Survey Sampling in the Pet Care Industry

By Joshua Candamo

It is time for the pet care industry to storm into the era of statistics. Statistics is the science that focuses on the collection, organization, and analysis of numerical facts, which we call data. The study of statistics involves math and calculations of numbers, but it also relies heavily on how the numbers are chosen and how the statistics are interpreted.

More and more, pet care providers are leaning towards decision-making processes that are increasingly dependent upon the collection and interpretation of data. Data is available in most cases, so the problem comes down to properly evaluate the data in order to effectively use it to improve business practices.

Sampling is the technique of selecting a representative part of a population. Sometimes, the entire population will be sufficiently small, and it can be included completely in the study. This type of research is called a census study because data is gathered on every member of the population. In general, the population is too large to attempt to survey it entirely. A carefully chosen portion of the population (a sample) can be used to represent the entire population. So, how big should the sample size be? Well… the bigger the better, but at the very least 30 members from the population are required. A trade-off exists between maximizing confidence and minimizing costs. The main idea is that a sample reflects the general characteristics of the population from which it was drawn.

There are many ways to methodically perform the sample selection. The key is to find what is called an unbiased sample. A sample is unbiased if every member of the population has an equal chance of being selected. The simple random sampling is a technique that can be used to generate an unbiased sample for surveys. Let's say we want to find out how likely our daycare customers are to buy a buy-1-get-1-free bath offer. First, obtain a list of all your current daycare clients, and then using a sequence of numbers from a random numbers table, select say 10-30% of names on that list. The selected names are a random sample of the daycare population.

Once the sample is selected, is time to gather the data. As a rule of thumb, use a mix of interviewers, make sure questions are consistent, avoid jargon, and keep to short and simple questions. The survey results not only depend on the questions asked, but also in the way the questions are asked.

After the data has been collected we are ready to compute statistics that characterize the population in general. If you were to measure the weight of the puppies in a litter of golden retrievers, you would clearly find that they all in fact have different weights. So, it is impossible to report the weight of puppies for that litter, instead the best you can do is to report a typical weight and give some estimate of the range of variation above and below that typical weight. For that purpose we compute two statistics called mean and standard deviation.

To find the mean (a.k.a. average) add up all the values in a set of data and then divide that sum by the number of values in the dataset. Let's say a golden retriever litter has 3 puppies. The puppies weight 4, 5, and 9 pounds respectively. The mean puppy weight is the following:

$$mean = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{4+5+9}{3} = 6$$

The standard deviation measures how widely values are dispersed from the mean. The standard deviation for the weight of the puppies in the litter is:

$$s.d. = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - mean)^2}$$

$$s.d. = \sqrt{\frac{(4-6)^2 + (5-6)^2 + (9-6)^2}{3-1}} = 2.7$$

Once we have some statistics in our hands, we can estimate the accuracy of the survey. We want the sample to be as representative to the

population as possible; however, a sampling error will always exist. The sampling error has to do with the differences between the sample and the population, which are due solely to the particular units that were selected. In simple terms, we want to know how confident we can be about the results of our survey. If we want to be 95% certain of our results, we pick the *z*-score corresponding to 95% from the following table (z=1.96).

Confidence Interval Values
• 99% Confidence interval, z = 2.575 • 95% Confidence interval, z = 1.96 • 90% Confidence interval, z = 1.645 • 80% Confidence interval, z = 1.28

And substitute z in the next formula.

$$Error = z\left(\frac{s.d.}{\sqrt{N}}\right)$$

The *s.d.* above represents the standard deviation of the population (not the sample). Commonly, pilot studies are set to approximate it. Finally, compute the confidence interval using:

$$Interval = mean \pm Error$$

Let's use an illustrative example to clarify. The front desk of a boarding kennel was recorded for 32 randomly selected customers. The time that took the customer to leave from the moment he entered was measured. The measured times had a mean of 520 seconds and a standard deviation of 270 seconds. We want to construct a 95% confidence interval for the population mean.

$$Error = 1.96\left(\frac{270}{\sqrt{32}}\right) = 94$$

$$Interval_{95\%} = (520 - 94\ ,\ 520 + 94) = (426\ ,\ 614)$$

So, what does the confidence interval represent? In rough terms, it gives an estimate of the margin of error of the study. In other words, if we make 100 surveys with the same sample size, in 95 of them the true population average will be within the confidence interval. How is that useful? We can determine the sample size needed for future studies. Determining sample size is a very important issue because a large sample may waste time, resources and money, while samples that are too small may lead to inaccurate results. Let's follow up on the previous example.

Say that an error of 94 seconds is too imprecise. We would like instead to have an error of only 60 seconds. How many customers do we need to sample to compute the 95% confidence interval? Let's find out:

$$N = z^2\left(\frac{s.d.}{Error}\right)^2 = 1.96^2\left(\frac{270}{60}\right)^2 = 78$$

As expected, as the sampling error decreases the sample size required is larger. So, by sampling 46 more customers than the previous study we can reduce the sampling error by 64%, achieving a 95% confidence interval with an error of 60 seconds.

The pet care industry realizes the importance of information. Statistics is a mathematical science that enables the use of information to improve decision making. It is all about interpretation…. It's about asking why… Analysis is definitely not about patterns or numbers, is about meaning.

About the Author

Joshua Candamo has been involved in the boarding kennel industry, breeding, and show handling all his life. Mr. Candamo is a Computer Science PhD candidate at the University of South Florida, working in the areas of Artificial Intelligence, Pattern Recognition, and Computer Vision. Joshua has been invited all over the United States to give seminars about applying technology and statistics to the boarding kennel industry, and talks about his research applications.