4.4 Confidence Intervals for Comparing Means and Proportions of Two Populations

Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, relative to two populations, with means $\mu_1 = E(X_{(1)}), \mu_2 = E(X_{(2)})$ and variances $\sigma_1^2 = V(X_{(1)}), \sigma_2^2 = V(X_{(2)})$, respectively.
We draw from both populations random samples of sizes n_1 and n_2, respectively, that are **independent**. Denote the two sets of random variables by $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$. Then we have two sample means and two sample variances, given by

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}, \quad \overline{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}$$

and

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_{1i} - \overline{X}_1\right)^2, \;\; s_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2,$$

respectively. In addition, denote by

$$s_p^2 = \frac{\displaystyle\sum_{i=1}^{n_1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

the *pooled variance* of the two samples, i.e. a variance that considers ("pools") the sample data from both samples.

Recall that when comparing the means of two populations, we estimate their difference. The formulas for finding confidence intervals for the difference of means $\mu_1 - \mu_2$ are based on the following results.

Proposition 4.1. *Assume $X_{(1)} \in N(\mu_1, \sigma_1)$ and $X_{(2)} \in N(\mu_2, \sigma_2)$ or that the samples are large enough ($n_1 + n_2 > 40$). Then*

a) $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \in N(0,1)$; b) $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \in T(n_1 + n_2 - 2)$;

c) $T^* = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \in T(n)$, *where* $\dfrac{1}{n} = \dfrac{c^2}{n_1 - 1} + \dfrac{(1-c)^2}{n_2 - 1}$ *and* $c = \dfrac{\dfrac{s_1^2}{n_1}}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$.

CI for the difference of means

Case σ_1, σ_2 known

If either $X_{(1)} \in N(\mu_1, \sigma_1),\ X_{(2)} \in N(\mu_2, \sigma_2)$ or the samples are large enough ($n_1 + n_2 > 40$) and σ_1, σ_2 are known, then by Proposition 4.1, we can use the pivot

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in N(0, 1).$$

With the same line of computations as before, we find a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ as

$$\mu_1 - \mu_2 \ \in \ \left[\overline{X}_1 - \overline{X}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \overline{X}_1 - \overline{X}_2 - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right], \tag{4.1}$$

or, using symmetry,

$$\left[\overline{X}_1 - \overline{X}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]. \tag{4.2}$$

Case $\sigma_1 = \sigma_2$ unknown

Assume that either $X_{(1)} \in N(\mu_1, \sigma_1),\ X_{(2)} \in N(\mu_2, \sigma_2)$ or the samples are large enough ($n_1 + n_2 > 40$). The population variances are *not* known anymore, but they are known to be equal. Then each is approximated by the pooled variance s_p^2. Then by Proposition 4.1, we use the pivot

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \in T(n_1 + n_2 - 2).$$

A $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is given by

$$\mu_1 - \mu_2 \ \in \ \left[\overline{X}_1 - \overline{X}_2 - t_{1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \overline{X}_1 - \overline{X}_2 - t_{\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right], \tag{4.3}$$

where the quantiles $t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}$ refer to the $T(n_1 + n_2 - 2)$ distribution. Again, by symmetry we can

write the CI in short as

$$\left[\overline{X}_1 - \overline{X}_2 \pm t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \tag{4.4}$$

Case σ_1, σ_2 unknown

Assuming that either $X_{(1)} \in N(\mu_1, \sigma_1),\ X_{(2)} \in N(\mu_2, \sigma_2)$ or the samples are large enough ($n_1 + n_2 > 40$), by Proposition 4.1, we use the pivot

$$T^* = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \in T(n),$$

where

$$\frac{1}{n} \;=\; \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1} \quad \text{and} \quad c \;=\; \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{4.5}$$

We find a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ as

$$\mu_1 - \mu_2 \;\in\; \left[\overline{X}_1 - \overline{X}_2 - t_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ \overline{X}_1 - \overline{X}_2 - t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right], \tag{4.6}$$

or, by symmetry,

$$\left[\overline{X}_1 - \overline{X}_2 \pm t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] \tag{4.7}$$

where the quantile $t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}$ refer to the $T(n)$ distribution, with n given above.

Example 4.2. An account on server A is more expensive than an account on server B. However, server A is faster. To see if it's optimal to go with the faster but more expensive server, a manager needs to know how much faster it is. A certain computer algorithm is executed 30 times on server A and 20 times on server B with the results given below. Find a 95% confidence interval for the

difference $\mu_1 - \mu_2$ between the mean execution times on server A and server B.

Server A			Server B		
n_1	$=$	30	n_2	$=$	20
$\overline{X}_1$	$=$	6.7 min	$\overline{X}_2$	$=$	7.5 min
s_1	$=$	0.6 min	s_2	$=$	1.2 min

Solution. The samples are large enough ($n_1 + n_2 = 50$), that we can use Proposition 4.1. Nothing is said about the population variances (that they might be known, or known to be equal). Also, the second sample standard deviation is twice as large as the first one, therefore, equality of population variances can hardly be assumed. We use the general case for unknown, unequal variances and use formula (4.7).

We want confidence level $1 - \alpha = 0.95$, so $\alpha = 0.05$ and $\alpha/2 = 0.025$. The parameter n in (4.5) is found to be $n = 25.3989 \approx 25$. For the $T(25)$ distribution, we find the quantile $t_{0.025} = -2.0595$. Them the 95% CI for the difference of means is

$$\left[\overline{X}_1 - \overline{X}_2 \pm t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] = \left[6.7 - 7.5 \pm 2.06\sqrt{\frac{0.6^2}{30} + \frac{1.2^2}{20}}\right] = [-0.8 \pm 0.505],$$

so,

$$\mu_1 - \mu_2 \in [-1.305,\ -0.295]$$

with probability 0.95. Since *all* values in the CI are negative, with high probability, it seems that $\mu_1 - \mu_2 < 0$, so indeed the first server seems to be faster, on average.

■

CI for the difference of means, paired data

In many applications, we want to compare the means of two populations, when two random samples (one from each population) are available, which *are not* independent, where each observation in one sample is naturally or by design *paired* with an observation in the other sample. As examples, consider:

− comparing average values of the same measurements made using two different devices,

− compare the health of the same group of patients in response to a certain treatment,

− compare the behaviour of some equipment under different temperature/pressure conditions, etc.

These are usually cases best described as "before and after" situations.

In such cases, both samples have the same length, n: $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2n}$. Then we consider the sample of their *differences*, $D_1, \ldots, D_n$, where $D_i = X_{1i} - X_{2i},\ i = \overline{1,n}$.
For this sample, we have

$$\begin{aligned} \overline{X}_d &= \frac{1}{n}\sum_{i=1}^{n} D_i, \text{ the sample mean and} \\ s_d^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \overline{X}_d\right)^2, \text{ the sample variance.} \end{aligned}$$

Then, it is known that when n is large enough ($n > 30$) or the two populations that the samples are drawn from have approximately Normal distributions $N(\mu_1, \sigma_1), N(\mu_2, \sigma_2)$, the statistic

$$T_d = \frac{\overline{X}_d - (\mu_1 - \mu_2)}{\frac{s_d}{\sqrt{n}}} \in T(n-1).$$

Thus, we can use it as a pivot to construct a CI for the difference of means. The same line of computations as before will lead to the $100(1-\alpha)\%$ CI for the difference of means:

$$\mu_1 - \mu_2 \in \left[\overline{X}_d - t_{1-\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}},\ \overline{X}_d - t_{\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}}\right] = \left[\overline{X}_d \pm t_{\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}}\right] = \left[\overline{X}_d \mp t_{1-\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}}\right]. \quad (4.8)$$

Remark 4.3. We can find *one-sided* CI's for the difference of means of paired data, using the same procedures (and appropriate quantiles) that were described in Lecture 6.

CI for the difference of proportions

To estimate the difference of two population proportions $p_1 - p_2$, based on two independent samples of sizes n_1 and n_2, respectively, we use the estimator $\overline{p}_1 - \overline{p}_2$ for which we know (from Lecture 5, Proposition 3.4) that

$$\begin{aligned} E\left(\overline{p}_1 - \overline{p}_2\right) &= p_1 - p_2, \\ V\left(\overline{p}_1 - \overline{p}_2\right) &= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}, \\ s^2\left(\overline{p}_1 - \overline{p}_2\right) &= \frac{\overline{p}_1 \overline{q}_1}{n_1} + \frac{\overline{p}_2 \overline{q}_2}{n_2} \end{aligned} \quad (4.9)$$

with $q_i = 1 - p_i, \overline{q}_i = 1 - \overline{p}_i, i = 1, 2$. Also, for large samples ($n_1 + n_2 > 40$), by a CLT,

$$Z = \frac{\overline{p}_1 - \overline{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\overline{p}_1\overline{q}_1}{n_1} + \frac{\overline{p}_2\overline{q}_2}{n_2}}} \in N(0, 1). \tag{4.10}$$

Using Z as a pivot, we construct a $100(1-\alpha)\%$ CI for the difference of population proportions $p_1 - p_2$ as

$$\left[\overline{p}_1 - \overline{p}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}\right]. \tag{4.11}$$

Example 4.4. A candidate prepares for the local elections. During his campaign, 42 out of 70 randomly selected people in town A and 59 out of 100 randomly selected people in town B showed they would vote for this candidate. Estimate the difference in support that this candidate is getting in towns A and B with 95% confidence. Can we state affirmatively that the candidate gets a stronger support in town A?

Solution. We have

$$\begin{aligned} n_1 &= 70,\ n_2 = 100, \\ \overline{p}_1 &= 42/70 = 0.6, \\ \overline{p}_2 &= 59/100 = 0.59. \end{aligned}$$

For the confidence interval, we want $1 - \alpha = 0.95$, so we compute the quantile $z_{0.025} = -1.96$. We find the 95% CI for the difference of proportions,

$$\left[0.6 - 0.59 \pm 1.96\sqrt{\frac{0.6 \cdot 0.4}{70} + \frac{0.59 \cdot 0.41}{100}}\right] = [0.01 \pm 0.15] = [-0.14,\ 0.16].$$

So, is the support stronger for the candidate in town A? On one hand, the estimator $\overline{p}_1 - \overline{p}_2 = 0.01$ suggests that the support is 1% higher in town A than in town B. On the other hand, the difference could appear positive just because of a sampling error. As we see, the 95% confidence interval includes a large range of *negative* values too. Therefore, the obtained data does not indicate affirmatively that the support in town A is stronger.

In the following sections, we will learn how to *test* if there is any significant difference between the two candidates, so that we can conclude for it or against it. ■

5 Hypothesis Testing

In the previous sections we have considered the basic ideas of parameter estimation in some detail. We attempted to approximate the value of some population parameter θ, based on a sample, *without* having any predetermined notion concerning the actual value of this parameter. We simply tried to ascertain its value, to the best of our ability, from the information given by a random sample. In contrast, **statistical hypothesis testing** is a method of making statistical inferences on some unknown population characteristic, when *there is* a preconceived notion concerning its value or its properties.

Based on a random sample, we can use Statistics to verify a various number of statements, such as:

– the average connection speed is as claimed by the internet service provider,

– the proportion of defective products is at most a certain percentage, as promised by the manufacturer,

– service times have a certain distribution, etc.

Testing statistical hypotheses has wide applications far beyond Mathematics or Computer Science. These methods can be used to prove efficiency of a new medical treatment, safety of a new automobile brand, innocence of a defendant, authorship of a document and so forth.

5.1 Basic Concepts

So, we will work with **statistical hypotheses**, about some characteristic X (relative to a population), whose pdf $f(x;\theta)$ depends on the parameter θ, which is to be estimated.

The method(s) used to decide whether a hypothesis is true or not (in fact, to decide whether to *reject* a hypothesis or not) make up the **hypothesis test**. To begin with, we need to state *exactly* what we are testing. Any hypothesis test will involve two theories, two hypotheses,

– the **null hypothesis**, denoted by H_0 and

– the **alternative (research) hypothesis**, denoted by H_1 (or H_a).

A null hypothesis is always an equality, showing absence of an effect or relation, some "normal" situation or some usual statement that people have believed in for years. The alternative is the opposite (in some way) of the null hypothesis, a "new" theory proposed by the researcher to "challenge" the old one. In order to overturn the common belief and to reject the null hypothesis, *significant* evidence is needed. Such evidence can only be provided by data. Only when such evidence is found, and when it *strongly* supports the alternative H_1, can the hypothesis H_0 be rejected in favor of H_1. The purpose of each test is to determine whether the data provides sufficient evidence *against* H_0

in favor of H_1. This is similar to a criminal trial. The jury are required to determine if the presented evidence against the defendant is sufficient and convincing. By default, i.e. the *presumption of innocence*, insufficient evidence leads to acquittal.

To determine the truth value of a hypothesis, we use a sample function called
− the **test statistic (TS)**.

The set of values of the test statistic for which we decide to *reject* H_0 is called
− the **rejection region (RR)** or **critical region (CR)**.

The purpose of the experiment is to decide if the evidence (the data from a sample) tends to rebut the null hypothesis (if the value of the test statistic is in the rejection region) or not (if that value falls outside the rejection region).

If the statistical hypothesis refers to the parameter(s) of the distribution of the characteristic X, then we have a **parametric** test, otherwise, a **nonparametric** test. For parametric tests, we will consider that the target parameter

$$\theta \in A = A_0 \cup A_1,\ A_0 \cap A_1 = \emptyset,$$

and then the two hypotheses will be set as

$$\begin{aligned} H_0 :&\quad \theta \in A_0 \\ H_1 :&\quad \theta \in A_1. \end{aligned}$$

If the set A_0 consists of one single value, $A_0 = \{\theta_0\}$, which completely specifies the population distribution, then the hypothesis is called **simple**, otherwise, it is called a **composite** hypothesis (and the same is true for A_1 and the alternative hypothesis). The null hypothesis will *always* be taken to be simple. Then the null hypothesis

$$H_0 : \theta = \theta_0$$

will have one of the alternatives

$$\begin{aligned} H_1 :&\quad \theta < \theta_0 \ \text{(left-tailed test)}, \\ H_1 :&\quad \theta > \theta_0 \ \text{(right-tailed test)}, \\ H_1 :&\quad \theta \neq \theta_0 \ \text{(two-tailed test)}. \end{aligned}$$

Remark 5.1. The first and one of the most important tasks in a hypothesis testing problem is to state the *relevant* null and alternative hypotheses to be tested. The null hypothesis is taken to be a simple hypothesis, but the *appropriate* alternate has to be *understood from the context*. We mentioned that H_1 is the opposite "in some way" of H_0. Let us clarify this.

1. Consider a problem in which a medicine which is believed to have the side effect of increasing the body temperature above normal, is tested. If the temperature values of a number of patients taking this medicine are considered, then for the mean temperature the relevant hypotheses would be

$$\begin{aligned} H_0: &\quad \mu = 37 \\ H_1: &\quad \mu > 37, \end{aligned}$$

since an average lower than or equal to 37^oC would mean the same thing in this context, the patients are fine. A problem would be a mean temperature *greater* than 37^oC. In this sense, H_0 and H_1 are "opposites" of each other.

2. To verify that the average broadband internet connection speed is 100 Mbps, we test the hypothesis

$$\begin{aligned} H_0: &\quad \mu = 100 \\ H_1: &\quad \mu \neq 100. \end{aligned}$$

However, if we worry about a *low* connection speed only, we can conduct a one-sided test of

$$\begin{aligned} H_0: &\quad \mu = 100 \\ H_1: &\quad \mu < 100. \end{aligned}$$

In this case, we only measure the amount of evidence supporting the one-sided alternative $H_1 : \mu < 100$. In the absence of such evidence, we gladly accept the null hypothesis.

Designing a hypothesis test means constructing the rejection region RR, such that for a given $\alpha \in (0, 1)$, the conditional probability, conditioned by H_0 being true, is

$$P(TS \in RR \mid H_0) = \alpha. \tag{5.1}$$

For any given hypothesis testing problem, we have the following possibilities:

Decision	Actual situation	
	H_0 true	H_1 true
Reject H_0	Type I error (prob. α)	Right decision
Not reject H_0	Right decision	Type II error (prob. β)

Table 1: Decisions and errors

In two of the cases, we make the right decision, in the other two, we make an error.
A **type I error** occurs when we reject a true null hypothesis and by (5.1), the probability of making such an error is

$$P(\text{type I error}) \;=\; P(\text{ reject } H_0 \mid H_0) \;=\; P(TS \in RR \mid H_0) \;=\; \alpha. \tag{5.2}$$

The value α is called **significance level** or **risk probability**.
A **type II error** happens when we fail to reject a false null hypothesis, and its probability is denoted by β,

$$P(\text{ type II error}) \;=\; P(\text{ not reject } H_0 \mid H_1) \;=\; P(TS \notin RR \mid H_1) \;=\; \beta. \tag{5.3}$$

Remark 5.2.
1. The rejection region and hence, the hypothesis test, are *not* uniquely determined by (5.1), just like confidence intervals.
2. Since both α and β represent risks of making an error, we would like to design tests such that both of their values are small. Unfortunately, making one of them very small will result in the other being unreasonably large. But, for almost all statistical tests, α and β will both decrease as the sample size increases.
3. In general, α is preset and a procedure is given for finding an appropriate rejection region.

5.2 General Framework, Z-Tests

Just like with confidence intervals, we start with the case where the test statistic has a $N(0,1)$ distribution, so we can better understand the ideas.

Let θ be a target parameter and let $\overline{\theta}$ be an unbiased estimator for θ ($E(\overline{\theta}) = \theta$), with standard error $\sigma_{\overline{\theta}}$, such that, under certain conditions, it is known that

$$Z \;=\; \frac{\overline{\theta} - \theta}{\sigma_{\overline{\theta}}} \;\left(= \;\frac{\overline{\theta} - E(\overline{\theta})}{\sigma(\overline{\theta})}\right) \tag{5.4}$$

has an approximately Standard Normal $N(0,1)$ distribution. We design a hypothesis testing procedure for θ the following way: for a given level of significance $\alpha \in (0,1)$, consider the hypotheses

$$H_0: \;\; \theta = \theta_0,$$

with one of the alternatives

$$H_1 : \begin{cases} \theta < \theta_0 \\ \theta > \theta_0 \\ \theta \neq \theta_0. \end{cases} \tag{5.5}$$

We will use the test statistic $TS = Z$ given by (5.4).
The **observed value of the test statistic** from the sample data is

$$TS_0 \;=\; TS(\theta = \theta_0). \tag{5.6}$$

In our case, this is

$$Z_0 \;=\; TS(\theta = \theta_0) \;=\; \frac{\overline{\theta} - \theta_0}{\sigma_{\overline{\theta}}}.$$

How to design the rejection region RR? Let us start with the left-tailed case. We need to determine the RR such that (5.1) holds. Intuitively, we reject H_0 if the observed value of the test statistic is *far* from the value specified in H_0, "far" in the sense of the alternative H_1, in this case *far to the left* of θ_0. So, we determine a rejection region of the form

$$RR \;=\; \{Z_0 \mid Z_0 \leq k_1\} \;=\; (-\infty, k_1].$$

We have

$$\begin{aligned} \alpha &= P(Z_0 \in RR \mid H_0) \\ &= P(Z_0 \leq k_1 \mid \theta = \theta_0) \\ &= P(Z_0 \leq k_1 \mid Z_0 \in N(0,1)). \end{aligned}$$

Now, we know that if $Z_0 \in N(0,1)$, $P(Z_0 \leq z_\alpha) = \alpha$, where z_α is the quantile of order α for the $N(0,1)$ distribution. Thus, we choose $k_1 = z_\alpha$ and

$$RR_{\text{left}} \;=\; \{Z_0 \leq z_\alpha\}. \tag{5.7}$$

Similarly, for a right-tailed test, we want to find a rejection region of the form

$$RR \;=\; \{Z_0 \mid Z_0 \geq k_2\} \;=\; [k_2, \infty),$$

so that

$$\begin{aligned}\alpha &= P(Z_0 \in RR \mid H_0)\\ &= P(Z_0 \geq k_2 \mid \theta = \theta_0)\\ &= P(Z_0 \geq k_2 \mid Z_0 \in N(0,1))\\ &= 1 - P(Z_0 < k_2 \mid Z_0 \in N(0,1)).\end{aligned}$$

Since $P(Z_0 < z_{1-\alpha}) = 1 - \alpha$, then $P(Z_0 \geq z_{1-\alpha}) = \alpha$ and so we choose $k_2 = z_{1-\alpha}$, the quantile of order $1 - \alpha$ for the $N(0,1)$ distribution and

$$RR_{\text{right}} = \{Z_0 \geq z_{1-\alpha}\}. \tag{5.8}$$

Finally, for a two-tailed test, we reject the null hypothesis if the observed value of the test statistic is far away from θ_0 *on either side*. That is, the rejection region should be of the form $RR = \{Z_0 \mid Z_0 \leq k_1 \text{ or } Z_0 \geq k_2\} = (-\infty, k_1] \cup [k_2, \infty)$. The rejection region should be chosen such that

$$P(Z_0 \leq k_1 \text{ or } Z_0 \geq k_2 \mid \theta = \theta_0) = \alpha,$$

or, equivalently,

$$P(k_1 < Z_0 < k_2 \mid Z_0 \in N(0,1)) = 1 - \alpha.$$

We encountered such problems before in the previous sections, when finding (two-sided) confidence intervals. As we did then, we will choose $k_1 = z_{\frac{\alpha}{2}}$ and $k_2 = z_{1-\frac{\alpha}{2}}$, so

$$RR_{\text{two}} = \{Z_0 \leq z_{\frac{\alpha}{2}} \text{ or } Z_0 \geq z_{1-\frac{\alpha}{2}}\}, \tag{5.9}$$

or, since the distribution of Z is symmetric and $z_{1-\frac{\alpha}{2}} > 0$,

$$\begin{aligned}RR_{\text{two}} &= \{Z_0 \leq -z_{1-\frac{\alpha}{2}} \text{ or } Z_0 \geq z_{1-\frac{\alpha}{2}}\}\\ &= \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\}.\end{aligned}$$

To summarize, the rejection regions for the three alternatives (5.5) are given by

$$RR: \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{Z_0 \leq z_{\frac{\alpha}{2}} \text{ or } Z_0 \geq z_{1-\frac{\alpha}{2}}\} = \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\}. \end{cases} \tag{5.10}$$

Remark 5.3.

1. Since a test statistic $Z \in N(0,1)$ was used, these are commonly known as ***Z*-tests**.

2. We will derive hypothesis tests for common parameters (mean, proportion, difference of means, ratio of variances, difference of proportions). The test statistics and their distributions will change, but the ideas and the principles will remain the same, as for the case we just described.

3. Notice from our derivation of the rejection region for a two-tailed test, that there is a strong relationship between confidence intervals and rejection regions: The values θ_0 of a target parameter θ in a $100(1-\alpha)\%$ CI $(\alpha \in (0,1))$, are precisely the values for which the test statistic falls *outside* the RR, and hence, for which the null hypothesis $\theta = \theta_0$ is *not* rejected at the significance level α. We say that the $100(1-\alpha)\%$ two-sided CI consists of all the *acceptable* values of the parameter, at the significance level α.

4. **Caution!** This is **not** saying that the rejection region is the complement of the confidence interval! The RR contains values for the *test statistic* TS, while the CI consists of values of the *parameter* θ.

Example 5.4. The number of monthly sales at a firm is known to have a mean of 20 and a standard deviation of 4 and all salary, tax and bonus figures are based on these values. However, in times of economical recession, a sales manager fears that his employees do not average 20 sales per month, but less, which could seriously hurt the company. For a number of 36 randomly selected salespeople, it was found that in one month they averaged 19 sales. At the 5% significance level, does the data confirm or contradict the manager's suspicion?

Solution. The question is about the *average* number of sales per month, so the test is for the population mean μ. Recall that if either the original population is approximately Normally distributed or the sample size is large (over 30) and σ is known, then

$$Z \;=\; \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \;\in\; N(0,1).$$

Since the sample size $n = 36 > 30$ and we know $\sigma = 4$, we can use a Z-test. The manager's suspicion is that the average is *less* than 20, which is supposed to be, so the two relevant hypotheses in this case are

$$\begin{aligned} H_0: &\quad \mu = 20 \\ H_1: &\quad \mu < 20, \end{aligned}$$

a left-tailed test.

A type I error would mean concluding that the average number of monthly sales is less than 20, when in fact, it is not; a type II error would be deciding that the average number of monthly sales is

20 (or higher), but it actually is not. We allow for the probability of a type I error (the significance level) to be $\alpha = 0.05$. The population standard deviation is known, $\sigma = 4$ and the sample mean is $\overline{X} = 19$.

The observed value of the test statistic is

$$Z_0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{19 - 20}{\frac{4}{6}} = -1.5.$$

The rejection region is, by (5.10),

$$RR = (-\infty, z_\alpha] = (-\infty, -1.645].$$

Since $Z_0 \notin RR$, we *do not reject* H_0. The evidence obtained from the data is not sufficient to reject it. In the absence of sufficient evidence, by default, we accept the null hypothesis. So, at the 5% significance level, the data *does not* confirm the manager's suspicion.

■

5.3 Significance Testing, P-Values

There is a problem that might occur in hypothesis testing: We preset α, the probability of a type I error and henceforth determine a rejection region. We get a value of the test statistic that *does not belong* to it, so we cannot reject the null hypothesis H_0, i.e. we accept it as being true. However, when we compute the probability of getting that value of the test statistic under the assumption that H_0 is true, we find it is *very small*, comparable with our preset α. So, we accept H_0, yet considering it to be true, we find that it is *very unlikely* (very improbable) that the test statistic takes the observed value we found for it. That makes us wonder if we set our RR right and if we didn't "accept" H_0 too easily, by hastily dismissing values of the test statistic that did not fall into our RR. So we should take a look at how "far-fetched" does the value of the test statistic seem, under the assumption that H_0 is true. If it seems really implausible to occur by chance, i.e. if its probability is *small*, then maybe we should reject the null hypothesis H_0 after all.

To avoid this situation, we perform what is called a **significance test**: for a given random sample $(X_1, \ldots, X_n)$, we still set up H_0 and H_1 as before and we choose an appropriate test statistic. Then, we compute the probability of observing a value *at least as extreme* (in the sense of the test conducted) of the test statistic TS as the value observed from the sample, TS_0, under the assumption that H_0 is true. This probability is called the critical value, the descriptive significance level, the

probability of the test, or, simply the ***P*-value** of the test. If it is small, we reject H_0, otherwise we do not reject it. The P-value is a numerical value assigned to the test, it depends only on the sample data and its distribution, but *not* on α.

In general, for the three alternatives (5.5), if TS_0 is the value of the test statistic TS under the assumption that H_0 is true and F is the cdf of TS, the P-value is computed by

$$P = \begin{cases} P(TS \leq TS_0 \mid H_0) & = & F(TS_0) \\ P(TS \geq TS_0 \mid H_0) & = & 1 - F(TS_0) \\ 2 \cdot \min\{P(TS \leq TS_0 \mid H_0), P(TS \geq TS_0 \mid H_0)\} & = & 2 \cdot \min\{F(TS_0), 1 - F(TS_0)\}. \end{cases} \tag{5.11}$$

Then the decision will be

$$\begin{array}{l} \text{if } \ P \leq \alpha, \text{ reject } \ H_0, \\ \text{if } \ P > \alpha, \text{ do not reject } \ H_0. \end{array} \tag{5.12}$$

So, more precisely, the P-value of a test is the smallest level at which we could have preset α and still have been able to reject H_0, or the lowest significance level that *forces* rejection of H_0, i.e. the *minimum rejection level*.

Remark 5.5.

1. Thus, we can avoid the costly computation of the rejection region (costly because of the quantiles) and compute the P-value instead. Then, we simply compare it to the significance level α. If α is above the P-value, we reject H_0, but if it is below that minimum rejection level, we can no longer reject the null hypothesis.

2. Hypothesis testing (determining the rejection region) and significance testing (computing the P-value) are two methods for testing *the same* thing (the same two hypotheses), so, of course, the outcome (the decision of rejecting or not H_0) will be *the same*, for the same data. Significance testing is preferable to hypothesis testing, especially from the computer implementation point of view, since it avoids the inversion of a cdf, which is, oftenly, a complicated improper integral.

Example 5.6. For the problem in Example 5.4, let us perform a significance test.

Solution. We tested a left-tailed alternative for the mean

$$\begin{aligned} H_0 &: \ \mu = 20 \\ H_1 &: \ \mu < 20. \end{aligned}$$

The population standard deviation was given, $\sigma = 4$, and for a sample of size $n = 36$, the sample

mean was $\overline{X} = 19$. For the test statistic

$$Z \;=\; \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \;\in\; N(0,1),$$

the observed value was

$$Z_0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \;=\; \frac{19 - 20}{\frac{4}{6}} \;=\; -1.5.$$

Now, we compute the P-value

$$P \;=\; P(Z \leq Z_0) \;=\; P(Z \leq -1.5) \;=\; 0.0668.$$

Since

$$\alpha \;=\; 0.05 \;<\; 0.0668 \;=\; P,$$

(is below the minimum rejection level), we do not reject H_0, so, at the 5% significance level, we conclude that the data contradicts the manager's suspicion.

■