EGS 2405- Geostatistics

Basic Statistics and Exploratory Data Analysis (EDA)

Lecturer: Mr H. Kipkulei, hkipkulei@jkuat.ac.ke
Technologist: Ms. Sarah Orado

Introduction

Recall from our previous class

Most **consequential decisions** of any **geostatistical** study are made early in the exploratory data analysis (EDA) – Which is our today's topic

Introduction

Survey/research data is structured as;

- Rectangular array (e.g., spreadsheet or database)
- One row/experimental subject and one column for each subject identifier
- Outcome variable, and explanatory variable.

```
1   4  2 1KE8   1   4  2   1 1296049 42022146844676101987105434431251405176420  1  1  1 11 1  102   100    1 5501 1  2 160  1   11 142 90   14 15221010000343131 7 11 0 1 6 1 1 11
1   4  2 2KE8   1   4  2   1 1296049 42022146844676101987105434431251405176420  1  1  1 11 1  102   100    1 5501 1  2 160  1   11 142 90   14 15221010000343131 7 11 0 1 6 1 1 11
1   4  2 3KE8   1   4  2   1 1296049 42022146844676101987105434431251405176420  1  1  1 11 1  102   100    1 5501 1  2 160  1   11 142 90   14 15221010000343131 7 11 0 1 6 1 1 11
1   4  2 4KE8   1   4  2   1 1296049 42022146844676101987105434431251405176420  1  1  1 11 1  102   100    1 5501 1  2 160  1   11 142 90   14 15221010000343131 7 11 0 1 6 1 1 11
1   7  2 1KE8   1   7  2   1 1296049 42022146844676121983100838511251405176420  1  1  1 11 1  102   100    1 5500 1  2  60  1   11 95     2 414996121110000343135 1  310 1 3 0 1 11
1  10  1 1KE8   1  10  1   1 1296049 42022146844677 11989106933411261405176420  1  1  1 11 1  103   100      5500 1  2  70  1   11 133211 814  5131110000353231 1  4 8 1 2 0 1 11
1  13  2 1KE8   1  13  2   1 1296049 42022146844679121982 99639511281405176420  1  1  1 11 1  904   100    1 5500 1  2  61  1   11 183152 412996121110000343135 1  310 1 8 1 1 11
1  13  2 2KE8   1  13  2   1 1296049 42022146844679121982 99639511281405176420  1  1  1 11 1  904   100    1 5500 1  2  61  1   11 183152 412996121110000343135 1  310 1 8 1 1 11
1  13  2 3KE8   1  13  2   1 1296049 42022146844679121982 99639511281405176420  1  1  1 11 1  904   100    1 5500 1  2  61  1   11 183152 412996121110000343135 1  310 1 8 1 1 11
1  13  2 4KE8   1  13  2   1 1296049 42022146844679121982 99639511281405176420  1  1  1 11 1  904   100    1 5500 1  2  61  1   11 183152 412996121110000343135 1  310 1 8 1 1 11
1  13  2 5KE8   1  13  2   1 1296049 42022146844679121982 99639511281405176420  1  1  1 11 1  904   100    1 5500 1  2  61  1   11 183152 412996121110000343135 1  310 1 8 1 1 11
1  20  2 1KE8   1  20  2   1 1296049 42022146844680 11992110530411291405176420  1  1  1 11 1  102   100    1 5501 1  2 140  1   11 123 31 81299612111001034333596 10 8 1 4 1 1 11
1  20  2 2KE8   1  20  2   1 1296049 42022146844680 11992110530411291405176420  1  1  1 11 1  102   100    1 5501 1  2 140  1   11 123 31 81299612111001034333596 10 8 1 4 1 1 11
1  26  2 1KE8   1  26  2   1 1296049 42022146844676111993112728311251405176420  1  1  1 11 1  103   100    1 5500 1  2  60  1   11 95     2 413996211010000353135 3  310 1 5 1 1 11
1  26  2 2KE8   1  26  2   1 1296049 42022146844676111993112728311251405176420  1  1  1 11 1  103   100    1 5500 1  2  60  1   11 95     2 413996211010000353135 3  310 1 5 1 1 11
1  26  2 3KE8   1  26  2   1 1296049 42022146844676111993112728311251405176420  1  1  1 11 1  103   100    1 5500 1  2  60  1   11 95     2 413996211010000353135 3  310 1 5 1 1 11
1  36  2 1KE8   1  36  2   1 1296049 42022146844677 31985102337511261405176420  1  1  1 11 1  103   100    1 5501 1  1  60  1   11 95     3 312996121110000353131 2  315 1 5 0 1 11
1  36  2 2KE8   1  36  2   1 1296049 42022146844677 31985102337511261405176420  1  1  1 11 1  103   100    1 5501 1  1  60  1   11 95     3 312996121110000353131 2  315 1 5 0 1 11
1  36  2 3KE8   1  36  2   1 1296049 42022146844677 31985102337511261405176420  1  1  1 11 1  103   100    1 5501 1  1  60  1   11 95     3 312996121110000353131 2  315 1 5 0 1 11
1  39  1 1KE8   1  39  1   1 1296049 42022146844676101990109031411251405176420  1  1  1 11 1  904   100    0 5500 1  2  90  1   11 153471 814180211100010353131 7  6 8 1 1 0 1 11
1  39  1 2KE8   1  39  1   1 1296049 42022146844676101990109031411251405176420  1  1  1 11 1  904   100    0 5500 1  2  90  1   11 153471 814180211100010353131 7  6 8 1 1 0 1 11
1  42  1 1KE8   1  42  1   1 1296049 42022146844677 41992110830411261405176420  1  1  1 11 2  904   100    2 5501 1  2  60  1   11  63152 414998211110000353131 4  310 1 3 1 1 11
1  52  1 1KE8   1  52  1   1 1296049 42022146844678101982 99439511271405176420  1  1  1 11 1  103   100      5500 1  2 960  1   11 95     1 814  5121000000343231 7 12 8 1 2 0 1 11
1  52  1 2KE8   1  52  1   1 1296049 42022146844678101982 99439511271405176420  1  1  1 11 1  103   100      5500 1  2 960  1   11 95     1 814  5121000000343231 7 12 8 1 2 0 1 11
1  55  2 1KE8   1  55  2   1 1296049 42022146844676 81987105234411251405176420  1  1  1 11 1  103   100    1 5501 1  2  60  1   11 95     2 414  5211100010353431 2  310 1 4 2 1 11
1  55  2 2KE8   1  55  2   1 1296049 42022146844676 81987105234411251405176420  1  1  1 11 1  103   100    1 5501 1  2  60  1   11 95     2 414  5211100010353431 2  310 1 4 2 1 11
1  65  2 1KE8   1  65  2   1 1296049 42022146844677 91997117324211261405176420  1  1  1 11 1  904   100    1 5500 1  2  60  1   11 95     2 414240211100000353131 1  310 1 3 1 1 11
1  68  2 1KE8   1  68  2   1 1296049 42022146844677 11983 99739511261405176420  1  1  1 11 1  103   100    1 5500 1  2  60  1   11 153152 414  5131110000353431 3  310 1 5 0 2 11
1  68  2 2KE8   1  68  2   1 1296049 42022146844677 11983 99739511261405176420  1  1  1 11 1  103   100    1 5500 1  2  60  1   11 153152 414  5131110000353431 3  310 1 5 0 2 11
1  68  2 3KE8   1  68  2   1 1296049 42022146844677 11983 99739511261405176420  1  1  1 11 1  103   100    1 5500 1  2  60  1   11 153152 414  5131110000353431 3  310 1 5 0 2 11
1  72  1 1KE8   1  72  1   1 1296049 42022146844678 71981 97940611271405176420  1  1  1 11 1  904   100      5501 1  2  60  1   11  53152 114120211110000343131 4  3 7 1 3 0 1 11
1  72  1 2KE8   1  72  1   1 1296049 42022146844678 71981 97940611271405176420  1  1  1 11 1  904   100      5501 1  2  60  1   11  53152 114120211110000343131 4  3 7 1 3 0 1 11
1  72  1 3KE8   1  72  1   1 1296049 42022146844678 71981 97940611271405176420  1  1  1 11 1  904   100      5501 1  2  60  1   11  53152 114120211110000343131 4  3 7 1 3 0 1 11
1  78  2 1KE8   1  78  2   1 1296049 42022146844677111993112728311261405176420  1  1  1 11 1  102   100    1 5500 1  2  60  1   11 103173 414 20221110000343131 2  316 1 3 1 1 11
1  81  2 1KE8   1  81  2   1 1296049 42022146844677 31983 99939511261405176420  1  1  1 11 1  904   100    1 5501 1  2  60  1   11 95     2 414180211010000353131 1  310 1 3 0 1 11
1  81  2 2KE8   1  81  2   1 1296049 42022146844677 31983 99939511261405176420  1  1  1 11 1  904   100    1 5501 1  2  60  1   11 95     2 414180211010000353131 1  310 1 3 0 1 11
```

DHS data 2022

Introduction

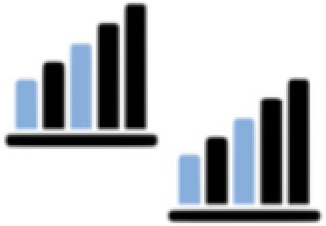
- Examining the spread sheet by eye is a tedious, boring and overwhelming process
- Exploratory data analysis (EDA) can be used to aid the situation

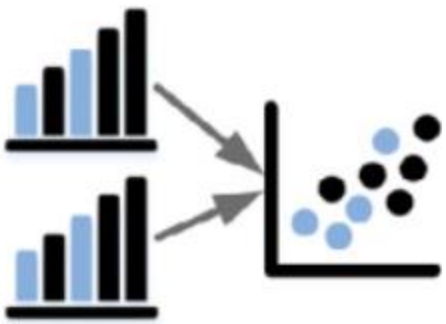
```
1  4 2 1KE8  1  4 2  1 1296049 42022146844676101987105434431251405176420  1  1  1 11 1  102  100  1 5501 1 2 160  1  11 142 90  14 1522101000343131 7 11 0 1 6 1 1 11
1  4 2 2KE8  1  4 2  1 1296049 42022146844676101987105434431251405176420  1  1  1 11 1  102  100  1 5501 1 2 160  1  11 142 90  14 1522101000343131 7 11 0 1 6 1 1 11
1  4 2 3KE8  1  4 2  1 1296049 42022146844676101987105434431251405176420  1  1  1 11 1  102  100  1 5501 1 2 160  1  11 142 90  14 1522101000343131 7 11 0 1 6 1 1 11
1  4 2 4KE8  1  4 2  1 1296049 42022146844676101987105434431251405176420  1  1  1 11 1  102  100  1 5501 1 2 160  1  11 142 90  14 1522101000343131 7 11 0 1 6 1 1 11
1  7 2 1KE8  1  7 2  1 1296049 42022146844676121983100838511251405176420  1  1  1 11 1  102  100  1 5500 1 2  60  1  11 95   2 4149961211100003431 35 1  310 1 3 0 1 11
1 10 1 1KE8  1 10 1  1 1296049 4202214684467 7 119891069334112614051764 20  1  1  1 11 1  103  100    5500 1 2  70  1  11 133211 814  5131110000353231 1  4 8 1 2 0 1 11
1 13 2 1KE8  1 13 2  1 1296049 42022146844679121982 996395112814051764 20  1  1  1 11 1  904  100  1 5500 1 2  61  1  11 183152 412996121110000343135 1  310 1 8 1 1 11
1 13 2 2KE8  1 13 2  1 1296049 42022146844679121982 996395112814051764 20  1  1  1 11 1  904  100  1 5500 1 2  61  1  11 183152 412996121110000343135 1  310 1 8 1 1 11
1 13 2 3KE8  1 13 2  1 1296049 42022146844679121982 996395112814051764 20  1  1  1 11 1  904  100  1 5500 1 2  61  1  11 183152 412996121110000343135 1  310 1 8 1 1 11
1 13 2 4KE8  1 13 2  1 1296049 42022146844679121982 996395112814051764 20  1  1  1 11 1  904  100  1 5500 1 2  61  1  11 183152 412996121110000343135 1  310 1 8 1 1 11
1 13 2 5KE8  1 13 2  1 1296049 42022146844679121982 996395112814051764 20  1  1  1 11 1  904  100  1 5500 1 2  61  1  11 183152 412996121110000343135 1  310 1 8 1 1 11
1 20 2 1KE8  1 20 2  1 1296049 4202214684468 0 11992110530411291405176420  1  1  1 11 1  102  100  1 5501 1 2 140  1  11 123 31 81299612111001034333596 10 8 1 4 1 1 11
1 20 2 2KE8  1 20 2  1 1296049 4202214684468 0 11992110530411291405176420  1  1  1 11 1  102  100  1 5501 1 2 140  1  11 123 31 81299612111001034333596 10 8 1 4 1 1 11
1 26 2 1KE8  1 26 2  1 1296049 42022146844676111993112728311251405176420  1  1  1 11 1  103  100  1 5500 1 2  60  1  11 95   2 413996211010000353135 3  310 1 5 1 1 11
1 26 2 2KE8  1 26 2  1 1296049 42022146844676111993112728311251405176420  1  1  1 11 1  103  100  1 5500 1 2  60  1  11 95   2 413996211010000353135 3  310 1 5 1 1 11
1 26 2 3KE8  1 26 2  1 1296049 42022146844676111993112728311251405176420  1  1  1 11 1  103  100  1 5500 1 2  60  1  11 95   2 413996211010000353135 3  310 1 5 1 1 11
1 36 2 1KE8  1 36 2  1 1296049 4202214684467 7 31985102337511261405176420  1  1  1 11 1  103  100  1 5501 1 1  60  1  11 95   3 312996121110000353131 2  315 1 5 0 1 11
1 36 2 2KE8  1 36 2  1 1296049 4202214684467 7 31985102337511261405176420  1  1  1 11 1  103  100  1 5501 1 1  60  1  11 95   3 312996121110000353131 2  315 1 5 0 1 11
1 36 2 3KE8  1 36 2  1 1296049 4202214684467 7 31985102337511261405176420  1  1  1 11 1  103  100  1 5501 1 1  60  1  11 95   3 312996121110000353131 2  315 1 5 0 1 11
1 39 1 1KE8  1 39 1  1 1296049 42022146844676101990109031411251405176420  1  1  1 11 1  904  100  0 5500 1 2  90  1  11 153471 814180211100010353131 7  6 8 1 1 0 1 11
1 39 1 2KE8  1 39 1  1 1296049 42022146844676101990109031411251405176420  1  1  1 11 1  904  100  0 5500 1 2  90  1  11 153471 814180211100010353131 7  6 8 1 1 0 1 11
1 42 1 1KE8  1 42 1  1 1296049 4202214684467 7 41992110830411261405176420  1  1  1 11 2  904  100  2 5501 1 2  60  1  11  63152 414998211110000353131 4  310 1 3 1 1 11
1 52 1 1KE8  1 52 1  1 1296049 42022146844678101982 994395112714051764 20  1  1  1 11 1  103  100    5500 1 2 960  1  11 95   1 814  5121000000343231 7 12 8 1 2 0 1 11
1 52 1 2KE8  1 52 1  1 1296049 42022146844678101982 994395112714051764 20  1  1  1 11 1  103  100    5500 1 2 960  1  11 95   1 814  5121000000343231 7 12 8 1 2 0 1 11
1 55 2 1KE8  1 55 2  1 1296049 4202214684467 6 81987105234411251405176420  1  1  1 11 1  103  100  1 5501 1 2  60  1  11 95   2 414  5211100010353431 2  310 1 4 2 1 11
1 55 2 2KE8  1 55 2  1 1296049 4202214684467 6 81987105234411251405176420  1  1  1 11 1  103  100  1 5501 1 2  60  1  11 95   2 414  5211100010353431 2  310 1 4 2 1 11
1 65 2 1KE8  1 65 2  1 1296049 4202214684467 7 91997117324211261405176420  1  1  1 11 1  904  100  1 5500 1 2  60  1  11 95   2 414240211100000353131 1  310 1 3 1 1 11
1 68 2 1KE8  1 68 2  1 1296049 4202214684467 7 11983 99739511261405176420  1  1  1 11 1  103  100  1 5500 1 2  60  1  11 153152 414  5131110000353431 3  310 1 5 0 2 11
1 68 2 2KE8  1 68 2  1 1296049 4202214684467 7 11983 99739511261405176420  1  1  1 11 1  103  100  1 5500 1 2  60  1  11 153152 414  5131110000353431 3  310 1 5 0 2 11
1 68 2 3KE8  1 68 2  1 1296049 4202214684467 7 11983 99739511261405176420  1  1  1 11 1  103  100  1 5500 1 2  60  1  11 153152 414  5131110000353431 3  310 1 5 0 2 11
1 72 1 1KE8  1 72 1  1 1296049 4202214684467 8 71981 97940611271405176420  1  1  1 11 1  904  100    5501 1 2  60  1  11  53152 114120211110000343131 4  3 7 1 3 0 1 11
1 72 1 2KE8  1 72 1  1 1296049 4202214684467 8 71981 97940611271405176420  1  1  1 11 1  904  100    5501 1 2  60  1  11  53152 114120211110000343131 4  3 7 1 3 0 1 11
1 72 1 3KE8  1 72 1  1 1296049 4202214684467 8 71981 97940611271405176420  1  1  1 11 1  904  100    5501 1 2  60  1  11  53152 114120211110000343131 4  3 7 1 3 0 1 11
1 78 2 1KE8  1 78 2  1 1296049 42022146844677111993112728311261405176420  1  1  1 11 1  102  100  1 5500 1 2  60  1  11 103173 414 2022111000034313 1 2  316 1 3 1 1 11
1 81 2 1KE8  1 81 2  1 1296049 4202214684467 7 31983 99939511261405176420  1  1  1 11 1  904  100  1 5501 1 2  60  1  11 95   2 414180211010000353131 1  310 1 3 0 1 11
1 81 2 2KE8  1 81 2  1 1296049 4202214684467 7 31983 99939511261405176420  1  1  1 11 1  904  100  1 5501 1 2  60  1  11 95   2 414180211010000353131 1  310 1 3 0 1 11
```

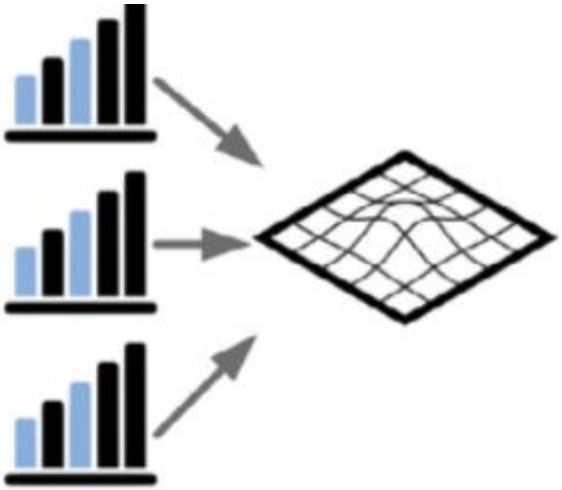
DHS data 2022

What is EDA

- Approach of **analyzing** data sets to summarize their main characteristics, often with visual methods.
- EDA helps to gain insight into a dataset before doing any formal statistical modelling/hypothesis testing







Purpose of EDA

- **Uncovers** the underlying structure of the dataset
- **Identifies** important variables
- **Detects** outliers and anomalies
- **Tests** underlying assumptions
- **Determines** relevant variables, their transformations, and interaction among variables with respect to the model to be built.
- **Highlights** missing data as may be relevant to building desired models.

Classifications of EDA

- EDA
  - Non-Graphical
    - Univariate
    - Multivariate (or usually bivariate)
  - Graphical
    - Univariate
    - Multivariate (or usually bivariate)

Univariate Non-graphical EDA

Univariate non-graphical EDA helps to:

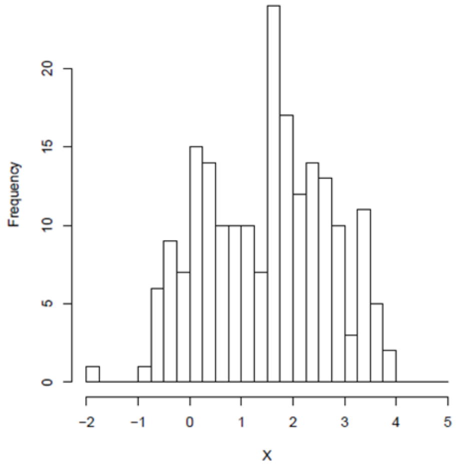
- Better appreciate the “sample distribution”
- Make some tentative conclusions about what population distribution(s) is/are compatible with the sample distribution

Categorical data

Statistic/College	H&SS	MCS	SCS	other	Total
Count	5	6	4	5	20
Proportion	0.25	0.30	0.20	0.25	1.00
Percent	25%	30%	20%	25%	100%

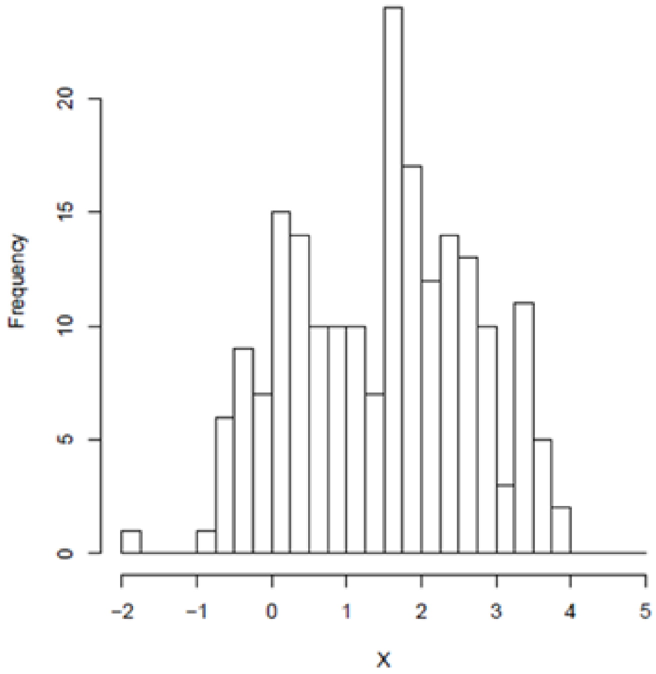
Count, proportion, percentage of students taking different subjects

Quantitative data



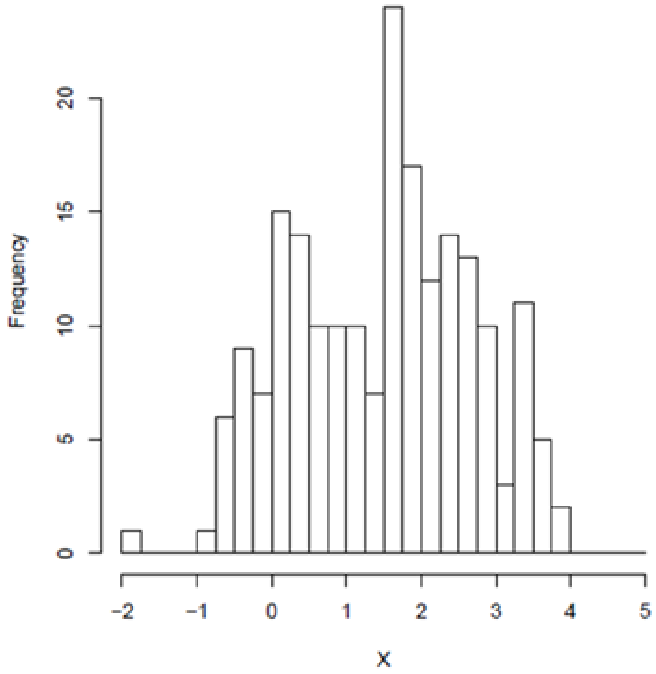
Univariate Non-graphical EDA

- Sample's distributional characteristics are seen qualitatively in the univariate graphical EDA technique of a histogram
- Think of univariate non-graphical EDA as telling you about aspects of the histogram of the distribution of the variable of interest for example SOC



Univariate Non-graphical EDA

- Sample's distributional characteristics are seen qualitatively in the univariate graphical EDA technique of a histogram
- Think of univariate non-graphical EDA as telling you about aspects of the histogram of the distribution of the variable of interest for example SOC



The central tendency

- Measures of central tendency provide information about where the center of a distribution is located.

- The most commonly used measures of center for numerical data are the
  - ❖ Mean
  - ❖ Median
  - ❖ Mode

The central tendency

- The mean is the simple arithmetic average:
- The sum of the values of a variable divided by the number of observations (n)

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

where

- n is the total number of observations
- x_i is the score of the ith observation
- Σ is the symbol of summation (pronounced sigma)
- $\bar{x}$ is the sample mean value

The central tendency

- The **mean** is the simple arithmetic average:
- The sum of the values of a variable divided by the number of observations (n)

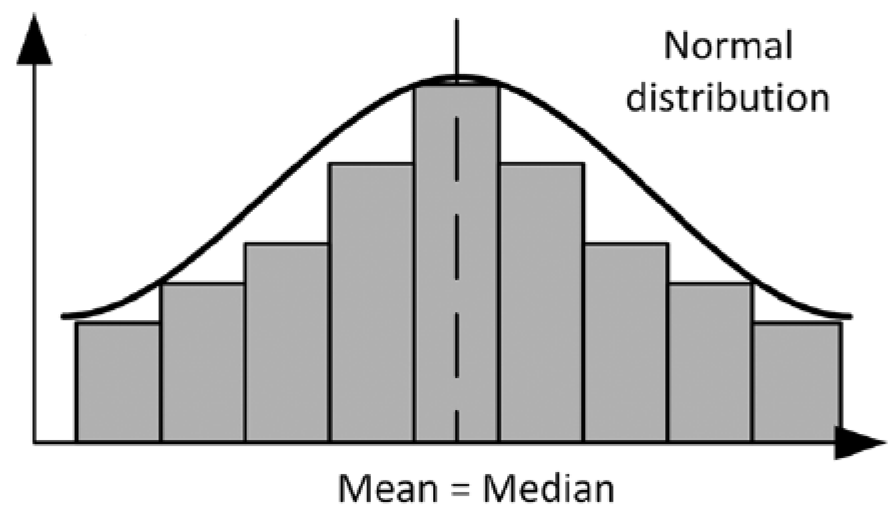
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{Equation (2.1)}$$

where

- n is the total number of observations
- x_i is the score of the ith observation
- Σ is the symbol of summation (pronounced sigma)
- $\bar{x}$ is the sample mean value

The central tendency

- The **median** is the value that divides the sorted scores form smaller to larger in half. It is a measure of center.
- Relevant for interval/ratio and ordinal
- The median overcomes the outlier problem
- When n is odd; single median. When n is even, there are two “middle values,”. Average is taken

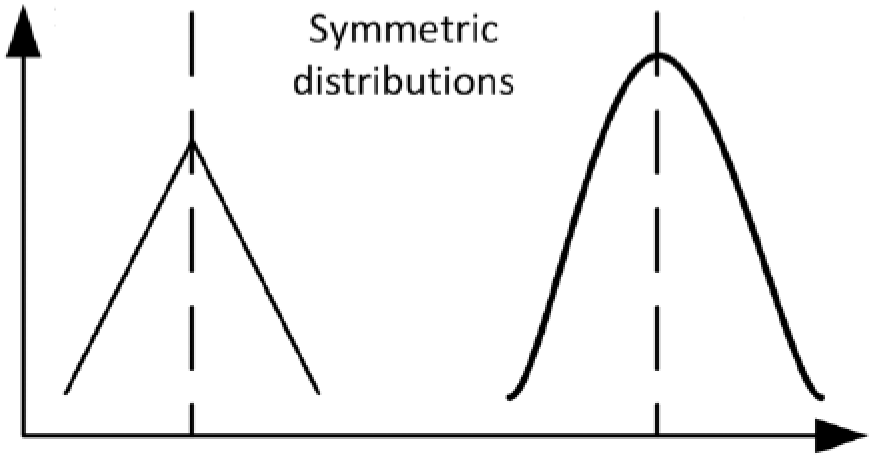


The median is located in the center of a normal distribution and coincides with the mean

The central tendency

- The mode is the most typical value.
- It implies that the frequency distribution has a single
- peak.
- For a **symmetric** distribution the mode, the mean and the median are in principle the same.
- For an **asymmetric** one

(mode - median) ~ 2 * (median - mode)



Measures of dispersion

- Measures of spread/variability/variation/diversity or dispersion provide information on how much the values of a variable differ among themselves and in relation to the mean.
- The most common measures are as follows (de Smith 2018):
    - ❖ Range
    - ❖ Deviation from the mean
    - ❖ Variance
    - ❖ Standard deviation
    - ❖ Standard distance
    - ❖ Percentiles and quartiles

Example with Meuse data

```
install.packages("sp")
library(sp)
data(meuse)
summary(meuse)
head(meuse)
dim(meuse) #To check the data dimension (RowsxColumns)
str(meuse) #To check the structure of the data (Continuous vs categorical)
mean(meuse$cadmium)
median(meuse$copper)
# define mode() function
mode = function() {
  # calculate mode of the copper concentration
  return(names(sort(-table(meuse$copper)))[1])
}
# call mode() function
mode()
```

Measures of dispersion

- A **range** is the difference between the largest and smallest values of the variable studied.

Range = xmax-xmin Equation (2.2)

- **Deviation** from the mean is the subtraction of the mean from each score, **Deviation**=$(x_i - \bar{x})$ Equation (2.3)
- The sum of all deviations is **zero** (sometimes, due to rounding up, the sum is very close to zero)

Measures of dispersion

Variance and standard deviation

- The variance of a set of values, which we denote S^2, is by definition

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2 \qquad \text{Equation (2.4)}$$

- The variance is the second moment about the mean.
- Like the mean, it is based on all of the observations, it can be treated algebraically, and it is little affected by sampling fluctuations.
- It is both additive and positive.
- Its square root is the **standard deviation, S**.
- To estimate population variance σ^2 from a sample, then N in Equation (2.4) is replaced by N-1.

Measures of dispersion

Coefficient of variation

- The SD is regarded in relative terms (relative variability).
- The CV is particularly useful when you want to compare results from **two different surveys** or tests that have different measures or values (hence different means)
  - CV = $(SD/\bar{x}) * 100$

Equation (2.5)

Measures of dispersion in R with Meuse data

```
var(meuse$copper)
sd(meuse$copper)

#anotherway of computing the SD

sqrt(var(meuse$copper))

#Coefficient of variance for copper conc in the region

((sd(meuse$copper))/(mean(meuse$copper)))*100
```

Measures of shape

Measures of shape describe how values (e.g., frequencies) are distributed across the intervals (bins) and are measured by **skewness** and **kurtosis**.

Skewness

The skewness measures the asymmetry of the observations. It is defined formally from the third moment about the mean:

$$m_3 = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^3 \qquad \text{Equation (2.6)}$$

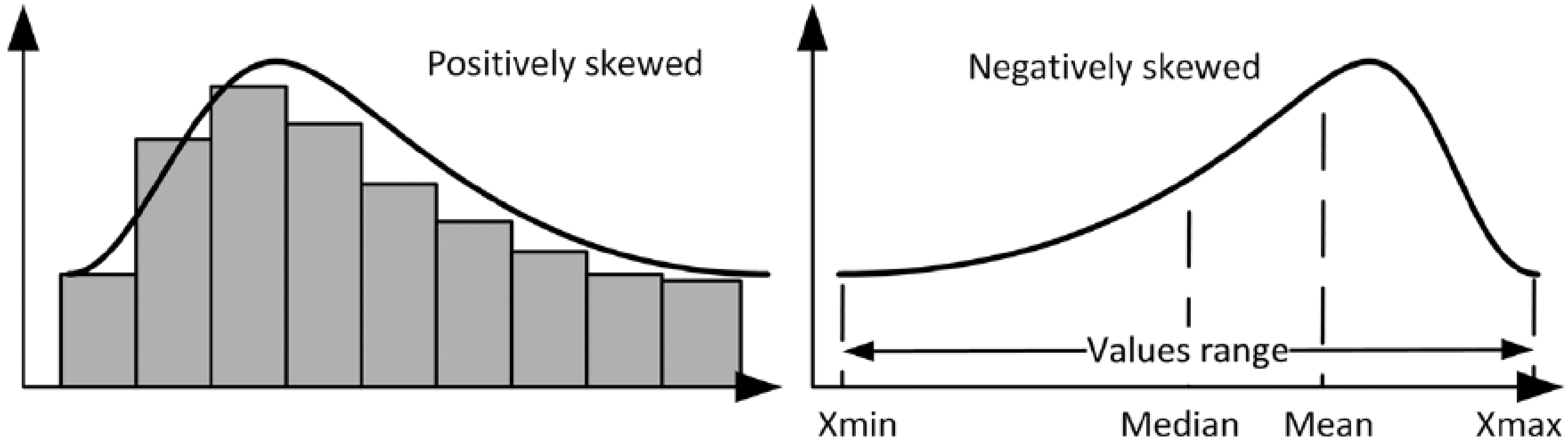
Measures of shape

- The coefficient of skewness is then

$$g_1 = \frac{m_3}{m_2\sqrt{m_2}} = \frac{m_3}{S^3} \qquad \text{Equation (2.7)}$$

- where m2 is the variance and S the standard deviation. Symmetric distributions have g1 = 0. N
- **Negative skewness** indicates that the mean of the data values is less than the median, and the data distribution is left-skewed.
- **Positive skewness** would indicate that the mean of the data values is larger than the median, and the data distribution is right-skewed

Measures of shape



Positively and Negatively skewed distributions

Measures of shape

- Kurtosis is obtained from the fourth moment about the mean

$$m_4 = \frac{1}{N}\sum_{i-1}^{N}(z_i - \bar{z})^4$$

Equation (2.8)

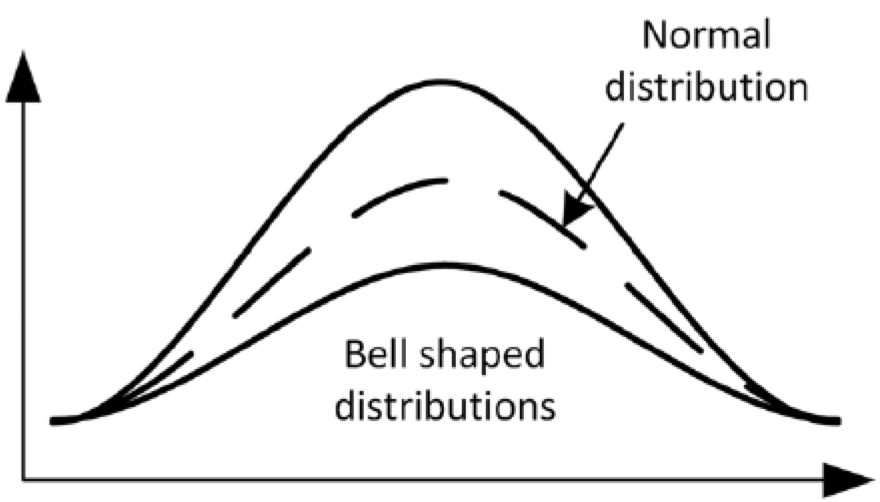
- The coefficient of Kurtosis is given by

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{m_4}{(S^2)^2} - 3.$$

Equation (2.9)

Measures of shape

- **Kurtosis**, from the graphical inspection perspective, is the degree of the peakedness or flatness of a distribution.
- A **zero kurtosis** (g=0) indicates a near-normal distribution peakedness.
- A negative kurtosis indicates a more flat distribution (lower than normal) (g<0)
- A **positive** kurtosis reveals a distribution with a higher peak than the normal distribution (g>0)



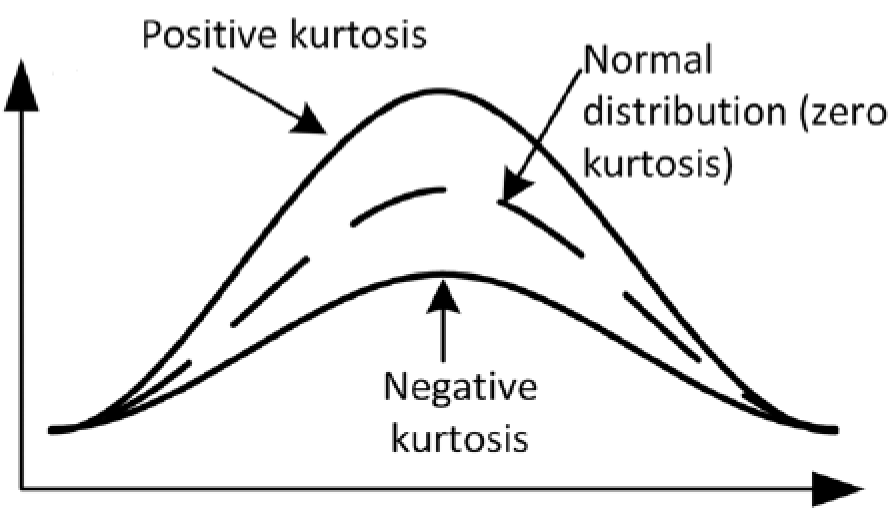
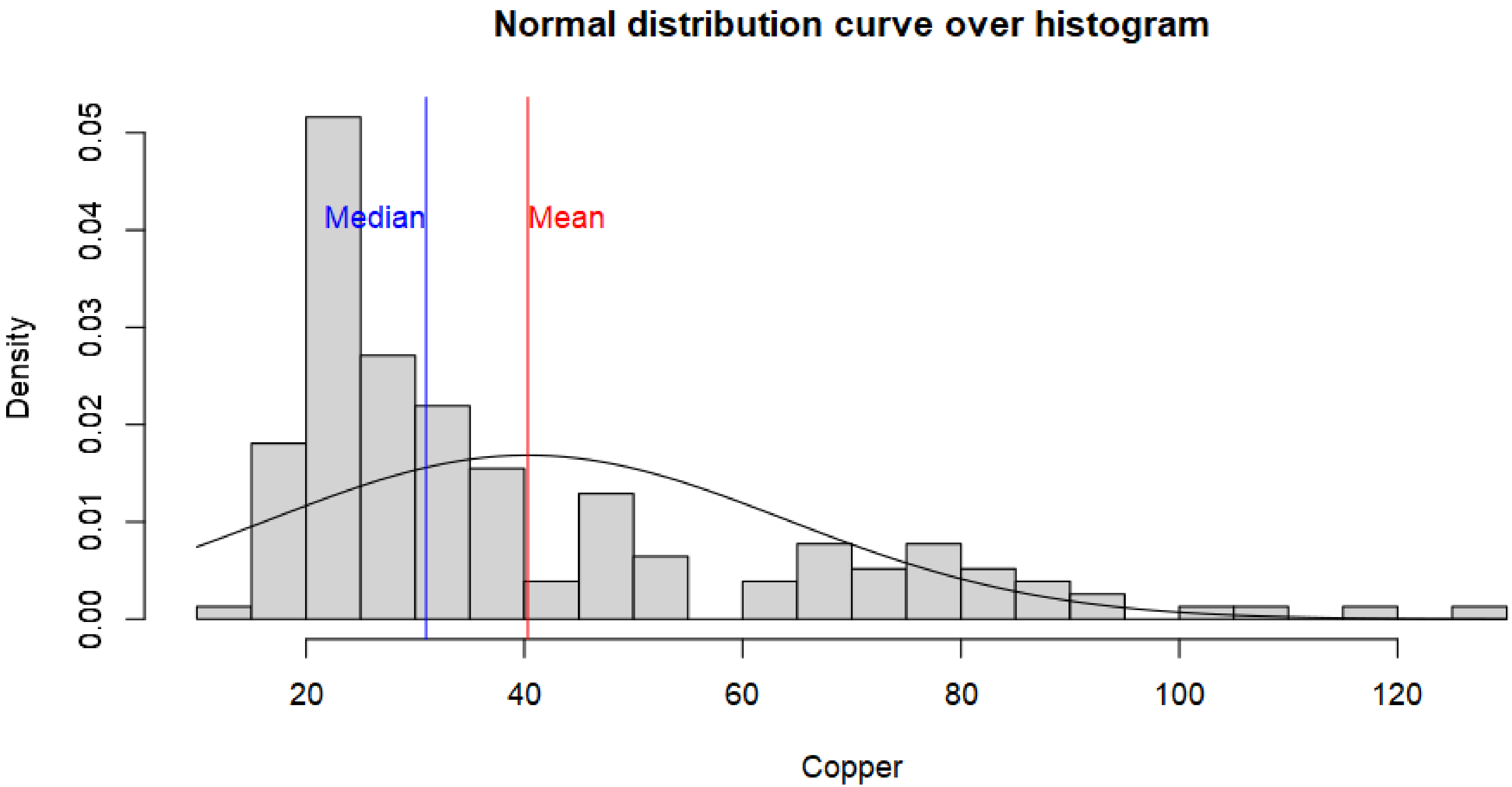


Illustration of dispersion in R with Meuse data



Measures of shape in R with Meuse data

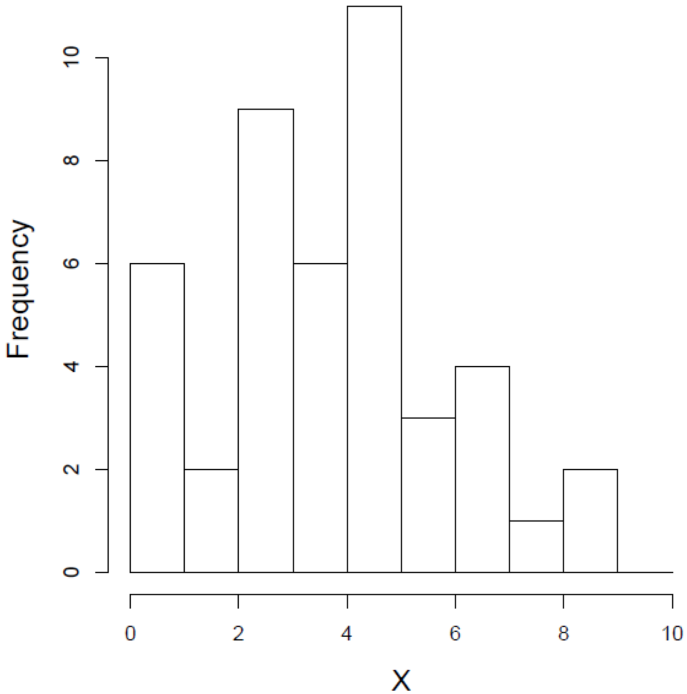
```
#Measures of shape
library(pacman)

pacman::p_load(e1071)
kurtosis(meuse$copper)

hist(meuse$copper, prob=TRUE,breaks=20,main="Normal distribution
curve over histogram", xlab= "Copper")
pacman::p_load(e1071)
pacman::p_load(e1071)
curve(dnorm(x, mean=mean(meuse$copper), sd=sd(meuse$copper)),
add=TRUE)
abline(v=mean(meuse$copper), col="red")
text(mean(meuse$copper),0.04,"Mean", col = "red", adj = c(0, -.1))
abline(v=median(meuse$copper), col="blue")
text(median(meuse$copper),0.04,"Median", col = "blue", adj = c(1, -.1))
```

Univariate graphical EDA

- It involves visualizing graphically at the distribution of the sample
- Graphical methods are more qualitative and involve a degree of subjective analysis.

Histograms

- The only one of these techniques that makes sense for categorical data is the histogram (basically just a bar plot of the tabulation of the data).

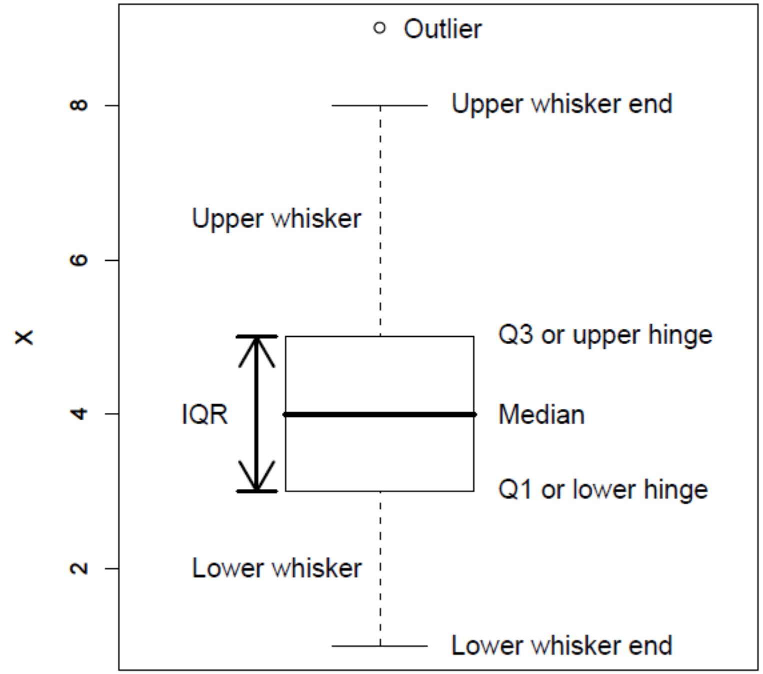


- Histograms are one of the best ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.

Univariate graphical EDA

Box-plots

- Boxplots are very good at presenting information about the central tendency, symmetry and skew, as well as outliers
- They can be misleading about aspects such as multimodality. One of the best uses of boxplots is in the form of side-by-side boxplots

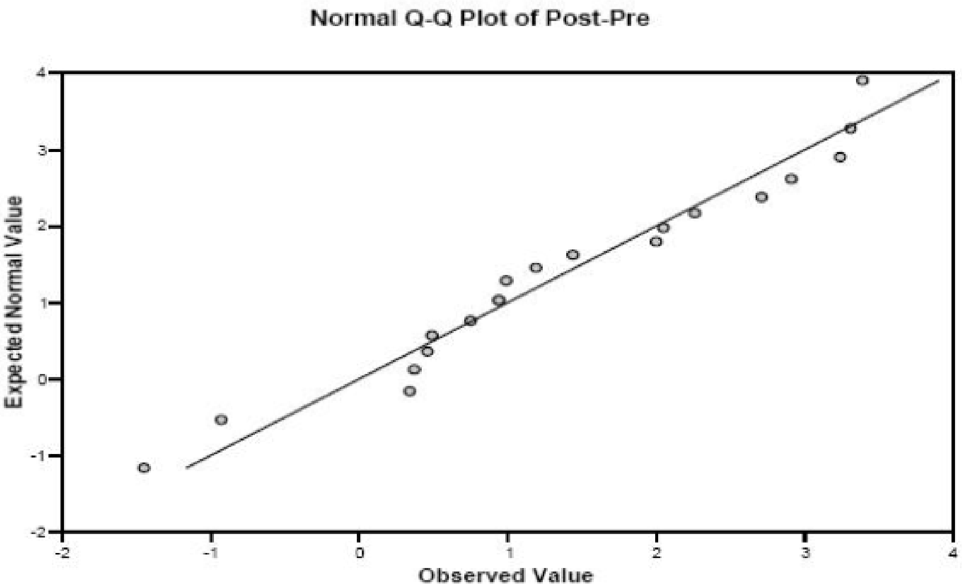


- Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers

Univariate graphical EDA

Quantile-normal plots

- QN plot or more generality the or **QQ plot**.
- Used to see how well a sample of data of size n matches a **Gaussian distribution** with mean and variance equal to the sample mean and variance.
- The quantile-normal plot we can detect left or right skew, positive or negative kurtosis, and bimodality.



Quantile-Normal plots allow detection of non-normality and diagnosis of skewness and kurtosis.

Multivariate non-graphical EDA

- Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either **cross-tabulation** or statistics.

Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

Sample Data for Cross-tabulation

Cross-tabulation for sample data

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Multivariate non-graphical EDA

Covariance and correlation

- The joint dispersion of two variables, z1 and z2, is termed as covariance C1;2.
- Covariance for a finite set of observations can be expressed as:

$$C_{1,2} = \frac{1}{N} \sum_{i=1}^{N} \{(z_1 - \bar{z_1})(z_2 - \bar{z_2})\} \qquad \text{Equation (2.10)}$$

- where $\bar{z_1}$ and $\bar{z_2}$ are the means of the two variables. This expression is analogous to the variance of a finite set of observations.

Covariance and correlation

- Covariance is affected by the scales on which the properties have been measured.
- This makes comparisons between different pairs of variables and sets of observations difficult unless measurements are on the same scale

- Pearson product-moment correlation coefficient, or simply the **correlation coefficient**, is often preferred. It refers specifically to linear correlation and it is a dimensionless value

Multivariate non-graphical EDA

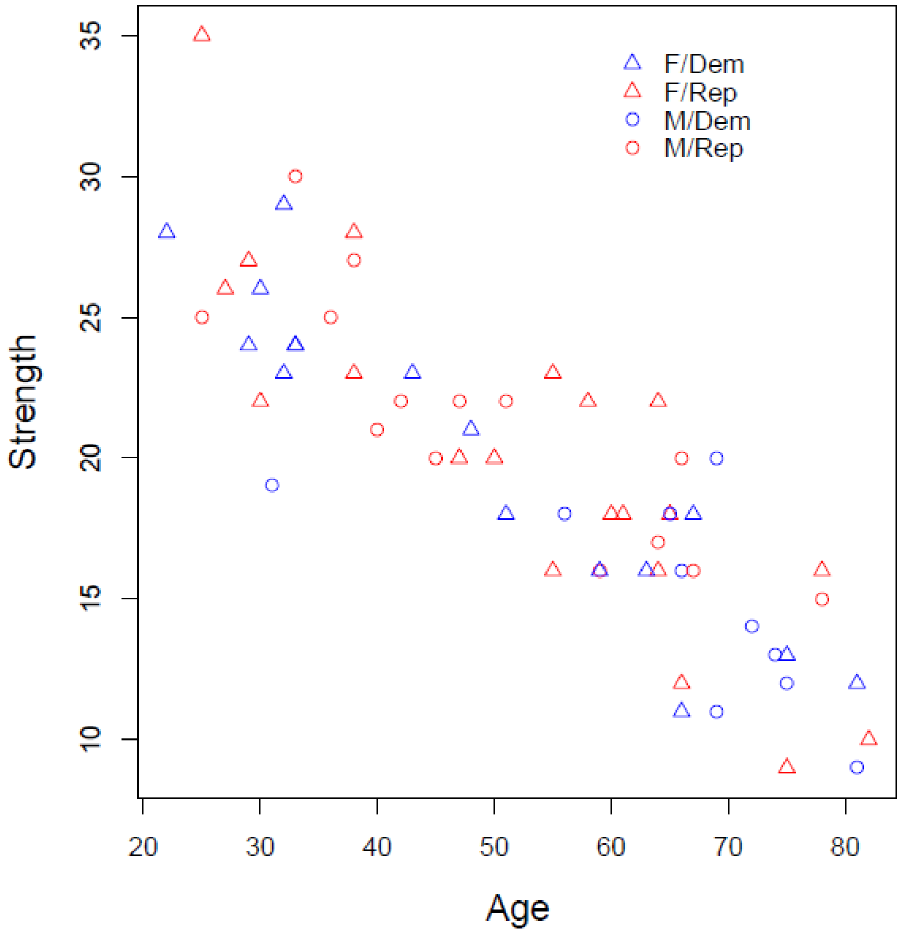
- The **correlation coefficient**, is given as.

$$\rho = \frac{C_{1,2}}{S_1 S_2} \qquad \text{Equation (2.11)}$$

- CC ranges between 1 and -1.
- If units with large values of one variable also have large values of the other then the two variables are **positively correlated**, $\rho>0$;
- if the large values of the one are matched by small values of the other then the two are **negatively correlated**, $\rho< 0$.
- If $\rho=0$ then there is **no linear relation**.

Multivariate graphical EDA

- For two quantitative variables, the basic graphical EDA technique is the **scatterplot**
- It has one variable on the x-axis, and one on the y-axis and a point for each case of the dataset.
- If one variable is **explanatory** and the other is **outcome**, it is a very, very strong convention to put the outcome on the y (vertical) axis



The normal distribution

- Normal distribution (also called **Gaussian distribution**) is the most commonly used distribution in statistics, as many physical phenomena are normally distributed (e.g., human weight and height).
- In a normal distribution, the values of a variable are more likely to be **closer** to the **mean**, while larger or smaller scores have low probabilities of occurring
- Many environmental variables, such as of the soil, are distributed in a way that approximates the normal distribution

The normal distribution

- Normal distribution is defined for a continuous random variable X in terms of the probability density function (pdf), f(x), as

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{Equation (2.12)}$$

where

μ is the mean of the population

σ is the population standard deviation

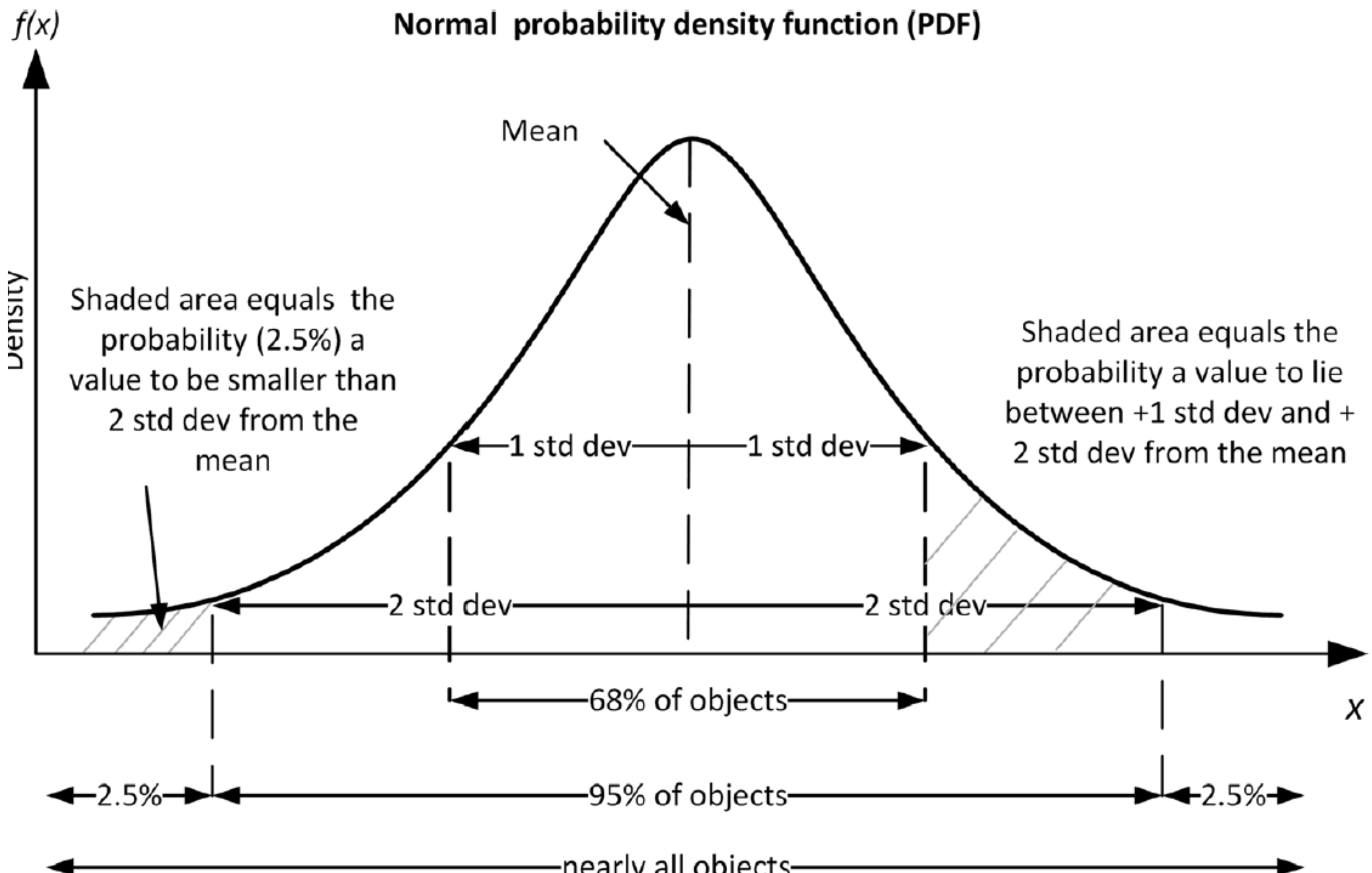
σ^2 is the population variance

x is the value of the variable

The normal distribution

- Normal distribution is **continuous** and **symmetrical**, with its peak at the mean of the distribution
- The ordinate f(x) at any given value of x is the probability density at x
- The total area under the curve is 1, the total probability of the distribution.
- The area under any portion of the curve, say between z1 and z2, represents the proportion of the distribution lying in that range

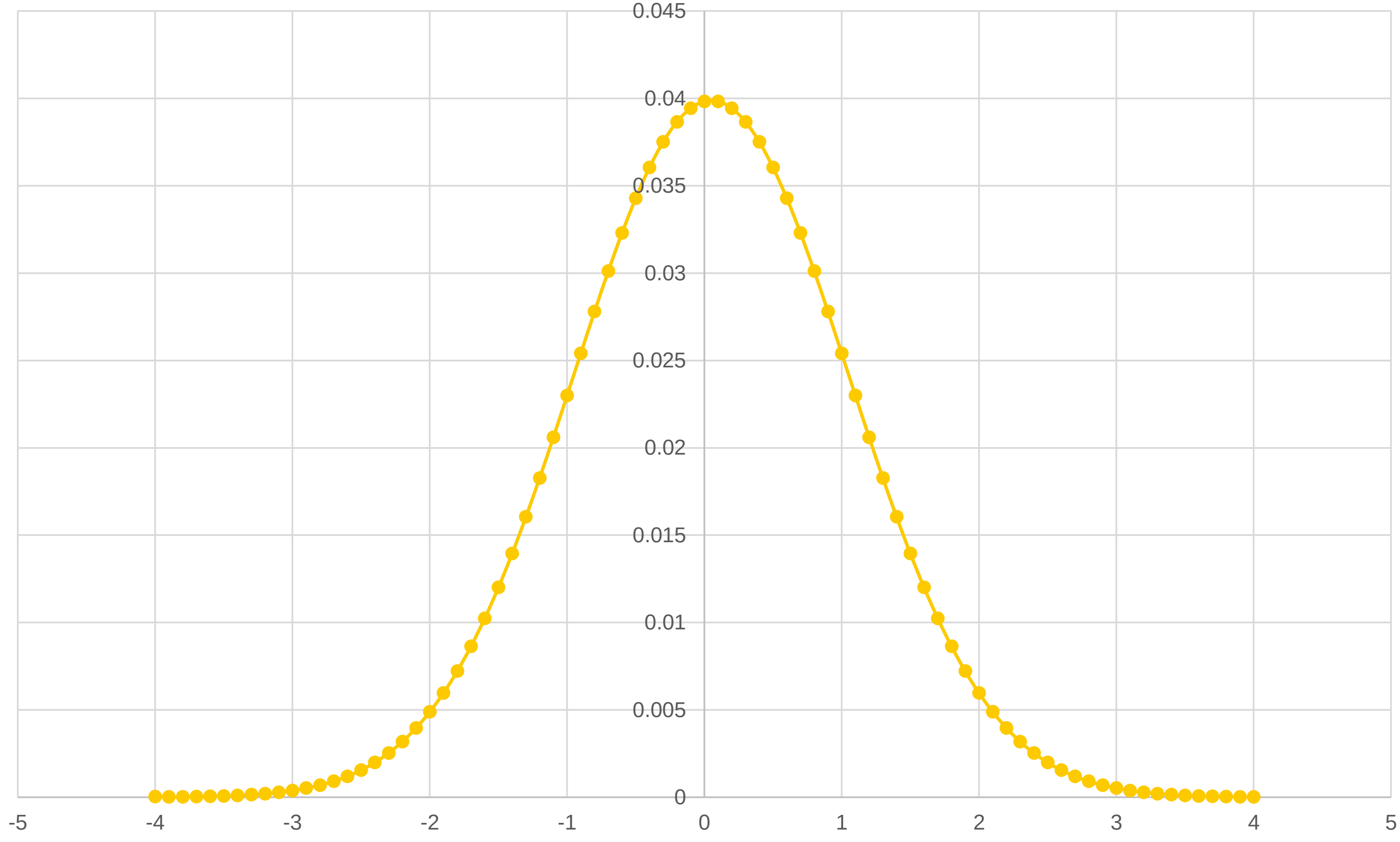
The normal distribution



The normal distribution

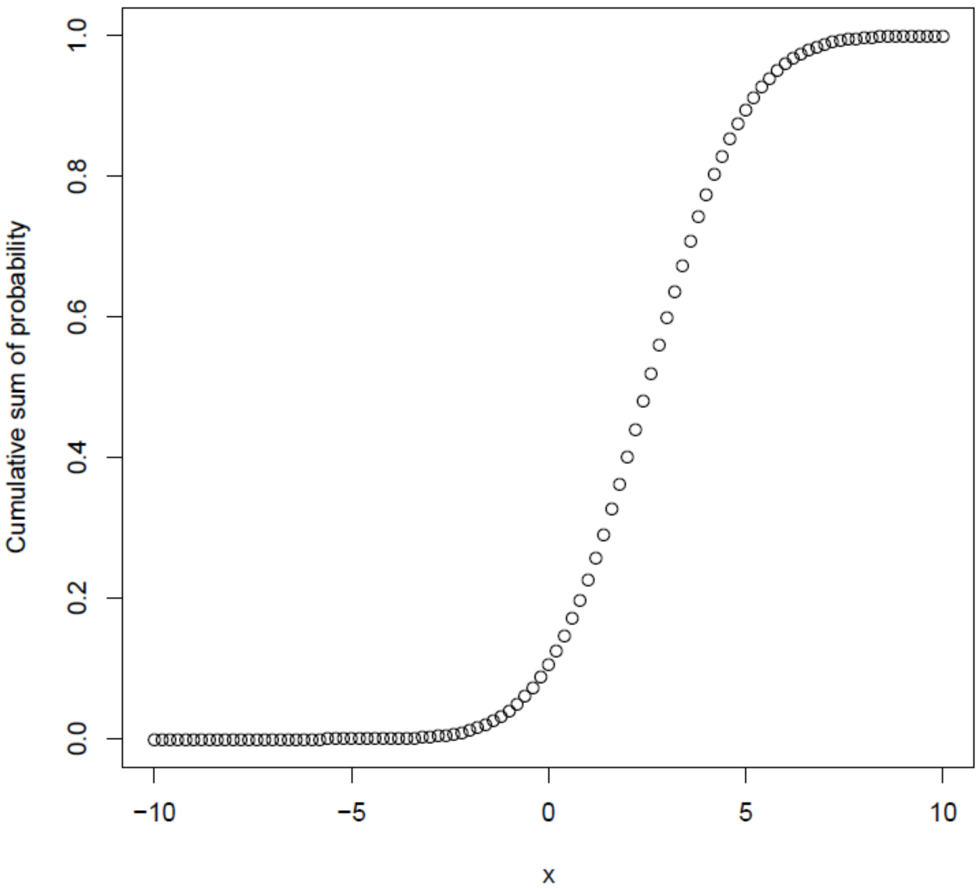
- Slightly more than two-thirds of the distribution lies within one standard deviation of the mean, i.e. between $\mu-\sigma$ AND $\mu+\sigma$;
- About 95% lies in the range $\mu-2\sigma$ AND $\mu+2\sigma$;
- About 99.73% lies within three standard deviations of the mean $\mu-3\sigma$ AND $\mu+3\sigma$;

The normal distribution- Illustration using excel



The Cumulative distribution function

- Pdf can be represented as a cumulative distribution
- In this representation the normal distribution is characteristically sigmoid
- The main use of the cumulative distribution function is that the probability of a values being less than a specified amount can be deduced



Data transformation

- To overcome the difficulties arising from departures from normality observations can be transformed to a new scale on which the distribution is more nearly normal.
- Further analysis can be done on the transformed data, and if necessary transform the results to the original scale at the end.

Logarithmic transformation

- The geometric mean of a dataset is given as

$$\bar{g} = \left\{ \prod_{i=1}^{N} z_i \right\}^{\frac{1}{N}}, \quad \text{Equation (2.13)}$$

- Therefore, the log transformation is

$$\log \bar{g} = \frac{1}{N} \sum_{i=1}^{N} \log z_i, \quad \text{Equation (2.14)}$$

- The logarithm may be either natural (ln) or common (log10). If by transforming the data Z_i= 1,2, 3……….N, we obtain log z with a normal distribution then the variable is said to be **lognormally distributed**.

Square root transformation

- Taking logarithms will often normalize, or at least make symmetric, distributions that are strongly positively skewed, i.e. have g1 > 1.
- Less pronounced positive skewness can be removed by taking square roots:

$$r = \sqrt{z}. \qquad \text{Equation (2.15)}$$

Angular transformation

- This is sometimes used for proportions in the range 0 to 1, or 0 to 100 if expressed as percentages.
- If p is the proportion then define

$$\phi = \sin^{-1}\sqrt{p}. \qquad \text{Equation (2.16}$$

- The desired transform is the angle whose sine is $\sqrt{p}$

Logit transformation

If, as above, p is a proportion (0 <p <1), then its logit is

$$l = \ln \frac{p}{1-p}$$ Equation (2.17)

- Note that the limits 0 and 1 are excluded; otherwise l would either go to -∞ or +∞.
- If you have proportions that include 0 or 1 then you must make some little adjustment to use the logit transformation.

Thank you for your attention! Questions?