Probability and Statistics

Radjesvarane ALEXANDRE

November 12, 2016

Chapter 1

Probability Spaces

Chapter 2

Random Variables

Chapter 3

2d Random Vectors

Chapter 4

Limit Theorems

Chapter 5

Introduction to descriptive statistics

5.1 Generalities around Statistics

The main goal is: from given data, try to learn, organize and use these data in order to make decision [which is also coined as inferential statistics] from uncertainty.

Example of statistical problems:

- A pharmaceutical company wants to know if a new drug is superior to other existing drugs
- Is there any relationship between GPA of a country and the height of people.
- Is there any relationship between amount of salary and happiness?

Basic Definitions

- Data: any recorded event
- Information: any acquired data
- Knowledge: useful data
- Population: set of all measurements of interest [for example voters, students of USTH, ...]
- Sample: a subset of measurements selected from the given population
- Variable: a property of an individual population unit (height of students, ...)

Example 5.1.1

We carry out the measurement of all heights of the 70 students from USTH attending Prob and Stats lectures.

Thus we have 70 height values x_1, ..., x_{70} in cm [these are the observations].

The statistical sample is simply these 70 values.

The statistical population would be for example the total height values of all USTH students (not only those attending Prob and Stats lectures).

The observational unit would be a typical student.

5.2 Some Data presentation

Frequency distribution

Example: if we have the statistical data

$$\begin{array}{ccccc} 3 & 2 & 2 & 3 & 2 \\ 4 & 4 & 1 & 2 & 2 \\ 4 & 3 & 2 & 0 & 2 \\ 2 & 1 & 3 & 3 & 1 \end{array}$$

then the frequency distribution is

$$X\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 3 & 8 & 5 & 3 \end{pmatrix}$$

We can consider frequency by class of data. In general a frequency distribution looks as:

$$X\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ f_1 & f_2 & f_3 & \cdots & f_n \end{pmatrix}$$

We may also consider relative frequency (that is frequency divided by the total frequency). Then the mode is the value of the piece of data with the greatest frequency.

A way to represent graphically statistical data is by using a pie chart.
We start from a given (ungrouped) frequency distribution:

$$X\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ f_1 & f_2 & f_3 & \cdots & f_n \end{pmatrix}$$

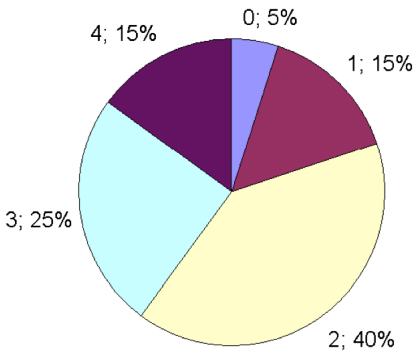
Then a pie chart is a disk divided into n circular sectors S_i with area equal to $f_i/F \times 100$ (where F is the total frequency) percents of the disk area.

Example 5.2.1

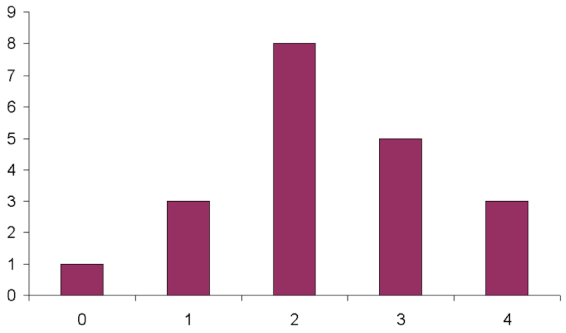
Consider the frequency distribution

$$X\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 3 & 8 & 5 & 3 \end{pmatrix}$$

then the disk is divided into 5 sectors with areas equal to 5%, 15%, 40%, 25%, 15% *of the total area.*

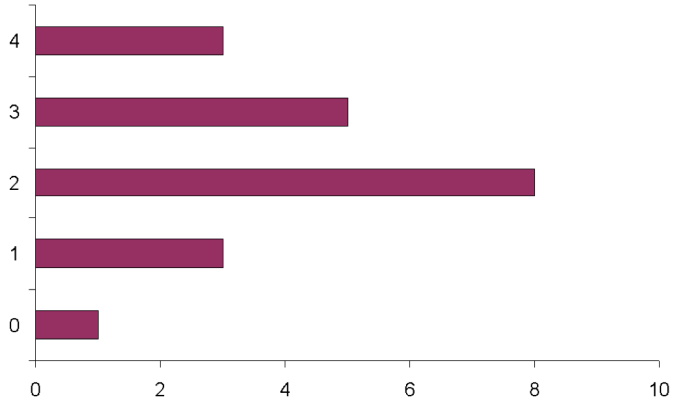


Another graphical way is a bar chart: this is a sequence of n rectangles of heights f_i. With the same previous example, we have



We can also introduce a stem and leaf diagram of this (ungrouped) frequency distribution: this is a set of n rectangles whose basis are equal and with heights being f_i but along the y axis.

For example with the same previous example, we have the following stem and leaf diagram:



Another way to present data is related to histograms (frequency and relative frequency distributions).

It is more simple to have to have in mind an example:

Weight Loss Data

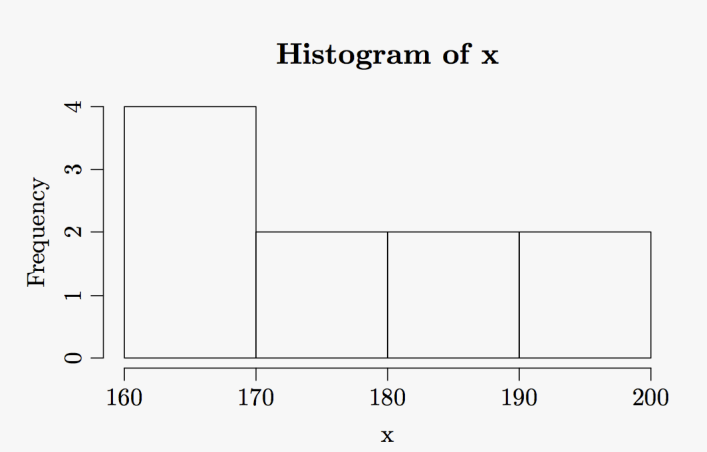
20.5	19.5	15.6	24.1	9.9
15.4	12.7	5.4	17.0	28.6
16.9	7.8	23.3	11.8	18.4
13.4	14.3	19.2	9.2	16.8
8.8	22.1	20.8	12.6	15.9

Weight Loss Data

class	boundaries	tally	class freq, f	rel. freq, f/n
1	5.0-9.0-		3	3/25 (.12)
2	9.0-13.0-		5	5/25 (.20)
3	13.0-17.0-		7	7/25 (.28)
4	17.0-21.0-		6	6/25 (.24)
5	21.0-25.0-		3	3/25 (.12)
6	25.0-29.0		1	1/25 (.04)
Totals			25	1.00

Let us set
- k for the number of classes (or bins)
- max for the largest measurement
- min for the smallest measurement
- n for the sample size
- ω for the class width usually increases with the number of data.

Another example is based on previous ones (the students). Using R software, we have



Here: we have equidistant class widths (same width for all classes). Moreover, we have just the number of individuals in each class. Wrt to density, we have

See below for other examples.
Cumulative distributions

$$F_n(x) = \frac{card\{x_i \leq x}{n}$$

5.3 Some statistical measurements

Given a set of n observations x_1, ..., x_n, the sample mean is the first key measure, giving the gravity center of these data. By definition

Definition 5.3.1

Given a set of n observations x_1, ..., x_n, the sample mean is defined by

$$\bar{x} \equiv \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Thus out of a population, if the set of observations is a meaningful sample, this number $\bar{x}$ will give an estimate of the population mean value.

Another measurement number is given by the median, wich also indicates the center of the data, but might be more useful in case of extreme values for example.

Definition 5.3.2

Given a set of n observations x_1, ..., x_n, first rearrange them from the smallest to the largest to get a new arranged set $x_{(1)}, ...x_{(n)}$. Then

- If n is odd, the sample median is the observation in position $\frac{n+1}{2}$

$$median = x_{(\frac{n+1}{2})}$$

- If n is even, the sample median is an average

$$median = x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}$$

The mode is the value of x (observation) that occurs with the greatest frequency.

Example 5.3.1

Students heights: suppose that we have a random sample of 10 students among you:

168 161 167 179 184 166 198 187 191 179

Then we find that

$$\bar{x} = 178$$

For the sample median, rearranging and because 10 is an even number, we find that the median is 179.

If we remove from the sample the 198 cm person, that is corresponding to $n = 9$, we find

$$\bar{x} = 175.78$$

while we get that the sample median is 179.

Note that the mean changes a lot, which is not the case of the median, which is a more robust quantity.

The median is the point "dividing" the data into two halves. We can generalize the number of parts and introduce the notion of percentiles or quantiles.

Definition 5.3.3

Given a set of n observations x_1, ..., x_n, the quantile of order p or the $100p$-th percentile (with p given less than 1) is defined as follows:

- first rearrange them from the smallest to the largest to get a new arranged set $x_{(1)}, ...x_{(n)}$

- then compute pn.

- if pn is an integer, then

$$p\text{-th quantile} = (x_{(np)} + x_{(np+1)}))/2$$

- if pn is not an integer, then

$$p\text{-th quantile} = x_{[np]}$$

where $[np]$ is the smallest integer larger than np.

Most used are the quartiles, splitting the data into quarters:

$$0,25\ ,0,50,\ 0,75,\ 1$$

The 0% percentile is the the smallest observation while the 100% percentile is the largest boservation.

Definition 5.3.4

We set

$$Q_1 = \textit{ lower quartile } = 0.25\ \textit{quantile} = \ 25\%\textit{percentile}$$

$$Q_2 = \textit{ median } = 0.50\ \textit{quantile} = \ 50\%\textit{percentile}$$

$$Q_3 = \textit{ upper quartile } = 0.75\ \textit{quantile} = \ 75\%\textit{percentile}$$

Example 5.3.2

Wo go back to the previous example with $n = 10$. *We have seen that* $Q_2 = 179$. *We want to compute* Q_1 *and* Q_3.

Here $p = 0.25$. *Then* $np = 2.5$.

Then $Q_1 = x_{([2.5])} = x_{(3)} = 167$. *For the upper quartile, since* $n.0.75 = 7.5$, *we get that* $Q_3 = x_{([7.5])} = x_{(8)} = 187$.

We have also that the 0% *percentile is the min, that is* 161, *and the* 100% *percentile is the max, that is* 198.

For the 0.10 *quantile, we have as* $p = 0.10$ *that* $np = 1$ *and then the* 0.10 *quantile is given by the average between* $x_{(1)}$ *and* $x_{(2)}$, *that is* 163.5.

The above notions measure the mean or classe means. We wish now to introduce measures of variability wrt to these quantities: that is, how far are from these quantities?

Definition 5.3.5

Given a set of n *observations* x_1, ..., x_n. *The sample variance is the number*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The sample standard deviation is simply s.
The coefficient of variation is

$$V = \frac{s}{\bar{x}}.100$$

Caution: In the definition of s^2, usuall $n-1$ is replaced by n. We took the above definition to fit the use of R software. There is also a deep mathematical explanation that we will not explain here.

Note that

$$s^2 = \frac{1}{n-1}[\sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2]$$

Interpretation: using Chebyshev's inequality, and regardless of the precise shape of the data, given a number $k \geq 1$, and a set of n observations x_1, ..., x_n, at least $1 - \frac{1}{k^2}$ of the measurements lie within k standard deviations of their mean sample.

Example 5.3.3

If we have a set of grades with $\bar{x} = 75$ and $s = 6$. Then:
- if $k = 1$, at least 0% of all grades lie in $[69, 81]$;
- if $k = 2$, at least 75% of all grades lie in $[63, 87]$;
- if $k = 3$, at least 88% of all grades lie in $[57, 93]$.

Finally

Definition 5.3.6

Given a set of n observations x_1, ..., x_n. The Range of the data is

$$Range = max - min$$

The Inter Quartile Range (IQR) is the middle 50% of data

$$IQR = Q_3 - Q_1$$

Example 5.3.4

We go back again to our previous data. We find that the sample variance is

$$s^2 = 149.1111$$

while the sample standard deviation is

$$s = 12.21$$

which can be interpreted as follows: students are on average around 12 cm away from the mean heigth of 178 cm.

One finds also that $Range = 37$ and $IQR = 20$. So 50% of students in this sample lie within 20 cm.

Comparing two sets of data

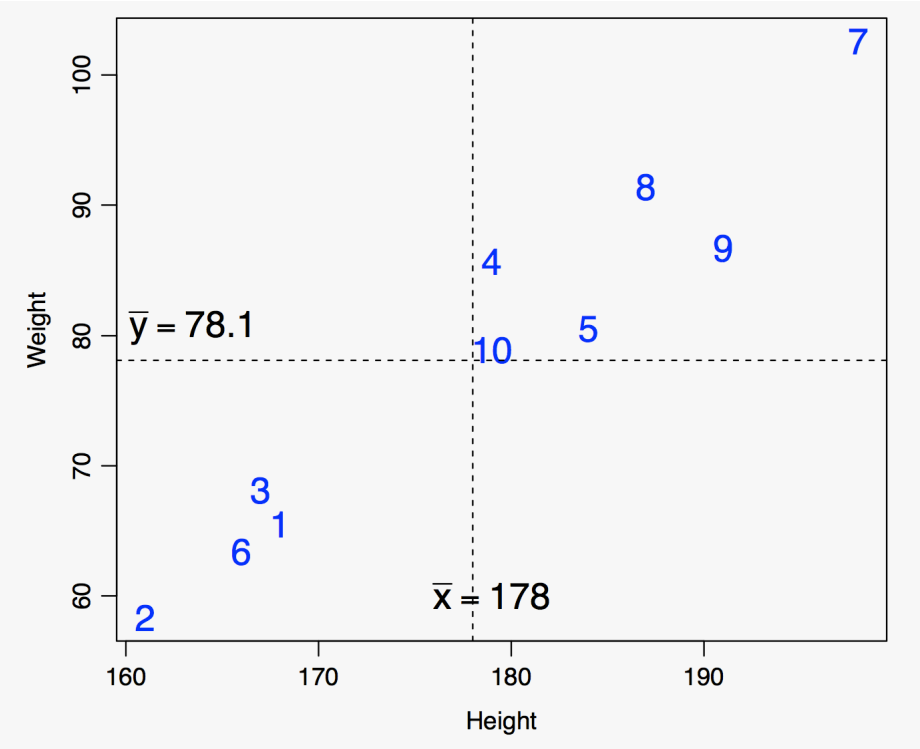
Now we want to compare different sets of observations. For example, imagine that for the same students, we are observing their heigths and their weigths.

Example 5.3.5

Let be given the table

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

We can use a "scatter plot"



Definition 5.3.7

Given two sets of observations x_i and y_j, $1 \leq i, j \leq n$. The sample covariance is given by

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i = \bar{y})$$

The sample correlation coefficient is defined by

$$r = \frac{s_{xy}}{s_x.s_y}$$

where s_x and s_y are the sample standard deviation for x and y.

Example 5.3.6

From the previous examples, we have

$$\bar{x} = 178,\ \bar{y} = 78.1$$

Computing, we find that $s_{xy} = 165.9$, $s_x = 12.21$, $s_y = 14.07$ and thus $r = 0.97$.

Without details, let us note that:
- $-1 \leq r \leq 1$
- $r = \pm 1$ iff all points are on a same line in the scatter plot.
- $r > 0$ means that the general trend is positive.

5.4 Using R software with Rstudio

Rstudio appears very similar graphically to MatLab at first glance. Eventually the use of R software can be simplified through the use of scripts.

There are many tutorials on the web, as for example many videos. Please take some time to watch them.

```
> ## additon of two numbers

> 2+3

[1] 5

> ## assignement

> y<- 3

> x<- c(1,4,6,2)
> x

[1] 1 4 6 2

> x <- 1:10
> x
[1] 1 2 3 4 5 6 7 8 9 10

> x<- seq(0,1,by=0.1)
> x

>?seq
```

Usual statistics functions uder R
- mean (x) for the mean value of the vector x
- var (x) for its variance
- sd (x) for its standard deviation
- median (x) for its median

- quantile (x,p): finds the p-th quantile
- cov(x,y) for the covariance
- cor(x,y) for the correlation

```
> ## sample mean and median

>  x<- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
> mean(x)

[1] 179

> median(x)

[1] 179

>## sample variance and standard deviation

> var(x)

[1] 149.1111

> sqrt(var(x))

[1] 12.21111

>sd(x)

[1] 12.21111



>## sample quartiles

> quantile(x,type=2)

0%  25%  50%  75%  100%
161  167  179  187   198
```

Here the option ″type =2″ is used to fit the definition adopted here.
Sample quantiles 0%, 10%, ..., 90%, 100%

```
> quantile (x,probs -seq(0,1,by=0.10),type=2)
```

For histograms, summarizing the data using rectangles displaying either frequencies or proportions of normalized frequencies:

- divide the range of data into bins of equal width (not always)
- count the number of observations in each class
- draw the histogram rectangles (in terms of frequencies or percents by area)

```
>## histogram of heights
>hist(x)

> ## density histogram of heights
> hist(x, freq=FALSE, col="red",nclass=8)


>## age of the presidents of the United States at the time of their inauguration

> age<- c(57,61,57,57,58,57,61,54,68,51,49,64,50,48,65,52,56,46,54,49,51,47,55,55,
54,42,51,56,55,51,54,51,60,61,43,55,56,61,52,69,64,46,54,47)
> hist(age, main = c("Age of Presidents at the Time of Inauguration"))



> ## empirical cumulative distribution function

> plot(ecdf(x),verticals=TRUE)



>## from smallest to largets by using sort (.)

>## back to the age of presidents


>## match these up with the integral multiples of the 1 over the number of observations

> plot(sort(age),1:length(age)/length(age),type="s",ylim=c(0,1),
main = c("Age of Presidents at the Time of Inauguration"),
sub=("Empiricial Cumulative Distribution Function"),
xlab=c("age"),ylab=c("cumulative fraction"))
```

Box Plot: depicts the five quartiles ($min, Q_1,$ $median$, $Q_3,$ max) together with abow from Q_1 to Q_3 to emphasize the IQR

```
>## basic bowplot of the heights (range =0 for "basic")

> boxplot (x,range=0, col="red",main="Basic bowplot")

>text(1.3, quantile(x), c("min", "Q1","median","Q3#,"max")
```

Another example when adding an extreme observation says 235 to the heights data.

```
> bowpolot(c(x,2345),col="red",main="modified boxplot")
> text(1.4, quantile(c(x,235)),c("min","Q1","median","Q3","max"),col="blue")
```

Another example

Males	152	171	173	173	178	179	180	180	182	182	182	185
	185	185	185	185	186	187	190	190	192	192	197	
Females	159	166	168	168	171	171	172	172	173	174	175	175
	175	175	175	177	178							

```
>males <- c(152, 171, 173, 173, 178, 179, 180, 180, 182, 182, 182, 185,
            185 ,185, 185, 185 ,186 ,187 ,190 ,190, 192, 192, 197)

>females <- c(159, 166, 168 ,168 ,171 ,171 ,172, 172, 173, 174 ,175 ,175,
            175, 175, 175, 177, 178)

> boxplot(list(males,females),col=2:3, names=c("males","females"))
```

If we have a csv type file of data as for example this one: –¿ studentheights.csv, then under R, we can:

```
>?studentheights <- read.table("studentheights.csv", sep = ";", dec = ".",
                               header = TRUE)
```

from which we get an object of type data.frame, that is typical data sets in R.

```
> ## Have a look at the first 6 rows of the data:
> head(studentheights)
```

```
> ## Get a summary of each column/variable in the data:
> summary(studentheights)
```

then we can get for example

```
> boxplot(Height ~ Gender, data = studentheights, col=2:3)
```

Here the tilde symbol means height as a function of gender.
Now R has already some data sets available.

```
>?mtcars
```

Now we want to plot the gasoline use (mpg=miles pr gallon) vs the weigth (wt): 1 mile being 1.6 km, 1 g nearly 3.8 l.

```
> ## To make 2 plots on a single plot-region:

> par(mfrow=c(1,2))

> ## First the default version:

> plot(mtcars$wt, mtcars$mpg)

> ## Then a nicer version:

> plot(mpg ~ wt, xlab = "Car Weight (1000lbs)", data = mtcars,
     ylab = "Miles pr. Gallon", col = factor(am),
     sub = "Red: manual transmission", main = "Inverse fuel usage vs. size")
```

```
> ## Barplot:

> barplot(table(studentheights$Gender), col=2:3)
```

```
>#  simple bar graph

> males<- c(58,18,16,7,1)

>barplot(males)

>#description of females

> females <- c(0,71,27,0,2)
> hiv<-array(c(males,females), dim=c(5,2))

>#  Generate side-by-side bar graphs and create a legend

> barplot(hiv, main="Proportion of AIDS Cases by Sex and Transmission Category
    + Diagnosed - USA, 2005", ylab= "percent",  beside=TRUE,
    + names.arg = c("Males", "Females"),col=colors)
    > legend("topright", c("Male-male contact","Injection drug use (IDU)",
    + "High-risk heterosexual contact","Male-male contact and IDU","Other"),
    + cex=0.8,fill=colors)



> ## Pie chart:
> pie(table(studentheights$Gender), cex=1, radius=1)



>## pie chart for the proportion of AIDS cases among US males by transmission category

> males<- c(58,18,16,7,1)
> pie(males)

> ## some colors ideal for black and white print

> colors <- c("white","grey70","grey90","grey50","black")

>#  percentage for each category

>  male_labels <- round(males/sum(males)*100, 1)

>#  number 1 indicates rounded to one decimal place
```

```
> male_labels <- paste(male_labels, "\%", sep=" ")

>#adds a space and a percent sign

>#Create a pie chart with defined heading and custom

>#colors and labels and create a legend


> pie(males, main="Proportion of AIDS Cases among Males by  Transmission Category
    + Diagnosed - USA, 2005", col=colors, labels=male_labels, cex=0.8)
    > legend("topright", c("Male-male contact","Injection drug use (IDU)",
    + "High-risk heterosexual contact","Male-male contact and IDU","Other"),
+ cex=0.8,fill=colors)

>#entry cex=0.8 indicates that the legend has a type set that is 80\% of

>#the font size of the main title
```

To get a summary of data

```
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  42.00   51.00   54.50   54.64   57.25   69.00


>#then

> boxplot(age, main = c("Age of Presidents at the Time of Inauguration"))
```