ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 6: The Normal Distribution

William Marsh

Quiz (1,2,3)

Outline

- Aim: Introduce ‘normal’ distribution

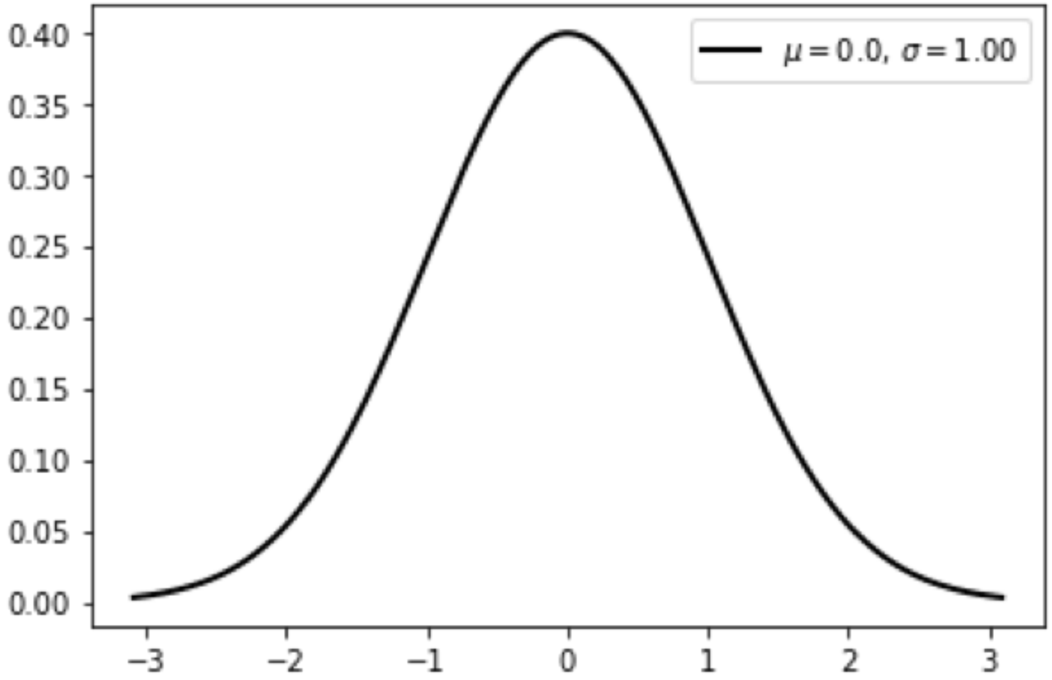
- The Normal distribution
- Variance and standard deviation
- Z score
- Normality testing and QQ plots

Introducing the Normal Distribution

‘Normal’ is NOT normal

Normal; Bell Curve

- Origin: measurement error
- Names
  - Normal
  - Gaussian
  - ‘Bell’ curve
- Symmetric around mean



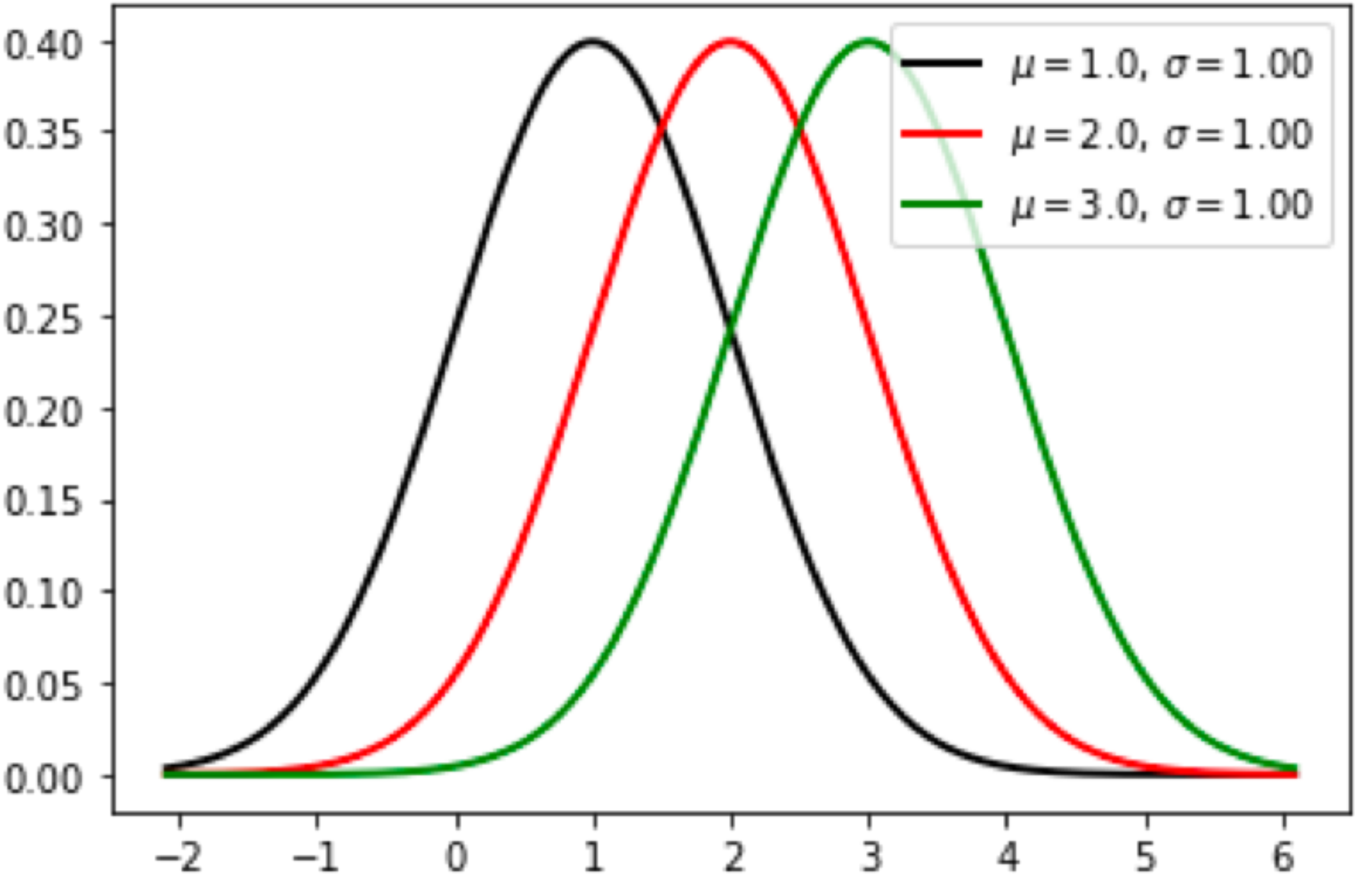
- Two parameters
  - Mean: where the centre is
  - Standard deviation: how wide distribution is

Parameters

- Two parameters
  - Mean: where the centre is
  - Standard deviation: how wide distribution is
- Mean: μ
- Standard deviation: σ

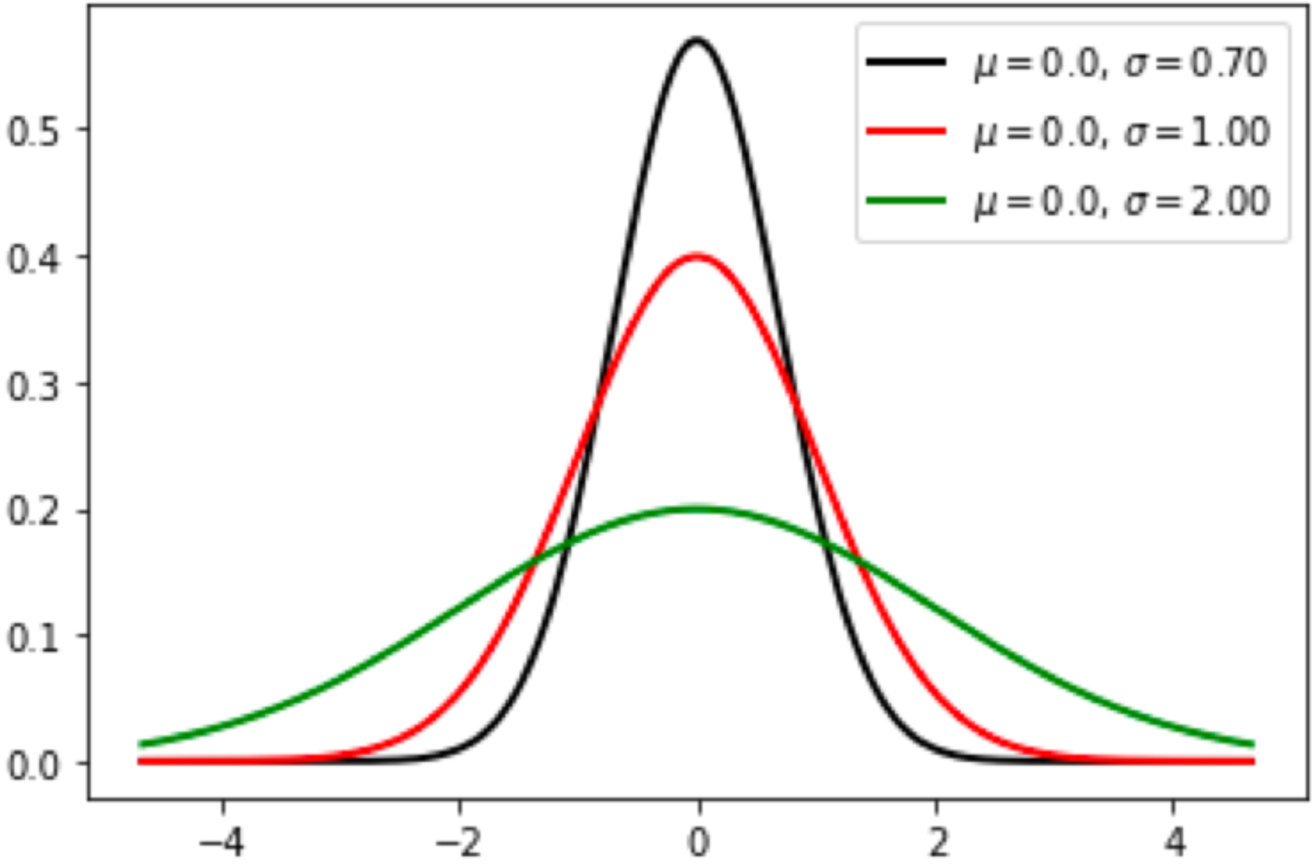
Mean

- Mean: μ
  - Same meaning as before
  - Mean, medium, mode are all equal



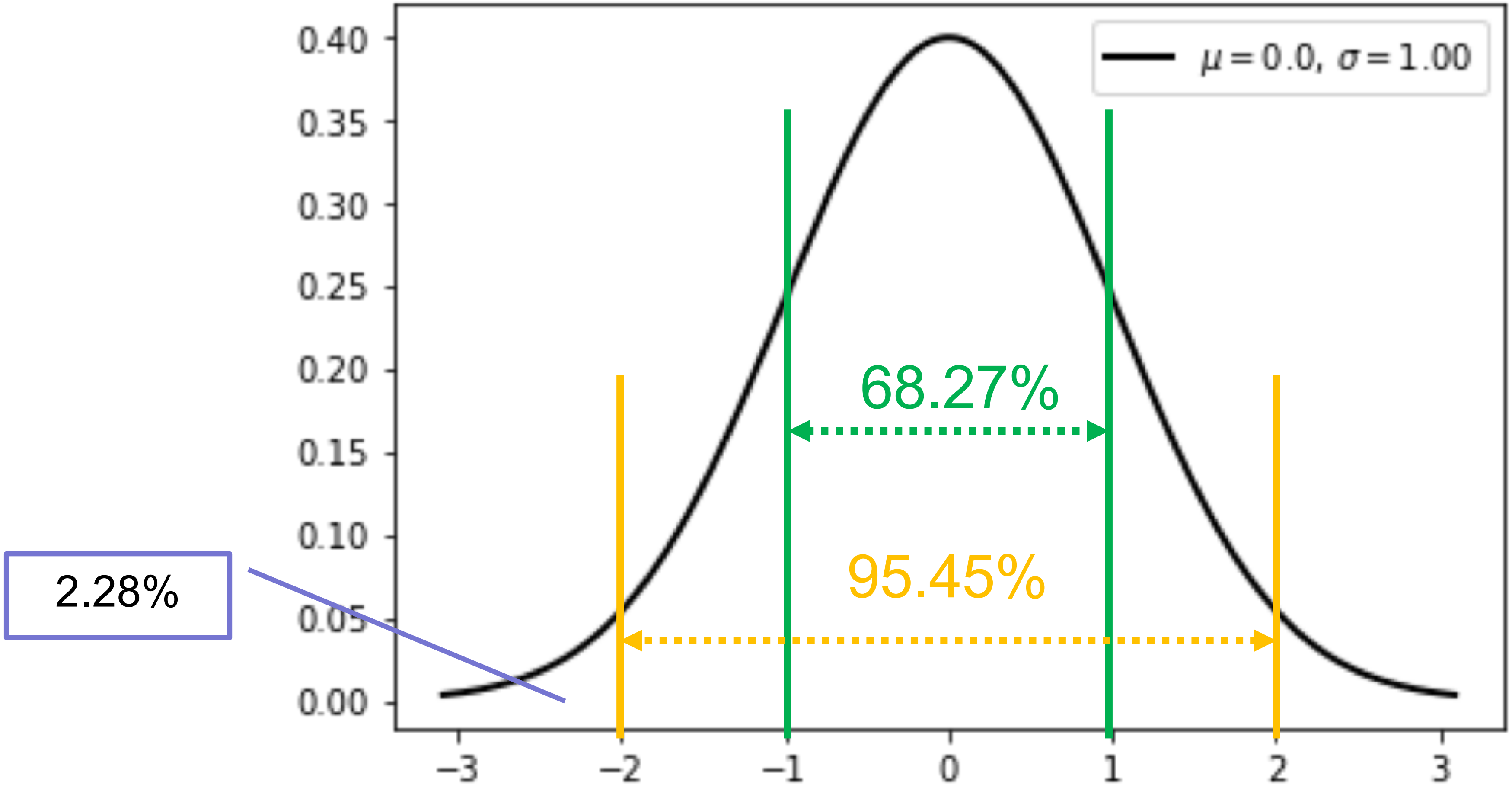
Standard Derivation

- Standard deviation: σ
  - How far the distribution stretches on either side of the mean



Where are most cases?

- Recall: area corresponds to probability



2.28%

Every lecture will have a ‘learning reflection’ slide

Should a ‘Data Analysis’ be Readable?

The notebook format allows us to create a program that is a document

Data Analysis: Telling a Story

What to Cover

- Looking at the data
  - Variable types
  - Ranges and distributions
- Relationship
  - Scatter and correlations
  - Group means
  - Conditional probabilities
- Modelling
- Statistical tests
- Conclusions

Document Structure

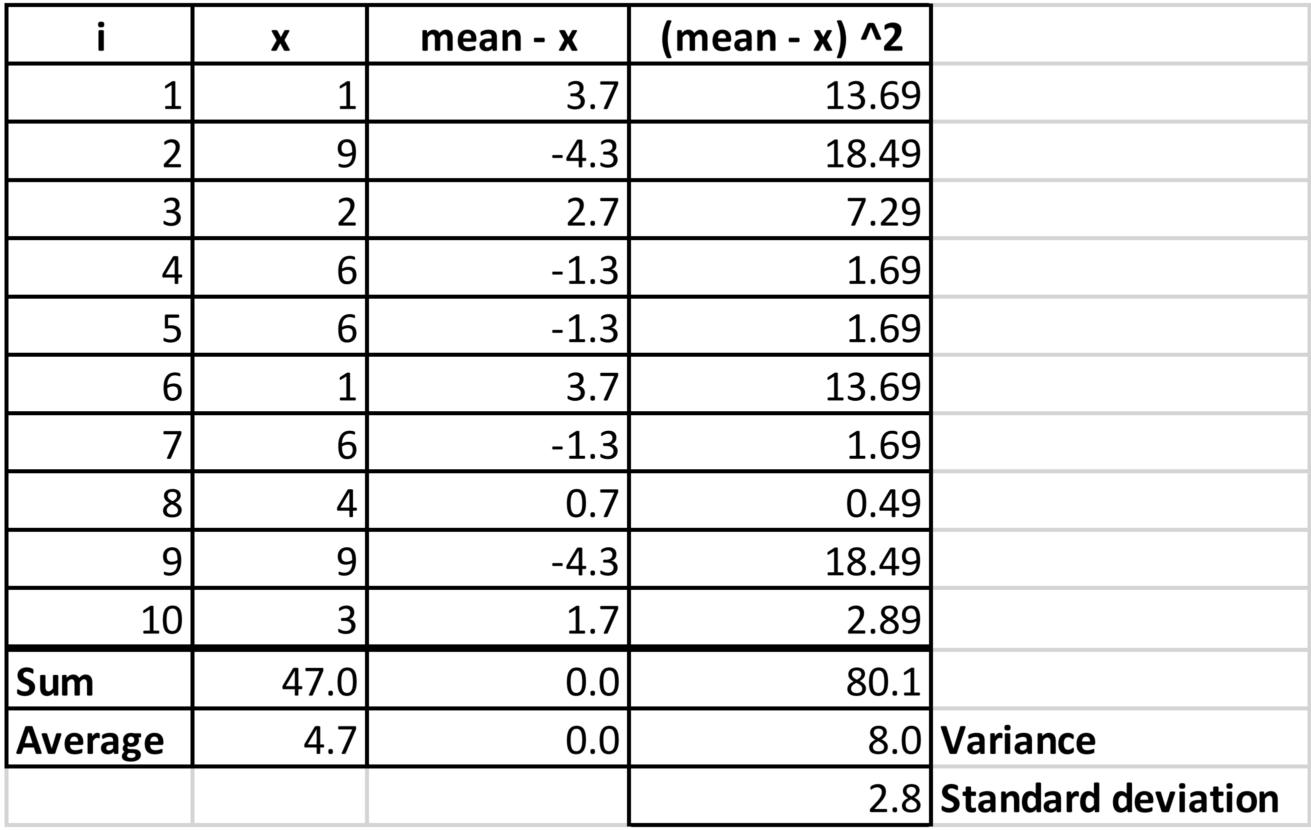
- Title
- Table of contents
- Section headers and sub-heading
- Short code cells
- Narrative: using markdown

Variance and Standard Deviation

How Wide is My Distribution?

- Idea: average distance from the mean
  - average of (x – mean)
- Problem
  - Some data points x > mean
  - Some data points x < mean
  - Average of difference is zero
- Resolution
  - Variance = Average ($(x - \text{mean})^2$)
  - Standard deviation = square root (Variance)

Mean and Variance



i	x	mean - x	(mean - x) ^2	
1	1	3.7	13.69	
2	9	-4.3	18.49	
3	2	2.7	7.29	
4	6	-1.3	1.69	
5	6	-1.3	1.69	
6	1	3.7	13.69	
7	6	-1.3	1.69	
8	4	0.7	0.49	
9	9	-4.3	18.49	
10	3	1.7	2.89	
Sum	47.0	0.0	80.1	
Average	4.7	0.0	8.0	**Variance**
			2.8	**Standard deviation**

- Mean(xs) = sum(xs) / N
- Variance = Mean ((x – mean)^2)
- Standard derivation = Variance$^{1/2}$

$$\mu = \frac{1}{N}\sum_{i=1}^{i=N} x_i$$

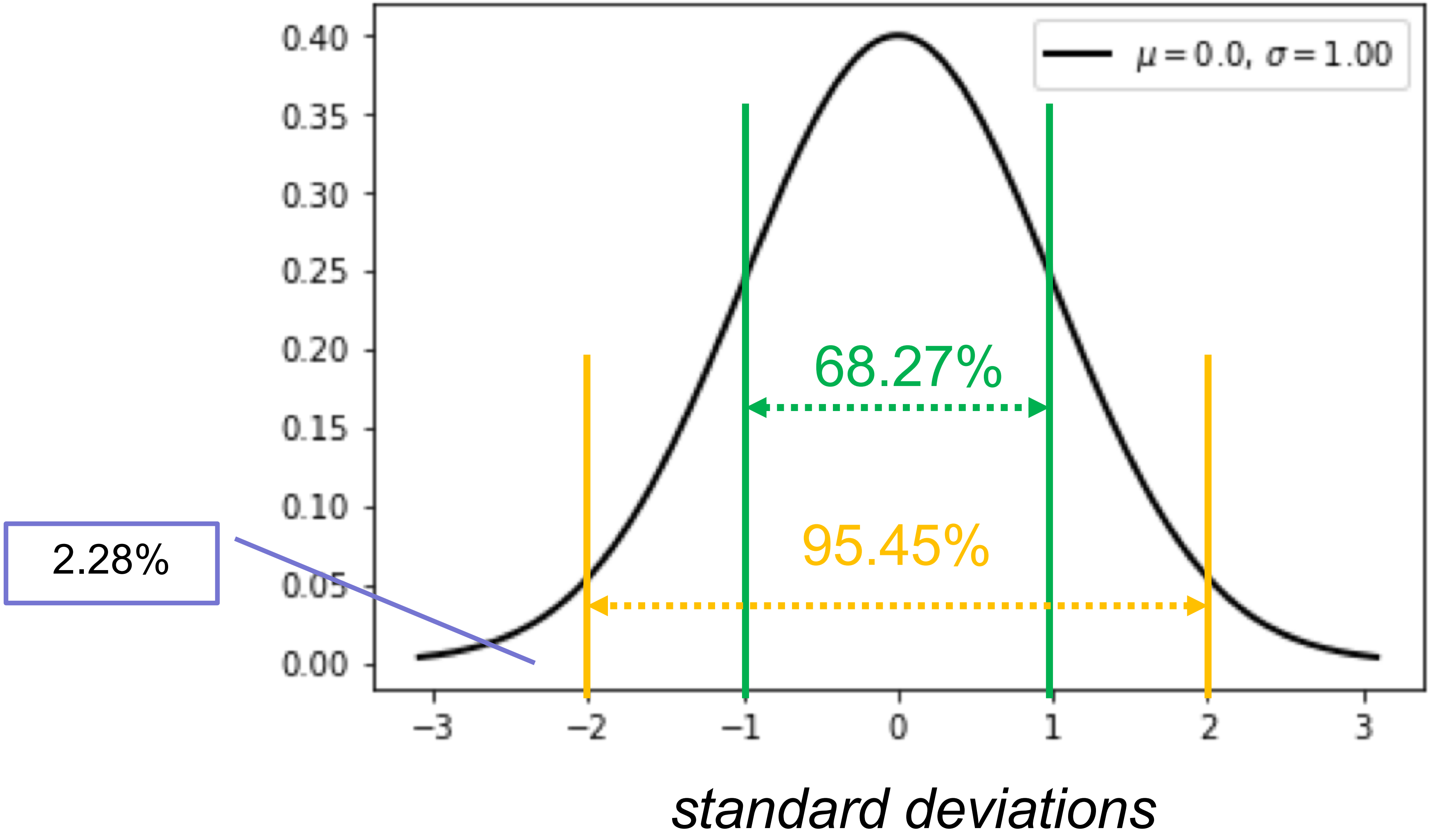
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N} (x-\mu)^2}$$

Mean and Variance

- Mean (μ) and variance ($\nu = \sigma^2$) are parameters of the normal distribution
- Any distribution has a mean and variance

Standard Deviations from the Normal

- Recall: area corresponds to probability



The Financial Crisis

*On 13 August 2007, The Financial Times reported Viniar's explanation of why two large hedge funds managed by Goldman Sachs had both lost over a quarter of their value in a week, requiring the injection of $3 billion to support them. Viniar ascribed the events to a series of exceptional events: "**We were seeing things that were 25 standard deviation moves, several days in a row**". This has since been used to illustrate the problems of inappropriate mathematical models in finance, especially those based on the assumption of Normality.*

From https://en.wikipedia.org/wiki/David_Viniar

- David Viniar was the CFO at Goldman Sachs
- Distribution of loss assumed to be 'normal'
  - Very large losses very improbable
  - 'Fat tails' – created by correlated events

Normal Formula

- When μ = 1 and σ = 1

Density

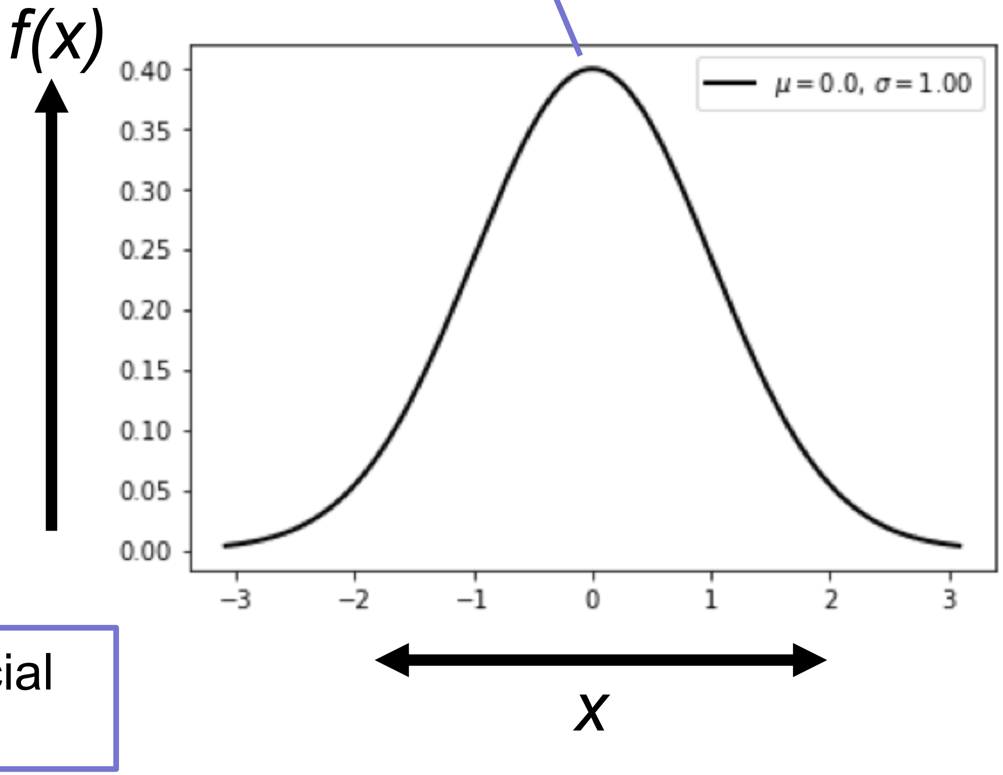
Decreases as |x| increases

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

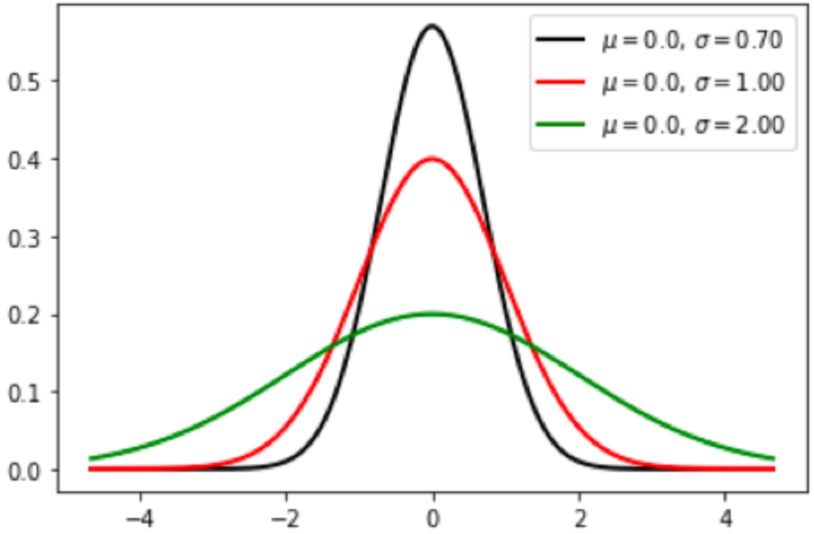
Max when x = 0 of $\frac{1}{\sqrt{2\pi}}$

Normalising constant: area must total 1

e is a special number



Normal Formula II



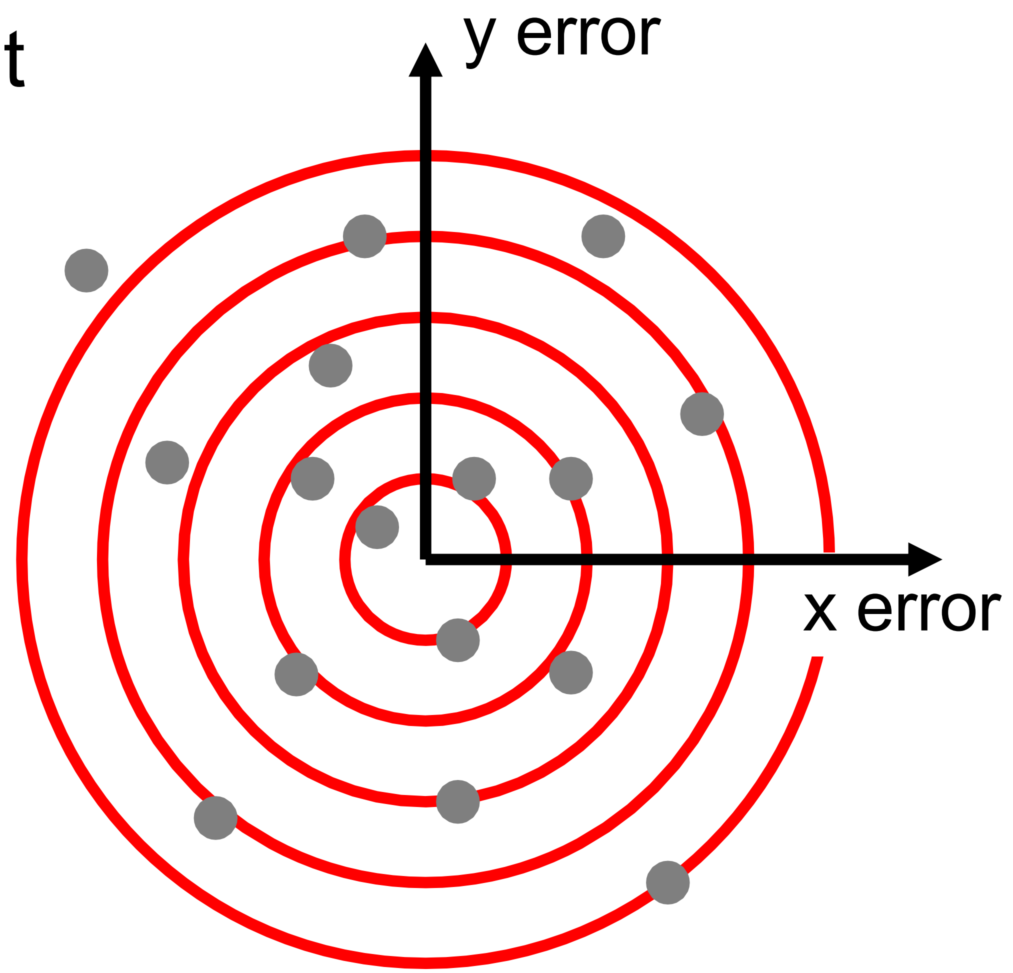
- General μ and σ
- Family of curves

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

- Z score $z = \frac{x-\mu}{\sigma}$
  - Converts x to a standardised value = ‘standard derivations from the mean’

Where Does Normal Come From?

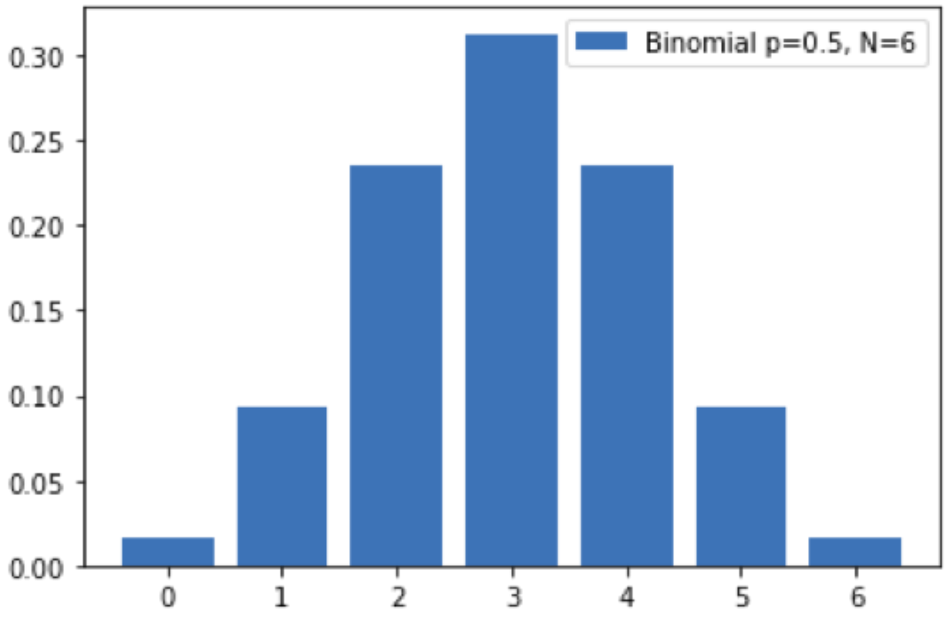
- Imagine an infinite target
  - Aiming at centre
  - Probability is area x density
- Assumptions
  - Error in x independent of error in y
  - Density depends only on distance from aim
- See youtube https://www.youtube.com/watch?v=cTyPuZ9-JZ0

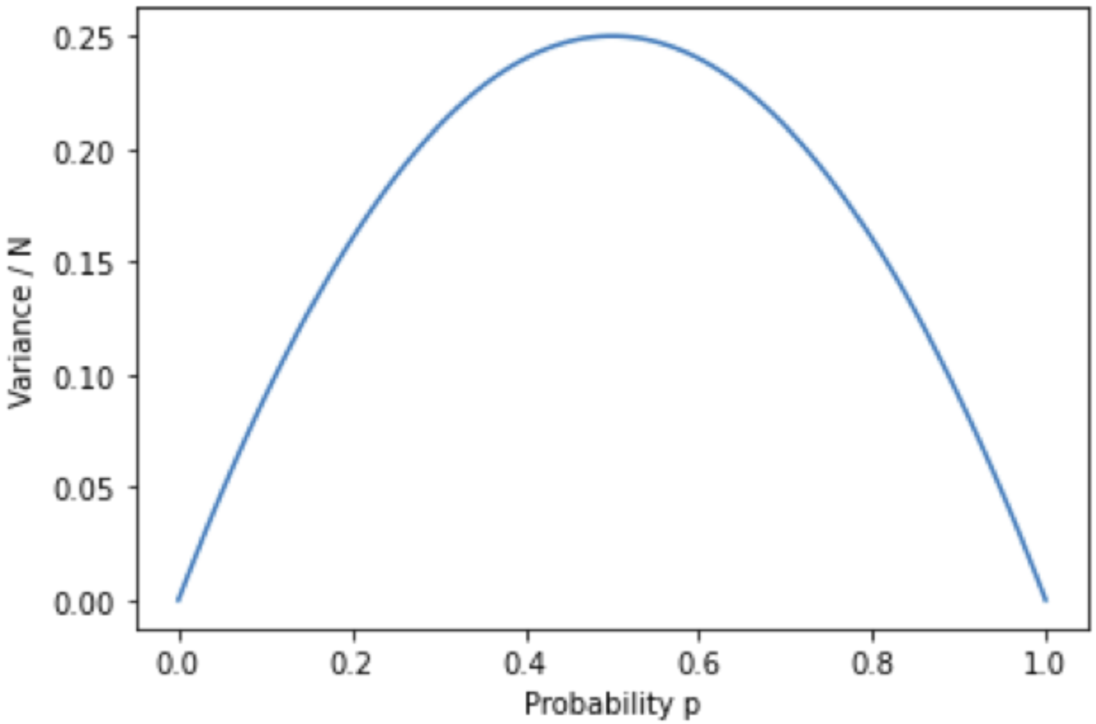


Mean and Variance of Binomial

Binomial(p, n)

- Number of trials = n
- Probability = p
- Mean = n.p
  - Mean is ‘expected value’
- Variance = n.p.(1-p)





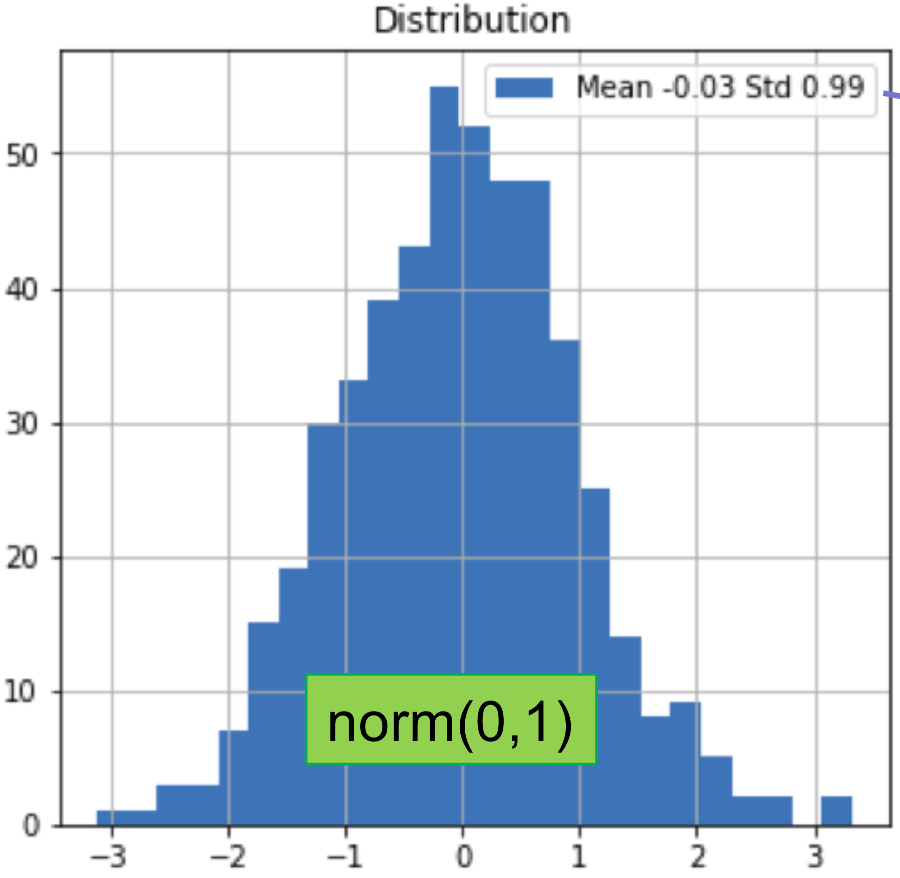
Quiz

Is a Distribution Normal?

QQ plot

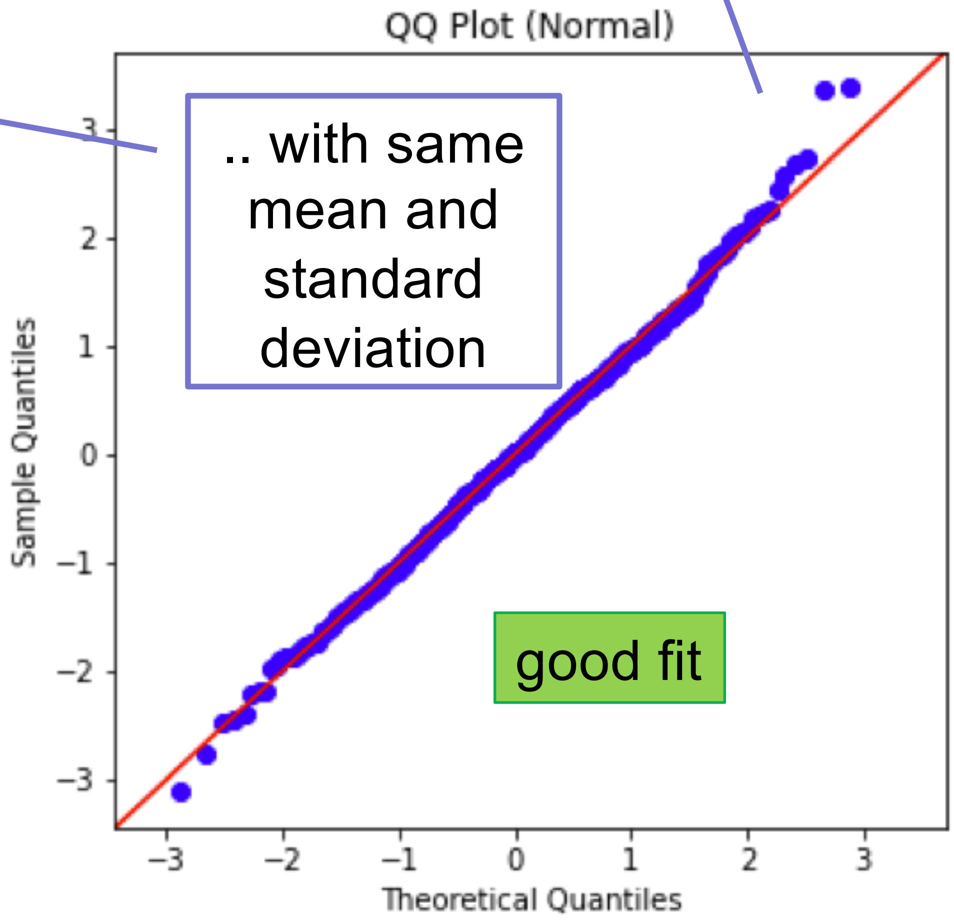
- Compare quantiles of set of values against Normal

Distribution of a sample



norm(0,1)

Quantiles of sample versus quantiles of normal ...

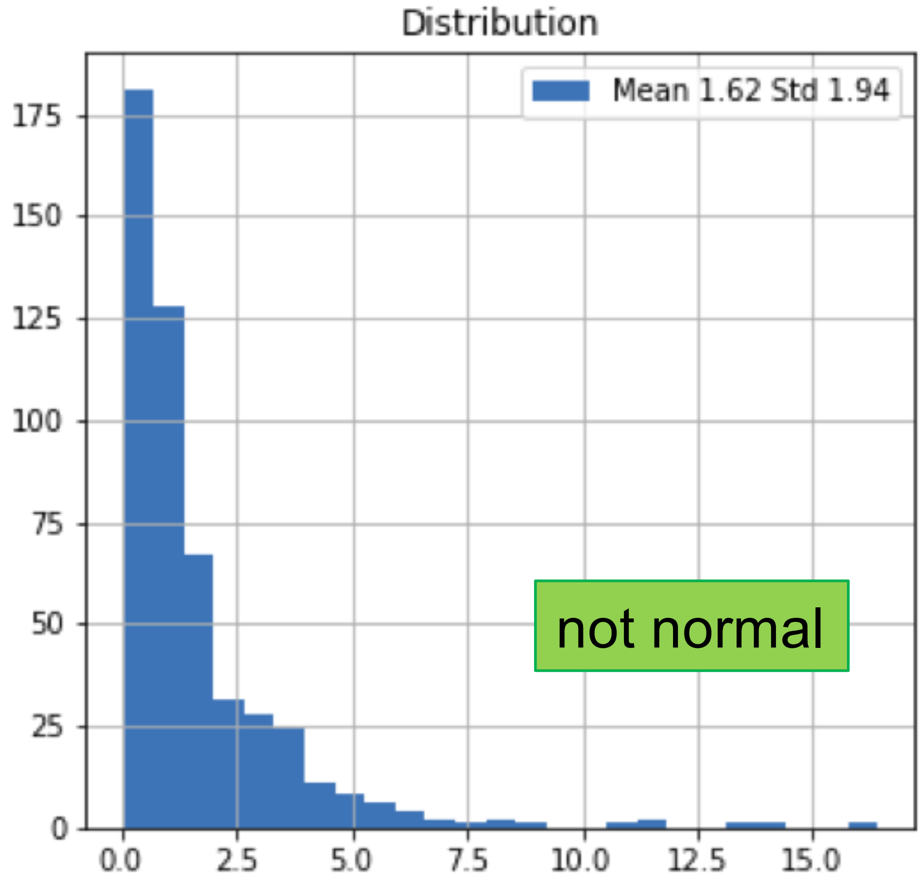


.. with same mean and standard deviation

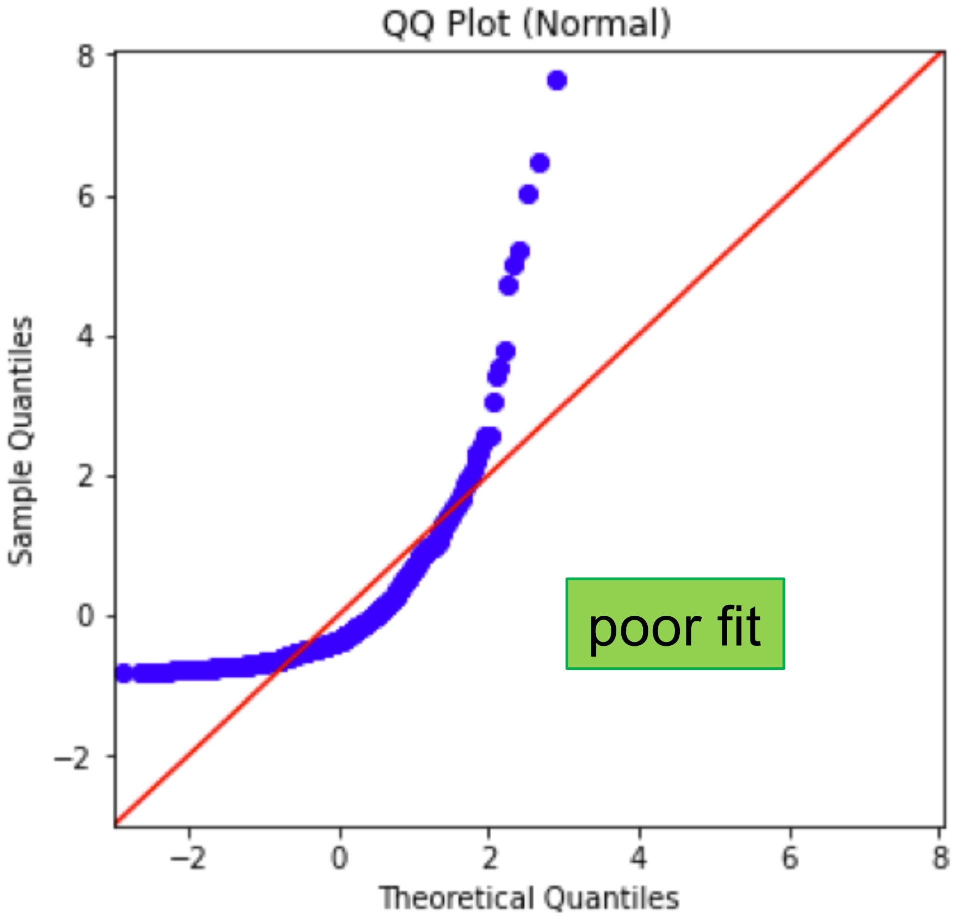
good fit

QQ plot

- Positive skew



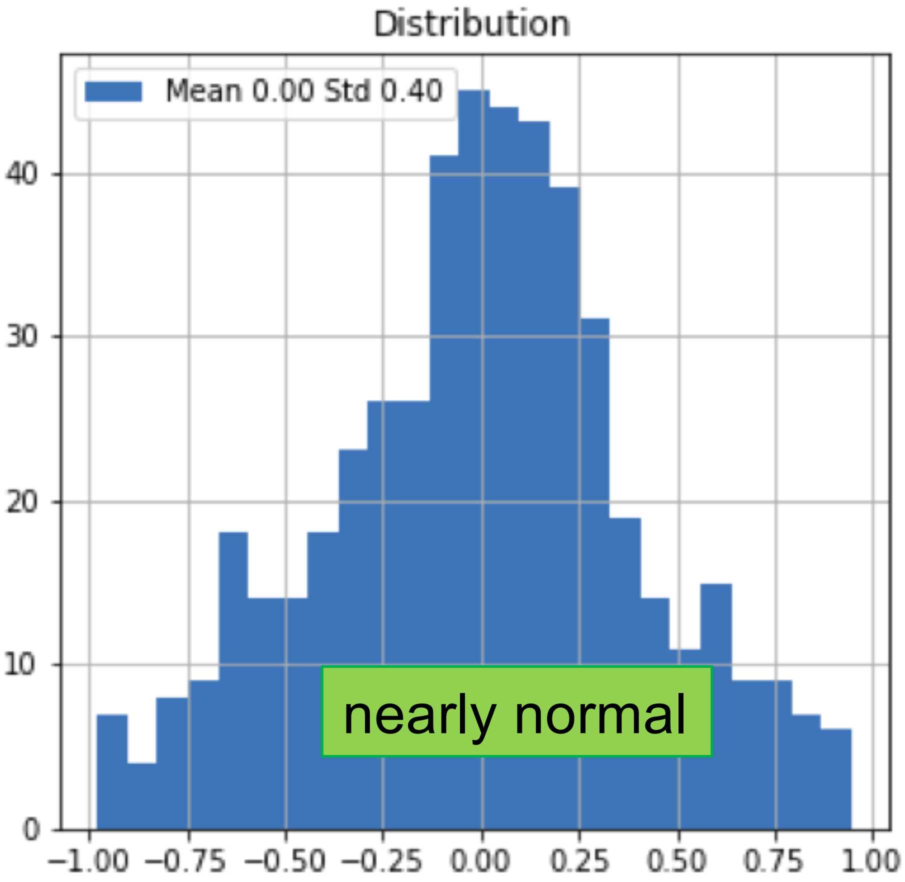
not normal



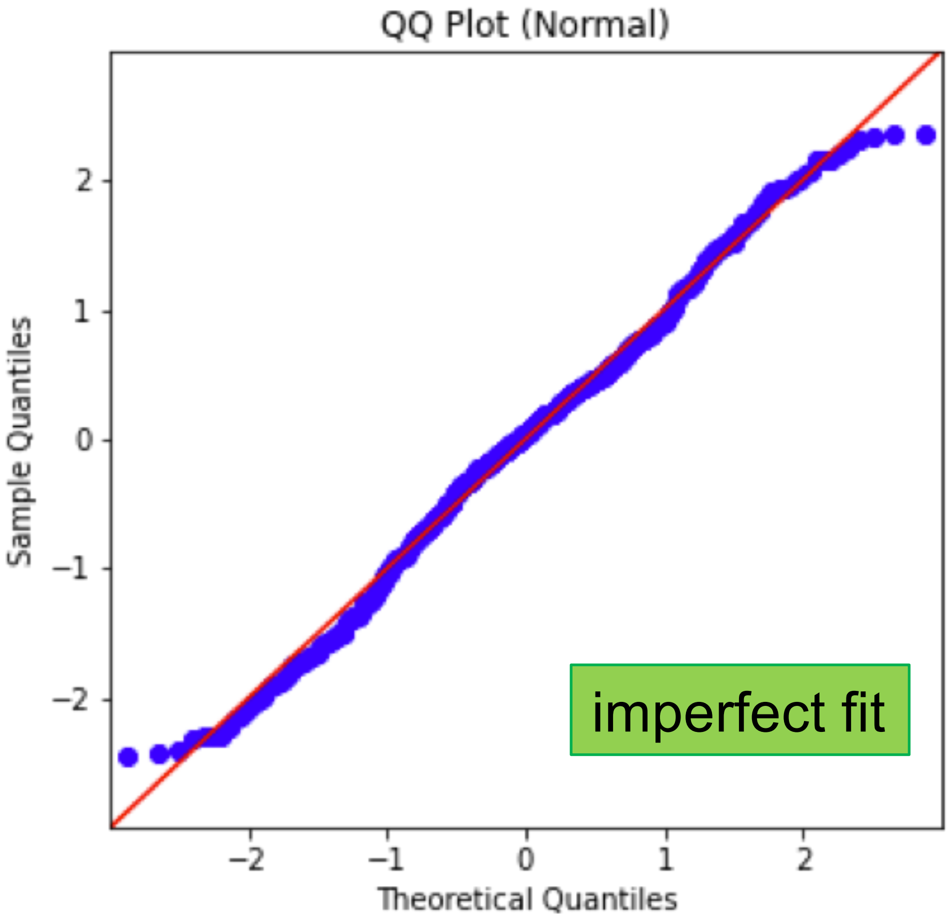
poor fit

QQ plot

Fat tails – less spread



nearly normal



imperfect fit

Summary

- Normal (or Gaussian) distribution
  - Symmetric
  - Two parameters mean and variance (std dev)
  - Arises from ‘errors’ or ‘combined variation’

- Other distributions also have a Variance
  - How spread out is the distribution?
  - Variance = (Standard deviation)^2

- QQPlot uses quantiles to see whether a data fits a distribution (such as normal)

Recommended video: https://www.youtube.com/watch?v=RKdB1d5-OE0