4.5 Confidence Intervals for Proportions

Recall (from Lecture 5) that a *population proportion* is

$$p \;=\; P(i \in A),$$

where A is a subpopulation.
Based on a random sample $X_1, \ldots, X_n$, we define the *sample proportion* as

$$\overline{p} \;=\; \frac{\text{number of sampled items from } A}{n}.$$

Then

$$\begin{aligned} E(\overline{p}) &= p, \\ V(\overline{p}) &= \frac{p(1-p)}{n} = \frac{pq}{n}. \end{aligned} \tag{4.1}$$

So $\overline{p}$ is an absolutely correct estimator for p and by a CLT,

$$Z \;=\; \frac{\overline{p} - p}{\sqrt{\dfrac{pq}{n}}} \tag{4.2}$$

converges in distribution to a Standard Normal $N(0,1)$ variable, as $n \to \infty$.
Now, as p is unknown, we estimate the standard error $\sigma_{\overline{p}} = \sqrt{V(\overline{p})} = \sqrt{\dfrac{p(1-p)}{n}}$ by

$$s_{\overline{p}} \;=\; \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}.$$

So, again, for large samples ($n > 30$), we can use

$$Z = \frac{\overline{p} - p}{\sqrt{\dfrac{\overline{p}(1-\overline{p})}{n}}} \in N(0,1)$$

as a pivot to construct a confidence interval for p.
For a given confidence level $1-\alpha$, with the same computations as before, we obtain a $100(1-\alpha)\%$

CI for the population proportion p as

$$\left[\overline{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right]. \tag{4.3}$$

Selecting the sample size

Just as we did for the population mean (in the case of known variance), we can derive a formula for the sample size that will provide a certain precision of our interval estimator. The length of the CI in (4.3) is

$$2\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} z_{1-\frac{\alpha}{2}}.$$

Notice that for any $\overline{p} \in (0,1)$, we have

$$\overline{p}(1-\overline{p}) \ \leq \ \frac{1}{4}.$$

Then to get a desired precision

$$2\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} z_{1-\frac{\alpha}{2}} \quad \leq \quad \Delta,$$

we solve

$$2 \cdot \frac{1}{2} \frac{1}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \quad \leq \quad \Delta,$$

for n. We get

$$n \quad \geq \quad \left(\frac{z_{1-\frac{\alpha}{2}}}{\Delta}\right)^2. \tag{4.4}$$

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 (again, from Lecture 4) that

$$E\left(\overline{p}_1 - \overline{p}_2\right) \quad = \quad p_1 - p_2,$$

$$\begin{aligned} V\left(\overline{p}_1-\overline{p}_2\right) &= \frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}, \\ s^2\left(\overline{p}_1-\overline{p}_2\right) &= \frac{\overline{p}_1\overline{q}_1}{n_1}+\frac{\overline{p}_2\overline{q}_2}{n_2} \end{aligned} \tag{4.5}$$

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,

$$Z = \frac{\overline{p}_1-\overline{p}_2-(p_1-p_2)}{\sqrt{\frac{\overline{p}_1\overline{q}_1}{n_1}+\frac{\overline{p}_2\overline{q}_2}{n_2}}} \in N(0,1). \tag{4.6}$$

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

$$\left[\overline{p}_1-\overline{p}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1}+\frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}\right]. \tag{4.7}$$

Example 4.1. A company has to accept or reject a large shipment of items. For quality control purposes, they collect a sample of 200 items and find 12 defective items in it.
a) Find a 99% confidence interval for the proportion of defective items in the whole shipment.
b) How many items should be tested to ensure a 99% confidence interval of length at most 0.05?

Solution. The sample is large enough and we have

$$\overline{p} = \frac{12}{200} = 0.06.$$

For $1-\alpha = 0.99$, $\alpha = 0.01$, $\alpha/2 = 0.005$, the quantile is

$$z_{0.005} = -2.576.$$

Then the 99% confidence interval for the proportion of defective items is

$$\left[\overline{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right] = \left[0.06 \pm 2.576\sqrt{\frac{0.06\cdot 0.94}{200}}\right] = [0.017,\ 0.103].$$

So, with 99% confidence, the percentage of defective items is between 1.7% and 10.3%.

b) The length of the 99% CI we found is 0.086. For a margin of $\Delta \leq 0.05$ of the 99% CI, we

need a sample size of

$$n \;\geq\; \left(\frac{z_{0.995}}{\Delta}\right)^2 \;=\; \left(\frac{2.576}{0.05}\right)^2 \;=\; 2653.898 \;\approx\; 2654.$$

■

Example 4.2. Two candidates prepare for the local elections. During a phone poll, 42 out of 70 randomly selected people said they would vote for candidate A and 59 out of 100 randomly selected people said they preferred candidate B and would vote for him. Estimate the difference in support for the two candidates with 95% confidence. Can we state affirmatively that candidate A gets a stronger support than candidate B?

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 candidate A? On one hand, the estimator $\overline{p}_1 - \overline{p}_2 = 0.01$ suggests that the support is 1% higher for candidate A than for 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 for candidate 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" 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, 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

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

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

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

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 usually 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,

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

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), as was the case with 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

$$RR \;=\; \{Z_0 \mid Z_0 \leq k_1\} \;=\; (-\infty, k1].$$

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 section, 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 all the common parameters (mean, variance, 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.

Here, the *population* would be the number of monthly sales at this firm, of *all* the employees, for any period of time.

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.

Now, which alternative is appropriate? The manager's suspicion is that the average is *less* than 20, which is supposed to be. If that average is 20, everything is ok. If it is *greater* than 20, even

better! A problem is if the average is *less* than 20, so the two hypotheses to be tested are

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

a left-tailed test. Here, "greater than or equal to 20" go together, they are in the same category and the opposite is "less than 20". However, the null hypothesis states *only* equality to 20, but we keep in the back of our minds that "$\mu > 20$" falls in the same category.

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.

■