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

Two aims:
- Describe
- Predict

Statistical variables

- Certain properties of individuals, communities, countries etc. are unique and contingent factors etc.; it is hard to reduce them to a ‘number’ or a ‘category’

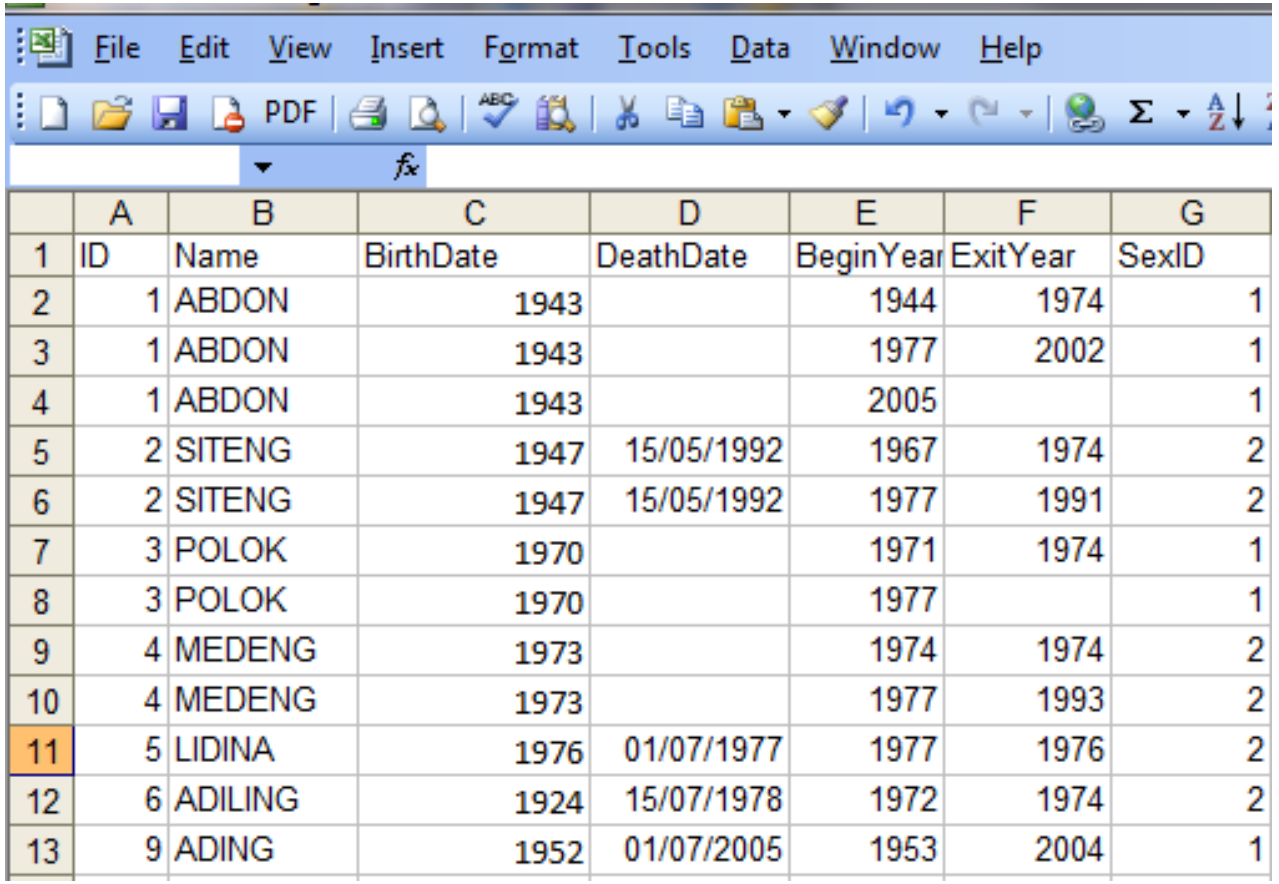
- But certain properties of study subjects/objects can be coded as v*ariables*

- A variable can be seen as a dimension of available information
  - continuous (measurement): income, height, life expectancy
  - discrete (count): number of children or wives
  - ranked (order): birth order, marriage order
  - categorical: gender, country, religion etc.

Datasets

- We typically analyse collections of data, or *datasets*
  - Variables: columns
  - Cases: rows

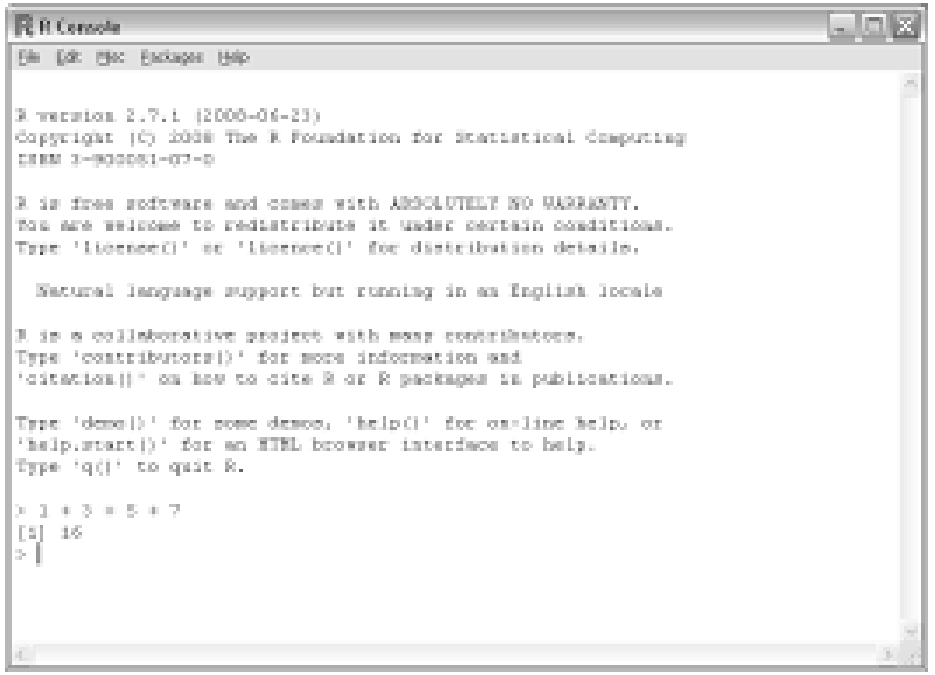
- Creating and analysing datasets
  - created in *Excel* (easy)
  - imported into *RStudio* for analysis (as .xlsx or .csv or .txt files)

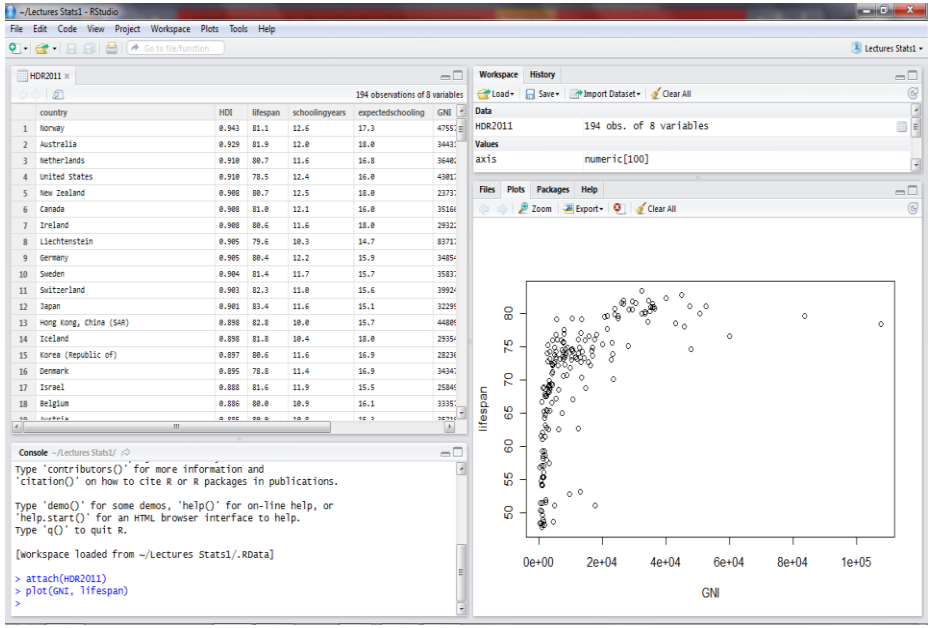


	A	B	C	D	E	F	G
1	ID	Name	BirthDate	DeathDate	BeginYear	ExitYear	SexID
2	1	ABDON	1943		1944	1974	1
3	1	ABDON	1943		1977	2002	1
4	1	ABDON	1943		2005		1
5	2	SITENG	1947	15/05/1992	1967	1974	2
6	2	SITENG	1947	15/05/1992	1977	1991	2
7	3	POLOK	1970		1971	1974	1
8	3	POLOK	1970		1977		1
9	4	MEDENG	1973		1974	1974	2
10	4	MEDENG	1973		1977	1993	2
11	5	LIDINA	1976	01/07/1977	1977	1976	2
12	6	ADILING	1924	15/07/1978	1972	1974	2
13	9	ADING	1952	01/07/2005	1953	2004	1

R and *RStudio*

- *R* is a free, command-line statistical software
- *RStudio* is a very user-friendly *R* interface
  - additional buttons, separate panels etc.
- 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)*

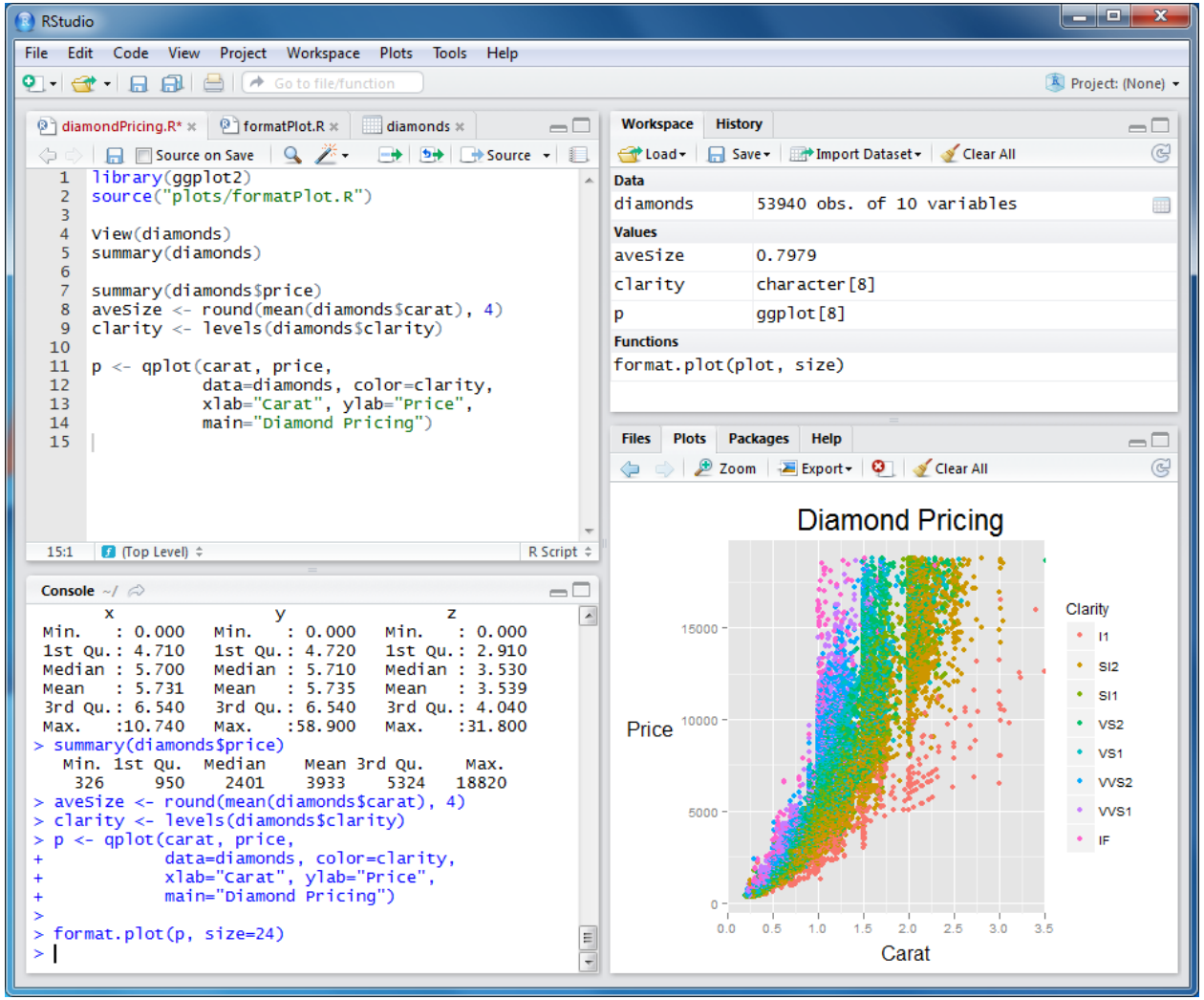




RStudio

RStudio interface has four panels:

- Console/input panel (bottom left)
  - commands entered after '>' prompt
  - '+' to continue command on separate line
  - '#' for comments, notes
- Source panel (top left)
  - where commands and workspace are edited and datasets viewed
- 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



R as a calculator

- *R* is a calculator; try

```
> 3 + 2
> 3 – 2
> 3*2
> 3/2
> 3^2
> 3**2
> sqrt(16)
```

16**0.5

Exercises:

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

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

Defining values, vectors

- Variables, vectors can be defined with "makeshift arrow" operator " <- " (or " -> ")
  - Try:

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

or

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

- Tip: Use up and down arrow keys to navigate through command history

Exercises:

a) Define x as a vector with five values
b) Define y as a vector with five values
c) Calculate x + y and x*y
d) Now define x as 5: what happens?
e) Recalculate x + y and x*y

f) Now make a data frame (a data file) with x and y as columns (with arbitrary names). Code:

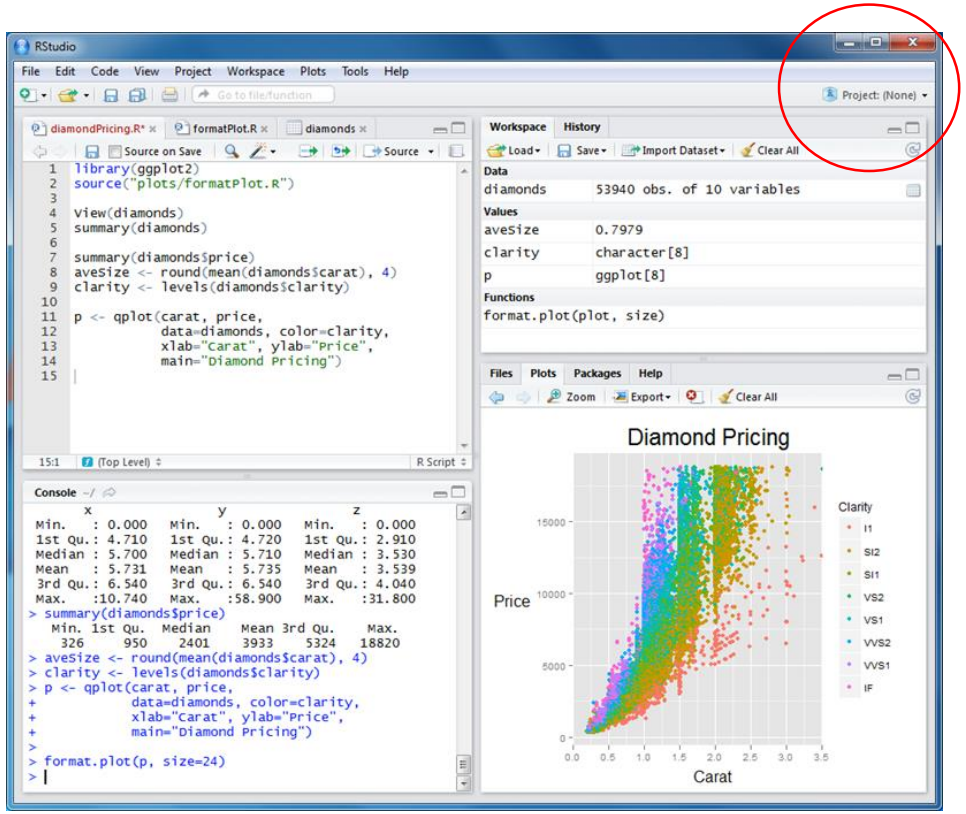
```
> file1 <- data.frame(mycol1=x, mycol2=y)
```

Redefine: file1[3,2] <- 1

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

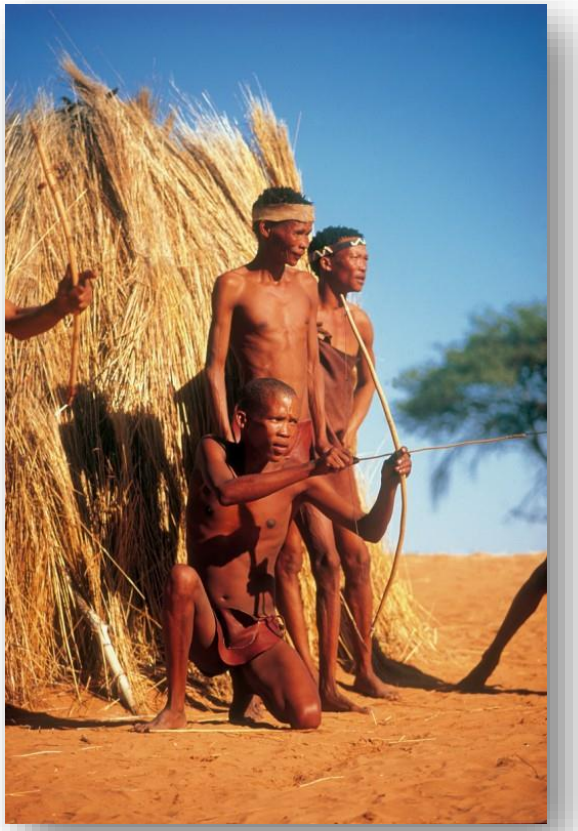
- Importing a file creates a copy of the original file in the workspace of an *R* session
  - (Modifications to imported file do not affect original Excel file)

- Download file *KungCensus.csv* file from our Moodle page

  https://moodle.ucl.ac.uk

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

- Good practice: select the shortest possible descriptive file name
- If original file is too long/weird, change it in import tab



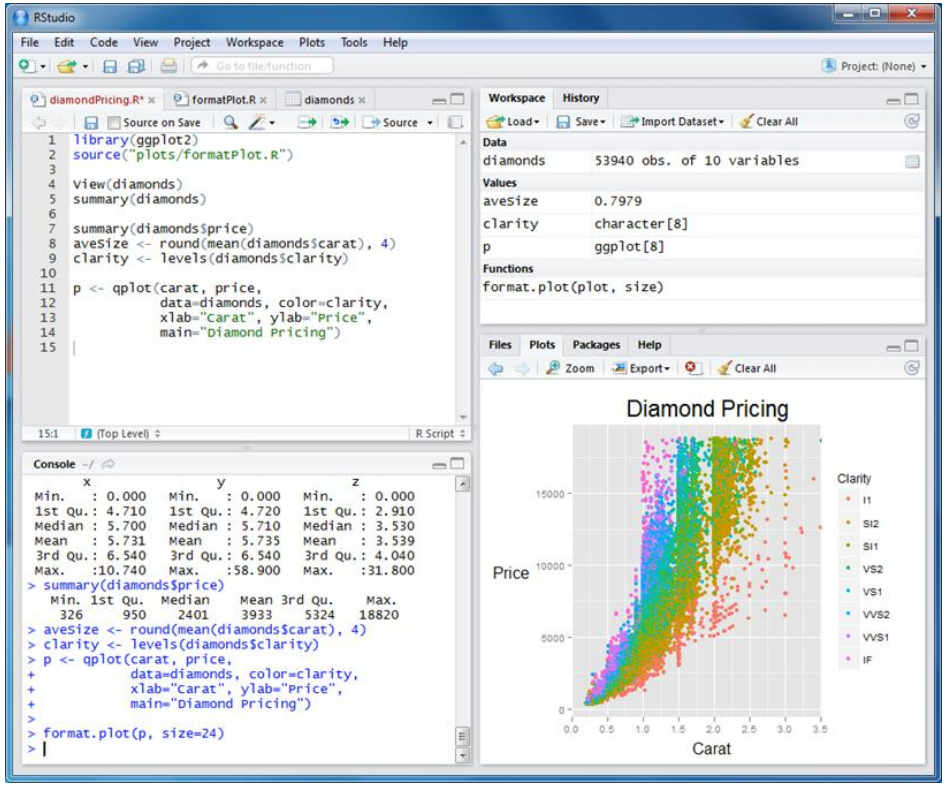
Importing dataset and command script

Command script

- We create scripts on the source panel

- We can also open an ‘R script’ file with code already written to run analyses.
  - download file *R code, Lecture 1.R* and open it in source panel

  - ps: this is not the same as importing!
    - you *import* datasets but *open* script files



Descriptive statistics: mean

- Descriptive statistics provide summaries of variables

- The sample **mean** is the most informative sample summary
- Mean is the sum of sample values divided by sample size
- Mean may represent a sample and be different from any sample values
  - mean fecundity in the UK is about 1.8 children per woman

- Running a command*:*
  - in the R script file, write command below
  - then click on the line, and click 'run'

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

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?

A/ [1] NA
— missing values in the column
B/ Error in mean(weight) : object 'weight' not found
— unnamed file

Notes

na.rm=T

- Not available, remove, true
- Parameter required in some but not all functions
- Removes NA (missing data)

file$variable

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

Range

- 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

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

- The median is the sample ‘mid-point’
  - i.e. half the !Kung have weights below median, and half above

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

- A quartile divides sample into *quarters*
  - 25% of sample below 1st quartile
  - 50% below 2nd quartile (=median)
  - 75% below 3rd quartile

Summary of a variables and whole files

- 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)
```

Exercises:

a) What is the code below doing?

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

b) And this?

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

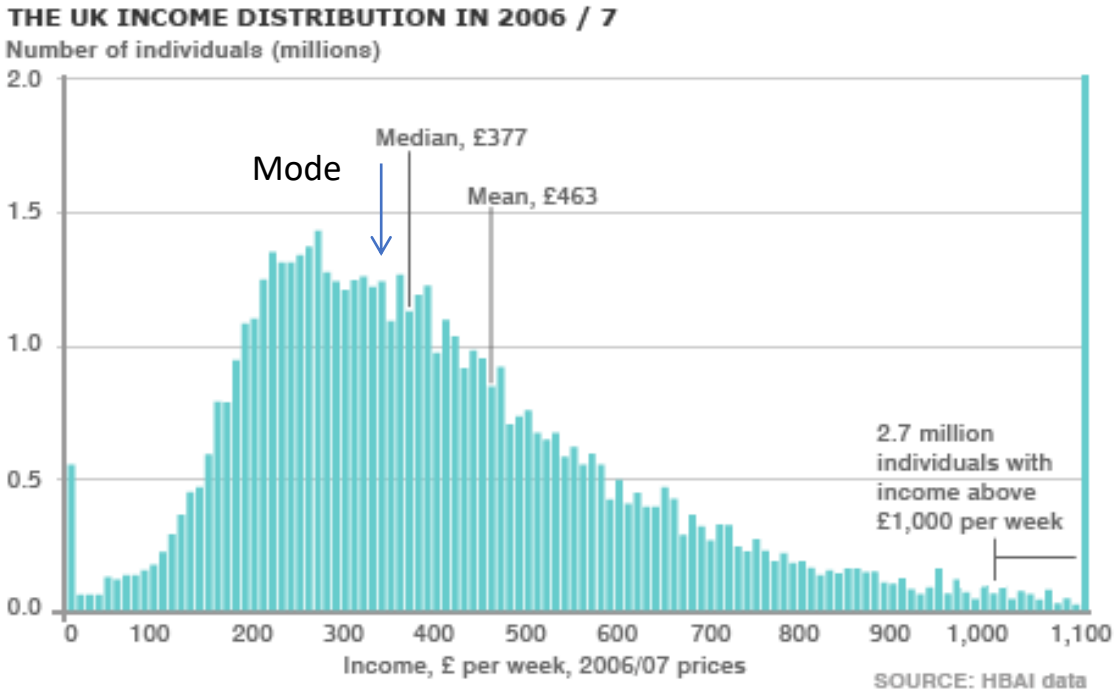
Sorted by frequency: low > high

c) And this one?

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

Differences between mean, median and mode

- Mean, median, and mode may differ in the same population
  - e.g. extreme cases may significantly alter mean



Measures of dispersal: variance

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

- The most common measures of 'dispersal' are *variance* and *standard deviation*

- 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?
  - 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

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

- *Standard deviation* (σ) is the square root of variance
  - Interpretation is straightforward: *sd* is the expected deviation of any random case from mean

```
> 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 (=variable *kids*)

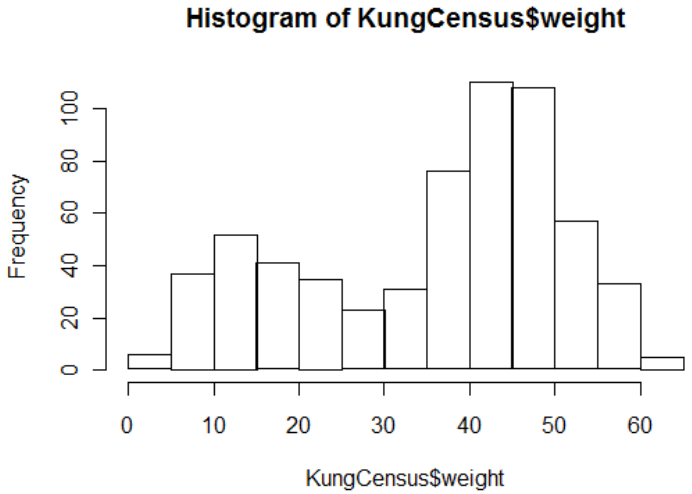
b) Estimate standard deviation of variable *kids*; what does that mean?

Visualising distributions: histograms

- Histograms help to visualise distribution of a variable

- Let's plot the distribution of weight

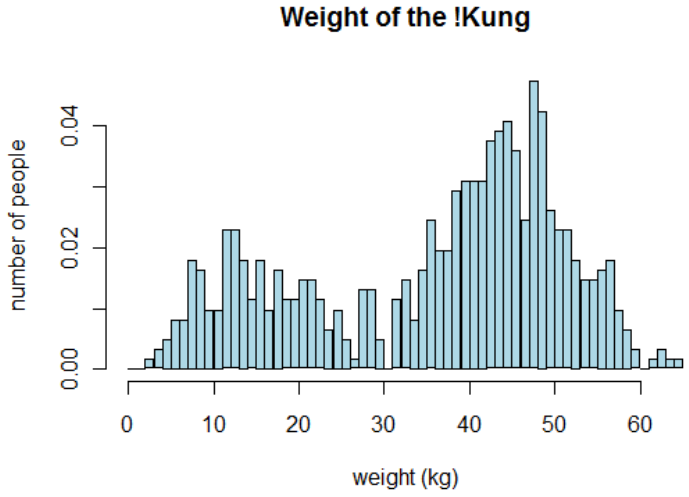
```
> hist(KungCensus$weight)
```



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

Sequence

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



Plots

- Plots are a useful way of representing the relationship between two variables
- They allow you to see all sample points
- For example, weight should increase until adult age
- We can plot weight against age

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

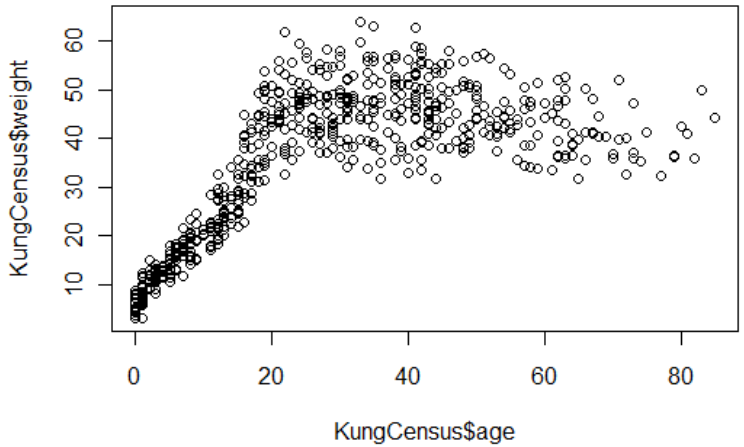
Means “by” , graph shows Y by X

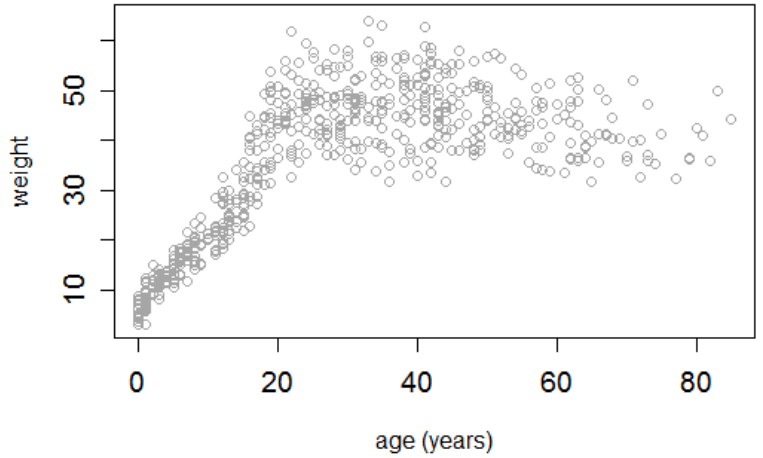
- A better-looking plot:

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

Size of printing statistical signs, default 1

Grey 1- 100, black to white





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)
```

Starting point, ending point and step

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)
```

Weight on the left, Y-axis

and

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

Weight at the bottom, X-axis

How does " , " instead of " ~ " change the output?

e) Plot in grey50

tip: run command *colours()* colours(170) tips: by name or order

References, help, bibliography

- Book:

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

(very useful guide to our course)

- *R* help files (Plots panel in *RStudio*)

- Online resources:

- http://stat.ethz.ch/R-manual/R-patched/library/base/html/00Index.html
- http://www.statmethods.net/
- http://https://stats.stackexchange.com/