LA's BEST – Sampling activity

Juan Pablo Lewinger

6/16/2022

Goal

The goal of this lab is to reinforce the key concepts of **population** and **sample** through a hands-on activity. You will sample from a large population of glass beads and record their attributes. You will then enter the data into R and generate plots and descriptive statistics to summarize and visualize the data. You will then use this dataset to illustrate the estimation of means, proportions, and measures of spread.

Sample

You will be presented with a container with over 3,000 glass beads (the kind used for making jewelry) of different sizes, shapes and colors. These will represent the population you want to learn about. Because it would be extremely time consuming and tedious to record the characteristics of every single individual bead in the container, we will instead take a small sample of beads to make inferences about the entire population. This is analogous to what we do in Biostatistics when we want to learn about a particular population of individuals (e.g. prostate cancer patients in LA county, California children aged 5-13, etc): we take a sample rather than examining every individual in the population (which would be extremely expensive and time consuming). As we know, we can learn about a large population using a sample. You will learn about how large a sample is needed for estimating certain population parameters in a subsequent class.

Working in groups you will take a sample of ~30 beads and record their shape, color, weight and size. It's critical that the sampling is done, as much as possible, to resemble truly random sampling; otherwise biases can crop in that will invalidate the inferences we make about the population of beads. So, make sure the jar is well mixed and use the scoop provided to take the sample (using your hand could lead to biases as it could make it physically easier to grab, or unconsciously prefer, larger or smaller beads, beads of certain shapes or certain colors).

Shapes are: round (spherical shape), rondelle (spherical with flatenned poles), barrel nugget, flat, teardrop, heishi (disk).

Colors are: celeste (light blue), emerald (green), and aquamarine (blue)

Weight: You'll be provided with a precision scale and a small tray to weigh the beads.

Size: You will be provided with a caliper (a precision tool to measure length) to measure their length. You should measure the longest length possible of the bead (in mathematics this is referred to as the diameter, even when it's the object is not circular or spherical)

Create a data file

Keep a paper record of the collected data to have something to refer to in case there are errors during data entry. Create a file with the data using a spreadsheet such as Excel or a text editor like textEdit or Notepad. Use one row/line to represent each bead in your sample and one column to represent each variable (color, shape, size, and weight). Use the first row for the variable names. If using Excel or other spreadsheet, save the data as a comma separated file (.csv). If using a text editor make sure you leave white space between entries and save the file as a regular text file.

Enter the data into R

For a .csv file you can use:

```
setwd('the path to the directory where your file resides')
beads <- read.csv("beads.csv")
```

For a .txt file you can use:

```
setwd('the path to the folder where your file resides')
beads <- read.table("beads.txt, header=TRUE")
```

Visualize the data

1. Create appropriate univariate graphs (histograms, boxplots, barplots) to visualize each of the variables in your dataset

Comment on the shape of the distributions in your sample. How would you expect them to compare to the distribution in the entire population of beads?

tip: If you want multiple graphs on the same plot you can split the plotting area using the option:

```
par(mfrow=c(3,2), mar(c(2,2,2,2)))
```

This tells R to subdivide the plotting region into 6 panels organized into 3 rows and 2 columns, using margins of size 2 (measured in lines) within each panel (small margin is chosen so that plotting areas are not too small). Each new plot you create with the plotting function 'plot()' will be plotted in the next available panel. After all panels are used up, any new plot will be plotted over the old ones.

Summarize the data

1. Compute the mean, standard deviation (sd) and interquantile range (IQR) for each of the variables in your data for which they are appropriate
2. Generate tables for each of the variables in your data for which they are appropriate

Estimate population parameters

3. Estimate the mean weight and mean diameter in the entire 'population' of beads (~ 3,000 beads in the original container). Compute 95% confidence intervals.
4. Estimate the proportion of beads of each color in the entire 'population' of beads. Compute 95% confidence intervals.