Single-sample hypothesis testing about a proportion

Example:

- We know from large databases of medical records that, among patients diagnosed with lung cancer when they are 40 years of age or older, the proportion that survive for 5 years after diagnosis is 0.082.
- We are interested in determining whether the proportion of 5-year survivors is the same in the population of patients diagnosed with lung cancer before age 40.
- The parameter of interest is the population proportion \( p \) in the population diagnosed with lung cancer before age 40.
- We will get data on a sample of persons under 40 who have been diagnosed with lung cancer.

Hypotheses

The null hypothesis says that the population proportion \( p \) in those diagnosed before age 40 is the same as the known proportion in those diagnosed at a later age.

\[
H_0 : p = 0.082
\]

The alternative hypothesis is two-sided because we do not know in advance in which direction a difference might go. (Younger people in general are more likely to survive for 5 years than older people, but perhaps a more severe form of lung cancer occurs in younger people.)

\[
H_a : p \neq 0.082
\]

Significance level

We choose to do our test at the \( \alpha = .05 \) significance level.

Data

From a 1991 article in the journal Cancer, we obtain data on a sample of 52 person diagnosed with lung cancer at age 40 or younger. Only 6 of them survived for 5 years after diagnosis.

The sample proportion was

\[
\hat{p} = \frac{6}{52} = 0.115
\]

The test statistic

The \( z \) test statistic is:

\[
z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}
\]

\[
= \frac{0.115 - 0.082}{\sqrt{\frac{0.082(1-0.082)}{52}}}
\]

\[
= 0.87
\]
The p-value

Because the test is two-sided, the p-value is the area under the standard normal curve more than 0.87 away from 0 in either direction. Table A tells us that the area to the left of -0.87 is 0.192. The p-value is twice this area:

\[ p = 2(0.192) = 0.384 \]

Conclusion

Can we reject the null hypothesis that \( p = 0.082 \)?

A proportion of survivors as far from 0.082 as what we found would happen 38% of the time if a sample of 52 patients were drawn from a population in which the true proportion of survivors was 0.082. Our result does not show that the proportion of 5-year lung cancer survivors is different in the population of patients diagnosed before age 40 from in the population diagnosed at age 40 or later.

The 95% confidence interval for the proportion \( p \) of patients diagnosed with lung cancer before age 40 who will survive 5 years is:

\[ \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.115 \pm 1.96 \sqrt{\frac{(0.115)(1-0.115)}{52}} \]

\[ = 0.115 \pm 0.087 \]

\[ = (0.028, 0.202) \]

Choosing the sample size for a desired margin of error

- Recall that the margin of error is the quantity that we add to and subtract from a point estimate in order to compute the right and left endpoints of a confidence interval.
- For a proportion, the confidence interval is

\[ \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

- so the margin of error is

\[ z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]
• Since we don’t know in advance what \( \hat{p} \) is going to be, we have to guess it. Call our guess \( p^* \). Some ways to make an “educated guess”:
  – Use a pilot study or past experience with similar studies.
  – Use \( p^* = 0.5 \). This is conservative, since it will give the largest possible margin of error.
• Then if \( m \) is the desired margin of error, the required sample size \( n \) is:
\[
    n = \left( \frac{z^*}{m} \right)^2 p^*(1 - p^*)
\]

• How would the sample size change if you had no previous information about what proportion to expect?

Example:
• PTC is a substance that has a strong bitter taste for some people and is tasteless for others. The ability to taste PTC is inherited. About 75% of Italians can taste PTC, for example.
• You want to estimate the proportion of Americans with at least one Italian grandparent who can taste PTC.
• Starting with the 75% estimate for Italians, how large a sample must you test in order to estimate the proportion of PTC tasters within \( \pm 0.04 \) with 95% confidence?

Sample size calculation for a hypothesis test regarding a single population proportion

• Consider a one-sided test:
\[
    H_0 : p = p_0 \\
    H_a : p < p_0
\]
• To compute sample size, we need to specify:
  – the significance level \( \alpha \)
  – a specific alternative hypothesis \( p = p_1 \)
  – the power \( 1 - \beta \)
• Then the sample size \( n \) is
\[
    n = \left( \frac{z_{1-\alpha} \sqrt{p_0(1-p_0)} + z_{1-\beta} \sqrt{p_1(1-p_1)}}{(p_1 - p_0)} \right)^2
\]
**Example:**

- Suppose in the PTC example that instead of just estimating \( p \) in Americans with at least one Italian grandparent, we wished to test the hypotheses:

\[
H_0 : p = .75 \\
H_a : p < .75
\]

- We choose \( \alpha = .05 \).
  - We would not consider the difference to be scientifically meaningful unless the true \( p \) were .60 or less, so we set \( p_1 = .6 \).
  - We want 90% power if the true \( p \) is .6.

- According to Table A
  
  \[
  z_{1-\alpha} = 1.645 \\
  z_{1-\beta} = 1.28
  \]

- So our sample size is

\[
\begin{align*}
n &= \frac{z_{1-\alpha} \sqrt{p_0(1-p_0) + z_{1-\beta} \sqrt{p_1(1-p_1)}}}{(p_1 \pm p_0)}^2 \\
&= \frac{1.645 \sqrt{.75(.25) + 1.28 \sqrt{.6(.4)}}}{(.6 - .75)}^2 \\
&= 8.929^2 \\
&= 79.73 \text{ or } 80
\end{align*}
\]

For a two-sided test, use \( z_{1-\frac{\alpha}{2}} \) instead of \( z_{1-\alpha} \) in the formula:

\[
n = \left[ \frac{z_{1-\alpha} \sqrt{p_0(1-p_0) + z_{1-\beta} \sqrt{p_1(1-p_1)}}}{(p_1 - p_0)} \right]^2
\]

In our example, this would be:

\[
n = \left[ \frac{1.96 \sqrt{.75(.25) + 1.28 \sqrt{.6(.4)}}}{(.6 - .75)} \right]^2 \\
&= 9.838^2 \\
&= 96.8 \text{ or } 97