

# Basic Statistics and Probability

## Chapter 8: Inferences Based on a Single Sample: Tests of Hypothesis

- ▶ Elements of a Test of Hypothesis
- ▶ Hypothesis Tests in Normal Populations
- ▶ Tests in Non-Normal Populations
- ▶ The  $p$ -Value
- ▶ Some Remarks

# Elements of a Test of Hypothesis

## Example 8.1: Chips Ahoy. Source: DASL

In 1998, as an advertising campaign, the Nabisco Company announced a “1000 Chips Challenge”, claiming that every 18-ounce bag of their Chips Ahoy! cookies contained at least 1000 chocolate chips. Dedicated statistics students at the Air Force Academy randomly selected bags of cookies and counted the chocolate chips. The data report their counts.

1219	1214	1087	1200
1419	1121	1325	1345
1244	1258	1356	1132
1191	1270	1295	1135

Data and description at:

<https://dasl.datadescription.com/datafile/chips-ahoy>

## Example 8.1 (Chips Ahoy)

$X$  = number of chocolate chips in an 18-oz bag of Chips Ahoy!

$\mu = E(X) =$   $\begin{matrix} \text{expected} \\ \text{average} \end{matrix}$  number of chocolate chips in a bag

We want to test which of the following two hypotheses is true:

- $\mu < 1000$  (the claim of the company is not true)
- $\mu \geq 1000$  (the claim of the company is true)

A **statistical hypothesis** is a statement about the numerical value of a population parameter.

There are just two (mutually exclusive) hypotheses in a statistical hypothesis test. Typically one hypothesis will establish a range of possible values for the parameter ( $[1000, \infty)$ ) and the other hypothesis will contain the rest of the possible values ( $[0, 1000)$ )

One of the hypotheses is called **null hypothesis** and denoted by  $H_0$ .

The null hypothesis  $H_0$  represents the statement that will be assumed to be true unless the data provide convincing evidence that it is false.

The other hypothesis is the **alternative (or research) hypothesis** and it is denoted by  $H_1$  (or also by  $H_a$ ).

The alternative hypothesis  $H_1$  represents the statement that will be accepted only if the data provide convincing evidence of its truth. This usually represents the values of a population parameter for which the researcher wants to gather evidence to support.

### **Example 8.1 (Chips Ahoy)**

Intuitively...

To decide that there is enough evidence supporting the alternative hypothesis we construct a **test statistic**, which measures the “discrepancy” between the sample information and the null hypothesis. When the test statistic is “significantly large”, we reject the null hypothesis.

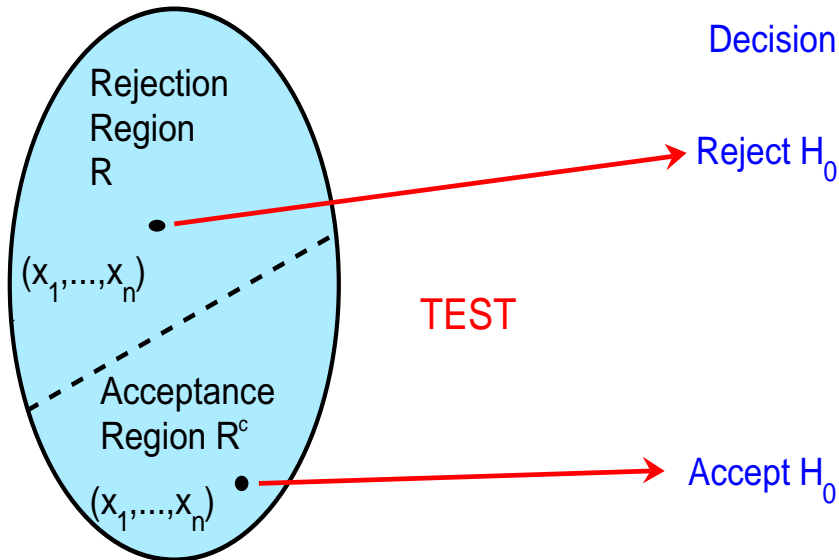
### **Example 8.1 (Chips Ahoy)**

Rigorously...

The test statistic is used to construct an automatised decision rule, such that the decision of accepting  $H_0$  or rejecting  $H_0$  is assigned to every possible value of a sample  $x_1, \dots, x_n$  of  $X$ .

Such an decision rule is given by the **rejection region**  $R$ . The rejection region is usually given in terms of the test statistic.

# Sampling Space



To obtain an explicit mathematical expression for the rejection region  $R$ , we first study the criteria and conditions that  $R$  has to fulfil.

Possible errors in a hypothesis test:

**Type I error:** Reject  $H_0$  when, in fact,  $H_0$  is true.

**Type II error:** Accept  $H_0$  when  $H_0$  is false.

How the region  $R$  is constructed:

**1. Bound the maximum probability of Type I error.**

- Fix a **significance level**  $\alpha \in (0, 1)$ . Typically, by default,  $\alpha = 0.05$ .
- Define the **size** of a hypothesis test as the maximum probability of Type I error:  $\max P(\text{Error of Type I}) = \max P(R|H_0 \text{ true})$ .
- Look for  $R$  such that  $\max P(\text{Error of Type I}) \leq \alpha$ .

**2. Minimize the maximum probability of Type II error.**

Of all the possible rejection regions satisfying 1, we try to choose one with low probability of Type II error.

## Example 8.1 (Chips Ahoy)

$H_0$  :  $\mu \geq 1000$  (the claim of the company is true)

$H_1$  :  $\mu < 1000$  (the claim of the company is not true)

$$P(\text{Error of Type I}) = P_{\mu}(R) \quad \text{for } \mu \geq 1000$$

$$\max P(\text{Error of Type I}) = \max_{\mu \geq 1000} P_{\mu}(R) \leq \alpha$$

$$P(\text{Error of Type II}) = P_{\mu}(R^c) = \beta(\mu) \quad \text{for } \mu < 1000$$

Table 8.1 Conclusions and Consequences for a Test of Hypothesis		
Conclusion	True State of Nature	
	$H_0$ True	$H_a$ True
Accept $H_0$ (Assume $H_0$ True)	Correct decision	Type II error (probability $\beta$ )
Reject $H_0$ (Assume $H_a$ True)	Type I error (probability $\alpha$ )	Correct decision



We bound by  $\alpha$  the probability of rejecting  $H_0$  when it is true.



We are “conservative” with the null hypothesis  $H_0$ :  
we only reject it when there is a lot of sample evidence against it.



In the alternative hypothesis  $H_1$  we state the result for which  
we need or want a lot of sample evidence before accepting it.

If the condition in the mathematical expression of the rejection region  $R$  holds, then we say that there is enough statistical or sample evidence to reject the null hypothesis  $H_0$ . We conclude that the alternative hypothesis is true.

The hypothesis-testing process will lead to this conclusion incorrectly (Type I error) only  $100\alpha\%$  of the time when  $H_0$  is true.

If the condition of the rejection region  $R$  is not fulfilled, we do not reject  $H_0$ . We do not conclude that the null hypothesis is true. We just say that there is not enough evidence to reject  $H_0$ .

The methodology for testing hypotheses varies depending on the target population parameter. In this chapter, we develop methods for testing a population mean, a population proportion, and (optionally) a population variance.

### Determining the Target Parameter

Parameter	Key Words or Phrases	Type of Data
$\mu$	Mean; average	Quantitative
$p$	Proportion; percentage; fraction; rate	Qualitative
$\sigma^2$	Variance; variability; spread	Quantitative

# Hypothesis Tests in Normal Populations

## Tests on the expectation $\mu$ of a normal population

Let  $x_1, \dots, x_n$  be a sample from  $X \sim N(\mu, \sigma)$  with  $\sigma$  unknown.

Possible tests on  $\mu$  (*t-tests*):

$$\begin{array}{ll} \text{Two-tailed test} & \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array} & R = \{(x_1, \dots, x_n) : |t| \geq t_{n-1; \alpha/2}\} \end{array}$$

$$\begin{array}{ll} \text{One-tailed test} & H_0 : \mu \leq \mu_0 \\ \text{(upper-tailed)} & H_1 : \mu > \mu_0 & R = \{(x_1, \dots, x_n) : t \geq t_{n-1; \alpha}\} \end{array}$$

$$\begin{array}{ll} \text{One-tailed test} & H_0 : \mu \geq \mu_0 \\ \text{(lower-tailed)} & H_1 : \mu < \mu_0 & R = \{(x_1, \dots, x_n) : t \leq t_{n-1; 1-\alpha}\} \end{array}$$

where  $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$  is the test statistic.

## Example 8.1 (Chips Ahoy)

### Example 8.2: (Road Paint)

Paint used to paint lines on roads must reflect enough light to be clearly visible at night. Let  $\mu$  denote the mean reflectometer reading for a new type of paint under consideration. A test of  $H_0 : \mu \leq 20$  versus  $H_1 : \mu > 20$  based on a sample of 15 observations gave  $t = 3.2$ . What conclusion is appropriate at each of the following significance levels?

$$\alpha = 0.05 \quad \alpha = 0.01 \quad \alpha = 0.001$$

How to do a  $t$ -test with R:

### Example 8.1 (Chips Ahoy)

```
NumChips = scan("chips-ahoy.txt",skip=1)
```

Read 16 items

```
t.test(x=NumChips,mu=1000,alternative="less")
```

One Sample  $t$ -test

data: NumChips

$t = 10.105$ ,  $df = 15$ ,  $p\text{-value} = 1$

alternative hypothesis: true mean is less than 1000

95 percent confidence interval:

-Inf 1279.508

sample estimates:

mean of  $x$

1238.188

How strange! R does not give a solution to the test. Or it does?  
Let's introduce the  $p$ -value concept (page 20).

# Tests in Non-Normal Populations

## Large-sample tests for the expectation of any distribution

Let  $x_1, \dots, x_n$  be a sample of size  $n \geq 20$  from a r.v.  $X$  with  $E(X) = \mu$ .

Using the Central Limit Theorem we have the following **z-tests**:

$$\begin{array}{ll} H_0 : \mu = \mu_0 & R = \{(x_1, \dots, x_n) : |z| > z_{\alpha/2}\} \\ H_1 : \mu \neq \mu_0 & \end{array}$$

$$\begin{array}{ll} H_0 : \mu \leq \mu_0 & R = \{(x_1, \dots, x_n) : z > z_{\alpha}\} \\ H_1 : \mu > \mu_0 & \end{array}$$

$$\begin{array}{ll} H_0 : \mu \geq \mu_0 & R = \{(x_1, \dots, x_n) : z < -z_{\alpha}\} \\ H_1 : \mu < \mu_0 & \end{array}$$

where  $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$  is the test statistic.

**Example 8.3:** Test the hypothesis that a population has a mean blood glucose level of 100. Suppose we select a random sample of 30 individuals from this population ( $\bar{x} = 98.1$ ,  $s^2 = 126$ ).



## Large-sample tests on a proportion $p$

Let  $x_1, \dots, x_n$  be a sample of size  $n \geq 20$  from a r.v.  
 $X \sim \text{Bernoulli}(p)$ .

Using the Central Limit Theorem we have the following **z-tests**:

$$\begin{aligned} H_0 : p &= p_0 & R &= \{|z| > z_{\alpha/2}\} \\ H_1 : p &\neq p_0 \end{aligned}$$

$$\begin{aligned} H_0 : p &\leq p_0 & R &= \{z > z_{\alpha}\} \\ H_1 : p &> p_0 \end{aligned}$$

$$\begin{aligned} H_0 : p &\geq p_0 & R &= \{z < z_{1-\alpha}\} \\ H_1 : p &< p_0 \end{aligned}$$

where  $z = \frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$  is the test statistic.

### **Example 8.4: : Most Like It Hot**

“Most Like it Hot” is the title of a press release issued by the Pew Research Center (March 18, 2009, [www.pewsocialtrends.org](http://www.pewsocialtrends.org)). The press release states that “by an overwhelming margin, Americans want to live in a sunny place.” This statement is based on data from a nationally representative sample of 2260 adult Americans. Of those surveyed, 1288 indicated that they would prefer to live in a hot climate rather than a cold climate. Do the sample data provide convincing evidence that a majority of all adult Americans prefer a hot climate over a cold climate?

## Example 8.4: Most Like It Hot

```
prop.test(1288,2260,p=0.5,alternative="greater",correct=FALSE)
```

1-sample proportions test without continuity correction

```
data: 1288 out of 2260, null probability 0.5
X-squared = 44.1841, df = 1, p-value = 1.495e-11
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.552708 1.000000
sample estimates:
      p
0.5699115
```

The *continuity correction* adds an extra term to the test statistic to correct the error attained when approximating a discrete distribution by a continuous one. The correction adjusts the probability of Type I error (which “inflates” when employing the normal approximation with smaller sample sizes). For instance, in the case of the test  $H_0 : p = p_0$ , the rejection region is  $R = \{|z| > z_{\alpha/2}\}$  where the test statistic is

without correction

$$Z = \frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

with correction

$$Z = \frac{|\bar{x} - p_0| - \frac{1}{2n}}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

If the significance level  $\alpha$  of a test decreases, then we are more conservative with  $H_0$ .

Given a hypothesis test and a specific sample, the **p-value of the test** is **the lowest of the  $\alpha$  values for which the null hypothesis  $H_0$  is rejected.**

The observed significance level, or p-value, also coincides with the probability (assuming  $H_0$  is true) of observing a value of the test statistic that is at least as contradictory to the null hypothesis, and supportive of the alternative hypothesis, as the actual one computed from the sample data.

The lower is the p-value the more evidence in favour of  $H_1$ .

Statistical computer software (R, SPSS, Excel, Matlab, ...) does not carry out a test for a specified value of  $\alpha$ , but gives us just the p-value.

### Example 8.5:

A laboratory quality assurance study was carried out to look for signs of systematic bias in a lab's method for measuring total organic carbon (TOC), a measure of water quality. Certified standard solutions having 50 mg/L TOC were randomly inserted into the lab's normal work stream. The lab analysts were unaware of the presence of these standard solutions.

Here are the measurement readings for  $n = 16$  of the standard solutions inserted into the lab's work stream:

50.3	51.2	50.5	50.2	49.9	50.2	50.3	50.5
49.3	50.0	50.4	50.1	51.0	49.8	50.7	50.6

If there was no systematic bias, a 50 mg/L standard solution would, on average, give a 50 mg/L measurement reading. Test for the existence of bias when  $\alpha = 0.1$ ,  $\alpha = 0.05$  and  $\alpha = 0.01$ .

## Remark 1:

In the McClave and Sincich manual and other books the following notation is used for one-tailed tests:

$$\begin{array}{ll} H_0 : \mu = \mu_0 & H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 & H_1 : \mu < \mu_0 \end{array}$$

instead of the one used in these notes (and many other books)

$$\begin{array}{ll} H_0 : \mu \leq \mu_0 & H_0 : \mu \geq \mu_0 \\ H_1 : \mu > \mu_0 & H_1 : \mu < \mu_0 \end{array}$$

## Remark 2:

What can be done to carry out a test on  $\mu$  if the population histogram departs greatly from normal and the sample size is “small” ( $n < 20$ )? Use one of the nonparametric statistical methods of Chapter 14.