Lecture 8- Spring 2025

Villas-Boas

Lecture 8 EEP 118 Spring 2025

In [1]:
```
# 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
```

```
Installing package into ‘/srv/r’
(as ‘lib’ is unspecified)

Installing package into ‘/srv/r’
(as ‘lib’ is unspecified)

also installing the dependencies ‘mnormt’, ‘GPArotation’



psych installed
```

In [2]:
```
#-------------------------------------------
#1. Read in data
#-------------------------------------------
my_data2025 <- read_dta("dataLecture82025.dta")
head(my_data2025)
```

A tibble: 6 × 10

var1	var2	var3	var4	var5	correct1	correct2	isSoccer
<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<c
2/11/25 9:44	Omitting unemployment in the regression will not affect beta2hat (for percent female)	Omitting rural area in the regression we will over estimate betahat_percent_wite	yes	no	1	1	
2/11/25 11:11	Omitting unemployment in the regression will not affect beta2hat (for percent female)	Omitting rural area in the regression we will over estimate betahat_percent_wite	yes	yes	1	1	
2/11/25 11:28	Omitting unemployment in the regression will not affect beta2hat (for percent female)	Omitting rural area in the regression we will over estimate betahat_percent_wite	yes	no	1	1	
2/11/25 11:32	Omitting unemployment in the regression will over estimate beta2hat (for percent female)	Omitting rural area in the regression we will over estimate betahat_percent_wite	yes	no	0	1	
2/11/25 11:37	Omitting unemployment in the regression will not affect beta2hat (for percent female)	Omitting rural area in the regression we will under estimate of betahat_percent_white	yes	yes	1	0	
2/11/25 11:38	Omitting unemployment in the regression will under estimate beta2hat (for percent female)	Omitting rural area in the regression we will over estimate betahat_percent_wite	no	no	0	1	

In [3]:
```
#number of observations
nobserv2025<-nrow(my_data2025)

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

20

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 [4]:
```
(phat2025<-mean(my_data2025$correct1and2))
```

0.65

In [5]:
```
#and compute the variance of phat2025

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

#show it
var_phat2025
```

0.011375

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

In [6]:
```
#get the standard error, se,  of phat2025 is the square root of the variance

se_phat2025<-sqrt(var_phat2025)
se_phat2025
```

0.106653645038508

95% confidence interval for p

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

where $ct^{95\%}$ is the two-tailed critical value for a t distribution with 95% probability between $-ct$ and ct.

To get the critical value ct, we use a t with (N-1)=19 degrees of freedom because the sample size is only N=20.

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

If the sample size was large, like in the beach data set, N=80, we would use a normal 0,1 and the critical there would be 1.96.

How do we read the t table?

we will do this formally in a future lecture, but a preview

No description has been provided for this image

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

critical value df=19 is 2.093 from the t table two tailed, 5 percent column, and row 19

Note that using 1.96 or the correct 2.093 matters very little in the end, so most people, to construct a CI fast, use 2 as the general critical value...

In class, I will ask you to get the correct critical values so you can practice to read normal and t statistical tables.

In [7]:
```
#the lower part of the 95 % confidence interval is

ci95_l2025<-phat2025 - ( 2.093 * se_phat2025 )
ci95_l2025
```

0.426773920934403

In [8]:
```
#the upper part of the 95 % confidence interval is

ci95_u2025<-phat2025 + ( 2.093 * se_phat2025 )
ci95_u2025
```

0.873226079065597

In [9]:
```
ci95percent2025=cbind(ci95_l2025,ci95_u2025)
ci95percent2025
```

```
#will give you
#         ci95_l 2025    ci95_u2025
#[1,]       0.426          0.873
```

A matrix: 1 × 2 of type dbl

ci95_l2025	ci95_u2025
0.4267739	0.8732261

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 [10]:
```
#------------------------------------------
#1. Read in data
#------------------------------------------
my_data <- read_dta("data2024.dta")
head(my_data)
```

A tibble: 6 × 9

timestamp	went2class	soccerfan	correct1	correct2	correctboth	numberCorrect	we
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	
2/6/2024 10:53:19	yes	yes	1	1	1	2	
2/6/2024 11:01:25	yes	yes	1	0	0	1	
2/6/2024 11:11:24	yes	yes	1	1	1	2	
2/6/2024 11:11:28	yes	yes	1	0	0	1	
2/6/2024 11:11:52	yes	yes	1	0	0	1	
2/6/2024 11:11:53	yes	no	1	1	1	2	

In [11]:
```
#describe data
describe(my_data,skew = FALSE)
```

A psych: 9 × 9

	vars	n	mean	sd	median	min	max	range	
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
timestamp*	1	108	53.8055556	30.9529212	53.5	1	107	106	2
went2class*	2	108	1.9351852	0.2473466	2.0	1	2	1	0
soccerfan*	3	108	1.5185185	0.5019864	2.0	1	2	1	0
correct1	4	108	0.9722222	0.1651017	1.0	0	1	1	(
correct2	5	108	0.5555556	0.4992206	1.0	0	1	1	0
correctboth	6	108	0.5555556	0.4992206	1.0	0	1	1	0
numberCorrect	7	108	1.5277778	0.5546462	2.0	0	2	2	0
went2Class	8	108	0.9351852	0.2473466	1.0	0	1	1	0
isSoccerFan	9	108	0.5185185	0.5019864	1.0	0	1	1	0

In [12]:
```
# what is the proportion of correct question 1?
mean(mean(my_data$correct1))
```

0.972222222222222

In [13]:
```
#what is the proportion of correct question2?
mean(mean(my_data$correct2))
```

0.555555555555556

create a new column correct 1 and 2

In [14]:
```
#what is the proportion of both correct in general?
my_data$correct1and2<-my_data$correct1*my_data$correct2
mean(mean(my_data$correct1and2))
```

0.555555555555556

In [15]:
```
#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 [16]:
```
#let phat be the estimated proportion of both correct in general
phat<-mean(my_data$correctboth)
#show it
phat
```

0.555555555555556

In [17]:
```
#number of observations
nobserv<-nrow(my_data)

#answer is 108
nobserv
```

108

In [18]:
```
#and compute the variance of phat
```

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

#show it
var_phat
```

0.00228623685413809

In [19]:
```
#se of phat is the square root of the variance

se_phat<-sqrt(var_phat)
se_phat
```

0.0478146092124372

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

critical value df=108 is approx 1.96, two tailed, 5 percent column, and row between 100 and 1000

In [20]:
```
#the lower part of the 95 % confidence interval is

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

0.461838921499179

In [21]:
```
#the upper part of the 95 % confidence interval is

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

0.649272189611933

In [22]:
```
ci95percent=cbind(ci95_l,ci95_u)
ci95percent

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

A matrix: 1 × 2 of type dbl

ci95_l	ci95_u
0.4618389	0.6492722

THE END DA Lecture 8