ENWONGO EKAH

PLCY610 PS2

ENWONGO EKAH

2/17/2021

R Markdown

Step 1 : Clear All

```
rm(list = ls())
```

Step 2: Import Data

```
library(readr)
library(tidyverse)

demo <- read_csv("C:/Users/enwon/OneDrive/Desktop/PLCY610/PS2/demo.csv")

##
## -- Column specification --------------------------------------------------------
## cols(
##   country = col_character(),
##   polity2 = col_double(),
##   gdp = col_double(),
##   regime = col_double(),
##   wealth = col_double()
## )

library(dplyr)
HW2<- as_tibble(demo)
```

Problem 1

Question 1: Let X ={0,6,8,12}and y ={10.97,23.90,19.34,18.29,24.32} 1a. Calculate x and y as well as s(x) and s(y)

1a. Calculate the mean of x and y as well as the standard deviation of x and y

ENWONGO EKAH

X= $\bar{x} = \frac{1}{n}\sum xi$

$\bar{x} = \frac{0 + 6 + 8 + 10 + 12}{5}$

$= \frac{36}{5}$

$\bar{x} = 7.2$

The mean for x =7.2

Now let's find the mean of Y

Y= $\bar{y} = \frac{1}{n}\sum xi$

$\bar{y} = \frac{10.97 + 23.90 + 19.34 + 18.29 + 24.32}{5} = \frac{96.82}{5}$

$\bar{y} = 19.36$

The mean of Y= 19.36

Now, The standard deviation of X and Y

Solution:

s= $\sqrt{\frac{1}{n-1}}\sum(x1 - \bar{x})^2$

s(x)= $\frac{(0-7.2)^2 + (6-7.2)^2 + (8-7.2)^2 + (10-7.2)^2 + (12-7.2)^2}{5-1}$

$= \frac{51.84 + 1.44 + 0.64 + 7.84 + 23.04}{4}$

$= \frac{84.8}{4}$

$= \sqrt{21.2}$

s(x) = 4.60

The standard deviation for x is 4.60.

Now, we calculate the standard deviation of Y.

Solution:

ENWONGO EKAH

$$s=\sqrt{\frac{1}{n-1}}\sum(yi-\overline{y})^2$$

$$\frac{(10.97-19.36)^2+(23.90-19.36)^2+(19.34-19.36)^2+(18.29-19.36)^2+(24.32-19.36)^2}{5-1}$$

$$s(y)=\frac{70.39+20.61+0.0004+1.1449+24.60}{4}$$

$$=\frac{116.74}{4}$$

$$=\sqrt{29.18}$$

s(y) = 5.40

The standard deviation of y=5.40

Question 1b: Calculate the correlation of X and Y

Solution:

$$r=\frac{1}{n-1}\sum\frac{xi-\bar{x}}{sx}*\frac{yi-\bar{y}}{sx}$$

$$\frac{1}{4}*\left(\frac{0-7.2}{4.6}\right)*\frac{10.97-19.36}{5.4}+$$

$$\left(\frac{6-7.2}{4.6}\right)*\left(\frac{23.90-19.36}{5.4}\right)+\left(\frac{8-7.2}{4.6}\right)*\left(\frac{19.34-19.36}{5.4}\right)+\frac{10-7.2}{4.6}\right)*\left(\frac{18.29-19.36}{5.4}\right)+($$

$$\frac{12-7.2}{4.6}\right)*\left(\frac{18.29-19.36}{5.4}\right)$$

Correlation (r)= 0.76

1c: Describe the direction and strength of the correlation.

Solution: The correlation is positive and strong.

Problem 2

ENWONGO EKAH

2a Cases/Units of Analysis

Solution: Countries are the units of observation.

2b. The Summary Statistics

Solution: Summary Statistics of Variables except country name; Mean, Standard deviation, median, and first and third quartiles for Polity2,gdp,regime,and wealth

Summary Statistics

Variables	Average	Standard Deviation	Median	First Quartile	Third Quartile
Polity 2	2.323809	7.3648010	5	-7	10
GDP	4832.295238	4945.5916849	2777	1182	6673
Regime Type	2.190476	0.8781015	2	1	3
Wealth	1.904762	0.7786059	2	1	3

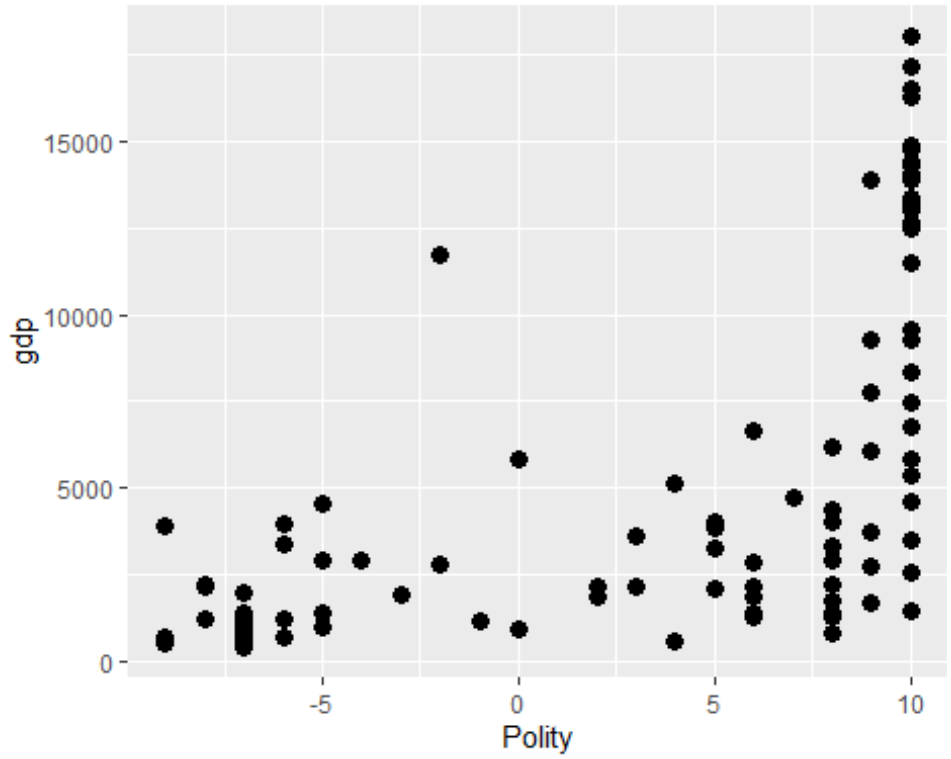
Code:

```
set1<-HW2 %>%
  select(polity2, gdp, regime, wealth)

Variables= c("Polity 2","GDP", "Regime Type", "Wealth")
Mean = c(mean(set1$polity2),mean(set1$gdp), mean(set1$regime), mean(set1$wealth))
Sd = c(sd(set1$polity2), sd(set1$gdp), sd(set1$regime), sd(set1$wealth))
Median = c(median(set1$polity2),median(set1$gdp),median(set1$regime),median(set1$wealth))
Quantile1 = c(quantile(set1$polity2, .25), quantile(set1$gdp, .25), quantile(set1$regime, .25),
quantile(set1$wealth, .25))
Quantile3 = c(quantile(set1$polity2, .75), quantile(set1$gdp, .75), quantile(set1$regime, .75),
quantile(set1$wealth, .75))
# dataframe
tab1.df<- data.frame(Variables,Mean, Sd, Median, Quantile1, Quantile3)
#check
is.data.frame(tab1.df)
```

```
#Change row and column names
colnames(tab1.df)<-c("Variables", "Average", "Standard Deviation", "Median", "First Quartile",
"Third Quartile")
library("knitr")
tabx<-kable(tab1.df, format = "simple", caption = "Summary Statistics")
#print table
#using lemon
library(lemon)

knit_print.table <- lemon_print
tabx
```

2c. Scatterplot of gdp and polity2

Solution: Yes there is a relationship. A statistical relationship between two variables is referred to as their correlation.



Code:

```
library(ggplot2)
HW2 %>%
ggplot(
    aes(x=polity2, y=gdp))+
geom_point(size=3) +
xlab('Polity')

ylab('GDP')

 labs(x="polity 2 score",y="GDP", title = "Figure 1:scatterplot of GDP/Capita on Polity Score")
```

2d. Find the Correlation Between gdp and polity2.

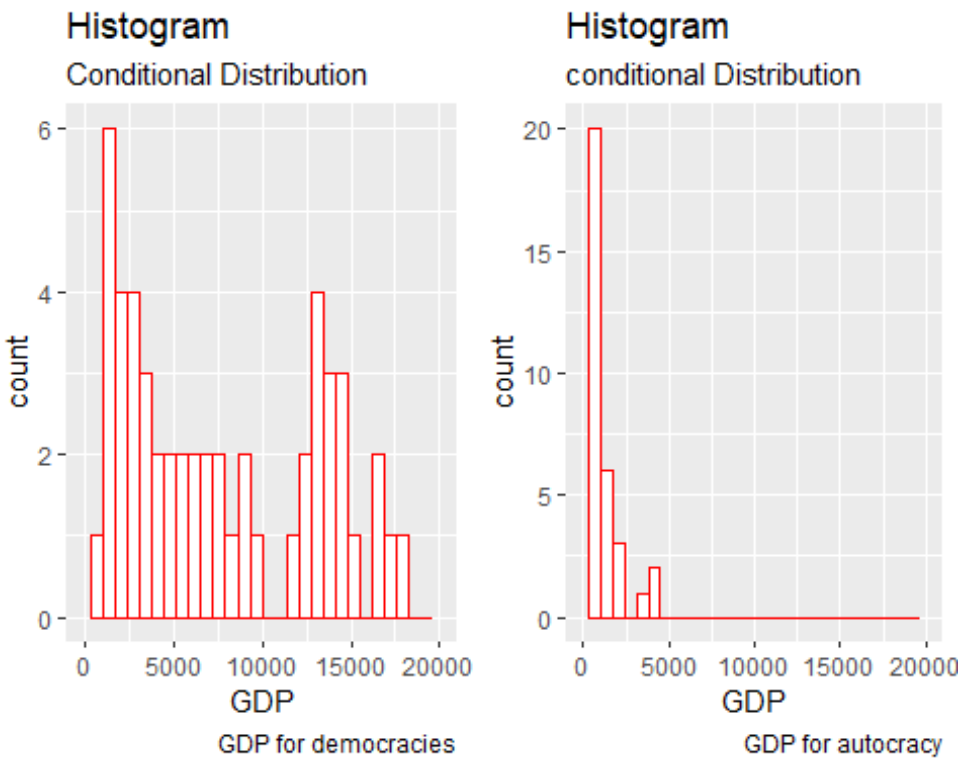
Solution: The correlation is 0.64. It indicates a strong positive correlation between gdp and polity2. It is not a test of causality because correlation is not equivalent to causation Because we can see a relationship between two variables does not necessarily mean that one causes the other.

Code:

```
cor1<-cor(HW2$gdp, HW2$polity2)
print(cor1)

## [1] 0.6359307
```

2e. Histogram of Conditional distribution of GDP for autocracies and Conditional distribution of GDP for democracy

Solution: The conditional distribution of gdp for democracies is bimodal and data seems to be spread more to the left side while gdp for autocracies is unimodal.



Code:

```
autocracy_HW2 = HW2[HW2$regime==1,]
democracy_HW2 = HW2[HW2$regime==3,]

autocracy1_HW2 =
HW2 %>%
filter(regime==1)
#Histograms

library(gridExtra)

plot1 = democracy_HW2 %>%
ggplot(aes(x=gdp)) +  # x = Column_name
geom_histogram(color="red", fill="white")+ xlim(0, 20000)+
labs(x="GDP",y="count",title="Histogram ",subtitle ="Conditional Distribution",caption="GDP for democracies")
```

```
plot2 = autocracy_HW2 %>%
ggplot(aes(x=gdp)) +  # x = Column_name
geom_histogram(color="red", fill="white")+ xlim(0, 20000)+
labs(x="GDP",y="count",title="Histogram",subtitle = "conditional Distribution",caption ="GDP for autocracy")


grid.arrange(plot1, plot2, ncol=2)
```

2f. Create a two- way table with frequencies for wealth and regime

Solution: Two Way Frequency Table for Wealth and regime

```
##       regime
## wealth  1  2  3 Sum
##    1   26  5  6  37
##    2    6 15 20  41
##    3    0  1 26  27
##    Sum 32 21 52 105
```

Code:

```
library(gt)
library(gtsummary)
library(ggplot2)
library(lemon)
tab1<-HW2  %>%
select(wealth,regime) %>%
table() %>%
  addmargins()
  kable.opts=list(caption='Table 2 : 2-way Table showing Joint and Marginal Frequencies for Regime Type and Wealth ')
knit_print.table <- lemon_print
tab1
```

2g. Create two more two-way tables for wealth and regime. First,present the joint distribution of regime and wealth. Next,present the conditional distributions of wealth for each regime type.

solution: Joint distribution of regime and wealth and the conditional distribution of wealth for each regime type

Characteristic	1	2	3	**Total**
Wealth				
1	26 (25%)	5 (4.8%)	6 (5.7%)	37 (35%)
2	6 (5.7%)	15 (14%)	20 (19%)	41 (39%)
3	0 (0%)	1 (1.0%)	26 (25%)	27 (26%)
Total	32 (30%)	21 (20%)	52 (50%)	105 (100%)

Code:

```
#Joint
library(ggplot2)
library(lemon)
pubtab2<- HW2 %>%
  tbl_cross(
    row = wealth,
    col = regime,
    percent = "cell" ) %>%
  as_kable()
pubtab2
```

```
#Conditional for wealth GIVEN regime- for conditional we want to divide by the condition (regime)
#regime is on columns
```

```
##       regime
## wealth      1         2         3
##      1 0.8125 0.23809524 0.1153846
##      2 0.1875 0.71428571 0.3846154
##      3 0.0000 0.04761905 0.5000000
```

Code:

```
prop.1<-with(HW2, table(wealth,regime)) %>% prop.table(margin=2)
prop.1
```

2h. What proportion of countries that are middle income anocracies?

Solution :14% Joint probability of two events occuring simultaneously.

2i. What is the marginal proportion of autocratic countries

Solution : 30% The marginal distribution is univariate. it examines the distribution of the categorical variables individually. The probability of an event in Variable A(regime) occuring regardless of an event on Variable B(wealth).

Code:

```
HW2 %>%
 count(regime) %>%
 mutate(marginalprobalities=n/sum(n)*100)
```

```
## # A tibble: 3 x 3
##   regime     n marginalprobalities
## *  <dbl> <int>               <dbl>
## 1      1    32                30.5
## 2      2    21                20
## 3      3    52                49.5
```

```
 percent = ("sum" )
```

2j. What is the conditional proportion of rich countries given that they are democratic?

Solution:50% We are finding the conditional distribution of wealth given regime type. In this case what proption of countries are rich because they are a democracy.

Question 3

3a. Generate 1000 draws of x

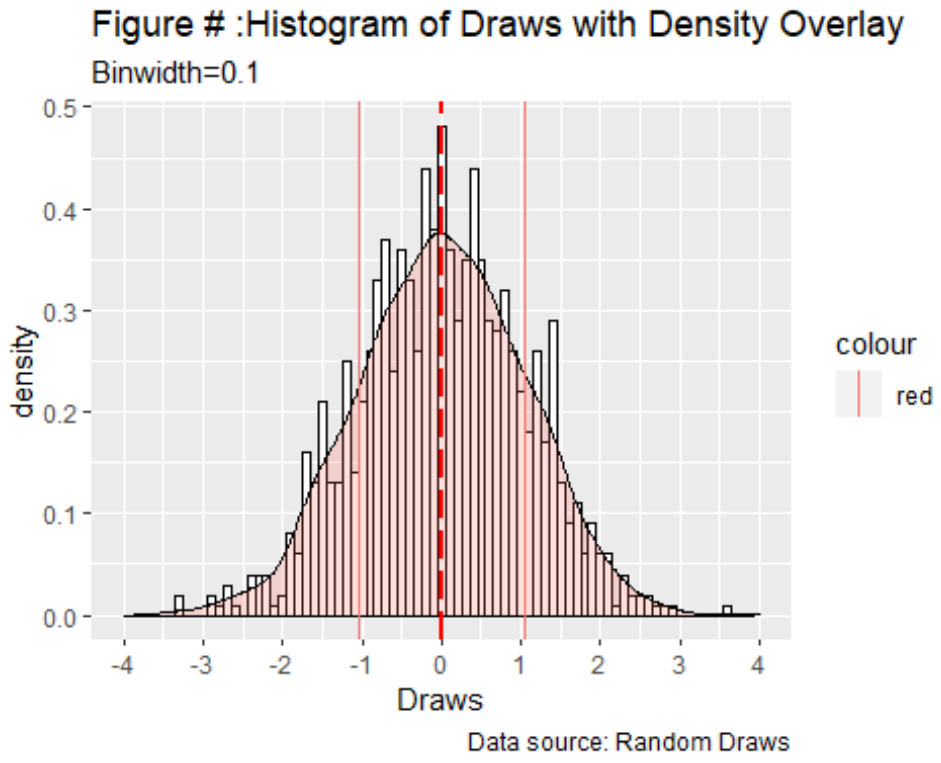
solution; Taking random draws of 1000

Code:

```
draws_1 <- rnorm(1000,0,1)
```

3b. Generate a density plot and describe the shape of the distribution

Answer; The distribution looks like a bell curve, symmetric distribution. (I could say it is not perfect but if we increase the number of samples then it would become more smoother). Yes, it is as expected because we are drawing sample from a standard normal distribution .



Code:

```
rand.df<-as.data.frame(draws_1)
draws.hist<-ggplot(rand.df ,aes(x=draws_1))+
geom_histogram( aes(y=..density..),binwidth = 0.1, color="black", fill="white") +
  scale_x_continuous(name = "Draws", n.breaks = 8, limits = c(-4,4))+
  geom_density(alpha = .2, fill="#FF6655")+
labs(title = "Figure # :Histogram of Draws with Density Overlay", subtitle = "Binwidth=0.1",
caption = "Data source: Random Draws")+
geom_vline(aes(xintercept = mean(draws_1)),
             colour = "red", linetype ="longdash", size = .8)+
  geom_vline(aes(xintercept= sd(draws_1), colour="red"))+
   geom_vline(aes(xintercept=(-1* sd(draws_1)), colour="red"))
draws.hist
```

3c. Calculate the mean and standard deviation. Are they exactly 0 and 1,respectively? if not,why not?

solution: Because we are drawing randomly from a standard normal distribution and not a population. The mean and standard deviation from a standard normal distribution is expected to be 0 and 1 respectively. Here, the mean of this sample is a random variable. Each sample I draw has its own mean and standard deviation value and each value is different. The mean ranging from -0 to 0 and the standard deviation approximately 1.

```
## [1] 0.01155188
```

```
## [1] 1.037894
```

Codes:

```
mean(draws_1)

sd(draws_1)
```

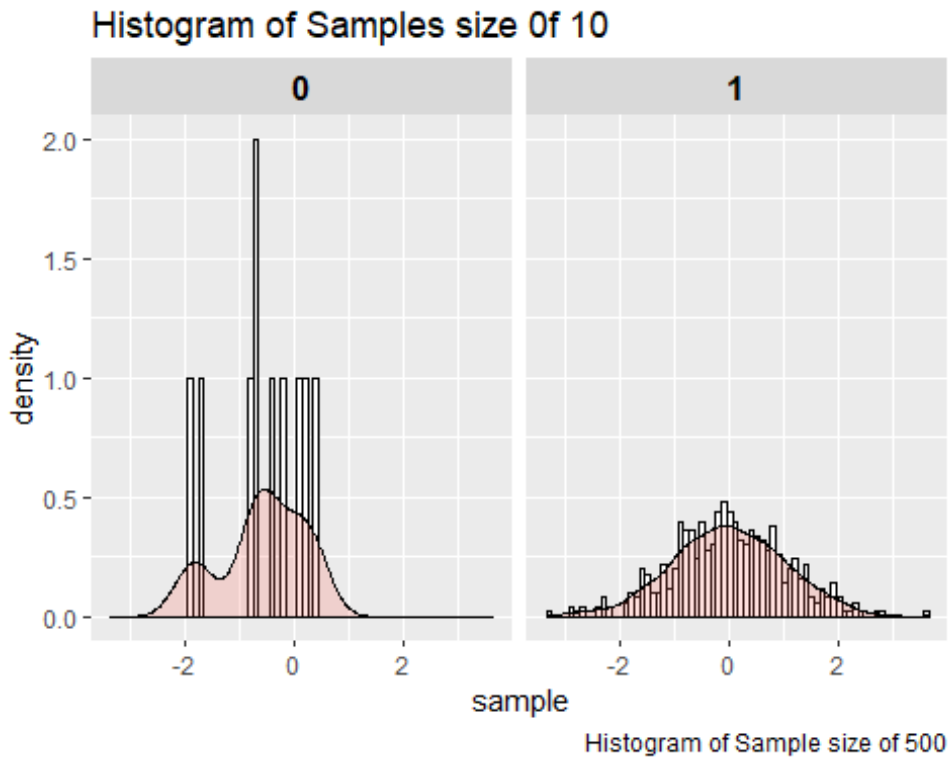
3d. From those 1,000 draws, draw random samples of size 10 and 500

solution: Sample draw of size 10 and 500

```
small = sample(draws_1,10)
large = sample(draws_1,500)
```

3e. Plot a histogram of each of the samples solution: Histogram of Sample draws of 10 and 500

```
overall_sample= c(small, large)
samp_size= ifelse(overall_sample==small,0,1)
newdat<-data.frame(overall_sample, samp_size)
ggplot(newdat,aes(x=overall_sample))+ labs(x="sample",y="density",title = "Histogram of
Samples size 0f 10",caption ="Histogram of Sample size of 500")+
 geom_histogram(aes(y=..density..),binwidth = 0.1, color="black", fill="white")+
  geom_density(alpha = .2, fill="#FF6655")+
  facet_wrap(~samp_size)+
            theme(strip.text.x = element_text(size = 12, color = "black", face = "bold"))
```



3f. What conclusions do you draw about the bias and efficiency of each sampling procedure(e.g samples of 10 vs 500)

Solution: Bias; Because both samples were randomly selected from standard normal distribution, they are unbiased. Efficiency; Larger samples are closely approximate the population. It is less of an inference if the sample is large. Choosing 10 draws to represent an entire population even if they are chosen completely at random will often result a sample that is very unrepresentative. Smaller samples tend to have larger standard errors and larger samples tend to have smaller standard errors. It is always better to strive to get a sample as large as possible.

Problem 4

4a. What is the population that the article is alluding to?

Solution; The United States working Population.

4b. What is the independent and dependent variable?

Solution: The independent variable is Education and dependent variable is wage/income. It is a positive relationship.

4c. What kind of data is mobilized by the author? Solution: The author used observational data.

4d. Does the evidence presented indicate a causal relationship?

Solution: No, the evidence presented does not indicate a causal relationship. There is likely to be some alternative explanation for a relationship between the independent variable(education)and the dependent variable(wage) because we did not randomly assign participants to experimental conditions.There may be a third variable,that may affect the outcome. The explanatory variable is confounded with lurking variables. It could be demand and supply,discrimination or even cost of living.

The author used existing data collected by other people that were made available for use. Some of those sources may be reliable or not. Was the data sample biased? We don't know if the research was rushed. Did the author use intercoder reliability to check the data? What influenced this research. There could also be personal bias and poor information. The author could have missed a variable.

4e. Could this question be addressed in an RCT format? If so, how? What might be some ethical issues involved?

Solution: Everything is always randomized in statistics. Yes, this question can be addressed in an RCT format. There is no inherent format by the person creating the experiment randomly.The simplest use of matching is the matched pair design,which compares just two treatments.The subjects are matched in pairs. The idea is that matched subjects are more similar than unmatched subjects, so that comparing responses within a number of pairs is more efficient than comparing the responses of groups of randomly assigned subject. For RCT to be ethically possible,many different conditions must be met. They include social values, comprehensive informed consent,conflicts of interest, Institutional

Review Board (IRB)approval and protection of participants right. Another ethical issue could relate to the information you might uncover and actions you should take.