1 Population and Sample Proportion

Consider categorical data for a population of size N. If M individuals from the population belong to a certain group, we say that the proportion of the population that belongs to this group is $p = M/N$.

Now suppose that a sample of size m is randomly selected and k individuals from the sample belong to the group in question. We say that the proportion of the sample that belongs to this group is $\bar{p} = m/n$. The sample proportion may or may not equal the population proportion.

Since $\bar{p}$ was obtained through a random process, it is a random variable. Therefore, it has a set of possible values, a probability distribution, an expected value or mean, a variance, and a standard deviation. Since $\bar{p}$ represents a proportion, its set of possible values is limited to the interval between 0 and 1. We let $\mu_{\bar{p}}$ denote the mean of $\bar{p}$ and we let $\sigma_{\bar{p}}$ denote the standard deviation of $\bar{p}$.

It turns out that the mean and standard deviation of the sample proportion are related to the population proportion in the following way:

$$\mu_{\bar{p}} = p$$

That is, the mean or expected value of the sample proportion is the same as the population proportion. Notice that this does not depend on the sample size or the population size.

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \underbrace{\sqrt{\frac{N-n}{N-1}}}_{\text{FPCF}}$$

The finite population correction factor appears again. We can ignore it in the same three cases that we did when considering the sample mean. Observe that, as the sample size n increases, the standard deviation of the sample proportion gets smaller. That is, as the sample size increases, the sample proportion becomes more likely to be closer to the population proportion.

Notice that we have not said anything about the distribution of $\bar{p}$ so far other than its mean and standard deviation. For all we know at this point, it could follow a normal distribution, or a uniform distribution, or any distribution really. We will give a more precise description of the distribution of $\bar{p}$ later.

As an example, suppose that a family has five people, A, B, C, D, and E. A and D are women and B, C, and E are men. This is our population data. The proportion of the population which is men is $p = 3/5$.

Now suppose that we obtain a simple random sample of 2 people from the family, without replacement. That is, the sample must consist of 2 different people. From Lecture 7, we know that there are ${}_5C_2 = \dfrac{5 \cdot 4}{2 \cdot 1} = 10$ possible ways of doing this. Each pair of people is equally likely to occur, with probability 1/10. For each different sample, we will get a (perhaps) different value for $\bar{p}$, the proportion of men in the sample. For example, if the sample consists of people A and B, then $\bar{p}$ is 1/2. We can then fill in the rest of the table below.

sample	$\bar{p}$
A,B	1/2
A,C	1/2
A,D	0/2
A,E	1/2
B,C	2/2
B,D	1/2
B,E	2/2
C,D	1/2
C,E	2/2
D,E	1/2

In the second column, we see all the possible values of $\bar{p}$. The probability distribution of $\bar{x}$ is:

k	$P(\bar{p} = k)$
0/2	1/10
1/2	6/10
2/2	3/10

From this, we can use techniques of Lectures 8 and 9 to compute the mean and standard deviation of $\bar{p}$ using a table. For example, the next step in computing the mean would be to compute the values of $kP(\bar{p} = k)$ for all the possible values k. The mean would then be the sum of those values. We could compute the standard deviation from the variance. Computing the variance requires several more columns.

Since $\bar{p}$ is a sample proportion, we don't actually need to use these old techniques here. We can use formulas to compute the mean and standard deviation of the sample proportion. The mean of $\bar{p}$ is simply the population proportion, so

$$\mu_{\bar{p}} = p = 3/5.$$

Since our population is small and the sample is 40% of the population and the sample did not allow replacement, we must include the FPCF in our computation of the standard deviation of the sample proportion, so

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{(3/5)(1-3/5)}{2}}\sqrt{\frac{5-2}{5-1}} = \frac{3}{10}.$$

2 Central Limit Theorem for Proportion

Consider a population of size N with proportion p belonging to a certain category. Also consider the proportion $\bar{p}$ belonging to the same category from a randomly selected sample of size n. Provided that $np \geq 5$ and $n(1-p) \geq 5$, $\bar{p}$ is normally distributed.

3 Main Ideas for Confidence Intervals for Proportion

The problem of estimating the population proportion using the sample proportion is analogous to the problem of estimating the population mean using the sample mean. Under the conditions described above, $\bar{p}$ is normally distributed with mean p and standard deviation $\sigma_{\bar{p}}$. Therefore, $\dfrac{\bar{p}-p}{\sigma_{\bar{p}}}$ follows a standard normal distribution.

Once again, we can convert between:

- confidence levels and z-scores
- z-scores and maximum sampling errors E (provided that $\sigma_{\bar{p}}$ is known)
- maximum sampling errors E and confidence intervals

Computation of $\sigma_{\bar{p}}$ requires knowledge of $\bar{p}$. Typically, $\bar{p}$ is unknown, and we approximate $\sigma_{\bar{p}}$ with

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\sqrt{\frac{N-n}{N-1}},$$

which uses $\bar{p}$ in places where the formula for $\sigma_{\bar{p}}$ used p. Provided that $n\bar{p} \geq 5$ and $n(1-\bar{p}) \geq 5$, $\dfrac{\bar{p}-p}{s_{\bar{p}}}$ follows a standard normal distribution.

To get a maximum sampling error E from a z score, we use the formula

$$E = zs_{\bar{p}}.$$

To get a z-score from a maximum sampling error E, we us the inverse formula

$$z = \frac{E}{\bar{p}}.$$

If the sample size is unknown, but we know the z-score and the maximum sampling error E, we can force the appropriate relationship between z and E to hold by setting

$$E = zs_{\bar{p}} = z\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}.$$

This assumes that the FPCF is not needed. Dividing both sides by z and then squaring both sides, we find that

$$\left(\frac{E}{z}\right)^2 = \frac{\bar{p}(1-\bar{p})}{n}.$$

Therefore,

$$n = \left(\frac{z}{E}\right)^2 \bar{p}(1-\bar{p}).$$

Be careful to express proportions as decimals or fractions (not as percents) when doing computations.

4 Confidence Levels from Confidence Intervals

Suppose we have 3 presidential candidates, A, B, and C. We randomly select 160 people. Of them, 120 vote for candidate A. How confident can we be that between 70% and 80% of the population will vote for candidate A?

For this sample, $\bar{p} = 120/160 = 0.75$ and Observe that $np = 120$ and $n(1-p) = 40$. Since these are both at least 5, the central limit theorem tells us that $\dfrac{\bar{p}-p}{s_{\bar{p}}}$ has a standard normal distribution.

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{0.75(1-0.75)}{160}} = 0.0342.$$

First we find the maximum sampling error E. The confidence interval is 0.7 to 0.8, so the maximum sampling error is $E = 0.05$.

Second we find the z-score using the formula

$$z = \frac{E}{s_{\bar{p}}} = \frac{0.05}{0.0342} = 1.46.$$

Third we find the corresponding confidence level. We look up $z = 1.46$ on the normal table and find that it corresponds to an area of 0.4279. This means that the confidence level is $2 \cdot 0.4279 = .8558$. That is, we can be 85.58% confident that the true population proportion is between 0.7 and 0.8.

5 Confidence Intervals from Confidence Levels

Suppose we have 3 presidential candidates, A, B, and C. We randomly select 160 people. Of them, 120 vote for candidate A. Find a 95% confidence interval for the proportion of the population that will vote for candidate A.

Again we know that $\bar{p} = 0.75$ and $s_{\bar{p}} = 0.0342$.

First we find the z-score corresponding to a 95% confidence level. By the 68-95-99.7 rule, $z = 2$.

Second we find the maximum sampling error E using the formula

$$E = zs_{\bar{p}} = 2 \cdot 0.0342 = 0.0684.$$

Third we find the corresponding confidence interval. The left endpoint is $0.75 - 0.0684 = 0.682$ and the right endpoint is $0.75 + 0.0684 = 0.818$.

6 Sample Size Determination

Suppose that, in our sample, 75% of people vote for candidate A. How large must our sample size be so that we can be 99.7% confident that between 72% and 78% of the population vote for A?

First we find the maximum sampling error E. The confidence interval is 0.72 to 0.78, so the maximum sampling error is $E = 0.03$.

Second we find the z-score corresponding to a 99% confidence level. That is, we look up an area of 0.495 on the standard normal table and we find $z = 2.575$.

Third we find the sample size using the formula

$$n = \left(\frac{z}{E}\right)^2 \bar{p}(1-\bar{p}) = \left(\frac{2.575}{0.03}\right)^2 (0.75)(1-0.75) = 1381.3.$$

Since n represents the number of individuals in our sample, it does not make sense to say $n = 1381.3$. A sample size of 1381 is not big enough to guarantee that our confidence interval has the desired confidence level. However, a sample size of 1382 is.