University of East London

Pioneering Futures Since 1898

Advanced Data Analysis

Dr Amin Karami

a.karami@uel.ac.uk

www.aminkarami.com

CN5209 – Week 4

21 October 2019

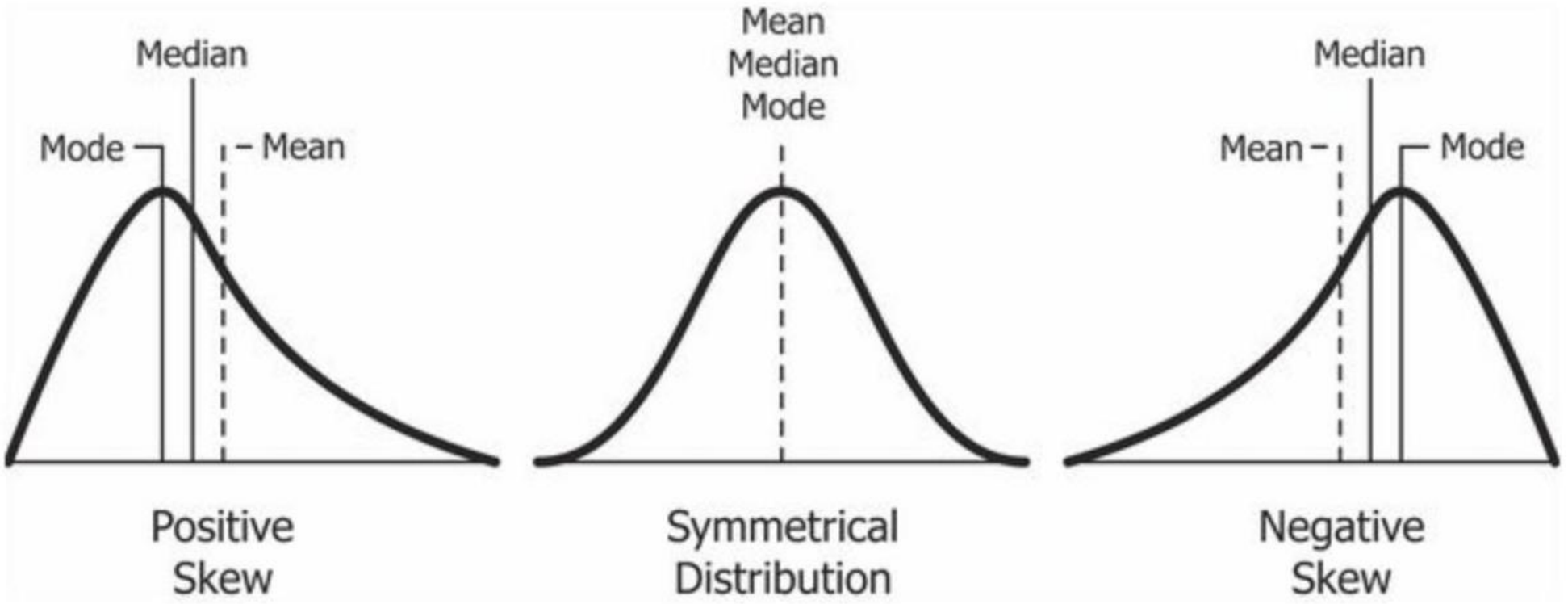
Outline

- Data Distribution/Shape by Skewness and Kurtosis
- Generate random numbers (discrete and continuous)
- Correlation analysis by Variance, Covariance, and Correlation Coefficient

Learning Outcomes

- Understanding data distribution/shape
- To be able to generate random numbers
- Understanding the relation/correlation between variables through correlation analysis

Skewness and Averaging



Right skewed

Left skewed

Skewness

- **Skewness** is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the right of the mean than to the left. If skewness is positive, the data are spread out more to the left. The skewness of the normal distribution (or any perfectly symmetric distribution) is zero.

$$g_1 = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

When is the skewness too much?

- **-0.5 < Skewness <0.5** → data are fairly symmetrical.
- **-1 < Skewness <-0.5** : negatively skewed → (moderately skewed)
- **0.5 < Skewness < 1** : positively skewed → (moderately skewed)
- **-1 > Skewness or Skewness > 1** → (highly skewed)

Example:

Suppose we have house values ranging from $100k to $1,000,000 with the average being $500,000.

1. If the peak of the distribution was left of the average value, portraying a positive skewness in the distribution. It would mean that many houses were being sold for less than the average value, i.e. $500k.
2. If the peak of the distributed data was right of the average value, that would mean a negative skew. This would mean that the houses were being sold for more than the average value.

Skewness in MATLAB

```
X = randn([5 4])
X =
   1.1650   1.6961 -1.4462 -0.3600
   0.6268   0.0591 -0.7012 -0.1356
   0.0751   1.7971  1.2460 -1.3493
   0.3516   0.2641 -0.6390 -1.2704
  -0.6965   0.8717  0.5774  0.9846
```

- **First way:**

```
y = skewness(X)
y =
   -0.2933   0.0482   0.2735   0.4641
```

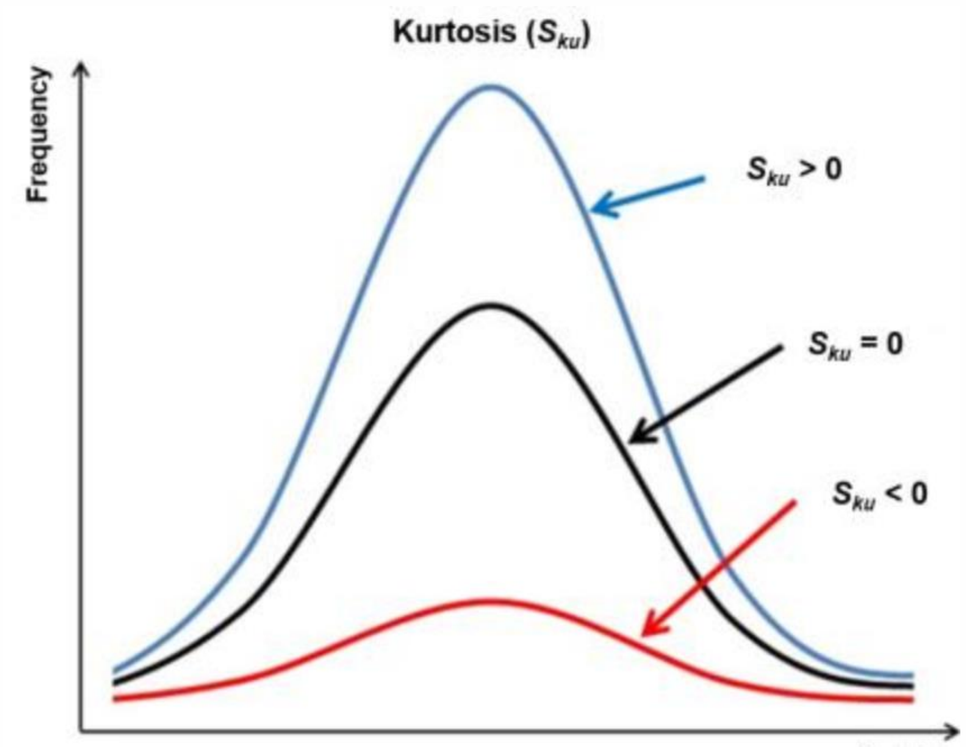
- **Second way:**

```
>> (sum((X-mean(X)).^3)./length(X)) ./ (var(X,1).^1.5)

ans =

   -0.2932    0.0482    0.2735    0.4642

fx >>
```

University of East London

Kurtosis

- **Kurtosis** is about the tails of the distribution, not the peakedness or flatness. It is the measure of outliers present in the distribution.
- The kurtosis of the normal distribution is 3. Distributions that are more outlier-prone than the normal distribution have kurtosis greater than 3; distributions that are less outlier-prone have kurtosis less than 3.



$$g_2 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

Kurtosis in MATLAB

- **First way:**

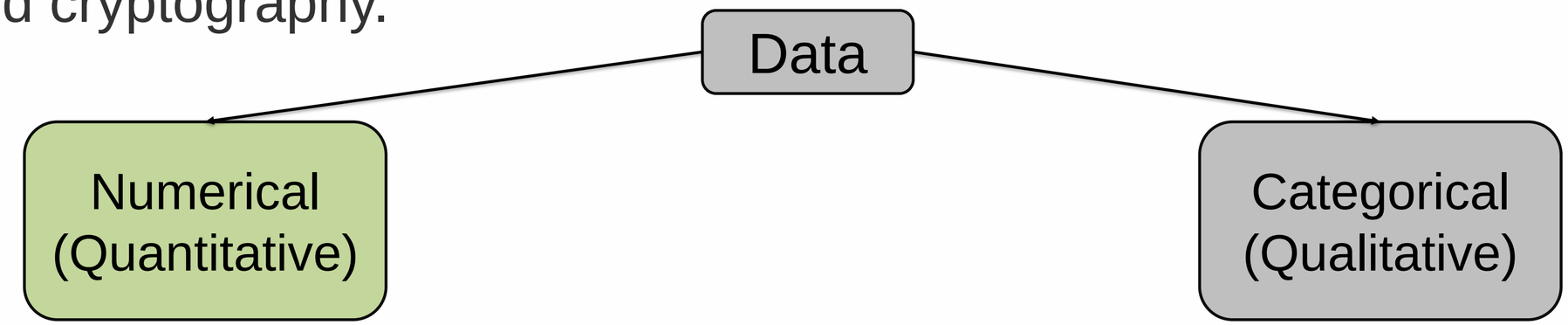
```
X = randn([5 4])
X =
  1.1650  1.6961  -1.4462  -0.3600
  0.6268  0.0591  -0.7012  -0.1356
  0.0751  1.7971   1.2460  -1.3493
  0.3516  0.2641  -0.6390  -1.2704
 -0.6965  0.8717   0.5774   0.9846

k = kurtosis(X)
k =
  2.1658  1.2967  1.6378  1.9589
```

- **Second way:**

```
>> (sum((X-mean(X)).^4)./length(X)) ./ (var(X,1).^2)

ans =

    2.1659    1.2967    1.6378    1.9589

fx >>
```

Generate Random Numbers

- A random number is a number generated using a large set of numbers and a mathematical algorithm which gives equal probability to all numbers occurring in the specified distribution. E.g., testing a system and cryptography.



- Discrete (only integer values):
  - **`randi`**
  - **`randperm`**
- Continuous (any value):
  - **`rand`**
  - **`randn`**
  - **`unifrnd`**

- Nominal
(named categories)
- Ordinal
(categories with an implied order)

`randi` function

`X = randi(imax,n)` returns a `n-by-n` matrix of pseudorandom integers drawn from the discrete uniform distribution on the interval `[1,imax]`.

- Generate a 10-by-1 column vector of uniformly distributed random integers from the sample interval [-5,5].
- Generate a 5-by-5 matrix of random integers between 1 and 10.

```
r = randi(10,5)
```

```
r = 5×5

     9     1     2     2     7
    10     3    10     5     1
     2     6    10    10     9
    10    10     5     8    10
     7    10     9    10     7
```

```
>> r= randi([-5,5],[10,1])

r =

     3
     4
    -4
     5
     1
    -4
    -2
     1
     5
     5
```

`randperm` function

- `randperm(n,k)` returns a row vector containing `k` unique integers selected randomly from `1` to `n` inclusive.

```
>> randperm(10,5)

ans =

     3     4     1     9    10
```

```
>> randperm(8)

ans =

     8     6     7     3     5     4     1     2

>> randperm(8)

ans =

     2     4     5     3     8     7     1     6
```

Probability Distributions

- In probability theory and statistics, a probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.

1. **Uniform** Distribution
2. **Normal** Distribution
3. Poisson Distribution
4. Binomial Distribution
5. Bernoulli Distribution
6. Exponential Distribution

`rand/randn`: probability distributions

- `rand(n)` returns a `n-by-n` matrix of uniformly distributed random numbers between `[0 1]`.

```
>> rand(3)

ans =

    0.9502    0.3816    0.1869
    0.0344    0.7655    0.4898
    0.4387    0.7952    0.4456
```

```
>> rand(2,5)

ans =

    0.5472    0.1493    0.8407    0.8143    0.9293
    0.1386    0.2575    0.2543    0.2435    0.3500
```

- `randn(n)` returns a `n-by-n` matrix of normally distributed random numbers between `[-1 1]`.

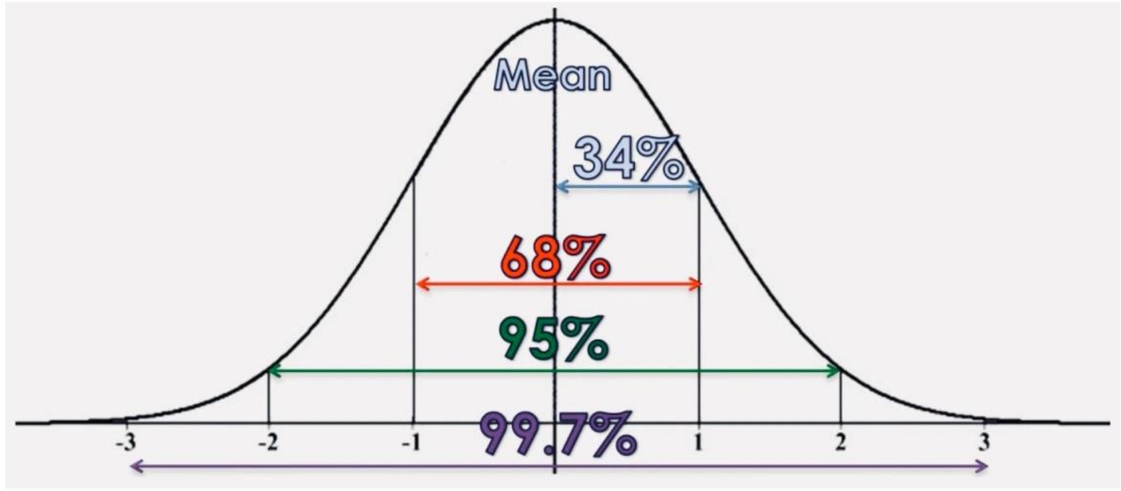
```
>> randn(3)

ans =

    0.6007   -0.0068    0.3714
   -1.2141    1.5326   -0.2256
   -1.1135   -0.7697    1.1174
```

```
>> randn(2,5)

ans =

   -1.4023    0.4882   -0.1961    0.2916    1.5877
   -1.4224   -0.1774    1.4193    0.1978   -0.8045
```

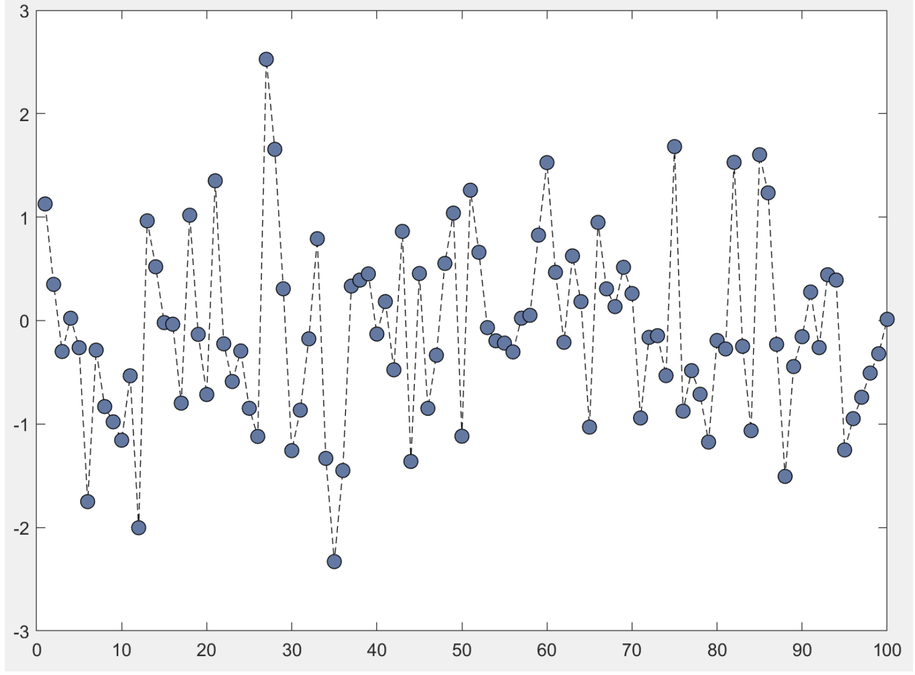
University of East London

Normal Distribution

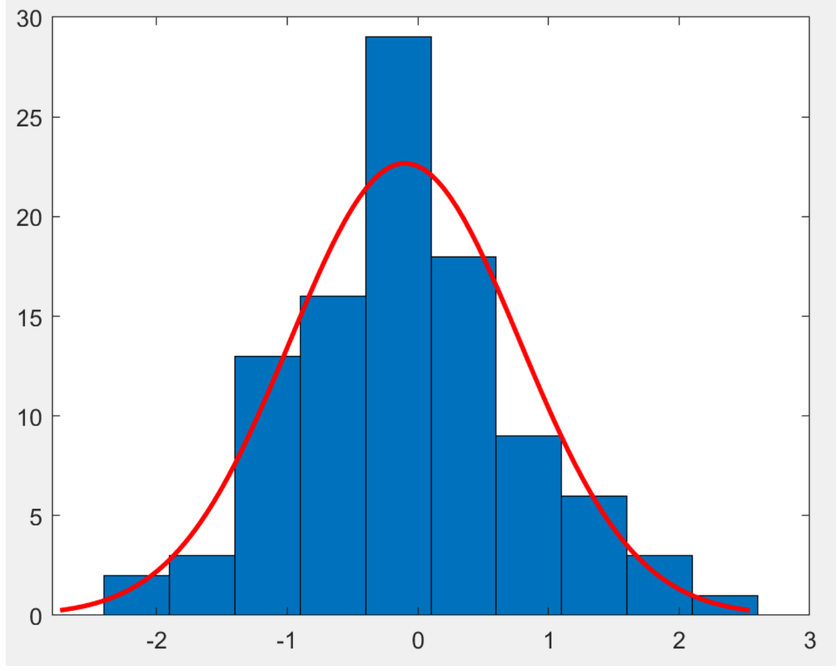
- The standard normal distribution has three properties:
  - The graph is bell-shaped.
  - The mean is 0 (μ = 0).
  - The standard deviation is 1 (σ = 1).




```
y=randn(1,100)
```



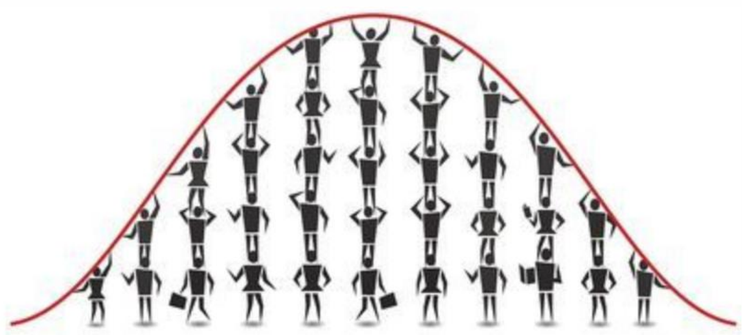
Distribution



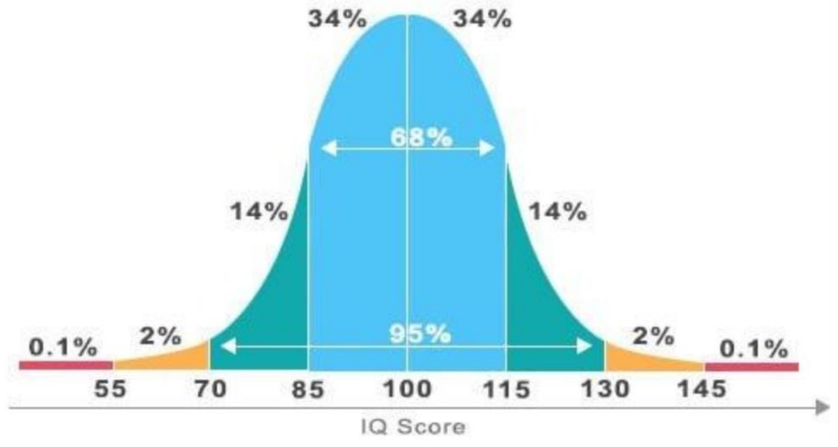
University of East London

Real life examples for Normal Dist.

- **Height:** The number of people taller and shorter than the average height people is almost equal, and a very small number of people are either extremely tall or extremely short.



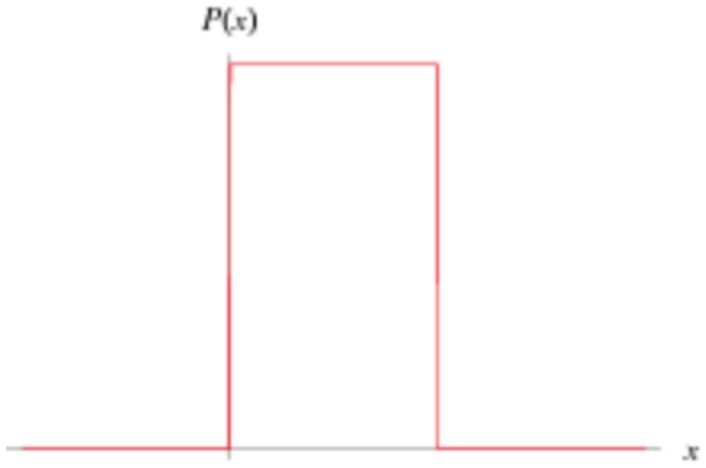
- **IQ Test:** IQ of a majority of the people in the population lies in the normal range whereas the IQ of the rest of the population lies in the deviated range.



- Other examples: Technical stock market, Income distribution in economy, Shoe size, Birth weight, Students' marks, etc.

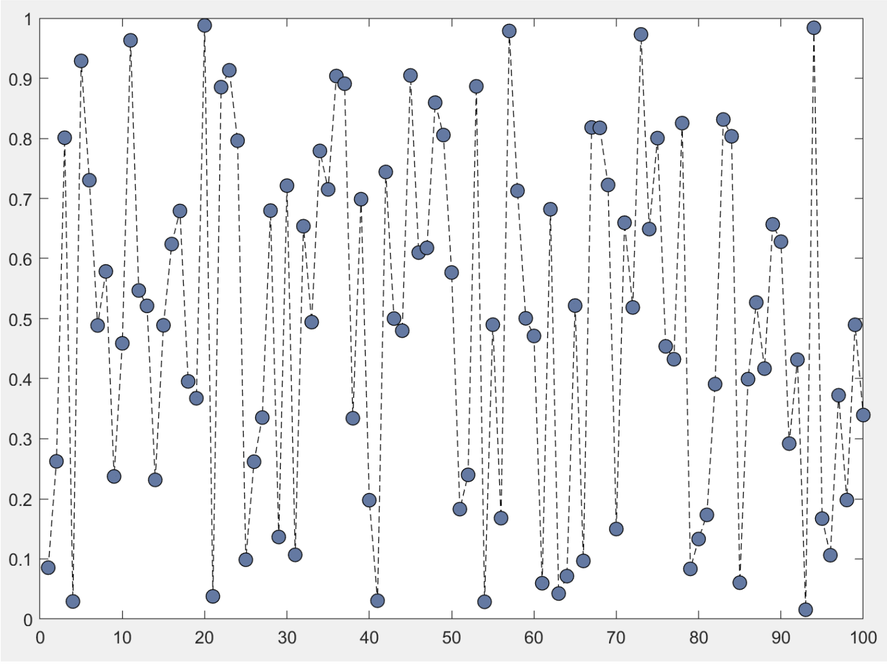
Uniform (Rectangular) Distribution

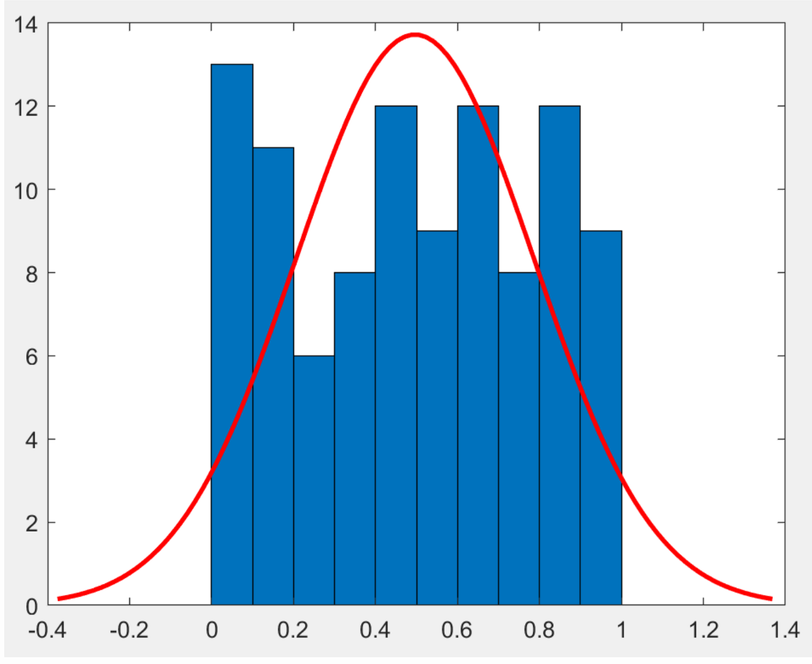
- A continuous random variable has a uniform distribution if its values are spread evenly over the range of probabilities. The graph of a uniform distribution results in a rectangular shape.



Distribution

```
y=rand(1,100)
```





University of East London

Real life example for Uniform Dist.

- In a **uniform distribution,** all values between two boundaries occur roughly equally. If you roll a six-sided die, you're equally likely to get 1, 2, 3, 4, 5, or 6. If you rolled it 6,000 times, you'd probably get roughly 1,000 of each result.

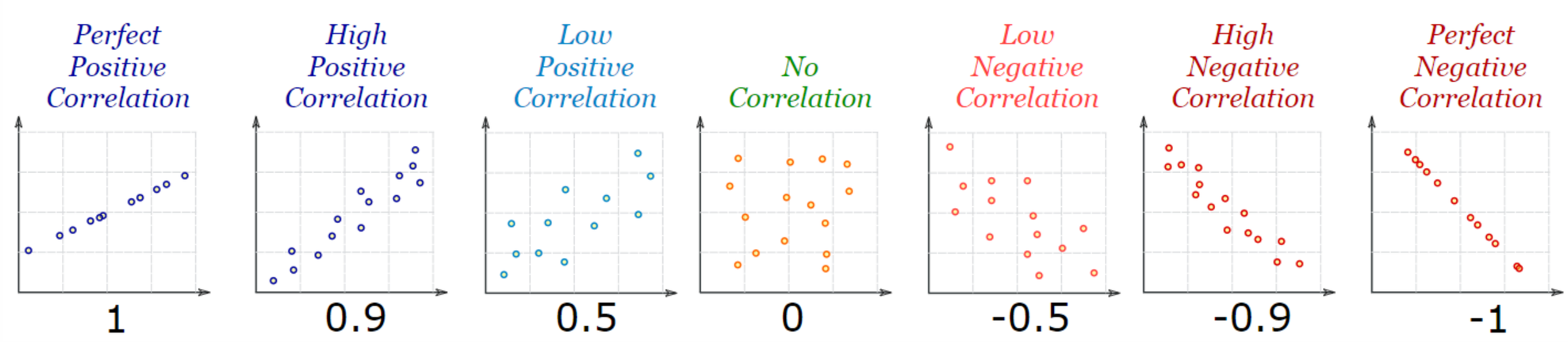
Correlation

- It is a statistical technique used to determine the degree to which two variables are related.
- Is there a relationship between `x` and `y`?
- What is the strength of this relationship? `Pearson's`
- Can we describe this relationship and use this to predict `y` from `x`? `Regression` (how well a certain independent variable predict dependent variable)
- Is the relationship we have described statistically significant? `T-test`

University of East London

Correlation

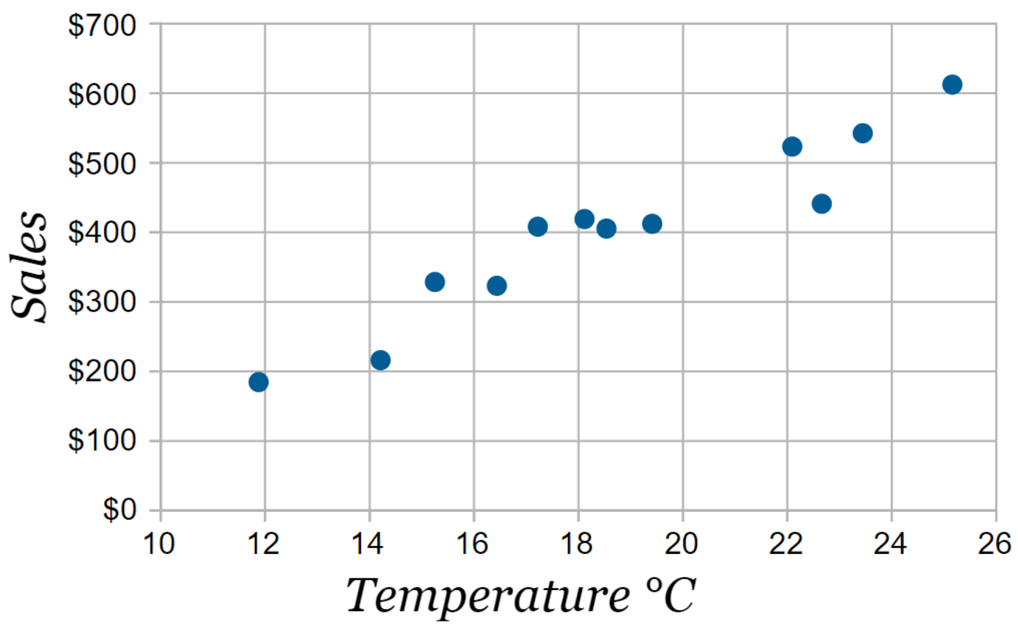
- Correlation is **Positive** when the values increase together.
- Correlation is **Negative** when one value decreases as the other increases.



1: is a perfect positive correlation
0: is no correlation (the values don't seem linked at all)
-1: is a perfect negative correlation

Example

The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days.



Correlation: **0.9575**

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

Example

Variance:

• Gives information on variability of a single variable.

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

Covariance:

• Gives information on the degree to which two variables vary together.

• Note how similar the covariance is to variance: the equation simply multiplies x's error scores by y's error scores as opposed to squaring x's error scores.

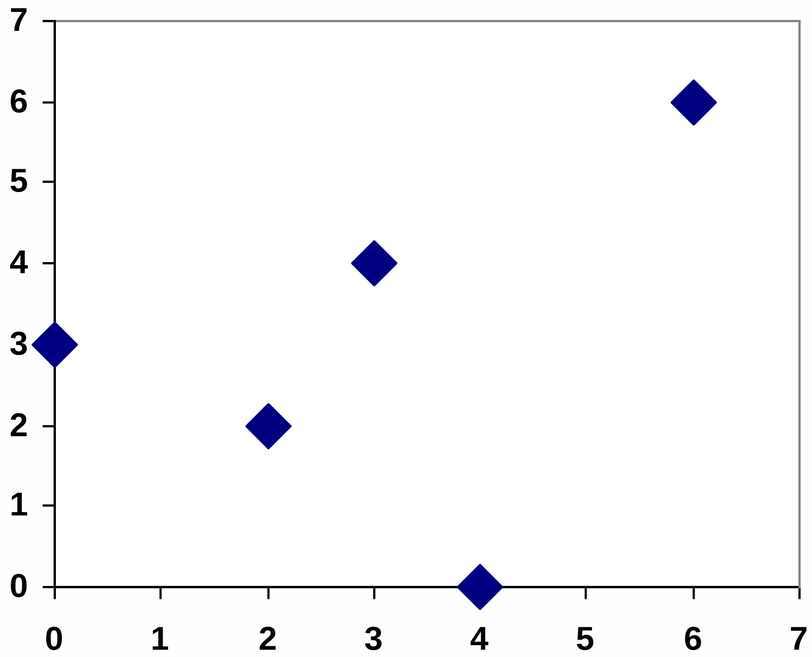
$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Covariance

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- When X ↑ and Y ↑: cov (x,y) = +
- When X ↓ and Y ↑: cov (x,y) = -
- When no constant relationship: cov (x,y) = 0

Covariance: example



x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
0	3	-3	0	0
2	2	-1	-1	1
3	4	0	1	0
4	0	1	-3	-3
6	6	3	3	9
$\bar{x} = 3$	$\bar{y} = 3$			$\sum = 7$

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}))}{n-1} = \frac{7}{4} = 1.75$$

What does 1.75 mean?

University of East London

How Covariance relies on Variance?

	High variance data				Low variance data		
Subject	**x**	**y**	**x error * y error**		**x**	**y**	**X error * y error**
1	101	100	2500		54	53	9
2	81	80	900		53	52	4
3	61	60	100		52	51	1
4	51	50	0		51	50	0
5	41	40	100		50	49	1
6	21	20	900		49	48	4
7	1	0	2500		48	47	9
Mean	51	50			51	50	
Sum of x error * y error :			7000		**Sum of x error * y error :**		28
Covariance:			**1166.67**		**Covariance:**		**4.67**

Solution for Covariance

- Covariance does not really tell us anything
  - **Solution**: standardise this measure

- **Pearson**: standardises the covariance value. It called `correlation coefficient` method.
  - Divides the covariance by the multiplied standard deviations of X and Y:

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

Correlation Coefficient methods

`'type'` — Correlation coefficient
`'Pearson'` (default) | `'Kendall'` | `'Spearman'`

Correlation coefficient to compute, specified as the comma-separated pair consisting of `'type'` and one of the following:

`'Pearson'`	Pearson's linear correlation coefficient
`'Kendall'`	Kendall's rank correlation coefficient (τ)
`'Spearman'`	Spearman's rank correlation coefficient (ρ)

An example of a Negative Correlation

```
>> Math = randi([40 100], 1,15)

Math =

    68    54    91    51    53    50    53    66    58    96    66    51    95    99    66

>> Physics = randi([40 100], 1,15)

Physics =

    46    55    64    76    55    76    83    53    47    58    59    65    70    45    56
```

```
>> corr(Math', Physics','type','pearson')

ans =

   -0.2860

>> corr(Math', Physics','type','kendall')

ans =

   -0.3350

>> corr(Math', Physics','type','spearman')

ans =

   -0.4820
```

- We have two classes of math and physics with final year marks (assessment). we are going to find out the relationship between two assessments.
- Is it positive/negative correlation?

Summary

- Explained data distributions
- Discussed random numbers
- Exercised correlation between variables
- Discussed correlation coefficient methods

University of East London