OQC

Oxford Q-Step Centre

LAB SESSION 2

Tutor.name@politics.ox.ac.uk

OQC
Oxford
Q-Step
Centre

Today

- Homework Lab 1
- Working with the Lijphart data-set
- Getting an overview of your data
- Descriptive Statistics
- Correlations

OQC
Oxford
Q-Step
Centre

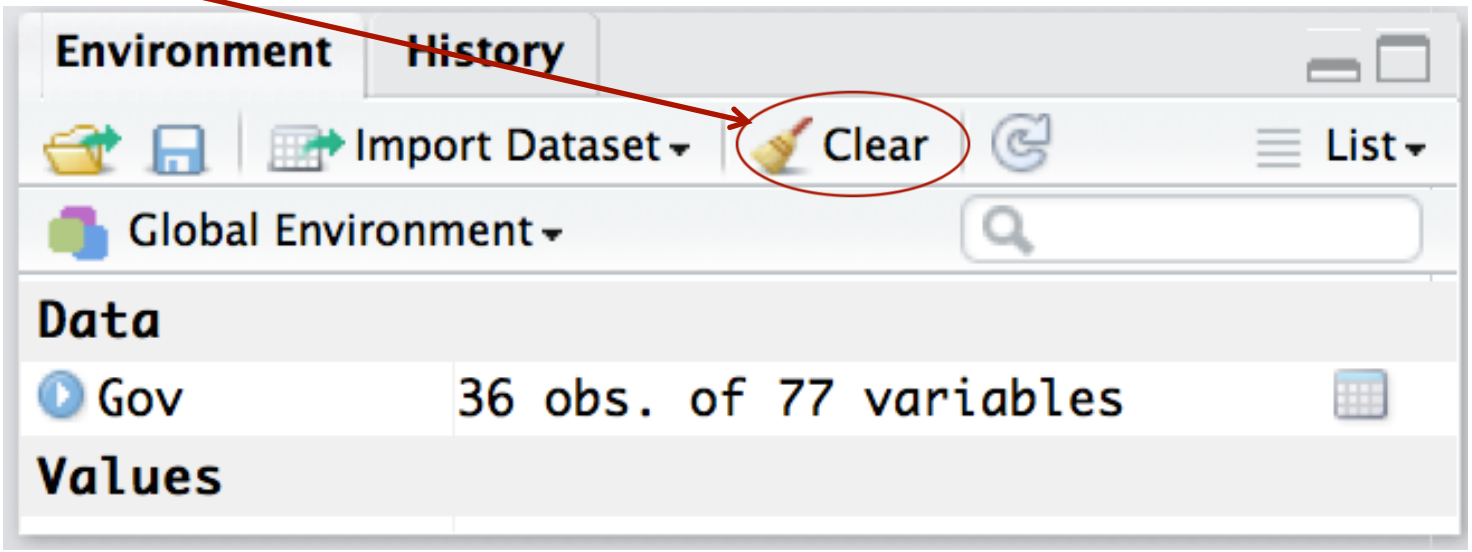
Log-in details

Login:	**Password:**
q-step-01	q-step-01
q-step-02	q-step-02
……	……
……	……
q-step-10	q-step-10
q-step-11	q-step-11

OQC
Oxford
Q-Step
Centre

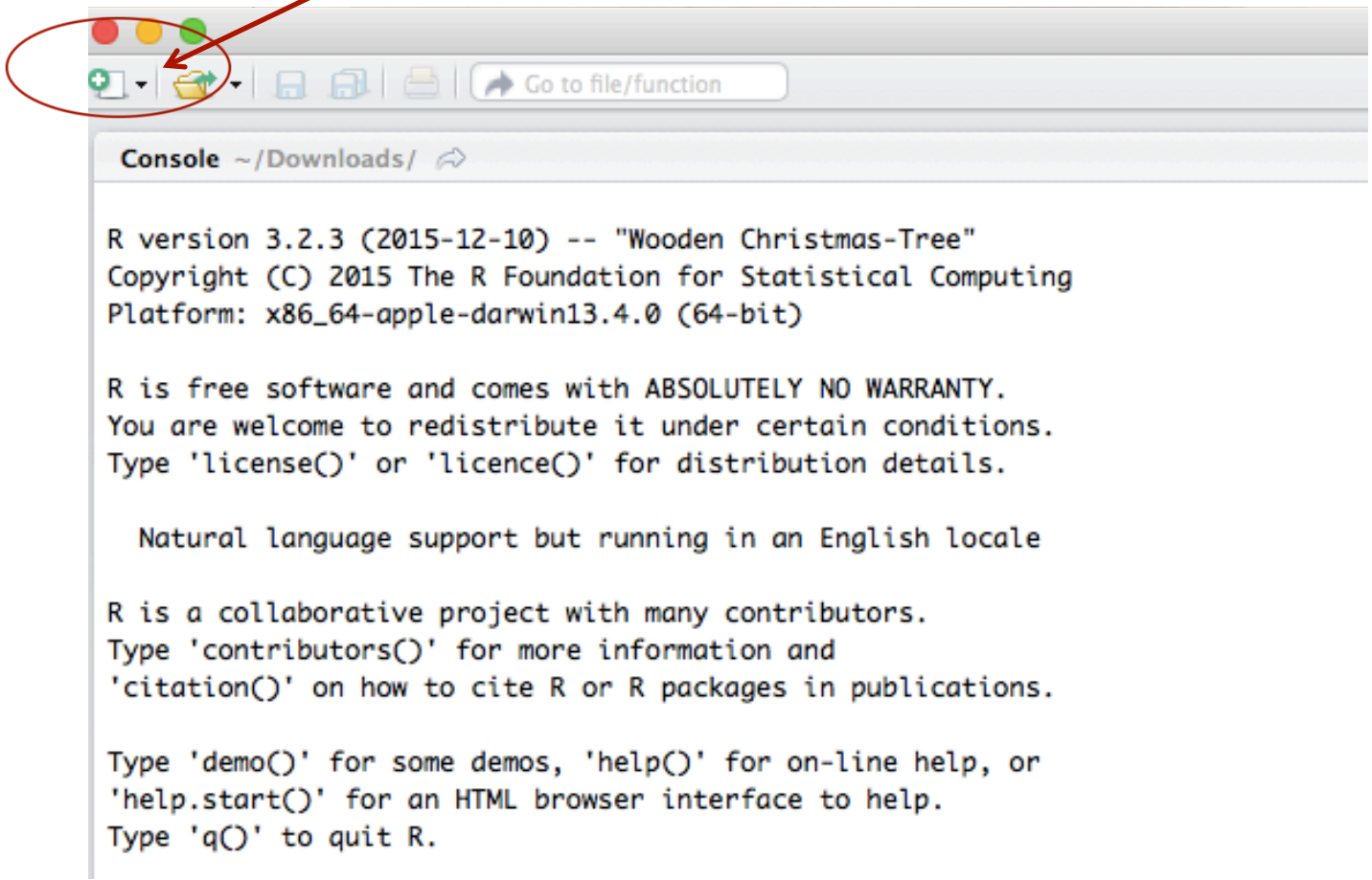
First: clean your workspace!

- **Command:** CTRL + L
- **Or:**



Second: open a new R Script

OQC
Oxford
Q-Step
Centre



Homework Lab 1

OQC
Oxford
Q-Step
Centre

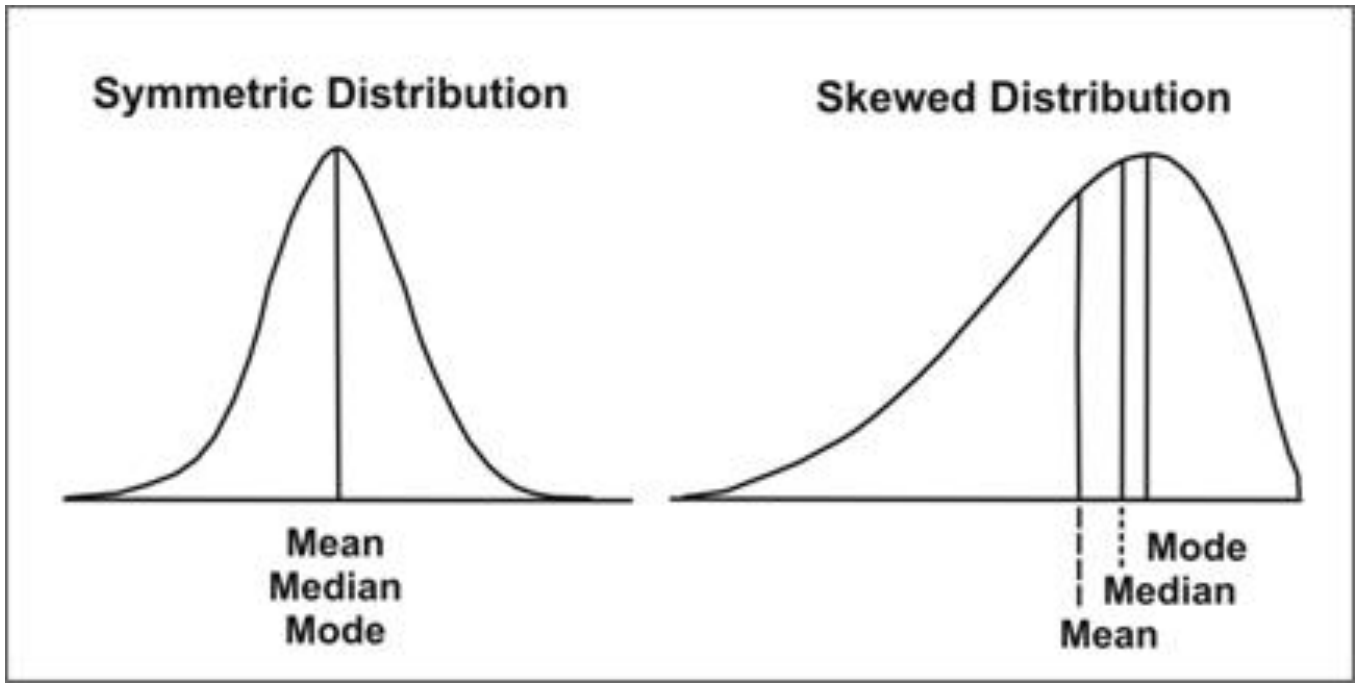
- Create a vector called x that contains the numbers 1 to 50
  - X <- 1:50
- Create a logical vector y that takes the value TRUE if x is smaller than 25
  - Y <- x<25
  - Y <- ifelse(x <25, TRUE, FALSE)
- Create a numeric vector that contains the words: My name is [your name]
  - My_name <- c("My", "name", "is", "name")
- How do you display all variable names for data-set cars
  - Names(cars)
- matrix <- matrix(1:12, nrow = 3, ncol = 4)
- party <- c("Conservative", "Labour", "LibDem")
- my_data <- data.frame(party, matrix)
- cnames <- c("Party", "Leadername", "Leader Resigned", "Voteshare", "Number MPs")
- colnames(my_data) <- cnames

Measures of Central Tendency

OQC
Oxford
Q-Step
Centre

- **Mean** = average response ; it is the sum of the measurements divided by the number of observations
  - it is affected by outliers
  - can only be calculated for numerical data (makes no sense for nominal data)
- **Median** = measurement that falls in the middle of the ordered sample
  - not affected by outliers
  - can be calculated for ordinal or interval data
- Note:
  - If a sample is perfectly symmetrical: mean = median
  - Otherwise we can speak of a "skewed" sample, a left skewed sample has a tail on the left - and the mean will lie to the left of the median (towards the skew / tail)

OQC
Oxford
Q-Step
Centre



Percentiles and Quantiles

OQC
Oxford
Q-Step
Centre

- **Percentiles** = the "p"th percentile is a number such that p% of the sample falls below it (and 100- p above) – thus the 10% percentile implies that 10% of the sample falls below this value.
  - Note that this makes the median the 50% percentile!]
- **Quantiles** = usually the "upper" and "lower" 25% of your sample, thus the point where 25% and 75% of your sample fall.
  - The median falls exactly between the upper and lower percentiles (at 50%) – a quarter of your sample falls below the lower quantile and a quarter falls above the upper quantile.

Variance and Standard Deviation

OQC
Oxford
Q-Step
Centre

- **Variance** = the average of the squared deviations.
  - i.e. it approximates the average of the distance from the mean (note we square it because we want the nominal distance, and by squaring the distances we get positive values only)
- **Standard deviation** = the square root of the variance.
  - i.e. the variance gives you the average of the 'squared distance from the mean' – which is difficult to interpret – thus by taking the square root of this value we get a measure of the average distance from the mean that is easier to interpret

→ measures of dispersion of the data – i.e. are the observations centred around the mean or is there much variance?

Pearson's Correlation Coefficient

OQC
Oxford
Q-Step
Centre

- A measure of the strength of the association / correlation between two continuous variables
- Linear association
- **NULL Hypothesis**: hypothesis of no effect; i.e. no association
- **ALTERNATIVE Hypothesis:** the hypothesis that contradicts the null hypothesis; i.e. there is an **association** between variable x and y.

REMEMBER: correlation != causation + does not capture non-linear relations

OQC
Oxford
Q-Step
Centre

Interpretation

- Correlation or Association = non-directional test;
  - Thus – we **do not test** the effect of x on y
  - Or – the **causal relation** between x and y
  - But – the co-variation of x and y

- When we run a correlation test :
  - Null hypothesis = no linear association between var1 and var2
  - Alternative hypothesis = there is a linear association between var1 and var2

Interpretation

OQC
Oxford
Q-Step
Centre

- Pearson's Coefficient can range between -1 and +1
  - -1 = complete negative correlation
  - 0 = no association
  - + 1 = complete positive correlation
- The size of the coefficient tells you the **strength** of the association
- **The P-value** tells you whether the observed test statistic [the correlation coefficient]is consistent with what we would expect given that the null hypothesis is true
- The **smaller the p-value** the more strongly the data contradicts the null hypothesis
  - Thus: small p-value; we reject the H0 hypothesis
  - We usually reject the null hypothesis when **$p \leq 0.05$**

Interpretation

OQC
Oxford
Q-Step
Centre

Pearson's product-moment correlation

data: Gov$hdi_2010 and Gov$women2010

t = 2.447, df = 34, p-value = 0.01973

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.06693071 0.63479253

sample estimates:

cor : 0.3869576

The smaller the p-value the less likely it is that H0 = true

Mostly P <0.05 = significant

Ha = correlation is **not** 0
(i.e. there is a correlation)

95% confidence interval around your correlation estimate

Strength of the association

Positive : 0.38

Homework Assignment

OQC
Oxford
Q-Step
Centre

- Create a scatterplot that show human development index (hdi_2010) on the x-axis and representation of women in parliament in 2010 (women2010) on the y-axis

- What is the correlation between these two variables? Is the correlation statistically significant?

OQC
Oxford
Q-Step
Centre

SEE YOU IN WEEK 6