Lecture 8- Spring 2025

Villas-Boas

Lecture 8 EEP 118 Spring 2025

In [ ]:
```
# Load the 'pacman' package
install.packages("pacman")
library(pacman)
#packages to use load them now using the pacman "manager"
p_load(dplyr, readr)
#Another great feature of p_load(): if you try to load a package that is not
p_load(ggplot2)

#set scientific display off, thank you Roy
options(scipen=999)

# Loading packages
pacman::p_load(lfe, lmtest, haven, sandwich, tidyverse,psych)
# lfe for running fixed effects regression
# lmtest for displaying robust SE in output table
# haven for loading in dta files
# sandwich for producing robust Var-Cov matrix
# tidyverse for manipulating data and producing plots
# psych for using describe later on
```

In [ ]:
```
#-------------------------------------------
#1. Read in data
#-------------------------------------------
my_data2025 <- read_dta("dataLecture82025.dta")
head(my_data2025)
```

In [ ]:
```
#number of observations
nobserv2025<-nrow(my_data2025)

#answer is 20  (this is your response rate this year)
nobserv2025
```

Let us construct the 95% confidence interval for the true proportion os answering both questions 1 and 2 correctly

to do that, we need the sample average of p, which we call phat = number answering correctly divided by sample size N

$\hat{p} =\frac {number \ correct}{N}$

and we also need the std error of the sample mean proportion that is equal to the square root of the variance of $\hat{p}$

where the estimated variance of $\hat{p}$

is $\hat{var}(\hat{p})=\frac{ \hat{p} \ (1-\hat{p}) }{ N}$

Get the sample estimate of $\hat{p}$

In [ ]:
```
(phat2025<-mean(my_data2025$correct1and2))
```

In [ ]:
```
#and compute the variance of phat2025

var_phat2025<-phat2025*(1-phat2025)/nobserv2025

#show it
var_phat2025
```

Get $se(\hat{p})$, the sample estimated Standard error of $\hat{p}$

In [ ]:
```
#get the standard error, se,  of phat2025 is the square root of the variance

se_phat2025<-sqrt(var_phat2025)
se_phat2025
```

95% confidence interval for p

$ \ \hat{p} \ - \ c^{95\%} \ se(\hat{p}) \\ ; \ \ \hat{p} \ + \ c^{95\%} \ se(\hat{p}) \ $

where $ct^{95\%}$ is the two-tailed critical value for a N(0,1) distribution, that is, 1.96.

So the probability that the random CI= (phat– c se_phat , phat + c se_phat) includes the true value of p is 95%.

Derive a 95% confidence interval for p2025 and interpret in a sentence.

In [ ]:
```
#the lower part of the 95 % confidence interval is

ci95_l2025<-phat2025 - ( 1.96 * se_phat2025 )
ci95_l2025
```

In [ ]:
```
#the upper part of the 95 % confidence interval is
```

```
ci95_u2025<-phat2025 + ( 1.96 * se_phat2025 )
ci95_u2025
```

In [ ]:
```
ci95percent2025=cbind(ci95_l2025,ci95_u2025)
ci95percent2025

#will give you
#         ci95_l 2025    ci95_u2025
#[1,]       0.441          0.859
```

What would be the probability of guessing each question right?

Since there are three options, the probability of a guess is 1/3.

What is the probability that students guess both questions right?

It is 1/3 * 1/3 = 1/9=0.111

Does the Confidence interval we just created, that we are 95% sure contains the true proportion of students that answer both questions right, contain 0.111?

The answer is no.

You will learn then that we reject with 95% confidence that the students are not guessing both questions right (corresponds to p=0.11), since the 95% confid interval for the true p does not contain 0.11.

How wrong can we be, based on this analysis? 5% of the times we can be wrong, we are 95% confident...

There was some thinking going on in the answers, great job!

you were not just guessing...! We reject guessing based on your answers!!!

the end during Lecture

now do DA Lecture 8

do the same with data2024.dta

In [ ]:
```
#-------------------------------------------
#1. Read in data
#-------------------------------------------
my_data <- read_dta("data2024.dta")
head(my_data)
```

In [ ]:
```
#describe data
describe(my_data,skew = FALSE)
```

In [ ]:
```
# what is the proportion of correct question 1?
mean(mean(my_data$correct1))
```

In [ ]:
```
#what is the proportion of correct question2?
mean(mean(my_data$correct2))
```

create a new column correct 1 and 2

In [ ]:
```
#what is the proportion of both correct in general?
my_data$correct1and2<-my_data$correct1*my_data$correct2
mean(mean(my_data$correct1and2))
```

In [ ]:
```
#answer [1] 0.5555556
```

Let us construct the 95% confidence interval for the true proportion os answering both questions 1 and 2 correctly

to do that, we need the sample average of p, which we call

phat = number answering correctly divided by sample size N

$\hat{p} =\frac {number \ correct}{N}$

and we also need the std error of the sample mean proportion that is equal to the square root of the variance of $\hat{p}$

where the estimated variance of $\hat{p}$

is $\hat{var}(\hat{p})=\frac{ \hat{p} \ (1-\hat{p}) }{ N}$

In [ ]:
```
#let phat be the estimated proportion of both correct in general
phat<-mean(my_data$correctboth)
#show it
phat
```

In [ ]:
```
#number of observations
nobserv<-nrow(my_data)
```

```
#answer is 108
nobserv
```

In [ ]:
```
#and compute the variance of phat

var_phat<-phat*(1-phat)/nobserv

#show it
var_phat
```

In [ ]:
```
#se of phat is the square root of the variance

se_phat<-sqrt(var_phat)
se_phat
```

Derive a 95% confidence interval for p and interpret in a sentence.

critical value is approx 1.96, two-tailed, 5 percent for a N(0,1).

In [ ]:
```
#the lower part of the 95 % confidence interval is

ci95_l<-phat - ( 1.96 * se_phat )
ci95_l
```

In [ ]:
```
#the upper part of the 95 % confidence interval is

ci95_u<-phat + ( 1.96 * se_phat )
ci95_u
```

In [ ]:
```
ci95percent=cbind(ci95_l,ci95_u)
ci95percent

#will give you
#         ci95_l     ci95_u
#[1,] 0.4618389 0.6492722
```

THE END DA Lecture 8