First Project - Basic Version

Contents

Background

We hope to show that when we estimate a population's mean and standard deviation by using a sample's mean and standard deviation, the estimates are typically close to - but not exactly equal to - the actual population's parameters.

In research, we often encounter data that contains randomness. For example, when testing a medicine, some patients survive and others die whether or not the medicine is administered. This randomness can be simulated with a random-number generator.

In statistics, we will think of the random-number generator as the "population". This is because we often focus on a specific random-number generating process: a simple random sample (with replacement) from a population.

Research often involves describing a population based on a relatively small sample. For example, a scientist may want to characterize the weight of a new species of tree frog. That scientist cannot capture all of those tree frogs. Instead, that scientist tries to collect a simple random sample of those tree frogs. Then, based on the sample, the scientist infers knowledge about the population. A basic inference is to report a "best guess" of the population's mean and standard deviation. That best guess is the sample's mean and standard deviation.

We will simulate this process. However, unlike the scientist, we know the population exactly. This will allow us to explore how the best guesses are probably close, but not exact. An interested student might be see (and report) what happens when they repeat this process or what happens when the sample size gets larger (these additions would make the paper better, but they are not requirements).

Outline

- Choose a random-number generator (population) with identifiable mean and standard deviation.
- Determine population mean.
- Determine population standard deviation.
- Sample from the population ($n \geq 20$) with replacement to get independent identically distributed random numbers.
- Determine sample mean.
- Determine sample standard deviation.
- Compare the sample statistics to the population parameters using relative difference.
- Determine the z-score of the sample mean.

You will write a paper detailing your process. You will describe which population you chose (and why, if there is an interesting reason). You will report the required values in a clear manner. Visualizations (histograms, density curves, spinners, etc.) are helpful in clearly displaying the population and sample (but I'd also like the raw measurements). A frequency distribution (as a table) could also be useful. You can also just use English to report the results.

Your intended audience is someone who knows statistics, but has not read the project description. So, it is up to you to make it into a coherent report. It might help to simulate something you care about. You probably should describe why you would do this simulation. You are simulating the process of using sample mean and sample standard deviation as best-guess estimators for the population parameters. You are showing that these best-guess estimators tend to be near, but not exactly equal to, the actual population parameters.

To make it more interesting, you can repeat the simulation multiple times. You could also repeat the simulation with larger sample sizes. By using repetition and various sample sizes, you should see that larger samples tend to give better estimates. You could also repeat the simulation with a different population: one with a larger standard deviation will show the sample means tends to have a larger relative difference, but the z-scores should still mostly be between -2 and 2.

Step 1: pick a population

You need to pick a random number generator whose population mean and population standard deviation are known.

Population (distribution)	Mean (μ)	Standard deviation (σ)	Example	Generator in R	Generator in spreadsheet
Dice (discrete uniform from 1 to N)	$\mu = \frac{N+1}{2}$	$\sigma = \sqrt{\frac{N^2-1}{12}}$	$N = 6$ $\mu = \frac{6+1}{2} = 3.5$ $\sigma = \sqrt{\frac{6^2-1}{12}} = 1.708$	`sample(1:6,size=20,replace=TRUE)`	
Discrete uniform from A to B	$\mu = \frac{A+B}{2}$	$\sigma = \sqrt{\frac{(B-A+1)^2-1}{12}}$	$A = 30$ $B = 50$ $\mu = \frac{30+50}{2} = 40$ $\sigma = \sqrt{\frac{(50-30+1)^2-1}{12}} = 6.055$	`sample(30:50,size=20,replace=TRUE)`	
Bernoulli distribution (0s and 1s with weighted coin, p chance of 1)	$\mu = p$	$\sigma = \sqrt{p(1-p)}$	$p = 0.6$ $\mu = 0.6$ $\sigma = \sqrt{0.6(1-0.6)} = 0.4898979$	`sample(0:1,20,prob=c(0.4,0.6),replace=TRUE)`	

Population (distribution)	Mean (μ)	Standard deviation (σ)	Example	Generator in R	Generator in spreadsheet
Normal distribution	μ	σ	$\mu = 40$ $\sigma = 3$	`rnorm(n=20,mean=40,sd=3)`	`=NORM.INV(RAND(),0,1)`
Geometric distribution with parameter p	$\mu = \frac{1-p}{p}$	$\sigma = \sqrt{\frac{1-p}{p^2}}$	$p = 0.4$ $\mu = \frac{1-0.4}{0.4} = 1.5$ $\sigma = \sqrt{\frac{1-0.4}{0.4^2}} = 1.9364917$	`rgeom(20,0.4)`	`=ceiling(log(1-rand())/log(1`
Lottery machine with numbers $X_1, X_2, ..., X_N$	$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$	$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}$	$\mathbf{X} = \{1,1,5,7,8,8\}$ $\mu = \frac{1+1+5+7+8+8}{6} = 5$ $\sigma = \sqrt{\frac{(1-5)^2+(1-5)^2+(5-5)^2+(7-5)^2+(8-5)^2+(8-5)^2}{6}}$	`sample(c(1,1,5,7,8,8),20,T)`	`=INDEX(array;RANDBETWEEN(1;N(array)))` in spreadsheet
Discrete probability distribution with numbers $X_1, X_2, ..., X_N$ and probabilities $P_1, P_2, ..., P_N$	$\mu = \sum_{i=1}^{N} P_i \cdot X_i$	$\sigma = \sqrt{\sum_{i=1}^{N} P_i \cdot (X_i-\mu)^2}$	$\mathbf{X} = \{1,5,7,8\}$ $\mathbf{P} = \left\{\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{3}\right\}$ $\mu = \frac{1}{3}\cdot 1 + \frac{1}{6}\cdot 5 + \frac{1}{6}\cdot 7 + \frac{1}{3}\cdot 8 = 5$ $\sigma = \sqrt{\frac{1}{3}\cdot(1-5)^2 + \frac{1}{6}\cdot(5-5)^2 + \frac{1}{6}\cdot(7-5)^2 + \frac{1}{3}\cdot(8-5)^2} = 3$	`sample(c(1,5,7,8),20,T,c(1/3,1/6,1/6,1/3))`	`=INDEX(array;MATCH(RAND();CUMSUM(probabilities);1))` in spreadsheet
Exponential distribution with rate λ	$\mu = \frac{1}{\lambda}$	$\sigma = \frac{1}{\lambda}$	$\lambda = 0.4$ $\mu = \frac{1}{0.4} = 2.5$ $\sigma = \frac{1}{0.4} = 2.5$	`rexp(20,0.4)`	`=-LN(1-RAND())/0.4`

Other possibilities include:

- binomial

  - `rbinom(n,size,prob)`
- poisson
  - `rpois(n,lambda)`
- Student's t distribution
  - `rt(n,df)`

The general formulas are provided here (but easier, more specific formulas are provided later).

Mean of discrete population (general formula) with probability function $\mathbb{P}[x]$:

$$\mu = \sum x \cdot \mathbb{P}[x]$$

Standard deviation of discrete population (general formula):

$$\sigma = \sqrt{\sum \mathbb{P}[x] \cdot (x-\mu)^2}$$

Mean of continuous population (general formula) with density function $f[x]$:

$$\mu = \int_{-\infty}^{\infty} x \cdot f[x] \cdot \mathrm{d}x$$

Standard deviation of continuous population (general formula):

$$\sigma = \sqrt{\int f[x] \cdot (x-\mu)^2 \cdot \mathrm{d}x}$$

In this class, you will be expected to use those discrete general formulas (but not the continuous versions). For this project, you will only need to use the general discrete formulas if you make a spinner with unequally sized regions (or use a computer to mimic a spinner with unequally sized regions).

Dice

Rolling a die is equivalent to a special case of the discrete uniform distribution. If a die has N sides labeled 1 through N, then the following formulas will calculate the population mean and population standard deviation.

$$\mu_{\text{dice}} = \frac{N+1}{2}$$

$$\sigma_{\text{dice}} = \sqrt{\frac{N^2-1}{12}}$$

So, for example, if you have an 8-sided die.

$$\mu_{\text{eight-sided-die}} = \frac{8+1}{2} = 4.5$$

$$\sigma_{\text{eight-sided-die}} = \sqrt{\frac{8^2-1}{12}} = 2.2912878 \approx 2.291$$

Lottery machine, cards, or spinner with equally-sized regions

If you have N equally-likely outcomes $(x_1, x_2, \ldots, x_N)$, then

$$\mu = \frac{\sum x}{N}$$
$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

Notice we do **not** use Bessel's correction for the population. In your deck, there can be repeats. For example, your deck (population) could have the following cards (ignoring suits):

$$1, 1, 5, 7, 8, 8$$

Then,

$$\mu = \frac{1+1+5+7+8+8}{6} = 5$$

and

$$\sigma = \sqrt{\frac{(1-5)^2+(1-5)^2+(5-5)^2+(7-5)^2+(8-5)^2+(8-5)^2}{6}} = 3$$

Common continuous distributions

Standard uniform The standard uniform distribution generates numbers between 0 and 1. Any spreadsheet will produce standard uniform values with the `RAND()` function. In R, you can use the `runif()` function.

Uniform A continuous uniform distribution is often a fundamental component to a computer's capabilities.

In most spreadsheets, you can use `RANDBETWEEN(bottom; top)`.

Other

If you use a lottery machine (numbers in a hat) or cards, remember you need to replace the number and reshuffle after each draw to get independent identically distributed numbers.

You will choose a well-characterized population (random number generator) to sample from. Here, well-characterized means you know the population mean and population standard deviation.

Some possible populations are dice, a spinner, a deck of cards, a lottery machine (numbers in a hat), or (probably easiest) a computer's random number generation. We will want independent identically distributed random numbers. This means if you use cards or a lottery machine, then you need to replace the number into the pile and reshuffle before drawing again.

Before sampling, you will first calculate the following parameters of your population:

- population mean
- population standard deviation

Then, we will sample from the population, using a sample size of at least 20. We will compare each estimate with actual parameter by calculating the relative difference.

$$\text{relative difference} = \frac{\text{estimate} - \text{parameter}}{\text{parameter}}$$

You will need at least 20 measurements. Then,

- Summarize the measurements by producing the following:
  - the sample mean (and appropriate relative difference)
  - sample standard deviation (and appropriate relative difference)
  - a relative frequency distribution and/or histogram of the measurements