Rensselaer

why not change the world?®

Lab exercises: installing R/ Rstudio, beginning to work with data: Data Frames, populations, filtering, distributions...

Ahmed Eleish

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960 MGMT4962/6962/ BCBP4962

Group 1, Lab 01, September 13th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Add/Drop Deadline: Todayyyy!!!!!

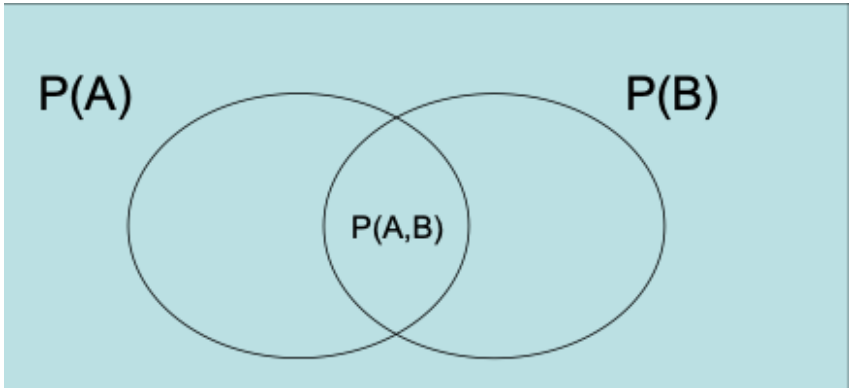
Probability

Conditional Probability: *probability of event A occurring, given that event B occurred.*

$P(A|B) = \frac{P(A,B)}{P(B)}$ ** = *Probability of A, given B ; P(B)>0*

~ Probability of effect given cause

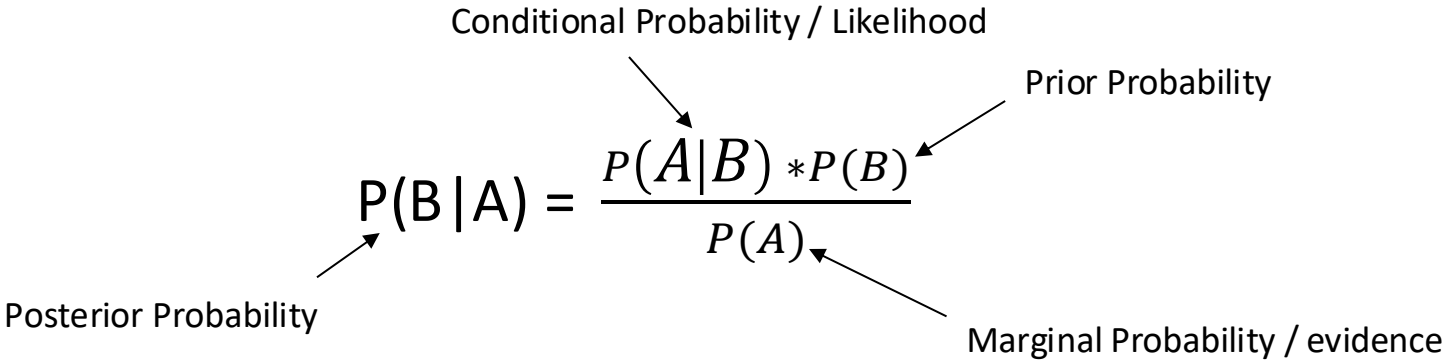
*** P(A,B) = P(A) * P(B)*



https://en.wikipedia.org/wiki/Conditional_probability

Bayes Theorem

- The relationship between conditional probabilities, P(B|A) and P(A|B) can be expressed using the Bayes Theorem.



~ Probability of cause given effect

Reference: https://en.wikipedia.org/wiki/Bayes%27_theorem

The Complement Rule

$P(\text{not } A) = 1 - P(A)$

$P(B|\text{not } A) = 1 - P(\text{not } B|\text{not } A)$

Hypothesis

1. Write the original claim and identify whether it is the null hypothesis or the alternative hypothesis.
2. Write the null and alternative hypotheses. Use the alternative hypothesis to identify the type of test.
3. Write down all information from the problem.
4. Find the critical value using the tables
5. Compute the test statistic
6. Make a decision to reject or fail to reject the null hypothesis. A figure showing the critical value and test statistic may be useful.
7. Write the conclusion.

Null and Alternate Hypotheses

- What are you exploring?
- H0 - null
- H1 – alternative
- If a given claim contains equality, or a statement of no change from the given or accepted condition, then it is the null hypothesis, otherwise, if it represents change, it is the alternative hypothesis.

- Stock market performance / trends versus unusual events (crash/ boom):
- Election results are predictable from exit polls
- It never snows in Troy in January
- Students will attend their scheduled classes

Lab tasks

- Installations
- Data Frames in R
- Exploring data and their distributions
- Fitting distributions

Objectives for today

- Get familiar with:
  - R (Data Frames)
  - Populations / Filtering
  - Distributions
  - Fitting

Gnu R

- http://lib.stat.cmu.edu/R/CRAN/ - load this first
- http://cran.r-project.org/doc/manuals/
- http://cran.r-project.org/doc/manuals/R-lang.html
- R Studio – (see R-intro.html too)

https://www.rstudio.com/products/rstudio/ (desktop version)

- Manuals - Libraries – at the command line – library(), or select the packages tab, and check/ uncheck as needed

Creating Data Frames in R

Using the data.frames() in R, we can create the data frames Read the documentation for data.frame in Rstudio,
Using the **help(data.frame)**

```
# Creating a dataframe
# Example: RPI Weather dataframe.

days <- c('Mon', 'Tue','Wed','Thur','Fri','Sat','Sun') #  days
temp <- c(28,30.5,32,31.2,29.3,27.9,26.4) # Temperature in F' during the winter :)
snowed <- c('T','T','F','F','T','T','F') # Snowed on that day: T = TRUE, F= FALSE
help("data.frame")
RPI_Weather_Week <- data.frame(days,temp,snowed) # creating the dataframe using the data.frame() function

RPI_Weather_Week
head(RPI_Weather_Week) # head of the data frame, NOTE: it will show only 6 rows, usually head() function shows the
# first 6 rows of the dataframe, here we have only 7 rows in our dataframe.

str(RPI_Weather_Week) # we can take a look at the structure of the dataframe using the str() function.

summary(RPI_Weather_Week) # summary of the dataframe using the summary() function
```

Data frames

```
RPI_Weather_Week[1,] # showing the 1st row and all the columns
RPI_Weather_Week[,1] # showiing the 1st coulmn and all the rows

RPI_Weather_Week[,'snowed']
RPI_Weather_Week[,'days']
RPI_Weather_Week[,'temp']
RPI_Weather_Week[1:5,c("days","temp")]
RPI_Weather_Week$temp
subset(RPI_Weather_Week,subset=snowed==TRUE)

sorted.snowed <- order(RPI_Weather_Week['snowed'])
sorted.snowed
RPI_Weather_Week[sorted.snowed,]
```

```r
# RPI_Weather_Week[descending_snowed,]
dec.snow <- order(-RPI_Weather_Week$temp)
dec.snow
# Creating Dataframes
# creating an empty dataframe
empty.DataFrame <- data.frame()
v1 <- 1:10
v1
letters
v2 <- letters[1:10]
df <- data.frame(col.name.1 = v1,col.name.2 = v2)
df
# importing data and exporting data
# writing to a CSV file:
write.csv(df,file = 'saved_df1.csv')
df2 <- read.csv('saved_df1.csv')
df2
```

Table: Matlab/R/scipy-numpy

http://hyperpolyglot.org/numerical-analysis

Exercises – importing data

- Rstudio
  - read in csv file (two ways to do this) - filename.csv
  - Read in excel file (directly or by csv convert) - filename.xls
  - Plot some variables
  - Commonalities among them?
- Also for other datasets, enter these in the R command window panel or cmd line

> data()
> help(data)

Files

Here -> https://rpi.box.com/s/2seknncutn7juxq95pvsmdk9l9wkopxm

- epi2024results06022024.csv
- Get the data read in (in e.g. use “EPI_data” for the object (in R))

> EPI_data <- read.csv(“<path>/epi_2024_results_DAF24.csv”)
Note: replace default data frame name – cannot start with numbers! Munging has begun! (ugh)

# Note: replace <path> with either a directory path or use:
setwd(“<path>”)

> View(EPI_data)

Tips (in R)

> attach(EPI_data) # sets the ‘default’ object

> EPI.new # prints out values EPI_data$EPI.new

```
 [1] "30.7" "52.1" "41.9" "39.7" "55.5" "46.8" "44.7" "63"   "69"   "40.4" "56"   "35.9"
[13] "27.8" "53.1" "58.1" "66.7" "47.4" "37.4" "43.3" "44.9" "45.6" "49"   "53"   "48.5"
[25] "56.3" "41.5" "33"   "37.9" "31"   "38.1" "61.1" "38.3" "35.2" "50"   "35.5" "49.4"
[37] "37.9" "55.5" "42.5" "62.6" "52.3" "54"   "65.6" "39"   "67.9" "32.2" "49.2" "47.6"
[49] "51.2" "43.8" "41.5" "41.6" "28.6" "75.3" "38.5" "35.8" "45.8" "73.7" "67.1" "53.1"
```

> tf <- is.na(EPI.new) # records True values if the value is NA

> E <- EPI.new[!tf] # filters out NA values, new array

Data Cleaning, etc.

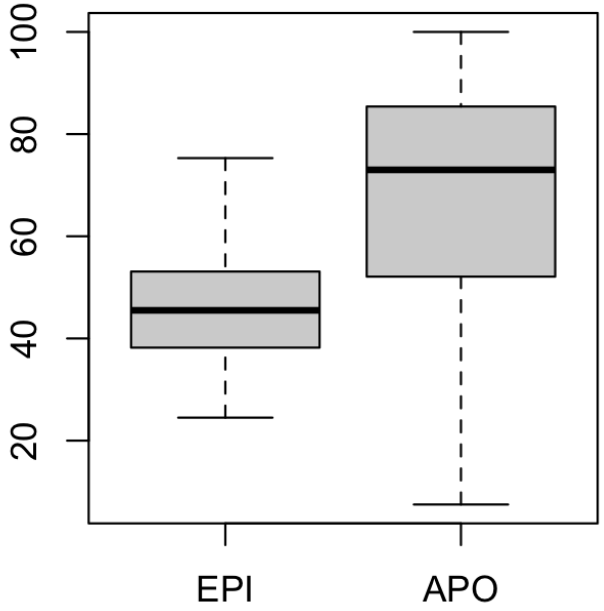
- Bad values, outliers, corrupted entries, thresholds ...
- Noise reduction – low-pass filtering, binning
- REMEMBER: when you clean you MUST record what you did (and why) and save copies of data/code pre- and post-operations...

Exercise 1: exploring the distribution

```
> summary(EPI.new) # stats
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 32.10 48.60 59.20 58.37 67.60 93.50 68
> fivenum(EPI.new,na.rm=TRUE) [1] 32.1 48.6 59.2 67.6 93.5
> stem(EPI.new) # stem and leaf plot
> hist(EPI.new)
> hist(EPI.new, seq(20., 80., 1.0), prob=TRUE)
> lines(density(EPI.new,na.rm=TRUE,bw=1.)) # or try bw="SJ"
> rug(EPI.new)
#Use help(<command>), e.g. > help(stem)
```

Comparing distributions

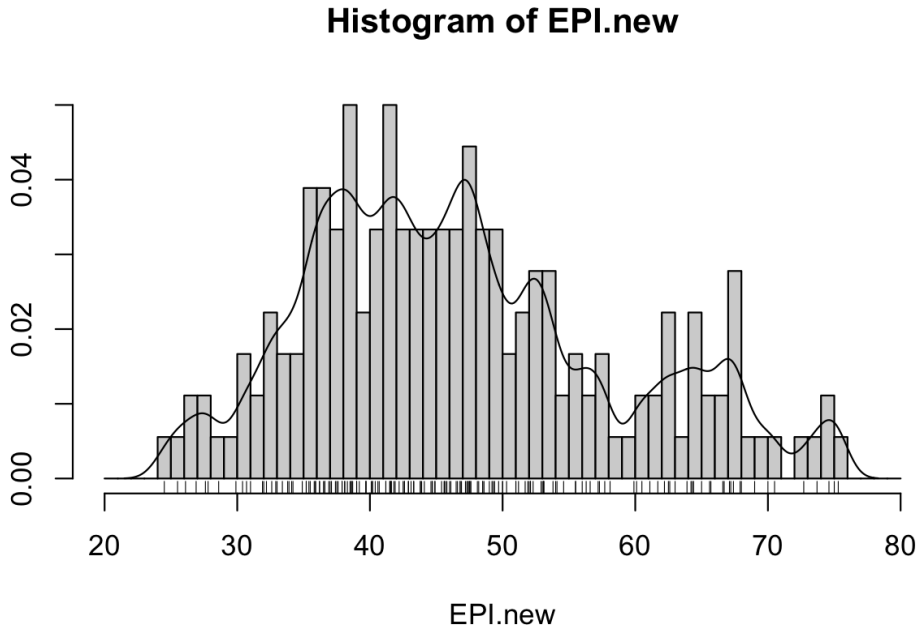
> boxplot(EPI.new, APO.new)



> hist(EPI.new, seq(20., 80., 1.0), prob=TRUE)

> lines (density(EPI.new,na.rm=TR UE,bw=1.))

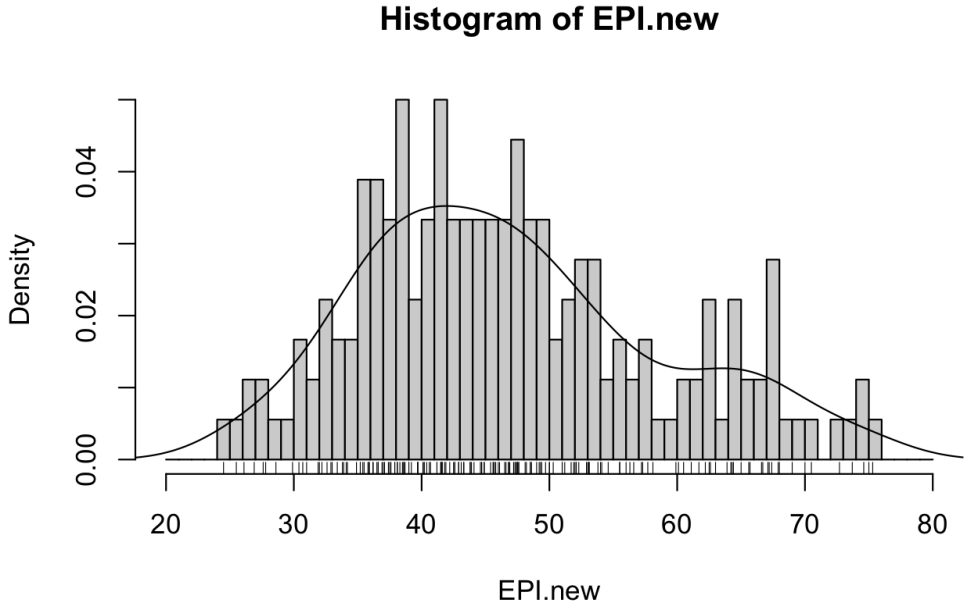
> rug(EPI.new)



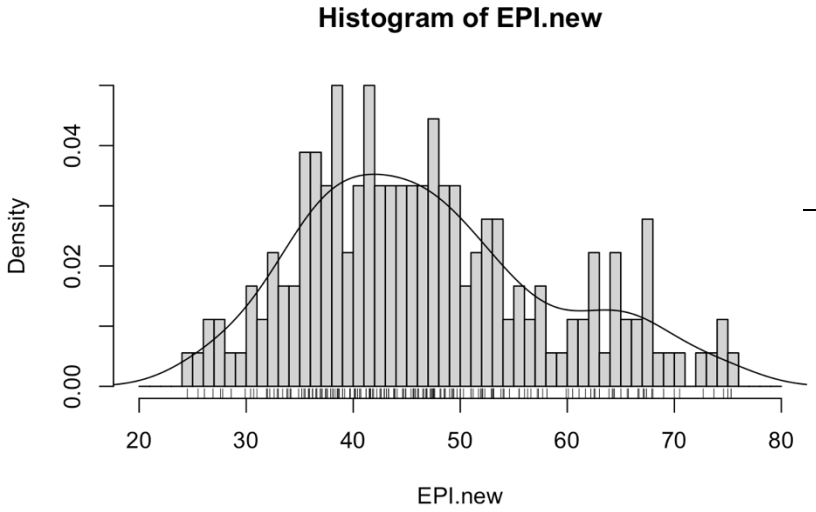
> hist(EPI.new, seq(20., 80., 1.0), prob=TRUE)

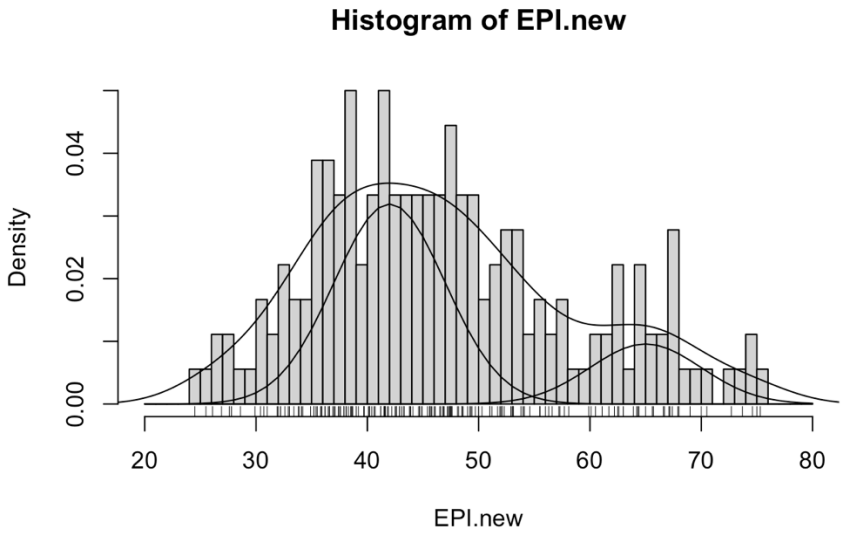
> lines (density(EPI.new,na.rm=TRUE,bw="SJ"))

> rug(EPI.new)

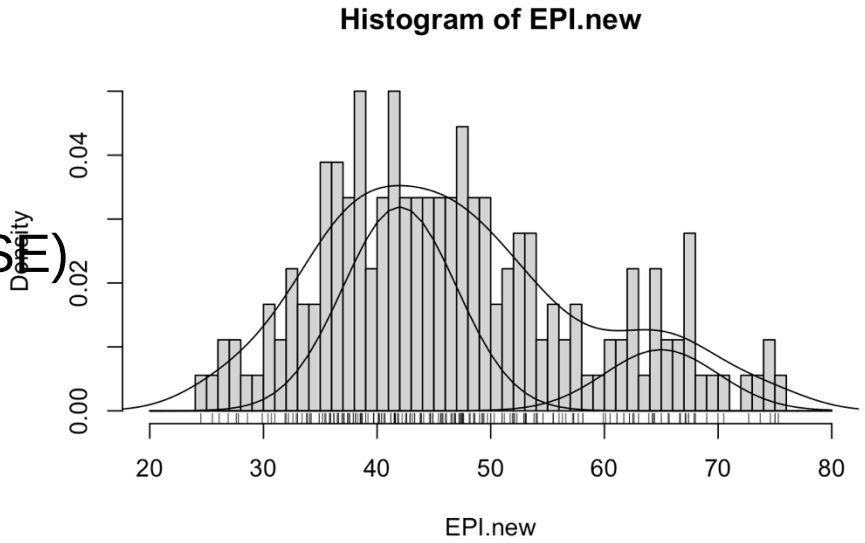


Why are histograms so unsatisfying?





```
> x<-seq(20,80,1)

> q<- dnorm(x,mean=42, sd=5,log=FALSE)

> lines(x,q)

> lines(x,.4*q)

> q<-dnorm(x,mean=65, sd=5,log=FALSE)

> lines(x,.12*q)
```



Exercise 2: fitting a distribution beyond histograms

- Cumulative density function?

```
> plot(ecdf(EPI.new), do.points=FALSE, verticals=TRUE)
```

- Quantile-Quantile?

```
> qqnorm(EPI.new); qqline(EPI.new)
```

- Make a Q-Q plot against the generating distribution by:

```
> qqplot(rnorm(250), EPI.new, xlab = "Q-Q plot for norm dsn")
> qqline(EPI.new)

> qqplot(rt(250, df = 5), EPI.new, xlab = "Q-Q plot for t dsn")
> qqline(EPI.new)
```

Exercise 2a: fitting a distribution

- Your exercise: do the same exploration and fitting for another 2 variables in the EPI dataset, i.e. primary variables (APO, WRS, etc.)

Push your Lab code to Github

- Please create a folder for Lab1 in your Github repository and push your code/plots.

Next Class:

September 17th - Introduction to Analytic Methods, Types of Data Mining for Analytics, Data filtering, hypothesis exploration, visual analysis

Thanks!

Enjoy the carnival today!!!