Short review

Let us recall: we have a population characteristic X, whose pdf $f(x;\theta)$ depends on θ, the target parameter to be estimated. The estimation is done based on a sample of size n, i.e. sample variables $X_1, X_2, \ldots, X_n$ that are *iid*, with the same pdf as X.

We set up two hypotheses, the *null* hypothesis, always simple, i.e.

$$H_0 : \theta = \theta_0$$

and one of the *alternative* hypotheses

$$\begin{aligned} H_1 :&\quad \theta < \theta_0 \ \text{(left-tailed test)},\\ H_1 :&\quad \theta > \theta_0 \ \text{(right-tailed test)},\\ H_1 :&\quad \theta \neq \theta_0 \ \text{(two-tailed test)}. \end{aligned} \tag{5.1}$$

We want to decide if H_0 is *rejected* (in favor of H_1) or *not rejected* (accepted). We use a *test statistic* TS (with the same properties as the pivot in CI's) and a *rejection (critical) region* RR, such that for a given *significance level* $\alpha \in (0,1)$,

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

The probability of a *type II error* is

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

In general, the significance level α is preset and a procedure is given for finding an appropriate rejection region, such that β is also reasonably small.

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.

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 (5.1), 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

$$P = \begin{cases} P(TS \leq TS_0 \mid H_0) & = & F(TS_0) \\ P(TS \geq TS_0 \mid H_0) & = & 1 - F(TS_0) \\ 2 \cdot \min\{P(TS \leq TS_0 \mid H_0), P(TS \geq TS_0 \mid H_0)\} & = & 2 \cdot \min\{F(TS_0), 1 - F(TS_0)\}. \end{cases} \tag{5.3}$$

Then the decision will be

$$\begin{array}{l} \text{if } \ P \leq \alpha, \text{ reject } \ H_0, \\ \text{if } \ P > \alpha, \text{ do not reject } \ H_0. \end{array} \tag{5.4}$$

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 5.1.

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 5.2. Recall the problem in Example 5.4 (Lecture 10): 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? Let us now perform a significance test.

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

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

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.

■

5.4 Tests for Proportions

Tests for a population proportion, $\theta = p$

Let us recall that, when estimating a population proportion p, if the sample size is large enough ($n > 30$), then the variable

$$Z \;=\; \frac{\overline{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \;\in\; N(0,1), \tag{5.5}$$

where $\overline{p}$ is the sample proportion. So this case fits the general Z-test framework.
To test

$$H_0 : \;\; p = p_0,$$

with one of the alternatives

$$H_1 : \;\; \begin{cases} p < p_0 \\ p > p_0 \\ p \neq p_0. \end{cases}, \tag{5.6}$$

we use the test statistic $TS = Z$ from (5.5). Then, as before, at the $\alpha \in (0,1)$ significance level, the rejection region for each test will be given by

$$RR : \;\; \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\}, \end{cases} \tag{5.7}$$

and the P-value will be computed as

$$P = \;\; \begin{cases} P(Z \leq Z_0 \mid H_0) &= \Phi(Z_0) \\ P(Z \geq Z_0 \mid H_0) &= 1 - \Phi(Z_0) \\ P(|Z| \geq |Z_0| \mid H_0) &= 2\left(1 - \Phi(|Z_0|)\right), \end{cases} \tag{5.8}$$

since $N(0,1)$ is symmetric, where

$$\Phi(x) \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

is Laplace's function, the cdf for the Standard Normal $N(0,1)$ distribution.

Example 5.3. A company is receiving a large shipment of items. For quality control purposes, they collect a sample of 200 items and find 24 defective ones in it.
a) The manufacturer claims that at most 1 in 10 items in the shipment is defective. At the 5% significance level, does the data confirm or contradict his claim?
b) Find the P-value of the test in part a).
c) Find the probability of a type II error, β, for testing

$$\begin{aligned} H_0 &: p = 0.1 \\ H_1 &: p = 0.15. \end{aligned}$$

d) What sample size would ensure that both $\alpha = \beta = 0.05$ for the test in part c)?

Solution.
We have a sample of size $n = 200$ for which the sample proportion is

$$\overline{p} \;=\; \frac{24}{200} \;=\; \frac{3}{25} \;=\; 0.12.$$

a) The manufacturer claims that *at most* 1 in 10 items is defective, i.e. that $p \leq 0.1$. So, we are testing a *right*- tailed alternative

$$\begin{aligned} H_0 &: p = 0.1 \\ H_1 &: p > 0.1. \end{aligned}$$

If we decide to reject H_0, that means the data *contradicts* the manufacturer's claim, whereas if we do not reject it, it means the data is insufficient to contradict his claim, so we consider it to be true. We have a significance level $\alpha = 0.05$, so for the rejection region we need the quantile

$$z_{1-\alpha} \;=\; z_{0.95} \;=\; 1.645$$

and the rejection region is

$$RR \;=\; [1.645, \infty).$$

The test statistic is

$$Z \;=\; \frac{\overline{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$$

and its observed value is

$$Z_0 \;=\; \frac{0.12 - 0.1}{\sqrt{\dfrac{0.1 \cdot 0.9}{200}}} \;=\; 0.943.$$

Since $Z_0 \notin RR$, we *do not* reject H_0 at this significance level, i.e. conclude that the data seems to confirm the manufacturer's claim that at most 10% of items are defective. Notice that even though the sample proportion was 0.12, *bigger* than 0.1, the inference on the *entire* population proportion is that it *does not exceed* 0.1 (data from a sample may be misleading, if it is not used properly ...)

b) The P-value is

$$P \;=\; P(Z \geq Z_0) \;=\; 1 - P(Z \leq 0.943) \;=\; 1 - \Phi(0.843) \;=\; 0.173.$$

Since

$$\alpha \;=\; 0.05 \;<\; 0.173 \;=\; P,$$

the decision is to *not reject* the null hypothesis. i.e. accept the manufacturer's claim.
Notice that the significance test tells us more! Since the P-value is so large (remember, it is comparable to a probability of an *error*, so a *small* quantity), not only at the 5% significance level we decide to accept H_0, but at *any* reasonable significance level the decision would be the same. That means that the data *strongly* suggests that H_0 is true and should not be rejected. So, even more we see that we should be careful not to extrapolate the property of one sample to the entire population.

c) We are now testing two simple hypotheses

$$\begin{aligned} H_0 \;&:\; p = 0.1 \;= p_0 \\ H_1 \;&:\; p = 0.15 \;=\; p_1. \end{aligned}$$

From part a), we found the rejection region

$$Z_0 \;\geq\; z_{0.95},$$

which means

$$\frac{\overline{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \;\geq\; z_{0.95}.$$

If we rewrite this inequality (the way we did when we found confidence intervals), we find that H_0 is rejected if

$$\overline{p} \ \geq \ p_0 + z_{0.95}\sqrt{\frac{p_0(1-p_0)}{n}} \ = \ 0.1349.$$

So, we *reject* H_0 if $\overline{p} \geq 0.1349$. Then we *do not reject* the null hypothesis if

$$\overline{p} \ < \ 0.1349.$$

Also, recall that the test statistic $Z \ = \ \dfrac{\overline{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}}$ has a $N(0,1)$ distribution, when we replace p by its *true* value. So, if we assume that the true value is $p = p_1$, then

$$Z_1 \ = \ \frac{\overline{p}-p_1}{\sqrt{\dfrac{p_1(1-p_1)}{n}}} \in N(0,1).$$

Now, let us proceed to compute $\beta(p_1)$. We have

$$\begin{aligned}
\beta(p_1) \ &= \ P(\text{not reject } \ H_0 \mid H_1) \ = \ P(\overline{p} < 0.1349 \mid p = p_1) \\
&= \ P\left(\frac{\overline{p}-p_1}{\sqrt{\dfrac{p_1(1-p_1)}{n}}} < \frac{0.1349-p_1}{\sqrt{\dfrac{p_1(1-p_1)}{n}}} \ \middle| \ p = p_1 \right) \\
&= \ P(Z_1 < -0.598 \mid Z_1 \in N(0,1)) \\
&= \ \Phi(-0.598) \ = \ 0.275,
\end{aligned}$$

very large! Such a probability of error is *not* acceptable, under any circumstances!

d) So, let us see how large should the sample size be so that the probability of type II error becomes acceptable, i.e. $\beta = 0.05$.
The idea stems from the computations we did in part c) to find β. Now, we go backwards: presetting *both* α and β, we find that cutoff value for $\overline{p}$ (that was 0.1349 before), say k, from both conditions. Setting them equal, we solve for n.
So, on one hand, we have

$$\alpha \ = \ P(\text{ reject } \ H_0 \mid H_0) \ = \ P(\overline{p} \geq k \mid p = p_0)$$

$$
\begin{aligned}
&= P\left(\frac{\overline{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq \frac{k-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \;\middle|\; p=p_0\right)\\
&= P\left(Z_0 \geq \frac{k-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \;\middle|\; Z_0 \in N(0,1)\right)\\
&= P(Z_0 \geq z_{1-\alpha} \mid Z_0 \in N(0,1)),
\end{aligned}
$$

since the quantile $z_{1-\alpha}$ is the *only* value with the property that $P(Z_0 > z_{1-\alpha}) = \alpha$, for a $N(0,1)$ variable. Then,

$$
\frac{k-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = z_{1-\alpha}, \text{ i.e. } k = p_0 + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}.
$$

On the other hand, as we did in part c), we have

$$
\begin{aligned}
\beta &= P(\text{ not reject } H_0 \mid H_1) = P(\overline{p} < k \mid p = p_1)\\
&= P\left(\frac{\overline{p}-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} < \frac{k-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \;\middle|\; p=p_1\right)\\
&= P\left(Z_1 < \frac{k-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \;\middle|\; Z_1 \in N(0,1)\right)\\
&= P(Z_1 < z_\beta \mid Z_1 \in N(0,1)),
\end{aligned}
$$

so

$$
\frac{k-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} = z_\beta, \text{ i.e. } k = p_1 + z_\beta\sqrt{\frac{p_1(1-p_1)}{n}}.
$$

From the two equations, we find

$$
p_0 + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}} = p_1 + z_\beta\sqrt{\frac{p_1(1-p_1)}{n}},
$$

$$
\begin{aligned}
z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}} - z_{\beta}\sqrt{\frac{p_1(1-p_1)}{n}} &= p_1 - p_0, \\
\frac{1.0808}{\sqrt{n}} &= 0.05,
\end{aligned}
$$

so

$$
n = \left(\frac{1.0809}{0.05}\right)^2 = 467.34
$$

Thus, a sample of size at least $n = 468$ items should be selected in order to ensure that both probabilities of error are at most 0.05. ■

Tests for a comparing population proportions, $\theta = p_1 - p_2$

Similarly, in this case, if the samples are large enough ($n_1 + n_2 > 40$), then the variable

$$
Z = \frac{\overline{p}_1 - \overline{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \in N(0,1), \tag{5.9}
$$

where $\overline{p}_1$ and $\overline{p}_2$ are the two sample proportions. To test

$$
\begin{aligned}
H_0 &: \ p_1 - p_2 = 0, \text{ versus} \\
H_1 &: \begin{cases} p_1 - p_2 < 0 \\ p_1 - p_2 > 0 \\ p_1 - p_2 \neq 0, \end{cases}
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
H_0 &: \ p_1 = p_2, \text{ versus} \\
H_1 &: \begin{cases} p_1 < p_2 \\ p_1 > p_2 \\ p_1 \neq p_2, \end{cases}
\end{aligned} \tag{5.10}
$$

we use $TS = Z$ from (5.9) as test statistic. Let us see what the observed value Z_0 would be. Since under the null hypothesis, $p_1 = p_2$, it makes sense to estimate both proportions in (5.9) by the *overall* proportion

$$
\hat{p} = \frac{n_1 \overline{p}_1 + n_2 \overline{p}_2}{n_1 + n_2}, \tag{5.11}
$$

called the **pooled proportion** (a proportion that takes into account data from both samples). Then the observed value of the test statistic Z_0 is

$$Z_0 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{5.12}$$

The rejection regions and P-values for the three alternatives are then given by equations (5.7)-(5.8), with Z_0 from (5.12).

Example 5.4. Suppose now that the company in Example 5.3 is trying a new supplier. A sample of 150 items produced by the second supplier contains 21 defective parts. Considering that now 14% of items are defective (and with the first supplier the percentage was 12%), the company is in a serious bind. At the 5% significance level, does the new supplier seem worse than the first one?

Solution. For the first supplier the data was $n_1 = 200$, $\overline{p}_1 = 0.12$, for the new one, we have $n_2 = 150$ and $\overline{p}_2 = 0.14$, this is why the company is afraid the second supplier may be worse than the first one. Now, "worse" would mean that for the entire populations the proportions satisfy $p_1 < p_2$. So, we perform a *left-tailed* test

$$\begin{aligned} H_0 &: p_1 = p_2 \\ H_1 &: p_1 < p_2. \end{aligned}$$

For a left-tailed test and significance level $\alpha = 0.05$, the rejection region is

$$RR = (-\infty, z_{0.05}] = (-\infty, -1.654].$$

The pooled proportion from (5.11) is

$$\hat{p} = \frac{n_1\overline{p}_1 + n_2\overline{p}_2}{n_1 + n_2} = \frac{24 + 21}{350} = 0.1286.$$

Then the observed value of the test statistic (from (5.12)) is

$$Z_0 = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -0.5531.$$

Since $Z_0 \notin RR$, we do not reject the null hypothesis, i.e. we conclude that overall, the second supplier is *not* worse than the first one.
For significance testing, the P-value of the test is

$$P \;=\; P(Z \leq Z_0) \;=\; P(Z \leq -0.5531) \;=\; \Phi(-0.5531) \;=\; 0.29,$$

again, *very* large, much larger than this (or any reasonable) α, so the decision is to not reject H_0, a decision that seems *strongly* supported by the data.

■