MOD550
Fundaments of Machine Learning for and with Engineering Applications

University of Stavanger

01 – Uni- and Bivariate Statistics

Reidar B Bratvold

and

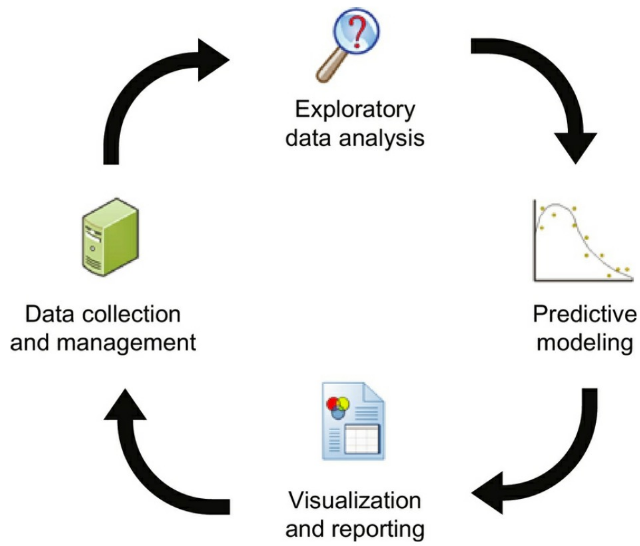
Enrico Riccardi

1

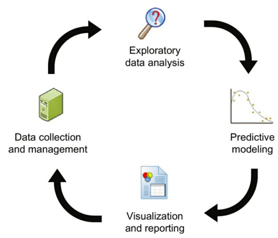
Introduction

2

Integrated Data Analysis Cycle

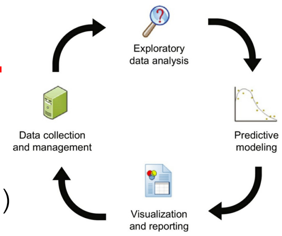


Integrated Data Analysis Cycle



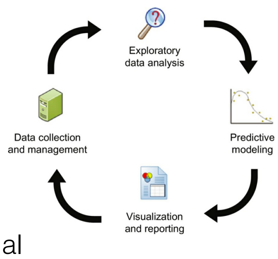
- Data collection and management.
  - Involves the acquisition and aggregation of data from multiple sources (application dependent but could be cores, well logs, and production records), possibly in multiple forms (e.g., numbers and text)
  - The data also undergo a QA/QC process to ensure the traceability and accuracy of each data record
  - Finally, the data have to be made easily available for visualization and analysis. This involves "data cleaning"
- Exploratory data analysis.
  - The goal of this step is to develop a preliminary understanding of the data in terms of the characteristics of individual variables and the relationship among various variables.
  - Other objectives include identifying key variables of interest, formulating questions for digging deeper into the data, and selecting techniques that will be used for detailed analysis.

Integrated Data Analysis Cycle



- Predictive modeling.
  - Begin with unsupervised learning, where the issues of redundancy among the independent variables and possible reduction in data dimensionality (without losing any information) are addressed
  - Supervised learning, where observed values of a response variable are used to train a model between the independent variables (i.e., predictors) and the dependent variable (i.e., response)
  - This predictive model can then be used to answer questions posted in the previous step

Integrated Data Analysis Cycle



- Visualization and reporting.
  - The ultimate goal of any modeling and/or analysis is to provide input for a decision by transferring information to decision-makers
  - Necessary to capture what has been learned in the form of visual summaries, compact reports, or decision-support tools that can be used to answer "what-if" type questions (sensitivity analysis)
  - The use of insights from predictive modeling to identify what new data should be collected and the kinds of questions to pursue in the future

Exploratory Data Analysis

- Concerned with summarizing and visualizing data as a starting point for more detailed analyses
- We restrict ourselves to numerical data (as opposed to text or images) and note that:
  - data can be univariate or multivariate,
  - data can be categorical or numerical,
  - random variables can have more than one value, and
  - distributions capture the values taken by variables, and the frequency with each specific value occurs.

Random Variables and their types

Random Variables

- A random variable is a real valued function that assigns a value to each outcome in the sample space.
- A random variable (RV) can be either discrete or continuous.
  - Discrete RV: the number of failures in a wind turbine in a given month
  - Continuous RV: the wind speed at a given location
- The probability mass function (PMF), p, of a discrete RV, X, denotes the probability that the RV is equal to a specified value, a.

$$p(a) = p(X = a)$$

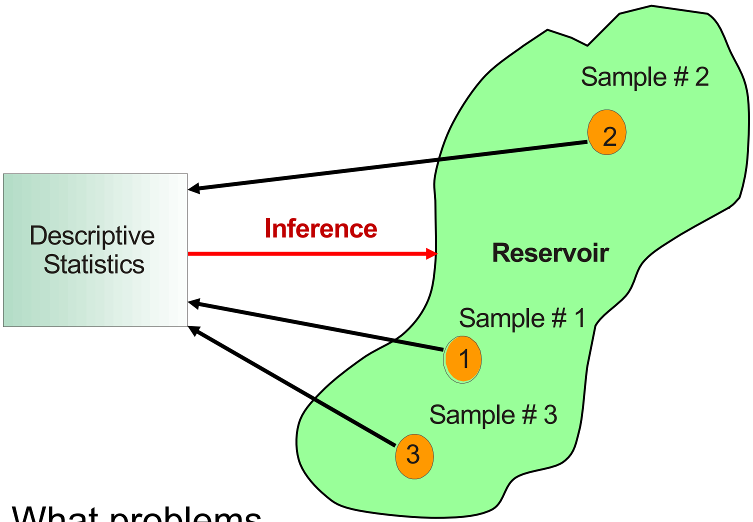
- Similarly, the cumulative distribution function (CDF), F, denotes the probability that X will take on values equal to or less than a.

$$F(A) = P(X \leq a) = \sum_i p(a_i) \; with \;\; a_i \leq a$$

Sampling

- A subset of the population
- Used to develop an understanding of the population's behaviour
- Most commonly used to predict future behaviour
- An area of specialization within statistics

Descriptive versus Inferential Process



What problems might arise?

1. Randomly sample this population using a finite number of samples, e.g., by drilling and coring wells
2. Analyze these values, e.g., permeability, to determine the proportion of samples with permeability greater than 10 mD (e.g., 65%)
3. Determine the representativeness of this result for the entire population (e.g., 95% certain that margin of error is ± 6%).

Common Sampling Questions Include

- What Are the Effective Sampling Methods for Predicting Equipment Failure in Renewable Energy Facilities?
  - This could involve time-series data and predicting maintenance needs or equipment lifespan
- How to Sample and Analyze Energy Consumption Data from Households Using Solar Panels to Determine the Efficiency of Solar Energy?
  - Focus on understanding usage patterns and how they correlate with solar energy production
- Where should we locate appraisal wells and how should we adjust results for the “non-random” choice?
- Was the change in production a result of operations or merely a chance fluctuation?

Some Applications

- Computing summary statistics (e.g., mean and variance)
- Determining conditional probabilities of cause-effect relationships
- Calculating correlation and rank correlation coefficients between two variables
- Visualizing univariate, bivariate, and multivariate data
- Estimating probability coverage levels for different distributions
- Analyzing behavior of normal and lognormal distributions
- Calculating confidence interval and sampling distribution for the mean
- Testing for significance of difference in means
- Comparing two different distributions for statistical equivalence
- Fitting simple and multiple linear regression models to observed data
- Developing a nonparametric regression model from given data
- Reducing data dimensionality with principal component analysis
- Grouping data with k-means and hierarchical clustering
- Identifying classification boundary between clusters using discriminant analysis
- Developing distributions from data, limited knowledge, or subjective judgment
- Translating model input uncertainty into uncertainty in model predictions using Monte Carlo simulation
- Analyzing input-output dependencies from Monte Carlo simulation results

Typical Data for Statistical Analysis

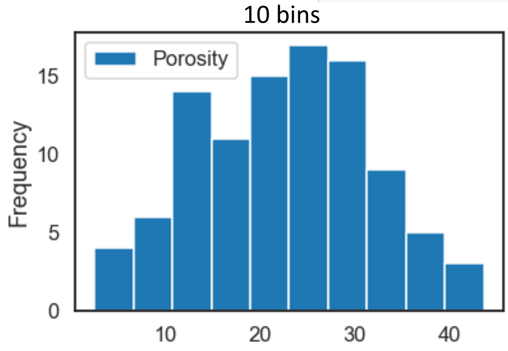
Turbine	Height	X	Y	Wind Speed	Air Density	Temperature	Power Output	Rotor Diameter	Hub Height	Air Pressure	Turbulence Intensity
WT-1	80	752.1	3945	7.5	1.225	15	1500	82	80	1013	0.1
WT-1	80	752.2	3945	8	1.223	15	1600	82	80	1012	0.12
WT-1	80	752.3	3945	7.8	1.224	16	1550	82	80	1013	0.11
WT-2	90	753.5	3946	6.5	1.226	14	1400	85	90	1012	0.15
WT-2	90	753.6	3946	7	1.225	14	1500	85	90	1011	0.13
WT-2	90	753.7	3946	7.2	1.227	14	1520	85	90	1012	0.14

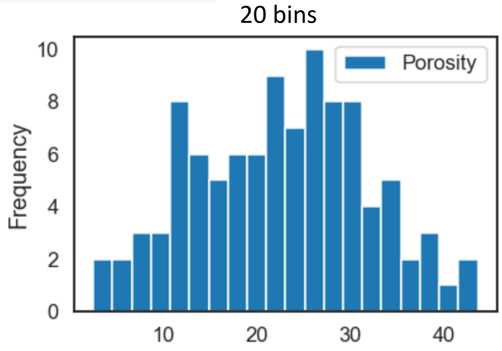
Goals of Exploratory Data Analysis

- understand the data: statistical versus geological populations
- ensure data quality – data cleaning
- condense information

Histogram – Python using Matplotlib

```
import matplotlib.pyplot as plt

# Set the number of bins
num_bins = 20  # or any other number you wish to set

plt.hist(df['Porosity'], bins=num_bins, edgecolor='black')

plt.title('Histogram of Porosity')
plt.xlabel('Value')
plt.ylabel('Frequency')

# Show the plot
plt.show()
```





Univariate Statistics Outline – Describing Sample Data

- Displaying Data
  - Histograms, frequency plots, cumulatives
- Measures of Location
  - Mean median mode
  - Quartiles, Percentiles, Quantiles
- Measures of Dispersion (Spread)
  - MAD, standard deviation (sd), variance (Var), interquartile ranges, Coefficient of Variation (CV)
- Measures of Shape
  - Skewness & kurtosis
- Summarizing Distributions

Central tendency of random variable (mean, median, mode)

17

Measures of Location: Central Tendency: Mean

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

if the data represent a random sample, i.e., each point weighted equally by $1/n$

- Every element in the data set contributes to the value of the mean
- An average provides a common measure for comparing one set of data with others
- The mean is influenced by the extreme values in the data set
- The mean may not be an actual element of the data set – example?
- The sum of all deviations from the mean is zero, and the sum of squared deviations is minimized when those deviations are measured from the mean

18

Arithmetic, Geometric & Harmonic means

- Arithmetic
  - Mean of raw data

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

- Geometric
  - n^{th} root of product
  - Mean of logarithms

$$\bar{g}_x = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

$$= Exp\left(\frac{1}{n}\sum_{i=1}^{n} ln(x_i)\right)$$

- Harmonic
  - Mean of inverses

$$\bar{h}_x = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$$

Measures of Location: Central Tendency: Median

- The central value in a data set when the data points are put in ascending or descending order

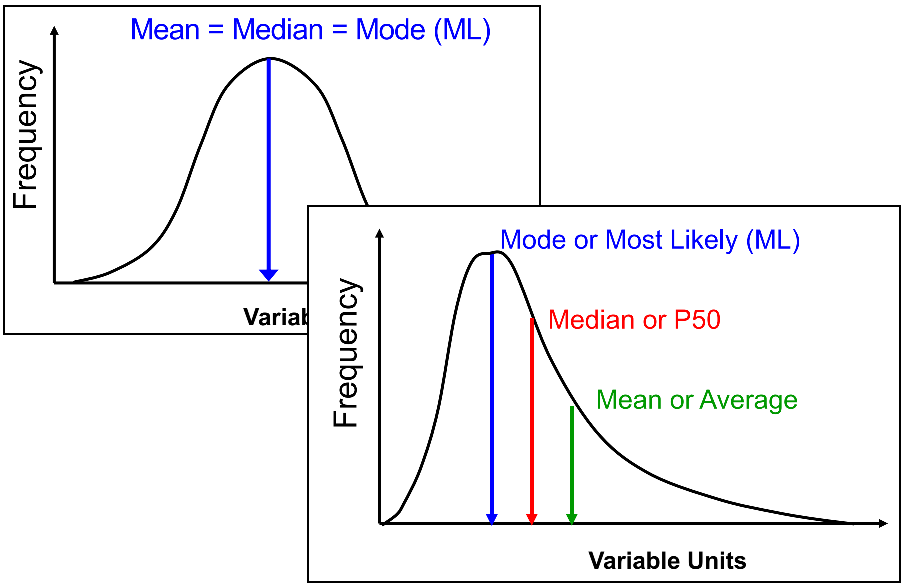
$$median = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ x_{n/2} + x_{(n/2)+1} & \text{if } n \text{ is even} \end{cases}$$

  - Average of middle two data points if n is even
  - On a cumulative frequency plot, the value on the x-axis that corresponds to 50% on the y-axis
- Not influenced by extreme values - therefore robust
  - Makes the median useful in describing the central tendency of data-sets where one extreme has not been well sampled. Or if there are dubious extreme values.
- May not be an actual value of the data set (n even)
- For a perfectly symmetrical data set, the mean = median

Measures of Location: Central Tendency: Mode

- The most frequently occurring data element
  - The most likely or most probable value (for a pmf)
- A data set may have more than one mode and is called bimodal when two data elements occur an equal number of times
  - If a data set has more than two modes, the worth of the mode becomes questionable
- The mode is unaffected by extreme values (see comments on median)
- The mode does not take into account all the values in the data set => may be misleading as a measure of central tendency
- A mode is always a data element in the set
- For a perfectly symmetrical data set: the mean, the median and the mode are the same

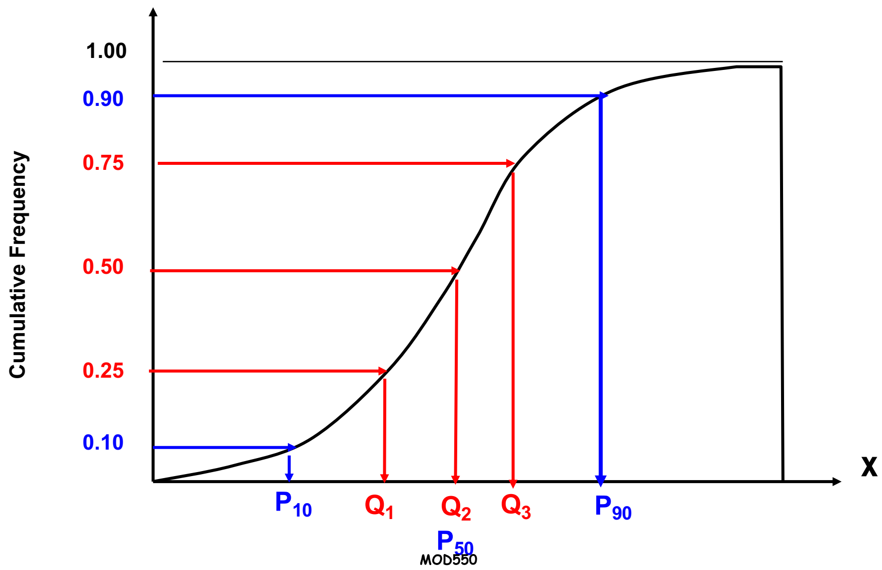
Mean, Median and Modes for Symmetric and Asymmetric Distributions



Measures of Location: Quantiles

- Quartiles — What is another name for the 2nd quartile?
  - in the same way that the median splits data into two halves, the quartiles split the data into quarters. If data values are arranged in increasing order, then a quarter of data fall below the first quartile, and a quarter of data falls above the third quartile.
- Deciles
  - splits the data into tenths; one tenth of the data falls below the first or lowest decile, two tenths fall below second decile. Fifth decile corresponds to the median.
- Percentiles
  - splits the data into hundredths; 25^{th} percentile is the same as the first quartile, and 50^{th} percentile is the same as the median, and 75th quantile is the same as 3rd quartile. Often referred to as P25, P50, etc. What we typically use.
- Quantiles
  - are a generalization of splitting data into any fraction.

Use of Cumulative Frequency to calculate Quantiles



Dispersion (Spread) of random variable

25

Sample Measures of Dispersion (Spread)

- Range

  $R = maximum - minimum$

- Inter-quartile Range

  $IQR = Q3 - Q1$

- Mean Deviation from the Mean?

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

- Mean Absolute Deviation

$$MAD = \sum_{i=1}^{n} |x_i - \bar{x}| / n$$

MOD550 26

26

Sample Measures of Dispersion: Variance & Standard Deviation

Variance is the average of squared differences between the sample data points and their mean

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

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

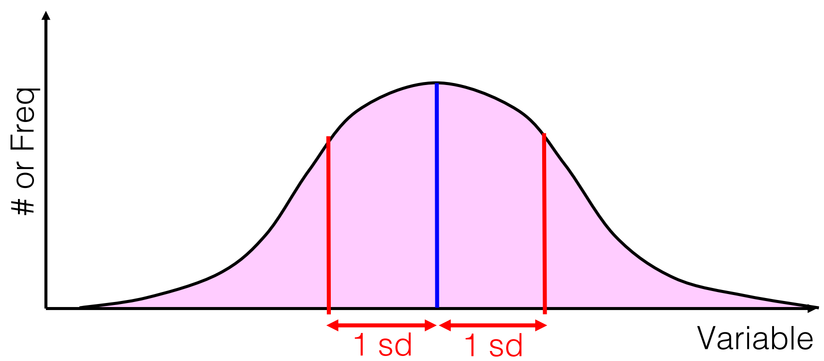
where $s_x{}^2$ = variance
s_x = standard deviation
n = sample size
x_i = i^{th} data sample

MOD550 27

27

Measures of Dispersion: Standard Deviation (SD)

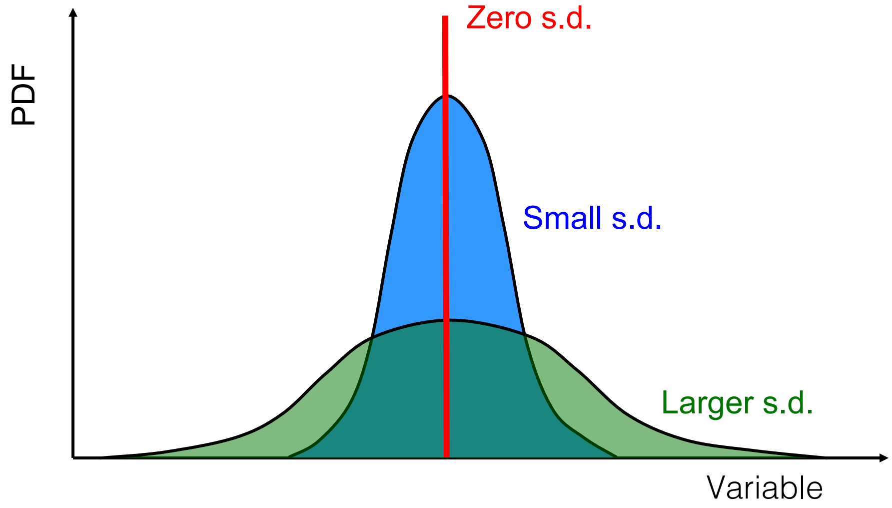
$sd = \sqrt{Var}$ ~ average squared difference from the mean



MOD550 28

28

Measures of Dispersion:
Variance (Var) and Standard Deviation (s.d.)



Population versus Sample Statistics

Key Differences Between Population and Sample

- Population: Includes all possible observations.
- Sample: A subset of the population.
- Sample statistics differ slightly due to estimation adjustments.

Mean (Expected Value)

- Population Mean (μ):

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

- Sample Mean ($\bar{X}$):

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Key Difference:
  - The sample mean is an estimator of the population mean.

31

Variance (Measure of Spread)

- Population Variance (σ^2):

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

- Sample Variance (s^2):

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

- Key Difference:
  - Sample variance uses $(n-1)$ for unbiased estimation.

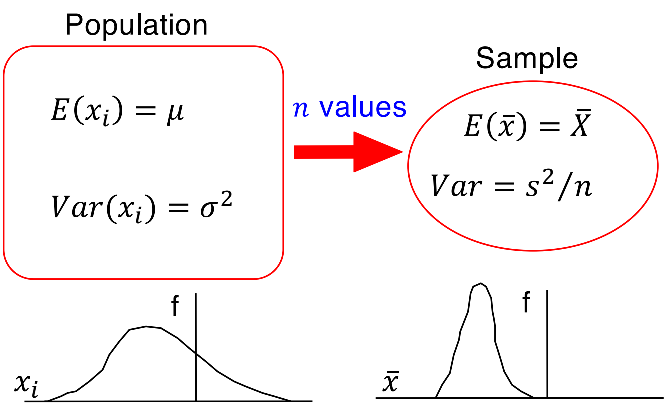
32

Comparison: Population vs. Sample

Measure	Population Formula	Sample Formula	Key Difference
Mean	$\mu = \frac{1}{N}\sum X_i$	$\bar{X} = \frac{1}{n}\sum X_i$	Sample mean estimates μ
Variance	$\sigma^2 = \frac{1}{N}\sum(X_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$	Sample variance divides by $n-1$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$	Sample standard deviation uses s^2

33

Variance of the sample will be smaller than the variance of the population



34

Why Use $(n-1)$ for Sample Variance?

- If we use n instead of (n-1), the sample variance underestimates population variance.
- Bessel's correction ensures an unbiased estimator.
- It adjusts for reduced variability in samples.

35

Key Takeaways

- Population includes all data, sample is a subset.
- Sample statistics estimate population parameters.
- Bessel's correction adjusts sample variance.
- Understanding these differences improves statistical accuracy.

36

Questions and Discussion

- Feel free to ask any questions!

37

Effect of Sample Size (random sample): Standard Error (SE) of the mean

$$SE_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

This is the *sd* of a series of different samples (one mean is computed from each sample)

It is an estimate of the uncertainty in the population mean

	Sample Standard Deviation		
	Porosity %	Perm. md	Shale Freq #/m
	5	100	0.05
Sample Size	**s.d. of Sample Mean**		
2	3.54	70.71	0.035
5	2.24	44.72	0.022
10	1.58	31.62	0.016
20	1.12	22.36	0.011
50	0.71	14.14	0.007
100	0.50	10.00	0.005
200	0.35	7.07	0.004
500	0.22	4.47	0.002
1000	0.16	3.16	0.002

MOD550 38

38

Measures of Dispersion: Coefficient of Variability

$$CV = \frac{s_X}{\bar{X}}$$

- A CV of greater than 1 can indicate the presence of extreme values
- Used as a measure of heterogeneity
- Graphical displays (PDF, CDF, Box-Plot) are more useful

CDF, PDF and PMF

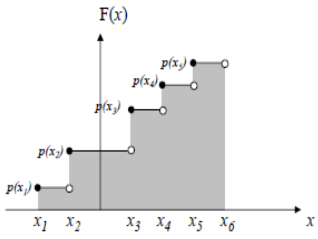
Discrete and Continuous CDF

Discrete CDF

For a discrete r.v. that attains values $x_1, x_2, \dots$ with probability $p_i = P(x_i)$, the CDF is discontinuous at x_i and constant in between.

$$F(x) = \mathrm{P}(\mathrm{X} \le x) = \sum_{x_i \le x} P(X = x_i) = \sum_{x_i \le x} p(x_i)$$

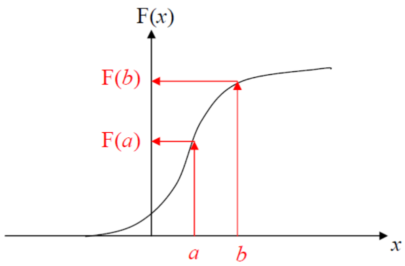
$$With \quad \sum_{i=1}^{N} p(x_i) = 1$$



Continuous CDF

$$\mathrm{F}(X \le a) = \int_{-\infty}^{a} f_X(x)dx$$

$$With \quad \mathrm{F}(X \le \infty) = \int_{-\infty}^{\infty} f_X(x)dx = 1$$



Empirical CDF (Generate CDF from Data)

1) Sort n data points in an ascending order such that $\mathrm{x}_1 \le \mathrm{x}_2 \le \mathrm{x}_3 \le \cdots \le \mathrm{x}_\mathrm{n}$.

2) Assign a rank (i) to each data point.

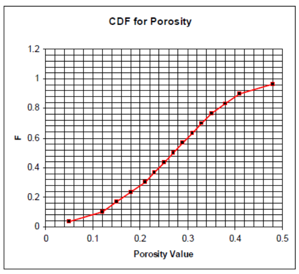
3) Assign a probability $\mathrm{F_i}$ to event $\mathrm{X} \le \mathrm{x_i}$ using:

$$F_i = P(X \le X_i) = \frac{\left(i - \frac{1}{2}\right)}{n} \quad or \quad \frac{i}{n+1}$$

4) Plot x_i versus F_i

Example

X_i	Xi_Sorted	Rank (i)	F_i = (i-1/2)/n
0.12	0.05	1	0.03
0.33	0.12	2	0.10
0.15	0.15	3	0.17
0.29	0.18	4	0.23
0.38	0.21	5	0.30
0.18	0.23	6	0.37
0.41	0.25	7	0.43
0.25	0.27	8	0.50
0.21	0.29	9	0.57
0.05	0.31	10	0.63
0.27	0.33	11	0.70
0.23	0.35	12	0.77
0.35	0.38	13	0.83
0.48	0.41	14	0.90
0.31	0.48	15	0.97

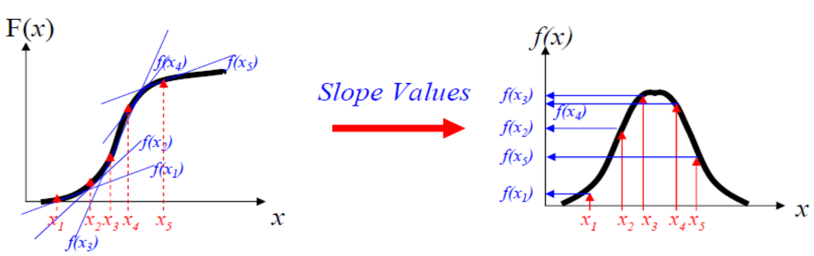


Continuous PDF and Discrete PMF

Continuous PDF

The PDF $f(x)$ is a non-negative function that characterizes the relative probability (frequency of occurrence) of realization values for a rv at a neighbourhood of a point:

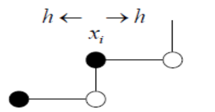
$$P(x \le X \le x+\Delta x) = F(x+\Delta x) - F(x) = \int_x^{x+\Delta x} f_X(x)dx \longrightarrow \lim_{\Delta x \to 0} \frac{F(x+\Delta x)-F(x)}{\Delta x} = \frac{dF(x)}{dx} = f_X(x)$$



Note: $f(x)$ does not represent probability values. $f(x)$ represents density values.

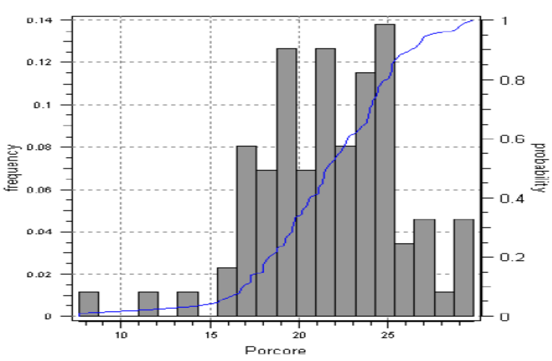
Discrete PMF

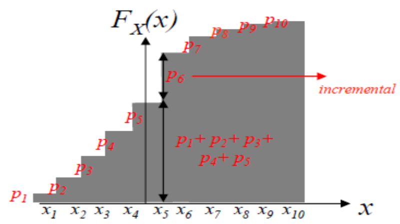
Discrete PMF

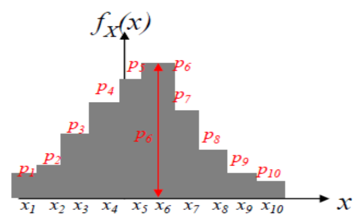


$$P(x_i - h \le X \le x_i + h) = F(x_i + h) - F(x_i - h) = p_i$$

$$h \to 0 \qquad\qquad h \to 0$$







Empirical PMF (Generate PMF from Data)

1) Sort 'n' data points in an ascending order such that $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n$.

2) Divide data range in to reasonable number of bins. (n_{bins})

3) Count the number of data in each bin (category) i to find f_i and compute the probability of each bin, $P_i = \frac{f_i}{n}$

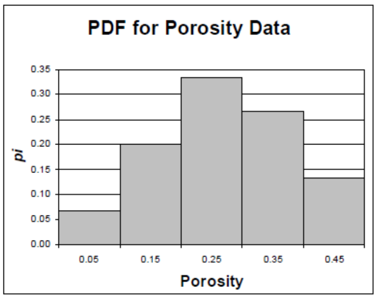
4) Plot bin mid-range $(x_i < X < x_i + \Delta x)$ versus p_i

x_i	x_i Sorted
0.12	0.05
0.33	0.12
0.15	0.15
0.29	0.18
0.38	0.21
0.18	0.23
0.41	0.25
0.25	0.27
0.21	0.29
0.05	0.31
0.27	0.33
0.23	0.35
0.35	0.38
0.48	0.41
0.31	0.48

→

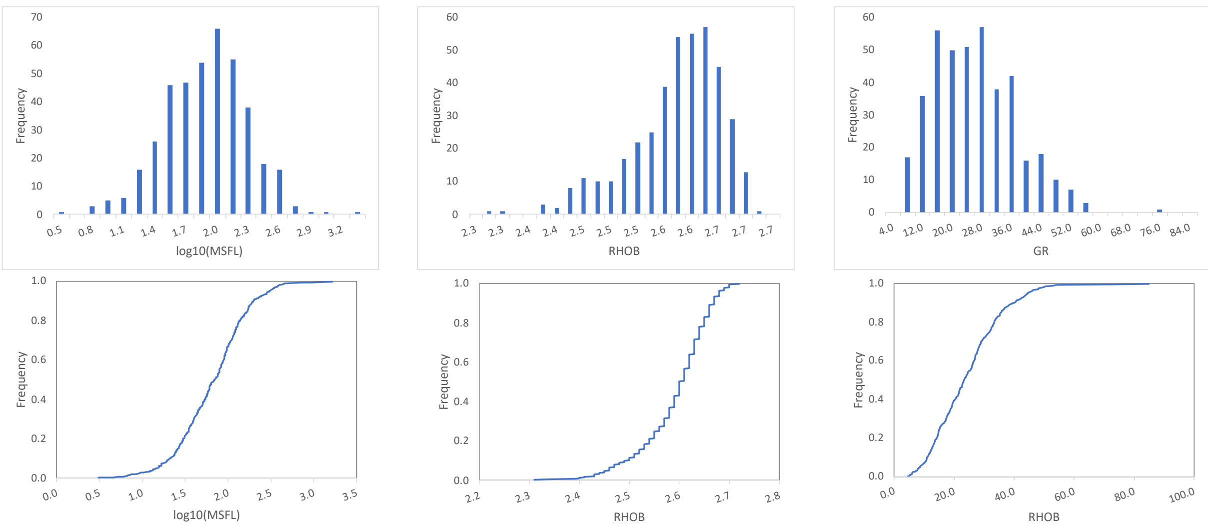
Bin Range	Bin Midsize	f_i	p_i
0.0-0.1	0.05	1	0.07
0.1-0.2	0.15	3	0.20
0.2-0.3	0.25	5	0.33
0.3-0.4	0.35	4	0.27
0.4-0.5	0.45	2	0.13

→



Distributions – PMFs and CDFs

Symmetric | Right-Skewed | Left-Skewed

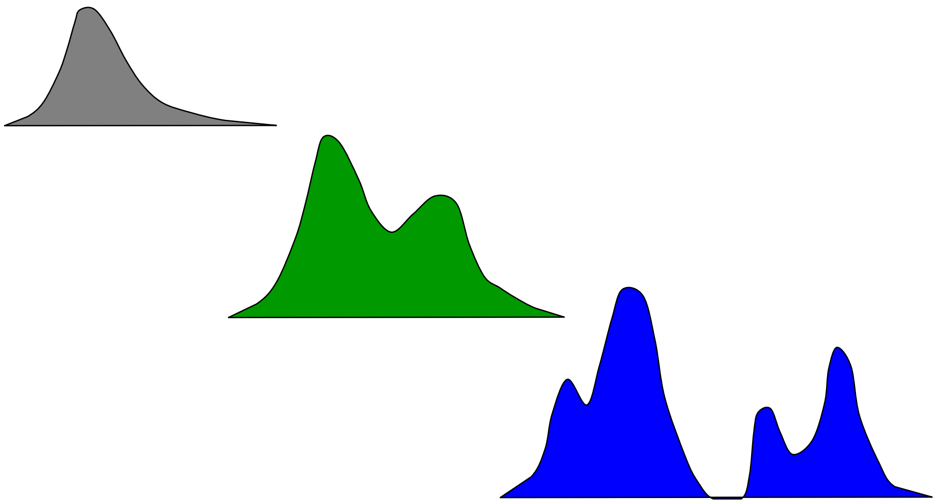


Measured of Shape and Box Plots

48

Measures of Shape: Modality

Number of Modes: unimodal, bimodal, polymodal



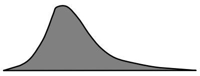
49

Measures of Shape: Skewness

Measure of symmetry in the distribution of the data values

$$Sk = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{s^3}$$

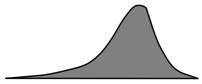
Positive - Values clustered toward the lower end; tail extends to the right



$Sk > 0$

Zero – Symmetric distribution

Negative - Values clustered toward the higher end; tail extends to the left



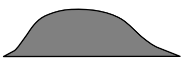
$Sk < 0$

Measures of Shape: Kurtosis

Measures the “flatness” or “peakedness” of the distribution

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4}{s^4} - 3$$

Negative – flatter, more extreme values



Positive – more peaked, fewer extreme values

Normal distribution has zero kurtosis

Example: Sample of 100 Porosities

30.3	29.7	16.9	9.2	21.1	23.5	17.8	26.3	28.3	30.9
39.8	27.4	19.1	20.9	5.1	35.6	22.8	34.2	17.9	23.4
37.5	29.4	29.3	25.5	16.2	19.5	28.2	28.1	26.8	38.1
14.7	21.4	31.7	24.3	26.5	34.9	14.3	5.7	22.2	37.0
23.7	26.0	29.6	28.4	11.5	17.8	22.1	23.0	7.6	13.3
25.0	29.9	26.1	15.1	10.8	26.3	26.0	18.4	20.7	22.4
33.8	29.2	31.9	34.6	11.3	24.4	9.5	4.1	15.8	27.2
12.0	24.0	39.1	12.9	42.1	35.1	11.7	14.7	43.6	12.2
20.5	26.9	20.1	29.5	31.5	32.5	16.5	17.3	21.2	13.0
7.8	9.1	25.9	8.0	2.5	21.9	11.1	28.3	12.4	18.3

Example: Statistics of Sample of 100 porosity data

Stats	
n	100
Min (=P0)	2.53
P5	7.78
Q1	15.59
Mode	#N/A
Median (=Q2)	23.17
Mean	22.59
Q3	29.19
P95	37.52
Max (=P100)	43.63

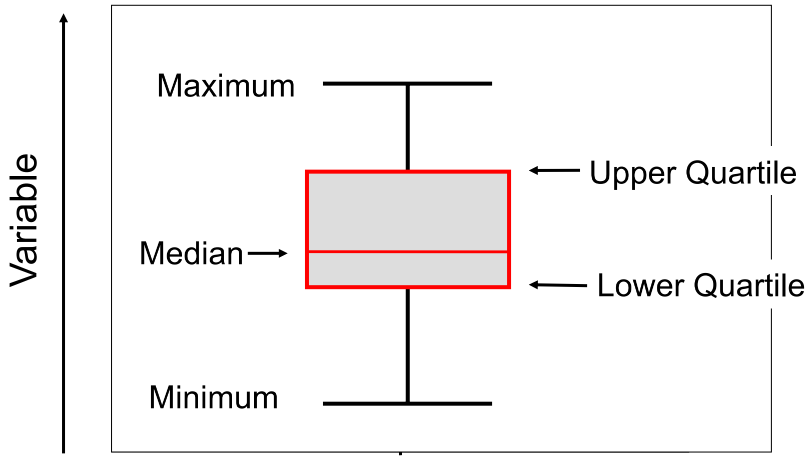
Why?

Example: Statistics of Sample of 100 porosity data - Python

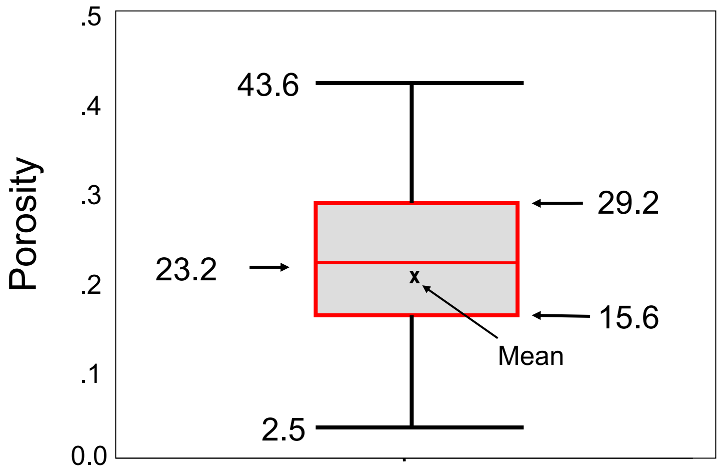
```
df.describe()
```

	Porosity
count	100.000000
mean	22.587709
std	9.240477
min	2.534500
25%	15.591430
50%	23.173575
75%	29.190603
max	43.634960

Visual Summary of Sample Statistics: Box Plots - Simple Version



Example: Visual Summary of Sample Statistics. Box Plots – 100 Porosity Data



MOD550 56

56

Summary - Univariate Data Descriptors

- Mean => expected value (central tendency)
  - $E(x) = \sum p_i x_i = (1/N) \sum x_i$
  - p_i *is relative frequency*
- Variance => spread around mean
  - $V(x) = \sum p_i [x_i - E(x)]^2$
    $= (1 / N) \sum [x_i - E(x)]^2$
- Standard Deviation => square root of variance
  - $\sigma(x) = \sqrt{\sum p_i [x_i - E(x)]^2}$
    $= \sqrt{(1 / N) \sum [x_i - E(x)]^2}$
- Coefficient of variation => normalized spread
  - $CV(x) = \sigma(x)/E(x)$
    (also expressed as percent)
- Median => mid-point of distribution
- Mode => most likely (frequently occurring) value
- Skewness => degree of asymmetry in PDF
- Kurtosis => degree of peakedness in PDF

MOD550 57

57

Bivariate Statistics

58

Renewable Energy Variables that are Related

- Solar Irradiance and Panel Efficiency:
  - Dictates how much solar power can be harnessed at a location
- Wind Speed and Turbine Efficiency:
  - Indicate the potential for wind energy generation.
- Panel or Turbine Density and Material Composition:
  - Reflects how closely packed energy generation units are and what materials they are made from.
- Energy Storage Density and Charge-Discharge Efficiency:
  - Affect how energy is stored and how quickly it can be deployed.
- Renewable Plant Area and Energy Yield:
  - Shows how the physical size of a renewable energy plant impacts its total energy output.
- Biomass Crop Yield and Harvesting Cycle Time:

59

Common Oil & Gas Variables that are Related

- Porosity and permeability
- Porosity and water saturation
- Density and molecular weight
- Oil density and viscosity
- Formation thickness and productivity
- Sand-body width and thickness

Scatterplot – Read and Display Data for CO_2 Injection

In [7]:
```
## Read the dataset

file_name = '../data/regression.csv'
df = pd.read_csv(file_name)
df.head()
```

Out[7]:

	Well	Net_Pay	Well_Injection_Potential
0	A-1	65	2250
1	A-2	45	2450
2	A-3	28	2000
3	A-4	47	1820
4	A-5	12	680

Scatterplot – Read and Display Data for CO2 Injection

Out[81]:

	Well	Net_Pay	Well_Potential
0	A-1	65	225
1	A-2	45	245
2	A-3	28	200
3	A-4	47	182
4	A-5	12	68

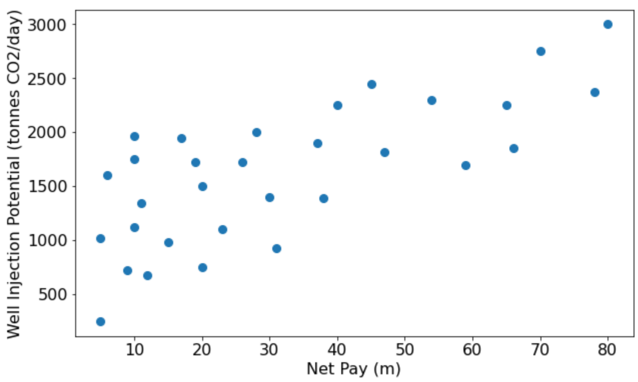
In [87]:
```
fig, ax = plt.subplots(figsize=(8,6))
ax.scatter(x=df['Net_Pay'], y=df['Well_Potential'], marker='o', c='r', edgecolor='b')
ax.set_xlabel('Net Pay (ft)')
ax.set_ylabel('Initial Well Potential (BOPD)')
```


Graphing Bivariate Data

A scatterplot between two variables is the simplest way of graphically displaying their relationship.

```
fig, ax = plt.subplots(figsize=(10,6))
ax.scatter(x=df['Net_Pay'], y=df['Well_Injection_Potential'], marker='o')
ax.set_xlabel('Net Pay (m)')
ax.set_ylabel('Well Injection Potential (tonnes CO2/day)')
```

```
Text(0, 0.5, 'Well Injection Potential (tonnes CO2/day)')
```



Covariance - Describing the Relationship between two Variables

- The covariance or joint variance between two random variables is an extension of the concept of variance and is defined as

$$Cov[XY] = \sigma_{xy} = E[(X-\bar{X})(Y-\bar{Y})] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y})$$

$$= \frac{N}{N-1}\{E[XY] - E[X]E[Y]\}$$

- Generalization of variance.
- Consider the covariance of a variable with itself

$$Cov[XX] = \sigma_{xx} = E[(X-\bar{X})(X-\bar{X})] = Var[X]$$

- Variance: positive
- Covariance: positive or negative

64

Correlation Analysis

- The correlation between two random variables is a measure of the strength of their linear relationship.
- Parametric Correlation:
  - Measures a linear (Pearson) dependence between two variables (x and y) is known as a parametric correlation test because it depends on the distribution of the data.
- Non-Parametric Correlation:
  - Kendall (tau) and Spearman (rho), which are rank-based correlation coefficients, are known as non-parametric correlation.
- There are several NumPy, SciPy, and Pandas correlation functions and methods that you can use to calculate these coefficients.
- Use Matplotlib to conveniently illustrate the results.

65

Correlation (Pearson's r Value)

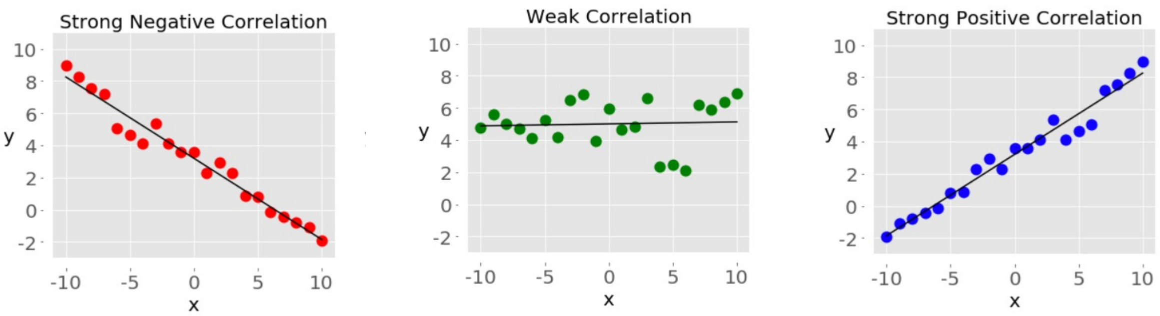
- The correlation coefficient (r) between two random variables is a measure of the strength of their linear relationship.
- It is closely linked to the concept of covariance and is defined as

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i - \bar{X}}{\sigma_x}\right)\left(\frac{(y_i - Y)}{\sigma_y}\right)$$

- **Type equation here.** r ranges between -1 (indicating perfectly negative correlation) and $+1$ (indicating perfectly positive correlation).
- The sign indicates the direction of the trend (i.e., positive or negative), and the absolute value quantifies the strength of the relationship.
- The concept of correlation strictly applies for a monotonic relationship.

66

Correlation: Examples



Pearson's r Value	Correlation Between x and y
$= 1$	perfect positive linear relationship
> 0	positive correlation
$= 0$	independent
< 0	negative correlation
$= -1$	Perfect negative linear relationship

67

Example: NumPy Pearson Correlation Calculation

Pearson's correlation coefficient = r

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

```
from numpy import cov, var
x = df['Net_Pay']
y = df['Well_Injection_Potential']
sigma_xy = cov(x, y)[0,1]
sigma_x = np.sqrt(var(x, ddof=1))
sigma_y = np.sqrt(var(y, ddof=1))

print(f'The covariance between Net Pay and Well Injection Potential is {sigma_xy:.1f}')
print('The standard deviations are:')
print(f'Net Pay = {sigma_x:.2f}')
print(f'Well Injection Potential = {sigma_y:.2f}')
```

```
The covariance between Net Pay and Well Injection Potential is 10935.0
The standard deviations are:
Net Pay = 23.03
Well Injection Potential = 647.15
```

Now calculate the Pearson correlation coefficient r

```
sigma_xy / (sigma_x * sigma_y)
f'The Pearson correlation coefficient, \u03C1, value between Net Pay and Well Injection Potential is = {rho:.2f}'
```

The Pearson correlation coefficient, r, value between Net Pay and Well Injection Potential is = 0.73

68

NumPy Pearson Correlation Calculation

Why `sigma_xy = cov(x, y)[0, 1]`?

The `[0, 1]` in the line `sigma_xy = cov(x, y)[0, 1]` is used to access a specific element from the covariance matrix returned by the `cov` function from NumPy.

When you calculate the covariance between two sets of values (in your case, `x` and `y`), the `cov` function returns a covariance matrix. This matrix is a 2x2 matrix when dealing with two variables, structured as follows:

$cov(x, x), cov(x, y)$

$cov(y, x), cov(y, y)$

- `cov(x, x)` is the variance of `x`.
- `cov(y, y)` is the variance of `y`.
- `cov(x, y)` and `cov(y, x)` are the same and represent the covariance between `x` and `y`.

The `[0, 1]` is used to access the element in the first row and second column of this matrix, which is `cov(x, y)`, the covariance between `x` and `y`. Similarly, `[1, 0]` would also give you the same value since the covariance matrix is symmetric.

So, in summary, you use `[0, 1]` to extract the actual covariance value between `x` and `y` from the covariance matrix.

69

NumPy Pearson Correlation Calculation

Why `sigma_x = np.sqrt(var(x, ddof=1))` ?

The `ddof=1` parameter in the variance function (`var`) is used to specify the "Delta Degrees of Freedom." This parameter determines how the variance is normalized.

1. **Default Behavior (`ddof=0`)**: By default, when `ddof` is set to 0, the variance is normalized by `N` , where `N` is the number of observations. This is known as the population variance. It assumes that the data set represents the entire population and, therefore, divides by the total number of observations.
2. **Sample Variance (`ddof=1`)**: When you set `ddof=1` , the variance is normalized by `N-1` , where `N` is the number of observations. This is known as the sample variance. It's used when the data set is a sample of the entire population, not the entire population itself. Dividing by `N-1` corrects the bias in the estimation of the population variance from a sample.

In many practical situations, especially in statistics and data science, you often work with a sample of data rather than the entire population. Using `ddof=1` provides an unbiased estimator of the population variance based on the sample. This adjustment is particularly important in smaller data samples, where the difference between `N` and `N-1` can have a more pronounced effect on the variance calculation.

In summary, `ddof=1` is used to calculate the sample variance, which is a more accurate estimate of the true population variance when working with sample data.

70

Correlation Coefficient - Interpretation

- What do I do with this number?
- 0.734 indicates that the two variables are partially correlated
- A correlation coefficient of 0.734 means that 53.85% $(= 0.734^2)$ of the variation in the y-variable is explained by the variability in the x-variable.
- r is a measure of how close the points come to falling on a straight line

Correlation	Negative	Positive
Small	−0.29 to −0.10	0.10 to 0.29
Medium	−0.49 to −0.30	0.30 to 0.49
Large	−1.00 to −0.50	0.50 to 1.00

71

71

Correlation Coefficient - Interpretation

Since r is a measure of how close the points come to falling on a straight line it is an indicator of how successful we might be in predicting one variable from another (see later – regression)

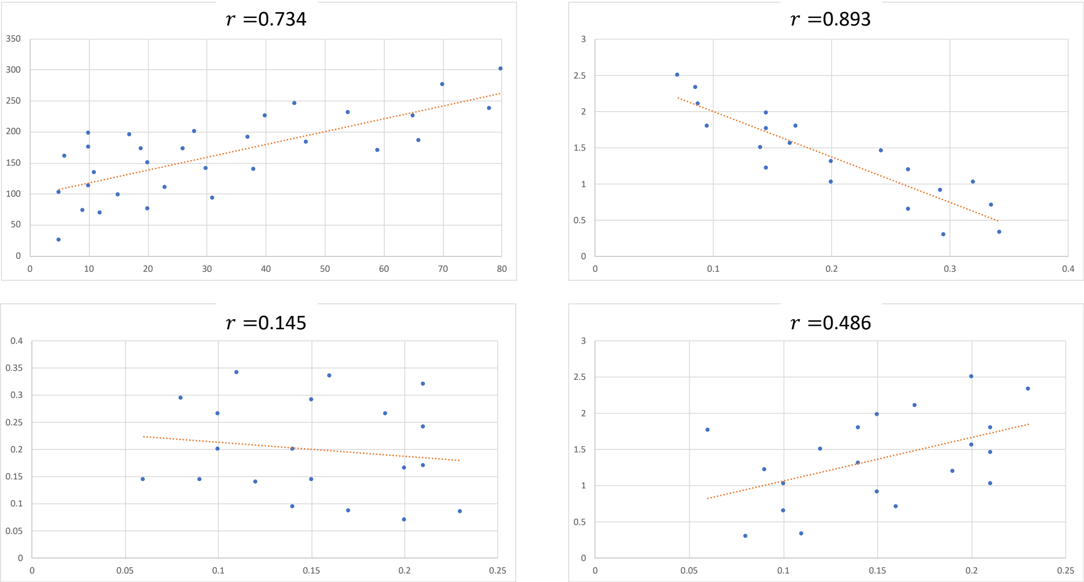
– If r is high, then for a given value of one variable, then we know that the other variable is restricted to only a small range of values

– If r is low, then knowing the value of one variable does not give us much information on the other

72

72

The Value of r is inversely proportional to the degree of scatter around the underlying liner trend

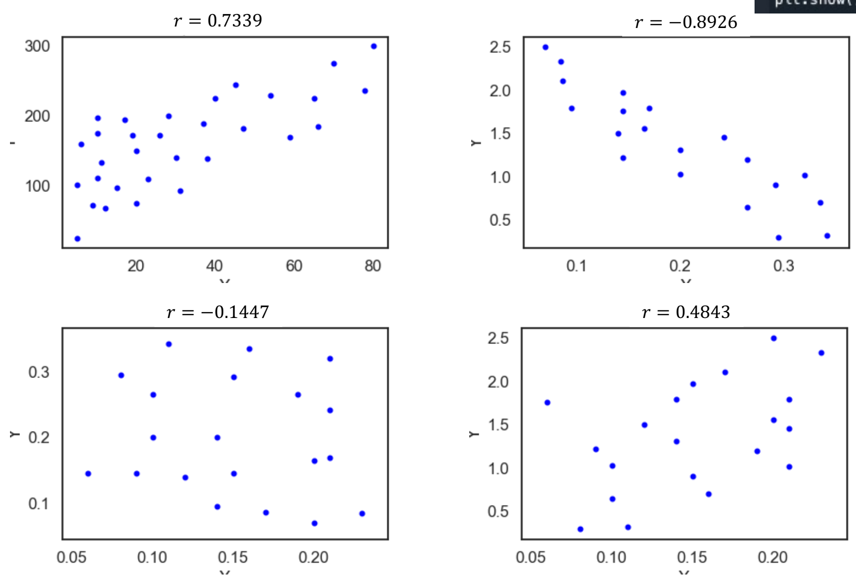


73

73

Python:

```
sP = dfI['X'].corr(dfI['Y'])
sS = dfI['X'].corr(dfI['Y'], method="pearson")
dfI.plot(kind='scatter',x='X',y='Y',color='blue')
plt.title(r'$\rho$ = {:.4f}'.format(sP))
plt.savefig('ScatterI.png')
plt.show()
```



Note that these are the Pearson r values and not the Spearman ρ values.

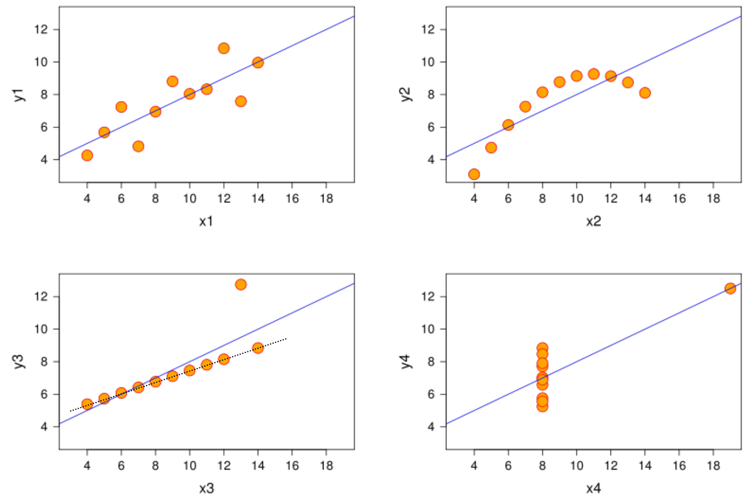
74

74

Pearson Correlation – Limited to Linear Correlations

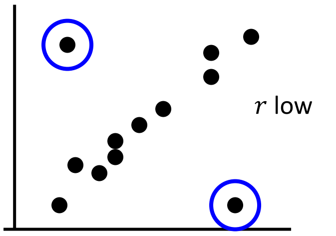
Anscombe's Quartet - Four different pairs of variables

– the same mean (7.5), standard deviation (4.12), correlation (0.81) and regression line ($y = 3 + 0.5x$).

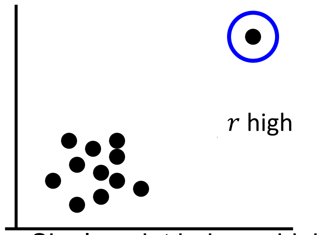


75

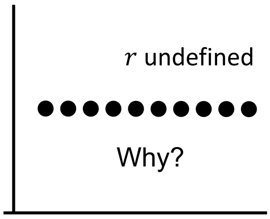
Correlation: Problem areas – always look at the scatter plot

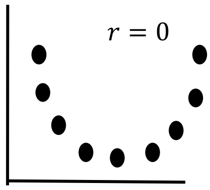


Outliers significantly reduce correlation coefficient



Single point induces high correlation coefficient



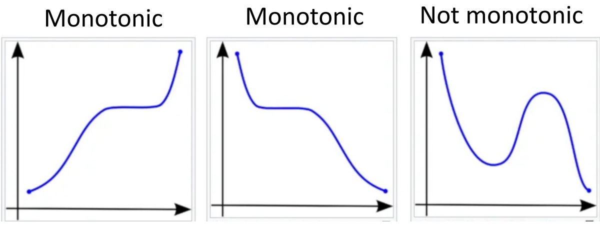


No linear correlation, but clearly NOT independent

76

Spearman Rank Correlation

- Wikipedia:
  - In statistics, Spearman's rank correlation coefficient or Spearman's ρ, named after Charles Spearman is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.
- The Spearman correlation evaluates a monotonic relationship between two variables — Continuous or Ordinal and it is based on the ranked values for each variable rather than the raw data.



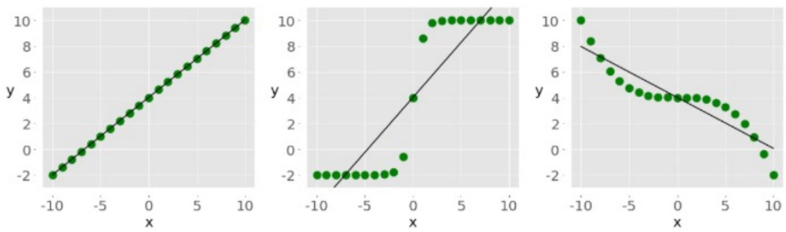
77

Spearman Rank Correlation

- Rank correlation compares the ranks (orderings) of the data related to two variables.
- If the orderings are similar, then the correlation is strong, positive, and high.
- However, if the orderings are close to reversed, then the correlation is strong, negative, and low.
- In other words, rank correlation is concerned only with the order of values, not with the particular values from the dataset.

Pearson Linear (r) versus Spearman Rank ρ Correlation

To illustrate the difference between linear and rank correlation, consider the following figure:



- Left plot $r = 1$
- Central plot $r > 0$
- Right plot $r < 0$
- When you look only at the orderings or ranks, all three relationships are perfect, i.e., $\rho = 1$ or -1 !

Spearman Rank Correlation

- Calculated the same way as the Pearson correlation coefficient but using ranks instead of values.
  - Denoted with the Greek letter rho (ρ) and called Spearman's rho.
- Facts about the Spearman correlation coefficient:
  - It can take a real value in the range $-1 \le \rho \le 1$
  - Max value $\rho = 1$ corresponds to the case when there's a monotonically increasing function between x and y.
    - => larger x values correspond to larger y values and vice versa.
  - Min value $\rho = -1$ corresponds to the case when there's a monotonically decreasing function between x and y.
  - Calculate Spearman's ρ in Python in a very similar way as for Pearson's r

$$\rho = \frac{\sigma_{xy(rank)}}{\sigma_{x(rank)}\sigma_{y(rank)}} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{R_{x,i} - \bar{R}_x}{\sigma_{R_x}}\right)\left(\frac{R_{y,i} - \bar{R}_y}{\sigma_{R_y}}\right)$$

80

Example: SciPy Spearman Correlation Calculation

Spearman's ρ

$$\rho = \frac{\sigma_{rank_{xy}}}{\sigma_{rank_x}\sigma_{rank_y}}$$

```
sigma_xy_rank = cov(x_rank, y_rank)[0,1]
sigma_x_rank = np.sqrt(var(x_rank, ddof=1))
sigma_y_rank = np.sqrt(var(y_rank, ddof=1))

print(f'The covariance between Net Pay and Well Injection Potential is {sigma_xy_rank:.1f}')
print('The standard deviations are:')
print(f'Net Pay = {sigma_x_rank:.2f}')
print(f'Well Injection Potential = {sigma_y_rank:.2f}')
```

```
The covariance between Net Pay and Well Injection Potential is 56.0
The standard deviations are:
Net Pay = 9.09
Well Injection Potential = 9.09
```

```
rho_spearman = sigma_xy_rank / (sigma_x_rank * sigma_y_rank)
print(f'The Spearman correlation coefficient, \u03C1, value between Net Pay and Well Injection Potential is = {rho_s
```

```
The Spearman correlation coefficient, ρ, value between Net Pay and Well Injection Potential is = 0.678
```

81

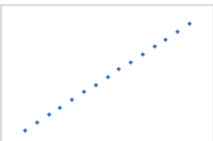
Example: SciPy Spearman Correlation Calculation

Can also use scipy

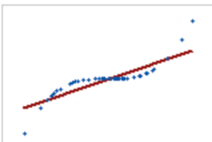
```
spearman_correlation_coefficient = scipy.stats.spearmanr(x, y)

# Extract the correlation coefficient
spearman_rho_scipy = spearman_correlation_coefficient[0]

print(f'The Scipy calculated Spearman correlation coefficient, ρ, = {spearman_rho_scipy:.3f}')
```

```
The Scipy calculated Spearman correlation coefficient, ρ, = 0.678
```


Comparison of Pearson and Spearman Coefficients

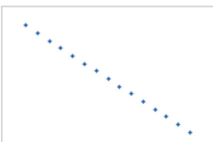


Pearson = +1, Spearman = +1

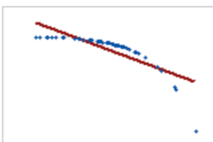


Pearson = +0.851, Spearman = +1

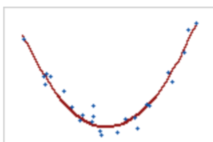
Pearson = −0.093, Spearman = −0.093



Pearson = −1, Spearman = −1



Pearson = −0.799, Spearman = −1



Pearson = 0, Spearman ~ 0.9 (or -0.9)

83

Kendall tau (τ) Correlation

- Kendall's tau and Spearman's rank correlations assess statistical associations based on the ranks of the data.
- Kendall tau correlation (non-parametric) is an alternative to Pearson's correlation (parametric) when the data has failed one or more assumptions of the test.
- Kendall tau is also the best alternative to Spearman correlation (non-parametric) when the sample size is small and has many tied ranks.
- Kendall rank correlation is used to test the similarities in the ordering of data when it is ranked by quantities.
  - Other types of correlation coefficients use the observations as the basis of the correlation,
  - Kendall's correlation coefficient uses pairs of observations and determines the strength of association based on the patter on concordance and discordance between the pairs.

84

Correlation Does NOT Indicate Causation

- Note that correlation does not indicate causation.
- It quantifies the strength of the relationship between the features of a dataset.
- Sometimes, the association is caused by a factor common to several features of interest.

85

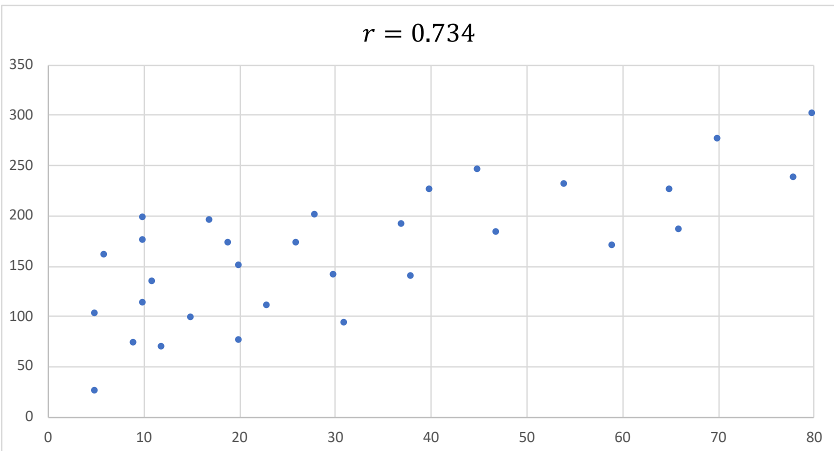
Graphing Bivariate Data

- A scatterplot between two variables is the simplest way of graphically displaying their relationship.
- The strength of linear association, if any, is given by the absolute value of the Pearson r value
- The sign of r indicates whether the correlation is positive or negative.

86

86

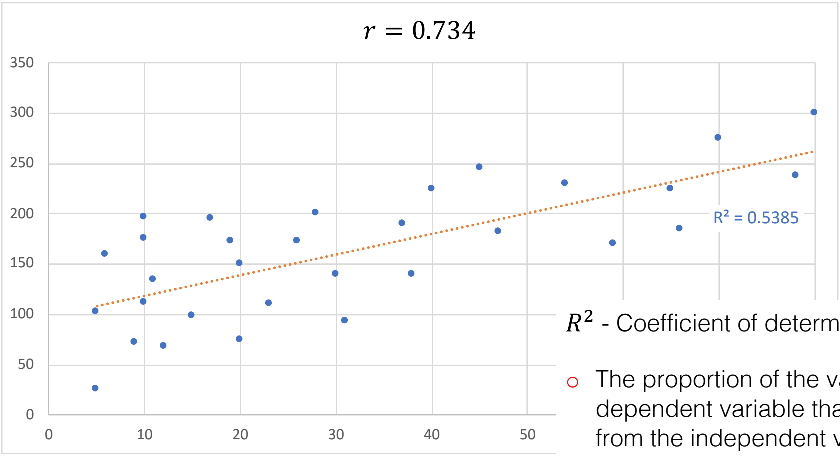
Scatterplot



87

87

Scatterplot



R^2 - Coefficient of determination

- The proportion of the variance in the dependent variable that is predictable from the independent variable(s)

$$\sqrt{R^2} = \sqrt{0.5385} = 0.734 = r$$

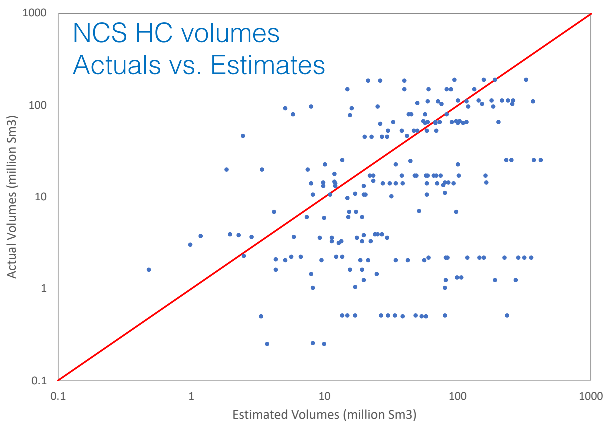
88

88

Scatterplots

- Bivariate display, typically
  - two covariates (e.g. porosity and permeability) at same location
  - the same variable at different locations - separated by some distance vector
  - estimated value versus true value
- Good for spotting aberrant data

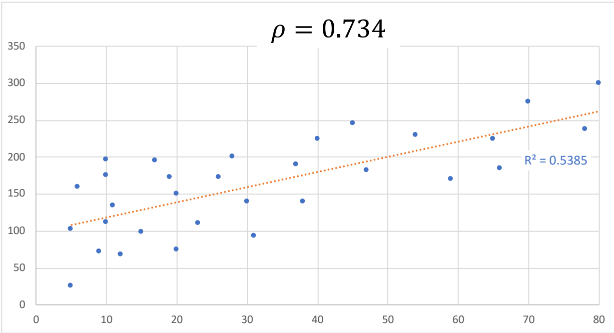
Number of data points		202
X Variable: mean		72.52
	variance	11308.07
Y Variable: mean		36.05
	variance	2243.62
Correlation		0.19



89

89

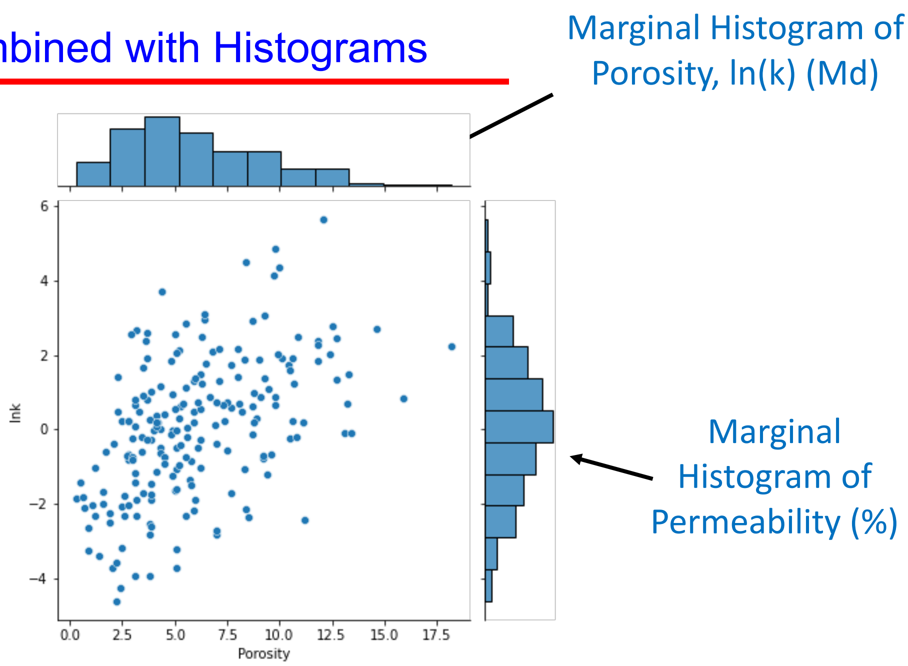
ρ is NOT Equal to the Slope of the Regression Line



90

90

Scatterplots Combined with Histograms



91

91

Bivariate Gaussian PDF

Bivariate normal densities

$$f_{X_1X_2}(x_1,x_2)=\frac{\exp\left\{-\frac{1}{2(1-\rho_{XY}^2)}\left[\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2-2\rho_{XY}\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)+\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2\right]\right\}}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho_{XY}^2}}=$$

Covariance Matrix

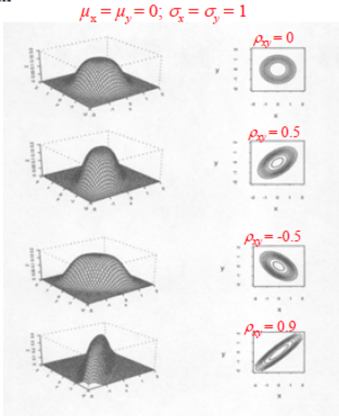
$$\mathbf{C}_{XX}=\begin{bmatrix} Cov(X_1,X_1) & Cov(X_1,X_2) \\ Cov(X_2,X_1) & Cov(X_2,X_2) \end{bmatrix}=\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}=$$

For the above example

$$\mathbf{C}_{XX}=\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}=\begin{bmatrix} \sigma_{X_1}^2 & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2 \end{bmatrix}$$

Note that mean and covariance fully define a Gaussian PDF, which makes Gaussians mathematically desirable PDFs!

$\mu_x=\mu_y=0;\ \sigma_x=\sigma_y=1$



92

Correlation versus Dependence

- Uncorrelated Random Variables:
  - Random variables are uncorrelated if there is no linear relationship between them
  - Mathematically, two random variables X and $Y$are uncorrelated if their covariance is zero. That is, $Cov(X,Y)\ =\ 0$
  - Uncorrelation refers to the absence of a linear relationship
- Independent Random Variables:
  - Random variables are independent if the occurrence of one variable does not affect the probability distribution of the other
  - Mathematically, X and $Y$are independent if, $P(X=x\ and\ Y=y)=P(X=x)\cdot P(Y=y)$, for all x and y
  - If two variables are independent, they must be uncorrelated
  - However, the opposite is not true: uncorrelated random variables are not necessarily independent. Two variables can be uncorrelated but still dependent through some non-linear relationship

93

Causation Implies Dependency

- Statistical Dependency:
  - Variables are causally related must be statistically dependent
  - Knowing the value of one variable gives some information about the other
  - For instance, if smoking causes lung cancer, then knowing whether a person smokes changes the probability of them having lung cancer
- Causal Dependency
  - Causation is a specific type of dependency where one variable (the cause) directly affects another variable (the effect)
  - This goes beyond mere association or correlation and implies a direct or indirect mechanism through which the cause influences the effect

While causation implies dependency, dependency does not necessarily imply causation

94

94

Dependency Does not Imply Causation

- Correlation without Causation:
  - Two variables might be correlated (and hence dependent) due to a coincidence, a lurking variable, or a confounding factor
  - This is the classic scenario where correlation does not imply causation
- Common Cause:
  - Two variables might be dependent because they are both influenced by a third variable
  - This does not mean that one of the two variables causes the other, but rather that they have a common cause
  - Example:
    - In most countries, as ice cream sales increase, the number of drowning incidents also increases
    - It would be wrong to infer from this that ice cream eating leads to an increased risk of drowning

95

95

Correlation is not Causation

96

96