Dr Isabella Gollini

Probability and Statistics

11 Proportions

11.1 The sampling distribution of the proportion

As discussed in Section 4.3, one of the situations in which the binomial distribution arises is where, in a large (or infinite) population, a proportion p of individuals have a specified characteristic. When we take a random sample of size n from the population, the number of individuals X in the sample who have the characteristic has a binomial distribution, $X \sim B(n, p)$.

It follows from the properties of the binomial distribution that $E(X) = np$ and $\text{var}(X) = npq$, where $q = 1 - p$.

The proportion $\hat{p}$ of individuals in the sample who have the characteristic is given by

$$\hat{p} = \frac{X}{n}.$$

Hence

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{np}{n} = p \tag{1}$$

and

$$\text{var}(\hat{p}) = \text{var}\left(\frac{X}{n}\right) = \frac{\text{var}(X)}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}\ . \tag{2}$$

So $\hat{p}$ is an unbiased estimator of p with standard deviation $\sqrt{pq/n}$.

The exact probability distribution of $\hat{p}$ is a scaled version of the $B(n, p)$ distribution, but in practice an approximating normal distribution is often used. One way of deriving this approximation is by use of the Central Limit Theorem stated in Section 7.3.

If for $1 \leq i \leq n$ we define the independently and identically distributed r.v.s X_i by

$$X_i = \begin{cases} 1 & \text{if the } i\text{th sample member has the characteristic} \\ 0 & \text{if the } i\text{th sample member does not have the characteristic} \end{cases}$$

then the number of individuals in the sample who have the characteristic is $X \equiv \sum_{i=1}^{n} X_i$ and $\hat{p} = \sum_{i=1}^{n} X_i/n = \bar{X}$.

Applying the Central Limit Theorem, for large n, $\hat{p}$ has approximately a normal distribution,

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right)\ . \tag{3}$$

- Note that, for $1 \leq i \leq n, X_i \sim B(1, p)$, so that $E(X_i) = p$ and $\text{var}(X_i) = pq$. It follows that $E(\hat{p}) \equiv E(\bar{X}) = p$ and $\text{var}(\hat{p}) \equiv \text{var}(\bar{X}) = pq/n$, which is just another way of seeing the results of Equations (1) and (2).

- Sometimes the rule of thumb is used that the normal approximation is adequate provided that $np > 5$ and $nq > 5$. Since in practice p and q are unknown, the rule is used in the form $n\hat{p} > 5$ and $n\hat{q} > 5$, where $\hat{q} = 1 - \hat{p}$.

It follows from the result of Equation (3) that, at least approximately,

$$z \equiv \frac{\hat{p} - p}{\sqrt{pq/n}} \sim N(0,1) \ . \tag{4}$$

11.2 Estimation and confidence intervals for proportions

Consider again a random sample of size n from a large or infinite population in which a proportion p of individuals have a specified characteristic. Let $\hat{p}$ denote the proportion of individuals in the sample who have the specified characteristic.

As we saw from Equations (1) and (2), $\hat{p}$ is an unbiased estimator of p with standard deviation $\sqrt{pq/n}$. However, in the practical estimation problem, the values of p and q are unknown to us and we replace them by their estimates $\hat{p}$ and $\hat{q} \equiv 1 - \hat{p}$, respectively, to obtain the *standard error* of $\hat{p}$ as an estimator of p,

$$\sqrt{\frac{\hat{p}\hat{q}}{n}} \ . \tag{5}$$

From the result of Equation (4) we have that

$$\Pr\left(-x(50\alpha) < \frac{\hat{p} - p}{\sqrt{pq/n}} < x(50\alpha) \right) = 1 - \alpha \ ,$$

where $x(50\alpha)$ is the $50\alpha\%$ percentage point of the standard normal distribution. Rearranging the terms within the brackets on the left hand side, we obtain the result that, whatever the value of p, as we repeatedly take random samples of size n,

$$\Pr\left(\hat{p} - x(50\alpha)\sqrt{\frac{pq}{n}} < p < \hat{p} + x(50\alpha)\sqrt{\frac{pq}{n}} \right) = 1 - \alpha \ . \tag{6}$$

Now given a particular observed value of the sample proportion $\hat{p}$, a $100(1-\alpha)\%$ confidence interval for p is given by

$$\left(\hat{p} - x(50\alpha)\sqrt{\frac{\hat{p}\hat{q}}{n}} \ , \ \hat{p} + x(50\alpha)\sqrt{\frac{\hat{p}\hat{q}}{n}} \right) \ , \tag{7}$$

where, as a further approximation, in the limits on the left hand side of Equation (6) we have replaced the unknown p and q by their estimates $\hat{p}$ and $\hat{q}$, respectively.

Example

A sample of 200 voters in a certain constituency has been taken, and 73 of them say that they are planning to vote for the Conservative party at the next general election. What proportion of the voters in the constituency as a whole are planning to vote Conservative (or at least would say they are)? We assume that the sample is a random sample (which in practice might be a questionable assumption).

The estimated population proportion is $\hat{p} = 73/200 = 0.365$ or 36.5%. Using the formula (5), the associated standard error is

$$\sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.365)(0.635)}{200}} = 0.034 \ .$$

In R:

```
x <- 73
n <- 200
phat <- x / n
phat

## [1] 0.365

qhat <- 1 - phat
se <- sqrt(phat * qhat / n)
se

## [1] 0.03404225
```

From the expression (7) with $\alpha = 0.05$, and recalling that $x(2.5) = 1.96$, a 95% confidence interval for the population proportion p is given by

$$0.365 \pm (1.96)(0.034),$$

i.e.,

$$(0.298, 0.432) \ .$$

So a 95% confidence interval includes a range of values from about 30% to 43% of the voters in the constituency who are planning to vote Conservative.

In R:

```
alpha <- 0.05
zval95 <- qnorm(1 - alpha / 2)
zval95

## [1] 1.959964

c(phat - zval95 * se,
  phat + zval95 * se)

## [1] 0.2982784 0.4317216
```

11.3 Hypothesis testing for a proportion

Given the sample data, we may wish to test a hypothesis about the proportion p of individuals in the population who have the specified characteristic.

Given some particular value p_0, we might test the null hypothesis

$$H_0 : p = p_0$$

against the two-sided alternative hypothesis

$$H_1 : p \neq p_0$$

or against a one-sided alternative $H_1 : p > p_0$ or $H_1 : p < p_0$.

We shall reject H_0 if the difference $\hat{p} - p_0$ is large enough. Using the result of Equation (4), and writing $q_0 = 1 - p_0$, we use the test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \ , \tag{8}$$

which under H_0 has approximately the standard normal distribution.

Example

Suppose that we have tossed a coin 1000 times and observed 532 heads. Does this provide significant evidence that the coin is biased?

We are now sampling from an infinite population. Let p denote the long-term proportion of heads. The null hypothesis is that the coin is fair, i.e., that the long-term proportion of heads is $p_0 = 1/2$, so that we have $H_0 : p = 1/2$. We adopt the two-sided alternative hypothesis $H_1 : p \neq 1/2$. The observed proportion of heads is $\hat{p} = 532/1000 = 0.532$. Using the test statistic of Equation (8),

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0.532 - 0.5}{\sqrt{(0.5)(0.5)/1000}} = 2.024 \ .$$

Corresponding to the two-sided alternative hypothesis, we have a two-tail test. Using Table 4, the p-value that corresponds to the calculated z-value is

$$2(1 - \Phi(2.024)) = 2(1 - 0.9785) = 0.043 \ .$$

Since the p-value $p < 0.05$, we reject the null hypothesis that the coin is fair at the 5% significance level. So there is strong evidence that the coin is biased.

- Note that we are using the notation p for two different entities, (i) the population proportion and (ii) the p-value of the test statistic.

An alternative way of arriving at the same conclusion is to note that our calculated z-value exceeds the 2.5% percentage point of the standard normal distribution, $x(2.5) = 1.96$.

The test may readily be carried out using `R` by using the function `prop.test`. We have to insert the values of x, n and p in this order. The argument `p` is optional, and if it not specified the hypothesized probability of success is equal to $\frac{1}{2}$.

In this function, a continuity correction is applied. Since the number of successes is an integer, using a continuous Normal distribution to approximate the sampling distribution of the estimate of p may not be entirely accurate unless n is large. A continuity correction will compensate for that.

In order to get the same results as in the calculations above that didn't make use of the continuity correction we use the argument `correct = FALSE`.

```r
prop.test(532, 1000,
  correct = FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  532 out of 1000, null probability 0.5
## X-squared = 4.096, df = 1, p-value = 0.04298
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5010103 0.5627448
## sample estimates:
##     p
## 0.532
```

The function `prop.test` computes a chi-squared statistic, `X-squared` that is equivalent to the squared of the z-test statistic: $z = \sqrt{\texttt{X-squared}} = \sqrt{4.096} = 2.023858$, the p-value is the same as above.

In addition, for the given coin, a 95% confidence interval for the long-term proportion of heads is shown to be given by (0.5010,0.5627).

`R` uses a different method for calculating confidence intervals for one proportion, and this leads to a slightly different result to the one obtained by using equation (7).

```r
x <- 532
n <- 1000
phat <- x / n
qhat <- 1 - phat
se <- sqrt(phat * qhat / n)
alpha <- 0.05
zval95 <- qnorm(1 - alpha / 2)
c(phat - zval95 * se,
  phat + zval95 * se)

## [1] 0.5010738 0.5629262
```

11.4 Comparing two proportions

Suppose that we are considering two large populations, Population 1 and Population 2, where p_1 is the proportion of individuals in Population 1 who have a specified characteristic and p_2 is the proportion of individuals in Population 2 who have the characteristic, with $q_i = 1 - p_i$ $(i = 1, 2)$. Suppose further that we take a random sample of size n_1 from Population 1 and, independently, a random sample of size n_2 from Population 2. Let $\hat{p}_1$ and $\hat{p}_2$, respectively, be the sample proportions in the two samples.

	Population 1	Population 2
population proportion	p_1	p_2
sample size	n_1	n_2
number in sample with specified characteristic	X_1	X_2
sample proportion	$\hat{p}_1 = X_1/n_1$	$\hat{p}_2 = X_2/n_2$

An unbiased estimate of the difference $p_1 - p_2$ is given by $\hat{p}_1 - \hat{p}_2$. Furthermore,

$$\text{var}(\hat{p}_1 - \hat{p}_2) = \text{var}(\hat{p}_1) + \text{var}(\hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \ ,$$

which is estimated by

$$\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2} \ ,$$

where $\hat{q}_i = 1 - \hat{p}_i$ $(i = 1, 2)$. Thus the standard error of $\hat{p}_1 - \hat{p}_2$ is given by

$$\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \ .$$

Again using a normal approximation, a $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is given by

$$\left(\hat{p}_1 - \hat{p}_2 - x(50\alpha)\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \ , \ \hat{p}_1 - \hat{p}_2 + x(50\alpha)\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right) \ . \tag{9}$$

If, instead, we wish to test the hypothesis that the proportions in the two populations are the same,

$$H_0 : p_1 = p_2 \ ,$$

then under H_0 the two population proportions have a common value p whose value may be estimated by the pooled proportion $\hat{p}$ of individuals who have the characteristic, aggregated over both samples. If X_1 and X_2 represent the numbers of individuals who have the characteristic in Sample 1 and Sample 2, respectively, then

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} \ . \tag{10}$$

We shall reject H_0 if the difference $\hat{p}_1 - \hat{p}_2$ is large enough. Under H_0,

$$\text{var}(\hat{p}_1 - \hat{p}_2) = \text{var}(\hat{p}_1) + \text{var}(\hat{p}_2) = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

and the standard error of $\hat{p}_1 - \hat{p}_2$ is given by

$$\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} ,$$

where $\hat{p}$ is as given in Equation (10) and $\hat{q} = 1 - \hat{p}$. The test statistic for testing H_0 is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\,(1/n_1 + 1/n_2)}} , \tag{11}$$

which under H_0 has approximately the standard normal distribution.

Example (continued)

In an earlier example we considered a sample of 200 voters in a certain constituency, where 73 of them say that they are planning to vote for the Conservative party at the next general election. Suppose that a sample of size 100 is taken in a neighbouring constituency and 43 of those sampled say that they are planning to vote Conservative. Is there any significant evidence that the proportions of those planning to vote Conservative differ between the two constituencies?

Let p_1 and p_2, respectively, denote the proportions of voters in the two constituencies who are planning to vote Conservative. The estimated population proportions for the two constituencies are $\hat{p_1} = 73/200 = 0.365$ and $\hat{p_2} = 43/100 = 0.43$, respectively.

We test the null hypothesis

$$H_0 : p_1 = p_2$$

against the alternative

$$H_1 : p_1 \neq p_2 .$$

Under H_0 both constituencies have the same proportion p of voters planning to vote Conservative, which is estimated by the sample proportion aggregated over both samples as in Equation (10),

$$\hat{p} = \frac{73 + 43}{200 + 100} = \frac{116}{300} = 0.387 .$$

Using Equation (11), the value of the test statistic for testing H_0 is

$$z = \frac{0.365 - 0.43}{\sqrt{(0.387)(0.613)\,(1/200 + 1/100)}} = -1.090 .$$

Using a two-tail test and Table 4, the corresponding p-value is

$$2(1 - \Phi(1.090)) = 2(1 - 0.8621) = 0.276 .$$

Since the p-value $p > 0.20$, we do not reject the null hypothesis at the 5% significance level or even at the 20% significance level. (More simply, note that that $|z| < 1.2816 \equiv x(10)$.) So there is no strong evidence of any difference between the constituencies in the proportion of voters who plan to vote Conservative.

In `R` the function `prop.test` is used to carry out the test using the test statistic as specified in Equation (10).

The first argument of the function `prop.test` is a vector of length 2 containing the number of events in the two samples. The second argument is a vector of length 2 containing the number of trials in the two samples.

```
x1 <- 73
n1 <- 200

x2 <- 43
n2 <- 100

prop.test(c(x1, x2), c(n1, n2),
  correct = FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 1.1877, df = 1, p-value = 0.2758
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.18275902  0.05275902
## sample estimates:
## prop 1 prop 2
##  0.365  0.430
```

In addition a 95% confidence interval for $p_1 - p_2$, using the expression (9), is found. In the present case (-0.183, 0.053) is a 95% confidence interval for $p_1 - p_2$.

Extra Example

A machine in a factory must be repaired if it produces more than 10% defective items. A random sample of 100 items contains 15 defectives. The supervisor says that the machine must be repaired.
Does the sample evidence support his decision? Use a test with level $\alpha = .01$.