

Rodrigo Becerril Ferreyra  
 E E 381 Section 12  
 Lab 5  
 18 November 2020

## Introduction

The purpose of this lab is to experience random sampling from a population, and using a sample to estimate properties about the entire population. Specifically, we were tasked with creating samples  $\bar{X}$  of various sizes, finding their mean  $\mu_{\bar{X}}$ , and comparing their values to the population mean  $\mu$  (known beforehand).

The probability that the population mean  $\mu$  is in between two values satisfies the following equation:

$$P\left(\bar{X} - z \frac{\hat{S}}{\sqrt{n}} \leq \mu \leq \bar{X} + z \frac{\hat{S}}{\sqrt{n}}\right) = F(z), \quad (1)$$

where  $\bar{X} \pm z \frac{\hat{S}}{\sqrt{n}}$  are the upper and lower bounds of  $\mu$  and  $F(z)$  represents the cumulative distribution function (CDF) of the distribution in question.  $F(z)$  represents the probability that  $\mu$  lies between the two bounds. Common choices for  $F(z)$  are 0.95 and 0.99; the values of  $z$  can be found using common tables, and are listed below for convenience.

$F(z)$	Standard Normal Distribution $\mu = 0, \sigma = 1$	Student's t-Distribution $\nu = 4$
0.95	$z = 1.96$	$z = 2.78$
0.99	$z = 2.58$	$z = 4.60$

Table 1: Table of common values for  $F(z)$ .

The population for this lab was generated using a normal distribution with a mean of  $\mu = 55$  g and a standard deviation of  $\sigma = 5$  g. The population size is 1 500 000.

## 1 Problem 1

### 1.1 Question

In this problem, the task is to take 200 samples, with each sample's size increasing each time (for example, the first sample has a size of 1, the second sample has a size of 2, etc.). The mean of each sample was taken, and plotted against its sample size. It is worth noting that the sampling was done with replacement; ergo, it is possible that a certain value was selected more than once.

## 1.2 Results

The graphs of the sample means vs sample sizes compared with the theoretical upper and lower bounds given by  $g_{99}(\text{size}) = \mu \pm \frac{2.58\sigma}{\sqrt{\text{size}}}$  and  $g_{95}(\text{size}) = \mu \pm \frac{1.96\sigma}{\sqrt{\text{size}}}$  are listed below<sup>1</sup>. Note that the data (points on the plot) are the same; the only difference between the two plots are the upper and lower bounds of the mean. For the first plot, it can be assumed that 99% of the points are within the boundaries; for the second plot, 95% of the points are within the boundaries.

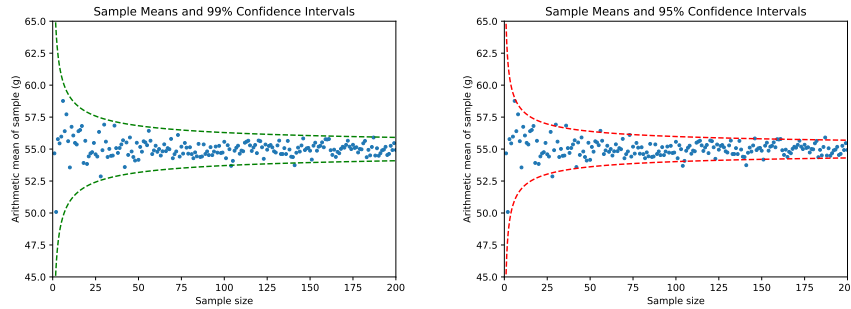


Figure 1: Scatter plots of sample mean vs. sample size.

Note that as the sample size approaches infinity, the sample mean approaches  $\mu$ . This problem took about 0.168 s to complete.

## 2 Problem 2

### 2.1 Question

In this problem, the task is to compare the use of  $z$ -values (see Table 1) in order to estimate the upper and lower limits of the population mean. This was done for various sample sizes, using both normal and Student's  $t$ -distributions.

In this lab, a large sample size is defined as having greater than or equal to 30 elements, while a small sample size is defined as having less than 30 elements. For small sample sizes, the  $t$ -distribution is a better estimator of the population mean, while the normal distribution is a better estimator for large samples. This can be seen in Table 2.

The process to obtain the values in Table 2 is as follows: first, the mean and sample standard deviation of the sample were obtained using the following

<sup>1</sup>The values 2.58 and 1.96 were selected from Table 1.

equations:

$$\bar{X} = \frac{1}{\text{size}} \sum X \quad (2)$$

$$\hat{S} = \sqrt{\frac{1}{\text{size} - 1} \sum (X - \bar{X})^2} \quad (3)$$

Using the mean calculated in (2) and the sample standard deviation calculated in (3), we can calculate the upper and lower bounds as follows:

$$\mu_{\text{lower}} = \bar{X} - z \frac{\hat{S}}{\sqrt{n}} \quad (4)$$

$$\mu_{\text{upper}} = \bar{X} + z \frac{\hat{S}}{\sqrt{n}} \quad (5)$$

where  $z$  is taken appropriately from Table 1.

In Problem 2, 10 000 tests were conducted for each sample size. If the expression  $\mu \in [\mu_{\text{lower}}, \mu_{\text{upper}}]$  holds true, then that test is considered a success; if not, then the test is a failure. The ratio of the number of successes to the total number of tests is shown in Table 2.

## 2.2 Results

Sample size ( $n$ )	$F(z) = 0.95$ Normal	$F(z) = 0.99$ Normal	$F(z) = 0.95$ Student's t	$F(z) = 0.99$ Student's t
5	87.35%	93.48%	94.79%	98.94%
40	94.18%	98.59%	99.10%	100.0%
120	94.65%	98.91%	99.27%	100.0%

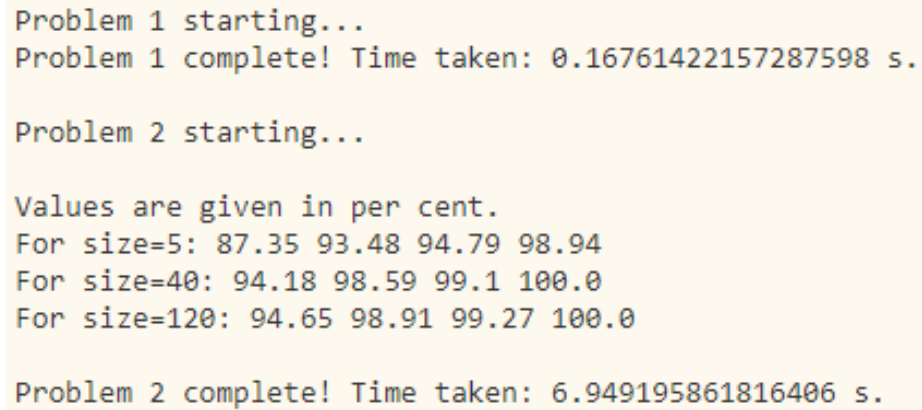
Table 2: Success rate for different sample sizes.

According to the definition of small and large sample sizes for this lab, it is inappropriate to use a value of  $z$  associated with the normal CDF for size = 5; this can be seen clearly, as 0.8735 is not close to 0.95, and 0.9348 is not close to 0.99. The values for the Student's t-distribution are much closer (0.9479 is close to 0.95 and 0.9894 is close to 0.99), which makes it the superior choice for small sample sizes. This fact is reversed, however, when size = 40 and size = 120. In these cases, the normal distribution values are much closer to their expected values, but the t-distribution values are much higher, and sometimes these values are 1, meaning that, in these cases,  $\mu_{\text{lower}}$  was very small and  $\mu_{\text{upper}}$  was very big such that they tell no information on the location of the population mean; in other words, the range of  $[\mu_{\text{lower}}, \mu_{\text{upper}}]$  is too broad. Therefore, it is best to use a normal distribution CDF value for large sample sizes.

This problem took about 6.95 s to compute.

### 3 Media

The following is an image displaying the output of the `main.py` file.

A screenshot of a terminal window with a light yellow background. The text is in a monospaced font. It shows the execution of a program with two problems. Problem 1 is completed quickly, while Problem 2 takes significantly longer and outputs a table of values for different sizes.

```
Problem 1 starting...
Problem 1 complete! Time taken: 0.16761422157287598 s.

Problem 2 starting...

Values are given in per cent.
For size=5: 87.35 93.48 94.79 98.94
For size=40: 94.18 98.59 99.1 100.0
For size=120: 94.65 98.91 99.27 100.0

Problem 2 complete! Time taken: 6.949195861816406 s.
```

Figure 2: Output from source file.