Social Sciences Intro to Statistics

Week 4.1 Sampling Distribution

Week 4: Learning goal - Apply understanding of central limit theorem and sampling distributions towards how to evaluate inferential statistics in R.

Introduction

Lecture overview:

- Central Limit Theorem
- Sampling Distributions

Load packages:

```r
library(tidyverse)
library(ggplot2)
library(labelled)
library(patchwork)

# Load ipeds dataset from course website
load(url('https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/data/ipeds/outpu
```

```
#> Rows: 965
#> Columns: 38
#> $ instnm           <chr> "Alabama A & M University", "University of Alabama a~
#> $ unitid           <dbl> 100654, 100663, 100706, 100724, 100751, 100830, 1008~
#> $ opeid6           <chr> "001002", "001052", "001055", "001005", "001051", "0~
#> $ opeid            <chr> "00100200", "00105200", "00105500", "00100500", "001~
#> $ control          <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 2, ~
#> $ c15basic         <dbl+lbl> 18, 15, 16, 19, 16, 18, 16, 20, 18, 18, 19, 18, ~
#> $ stabbr           <chr+lbl> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", ~
#> $ city              <chr> "Normal", "Birmingham", "Huntsville", "Montgomery", ~
#> $ zip               <chr> "35762", "35294-0110", "35899", "36104-0271", "35487~
#> $ locale            <dbl+lbl> 12, 12, 12, 12, 13, 12, 13, 12, 23, 43, 21, 13, ~
#> $ region            <dbl+lbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
#> $ tuit_grad_res     <dbl> 10128, 8424, 10632, 7416, 11100, 7812, 10386, 15325,~
#> $ fee_grad_res      <dbl> 1414, 0, 1054, 2740, 690, 766, 1784, 900, 1000, 190,~
#> $ tuit_grad_nres    <dbl> 20160, 19962, 24430, 14832, 31460, 17550, 31158, 153~
#> $ fee_grad_nres     <dbl> 1414, 0, 1054, 2740, 690, 766, 1784, 900, 1000, 190,~
#> $ tuit_md_res       <dbl> NA, 31198, NA, NA, 31198, NA, NA, NA, NA, NA, NA, NA~
#> $ fee_md_res        <dbl> NA, 3464, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, NA,~
#> $ tuit_md_nres      <dbl> NA, 62714, NA, NA, 62714, NA, NA, NA, NA, NA, NA, NA~
#> $ fee_md_nres       <dbl> NA, 3464, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, NA,~
#> $ tuit_law_res      <dbl> NA, NA, NA, NA, 24080, NA, NA, 39000, NA, NA, NA, NA~
#> $ fee_law_res       <dbl> NA, NA, NA, NA, 300, NA, NA, 325, NA, NA, NA, NA, 65~
#> $ tuit_law_nres     <dbl> NA, NA, NA, NA, 44470, NA, NA, 39000, NA, NA, NA, NA~
#> $ fee_law_nres      <dbl> NA, NA, NA, NA, 300, NA, NA, 325, NA, NA, NA, NA, 65~
#> $ books_supplies    <dbl> 1600, 1200, 2416, 1600, 800, 1200, 1200, 1800, 998, ~
#> $ roomboard_off     <dbl> 9520, 14330, 11122, 7320, 14426, 10485, 14998, 8020,~
#> $ oth_expense_off   <dbl> 3090, 6007, 4462, 5130, 4858, 4030, 6028, 4600, 3318~
#> $ tuitfee_grad_res  <dbl> 11542, 8424, 11686, 10156, 11790, 8578, 12170, 16225~
#> $ tuitfee_grad_nres <dbl> 21574, 19962, 25484, 17572, 32150, 18316, 32942, 162~
#> $ tuitfee_md_res    <dbl> NA, 34662, NA, NA, 31198, NA, NA, NA, NA, NA, NA, NA~
#> $ tuitfee_md_nres   <dbl> NA, 66178, NA, NA, 62714, NA, NA, NA, NA, NA, NA, NA~
#> $ tuitfee_law_res   <dbl> NA, NA, NA, NA, 24380, NA, NA, 39325, NA, NA, NA, NA~
#> $ tuitfee_law_nres  <dbl> NA, NA, NA, NA, 44770, NA, NA, 39325, NA, NA, NA, NA~
#> $ coa_grad_res      <dbl> 25752, 29961, 29686, 24206, 31874, 24293, 34396, 306~
#> $ coa_grad_nres     <dbl> 35784, 41499, 43484, 31622, 52234, 34031, 55168, 306~
#> $ coa_md_res        <dbl> NA, 56199, NA, NA, 51282, NA, NA, NA, NA, NA, NA, NA~
#> $ coa_md_nres       <dbl> NA, 87715, NA, NA, 82798, NA, NA, NA, NA, NA, NA, NA~
#> $ coa_law_res       <dbl> NA, NA, NA, NA, 44464, NA, NA, 53745, NA, NA, NA, NA~
#> $ coa_law_nres      <dbl> NA, NA, NA, NA, 64854, NA, NA, 53745, NA, NA, NA, NA~
#> Rows: 200
#> Columns: 4
#> $ norm_dist    <dbl> 42.70513, 50.24400, 61.29008, 45.47494, 44.74406, 47.9912~
#> $ rskew_dist   <dbl> 0.34451771, 0.31359906, 0.09375337, 0.05581678, 0.0744584~
#> $ lskew_dist   <dbl> 0.6554823, 0.6864009, 0.9062466, 0.9441832, 0.9255415, 0.~
#> $ stdnorm_dist <dbl> -1.45897348, 0.04880097, 2.25801577, -0.90501164, -1.0511~
#> [1] 32528.35
#> [1] 31620.8
```

Central Limit Theorem - What Is It?

Central Limit Theorem (CLT) describes the behavior we observe on the average of a large number of independent and identically distributed random variables. It states that, regardless of the shape of the original distribution, the distribution of the sum (or average) of these variables approaches a normal (Gaussian) distribution as the sample size increases, provided that the sample size is sufficiently large. Most count n=30 or more as a “large” sample size.

A key point is the idea that as the sample size increases, the distribution of the sample mean approaches a normal distribution. This holds true regardless of the shape of the original distribution.

When we start to take many, many samples, the average of the samples (also know as our sample mean) will start to look like a normal distribution.

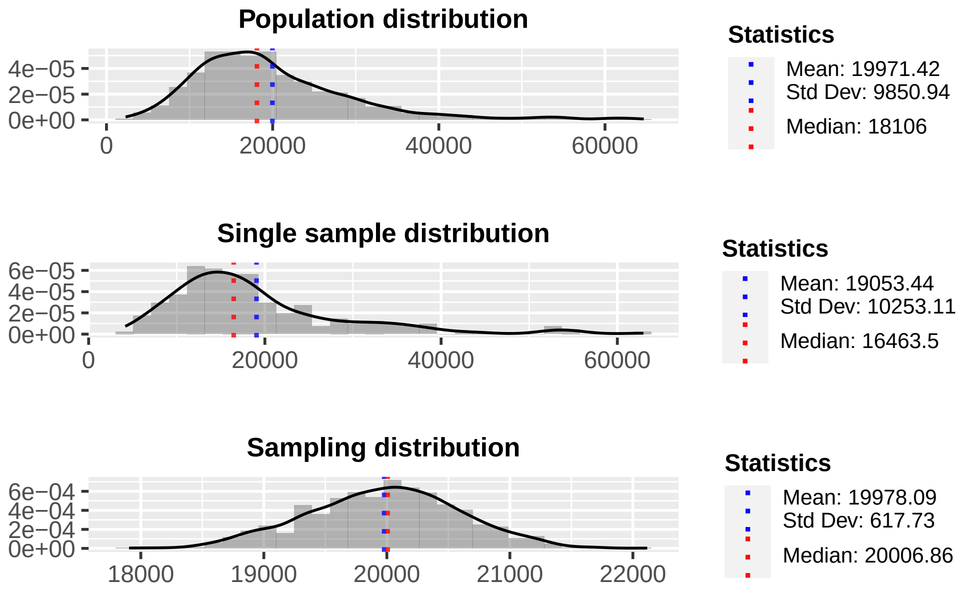
As the sample size gets bigger or approaches infinity, we would see a frequency plot that is extremely close to a perfect normal distribution.

Why is central limit theorem important?

Central limit theorem is important when we conduct hypothesis tests about a population parameter (e.g., about a population mean, about a population regression coefficient), based on the sampling distribution of the relevant sample statistic. Even if the underlying variable of interest has a skewed population, the sampling distribution of the sample statistic (e.g., sample mean) will have a normal distribution, because of CLT.

If the sampling distribution has a normal distribution, then we know the percent of observations that are a certain number of standard deviations from the mean.

```
# Show central limit theorem using a very skewed variable: non-resident, grad school cost of
plot_distribution(df_ipeds_pop$tuitfee_grad_nres, plot_title = 'Population distribution') +
  plot_distribution(df_ipeds_sample$tuitfee_grad_nres, plot_title = 'Single sample distribut
  plot_distribution(get_sampling_distribution(df_ipeds_pop$tuitfee_grad_nres),
                    plot_title = 'Sampling distribution') +
  plot_layout(ncol = 1)
```



Central limit theorem using interactive simulation

- Sampling distribution simulation LINK

Sampling distribution

Sampling distribution is the distribution of the statistic that we get when a population is repeatedly sample. In other words, it describes the data chosen for a sample from the population. It is the fundamental concept of inferential statistics, and how we can understand the population parameters when we cannot necessarily measure the whole population.

Briefly, recall the goal of inferential statistics:

- We want to make statements about population parameters, for example the population mean of some variable x, μ_x
  - But we usually cannot obtain data on the entire population
- Therefore, we collect a representative (random) sample from the population
- We calculate "estimates" based on this sample data. These estimates are our best guess abut the value of population parameters
- For example, the sample mean $\bar{x}$ is our best guess of the population mean μ_x

Usually, we collect a single sample from the population. How do we know if the sample we collected is representative of the underlying population we want to make statements about?

- This is a problem that statisticians have thought a lot about

For example, for our variable `norm_dist` from the data frame `df_generated_pop`, we randomly draw `30` observations from a population of `10,000` observations

```
set.seed(321)
norm_dist_s1 <- sample(x = df_generated_pop$norm_dist, size = 30)

mean(df_generated_pop$norm_dist)
#> [1] 49.98631
mean(norm_dist_s1) # mean of our sample
#> [1] 50.19565
```

But, what if we had obtained a different random sample?

```
set.seed(123)
norm_dist_s1 <- sample(x = df_generated_pop$norm_dist, size = 30)

mean(df_generated_pop$norm_dist)
#> [1] 49.98631
mean(norm_dist_s1) # mean of our sample
#> [1] 49.22062

# remove object norm_dist_s1
rm(norm_dist_s1)
```

So we can see that the sample mean, $\bar{x}$, changes from sample to sample.

Imagine if we take 1,000 random samples of size `n` (e.g., `30`) from a population.

- For each random sample, we calculate the sample mean, and record the value of the sample mean.
- We would have 1,000 observations, where each observation is the value of a sample mean.
- If we plotted these 1,000 observations, it would give us a distribution of sample means.
- More specifically, this would give us the "sampling distribution" of the sample mean.

Sampling distribution (of the sample mean)

- The sampling distribution of the sample mean is a relative frequency distribution where each observation is the sample mean of a single random sample from a population.

- The sampling distribution shows how the value of the sample mean varies from sample to sample.

Excellent website for understanding how the sampling distribution works

- This very useful website does interactive simulations that show how the sampling distribution works LINK
  - **Please** spend 5 minutes playing around with this website; this is really the most important concept in statistics

The sampling distribution of any statistic

- So far, we have discussed the sampling distribution of the sample mean, but a sampling distribution can be created for any sample statistic (e.g., median, min, max, regression coefficient)
- Once we get to the unit on regression, we'll be thinking about the sampling distribution of a regression coefficient. But the underlying concepts will be exactly the same as the sampling distribution of the sample mean.