

Using Sample Data to Draw Conclusions about Population Means

Your instructor has a handout of 100 little boxes. This set of 100 boxes is the population that we will pretend to be interested in today. Our goal is to learn what we can about the mean/average area of *all* the boxes in the population by selecting and observing only a simple random sample of 10 boxes. In a real problem, our population of interest might be all geese in North America, all trees in Alaska, all farmers in Iowa, etc. We will use the population of boxes to illustrate important statistical concepts because this population is much easier to work with (at least in our classroom) than many more interesting populations.

1. Use the random digit sheet to draw a simple random sample of 10 boxes and find the area (the number of small cubes) of each box in the sample. (Note: To use the random digit table, all boxes in the population should be labeled with the same number of digits. So box 1 can be labeled “box 01”, box 9 can be labeled “box 09”, box 10 is just “box 10”, box 99 is “box 99”, and so on. Box 100 can be labeled “box 00”. Then all the boxes in the population have labels consisting of two digits. Pick any line in the random digit table and read off two numbers at a time. By reading off ten pairs of numbers, you will randomly select ten boxes from the population with the corresponding labels.)
2. Compute \bar{Y} the average area of the sampled boxes. We use the statistic \bar{Y} to estimate μ = the mean/average area of all 100 boxes in the population.
3. One of the interesting facts is that we can use information in the sample to obtain information on how far our statistic (\bar{Y}) is likely to be from the population parameter (μ). We begin by estimating the population standard deviation (σ) of the areas of all 100 boxes in the population by the sample standard deviation (s). Compute s , the standard deviation of the areas for the 10 boxes in your sample. ¹
4. Statistical theory dictates that the standard deviation of a sample average \bar{Y} is σ/\sqrt{n} , where σ denotes the population standard deviation and n denotes the number of observations drawn randomly from the population of interest. The **estimated** standard deviation of a statistic is called its **standard error**. Give the formula for the standard error of a sample average \bar{Y} . ²
5. Compute the standard error for \bar{Y} in this example. This value is an estimate of the “typical” size of the error made when estimating μ by \bar{Y} .
6. The quantity

$$t = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

is called a t -ratio. Note that a t -ratio will vary from sample to sample because it depends on statistics (\bar{Y} and s) that change from sample to sample. Statistical theory dictates that under fairly general conditions discussed in Chapter 3, the distribution of the t -ratio will be approximated by a Student’s t -distribution with $n - 1$ degrees of freedom. Table A.2 in the back of your text gives information about the proportion of observations from a Student’s t -distribution that will be less than various values. (Usually we will just say t -distribution rather than Student’s t -distribution.)

¹Since not all 100 boxes in the population have the same area, σ represents a population characteristic/parameter assessing how much spread/variability there is among the box areas in the population of 100 boxes, i.e., how much individual box areas in the population of 100 boxes deviate from the population mean/average μ (some boxes will have more or less area compared to μ). Since we don’t know σ (much like we don’t know μ), we use the amount of spread s in the data values (10 box areas) to estimate the amount of spread σ in the population (100 box areas).

²Again because random samples can vary, the values of the statistic \bar{Y} (i.e., sample mean box area) can vary depending on which observations (boxes) are drawn into the sample from the population (i.e., each student will have a single \bar{Y} value from her/his sample, these may vary). The potential spread/variability in possible \bar{Y} values across *all possible samples* is σ/\sqrt{n} which depends on how much variability (σ) there is in individual observations in the population (i.e., 100 box areas) as well as the sample size (n). Generally speaking, we would like the spread (σ/\sqrt{n}) in all possible \bar{Y} values that could occur in sampling to be *small*; a small σ/\sqrt{n} entails that, when we take a single sample, we expect our sample mean \bar{Y} from our single sample to be close to the population mean μ . Since we don’t know the spread σ in individual values in the population, we don’t know the spread σ/\sqrt{n} in all possible \bar{Y} values, so we estimate this with the standard error given by s/\sqrt{n} (namely, replace σ with an estimate s).

- (a) What proportion of observations from a t -distribution with 1 degree of freedom will be less than 6.314?
- (b) What proportion of observations from a t -distribution with 20 degrees of freedom will be less than 1.064?
- (c) What proportion of observations from a t -distribution with 20 degrees of freedom will be less than -1.725?
- (d) What proportion of observations from a t -distribution with 20 degrees of freedom will be larger than 1.725?

7. Based on our observed sample and the information in the previous problem, do you think the mean/average area of all 100 boxes in the population is 3? That is, does $\mu = 3$ seem plausible?

8. Formally we can test the null hypothesis $H_0 : \mu = 3$ against the alternative $H_a : \mu > 3$.

- (a) Our **test statistic** is the t -ratio that we get by plugging the value of μ according to the null hypothesis in for μ in the t -ratio equation. Compute the test statistic.
- (b) Our **p -value** is the probability - computed assuming the null hypothesis to be true - of observing a t -ratio that provides at least as much evidence in favor of the alternative hypothesis as the test statistic that we computed (i.e., the probability of a test statistic more extreme (larger in this case) than the t -ratio we observed). Compute the p -value for this problem.

(c) Is there evidence that the mean/average area of all 100 boxes in the population is greater than 3? (See information about interpreting p -values on page 47.)

(d) We have just carried out what is known as a **one-sided hypothesis test**. A **two-sided hypothesis test** is concerned about alternatives on both sides of the null hypothesis ($H_a : \mu \neq 3$). The p -value for a two-sided test is just twice the p -value of a one-sided test. Write the p -value for the two-sided test below.

9. Rather than checking many individual values for μ to see if they seem plausible, we can compute a confidence interval for μ . A confidence interval directly provides a range of plausible values for μ . A $100(1 - \alpha)\%$ confidence interval for a population mean μ is given by

$$\bar{Y} \pm t_{n-1}^{(1-\alpha/2)} \cdot \frac{s}{\sqrt{n}},$$

where $t_{n-1}^{(1-\alpha/2)}$ is the t -ratio larger than $100(1 - \alpha/2)\%$ of t -ratios from a t -distribution with $n - 1$ degrees of freedom. Compute a 95% confidence interval for μ .

10. The percentage attached to a confidence interval indicates the percentage of samples that will give a confidence interval that contains the true values of μ . In other words, when your class drew a random sample of 10 boxes from the population of 100 boxes, there was 95% chance that the sample selected would lead to a confidence interval (as in question 10) that contains the population mean $\mu =$ the mean/average area of all boxes in the population. Your instructor knows μ because she computed it by examining all boxes in the population. Was your sample one of the 95% that produces a good confidence interval, or did you get unlucky?