Statistics with R and RStudio

Dr Lucio Vinicius

Department of Anthropology

l.vinicius@ucl.ac.uk

Lecture 1

Introduction to Statistics and *R:* Descriptive Statistics

描述

About me

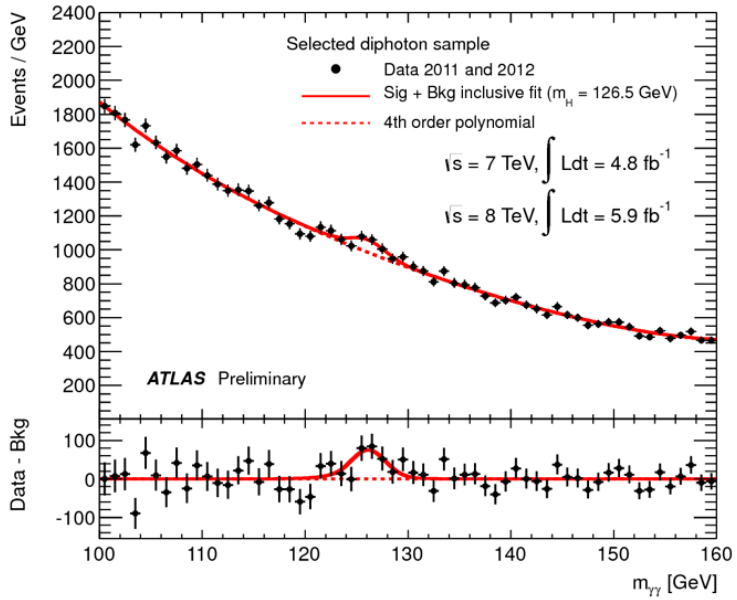
Statistics: a tool and way of thinking

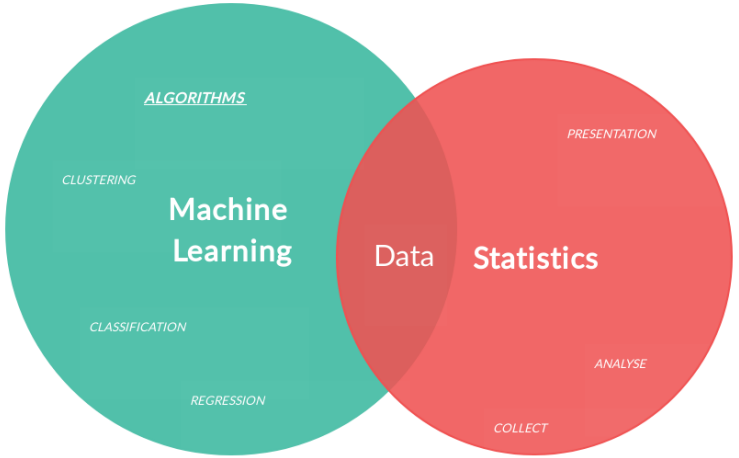
- Statistics is a way of thinking based on the analysis of data
  - beyind the certainty of mathematical demonstrations (that do not data for validation)
  - beyond qualitative assessments (that express views or preferences not grounded on data analysis)

- Statistical methods extract probabilities and most likely outcomes from collection and observation of data
  - Statistical results are not ‘true’: they reflect the most likely outcomes

- Applications:
  - Pure sciences: natural and social etc.
  - Applied sciences: medical, public health, engineering, business etc.
  - Technical extensions: stock market analysis, fintech, business, machine learning and AI…virtually all fields where the purpose is to learn from data

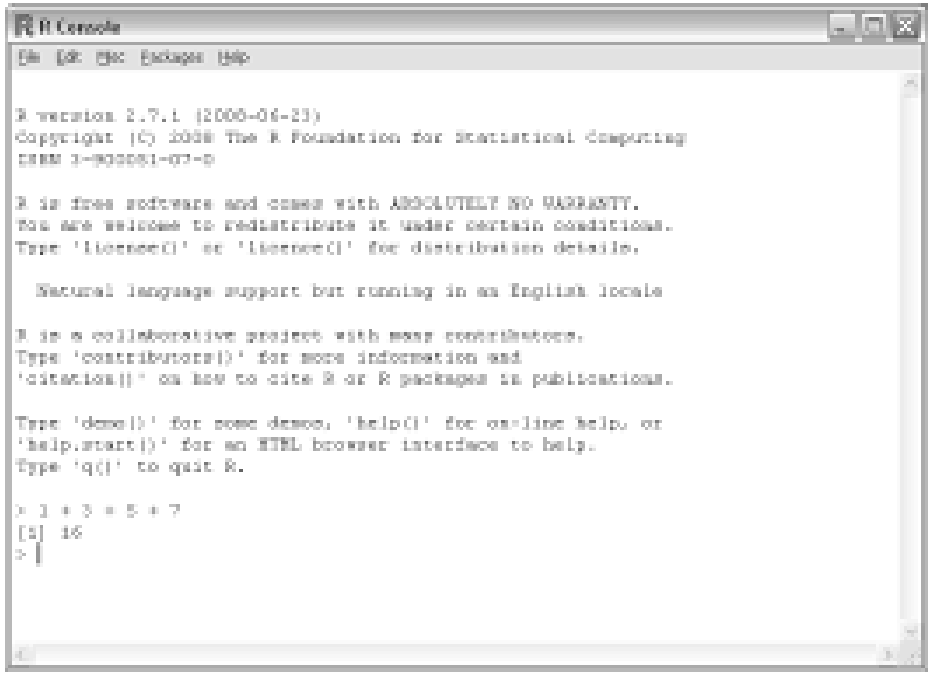
- Conclusion: if you want to do any of the above, you need statistics!

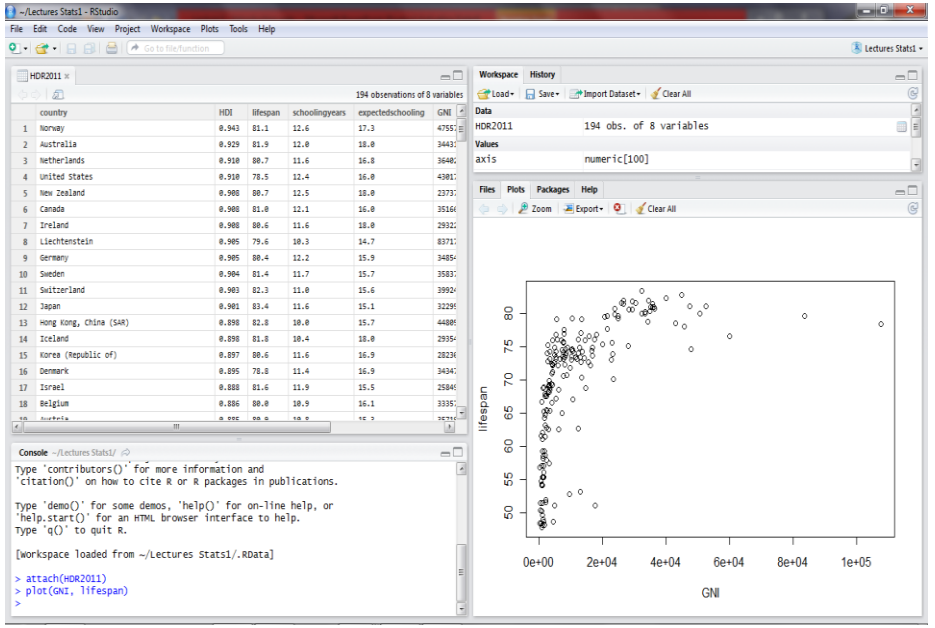




R and *RStudio*

- *R* is a free, command-line statistical software
- *RStudio* is a very user-friendly *R* interface
- Installing *RStudio* on your laptop
  - (1) download and install *R*
  - http://cran.ma.imperial.ac.uk
  - (2) download and install *RStudio*
  - http://rstudio.org/download/desktop
  - (3) start *RStudio* only (this will launch *R)*
    - Never start *R* itself! Only use *RStudio*





RStudio

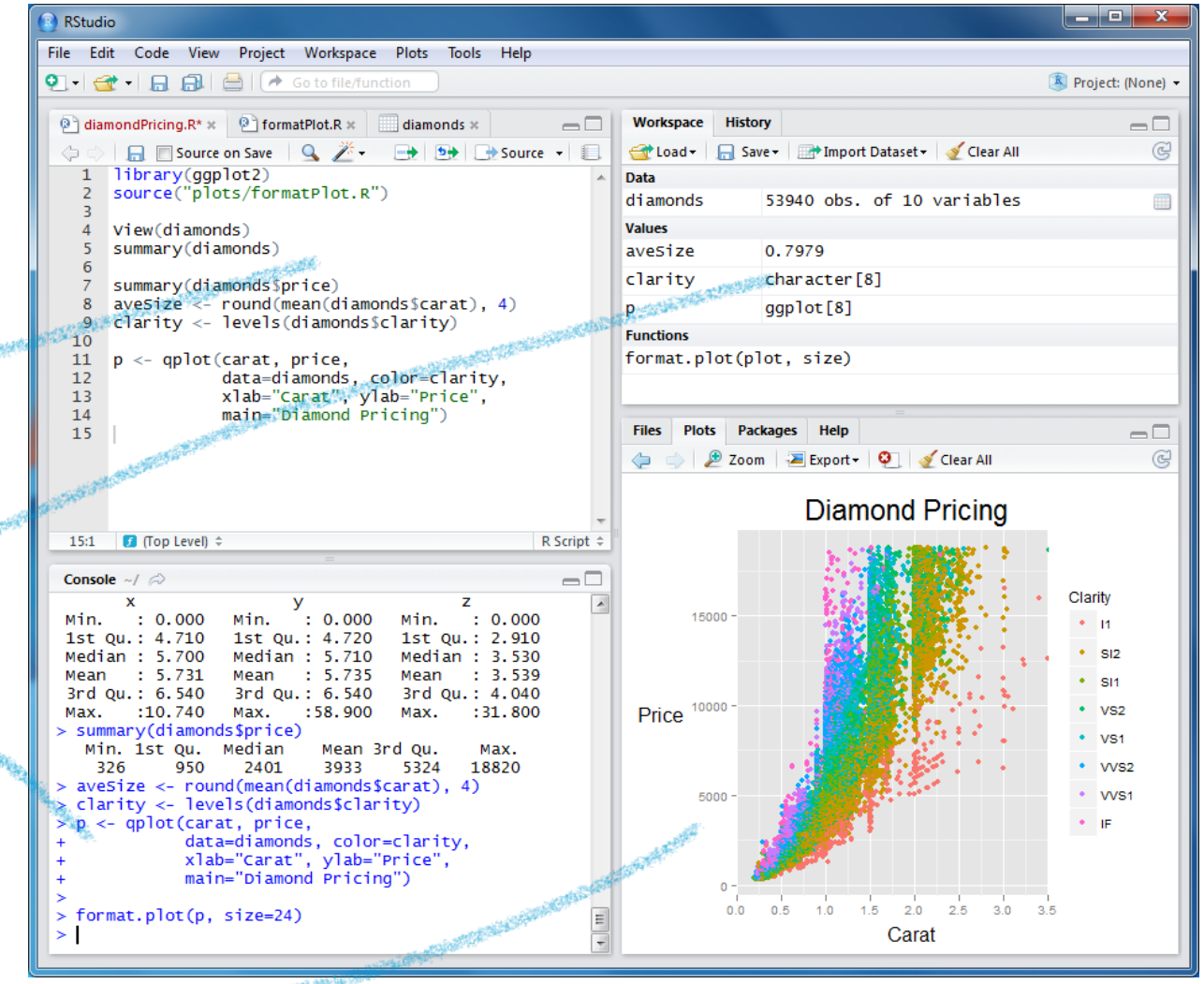
- *RStudio* interface has four panels
  - if you see only three panels the first time you launch Rstudio, just click on expander button

- Console/input panel (bottom left)
  - commands entered after '>' prompt
  - '+' to continue command on separate line
  - '#' (hash button) for comments, notes

- Source panel (top left)
  - to edit code, visualise datasets

- Environment/workspace/history panel (top right)
  - registers all command history and all data currently held in memory

- Plots panel (bottom right)
  - plots
  - new packages installed, help, files
  - search, help files



RStudio as a calculator

- *RStudio* is a calculator; try

  > 3 + 2

  > 3 – 2

  > 3*2

  > 3/2

  > 3^2

  > 3**2

  > sqrt(16)

```
log10(100)   log2(32)
log(16,base = 2)
log(x)是以e为底
```

Exercises:

a) what is the function exp(x)?
b) what is the function log(x)?
c) how to estimate log in base 10 and base 2 in R?
d) can you think of another way of calculating sqrt(16)?

```
16**0.5
```

Tip: start using the RStudio help and search (bottom right panel),
or

> ?log 查帮助文档

Defining values, vectors

- Variables, vectors, data frames etc. can also be defined with operator " <- " (or " -> ")
  - Try:

```
> x <- 2
> x
```

and

```
> y <- c(1, 2, 3)
> y
```

- Tip: Use **up** and **down arrow keys** to navigate through command history
  - Note: if you are copying and pasting from console panel, you are wasting time!

Exercises:

a) Create vector x with five values
b) Create vector y with five values
c) Calculate x + y and x*y
d) Now redefine x as 5: what happens?
e) Recalculate x + y and x*y
f) Now create a data frame (a data file) with x and y as columns (with arbitrary names).

Code:

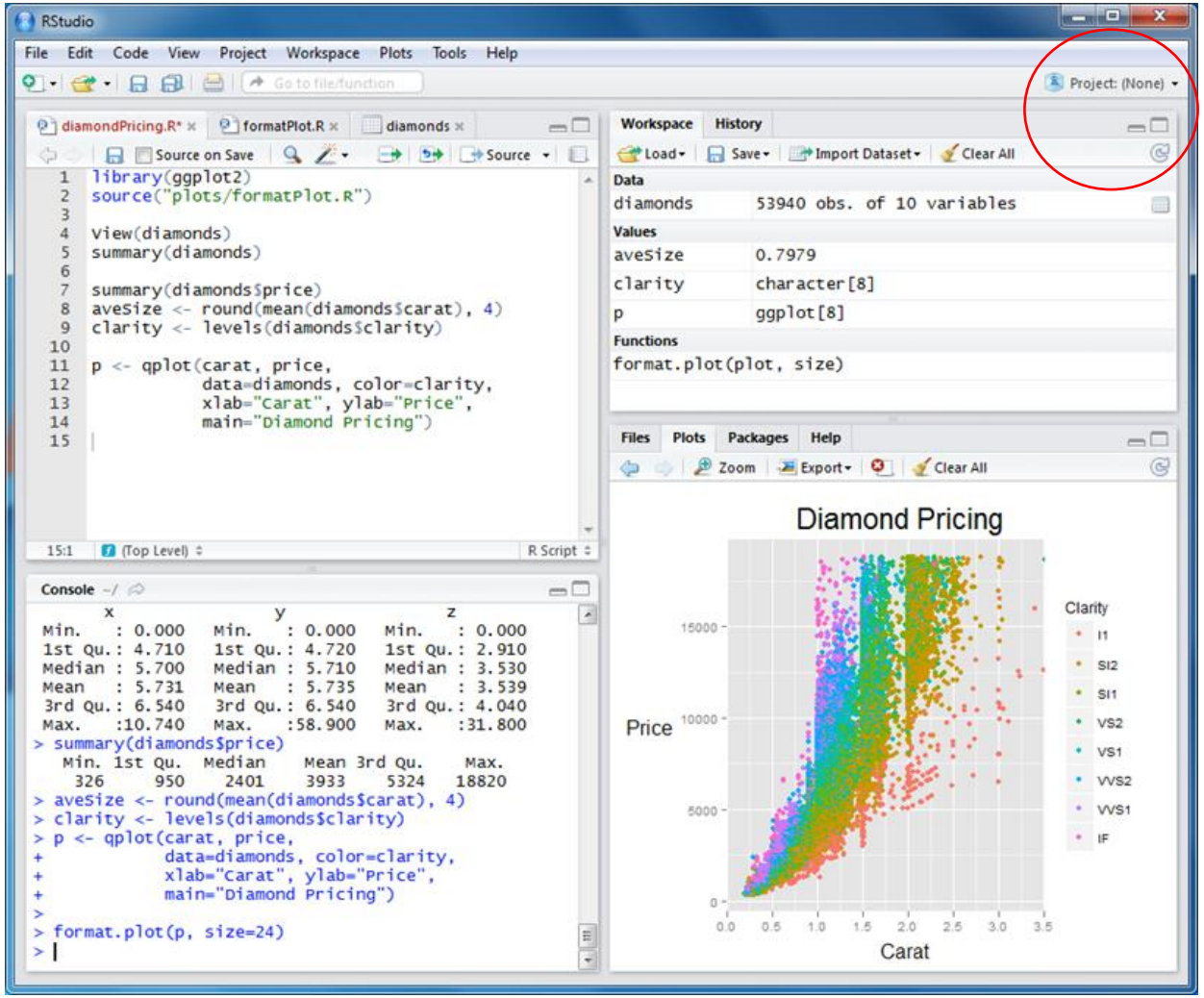
```
> file1 <- data.frame(mycol1=x, mycol2=y)
```

data.frame是构造数据框

Naming project and creating a folder

- To organise your work and files, create a new project (e.g. project ‘R course UCL’)

- Select ‘New Project’, top right
  - choose project name
  - choose location for the folder that will contain project files



Importing dataset and command script

Dataset

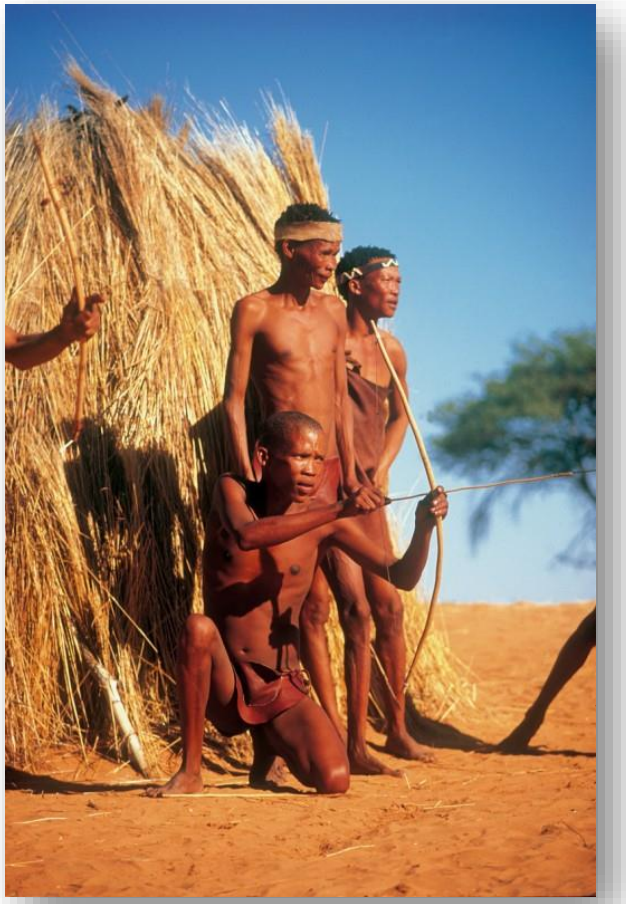
- Importing a file creates a copy of the original file in the workspace of an *R* session
  - (Modifications to imported file do not affect original Excel file)

- Download file *KungCensus.xlsx* file from our Moodle page
  - (or .csv version)

- Note: In *R*, file names are case-sensitive

- Good practice: select the shortest possible file name
- If original file is too long/weird, change it in import tab
- Or create a new data frame (and after that, delete original one):

```
> f1 <- Old_file_with_very_long_name
```



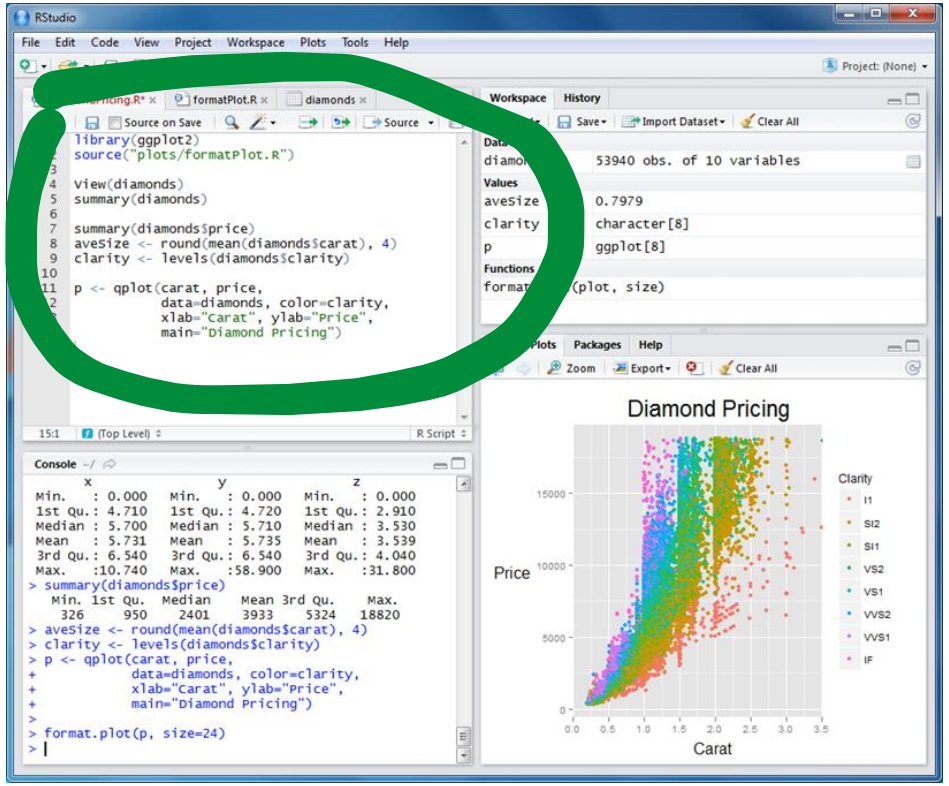
Importing dataset and command script

Command script

- We create a new script in the source panel (top left)

右上角import dataset导入数据集

- But we can also open an ‘R script’ file with code already written to run analyses.
  - download file *R code, Lecture 1.R* from Moodle page, and open it in source panel
  - ps: this is not the same as importing!
    - you *import* datasets in environment panel (top right)
    - but *open* script files in source panel (top left)



总和
总值/大小

Descriptive statistics: mean

- At the most basic level, descriptive statistics provide summaries of variables
  - predictive statistics (later) estimates probabilities of hypotheses etc.)
- The sample **mean** is the most informative sample summary
- The mean is the sum of sample values divided by sample size
- The mean may differ from any individual values
  - mean fecundity in the UK is about 1.7 children per woman; but no woman can have 1.7 children!
- Calculating mean value of a variable (column from a data frame):

```
> mean(KungCensus$weight, na.rm=T)
[1] 35.76768
```

平均值可能与任何单个值都不同，比如1.7个孩子

Exercises:

a) Try to calculate

```
> mean(KungCensus$weight)
```

What happens?

b) Try to calculate

```
> mean(weight)
```

What happens?

c) What is the mean height of !Kung people?

Notes:

na.rm=T

去除值为NA的项

- =not available, remove, true
- Removes NAs (missing data)
- Parameter required by some but not all functions

参数，界限，范围

file$variable

同时的，同步的

- *R* can work simultaneously on different datasets
- you must indicate which file a given variable is from

必须指出给定变量来自哪个文件

Range 给出最大最小值

- We can also look at variable range, or the minimum and maximum weight values

```
> range(KungCensus$weight, na.rm=T)
[1]  2.948348 64.750258
```

- This suggests significant variation around the mean weight of 35.8kg

Median 中位数

- Another measure of central tendency is the median; this is another attempt at capturing an ‘average’ weight

- The median is the sample ‘mid-point’: or the measure right in the middle of the distribution
  - i.e. half the people have weights below median, and half above

```
> median(KungCensus$weight, na.rm=T)
[1] 40.49726
```

四分位数

- A quartile divides sample into *quarters* 四个数分别为25%，50%，75%，100%
  - 25% of sample below 1st quartile
  - 50% below 2nd quartile (=median)
  - 75% below 3rd quartile
  - 100% below fourth quartile (=maximum value)

Summary of variables and files

summary()函数给出了最小最大平均中位数四分位数，和NA个数

- function *summary( )* produces min, max, mean, median, quartiles, and NAs (missing cases)

```
> summary(KungCensus$weight)
Min.   1st Qu.  Median  Mean   3rd Qu.  Max.   NA's
2.948  22.000  40.500  35.770  47.460  64.750 230.000
```

You can also summarise the whole dataset

```
> summary(KungCensus)
```

```
x<-c(12,3,12,4)
tmp<-table(x)
index<-which.max(tmp)
tmp[index]
```

Exercises:

a) What is the code below doing?

```
> table(KungCensus$weight)
```

统计数量

b) And this?

```
> sort(table(KungCensus$weight)))
```

按照统计出来每一项的个数
从小到大排序

c) And this one?

```
> sort(table(as.integer(kc$weight)))
```

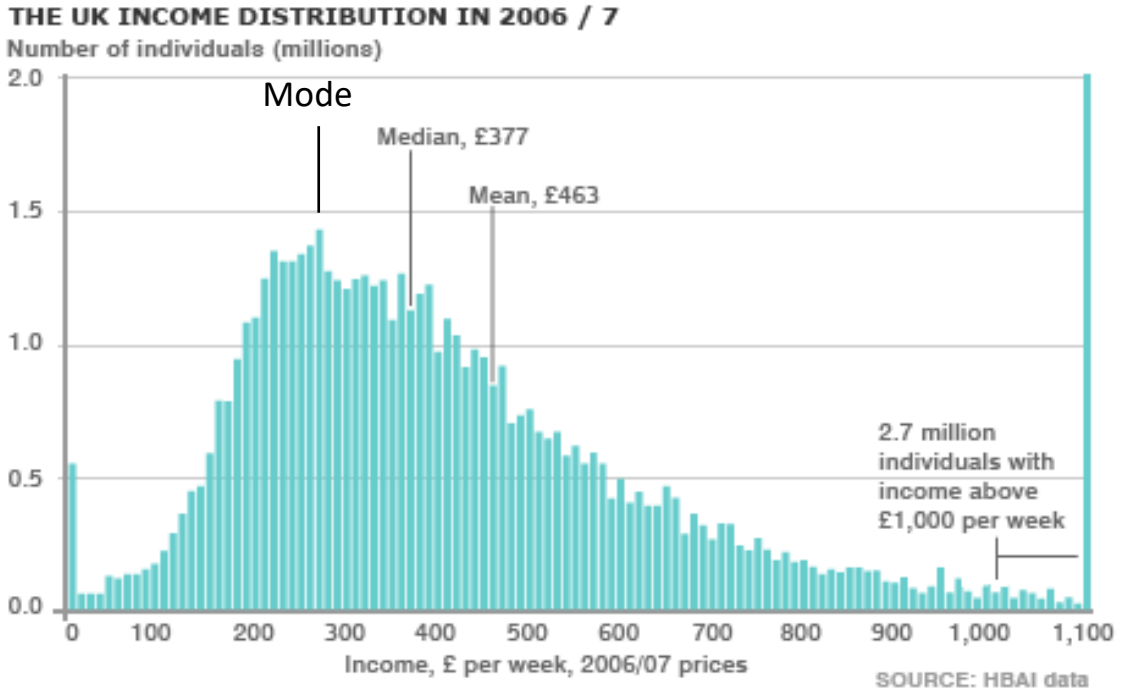
as.integer是直接把数据的小数点后丢掉

d) What is the mode (=the most frequent value of the variable)? mode()函数显示数据类型

一组数据中的众数不止一个

Differences between mean, median and mode

- Mean, median, and mode may differ in the same population
  - extreme cases may significantly alter the mean 极端情况可能会显著改变平均值



Measures of dispersal: variance 分散度：方差

- Measures of central tendency may provide an incomplete and misleading description of populations,
  - they must be supplemented with info on *variation* around central trend

- The most common measures of 'dispersal' are
  - *variance* 方差
  - *standard deviation* 标准差

标准方差

- Sample ***variance*** (σ^2, sigma squared) measures mean *squared* deviation of all observations x_i from the mean μ:

$$\sigma^2 = \frac{1}{n}\sum_i^n (x_i - \mu)^2$$

- Why squared deviation? 平方是为了消除偏差，不然可能这玩意是0
  - To eliminate sign (plus or minus); otherwise total sum may be zero even when there is variation around mean

- How much total variation around mean weight?

```
> var(KungCensus$weight, na.rm=T)
[1] 229.7117
```

Standard deviation 标准差

- Variance is an important measure, but its interpretation is not very intuitive

方差的平均根

- *Standard deviation* (σ) is the square root of variance
  - interpretation is straightforward: *sd* is the expected deviation from the mean by any selected case in the sample

标准差是样本中选中的样例与平均值的预期偏差

```
> sd(KungCensus$weight, na.rm=T)
[1] 15.15624
```

- What does a standard deviation of 15.15 kg signify?
  - if you select a random person from the sample, you expect it to deviate by 15.15 kg from the mean of 35.76 kg (i.e. ~43% deviation from mean)
  - sd is a measure of dispersal around the mean
  - important: the larger the standard deviation, the less representative of the average case in the sample the mean is

Exercises:

a) Estimate the variance in offspring number in the Kung population (=variable *kids*)

b) Estimate standard deviation of variable *kids*; what does that mean?

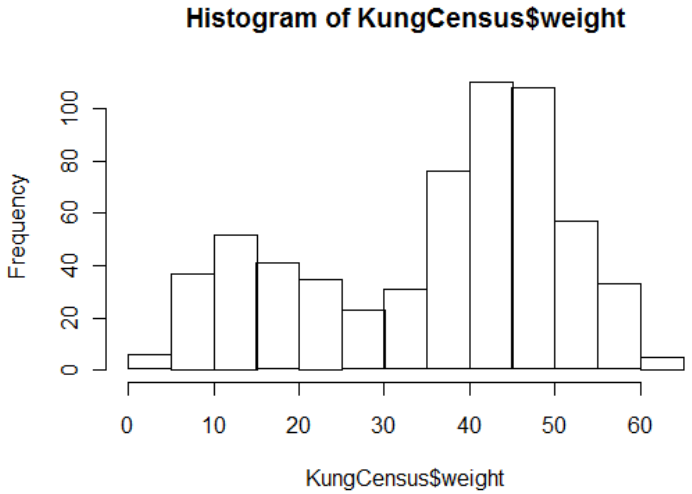
标准差越大，平均值越不能代表样本中的平均情况

Visualising distributions: histograms 直方图

- Histograms help to visualise distribution of a variable

- Let's plot the distribution of weight

  > hist(KungCensus$weight)



freq用于设置纵轴表示频数还是概率密度，freq=T表示频数(frequency)，F表示概率密度(默认，density) probability和freq相反

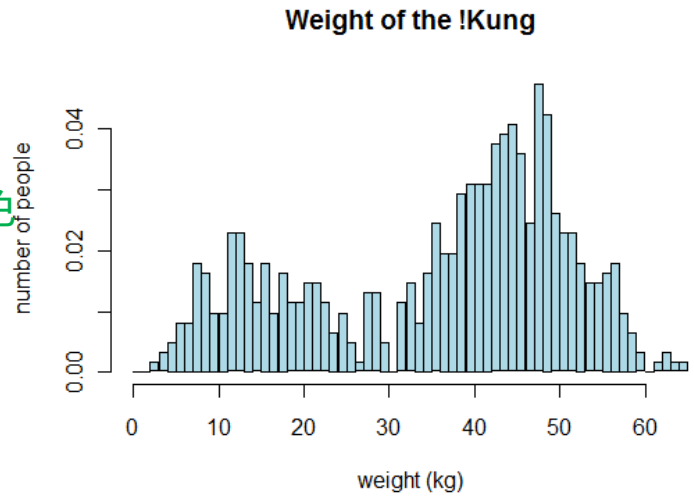
  - plot above provides basic info, but we can add our choice of plot title, axis title, x-axis breaks (subdivisions) in x axis etc.:

```
>hist(KungCensus$weight,
breaks=seq(0,65,1), col="lightblue",
main="Weigth of !Kung",
xlab="weight (kg)", ylab="number of people")
```

颜色，用colour()可以查看所有颜色

图的题目

x坐标和y坐标的题目



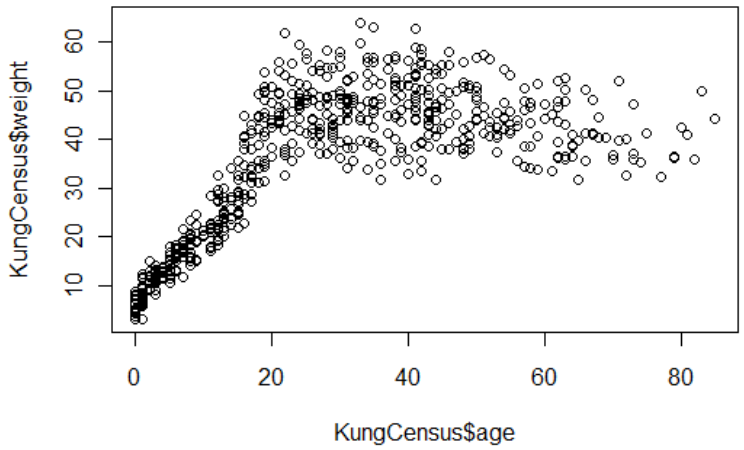
Plots

```
plot(forehead~sample,pch=15,col="DarkTurquoise",cex.axis=1.5,cex.lab=1.5,cex.main=1.5,
ylim=c(0,400),ylab="Number of active sweat glands per cm2",main="Number of active sweat
glands per cm2 in forehead, forearm and back") #pch表示散点用什么形状表示，col表示颜色，
ylim表示Y轴范围，ylab表示Y轴标题，main表示图片标题,cex.axis表示修改坐标轴刻度字体大小，
cex.lab表示修改坐标轴名称字体大小，cex.main表示修改标题字体大小
```

- Plots are a useful way of representing the relationship between two variables

可以看到所有样本点

- They allow you to see all sample points



- For example, weight should increase until adult age

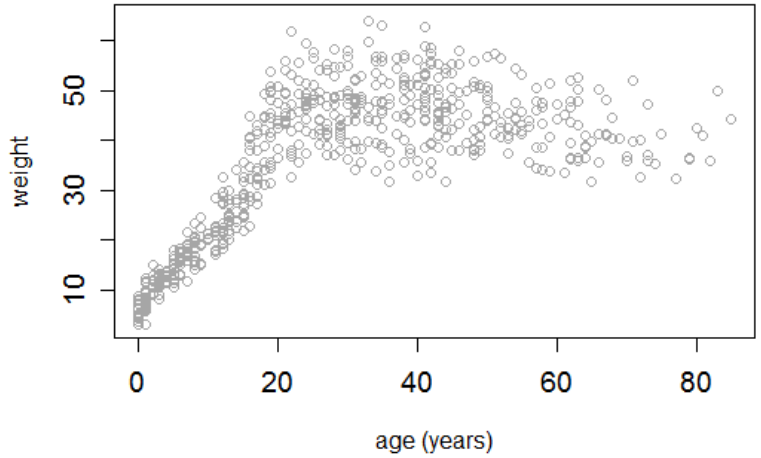
纵坐标~横坐标

- We can plot weight against age

```
> plot(weight ~ age, data=KungCensus)
```

- A better-looking plot:

```
> plot(weight ~ age, cex.axis=1.2, col="grey65",
xlab="age (years)", ylab="weight“,
data=KungCensus)
```



Exercises:

a) Create a histogram of *age* distribution

- -create a basic histogram first
- -then create a better plot with an appropriately named axis
- -which range should it cover?
- -how many breaks in *x* axis should it have?

b) Produce a plot of height by age

- -create basic plot
- -create a more sophisticated plot (with colours, main title etc.)

Exercises

c) Run command

```
> seq(0, 65, 1)
```

Now change each of the three values separately. What is the function of 0, 65 and 1 in the code?

d) Compare

```
> plot(weight ~ height, data=KungCensus)
```

and

```
>plot(weight, height, data=KungCensus)
```

这种不行，可以改成kc$weight就可以用逗号，改后weight是x坐标，height是y坐标

How does “ , “ instead of “ ~ “ change the output?
(note: always use “ ~ “)

e) Plot in grey50
tip: run command *colours()*
Note: to save your file, use Export function in plot panel (bottom right)

References, help, bibliography

Books

- Dalgaard, P. 2008. *Introductory Statistics with R.*

(useful guide to our course)

- *R for Data Science*

https://r4ds.had.co.nz/index.html

(very good online intro to R plotting, programming etc, but not statistics)

- *R* help files (Plots panel in *RStudio*)
- Other online resources:

- http://stat.ethz.ch/R-manual/R-patched/library/base/html/00Index.html
- http://www.statmethods.net/
- http://https://stats.stackexchange.com/ (search for anything in R)