INDEX[KPC1]

INTRODUCTION

This course explores the fascinating world of data analysis and interpretation. Statistics is the science of collecting, organizing, analyzing [KPC2]and drawing conclusions from data. Whether we are looking to understand trends, make decisions based on data, or gain insights into complex phenomena, statistics provide [KPC3]the tools we need to make sense of the world around us.

On this course you will learn key statistical concepts such as probability, distributions, time series and their applications. By the end of the course, you will be able to analyze datasets confidently, understand variability and apply statistical reasoning to solve real-world problems.

Whatever your field of study or career aspirations, a solid foundation in statistics is essential in today's data-driven world. We look forward to guiding you through this journey as you develop the skills you need to become a competent and critical user of data.

UNIT 1: Descriptive Statistics Basics

In this first unit we delve into the field of Descriptive Statistics, which is the foundation for understanding and summarizing data. Descriptive statistics involves methods for collecting, organizing and presenting data in a meaningful way in order to provide insights into patterns and trends.

A key focus of this course [KPC4]will be the graphical representation of data. You will learn how to use visual tools such as histograms, bar charts, pie charts, box plots and scatterplots to communicate data effectively. These graphs and charts help transform raw data into intuitive visualizations, making it easier to identify relationships, spot outliers and understand the overall distribution of data.

By the end of this unit you will be equipped with the skills you need to summarize data using numerical measures (such as mean, median and standard deviation) and present your findings visually in ways that are clear and informative. These skills are crucial for decision-making in fields such as business, science, health and social research.

First of all, let N and n be different variables to take into account from the start. It is extremely important to define our first concepts to provide us with the vocabulary we need to begin our course:

Population (N) is the collection of all items of interest.
Sample (n) is the observed subset of the population.
A parameter is a numerical measure that describes a specific characteristic of the population.
A statistic is a numerical measure that describes a specific characteristic of the sample.

Sample representative of the population:

- Probabilistic sampling (random, systematic, etc.)
- Non-probabilistic sampling (convenience, opinion)

Random sampling involves taking into account a random variable.

Descriptive statistics is based on collecting data (surveys), presenting data (tables/graphs) and summarizing data (sample means).

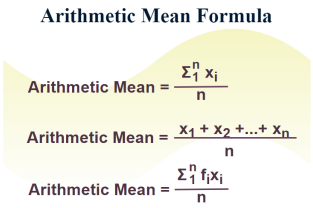
Inference: drawing conclusions from a population based on sample results.

Variable: a characteristic or feature of an element or item of interest.

Variables are data that may be categorical (nominal, ordinal) or numerical (discrete, continuous).

Data are used to make predictions, forecasts and estimates to help in decision making.

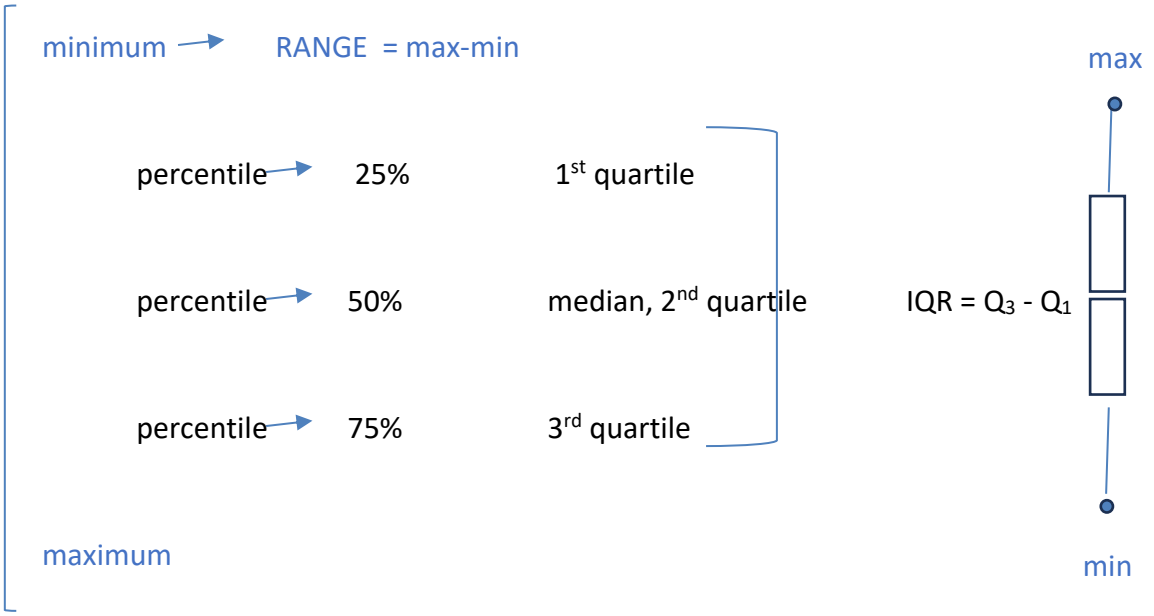
The arithmetic mean is an arithmetic average. It is the sum of values divided by the number of values. The arithmetic mean is the most common measure of central tendency.



The median is the "middle number" [KPC5]in an ordered list (50% above, 50% below).

The mode is the value that is repeated the most.

Percentiles and quartiles indicate the position of a value relative to the entire set of data.
Quartiles divide the distribution into 4 parts.
Deciles divide the distribution into 10 parts.
Percentiles divide the distribution into 100 parts.



The range is the difference between the highest value and the lowest value.

The interquartile range is the range comprising the middle 50% of the data.

The Z-score measures relative variation for specific [KPC6]values.

Variance is the average of the squared deviations of values from the mean.

The standard deviation is the square root of the variance:

Stand. Dev. = σ = s = $\sqrt{Var}$

The Gini index measures the degree of equality or inequality of a distribution between p (cumulative relative frequency) and q (cumulative relative amount x[KPC7] frequency).

GI=0 indicates the lowest [KPC8]concentration.
GI=1 indicates the highest concentration and the lowest [KPC9]uniformity.

$$G = \frac{\sum (p_i - q_i)}{\sum p_i}$$

EXERCISES/PROBLEMS

Problem 1. (1.5 points) A researcher is inserted in studying the effect on gas consumption after the prices rise. Data about gas consumption (in m^3) in a working class neighbourhood in Valencia was collected in the following table:

Gas consumption (m^3)	50	100	150	200	300	400
Number of homes	10	15	28	42	7	1

In addition, it is known that in a high class neighbourhood, the average gas consumption is 190 m^3 with a standard deviation of 77.35 m^3.

a) In which neighbourhood is the mean more representative (that is, has lower relative dispersion)?
b) The gas consumption in a particular home from the working class neighbourhood is 150 m^3 and the consumption in a specific home from the high class neighbourhood is 165 m^3. Which home had a higher consumption in relative terms?
c) A concentration measure about the gas consumption in the working class neighbourhood is 0.1584. In the high class neighbourhood, the highest consumers' homes are the 30% of homes and consume the 40% of the total consumption, while the less consumers' homes are the 40% of homes and consume 35% of the total gas consumption. In which neighbourhood the gas consumption is less concentrated?

Problem 1: Gas consumption

Working class

a = ?

$\bar{x} = \frac{(50*10 + 100*15\ldots.)}{10 + 15 + 28 + 42 \ldots}$

$= 166.01\ m^3$

High class

$a = 77.35\ m^3$

$x = 190\ m^3$

Gas consumption	50	100	150	200	300	400
Number of homes	10	15	28	42	7	1

a) In which neighborhood is the mean more representative (i.e. which neighborhood has the lower relative dispersion)?

Coefficient of variation[KPC10]

Working class

$a = \sqrt{\sum(X1 - X)2}\ / N$

High class

CV HG = $a/\bar{x}$ = 77.35/ 190
= 40.71%

$a = \sqrt{(50 - 166.01)2} + \cdots$

$a = \sqrt{4233.67}$

$a = 65.06\ m^3$

$CV_{WC} = a/\bar{x} = 65.06 / 166.01$
$= 39\%$

The working class neighborhood has the more representative mean.

b) Which neighborhood has the higher consumption in relative terms?

$Z\text{- score} = \frac{X_i - \bar{x}}{a}$

Working class	High class
$X_i = 150\ m^3$	$X_i = 165\ m^3$
$Z_{WC} = (150-166.01)/65.06 = -0.24$	$Z_{HC} = (165-190)/77.35 = -0.32$

The high class neighborhood has the higher consumption in relative terms.

c) In which neighborhood is gas consumption less concentrated?

P_i Q_i

Categories	Freq.	Accum.	Consumption	Accum.	$p_i - q_i$
Low	0.4	0.4	0.35	0.35	0.05
Medium	0.3	0.7	0.25	0.6	0.01
High	0.3	1	0.4	1	0
	1				

Working class $I_G = 0.1584$

High class $I_G = = \sum_{i=1}^{n-1} (pi - qi) / P_i = (0.05 + 0.01) / (0.4 + 0.7) = 0.11$

Gas consumption is less concentrated in the high class neighborhood.

Problem 1 (1.5 points) A study about daily sales of two firms (A and B) in the food sector in Comunitat Valenciana region has been conducted. There is information available: A firm has a mean daily sales of 200 thousand € and a standard deviation of 10 thousand €, while B firm has a mean daily sales of 400 thousand € with a standard deviation of 40 thousand €.

a) Which firm have more homogeneous daily sales (that is, lower relative dispersion)?

b) One day, A firm has sales equal to 190 and B firm has sales equal to 350. Which firm has had higher sales in relative terms?

c) Salaries distribution has been analyzed. A firm salaries are classified in 3 categories (low, medium and high). The result is that 10% of workers are in the high category and own 30% of the total salaries amount, while half of workers are in the low category and own 20% of the total salaries amount. In case of B firm, salaries' distribution Gini Index is equal to 0.3. Which firm has a more concentrated salaries distribution? Justify your answer.

c) The distribution of salaries has been analyzed. Firm A's salaries are divided into three categories (low, medium and high): 10% of its workers are in the high category and earn 30% of the total amount of salaries, while 50% of its workers are in the low category and earn 20% of the total amount of salaries. The Gini index for the distribution of salaries in firm B is 0.3. In which firm is the distribution of salaries more concentrated? Justify your answer.

Company A

p_i q_i

Categories	Freq.	Accum.		Salaries	Accum.	$p_i - q_i$
Low	0.5	0.5		0.2	0.2	0.3
Medium	0.4	0.9		0.5	0.7	0.2
High	0.1	1		0.3	1	0

$I_{G\,(A)} = \sum_{i=1}^{n-1} pi - qi \;/\; \sum_{i=1}^{n-1} pi = \frac{0.3 + 0.2}{0.5 + 0.4} = 0.35$ [KPC11]

Company B [KPC12]

$I_{G\,(B)} = 0.3$

$I_{G\,(A)}$ → has more inequality and a more highly concentrated distribution. [KPC13]

UNIT 2: BIVARIATE DATA ANALYSIS

THEORY

This second unit explores the concept of bivariate data analysis, which examines the relationship between two variables. Understanding how two variables interact is crucial for identifying correlations, trends and potential cause-and-effect relationships.

On this course we will cover several techniques for analyzing and interpreting bivariate data, such as scatterplots, correlation coefficients and regression analysis. You will learn how to visually represent the relationship between two variables and quantify their strength and direction. These methods will enable you to draw meaningful conclusions about the connection between datasets.

Bivariate analysis is widely used across many fields. Whether we are studying the link between advertising and sales in business, the relationship between age and income in economics, or scientific relationships such as height and weight in biology, this type of analysis is key to making data-driven decisions.

Bivariate analysis can be summarized as the analysis of relationships between variables (x,y) to determine the direction and degree (strength) of those relationships (if they exist).

CROSS TABLES (rows and columns)

Cross tables display the number of observations for all combinations of values for two categorical variables.
The observations are ordered in pairs.

x/y	y(1)	y(j)	y(s)	n(i)
x(1)	n(1,1)	n(1,j)	n(1,s)	n(1)
x(i)	n(i,1)	n(i,j)	n(i,s)	n(i)
x(r)	n(r,1)	n(r,j)	n(r,s)	n(r)
n(j)	n(1)	n(j)	n(s)	N

n(i,j): joint frequency for x(i) and y(j)
n(i): marginal frequency for x(i) and n(i)
n(j): marginal frequency for y(j)

f(i/j) = n(ij)/n(j) = joint frequency / conditioned marginal frequency

There can be as many conditioned variables as values the variables take.

(x,y) are independent variables if the frequency of a conditioned variable is equal to the non-conditioned frequency or if the relative joint frequency is equal to the multiplication of both relative marginal frequencies.

X	f	f_r	$X_r * f_r$
X_1	n_1	$n_1 / N = f_1$	$X_r * f_r$
X_2	n_2	$n_2 / N = f_2$	$X_r * f_r$
X_3	n_3	$n_{13} / N = f_3$	$X_r * f_r$
.	.	.	.
.	.	.	.
.	.	.	.
X_r	n_r	$n_r / N = f_r$	$X_r * f_r$

The population covariance measures the direction of the linear relationship between two variables:

- Cov(x,y) > 0: positive sign, direct relationship, x and y tend to move in the same direction.
- Cov(x,y) < 0: negative sign, inverse relationship, x and y tend to move in opposite directions.
- Cov(x,y) = 0: zero, x and y have no linear relationship but they can have a different type of relationship.

If x and y are independent, then cov(x,y) = 0.

Covariance Formula

For Population

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

The correlation coefficient measures the direction and strength of the linear relationship between two variables:

- r(xy) = -1: this indicates a perfectly negative linear relationship; a decreasing straight line closer to -1 indicates a stronger relationship.
- r(xy) = 1: this indicates a perfectly positive linear relationship; an increasing straight line closer to 1 indicates a stronger relationship.
- r(xy) = 0: this indicates that there is no linear relationship; the closer r(xy) is to 0 indicates a weaker linear relationship[KPC14].

If x and y are independent, r(xy) = 0

$$\rho_{xy} = \frac{Cov(r_x, r_y)}{\sigma_x \sigma_y}$$

Linear transformation of the x variable into a y variable:

y = a + bx

Mean: $\bar{y} = a + \bar{bx}$
Variance: Sy^2 = b^2 x Sx^2

Standardizing the variable: $y = x - \bar{x}/Sx$

A standardized variable is one whose mean is subtracted and divided by its standard deviation. The mean of a standardized variable is always 0 and the standard deviation of a standardized variable is always 1.

EXERCISES

The manager of a used-car dealership is very interested in the resale price of used cars. The manager feels that the age of the car is important in determining the resale value. He collects data on the value of the car (in thousands of dollars) and the age of the car (in years):

Age (variable X)	Value (variable Y)
3	50
5	43
7	35
8	40
11	15

a) Compute the covariance.
b) Compute the correlation coefficient.

Solution:

X_i	Y_i	$X_i - \bar{x}$	$Y_i - \bar{Y}$	$(X_i - \bar{x})^2$	$(Yi - Y)^2$	$(X_i - \bar{x})^* (Yi - Y)$
3	50	-3.8	13.4	14.44	179.59	-50.92
5	43	-1.8	6.4	3.24	40.96	-11.52
7	35	0.2	-1.6	0.04	2.56	-0.32
8	40	1.2	3.4	1.44	11.56	4.08
11	15	4.2	-21.6	17.64	466.56	-90.72
34	183			36.8	701.2	-149.4

a) Compute the covariance

Average $\bar{x}$ = $\sum X_i / n = 34 / 5 = 6.8$

Average $\bar{y}$ = $\sum Y_i / n = 183 / 5 = 36.6$

Variance (x) = $\sum(X_i - \bar{x})^2 / n = 36.8 / 5 = 7.36$

σ (x) = $\sqrt{7.36} = 2.71$

Variance (y) = $\sum(Yi - \bar{y})^2 / n = 701.2 / 5 = 140.24$

σ (y) = $\sqrt{140.24} = 11.84$

Covariance $= \sum(X_i - \bar{x})(Y_i - \bar{y}) / n$

CoV (x,y) = -149.4/ 5 = -29.88

Interpretation: the relationship between age and value is negative because -29.88 < 0

b) Compute the correlation coefficient

Correlation coefficient $R = CoV(x,y) / \sigma_x \sigma_y = -29.88 / 2.71 \times 11.84 = -0.93$

Because the correlation coefficient is close to -1, the negative relationship is strong[KPC15].

UNIT 3: REGRESSION

Regression is a statistical method that is used to model the relationship between a dependent variable (the outcome you want to predict) and one or more independent variables (the factors you think influence the outcome).

Key points:

- **Purpose:** to understand and quantify the relationship between the variables.
- **Types:** simple linear regression (one independent variable), multiple linear regression (multiple independent variables), and non-linear regression (the relationship is not a straight line).
- **Applications:** regression is widely used in fields such as finance, economics, social sciences and natural sciences.
- **How it works:** regression models create an equation that best fits the data, thus allowing you to make predictions based on new data.

Example introduction

Regression analysis is a powerful statistical technique employed to explore the relationship between a dependent variable and one or more independent variables. By constructing mathematical models, regression enables us to make predictions, assess the strength of these relationships, and gain insights into the underlying factors influencing the outcome of interest.

THEORY

Dependent variable (x) and independent variable (y)
Endogenous variable and exogenous variable

Regression attempts to fit the linear relationship between two variables:
y = a + bx

- b(slope line) = Cov(x,y)/Sx
- a = y - bx

If r(xy) = 1 or r(xy) = -1, the linear relationships are perfect and both regression lines coincide.
The slope of the line is b: it is the change in y when x changes by 1 unit.

Regression analysis can have different objectives:

- to predict the value of the dependent variable based on a value of 1 for the independent variable[KPC16];
- to explain how changes to the independent variable affect dependent variables[KPC17].

The residual is the difference between the real value of y and the predicted value for y.

Coefficients a and b are obtained as the values that minimize the sum of the squared residuals.

There is a relationship between correlation and regression:
b x b' = (S(xy)/S(x)^2) x (S(xy)/S(y)^2) = r(xy)^2

GOODNESS OF FIT

In simple terms, *goodness of fit* in a regression model is a measure of how well the model fits the real data. It tells us how well the regression line (or curve) drawn on the data represents the actual relationship between the variables.

Why is it important?

- **Model validity:** a good fit indicates that the model is valid and can be used to make reasonably accurate predictions.
- **Model comparison:** this enables us to compare different models and select the one that best fits the data.
- **Interpretation of results:** this helps us to interpret the results of the regression and understand the importance of the independent variables.

How is it measured?

Several statistics are used to measure goodness of fit. The most common statistics are:

- **R-squared (R^2):**

- R^2 explains the proportion of the variability of the dependent variable that is explained by the model.
- It ranges from 0 to 1.
- A value close to 1 indicates good fit, while a value close to 0 indicates poor fit.

- **Root Mean Squared Error (RMSE):**
  - RMSE measures the average difference between the values predicted by the model and the actual values.
  - Low RMSE values indicate good fit.
- **Standard Error of the Estimate (SEE[KPC18]):**
  - SEE is similar to RMSE but is expressed in the same units as the dependent variable.

What factors affect goodness of fit?

- **The number of independent variables:** as the number of variables increases, R^2 tends to increase but this does not always mean a better model.
- **Model complexity:** although more complex models can fit the data better, they can also overfit.
- **Data quality:** data with errors or outliers can significantly affect the model's fit.

How can we improve goodness of fit?

- **Variable selection:** include only those variables that are truly relevant to the model.
- **Variable transformation:** transforming variables (e.g., taking logarithms) can sometimes improve the fit.
- **Use of non-linear models:** if the relationship between the variables is not linear, a non-linear model may be more appropriate.
- **Consideration of interactions:** the relationship between two variables can sometimes depend on a third variable.

In summary, goodness of fit is a fundamental tool in regression analysis. By understanding and evaluating the fit of a model, we can make more informed decisions about its validity and usefulness.

SST = SSR + SSE
(SUM TOTAL SQUARES = SUM SQUARES REGRESSION + SUM SQUARES ERROR)

TOTAL VARIABILITY OF Y = EXPLAINED COMPONENT + ERROR COMPONENT

TOTAL VARIABILITY OF Y measures the variation in the x(i) values and their mean $\bar{y}$.
EXPLAINED COMPONENT explains the variation due to the linear relationship between X and Y.
ERROR COMPONENT is any variation due to the linear relationship between X and Y or any unexplained deviation of points from the regression line[KPC19].

COEFFICIENT OF DETERMINATION R^2

The coefficient of determination measures how good the model is: it is the part of the total variation in y that is explained by x.
0<=R^2<=1 THE LARGER THE R^2, THE BETTER THE REGRESSION
R^2 = r(xy)^2

EXERCISES

Problem 2 (1.5 points) Net income per capita (X variable) and the expenditure in leisure and culture (Y variable) are available for Spanish regions (autonomous communities) in 2021, both measured in euros. INE data yield the following summary statistics:

$\bar{x}=12133.11$	$\bar{y}=499.12$	$S_X=1907.32$	$S_Y=104.04$	$S_{XY}=124718.49$

a) Analyze the linear relationship between both variables. Compute the appropriate statistic and interpret direction and strength of this relationship.

b) Obtain the regression line that explains the expenditure in leisure and culture as a function of net income per capita. Interpret the coefficients.

c) Predict the expenditure in leisure and culture for a region (autonomous community) that has a net income per capita of 12500 euros. How reliable is this prediction? Compute a measure of the goodness of fit and interpret the results.

Solution

X → Net income per capita

Y → Expenditure in leisure and culture

$\bar{x}$ = 12,133.11[KPC20]

Y = 499.12

S_x = 1907.32 S_y = 104.04 S_{xy} = 124,718.49

a) Analyze the linear relationship between the two variables. Compute the appropriate statistic and interpret the direction and strength of this relationship.

Correlation coefficient

ρ = CoV (x,y)/ a_x a_y = S_{xy} / S_x S_y = 124,718.49 / 1907.32 x 104.04 = 0.6285 = 62.85%

The positive relationship is strong because:

$-1 \leq \rho \leq 1$

Strong negative relationship — Strong positive relationship

b) Obtain the regression line that explains the expenditure in leisure and culture as a function of net income per capita.
Interpret the coefficients.

regression line: $y = a + b\,X$ — a = Y – intercept — b = slope

$b = CoV(x,y)/\,Var(x) = S_{xy} / S_x = 124{,}718.49 / (104.04)^2 = 0.0343$

$a = \bar{y} - b\bar{x} = 499{,}12 - 0.0343 \times 12{,}133.11 = 82.95$

The regression line is $y = 82.95 + 0.0343\,X$

c) Predict the expenditure in leisure and culture for a region that has a net income per capita of 12,500 €.
How reliable is this prediction? Compute a measure of the goodness of fit and interpret the results.

Predict "y" when x = 12,500 so $y = 82.95 + 0.0343 \times 12{,}500 = 511.70$ €

Reliability : $R^2 = \rho_{xy}^2 = (0.6285)^2 = 0.305$ GOF

39.5% of y is explained by the linear relationship with X, so reliability is low.

UNIT 4: TIME SERIES

A **time series** is a sequence of data points collected at specific time intervals. These data points represent measurements of a variable over time: for example, the daily closing prices of a stock, monthly sales figures, and annual temperature readings are all time series.

Key characteristics of time series:

- **Time dependence:** each data point is influenced by the previous data points.
- **Order matters:** the sequence of the data is crucial for analysis.
- **Trends:** time series often exhibit trends such as increasing or decreasing values over time.
- **Seasonality:** many time series have seasonal patterns, i.e. repeating patterns at regular intervals (e.g. yearly, monthly).
- **Cyclicity:** longer-term fluctuations that do not have a fixed period are known as cycles.
- **Noise:** random variations not explained by trend, seasonality or cycle.

Why study time series?

- **Forecasting:** predicting future values of the variable based on past observations.
- **Anomaly detection:** identifying unusual patterns or outliers.
- **Understanding causal relationships:** exploring how changes in one variable affect another over time.
- **Evaluating the impact of interventions:** assessing the effects of specific events or policies.

Common time series analysis techniques:

- **Visualization:** plotting the data to identify trends, seasonality and other patterns.
- **Stationarity:** checking whether the statistical properties of the series remain constant over time.
- **Decomposition:** breaking down the time series into its components (trend, seasonality, cycle and noise).
- **ARIMA models:** ARIMA (AutoRegressive Integrated Moving Average) models are used to model and forecast stationary time series.
- **Exponential smoothing:** a family of forecasting methods that assign exponentially decreasing weights to older observations.
- **SARIMA models:** Seasonal ARIMA models used for time series with seasonality.
- **Other models:** many other specialized models for time series analysis exist, including GARCH models for volatility, and state space models.

Applications of time series analysis:

- **Economics:** forecasting GDP, inflation and stock prices.
- **Finance:** analyzing asset returns and risk management.
- **Environmental science:** studying climate change and weather patterns.

- **Marketing:** forecasting sales and analyzing customer behavior.
- **Engineering:** monitoring industrial processes and quality control.

In conclusion, [KPC21]time series analysis is a powerful tool for understanding and modeling data that change [KPC22]over time. By understanding the underlying patterns and trends in time series data, we can make more informed decisions and predictions.

From a practical point of view, time series have a trend component, which means that data are taken over a long period of time to observe an increase or decrease and determine whether the trend goes upwards or downwards. Time series also have a seasonal component, which means that we can observe the behavior of the data over a period of time (usually a year). Time series also have a cyclical component that is measured peak to peak or trough to trough[KPC23]. Last but not least, time series have an unpredictable component since they can adapt to any behavior at any time due to random variations.

In mathematical terms we can differentiate between the additive model and the multiplicative model of a time series.

Additive model: T + S + C + I
Multiplicative model: T*S*C*I
where T = temporality, S=seasonality, C=the cyclical component of time, and I=unpredictability (irregular component).

Another important topic is the moving average. This shows us an average of the pattern of movement over time. The result depends on the value we choose for M in the following equation

$$X_t^* = \frac{1}{2m+1}\sum_{j=-m}^{m} X_{t+j} \quad (t = m+1, m+2, \ldots, n-m)$$

For example, if our period of time for the equation is five years, we put m=2 [KPC24]in order to obtain in the denominator 5[KPC25].

Centered moving average is another measure to take into account. This measure uses previous and future data to calculate the average at a specific time.

To obtain a centered moving average we need a period s where s is an even number (quarterly data=4 and monthly data=12).

To obtain the centered moving average we solve the following equation:

$$x_t^* = \frac{x_{t-0.5}^* + x_{t+0.5}^*}{2} \qquad (t = \frac{s}{2}+1, \frac{s}{2}+2, \ldots, n-\frac{s}{2})$$

The seasonal impact is also assessed using the seasonal index method to observe increases or decreases in the series.

We obtain this index by dividing the current sales value by the centered moving average for that period.

We use the following equation:

$$100\,\frac{x_t}{x_t^*}$$

Exponential smoothing is conducted through weighted mean averages.

It is used for short-term forecasting.[KPC26]

The criteria for the weight, α (the smoothing coefficient), are as follows:

It is chosen subjectively.
It ranges from 0 to 1.
A smaller α provides more smoothing and a larger α provides less smoothing.
A low value (closer to 0) is selected to smooth out unwanted cyclical and irregular components.
A high value (closer to 1) is selected for forecasting, especially for smoother time series.

Forecasting seasonal time series. Here we assume time series as a period named **s.** We use the Holt-Winters method with recursive estimates from the previously obtained series.

The estimates use three factors: a level factor, a trend factor, and a multiplicative seasonal factor.

First we must obtain the smoothed series, which are calculated from:

$$\hat{x}_t = (1-\alpha)\,\hat{x}_{t-1} + (\alpha)x_t$$

From the time n, the forecasts are therefore:

$$\hat{x}_{n+h} = \hat{x}_n$$

where:

= exponentially smoothed value for period t.

= exponentially smoothed value already computed for period i - 1[KPC27].

x_t = observed value in period t.

α = weight (smoothing coefficient); $0 < \alpha < 1$.

This is better visualized by means of a numerical example:

Suppose we use weight $\alpha = 0.2$, then: $\hat{x}_t = (1-0.2)\hat{x}_{t-1} + 0.2x_t$

Time Period (t)	**Sales (X_t)**	**Forecast from prior period ( )**[KPC28]	**Exponentially Smoothed Value for this period ( )**[KPC29]
1	23	--	23
2	40	23	(.8)(23)+(.2)(40)=26.4
3	25	26.4	(.8)(26.4)+(.2)(25)=26.12
4	27	26.12	(.8)(26.12)+(.2)(27)=26.296
5	32	26.296	(.8)(26.296)+(.2)(32)=27.437
6	48	27.437	(.8)(27.437)+(.2)(48)=31.549
7	33	31.549	(.8)(31.549)+(.2)(33)=31.840
8	37	31.840	(.8)(31.840)+(.2)(37)=32.872
9	37	32.872	(.8)(32.872)+(.2)(37)=33.697
10	50	33.697	(.8)(33.697)+(.2)(50)=36.958
etc.	etc.	etc.	etc.

Forecasting time period (t+1).
The smoothed value in the current period is used as the forecast value for the next period t+1.

Forecasting autoregressive models:

Consider time series observations $x_1, x_2, \ldots, x_t$

Suppose that an autoregressive model of order p has been fitted to these data.

$$x_t = \hat{\gamma} + \hat{\varphi}_1 x_{t-1} + \hat{\varphi}_2 x_{t-2} + \cdots + \hat{\varphi}_p x_{t-p} + \varepsilon_t$$

At time n we obtain forecasts of future values of the series from:

$$\hat{x}_{n+h} = \hat{\gamma} + \hat{\varphi}_1 \hat{x}_{n+h-1} + \hat{\varphi}_2 \hat{x}_{n+h-2} + \cdots + \hat{\varphi}_p \hat{x}_{n+h-p} \qquad (h = 1,2,3,\ldots)$$

where for $h > 0$, [KPC30] is the forecast of X_{t+h} at time n and for $h \leq 0$, [KPC31] is simply the observed value of X_{t+h}

Box-Jenkins Methodology

1. Summary statistics from the available data are used to select a specific model that may be appropriate from the general class.
2. The specific model chosen will have some unknown coefficients. These must be estimated from the available data using techniques such as least squares.
3. Checks are applied to determine whether the estimated model provides an adequate representation of the available time-series data.

If stage three reveals any inadequacies, alternative specifications can be used and the process is iterated until a satisfactory model is found.
The advantages of the Box-Jenkins approach to forecasting are
- flexibility: a wide range of predictors is available and their choice is based on data evidence; and
- results: compared to other methods that use actual economic and business time series, it usually performs very well.

INDEX NUMBERS

Index numbers enable relative comparisons over time.
Index numbers are reported relative to a Base Period Index.
The Base Period Index = 100 by definition.
Index numbers are used [KPC32]for an individual item or measurement.

Another important topic in relation to index numbers is the single price index.
The single price index considers observations over time on the price of a single item.

To create a price index, one time period is chosen as a base and the price for every period is expressed as a percentage of the base period price.
Let p_0 denote the price in the base period and **let p_1 be the price in a second period**.
The price index for this second period is:

$$100\left(\frac{p_1}{p_0}\right)$$

Calculation of a single price index:

Year	Price	Index (base year = 2005)
2000	272	85.0
2001	288	90.0

$$I_{2001} = 100\frac{P_{2001}}{P_{2005}} = (100)\frac{288}{320} = 90$$

The aggregate price index is a measure established to determine the rate of change from a base period for a group of items. We differentiate between an unweighted aggregate price index and a weighted aggregate prices index, which includes the Laspeyres [KPC33]index[KPC34].

- Unweighted aggregate price index for period t for a group of K items:
-

i = item

t = time period

K = total number of items

$$100\left(\frac{\sum_{i=1}^{K} p_{ti}}{\sum_{i=1}^{K} p_{0i}}\right)$$

$\sum_{i=1}^{K} p_{ti}$ sum of the prices for the group of items at time t

$\sum_{i=1}^{K} p_{0i}$ sum of the prices for the group of items in time period 0

4.2 [KPC35]TREND EQUATIONS I[KPC36] CHANGES IN SCALE

How is the theory of trend equations applied?

Origin of changes in trend equations.

This is best explained with an example.

From a time-series the annual trend is known: $y_t^* = 12000 + 288t$,
where y are the annual sales (in tens of €) of a firm and t is the time (in years) starting in 2015. Seasonal indexes for quarters are known (assuming a multiplicative model).

a) Change the annual trend while taking 2020 as the origin.

a) [KPC37]Since the annual trend equation originates in 2015, for that year t = 0.
For 2020, t = 2020-2015 = 5. If we substitute this in the trend equation:

$$y_5^* = 12000 + 288 \text{ x } 5 = 13440$$

The new annual trend equation with origin in 2020 is:
$y_{t'}^{**} = 13440 + 288t'$, where $t' = 0 \rightarrow 2020$.

b) Find the quarterly trend while taking the central quarter of year 2023 as the origin.

b) By taking the annual trend obtained in a), we obtain the quarterly trend equation by dividing coefficient a by 4 and coefficient b by 4^2:

$$y_{t''}^{***} = \frac{13440}{4} + \frac{288}{4^2}t'' = 3360 + 18t''$$

where t'' are quarters with origins in the central quarter of 2020.
Since the origin of the new annual trend is in 2020, when we obtain the quarterly trend equation, this trend is located in the central quarter of 2020 (the original quarter is made up of the second half of the second quarter and the first half of the third quarter in 2020).

c) Make a seasonally adjusted forecast for the fourth quarter of 2024 (adapted from Exercise 3.1.9 on page 92 in the handbook by Murgui et al. (2002)).

Taking the quarterly trend equation obtained in b, we need to count quarters from the central quarter in 2020 to the fourth quarter in 2024, i.e.: $Nº\ years\ =\ 2024 - 2020 = 4\ years$

$$Nº\ quarters\ = 4\ years \text{ x } 4\ quarters + 1.5\ quarters = 17.5\ quarters$$

Now we can make the forecast for the fourth quarter of 2024:

$$y_{2024-4Q}^{***} = 3360 + 18 \text{ x } 17.5 = 3675$$

To correct this forecast to a seasonally adjusted forecast:

$$\hat{y}_{2024-4Q} = y_{2024-4Q}^{***} \cdot SI_4$$

where SI_4 is the seasonal index for the fourth quarter of the year, which is obtained as follows: $SI_4 = 4 - SI_1 - SI_2 - SI_3 = 4 - 0.7 - 0.9 - 1.3 = 1.1$
The expected sales forecast, in tens of euros for the fourth quarter of 2024, is therefore:
$\hat{y}_{2024-4Q} = y_{2024-4Q}^{***} \text{ x } SI_4 = 3675 \text{ x } 1.1 = 4042.5$

4.3 APPLICATION: Exercises.

The evolution of labor costs over time can be described by the following annual trend:

$$T = 1405 + 47t$$

with t=0 in 2001.

a) **Find the quarterly trend while setting the origin in the first quarter of 2012.**
b) **Predict the value of labor costs for the fourth quarter of 2015. Adjust this prediction using the following seasonal indices:**

I	II	III	IV
0.966	1.016	0.955	?

c) **Find the real rate of change of the annual trend in labor costs between 2001 and 2013. Use the CPI to deflate the trend in the values of labor costs.**

Year	CPI (2001)	CPI (2006)
2001	100.0	
2006	117.6	100.0
2011		112.1
2013		116.5

$T = 1405 + 47t$ with t = 0 in 2001

a) Find the quarterly trend while setting the origin in the first quarter of 2012.

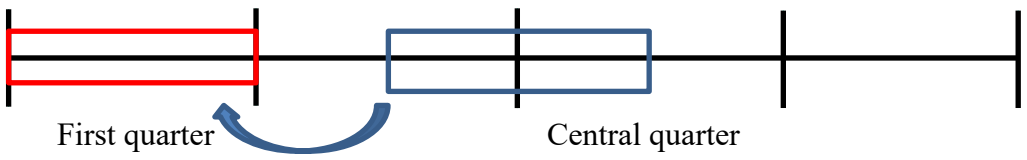
From this annual trend equation we obtain the quarterly trend equation by dividing coefficient a by 4 and coefficient b by 4^2:

$$T = \frac{1405}{4} + \frac{47}{4^2}t' = 351.25 + 2.94t'$$

where t' are quarters and the origin is in the central quarter of 2001.

Taking the quarterly trend equation, we need to count quarters from the central quarter in 2001 to the first quarter of 2012:

$11\ years$ x $4\ quarters - 1.5\ quarters = 42.5\ quarters$

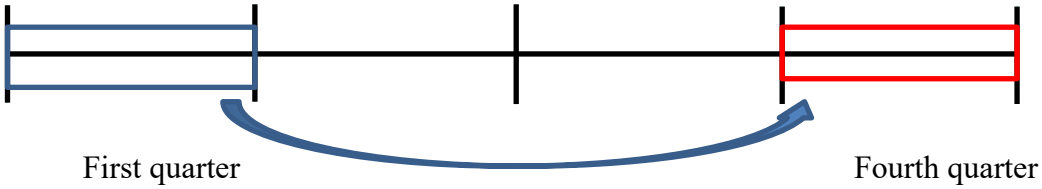


$$T = 351.25 + 2.94t' = 351.25 + 2.94 \cdot 42.5 = 476.2$$

So the quarterly trend equation with origin in the first quarter of 2012 is:
$T = 476.2 + 2.94t''$
where t'' are quarters and the origin is in the first quarter of 2012.

b) Prediction for the fourth quarter of 2015, adjusted by seasonality.

We need to count quarters from the first quarter of 2012 to the fourth quarter of 2015:
$3\ years$ x $4\ quarters + 3\ quarters = 15\ quarters$



The prediction for the fourth quarter of 2015 is:

$T_{4-2015} = 476.2 + 2.94t'' = 476.2 + 2.94 \text{ x } 15 = 520.3$

The seasonal index for the fourth quarter is:

$SI_4 = 4 - SI_1 - SI_2 - SI_3 = 4 - 0.966 - 1.016 - 0.955 = 1.063$

The seasonally adjusted prediction is:

$T_{4-2015}^{seasonally\ adjusted} = T_{4-2015} \text{ x } SI_4 = 520.3 \text{ x } 1.063 = 553.08$

c) Real rate of change of the annual trend in labor costs between 2001 and 2013.

Prediction for 2001 (t = 0): $T = 1405 + 47 \text{ x } 0 = 1405$
Prediction for 2013 (t = 12): $T = 1405 + 47\text{x } 12 = 1969$

Now we have to complete the series of the CPI:

Year	CPI (2001)	CPI (2006)
2001	100.0	
2006	117.6	100.0
2011	$\frac{117.6 \text{ x } 112.1}{100} = 131.83$	112.1
2013	$\frac{117.6 \text{ x } 116.5}{100} = 137$	116.5

Real labor costs in 2001: $\frac{1405}{1} = 1405$

Real labor costs in 2013: $\frac{1969}{1.37} = 1437.23$

The real rate of change is therefore:

$$r_t = \left(\frac{Y_t}{Y_{t-1}} - 1\right) \text{x } 100 = \left(\frac{1437.23}{1405} - 1\right) \cdot 100 = 2.29\%$$

ANOTHER PROBLEM

The annual income (in thousand euros) of an automobile sales firm can be expressed by the following annual trend equation: $T_t = 1200 + 160 \text{ x } t$, with origin in 2010. The quarterly seasonal indices in this sector are:

I	II	III	IV
0.6	1.1	1.4	?

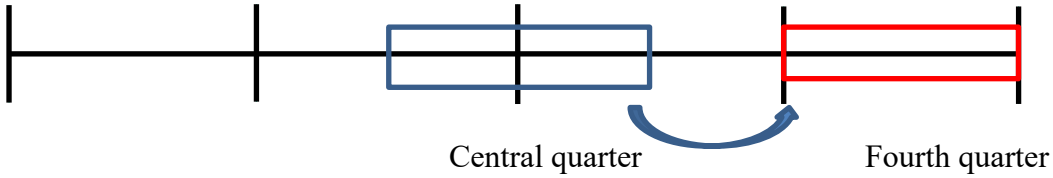
a) Make a seasonally adjusted forecast for income in the fourth quarter of 2014.
b) Taking the annual trend equation, compute the variation rate in real terms (constant euros) of annual income in the period 2011–2013 while taking into account that the Consumer Price Index (base 2011) for 2013 is 103.89.

Y: annual income (in thousand euros) of an automobile sales firm.
$T_t^* = 1200 + 160t$ with origin in 2010.

a) Make a seasonally adjusted forecast for income in the fourth quarter of 2014.

The annual trend equation is:
$T_t^* = 1200 + 160t$
We then compute the quarterly trend equation by dividing coefficient a by 4 and coefficient b by 4^2:
$T_{t'}^* = \frac{1200}{4} + \frac{160}{4^2}t' = 300 + 10t'$
where t' are quarters with origin in the central quarter of 2010.

Prediction for the fourth quarter of 2014.
First we need to compute the quarters from the central quarter of 2010 to the fourth quarter of 2014:
$4\ years$ x $4\ quarters + 1.5\ quarters = 17.5\ quarters$
Year 2014



Prediction for the fourth quarter of 2014:
$T^*_{17.5} = 300 + 10 \cdot 17.5 = 475$

Under a multiplicative model the seasonally adjusted prediction of annual income in fourth quarter of 2014 is:

$$\hat{T}_{4-14} = T^*_{17.5} \text{ x } SI_4$$

where SI_4 is the seasonal index for the fourth quarter in the year. We do not know its value but it can be computed from all seasonal indexes as follows:

$$SI_4 = 4 - SI_1 - SI_2 - SI_3 = 4 - 0.6 - 1.1 - 1.4 = 0.9$$

The predicted annual income for the fourth quarter of 2014 is therefore:
$\hat{T}_{4-14} = T^*_{17.5} \text{ x } SI_4 = 475 \cdot 0.9 = 427.5$

b) Variation rate in real terms (constant euros) of the annual income in the period 2011–2013 while taking into account that the CPI (base 2011) in 2013 is 103.89.

Annual income 2011: $T^*_t = 1200 + 160 \cdot 1 = 1360$
Annual income 2013: $T^*_t = 1200 + 160 \cdot 3 = 1680$

Real annual income 2011 (in constant euros): $\frac{1360}{1} = 1360$
Real annual income 2013 (in constant euros): $\frac{1680}{1.0389} = 1617.1$

In real terms the variation rate of annual income for 2011-2013 is therefore:

$$r_t = \left(\frac{Y_t}{Y_{t-1}} - 1\right) \text{x } 100 = \left(\frac{1617.1}{1360} - 1\right) \text{ x } 100 = 18.90\%$$

The real variation rate of annual income for 2011-2013 is 18.90%.

UNIT 5. PROBABILITY BASICS

Probability is a branch of mathematics that deals with the study of chance and uncertainty. It helps us to quantify the likelihood of events occurring, such as guessing the flip of a coin or predicting the weather.

Key Concepts:

- **Experiment:** a procedure that results in an observable outcome.
- **Outcome:** a possible result of an experiment.
- **Sample space:** the set of all possible outcomes of an experiment.
- **Event:** a subset of the sample space.
- **Probability:** a numerical measure of the likelihood of an event occurring.

Probability Scale:

Probabilities are expressed as numbers between 0 and 1:

- **0:** an impossible event (e.g., getting 7 when rolling a dice).
- **1:** a certain event (e.g., getting heads or tails when flipping a coin).
- **0.5:** an equally likely event (e.g., drawing a red card from a standard deck).

Types of Probability:

- **Classical probability:** based on equally likely outcomes.
- **Empirical probability:** based on observed data.
- **Subjective probability:** based on personal belief or judgment.

Basic Probability Rules:

- **Rule of total probability:** the probability of an event is the sum of the probabilities of all mutually exclusive outcomes that lead to the event.
- **Rule of complementary events:** the probability of an event not occurring is 1 minus the probability of the event occurring.
- **Conditional probability:** the probability of one event occurring given that another event has already occurred.

Applications of Probability:

- **Statistics:** inferring conclusions from data.
- **Finance:** risk management, investment analysis.
- **Insurance:** calculating premiums.
- **Quality control:** assessing product reliability.
- **Gambling:** understanding odds.

In this unit we will explore these concepts in greater detail and learn how to calculate probabilities for various scenarios. First we must learn the basics:

Basic terms

Random Experiment: a process leading to an uncertain outcome.
Basic Outcome: a possible outcome of a random experiment.
Sample Space (S): the collection of all possible outcomes of a random experiment.
Event (E): any subset of basic outcomes from the sample space.

Intersection of events: if A and B are two events in a sample space S, then the intersection, A ∩ B, is the set of all outcomes in S that belong to both A and B.

Mutually exclusive: A and B are Mutually Exclusive events if they have no basic outcomes in common

i.e., the set A ∩ B is empty.

Union of events: if A and B are two events in a sample space S, then the union, A U B, is the set of all outcomes in S that belong to either A or B.

Collectively exhaustive: E_1 U E_2 U . . . U E_k = S.

The Complement of an event A is the set of all basic outcomes in the sample space that do not belong to A. The complement is denoted A[KPC38].

ASSESSING PROBABILITY

1. CLASSICAL PROBABILITY

All outcomes present in the sample space are equally likely to happen.

$$P(A) = \frac{N_A}{N} = \frac{\text{number of outcomes that satisfy the event A}}{\text{total number of outcomes in the sample space}}$$

To count the possible outcome [KPC39]to determine the combinations (n) taken at time k[KPC40]:

$$C_k^n = \frac{n!}{k!(n-k)!}$$

Permutations: the number of possible arrangements when x objects are to be selected from a total of n objects and arranged in order [with (n – x) objects left over]:

$$P_x^n = n(n-1)(n-2)\ldots(n-x+1) = \frac{n!}{(n-x)!}$$

A numerical example:

Suppose that two letters are to be selected from A, B, C, D. How many combinations are possible (i.e., order is not important)?
Solution: the number of combinations is

$$C_2^4 = \frac{4!}{2!(4-2)!} = 6$$

2.RELATIVE FREQUENCY PROBABILITY

The limit of the proportion of times that an event A occurs in a large number of trials, *n*.

$$P(A) = \frac{n_A}{n} = \frac{\text{number of events in the population that satisfy event A}}{\text{total number of events in the population}}$$

3.SUBJECTIVE PROBABILITY

An individual opinion or belief about the probability of occurrence.

PROBABILITY POSTULATES

If A is any event in the sample space S, then

$$0 \leq P(A) \leq 1$$

Let A be an event in S, and let O_i denote the basic outcomes. Then

$$P(A) = \sum_{A} P(O_i)$$

This means that the summation is over all the basic outcomes in A)

P(S) = 1

The Complement rule

$$P(\overline{A}) = 1 - P(A)$$

The Addition rule

The probability of the union of two events is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

CONDITIONAL PROBABILITY

The probability of one event occurring when knowing that the other event has occurred:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

This is the conditional probability of A occurring given that B has occurred.

Statistical independence: when one probability (let's say A) is not affected by the other probability (let's say B).

The Odds in favor of a particular event are given by the ratio of the probability of the event divided by the probability of its complement.

The odds in favor of A are:

$$\text{odds} = \frac{P(A)}{1-P(A)} = \frac{P(A)}{P(\overline{A})}$$

Bayes' Theorem:

$$P(B_1 \mid A_1) = \frac{P(A_1 \mid B_1)P(B_1)}{P(A_1)}$$

and

$$P(A_1 \mid B_1) = \frac{P(B_1 \mid A_1)P(A_1)}{P(B_1)}$$

EXAMPLE:

A drilling company has estimated a 40% chance of striking oil for their new well. A detailed test has been scheduled for more information. Historically, 60% of successful wells have had detailed tests, and 20% of unsuccessful wells have had detailed tests.
Given that the well has been scheduled for a detailed test, what is the probability that it will be successful?

Let S = successful well.
Let U = unsuccessful well.
P(S) = 0.4, P(U) = 0.6 (prior probabilities).
Define the detailed test event as D.
Conditional probabilities:
P(D|S) = 0.6 and P(D|U) = 0.2.
The goal is to find P(S|D).

$$P(S \mid D) = \frac{P(D \mid S)P(S)}{P(D \mid S)P(S) + P(D \mid U)P(U)}$$
$$= \frac{(.6)(.4)}{(.6)(.4) + (.2)(.6)}$$
$$= \frac{.24}{.24 + .12} = .667$$

EXERCISES & PROBLEMS

Problem 4 (1.5 points) The discrete random variable X can take values from 1 to 4. The following information is available regarding its probability distribution function:

X	1	2	3	4
$P(X=x)$	0.2		0.4	?

Besides, it is known that the Cumulative Distribution Function (CDF), $F(2)=0.3$.

a) Explain the probability function of X variable detailing the properties it must hold.
b) Variable Y is defined as $Y=X+1$. Find the probability of $Y=6$.
c) The expected value of X is equal to 2.8. Find the variance of X.

SOLUTION

x(i)	P(i)
1	0.2
2	
3	0.4
4	?

a) Explain the probability function of X while detailing the properties it must hold.

F(2) = P(X<=2) = 0.3
P(X = 1) + P(X = 2) = 0.3

P(X = 2) = 0.3 - P(X = 1) = 0.3 – 0.2 = 0.1
P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1
0.2 + 0.1 + 0.4 + ? = 1
? = 0.3

b) Variable Y is defined as Y = x + 1. Find the probability of Y = 6.

Y = x + 1
P(Y = 6) = P(X + 1 = 6)
P(X = 5) = 0

c) The expected value of X is 2.8. Find the variance of X.

X(i)	P(i)
1	0.2
2	0.1
3	0.4
4	0.3

XiPi	Xi - $\bar{x}$	(X(i) - $\bar{x}$)^2	(X(i) - $\bar{x}$)^2 P(i)
0.2	-2.6	6.76	1.352
0.2	-2.6	6.76	0.676
1.2	-1.6	2.56	1.024
1.2	-1.6	2.56	0.768

ANOTHER EXAMPLE:

A = "to get an even number"
B = "to get 4 points"

P(A) = 3/6 = ½ = 50%
P(B) = ⅙

P(A(∪)B) = P(A) = 50%

P(A(∩)B) = P(B) = ⅙

P(B/A) = ⅓

P(B/A) = P(A(∩)B/P(A) = P(B)/P(A) = 2/6 = ⅓

UNIT 6. Discrete probability distributions

A discrete probability distribution describes the probability of each possible value of a discrete random variable. A discrete random variable is a variable that can take on only a countable number of values: for example, the number of heads in a coin toss, the number of defective items in a batch, and the number of customers entering a store per hour.

Key components of a discrete probability distribution:

- **Random variable:** the variable of interest.
- **Sample space:** the set of all possible values the random variable can take.
- **Probability function:** a function that assigns a probability to each possible value of the random variable.

Properties of a discrete probability distribution:

- **Non-negativity:** the probability of any value must be between 0 and 1.
- **Exhaustiveness:** the sum of the probabilities for all possible values must equal 1.

Common discrete probability distributions are:

- **The Bernoulli distribution**, which describes the probability of a single binary outcome (e.g., success or failure).
- **The Binomial distribution**, which describes the number of successes in a fixed number of independent Bernoulli trials.
- **The Poisson distribution**, which describes the number of events occurring in a fixed interval of time or space.
- **The Geometric distribution**, which describes the number of trials needed to obtain the first success in a series of Bernoulli trials.

- **The Hypergeometric distribution**, which describes the number of successes in a fixed number of draws without replacement from a finite population.
- **The Negative Binomial distribution**, which describes the number of failures before a fixed number of successes occur in a series of Bernoulli trials.

Why are discrete probability distributions important?

- **For modeling real-world phenomena** – they can be used to model various real-world situations involving discrete random variables.
- **For making predictions** –they can be used to calculate the probabilities of different outcomes and make informed decisions.
- **For statistical inference** – they are used as building blocks for statistical inference techniques.

In this unit some of the most common discrete probability distributions are explained in greater detail.

6.1 General theory for distributions.

Random variables can be divided into discrete variables and continuous variables (see unit 7).

Discrete variables take no more than one value.
Continuous variables can take any value in an interval.

Examples of discrete variables: roll a dice twice or toss a coin several times.

Distribution Function: Cumulative probability function, i.e.

$$F(x_0) = P(X \leq x_0)$$

The probability that x does not exceed the value x0. Example:

x Value	P(x)	F(x)
0	0.25	0.25
1	0.50	0.75
2	0.25	1.00

$$E[X] = \mu = \sum_{x} xP(x)$$

Expected value (mean) and variance.

From the previous example of tossing a coin:

E(x) = (0 x .25) + (1 x .50) + (2 x .25)
= 1.0

Variance:

$$\sigma = \sqrt{(0-1)^2(.25)+(1-1)^2(.50)+(2-1)^2(.25)} = \sqrt{.50} = .707$$

6.2 The Bernoulli distribution

The Bernoulli distribution is the simplest discrete probability distribution. It models the outcome of a single binary experiment, such as flipping a coin or answering a yes/no question.

Consider only two outcomes: "success" or "failure".
Let P denote the probability of success.
Let 1 – P denote the probability of failure.
Define the random variable X:
x = 1 if success, x = 0 if failure.

The Bernoulli probability function is therefore:

$$P(0) = (1-P) \quad \text{and} \quad P(1) = P$$

Mean: P

$$\mu_x = E[X] = \sum_{X} xP(x) = (0)(1-P)+(1)P = P$$

Variance: P(1-P)

$$\sigma_x^2 = E[(X-\mu_x)^2] = \sum_x (x-\mu_x)^2 P(x)$$

$$= (0-P)^2(1-P) + (1-P)^2 P = P(1-P)$$

Example of a binomial distribution:

Denote the use of combinational numbers.
P(x) = the probability of **x** successes in **n** trials with the probability of success **P** on each trial.

x = the number of 'successes' in the sample
(x = 0, 1, 2, ..., n)
n = the sample size (the number of independent trials or observations)
P = the probability of "success".

$$P(x=1) = \frac{n!}{x!(n-x)!} P^x (1-P)^{n-x}$$

$$= \frac{5!}{1!(5-1)!} (0.1)^1 (1-0.1)^{5-1}$$

$$= (5)(0.1)(0.9)^4$$

$$= .32805$$

Mean: n*p

Variance: nP (1-P)

6.2 The Binomial distribution

The binomial distribution models the number of successes in a fixed number of independent Bernoulli trials. Each trial has the same probability of success.

- **Parameters:**
  - **n:** the number of trials.
  - **p:** the probability of success in each trial.
- **Possible values:** 0, 1, 2, ..., n.
- **Probability density function:**
  - P(X = k) = C(n, k) * p^k * (1-p)^(n-k)
    - where C(n, k) is the binomial coefficient (the number of ways to choose k successes from n trials).

Relationship between the Bernoulli and Binomial Distributions:

- A binomial distribution with n = 1 is equivalent to a Bernoulli distribution.

Common applications:

- **Quality control:** counting the number of defective items in a sample.
- **Opinion polls:** determining the proportion of people who support a particular candidate or issue.
- **Genetics:** modeling the inheritance of traits.
- **Sports:** analyzing the probability of winning a game or tournament.

6.3 The Poisson distribution

The Poisson distribution is a discrete probability distribution that models the number of events occurring in a fixed interval of time or space. It is often used to model rare events that occur randomly and independently.

Key characteristics:

- **Discrete:** the random variable can take on only integer values.
- **Rare events:** the probability of a single event occurring in a small interval is small.
- **Independence:** the occurrence of one event does not affect the probability of another event occurring.
- **Constant rate:** the average rate of events per unit of time or space is constant.

Parameter:

- **λ:** the average rate of events per unit of time or space.

Probability function:

The probability function of a Poisson distribution is given by:

$P(X = k) = (\lambda^k * e^{(-\lambda)}) / k!$

where:

- **X** is the random variable representing the number of events.
- **k** is the number of events.
- **λ** is the average rate of events.
- **e** is Euler's number (approximately 2.71828).
- **k!** is the factorial of k.

Mean and Variance:

- **Mean:** $\mu = \lambda$
- **Variance:** $\sigma^2 = \lambda$

Applications:

- **Quality control:** counting the number of defects in a product.
- **Telecommunications:** modeling the number of phone calls received in a given time period.
- **Insurance:** modeling the number of claims filed in a year.
- **Biology:** modeling the number of mutations in a DNA sequence.
- **Physics:** modeling the number of radioactive decays in a given time period.

Example

If the average number of customers arriving at a store per hour is 5, then the number of customers arriving in an hour follows a Poisson distribution with $\lambda = 5$. You can use the Poisson function to calculate the probability of any number of customers arriving in an hour.

In summary, the Poisson distribution is a useful tool for modeling the occurrence of rare events. It has many applications in numerous fields from statistics and engineering to biology and physics.

UNIT 7. Continuous Probability Distributions

A continuous probability distribution describes the probability of a continuous random variable taking on a specific value within a given range. Unlike discrete random variables, continuous random variables can take on any value within a specified interval. Examples include height, weight, time and temperature.

A continuous random variable is a value that can assume any number in the interval[KPC41].

Properties of a continuous probability distribution:

- **Non-negativity:** the density function must be non-negative for all values.
- **Total probability:** the integral of the DENSITY FUNCTION over the entire sample space must equal 1.

Common continuous probability distributions:

- **Uniform distribution** describes a random variable that is equally likely to take on any value within a specified interval.
- **Normal distribution**, also known as the Gaussian distribution, is a bell-shaped curve that is widely used to model many real-world phenomena.
- **Exponential distribution** describes the time between events in a Poisson process.
- **Gamma distribution** is a generalization of the exponential distribution.
- **Beta distribution** describes probabilities or proportions.
- **Chi-square distribution** is used in statistical tests for goodness-of-fit and independence.
- **The Student's t-distribution** is used for hypothesis testing and confidence intervals when the sample size is small or the population standard deviation is unknown.
- **F-distribution** is used to compare the variances of two populations.

Why are continuous probability distributions important?

- **For modeling real-world phenomena** – they can be used to model various real-world situations involving continuous random variables.
- **For making predictions** – they can be used to calculate the probabilities of different outcomes and make informed decisions.
- **For statistical inference** – they are used as building blocks for statistical inference techniques.

Calculations of probabilities:

$$P(a < X < b) = F(b) - F(a)$$

Probability density function(integral)

$$F(x_0) = \int_{x_m}^{x_0} f(x)dx$$

7.1 The Uniform distribution

A uniform distribution is a type of probability distribution in which all possible outcomes are equally likely. In other words, each value within a specified range has the same probability of occurring.

Key characteristics:

- **Continuous or discrete:** uniform distributions can be either continuous or discrete.
- **Equal probability:** all outcomes have the same probability.
- **Range:** the distribution is defined over a specific range (a, b).
- **Shape:** the graph of a uniform distribution is a rectangle.

Probability density function or simply density function:

For a continuous uniform distribution, the density function is:

f(x) = 1 / (b - a) for $a \leq x \leq b$
  0 otherwise

Cumulative distribution function:

The distribution function of a continuous uniform distribution is:

F(x) = 0 for $x < a$
  (x - a) / (b - a) for $a \leq x \leq b$
  1 for $x > b$

a = maximum value
b = minimum value

Mean and variance:

- **Mean:** (a + b) / 2
- **Variance:** (b - a)^2 / 12

Examples:

- **Rolling a dice:** each outcome (1, 2, 3, 4, 5 or 6) has a probability of 1/6.
- **Randomly selecting a number between 0 and 1**[KPC42]**:** any number within this range has an equal probability of being chosen.
- **Waiting times for a bus:** if buses arrive at a constant rate, the waiting time between buses follows a uniform distribution.

Applications:

- **Simulation:** generating random numbers that are uniformly distributed.
- **Sampling:** selecting random samples from a population.
- **Hypothesis testing:** determining the probability of observing a particular outcome under the assumption of a uniform distribution.
- **Modeling:** modeling processes where all outcomes are equally likely.

In summary, uniform distributions are a simple but useful tool for modeling random events where all outcomes are equally probable.

7.2 The Normal distribution

The Normal distribution approximates the probability distribution of a width range[KPC43].

Formula: $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Where e = the mathematical constant approximated by 2.71828
π = the mathematical constant approximated by 3.14159
μ = the population mean
σ^2 = the population variance
x = any value of the continuous variable, $-\infty < x <$

Any normal distribution can be transformed into z-standardized form by applying this formula with mean and variance:

$$Z = \frac{X-\mu}{\sigma}$$

Example from the PowerPoint: if X is distributed normally with a mean of 100 and a standard deviation of 50, the Z value for X = 200 is

$$Z = \frac{X - \mu}{\sigma} = \frac{200 - 100}{50} = 2.0$$

This means that X = 200 is two standard deviations (two increments of 50 units) above the mean of 100.

However, we will see further applications with more examples.

7.3 The Exponential distribution

The exponential distribution is a continuous probability distribution that models the time between events in a Poisson process. A Poisson process is a random process in which events occur continuously and independently at a constant average rate.

Key characteristics:

- **Continuous:** the random variable can take on any non-negative real value.
- **Time between events:** the distribution models the time elapsed between events.
- **Memoryless property:** the probability of an event occurring in the next time interval is independent of the time that has already elapsed.
- **Constant rate:** the average rate of events per unit of time is constant.

Parameter:

- **λ** is the rate parameter representing the average number of events per unit of time.

Probability density function:

The density function of an exponential distribution is given by:

f(x) = λ * e^(-λx) for x ≥ 0
0 otherwise

where:

- **x** is the random variable representing the time between events.
- **λ** is the rate parameter.

- **e** is Euler's number (approximately 2.71828).

Cumulative distribution function:

This is given by:

$F(x) = 1 - e^{(-\lambda x)}$ for $x \geq 0$

Mean and Variance:

- **Mean:** $\mu = 1 / \lambda$
- **Variance:** $\sigma^2 = 1 / \lambda^2$

Applications:

- **Reliability engineering:** modeling the time to failure of a component.
- **Queueing theory:** modeling the time between arrivals at a service facility.
- **Survival analysis:** modeling the time to death or failure of an individual or system.
- **Physics:** modeling the decay time of radioactive particles.
- **Finance:** modeling the time between stock price changes.

Example 1. If the average number of customers arriving at a store per hour is 5, then the time between customer arrivals follows an exponential distribution with $\lambda = 5$. Use the exponential density and distribution functions to calculate the probability of a customer arriving within a certain time period.

Example 2. Customers arrive at the service counter at a rate of 15 per hour. What is the probability that the arrival time between consecutive customers is less than three minutes?

The mean number of arrivals per hour is 15, so $\lambda = 15$.
Three minutes is .05 hours.
P (arrival time < .05) = $1 - e^{-\lambda X} = 1 - e^{-(15)(.05)} = 0.5276$.
There is therefore a 52.76% probability that the arrival time between successive customers is less than three minutes.