

Institute of  
Data



2020



# Data Science and AI

Module 1

Part 1:

---

## Mathematics & Statistics

---



# Agenda: Module 1 Part 1

- Linear algebra
- Calculus
- Multivariable calculus
- Statistics
- Probability



# What is Linear Algebra and *why is it important for Data Science?*

- Linear Algebra is the branch of mathematics that deals with ***linear equations*** and their representations.
- Linear Algebra is used extensively in science and engineering to '***model***' many systems such as in economics, health and finance.
- Although many systems are '***non-linear***', Linear Models can be effective ***first-order approximation***. This is crucial, because non-linear models are ***very difficult*** to represent and manipulate.
- One excellent way to better understand Linear Algebra is use the ***geometrical representations*** of its constructs.
- Another way to better understand Linear Algebra is through ***programming***. This is crucial for a Data Scientist.
- Key constructs of Linear Algebra are: Scalar, Vector, Matrix and Tensor.



# Linear Algebra

- Vectors
  - definitions
  - vector arithmetic (adding, subtracting and multiplying vectors), dot products and cross products
- Matrices
  - definitions
  - matrix arithmetic
  - inverses, determinants, transposes
- Solving systems of linear equations
- Eigenvalues and eigenvectors



# Linear Algebra

- Vectors
  - definitions
  - vector arithmetic (adding, subtracting and multiplying vectors), dot products and cross products
- Matrices
  - definitions
  - matrix arithmetic
  - inverses, determinants, transposes
- Solving systems of linear equations
- Eigenvalues and eigenvectors



# Mapping and usage of Linear Algebra in Data Science

| Concept | Definition  | Mapping to Data Science   | Examples  |
|---------|---|---|---|
| Scalar  | A 'zero-dimensional' dataset.<br>A number, value, magnitude.<br>Geometrically, it's <b>a point on on a line.</b>                          | A single data point   | Age of a customer   |
| Vector  | A one-dimension dataset. A two or more values.<br>Geometrically it represent a vector in a plane that has <b>magnitude and direction.</b> | A number of data points (usually about a single entity)   | Attributes (or <b>features</b> ) of one customer:<br>Age, income, marital status, postcode, ..., etc<br>In Deep Learning a vector could be the input to a Neural Network. |
| Matrix  | A two-dimensional dataset.<br>Geometrically, it represents a <b>transformation</b> of two or more vectors.                                | A set of observations for multiple entities.<br>A transformation of a dataset from one representation to another. | Information about all customers.<br>In Deep Learning a matrix may represents the mapping and weights on hidden layer.   |
| Tensor  | An n-dimensional dataset.   | A number of sets of observations  | Information about all customers.<br>TensorFlow is built around tensors.   |



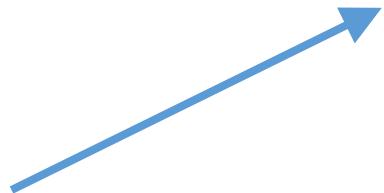
# Vectors

*def:* ?

a directed quantity

*examples:* ?

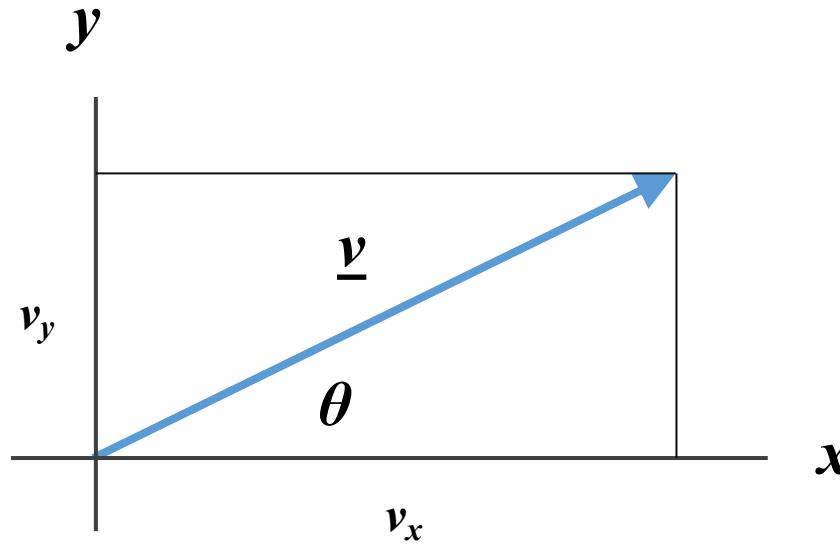
- displacement (*not* length)
- velocity (*not* speed)
- acceleration
- force
- weight (*not* mass)



dimensionality > 1



# Vector Decomposition: 2D



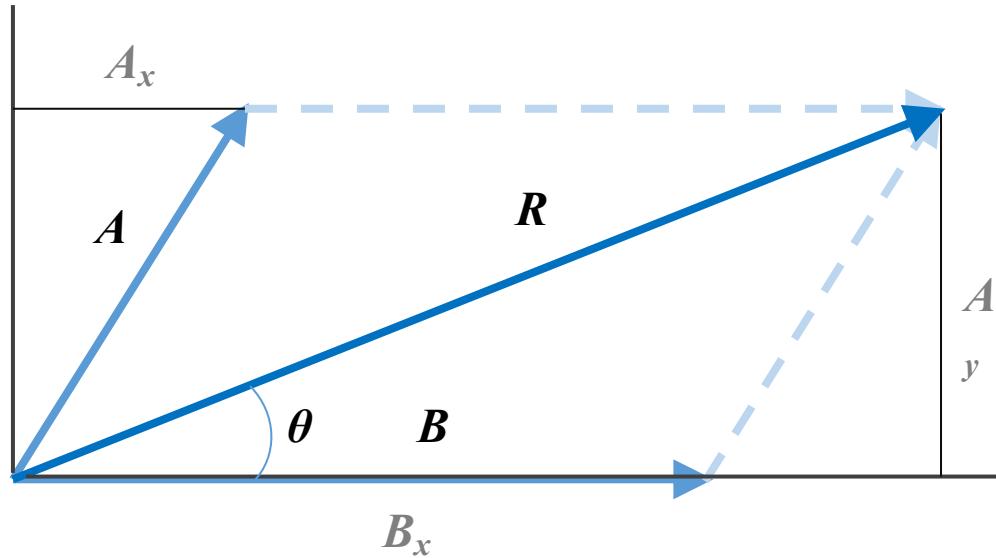
$$v_x = v \cos \theta$$

$$v_y = v \sin \theta$$

$$|\underline{v}| = (v_x^2 + v_y^2)^{1/2}$$



# Vector Addition



in general:

$$R_x = A_x + B_x$$

$$R_y = A_y + B_y$$

in this example:

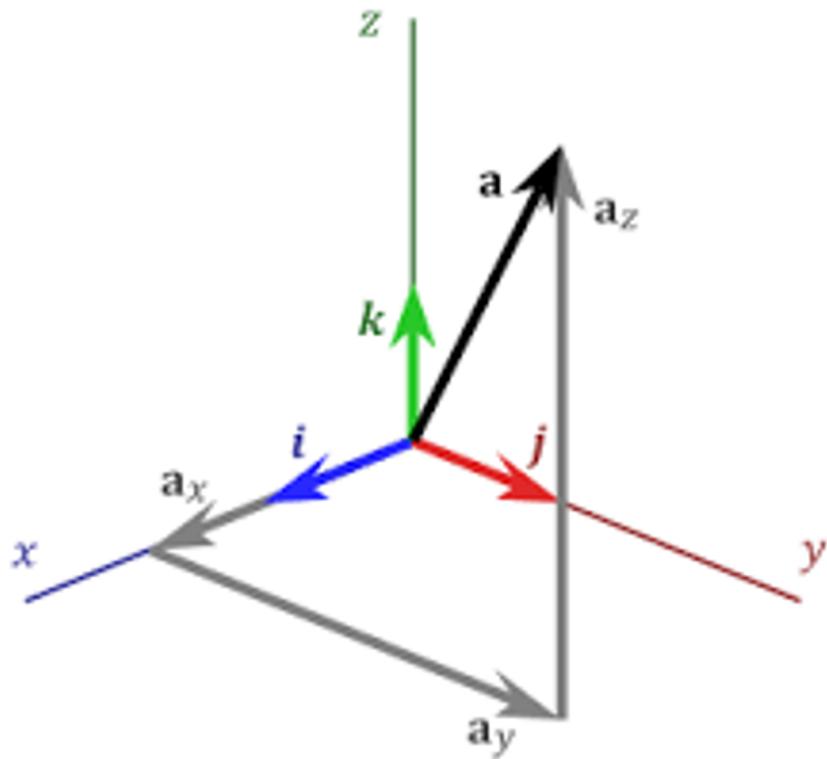
$$R_x = A \cos \theta + B$$

$$R_y = A \sin \theta$$

$$\theta = \tan^{-1} (R_y / R_x)$$



# 3D Vectors



$$|\underline{a}| = (\underline{a}_x^2 + \underline{a}_y^2 + \underline{a}_z^2)^{1/2}$$

Note:

$i, j, k$  are unit vectors



# Scalar Multiplication of Vectors

*aka* inner product, dot product  
result is a *scalar*

$$\begin{aligned}\mathbf{a} \bullet \mathbf{b} &= (a_1, a_2, \dots, a_n) \bullet (b_1, b_2, \dots, b_n) \\ &= a_1b_1 + a_2b_2 + \dots + a_nb_n \\ &= |\mathbf{a}| |\mathbf{b}| \cos(\theta)\end{aligned}$$



# Vector Multiplication of Vectors

aka cross product

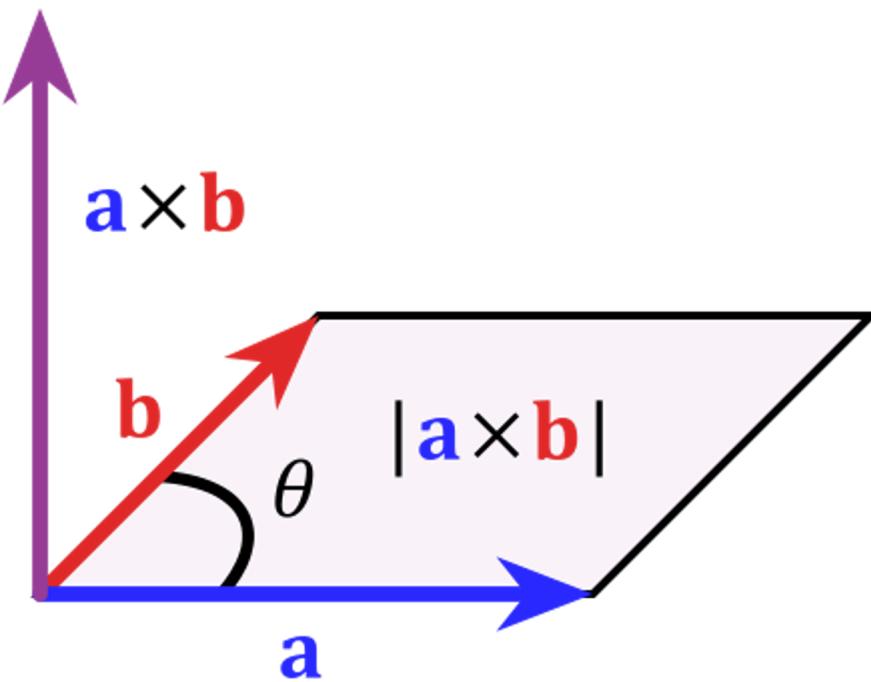
result is a *vector*

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_x & u_y & u_z \\ v_x & v_y & v_z \end{vmatrix}$$

$$= (u_y v_z - u_z v_y) \mathbf{i} + (u_z v_x - u_x v_z) \mathbf{j} + (u_x v_y - u_y v_x) \mathbf{k}$$



## cross product – cont'd



magnitude of  $\mathbf{a} \times \mathbf{b}$   
= area of parallelogram



# Vector Operations

Entry-wise multiplication:

*aka* Hadamard product

result is a *vector*

$$\mathbf{a} * \mathbf{b} = (a_1 b_1, a_2 b_2, \dots a_n b_n)$$

Transpose

$$(a_1, a_2, \dots a_n)^T = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$



# Matrices

*def:* ?

a rectangular array of numbers

$$\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ a_{21} & \cdots & a_{2m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}$$



# Matrix Arithmetic

addition

$$\begin{aligned} A + B &= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \\ &= \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{pmatrix} \end{aligned}$$



# Matrix Arithmetic

multiplication

$$\begin{aligned} A B &= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \\ &= \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix} \end{aligned}$$

in general:

$A$  is  $m \times n$     $B$  is  $n \times p$

$A B$  is  $m \times p$



# Matrix Operations 3

Transpose

$$\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ a_{21} & \cdots & a_{2m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{pmatrix}$$

the rows of  $A^T$  are the columns of  $A$



# *Determinant of a Matrix*

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\det(A) = |A| = a_{11}a_{22} - a_{12}a_{21}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$|A| = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$



# Identity Matrix

Define the  $n \times n$  identity  $I_n$  :

$$A I_n = I_n A = A$$

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots \\ 0 & 0 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix}$$



# Matrix Inversion

Define the inverse  $A^{-1}$  of an invertible matrix  $A$ :

$$A A^{-1} = I_n = A^{-1}A$$

only exists if...

$A$  is  $n \times n$

$\det(A) \neq 0$

Methods:

- Gaussian (Gauss-Jordan) elimination
- LU decomposition (orthogonalisation)
- Eigen decomposition



# Solving Systems of Linear Equations

Simultaneous linear equations in  $n$  unknowns ( $x_i$ ):

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$

⋮

$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m$$

in matrix form:

$$A\mathbf{x} = \mathbf{b}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$



## Solving Systems of Linear Equations – cont'd

Problem:

$$A \mathbf{x} = \mathbf{b}$$

Solution:

$$\mathbf{x} = A^{-1} \mathbf{b}$$

So, we only need to invert the matrix of coefficients  $A$  and multiply with vector  $\mathbf{b}$



# Eigenvalues and Eigenvectors

*def:* A nonzero vector that scales linearly in response to a linear transformation

$$T(\boldsymbol{v}) = \lambda \boldsymbol{v}$$

$T$  is a linear transformation

$\lambda$  is a scalar = ‘eigenvalue’ (*aka* characteristic value, characteristic root)

$\boldsymbol{v}$  is a vector = ‘eigenvector’ (*aka* characteristic vector)

eigenbasis: a set of eigenvectors of  $T$  that forms a basis of the domain of  $T$



# *What is bigger than a matrix?*

tensor

- represented by an  $n$ -dimensional array
- examples
  - stress tensor (mechanics)
  - spacetime tensor (general relativity)



# Lab 1.1.1: Vector and Matrix Operations

Purpose:

- To apply the definitions of vector and matrix operations by designing code that implements them.

Materials:

- 'Lab 1.1.docx'
- See Notebook:
  - 'Lab 1.1 – Notebook – Linear Algebra'





# Calculus

- Limits and continuity
- Taking derivatives
- Integration
- Sequences and series



# What is Calculus and why is it important for Data Scientists?

- Calculus is the mathematical study of **continuous change**. It is used extensively in many science and engineering domains such as business, economics and medicine.
- All key concepts in calculus can be mapped directly to **geometrical concepts**. For example, differentiation is the slope of a curve and integration is the area under a curve.
- Calculus is usually used with linear algebra to find the "**best fit**" linear approximation for a set of points in a domain. Therefore it is essential for Data Science as it underpins all model optimisation.

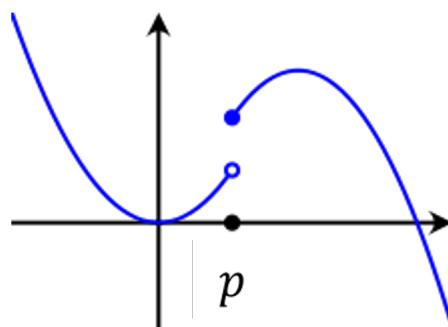


# Limits and Continuity

If a function  $f(x)$  approaches a value  $L$  ('limit') as  $x$  approaches  $p$ , then

$$\lim_{x \rightarrow p} f(x) = L$$

Note: The limits of a discontinuous function are directional



$$\lim_{x \rightarrow p^-} f(x) \neq \lim_{x \rightarrow p^+} f(x)$$



# Limit Theorems

$$\lim_{n \rightarrow \infty} k a_n = k \lim_{n \rightarrow \infty} a_n$$

$$\lim_{n \rightarrow \infty} (a_n \pm b_n) = \lim_{n \rightarrow \infty} a_n \pm \lim_{n \rightarrow \infty} b_n$$

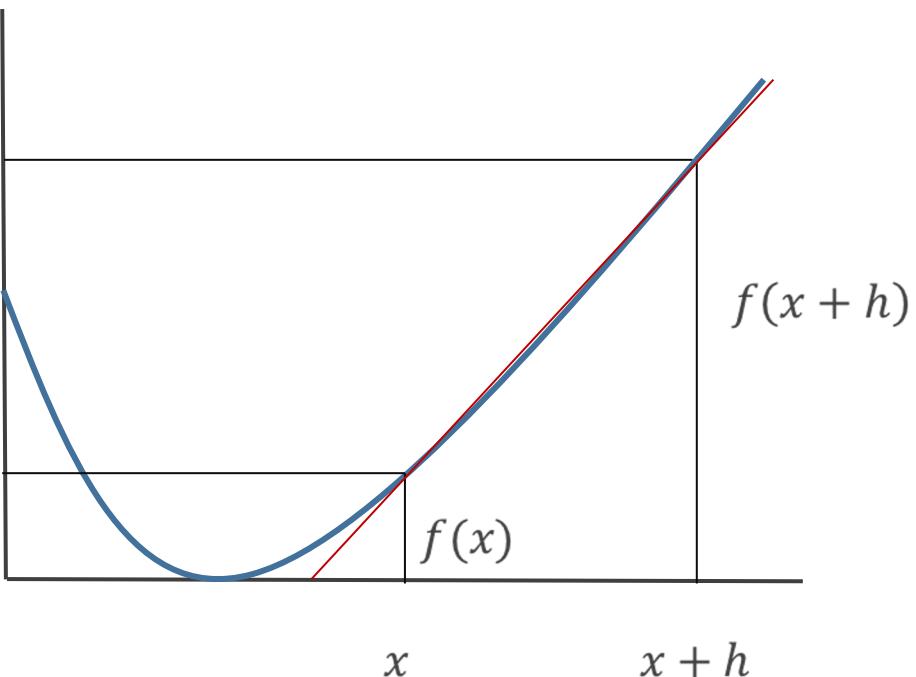
$$\lim_{n \rightarrow \infty} (a_n b_n) = \lim_{n \rightarrow \infty} a_n \lim_{n \rightarrow \infty} b_n$$

$$\lim_{n \rightarrow \infty} \left( \frac{a_n}{b_n} \right) = \frac{\lim_{n \rightarrow \infty} a_n}{\lim_{n \rightarrow \infty} b_n}$$



# Differentiation

Rate of change of a continuous function  $f(x)$ :



Derivative of  $f(x)$ :

$$\frac{d}{dx}f(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$



# Rules of Differentiation

Let  $f, g, h$  be functions of  $x$ , and let  $a, b$  be constants ...

Linearity:

$$\frac{d(af+bg)}{dx} = a \frac{df}{dx} + b \frac{dg}{dx}$$

Product rule:

$$\frac{d(fg)}{dx} = g \frac{df}{dx} + f \frac{dg}{dx}$$

Chain rule:

if

$$h(x) = f(g(x))$$

then

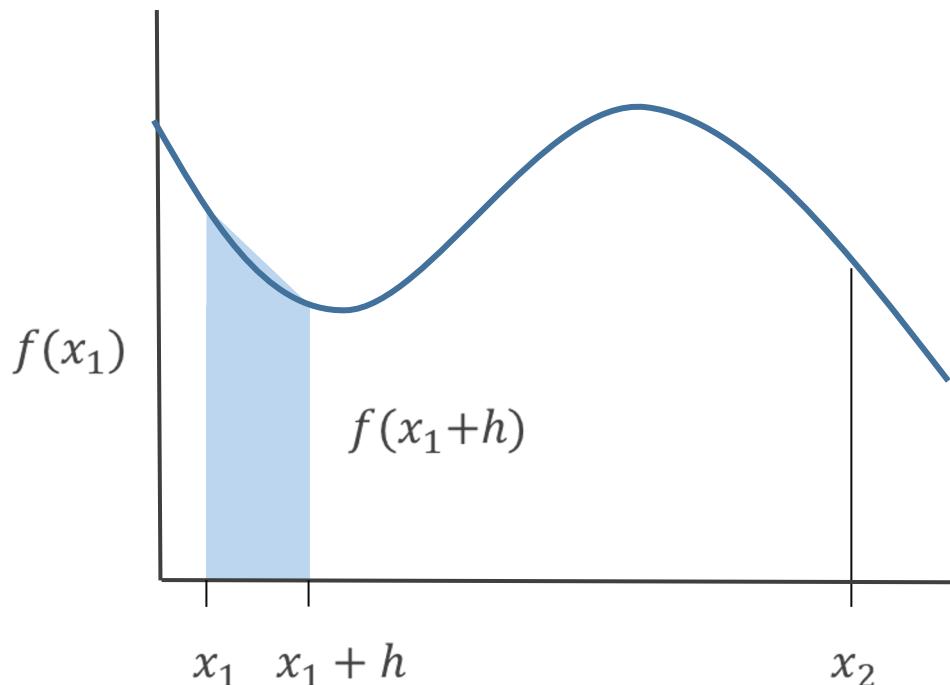
$$\frac{dh}{dx} = \frac{dh}{dg} \frac{dg}{dx}$$



# Integration

Area under a continuous function  $f(x)$ :

$$F(x) = \lim_{h \rightarrow 0} \sum [f(x_i) + f(x_i + h)] h/2$$



If  $f(x) = dF/dx$

then the integral of  $f(x)$  between  $x_1$  and  $x_2$  is:

$$\int_{x_1}^{x_2} f(x) dx = F(x_2) - F(x_1)$$



# Sequences and Series

Infinite sequence

$$a_n = f(n)$$

typically:  $0 \leq n \leq \infty$

$-\infty \leq n \leq \infty$

*examples:*

$$a_n = n$$

$$a_n = 1/n$$

$$a_n = 1/n^2$$

$$a_n = (-1)^n$$



# Multivariate Calculus

- Partial derivatives
- Multivariate differentiation
- Multivariate integration
- Optimising multivariate functions



# Partial Derivatives

If  $f$  is a function of several variables, we can calculate the partial derivative with respect to any single variable by treating the others as constants:

example:

$$f(x, y) = 3x^2 + 2xy$$

$$\frac{\partial f}{\partial x} = 6x + 2y$$

$$\frac{\partial f}{\partial y} = 2x$$



## Partial Derivatives – cont'd

For a function operating on a 3-dimensional Euclidean space, the partial derivatives define the ***gradient*** of the function:

$$\nabla f = \left( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right)$$

The *del* operator is often written as:

$$\nabla = \hat{i} \frac{\partial}{\partial x} + \hat{j} \frac{\partial}{\partial y} + \hat{k} \frac{\partial}{\partial z}$$



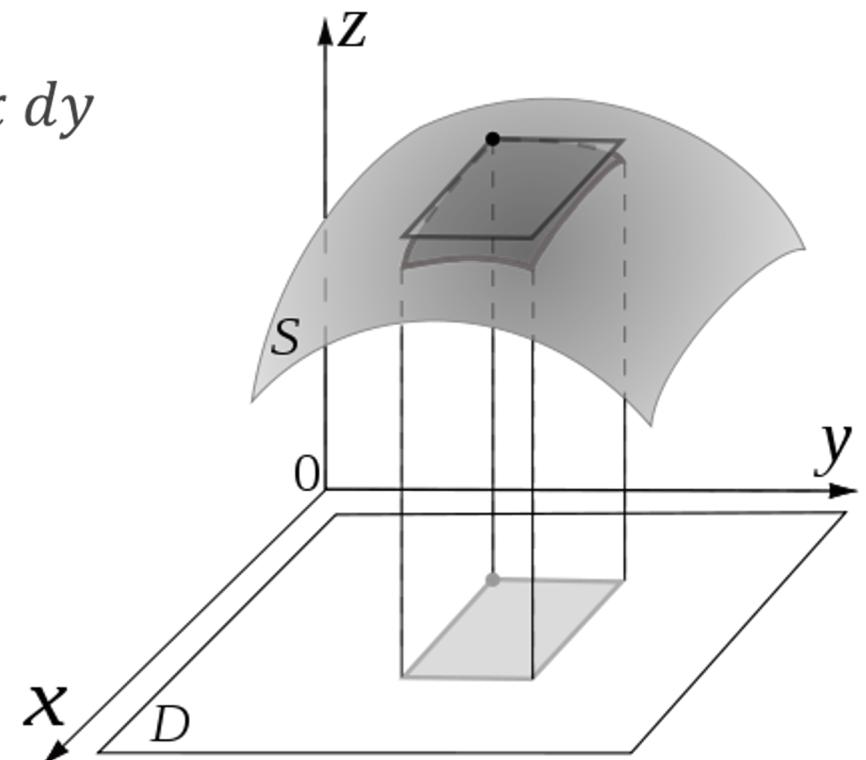
# Multivariate Integration

Example: Surface area

$$A_S = \iint_{R_{xy}} \sqrt{\left(\frac{df}{dx}\right)^2 + \left(\frac{df}{dy}\right)^2 + 1} \ dx \ dy$$

Cartesian:  $ds = dx \ dy$

Polar:  $ds = r \ dr \ d\theta$





# Multivariate Optimisation

Given a function  $f: A \rightarrow \mathbb{R}$

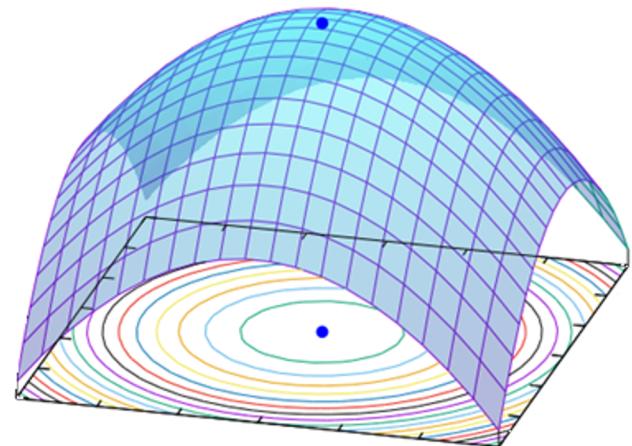
minimisation:

find  $x_0 \in A$  such that  $f(x_0) \leq f(x) \forall x \in A$

maximisation:

find  $x_0 \in A$  such that  $f(x_0) \geq f(x) \forall x \in A$

$f$  = [objective | loss | cost | utility | fitness | energy]  
function





# Lab 1.1.2: Differentiation and integration in Python

## Purpose:

- Use python to define limits, derivative and integral of a function (for example,  $f(x) = x^{**}2$ ).

## Materials:

- See Notebooks
  - Calculus - Limits
  - Calculus – Derivative
  - Calculus – Integral

## Note:

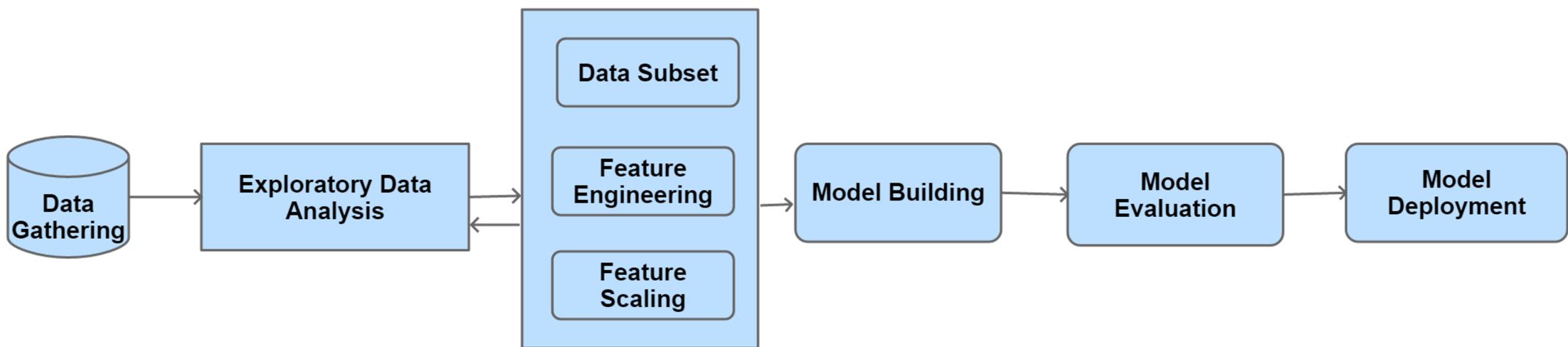
- There may not be enough time to complete this lab in the class.  
Please complete it as a part of your homework.  
This should apply to all labs.





# Discussion

- Why do data scientists need to be proficient at calculus, infinite series, and linear algebra?
- Considering the illustrative Data Science process, where would you use calculus?





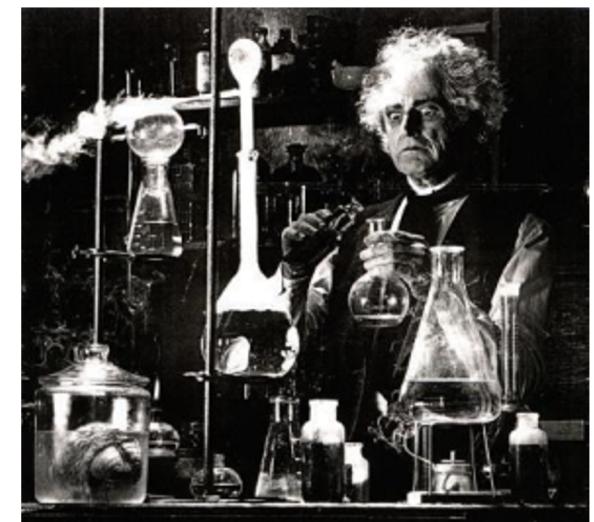
# Homework: Optimisation

## Purpose:

- To introduce several methods of multivariate optimisation via working examples.

## Materials:

- An Interactive Tutorial on Numerical Optimization
  - <https://www.benfrederickson.com/numerical-optimization/>
- Prepare to discuss:
  - trade-off: convergence speed vs accuracy
    - faster convergence requires lower resolution
    - prefer methods with adaptive resolution





# Statistics

- Statistical Thinking
- Categorical data
- Continuous variables
- Summarising quantitative data
- Modelling data distributions
- Confidence intervals
- Significance tests and hypothesis testing
- Statistical Inference



# Why statistics is important for a Data Scientist?

- **Statistical Thinking** is an essential component of a data-driven mindset which is crucial for a Data Scientist
  - Statistical analysis must start with the appropriate **data** (sample)
  - Statistical Inference (reasoning) should start with measurement, ideally, via **controlled experiments**
  - Statistics uses samples (a small subset of the population) and therefore always has a degree of **uncertainty**
  - Sampling must be **random, and preferably, independent**
- The best way to learn statistics is by **experimenting with data using Python code and visualisation**



# Statistics – Part 1

- Analysing categorical data



# Categorical Variables

## Examples

- FALSE / TRUE (alt: 0 / 1)
- colour
- size
- class
  - e.g. species, occupation, degree program, disease category
- tier
  - e.g. age range, income range, frequency range



# Analysing Categorical Variables

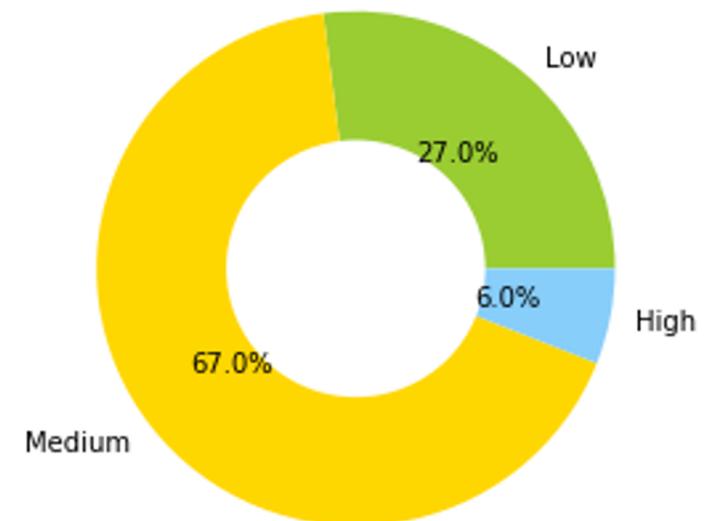
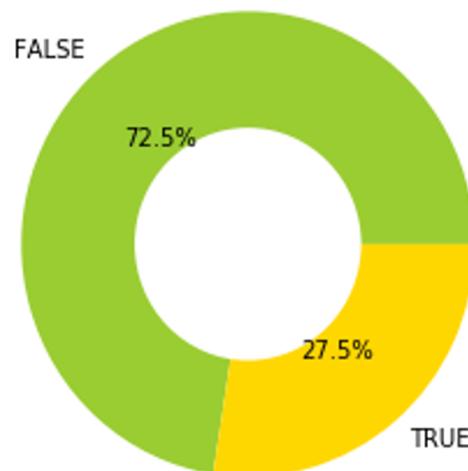
Frequency Tables (aka contingency tables)

*category incidence for a single variable within a population*

| Passed Exam |    |
|-------------|----|
| FALSE       | 37 |
| TRUE        | 14 |

| Income Bracket |      |
|----------------|------|
| Low            | 0.27 |
| Medium         | 0.67 |
| High           | 0.06 |

Donut Charts





# Analysing Categorical Variables – cont'd

## Two-Way Frequency Tables

- *for a Single Variable within Two Populations*

| Income Bracket: | Low | Medium | High | Total |
|-----------------|-----|--------|------|-------|
| Male            | 27  | 75     | 6    | 108   |
| Female          | 32  | 59     | 3    | 94    |
| TOTAL           | 59  | 134    | 9    | 202   |

- totals row, column: marginal frequencies (*aka* marginal distribution)



# Analysing Categorical Variables – cont'd

Dummy Variables (*aka* dummy coding)

- *allows categorical variables to be treated like continuous variables*

| Passed Exam |    |
|-------------|----|
| 0           | 37 |
| 1           | 14 |

| Treatment | T1 | T2 |
|-----------|----|----|
| Control   | 0  | 0  |
| Drug 1    | 1  | 0  |
| Drug 2    | 0  | 1  |



# Statistics – Part 2

- Continuous variables
- Summarising quantitative data



# Continuous Variables

## Examples

- height
- dose
- temperature
- concentration
- revenue
- clicks

## “Continuous”?

- variability is treated as infinite
  - precision is determined by data acquisition methodology
- range usually has practical limits
  - outliers can be defined statistically or heuristically
- *frequency (contingency) is not meaningful*



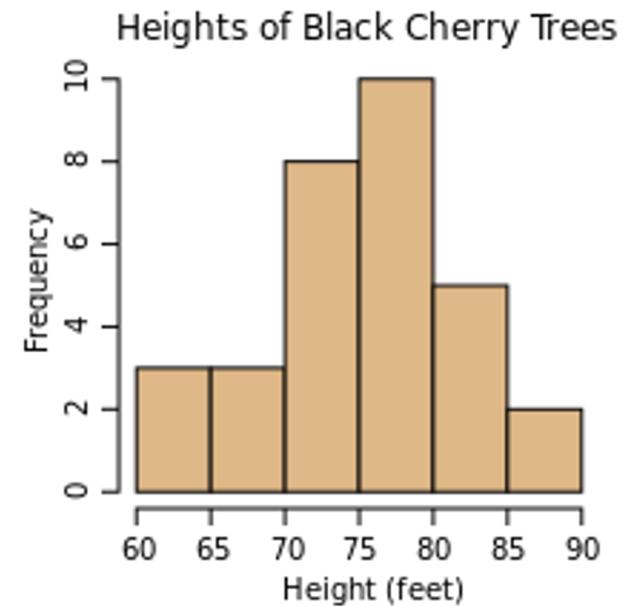
# Analysing Continuous Variables

Distribution of binned data:

- choose an appropriate bin width
- ‘cut’ the data into bins
- count the number of samples that fall into each bin

*what is the resulting plot called?*

- histogram





# Summarising Quantitative Data

Measuring the centre of the data

mean

the average value of the variable

median

the value that separates the 50% lowest values from the rest

mode

the most frequently occurring value



# Summarising Quantitative Data – cont'd

## Quantiles

- inverse of binning data for a histogram:
  - specify proportions of samples we want in each bin
  - compute bin boundaries that correspond

example: 4 quantiles from a random sample (mean = 0, variance = 1):

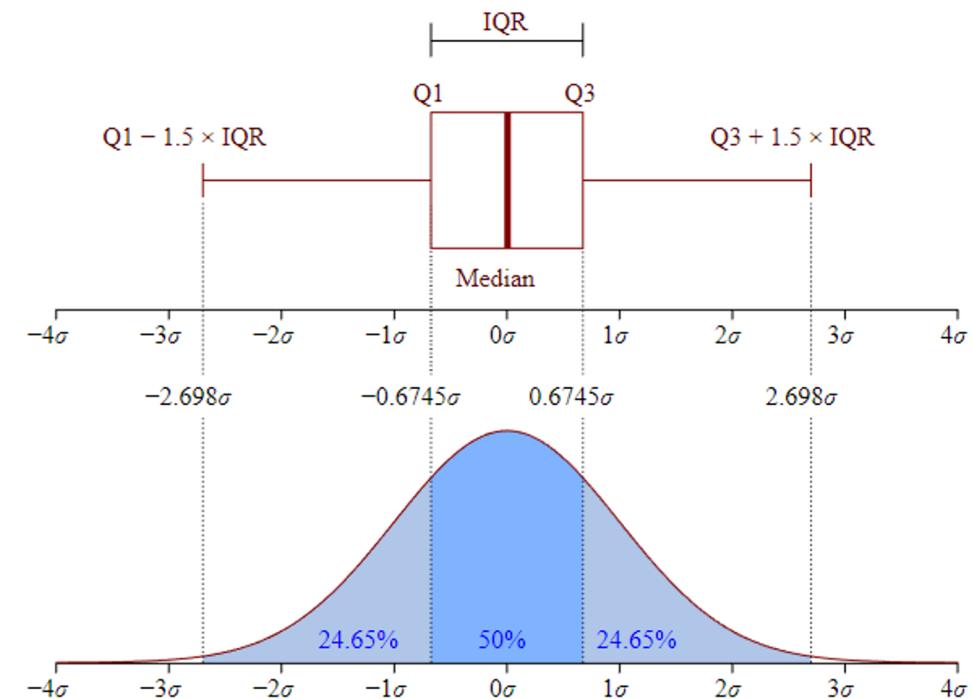
| Quantile      | Boundaries       | Count |
|---------------|------------------|-------|
| (0 - 0.25]    | (-3.135, -1.61]  | 53    |
| (0.25 - 0.50] | (-1.61, -0.0913] | 389   |
| (0.50 - 0.75] | (-0.0913, 1.427] | 476   |
| (0.75 - 1.0]  | (1.427, 2.946]   | 82    |



# Summarising Quantitative Data – cont'd

## Interquartile range (IQR)

- $IQR = [0.25, 0.75]$
- box plots are drawn with whiskers extending  $1.5 \times IQR$  beyond the 0.25 and 0.75 quantiles (i.e. the 1<sup>st</sup> and 3<sup>rd</sup> quartiles)
- **outliers** are typically defined as lying outside this range



By Jhguch at en.wikipedia, CC BY-SA 2.5,  
<https://commons.wikimedia.org/w/index.php?curid=14524285>



# Summarising Quantitative Data – cont'd

Moments of a Sample

mean

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

variance

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

skewness

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^3$$

kurtosis

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^4 - 3$$



# Lab 1.1.3: Simple data visualisation

## Purpose:

- Use various plot types to visualise statistical observations.

## Materials:

- Notebook: 'Statistics – part 1'

## Note:

- There may not be enough time to complete this lab in the class.  
Please complete it as a part of your homework.  
This should apply to all labs.





# Statistics – Part 3

- Modelling data distributions



# Summarising Quantitative Data – cont'd

Summary statistics

standard deviation

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

square root of variance

same units as mean



# Modelling Data Distributions

Sample vs Population

$\mu$  = mean of population

$\bar{x}$  = mean of sample

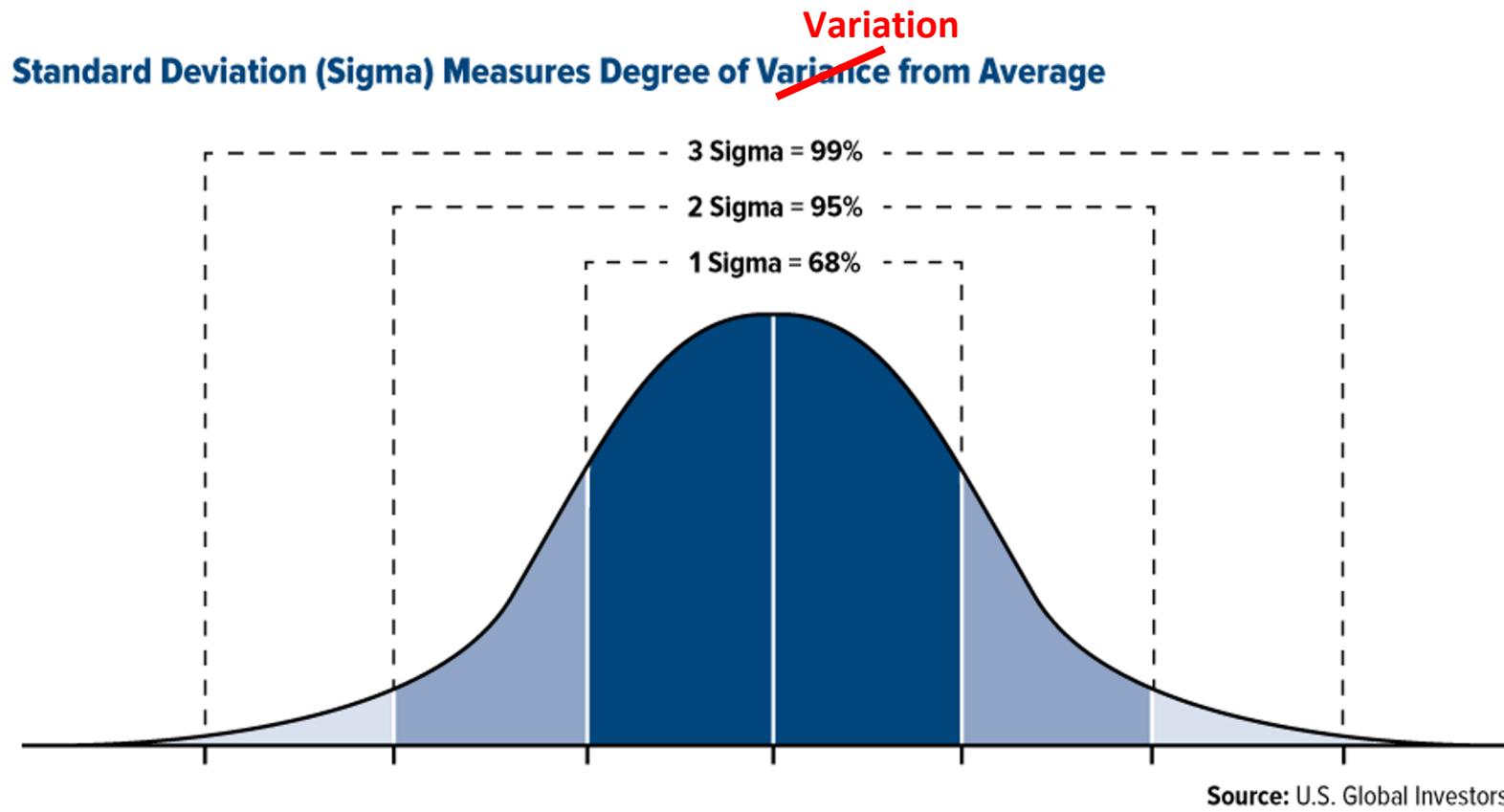
$\sigma$  = standard deviation of population

$s$  = standard deviation of sample



# Modelling Data Distributions – cont'd

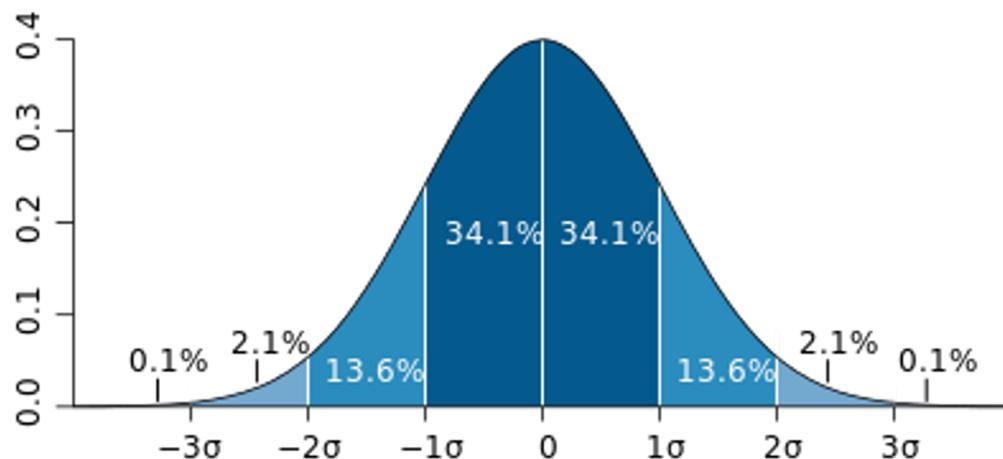
Mean and Standard Deviation of a Population



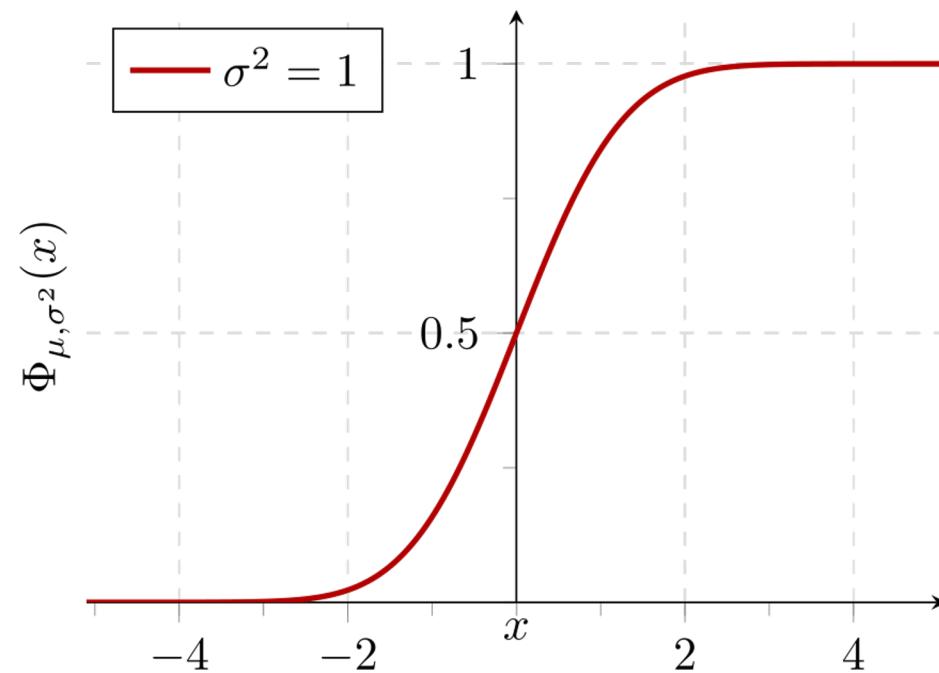


# Modelling Data Distributions – cont'd

Probability Density Function



Cumulative Probability



By M. W. Toews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, <https://commons.wikimedia.org/w/index.php?curid=1903871>

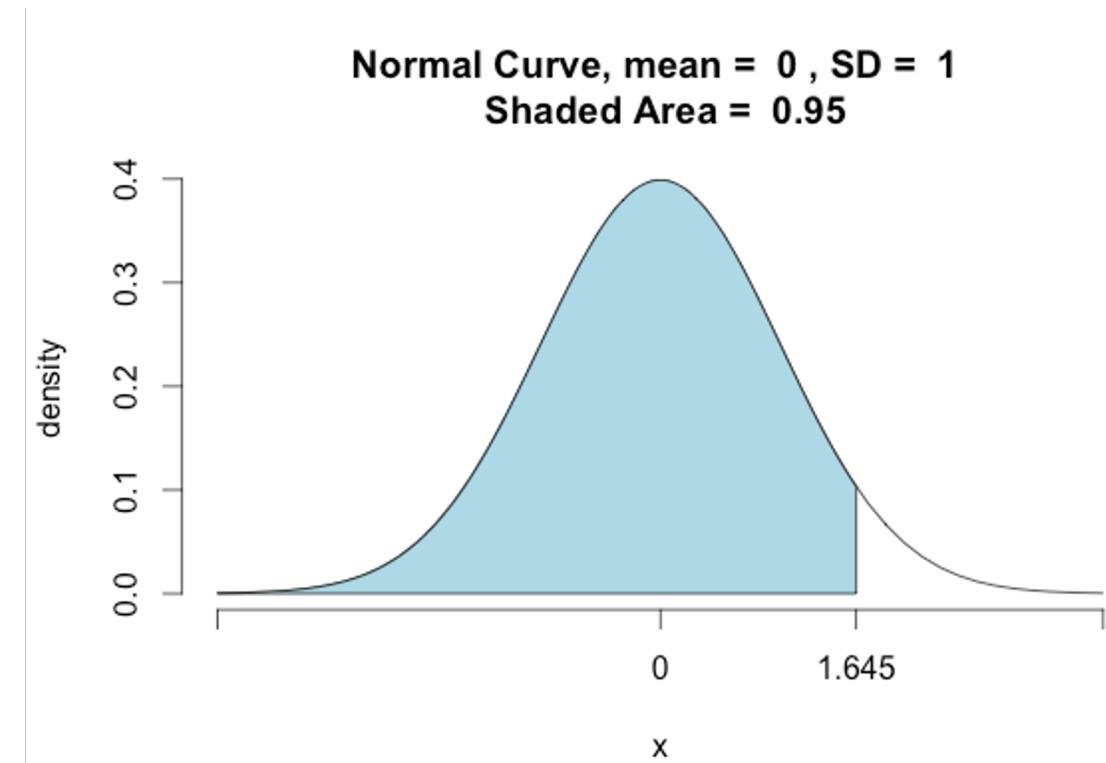


## Modelling Data Distributions – cont'd

z-score

measures how far a sample lies  
from the population mean:

$$z = \frac{x - \mu}{\sigma}$$





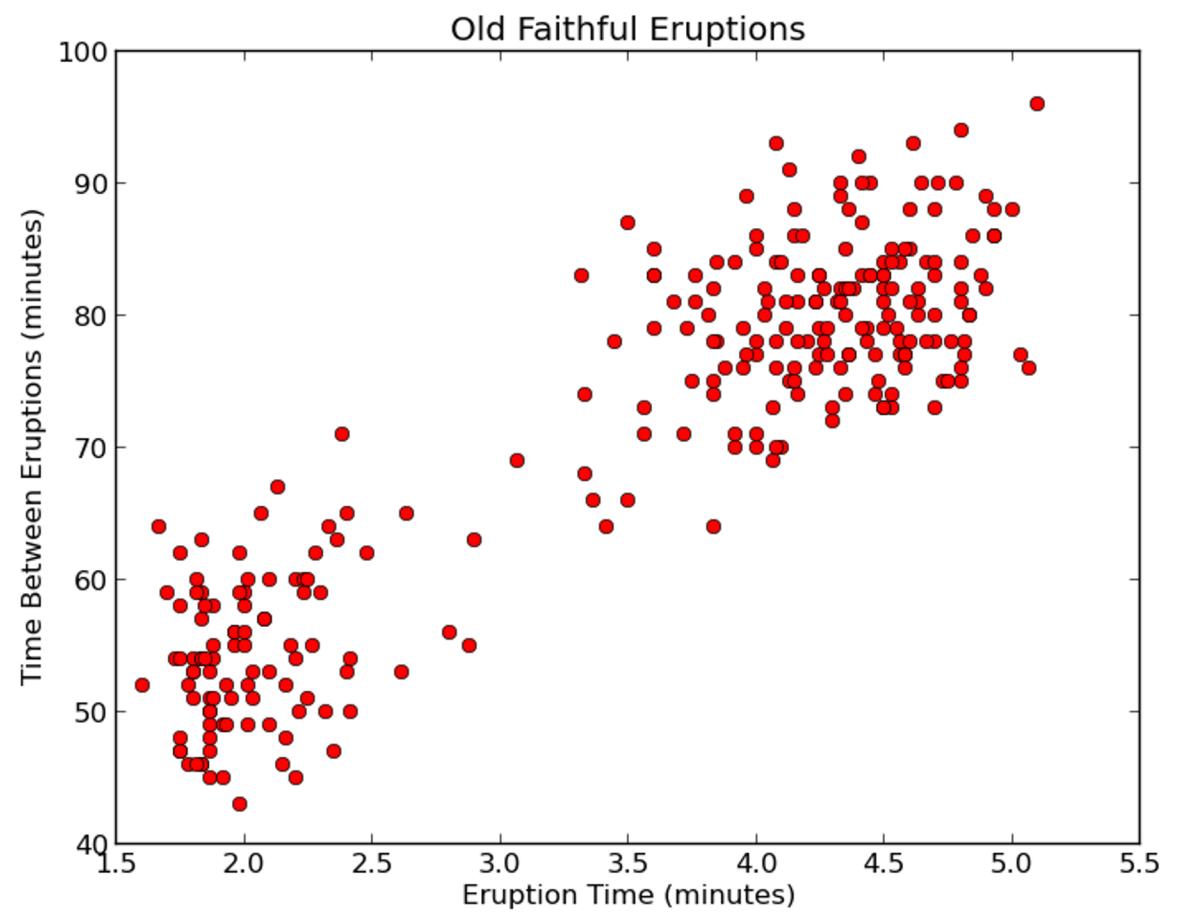
# Statistics – Part 4

- Exploring bivariate numerical data



# Scatter Plots

- 2D: plots one variable against another
- demonstrates a relation (or lack thereof) between two variables
- *assumption: data pairs are sampled simultaneously*





# Correlation

Pearson correlation coefficient

measures strength of covariance between one variable and another:

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

- large when big variations in  $y$  correspond to big variations in  $x$
- small when small variations in  $y$  cancel out big variations in  $x$  (or vice versa)

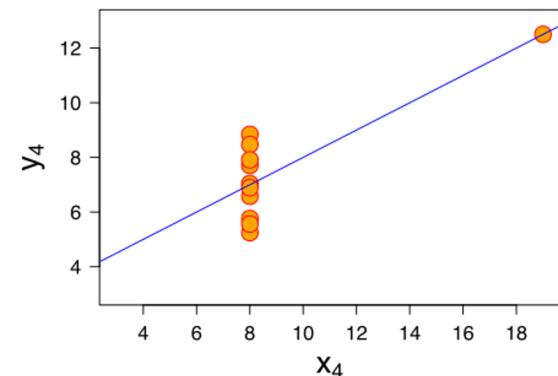
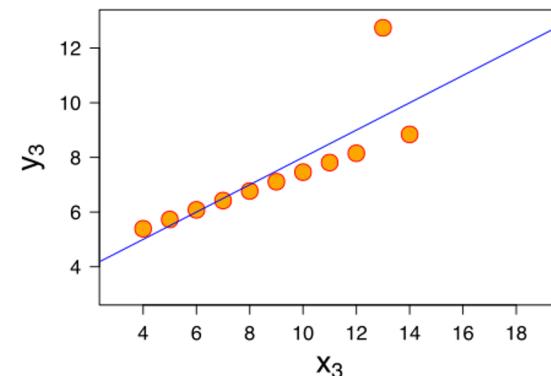
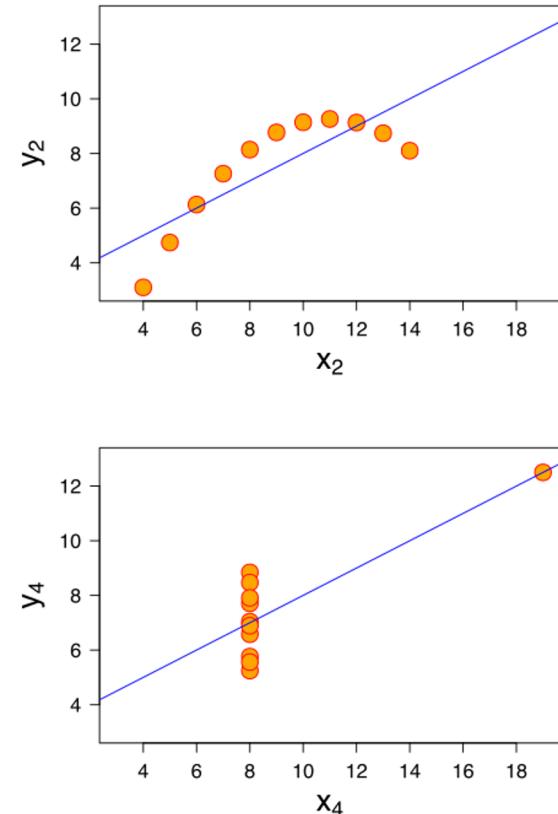
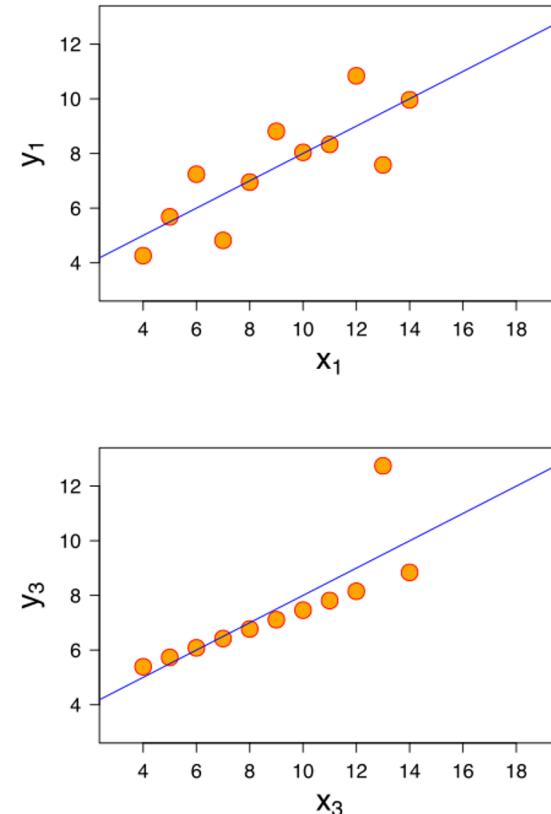


# Correlation

Anscomb's quartet

four very different sets of 11 data pairs, each with  $r_{xy} = 0.816$

- correlation coefficient
  - assumes a *linear* relation
  - does not completely characterise the relationship between  $x$  and  $y$



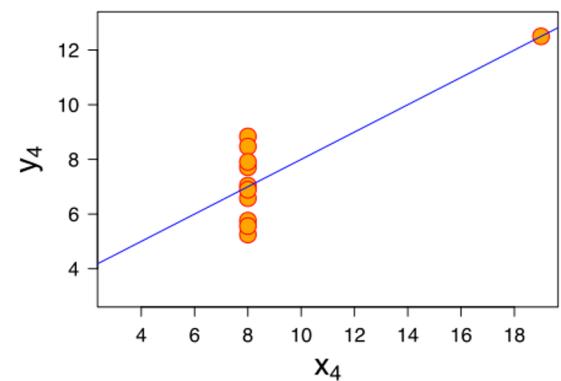
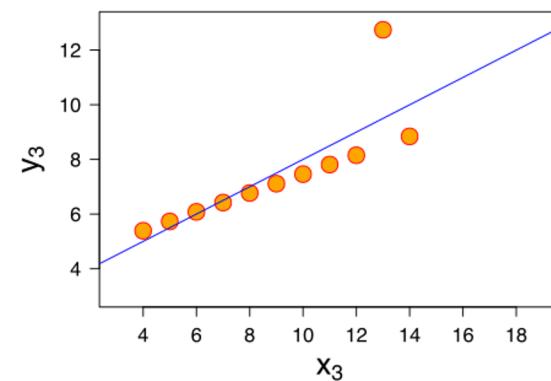
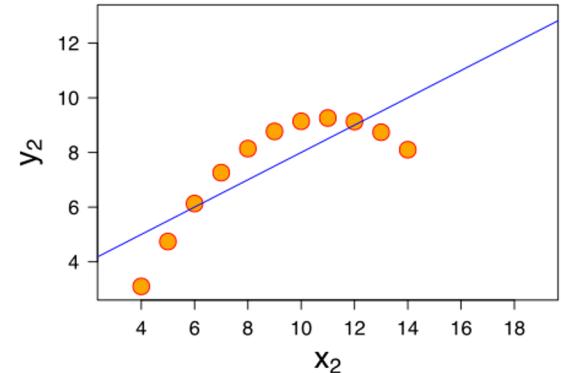
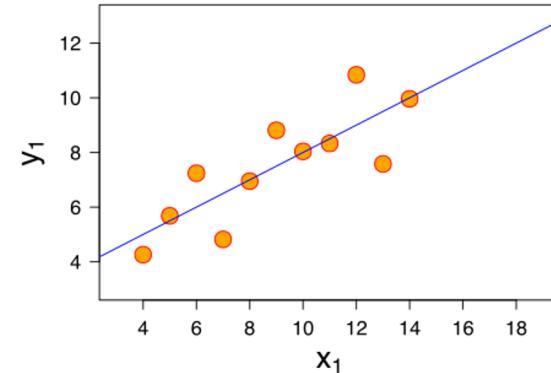
By Anscombe.svg: SchutzDerivative works of this file:(label using subscripts): Avenue - Anscombe.svg, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=9838454>



# Trend Lines

best (linear) fit to a 2D scatter plot

- the line that minimises error  
*by some criterion*
- line is specified by
  - slope
  - intercept



By Anscombe.svg: SchutzDerivative works of this file:(label using subscripts): Avenue - Anscombe.svg, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=9838454>



# Least-Squares Regression

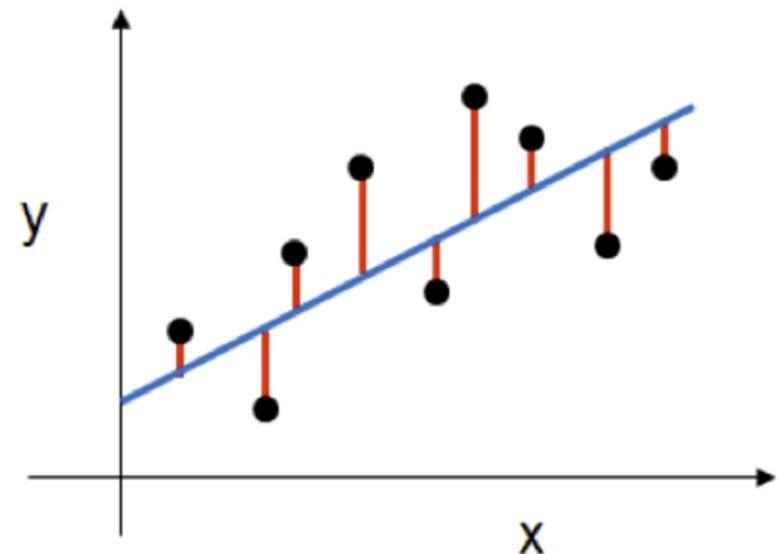
*def: residual*

the difference between the observed  
and predicted values:

$$\varepsilon_i = y_i - \hat{y}_i$$

- least-squares criterion:

$$err = \sum_{i=1}^n \varepsilon_i^2$$





# Least-Squares Regression

minimise:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

solution:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



# Statistics – Part 5

- Random variables



# Random Variables

## *Discrete random variables*

- range is finite or countably infinite
- distribution can be described by a ***probability mass function***
  - assigns a probability to each value in the image of  $X$

## *Continuous random variables*

- distribution can be described by a ***probability density function***
  - assigns a probability to each specified interval over the range of  $X$



# Transforming random variables

*standardising* variables for analysis:

- *centering* (subtracting the mean)
- *scaling* (dividing by SD)
- allows standard tables to be used to compute percentiles of the sample distribution and probabilities of sampled values

$$x' = \frac{x - \bar{x}}{s}$$

recall: z-score      $z = \frac{x-\mu}{\sigma}$



## Transforming random variables – cont'd

*offset*

$$x' = x + 1$$

- datasets based on counts (binning) may contain zeros
  - examples:
    - calls received in each minute at a call centre
    - instances of a keyword in a corpus
  - if the method relies on the logarithm of the count (which many do), it will blow up for  $x_i = 0$



# Transforming random variables – cont'd

## *logarithmic rescaling*

$$x' = \log(x)$$

- datasets with large dynamic range
  - examples:
    - lifetime value of customer
  - algorithm could be skewed if a small amount data with large values dominates a large amount of data with small values



# Transforming random variables – cont'd

## *Box-Cox transformation*

$$x' = x^\lambda \quad \lambda \in \{0, \pm 0.5, \pm 1, \pm 2, \pm 3\}$$

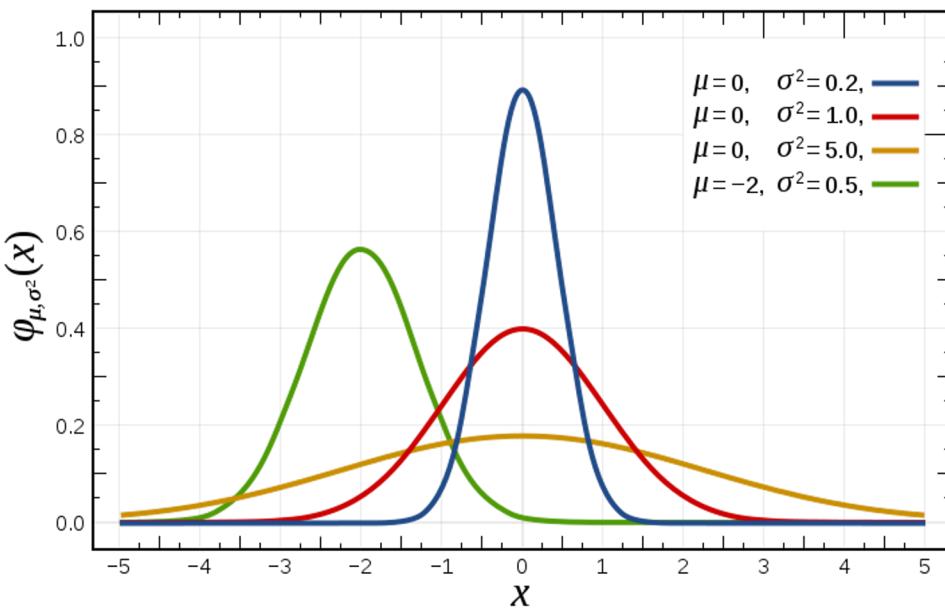
- datasets high skewness or kurtosis
- try different values of  $\lambda$ 
  - choose the one that gives the most normal distribution



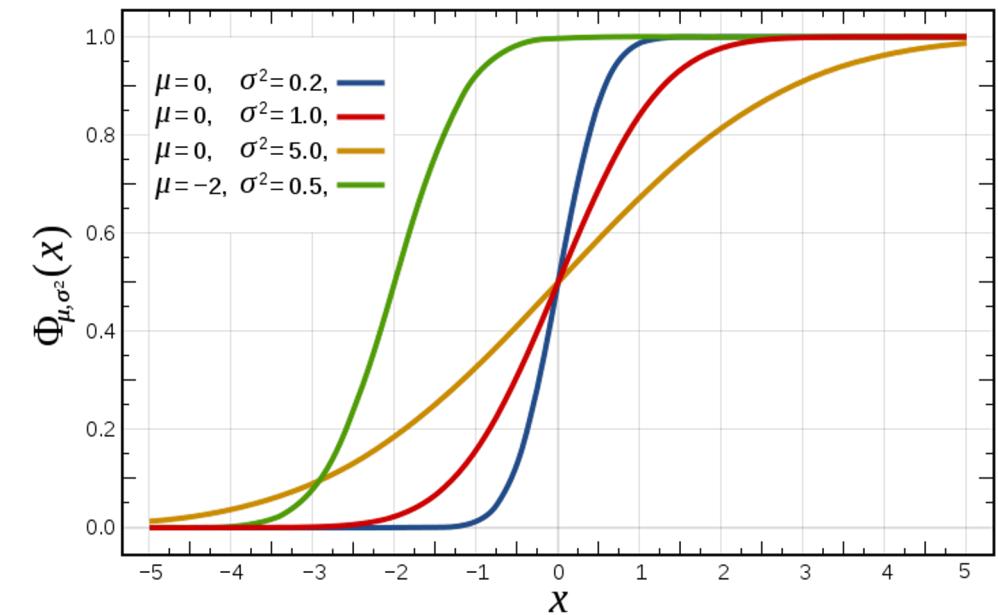
# Normal Distribution

*aka Gaussian distribution, bell curve*

$$PDF = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$



*CDF*



By Inductiveload - self-made, Mathematica, Inkscape, Public Domain, <https://commons.wikimedia.org/w/index.php?curid=3817954>

© 2020 Institute of Data



# Other types of Probability Distributions

- Bernoulli distribution
  - The outcome of a single Bernoulli trial (e.g. success/failure, yes/no)
- Binomial distribution
  - The number of "positive occurrences" (e.g. successes, yes votes, etc.) given a fixed total number of independent occurrences
- Geometric distribution
  - Binomial-type observations but where the quantity of interest is the number of failures before the first success; a special case of the negative binomial distribution

# Statistics – Part 6

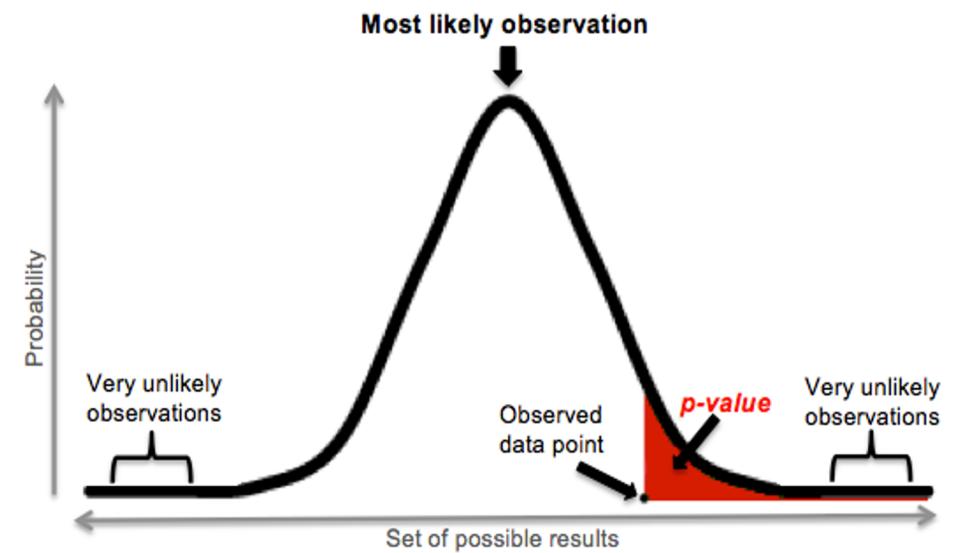
- Confidence intervals
- Significance tests and hypothesis testing
- Inference
- ANOVA



# p-Value

P-value measures the probability that a more extreme-valued sample could be randomly drawn from the distribution.

$$p(x > x') = 1 - p(x \leq x')$$



A **p-value** (shaded red area) is the probability of an observed (or more extreme) result arising by chance



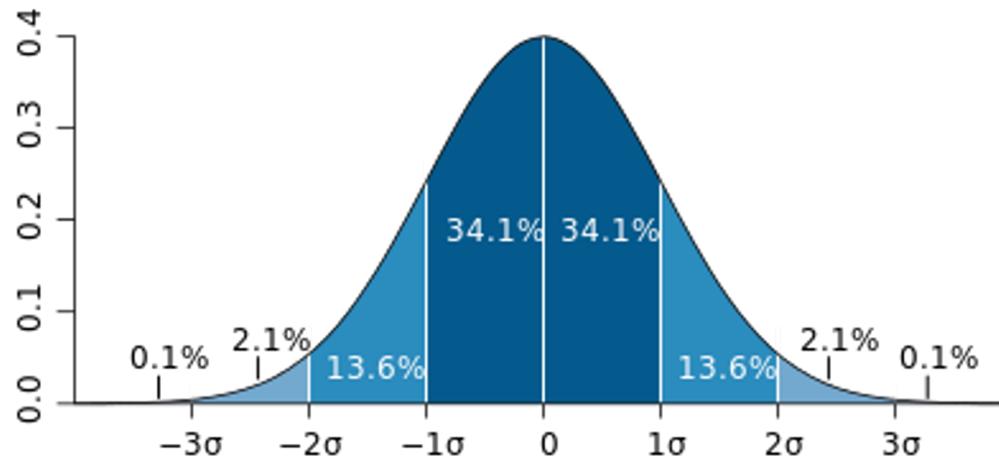
# Confidence Intervals

recall: *z-score*

measures how far a sample lies from the population mean:

$$z = \frac{x-\mu}{\sigma}$$

normal distribution:



| mean $\pm$ | % population |
|------------|--------------|
| $1\sigma$  | 68.2         |
| $2\sigma$  | 95.4         |
| $3\sigma$  | 99.7         |



## Confidence Intervals – cont'd

We define confidence intervals in terms of target probability bands:

| confidence interval | mean $\pm$      | p-value |
|---------------------|-----------------|---------|
| 0.68                | $\sim 1 \sigma$ | 0.32    |
| 0.95                | $\sim 2 \sigma$ | 0.05    |
| 0.99                | $\sim 3 \sigma$ | 0.01    |

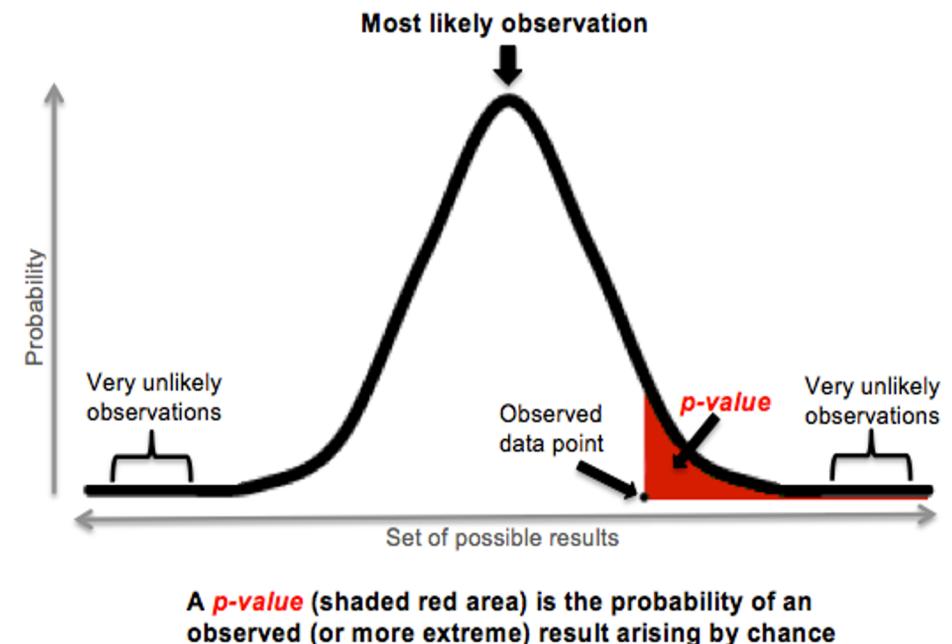


# Significance Tests

- Given a specified confidence interval, is a particular sample close enough to the mean that we can confidently presume that comes from the same population?
- Conversely, can we say that the new sample is *different* enough that it probably comes from a different population?

*example:*

if we choose  $p = 0.05$ , then  
a new sample  $x^o > \bar{x} + 2\sigma$   
then we can say that  $x^o$  is  
significantly greater than  $\bar{x}$   
(for a 95% confidence interval)



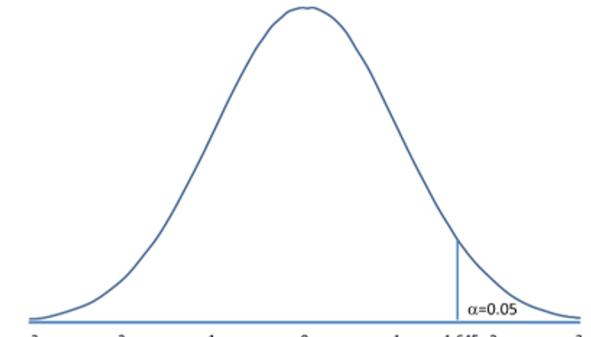


# One-Tailed Test vs Two-Tailed Test

**One-tailed test:** is B greater than A?

a 95% confidence interval would mean we are interested in the last 5% of the right tail

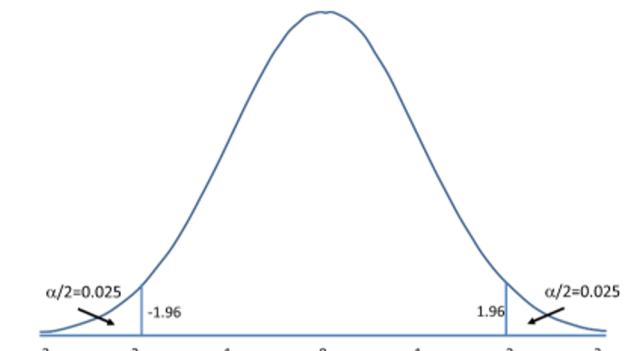
*Nb. for “Is B less than A” we would be looking at the left tail instead of the right.*



standard  
normal  
distribution

**Two-tailed test:** is B different from A?

a 95% confidence interval would mean we are interested in the last 2.5% of each tail





# Standard Error of the Mean

Corrects the standard deviation for a finite population (i.e. an acquired dataset):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cong \frac{s}{\sqrt{n}}$$

$\sigma$  is the population SD  
 $s$  is the sample SD

## ***Central Limit Theorem:***

- as sample size increases, SEM approaches the population mean



# Student's *t*-Test

Corrects the z-score for a finite population (i.e. an acquired dataset):

$$t = \frac{z}{\sigma_{\bar{x}}} = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}}$$

$z$  is the z-score

$\sigma_{\bar{x}}$  is the standard error of the mean

$\mu, \sigma$  are the population mean, SD

$\bar{x}, s$  are the sample mean, SD



# Null Hypothesis

If we want to test whether  $B$  is different from  $A$ , we first assume that it is not. Then we test to see if the difference between  $A$  and  $B$  is likely to occur by random chance.

If the difference between  $A$  and  $B$  exceeds the confidence interval, we reject the null hypothesis, and infer that  $B$  is not from the same population.



# ANOVA

- for comparing multiple groups, repeated application of the  $t$ -test would randomly give rise to apparent significance
- ANOVA avoids this error by introducing the  $F$ -test (analogous to the  $t$ -test but for more than 2 groups)
- You can use SciPy to estimate variations between two or more groups



# Probability

- Basic theoretical probability
- Bayesian inference
- Probability using sample spaces
- Basic set operations
- Permutations and combinations
- Conditional probability and independence



# Probability

If  $A, B$  are independent events, the likelihood of ...

$A$  occurring =  $P(A)$

$A$  not occurring =  $1 - P(A)$

both occurring (and) =  $P(A \cap B) = P(A) P(B)$

either occurring (or) =  $P(A \cup B) = P(A) + P(B)$

$A$  occurring if  $B$  occurs (conditional) =  $P(A|B) = \frac{P(A \cap B)}{P(B)}$



# Sample Space

*def:* The set of all possible outcomes of an experiment

- Ordered: sequence is important
- Unordered: sequence is ignored

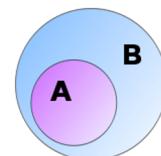
$$\Omega = \{s_1, s_2, \dots, s_n\}$$

$$P(A) = \frac{\text{number of outcomes in event } A}{\text{number of outcomes in sample space } \Omega}$$



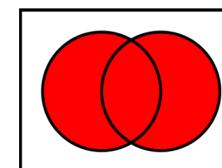
# Set Operations

$$A \subseteq B$$



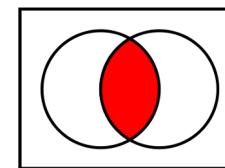
subset

$$A \cup B$$



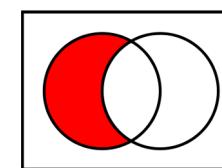
union

$$A \cap B$$



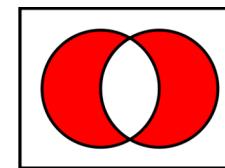
intersection

$$A - B$$



relative complement of  $B$  in  $A$

$$A \Delta B$$



symmetric difference of  $A$  and  $B$



# Permutations and Combinations

Permutation: an ordered set

Combination: an unordered set

number of  $k$ -permutations of  $n$ :

$$P(n, k) = n(n - 1)(n - 2) \cdots (n - k + 1)$$

number of  $k$ -combinations of  $n$ :

$$C(n, k) = \frac{n!}{(n - k)! k!}$$



# Bayes' inference theorem

- Bayes' inference theorem used to update the probability for a hypothesis as more evidence or information becomes available.
- Theorem:
  - $P(H|E) = P(E|H).P(H)/P(E)$
- Definition:
  - $P(A|B)$ : The probability of a event A given B



# Lab 1.1.4: Applying statistical thinking using Python

## Purpose:

- Explore how to use Python (and related packages) to apply Statistical Thinking on data.

## Materials:

- Notebook: ‘Statistics – part 2’

## Note:

- There may not be enough time to complete this lab in the class.  
Please complete it as a part of your homework.  
This should apply to all labs.



# Questions?

# End of Presentation!