Chapter 20, Fourth Edition : One Sample Proportions

Note : Here the 4-th and 5-th editions of the text have different chapters, but the material is the same.

There are many problems where the interest is in proportions of a population that fall into a given category. Opinion polls are commonly reported in the news, are undertaken on university campuses and are measures of the effectiveness of a medical drug or or other medical treatment.

The data in such a study consists of a simple random sample of individuals. These may be chosen for example as (i) a random sample of voters or other relevant population, (ii) randomly chosen individuals from the population of patients who undergo some medical intervention such as surgery or drug therapy.

For a random sample of size n, each data point or individual will then fall into one of two categories, which we may generically call *success* or *failure*. For coin tossing these categories are *heads* or *tails*. In an opinion poll these may be *support* or *do not support* the phrase or question being asked.

This data is from an EKOS political poll given in November 5, 2009. This information is taken from the URL http://www.scribd.com/doc/22161509/EKOS-national-opinion-poll-November-5-2009

There were $n = 3327$ observations or individuals in the sample. The support and percent support are amonsgt all voters in the sample, not just amongst the decided voters.

Table 1: EKOS Poll of November 5, 2009 for Voter Support for Canadian Federal Parties

Party	Conservative	Liberal	NDP	Green	Bloc Quebecois	Do not know
Support	1038	744	453	278	261	553
Percent	31.2	22.4	13.6	8.4	7.8	16.6

In this course we are only studying inference for a single population proportions problem. From Table **??** we could for example obtain such data, for example for support amongst Canadian voters for the Conservative Party. Let p represent the proportion of Canadian voters who support the Conservatives.

We have a sample of size $n = 3327$ and the observed number supporting the Conservatives as 1038. The observed proportion in the sample who support the Conservatives is then

$$\hat{p} = \frac{1038}{3327} = 0.312$$

The random variable $\hat{P}$ has a sampling distribution given by

$$\hat{P} \sim \text{ Normal , mean } = p, \text{ standard deviation } = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \tag{1}$$

This sampling distribution is approximate and it is a result of the Central Limit Theorem (Chapter 11) but beyond our discussion.

The random variable $\hat{P}$ is an unbiased estimator of the population parameter p. The standard deviation of $\hat{P}$ is given by

$$\text{standard deviation } = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

In some our calculations below we will also use

$$\text{SE} = \text{standard error } = \frac{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})}}{\sqrt{n}} \tag{2}$$

where $\hat{p}_{\text{obs}}$ is the observed value of the statistic $\hat{P}$, that is the observed sample proportion. Notice that the standard error is the same formula as the standard deviation but with the estimate $\hat{p}_{\text{obs}}$ substituted in place of p.

How can we use this sampling distribution to obtain a confidence interval for the population parameter p? Let us review the procedure we used earlier for our confidence interval. Specifically we use the 95% confidence interval.

From the sampling distribution given by equation (**??**) we obtain (writing Pr for probability for ease of reading)

$$\begin{aligned} 0.95 &= Pr(-z^* \leq \frac{\hat{P}-p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} \leq z^*) \\ &= P(-z^* \leq Z \leq z^*) \end{aligned}$$

where in the last line Z is standard normal. Thus from Table A we find $z^* = 1.96$. Next we use SE = standard error in place of the standard deviation; see equation (**??**) for this formula. The 95% confidence interval for the population parameter p is now given by solving for p from

$$-1.96 \leq \frac{\hat{p}_{\text{obs}}-p}{\frac{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})}}{\sqrt{n}}} \leq 1.96$$

This gives

$$\hat{p}_{\text{obs}} - 1.96 * \frac{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})}}{\sqrt{n}} \leq p \leq \hat{p}_{\text{obs}} + 1.96 * \frac{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})}}{\sqrt{n}}$$

or equivalently

$$\hat{p}_{\text{obs}} \pm 1.96 * \frac{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})}}{\sqrt{n}}$$

This is also

$$\hat{p}_{\text{obs}} \pm 1.96 * SE$$

which is completely analogous to our 95% confidence interval for a population mean.

The general $100(1-\alpha)\%$ confidence interval is

$$\hat{p}_{\text{obs}} \pm z^* \times \frac{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})}}{\sqrt{n}} \tag{3}$$

where z^* is the $(1-\alpha/2)$ critical value from the standard normal distribution, and is read from Table A.

What are the possible or reasonable values of the level of support for the Conservatives based on the EKOS poll? The 95% confidence interval for the voter support for Conservatives amongst Canadian voters is

$$\begin{aligned}
\hat{p}_{\text{obs}} \pm 1.96 * \frac{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})}}{\sqrt{n}} &= .312 \pm 1.96 * \frac{\sqrt{.312(1-.312)}}{\sqrt{3327}} \\
&= .312 \pm 1.96 * \frac{\sqrt{.312(1-.312)}}{\sqrt{3327}} \\
&= .312 \pm 1.96 * \sqrt{0.000064519} \\
&= .312 \pm 1.96 * 0.008032396 \\
&= .312 \pm 1.96 * 0.0157 \\
&= [.296, .328]
\end{aligned}$$

Thus we conclude that the voter support is between 29.6% and 32.8% at confidence level 95%.

At home verify that the 95% confidence interval for Liberal support is $0.224 \pm 1.96 * .00723 = .224 \pm 0.014$, or equivalently the interval $[0.210, .238]$.

How big should the sample size be for a specified margin of error? For a given confidence level 100*(1 - α) determine the critical value z^*. Suppose that we now the *true* or guessed value of p is given by p^*. For a given margin m of error we then solve for n from

$$m = z^* \times \frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}$$

This gives

$$n = \frac{z^*}{m} p^*(1-p^*)$$

If we do not known p^* (which is typical) then we use the worst case which correpsonds to $p^* = 0.5$. We then take the sample size n to be the next integer bigger than

$$n = \frac{z^*}{m} 0.5(1-0.5)$$

that is we round up. Table **??** gives these sample sizes for various margins of error. A sample size of 1500 or so is often affordable for an opinion polling company and its clients, so this sample size to produce a margin of error of 2.5 percentage points 19 times out of 20 (that is a 95% confidence interval with margin of error 0.025) is typically used.

Table 2: Sample Size for 95% Confidence Interval Margin of Error m

m	n
.03	1067
.025	1537
.02	2401
.01	9604

The confidence intervals are based ont he normal approximation to the sampling distribution of $\hat{P}$. Unless p is quite small (near 0) or quite large (near 1), this approximation works well for n of about size 20 or 30, or larger. When n is small there are better normal approximations. We will not consider these in our lecture.

Table 3: Table of Rejection Region forms and p-Value forms

H_a	determine z^*	p-value
$H_a : p < p_0$	$P(Z < -z^*) = \alpha$	$P(Z < z_{\text{obs}})$
$H_a : p > p_0$	$P(Z > z^*) = \alpha$	$P(Z > z_{\text{obs}})$
$H_a : p \neq p_0$	$P(\lvert Z\rvert > z^*) = \alpha$	$P(\lvert Z\rvert > \lvert z_{\text{obs}}\rvert)$

Hypothesis Testing for Proportions

Consider a null hypothesis $H_0 : p = p_0$ versus an alternative H_a. The alternative may be one sided or two sided, which will be used to determine the form of the rejection region for our decision rule.

The test statistic will be of the form as determined by equation (**??**) when the null hypothesis is true. Therefore we *know* in this case that $p = p_0$. Thus our test statistic is

$$Z = \frac{\hat{P} - p_0}{\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}} \tag{4}$$

We now proceed as we had done in our previous work on hypothesis testing. This is summarized in Table **??**.

In politics it is often of interest to know if their party support is above or below some threshold. For example the Conservative Part wishes to know if their support exceeds say one third proportion amongst Canadian voters. If they exceed this level of support they may wish to call an election as the governing party. This relevant null hypothesis is

$$H_0 : p = p_0 = \frac{1}{3}$$

since in this case and with no evidence they would not go into an election. The alternative is a one sided alternative

$$H_a : p > \frac{1}{3}$$

We use the data above. The observed sample proportion is $\hat{p}_{\text{obs}} = \frac{1038}{3327} = 0.312/$

The observed value of the test statistic is then

$$\begin{aligned} Z_{\text{obs}} &= \frac{\hat{p}_{\text{obs}} - p_0}{\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}} \\ &= \frac{.312 - .333}{\frac{\sqrt{.333(1-.333)}}{\sqrt{3327}}} \\ &= -0.719 \end{aligned}$$

We thus have observed p-value

$$\text{p-value} = P(Z > z_{\text{obs}}) = P(Z > -0.719) = .778$$

This is quite large and so there is no evidence against the null hypothesis in favour of the the alternative. Thus we conclude there is no evidence against $p = \frac{1}{3}$.