– A Faculty Development Program on Data Analytics in R –

25-29 May 2020

Sponsored by

All India Council for Technical Education

and

AICTE Training and Learning Academy

Organized by

Indian Institute of Information Technology Vadodara

`https://atal-da.iiitvadodara.ac.in`

{atal-ai@iiitvadodara.ac.in}

Data Analytics in R

A Five Day Faculty Development Program
Hands-on

Pratik Shah

26 May 2020

https://atal-da.iiitvadodara.ac.in

Warm Up

Warm up Exercise

Example (P-value)

Suppose that an engineer encounters data from a manufacturing process in which 100 items are observed and 10 are found to be defective. It is expected and anticipated that occasionally there will be defective items. However, it has been determined that in the long run, the company can only tolerate 5% defects. What would be the engineer's reaction?

Warm up Exercise Cont.I

[-R-] (First-Few-R-Commands)

```
> ?  ‘‘string’’ or help(‘‘string’’)
> help(‘‘pbinom’’)
> help.search(‘‘dbinom’’)
> pe=0.05
> dbinom(10,100,pe)
> n=100; x=10;
> X<-seq(x,n,1)
> Px<-dbinom(X,n,pe)
> plot(dbinom(seq(0,n,1),n,pe))
> P=1-sum(dbinom(X,n,pe))
> P=1-pbinom(x-1,n,pe)
```

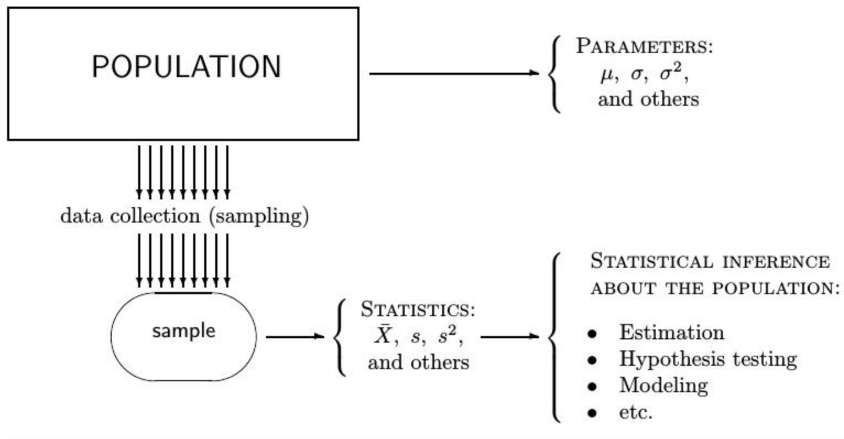
An Interesting Exercise

Example (Conditional Probability)

Given my recorded arrival time-stamps (workspace folder), estimate the probability of my being present in the institute at $T = t$ time instance. Assume that if I have arrived, I will be available in my office, when should you come to my office to meet me to ensure that we meet with 0.9 probability.

Suppose you are able to observe the light in my office. Assume that if I am in my office the probability that the light is on is 0.90 and if I am not available in my office the probability that the light is on is 0.01. Derive the conditional probability of me being available in my office given that the light in my office is on (off).

Population Parameters and Sample Statistics



Why Probability and Statistics?

Probability and Statistical Inference

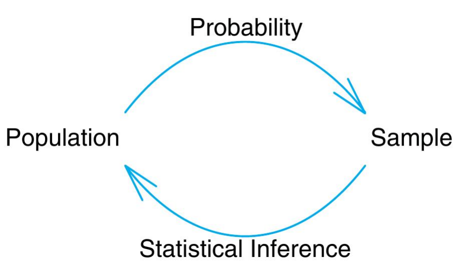
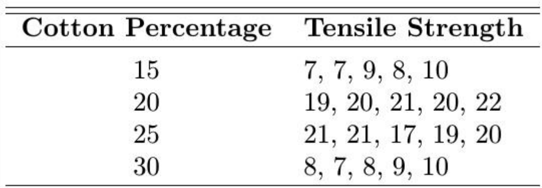


Figure 1: (1) The sample along with inferential statistics allows us to draw conclusions about the population, with inferential statistics making clear use of elements of probability. (2) Elements in probability allow us to draw conclusions about characteristics of hypothetical data taken from the population, based on known feature of the population.

Not matter how bad your data is, there is no harm in looking at it!

Example

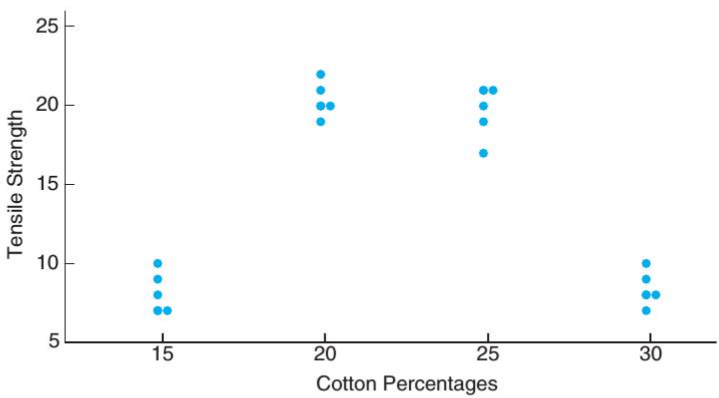
Consider the data of an experiment where cloth specimen that contain various percentages of cotton are produced



Cotton Percentage	Tensile Strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

Before you do anything with a data set, LOOK AT IT!

Scatter Plot



Some-More-R-Commands

[-R-]

```
> x<-read.table(‘‘cotton_strength.txt’’)
> plot(x$V1,x$V2)
> summary(x)
> hist(x$V2)
> x[7,2]
> u<- x$V2
> v <- x$V1; v2 <- v*v
> qm <- lm(u ∼ v+v2); summary(qm)
> vs <- seq(5,40,1);
> us <- predict(qm,list(v=vs, v2=vs*vs))
> plot(v,u); lines(vs,us)
```

Why R?

Data Frames

Time Series in Data Frame

[-R-] (Generating and Visualizing TS)

```
> n <- 500; t<-seq(1,n,1);
> o1 <- rnorm(n); o2 <- 2+rnorm(n);
> tsd <-
data.frame(id=c(rep(1,n),rep(2,n)),day=c(t,t),ts=c(o1,o2)
> library(ggplot2)
> ggplot(tsd,aes(x=day,y=ts,group=id,color=id))+
geom_line()
> ggplot(tsd,aes(day,ts,color=id)+ geom_line()+
facet_wrap(vars(id))
> ggplot(tsd,aes(id,ts,group=id))+ geom_boxplot()
> ggplot(tsd,aes(id,ts,color=id,group=id))+
geom_boxplot()+ geom_jitter(alpha=0.3)
```

Time Series Cont..

[-R-] (Manipulating TS)

```
> str(tsd)
> library(tidyr)
> tsd_spread <- spread(tsd,key=id,value=ts)
> str(tsd_spread)
> tsd_gather <-
gather(tsd_spread,key=IdNo,value=ME,-t)
> str(tsd_gather)
```

Data Summary in R

- Five point summary

Data Summary in R

- Five point summary
- Box plot

Data Summary in R

- Five point summary
- Box plot
- Stem-leaf plot

Data Summary in R

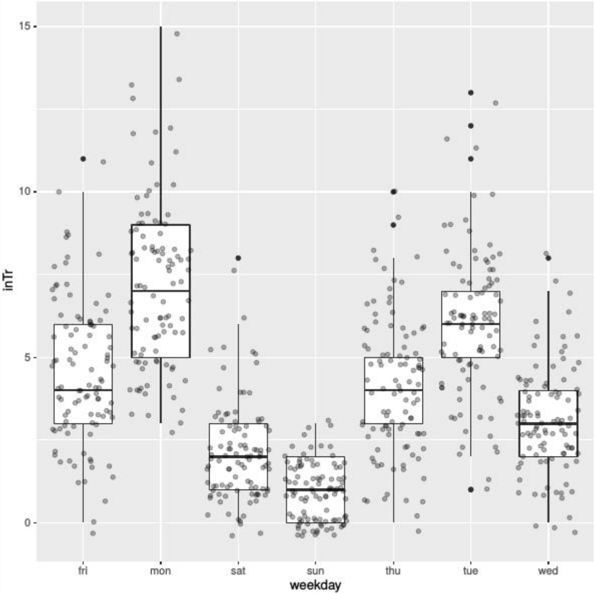
- Five point summary
- Box plot
- Stem-leaf plot
- Probability plot, histogram

How to get hold of Summary

[-R-] (Seeing is believing)

```
> library(tidyr)
> library(ggplot2)
> x<-read.csv("internet_traffic.csv")
> xt<-gather(x,key=weekday,value=inTr,-X)
> ggplot(xt,mapping=aes(x=weekday,y=inTr))+
geom_boxplot()+geom_jitter(alpha=.3)
>library(dplyr)
>summarize(group_by(xt,weekday),mean(inTr))
```

Internet Traffic on Weekdays



Acknowledgement

These slides are prepared primarily based on the following books, The art of R Programming by Matloff [Matloff, 2011], Probability and Statistics for Engineers and Scientists by Walpole et al. [Walpole et al., 2007], and Probability and Statistics by Michael Baron [Baron, 2013].

The best way to begin exploration is to **read a good book** and **not** Google!

Readings

Baron, M. (2013).
Probability and Statistics for Computer Scientists, Second Edition.
Chapman & Hall/CRC, 2nd edition.

Matloff, N. (2011).
The Art of R Programming: A Tour of Statistical Software Design.
No Starch Press, San Francisco, CA, USA, 1st edition.

Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. (2007).
Probability & statistics for engineers and scientists.
Pearson Education, Upper Saddle River, 8th edition.

Questions?