Assignment 1

Introduction to Applied Statistics and Data Science

Sophie Ennis & Hannah Yoon

2025-02-02

rm(list=ls())

Load Libraries

library(tidyverse)  
library(rio)

Import GSS Cleaned Data

setwd("C:/Users/yoonh/OneDrive - The University of Chicago/Courses/Winter 2025/SOCI 30617 - Intro to Applied Stats and Data Science/R Demonstrations/")  
gss <- import("GSS2018cleaned.rdata")

# QUESTION 1

### Consider each of the following cleaned variables: OccPres, Working, Degree, and Age. For each of the four…

## Part A

### Identify the scale of the variable.

Variable OccPres is interval-scale because it is quantitative, ordered and equally spaced, but has no true zero.

Variable Working is nominal because it is categorical and unordered.

Variable Degree is ordinal because it is categorical and has a sense of order.

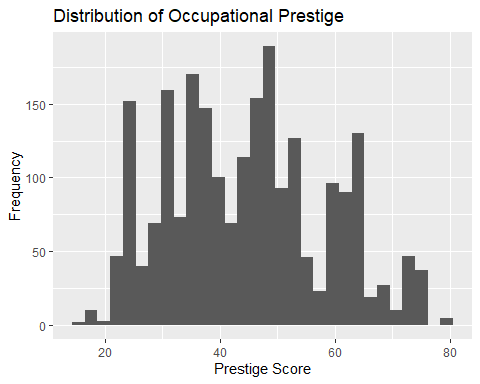
Variable Age is ratio-scale because it is an interval-scale variable with a true zero (true zero point at birth).

## Part B

### Provide a graphic display of the distribution.

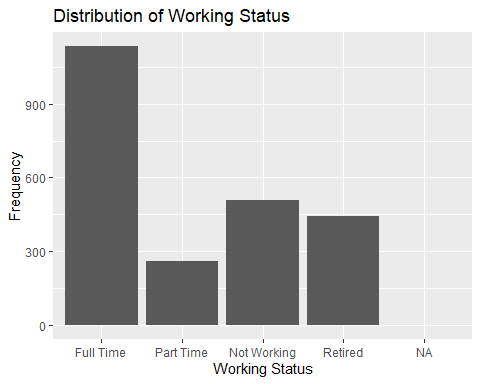
#### 1b – OccPres Histogram

ggplot(gss, aes(OccPres)) +   
 geom\_histogram(bins=30) +   
 labs(   
 title = "Distribution of Occupational Prestige",   
 x = "Prestige Score",   
 y = "Frequency")



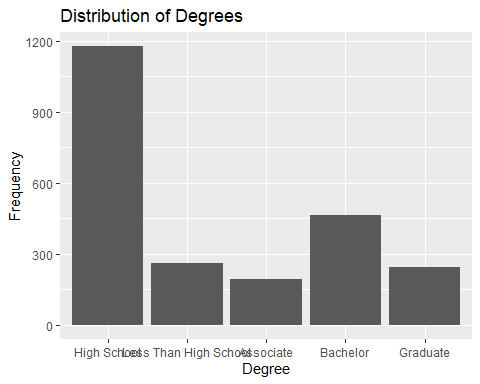
#### 1b – Working

ggplot(gss, aes(Working)) +   
 geom\_bar() +   
 labs(   
 title = "Distribution of Working Status",   
 x = "Working Status",   
 y = "Frequency")



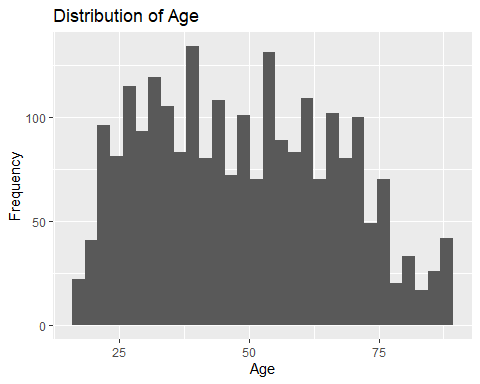
#### 1b – Degree

ggplot(gss, aes(Degree)) +   
 geom\_bar() +   
 labs(   
 title = "Distribution of Degrees",   
 x = "Degree",   
 y = "Frequency")



#### 1b – Age

ggplot(gss, aes(Age)) +   
 geom\_histogram(bins=30) +   
 labs(   
 title = "Distribution of Age",   
 x = "Age",   
 y = "Frequency")



## Part C

### Based on the scale, provide a relevant measure of central tendency. For continuous variables, provide an appropriate measure of dispersion and describe the shape of the distribution.

Since both OccPres and Age distributions are normal (supported by the small difference between the mean and median values), we chose to present the mean in the table. We chose standard deviation as the appropriate measure of dispersion (rather than IQR or range) because it takes every data point into account in the calculation of the spread. Working and Degree are not continuous, so we present the mode (the category that appears most often) as the relevant measure of central tendency.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | OccPres | Working | Degree | Age |
| Measure of Central Tendency | Mean: 44.68 | Mode:  Full Time (1134) | Mode:  High School (1178) | Mean: 48.97 |
| Measure of Dispersion | SD: 13.64 | N/A | N/A | SD: 18.06 |
| Shape | Normal Distribution | N/A | N/A | Normal Distribution |

#### Question 1c – OccPres

The shape of the data is a normal distribution, as the mean (44.68) and the median (45) are very close. The dispersion can be described by the standard deviation, which is 13.64.

summary(gss$OccPres)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 16.00 35.00 45.00 44.68 53.00 80.00 100

mean(gss$OccPres, na.rm = TRUE)

## [1] 44.68238

median(gss$OccPres, na.rm = TRUE)

## [1] 45

sd(gss$OccPres, na.rm = TRUE)

## [1] 13.64499

IQR(gss$OccPres, na.rm = TRUE)

## [1] 18

range(gss$OccPres, na.rm = TRUE)

## [1] 16 80

#### Question 1c – Working

The mode of the data is the category Full Time, which has a frequency of 1134.

summary(gss$Working)

## Full Time Part Time Not Working Retired NA's   
## 1134 259 508 445 2

#### Question 1c – Degree

The mode of the data is the category High School, which has a frequency of 1178.

summary(gss$Degree)

## High School Less Than High School Associate   
## 1178 262 196   
## Bachelor Graduate   
## 465 247

#### Question 1c – Age

The shape of the data is a normal distribution, as the mean (48.97) and the median (48) are very close. The dispersion can be described by the standard deviation, which is 18.06.

summary(gss$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 18.00 34.00 48.00 48.97 63.00 89.00 7

mean(gss$Age, na.rm = TRUE)

## [1] 48.97138

median(gss$Age, na.rm = TRUE)

## [1] 48

sd(gss$Age, na.rm = TRUE)

## [1] 18.06088

IQR(gss$Age, na.rm = TRUE)

## [1] 29

range(gss$Age, na.rm = TRUE)

## [1] 18 89

# QUESTION 2

### Max Weber is famous (in part) for his treatise on how religious beliefs shape economic outcomes. But how might we treat religious belief in a statistical analysis? Describe ways in which religious belief could be operationalized as…

#### A Nominal Variable

Religious belief could be operationalized as a nominal variable through the measurement of the type of religion that an individual practices. Type of religion (such as Christianity, Judaism, Islam, etc.) is categorical and unordered.

#### An Ordinal Variable

Religious belief could be operationalized as an ordinal variable through the measurement of religiosity. Religiosity would be an ordinal variable because, like a Likert scale, it would have categories that have a natural order (e.g., very not religious, not religious, neutral, religious, very religious).

#### A Ratio-Scale Variable

Religious belief could be operationalized as a ratio-scale variable through the measurement of religious service attendance. This would be a ratio-scale variable because the number of services an individual attends (e.g., church attendance) could be measured, and the number of services attended includes 0 for those who do not attend religious services. (Number of services is quantitative, ordered, and equally spaced.) This inclusion of a true 0 makes this a ratio-scale variable.

# QUESTION 3

### In a Gallup poll in 2023, a plurality of Americans stated that two children were an optimal number in a family. We wish to know if American families have, on average, two children.

## Part A

### State the null and alternative hypotheses.

**Null Hypothesis:** American families have, on average, two children.

H0 : μ = 2 Children

**Alternative Hypothesis:** American families do NOT have, on average, two children.

Ha : μ ≠ 2 Children

## Part B

### Test your hypothesis using the z-distribution. Provide your test statistic, a p-value, and the conclusion of your hypothesis test. Use a critical value of .05.

#### 3b – z-distribution

x\_bar <- mean(gss$NChildren, na.rm = TRUE)   
  
sample\_sd <- sd(gss$NChildren, na.rm = TRUE)   
  
n <- length(gss$NChildren[!is.na(gss$NChildren)])   
  
mu <- 2  
  
z\_score <- (x\_bar - mu)/(sample\_sd/(sqrt(n)))  
  
z\_score

## [1] -4.182071

p\_value <- 2\*pnorm(z\_score)  
  
p\_value

## [1] 2.888659e-05

The test statistic is -4.182. The p-value is 0.0000288.  
Conclusion: Since the p-value is less than the critical value of 0.05, we reject the null hypothesis (in favor of the alternative hypothesis). From this, we know that American families do NOT have, on average, two children.

## Part C

### Test your hypothesis using the t-distribution. Provide your test statistic, a p-value, and the conclusion of your hypothesis test. Use a critical value of .05.

#### 3b – t-distribution

t.test(gss$NChildren, mu = 2)

##   
## One Sample t-test  
##   
## data: gss$NChildren  
## t = -4.1821, df = 2343, p-value = 2.995e-05  
## alternative hypothesis: true mean is not equal to 2  
## 95 percent confidence interval:  
## 1.787561 1.923190  
## sample estimates:  
## mean of x   
## 1.855375

The test statistic, using the t-distribution, is -4.182. The p-value is 0.00002995.

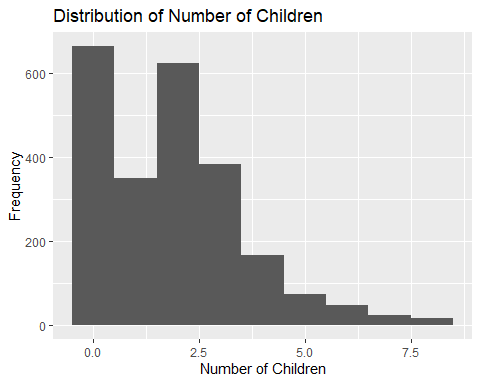
Conclusion: Since the p-value is less than the critical value of 0.05, we reject the null hypothesis (in favor of the alternative hypothesis). From this, we know that American families do NOT have, on average, two children.

# QUESTION 4

### Examine a histogram for the number of children in the sample. Describe the shape. Explain whether the shape of the sample distribution makes us concerned about our point estimates of the mean of this sample. Explain how this relates to your findings in part 3.

#### NChildren Histogram

ggplot(gss, aes(NChildren)) +   
 geom\_histogram(binwidth=1) +   
 labs(   
 title = "Distribution of Number of Children",   
 x = "Number of Children",   
 y = "Frequency")



#### Sample Size

n <- length(gss$NChildren[!is.na(gss$NChildren)])   
n

## [1] 2344

The shape is right-skewed. The shape of the sample distribution makes us concerned about our point estimates of the mean of this sample. When conducting hypothesis tests, we do it under the assumption that sample comes from a normal population distribution. However, if the sample data displays a right-skewed distribution, there is a possibility that the population distribution is not normal. However, the Central Limit Theorem (CLT) states that for random sampling with a large sample size (n = 2344), the sampling distribution of the sample mean is approximately a normal distribution – and it is true regardless of the population distribution. Thus, even if the assumption that the normal population distribution is violated, due to the CLT, we can assume that our point estimates of the mean and the hypothesis test remains valid.

# QUESTION 5

### Calculate the following confidence intervals for the number of children using the z distribution:

## Part A

### A 90% Confidence Interval

length(na.omit(gss$NChildren))

## [1] 2344

mean(gss$NChildren, na.rm = T)

## [1] 1.855375

sd(gss$NChildren, na.rm = T)

## [1] 1.674285

mean(gss$NChildren, na.rm = T) + (1.65 \* ( sd(gss$NChildren, na.rm = T) / sqrt (length(na.omit(gss$NChildren)))))

## [1] 1.912436

mean(gss$NChildren, na.rm = T) - (1.65 \* ( sd(gss$NChildren, na.rm = T) / sqrt (length(na.omit(gss$NChildren)))))

## [1] 1.798315

The 90% confidence interval for the number of children is from 1.798 to 1.912.

## Part B

### A 95% Confidence Interval

mean(gss$NChildren, na.rm = T) + (1.96 \* ( sd(gss$NChildren, na.rm = T) / sqrt (length(na.omit(gss$NChildren)))))

## [1] 1.923156

mean(gss$NChildren, na.rm = T) - (1.96 \* ( sd(gss$NChildren, na.rm = T) / sqrt (length(na.omit(gss$NChildren)))))

## [1] 1.787595

The 95% confidence interval for the number of children is from 1.787 to 1.923.

## Part C

### A 99% Confidence Interval

mean(gss$NChildren, na.rm = T) + (2.58 \* ( sd(gss$NChildren, na.rm = T) / sqrt (length(na.omit(gss$NChildren)))))

## [1] 1.944597

mean(gss$NChildren, na.rm = T) - (2.58 \* ( sd(gss$NChildren, na.rm = T) / sqrt (length(na.omit(gss$NChildren)))))

## [1] 1.766154

The 99% confidence interval for the number of children is from 1.766 to 1.944.

## Part D

### Explain the advantage(s) and disadvantage(s) of using a higher level of confidence, especially in terms of hypothesis testing.

A higher level of confidence means that the interval will be wider. This means that there is a greater certainty that the true population parameter will be within the confidence interval (we are more sure that the interval will enclose the parameter), which is an advantage. However, a disadvantage is that we are increasing the interval at the cost of precision – we are less specific about the value of the parameter with a higher level of confidence because we have a larger range.

# QUESTION 6

### Many occupations in the United States are disproportionately male or female. For example, elementary school teachers are mostly women, while truck drivers are mostly men. Test whether men or women have equal occupational prestige.

## Part A

### State your null and alternative hypothesis.

**Null Hypothesis:** Men and women have equal occupational prestige.

H0 : μ(Men’s Occupational prestige)  = μ(Women’s Occupational Prestige)  OR

μ(Men’s Occupational prestige)  – μ(Women’s Occupational Prestige)  = 0

**Alternative Hypothesis:** Men and women do NOT have equal occupational prestige

μ(Men’s Occupational prestige)  ≠ μ(Women’s Occupational Prestige)  OR

μ(Men’s Occupational prestige)  – μ(Women’s Occupational Prestige)  ≠ 0

## Part B

### Test your hypothesis with the appropriate statistical test. Provide your test statistic, a p-value, and the conclusion of your hypothesis test. Use a critical value of .05.

t.test(gss$OccPres ~ gss$Sex)

##   
## Welch Two Sample t-test  
##   
## data: gss$OccPres by gss$Sex  
## t = -0.057819, df = 2185.4, p-value = 0.9539  
## alternative hypothesis: true difference in means between group Female and group Male is not equal to 0  
## 95 percent confidence interval:  
## -1.165410 1.098656  
## sample estimates:  
## mean in group Female mean in group Male   
## 44.66721 44.70059

The appropriate statistical test is the t test because we are comparing two means – the mean of men’s and women’s occupational prestige. The test statistic is -0.058 and the p-value is 0.954. The p value is greater than the critical value of 0.05, so we fail to reject the null hypothesis, indicating that the mean of men’s occupational prestige is equal to the mean of women’s occupational prestige.

## Part C

### State the assumption(s) under which your hypothesis test will be valid. Speculate on whether these assumptions are likely met by our data, and why.

#### Sample Size

n <- length(gss$Sex[!is.na(gss$Sex)])   
n

## [1] 2348

Since we are completing a t-test, the assumptions under which our hypothesis test will be valid are:

1. Randomization for the collection of the sample data
   1. The GSS is a random sample that is representative of the U.S. population.
2. Normal population distribution
   1. The t-test assumes that our sample data comes from a normal population distribution. However, we do not know the distribution of the population. Regardless, for random sampling with large sample sizes, the sampling distribution of the sample mean (in this case, the difference in mean of women & men’s occupational prestige) is approximately a normal distribution according to the Central Limit Theorem. This is true regardless of the shape of population distribution, and since our sample size is large, we are safe to complete the t-test even if this assumption is not true.

# QUESTION 7

### Although the value of four-year college degrees has been questioned in recent years, it is still considered an important predictor of economic success. Create a binary variable for whether or not a respondent has at least a Bachelor’s degree. Test whether members of the major racial groups defined by the cleaned Race variable are equally likely to hold Bachelor’s degrees.

## Part A

### State your null and alternative hypothesis.

Null Hypothesis

H0 : Race and having a Bachelor’s degree are statistically independent

Alternative Hypothesis

Ha : Race and having a Bachelor’s degree are statistically dependent

## Part B

### Test your hypothesis with the appropriate statistical test. Provide your test statistic, a p-value, and the conclusion of your hypothesis test. Use a critical value of .05.

gss$Bachelors <- ifelse (gss$Degree== "Bachelor", 1, 0)   
  
table(gss$Race, gss$Bachelors)

##   
## 0 1  
## White 1335 358  
## Black 318 67  
## Other 230 40

crosstab\_data <- table(gss$Race, gss$Bachelors)   
  
chisq.test(crosstab\_data)

##   
## Pearson's Chi-squared test  
##   
## data: crosstab\_data  
## X-squared = 7.5491, df = 2, p-value = 0.02295

The appropriate statistical test to test this hypothesis is the chi-square test because we want to test if categories are associated.

The test statistic is 7.549 and the p-value is 0.022.

The p value is less than the critical value of 0.05, so we reject the null hypothesis in favor of the alternative hypothesis, indicating that there is an association between an individual’s race and whether or not they have a Bachelor’s degree.