R Project 2:
Sampling Distributions

Data Analytics I – ADS522

James Sears

BAY PATH UNIVERSITY 1897

Applied Data Science Program, Graduate School, Longmeadow, MA

Submitted To: Professor Aja Shabana

Spring 2021

Jim Sears
ADS522Z2 – Data Analytics I
R Project 2

Introduction

The objectives of this project are to explore the binomial distribution and its normal approximation and simulate a sampling distribution of a sample proportion. The following scenario will be used to demonstrate these objectives:

A store owner estimates that the number of customers who leave the store without making a purchase is 15%.

Binomial Distribution

A discrete probability experiment is where the random variable (a numerical measure of the outcome) is a finite number or countable number of values. If the experiment only has two mutually exclusive (disjoint) outcomes, and meets several other conditions, it is a specific type of discrete probability experiment known as a binomial probability experiment. The scenario above that estimates that 15% of customers leave a store without making a purchase qualifies as a binomial experiment as it meets the required conditions:

1. The experiment is performed a fixed number of times (trials):
   The same number of customers (trials) are analyzed in each experiment.
2. The trials are independent:
   Assuming that there is unlimited supply, the purchasing behavior of any customer or customers does not affect that of any other customer or customers.
3. For each trial, there are two mutually exclusive (or disjoint) outcomes:
   A customer can only leave the store after making a purchase (failure) or without making a purchase (success).
4. The probability of success is fixed for each trial of the experiment:
   The estimated proportion of customers that leave the store without making a purchase (success) is set at 15% for all trials; there is always a 15% probability of success.

As the value of a random variable (X) in a binomial probability experiment is determined by chance, probabilities can be assigned to each of the possible values of that random variable. These discrete values and their corresponding probabilities are known as the binomial probability distribution. In a probability distribution, each individual probability is between zero and one, and the sum of all the probabilities is one. Further, the binomial probability distribution has a specific function to calculate the probability of success for any value of random variable:

$$P\left(x\right) = {}_{n}C_{x}p^{x}\left(1-p\right)^{n-x}$$

where x = number of successful outcomes (the number of customers of that leave the store without making a purchase), *n* = the number of trials in the experiment (the fixed total number of customers analyzed), and *p* = the probability of success (the 15% chance that a customer leaves the store without making a purchase).

If the number of trials is set at 10 customers, the probability of 0 of those 10 customers leaving the store without making a purchased is calculated to be 0.1969. The probability for 1 customer is 0.1298, for 2

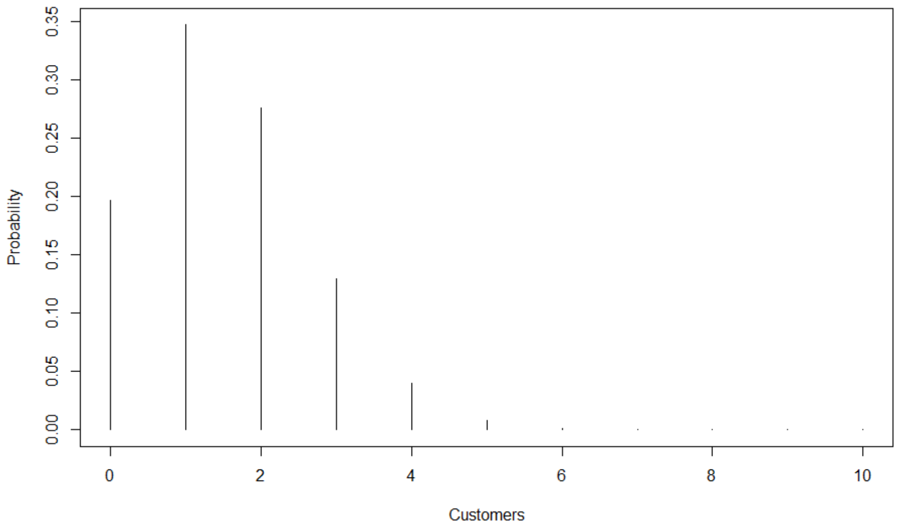

Figure 1a. Binomial probability distribution of 10 (n) customers where p = 0.15. Mean (μ) = 1.5, standard deviation (σ) = 1.13.

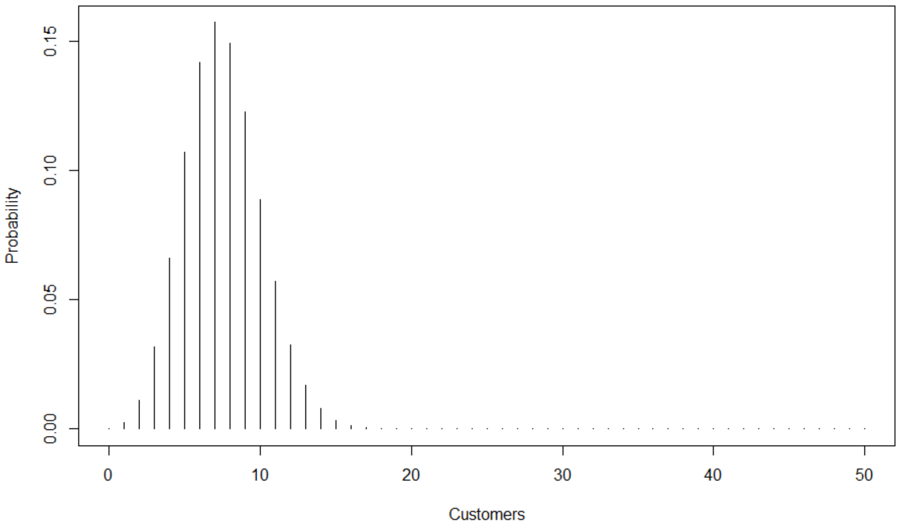

Figure 1b. Binomial probability distribution of 50 (n) customers where p = 0.15. Mean (μ) = 7.5, standard deviation (σ) = 2.52.

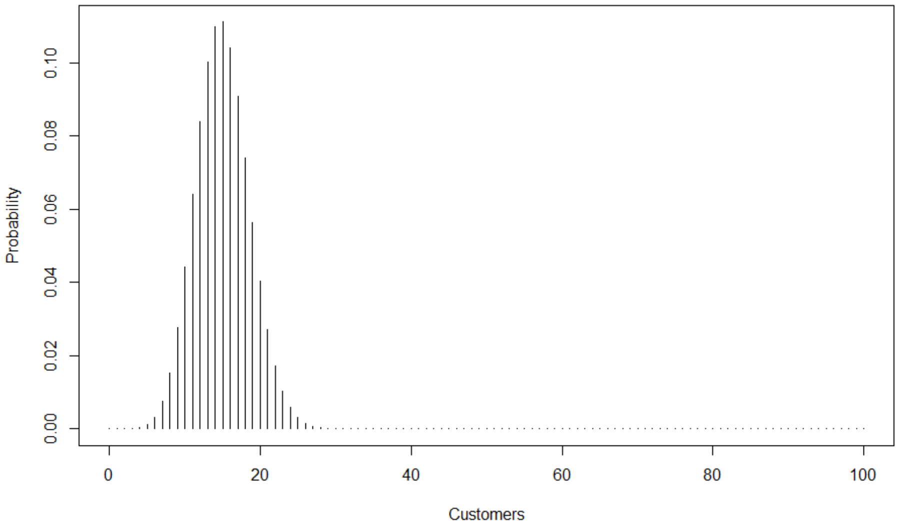

Figure 1c. Binomial probability distribution of 100 (n) customers where p = 0.15. Mean (μ) = 15, standard deviation (σ) = 3.57.

customers is 0.2759, and so on. The probability for each discrete number of customers to leave the store without making a purchase is best demonstrated graphically in Figure 1a.

It is noted that the highest probabilities are 0.3474 and 0.2759 which correspond with 1 and 2 customers, respectively. This is because these values of x are the closest to the mean (μ), which can be calculated as $\mu = np$, or the number of trials multiplied by the probability of success. Therefore, the mean (μ) of 10 trials (n) with a 0.15 probability of success (p) is 1.5 (between 1 and 2). The mean represents the outcome or value expected the most if the experiment were repeated multiple times. (It should be noted that the mean is not always a whole number and therefore may not be represented exactly in a binomial, or any discrete, probability distribution.)

The standard deviation represents the dispersion, or spread, of probability distribution. For a binomial probability distribution, this is calculated as the square root of the product of the mean (number of trials multiplied by the probability of success) multiplied by the probability of failure: $\sqrt{}(\mu \bullet (1 - p))$.

The functions that calculate the mean and the standard deviation both demonstrate a direct relationship with the number of trials (n). That is, assuming the probability (0.15) is kept the same, as the number of trials increases, so do the mean and standard deviation.

As the number of trials (customers in the experiment) increase, so does the mean, or expected number of customers that leave without making a purchase. Figures 1b and 1c demonstrate the binomial probability distributions for 50 and 100 trials, respectively. Again, the values of x with the largest probabilities (tallest bars) are closest or equal to the calculated mean, or expected value. Table 1 demonstrates the increasing value of the mean as the number of trials increases.

Table 1. Calculated mean and standard deviation of n trials.

Trials	n	10	50	100
Probability	p	0.15	0.15	0.15
Mean	μ	1.5	7.5	15
Standard Deviation	σ	1.13	2.52	3.57

Increasing the number of customers (trials, n) means that there are more possible outcomes of the experiment. Therefore, it is reasonable to assume that the spread of the values about the mean will increase. This is also demonstrated by Table 1 that also shows the calculated standard deviations increasing with the increasing number of trials.

Normal Approximation of the Binomial Distribution

The shape of the binomial probability distribution should also be noted. Figure 1 demonstrates the changing shape of the binomial probability distribution as the number of trials increases. The distribution with 10 trials (Figure 1a) is right skewed, which is expected with a binomial distribution with a probability that is less than 0.5. However, the distribution with 50 trials (Figure 1b) is only slightly right skewed, and the distribution with 100 trials (Figure 1c) is nearly symmetric. This similarity to a symmetric distribution is an important distinction in that, if other certain qualifications are met, the properties of a normal distribution will apply.

These properties include a symmetry about the mean; an equal value for the mean, median and mode; and the application of the Empirical Rule that states that approximately 68% of the observations will lie within one standard deviation of the mean, 95% within two standard deviations, 99.7% within three standard deviations, and that any observations outside of two standard deviations (5%) are considered unusual. These characteristics not only simplify the calculations for probabilities, but they also enable statistical inferences such as probabilities and proportions of the population, rather than just a probability of a sample.

A probability density function (PDF) is an equation used to compute probabilities of continuous random variables. A continuous variable differs from discrete one in that it has an infinite number of possible values that are not countable. The normal density function specifically calculates the probabilities for any value of the continuous random variable of a normal, or symmetric, distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

where μ = the mean and σ = the standard deviation. As this function is applied to continuous variable, the resulting plot is a line, rather than the bars of the binomial probability distribution. Specifically, the line is a bell-shaped curve that maintains some similar properties to the binomial distribution function. These include the height of the curve must be greater than or equal to zero for all possible values of the random variable, and that the total area under the curve is equal to one. Further the area under the normal curve for any interval of values of the random variable represents either the proportion of the population with the characteristic described by the interval of values or the probability that a randomly selected individual from the population will have the characteristic described by the interval of values.

Normal distribution curves were generated using the calculated means and standard deviations of the binomial probability distributions. The resulting Figure 2 shows the normal curve has a better "fit" of the binomial distribution with a greater number of trials. This demonstrates the rule of thumb that if the sample is large enough, the probability distribution will be approximately normal. The mathematical representation of large is $np(1 - p) \geq 10$. For this scenario, this calculates to 1.275 for 10 customers, 6.375 for 50 customers, and 12.75 for 100 customers. The latter is the only calculated value that exceeds the threshold of 10. Not coincidentally, the corresponding Figure 2c best "fits" the normal curve for the three trial scenarios presented. (In this scenario, it is calculated that a minimum of 79 trials is required to meet the threshold of 10).

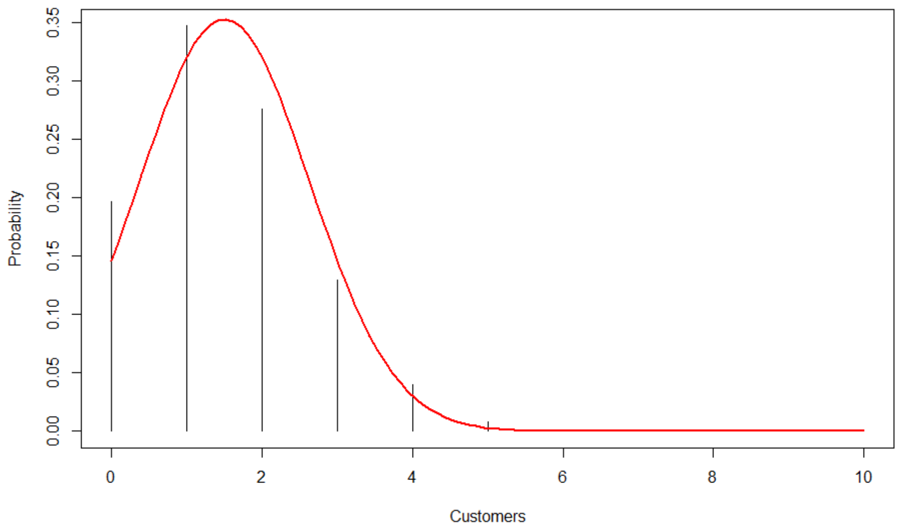

Figure 2a. Binomial (black) and Normal (red) probability distribution of 10 (*n*) customers where p = 0.15. Mean (μ) = 1.5, standard deviation (σ) = 1.13.

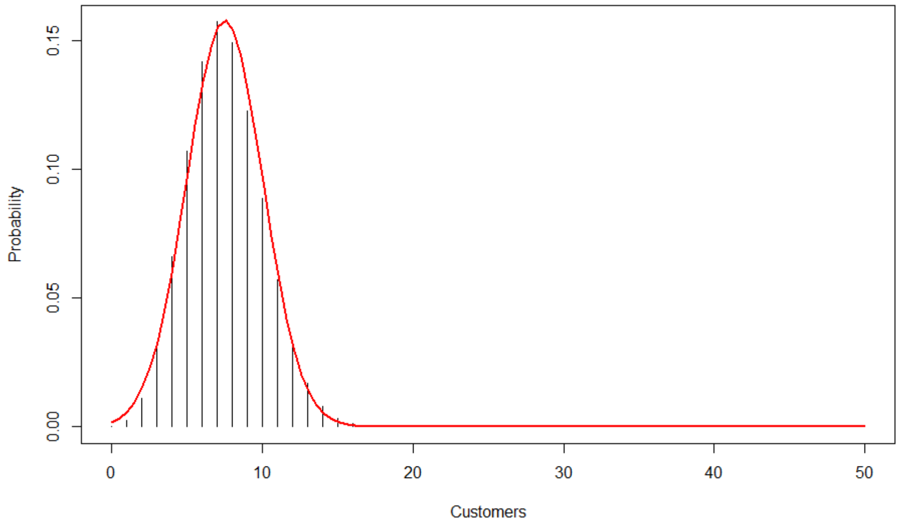

Figure 2b. Binomial (black) and Normal (red) probability distribution of 50 (*n*) customers where p = 0.15. Mean (μ) = 7.5, standard deviation (σ) = 2.52.

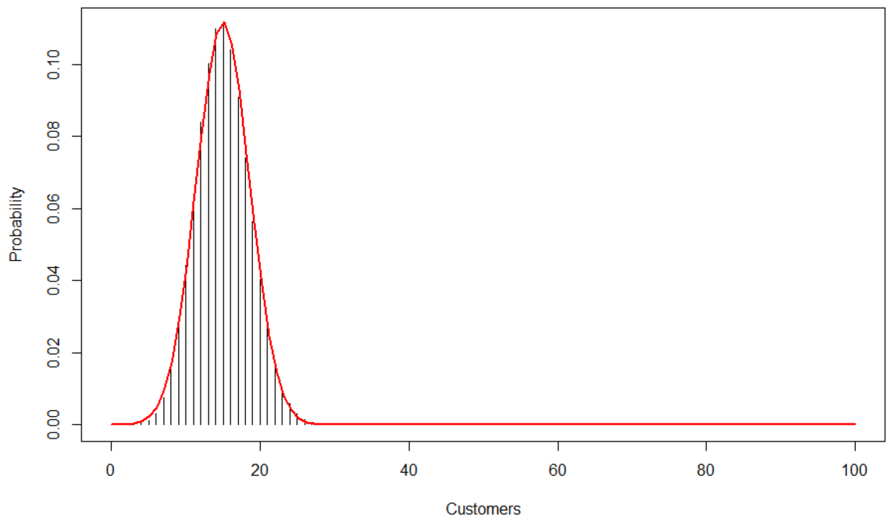

Figure 2c. Binomial (black) and Normal (red) probability of 100 (n) customers where p = 0.15. Mean (μ) = 15, standard deviation (σ) = 3.57.

To further demonstrate this point, the probabilities of each of the discrete values of the random variable based on a normal distribution can be calculated and compared to those of the binomial distributions. A graphical representation of those calculations for 10, 50 and 100 customers can be seen in Figure 3. It should be noted that the calculated normal approximations are closer to the actual binomial probability calculations as the number of trials increases. Table 2 provides the calculations for each trial size for a selection of values of the random variable:

Table 2a. Selections of binomial and normal probability calculations of 10 (*n*) customers where *p* = 0.15. Mean (μ) = 1.5, standard deviation (σ) = 1.13.

x	Binomial	Normal	Difference
0	0.1969	0.1462	0.0507
1	0.3474	0.3203	0.0271
2	0.2759	0.3203	-0.0444
3	0.1298	0.1462	-0.0164

Table 2b. Selections of binomial and normal probability calculations of 50 (*n*) customers where *p* = 0.15. Mean (μ) = 7.5, standard deviation (σ) = 2.52.

x	Binomial	Normal	Difference
5	0.1072	0.0968	0.0105
6	0.1419	0.1324	0.0095
7	0.1575	0.1549	0.0025
8	0.1493	0.1549	-0.0056
9	0.1230	0.1324	-0.0095
10	0.0890	0.0968	-0.0078

Table 2c. Selections of binomial and normal probability calculations of 100 (*n*) customers where *p* = 0.15. Mean (μ) = 15, standard deviation (σ) = 3.57.

x	Binomial	Normal	Difference
13	0.1001	0.0955	0.0046
14	0.1098	0.1074	0.0024
15	0.1111	0.1117	-0.0006
16	0.1041	0.1074	-0.0033
17	0.0908	0.0955	-0.0047

As the number of customers increases, the difference between the two calculations decreases. This further concludes that a binomial distribution will better resemble a normal distribution as the number of trials increases, but only when the number of trials is sufficiently large. And as stated above, when these conditions are met, the normal probability density function applies, which implies that the area under the normal curve for any interval of values of the random variable represents either the proportion of the population with the characteristic described by the interval of values or the probability that a randomly selected individual from the population will have the characteristic described by the interval of values.

The application of the properties of a normal distribution – such as the Empirical Rule – enables the ability to make confident predictions of a population of customers. This is a powerful tool for the owner of the store. For example, they likely know the average amount of money that a customer spends; knowing the proportion of customers that will make a purchase will enable them to calculate the number of customers they need to attract to the store to meet certain sales goals, sell a specific quantity of inventory that is perishable, and so on.

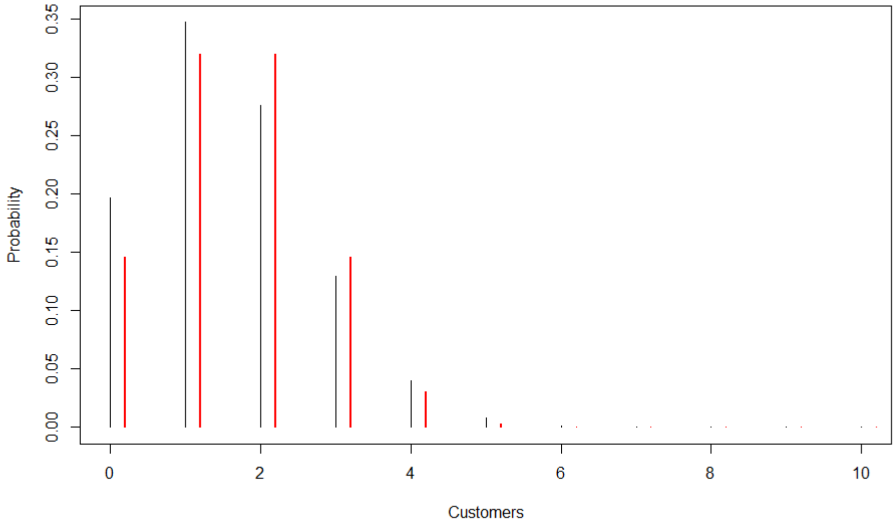

Figure 3a. Binomial probability distribution (black) and Normal approximation (red) of 10 (*n*) customers where p = 0.15. Mean (*μ*) = 1.5, standard deviation (*σ*) = 1.13.

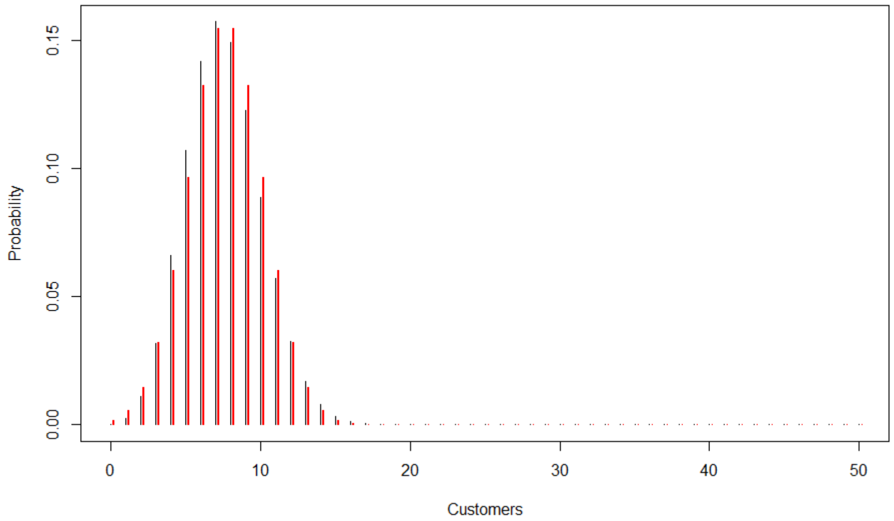

Figure 3b. Binomial probability distribution (black) and Normal approximation (red) of 50 (n) customers where p = 0.15. Mean (μ) = 7.5, standard deviation (σ) = 2.52.

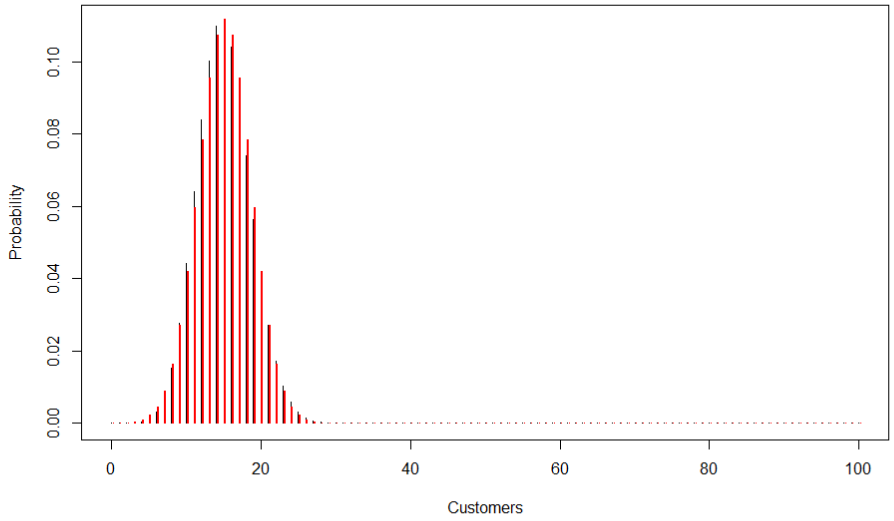

Figure 3c. Binomial probability distribution (black) and Normal approximation (red) of 100 (n) customers where p = 0.15. Mean (μ) = 15, standard deviation (σ) = 3.57.

Sampling Distributions of a Binomial Experiment

The previous sections discussed the binomial probability distribution and the application of the normal approximation and its properties. It should be noted that the calculations and graphical representations of these concepts are theoretical. In the scenario above, the store owner states that 15% of the customers that come into their store do not make a purchase and therefore when 100 customers were sampled, the mean was calculated as 15 customers. However, this does not necessarily mean that, in reality, every time 100 customers are sampled, exactly 15 will leave the store without making a purchase. That number will vary from sample to sample; it could be 10, or 14, or 16, or 20, and so on. It can therefore be said that the statistics of the repetition of a probability distribution are random variables themselves, and therefore can be associated with a probability distribution as well.

In the case of the mean, the sampling distribution of the sample mean $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$ computed from a sample size (trials) *n* from a population with a mean μ and standard deviation σ. When the population is normally distributed, the mean of the sample distribution is the same as mean of the population. This is reasonable as the mean is the outcome that is expected the most. The standard deviation of the sample mean is equal to the standard deviation of the population divided by the square root of the sample size. This inverse relationship in which the standard deviation of the mean decreases as the sample size increases is reasonable because the larger a sample is, extreme values within the sample are more likely to be offset by another and therefore the mean of the sample will tend toward the mean of the population. The standard deviation of the sample mean is also referred to as the standard error. Further, an observed sample mean that is within one standard deviation of the population mean can be interpreted as a sample that is typical, or statistically representative of the population.

In the scenario of the store, the distribution was described as right skewed as it is a binomial distribution with a probability that is less than 0.5. As this is not a normal distribution, a sufficient sample size is required for the distribution of the sample mean to be considered approximately normal. This is due to the Central Limit Theorem which states that regardless of the shape of the underlying population, the sampling distribution of the sample mean becomes approximately normal as the sample size increases; as a rule of thumb, skewed distributions require a sample size of at least 30.

Like the sampling distribution of the mean, the sampling distribution of the sample proportion – the proportion of individuals in a sample who have a specified characteristic – can also be associated with a probability distribution. In this binomial experiment, the sample proportion ($\hat{p}$) is simply the number of individuals who leave the store without making a purchase (x) divided by the number of customers surveyed or sampled (trials, *n*). The mean proportion of the sampling distribution ($\mu_{\hat{p}}$) is therefore the mean of the population (μ, or *np*) divided by the number of customers surveyed or sampled (trials, *n*); $\mu_{\hat{p}} = np / n = p$; $\mu_{\hat{p}} = p$. And the standard deviation of the proportion of the sampling distribution ($\sigma_{\hat{p}}$) is therefore the standard deviation of the population (σ) divided by the number of customers surveyed or sampled (trials, *n*); $\sigma_{\hat{p}} = \surd(\mu \bullet (1 - p)) / n = \surd(((np \bullet (1 - p)) / n^2) = \surd((p \bullet (1 - p)) / n)$.

Just as the mean of every sample is not expected to always be equal to the mean of the population, the sample proportion also varies from sample to sample. In the scenario of the store where the owner, the expected proportion of customers who will leave the store without making a purchase is 15%, or 0.15. If he or she were to survey equal size groups of customers over an over, the expectation would be that the

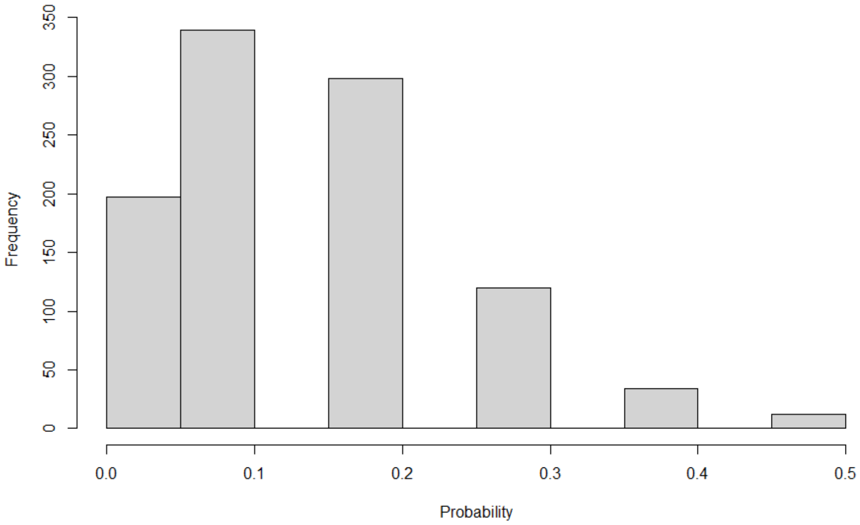

Figure 4a. Sample proportion distribution of 1,000 samples where the number of trials (*n*) = 10 and the population's probability (*p*) = 0.15 and theoretical proportional mean (*μ/n*) = 0.15 and standard deviation (*σ/n)* = 0.1129. The resulting sampling distribution's mean ($\mu_{\hat{p}}$) = 0.1491 and standard deviation ($\sigma_{\hat{p}}$) = 0.1110.

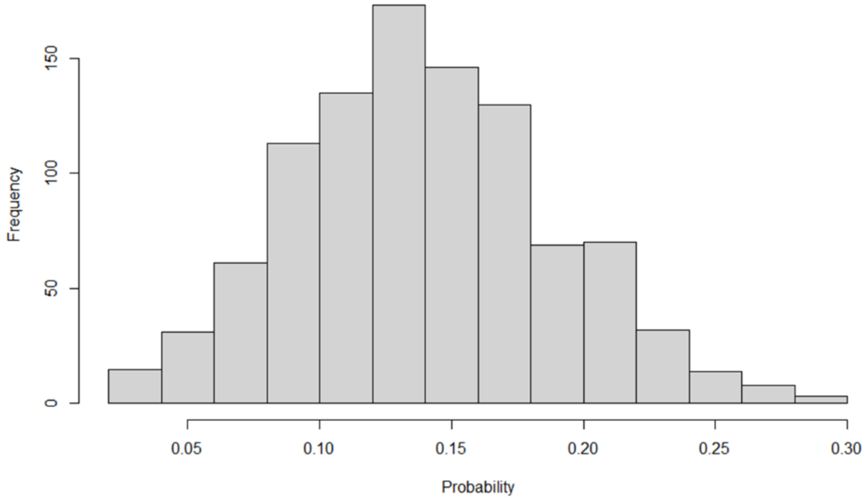

Figure 4b. Sample proportion distribution of 1,000 samples where the number of trials (*n*) = 50 and the population's probability (*p*) = 0.15 and theoretical proportional mean (*μ/n*) = 0.15 and standard deviation (*σ/n*) = 0.0505. The resulting sampling distribution's mean ($\mu_{\hat{p}}$) = 0.1494 and standard deviation ($\sigma_{\hat{p}}$) = 0.0495.

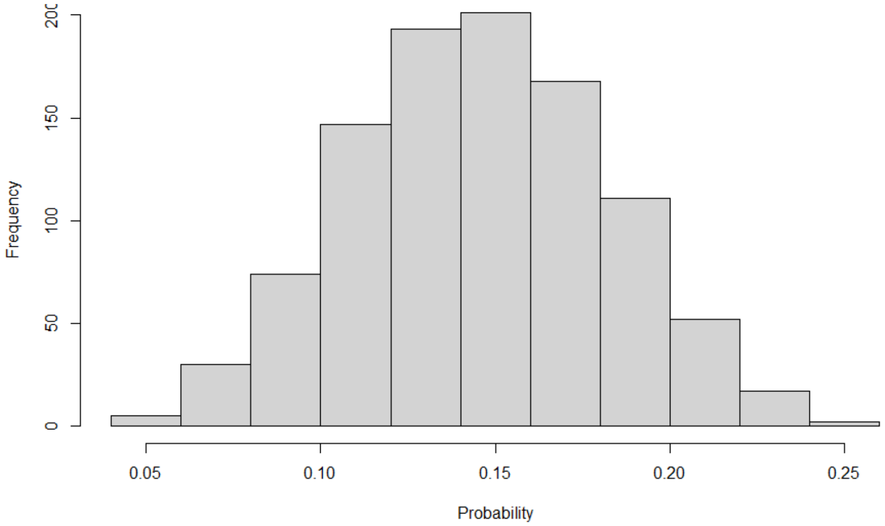

Figure 4c. Sample proportion distribution of 1,000 samples where the number of trials (*n*) = 100 and the population's probability (*p*) = 0.15 and theoretical proportional mean (*μ/n*) = 0.15 and standard deviation (*σ/n*) = 0.0357. The resulting sampling distribution's mean ($\mu_{\hat{p}}$) = 0.1504 and standard deviation ($\sigma_{\hat{p}}$) = 0.0366.

mean proportion would be 15% with some degree of error. As that is not a very practical survey to conduct, this scenario is simulated 1,000 times with sample sizes of 10, 50 and 100 (Figure 4). The mean and standard deviation of these simulations are summarized in Table 3 along with their theoretical proportional mean and standard deviation based on the population.

In each scenario, regardless of the sample size, the mean sample proportion is approximately equal to the theoretical mean of 0.15. However, the standard deviation, or variation of the observations about the mean, decreases as the sample size increases. It can therefore be concluded that a larger sample size results in a proportional mean that is more likely to resemble that of the population.

Table 3. Calculated theoretical (population) and sample simulated proportional mean and standard deviation of n trials.

Trials	n	10	50	100
Probability	p	0.15	0.15	0.15
Proportional Mean (population)	$\mu / n = p$	0.1500	0.1500	0.1500
Proportional Mean (sample simulation)	$\hat{\mu}_{\hat{p}}$	0.1491	0.1494	0.1504
Proportional Standard Deviation (population)	σ / n	0.1129	0.0505	0.0357
Proportional Standard Deviation (sample simulation)	$\sigma_{\hat{p}}$	0.1110	0.0495	0.0366

It should also be noted that the distribution of sample proportions is skewed right when the sample size is 10, slightly skewed right when it is 50, and approximately normal when it is 100. This once again demonstrates the rule of thumb that if the sample is large enough, the probability distribution will be approximately normal. The mathematical representation of large is $np(1 – p) \geq 10$. In this scenario, it is calculated that a minimum of 79 trials is required to meet the threshold of 10.

Therefore, if the store owner surveys or samples at least 79 customers, they can expect a normal distribution of the proportion of customers that leave the store with or without making a purchase. Invoking the Empirical Rule, they can expect that 11.4-18.6% of their customers to leave the store without making a purchase in 68% of their customer samples (15% ± 3.57%; proportional mean ± one standard deviation, aka standard error). They can also expect 95% of their samples to include 7.9-22.1% of customers to leave without making a purchase. Any sample below 7.9% or above 22.1% would be considered unusual as it represents only 5% of expected samples. These statistically significant conclusions about their business enable them to make insights that can improve their sales if they continue taking samples periodically.

For example, if they find that less than 7.9% of customers (more than 2 standard deviations from the mean) leave without making a purchase, this should be considered unusual, but with a positive connotation. A period with an improved customer conversion rate should be explored to determine if it can be repeated. There may be a sales representative that is particularly successful, and their methods could be used to train other employees. Or there may have been a product that sold well or a promotion that worked particularly well.

Conversely, if the owner finds a sample where less than 22.1% of customers leave without making a purchase, they may want to investigate the cause of the poor sales conversion rate. Perhaps their pricing is out of line, or maybe there is a substandard sales representative to blame, or maybe even a lapse in their store security that is inviting shoplifters. And if the owner consistently finds samples that are unusual, they may have to recalculate the mean, and conclude that their customer conversion rate is now less than they previously determined. Then they must find ways to improve their customer conversion rate.

References

Sullivan III, Michael. Statistics: Informed Decisions Using Data; 6th Edition, ISBN-13: 9780135780121.