Foundations of Geometric Methods in Data Analysis : course notes

Frederic.Cazals@inria.fr

January 18, 2017

Contents

Chapter 1

Introduction to statistical hypothesis testing and two-sample tests

1.1 Hypothesis, errors, and test statistics

1.1.1 Hypothesis

We shall analyze collections of samples, using the following notations:

- $X^{(n)} = \{X_1, \ldots, X_n\}$: a random vector consisting of n iid copies of a RV X taking values in $\mathbb{R}d$. A realization $x^{(n)} = \{x_1, \ldots, x_n\}$ of this vector is called a *sample*. The measure associated to the RV X is denoted μ_X.
- $Y^{(n)} = \{Y_1, \ldots, Y_n\}$: a random vector consisting of n iid copies of a RV Y taking values in $\mathbb{R}d$. A realization of this vector is denoted $y^{(n)} = \{y_1, \ldots, y_n\}$. The measure associated to the RV Y is denoted μ_Y.

Definition. 1. *A statistical hypothesis is a set of probability distributions for a random sample. The hypothesis is termed* simple *if the distribution is uniquely determined, and* composite *otherwise.*

One typically deals with two sets of hypothesis:

Definition. 2. *Testing binary hypothesis involves two hypothesis:*

- *The null hypothesis, denoted H_0, expressing the belief / the default / no effect. The corresponding parameters are denoted Θ_0.*
- *The alternative hypothesis, denoted H_1, whose parameters are denoted $\Theta_1 \subset \Theta_0^c$.*

Note that if H_0 is simple, the set Θ_0 reduces to a singleton denoted θ_0—and likewise for H_1.

Example 1. *(Simple versus composite hypothesis, one-sided versus two-sided) Consider a sample $x^{(n)}$ of real values, whose median θ is to be estimated. Here are several options:*

- *Two simple hypothesis:*
  - *H_0: $H_0 : \theta = \theta_0$*
  - *H_1: $H_1 : \theta = \theta_1$*
- *One simple and one composite (two-sided) hypothesis:*

  - H_0: $H_0 : \theta = \theta_0$
  - H_1: $H_1 : \theta \neq \theta_0$
- *One simple and one composite (one-sided) hypothesis:*
  - H_0: $H_0 : \theta = \theta_0$
  - H_1: $H_1 : \theta > \theta_0$

Example 2. *(TST) The so-called* two-sample test *problem. Consider two samples $x^{(n)}$ and $y^{(n)}$. One wishes to test whether the two distributions underlying the data are identical, that is*

- H_0: $\mu_X = \mu_Y$
- H_1: $\mu_X \neq \mu_Y$

In testing hypothesis, one needs to choose between H_0 and H_1. In fact, the problem can be rephrased as follows:

Do the samples carry enough evidence to reject the null H_0?

1.1.2 Errors

In testing an hypothesis, two erroneous situations may occur (see also Table 1.1):

Definition. 3. *(Type I error) A situation where H_0 is rejected, while it is true. The probability of this event is denoted α.*

Definition. 4. *(Type II error) A situation where H_0 is accepted, while it is false. The probability of this event is denoted β.*

	accept H_0	reject H_0
H_0 true	OK	type I error (α)
H_1 true	type II error (β)	OK

Table 1.1: **Type I and type II errors**

Example 3. *In a court, consider a person undergoing a trial. Assume that the null hypothesis H_0 is* the person is innocent*. Then:*

- *Type I error: the person is innocent but convicted.*
- *Type II error: the person is guilty but freed.*

1.1.3 Test statistics

The goal of hypothesis testing is to design so-called test statistics, and to analyze their errors. We define:

Definition. 5. *A* test statistic *is a real-valued function defined on the sample space, that is $T(X^{(n)}) \to \mathbb{R}$.*

As a simple example, one may consider the sample mean $T(X^{(n)}) = \overline{X}$.

1.2 Tests: the Neyman-Person approach

In this section, we consider simple hypothesis only.

1.2.1 Power and consistency

Consider two simple hypothesis:

- $H_0 : \theta = \theta_0$
- $H_1 : \theta = \theta_1$

Under the null, we assume that the distribution of T is given in parametric form by the real valued function $g(t \mid \theta_0)$.

To design a statistical test, one chooses a threshold α, from which regions of acceptance / rejection are derived. The simplest way to define these regions, given α, is to find the value t_α such that the tail of the distribution beyond t_α has weight α (Fig. 1.1) (see also section 1.3).

Once these regions have been defined, one accepts or rejects the null as follows:

- Compute the statistic $T(X_0)$
- Reject the null if $T(X_0)$ falls within region R.

Rejecting the null corresponds to the following rationale:

- Either the null hypothesis is not correct,
- Or it is so, but a rare event has been observed.

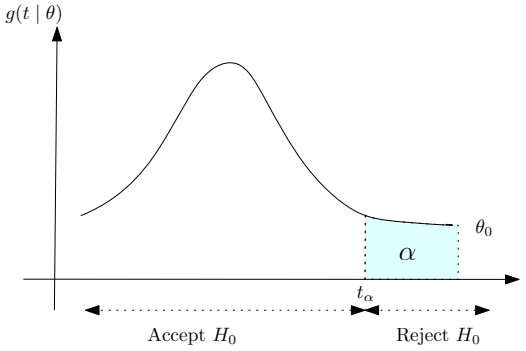

Figure 1.1: **Defining the acceptance/rejection regions.** Here, the regions are defined from the value t_α such that the weight of the tail after t_α is α. Note that one could also have assigned the weight α to the two tails, in which case there is not a unique way to proceed since α can be split into uneven parts.

If H_0 is true, the type I error is also called the size of the test:

Definition. 6. *(Size of test) The size of the test is defined by*

$$\alpha = \int_R g(t|\theta_0)dt. \tag{1.1}$$

If the alternative hypothesis H_1 is true, the test statistic T has a different distribution, whose density is denoted $g(t \mid \theta_1)$. The probability of making an erroneous acceptance is that case is given by:

$$\beta = \int_A g(t|\theta_1), \tag{1.2}$$

from which one defines:

Definition. 7. *(Power of test) The power of the test is the probability of rejecting the null when H_1 is true, that is*

$$P(\theta_1) = 1 - \beta = \int_R g(t|\theta_1)dt. \tag{1.3}$$

Alternatively, one uses the type II error function $\beta(\theta) = 1 - P(\theta)$.

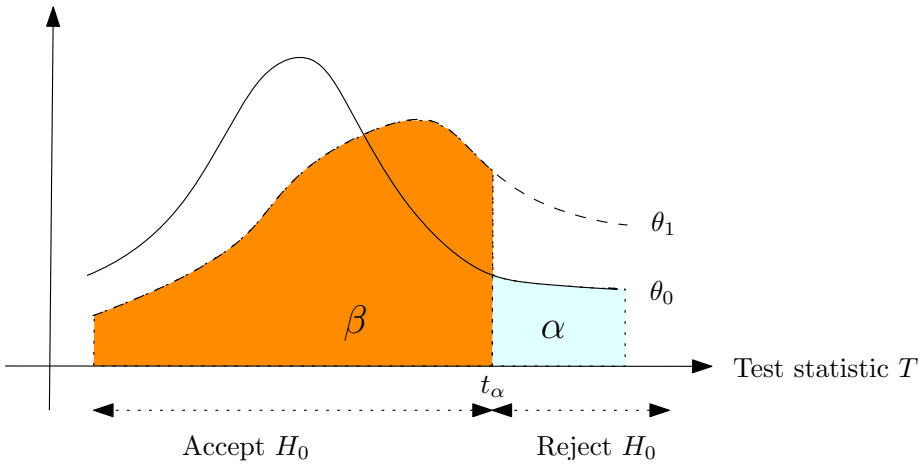

Figure 1.2: **Type I and II errors for simple hypothesis**. Note that in reducing the type I error α (by shifting t_α to the right) yields an increase the type II error β.

As seen from Fig. 1.2, the interpretation of type I and II errors is as follows:

- type I error: occurs with proba. α under H_0,
- type II error: occurs with proba. β under H_1.

1.2.2 The power function and uniformly most powerful tests

Consider a test associated to a given rejection region. In varying the parameter θ_1 of the alternative hypothesis, each value of θ_1 yields a different power value. The corresponding curve is called the *power function* (Fig. 1.3). Note that on such a curve, θ stand for any value of θ_1.

We consider the power $P(\theta_1)$ of the test as a function of θ (Fig. 1.3):

- the power should be large when θ_1 far from θ_0: in such a case, one does not want to accept the null.
- on the other hand, a lower power can be tolerated for θ_1 near θ_0: indeed, if one accepts the null, θ_0 is not so different from θ_1.

Of course, different tests (i.e. associated to different rejection regions) yield different power curves (Fig. 1.4). Whence the importance of tests maximizing the power:

Definition. 8. *A* most powerful test *of size α for testing θ_1 against θ_0 is a test associated with a region R, such that for every region R' for which the error I is also less than α, one has:*

$$\int_{R'} g(t|\theta_1)dt \leq \int_{R} g(t|\theta_1). \tag{1.4}$$

If the previous condition holds for every value of θ except θ_0, the test using region R is termed uniformly most powerful *(UMP).*

From a pictorial standpoint, note the a UMP test has a power curve which *dominates* those associated with other regions R'.

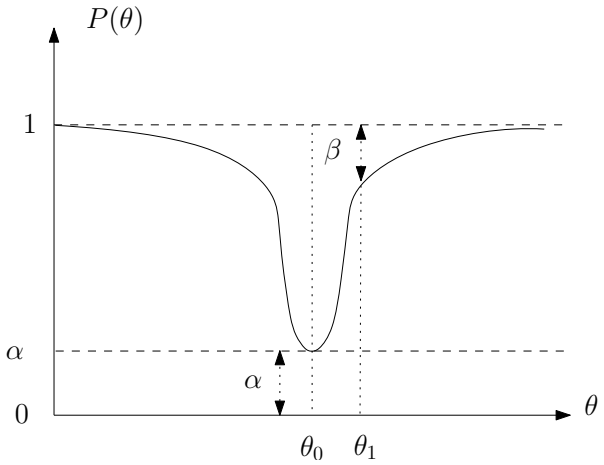

Figure 1.3: **An ideal power function.** A small power can be tolerated for θ_1 near θ_0: if one accepts the null, θ_0 is not so different from θ_1.

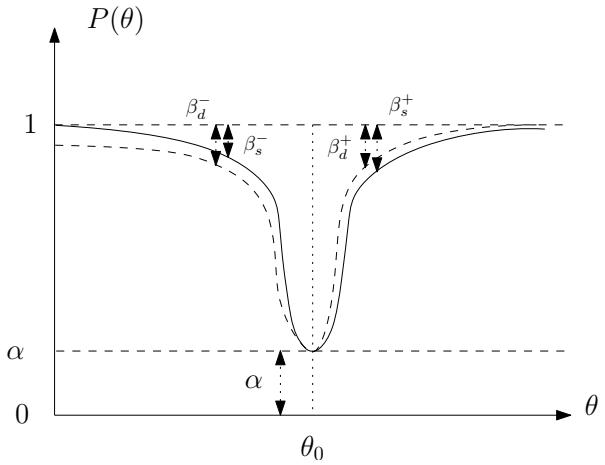

Figure 1.4: **Power curves for two tests.** We consider two tests, each characterized by a rejection region. Varying θ_1 yields the two power curves. For $\theta_1 < \theta_0$, the type II error is less for the solid curve, while the opposite holds for $\theta_1 > \theta_0$.

1.2.3 One-sided versus two-sided tests

In example 1, we introduced the notion of one-sided versus two-sided test. For example, if the parameter θ is the median of a sample, the null may state that the median is equal to zero, while the alternative may stipulate that the median is strictly positive.

As seen on Fig. 1.5, one and two-sided tests use different rejections regions.

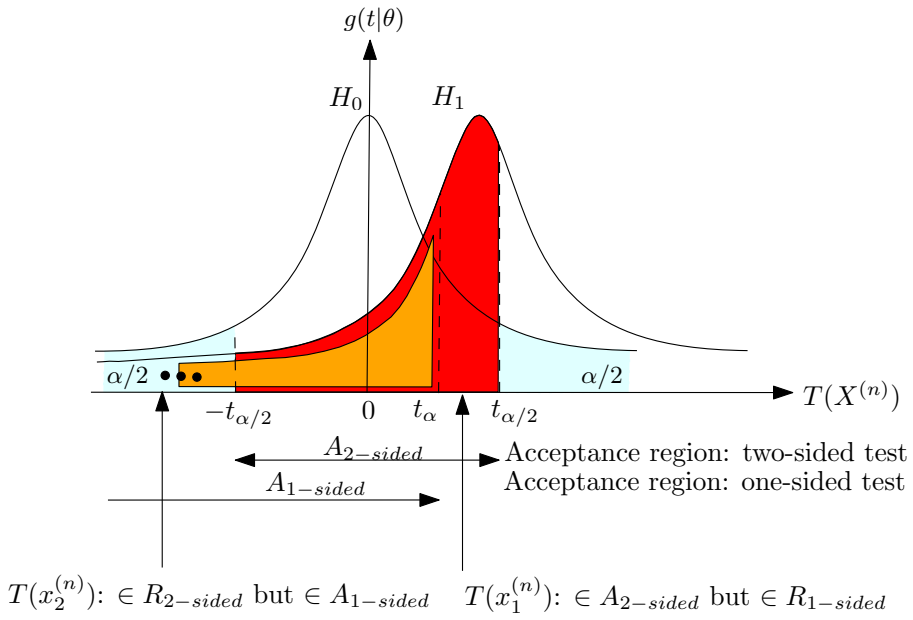

Figure 1.5: **One-sided versus two-sided tests, interpretation in terms of power: illustration on the problem of testing the median of a sample.** For the sake of simplicity, the distribution under the null is assumed to be symmetric. **(Two-sided test)** At type I error threshold α, the accept region is $[-t_{\alpha/2}, t_{\alpha/2}]$, and the associated type II error is the weight of the red region. **(One-sided test)** At type I error threshold α, the accept region is $[-\infty, t_\alpha]$, and the associated type II error is the weight of the orange region.

In using a one-sided test:

- one actually only uses one side of the null distribution (and whence the observed values located on that side), to make the decision.
- one increases the power on one side: on Fig. 1.5, the test with $T(x_1^{(n)})$ gets rejected (one-sided) rather than accepted (two-sided). However, this increase in power is at the detriment of statistics on the left-hand side: $T(x_2^{(n)})$ gets accepted (one-sided) instead of rejected (two-sided).

In short, one needs to remember the following:

- It is the null hypothesis which shapes the distribution of the test statistic;
- But it is the alternative hypothesis which shapes the acceptance/rejection regions.

1.3 Designing parametric tests

The classical methods to design tests are the following ones:

- Likelihood ratio tests (LRT),
- Bayesian tests,
- Union-intersection and intersection-union tests.

The reader is referred to [CB01, LR05]. In the sequel, we merely give one illustrate of LRT.

1.3.1 Likelihood ratio tests (LRT)

We define:

Definition. 9. *Consider two hypothesis H_0 and H_1, characterized by parameter domain Θ_0 and Θ_1, respectively. The LRT test statistic is defined by*

$$\lambda(x^{(n)}) = \frac{\sup_{\Theta_0} L(\theta \mid x^{(n)})}{\sup_{\Theta} L(\theta \mid x^{(n)})}. \tag{1.5}$$

A LRT is any test whose rejection region is of the form $\{x^{(n)} : \lambda(x^{(n)}) \leq c\}$, with $0 \leq c \leq 1$.

To understand the intuition behind Eq. (1.5), observe that the parameter domain used in the denominator contains that of the numerator. Thus, the ratio is small when there are parameter points in the alternative hypothesis, yielding a much higher likelihood than that obtained with the best parameter point(s) in the null hypothesis.

Example 4. *(LRT for a Bernoulli random variable.) Assume that the $X_i \sim \mathcal{B}(p)$, and consider the sum $Y = \sum X_i$. Let us test the hypothesis*

- *H_0: $p = p_0$*
- *H_1: $p \neq p_o$*

.

The likelihood function is determined by Y, namely $L(p \mid x^{(n)}) = p^y(1-p)^{n-y}$, with y the number of ones obtained. Therefore:

$$\lambda(y) = \frac{p_0^y(1-p_0)^{n-y}}{\max_{p\in[0,1]} p_0^y(1-p_0)^{n-y}}$$

Using the maximum likelihood estimator (MLE) $\hat{p} = y/n$ for the denominator of Eq. (1.5), we finally obtain:

$$\lambda(y) = (\frac{n\theta_0}{y})^y(\frac{n(1-\theta_0)}{n-y})^{n-y}.$$

Remark 1. *LRT are parametric tests. Indeed, the likelihood depends on the particular parametric model used.*

1.4 Statistical significance via p-values – The Fisherian approach

1.4.1 The p-value

The following definition is illustrated on Fig. 1.6:

Definition. 10. *(p-value) Let T be some test statistic. Upon observing a data $x^{(n)}$, the p-value is the conditional probability, under the null hypothesis H_0, to observe a value as extreme as that associated with x_0:*

$$p = \mathbb{P}\left[T(X^{(n)}) \geq T(x^{(n)}) \mid H_0\right]. \tag{1.6}$$

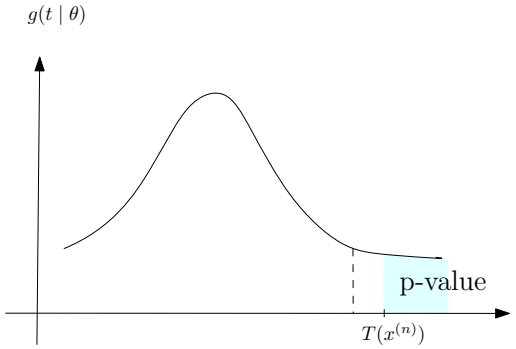

Figure 1.6: **p-value: example.** Given a test statistic T, a p-value can be constructed by computing the probability to get an observation more extreme than the one observed. A p-value is of special interest if the rejection region is large (Fig. 1.1), yielding a large type I error.

Note that for the sake of simplification, one often writes T instead of $T(X^{(n)})$, as the semantics is clear from the context. Likewise, for the sake of conciseness, the conditioning on the null hypothesis is often forgotten.

Remark 2. *There is a more general way to introduce p-values, namely using an upper bound on the type I error [CB01]. More precisely, a p-value $p(\cdot)$ is called* valid *if $\forall \theta \in \Theta_0$, and every $0 \leq \alpha \leq 1$, one has*

$$\mathbb{P}_\theta\left[p(X^{(n)}) \leq \alpha\right] \leq \alpha \tag{1.7}$$

In that case, definition 10 becomes a theorem. The reader is referred to [CB01, Chapter 8] for further details. This approach is of interest in particular in the sequential setting, where the number of samples to process is unknown a priori. One recent example of such tests can be found in [LC15]. Another example of generalized p-value is that of used for resampling and permutations tests [Har12].

Rationale in observing a small p-value:

- either a rare event has been observed (tail of the null distribution),
- or the null is false.

Observation. 1. *Consider a test statistic T whose associated density is smooth, and whose associated cumulated distribution function $F_0(T)$ is strictly increasing. Under the null hypothesis, the associated p-value is a uniform random variable $\mathcal{U}(0,1)$.*

Proof. Let F_0^{-1} be the inverse of the cdf of T. By definition of the p-value, we have:

$$\text{p-value} = Z = 1 - F_0(T).$$

Therefore

$$\begin{aligned}
\mathbb{P}\left[Z \leq t | H_0\right] &= \mathbb{P}\left[F_0(T) \geq 1 - t | H_0\right] \\
&= 1 - \mathbb{P}\left[F_0(T) \leq 1 - t | H_0\right] \\
&= 1 - \mathbb{P}\left[F_0^{-1}(F_0(T)) \leq F_0^{-1}(1-t) | H_0\right] && (\text{ since } F_0^{-1} \text{ is increasing}) \\
&= 1 - \mathbb{P}\left[T \leq F_0^{-1}(1-t) | H_0\right] && (\text{by def. of the inverse}) \\
&= 1 - F_0(F_0^{-1}(1-t)) = t.
\end{aligned}$$

□

Remark 3. *The previous observation is a particular case of the so-called* probability integral transform. *The strictly increasing assumption is actually unnecessary, as can be seen from the proof in [CB01, Section 2.1].*

Remark 4. *In section 1.2.3, one-sided versus two-sided tests have been discussed. This discussion is also valid in using p-values to assess a null hypothesis. In testing a one-sided hypothesis, only one side of the distribution of the test statistic is used.*

1.4.2 Examples

Example 5. *(Estimating the median of a set of real values) In the following, we present two-sided and one-sided tests, and discuss the rationale is choosing their rejection regions.*

Consider $n = 10$ real values X_i, say 3 negatives and 7 positives.

Case 1: two-sided test *We wish to test the following on the median θ:*

- $H_0 : \theta = 0$ *(simple hypothesis)*
- $H_1 : \theta \neq 0$ *(composite hypothesis)*

Denoting θ the median, consider the RV $Y = \sum_i \mathbf{1}_{X_i > \theta}$. Under the null, Y is binomial $\mathcal{B}(n, 1/2)$:

$$\mathbb{P}\left[T(X^{(n)}) = k\right] = \binom{n}{k} (1/2)^{n-k} (1/2)^k. \tag{1.8}$$

The expectation is $n/2$, namely 5 in our example with $n = 10$. Since evidence against the null is provided by either large or small values of Y, we define the statistic

$$T = \mid Y - 5 \mid .$$

Since 7 positive values were observed, we have $T(x^{(n)}) = 2$ and

$$\mathbb{P}\left[T(X^{(n)}) \geq 2\right] = \mathbb{P}\left[Y(X^{(n)}) \geq 7\right] + \mathbb{P}\left[Y(X^{(n)}) \leq 3\right] \sim 0.34.$$

That is, there is not enough evidence to reject the null at a threshold $\alpha = 0.05$.

Case 2: one-sided test *Assume that we now wish to test*

- $H_0 : \theta = 0$ *(simple hypothesis)*
- $H_1 : \theta > 0$ *(composite hypothesis)*

In that case, the alternative hypothesis is composite, and the test one-sided since evidence is sought for large values (See also Fig. 1.5). Thus, the associated 1-sided p-value is:

$$p = \mathbb{P}\left[T(X^{(n)}) \geq 7\right] \sim 0.17,$$

and one does not reject the null either.
Rmk: *by seeking evidence for large values only, one reduces the p-value, and increases the chances to reject the null at a given α level. This is similar to shifting the reject interval on the left hand side, as seen on Fig.* **??**.

1.4.3 Pitfalls and potential problems in using p-values

Some important facts calling for care in using p-values [Wag07].

Comparing p-values. A test is run on two samples A and B, and one gets p-values $p_A = 0.01$ and $p_B = 0.001$. Can anything be said from these two p-values and the associated samples?

The answer is No:

- On the one hand, a p-value can only be interpreted under the null. If one rejects the null, the p-value does not have any meaning.
- On the other hand, if the null holds, the p-value is a uniform random variable $\mathcal{U}(0,1)$.

Concluding: comparing two p-values does not make sense.

Dependence on data which were not observed. Consider a real valued discrete random variable X taking 5 values. We wish to assess two null hypothesis H_0 and $H_0^{'}$ for that RV, using the test statistic $T(X) = X$. Since a null hypothesis is a probability distribution for the data, assume that one has the two distributions displayed in Table 1.2.

Assuming that the value $x = 2$ has been observed, one gets the following two p-values:

$$p_0 = \mathbb{P}\left[X \geq 2|H_0\right] = 0.11 \text{ and } p_0^{'} = \mathbb{P}\left[X \geq 2|H_0^{'}\right] = 0.05$$

That is, $x = 2$ provides a significant evidence against $H_0^{'}$ at $\alpha = 0.05$ level, but does not provide evidence against H_0 at $\alpha = 0.1$ level. Strikingly, $x = 0.4$ is equally likely in both cases.

In this example, it is the non-observed data ($x = 3$ and $x = 4$) which influence the p-value. This is rather counter-intuitive.

x	0	1	2	3	4
$H_0(x)$	0.75	0.14	0.04	0.037	0.033
$H_0^{'}(x)$	0.70	0.25	0.04	0.005	0.005

Table 1.2: **Dependence of p-values on non-observed data.**

Dependence on the intentions - sampling plan. Two researchers want to test whether a wine tester can distinguish two wines, say Pommard versus Gevrey-Chambertin. The null hypothesis is that the tester makes decision at random.

Two statisticians believe that individual decisions are made by coin tossing. But they use different sampling plans to run a test.

Researcher 1: gives a predefined number i.e. 12 glasses of wines, and counts Correct and Wrong answers.

A binomial model with parameter θ is used to count the number C of correct answers, that is

$$\mathbb{P}\left[C(X^{(n)}) = k|\theta\right] = \binom{12}{k}\theta^k(1-\theta)^{n-k}.$$

Assume that the following sequence of 12=9 (correct) + 3 (erroneous) answers has been observed:

$$CCCCCWWCCCCW \tag{1.9}$$

Under H_0 and $\theta = 1/2$, and one expects $C = 6$. Since evidence against the null is provided by low and large values of C, consider the test statistic $T =\mid C - 6 \mid$.

The observed value is $T(x^{(n)}) =\mid 9 - 6 \mid= 3$. The 2-sided p-value is

$$p = \mathbb{P}\left[T(X^{(n)}) \geq 3|\theta = 1/2\right] = \mathbb{P}\left[C(X^{(n)}) \geq 9|\theta = 1/2\right] + \mathbb{P}\left[C(X^{(n)}) \leq 3|\theta = 1/2\right] \sim 0.146.$$

Researcher 2: the researcher does not define a priori a number of glasses to test, but decides to stop upon observing the 3rd wrong answer.

The test statistic T is taken as the number of glasses needed to reach k wrong guesses. One has:

$$\mathbb{P}\left[T(X^{(n)}) = n|\theta\right] = \binom{n-1}{k-1}\theta^{n-k}(1-\theta)^k.$$

If the same 12 answers are observed, the p-value is now

$$p = \mathbb{P}\left[T(X^{(n)}) \geq 12|\theta = 1/2\right] = \sum_{n\geq 12}\binom{n-1}{2}\frac{1}{2}^n \sim 0.033.$$

As a conclusion: the same data, under different sampling plans, yields different p-values.

As this example shows, the question to be answered has to be pondered upon carefully beforehand.

1.5 A non parametric two-sample test: the Mann-Whitney test

1.5.1 The Mann-Whitney test

In this section, we consider the simple case where the random variable is real valued. In that case, we wish to compare n realization of X against m realizations of Y.

Idea. The two-sample test has been introduced in example 2, and consists in checking whether two distributions match.

- Compute the sorted list $L = x^{(n)} \cup y^{(n)}$ of the $n + m$ values
- Compute the ranks of the x_is
- Reject the null if the sum of ranks is too small and/or too large (depends on whether the test is one sided or not)

More formally, define (Fig. 1.7):

$$R_X = \sum_i \text{rank}(x_i, L) \text{ and } \Sigma_X = \frac{n(n+1)}{2} \tag{1.10}$$

$$R_Y = \sum_i \text{rank}(y_j, L) \text{ and } \Sigma_Y = \frac{m(m+1)}{2} \tag{1.11}$$

$$\tag{1.12}$$

Test statistic: U_X. The statistic of interest is the number of times an element from Y precedes an element from X, and vice versa (Fig. 1.7):

$$\begin{cases} U_X = \sum_{x_i} \#(y_j < x_i), \\ U_Y = \sum_{y_j} \#(x_i < y_j). \end{cases} \tag{1.13}$$

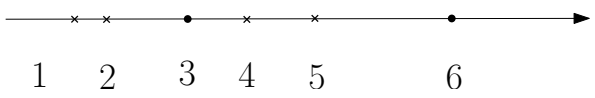

Figure 1.7: **Sum of ranks for the Mann-Whitney test: illustration for two populations $x^{(n)}$ and $y^{(n)}$ of size $n = 4$ and $m = 2$.** One has: Population X: $R_X = 1 = 2 + 4 + 5 = 12$; $\Sigma_X = 4 \cdot 5/2 = 10$; $U_X = 2$. Population Y: $R_Y = 3 + 6 = 9$; $\Sigma_Y = 2 \cdot 3/2 = 3$; $U_Y = 6$. Note that $U_X + U_Y = 2 + 6 = 8 = nm$.

Observation. 2. *One has*

$$\begin{cases} U_X = R_X - \Sigma_X, \\ U_Y = R_Y - \Sigma_Y, \\ U_X + U_Y = nm. \end{cases} \tag{1.14}$$

Proof. To prove the first equality (or the second), observe that:

$$R_X = \sum_{x_i} \text{rank}(x_i, L) = \sum_{x_i} \big(\#(y_j < x_i) + \text{rank}(x_i, X) \big) \Leftrightarrow R_X = U_X + \Sigma_X.$$

For the third claim, consider the sum of the ranks in L, that is:

$$\begin{aligned} R_L = R_X + R_Y &= U_X + \Sigma_X + U_Y + \Sigma_Y \\ &= \frac{(n+m)(n+m+1)}{2} \\ &= nm + \frac{n(n+1)}{2} + \frac{m(m+1)}{2} \\ &= nm + \Sigma_X + \Sigma_Y. \end{aligned}$$

□

The following is also immediate:

Observation. 3. *Under the model of equally likely permutations, one has:*

$$\mathbb{E}[U_X] = \frac{nm}{2}. \tag{1.15}$$

Proof. Since $\mathbb{E}[U_X] = \mathbb{E}[R_X] - n(n+1)/2$, we compute $\mathbb{E}[R_X]$, the expectation of the ranks of the values of the x_is. Consider the RV $R_k = 1$ if the k-th position of the ranked list is occupied by an x_i, and 0 otherwise. Note that the expectation of R_k is $k/(n+m)$. We have

$$R_X = \sum_{k=1,\ldots,n+m} R_k,$$

and by the linearity of expectation

$$\mathbb{E}[R_X] = \sum_{k=1,\ldots,n+m} \mathbb{E}[R_k].$$

But since $\mathbb{E}[R_k] = k/(n+m)$, we get

$$\mathbb{E}[R_X] = \sum_{k=1,\dots,n+m} k/(n+m) = \frac{(n+m)(n+m+1)}{2}\frac{n}{n+m} = \frac{n(n+m+1)}{2}.$$

The claim follows. □

Distribution of U_X. Consider U_X. Since we only care for the values of Y_js smaller than values of X_is, assume that $X \equiv 1$ and $Y \equiv 0$. In that case:

$$U \equiv U_X = \#0s \text{ preceding } 1s.$$

There are $\binom{n+m}{n}$ different strings of length $n+m$. Denote

$$p_{n,m}(U) : \#\text{ strings with } n \text{ ones and } m \text{ zeros, achieving the statistic } U. \tag{1.16}$$

Since a string on length $n+m$ is obtained by adding a 0 or a 1 to a string of length $n+m-1$:

- If a 0 is added: # 0 preceding 1 does not change
- If a 1 is added: the m zeros score one more.

Whence the following recurrence

$$\begin{cases} p_{n,m}(U) = p_{n,m-1}(U) + p_{n-1,m}(U-m) // \text{resp.: add a 0 and a 1 at the end} \\ p_{n,m} = 0 \text{ if } U < 0 \\ p_{n,0} \text{ and } p_{0,m} = 0 \text{ if } U \neq 0, \text{ and 1 if } U = 0. \end{cases}$$

One also has the following approximation theorem [Kee62], which can be proved in particular using asymptotic analysis [FS09]:

Theorem. 1. *When $n \to \infty, m \to \infty$, $U \sim \mathcal{N}(\mu, \sigma)$, with $\mu = n \times m/2$, and $\sigma^2 = n \times m \times (m+m+1)/12$.*

Test and p-value. The statistic U_X can be used to test three alternative hypothesis:

- $X > Y$: realization of X will have ranks larger than those from Y, and U_X will be large–the maximum being nm. The right tail from the distribution of U_X is needed to compute the p-value.
- $X < Y$: realizations of Y will have ranks smaller than those from Y, and U_X will be small–the minimum being 0. The left tail from the distribution of U_X is needed to compute the p-value.
- $X \neq Y$: realizations of X will have ranks smaller or larger than those from Y, and U_X may be small or large. Both tails from the distribution of U_X are needed to compute the p-value.

The intervals of accept/reject, or the p-values are derived from the distribution of U_X.

Null hypothesis and effect size. If one reject the null: what is the magnitude of the difference. The following comes with the U test:

Definition. 11. *The Hodges–Lehmann estimate Δ is the median of all pairwise differences between an observation from the first population and an observation from the second population.*

1.5.2 Applications of the Mann-Whitney test to ROC curves

ROC curves aim at assessing the ability of a classifier (consisting of a real value) to predict a binary classification. The individuals to be classified are termed the *positives* and *negatives*.

The idea is rather simple (Fig. 1.8). Consider $n+m$ individual, say n positives and m negatives. One processes the data by increasing test statistic, one at a time, and at each step, predicts the values below (or above) the current one as positives. In doing so, an individual predicted as positive is either a true positive, or a false positive.

In fact, at all times, positives P and negatives N split as follows, where the prefixes T and F stand for True and False, respectively:

$$P = TP + FN, N = TN + FP, \tag{1.17}$$

From this one defines

$$\begin{cases} \text{sensitivity} = \frac{TP}{P}, \text{specificity} = \frac{TN}{N} \\ \text{false alarm rate} = 1 - \text{specificity} = \frac{FP}{N}. \end{cases} \tag{1.18}$$

The performance of the classifier is given by the Area Under the Curve (AUC) (Fig. 1.8).

To assess it, denote f_i the number of FP seen upon processing the i-th TP x_i.

Upon processing x_i, a vertical step is made, and x_i contributes the following area to AUC:

$$A_i = \frac{1}{n}\frac{(m - f_i)}{m}.$$

The AUC is therefore computed by adding this are for all TP, namely:

$$AUC = \sum_{i=1,\ldots,n} \frac{1}{nm}(m - f_i) = 1 - \frac{1}{nm}\sum_i f_i.$$

Thus, $U = \sum_i f_i$ is the total number of Negatives which precede Positives, and one has

$$U = nm(1 - AUC).$$

In other words, once AUC is known, one derives the statistic of the Mann-Whitney test. From this statistic, one assesses whether the AUC is significant or not, at a given level.

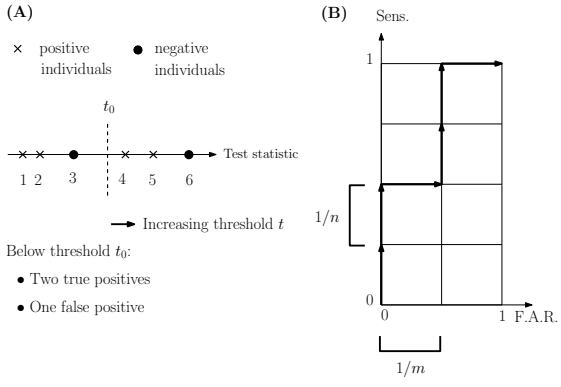

Figure 1.8: **Connexion between the Mann-Whitney test, and ROC curves** In this example, it is assumed that there are $n = 4$ positives,a and $m = 2$ negatives. **(A)** Varying the threshold amounts to predicting individuals as true positives or false positives **(B)** In varying the threshold, one traces a polyline, the ROC curves, in the plane *false alarm rate* $\times$ *sensitivity*. The performance of the classifier is assessed by the area under the curve (AUC).

1.6 Resampling (Monte Carlo, permutation) tests

The previous sections resorted to two types of constructions to build test statistics: either parametric expressions, or enumeration of permutations (Mann-Whitney test). This section generalizes this latter approach, introducing so-called resampling tests. Such tests were introduced in [Dwa57], [HT89] and [Ode91]. More recently, the selected computational aspects were investigated in [BZ00], and the various variants discussed in [PS10]. Mathematical aspects, and in particular the validity of p-values, were studied in [Har12].

1.6.1 Resampling tests: the idea

The technique presented in this section is of special interest in two cases:

- **Parametric bootstrap.** Assume that an analytical expression is given to model the observations (e.g., real valued data follow a Gaussian or any other parametric distribution), but that no analytical expression is known for the distribution of the test statistic under the null. In that case, if one knows how to generate alternate datasets, called *random datasets* in the sequel, the distribution of the test statistic can be inferred. More precisely, assume that I random datasets have been generated. Denoting $T_o = T(x^{(n)})$ the test statistic on the observed data, an an intuitive way to estimate a p-value is:

$$\hat{p} = \frac{1}{I}\sum_{i=1}^{I} \mathbf{1}_{T_j \geq T_0} \equiv \frac{b}{I}. \tag{1.19}$$

  Note that b stands for the number of random datasets yielding a test statistic larger than that obtained on the observed dataset.

- **Permutation tests.** Assume that one does not have the ability to generate random data sets—which is in particular the case if no parametric model is available to model the data. In that case, random datasets can be created using permutations—see examples below. Upon evaluating the test statistic for each of them, one can resorts to equation (1.19) or a variant as well.

A key question is to assess whether or not the type I error mastered, that is whether $\mathbb{P}[\hat{p} \leq \alpha] \leq \alpha$. Because $\hat{p} < \alpha$ occurs with positive probability even when $p > \alpha$, one may indeed have

$$\mathbb{P}[\hat{p} < \alpha] > \alpha.$$

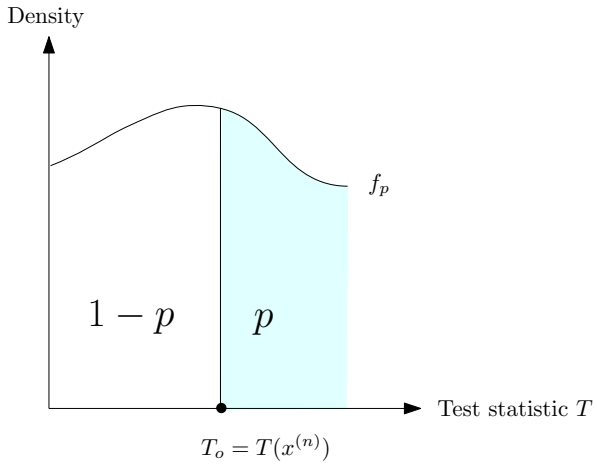

Figure 1.9: **p-value from resampling** The unknown p-value associated with a sample $x^{(n)}$ is denoted $p = \mathbb{P}\left[T(X^{(n)}) \geq T(x^{(n)})|H_0\right]$. A **resampling** plan aims at generating I random datasets, from which an estimate $\hat{p}$ for p is obtained.

1.6.2 Level of a resampling test: basic derivation

Denote p the p-value associated with the observation $x^{(n)}$, that is $p = \mathbb{P}\left[T(X^{(n)}) \geq T_o | H_0\right]$.

From Eq. (1.19), denote B the random variable equal to the number of random samples yielding a test statistic $\geq T_o$. This RV takes values in $0, \ldots, I$. Moreover, assuming that the random datasets are independent, this RV follows a binomial distribution $\mathcal{B}(I, p)$ (Fig. 1.9).

$$\mathbb{P}\left[\hat{p} = \frac{b}{I} | p\right] = \mathbb{P}\left[B = b | p\right] = \binom{I}{b} p^b (1-p)^{I-b}. \tag{1.20}$$

Therefore, denoting $F(p)$ the distribution of the unknown p-value, removing the conditioning from the previous equation yields

$$\mathbb{P}\left[\hat{p} = \frac{b}{I}\right] = \mathbb{P}\left[B = b\right] = \int_0^1 \binom{I}{b} p^b (1-p)^{I-b} dF(p). \tag{1.21}$$

Consider now the case here the test statistic T is continuous. In that case, under the null, the distribution of p is a uniform distribution $\mathcal{U}(0, 1)$. Therefore, a simple integration yields

Observation. 4. *For a continuous test statistic, under the null hypothesis, one has*

$$\mathbb{P}\left[\hat{p} = \frac{b}{I}\right] = \frac{1}{I+1}. \tag{1.22}$$

Corollary. 1. *The type I error rate in using $\hat{p} = b/I$ as p-value estimate is given by*

$$\mathbb{P}\left[\hat{p} \leq \alpha\right] = \frac{\lfloor I\alpha \rfloor + 1}{I+1}. \tag{1.23}$$

Proof. Observe that $\lfloor I\alpha \rfloor / I \leq \alpha \leq \lceil I\alpha \rceil / I$. Thus, using Eq. (1.22), we get

$$\mathbb{P}\left[\hat{p} \leq \alpha\right] = \sum_{b=0}^{\lfloor I\alpha \rfloor} \mathbb{P}\left[\hat{p} = \frac{b}{I}\right]. \tag{1.24}$$

Combining the expression of Eq. (1.22) and the previous expression yields the result. □

The expression of Eq. (1.23) calls for the following comments (see also Fig. 1.10):

- The error rate is generally above the nominal rate especially for small values of α (Fig. 1.10. Thus, using b/I as estimator does not yield a valid p-value (see also Eq. 1.7).
- The error rate is always larger than $1/(m+1)$; that is, it always exceeds α when $\alpha < 1/(m+1)$.
- In fact, a test with controlled type I error requires using directly the observations, as wee shall see below.

Figure 1.10: **Permutation Tests: Controlling the Type I Error.** The type I error of Eq. (1.23) is generally above the the nominal rate (x-axis), and does not yield a valid p-value.

Remark 5. *Comments on permutation tests:* `http://stats.stackexchange.com/questions/20217/bootstrapping-vs-permutation-tests` *nb: Wilcoxon special case of permutation test*

1.6.3 The case of parametric bootstrap

This case is generally referred to as Monte Carlo tests. The basic assumption is that it is possible to generate independent random datasets under the null – which is possible since by assumption an analytical expression for the data is taken for granted.

We have seen that the expression of Eq. (1.23) does not yield a test with controlled type I error. In fact, it can be shown that the following is a valid p-value [Har12]:

$$\hat{p} = \frac{b+1}{I+1}. \tag{1.25}$$

Remark 6. *Note that the previous expression uses the number b of random samples yielding a test statistic more extreme than the observed test statistic T_o, and not $\lfloor I\alpha \rfloor$ as in Eq. (1.23).*

1.6.4 The case of permutations tests

In a permutation test, the random dataset is obtained using permutations acting on the observed data. We provide two examples.

Example 1: two-sample test. Consider two datasets $x^{(n)}$ and $y^{(n)}$. Merging them and associated labels to the individual observation yields the *pooled dataset*:

$$D = (X_1, \ldots, X_n, Y_1, \ldots, Y_m) \tag{1.26}$$

$$L = (0, \ldots, 0, 1, \ldots, 1) \tag{1.27}$$

Under the null, the X_is and Y_js can be exchanged, so that it makes sense to create random datasets

$$(D, \pi(L)), \text{ for a permutation, out of the } \binom{n+m}{n} \text{ possible ones.}$$

The test statistic can then be computed for each such dataset, and the p-value estimated as done previously.

Example 2: correlation. Assume now that the linear correlation coefficient needs to be evaluated, for two real-valued datasets $x^{(n)}$ and $y^{(n)}$. The reference i.e. observed dataset consists of the pairs $\{(x_i, y_i)\}$. Then, one creates random datasets by shuffling the y_is onto the x_i, that is, each random dataset requires a random permutation of the y_is.

Remark 7. *In using resampling, a loss of power may be incurred if too few randomized datasets are used. To get around this difficulty, extrapolation of the e power function may be used [BZ00].*

Permutations without replacement. In using permutations, we first examine the following case:

- The permutations used are all different, and different from that corresponding to the observed data.
- The test statistic is injective (distinct permutations yield distinct values), and these values are equally likely under the permutation distribution. (NB: this hypothesis mimics that used for the Monte Carlo version, in the case of the parametric bootstrap.)

In that case, the formula used is also the one of Eq. (1.25), see [PS10].

Permutations with replacement. A more involved case is that where permutations are randomly drawn with replacement, since random permutations may include repetitions–including the original data.

To handle this case, the reader is referred to [PS10] and also to [Dwa57].

1.7 Concluding remarks

This class focused on key concepts. Some critical topics were not touched upon, and in particular:

- **Effect sizes.** In testing a statistical hypothesis, one summarizes the information into a bit (accept, reject). A central problem is to assess the magnitude of an effect. This topic is generally covered by estimating intervals, and by computing so-called effect sizes.
- **Multivariate two-sample tests.** To perform two-sample tests on multivariate data, of particular interest is the *Maximum Mean Discrepancy (MMD)* test [GBR+12]. This test uses a *kernel*, and recent developments have addressed the automatic detected of the most relevant scale to compare two datasets.
- **Feedback.** In the context of two-sample tests, if a difference between two populations is observed, a natural question is to understand the nature of the difference (which sample points contribute to the difference, do these form clusters, etc). A recent contribution to this problem is [CL15].
- **Feature selection.** Also in the context of two-sample tests, an important problem is the selection of features which do convey the difference between the populations compared. A recent contribution to this problem is [MJ15].
- **Applications to inference and shape analysis.** Resampling tests and persistence (barcodes) were recently combined [Bub15], yielding statistical tests focusing on topological invariants.

Bibliography

[Bub15] P. Bubenik. Statistical topological data analysis using persistence landscapes. *J. of Machine Learning Research*, 16:77–102, 2015.

[BZ00] D. Boos and J. Zhang. Monte carlo evaluation of resampling-based hypothesis tests. *Journal of the American Statistical Association*, 95(450):486–492, 2000.

[CB01] G. Casella and R. Berger. *Statistical inference*. Duxbury Press, 2001.

[CL15] F. Cazals and A. Lhéritier. Beyond two-sample-tests: Localizing data discrepancies in high-dimensional spaces. In P. Gallinari, J. Kwok, G. Pasi, and O. Zaiane, editors, *IEEE/ACM International Conference on Data Science and Advanced Analytics*, Paris, 2015. Preprint: Inria tech report 8734.

[Dwa57] M. Dwass. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28(1):181–187, 1957.

[FS09] P. Flajolet and R. Sedgewick. *Analytic combinatorics*. cambridge University press, 2009.

[GBR+12] A. Gretton, K.M. Borgwardt, J.R. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[Har12] M.T. Harrison. Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika*, 99(1):57–69, 2012.

[HT89] Peter Hall and DM Titterington. The effect of simulation order on level accuracy and power of monte carlo tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 459–467, 1989.

[Kee62] E.S. Keeping. *Introduction to statistical inference*, volume 26. Courier Corporation, 1962.

[LC15] A. Lhéritier and F. Cazals. A sequential non-parametric two-sample test. *Submitted*, 2015. Preprint: Inria tech report 8704.

[LR05] E.L. Lehmann and J.P. Romano. *Testing statistical hypotheses*. Springer Science+ Business Media, 2005.

[MJ15] J.W. Mueller and T. Jaakkola. Principal differences analysis: Interpretable characterization of differences between distributions. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[Ode91] N.L. Oden. Allocation of effort in monte carlo simulation for power of permutation tests. *Journal of the American Statistical Association*, 86(416):1074–1076, 1991.

[PS10] B. Phipson and G.K. Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

[Wag07] E-J. Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, 14(5):779–804, 2007.