

Inference for a Single Proportion¹

by Yann Lamontagne

Example: An open source advocate claims Mozilla Thunderbird is used by at least 75% of internet users. A random sample of 300 internet users included 216 that use Mozilla Thunderbird.

- State the hypotheses.
- Verify the conditions to test the hypotheses.
- Define the rejection region with respect to the sampling distribution and the rejection region with respect to the standard normal distribution at 5% significance.
- Test the hypothesis using the sample proportion, the test statistic and p-value at 5% significance.
- Determine the probability of committing an error (Type I or Type II) in the hypothesis test assuming that the true proportion of the population is $p_a = 0.68$.
- Determine the power of the test assuming that the true proportion of the population is $p_a = 0.68$.
- Determine the sample size to control the error at $\alpha = \beta = 0.05$ assuming that the true proportion of the population is $p_a = 0.68$.

a. State the hypotheses.

The text gives the following definition

The *null hypothesis* (H_0) often represents either a skeptical perspective or a claim to be tested. The *alternative hypothesis* (H_a) represents an alternative claim under consideration and is often represented by a range of possible parameter values.[OS]

In this problem:

$H_0 : p = 0.75$ (\geq), that is Mozilla Thunderbird is used by at least 75% of internet users.

$H_a : p < 0.75$, that is Mozilla Thunderbird is used by less than 75% of internet users.

Remember that the null hypothesis must be the hypothesis containing the equality, as the sampling distribution on which the hypothesis test is performed uses the population parameter from the null hypothesis.

b. Verify the conditions to test the hypotheses.

The conditions for performing an inference for a single proportion using a normal distribution are verified:

- The sample is assumed to be independent as the sample is less than 10% of internet users.
- Since $np_0 = 300(0.75) = 225 \geq 10$ and $n(1 - p_0) = 75 \geq 10$ the binomial distribution can be approximated using a normal distribution.

c. Define the rejection region with respect to the sampling distribution and the rejection region with respect to the standard normal distribution at 5% significance.

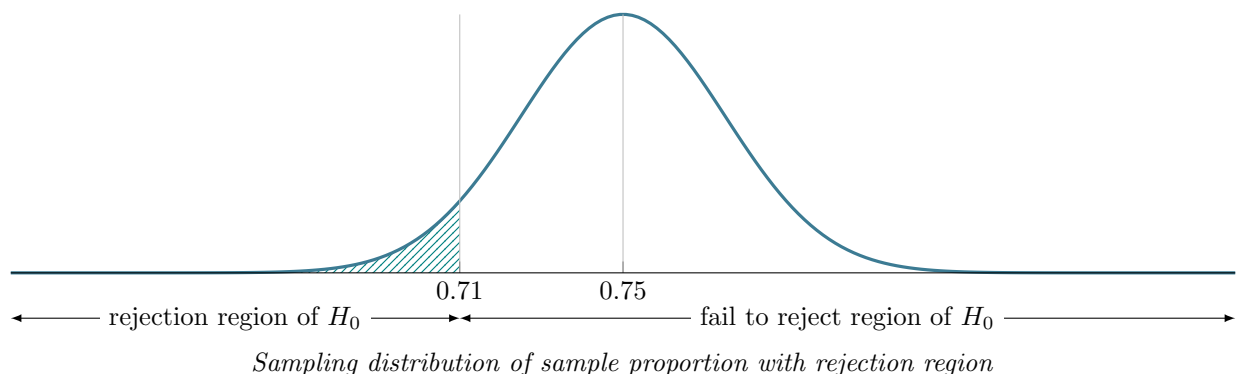
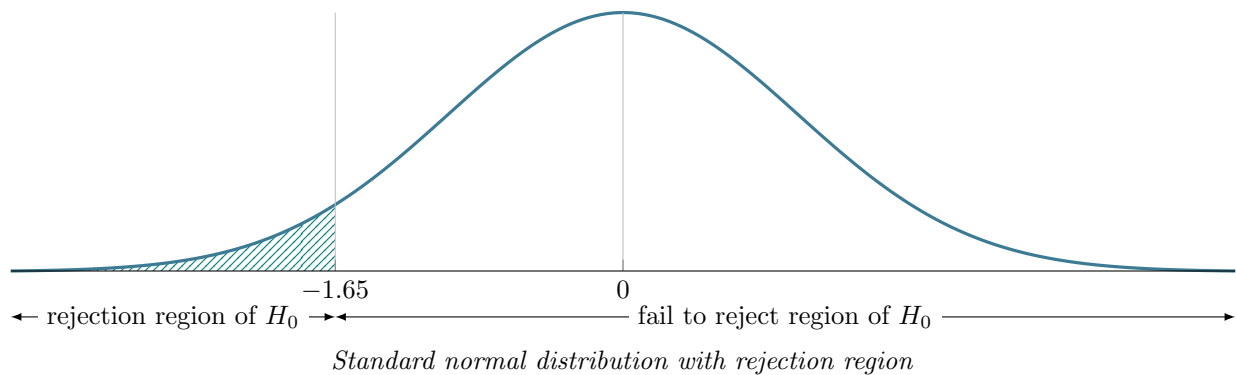
The sampling distribution for the null hypothesis of sample size $n = 300$ is centered at $\mu_{\hat{p}} = p_0 = 0.75$ and has standard error $SE = \sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.75(0.25)}{300}} = 0.025$.

The hypotheses determine that the rejection region is the left tail. The rejection region for the standard normal distribution is the interval $(-\infty, -Z_\alpha) = (-\infty, -1.645)$. The rejection region for the sampling distribution is $(-\infty, \hat{p}_l)$ where

$$-Z_\alpha = \frac{\hat{p}_l - p_0}{SE}$$
$$\hat{p}_l = p_0 - Z_\alpha SE = 0.75 - 1.645(0.025) = 0.71$$

Therefore the fail to reject region, “acceptance region”, for the standard normal distribution is the interval $[-Z_\alpha, \infty) = [-1.645, \infty)$. The fail to reject region, “acceptance region”, for the sampling distribution is the interval $[\hat{p}_l, \infty) = [0.71, \infty)$.

¹Licensed under the GNU Free Documentation License or the Creative Commons Attribution-ShareAlike License. November 17, 2016



d. Test the hypothesis using the sample proportion, the test statistic and p-value at 5% significance.

The below illustrates three different ways of performing a hypothesis claim:

- *Using the sample proportion*

It is clear that the sample proportion $\hat{p} = \frac{216}{300} = 0.72$ lies in the fail to reject region. Hence one fails to reject the null hypothesis. But one should note that “even if we fail to reject the null hypothesis, we typically do not accept the null hypothesis as true. Failing to find strong evidence for the alternative hypothesis is not equivalent to accepting the null hypothesis,”[OS] although stating that “we accept the null hypothesis” will be allowed on the basis that it implies that the sample proportion was not enough evidence to reject the null hypothesis.

- *Using the test statistic*

The test statistic is:

$$Z = \frac{\hat{p} - p_0}{SE} = \frac{\frac{216}{300} - 0.75}{0.025} = -1.2$$

Observe that the test statistic lies in the fail to reject region of the standard normal distribution. Hence again it follows that the sample proportion fails to reject the null hypothesis.

- *Using the p-value*

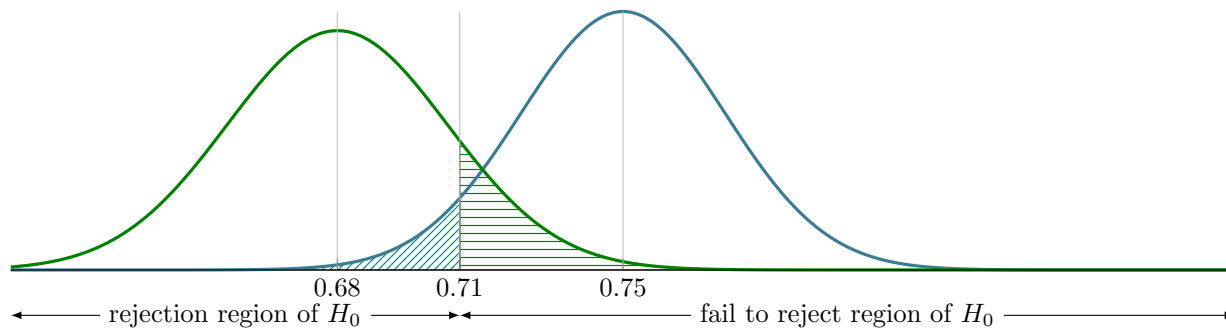
The *p-value* is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true.[OS]

In this problem, the p-value = $P(Z < -1.2) = 0.1151$. Since p-value $\geq \alpha$ one fails to reject the null hypothesis.

e. Determine the probability of committing an error (Type I or Type II) in the hypothesis test assuming that the true proportion of the population is $p_a = 0.68$.

Because the null hypothesis was not rejected, there is a possibility of committing a type II error. A type II error is committed when one fails to reject the null hypothesis given that the null hypothesis is false. Assuming an alternate proportion $p_a = 0.68$ as the true proportion, it is possible to compute the probability of committing a type II error.

Let \hat{P}_a be the random variable associated to the sampling distribution of sample proportions \hat{p}_a of sample size $n = 300$ and true population proportion assumed as $p_a = 0.68$. \hat{P}_a is centered at $\mu_{\hat{p}_a} = p_a = 0.68$ and has standard error $SE_a = \sigma_{\hat{p}_a} = \sqrt{\frac{p_a(1-p_a)}{n}} = \sqrt{\frac{0.68(0.32)}{300}} = 0.027$.



The probability of committing a type II error is the probability that a sample proportion of \hat{P}_a lies in the fail to reject region of H_0 .

$$\begin{aligned}\beta &= P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) \\ &= P(\hat{P}_a > 0.71) \\ &= P\left(Z > \frac{0.71 - p_a}{SE_a}\right) \\ &= P\left(Z > \frac{0.71 - 0.68}{0.027}\right) \\ &= P(Z > 1.11) = 1 - P(Z < 1.11) = 1 - 0.8665 = 0.1335\end{aligned}$$

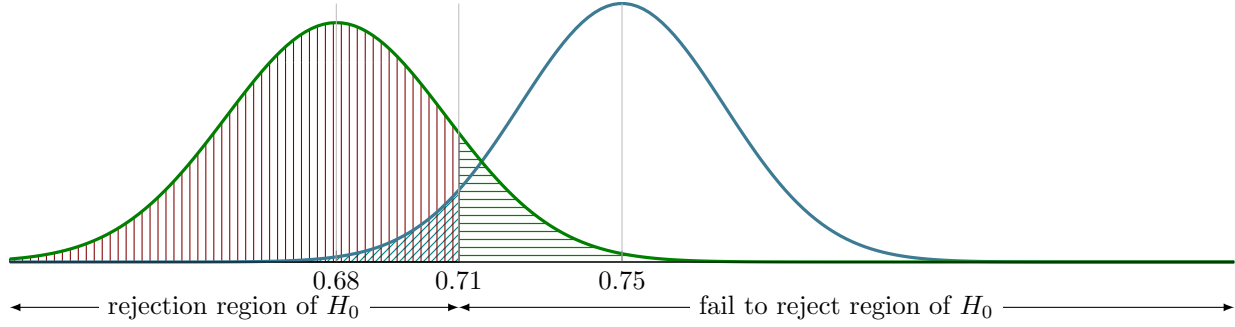
Therefore the probability of committing a type II error is about 13%, in the case where true population proportion is $p_a = 0.68$. The probability of committing a type II error is the area represented by horizontal stripes in the above diagram.

f. Determine the power of the test assuming that the true proportion of the population is $p_a = 0.68$.

Once the probability of a type II error is computed, one can determine the power of the test. That is

$$\begin{aligned}\text{power of a test} &= P(\text{reject } H_0 \mid H_0 \text{ is false}) \\ &= 1 - P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) \\ &= 1 - \beta = 1 - 0.13 = 0.87\end{aligned}$$

The larger the power of a test, the smaller the probability of committing a type II error.



The power of the test is the probability that a sample proportion of \hat{P}_a lies in the rejection region of H_0 . The probability is represented by the area with red vertical stripes in the above diagram.

g. Determine the sample size to control the error at $\alpha = \beta = 0.05$ assuming that the true proportion of the population is $p_a = 0.68$.

It is possible to control type I and type II error by changing the sample size of the sample proportions.

Comparing \hat{p}_l for both sampling distributions is required to determine the sample size. \hat{p}_l is the proportion which determines the boundary between the reject region and the fail to reject region of H_0 .

For the H_0 :

$$-Z_\alpha = \frac{\hat{p}_l - p_0}{SE}$$

$$\hat{p}_l = p_0 - Z_\alpha SE = 0.75 - 1.645 \sqrt{\frac{0.75(1 - 0.75)}{n}}$$

For \hat{P}_a :

$$Z_\beta = \frac{\hat{p}_l - p_a}{SE_a}$$

$$\hat{p}_l = p_0 + Z_\beta SE_a = 0.68 + 1.645 \sqrt{\frac{0.68(1 - 0.68)}{n}}$$

Set the two above formulas for \hat{p}_l equal and solve for n :

$$0.75 - 1.645 \sqrt{\frac{0.75(1 - 0.25)}{n}} = 0.68 + 1.645 \sqrt{\frac{0.68(1 - 0.68)}{n}}$$

$$0.75 - 0.68 = 1.645 \sqrt{\frac{0.68(1 - 0.68)}{n}} + 1.645 \sqrt{\frac{0.75(1 - 0.75)}{n}}$$

$$\sqrt{n} = 1.645 \frac{\sqrt{0.68(1 - 0.68)} + \sqrt{0.75(1 - 0.75)}}{0.75 - 0.68}$$

$$n = \left(1.645 \frac{\sqrt{0.68(1 - 0.68)} + \sqrt{0.75(1 - 0.75)}}{0.75 - 0.68} \right)^2$$

$$n = 446.81$$

Hence with a sample size of 447 both types of error are controlled at 5% assuming that the true population parameter is $p_a = 0.68$.

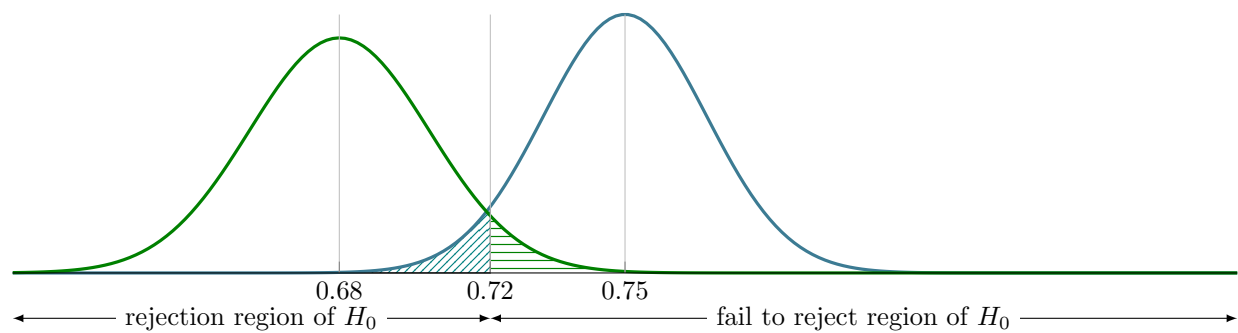
If the sample size is $n = 447$ then

$$SE = \sigma_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.75(0.25)}{447}} = 0.020$$

$$\hat{p}_l = p_0 - Z_\alpha SE = 0.75 - 1.645(0.020) = 0.717$$

$$SE_a = \sigma_{\hat{p}_a} = \sqrt{\frac{p_a(1 - p_a)}{n}} = \sqrt{\frac{0.68(0.32)}{447}} = 0.022$$

which results in the following sampling distributions.



References

- [OS] D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel. *OpenIntro Statistics*, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.