Estimation of a Population Proportion

Example: We would like to estimate p, the probability that a person under the age of 40 who is diagnosed with lung cancer survives for at least 5 years

A random sample of $n = 70$ individuals is selected from the population

The number of "successes" X has a binomial distribution

$\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = npq$

It is found that only $X = 8$ patients out of the 70 survive for 5 years

The 5-year survival probability is estimated by the sample proportion of individuals who survive for 5 years

$$\begin{aligned} \widehat{p} &= \frac{X}{n} \\ &= \frac{8}{70} \\ &= 0.114 \end{aligned}$$

Note that the sample proportion $\widehat{p}$ is actually a sample mean

Let Y be a Bernoulli random variable that takes the value 1 if an individual survives for 5 years after diagnosis and 0 otherwise

Then $X = y_1 + y_2 + \ldots + y_{70} = \sum_{i=1}^{70} y_i$

and $\widehat{p} = [\sum_{i=1}^{70} y_i]/n$

If repeated samples of size 70 are selected from the population, what can be said about the distribution of sample proportions?

The distribution is called a **sampling distribution of proportions**

The mean is

$$\begin{aligned} \mathsf{E}(\widehat{p}) &= \mathsf{E}\left(\frac{X}{n}\right) \\ &= \left(\frac{1}{n}\right)\mathsf{E}(X) \\ &= p \end{aligned}$$

and the variance is

$$\begin{aligned} \mathsf{Var}(\widehat{p}) &= \mathsf{Var}\left(\frac{X}{n}\right) \\ &= \left(\frac{1}{n}\right)^2 \mathsf{Var}(X) \\ &= \frac{pq}{n} \end{aligned}$$

Applying the central limit theorem, the shape of the sampling distribution is approximately normal provided that n is large enough

$$\widehat{p} \sim \mathsf{N}(p, pq/n)$$

Note: We could have used the normal approximation to the binomial distribution

Since the distribution of X can be approximated by the distribution of a $\mathsf{N}(np, npq)$ random variable and $\widehat{p} = X/n$,

$$\widehat{p} \sim \mathsf{N}(p, pq/n)$$

How large does n need to be?

If p is known, we must have $npq \geq 5$

Since p is not known, we use $n\,\widehat{p}\,\widehat{q} \geq 5$

For the lung cancer 5-year survival data,
$n\,\widehat{p}\,\widehat{q} = 70(0.114)(0.886) = 7.1$

Since the sample size is large enough, the distribution of $\widehat{p}$ can be assumed to be normal

We use this information to find an interval estimate for p

First note that because $\widehat{p} \sim \mathrm{N}(p, pq/n)$,

$$z \;=\; \frac{\widehat{p} - p}{\sqrt{pq/n}}$$

has a standard normal distribution

$$\mathrm{P}(-1.96 \leq z \leq 1.96) \;=\; 0.95$$

Substituting for z,

$$\mathrm{P}\left(-1.96 \leq \frac{\widehat{p}-p}{\sqrt{pq/n}} \leq 1.96\right) \;=\; 0.95$$

Isolating p in the center of the inequality, we find that

$$\mathrm{P}\left(\widehat{p} - 1.96\sqrt{\frac{pq}{n}} \leq p \leq \widehat{p} + 1.96\sqrt{\frac{pq}{n}}\right) \;=\; 0.95$$

Therefore,

$$\left(\widehat{p} - 1.96\,\sqrt{\frac{pq}{n}}\,,\, \widehat{p} + 1.96\,\sqrt{\frac{pq}{n}}\right)$$

is a 95% confidence interval for p

There is a problem – this confidence interval depends on the true population mean p, which is not known

p must be estimated by $\widehat{p}$

Consequently,

$$\left(\widehat{p} - 1.96\sqrt{\frac{\widehat{p}\,\widehat{q}}{n}}\,,\, \widehat{p} + 1.96\sqrt{\frac{\widehat{p}\,\widehat{q}}{n}}\right)$$

is an approximate 95% confidence interval for p

An approximate 95% confidence interval for the proportion of individuals under the age of 40 who survive at least 5 years after being diagnosed with lung cancer is

$$\left(.114 - 1.96\sqrt{\frac{(.114)(.886)}{70}}\,,.114 + 1.96\sqrt{\frac{(.114)(.886)}{70}}\right)$$

or

$$(0.041, 0.188)$$

A general $100\% \times (1-\alpha)$ confidence interval for p takes the form

$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}\,\widehat{q}}{n}}\,,\, \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}\,\widehat{q}}{n}}\right)$$

where $z_{\alpha/2}$ is the value that cuts off the upper $\alpha/2 \times$ 100% of the standard normal curve

What do we do if the normal approximation to the binomial distribution cannot be applied?

An exact method for constructing a confidence interval can be used instead

This method is based on the binomial distribution itself

An exact $100\% \times (1-\alpha)$ confidence interval for p takes the form (p_1, p_2) where p_1 and p_2 satisfy the equations

$$\begin{aligned} \mathsf{P}(X \geq x \mid p = p_1) &= \sum_{k=x}^{n} \binom{n}{k} {p_1}^k {q_1}^{n-k} \\ &= \frac{\alpha}{2} \end{aligned}$$

and

$$\begin{aligned} \mathsf{P}(X \leq x \mid p = p_2) &= \sum_{k=0}^{x} \binom{n}{k} {p_2}^k {q_2}^{n-k} \\ &= \frac{\alpha}{2} \end{aligned}$$

X represents the number of successes

What is an exact 95% confidence interval for the proportion of individuals under the age of 40 who survive for at least 5 years after being diagnosed with lung cancer?

We must find p_1 such that

$$\begin{aligned} \mathrm{P}(X \geq 8 \mid p = p_1) &= \sum_{k=8}^{70} \binom{70}{k} {p_1}^k {q_1}^{70-k} \\ &= 0.025 \end{aligned}$$

and p_2 such that

$$\begin{aligned} \mathrm{P}(X \leq 8 \mid p = p_2) &= \sum_{k=0}^{8} \binom{70}{k} {p_2}^k {q_2}^{70-k} \\ &= 0.025 \end{aligned}$$

```
. cii 70 8
                                             -- Binomial Exact --
Variable |  Obs      Mean  Std. Err.   [95% Conf. Interval]
---------+--------------------------------------------------
         |   70  .1142857  .0380272    .0506557    .2128366
```

Recall that the 95% confidence interval based on the normal approximation was

$$(0.041, 0.188)$$

Example: We would like to estimate the probability that a child between the ages of 6 and 19 exhibits wheezing — a common respiratory symptom — given that his or her mother smokes in the home

In a sample of 400 children whose mothers smoke (but who do not themselves smoke), 42 children reported episodes of wheezing

A point estimate of the probability p is

$$\begin{aligned} \widehat{p} &= \frac{42}{400} \\ &= 0.105 \end{aligned}$$

Note: $n\,\widehat{p}\,\widehat{q} = (400)(0.105)(0.895) = 37.6$, which is considerably greater than 5

Using the normal approximation to the binomial distribution, an approximate 95% confidence interval for p is

$$\left(.105 - 1.96\sqrt{\frac{(.105)(.895)}{400}}, .105 + 1.96\sqrt{\frac{(.105)(.895)}{400}}\right)$$

or

$$(0.075, 0.135)$$

An exact 95% confidence interval for p is

$$(0.077, 0.139)$$

The two intervals are much closer together in this case

Given that no one living in the home smokes, the probability that a child between the ages of 6 and 19 exhibits wheezing is 0.08

Is the proportion of children who wheeze in homes where the mother smokes equal to the proportion who wheeze when no one in the home smokes?

Note that the 95% confidence interval for p contains the value 0.08

Therefore, 0.08 is a plausible value for p — it is possible that the two proportions are in fact the same

Hypothesis Testing for a Proportion

Example: We are interested in the probability of developing asthma over a given one-year period for children 0 to 4 years of age whose mothers smoke in the home

In the general population of 0 to 4-year-olds, the annual incidence of asthma is 1.4%

If 10 cases of asthma are observed over a single year in a sample of 500 children whose mothers smoke, is this compatible with an underlying probability of $p_0 = 0.014$?

We assume that cigarette smoke in the home cannot reduce the incidence of asthma

$$H_0 : p = p_0 = 0.014$$

$$H_A : p = p_1 > 0.014$$

H_0 would be rejected if $\widehat{p} = x/n$ is too big

One method of hypothesis testing relies on the normal approximation to the binomial distribution (central limit theorem)

This approximation is reasonable if $np_0q_0 \geq 5$

Under H_0,

$$\widehat{p} \sim \mathsf{N}\left(p_0, \frac{p_0q_0}{n}\right)$$

Therefore,

$$z = \frac{\widehat{p} - p_0}{\sqrt{p_0q_0/n}}$$

has a standard normal distribution

This is the test statistic

To conduct a one-sided test at the α level of significance, H_0 is rejected if $z > z_\alpha$

If $\alpha = 0.05$, then H_0 would be rejected for $z > 1.645$

For the asthma data, note that $n\,p_0\,q_0 = 500(0.014)(0.986) = 6.9$

Using the normal approximation,

$$\begin{aligned} \widehat{p} &= \frac{10}{500} \\ &= 0.02 \end{aligned}$$

and

$$\begin{aligned} \mathrm{z} &= \frac{0.02 - 0.014}{\sqrt{(0.014)(0.986)/500}} \\ &= 1.14 \end{aligned}$$

Since $1.14 < 1.645$, we do not reject H_0

We do not have sufficient evidence to conclude that the probability of developing asthma for children whose mothers smoke in the home is different from the probability in the general population

This is the critical value method of hypothesis testing — the p-value method could also be used

The p-value is the probability of obtaining a sample proportion as extreme or more extreme than the observed proportion $\widehat{p}$, given that H_0 is true

The area under the standard normal curve to the right of 1.14 is 0.1271

Therefore, $p = 0.1271$

Since $p > 0.05$, we again fail to reject H_0

An exact method of hypothesis testing uses the binomial distribution itself, rather than the normal approximation

If $\widehat{p} \leq p_0$, then the p-value for a one-sided test is

$$\begin{aligned} p &= \mathsf{P}(\leq x \text{ out of } n \mid H_0) \\ &= \sum_{k=0}^{x} \binom{n}{k} p_0^k \, q_0^{n-k} \end{aligned}$$

If $\widehat{p} > p_0$, then

$$\begin{aligned} p &= \mathsf{P}(\geq x \text{ out of } n \mid H_0) \\ &= \sum_{k=x}^{n} \binom{n}{k} p_0^k \, q_0^{n-k} \end{aligned}$$

For the asthma example, $\widehat{p} = 0.02$ is greater than $p_0 = 0.014$

Therefore,

$$\begin{aligned} p &= \mathrm{P}(\geq 10 \text{ out of } 500 \mid p = 0.014) \\ &= 1 - \mathrm{P}(< 10 \mid p = 0.014) \\ &= 1 - \sum_{k=0}^{9} \binom{500}{k} (0.014)^k (0.986)^{500-k} \\ &= 0.1681 \end{aligned}$$

```
. bitesti 500 10 .014

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
      500          10            7     0.01400      0.02000

  Pr(k >= 10)            = 0.168070  (one-sided test)
  Pr(k <= 10)            = 0.902981  (one-sided test)
  Pr(k <= 3 or k >= 10)  = 0.248373  (two-sided test)
```

Again we cannot reject $H_0 : p = 0.014$ at the 0.05 level

Example: We are interested in studying the cognitive abilities of children weighing less than 1500 grams at birth who experience perinatal growth failure, a condition preventing proper development

In the general population of children exhibiting normal growth in the perinatal period, 3.2% have an IQ score below 70 when they reach the age of 8 years

Is this also true for children who experience perinatal growth failure?

$$H_0 : p = 0.032$$

$$H_A : p \neq 0.032$$

We wish to conduct the two-sided test at the 0.01 level of significance

A random sample of 33 children with perinatal growth failure is selected

At the age of 8 years, 8 of the children have an IQ score below 70

The sample proportion is

$$\begin{aligned} \widehat{p} &= \frac{8}{33} \\ &= 0.242 \end{aligned}$$

Note that $n\,p_0\,q_0 = 33(0.032)(0.968) = 1.0$

Therefore, the normal approximation to the binomial distribution cannot be applied

We must use the exact binomial test

```
. bitesti 33 8 .032

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
       33           8        1.056       0.03200      0.24242

  Pr(k >= 8) = 0.000007  (one-sided test)
  Pr(k <= 8) = 0.999999  (one-sided test)
  Pr(k >= 8) = 0.000007  (two-sided test)
```

Since $p < 0.00001$ which is less than $\alpha = 0.01$, we reject the null hypothesis

For children who suffer from perinatal growth failure, the proportion who have an IQ score below 70 at the age of 8 years is not equal to 0.032, and is in fact higher

Power and sample size calculations can be performed for a binomial proportion p if the normal approximation applies

For a two-sided test of $H_0 : p = p_0$ versus $H_A : p \neq p_0$ for the specific alternative $p = p_1$, the power of the test is

$$\Phi\left(\sqrt{\frac{p_0 q_0}{p_1 q_1}}\left[-z_{\alpha/2} + \frac{|\, p_0 - p_1 \,|\sqrt{n}}{\sqrt{p_0 q_0}}\right]\right)$$

To achieve a power of $1 - \beta$, the sample size required is

$$n = \frac{p_0 q_0 (z_{\alpha/2} + z_\beta \sqrt{p_1 q_1 / p_0 q_0})^2}{(p_1 - p_0)^2}$$

Example: Among persons over the age of 40 who are diagnosed with lung cancer, 8.2% survive for at least 5 years

We wish to determine whether this proportion is the same for individuals under the age of 40 at the time of diagnosis

To conduct a two-sided test at the $\alpha = 0.05$ level of significance, the null and alternative hypotheses would be

$$H_0 : p = 0.082$$

and

$$H_A : p \neq 0.082$$

If the true population proportion of persons under the age of 40 at the time of diagnosis who survive 5 years is as high as 15%, we want to risk only a 10% chance of failing to reject H_0

$\Rightarrow$ Difference between p_0 and p_1 is 6.8%

P(type II error) $= 0.10$ and power $= 0.90$

Under the null hypothesis, $p_0 = 0.082$ and $q_0 = 0.918$

Under the alternative hypothesis, $p_1 = 0.15$ and $q_1 = 0.85$

$\alpha = 0.05 \;\Rightarrow\; z_{\alpha/2} = z_{0.025} = 1.96$

$\beta = 0.10 \;\Rightarrow\; z_{\beta} = z_{0.10} = 1.28$

Using the sample size formula,

$$\begin{aligned} n &= \frac{p_0 q_0 \left(z_{\alpha/2} + z_\beta \sqrt{p_1 q_1 / p_0 q_0}\right)^2}{(p_1 - p_0)^2} \\ &= \frac{(.082)(.918)\left(1.96 + 1.28\sqrt{\frac{(.15)(.85)}{(.082)(.918)}}\right)^2}{(.15 - .082)^2} \\ &= 214.0 \end{aligned}$$

Therefore, a sample of size 214 would be required

What would be the power of the test if the sample size were only $n = 100$?

In this case,

$$\begin{aligned}
\text{power} &= \Phi\left(\sqrt{\frac{p_0 q_0}{p_1 q_1}}\left[-z_{\alpha/2} + \frac{\mid p_0 - p_1 \mid \sqrt{n}}{\sqrt{p_0 q_0}}\right]\right) \\
&= \Phi\left(\sqrt{\frac{(.082)(.918)}{(.15)(.85)}}\left[-1.96 + \frac{|.082-.15|\sqrt{100}}{\sqrt{(.082)(.918)}}\right]\right) \\
&= \Phi(0.40) \\
&= 0.6554
\end{aligned}$$

The probability of rejecting the null hypothesis given that it is false and the true population proportion is actually 0.15 is approximately 66%

Hypothesis Testing for Two Proportions

For the most part, we have been applying the techniques of hypothesis testing to either continuous or ordinal data

What about nominal data?

We have performed one-sample tests for binomial proportions

This method can be generalized to allow the comparison of two proportions

Example: We are interested in determining whether the advice given by a physician during a routine physical examination is effective in encouraging patients to stop smoking

In a study of current smokers, one group of patients were given a brief talk about the hazards of smoking and were encouraged to quit

A second group received no smoking advice

All patients had a follow-up exam

We wish to test

$$H_0 : p_1 = p_2$$

against the alternative

$$H_A : p_1 \neq p_2$$

The two proportions come from independent populations

Among the 114 individuals who received advice to stop smoking, 11 reported that they had quit

$$\begin{aligned} \widehat{p}_1 &= \frac{x_1}{n_1} \\ &= \frac{11}{114} \\ &= 9.6\% \end{aligned}$$

Among the 96 patients who received no advice, 7 quit smoking

$$\begin{aligned} \widehat{p}_2 &= \frac{x_2}{n_2} \\ &= \frac{7}{96} \\ &= 7.3\% \end{aligned}$$

The null hypothesis would be rejected if $| \widehat{p}_1 - \widehat{p}_2 |$ is large

Assume that the sample sizes are large enough, and the normal approximation to the binomial distribution is valid

Therefore,

$$\widehat{p}_1 \sim \mathsf{N}\left(p_1\,,\,\frac{p_1 q_1}{n_1}\right)$$

and

$$\widehat{p}_2 \sim \mathsf{N}\left(p_2\,,\,\frac{p_2 q_2}{n_2}\right)$$

Furthermore,

$$\widehat{p}_1 - \widehat{p}_2 \sim \mathsf{N}\left(p_1 - p_2\,,\,\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

If the null hypothesis is true, then $p_1 = p_2 = p$

In this case

$$\widehat{p}_1 \sim \mathsf{N}\left(p\,,\,\frac{pq}{n_1}\right) \quad \text{and} \quad \widehat{p}_2 \sim \mathsf{N}\left(p\,,\,\frac{pq}{n_2}\right),$$

and

$$\widehat{p}_1 - \widehat{p}_2 \sim \mathsf{N}\left(0\,,\,\frac{pq}{n_1} + \frac{pq}{n_2}\right)$$

Therefore,

$$\begin{aligned} z &= \frac{(\widehat{p}_1 - \widehat{p}_2) - 0}{\sqrt{(pq/n_1) + (pq/n_2)}} \\ &= \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \end{aligned}$$

is the outcome of a standard normal random variable

Note that the ratio z depends upon the unknown proportion p

This proportion can be estimated by the weighted average of the two sample proportions

$$\begin{aligned} \widehat{p} &= \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2} \\ &= \frac{x_1 + x_2}{n_1 + n_2} \end{aligned}$$

Therefore, to test the null hypothesis $p_1 = p_2$ versus the alternative $p_1 \neq p_2$, we calculate the test statistic

$$\mathrm{z} = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\,\widehat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

We reject H_0 at the α level of significance if
$\mathrm{z} > z_{\alpha/2}$ or $\mathrm{z} < -z_{\alpha/2}$

This test should be used only if $n_1\,\widehat{p}\,\widehat{q} \geq 5$ and $n_2\,\widehat{p}\,\widehat{q} \geq 5$

For the smoking data, the estimated common proportion $\widehat{p}$ is

$$\begin{aligned} \widehat{p} &= \frac{x_1 + x_2}{n_1 + n_2} \\ &= \frac{11 + 7}{114 + 96} \\ &= 0.086 \end{aligned}$$

Since $114(.086)(.914) = 8.96 > 5$ and $96(.086)(.914) = 7.55 > 5$, the normal approximation can be applied

To test the null hypothesis

$$H_0 : p_1 = p_2,$$

we calculate the test statistic

$$
\begin{aligned}
z &= \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\,\widehat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\
&= \frac{0.096 - 0.073}{\sqrt{(.086)(.914)\left(\frac{1}{114} + \frac{1}{96}\right)}} \\
&= 0.59
\end{aligned}
$$

The area under the standard normal curve to the right of 0.59 is 0.2776

Therefore, $p = 0.5552$

We are unable to reject H_0 at the 0.05 level

The samples do not provide evidence that the proportion of patients who quit smoking differs in the two groups

The quantity $\widehat{p}_1 - \widehat{p}_2$ is a point estimate for the true difference in population proportions $p_1 - p_2$

We can also construct a confidence interval

First note that, if the normal approximation is valid,

$$\widehat{p}_1 - \widehat{p}_2 \sim \mathsf{N}\left(p_1 - p_2\,, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

This expression does not assume that H_0 is true ($p_1 = p_2$)

An approximate 95% confidence interval takes the form

$$(\widehat{p}_1 - \widehat{p}_2) \pm 1.96\sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_2\widehat{q}_2}{n_2}}$$

For the smoking data,

$$\begin{aligned}\sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1}+\frac{\widehat{p}_2\widehat{q}_2}{n_2}} &= \sqrt{\frac{(.096)(.904)}{114}+\frac{(.073)(.927)}{96}} \\ &= 0.0383\end{aligned}$$

Therefore, a 95% confidence interval for $p_1 - p_2$ is

$$(0.096 - 0.073) \pm 1.96(0.0383)$$

or

$$(-0.052\,,\, 0.0980)$$

This interval contains the value 0, and thus is consistent with the test of hypothesis conducted at the 0.05 level