MAILAM Engineering College

Approved by AICTE, New Delhi, affiliated to Anna University, Chennai, Accredited by NBA & TCS

AD3491 FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

[REGULATION-2021]

STUDY MATERIAL

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

NAME OF THE STUDENT:……

REGISTER NUMBER:……

YEAR / SEM:……

ACADEMIC YEAR:……

PREPARED BY

➔ **Dr. S. ARTHEESWARI, Prof. & HEAD/AI&DS**

MAILAM
Engineering College
Approved by AICTE, New Delhi, affiliated to Anna University, Chennai, Accredited by NBA & TCS

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

AD3491 FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II Yr. / IV SEM

SYLLABUS

COURSE OBJECTIVES:

- To understand the techniques and processes of data science
- To apply descriptive data analytics
- To visualize data for various applications
- To understand inferential data analytics
- To analysis and build predictive models from data

UNIT I INTRODUCTION TO DATA SCIENCE **08**

Need for data science – benefits and uses – facets of data – data science process – setting the research goal – retrieving data – cleansing, integrating, and transforming data – exploratory data analysis – build the models – presenting and building applications.

UNIT II DESCRIPTIVE ANALYTICS **10**

Frequency distributions – Outliers –interpreting distributions – graphs – averages – describing variability – interquartile range – variability for qualitative and ranked data - Normal distributions – z scores –correlation – scatter plots – regression – regression line – least squares regression line – standard error of estimate – interpretation of r^2 – multiple regression equations – regression toward the mean.

UNIT III INFERENTIAL STATISTICS **09**

Populations – samples – random sampling – Sampling distribution- standard error of the mean - Hypothesis testing – z-test – z-test procedure –decision rule – calculations – decisions – interpretations - one-tailed and two-tailed tests – Estimation – point estimate – confidence interval – level of confidence – effect of sample size.

UNIT IV ANALYSIS OF VARIANCE **09**

t-test for one sample – sampling distribution of t – t-test procedure – t-test for two independent samples – p-value – statistical significance – t-test for two related samples.

F-test – ANOVA – Two-factor experiments – three f-tests – two-factor ANOVA – Introduction to chi-square tests.

UNIT V PREDICTIVE ANALYTICS **09**

Linear least squares – implementation – goodness of fit – testing a linear model – weighted resampling. Regression using StatsModels – multiple regression – nonlinear relationships – logistic regression – estimating parameters – Time series analysis – moving averages – missing values – serial correlation – autocorrelation. Introduction to survival analysis.

TOTAL: 45 PERIODS

COURSE OUTCOMES:

CO1: Explain the data analytics pipeline

CO2: Describe and visualize data

CO3: Perform statistical inferences from data

CO4: Analyze the variance in the data

CO5: Build models for predictive analytics

TEXT BOOKS

1. David Cielen, Arno D. B. Meysman, and Mohamed Ali, "Introducing Data Science", Manning Publications, 2016. (first two chapters for Unit I).
2. Robert S. Witte and John S. Witte, "Statistics", Eleventh Edition, Wiley Publications, 2017.
3. Jake VanderPlas, "Python Data Science Handbook", O'Reilly, 2016.

REFERENCES

1. Allen B. Downey, "Think Stats: Exploratory Data Analysis in Python", Green Tea Press, 2014.
2. Sanjeev J. Wagh, Manisha S. Bhende, Anuradha D. Thakare, "Fundamentals of Data Science", CRC Press, 2022.
3. Chirag Shah, "A Hands-On Introduction to Data Science", Cambridge University Press, 2020.
4. Vineet Raina, Srinath Krishnamurthy, "Building an Effective Data Science Practice: A Framework to Bootstrap and Manage a Successful Data Science Practice", Apress,2021.

Prepared by –
Dr. S. Artheeswari Prof. & Head

PRINCIPAL

UNIT I – INTRODUCTION TO DATA SCIENCE

SYLLABUS:

Need for data science – benefits and uses – facets of data – data science process – setting the research goal – retrieving data – cleansing, integrating, and transforming data – exploratory data analysis – build the models – presenting and building applications.

PART A

1. What is Bigdata?

- Big data is a ***huge volume, high velocity and variety of data*** that cannot be processed by traditional processing system.
- They are characterized by the 7 Vs: ***velocity, variety, volume, variability, visualization, value and veracity.***

2. What are the Characteristics of Bigdata?

- Velocity - refers to the speed of data processing
- Volume - refers to the amount of data
- Value - refers to the benefits that the organization derives from the data.
- Variety - refers to the different types of big data.
- Veracity - refers to the accuracy of your data.
- Validity – refers to the relevance of data for the intended purpose.
- Volatility – refers to constantly changing
- Visualization - Visualization refers to showing your big data-generated insights
- through visual representations such as charts and graphs.

3. Define Data Science.

- Data science is the field of study of data, using modern scientific techniques, statistical methods and algorithms to derive insights from huge volume of data and to create business and IT strategies.
- It deals about where the data comes from, what it represents, and the ways by which it can be transformed into valuable inputs and resources

4. What are the benefits and uses of Bigdata

- Commercial Companies
- Human Resource professionals
- Financial institutions
- Governmental organizations
- Nongovernmental organizations (NGOs)
- Universities

5. List out the Facets of data.

The facets of data are categorized below,

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

6. Define Structured data.

- Structured data is data that ***depends on a data model*** and ***resides in a fixed field*** within a record.
- It's ***easy to store*** structured data in tables within databases or Excel files.
- SQL, or Structured Query Language, is the preferred way to manage and query data that resides in databases.
- **Example:** Excel files

7. Define unstructured data

- Unstructured data is data **that isn't easy to fit into a data model** because the content is context-specific or varying.
- **Example:** Email

8. What is Machine Generated Data?

- Machine-generated data is information that's automatically created by a computer, process, application, or other machine without human intervention.
- The analysis of machine data relies on highly scalable tools, due to its high volume and speed.
- **Examples:** web server logs, call detail records, network event logs, and telemetry

9. What is Streaming Data?

- The data flows into the system in a continuous manner when an event happens instead of being loaded into a data store in a batch.
- **Examples** - "What's trending" on Twitter, live sporting or music events, and the stock market.

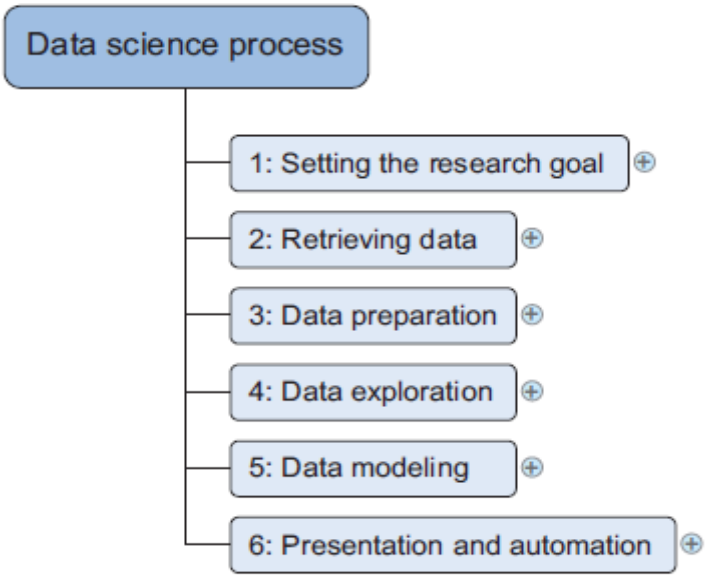
10. Define Graph based or Network data

- "Graph" points to mathematical graph theory.

- In graph theory, a graph is a ***mathematical structure to model pair-wise relationships between objects.***
- Graph or network data is, a data that focuses on the ***relationship or adjacency of objects.***
- The graph structures use nodes, edges, and properties to represent and store graphical data.
- **Graph databases** are used to store graph-based data and are queried with specialized **query languages such as SPARQL**.
- **Example: social media websites**
  - For instance, on **LinkedIn** you can see who you know at which company.
  - Your follower list on **Twitter** is another example of graph-based data.

11. List out the steps in Data Science Process

The data science process typically consists of six steps.



12. What is meant by Project Charter?

- All the information which are related to research goal is best collected in a project charter.
- A project charter requires teamwork, and input covers at least the following:
  - A clear research goal
  - The project mission and context
  - How to perform analysis
  - What resources to use
  - Proof that it's an achievable project, or proof of concepts
  - Deliverables and a measure of success
  - A timeline

13. How to retrieving the data in Data Science process?

- The second step is to ***collect data*** by finding suitable data and getting access to the data from the ***data owner***.
- Data can also be delivered by ***third-party companies*** and take many forms ranging from Excel spreadsheets to different types of databases.
- The result is data in its ***raw form***, which probably needs polishing and transformation before it becomes usable.

14. What is Data Repositories?

- A data repository is also known as a **data library or data archive**.
- The **data repository** is a large database infrastructure — several databases — that collect, manage, and store data sets for data analysis, sharing and reporting.
- Example: Database, Data Warehouse, Data mart, Data Lake.

15. Difference between Data Marts and Data warehouse.

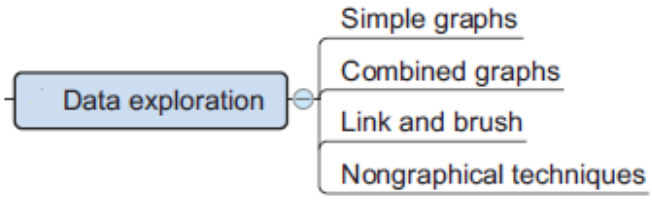
Data Warehouse	Data Mart
Data Warehouse stores a large amount of data which is collected from different sources	Data Mart contains only the specific data from data warehouse, which is required by the company for analysis
Data Warehouse is focused on all departments in an organization	Data Mart focuses on a specific group.
Data Warehouse designing process is complicated	Data Mart process is easy to design.
Data Warehouse takes a long time for data handling	Data Mart takes a short time for data handling.
Data Warehouse size range is 100 GB to 1 TB+	Data Mart size is less than 100 GB.

16. Define Data Lake.

- A **data lake** is a large data repository that stores unstructured data that is classified and tagged with metadata.

17. What is Exploratory Data Analysis (EDA)?

- Data exploration is concerned with building a ***deeper understanding*** of the data to know ***how variables interact with each other***, the distribution of the data, and whether there are outliers.



18. Define Data Modeling.

- Building a model is an **iterative process** that involves selecting the variables for the model, executing the model, and model diagnostics.
- Models consist of the following main steps:
  - Selection of a modeling technique and variables to enter in the model
  - Execution of the model
  - Diagnosis and model comparison

19. Define linking and brushing technique.

- With brushing and linking can ***combine and link different graphs and tables***
  so changes in one graph are automatically transferred to the other graphs.

20. What is Histogram and Boxplot?

- In a **histogram** a variable is cut into discrete categories and the number of occurrences in each category are summed up and shown in the graph.
- The **boxplot**, doesn't show how many observations are present but does offer an impression of the distribution within categories.
- It can show the maximum, minimum, median, and other characterizing measures at the same time.

21. Define Presentation and automation steps in Data Science process.

- Finally ***presenting the results*** to the business.
- These results can take many forms, ranging from presentations to research reports.
- Sometimes need to ***automate the execution*** of the process because the business will use the insights gained in another project or enable an operational process to use the outcome from the model.

22. Discuss the three sub-phases of Data preparation.

- This includes transforming the data from a raw form into data that's directly usable in your models.

- This phase consists of three sub-phases:
  - **i)** ***Data cleansing*** removes false values from a data source and inconsistencies across data sources,
  - **ii)** ***Data integration*** enriches data sources by combining information from multiple data sources, and
  - **iii)** ***Data transformation*** ensures that the data is in a suitable format for use in your models.

23. Define common errors that occur during cleansing data.

Error description	Possible solution
Errors pointing to false values within one data set	
Mistakes during data entry	Manual overrules
Redundant white space	Use string functions
Impossible values	Manual overrules
Missing values	Remove observation or value
Outliers	Validate and, if erroneous, treat as missing value (remove or insert)
Errors pointing to inconsistencies between data sets	
Deviations from a code book	Match on keys or else use manual overrules
Different units of measurement	Recalculate
Different levels of aggregation	Bring to same level of measurement by aggregation or extrapolation

24. Define outlier.

- An outlier is an observation that seems to be distant from other observations or, more specifically, one observation that follows a different logic or generative process than the other observations.
- The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.

PART B

1. Give the description about data science and its applications, also discuss the benefits and uses of Data Science and Big Data.

Contents

- Big Data
- Data Science
- Benefits and Uses:
  1. Commercial Companies
  2. Human Resource Professionals
  3. Financial Institutions
  4. Government Organizations
  5. Non-governmental organizations (NGOs)
  6. Universities
- Data Science Tools
- Real Time Applications of Data Science

Data

- Data is a ***collection of discrete states*** that convey information, describing quantity, quality, fact and statistics.

Big data

- Big data is a ***huge volume, high velocity and variety of data*** that cannot be processed by traditional processing system.
- They are characterized by the 7 Vs: ***velocity, variety, volume, variability, visualization, value and veracity.***

Data science

- Data science is the field of ***study of data***, using ***modern scientific techniques, statistical methods*** *and* ***algorithms*** to derive ***insights*** from huge volume of data and to create business and IT strategies.
- It deals about ***where*** the data comes from, ***what*** it represents, and the ***ways*** by which it can be transformed into ***valuable inputs and resources***

Benefits and uses of data science

1. Commercial Companies

- ***Commercial companies*** use data science to gain insights into their customers, processes, staff, completion, and products.

- Many companies use data science to offer customers a better user experience, cross-sell, up-sell, and personalize their offerings.
- **Example:**
  - ***Google AdSense*** - collects data from internet users so relevant commercial messages can be matched to the person browsing the internet.
  - ***MaxPoint*** - example of real-time personalized advertising.

2. Human Resource Professionals

- ***Human resource professionals*** use people analytics and text mining to screen candidates, monitor the mood of employees, and study informal networks among co-workers.

3. Financial Institutions

- ***Financial institutions*** use data science to predict stock markets, determine the risk of lending money, and learn how to attract new clients for their services.

4. Government Organizations

- ***Governmental organizations*** are also aware of data's value.
- **Example:**
  - ***Data.gov*** is the home of the US Government's open data.

5. Non-governmental organizations (NGOs)

- ***Non-governmental organizations (NGOs)*** use it to raise money and defend their causes.

- **Example:**
  - The ***World Wildlife Fund (WWF),*** employs data scientists to increase the effectiveness of their fund raising efforts.
  - ***DataKind*** is a data scientist group that devotes it's time to the benefit of mankind.

6. Universities

- ***Universities*** use data science in their research to enhance the study experience of their students.
- **Example:**
  - The rise of ***massive open online courses (MOOC)*** produces a lot of data, which allows universities to study.
  - ***Coursera, Udacity,*** *and* ***edX****.*

Data Science Tools

1. SAS	-	processing Statistical operations
2. Apache Spark	-	handles batch processing and stream processing
3. BigML	-	processing Machine Learning Algorithms
4. MATLAB	-	processing Mathematical Information
5. Tableau	-	Data Visualization Software

6. Jupyter	-	Used for writing code in Python.
7. MatplotLib	-	Library for plotting and visualization in python.
8. NLTK	-	Natural Language Processing
9. Tensor flow	-	Machine Learning Algorithm
10. Numpy	-	Numerical python for Data Analysis
11. Scipy	-	Scientific python for scientific and technical Computations
12. Pandas	-	Used for Data Analysis

Real Time Applications of Data Science

- Fraud and Risk Detection
- Healthcare
  - Medical Image Analysis
  - Medical Drug Development
  - Virtual Assistance for patients and customer support
- Internet Search
- Target Advertising
- Website Recommendation
- Speech Recognition
- Gaming
- Augmented Reality
- Robotics

2. List and explain the facets of data or different types of data or categories of data.

Contents

1. Structured
2. Unstructured
3. Natural Language
4. Machine-generated
5. Graph-based
6. Audio, video, and images
7. Streaming

➢ **Categories of data:**

1. Structured data

- Structured data is data that ***depends on a data model*** and ***resides in a fixed field*** within a record.
- It's ***easy to store*** structured data in tables within databases or Excel files.

- SQL, or Structured Query Language, is the preferred way to manage and query data that resides in databases.

  Example: Refer Figure 1.1

	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843	White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
19	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Figure 1.1 An Excel table is an example of structured data.

2. Unstructured data

- Unstructured data is data that isn't easy to fit into a data model because the content is ***context-specific or varying***.
- **Example** - regular email. (Figure 1.2).

Figure 1.2 Email is simultaneously an example of unstructured data and natural language data.

- In Figure 1.2, email contains structured elements such as the sender, title, and body text, it's a challenge to find the number of people who have written an email complaint about a specific employee because so many ways exist to refer to a person, for example.

3. Natural language

- Natural language is a ***special type of unstructured data***; it's challenging to process because it ***requires knowledge of specific data science techniques*** and linguistics.
- The natural language processing community had success in entity recognition, topic recognition, summarization, text completion, and sentiment analysis, but models trained in one domain don't generalize well to other domains.

4. Machine-generated data

- Machine-generated data is ***information that's automatically created*** by a computer, process, application, or other machine without human intervention.
- The analysis of machine data relies on highly scalable tools, due to its high volume and speed.
- **Examples** - web server logs, call detail records, network event logs, and telemetry (Figure 1.3).

```
trace:
CSIPERF:TXCOMMIT;273983
2014-11-28 11:36:13, Info                  CSI    00000157 Creating NT transaction (seq
70), objectname [6]"(null)"
2014-11-28 11:36:13, Info                  CSI    00000158 Created NT transaction (seq 70)
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:13, Info                  CSI    00000159@2014/11/28:10:36:13.764
Beginning NT transaction commit...
2014-11-28 11:36:14, Info                  CSI    0000015a@2014/11/28:10:36:14.094 CSI perf
trace:
CSIPERF:TXCOMMIT;386259
2014-11-28 11:36:14, Info                  CSI    0000015b Creating NT transaction (seq
71), objectname [6]"(null)"
2014-11-28 11:36:14, Info                  CSI    0000015c Created NT transaction (seq 71)
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:14, Info                  CSI    0000015d@2014/11/28:10:36:14.106
Beginning NT transaction commit...
2014-11-28 11:36:14, Info                  CSI    0000015e@2014/11/28:10:36:14.428 CSI perf
trace:
CSIPERF:TXCOMMIT;375581
```

Figure 1.3 Example of machine-generated data

- The machine data in figure 1.3 would fit nicely in a classic table-structured database.
- This isn't the best approach for highly interconnected or "networked" data, where the relationships between entities have a valuable role to play.

5 Graph-based or network data

- "Graph" points to mathematical graph theory.
- In graph theory, a graph is a ***mathematical structure to model pair-wise relationships between objects.***
- Graph or network data is, a data that focuses on the ***relationship or adjacency of objects.***
- The graph structures use nodes, edges, and properties to represent and store graphical data.

- Graph-based data is a natural way ***to represent social networks***, and its structure allows to calculate specific metrics such as the influence of a person and the shortest path between two people.
- **Example**: graph-based data can be found on many social media websites such as Follower list on Twitter. (figure 1.4).

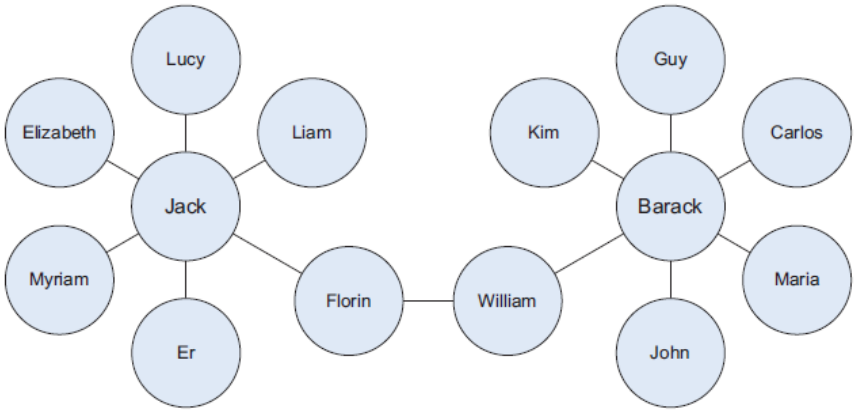


Figure 1.4 Friends in a social network are an example of graph-based data.

- Graph databases are used to store graph-based data and are queried with specialized query languages such as SPARQL.

6. Audio, image, and video

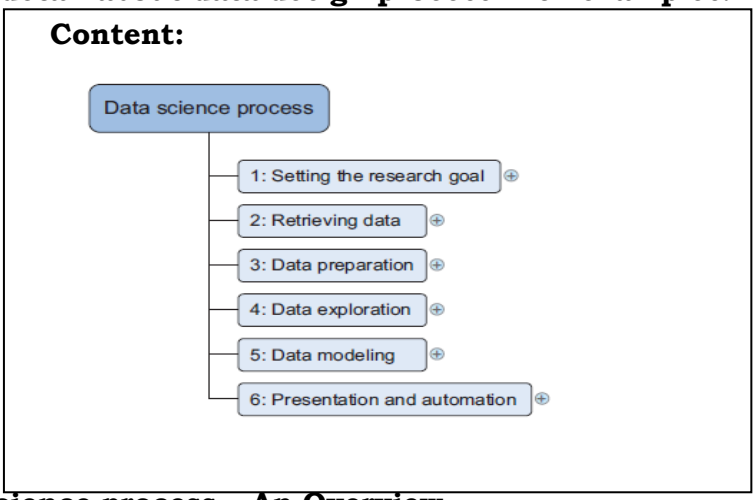
- Audio, image, and video are data types that pose ***specific challenges*** to a data scientist.
- Tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers.
- High-speed cameras at stadiums will capture ball and athlete movements to calculate in real time, for example, the path taken by a defender relative to two baselines.
- Recently a company called DeepMind succeeded at creating an algorithm that's capable of learning how to play video games.
- This algorithm takes the video screen as input and learns to interpret everything via a complex process of deep learning.
- This prompted Google to buy the company for their own Artificial Intelligence (AI) development plans.

7. Streaming data

- The data flows into the system in a continuous manner when an event happens instead of being loaded into a data store in a batch.
- **Examples** - "What's trending" on Twitter, live sporting or music events, and the stock market.

3 Explain in detail about data design process with examples.

Content:



- **The data science process – An Overview**

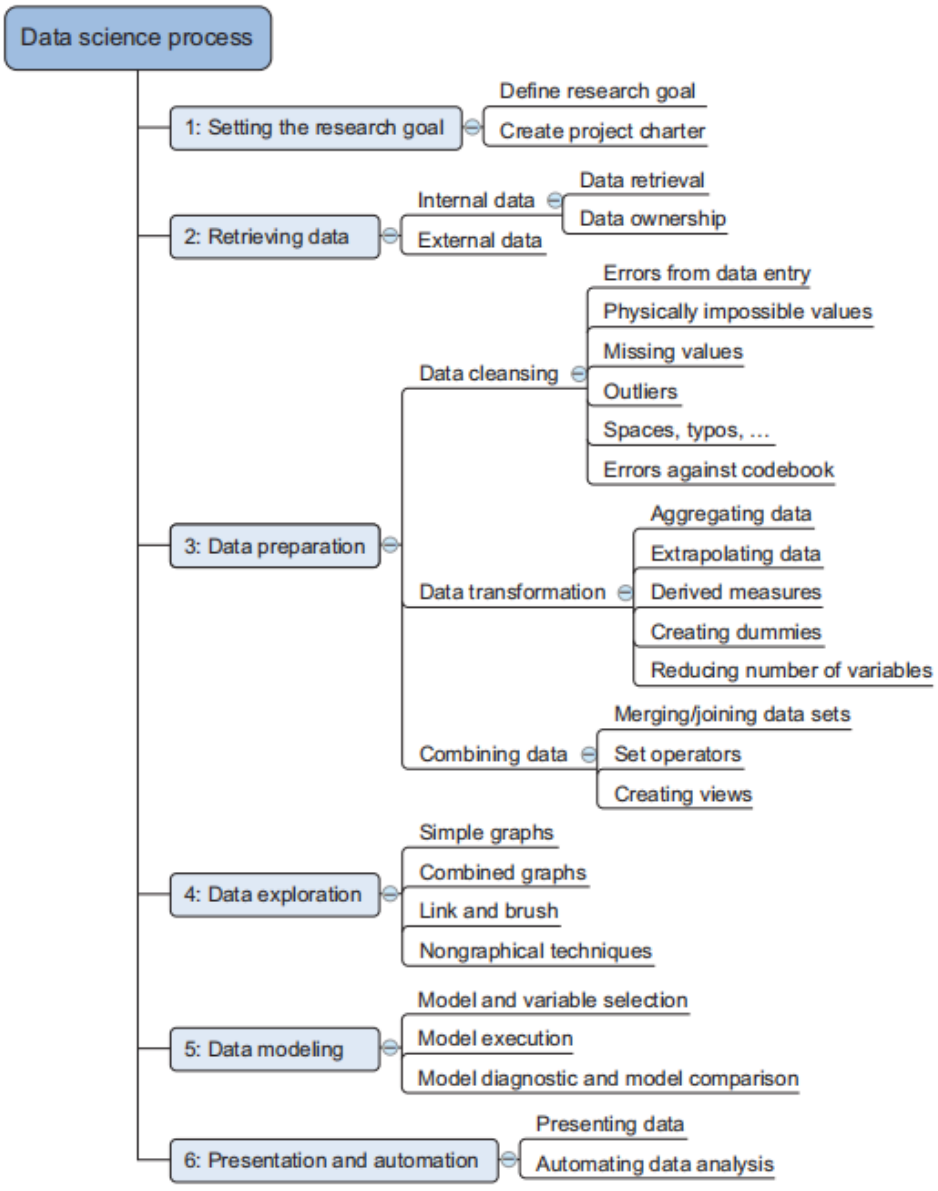
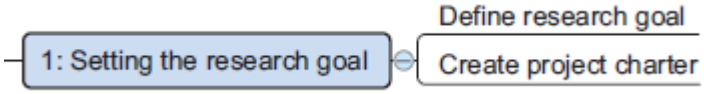


Figure 1.5: Steps of Data Science Process

- The data science process typically consists of ***six steps***, as shown in figure 1.5

1. Setting the research goal



- The first step of this process is ***defining a research goal*** by creating a ***project charter***.
- A project charter requires teamwork, and input covers at least the following:
  - A clear research goal
  - The project mission and context
  - How to perform analysis
  - What data and resources to use
  - Proof that it's an achievable project, or proof of concepts
  - Deliverables and a measure of success
  - A timeline

2 Retrieving data

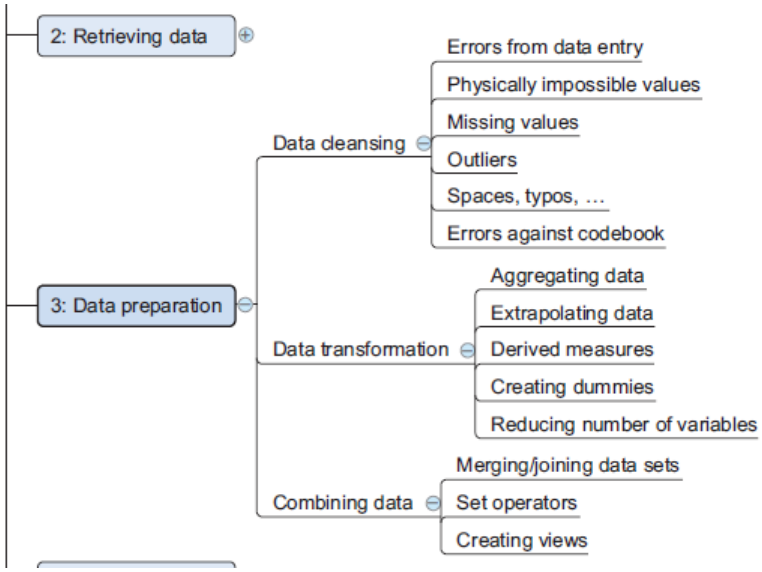
- The second step is to ***collect data*** by finding suitable data and getting access to the data from the ***data owner***.
  - *Start with data stored within the company*
    - The data can be stored in official data repositories such as databases, data marts, data warehouses, and data lakes maintained by a team of IT professionals.
    - The primary goal of a database is data storage, while a data warehouse is designed for reading and analyzing that data.
    - A data mart is a subset of the data warehouse and geared toward serving a specific business unit.
    - While data warehouses and data marts are home to preprocessed data, data lakes contains data in its natural or raw format which probably needs polishing and transformation before it becomes usable..
  - *Don't be afraid to shop around*
    - Many companies specialize in collecting valuable information.
    - Data can also be delivered by third-party companies and take many forms ranging from Excel spreadsheets to different types of databases. Refer Table 1.2

Table 1.2 – Open Data Sites

Open data site	Description
Data.gov	The home of the US Government's open data
https://open-data.europa.eu/	The home of the European Commission's open data
Freebase.org	An open database that retrieves its information from sites like Wikipedia, MusicBrains, and the SEC archive
Data.worldbank.org	Open data initiative from the World Bank
Aiddata.org	Open data for international development
Open.fda.gov	Open data from the US Food and Drug Administration

- *Do data quality checks to prevent problems later*
  - Expect to spend a good portion of your project time doing data correction and cleansing, sometimes up to 80%.

3 Data preparation



- Data collection is an error-prone process; this phase ***enhance the quality of the data and prepare it for use*** in subsequent steps.
- This phase consists of three sub-phases:
  1. **Data cleansing -** Data cleansing is a sub process of the data science that removes false values from a data source and inconsistencies across data sources,.

     Types of errors

     - *Interpretation error* – Taking value for granted.
       Example: person's age is greater than 300 years
     - *Inconsistencies* – between data sources and standardized value.
       Example: putting "Female" in one table and "F" in another

Common Errors

Table 1.3 – Common Errors

Error description	Possible solution
Errors pointing to false values within one data set	
Mistakes during data entry	Manual overrules
Redundant white space	Use string functions
Impossible values	Manual overrules
Missing values	Remove observation or value
Outliers	Validate and, if erroneous, treat as missing value (remove or insert)
Errors pointing to inconsistencies between data sets	
Deviations from a code book	Match on keys or else use manual overrules
Different units of measurement	Recalculate
Different levels of aggregation	Bring to same level of measurement by aggregation or extrapolation

1. Data Entry Errors

- Data collection and data entry are error-prone processes.
- They often require human intervention, and because humans are only human, they make typos or lose their concentration for a second and introduce an error into the chain.
- Example

```
if x == "Godo":
        x = "Good"
if x == "Bade":
        x = "Bad"
```


2. Redundant Whitespace

Whitespaces tend to be hard to detect but cause errors. Fixing redundant whitespaces is luckily easy enough in most programming languages. They all provide string functions that will remove the leading and trailing whitespaces.

Example:

Python the strip() function is used to remove leading and trailing spaces.

3. Impossible Values And Sanity Checks

Sanity checks are another valuable type of data check.

Example:

Sanity checks can be directly expressed with rules:

check = 0 <= age <= 120

4. Outliers

An outlier is an observation that seems to be distant from other observations or, more

specifically, one observation that follows a different logic or generative process than
the other observations. The easiest way to find outliers is to use a plot or a table with
the minimum and maximum values.
An example is shown in figure 1.6.

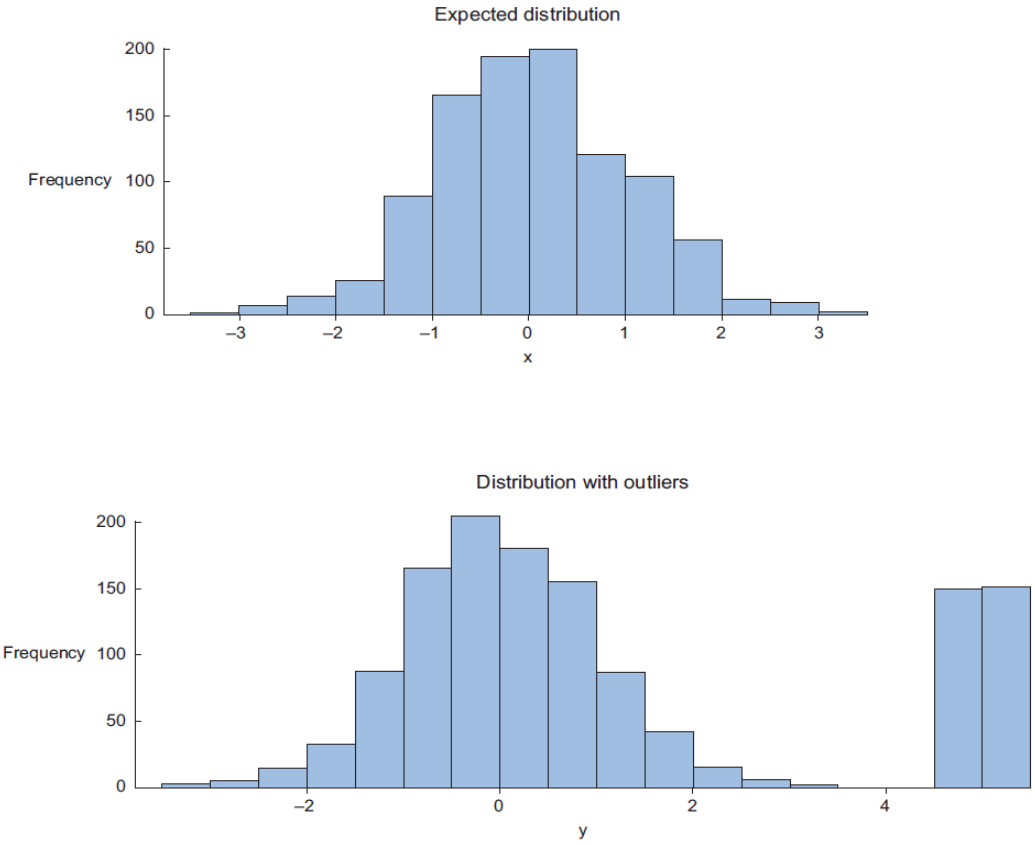


Figure 1.6 Distribution plots are helpful in detecting outliers and helping you understand the variable.

5. Dealing With Missing Values

Missing values aren't necessarily wrong, but still need to handle them separately;

Table 1.4 An overview of techniques to handle missing data

Technique	Advantage	Disadvantage
Omit the values	Easy to perform	You lose the information from an observation
Set value to `null`	Easy to perform	Not every modeling technique and/or implementation can handle `null` values
Impute a static value such as 0 or the mean	Easy to perform You don't lose information from the other variables in the observation	Can lead to false estimations from a model
Impute a value from an estimated or theoretical distribution	Does not disturb the model as much	Harder to execute You make data assumptions
Modeling the value (nondependent)	Does not disturb the model too much	Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute You make data assumptions

6. Deviations From A Code Book

- Detecting errors in larger data sets against a code book or against standardized values can be done with the help of set operations.
- A code book is a description of your data, a form of metadata.
- It contains things such as the number of variables per observation, the number of observations, and what each encoding within a variable means.
- (For instance "0" equals "negative", "5" stands for "very positive".)
- A code book also tells the type of data looking at: is it hierarchical, graph, something else

7. Different Units Of Measurement

- When integrating two data sets, should pay attention to their respective units of measurement.
- An example of this would be when studying the prices of gasoline in the world, gather data from different data providers.
- Data sets can contain prices per gallon and others can contain prices per liter.
- A simple conversion will do the trick in this case.

8. Different Levels Of Aggregation

- Having different levels of aggregation is similar to having different types of measurement.

- An example of this would be a data set containing data per week versus one containing data per work week.
- This type of error is generally easy to detect, and summarizing (or the inverse, expanding) the data sets will fix it.
- After cleaning the data errors, combine information from different data sources.

Correct errors as early as possible

- Data should be cleansed when acquired for many reasons:
  - Decision-makers may make costly mistakes on information based on incorrect data from applications that fail to correct for the faulty data.
  - If errors are not corrected early on in the process, the cleansing will have to be done for every project that uses that data.
  - Data errors may point to a business process that isn't working as designed.
  - Data errors may point to defective equipment, such as broken transmission lines and defective sensors.
  - Data errors can point to bugs in software or in the integration of software that may be critical to the company.

Combining data from different data sources

The different ways of combining data

- The first operation is joining: enriching an observation from one table with information from another table.
- The second operation is appending or stacking: adding the observations of one table to those of another table.

1. Joining Tables

- Joining tables allows to combine the information of one observation found in one table with the information that found in another table.
- To join tables, use variables that represent the same object in both tables, such as a date, a country name,.
- These common fields are known as keys.
- When these keys also uniquely define the records in the table they are called primary keys

Example:

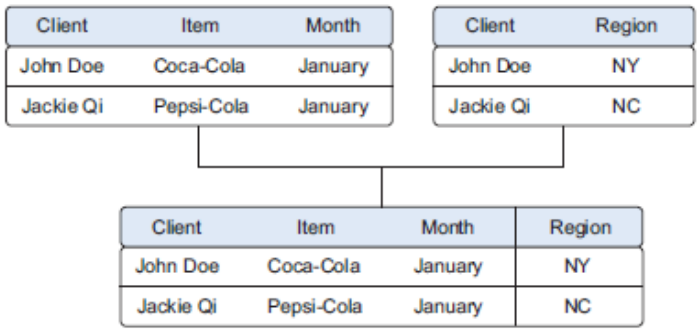


Figure 1.6 : Joining two tables on the Item and Region keys

In figure 1.6, both tables contain the client name, and this makes it easy to enrich the client expenditures with the region of the client.

2. Appending or stacking:

- Appending or stacking tables is effectively adding observations from one table to another table.
- The equivalent operation in set theory would be the union, and this is also the command in SQL, the common language of relational databases.
- Other set operators are also used in data science, such as set difference and intersection.

Example:

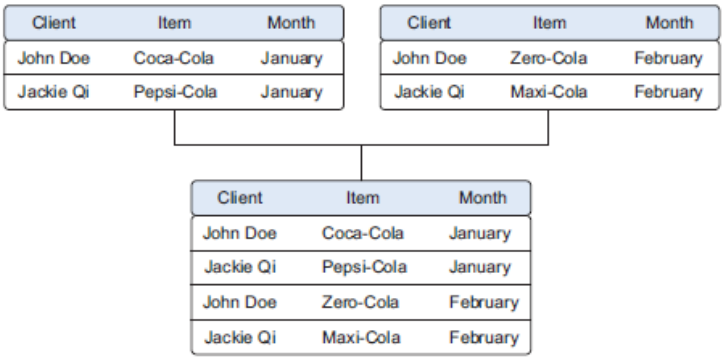


Figure 1.7: Appending tables

In figure 1.7, Appending data from tables is a common operation but requires an equal structure in the tables being appended.

3. *View*

- Views are kind of ***virtual tables***.
- Can create a view by selecting fields from one or more tables present in the database.
- A View can either have all the rows of a table or specific rows based on certain condition.

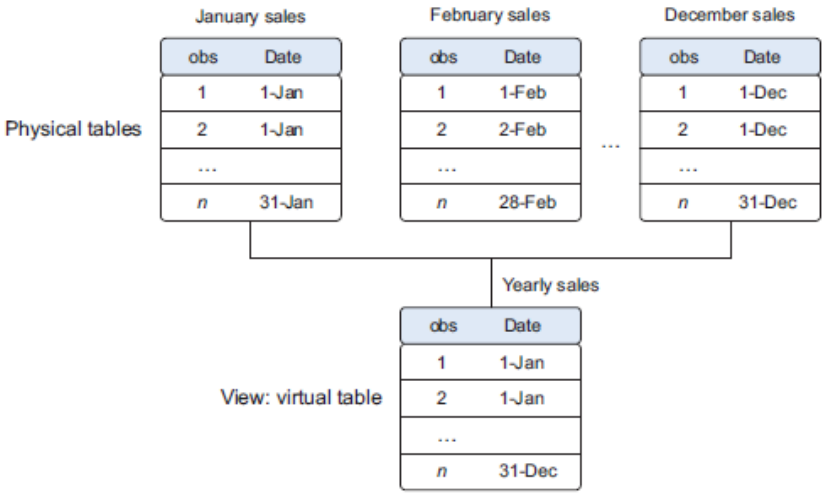


Figure 1.8: Views

4 Data transformation

- Certain models require their data to be in a certain shape.
- Ensures that the data is in a suitable format for use in data models.
- Taking the log of the independent variables simplifies the estimation problem dramatically.

Example – Refer Figure 1.9

Relationships between an input variable and an output variable aren't always linear.

x	1	2	3	4	5	6	7	8	9	10
log(x)	0.00	0.43	0.68	0.86	1.00	1.11	1.21	1.29	1.37	1.43
y	0.00	0.44	0.69	0.87	1.02	1.11	1.24	1.32	1.38	1.46

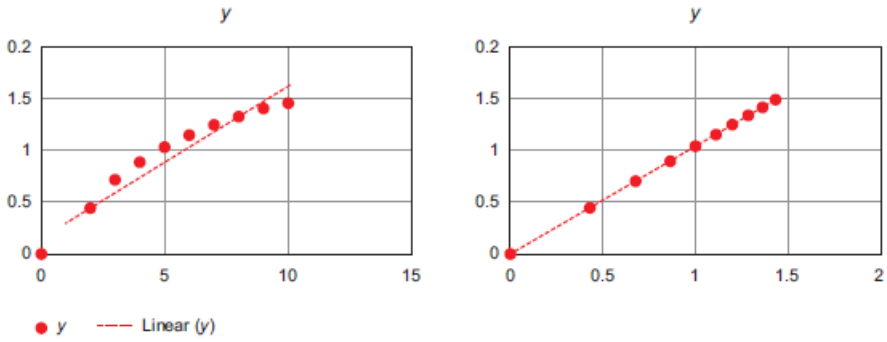
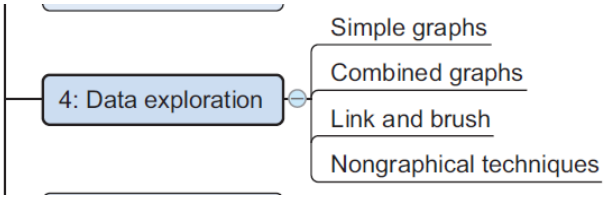


Figure 1.9: Transformation

Figure 1.9 Transforming x to log x makes the relationship between x and y linear (right), compared with the non-log x (left).

5. Data exploration or EDA (Exploratory Data Analysis)

- Data exploration is concerned with building a ***deeper understanding*** of the data to know how variables interact with each other, the distribution of the data, and whether there are outliers.
- The visualization techniques used in this phase range from simple line graphs or histograms, to more complex diagrams such as Sankey and network graphs.



- **Graphs:** - **Simple and Combined Graphs**

  In figure 1.11 - From top to bottom, a ***bar chart, a line plot, and a Distribution*** is some of the graphs used in exploratory analysis.

- **Brushing and linking**.

  With brushing and linking can ***combine and link different graphs and tables*** so changes in one graph are automatically transferred to the other graphs.

Figure 1.11 - Graphs used in exploratory analysis

Histogram

- In a histogram a variable is cut into discrete categories and the number of occurrences in each category are summed up and shown in the graph.

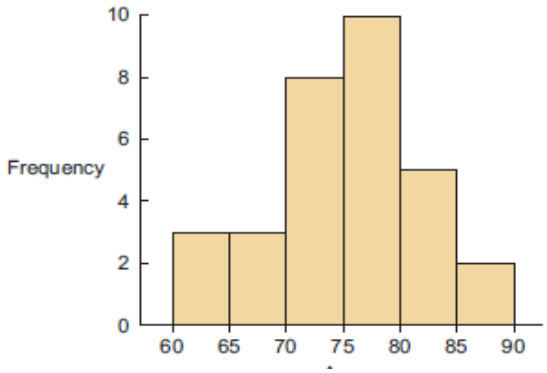


Figure 1.12 - Example Histogram

- Example – Figure 1.12 shows the number of people in the age groups of 5-year intervals

The boxplot

- The boxplot, offers an impression of the distribution within categories.
- It can ***show the maximum, minimum, median***, and other characterizing measures at the same time.

Example:

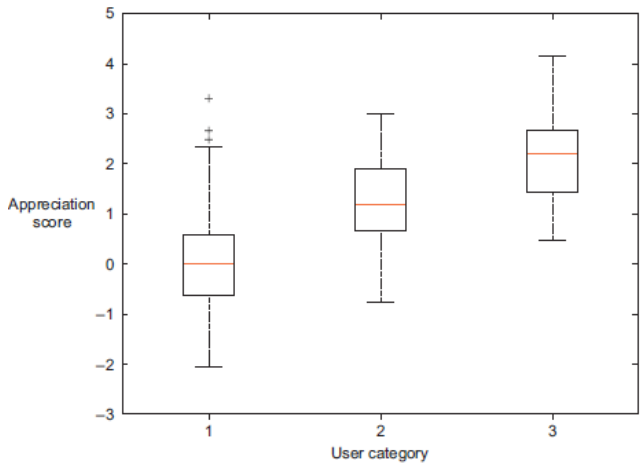
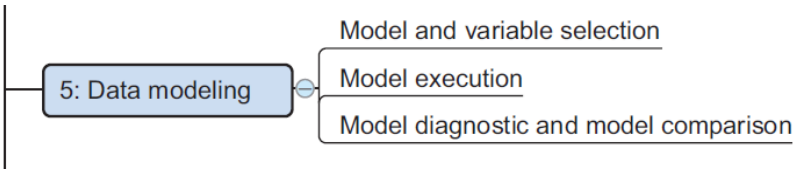


Figure 1.13 - Boxplot

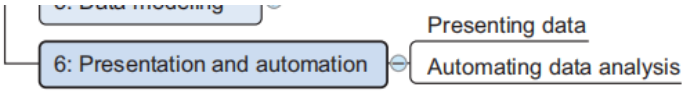
In figure 1.13 each user category has a distribution of the appreciation each has for a certain picture on a photography website.

6 Data modeling or model building



- Building a model is an iterative process that involves selecting the ***variables for the model, executing the model, and model diagnostics***.
- Models consist of the following main steps:
  1. Selection of a modeling technique and variables to enter in the model
  2. Execution of the model
  3. Diagnosis and model comparison

7 Presentation and automation



- Finally ***presenting the results*** to the business.

- These results can take many forms, ranging from presentations to research reports.
- Sometimes need to ***automate the execution*** of the process because the business will use the insights gained in another project or enable an operational process to use the outcome from the model.

UNIT 2 – DESCRIPTIVE ANALYTICS

SYLLABUS:

Frequency distributions – Outliers –interpreting distributions – graphs – averages – describing variability – interquartile range – variability for qualitative and ranked data - Normal distributions – z scores – correlation – scatter plots – regression – regression line – least squares regression line – standard error of estimate – interpretation of r^2 – multiple regression equations – regression toward the mean.

PART A

1. What is Statistics?

- Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data.

2. What are the categories of Statistics? Define it.

- **Descriptive statistics**
  Statistics provides tools—tables, graphs, averages, ranges, correlations—for organizing and summarizing the inevitable variability in collections of actual observations or scores.
- **Inferential statistics.**
  Statistics provides tools—a variety of tests and estimates—for generalizing beyond collections of actual observations.

3. Define Populations and Samples.

- Population refers to any complete collection of observations or potential observations.
- Sample refers to any smaller collection of actual observations drawn from a population

4. What is Random Sampling in Statistics?

- Random sampling is a procedure designed to ensure that each potential observation in the population has an equal chance of being selected in a survey.

5. What is Random Assignment in Statistics?

- Random Assignment is a procedure designed to ensure that each person has an equal chance of being assigned to any group in an experiment.

6. Define Data in Statistics.

- A collection of actual observations or scores in a survey or an experiment

7. What are the types of data in statistical analysis?

- Qualitative data
- Ranked data
- Quantitative data

8. Define Qualitative data

- A set of observations where any single observation is a word, letter, or numerical code that represents a class or category.

9. Define Ranked data

- A set of observations where any single observation is a number that indicates relative standing

10. Define Ranked data

- A set of observations where any single observation is a number that represents an amount or a count.

11. What are Levels of Measurement?

- Levels of measurement specify the extent to which a number (or word or letter) actually represents some attribute and, therefore, has implications for the appropriateness of various arithmetic operations and statistical procedures.
- There are three levels of measurement
    - nominal
    - ordinal
    - interval/ratio

12. What is nominal measurement?

- Nominal measurement is classification—that is, sorting observations into different classes or categories.
- Words, letters, or numerical codes reflect only differences in kind, not differences in amount.
- Examples of nominal measurement include classifying mood disorders as manic, bipolar, or depressive.

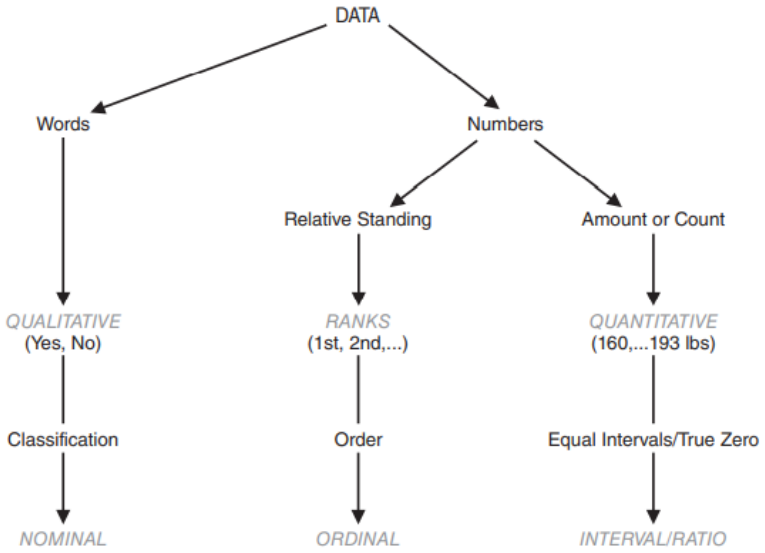
13. What is Ordinal measurement?

- Ordinal measurement is order.
- The relative standing of ranked data that reflects differences in degree based on the order.
- For example, it's inappropriate to conclude that the arithmetic means of ranks 1 and 3 equals rank 2, since this assumes that the actual distance between ranks 1 and 2 equals the distance between ranks 2 and 3.

14. What is Interval/Ratio measurement?

- Interval/ratio measurement is equal intervals and a true zero.
- Amounts or counts of quantitative data reflect differences in degree based on equal intervals and a true zero.
- For example, a reading of 0 on the Fahrenheit temperature scale does not reflect the complete absence of heat—that is, the absence of any molecular motion.

15. Give the Pictorial representation of types of data and levels of measurement.



16. What is a Variable in statistical analysis?

- A variable is a characteristic or property that can take on different values.

17. Define Constant in statistical analysis.

- A Constant is a characteristic or property that can take on only one value.

18. What is meant by Discrete and Continuous Variables?

- A discrete variable consists of isolated numbers separated by gaps.
- Examples include most counts, such as the number of children in a family.
- A continuous variable consists of numbers whose values, at least in theory, have no restrictions.
- Examples include amounts, such as weights of male statistics students.

19. What is meant by Independent and Dependent Variables?

- **Independent Variable**
  - An independent variable is a treatment manipulated by the investigator.

- **Dependent Variable**
  - A dependent Variable is a variable that is believed to have been influenced by the independent variable.

20. What is frequency distribution and give usage?

- A frequency distribution is a collection of observations produced by sorting observations into classes and showing their frequency (f) of occurrence in each class.
- A frequency distribution helps us to detect any pattern in the data (assuming a pattern exists) by superimposing some order on the inevitable variability among observations.

21. Give guidelines for frequency distributions rules.

Essential

1. Each observation should be included in one, and only one, class.
2. List all classes, even those with zero frequencies.
3. All classes should have equal intervals.

Optional

4. All classes should have both an upper boundary and a lower boundary.
5. Select the class interval from convenient numbers, such as 1, 2, 3, . . . 10, particularly 5 and 10 or multiples of 5 and 10.
6. The lower boundary of each class interval should be a multiple of the class interval.
7. Aim for a total of approximately 10 classes.

22. Define outliers.

- An outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data graph or dataset.

23. What is frequency distribution for Ungrouped Data and grouped Data?

- **Frequency Distribution for Ungrouped Data**
  - A frequency distribution produced whenever observations are sorted into classes of single values.
- **Frequency Distribution for Grouped Data**
  - A frequency distribution produced whenever observations are sorted into classes of more than one value

24. Define Unit of Measurement.

- The smallest possible difference between scores

25. Define Real Limits of Class Intervals.

- Located at the midpoint of the gap between adjacent tabled boundaries.

26. What is Relative Frequency Distribution?

- A frequency distribution showing the frequency of each class as a fraction of the total frequency for the entire distribution.

27. What is Cumulative Frequency Distribution?

- A frequency distribution showing the total number of observations in each class and all lower-ranked classes.

28. Define Percentile Rank of an Observation.

- Percentage of scores in the entire distribution with equal or smaller values than that score.

29. What do you mean by correlation?

- Correlation is a statistical measure that expresses the extent to which two variables are linearly related.
- A correlation reflects the strength and/or direction of the relationship between two (or more) variables.
- The direction of a correlation can be either positive or negative.

30. What are the three features of a correlation?

- Correlations have three important characteristics.
  The direction of the relationship,
  the form (shape) of the relationship,
  the degree (strength) of the relationship between two variable

31. What are the 4 types of correlation?

- Pearson correlation,
- Kendall rank correlation,
- Spearman correlation,
- Point-Biserial correlation.

32. What is importance of correlation?

- The correlation coefficient helps in measuring the extent of the relationship between two variables.
- Correlation analysis facilitates the understanding of economic behavior and helps in locating the critically important variables on which others depend.

33.Why correlation is used in research?

- A correlational research design investigates relationships between variables without the researcher controlling or manipulating any of them.
- A correlation reflects the strength and/or direction of the relationship between two (or more) variables. The direction of a correlation can be either positive or negative.

34.What are the benefits of correlation in statistics?

- The main benefits of correlation analysis are that it helps companies determine which variables they want to investigate further, and it allows for rapid hypothesis testing.
- The main type of correlation analysis uses Pearson's r formula to identify the degree of the linear relationship between two variables

35.What is positive relationship?

- Low values are paired with relatively low values, and relatively high values are paired with relatively high values, the relationship is positive.

36.What is negative relationship?

- Low values are paired with relatively high values, and relatively high values are paired with relatively low values, the relationship is negative.

37.What is Little or no relationship?

- A dot cluster that lacks any apparent slope, reflects little or no relationship.

38.What is strong or weak relationship?

- Having established that a relationship is either positive or negative, note how closely the dot cluster approximates a straight line.
- The more closely the dot cluster approximates a straight line, the stronger (the more regular) the relationship will be.

39.What is Perfect Relationship?

- A dot cluster that equals (rather than merely approximates) a straight line reflects a perfect relationship between two variables. In practice, perfect relationships are most unlikely.

40.What is Linear Relationship?

- A relationship that can be described best with a straight line.

41.How do you describe correlation on a scatter plot?

- We often see patterns or relationships in scatter plots.
- When the y variable tends to increase as the x variable increases, there is a positive correlation between the variables.
- When the y variable tends to decrease as the x variable increases, there is a negative correlation between the variables.

42.Define correlation coefficient.

- A **correlation coefficient** is a number between –1 and 1 that describes the relationship between pairs of variables.

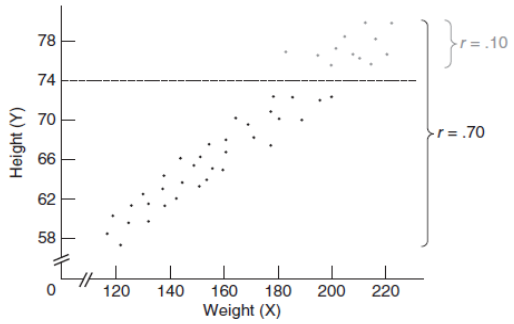
43.Write short note on Pearson Correlation Coefficient (r).

- A number between –1.00 and+1.00 that describes the linear relationship between pairs of quantitative variables.

44.What are the key properties of r?

- The Pearson correlation coefficient, r, can equal any value between –1.00 and +1.00. Furthermore, the following two properties apply:
  - The sign of r indicates the type of linear relationship, whether positive or negative.
  - The numerical value of r, without regard to sign, indicates the strength of the linear relationship.

45.Draw the graph for Effect of range restriction on the value of r.



46.What is the correlation coefficient computational formula for r

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

$$SP_{xy} = \Sigma\left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}$$

47.What is Correlation Matrix ?

Table showing correlations for all possible pairs of variables.

48. What is regression or correlation?

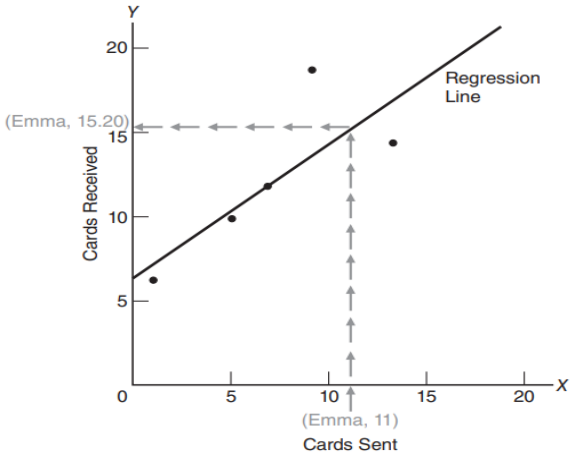
- The most commonly used techniques for investigating the relationship between two quantitative variables are correlation and linear regression.
- Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation

49.What is mean by regression line?

- The regression line is a straight line rather than a curved line because of the linear relationship between cards sent and cards received.
- Used to predict the value of some continuous response variable.

50.What is mean by predictive errors?

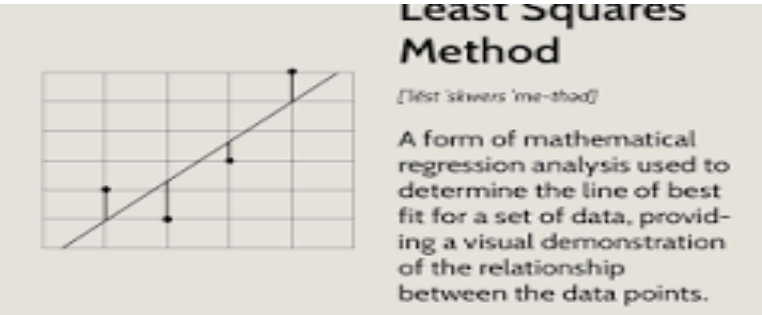
- In statistics, **prediction error** refers to the difference between the predicted values made by some model and the actual values.



51.What is Logistic Regression?

- Used to predict the value of some binary response variable. One common way to measure the prediction error of a logistic regression model is with a metric known as the total misclassification rate.

52.. What is a least squares regression line?

- 

- If the data shows a linear relationship between two variables, the line that best fits this
- linear relationship is known as a least-squares regression line, which minimizes the vertical
- distance from the data points to the regression line

53. Write Least Squares Regression Equation.

- An equation pinpoints the exact least squares regression line for any scatterplot. Most generally, this equation reads

$$Y' = bX + a$$

54. What is Least Squares Regression Equation ?

- The equation that minimizes the total of all squared prediction errors for known Y scores in the original correlation analysis.

55. How will you Find Values of b and a in least square regression equation?

SOLVING FOR *b*

$$b = r\sqrt{\frac{SS_y}{SS_x}}$$

SOLVING FOR *a*

$$a = \overline{Y} - b\overline{X}$$

56. What is Standard Error of Estimate (sy|x)?

- A rough measure of the average amount of predictive error

57. Write the formula to Finding the Standard Error of Estimate (Definition formula)

$$s_{y|x} = \sqrt{\frac{SS_{y|x}}{n-2}} = \sqrt{\frac{\sum(Y-Y')^2}{n-2}}$$

58. Write the formula to Finding the Standard Error of Estimate (Computation formula)

$$s_{y|x} = \sqrt{\frac{SS_y\left(1-r^2\right)}{n-2}}$$

59. Write short notes on squared correlation coefficient of r^2?

- The squared correlation coefficient, r2 , provides us with not only a key interpretation of the correlation coefficient but also a measure of predictive accuracy that supplements the standard error of estimate, sy|x.

60. Write notes on Multiple Regression Equation .

- A least squares equation that contains more than one predictor or X variable.

61. What is mean by Regression Toward the Mean ?

- A tendency for scores, particularly extreme scores, to shrink toward the mean

62. Elucidate Regression Fallacy .

- Regression Fallacy Occurs whenever regression toward the mean is interpreted as a real, rather than a chance, effect.

PART B

1. Explain in detail about types of Data.

THREE TYPES OF DATA

- **Qualitative data** consist of words (Yes or No), letters (Y or N), or numerical codes (0 or 1) that represent a class or category.
- **Ranked data** consist of numbers (1st, 2nd, . . . 40th place) that represent relative standing within a group.
- **Quantitative data** consist of numbers (weights of 238, 170, . . . 185 lbs) that represent an amount or a count.

Example **Table 2.1**

QUANTITATIVE DATA: WEIGHTS (IN POUNDS) OF MALE STATISTICS STUDENTS

160	168	133	170	150	165	158	165
193	169	245	160	152	190	179	157
226	160	170	180	150	156	190	156
157	163	152	158	225	135	165	135
180	172	160	170	145	185	152	
205	151	220	166	152	159	156	
165	157	190	206	172	175	154	

- The weights reported by **53 male** students in Table 2.1 are **quantitative data**, since any single observation, such as 160 lbs, represents an amount of weight.
- If the weights had been replaced with ranks, beginning with a rank of 1 for the **lightest weight of 133 lbs** and ending with a **rank of 53** for the **heaviest weight of 245 lbs**, these numbers would have been **ranked data**.

Table 2.2

QUALITATIVE DATA: "DO YOU HAVE A FACEBOOK PROFILE?" YES (Y) OR NO (N) REPLIES OF STATISTICS STUDENTS

Y	Y	Y	N	N	Y	Y	Y
Y	Y	Y	N	N	Y	Y	Y
N	Y	N	Y	Y	Y	Y	Y
Y	Y	N	Y	N	Y	N	Y
Y	N	Y	N	N	Y	Y	Y
Y	Y	N	Y	Y	Y	Y	Y
N	N	N	N	Y	N	N	Y
Y	Y	Y	Y	Y	N	Y	N
Y	Y	Y	Y	N	N	Y	Y
N	Y	N	N	Y	Y	Y	Y
	Y	Y	N				

- Finally, the Y and N replies of students in **Table 2.2** are **qualitative data**, since any single observation is a letter that represents a class of replies.

2. Explain in detail about Types of Variables?

TYPES OF VARIABLES

Definition

- A **variable** is a characteristic or property that can take on different values.
- Accordingly, the weights not only as quantitative data but also as observations for a **quantitative variable**, since the various weights take on different numerical values.
- By the same token, for a **qualitative variable,** since the replies to the Facebook profile question take on different values of either Yes or No.
- Any single observation is **constant,** since it takes on only one value.

Discrete and Continuous Variables

- Quantitative variables can be divided into
  - Discrete
  - Continuous
- A **discrete variable** consists of isolated numbers separated by gaps.
  - **Example**
    - **counts,** such as the number of children in a family;
    - The number of foreign countries you have visited;

- A **continuous variable** consists of numbers whose values have no restrictions.

  Example
  - **amounts,** such as weights of male statistics students;
  - **durations,** such as the reaction times of grade school children to a fire alarm;
  - **Standardized test scores,** such as those on the Scholastic Aptitude Test (SAT).
- **Approximate Numbers**
  - Values for continuous variables be rounded off, the resulting numbers are **approximate,** never exact.
  - **Example**, the weights of the male statistics students are approximate because they have been rounded to the nearest pound.
  - A student whose weight is listed as 150 lbs could actually weigh between 149.5 and 150.5 lbs.

Independent and Dependent Variables

- Presence or absence of a relationship between two or more variables.
- An **experiment** is a study in which the investigator decides who receives the special treatment.

Independent Variable

- An **independent variable** is the treatment manipulated by the investigator.

Dependent Variable

- A variable that is believed to have been influenced by the independent variable.

Observational Studies

- An **observational study** focuses on detecting relationships between variables not manipulated by the investigator,

Example,

- The independent variable is qualitative, with nominal measurement, whereas the dependent variable (number of communication breakdowns) is quantitative.

Confounding Variable

- An uncontrolled variable that compromises the interpretation of a study is known as a **confounding variable.**

3. Explain in detail about Describing Data with Tables and Graphs.

DESCRIBING DATA WITH TABLES AND GRAPHS

Tables

- Frequency distributions for quantitative data
- Outliers
- Relative frequency distributions
- Cumulative frequency distributions
- Frequency distributions for qualitative (nominal) data
- Interpreting distributions constructed by others

Graphs

- Graphs for quantitative data
- Typical shapes
- A graph for qualitative (nominal) data

FREQUENCY DISTRIBUTIONS FOR QUANTITATIVE DATA

- Definition
- Frequency Distribution
  - Ungrouped Data
  - Grouped Data
- Guidelines
- Example Problems
- Types of Frequency Distribution
  - Relative FD
  - Cumulative FD
- Frequency Distributions For Qualitative (Nominal) Data

- **Definition**
  - **A *frequency distribution* is a collection of observations produced by sorting observations into classes and showing their frequency (*f*) of occurrence in each class.**

- **Frequency Distribution for Ungrouped Data**

  A frequency distribution produced whenever observations are sorted into classes of single values. **, refer Table 2.3.**

Table 2.3

FREQUENCY DISTRIBUTION (UNGROUPED DATA)

WEIGHT	*f*
245	1
244	0
243	0
242	0
*	
*	
*	
161	0
160	4
159	1
158	2
157	3
*	
*	
*	
136	0
135	2
134	0
133	1
Total	53

- Frequency distributions for ungrouped data are much more informative when the number of possible values is less than about 20.
- Otherwise, if there are 20 or more possible values, using a frequency distribution for grouped data

Table 2.4

3	7	2	7	8
3	1	4	10	3
2	5	3	5	8
9	7	6	3	7
8	9	7	3	6

- **Example** - In Table 2.4 Students in a theater arts appreciation class rated the classic film, *The Wizard of Oz* on a 10-point scale, ranging from 1 (poor) to 10 (excellent), since the number of possible values is relatively small—only 10—it's appropriate to construct a frequency distribution for ungrouped data.

  Solution Refer Table 2.5

Table 2.5

RATING	TALLY*	*f*
10	/	1
9	//	2
8	///	3
7	𝍸	5
6	//	2
5	//	2
4	/	1
3	𝍸 /	6
2	//	2
1	/	1
	Total	25

➢ **Frequency Distribution for Grouped Data**

- A frequency distribution produced whenever observations are sorted into classes of more than one value.
- The general structure of this frequency distribution is
  - Data are grouped into class intervals with 10 possible values each in Table 2.6.
  - The bottom class includes the smallest observation (133), and the top class includes the largest observation (245).
  - The distance between bottom and top is occupied by an orderly series of classes.
  - The frequency (*f*) column shows the frequency of observations in each class and, at the bottom, the total number of observations in all classes.

- **Example – Refer Table 2.6**

Table 2.6

FREQUENCY DISTRIBUTION (GROUPED DATA)	
WEIGHT	***f***
240–249	1
230–239	0
220–229	3
210–219	0
200–209	2
190–199	4
180–189	3
170–179	7
160–169	12
150–159	17
140–149	1
130–139	3
Total	53

➢ **GUIDELINES**

- The "Guidelines for Frequency Distributions" box lists seven rules for producing a well-constructed frequency distribution.
- The first three rules are essential and should not be violated. The last four rules are optional and can be modified or ignored.

GUIDELINES FOR FREQUENCY DISTRIBUTIONS

Essential

1. ***Each observation should be included in one, and only one, class.***
   Example: 130–139, 140–149, 150–159, etc. It would be incorrect to use 130–140, 140–150, 150–160, etc., in which, because the boundaries of classes overlap, an observation of 140 (or 150) could be assigned to either of two classes.
2. ***List all classes, even those with zero frequencies.***
   Example: Listed in Table 2.2 is the class 210–219 and its frequency of zero. It would be incorrect to skip this class because of its zero frequency.
3. ***All classes should have equal intervals.***
   Example: 130–139, 140–149, 150–159, etc. It would be incorrect to use 130–139, 140–159, etc., in which the second class interval (140–159) is twice as wide as the first class interval (130–139).

Optional

4. ***All classes should have both an upper boundary and a lower boundary.***
   Example: 240–249. Less preferred would be 240–above, in which no maximum value can be assigned to observations in this class.
5. ***Select the class interval from convenient numbers, such as 1, 2, 3, . . . 10, particularly 5 and 10 or multiples of 5 and 10.***
   Example: 130–139, 140–149, in which the class interval of 10 is a convenient number. Less preferred would be 130–142, 143–155, etc., in which the class interval of 13 is not a convenient number.
6. ***The lower boundary of each class interval should be a multiple of the class interval.***
   Example: 130–139, 140–149, in which the lower boundaries of 130, 140, are multiples of 10, the class interval. Less preferred would be 135–144, 145–154, etc., in which the lower boundaries of 135 and 145 are not multiples of 10, the class interval.
7. ***Aim for a total of approximately 10 classes.***

Gaps between Classes

- The size of the gap should always equal one **unit of measurement;** that is, it should always equal the smallest possible difference between scores within a particular set of data.
- The smallest class interval would be 130.0–139.9 (not 130–139), and the next class interval would be 140.0–149.9 (not 140–149), and so on.

Real Limits of Class Intervals

- The **real limits** are located at the midpoint of the gap between adjacent tabled boundaries;
- That is, one-half of one unit of measurement below the lower tabled boundary and one-half of one unit of measurement above the upper tabled boundary.

Example,

- The real limits for 140–149 are 139.5 (140 minus one-half of the unit of measurement of 1) and 149.5 (149 plus one-half of the unit of measurement of 1), and the actual width of the class interval would be 10 (from 149.5 -139.5 = 10).

Example 2.1

The IQ scores for a group of 35 high school dropouts are as follows:

Table 2.7

91	85	84	79	80
87	96	75	86	104
95	71	105	90	77
123	80	100	93	108
98	69	99	95	90
110	109	94	100	103
112	90	90	98	89

(a) Construct a frequency distribution for grouped data.

(b) Specify the real limits for the lowest class interval in this frequency Distribution.

Solution

a. Construction of a Frequency Distribution

1. **Find the range, that is,** the difference between the largest and smallest observations.
   Example- From Table 2.7:
   The range of weights in the above table is 123 – 69 = 54.
2. **Find the class interval required to span the range** by dividing the range by the desired number of classes (ordinarily 10).

$$\text{Class interval} = \frac{\text{range}}{\text{desired number of classes}}$$

   Example: Class interval = 54/10=5.4
3. **Round off to the nearest convenient interval** (such as 1, 2, 3, . . . 10, particularly 5 or 10 or multiples of 5 or 10).
   Example: nearest convenient interval = 5
4. **Determine where the lowest class should begin.** (Ordinarily, this number should be a multiple of the class interval.)
   Example: the smallest score is 69, and therefore the lowest class should begin at 65, since 65 is a multiple of 5 (the class interval).
5. **Determine where the lowest class should end** by adding the class interval to the lower boundary and then subtracting one unit of measurement.
   Example: Add 5 to 65 and then subtract 1, the unit of measurement, to obtain 69—the number at which the lowest class should end.

6. **Working upward, list as many equivalent classes as are required to include the largest observation.**
   Example: 65-69, 70 – 74, . . . 120 - 124, so that the last class includes 123, the largest score.
7. **Indicate with a tally the class in which each observation falls.**
   For example, the first score in Table 2.6, 160, produces a tally next to 160–169; the next score, 193, produces a tally next to 190–199; and so on.
8. **Replace the tally count for each class with a number—the frequency (f)—and show the total of all frequencies.** (Tally marks are not usually shown in the final frequency distribution.)
9. **Supply headings for both columns and a title for the table.**
   Output – Refer Table 2.8

Table 2.8

IQ	TALLY*	f
120–124	/	1
115–119		0
110–114	//	2
105–109	///	3
100–104	~~////~~	4
95–99	~~////~~ /	6
90–94	//// //	7
85–89	////	4
80–84	///	3
75–79	///	3
70–74	/	1
65–69	/	1
	Total	35

Solution b - the real limits for the lowest class interval in this frequency Distribution **-** 64.5 – 69.5

Example 2.2

What are some possible poor features of the following frequency Distribution?

ESTIMATED WEEKLY TV VIEWING TIME (HRS) FOR 250 SIXTH GRADERS

VIEWING TIME	f
35–above	2
30–34	5
25–30	29
20–22	60
15–19	60
10–14	34
5–9	31
0–4	29
Total	250

Solution

- Not all observations can be assigned to one and only one class (because of gap between 20–22 and 25–30 and overlap between 25–30 and 30–34).
- All classes are not equal in width (25–30 versus 30–34).
- All classes do not have both boundaries (35–above).

➢ **Types of Frequency distribution**

1. Relative Frequency Distributions

- Relative frequency distributions shows the frequency of each class as a part or fraction of the total frequency for the entire distribution.

Constructing Relative Frequency Distributions

- To convert a frequency distribution into a relative frequency distribution, divide the frequency for each class by the total frequency for the entire distribution.

Percentages or Proportions

- A proportion always varies between 0 and 1, whereas a percentage always varies between 0 percent and 100 percent.
- To convert the relative frequencies from proportions to percentages, multiply each proportion by 100;
- That is, move the decimal point two places to the right.
- For example, multiply .06 by 100 to obtain 6 percent.

Example – Table 2.9

RELATIVE FREQUENCY DISTRIBUTION

WEIGHT	*f*	RELATIVE *f*
240–249	1	.02
230–239	0	.00
220–229	3	.06
210–219	0	.00
200–209	2	.04
190–199	4	.08
180–189	3	.06
170–179	7	.13
160–169	12	.23
150–159	17	.32
140–149	1	.02
130–139	3	.06
Total	53	1.02*

** The sum does not equal 1.00 because of rounding-off errors.*

2. Cumulative Frequency Distributions

- Cumulative frequency distributions show the total number of observations in each class and in all lower-ranked classes.
- Cumulative frequencies are usually converted, to cumulative percentages. Cumulative percentages are often referred to as percentile ranks.

Constructing Cumulative Frequency Distributions

- To convert a frequency distribution into a cumulative frequency distribution, add to the frequency of each class to the sum of the frequencies of all classes ranked below it.
- Begin with the lowest-ranked class in the frequency distribution and work upward, finding the cumulative frequencies in ascending order.

Example – Table 2.10

CUMULATIVE FREQUENCY DISTRIBUTION

WEIGHT	*f*	CUMULATIVE *f*	CUMULATIVE PERCENT
240–249	1	53	100
230–239	0	52	98
220–229	3	52	98
210–219	0	49	92
200–209	2	49	92
190–199	4	47	89
180–189	3	43	81
170–179	7	40	75
160–169	12	33	62
150–159	17	21	40
140–149	1	4	8
130–139	3	3	6
Total	53		

- The cumulative frequency for the class 130–139 is 3, since there are no classes ranked lower.
- The cumulative frequency for the class 140–149 is 4, since 1 is the frequency for that class and 3 is the frequency of all lower-ranked classes.
- The cumulative frequency for the class 150–159 is 21, since 17 is the frequency for that class and 4 is the sum of the frequencies of all lower-ranked classes.

Cumulative Percentages

- To obtain this cumulative percentage (75%), the cumulative frequency of 40 for the class 170–179 should be divided by the total frequency of 53 for the entire distribution.

Example 2.3

GRE scores for a group of graduate school applicants are distributed as follows:

GRE	*f*
725–749	1
700–724	3
675–699	14
650–674	30
625–649	34
600–624	42
575–599	30
550–574	27
525–549	13
500–524	4
475–499	2
Total	200

Convert the distribution of GRE scores shown to a cumulative frequency distribution and cumulative percent.

Solution

GRE	(a) CUMULATIVE *f*	(b) CUMULATIVE PERCENT(%)
725–749	200	100
700–724	199	100
675–699	196	98
650–674	182	91
625–649	152	76
600–624	118	59
575–599	76	38
550–574	46	23
525–549	19	10
500–524	6	3
475–499	2	1

Percentile Ranks

- Cumulative percentages are referred to as percentile ranks.
- The percentile rank of a score indicates the percentage of scores in the entire distribution with similar or smaller values than that score.

Types of Percentile Rank

- The assignment of exact percentile ranks requires that cumulative percentages be obtained from frequency distributions for ungrouped data.
- The assignment of approximate percentile ranks requires that cumulative percentages be obtained from frequency distributions for grouped data.

Example 2.4

Referring to Table 2.10, find the approximate percentile rank of any weight in the class 200–209.

Solution –

The approximate percentile rank for weights between 200 and 209 lbs is 92 (because 92 is the cumulative percent for this interval).

➢ FREQUENCY DISTRIBUTIONS FOR QUALITATIVE (NOMINAL) DATA

- When, among a set of observations, any single observation is a word, letter, or numerical code, the data are qualitative.

Example 2.5

Movie ratings reflect ordinal measurement because they can be ordered from most to least restrictive: NC-17, R, PG-13, PG, and G. The ratings of some films shown recently in San Francisco are as follows:

PG	PG	PG	PG-13	G
G	PG-13	R	PG	PG
R	PG	R	PG	R
NC-17	NC-17	PG	G	PG-13

(a) Construct a frequency distribution.

(b) Convert to relative frequencies, expressed as percentages.

(c) Construct a cumulative frequency distribution.

(d) Find the approximate percentile rank for those films with a PG rating

MOVIE RATINGS	(a) *f*	(b) RELATIVE *f(%)*	(c) CUMULATIVE *f*
NC-17	2	10	20
R	4	20	18
PG-13	3	15	14
PG	8	40	11
G	3	15	3
Totals	20	100%	

(d) Percentile rank for films with a PG rating is 55 (from $\frac{11}{20}$ multiplied by 100).

4. How the frequency distributions are interpreted by others?

Interpreting Distributions

- When inspecting a distribution for the first time, train to look at the entire table, not just the distribution.
- Read the title, column headings, and any footnotes.
- Identify from where do the data come from and Is a source cited.
- Next, focus on the form of the frequency distribution. Is it well constructed?
- For quantitative data, check does the total number of classes seem to avoid either over- or under-summarizing the data.
- After these preliminaries, inspect the content of the frequency distribution.
- What is the approximate range? Does it seem reasonable?
- Disregard the inevitable irregularities that accompany a frequency distribution and focus on its overall appearance or shape.
- Do the frequencies arrange themselves around a single peak (high point) or several peaks?
- Is the distribution fairly balanced around its peak?
- When interpreting distributions, including distributions constructed by someone else, keep an open mind.

5. Give a brief summary about outliers.

OUTLIERS

- Are a very extreme score

Handling Outliers

- **Check for Accuracy**
  - Whenever encounter an outrageously extreme value, such as a GPA of 0.06, attempt to verify its accuracy.

- For instance, was a respectable GPA of 3.06 recorded erroneously as 0.06?
- If the outlier survives an accuracy check, it should be treated as a legitimate score.

- **Might Exclude from Summaries**
  - To segregate an outlier from any summary of the data
- **Might Enhance Understanding**
  - Insofar as a valid outlier can be viewed as the product of special circumstances, it might help you to understand the data.

Example 2.6:

Identify any outliers in each of the following sets of data collected from nine college students

SUMMER INCOME	AGE	FAMILY SIZE	GPA
$6,450	20	2	2.30
$4,820	19	4	4.00
$5,650	61	3	3.56
$1,720	32	6	2.89
$600	19	18	2.15
$0	22	2	3.01
$3,482	23	6	3.09
$25,700	27	3	3.50
$8,548	21	4	3.20

Solution

- Outliers are a summer income of $25,700; an age of 61; and a family size of 18. No outliers for GPA.

6. Discuss in detail about graphs and graphs for Quantitative data.

GRAPHS

- Graphs for Quantitative Data
  - Histograms
  - Frequency Polygon
  - Stem and Leaf Displays
- Graphs for Qualitative Data
  - Bar Graph
- Shapes
  - Normal
  - Bimodal
  - Positively Skewed
  - Negatively Skewed

➢ GRAPHS FOR QUANTITATIVE DATA

1. Histograms

- A bar-type graph for quantitative data.
- The common boundaries between adjacent bars emphasize the Continuity of the data, as with continuous variables.

Important features of histograms

- Equal units along the horizontal axis (the X axis) reflect the various class intervals of the frequency distribution.
- Equal units along the vertical axis (the Y axis) reflect increases in frequency.
- The intersection of the two axes defines the origin at which both numerical scales equal 0.
- Numerical scales always increase from left to right along the horizontal axis and from bottom to top along the vertical axis.
- The body of the histogram consists of a series of bars whose heights reflect the frequencies for the various classes.
- Example **Figure 2.1**

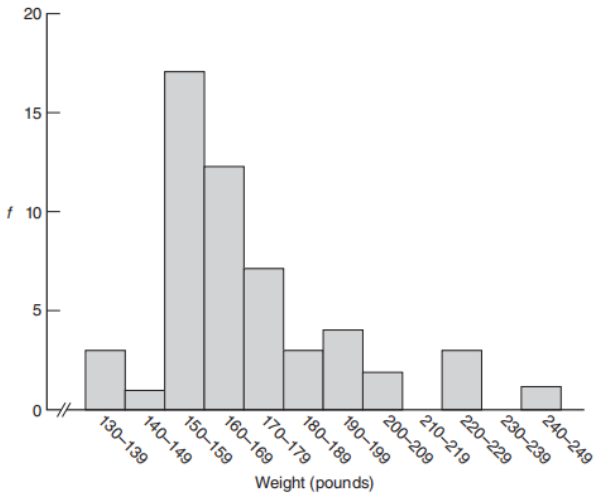


Figure 2.1

2. Frequency Polygon

- Frequency Polygon or a line graph for quantitative data emphasizes the continuity of continuous variables.
- Frequency polygons may be constructed directly from frequency distributions.

Transformation steps from histogram to frequency polygon

1. This panel shows the histogram for the weight distribution.
2. Place dots at the midpoints of each bar top or, at midpoints for classes on the horizontal axis, and connect them with straight lines.
3. First, extend the upper tail to the midpoint of the first unoccupied class on the upper flank of the histogram. Then extend the lower tail to the

midpoint of the first unoccupied class on the lower flank of the histogram. Now all of the area under the frequency polygon is enclosed completely.

4. Finally, erase all of the histogram bars, leaving only the frequency polygon.
5. Example **Figure 2.2**

EXAMPLE

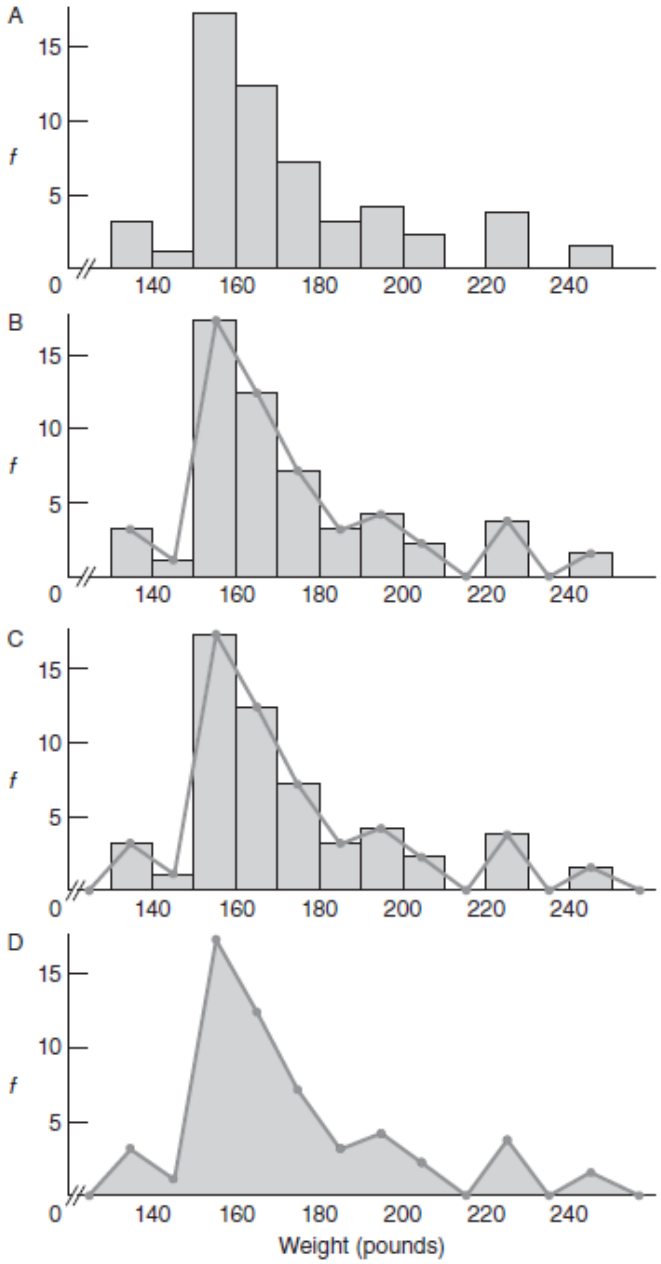


Figure 2.2 Transition from histogram to frequency polygon.

3. Stem and Leaf Displays

- A device for sorting quantitative data on the basis of leading and trailing digits
- Summarizing quantitative data is a stem and leaf display.

Constructing a Display

- To construct the stem and leaf display for these data, first note that, when counting by tens, the weights range from the 130s to the 240s.
- Arrange a column of numbers, the stems, beginning with 13 (representing the 130s) and ending with 24 (representing the 240s).
- Draw a vertical line to separate the stems, which represent multiples of 10, from the space to be occupied by the leaves, which represent multiples of 1.
- Next, enter each raw score into the stem and leaf display.

Interpretation

- The weight data have been sorted by the stems. All weights in the 130s are listed together; all of those in the 140s are listed together, and so on.

CONSTRUCTING STEM AND LEAF DISPLAY FROM WEIGHTS OF MALE STATISTICS STUDENTS

RAW SCORES					STEM AND LEAF DISPLAY
160	165	135	175		
193	168	245	165	13	3 5 5
226	169	170	185	14	5
152	160	156	154	15	2 7 1 7 8 0 2 0 2 6 9 8 2 6 4 7 6
180	170	160	179	16	0 3 5 8 9 0 0 0 6 5 5 5
205	150	225	165	17	2 0 0 0 2 5 9
163	152	190	206	18	0 0 5
157	160	159	165	19	3 0 0 0
151	190	172	157	20	5 6
157	150	190	156	21	
220	133	166	135	22	6 0 5
145	180	158		23	
158	152	152		24	5
172	170	156			

As suggested by the shaded coding in **Table 2.11**, the first raw score of 160 reappears as a leaf of 0 on a stem of 16. The next raw score of 193 reappears as a leaf of 3 on a stem of 19, and the third raw score of 226 reappears as a leaf of 6 on a stem of 22, and so on, until each raw score reappears as a leaf on its appropriate stem.

Example 2.7

Construct a stem and leaf display for the following IQ scores obtained from a group of four-year-old children.

120	98	118	117	99	111
126	85	88	124	104	113
108	141	123	137	78	96
102	132	109	106	143»	

Solution

Stem	Leaf				
7	8				
8	5	8			
9	8	9	6		
10	8	2	9	6	4
11	8	7	1	3	
12	0	6	3	4	
13	2	7			
14	1	3			

➢ A GRAPH FOR QUALITATIVE (NOMINAL) DATA

- A bar-type graph for qualitative data. Gaps between adjacent bars emphasize the discontinuous nature of the data.

Bar Graph

- Equal segments along the vertical axis reflect increases in frequency. The body of the bar graph consists of a series of bars whose heights reflect the frequencies for the various words or classes. Refer Figure 2.7.

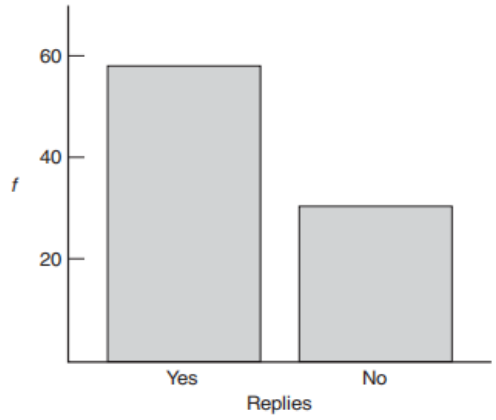


Figure 2.7 Bar Graph

CONSTRUCTING GRAPHS

1. **Decide on the appropriate type of graph,** recalling that histograms and frequency polygons are appropriate for quantitative data, while bar graphs are appropriate for qualitative data**.**
2. **Draw the horizontal axis, then the vertical axis,** remembering that the vertical axis should be about as tall as the horizontal axis is wide.
3. **Identify the string of class intervals that eventually will be superimposed on the horizontal axis.** For qualitative data or ungrouped quantitative data, just use the classes suggested by the data. For grouped quantitative data, creating a set of class intervals for a frequency distribution.
4. **Superimpose the string of class intervals** (with gaps for bar graphs) along the entire length of the horizontal axis.
5. **Along the entire length of the vertical axis, superimpose a progression of convenient numbers,** beginning at the bottom with 0 and ending at the top with a number as large as or slightly larger than the maximum observed frequency.
6. **Using the scaled axes, construct bars** (or dots and lines) to reflect the frequency of observations within each class interval.
7. **Supply labels for both axes and a title** for the graph.

➢ **TYPICAL SHAPES**

- Whether expressed as a histogram, a frequency polygon, or a stem and leaf display, an important characteristic of a frequency distribution is its shape
- The more typical shapes for smoothed frequency polygons

Normal

- The familiar bell-shaped silhouette of the normal curve can be superimposed on many frequency distributions, scores on standardized tests, and even the popping times of individual kernels in a batch of popcorn. Figure 2.3 represents normal curve.

Figure 2.3 Normal

Bimodal

- Any distribution that approximates the bimodal shape reflect the coexistence of two different types of observations in the same distribution.
- For instance, the distribution of the ages of residents in a neighborhood consisting largely of either new parents or their infants has a bimodal shape. Figure 2.4 represents bimodal curve.

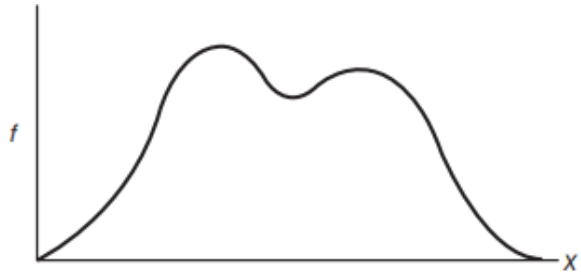


Figure 2.4 Bimodal

Positively Skewed Distribution

- A distribution that includes a few extreme observations in the positive direction (to the right of the majority of observations).
- Figure 2.5 represents positively skewed distribution curve.

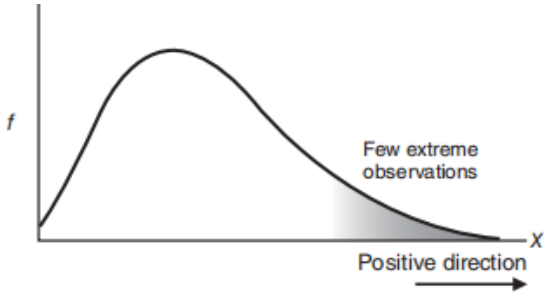


Figure 2.5 Positively Skewed Distribution

Negatively Skewed Distribution

- A distribution that includes a few extreme observations in the negative direction (to the left of the majority of observations)
- Figure 2.6 represents positively skewed distribution curve.

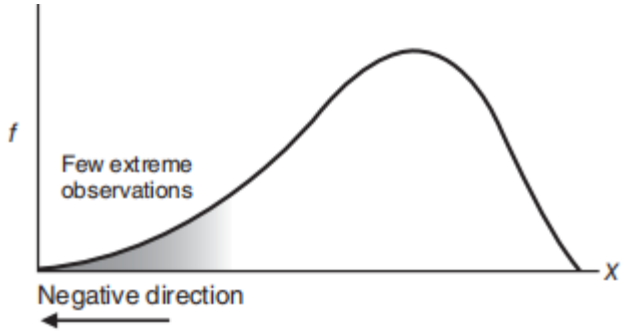


Figure 2.6 Negatively Skewed Distribution

Example 2.8

Describe the probable shape—normal, bimodal, positively Skewed, or negatively skewed—for each of the following distributions:

(a) female beauty contestants' scores on a masculinity test, with a higher score indicating a greater degree of masculinity

(b) scores on a standardized IQ test for a group of people selected from the general population

(c) test scores for a group of high school students on a very difficult college - level math exam

(d) reading achievement scores for a third-grade class consisting of about equal numbers of regular students and learning-challenged students

(e) scores of students at the Eastman School of Music on a test of music aptitude (designed for use with the general population)

Solution

(a) Positively skewed
(b) Normal
(c) Positively skewed
(d) Bimodal
(e) Negatively skewed

7 Explain in detail about describing data with Averages.

Averages or Measures of central tendency

- Averages for Quantitative Data
  - Mode
  - Median
  - Mean
- Averages for Qualitative Data
- Averages for Ranked Data

MEASURES OF CENTRAL TENDENCY

- Numbers or words that attempt to describe, most generally, the middle or typical value for a distribution.

➢ **AVERAGES FOR QUANTITATIVE DATA**

MODE

- The mode reflects the value of the most frequently occurring score.
- Distributions with two obvious peaks, even though they are not exactly the same height, are referred to as bimodal.
- Distributions with more than two peaks are referred to as multimodal.
- Refer Figure 2.8

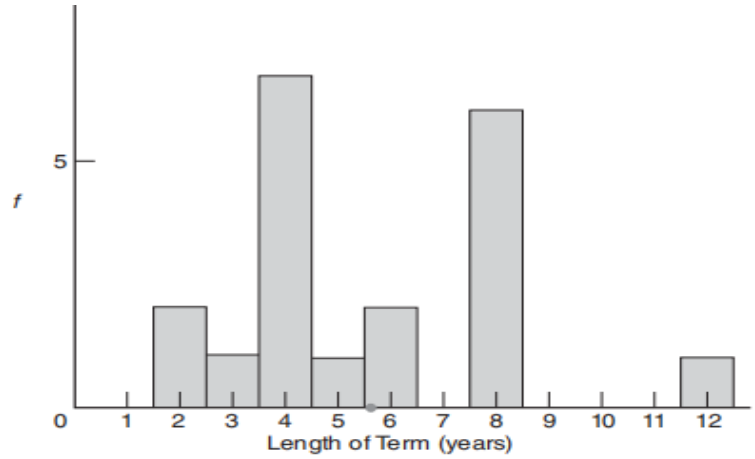


Figure 2.8 - Modes

Example 2.9:

Determine the mode for the following retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63.

Answer - mode = 63

Example 2.10:

The owner of a new car conducts six gas mileage tests and obtains the following results, expressed in miles per gallon: 26.3, 28.7, 27.4, 26.6, 27.4, 26.9. Find the mode for these data.

Answer - mode = 27.4

MEDIAN

- The median reflects the middle value when observations are ordered from least to most.

FINDING THE MEDIAN

A. INSTRUCTIONS

1. Order scores from least to most.
2. Find the middle position by adding one to the total number of scores and dividing by 2.
3. If the middle position is a whole number, as in the left-hand panel below, use this number to count into the set of ordered scores.
4. The value of the median equals the value of the score located at the middle position.

5. If the middle position is not a whole number, as in the right-hand panel below, use the two nearest whole numbers to count into the set of ordered scores.
6. The value of the median equals the value midway between those of the two middlemost scores; to find the midway value, add the two given values and divide by 2.

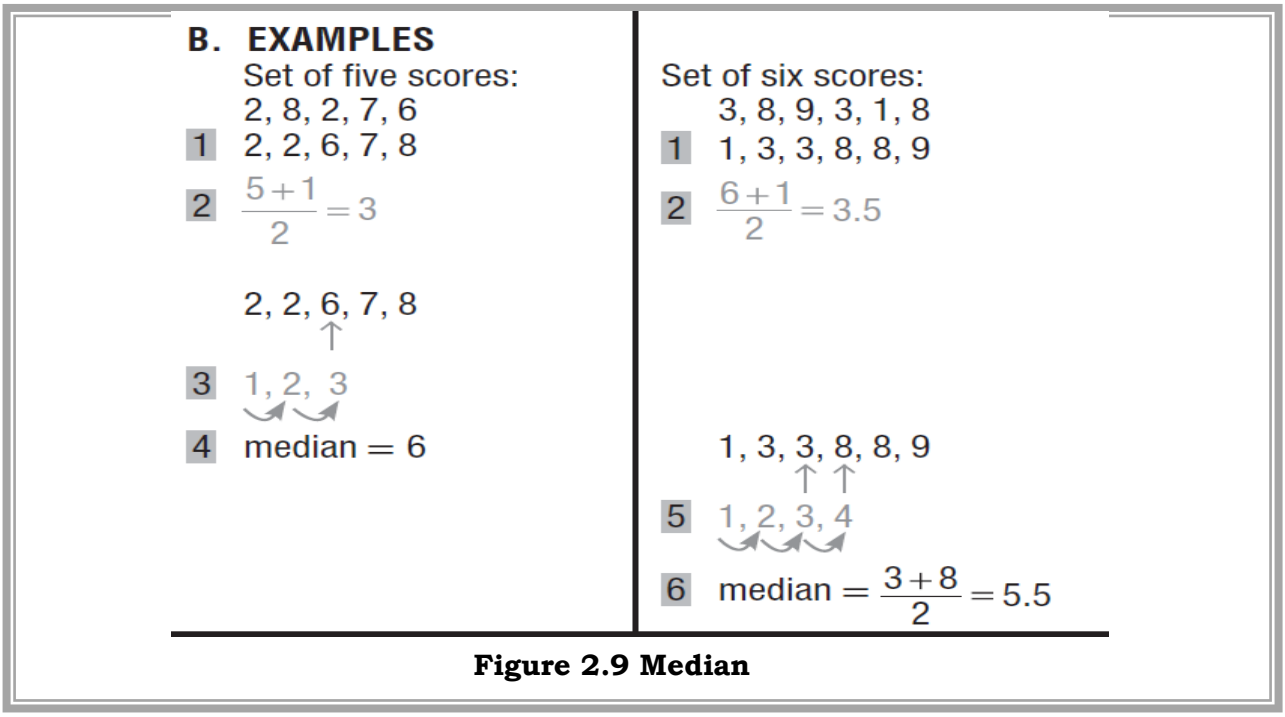


Figure 2.9 Median

Example 2.11:

Find the mean for the following retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63.

$$\text{mean} = \frac{672}{11} = 6109$$

Example 2.12:

Find the median for the following gas mileage tests: 26.3, 28.7, 27.4, 26.6, 27.4, 26.9.

median = 27.15 (halfway between 26.9 and 27.4)

MEAN

- The mean is found by adding all scores and then dividing by the number of scores.

Mean = sum of all scores / number of scores

Types of mean

- population mean
- sample mean $\bar{X}$

Sample Mean $\bar{X}$

- Sample Size (n) - The total number of scores in the sample.
- Sample Mean is obtained by dividing the sum of all scores in the sample by the number of scores in the sample.

SAMPLE MEAN

$$\bar{X} = \frac{\sum X}{n}$$

- "X-bar equals the sum of the variable X divided by the sample size n."

Population Mean (μ)

- Population Size (N) - The total number of scores in the population.
- Population Mean (μ) is obtained by dividing the sum for all scores in the population by the number of scores in the population.
- The population mean is represented by μ (pronounced "mu"),

POPULATION MEAN

$$\mu = \frac{\sum X}{N}$$

Where the uppercase letter N refers to the population size.

Example 2.13

Find the mean for the following retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63.

$$\text{mean} = \frac{672}{11} = 61.09$$

Find the mean for the following gas mileage tests: 26.3, 28.7, 27.4, 26.6, 27.4, 26.9.

$$\text{mean} = \frac{163.3}{6} = 27.22$$

Mean as Balance Point

- The mean serves as the balance point for its frequency distribution.

If Distribution Is Not Skewed

- When a distribution of scores is not too skewed, the values of the mode, median, and mean are similar, and any of them can be used to describe the central tendency of the distribution.

If Distribution Is Skewed

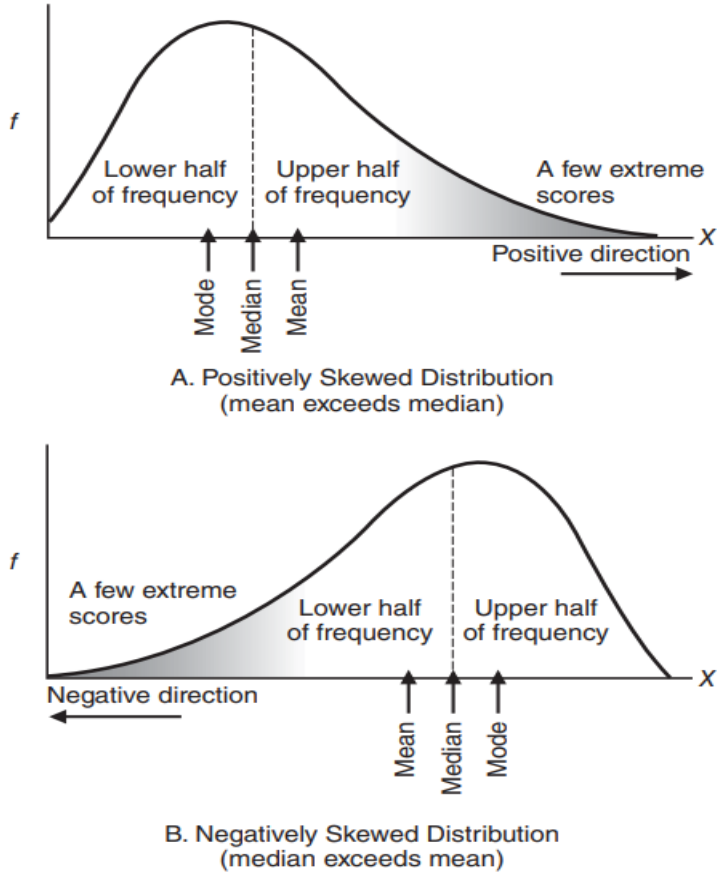


Figure 2.10 - Mode, median, and mean in positively and negatively skewed distributions.

Example 2.14

Indicate whether the following skewed distributions are positively skewed because the mean exceeds the median or negatively skewed because the median exceeds the mean.

a. a distribution of test scores on an easy test, with most students scoring high and a few students scoring low

Solution

- *negatively skewed because the median exceeds the mean*

b. a distribution of ages of college students, with most students in their late teens or early twenties and a few students in their fifties or sixties

Solution

- *positively skewed because the mean exceeds the median*

c. a distribution of loose change carried by classmates, with most carrying less than $1 and with some carrying $3 or $4 worth of loose change

Solution - positively skewed

d. a distribution of the sizes of crowds in attendance at a popular movie theater, with most audiences at or near capacity

Solution - negatively skewed

- **AVERAGES FOR QUALITATIVE DATA**
  - The mode always can be used with qualitative data.
  - The median can be used whenever it is possible to order qualitative data from least to most because the level of measurement is ordinal.

Example 2.15

College students were surveyed about where they would most like to spend their spring break: Daytona Beach (DB), Cancun, Mexico (C), South Padre Island (SP), Lake Havasu (LH), or other (O). The results were as follows:

DB	DB	C	LH	DB
C	SP	LH	DB	O
O	SP	C	DB	LH
DB	C	DB	O	DB

Find the mode and, if possible, the median.

Solution –

mode = DB (Daytona Beach)

Impossible to find the median when qualitative data are unordered, with only nominal measurement.

- **AVERAGES FOR RANKED DATA**
  - When the data consist of a series of ranks, with its ordinal level of measurement, the median rank always can be obtained.
  - It's simply the middlemost or average of the two middlemost ranks.

8 Explain in detail about the Describing data with Variability.

Measures of Variability

- Variability for Quantitative Data
  - Range
  - Variance
  - Standard Deviation
- Variability for Qualitative Data
- Variability for Ranked Data

MEASURES OF VARIABILITY

- Variability is the description of the amount by which scores are dispersed or scattered in a distribution.
- There are many ways to describe variability or spread including:
  - Range
  - Variance
  - Standard Deviation

➢ VARIABILITY FOR QUANTITATIVE DATA

RANGE

- It is the difference between the largest and smallest scores.

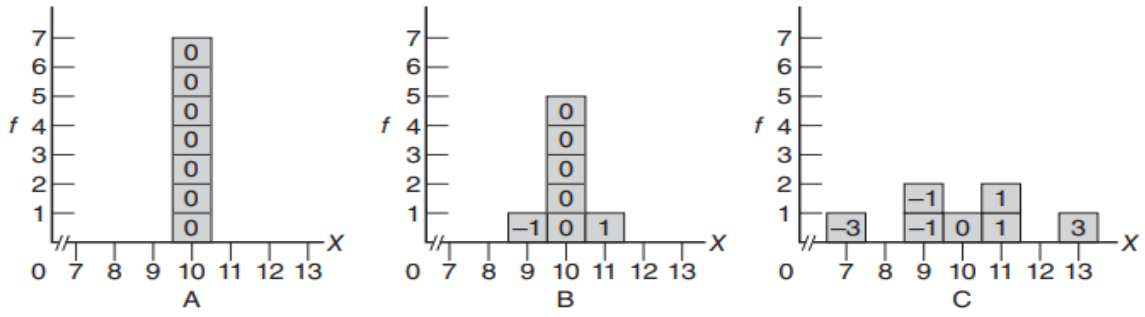


FIGURE 2.11 - Three distributions with the same mean (10) but different amounts of variability. Numbers in the boxes indicate distances from the mean.

In Figure 2.11, distribution A, the least variable, has the smallest range of 0 (from 10 to 10); distribution B, the moderately variable, has an intermediate range of 2 (from 11 to 9); and distribution C, the most variable, has the largest range of 6 (from 13 to 7),

VARIANCE

- The variance is the mean of all squared deviation scores.
- A deviation from the mean is how far a score lies from the mean.
- Variance is the square of the standard deviation.

Example

To get variance, square the standard deviation.

s = 95.5

s2 = 95.5 x 95.5 = 9129.14

The variance of your data is 9129.14.

- Variance reflects the degree of spread in the data set.
- The more spread the data, the larger the variance is in relation to the mean.

SUM OF SQUARES (SS)

- Sum of Squares (SS) - The sum of squared deviation scores.

- **Sum of Squares Formulas for Population**

SUM OF SQUARES (*SS*) FOR POPULATION (DEFINITION FORMULA)

$$SS = \sum (X - \mu)^2$$

Where SS represents the sum of squares, Σ directs us to sum over the expression to its right, and $(X - \mu)^2$ denotes each of the squared deviation scores.

- **Steps:**
  1. Subtract the population mean, μ, from each original score, X, to obtain a deviation score, X − μ.
  2. Square each deviation score, $(X - \mu)^2$, to eliminate negative signs.
  3. Sum all squared deviation scores, $\Sigma (X - \mu)^2$

SUM OF SQUARES (*SS*) FOR POPULATION (COMPUTATION FORMULA)

$$SS = \sum X^2 - \frac{(\sum X)^2}{N}$$

- where $\sum X^2$, the sum of the squared X scores, is obtained by first squaring each X score and then summing all squared X scores;
- ∑ X2 , the square of sum of all X scores, is obtained by first adding all X scores and then squaring the sum of all X scores;
- N is the population size

- **Sum of Squares Formulas for Sample**

SUM OF SQUARES (*SS*) FOR SAMPLE (DEFINITION FORMULA)

$$SS = \Sigma(X - \overline{X})^2$$

(COMPUTATION FORMULA)

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

where X, the sample mean, replaces μ, the population mean, and n, the sample size, replaces N, the population size.

STANDARD DEVIATION

- The square root of the mean of all squared deviations from the mean, that is,
  standard deviation = √variance

Standard Deviation for Population σ

variance = sum of all squared deviation scores /number of scores

VARIANCE FOR POPULATION

$$\sigma^2 = \frac{SS}{N}$$

where σ^2, represents the population variance, SS is the sum of squared deviations for the population, and N is the population size.

STANDARD DEVIATION FOR POPULATION

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$$

where σ represents the population standard deviation.

A. COMPUTATION SEQUENCE

Assign a value to N **1** representing the number of X scores
Sum all X scores **2**
Obtain the mean of these scores **3**
Subtract the mean from each X score to obtain a deviation score **4**
Square each deviation score **5**
Sum all squared deviation scores to obtain the sum of squares **6**
Substitute numbers into the formula to obtain population variance, σ^2 **7**
Take the square root of σ^2 to obtain the population standard deviation, σ **8**

B. DATA AND COMPUTATIONS

X	**4** $X-\mu$	**5** $(X-\mu)^2$
13	3	9
10	0	0
11	1	1
7	–3	9
9	–1	1
11	1	1
9	–1	1

1 $N = 7$ **2** $\Sigma X = 70$ **6** $SS = \Sigma (X-\mu)^2 = 22$

3 $\mu = \frac{70}{7} = 10$

7 $\sigma^2 = \frac{SS}{N} = \frac{22}{7} = 3.14$ **8** $\sigma = \sqrt{\frac{SS}{N}} = \sqrt{\frac{22}{7}} = \sqrt{3.14} = 1.77$

Table 2.22: Calculation of Population Standard Deviation Σ (Definition Formula)

A. COMPUTATIONAL SEQUENCE

Assign a value to N representing the number of X scores **1**
Sum all X scores **2**
Square the sum of all X scores **3**
Square each X score **4**
Sum all squared X scores **5**
Substitute numbers into the formula to obtain the sum of squares, SS **6**
Substitute numbers into the formula to obtain the population variance, σ^2 **7**
Take the square root of σ^2 to obtain the population standard deviation, σ **8**

B. DATA AND COMPUTATIONS

X	**4** X^2
13	169
10	100
11	121
7	49
9	81
11	121
9	81

1 $N = 7$ **2** $\sum X = 70$ **5** $\sum X^2 = 722$
3 $(\sum X)^2 = 4900$

6 $$SS = \sum X^2 - \frac{(\sum X)^2}{N} = 722 - \frac{4900}{7} = 722 - 700 = 22$$

7 $$\sigma^2 = \frac{SS}{N} = \frac{22}{7} = 3.14$$ **8** $$\sigma = \sqrt{\frac{SS}{N}} = \sqrt{\frac{22}{7}} = \sqrt{3.14} = 1.77$$

Table 2.23: Calculation Of Population Standard Deviation (Σ) (Computation Formula)

Standard Deviation for Sample (s)

VARIANCE FOR SAMPLE

$$s^2 = \frac{SS}{n-1}$$

STANDARD DEVIATION FOR SAMPLE

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$$

A. COMPUTATION SEQUENCE

Assign a value to n **1** representing the number of X scores
Sum all X scores **2**
Obtain the mean of these scores **3**
Subtract the mean from each X score to obtain a deviation score **4**
Square each deviation score **5**
Sum all squared deviation scores to obtain the sum of squares **6**
Substitute numbers into the formula to obtain the sample variance, s^2 **7**
Take the square root of s^2 to obtain the sample standard deviation, s **8**

B. DATA AND COMPUTATIONS

X	**4** $X - \bar{X}$	**5** $(X - \bar{X})^2$
7	4	16
3	0	0
1	–2	4
0	–3	9
4	1	1

1 $n = 5$ **2** $\Sigma X = 15$ **6** $SS = \Sigma(X - \bar{X})^2 = 30$

3 $\bar{X} = \frac{15}{5} = 3$

7 $s^2 = \frac{SS}{n-1} = \frac{30}{4} = 7.50$ **8** $s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{30}{4}} = \sqrt{7.50} = 2.74$

Table 2.24 Calculation of Sample Standard Deviation (S) (Definition Formula)

A. COMPUTATIONAL SEQUENCE

Assign a value to n representing the number of X scores **1**
Sum all X scores **2**
Square the sum of all X scores **3**
Square each X score **4**
Sum all squared X scores **5**
Substitute numbers into the formula to obtain the sum of squares, SS **6**
Substitute numbers into the formula to obtain the sample variance, s^2 **7**
Take the square root of s^2 to obtain the sample standard deviation, s **8**

B. DATA AND COMPUTATIONS

X	**4** X^2
7	49
3	9
1	1
0	0
4	16

1 $n = 5$ **2** $\Sigma X = 15$ **5** $\Sigma X^2 = 75$

3 $(\Sigma X)^2 = 225$

6 $SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 75 - \frac{225}{5} = 75 - 45 = 30$

7 $s^2 = \frac{SS}{n-1} = \frac{30}{4} = 7.50$ **8** $s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{30}{4}} = \sqrt{7.50} = 2.74$

Table 2.25 Calculation of Sample Standard Deviation (S) (Computation Formula)

➢ **DEGREES OF FREEDOM (df)**

- Degrees of freedom (df) refers to the number of values that are free to vary, given one or more mathematical restrictions
- Degrees of freedom rewrite the formulas for the sample variance and standard deviation:

VARIANCE FOR SAMPLE

$$s^2 = \frac{SS}{n-1} = \frac{SS}{df}$$

STANDARD DEVIATION FOR SAMPLE

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$$

where s^2 and s represent the sample variance and standard deviation, SS is the sum of squares and df is the degrees of freedom and equals n − 1.

➢ **INTERQUARTILE RANGE (IQR)**

- The interquartile range (IQR), is the range for the middle 50 percent of the scores.

A. INSTRUCTIONS

1. Order scores from least to most.
2. To determine how far to penetrate the set of ordered scores, begin at either end, then add 1 to the total number of scores and divide by 4. If necessary, round the result to the nearest whole number.
3. Beginning with the largest score, count the requisite number of steps (calculated in step 2) into the ordered scores to find the location of the third quartile.
4. The third quartile equals the value of the score at this location.
5. Beginning with the smallest score, again count the requisite number of steps into the ordered scores to find the location of the first quartile.
6. The first quartile equals the value of the score at this location.
7. The IQR equals the third quartile minus the first quartile.

B. EXAMPLE

1. 7, 9, 9, 10, 11, 11, 13
2. (7 + 1)/4 = 2
3. 7, 9, 9, 10, 11, 11, 13
   ↑
   2 1
4. third quartile = 11
5. 7, 9, 9, 10, 11, 11, 13
   ↑
   1 2
6. first quartile = 9
7. IQR = 11 − 9 = 2

Table 2.26: Calculation of the IQR

Note: Measures of variability are virtually nonexistent for qualitative and ranked data.

Example 2.16

Using the computation formula for the sum of squares, Calculate the population standard deviation for the scores in (a) and the sample standard deviation for the scores in (b).

(a) 1, 3, 7, 2, 0, 4, 7, 3

$$\sigma = \sqrt{\frac{137 - \frac{729}{8}}{8}} = \sqrt{5.73} = 2.39$$

(b) 10, 8, 5, 0, 1, 1, 7, 9, 2

$$s = \sqrt{\frac{325 - \frac{1849}{9}}{9 - 1}} = \sqrt{14.95} = 3.87$$

Example 2.17

Days absent from school for a sample of 10 first-grade children are: 8, 5, 7, 1, 4, 0, 5, 7, 2, 9.

a) Before calculating the standard deviation, decide whether the definitional or computational formula would be more efficient. Why?

Solution - computation formula since the mean is not a whole number.

b) Use the more efficient formula to calculate the sample standard deviation.

$$s = \sqrt{\frac{314 - \frac{2304}{10}}{10 - 1}} = \sqrt{9.28} = 3.05$$

Example 2.18

As a first step toward modifying his study habits, Phil keeps daily records of his study time.

a. During the first two weeks, Phil's mean study time equals 20 hours per week. If he studied 22 hours during the first week, how many hours did he study during the second week?

Solution - 18 hours

b. During the first four weeks, Phil's mean study time equals 21 hours. If he studied 22, 18, and 21 hours during the first, second,

and third weeks, respectively, how many hours did he study during the fourth week?

Solution - 23 hours

c. If the information in (a) and (b) is to be used to estimate some unknown population characteristic, the notion of degrees of freedom can be introduced. How many degrees of freedom are associated with (a) and (b)?

Solution - df = 1 in (a) and df = 3 in (b)

d. Describe the mathematical restriction that causes a loss of degrees of freedom in (a) and (b).

Solution - When all observations are expressed as deviations from their mean, the sum of all deviations must equal zero.

Example 2.19

Determine the values of the range and the IQR for the following sets of data.

a. Retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63

range = 25; IQR = 65 – 60 = 5

b. Residence changes: 1, 3, 4, 1, 0, 2, 5, 8, 0, 2, 3, 4, 7, 11, 0, 2, 3, 4

range = 11; IQR = 4 – 1 = 3

9 Discuss in detail about Normal Distribution, Normal Curve, Z - Scores and list out the properties of normal Curve?

Normal Curve

- A theoretical curve noted for its symmetrical bell-shaped form.

Important properties of the normal curve:

- The normal curve is a theoretical curve defined for a continuous variable, in symmetrical bell-shaped form**.**
- Because the normal curve is symmetrical, its lower half is the mirror image of its upper half.
- Being bell shaped, the normal curve peaks above a point midway along the horizontal spread and then tapers off gradually in either direction from the peak.
- The values of the mean, median and mode, located at a point midway along the horizontal spread, are the same for the normal curve.

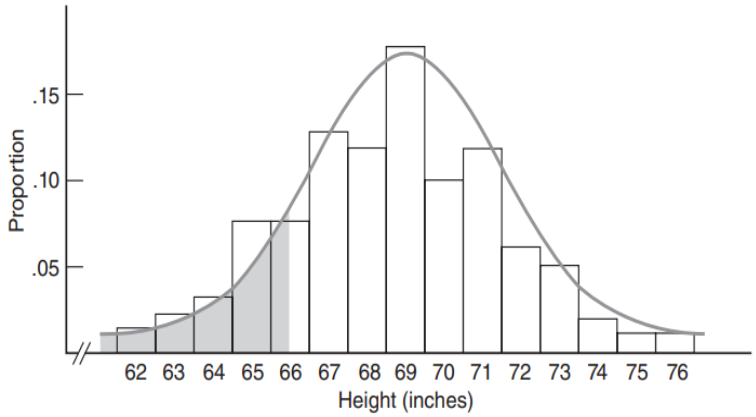
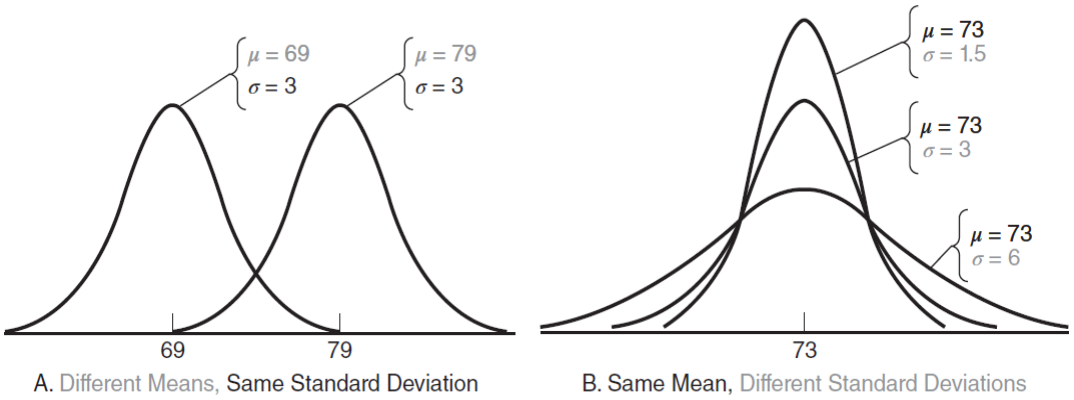


FIGURE 2.12 Normal curve superimposed on the distribution of heights.

Different Normal Curves



Standard Normal Curve.

- Standard Normal Curve is the **one normal curve** for which a table is actually available.
- **Standard Normal Curve** is the tabled normal curve for z scores, with a mean of 0 and a standard deviation of 1.
- To verify (rather than prove) that the mean of a standard normal distribution equals 0, replace X in the z score formula with μ, the mean of any (nonstandard) normal distribution, and then solve for z:

$$\text{Mean of } z = \frac{X-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = \frac{0}{\sigma} = 0$$

- To verify that the standard deviation of the standard normal distribution equals 1, replace X in the z score formula with μ + 1σ, the value corresponding to one standard deviation above the mean for any (nonstandard) normal distribution, and then solve for z:

$$\text{Standard deviation of } z = \frac{X-\mu}{\sigma} = \frac{\mu+1\sigma-\mu}{\sigma} = \frac{1\sigma}{\sigma} = 1$$

- Although there is an infinite number of different normal curves, each with its own mean and standard deviation, there is only one standard normal curve, with a mean of 0 and a standard deviation of 1.

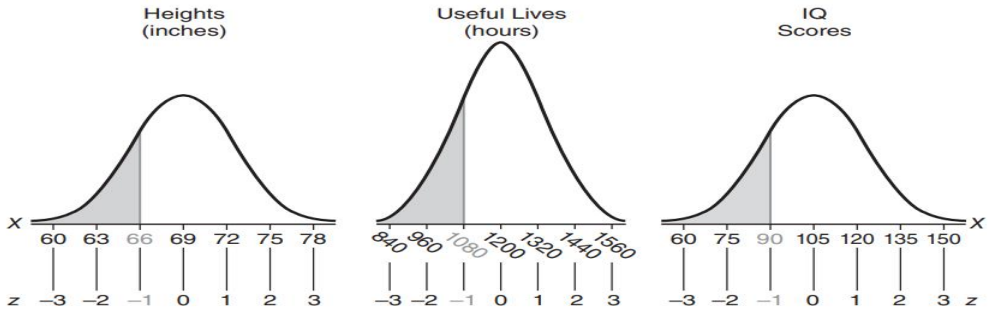


FIGURE 2.13 Converting three normal curves to the standard normal curve.

- **Figure 2.13** illustrates the emergence of the standard normal curve from three different normal curves: that for the men's heights, with a mean of 69 inches and a standard deviation of 3 inches; that for the useful lives of 100-watt electric light bulbs, with a mean of 1200 hours and a standard deviation of 120 hours; and that for the IQ scores of fourth graders, with a mean of 105 points and a standard deviation of 15 points.
- Converting all original observations into z scores leaves the normal shape intact but not the units of measurement. Shaded observations of 66 inches, 1080 hours, and 90 IQ points all reappear as a z score of –1.00.

10 Discuss in detail about z SCORES.

- A z score is a unit-free, standardized score that, indicates how many standard deviations a score is above or below the mean of its distribution.

z SCORE

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

- where X is the original score and μ and σ are the mean and the standard deviation, respectively, for the normal distribution of the original scores.
- A z score consists of two parts:
  1. a positive or negative sign indicating whether it's above or below the mean;
  2. a number indicating the size of its deviation from the mean in standard deviation units.

Example

- A z score of 2.00 always signifies that the original score is exactly two standard deviations above its mean.
- Similarly, a z score of –1.27 signifies that the original score is exactly 1.27 standard deviations below its mean.

- A z score of 0 signifies that the original score coincides with the mean.

Example 2.20

Express each of the following scores as a z score:

a. Margaret's IQ of 135, given a mean of 100 and a standard deviation of 15

b. a score of 470 on the SAT math test, given a mean of 500 and a standard deviation of 100

c. a daily production of 2100 loaves of bread by a bakery, given a mean of 2180 and a standard deviation of 50

d. Sam's height of 69 inches, given a mean of 69 and a standard deviation of 3

e. a thermometer-reading error of –3 degrees, given a mean of 0 degrees and a standard deviation of 2 degrees

(a) 2.33 (b) –0.30 (c) –1.60 (d) 0.00 (e) –1.50

Standard Normal Table

- The standard normal table consists of columns of z scores coordinated with columns of proportions. In a typical problem, access to the table is gained through a z score, such as –1.00.

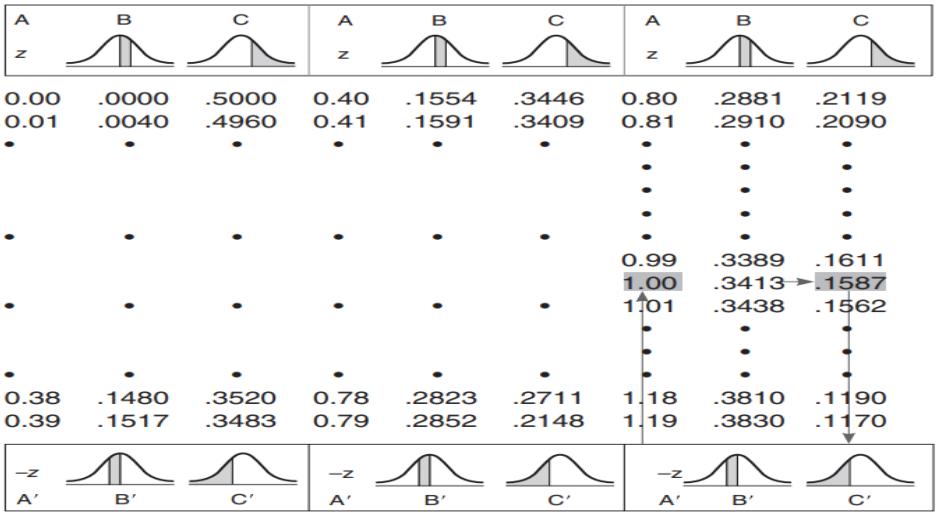


TABLE 2.27 Proportions (Of Areas) Under The Standard Normal Curve For Values Of Z

Using the Top Legend of the Table

- **Table 2.27** shows an abbreviated version of the standard normal curve, The columns are arranged in sets of three, designated as A, B, and C in the legend at the top of the table.

- When using the top legend, all entries refer to the upper half of the standard normal curve. The entries in column A are z scores, beginning with 0.00 and ending with 4.00.
- Given a z score of zero or more, columns B and C indicate how the z score splits the area in the upper half of the normal curve.
- The shading in the top legend, column B indicates the proportion of area between the mean and the z score, and column C indicates the proportion of area beyond the z score, in the upper tail of the standard normal curve.

Using the Bottom Legend of the Table

- The columns are designated as A′, B′, and C′ in the legend at the bottom of the table. When using the bottom legend, all entries refer to the lower half of the standard normal curve.
- Imagine that the nonzero entries in column A′ are negative z scores, beginning with –0.01 and ending with –4.00.
- Given a negative z score, columns B′ and C′ indicate how that z score splits the lower half of the normal curve.
- The shading in the bottom legend of the table, column B′ indicates the proportion of area between the mean and the negative z score, and column C′ indicates the proportion of area beyond the negative z score, in the lower tail of the standard **normal curve.**

Example 2.21

Using Table A in Appendix C, find the proportion of the total area identified with the following statements:

(a) above a z score of 1.80

(b) between the mean and a z score of –0.43

(c) below a z score of –3.00

(d) between the mean and a z score of 1.65

(e) between z scores of 0 and –1.96

Solution : (a) .0359 (b) .1664 (c) .0013 (d) .4505 (e) .4750

11 Explain in detail about finding proportions and Finding Scores.

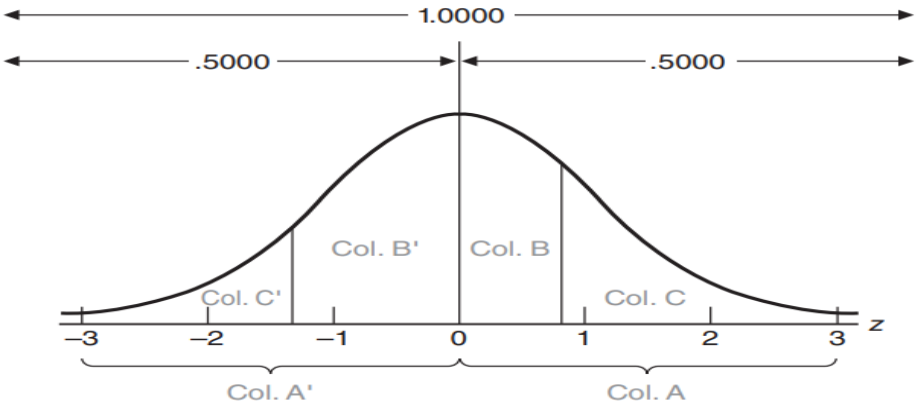


FIGURE 2.14 Normal Curve Problems

- When using the standard normal table, the corresponding proportions in columns B and C (or columns B′ and C′) always sum to .5000.
- Similarly, the total area under the normal curve always equals 1.0000, the sum of the proportions in the lower and upper halves, that is, .5000 + .5000.
- Finally, although a z score can be either positive or negative, the proportions of area under the curve are always positive or zero but never negative.

Steps for finding proportions

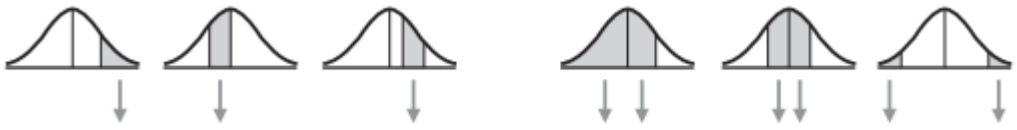
DOING NORMAL CURVE PROBLEMS

Read the problem carefully to determine whether a proportion or a score is to be found.

----------FINDING PROPORTIONS ----------

1. Sketch the normal curve and shade in the target area.

Examples: One Area Two Areas



2. Plan the solution in terms of the normal table.

3. Convert *X* to z: $z = \frac{X - \mu}{\sigma}$

4. Find the target area by entering either column A or A' with *z*, and noting the corresponding proportion from column B, C, B', or C'.

Finding proportion for One Score

- Find the proportion who are shorter than exactly 66 inches, given that the distribution of heights approximates a normal curve with a mean of 69 inches and a standard deviation of 3 inches.

1. Sketch a normal curve and shade in the target area. Being less than the mean of 69, 66 is located to the left of the mean.

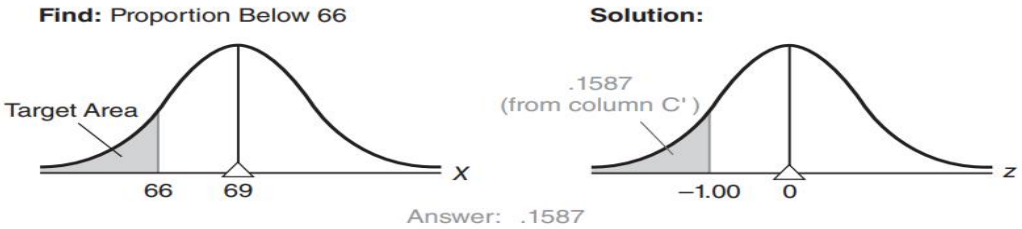


FIGURE 2.15: Finding proportions.

2. Plan your solution according to the normal table.
3. Convert X to z. Express 66 as a z score:

$$z = \frac{X - \mu}{\sigma} = \frac{66 - 69}{3} = \frac{-3}{3} = -1$$

4. Find the target area. and note the corresponding proportion of .1587 in column C':

Example 2.22

Assume that GRE scores approximate a normal curve with a mean of 500 and a standard deviation of 100.

(a) Sketch a normal curve and shade in the target area described by each of the following statements:

(i) less than 400

(ii) more than 650

(iii) less than 700

(b) Plan solutions (in terms of columns B, C, B′, or C′ of the standard normal table, as well as the fact that the proportion for either the entire upper half or lower half always equals .5000) for the target areas in part (a).

(c) Convert to z scores and find the proportions that correspond to the target areas in part (a)

(a_1) 	(b_1) C′	(c_1) $z = -1.00$ answer = .1587
(a_2) 	(b_2) C	(c_2) $z = 1.50$ answer = .0668
(a_3) 	(b_3) .5000 + B	(c_3) $z = 2.00$ answer = .5000 + .4772 = .9772

Finding Proportions between Two Scores

- Assume that, the gestation periods for human fetuses approximate a normal curve with a mean of 270 days (9 months) and a standard deviation of 15 days. What proportion of gestation periods will be between 245 and 255 days?

1. Sketch a normal curve and shade in the target area, as in the top panel of Figure 5.7.

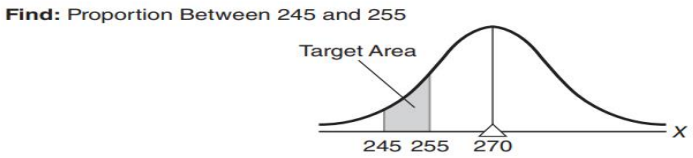
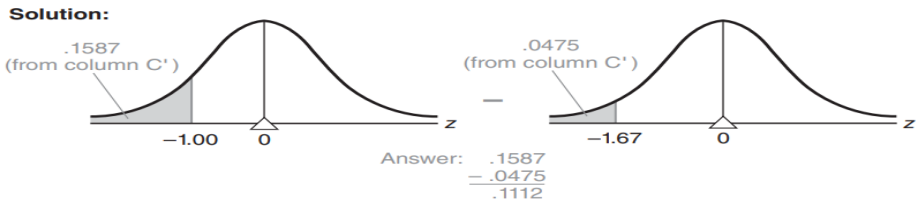


FIGURE 2.16: Finding proportions

2. Plan your solution according to the normal table. The basic idea is to identify the target area with the difference between two overlapping areas whose values can be read from column C′ of Table A. The larger area (less than 255 days) contains two sectors: the target area (between 245 and 255 days) and a remainder (less than 245 days). The smaller area contains only the remainder (less than 245 days). Subtracting the smaller area (less than 245 days) from the larger area (less than 255 days), therefore, eliminates the common remainder (less than 245 days), leaving only the target area (between 245 and 255 days).



3. Convert X to z by expressing 255 as

$$z = \frac{255 - 270}{15} = \frac{-15}{15} = -1.00$$

and by expressing 245 as

$$z = \frac{245 - 270}{15} = \frac{-25}{15} = -1.67$$

4. Find the target area.

Finding Proportions beyond Two Scores

- Assume that high school students' IQ scores approximate a normal distribution with a mean of 105 and a standard deviation of 15. What proportion of IQs are more than 30 points either above or below the mean?

1. Sketch a normal curve and shade in the two target areas, as in the top panel of **Figure 2.17**

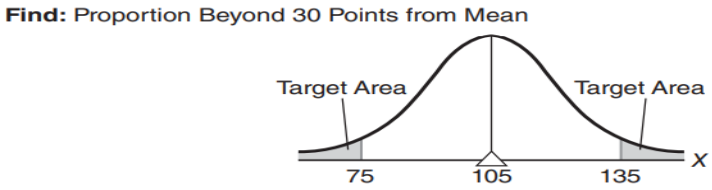
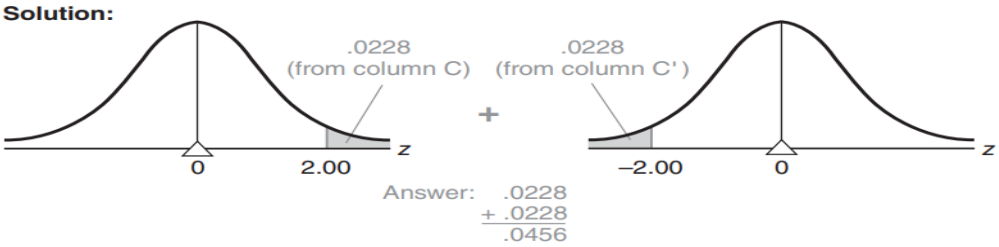


Figure 2.17 Finding proportions

2. Plan your solution according to the normal table.



3. Convert X to z by expressing IQ scores of 135 and 75 as

$$z = \frac{135-105}{15} = \frac{30}{15} = 2.00$$

$$z = \frac{75-105}{15} = \frac{-30}{15} = -2.00$$

4. Find the target area. In Table A, locate a z score of 2.00 in column A, and note the corresponding proportion of .0228 in column C. Because of the symmetry of the normal curve, you need not enter the table again to find the proportion below z score of –2.00. Instead, merely double the above proportion of .0228 to obtain .0456, which represents the proportion of students with IQs more than 30 points either above or below the mean.

Example 2.23:

Assume that SAT math scores approximate a normal curve with a mean of 500 and a standard deviation of 100.

(a) Sketch a normal curve and shade in the target area(s) described by each of the following statements:

(i) more than 570

(ii) less than 515

(iii) between 520 and 540

(iv) between 470 and 520

(v) more than 50 points above the mean

(vi) more than 100 points either above or below the mean

(vii) within 50 points either above or below the mean</RLLNL>

(b) Plan solutions (in terms of columns B, C, B′ , and C′) for the target areas in part (a).

(c) Convert to z scores and find the target areas in part (a).

Solution:

(a_1) 	(b_1) C	(c_1) $z = 0.70$ .2420
(a_2) 	(b_2) .5000 + B	(c_2) $z = 0.15$ .5000 + .0596 = .5596
(a_3) 	(b_3) larger B – smaller B or larger C – smaller C	(c_3) $z = 0.20$; $z = 0.40$ .1554 – .0793 = .0761 or .4207 – .3446 = .0761
(a_4) 	(b_4) B′ + B	(c_4) $z = –0.30$; $z = 0.20$ .1179 + .0793 = .1972
(a_5) 	(b_5) C	(c_5) $z = 0.50$ .3085
(a_6) 	(b_6) C′ + C or 2(C)	(c_6) $z = –1.00$; $z = 1.00$ .1587 + .1587 = .3174
(a_7) 	(b_7) B′ + B or 2(B)	(c_7) $z = –0.50$; $z = 0.50$ .1915 + .1915 = .3830

Table A[a]
PROPORTIONS (OF AREA) UNDER THE STANDARD NORMAL CURVE FOR VALUES OF *z*

A z	B	C
0.00	.0000	.5000
0.01	.0040	.4960
0.02	.0080	.4920
0.03	.0120	.4880
0.04	.0160	.4840
0.05	.0199	.4801
0.06	.0239	.4761
0.07	.0279	.4721
0.08	.0319	.4681
0.09	.0359	.4641
0.10	.0398	.4602
0.11	.0438	.4562
0.12	.0478	.4522
0.13	.0517	.4483
0.14	.0557	.4443
0.15	.0596	.4404
0.16	.0636	.4364
0.17	.0675	.4325
0.18	.0714	.4286
0.19	.0753	.4247
0.20	.0793	.4207
0.21	.0832	.4168
0.22	.0871	.4129
0.23	.0910	.4090
0.24	.0948	.4052
0.25	.0987	.4013
0.26	.1026	.3974
0.27	.1064	.3936
0.28	.1103	.3897
0.29	.1141	.3859
0.30	.1179	.3821
0.31	.1217	.3783
0.32	.1255	.3745
0.33	.1293	.3707
0.34	.1331	.3669
0.35	.1368	.3632
0.36	.1406	.3594
0.37	.1443	.3557
0.38	.1480	.3520
0.39	.1517	.3483
0.40	.1554	.3446
0.41	.1591	.3409
0.42	.1628	.3372
0.43	.1664	.3336
0.44	.1700	.3300
0.45	.1736	.3264
0.46	.1772	.3228
0.47	.1808	.3192
0.48	.1844	.3156
0.49	.1879	.3121
0.50	.1915	.3085
0.51	.1950	.3050
0.52	.1985	.3015
0.53	.2019	.2981
0.54	.2054	.2946
0.55	.2088	.2912
0.56	.2123	.2877
0.57	.2157	.2843
0.58	.2190	.2810
0.59	.2224	.2776
0.60	.2257	.2743
0.61	.2291	.2709
0.62	.2324	.2676
0.63	.2357	.2643
0.64	.2389	.2611
0.65	.2422	.2578
0.66	.2454	.2546
0.67	.2486	.2514
0.68	.2517	.2483
0.69	.2549	.2451
0.70	.2580	.2420
0.71	.2611	.2389
0.72	.2642	.2358
0.73	.2673	.2327
0.74	.2704	.2296
0.75	.2734	.2266
0.76	.2764	.2236
0.77	.2794	.2206
0.78	.2823	.2177
0.79	.2852	.2148
0.80	.2881	.2119
0.81	.2910	.2090
0.82	.2939	.2061
0.83	.2967	.2033
0.84	.2995	.2005
0.85	.3023	.1977
0.86	.3051	.1949
0.87	.3078	.1922
0.88	.3106	.1894
0.89	.3133	.1867
0.90	.3159	.1841
0.91	.3186	.1814
0.92	.3212	.1788
0.93	.3238	.1762
0.94	.3264	.1736
0.95	.3289	.1711
0.96	.3315	.1685
0.97	.3340	.1660
0.98	.3365	.1635
0.99	.3389	.1611
1.00	.3413	.1587
1.01	.3438	.1562
1.02	.3461	.1539
1.03	.3485	.1515
1.04	.3508	.1492
1.05	.3531	.1469
1.06	.3554	.1446
1.07	.3577	.1423
1.08	.3599	.1401
1.09	.3621	.1379
1.10	.3643	.1357
1.11	.3665	.1335
1.12	.3686	.1314
1.13	.3708	.1292
1.14	.3729	.1271
1.15	.3749	.1251
1.16	.3770	.1230
1.17	.3790	.1210
1.18	.3810	.1190
1.19	.3830	.1170
1.20	.3849	.1151
1.21	.3869	.1131
1.22	.3888	.1112
1.23	.3907	.1093
1.24	.3925	.1075
1.25	.3944	.1056
1.26	.3962	.1038
1.27	.3980	.1020
1.28	.3997	.1003
1.29	.4015	.0985
1.30	.4032	.0968
1.31	.4049	.0951
1.32	.4066	.0934
1.33	.4082	.0918
1.34	.4099	.0901
1.35	.4115	.0885
1.36	.4131	.0869
1.37	.4147	.0853
1.38	.4162	.0838
1.39	.4177	.0823
1.40	.4192	.0808
1.41	.4207	.0793
1.42	.4222	.0778
1.43	.4236	.0764
1.44	.4251	.0749
1.45	.4265	.0735
1.46	.4279	.0721
1.47	.4292	.0708
1.48	.4306	.0694
1.49	.4319	.0681
1.50	.4332	.0668
1.51	.4345	.0655
1.52	.4357	.0643
1.53	.4370	.0630
1.54	.4382	.0618
1.55	.4394	.0606
1.56	.4406	.0594
1.57	.4418	.0582
1.58	.4429	.0571
1.59	.4441	.0559
1.60	.4452	.0548
1.61	.4463	.0537
1.62	.4474	.0526
1.63	.4484	.0516
1.64	.4495	.0505
1.65	.4505	.0495
1.66	.4515	.0485
1.67	.4525	.0475
−z A′	B′	C′

Table A[a] (Continued)

PROPORTIONS (OF AREA) UNDER THE STANDARD NORMAL CURVE FOR VALUES OF *z*

A z	B	C
1.68	.4535	.0465
1.69	.4545	.0455
1.70	.4554	.0446
1.71	.4564	.0436
1.72	.4573	.0427
1.73	.4582	.0418
1.74	.4591	.0409
1.75	.4599	.0401
1.76	.4608	.0392
1.77	.4616	.0384
1.78	.4625	.0375
1.79	.4633	.0367
1.80	.4641	.0359
1.81	.4649	.0351
1.82	.4656	.0344
1.83	.4664	.0336
1.84	.4671	.0329
1.85	.4678	.0322
1.86	.4686	.0314
1.87	.4693	.0307
1.88	.4699	.0301
1.89	.4706	.0294
1.90	.4713	.0287
1.91	.4719	.0281
1.92	.4726	.0274
1.93	.4732	.0268
1.94	.4738	.0262
1.95	.4744	.0256
1.96	.4750	.0250
1.97	.4756	.0244
1.98	.4761	.0239
1.99	.4767	.0233
2.00	.4772	.0228
2.01	.4778	.0222
2.02	.4783	.0217
2.03	.4788	.0212
2.04	.4793	.0207
2.05	4798	.0202
2.06	.4803	.0197
2.07	.4808	.0192
2.08	.4812	.0188
2.09	.4817	.0183
2.10	.4821	.0179
2.11	.4826	.0174
2.12	.4830	.0170
2.13	.4834	.0166
2.14	.4838	.0162
2.15	.4842	.0158
2.16	.4846	.0154
2.17	.4850	.0150
2.18	.4854	.0146
2.19	.4857	.0143
2.20	.4861	.0139
2.21	.4864	.0136
2.22	.4868	.0132
2.23	.4871	.0129
2.24	.4875	.0125
2.25	.4878	.0122
2.26	.4881	.0119
2.27	.4884	.0116
2.28	.4887	.0113
2.29	.4890	.0110
2.30	.4893	.0107
2.31	.4896	.0104
2.32	.4898	.0102
2.33	.4901	.0099
2.34	.4904	.0096
2.35	.4906	.0094
2.36	.4909	.0091
2.37	.4911	.0089
2.38	.4913	.0087
2.39	.4916	.0084
2.40	.4918	.0082
2.41	.4920	.0080
2.42	.4922	.0078
2.43	.4925	.0075
2.44	.4927	.0073
2.45	.4929	.0071
2.46	.4931	.0069
2.47	.4932	.0068
2.48	.4934	.0066
2.49	.4936	.0064
2.50	.4938	.0062
2.51	.4940	.0060
2.52	.4941	.0059
2.53	.4943	.0057
2.54	.4945	.0055
2.55	.4946	.0054
2.56	.4948	.0052
2.57	.4949	.0051
2.58	.4951	.0049
2.59	.4952	.0048
2.60	.4953	.0047
2.61	.4955	.0045
2.62	.4956	.0044
2.63	.4957	.0043
2.64	.4959	.0041
2.65	.4960	.0040
2.66	.4961	.0039
2.67	.4962	.0038
2.68	.4963	.0037
2.69	.4964	.0036
2.70	.4965	.0035
2.71	.4966	.0034
2.72	.4967	.0033
2.73	.4968	.0032
2.74	.4969	.0031
2.75	.4970	.0030
2.76	.4971	.0029
2.77	.4972	.0028
2.78	.4973	.0027
2.79	.4974	.0026
2.80	.4974	.0026
2.81	.4975	.0025
2.82	.4976	.0024
2.83	.4977	.0023
2.84	.4977	.0023
2.85	.4978	.0022
2.86	.4979	.0021
2.87	.4979	.0021
2.88	.4980	.0020
2.89	.4981	.0019
2.90	.4981	.0019
2.91	.4982	.0018
2.92	.4982	.0018
2.93	.4983	.0017
2.94	.4984	.0016
2.95	.4984	.0016
2.96	.4985	.0015
2.97	.4985	.0015
2.98	.4986	.0014
2.99	.4986	.0014
3.00	.4987	.0013
3.01	.4987	.0013
3.02	.4987	.0013
3.03	.4988	.0012
3.04	.4988	.0012
3.05	.4989	.0011
3.06	.4989	.0011
3.07	.4989	.0011
3.08	.4990	.0010
3.09	.4990	.0010
3.10	.4990	.0010
3.11	.4991	.0009
3.12	.4991	.0009
3.13	.4991	.0009
3.14	.4992	.0008
3.15	.4992	.0008
3.16	.4992	.0008
3.17	.4992	.0008
3.18	.4993	.0007
3.19	.4993	.0007
3.20	.4993	.0007
3.21	.4993	.0007
3.22	.4994	.0006
3.23	.4994	.0006
3.24	.4994	.0006
3.25	.4994	.0006
3.30	.4995	.0005
3.35	.4996	.0004
3.40	.4997	.0003
3.45	.4997	.0003
3.50	.4998	.0002
3.60	.4998	.0002
3.70	.4999	.0001
3.80	.4999	.0001
3.90	.49995	.00005
4.00	.49997	.00003

−z A′	B′	C′

12. Explain in detail about correlations and the types of relationships in correlation.

CONTENTS:

- **Correlation**
- **Types of relationship**
  - Positive Relationship
  - Negative Relationship
  - Little or no relationship
- **Correlation Coefficient r**
  - **Key properties of r**
- **Computational formula for Correlation Coefficient**
- **Computational Sequence**
- **Data and Computation**

Correlation

- Two variables are related if pairs of scores show an orderliness that can be depicted graphically with a scatter plot and numerically with a correlation coefficient.

Table 2.8 Greeting Cards Sent and Received by five Friends

	NUMBER OF CARDS	
FRIEND	SENT	RECEIVED
Andrea	5	10
Mike	7	12
Doris	13	14
Steve	9	18
John	1	6

- **Types of relationship:**
  1. Positive Relationship
  2. Negative Relationship
  3. Little or no relationship

To illustrate the types of relationships refer Table 2.8 Greeting cards sent and received by friends.

Positive relationship:

- Two variables are positively related if pairs of scores tend to occupy similar relative positions (high with high and low with low) in their respective distributions.

- Trends among pairs of scores can be detected most easily by constructing a list of paired scores in which the scores along one variable are arranged from largest to smallest.
- In Table 2.9, the five pairs of scores are arranged from the largest (13) to the smallest (1) number of cards sent.
- Relatively low values are paired with relatively low values, and relatively high values are paired with relatively high values, the relationship is positive.

Table 2.9 Positive Relationship

A. POSITIVE RELATIONSHIP

FRIEND	SENT	RECEIVED
Doris	13	14
Steve	9	18
Mike	7	12
Andrea	5	10
John	1	6

Negative relationship:

- Two variables are negatively related if pairs of scores tend to occupy dissimilar relative positions (high with low and vice versa) in their respective distributions.
- Occurs insofar as pairs of scores tend to occupy dissimilar relative positions (high with low and vice versa) in their respective distributions the relationship is negative.
- Notice the pattern among the pairs in Table 2.10. Now there is a pronounced tendency for pairs of scores to occupy dissimilar and opposite relative positions in their respective distributions.
- This relationship implies that relatively low values are paired with relatively high values, and relatively high values are paired with relatively low values, the relationship is negative.

Table 2.10 Negative Relationship

B. NEGATIVE RELATIONSHIP

FRIEND	SENT	RECEIVED
Doris	13	6
Steve	9	10
Mike	7	14
Andrea	5	12
John	1	18

Little or No Relationship

- No regularity is apparent among the pairs of scores in Table 2.11.

- For instance, although both Andrea and John sent relatively few cards (5 and 1, respectively), Andrea received relatively few cards (6) and John received relatively many cards (14).
- Given this lack of regularity, if any, relationship exists between the two variables it is little or no relationship

Table 2.12 Little or No Relationship

FRIEND	SENT	RECEIVED
Doris	13	10
Steve	9	18
Mike	7	12
Andrea	5	6
John	1	14

Example – 2.34

Indicate whether the following statements suggest a positive or negative relationship:

a. More densely populated areas have higher crime rates.

b. Schoolchildren who often watch TV perform more poorly on academic achievement tests.

c. Heavier automobiles yield poorer gas mileage.

d. Better-educated people have higher incomes.

e. More anxious people voluntarily spend more time performing a simple repetitive task.

Solution:

a. Positive. The crime rate is higher, square mile by square mile, in densely populated cities than in sparsely populated rural areas.

b. Negative. As TV viewing increases, performance on academic achievement tests tends to decline.

c. Negative. Increases in car weight are accompanied by decreases in miles per gallon.

d. Positive. Increases in educational level—grade school, high school, college—tend to be associated with increases in income.

e. Positive. Highly anxious people willingly spend more time performing a simple repetitive task than do less anxious people.

CORRELATION COEFFICIENT R

- A correlation coefficient is a number between –1 and 1 that describes the relationship between pairs of variables.

Key Properties of r

- Named in honor of the British scientist Karl Pearson, the Pearson correlation coefficient, r, can equal any value between –1.00 and +1.00.

- Furthermore, the following two properties apply:
  1. The sign of r indicates the type of linear relationship, whether positive or negative.

     A number with a plus sign (or no sign) indicates a positive relationship, and a number with a minus sign indicates a negative relationship.
  2. The numerical value of r, indicates the strength of the linear relationship.

     The more closely a value of r approaches either –1.00 or +1.00, the stronger (more regular) the relationship.

     The more closely the value of r approaches 0, the weaker (less regular) the relationship.

➢ **COMPUTATIONAL FORMULA FOR CORRELATION COEFFICIENT**

Calculate a value for r by using the following computation formula:

CORRELATION COEFFICIENT (COMPUTATION FORMULA)

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

Where

- SP_{xy}, - Sum of the products for each pair of deviation scores defined as

SUM OF PRODUCTS (DEFINITION AND COMPUTATION FORMULAS)

$$SP_{xy} = \Sigma\left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}$$

- SS_x and SS_y - summing the squared deviation scores for either X or Y is defines as

$$SS_x = \Sigma\left(X - \bar{X}\right)^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$$SS_y = \Sigma\left(Y - \bar{Y}\right)^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$

- Notice in Formula that, since the terms in the denominator must be positive, only the sum of the products, SP_{xy}, determines whether the value of *r* is positive or negative.
- Furthermore, the size of SP_{xy} mirrors the strength of the relationship; stronger relationships are associated with larger positive or negative sums of products.

➢ **Computational Sequence**

- Assign a value to n (1), representing the number of pairs of scores.
- Sum all scores for X (2) and for Y (3).
- Find the product of each pair of X and Y scores (4), one at a time, then add all of these products (5).
- Square each X score (6), one at a time, then add all squared X scores (7).
- Square each Y score (8), one at a time, then add all squared Y scores (9).
- Substitute numbers into formulas (10) and solve for SP_{xy}, SS_x, and SS_y.
- Substitute into formula (11) and solve for r.

➢ **Data and Computation**

	CARDS		4	6	8
FRIEND	SENT, X	RECEIVED, Y	XY	X^2	Y^2
Doris	13	14	182	169	196
Steve	9	18	162	81	324
Mike	7	12	84	49	144
Andrea	5	10	50	25	100
John	1	6	6	1	36

1 $n = 5$ 2 $\Sigma X = 35$ 3 $\Sigma Y = 60$ 5 $\Sigma XY = 484$ 7 $\Sigma X^2 = 325$ 9 $\Sigma Y^2 = 800$

10 $$SP_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 484 - \frac{(35)(60)}{5} = 484 - 420 = 64$$

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 325 - \frac{(35)^2}{5} = 325 - 245 = 80$$

$$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 800 - \frac{(60)^2}{5} = 800 - 720 = 80$$

11 $$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{64}{\sqrt{(80)(80)}} = \frac{64}{80} = .80$$

Example 2.35:

Supply a verbal description for each of the following correlations.

a. an r of –.84 between total mileage and automobile resale value

b. an r of –.35 between the number of days absent from school and performance on a math achievement test

c. an r of .03 between anxiety level and college GPA

d. an r of .56 between age of schoolchildren and reading comprehension

Solution:

a. Cars with more total miles tend to have lower resale values.

b. Students with more absences from school tend to score lower on math achievement tests.

c. Little or no relationship between anxiety level and college GPA.

d. Older school children tend to have better reading comprehension.

Example 2.36:

Couples who attend a clinic for first pregnancies are asked to estimate (Independently of each other) the ideal number of children. Given that X and Y represent the estimates of females and males, respectively, the results are as follows:

COUPLE	X	Y
A	1	2
B	3	4
C	2	3
D	3	2
E	1	0
F	2	3

Calculate a value for r, using the computation formula .

Solution:

$$r = \frac{4}{\sqrt{(4)(9.33)}} = .65$$

13. Explain in detail about Scatterplots and diagrammatically represent the relationship between variables using scatterplot.

CONTENTS:

- **Scatterplot**
- **Construction of Scatterplot**
- **Types of relationship in Scatterplot**
  - Positive Relationship
  - Negative Relationship
  - Little or no Relationship
  - Strong or Weaker Relationship
  - Linear Relationship
  - Perfect Relationship
  - Curvilinear Relationship

Scatterplot

- A scatterplot is a graph containing a cluster of dots that represents all pairs of scores.

Construction of Scatterplot

- To construct a scatterplot, scale each of the two variables along the horizontal (X) and vertical (Y) axes, and use each pair of scores to locate a dot within the scatterplot.
- For example, the pair of numbers for Mike, 7 and 12, define points along the X and Y axes, respectively.
- Using these points to anchor lines perpendicular (at right angles) to each axis, locate Mike's dot where the two lines intersect.
- Repeat this process, with imaginary lines, for each of the four remaining pairs of scores to create the scatterplot refer Figure 3.1

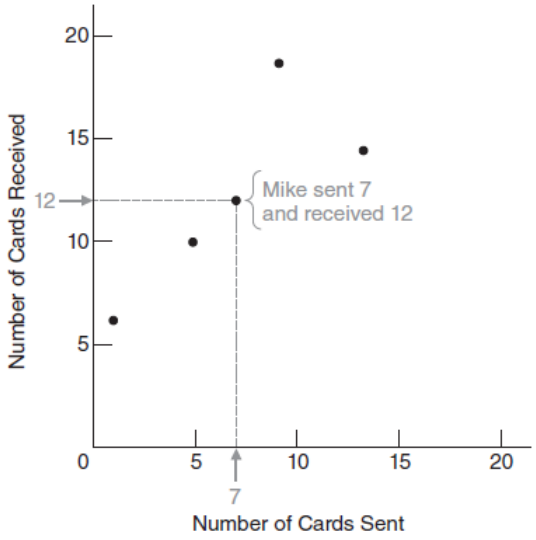


Figure 2.14 – Scatterplot for Greeting Card Exchange

- Figure 2.14 has shown the basic idea of correlation and the construction of a scatterplot.

Types of Relationship in Scatterplot

- Positive Relationship
- Negative Relationship
- Little or no Relationship
- Strong or Weaker Relationship
- Linear Relationship
- Perfect Relationship
- Curvilinear Relationship

Positive Relationship

- A dot cluster that has a slope from the lower left to the upper right, reflects a positive relationship.
- Small values of one variable are paired with small values of the other variable, and large values are paired with large values.
- In Figure 2.15, short people tend to be light, and tall people tend to be heavy.

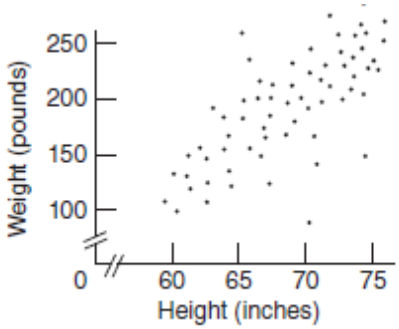


Figure 2.15 – Positive Relationship

Negative Relationship

- A dot cluster that has a slope from the upper left to the lower right, reflects a negative relationship.
- Small values of one variable tend to be paired with large values of the other variable, and vice versa.
- Figure 2.16 reflects the relationship between life expectancy and Heavy Smoking.

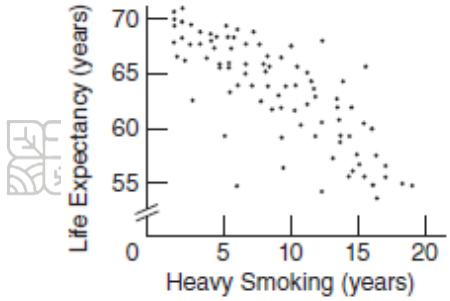


Figure 2.16 – Negative Relationship

Little or No Relationship

- A dot cluster that lacks any apparent slope, reflects little or no relationship.
- Small values of one variable are just as likely to be paired with small, medium, or large values of the other variable.
- In Figure 2.17, dots are seen about in an irregular fashion, suggesting that there is little or no relationship between the height of young adults and their life expectancies.

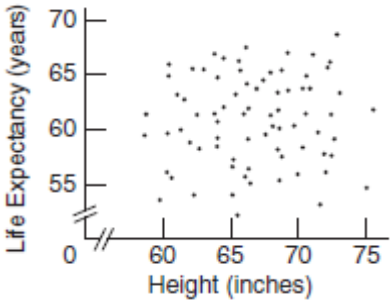


Figure 2.17 – Little or No Relationship

Strong or Weak Relationship

- The more closely the dot cluster approximates a straight line, the stronger the relationship will be.
- The more scattered the dot cluster approximates a weaker the relationship will be.

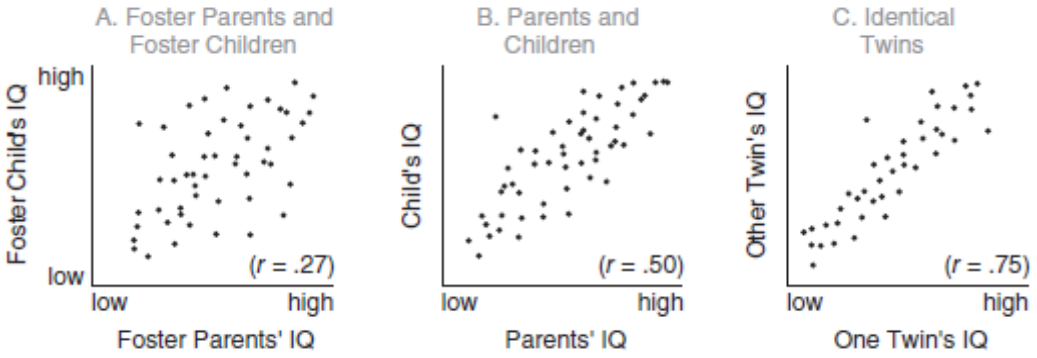


Figure 2.18 – Three Positive Relationships

- Figure 2.18 shows a series of scatterplots, each representing a different positive relationship between IQ scores for pairs of people whose backgrounds reflect different degrees of genetic overlap, ranging from minimum overlap between foster parents and foster children to maximum overlap between identical twins.
- Notice that the dot cluster more closely approximates a straight line for people with greater degrees of genetic overlap—for parents and children in panel B of Figure 3.5 and even more so for identical twins in panel C.

Linear Relationship

- A relationship that can be described best with a straight line.

Perfect Relationship

- A dot cluster that equals (rather than merely approximates) a straight line reflects a perfect relationship between two variables.

Curvilinear Relationship

- Sometimes a dot cluster approximates a bent or curved line, as in Figure 2.19, and therefore reflects a curvilinear relationship.

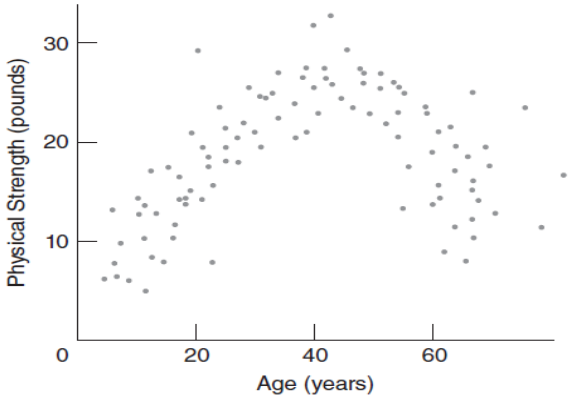
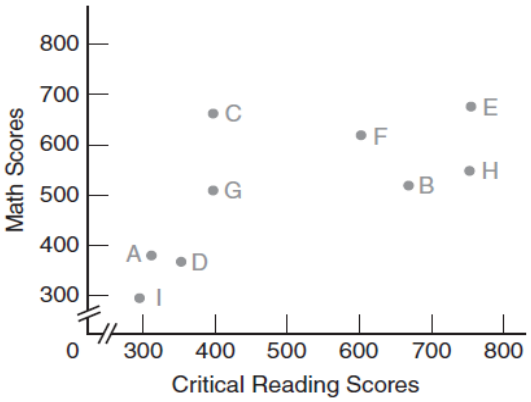


Figure 2.19 – Curvilinear Relationship

Example 2.37:

Critical reading and math scores on the SAT test for students A, B, C, D, E, F, G, and H are shown in the following scatterplot:



a. Which student(s) scored about the same on both tests?

b. Which student(s) scored higher on the critical reading test than on the math test?

c. Which student(s) will be eligible for an honors program that requires minimum scores of

d. 700 in critical reading and 500 in math?

e. Is there a negative relationship between the critical reading and math scores?

Solution:

a. I, D, F

b. B, H, E

c. E, H

d. No. The relationship is positive.

14.Explain in detail about Regression Line.

CONTENTS:

- **Regression Line**
- **Example**
- **Placement of Line**
- **Predictive Errors**
- **Total Predictive Errors**

Regression Line

- The regression line is a straight line rather than a curved line because of the linear relationship between two variables.

Example: Rough Prediction

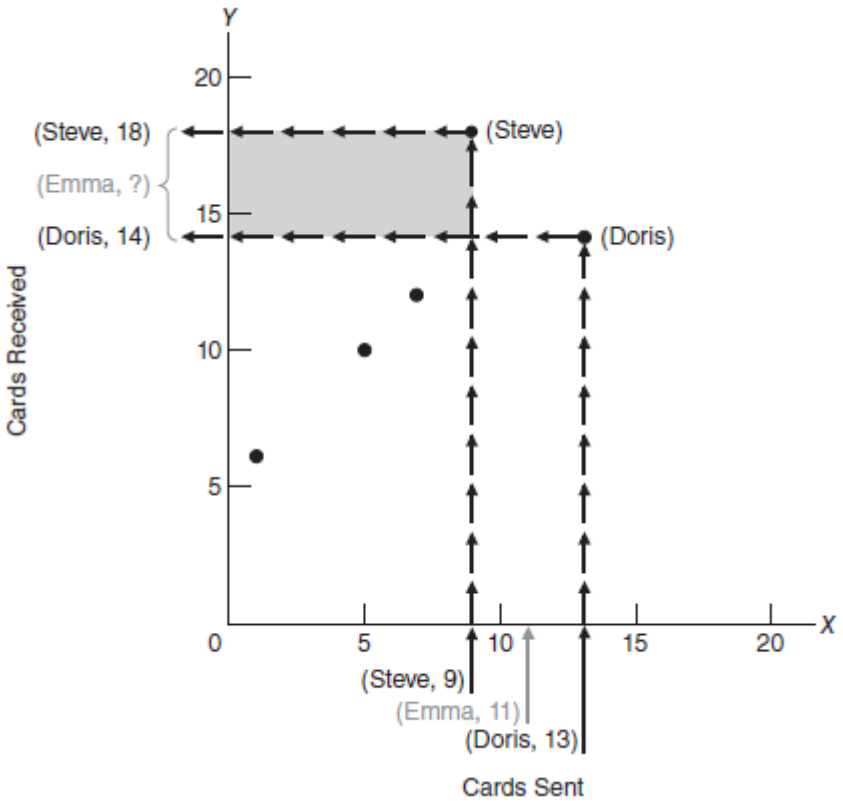


Figure 2.20 – A rough prediction

- To obtain a slightly more precise prediction for Emma, refer to the scatter plot for the original five friends shown in **Figure 2.20.**
- Notice that Emma's plan to send 11 cards locates her along the *X* axis between the 9 cards sent by Steve and the 13 sent by Doris.

- Using the dots for Steve and Doris as guides, construct two strings of arrows, one beginning at 9 and ending at 18 for Steve and the other beginning at 13 and ending at 14 for Doris.
- Focusing on the interval along the *Y* axis between the two strings of arrows, could predict that Emma's return should be between 14 and 18 cards, the numbers received by Doris and Steve.

Prediction using Regression Line

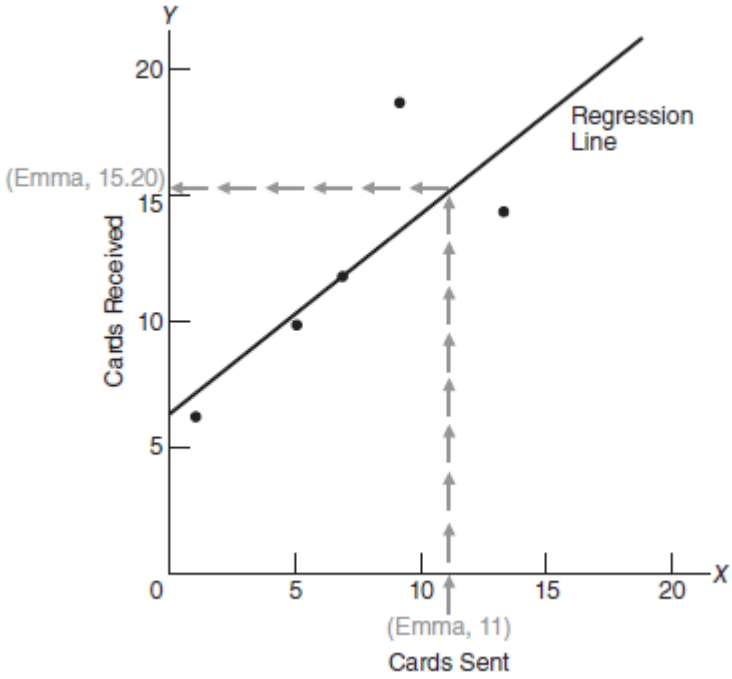


Figure 2.21 - Prediction using the Regression Line

- All five dots contribute to the more precise prediction, illustrated in Figure 2.21, that Emma will receive 15.20 cards.
- The solid line designated as the regression line in Figure 2.21, which guides the string of arrows, beginning at 11, toward the predicted value of 15.20.
- The regression line is a straight line rather than a curved line because of the linear relationship between cards sent and cards received.

Predictive Errors

- Figure 2.22 illustrates the predictive errors that would have occurred if the regression line had been used to predict the number of cards received by the five friends.

- Solid dots reflect the actual number of cards received, and open dots, always located along the regression line, reflect the predicted number of cards received

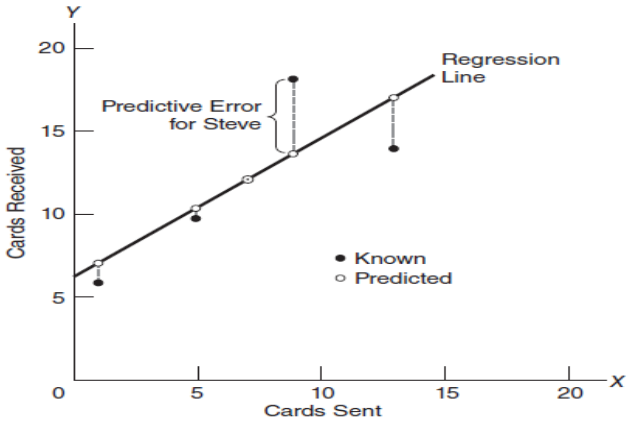


Figure 2.22 - Predictive Errors

Total Predictive Error

- It is desirable for the regression line to be placed in a position that minimizes the total predictive error.

15.Describe about Least squares Regression Line.

CONTENTS:

- **Least squares Regression Line**
- **Least Squares Regression Equation**
- **Computational Sequence**
- **Computations**

Least Squares Regression Line

- The regression line is often referred to as the least squares regression line to minimize the total predictive error thereby providing a more favorable prognosis for the predictions.

Least Squares Regression Equation

- An equation pinpoints the exact least squares regression line for any scatterplot.
- The equation that minimizes the total of all squared prediction errors for known Y scores in the original correlation analysis.

LEAST SQUARES REGRESSION EQUATION

$$Y' = bX + a$$

- where Y´ represents the predicted value;
- X represents the known value;
- b and a represent numbers calculated from the original correlation analysis

Finding Values of b and a

- The expression for b reads:

SOLVING FOR *b*

$$b = r\sqrt{\frac{SS_y}{SS_x}}$$

  - where r represents the correlation between X and Y;
  - SS_y represents the sum of squares for all Y scores;
  - SS_x represents the sum of squares for all X scores

- The expression for a reads:

SOLVING FOR *a*

$$a = \overline{Y} - b\overline{X}$$

where Y and X refer to the sample means for all Y and X scores, respectively, and b is defined by the preceding expression.

Computational Sequence

- Determine values of SS_x, SS_y, and r (1) by referring to the original correlation analysis
- Substitute numbers into the formula (2) and solve for b.
- Assign values to X and Y (3) by referring to the original correlation analysis
- Substitute numbers into the formula (4) and solve for a.
- Substitute numbers for b and a in the least squares regression equation (5).

Computations

1 $SS_x = 80^*$
$SS_y = 80^*$
$r = .80$

2 $b = r\sqrt{\frac{SS_Y}{SS_X}} = .80\sqrt{\frac{80}{80}} = .80$

3 $\overline{X} = 7^{**}$
$\overline{Y} = 12^{**}$

4 $a = \overline{Y} - (b)(\overline{X}) = 12 - (.80)(7) = 12 - 5.60 = 6.40$

5 $Y' = (b)(X) + a$
$= (.80)(X) + 6.40$

where .80 and 6.40 represent the values computed for b and a, respectively.

Example 2.37:

Assume that an r of .30 describes the relationship between educational level (highest grade completed) and estimated number of hours spent reading each week. More specifically:

EDUCATIONAL LEVEL (X)	WEEKLY READING TIME (Y)
$\overline{X} = 13$	$\overline{Y} = 8$
$SS_x = 25$	$SS_y = 50$
$r = .30$	

(a) Determine the least squares equation for predicting weekly reading time from educational level.

(b) Faith's education level is 15. What is her predicted reading time?

(c) Keegan's educational level is 11. What is his predicted reading time?

Solution:

(a) $b = \sqrt{\frac{50}{25}}(.30) = .42;\ a = 8 - (.42)(13) = 2.54$

(b) $Y' = (.42)(15) + 2.54 = 8.84$

(c) $Y' = (.42)(11) + 2.54 = 7.16$

16.Explain in detail about Standard error of estimate.

CONTENTS:

- **Standard Error of Estimate**
- **Finding the Standard Error of Estimate**
- **Definition Formula**
- **Computation Formula**
- **Computational Sequence**
- **Computation**
- **Importance of r**

Standard Error of Estimate

- The task is to estimate the amount of error associated with the predictions.

Finding the Standard Error of Estimate

- The standard error of estimate and symbolized as $S_{y|x}$, for any sample standard deviation, that is, the square root of a sum of squares term divided by its degrees of freedom.
- The symbol $S_{y|x}$ is read as "S sub y given x."

Definition Formula

- The formula for $S_{y|x}$ reads:

STANDARD ERROR OF ESTIMATE (DEFINITION FORMULA)

$$s_{y|x} = \sqrt{\frac{SS_{y|x}}{n-2}} = \sqrt{\frac{\Sigma(Y-Y')^2}{n-2}}$$

where

$SS_{y|x}$, represents the sum of the squares for predictive errors,

$Y - Y'$, and the degrees of freedom term in the denominator,

$n - 2$, reflects the loss of two degrees of freedom

Computation Formula

STANDARD ERROR OF ESTIMATE (COMPUTATION FORMULA)

$$s_{y|x} = \sqrt{\frac{SS_y(1-r^2)}{n-2}}$$

where SS_y is the sum of the squares for Y scores

$$SS_y = \Sigma(Y-\bar{Y}) = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$

r is the correlation coefficient

Computational Sequence

- Assign values to SS_y and r (1) by referring to previous work with the least squares regression equation.
- Substitute numbers into the formula (2) and solve for $s_{y|x}$.

Computation

1 $SS_y = 80$

$r = .80$

2 $s_{y|x} = \sqrt{\frac{SS_y(1-r^2)}{n-2}} = \sqrt{\frac{80\left(1-[.80]^2\right)}{5-2}} = \sqrt{\frac{80(.36)}{3}} = \sqrt{\frac{28.80}{3}} = \sqrt{9.60}$

$= 3.10$

Importance of r

- Let's substitute a few extreme values for r, the sum of squares for predictive errors, $SS_{y|x}$.
- Substituting a value of 1 for r,

$$SS_{y|x} = SS_y(1-r^2) = SS_y[1-(1)^2] = SS_y[1-1] = SS_y[0] = 0$$

when predictions are based on perfect relationships, the sum of squares for predictive errors equals zero, and there is no predictive error.

- Substituting a value of 0 for r,

$$SS_{y|x} = SS_y(1-r^2) = SS_y[1-(0)^2] = SS_y[1-0] = SS_y[1] = SS_y$$

when predictions are based on a nonexistent relationship, the sum of squares for predictive errors equals SS_y, and there is no reduction in predictive error.

17. Explain in detail about Interpretation of r^2

CONTENTS:

- **Squared correlation coefficient, r^2**
- **Two kinds of predictive errors**
- **Repetitive Prediction of the Mean**
- **Predictive Errors**
- **Error Variability (Sum of Squares)**
- **Proportion of Predicted Variability**

Squared correlation coefficient, r^2

- The squared correlation coefficient, r^2, defines the interpretation of the correlation coefficient and also a measure of predictive accuracy that supplements the standard error of estimate, $s_{y|x}$.

Two kinds of predictive errors

- repetitive prediction of the mean
- due to the regression equation.

Repetitive Prediction of the Mean

- Given the present restricted circumstances, statisticians recommend repetitive predictions of the mean, Y, for a variety of reasons, although the predictive error for any individual might be quite large, the sum of all of the resulting predictive errors always equals zero

Predictive Errors

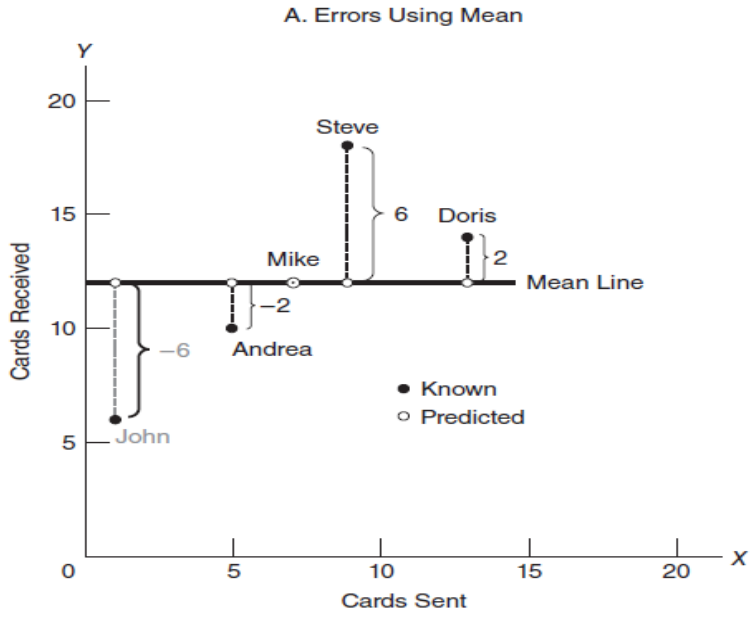


Figure 2.23 – Predictive Error using mean

The figure 2.23 shows the predictive errors for all five friends when the mean for all five friends, Y, of 12 is always used to predict each of their five Y scores.

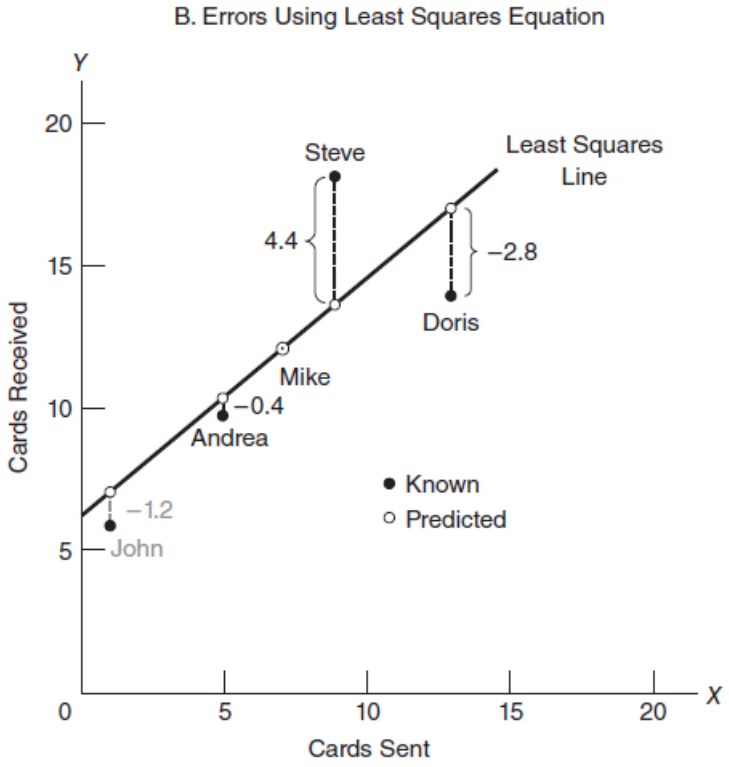


Figure 2.24– Predictive Error using Least Square Equation

- The figure 3.12 shows the corresponding predictive errors for all five friends when a series of different Y′ values, obtained from the least squares equation is used to predict each of their five Y scores.
- For example, Figure 3.11 shows the error for John when the mean for all five friends, Y, of 12 is used to predict his Y score of 6. Shown as a broken vertical line, the error of −6 for John (from Y – Y' = 6 − 12 = −6) indicates that Y overestimates John's Y score by 6 cards.
- Figure 2.24 shows a smaller error of −1.20 for John when a Y′ value of 7.20 is used to predict the same Y score of 6.
- This Y' value of 7.20 is obtained from the least squares equation,

$$\begin{aligned} Y' &= .80(X) + 6.40 \\ &= .80(1) + 6.40 \\ &= 7.20 \end{aligned}$$

- Positive and negative errors indicate that Y scores are either above or below their corresponding predicted scores.
- Overall, errors are smaller when a customized prediction of Y′ from the least squares equation than the repetitive prediction of Y.

Error Variability (Sum of Squares)

- The sum of squares of any set of deviations, called errors, can be calculated by first squaring each error, then summing all squared errors.
- The error variability for the repetitive prediction of the mean can be designated as SS_y , since each Y score is expressed as a squared deviation from Y and then summed, that is

$$SS_y = \sum(Y - \bar{Y})^2$$

- Using the errors for the five friends shown in Figure 3.11, this becomes

$$SS_y = [(-6)^2 + (-2)^2 + 0^2 + 6^2 + 2^2] = 80$$

- The error variability for the customized predictions from the least squares equation can be designated as $SS_{y|x}$, since each Y score is expressed as a squared deviation from its corresponding Y' and then summed, that is

$$SS_{y|x} = \sum(Y - Y')^2$$

Using the errors for the five friends shown in Figure 3.12, we obtain:

$$SS_{y|x} = [(-1.2)^2 + (-0.4)^2 + 0^2 + (4.4)^2 + (-2.8)^2] = 28.8$$

Proportion of Predicted Variability

- To obtain an SS measure of the actual gain in accuracy due to the least squares predictions, subtract the residual variability from the total variability, that is, subtract $SS_{y|x}$ from SS_y , to obtain

$$SS_y - SS_{y|x} = 80 - 28.8 = 51.2$$

- To express this difference, 51.2, as a gain in accuracy relative to the original error variability for the repetitive prediction of Y, divide the above difference by SS_y , that is,

$$\frac{SS_y - SS_{y|x}}{SS_y} = \frac{80 - 28.8}{80} = \frac{51.2}{80} = .64$$

- This result, .64 or 64 percent, represents the proportion or percent gain in predictive accuracy when the repetitive prediction of Y is replaced by a series of customized Y′ predictions based on the least squares equation.
- In other words, .64 or 64 percent represents the proportion or percent of the total variability of SSy that is predictable from its relationship with the X variable.
- The square of the correlation coefficient, r^2, always indicates the proportion of total variability in one variable that is predictable from its relationship with the other variable.

- Expressing the equation for r^2 in symbols,

r^2 INTERPRETATION

$$r^2 = \frac{SS_{Y'}}{SS_Y} = \frac{SS_Y - SS_{Y|X}}{SS_Y}$$

- where the sum of squares term, $SS_{y'}$, is simply the variability explained by or predictable from the regression equation, that is,

$$SS_{y'} = \sum(Y' - \overline{Y})^2$$

- Accordingly, r^2 provides us with a straightforward measure of the worth of our least squares predictive effort.

Problem 2.38:

- ***Assume that an r of .30 describes the relationship between educational level and estimated hours spent reading each week.***
  - ***a. According to r^2, what percent of the variability in weekly reading time is predictable from its relationship with educational level?***
  - ***b. What percent of variability in weekly reading time is not predictable from this relationship?***
  - ***c. Someone claims that 9 percent of each person's estimated reading time is predictable from the relationship. What is wrong with this claim?***

Solution:

(a) 9 percent predicted.
(b) 91 percent not predicted.
(c) 9 percent refers to the variability of all estimated reading times.

Problem 2.39:

As the correlation between the IQ scores of parents and children is .50, and that between the IQ scores of foster parents and foster children is .27.

- ***a. Does this signify, therefore, that the relationship between foster parents and foster children is about one-half as strong as the relationship between parents and children?***
- ***b. Use r^2 to compare the strengths of these two correlations.***

Solution:

(a) No

(b) The r^2 of .25 for parents and children is about four times greater than the r^2 of .07 for foster parents and foster children.

18. Explain in detail about Multiple regression equations.

- A least squares equation that contains more than one predictor or X variable
- For instance, a serious effort to predict college GPA might culminate in the following equation:

$$Y' = .410(X_1) + .005(X_2) + .001(X_3) + 1.03$$

where Y′ represents predicted college GPA

X1, X2, and X3 refer to high school GPA, IQ score, and SAT score, respectively.

- By capitalizing on the combined predictive power of several predictor variables, these multiple regression equations supply more accurate predictions for Y′ than could be obtained from a simple regression equation.

19. Explain in detail about Regression toward the mean.

- Regression toward the mean refers to a tendency for scores, particularly extreme scores, to shrink toward the mean.
- This tendency often appears among subsets of observations whose values are extreme and at least partly due to chance.
- Regression toward the mean appears among subsets of extreme observations for a wide variety of distributions.

Table 2.13 – Regression towards mean

REGRESSION TOWARD THE MEAN: BATTING AVERAGES OF TOP 10 HITTERS IN MAJOR LEAGUE BASEBALL DURING 2014 AND HOW THEY FARED DURING 2015

	BATTING AVERAGES*		
TOP 10 HITTERS (2014)	2014	2015	REGRESS TOWARD MEAN?
1. J. Alture	.341	.313	Yes
2. V. Martinez	.335	.282	Yes
3. M. Brantley	.327	.310	Yes
4. A. Beltre	.324	.287	Yes
5. J. Abreu	.317	.290	Yes
6. R. Cano	.314	.287	Yes
7. A. McCutchen	.314	.292	Yes
8. M. Cabrera	.313	.338	No
9. B. Posey	.311	.318	No
10. B. Revere	.306	.306	No

- Table 2.13 lists the top 10 hitters in the major leagues during 2014 and shows how they fared during 2015.
- Notice that 7 of the top 10 batting averages regressed downward, toward 260s, the approximate mean for all hitters during 2015.
- Hitters among the top 10 in 2014, who were not among the top 10 in 2015, were replaced by other mostly above-average hitters, who also were very lucky during 2015.
- Observed regression toward the mean occurs for individuals or subsets of individuals, not for entire groups.

The Regression Fallacy

- The regression fallacy is committed whenever regression toward the mean is interpreted as a real, rather than a chance, effect.
- A classic example of the regression fallacy occurred in an Israeli Air Force study of pilot training Some trainees were praised after very good landings, while others were reprimanded after very bad landings.
- On their next landings, praised trainees did more poorly and reprimanded trainees did better.
- It was concluded, therefore, that praise hinders but a reprimand helps performance.
- It's reasonable to assume that, in addition to skill, chance plays a role in landings.

Avoiding the Regression Fallacy

- The regression fallacy can be avoided by splitting the subset of extreme observations into two groups.
- One group of trainees would continue to be praised after very good landings and reprimanded after very poor landings.
- A second group of trainees would receive no feedback whatsoever after very good and very bad landings.
- In effect, the second group would serve as a control for regression toward the mean, since any shift toward the mean on their second landings would be due to chance.

Key Equations

PREDICTION EQUATION

$$Y' = bx + a$$

$$\text{where } b = r\sqrt{\frac{SS_Y}{SS_X}}$$

$$\text{and } a = \overline{Y} - b\overline{X}$$

CORRELATION COEFFICIENT

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

$$\text{where } SP_{xy} = \sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right) = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}$$

BOOK EXERCISES

Progress Check *6.1 Indicate whether the following statements suggest a positive or negative relationship:

(a) More densely populated areas have higher crime rates.

(b) Schoolchildren who often watch TV perform more poorly on academic achievement tests.

(c) Heavier automobiles yield poorer gas mileage.

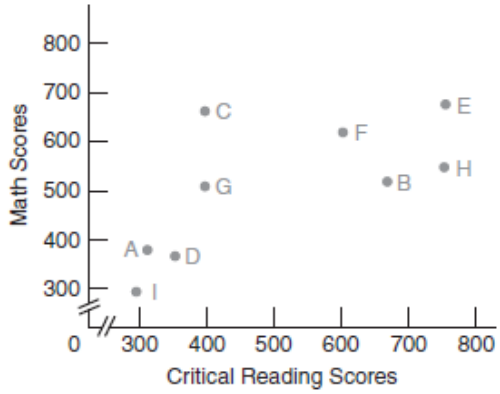
(d) Better-educated people have higher incomes.

(e) More anxious people voluntarily spend more time performing a simple repetitive task.

Solution

6.1 **(a)** Positive. The crime rate is higher, square mile by square mile, in densely populated cities than in sparsely populated rural areas.
(b) Negative. As TV viewing increases, performance on academic achievement tests tends to decline.
(c) Negative. Increases in car weight are accompanied by decreases in miles per gallon.
(d) Positive. Increases in educational level—grade school, high school, college—tend to be associated with increases in income.
(e) Positive. Highly anxious people willingly spend more time performing a simple repetitive task than do less anxious people.

Progress Check *6.2 Critical reading and math scores on the SAT test for students A, B, C, D, E, F, G, and H are shown in the following scatterplot:



(a) Which student(s) scored about the same on both tests?

(b) Which student(s) scored higher on the critical reading test than on the math test?

(c) Which student(s) will be eligible for an honors program that requires minimum scores of 700 in critical reading and 500 in math?

(d) Is there a negative relationship between the critical reading and math scores?

Solution

6.2 **(a)** I, D, F **(c)** E, H
(b) B, H, E **(d)** No. The relationship is positive.

Progress Check *6.3 Supply a verbal description for each of the following correlations. (If necessary, visualize a rough scatterplot for *r*, using the scatterplots in Figure 6.3 as a frame of reference.)

(a) an *r* of –.84 between total mileage and automobile resale value

(b) an *r* of –.35 between the number of days absent from school and performance on a math achievement test

(c) an *r* of .03 between anxiety level and college GPA

(d) an *r* of .56 between age of schoolchildren and reading comprehension

Solution

6.3 **(a)** Cars with more total miles tend to have lower resale values.
(b) Students with more absences from school tend to score lower on math achievement tests.
(c) Little or no relationship between anxiety level and college GPA.
(d) Older schoolchildren tend to have better reading comprehension.

Progress Check *6.4 Speculate on whether the following correlations reflect simple cause-effect relationships or more complex states of affairs. (**Hint:** A cause-effect relationship implies that, if all else remains the same, any change in the causal variable should always produce a predictable change in the other variable.)

(a) caloric intake and body weight

(b) height and weight

(c) SAT math score and score on a calculus test

(d) poverty and crime

Solution

6.4 **(a)** simple cause-effect **(b)** complex **(c)** complex **(d)** complex

Progress Check *6.5 Couples who attend a clinic for first pregnancies are asked to estimate (independently of each other) the ideal number of children. Given that *X* and *Y* represent the estimates of females and males, respectively, the results are as follows:

COUPLE	*X*	*Y*
A	1	2
B	3	4
C	2	3
D	3	2
E	1	0
F	2	3

Calculate a value for *r*, using the computation formula (6.1).

Solution

$$\textbf{6.5} \quad r = \frac{4}{\sqrt{(4)(9.33)}} = .65$$

*6.10 On the basis of an extensive survey, the California Department of Education reported an r of –.32 for the relationship between the amount of time spent watching TV and the achievement test scores of schoolchildren. Each of the following statements represents a possible interpretation of this finding. Indicate whether each is True or False.

(a) *Every* child who watches a lot of TV will perform poorly on the achievement tests.

(b) Extensive TV viewing causes a decline in test scores.

(c) Children who watch little TV will tend to perform well on the tests.

(d) Children who perform well on the tests will tend to watch little TV.

(e) If Gretchen's TV-viewing time is reduced by one-half, we can expect a substantial improvement in her test scores.

(f) TV viewing could not possibly cause a decline in test scores.

Answers on page 428.

Solution

6.10 **(a)** False. This statement would be true only if a perfect negative relationship (–1.00) described the relationship between TV viewing time and test scores.
(b) False. Correlation does not necessarily signify cause-effect.
(c) True
(d) True
(e) False. See **(b)**.
(f) False. Although correlation does not necessarily signify cause-effect, it opens the *possibility* of cause-effect.

Progress Check *7.1 To check your understanding of the first part of this chapter, make predictions using the following graph.

(a) Predict the approximate rate of inflation, given an unemployment rate of 5 percent.

(b) Predict the approximate rate of inflation, given an unemployment rate of 15 percent.

Solution

7.1 **(a)** approximately 5–6 percent
(b) approximately 2–3 percent

Progress Check *7.2 Assume that an r of .30 describes the relationship between educational level (highest grade completed) and estimated number of hours spent reading each week. More specifically:

EDUCATIONAL LEVEL (X)	WEEKLY READING TIME (Y)
$\bar{X} = 13$	$\bar{Y} = 8$
$SS_x = 25$	$SS_y = 50$
$r = .30$	

(a) Determine the least squares equation for predicting weekly reading time from educational level.

(b) Faith's education level is 15. What is her predicted reading time?

(c) Keegan's educational level is 11. What is his predicted reading time?

Solution

7.2 **(a)** $b = \sqrt{\frac{50}{25}}(.30) = .42$; $a = 8 - (.42)(13) = 2.54$

(b) $Y' = (.42)(15) + 2.54 = 8.84$

(c) $Y' = (.42)(11) + 2.54 = 7.16$

Progress Check *7.3

(a) Calculate the standard error of estimate for the data in Question 7.2 on page 132, assuming that the correlation of .30 is based on $n = 35$ pairs of observations.

(b) Supply a rough interpretation of the standard error of estimate.

Solution

7.3 **(a)** $s_{Y|X} = \sqrt{\frac{50\left(1-[.30]^2\right)}{35-2}} = \sqrt{\frac{50(.91)}{33}} = \sqrt{1.38} = 1.17$

(b) Roughly indicates the average amount by which the prediction is in error.

Progress Check *7.4 Assume that an r of .30 describes the relationship between educational level and estimated hours spent reading each week.

(a) According to r^2, what percent of the variability in weekly reading time is predictable from its relationship with educational level?

(b) What percent of variability in weekly reading time is not predictable from this relationship?

(c) Someone claims that 9 percent of *each* person's estimated reading time is predictable from the relationship. What is wrong with this claim?

Solution

7.4 **(a)** 9 percent predicted.
(b) 91 percent not predicted.
(c) 9 percent refers to the variability of *all* estimated reading times.

Progress Check *7.5 As indicated in Figure 6.3 on page 111, the correlation between the IQ scores of parents and children is .50, and that between the IQ scores of foster parents and foster children is .27.

(a) Does this signify, therefore, that the relationship between foster parents and foster children is about one-half as strong as the relationship between parents and children?

(b) Use r^2 to compare the strengths of these two correlations.

Solution

7.5 **(a)** No
(b) The r^2 of .25 for parents and children is about four times greater than the r^2 of .07 for foster parents and foster children.

Progress Check *7.6 After a group of college students attended a stress-reduction clinic, declines were observed in the anxiety scores of those who, prior to attending the clinic, had scored high on a test for anxiety.

(a) Can this decline be attributed to the stress-reduction clinic? Explain your answer.

(b) What type of study, if any, would permit valid conclusions about the effect of the stress-reduction clinic?

Solution

7.6 **(a)** No, because the observed decline could be due to regression toward the mean, given that the students scored high on the anxiety test prior to attending the clinic.
(b) An experiment where students who score high on the anxiety test are randomly assigned either to attend the stress-reduction clinic or to be in a control group.

*7.11 In studies dating back over 100 years, it's well established that regression toward the mean occurs between the heights of fathers and the heights of their *adult* sons. Indicate whether the following statements are true or false.

(a) Sons of tall fathers will tend to be shorter than their fathers.

(b) Sons of short fathers will tend to be taller than the mean for all sons.

(c) Every son of a tall father will be shorter than his father.

(d) Taken as a group, adult sons are shorter than their fathers.

(e) Fathers of tall sons will tend to be taller than their sons.

(f) Fathers of short sons will tend to be taller than their sons but shorter than the mean for all fathers.

Solution

7.11 **(a)** True

(b) False. Sons of short fathers will tend to be taller than their fathers but still shorter than the mean for all sons.

(c) False. Regression toward the mean is only a tendency, so there will be exceptions.

(d) False. Taken as an entire group, adult sons will be as tall as their fathers. (In fact, a comparison of entire groups might reveal that sons tend to be slightly taller because of an improvement in nutrition across generations.)

(e) False. *Given the subset of tall sons,* their fathers will tend to be shorter because of regression toward the mean.

(f) True

UNIT III – INFERENTIAL STATISTICS

SYLLABUS:

Populations – samples – random sampling – Sampling distribution- standard error of the mean - Hypothesis testing – z-test – z-test procedure –decision rule – calculations – decisions – interpretations - one-tailed and two-tailed tests – Estimation – point estimate – confidence interval – level of confidence – effect of sample size.

PART A

1. Define Population and it types.

- **Population**
  - Any complete set of observations (or potential observations).

Types of Population

- **Real Populations**
  - A *real* population is one in which all potential observations are accessible at the time of sampling.
- **Hypothetical Populations**
  - A *hypothetical* population is one in which all potential observations are not accessible at the time of sampling.

2. Define Sample and Random Sampling.

- **Sample**
  - Any subset of observations from a population.
  - The sample size is small relative to the population size.
- **Random Sampling**
  - A selection process that guarantees all potential observations in the population have an equal chance of being selected.
  - Inferential statistics requires that samples be random.

3. Define the term probability.

- **Probability**
  - The proportion or fraction of times that a particular event is likely to occur.

4. What is meant by Mutually Exclusive Events? State the Addition Rule for Mutually Exclusive Events

Mutually Exclusive Events

- Events that cannot occur together.

Addition Rule

- Add together the separate probabilities of several mutually exclusive events to find the probability that any one of these events will occur.

ADDITION RULE FOR MUTUALLY EXCLUSIVE EVENTS

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$$

where Pr( ) refers to the probability of the event in parentheses and A and B are mutually exclusive events.

5. What is meant by Dependent and Independent Events? State the Multiplication Rule for Independent Events.

Dependent Events

- When the occurrence of one event affects the probability of the other event, these events are dependent.
- Although the heights of randomly selected pairs of men are independent, the heights of brothers are dependent.

Independent Events

- The occurrence of one event has no effect on the probability that the other event will occur.

Multiplication Rule

- Multiply together the separate probabilities of several independent events to find the probability that these events will occur together.

MULTIPLICATION RULE FOR INDEPENDENT EVENT

$$\Pr(A \text{ and } B) = [\Pr(A)][\Pr(B)]$$

where A and B are independent events.

6. Define Conditional Probability and Alternative Approach to Conditional Probabilities

Conditional Probability

- The probability of one event, given the occurrence of another event.

Alternative Approach to Conditional Probabilities

- Conditional probabilities can be easily misinterpreted.
- Convert probabilities to frequencies (which, for example, total 100); solve the problem with frequencies; and then convert the answer back to a probability

7. Define sampling distribution of the mean.

- The sampling distribution of the mean refers to the probability distribution of means for all possible random samples of a given size from some population.

8. Narrate the symbols used for the mean and standard deviation of three types of Distributions.

SYMBOLS FOR THE MEAN AND STANDARD DEVIATION OF THREE TYPES OF DISTRIBUTIONS

TYPE OF DISTRIBUTION	MEAN	STANDARD DEVIATION
Sample	$\bar{X}$	s
Population	μ	σ
Sampling distribution of the mean	$\mu_{\bar{X}}$	$\sigma_{\bar{X}}$ (standard error of the mean)

9. Define mean of all sample means.

➢ **MEAN OF ALL SAMPLE MEANS** $(\mu_{\bar{X}})$

- The mean of all sample means always equals the population mean.

MEAN OF THE SAMPLING DISTRIBUTION

$$\mu_{\bar{X}} = \mu$$

where $(\mu_{\bar{X}})$ represents the mean of the sampling distribution and μ represents the mean of the population.

10. Define Standard error of the mean.

➢ **STANDARD ERROR OF THE MEAN** $(\sigma_{\bar{X}})$

- The distribution of sample means also has a standard deviation, referred to as the standard error of the mean.
- The standard error of the mean equals the standard deviation of the population divided by the square root of the sample size.

STANDARD ERROR OF THE MEAN

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

11. Define Shape of the sampling distribution or state the central limit theorem.

➢ **SHAPE OF THE SAMPLING DISTRIBUTION**

Central Limit Theorem

- The central limit theorem states that, regardless of the shape of the population, the shape of the sampling distribution of the mean approximates a normal curve if the sample size is sufficiently large.

12. Define Hypothesis Testing and its types.

Hypothesis Testing

- Hypothesis testing is a statistical method used to determine if there is enough evidence in a sample data to draw conclusions about a population.
- It is used to estimate the relationship between 2 statistical variables.
- It involves formulating two competing hypotheses, the null hypothesis (H0) and the alternative hypothesis (H1), and then collecting data to assess the evidence.
- Hypothesis testing evaluates two mutually exclusive population statements to determine which statement is most supported by sample data.

13. Defining Null Hypothesis and Alternate Hypothesis

- **Null hypothesis (H0):**

  In statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured cases or no relationship among groups. In other words, it is a basic assumption or made based on the problem knowledge.

  Example:

  A company's mean production is 50 units/per day

  H0: = 50.

- **Alternative hypothesis (H1):**

  The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis.

  Example:

  A company's production is not equal to 50 units/per day i.e.

  H1: 50.

14. Explain testing of Null Hypothesis. Define Common Outcome and Rare Outcome.

Testing Null Hypothesis

- The null hypothesis is tested by determining whether the one observed sample mean qualifies as a common outcome or a rare outcome in the hypothesized sampling distribution

Common Outcomes

- An observed sample mean qualifies as a common outcome if the difference between its value and that of the hypothesized population mean is small enough to be viewed as a probable outcome under the null hypothesis.
- There is no compelling reason for rejecting the null hypothesis, it is retained.

Rare Outcomes

- An observed sample mean qualifies as a rare outcome if the difference between its value and the hypothesized population mean is too large to be reasonably viewed as a probable outcome under the null hypothesis.

15. Discuss z test for a population mean.

Z TEST FOR A POPULATION MEAN

- A hypothesis test that evaluates how far the observed sample mean deviates, in standard error units, from the hypothesized population mean.
- This z test is accurate only when
  (1) the population is normally distributed or the sample size is large enough to satisfy the requirements of the central limit theorem
  (2) the population standard deviation is known.

16. List the z - test step by step procedure

Step 1 - State the research problem.
Step 2 - Identify the statistical hypotheses.
Step 3 - Specify a decision rule.
Step 4 - Calculate the value of the observed *z*.
Step 5 - Make a decision.
Step 6 - Interpret the decision.

17. Define Critical z Score

- A z score that separates common from rare outcomes and hence dictates whether H0 should be retained or rejected.

z RATIO FOR A SINGLE POPULATION MEAN

$$z = \frac{\bar{X} - \mu_{hyp}}{\sigma_{\bar{x}}}$$

18. Define Level of Significance (α)

- The degree of rarity required of an observed outcome in order to reject the null hypothesis (H0).

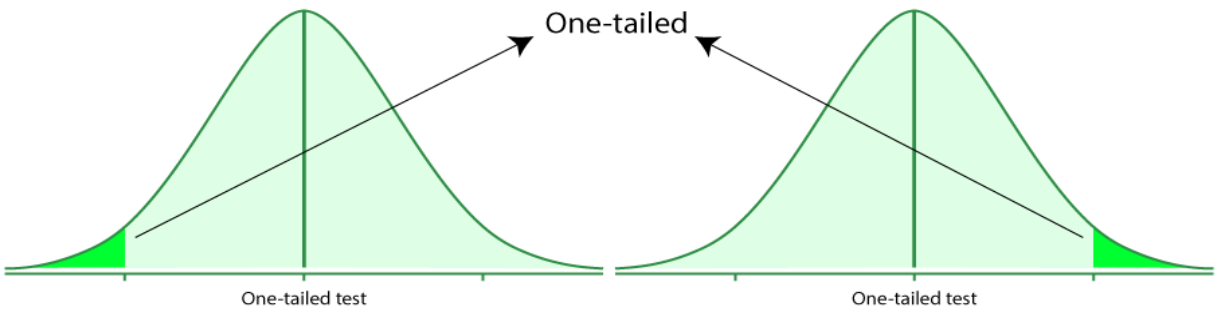
TYPE OF TEST	LEVEL OF SIGNIFICANCE (α) .05	.01
Two-tailed or nondirectional test (H_0: μ = some number) (H_1: $\mu \neq$ some number)	±1.96	±2.58
One-tailed or directional test, lower tail critical (H_0: $\mu \geq$ some number) (H_1: $\mu <$ some number)	–1.65	–2.33
One-tailed or directional test, upper tail critical (H_0: $\mu \leq$ some number) (H_1: $\mu >$ some number)	+1.65	+2.33

19. What is the use of one-tailed and two – tailed tests in hypothesis testing? When to use it?

- One and Two-Tailed Tests are ways to identify the relationship between the statistical variables.
- For checking the relationship between variables in a single direction (Left or Right direction), use a one-tailed test.
- A two-tailed test is used to check whether the relations between variables are in any direction or not.

20. Define One-Tailed or Directional Test

- A one-tailed test is based on a uni-directional hypothesis where the area of rejection is on only one side of the sampling distribution.
- It determines whether a particular population parameter is larger or smaller than the predefined parameter. It uses one single critical value to test the data.



21. Define Two-Tailed or Non-directional Test

- Rejection regions are located in both tails of the sampling distribution.
- For checking whether the sample is greater or less than a range of values, use the two-tailed testing.
- It is used for null hypothesis testing.

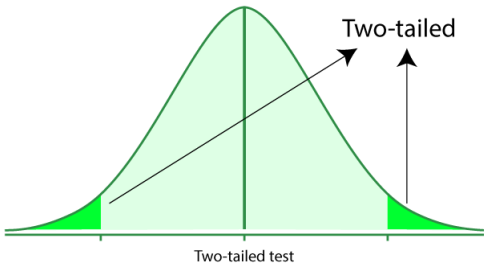


Figure 3.9 – Two tailed test

22. Define Point Estimate.

POINT ESTIMATE

- A single value that represents some unknown population characteristic, such as the population mean.
- The best single point estimate for the unknown population mean is simply the observed value of the sample mean.

23. Define Confidence interval

CONFIDENCE INTERVAL (CI) FOR μ

- A confidence interval for μ uses a range of values that, with a known degree of certainty, includes an unknown population characteristic, such as a population mean.

Confidence Interval for μ Based on z

CONFIDENCE INTERVAL FOR μ (BASED ON z)

$$\bar{X} \pm (z_{conf})(\sigma_{\bar{X}})$$

where

$\bar{X}$ represents the sample mean;

z_{conf} represents a number from the standard normal table that satisfies the confidence specifications for the confidence interval; and

$\sigma_{\bar{X}}$ represents the standard error of the mean.

24. Define level of confidence

- The **level of confidence** indicates the percent of time that a series of confidence intervals includes the unknown population characteristic, such as the population mean.
- Any level of confidence may be assigned to a confidence interval merely by substituting an appropriate value for *zconf* in Formula
- Although many different levels of confidence have been used, 95 percent and 99 percent are the most prevalent.

25. Which is efficient hypothesis tests or confidence intervals?

- Hypothesis tests merely indicate whether or not an effect is present, whereas Confidence intervals indicate the possible size of the effect.
- Confidence intervals tend to be more informative than hypothesis tests.

PART B

1. Give a detailed introduction about Population Sample and Probability.

- **Population**
  - Any complete set of observations (or potential observations).

Types of Population

- **Real Populations**
  - A *real* population is one in which all potential observations are accessible at the time of sampling.
- **Hypothetical Populations**
  - A *hypothetical* population is one in which all potential observations are not accessible at the time of sampling.

- **Sample**
  - Any subset of observations from a population.
  - The sample size is small relative to the population size.

Example 3.1

For each of the following pairs, indicate with a Yes or No whether the relationship between the first and second expressions could describe that between a sample and its population, respectively.

(a) students in the last row; students in class

(b) citizens of Wyoming; citizens of New York

(c) 20 lab rats in an experiment; all lab rats, similar to those used, that could undergo the same experiment

(d) all U.S. presidents; all registered Republicans

(e) two tosses of a coin; all possible tosses of a coin

Solution

(a) Yes

(b) No. Citizens of Wyoming aren't a subset of citizens of New York.

(c) Yes

(d) No. All U.S. presidents aren't a subset of all registered Republicans.

(e) Yes

Example 3.2

Identify all of the expressions from Example 3.1 that involve a hypothetical population.

Solution

Expressions in 8.1(c) and 8.1(e) involve hypothetical populations.

- **Random Sampling**
  - A selection process that guarantees all potential observations in the population have an equal chance of being selected.
  - Inferential statistics requires that samples be random.

Example 3.3

Indicate whether each of the following statements is True or False.

A random selection of 10 playing cards from a deck of 52 cards implies that

(a) the random sample of 10 cards accurately represents the important features of the whole deck.

(b) each card in the deck has an equal chance of being selected.

(c) it is impossible to get 10 cards from the same suit (for example, 10 hearts).

(d) any outcome, however unlikely, is possible.

Solution

a. False. Sometimes, just by chance, a random sample of 10 cards fails to represent the important features of the whole deck.

b. True

c. False. Although unlikely, 10 hearts could appear in a random sample of 10 cards.

d. True

- **Tables Of Random Numbers**
  - Tables of random numbers can be used to obtain a random sample.
  - These tables are generated by a computer designed to equalize the occurrence of any one of the 10 digits: 0, 1, 2, . . . , 8, 9.

Example 3.4

Describe how you would use the table of random numbers to take

a. a random sample of five statistics students in a classroom where each of nine rows consists of nine seats.

b. a random sample of size 40 from a large directory consisting of 3041 pages, with 480 lines per page.

Solution

a. There are many ways. For instance, consult the tables of random numbers, using the first digit of each 5-digit random number to identify the row (previously labelled 1, 2, 3, and so on), and the second digit of the same random number to locate a particular student's seat within that row. Repeat this process until five students have been identified. (If the classroom is larger, use additional digits so that every student can be sampled.)

b. *Once again, there are many ways. For instance, use the initial 4 digits of each random number (between 0001 and 3041) to identify the page number of the telephone directory and the next 3 digits (between 001 and 480) to identify the particular line on that page. Repeat this process, using 7-digit numbers, until 40 telephone numbers have been identified.*

➢ **Probability**

- The proportion or fraction of times that a particular event is likely to occur.

Mutually Exclusive Events

- Events that cannot occur together.

Addition Rule

- Add together the separate probabilities of several mutually exclusive events to find the probability that any one of these events will occur.

ADDITION RULE FOR MUTUALLY EXCLUSIVE EVENTS

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$$

where Pr( ) refers to the probability of the event in parentheses and A and B are mutually exclusive events.

Example 3.5

Assuming that people are equally likely to be born during any One of the months, what is the probability of Jack being born during

(a) June?

(b) any month other than June?

(c) either May or June?

Solution

(a) $\frac{1}{12}$ **(b)** $\frac{11}{12}$ **(c)** $\frac{2}{12}$

Independent Events

- The occurrence of one event has no effect on the probability that the other event will occur.

Multiplication Rule

- Multiply together the separate probabilities of several independent events to find the probability that these events will occur together.

MULTIPLICATION RULE FOR INDEPENDENT EVENT

$$\Pr(A \text{ and } B) = [\Pr(A)][\Pr(B)]$$

where A and B are independent events.

Example 3.6

Assuming that people are equally likely to be born during any of the months, and also assuming (possibly over the objections of astrology fans) that the birthdays of married couples are independent, what's the probability of

(a) the husband being born during January and the wife being born during February?

(b) both husband and wife being born during December?

(c) both husband and wife being born during the spring (April or May)? (Hint: First, find the probability of just one person being born during April or May.)

Solution

(a) $\left(\frac{1}{12}\right)\left(\frac{1}{12}\right) = \left(\frac{1}{144}\right)$

(b) $\left(\frac{1}{12}\right)\left(\frac{1}{12}\right) = \left(\frac{1}{144}\right)$

(c) $\left(\frac{2}{12}\right)\left(\frac{2}{12}\right) = \left(\frac{4}{144}\right)$

Dependent Events

- When the occurrence of one event affects the probability of the other event, these events are dependent.
- Although the heights of randomly selected pairs of men are independent, the heights of brothers are dependent.

Conditional Probability

- The probability of one event, given the occurrence of another event.

Alternative Approach to Conditional Probabilities

- Conditional probabilities can be easily misinterpreted.
- Convert probabilities to frequencies (which, for example, total 100); solve the problem with frequencies; and then convert the answer back to a probability

Example –

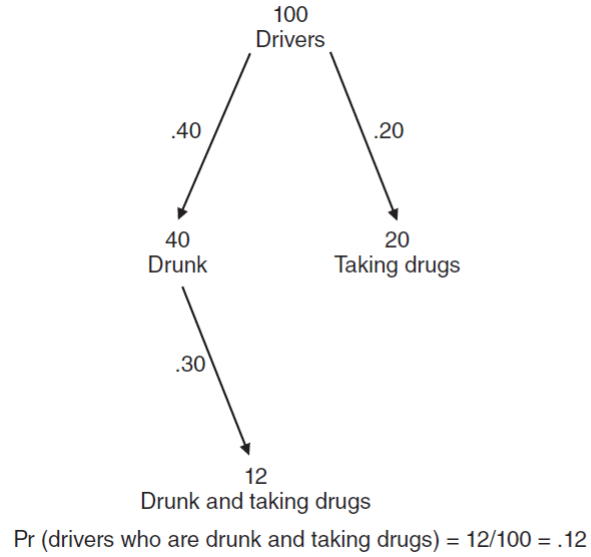


Figure 3.1 – A frequency analysis of 100 drivers who caused fatal accidents

Figure 3.1 shows a frequency analysis for the 100 drivers involved in fatal accidents.

Working from the top down, notice that among the 100 drivers, 40 are drunk (from .40 × 100 = 40) and 20 take drugs (from .20 × 100 = 20). Also notice that 12 of the 40 drunk drivers also take drugs (from .30 × 40 = 12). Now, it is fairly straightforward to establish that the probability of drivers both being drunk *and* taking drugs. It is simply the number of drivers who are drunk and take drugs, 12, divided by the total number of drivers, 100, that is, 12/100 =.12, which, of course, is the same as the previous answer.

Once a frequency analysis has been done, it often is easy to answer other questions.

For example,

"What is the conditional probability of being drunk, given that the driver takes illegal drugs?"

Referring to Figure 3.1, divide the number of drivers who are drunk and take drugs, 12, by the number of drivers who take drugs, 20, that is, 12/20 = .60.

Example 3.7

Among 100 couples who had undergone marital counselling, 60 couples described their relationships as improved, and among this latter group, 45 couples had children. The remaining couples described their relationships as unimproved, and among this group, 5 couples had children. (Hint: Using a frequency analysis, begin with the 100 couples, first branch into the number of couples with improved and unimproved relationships, then under each of these numbers,

branch into the number of couples with children and without children. Enter a number at each point of the diagram before proceeding.)

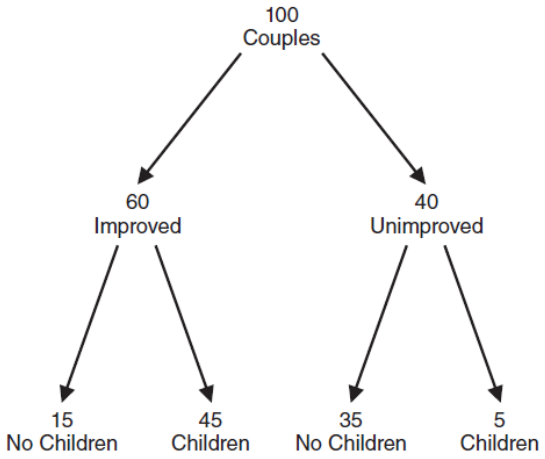
a. What is the probability of randomly selecting a couple who described their relationship as improved?

b. What is the probability of randomly selecting a couple with children?

c. What is the conditional probability of randomly selecting a couple with children, given that their relationship was described as improved?

d. What is the conditional probability of randomly selecting a couple without children, given that their relationship was described as not improved?

e. What is the conditional probability of an improved relationship, given that a couple has children?



(a) $\frac{60}{100} = .60$

(b) $\frac{45+5}{100} = \frac{50}{100} = .50$

(c) $\frac{45}{60} = .75$

(d) $\frac{35}{40} = .875$

(e) $\frac{45}{45+5} = \frac{45}{50} = .90$

2. Discuss in detail about sampling distribution and creating sampling distribution in inferential statistics.

- Sampling distribution of the mean
- Creating a sampling distribution
- Mean of all sample means $(\mu_{\bar{X}})$
- Standard error of the mean $(\sigma_{\bar{X}})$
- Shape of the sampling distribution

➢ SAMPLING DISTRIBUTION OF THE MEAN

- The sampling distribution of the mean refers to the probability distribution of means for all possible random samples of a given size from some population.

SYMBOLS FOR THE MEAN AND STANDARD DEVIATION OF THREE TYPES OF DISTRIBUTIONS

TYPE OF DISTRIBUTION	MEAN	STANDARD DEVIATION
Sample	$\bar{X}$	s
Population	μ	σ
Sampling distribution of the mean	$\mu_{\bar{X}}$	$\sigma_{\bar{X}}$ (standard error of the mean)

➢ CREATING A SAMPLING DISTRIBUTION

- Imagine small population of four observations with values of 2, 3, 4, and 5, as shown in **Figure 3.2.**

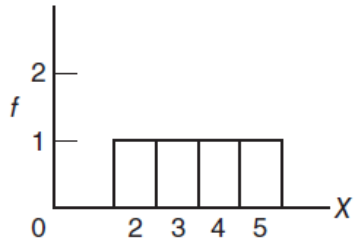


Figure 3.2 - Graph of a miniature population.

- Itemize all possible random samples, each of size two, that could be taken from this population.
- There are four possibilities on the first draw from the population and also four possibilities on the second draw from the population, as indicated in **Table 3.1.***

- The two sets of possibilities combine to yield a total of 16 possible samples.
- Table 3.1 also lists a sample mean (found by adding the two observations and dividing by 2) and its probability of occurrence (expressed as 1⁄16, since each of the 16 possible samples is equally likely).

Table 3.1 - All possible samples of size two from a miniature population

ALL POSSIBLE SAMPLES OF SIZE TWO FROM A MINIATURE POPULATION

	ALL POSSIBLE SAMPLES	MEAN ($\bar{X}$)	PROBABILITY
(1)	2,2	2.0	$\frac{1}{16}$
(2)	2,3	2.5	$\frac{1}{16}$
(3)	2,4	3.0	$\frac{1}{16}$
(4)	2,5	3.5	$\frac{1}{16}$
(5)	3,2	2.5	$\frac{1}{16}$
(6)	3,3	3.0	$\frac{1}{16}$
(7)	3,4	3.5	$\frac{1}{16}$
(8)	3,5	4.0	$\frac{1}{16}$
(9)	4,2	3.0	$\frac{1}{16}$
(10)	4,3	3.5	$\frac{1}{16}$
(11)	4,4	4.0	$\frac{1}{16}$
(12)	4,5	4.5	$\frac{1}{16}$
(13)	5,2	3.5	$\frac{1}{16}$
(14)	5,3	4.0	$\frac{1}{16}$
(15)	5,4	4.5	$\frac{1}{16}$
(16)	5,5	5.0	$\frac{1}{16}$

- When cast into a relative frequency or probability distribution, as in **Table 3.2,** the 16 sample means constitute the sampling distribution of the mean, previously defined as the probability distribution of means for all possible random samples of a given size from some population.
- Not all values of the sample mean occur with equal probabilities in Table 3.2 since some values occur more than once among the 16 possible samples.
- For instance, a sample mean value of 3.5 appears among 4 of 16 possibilities and has a probability of 4⁄16.

Table 3.2 – Sampling Distribution of the Mean (samples of size Of two from a miniature population)

SAMPLING DISTRIBUTION OF THE MEAN (SAMPLES OF SIZE TWO FROM A MINIATURE POPULATION)	
SAMPLE MEAN ($\bar{X}$)	PROBA-BILITY
5.0	$\frac{1}{16}$
4.5	$\frac{2}{16}$
4.0	$\frac{3}{16}$
3.5	$\frac{4}{16}$
3.0	$\frac{3}{16}$
2.5	$\frac{2}{16}$
2.0	$\frac{1}{16}$

- **Probability of a Particular Sample Mean**
  - The distribution in Table 3.2 can be consulted to determine the probability of obtaining a particular sample mean or set of sample means.
  - The probability of a randomly selected sample mean of either 5.0 or 2.0 equals 1⁄16 + 1⁄16 = 2⁄16 = .1250.
  - This type of probability statement, based on a sampling distribution, assumes an essential role in inferential statistics
  - Refer Figure 3.3

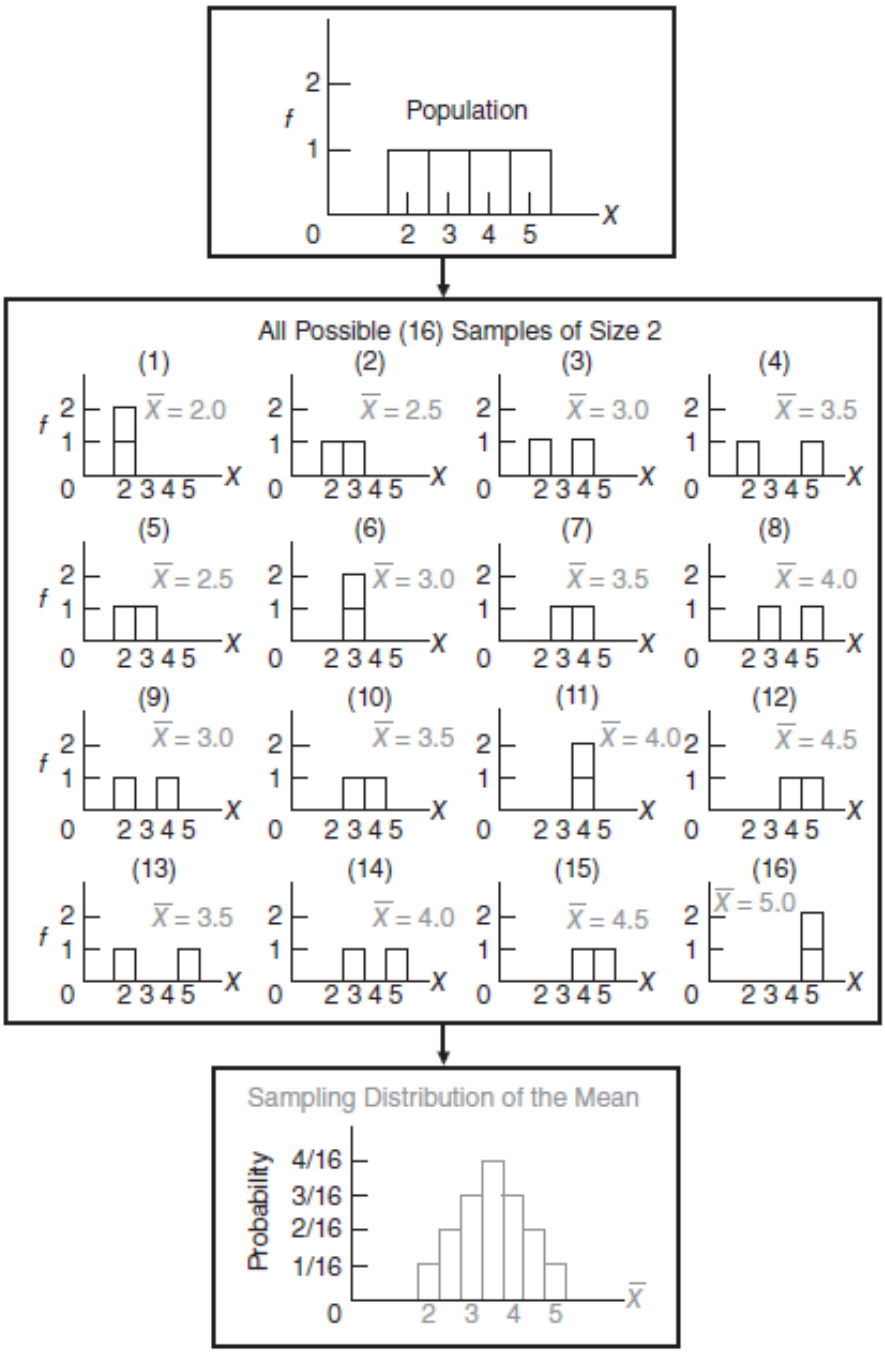


FIGURE 3.3

Emergence of the sampling distribution of the mean from all possible samples.

Example 3.8

Without peeking, list the special symbols for the mean of the population

(a) mean of the sampling distribution of the mean

(b) mean of the sample

(c) standard error of the mean

(d) standard deviation of the sample

(e) standard deviation of the population (f) .

(a) μ **(b)** $\mu_{\overline{X}}$ **(c)** $\overline{X}$ **(d)** $\sigma_{\overline{X}}$ **(e)** s **(f)** σ

Example 3.9

Imagine a very simple population consisting of only five observations: 2, 4, 6, 8, 10.

(a) List all possible samples of size two.

(1) 2,2	(6) 4,2	(11) 6,2	(16) 8,2	(21) 10,2
(2) 2,4	(7) 4,4	(12) 6,4	(17) 8,4	(22) 10,4
(3) 2,6	(8) 4,6	(13) 6,6	(18) 8,6	(23) 10,6
(4) 2,8	(9) 4,8	(14) 6,8	(19) 8,8	(24) 10,8
(5) 2,10	(10) 4,10	(15) 6,10	(20) 8,10	(25) 10,10

(b) Construct a relative frequency table showing the sampling distribution of the mean.

$\overline{X}$	PROBABILITY
10	1/25
9	2/25
8	3/25
7	4/25
6	5/25
5	4/25
4	3/25
3	2/25
2	1/25

➢ MEAN OF ALL SAMPLE MEANS $(\mu_{\bar{X}})$

- The mean of all sample means always equals the population mean.

MEAN OF THE SAMPLING DISTRIBUTION

$$\mu_{\bar{X}} = \mu$$

where $(\mu_{\bar{X}})$ represents the mean of the sampling distribution and μ represents the mean of the population.

Example 3.10

Indicate whether the following statements are True or False.

The mean of all sample means, $(\mu_{\bar{X}})$***, . . .***

(a) always equals the value of a particular sample mean.

(b) equals 100 if, in fact, the population mean equals 100.

(c) usually equals the value of a particular sample mean.

(d) is interchangeable with the population mean.

a. False. It always equals the value of the population mean.
b. True
c. False. Because of chance, most sample means tend to be either larger or smaller than the mean of all sample means.
d. True

➢ STANDARD ERROR OF THE MEAN $(\sigma_{\bar{X}})$

- The distribution of sample means also has a standard deviation, referred to as the standard error of the mean.
- The standard error of the mean serves as a special type of standard deviation that measures variability in the sampling distribution.
- A rough measure of the average amount by which sample means deviate from the mean of the sampling distribution or from the population mean.
- The standard error of the mean equals the standard deviation of the population divided by the square root of the sample size.

STANDARD ERROR OF THE MEAN

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Example 3.10

Indicate whether the following statements are True or False. The standard error of the mean, $(\sigma_{\bar{X}})$ ***, . . .***

(a) roughly measures the average amount by which sample means deviate from the population mean.

(b) measures variability in a particular sample.

(c) increases in value with larger sample sizes.

(d) equals 5, given that σ = 40 and n = 64.

(a) True

(b) False. It measures variability among sample means.

(c) False. It decreases in value with larger sample sizes.

(d) True

- **SHAPE OF THE SAMPLING DISTRIBUTION**

Central Limit Theorem

- the central limit theorem states that, regardless of the shape of the population, the shape of the sampling distribution of the mean approximates a normal curve if the sample size is sufficiently large.
- **Example -** For the two non-normal populations in the top panel of **Figure 3.4**, the shapes of the sampling distributions in the middle panel show essentially the same preliminary drift toward normality when the sample size equals only 2, while the shapes of the sampling distributions in the bottom panel closely approximate normality when the sample size equals 25.

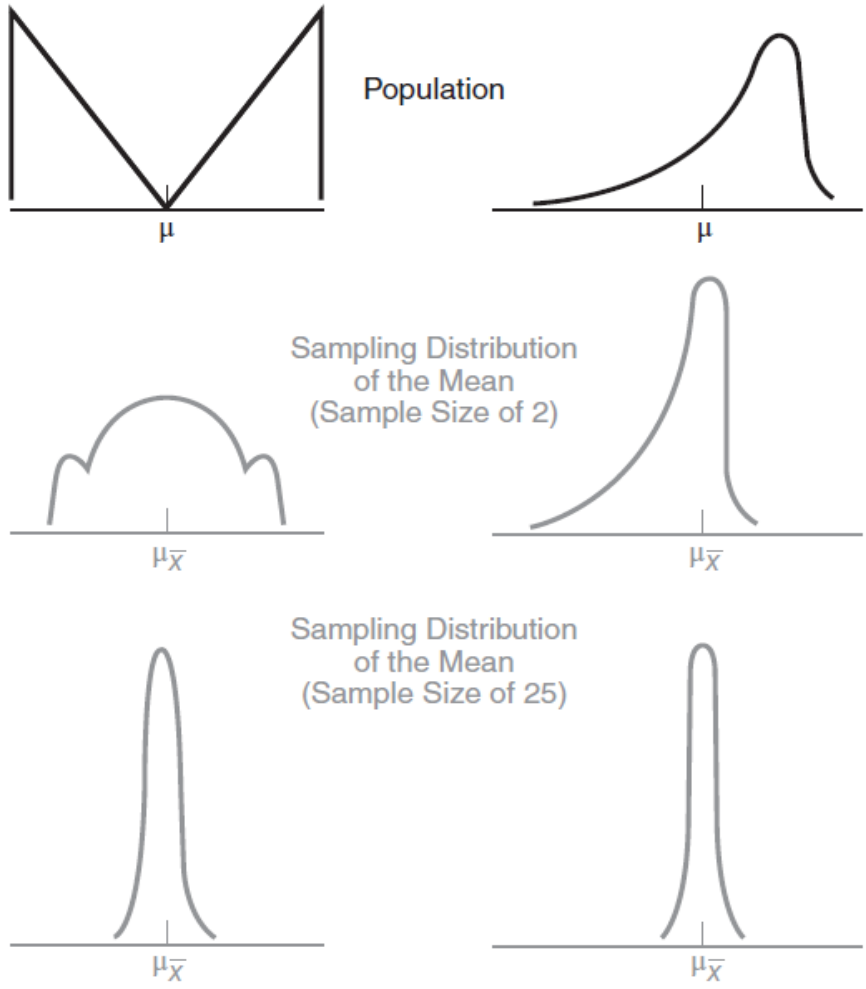


Figure 3.4 – Effect of Central limit theorem

Example 3.11

Indicate whether the following statements are True or False. The central limit theorem

a. states that, with sufficiently large sample sizes, the shape of the population is normal.

b. states that, regardless of sample size, the shape of the sampling distribution of the mean is normal.

c. ensures that the shape of the sampling distribution of the mean equals the shape of the population.

d. applies to the shape of the sampling distribution—not to the shape of the population and not to the shape of the sample.

a. *False. The shape of the population remains the same regardless of sample size.*
b. *False. It requires that the sample size be sufficiently large—usually between 25 and 100.*
c. *False. It ensures that the shape of the sampling distribution approximates a normal curve, regardless of the shape of the population (which remains intact).*
d. *True*

3. Explain in detail about Hypothesis Testing and its types.

Hypothesis Testing

- Hypothesis testing is a statistical method used to determine if there is enough evidence in a sample data to draw conclusions about a population.
- It is used to estimate the relationship between 2 statistical variables.
- It involves formulating two competing hypotheses, the null hypothesis (H0) and the alternative hypothesis (Ha), and then collecting data to assess the evidence.
- Hypothesis testing evaluates two mutually exclusive population statements to determine which statement is most supported by sample data.

Defining Hypotheses

- **Null hypothesis (H0):**
  In statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured cases or no relationship among groups. In other words, it is a basic assumption or made based on the problem knowledge.
  Example:
  A company's mean production is 50 units/per day
  H0: = 50.
- **Alternative hypothesis (H1):**
  The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis.
  Example:
  A company's production is not equal to 50 units/per day i.e.
  H1: 50.

Key Terms of Hypothesis Testing

- **Level of significance**:
  - It refers to the degree of significance to accept or reject the null hypothesis. 100% accuracy is not possible for accepting a hypothesis, so, therefore, select a level of significance that is usually 5%.
  - This is normally denoted with and generally, it is 0.05 or 5%, which means the output should be 95% confident to give a similar kind of result in each sample.
- **P-value:**
  - The P value, or calculated probability, is the probability of finding the observed/extreme results when the null hypothesis(H0) of a study-given problem is true.
  - If P-value is less than the chosen significance level then reject the null hypothesis i.e. accept that the sample claims to support the alternative hypothesis.
- **Test Statistic:**
  - The test statistic is a numerical value calculated from sample data during a hypothesis test, used to determine whether to reject the null hypothesis.
  - It is compared to a critical value or p-value to make decisions about the statistical significance of the observed results.
- **Critical value**:
  - The critical value in statistics is a threshold or cutoff point used to determine whether to reject the null hypothesis in a hypothesis test.
- **Degrees of freedom:**
  - Degrees of freedom are associated with the variability or freedom one has in estimating a parameter.
  - The degrees of freedom are related to the sample size and determine the shape.

Testing Null Hypothesis

- The null hypothesis is tested by determining whether the one observed sample mean qualifies as a common outcome or a rare outcome in the hypothesized sampling distribution of Figure 3.5.

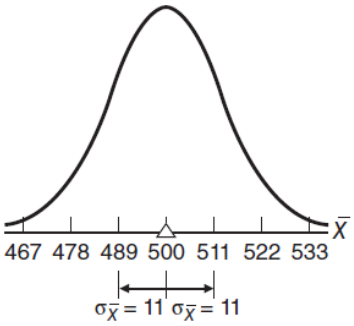


Figure 3.5. - Hypothesized sampling distribution of the mean centred about a hypothesized population mean of 500.

- **Common Outcomes**
  - An observed sample mean qualifies as a common outcome if the difference between its value and that of the hypothesized population mean is small enough to be viewed as a probable outcome under the null hypothesis.
  - There is no compelling reason for rejecting the null hypothesis, it is retained.
- **Rare Outcomes**
  - An observed sample mean qualifies as a rare outcome if the difference between its value and the hypothesized population mean is too large to be reasonably viewed as a probable outcome under the null hypothesis.

Boundaries for Common and Rare Outcomes

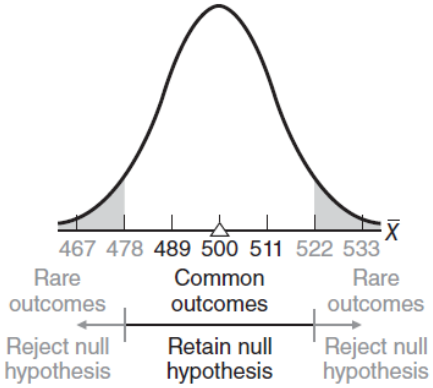


Figure 3.6 - One possible set of common and rare outcomes (values of X).

Figure 3.6 shows one possible set of boundaries for common and rare outcomes, expressed in values of *X*.

If the one observed sample mean is located between 478 and 522, it will qualify as a common outcome, and the null hypothesis will be retained.

If, however, the one observed sample mean is greater than522 or less than 478, it will qualify as a rare outcome, and the null hypothesis will be rejected.

4. Discuss in detail about z test for a population mean and z test procedure.

Converting a Raw Score to *z*

- To convert a raw score into a standard score, express the raw score as a distance from its mean (by subtracting the mean from the raw score), and then split this distance into standard deviation units (by dividing with the standard deviation).

$$Standard\ score = \frac{raw\ score - mean}{standard\ deviation}$$

Converting a Sample Mean to *z*

z RATIO FOR A SINGLE POPULATION MEAN

$$z = \frac{\overline{X} - \mu_{hyp}}{\sigma_{\overline{x}}}$$

where

$\overline{X}$ - observed sample mean;

μ_{hyp} - the hypothesized population mean

$\sigma_{\overline{x}}$ - the standard error of the mean

Z TEST FOR A POPULATION MEAN

- A hypothesis test that evaluates how far the observed sample mean deviates, in standard error units, from the hypothesized population mean.
- This z test is accurate only when
  (1) the population is normally distributed or the sample size is large enough to satisfy the requirements of the central limit theorem
  (2) the population standard deviation is known.

Example 3.12

Calculate the value of the z test for each of the following situations:

(a) $\overline{X} = 566; \sigma = 30; n = 36; \mu_{hyp} = 560$

(b) $\overline{X} = 24; \sigma = 4; n = 64; \mu_{hyp} = 25$

(c) $\overline{X} = 82; \sigma = 14; n = 49; \mu_{hyp} = 75$

(d) $\overline{X} = 136; \sigma = 15; n = 25; \mu_{hyp} = 146$

$$\textbf{(a)}\quad z=\frac{566-560}{30/\sqrt{36}}=\frac{6}{5}=1.20$$

$$\textbf{(b)}\quad z=\frac{24-25}{4/\sqrt{64}}=\frac{-1}{.5}=-2.00$$

$$\textbf{(c)}\quad z=\frac{82-75}{14/\sqrt{49}}=\frac{7}{2}=3.50$$

$$\textbf{(d)}\quad z=\frac{136-146}{15/\sqrt{25}}=\frac{-10}{3}=-3.33$$

Z - TEST STEP BY STEP PROCEDURE

Step 1 - State the research problem.

- State the problem to be resolved by the investigation.

Step 2 - Identify the statistical hypotheses.

- The statistical hypotheses consist of a null hypothesis (*H*0) and an alternative (or research) hypothesis (*H*1).

Null Hypothesis (H0)

- A statistical hypothesis that usually asserts that nothing special is happening with respect to some characteristic of the underlying population.

$$H_0 : \mu = 500$$

Where μ is the population mean

Alternative Hypothesis (H1)

- The opposite of the null hypothesis.

$$H_1 : \mu \neq 500$$

- Depending on the outcome of the hypothesis test, *H*0 will either be retained or rejected.

Step 3 - Specify a decision rule.

- This rule indicates precisely when *H*0 should be rejected.

Step 4 - Calculate the value of the observed *z*.

- Express the one observed sample mean as an observed *z*,

Critical z Score

- A z score that separates common from rare outcomes and hence dictates whether H0 should be retained or rejected.

z RATIO FOR A SINGLE POPULATION MEAN

$$z = \frac{\overline{X} - \mu_{hyp}}{\sigma_{\overline{x}}}$$

Level of Significance (α)

- The degree of rarity required of an observed outcome in order to reject the null hypothesis (H0).

TYPE OF TEST	LEVEL OF SIGNIFICANCE (α) .05	.01
Two-tailed or nondirectional test (H_0: μ = some number) (H_1: $\mu \neq$ some number)	±1.96	±2.58
One-tailed or directional test, lower tail critical (H_0: $\mu \geq$ some number) (H_1: $\mu <$ some number)	–1.65	–2.33
One-tailed or directional test, upper tail critical (H_0: $\mu \leq$ some number) (H_1: $\mu >$ some number)	+1.65	+2.33

Step 5 - Make a decision.

- Either retain or reject H_0 at the specified level of significance, justifying this decision by noting the relationship between observed and critical *z* scores.

Retaining H_0 is a Weak Decision

- H_0 is retained whenever the observed z qualifies as a common outcome on the assumption that H_0 is true.

Rejecting H_0 is a Strong Decision

- H_0 is rejected whenever the observed z qualifies as a rare outcome on the assumption that H_0 is true.

Step 6 - Interpret the decision.

- Using words, interpret the decision in terms of the original research problem.
- Rejection of the null hypothesis supports the research hypothesis, while retention of the null hypothesis fails to support the research hypothesis.

HYPOTHESIS TEST SUMMARY: *z* TEST FOR A POPULATION MEAN (SAT SCORES)

Research Problem

Does the mean SAT math score for all local freshmen differ from the national average of 500?

Statistical Hypotheses

$$H_0 : \mu = 500$$
$$H_1 : \mu \neq 500$$

Decision Rule

Reject H_0 at the .05 level of significance if $z \geq 1.96$ or if $z \leq -1.96$.

Calculations

Given

$$\bar{X} = 533;\ \mu_{hyp} = 500;\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{110}{\sqrt{100}} = 11$$

$$z = \frac{533 - 500}{11} = 3$$

Decision

Reject H_0 at the .05 level of significance because $z = 3$ exceeds 1.96.

Interpretation

The mean SAT math score for all local freshmen does not equal—it exceeds—the national average of 500.

Example 3.13

Indicate what's wrong with each of the following statistical hypotheses:

(a) $H_0: \mu = 155$
$H_1: \mu \neq 160$

(b) $H_0: \bar{X} = 241$
$H_1: \bar{X} \neq 241$

(a) Different numbers appear in *H*0 and *H*1.
(b) Sample means (rather than population means) appear in *H*0 and *H*1.

Example 3.14

First using words, then symbols, identify the null hypothesis for each of the following situations.

a. A school administrator wishes to determine whether sixth-grade boys in her school district differ, on average, from the national norms of 10.2 pushups for sixth-grade boys.

b. A consumer group investigates whether, on average, the true weights of packages of ground beef sold by a large supermarket chain differ from the specified 16 ounces.

c. A marriage counselor wishes to determine whether, during a standard conflict-resolution session, his clients differ, on average, from the 11 verbal interruptions reported for "welladjusted couples."

(a) Sixth-grade boys in her school district average 10.2 pushups.
H0: μ = 10.2

(b) On average, weights of packages of ground beef sold by a large supermarket chain equal 16 ounces.
H0: μ = 16

(c) The marriage counselor's clients average 11 interruptions per session.
H0: μ = 11

Example 3.15

For each of the following situations, indicate whether H_0 should be retained or rejected and justify your answer by specifying the precise relationship between observed and critical z scores. Should H_0 be retained or rejected, given a hypothesis test with critical z scores of ± 1.96 and

(a) $z = 1.74$ **(b)** $z = 0.13$ **(c)** $z = -2.51$

a. Retain H_0 at the .05 level of significance because z = 1.74 is less positive than 1.96.
b. Retain H_0 at the .05 level of significance because z = 0.13 is less positive than 1.96.
c. Reject H_0 at the .05 level of significance because z = −2.51 is more negative than –1.96.

Example 3.16

According to the American Psychological Association, members with a doctorate and a full-time teaching appointment earn, on thaverage, $82,500 per year, with a standard deviation of $6,000. An investigator wishes to determine whether $82,500 is also the mean salary for all female members with a doctorate and a full-time teaching appointment. Salaries are obtained for a random sample of 100 women from this population, and the mean salary equals $80,100.

(a) Someone claims that the observed difference between $80,100 and $82,500 is large enough by itself to support the conclusion that female members earn less than male members. Explain why it is important to conduct a hypothesis test.

(b) The investigator wishes to conduct a hypothesis test for what population?

(c) What is the null hypothesis, H0?

(d) What is the alternative hypothesis, H1?

(e) Specify the decision rule, using the .05 level of significance.

(f) Calculate the value of z. (Remember to convert the standard deviation to a standard error.)

(g) What is your decision about H0?

(h) Using words, interpret this decision in terms of the original problem.

(a) The observed difference between $80,100 and $82,500 cannot be interpreted at face value, as it could have happened just by chance. A hypothesis test permits us to evaluate the effect of chance by measuring the observed difference relative to the standard error of the mean.

(b) All female members of the APA with a Ph.D. degree and a full-time teaching appointment.

(c) $H_0: \mu = 82{,}500$

(d) $H_1: \mu \neq 82{,}500$

(e) Reject H_0 at the .05 level of significance if $z \geq 1.96$ *or* $z \leq -1.96$

(f) $$z = \frac{80{,}000 - 82{,}500}{\frac{6000}{\sqrt{100}}} = \frac{-2{,}400}{600} = -4.00$$

(g) Reject H_0 at the .05 level of significance because $z = -4.00$ is more negative than -1.96.

(h) The average salary of all female APA members (with a Ph.D. and a full-time teaching appointment) is less than $82,500.

5. Discuss and differentiate between one-tailed and two – tailed tests in hypothesis testing.

- One and Two-Tailed Tests are ways to identify the relationship between the statistical variables.
- For checking the relationship between variables in a single direction (Left or Right direction), use a one-tailed test.
- A two-tailed test is used to check whether the relations between variables are in any direction or not.

One-Tailed or Directional Test

- A one-tailed test is based on a uni-directional hypothesis where the area of rejection is on only one side of the sampling distribution.
- It determines whether a particular population parameter is larger or smaller than the predefined parameter. Refer Figure 3.7 and Figure 3.8
- It uses one single critical value to test the data.

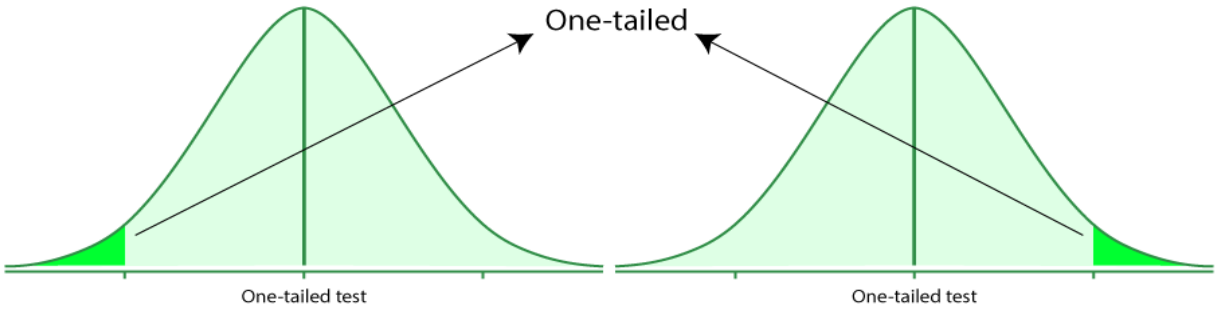


Figure 3.7 – One tailed test

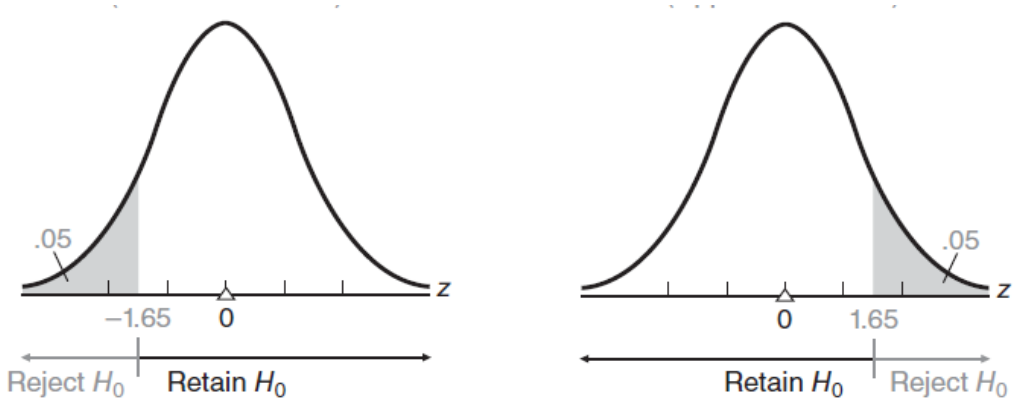


Figure 3.8 a. One-Tailed or Directional Test (Lower Tail Critical)
Figure 3.8 b. One-Tailed or Directional Test (Upper Tail Critical)

- Figure 3.8 a, illustrates a rejection region that is associated with only the lower tail of the hypothesized sampling distribution.
- The corresponding decision rule, with its critical z of –1.65, is referred to as a one-tailed or directional test with the lower tail critical.
- Figure 3.8 b, illustrates one-tailed or directional test with the upper tail critical. This one-tailed test is the mirror image of the previous test.
- The corresponding decision rule, with its critical z of 1.65, is referred to as a one-tailed or directional test with the upper tail critical.

Two-Tailed or Non-directional Test

- Rejection regions are located in both tails of the sampling distribution.
- For checking whether the sample is greater or less than a range of values, use the two-tailed testing.
- It is used for null hypothesis testing.

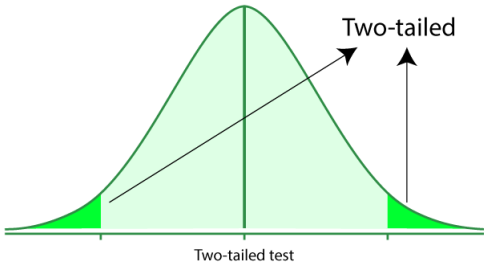


Figure 3.9 – Two tailed test

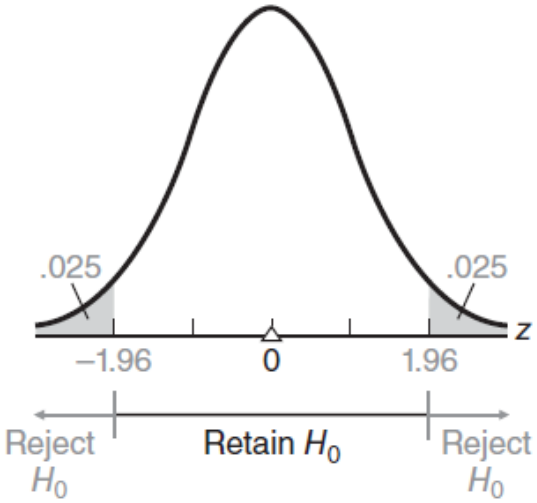


Figure 3.10 – Two-Tailed or Nondirectional Test

- Figure 3.10 shows rejection regions that are associated with both tails of the hypothesized sampling distribution.
- The corresponding decision rule, with its pair of critical z scores of ±1.96, is referred to as a two-tailed or nondirectional test.

Difference Between One and Two-Tailed Test:

One-Tailed Test	Two-Tailed Test
A test of any statistical hypothesis, where the alternative hypothesis is **one-tailed** either right-tailed or left-tailed.	A test of a statistical hypothesis, where the alternative hypothesis is **two-tailed**.
For one-tailed, use either **>** or **<** sign for the alternative hypothesis.	For two-tailed, use **≠** sign for the alternative hypothesis.
When the alternative hypothesis specifies a direction then use a one-tailed test.	If no direction is given then use a two-tailed test.
Critical region lies entirely on either the right side or left side of the sampling distribution.	Critical region is given by the portion of the area lying in both the tails of the probability curve of the test statistic.
Here, the Entire **level of significance (α)** i.e. 5% has either in the left tail or right tail.	It splits the **level of significance (α)** into half.
Rejection region is either from the left side or right side of the sampling distribution.	Rejection region is from both sides i.e. left and right of the sampling distribution.
It checks the relation between the variable in a single direction.	It checks the relation between the variables in any direction.
It is used to check whether the one mean is different from another mean or not.	It is used to check whether the two mean different from one another or not.

Example 3.17

For each of the following situations, indicate whether H0 should be retained or rejected.

Given a one-tailed test, lower tail critical with a = .01, and

(a) z = – 2.34 (b) z = – 5.13 (c) z = 4.04

Given a one-tailed test, upper tail critical with a = .05, and

(d) z = 2.00 (e) z = – 1.80 (f) z = 1.61

a. Reject *H*0 at the .01 level of significance because $z = -2.34$ is more negative than –2.33.
b. Reject *H*0 at the .01 level of significance because $z = -5.13$ is more negative than –2.33.
c. Retain *H*0 at the .01 level of significance because $z = 4.04$ is *less negative* than –2.33. (The value of the observed *z* is in the direction of no concern.)
d. Reject *H*0 at the .05 level of significance because $z = 2.00$ is more positive than 1.65.
e. Retain *H*0 at the .05 level of significance because $z = -1.80$ is *less positive* than 1.65. (The value of the observed *z* is in the direction of no concern.)
f. Retain *H*0 at the .05 level of significance because $z = 1.61$ is less positive than 1.65.

Example 3.18

Specify the decision rule for each of the following situations (referring to Table to find critical z values):

(a) a two-tailed test with a = .05

(b) a one-tailed test, upper tail critical, with a = .01

(c) a one-tailed test, lower tail critical, with a = .05

(d) a two-tailed test with a = .01

a. Reject *H*0 at the .05 level of significance if *z* equals or is more positive than 1.96 of if *z* equals or is more negative than –1.96.
b. Reject *H*0 at the .01 level of significance if *z* equals or is more positive than 2.33.
c. Reject *H*0 at the .05 level of significance if *z* equals or is more negative than –1.65.
d. Reject *H*0 at the .01 level of significance if *z* equals or is more positive than 2.58 or if *z* equals or is more negative than –2.58.

6. Discuss in detail about Estimation.

➢ **POINT ESTIMATE**

- A single value that represents some unknown population characteristic, such as the population mean.
- The best single point estimate for the unknown population mean is simply the observed value of the sample mean.

Example 3.19

A random sample of 200 graduates of U.S. colleges reveals a mean annual income of $62,600. What is the best estimate of the unknown mean annual income for all graduates of U.S. colleges?

$62,600

➢ **CONFIDENCE INTERVAL (CI) FOR μ**

- A confidence interval for μ uses a range of values that, with a known degree of certainty, includes an unknown population characteristic, such as a population mean.

Confidence Interval for μ Based on z

CONFIDENCE INTERVAL FOR μ (BASED ON z)

$$\bar{X} \pm (z_{conf})(\sigma_{\bar{X}})$$

where

$\bar{X}$ represents the sample mean;

z_{conf} represents a number from the standard normal table that satisfies the confidence specifications for the confidence interval; and

$\sigma_{\bar{X}}$ represents the standard error of the mean.

Example 3.20

Reading achievement scores are obtained for a group of fourth graders. A score of 4.0 indicates a level of achievement appropriate for fourth grade, a score below 4.0 indicates underachievement, and a score above 4.0 indicates overachievement. Assume that the population standard deviation equals 0.4. A random sample of 64 fourth graders reveals a mean achievement score of 3.82.

a. Construct a 95 percent confidence interval for the unknown population mean. (Remember to convert the standard deviation to a standard error.)

b. Interpret this confidence interval; that is, do you find any consistent evidence either of overachievement or of underachievement?

(a) $3.82 \pm 1.96\left(\frac{.4}{\sqrt{64}}\right) = \begin{cases}3.92\\3.72\end{cases}$

(b) Can claim, with 95 percent confidence, that the interval between 3.72 and 3.92 includes the true population mean reading score for the fourth graders. All of these values suggest that, on average, the fourth graders are underachieving

Example 3.21

Before taking the GRE, a random sample of college seniors received special training on how to take the test. After analysing their scores on the GRE, the investigator reported a dramatic gain, relative to the national average of 500, as indicated by a 95 percent confidence interval of 507 to 527. Are the following interpretations true or false?

a. About 95 percent of all subjects scored between 507 and 527.
b. The interval from 507 to 527 refers to possible values of the population mean for all students who undergo special training.
c. The true population mean definitely is between 507 and 527.
d. This particular interval describes the population mean about 95 percent of the time.
e. In practice, we never really know whether the interval from 507 to 527 is true or false.
f. We can be reasonably confident that the population mean is between 507 and 527.

a. False. We can be 95 percent confident that the mean for all subjects will be between 507 and 527.
b. True
c. False. We can be reasonably confident—but not absolutely confident—that the true population mean lies between 507 and 527.
d. False. This particular interval either describes the one true population mean or fails to describe the one true population mean.
e. True
f. True

- **LEVEL OF CONFIDENCE**
  - The **level of confidence** indicates the percent of time that a series of confidence intervals includes the unknown population characteristic, such as the population mean.
  - Any level of confidence may be assigned to a confidence interval merely by substituting an appropriate value for *zconf* in Formula

Choosing a Level of Confidence

- Although many different levels of confidence have been used, 95 percent and 99 percent are the most prevalent.

➢ EFFECT OF SAMPLE SIZE

- The larger the sample size, the smaller the standard error and, hence, the more precise (narrower) the confidence interval will be.
- Indeed, as the sample size grows larger, the standard error will approach zero and the confidence interval will shrink to a point estimate.
- Given this perspective, the sample size for a confidence interval, unlike that for a hypothesis test, never can be too large.
- Factors to select the sample size
  i. **Experience** – Small samples can result in wide confidence interval and risk of errors.
  ii. **Confidence Level** – Larger the confidence level, larger the sample size.

Example 3.22

On the basis of a random sample of 120 adults, a pollster reports, with 95 percent confidence, that between 58 and 72 percent of all Americans believe in life after death.

a. If this interval is too wide, what, if anything, can be done with the existing data to obtain a narrower confidence interval?

b. What can be done to obtain a narrower 95 percent confidence interval if another similar investigation is being planned?

a. Switch to an interval having a lesser degree of confidence, such as 90 percent or 75 percent.
b. Increase the sample size.

➢ HYPOTHESIS TESTS OR CONFIDENCE INTERVALS?

- Hypothesis tests merely indicate whether or not an effect is present, whereas Confidence intervals indicate the possible size of the effect.
- Confidence intervals tend to be more informative than hypothesis tests.

Example 3.23

In a recent scientific sample of about 900 adult Americans, 70 percent favour stricter gun control of assault weapons, with a margin of error of ±4 percent for a 95 percent confidence interval. Therefore, the 95 percent confidence interval equals 66 to 74 percent. Indicate whether the following interpretations are true or false:

a. The interval from 66 to 74 percent refers to possible values of the sample percent.

b. The true population percent is between 66 and 74 percent.

c. In the long run, a series of intervals similar to this one would fail to include the population percent about 5 percent of the time.

d. We can be reasonably confident that the population percent is between 66 and 74 percent.

(a) False. The interval from 66 to 74 percent refers to possible values of the population proportion.

(b) False. Can be reasonably confident—but not absolutely confident—that the true population proportion is between 66 and 74 percent.

(c) True

(d) True

Example 3.23

For the population at large, the Wechsler Adult Intelligence Scale is designed to yield a normal distribution of test scores with a mean of 100 and a standard deviation of 15. School district officials wonder whether, on the average, an IQ score different from 100 describes the intellectual aptitudes of all students in their district. Wechsler IQ scores are obtained for a random sample of 25 of their students, and the mean IQ is found to equal 105. Using the step-by-step procedure, test the null hypothesis at the .05 level of significance.

Research Problem

Does the mean IQ of all students in the district differ from 100?

Statistical Hypotheses

$H_0: \mu = 100$

$H_1: \mu \neq 100$

Decision Rule

Reject H_0 at the .05 level of significance if z equals or is more positive than 1.96 or if z equals or is more negative than –1.96.

Calculations

Given that $\bar{X} = 105$; $\sigma_{\bar{X}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$

$$z = \frac{105 - 100}{3} = \frac{5}{3} = 1.67$$

Decision

Retain H_0 at the .05 level of significance because $z = 1.67$ is less positive than 1.96.

Interpretation

There is no evidence that the mean IQ of all students differs from 100.

Example 3.24

Consult the power curves in Figure 11.7 to estimate the approximate detection rates, rounded to the nearest tenth, for the following situations:

(a) a three-point effect, with a sample size of 29

(b) a six-point effect, with a sample size of 13

(c) a twelve-point effect, with a sample size of 13

(a) .3

(b) .4

(c) .9

Example 3.25

An investigator consults a chart to determine the sample size required to detect an eight-point effect with a probability of .80. What happens to this detection rate of .80—will it actually be smaller, the same, or larger—if, unknown to the investigator, the true effect actually equals

(a) twelve points?

(b) five points?

a. The power for the 12-point effect is *larger* than .80 because the true sampling distribution is shifted further into the rejection region for the false *H*0.
b. The power for the 5-point effect is *smaller* than .80 because the true sampling distribution is shifted further into the retention region for the false *H*0.

Example 3.26

In Question 10.5 on page 191, it was concluded that, the mean salary among the population of female members of the American Psychological Association is less than that ($82,500) for all comparable members who have a doctorate and teach full time.

(a) Given a population standard deviation of $6,000 and a sample mean salary of $80,100 for a random sample of 100 female members, construct a 99 percent confidence interval for the mean salary for all female members.

(b) Given this confidence interval, is there any consistent evidence that the mean salary for all female members falls below $82,500, the mean salary for all members?

$$\textbf{(a)}\quad 80{,}100 \pm 2.58\left(\frac{6{,}000}{\sqrt{100}}\right) = \begin{cases} 81{,}648 \\ 78{,}552 \end{cases}$$

(b) can claim, with 99 percent confidence, that the interval between $78,552 and $81,648 includes the *true population mean* salary for all female members of the American Psychological Association. All of these values suggest that, on average, females' salaries are less than males' salaries.

Example 3.27

Imagine that one of the following 95 percent confidence intervals estimates the effect of vitamin C on IQ scores:

95% CONFIDENCE INTERVAL	LOWER LIMIT	UPPER LIMIT
1	100	102
2	95	99
3	102	106
4	90	111
5	91	98

(a) Which one most strongly supports the conclusion that vitamin C increases IQ scores?

(b) Which one implies the largest sample size?

(c) Which one most strongly supports the conclusion that vitamin C decreases IQ scores?

(c) Which one would most likely stimulate the investigator to conduct an additional experiment using larger sample sizes?

(a) 3 **(b)** 1 **(c)** 5 **(d)** 4

UNIT IV – ANALYSIS OF VARIANCE

SYLLABUS:

t-test for one sample – sampling distribution of t – t-test procedure – t-test for two independent samples – p-value – statistical significance – t-test for two related samples. F-test – ANOVA – Two-factor experiments – three f-tests – two-factor ANOVA –Introduction to chi-square tests.

PART A

1. Define Sampling Distribution of t.

- The distribution that would be obtained if a value of t were calculated for each sample mean for all possible random samples of a given size from some population.

2. Define Degree of Freedom.

- Degrees of freedom (df) refers to the number of values free to vary when, for example, sample variability is used to estimate the unknown population variability.

DEGREES OF FREEDOM (ONE SAMPLE)

$$df = n - 1$$

where *df* represents degrees of freedom and *n* equals the sample size.

3. What is t-test or t-ratio?

A replacement for the z ratio whenever the unknown population standard deviation must be estimated.

***t* RATIO FOR A SINGLE POPULATION MEAN**

$$t = \frac{sample\ mean - hypothesized\ population\ mean}{estimated\ standard\ error} = \frac{\bar{X} - \mu_{hyp}}{s_{\bar{X}}}$$

with its *t* sampling distribution and $n - 1$ degrees of freedom.

4. Formulate the estimation of standard error and estimated standard error of the mean.

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

where $s_{\bar{X}}$ represents the estimated standard error of the mean;
n equals the sample size; and
s has been defined as

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$$

where *s* is the sample standard deviation;

df refers to the degrees of freedom; and
SS has been defined as

$$SS = \Sigma(X - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

- This new version of the standard error, the **estimated standard error of the mean,** is used whenever the unknown population standard deviation must be estimated.

5. Discuss the steps in Calculation for the t Test.

Panel I

- This panel generates values for the sample mean, *X*, and the sample standard deviation, *s*.

Panel II

- Dividing the sample standard deviation, *s*, by the square root of the sample size, *n*, gives the value for the estimated standard error, $s_{\bar{X}}$.

Panel III

- Finally, dividing the difference between the sample mean, *X*, and the null hypothesized value, μ_{hyp}, by the estimated standard error, $s_{\bar{X}}$, yields the value of the *t* ratio.

6. Define confidence intervals for μ based on t.

- When the population standard deviation is unknown and, therefore, must be estimated, as in the present case, *t* replaces *z* in the new formula for a confidence interval:

CONFIDENCE INTERVAL FOR μ BASED ON *t*

$$\bar{X} \pm (t_{conf})(s_{\bar{X}})$$

where *X* represents the sample mean;
t_{conf} represents a number (distributed with $n - 1$ degrees of freedom) from the *t* tables, which satisfies the confidence specifications for the confidence interval; and
$s_{\bar{X}}$ represents the estimated standard error of the mean.

7. Define two independent samples.

- Observations in each sample are based on different (and unmatched) subjects.
- When samples are independent, observations in one sample are not paired, on a one-to-one basis, with observations in the other sample.

8. Define Sampling Distribution of $\bar{X}_1 - \bar{X}_2$

- Differences between sample means based on all possible pairs of random samples from two underlying populations.
- It represents the entire spectrum of differences between sample means based on all possible pairs of random samples from the two underlying populations.

9. Define Mean of the Sampling Distribution, $\mu_{\bar{X}_1 - \bar{X}_2}$

- The mean of the new sampling distribution of $\bar{X}_1 - \bar{X}_2$ equals the difference between population means, that is,

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

Where $\mu_{\bar{X}_1 - \bar{X}_2}$ is the mean of the new sampling distribution and $\bar{X}_1 - \bar{X}_2$ is the difference between population means.

10. Define Standard Error of the Sampling Distribution $\sigma_{\bar{X}_1 - \bar{X}_2}$

- A rough measure of the average amount by which any sample mean difference deviates from the difference between population means.

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where

$\sigma_{\bar{X}_1 - \bar{X}_2}$ is the new standard error, σ_1^2 and σ_2^2 are the two population variances, and n1 and n2 are the two sample sizes.

11. Define t – ratio for two population means or two independent samples.

***t* RATIO FOR TWO POPULATION MEANS (TWO INDEPENDENT SAMPLES)**

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)_{hyp}}{s_{\bar{X}_1 - \bar{X}_2}}$$

12. List the steps for calculating t – ratio for two population means or two independent samples.

Panel I

- Requiring the most computational effort, this panel produces values for the two sample means, X_1 and X_2, and for the two sample sums of squares, SS_1 and SS_2,

Panel II - Pooled Variance Estimate, s_p^2

- The most accurate estimate of the population variance (assumed to be the same for both populations) based on a combination of two sample sums of squares and their degrees of freedom.

Panel III - Estimated Standard Error, $s_{\bar{X}_1-\bar{X}_2}$

- The standard deviation of the sampling distribution for the difference between means whenever the unknown variance common to both populations must be estimated.

Panel IV

- Finally, dividing the difference between the two sample means, $\bar{X}_1 - \bar{X}_2$, and the null hypothesized population mean difference, $(\mu_1 - \mu_2)_{hyp}$, (of zero) by the estimated standard error, $s_{\bar{X}_1-\bar{X}_2}$, generates a value for the t ratio.

13. Define about p-values.

p-value

- The p-value for a test result represents the degree of rarity of that result, given that the null hypothesis is true.
- Smaller p-values tend to discredit the null hypothesis and to support the research hypothesis.
- The *p*-value represents the proportion of area, beyond the observed result, in the tail of the sampling distribution.

14. Which should be used Level of Significance or *p*-Value?

- Specified *before* the test result has been observed, the level of significance describes a degree of rarity that, if attained subsequently by the test result, triggers the decision to reject *H*0.
- Specified *after* the test result has been observed, a *p*-value describes the most impressive degree of rarity actually attained by the test result.

15. What is Statistical Significance?

- **Statistical significance** between pairs of sample means implies only that the null hypothesis is probably false, and not whether it's false because of a large or small difference between population means.

16. Define t-test for two related samples.

- The null hypothesis for two related samples can be tested with a *t* ratio.

$$t = \frac{(sample\ mean\ difference) - (hypothesized\ population\ mean\ difference)}{estimated\ standard\ error}$$

t RATIO FOR TWO POPULATION MEANS (TWO RELATED SAMPLES)

$$t = \frac{\bar{D} - \mu_{D_{hyp}}}{s_{\bar{D}}}$$

which has a *t* sampling distribution with $n - 1$ degrees of freedom, D represents the sample mean of the difference scores; *Dhyp* represents the hypothesized population mean (of zero) for the difference scores; and $s_{\bar{D}}$ represents the estimated standard error of $\bar{D}$,

17. List the steps in calculations for the *t* test

Panel I

- Panel I involves most of the computational labour, and it generates values for the sample mean difference, *D*, and the sample standard deviation for the difference scores, s_D.

Panel II

- Dividing the sample standard deviation, *sD*, by the square root of its sample size, *n*, gives the estimated standard error, s_D .

Panel III

- Finally, dividing the difference between the sample mean, *D*, and the null hypothesized value, *Dhyp*(of zero), by the estimated standard error, s_D, culminates in the value for the *t* ratio.

18. What is f – test?

- F reflects the ratio of the observed differences between all sample means (measured as variability between groups) in the numerator and the estimated error term or pooled variance estimate (measured as variability within groups) in the denominator term, that is,

F RATIO

$$F = \frac{variability\ between\ groups}{variability\ within\ groups}$$

19. What is ANOVA? Discuss in detail about one factor ANOVA.

Analysis of Variance (ANOVA)

- When data are quantitative, an overall test of the null hypothesis for more than two population means is known as analysis of variance.
- An overall test of the null hypothesis for more than two population means.

One-Factor ANOVA

- The simplest type of ANOVA that tests for differences among population means categorized by only one independent variable.
- In a one-factor ANOVA, a single null hypothesis is tested with one F ratio.

20. Define Two-Factor ANOVA

- A more complex type of analysis that tests whether differences exist among population means categorized by two factors or independent variables.
- In two-factor ANOVA, three different null hypotheses are tested, one at a time, with three F ratios: Fcolumn, Frow, and Finteraction

21. What is three f tests?

$$F_{column} = \frac{\text{Between columns}}{\text{Within cells}}$$

$$F_{row} = \frac{\text{Between rows}}{\text{Within cells}}$$

$$F_{interaction} = \frac{\text{Interaction}}{\text{Within cells}}$$

22. What is Chi-Square Test?

- A Chi-square test is a hypothesis testing method.
- There are two commonly used Chi-square tests:
  - Chi-square goodness of fit test
  - Chi-square test of independence.

Chi-Square Test Forumla

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where,

- **χ 2** = Chi-Square value
- **Oi =** Observed frequency
- **Ei** = Expected frequency

PART B

1. **Discuss in detail about t-test for one sample – sampling distribution of t and t – test procedure with a case study – Gas Mileage Investigation.**

- **Sampling Distribution of t**
  - The distribution that would be obtained if a value of t were calculated for each sample mean for all possible random samples of a given size from some population.
- **Degrees of freedom**
  - Degrees of freedom (df) refers to the number of values free to vary when, for example, sample variability is used to estimate the unknown population variability.

DEGREES OF FREEDOM (ONE SAMPLE)

$$df = n - 1$$

where *df* represents degrees of freedom and *n* equals the sample size.

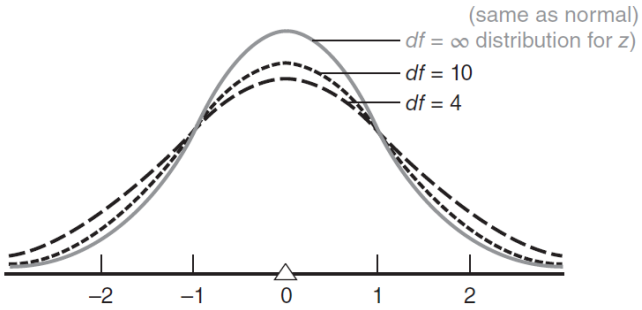


Figure 4.1 - Various t distributions.

- **Figure 4.1** shows three *t* distributions.
- When there is an infinite (∞) number of degrees of freedom, the distribution of *t* is the same as the standard normal distribution of *z*.
- Notice that even with only four or ten degrees of freedom, a *t* distribution shares a number of properties with the normal distribution.
- All *t* distributions are symmetrical, unimodal, and bell-shaped, with a dense concentration that peaks in the middle (when *t* equals 0) and tapers off both to the right and left of the middle (as *t* becomes more positive or negative, respectively).

HYPOTHESIS TEST SUMMARY: *t* TEST FOR A POPULATION MEAN (GAS MILEAGE INVESTIGATION)

Research Problem

Does the mean gas mileage for some population of cars drop below the legally required minimum of 45 mpg?

Statistical Hypotheses

$$H_0 : \mu \geq 45$$
$$H_1 : \mu < 45$$

Decision Rule

Reject H_0 at the .01 level of significance if $t \leq -3.365$ (from Table B, Appendix C, given $df = n - 1 = 6 - 1 = 5$).

Calculations

Given $\bar{X} = 43$, $s_{\bar{X}} = 0.89$

(See Table 13.1 on page 240 for computations.),

$$t = \frac{43 - 45}{0.89} = -2.25$$

Decision

Retain H_0 at the .01 level of significance because $t = -2.25$ is less negative than -3.365.

Interpretation

The population mean gas mileage *could* equal the required 45 mpg or more. The manufacturer shouldn't be penalized.

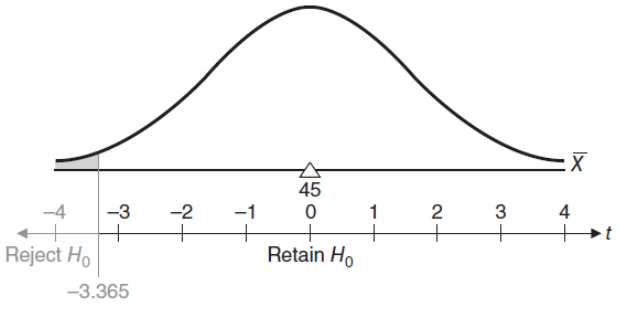


Figure 4.2

Hypothesized sampling distribution of t (gas mileage investigation).

Problem 4.1

Find the critical t values for the following hypothesis tests:

(a) two-tailed test, a = .05, df = 12

(b) one-tailed test, lower tail critical, a = .01, df = 19

(c) one-tailed test, upper tail critical, a = .05, df = 38

(d) two-tailed test, a = .01, df = 48

(a) ±2.179 **(c)** 1.697

(b) –2.539 **(d)** ±2.704

- ***t* TEST**

t Ratio

- A replacement for the z ratio whenever the unknown population standard deviation must be estimated.

t RATIO FOR A SINGLE POPULATION MEAN

$$t = \frac{\text{sample mean} - \text{hypothesized population mean}}{\text{estimated standard error}} = \frac{\bar{X} - \mu_{hyp}}{s_{\bar{X}}}$$

with its *t* sampling distribution and $n - 1$ degrees of freedom.

- **ESTIMATING THE STANDARD ERROR ($s_{\bar{X}}$)**

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

where $s_{\bar{X}}$ represents the estimated standard error of the mean;
n equals the sample size; and
s has been defined as

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$$

where s is the sample standard deviation;
df refers to the degrees of freedom; and
SS has been defined as

$$SS = \Sigma\left(X - \bar{X}\right)^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

- This new version of the standard error, the **estimated standard error of the mean,** is used whenever the unknown population standard deviation must be estimated.

Problem 4.2

A consumers' group randomly samples 10 "one-pound" packages of ground beef sold by a supermarket. Calculate (a) the mean and (b) the estimated standard error of the mean for this sample, given the following weights in ounces: 16, 15, 14, 15, 14, 15, 16, 14, 14, 14.

(a) $\bar{X} = \frac{147}{10} = 14.7$

(b) $s = \sqrt{\frac{2167 - \frac{21609}{10}}{10-1}} = \sqrt{\frac{6.10}{9}} = \sqrt{.68} = .82$

$s_{\bar{X}} = \frac{.82}{\sqrt{10}} = \frac{.82}{3.16} = .26$

➢ **CALCULATIONS FOR THE *t* TEST**

Panel I

- This panel generates values for the sample mean, *X*, and the sample standard deviation, *s*.
- The sample standard deviation is obtained by first using Formula

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

and after dividing the sum of squares, SS, by its degrees of freedom, n − 1, extracting the square root.

Panel II

- Dividing the sample standard deviation, *s*, by the square root of the sample size, *n*, gives the value for the estimated standard error, $s_{\bar{X}}$.

Panel III

- Finally, dividing the difference between the sample mean, *X*, and the null hypothesized value, μ_{hyp}, by the estimated standard error, $s_{\bar{X}}$, yields the value of the *t* ratio.

t RATIO FOR A SINGLE POPULATION MEAN

$$t = \frac{sample\ mean - hypothesized\ population\ mean}{estimated\ standard\ error} = \frac{\overline{X} - \mu_{hyp}}{s_{\overline{X}}}$$

CALCULATIONS FOR THE *t* TEST (GAS MILEAGE INVESTIGATION)

I. FINDING $\bar{X}$ AND s

(a) Computational sequence:
Assign a value to n (1).
Sum all X scores (2).
Substitute numbers in the formula (3) and solve for $\bar{X}$.
Square each X score (4), one at a time, and then add all squared X scores (5).
Substitute numbers in the formula (6) and solve for s (7).

(b) Data and computations:

X	4 X^2
40	1600
44	1936
46	2116
41	1681
43	1849
44	1936

1 $n = 6$ 2 $\Sigma X = 258$ 5 $\Sigma X^2 = 11118$

3 $$\bar{X} = \frac{\Sigma X}{n} = \frac{258}{6} = 43$$

6 $$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 11118 - \frac{(258)^2}{6} = 11118 - \frac{66564}{6} = 11118 - 11094 = 24$$

7 $$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{24}{6-1}} = \sqrt{4.8} = 2.19$$

II. FINDING $s_{\bar{X}}$

(a) Computational sequence:
Substitute the numbers obtained above in the formula 8 and solve for $s_{\bar{X}}$.

(b) Computations:

8 $$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{2.19}{\sqrt{6}} = \frac{2.19}{2.45} = 0.89$$

III. FINDING THE OBSERVED *t*

(a) Computational sequence:
Assign a value to μ_{hyp} 9, the hypothesized population mean.
Substitute the numbers obtained above in the formula 10 and solve for t.

(b) Computations:

9 $$\mu_{hyp} = 45$$

10 $$t = \frac{\bar{X} - \mu_{hyp}}{s_{\bar{X}}} = \frac{43 - 45}{0.89} = \frac{-2}{0.89} = -2.25$$

Problem 4.3

The consumers' group suspects that a supermarket makes extra money by supplying less than the specified weight of 16 ounces in its "one-pound" packages of ground beef. Given that a random sample of 10 packages yields a mean of 14.7 ounces and an estimated standard error of the mean of 0.26 ounce, use the customary step-by-step procedure to test the null hypothesis at the .05 level of significance with t.

Research Problem

Does the mean weight for all packages of ground beef drop below the specified weight of 16 ounces?

Statistical Hypothesis

$H_0: \mu \geq 16$

$H_1: \mu < 16$

Decision Rule

Reject H_0 at the .05 level of significance if $t \leq -1.833$ given $df = 10 - 1 = 9$.

Calculations

$$t = \frac{14.7 - 16}{.26} = -5.00$$

Decision

Reject H_0.

Interpretation

The mean weight for all packages drops below the specified weight of 16 ounces.

➢ CONFIDENCE INTERVALS FOR μ BASED ON t

- When the population standard deviation is unknown and, therefore, must be estimated, as in the present case, t replaces z in the new formula for a confidence interval:

CONFIDENCE INTERVAL FOR μ BASED ON t

$$\overline{X} \pm (t_{conf})(s_{\overline{X}})$$

where X represents the sample mean;

t_{conf} represents a number (distributed with $n - 1$ degrees of freedom) from the t tables, which satisfies the confidence specifications for the confidence interval; and

$s_{\overline{X}}$ represents the estimated standard error of the mean.

Problem 4.3

The consumers' group concludes that, in spite of the claims of the supermarket, the mean weight of its "one-pound" packages of ground beef drops below the specified 16 ounces even when chance sampling variability is taken into account.

(a) Construct a 95 percent confidence interval for the true weight of all "one-pound" packages of ground beef.

(b) Interpret this confidence interval.

(a) $14.7 \pm (2.26)(.26) = \begin{cases} 15.29 \\ 14.11 \end{cases}$

(b) We can be 95 percent confident that the interval between 14.11 and 15.29 ounces includes the true population mean weight for all packages.

2. Discuss in detail about t-test for two independent samples using the case study – EPO Experiment.

TWO INDEPENDENT SAMPLES

- Observations in each sample are based on different (and unmatched) subjects.
- When samples are independent, observations in one sample are not paired, on a one-to-one basis, with observations in the other sample.

Example 4.4

Identifying the treatment group with μ1, specify both the null and alternative hypotheses for each of the following studies. Select a directional alternative hypothesis only when a word or phrase justifies an exclusive concern about population mean differences in a particular direction.

a. After randomly assigning migrant children to two groups, a school psychologist determines whether there is a difference in the mean reading scores between groups exposed to either a special bilingual or a traditional reading program.

b. On further reflection, the school psychologist decides that, because of the extra expense of the special bilingual program, the null hypothesis should be rejected only if there is evidence that reading scores are improved, on average, for the group exposed to the special bilingual program.

c. An investigator wishes to determine whether, on average, cigarette consumption is reduced for smokers who chew caffeine gum. Smokers in attendance at an antismoking workshop are randomly assigned to two groups—one that chews caffeine gum and one that does not—and their daily cigarette consumption is monitored for six months after the workshop.

d. A political scientist determines whether males and females differ, on average, about the amount of money that, in their opinion, should be spent by the U.S. government on homeland security. After being informed about the size of the current budget for homeland security, in billions of dollars, randomly selected males and females are asked to indicate the percent by which they would alter this amount—for example, –8 percent for an 8 percent reduction, 0 percent for no change, 4 percent for a 4 percent increase.

(a) $H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 \neq 0$

(b) $H_0: \mu_1 - \mu_2 \leq 0$
$H_1: \mu_1 - \mu_2 > 0$

(c) $H_0: \mu_1 - \mu_2 \geq 0$
$H_1: \mu_1 - \mu_2 < 0$

(d) $H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 \neq 0$

Sampling Distribution of $\bar{X}_1 - \bar{X}_2$

- Differences between sample means based on all possible pairs of random samples from two underlying populations.
- It represents the entire spectrum of differences between sample means based on all possible pairs of random samples from the two underlying populations.

Mean of the Sampling Distribution, $\mu_{\bar{X}_1 - \bar{X}_2}$

- The mean of the new sampling distribution of $\bar{X}_1 - \bar{X}_2$ equals the difference between population means, that is,

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

Where $\mu_{\bar{X}_1 - \bar{X}_2}$ is the mean of the new sampling distribution and $\bar{X}_1 - \bar{X}_2$ is the difference between population means.

Standard Error of the Sampling Distribution $\sigma_{\bar{X}_1 - \bar{X}_2}$

- A rough measure of the average amount by which any sample mean difference deviates from the difference between population means.

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where

$\sigma_{\bar{X}_1 - \bar{X}_2}$ is the new standard error, σ_1^2 and σ_2^2 are the two population variances, and n1 and n2 are the two sample sizes.

t TEST

t Ratio

The null hypothesis can be tested using a *t* ratio. Expressed in words,

$$t = \frac{(\textit{difference between sample means}) - (\textit{hypothesized difference between population means})}{\textit{estimated standard error}}$$

Expressed in symbols,

t RATIO FOR TWO POPULATION MEANS (TWO INDEPENDENT SAMPLES)

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)_{hyp}}{s_{\bar{X}_1 - \bar{X}_2}}$$

Example 4.5

Find the critical t values for each of the following hypothesis tests:

(a) two-tailed test; a = .05; n1 = 12; n2 = 11

(b) one-tailed test, upper tail critical; a = .05; n1 = 15; n2 = 13

(c) one-tailed test, lower tail critical; a = .01; n1 = n2 = 25

(d) two-tailed test; a = .01; n1 = 8; n2 = 10

(a) ±2.080
(b) 1.706
(c) –2.423
(d) ±2.921

CALCULATIONS FOR THE t TEST

Panel I

- Requiring the most computational effort, this panel produces values for the two sample means, X_1 and X_2, and for the two sample sums of squares, SS_1 and SS_2, where

$$SS_1 = \sum X_1^2 - \frac{\left(\sum X_1\right)^2}{n_1}$$

and

$$SS_2 = \sum X_2^2 - \frac{\left(\sum X_2\right)^2}{n_2}$$

Panel II

Pooled Variance Estimate, s_P^2

- The most accurate estimate of the population variance (assumed to be the same for both populations) based on a combination of two sample sums of squares and their degrees of freedom.

POOLED VARIANCE ESTIMATE, s_P^2

$$s_p^2 = \frac{SS_1 + SS_2}{df} = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$$

Panel III

Estimated Standard Error, $s_{\bar{X}_1 - \bar{X}_2}$

- The standard deviation of the sampling distribution for the difference between means whenever the unknown variance common to both populations must be estimated.

- The **estimated standard error**, $s_{\bar{X}_1 - \bar{X}_2}$, is calculated by substituting the pooled variance, s_P^2, twice, once as an estimate for σ_1^2 and once as an estimate for σ_2^2; then dividing each term by its sample size, either n1 or n2; and finally, taking the square root of the entire expression, that is,

ESTIMATED STANDARD ERROR, $s_{\bar{X}_1 - \bar{X}_2}$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

Panel IV

- Finally, dividing the difference between the two sample means, $\bar{X}_1 - \bar{X}_2$, and the null hypothesized population mean difference, $(\mu_1 - \mu_2)_{hyp}$, (of zero) by the estimated standard error, $s_{\bar{X}_1 - \bar{X}_2}$, generates a value for the t ratio.

***t* RATIO FOR TWO POPULATION MEANS (TWO INDEPENDENT SAMPLES)**

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)_{hyp}}{s_{\bar{X}_1 - \bar{X}_2}}$$

CALCULATIONS FOR THE *t* TEST: TWO INDEPENDENT SAMPLES (EPO EXPERIMENT)

I. FINDING SAMPLE MEANS AND SUMS OF SQUARES, $\bar{X}_1$, $\bar{X}_2$, SS_1, AND SS_2

(a) Computational sequence:
Assign a value to n_1 (1).
Sum all X_1 scores (2).
Substitute numbers in the formula (3) and solve for $\bar{X}_1$.
Square each X_1 score (4), one at a time, and then add all squared X_1 scores (5).
Substitute numbers in the formula (6) and solve for SS_1.
Repeat this entire computational sequence for n_2 and X_2 and solve for $\bar{X}_2$ and SS_2.

(b) Data and computations:

ENDURANCE SCORES (MINUTES)

EPO		CONTROL	
X_1	X_1^2	X_2	X_2^2
12	144	7	49
5	25	3	9
11	121	4	16
11	121	6	36
9	81	3	9
18	324	13	169

1 $n_1 = 6$ 2 $\sum X_1 = 66$ 5 $\sum X_1^2 = 816$ $n_2 = 6$ $\sum X_2 = 36$ $\sum X_2^2 = 288$

3 $\bar{X}_1 = \dfrac{\sum X_1}{n_1} = \dfrac{66}{6} = 11$ $\bar{X}_2 = \dfrac{\sum X_2}{n_2} = \dfrac{36}{6} = 6$

6 $SS_1 = \sum X_1^2 - \dfrac{(\sum X_1)^2}{n_1}$ $SS_2 = \sum X_2^2 - \dfrac{(\sum X_2)^2}{n_2}$

$= 816 - \dfrac{(66)^2}{6}$ $= 288 - \dfrac{(36)^2}{6}$

$= 816 - 726$ $= 288 - 216$

$= 90$ $= 72$

II. FINDING THE POOLED VARIANCE, s_p^2

(a) Computational sequence:
Substitute numbers obtained above in the Formula (7) and solve for s_p^2.

(b) Computations:

7 $$s_p^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} = \frac{90 + 72}{6 + 6 - 2} = \frac{162}{10} = 16.2$$

III. FINDING THE STANDARD ERROR, $s_{\bar{X}_1-\bar{X}_2}$

(a) Computational sequence:
Substitute numbers obtained above in the formula (8) and solve for $s_{\bar{X}_1-\bar{X}_2}$.

(b) Computations:

8 $$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_P^2}{n_1}+\frac{s_P^2}{n_2}} = \sqrt{\frac{16.2}{6}+\frac{16.2}{6}} = \sqrt{\frac{32.4}{6}} = \sqrt{5.4} = 2.32$$

IV. FINDING THE OBSERVED *t* RATIO

(a) Computational sequence:
Substitute numbers obtained above in the formula (9), as well as a value of 0 for the expression $(\mu_2 - \mu_2)_{hyp}$ and solve for *t*.

(b) Computations:

9 $$t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)_{hyp}}{s_{\bar{X}_1-\bar{X}_2}} = \frac{(11-6)-0}{2.32} = \frac{5}{2.32} = 2.16$$

Example 4.6

A psychologist investigates the effect of instructions on the time required to solve a puzzle. Each of 20 volunteers is given the same puzzle to be solved as rapidly as possible. Subjects are randomly assigned, in equal numbers, to receive two different sets of instructions prior to the task. One group is told that the task is difficult (X1), and the other group is told that the task is easy (X2). The score for each subject reflects the time in minutes required to solve the puzzle. Use a t to test the null hypothesis at the .05 level of significance.

SOLUTION TIMES

"DIFFICULT" TASK	"EASY" TASK
5	13
20	6
7	6
23	5
30	3
24	6
9	10
8	20
20	9
12	12

Research Problem

Is there a difference, on average, between the puzzle-solving times required by subjects who are told that the puzzle is difficult and those required by subjects who are told that the puzzle is easy?

Statistical Hypotheses

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

Decision Rule

Reject H_0 at the .05 level of significance if $t \geq 2.101$ or $t \leq -2.101$, given $df = 10 + 10 - 2 = 18$.

Calculations

$$\bar{X}_1 = \frac{158}{10} = 15.8 \qquad \bar{X}_2 = \frac{90}{10} = 9.0$$

$$SS_1 = 3168 - \frac{(158)^2}{10} = 671.6 \qquad SS_2 = 1036 - \frac{(90)^2}{10} = 226$$

$$s_p^2 = \frac{671.6 + 226}{10 + 10 - 2} = 49.87 \qquad s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{49.87}{10} + \frac{49.87}{10}} = 3.16$$

$$t = \frac{(15.8 - 9.0) - 0}{3.16} = 2.15$$

Decision

Reject H_0 at the .05 level of significance because $t = 2.15$ exceeds 2.101.

Interpretation

Puzzle-solving times are longer, on average, for subjects who are told that the puzzle is difficult than for those who are told that the puzzle is easy.

3. Discuss in detail about p-values.

p-value

- The p-value for a test result represents the degree of rarity of that result, Given that the null hypothesis is true.
- Smaller p-values tend to discredit the null hypothesis and to support the research hypothesis.
- The *p*-value represents the proportion of area, beyond the observed result, in the tail of the sampling distribution.
- In the left panel of Figure 4.3, a relatively deviant (from zero) observed *t* is associated with a small *p*-value that makes the null hypothesis suspect, while in the right panel, a relatively non-deviant observed *t* is associated with a large *p*-value that does not make the null hypothesis suspect.

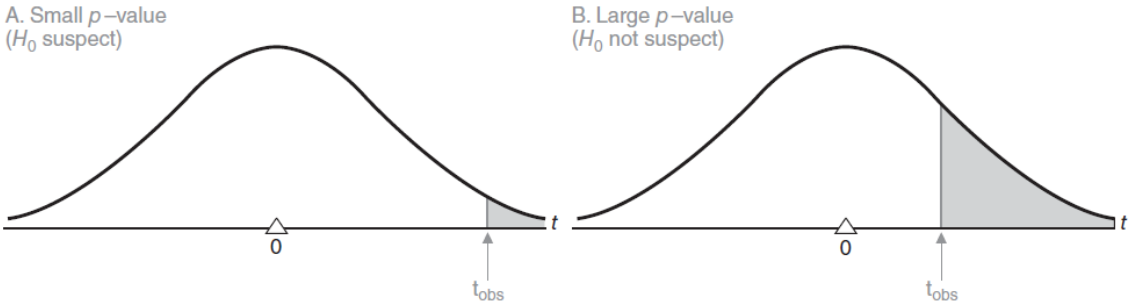


Figure 4.3
Shaded sectors showing small and large p-values.

- Figure 4.3 illustrates *one-tailed p*-values that are appropriate whenever the investigator has an interest only in deviations in a particular direction, as with a one-tailed hypothesis test.
- Otherwise, *two-tailed p*-values are appropriate.
- Two-tailed *p*-values would require equivalent shaded areas to be located in *both* tails of the sampling distribution, and the resulting two-tailed *p*-value would be twice as large as its corresponding one-tailed *p*-value.

Level of Significance or *p*-Value?

- Specified *before* the test result has been observed, the level of significance describes a degree of rarity that, if attained subsequently by the test result, triggers the decision to reject *H*0.
- Specified *after* the test result has been observed, a *p*-value describes the most impressive degree of rarity actually attained by the test result.

Example 4.7

Find the approximate p-value for each of the following test results:

(a) one-tailed test, upper tail critical; df = 12; t = 4.61

(b) one-tailed test, lower tail critical; df = 19; t = –2.41

(c) two-tailed test; df = 15; t = 3.76

(d) two-tailed test; df = 42; t = 1.305

(e) one-tailed test, upper tail critical; df = 11; t = –4.23 (Be careful!)

(a) $p < .001$ **(d)** $p > .05$
(b) $p < .05$ **(e)** $p > .05$
(c) $p < .01$

Example 4.8

Indicate which member of each of the following pairs of p-values describes the more rare test result:

(a1) p > .05 (a2) p < .05

(b1) p < .001 (b2) p < .01

(c1) p < .05 (c2) p < .01

(d1) p < .10 (d2) p < .20

(e1) p = .04 (e2) p = .02

a2, b1, c2, d1, e2

Example 4.9

Treating each of the p-values in the previous exercise separately, indicate those that would cause you to reject the null hypothesis at the .05 level of significance.

a2, b1, b2, c1, c2, e1, e2

4. What is Statistical Significance? Discuss in detail.

- Tests of hypotheses often are referred to as tests of significance, and test results are described as being statistically significant (if the null hypothesis has been rejected) or as not being statistically significant (if the null hypothesis has been retained).
- **Statistical significance** between pairs of sample means *implies only that the null hypothesis is probably false, and not whether it's false because of a large or small difference between population means.*
- *Rejecting the null hypothesis* always refers to the population, such as rejecting the hypothesized zero difference between two population means, while *statistically significant* always refers to the sample, such as assigning statistical significance to the observed difference between two sample means.

- Using excessively large sample sizes can produce statistically significant results that lack importance.
- Statistical significance merely indicates that an observed effect, such as an observed difference between the sample means, is sufficiently large, relative to the standard error, to be viewed as a rare outcome.
- (Statistical significance also implies that the observed outcome is *reliable,* that is, it would reappear as a similarly rare outcome in a repeat experiment.)
- Rejecting *H*0 at, for instance, the .05 level of significance, signifies that the probability of the observed, or a more extreme, result is less than or equal to .05 *assuming H*0 *is true*. This is a conditional probability that takes the form:

Pr (the observed result, given H0 is true) .05.

- The probability of .05 depends entirely on the *assumption* that *H*0 is true since that probability of .05 originates from the hypothesized sampling distribution centered about *H*0.

5. Discuss in detail about t-test for two related samples with case study.

***t* TEST**

- The null hypothesis for two related samples can be tested with a *t* ratio.

$$t = \frac{(sample\ mean\ difference) - (hypothesized\ population\ mean\ difference)}{estimated\ standard\ error}$$

t RATIO FOR TWO POPULATION MEANS (TWO RELATED SAMPLES)

$$t = \frac{\bar{D} - \mu_{D_{hyp}}}{s_{\bar{D}}}$$

which has a *t* sampling distribution with $n - 1$ degrees of freedom, *D* represents the sample mean of the difference scores; *Dhyp* represents the hypothesized population mean (of zero) for the difference scores; and $s_{\bar{D}}$ represents the estimated standard error of $\bar{D}$,

HYPOTHESIS TEST SUMMARY
t Test for Two Population Means: Repeated Measures (EPO Experiment)

Research Problem

When patients are measured twice, once with and once without EPO, does the population mean difference score show greater endurance due to EPO?

Statistical Hypotheses

$$H_0: \mu_D \leq 0$$
$$H_1: \mu_D > 0$$

Decision Rule

Reject H_0 at the .05 level of significance if $t \geq 2.015$ (from Table B in Appendix C, given that $df = n - 1 = 6 - 1 = 5$).

Calculations

$$t = \frac{5-0}{0.68} = 7.35 = \text{(See Table 15.1 for all computations.)}$$

Decision

Reject H_0 at the .05 level of significance because the calculated *t* of 7.35 exceeds 2.015.

Interpretation

There is evidence that when patients are measured twice, EPO is found to increase the mean endurance score.

CALCULATIONS FOR THE *t* TEST

The three panels show the computational steps that produce a *t* of 7.35 in the current experiment.

Panel I

Panel I involves most of the computational labour, and it generates values for the sample mean difference, *D*, and the sample standard deviation for the difference scores, s_D.

To obtain the sample standard deviation, first use a variation on the computation formula for the sum of squares where *X* has been replaced with *D*, that is,

$$SS_D = \Sigma D^2 - \frac{(\Sigma D)^2}{n}$$

and then, after dividing the sum of squares, *SSD*, by its degrees of freedom, $n - 1$, extract the square root, that is,

SAMPLE STANDARD DEVIATION, s_D

$$s_D = \sqrt{\frac{SS_D}{n-1}}$$

Panel II

Dividing the sample standard deviation, *sD*, by the square root of its sample size, *n*, gives the estimated standard error, $s_{\bar{D}}$, that is,

ESTIMATED STANDARD ERROR, $s_{\bar{D}}$

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n}}$$

Panel III

Finally, , dividing the difference between the sample mean, *D*, and the null hypothesized value, *Dhyp*(of zero), by the estimated standard error, $s_{\bar{D}}$, culminates in the value for the *t* ratio.

***t* RATIO FOR TWO POPULATION MEANS (TWO RELATED SAMPLES)**

$$t = \frac{\bar{D} - \mu_{D_{hyp}}}{s_{\bar{D}}}$$

CALCULATIONS FOR THE *t* TEST: REPEATED MEASURES (EPO EXPERIMENT)

I. FINDING THE MEAN AND STANDARD DEVIATION, $\bar{D}$ AND s_D

(a) Computational sequence:
Assign a value to n, the number of difference scores (1)
Subtract X_2 from X_1 to obtain D (2)
Sum all D scores (3)
Substitute numbers in the formula (4) and solve for $\bar{D}$
Square each D score (5), one at a time, and then add all squared D scores (6)
Substitute numbers in the formula (7) for SS_D, and then solve for s_D (8)

(b) Data and computations:

ENDURANCE SCORES (MINUTES)

PATIENT	EPO X_1	CONTROL X_2	DIFFERENCE SCORES 2 D	5 D^2
1	12	7	5	25
2	5	3	2	4
3	11	4	7	49
4	11	6	5	25
5	9	3	6	36
6	18	13	5	25
1 $n = 6$			3 $\Sigma D = 30$	6 $\Sigma D^2 = 164$

$$4\ \bar{D} = \frac{\Sigma D}{n} = \frac{30}{6} = 5$$

$$7\ SS_D = \Sigma D^2 - \frac{(\Sigma D)^2}{n} = 164 - \frac{(30)^2}{6} = 164 - 150 = 14$$

$$8\ s_D = \sqrt{\frac{SS_D}{n-1}} = \sqrt{\frac{14}{6-1}} = \sqrt{2.8} = 1.67$$

II. FINDING THE STANDARD ERROR, $s_{\bar{D}}$

(a) Computational sequence:
Substitute numbers obtained above in the formula 9 and solve for $s_{\bar{D}}$

(b) Computations:

$$9\ s_{\bar{D}} = \frac{s_D}{\sqrt{n}} = \frac{1.67}{\sqrt{6}} = \frac{1.67}{2.45} = 0.68$$

III. FINDING THE OBSERVED *t* RATIO

(a) Computational sequence:
Substitute numbers obtained above in the formula 10, as well as a value of 0 for $\mu_{D_{hyp}}$, and solve for t.

(b) Computations:

$$10\ t = \frac{\bar{D} - \mu_{D_{hyp}}}{s_{\bar{D}}} = \frac{5-0}{0.68} = 7.35$$

Example 4.10

An investigator tests a claim that vitamin C reduces the severity of common colds. To eliminate the variability due to different family environments, pairs of children from the same family are randomly assigned to either a treatment group that receives vitamin C or a control group that receives fake vitamin C. Each child estimates, on a 10-point scale, the severity of their colds during the school year. The following scores are obtained for ten pairs of children:

	ESTIMATED SEVERITY	
PAIR NUMBER	VITAMIN C (X_1)	FAKE VITAMIN C (X_2)
1	2	3
2	5	4
3	7	9
4	0	3
5	3	5
6	7	7
7	4	6
8	5	8
9	1	2
10	3	5

Using t, test the null hypothesis at the .05 level of significance.

Research Problem

When schoolchildren are matched for home environment, does vitamin C consumption reduce the mean estimated severity of common colds?

Statistical Hypotheses

$H_0: \mu_D \geq 0$

$H_1: \mu_D < 0$

Decision Rule

Reject H_0 at the .05 level of significance if $t \leq -1.833$, given $df = 10 - 1 = 9$.

Calculations

$$\bar{D} = \frac{-15}{10} = -1.5 \quad SS_D = 37 - \frac{(15)^2}{10} = 14.5$$

$$s_D = \sqrt{\frac{14.5}{10-1}} = \sqrt{1.61} = 1.27 \quad s_{\bar{D}} = \frac{1.27}{\sqrt{10}} = 0.40$$

$$t = \frac{-1.5 - 0}{.40} = -3.75$$

Decision

Reject H_0 at the .05 level of significance because $t = -3.75$ is more negative than -1.833.

Interpretation

When schoolchildren are matched for home environment, vitamin C consumption reduces the mean estimated severity of common colds.

6. What is f – test? Discuss in detail about the purpose of f-test with the case study?

- In the two-sample case, t reflects the ratio between the observed difference between the two sample means in the numerator and the estimated standard error in the denominator.
- For three or more samples, the null hypothesis is tested with a new ratio, the F ratio.
- Essentially, F reflects the ratio of the observed differences between all sample means (measured as variability between groups) in the numerator and the estimated error term or pooled variance estimate (measured as variability within groups) in the denominator term, that is,

F RATIO

$$F = \frac{variability\ between\ groups}{variability\ within\ groups}$$

- Like t, F has its own family of sampling distributions that can be consulted to test the null hypothesis. The resulting test is known as an F test.
- An F test of the null hypothesis is based on the notion that if the null Hypothesis is true, both the numerator and the denominator of the F ratio would tend to be about the same, but if the null hypothesis is false, the numerator would tend to be larger than the denominator.

If Null Hypothesis Is True

- If the null hypothesis is true (because there is no treatment effect due to Different sleep deprivation periods), the two estimates of variability (between and within groups) would reflect only random error. In this case,

$$F = \frac{random\ error}{random\ error}$$

- Except for chance, estimates in both the numerator and the denominator are similar, and generally, *F* varies about a value of 1.

If Null Hypothesis Is False

- If the null hypothesis is false (because there is a treatment effect due to different sleep deprivation periods), both estimates still would reflect random error, but the estimate for between groups would also reflect the treatment effect. In this case,

$$F = \frac{random\ error + treatment\ effect}{random\ error}$$

- When the null hypothesis is false, the presence of a treatment effect tends to cause a chain reaction:
  - The observed differences between group means tend to be large, as does the variability between groups.
    Accordingly, the numerator term tends to exceed the denominator term, producing an F whose value is larger than 1.
- When the null hypothesis is false because of a *large* treatment effect, there is an even more pronounced chain reaction,
  - beginning with very large observed differences between group means and ending with an F whose value tends to be *considerably* larger than 1.

HYPOTHESIS TEST SUMMARY

One-Factor *F* Test (Sleep Deprivation Experiment, Outcome B)

Research Problem
On average, are subjects' aggression scores in a controlled social situation affected by sleep deprivation periods of 0, 24, or 48 hours?

Statistical Hypotheses

H_0: $\mu_0 = \mu_{24} = \mu_{48}$
H_1: H_0 is false.

Decision Rule
Reject H_0 at the .05 level of significance if $F \geq 5.14$ (from Table C, Appendix C, given $df_{between} = 2$ and $df_{within} = 6$).

Calculations

$F = 7.36$ (See Tables 16.3 and 16.6 for additional details.)

Decision
Reject H_0 at the .05 level of significance because $F = 7.36$ exceeds 5.14.

Interpretation
Hours of sleep deprivation affect the subjects' mean aggression scores in a controlled social situation.

- Full-fledged *F* tests for Outcomes A and B agree with the earlier intuitive decisions.
- Given the .05 level of significance, the null hypothesis should be retained for Outcome A, since the observed *F* of 0.75 is smaller than the critical *F* of 5.14.
- However, the null hypothesis should be rejected for Outcome B, since the observed *F* of 7.36 exceeds the critical *F*.

Example 4.11

If the null hypothesis is true, both the numerator and denominator of the F ratio would reflect only (a) . If the null hypothesis is false, the numerator of the F ratio would also reflect the (b). If the null hypothesis is false because of a large treatment effect, the value of F would tend to be considerably larger than (c).

(a) random error
(b) treatment effect
(c) one

Example 4.12

Find the critical values for the following F tests:

(a) a = .05, dfbetween = 1, dfwithin = 18

(b) a = .01, dfbetween = 3, dfwithin = 56

(c) a = .05, dfbetween = 2, dfwithin = 36

(d) a = .05, dfbetween = 4, dfwithin = 95

(a) 4.41 **(c)** 3.26
(b) 4.16 **(d)** 2.48

Example 4.13

Find the approximate p-value for the following observed F ratios, where the numbers in parentheses refer to the degrees of freedom in the numerator and denominator, respectively.

(a) F (2, 11) = 4.56

(b) F (1, 13) = 11.25

(c) F (3, 20) = 2.92

(d) F (2, 29) = 3.66

(a) $p < .05$ **(c)** $p > .05$
(b) $p < .01$ **(d)** $p < .05$

7. What is ANOVA? Discuss in detail about one factor ANOVA.

Analysis of Variance (ANOVA)

- When data are quantitative, an overall test of the null hypothesis for more than two population means is known as analysis of variance.
- An overall test of the null hypothesis for more than two population means.

One-Factor ANOVA

- The simplest type of ANOVA that tests for differences among population means categorized by only one independent variable.

Two Possible Outcomes

Example

TWO POSSIBLE OUTCOMES OF A SLEEP-DEPRIVATION EXPERIMENT: AGGRESSION SCORES

	OUTCOME A HOURS OF SLEEP DEPRIVATION			
	ZERO	TWENTY-FOUR	FORTY-EIGHT	
	3	4	2	
	5	8	4	
	7	6	6	
Group mean:	5	6	4	Grand mean = 5

	OUTCOME B HOURS OF SLEEP DEPRIVATION			
	ZERO	TWENTY-FOUR	FORTY-EIGHT	
	0	3	6	
	4	6	8	
	2	6	10	
Group mean:	2	5	8	Grand mean = 5

Table shows two fictitious experimental outcomes that, when analysed with ANOVA, produce different decisions about the null hypothesis: It is retained for one outcome but rejected for the other.

TWO SOURCES OF VARIABILITY

Differences between Group Means

- Differences of 5, 6, and 4 appear between group means in Outcome A, and these relatively small differences might reflect only chance.
- Even though the null hypothesis is true (because sleep deprivation does not affect the subjects' aggression scores), group means tend to differ merely because of chance sampling variability.
- It's reasonable to expect, therefore, that the null hypothesis for Outcome A should not be rejected.
- There appears to be a lack of evidence that sleep deprivation affects the subjects' aggression scores in Outcome A.
- On the other hand, differences of 2, 5, and 8 appear between the group means for Outcome B, and these relatively large differences might not be attributable to chance.
- Instead, they indicate that the null hypothesis probably is false (because sleep deprivation affects the subjects' aggression scores). It's reasonable to expect, therefore, that the null hypothesis for Outcome B should be rejected.

  There appears to be evidence of a **treatment effect,** that is, the existence of at least one difference between the population means defined by the independent variable (sleep deprivation).

Two-Factor ANOVA

- A more complex type of analysis that tests whether differences exist among population means categorized by two factors or independent variables.

Example

- For computational simplicity, assume that the social psychologist randomly
- assigns two subjects to be tested (one at a time) with crowds of either zero, two, or four people and either the nondangerous or dangerous conditions.
- The resulting six groups, each consisting of two subjects, represent all possible combinations of the two factors.*

OUTCOME OF TWO-FACTOR EXPERIMENT (REACTION TIMES IN MINUTES)

	CROWD SIZE						
DEGREE OF DANGER	**ZERO**		**TWO**		**FOUR**		**ROW MEAN**
Dangerous	8 8	8	8 6	7	10 8	9	8
Nondangerous	9 11	10	15 19	17	24 18	21	16
Column mean		9		12		15	Grand mean = 12

Note: Shaded numbers are means.

The shaded numbers represent four different types of means:

1. The three column means (9, 12, 15) represent the mean reaction times for Each crowd size when degree of danger is ignored. Any differences among these column means not attributable to chance are referred to as the main effect of crowd size on reaction time. In ANOVA, **main effect** always refers to the effect of a single factor, such as crowd size, when any other factor, such as degree of danger, is ignored.

2. The two row means (8, 16) represent the mean reaction times for degree of danger when crowd size is ignored. Any difference between these row means not attributable to chance is referred to as the main effect of degree of danger on reaction time.

3. The mean of the reaction times for each group of two subjects yields the six means (8, 7, 9, 10, 17, 21) for each combination of the two factors. Often referred to as cell means or treatment-combination means, these means reflect not only the main effects for crowd size and degree of danger described earlier but, more importantly, any effect due to the interaction between crowd size and degree of danger, as described below.

4. Finally, the one mean for all three column means—or for both row means—yields the overall or grand mean (12) for all subjects in the study.

Main Effect

- The effect of a single factor when any other factor is ignored.

Graphs for Main Effects

- The slanted line in panel A of Figure 4.4 depicts the large differences between column means, that is, between mean reaction times for subjects, regardless of degree of danger, with crowds of zero, two, and four people.
- The relatively steep slant of this line suggests that the null hypothesis for crowd size might be rejected.
- The steeper the slant is, the larger the observed differences between column means and the greater the suspected main effect of crowd size.
- On the other hand, a fairly level line in panel A of Figure 4.4 would have reflected the relative absence of any main effect due to crowd size.
- The slanted line in panel B of Figure 4.4 depicts the large difference between row means, that is, between mean reaction times for dangerous and non dangerous conditions, regardless of crowd size.
- The relatively steep slope of this line suggests that the null hypothesis for degree of danger also might be rejected; that is, there might be a main effect due to degree of danger.

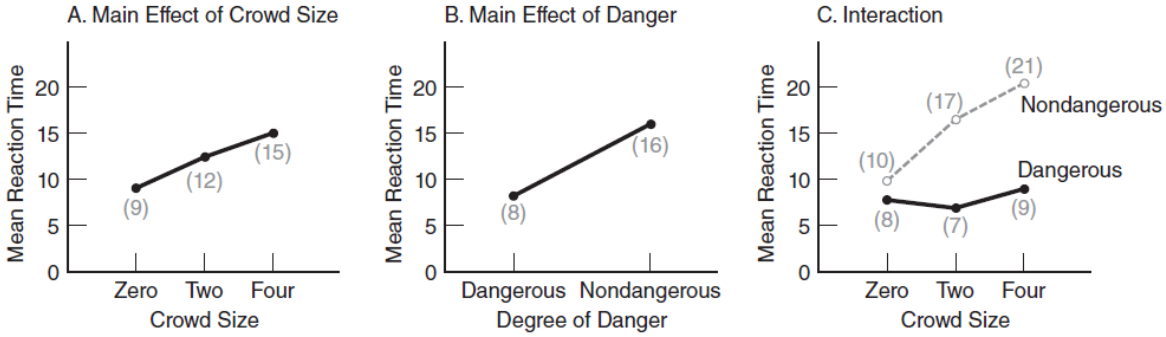


Figure 4.4 depicts the large differences between column means

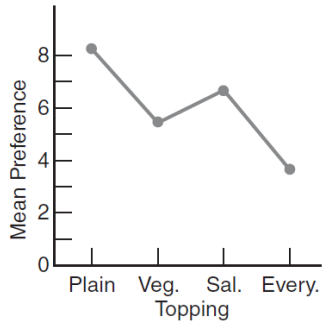
Example 4.14

A college dietician wishes to determine whether students prefer a particular pizza topping (either plain, vegetarian, salami, or everything) and one type of crust (either thick or thin). A total of 160 volunteers are randomly assigned to one of the eight cells in this two-factor experiment. After eating their assigned pizza, the 20 subjects in each cell rate their preference on a scale ranging from 0 (inedible) to 10 (the best). The results, in the form of means for cells, rows, and columns, are as follows:

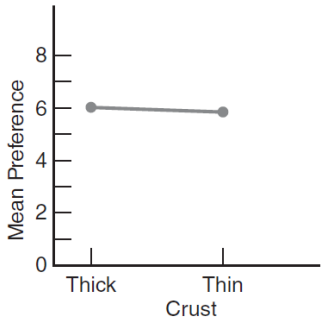
MEAN PREFERENCE SCORES OR PIZZA AS A FUNCTION OF TOPPING AND CRUST

	TOPPING				
CRUST	PLAIN	VEGETARIAN	SALAMI	EVERYTHING	ROW
Thick	7.2	5.7	4.8	6.1	6.0
Thin	8.9	4.8	8.4	1.3	5.9
Column	8.1	5.3	6.6	3.7	

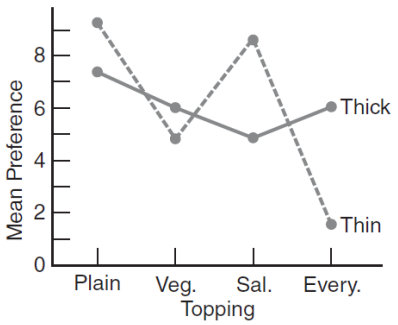
Construct graphs for each of the three possible effects, and use this information to make preliminary interpretations about pizza preferences. Ordinarily, of course, you would verify these speculations by performing an ANOVA—a task that cannot be performed for these data, since only means are supplied.



Interpretation: H_0 for topping is suspect.



Interpretation: H_0 for crust is not suspect.



Interpretation: H_0 for interaction is suspect.

THREE F TESTS

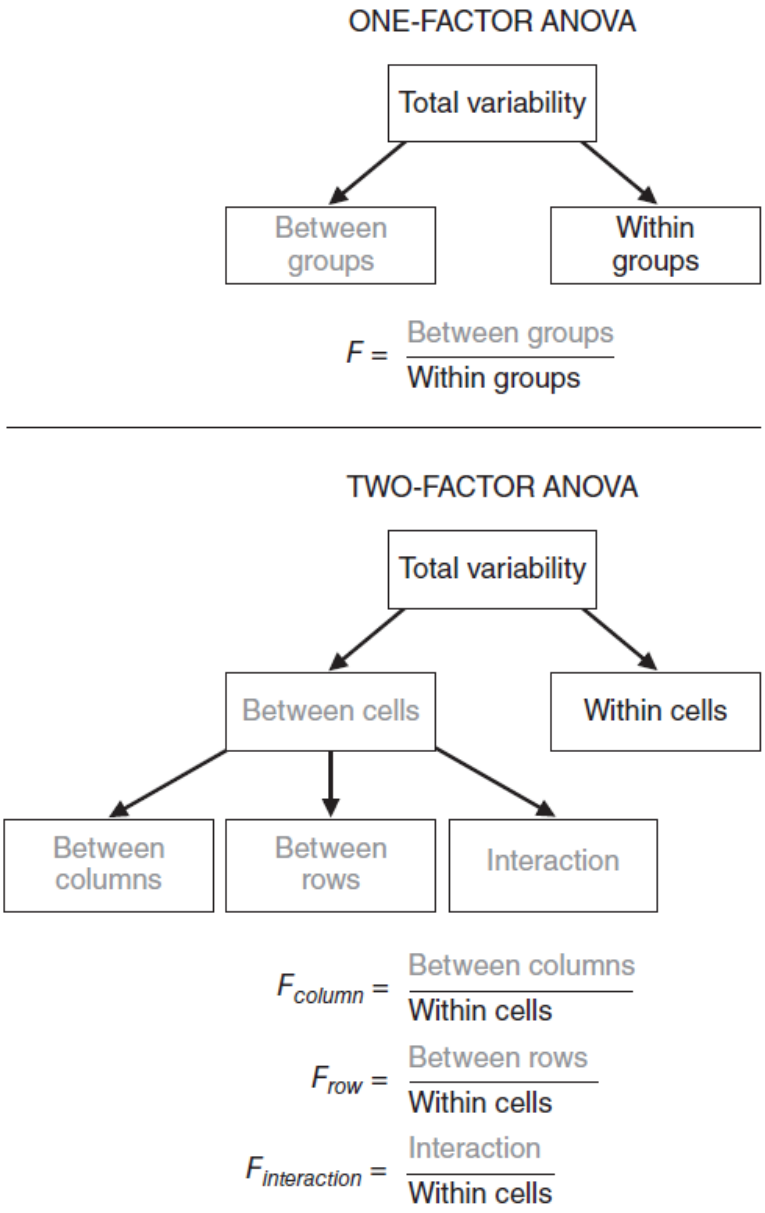


FIGURE 4.5 - Sources of variability and F ratios in one- and two-factor ANOVAs.

- As suggested in **Figure 4.**5, F ratios in both a one- and a two-factor ANOVA always consist of a numerator (shaded) that measures some aspect of variability between groups or cells and a denominator that measures variability within groups or cells.
- In a one-factor ANOVA, a single null hypothesis is tested with one F ratio.
- **In two-factor ANOVA, three different null hypotheses are tested, one at a time, with three F ratios: Fcolumn, Frow, and Finteraction.**
- The numerator of each of these three F ratios reflects a different aspect of variability between cells:
    - variability between columns (crowd size),
    - variability between rows (degree of danger),
    - interaction—any remaining variability between cells not attributable to either variability between columns (crowd size) or rows (degree of danger).
- The shaded numerator terms for the three F ratios in the bottom panel of Figure 4.5 estimate random error and, if present, a treatment effect (for subjects treated differently by the investigator).
- The denominator term always estimates only random error (for subjects treated similarly in the same cell).
- In practice, a sufficiently large F value is viewed as rare, given that the null hypothesis is true, and therefore, it leads to the rejection of the null hypothesis.
- Otherwise, the null hypothesis is retained.

Test Results for Two-Factor Experiment

As indicated in the boxed summary for the hypothesis test for a smoke alarm experiment, test results agree with our preliminary interpretations based on graphs. Each of the three null hypotheses is rejected at the .05 level of significance. The significant main effects indicate that crowd size and degree of danger, in turn, influence the reaction times of subjects to smoke. The significant interaction, however, indicates that the effect of crowd size on reaction times differs for nondangerous and dangerous conditions.

HYPOTHESIS TEST SUMMARY
Two-Factor ANOVA (Smoke Alarm Experiment)

Research Problem

Do crowd size and degree of danger, as well as the interaction of these two factors, influence the subjects' mean reaction times to potentially dangerous smoke?

Statistical Hypotheses

H_0: no main effect due to columns or crowd size (or $\mu_0 = \mu_2 = \mu_4$).

H_0: no main effect due to rows or degree of danger (or $\mu_{dangerous} = \mu_{nondangerous}$).

H_0: no interaction.

H_1: H_0 is not true.

(Same H_1 accommodates each H_0.)

Decision Rule

Reject H_0 at the .05 level of significance if F_{column} or $F_{interaction} \geq 5.14$ (from Table C in Appendix C, given 2 and 6 degrees of freedom) and if $F_{row} \geq 5.99$ (given 1 and 6 degrees of freedom).

Calculations

$$F_{column} = 6.75$$
$$F_{row} = 36.02$$
$$F_{interaction} = 5.25$$

(See Tables 18.3 and 18.6 for more details.)

Decision

Reject all three null hypotheses at the .05 level of significance because $F_{column} = 6.75$ exceeds 5.14; $F_{row} = 36.02$ exceeds 5.99; and $F_{interaction} = 5.25$ exceeds 5.14.

Interpretation

Both crowd size and degree of danger influence the subjects' mean reaction times to smoke. The interaction indicates that the influence of crowd size depends on the degree of danger. It appears that the mean reaction times increase with crowd size for nondangerous but not for dangerous conditions.

8. What is Chi-Square Test?

- A Chi-square test is a hypothesis testing method.
- Two common Chi-square tests involve checking if observed frequencies in one or more categories match expected frequencies.
- The chi-square test is a statistical test used to determine if there is a significant association between two categorical variables.
- It is a non-parametric test, meaning it does not make assumptions about the underlying distribution of the data.
- It compares the observed frequencies of the categories in a contingency table with the expected frequencies that would occur under the assumption of independence between the variables.
- The test calculates a chi-square statistic, which measures the discrepancy between the observed and expected frequencies.

Chi-Square Test Forumla

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where,

- **χ 2** = Chi-Square value
- **Oi =** Observed frequency
- **Ei** = Expected frequency

Types of Chi-square tests

- There are two commonly used Chi-square tests:
  - the Chi-square goodness of fit test
  - the Chi-square test of independence.

	Chi-Square Goodness of Fit Test	Chi-Square Test of Independence
Number of variables	One	Two
Purpose of test	Decide if one variable is likely to come from a given distribution or not	Decide if two variables might be related or not
Example	Decide if bags of candy	Decide if movie goers' decision to

	have the same number of pieces of each flavor or not	buy snacks is related to the type of movie they plan to watch
Hypotheses in example	H_o: proportion of flavors of candy are the same H_a: proportions of flavors are not the same	H_o: proportion of people who buy snacks is independent of the movie type H_a: proportion of people who buy snacks is different for different types of movies
Theoretical distribution used in test	Chi-Square	Chi-Square
Degrees of freedom	Number of categories minus 1 In our example, number of flavors of candy minus 1	Number of categories for first variable minus 1, multiplied by number of categories for second variable minus 1 In our example, number of movie categories minus 1, multiplied by 1 (because snack purchase is a Yes/No variable and 2-1 = 1)

Steps to perform a Chi-square test

For both the Chi-square goodness of fit test and the Chi-square test of independence, the same analysis steps, listed below.

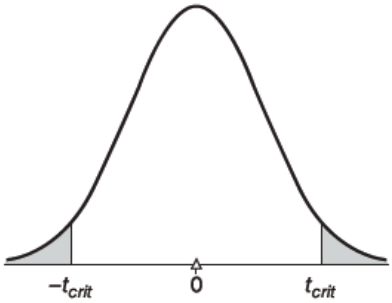
1. Define your null and alternative hypotheses before collecting your data.
2. Decide on the alpha value. For example, suppose set α=0.05 when testing for independence. Here, have decided on a 5% risk of concluding the two variables are independent when in reality they are not.
3. Check the data for errors.
4. Check the assumptions for the test.
5. Perform the test and draw your conclusion.

Properties of Chi-Square Test

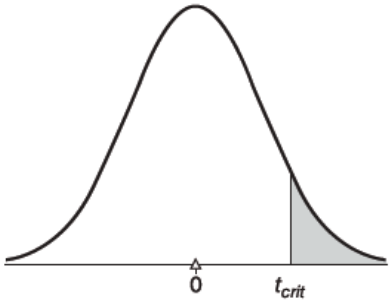
The chi-square test possesses several important properties that make it a valuable statistical tool:

Aspect	Description
Non-parametric Test	The chi-square test is non-parametric, making no assumptions about the data's underlying distribution. Applicable to categorical data.
Test for Independence	Examines association between categorical variables, determining significance of relationship or dependency, not strength or direction.
Goodness of Fit Test	Assesses how well observed data fit an expected distribution, comparing observed frequencies to expected frequencies.
Chi-Square Statistic	Measures discrepancy between observed and expected frequencies in a contingency table, indicating association or goodness of fit.
Degrees of Freedom	Depend on the number of categories in variables. Determine critical values and influence test result interpretation.
Null and Alternative Hypotheses	Null hypothesis assumes no association or difference, alternative hypothesis suggests presence of association or difference.
Test Statistic and P-value	Produces test statistic and corresponding p-value. Compare test statistic to critical value, p-value indicates probability under null hypothesis.
Interpretation	Null hypothesis rejected if test statistic exceeds critical value or p-value is less than chosen significance level. Indicates significant association or deviation.

Table B[a]
CRITICAL VALUES OF *t*



Two-tailed or Nondirectional Test
LEVEL OF SIGNIFICANCE

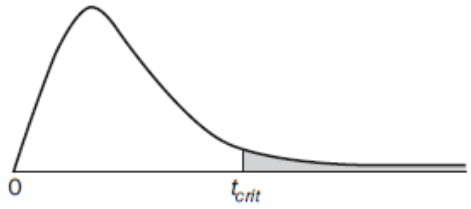


One-tailed or Directional Test
LEVEL OF SIGNIFICANCE

	$p > .05$	$p < .05$	$p < .01$	$p < .001$
df	.05*	.01**	.001	
1	12.706	63.657	636.62	
2	4.303	9.925	31.598	
3	3.182	5.841	12.924	
4	2.776	4.604	8.610	
5	2.571	4.032	6.869	
6	2.447	3.707	5.959	
7	2.365	3.499	5.408	
8	2.306	3.355	5.041	
9	2.262	3.250	4.781	
10	2.228	3.169	4.587	
11	2.201	3.106	4.437	
12	2.179	3.055	4.318	
13	2.160	3.012	4.221	
14	2.145	2.977	4.140	
15	2.131	2.947	4.073	
16	2.120	2.921	4.015	
17	2.110	2.898	3.965	
18	2.101	2.878	3.922	
19	2.093	2.861	3.883	
20	2.086	2.845	3.850	
21	2.080	2.831	3.819	
22	2.074	2.819	3.792	
23	2.069	2.807	3.767	
24	2.064	2.797	3.745	
25	2.060	2.787	3.725	
26	2.056	2.779	3.707	
27	2.052	2.771	3.690	
28	2.048	2.763	3.674	
29	2.045	2.756	3.659	
30	2.042	2.750	3.646	
40	2.021	2.704	3.551	
60	2.000	2.660	3.460	
120	1.980	2.617	3.373	
∞	1.960	2.576	3.291	

	$p > .05$	$p < .05$	$p < .01$	$p < .001$
df	.05	.01	.001	
1	6.314	31.821	318.31	
2	2.920	6.965	22.326	
3	2.353	4.541	10.213	
4	2.132	3.747	7.173	
5	2.015	3.365	5.893	
6	1.943	3.143	5.208	
7	1.895	2.998	4.785	
8	1.860	2.896	4.501	
9	1.833	2.821	4.297	
10	1.812	2.764	4.144	
11	1.796	2.718	4.025	
12	1.782	2.681	3.930	
13	1.771	2.650	3.852	
14	1.761	2.624	3.787	
15	1.753	2.602	3.733	
16	1.746	2.583	3.686	
17	1.740	2.567	3.646	
18	1.734	2.552	3.610	
19	1.729	2.539	3.579	
20	1.725	2.528	3.552	
21	1.721	2.518	3.527	
22	1.717	2.508	3.505	
23	1.714	2.500	3.485	
24	1.711	2.492	3.467	
25	1.708	2.485	3.450	
26	1.706	2.479	3.435	
27	1.703	2.473	3.421	
28	1.701	2.467	3.408	
29	1.699	2.462	3.396	
30	1.697	2.457	3.385	
40	1.684	2.423	3.307	
60	1.671	2.390	3.232	
120	1.658	2.358	3.160	
∞	1.645	2.326	3.090	

Table C[a]
CRITICAL VALUES OF *F*



FINDING *p*-VALUE
If observed *F* is
. . .smaller than light number, $p > .05$
. . .between light and dark numbers, $p < .05$
. . .larger than dark number, $p < .01$

.05 level of significance (light numbers)
.01 level of significance (dark numbers)

DEGREES OF FREEDOM IN NUMERATOR

DEGREES OF FREEDOM IN DENOMINATOR	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161 **4,052**	200 **4,999**	216 **5,403**	225 **5,625**	230 **5,764**	234 **5,859**	237 **5,928**	239 **5,981**	241 **6,022**	242 **6,056**	243 **6,082**	244 **6,106**	245 **6,142**	246 **6,169**	248 **6,208**	249 **6,234**	250 **6,258**	251 **6,286**	252 **6,302**	253 **6,323**	253 **6,334**	254 **6,352**	254 **6,361**	254 **6,366**
2	18.51 **98.49**	19.00 **99.00**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.36 **99.34**	19.37 **99.36**	19.38 **99.38**	19.39 **99.40**	19.40 **99.41**	19.41 **99.42**	19.42 **99.43**	19.43 **99.44**	19.44 **99.45**	19.45 **99.46**	19.46 **99.47**	19.47 **99.48**	19.47 **99.48**	19.48 **99.49**	19.49 **99.49**	19.49 **99.49**	19.50 **99.50**	19.50 **99.50**
3	10.13 **34.12**	9.55 **30.82**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.88 **27.67**	8.84 **27.49**	8.81 **27.34**	8.78 **27.23**	8.76 **27.13**	8.74 **27.05**	8.71 **26.92**	8.69 **26.83**	8.66 **26.69**	8.64 **26.60**	8.62 **26.50**	8.60 **26.41**	8.58 **26.35**	8.57 **26.27**	8.56 **26.23**	8.54 **26.18**	8.54 **26.14**	8.53 **26.12**
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.54**	5.93 **14.45**	5.91 **14.37**	5.87 **14.24**	5.84 **14.15**	5.80 **14.02**	5.77 **13.93**	5.74 **13.83**	5.71 **13.74**	5.70 **13.69**	5.68 **13.61**	5.66 **13.57**	5.65 **13.52**	5.64 **13.48**	5.63 **13.46**
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.45**	4.82 **10.27**	4.78 **10.15**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**	4.64 **9.77**	4.60 **9.68**	4.56 **9.55**	4.53 **9.47**	4.50 **9.38**	4.46 **9.29**	4.44 **9.24**	4.42 **9.17**	4.40 **9.13**	4.38 **9.07**	4.37 **9.04**	4.36 **9.02**
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**	3.96 **7.60**	3.92 **7.52**	3.87 **7.39**	3.84 **7.31**	3.81 **7.23**	3.77 **7.14**	3.75 **7.09**	3.72 **7.02**	3.71 **6.99**	3.69 **6.94**	3.68 **6.90**	3.67 **6.88**
7	5.59 **12.25**	4.47 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **7.00**	3.73 **6.84**	3.68 **6.71**	3.63 **6.62**	3.60 **6.54**	3.57 **6.47**	3.52 **6.35**	3.49 **6.27**	3.44 **6.15**	3.41 **6.07**	3.38 **5.98**	3.34 **5.90**	3.32 **5.85**	3.29 **5.78**	3.28 **5.75**	3.25 **5.70**	3.24 **5.67**	3.23 **5.65**
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.19**	3.44 **6.03**	3.39 **5.91**	3.34 **5.82**	3.31 **5.74**	3.28 **5.67**	3.23 **5.56**	3.20 **5.48**	3.15 **5.36**	3.12 **5.28**	3.08 **5.20**	3.05 **5.11**	3.03 **5.06**	3.0 **5.00**	2.98 **4.96**	2.96 **4.91**	2.94 **4.88**	2.93 **4.86**
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.62**	3.23 **5.47**	3.18 **5.35**	3.13 **5.26**	3.10 **5.18**	3.07 **5.11**	3.02 **5.00**	2.98 **4.92**	2.93 **4.80**	2.90 **4.73**	2.86 **4.64**	2.82 **4.56**	2.80 **4.51**	2.77 **4.45**	2.76 **4.41**	2.73 **4.36**	2.72 **4.33**	2.71 **4.31**
10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.14 **5.21**	3.07 **5.06**	3.02 **4.95**	2.97 **4.85**	2.94 **4.78**	2.91 **4.71**	2.86 **4.60**	2.82 **4.52**	2.77 **4.41**	2.74 **4.33**	2.70 **4.25**	2.67 **4.17**	2.64 **4.12**	2.61 **4.05**	2.59 **4.01**	2.56 **3.96**	2.55 **3.93**	2.54 **3.91**
11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	3.01 **4.88**	2.95 **4.74**	2.90 **4.63**	2.86 **4.54**	2.82 **4.46**	2.79 **4.40**	2.74 **4.29**	2.70 **4.21**	2.65 **4.10**	2.61 **4.02**	2.57 **3.94**	2.53 **3.86**	2.50 **3.80**	2.47 **3.74**	2.45 **3.70**	2.42 **3.66**	2.41 **3.62**	2.40 **3.60**
12	4.75 **9.33**	3.88 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.92 **4.65**	2.85 **4.50**	2.80 **4.39**	2.76 **4.30**	2.72 **4.22**	2.69 **4.16**	2.64 **4.05**	2.60 **3.98**	2.54 **3.86**	2.50 **3.78**	2.46 **3.70**	2.42 **3.61**	2.40 **3.56**	2.36 **3.49**	2.35 **3.46**	2.32 **3.41**	2.31 **3.38**	2.30 **3.36**
13	4.67 **9.07**	3.80 **6.70**	3.41 **5.74**	3.18 **5.20**	3.02 **4.86**	2.92 **4.62**	2.84 **4.44**	2.77 **4.30**	2.72 **4.19**	2.67 **4.10**	2.63 **4.02**	2.60 **3.96**	2.55 **3.85**	2.51 **3.78**	2.46 **3.67**	2.42 **3.59**	2.38 **3.51**	2.34 **3.42**	2.32 **3.37**	2.28 **3.30**	2.26 **3.27**	2.24 **3.21**	2.22 **3.18**	2.21 **3.16**

Table C[a] (Continued)
CRITICAL VALUES OF *F*

FINDING *p*-VALUE
If observed *F* is
... smaller than light number, $p > .05$
... between light and dark numbers, $p < .05$
... larger than dark number, $p < .01$

DEGREES OF FREEDOM IN NUMERATOR

DEGREES OF FREEDOM IN DENOMINATOR	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.9**	**2.89**	**2.80**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**	**3.45**	**3.37**	**3.25**	**3.18**	**3.10**	**3.01**	**2.96**	**2.89**	**2.86**	**2.80**	**2.77**	**2.75**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**	**3.08**	**3.00**	**2.92**	**2.86**	**2.79**	**2.76**	**2.70**	**2.67**	**2.65**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**	**3.00**	**2.91**	**2.83**	**2.78**	**2.71**	**2.68**	**2.62**	**2.59**	**2.57**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**	**2.92**	**2.84**	**2.76**	**2.70**	**2.63**	**2.60**	**2.54**	**2.51**	**2.49**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**	**2.86**	**2.77**	**2.69**	**2.63**	**2.56**	**2.53**	**2.47**	**2.44**	**2.42**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**	**2.80**	**2.72**	**2.63**	**2.58**	**2.51**	**2.47**	**2.42**	**2.38**	**2.36**
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**	**3.02**	**2.94**	**2.83**	**2.75**	**2.67**	**2.58**	**2.53**	**2.46**	**2.42**	**2.37**	**2.33**	**2.31**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**	**2.97**	**2.89**	**2.78**	**2.70**	**2.62**	**2.53**	**2.48**	**2.41**	**2.37**	**2.32**	**2.28**	**2.26**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**	**2.93**	**2.85**	**2.74**	**2.66**	**2.58**	**2.49**	**2.44**	**2.36**	**2.33**	**2.27**	**2.23**	**2.21**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**	**2.89**	**2.81**	**2.70**	**2.62**	**2.54**	**2.45**	**2.40**	**2.32**	**2.29**	**2.23**	**2.19**	**2.17**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.86	1.82	1.78	1.76	1.72	1.70	1.69
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**	**2.86**	**2.77**	**2.66**	**2.58**	**2.50**	**2.41**	**2.36**	**2.28**	**2.25**	**2.19**	**2.15**	**2.13**

Table C[a] (Continued)
CRITICAL VALUES OF *F*

27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**	**2.83**	**2.74**	**2.63**	**2.55**	**2.47**	**2.38**	**2.33**	**2.25**	**2.21**	**2.16**	**2.12**	**2.10**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**	**2.80**	**2.71**	**2.60**	**2.52**	**2.44**	**2.35**	**2.30**	**2.22**	**2.18**	**2.13**	**2.09**	**2.06**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**	**2.77**	**2.68**	**2.57**	**2.49**	**2.41**	**2.32**	**2.27**	**2.19**	**2.15**	**2.10**	**2.06**	**2.03**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**	**2.74**	**2.66**	**2.55**	**2.47**	**2.38**	**2.29**	**2.24**	**2.16**	**2.13**	**2.07**	**2.03**	**2.01**
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
	7.50	**5.34**	**4.46**	**3.97**	**3.66**	**3.42**	**3.25**	**3.12**	**3.01**	**2.94**	**2.86**	**2.80**	**2.70**	**2.62**	**2.51**	**2.42**	**2.34**	**2.25**	**2.20**	**2.12**	**2.08**	**2.02**	**1.98**	**1.96**
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**	**2.82**	**2.76**	**2.66**	**2.58**	**2.47**	**2.38**	**2.30**	**2.21**	**2.15**	**2.08**	**2.04**	**1.98**	**1.94**	**1.91**
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**	**2.62**	**2.54**	**2.43**	**2.35**	**2.26**	**2.17**	**2.12**	**2.04**	**2.00**	**1.94**	**1.90**	**1.87**
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2.75**	**2.69**	**2.59**	**2.51**	**2.40**	**2.32**	**2.22**	**2.14**	**2.08**	**2.00**	**1.97**	**1.90**	**1.86**	**1.84**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**	**2.56**	**2.49**	**2.37**	**2.29**	**2.20**	**2.11**	**2.05**	**1.97**	**1.94**	**1.88**	**1.84**	**1.81**
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**	**2.54**	**2.46**	**2.35**	**2.26**	**2.17**	**2.08**	**2.02**	**1.94**	**1.91**	**1.85**	**1.80**	**1.78**
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**	**2.52**	**2.44**	**2.32**	**2.24**	**2.15**	**2.06**	**2.00**	**1.92**	**1.88**	**1.82**	**1.78**	**1.75**
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**	**2.50**	**2.42**	**2.30**	**2.22**	**2.13**	**2.04**	**1.98**	**1.90**	**1.86**	**1.80**	**1.76**	**1.72**
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**	**2.48**	**2.40**	**2.28**	**2.20**	**2.11**	**2.02**	**1.96**	**1.88**	**1.84**	**1.78**	**1.73**	**1.70**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**	**2.46**	**2.39**	**2.26**	**2.18**	**2.10**	**2.00**	**1.94**	**1.86**	**1.82**	**1.76**	**1.71**	**1.68**
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41
	7.12	**5.01**	**4.16**	**3.68**	**3.37**	**3.15**	**2.98**	**2.85**	**2.75**	**2.66**	**2.59**	**2.53**	**2.43**	**2.35**	**2.23**	**2.15**	**2.06**	**1.96**	**1.90**	**1.82**	**1.78**	**1.71**	**1.66**	**1.64**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**	**2.40**	**2.32**	**2.20**	**2.12**	**2.03**	**1.93**	**1.87**	**1.79**	**1.74**	**1.68**	**1.63**	**1.60**
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37
	7.04	**4.95**	**4.10**	**3.62**	**3.31**	**3.09**	**2.93**	**2.79**	**2.70**	**2.61**	**2.54**	**2.47**	**2.37**	**2.30**	**2.18**	**2.09**	**2.00**	**1.90**	**1.84**	**1.76**	**1.71**	**1.64**	**1.60**	**1.56**

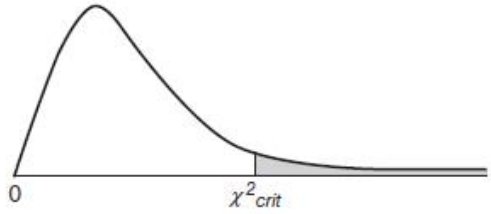
Table C[a] (Continued)
CRITICAL VALUES OF *F*

FINDING *p*-VALUE
If observed *F* is
... smaller than light number, $p > .05$
... between light and dark numbers, $p < .05$
... larger than dark number, $p < .01$

DEGREES OF FREEDOM IN NUMERATOR

DEGREES OF FREEDOM IN DENOMINATOR	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**	**2.35**	**2.28**	**2.15**	**2.07**	**1.98**	**1.88**	**1.82**	**1.74**	**1.69**	**1.62**	**1.56**	**1.53**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**	**2.32**	**2.24**	**2.11**	**2.03**	**1.94**	**1.84**	**1.78**	**1.70**	**1.65**	**1.57**	**1.52**	**1.49**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**	**2.26**	**2.19**	**2.06**	**1.98**	**1.89**	**1.79**	**1.73**	**1.64**	**1.59**	**1.51**	**1.46**	**1.43**
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.79**	**2.65**	**2.56**	**2.47**	**2.40**	**2.33**	**2.23**	**2.15**	**2.03**	**1.94**	**1.85**	**1.75**	**1.68**	**1.59**	**1.54**	**1.46**	**1.40**	**1.37**
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22
	6.81	**4.75**	**3.91**	**3.44**	**3.14**	**2.92**	**2.76**	**2.62**	**2.53**	**2.44**	**2.37**	**2.30**	**2.20**	**2.12**	**2.00**	**1.91**	**1.83**	**1.72**	**1.66**	**1.56**	**1.51**	**1.43**	**1.37**	**1.33**
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.90**	**2.73**	**2.60**	**2.50**	**2.41**	**2.34**	**2.28**	**2.17**	**2.09**	**1.97**	**1.88**	**1.79**	**1.69**	**1.62**	**1.53**	**1.48**	**1.39**	**1.33**	**1.28**
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
	6.70	**4.66**	**3.83**	**3.36**	**3.06**	**2.85**	**2.69**	**2.55**	**2.46**	**2.37**	**2.29**	**2.23**	**2.12**	**2.04**	**1.92**	**1.84**	**1.74**	**1.64**	**1.57**	**1.47**	**1.42**	**1.32**	**1.24**	**1.19**
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
	6.66	**4.62**	**3.80**	**3.34**	**3.04**	**2.82**	**2.66**	**2.53**	**2.43**	**2.34**	**2.26**	**2.20**	**2.09**	**2.01**	**1.89**	**1.81**	**1.71**	**1.61**	**1.54**	**1.44**	**1.38**	**1.28**	**1.19**	**1.11**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.86	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
	6.64	**4.00**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**	**1.79**	**1.69**	**1.59**	**1.52**	**1.41**	**1.36**	**1.25**	**1.15**	**1.00**

CRITICAL VALUES OF χ^2



LEVEL OF SIGNIFICANCE

	$p > .10$	$p < .10$	$p < .05$	$p < .01$	$p < .001$
df	**.10**	**.05**	**.01**	**.001**
1	2.71	3.84	6.64	10.83
2	4.60	5.99	9.21	13.82
3	6.25	7.81	11.34	16.27
4	7.78	9.49	13.28	18.47
5	9.24	11.07	15.09	20.52
6	10.64	12.59	16.81	22.46
7	12.02	14.07	18.48	24.32
8	13.36	15.51	20.09	26.12
9	14.68	16.92	21.67	27.88
10	15.99	18.31	23.21	29.59
11	17.28	19.68	24.72	31.26
12	18.55	21.03	26.22	32.91
13	19.81	22.36	27.69	34.53
14	21.06	23.68	29.14	36.12
15	22.31	25.00	30.58	37.70
16	23.54	26.30	32.00	39.25
17	24.77	27.59	33.41	40.79
18	25.99	28.87	34.80	42.31
19	27.20	30.14	36.19	43.82
20	28.41	31.41	37.57	45.32
21	29.62	32.67	38.93	46.80
22	30.81	33.92	40.29	48.27
23	32.01	35.17	41.64	49.73
24	33.20	36.42	42.98	51.18
25	34.38	37.65	44.31	52.62
26	35.56	38.88	45.64	54.05
27	36.74	40.11	46.96	55.48
28	37.92	41.34	48.28	56.89
29	39.09	42.56	49.59	58.30
30	40.26	43.77	50.89	59.70
40	51.80	55.76	63.69	73.40
50	63.17	67.50	76.15	86.66
60	74.40	79.08	88.38	99.61
70	85.53	90.53	100.42	112.32

UNIT V – PREDICTIVE ANALYTICS

SYLLABUS:

Linear least squares – implementation – goodness of fit – testing a linear model – weighted resampling. Regression using StatsModels – multiple regression – nonlinear relationships – logistic regression – estimating parameters – Time series analysis – moving averages – missing values – serial correlation – autocorrelation. Introduction to survival analysis.

PART A

1. Define predictive analytics.

- Predictive analytics is the process of using data to forecast future outcomes.
- The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior.
- Data scientists use historical data as their source and utilize various regression models and machine learning techniques to detect patterns and trends in the data.

2. List the Steps in Predictive Analytics

1. Define the problem
2. Acquire and organize data
3. Pre-process data
4. Develop predictive models
5. Validate and deploy results

3. What are the Predictive Analytics Techniques available? List the techniques used for Predictive Analytics.

1. Regression analysis
2. Decision trees
3. Neural networks

4. List the uses and examples of predictive analytics

- Fraud detection
- Conversion and purchase prediction
- Risk reduction
- Operational improvement
- Customer segmentation
- Maintenance forecasting

5. Define Least squares fit

- A "linear fit" is a line intended to model the relationship between variables.
- A "least squares" fit is one that minimizes the mean squared error (MSE) between the line and the data.

6. Define Residuals

- The deviation of an actual value from a model.
- The difference between the actual values and the fitted line.
- thinkstats2 provides a function that computes residuals:

```
def Residuals(xs, ys, inter, slope):
xs = np.asarray(xs)
ys = np.asarray(ys)
res = ys - (inter + slope * xs)
return res
```

It returns the differences between the actual values and the fitted line.

7. What is Goodness of fit in predictive analytics?

Goodness of fit

- A goodness-of-fit is a statistical test that tries to determine whether a set of observed values match those expected under the applicable model.
- They can show whether your sample data fit an expected set of data from a population with normal distribution.

8. Mention the types of goodness-of-fit tests

- The **chi-square test** determines if a relationship exists between categorical data.
  Variables must be mutually exclusive in order to qualify for the chi-square test for independence. And the chi goodness-of-fit test should not be used for data that is continuous.
- The **Kolmogorov-Smirnov test** determines whether a sample comes from a specific distribution of a population.

9. What are the different ways to measure the quality of a linear model, or goodness of fit?

- Standard deviation of the residuals
- Coefficient of determination, usually denoted R^2 and called "R-squared":

```
def CoefDetermination(ys, res):
return 1 - Var(res) / Var(ys)
```

Var(res) is the MSE of guesses using the model, Var(ys) is the MSE without it.

10. Differentiate Goodness-of-Fit Test vs. Independence Test

- Goodness-of-fit test and independence test are both statistical tests used to assess the relationship between variables.
- A goodness-of-fit test is used to evaluate how well a set of observed data fits a particular probability distribution.
- An independence test is used to assess the relationship between two variables. It is used to test whether there is any association between two variables.
- The primary purpose of an independence test is to see whether a change in one variable is related to a change in another variable.
- An independence test is pointed towards two specific variables. A goodness-of-fit test is used on an entire set of observed data to evaluate the appropriateness of a specific model.

11. Define Regression and list its types.

Regression

- The linear least squares fit is an example of regression, which is fitting any kind of model to any kind of data.
- The goal of regression analysis is to describe the relationship between one set of variables, called the dependent variables, and another set of variables, called independent or explanatory variables.
- When there is only one dependent and one explanatory variable, that's **simple regression**.
- If there is more than one dependent variable with more than one explanatory variable, that's **multivariate regression**.
- If the relationship between the dependent and explanatory variable is linear, that's **linear regression**.

12. Define StatsModels and mention its purpose.

- statsmodels provides two interfaces (APIs); the "formula" API uses strings to identify the dependent and explanatory variables.
  It uses a syntax called patsy; in this example, the ~ operator separates the dependent variable on the left from the explanatory variables on the right.
- smf.ols takes the formula string and the DataFrame, live, and returns an OLS object that represents the model.
  The name ols stands for "**ordinary least squares**."
  Given a sequence of values for y and sequences for x1 and x2, find the parameters, β_0, β_1, and β_2, that minimize the sum of ε^2. This process is called ordinary least squares.

13. How to implement Regression in Python using StatsModels?

1. Step 1: Import packages.
2. Step 2: Loading data.
3. Step 3: Setting a hypothesis.
4. Step 4: Fitting the model
5. Step 5: Summary of the model.

14. Define R- squared value, F- statistic and Predictions.

R- squared value

- R-squared value ranges between 0 and 1.
- An R-squared of 100 percent indicates that all changes in the dependent variable are completely explained by changes in the independent variable(s).

F- statistic:

- The F statistic simply compares the combined effect of all variables.

Predictions:

- If significance level (alpha) to be 0.05, reject the null hypothesis and accept the alternative hypothesis as p<0.05. so, say that there is a relationship between head size and brain weight.

15. Define multiple linear regression or Multiple Regression using Statsmodels in Python.

Multiple linear regression (MLR)

- Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.
- MLR is used extensively in econometrics and financial inference.

Formula and Calculation of Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

16. Define logistic regression.

- If the dependent variable is boolean, the generalized model is called logistic regression.
- Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation.
- The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.
- Logistic regression is commonly used in binary classification problems where the outcome variable reveals either of the two categories (0 and 1).

17. Define Sigmoid Function

- Logistic regression uses a logistic function called a sigmoid function to map predictions and their probabilities. Refer figure 5.3 for Sigmoid function.
- The sigmoid function refers to an S-shaped curve that converts any real value to a range between 0 and 1.
- If the output of the sigmoid function (estimated probability) is greater than a predefined threshold on the graph, the model predicts that the instance belongs to that class.
- If the estimated probability is less than the predefined threshold, the model predicts that the instance does not belong to the class.

The sigmoid function is referred to as an activation function for logistic regression and is defined as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

where,

e = base of natural logarithms

value = numerical value one wishes to transform

18. List the types of Logistic Regression with Examples

Logistic regression is classified into binary, multinomial, and ordinal.

Binary logistic regression

- Binary logistic regression predicts the relationship between the independent and binary dependent variables.
- Some examples of the output of this regression type may be, success/failure, 0/1, or true/false.

Examples:

1. Deciding on whether or not to offer a loan to a bank customer: Outcome = yes or no.
2. Evaluating the risk of cancer: Outcome = high or low.

3. Predicting a team's win in a football match: Outcome = yes or no.

Multinomial logistic regression

- A categorical dependent variable has two or more discrete outcomes in a multinomial regression type.
  - This implies that this regression type has more than two possible outcomes.

Ordinal logistic regression

- Ordinal logistic regression applies when the dependent variable is in an ordered state (i.e., ordinal). The dependent variable (y) specifies an order with two or more categories or levels.

19. Define time series and time series analysis.

➢ **Time Series**

- A time series is a sequence of measurements from a system that varies in time.
- An ordered sequence of values of a variable at equally spaced time intervals.

➢ **Time Series Analysis**

- Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time.
- In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly.
- Time series analysis has become a crucial tool for companies looking to make better decisions based on data.

20. Mention the components of Time Series Data

- **Trends:** Long-term increases, decreases, or stationary movement
- **Seasonality:** Predictable patterns at fixed intervals
- **Cycles:** Fluctuations without a consistent period
- **Noise:** Residual unexplained variability

21. List the different types of data used for predictive analysis.

➢ **Types of Data**

- **Time Series Data:** Comprises observations collected at different time intervals. It's geared towards analyzing trends, cycles, and other temporal patterns.
- **Cross-Sectional Data:** Involves data points collected at a single moment in time. Useful for understanding relationships or comparisons between different entities or categories at that specific point.

Pooled Data: A combination of Time Series and Cross-Sectional data. This hybrid enriches the dataset, allowing for more nuanced and comprehensive analyses.

22. Mention the different types of time series analysis.

- **Time Series Analysis Types**
  - Classification
  - Curve fitting
  - Descriptive analysis
  - Explanative analysis
  - Exploratory analysis
  - Forecasting
  - Intervention analysis
  - Segmentation.

23. List the Time Series Analysis Techniques

- Moving Average
- Exponential Smoothing
- Autoregression
- Decomposition
- Time Series Clustering
- Wavelet Analysis
- Intervention Analysis
- Box-Jenkins ARIMA models
- Box-Jenkins Multivariate models
- Holt-Winters Exponential Smoothing

24. List the Advantages of Time Series Analysis

1. Data Cleansing
2. Understanding Data
3. Forecasting
4. Identifying Trends and Seasonality
5. Visualizations
6. Efficiency
7. Risk Assessment

25. List the Challenges of Time Series Analysis

1. Limited Scope
2. Noise Introduction
3. Interpretation Challenges
4. Generalization Issues

5. Model Complexity
6. Non-Independence of Data
7. Data Availability

26. Define Serial Correlation and Auto Correlation in Time Series Analysis.

Serial Correlation

- Serial correlation is the relationship between a given variable and a lagged version of itself over various time intervals.
- It measures the relationship between a variable's current value given its past values.
- A variable that is serially correlated indicates that it may not be random.
- Serial correlation occurs in a time series when a variable and a lagged version of itself (for instance a variable at times T and at T-1) are observed to be correlated with one another over periods of time.
- **lag**: The size of the shift the time series by an interval in a serial correlation or autocorrelation.

Autocorrelation

- Autocorrelation, refers to the degree of correlation of the same variables between two successive time intervals.
- Autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals.
- Autocorrelation measures the relationship between a variable's current value and its past values.

27. Define Survival Analysis and Survival Curve.

Survival Analysis

- Survival analysis is a field of statistics that focuses on analysing the expected time until a certain event happens.
- Survival analysis can be used for analysing the results of that treatment in terms of the patients' life expectancy.
- The term `survival time' specifies the length of time taken for failure to occur.

Survival curves

- The fundamental concept in survival analysis is the survival curve, S(t), which is a function that maps from a duration, t, to the probability of surviving longer than t, it's just the complement of the CDF:

$$S(t) = 1 - CDF(t)$$

where CDF(t) is the probability of a lifetime less than or equal to t.

28. Define missing value and narrate the reason for missing value.

Missing Value

- Missing data is defined as the values or data that is not stored for some variable/s in the given dataset.

Reason for Missing Values

- Past data might get corrupted due to improper maintenance.
- Observations are not recorded for certain fields due to some reasons. There might be a failure in recording the values due to human error.
- The user has not provided the values intentionally
- Item nonresponse: This means the participant refused to respond.

29. Why the missing data should be handled?

- The missing data will decrease the predictive power of the model. If the algorithms are applied with missing data, then there will be bias in the estimation of parameters.
- The results are not confident if the missing data is not handled properly.

30. List the types of Missing Values

Type	*Definition*
Missing completely at random (MCAR)	Missing data are randomly distributed across the variable and unrelated to other variables.
Missing at random (MAR)	Missing data are not randomly distributed but they are accounted for by other observed variables.
Missing not at random (MNAR)	Missing data systematically differ from the observed values.

31. List the methods for identifying missing data

Functions	Descriptions
.isnull()	This function returns a pandas dataframe, where each value is a boolean value True if the value is missing, False otherwise.
.notnull()	Similarly to the previous function, the values for this one are False if either NaN or None value is detected.
.info()	This function generates three main columns, including the "Non-Null Count" which shows the number of non-missing values for each column.
.isna()	This one is similar to isnull and notnull. However it shows True only when the missing value is NaN type.

PART B

1. Give a brief introduction about predictive analytics.

- **Predictive analytics**
  - Predictive analytics is the process of using data to forecast future outcomes.
  - The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior.
  - Data scientists use historical data as their source and utilize various regression models and machine learning techniques to detect patterns and trends in the data.

- **Steps in Predictive Analytics**
  - The workflow for building predictive analytics frameworks follows five basic steps:

1. **Define the problem:**
   - A prediction starts with a good thesis and set of requirements.
   - A distinct problem to solve will help determine what method of predictive analytics should be used.
2. **Acquire and organize data:**
   - An organization may have decades of data to draw upon, or a continual flood of data from customer interactions.
   - Before predictive analytics models can be developed, data flows must be identified, and then datasets can be organized in a repository such as a data warehouse like BigQuery.
3. **Pre-process data:**
   - To prepare the data for the predictive analytics models, it should be cleaned to remove anomalies, missing data points, or extreme outliers, any of which might be the result of input or measurement errors.
4. **Develop predictive models:**
   - Data scientists have a variety of tools and techniques to develop predictive models depending on the problem to be solved and nature of the dataset.
   - Machine learning, regression models, and decision trees are some of the most common types of predictive models.
5. **Validate and deploy results:**
   - Check on the accuracy of the model and adjust accordingly.
   - Once acceptable results have been achieved, make them available to stakeholders via an app, website, or data dashboard.

- **Predictive Analytics Techniques**

Predictive analytics tends to be performed with three main types of techniques:

1. **Regression analysis**
   - Regression is a statistical analysis technique that estimates relationships between variables.
   - Regression is useful to determine patterns in large datasets to determine the correlation between inputs.
   - Regression is often used to determine how one or more independent variables affects another, such as how a price increase will affect the sale of a product.
2. **Decision trees**
   - Decision trees are classification models that place data into different categories based on distinct variables.
   - The model looks like a tree, with each branch representing a potential choice, with the leaf of the branch representing the result of the decision.
3. **Neural networks**
   - Neural networks are machine learning methods that are useful in predictive analytics when modeling very complex relationships.
   - Neural networks are best used to determine nonlinear relationships in datasets, especially when no known mathematical formula exists to analyze the data.
   - Neural networks can be used to validate the results of decision trees and regression models.

- **Uses and examples of predictive analytics**
  - Predictive analytics can be used to streamline operations, boost revenue, and mitigate risk for almost any business or industry, including banking, retail, utilities, public sector, healthcare, and manufacturing.

Fraud detection

- Predictive analytics examines all actions on a company's network in real time to pinpoint abnormalities that indicate fraud and other vulnerabilities.

Conversion and purchase prediction

- Companies can take actions, like retargeting online ads to visitors, with data that predicts a greater likelihood of conversion and purchase intent.

Risk reduction

- Credit scores, insurance claims, and debt collections all use predictive analytics to assess and determine the likelihood of future defaults.

Operational improvement

- Companies use predictive analytics models to forecast inventory, manage resources, and operate more efficiently.

Customer segmentation

- By dividing a customer base into specific groups, marketers can use predictive analytics to make forward-looking decisions to tailor content to unique audiences.

Maintenance forecasting

- Organizations use data to predict when routine equipment maintenance will be required and can then schedule it before a problem or malfunction arises.

2. Explain linear least squares and its implementation in detail.

Least squares fit

- A "linear fit" is a line intended to model the relationship between variables.
- A "least squares" fit is one that minimizes the mean squared error (MSE) between the line and the data.
- The more general problem is that of fitting a straight line to a collection of pairs of observations (x, y)

$$y = \beta_0 + \beta_1 x,$$

- The most commonly used method for finding a model is that of least squares estimation.
- It is supposed that x is an independent (or predictor) variable which is known exactly, while y is a dependent (or response) variable.
- The least squares (LS) estimates for β_0 and β_1 are those for which the predicted values of the curve minimize the sum of the squared deviations from the observations.
- That is the problem is to find the values of β_0 and β_1 that minimize the residual sum of squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Implementation

thinkstats2 provides simple functions that demonstrate linear least squares:

```
def LeastSquares(xs, ys):
meanx, varx = MeanVar(xs)
meany = Mean(ys)
slope = Cov(xs, ys, meanx, meany) / varx
inter = meany - slope * meanx
return inter, slope
```

LeastSquares takes sequences xs and ys and returns the estimated parameters inter and slope.

thinkstats2 also provides FitLine, which takes inter and slope and returns the fitted line for a sequence of xs.


```
def FitLine(xs, inter, slope):
fit_xs = np.sort(xs)
fit_ys = inter + slope * fit_xs
return fit_xs, fit_ys
```

Residuals

- The deviation of an actual value from a model.
- The difference between the actual values and the fitted line.
- thinkstats2 provides a function that computes residuals:

```
def Residuals(xs, ys, inter, slope):
xs = np.asarray(xs)
ys = np.asarray(ys)
res = ys - (inter + slope * xs)
return res
```

Residuals takes sequences xs and ys and estimated parameters inter and slope. It returns the differences between the actual values and the fitted line.

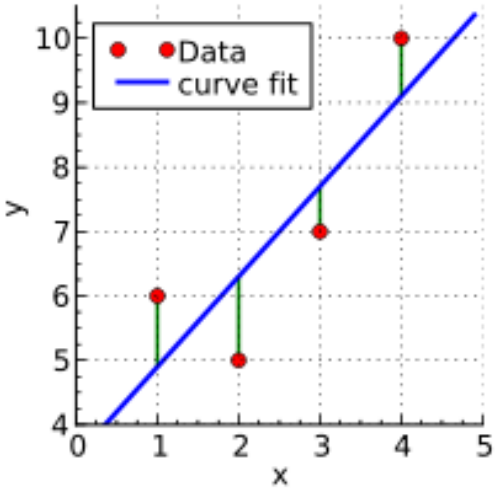


Figure 5.1 – Linear Least Square

- A plot in the figure 5.1 depicts the data points (in red), the least squares line of best fit (in blue), and the residuals (in green)
- The parameters slope and inter are estimates based on a sample; like other estimates, they are vulnerable to sampling bias, measurement error, and sampling error.
- Sampling bias is caused by non-representative sampling, measurement error is caused by errors in collecting and recording data, and sampling error is the result of measuring a sample rather than the entire population.

3. Explain in detail about Goodness of fit.

Goodness of fit

- A goodness-of-fit is a statistical test that tries to determine whether a set of observed values match those expected under the applicable model.
- They can show whether sample data fit an expected set of data from a population with normal distribution.

Types of goodness-of-fit tests

- The **chi-square test** determines if a relationship exists between categorical data.

  Variables must be mutually exclusive in order to qualify for the chi-square test for independence. And the chi goodness-of-fit test should not be used for data that is continuous.

  Goodness of fit is a measure of how well a statistical model fits a set of observations.

  - When goodness of fit is high, the values expected based on the model are close to the observed values.
  - When goodness of fit is low, the values expected based on the model are far from the observed values.
- The **Kolmogorov-Smirnov test** determines whether a sample comes from a specific distribution of a population.

To conduct the test, need a certain variable, along with an assumption of how it is distributed.

- The **observed values**, which are derived from the actual data set
- The **expected values**, which are taken from the assumptions made
- The **total number of categories** in the set

Ways to measure the quality of a linear model, or goodness of fit.

- **Standard deviation of the residuals** - Std(res) is the root mean squared error (RMSE) of predictions.
- **Coefficient of determination**, usually denoted R^2 and called "R-squared":

```
def CoefDetermination(ys, res):
    return 1 - Var(res) / Var(ys)
```

Var(res) is the MSE of guesses using the model, Var(ys) is the MSE without it.

Importance of Goodness-of-Fit Tests

- Provide a way to assess how well a statistical model fits a set of observed data.
- To determine whether the observed data are consistent with the assumed statistical model
- Useful in choosing between different models which may better fit the data.
- Help to identify outliers or market abnormalities that may be affecting the fit of the model
- Provide information about the variability of the data and the estimated parameters of the model.
- Can be useful for making predictions and understanding the behavior of the system being modeled.

Goodness-of-Fit Test vs. Independence Test

- Goodness-of-fit test and independence test are both statistical tests used to assess the relationship between variables.
- A goodness-of-fit test is used to evaluate how well a set of observed data fits a particular probability distribution.

- An independence test is used to assess the relationship between two variables. It is used to test whether there is any association between two variables.
- The primary purpose of an independence test is to see whether a change in one variable is related to a change in another variable.
- An independence test is pointed towards two specific variables. A goodness-of-fit test is used on an entire set of observed data to evaluate the appropriateness of a specific model.

4. Discuss in detail about Regression using StatsModels.

Regression

- The linear least squares fit is an example of regression, which is fitting any kind of model to any kind of data.
- The goal of regression analysis is to describe the relationship between one set of variables, called the dependent variables, and another set of variables, called independent or explanatory variables.
- When there is only one dependent and one explanatory variable, that's simple regression.
- If there is more than one dependent variable with more than one explanatory variable, that's multivariate regression.
- If the relationship between the dependent and explanatory variable is linear, that's linear regression.
- For example, if the dependent variable is y and the explanatory variables are x1 and x2, linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where β_0 is the intercept, β_1 is the parameter associated with x1, β_2 is the parameter associated with x2, and ε is the residual.

StatsModels

- statsmodels provides two interfaces (APIs); the "formula" API uses strings to identify the dependent and explanatory variables.
  It uses a syntax called patsy; in this example, the ~ operator separates the dependent variable on the left from the explanatory variables on the right.
- smf.ols takes the formula string and the DataFrame, live, and returns an OLS object that represents the model.

The name ols stands for "**ordinary least squares**."
Given a sequence of values for y and sequences for x1 and x2, find the parameters, β_0, β_1, and β_2, that minimize the sum of ε_2. This process is called ordinary least squares.

- The fit method fits the model to the data and returns a RegressionResults object that contains the results.

Stepwise Implementation in Python

Step 1: Import packages.

Step 2: Loading data.

Step 3: Setting a hypothesis.

Step 4: Fitting the model

statsmodels.regression.linear_model.OLS() method is used to get ordinary least squares, and fit() method is used to fit the data in it. The ols method takes in the data and performs linear regression.

inpendent_columns ~ dependent_column:

left side of the ~ operator contains the independent variables and right side of the operator contains the name of the dependent variable or the predicted column.

Step 5: Summary of the model.

All the summary statistics of the linear regression model are returned by the model.summary() method.

Example Program

```
# import packages
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf

# loading the csv file
df = pd.read_csv('headbrain1.csv')
print(df.head())

# fitting the model
df.columns = ['Head_size', 'Brain_weight']
model = smf.ols(formula='Head_size ~ Brain_weight', data=df).fit()

# model summary
print(model.summary())
```

Output

OLS Regression Results

Dep. Variable:	GRADE	R-squared:	0.416
Model:	OLS	Adj. R-squared:	0.353
Method:	Least Squares	F-statistic:	6.646
Date:	Thu, 14 Dec 2023	Prob (F-statistic):	0.00157
Time:	14:55:37	Log-Likelihood:	-12.978
No. Observations:	32	AIC:	33.96
Df Residuals:	28	BIC:	39.82
Df Model:	3		
Covariance Type:	nonrobust		

R- squared value:

- R-squared value ranges between 0 and 1.
- An R-squared of 100 percent indicates that all changes in the dependent variable are completely explained by changes in the independent variable(s).

F- statistic:

- The F statistic simply compares the combined effect of all variables.

Predictions:

- If significance level (alpha) to be 0.05, reject the null hypothesis and accept the alternative hypothesis as $p<0.05$. so, say that there is a relationship between head size and brain weight.

5. Discuss in detail about multiple linear regression or Multiple Regression using Statsmodels in Python.

Multiple linear regression (MLR)

- Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.
- MLR is used extensively in econometrics and financial inference.

Formula and Calculation of Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

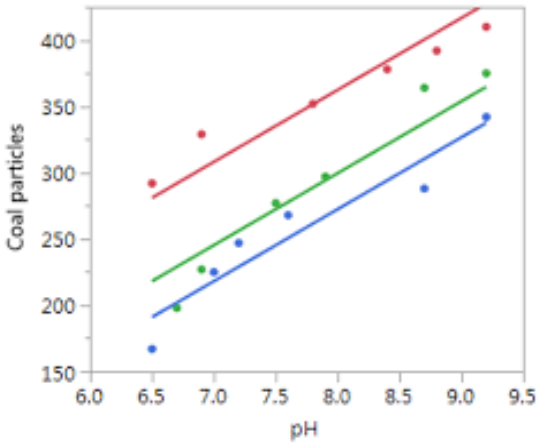
y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)



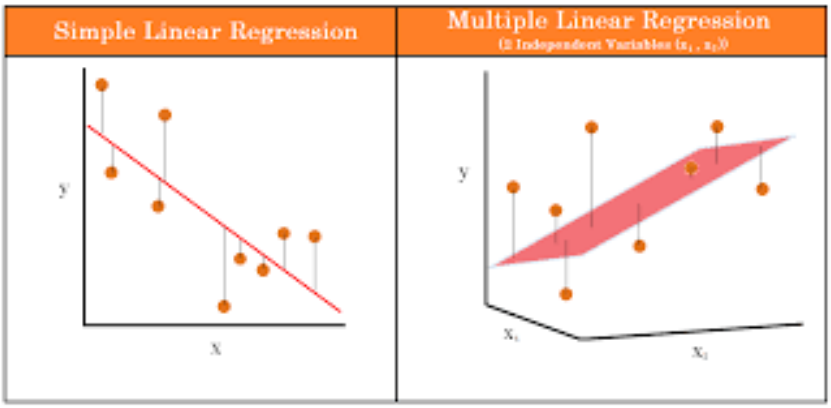


Figure 5.2 – Simple Linear Regression Vs Multiple Linear Regression

Example

```
import statsmodels.api as sm
X = advertising[['TV','Newspaper','Radio']]
y = advertising['Sales']

# Add a constant to get an intercept
X_train_sm = sm.add_constant(X_train)

# Fit the resgression line using 'OLS'
lr = sm.OLS(y_train, X_train_sm).fit()
print(lr.summary())
```

Output

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.910
Model:	OLS	Adj. R-squared:	0.909
Method:	Least Squares	F-statistic:	461.2
Date:	Wed, 29 Sep 2021	Prob (F-statistic):	4.73e-71
Time:	19:09:53	Log-Likelihood:	-270.60
No. Observations:	140	AIC:	549.2
Df Residuals:	136	BIC:	561.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.3346	0.357	12.139	0.000	3.628	5.041
TV	0.0538	0.002	34.539	0.000	0.051	0.057
Newspaper	0.0063	0.007	0.902	0.369	-0.008	0.020
Radio	0.1100	0.010	10.609	0.000	0.090	0.131

Omnibus:	18.669	Durbin-Watson:	2.069
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31.404
Skew:	-0.643	Prob(JB):	1.52e-07
Kurtosis:	4.932	Cond. No.	443.

Understanding the results:

- Rsq value is 91% which is good. It means that the degree of variance in Y variable is explained by X variables
- Adj Rsq value is also good although it penalizes predictors more than Rsq
- After looking at the p values we can see that 'newspaper' is not a significant X variable since p value is greater than 0.05
- The coef values are good as they fall in 5% and 95%, except for the newspaper variable.

6. Discuss in detail about logistic regression with suitable case study.

LOGISTIC REGRESSION

- Linear regression can be generalized to handle other kinds of dependent variables.
- If the dependent variable is boolean, the generalized model is called logistic regression.
- If the dependent variable is an integer count, it's called Poisson regression.
- Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation.
- The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.
- Logistic regression is commonly used in binary classification problems where the outcome variable reveals either of the two categories (0 and 1).

Example:

1. **Determine the probability of heart attacks**:
   With the help of a logistic model, medical practitioners can determine the relationship between variables such as the weight, exercise, etc., of an individual and use it to predict whether the person will suffer from a heart attack or any other medical complication.
2. **Identifying spam emails**:
   Email inboxes are filtered to determine if the email communication is promotional/spam by understanding the predictor variables and applying a logistic regression algorithm to check its authenticity.

Sigmoid Function

- Logistic regression uses a logistic function called a sigmoid function to map predictions and their probabilities. Refer figure 5.3 for Sigmoid function.
- The sigmoid function refers to an S-shaped curve that converts any real value to a range between 0 and 1.
- If the output of the sigmoid function (estimated probability) is greater than a predefined threshold on the graph, the model predicts that the instance belongs to that class.
- If the estimated probability is less than the predefined threshold, the model predicts that the instance does not belong to the class.
- The sigmoid function is referred to as an activation function for logistic regression and is defined as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

where,

e = base of natural logarithms
value = numerical value one wishes to transform

The following equation represents logistic regression:

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

- x = input value
- y = predicted output
- b0 = bias or intercept term
- b1 = coefficient for input (x)

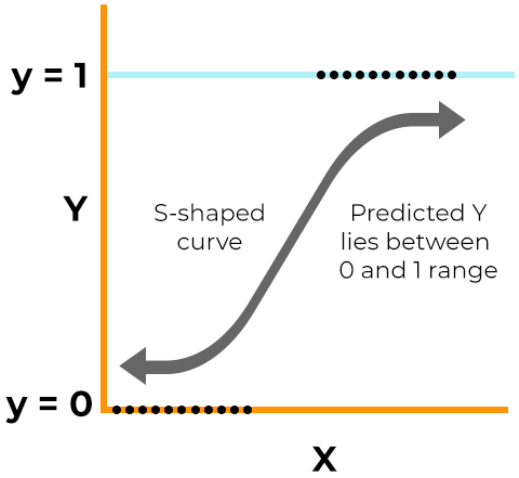


Figure 5.3 – Sigmoid Function

Key Assumptions for implementing Logistic Regression

1. The dependent/response variable is binary or dichotomous

- The first assumption of logistic regression is that response variables can only take on two possible outcomes – pass/fail, male/female, and malignant/benign.

2. Little or no multicollinearity between the predictor/explanatory variables

- This assumption implies that the predictor variables (or the independent variables) should be independent of each other. Multicollinearity relates to two or more highly correlated independent variables.

3. Linear relationship of independent variables to log odds

- Log odds refer to the ways of expressing probabilities. Log odds are different from probabilities. Odds refer to the ratio of success to failure, while probability refers to the ratio of success to everything that can occur.
- For example, consider that you play twelve tennis games with your friend. Here, the odds of you winning are 5 to 7 (or 5/7), while the probability of you winning is 5 to 12 (as the total games played = 12).

4. Prefers large sample size

- Logistic regression analysis yields reliable, robust, and valid results when a larger sample size of the dataset is considered.

5. Problem with extreme outliers

- Another critical assumption of logistic regression is the requirement of no extreme outliers in the dataset.

6. Consider independent observations

- This assumption states that the dataset observations should be independent of each other. The observations should not be related to each other or emerge from repeated measurements of the same individual type.

Types of Logistic Regression with Examples

Logistic regression is classified into binary, multinomial, and ordinal.

Binary logistic regression

- Binary logistic regression predicts the relationship between the independent and binary dependent variables.
- Some examples of the output of this regression type may be, success/failure, 0/1, or true/false.

Examples:

4. Deciding on whether or not to offer a loan to a bank customer: Outcome = yes or no.
5. Evaluating the risk of cancer: Outcome = high or low.
6. Predicting a team's win in a football match: Outcome = yes or no.

Multinomial logistic regression

- A categorical dependent variable has two or more discrete outcomes in a multinomial regression type.
- This implies that this regression type has more than two possible outcomes.

Examples:

1. Let's say you want to predict the most popular transportation type for 2040. Here, transport type equates to the dependent variable, and the possible outcomes can be electric cars, electric trains, electric buses, and electric bikes.
2. Predicting whether a student will join a college, vocational/trade school, or corporate industry.
3. Estimating the type of food consumed by pets, the outcome may be wet food, dry food, or junk food.

Ordinal logistic regression

- Ordinal logistic regression applies when the dependent variable is in an ordered state (i.e., ordinal). The dependent variable (y) specifies an order with two or more categories or levels.

Examples: Dependent variables represent,

1. Formal shirt size: Outcomes = XS/S/M/L/XL
2. Survey answers: Outcomes = Agree/Disagree/Unsure
3. Scores on a math test: Outcomes = Poor/Average/Good

Logistic regression works in the following steps:

1. **Prepare the data: T**he data should be in a format where each row represents a single observation and each column represents a different variable. The target variable (the variable you want to predict) should be binary (yes/no, true/false, 0/1).
2. **Train the model:** We teach the model by showing it the training data. This involves finding the values of the model parameters that minimize the error in the training data.
3. **Evaluate the model:** The model is evaluated on the held-out test data to assess its performance on unseen data.
4. **Use the model to make predictions:** After the model has been trained and assessed, it can be used to forecast outcomes on new data.

ESTIMATING PARAMETERS

Given a probability, compute the odds like this:

```
o = p / (1-p)
```

Given odds in favor, convert to probability like this:

```
p = o / (o+1)
```


Logistic regression is based on the following model:

$$\log o = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Where o is the odds in favor of a particular outcome;

Suppose having estimated the parameters β_0, β_1, and β_2 And given values for x1 and x2 can compute the predicted value of log o, and then convert to a probability:

o = np.exp (log_o)

p = o / (o+1)

The usual goal is to find the maximum-likelihood estimate (MLE), which is the set of parameters that maximizes the likelihood of the data.

Example

Suppose the following data:

```
>>> y = np.array([0, 1, 0, 1])
>>> x1 = np.array([0, 0, 0, 1])
>>> x2 = np.array([0, 1, 1, 1])
```

And start with the initial guesses

$\beta_0 = -1.5,\ \beta_1 = 2.8,\ \text{and}\ \beta_2 = 1.1$:

```
>>> beta = [-1.5, 2.8, 1.1]
```

Then for each row can compute log_o:

```
>>> log_o = beta[0] + beta[1] * x1 + beta[2] * x2
[-1.5 -0.4 -0.4 2.4]
```

And convert from log odds to probabilities:

```
>>> o = np.exp(log_o)
[ 0.223 0.670 0.670 11.02 ]

>>> p = o / (o+1)
[ 0.182 0.401 0.401 0.916 ]
```


Notice that when log_o is greater than 0, o is greater than 1 and p is greater than 0.5.
The likelihood of an outcome is p when y==1 and 1-p when y==0.

If think the probability of a boy is 0.8 and the outcome is a boy, the likelihood is 0.8; if the outcome is a girl, the likelihood is 0.2.

Compute that like this:

```
>>> likes = y * p + (1-y) * (1-p)
[ 0.817 0.401 0.598 0.916 ]
```

The overall likelihood of the data is the product of likes:

```
>>> like = np.prod(likes)
0.18
```

For these values of beta, the likelihood of the data is 0.18. The goal of logistic regression is to find parameters that maximize this likelihood.

IMPLEMENTATION

StatsModels provides an implementation of logistic regression called logit, named for the function that converts from probability to log odds.

```
import statsmodels.formula.api as smf
model = smf.logit('boy ~ agepreg', data=df)
results = model.fit()
SummarizeResults(results)
```

The result is a Logit object that represents the model.
It contains attributes called endog and exog that contain the endogenous variable, another name for the dependent variable, and the exogenous variables, another name for the explanatory variables.
The result of model.fit is a BinaryResults object,

7. Discuss in detail about time series analysis with a suitable case study.

- **Time Series**
  - A time series is a sequence of measurements from a system that varies in time.
  - An ordered sequence of values of a variable at equally spaced time intervals.
- **Time Series Analysis**
  - Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time.
  - In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly.
  - Time series analysis has become a crucial tool for companies looking to make better decisions based on data.
- **Examples of time series analysis** in action include:
  - Weather data
  - Rainfall measurements
  - Temperature readings
  - Heart rate monitoring (EKG)
  - Brain monitoring (EEG)
  - Quarterly sales
  - Stock prices
  - Automated stock trading
  - Industry forecasts

- **Components of Time Series Data**
  - **Trends:** Long-term increases, decreases, or stationary movement
  - **Seasonality:** Predictable patterns at fixed intervals
  - **Cycles:** Fluctuations without a consistent period
  - **Noise:** Residual unexplained variability

- **Types of Data**
  - **Time Series Data:** Comprises observations collected at different time intervals. It's geared towards analyzing trends, cycles, and other temporal patterns.
  - **Cross-Sectional Data:** Involves data points collected at a single moment in time. Useful for understanding relationships or comparisons between different entities or categories at that specific point.

    Pooled Data: A combination of Time Series and Cross-Sectional data. This hybrid enriches the dataset, allowing for more nuanced and comprehensive analyses.

- **Time Series Analysis Types**
  - **Classification:** Identifies and assigns categories to the data.
  - **Curve fitting:** Plots the data along a curve to study the relationships of variables within the data.
  - **Descriptive analysis:** Identifies patterns in time series data, like trends, cycles, or seasonal variation.
  - **Explanative analysis:** Attempts to understand the data and the relationships within it, as well as cause and effect.
  - **Exploratory analysis:** Highlights the main characteristics of the time series data, usually in a visual format.
  - **Forecasting:** Predicts future data. This type is based on historical trends. It uses the historical data as a model for future data, predicting scenarios that could happen along future plot points.
  - **Intervention analysis:** Studies how an event can change the data.
  - **Segmentation:** Splits the data into segments to show the underlying properties of the source information.

- **Time Series Analysis Techniques**
  - **Moving Average**: Useful for smoothing out long-term trends. It is ideal for removing noise and identifying the general direction in which values are moving.

- **Exponential Smoothing**: Suited for univariate data with a systematic trend or seasonal component. Assigns higher weight to recent observations, allowing for more dynamic adjustments.
- **Autoregression**: Leverages past observations as inputs for a regression equation to predict future values. It is good for short-term forecasting when past data is a good indicator.
- **Decomposition**: This breaks down a time series into its core components—trend, seasonality, and residuals—to enhance the understanding and forecast accuracy.
- **Time Series Clustering**: Unsupervised method to categorize data points based on similarity, aiding in identifying archetypes or trends in sequential data.
- **Wavelet Analysis**: Effective for analyzing non-stationary time series data. It helps in identifying patterns across various scales or resolutions.
- **Intervention Analysis**: Assesses the impact of external events on a time series, such as the effect of a policy change or a marketing campaign.
- **Box-Jenkins ARIMA models**: Focuses on using past behavior and errors to model time series data. Assumes data can be characterized by a linear function of its past values.
- **Box-Jenkins Multivariate models**: Similar to ARIMA, but accounts for multiple variables. Useful when other variables influence one time series.
- **Holt-Winters Exponential Smoothing**: Best for data with a distinct trend and seasonality. Incorporates weighted averages and builds upon the equations for exponential smoothing.

➢ **The Advantages of Time Series Analysis**

1. **Data Cleansing**: Time series analysis techniques such as smoothing and seasonality adjustments help remove noise and outliers, making the data more reliable and interpretable.
2. **Understanding Data**: Models like ARIMA or exponential smoothing provide insight into the data's underlying structure. Autocorrelations and stationary measures can help understand the data's true nature.
3. **Forecasting**: One of the primary uses of time series analysis is to predict future values based on historical data. Forecasting is invaluable for business planning, stock market analysis, and other applications.
4. **Identifying Trends and Seasonality**: Time series analysis can uncover underlying patterns, trends, and seasonality in data that might not be apparent through simple observation.
5. **Visualizations**: Through time series decomposition and other techniques, it's possible to create meaningful visualizations that clearly show trends, cycles, and irregularities in the data.

6. **Efficiency**: With time series analysis, less data can sometimes be more. Focusing on critical metrics and periods can often derive valuable insights without getting bogged down in overly complex models or datasets.
7. **Risk Assessment**: Volatility and other risk factors can be modeled over time, aiding financial and operational decision-making processes.

- **Challenges of Time Series Analysis**

1. **Limited Scope**: Time series analysis is restricted to time-dependent data. It's not suitable for cross-sectional or purely categorical data.
2. **Noise Introduction**: Techniques like differencing can introduce additional noise into the data, which may obscure fundamental patterns or trends.
3. **Interpretation Challenges**: Some transformed or differenced values may need more intuitive meaning, making it easier to understand the real-world implications of the results.
4. **Generalization Issues**: Results may only sometimes be generalizable, primarily when the analysis is based on a single, isolated dataset or period.
5. **Model Complexity**: The choice of model can greatly influence the results, and selecting an inappropriate model can lead to unreliable or misleading conclusions.
6. **Non-Independence of Data**: Unlike other types of statistical analysis, time series data points are not always independent, which can introduce bias or error in the analysis.
7. **Data Availability**: Time series analysis often requires many data points for reliable results, and such data may not always be easily accessible or available.

8. Explain in detail about Time Series Analysis Technique – Moving Average and exponentially-weighted moving average with an example.

- **Moving Average**

- A moving average divides the series into overlapping regions, called windows, and computes the average of the values in each window.
- One of the simplest moving averages is the rolling mean, which computes the mean of the values in each window.
- For example, if the window size is 3, the rolling mean computes the mean of values 0 through 2, 1 through 3, 2 through 4, etc.
- pandas provides rolling_mean, which takes a Series and a window size and returns a new Series.

>>> series = np.arange(10)

array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])

>>> pandas.rolling_mean(series, 3)
array([nan, nan, 1, 2, 3, 4, 5, 6, 7, 8])

- The first two values are nan; the next value is the mean of the first three elements, 0, 1, and 2. The next value is the mean of 1, 2, and 3. And so on.
- The rolling mean seems to do smoothing out the noise and extracting the trend.

➢ **Exponentially-weighted moving average (EWMA)**

- The Exponentially Weighted Moving Average (EWMA) is a quantitative or statistical measure used to model or describe a time series.
- The moving average is designed as such that older observations are given lower weights. The weights fall exponentially as the data point gets older – hence the name exponentially weighted.
- An alternative is the exponentially-weighted moving average (EWMA), which has two advantages.
- First, it computes a weighted average where the most recent value has the highest weight and the weights for previous values drop off exponentially.
- Second, the pandas implementation of EWMA handles missing values better.

EWMA Formula

$$EWMA_t = \alpha * r_t + (1 - \alpha) * EWMA_{t-1}$$

Where:

Alpha = The weight decided by the user
r = Value of the series in the current period

ewma = pandas.ewma(reindexed.ppg, span=30)
thinkplot.Plot(ewma.index, ewma)

- The span parameter corresponds roughly to the window size of a moving average; it controls how fast the weights drop off, so it determines the number of points that make a non-negligible contribution to each average.

- Figure 5.1 (right) shows the EWMA for the same data.

- It is similar to the rolling mean, where they are both defined, but it has no missing values, which makes it easier to work with.

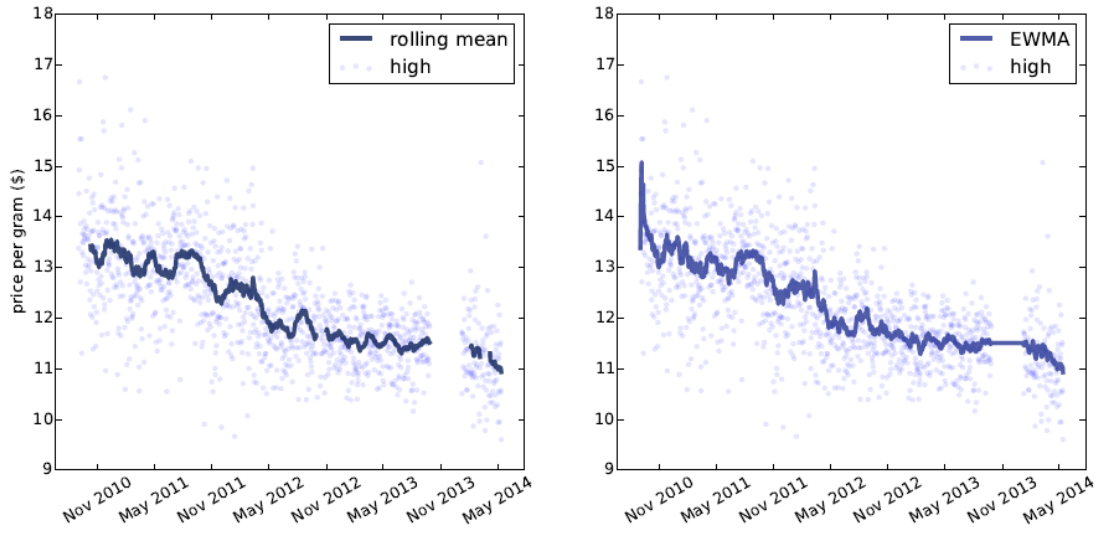


Figure 5.4: Daily price and a rolling mean (left) and exponentially-weighted moving average (right).

Missing values

- A simple and common way to fill missing data is to use a moving average.
- The Series method fillna:

reindexed.ppg.fillna(ewma, inplace=True)

Wherever reindexed.ppg is nan, fillna replaces it with the corresponding value from ewma. The inplace flag tells fillna to modify the existing Series rather than create a new one.

9. Discuss in detail about Serial Correlation and Auto Correlation in Time Series Analysis with suitable example.

Serial Correlation

- Serial correlation is the relationship between a given variable and a lagged version of itself over various time intervals.
- It measures the relationship between a variable's current value given its past values.
- A variable that is serially correlated indicates that it may not be random.
- Serial correlation occurs in a time series when a variable and a lagged version of itself (for instance a variable at times T and at T-1) are observed to be correlated with one another over periods of time.
- **lag**: The size of the shift the time series by an interval in a serial correlation or autocorrelation.

- One example of serial correlation is found in **stock prices.**
- Stock prices tend to go up and down together over time, which is said to be "serially correlated." This means that if stock prices go up today, they will also go up tomorrow. Similarly, if stock prices go down today, they are likely to go down tomorrow.
- The degree of serial correlation can be measured using the **autocorrelation coefficient.**
- The autocorrelation coefficient measures how closely related a series of data points are to each other.

Types of Serial Correlation

Positive Serial Correlation

- Positive serial correlation occurs when a positive error for one observation increases the chance of a positive error for another observation.
- In other words, if there is a positive error in one period, there is a greater likelihood of a positive error in the next period as well.
- Positive serial correlation also means that a negative error for one observation increases the chance of a negative error for another observation.
- So, if there is a negative error in one period, there is a greater likelihood of a negative error in the next period. Refer Figure 5.5

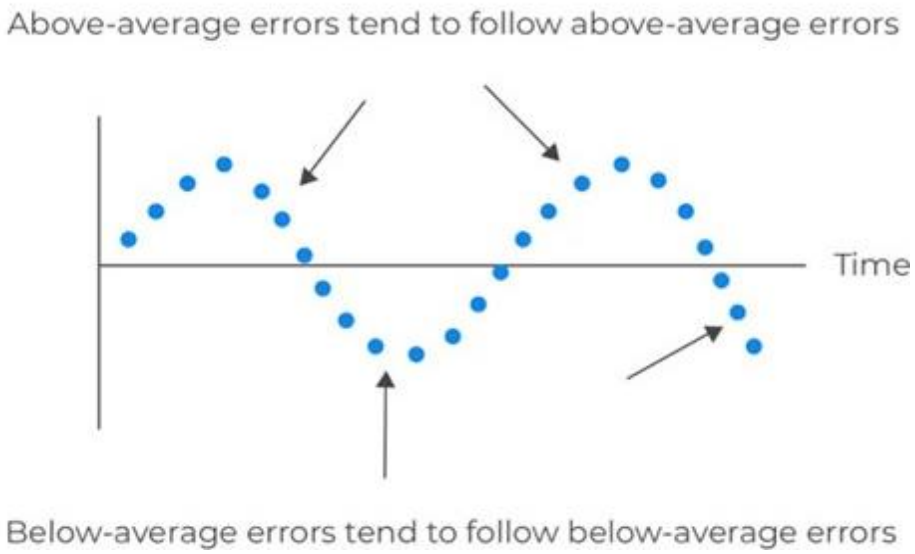


Figure 5.5 – Positive Serial Correlation

Negative Serial Correlation

- A negative serial correlation occurs when a positive error for one observation increases the chance of a negative error for another observation.

- In other words, if there is a positive error in one period, there is a greater likelihood of a negative error in the next period.
- A negative serial correlation also means that a negative error for one observation increases the chance of a positive error for another observation.
- So, if there is a negative error in one period, there is a greater likelihood of a positive error in the next period. Refer Figure 5.6

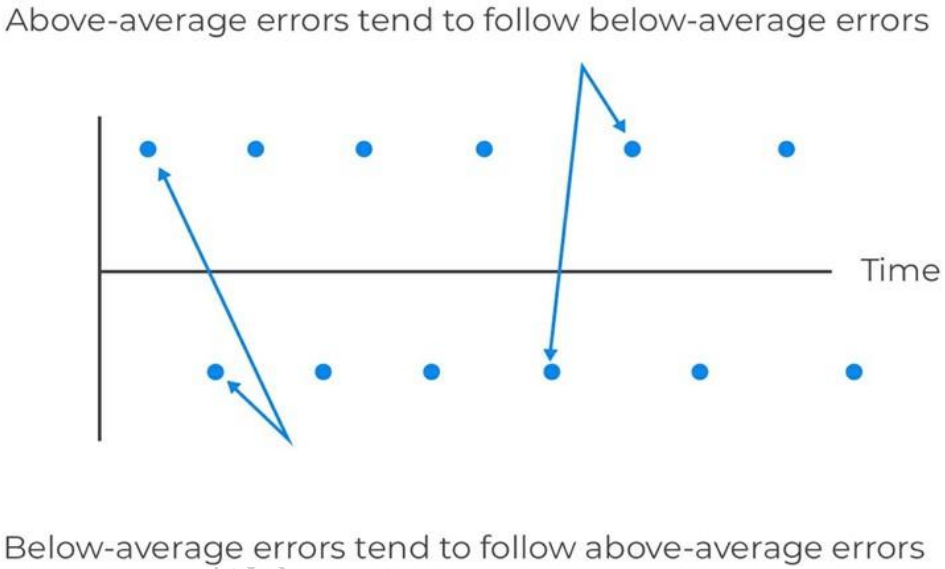


Figure 5.6 – Negative Serial Correlation


```
def SerialCorr(series, lag=1):
xs = series[lag:]
ys = series.shift(lag)[lag:]
corr = thinkstats2.Corr(xs, ys)
return corr
```

After the shift, the first lag values are nan, so I use a slice to remove them before computing Corr.

Testing for Serial Correlation

Durbin-Watson Test

- The Durbin-Watson test is a statistical test used to determine whether or not there is a serial correlation in a data set.
- It tests the null hypothesis of no serial correlation against the alternative positive or negative serial correlation hypothesis.
- The test is named after James Durbin and Geoffrey Watson, who developed it in 1950.

The Durbin-Watson Statistic (DW) is approximated by:

$$ DW = 2(1 - r) $$

Where:
\(r\) is the sample correlation between regression residuals from one period and the previous period.

- The test statistic can take on values ranging from 0 to 4.
- A value of 2 indicates no serial correlation, a value between 0 and 2 indicates a positive serial correlation, and a value between 2 and 4 indicates a negative serial correlation:

- If there is no autocorrelation, the regression errors will be uncorrelated, and thus \(DW = 2\)

$$ DW = 2(1 - r) = 2(1 - 0) = 2 $$

- For positive serial autocorrelation, \(DW < 2\).
  For example, if serial correlation of the regression residuals = 1,
  \(DW = 2(1 - 1) = 0\).

- For negative autocorrelation, \(DW > 2\).
  For example, if serial correlation of the regression residual = −1, \(DW = 2(1 - (-1)) = 4\).

- To reject the null hypothesis of no serial correlation, need to find a critical value lower than the calculated value of d*.

Define \(d_l\) as the lower value and \(d_u\) as the upper value:

- If the DW statistic is less than \(d_l\), we reject the null hypothesis of no positive serial correlation.
- If the DW statistic is greater than \((4 - d_l)\), we reject the null hypothesis, indicating a significant negative serial correlation.
- If the DW statistic falls between \(d_l\) and \(d_u\), the test results are inconclusive.
- If the DW statistic is greater than \(d_u\), we fail to reject the null hypothesis of no positive serial correlation.
- Refer Figure 5.7

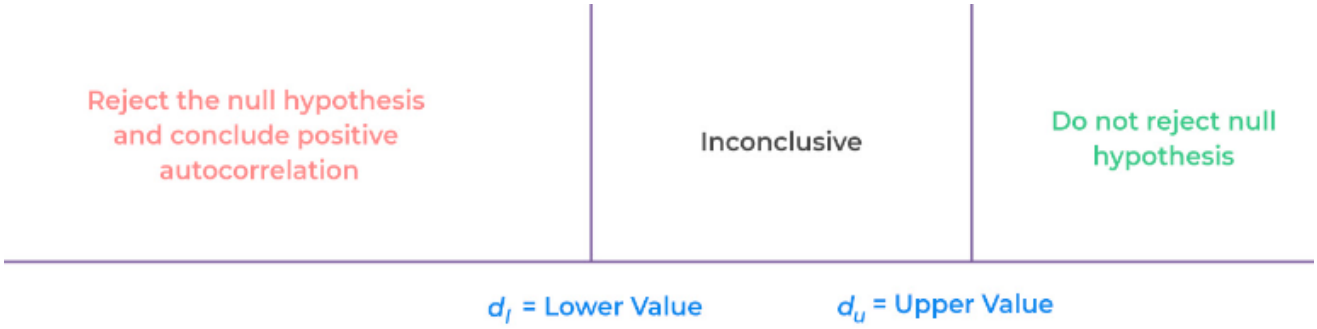


Figure 5.7 - Durbin-Watson Test for Serial Correlation

Reject H_0 if $d < d_L$.
Do not reject H_0 if $d > d_U$.
Test inconclusive if $d_L < d < d_U$.

Example 5.1:

The Durbin-Watson Test for Serial Correlation

Consider a regression output with two independent variables that generate a DW statistic of 0.654. Assume that the sample size is 15. Test for serial correlation of the error terms at the 5% significance level.

Solution

From the Durbin-Watson table with $n = 15$ and $k = 2$,

$d_l = 0.95$ and $d_u = 1.54$.

Since $d = 0.654 < 0.95 = d_l$,

Reject the null hypothesis and conclude that there is significant positive autocorrelation.

Example 5.2

Consider a regression model with 80 observations and two independent variables. Assume that the correlation between the error term and the first lagged value of the error term is 0.18. The most appropriate decision is:

A. reject the null hypothesis of positive serial correlation.

B. fail to reject the null hypothesis of positive serial correlation.

C. declare that the test results are inconclusive.

Solution

The correct answer is **C**.

The test statistic is:

$$ DW \approx 2(1 − r) = 2(1 − 0.18) = 1.64 $$

The critical values from the Durbin Watson table with $n = 80$ and $k = 2$ is $d_l = 1.59$ and $d_u = 1.69$.

Because 1.69 > 1.64 > 1.59, determine the test results are inconclusive.

10. Discuss in detail about Autocorrelation. And differentiate between serial correlation and autocorrelation.

➢ **Autocorrelation**

- Autocorrelation, refers to the degree of correlation of the same variables between two successive time intervals.
- Autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals.
- Autocorrelation measures the relationship between a variable's current value and its past values.
- The value of autocorrelation ranges from -1 to 1.
- An autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of -1 represents a perfect negative correlation.
- A value between -1 and 0 represents negative autocorrelation.
- A value between 0 and 1 represents positive autocorrelation.
- Autocorrelation gives information about the trend of a set of historical data so that it can be useful in the technical analysis

Types of Autocorrelation

- **Positive autocorrelation**

  The observations with positive autocorrelation can be plotted into a smooth curve. By adding a regression line, it can be observed that a positive error is followed by another positive one, and a negative error is followed by another negative one. Refer Figure 5.8

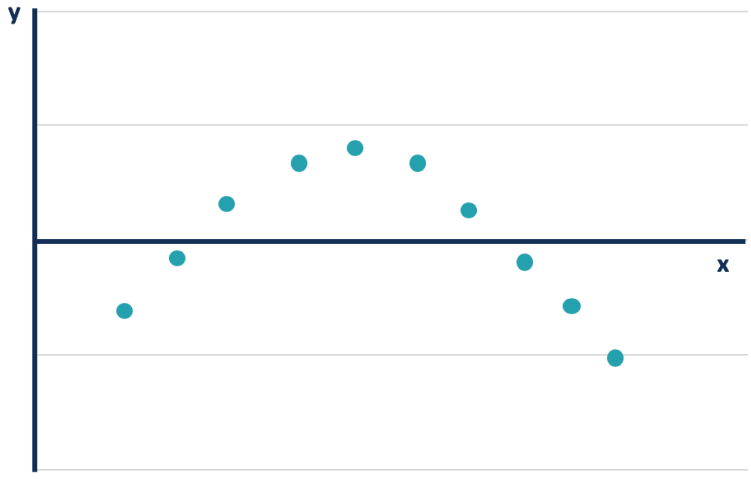


Figure 5.8 – Positive Autocorrelation

- **Negative autocorrelation**

  Conversely, negative autocorrelation represents that the increase observed in a time interval leads to a proportionate decrease in the lagged time interval. By plotting the observations with a regression

line, it shows that a positive error will be followed by a negative one and vice versa. Refer Figure 5.9

Figure 5.9 – Negative Autocorrelation

- Autocorrelation can be applied to different numbers of time gaps, which is known as lag.
- A lag 1 autocorrelation measures the correlation between the observations that are a one-time gap apart.
- For example, to learn the correlation between the temperatures of one day and the corresponding day in the next month, a lag 30 autocorrelation should be used (assuming 30 days in that month).
- Autocorrelation refers to the correlation between a time series variable and its own lagged values over time. In other words, it measures the degree of similarity between observations of a variable at different points in time.
- Autocorrelation is an important concept in time series analysis as it helps to identify patterns and relationships within the data.
- Positive autocorrelation occurs when a time series variable is correlated with its past values, while negative autocorrelation occurs when it is correlated with its future values.
- Zero autocorrelation indicates that there is no correlation between the variable and its lagged values.

Benefits of Autocorrelation

- Autocorrelation has several benefits in time series analysis:
- **Identifying patterns** – Autocorrelation helps to identify patterns in the time series data, which can provide insights into the behavior of the variable over time.
- **Model selection** – Autocorrelation can be used to select appropriate models for time series analysis.

- **Forecasting** – Autocorrelation can help to forecast future values of a time series variable.
- **Validating assumption**s – Autocorrelation can be used to validate assumptions of statistical models.
- **Hypothesis testing** – Autocorrelation can affect the results of hypothesis tests, such as t-tests and F-tests. By

Test for Autocorrelation

- Autocorrelation can be assessed using a variety of statistical techniques such as the autocorrelation function (ACF), partial autocorrelation function (PACF), and the Durbin-Watson statistic.
- These methods help to quantify the strength and direction of the autocorrelation and can be used to model and forecast time series data.

- The Durbin-Watson statistic is commonly used to test for autocorrelation.
- It can be applied to a data set by statistical software. T
- he outcome of the Durbin-Watson test ranges from 0 to 4.
- An outcome closely around 2 means a very low level of autocorrelation.
- An outcome closer to 0 suggests a stronger positive autocorrelation, and an outcome closer to 4 suggests a stronger negative autocorrelation.

- The autocorrelation function (ACF) assesses the correlation between observations in a time series for a set of lags. The ACF for time series y is given by:

$$\textbf{Corr } (y_t, y_{t-k}),\ k=1,2,....$$

Analysts typically use graphs to display this function.

Computation of Autocorrelation in Python

The pandas.Series.autocorr() function lets you compute the lag-N (default=1) autocorrelation on a given series.

Code Snippet:

df['series'].autocorr(lag=1)

- **Serial Correlation Versus Autocorrelation**

1. Serial correlation is a statistical concept that refers to the correlation between a variable and itself over time. It is used to measure the degree to which a variable's values at one point in time are related to its values at another point in time. Serial correlation is often used in time-series analysis to detect patterns in data and to test whether a model is appropriate for the data.

2. Autocorrelation is a specific type of serial correlation that measures the correlation between a variable and its lagged values. In other words, autocorrelation measures the degree to which a variable's values at one point in time are related to its values at previous points in time. Autocorrelation is often used to assess whether a time-series model is appropriate for the data.
3. Serial correlation is a more general term that refers to the correlation between a variable and itself over time, whereas autocorrelation specifically refers to the correlation between a variable and its lagged values.
4. In terms of applications, serial correlation is often used to analyze patterns in data over time, such as trends and seasonality, while autocorrelation is often used in time-series analysis to assess the fit of a model and to make predictions about future values.
5. For example, in finance, serial correlation might be used to analyze the daily returns of a stock or portfolio over time to detect trends and seasonality. Autocorrelation might be used to test whether a time-series model is appropriate for the data and to make predictions about future returns based on past values.

11. Give a brief introduction about Survival Analysis.

Survival Analysis

- Survival analysis is a field of statistics that focuses on analysing the expected time until a certain event happens.
- Survival analysis can be used for analysing the results of that treatment in terms of the patients' life expectancy.
- The term `survival time' specifies the length of time taken for failure to occur.
- Survival analysis is used to analyse data in which the time until the event is of interest.
- The response is often referred to as a failure time, survival time, or event time.
- This branch of statistics developed around measuring the effects of medical treatment on patients' survival in clinical trials.
- Examples
    - Time until tumour recurrence
    - Time until a machine part fails

Survival curves

- The fundamental concept in survival analysis is the survival curve, S(t), as in figure 5.10, which is a function that maps from a duration, t, to the probability of surviving longer than t, it's just the complement of the CDF:

  S(t) = 1 − CDF(t)

  where CDF(t) is the probability of a lifetime less than or equal to t.

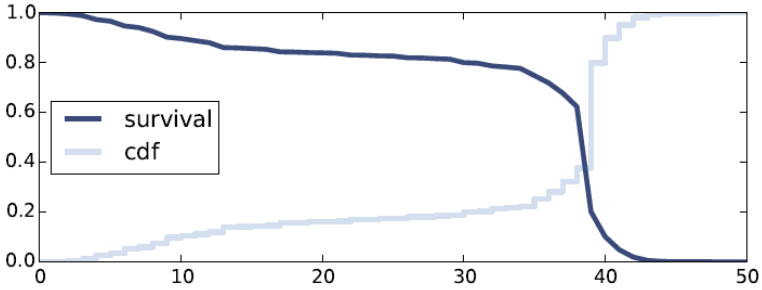


Figure 5.10 - Survival curves

- For example, in the NSFG dataset, given the duration of 11189 complete pregnancies.
- Can read this data and compute the CDF:

```
preg = nsfg.ReadFemPreg()
complete = preg.query('outcome in [1, 3, 4]').prglngth
cdf = thinkstats2.Cdf(complete, label='cdf')
```

- The outcome codes 1, 3, 4 indicate live birth, stillbirth, and miscarriage.
- The DataFrame method query takes a boolean expression and evaluates it for each row, selecting the rows that yield True.

```
class SurvivalFunction(object):
    def __init__(self, cdf, label=''):
        self.cdf = cdf
        self.label = label or cdf.label
@property
def ts(self):
        return self.cdf.xs
@property
def ss(self):
        return 1 - self.cdf.ps
```

- SurvivalFunction provides two properties:
  - ts, which is the sequence of lifetimes,
  - ss, which is the survival curve.

- From the survival curve can derive the hazard function; for pregnancy lengths, the hazard function maps from a time, t, to the fraction of pregnancies that continue until t and then end at t.

$$\lambda(t) = \frac{S(t) - S(t+1)}{S(t)}$$

- The numerator is the fraction of lifetimes that end at t, which is also PMF(t).

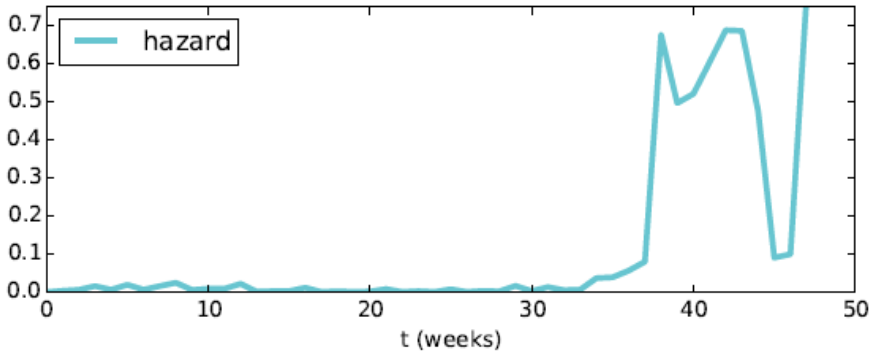


Figure 5.5 - Hazard curve

Censoring

- In longitudinal studies exact survival time is only known for those individuals who show the event of interest during the follow-up period. These individuals are called censored observations.
- The following terms are used in relation to censoring:
- **Right censoring:** a subject is right censored if it is known that failure occurs some time after the recorded follow-up period.
- **Left censoring:** a subject is left censored if it is known that the failure occurs some time before the recorded follow-up period.
- **Interval censoring:** a subject is interval censored if it is known that the event occurs between two times, but the exact time of failure is not known.

Truncation

- A truncation period means that the outcome of interest cannot possibly occur.
- A censoring period means that the outcome of interest may have occurred.
- There are two types of truncation:
- **Left truncation:** a subject is left truncated if it enters the population at risk some stage after the start of the follow-up period.

- **Right truncation:** a subject is right truncated if it leaves the population at risk some stage after the study start.

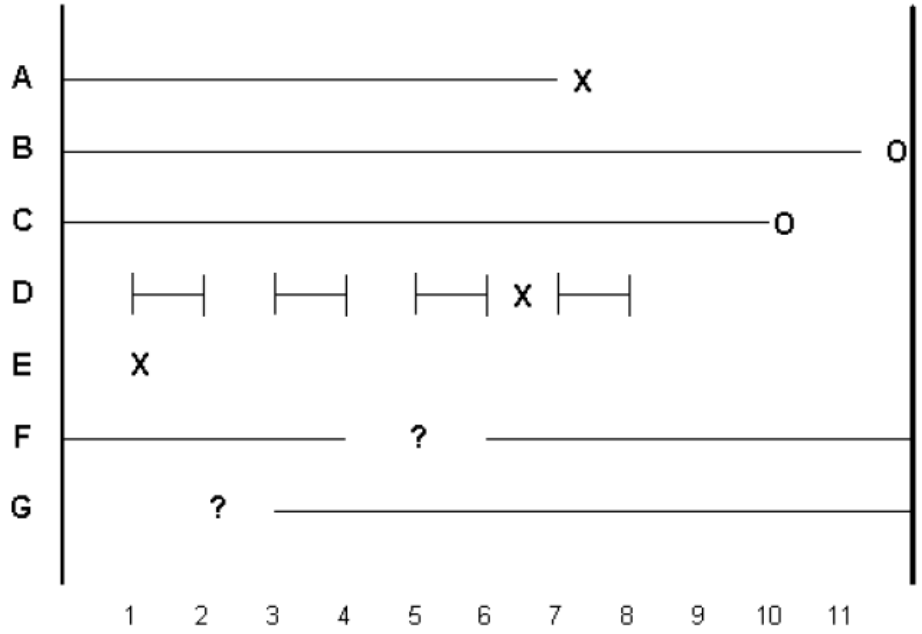


Figure 5.6: Left-, right-censoring, and truncation

- An `X' indicates that the subject has experienced the outcome of interest; a `O' indicates censoring.
- Subject A experiences the event of interest on day 7. Subject B does not experience the event during the study period and is right censored on day 12 (this implies that subject B experienced the event sometime
- after day 12).
- Subject C does not experience the event of interest during its period of observation and is censored on day 10.
- Subject D is interval censored: this subject is observed intermittantly and experiences the event of interest sometime between days 5 { 6 and 7 { 8. Subject E is left censored | it has been found to have already experienced the event of interest when it enters the study on day 1.
- Subject F is interval truncated: there is no way possible that the event of interest could occur to this individual between days 4 { 6.
- Subject G is left truncated: there is no way possible that the event of interest could have occurred before the subject enters the study on day 3.

12. What are the Effective Strategies for Handling Missing Values in Data Analysis?

Missing Value

- Missing data is defined as the values or data that is not stored for some variable/s in the given dataset.
- Below is a sample of the missing data from the Titanic dataset.
- The columns 'Age' and 'Cabin' have some missing values.

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

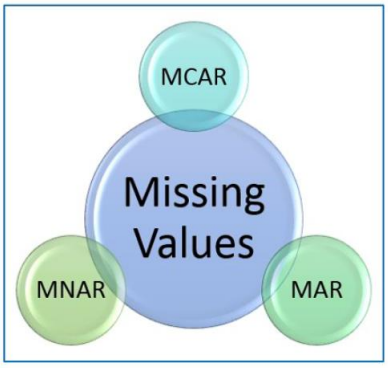
Reason for Missing Values

- Past data might get corrupted due to improper maintenance.
- Observations are not recorded for certain fields due to some reasons. There might be a failure in recording the values due to human error.
- The user has not provided the values intentionally
- Item nonresponse: This means the participant refused to respond.

Reason to handle missing data

- The missing data will decrease the predictive power of the model. If the algorithms are applied with missing data, then there will be bias in the estimation of parameters.
- The results are not confident if the missing data is not handled properly.

Types of Missing Values



Type	*Definition*
Missing completely at random (MCAR)	Missing data are randomly distributed across the variable and unrelated to other variables.
Missing at random (MAR)	Missing data are not randomly distributed but they are accounted for by other observed variables.
Missing not at random (MNAR)	Missing data systematically differ from the observed values.

Missing Completely At Random (MCAR)

- In MCAR, the probability of data being missing is the same for all the observations.
- In this case, there is no relationship between the missing data and any other values observed or unobserved within the given dataset.
- That is, missing values are completely independent of other data. There is no pattern.
- In the case of MCAR data, the value could be missing due to human error, some system/equipment failure, loss of sample, or some unsatisfactory technicalities while recording the values.
- **For Example**, suppose in a library there are some overdue books. Some values of overdue books in the computer system are missing. The reason might be a human error, like the librarian forgetting to type in the values.

Missing At Random (MAR)

- MAR data means that the reason for missing values can be explained by variables which have complete information, as there is some relationship between the missing data and other values/data.
- In this case, the data is not missing for all the observations.
- It is missing only within sub-samples of the data, and there is some pattern in the missing values.
- **For example,** if you check the survey data, you may find that all the people have answered their 'Gender,' but 'Age' values are **mostly** missing for people who have answered their 'Gender' as 'female.' (The reason being most of the females don't want to reveal their age.)
- So, the probability of data being missing depends only on the observed value or data. In this case, the variables 'Gender' and 'Age' are related.

The reason for missing values of the 'Age' variable can be explained by the 'Gender' variable, but you cannot predict the missing value itself.

Missing Not At Random (MNAR)

- Missing values depend on the unobserved data.
- If there is some structure/pattern in missing data and other observed data **can not explain** it, then it is considered to be Missing Not At Random (MNAR).
- If the missing data does not fall under the MCAR or MAR, it can be categorized as MNAR.
- It can happen due to the reluctance of people to provide the required information.
- A specific group of respondents may not answer some questions in a survey.

Methods for identifying missing data

Functions	Descriptions
.isnull()	This function returns a pandas dataframe, where each value is a boolean value True if the value is missing, False otherwise.
.notnull()	Similarly to the previous function, the values for this one are False if either NaN or None value is detected.
.info()	This function generates three main columns, including the "Non-Null Count" which shows the number of non-missing values for each column.
.isna()	This one is similar to isnull and notnull. However it shows True only when the missing value is NaN type.

Approach to handle missing values in a dataset.

- Deleting Rows with missing values
- Impute missing values for continuous variable
- Impute missing values for categorical variable
- Other Imputation Methods
- Using Algorithms that support missing values

- Prediction of missing values
- Imputation using Deep Learning Library — Datawig

Reg. No. : 4

Question Paper Code : 20014

B.E./B.Tech. DEGREE EXAMINATIONS, NOVEMBER/DECEMBER 2023.

Third/Fourth Semester

Artificial Intelligence and Data Science

AD 3491 — FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

(Common to : Computer Science and Business Systems)

(Regulations 2021)

Time : Three hours | Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. List down any five skills required for a data analyst.
2. Outline the significance of Exploratory Data Analysis (EDA)?
3. Tabulate the differences between univariate, bivariate, and multivariate analysis. Give examples.
4. Give an example of a data set with a non-Gaussian distribution.
5. Explain the term 'Normal Distribution'.
6. Brief about the Type I and Type II errors in Statistics. Identify the relationship between standard error and margin of error.
7. With an assumption of a null hypothesis as correct, what does it mean when the p-values are high and low?
8. Define the term one-factor ANOVA.
9. Outline a few approaches to detect outliers? Explain different ways to deal with it.
10. Give an approach to handle missing values in a dataset.

PART B — (5 × 13 = 65 marks)

11. (a) Brief about exploratory analysis in Dataset analysis and knowledge discovery process. (13)

Or

(b) (i) Outline the purpose of data cleansing. How missing and nullified data attributes are handled and modified during preprocessing stage? (6)

(ii) Explain Data Analytic life cycle. Brief about Regression Analysis. (7)

12. (a) (i) Indicate whether each of the following distribution is positively or negatively skewed. The distribution of

(1) Incomes of tax payers have a mean of $48,000 and a median of $43,000.

(2) GPAs for all students at some college have a mean of 3.01 and a median of 3.20 (6)

(ii) Consider the following number of online examination attempt 15 students :

2, 17, 5, 3, 28, 7, 5, 8, 5, 6, 2 , 12, 10, 4, 3

(1) Find the mode, median and mean for these data.

(2) Draw the distribution for balanced, positively skewed or negatively skewed. (7)

Or

(b) (i) Assume that SAT math scores approximates a normal curve with a mean of 500 and a standard deviation of 100.

Sketch a normal curve and shade in the target area(s) described by each of the following statements:

- More than 570
- Less than 515
- Between 520 and 540. (5)

(ii) Assume that the burning times of electric light bulbs approximate a normal curve with a mean of 1200 hours and a standard deviation

of 120 hours. If a large number of new lights are installed at the same time (possibly along a newly opened freeway), at what time will

- 1 percent fails? (2)
- 50 percent fail? (2)
- 95 percent fail? (4)

13. (a) (i) Among 100 couples who had undergone marital counseling, 60 couples described their relationships as improved, and among this latter group, 45 couples had children. The remaining couples described their relationships as unimproved, and among this group, 5 couples had children.

(1) What is the probability of randomly selecting a couple who described their relationship as improved? (2)

(2) What is the probability of randomly selecting a couple with children? (2)

(3) What is the conditional probability of randomly selecting a couple with children, given that their relationship was described as improved? (2)

(ii) The probability of a boy being born equals 0.50, or 1/2, as does the probability of a girl being born. For a randomly selected family with two children, what's the probability of

(1) Two boys, that is, a boy and a boy? (3)

(2) Two girls? (2)

(3) Either two boys or two girls? (2)

Or

(b) (i) The normal range for a widely accepted measure of body size, the Body Mass Index (BMI), ranges from 18.5 to 25. Using the midrange BMI score of 21.75 as the null hypothesized value for the population mean, test this hypothesis at the .01 level of significance given a random sample of 30 weight-watcher participants who show a mean BMI = 22.2 and a standard deviation of 3.1. (6)

(ii) State any two reasons why the research hypothesis is not tested directly. Explain them in brief. (7)

14. (a) (i) Twenty-three overweight male volunteers are randomly assigned to three different treatment programs designed to produce a weight loss by focusing on either diet, exercise, or the modification of eating behavior. Weight changes were recorded, to the nearest pound, for all participants who completed the two-month experiment. Positive scores signify a weight drop; negative scores, a weight gain.

Weight Change

Diet	Exercise	Behavior Modification
3	–1	7
4	8	1
0	4	10
–3	2	0
5	2	18
10	–3	12
3	–	4
0	–	6
–	–	5
T 22	12	63
n 8	6	9

$\Sigma X = G = 97$; $N = 23$ $\Sigma X^2 = 961$

Summarize the results with an ANOVA table. (6)

(ii) The F test describes the ratio of two sources of variability : that for subjects treated differently and that for subjects treated similarly. Is there any sense in the which the t test for two independent groups can be viewed likewise? (7)

Or

(b) (i) Brief about Partial squared curvilinear correlation. What is its purpose? (6)

(ii) Brief about TUKEY'S HSD Test. Additionally, explain in brief about two-factor ANOVA. (7)

15. (a) (i) In Statistics, what happens when the goodness of fit test score is low? (6)

(ii) Given the following dataset of employee, Using regression analysis, find the expected salary of an employee if the age is 45. (7)

Age	Salary
54	67000
42	43000
49	55000
57	71000
35	25000

Or

(b) (i) Define autocorrelation and how is it calculated? What does the negative correlation convey? (6)

(ii) What is the philosophy of Logistic regression? What kind of model it is? What does logistic Regression predict? Tabulate the cardinal differences of Linear and Logistic Regression. (7)

PART C — ($1 \times 15 = 15$ marks)

16. (a) (i) Discuss the role of sampling distribution in inferential statistics. (8)

(ii) Indicate whether each of the following distributions is positively or negatively skewed. The distribution of

- Incomes of taxpayers have a mean of \$48,000 and a median of \$43,000.
- GPAs for all students at some college have a mean of 3,01 and a median of 3.20.
- Daily TV viewing times for preschool children has a mean of 55 minutes and a median of 73 minutes. (7)

Or

(b) (i) Assume that we have a stream of items of large and unknown length that we can only iterate over once. Devise an effective sampling algorithm that randomly chooses an item from this stream such that each item is equally likely to be selected. (8)

(ii) During their first swim through a water maze, 15 laboratory rats made the following number of errors (blind alleyway entrances): 2, 17, 5, 3, 28, 7, 5, 8, 5, 6, 2, 12, 10, 4, 3.

- Find the mode, median and mean for these data.
- Without constructing a frequency distribution or graph, would you characterize the shape of this distribution as balanced, positively skewed or negatively skewed? (7)

ANNA UNIVERSITY QP NOV/DEC 2023

SOLVED ANSWERS

PART A

1. List down any five skills required for a data analyst.

1. Structured Query Language, or SQL, to communicate with databases.
2. Statistical programming languages, like R or Python, to perform advanced analyses.
3. Machine learning for developments in data science.
4. Probability and statistics the field of math and science concerned with collecting, analyzing, interpreting, and presenting data.
5. Data management refers to collecting, organizing, and storing data in an efficient, secure, and cost-effective way.
6. Statistical visualization Gleaning insights from data is only one part of the data analysis process.
7. Econometrics to apply statistical and mathematical data models to economics to help forecast future trends based on historical data.

2. Outline the significance of Exploratory Data Analysis (EDA)?

- The main purpose of EDA is to help look at data before making any assumptions.
- It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

3. Tabulate the differences between univariate, bivariate, and multivariate analysis. Give examples.

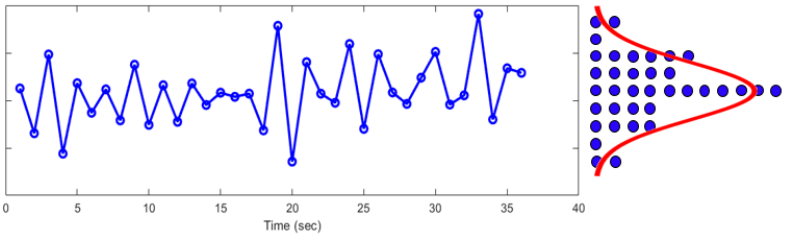
Univariate VS Bivariate VS Multivariate analysis

Univariate	Bivariate	Multivariate
Univariate statistics summarize only one variable at a time.	Bivariate statistics compare two variables.	Multivariate statistics compare more than two variables.
It does not deal with causes or relationships	It deals with causes and relationships and the analysis is done	It also deals with causes and relationships and the analysis is done
It does not contain any dependent variable.	It contain only one dependent variable.	It is similar to bivariate but contains more than one dependent variable.
The purpose of univariate analysis is to describe	The purpose of univariate analysis is to explain.	The purposes of multivariate data analysis is to study the relationships among the P attributes, classify the n collected samples into homogeneous groups, and make inferences about the underlying populations from the sample.
The example of a univariate data can be height.	Example of bivariate data can be temperature and ice cream sales in summer season.	Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

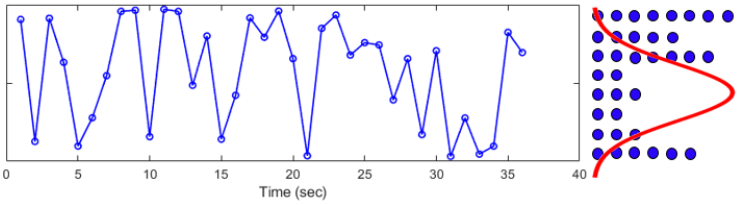
4. Give an example of a data set with a non-Gaussian distribution.

- Sample time series for two measured variables: one is Gaussian-distributed (top), and the other is not (bottom).
- On the right, the measurements in a histogram. This can help to check if a variable is Gaussian or not.
- Non-Gaussian distributed time series data arise when the mean or noise statistics vary with time.

- If the mean varies with time, the variable could be non-stationary / time-varying (its trend changes with time), auto- or cross-correlated (it changes depending on its previous value or the values of other variables), or its value is computed from the values of other Gaussian variables but in a nonlinear way.



The Gaussian distribution (red) fits the data well.



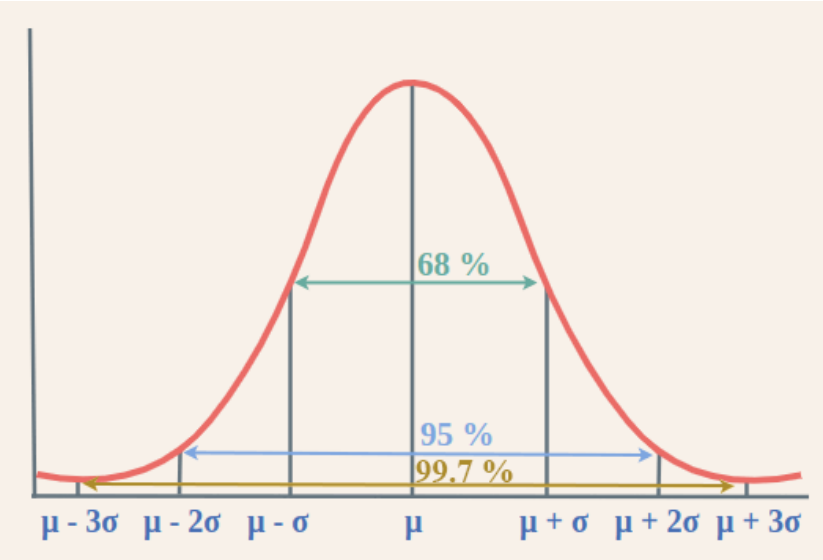
The Gaussian distribution (red) **does not** fit the data well.

5. Explain the term ˜Normal Distribution"

- Normal distribution the **Normal Distribution**, also called the **Gaussian Distribution**,, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
- In graphical form, the normal distribution appears as a "bell curve".

Empirical Rule states that,

- 68% of the data approximately fall within one standard deviation of the mean, i.e. it falls between {Mean – One Standard Deviation, and Mean + One Standard Deviation}
- 95% of the data approximately fall within two standard deviations of the mean, i.e. it falls between {Mean – Two Standard Deviation, and Mean + Two Standard Deviation}
- 99.7% of the data approximately fall within a third standard deviation of the mean, i.e. it falls between {Mean – Third Standard Deviation, and Mean + Third Standard Deviation}



6. Brief about the Type I and Type II errors in Statistics. Identify the relationship between standard error and margin of error.

Type I and Type II errors

- A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

α= P(TypeI Error) and β = P(TypeII Error)

Example

- Type I error (false positive): the test result says you have coronavirus, but you actually don't.
- Type II error (false negative): the test result says you don't have coronavirus, but you actually do.

Relationship between standard error and margin of error

- A margin of error is a statistical measure that accounts for the degree of error received from the outcome of the research sample.
- Standard error measures the accuracy of the representation of the population sample to the mean using the standard deviation of the data set.
- Standard error and standard deviation are both measures of variability:
- The standard deviation describes variability within a single sample.
- The standard error estimates the variability across multiple samples of a population.

7. With an assumption of a null hypothesis as correct, what does it mean when the p-values are high and low?

- **Low P values:** data are unlikely with a true null.
- A low p value means that the sample result would be unlikely if the null hypothesis were true and leads to the rejection of the null hypothesis.
- **High P values:** data are likely with a true null.
- A high p value means that the sample result would be likely if the null hypothesis were true and leads to the retention of the null hypothesis.

8. Define the term one-factor ANOVA.

- The most basic method is the **single-factor analysis of variance**, which is also known as the **one-way ANOVA** simply because this method contains just one factor (single factor).
- A single factor with a maximum of two levels can still be analyzed using the t-test or z-test or other appropriate tests.
- However, the single factor with more than two levels will need ANOVA with advanced methods depending on the experimental situations.
- The most basic single factor with more than two levels is the **completely randomized design (CRD).**

9. Outline a few approaches to detect outliers? Explain different ways to deal with it.

- **Outliers** are values at the extreme ends of a dataset.
- **Outliers** are extreme values that differ from most other data points in a dataset.
- They can have a big impact on statistical analyses and skew the results of any hypothesis tests.
- It's important to carefully identify potential outliers dataset and deal with them in an appropriate manner for accurate results.
- There are **four ways** to identify or detect outliers:
  1. **Sorting method** - can **sort** quantitative variables from low to high and scan for extremely low or extremely high values.
  2. **Data visualization method -** can use software to **visualize** data with a box plot, or a box-and-whisker plot
  3. **Statistical tests** - applying **statistical tests** or procedures to identify extreme values.
  4. **Interquartile range method - t**he range of the middle half of dataset.

Different ways to deal with outliers

- **Retain outliers**

  Keeping outliers is usually the better option when not sure if they are errors.

- **Remove outliers**

  Deleting extreme values from dataset before performing statistical analyses.

10. Give an approach to handle missing values in a dataset.

- Deleting Rows with missing values
- Impute missing values for continuous variable
- Impute missing values for categorical variable
- Other Imputation Methods
- Using Algorithms that support missing values
- Prediction of missing values
- Imputation using Deep Learning Library — Datawig

PART B

11.(a) Brief about exploratory analysis in Dataset analysis and knowledge discovery process. (13)

Refer Unit 1 – Page No. 22

OR

11 (b) (i) Outline the purpose of data cleansing. How missing and nullified data attributes are handled and modified during preprocessing stage? (6)

Refer Unit 1 – Page No. 15

(ii) Explain Data Analytic life cycle. Brief about Regression Analysis. (7)

Life Cycle of Data Analytics

Phase 1: Discovery -

- The data science team is trained and researches the issue.
- Create context and gain understanding.
- Learn about the data sources that are needed and accessible to the project.
- The team comes up with an initial hypothesis, which can be later confirmed with evidence.

Phase 2: Data Preparation -

- Methods to investigate the possibilities of pre-processing, analysing, and preparing data before analysis and modelling.

- Data preparation tasks can be repeated and not in a predetermined sequence.
- Some of the tools used commonly for this process include - Hadoop, Alpine Miner, Open Refine, etc.

Phase 3: Model Planning -

- The team studies data to discover the connections between variables. Later, it selects the most significant variables as well as the most effective models.
- In this phase, the data science teams create data sets that can be used for training for testing, production, and training goals.
- The team builds and implements models based on the work completed in the modelling planning phase.
- Some of the tools used commonly for this stage are MATLAB and STASTICA.

Phase 4: Model Building -

- The team creates datasets for training, testing as well as production use.
- The team is also evaluating whether its current tools are sufficient to run the models or if they require an even more robust environment to run models.
- Tools that are free or open-source or free tools Rand PL/R, Octave, WEKA.
- Commercial tools - MATLAB, STASTICA.

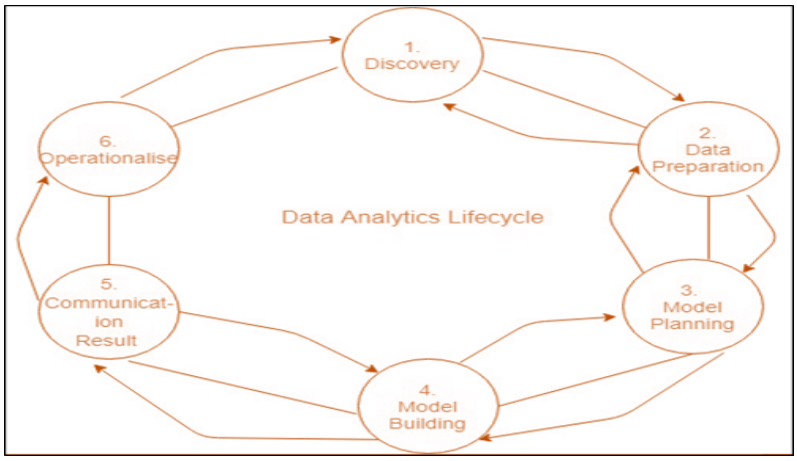
Phase 5: Communication Results -

- Following the execution of the model, team members will need to evaluate the outcomes of the model to establish criteria for the success or failure of the model.
- The team is considering how best to present findings and outcomes to the various members of the team and other stakeholders while taking into consideration cautionary tales and assumptions.
- The team should determine the most important findings, quantify their value to the business and create a narrative to present findings and summarize them to all stakeholders.

Phase 6: Operationalize -

- The team distributes the benefits of the project to a wider audience. It sets up a pilot project that will deploy the work in a controlled manner prior to expanding the project to the entire enterprise of users.
- This technique allows the team to gain insight into the performance and constraints related to the model within a production setting at a small scale and then make necessary adjustments before full deployment.

- The team produces the last reports, presentations, and codes.
- Open source or free tools such as WEKA, SQL, MADlib, and Octave.



REGRESSION ANALYSIS

- Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables.
- It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.
- Regression analysis includes several variations, such as linear, multiple linear, and nonlinear.
- The most common models are simple linear and multiple linear.
- Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Types of Regression Analysis Techniques

- Linear Regression.
- Logistic Regression.
- Ridge Regression.
- Lasso Regression.
- Polynomial Regression.
- Bayesian Linear Regression.

Advantage of regression

- Regression is a helpful statistical concept that helps facilitate decision-making by determining the correlation between a dependent variable and one or more independent variables.

Disadvantage of regression

- Sensitive to outliers and noise

Limitations of regression

- The relationship between x and y should be linear.
- The observations in a sample must be independent of each other

12 (a) (i) Indicate whether each of the following distribution is positively or negatively skewed.

a. The distribution of Incomes of tax payers have a mean of $48,000 and a median of $43,000.

b. GPAs for all students at some college have a mean of 3.01 and a median of 3.20 **(6)**

(a) Incomes of taxpayers: The distribution is positively skewed. This is because the mean ($48,000) is greater than the median ($43,000), indicating that there are a few high-income earners pulling the mean up.

(b) GPAs for all students at some college: The distribution is negatively skewed. The mean (3.01) is less than the median (3.20), suggesting that there are a few students with lower GPAs pulling the mean down.

(ii) Consider the following number of online examination attempt 15 students :

2, 17, 5, 3, 28, 7, 5, 8, 5, 6, 2 , 12, 10, 4, 3

a) Find the mode, median and mean for these data.

b) Draw the distribution for balanced, positively skewed or negatively skewed. **(7)**

Solution:

Scores of 15 students are:

2, 17, 5, 3, 28, 7, 5, 8, 5, 6, 2, 12, 10, 4, 3

Arranging scores in ascending order,

2, 2, 3, 3, 4, 5, 5, 5, 6, 7, 8, 10, 12, 17, 28

Mode - Mode of a given data is that value of observation that has the maximum occurrence or repetition i.e., most frequent.

Therefore, 5 has the highest occurrence.

∴ Mode = 5

Median = Middlemost observation (in this case, 8th observation)

∴ Median = 5

Mean = (Sum of all the values / total number of values)

= 117/15 = 8

The distribution is positively skewed. This is because the mean is greater than the median.

OR

(b) (i) Assume that SAT math scores approximates a normal curve with a mean of 500 and a standard deviation of 100. Sketch a normal curve and shade in the target area(s) described by each of the following statements:

- **More than 570**
- **Less than 515**
- **Between 520 and 540. (5)**

(ii) Assume that the burning times of electric light bulbs approximate a normal curve with a mean of 1200 hours and a standard deviation of 120 hours. If a large number of new lights are installed at the same time (possibly along a newly opened freeway), at what time will

- **1 percent fails? (2)**
- **50 percent fail? (2)**
- **95 percent fail? (4)**

Solution

Let the burning times of electric bulbs be X.

The mean and standard deviation of the burning time is 1200 hours and 120 hours, respectively.

The burning time will follow a normal distribution.

That is,

X follows N(mean=1200, sd=120)

To find x such that P(z is greater than x) = 0.5

At probability 0.5 the z-score is 2.33.

Using z-score formula the x value is as follows:

z = (x-mean)/sd

2.33 = (x-1200)/120

x = 2.33(120) +1200

x =1479.6

Therefore, the burning time at the upper 50 percent is 1479.6 hours.

To find x such that P(z is less than x) = 0.01

At probability 0.01 the z-score is -2.33.

Using z-score formula the x value is as follows:

z = (x-mean)/sd

-2.33 = (x-1200)/120

x = -2.33(120) +1200

x =920.4

Therefore, the burning time at the lower 1 percent is 920.4 hours.

13. (a) (i) Among 100 couples who had undergone marital counseling, 60 couples described their relationships as improved, and among this latter group, 45 couples had children. The remaining couples described their relationships as unimproved, and among this group, 5 couples had children.

a) What is the probability of randomly selecting a couple who described their relationship as improved? (2)

b) What is the probability of randomly selecting a couple with children? (2)

c) What is the conditional probability of randomly selecting a couple with children, given that their relationship was described as improved? (2)

(d) What is the conditional probability of an improved relationship, given that a couple has children?

Solution

(a) The probability of a couple describing their relationship as improved is 0.6 or 60%.

(b) The probability of a couple having children is 0.5 or 50%.

(c) The conditional probability of a couple having children given that their relationship was described as improved is 0.75 or 75%.

(d) The conditional probability of an improved relationship, given a couple has children, is 0.9 or 90%.

Step 1: Calculate the total number of Couples

- The total number of couples is given as 100. So, this will act as the denominator for all the probability calculations.

Step 2: Probability of a Couple describing their relationship as Improved

- Out of the total 100 couples, 60 couples described their relationships as improved. So, the probability of randomly selecting a couple who described their relationship as improved is calculated as: 60/100=0.6 or 60%.

Step 3: Probability of a Couple having Children

- Out of the total 100 couples, 45+5=50 couples had children. So, the probability of randomly selecting a couple with children is calculated as: 50/100=0.5 or 50%.

Step 4: Conditional Probability of a Couple having Children given their relationship is improved

- Out of the 60 couples who described their relationships as improved, 45 couples had children. So, the conditional probability of randomly selecting a couple with children, given their relationship was described as improved, is calculated as: 45/60=0.75 or 75%.

Step 5: Conditional Probability of an improved relationship, given a couple has children

- Out of the 50 couples who had children, 45 couples described their relationships as improved. So, the conditional probability of an improved relationship, given a couple has children, is calculated as: 45/50=0.9 or 90%.

(ii) The probability of a boy being born equals 0.50, or 1/2, as does the probability of a girl being born. For a randomly selected family with two children, what's the probability of

(1) Two boys, that is, a boy and a boy? (3)

(2) Two girls? (2)

(3) Either two boys or two girls? (2)

Step 1: Analyze the probability of each event

- According to the problem, the probability for a child to be a boy or a girl is 0.50 or 1/2. Since having a boy or a girl are independent events, the probability of having two boys or two girls is the product of their individual probabilities (multiplication rule).

Step 2: Compute the probability of two boys

- The probability of having a boy is 1/2 and since the births are independent, the probability of having two boys is (1/2)∗(1/2)=1/4.

Step 3: Compute the probability of two girls

- Similar to Step 2, the probability of having two girls is (1/2)∗(1/2)=1/4.

Step 4: Compute the probability of either two boys or two girls

- Two boys and two girls are mutually exclusive events, meaning they can't occur simultaneously. Hence, the probability of having either two boys or two girls is the sum of their individual probabilities (addition rule). So, the probability is 1/4+1/4=1/2

OR

13. (b) (i) The normal range for a widely accepted meapre of body size, the Body Mass Index (BMI), ranges from 18.5 to Using the midrange BMI score of 21.75 as the null hypothesized value for the population mean, test this hypothesis at the .01 level of significance given a random sample of,30 weight-watcher participants who show a mean BMI = 22. /and a standard deviation of 3.1. (6)

Solution

- Not reject the null hypothesis
- Significance level of .01, the p-value will be 0.433, which is greater than the level of significance of .01.
- No difference between the BMI and body size of the 30 weight-watchers participants and the general population from which they were drawn.

(ii) State any two reasons why the research hypothesis is not tested directly. Explain them in brief. (7)

- The research hypothesis is not tested directly because it is difficult to prove a specific effect or relationship exists. Instead, K researchers test the null hypothesis, which states that there is no effect or relationship.

14. (a) (i) Twenty-three overweight male volunteers are randomly assigned to three different treatment programs designed to produce a weight loss by focusing on either diet, exercise, or the modification of eating behavior. Weight changes were recorded, to the nearest pound, for all participants who completed the two-month experiment. Positive scores signify a weight drop; negative scores, a weight gain.

	Weight Change		
	Diet	**Exercise**	**Behavior Modification**
	3	**—1**	**7**
	4	**8**	**1**
	0	**4**	**10**
	—3	**2**	**0**
	5	**2**	**18**
	10	**—3**	**12**
	3		**4**
	0		**6**
			5
T	**22**	**12**	**63**
N	**8**	**6**	**9**

EX = G = 97; N= 23 EX^2 = 961

Summarize the results with an ANOVA table. (6)

Step 1: Create ANOVA table

- First, let's set up an ANOVA table.
- Need five columns labeled
  - Source of Variation (which includes Between Groups, Within Groups, and Total),
  - Sum of Squares (SS),
  - Degrees of Freedom (df),
  - Mean Square (MS), and
  - F-ratio (F).

To fill out the table, will need the means and overall count, plus the method of weight loss.

Note that degrees of freedom for Between Groups is the number of groups minus 1 and for Within Groups is the total size minus the number of groups.

ANOVA

ANOVA – Weight Change (lb)

Cases	Sum of Squares	df	Mean Square	F	p
Program	116.413	2	58.207	2.673	0.094
Residuals	435.500	20	21.775		

Note. Type III Sum of Squares

(ii) The F test describes the ratio of two sources of variability: that for subjects treated differently and that for subjects treated similarly. Is there any sense in the which the t test for two independent groups can be viewed likewise? (7)

Short Answer

- Yes, the t test for two independent groups can be viewed similarly to the F test in the sense that both are examining variability.
- They differ in their specifics: the t test is comparing the mean difference against the variability within groups, while the F test is comparing the variability between groups against the variability within groups.

Step by step solution

Step 1: Understanding The F Test

The F test is normally used in the context of Analysis of Variance (ANOVA) to compare the variances between different groups. It is essentially a ratio of two estimates of variance: the variance between groups (numerator), and the variance within groups (denominator). If the between-group variance is significantly greater than the within-group variance, it would suggest that the means of the groups differ.

Step 2: Understanding The t Test

The t-test for two independent groups is used to compare the means of those groups to determine if they are significantly different. The t-test is calculated using the mean difference between the two groups (numerator) and the variability within the groups (denominator).

Step 3: Identifying the Link between The t Test and The F Test

In the t-test, we are technically comparing variability too, though we are specifically interested in whether the variability in group means is greater than what we would expect by chance. In the F test, we're more broadly comparing variability to examine if the amount of variability between group means is larger than the variability within groups. So it can be said that there is a link between them, but they're not quite serving the same purpose.

OR

14. (b) (i) Brief about Partial squared curvilinear correlation. What is its purpose? (6)

- The partial correlation coefficient (or partial correlation) is a statistical measure that quantifies the linear association between two variables, while controlling for the effects of one or more other variables.
- It measures the strength of the association between two variables, while accounting for the effects of other variables that may also be related to both of them.
- The formula for the partial correlation coefficient is similar to the formula for the Pearson correlation coefficient, but it includes a term that adjusts for the effects of other variables.
- The formula can be represented as:

$$r_{xy.z} = (r_{xy} - r_{xz} * r_{yz}) / \text{sqrt}((1-r_{xz}^2)(1-r_{yz}^2))$$

- where x, y and z are variables, r_{xy} is the correlation coefficient between x and y, r_{xz} is the correlation coefficient between x and z, and r_{yz} is the correlation coefficient between y and z.

- The value of the partial correlation coefficient ranges from -1 to 1, where a value of -1 indicates a perfect negative association, a value of 0 indicates no association, and a value of 1 indicates a perfect positive association.
- The partial correlation coefficient is a measure of association and not a measure of causality.
- A high partial correlation coefficient does not imply that one variable causes the other, only that the two variables have a strong association when controlling for the effect of other variables.

(ii) Brief about TUKEY'S HSD Test. Additionally, explain in brief about two-factor ANOVA. (7)

TUKEY'S HSD Test

- The Tukey HSD ("honestly significant difference" or "honest significant difference") test is a statistical tool used to determine if the relationship between two sets of data is statistically significant – that is, whether there's a strong chance that an observed numerical change in one value is causally related to an observed change in another value.
- The Tukey test is a way to test an experimental hypothesis.
- The Tukey's honestly significant difference test (Tukey's HSD) is used to test differences among sample means for significance.
- The Tukey's HSD tests all pairwise differences while controlling the probability of making one or more Type I errors.
- The Tukey's HSD test is one of several tests designed for this purpose and fully controls this Type I error rate.
- The value of the Tukey test is given by taking the absolute value of the difference between pairs of means and dividing it by the standard error of the mean (SE) as determined by a one-way ANOVA test.
- The SE is in turn the square root of (variance divided by sample size).

Two-factor ANOVA.

- ANOVA (Analysis of Variance) is a statistical test used to analyze the difference between the means of more than two groups.
- A two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables.
- The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors).

- The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.
- For example, could use a two-way ANOVA to understand whether there is an interaction between gender and educational level on test anxiety amongst university students, where gender (males/females) and education level (undergraduate/postgraduate) are independent variables, and test anxiety is dependent variable.
- Alternately, may want to determine whether there is an interaction between physical activity level and gender on blood cholesterol concentration in children, where physical activity (low/moderate/high) and gender (male/female) are independent variables, and cholesterol concentration is dependent variable.
- The interaction term in a two-way ANOVA informs whether the effect of one of independent variables on the dependent variable is the same for all values of your other independent variable (and vice versa).

How does the ANOVA test work?

- ANOVA tests for significance using the F test for statistical significance.
- The F test is a groupwise comparison test, which means it compares the variance in each group mean to the overall variance in the dependent variable.
- If the variance within groups is smaller than the variance between groups, the F test will find a higher F value, and therefore a higher likelihood that the difference observed is real and not due to chance.

- A two-way ANOVA with interaction tests three null hypotheses at the same time:
    - There is no difference in group means at any level of the first independent variable.
    - There is no difference in group means at any level of the second independent variable.
    - The effect of one independent variable does not depend on the effect of the other independent variable

15. (a) (i) In Statistics, what happens when the goodness of fit test score is low? (6)

Refer Unit 5 – Page No. 15

(ii) Given the following dataset of employee, Using regression analysis, find the expected salary of an employee if the age is 45. (7)

Age	Salary
54	67000
42	43000
49	55000
57	71000
35	25000

OR

(b) (i) Define autocorrelation and how is it calculated? What does the negative correlation convey? (6)

Refer Unit 5 – Page No. 38

(ii) What is the philosophy of Logistic regression? What kind of model it is? What does logistic Regression predict? Tabulate the cardinal differences of Linear and Logistic Regression. (7)

Refer Unit 5 – Page No. 21

Tabulate the cardinal differences of Linear and Logistic Regression.

Linear Regression	Logistic Regression
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
Linear regression, predict the value of continuous variables.	logistic Regression, predict the values of categorical variables.
linear regression, find the best fit line, by which we can easily predict the output.	Logistic Regression, find the S-curve by which we can classify the samples.

Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.
The output for Linear Regression must be a continuous value, such as price, age, etc.	The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc.
In Linear regression, it is required that relationship between dependent variable and independent variable must be linear.	In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable.
In linear regression, there may be collinearity between the independent variables.	In logistic regression, there should not be collinearity between the independent variable.

PART C — (1 x 15 = 15 marks)

16. (a) (i) Discuss the role of sampling distribution in inferential statistics. (8)

Refer Unit 3 – Page No. 15

(ii) Indicate whether each of the following distributions is positively or negatively skewed. The distribution of

- **Incomes of taxpayers have a mean of $48,000 and a median of $43,000.**
  Mean (48,000) > Median (43,000) Positively Skewed
- **GPAs for all students at some college have a mean of 3,01 and a median of 3.20.**
  Mean (3.01) < Median (3.20) Negatively Skewed
- **Daily TV viewing times for preschool children has 'a mean of 55 minutes and a median of 73 minutes. (7)**
  Mean (55 minutes) < Median (73 minutes) Negatively Skewed

OR

(b) (i) Assume that we have a stream of items of large and unknown length that we can only iterate over once. Devise an effective sampling algorithm that randomly chooses an item from this stream such that each item is equally likely to be selected. (8)

Reservoir Sampling

Let us assume we have to sample 5 objects out of an infinite stream such that each element has an equal probability of getting selected.

import randomdef generator(max):

number = 1

while number < max:

```
number += 1
yield number# Create as stream generator
stream = generator(10000)# Doing Reservoir Sampling from the stream
k=5
reservoir = []
for i, element in enumerate(stream):
if i+1<= k:
reservoir.append(element)
else:
probability = k/(i+1)
if random.random() < probability:
# Select item in stream and remove one of the k items already selected
reservoir[random.choice(range(0,k))] = elementprint(reservoir)
```

[1369, 4108, 9986, 828, 5589]

So, let us think of a stream of only 3 items, and we have to keep 2 of them.

We see the first item, and we hold it in the list as our reservoir has space.

We see the second item, and we hold it in the list as our reservoir has space.

We see the third item. We choose the third item to be in the list with probability 2/3.

Let us now see the probability of first item getting selected:

The probability of removing the first item is the probability of element 3 getting selected multiplied by the probability of Element 1 getting randomly chosen as the replacement candidate from the 2 elements in the reservoir.

That probability is:

$$2/3*1/2 = 1/3$$

Thus the probability of 1 getting selected is:

$$1–1/3 = 2/3$$

We can have the exact same argument for the Second Element and we can extend it for many elements.

Thus each item has the same probability of getting selected: 2/3 or in general k/n.

Appendix

A

Durbin-Watson Significance Tables

The Durbin-Watson test statistic tests the null hypothesis that the residuals from an ordinary least-squares regression are not autocorrelated against the alternative that the residuals follow an AR1 process. The Durbin-Watson statistic ranges in value from 0 to 4. A value near 2 indicates non-autocorrelation; a value toward 0 indicates positive autocorrelation; a value toward 4 indicates negative autocorrelation.

Because of the dependence of any computed Durbin-Watson value on the associated data matrix, exact critical values of the Durbin-Watson statistic are not tabulated for all possible cases. Instead, Durbin and Watson established upper and lower bounds for the critical values. Typically, tabulated bounds are used to test the hypothesis of zero autocorrelation against the alternative of *positive* first-order autocorrelation, since positive autocorrelation is seen much more frequently in practice than negative autocorrelation. To use the table, you must cross-reference the sample size against the number of regressors, excluding the constant from the count of the number of regressors.

The conventional Durbin-Watson tables are not applicable when you do not have a constant term in the regression. Instead, you must refer to an appropriate set of Durbin-Watson tables. The conventional Durbin-Watson tables are also not applicable when a lagged dependent variable appears among the regressors. Durbin has proposed alternative test procedures for this case.

Statisticians have compiled Durbin-Watson tables from some special cases, including:

- Regressions with a full set of quarterly seasonal dummies.
- Regressions with an intercept and a linear trend variable (CURVEFIT MODEL=LINEAR).
- Regressions with a full set of quarterly seasonal dummies and a linear trend variable.

In addition to obtaining the Durbin-Watson statistic for residuals from REGRESSION, you should also plot the ACF and PACF of the residuals series. The plots might suggest either that the residuals are random, or that they follow some ARMA process. If the residuals resemble an AR1 process, you can estimate an appropriate regression using the AREG procedure. If the residuals follow any ARMA process, you can estimate an appropriate regression using the ARIMA procedure.

In this appendix, we have reproduced two sets of tables. Savin and White (1977) present tables for sample sizes ranging from 6 to 200 and for 1 to 20 regressors for models in which an intercept is included. Farebrother (1980) presents tables for sample sizes ranging from 2 to 200 and for 0 to 21 regressors for models in which an intercept is not included.

Let's consider an example of how to use the tables. In Chapter 9, we look at the classic Durbin and Watson data set concerning consumption of spirits. The sample size is 69, there are 2 regressors, and there is an intercept term in the model. The Durbin-Watson test statistic value is 0.24878. We want to test the null hypothesis of zero autocorrelation in the residuals against the alternative that the residuals are positively autocorrelated at the 1% level of significance. If you examine the Savin and White tables (Table A.2 and Table A.3), you will not find a row for sample size 69, so go to the next *lowest* sample size with a tabulated row, namely *N=65*. Since there are two regressors, find the column labeled *k=2*. Cross-referencing the indicated row and column, you will find that the printed bounds are dL = 1.377 and dU = 1.500. If the observed value of the test statistic is less than the tabulated lower bound, then you should reject the null hypothesis of non-autocorrelated errors in favor of the hypothesis of positive first-order autocorrelation. Since 0.24878 is less than 1.377, we reject the null hypothesis. If the test statistic value were greater than dU, we would not reject the null hypothesis.

A third outcome is also possible. If the test statistic value lies between dL and dU, the test is inconclusive. In this context, you might err on the side of conservatism and not reject the null hypothesis.

For models with an intercept, if the observed test statistic value is greater than 2, then you want to test the null hypothesis against the alternative hypothesis of negative first-order autocorrelation. To do this, compute the quantity 4-d and compare this value with the tabulated values of dL and dU as if you were testing for positive autocorrelation.

When the regression does not contain an intercept term, refer to Farebrother,Äôs tabulated values of the ,Äúminimal bound,,Äù denoted dM (Table A.4 and Table A.5), instead of Savin and White,Äôs lower bound dL. In this instance, the upper bound is

the conventional bound dU found in the Savin and White tables. To test for negative first-order autocorrelation, use Table A.6 and Table A.7.

To continue with our example, had we run a regression with no intercept term, we would cross-reference N equals 65 and k equals 2 in Farebrother,Äôs table. The tabulated 1% minimal bound is 1.348.

Table A-1

Models with an intercept (from Savin and White)

Durbin-Watson Statistic: 1 Per Cent Significance Points of dL and dU

	k'*=1		k'=2		k'=3		k'=4		k'=5		k'=6		k'=7		k'=8		k'=9		k'=10	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.390	1.142	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.435	1.036	0.294	1.676	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.497	1.003	0.345	1.489	0.229	2.102	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453	0.124	2.892	-----	-----	-----	-----	-----	-----	-----	-----
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.665	0.105	3.053	-----	-----	-----	-----	-----	-----
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838	0.090	3.182	-----	-----	-----	-----
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	0.183	2.667	0.122	2.981	0.078	3.287	-----	-----
15	0.811	1.070	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101	0.068	3.374
16	0.844	1.086	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944	0.094	3.201
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811	0.127	3.053
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.697	0.160	2.925
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597	0.196	2.813
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510	0.232	2.174
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434	0.268	2.625
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691	0.587	1.849	0.510	2.015	0.437	2.188	0.368	2.367	0.304	2.548
23	1.017	1.186	0.938	1.290	0.858	1.407	0.777	1.535	0.699	1.674	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308	0.340	2.479
24	1.037	1.199	0.959	1.298	0.881	1.407	0.805	1.527	0.728	1.659	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255	0.375	2.417
25	1.055	1.210	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645	0.682	1.776	0.610	1.915	0.540	2.059	0.473	2.209	0.409	2.362
26	1.072	1.222	1.000	1.311	0.928	1.410	0.855	1.517	0.782	1.635	0.711	1.759	0.640	1.889	0.572	2.026	0.505	2.168	0.441	2.313
27	1.088	1.232	1.019	1.318	0.948	1.413	0.878	1.514	0.808	1.625	0.738	1.743	0.669	1.867	0.602	1.997	0.536	2.131	0.473	2.269
28	1.104	1.244	1.036	1.325	0.969	1.414	0.901	1.512	0.832	1.618	0.764	1.729	0.696	1.847	0.630	1.970	0.566	2.098	0.504	2.229
29	1.119	1.254	1.053	1.332	0.988	1.418	0.921	1.511	0.855	1.611	0.788	1.718	0.723	1.830	0.658	1.947	0.595	2.068	0.533	2.193
30	1.134	1.264	1.070	1.339	1.006	1.421	0.941	1.510	0.877	1.606	0.812	1.707	0.748	1.814	0.684	1.925	0.622	2.041	0.562	2.160
31	1.147	1.274	1.085	1.345	1.022	1.425	0.960	1.509	0.897	1.601	0.834	1.698	0.772	1.800	0.710	1.906	0.649	2.017	0.589	2.131
32	1.160	1.283	1.100	1.351	1.039	1.428	0.978	1.509	0.917	1.597	0.856	1.690	0.794	1.788	0.734	1.889	0.674	1.995	0.615	2.104
33	1.171	1.291	1.114	1.358	1.055	1.432	0.995	1.510	0.935	1.594	0.876	1.683	0.816	1.776	0.757	1.874	0.698	1.975	0.641	2.080
34	1.184	1.298	1.128	1.364	1.070	1.436	1.012	1.511	0.954	1.591	0.896	1.677	0.837	1.766	0.779	1.860	0.722	1.957	0.665	2.057
35	1.195	1.307	1.141	1.370	1.085	1.439	1.028	1.512	0.971	1.589	0.914	1.671	0.857	1.757	0.800	1.847	0.744	1.940	0.689	2.037
36	1.205	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.987	1.587	0.932	1.666	0.877	1.749	0.821	1.836	0.766	1.925	0.711	2.018
37	1.217	1.322	1.164	1.383	1.112	1.446	1.058	1.514	1.004	1.585	0.950	1.662	0.895	1.742	0.841	1.825	0.787	1.911	0.733	2.001
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.584	0.966	1.658	0.913	1.735	0.860	1.816	0.807	1.899	0.754	1.985
39	1.237	1.337	1.187	1.392	1.137	1.452	1.085	1.517	1.033	1.583	0.982	1.655	0.930	1.729	0.878	1.807	0.826	1.887	0.774	1.970
40	1.246	1.344	1.197	1.398	1.149	1.456	1.098	1.518	1.047	1.583	0.997	1.652	0.946	1.724	0.895	1.799	0.844	1.876	0.749	1.956
45	1.288	1.376	1.245	1.424	1.201	1.474	1.156	1.528	1.111	1.583	1.065	1.643	1.019	1.704	0.974	1.768	0.927	1.834	0.881	1.902
50	1.324	1.403	1.285	1.445	1.245	1.491	1.206	1.537	1.164	1.587	1.123	1.639	1.081	1.692	1.039	1.748	0.997	1.805	0.955	1.864
55	1.356	1.428	1.320	1.466	1.284	1.505	1.246	1.548	1.209	1.592	1.172	1.638	1.134	1.685	1.095	1.734	1.057	1.785	1.018	1.837
60	1.382	1.449	1.351	1.484	1.317	1.520	1.283	1.559	1.248	1.598	1.214	1.639	1.179	1.682	1.144	1.726	1.108	1.771	1.072	1.817
65	1.407	1.467	1.377	1.500	1.346	1.534	1.314	1.568	1.283	1.604	1.251	1.642	1.218	1.680	1.186	1.720	1.153	1.761	1.120	1.802
70	1.429	1.485	1.400	1.514	1.372	1.546	1.343	1.577	1.313	1.611	1.283	1.645	1.253	1.680	1.223	1.716	1.192	1.754	1.162	1.792
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.586	1.340	1.617	1.313	1.649	1.284	1.682	1.256	1.714	1.227	1.748	1.199	1.783
80	1.465	1.514	1.440	1.541	1.416	1.568	1.390	1.595	1.364	1.624	1.338	1.653	1.312	1.683	1.285	1.714	1.259	1.745	1.232	1.777
85	1.481	1.529	1.458	1.553	1.434	1.577	1.411	1.603	1.386	1.630	1.362	1.657	1.337	1.685	1.312	1.714	1.287	1.743	1.262	1.773
90	1.496	1.541	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636	1.383	1.661	1.360	1.687	1.336	1.714	1.312	1.741	1.288	1.769
95	1.510	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.641	1.403	1.666	1.381	1.690	1.358	1.715	1.336	1.741	1.313	1.767
100	1.522	1.562	1.502	1.582	1.482	1.604	1.461	1.625	1.441	1.647	1.421	1.670	1.400	1.693	1.378	1.717	1.357	1.741	1.335	1.765
150	1.611	1.637	1.598	1.651	1.584	1.665	1.571	1.679	1.557	1.693	1.543	1.708	1.530	1.722	1.515	1.737	1.501	1.752	1.486	1.767
200	1.664	1.684	1.653	1.693	1.643	1.704	1.633	1.715	1.623	1.725	1.613	1.735	1.603	1.746	1.592	1.757	1.582	1.768	1.571	1.779

*k' is the number of regressors excluding the intercept

	k'*=11		**k'=12**		**k'=13**		**k'=14**		**k'=15**		**k'=16**		**k'=17**		**k'=18**		**k'=19**		**k'=20**	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
16	0.060	3.446	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
17	0.084	3.286	0.053	3.506	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
18	0.113	3.146	0.075	3.358	0.047	3.557	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
19	0.145	3.023	0.102	3.227	0.067	3.420	0.043	3.601	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
20	0.178	2.914	0.131	3.109	0.092	3.297	0.061	3.474	0.038	3.639	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
21	0.212	2.817	0.162	3.004	0.119	3.185	0.084	3.358	0.055	3.521	0.035	3.671	-----	-----	-----	-----	-----	-----	-----	-----
22	0.246	2.729	0.194	2.909	0.148	3.084	0.109	3.252	0.077	3.412	0.050	3.562	0.032	3.700	-----	-----	-----	-----	-----	-----
23	0.281	2.651	0.227	2.822	0.178	2.991	0.136	3.155	0.100	3.311	0.070	3.459	0.046	3.597	0.029	3.725	-----	-----	-----	-----
24	0.315	2.580	0.260	2.744	0.209	2.906	0.165	3.065	0.125	3.218	0.092	3.363	0.065	3.501	0.043	3.629	0.027	3.747	-----	-----
25	0.348	2.517	0.292	2.674	0.240	2.829	0.194	2.982	0.152	3.131	0.116	3.274	0.085	3.410	0.060	3.538	0.039	3.657	0.025	3.766
26	0.381	2.460	0.324	2.610	0.272	2.758	0.224	2.906	0.180	3.050	0.141	3.191	0.107	3.325	0.079	3.452	0.055	3.572	0.036	3.682
27	0.413	2.409	0.356	2.552	0.303	2.694	0.253	2.836	0.208	2.976	0.167	3.113	0.131	3.245	0.100	3.371	0.073	3.490	0.051	3.602
28	0.444	2.363	0.387	2.499	0.333	2.635	0.283	2.772	0.237	2.907	0.194	3.040	0.156	3.169	0.122	3.294	0.093	3.412	0.068	3.524
29	0.474	2.321	0.417	2.451	0.363	2.582	0.313	2.713	0.266	2.843	0.222	2.972	0.182	3.098	0.146	3.220	0.114	3.338	0.087	3.450
30	0.503	2.283	0.447	2.407	0.393	2.533	0.342	2.659	0.294	2.785	0.249	2.909	0.208	3.032	0.171	3.152	0.137	3.267	0.107	3.379
31	0.531	2.248	0.475	2.367	0.422	2.487	0.371	2.609	0.322	2.730	0.277	2.851	0.234	2.970	0.193	3.087	0.160	3.201	0.128	3.311
32	0.558	2.216	0.503	2.330	0.450	2.446	0.399	2.563	0.350	2.680	0.304	2.797	0.261	2.912	0.221	3.026	0.184	3.137	0.151	3.246
33	0.585	2.187	0.530	2.296	0.477	2.408	0.426	2.520	0.377	2.633	0.331	2.746	0.287	2.858	0.246	2.969	0.209	3.078	0.174	3.184
34	0.610	2.160	0.556	2.266	0.503	2.373	0.452	2.481	0.404	2.590	0.357	2.699	0.313	2.808	0.272	2.915	0.233	3.022	0.197	3.126
35	0.634	2.136	0.581	2.237	0.529	2.340	0.478	2.444	0.430	2.550	0.383	2.655	0.339	2.761	0.297	2.865	0.257	2.969	0.221	3.071
36	0.658	2.113	0.605	2.210	0.554	2.310	0.504	2.410	0.455	2.512	0.409	2.614	0.364	2.717	0.322	2.818	0.282	2.919	0.244	3.019
37	0.680	2.092	0.628	2.186	0.578	2.282	0.528	2.379	0.480	2.477	0.434	2.576	0.389	2.675	0.347	2.774	0.306	2.872	0.268	2.969
38	0.702	2.073	0.651	2.164	0.601	2.256	0.552	2.350	0.504	2.445	0.458	2.540	0.414	2.637	0.371	2.733	0.330	2.828	0.291	2.923
39	0.723	2.055	0.673	2.143	0.623	2.232	0.575	2.323	0.528	2.414	0.482	2.507	0.438	2.600	0.395	2.694	0.354	2.787	0.315	2.879
40	0.744	2.039	0.694	2.123	0.645	2.210	0.597	2.297	0.551	2.386	0.505	2.476	0.461	2.566	0.418	2.657	0.377	2.748	0.338	2.838
45	0.835	1.972	0.790	2.044	0.744	2.118	0.700	2.193	0.655	2.269	0.612	2.346	0.570	2.424	0.528	2.503	0.488	2.582	0.448	2.661
50	0.913	1.925	0.871	1.987	0.829	2.051	0.787	2.116	0.746	2.182	0.705	2.250	0.665	2.318	0.625	2.387	0.586	2.456	0.548	2.526
55	0.979	1.891	0.940	1.945	0.902	2.002	0.863	2.059	0.825	2.117	0.786	2.176	0.748	2.237	0.711	2.298	0.674	2.359	0.637	2.421
60	1.037	1.865	1.001	1.914	0.965	1.964	0.929	2.015	0.893	2.067	0.857	2.120	0.822	2.173	0.786	2.227	0.751	2.283	0.716	2.338
65	1.087	1.845	1.053	1.889	1.020	1.934	0.986	1.980	0.953	2.027	0.919	2.075	0.886	2.123	0.852	2.172	0.819	2.221	0.789	2.272
70	1.131	1.831	1.099	1.870	1.068	1.911	1.037	1.953	1.005	1.995	0.974	2.038	0.943	2.082	0.911	2.127	0.880	2.172	0.849	2.217
75	1.170	1.819	1.141	1.856	1.111	1.893	1.082	1.931	1.052	1.970	1.023	2.009	0.993	2.049	0.964	2.090	0.934	2.131	0.905	2.172
80	1.205	1.810	1.177	1.844	1.150	1.878	1.122	1.913	1.094	1.949	1.066	1.984	1.039	2.022	1.011	2.059	0.983	2.097	0.955	2.135
85	1.236	1.803	1.210	1.834	1.184	1.866	1.158	1.898	1.132	1.931	1.106	1.965	1.080	1.999	1.053	2.033	1.027	2.068	1.000	2.104
90	1.264	1.798	1.240	1.827	1.215	1.856	1.191	1.886	1.166	1.917	1.141	1.948	1.116	1.979	1.091	2.012	1.066	2.044	1.041	2.077
95	1.290	1.793	1.267	1.821	1.244	1.848	1.221	1.876	1.197	1.905	1.174	1.943	1.150	1.963	1.126	1.993	1.102	2.023	1.079	2.054
100	1.314	1.790	1.292	1.816	1.270	1.841	1.248	1.868	1.225	1.895	1.203	1.922	1.181	1.949	1.158	1.977	1.136	2.006	1.113	2.034
150	1.473	1.783	1.458	1.799	1.444	1.814	1.429	1.830	1.414	1.847	1.400	1.863	1.385	1.880	1.370	1.897	1.355	1.913	1.340	1.931
200	1.561	1.791	1.550	1.801	1.539	1.813	1.528	1.824	1.518	1.836	1.507	1.847	1.495	1.860	1.484	1.871	1.474	1.883	1.462	1.896

*k' is the number of regressors excluding the intercept

Table A-2
Models with an intercept (from Savin and White)

Durbin-Watson Statistic: 5 Per Cent Significance Points of dL and dU

	k'*=1		k'=2		k'=3		k'=4		k'=5		k'=6		k'=7		k'=8		k'=9		k'=10	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.610	1.400	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.700	1.356	0.467	1.896	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.763	1.332	0.559	1.777	0.367	2.287	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004	-----	-----	-----	-----	-----	-----	-----	-----
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149	-----	-----	-----	-----	-----	-----
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266	-----	-----	-----	-----
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	-----	-----
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.750	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.013	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.959	0.874	2.071	0.798	2.188	0.723	2.309	0.649	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.681	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.796	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.079	1.891	1.015	1.978	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.876	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.197
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.838	1.369	1.874	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.747	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.846	1.608	1.862	1.593	1.877
200	1.758	1.779	1.748	1.789	1.738	1.799	1.728	1.809	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

*k' is the number of regressors excluding the intercept

	k'*=11		k'=12		k'=13		k'=14		k'=15		k'=16		k'=17		k'=18		k'=19		k'=20	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
16	0.098	3.503	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
17	0.138	3.378	0.087	3.557	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
18	0.177	3.265	0.123	3.441	0.078	3.603	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
19	0.220	3.159	0.160	3.335	0.111	3.496	0.070	3.642	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
20	0.263	3.063	0.200	3.234	0.145	3.395	0.100	3.542	0.063	3.676	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
21	0.307	2.976	0.240	3.141	0.182	3.300	0.132	3.448	0.091	3.583	0.058	3.705	-----	-----	-----	-----	-----	-----	-----	-----
22	0.349	2.897	0.281	3.057	0.220	3.211	0.166	3.358	0.120	3.495	0.083	3.619	0.052	3.731	-----	-----	-----	-----	-----	-----
23	0.391	2.826	0.322	2.979	0.259	3.128	0.202	3.272	0.153	3.409	0.110	3.535	0.076	3.650	0.048	3.753	-----	-----	-----	-----
24	0.431	2.761	0.362	2.908	0.297	3.053	0.239	3.193	0.186	3.327	0.141	3.454	0.101	3.572	0.070	3.678	0.044	3.773	-----	-----
25	0.470	2.702	0.400	2.844	0.335	2.983	0.275	3.119	0.221	3.251	0.172	3.376	0.130	3.494	0.094	3.604	0.065	3.702	0.041	3.790
26	0.508	2.649	0.438	2.784	0.373	2.919	0.312	3.051	0.256	3.179	0.205	3.303	0.160	3.420	0.120	3.531	0.087	3.632	0.060	3.724
27	0.544	2.600	0.475	2.730	0.409	2.859	0.348	2.987	0.291	3.112	0.238	3.233	0.191	3.349	0.149	3.460	0.112	3.563	0.081	3.658
28	0.578	2.555	0.510	2.680	0.445	2.805	0.383	2.928	0.325	3.050	0.271	3.168	0.222	3.283	0.178	3.392	0.138	3.495	0.104	3.592
29	0.612	2.515	0.544	2.634	0.479	2.755	0.418	2.874	0.359	2.992	0.305	3.107	0.254	3.219	0.208	3.327	0.166	3.431	0.129	3.528
30	0.643	2.477	0.577	2.592	0.512	2.708	0.451	2.823	0.392	2.937	0.337	3.050	0.286	3.160	0.238	3.266	0.195	3.368	0.156	3.465
31	0.674	2.443	0.608	2.553	0.545	2.665	0.484	2.776	0.425	2.887	0.370	2.996	0.317	3.103	0.269	3.208	0.224	3.309	0.183	3.406
32	0.703	2.411	0.638	2.517	0.576	2.625	0.515	2.733	0.457	2.840	0.401	2.946	0.349	3.050	0.299	3.153	0.253	3.252	0.211	3.348
33	0.731	2.382	0.668	2.484	0.606	2.588	0.546	2.692	0.488	2.796	0.432	2.899	0.379	3.000	0.329	3.100	0.283	3.198	0.239	3.293
34	0.758	2.355	0.695	2.454	0.634	2.554	0.575	2.654	0.518	2.754	0.462	2.854	0.409	2.954	0.359	3.051	0.312	3.147	0.267	3.240
35	0.783	2.330	0.722	2.425	0.662	2.521	0.604	2.619	0.547	2.716	0.492	2.813	0.439	2.910	0.388	3.005	0.340	3.099	0.295	3.190
36	0.808	2.306	0.748	2.398	0.689	2.492	0.631	2.586	0.575	2.680	0.520	2.774	0.467	2.868	0.417	2.961	0.369	3.053	0.323	3.142
37	0.831	2.285	0.772	2.374	0.714	2.464	0.657	2.555	0.602	2.646	0.548	2.738	0.495	2.829	0.445	2.920	0.397	3.009	0.351	3.097
38	0.854	2.265	0.796	2.351	0.739	2.438	0.683	2.526	0.628	2.614	0.575	2.703	0.522	2.792	0.472	2.880	0.424	2.968	0.378	3.054
39	0.875	2.246	0.819	2.329	0.763	2.413	0.707	2.499	0.653	2.585	0.600	2.671	0.549	2.757	0.499	2.843	0.451	2.929	0.404	3.013
40	0.896	2.228	0.840	2.309	0.785	2.391	0.731	2.473	0.678	2.557	0.626	2.641	0.575	2.724	0.525	2.808	0.477	2.829	0.430	2.974
45	0.988	2.156	0.938	2.225	0.887	2.296	0.838	2.367	0.788	2.439	0.740	2.512	0.692	2.586	0.644	2.659	0.598	2.733	0.553	2.807
50	1.064	2.103	1.019	2.163	0.973	2.225	0.927	2.287	0.882	2.350	0.836	2.414	0.792	2.479	0.747	2.544	0.703	2.610	0.660	2.675
55	1.129	2.062	1.087	2.116	1.045	2.170	1.003	2.225	0.961	2.281	0.919	2.338	0.877	2.396	0.836	2.454	0.795	2.512	0.754	2.571
60	1.184	2.031	1.145	2.079	1.106	2.127	1.068	2.177	1.029	2.227	0.990	2.278	0.951	2.330	0.913	2.382	0.874	2.434	0.836	2.487
65	1.231	2.006	1.195	2.049	1.160	2.093	1.124	2.138	1.088	2.183	1.052	2.229	1.016	2.276	0.980	2.323	0.944	2.371	0.908	2.419
70	1.272	1.987	1.239	2.026	1.206	2.066	1.172	2.106	1.139	2.148	1.105	2.189	1.072	2.232	1.038	2.275	1.005	2.318	0.971	2.362
75	1.308	1.970	1.277	2.006	1.247	2.043	1.215	2.080	1.184	2.118	1.153	2.156	1.121	2.195	1.090	2.235	1.058	2.275	1.027	2.315
80	1.340	1.957	1.311	1.991	1.283	2.024	1.253	2.059	1.224	2.093	1.195	2.129	1.165	2.165	1.136	2.201	1.106	2.238	1.076	2.275
85	1.369	1.946	1.342	1.977	1.315	2.009	1.287	2.040	1.260	2.073	1.232	2.105	1.205	2.139	1.177	2.172	1.149	2.206	1.121	2.241
90	1.395	1.937	1.369	1.966	1.344	1.995	1.318	2.025	1.292	2.055	1.266	2.085	1.240	2.116	1.213	2.148	1.187	2.179	1.160	2.211
95	1.418	1.930	1.394	1.956	1.370	1.984	1.345	2.012	1.321	2.040	1.296	2.068	1.271	2.097	1.247	2.126	1.222	2.156	1.197	2.186
100	1.439	1.923	1.416	1.948	1.393	1.974	1.371	2.000	1.347	2.026	1.324	2.053	1.301	2.080	1.277	2.108	1.253	2.135	1.229	2.164
150	1.579	1.892	1.564	1.908	1.550	1.924	1.535	1.940	1.519	1.956	1.504	1.972	1.489	1.989	1.474	2.006	1.458	2.023	1.443	2.040
200	1.654	1.885	1.643	1.896	1.632	1.908	1.621	1.919	1.610	1.931	1.599	1.943	1.588	1.955	1.576	1.967	1.565	1.979	1.554	1.991

*K' is the number of regressors excluding the intercept

Table A-3

Models with no intercept (from Farebrother): Positive serial correlation

Durbin-Watson One Per Cent Minimal Bound

N	K=0	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10	K=11	K=12	K=13	K=14	K=15	K=16	K=17	K=18	K=19	K=20	K=21
2	0.001	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
3	0.034	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
4	0.127	0.022	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
5	0.233	0.089	0.014	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
6	0.322	0.175	0.065	0.010	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.398	0.253	0.135	0.049	0.008	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.469	0.324	0.202	0.106	0.038	0.006	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.534	0.394	0.268	0.164	0.086	0.031	0.005	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.591	0.457	0.333	0.223	0.136	0.070	0.025	0.004	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.643	0.515	0.394	0.284	0.189	0.114	0.059	0.021	0.003	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
12	0.691	0.568	0.451	0.341	0.244	0.161	0.097	0.050	0.018	0.003	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
13	0.733	0.617	0.503	0.396	0.298	0.212	0.139	0.083	0.043	0.015	0.002	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
14	0.773	0.662	0.552	0.448	0.350	0.262	0.185	0.121	0.072	0.037	0.013	0.002	0.000	-----	-----	-----	-----	-----	-----	-----	-----	-----
15	0.809	0.703	0.598	0.496	0.400	0.311	0.232	0.163	0.107	0.063	0.032	0.011	0.002	0.000	-----	-----	-----	-----	-----	-----	-----	-----
16	0.842	0.741	0.640	0.541	0.447	0.358	0.278	0.206	0.145	0.094	0.056	0.028	0.010	0.002	0.000	-----	-----	-----	-----	-----	-----	-----
17	0.873	0.776	0.679	0.583	0.491	0.404	0.323	0.249	0.184	0.129	0.084	0.050	0.025	0.009	0.001	0.000	-----	-----	-----	-----	-----	-----
18	0.901	0.808	0.715	0.623	0.533	0.447	0.366	0.292	0.225	0.166	0.116	0.075	0.044	0.023	0.008	0.001	0.000	-----	-----	-----	-----	-----
19	0.928	0.839	0.749	0.660	0.572	0.488	0.408	0.333	0.265	0.204	0.150	0.105	0.068	0.040	0.020	0.007	0.001	0.000	-----	-----	-----	-----
20	0.952	0.867	0.780	0.694	0.609	0.527	0.448	0.374	0.304	0.241	0.185	0.136	0.095	0.062	0.036	0.018	0.006	0.001	0.000	-----	-----	-----
21	0.976	0.893	0.810	0.727	0.644	0.564	0.486	0.413	0.343	0.279	0.221	0.169	0.124	0.087	0.056	0.033	0.017	0.006	0.001	0.000	-----	-----
22	0.997	0.918	0.838	0.757	0.677	0.599	0.523	0.450	0.381	0.316	0.257	0.203	0.155	0.114	0.079	0.051	0.030	0.015	0.005	0.001	0.000	-----
23	1.018	0.942	0.864	0.786	0.709	0.632	0.558	0.486	0.417	0.352	0.292	0.237	0.187	0.143	0.104	0.073	0.047	0.027	0.014	0.005	0.001	0.000
24	1.037	0.964	0.889	0.813	0.738	0.664	0.591	0.520	0.452	0.387	0.327	0.270	0.219	0.172	0.131	0.096	0.067	0.043	0.025	0.013	0.004	0.001
25	1.056	0.984	0.912	0.839	0.766	0.693	0.622	0.553	0.486	0.421	0.361	0.304	0.251	0.203	0.160	0.122	0.089	0.062	0.040	0.023	0.012	0.004
26	1.073	1.004	0.934	0.863	0.792	0.722	0.652	0.584	0.518	0.454	0.394	0.336	0.283	0.233	0.189	0.148	0.113	0.083	0.057	0.037	0.022	0.011
27	1.089	1.023	0.955	0.886	0.817	0.749	0.681	0.614	0.549	0.486	0.426	0.368	0.314	0.264	0.218	0.176	0.138	0.105	0.077	0.053	0.034	0.020
28	1.105	1.040	0.974	0.908	0.841	0.774	0.708	0.643	0.579	0.517	0.457	0.400	0.345	0.294	0.247	0.204	0.164	0.129	0.098	0.071	0.050	0.032
29	1.120	1.057	0.993	0.929	0.864	0.798	0.734	0.670	0.607	0.546	0.487	0.430	0.376	0.324	0.276	0.232	0.191	0.154	0.120	0.091	0.067	0.046
30	1.134	1.073	1.011	0.948	0.885	0.822	0.759	0.696	0.635	0.574	0.516	0.460	0.405	0.354	0.305	0.260	0.217	0.179	0.144	0.113	0.086	0.062
31	1.147	1.088	1.028	0.967	0.905	0.844	0.782	0.721	0.661	0.602	0.544	0.488	0.434	0.383	0.334	0.288	0.244	0.205	0.168	0.135	0.106	0.080
32	1.160	1.103	1.044	0.985	0.925	0.865	0.805	0.745	0.686	0.628	0.571	0.516	0.462	0.411	0.362	0.315	0.271	0.230	0.193	0.158	0.127	0.100
33	1.173	1.117	1.060	1.002	0.944	0.885	0.826	0.768	0.710	0.653	0.597	0.542	0.489	0.438	0.389	0.342	0.298	0.256	0.218	0.182	0.149	0.120
34	1.185	1.130	1.075	1.018	0.961	0.904	0.847	0.790	0.733	0.677	0.622	0.568	0.516	0.465	0.416	0.369	0.324	0.282	0.243	0.206	0.172	0.141
35	1.196	1.143	1.089	1.034	0.978	0.923	0.867	0.811	0.755	0.700	0.646	0.593	0.541	0.491	0.442	0.395	0.350	0.308	0.268	0.230	0.195	0.163
36	1.207	1.155	1.102	1.049	0.995	0.940	0.886	0.831	0.777	0.723	0.669	0.617	0.566	0.516	0.467	0.421	0.376	0.333	0.292	0.254	0.218	0.185
37	1.217	1.167	1.116	1.063	1.010	0.957	0.904	0.850	0.797	0.744	0.692	0.640	0.590	0.540	0.492	0.446	0.401	0.358	0.317	0.278	0.241	0.207
38	1.228	1.178	1.128	1.077	1.026	0.974	0.921	0.869	0.817	0.765	0.713	0.663	0.613	0.564	0.516	0.470	0.425	0.382	0.341	0.302	0.265	0.230
39	1.237	1.189	1.140	1.090	1.040	0.989	0.938	0.887	0.836	0.785	0.734	0.684	0.635	0.587	0.540	0.494	0.449	0.406	0.365	0.325	0.288	0.252
40	1.247	1.200	1.152	1.103	1.054	1.004	0.954	0.904	0.854	0.804	0.754	0.705	0.657	0.609	0.562	0.517	0.473	0.430	0.388	0.349	0.311	0.275
45	1.289	1.247	1.204	1.160	1.116	1.071	1.026	0.981	0.936	0.890	0.845	0.800	0.755	0.710	0.666	0.623	0.581	0.539	0.499	0.459	0.421	0.384
50	1.325	1.287	1.248	1.208	1.168	1.128	1.087	1.046	1.004	0.963	0.921	0.880	0.838	0.797	0.756	0.715	0.675	0.636	0.597	0.559	0.521	0.485
55	1.356	1.321	1.286	1.250	1.213	1.176	1.139	1.101	1.063	1.025	0.987	0.948	0.910	0.872	0.833	0.796	0.758	0.721	0.684	0.647	0.611	0.576
60	1.383	1.351	1.319	1.285	1.252	1.218	1.183	1.149	1.114	1.078	1.043	1.008	0.972	0.936	0.901	0.865	0.830	0.795	0.760	0.725	0.691	0.657
65	1.408	1.378	1.348	1.317	1.286	1.254	1.222	1.190	1.158	1.125	1.092	1.059	1.026	0.993	0.960	0.927	0.894	0.861	0.828	0.795	0.762	0.730
70	1.429	1.401	1.373	1.345	1.316	1.286	1.257	1.227	1.197	1.166	1.136	1.105	1.074	1.043	1.012	0.981	0.950	0.919	0.888	0.857	0.826	0.795
75	1.448	1.423	1.396	1.369	1.342	1.315	1.287	1.260	1.231	1.203	1.174	1.146	1.117	1.088	1.058	1.029	1.000	0.971	0.941	0.912	0.883	0.854
80	1.466	1.442	1.417	1.392	1.367	1.341	1.315	1.289	1.262	1.236	1.209	1.182	1.155	1.127	1.100	1.072	1.045	1.017	0.989	0.962	0.934	0.907
85	1.482	1.459	1.436	1.412	1.388	1.364	1.340	1.315	1.290	1.265	1.240	1.214	1.189	1.163	1.137	1.111	1.085	1.059	1.033	1.006	0.980	0.954
90	1.497	1.475	1.453	1.431	1.408	1.385	1.362	1.339	1.315	1.292	1.268	1.244	1.220	1.195	1.171	1.146	1.121	1.097	1.072	1.047	1.022	0.997
95	1.510	1.490	1.469	1.448	1.426	1.405	1.383	1.361	1.338	1.316	1.293	1.271	1.248	1.225	1.201	1.178	1.155	1.131	1.108	1.084	1.060	1.037
100	1.523	1.503	1.483	1.463	1.443	1.422	1.402	1.381	1.359	1.338	1.317	1.295	1.273	1.251	1.229	1.207	1.185	1.162	1.140	1.118	1.095	1.072
150	1.611	1.598	1.585	1.571	1.558	1.544	1.530	1.516	1.502	1.488	1.474	1.460	1.445	1.431	1.416	1.402	1.387	1.372	1.357	1.342	1.327	1.312
200	1.664	1.654	1.644	1.634	1.624	1.613	1.603	1.593	1.582	1.572	1.561	1.551	1.540	1.529	1.519	1.508	1.497	1.486	1.475	1.434	1.453	1.442

Table A-4

Models with no intercept (from Farebrother): Positive serial correlation

Durbin-Watson Five Per Cent Minimal Bound

N	K=0	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10	K=11	K=12	K=13	K=14	K=15	K=16	K=17	K=18	K=19	K=20	K=21
2	0.012	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
3	0.168	0.006	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
4	0.355	0.105	0.004	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
5	0.478	0.248	0.070	0.002	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
6	0.584	0.358	0.180	0.050	0.002	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.677	0.462	0.275	0.136	0.037	0.001	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.754	0.556	0.371	0.217	0.106	0.029	0.001	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.820	0.635	0.460	0.303	0.175	0.085	0.023	0.001	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.877	0.706	0.539	0.385	0.251	0.143	0.069	0.019	0.001	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.927	0.768	0.610	0.460	0.326	0.211	0.120	0.058	0.016	0.001	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
12	0.972	0.823	0.674	0.530	0.397	0.279	0.180	0.101	0.049	0.013	0.001	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
13	1.012	0.872	0.731	0.593	0.464	0.345	0.241	0.154	0.087	0.042	0.011	0.001	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
14	1.047	0.916	0.783	0.651	0.525	0.408	0.302	0.210	0.134	0.075	0.036	0.010	0.001	-----	-----	-----	-----	-----	-----	-----	-----	-----
15	1.079	0.955	0.829	0.704	0.583	0.467	0.361	0.266	0.185	0.118	0.066	0.031	0.008	0.001	-----	-----	-----	-----	-----	-----	-----	-----
16	1.109	0.992	0.872	0.752	0.635	0.523	0.418	0.322	0.237	0.164	0.104	0.058	0.028	0.007	0.000	-----	-----	-----	-----	-----	-----	-----
17	1.136	1.024	0.911	0.797	0.684	0.575	0.472	0.376	0.288	0.211	0.146	0.093	0.052	0.025	0.007	0.000	-----	-----	-----	-----	-----	-----
18	1.160	1.055	0.946	0.837	0.729	0.624	0.523	0.427	0.339	0.260	0.190	0.131	0.083	0.046	0.022	0.006	0.000	-----	-----	-----	-----	-----
19	1.183	1.082	0.979	0.875	0.771	0.669	0.570	0.476	0.388	0.307	0.235	0.171	0.118	0.075	0.041	0.020	0.005	0.000	-----	-----	-----	-----
20	1.204	1.108	1.010	0.910	0.810	0.711	0.615	0.523	0.436	0.354	0.280	0.213	0.156	0.107	0.067	0.037	0.018	0.005	0.000	-----	-----	-----
21	1.224	1.132	1.038	0.942	0.846	0.751	0.657	0.567	0.481	0.400	0.324	0.256	0.195	0.142	0.097	0.061	0.034	0.016	0.004	0.000	-----	-----
22	1.242	1.154	1.064	0.972	0.879	0.787	0.697	0.609	0.524	0.443	0.368	0.298	0.235	0.178	0.130	0.089	0.056	0.031	0.015	0.004	0.000	-----
23	1.259	1.175	1.088	1.000	0.911	0.822	0.734	0.648	0.565	0.485	0.410	0.339	0.274	0.216	0.164	0.119	0.081	0.051	0.028	0.014	0.004	0.000
24	1.275	1.194	1.111	1.026	0.940	0.854	0.769	0.685	0.604	0.525	0.450	0.380	0.314	0.254	0.199	0.151	0.110	0.075	0.047	0.026	0.012	0.003
25	1.290	1.212	1.132	1.050	0.967	0.884	0.802	0.720	0.641	0.563	0.489	0.419	0.353	0.291	0.235	0.184	0.140	0.101	0.069	0.044	0.024	0.011
26	1.304	1.229	1.152	1.073	0.993	0.913	0.833	0.753	0.676	0.600	0.527	0.457	0.390	0.328	0.271	0.218	0.171	0.130	0.094	0.064	0.040	0.022
27	1.318	1.245	1.171	1.094	1.017	0.940	0.862	0.785	0.709	0.635	0.563	0.493	0.427	0.365	0.306	0.252	0.203	0.159	0.120	0.087	0.060	0.037
28	1.330	1.260	1.188	1.115	1.040	0.965	0.889	0.815	0.741	0.668	0.597	0.529	0.463	0.400	0.341	0.286	0.236	0.190	0.148	0.112	0.081	0.055
29	1.342	1.275	1.205	1.134	1.062	0.989	0.916	0.843	0.770	0.699	0.630	0.562	0.497	0.435	0.376	0.320	0.268	0.221	0.177	0.139	0.105	0.076
30	1.354	1.288	1.221	1.152	1.082	1.011	0.940	0.869	0.799	0.729	0.661	0.595	0.530	0.468	0.409	0.353	0.301	0.252	0.207	0.166	0.130	0.098
31	1.365	1.301	1.236	1.169	1.101	1.033	0.964	0.895	0.826	0.758	0.691	0.626	0.562	0.501	0.442	0.386	0.333	0.283	0.237	0.195	0.156	0.122
32	1.375	1.313	1.250	1.185	1.120	1.053	0.986	0.919	0.852	0.785	0.720	0.653	0.593	0.532	0.474	0.418	0.364	0.314	0.267	0.223	0.183	0.147
33	1.385	1.325	1.264	1.201	1.137	1.072	1.007	0.942	0.876	0.811	0.747	0.684	0.623	0.563	0.504	0.449	0.395	0.344	0.297	0.252	0.211	0.173
34	1.394	1.336	1.277	1.216	1.153	1.091	1.027	0.963	0.900	0.836	0.774	0.712	0.651	0.592	0.534	0.479	0.425	0.374	0.326	0.280	0.238	0.199
35	1.403	1.347	1.289	1.230	1.169	1.108	1.046	0.984	0.922	0.860	0.799	0.738	0.678	0.620	0.563	0.508	0.455	0.404	0.355	0.309	0.266	0.225
36	1.412	1.357	1.301	1.243	1.184	1.125	1.064	1.004	0.943	0.883	0.823	0.763	0.705	0.647	0.591	0.536	0.483	0.432	0.384	0.337	0.293	0.252
37	1.420	1.367	1.312	1.256	1.199	1.141	1.082	1.023	0.964	0.905	0.846	0.787	0.730	0.673	0.618	0.564	0.511	0.460	0.412	0.365	0.321	0.279
38	1.428	1.376	1.323	1.268	1.212	1.156	1.099	1.041	0.983	0.925	0.868	0.811	0.754	0.698	0.644	0.590	0.538	0.488	0.439	0.392	0.347	0.305
39	1.436	1.385	1.333	1.280	1.225	1.170	1.114	1.058	1.002	0.945	0.889	0.833	0.778	0.723	0.669	0.616	0.564	0.514	0.466	0.419	0.374	0.331
40	1.443	1.394	1.343	1.291	1.238	1.184	1.130	1.075	1.020	0.965	0.909	0.854	0.800	0.746	0.693	0.641	0.590	0.540	0.492	0.445	0.400	0.357
45	1.476	1.432	1.387	1.341	1.294	1.246	1.197	1.148	1.099	1.049	1.000	0.950	0.900	0.851	0.802	0.753	0.706	0.658	0.612	0.567	0.523	0.480
50	1.504	1.464	1.424	1.382	1.340	1.297	1.253	1.209	1.164	1.120	1.075	1.029	0.984	0.939	0.894	0.849	0.804	0.760	0.717	0.674	0.631	0.590
55	1.528	1.492	1.455	1.417	1.379	1.340	1.300	1.260	1.219	1.179	1.138	1.096	1.055	1.013	0.972	0.930	0.889	0.848	0.807	0.766	0.726	0.687
60	1.549	1.516	1.482	1.447	1.412	1.376	1.340	1.303	1.266	1.229	1.191	1.153	1.115	1.077	1.038	1.000	0.962	0.923	0.885	0.847	0.810	0.772
65	1.568	1.537	1.505	1.474	1.441	1.408	1.375	1.341	1.307	1.272	1.238	1.202	1.167	1.132	1.096	1.061	1.025	0.989	0.953	0.918	0.882	0.847
70	1.584	1.555	1.526	1.497	1.467	1.436	1.405	1.374	1.342	1.310	1.278	1.245	1.213	1.180	1.147	1.113	1.080	1.047	1.013	0.980	0.947	0.914
75	1.599	1.572	1.545	1.517	1.489	1.461	1.432	1.403	1.373	1.344	1.313	1.283	1.253	1.222	1.191	1.160	1.129	1.098	1.066	1.035	1.004	0.972
80	1.612	1.587	1.561	1.536	1.509	1.483	1.456	1.429	1.401	1.373	1.345	1.317	1.288	1.259	1.230	1.201	1.172	1.143	1.113	1.084	1.054	1.025
85	1.624	1.600	1.576	1.552	1.527	1.502	1.477	1.452	1.426	1.400	1.373	1.347	1.320	1.293	1.266	1.238	1.211	1.183	1.155	1.128	1.100	1.072
90	1.635	1.613	1.590	1.567	1.544	1.520	1.497	1.472	1.448	1.423	1.399	1.373	1.348	1.323	1.297	1.271	1.245	1.219	1.193	1.167	1.141	1.114
95	1.645	1.624	1.603	1.581	1.559	1.537	1.514	1.491	1.468	1.445	1.422	1.398	1.374	1.350	1.326	1.301	1.277	1.252	1.227	1.202	1.177	1.152
100	1.654	1.634	1.614	1.593	1.573	1.551	1.530	1.508	1.487	1.465	1.442	1.420	1.397	1.374	1.352	1.328	1.305	1.282	1.258	1.235	1.211	1.187
150	1.720	1.706	1.693	1.679	1.666	1.652	1.638	1.624	1.609	1.595	1.580	1.566	1.551	1.536	1.521	1.506	1.491	1.476	1.461	1.445	1.430	1.414
200	1.759	1.748	1.738	1.728	1.718	1.708	1.697	1.687	1.676	1.666	1.655	1.644	1.633	1.622	1.611	1.600	1.589	1.578	1.567	1.556	1.544	1.533

Table A-5

Models with no intercept (from Farebrother): Negative serial correlation

Durbin-Watson Ninety Five Per Cent Minimal Bound

N	K=0	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10	K=11	K=12	K=13	K=14	K=15	K=16	K=17	K=18	K=19	K=20	K=21
2	1.988	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
3	2.761	0.994	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
4	2.871	1.836	0.582	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
5	2.857	2.178	1.267	0.380	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
6	2.844	2.320	1.655	0.917	0.266	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	2.828	2.398	1.871	1.283	0.690	0.197	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	2.805	2.453	2.008	1.521	1.017	0.537	0.151	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	2.783	2.483	2.110	1.687	1.251	0.823	0.429	0.120	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	2.762	2.501	2.181	1.816	1.427	1.044	0.678	0.350	0.097	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	2.742	2.511	2.231	1.913	1.569	1.218	0.881	0.567	0.291	0.080	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
12	2.723	2.516	2.268	1.987	1.682	1.364	1.049	0.752	0.481	0.245	0.068	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
13	2.705	2.518	2.296	2.044	1.771	1.484	1.193	0.911	0.649	0.413	0.210	0.058	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
14	2.688	2.517	2.316	2.090	1.843	1.582	1.316	1.051	0.797	0.565	0.358	0.181	0.050	-----	-----	-----	-----	-----	-----	-----	-----	-----
15	2.672	2.515	2.332	2.126	1.902	1.664	1.419	1.172	0.931	0.703	0.497	0.314	0.158	0.043	-----	-----	-----	-----	-----	-----	-----	-----
16	2.657	2.512	2.344	2.155	1.950	1.732	1.506	1.276	1.049	0.829	0.624	0.439	0.277	0.139	0.038	-----	-----	-----	-----	-----	-----	-----
17	2.644	2.508	2.353	2.179	1.990	1.789	1.580	1.367	1.153	0.944	0.743	0.557	0.391	0.246	0.124	0.034	-----	-----	-----	-----	-----	-----
18	2.631	2.504	2.359	2.199	2.024	1.838	1.644	1.445	1.244	1.045	0.852	0.669	0.501	0.351	0.220	0.110	0.030	-----	-----	-----	-----	-----
19	2.618	2.499	2.364	2.215	2.053	1.880	1.699	1.513	1.324	1.136	0.951	0.773	0.605	0.452	0.316	0.198	0.099	0.027	-----	-----	-----	-----
20	2.607	2.494	2.368	2.228	2.077	1.916	1.747	1.573	1.395	1.216	1.040	0.868	0.704	0.550	0.410	0.286	0.179	0.090	0.025	-----	-----	-----
21	2.596	2.489	2.370	2.239	2.098	1.947	1.789	1.625	1.457	1.289	1.120	0.955	0.796	0.644	0.502	0.373	0.260	0.162	0.081	0.022	-----	-----
22	2.585	2.484	2.372	2.249	2.116	1.974	1.825	1.671	1.513	1.353	1.193	1.034	0.880	0.731	0.591	0.460	0.341	0.238	0.148	0.074	0.020	-----
23	2.575	2.479	2.373	2.257	2.131	1.998	1.858	1.712	1.563	1.411	1.258	1.107	0.957	0.813	0.674	0.544	0.422	0.313	0.218	0.136	0.068	0.019
24	2.566	2.474	2.373	2.263	2.145	2.019	1.886	1.749	1.607	1.463	1.318	1.172	1.029	0.888	0.753	0.623	0.502	0.389	0.289	0.201	0.125	0.062
25	2.557	2.470	2.373	2.269	2.156	2.037	1.912	1.782	1.647	1.510	1.371	1.232	1.094	0.958	0.826	0.699	0.578	0.465	0.360	0.267	0.185	0.115
26	1.073	1.004	0.934	0.863	0.792	0.722	0.652	0.584	0.518	0.454	0.394	0.336	0.283	0.233	0.189	0.148	0.113	0.083	0.057	0.037	0.022	0.011
27	1.089	1.023	0.955	0.886	0.817	0.749	0.681	0.614	0.549	0.486	0.426	0.368	0.314	0.264	0.218	0.176	0.138	0.105	0.077	0.053	0.034	0.020
28	1.105	1.040	0.974	0.908	0.841	0.774	0.708	0.643	0.579	0.517	0.457	0.400	0.345	0.294	0.247	0.204	0.164	0.129	0.098	0.071	0.050	0.032
29	1.120	1.057	0.993	0.929	0.864	0.798	0.734	0.670	0.607	0.546	0.487	0.430	0.376	0.324	0.276	0.232	0.191	0.154	0.120	0.091	0.067	0.046
30	1.134	1.073	1.011	0.948	0.885	0.822	0.759	0.696	0.635	0.574	0.516	0.460	0.405	0.354	0.305	0.260	0.217	0.179	0.144	0.113	0.086	0.062
31	1.147	1.088	1.028	0.967	0.905	0.844	0.782	0.721	0.661	0.602	0.544	0.488	0.434	0.383	0.334	0.288	0.244	0.205	0.168	0.135	0.106	0.080
32	1.160	1.103	1.044	0.985	0.925	0.865	0.805	0.745	0.686	0.628	0.571	0.516	0.462	0.411	0.362	0.315	0.271	0.230	0.193	0.158	0.127	0.100
33	1.173	1.117	1.060	1.002	0.944	0.885	0.826	0.768	0.710	0.653	0.597	0.542	0.489	0.438	0.389	0.342	0.298	0.256	0.218	0.182	0.149	0.120
34	1.185	1.130	1.075	1.018	0.961	0.904	0.847	0.790	0.733	0.677	0.622	0.568	0.516	0.465	0.416	0.369	0.324	0.282	0.243	0.206	0.172	0.141
35	1.196	1.143	1.089	1.034	0.978	0.923	0.867	0.811	0.755	0.700	0.646	0.593	0.541	0.491	0.442	0.395	0.350	0.308	0.268	0.230	0.195	0.163
36	1.207	1.155	1.102	1.049	0.995	0.940	0.886	0.831	0.777	0.723	0.669	0.617	0.566	0.516	0.467	0.421	0.376	0.333	0.292	0.254	0.218	0.185
37	1.217	1.167	1.116	1.063	1.010	0.957	0.904	0.850	0.797	0.744	0.692	0.640	0.590	0.540	0.492	0.446	0.401	0.358	0.317	0.278	0.241	0.207
38	1.228	1.178	1.128	1.077	1.026	0.974	0.921	0.869	0.817	0.765	0.713	0.663	0.613	0.564	0.516	0.470	0.425	0.382	0.341	0.302	0.265	0.230
39	1.237	1.189	1.140	1.090	1.040	0.989	0.938	0.887	0.836	0.785	0.734	0.684	0.635	0.587	0.540	0.494	0.449	0.406	0.365	0.325	0.288	0.252
40	1.247	1.200	1.152	1.103	1.054	1.004	0.954	0.904	0.854	0.804	0.754	0.705	0.657	0.609	0.562	0.517	0.473	0.430	0.388	0.349	0.311	0.275
45	1.289	1.247	1.204	1.160	1.116	1.071	1.026	0.981	0.936	0.890	0.845	0.800	0.755	0.710	0.666	0.623	0.581	0.539	0.499	0.459	0.421	0.384
50	1.325	1.287	1.248	1.208	1.168	1.128	1.087	1.046	1.004	0.963	0.921	0.880	0.838	0.797	0.756	0.715	0.675	0.636	0.597	0.559	0.521	0.485
55	1.356	1.321	1.286	1.250	1.213	1.176	1.139	1.101	1.063	1.025	0.987	0.948	0.910	0.872	0.833	0.796	0.758	0.721	0.684	0.647	0.611	0.576
60	1.383	1.351	1.319	1.285	1.252	1.218	1.183	1.149	1.114	1.078	1.043	1.008	0.972	0.936	0.901	0.865	0.830	0.795	0.760	0.725	0.691	0.657
65	1.408	1.378	1.348	1.317	1.286	1.254	1.222	1.190	1.158	1.125	1.092	1.059	1.026	0.993	0.960	0.927	0.894	0.861	0.828	0.795	0.762	0.730
70	1.429	1.401	1.373	1.345	1.316	1.286	1.257	1.227	1.197	1.166	1.136	1.105	1.074	1.043	1.012	0.981	0.950	0.919	0.888	0.857	0.826	0.795
75	1.448	1.423	1.396	1.369	1.342	1.315	1.287	1.260	1.231	1.203	1.174	1.146	1.117	1.088	1.058	1.029	1.000	0.971	0.941	0.912	0.883	0.854
80	1.466	1.442	1.417	1.392	1.367	1.341	1.315	1.289	1.262	1.236	1.209	1.182	1.155	1.127	1.100	1.072	1.045	1.017	0.989	0.962	0.934	0.907
85	1.482	1.459	1.436	1.412	1.388	1.364	1.340	1.315	1.290	1.265	1.240	1.214	1.189	1.163	1.137	1.111	1.085	1.059	1.033	1.006	0.980	0.954
90	1.497	1.475	1.453	1.431	1.408	1.385	1.362	1.339	1.315	1.292	1.268	1.244	1.220	1.195	1.171	1.146	1.121	1.097	1.072	1.047	1.022	0.997
95	1.510	1.490	1.469	1.448	1.426	1.405	1.383	1.361	1.338	1.316	1.293	1.271	1.248	1.225	1.201	1.178	1.155	1.131	1.108	1.084	1.060	1.037
100	1.523	1.503	1.483	1.463	1.443	1.422	1.402	1.381	1.359	1.338	1.317	1.295	1.273	1.251	1.229	1.207	1.185	1.162	1.140	1.118	1.095	1.072
150	1.611	1.598	1.585	1.571	1.558	1.544	1.530	1.516	1.502	1.488	1.474	1.460	1.445	1.431	1.416	1.402	1.387	1.372	1.357	1.342	1.327	1.312
200	1.664	1.654	1.644	1.634	1.624	1.613	1.603	1.593	1.582	1.572	1.561	1.551	1.540	1.529	1.519	1.508	1.497	1.486	1.475	1.464	1.453	1.442

Table A-6

Models with no intercept (from Farebrother): Negative serial correlation

Durbin-Watson Ninety Nine Per Cent Minimal Bound

N	K=0	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10	K=11	K=12	K=13	K=14	K=15	K=16	K=17	K=18	K=19	K=20	K=21
2	1.999	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
3	2.951	0.999	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
4	3.221	1.967	0.586	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
5	3.261	2.462	1.359	0.382	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
6	3.235	2.682	1.878	0.983	0.268	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	3.198	2.776	2.177	1.459	0.740	0.198	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	3.166	2.817	2.347	1.776	1.158	0.576	0.153	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	3.133	2.837	2.448	1.983	1.465	0.937	0.460	0.121	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	3.101	2.847	2.514	2.121	1.684	1.224	0.773	0.375	0.098	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	3.071	2.847	2.560	2.220	1.842	1.441	1.035	0.647	0.312	0.081	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
12	3.043	2.843	2.592	2.294	1.961	1.607	1.244	0.885	0.549	0.263	0.069	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
13	3.017	2.836	2.612	2.349	2.054	1.737	1.410	1.082	0.764	0.471	0.225	0.059	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
14	2.992	2.828	2.626	2.391	2.127	1.842	1.544	1.244	0.948	0.666	0.409	0.195	0.051	-----	-----	-----	-----	-----	-----	-----	-----	-----
15	2.969	2.818	2.635	2.423	2.185	1.928	1.656	1.379	1.104	0.837	0.585	0.358	0.170	0.044	-----	-----	-----	-----	-----	-----	-----	-----
16	2.948	2.808	2.640	2.447	2.231	1.997	1.749	1.494	1.237	0.985	0.743	0.517	0.316	0.150	0.039	-----	-----	-----	-----	-----	-----	-----
17	2.927	2.797	2.643	2.466	2.269	2.055	1.827	1.591	1.351	1.114	0.883	0.664	0.461	0.281	0.133	0.035	-----	-----	-----	-----	-----	-----
18	2.908	2.787	2.644	2.480	2.299	2.102	1.893	1.675	1.451	1.227	1.007	0.796	0.597	0.413	0.251	0.119	0.031	-----	-----	-----	-----	-----
19	2.890	2.776	2.643	2.492	2.324	2.142	1.948	1.746	1.538	1.327	1.118	0.915	0.721	0.539	0.372	0.226	0.107	0.028	-----	-----	-----	-----
20	2.874	2.766	2.641	2.500	2.344	2.176	1.996	1.807	1.613	1.415	1.217	1.022	0.834	0.656	0.489	0.337	0.204	0.096	0.025	-----	-----	-----
21	2.858	2.756	2.638	2.506	2.361	2.204	2.036	1.861	1.678	1.492	1.305	1.119	0.937	0.763	0.598	0.446	0.307	0.185	0.087	0.023	-----	-----
22	2.842	2.746	2.635	2.511	2.375	2.228	2.071	1.907	1.736	1.561	1.384	1.207	1.032	0.862	0.700	0.548	0.408	0.280	0.169	0.080	0.021	-----
23	2.828	2.736	2.631	2.515	2.387	2.249	2.102	1.947	1.786	1.621	1.454	1.285	1.118	0.954	0.796	0.645	0.504	0.374	0.257	0.155	0.073	0.019
24	2.814	2.727	2.627	2.517	2.396	2.267	2.128	1.983	1.831	1.675	1.516	1.356	1.196	1.038	0.884	0.736	0.596	0.465	0.345	0.237	0.143	0.067
25	2.801	2.717	2.623	2.518	2.404	2.282	2.151	2.014	1.871	1.723	1.572	1.420	1.267	1.115	0.966	0.821	0.683	0.552	0.430	0.319	0.218	0.132
26	2.789	2.709	2.618	2.519	2.411	2.295	2.171	2.042	1.906	1.766	1.623	1.478	1.331	1.186	1.042	0.901	0.765	0.635	0.512	0.399	0.295	0.202
27	2.777	2.700	2.614	2.519	2.416	2.306	2.189	2.066	1.938	1.805	1.669	1.530	1.390	1.250	1.111	0.975	0.842	0.714	0.592	0.477	0.371	0.274
28	2.766	2.692	2.609	2.519	2.421	2.316	2.205	2.088	1.966	1.839	1.710	1.577	1.444	1.309	1.176	1.043	0.914	0.788	0.667	0.553	0.445	0.346
29	2.755	2.684	2.604	2.518	2.425	2.325	2.219	2.107	1.991	1.871	1.747	1.621	1.493	1.364	1.235	1.107	0.981	0.858	0.739	0.625	0.517	0.416
30	2.745	2.676	2.600	2.517	2.428	2.332	2.231	2.125	2.014	1.899	1.781	1.660	1.537	1.414	1.290	1.166	1.044	0.924	0.807	0.695	0.587	0.485
31	2.735	2.668	2.595	2.515	2.430	2.339	2.242	2.140	2.035	1.925	1.812	1.696	1.579	1.460	1.340	1.221	1.102	0.986	0.872	0.761	0.654	0.552
32	2.725	2.661	2.590	2.514	2.432	2.344	2.252	2.155	2.053	1.948	1.840	1.729	1.616	1.502	1.387	1.272	1.157	1.043	0.932	0.823	0.718	0.617
33	2.716	2.654	2.586	2.512	2.433	2.349	2.260	2.167	2.070	1.970	1.866	1.759	1.651	1.541	1.430	1.319	1.208	1.097	0.989	0.882	0.779	0.678
34	2.707	2.647	2.581	2.510	2.434	2.353	2.268	2.179	2.086	1.989	1.889	1.787	1.683	1.577	1.470	1.363	1.255	1.148	1.042	0.938	0.836	0.737
35	2.699	2.640	2.576	2.508	2.435	2.357	2.275	2.189	2.100	2.007	1.911	1.813	1.713	1.611	1.507	1.404	1.299	1.196	1.093	0.991	0.891	0.794
36	2.690	2.634	2.572	2.506	2.435	2.360	2.281	2.199	2.113	2.023	1.931	1.837	1.740	1.642	1.542	1.442	1.341	1.240	1.140	1.041	0.943	0.847
37	2.683	2.627	2.567	2.503	2.435	2.363	2.287	2.207	2.124	2.038	1.950	1.859	1.765	1.670	1.574	1.477	1.379	1.282	1.184	1.088	0.992	0.898
38	2.675	2.621	2.563	2.501	2.435	2.365	2.292	2.215	2.135	2.052	1.967	1.879	1.789	1.697	1.604	1.510	1.416	1.321	1.226	1.132	1.039	0.947
39	2.667	2.615	2.559	2.499	2.435	2.367	2.296	2.222	2.145	2.065	1.982	1.898	1.811	1.722	1.632	1.541	1.450	1.358	1.266	1.174	1.083	0.993
40	2.660	2.609	2.555	2.496	2.434	2.369	2.300	2.229	2.154	2.077	1.997	1.915	1.831	1.746	1.659	1.570	1.482	1.392	1.303	1.213	1.124	1.036
45	2.628	2.583	2.535	2.484	2.430	2.374	2.315	2.253	2.190	2.124	2.056	1.986	1.914	1.841	1.767	1.691	1.614	1.537	1.459	1.381	1.302	1.224
50	2.600	2.559	2.516	2.471	2.424	2.374	2.323	2.269	2.214	2.157	2.098	2.037	1.975	1.911	1.847	1.781	1.714	1.646	1.578	1.509	1.439	1.370
55	2.575	2.538	2.500	2.459	2.417	2.373	2.327	2.280	2.231	2.180	2.128	2.075	2.020	1.964	1.907	1.849	1.790	1.730	1.669	1.608	1.546	1.484
60	2.553	2.519	2.484	2.448	2.409	2.370	2.329	2.286	2.242	2.197	2.151	2.103	2.054	2.004	1.954	1.902	1.849	1.796	1.742	1.687	1.631	1.576
65	2.534	2.503	2.470	2.437	2.402	2.366	2.329	2.290	2.250	2.210	2.168	2.125	2.081	2.036	1.990	1.944	1.896	1.848	1.799	1.750	1.700	1.650
70	2.516	2.487	2.458	2.427	2.395	2.361	2.327	2.292	2.256	2.219	2.181	2.142	2.102	2.061	2.020	1.977	1.934	1.891	1.846	1.802	1.756	1.710
75	2.500	2.473	2.446	2.417	2.387	2.357	2.325	2.293	2.260	2.226	2.191	2.155	2.118	2.081	2.043	2.005	1.965	1.926	1.885	1.844	1.802	1.760
80	2.486	2.461	2.436	2.408	2.380	2.352	2.323	2.293	2.262	2.231	2.198	2.165	2.132	2.098	2.063	2.027	1.991	1.954	1.917	1.879	1.841	1.803
85	2.473	2.449	2.425	2.399	2.374	2.347	2.320	2.292	2.263	2.234	2.204	2.174	2.143	2.111	2.079	2.046	2.012	1.979	1.944	1.909	1.874	1.838
90	2.460	2.438	2.415	2.391	2.367	2.342	2.317	2.291	2.264	2.237	2.209	2.181	2.152	2.122	2.092	2.061	2.030	1.999	1.967	1.935	1.902	1.869
95	2.449	2.428	2.406	2.384	2.361	2.338	2.314	2.289	2.264	2.239	2.212	2.186	2.159	2.131	2.103	2.075	2.046	2.016	1.986	1.956	1.926	1.895
100	2.438	2.418	2.398	2.377	2.355	2.333	2.310	2.287	2.264	2.240	2.215	2.190	2.165	2.139	2.113	2.086	2.059	2.031	2.003	1.975	1.946	1.917
150	2.363	2.349	2.336	2.322	2.308	2.294	2.279	2.265	2.250	2.235	2.220	2.204	2.188	2.173	2.156	2.140	2.124	2.107	2.090	2.073	2.056	2.039
200	2.317	2.307	2.296	2.286	2.276	2.265	2.255	2.244	2.233	2.222	2.211	2.200	2.189	2.177	2.166	2.154	2.142	2.131	2.119	2.106	2.094	2.082