

Sampling Distributions and Inference

What Do We Know and How Do We Know It?

(Student Pages)

In statistical inference we are generally interested in identifying a population parameter from among all possible values. The ability to describe one or more characteristics determined by the parameter is vital. Typically, inference compares an observed sample with these known characteristics of possible populations to judge the likelihood of that sample's occurring.

In this setting, then, it is important to know the candidate populations and to know characteristics distinguishing one population from another. The following sequence of activities is restricted to populations consisting of only two types of elements, "success" and "failure". The goal of the inference is to estimate the proportion, p , of the population that are successes.

Problem Setting:

Suppose that an automobile manufacturer wishes to have 60% of its vehicles be free of warranty-related repairs during the first three years after the car is sold. If too few cars go three years without repairs, the company will be branded as selling "lemons." If too many cars need no repairs, opportunities for income through the service department of dealerships diminishes. One dealership decides to check this performance by examining the service records of 20 randomly selected cars sold three years ago. They find that exactly 5 of these 20 cars have needed no repairs covered under warranty. Is this evidence that the 60% goal for all car sales has not been met? What is a reasonable estimate of the actual percentage of all cars sold that need no covered repairs during the three-year period?

The count of cars in the sample that actually needed no repairs during the three-year period is a sample statistic. We will refer to it as the "number of successes" and denote it by x . Thus, for this scenario, $x = 5$. The dealer needs a statistical estimate of the proportion, p , of successes among *all* cars sold. This proportion is a property of the population.

We will examine this scenario through a sequence of simulations, building from physical to numerical to calculator or computer based. The primary goal is to see what "should" happen for known populations (known values of p). The resulting sampling distributions will then permit estimation of the actual (but unknown) proportion p . A key idea is that each population gives rise to a predictable distribution of the sample statistic, x .

In simple terms, the dealer wants to know:

- whether the observed result should be surprising if the 60% rate is accurate,
- just how surprising the observed result is if the 60% rate is correct,
- for what actual rates the observed result would *not* be surprising.

Keep in mind, too, that although the observed number of "good" cars seems surprisingly low, it would also be of interest to the dealer if that number were surprisingly high.

Physical Simulation:

Understand that the situation you are simulating is that of selecting 20 cars, independently, from a population in which the proportion of “good” cars (successes) is p and counting the number, x , in your sample that are good.

Mathematically, then, each run must involve selecting 20 “somethings” in such a way that no selection influences other selections, and each selection must have probability p of being a success. After all 20 are selected, the desired “output” value is the count of the number of successes in the sample.

One physical mechanism for implementing this simulation is a bag of candies of various colors. Red can represent success with all other colors representing failure. The mix of colors in your bag will define what the actual population proportion, p , is for your simulation. Selecting one candy represents checking the repair record of one car. In order to keep p constant from one “car” to the next you will need to replace the candy after its color is checked.

You will be given a bag containing a mix of candies, some red and some not. Each of your classmates has a different population with a different value for p . Each bag is labelled with its value of p .

Select a sample of size 20 (with replacement) from your bag.

Record the number, x , of candies in the sample that are red.

Repeat the sampling 9 more times (so you have 10 values of x).

Make a dot plot of your observed x values.

Interpret your plot. Is a sample of size 20 with 5 reds likely to have come from your population?

Numerical Simulation:

Since you are simulating the process of sampling from a population, the physical simulation is conceptually very easy. You really *are* doing the sampling; only the items in the population have been changed for ease of handling. However, the method has a couple of major drawbacks.

First, physical simulation is slow. It takes time to complete enough runs for the distribution of x to be informative. Ten runs probably are not enough to get a sense of what all possibilities might be.

Even worse, though, is the lack of flexibility in selecting values of p to use in the simulation. Each new value of p that you wish to examine requires a new bag, carefully counted to contain the correct proportion. Numerical simulation addresses these problems.

A new mechanism for your simulation uses random integers from 001 through 100 to represent cars, with integers 001 through $100p$ being successes and integers $100p+1$ through 100 being failures. For example, with $p = 0.6$, then [001, 060] represents successes and [061, 100] represents failures. Then select 20 integers from [001, 100]. Count how many of the 20 are less than or equal to $100p$ to get the simulated x value.

Use a random number table to execute this simulation mechanism for your assigned value of p 10 times.

Make a dot plot of your x values and interpret it. Is a sample of size 20 with 5 successes likely to have come from your population?

This simulation mechanism permits much greater flexibility in choosing the population (that is, the value of p) on which to base the simulated the sampling. Still, only values of p for which $100p$ is an integer may be used. In addition, using a random number table remains pretty slow.

Calculator Simulation:

The mathematical description, and the mechanism for realizing a valid simulation, given in the preceding section are not limited to random number tables. Calculators and computers usually have built-in random number generators. In fact, many calculators allow you to select all 20 values, store them to a list, and count the successes all at once. For example, the TI-83 command

$$\text{randInt}(1,100,20) \rightarrow \text{L1}:\text{sum}(\text{L1} \leq 100*0.6)$$

effects one run of one sample of size 20 using $p = 0.6$, storing the sample in list L1 and reporting the number of successes on the home screen.

Use appropriate calculator (or computer) commands to carry out 100 runs with your assigned value of p .

Make a dot plot or histogram of x .

Interpret your plot. Is a sample of size 20 with 5 successes likely to have come from your population? About how likely is it that a sample of size 20 with 5 or fewer successes would be drawn from your population? About how likely are samples with 16 or more successes?

For your population, what values of x seem reasonably likely to occur for samples of size 20? How did you decide?

Calculator Simulation II:

You have simulated sampling 20 items from a population in a variety of ways. Each simulation, though, has retained the elements of the original situation. Regardless of the method used you could identify each individual sample and its 20 items, identify the successes, and count them.

Now that you understand how that simulation process works, you can turn the details over to technology in order to gain the speed necessary to permit looking at simulations involving many more runs. “Output values” (x) collected according to the mathematical conditions described in the Numerical Simulations section are said to be binomially distributed. Many calculators and computers accept commands that generate binomially distributed values directly. For example, the TI-83 command:

$$\text{randBin}(20,0.6,300) \rightarrow \text{L1}$$

generates 300 random x values corresponding to samples of size 20 from a population in which $p = 0.6$ and stores the x s in list L1. The tradeoff for greater speed and more runs is loss of detail within each run. Now, though you can tell how many successes appear in each sample of 20, you can no longer see each set of 20 or know which items were the successes.

Use your calculator or computer to generate at least 300 runs simulating the sampling process from your assigned value of p . (More runs are better, but everyone in the class should use the same number for easier comparisons.)

Make a histogram of the resulting x values.

Interpret your plot. Is a sample of size 20 with 5 successes likely to have come from your population? About how likely is it that a sample of size 20 with 5 or fewer successes would be drawn from your population? About how likely are samples with 16 or more successes?

For your population, what values of x seem reasonably likely to occur for samples of size 20? How did you decide?

The Logic of Inference:

The initial stated purpose of inference is to identify an unknown population by examining a sample from it. You and your classmates have examined many different populations (different values of p). Thus, you have different simulated histograms of x values representing samples from your populations. It is useful to compare the results of your entire class to see how different populations lead to different sampling distributions of x . Remember, the initial idea was that these sampling distributions might be able to help identify your populations if only sample information were known. Knowing p determines the distribution of x s; knowing x tells you what population(s) it might have come from.

On a separate sheet of paper, or on your calculator screen, and using a common scale assigned by your teacher, make a histogram of your simulated x values.

Add a vertical line on your histogram at $x = 5$.

Place your completed graph on a table or chalkboard tray as instructed by your teacher.

When all graphs are in place, walk along the entire display to see how “ $x = 5$ ” is related to each sampling distribution (thus, to each population).

Summarize the similarities and differences among the sampling distributions based on the various populations. What role does p appear to play?

Using the agreed-upon decision criterion, determine whether the observed $x = 5$ is likely to have come from your bag. This kind of decision is known formally as a two-sided hypothesis test, a topic you will study soon.

List all the populations (values of p) from your class for which the observation $x = 5$ was judged to be likely. This set corresponds to what is known formally as a confidence interval for p based on your sample value.