $\vec{x}$. The p value can consequently be interpreted as the probability of observing a result that is *more extreme* than what we observe in the data *if the null hypothesis holds*.

A hypothesis of the form $\theta = \theta_0$ is known as a **simple** hypothesis. If a hypothesis is of the form $\theta \in \mathcal{S}$ for a certain set $\mathcal{S}$ then the hypothesis is **composite**. For a composite null hypothesis $\theta \in \mathcal{H}_0$ we redefine the size and the p value in the following way,

$$\alpha = \sup_{\theta \in \mathcal{H}_0} \mathrm{P}_\theta\left(T(\vec{X}) \geq \eta\right), \tag{11.8}$$

$$p = \sup_{\theta \in \mathcal{H}_0} \mathrm{P}_\theta\left(T(\vec{X}) \geq T(\vec{x})\right). \tag{11.9}$$

In order to characterize the power of the test for a certain significance level, we compute the power function.

Definition 11.2.1 (Power function)**.** *Let P_θ be the probability measure parametrized by θ and let $\mathcal{R}$ the rejection region for a test based on the test statistic $T(\vec{x})$. The power function of the test is defined as*

$$\beta(\theta) := P_\theta\left(T(\vec{X}) \in \mathcal{R}\right) \tag{11.10}$$

Ideally we would like $\beta(\theta) \approx 0$ for $\theta \in \mathcal{H}_0$ and $\beta(\theta) \approx 1$ for $\theta \in \mathcal{H}_1$.

Example 11.2.2 (Coin flip)**.** We are interested in checking whether a coin is biased towards heads. The null hypothesis is that for each coin flip the probability of obtaining heads is $\theta \leq 1/2$. Consequently, the alternative hypothesis is $\theta > 1/2$. Let us consider a test statistic equal to the number of heads observed in a sequence of n iid flips,

$$T(\vec{x}) = \sum_{i=1}^{n} 1_{\vec{x}_i=1}, \tag{11.11}$$

where $\vec{x}_i$ is one if the ith coin flip is heads and zero otherwise. A natural rejection region is

$$T(\vec{x}) \geq \eta. \tag{11.12}$$

In particular, we consider two possible thresholds

1. $\eta = n$, i.e. we only reject the null hypothesis if *all* the coin flips are heads,
2. $\eta = 3n/5$, i.e. we reject the null hypothesis if at least three fifths of the coin flips are heads.

What test should we use if the number of coin flips is 5, 50 or 100? Do the tests have a 5% significance level? What is the power of the tests for these values of n?

To answer these questions, we compute the power function of the test for both options. If $\eta = n$,

$$\beta_1(\theta) = P_\theta\left(T(\vec{X}) \in \mathcal{R}\right) \tag{11.13}$$

$$= \theta^n. \tag{11.14}$$

If $\eta = 3n/5$,

$$\beta_2(\theta) = \sum_{k=3n/5}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k}. \tag{11.15}$$

Figure 11.1 shows the two power functions. If $\eta = n$, then the test has a significance level of 5% for the three values of n. However the power is very low, especially for large n. This makes sense: even if the coin is pretty biased the probability of n heads is extremely low. If $\eta = 3n/5$, then for $n = 5$ the test has a significance level way above 5%, since even if the coin is not biased the probability of observing 3 heads out of 5 flips is quite high. However for large n the test has much higher power than the first option. If the bias of the coin is above 0.7 we reject the null hypothesis with high probability.

△

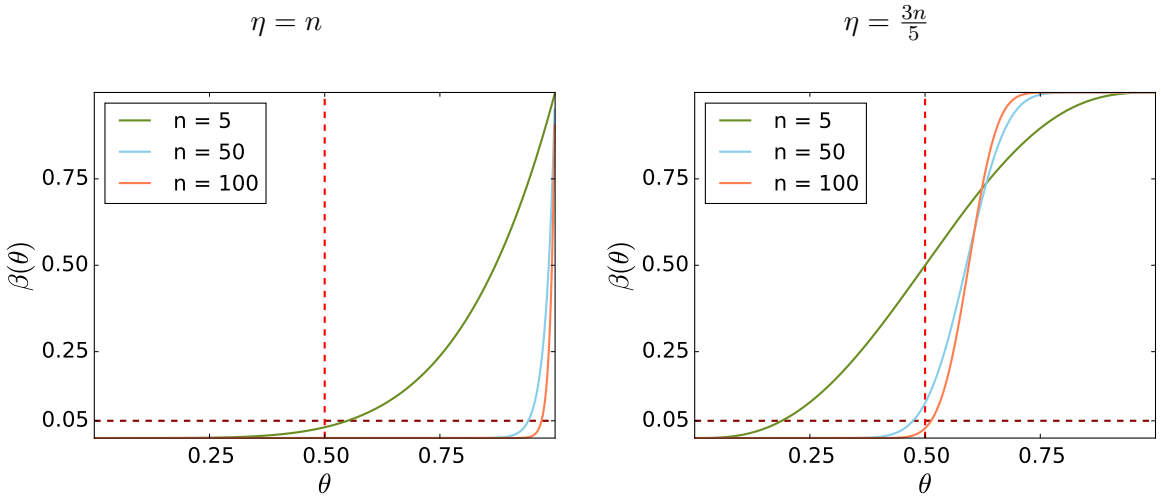


Figure 11.1: Power functions for the tests described in Example 11.2.2.

A systematic method for building tests under parametric assumptions is to threshold the ratio between the likelihood of the data under the null hypothesis and the likelihood of the data under the alternative hypothesis. If this ratio is high, the data are compatible with the null hypothesis, so it should not be rejected.

Definition 11.2.3 (Likelihood-ratio test)**.** *Let $\mathcal{L}_{\vec{x}}(\theta)$ denote the likelihood function corresponding to a data vector $\vec{x}$. $\mathcal{H}_0$ and $\mathcal{H}_1$ are the sets corresponding to the null and alternative hypotheses respectively. The likelihood ratio is*

$$\Lambda(\vec{x}) := \frac{\sup_{\theta \in \mathcal{H}_0} \mathcal{L}_{\vec{x}}(\theta)}{\sup_{\theta \in \mathcal{H}_1} \mathcal{L}_{\vec{x}}(\theta)}. \tag{11.16}$$

A likelihood-ratio test has a rejection region of the form $\{\Lambda(\vec{x}) \leq \eta\}$, for a constant threshold η.

Example 11.2.4 (Gaussian with known variance)**.** Imagine that you have some data that are well modeled as iid Gaussian with a known variance σ. The mean is unknown and we are interested in establishing that it is *not* equal to a certain value μ_0. What is the corresponding likelihood-ratio test and how should be set the threshold so that we have a significance level α?

First, from Example 9.6.4 the sample mean achieves the maximum of the likelihood function of a Gaussian

$$\text{av}(\vec{x}) := \arg\max_{\mu} \mathcal{L}_{\vec{x}}(\mu, \sigma) \tag{11.17}$$

for any value of σ. Using this result, we have

$$\Lambda(\vec{x}) = \frac{\sup_{\mu \in \mathcal{H}_0} \mathcal{L}_{\vec{x}}(\mu)}{\sup_{\mu \in \mathcal{H}_1} \mathcal{L}_{\vec{x}}(\mu)} \tag{11.18}$$

$$= \frac{\mathcal{L}_{\vec{x}}(\mu_0)}{\mathcal{L}_{\vec{x}}(\text{av}(\vec{x}))}. \tag{11.19}$$

Plugging in the expressions for the likelihood we obtain,

$$\Lambda\left(\vec{x}\right) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\left(\vec{x}_i - \text{av}\left(\vec{x}\right)\right)^2 - \left(\vec{x}_i - \mu_0\right)^2\right)\right\} \tag{11.20}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left(-2\,\text{av}\left(\vec{x}\right)\sum_{i=1}^{n}\vec{x}_i + n\,\text{av}\left(\vec{x}\right)^2 - 2\mu_0\sum_{i=1}^{n}\vec{x}_i + n\mu_0^2\right)\right\} \tag{11.21}$$

$$= \exp\left\{-\frac{n\left(\text{av}\left(\vec{x}\right) - \mu_0\right)^2}{2\sigma^2}\right\}. \tag{11.22}$$

Taking logarithms, the test is of the form

$$\frac{\left|\text{av}\left(\vec{x}\right) - \mu_0\right|}{\sigma} \geq \sqrt{\frac{-2\log\eta}{n}}. \tag{11.23}$$

The sample mean of n independent Gaussian random variables with mean μ_0 and variance σ^2 is Gaussian with mean μ_0 and variance σ^2/n, which implies

$$\alpha = \text{P}_{\mu_0}\left(\left|\frac{\text{av}\left(\vec{X}\right) - \mu_0}{\sigma/\sqrt{n}}\right| \geq \sqrt{-2\log\eta}\right) \tag{11.24}$$

$$= 2\,Q\left(\sqrt{-2\log\eta}\right). \tag{11.25}$$

If we fix a desired size α then the test becomes

$$\frac{\left|\text{av}\left(\vec{x}\right) - \mu_0\right|}{\sigma} \geq \frac{Q^{-1}\left(\alpha/2\right)}{\sqrt{n}}. \tag{11.26}$$

△

A motivating argument to employ the likelihood-ratio test is that if the null and alternative hypotheses are simple, then it is optimal in terms of power.

Lemma 11.2.5 (Neyman-Pearson Lemma)**.** *If both the null hypothesis and the alternative hypothesis are simple, i.e. the parameter θ can only have two values θ_0 and θ_1, then the likelihood-ratio test has the highest power among all tests with a fixed size.*

Proof. Recall that the power is the probability of rejecting the null hypothesis if it does not hold. If we denote the rejection region of the likelihood-ratio test by $\mathcal{R}_{LR}$ then its power is

$$P_{\theta_1}\left(\vec{X} \in \mathcal{R}_{LR}\right). \tag{11.27}$$

Assume that we have another test with rejection region $\mathcal{R}$. Its power is equal to

$$P_{\theta_1}\left(\vec{X} \in \mathcal{R}\right). \tag{11.28}$$

To prove that the power of the likelihood-ratio test is larger we only need to establish that

$$P_{\theta_1}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) \geq P_{\theta_1}\left(\vec{X} \in \mathcal{R}_{LR}^c \cap \mathcal{R}\right). \tag{11.29}$$

Let us assume that the data are continuous random variables (the argument for discrete random variables is practically the same) and that the pdf when the null and alternative hypotheses hold are f_{θ_0} and f_{θ_1} respectively. By the definition of the rejection region of the likelihood-ratio test, if $\Lambda(\vec{x}) \in \mathcal{R}_{LR}$

$$f_{\theta_1}(\vec{x}) \geq \frac{f_{\theta_0}(\vec{x})}{\eta}, \tag{11.30}$$

whereas if $\Lambda(\vec{x}) \in \mathcal{R}_{LR}^c$

$$f_{\theta_1}(\vec{x}) \leq \frac{f_{\theta_0}(\vec{x})}{\eta}. \tag{11.31}$$

If both tests have size α then

$$P_{\theta_0}\left(\vec{X} \in \mathcal{R}\right) = \alpha = P_{\theta_0}\left(\vec{X} \in \mathcal{R}_{LR}\right). \tag{11.32}$$

and consequently

$$\begin{aligned} P_{\theta_0}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) &= P_{\theta_0}\left(\vec{X} \in \mathcal{R}_{LR}\right) - P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}\right) && (11.33)\\ &= P_{\theta_0}\left(\vec{X} \in \mathcal{R}\right) - P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}\right) && (11.34)\\ &= P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}^c\right). && (11.35) \end{aligned}$$

Now let us prove that (11.29) holds,

$$\begin{aligned} P_{\theta_1}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) &= \int_{\vec{x} \in \mathcal{R}^c \cap \mathcal{R}_{LR}} f_{\theta_1}(\vec{x})\, \mathrm{d}\vec{x} && (11.36)\\ &\geq \frac{1}{\eta} \int_{\vec{x} \in \mathcal{R}^c \cap \mathcal{R}_{LR}} f_{\theta_0}(\vec{x})\, \mathrm{d}\vec{x} \qquad \text{by (11.30)} && (11.37)\\ &= \frac{1}{\eta} P_{\theta_0}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) && (11.38)\\ &= \frac{1}{\eta} P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}^c\right) \qquad \text{by (11.35)} && (11.39)\\ &= \frac{1}{\eta} \int_{\vec{x} \in \mathcal{R} \cap \mathcal{R}_{LR}^c} f_{\theta_0}(\vec{x})\, \mathrm{d}\vec{x} && (11.40)\\ &\geq \int_{\vec{x} \in \mathcal{R} \cap \mathcal{R}_{LR}^c} f_{\theta_1}(\vec{x})\, \mathrm{d}\vec{x} \qquad \text{by (11.31)} && (11.41)\\ &= P_{\theta_1}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}^c\right). && (11.42) \end{aligned}$$

□

11.3 Nonparametric testing: The permutation test

In practical situations we may not able to design a parametric model that is adequate for our data. Nonparametric tests are hypothesis tests that do not assume that the data follow

any distribution with a predefined form. In this section we describe the permutation test, a nonparametric test that can be used to compare two data sets $\vec{x}_A$ and $\vec{x}_B$ in order to evaluate conjectures of the form $\vec{x}_A$ *is sampled from a distribution that has a higher mean than* $\vec{x}_B$ or $\vec{x}_B$ *is sampled from a distribution that has a higher variance than* $\vec{x}_A$. The null hypothesis is that the two data sets are actually sampled from the same distribution.

The test statistic in a permutation test is the difference between the values of a test statistic of interest t evaluated on the two data sets

$$t_{\text{diff}}\left(\vec{x}\right) := t\left(\vec{x}_A\right) - t\left(\vec{x}_B\right), \tag{11.43}$$

where $\vec{x}$ are all the data merged together. Our goal is to test whether $t\left(\vec{x}_A\right)$ is larger than $t\left(\vec{x}_B\right)$ at a certain significance level. The corresponding rejection region is of the form $\mathcal{R} := \{t \mid t \geq \eta\}$. The problem is how to fix the threshold so that the test has the desired significance level.

Imagine that we randomly permute the labels A and B in the merged data set $\vec{x}$. As a result, some of the data that were labeled as A will be labeled as B and vice versa. If we recompute $t_{\text{diff}}\left(\vec{x}\right)$ we will obviously obtain a different value. However, the *distribution* of the random variable $t_{\text{diff}}(\vec{X})$ under the hypothesis that the data are sampled from the same distribution has *not* changed. Indeed, the null hypothesis implies that the distribution of any function of $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n$ that only depends on the class assigned to each variable is *invariant to permutations*. More formally, the random sequence is **exchangeable** with respect to such functions.

Consider the value of t_{diff} for all the possible permutations of the labels: $t_{\text{diff},1}, t_{\text{diff},2}, \ldots t_{\text{diff},n!}$. If the null hypothesis holds, then it would be surprising to find that $t_{\text{diff}}\left(\vec{x}\right)$ is larger than most of the $t_{\text{diff},i}$. In fact, under the null hypothesis, the random variable $t_{\text{diff}}(\vec{X})$ is uniformly distributed in the set $\{t_{\text{diff},1}, t_{\text{diff},2}, \ldots t_{\text{diff},n!}\}$, so that

$$\mathrm{P}\left(t_{\text{diff}}(\vec{X}) \geq \eta\right) = \frac{1}{n!}\sum_{i=1}^{n!} 1_{t_{\text{diff},i} \geq \eta}. \tag{11.44}$$

This is exactly to the size of the test. We can therefore compute the p value of the observed statistic $t_{\text{diff}}\left(\vec{x}\right)$ as

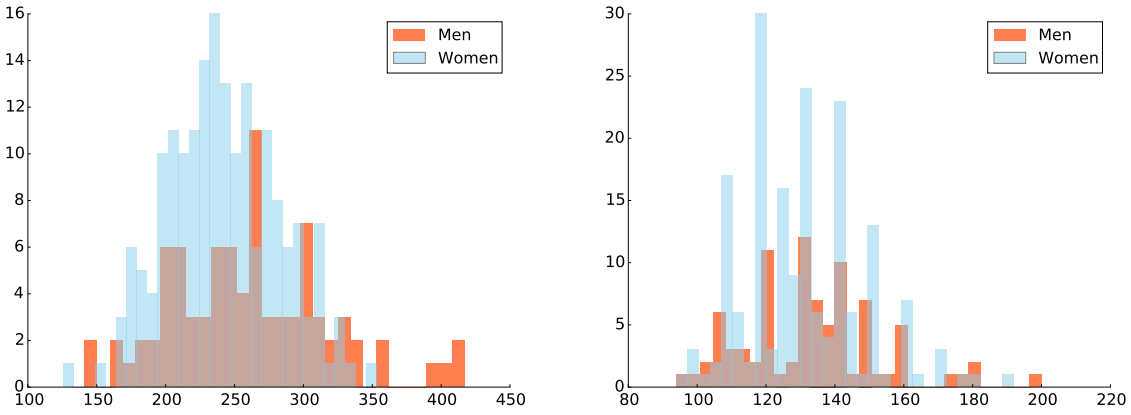
$$p = \mathrm{P}\left(t_{\text{diff}}(\vec{X}) \geq t_{\text{diff}}\left(\vec{x}\right)\right) \tag{11.45}$$

$$= \frac{1}{n!}\sum_{i=1}^{n!} 1_{t_{\text{diff},i} \geq t_{\text{diff}}(\vec{x})}. \tag{11.46}$$

In words, the p value is the fraction of permutations that yield a more extreme test statistic than the one we observe. Unfortunately, it is often challenging to compute (11.46) exactly. Even for moderately sized data sets the number of possible permutations is usually too large (for example, $40! > 8\,10^{47}$) for it to be computationally tractable. In such cases the p value can be approximated by sampling a large number of permutations and making a Monte Carlo approximation of (11.46) with its average.

Before looking at an example, let us review the steps to be followed when applying a permutation test.

1. Choose a conjecture as to how $\vec{x}_A$ and $\vec{x}_B$ are different.



Cholesterol Blood pressure

Figure 11.2: Histograms of the cholesterol and blood-pressure for men and women in Example 11.3.1.

2. Choose a test statistic t_{diff}.
3. Compute $t_{\text{diff}}\left(\vec{x}\right)$.
4. Permute the labels m times and compute the corresponding values of t_{diff}: $t_{\text{diff},1}$, $t_{\text{diff},2}$, $\ldots t_{\text{diff},m}$.
5. Compute the approximate p value

$$p = \mathrm{P}\left(t_{\text{diff}}(\vec{X}) \geq t_{\text{diff}}\left(\vec{x}\right)\right) \tag{11.47}$$

$$= \frac{1}{m}\sum_{i=1}^{m} 1_{t_{\text{diff},i} \geq t_{\text{diff}}(\vec{x})} \tag{11.48}$$

and reject the null hypothesis if it is below a predefined limit (typically 1% or 5%).

Example 11.3.1 (Cholesterol and blood pressure)**.** A scientist want to determine whether men have higher cholesterol and blood pressure. She gathers data from 86 men and 182 women. Figure 11.2 shows the histograms of the cholesterol and blood-pressure for men and women. From the histograms it seems that men have higher levels of cholesterol and blood pressure. The sample mean for cholesterol is 261.3 mg/dl amongst men and 242.0 mg/dl amongst women. The sample mean for blood pressure is 133.2 mmHg amongst men and 130.6 mmHg amongst women.

In order to quantify whether these differences are significant we compute the sample permutation distribution of the difference between the sample means using 10^6 permutations. To make sure that the results are stable, we repeat the procedure three times. The results are shown in Figure 11.3. For cholesterol, the p value is around 0.1%, so we have very strong evidence against the null hypothesis. In contrast, the p value for blood pressure is 13%, so the results are not very conclusive, we cannot reject the possibility that the difference is merely due to random fluctuations.

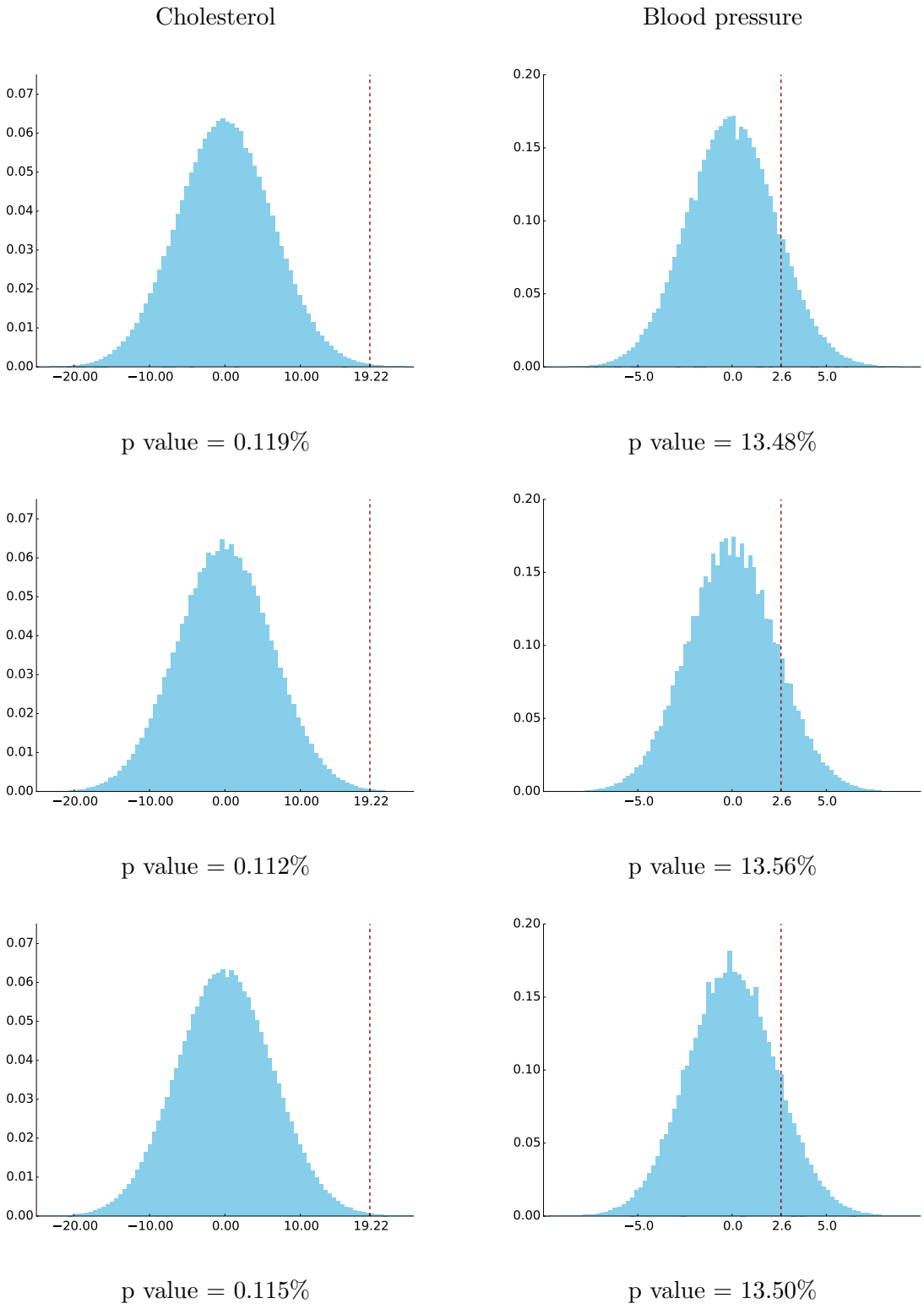


Figure 11.3: Approximate distribution under the null hypothesis of the difference between the sample means of cholesterol and blood pressure in men and women. The observed value for the test statistic is marked by a dashed line.

△

11.4 Multiple testing

In some applications, it is common to conduct many simultaneous hypothesis tests. For example, in computational genomics a researcher might be interested in testing whether any gene within a group of several thousand is relevant to a certain disease. If we apply a hypothesis test with size α in this setting, then the probability of obtaining a false positive for a particular gene is α. Now, assume that we test n genes and that the events *gene i is a false positive*, $1 \leq i \leq n$ are all mutually independent. The probability of obtaining at least one false positive is

$$\begin{aligned} \mathrm{P}\left(\text{at least one false positive}\right) &= 1 - \mathrm{P}\left(\text{no false positives}\right) && (11.49)\\ &= 1 - (1-\alpha)^n . && (11.50) \end{aligned}$$

For $\alpha = 0.01$ and $n = 500$ this probability is equal to 0.99! If we want to control the probability of making a Type I error we must take into account that we are carrying out multiple tests at the same time. A popular procedure to do this is **Bonferroni's method**.

Definition 11.4.1 (Bonferroni's method)**.** *Given n hypothesis tests, compute the corresponding p values $p_1, \ldots, p_n$. For a fixed significance level α reject the ith null hypothesis if*

$$p_i \leq \frac{\alpha}{n}. \tag{11.51}$$

The following lemma shows that the method guarantees that the desired significance level holds simultaneously for all the tests.

Lemma 11.4.2. *If we apply Bonferroni's method, the probability of making a Type I error is bounded by α.*

Proof. The result follows directly from the union bound, which controls the probability of a union of events with the sum of their individual probabilities.

Theorem 11.4.3 (Union bound)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $S_1, S_2, \ldots$ a collection of events in $\mathcal{F}$. Then*

$$\mathrm{P}\left(\cup_i S_i\right) \leq \sum_i \mathrm{P}\left(S_i\right). \tag{11.52}$$

Proof. Let us define the sets:

$$\tilde{S}_i = S_i \cap \cap_{j=1}^{i-1} S_j^c. \tag{11.53}$$

It is straightforward to show by induction that $\cup_{j=1}^n S_j = \cup_{j=1}^n \tilde{S}_j$ for any n, so $\cup_i S_i = \cup_i \tilde{S}_i$. The sets $\tilde{S}_1$, $\tilde{S}_2$, ... are disjoint by construction, so

$$\begin{aligned} \mathrm{P}\left(\cup_i S_i\right) = \mathrm{P}\left(\cup_i \tilde{S}_i\right) &= \sum_i \mathrm{P}\left(\tilde{S}_i\right) \quad \text{by Axiom 2 in Definition 1.1.4} && (11.54)\\ &\leq \sum_i \mathrm{P}\left(S_i\right) \quad \text{because } \tilde{S}_i \subseteq S_i. && (11.55) \end{aligned}$$

□

Applying the bound,

$$\begin{aligned} \mathrm{P}\left(\text{Type I error}\right) &= \mathrm{P}\left(\cup_{i=1}^{n} \text{Type I error for test } i\right) && (11.56)\\ &\leq \sum_{i=1}^{n} \mathrm{P}\left(\text{Type I error for test } i\right) \qquad \text{by the union bound} && (11.57)\\ &= n \cdot \frac{\alpha}{n} = \alpha. && (11.58) \end{aligned}$$

□

Example 11.4.4 (Clutch (continued))**.** If we apply the test in Example 11.1.4 to 10 players, the probability that one of them seems to be clutch just due to chance increases substantially. To control for this, by Bonferroni's method we must divide the p values of the individual tests by 10. As a result, to maintain a significance level of 0.05 we would require that each player score more points per minute during the last quarter in 17 of the 20 games instead of 15 (see Table 11.2) in order to reject the null hypothesis.

△