INTRODUCTION TO THE TEST OF HYPOTHESES

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

DA_2022 Lecture n. 10, Rome April 3trd 2022

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

outline

- the material for this topic can be found in chap. 6 of W&S READ CAREFULLY THE ENTIRE CHAPTER

General methodological remark

- $H\emptyset$ and $H1$ are thought to be alternative
  As events they are thought to be not compatible
- i.e. we would like to assume that
  $P(H\emptyset) = 1 - P(H1)$

$$\begin{cases} P(H\emptyset \wedge H1) = 0 \\ P(H\emptyset \vee H1) = 1 \end{cases}$$

- In the real set: it could well be that both $H\emptyset$ and $H1$ are true and it is just a matter of chance, insight, experience of the experimenter to isolate the good alternative hypotheses
- In the ordinary situation there is not such drama... testing of hypotheses is not a matter for geniuses but for professionals

To better understand hypothesis testing, consider the polio vaccine developed by Jonas Salk. In 1954, Salk's vaccine was tested on elementary-school students across the United States and Canada. In the study, 401,974 students were divided randomly into two groups: kids in one group received the vaccine, whereas those in the other group (the control group) were injected with saline solution instead. The students were unaware of which group they were in. Of those who received the vaccine, 0.016% developed paralytic polio during the study, whereas 0.057% of the control group developed the disease (Brownlee 1955). The vaccine seemed to reduce the rate of disease by two-thirds, but the difference between groups was quite small, only about four cases per 10,000. Did the vaccine work, or did such a small difference arise purely by chance?

Hypothesis testing uses probability to answer this question. The null hypothesis is that the vaccine didn't work, and that any observed difference between groups happened only by chance. Evaluating the null hypothesis involved calculating the probability, under the assumption that the vaccine has no effect, of getting a difference between groups as big or bigger than that observed. This probability turned out to be very small. Even though the rate of disease was not hugely different between the vaccine and control groups, the Salk vaccine trial was so large (over 400,000 participants) that it was able to demonstrate a real difference. Thus, the "null" hypothesis was rejected. The vaccine had an effect, sparing many kids from disease, which was borne out by the success of the vaccine in the ensuing decades.

Hypothesis testing compares data to what we would expect to see if a specific null hypothesis were true. If the data are too unusual, compared to what we would expect to see if the null hypothesis were true, then the null hypothesis is rejected.

Making and using statistical hypotheses

Formal hypothesis testing begins with clear statements of two hypotheses—the null and alternative hypotheses—about a population. The null hypothesis is the default, whereas the alternative hypothesis usually includes every other possibility except the one stated in the null hypothesis. One of the two hypotheses is true, and the other must be false. We analyze the data to help determine which is which.

Both statistical hypotheses, the null and the alternative, are simple statements about a population. They are not to be confused with scientific hypotheses, which are statements about the existence and possible causes of natural phenomena. Scientists design experiments and observational studies to test predictions of scientific hypotheses. When applied to the resulting data, statistical hypotheses help to decide which predictions of these scientific hypotheses are met and which are not met.

Slides from 3 to___from M.C. Whitlock and D. Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015)

Null hypothesis H_0

The **null hypothesis** is a specific claim about the value of a population parameter. It is made for the purposes of argument and often embodies the skeptical point of view. Often, the null hypothesis is that the population parameter of interest is zero (i.e., no effect, no preference, no correlation, or no difference). In general, the null hypothesis is a statement that would be interesting to *reject*. For example, if we can reject the statement, "Medication X does not affect the average life span of patients suffering from illness Y," then we have learned something useful—that such patients do in fact live longer—or shorter—lives on average when taking medication X. Rejecting the null hypothesis would provide support for the scientific hypothesis that predicted a beneficial effect of medication X, whereas failing to reject the null hypothesis would not provide support.

The ***null hypothesis*** is a specific statement about a population parameter made for the purposes of argument. A good null hypothesis is a statement that would be interesting to reject.

Alternative hypothesis

In general is a POSITIVE/interesting statement

Every null hypothesis is paired with an **alternative hypothesis** (abbreviated H_A) that usually represents all other feasible parameter values except that stated in the null hypothesis. The alternative hypothesis typically includes possibilities that are biologically more interesting than that stated in the null hypothesis. The alternative hypothesis often includes parameter values predicted by a scientific hypothesis being evaluated. For this reason the alternative hypothesis is often, but not always, the statement that the researcher hopes is true.

The ***alternative hypothesis*** includes all other feasible values for the population parameter besides the value stated in the null hypothesis.

Somehow the alternative hypothesis alludes to a symmetry breaking:
…Haha! There is something observable, in spite of the skepticism

To reject or not to reject

Crucially, null and alternative hypotheses do not have equal standing. The null hypothesis is the only statement being tested with the data. If the data are consistent with the null hypothesis, then we say we have failed to reject it (we never "accept" the null hypothesis). If the data are inconsistent with the null hypothesis, we reject it and say the data support the alternative hypothesis.

Rejecting H_0 means that we have ruled out the null hypothesized value. It also tells us in which direction the true value likely lies, compared to the null hypothesized value. But rejecting a hypothesis by itself reveals nothing about the magnitude of the population parameter. We use estimation to provide magnitudes.

Hypothesis testing: an example

To show you the basic concepts and terminology of hypothesis testing, we'll take you through all the steps by using an example. Our goal is to illuminate the basic process without distraction from the details of the probability calculations. We'll get to plenty of the details in later chapters.

Four basic steps are involved in hypothesis testing:

1. State the hypotheses.
2. Compute the test statistic.
3. Determine the *P*-value.
4. Draw the appropriate conclusions.

We'll define the new terms we just used in this section.

Example 6.2 tests a hypothesis about a proportion, but hypothesis testing can address a wide variety of quantities, such as means, variances, differences in means, correlations, and so on. We'll try to emphasize the general over the specific here. Further details of how to test hypotheses about proportions are discussed in Chapter 7.

EXAMPLE 6.2 The right hand of toad

Humans are predominantly right-handed. Do other animals exhibit handedness as well? Bisazza et al. (1996) tested the possibility of handedness in European toads, *Bufo bufo*, by sampling and measuring 18 toads from the wild. We will assume that this was a random sample. The toads were brought to the lab and subjected one at a time to the same indignity: a balloon was wrapped around each individual's head. The researchers then recorded which forelimb each toad used to remove the balloon. It was found that individual toads tended to use one forelimb more than the other. At this point the question became: do right-handed and left-handed toads occur with equal frequency in the toad population, or is one type more frequent than the other, as in the human population?

Hintau Aliaksei/Shutterstock

Of the 18 toads tested, 14 were right-handed and four were left-handed. Are these results evidence of a predominance of one type of handedness in toads?

Stating the hypotheses

The number of interest is the proportion of right-handed toads in the *population*. Let's call this proportion p. The default statement, the null hypothesis, is that the two types of handedness are equally frequent in the population, in which case $p = 0.5$.

H_0: Left- and right-handed toads are *equally frequent* in the population (i.e., $p = 0.5$).

This is a specific statement about the state of the toad population, one that would be interesting to prove wrong. If this null hypothesis is wrong, then toads, like humans, on average favor one hand over the other. This statement establishes the alternative hypothesis:

H_A: Left- and right-handed toads are *not equally frequent* in the population (i.e., $p \neq 0.5$).

The alternative hypothesis is **two-sided.** This just means that the alternative hypothesis allows for two possibilities: that p is greater than 0.5 (in which case right-handed toads outnumber left-handed toads in the population), or that p is less than 0.5 (i.e., left-handed toads predominate). Neither possibility can be ruled out before gathering the data, so both should be included in the alternative hypothesis.

In a ***two-sided*** (or two-tailed) test, the alternative hypothesis includes parameter values on both sides of the parameter value specified by the null hypothesis.

The test statistic

The **test statistic** is a number calculated from the data that is used to evaluate how compatible the results are with those expected under the null hypothesis.

The ***test statistic*** is a number calculated from the data that is used to evaluate how compatible the data are with the result expected under the null hypothesis.

For the toad study, we use the observed number of right-handed toads as our test statistic. On average, if the null hypothesis were correct, we would expect to observe nine right-handed toads out of the 18 sampled (and nine left-handed toads, too). Instead, we observed 14 right-handed toads out of the 18 sampled. Fourteen, then, is the value of our test statistic.

The null distribution

Unfortunately, data do not always perfectly reflect the truth. Because of the effects of chance during sampling, we don't really expect to see exactly nine right-handed toads when we sample 18 from the population, even if the null hypothesis is true. There is usually a discrepancy, due to chance, between the observed result and that expected under H_0. The mismatch between the data and the expectation under H_0 can be quite large, even when H_0 is true, particularly if there are not many data. To decide whether the data are compatible with the null hypothesis, we must calculate the probability of a mismatch as extreme as or more extreme than that observed, assuming that the null hypothesis is true.

To obtain this probability, we need to determine the sampling distribution of the test statistic *assuming that the null hypothesis is true*. We need to determine what values of the test statistic are possible under H_0 and their associated probabilities. The probability distribution of values for the test statistic, assuming the null hypothesis is true, is called the "sampling distribution under H_0" or, more simply, the **null distribution.**

The ***null distribution*** is the sampling distribution of outcomes for a test statistic under the assumption that the null hypothesis is true.

This null histogram/distribution has been generated by simulating on a computer a large series of tosses of 18 fair coins and binning the number of heads (right-handed toads)

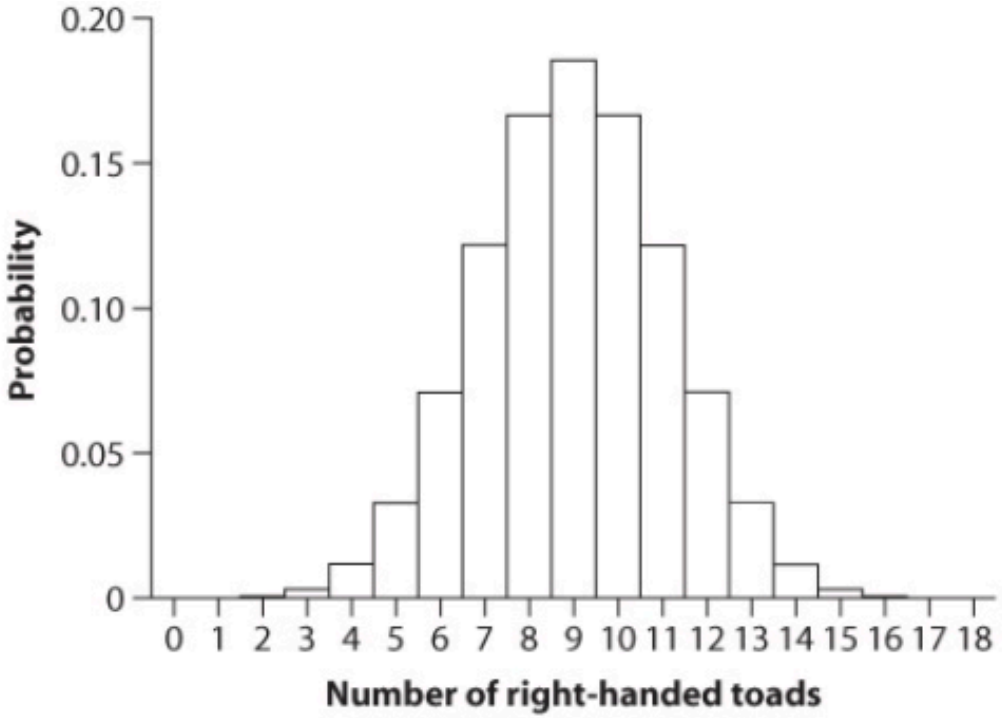


Figure 6.2-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 6.2-1 The null distribution for the test statistic, the number of right-handed toads out of 18 sampled.

This histogram distribution has been computed assuming a Binomial probabilistic model for H0 see chap 7 in WS
THE TEST STATISTIC in this case is 14

Quantifying uncertainty: the *P*-value

The probability of obtaining the data (or data that are an even worse match to the null hypothesis), assuming the null hypothesis, is called the ***P*-value**. If the *P*-value is small, then the null hypothesis is inconsistent with the data and we reject it.[1] Otherwise, we do not reject the null hypothesis. In general, the smaller the *P*-value, the stronger is the evidence against the null hypothesis.

> The ***P*-value** is the probability of obtaining the data (or data showing as great or greater difference from the null hypothesis) if the null hypothesis were true.

The *P*-value is *not* the probability that the null hypothesis is true. (Hypotheses are not outcomes of random trials and so do not have probabilities.) The *P*-value refers to the probability of a specific event when sampling data under the null hypothesis: it is the probability of obtaining a result as extreme as or more extreme than that observed.

In practice, we calculate the *P*-value from the null distribution for the test statistic, shown for the toad data in Figure 6.2-2.

In this example (two sided hypothesis) the red area sum up to 0.031

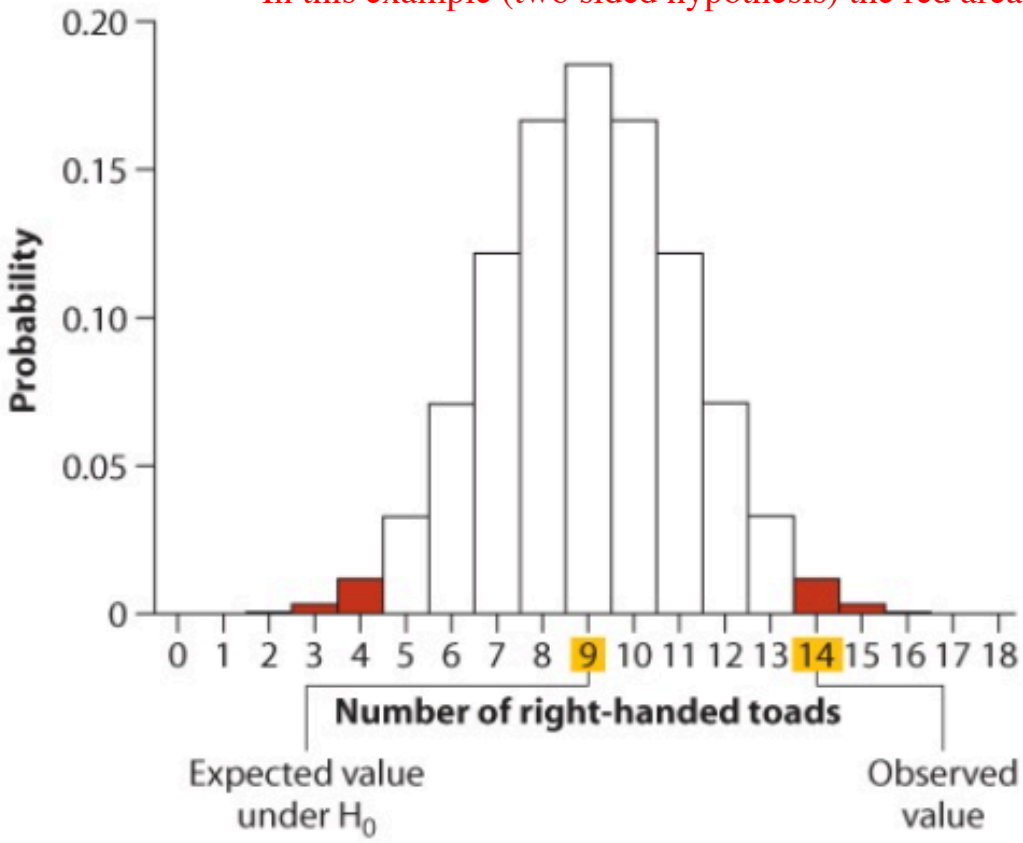


Figure 6.2-2
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 6.2-2 The null distribution for the number of right-handed toads out of the 18 sampled. Outcomes in red are values as different as, or more different from, the expectation under H_0 than 14, the number observed in the data.

IN DETAIL: EVALUATE THE P-Value

Based on the data in Figure 6.2-2, the probability of 14 or more right-handed toads, assuming the null hypothesis is true, is

Pr[14 or more right-handed toads]=Pr[14]+Pr[15]+Pr[16]+Pr[17]+Pr[18]=0.0155,

$$\begin{aligned}\Pr[\textit{14 or more right-handed toads}] &= \Pr[14] + \Pr[15] + \Pr[16] + \Pr[17] + \Pr[18] \\ &= 0.0155,\end{aligned}$$

where Pr[14] is the probability of exactly 14 right-handed toads. We can add the probabilities of 14, 15, 16, 17, and 18 because each outcome is mutually exclusive. This sum is not the *P*-value, though, because it does not yet include the equally extreme results at the left tail of the null distribution—that is, those outcomes involving a predominance of left-handed toads. The quickest way to include the probabilities of the equally extreme results at the other tail is to take the above sum and multiply by two:

P=2×(Pr[14]+Pr[15]+Pr[16]+Pr[17]+Pr[18])=2×0.0155=0.031.

$$\begin{aligned}P &= 2 \times (\Pr[14] + \Pr[15] + \Pr[16] + \Pr[17] + \Pr[18]) \\ &= 2 \times 0.0155 \\ &= 0.031.\end{aligned}$$

This number is our *P*-value. In other words, the probability of an outcome as extreme as or more extreme than 14 right-handed toads out of 18 toads sampled is $P = 0.031$, assuming that the null hypothesis is true.

Draw the appropriate conclusion

Having calculated the *P*-value, what conclusion can we draw from it? On page 157, we said that if *P* is "small," we reject the null hypothesis; otherwise, we do not reject H_0. But what value of *P* is small enough? By convention in most areas of biological research, the boundary between small and not-small *P*-values is 0.05. That is, if *P* is less than or equal to 0.05, then we reject the null hypothesis; if $P > 0.05$, we do not reject it.

The *P*-value for the toad data, $P = 0.031$, is indeed less than 0.05, so we reject the null hypothesis that left-handed and right-handed toads are equally frequent in the toad population. We conclude from these data that most of the toads in the population are right-handed.

This decision threshold for *P* (i.e., $P = 0.05$) is called the **significance level**, which is signified by α (the lowercase Greek letter alpha). In biology, the most widely used significance level is $\alpha = 0.05$, but you will encounter some studies that use a different value for α. After $\alpha = 0.05$, the next most commonly used signifi-cance level is $\alpha = 0.01$. In Section 6.3, we explain the consequences of choosing a significance level and consider why $\alpha = 0.05$ is the most common choice.[2]

The ***significance level, α***, is a probability used as a criterion for rejecting the null hypothesis. If the *P*-value is less than or equal to α, then the null hypothesis is rejected. If the *P*-value is greater than α, then the null hypothesis is *not* rejected.

Type I and Type II errors

There are two kinds of errors in hypothesis testing, prosaically named Type I and Type II.

Rejecting a true null hypothesis is a **Type I error**. Failing to reject a false null hypothesis is a **Type II error**. Both types of error are summarized in Table 6.3-1.

Type I error is rejecting a true null hypothesis. The significance level α sets the probability of committing a Type I error.

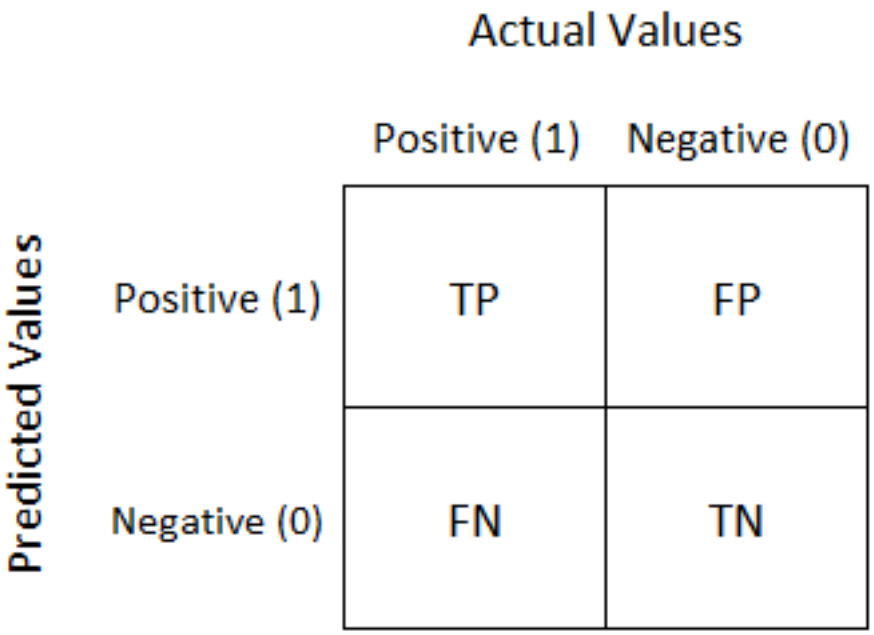
Type II error is failing to reject a false null hypothesis.

TABLE 6.3-1 Types of error in hypothesis testing.

		Reality
Conclusion	H_0 true	H_0 false
Reject H_0	Type I error	Correct
Do not reject H_0	Correct	Type II error

The significance level, α, gives us the probability of committing a Type I error. If we go along with convention and use a significance level of $\alpha = 0.05$, then we reject H_0 whenever P is less than or equal to 0.05. This means that, if the null hypothesis were true, we would reject it mistakenly one time in 20. Biologists typically regard this as an acceptable error rate.

The confusion matrix associated to a binary classification problem (https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62)



T. Fawcett / Pattern Recognition Letters 27 (2006) 861–874

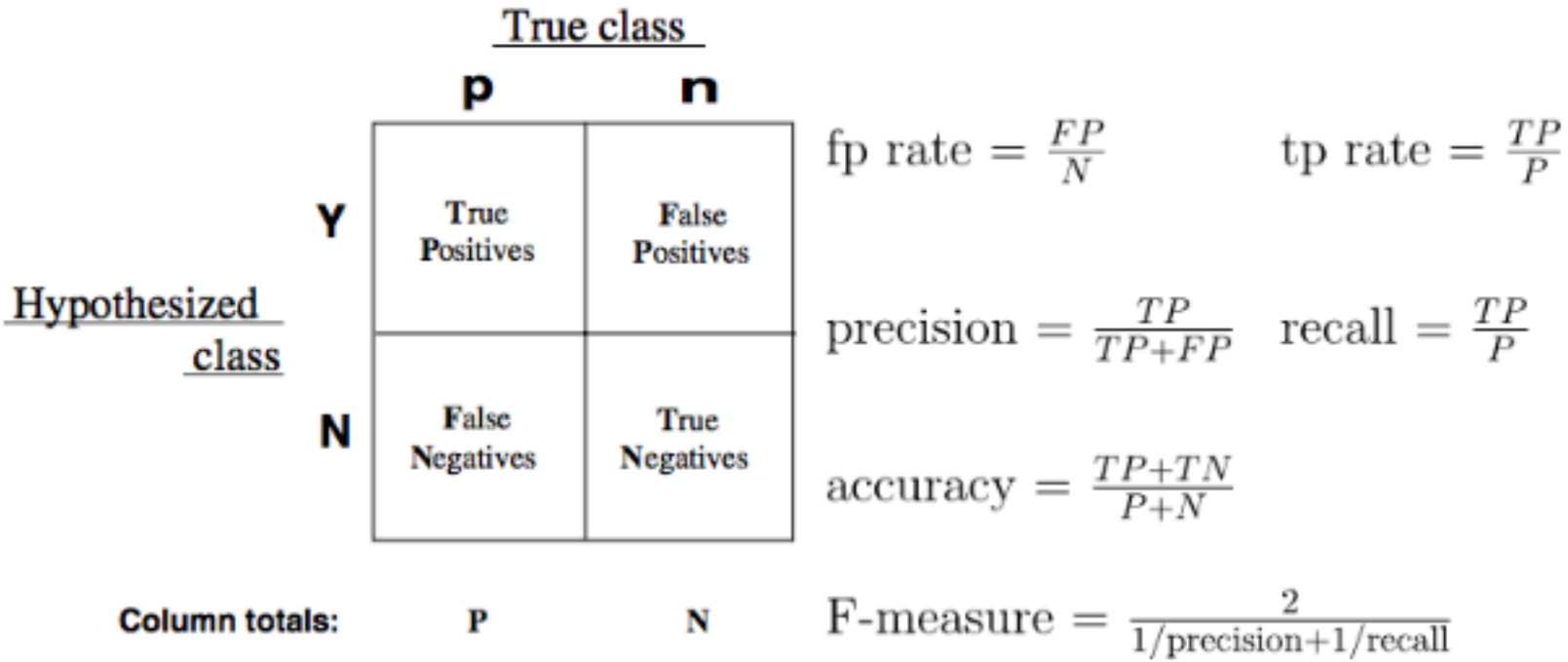
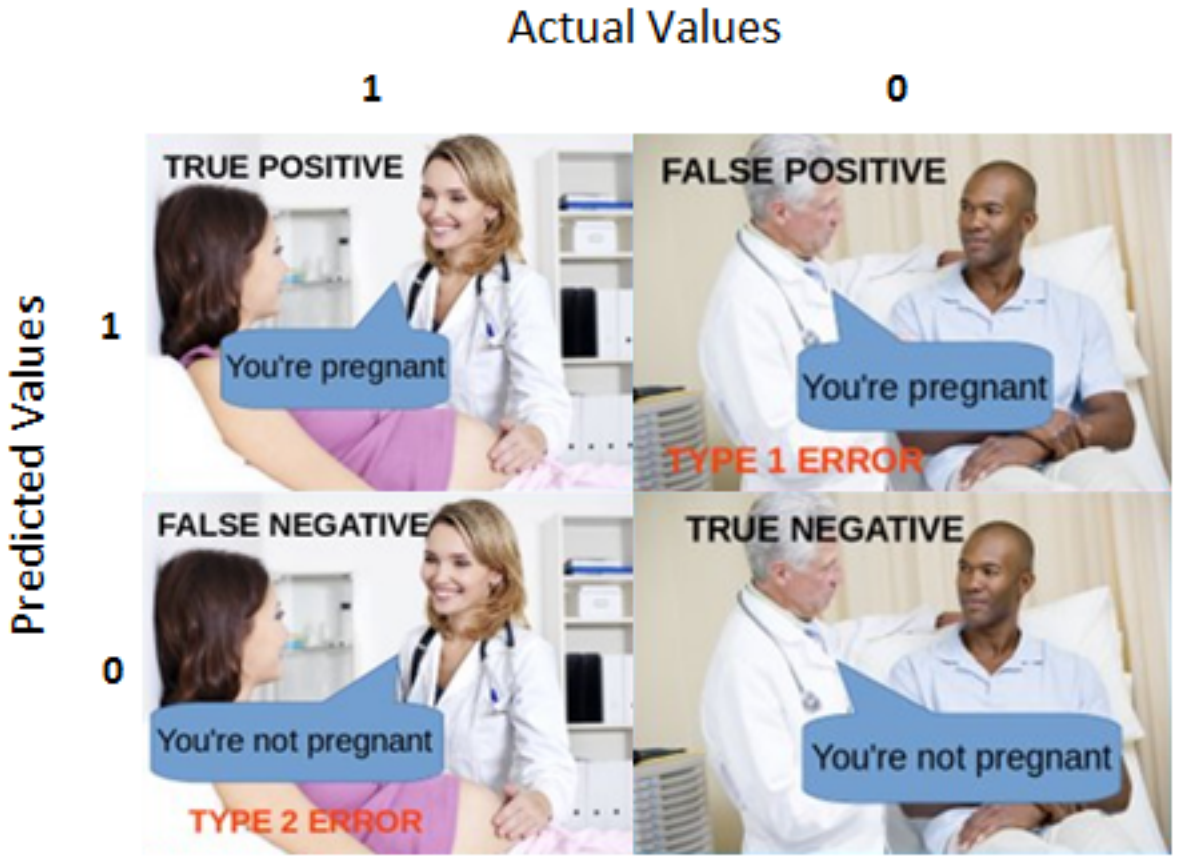


Fig. 1. Confusion matrix and common performance metrics calculated from it.

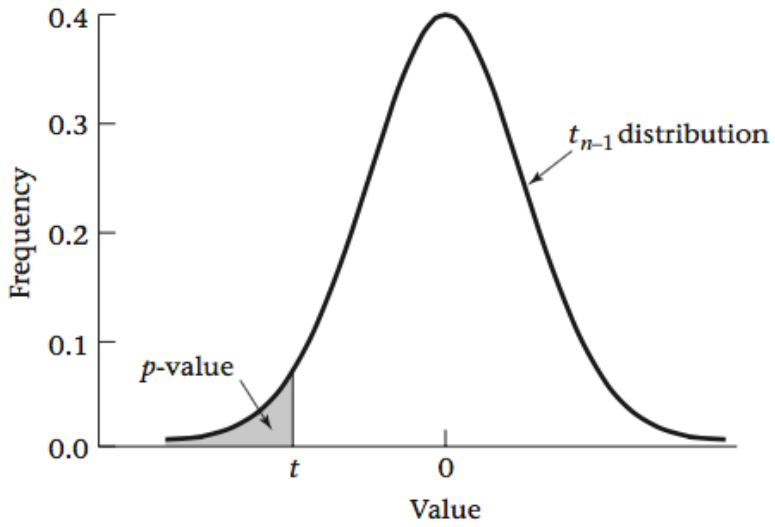

Actual Values
1
0
Predicted Values
1
0
TRUE POSITIVE
You're pregnant
FALSE POSITIVE
You're pregnant
TYPE 1 ERROR
FALSE NEGATIVE
You're not pregnant
TYPE 2 ERROR
TRUE NEGATIVE
You're not pregnant

True Positive:
Interpretation: You predicted positive and it's true.
You predicted that a woman is pregnant and she actually is.
True Negative:
Interpretation: You predicted negative and it's true.
You predicted that a man is not pregnant and he actually is not.
False Positive: (Type 1 Error)
Interpretation: You predicted positive and it's false.
You predicted that a man is pregnant but he actually is not.
False Negative: (Type 2 Error)
Interpretation: You predicted negative and it's false.
You predicted that a woman is not pregnant but she actually is.

DEFINITION 7.13 The ***p*-value** for any hypothesis test is the α level at which we would be indifferent between accepting or rejecting H_0 given the sample data at hand. That is, the *p*-value is the α level at which the given value of the test statistic (such as t) is on the borderline between the acceptance and rejection regions.

DEFINITION 7.14 The ***p*-value** can also be thought of as the probability of obtaining a test statistic as extreme as or more extreme than the actual test statistic obtained, given that the null hypothesis is true.

Graphic display of a *p*-value



We know that under the null hypothesis, the t statistic follows a t_{n-1} distribution. Hence, the probability of obtaining a t statistic that is no larger than t under the null hypothesis is $Pr(t_{n-1} \leq t) = p$-value, as shown in Figure 7.1.

EQUATION 7.4

Guidelines for Judging the Significance of a *p*-Value

If $.01 \le p < .05$, then the results are *significant.*
If $.001 \le p < .01$, then the results are *highly significant.*
If $p < .001$, then the results are *very highly significant.*
If $p > .05$, then the results are considered *not statistically significant* (sometimes denoted by NS).
However, if $.05 < p < .10$, then a trend toward statistical significance is sometimes noted.

EQUATION 7.5

Determination of Statistical Significance for Results from Hypothesis Tests

Either of the following methods can be used to establish whether results from hypothesis tests are statistically significant:

(1) The test statistic t can be computed and compared with the critical value $t_{n-1,\,\alpha}$ at an α level of .05. Specifically, if $H_0: \mu = \mu_0$ vs. $H_1: \mu < \mu_0$ is being tested and $t < t_{n-1,.05}$, then H_0 is rejected and the results are declared *statistically significant* ($p < .05$). Otherwise, H_0 is accepted and the results are declared *not statistically significant* ($p \ge .05$). We have called this approach the **critical-value method** (see Definition 7.12).

(2) The exact p-value can be computed and, if $p < .05$, then H_0 is rejected and the results are declared *statistically significant.* Otherwise, if $p \ge .05$, then H_0 is accepted and the results are declared *not statistically significant.* We will refer to this approach as the ***p*-value method.**