Chapter 5

Measurement of a proportion

In may situations a scientist has to measure the proportion of two things in a given system. As an example, take the classic Mendel experiment, where he cross-bred a tall variety and a short variety of pea plants and measured the proportion of the tall variety among the offspring. Similarly a chemist may have to measure the proportion of two chemical reagents as a reaction progresses. A geologist may have to measure the proportion of aluminium and silicon in a class of rocks. An anthropologist trying to work out the migration patterns in ancient times may have to measure the proportion of people in an area carrying a particular genetic marker. Then there is the well documented case of a species of moths evolving from white to brown due to selection pressure, because industrial pollution in Britain turned the tree barks brown. A scientist studying such an evolution in the field will have to measure the proportion of brown moths as time progresses. In all these cases, there is an objective proportion 'out there', and the scientist has to get as close to that as possible, by taking a small number of samples.

Two-valued distribution

To attack the problem, let us consider the case where something can take two values, like Mendel's experiment where the plant

can either be tall or short, a moth can be either white or brown.

Observing one at a time

Suppose we gather data about the population by observing one at a time. Let us call each observation as Y, which can take values 1 and 0. In the example of the brown and white moths,

$$Y = \begin{cases} 1 & \text{if the moth is brown} \\ 0 & \text{if the moth is white} \end{cases}$$

If we are trying to measure the proportion of brown moths in the population, 1 can be interpreted as 'success' and 0 as 'failure'. Suppose the actual proportion of these two in nature is that 1 occurs in 60% of the cases and 0 occurs in 40%. In general, assume that the probability of having $Y = 1$ is p and that of having $Y = 0$ is $(1-p)$.

What is the mean of Y? It is obtained as the weighted sum:

$$\mu_Y = p \times 1 + (1-p) \times 0 = p$$

Thus, for the specific example taken, the mean of Y is 0.6. Notice that no individual being studied can assume the mean value (they can only be either 0 or 1).

What is the variance of the distribution? This is calculated as the weighted sum of the squares of the distance from the mean. The distance of 1 from the mean is $(1-0.6)$ and of 0 from the mean is $(0-0.6)$. Thus, the variance is,

$$\sigma_Y^2 = 0.6(1-0.6)^2 + 0.4(-0.6)^2 = 0.24$$

In general,

$$\begin{aligned} \sigma_Y^2 &= p(1-p)^2 + (1-p)p^2 \\ &= p(1-2p+p^2) + (1-p)p^2 \\ &= p - 2p^2 + p^3 + p^2 - p^3 \end{aligned}$$

$$= \quad p(1-p) \tag{5.1}$$

Therefore the standard deviation is

$$\sigma_Y = \sqrt{p(1-p)} \tag{5.2}$$

This distribution is called the Bernoulli distribution.

Taking n samples at a time

In practice, it is not possible to make a very large number of observations, one at a time. Instead, a scientist has to take a handleable number of 'samples' from the population and has to record the observations. In this case one is trying to estimate the proportion of 1 in the population from the proportion of 1 observed in the sample.

Suppose you take 10 samples. If $p = 0.6$, will you get 6 ones and 4 zeros in the 10 samples? No. You may get any other proportion because of chance factors.

Now, what is the probability of getting 6 ones and 4 zeros?

You might get 6 ones if you had got it in the order: 0011101011. What is the probability of geting this sequence? Noticing that the probability of getting 1 is 0.6 and that of getting 0 is 0.4, we may write:

$$\begin{aligned} &P(\text{appearance of the sequence } 0011101011) \\ &= 0.4 \times 0.4 \times 0.6 \times 0.6 \times 0.6 \times 0.4 \times 0.6 \times 0.4 \times 0.6 \times 0.6 \\ &= 0.6^6 \times 0.4^4 \approx 0.0012 \end{aligned}$$

However, 6 1's and 4 0's can be obtained in many other ways. In all these cases the probability of getting that sequence will be the same, $0.6^6 \times 0.4^4$. Therefore to obtain the probability of getting 6 ones and 4 zeros, all these will have to be added up.

In how many ways can you get 6 ones in 10 samples? We know that it is '10 choose 6'. Thus the probability of getting 6 ones in 10

samples is

$$\begin{aligned} P(\text{6 1's in 10 samples}) &= \binom{10}{6} 0.6^6 \times 0.4^4 \\ &= \frac{10!}{6! \times (10-6)!} 0.6^6 \times 0.4^4 \approx 0.2508 \end{aligned}$$

This means, even though the probability of getting 1 is 0.6, if you draw 10 samples you will get 6 1's in only 25% of the cases.

In a similar way, you can find out the probability of getting other proportions. For example, the probability of getting 3 ones and 7 zeros will be

$$P(\text{3 1's in 10 samples}) = \binom{10}{3} 0.6^3 \times 0.4^7 \approx 0.0424$$

Now if we ask, what will be the distribution of the occurrence of ones in a samples of size 10, we may easily plot it from the above values. The resulting distribution is the binomial distribution. If you consider increasing the sample size, more and more possibilities will appear, as a result of which the graph will get smoother, and in the limit you will get a normal distribution. What will be the mean and standard deviation of this normal distribution?

Let X be the number of 1's observed in n samples. If you repeatedly take such n samples, in each case you will get different values of X. What will be the mean and standard deviation of this distribution? Given $p = 0.6$, if you take 100 samples, on an average 60 of them will be expected to be 1. Thus it stands to reason that the expected value of X should be np. Since X is a collection of n individual observations, i.e., Y,

$$X = Y + Y + \cdots n \text{ times}$$

Therefore the mean of X is the sum of the means of Y

$$\mu_X = \mu_Y + \mu_Y + \cdots n \text{ times} = np$$

The variance will similarly be n times the variance of Y, i.e.,

$$\mathrm{Var}(X) = n \cdot \mathrm{Var}(Y) = np(1-p)$$

and therefore the standard deviation is $\sigma_X = \sqrt{np(1-p)}$.

Estimating the distribution of the proportion

So you have made n observations and have obtained X 1's. The obtained 'sample proportion' of 1 in n observations is

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

You will get different values of $\hat{p}$ in different trials. How can you then infer the mean proportion in the population?

Notice that if you do a large number of trials, you will get a distribution of $\hat{p}$. The mean of that distribution will be

$$\mu_{\hat{p}} = \frac{\mu_X}{n} = \frac{np}{n} = p$$

And the standard deviation will be

$$\sigma_{\hat{p}} = \frac{\sigma_X}{n} = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$$

Will it be a normal distribution? It may not. For example, if the proportion of 1 in the population is very close to zero, the distribution will have a longer tail to the right (i.e., it is right skewed). But this can be offset by increasing the sample size because that will result in tighter distribution with a smaller standard deviation since $\sigma_{\hat{p}}$ is inversely proportional to $\sqrt{n}$.

There is a rule of thumb that if the probability p times the number of samples is greater than 10, and $(1-p)$ times the number of samples is also greater than 10, then the distribution of $\hat{p}$ is approximately a normal distribution. This give a great advantage in estimating the population mean of the proportion of 1 with a certain confidence. Notice that this thumb rule also

indicates the number of samples to be taken in order for this method to work.

Example 5.1: In the example of the population of moths in which 60% are brown and 40% are white, if you take 20 samples, you get these two numbers as $0.4 \times 20 = 8$ and $0.6 \times 20 = 12$. Since both are not greater than 10, you will not get a normal distribution of the samples means. You need to take a bigger sample size. □

Another point about the sample size: The population should be at least 10 times the sample size. If you are making observations on a dwindling population, like a species close to extinction, then the method will not work.

If you can ensure that $\hat{p}$ has more or less a normal distribution, a lot of information can be extracted from the data obtained from sampling. Let us illustrate it with an example.

Example 5.2: Suppose there is a population of tall and short pea plants in an experimental plot. You have sampled 50 plants and have found that 33 of them are tall. With what level of confidence can you state that the actual population proportion of tall plants lies in the range between 0.64 and 0.68?

Solution: The observed sample proportion is $\hat{p} = 33/50 = 0.66$.

The sample standard deviation is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Since we do not know the population proportion μ_p, the best we can do is to estimate it by using the sample proportion. Thus,

$$\sigma_{\hat{p}} = \sqrt{\frac{0.66(1-0.66)}{50}} \approx 0.067$$

Since $0.66 \times 50 = 33$ and $(1-0.66) \times 50 = 17$, i.e., both are above 10, if we could repeat the observations taking 50 samples at a time

the distribution of the proportions μ_s would be approximately a normal distribution with standard deviation close to 0.067.

Now, the range given is 0.66 ± 0.02. Therefore it is actually asking: What is the probability that the actual population proportion μ_p will lie within 0.02 of $\hat{p}$? This is the same as asking: What is the probability that the sample proportion is within 0.02 of population proportion?

In order to obtain the required confidence interval, one will have to obtain the probability

$$\begin{aligned}
&P(\mu_p \text{ is within } 0.02 \text{ of } \hat{p}) \\
&= P(\hat{p} \text{ is within } 0.02 \text{ of } \mu_p) \\
&= P(\hat{p} \text{ is within } 0.02/0.067 = 0.30 \text{ standard deviations of } \mu_p)
\end{aligned}$$

From the z-table we find that the area under the normal curve to the left of 0.15 standard deviations is 0.6179. Therefore the required area is $(0.6179 - 0.5) \times 2 = 0.2358$ or 23.58%.

Therefore one can state that that the population proportion lies in the range [0.64, 0.68] with only 23.58% confidence level. □

Example 5.3: What if, in the above example, we took 500 samples instead of 50? Suppose in this case we get 330 1's, i.e., the observed proportion remains the same

$$\hat{p} = 330/500 = 0.66$$

But the standard deviation reduces to

$$\sigma_{\hat{p}} = \sqrt{\frac{0.66(1-0.66)}{500}} \approx 0.021$$

In that case,

$$\begin{aligned}
&P(\mu_p \text{ is within } 0.02 \text{ of } \hat{p}) \\
&= P(\hat{p} \text{ is within } 0.02 \text{ of } \mu_p) \\
&= P(\hat{p} \text{ is within } 0.02/0.021 = 0.94 \text{ standard deviations of } \mu_p)
\end{aligned}$$

For $z = 0.94$, the area under the normal curve to the left of that value is 0.8264. Therefore the required confidence interval is $(0.8264 - 0.5) \times 2 = 0.6524$ or 65.24%. □

Example 5.4: Suppose an Institute decides to block internet access after midnight if it finds that more than 30% students are using the internet for non-academic purposes and are losing sleep. A survey is conducted on 150 students and finds that 57 of them are using the internet for 'wrong' purposes. Will it be logical to recommend blocking the internet access ('logical' implies confidence level of 95%)?

Solution: Here we are trying to assess the correctness of the statement that the proportion p of students engaging in unproductive practices on the internet after midnight exceeds 0.3. The null and alternative hypotheses are

$$\begin{aligned} H_0 &: p \leq 0.3 \\ H_1 &: p > 0.3 \end{aligned}$$

We start by assuming the null hypothesis, and check for the probability of getting the result $\hat{p} = 57/150 = 0.38$. If the probability is less than 5%, we reject the null hypothesis.

Since 0.38 is bigger than 0.3, the probability will be highest if we assume the highest population proportion satisfying the null hypothesis. Thus, we assume

$$\mu_{\hat{p}} = 0.3$$

Since both 0.3×150 and 0.7×150 are bigger than 10, the sampling distribution of proportions will be approximately a normal distribution with mean 0.3 and standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.3 \times 0.7}{150}} = 0.037$$

For $\hat{p} = 0.38$,

$$z = \frac{\hat{p} - \mu_{\bar{p}}}{\sigma_{\hat{p}}} = (0.38 - 0.3)/0.037 = 2.14,$$

i.e., $\hat{p}$ is 2.14 standard deviations away from the mean.

Is the probability of getting this z value less than 5%? To check, we refer to the z table, which shows that the value of z corresponding to 95% confidence level (i.e., 95% of the area under the curve) occurs at $z = 1.65$.

We have got a z value of 2.14. This means that the probability of getting that z value is less than 5%. Therefore we can reject the null hypothesis and can conclude that the proportion of students engaging in inappropriate practices exceeds 30%. □

Exercise

1. Suppose there is a population of tall and short pea plants in an experimental plot. You have sampled 100 plants and have found that 66 of them are tall. With what level of confidence can you state that the actual population proportion of tall plants lies in the range between 0.65 and 0.67?

2. Suppose there are two strains in a species of bacteria – one benign and the other virulent. You have read a paper that says more than 30% of the bacteria are virulent. You repeat the experiment and culture the bacteria. Then you pick up a bit of the culture and find 150 bacteria in it. On closer inspection, you find that 57 of them are virulent. Does your observation support the report in the paper with 95% level of confidence?