8.5: Calculating the Sample Size n- Continuous and Binary Random Variables

Continuous Random Variables

Usually,► we have no control over the sample size of a data set. However, if we are able to set the sample size, as in cases where we are taking a survey, it is very helpful to know just how large it should be to provide the most information. Sampling can be very costly in both time and product. Simple telephone surveys will cost approximately $30.00 each, for example, and some sampling requires the destruction of the product.

If we go back to our standardizing formula for the sampling distribution for means, we can see that it is possible to solve it for n. If we do this we have $(\overline{X}-\mu)$ in the denominator.

$$n=\frac{Z_\alpha^2\sigma^2}{(\overline{X}-\mu)^2}=\frac{Z_\alpha^2\sigma^2}{e^2}$$

Because we have not taken a sample yet we do not know any of the variables in the formula except that we can set Z_α to the level of confidence we desire just as we did when determining confidence intervals. If we set a predetermined acceptable error, or tolerance, for the difference between $\overline{X}$ and μ, called e in the formula, we are much further in solving for the sample size n. We still do not know the population standard deviation, σ. In practice, a pre-survey is usually done which allows for fine-tuning the questionnaire and will give a sample standard deviation that can be used. In other cases, previous information from other surveys may be used for σ in the formula. While crude, this method of determining the sample size may help in reducing cost significantly. It will be the actual data gathered that determines the inferences about the population, so caution in the sample size is appropriate calling for high levels of confidence and small sampling errors.

When the population standard deviation is unknown, it can be estimated by taking the difference of the maximum value and the minimum value, the range divided by 6. The Empirical rule indicates that 99.7% of the data is between 3 standard deviations.

$$\mu+3\sigma-(\mu-3\sigma)=6\sigma \tag{8.5.1}$$

$$est.\,\sigma=\frac{max-min}{6} \tag{8.5.2}$$

$$n=\frac{Z_\alpha^2\sigma^2}{(\overline{X}-\mu)^2}=\frac{Z_\alpha^2(range/6)^2}{e^2}$$

Binary Random Variables

What was done in cases when looking for the mean of a distribution can also be done when sampling to determine the population parameter p for proportions. Manipulation of the standardizing formula for proportions gives:

$$n=\frac{Z_\alpha^2\mathrm{pq}}{e^2}$$

where $e=(p'-p)$, and is the acceptable sampling error, or tolerance, for this application. This will be measured in percentage points.

In this case the very object of our search is in the formula, p, and of course q because $q=1-p$. This result occurs because the binomial distribution is a one parameter distribution. If we know p then we know the mean and the standard deviation. Therefore, p shows up in the standard deviation of the sampling distribution which is where we got this formula. If, in an abundance of caution, we substitute 0.5 for p we will draw the largest required sample size that will provide the level of confidence specified by $Z\alpha$ and the tolerance we have selected. This is true because of all combinations of two fractions that add to one, the largest multiple is when each is 0.5. Without any other information concerning the population parameter p, this is the common practice. This may result in oversampling, but certainly not under sampling, thus, this is a cautious approach.

There is an interesting trade-off between the level of confidence and the sample size that shows up here when considering the cost of sampling. Table $8.5.1$ shows the appropriate sample size at different levels of confidence and different level of the acceptable error, or tolerance.

Table 8.5.1

Required sample size (90%)	Required sample size (95%)	Tolerance level
1691	2401	2%
752	1067	3%
271	384	5%
68	96	10%

This table is designed to show the maximum sample size required at different levels of confidence given an assumed $p = 0.5$ and $q = 0.5$ as discussed above.

The acceptable error, called tolerance in the table, is measured in plus or minus values from the actual proportion. For example, an acceptable error of 5% means that if the sample proportion was found to be 26 percent, the conclusion would be that the actual population proportion is between 21 and 31 percent with a 90 percent level of confidence if a sample of 271 had been taken. Likewise, if the acceptable error was set at 2%, then the population proportion would be between 24 and 28 percent with a 90 percent level of confidence, but would require that the sample size be increased from 271 to 1,691. If we wished a higher level of confidence, we would require a larger sample size. Moving from a 90 percent level of confidence to a 95 percent level at a plus or minus 5% tolerance requires changing the sample size from 271 to 384. A very common sample size often seen reported in political surveys is 384. With the survey results it is frequently stated that the results are good to a plus or minus 5% level of "accuracy".

Example 8.5.9

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Answer

Solution 8.9

From the problem, we know that the acceptable error, e, is **0.03** (3%=0.03) and $z_{\frac{\alpha}{2}} Z_{0.05} = 1.645$ because the confidence level is 90%. The acceptable error, e, is the difference between the actual population proportion p, and the sample proportion we expect to get from the sample.

However, in order to find n, we need to know the estimated (sample) proportion p'. Remember that $q' = 1 - p'$. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because $p'q' = (0.5)(0.5) = 0.25$ results in the largest possible product. (Try other products: $(0.6)(0.4) = 0.24; (0.3)(0.7) = 0.21; (0.2)(0.8) = 0.16$ and so on). The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$n = \frac{z^2 p' q'}{e^2}$ gives $n = \frac{1.645^2 (0.5)(0.5)}{0.03^2} = 751.7$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Exercise 8.5.9

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?