Statistical Modeling - 24DS636 (2024-25)

Abhijith M S

2025-03-29

Table of contents

Course Introduction

The website contains course contents of "Statistical Modeling" offered by Abhijith M S, PhD to Masters students pursuing M.Tech in Data Science, during the even semester of the academic year 2024-25.

Syllabus

(As given in the curriculum)

- Probability, Random Variables & Probability Distributions.
- Sampling, analysis of sample data-Empirical Distributions, Sampling from a Population Estimation, confidence intervals, point estimation–Maximum Likelihood, Probability mass functions, Modeling distributions, Hypothesis testing- Z, t, Chi-Square.
- ANOVA & Designs of Experiments - Single, Two factor ANOVA, Factorials ANOVA models.
- Linear least squares, Correlation & Regression Models-linear regression methods, Ridge regression, LASSO, univariate and Multivariate Linear Regression, probabilistic interpretation, Regularization, Logistic regression, locally weighted regression.
- Exploratory data analysis, Time series analysis, Analytical methods – ARIMA and SARIMA.

Evaluations: A Tentative Timeline

- Best two marks out of three quizzes (Total = 20 marks)
- Quiz-1 (10 marks): (January First week)
- Quiz-2 (10 marks):(March First week)
- Quiz-3 (10 marks):(April First week)
- Assignments (Total = 30 marks)
- Assignment-1 (10 marks):(Submission: End of January)
- Assignment-2 (10 marks):(Submission: End of March)

Course Introduction

- Project Review - 1 (10 marks):(February second week)
- Mid Sem (Total = 20 marks)
- Mid-Semester Exam (20 marks):(Feb first week, as per Academic calender)
- End Sem (Total = 30 marks)
- End-Semester Project Presentation (20 marks):(April second week, as per Academic calender) \end{itemize}$

Contact: ms_abhijith@cb.amrita.edu

1 Descriptive statistics

- Descriptive statistics deals with methods to describe and summarize data.
- Describing of data is effectively done through tables or graphs. Those often reveal important features such as the range, the degree of concentration, and the symmetry of the data.
- The summary of data is expressed through numerical quantities (summary statistics) whose values are determined by the data.

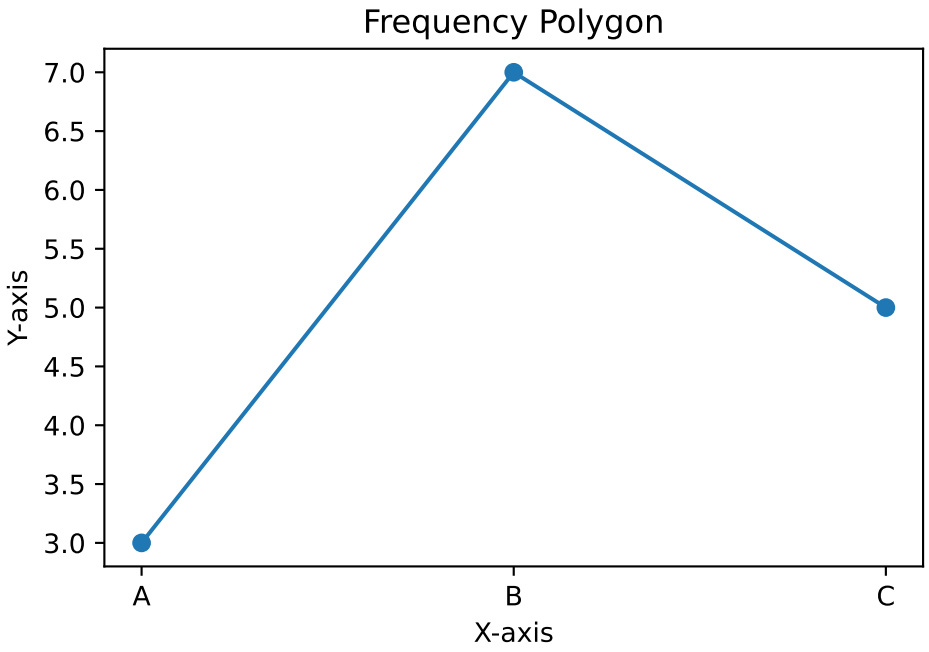
1.0.1 Describing Data sets

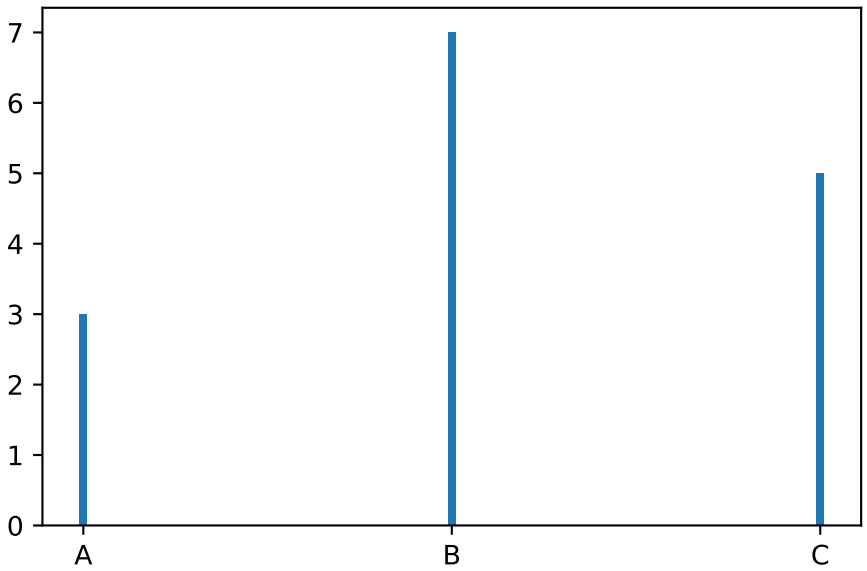
1.0.2 Frequency Tables and Graphs

- A data set having a relatively small number of distinct values can be conveniently presented in a frequency table.
- Data from a frequency table can be graphically represented by:
  - Line Graph
  - Bar Graph
  - Frequency Polygon

```
import numpy as np
import matplotlib.pyplot as plt

# Sample data
data = np.array(['A', 'B', 'C', 'A', 'A', 'B','B','B','B','B','B','C','C','C','C'])

# Calculate frequencies
values, frequencies = np.unique(data, return_counts=True)

# Line Graph
plt.plot(values, frequencies, marker='o')
plt.title('Frequency Polygon')
plt.xlabel('X-axis')
plt.ylabel('Y-axis')
```

```
plt.show()


plt.bar(values, frequencies, width=0.02)
plt.show()
```





1.0.3 Relative Frequency Tables and Graphs

- Consider a data set consisting of n values. If f is the frequency of a particular value, then the ratio f /n is called its relative frequency.

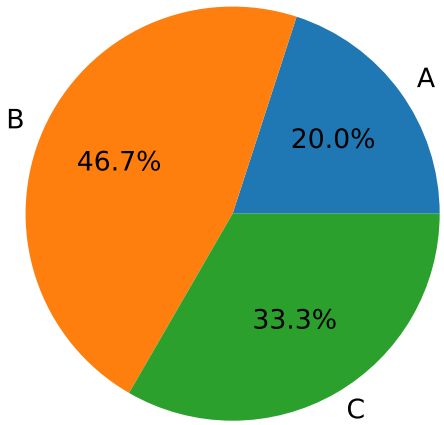
- That is, the relative frequency of a data value is the proportion of the data that have that value.

- The relative frequencies can be represented graphically by:

  - relative frequency line
  - relative frequency bar graph
  - relative frequency polygon
  - pie chart: A pie chart is often used to indicate relative frequencies when the data are not numerical in nature. A circle is constructed and then sliced into different sectors with areas proportional to the respective relative frequencies.

```
import numpy as np
import matplotlib.pyplot as plt

# Sample data
data = np.array(['A', 'B', 'C', 'A', 'A', 'B','B','B','B','B','B','C','C','C','C'])

# Calculate frequencies
values, frequencies = np.unique(data, return_counts=True)

relative_frequencies = frequencies/len(data)
print(relative_frequencies)


plt.pie(relative_frequencies, labels = values, autopct='%1.1f%%')
plt.show()
```

```
[0.2        0.46666667 0.33333333]
```

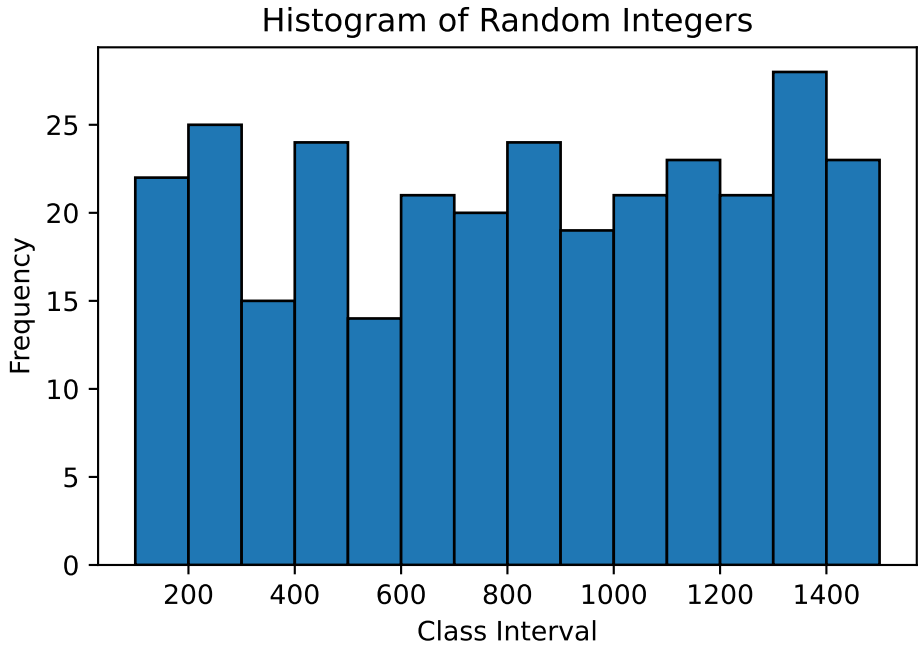


1.0.4 Grouped Data, Histograms, Ogives, and Stem and Leaf Plots

- For some data sets the number of distinct values is too large to utilize frequency tables.
- Instead, in such cases, it is useful to divide the values into groupings, or class intervals, and then plot the number of data values falling in each class interval.
- The number of class intervals chosen should be a trade-off between:
  - choosing too few classes at a cost of losing too much information about the actual data values in a class.
  - choosing too many classes, which will result in the frequencies of each class being too small.
- It is common, although not essential, to choose class intervals of equal length.
- The endpoints of a class interval are called the class boundaries.
- We will adopt the left-end inclusion convention, which stipulates that a class interval contains its left-end but not its right-end boundary point.
- Thus, for instance, the class interval 20-30 contains all values that are both greater than or equal to 20 and less than 30.
- A bar graph plot of class data, with the bars placed adjacent to each other, is called a histogram.
- The vertical axis of a histogram can represent either the class frequency or the relative class frequency.

```
import numpy as np
import matplotlib.pyplot as plt

# Set seed for reproducibility
np.random.seed(50)

# Generate 300 random integers between 100 and 1500
random_integers = np.random.randint(100, 1501, size=300)

# Create bins for the class intervals
bins = np.arange(100, 1600, 100)
print(bins)

# Plot histogram
plt.hist(random_integers, bins=bins, edgecolor='black')
plt.title('Histogram of Random Integers')
plt.xlabel('Class Interval')
plt.ylabel('Frequency')
plt.show()
```

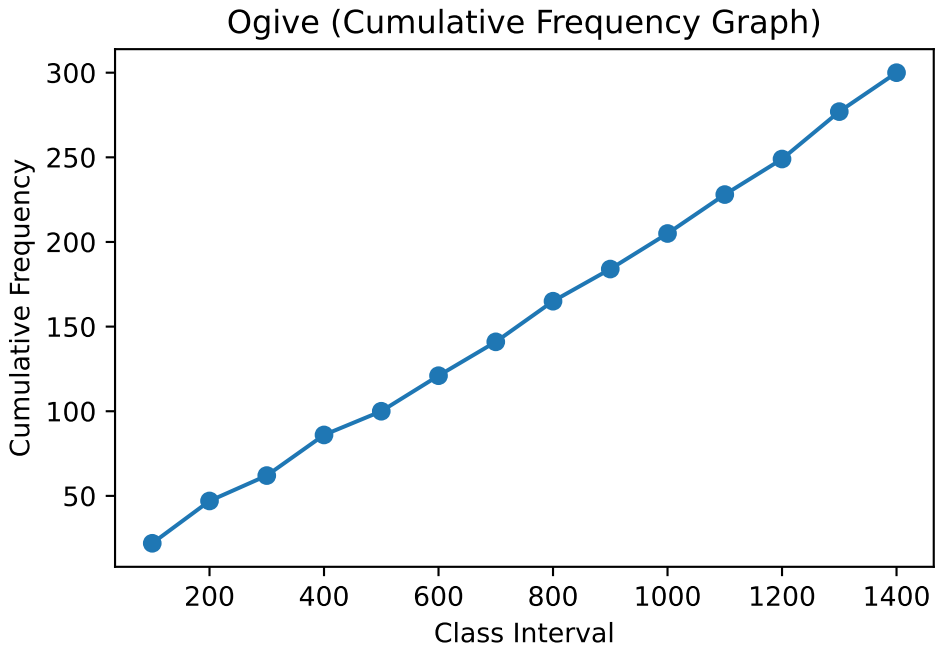
```
[ 100  200  300  400  500  600  700  800  900 1000 1100 1200 1300 1400
 1500]
```



- We are sometimes interested in plotting a cumulative frequency (or cumulative relative frequency) graph.
- A point on the horizontal axis of such a graph represents a possible data value; its corresponding vertical plot gives the number (or proportion) of the data whose values are less than or equal to it.
- A cumulative frequency plot is called an ogive.

```
import numpy as np
import matplotlib.pyplot as plt

# Set seed for reproducibility
np.random.seed(50)

# Generate 300 random integers between 100 and 1500
random_integers = np.random.randint(100, 1501, size=300)

# Create bins for the class intervals
bins = np.arange(100, 1600, 100)
print(bins)

histograms = np.histogram(random_integers, bins=bins)[0]
print(histograms)
cumulativeSum = np.cumsum(histograms)
print(cumulativeSum)


plt.plot(bins[:-1], cumulativeSum, marker='o', linestyle='-')
plt.title('Ogive (Cumulative Frequency Graph)')
plt.xlabel('Class Interval')
plt.ylabel('Cumulative Frequency')
#plt.grid(True)
plt.show()
```

```
[ 100  200  300  400  500  600  700  800  900 1000 1100 1200 1300 1400
 1500]
[22 25 15 24 14 21 20 24 19 21 23 21 28 23]
[ 22  47  62  86 100 121 141 165 184 205 228 249 277 300]
```



- An efficient way of organizing a small-to moderate-sized data set is to utilize a stem and leaf plot.
- Such a plot is obtained by first dividing each data value into two parts - its stem and its leaf.
- For instance, if the data are all two-digit numbers, then we could let the stem part of a data value be its tens digit and let the leaf be its ones digit.
- Thus, for instance, the value 62 is expressed as

```
import numpy as np

# Sample data
data = np.array([62, 67, 63, 68, 69, 61, 64, 65, 66, 60, 75, 74, 76, 78, 90, 92, 34, 36, 56, 57, 45, 53, 52, 59, 73, 74,

# Create stem and leaf plot
stem_leaf = {}

for number in data:
    stem = number // 10
    leaf = number % 10
    if stem in stem_leaf:
        stem_leaf[stem].append(leaf)
    else:
```

```
        stem_leaf[stem] = [leaf]

# Print stem and leaf plot
for stem, leaves in sorted(stem_leaf.items()):
    print(f"{stem} | {' '.join(map(str, sorted(leaves)))}")
```

```
2 | 0
3 | 4 4 6 6
4 | 5
5 | 2 3 6 7 9
6 | 0 1 2 3 4 5 6 7 8 9
7 | 3 4 4 5 6 8 9
8 | 0 1 5
9 | 0 2
```

2 Parameter Estimation

2.1 Point Estimation: Maximum Likelihood Estimators

(Please refer the book titled "Introduction to Probability and Statistics for Engineers and Scientists" by Sheldon M Ross for more details)

2.2 Interval Estimates

- Consider a sample $X_1, X_2, \ldots, X_n$ drawn from a known distribution with an unknown mean μ.
- It is established that the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ serves as the maximum likelihood estimator for μ.
- However, the sample mean $\bar{X}$ is not expected to be exactly equal to μ, but rather close to it.
- Therefore, instead of providing a single point estimate, it is often more useful to specify an interval within which we are confident that μ lies.
- To determine such an interval estimator, we utilize the probability distribution of the point estimator.

2.2.1 Confidence Intervals for the Mean of a normal population with known Variance

- Consider a sample $X_1, X_2, \ldots, X_n$ drawn from a normal distribution with an unknown mean μ and a known variance σ^2.
- The point estimator $\bar{X}$ is normal with mean μ and variance σ^2/n.
- Therefore, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows a standard normal distribution.

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.9750 - 0.0250 = 0.95$$

- Rewriting the above equation:

What to do

Consider that I want to find an interval around $\bar{X}$ such that the actual population mean μ falls within the interval, say 95 % of the times.

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(1.96\frac{\sigma}{\sqrt{n}} > \mu - \bar{X} > -1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} < \mu - \bar{X} < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- We have P(Z < -1.96) = 0.025, similarly P(Z > 1.96) = 0.025. Usually 1.96 is represented generally as $z_{0.025}$. Thus, P(Z < -$z_{0.025}$) = 0.025 and P(Z > $z_{0.025}$) = 0.025.
- Hence, 100(1-0.05) percent confidence interval for the mean of a normal population with known variance is:

$$P\left(\bar{X} - z_{0.025}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{0.025}\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{X} - z_{0.05/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{0.05/2}\frac{\sigma}{\sqrt{n}}\right) = (1 - 0.05)$$

- For a confidence level of $100(1-\alpha)$ percent, the corresponding critical value from the standard normal distribution is $z_{\alpha/2}$.

- The $100(1-\alpha)$ percent confidence interval for μ is given by:

$$\mu \in \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{2.1}$$

- The interval as given in Equation 2.1 is called a two-sided confidence interval.
- Also the term $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is called the margin of error.

Derivation of two-sided confidence interval

- To find 100(1-α) percent confidence interval of mean (μ), we have;

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha}/2\right) = 1-\alpha$$

- Doing the same manipulations we did earlier for obtaining the 95percent confidence interval we can obtain:

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

- The above equation give us the required confidence interval, as given in Equation 2.1.

What if !?

What if we are interested in one sided confidence intervals !!?

One-sided Upper Confidence Iterval

- To determine such an interval, for a standard normal random variable Z, we have;

$$P\left(Z < 1.645\right) = 0.95$$

- Thus,

$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.645\right) = 0.95$$

$$P\left(\mu - \bar{X} > -1.645\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\mu > \bar{X} - 1.645\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- Thus a 95 percent one-sided upper confidence interval for μ is

$$\mu \in \left(\bar{X} - 1.645\frac{\sigma}{\sqrt{n}}, \infty\right)$$

or in other words; 100(1-0.05) percent one-sided upper confidence interval for μ is

$$\mu \in \left(\bar{X} - z_{0.05}\frac{\sigma}{\sqrt{n}}, \infty\right)$$

Oneside interval!

Can you think of another one sided confidence interval?

One-sided lower confidence interval

- We have

$$P\left(Z > -1.645\right) = 0.95$$

- Proceed just like in the previous case and you will find a 100(1-0.05) percent one-sided lower confidence interval for μ as;

$$\mu \in \left(-\infty, \bar{X} + z_{0.05}\frac{\sigma}{\sqrt{n}}\right)$$

- In general, 100(1-α) percent one-sided upper confidence interval for μ is given in Equation 2.2.

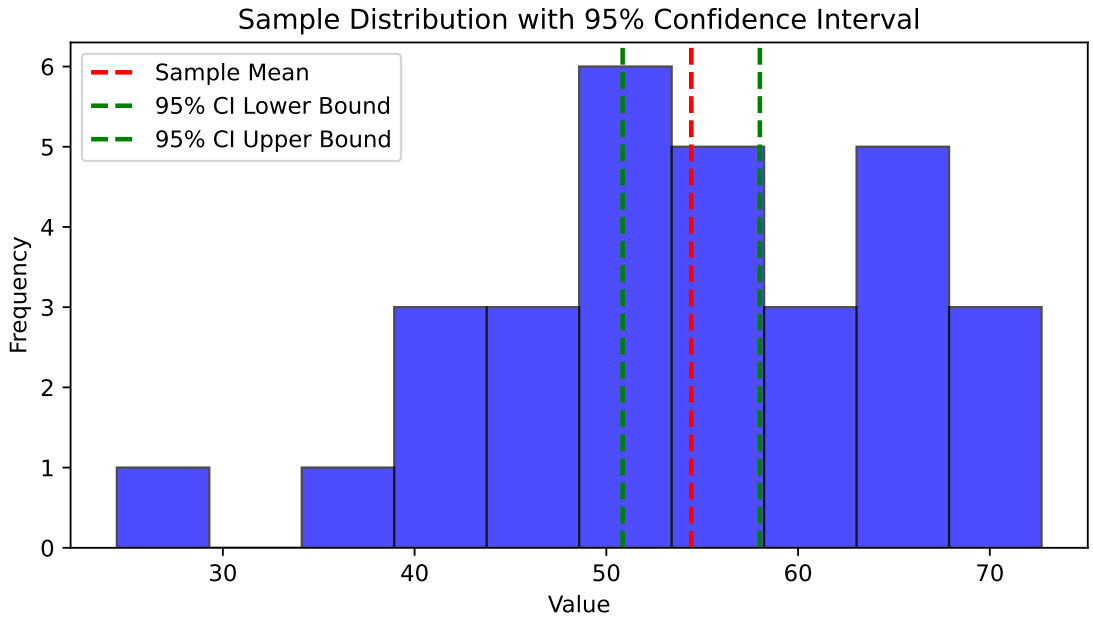
$$\mu \in \left(\bar{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right) \tag{2.2}$$

- Also, 100(1-α)percent one-sided lower confidence interval for μ is given in Equation 2.3.

$$\mu \in \left(-\infty, \bar{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right) \tag{2.3}$$

- The python code below creates a sample and find 95% confidence interval for the mean if the population standard deviation is assumed to be 10. Other values are specified in the code.

```
import numpy as np
import matplotlib.pyplot as plt

# Parameters
mu = 50  # true mean
sigma = 10  # known standard deviation
n = 30  # sample size
alpha = 0.05  # significance level

# Generate a sample
np.random.seed(0)
sample = np.random.normal(mu, sigma, n)
sample_mean = np.mean(sample)

# Calculate the confidence interval
z = 1.96  # z-value for 95% confidence
margin_of_error = z * (sigma / np.sqrt(n))
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_of_error)

# Plot the sample and confidence interval
plt.figure(figsize=(8, 4))
plt.hist(sample, bins=10, alpha=0.7, color='blue', edgecolor='black')
plt.axvline(sample_mean, color='red', linestyle='dashed', linewidth=2, label='Sample Mean')
plt.axvline(confidence_interval[0], color='green', linestyle='dashed', linewidth=2, label='95% CI Lower Bound')
plt.axvline(confidence_interval[1], color='green', linestyle='dashed', linewidth=2, label='95% CI Upper Bound')
plt.title('Sample Distribution with 95% Confidence Interval')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.legend()
plt.show()

print(f"Sample Mean: {sample_mean}")
print(f"95% Confidence Interval: {confidence_interval}")
```



Sample Mean: 54.42856447263174
95% Confidence Interval: (50.85011043026466, 58.007018514998826)

Problem

Suppose that when a signal having value μ is transmitted from location A the value received at location B is normally distributed with mean μ and variance 4. That is, if μ is sent, then the value received is μ + N where N, representing noise, is normal with mean 0 and variance 4. To reduce error, suppose the same value is sent 9 times. If the successive values received are 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5;
(a). construct a 95 percent two-sided confidence interval for μ.
(b). construct 95 percent one-sided upper and lower confidence intervals for μ.

Problem

Suppose a quality control manager at a factory wants to ensure that the average weight of a product is at least 500 grams. They take a random sample of 30 products and find the sample mean weight to be 495 grams with a standard deviation of 10 grams. Help the manager to estimate the minimum average weight of the products with 95% confidence.

2.2.2 Confidence Intervals for the Mean of a normal population with unknown Variance

- If you recollect the discussion we had about the sample mean from a normal population with unknown variance we saw that variable t_{n-1} given by:

$$t_{n-1} = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

has a t-distribution with n-1 degrees of freedom.

- Because of the symmetry of the t-distribution we can write for any $\alpha \in (0, 1/2)$;

$$P\left(-t_{\alpha/2,n-1} < \sqrt{n}\frac{\bar{X} - \mu}{S} < t_{\alpha/2,n-1}\right) = 1 - \alpha$$

$$P\left(-\bar{X} - t_{\alpha/2,n-1}\frac{\sqrt{n}}{S} < -\mu < -\bar{X} + t_{\alpha/2,n-1}\frac{\sqrt{n}}{S}\right) = 1 - \alpha$$

$$P\left(\bar{X} + t_{\alpha/2,n-1}\frac{\sqrt{n}}{S} > \mu > \bar{X} - t_{\alpha/2,n-1}\frac{\sqrt{n}}{S}\right) = 1 - \alpha$$

$$P\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

- If the sample mean is $\bar{X}$ and sample standard deviation S, then we can say that with 100(1-α) percent confidence that

$$\mu \in \left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)$$

- In this case 100(1-α) percent one-sided upper confidence interval can be obtained from the fact that:

$$P\left(\sqrt{n}\frac{(\bar{X} - \mu)}{S} < t_{\alpha,n-1}\right) = 1 - \alpha$$

$$P\left(\mu > \bar{X} - \frac{S}{\sqrt{n}}t_{\alpha,n-1}\right) = 1 - \alpha$$

- Thus $100(1 - \alpha)$ percent one-sided upper confidence interval for the mean in this case is given by;

$$\mu \in \left(\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha,n-1}, \infty \right)$$

- Thus $100(1-\alpha)$ percent one-sided lower confidence interval for the mean in this case is given by;

$$\mu \in \left(-\infty, \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha,n-1} \right)$$

Problem

Let us again consider the previous problem but let us now suppose that when the value μ is transmitted at location A then the value received at location B is normal with mean μ and variance σ^2 but with σ^2 being unknown. If 9 successive values are, 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, and 10.5, compute a 95 percent confidence interval for μ.

2.2.3 Confidence Intervals for the Variance of a Normal Distribution

- If we are sampling from a normal distribution with unknown mean and unknown variance then;

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

follows a chi-squared distribution.

- We have

$$P\left(\chi^2_{1-\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1} \right) = 1-\alpha$$

$$P\left(\chi^2_{1-\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1} \right) = 1-\alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \right) = 1-\alpha$$

- Hence, 100(1-α) percent two-sided confidence interval for the variance in this case;

$$\sigma^2 \in \left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \right)$$

- The 100(1-α) percent one-sided upper and lower confidence intervals in this case will be respectively;

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}, \infty \right)$$

and

$$\left(0, \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}} \right)$$

Problem

A standardized procedure is expected to produce washers with very small deviation in their thicknesses. Suppose that 10 such washers were chosen and measured. If the thicknesses of these washers were, in inches; .123, .133, .124, .125, .126, .128, .120, .124, .130, and .126. What is a 90 percent confidence interval for the standard deviation of the thickness of a washer produced by this procedure?

All problems and most part of text are taken from Ross (2009) .

3 Hypothesis Testing

Problem

A quality control manager at a factory wants to ensure that the average weight of a product is at least 500 grams. They take a random sample of 30 products and find the sample mean weight to be 495 grams with a standard deviation of 10 grams. The manager wants to test if the average weight of the products is significantly less than 500 grams at a 5% significance level.

- Null Hypothesis (H_0): The average weight of the products is at least 500 grams ($\mu \geq 500$).
- Alternative Hypothesis (H_1): The average weight of the products is less than 500 grams ($\mu < 500$).

Examples Highlighting the Need for Hypothesis Testing

1. Medical Research:
   - Scenario: A pharmaceutical company develops a new drug intended to lower blood pressure.
   - Hypothesis Testing: The null hypothesis (H_0) might state that the new drug has no effect on blood pressure, while the alternative hypothesis (H_1) states that the drug does lower blood pressure. Hypothesis testing helps determine if the observed effects in clinical trials are statistically significant or if they could have occurred by random chance.
2. Quality Control:
   - Scenario: A factory produces light bulbs, and the quality control team wants to ensure that the average lifespan of the bulbs is 1000 hours.
   - Hypothesis Testing: The null hypothesis (H_0) could be that the mean lifespan of the bulbs is 1000 hours. The alternative hypothesis (H_1) might be that the mean lifespan is not 1000 hours. Hypothesis testing helps the team decide whether to accept the production process or take corrective actions.
3. Marketing:

- Scenario: A company launches a new advertising campaign and wants to know if it has increased sales.
- Hypothesis Testing: The null hypothesis (H_0) might state that the advertising campaign has no effect on sales, while the alternative hypothesis (H_1) states that the campaign has increased sales. Hypothesis testing helps the company determine if the increase in sales is statistically significant.

4. Education:

   - Scenario: An educator wants to test if a new teaching method is more effective than the traditional method.
   - Hypothesis Testing: The null hypothesis (H_0) could be that there is no difference in effectiveness between the new and traditional methods. The alternative hypothesis (H_1) might be that the new method is more effective. Hypothesis testing helps in making data-driven decisions about adopting new teaching strategies.

5. Environmental Science:

   - Scenario: Researchers want to determine if a new policy has reduced pollution levels in a city.
   - Hypothesis Testing: The null hypothesis (H_0) might state that the policy has no effect on pollution levels, while the alternative hypothesis (H_1) states that the policy has reduced pollution levels. Hypothesis testing helps in evaluating the effectiveness of environmental policies.

These examples illustrate how hypothesis testing is a crucial tool in various fields for making informed decisions based on data.

Important Terminology

- Null Hypothesis (H_0): The hypothesis that there is no effect or no difference. It is the default assumption that any observed effect is due to random chance. It is the hypothesis that researchers aim to test against.
- Alternative Hypothesis (H_1 or H_a): The hypothesis that there is an effect or a difference. It is what researchers want to prove.
- Test Statistic: A standardized value that is calculated from sample data during a hypothesis test. It is used to decide whether to reject the null hypothesis.
- P-value: The probability of obtaining test results at least as extreme as the observed results, assuming that the null hypothesis is true. A smaller p-value

indicates stronger evidence against the null hypothesis.

- Significance Level (α): A threshold set by the researcher which the p-value must be below in order to reject the null hypothesis. Common significance levels are 0.05, 0.01, and 0.10.
- Critical Value: The value that the test statistic must exceed in order to reject the null hypothesis. It is determined based on the significance level and the distribution of the test statistic.
- Power of a Test: The probability that the test correctly rejects a false null hypothesis (i.e., it does not make a type II error). Higher power indicates a greater ability to detect an effect when there is one.
- Type I Error: The error made when the null hypothesis is true, but is incorrectly rejected. The probability of making a type I error is denoted by α.
- Type II Error: The error made when the null hypothesis is false, but is incorrectly accepted. The probability of making a type II error is denoted by β.
- Confidence Interval: A range of values derived from the sample data that is likely to contain the population parameter. It provides an estimate of the parameter with a certain level of confidence (e.g., 95%).
- One-tailed Test: A hypothesis test in which the region of rejection is on only one side of the sampling distribution. It tests for the possibility of the relationship in one direction.
- Two-tailed Test: A hypothesis test in which the region of rejection is on both sides of the sampling distribution. It tests for the possibility of the relationship in both directions.

Introduction

- A statistical hypothesis is typically a statement regarding a set of parameters of a population distribution.
- It is termed a hypothesis because its true value is unknown.
- The main challenge is to devise a method to determine whether the values of a random sample from this population align with the hypothesis.
- Consider a population with distribution F_θ, where θ is unknown.
- We aim to test a specific hypothesis about θ.

- This hypothesis is denoted by H_0 and is referred to as the null hypothesis.
- For instance, if F_θ is a normal distribution function with mean θ and variance equal to 1, two possible null hypotheses about θ are:

$$H_0 : \theta = 1$$

$$H_0 : \theta > 1$$

$$H_0 : \theta \leq 1$$

- It is important to note that the null hypothesis in the first case fully specifies the population distribution.
- Whereas the null hypothesis in the second and third cases do not.

> **Simple and Composite Hypotheses**
>
> - A hypothesis that fully specifies the population distribution when true is known as a simple hypothesis. (Eg; $H_0 : \theta = 1$)
> - A hypothesis that does not fully specifies the population distribution is referred to as a composite hypothesis. (Eg; $H_0 : \theta > 1$, H_{0}: 1)

Testing a Null Hypothesis

- To test a specific null hypothesis H_0, we take a sample of size n from the population, say $X_1, X_2, \ldots, X_n$.
- Based on these n values, we decide whether to accept or reject H_0.
- We define a region C in the n-dimensional space. This region is called the critical region.
- If the sample $X_1, X_2, \ldots, X_n$ falls within the critical region C, we reject H_0. Otherwise, we accept H_0.
- In simple terms, the critical region C helps us determine the outcome of the statistical test.

$$accepts \ H_0 \ if \ (X_1, X_2, ..., X_n) \notin C$$

and

$$rejects \ H_0 \ if \ (X_1, X_2, ..., X_n) \in C$$

Types of Errors in Hypothesis Testing

- When developing a procedure for testing a given null hypothesis H_0, it is crucial to recognize that two different types of errors can occur.
- A type I error occurs if the test incorrectly rejects H_0 when it is actually true.
- A type II error occurs if the test incorrectly accepts H_0 when it is actually false.

Note

The goal of a statistical test for H_0 is not to definitively determine its truth but to assess if the data is consistent with H_0.

Significance Level and Classical Approach

- H_0 should be rejected only if the observed data is highly unlikely under H_0.
- The classical method involves specifying a value α, known as the level of significance.
- The test is designed so that the probability of rejecting H_0 when it is true does not exceed α.
- Common choices for α are 0.1, 0.05, and 0.005.
- This approach ensures that the probability of a type I error (incorrectly rejecting H_0) is controlled and does not exceed the chosen α.

Example

- For instance, consider testing the hypothesis that the mean of a normal distribution with parameters $(\theta, 1)$ is equal to 1.
- The test rejects the null hypothesis if the point estimate of θ (i.e., the sample mean) deviates more than $\frac{1.96}{\sqrt{n}}$ from 1.

- As we will discuss in the next section, the value $\frac{1.96}{\sqrt{n}}$ is selected to achieve a significance level of $\alpha = 0.05$.

4 Hypothesis Tests Concerning the mean of a normal population

4.1 With known Variance (Z-test)

- Let $X_1, X_2, \ldots, X_n$ be a sample of size n from a normal distribution with an unknown mean μ and a known variance σ^2.
- We are interested in testing the null hypothesis:

$$H_0 : \mu = \mu_0$$

- Against the alternative hypothesis:

$$H_1 : \mu \neq \mu_0$$

- Where μ_0 is a specified constant.
- Since $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a natural point estimator of μ, it is reasonable to accept H_0 if $\bar{X}$ is not too far from μ_0.
- Thus, the critical region of the test would be of the form:

$$C = \{X_1, \ldots, X_n : |\bar{X} - \mu_0| > c\}$$

for some suitably chosen value c.

- To ensure that the test has a significance level α, we must determine the critical value c in the above equation such that the type I error is equal to α. This means c must satisfy:

$$P_{\mu_0}\{|\bar{X} - \mu_0| > c\} = \alpha$$

where P_{μ_0} denotes that the probability is computed under the assumption that population mean, $\mu = \mu_0$.

- When $\mu = \mu_0$, $\bar{X}$ follows a normal distribution with mean μ_0 and variance $\frac{\sigma^2}{n}$. Therefore, the standardized variable Z defined by:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

will have a standard normal distribution.

- The probability of a type I error is given by:

$$P\left(|\bar{X} - \mu_0| > c\right) = \alpha$$

- Equivalently, this can be written as:

$$2P\left(Z > \frac{c\sqrt{n}}{\sigma}\right) = \alpha$$

- Where Z is a standard normal random variable. We know that:

$$P\left(Z > z_{\alpha/2}\right) = \frac{\alpha}{2}$$

- Therefore, we have:

$$\frac{c\sqrt{n}}{\sigma} = z_{\alpha/2}$$

- Solving for c, we get:

$$c = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

- Thus, the test at significance level α is to reject H_0 if:

$$|\bar{X} - \mu_0| > \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

- And accept H_0 otherwise. Equivalently, we can reject H_0 if:

$$\sqrt{n}\frac{|\bar{X} - \mu_0|}{\sigma} > z_{\alpha/2}$$

- And accept H_0 if:

$$\sqrt{n}\frac{|\bar{X}-\mu_0|}{\sigma} \leq z_{\alpha/2}$$

Problem

If a signal of value μ is sent from location A, then the value received at location B is normally distributed with mean μ and standard deviation 2. That is, the random noise added to the signal is an N(0, 4) random variable. There is reason for the people at location B to suspect that the signal value $\mu = 8$ will be sent today. Test this hypothesis if the same signal value is independently sent five times and the average value received at location B is X = 9. 5.

4.1.1 Choosing the Significance Level

- The appropriate significance level α depends on the specific context and consequences of the hypothesis test.
- If rejecting the null hypothesis H_0 would lead to significant costs or consequences, a more conservative significance level (e.g., 0.05 or 0.01) should be chosen.
- If there is a strong initial belief that H_0 is true, strict evidence is required to reject H_0, implying a lower significance level.
- The test can be described as follows: For an observed value of the test statistic $\sqrt{n}\frac{|\bar{X}-\mu_0|}{\sigma}$, denoted as v, reject H_0 if the probability of the test statistic being as large as v under H_0 is less than or equal to α.
- This probability is known as the p-value of the test. H_0 is accepted if α is less than the p-value and rejected if α is greater than or equal to the p-value.
- In practice, the significance level is sometimes not set in advance. Instead, the p-value is calculated from the data, and decisions are made based on the p-value.
- If the p-value is much larger than any reasonable significance level, H_0 is accepted. Conversely, if the p-value is very small, H_0 is rejected.

4.1.2 Hypothesis Testing Summary: Z- Test

Table 4.1: Caption: Summary of hypothesis testing for a sample from a $N(\mu, \sigma^2)$ population with known σ^2.

Sample and Population	Details
Sample	$\{X_1, X_2, ..., X_n\}$

Sample and Population	Details
Population	$N(\mu, \sigma^2)$
Known Parameter	σ^2
Sample Mean	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Significance Level	α

Hypothesis	Test Statistic (TS)	Reject if	p-Value if TS = t
$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/\sigma$	$\lvert TS\rvert > z_{\alpha/2}$	$2P\{Z \geq \lvert t\rvert\}$
$H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/\sigma$	$TS > z_\alpha$	$P\{Z \geq t\}$
$H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/\sigma$	$TS < -z_\alpha$	$P\{Z \leq t\}$

Problem

Imagine you're the quality control manager at a company that prides itself on the precision of its product weights. The company claims that the average weight of their product is exactly 100 grams. But, as a diligent manager, you decide to put this claim to the test. You randomly select a sample of 30 products and measure their weights. To your surprise, the average weight of your sample is 110 grams! Now, you need to determine if this difference is statistically significant or just a fluke. Assume the population standard deviation as 15 grams.

```
import numpy as np
from scipy import stats

# Given data
sample_mean = 495
population_mean = 500
std_dev = 10
sample_size = 30
alpha = 0.05

# Calculate the Z-score
z_score = (sample_mean - population_mean) / (std_dev / np.sqrt(sample_size))

# Calculate the p-value
p_value = stats.norm.cdf(z_score)
```

```
# Determine if we reject the null hypothesis
reject_null = p_value < alpha

# Output the results
print(f"Z-score: {z_score}")
print(f"P-value: {p_value}")
print(f"Reject the null hypothesis: {reject_null}")
```

Z-score: -2.7386127875258306
P-value: 0.00308494966027208
Reject the null hypothesis: True

4.2 With unknown Variance (T-test)

- Let $X_1, X_2, \ldots, X_n$ be a sample of size n from a normal distribution with an unknown mean μ and a unknown variance.
- Say, we are interested in testing the null hypothesis:

$$H_0 : \mu = \mu_0$$

- Against the alternative hypothesis:

$$H_1 : \mu \neq \mu_0$$

- Where μ_0 is a specified constant.
- In the previous case (with known variance), for a significance level (α) we accepted the null hypothesis if:

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \leq z_{\alpha/2}$$

- But in this case, σ is unknown.
- We know that the statistic, T, as given below has a t-distribution with n-1 degrees of freedom when $\mu = \mu_0$.

$$T = \frac{\bar{X} - \mu_0}{S\sqrt{n}}$$

where S is the sample standard deviation.

- Hence here with H_0: $\mu = \mu_0$ and H_1; $\mu \neq \mu_0$; analogous to the Z-test here in T-test we can:

two-sided t-test

- reject the null hypothesis (H_0) if:

$$\left|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}\right| > t_{\alpha/2}$$

- accept H_0 if:

$$\left|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}\right| \leq t_{\alpha/2}$$

4.2.1 Hypothesis Testing Summary: T- Test

Table 4.3: Caption: Summary of hypothesis testing for a sample from a $N(\mu, \sigma^2)$ population with unknown σ^2.

Sample and Population	Details
Sample	$\{X_1, X_2, ..., X_n\}$
Population	$N(\mu, \sigma^2)$
Sample Mean	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Sample Variance	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
Significance Level	α

Hypothesis	Test Statistic (TS)	Reject if	p-Value if TS = t
$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/S$	$\lvert TS\rvert > t_{\alpha/2,n-1}$	$2P\{T_{n-1} \geq \lvert t\rvert\}$
$H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/S$	$TS > t_{\alpha,n-1}$	$P\{T_{n-1} \geq t\}$
$H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/S$	$TS < -t_{\alpha,n-1}$	$P\{T_{n-1} \leq t\}$

T_{n-1} is a t-random variable with (n - 1) degrees of freedom: $P(T_{n-1} > t_{\alpha,n-1}) = \alpha$.

Problem

A public health official claims that the mean home water use is at most 350 gallons a day. To verify this claim, a study of 20 randomly selected homes was instigated with the result that the average daily water uses of these 20 homes were as follows:
340 344 362 375 356 386 354 364 332 402 340 355 362 322 372 324 318 360 338 370
Do the data contradict the official's claim?

```
import numpy as np
from scipy import stats

# Given data
data = [340, 344, 362, 375, 356, 386, 354, 364, 332, 402, 340, 355, 362, 322, 372, 324, 318, 360, 338, 370]
sample_mean = np.mean(data)
sample_std = np.std(data, ddof=1)
sample_size = len(data)
population_mean = 350
alpha = 0.05

# Calculate the T-score
t_score = (sample_mean - population_mean) / (sample_std / np.sqrt(sample_size))

# Calculate the p-value
p_value = 2 * (1 - stats.t.cdf(np.abs(t_score), df=sample_size-1))

# Determine if we reject the null hypothesis
reject_null = p_value < alpha

# Output the results
print(f"Sample Mean: {sample_mean}")
print(f"Sample Standard Deviation: {sample_std}")
print(f"T-score: {t_score}")
print(f"P-value: {p_value}")
print(f"Reject the null hypothesis: {reject_null}")
```

```
Sample Mean: 353.8
Sample Standard Deviation: 21.847798877449275
T-score: 0.7778411328447066
P-value: 0.4462410900531899
Reject the null hypothesis: False
```

5 Hypothesis Tests Concerning the variance of a normal population

- Let $X_1, X_2, \ldots, X_n$ be a sample of size n from a normal distribution with an unknown mean μ and variance σ^2.
- We are interested in testing the null hypothesis:

$$H_0 : \sigma^2 = \sigma_0^2$$

- Against the alternative hypothesis:

$$H_1 : \sigma^2 \neq \sigma_0^2$$

- Where σ_0^2 is a specified constant.
- We know from the discussion on sampling distribution, $\frac{(n-1)S^2}{\sigma_0^2}$ has a chi-squared distribution with (n-1) degrees of freedom.

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

- Also,

$$P_{H_0}\left(\chi^2_{1-\alpha/2,\, n-1} \leq \frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{\alpha/2,\, n-1}\right) = 1 - \alpha$$

- In this case the test statistic (TS) is:

$$TS = \frac{(n-1)S^2}{\sigma_0^2}$$

- The p-value for this case is:

$$p - value = 2\ \min\left(P(\chi^2_{n-1} < TS), 1 - P(\chi^2_{n-1} < TS)\right)$$

5.0.1 Hypothesis Testing Summary: chi-square Test

Table 5.1: Caption: Summary of hypothesis testing for a sample from a $N(\mu, \sigma^2)$ population.

Sample and Population	Details
Sample	$\{X_1, X_2, ..., X_n\}$
Population	$N(\mu, \sigma^2)$
Sample Mean	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Sample Variance	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
Significance Level	α

Hypothesis	Test Statistic (TS)	Reject if	p-Value if TS = t
$H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$	$\frac{(n-1)S^2}{\sigma_0^2}$	$TS \notin \left[\chi^2_{1-\alpha/2,\, n-1},\ \chi^2_{\alpha/2,\, n-1}\right]$	$2\ \min\left(P(\chi^2_{n-1} < t), 1 - P(\chi^2_{n-1} < t\right.$
$H_0 : \sigma^2 \leq \sigma_0^2$ vs $H_1 : \sigma^2 > \sigma_0^2$	$\frac{(n-1)S^2}{\sigma_0^2}$	$TS > \chi^2_{\alpha,\, n-1}$	$P\left(\chi^2_{n-1} \geq t\right)$
$H_0 : \sigma^2 \geq \sigma_0^2$ vs $H_1 : \sigma^2 < \sigma_0^2$	$\frac{(n-1)S^2}{\sigma_0^2}$	$TS < \chi^2_{1-\alpha,\, n-1}$	$P\left(\chi^2_{n-1} \leq t\right)$

Problem

A machine that automatically controls the amount of ribbon on a tape has recently been installed. This machine will be judged to be effective if the standard deviation σ of the amount of ribbon on a tape is less than .15 cm. If a sample of 20 tapes yields a sample variance of $S^2 = .025$ cm^2, are we justified in concluding that the machine is ineffective? Assume the level of significance as 0.05.

```
import numpy as np
from scipy.stats import chi2

# Given data
sample_variance = 0.025
sample_size = 20
population_variance = 0.15**2
alpha = 0.05
```

```
# Calculate the test statistic
test_statistic = (sample_size - 1) * sample_variance / population_variance

# Calculate the critical values
chi2_critical_low = chi2.ppf(alpha / 2, df=sample_size - 1)
chi2_critical_high = chi2.ppf(1 - alpha / 2, df=sample_size - 1)

# Calculate the p-value
p_value = 1 - chi2.cdf(test_statistic, df=sample_size - 1)

# Determine if we reject the null hypothesis
reject_null = test_statistic < chi2_critical_low or test_statistic > chi2_critical_high

# Output the results
print(f"Test Statistic: {test_statistic}")
print(f"Chi-square Critical Low: {chi2_critical_low}")
print(f"Chi-square Critical High: {chi2_critical_high}")
print(f"P-value: {p_value}")
print(f"Reject the null hypothesis: {reject_null}")
```

Test Statistic: 21.111111111111114
Chi-square Critical Low: 8.906516481987971
Chi-square Critical High: 32.85232686172969
P-value: 0.33069403418551535
Reject the null hypothesis: False

Additional Topic

Equality of Means of two normal populations

- Comparing the means of two different normal populations is common in hypothesis testing.
- Example scenarios include comparing average test scores of students from two schools or average lifespans of two brands of light bulbs.
- Use a two-sample z-test to test the equality of means of two normal populations with unequal variances.
- Null Hypothesis (H_0): The means of the two populations are equal.
- Alternative Hypothesis (H_1): The means of the two populations are not equal.
- Test statistic for the two-sample z-test:

$$z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_X}}} = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

- where,
  - $\bar{X}$ and $\bar{Y}$: Sample means
  - σ_X^2 and σ_Y^2: Population variances
  - n_X and n_Y: Sample sizes of the two groups
- Compare the calculated z-value to the critical z-value from the standard normal distribution table with the chosen significance level (α).
- If the calculated z-value is greater than the critical z-value, reject the null hypothesis and conclude a significant difference between the means of the two populations.

> **!** Important
>
> Additionally, refer to section 8.4, titled “TESTING THE EQUALITY OF MEANS OF TWO NORMAL POPULATIONS,” in “INTRODUCTION TO PROBABILITY AND STATISTICS FOR ENGINEERS AND SCIENTISTS” by Sheldon M. Ross.

6 Analysis of Variance (ANOVA)

6.1 Introduction

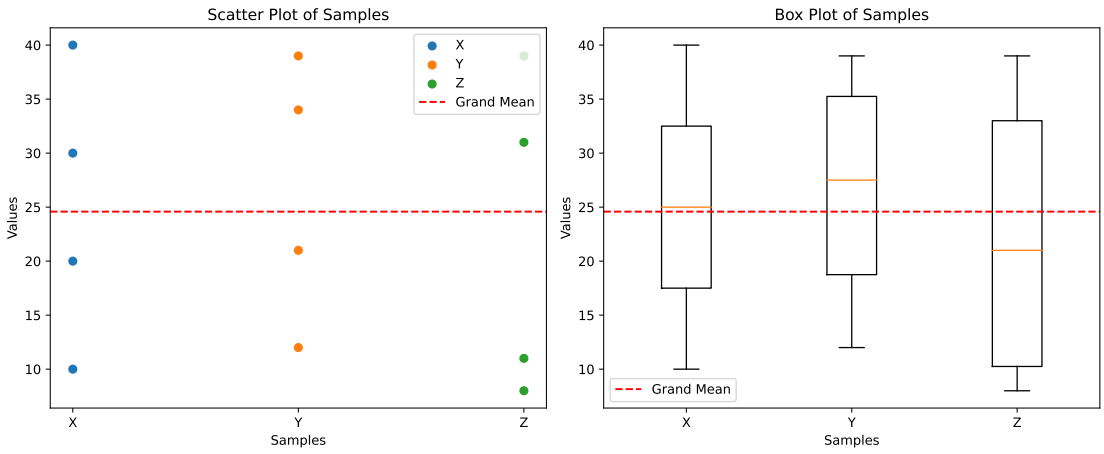
Here, we will perform an Analysis of Variance (ANOVA) to determine if there are any statistically significant differences between the means of three or more independent groups.

Comparison of means of three samples

- Consider three independent samples:
  - X: 10, 20, 30, 40
  - Y: 12, 21, 34, 39
  - Z: 8, 11, 31, 39
- We want to test if the means of these three samples are significantly different from each other.
- Calculate the mean of each sample:
  - Mean of X: $(10 + 20 + 30 + 40)/4 = 25$
  - Mean of Y: $(12 + 21 + 34 + 39)/4 = 26.5$
  - Mean of Z: $(8 + 11 + 31 + 39)/4 = 22.25$
- Calculate the overall mean (grand mean) of all the samples combined:
  - Grand Mean: $(10+20+30+40+12+21+34+39+8+11+31+39)/12 = 24.58$

```
import numpy as np
import matplotlib.pyplot as plt

# Data
X = np.array([10, 20, 30, 40])
Y = np.array([12, 21, 34, 39])
Z = np.array([8, 11, 31, 39])

# Grand mean
```

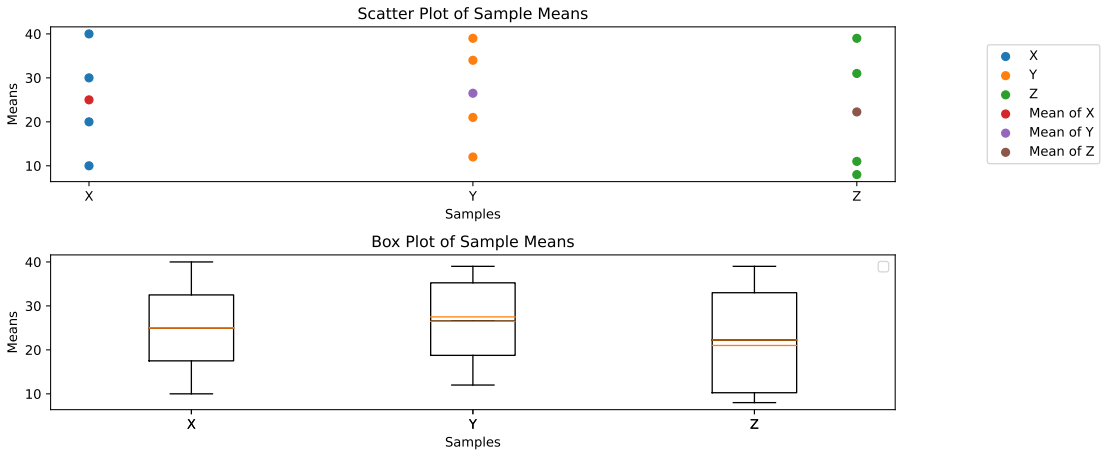
```
grand_mean = np.mean(np.concatenate([X, Y, Z]))

# Scatter plot
plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
plt.scatter(np.ones_like(X), X, label='X')
plt.scatter(np.ones_like(Y) * 2, Y, label='Y')
plt.scatter(np.ones_like(Z) * 3, Z, label='Z')
plt.xticks([1, 2, 3], ['X', 'Y', 'Z'])
plt.xlabel('Samples')
plt.ylabel('Values')
plt.title('Scatter Plot of Samples')
plt.axhline(grand_mean, color='red', linestyle='--', label='Grand Mean')
plt.legend()

# Box plot
plt.subplot(1, 2, 2)
plt.boxplot([X, Y, Z], labels=['X', 'Y', 'Z'])
plt.xlabel('Samples')
plt.ylabel('Values')
plt.title('Box Plot of Samples')
plt.axhline(grand_mean, color='red', linestyle='--', label='Grand Mean')
plt.legend()

plt.tight_layout()
plt.show()
```



```
import numpy as np
import matplotlib.pyplot as plt

# Data
X = np.array([10, 20, 30, 40])
Y = np.array([12, 21, 34, 39])
Z = np.array([8, 11, 31, 39])

# Means
mean_X = np.mean(X)
mean_Y = np.mean(Y)
mean_Z = np.mean(Z)

# Grand mean
grand_mean = np.mean(np.concatenate([X, Y, Z]))

# Scatter plot
plt.figure(figsize=(12, 5))

plt.subplot(2, 1, 1)
plt.scatter(np.ones_like(X), X, label='X')
plt.scatter(np.ones_like(Y) * 2, Y, label='Y')
plt.scatter(np.ones_like(Z) * 3, Z, label='Z')
plt.scatter([1], [mean_X], label='Mean of X')
plt.scatter([2], [mean_Y], label='Mean of Y')
plt.scatter([3], [mean_Z], label='Mean of Z')
plt.xticks([1, 2, 3], ['X', 'Y', 'Z'])
plt.xlabel('Samples')
plt.ylabel('Means')
plt.title('Scatter Plot of Sample Means')
#plt.axhline(grand_mean, color='red', linestyle='--', label='Grand Mean')
#plt.legend(outside)
plt.legend(loc='center right', bbox_to_anchor=(1.25, 0.5))


# Box plot
plt.subplot(2, 1, 2)
plt.boxplot([X, Y, Z], labels=['X', 'Y', 'Z'])
plt.boxplot([[mean_X], [mean_Y], [mean_Z]], labels=['X', 'Y', 'Z'])
plt.xlabel('Samples')
plt.ylabel('Means')
plt.title('Box Plot of Sample Means')
#plt.axhline(grand_mean, color='red', linestyle='--', label='Grand Mean')
```

```
plt.legend()

plt.tight_layout()
plt.show()
```

```
No artists with labels found to put in legend.  Note that artists whose label start with an underscore are
```



6.2 One-way Analysis of Variance

- Consider m independent samples, each of size n.
  - The members of the ith sample are denoted as: $X_{i1}, X_{i2}, \ldots, X_{in}$.
  - Each X_{ij} is a normal random variable with:
    - Unknown mean: μ_i.
    - Unknown variance: σ^2.
  - Mathematically: $X_{ij} \sim N(\mu_i, \sigma^2)$, where:
    - $i = 1, \ldots, m$.
    - $j = 1, \ldots, n$.
- Hypothesis Testing:
  - Null Hypothesis (H_0): $\mu_1 = \mu_2 = \cdots = \mu_m$. (All population means are equal.)

- Alternative Hypothesis (H_1):
  Not all means are equal.
  (At least two means differ.)

Interpretation:

- Imagine m different treatments.
- Applying treatment i to an item results in a normal random variable with:
  - Mean: μ_i.
  - Variance: σ^2.
- Goal: Test if all treatments have the same effect.
- Method:
  - Apply each treatment to a different sample of n items.
  - Analyze the results to compare the means.

- Since there are a total of nm independent normal random variables X_{ij}:
  - The sum of the squares of their standardized versions follows a chi-square distribution with nm degrees of freedom.
  - Mathematically:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-E[X_{ij}])^2}{\sigma^2}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-\mu_i)^2}{\sigma^2}\sim\chi^2_{nm} \tag{6.1}$$

- To estimate the m unknown parameters $\mu_1,\dots,\mu_m$:
  - Let X_i denote the sample mean of the ith sample:

$$\bar{X}_i=\sum_{j=1}^{n}\frac{X_{ij}}{n}$$

  - Here, $\bar{X}_i$ is the estimator of the population mean μ_i for $i=1,\dots,m$.

- Substituting the estimators $\bar{X}_i$ for μ_i in Equation Equation 6.1:
  - The resulting variable:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-\bar{X}_i)^2}{\sigma^2} \tag{6.2}$$

    follows a chi-square distribution with nm − m degrees of freedom.
    (One degree of freedom is lost for each estimated parameter.)

- Define:
$$SSW = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$$
  - The variable in Equation 6.2 becomes $\frac{SSW}{\sigma^2}$.
- Since the expected value of a chi-square random variable equals its degrees of freedom:
  - Taking the expectation of the variable in Equation Equation 6.2 gives:
$$E\left[\frac{SSW}{\sigma^2}\right] = nm - m$$
  - Equivalently:
$$E\left[\frac{SSW}{nm - m}\right] = \sigma^2$$
- We thus have our first estimator of σ^2, namely, SSW /(nm - m). The statistic SSW is called the Sum of Squares Within Samples aka (within samples sum of squares or sum of squares within groups).
- Also, note that this estimator was obtained without assuming anything about the truth or falsity of the null hypothesis.
- Our second estimator of σ^2 is valid only when the null hypothesis is true.
  - Assume H_0 is true, meaning all population means μ_i are equal, i.e., $\mu_i = \mu$ for all i.
  - Under this assumption:
    - The sample means $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_m$ are normally distributed with:
      - Mean: μ.
      - Variance: $\frac{\sigma^2}{n}$.
- We have;
$$\frac{\bar{X}_{i.} - \mu}{\sqrt{\sigma^2/n}} = \frac{\sqrt{n}(\bar{X}_{i.} - \mu)}{\sigma}$$
follows a standard normal distribution; hence,
$$n\sum_{i=1}^{m}\frac{(\bar{X}_{i.} - \mu)^2}{\sigma^2} \sim \chi^2_m \tag{6.3}$$
follows a chi-square distribution with m degrees of freedom when H_0 is true.
- When all population means are equal to μ, the estimator of μ is the average of all nm data values, denoted as $\bar{X}_{..}$:
$$\bar{X}_{..} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij}}{nm} = \frac{\sum_{i=1}^{m}\bar{X}_{i.}}{m}$$

- Substituting $\bar{X}_{..}$ for the unknown parameter μ in Equation 6.3:
  - When H_0 is true, the resulting quantity:

$$n\sum_{i=1}^{m}\frac{(\bar{X}_{i.}-\bar{X}_{..})^2}{\sigma^2}$$

- The resulting quantity:

$$n\sum_{i=1}^{m}\frac{(\bar{X}_{i.}-\bar{X}_{..})^2}{\sigma^2}$$

  will be a chi-square random variable with m - 1 degrees of freedom.

- Define SSb as:

$$SSb = n\sum_{i=1}^{m}(\bar{X}_{i.}-\bar{X}_{..})^2 \tag{6.4}$$

  - It follows that, when H_0 is true:

$$SSb/\sigma^2 \sim \chi^2_{m-1}$$

- From the above, we derive that when H_0 is true:

$$E[SSb]/\sigma^2 = m-1$$

  - Equivalently:

$$E[SSb/(m-1)] = \sigma^2 \tag{6.5}$$

- Therefore, when H_0 is true, $SSb/(m-1)$ is also an estimator of σ^2.
- The quantity SSb Equation 6.4 is called the sum of squares between samples or between samples sum of squares or sum of squares between groups.
- Note that $E[SSb/(m-1)]$ is an estimate of σ^2 only if the hypothesis is true.
- Thus we have shown that;
  - $\frac{SSW}{nm-m}$ always estimates σ^2.
  - $\frac{SSb}{m-1}$ estimates σ^2 when H_0 is true.
- Because it can be shown that $\frac{SSb}{m-1}$ will tend to exceed σ^2 when H_0 is not true, now it is reasonable to let the test statistic be given by

$$TS = \frac{\frac{SSb}{m-1}}{\frac{SSW}{nm-m}}$$

and to reject H_0 when TS is sufficiently large.

How to do the hypothesis testing

- To determine how large TS needs to be to justify rejecting H_0, we use the fact that:
  - If H_0 is true, then SSb and SSW are independent.
  - It follows that, when H_0 is true, TS has an F-distribution with $m-1$ numerator and $nm-m$ denominator degrees of freedom.
  - Let $F_{m-1,nm-m,\alpha}$ denote the $100(1-\alpha)$ percentile of this distribution — that is, $P\{F_{m-1,nm-m} > F_{m-1,nm-m,\alpha}\} = \alpha$.
  - We use the notation $F_{r,s}$ to represent an F-random variable with r numerator and s denominator degrees of freedom.
- The significance level α test of H_0 is as follows:
  - Reject H_0 if $\frac{\frac{SSb}{m-1}}{\frac{SSW}{nm-m}} > F_{m-1,nm-m,\alpha}$.
  - Do not reject H_0, otherwise.

> **Note**
>
> - The following algebraic identity, known as the sum of squares identity, is useful for simplifying computations, especially when done by hand:
>
>   The Sum of Squares Identity:
>
> $$\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2 = nm\bar{X}_{\cdot\cdot}^2 + SSb + SSW$$
>
>   - Here:
>   * $\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2$ represents the total sum of squares of all observations.
>   * $nm\bar{X}_{\cdot\cdot}^2$ is the contribution from the grand mean ($\bar{X}_{\cdot\cdot}$).
>   * SSb (Sum of Squares Between) measures the variation between sample means.
>   * SSW (Sum of Squares Within) measures the variation within each sample.
> - This identity helps decompose the total variability in the data into components that can be analyzed separately.
> - When performing calculations by hand, the quantity SSb (Sum of Squares Between) should be computed first. It is defined as:
>
> $$SSb = n\sum_{i=1}^{m}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

- Once SSb has been calculated, SSW (Sum of Squares Within) can be determined using the sum of squares identity. To do this:

  1. Compute the total sum of squares:

  $$\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2$$

  2. Compute the term involving the grand mean:

  $$nm\bar{X}_{..}^2$$

  3. Use the sum of squares identity to find SSW:

  $$SSW = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2 - nm\bar{X}_{..}^2 - SSb$$

- This approach simplifies the computation process by breaking it into manageable steps and leveraging the sum of squares identity.

Alternative Method

Another useful identity involving the sum of squares is:
Error Sum of Squares (SS_{total}): $SS_{total} = SSb + SSW$
where Total Sum of Squares (SS_{total}): $SS_{total} = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{..})^2$

- This identity can also be used for the anova analysis while performing the computations by hand.
- This approach too simplifies the computation process by breaking it into manageable steps and leveraging the sum of squares identity.

Summary of one-way analysis of variance

Source of Variation	Sum of Squares	Degrees of Freedom
Between samples	$SS_b = n\sum_{i=1}^{m}(\bar{X}_{i.} - \bar{X}_{..})^2$	$m-1$
Within samples	$SS_W = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{i.})^2$	$nm-m$

Value of Test Statistic, $TS = \frac{SS_b/(m-1)}{SS_W/(nm-m)}$

Significance level α test:
reject H_0 if $TS \geq F_{m-1,nm-m,\alpha}$
do not reject otherwise

If $TS = v$, then p-value $= P\{F_{m-1,nm-m} \geq v\}$

Problem

An auto rental firm is using 15 identical motors that are adjusted to run at a fixed speed to test 3 different brands of gasoline. Each brand of gasoline is assigned to exactly 5 of the motors. Each motor runs on 10 gallons of gasoline until it is out of fuel. The following data represents the total mileages achieved by different motors using three types of gas:

- Gas 1: $220, 251, 226, 246, 260$
- Gas 2: $244, 235, 232, 242, 225$
- Gas 3: $252, 272, 250, 238, 256$

Test the hypothesis that the average mileage is not affected by the type of gas used. (In other words, determine if there is a significant difference in the mean mileages for the three types of gas.)

Multiple Comparisons of Sample Means

- When the null hypothesis of equal means is rejected, we are often interested in comparing the different sample means $\mu_1, \ldots, \mu_m$.
- One commonly used procedure for this purpose is known as the T-method.
- For a specified significance level α, this method provides joint confidence intervals for all $\binom{m}{2}$ differences $\mu_i - \mu_j$ (where $i \neq j$, and $i, j = 1, \ldots, m$), ensuring that with probability $1 - \alpha$, all confidence intervals will contain their respective differences $\mu_i - \mu_j$.
- The T-method is based on the following result:
- With probability $1 - \alpha$, for every $i \neq j$:

$$X_{i.} - X_{j.} - W < \mu_i - \mu_j < X_{i.} - X_{j.} + W$$

Where: - $X_{i.}$ and $X_{j.}$ are the sample means for groups i and j, respectively. - W is the critical value derived from the Studentized range distribution, adjusted for multiple comparisons.

where

$$W = \sqrt{\frac{1}{n}} \cdot C(m, nm - m, \alpha) \cdot \sqrt{\frac{SSW}{(nm - m)}}$$

and the values of $C(m, nm - m, \alpha)$ are provided for $\alpha = 0.05$ and $\alpha = 0.01$.

Problem

A college administrator claims that there is no difference in first-year grade point averages for students entering the college from any of three different city high schools. The following data provide the first-year grade point averages of 12 randomly chosen students, 4 from each of the three high schools. At the 5 percent level of significance, do these data disprove the administrator's claim? If so, determine confidence intervals for the difference in means of students from the different high schools, such that we can be 95 percent confident that all of the interval statements are valid.
School 1: 3.2, 3.4, 3.3, 3.5
School 2: 3.4, 3.0, 3.7, 3.3
School 3: 2.8, 2.6, 3.0, 2.7

6.2.1 One-Way Analysis of Variance with Unequal Sample Sizes

Source of Variation	Sum of Squares	Degrees of Freedom
Between samples	$SS_b = \sum_{i=1}^{m} n_i(\bar{X}_{i.} - \bar{X}_{..})^2$	$m-1$
Within samples	$SS_W = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{i.})^2$	$N-m$

where, N $= \sum_{i=1}^{m} n_i$

Value of Test Statistic, $TS = \frac{SS_b/(m-1)}{SS_W/(N-m)}$

Significance level α test:
reject H_0 if $TS \geq F_{m-1,N-m,\alpha}$
do not reject otherwise

If $TS = v$, then p-value $= P\{F_{m-1,N-m} \geq v\}$

Problem

Test the hypothesis that the following three independent samples are drawn from the same normal probability distribution.
Sample 1: $35, 37, 29, 27, 30$
Sample 2: $29, 38, 34, 30, 32$
Sample 3: $44, 52, 56$
Use a statistical test to determine whether the means of these samples are significantly different.

6.3 Two-way Analysis of Variance

Introduction

Example

Consider that four different examinations were administered to each of 5 students, with the scores shown in the table below. Each of the 20 data points is influenced by two factors: the exam and the student whose score on that exam is being recorded.

- The exam factor has 4 possible levels.
- The student factor has 5 possible levels.
- Data Table:

Exam	Student 1	Student 2	Student 3	Student 4	Student 5
1	75	73	60	70	86
2	78	71	64	72	90
3	80	69	62	70	85
4	73	67	63	80	92

- In general, suppose there are m possible levels of the first factor (row factor) and n possible levels of the second factor (column factor). Let X_{ij} denote the value obtained when:
  - The first factor is at level i.
  - The second factor is at level j.
- The data can be represented in an array format:

$$\begin{matrix} X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \dots & X_{mj} & \dots & X_{mn} \end{matrix}$$

- Here, the first factor is referred to as the row factor, and the second factor is referred to as the column factor.
- As in the case of one-way analysis assume that the data X_{ij} $(i = 1, \dots, m,\ j = 1, \dots, n)$ are independent normal random variables with a common variance σ^2. However, unlike the one-way analysis, where only a single factor affected the mean value of a data point, we now assume that the mean value of the data depends additively on both its row and column factors.

Summary of Two-way analysis of variance

Sum of Squares	Degrees of Freedom
Row	$SS_r = n\sum_{i=1}^{m}(\bar{X}_{i.} - \bar{X}_{..})^2$
Column	$SS_c = m\sum_{j=1}^{n}(\bar{X}_{.j} - \bar{X}_{..})^2$
Error	$SS_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$

Let $N = (m-1)(n-1)$

Null Hypothesis	Test Statistic	Significance Level α Test	p-value if TS = v
Row factor has no effect	$\frac{SS_r/(m-1)}{SS_e/N}$	Reject if $TS \geq F_{m-1,N,\alpha}$	$P\{F_{m-1,N} \geq v\}$
Column factor has no effect	$\frac{SS_c/(n-1)}{SS_e/N}$	Reject if $TS \geq F_{n-1,N,\alpha}$	$P\{F_{n-1,N} \geq v\}$

where:

- Row Mean ($\bar{X}_{i}.$):

$$\bar{X}_{i.} = \frac{\sum_{j=1}^{n} X_{ij}}{n} = \text{Average of the values in row } i$$

- Column Mean ($\bar{X}_{.j}$):

$$\bar{X}_{.j} = \frac{\sum_{i=1}^{m} X_{ij}}{m} = \text{Average of the values in column } j$$

- Grand Mean ($\bar{X}_{..}$):

$$\bar{X}_{..} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}}{nm} = \text{Average of all data values}$$

Problem 1

A researcher aims to investigate the impact of two factors on crop yield:

1. Factor A (Fertilizer Type): Two levels - Type 1 ($T1$) and Type 2 ($T2$).
2. Factor B (Irrigation Level): Three levels - Low (L), Medium (M), and High

(H).

The researcher measures the yield (in kg) for each combination of factors. Due to the high cost of the experiment, only one observation is recorded for each combination. The data is as follows:

Fertilizer	Irrigation	Yield (kg)
Type 1 ($T1$)	Low (L)	55
Type 1 ($T1$)	Medium (M)	60
Type 1 ($T1$)	High (H)	70
Type 2 ($T2$)	Low (L)	55
Type 2 ($T2$)	Medium (M)	65
Type 2 ($T2$)	High (H)	75

Note

Here also we have an useful identity for the sum of squares
Total Sum of Squares (SS_{total}): $SS_{total} = SS_{error} + SS_r + SS_c$
where Total Sum of Squares (SS_{total}): $SS_{total} = \sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - \bar{X}_{..})^2$

Supplementary Topics

Degrees of Freedom (dof)

- The statement "One degree of freedom is lost for each estimated parameter" is a key concept in statistics, especially in hypothesis testing and parameter estimation. Here's a detailed explanation:

What Are Degrees of Freedom?

- Degrees of freedom (df) represent the number of independent pieces of information available to estimate a parameter or test a hypothesis.
- In simpler terms, it's the number of values in a calculation that are free to vary after certain constraints (like estimating parameters) are applied.

Why Are Degrees of Freedom Lost?

- When you estimate parameters (e.g., population means, variances) from sample data, you use the data itself to calculate these estimates.
- Each time you estimate a parameter, you impose a constraint on the data, reducing the number of independent pieces of information available.

- Example: If you estimate the sample mean ($\bar{X}$) from a dataset, you use the data to calculate $\bar{X}$. This means one piece of information (one degree of freedom) is "used up" in estimating $\bar{X}$, and the remaining data points are no longer fully independent.

Application in the Context of the Problem

- In the given problem:
  1. You have m samples, each of size n, and you estimate the mean of each sample (μ_i) using the sample mean ($\bar{X}_{i.}$).
  2. Each time you estimate a mean (μ_i), you lose one degree of freedom because the data is used to calculate that estimate.
  3. Since you estimate m means ($\mu_1, \mu_2, \ldots, \mu_m$), you lose m degrees of freedom in total.

Mathematical Explanation

- Initially, the sum of squares:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-\mu_i)^2}{\sigma^2}$$

  follows a chi-square distribution with nm degrees of freedom (since there are nm independent observations).

- However, when you replace the true means (μ_i) with their estimates ($\bar{X}_{i.}$), you lose m degrees of freedom (one for each estimated mean). This is because the estimates are derived from the data, reducing the independence of the observations.

- As a result, the modified sum of squares:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-\bar{X}_{i.})^2}{\sigma^2}$$

  follows a chi-square distribution with nm - m degrees of freedom.

Why Does This Matter?

- Degrees of freedom affect the shape and critical values of the chi-square distribution, which is used for hypothesis testing.
- Losing degrees of freedom accounts for the fact that estimating parameters introduces uncertainty into the analysis.
- This adjustment ensures that statistical tests (e.g., ANOVA) are accurate and reliable.

Summary

- Degrees of freedom represent independent information in the data.
- Each estimated parameter (e.g., a mean) reduces the degrees of freedom by 1 because the data is used to calculate the estimate.
- In the problem, estimating m means reduces the degrees of freedom from nm to nm - m, ensuring the chi-square distribution is correctly applied.

7 Regression

7.1 Linear Regression

- Many engineering and scientific problems aim to determine relationships between variables.
  - Example: In a chemical process, the relationship between output, temperature, and catalyst amount is of interest.
  - Knowing this relationship allows predicting outputs for different temperature and catalyst values.
- Typically, there is a single response variable Y (dependent variable) that depends on a set of input variables $x_1, x_2, \dots, x_r$ (independent variables).
- The simplest relationship between Y and $x_1, x_2, \dots, x_r$ is a linear relationship:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r \tag{7.1}$$

  - Here, $\beta_0, \beta_1, \dots, \beta_r$ are constants.
- In practice, exact predictions are rarely possible due to random errors.
  - The relationship is better expressed as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + e \tag{7.2}$$

    * e represents the random error, assumed to have a mean of 0.
- Alternatively, the relationship can be expressed in terms of the expected value:

$$E[Y \mid \mathrm{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$

  - $\mathrm{x} = (x_1, x_2, \dots, x_r)$ is the set of independent variables.
  - $E[Y \mid \mathrm{x}]$ is the expected response given the inputs x.
- The Equation 7.2 is called a linear regression equation.
  - It describes the regression of Y on the independent variables $x_1, x_2, \dots, x_r$.

  - The quantities $\beta_0, \beta_1, \ldots, \beta_r$ are called regression coefficients, and must usually be estimated from a set of data.

- A regression equation can be classified based on the number of independent variables:
  - Simple regression equation: Contains a single independent variable ($r = 1$).
  - Multiple regression equation: Contains many independent variables.

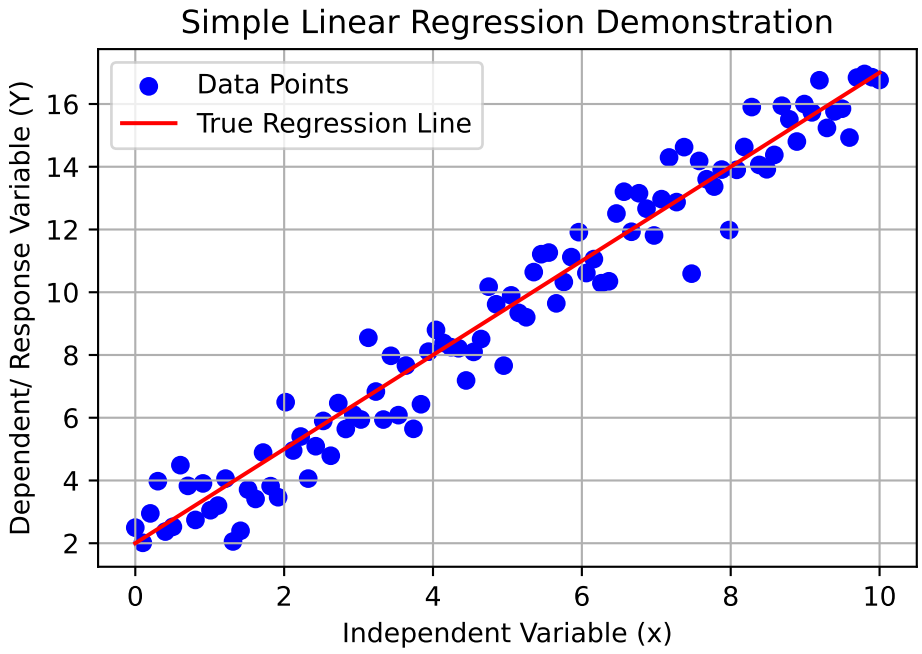
- A simple linear regression model assumes a linear relationship between the mean response and a single independent variable.
  - It is expressed as:

$$Y = \alpha + \beta x + e$$

    * x is the value of the independent variable (also called the input level).
    * Y is the response.
    * e represents the random error, assumed to be a random variable with mean 0.
    * Regression analysis is used to find the mean value of the dependent variable Y given the independent variables x.
      · The mean value of Y given x is denoted as $E[Y \mid \mathrm{x}]$.
      · By estimating the regression coefficients $\beta_0, \beta_1, \ldots, \beta_r$, we can predict the mean response for any given set of input variables.
      · This is particularly useful in understanding the central tendency of the response variable and making informed decisions based on the predicted mean.

```
import numpy as np
import matplotlib.pyplot as plt

# Set random seed for reproducibility
np.random.seed(42)

# Parameters for the linear regression model
alpha = 2.0  # Intercept
beta = 1.5   # Slope
num_samples = 100  # Number of data points

# Generate synthetic data
x = np.linspace(0, 10, num_samples)  # Independent variable (input)
e = np.random.normal(0, 1, num_samples)  # Random error with mean 0
```

```
Y = alpha + beta * x + e  # Dependent variable (response)

# Plot the data points
plt.scatter(x, Y, color='blue', label='Data Points')

# Plot the true regression line (without error)
plt.plot(x, alpha + beta * x, color='red', label='True Regression Line')

# Add labels and title
plt.xlabel('Independent Variable (x)')
plt.ylabel('Dependent/ Response Variable (Y)')
plt.title('Simple Linear Regression Demonstration')
plt.legend()

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



7.1.1 Least Squares Estimators Of The Regression Parameters

- As mentioned previouslly, a simple linear regression model assumes a linear relationship between the mean response and a single independent variable.

  - It is expressed as:
$$Y = \alpha + \beta x + e$$
    * x is the value of the independent variable (also called the input level).
    * Y is the response.
    * e represents the random error, assumed to be a random variable with mean 0.

- Suppose the responses Y_i corresponding to the input values x_i for $i = 1, \dots, n$ are observed and used to estimate α and β in a simple linear regression model.
  - Let A be the estimator of α and B be the estimator of β.
  - The estimated response for the input x_i is $A + Bx_i$.
  - The difference between the actual response Y_i and the estimated response is $(Y_i - A - Bx_i)$ is called residual.

- The sum of squares of residuals or residual sum of squares (SSR) is given by:
$$SSR = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

- The method of least squares chooses A and B to minimize SSR.
  - To find the minimizing values, differentiate SSR with respect to A and B:
$$\frac{\partial SSR}{\partial A} = -2 \sum_{i=1}^{n} (Y_i - A - Bx_i)$$
$$\frac{\partial SSR}{\partial B} = -2 \sum_{i=1}^{n} x_i (Y_i - A - Bx_i)$$
  - Setting the partial derivatives to zero yields the normal equations:
$$\sum_{i=1}^{n} Y_i = nA + B \sum_{i=1}^{n} x_i$$
$$\sum_{i=1}^{n} x_i Y_i = A \sum_{i=1}^{n} x_i + B \sum_{i=1}^{n} x_i^2$$

- Let $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ and $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.
  - The first normal equation can be rewritten as:
$$A = \overline{Y} - B\overline{x} \tag{7.3}$$

- Substituting $A = \overline{Y} - B\overline{x}$ into the second normal equation:

$$\sum_{i=1}^{n} x_i Y_i = (\overline{Y} - B\overline{x})n\overline{x} + B\sum_{i=1}^{n} x_i^2$$

Simplifying:

$$B\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y}$$

Solving for B:

$$B = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

- Using Equation 7.3 and the fact that $n\overline{x} = \sum_{i=1}^{n} x_i$, we obtain the estimators for α and β.

Least Squares Estimators

- The least squares estimators of β and α for the data set (x_i, Y_i), where $i = 1, \ldots, n$, are given by:
  - The estimator for β (B):

$$B = \frac{\sum_{i=1}^{n} x_i Y_i - \overline{x}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

  - The estimator for α (A):

$$A = \overline{Y} - B\overline{x}$$

  - Here:
    - $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is the mean of the independent variable x.
    - $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ is the mean of the dependent variable Y.

Problem

The raw material used in the production of a certain synthetic fiber is stored in a location without humidity control. Measurements of the relative humidity in the storage location and the moisture content of a sample of the raw material were taken over 15 days, resulting in the following data (in percentages):

Relative Humidity (%)	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture Content (%)	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Use linear regression to express moisture content as a function of the relative humidity.

The Coefficient of Determination and the Sample Correlation Coefficient

- Suppose we want to measure the amount of variation in the set of response values $Y_1, \dots, Y_n$ corresponding to the input values $x_1, \dots, x_n$.
  - A standard measure of variation in statistics is given by:

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

  ('SST' also refers to 'Total Sum of Squares')

  - Here, $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ is the mean of the response values.
  - If all Y_i are equal (i.e., $Y_i = \overline{Y}$ for all i), then $SST = 0$.
- The variation in the response values Y_i arises from two factors:
  1. Different input values: The input values x_i are different, leading to different mean values for the responses Y_i.
  2. Inherent variance: Even after accounting for the input values, each response Y_i has a variance σ^2 and will not exactly equal the predicted value at its input x_i.
- To determine how much variation is due to the different input values and how much is due to inherent variance, consider:
  - The quantity:

$$SSR = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

    measures the remaining variation in the response values after accounting for the input values.

– The difference:

$$SST - SSR$$

represents the variation in the response values explained by the different input values.

* The coefficient of determination (R^2) is defined as the proportion of the total variation in the response values that is explained by the input values:

$$R^2 = \frac{SST - SSR}{SST}$$

  · R^2 ranges from 0 to 1.
  · An R^2 value of 1 indicates that the regression model perfectly explains the variation in the response values.
  · An R^2 value of 0 indicates that the regression model does not explain any of the variation in the response values.

* The sample correlation coefficient (r) measures the strength and direction of the linear relationship between the independent variable x and the dependent variable Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \sqrt{R^2}$$

  · Note that the sign of r is the sign of B.
  · r ranges from -1 to 1.
  · An r value of 1 indicates a perfect positive linear relationship.
  · An r value of -1 indicates a perfect negative linear relationship.
  · An r value of 0 indicates no linear relationship.

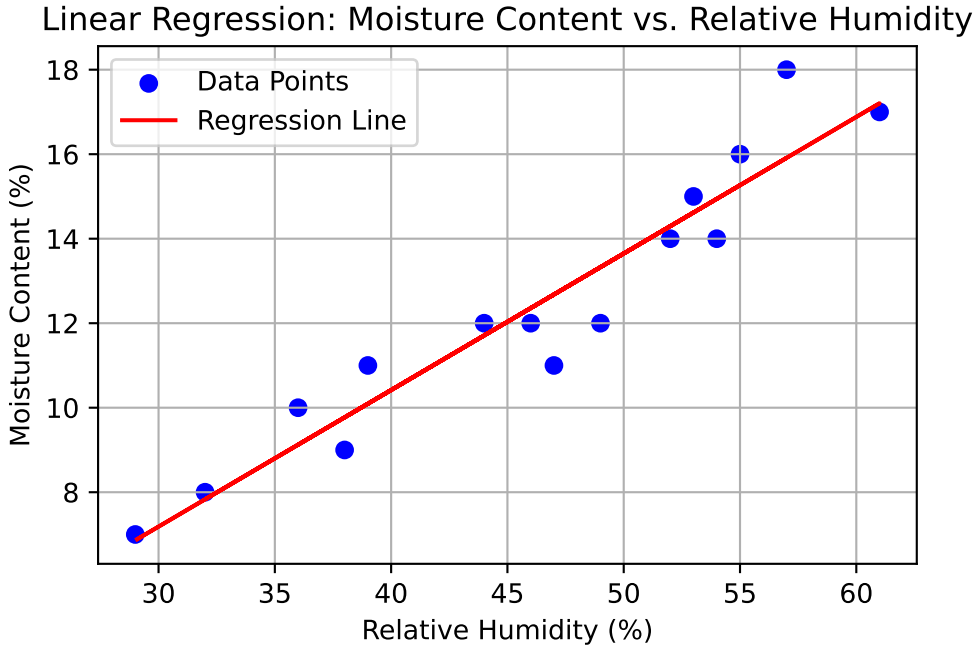
Problem

Find the coefficient of determination and the sample correlation coefficient of the previous regression problem.

```
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import matplotlib.pyplot as plt

# Data for relative humidity and moisture content
humidity = np.array([46, 53, 29, 61, 36, 39, 47, 49, 52, 38, 55, 32, 57, 54, 44]).reshape(-1, 1)
moisture = np.array([12, 15, 7, 17, 10, 11, 11, 12, 14, 9, 16, 8, 18, 14, 12])
```

```
# Fit the linear regression model
model = LinearRegression().fit(humidity, moisture)

# Print the coefficients
print(f"Intercept: {model.intercept_}")
print(f"Coefficient: {model.coef_[0]}")

# Predict the moisture content
moisture_pred = model.predict(humidity)

# Calculate the coefficient of determination (R^2)
R_squared = r2_score(moisture, moisture_pred)
print(f"Coefficient of Determination (R^2): {R_squared}")

# Calculate the sample correlation coefficient (r)
correlation_coefficient = np.sqrt(R_squared)
print(f"Sample Correlation Coefficient (r): {correlation_coefficient}")

# Plot the data points
plt.scatter(humidity, moisture, color='blue', label='Data Points')

# Plot the regression line
plt.plot(humidity, moisture_pred, color='red', label='Regression Line')

# Add labels and title
plt.xlabel('Relative Humidity (%)')
plt.ylabel('Moisture Content (%)')
plt.title('Linear Regression: Moisture Content vs. Relative Humidity')
plt.legend()

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

Intercept: -2.5104576516877177
Coefficient: 0.32320356181404014
Coefficient of Determination (R^2): 0.9113639730826797
Sample Correlation Coefficient (r): 0.9546538498757965



7.1.2 Multiple Linear Regression

- Overview:
  - In most applications, the response Y of an experiment is better predicted using multiple independent input variables rather than a single one.
  - A typical scenario involves k input variables, and the response Y is related to them by the equation:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

    where:
    - x_j is the level of the j^{th} input variable $(j = 1, \ldots, k)$,
    - e is a random error term, assumed to be normally distributed with mean 0 and constant variance σ^2.
- Parameters:
  - The parameters $\beta_0, \beta_1, \ldots, \beta_k$ and σ^2 are unknown and must be estimated from the data.
  - The data consists of observed values $Y_1, Y_2, \ldots, Y_n$, where each Y_i corresponds to a set of input levels $x_{i1}, x_{i2}, \ldots, x_{ik}$.
- Expected Value:
  - The expected value of Y_i is given by:

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

Determining the Least Squares Estimators

- Least Squares Estimation:
  - Let $B_0, B_1, \ldots, B_k$ denote estimators of $\beta_0, \beta_1, \ldots, \beta_k$.
  - The sum of squared differences between the observed Y_i and their estimated expected values is:

$$\sum_{i=1}^{n}(Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik})^2$$

  - The least squares estimators are the values of $B_0, B_1, \ldots, B_k$ that minimize the above sum of squared differences.
- Method:
  - To determine the least squares estimators $B_0, B_1, \ldots, B_k$, we take partial derivatives of the sum of squared differences with respect to each estimator:
    - First, with respect to B_0,
    - Then, with respect to B_1,
    - And so on, up to B_k.
  - By setting these partial derivatives equal to 0, we obtain a system of $k+1$ equations.
- System of Equations:
  - The partial derivatives yield the following equations:

$$\sum_{i=1}^{n}(Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i1}(Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i2}(Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik}) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik}(Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik}) = 0$$

- Normal Equations:
  - Rewriting the above equations, the least squares estimators $B_0, B_1, \ldots, B_k$ satisfy the following normal equations:

$$\sum_{i=1}^{n} Y_i = nB_0 + B_1 \sum_{i=1}^{n} x_{i1} + B_2 \sum_{i=1}^{n} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{ik}$$

$$\sum_{i=1}^{n} x_{i1} Y_i = B_0 \sum_{i=1}^{n} x_{i1} + B_1 \sum_{i=1}^{n} x_{i1}^2 + B_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{i1} x_{ik}$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik} Y_i = B_0 \sum_{i=1}^{n} x_{ik} + B_1 \sum_{i=1}^{n} x_{ik} x_{i1} + B_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{ik}^2$$

- These equations form a linear system that can be solved to find the least squares estimators $B_0, B_1, \ldots, B_k$.

7.1.3 Normal Equations in Matrix Notation

- Normal Equations:
  - The normal equations for the least squares estimators $B_0, B_1, \ldots, B_k$ are given by:

$$\sum_{i=1}^{n} x_{i1} Y_i = B_0 \sum_{i=1}^{n} x_{i1} + B_1 \sum_{i=1}^{n} x_{i1}^2 + B_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{i1} x_{ik}$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik} Y_i = B_0 \sum_{i=1}^{n} x_{ik} + B_1 \sum_{i=1}^{n} x_{ik} x_{i1} + B_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{ik}^2$$

- Matrix Notation:
  - To simplify the solution of the normal equations, we introduce matrix notation.
  - Let:
    - Y be the response vector:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

    - X be the design matrix:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

7.1.4 Matrix Representation of Multiple Regression

* β be the parameter vector:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

* e be the error vector:

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- Dimensions:
  - Y is an $n \times 1$ matrix (response vector),
  - X is an $n \times p$ matrix (design matrix), where $p = k + 1$,
  - β is a $p \times 1$ matrix (parameter vector),
  - e is an $n \times 1$ matrix (error vector).
- Multiple Regression Model:
  - The multiple regression model can be written in matrix form as:

$$Y = X\beta + e$$

- Least Squares Estimators:
  - Let B be the matrix of least squares estimators:

$$B = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_k \end{bmatrix}$$

  - The normal equations can be written in matrix form as:

$$X^T Y = X^T X B$$

    where:
    * X^T is the transpose of the design matrix X,
    * $X^T X$ is a $p \times p$ matrix,
    * $X^T Y$ is a $p \times 1$ vector.
- Solution:
  - The least squares estimators B can be obtained by solving the matrix equation:

$$B = (X^T X)^{-1} X^T Y$$

    where $(X^T X)^{-1}$ is the inverse of the matrix $X^T X$.

Residuals and Sum of Squared Residuals (SSR)

- Residuals:
  - Let r_i denote the i^{th} residual, which is the difference between the observed response Y_i and the predicted value from the regression model:

$$r_i = Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik}, \quad i = 1, \ldots, n$$

  - The residual vector r is defined as:

$$r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix}$$

  - In matrix notation, the residual vector can be expressed as:

$$r = Y - XB$$

    where:
    - Y is the response vector,
    - X is the design matrix,
    - B is the matrix of least squares estimators.
- Sum of Squared Residuals (SSR):
  - The sum of squared residuals (SSR) is given by:

$$SSR = \sum_{i=1}^{n} r_i^2$$

  - In matrix notation, the SSR can be written as:

$$SSR = r^T r$$

    where r^T is the transpose of the residual vector r.
  - Substituting $r = Y - XB$, we get:

$$SSR = (Y - XB)^T (Y - XB)$$

  - Expanding the matrix product:

$$SSR = [Y^T - (XB)^T](Y - XB)$$

$$SSR = Y^T Y - Y^T XB - B^T X^T Y + B^T X^T XB$$

- Simplification Using Normal Equations:
  - From the normal equations, we know:

$$X^T XB = X^T Y$$

  - Substituting this into the expression for SSR:

$$SSR = Y^T Y - Y^T XB - B^T X^T Y + B^T X^T XB$$

$$SSR = Y^T Y - Y^T XB$$

    (since $B^T X^T Y = B^T X^T XB$ from the normal equations).

- Transpose Property:
  - Since $Y^T XB$ is a scalar (a 1×1 matrix), it is equal to its transpose:

$$Y^T XB = (Y^T XB)^T = B^T X^T Y$$

  - Therefore, the expression for SSR simplifies further to:

$$SSR = Y^T Y - B^T X^T Y$$

- Final Computational Formula for SSR:
  - The sum of squared residuals (SSR) can be computed using the formula:

$$SSR = Y^T Y - B^T X^T Y$$

  - This formula is computationally efficient but requires careful handling to avoid roundoff errors.

- Interpretation:
  - The term $Y^T Y$ represents the total variation in the response variable Y.
  - The term $B^T X^T Y$ represents the explained variation by the regression model.
  - The difference $Y^T Y - B^T X^T Y$ gives the unexplained variation (SSR), which is minimized in the least squares method.

Coefficient of Multiple Determination (R^2)

- Definition:
  - The coefficient of multiple determination, denoted by R^2, measures the proportion of the total variation in the response variable Y that is explained by the regression model.

  - It is defined as:
$$R^2 = 1 - \frac{SSR}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$
    where:
    * SSR is the sum of squared residuals (unexplained variation),
    * $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is the total sum of squares (SST) (total variation in Y),
    * $\overline{Y}$ is the mean of the observed response values Y_i.

- Range of R^2:
  - R^2 ranges between 0 and 1:
    * $R^2 = 0$: The regression model explains none of the variation in Y.
    * $R^2 = 1$: The regression model explains all of the variation in Y.

- Formula in Terms of SST and SSR:
  - The formula for R^2 can also be written as:
$$R^2 = \frac{SST - SSR}{SST}$$
    where:
    * $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is the total sum of squares,
    * $SSR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ is the sum of squared residuals.

- Usefulness:
  - R^2 is a useful measure of the goodness of fit of the regression model.
  - A higher R^2 indicates that the model explains a larger proportion of the variation in the response variable Y.

Probabilistic Perspective of Linear Regression

- Overview:
  - Linear regression can also be viewed from a probabilistic perspective.
  - In this view, the response variable Y is considered a random variable with a probability distribution that depends on the input variables $x_1, x_2, \ldots, x_r$.

- Assumptions:
  - The response variable Y is normally distributed with mean μ_Y and variance σ^2.
  - The mean μ_Y is a linear function of the input variables:
$$\mu_Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$
  - The variance σ^2 is constant and does not depend on the input variables.

- Model:
  - The linear regression model can be written as:

    $$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \epsilon$$

    where ϵ is a random error term that follows a normal distribution with mean 0 and variance σ^2:

    $$\epsilon \sim N(0, \sigma^2)$$

- Likelihood Function:
  - Given a set of observed data points (x_i, Y_i) for $i = 1, \ldots, n$, the likelihood function represents the probability of observing the data given the parameters $\beta_0, \beta_1, \ldots, \beta_r$ and σ^2.
  - The likelihood function is given by:

    $$L(\beta_0, \beta_1, \ldots, \beta_r, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_r x_{ir})^2}{2\sigma^2}\right)$$

- Log-Likelihood Function:
  - The log-likelihood function is the natural logarithm of the likelihood function:

    $$\log L(\beta_0, \beta_1, \ldots, \beta_r, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_r x_{ir})^2$$

- Maximum Likelihood Estimation (MLE):
  - The parameters $\beta_0, \beta_1, \ldots, \beta_r$ and σ^2 can be estimated by maximizing the log-likelihood function.
  - The maximum likelihood estimators (MLEs) of $\beta_0, \beta_1, \ldots, \beta_r$ are the same as the least squares estimators.
  - The MLE of σ^2 is given by:

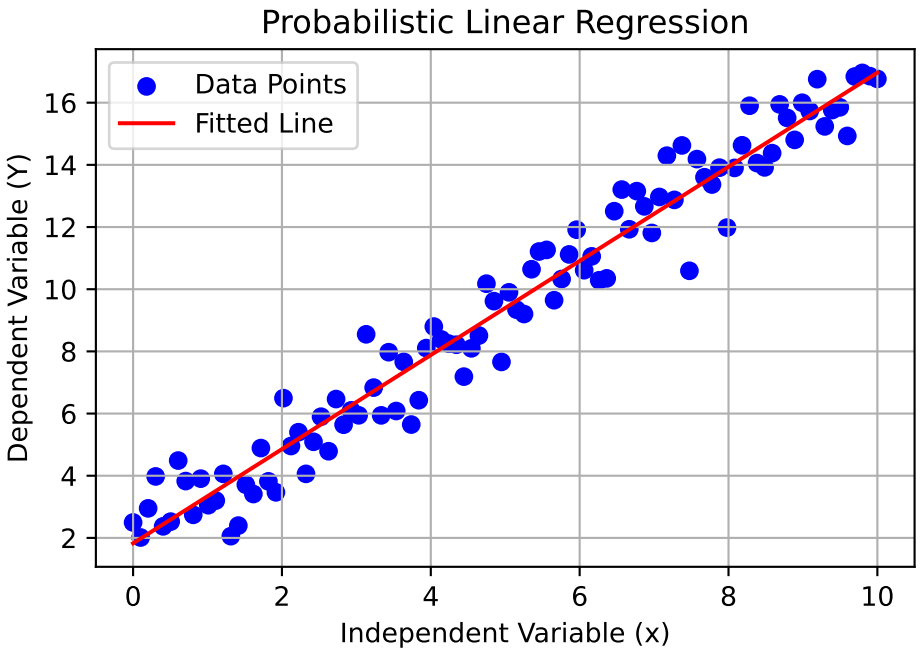
    $$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

    where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_r x_{ir}$ are the predicted values from the regression model.

- Inference:
  - The probabilistic perspective allows for statistical inference about the regression parameters.
  - Confidence intervals and hypothesis tests can be constructed for the parameters $\beta_0, \beta_1, \ldots, \beta_r$ and σ^2.

```
import numpy as np
from scipy.stats import norm

# Generate synthetic data
np.random.seed(42)
n = 100
x = np.linspace(0, 10, n)
beta_0 = 2.0
beta_1 = 1.5
sigma = 1.0
epsilon = np.random.normal(0, sigma, n)
Y = beta_0 + beta_1 * x + epsilon

# Maximum Likelihood Estimation (MLE) for beta_0, beta_1, and sigma^2
X = np.vstack([np.ones(n), x]).T

beta_hat = np.linalg.inv(X.T @ X) @ X.T @ Y

Y_hat = X @ beta_hat
sigma_hat = np.sqrt(np.sum((Y - Y_hat) ** 2) / n)

print(f"Estimated beta_0: {beta_hat[0]}")
print(f"Estimated beta_1: {beta_hat[1]}")
print(f"Estimated sigma^2: {sigma_hat ** 2}")

# Plot the data and the fitted line
import matplotlib.pyplot as plt

plt.scatter(x, Y, color='blue', label='Data Points')
plt.plot(x, Y_hat, color='red', label='Fitted Line')
plt.xlabel('Independent Variable (x)')
plt.ylabel('Dependent Variable (Y)')
plt.title('Probabilistic Linear Regression')
plt.legend()
plt.grid(True)
plt.show()
```

```
Estimated beta_0: 1.8271871459226277
Estimated beta_1: 1.5137932673366563
Estimated sigma^2: 0.8149047134980785
```



7.2 Logistic Regression

- Overview:
  - Logistic regression is used when the response variable is categorical, typically binary (0 or 1).
  - It models the probability that a given input point belongs to a particular category.
- Logistic Function:
  - The logistic function (also called the sigmoid function) is used to model the probability:

$$P(Y = 1 \mid \mathrm{x}) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r))}$$

- Logistic Regression Model:
  - The logistic regression model can be written as:

$$\log\left(\frac{P(Y = 1 \mid \mathrm{x})}{1 - P(Y = 1 \mid \mathrm{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$

  - Here, the left-hand side is the log-odds (logit) of the probability of the response being 1.

- Estimation of Parameters:
  - The parameters $\beta_0, \beta_1, \ldots, \beta_r$ are estimated using the method of maximum likelihood.
  - The likelihood function for logistic regression is given by:

$$L(\beta_0, \beta_1, \ldots, \beta_r) = \prod_{i=1}^{n} P(Y_i \mid \mathrm{x}_i)^{Y_i} (1 - P(Y_i \mid \mathrm{x}_i))^{1-Y_i}$$

- Log-Likelihood Function:
  - The log-likelihood function is:

$$\log L(\beta_0, \beta_1, \ldots, \beta_r) = \sum_{i=1}^{n} [Y_i \log P(Y_i \mid \mathrm{x}_i) + (1 - Y_i) \log(1 - P(Y_i \mid \mathrm{x}_i))]$$

- Fitting the Model:
  - The parameters are estimated by maximizing the log-likelihood function using numerical optimization techniques.

Checking the goodness of fit

7.2.1 Confusion Matrix, Accuracy, and ROC Curve

- Confusion Matrix:
  - A confusion matrix is a table used to evaluate the performance of a classification model.
  - It summarizes the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.
  - The matrix helps in understanding the types of errors the model is making.
- Accuracy:
  - Accuracy is a metric that measures the proportion of correct predictions made by the model.
  - It is calculated as the sum of true positives and true negatives divided by the total number of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  - While accuracy is a useful metric, it may not be sufficient for imbalanced datasets where one class is more frequent than the other.

8 Calculate sensitivity (recall) and specificity

sensitivity = TP / (TP + FN) specificity = TN / (TN + FP)

print(f”Sensitivity (Recall): {sensitivity}“) print(f”Specificity: {specificity}“)

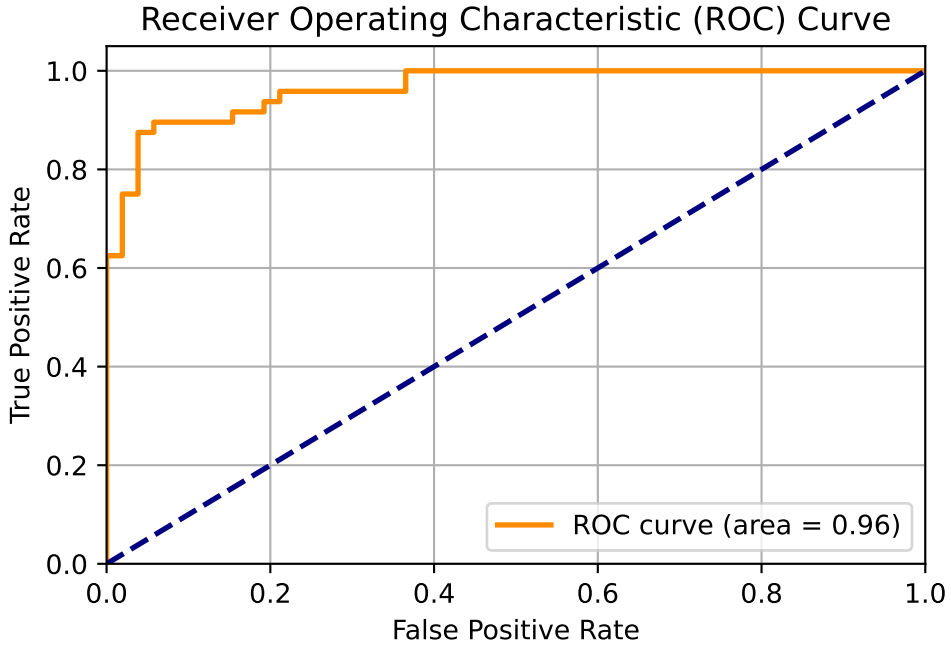
- ROC Curve:
  - The Receiver Operating Characteristic (ROC) curve is a graphical representation of a classification model’s performance.
  - It plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
  - The area under the ROC curve (AUC) is a single scalar value that summarizes the model’s ability to discriminate between positive and negative classes.
  - An AUC value of 1 indicates perfect classification, while an AUC value of 0.5 indicates no discriminative power (equivalent to random guessing).

```
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, roc_curve, auc
import matplotlib.pyplot as plt

# Generate synthetic data for logistic regression
np.random.seed(42)
n = 100
x = np.linspace(0, 10, n)
beta_0 = -5.0
beta_1 = 1.0
prob = 1 / (1 + np.exp(-(beta_0 + beta_1 * x)))
Y = np.random.binomial(1, prob, n)

# Reshape x for sklearn
x = x.reshape(-1, 1)

# Fit the logistic regression model
model = LogisticRegression().fit(x, Y)

# Predict probabilities
Y_prob = model.predict_proba(x)[:, 1]
```

8 Calculate sensitivity (recall) and specificity

```
# Predict class labels
Y_pred = model.predict(x)

# Calculate confusion matrix
conf_matrix = confusion_matrix(Y, Y_pred)
print("Confusion Matrix:")
print(conf_matrix)

# Calculate accuracy
accuracy = accuracy_score(Y, Y_pred)
print(f"Accuracy: {accuracy}")

# Calculate ROC curve and AUC
fpr, tpr, _ = roc_curve(Y, Y_prob)
roc_auc = auc(fpr, tpr)
print(f"AUC: {roc_auc}")

# Plot ROC curve
plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (area = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.grid(True)
plt.show()
```

Confusion Matrix:
[[47 5]
 [5 43]]
Accuracy: 0.9
AUC: 0.9647435897435896


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.96)

9 Time series Analysis

9.1 Introduction

- Time Series Data:
  - A sequence of observations in chronological order (e.g., daily log returns on a stock, monthly Consumer Price Index (CPI) values).
  - Often assumes equally spaced data with a discrete-time index, though real-world data may deviate:
    * Daily stock returns exclude weekends/holidays.
    * Monthly CPI values are spaced by months (unequal days), but treated as equally spaced for simplicity.
- Time Series Models:
  - Studied for applications in econometrics, business forecasting, and scientific fields.
  - Aim to capture patterns, dependencies, or trends in sequential data.
- Stochastic Process:
  - A sequence of random variables, representing the theoretical/population analog of a time series.
  - A time series is a sample from a stochastic process.
  - "Stochastic" = random.
- Stationarity:
  - A key property for achieving parsimony (simplicity) in time series models.
  - Assumes distributional invariance over time (e.g., constant mean, variance, and autocorrelation structure).

9.2 Stationary processes

- Stationary Processes:
  - Definition: Probability models for time series with time-invariant behavior.
  - Examples:
    - Financial time series (e.g., log returns) where changes may be stationary even if the series itself is not.
    - Seasonal demand (e.g., sunscreen, winter coats) with recurring patterns over shorter periods.
- Strict Stationarity:
  - All aspects of the process's distribution remain unchanged under time shifts.
  - Mathematically: For any m and n, $(Y_1, \dots, Y_n)$ and $(Y_{1+m}, \dots, Y_{n+m})$ have identical distributions.
  - A strong assumption requiring all statistical properties (mean, variance, quantiles, etc.) to be time-invariant.
- Weak Stationarity (Covariance Stationarity):
  - Requires:
    1. Constant mean: $E(Y_t) = \mu$ for all t.
    2. Constant variance: $\text{Var}(Y_t) = \sigma^2$ for all t.
    3. Covariance depends only on lag: $\text{Cov}(Y_t, Y_s) = \gamma(|t-s|)$ for some function $\gamma(h)$.
  - Example: $\text{Cov}(Y_2, Y_5) = \text{Cov}(Y_7, Y_{10})$ if $|2-5| = |7-10| = 3$.
- Autocovariance and Autocorrelation:
  - Autocovariance function: $\gamma(h) = \text{Cov}(Y_t, Y_{t+h})$, with $\gamma(h) = \gamma(-h)$.
  - Autocorrelation function: $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\gamma(h)}{\sigma^2}$.
  - $\gamma(0) = \sigma^2$ (variance at lag 0).
- Applications and Benefits:
  - Enables parsimonious modeling (fewer parameters) for time series data.
    - Example: All Y_t share a common mean μ, estimated accurately by $\overline{Y}$.
  - Non-stationary series (e.g., stock prices) may have stationary changes (e.g., log returns), allowing modeling.
- Assessing Stationarity:
  - Visual inspection:

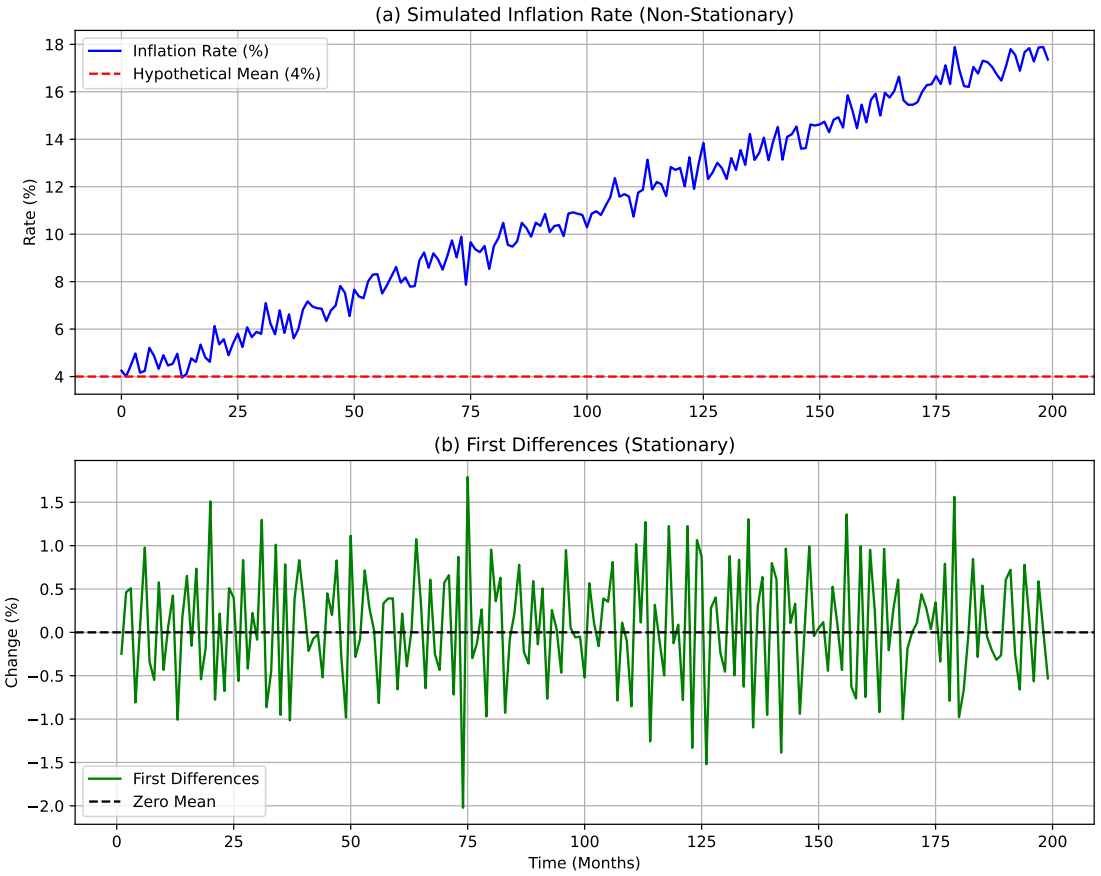
    - Time series plot should show mean-reversion (oscillation around a fixed level).
    - Non-stationary series "wander" without returning to a fixed level.
  - Statistical tests: Formal tests (e.g., ADF test) to evaluate stationarity.
  - Sample autocorrelation function (ACF): Supplementary tool for analysis.
- Challenges:
  - Stationarity is an assumption, not guaranteed.
  - Even with stationarity, uncertainty remains (e.g., limited data for estimating μ in non-stationary series).

-Dealing with non-stationary data

- Example-1:
  - One-month annualized inflation rate (%) is shown in the code below.
  - Appears to wander without clear mean reversion.
  - Ambiguity in stationarity requires further investigation.
  - First Differences:
    - Changes between consecutive months (first differences) oscillate around 0%.
    - Differenced series exhibits stationary behavior (stable mean/variance).
- Key Observations:
  - Non-stationary original series may require differencing for modeling.
  - Differenced series ($\Delta Y_t = Y_t - Y_{t-1}$) simplifies analysis by removing trends.

```
import numpy as np
import matplotlib.pyplot as plt

# Simulate inflation rate data (non-stationary) with a trend and noise
np.random.seed(42)
n = 200  # Number of months
time = np.arange(n)
trend = 0.02 * time  # Simulated upward trend
noise = np.random.normal(0, 0.5, n)
```

```
inflation_rate = 4 + 0.05 * time + trend + noise  # Non-stationary series

# Compute first differences (stationary)
diff_inflation = np.diff(inflation_rate)

# Plotting
fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(10, 8))

# Original series (non-stationary)
ax1.plot(time, inflation_rate, color='blue', label='Inflation Rate (%)')
ax1.axhline(y=4, color='red', linestyle='--', label='Hypothetical Mean (4%)')
ax1.set_title('(a) Simulated Inflation Rate (Non-Stationary)')
ax1.set_ylabel('Rate (%)')
ax1.legend()
ax1.grid(True)

# Differenced series (stationary)
ax2.plot(time[1:], diff_inflation, color='green', label='First Differences')
ax2.axhline(y=0, color='black', linestyle='--', label='Zero Mean')
ax2.set_title('(b) First Differences (Stationary)')
ax2.set_xlabel('Time (Months)')
ax2.set_ylabel('Change (%)')
ax2.legend()
ax2.grid(True)

plt.tight_layout()
plt.show()
```

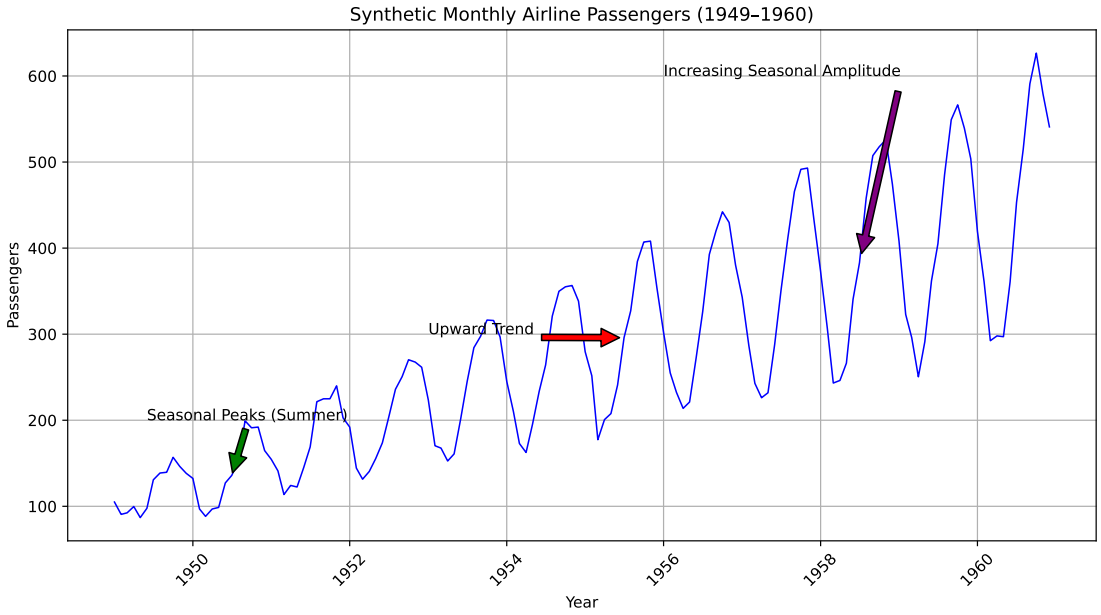


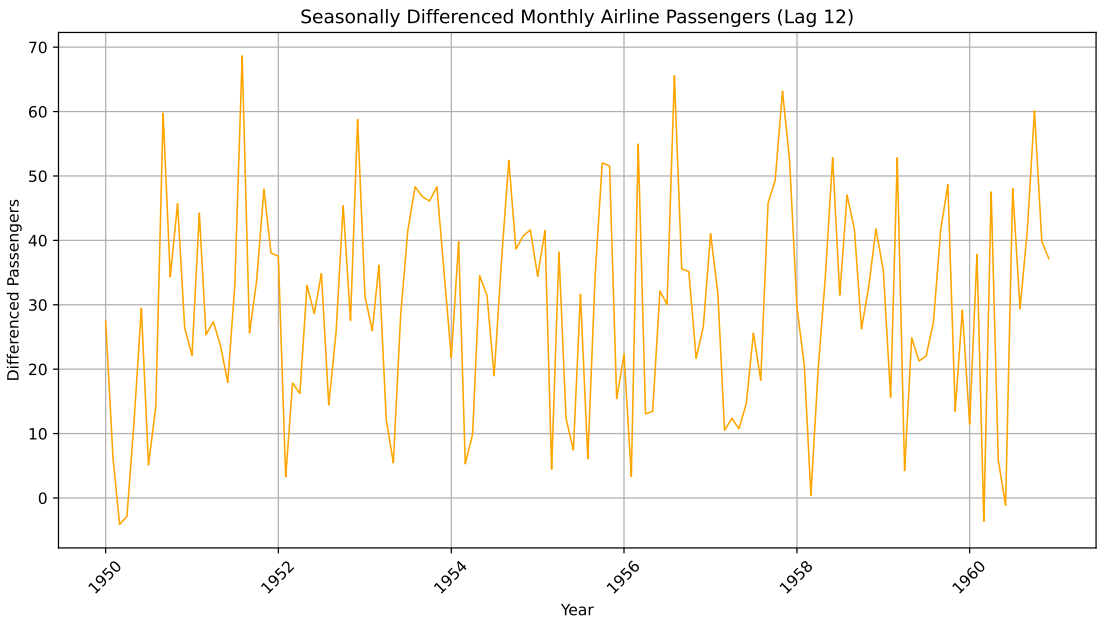
- Example-2:
  - Key Features:
    1. Upward Trend: Persistent increase in passenger numbers over time.
    2. Seasonal Variation:
       * Peaks in summer months (e.g., July, August).
       * Troughs in winter months (e.g., December, January).
    3. Increasing Seasonal Amplitude: Seasonal fluctuations grow larger over time.
  - Nonstationarity: The combination of trend, seasonality, and increasing variability violates stationarity assumptions.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
import matplotlib.dates as mdates

# Generate synthetic monthly data (1949–1960)
np.random.seed(42)
dates = pd.date_range(start="1949-01-01", end="1960-12-01", freq="MS")  # Ensure full date for

# Define n_months correctly
n_months = len(dates)

# 1. Upward trend component (linear increase)
trend_slope = 2.5  # Passengers increase by ~2.5 per month
trend = 100 + trend_slope * np.arange(n_months)  # Baseline trend

# 2. Seasonal component (peaks in summer, troughs in winter)
# Amplitude increases over time (non-stationary seasonality)
seasonal_amplitude = 0.5 * np.arange(n_months)  # Growing amplitude
month_indices = dates.month.values - 1  # 0-based months (0=Jan, 11=Dec)
seasonal = 20 * np.sin(2 * np.pi * (month_indices - 6) / 12)  # Peaks in July (month 7)
seasonal *= (1 + 0.05 * np.arange(n_months))  # Amplify seasonality over time

# 3. Random noise
noise = np.random.normal(0, 10, n_months)

# Combine components
passengers = trend + seasonal + noise

# Create DataFrame
df = pd.DataFrame({"Passengers": passengers}, index=dates)

# Plot
plt.figure(figsize=(12, 6))
plt.plot(df.index, df["Passengers"], color="blue", linewidth=1)
plt.title("Synthetic Monthly Airline Passengers (1949–1960)")
plt.xlabel("Year")
plt.ylabel("Passengers")
plt.grid(True)

# Convert annotation dates to full datetime format
annotate_dates = pd.to_datetime(["1955-07-01", "1950-07-01", "1958-07-01"])

# Annotate features
plt.annotate(
    "Upward Trend",
```

```
    xy=(annotate_dates[0], df.loc[annotate_dates[0], "Passengers"]),
    xytext=(pd.Timestamp("1953-01-01"), 300),
    arrowprops=dict(facecolor="red", shrink=0.05),
)
plt.annotate(
    "Seasonal Peaks (Summer)",
    xy=(annotate_dates[1], df.loc[annotate_dates[1], "Passengers"]),
    xytext=(pd.Timestamp("1949-06-01"), 200),
    arrowprops=dict(facecolor="green", shrink=0.05),
)
plt.annotate(
    "Increasing Seasonal Amplitude",
    xy=(annotate_dates[2], df.loc[annotate_dates[2], "Passengers"]),
    xytext=(pd.Timestamp("1956-01-01"), 600),
    arrowprops=dict(facecolor="purple", shrink=0.05),
)

# Format x-axis dates
plt.gca().xaxis.set_major_formatter(mdates.DateFormatter('%Y'))
plt.xticks(rotation=45)

# Differencing to remove seasonality (lag 12)
seasonal_diff = df["Passengers"].diff(12).dropna()

# Plot the differenced series
plt.figure(figsize=(12, 6))
plt.plot(seasonal_diff, color="orange", linewidth=1)
plt.title("Seasonally Differenced Monthly Airline Passengers (Lag 12)")
plt.xlabel("Year")
plt.ylabel("Differenced Passengers")
plt.grid(True)

# Format x-axis dates
plt.gca().xaxis.set_major_formatter(mdates.DateFormatter('%Y'))
plt.xticks(rotation=45)

plt.show()
```





9.3 Differencing to Make Time Series Stationary

Differencing is a technique used to transform a non-stationary time series into a stationary one. A stationary time series has a constant mean, variance, and autocorrelation over time, which is crucial for many time series forecasting methods.

9.3.1 First Order Differencing

First order differencing involves subtracting the previous observation from the current observation. This helps to remove trends and stabilize the mean of the time series (Example-1 involving the inflation rates has the same scenario).

For a time series Y_t, the first order differenced series Y_t' is calculated as:

$$Y_t' = Y_t - Y_{t-1}$$

9.3.2 Higher Order Differencing

If the first order differencing is not sufficient to make the series stationary, higher order differencing can be applied. This involves differencing the differenced series.

For example, the second order differenced series Y_t'' is calculated as:

$$Y_t'' = Y_t' - Y_{t-1}'$$
$$Y_t'' = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$$
$$Y_t'' = Y_t - 2Y_{t-1} + Y_{t-2}$$

9.3.3 Seasonal Differencing

Seasonal differencing is used to remove seasonal patterns. This involves subtracting the observation from the same season in the previous cycle.

For a time series with a seasonal period s, the seasonally differenced series Y_t' is calculated as: $Y_t' = Y_t - Y_{t-s}$

By applying these differencing techniques, we can transform a non-stationary time series into a stationary one, making it suitable for further analysis and forecasting.

9.4 Autoregressive Models: AR(1), AR(p), and Model Selection with PACF

- A series displays autoregressive (AR) behavior if it apparently feels a "restoring forcerestoring force" that tends to pull it back toward its mean.
- Applications:
  - Economics: Modeling GDP, inflation rates, and stock prices.
  - Weather Forecasting: Predicting temperature and precipitation.
  - Engineering: Signal processing and control systems.
  - Healthcare: Analyzing patient vital signs and disease progression.

9.4.1 AR(1) Model

Definition:
A first-order autoregressive model where the current value depends linearly on its immediate past value and a stochastic term.

Formula:

$$(Y_t - \mu) = \phi_1 (Y_{t-1} - \mu) + \epsilon_t, \quad \epsilon_t \sim \text{WN}(0, \sigma^2)$$

Also written as;

$$Y_t = c + \phi_1 Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim \text{WN}(0, \sigma_\epsilon^2)$$

where c is the intercept.

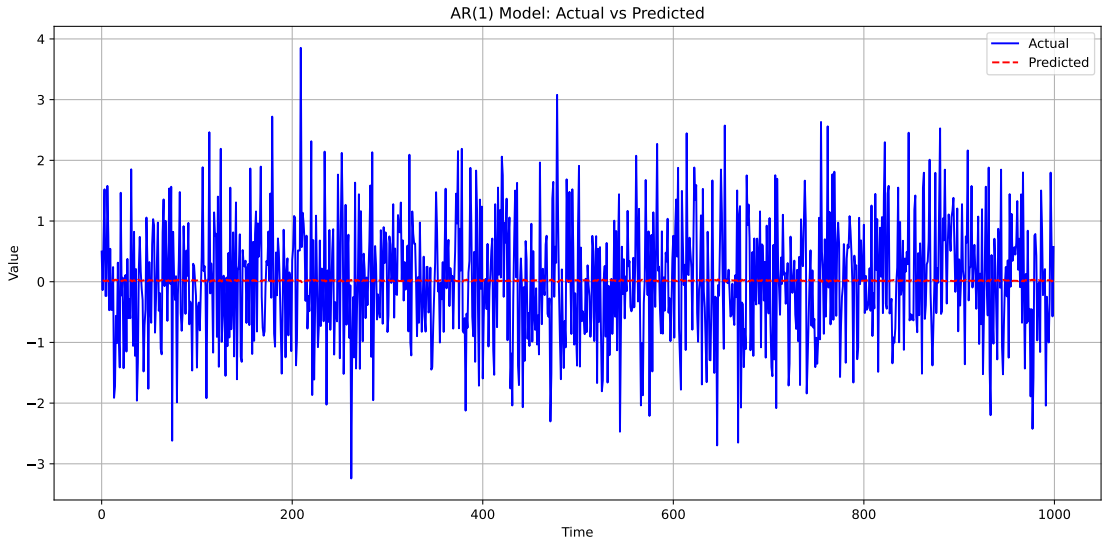
Key Properties:

- Stationarity: Requires $|\phi_1| < 1$.
- ACF: Exponentially decays.
- PACF: Cuts off abruptly after lag 1.

Python Example:

```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.ar_model import AutoReg
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# Generate dummy time series data
np.random.seed(42)
n = 1000
data = np.random.randn(n)

# Fit AR(1) model
model = AutoReg(data, lags=1)
model_fit = model.fit()
predictions = model_fit.predict(start=1, end=n-1)

# Plot the actual and predicted time series
plt.figure(figsize=(12, 6))
plt.plot(data, label='Actual', color='blue')
plt.plot(np.arange(1, n), predictions, label='Predicted', color='red', linestyle='--')
plt.title('AR(1) Model: Actual vs Predicted')
```

```
plt.xlabel('Time')
plt.ylabel('Value')
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```



9.4.2 Autoregressive Model of Order p: AR(p)

9.4.2.1 Definition

The AR(p) model is a linear regression of the current value of a time series against its own past p values.

Formula:

$$(Y_t - \mu) = \phi_1 (Y_{t-1} - \mu) + \phi_2 (Y_{t-2} - \mu) + \cdots + \phi_p \left(Y_{t-p} - \mu\right) + \epsilon_t$$

where: - $\phi_1, \ldots, \phi_p$: Autoregressive coefficients - ϵ_t: White noise error term $\sim \text{WN}(0, \sigma^2)$

9.4.2.2 Key Properties

1. Stationarity:

- Requires roots of the characteristic equation

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

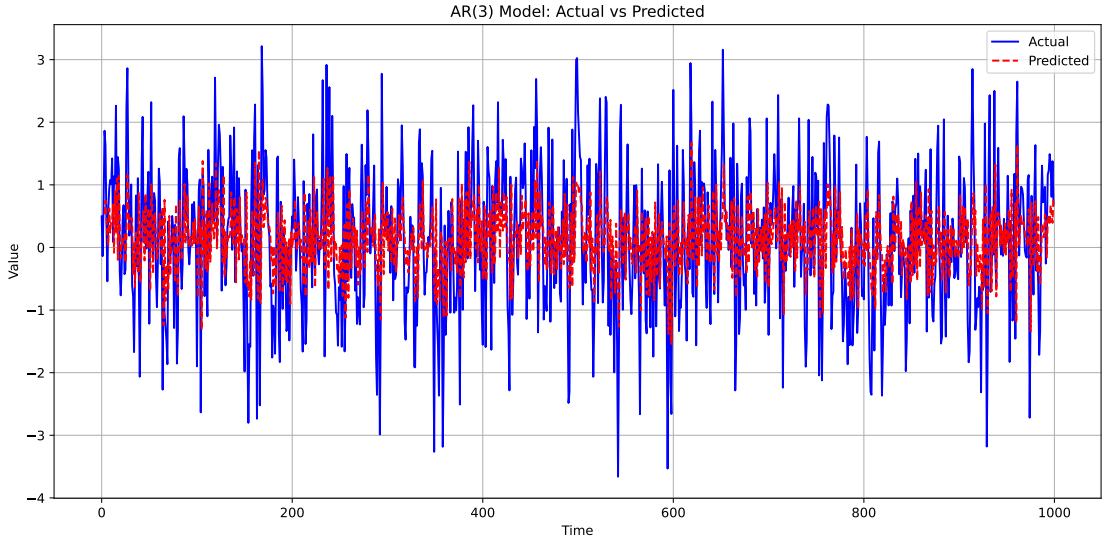
to lie outside the unit circle.

2. ACF: Decays gradually (exponentially or oscillating).
3. PACF: Cuts off abruptly after lag p.

9.4.2.3 Python Implementation

```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.ar_model import AutoReg
from statsmodels.graphics.tsaplots import plot_pacf

# Generate dummy time series data for AR(p) process
np.random.seed(42)
n = 1000
p = 3  # Order of the AR process
phi = [0.5, -0.3, 0.2]  # AR coefficients
data = np.random.randn(n)

# Simulate AR(p) process
for t in range(p, n):
  data[t] = phi[0] * data[t-1] + phi[1] * data[t-2] + phi[2] * data[t-3] + np.random.randn()

# Fit AR(p) model
model = AutoReg(data, lags=p)
model_fit = model.fit()
predictions = model_fit.predict(start=p, end=n-1)

# Plot the actual and predicted time series
plt.figure(figsize=(12, 6))
plt.plot(data, label='Actual', color='blue')
plt.plot(np.arange(p, n), predictions, label='Predicted', color='red', linestyle='--')
plt.title(f'AR({p}) Model: Actual vs Predicted')
plt.xlabel('Time')
plt.ylabel('Value')
plt.legend()
plt.grid(True)
```

```
plt.tight_layout()
plt.show()
```



9.4.3 Identify Order p Using PACF

9.4.3.1 What is PACF?

- The Partial Autocorrelation Function (PACF) measures the correlation between Y_t and Y_{t-h} after removing the effects of the intermediate lags $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-h+1}$.
- For AR(p) models, the PACF helps identify the order p.

9.4.3.2 Method to Determine p

1. Plot the PACF of the time series.
2. Identify Significant Spikes:
   - Significant spikes (outside the confidence band) at lags $1, 2, \ldots, p$.
   - PACF cuts off (becomes insignificant) after lag p.
3. Interpretation:
   - If PACF drops to near zero after lag p, the process is likely AR(p).

9.4.3.3 ACF and PACF

Autocorrelation Function (ACF):

The ACF measures the correlation between observations of a time series separated by h time units (lags).

$$ACF = \rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

where: - $\rho(h)$ is the autocorrelation at lag h. - $\gamma(h) = \text{Cov}(Y_t, Y_{t+h})$ is the autocovariance at lag h. - $\gamma(0) = \text{Var}(Y_t)$ is the variance of the series.

Partial Autocorrelation Function (PACF): The PACF measures the correlation between Y_t and Y_{t-h} after removing the effects of the intermediate lags $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-h+1}$.

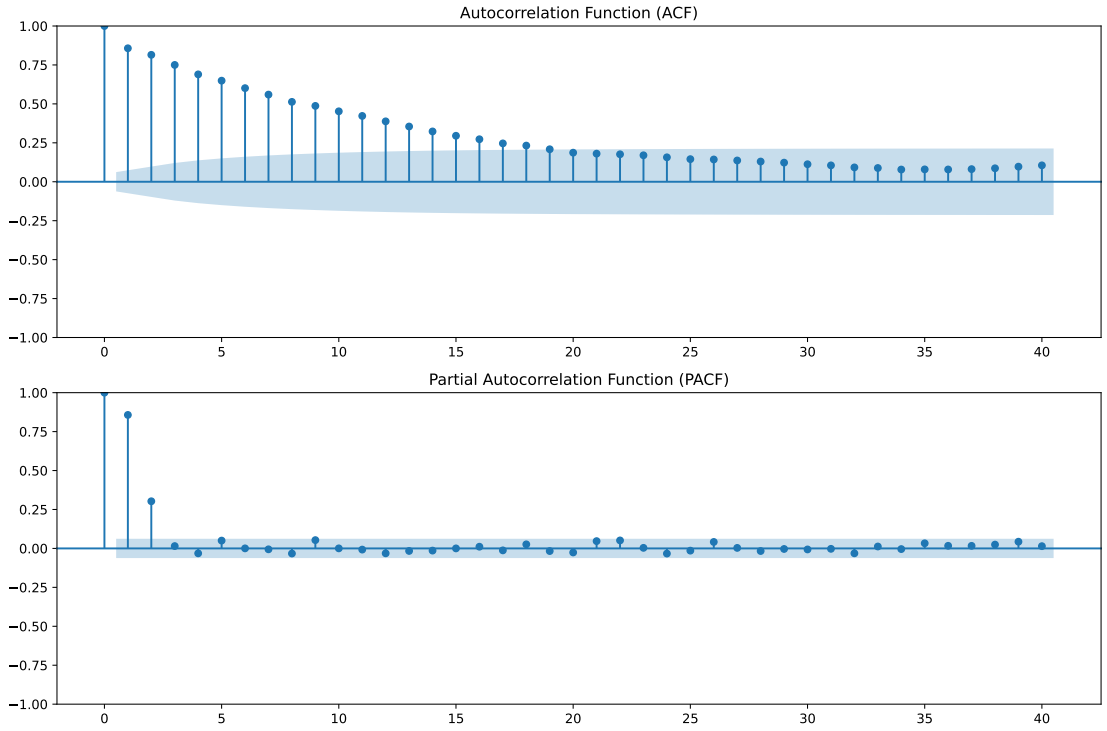
Relation to Correlation Factor in Regression:

- In regression analysis, the correlation factor (or coefficient) measures the strength and direction of the linear relationship between two variables. Similarly, the ACF and PACF measure the linear relationship between lagged values of a time series.
- ACF: Analogous to the correlation coefficient in regression, it measures the linear relationship between Y_t and Y_{t-h}.
- PACF: Analogous to the partial correlation coefficient in multiple regression, it measures the linear relationship between Y_t and Y_{t-h} after accounting for the linear relationships with intermediate lags.
- By examining the ACF and PACF plots, we can identify the appropriate lag structure for autoregressive models, similar to how we use correlation and partial correlation in regression to identify relevant predictors.

9.4.3.4 Example: Identifying p

```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# Generate AR(2) time series data
np.random.seed(42)
n = 1000
```

```
phi1 = 0.6
phi2 = 0.3
data = np.zeros(n)
data[0] = np.random.randn()
for t in range(1, n):
    data[t] = phi1 * data[t-1] + phi2 * data[t-2] + np.random.randn()

# Plot ACF and PACF
fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(12, 8))

# ACF plot
plot_acf(data, ax=ax1, lags=40)
ax1.set_title('Autocorrelation Function (ACF)')

# PACF plot
plot_pacf(data, ax=ax2, lags=40)
ax2.set_title('Partial Autocorrelation Function (PACF)')

plt.tight_layout()
plt.show()
```



Output Interpretation

Key Observations:

1. Significant PACF Spikes:
    - Lag 1: Partial autocorrelation coefficient exceeds the confidence band (statistically significant).
    - Lag 2: Another significant spike outside the confidence band.
    - These spikes indicate autoregressive terms at lags 1 and 2.
2. Cutoff After Lag 2:
    - For lags $h > 2$, PACF values fall within the confidence band (gray dashed lines at $\pm\frac{1.96}{\sqrt{n}}$).
    - No significant partial autocorrelation beyond lag 2.

Conclusion:

- The PACF plot suggests an AR(2) model is appropriate ($p = 2$).
- Further lags ($p > 2$) do not contribute meaningfully to the model.

Important

AR (Autoregressive) Models Definition:
Models where the current value depends on its own past values.
Formula:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t$$

Strengths:
- Captures persistence/momentum (e.g., temperature trends, stock price inertia).
- Simple to interpret (linear dependence on past values).
Limitations:
- Struggles with short-lived shocks (e.g., sudden market crashes).
- Requires high order p to model complex dependencies $\rightarrow$ risk of overfitting.

Ljung-Box Test

The Ljung-Box test is a statistical test used to determine whether a time series exhibits significant autocorrelation at multiple lags. It is particularly useful for diagnosing the adequacy of time series models by checking if the residuals are uncorrelated (i.e., white noise).

Hypotheses

- Null Hypothesis (H_0): The data are independently distributed (no autocorrelation).
- Alternative Hypothesis (H_a): The data exhibit significant autocorrelation.

Test Statistic

The Ljung-Box test statistic is given by:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k}$$

where: - n: Number of observations. - m: Number of lags being tested. - $\hat{\rho}_k$: Sample autocorrelation at lag k.
Under the null hypothesis, Q follows a χ^2 distribution with m degrees of freedom.

Interpretation

1. If p-value $>$ significance level (α):

- Fail to reject H_0.
- Residuals are uncorrelated (white noise).

2. If p-value significance level (α):

- Reject H_0.
- Residuals exhibit significant autocorrelation.

Applications

- Model Diagnostics: Used to check if the residuals of a fitted time series model are white noise.
- Goodness-of-Fit: Helps assess whether the chosen model adequately captures the structure of the data.

Limitations

- Sensitive to the choice of m (number of lags).
- Assumes the model parameters are known (may lead to biased results if parameters are estimated).

The Ljung-Box test is a critical tool for validating time series models, ensuring that the residuals are free of autocorrelation and suitable for forecasting.

9.5 Moving Average Models: MA(1), MA(q) models and the choice of q.

A series displays moving-average (MA) model behavior if it apparently undergoes random "shocks" whose effects are felt in two or more consecutive periods.

9.5.1 MA(1) Model

9.5.1.1 Definition

The MA(1) (Moving Average of Order 1) model is a time series model where the current value depends linearly on the current random shock and the previous random shock.

Formula:

$$Y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}$$

- μ: Mean of the series.
- ϵ_t: White noise error term at time t ($\epsilon_t \sim \text{WN}(0, \sigma^2)$).
- θ: Coefficient of the lagged error term.

9.5.1.2 Key Properties

1. Stationarity:
    - MA(1) is always stationary (no restrictions on θ).
2. Autocorrelation Function (ACF):
    - Spikes at lag 1: $\rho(1) = \frac{\theta}{1+\theta^2}$.
    - $\rho(h) = 0$ for $h \geq 2$.
3. Partial Autocorrelation Function (PACF):
    - Decays exponentially (does not cut off abruptly).

9.5.1.3 Example Use Cases

- Stock Returns: Unexpected news (shocks) affecting returns for two days.
- Inventory Management: A supply disruption (shock) impacting current and next month's inventory levels.
- Weather Data: A temperature anomaly affecting consecutive days.

9.5.2 Moving Average Model of Order q: MA(q)

9.5.2.1 Definition

The MA(q) model is a linear regression of the current value of a time series against the past q error terms (shocks).
Formula:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

where: - μ: Mean of the series. - ϵ_t: White noise error term at time t ($\epsilon_t \sim \text{WN}(0, \sigma^2)$). - $\theta_1, \ldots, \theta_q$: Moving average coefficients.

9.5.2.2 Key Properties

1. Stationarity:

- MA(q) is always stationary (no restrictions on θ coefficients).

2. ACF:

- Spikes at lags $1, 2, \ldots, q$.
- ACF cuts off abruptly after lag q.

3. PACF:

- Decays gradually (exponentially or oscillating).

! Invertibility Property of MA Models

9.5.2.3 Definition

The invertibility property of a Moving Average (MA) model ensures that the model can be expressed equivalently as an infinite-order Autoregressive (AR) model. This property is crucial for the uniqueness of the model parameters and for meaningful interpretation.

9.5.2.4 Mathematical Condition

For an MA(q) model:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

The invertibility condition requires that the roots of the characteristic equation:

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

lie outside the unit circle (i.e., $|z| > 1$).

9.5.2.5 Additional Points

- AR(1) can always be rewritten as MA(∞) if it is stationary.
- AR(p) has an MA(∞) representation under stationarity conditions.
- Invertibility ensures that an MA(q) process has an equivalent AR(∞) representation.
- Both properties are crucial in time series modeling to ensure proper interpretation and forecasting.

9.5.2.6 Example

For an MA(1) model:

$$Y_t = \mu + \epsilon_t + \theta \epsilon_{t-1}$$

The invertibility condition is $|\theta| < 1$. If this condition is violated, the model may not have a meaningful AR representation.

9.6 ARMA(p, q) Model

9.6.1 Definition

The ARMA(p, q) (Autoregressive Moving Average) model combines the AR(p) and MA(q) models to describe a time series using both autoregressive and moving average components.

Formula:

$$(Y_t - \mu) = \phi_1 (Y_{t-1} - \mu) + \phi_2 (Y_{t-2} - \mu) + \cdots + \phi_p \left(Y_{t-p} - \mu\right) + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

or,

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

where: - c: Constant term (intercept). - $\phi_1, \ldots, \phi_p$: Autoregressive coefficients. - $\theta_1, \ldots, \theta_q$: Moving average coefficients. - ϵ_t: White noise error term $\sim \text{WN}(0, \sigma^2)$.

9.6.2 Key Properties

1. Stationarity:

- Requires the AR component to satisfy stationarity conditions (roots of the AR characteristic equation must lie outside the unit circle).

2. Invertibility:

- Requires the MA component to satisfy invertibility conditions (roots of the MA characteristic equation must lie outside the unit circle).

3. ACF and PACF:

- ACF and PACF exhibit a combination of behaviors from AR(p) and MA(q) models:
- ACF may decay gradually or cut off after lag q.
- PACF may decay gradually or cut off after lag p.

The Backwards Operator

- The backshift operator B, also known as the lag operator, is a concise notation useful for describing ARMA and ARIMA models. It is defined as:

$$BY_t = Y_{t-1}$$

- More generally, for any integer h:

$$B^h Y_t = Y_{t-h}$$

- This means that B shifts the time series back by one unit, while B^h shifts it back by h units. Additionally, for any constant c, the operator satisfies:

$$Bc = c$$

since a constant does not vary with time.

- The ARMA model can be represented more compactly using the backward operator as:

$$\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)(Y_t - \mu) = \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right)\epsilon_t$$

The Differencing Operator

- The differencing operator is a mathematical tool used to transform a time series by removing trends. It is defined as:

$$\Delta = 1 - B$$

where B is the backshift operator.

- Applying the differencing operator to a time series Y_t gives:

$$\Delta Y_t = Y_t - BY_t = Y_t - Y_{t-1}$$

- This operation computes the difference between consecutive observations, effectively removing linear trends and stabilizing the mean of the series.

- Differencing can be applied iteratively. For instance, the second-order differencing operator is defined as:

$$\Delta^2 Y_t = \Delta(\Delta Y_t) = \Delta(Y_t - Y_{t-1}) = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}.$$

In general, the kth-order differencing operator, Δ^k, can be expressed using the binomial expansion:

$$\Delta^k Y_t = (1 - B)^k Y_t = \sum_{i=0}^{k} \binom{k}{i} (-1)^i Y_{t-i},$$

where B is the backshift operator, and $\binom{k}{i}$ represents the binomial coefficient.

9.7 ARIMA(p, d, q) Model

9.7.1 Definition

- The ARIMA(p, d, q) (Autoregressive Integrated Moving Average) model is an extension of the ARMA(p, q) model that incorporates differencing to handle non-stationary time series.
- We can also say, a time series Y_t is considered an ARIMA(p, d, q) process if the d-th differenced series, $\Delta^d Y_t$, follows an ARMA(p, q) model.

Formula:
The ARIMA model can be expressed as:

$$\Delta^d Y_t = c + \phi_1 \Delta^d Y_{t-1} + \cdots + \phi_p \Delta^d Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

where: - c: Intercept - p: Order of the autoregressive (AR) component. - d: Degree of differencing (number of times the series is differenced to achieve stationarity). - q: Order of the moving average (MA) component. - Δ^d: Differencing operator applied d times.

9.7.2 Key Components

1. Autoregressive (AR) Component:
   Captures the relationship between an observation and its lagged values.
2. Differencing (I):
   Removes trends and makes the series stationary.

3. Moving Average (MA) Component:
   Models the relationship between an observation and past error terms.

9.7.3 Steps to Build an ARIMA Model

1. Check Stationarity:

- Use visual inspection, or statistical tests (e.g., KPSS test, Augmented Dickey-Fuller test (also known as ADF test is based on unit root test)).
- If non-stationary, apply differencing until stationarity is achieved.
- d is determined by the number of differencing steps required for stationarity.

2. Determine Parameters (p, d, q):

- Usually AIC or BIC are employed to identify p and q.

3. Fit the Model:

- Estimate the parameters using Maximum likelihood estimation.

4. Validate the Model:

- Check residuals for randomness (white noise).
- Use metrics like AIC, BIC, or cross-validation for model selection.

4. Validate the Model:

- Check residuals for randomness (white noise) using diagnostic plots (e.g., ACF/PACF of residuals) and statistical tests (e.g., Ljung-Box test).
- Use metrics like AIC, BIC, or cross-validation for model selection.

5. Forecast:

- Use the fitted model to make predictions.

Unit Roots Test

- Determining whether a time series is stationary or non-stationary can be challenging. Hypothesis testing provides a systematic approach to address this question.

9.7.4 What is a Unit Root?

- A unit root refers to a characteristic of a time series where the value of an autoregressive coefficient equals 1, leading to non-stationarity. Specifically, in an ARMA(p, q) process, the model can be expressed as:

$$(Y_t - \mu) = \phi_1(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) + \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q}$$

- If any of the roots of the characteristic equation for the AR component lie on the unit circle (e.g., $\phi_1 = 1$), the series exhibits a unit root and is non-stationary.
- The stationarity condition for the time series $\{Y_t\}$ requires that all roots of the characteristic polynomial:

$$1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p$$

must have absolute values greater than one (i.e., lie outside the unit circle).

Note

- AR(0), MA(0), ARMA(0,0), and ARIMA(0,0,0) refers to white noise WN(0, σ^2)
- AR(1), ARMA(1,0) and ARIMA(1,0,0) refers to the first order autoregression process.
- Similarly MA(1), ARMA(0,1) and ARIMA (0,0,1) refers to the first order moving average process.
- ARIMA(p, 0, q) model is the same as an ARMA(p, q) model.

9.8 SARIMA Model Equation

The Seasonal AutoRegressive Integrated Moving Average (SARIMA) model extends ARIMA by incorporating seasonal components. It is represented as:

$$SARIMA(p,d,q) \times (P,D,Q,m)$$

where: - (p,d,q) are the non-seasonal parameters: - p: Order of the autoregressive (AR) component. - d: Number of non-seasonal differences. - q: Order of the moving average (MA) component. - (P,D,Q,m) are the seasonal parameters: - P: Order of the seasonal AR component. - D: Number of seasonal differences. - Q: Order of the seasonal MA component. - m: Length of the seasonal cycle.

9.8.1 Full SARIMA Equation

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{P} \Phi_k y_{t-mk} + \sum_{l=1}^{Q} \Theta_l \epsilon_{t-ml} + \epsilon_t$$

Where:

- m is the length of the seasonality (e.g., $m = 12$ for monthly data with annual seasonality).
- P is the seasonal autoregressive order.
- Φ_k are the seasonal autoregressive coefficients.
- Q is the seasonal moving average order.
- Θ_l are the seasonal moving average coefficients.

$$(1 - \sum_{i=1}^{p} \phi_i B^i)(1 - \sum_{k=1}^{P} \Phi_k B^{mk})(1-B)^d(1-B^m)^D y_t = (1 + \sum_{j=1}^{q} \theta_j B^j + \sum_{l=1}^{Q} \Theta_l B^{ml})\epsilon_t$$

where: - B is the backshift operator, $B^k y_t = y_{t-k}$. - ϕ_i and Φ_k are the non-seasonal and seasonal AR coefficients, respectively. - θ_j and Θ_l are the non-seasonal and seasonal MA coefficients, respectively. - d and D represent non-seasonal and seasonal differencing orders. - ϵ_t is the error term.

The SARIMA model effectively captures both trend and seasonality in time series forecasting.

References

Ross, Sheldon. 2009. "Probability and Statistics for Engineers and Scientists." Elsevier, New Delhi 16: 32–33.