University of Technology - Iraq

Department of Computer Sciences – Information Systems

Branch

Data Analysis Methods

Fourth Class – Second Course

Lecturer: Asst. Prof. Dr. Hiba Basim Alwan

2024 – 2025

DATA ANALYSIS METHODS

Prepared by: Asst. Prof. Dr. Hiba Basim Alwan

2024-2025
UNIVERSITY OF TECHNOLOGY
Department of Computer Sciences – Information Systems Branch

CH. 1: Introduction

1.1. Data and Information

In everyday language, terms like information and data are often used interchangeably. Researchers use these terms in specific ways emphasising how useful each can be. Data is the ***facts or recorded measures of certain phenomena (things)*** while information is the ***data formatted (structured) to support decision-making or define the relationship between two facts***. There are various types of data as listed below:

- **Qualitative or Nominal:** Described by a word or phrase (e.g. blood group, colour). Qualitative data can be analysed by considering the frequencies of different categories. Two data analysis techniques for qualitative data are content analysis ***(which measures content changes over time and across media)*** and discourse analysis ***(which explores conversations in their social context)***.
- **Quantitative:** Described by a number (e.g. time till cure, number of calls arriving at a telephone exchange in 5 seconds). Two data analysis techniques for quantitative data are regression analysis ***(which examines relationships between two variables)*** and hypothesis analysis ***(which tests whether a hypothesis is true)***. Quantitative data can be:
    - **Discrete:** the data can only take one of a finite or countable number of values (e.g. a count)
    - **Continuous:** the data is a measurement which can take any value in an interval of the real line (e.g. a weight).
- **Ordinal:** This is an "in-between" case. Observations are not numbers but they can be ordered (e.g. much improved, improved, same, worse, much worse). Ordinal data can be analysed like qualitative data but requires special techniques called ***nonparametric methods***.

1.2. Data Analysis

Data analysis is ***the application of reasoning to understand the data gathered.*** The appropriate analytical technique for data analysis will be determined by ***management's information requirements, the characteristics of the research design, and the nature of the data gathered***.

The data analysis process usually involves organizing and summarizing the data. Tables, graphs, and numerical summaries allow increased understanding and provide an effective way to present data. Methods for organizing and summarizing data make up the branch of statistics called ***descriptive statistics***.

The data analysis process can be viewed as a sequence of steps that lead from planning to data collection to informed conclusions based on the resulting data. The process can be organized into the following six steps:

1. **Data Requirement Gathering:** First of all, you have to think about why you want to do this data analysis. All you need to find out the purpose or aim of analyzing data. You have to decide which type of data analysis you want to do. In this phase, you have to decide what to analyse and how to measure it, you have to understand why you are investigating and what measures you have to use to do this analysis.
2. **Data Collection:** After requirement gathering, you will get a clear idea about what things you have to measure and what should be your findings. Now it's time to collect your data based on requirements. Once you collect your data, remember that the collected data must be processed or organized for analysis. As you collect data from various sources, you must have to keep a log with a collection date and source of the data.
3. **Data Cleaning:** Now whatever data is collected may not be useful or irrelevant to your aim of analysis, hence it should be cleaned. The data which is collected may contain duplicate records, white spaces or errors. The data should be cleaned and error-free. This step must be done before analysis because based on data cleaning, your output of analysis will be closer to your expected outcome.
4. **Data Analysis:** Once the data is collected, cleaned, and processed, it is ready for analysis. As you manipulate data, you may find you have the exact information you need, or you might need to collect more data. During this step, you can use data analysis tools and software which will help you to understand, interpret, and derive conclusions based on the requirements.
5. **Data Interpretation:** After analysing your data, it's finally time to interpret your results. You can choose the way to express or communicate your data analysis either you can use simple words or maybe a table or chart. Then use the results of your data analysis process to decide your best course of action.

6 **Data Visualization:** Data visualization is very common in your day-to-day life; it often appears in the form of charts and graphs. In other words, data is shown graphically so that it will be easier for the human brain to understand and process it. Data visualization is often used to discover unknown facts and trends. By observing relationships and comparing datasets, you can find a way to find meaningful information.

1.3. Types of Data Analysis Methods

There are seven essential types of data analysis methods. They are:

- **Data Mining:** A method of analysis that is the umbrella term for engineering metrics and insights for additional value, direction, and context. By using exploratory statistical evaluation, data mining aims to identify dependencies, relations, data patterns, and trends to generate advanced knowledge. When considering how to analyse data, adopting a data mining mindset is essential to success, as such it is an area that is worth exploring in greater detail.
- **Text Analysis:** Text analysis, also known in the industry as text mining, is the process of taking large sets of textual data and arranging them in a way that makes it easier to manage. By working through this cleansing process in stringent detail, you will be able to extract the data that is truly relevant to your business and use it to develop actionable insights that will propel you forward.
- **Cluster Analysis:** The action of grouping a set of data elements in a way that said elements are more similar (in a particular sense) to each other than to those in other groups, hence the term “cluster.” Since there is no target variable when clustering, the method is often used to find hidden patterns in the data. The approach is also used to provide additional context to a trend or dataset.
- **Cohort Analysis:** This type of data analysis method uses historical data to examine and compare a determined segment of users' behaviour, which can then be grouped with others with similar characteristics. By using this data analysis methodology, it is possible to gain a wealth of insight into consumer needs or a firm understanding of a broader target group. A useful tool to start performing cohort analysis method is Google Analytics.

- **Regression Analysis:** The regression analysis uses historical data to understand how a dependent variable's value is affected when one (linear regression) or more independent variables (multiple regression) change or stay the same. By understanding each variable's relationship and how they developed in the past, you can anticipate possible outcomes and make better business decisions for example in the future.
- **Neural Network:** The neural network forms the basis for the intelligent algorithms of machine learning. It is a form of data-driven analytics that attempts, with minimal intervention, to understand how the human brain processes insights and predicts values. Neural networks learn from every data transaction, meaning that they evolve and advance over time. A typical area of application for neural networks is predictive data analysis.
- **Factor Analysis:** Factor analysis, also called "dimension reduction," is a type of data analysis used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. The aim here is to uncover independent latent variables, an ideal analysis method for streamlining specific data segments.

1.4. Main Analysis Categories

The main analysis categories are:

- **Text Analysis:** Also called "***Data Mining***" text analysis uses databases and data mining tools to discover patterns residing in large datasets. It transforms raw data into useful business information. Text analysis is arguably the most straightforward method of data analysis.
- **Exploratory Analysis:** Exploratory analysis answers the question, "How to explore data relationships." As its name suggests, the main aim of the exploratory analysis is to explore. Before it, there is still no notion of the relationship between the data and the variables. Once the data is investigated, the exploratory analysis enables you to find connections and generate hypotheses and solutions for specific problems. A typical area of application for exploratory analysis is data mining.

- **Diagnostic Analysis:** Diagnostic analysis answers the question, "Why did it happen?" Using insights gained from statistical analysis (more on that later!), analysts use diagnostic analysis to identify patterns in data. Ideally, the analysts find similar patterns that existed in the past, and consequently, use those solutions to resolve the present challenges hopefully.
- **Predictive Analysis:** Predictive analysis answers the question, "What will happen?" By using patterns found in older data as well as current events, analysts predict future events. While there's no such thing as 100 per cent accurate forecasting, the odds improve if the analysts have plenty of detailed information and the discipline to research it thoroughly.
- **Prescriptive Analysis:** Prescriptive analysis answers the question, "How will it happen?" Mix all the insights gained from the other data analysis types, and you have prescriptive analysis. Sometimes, an issue cannot be solved solely with one analysis type and instead requires multiple insights.
- **Statistical Analysis:** Statistical analysis answers the question, "What happened?" This analysis covers data collection, analysis, modelling, interpretation, and presentation using dashboards. The statistical analysis breaks down into two sub-categories:
  - **Descriptive Analysis:** it works with either complete or selections of summarized numerical data. It illustrates means and deviations in continuous data and percentages and frequencies in categorical data.
  - **Inferential Analysis:** it works with samples derived from complete data. An analyst can arrive at different conclusions from the same comprehensive data set just by choosing different samplings.

CH. 2: Descriptive and Inferential Statistics

2.1. Descriptive Statistics

Descriptive statistics: Describe basic characteristics and summarize the data straightforwardly and understandably, i.e., it describes characteristics of the population or sample. In other words, descriptive statistics describe the basic information or characteristics of a data set under study. Descriptive statistics just describes and summarizes data but does not allow us to conclude the whole population from which we took the sample. You are simply summarizing the data with charts, tables, and graphs. Descriptive statistics has two main types:

- Measures of central tendency (mean, median, and mode).
- Measures of dispersion or variation (variance, standard deviation, range).

2.2. Types of Distribution

There are many types of distributions, among them the following distributions are the most commonly used.

- **Frequency Distributions:** One of the most common ways to summarize a set of data is to construct a frequency table or frequency distribution. The process begins with recording the number of times a particular value of a variable occurs. This is the frequency of that value.
- **Percentage Distribution:** A frequency distribution organized into a table (or graph) that summarizes percentage values associated with particular values of a variable.
- **Probability Distribution:** It is the long-run relative frequency with which an event will occur. Inferential statistics uses the concept of a probability distribution, which is conceptually the same as a percentage distribution except that the data are converted into probabilities.

Example:

The number of accidents experienced by 80 machinists in a certain industry over one year was found to be as shown below. Construct a frequency table, percentage table, and probability table then draw a bar chart for these tables.

2	0	0	1	0	2	0	6	0	0	8	0	2	0	1
5	1	0	1	1	2	1	0	0	0	2	0	0	0	0

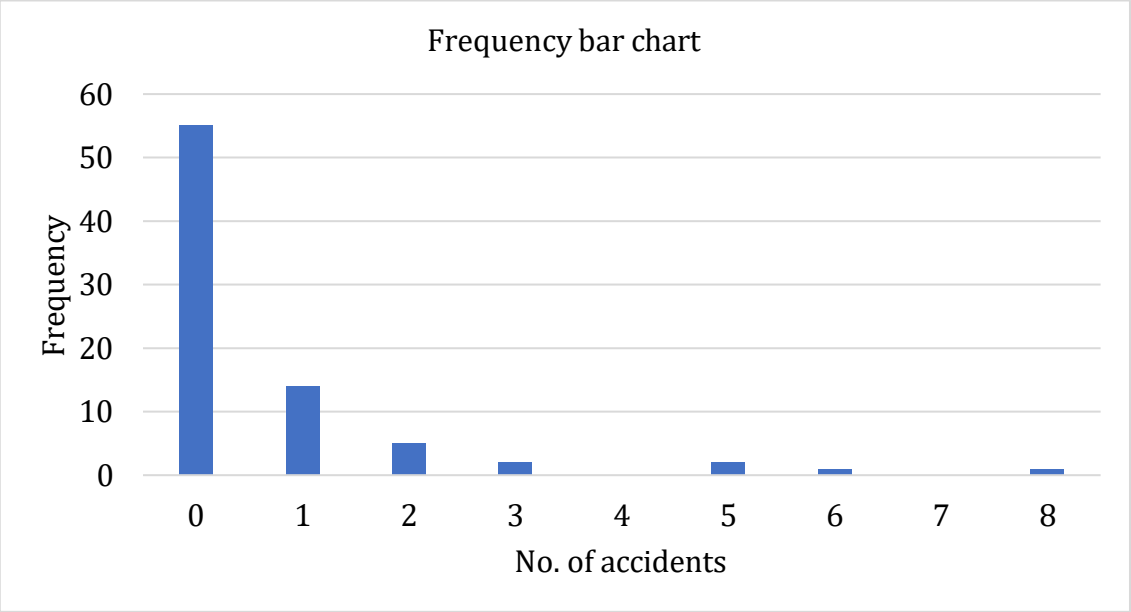
0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
0	0	0	5	1	0	0	0	0	0	0	0	0	1	1
0	3	0	0	1	1	0	0	0	2	0	1	0	0	0
0	0	0	0	0										

Solution:

Frequency table:

No. of accidents	Frequency
0	55
1	14
2	5
3	2
4	0
5	2
6	1
7	0
8	1
	80

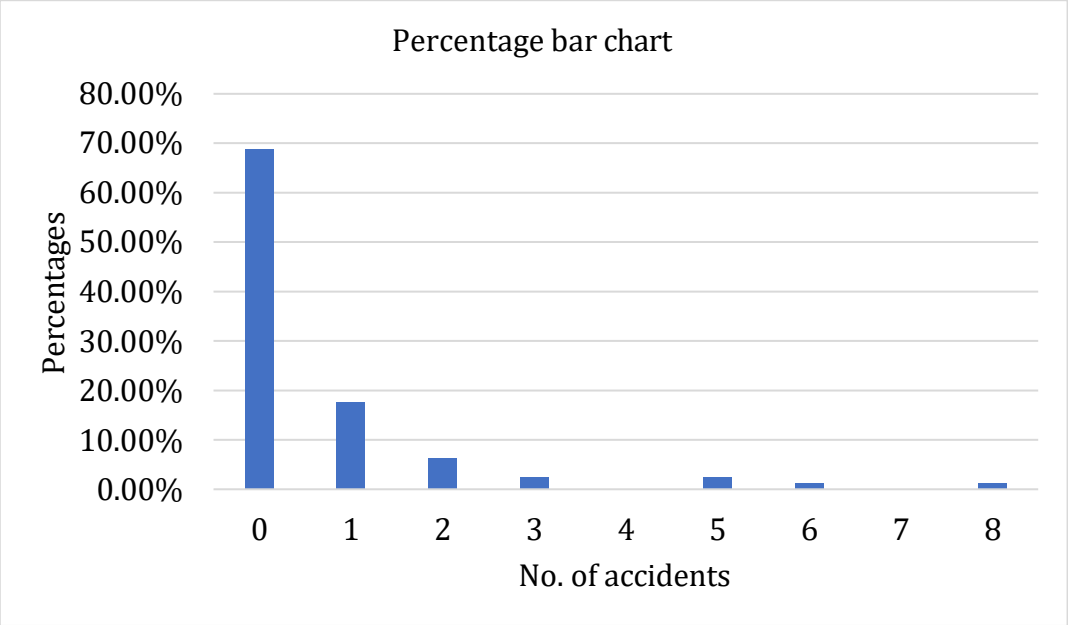
Frequency bar chart:



Percentage table:

No. of accidents	Percentage%	
0	55 / 80 × 100	68.75%
1	14 / 80 × 100	17.5%
2	5 / 80 × 100	6.25%
3	2 / 80 × 100	2.5%
4	0 /80 × 100	0%
5	2 / 80 × 100	2.5%
6	1 /80 × 100	1.25%
7	0 /80 × 100	0%
8	1 /80 × 100	1.25%
	80 / 80 × 100	100%

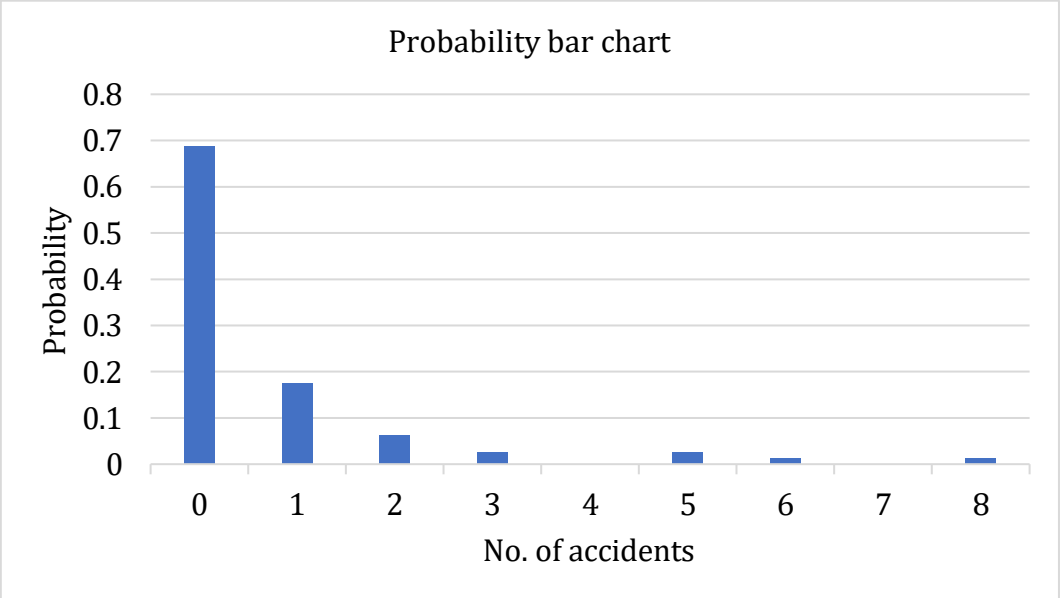
Percentage bar chart:



Probability table:

No. of accidents	Probability	
0	55 / 80	0.6875
1	14 / 80	0.175
2	5 / 80	0.0625
3	2 / 80	0.025
4	0 /80	0
5	2 / 80	0.025
6	1 /80	0.0125
7	0 /80	0
8	1 /80	0.0125
	80 / 80	1

Probability bar chart:



2.3. Inferential Statistics

Inferential statistics: it is used to make inferences or to project from a sample to an entire population.

Let's say you need to know the average weight of all the women in a city with a population of a million people. It is not easy to get the weight of each woman. This is where inferential statistics start playing. Inferential statistics can make conclusions about the whole population of women using data drawn from a sample or samples of it.

Inferential statistics studies the relationships between variables within a sample. Then make generalizations and even predictions about the relationship between those variables within the whole population. To do that, inferential statistics need some techniques, methods, and types of calculations. Some of the most important of them are:

- Linear regression analysis
- Logistic regression analysis
- Analysis of Variance (ANOVA)
- Analysis of Covariance (ANCOVA)
- Statistical significance (T-test)
- Correlation analysis
- Structural equation modelling
- Survival analysis
- Factor analysis
- Multidimensional scaling
- Cluster analysis
- Discriminant function analysis

2.3.1. Inferential Statistics through Hypothesis Test

Hypothesis testing is a ***form of inferential statistics that allows us to conclude an entire population based on a representative sample***. In this method, we test some hypotheses by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true. The goal of

hypothesis testing is to determine the likelihood that a population parameter, such as the mean, is likely to be true.

A hypothesis is a ***formal statement explaining some outcome***. In its simplest form, a hypothesis is a ***guess***. A sales manager may hypothesize that the salespeople who are highest in product knowledge will be the most productive. An advertising manager may hypothesize that if consumers' attitudes toward a product change in a positive direction, there will be an increase in consumption of the product. A human resource manager may hypothesize that job candidates with certain majors will be more successful employees.

A hypothesis is an empirically testable proposition. In other words, when one states a hypothesis, it should be written in a manner that can be supported or shown to be wrong through an empirical test.

2.3.2. Two Hypothesises

Hypothesis testing is a ***statistical analysis that uses sample data to assess two mutually exclusive theories about the properties of a population***. Statisticians call these theories the null hypothesis (***The null hypothesis, denoted by H_0, is a claim about a population characteristic that is initially assumed to be true***) and the alternative hypothesis (***The alternative hypothesis, denoted by H_a, is the competing claim***). In carrying out a test of H_0 versus H_a, the hypothesis H_0 will be rejected in favour of H_a only if sample evidence strongly suggests that H_0 is false. If the sample does not provide such evidence, H_0 will not be rejected. The two possible conclusions are then ***rejected H_0*** or ***fail to reject H_0***. The form of a null hypothesis is:

H_0: population characteristic = hypothesized value
where the hypothesized value is a specific number determined by the problem context.
The alternative hypothesis will have one of the following three forms:
H_a: population characteristic > hypothesized value **(one-tailed test, right-tailed test)**

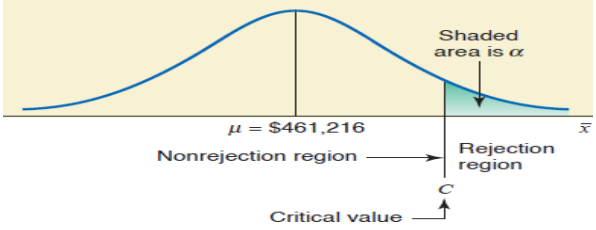


Fig. 1: A right-tailed test

H_a: population characteristic < hypothesized value **(one-tailed test, left-tailed test)**

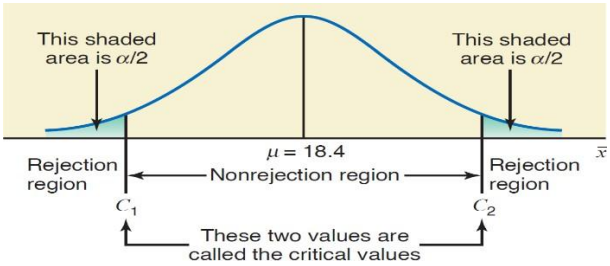


Fig. 2: A two-tailed test

H_a: population characteristic ≠ hypothesized value **(two-tailed test)**

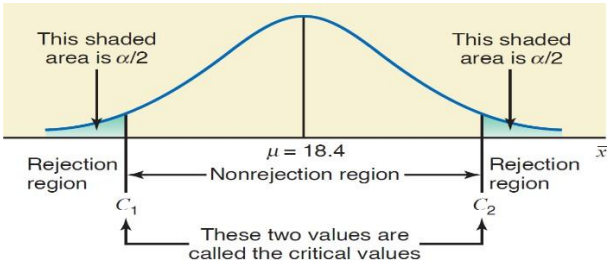


Fig. 3: A two-tailed test

2.3.3. Hypothesis Tests About μ: σ Known

This section explains how to perform a test of hypothesis for the population mean μ when the population standard deviation σ is known. Here there are three possible cases as follows:

Case I. If the following three conditions are fulfilled:

- The population standard deviation σ is known.
- The sample size is small (i.e., n < 30).
- The population from which the sample is selected is normally distributed.

then we use the normal distribution to perform a test of hypothesis about μ because the sampling distribution of $\bar{x}$ is normal with its mean equal to μ and the standard deviation equal to $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, assuming that $n/N \leq 0.5$.

Case II. If the following two conditions are fulfilled:

- The population standard deviation σ is known.
- The sample size is large (i.e., n ≥ 30).

then, again, we use the normal distribution to perform a test of the hypothesis about μ due to the central limit theorem, the sampling distribution of $\bar{x}$ is (approximately) normal with its mean equal to μ and the standard deviation equal to $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ assuming that $n/N \leq 0.5$.

Case III. If the following three conditions are fulfilled:

- The population standard deviation σ is known.
- The sample size is small (i.e., $n < 30$).
- The population from which the sample is selected is not normally distributed (or the shape of its distribution is unknown).

then we use a nonparametric method to perform a test of the hypothesis about μ.

The following chart summarizes the above three cases.

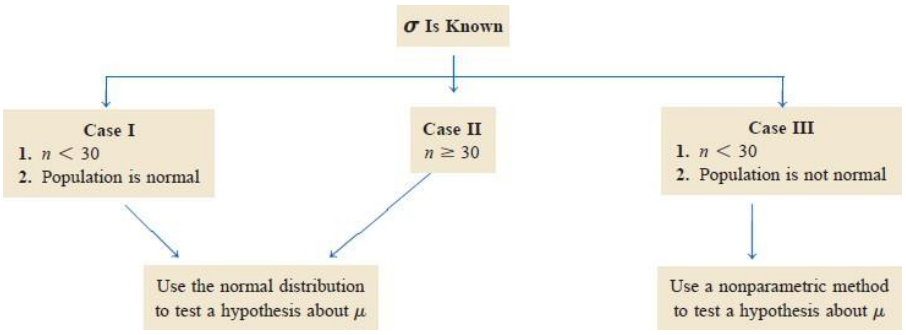


Fig. 4: Hypothesis Tests About μ: σ Known

Example: At Canon Food Corporation, it used to take an average of 90 minutes for new workers to learn a food processing job. Recently the company installed a new food processing machine. The supervisor at the company wants to find out if the meantime taken by new workers to learn the food processing procedure on this new machine is different from 90 minutes. A sample of 20 workers showed that it took, on average, 85 minutes for them to learn the food processing procedure on the new machine. It is known that the learning times for all new workers are normally distributed with a population standard deviation of 7 minutes. Find the *p*-value for the test that the mean learning time for the food processing procedure on the new machine is different from 90 minutes. What will your conclusion be if α = 0.01?

Solution: Let μ be the meantime (in minutes) taken to learn the food processing procedure on the new machine by all workers, and let $\bar{x}$ be the corresponding sample mean. From the given information,

n = 20, $\bar{x}$ = 85 minutes, σ = 7 minutes, and α = 0.01

To calculate the *p*-value and perform the test, we apply the following four steps:

Step 1. State the null and alternative hypotheses.

H_0: μ = 90 minutes

H_1: μ ≠ 90 minutes

Note that the null hypothesis states that the mean time for learning the food processing procedure on the new machine is 90 minutes, and the alternative hypothesis states that this time is different from 90 minutes.

Step 2. Select the distribution to use.

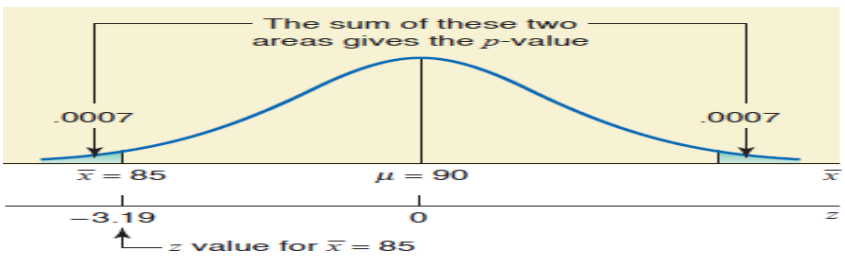
Here, the population standard deviation σ is known, the sample size is small (n < 30), but the population distribution is normal. Hence, the sampling distribution of $\bar{x}$ is normal with its mean equal to μ and the standard deviation equal to $\sigma_x = \frac{\sigma}{\sqrt{n}}$. Consequently, we will use the normal distribution to find the p-value and make the test.

Step 3. Calculate the p-value.

The ≠ sign in the alternative hypothesis indicates that the test is two-tailed. The p-value is equal to twice the area in the tail of the sampling distribution curve of $\bar{x}$ to the left of $\bar{x}$ = 85 as shown in Figure 7. To find this area, we first find the z value for $\bar{x}$ = 85 as follows:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{20}} = 1.56524758 \text{ mintues}$$

$$z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{85-90}{1.56524758} = -3.19$$



The area to the left of $\bar{x} = 85$ is equal to the area under the standard normal curve to the left of z = - 3.19. From the normal distribution table, the area to the left of z = - 3.19 is .0007. Consequently, the *p*-value is:

p-value = 2(0,0007) = 0,0014

Step 4. Make a decision.

Thus, based on the *p*-value of .0014, we can state that for any α (significance level) greater than 0.0014, we will reject the null hypothesis stated in Step 1, and for any α less than or equal to 0.0014, we will not reject the null hypothesis.

Because α = 0.01 is greater than the *p*-value of 0.0014, we reject the null hypothesis at this significance level. Therefore, we conclude that the mean time for learning the food processing procedure on the new machine is different from 90 minutes.

2.3.4. Hypothesis Tests About μ: σ Not Known

This section explains how to perform a test of hypothesis for the population mean μ when the population standard deviation σ is not known. Here there are three possible cases as follows:

Case I. If the following three conditions are fulfilled:

- The population standard deviation σ is not known.
- The sample size is small (i.e., n < 30).
- The population from which the sample is selected is normally distributed.

then we use the t distribution to perform a test of the hypothesis about μ.

Case II. If the following two conditions are fulfilled:

- The population standard deviation σ is not known.
- The sample size is large (i.e., n ≥ 30).

then, again, we use the t distribution to perform a test of the hypothesis about μ.

Case III. If the following three conditions are fulfilled:

- The population standard deviation σ is not known.
- The sample size is small (i.e., n < 30).

– The population from which the sample is selected is not normally distributed (or the shape of its distribution is unknown).

then we use a nonparametric method to perform a test of the hypothesis about μ.

The following chart summarizes the above three cases.

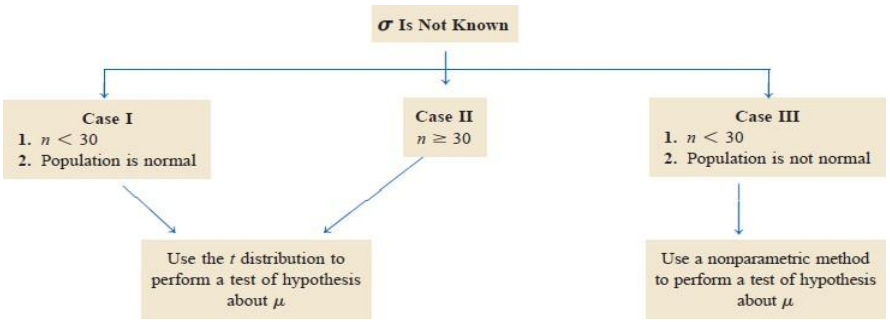


Fig. 5: Hypothesis Tests About μ: σ Not Known

Example: A psychologist claims that the mean age at which children start walking is 12.5 months. Carol wanted to check if this claim was true. She took a random sample of 18 children and found that the mean age at which these children started walking was 12.9 months with a standard deviation of .80 months. It is known that the ages at which all children start walking are approximately normally distributed. Find the *p*-value for the test that the mean age at which all children start walking is different from 12.5 months. What will your conclusion be if the significance level is 1%?

Solution: Let μ be the mean age at which all children start walking, and let $\bar{x}$ be the corresponding mean for the sample. From the given information:

n = 18 $\bar{x}$ = 12.9 months s = 0.8 month

The psychologist claims that the mean age at which children start walking is 12.5 months. To calculate the *p*-value and to make the decision, we apply the following four steps:

Step 1. State the null and alternative hypotheses.

We are to test if the mean age at which all children start walking is different from 12.5 months. Hence, the null and alternative hypotheses are:

H_0: μ = 12.5 (The mean walking age is 12.5 months.)

H_1: μ ≠ 12.5 (The mean walking age is different from 12.5 months.)

Step 2. Select the distribution to use.

In this example, we do not know the population standard deviation σ, the sample size is small (n < 30), and the population is normally distributed. Hence, it is the Case I mentioned at the beginning of this section. Consequently, we will use the t distribution to find the p-value for this test.

Step 3. Calculate the p-value.

The ≠ sign in the alternative hypothesis indicates that the test is two-tailed. To find the p-value, first, we find the degrees of freedom and the t-value for $\bar{x}$ = 12.9 months. Then, the p-value is equal to twice the area in the tail of the t distribution curve to the right of this t value for $\bar{x}$ = 12.9 months. This p-value is shown in Figure 10. We find this p-value as follows:

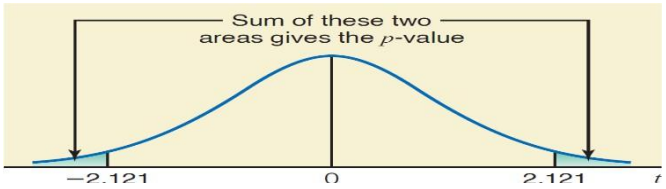
$$s_x = \frac{s}{\sqrt{n}} = \frac{0.8}{\sqrt{18}} = 0.18856181$$

$$t = \frac{\bar{x}-\mu}{s_{\bar{x}}} = \frac{12.9-12.5}{0.18856181} = 2.121 \; and \; df = n - 1 = 18 - 1 = 17$$

Now we can find the range for the p-value. To do so, we go to the attached table (the t distribution table) and find the row of df = 17. In this row, we find the two values of t that cover t = 2.121. From the attached table, for df = 17, these two values of t are 2.110 and 2.567. The test statistic t = 2.121 falls between these two values. Now look in the top row of this table to find the areas in the tail of the t distribution curve that correspond to 2.110 and 2.567. These two areas are 0.025 and 0.01, respectively. In other words, the area in the upper tail of the t distribution curve for df = 17 and t = 2.110 is 0.025, and the area in the upper tail of the t distribution curve for df = 17 and t = 2.567 is 0.01. Because it is a two-tailed test, the p-value for t = 2.121 is between 2(0.025) and 2(0.01) = 0.02, which can be written as:

0.02 < p-value < 0.05

Note that by using an attached table, we cannot find the exact p-value but only a range for it. If we have access to technology, we can find the exact p-value by using technology. If we use technology for this example, we will obtain a p-value of 0.049.



Step 4. Make a decision.

Thus, we can state that for any α greater than .05 (the upper limit of the p-value range), we will reject the null hypothesis. For any less than .02 (the lower limit of the p-value range), we will not reject the null hypothesis. However, if α is between 0.02 and 0.05, we cannot make a decision. Note that if we use technology, then the p-value we will obtain for this example is .049, and we can make a decision for any value of α. For our example, α = 0.01, which is less than the lower limit of the p-value range of 0.02. As a result, we fail to reject H_0 and conclude that the mean age at which all children start walking is not different from 12.5 months. As a result, we can state that the difference between the hypothesized population mean and the sample mean is so small that it may have occurred because of sampling error.

2.4. Permutation and Randomization Test

2.4.1. Permutation

The concept of permutations is very similar to that of combinations but with one major difference here the order of selection is important. Suppose there are three marbles in a jar red, green, and purple—and we select two marbles from these three. When the order of selection is not important, there are three ways (combinations) to do so. Those three ways are RG, RP, and GP, where R represents that a red marble is selected, G means a green marble is selected, and P indicates a purple marble is selected. In these three combinations, the order of selection is not important, and, thus, RG and GR represent the same selection. However, if the order of selection is important, then RG and GR are not the same selections, but they are two different selections. Similarly, RP and PR are two different selections, and GP and PG are two different selections. Thus, if the order in which the marbles are selected is important, then there are six selections—RG, GR, RP, PR, GP, and PG. These are called six ***permutations*** or ***arrangements***.

Permutations Notation Permutations give the total selections of x elements from n (different) elements in such a way that the order of selections is important. The notation

used to denote the permutations is $_nP_x$ which is read as "the number of permutations of selecting *x* elements from *n* elements." Permutations are also called **arrangements**.

The following formula is used to find the number of permutations or arrangements of selecting *x* items out of *n* items. Note that here, the *n* items should all be different.

$$_nP_x = \frac{n!}{(n-x)!}$$

Example: A club has 20 members. They are to select three office holders' president, secretary, and treasurer for next year. They always select these office holders by drawing 3 names randomly from the names of all members. The first person selected becomes the president, the second is the secretary, and the third one takes over as treasurer. Thus, the order in which 3 names are selected from the 20 names is important. Find the total arrangements of 3 names from these 20.

Solution: For this example,

n = total members of the club = 20

x = number of names to be selected = 3

Since the order of selections is important, we find the number of permutations or arrangements using the following formula:

$$_nP_x = \frac{n!}{(n-x)!} = \frac{20!}{(20-3)!} = \frac{20!}{17!} = 6840$$

Thus, there are 6840 permutations or arrangements for selecting 3 names out of 20.

2.4.2. Randomization Test

A modern randomization test uses software to shuffle data before computing values (for example, mean and median differences) and to compare the results after shuffling to the original data. Repeat the process thousands of times to generate a proportion similar to the p-value you get in a t-test. But randomization tests predate computers.

Randomization test ***is the random assignment of subject and treatments to groups; it is one device for equally distributing the effects of extraneous variables to all conditions.*** To conduct a randomization test, we follow the following steps:

1 Compute two means. Compute the mean of the two samples.

2 **Find the mean difference.** Compute the difference between means.

3 **Combine.** Combine both samples into one group of data.

4 **Shuffle.** Shuffle the order of the combined group.

5 **Select new samples.** Randomly sample the same number of values for two new samples as you did in the original data.

6 **Compute two new means.** Compute the means of the two new samples.

7 **Find the new mean difference.** Find the new difference between means.

8 **Compare mean differences.** If the absolute value of the difference after shuffling is more than or equal to the original mean difference, record 1 for this iteration; if not, give it 0.

9 **Iterate.** Repeat 1,000+ times (we usually use 10,000 iterations).

10 **Compute p.** The percentage of 1s is the p-value for the test. If the observed mean difference is unlikely to have happened by chance, this percentage will be small.

CH. 3: Regression and Analysis of Variance

3.1. Regression

3.1.1. Regression Analysis

We can divide analysis techniques into dependence techniques and interdependence techniques. A dependence technique makes a distinction between dependent and independent variables. An interdependence technique does not make this distinction and simply is concerned with how variables relate to one another.

Regression analysis is considered a dependence technique. Regression analysis is a technique for measuring the linear association between a dependent and an independent variable. Thus, with simple regression, a dependent (or criterion) variable, *Y*, is linked to an independent (or predictor) variable, *X.* Regression analysis attempts to predict the values of a dependent variable from specific values of the independent variable.

3.1.2. The Regression Equation

Linear regression is the most widely used statistical technique; it is a way to model a relationship between two sets of variables. The result is a linear regression equation that can be used to make predictions about data.

Simple linear regression is a measure of linear association that investigates straight-line relationships between a dependent variable and an independent variable. Simple linear regression investigates a *straight-line relationship* of the type:

$$Y = \alpha + \beta X$$

where *Y* is a dependent variable and *X* is an independent variable. Alpha (α) and beta (β) are two parameters that must be estimated so that the equation best represents a given set of data. These two parameters determine the height of the regression line and the angle of the line relative to the horizontal. When these parameters change, the line changes. Regression techniques have the job of estimating values for these parameters that make the line *fit* the observations the best.

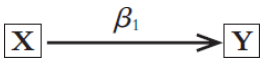
The result is simply a linear equation, or the equation for a line, just as in basic algebra! α represents the *Y-intercept* (where the line crosses the *Y*-axis) and β is the slope coefficient. The slope is the change in *Y* associated with a change of one unit in *X.* Slope

may also be thought of as rise over run: that is, how much *Y* rises (or falls if negative) for every one unit change in the *X*-axis.

3.1.3. Parameter Estimate Choices

The estimates for α and β are the key to regression analysis. In most research, the estimate of β is most important. The explanatory power of regression rests with β because this is where the direction and strength of the relationship between the independent and dependent variables are explained.

A *Y*-intercept term is sometimes referred to as a constant because α represents a fixed point. An estimated slope coefficient is sometimes referred to as a regression weight, regression coefficient, parameter estimate, or sometimes even as a *path* estimate. The term *path estimate* is a descriptive term adapted because of the way hypothesized causal relationships are often represented in diagrams:



For all practical purposes, these terms are used interchangeably.

Parameter estimates can be presented in either raw or standardized form. One potential problem with raw parameter estimates is the fact that they reflect the measurement scale range. So, if a simple regression involved distance measured with miles, very small parameter estimates may indicate a strong relationship. In contrast, if the very same distance is measured with centimetres, a very large parameter estimate would be needed to indicate a strong relationship.

Example: Find a linear regression equation for the following data:

Age X	Glucose level Y
43	99
21	65
25	79
42	75
57	87
59	81

Step 1:

Age x	Glucose level y	xy	x^2
43	99	4257	1849
21	65	1365	441
25	79	1975	625

42	75	3150	1764
57	87	4959	3249
59	81	4779	3481
247	486	20485	11409

Step 2:

$$Y = \alpha + \beta X$$

$$\alpha = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} = \frac{(485 \times 11409) - (247 \times 20485)}{(6 \times 11409) - (247)^2} = 65.1416$$

$$\beta = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{(6 \times 20485) - (247 \times 486)}{(6 \times 11409) - (247)^2} = 0.385225$$

$$Y = \alpha + \beta X$$

$$Y = 65.1416 + 0.385225\ X$$

3.2. Analysis of Variance

3.2.1. Analysis Of Variance (ANOVA)

So far, we have discussed tests for differences between the two groups. However, what happens when we have more than two groups? For example, what if we want to test and see if employee turnover differs across our five production plants? When the means of more than two groups or populations are to be compared, one-way ANOVA is the appropriate statistical tool. ANOVA involving only one grouping variable is often referred to as *one-way* ANOVA because only one independent variable is involved. Another way to define ANOVA is as the appropriate statistical technique to examine the effect of a less-than-interval independent variable on an at least interval-dependent variable. An independent samples *t*-test can be thought of as a special case of ANOVA in which the independent variable has only two levels. When more levels exist, the *t*-test alone cannot handle the problem. The statistical null hypothesis for ANOVA is stated as follows:

$$H_0: \mu_1 = \mu_2 = \mu_3 \ldots \mu_k$$

The symbol k is the number of groups or categories for an independent variable. In other words, all group means are equal. The substantive hypothesis tested in ANOVA is:

At least one group's mean is not equal to another group's mean.

As the term *analysis of variance* suggests, the problem requires comparing variances to make inferences about the means. The conditions for doing the ANOVA test are:

- The data are randomly sampled.
- The variances of each sample are assumed equal.
- The residuals are normally distributed.

3.2.2. Terminology

- **ANOVA.** It is a procedure used to test the null hypothesis that the means of three or more populations are all equal.
- **Grand Mean.** The mean of a variable overall observations.
- **Between Group Variance.** The sum of differences between the group mean and the grand mean is summed over all groups for a given set of observations.
- **Within Group Variance.** The sum of the differences between observed values and the group means for a given set of observations; also known as total error variance.
- **F-test.** A procedure used to determine whether there is more variability in the scores of one sample than in the scores of another sample.

Example: You have the data for 15 students divided into three groups and tested in the statistical course. Find whether the average scores differ between these groups or not if you know that the $\alpha = 0.05$.

G_1	G_2	G_3
75	80	70
77	82	72
79	84	74
81	86	76
83	88	78

Step 1: State the hypothesis

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

Step 2: Calculate Means

G_1	G_2	G_3
75	80	70
77	82	72
79	84	74
81	86	76
83	88	78

395/5 =79	420/5 = 84	370/5 = 74

Step 3: Calculate SS_W

	X_t	X_A	$(X_t - X_A)^2$
G_1	75	79	16
	77	79	4
	79	79	0
	81	79	4
	83	79	16
G2	80	84	16
	82	84	4
	84	84	0
	86	84	4
	88	84	16
G3	70	74	16
	72	74	4
	74	74	0
	76	74	4
	78	74	16
			$SS_W = 120$

Step 4: Calculate SS_B

	X_A	X_G	$(X_A - X_G)^2$
G_1	79	79	0
	79	79	0
	79	79	0
	79	79	0
	79	79	0
G_2	84	79	25
	84	79	25
	84	79	25
	84	79	25
	84	79	25
G_3	74	79	25
	74	79	25
	74	79	25
	74	79	25
	74	79	25
			$SS_B = 250$

Step 5: Make an ANOVA table

Source	SS	df	MS	F
SS_B	250	k – 1 = 2	$M_{SSB} = SS_B/df = 125$	$M_{SSB}/M_{SSW} = 12.5$
SS_W	120	N – k = 12	$M_{SSW} = SS_W/df = 10$	
SS_{tot}	370	N – 1 = 14		

Step 6: Make a decision

Calculated F value = 12.5

Tabulated F value = 3.89

If Calculated F value > Tabulated F value → reject H_0

12.5 > 3.89

Reject H_0 and accept H_1

Step 7: Post hoc tests

We will use the Tukey test because the sample sizes are equal.

Tukey = $q_\alpha \sqrt{\frac{M_{SS_W}}{N_A}} = 3.77\sqrt{\frac{10}{5}} = 5.33$

$Mean_{G1} – Mean_{G2} = 79 – 84 = – 5$

$Mean_{G1} – Mean_{G3} = 79 – 74 = 5$

$Mean_{G2} – Mean_{G3} = 84 – 74 = 10$

10 > 5.33

CH. 4: Supervised Learning

4.1. Introduction

Supervised Learning is a machine learning paradigm for acquiring the input-output relationship information of a system based on a given set of paired input-output training samples. Occasionally, it is also referred to as Learning with a Teacher. The goal of supervised learning is to build an artificial system that can learn the mapping between the input and the output and can predict the output of the system given new inputs.

The foremost advantage of supervised learning is that all classes or analogue outputs manipulated by the algorithm of this paradigm are meaningful to humans. And it can be easily used for discriminative pattern classification, and for data regression. But it also has several disadvantages. The first one is caused by the difficulty of collecting supervision or labels. When there is a huge volume of input data, it is prohibitively expensive, if not impossible, to label all of them. For example, it is not a trivial task to label a huge set of images for image classification. Second, as not everything in the real world has a distinctive label, there are uncertainties and ambiguities in the supervision or labels. For example, the margin for separating the two concepts of “hot” and “cold” is not distinct; and it is difficult to name an object that is a cross between a loveseat and a bed. These difficulties may limit the applications of the supervised learning paradigm in some scenarios. To overcome these limitations in practice, other learning paradigms, such as unsupervised learning, semi-supervised learning, reinforcement learning, active learning, or some mixed learning approaches can be considered.

Supervised learning enables a machine to learn human behaviour or object behaviour in certain tasks. The learned knowledge can then be used by the machine to perform similar actions on these tasks. Since the computing machinery may perform some input-output mappings much faster and more persistent than humans, machines equipped with a well-supervised learner can perform certain tasks much faster and more accurately than humans. On the other hand, because of the limitations in hardware, software, and algorithm designs, existing supervised learning algorithms still cannot match human learning ability on many complicated tasks. Supervised learning has been successfully used in areas such as information retrieval, data mining, computer vision, speech recognition, spam detection, bioinformatics, cheminformatics, and market analysis.

Classification is a supervised learning approach which is a significant field of research involving labelling an object to one of a group of classes, related to features of that object and it is considered one of the basic difficulties in numerous decision-making processes. Many decision-making processes are examples of classification difficulty or can be simply transformed into classification difficulty, for example, prognosis processes, diagnosis processes, and pattern recognition.

4.2. Supervised Learning Approaches

There are many approaches to supervised learning. Some of these approaches are explained below.

4.2.1. Linear Discriminant Analysis

Dimensionality reduction techniques are important in many applications related to machine learning, data mining, bioinformatics, biometrics and information retrieval. The main goal of the dimensionality reduction techniques is to reduce the dimensions by removing the redundant and dependent features by transforming the features from a higher dimensional space that may lead to a curse of dimensionality problem, to a space with lower dimensions. There are two major approaches of the dimensionality reduction techniques, namely, unsupervised and supervised approaches.

Linear Discriminant Analysis (LDA) is a very common technique for dimensionality reduction problems as a pre-processing step for machine learning and pattern classification applications.

The goal of the LDA technique is to project the original data matrix onto a lower dimensional space. To achieve this goal, four steps needed to be performed:

Step 1: Compute the Mean for all the given data.

Step 2: Compute the statistics for the given data.

- Mean vector (M_i).
- Covariance matrix (C_i).

Step 3: Compute within-class scatter matrix C.

Step 4: Generate discriminant functions (F_i).

Example: Compute the discrimination function for the following data then train it for X = [110 62 37.1]:

X_1	X_2	X_3	Y
138	76	38.2	1
140	75	39.2	1
145	88	38.6	1
136	89	39.8	1
129	95	37.9	1
133	82	38.3	1
138	73	38.4	1
110	62	37.1	0
115	63	37.2	0
98	62	36.8	0
120	65	37.0	0
118	68	37.1	0
102	58	37.3	0
106	65	36.9	0
111	57	37.0	0

Solution:

Step 1: Compute Mean for all the given data.

X_1	X_2	X_3	Y
138	76	38.2	1
140	75	39.2	1
145	88	38.6	1
136	89	39.8	1
129	95	37.9	1
133	82	38.3	1
138	73	38.4	1
110	62	37.1	0
115	63	37.2	0
98	62	36.8	0
120	65	37.0	0
118	68	37.1	0
102	58	37.3	0
106	65	36.9	0
111	57	37.0	0
122.6	**71.87**	**37.79**	

- **Step 2:** Compute the statistics for the given data (Mean vector (M_i) and Covariance matrix (C_i)).
  - Mean vector (M_1).

X_1	X_2	X_3	Y
138	76	38.2	1
140	75	39.2	1
145	88	38.6	1

136	89	39.8	1
129	95	37.9	1
133	82	38.3	1
138	73	38.4	1
137	**82.57**	**38.63**	

- Covariance matrix (C_1).

X_1	X_2	X_3	Y
138 – 122.6 = 15.4	76 – 71.87 = 4.13	38.2 – 37.79 = 0.413	1
140 – 122.6 = 17.4	75 – 71.87 = 3.13	39.2 – 37.79 = 1.413	1
145 – 122.6 = 22.4	88 – 71.87 = 16.13	38.6 – 37.79 = 0.813	1
136 – 122.6 = 13.4	89 – 71.87 = 17.13	39.8 – 37.79 = 2.013	1
129 – 122.6 = 6.40	95 – 71.87 = 23.13	37.9 – 37.79 = 0.113	1
133 – 122.6 = 10.4	82 – 71.87 = 10.13	38.3 – 37.79 = 0.513	1
138 – 122.6 = 15.4	73 – 71.87 = 1.13	38.4 – 37.79 = 0.613	1

$$C_1 = \frac{1}{7} \times \begin{bmatrix} 15.4 & 4.13 & 0.413 \\ 17.4 & 3.13 & 1.413 \\ 22.4 & 16.13 & 0.813 \\ 13.4 & 17.13 & 2.013 \\ 6.4 & 23.13 & 0.113 \\ 10.4 & 10.13 & 0.513 \\ 15.4 & 1.13 & 0.613 \end{bmatrix}^T \times \begin{bmatrix} 15.4 & 4.13 & 0.413 \\ 17.4 & 3.13 & 1.413 \\ 22.4 & 16.13 & 0.813 \\ 13.4 & 17.13 & 2.013 \\ 6.4 & 23.13 & 0.113 \\ 10.4 & 10.13 & 0.513 \\ 15.4 & 1.13 & 0.613 \end{bmatrix} = \begin{bmatrix} 229.65 & 139.96 & 13.09 \\ 139.96 & 174.19 & 8.89 \\ 13.09 & 8.89 & 1.08 \end{bmatrix}$$

- Mean vector (M_2).

X_1	X_2	X_3	Y
110	62	37.1	0
115	63	37.2	0
98	62	36.8	0
120	65	37.0	0
118	68	37.1	0
102	58	37.3	0
106	65	36.9	0
111	57	37.0	0
110	**62.5**	**37.05**	

- Covariance matrix (C_2).

X_1	X_2	X_3	Y
110 – 122.6 = –12.6	62 – 71.87 = –9.87	37.1 – 37.79 = –0.687	0
115 – 122.6 = –7.6	63 – 71.87 = –8.87	37.2 – 37.79 = –0.597	0
098 – 122.6 = –24.6	62 – 71.87 = –9.87	36.8 – 37.79 = –0.997	0
120 – 122.6 = –2.6	65 – 71.87 = –6.87	37.0 – 37.79 = –0.797	0
118 – 122.6 = –4.6	68 – 71.87 = –3.87	37.1 – 37.79 = –0.697	0
102 – 122.6 = –20.6	58 – 71.87 = –13.87	37.3 – 37.79 = –0.497	0
106 – 122.6 = –16.6	65 – 71.87 = –6.87	36.9 – 37.79 = –0.897	0
111 – 122.6 = –11.6	57 – 71.87 = –14.87	37.0 – 37.79 = –0.797	0

$$C_2 = \frac{1}{8} \times \begin{bmatrix} -12.6 & -9.87 & -0.687 \\ -7.60 & -8.87 & -0.597 \\ -24.6 & -9.87 & -0.997 \\ -2.60 & -6.87 & -0.797 \\ -4.60 & -3.87 & -0.697 \\ -20.6 & -13.87 & -0.497 \\ -16.6 & -6.87 & -0.897 \\ -11.6 & -14.87 & -0.797 \end{bmatrix}^T \times \begin{bmatrix} -12.6 & -9.87 & -0.687 \\ -7.60 & -8.87 & -0.597 \\ -24.6 & -9.87 & -0.997 \\ -2.60 & -6.87 & -0.797 \\ -4.60 & -3.87 & -0.697 \\ -20.6 & -13.87 & 0.497 \\ -16.6 & -6.87 & -0.897 \\ -11.6 & -14.87 & -0.797 \end{bmatrix}$$

$$C_2 = \begin{bmatrix} 210.51 & 130.31 & 9.67 \\ 130.31 & 99.55 & 6.87 \\ 9.67 & 6.87 & 0.58 \end{bmatrix}$$

Step 3: Compute within-class scatter matrix C.

$$C = \frac{\text{size of first sample}}{\text{size of all given data}} \times C_1 + \frac{\text{size of second sample}}{\text{size of all given data}} \times C_2$$

$$C = \frac{7}{7+8} \times \begin{bmatrix} 229.65 & 139.96 & 13.09 \\ 139.96 & 174.19 & 8.89 \\ 13.09 & 8.89 & 1.08 \end{bmatrix} + \frac{8}{7+8} \times \begin{bmatrix} 210.51 & 130.31 & 9.67 \\ 130.31 & 99.55 & 6.87 \\ 9.67 & 6.87 & 0.58 \end{bmatrix}$$

$$C = \begin{bmatrix} 219.44 & 134.81 & 11.27 \\ 134.81 & 134.38 & 7.81 \\ 11.27 & 7.81 & 0.81 \end{bmatrix}$$

Step 4: Generate discriminant functions (F_i).

$$F_i = M_i \times C^{-1} \times X^T - 0.5 \times M_i \times C^{-1} \times M_1^T + \ln(P_i)$$

$$F_1 = M_1 \times C^{-1} \times X^T - 0.5 \times M_1 \times C^{-1} \times M_1^T + \ln(P_1)$$

$$M_1 \times C^{-1} \times X^T = [137 \quad 82.57 \quad 38.63] \times \begin{bmatrix} 219.44 & 134.81 & 11.27 \\ 134.81 & 134.38 & 7.81 \\ 11.27 & 7.81 & 0.81 \end{bmatrix}^{-1} \times [110 \quad 62 \quad 37.1]^T = 4669.3382$$

$$M_1 \times C^{-1} \times M_1^T = [137 \quad 82.57 \quad 38.63] \times \begin{bmatrix} 219.44 & 134.81 & 11.27 \\ 134.81 & 134.38 & 7.81 \\ 11.27 & 7.81 & 0.81 \end{bmatrix}^{-1} \times [137 \quad 82.57 \quad 38.63]^T = 4696.9699$$

$$F_1 = 4669.3382 - 0.5 \times 4696.9699 + \ln\left[\frac{7}{7+8}\right] = 2320.0911$$

$$F_2 = M_2 \times C^{-1} \times X^T - 0.5 \times M_2 \times C^{-1} \times M_2^T + \ln(P_2)$$

$$M_2 \times C^{-1} \times X^T = [110 \quad 62.5 \quad 37.05] \times \begin{bmatrix} 219.44 & 134.81 & 11.27 \\ 134.81 & 134.38 & 7.81 \\ 11.27 & 7.81 & 0.81 \end{bmatrix}^{-1} \times [110 \quad 62 \quad 37.1]^T = 4636.7741$$

$$M_2 \times C^{-1} \times M_2^T = [110 \quad 62.5 \quad 37.05] \times \begin{bmatrix} 219.44 & 134.81 & 11.27 \\ 134.81 & 134.38 & 7.81 \\ 11.27 & 7.81 & 0.81 \end{bmatrix}^{-1} \times [110 \quad 62.5 \quad 37.05]^T = 4628.2068$$

$$F_2 = 4636.7741 - 0.5 \times 4628.2068 + \ln\left[\frac{8}{7+8}\right] = 2322.0421$$

$2320.0911 < 2322.0421$

$X = [110 \quad 62 \quad 37.1 \quad 0]$

4.2.2. Decision Tree

A decision tree is a hierarchical data structure that represents data through a divide-and-conquer strategy. In classification, the goal is to learn a decision tree that represents the training data such that labels for new examples can be determined. Decision trees are classifiers for instances represented as feature vectors (e.g. colour = ?; shape = ?; label = ?;). Nodes are tests for feature values, leaves specify the label, and at each node, there must be one branch for each value of the feature. One of the most common approaches to decision trees is the ID3 approach.

Example: Draw a decision tree for the following data by using the ID3 algorithm.

Instance	a_1	a_2	a_3	Classification
1	True	Hot	High	No
2	True	Hot	High	No
3	False	Hot	High	Yes
4	False	Cool	Normal	Yes
5	False	Cool	Normal	Yes
6	True	Cool	High	No
7	True	Hot	High	No
8	True	Hot	Normal	Yes
9	False	Cool	Normal	Yes
10	False	Cool	High	Yes

Solution:

$S_{a_1} = [6+, 4-]$ $\quad Entropy_{S_{a_1}} = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0.9709$

$S_{a_{1_{true}}} = [1+, 4-]$ $\quad Entropy_{S_{a_{1_{true}}}} = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.7219$

$S_{a_{1_{False}}} = [5+, 0-]$ $\quad Entropy_{S_{a_{1_{false}}}} = 0$

$Gain(S, a_1) = Entropy_{S_{a_1}} - \sum_{v\in(true,false)} \frac{|S_v|}{|S|} Entropy_{S_V}$

$Gain(S, a_1) = Entropy_{S_{a_1}} - \frac{5}{10} Entropy_{S_{a_{1_{true}}}} - \frac{5}{10} Entropy_{S_{a_{1_{False}}}}$

$Gain(S, a_1) = 0.9709 - \frac{5}{10} \times 0.7219 - \frac{5}{10} \times 0 = 0.6099$

$S_{a_2} = [6+, 4-]$ $\quad Entropy_{S_{a_2}} = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0.9709$

$S_{a_{2_{hot}}} = [2+, 3-]$ $\quad Entropy_{S_{a_{2_{hot}}}} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.9710$

$S_{a_{2_{wool}}} = [4+, 1-]$ $\quad Entropy_{S_{a_{2_{cool}}}} = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.7219$

$Gain(S, a_2) = Entropy_{S_{a_2}} - \sum_{v\in(hot,cool)} \frac{|S_v|}{|S|} Entropy_{S_V}$

$$\text{Gain}(S, a_2) = \text{Entropy}_{S_{a_2}} - \frac{5}{10}\text{Entropy}_{S_{a_2 \text{hot}}} - \frac{5}{10}\text{Entropy}_{S_{a_2 \boldsymbol{w}\text{ool}}}$$

$$\text{Gain}(S, a_2) = 0.9709 - \frac{5}{10} \times 0.9710 - \frac{5}{10} \times 0.7219 = 0.1245$$

$S_{a_3} = [6+, 4-]$ $\quad \text{Entropy}_{S_{a_3}} = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0.9709$

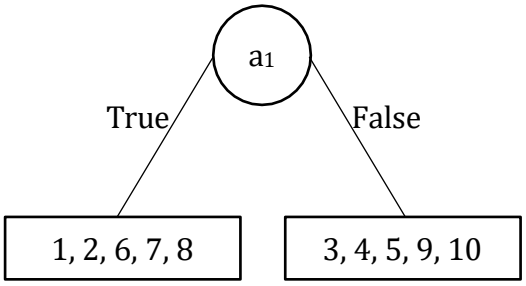
$S_{a_{3\text{high}}} = [2+, 4-]$ $\quad \text{Entropy}_{S_{a_3 \text{high}}} = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = 0.9183$

$S_{a_{3\text{normal}}} = [4+, 0]$ $\quad \text{Entropy}_{S_{a_3 \text{normal}}} = 0$

$$\text{Gain}(S, a_3) = \text{Entropy}_{S_{a_3}} - \sum_{v \in (\text{high,normal})} \frac{|S_v|}{|S|}\text{Entropy}_{S_V}$$

$$\text{Gain}(S, a_3) = \text{Entropy}_{S_{a_3}} - \frac{6}{10}\text{Entropy}_{S_{a_{3\text{high}}}} - \frac{6}{10}\text{Entropy}_{S_{a_{3\text{normal}}}}$$

$$\text{Gain}(S, a_3) = 0.9709 - \frac{6}{10} \times 0.9183 - \frac{4}{10} \times 0 = 0.4199$$



Yes

Instance	a_2	a_3	Classification
1	Hot	High	No
2	Hot	High	No
6	Cool	High	No
7	Hot	High	No
8	Hot	Normal	Yes

$S_{a_2} = [1+, 4-]$ $\quad \text{Entropy}_{S_{a_1}} = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.7219$

$S_{a_{2\text{hot}}} = [1+, 3-]$ $\quad \text{Entropy}_{S_{a_2 \text{hot}}} = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.8113$

$S_{a_{2\boldsymbol{w}\text{ool}}} = [0+, 1-]$ $\quad \text{Entropy}_{S_{a_2 \text{cool}}} = 0$

$$\text{Gain}(S, a_2) = \text{Entropy}_{S_{a_2}} - \sum_{v \in (\text{hot,cool})} \frac{|S_v|}{|S|}\text{Entropy}_{S_V}$$

$$\text{Gain}(S, a_2) = \text{Entropy}_{S_{a_2}} - \frac{4}{5}\text{Entropy}_{S_{a_2 \text{hot}}} - \frac{1}{5}\text{Entropy}_{S_{a_2 \boldsymbol{w}\text{ool}}}$$

$$\text{Gain}(S, a_2) = 0.7219 - \frac{4}{5} \times 0.8113 - \frac{1}{5} \times 0 = 0.0729$$

$S_{a_3} = [1+, 4-]$ $\quad \text{Entropy}_{S_{a_3}} = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.7219$

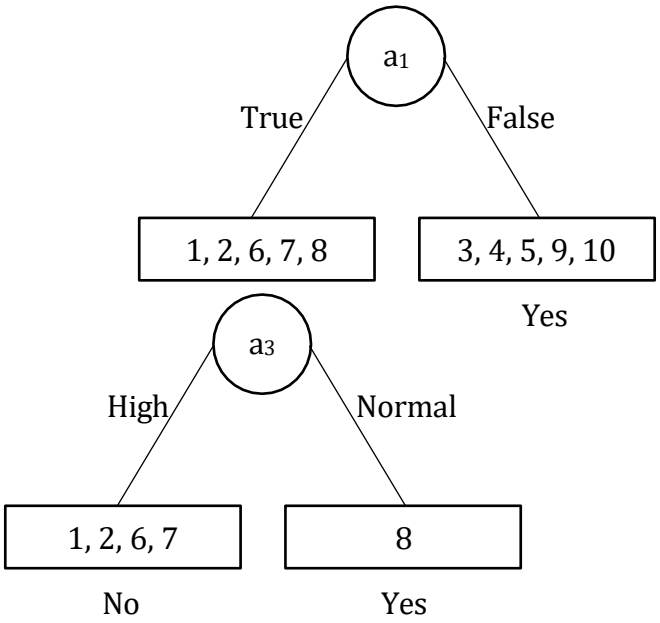
$S_{a_{3\text{high}}} = [0+, 4-]$ $\quad \text{Entropy}_{S_{a_3 \text{high}}} = 0$

$$S_{a_{3_{normal}}} = [1 + ,0] \qquad Entropy_{S_{a_{3}normal}} = 0$$

$$Gain(S, a_3) = Entropy_{S_{a_3}} - \sum_{v \in (high,normal)} \frac{|S_v|}{|S|} Entropy_{S_V}$$

$$Gain(S, a_3) = Entropy_{S_{a_3}} - \frac{4}{5} Entropy_{S_{a_{3_{high}}}} - \frac{1}{5} Entropy_{S_{a_{3_{normal}}}}$$

$$Gain(S, a_3) = 0.7219 - \frac{}{5} \times 0 - \frac{}{5} \times 0 = 0.7219$$



4.2.3. Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised learning algorithm that works best on smaller datasets but on complex ones. The SVM has been introduced as a successful statistical learning approach for classification. SVM has an extremely good generalization capability and strong theoretical foundation. Generalization capability can be defined as the ability of SVM to classify unknown data examples correctly through constructed SVM. This is achieved by learning SVM from training examples which is also known as SVM performance. The SVM manipulates the "curse of dimensionality", which means the computational complexity for the SVM training or testing is not affected by the feature space dimensionality.

There are two types of SVM: binary and multi-class. Binary SVM is the core of SVM. It is capable of distinguishing between two classes. Multi-class SVM expands binary SVM by being able to classify three or more classes. The main algorithm involved in multi-class SVM is to divide the classification problem to many binary problems with its own

classifier. Two main methods are related to SVM multi-class classification: One-Against-All (OAA) and One-Against-One (OAO).

Example: Suppose you have the following positively labelled data points

$$\{\begin{pmatrix}3\\1\end{pmatrix}, \begin{pmatrix}+3\\-1\end{pmatrix}, \begin{pmatrix}6\\1\end{pmatrix}, \begin{pmatrix}+6\\-1\end{pmatrix}\}$$

and the following negatively labelled data points:

$$\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}+0\\-1\end{pmatrix}, \begin{pmatrix}-1\\+0\end{pmatrix}\}$$

using SVM to discriminate between these two classes.

Solution:

Hyperplane equation y = w̃x + b

$\tilde{w} = \sum_i \alpha_i S_i$

$S_i = \{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}3\\1\end{pmatrix}, \begin{pmatrix}+3\\-1\end{pmatrix}\}$

$S_i \{\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}3\\1\end{pmatrix}, \begin{pmatrix}+3\\+1\end{pmatrix}\}$

$\alpha_1 S_1 S_1 + \alpha_2 S_2 S_1 + \alpha_3 S_3 S_1 = -1$

$\alpha_1 S_1 S_2 + \alpha_2 S_2 S_2 + \alpha_3 S_3 S_2 = +1$

$\alpha_1 S_1 S_3 + \alpha_2 S_2 S_3 + \alpha_3 S_3 S_3 = +1$

$$\alpha_1 \begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}+3\\+1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}3\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\end{pmatrix}\begin{pmatrix}3\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}-3\\1\end{pmatrix}\begin{pmatrix}3\\1\end{pmatrix} = +1$$

$$\alpha_1 \begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}+3\\+1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\end{pmatrix}\begin{pmatrix}+3\\+1\end{pmatrix} + \alpha_3 \begin{pmatrix}+3\\+1\end{pmatrix}\begin{pmatrix}+3\\+1\end{pmatrix} = +1$$

$\alpha_1(1+0+1) + \alpha_2(3+0+1) + \alpha_3(3+0+1) = -1$

$\alpha_1(3+0+1) + \alpha_2(9+1+1) + \alpha_3(9-1+1) = +1$

$\alpha_1(3+0+1) + \alpha_2(9-1+1) + \alpha_3(9+1+1) = +1$

$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$

$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = +1$

$4\alpha_1 + 9\alpha_2 + 11\alpha_3 = +1$

$\alpha_1 = -3.5$

$\alpha_2 = 0.75$

$\alpha_3 = 0.75$

$\tilde{w} = \sum_i \alpha_i S_i$

$$\tilde{w} = -3.5 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

$$\tilde{w} = \begin{pmatrix} +1 \\ +0 \\ -2 \end{pmatrix}$$

Hyperplane equation y = w̃x + b

$$\tilde{w} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, b = -2$$

4.2.4. Ensemble Methods: Random Forest

Ensemble learning is a model that makes predictions based on several different models. By combining individual models, the ensemble model tends to be more flexible (less bias) and less data-sensitive (less variance). It is an adaptive learning methodology to combines several algorithms to extract better results than individual performances. The two most popular ensemble methods are bagging and boosting.

- **Bagging:** Training a bunch of individual models in a parallel way. Each model is trained by a random subset of the data.
- **Boosting:** Training a bunch of individual models sequentially. Each model learns from mistakes made by the previous model.

Random forest is an ensemble model using bagging as the ensemble method and a decision tree as the individual model. Random forests use two types of randomness. First, in growing any given tree, a random sample of predictors is selected at each node in choosing the best split. A further layer of randomness is added by using a random sample of observations for growing each tree in the first place. In theory, using a random sample of observations and selecting random predictors at each node should reduce dependence between covariates and thus between the resulting trees. The steps for applying random forest are as follows:

Step 1: Select n (e.g., 1000) random subsets from the training set.
Step 2: Train n (e.g., 1000) decision trees.

- One random subset is used to train one decision tree.
- The optimal splits for each decision tree are based on a random subset of features (e.g., 10 features in total, randomly select 5 out of 10 features to split).

Step 3: Each tree predicts the records/candidates in the test set, independently.

Step 4: Make the final prediction. For each candidate in the test set, random forest uses the class (e.g., cat or dog) with the majority vote as this candidate's final prediction.

The advantages of the random forest algorithm can be summarized as follows:

- For applications in classification problems, the random forest algorithm will avoid the overfitting problem.
- For both classification and regression tasks, the same random forest algorithm can be used.
- The random forest algorithm can be used for identifying the most important features from the training dataset, in other words, feature engineering.
- In training data, they are less sensitive to outlier data.
- Parameters can be set easily and therefore, eliminate the need for pruning the trees variable importance and accuracy are generated automatically.

The disadvantages of the random forest algorithm can be summarized as follows:

- When using a random forest, more resources are required for computation.
- It consumes more time compared to a decision tree algorithm.

CH. 5: Unsupervised Learning

5.1. Introduction

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyse and cluster unlabelled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information makes it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition. Some of the most common real-world applications of unsupervised learning are:

- **News Sections:** Google News uses unsupervised learning to categorize articles on the same story from various online news outlets. For example, the results of a presidential election could be categorized under their label for "US" news.
- **Computer vision:** Unsupervised learning algorithms are used for visual perception tasks, such as object recognition.
- **Medical imaging:** Unsupervised machine learning provides essential features to medical imaging devices, such as image detection, classification and segmentation, used in radiology and pathology to diagnose patients quickly and accurately.
- **Anomaly detection:** Unsupervised learning models can comb through large amounts of data and discover atypical data points within a dataset. These anomalies can raise awareness around faulty equipment, human error, or breaches in security.
- **Customer personas:** Defining customer personas makes it easier to understand common traits and business clients' purchasing habits. Unsupervised learning allows businesses to build better buyer persona profiles, enabling organizations to align their product messaging more appropriately.
- **Recommendation Engines:** Using past purchase behaviour data, unsupervised learning can help to discover data trends that can be used to develop more effective cross-selling strategies. This is used to make relevant add-on recommendations to customers during the checkout process for online retailers.

Unsupervised learning models are utilized for three main tasks—clustering, association, and dimensionality reduction.

5.1.1. Clustering

Clustering is a data mining technique in which groups of unlabelled data are based on their similarities or differences. Clustering algorithms are used to process raw, unclassified data objects into groups represented by structures or patterns in the information. Clustering algorithms can be categorized into a few types, specifically exclusive, overlapping, hierarchical, and probabilistic.

- **Exclusive and Overlapping Clustering**

Exclusive clustering is a form of grouping that stipulates a data point can exist only in one cluster. This can also be referred to as "hard" clustering. The K-means clustering algorithm is an example of exclusive clustering.

> **K-means clustering** is a common example of an exclusive clustering method where data points are assigned into K groups, where K represents the number of clusters based on the distance from each group's centroid. The data points closest to a given centroid will be clustered under the same category. A larger K value will be indicative of smaller groupings with more granularity whereas a smaller K value will have larger groupings and less granularity. K-means clustering is commonly used in market segmentation, document clustering, image segmentation, and image compression.

Overlapping clusters differ from exclusive clustering in that they allow data points to belong to multiple clusters with separate degrees of membership. "Soft" or fuzzy k-means clustering is an example of overlapping clustering.

- **Hierarchical clustering**

Hierarchical clustering, also known as Hierarchical Cluster Analysis (HCA), is an unsupervised clustering algorithm that can be categorized in two ways; they can be agglomerative or divisive. Agglomerative clustering is considered a "bottom-up approach." Its data points are isolated as separate groupings initially, and then they are merged iteratively based on similarity until one cluster has been achieved. Four different methods are commonly used to measure similarity:

- **Ward's linkage:** This method states that the distance between two clusters is defined by the increase in the sum of squared after the clusters are merged.
- **Average linkage:** This method is defined by the mean distance between two points in each cluster.
- **Complete (or maximum) linkage:** This method is defined by the maximum distance between two points in each cluster.
- **Single (or minimum) linkage:** This method is defined by the minimum distance between two points in each cluster.

Euclidean distance is the most common metric used to calculate these distances; however, other metrics, such as Manhattan distance, are also cited in clustering literature.

Divisive clustering can be defined as the opposite of agglomerative clustering; instead, it takes a "top-down" approach. In this case, a single data cluster is divided based on the differences between data points. Divisive clustering is not commonly used, but it is still worth noting in the context of hierarchical clustering. These clustering processes are usually visualized using a dendrogram, a tree-like diagram that documents the merging or splitting of data points at each iteration.

– **Probabilistic clustering**

A probabilistic model is an unsupervised technique that helps to solve density estimation or "soft" clustering problems. In probabilistic clustering, data points are clustered based on the likelihood that they belong to a particular distribution. The Gaussian Mixture Model (GMM) is one of the most commonly used probabilistic clustering methods.

Gaussian Mixture Models are classified as mixture models, which means that they are made up of an unspecified number of probability distribution functions. GMMs are primarily leveraged to determine which Gaussian, or normal, probability distribution a given data point belongs to. If the mean or variance is known, then we can determine which distribution a given data point belongs to. However, in GMMs, these variables are not known, so we assume that a latent, or hidden, variable exists to cluster data points appropriately. While it is not required to use the Expectation-Maximization (EM) algorithm, it is commonly used to estimate the assignment probabilities for a given data point to a particular data cluster.

5.1.2. Associative Rule

An association rule is a rule-based method for finding relationships between variables in a given dataset. These methods are frequently used for market basket analysis, allowing companies to better understand relationships between different products. Understanding the consumption habits of customers enables businesses to develop better cross-selling strategies and recommendation engines. Examples of this can be seen in Amazon's "Customers Who Bought This Item Also Bought" or Spotify's "Discover Weekly" playlist. While there are a few different algorithms used to generate association rules, such as Apriori, Eclat, and FP-Growth, the Apriori algorithm is most widely used.

Apriori algorithms

Apriori algorithms have been popularized through market basket analyses, leading to different recommendation engines for music platforms and online retailers. They are used within transactional datasets to identify frequent itemsets, or collections of items, to identify the likelihood of consuming a product given the consumption of another product. For example, if I play Black Sabbath's radio on Spotify, starting with their song "Orchid", one of the other songs on this channel will likely be a Led Zeppelin song, such as "Over the Hills and Far Away." This is based on my prior listening habits as well as the ones of others. Apriori algorithms use a hash tree to count itemsets, navigating through the dataset in a breadth-first manner.

Example: You have the following data:

Transaction ID	Items
T_1	I_1, I_2, I_3
T_2	I_2, I_3, I_4
T_3	I_4, I_5
T_4	I_1, I_2, I_4
T_5	I_1, I_2, I_3, I_5
T_6	I_1, I_2, I_3, I_4

Find the frequent itemsets and generate association rules on this. Assume that the minimum support threshold is 50% and the minimum confidence threshold is 60%.

Solution:

Min_sup = minimum support threshold × number of transactions

Min_sup = 0.5 × 6 = 3

Items	Count
I_1	4
I_2	5
I_3	4
I_4	4
I_5	2

Items	Count
I_1	4
I_2	5
I_3	4
I_4	4

Items	Count
I_1, I_2	4
I_1, I_3	3
I_1, I_4	2
I_2, I_3	4
I_2, I_4	3
I_3, I_4	2

Items	Count
I_1, I_2	4
I_1, I_3	3
I_2, I_3	4
I_2, I_4	3

Items	Count
I_1, I_2, I_3	3
I_1, I_2, I_4	2
I_1, I_3, I_4	1
I_2, I_3, I_4	2

Only I_1, I_2, I_3 is frequent item

$I_1, I_2 \rightarrow I_3$

Confidence = support $\{I_1, I_2, I_3\}$ / support $\{I_1, I_2\}$ = $(3/4) \times 100 = 75\%$ accepted

$I_1, I_3 \rightarrow I_2$

Confidence = support $\{I_1, I_2, I_3\}$ / support $\{I_1, I_3\}$ = $(3/3) \times 100 = 100\%$ accepted

$I_2, I_3 \rightarrow I_1$

Confidence = support $\{I_1, I_2, I_3\}$ / support $\{I_2, I_3\}$ = $(3/4) \times 100 = 75\%$ accepted

$I_1 \rightarrow I_2, I_3$

Confidence = support $\{I_1, I_2, I_3\}$ / support $\{I_1\}$ = $(3/4) \times 100 = 75\%$ accepted

$I_2 \rightarrow I_1, I_3$

Confidence = support $\{I_1, I_2, I_3\}$ / support $\{I_2\}$ = (3/ 5) × 100 = 60% accepted

$I_3 \rightarrow I_1, I_2$

Confidence = support $\{I_1, I_2, I_3\}$ / support $\{I_3\}$ = (3/ 4) × 100 = 75% accepted

5.1.3. Dimensionality Reduction

Dimensionality reduction is commonly used in the pre-processing data stage, and there are a few different dimensionality reduction methods that can be used, such as:

- **Principal component analysis**

Principal Component Analysis (PCA) is a type of dimensionality reduction algorithm which is used to reduce redundancies and compress datasets through feature extraction. This method uses a linear transformation to create a new data representation, yielding a set of "principal components." The first principal component is the direction which maximizes the variance of the dataset. While the second principal component also finds the maximum variance in the data, it is completely uncorrelated to the first principal component, yielding a direction that is perpendicular, or orthogonal, to the first component. This process repeats based on the number of dimensions, where the next principal component is the direction orthogonal to the prior components with the most variance.

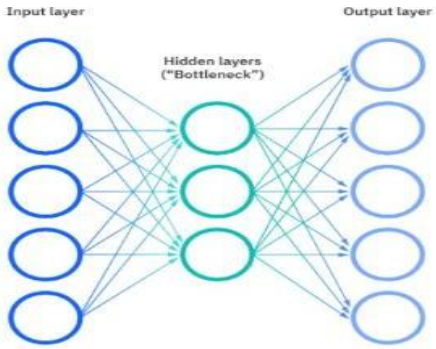
- **Singular value decomposition**

Singular Value Decomposition (SVD) is another dimensionality reduction approach which factorizes a matrix, A, into three, low-rank matrices. SVD is denoted by the formula, A = USVT, where U and V are orthogonal matrices. S is a diagonal matrix, and S values are considered singular values of matrix A. Similar to PCA, it is commonly used to reduce noise and compress data, such as image files.

- **Autoencoders**

Autoencoders leverage neural networks to compress data and then recreate a new representation of the original data's input. Looking at the image below, you can see that the hidden layer specifically acts as a bottleneck to compress the input layer before reconstructing it within the output layer. The stage from the input layer to the hidden

layer is referred to as “encoding” while the stage from the hidden layer to the output layer is known as “decoding.”



5.2. Mining Challenges for Big Data Analytics

Big Data Mining (BDM) is an approach that uses cumulative data mining or extraction techniques on large datasets/volumes of data. It is mainly focused on retrieving relevant and demanded information (or patterns) and thus extracting value hidden in data of an immense volume. BDM draws from the conventional data mining notation but also combines the aspects of big data, i.e. it enables to acquisition of useful information from databases or data streams that are huge in terms of “big data V’s”, like volume, velocity, and variety. The mining challenges for big data analytics can be summarized as follows:

- Generate business value.
- Provision of competitive advantage, and generation of new business ideas from big data insights.
- Data inconsistency and incompleteness, scalability, timeliness, uncertain, dynamic data, and data security.
- Data capture, storage, searching, sharing, analysis, and visualization.
- Lack of:
    - Large-scale data representation (for mining purposes).
    - Effective and efficient online large-scale machine learning techniques.
    - Data confidentiality mechanism.
- Also, the constant inflow of data to be mined can be recognized as a momentous challenge, as many mining algorithms do not provide proper sequences or patterns.
- Other challenges are associated with the mining process itself. The algorithms and techniques used for “classic” data mining e.g., data warehouses sometimes are not suited to be used with huge amounts of constantly incoming big data. This is so

because traditional data mining approaches start with a centralized data repository, able to store and process data. With the prodigious size and variety characterizing big data, such a centralized approach may not be used. There is a strong need for more distributed approaches capable of mining huge amounts of unstructured data.

CH. 6: Prescriptive Analytics

6.1. Introduction

Prescriptive analysis answers the question, “How will it happen?” Mix all the insights gained from the other data analysis types, and you have prescriptive analysis. Sometimes, an issue cannot be solved solely with one analysis type and instead requires multiple insights.

Machine learning algorithms are often used in prescriptive analytics to parse through large amounts of data faster—and often more efficiently—than humans can. Using “if” and “else” statements, algorithms comb through data and make recommendations based on a specific combination of requirements. For instance, if at least 50 per cent of customers in a dataset selected that they were “very unsatisfied” with your customer service team, the algorithm may recommend additional training.

It’s important to note that while algorithms can provide data-informed recommendations, they cannot replace human discernment. Prescriptive analytics is a tool to inform decisions and strategies and should be treated as such. Your judgment is valuable and necessary to provide context and guardrails to algorithmic outputs.

6.2. Creating Data for Analytics through Designed Experiment

Design Of Experiments (DOE) also referred to as Designed Experiments or Experimental Design – is defined as the systematic procedure carried out under controlled conditions to discover an unknown effect, to test or establish a hypothesis, or to illustrate a known effect. It involves determining the relationship between input factors affecting a process and the output of that process. It helps to manage process inputs to optimize the output. DOE is a powerful data collection and analysis tool that can be used in a variety of experimental situations.

It allows manipulating multiple input factors and determining their effect on a desired output (response). By changing multiple inputs at the same time, DOE helps to identify important interactions that may be missed when experimenting with only one factor at a time. We can investigate all possible combinations (full factorial) or only a portion of the possible combinations (fractional factorial).

While doing the interior design of a new house, the final effect of interior design will depend on various factors such as the colour of walls, lights, floors, placements of various objects in the house, sizes and shapes of the objects and many more. Each of these factors will have an impact on the outcome of interior decoration. While variation in each factor alone can impact, a variation in a combination of these factors at the same time also will impact the outcome.

Hence it needs to be studied how each of these factors impact the outcome, which are the critical factors impacting the most, which are the most important combination of these factors impacting the outcome significantly. The interior designer can plan and conduct some experiments.

A well-planned and executed experiment may provide a great deal of information about the effect on a response variable due to one or more factors. Many experiments involve holding certain factors constant and altering the levels of another variable. This "One Factor At a Time (OFAT)" approach to process knowledge is, however, inefficient when compared with changing multiple factor levels simultaneously.

A well-performed experiment may provide answers to the following such as:

- What are the key factors in a process? (both controllable and uncontrollable)
- At what settings would the process deliver acceptable performance?
- What is the key, main and interaction effects in the process?
- What settings would bring about less variation in the output?

6.2.1. Steps for Design of Experiments

We need to follow the below steps in sequence for conducting a DOE:

1. Define the problem(s).
2. Determine objective(s).
3. Brainstorm.
4. Design experiments.
5. Conduct experiments and collect data.
6. Analyse data.
7. Interpret results.

8. Verify predicted results.

6.2.2. Advantages of Design of Experiments

DOE has been in use for many years in the manufacturing industry. Below are some of the benefits/improvements we can expect from conducting DOEs:

- Reduce time to design/develop new products and processes.
- Improve the performance of existing processes.
- Improve reliability and performance of products.
- Achieve product and process robustness.
- Evaluation of materials, design alternatives, setting component and system tolerances, etc.

6.2.3. Purposes of Design of Experiments

DOE can be used for the following purposes:

1. Comparisons.
2. Variable Screening.
3. Transfer Function Exploration.
4. System Optimization.
5. System Robustness.

6.2.4. Common Design Types

Different designs have been used for different experiment purposes. The following list gives the commonly used design types.

- For comparison:
    - One-factor design.
- For variable screening:
    - Two-level factorial design.
    - Taguchi orthogonal array.
    - Plackett-Burman design.
- For transfer function identification and optimization:
    - Central composite design.

  - Box-Behnken design.
- For system robustness:
  - Taguchi robust design.

6.3. Creating Data for Analytics through Reinforcement Learning

Reinforcement Learning (RL) is defined as a machine learning method that is concerned with how software agents should take actions in an environment. RL is a part of the deep learning method that helps you to maximize some portion of the cumulative reward.

Here are some important terms used in RL:

- **Agent:** It is an assumed entity which performs actions in an environment to gain some reward.
- **Environment (e):** A scenario that an agent has to face.
- **Reward (R):** An immediate return given to an agent when he or she performs a specific action or task.
- **State (s):** State refers to the current situation returned by the environment.
- **Policy (π):** It is a strategy applied by the agent to decide the next action based on the current state.
- **Value (V):** It is an expected long-term return with a discount, as compared to the short-term reward.
- **Value Function:** It specifies the value of a state which is the total amount of reward. It is an agent which should be expected beginning from that state.
- **Model of the environment:** This mimics the behaviour of the environment. It helps you to make inferences to be made and also determine how the environment will behave.
- **Model-based methods:** It is a method for solving RL problems which use model-based methods.
- **Q value or action value (Q):** Q value is quite similar to value. The only difference between the two is that it takes an additional parameter as a current action.

6.3.1. Reinforcement Learning Algorithms

There are three approaches to implementing an RL algorithm. These algorithms are:

- Value-based: In a value-based RL method, you should try to maximize a value function **V(s)**. In this method, the agent is expecting a long-term return of the current states under policy **π**.
- Policy-based: In a policy-based RL method, you try to come up with such a policy that the action performed in every state helps you to gain maximum reward in the future. Two types of policy-based methods are:
    - Deterministic: For any state, the same action is produced by the policy π.
    - Stochastic: Every action has a certain probability, which is determined by the stochastic policy equation.
- Model-based: In this RL method, you need to create a virtual model for each environment. The agent learns to perform in that specific environment.

6.3.2. Characteristics of Reinforcement Learning

The characteristics of RL can be summarized as follows:

- There is no supervisor, only a real number or reward signal.
- Sequential decision-making.
- Time plays a crucial role in reinforcement problems.
- Feedback is always delayed, not instantaneous.
- Agent's actions determine the subsequent data it receives.

6.3.3. Types of Reinforcement Learning

Two types of RL methods are:

- Positive: It is defined as an event that occurs because of specific behaviour. It increases the strength and the frequency of the behaviour and impacts positively on the action taken by the agent. This type of reinforcement helps you to maximize performance and sustain change for a more extended period. However, too much reinforcement may lead to over-optimization of the state, which can affect the results.
- Negative: Negative reinforcement is defined as the strengthening of behaviour that occurs because of a negative condition which should have been stopped or avoided. It helps you to define the minimum standard of performance. However, the drawback of this method is that it provides enough to meet up the minimum behaviour.

6.3.4. Learning Models of Reinforcement Learning

There are two important learning models in RL. These models are:

- Markov Decision Process. The following parameters are used to get a solution:
  - Set of actions, A.
  - Set of states, S.
  - Reward, R.
  - Policy, n.
  - Value, V.
- Q-Learning: It is a value-based method of supplying information to inform which action an agent should take.

6.3.5. Applications Fields of Reinforcement Learning

Here are applications of RL:

- Robotics for industrial automation.
- Business strategy planning.
- Machine learning and data processing.
- It helps you to create training systems that provide custom instruction and materials according to the requirements of students.
- Aircraft control and robot motion control.

6.3.6. Reasons for using Reinforcement Learning

Here are prime reasons for using RL:

- It helps you to find which situation needs action.
- Helps you to discover which action yields the highest reward over a longer period.
- RL also provides the learning agent with a reward function.
- It also allows it to figure out the best method for obtaining large rewards.

6.3.7. When Not to Use Reinforcement Learning?

You cannot apply the RL model in all situations. Here are some conditions when you should not use the RL model:

- When you have enough data to solve the problem with a supervised learning method.
- You need to remember that RL is computing-heavy and time-consuming. In particular when the action space is large.

6.3.8. Challenges of Reinforcement Learning

Here are the major challenges you will face while doing RL:

- Feature/reward design which should be very involved.
- Parameters may affect the speed of learning.
- Realistic environments can have partial observability.
- Too much reinforcement may lead to an overload of states which can diminish the results.
- Realistic environments can be non-stationary.

6.4. Creating Data for Analytics through Active Learning

Active learning is the subset of machine learning in which a learning algorithm can query a user interactively to label data with the desired outputs. In active learning, the algorithm proactively selects the subset of examples to be labelled next from the pool of unlabelled data. The fundamental belief behind the active learner algorithm concept is that a machine learning algorithm could potentially reach a higher level of accuracy while using a smaller number of training labels if it were allowed to choose the data it wants to learn from.

Therefore, active learners are allowed to interactively pose queries during the training stage. These queries are usually in the form of unlabelled data instances and the request is to a human annotator to label the instance. This makes active learning part of the human-in-the-loop paradigm, where it is one of the most powerful examples of success.

Active learning is closer to traditional supervised learning. It is a type of semi-supervised learning, meaning models are trained using both labelled and unlabelled data. The idea behind semi-supervised learning is that labelling just a small sample of data might result in the same accuracy or better than fully labelled training data. The only challenge is determining what that sample is. Active learning machine learning is all about labelling

data dynamically and incrementally during the training phase so that the algorithm can identify what label would be the most beneficial for it to learn from.

6.4.1. Categories of Active Learning

- Stream-based selective sampling: the algorithm determines if it would be beneficial enough to query for the label of a specific unlabelled entry in the dataset. While the model is being trained, it is presented with a data instance and immediately decides if it wants to query the label. This approach has a natural disadvantage that comes from the lack of guarantee that the data scientist will stay within budget.
- Pool-based sampling: this is the most well-known scenario for active learning. In this sampling method, the algorithm attempts to evaluate the entire dataset before it selects the best query or set of queries. The active learner algorithm is often initially trained on a fully labelled part of the data which is then used to determine which instances would be most beneficial to insert into the training set for the next active learning loop. The downside of this method is the amount of memory it can require.
- Membership query synthesis: this does not apply to all cases, because it involves the generation of synthetic data. The active learner in this method is allowed to create their examples for labelling. This method is compatible with problems where it is easy to generate a data instance.

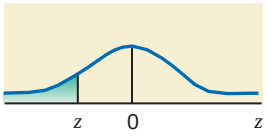
6.4.2. Where should we apply active learning?

Active learning can be used in the following situations:

- We have a very small amount or a huge amount of dataset.
- Annotation of the unlabelled dataset costs human effort, time, and money.
- We have access to limited processing power.

Table IV Standard Normal Distribution Table

The entries in this table give the cumulative area under the standard normal curve to the left of z with the values of z equal to 0 or negative.

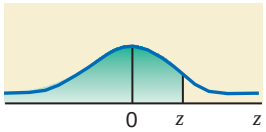


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

(continued on next page)

Table IV Standard Normal Distribution Table (continued from previous page)

The entries in this table give the cumulative area under the standard normal curve to the left of z with the values of z equal to 0 or positive.

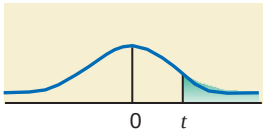


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

This is Table IV of Appendix C.

Table V The *t* Distribution Table

The entries in this table give the critical values of t for the specified number of degrees of freedom and areas in the right tail.



	Area in the Right Tail under the *t* Distribution Curve					
df	**.10**	**.05**	**.025**	**.01**	**.005**	**.001**
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
31	1.309	1.696	2.040	2.453	2.744	3.375
32	1.309	1.694	2.037	2.449	2.738	3.365
33	1.308	1.692	2.035	2.445	2.733	3.356
34	1.307	1.691	2.032	2.441	2.728	3.348
35	1.306	1.690	2.030	2.438	2.724	3.340

(continued on next page)

Table V The *t* Distribution Table (continued from previous page)

	Area in the Right Tail under the *t* Distribution Curve					
df	**.10**	**.05**	**.025**	**.01**	**.005**	**.001**
36	1.306	1.688	2.028	2.434	2.719	3.333
37	1.305	1.687	2.026	2.431	2.715	3.326
38	1.304	1.686	2.024	2.429	2.712	3.319
39	1.304	1.685	2.023	2.426	2.708	3.313
40	1.303	1.684	2.021	2.423	2.704	3.307
41	1.303	1.683	2.020	2.421	2.701	3.301
42	1.302	1.682	2.018	2.418	2.698	3.296
43	1.302	1.681	2.017	2.416	2.695	3.291
44	1.301	1.680	2.015	2.414	2.692	3.286
45	1.301	1.679	2.014	2.412	2.690	3.281
46	1.300	1.679	2.013	2.410	2.687	3.277
47	1.300	1.678	2.012	2.408	2.685	3.273
48	1.299	1.677	2.011	2.407	2.682	3.269
49	1.299	1.677	2.010	2.405	2.680	3.265
50	1.299	1.676	2.009	2.403	2.678	3.261
51	1.298	1.675	2.008	2.402	2.676	3.258
52	1.298	1.675	2.007	2.400	2.674	3.255
53	1.298	1.674	2.006	2.399	2.672	3.251
54	1.297	1.674	2.005	2.397	2.670	3.248
55	1.297	1.673	2.004	2.396	2.668	3.245
56	1.297	1.673	2.003	2.395	2.667	3.242
57	1.297	1.672	2.002	2.394	2.665	3.239
58	1.296	1.672	2.002	2.392	2.663	3.237
59	1.296	1.671	2.001	2.391	2.662	3.234
60	1.296	1.671	2.000	2.390	2.660	3.232
61	1.296	1.670	2.000	2.389	2.659	3.229
62	1.295	1.670	1.999	2.388	2.657	3.227
63	1.295	1.669	1.998	2.387	2.656	3.225
64	1.295	1.669	1.998	2.386	2.655	3.223
65	1.295	1.669	1.997	2.385	2.654	3.220
66	1.295	1.668	1.997	2.384	2.652	3.218
67	1.294	1.668	1.996	2.383	2.651	3.216
68	1.294	1.668	1.995	2.382	2.650	3.214
69	1.294	1.667	1.995	2.382	2.649	3.213
70	1.294	1.667	1.994	2.381	2.648	3.211
71	1.294	1.667	1.994	2.380	2.647	3.209
72	1.293	1.666	1.993	2.379	2.646	3.207
73	1.293	1.666	1.993	2.379	2.645	3.206
74	1.293	1.666	1.993	2.378	2.644	3.204
75	1.293	1.665	1.992	2.377	2.643	3.202
∞	1.282	1.645	1.960	2.326	2.576	3.090

This is Table V of Appendix C.

Table entry for p is the critical value F^* with probability p lying to its right.

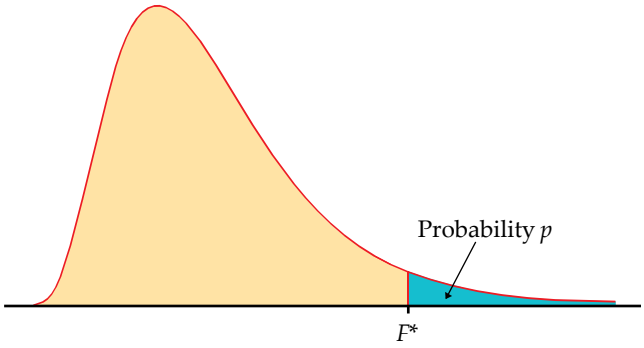


TABLE E

F critical values

Degrees of freedom in the denominator		p	Degrees of freedom in the numerator								
			1	2	3	4	5	6	7	8	9
		.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
		.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	1	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
		.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
		.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
		.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
		.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	2	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
		.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
		.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
		.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
		.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	3	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
		.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
		.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
		.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
		.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	4	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
		.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
		.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
		.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
		.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	5	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
		.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
		.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24
		.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
		.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	6	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
		.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
		.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69
		.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
		.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	7	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
		.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
		.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33

Table entry for p is the critical value F^* with probability p lying to its right.

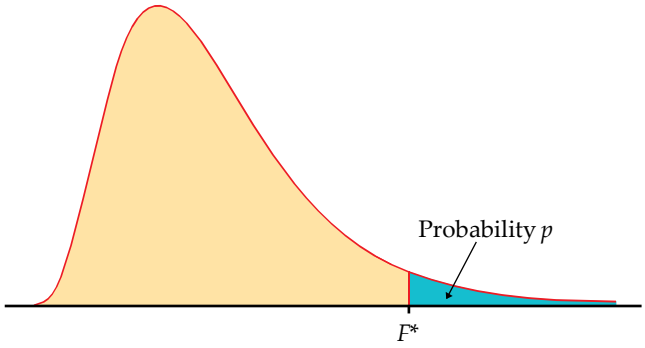


TABLE E

F critical values (continued)

Degrees of freedom in the numerator										
10	12	15	20	25	30	40	50	60	120	1000
60.19	60.71	61.22	61.74	62.05	62.26	62.53	62.69	62.79	63.06	63.30
241.88	243.91	245.95	248.01	249.26	250.10	251.14	251.77	252.20	253.25	254.19
968.63	976.71	984.87	993.10	998.08	1001.4	1005.6	1008.1	1009.8	1014.0	1017.7
6055.8	6106.3	6157.3	6208.7	6239.8	6260.6	6286.8	6302.5	6313.0	6339.4	6362.7
605621	610668	615764	620908	624017	626099	628712	630285	631337	633972	636301
9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.49
19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.48	19.49	19.49
39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.48	39.49	39.50
99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.48	99.49	99.50
999.40	999.42	999.43	999.45	999.46	999.47	999.47	999.48	999.48	999.49	999.50
5.23	5.22	5.20	5.18	5.17	5.17	5.16	5.15	5.15	5.14	5.13
8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.58	8.57	8.55	8.53
14.42	14.34	14.25	14.17	14.12	14.08	14.04	14.01	13.99	13.95	13.91
27.23	27.05	26.87	26.69	26.58	26.50	26.41	26.35	26.32	26.22	26.14
129.25	128.32	127.37	126.42	125.84	125.45	124.96	124.66	124.47	123.97	123.53
3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.76
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.70	5.69	5.66	5.63
8.84	8.75	8.66	8.56	8.50	8.46	8.41	8.38	8.36	8.31	8.26
14.55	14.37	14.20	14.02	13.91	13.84	13.75	13.69	13.65	13.56	13.47
48.05	47.41	46.76	46.10	45.70	45.43	45.09	44.88	44.75	44.40	44.09
3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.12	3.11
4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.44	4.43	4.40	4.37
6.62	6.52	6.43	6.33	6.27	6.23	6.18	6.14	6.12	6.07	6.02
10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.20	9.11	9.03
26.92	26.42	25.91	25.39	25.08	24.87	24.60	24.44	24.33	24.06	23.82
2.94	2.90	2.87	2.84	2.81	2.80	2.78	2.77	2.76	2.74	2.72
4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.75	3.74	3.70	3.67
5.46	5.37	5.27	5.17	5.11	5.07	5.01	4.98	4.96	4.90	4.86
7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.09	7.06	6.97	6.89
18.41	17.99	17.56	17.12	16.85	16.67	16.44	16.31	16.21	15.98	15.77
2.70	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.51	2.49	2.47
3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.32	3.30	3.27	3.23
4.76	4.67	4.57	4.47	4.40	4.36	4.31	4.28	4.25	4.20	4.15
6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.82	5.74	5.66
14.08	13.71	13.32	12.93	12.69	12.53	12.33	12.20	12.12	11.91	11.72

(*Continued*)

TABLE E

F critical values (continued)

Degrees of freedom in the denominator		Degrees of freedom in the numerator								
	p	1	2	3	4	5	6	7	8	9
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96
11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48
13	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98
14	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58
15	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26
16	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98
17	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	.010	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75

TABLE E

F critical values (continued)

Degrees of freedom in the numerator										
10	12	15	20	25	30	40	50	60	120	1000
2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.30
3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.02	3.01	2.97	2.93
4.30	4.20	4.10	4.00	3.94	3.89	3.84	3.81	3.78	3.73	3.68
5.81	5.67	5.52	5.36	5.26	5.20	5.12	5.07	5.03	4.95	4.87
11.54	11.19	10.84	10.48	10.26	10.11	9.92	9.80	9.73	9.53	9.36
2.42	2.38	2.34	2.30	2.27	2.25	2.23	2.22	2.21	2.18	2.16
3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.80	2.79	2.75	2.71
3.96	3.87	3.77	3.67	3.60	3.56	3.51	3.47	3.45	3.39	3.34
5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.48	4.40	4.32
9.89	9.57	9.24	8.90	8.69	8.55	8.37	8.26	8.19	8.00	7.84
2.32	2.28	2.24	2.20	2.17	2.16	2.13	2.12	2.11	2.08	2.06
2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.64	2.62	2.58	2.54
3.72	3.62	3.52	3.42	3.35	3.31	3.26	3.22	3.20	3.14	3.09
4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.08	4.00	3.92
8.75	8.45	8.13	7.80	7.60	7.47	7.30	7.19	7.12	6.94	6.78
2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.00	1.98
2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.51	2.49	2.45	2.41
3.53	3.43	3.33	3.23	3.16	3.12	3.06	3.03	3.00	2.94	2.89
4.54	4.40	4.25	4.10	4.01	3.94	3.86	3.81	3.78	3.69	3.61
7.92	7.63	7.32	7.01	6.81	6.68	6.52	6.42	6.35	6.18	6.02
2.19	2.15	2.10	2.06	2.03	2.01	1.99	1.97	1.96	1.93	1.91
2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.40	2.38	2.34	2.30
3.37	3.28	3.18	3.07	3.01	2.96	2.91	2.87	2.85	2.79	2.73
4.30	4.16	4.01	3.86	3.76	3.70	3.62	3.57	3.54	3.45	3.37
7.29	7.00	6.71	6.40	6.22	6.09	5.93	5.83	5.76	5.59	5.44
2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.85
2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.31	2.30	2.25	2.21
3.25	3.15	3.05	2.95	2.88	2.84	2.78	2.74	2.72	2.66	2.60
4.10	3.96	3.82	3.66	3.57	3.51	3.43	3.38	3.34	3.25	3.18
6.80	6.52	6.23	5.93	5.75	5.63	5.47	5.37	5.30	5.14	4.99
2.10	2.05	2.01	1.96	1.93	1.91	1.89	1.87	1.86	1.83	1.80
2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.24	2.22	2.18	2.14
3.15	3.05	2.95	2.84	2.78	2.73	2.67	2.64	2.61	2.55	2.50
3.94	3.80	3.66	3.51	3.41	3.35	3.27	3.22	3.18	3.09	3.02
6.40	6.13	5.85	5.56	5.38	5.25	5.10	5.00	4.94	4.77	4.62
2.06	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.82	1.79	1.76
2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.18	2.16	2.11	2.07
3.06	2.96	2.86	2.76	2.69	2.64	2.59	2.55	2.52	2.46	2.40
3.80	3.67	3.52	3.37	3.28	3.21	3.13	3.08	3.05	2.96	2.88
6.08	5.81	5.54	5.25	5.07	4.95	4.80	4.70	4.64	4.47	4.33
2.03	1.99	1.94	1.89	1.86	1.84	1.81	1.79	1.78	1.75	1.72
2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.12	2.11	2.06	2.02
2.99	2.89	2.79	2.68	2.61	2.57	2.51	2.47	2.45	2.38	2.32
3.69	3.55	3.41	3.26	3.16	3.10	3.02	2.97	2.93	2.84	2.76
5.81	5.55	5.27	4.99	4.82	4.70	4.54	4.45	4.39	4.23	4.08
2.00	1.96	1.91	1.86	1.83	1.81	1.78	1.76	1.75	1.72	1.69
2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.08	2.06	2.01	1.97
2.92	2.82	2.72	2.62	2.55	2.50	2.44	2.41	2.38	2.32	2.26
3.59	3.46	3.31	3.16	3.07	3.00	2.92	2.87	2.83	2.75	2.66
5.58	5.32	5.05	4.78	4.60	4.48	4.33	4.24	4.18	4.02	3.87

(*Continued*)

TABLE E

F critical values (continued)

Degrees of freedom in the denominator		Degrees of freedom in the numerator								
	p	1	2	3	4	5	6	7	8	9
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
19	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
20	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
21	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11
22	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
23	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89
24	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80
25	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
26	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64
27	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57

TABLE E

F critical values (continued)

Degrees of freedom in the numerator										
10	12	15	20	25	30	40	50	60	120	1000
1.98	1.93	1.89	1.84	1.80	1.78	1.75	1.74	1.72	1.69	1.66
2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.04	2.02	1.97	1.92
2.87	2.77	2.67	2.56	2.49	2.44	2.38	2.35	2.32	2.26	2.20
3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.78	2.75	2.66	2.58
5.39	5.13	4.87	4.59	4.42	4.30	4.15	4.06	4.00	3.84	3.69
1.96	1.91	1.86	1.81	1.78	1.76	1.73	1.71	1.70	1.67	1.64
2.38	2.31	2.23	2.16	2.11	2.07	2.03	2.00	1.98	1.93	1.88
2.82	2.72	2.62	2.51	2.44	2.39	2.33	2.30	2.27	2.20	2.14
3.43	3.30	3.15	3.00	2.91	2.84	2.76	2.71	2.67	2.58	2.50
5.22	4.97	4.70	4.43	4.26	4.14	3.99	3.90	3.84	3.68	3.53
1.94	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.68	1.64	1.61
2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.97	1.95	1.90	1.85
2.77	2.68	2.57	2.46	2.40	2.35	2.29	2.25	2.22	2.16	2.09
3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.61	2.52	2.43
5.08	4.82	4.56	4.29	4.12	4.00	3.86	3.77	3.70	3.54	3.40
1.92	1.87	1.83	1.78	1.74	1.72	1.69	1.67	1.66	1.62	1.59
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.94	1.92	1.87	1.82
2.73	2.64	2.53	2.42	2.36	2.31	2.25	2.21	2.18	2.11	2.05
3.31	3.17	3.03	2.88	2.79	2.72	2.64	2.58	2.55	2.46	2.37
4.95	4.70	4.44	4.17	4.00	3.88	3.74	3.64	3.58	3.42	3.28
1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.60	1.57
2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.91	1.89	1.84	1.79
2.70	2.60	2.50	2.39	2.32	2.27	2.21	2.17	2.14	2.08	2.01
3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.53	2.50	2.40	2.32
4.83	4.58	4.33	4.06	3.89	3.78	3.63	3.54	3.48	3.32	3.17
1.89	1.84	1.80	1.74	1.71	1.69	1.66	1.64	1.62	1.59	1.55
2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.88	1.86	1.81	1.76
2.67	2.57	2.47	2.36	2.29	2.24	2.18	2.14	2.11	2.04	1.98
3.21	3.07	2.93	2.78	2.69	2.62	2.54	2.48	2.45	2.35	2.27
4.73	4.48	4.23	3.96	3.79	3.68	3.53	3.44	3.38	3.22	3.08
1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.57	1.54
2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.86	1.84	1.79	1.74
2.64	2.54	2.44	2.33	2.26	2.21	2.15	2.11	2.08	2.01	1.94
3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.40	2.31	2.22
4.64	4.39	4.14	3.87	3.71	3.59	3.45	3.36	3.29	3.14	2.99
1.87	1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.59	1.56	1.52
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.82	1.77	1.72
2.61	2.51	2.41	2.30	2.23	2.18	2.12	2.08	2.05	1.98	1.91
3.13	2.99	2.85	2.70	2.60	2.54	2.45	2.40	2.36	2.27	2.18
4.56	4.31	4.06	3.79	3.63	3.52	3.37	3.28	3.22	3.06	2.91
1.86	1.81	1.76	1.71	1.67	1.65	1.61	1.59	1.58	1.54	1.51
2.22	2.15	2.07	1.99	1.94	1.90	1.85	1.82	1.80	1.75	1.70
2.59	2.49	2.39	2.28	2.21	2.16	2.09	2.05	2.03	1.95	1.89
3.09	2.96	2.81	2.66	2.57	2.50	2.42	2.36	2.33	2.23	2.14
4.48	4.24	3.99	3.72	3.56	3.44	3.30	3.21	3.15	2.99	2.84
1.85	1.80	1.75	1.70	1.66	1.64	1.60	1.58	1.57	1.53	1.50
2.20	2.13	2.06	1.97	1.92	1.88	1.84	1.81	1.79	1.73	1.68
2.57	2.47	2.36	2.25	2.18	2.13	2.07	2.03	2.00	1.93	1.86
3.06	2.93	2.78	2.63	2.54	2.47	2.38	2.33	2.29	2.20	2.11
4.41	4.17	3.92	3.66	3.49	3.38	3.23	3.14	3.08	2.92	2.78

(*Continued*)

TABLE E

F critical values (continued)

Degrees of freedom in the denominator		Degrees of freedom in the numerator								
	p	1	2	3	4	5	6	7	8	9
28	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50
29	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45
30	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
40	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
	.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02
50	.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76
	.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
	.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38
	.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
	.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82
60	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
100	.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69
	.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
	.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24
	.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
	.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44
200	.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66
	.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
	.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18
	.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
	.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26
1000	.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64
	.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
	.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13
	.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
	.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13

TABLE E

F critical values (continued)

Degrees of freedom in the numerator										
10	12	15	20	25	30	40	50	60	120	1000
1.84	1.79	1.74	1.69	1.65	1.63	1.59	1.57	1.56	1.52	1.48
2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.71	1.66
2.55	2.45	2.34	2.23	2.16	2.11	2.05	2.01	1.98	1.91	1.84
3.03	2.90	2.75	2.60	2.51	2.44	2.35	2.30	2.26	2.17	2.08
4.35	4.11	3.86	3.60	3.43	3.32	3.18	3.09	3.02	2.86	2.72
1.83	1.78	1.73	1.68	1.64	1.62	1.58	1.56	1.55	1.51	1.47
2.18	2.10	2.03	1.94	1.89	1.85	1.81	1.77	1.75	1.70	1.65
2.53	2.43	2.32	2.21	2.14	2.09	2.03	1.99	1.96	1.89	1.82
3.00	2.87	2.73	2.57	2.48	2.41	2.33	2.27	2.23	2.14	2.05
4.29	4.05	3.80	3.54	3.38	3.27	3.12	3.03	2.97	2.81	2.66
1.82	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.54	1.50	1.46
2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.74	1.68	1.63
2.51	2.41	2.31	2.20	2.12	2.07	2.01	1.97	1.94	1.87	1.80
2.98	2.84	2.70	2.55	2.45	2.39	2.30	2.25	2.21	2.11	2.02
4.24	4.00	3.75	3.49	3.33	3.22	3.07	2.98	2.92	2.76	2.61
1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.42	1.38
2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.66	1.64	1.58	1.52
2.39	2.29	2.18	2.07	1.99	1.94	1.88	1.83	1.80	1.72	1.65
2.80	2.66	2.52	2.37	2.27	2.20	2.11	2.06	2.02	1.92	1.82
3.87	3.64	3.40	3.14	2.98	2.87	2.73	2.64	2.57	2.41	2.25
1.73	1.68	1.63	1.57	1.53	1.50	1.46	1.44	1.42	1.38	1.33
2.03	1.95	1.87	1.78	1.73	1.69	1.63	1.60	1.58	1.51	1.45
2.32	2.22	2.11	1.99	1.92	1.87	1.80	1.75	1.72	1.64	1.56
2.70	2.56	2.42	2.27	2.17	2.10	2.01	1.95	1.91	1.80	1.70
3.67	3.44	3.20	2.95	2.79	2.68	2.53	2.44	2.38	2.21	2.05
1.71	1.66	1.60	1.54	1.50	1.48	1.44	1.41	1.40	1.35	1.30
1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.56	1.53	1.47	1.40
2.27	2.17	2.06	1.94	1.87	1.82	1.74	1.70	1.67	1.58	1.49
2.63	2.50	2.35	2.20	2.10	2.03	1.94	1.88	1.84	1.73	1.62
3.54	3.32	3.08	2.83	2.67	2.55	2.41	2.32	2.25	2.08	1.92
1.66	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.34	1.28	1.22
1.93	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.45	1.38	1.30
2.18	2.08	1.97	1.85	1.77	1.71	1.64	1.59	1.56	1.46	1.36
2.50	2.37	2.22	2.07	1.97	1.89	1.80	1.74	1.69	1.57	1.45
3.30	3.07	2.84	2.59	2.43	2.32	2.17	2.08	2.01	1.83	1.64
1.63	1.58	1.52	1.46	1.41	1.38	1.34	1.31	1.29	1.23	1.16
1.88	1.80	1.72	1.62	1.56	1.52	1.46	1.41	1.39	1.30	1.21
2.11	2.01	1.90	1.78	1.70	1.64	1.56	1.51	1.47	1.37	1.25
2.41	2.27	2.13	1.97	1.87	1.79	1.69	1.63	1.58	1.45	1.30
3.12	2.90	2.67	2.42	2.26	2.15	2.00	1.90	1.83	1.64	1.43
1.61	1.55	1.49	1.43	1.38	1.35	1.30	1.27	1.25	1.18	1.08
1.84	1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.33	1.24	1.11
2.06	1.96	1.85	1.72	1.64	1.58	1.50	1.45	1.41	1.29	1.13
2.34	2.20	2.06	1.90	1.79	1.72	1.61	1.54	1.50	1.35	1.16
2.99	2.77	2.54	2.30	2.14	2.02	1.87	1.77	1.69	1.49	1.22

Appendix H: Q Distribution Table

How to Use the Q Distribution Table

This table should be used only if the sample sizes in your Tukey's HSD analysis are equal. There are two sections of the table, one for the .05 significance level (H.1) and one for the .01 significance level (H.2). Select a significance level and the corresponding section of the table. Then, find the number of groups (samples) being compared in the top row of the table. This determines the column of the table you will use. Find the *df* for your data ($df = n - k$, where k is the number of groups [samples] and n is the number of observations in one of the samples) in the left-hand column of the table. This identifies the row of the table you will use. The critical value of Q for the HSD test is found at the intersection of the row and column you have identified. A difference between sample means as large or larger than the HSD you calculate using the table value of Q is significant at the selected level of significance.

Table H.1 Critical Values of Q (p = 0.05)

(*p*) = 0.05									
k / *df*	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	18.0	27.0	32.8	37.1	40.4	43.1	45.4	47.4	49.1
2	6.08	8.33	9.80	10.88	11.73	12.43	13.03	13.54	13.99
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56
□	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47

(Continued)

Table H.2 Critical Values of Q (p = 0.01)

$(p) = 0.01$									
df \ *k*	2	3	4	5	6	7	8	9	10
1	90.0	135	164	186	202	216	227	237	246
2	13.90	19.02	22.56	25.37	27.76	29.86	31.73	33.41	34.93
3	8.26	10.62	12.17	13.32	14.24	15.00	15.65	16.21	16.71
4	6.51	8.12	9.17	9.96	10.58	11.10	11.54	11.92	12.26
5	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10
7	4.95	5.92	6.54	7.00	7.37	7.68	7.94	8.17	8.37
8	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86
9	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21
11	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99
12	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54
15	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44
16	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09
24	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92
30	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76
40	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60
60	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45
120	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30
□	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16

Source: www.stat.wisc.edu/courses/st571-ane/tables/tableQ.pdf.