Social Sciences Intro to Statistics

Week 4.2 Fundamentals of Inferential Statistics

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

Introduction

Lecture overview:

- Population parameters
- Sample estimates
- Using sample estimates to make inferences about population parameters

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
```

Fundamentals of inferential statistics

Population parameters

Population parameter: A descriptive measure of the entire population. In other words, data that refers, summarizes, or describes an aspect of the population. Commonly used population parameters are:

- Number of cases: The number of data points we see in the population N
- Mean: The average of the population μ (mu)
- Standard deviation: On average how far each data point is from the mean. It also tells us how well the mean of a variable represents the central tendency of a population σ (Sigma)
- Variance: The average of the square of the deviations from the population's mean value σ^2 (Sigma-squared)
- Range: The difference between the largest and smallest value in the population

```r
# Let's test out finding some population parameters in the IPEDS population data frame

# Number of cases
nrow(df_ipeds_pop) # each row is a case
#> [1] 965

# Mean of the cost of book supplies in the population
mean(df_ipeds_pop$books_supplies)
#> [1] 1217.826

# Standard deviation of the cost of book supplies in the population
sd(df_ipeds_pop$books_supplies)
#> [1] 456.6553

# Variance of the cost of book supplies in the population
var(df_ipeds_pop$books_supplies)
#> [1] 208534.1

# Range of the cost of book supplies in the population
max(df_ipeds_pop$books_supplies, na.rm = TRUE) - min(df_ipeds_pop$books_supplies, na.rm = TRU
#> [1] 8470
range(df_ipeds_pop$books_supplies) # to see the smallest and largest values in the range
#> [1]    0 8470
```

Sample statistics

Sample statistic: A sample estimate used to make inferences about population parameters. A sample statistic can come from a single measurement or a series of measurements from the sample. Commonly used sample statistics are:

- Number of cases: The number of data points we see in the sample n (notice that the notations are different from N used for population parameter)
- Mean: The average of the sample X
- Standard deviation: On average how far each data point is from the mean. It also tells us how well the mean of a variable represents the central tendency of a sample S
- Variance: The average of the square of the deviations from the sample's mean value s^2
- Range: The difference between the largest and smallest value in the sample

```r
# Let's test out finding some sample statistics in the IPEDS sample data frame

# Number of cases
nrow(df_ipeds_sample) # each row is a case
#> [1] 200

# Mean of the cost of book supplies in the population
mean(df_ipeds_sample$books_supplies)
#> [1] 1186.685

# Standard deviation of the cost of book supplies in the population
sd(df_ipeds_sample$books_supplies)
#> [1] 401.8004

# Variance of the cost of book supplies in the population
var(df_ipeds_sample$books_supplies)
#> [1] 161443.6

# Range of the cost of book supplies in the population
max(df_ipeds_sample$books_supplies, na.rm = TRUE) - min(df_ipeds_sample$books_supplies, na.r
#> [1] 2400
range(df_ipeds_sample$books_supplies) # to see the smallest and largest values in the range
#> [1]    0 2400
```

Using sample estimates to make inferences about population parameters

Compared to what we found for the population parameters, do you see any similarities? Was this a good sample in comparison to the population?

Since it is not always possible to collect information or data on the entire population. We rely on the data collected from the samples to make inferences about the population.

From our examples above, we can see that the sample of 200 gave us a relatively close estimate on the mean, standard deviation, and variance of the population. We did not have the best estimates for number of cases or the range but that is understandable based on the probability of what was selected for our sample.