Süleyman Demirel University
Faculty of Economics

H.N.Aliyev

STATISTICS

Solved problems and exercises
Part 2

Almaty – 2009

Süleyman Demirel University
Faculty of Economics

H.N.Aliyev

STATISTICS

Solved problems and exercises
Part 2

Almaty – 2009

ББК 60.6
А 36

Рекомендовано к печати Ученым Советом Университета имени Сулеймана Демиреля.

2-nd edition
Aliyev H.N..
A 36 Statistic. Solved problems and exercises. Part 2. – Алматы, 2009 - 245с.

ISBN № 9965-9605-7-7

ББК 60.6

А 0702000000
00 (05)-05

ISBN № 9965-9605-7-7

CONTENTS

Chapter 1
Hypothesis testing

1.1. Introduction

Inferential statistics consists of methods that use sample results to help make decisions or predictions about a population. The point and interval estimation procedures are forms of statistical inference. Another type of statistical inference is hypothesis testing. In hypothesis testing we begin by stating a hypothesis about a population characteristic. This hypothesis, called the **null hypothesis**, is assumed to be true unless sufficient evidence can be found in a sample to reject it. The situation is quite similar to that in a criminal trial. The defendant is assumed to be innocent; if sufficient evidence to the contrary is presented, however, the jury will reject this hypothesis and conclude that the defendant is guilty.

In statistical hypothesis testing, often the null hypothesis is an assumption about the value of a population parameter. A sample is selected from the population, and a point estimate is computed. By comparing the value of the point estimate to the hypothesized value of the parameter we draw a conclusion with respect to whether or not there is a sufficient evidence to reject the null hypothesis. A decision is made and often a specific action is taken depending upon whether or not the null hypothesis about the population parameter is accepted or rejected.

1.1.2. Concepts of hypothesis testing

Let us consider example about coffee cans. A company may claim that, on average, its cans contain 100 grams of coffee. A government agency may want to test whether or not such cans contain, on average, 100 grams of coffee.

Suppose we take a sample of 50 cans of the coffee under investigation. We then find out that the mean amount of coffee in these 50 cans is 97 grams. Based on these results, can we state that on average, all such cans contain less than 100 grams of coffee and that the company is lying to the public?

Not until we perform a test of hypothesis. The reason is that the mean $\bar{x} = 97$ grams is obtained from the sample. The difference between 100 grams (the required amount for the population) and 97 grams (the observed

average amount for the sample) may have occurred only because of the sampling error. Another sample of 100 cans may give us a mean of 105 grams. Therefore, we make a test of hypothesis to find out how large the difference between 100 grams and 97 grams is and to investigate whether or not this difference has occurred as a result of chance alone. If 97 grams is the mean of all cans and not for only 100 cans, then we do not need to make a test of hypothesis. Instead, we can immediately state that the mean amount of coffee in all such cans is less than 100 grams. We perform a test of hypothesis only when we are making a decision about a population parameter based on the value of a sample statistic.

1.1.3. The null and alternative hypothesis

We will begin our general discussion by using θ to denote a population probability distribution parameter of interest, such as the mean, variance, or proportion. Our discussion begins with a hypothesis about the parameter that will be maintained unless there is strong contrary evidence. In statistical language it is called the **null hypothesis**.

For example, we might initially accept company's claim that on average, the contests of the cans weight at least 100 grams. Then after collecting sample data this hypothesis can be tested. If the null hypothesis is not true, then some alternative must be true. In carrying out a hypothesis test the investigator defines an **alternative hypothesis** against which the null hypothesis is tested.

For this coffee cans example a likely alternative is that on average can's weights are less than 100 grams. These hypotheses are chosen such that one or the other must be true. The null hypothesis will be denoted as H_0 and the alternative hypothesis as H_1.

Definition: A null hypothesis is a claim (or statement) about a population parameter that is assumed to be true until it is declared false.

Definition: An alternative hypothesis is a claim about population parameter that will be true if the null hypothesis is false.

Our analysis will be designed with the objective of seeking strong evidence to reject the null hypothesis and accept the alternative hypothesis. We will only reject the null hypothesis when there is a small probability that the null hypothesis is true. Thus rejection will provide strong evidence against H_0 and in favor of the alternative hypothesis, H_1. If we fail to reject H_0 then either H_0 is true or our evidence is not sufficient to reject H_0 and hence accept H_0. Thus we will be more comfortable with our decision if we reject H_0 and accept H_1.

A hypothesis, whether null or alternative, might specify a single value, say θ_0, for the population parameter θ. In that case, the hypothesis is said to be a simple hypothesis designated as

$$H_0 : \theta = \theta_0$$

That is read as, "The null hypothesis is that the population parameter θ is equal to the specific value θ_0".

Alternatively, a range of values might be specified for unknown parameter. We define such hypothesis as a composite hypothesis, and it will hold true for more than one value of the population parameter. In many applications, a simple null hypothesis, say

$$H_0 : \theta = \theta_0$$

is tested against a composite alternative. One possibility would be to test the null hypothesis against the general two-sided hypothesis

$$H_1 : \theta \neq \theta_0$$

In other cases, only alternatives on one side of the null hypothesis are of interest. For example, a government agency would be perfectly happy if the mean weight of coffee cans greater than 100 grams. Then we could write the null hypothesis as

$$H_0 : \theta \geq \theta_0$$

and the alternative hypothesis of interest might be

$$H_1 : \theta < \theta_0$$

We call these hypothesis one- sided composite alternatives.

Example:

A company intends to accept the product unless it has evidence to suspect that more than 10% of products are defective. Let θ denote the population proportion of defectives. The null hypothesis is that the proportion is less than 0.1, that is

$$H_0 : \theta \leq 0.1$$

and the alternative hypothesis is

$$H_1 : \theta > 0.1$$

The null hypothesis is that the product is of adequate quality overall, while the alternative is that the product is not adequate quality. In this case the product would only be rejected if there is strong evidence that there are more than 10% defectives.

Once we have specified a null hypothesis and alternative hypothesis and collected sample data, a decision concerning the null hypothesis must be made. We can either accept the null hypothesis or reject it in favor of the alternative. For good reasons many statisticians prefer not to use the term "accept the null hypothesis" and instead say "fail to reject". When we accept or fail to reject the null hypothesis, then either the hypothesis is true or our test procedure was not strong enough to reject and we have committed an error. When we use the term **accept a null hypothesis** that statement can be considered shorthand for failure to reject.

From our discussion of sampling distributions, we know that the sample mean is different from the population mean. With only a sample mean we can not be certain of the value of the population mean. Thus the decision rule we adopt will have some chance of reaching an erroneous conclusion. One error we call Type I error. **Type I** error is defined as the rejection of the null hypothesis when the null hypothesis is true. We will see that our decision rules will be defined so that the probability of rejecting a true null hypothesis, denoted as α, is "small". The probability, α, is defined as the **significance level** of the test. Since the null hypothesis is either accepted or rejected, it follows that the probability of accepting the null hypothesis when it is true is $(1-\alpha)$. The other possible error, called **Type II error**, arises when false null hypothesis is accepted. We say that for a particular decision rule, the probability of making such an error when the null hypothesis is false is denoted β. Then, the probability of rejecting a false null hypothesis is $(1-\beta)$ which is called the power of test.

Type I error

A type I error occurs when a true null hypothesis is rejected. The value α represents the probability of committing this type of error, that is

$$\alpha = P(H_0 \text{ is rejected} / H_0 \text{ is true})$$

The value α represents the significance level of the test.

Type II error

A Type II error occurs when a false null hypothesis is not rejected. The value β represents the probability of committing a Type II error, that is

$$\beta = P(H_0 \text{ is not rejected} / H_0 \text{ is false})$$

The value $(1-\beta)$ is called the power of the test. It represents the probability of not making a Type II error.

1.1.4. Tails of the test

In statistics, the rejection region for a hypothesis testing problem can be on both sides with non rejection region in the middle, or it can be on the left side or in the right side of the non rejection region. A test with two rejection regions is called a **two tailed test**, and a test with one rejection region is called a **one tailed test**. The one tailed test is called a **left tailed test** if the rejection region is in the left tail of the distribution curve, and a **right tailed test** if the rejection region is in the right tail of the distribution curve.

a) A two tailed test

Example:

The mean family size in a particular country was 3.75 in 1990. We want to check whether or not this mean has changed since 1990. The mean family size has changed if it has either increased or decreased during this period. This is an example of two tailed test. Let μ be the current mean family size for all families. We write the null and alternative hypothesis for this test as

$H_0 : \mu = 3.75$ (The mean family size has not changed)

$H_1 : \mu \neq 3.75$ (The mean family size has changed)

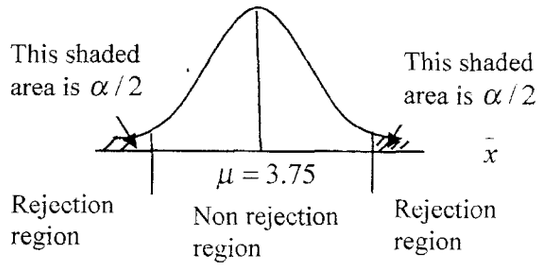

Fig.1.1

As shown in Figure 1.1, a two tailed test has two rejection regions, one in each tail of the distribution curve.

b) A left tailed test

Reconsider the example of the mean amount of coffee can produced by company. The company claims that these cans, on average, contain 100 grams of coffee. However, if these cans contain less than the claimed amount of coffee, then the company can be accused of cheating. Suppose that the government agency wants to test whether the amount of coffee can is less than 100 grams. Note that the key phrase this time is *less than*, which indicates a left tailed test. Let μ be the mean amount of coffee in all cans. The null and alternative hypothesis for this test are written as

$H_0 : \mu = 120$ grams (The mean is not less than 120 grams)

$H_1 : \mu < 120$ grams (The mean is less than 120 grams)

In this case, we can also write the null hypothesis as $H_0 : \mu \geq 120$ grams. This will not affect the result of the test as long as the sign in H_1 is *less than*. When the alternative hypothesis has a *less than* (<) sign, as in this case, the test is always left tailed. In a left tailed test the rejection region is always in the left tail of the distribution curve, as shown in Figure 1.2.

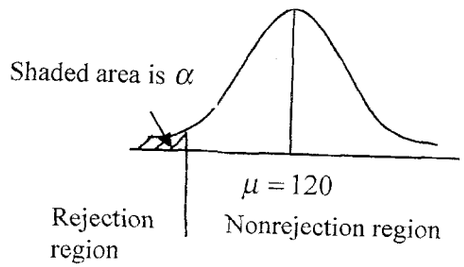

Fig.1.2

c) A right tailed test

Suppose that mean monthly income of all households was 45 500 tg in 2001. We want to test if current income of all households is higher than 45 500 tg. The key phrase in this case is *higher than,* which indicates a right tailed test. Let μ be the mean income of all households.

We write the null and alternative hypothesis for this test as

$H_0 : \mu = 45500$ (The current income is not higher than 45 500 tg)

$H_1 : \mu > 45500$ (The current income is higher than 45 500 tg)

In this case, we can also write the null hypothesis as $H_0 : \mu \leq 45500$, which states that current mean income is either equal to or less than 45 500 tg. Again, the result of the test will not be affected whether we use an *equal to* (=) or a *less or equal to* ($\leq$) sign in H_0 as long as the alternative hypothesis has a *greater than* (>) sign.

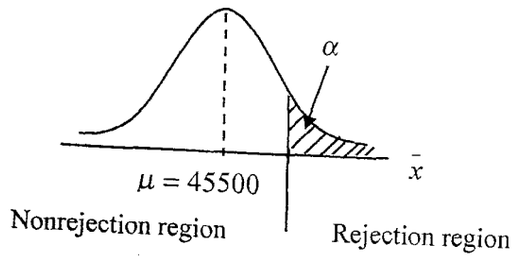

Fig.1.3

When an alternative hypothesis has a *greater than* (>) sign, the test is alway right tailed. As shown in the Fig. 1.3, in a right tailed test, the rejectio region is in the right tail of the distribution curve. The area of this rejectio region is equal to α, the significance level. We will reject H_0 if the value o

$\bar{x}$ obtained from the sample falls in the rejection region. Otherwise, we wil not reject H_0.

Remark: Note that the null hypothesis always has an *equal to* (=) or a *les or equal to* ($\leq$) or a *greater than or equal to* ($\geq$) sign and the alternativ hypothesis always has a *not equal to* ($\neq$) or a *greater than* (>) or a *less tha* (<) sign.

Exercises

1. Explain which of the following is a two tailed test, a left tailed test, or right tailed test.

a) $H_0 : \mu = 25, \qquad H_1 : \mu < 25$

b) $H_0 : \mu \leq 134, \qquad H_1 : \mu > 134$

c) $H_0 : \mu = 16, \qquad H_1 : \mu \neq 16$

Show the rejection and nonrejection regions for each of these cases b drawing a sampling distribution curve for the sample mean, assuming th sample size is large in each case.

2. Consider $H_0 : \mu = 35$, against $H_1 : \mu < 35$.

a) What type of error would you make if the null hypothesis is actually fals and you fail to reject it?

b) What type of error would you make if the null hypothesis is actually tru and you reject it?

3. For each of the following rejection regions, sketch the samplin distribution for z and indicate the location of rejection region.

a) $z > 2.05$; b) $z > 2.75$; c) $z < -1.28$;

d) $z < -2.13$; f) $z < -2.575$ or $z > 2.575$;

g) $z < -1.82$ or $z > 1.82$

4. Write the null hypothesis and alternative hypothesis for each of the following examples. Determine if each is a case of a two tailed, a left tailed, or a right tailed test.

a) To test whether or not the mean price of houses in a certain city is greater than $ 45 000.

b) To test if the mean number of hours spent working per week by students who hold jobs is different from 18 hours.

c) To test whether the mean life of a particular brand of auto batteries is less than 28 days.

d) To test if the mean amount of time taken by all workers to do a certain job is more than 45 minutes.

f) To test the mean age of all managers of companies is different from 40 years.

g) To test the mean time for an airline passenger to obtain his or her luggage, once luggage starts coming out the conveyer belt, is less than 180 seconds.

1.2. Tests of the mean of a normal distribution: Population variance known

In this and following sections we will present specific procedures for developing and implementing hypothesis test procedures with applications to business and economic problems.

We are given a random sample of n observations from a normal population with mean μ and known variance σ^2. If the observed sample mean is $\bar{x}$, then the test statistic is

$$T.S. = z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

and we can use the following tests with significance level α.

1. To test either null hypothesis

$H_0 : \mu = \mu_0$ or $H_0 : \mu \leq \mu_0$ against the alternative

$H_1 : \mu > \mu_0$

the decision rule is

Reject H_0 if $T.S. > z_\alpha$

2. To test either null hypothesis

$H_0 : \mu = \mu_0$ or $H_0 : \mu \geq \mu_0$ against the alternative

$H_1 : \mu < \mu_0$

the decision rule is

Reject H_0 if $T.S. < -z_\alpha$

3. To test the null hypothesis

$H_0 : \mu = \mu_0$ against the two sided alternative

$H_1 : \mu \neq \mu_0$

the decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$,

where $z_{\alpha/2}$ is the number for which

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

and Z is the standard normal distribution.

A statistical test of hypothesis procedure contains the following five steps:

1. State the null and alternative hypothesis
2. Select the distribution to use
3. Determine the rejection and nonrejection regions
4. Calculate the value of the test statistic
5. Make a decision.

Example:

A manufacturer of detergent claims that the content of boxes sold weigh average at least 160 grams. The distribution of weights is known to normal, with standard deviation of 14 grams. A random sample of 16 box yielded a sample mean weight of 158.9 grams. Test at the 10% significan level the null hypothesis that the population mean is at least 160 grams.

Solution:

Let μ be the mean average of all boxes and $\bar{x}$ be the corresponding me for the sample.

$n = 16$; $\sigma = 14$; $\bar{x} = 158.9$

The significance level is α is 0.1. That is, the probability of rejecting t null hypothesis when it is actually is true should not exceed 0.1. This is t probability of making a Type I error. We perform the test of hypothe using the five steps as follows.

Step 1. *State the null and alternative hypothesis*

We write the null and alternative hypothesis as

$H_0 : \mu \geq 160$ grams

$H_1 : \mu < 160$ grams

Step 2. *Select the distribution to use*

Since population standard deviation is known we will use $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$.

Step 3. *Determine the rejection and nonrejection regions*

The significance level is 0.1. The < sign indicates that the test is left tailed. We look for 0.9 from in the standard normal distribution table, (Table 1 of Appendix). The value of z is -1.28. (Fig. 1.4).

Step 4. *Calculate the value of the test statistic*

The decision to reject or not to reject the null hypothesis will depend on whether the evidence from the sample falls in the rejection or nonrejection region. If the value of the sample mean $\bar{x}$ falls in rejection region, we reject H_0. Otherwise we do not reject the null hypothesis. To locate the position of $\bar{x} = 158.9$ on the sampling distribution curve of $\bar{x}$ in Figure 1.4 we first calculate z value for $\bar{x} = 158.9$. This is called the *value of the test statistic.*

$$T.S. = z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{158.9 - 160}{14/\sqrt{16}} = -0.31$$

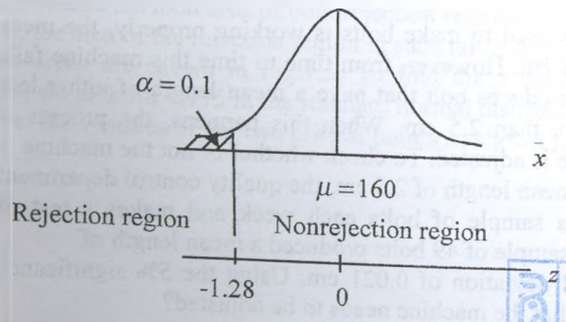

Fig 1.4

Step 5. *Make a decision*
In the final step we make a decision based on the value of the test statistic

$T.S. = z$ for $\bar{x}$ in previous step. This value of $z = -0.31$ is not less than the critical value of $z = -1.28$, and it falls in the nonrejection region. Hence we accept H_0 and conclude that based on sample information, it appears that the mean weight of all boxes is greater than 160 grams.
By accepting the null hypothesis we are stating that the difference between the sample mean $\bar{x} = 158.9$ and the hypothesized value of the population mean $\mu = 160$ is not too large and may occurred because of the chance or sampling error. There is a possibility that the mean weight is less than 160 grams, by the luck of the draw, we selected a sample with a mean that is not too far from required mean of 160 grams.

1.3. Tests of the mean of a normal distribution: Population variance unknown (Large sample size)

When the population standard deviation is unknown, we simply estimate σ with the value of the sample standard deviation s. We must consider separately the large sample ($n \geq 30$) and small sample size ($n < 30$) cases.
If the sample size n is large, the test procedure developed for the case when population variance is known can be employed when it is unknown, replacing σ^2 by the observed sample variance s. All the hypotheses and decision rules are stated in the same way as before (i.e. when σ^2 is known).

Example:
When a machine that is used to make bolts is working properly, the mean length of these bolts 2.5 cm. However, from time to time this machine falls out of alignment and produces bolt that have a mean length of either less than or 2.5 cm or more than 2.5 cm. When this happens, the process is stopped and the machine is adjusted. To check whether or not the machine is producing bolts with a mean length of 2.5 cm, the quality control department at the company takes a sample of bolts each week and makes a test of hypothesis. One such a sample of 49 bolts produced a mean length of 2.49 cm and a standard deviation of 0.021 cm. Using the 5% significance level, can we conclude that the machine needs to be adjusted?

Solution:

Let μ be the mean length bolts made on this machine and x be the corresponding mean for the sample.

$n = 49$; $\bar{x} = 2.49$ cm; $s = 0.021$ cm

The mean length of all bolts is supposed to be 2.5 cm. The significance level is α is 0.05. That is, the probability of rejecting the null hypothesis when it is actually is true should not exceed 0.05.

Step 1. *State the null and alternative hypothesis*
We are testing to find whether or not the machine needs to be adjusted. The machine will need an adjustment if the mean length of these bolts is either less than 2.5 cm or more than 2.5.
We write the null and alternative hypothesis as

$H_0 : \mu = 2.5$ cm (The machine does not need adjustment)

$H_1 : \mu \neq 2.5$ cm (The machine needs an adjustment)

Step 2. *Select the distribution to use*

Because the sample size is large ($n > 30$), the sampling distribution of $\bar{x}$ is approximately) normal. Consequently we will use $z = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ to make the test.

Step 3. *Determine the rejection and nonrejection regions*
The significance level is 0.05. The $\neq$ sign indicates that the test is two tailed with two rejection regions, one in each tail of the normal distribution curve of $\bar{x}$. Because the total area of both rejection regions is 0.05 (the significance level), the area of the rejection region in each tail is 0.025.
These areas are shown in Fig.1.5. To find the z values for these critical points, we look for 0.975 in the standard normal distribution table.
Hence, the z values of the two critical points as shown in Fig.1.5, are -1.96 and 1.96.

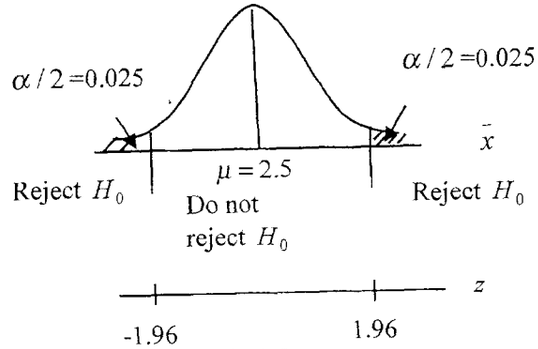

Fig.1.5

Step 4. *Calculate the value of the test statistic*

The value of $\bar{x}$ from the sample is 2.49. As σ is not known, we calculate z value as follows

$$T.S. = z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{2.49 - 2.5}{0.003} = -3.33$$

$z = -3.33$ is the value of the test statistic.

Step 5. *Make a decision*

The value of $z = -3.33$ is less than the critical value of $z = -1.96$, and it f
in the rejection region in the left tail. Hence we reject H_0 and conclude t
based on sample information, it appears that the mean length of all b
produced on this machine is not equal to 2.5 cm. Therefore, the mach
needs to be adjusted

By rejecting the null hypothesis we are stating that the difference betwi
the sample mean $\bar{x} = 2.49$ and the hypothesized value of the popula
mean $\mu = 2.5$ is too large and may not have occurred because of the cha
or sampling error. This difference seems to be real and, hence the ma
length of bolts is different from 2.5 cm. Note that the rejection of the
hypothesis does not necessarily indicate that the mean length of bolt
definitely different from 2.5 cm. It simply indicates that there is
evidence (from sample) that the mean length of bolts is not equal to 2.5 c

here is a possibility that the mean length of bolts equal to 2.5 cm. If so, we
ıve wrongfully rejected the null hypothesis H_0. This is Type I error and
obability of making such an error in this case is 0.05.

Exercises

Make the following tests of hypotheses.

$H_0 : \mu = 25$; $H_1 : \mu \neq 25$; $n = 81$; $\bar{x} = 28$; $s = 3$; $\alpha = 0.01$

$H_0 : \mu = 12$; $H_1 : \mu < 12$; $n = 45$; $\bar{x} = 11$; $\sigma = 4.5$; $\alpha = 0.05$

$H_0 : \mu = 40$; $H_1 : \mu > 40$; $n = 100$; $\bar{x} = 46$; $s = 7$; $\alpha = 0.1$

Consider $H_0 : \mu = 100$; against the two sided alternative $H_1 : \mu \neq 100$.

A random sample of 64 observations produced a sample mean of 98 and a
ndard deviation of 12. Using $\alpha = 0.01$, would you reject the null
pothesis?

Another random sample of 64 observations taken from the same
pulation produced a sample mean of 104 and a standard deviation of 10.
ing $\alpha = 0.01$, would you reject the null hypothesis?
mment on the results of parts a) and b).

A survey showed that people with a bachelor's degree earned average of
116 a year in 2001. A sample of 900 persons with a bachelor's degree
en recently by a researcher showed that the persons in this sample earned
average of $2345 a year with a standard deviation of $210. Test at 5%
nificance level whether people with a bachelor's degree currently earn an
rage of $2116 against the alternative that it is more than $2116 in a year.

The manufacturer of a certain brand of auto batteries claims that the
an life of these batteries is 45 month. A consumer protection agency that
nts to check this claim took a random sample of 36 such batteries and
nd the mean life for this sample is 43.75 month with a standard deviation
4 month. Using the 2.5% significance level, test the manufacturer claim
linst the alternative that the mean life of batteries is less than 45 month.

A random sample of 100 observations from a population with standard
viation 60 yielded a sample mean of 110.
Test the null hypothesis that $\mu = 100$ against the alternative hypothesis
t $\mu > 100$ using $\alpha = 0.05$. Interpret the results of the test.

b) Test the null hypothesis that $\mu = 100$ against the alternative hypothesi that $\mu \neq 100$ using $\alpha = 0.05$. Interpret the results of the test.

c) Compare the results of the two tests you conducted. Explain why th results differ.

6. In a random sample of 250 observations, the mean and standard deviatio are found to be 169.8 and 31.6, respectively. Is the claim that μ larger tha 169 substantiated by these data at the 10% level of significance?

7. From records, it is known that the duration of treating a disease by standard therapy has a mean of 15 days. It is claimed that a new therapy ca reduce the treatment time. To test this claim, the new therapy is tried on patients, and from the data of their times to recovery, the sample mean an standard deviation are found to be 14.6 and 3.0 days, respectively. Perform the hypothesis test using a 2.5% level of significance.

8. Suppose that you are to verify the claim that $\mu > 20$ on the basis of random sample of size 70, and you know that $\sigma = 5.6$.

a) If you set the rejection region to be $\bar{x} > 21.31$, what is the level significance of your test?

b) Find the numerical value of c so that the test $\bar{x} \geq c$ has a 5% level significance.

Answers

1. a) $T.S. = 9.00$; reject H_0; b) $T.S. = -1.49$; do not reject H_0; c) $T.S. = 8.5$ reject H_0; **2.** a) $T.S. = -1.33$; do not reject H_0; b) $T.S. = 3.20$; reject H_0; **3.** $T.S. = 32.71$; reject H_0; **4.** $T.S. = -1.87$; accept H_0; **5.** a) $z = 1.6$ reject H_0; b) $z = 1.67$; accept H_0; **6.** $T.S. = 0.4$; accept H_0; **7.** $z = -1.1$ H_0 is not rejected at $\alpha = 0.025$; **8.** a) $\alpha = 0.025$; b) c=21.10.

1.4. Hypothesis testing using the p –value approaches

In previous section, the value of the significance level α was elected before the test performed. Sometimes we may prefer not to redetermine α. Instead, we may want to find a value such that a given null ypothesis will be rejected for any α greater than this value and it will not be jected for any α smaller than this value. In this approach, we calculate the -value for the test, which is defined as the smallest level of significance at hich the given null hypothesis is rejected.

efinition:

he p-value is the smallest significance level at which the null hypothesis is jected.

sing the p-value approach, we reject the null hypothesis if

$$p\text{- value} < \alpha$$

nd we do not reject the null hypothesis if

$$p\text{- value} \geq \alpha$$

teps necessary for calculating the p-value for a test of hypothesis

Determine the value of the test statistic $T.S. = z$ corresponding to the result f the sampling experiment.

) If the test is one- tailed, the p-value is equal to the tail area beyond z in the ame direction as the alternative hypothesis. Thus, if the alternative ypothesis is of the form >, the p- value is the area to the right of, or above, ie observed z value. Conversely, if the alternative is of the form <, the - value is the area to the left of, or below, the observed z value. (Fig.1.6;1.7)

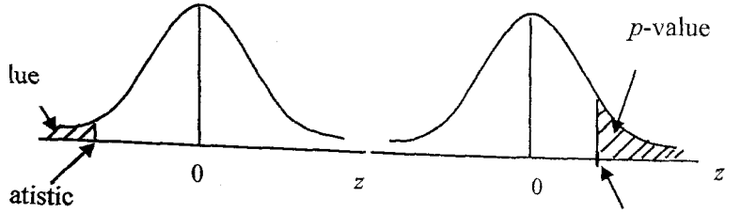

Fig.1.6. Left tailed test

Fig.1.7. Right tailed test

b) If the test is two tailed, the *p*-value is equal to twice the area beyond the observed *z*-value in the direction of the sign of *z*. That is, if *z* is positive, the *p*-value is twice the area to the right of, or above, the observed *z*- value. Conversely, if *z* is negative, the *p*-value is twice the area to the left of, or below, the observed *z*-value. (See Fig.1.8)

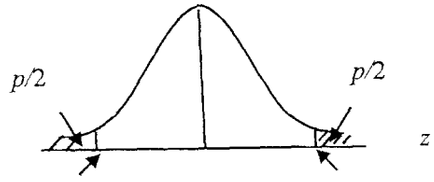

Fig.1.8. Finding the *p*-value for a two tailed test

Example:

The management of Health club claims that its members lose an average of 10kg or more within the first month after joining the club. A random sample of 36 members of this health club was taken and found that they lost an average of 9.2 kg within the first month of membership with standard deviation of 2.4kg. Find the *p*- value for this test.

Solution:

Let μ be the mean weight lost during the first month of membership by all members and $\bar{x}$ be corresponding mean for the sample.

Step 1. *State the null and alternative hypothesis*

$H_0 : \mu \geq 10$ (The mean weight lost is 10kg or more)

$H_1 : \mu < 10$ (The mean weight lost is less than 10kg)

Step 2. *Select the distribution to use*

Because the sample size is large we use the normal distribution to make the test and calculate *p*-value.

Step 3. Calculate the *p*-value.

The < sign in the alternative hypothesis indicates that test is left tailed. The *p*- value is given by the area in the left tail of the sampling distribution of $\bar{x}$ where $\bar{x}$ is less than 9.2. To find this area, we first find the *z* value for $\bar{x} = 9.2$ as follows

$$T.S. = z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{9.2 - 10}{2.4/\sqrt{36}} = -2.00$$

The area to the left of $\bar{x} = 9.2$ under the sampling distribution of $\bar{x}$ is equal to the area under the standard normal curve to the left of $z = -2.00$. The area to the left of $z = -2.00$ is 0.0228. Consequently,

$$p - \text{value} = 0.0228$$

Thus, based on the *p*- value of 0.0228 we can state that for any α (significance level) greater than 0.0228 we will reject the null hypothesis and for any α less than 0.0228 we will accept the null hypothesis.

Suppose we make the test for this example at $\alpha = 0.01$. Because $\alpha = 0.01$ is less than *p*-value of 0.0228, we will not reject the null hypothesis. Now suppose we make the test at $\alpha = 0.05$. Because $\alpha = 0.05$ is greater than the *p*-value of 0.0228, we will reject the null hypothesis.

Exercises

1. Find the *p*-value for each of the following hypothesis tests

a) $H_0 : \mu = 18$; $H_1 : \mu \neq 18$; $n = 50$; $\bar{x} = 20$; $s = 5$;

b) $H_0 : \mu = 15$; $H_1 : \mu < 15$; $n = 80$; $\bar{x} = 13.2$; $s = 5.5$;

c) $H_0 : \mu = 38$; $H_1 : \mu > 38$; $n = 35$; $\bar{x} = 40.6$; $s = 7.2$

2. Consider $H_0 : \mu = 29$; against the alternative $H_1 : \mu \neq 29$.

A random sample of 60 observations taken from this population produced a sample mean of 31.4 and a standard deviation of 8.

a) Calculate the *p*-value.

b) Considering the *p*-value of part a), would you reject the null hypothesis if the test were made at the significance level of 0.05?

c) Considering the *p*-value of part a), would you reject the null hypothesis if the test were made at the significance level of 0.01?

3. In a given situation, suppose H_0 was rejected at $\alpha = 0.05$. Answer th following questions as "yes", "no", or "can't tell" as the case may be.

a) Would H_0 also be rejected at $\alpha = 0.02$?

b) Would H_0 also be rejected at $\alpha = 0.10$?

c) Is the p-value smaller than 0.05?

4. In a problem of testing $H_0 : \mu = 75$ against $H_1 : \mu > 75$, the followin sample quantities are recorded.

$$n = 56; \qquad \bar{x} = 77.04; \qquad s = 6.80$$

a) State the test statistic and find the rejection region with $\alpha = 0.05$.

b) Calculate the test statistic and draw a conclusion with $\alpha = 0.05$.

c) Find the p-value and interpret the results.

Answers

1. a) 0.0046; b) 0.0017 ;c) 0.0162; **2.** a) 0.0204; b) yes, reject H_0; c) no, d not reject H_0 ;**3.** a) can't tell; b) yes; c) no; **4.** a) $T.S. = Z = \dfrac{\bar{x} - 75}{s/\sqrt{n}}$ $Z \geq 1.645$; b) $T.S. = 2.24$, H_0 is rejected at $\alpha = 0.05$; c) 0.0125;

1.5. Tests of the mean of a normal distribution: Population variance unknown. Small samples

Many times the size of a sample that is used to make test of hypothesis abo μ is small, that is, $n < 30$. If the population is (approximately) normall distributed, the population standard deviation σ is not known and the samp size is small ($n < 30$), then the normal distribution is replaced by t Student's t distribution to make a test of hypothesis about μ. In such a ca the random variable

$$t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has a Student's t distribution with $(n-1)$ degrees of freedom.

he value of test statistic t for the sample mean $\bar{x}$ is computed as

$$T.S. = t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

ıd we can use the following tests with significance level α.

To test either null hypothesis

$H_0 : \mu = \mu_0$ or $H_0 : \mu \leq \mu_0$ against the alternative

$H_1 : \mu > \mu_0$

e decision rule is

Reject H_0 if $T.S. > t_{n-1,\alpha}$

To test either null hypothesis

$H_0 : \mu = \mu_0$ or $H_0 : \mu \geq \mu_0$ against the alternative

$H_1 : \mu < \mu_0$

e decision rule is

Reject H_0 if $T.S. < -t_{n-1,\alpha}$

To test the null hypothesis

$H_0 : \mu = \mu_0$ against the two sided alternative

$H_1 : \mu \neq \mu_0$

e decision rule is

ject H_0 if $T.S. > t_{n-1,\alpha/2}$ or $T.S. < -t_{n-1,\alpha/2}$,

ere, $t_{n-1,\alpha}$ is the number for which

$$P(t_{n-1} > t_{n-1,\alpha}) = \alpha$$

ıere the random variable t_{n-1} follows a Student's t distribution with $-1)$ degrees of freedom.

xample:

ıe company that produces auto batteries claims that its batteries are good, r an average, for at least 64 days. A consumer protection agency tested 15 ch batteries to check this claim. It found the mean life of these 15 batteries be 62 days with a standard deviation of 3 days. At the 5% significance vel, can you conclude that the claim of the company is true? Assume that e life of such a battery has and approximate normal distribution.

Solution:

Let μ be the mean life of all batteries and $\bar{x}$ be the corresponding mean p the sample. Then from the given information,

$n = 15$; $\bar{x} = 62$ days; $s = 3$ days

The mean life of all batteries is supposed to be at least 64 days. T significance level is α is 0.05. That is, the probability of rejecting the n hypothesis when it is actually is true should not exceed 0.05.

Step 1. *State the null and alternative hypothesis*

We write the null and alternative hypothesis as

$H_0 : \mu \geq 64$ days (The mean life is at least 64 days)

$H_1 : \mu < 64$ days (The mean life is less than 64 days)

Step 2. *Select the distribution to use*

The sample size is small ($n = 15$), and the life of a battery is approximate normally distributed. Since population standard deviation is unknown, use the Student's t distribution to make the test.

Step 3. *Determine the rejection and nonrejection regions*

The significance level is 0.05. The < sign in the alternative test indicates th the test is left tailed with the rejection region in the left tail of the distribution curve.

Area in the left tail= $\alpha = 0.05$

Degree of freedom= $n - 1 = 15 - 1 = 14$,

From the Student's t distribution table (Table 2 of Appendix), the critic value of t for 14 degrees of freedom and an area 0.05 in the left t is -1.761. (Fig.1.9).

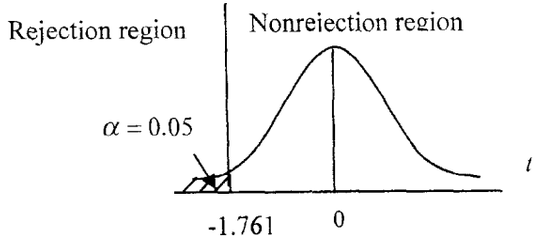

Fig.1.9

tep 4. *Calculate the value of the test statistic*

s σ is not known, and sample size is small, we calculate the t value as llows

$$T.S. = t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{62 - 64}{3/\sqrt{15}} = -2.50$$

tep 5. *Make a decision*

he value of $T.S. = t = -2.50$ is less than the critical value of $t = -1.761$, and falls in the rejection region. Therefore, we reject H_0 and conclude that the mple mean is too small compared to 62 days (company's claimed value f μ) and the difference between the two may not be attributed to chance lone. We can conclude that the mean life of company's batteries is less than 2 days.

emark: The conclusion of a t-test can also be strengthened by reporting e significance probability (p- value) of the observed statistic. Since the t ble provides only a few selected percentage points, we can get an idea bout the p-value but not its exact determination. For instance, the data in xample above gave an observed value $T.S. = t = -2.50$ with degree of reedom=14. Scanning the t table for $(n-1) = 14$, we notice that that 2.50 es between $t_{0.025}$ and $t_{0.010}$. Therefore, the p-value of $t = -2.50$ is higher than .025 but not as great as 0.010.

Exercises

. For each of the following examples of tests of hypothesis about μ, show ie rejection and nonrejection regions on the t distribution curve.

) A two tailed test with $\alpha = 0.2$ and $n = 14$

) A left tailed test with $\alpha = 0.005$ and $n = 23$

) A right tailed test with $\alpha = 0.025$ and $n = 14$

. Consider the null hypothesis $H_0 : \mu = 45$ about the mean of a population hat is normally distributed. Suppose a random sample of 20 observations is aken from this population to make this test. Using $\alpha = 0.05$ show the ejection and nonrejection regions and find critical value(s) for t for

) left tailed test; b) two tailed test; c) right tailed test

3. Consider $H_0 : \mu = 40$ versus $H_0 : \mu > 40$ for a population that is normal distributed.

a) A random sample of 16 observations taken from this population produc a sample mean of 45 and a standard deviation of 5. Using $\alpha = 0.025$, wou you reject the null hypothesis?

b) Another random sample of 16 observations taken from the sam population produced a sample mean of 41.9 and a standard deviation of Using $\alpha = 0.025$, would you reject the null hypothesis?

Comment on the result of parts a) and b).

4. Assuming that respective populations are normally distributed, make t following hypothesis tests.

a) $H_0 : \mu = 60$; $H_1 : \mu \neq 60$; $n = 14$; $\bar{x} = 56$; $s = 9$; $\alpha = 0.05$

b) $H_0 : \mu = 35$; $H_1 : \mu < 35$; $n = 24$; $\bar{x} = 29$; $s = 5.4$; $\alpha = 0.005$

c) $H_0 : \mu = 47$; $H_1 : \mu > 47$; $n = 18$; $\bar{x} = 51$; $s = 6$; $\alpha = 0.001$

5. A business school claims that students who complete a three month cour of typing course can type on average, at least 1200 words an hour.

A random sample of 25 students who completed this course typed, average, 1130 words an hour with a standard deviation of 85 words. Assum that the typing speeds for all students who complete this course have approximate normal distribution.

Using the 5% significance level, can you conclude, that the claim of t business school is true?

6. The supplier of home heating furnaces of a new model claims that t average efficiency of the new model is at least 60. Before buying the heating furnaces, a distributor wants to verify the supplier's claim is vali To this end, the distributor chooses a random sample of 9 heating furnaces a new model and measures their efficiency. The data are

63; 72; 64; 69; 59; 65; 66; 64; 65

Determine the rejection region of the test with $\alpha = 0.05$. Apply the test a state your conclusion.

7. A past study claims that adults spend an average of 18 hours a week o leisure activities. A researcher wanted to test this claim. He took a sample 10 adults and asked them about the time they spend per week on leisu activities. Their responses (in hours) were as follows

4; 25; 22; 38; 16; 26; 19; 23; 41; 33

ssume that the time spent on leisure activities by all adults is normally istributed. Using the 5% significance level, can you conclude that the claim f earlier study is true?

: According to the department of Labor, private sector workers earned, on verage $354.32 a week in 2001. A recently taken random sample of 400 rivate sector worker showed that they earn, on average, $362.50 a week ith a standard deviation of $72. Find p-value for the test with an alternative ypothesis that the current wean weekly salary of private sector workers is ifferent from $354.32.

: A manufacturer of a light bulbs claims that the mean life of these bulbs is t least 2500 hours. A consumer agency wanted to check whether or not this laim is true. The agency took a random sample of 36 such bulbs and tested iem. The mean life for the sample was found to be 2447 hours with a :andard deviation of 180 hours.

) Do you think that the sample information supports the company's claim? se $\alpha = 2.5\%$.

) What is the Type I error in this case? Explain. What is the probability of iaking this error?

) Will your conclusion of part a) change if the probability of making a ype I error is zero?

0. Given the eight sample observations 31, 29, 26, 33, 40, 28, 30, and 25, st the null hypothesis that the mean equals 35 versus the alternative that it oes not. Let $\alpha = 0.01$.

Answers

a) reject H_0 if $t < -1.350$ or $t > 1.350$; b) reject H_0 if $t < -2.819$; c)reject H_0 if $t > 2.160$; **2.** a) reject H_0 if $t < -1.729$; b) reject H_0 if either $t > 2.093$ r $t < -2.093$; c) reject H_0 if $t > 1.729$; **3.** a) $T.S. = t = 4.00$; reject H_0;) $T.S. = t = 1.086$; accept H_0; **4.** a) $T.S. = -1.663$; accept H_0; b) $T.S. = -5.443$; eject H_0; c) $T.S. = 2.828$; accept H_0; **5.** $T.S. = -4.118$; reject H_0; $t \geq 1.860$; $T.S. = 4.26$; H_0 is rejected at $\alpha = 0.05$; **7.** $T.S. = 2.692$; eject H_0; **8.** $T.S. = 2.27$; $p - \text{value} = 0.0232$; **9.** a) $T.S. = z = -1.77$; accept H_0;) 0.025; c) no; **10.** $T.S. = -2.85$; H_0 is not rejected.

1.6. Tests of the population proportion (Large sample)

Often we want to conduct test of hypothesis about a population proportion. This section presents the procedure to perform tests of hypothesis about t population proportion. p for large samples ($n \geq 40$). The procedure to ma such tests is similar in many respects to the one for the population mean μ.

The value of the test statistic $T.S. = z$ for the sample proportion $\hat{p}$ comput as

$$T.S. = z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

where $\hat{p}$ -is the sample proportion, and the value of p_0 used in this formu is the one used in the null hypothesis.
Then, if the number of sample observations is large and observed proporti is $\hat{p}$, the following tests have significance level α :

1. To test either null hypothesis

$H_0 : p = p_0$ or $H_0 : p \leq p_0$ against the alternative

$H_1 : p > p_0$

the decision rule is

Reject H_0 if $T.S. > z_\alpha$

2. To test either null hypothesis

$H_0 : p = p_0$ or $H_0 : p \geq p_0$ against the alternative

$H_1 : p < p_0$

the decision rule is

Reject H_0 if $T.S. < -z_\alpha$

3. To test the null hypothesis

$H_0 : p = p_0$ against the two sided alternative

$H_1 : p \neq p_0$

the decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$

)nce again, z_α is the number for which

$P(Z > z_\alpha) = \alpha$

nd Z is the standard normal distribution.

xample:

Ir. A and Mr. B are running for local public office in a large city. Mr. A ays that only 30% of the voters are in favor of a certain issue, a law to sell quor on Sundays. Mr. B doubts A's statement and believes that more than 0% favor such legislation. Mr. B pays for an independent organization to ake a study of this situation. In a random sample 400 voters, 160 favored ie legislation. What conclusions should the polling organization report to Ir. B?

olution:

et p_0 be proportion of all people who favor such legislation and $\hat{p}$ the orresponding sample proportion. Then from given information,

$= 400$; $p_0 = 0.30$; $\hat{p} = \frac{160}{400} = 0.40$. Let $\alpha = 0.05$.

he null and alternative hypotheses are as follows

$H_0 : p = p_0 = 0.30$

$H_1 : p > 0.30$

he decision rule is to reject the null hypothesis in favor of alternative if

$T.S. > z_\alpha$

$\alpha = 0.05$; $\alpha/2 = 0.025$.

$P(Z > z_{\alpha/2}) = P(Z > z_{0.025}) = 0.025$.

$P(Z > z_{0.025}) = F(z_{0.025}) = 0.975$ and

$z_{\alpha/2} = z_{0.025} = 1.645$

rom the given information we calculate the value of test statistic as

$$T.S. = z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.40 - 0.30}{\sqrt{0.30 \cdot 0.70/400}} = 4.36$$

ince $4.36 > 1.645$ we reject H_0. We make conclusion that more than 30% of oters are in favor of a law to sell liquor on Sundays.

Exercises

1. Make the following hypothesis tests about p.

a) $H_0 : p = 0.45$; $H_1 : p \neq 0.45$; $n = 100$; $\hat{p} = 0.48$; $\alpha =$

b) $H_0 : p = 0.72$; $H_1 : p < 0.72$; $n = 700$; $\hat{p} = 0.65$; $\alpha =$

c) $H_0 : p = 0.30$; $H_1 : p > 0.30$; $n = 200$; $\hat{p} = 0.34$; $\alpha =$

2. Consider $H_0 : p = 0.70$ versus $H_0 : p \neq 0.70$.

a) A random sample of 600 observations produced a sample proportio to 0.67. Using $\alpha = 0.01$, would you reject the null hypothesis?

b) Another random sample of 600 observations taken from the population produced a sample proportion of 0.76. Using $\alpha = 0.01$, wou reject the null hypothesis?

Comment on the result of parts a) and b).

3. A food company is planning to market a new type of ice cream. marketing this ice cream, the company wants to find what percentage people like it. The company's management has decided that it will this ice cream only if at least 35% of people like it. The company's re department selected a random sample of 400 persons and asked them this ice cream. Of these 400 persons, 128 said they liked it.

a) Testing at 2.5% significance level, can you conclude that the co should market this yogurt?

b) What will your decision be in part a) if the probability of making a error is zero?

4. A mail order company claims that at least 60% of all orders are within 48 hours. The quality control department took a sample of 500 and found that 310 of them were mailed within 48 hours of the placem the orders. Testing at 1% significance level, can you conclude th company's claim is true?

5. Let p=proportion of adults in a city who required a lawyer in the pas

a) Determine the rejection region for $\alpha = 0.05$ level test of $H_0 : p = 0.2$ against $H_1 : p > 0.25$.

b) If 65 persons in a random sample of 200 required lawyer service does the test conclude?

6. A magazine claims that 25% of its readers are university students. A random sample of 200 readers is taken and 42 of these readers are university students. Use $\alpha = 0.10$ level of significance to test the validity of the magazine's claim.

7. Suppose that in order to test the hypothesis that $p = 0.6$ against the alternative that $p < 0.6$, we decide to obtain a sample of size 100 and reject H_0 if we obtain fewer than 48 successes.

a) What is the approximate size of the Type I error?

b) If the value of p is really 0.5, what is the size of Type II error?

8. An educator wishes to test $H_0 : p = 0.3$ against $H_1 : p > 0.3$, where p-proportion of football players who graduate university in four years.

a) State the test statistic and the rejection region having $\alpha = 0.05$.

b) If 19 out of a random sample of 48 players graduated in four years, what does the test conclude? Also evaluate p-value.

9. The president of a company that produces national brand coffee claims that 40% of the people prefer to buy national brand coffee. A random sample of 700 people who buy coffee showed that 252 of them buy national brand coffee.

a) Using $\alpha = 0.01$, can you conclude that the percentage of people who buy national brand coffee is different from 40%?

b) Find the p-value for the test. Using this p-value, would you reject the null hypothesis at $\alpha = 0.05$? What if $\alpha = 0.02$?

Answers

1. a) $T.S. = z = 0.60$; do not reject H_0; b) $T.S. = -4.12$; reject H_0; c) $T.S. = 1.23$; do not reject H_0; **2.** a) $T.S. = -1.60$; do not reject H_0; b) $T.S. = 3.21$; reject H_0; **3.** a) $T.S. = -1.26$; do not reject H_0; b) do not reject H_0; **4.** $T.S. = 0.91$; accept H_0; **5.** a) $z \geq 1.645$; b) $T.S. = 2.45$; reject H_0; **6.** $T.S. = -1.31$; accept H_0; **7.** a) about 0.0071 b)approximately 0.6554; **8.** a) $z \geq 1.645$; b) $T.S. = 1.45$; accept H_0 for $\alpha = 0.05$; $p - \text{value} = 0.0735$; **9.** a) $T.S. = -2.16$; do not reject H_0;b) p-value=0.0308; reject H_0 at $\alpha = 0.05$; do not reject H_0 at $\alpha = 0.02$

1.7. Tests of the variance of a normal distribution

In addition to the need for tests based on the sample mean and sam proportion, there are a number of situations where we want to determine the population variance is a particular value or set of values. The basis developing particular tests lies in the fact that the random variable

$$\chi^2_{n-1} = \frac{(n-1)\cdot s^2}{\sigma^2}$$

follows a Chi-square distribution with $(n-1)$ degrees of freedom.

The value of the test statistic χ^2_{n-1} is calculated as

$$T.S. = \chi^2_{n-1} = \frac{(n-1)\cdot s^2}{\sigma^2}$$

We are given a random sample of n observations from a normally distribut population with variance σ^2. If we observe the sample variance s^2, then t following tests have significance level α :

1. To test either null hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{or} \quad H_0 : \sigma^2 \le \sigma_0^2 \qquad \text{against}$$

alternative

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

the decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > \chi^2_{n-1,\alpha}$$

2. To test either null hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{or} \quad H_0 : \sigma^2 \ge \sigma_0^2 \qquad \text{against}$$

alternative

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

the decision rule is

$$\text{Reject } H_0 \text{ if } T.S. < \chi^2_{n-1,1-\alpha}$$

3. To test the null hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{against the two sided alternative}$$

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

he decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > \chi^2_{n-1,\alpha/2} \quad \text{or} \quad T.S. < \chi^2_{n-1,1-\alpha/2}$$

where χ^2_{n-1} is a Chi-square random variable and $P(\chi^2_{n-1} > \chi^2_{n-1,\alpha}) = \alpha$.

Example:

/ariance of yearly earnings of all state employees for all 40 states is 49000 square dollars. A sample of 29 employees selected from state A roduced a variance of their earnings equal to $600 000 square dollars. Test t 5% significance level if the variance of yearly earnings of state employees n state A is different from $490 000 square dollars. Assume that the yearly arnings of all state employees in state A have an (approximate) normal istribution.

olution:

rom the given information,

$$n = 29; \qquad \alpha = 0.05; \qquad s^2 = 600000$$

he null and alternative hypotheses are

$$H_0 : \sigma^2 = 49000$$

$$H_1 : \sigma^2 \ne 49000$$

Ve use Chi square distribution to use. The decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > \chi^2_{n-1,\alpha/2} \quad \text{or} \quad T.S. < \chi^2_{n-1,1-\alpha/2}$$

$$\alpha/2 = 0.025; \qquad 1-\alpha/2 = 0.975; \qquad v = n-1 = 29-1 = 28$$

hen from Table 3 of appendix we obtain

$$\chi^2_{n-1,\alpha/2} = \chi^2_{28,0.025} = 44.461 \text{ and } \chi^2_{n-1,1-\alpha/2} = \chi^2_{28,0.975} = 15.308$$

he value of the test statistic is

$$T.S. = \chi^2_{n-1} = \frac{(n-1)\cdot s^2}{\sigma^2} = \frac{(29-1)\cdot(600000)}{490000} = 34.286$$

he value of the test statistic 34.286 is between the two critical values of χ^2_{n-1}, 15.308 and 44.461, and falls in the nonrejection region. Consequently ve fail to reject H_0 and conclude that the population variance of yearly arnings of all employees in state A is not different from 490000 square ollars.

Exercises

1. A sample of 24 observations selected from a normally distribut population produced a sample variance of 12.

a) Write the null hypothesis and alternative hypothesis, and decision rule test if the population variance is different from 10.

b) Using $\alpha = 0.05$, find the critical values of χ^2_{n-1}. Show the rejection a nonrejection regions on a Chi-square distribution curve.

c) Using the 5% significance level, will you reject the null hypothesis stat in part a)?

2. A sample of 25 observations selected from a normally distribut population produced a sample variance of 18. Using the 2.5% significan level, test hypothesis if the population variance is less than 25.

3. Usually people do not like waiting in line for service for a long time. A bank management does not want the variance of the waiting time for customers to be higher than 4.0 square minutes. A random sample of customers taken from this bank gave the variance of the waiting times equ to 7.9 square minutes. Test at 1% significance level if the variance of t waiting time for all customers at this bank is higher than 4.0 square minut Assume that the waiting time for all customers is normally distributed.

4. Test $H_0 : \sigma = 10$ against $H_1 : \sigma > 10$ with $\alpha = 0.05$ in each case

a) $n = 25$; $\sum_{i=1}^{25} (x_i - \bar{x})^2 = 4016$

b) $n = 15$; $s = 12$

5. A sample of seven observations taken from a population produced t following data

10; 8; 13; 15; 6; 8; 13

Assuming that the population from which this sample is selected is norma distributed, test at 2.5 significance level if the population variance i different from 10.

6. A drug manufacturer requires that the variance for a chemical contained the bottles of certain type of drug should not exceed 0.03 square grams. sample of 25 such bottles gave the variance for this chemical as 0.06 squ grams. Test at the 1% significance level if the variance of this chemical in such bottles exceeds 0.03 square grams. Assume that the amount of t chemical in all such bottles is (approximately) normally distributed. Find a interpret p-value of this test.

7. A random sample of ten students was asked, in hours, for time they spent studying in the week before final exams. The data are as follows:

28; 57; 42; 35; 61; 39; 55; 46; 49; 38

Assuming that the population distribution is normal, test at 5% significance level against two sided alternative the null hypothesis that the population standard deviation is 10 hours

8. Company claims that its employees earns a mean of at least \$40 000 in a year and that the population standard is no more than \$6 000. Earnings of a random sample of nine employees of this company produced

$$\sum_{i=1}^{9} x_i = 333 \quad \text{and} \quad \sum_{i=1}^{9} (x_i - \bar{x})^2 = 312$$

where x_i are measured in thousands of dollars and population distribution can be assumed to be normal. Test at 10% significance level the null hypothesis that the population standard deviation is at most \$6 000.

9. State whether each of the following statement is true or false

a) The significance level of a test is the probability that the null hypothesis is false.

b) A Type I error occurs when a true null hypothesis is rejected.

c) A null hypothesis is rejected at the 0.025 level, but is accepted at the 0.01 level. This means that p-value of the test lies between 0.01 and 0.025.

d) If a null hypothesis is rejected against an alternative at the 5% level, then using the same data, it must be rejected against that alternative at the 1% level.

f) If a null hypothesis is rejected against an alternative at the 1% level, then using the same data, it must be rejected against that alternative at the 5% level.

g) The p- value of a test is the probability that the null hypothesis is true.

Answers

1. b) reject H_0 if $\chi^2_{23} > 38.08$ or $\chi^2_{23} < 11.69$; b) $T.S. = \chi^2_{23} = 27.6$; do not reject H_0; **2.** $T.S. = \chi^2_{24} = 17.280$; do not reject H_0; **3.** $T.S. = \chi^2_{24} = 47.400$; reject H_0; **4.** a) $T.S. = 40.16$; reject H_0; b) $T.S. = 20.16$; accept H_0; **5.** $T.S. = 6.571$; do not reject H_0; **6.** $T.S. = 48.00$; reject H_0; reject H_0 for $\alpha = 0.0005$; **7.** $T.S. = \chi^2 = 9.999$; accept H_0; **8.** $T.S. = 8.67$; accept H_0; **9.** a) false; b) true; c) false; d) false; f) true; g) false.

1.8. Tests for the difference between two population means

1.8.1. Tests based on paired samples

Suppose that a random sample of n matched pairs of observations is obtain
from populations with means μ_x and μ_y. The observations will be deno
by $(x_1, y_1), (x_2, y_2), \ldots\ldots (x_n, y_n)$.
Let

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

and $$s_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} d_i^2 - n(\bar{d})^2}{n-1}}$$

denote the observed sample mean and standard deviation for the differences $d_i = x_i - y_i$. Let us denote difference between two populati means by $D_0 = \mu_x - \mu_y$. In this case test statistic will be calculated as

$$T.S. = \frac{\bar{d} - D_0}{s_{\bar{d}} / \sqrt{n}}$$

If the population differences is a normal distribution, then the following te have significance level α

1. To test either null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \qquad \text{or} \qquad H_0 : \mu_x - \mu_y \le D_0$$

against the alternative

$$H_1 : \mu_x - \mu_y > D_0$$

the decision rule is

Reject H_0 if $T.S. > t_{n-1,\alpha}$

2. To test either null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \qquad \text{or} \qquad H_0 : \mu_x - \mu_y \ge D_0$$

against the alternative

$$H_1 : \mu_x - \mu_y < D_0$$

e decision rule is

Reject H_0 if $T.S. < -t_{n-1,\alpha}$

To test the null hypothesis

$$H_0 : \mu_x - \mu_y = D_0$$ against the two sided alternative

$$H_1 : \mu_x - \mu_y \ne D_0$$

e decision rule is

Reject H_0 if $T.S. > t_{n-1,\alpha/2}$ or $T.S. < -t_{n-1,\alpha/2}$

ere, $t_{n-1,\alpha}$ is the number for which

$$P(t_{n-1} > t_{n-1,\alpha}) = \alpha$$

here the random variable t_{n-1} follows a Student's t distribution with
$n-1)$ degrees of freedom.

emark: When we want to test the null hypothesis that the two population eans are equal, we set $D_0 = 0$.

xample:

medical researcher wishes to determine if a pill has the undesirable side ffect of reducing the blood pressure of the user. The study involves cording the initial blood pressures of 7 college age adults. After they use ie pill regularly for three month, their blood pressures are again recorded. he researcher wishes to draw inferences about the effect of the pill on blood ressure from the information given in table

Before x_i	64	71	68	66	73	62	70
After y_i	60	66	66	69	63	57	62

)o the data substantiate the claim that use of the pill reduces the blood ressure? Use $\alpha = 0.01$. Assume that the population of paired differences las a normal distribution.

olution:

et d be the difference between the pressures before and after using pills.

$$d = before - after = x_i - y_i$$

The necessary calculations are shown in the following table

Before	After	Difference d	d^2
64	60	4	16
71	66	5	25
68	66	2	4
66	69	-3	9
73	63	10	100
62	57	5	25
70	62	8	64
		$\sum d = 31$	$\sum d^2 = 243$

The values of $\bar{d}$ and S_d are calculated as follows:

$$\bar{d} = \frac{\sum d}{n} = \frac{31}{7} = 4.43$$

$$S_d = \sqrt{\frac{1}{n-1}\left[\sum d^2 - n\cdot(\bar{d})^2\right]} = \sqrt{\frac{1}{6}(243 - 7\cdot 4.43^2)} = 4.198.$$

Let μ_x be the mean blood pressure for all adults before and μ_y -after the pill.

The null and alternative hypotheses are

$$H_0 : \mu_x - \mu_y = 0 \text{ (no difference)}$$

against

$$H_1 : \mu_x - \mu_y > 0 \text{ (mean decreases)}$$

The decision rule is that

Reject H_0 if $T.S. > t_{n-1,\alpha}$

$$T.S. = \frac{\bar{d} - D_0}{s_{\bar{d}}/\sqrt{n}} = \frac{4.43 - 0}{4.198/\sqrt{7}} = 2.792$$

$$t_{n-1,\alpha} = t_{6,0.01} = 3.14$$

Since 2.792<3.14, we accept H_0 and make conclusion at the level 0.0 using pills does not affect blood pressure.

Exercises

Perform the following tests of hypothesis assuming that the population of aired differences are normally distributed.

) $H_0 : \mu_d = D_0 = 0 ; H_1 : \mu_d = D_0 \neq 0 ; n = 9 ; \bar{d} = 6.7 ; s_d = 2.5 ; \alpha = 0.10$

) $H_0 : \mu_d = D_0 = 0 ; H_1 : \mu_d = D_0 > 0 ; n = 22 ; \bar{d} = 14.3 ; s_d = 6.4 ; \alpha = 0.05$

) $H_0 : \mu_d = D_0 = 0 ; H_1 : \mu_d = D_0 < 0 ; n = 17 ; \bar{d} = -9.3 ; s_d = 4.8 ; \alpha = 0.01$

It is claimed that an industrial safety program is effective in reducing the ss of working hours due to factory accidents. The following data are llected concerning the weekly hours due to accidents in nine plants both efore and after the safety program is installed

Before x_i	90	86	72	65	44	52	46	38	43
After y_i	85	87	70	62	44	53	42	35	46

o the data substantiate the claim? Use $\alpha = 0.05$.

ssume that the population of paired differences is (approximately) ormally distributed.

A company claims that the course it offers significantly increases the riting speed of secretaries. The following table gives the scores of 8 ecretaries before and after they attended this course

Before x_i	81	75	89	91	65	70	90	69
After y_i	97	72	93	110	78	69	115	75

Using the 5% significance level, can you conclude that attending this course ncreases the writing speed of secretaries?

ssume that the population of paired differences is (approximately) ormally distributed.

A random sample of nine employees was selected to test for the ffectiveness of hypnosis on their job performance. The following table ives the job performance ratings (on a scale of 1 to 4, with 1 being the

lowest and 4 being the highest) before and after these employe hypnosis.

Before x_i	2.3	2.8	3.1	2.7	3.4	2.6	2.8	2.5
After y_i	2.6	3.2	3.0	3.5	3.7	2.4	2.9	2.9

Test at the 5% significance level if there is an improvement in t performances of employees due to hypnosis.
Assume that the population of paired differences is (approximately) normally distributed.

Answers

1. a) $T.S. = 8.040$ reject H_0 ;b) $T.S. = 10.847$;reject H_0 ; c) $T.S. = -7.989$ reject H_0 ; **2.** $T.S. = 1.48$; do not reject H_0 ;**3.** $T.S. = -2.807$; reject H **4.** $T.S. = -2.236$; accept H_0 .

1.8.2. Tests based on independent samples (Known variance or large sample size)

Let us consider the case where we have independent random sample two normally distributed populations. The first population has mean variance σ_x^2 and we obtain a random sample of size n_x . The s population has mean μ_y and variance σ_y^2 and we obtain a random sam size n_y .

We know that if the sample means are denoted $\bar{x}$ and $\bar{y}$, then the ra variable

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

a standard normal distribution. If the population variances are known, ; for the difference between the population means can be based on this

lt. The value of the test statistic z for $(\bar{x} - \bar{y})$ is computed as

$$T.S. = z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

the following tests have a significance level α

'o test either null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{or} \quad H_0 : \mu_x - \mu_y \le D_0$$

inst the alternative

$$H_1 : \mu_x - \mu_y > D_0$$

decision rule is

Reject H_0 if $T.S. > z_\alpha$

'o test either null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{or} \quad H_0 : \mu_x - \mu_y \ge D_0$$

inst the alternative

$$H_1 : \mu_x - \mu_y < D_0$$

decision rule is

Reject H_0 if $T.S. < -z_\alpha$

'o test the null hypothesis

$$H_0 : \mu_x - \mu_y = D_0$$

against the two sided alternative

$$H_1 : \mu_x - \mu_y \ne D_0$$

decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$

nark: If the sample sizes are large $(n_x > 30; n_y > 30)$ then a good roximation at significance level α can be made if the population iances σ_x^2 and σ_y^2 are replaced by the sample variances s_x^2 and s_y^2.
ıddition the central limit theorem leads to good approximations even if populations are not normally distributed.

Example:

According to the Bureau of Labor Statistics, last year university instr earned an average $440 per month and college instructors earned an av of $420 per month. Assume that these mean earnings have been calculat samples of 400 and 600 instructors taken from the two popula respectively. Further assume that the standard deviations of monthly ea of the two populations are $50 and $63, respectively. Test at 1% signifi level if the mean monthly earnings of the two groups of the instructo different.

Solution:

From the information given above,

$$n_x = 400; \qquad \bar{x} = 440; \qquad \sigma_x = 50;$$

$$n_y = 600; \qquad \bar{y} = 420; \qquad \sigma_y = 63;$$

where the subscript x refers to university instructors and y-to co instructors. Let

μ_x = mean monthly earnings of all university instructors

μ_y = mean monthly earnings of all college instructors.

We are to test if the two population means are different. The nul alternative hypotheses are

$H_0 : \mu_x - \mu_y = 0$ (the monthly earnings are not differe

$H_1 : \mu_x - \mu_y \neq 0$ (the monthly earnings are different).

The decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$

First of all we find the value of $z_{\alpha/2}$. Since $\alpha/2 = 0.005$, the value of $z_{\alpha/2}$ (approximately) 2.58 and $-z_{\alpha/2} = -2.58$.

The value of the test statistic $T.S. = z$ is computed as follows:

$$T.S. = z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} = \frac{(440 - 420) - (0)}{\sqrt{\frac{50^2}{400} + \frac{63^2}{600}}} = 5.57.$$

$7 > 2.58$ and the value of test statistic $T.S. = z = 5.57$ falls in the rejection ion, we reject the null hypothesis H_0. Therefore, we conclude that the n monthly earnings of the two groups of instructors are different.

e that we can not say for sure that two means are different. All we can is that the evidence from the two samples is very strong that the responding population means are different.

Exercises

The following information is obtained from two independent samples ected from two populations

$n_1 = 155; \ \bar{x} = 5.58; \ s_x = 1.62$

$n_2 = 190; \ \bar{y} = 4.80; \ s_y = 1.52$

t at the 1% significance level if the two population means are the same inst the alternative that they are different.

Daily wage is $13.62 for transportation workers and $11.61 for factory rkers. Assume that these two estimates are based on random samples of 0 and 1200 workers taken, respectively, from the two populations. Also ume that the standard deviations of the two populations are $1.85 and 40, respectively.

Test at the 5% significance level if the mean daily wage of transportation rkers and factory workers are the same against the alternative that it is her for transportation workers.

What will your decision be in part a) if the probability of making a Type I or is zero. Explain.

A consulting firm was asked by a large insurance company to investigate business majors were better salespersons. A sample of 40 salespersons h a business degree showed that they sold an average of 10 insurance icies per week with a standard deviation of 1.80. Another sample of 45 espersons with a degree other than business showed that they sold an erage of 8.5 insurance policies per week with a standard deviation of 1.35. ing the 1% significance level, can you conclude that person with a siness degree are better salespersons than those who have a degree in other area?

4. The management at the bank A claims that the mean waiting time customers at its branches is less than that at the bank B, which i competitor. They took a sample of 200 customers from the bank A and that they waited an average of 4.60 minutes with a standard deviation minutes before being served. Another sample of 300 customers take the bank B showed that these customers waited an average of 4.85 m with a standard deviation of 1.5 minutes before being served.

a) Test at the 2.5% significance level if the claim of the management bank A is true.

b) Calculate the *p*-value. Based on this *p*-value, would you reject th hypothesis if $\alpha = 0.01$? What if $\alpha = 0.05$?

5. A production line is designed on the assumption that the differe mean assembly times for two operations is 5 minutes. Independent te the two assembly operations show the following results:

Operation A	Operation B
$n_1 = 100$	$n_2 = 50$
$\bar{x} = 14.8$ minutes	$\bar{y} = 10.4$ minutes
$s_x = 0.8$ minutes	$s_y = 0.6$ minutes

For $\alpha = 0.02$, test the hypothesis that the difference between the assembly times is 5 minutes.

6. An investigation was carried out to determine if women employees well paid as their male counterparts. Random samples of 75 males a females are selected. Their mean salaries were 45 530 and 44 620, sta deviations were 780 and 750, correspondingly. If you were to test th hypothesis that the mean salaries are equal against the two sided alter what would be the conclusion of your test with $\alpha = 0.05$?

7. For a random sample of 125 state companies, the mean number changes was 1.91 and the standard deviation was 1.32. For a random s of 86 private companies, the mean number of job changes was 0.21 a standard deviation was 0.53. Test the null hypothesis that the popu means are equal against the alternative that the mean number of job ch is higher in state companies than for private companies.

Answers

$S. = z = 4.56$; reject H_0; **2.** a) $T.S. = 28.27$; reject H_0; b) do not reject H_0; $S. = 4.30$; reject H_0; **4.** a) $T.S. = -2.06$; reject H_0; b) p-value=0.0197; not reject H_0 at $\alpha = 0.01$?; reject H_0 at $\alpha = 0.05$; **5.** $T.S. = -5.15$; ct H_0; **6.** $T.S. = 7$; reject H_0; **7.** $T.S. = 13$; reject H_0 at any level.

1.8.3. Tests based on independent samples (Population variances are unknown and equal)

Many times it may not be possible to take large samples from ulations to make inferences about the difference between two population ns. This section discusses how to test a hypothesis about the difference veen two population means when samples are small $< 30)$, $(n_y < 30)$ and independent. Our main assumption in this case is the two populations from which the two samples are drawn are proximately) normally distributed. If this assumption is true, and we w the population variances, we can still use the normal distribution to ce inferences about $(\mu_x - \mu_y)$ when samples are small and independent. vever, we usually do not know the population variances σ_x^2 and σ_y^2. In h cases, we replace the normal distribution by the Student's t distribution nake inferences about $(\mu_x - \mu_y)$ for small and independent samples. In section we will make one more assumption that the variances of the two ulations are equal. When the variances of the two populations are equal, can use σ^2 for both σ_x^2 and σ_y^2. Since σ^2 is unknown, we replace it by oint estimator s_p^2, which is called pooled sample variance.

w assume that we have independent random samples of size n_x and n_y ervations from normally distributed populations with means μ_x and and a common variance. The sample variances s_x^2 and s_y^2 are used to npute a pooled variance estimator

$$s_p^2 = \frac{(n_x - 1)\cdot s_x^2 + (n_y - 1)\cdot s_y^2}{(n_x + n_y - 2)}$$

The value of the test statistic z for $(\bar{x} - \bar{y})$ is computed as

$$T.S. = t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}}$$

and the following tests have a significance level α

1. To test either null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \qquad \text{or} \qquad H_0 : \mu_x - \mu_y \le D_0$$

against the alternative

$$H_1 : \mu_x - \mu_y > D_0$$

the decision rule is

Reject H_0 if $T.S. > t_{n_x+n_y-2,\alpha}$

2. To test either null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \qquad \text{or} \qquad H_0 : \mu_x - \mu_y \ge D_0$$

against the alternative

$$H_1 : \mu_x - \mu_y < D_0$$

the decision rule is

Reject H_0 if $T.S. < -t_{n_x+n_y-2,\alpha}$

3. To test the null hypothesis

$$H_0 : \mu_x - \mu_y = D_0$$

against the two sided alternative

$$H_1 : \mu_x - \mu_y \ne D_0$$

the decision rule is

Reject H_0 if $T.S. > t_{n_x+n_y-2,\alpha/2}$ or $T.S. < -t_{n_x+n_y-2,\alpha/2}$

Here, $t_{n_x+n_y-2,\alpha}$ is the number for which

$$P(t_{n_x+n_y-2} > t_{n_x+n_y-2,\alpha}) = \alpha$$

ere the random variable $t_{n_x+n_y-2,\alpha}$ follows a Student's t distribution with

$_x + n_y - 2)$ degrees of freedom.

ample:

sample of 12 cans of Brand A diet soda gave a mean number of calories of per can with a standard deviation of 2 calories. Another sample of 15 cans Brand B diet soda gave the mean number of calories of 24 per can with a ndard deviation of 3 calories. At the 1% significance level, are the mean mber of calories per can different for these two brands of diet soda? sume that the calories per can of diet soda are normally distributed for ch of the two brands and that the variances for the two populations are ual.

lution:

t μ_x and μ_y be the mean number of calories per can for diet soda of Brand

and Brand B, respectively, and let $\bar{x}$ and $\bar{y}$ be the means of respective mples. From the given information,

$$n_x = 12; \qquad \bar{x} = 22; \qquad s_x = 2;$$

$$n_y = 15; \qquad \bar{y} = 24; \qquad s_y = 3$$

he significance level is $\alpha = 0.01$.

'e are to test for the difference in the mean number of calories per can for vo brands. The null and alternative hypotheses are

$H_0 : \mu_x - \mu_y = 0$ (the mean number of calories are not different)

$H_1 : \mu_x - \mu_y \ne 0$ (the mean number of calories are different)

he decision rule is

Reject H_0 if $T.S. > t_{n_x+n_y-2,\alpha/2}$ or $T.S. < -t_{n_x+n_y-2,\alpha/2}$

$t_{n_x+n_y-2,\alpha/2} = t_{12+15-2,0.005} = t_{25,0.005} = 2.787$ and $-t_{25,0.005} = -2.787$.

he pooled estimate is

$$s_p^2 = \frac{(n_x - 1)\cdot s_x^2 + (n_y - 1)\cdot s_y^2}{(n_x + n_y - 2)} = \frac{(12-1)\cdot 2^2 + (15-1)\cdot 3^2}{(12+15-2)} = 6.8$$

The test statistic is then computed as

$$T.S. = t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}} \quad \frac{(22-24)-(0)}{\sqrt{\frac{6.8}{12} + \frac{6.8}{15}}} = -1.98$$

Because the value of test statistic $T.S. = t = -1.98$ for $(\bar{x} - \bar{y})$ falls i nonrejection region (Fig.1.10), we fail to reject the null hypot Consequently we conclude that there is no difference between the number of calories per can for the two brands of diet soda. The differe $\bar{x}$ and $\bar{y}$ observed for two samples may have occurred due to sampling only.

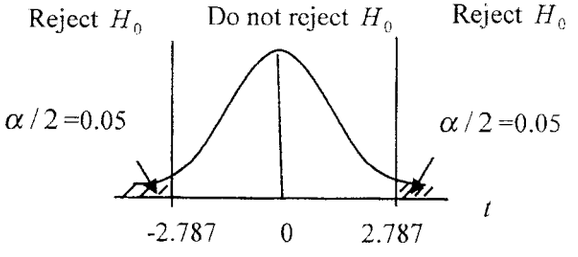

Fig.1.10

Exercises

1. The following information was obtained from two independent sa selected from two normally distributed populations with unknown but variances

$n_x = 20$; $\bar{x} = 33.75$; $s_x = 5.25$;

$n_y = 23$; $\bar{y} = 28.50$; $s_y = 4.55$

a) Test at 1% significance level if the two population means are differen
b) Test at 5% significance level if μ_x is different than μ_y.

he following summary statistics are recorded for independent random ples from two normally distributed populations with equal variances

Sample 1	Sample 2
$n_1 = 9$	$n_2 = 6$
$\bar{x} = 16.18$	$\bar{y} = 4.22$
$s_1 = 1.54$	$s_2 = 1.37$

the null hypothesis $\mu_1 - \mu_2 = 10$ against the alternative that $-\mu_2 > 10$ with $\alpha = 0.01$.

alary surveys of marketing and management majors show the following ting annual salary data

Marketing majors	management majors
$n_1 = 14$	$n_2 = 16$
$\bar{x} = \$14800$	$\bar{x}_2 = \$14300$
$s_1 = \$1000$	$s_2 = \$1400$

sider the test of the hypothesis that the mean annual salaries are the same both majors. For $\alpha = 0.05$ can you conclude that a difference exists in the n annual salary for the two majors?

professor took two samples, one of 21 males and another of 15 females n university students who were enrolled in business statistics at the same versity. He found that the mean score of male students in a mid-term mination in statistics was 75.3 with a standard deviation of 6.4, and the n score of female students was 78.3 with a standard deviation of 7.3. ume that the scores of all male and all female students are normally ributed with equal but unknown standard deviations.

t at the 2.5 significance level if the mean score in business statistics for male and female students are the same against the alternative that male dents have lower score than that for all female students.

The management of a supermarket wanted to investigate if the male tomers spend less money on average, than the female customers. A ple of 16 male customers who shopped at this supermarket showed that spent an average of \$55 with a standard deviation of \$12.50. Another ple of 22 female customers who shopped at the supermarket showed that spent an average of \$63 with a standard deviation of \$14.5. Assume that amounts of money spent at this supermarket by all male and female

customers are normally distributed with equal but unknown po variance. Test at the 5% significance level if the mean amount spe male and female customers are the same against the alternative th customers at this supermarket spend less than that of female custome

6. A bank has two branches. The quality department wanted to chec customers are equally satisfied with the service provided at the branches. Randomly selected customers asked to measure the satisfa services (on scale of 1 to 11, 1 being the lowest and 11 being the high A random sample of six customers from the branch A produced fo data:

9.50; 8.60; 8.59; 6.50; 4.79; 4.29

An independent random sample of six customers selected from the b produced following data:

10.21; 9.66; 7.67; 5.12; 4.88; 3.12

Stating any assumptions you need to make, test against two sided alt the null hypothesis that the two populations mean satisfaction index customers for the two branches are the same.

7. Given that $n_1 = 14$, $\bar{x} = 22$, $\sum (x_i - \bar{x})^2 = 30$, and $n_2 = 13$, $\bar{y} = 18$, $\sum (y_i - \bar{y})^2 = 24$. Test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ with $\alpha = 0.0$

8. A researcher wants to test the mean GPA (grade point averages) male and all female university students. She took a random sampl male students and 24 female students. She found that GPA's of t groups to be 2.62 and 2.74, respectively, with the corresponding st deviations equal to 0.43 and 0.38. Test at the 5% significance leve mean GPA's of the two populations are equal against two sided alterna Assume that the GPA's of all male and female students are no distributed with equal but unknown standard deviations.

Answers

1. a) $T.S. = t = 3.514$; reject H_0; b) $T.S. = t = 3.514$; reject H_0;

2. $T.S. = t = 2.52$; H_0 is not rejected; **3.** $T.S. = t = 1.11$; accept H_0;

4. $T.S. = -1.308$; accept H_0; **5.** $T.S. = -1.778$; reject H_0; **6.** We assu the values are normally distributed with equal variance; $T.S. = 0.183$; reject H_0 at 20% significance level; **7.** $T.S. = 7.071$; reject H_0;

8. $T.S. = -1.058$; accept H_0.

1.9. Tests for the difference between two population proportions (Large samples)

we will develop procedures for comparing two population proportions. will consider standard model with a random sample of n_x observations proportion $\hat{p}_x$ "successes" and an independent random sample of n_y rvations from population with proportion $\hat{p}_y$ "successes".

know that for large samples, proportions can be approximated as nally distributed random variables and as a result

$$Z = \frac{\left(\hat{p}_x - \hat{p}_y\right) - \left(p_x - p_y\right)}{\sqrt{\frac{p_x(1-p_y)}{n_x} + \frac{p_y(1-p_y)}{n_y}}}$$

a standard normal distribution.

want to test the hypothesis that the population proportions p_x and p_y are l. Denote their common value by p_0, then the value under this othesis

$$Z = \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\frac{p_0(1-p_0)}{n_x} + \frac{p_0(1-p_0)}{n_y}}}$$

ws a good approximation a standard normal distribution.

lly, unknown common proportion p_0 can be estimated by a pooled nator defined as

$$p_0 = \frac{n_x \cdot \hat{p}_x + n_y \cdot \hat{p}_y}{n_x + n_y} \quad \text{or} \quad p_0 = \frac{x_1 + y_1}{n_x + n_y}$$

re x_1 and x_2 are number of "successes" in n_x and n_y, respectively.

Which of these formulas is used to calculate p_0 depends on whether values of x_1 and y_1 or the values of $\hat{p}_x$ and $\hat{p}_y$ are known.

Testing equality of two population proportions:

We are given independent samples of size n_x and n_y with proportions successes $\hat{p}_x$ and $\hat{p}_y$. When we assume that the population proportions equal, an estimate of the common proportion is

$$p_0 = \frac{n_x \cdot \hat{p}_x + n_y \cdot \hat{p}_y}{n_x + n_y} \text{ or } p_0 = \frac{x_1 + y_1}{n_x + n_y}.$$

For large sample sizes $\left(n \cdot \hat{p} \cdot \hat{q} > 9\right)$ the value of the test statistic z is computed as

$$T.S. = z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{p_0(1-p_0)}{n_x} + \frac{p_0(1-p_0)}{n_y}}}$$

Then the following tests have significance level α:

1. To test either null hypothesis

$$H_0 : p_x - p_y = 0 \quad \text{or} \quad H_0 : p_x - p_y \le 0$$

against the alternative

$$H_1 : p_x - p_y > 0$$

the decision rule is

Reject H_0 if $T.S. > z_\alpha$

2. To test either null hypothesis

$$H_0 : p_x - p_y = 0 \quad \text{or} \quad H_0 : p_x - p_y \ge 0$$

against the alternative

$$H_1 : p_x - p_y < 0$$

the decision rule is

Reject H_0 if $T.S. < -z_\alpha$

To test the null hypothesis

$$H_0 : p_x - p_y = 0$$

against the two sided alternative

$$H_1 : p_x - p_y \neq 0$$

e decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$

xample:

company is planning to buy a few machines. Company is considering two pes of machines, but will buy all of the same type. The company selects ıe machine from each type and uses for a few days. A sample of 900 items oduced on machine A showed that 55 of them were defective. A sample of 00 items produced on machine B showed that 41 of them were defective. esting at 1% significance level, can we conclude based on the information om these samples that the proportions of the defective items produced on e two machines are different?

olution:

et p_x be the proportion of all items in all items produced on machine A, ıd p_y be the proportion of all items in all items produced on machine B.

et $\hat{p}_x$ and $\hat{p}_y$ be the corresponding sample proportions. Let x_1 and x_2 be the ımber of defective items in two samples respectively.

Machine A: $n_x = 900$; $x_1 = 55$

Machine B: $n_y = 700$; $y_1 = 41$

he two sample proportions are calculated as follows:

$$\hat{p}_x = \frac{x_1}{n_x} = \frac{55}{900} = 0.0611;$$

$$\hat{p}_y = \frac{y_1}{n_y} = \frac{41}{700} = 0.0586$$

he null and alternative hypotheses are

$H_0 : p_x - p_y = 0$ (the two proportions are equal)

$H_1 : p_x - p_y \neq 0$ (the two proportions are different)

he decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$

Let us check if the sample sizes are large:

$$\left(n_x \cdot \hat{p}_x \cdot \hat{q}_x > 9\right) = 900 \cdot 0.0611 \cdot 0.9389 = 51.63 > 9$$

$$\left(n_y \cdot \hat{p}_y \cdot \hat{q}_y > 9\right) = 700 \cdot 0.05860 \cdot 0.9414 = 38.62 > 9$$

Since the samples are large and independent we apply the no distribution to make a test.

The pooled sample proportion is

$$p_0 = \frac{x_1 + y_1}{n_x + n_y} = \frac{55 + 41}{900 + 700} = 0.06$$

The value of the test statistics is

$$T.S. = z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{p_0(1-p_0)}{n_x} + \frac{p_0(1-p_0)}{n_y}}} = \frac{(0.0611 - 0.0586)}{\sqrt{\frac{0.06 \cdot 0.94}{900} + \frac{0.06 \cdot 0.94}{700}}} = 0.20$$

Let us find the value of $z_{\alpha/2}$.

$\alpha = 0.01; \alpha/2 = 0.005$

$F_z(z_{\alpha/2}) = F_z(z_{0.005}) = 0.995$

$z_{0.005} = 2.58$ and $-z_{0.005} = -2.58$

The value of the test statistic $T.S. = z = 0.2089$ falls in the nonreje region. Consequently, we fail to reject the null hypothesis. As a result can conclude that proportions of defective items produced by two mach are not different.

Exercises

The following information is obtained from two independent samples ected from two populations

$n_x = 750;$ $\hat{p}_x = 0.56;$ $n_y = 600;$ $\hat{p}_y = 0.61;$

st at the 1% significance level if p_x is equal p_y, against the alternative t it is less.

A sample of 600 observations taken from the first population ve $x_1 = 320$. Another sample of 700 observations taken from the second pulation gave $y_1 = 370$. Show the rejection and nonrejection regions on sampling distribution of $\hat{p}_x - \hat{p}_y$ for $H_0 : p_x = p_y$ against the ernative $H_1 : p_x > p_y$ using significance level of 2.5%. Will you reject the ll hypothesis?

According to the statistics, 65% of single women and 80% of single men n cars. Assume that these estimates are based on random samples of 1700 ngles women and 1900 single men. At the 1% significance level, can you nclude that the proportion of single women who own cars is the same as e proportion of men who own cars against the alternative that it is less than e proportion of men who own cars? Also find p-value.

The management of a supermarket wanted to investigate if the percentage men and women who prefer to buy national brand products over the store and products are different. A sample of 600 men shoppers at the mpany's supermarkets showed that 246 of them prefer to buy national and products over the store brand products. Another sample of 700 women oppers showed that 266 of them prefer to buy national brand products over e store brand products. Testing at the 5% significance level, can you nclude that the proportion of all men and all women shoppers at these permarkets who prefer to buy national brand products over the store brand oducts are equal?

A sample of 500 male registered voters showed that 57% of them are in a vor of higher taxes on wealthy people. Another sample of 400 female gistered voters showed that 54% of them are in favor of higher taxes on ealthy people. Test at the 1% significance level if the proportion of all male

voters who are in favor of higher taxes on wealthy people is not diffe from that of female voters against two-sided alternative.

6. A medical researcher investigates if the smoking results in wrinkled s around the eyes. By observing 150 smokers and 250 nonsmokers, researcher finds that 95 of the smokers and 103 of the nonsmokers h prominent wrinkles around their eyes. Do these data substantiate the be that prominent wrinkles around eyes are more prevalent among smokers t nonsmokers? Answer by calculating p-value.

7. In a comparative study of two new drugs, A and B, 120 patients trea with drug A and 150 patients with drug B, and the following results w obtained

	Drug A	Drug B
Cured:	52	88
Not cured:	68	62
Total:	120	150

Do these results demonstrate statement that these two drugs have the sa effect against the alternative that higher cure rate with drug A? Test $\alpha = 0.05$.

8. According to a 2001 survey, 48% of managers "would choose the sa career if they were starting over again". In a similar survey conducted years ago, 60% of managers said that they "would choose the same caree they were starting over again". Assume that the 2001 survey is based o sample of 800 managers and the one done 10 years ago included 6 managers. Test at the 5% significance level if the proportion of all manag who "would choose the same career if they were starting over again" has changed against the alternative that it decreased during the past 10 years.

Answers

1. $T.S. = z = -1.85$; accept H_0; **2.** $T.S. = 0.17$; accept H_0; **3.** $T.S. = -9.8$ reject H_0; can reject H_0 at virtually any level; **4.** $T.S. = z = 1.10$; do not rej H_0; **5.** $T.S. = 0.90$; accept H_0; **6.** $T.S. = 4.28$; p-value=0.0002; claim strongly supported; **7.** $H_0 : p_A = p_B$; $H_1 : p_A < p_B$; $T.S. = -2.52$; reject H_0 **8.** $T.S. = -4.45$; reject H_0.

Chapter 2
Some nonparametric tests

2.1. Introduction

When the methods of statistical inference are based upon the ssumption that the population has a certain probability distribution, such as ie normal, the resulting collection of statistical tests and procedures is eferred to as *parametric methods.* In this chapter we will consider several atistical procedures that do not require knowledge of the form of the robability distribution form which the measurements come. The methods of atistical inference we will study here are called *nonparametric methods.* ince nonparametric methods do not require assumptions about the form of ie population distribution they are often referred to as *distribution free ethods.*

rom this discussion we see that one reason for using nonparametric ethods is that in some situations there is insufficient knowledge about the orm of the population distribution. Thus assumptions necessary for use of arametric tests can not be made.

second reason for using nonparametric methods concerns data easurement. Nonparametric methods are often applied to rank order or reference data. Preference data are the type of data generated when people xpress preference for one product over another, one service over another, tc. Parametric procedures can not be applied with these data, but onparametric ones can.

his chapter presents an introduction to some of the commonly used atistical procedures that can be classified as nonparametric or distribution ee methods. The emphasis will be on the type of problems that can be olved, how the statistical calculations are made, and how appropriate onclusions can be developed to assist management in the decision-making rocess.

2.2. 1. The Sign test for paired or matched samples

In Chapter 1 we considered z and t statistics for testing hypo
about a population mean. For both of them, the sample was selec
random from a *normal* distribution. The question is: How can we con
test of hypothesis when we have a small sample from a nonn
distribution?

The **Sign test** is a relatively simple and most frequently emp
nonparametric procedure for testing hypothesis about the central tende
a nonnormal probability distribution. The sign test is used in stud
identify if consumer preference exists for one of two products.
Suppose that paired or matched samples are taken from a population, an
differences equal to 0 are discharged, leaving n observations. The Sig
can be used to test the null hypothesis that the population median o
differences is 0. Let "+" indicates a positive difference, and "-" indic
negative difference. If the null hypothesis were true, our sequence of "+
"-" differences could be regarded as a random sample from a populati
which the probabilities for "+" and "-" were each 0.5. In that case
observations would constitute a random sample from a binomial popul
in which the probability of "+" was 0.5. Thus, if p denotes the
population proportion of "+"'s in the population (that is, the
proportion of positive differences), the null hypothesis is simply
The Sign test is then based on the fact that the number of pos
observations, S, in the sample has a binomial distribution (with p
under the null hypothesis).

$$H_0 : p = 0.5$$

Sign test for paired samples

Suppose that paired random samples are taken from a population and
differences equal to 0 are ignored. Calculate the difference for each pair
record the sign of this difference. The Sign test is used to test:

$$H_0 : p = 0.5$$

where p-is the proportion of nonzero observations in the population tha
positive. The test statistic S for the Sign test for paired samples is simply

S = the number of pairs with **positive** difference

S has a binomial distribution with $p = 0.5$ and n = the number of
ero differences.
determining the null and alternative hypotheses and finding a test
tic, the next step is to determine the p-value and to draw conclusions
on a decision rule.

Determining p- value for a Sign test

p-value for a Sign test is found using the binomial distribution with
ıe number of nonzero differences, S = the number of pairs with **positive**
ences and $p = 0.5$.

r right tailed test,

$$H_1 : p > 0.5, \quad \textbf{p-value} = P(x \geq S)$$

r left tailed test,

$$H_1 : p < 0.5, \quad \textbf{p-value} = P(x \leq S)$$

r two tailed test,

$$H_1 : p \neq 0.5, \quad 2 \cdot (p - \text{value})$$

iple:

study 8 individuals were asked to rate on a scale from 1 to 10 the test
oducts of two brands: Brand A and Brand B. The scores of the test
arison are shown in the following table

N	Brand A	Brand B
1	5	7
2	3	10
3	4	8
4	9	6
5	8	8
6	5	7
7	6	5
8	9	6

e data indicate an overall tendency to prefer the Brand B to the
l A?

Solution:
First of all, let us calculate differences

N	Brand A	Brand B	Difference (A-B)	Sign of differe
1	5	7	-2	-
2	3	10	-7	-
3	4	8	-4	-
4	9	6	3	+
5	8	8	0	0
6	5	7	-2	-
7	6	5	1	+
8	9	6	3	+

We are discarding those who rated the brands equally. In this examp
values for fifth person is omitted in future analysis and the effective s
size is reduced to $n = 7$. The only sample information on which our
based is that **three** of the seven tasters preferred the brand A. Hen
value of the Sign test is $S = 3$.
Let p-denotes the true proportion of "+"s in the population. Then th
hypothesis is

$H_0 : p = 0.5$ There is no overall tendency to prefer one Brand to the oth

A one tailed test is used to determine if there is an overall tendency to
the Brand B to the Brand A. The alternative hypothesis is that i
population, the majority of preferences are for Brand B. The alter
hypothesis is expressed as

$H_0 : p < 0.5$ Majority prefer the Brand B

The next step is the finding the p-value. If we denote by $P(x)$ the proba
of observing x "successes" ("+"s) in $n = 7$ binomial trials, each
probability of success 0.5, then the cumulative binomial probabil
observing tree or fewer "+"s can be obtained using binomial formula

$$p - \text{value} = P(x \le 3) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$
$$= C_0^7 (0.5)^0 (0.5)^7 + C_1^7 (0.5)^1 (0.5)^6 + C_2^7 (0.5)^2 (0.5)^5 + C_3^7 (0.5)^3 (0.5)^4$$
$$= 0.0078 + 0.0547 + 0.1641 + 0.2734 = 0.5000$$

this example p- value is 50%. We are unable to reject the null hypothesis
conclude that data is not sufficient to suggest that population have a
erence for Brand B. Since the p-value is the smallest significance level at
h the null hypothesis can be rejected, for this example, the null
othesis can be rejected at 50% or higher. It is unlikely that one would be
ing to accept such a high significance level. Again, we conclude that the
is not statistically significant to recommend that Brand B is preferred by
ority.

2.2.2. The sign test: Normal approximation (Large samples)

consequence of the central limit theorem, the normal distribution can be
to approximate the binomial distribution if the sample size is large
erts differ on the exact definition of "large". We suggest that the normal
oximation is acceptable if the sample size exceeds 20. With large n, the
mial distribution with $p = 0.5$ is close to the normal distribution with
n $n/2$ and standard deviation $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{0.25 \cdot n} = \sqrt{n/4}$.
test statistic is

$$T.S. = z = \frac{S - n/2}{\sigma} = \frac{S - n/2}{\sqrt{n/4}}$$

re

S-is the number of positive signs,
n-is the number of nonzero sample observations.
null hypothesis to be tested is that the proportion p- of nonzero
rvations in the population that are positive is 0.5; that is
o test either null hypothesis

$H_0 : p = 0.5$ against the alternative
$H_1 : p > 0.5$

decision rule is

Reject H_0 if $T.S. > z_\alpha$

o test either null hypothesis

$H_0 : p = 0.5$ against the alternative
$H_1 : p < 0.5$

decision rule is

Reject H_0 if $T.S. < -z_\alpha$

3. To test the null hypothesis

$H_0 : p = 0.5$ against the two sided alternative

$H_1 : p \neq 0.5$

the decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$

Example:

In a TV commercial, filmed live, 100 persons tested two brands of c say brand A and brand B and each selected their favorite. 56 p preferred coffee of brand A, 40 preferred coffee of brand B, and expressed on preference. Test at the 5% significance level the hypothesis that for the population, there is no overall preference for co brand A over the brand B.

Solution:

From the information given above we obtain that

$S = 56$; $n = 96$

To test if there is no overall preference in this population for one bra coffee over the other, the hypotheses are

$H_0 : p = 0.5$ (People have no preference for either brand of cof

$H_1 : p \neq 0.5$ (People have preference for one brand of coffee)

The decision rule is

Reject H_0 if $\dfrac{S - n/2}{\sqrt{n/4}} < -z_{\alpha/2}$ or $\dfrac{S - n/2}{\sqrt{n/4}} > z_{\alpha}$

The value of the test statistic is

$$\frac{S - n/2}{\sqrt{n/4}} = \frac{56 - 96/2}{\sqrt{96/4}} = -1.633$$

$\alpha = 0.05$; $z_{\alpha/2} = 1.96$ and $-z_{\alpha/2} = -1.96$.

Since -1.633 is not less than -1.96 we fail to reject the null hypothesis at 5% significance level we accept that there is no preference for either of coffee.

From the standard normal distribution it follows that the approximate

$p -$ value $= 0.1032$ or 10.32%.

nce the null hypothesis can be rejected at all significance levels greater n 10.32%.

Exercises

Two computer specialists estimated the amount of computer memory (in abytes) required by five different offices.

Office	Specialist A	Specialist B
1	4.2	6.1
2	6.3	6.7
3	3.1	3.0
4	2.2	2.9
5	7.2	10.9

e the Sign test to test the null hypothesis that two specialists estimations the same against the alternative that specialist B estimates higher than cialist A.

In a test of two chocolate chip cookie recipes, 13 out of 18 subjects ored recipe A. Using the sign test, find the significance probability when states that recipe A is preferable.

A firm attempting to determine if a difference exists in two manufacturing thods. A sample of 10 workers was selected, and each worker completed production task using each of the two production methods.

Worker	Method 1(minutes)	Method 2(minutes)
1	10.2	9.5
2	9.6	9.8
3	9.2	8.8
4	10.6	10.1
5	9.9	10.3
6	10.2	9.3
7	10.6	10.5
8	10.0	10.0
9	10.7	10.2
10	10.9	10.2

e the sign test and perform the null hypothesis that there is no overall eference for one method over the other.

4. A social researcher interviews 25 newly married couples. Each hu and wife are independently asked the question: "How many children you like to have?" The following data are obtained

Couple	Answer of Husband	Answer of Wife	Couple	Answer of Husband	Answer of Wife
1	3	2	14	2	1
2	2	2	15	3	2
3	2	1	16	2	2
4	2	3	17	0	0
5	5	1	18	1	2
6	0	1	19	2	1
7	0	2	20	3	2
8	1	3	21	4	3
9	2	2	22	3	1
10	3	1	23	0	0
11	4	2	24	2	3
12	1	2	25	2	2
13	3	3			

Use the Sign test with $\alpha = 0.05$ to test against two sided alternative th hypothesis that, for the population of families no difference in opi between husbands and wives.

5. A random sample of 80 sale managers was asked to predict whether year's sale would be higher than, lower than, or about the same as i current year. The results are shown below. Test the null hypothesis th opinion of managers is evenly divided on the question against a two alternative.

Prediction	Number
Higher	37
Lower	28
About the same	15

6. Of a random sample of 120 university students, 67 expected to achi better GPA than last year, 48 expected a lower GPA than last year, and 5 expected about the same GPA. Do these data present strong evidence for population of students they are divided evenly on the expectat against the alternative that more expect a lower GPA compared with year?

a random sample of 150 university instructors, 62 believed that nt's skills in solving problems increased over the last decade, 54 ed these skills had deteriorated and 4 saw no change. Evaluate the gth of the sample evidence suggesting that, for all university instructors, ers are divided evenly on the issue against the alternative that more ers believe that student's skills in solving problems have improved.

a coffee taste test 48 individuals stated a preference for one of two well- n brands. Results showed 28 favoring brand A, 16 favoring brand B, undecided. Use the sign test with $\alpha = 0.10$ to test the null hypothesis here is no difference in the preferences for the two brands of coffee st a two sided alternative.

Answers

value = 0.1874 or 18.74%; **2.** p-value = 0.0482 or 4.82%;
value = 0.1798 or 17.98%; **4.** $T.S. = 0.94$; accept H_0 ; **5.** $T.S. = 3.39$;
ue = 0.06%; **6.** $T.S. = 1.77$; p-value = 3.84%; **7.** $T.S. = 0.74$;
ue = 22.96%; **8.** $T.S. = 1.81$; reject H_0 .

2.3. The Wilcoxon signed test

2.3.1. The Wilcoxon signed test for paired samples (small sample size)

One disadvantage of the sign test is that it takes into accoun very limited amount information-namely, the signs of the differenc **Wilcoxon signed rank test** provides a method to use information ab magnitude of the differences between matched pairs. It is still a distr free test. Like many nonparametric test, it is based on ranks.

Table 2.1

Worker	Method I	Method II	Difference	Absolute value of difference	Rank (+)	Ra (-
1	10.2	9.5	0.7	0.7	8	
2	9.6	9.8	-0.2	0.2		
3	9.2	8.8	0.4	0.4	3.5	
4	10.6	10.1	0.5	0.5	5.5	
5	9.9	10.3	-0.4	0.4		3
6	10.2	9.3	0.9	0.9	10	
7	10.6	10.5	0.1	0.1	1	
8	10.0	10.0	0	0	--	
9	11.2	10.6	0.6	0.6	7	
10	10.7	10.2	0.5	0.5	5.5	
11	10.6	9.8	0.8	0.8	9	
					49.5	5

To demonstrate the use of the Wilcoxon signed ranked test let us cons manufacturing firm that is attempting to determine if a difference ex two production methods. A sample of 11 workers was selected, and worker completed the production task using each of the two prod methods. Each worker in the sample provides a pair of observatio shown in Table2.1. Table 2.1 also provides the difference in the comp times. A positive value indicates that Method I require more time, negative value indicates that Method II require more time. The stat question is whether or not the data indicate that the methods are signific

rent in terms of completion times. Thus the null and alternative othesis can be written as

H_0 : The two populations of task completion times are identical

H_1 : The two populations of task completion times are not identical

with the sign test, we ignore any difference of "0", so sample size in mple above is reduced to $n=10$. The nonzero absolute differences are ranked in ascending order of magnitude. That is, the smallest absolute e 0.1 is given a rank of "1". If two or more values are equal, they are ned the average of the next available ranks. In example above, absolute e of difference-0.4 occurs twice. The rink assigned to them is therefore verage of ranks 3 and 4-that is 3.5. The next absolute value-0.5 occurs e. The rank assigned to them is therefore the average of ranks 5 and 6- is 5.5. The next absolute value is assigned rank 7, and so on.

ranks for positive and negative differences are summed separately. The ler of these sums is the Wilcoxon Signed Rank Statistic *T.S.*

ce *T.S.*=5.5.

will now suppose that the population distribution of the paired rences is symmetric. The null hypothesis to be tested is that the center of distribution is 0. In example above, we are assuming that differences in ask completion times have a symmetric distribution, and we want to test her that distribution is centered on 0-that is no difference between task pletion times.

ff points for the distribution of this random variable are given in ndix (Table 4) for tests against a one sided alternative that the lation distribution of the paired differences is specified either to be red on some number bigger than 0 or to be centered on some number han 0. For sample size, n, the table shows, for selected probabilities α, umber T_α such that $P(T<T_\alpha)=\alpha$. In other words, the null hypothesis ected if *T.S.* is less than or equal to the corresponding number in the 4.

ample above, $T.S.=5.5$. For $n=10$ we find that the null hypothesis will ected for any significance level greater than $\alpha=0.005$.

Steps in the Wilcoxon Signed Rank test for paired samples

1. Calculate the differences
2. Discard (ignore) any difference of "0"
3. Find absolute value of differences
4. Rank the absolute value of differences in ascending ord
magnitude. Rank positive and negative differences in two di
columns
4. Assign tied absolute differences (if any ties) the average of the
they would receive if they were unequal but occurred in successive
5. Find separately sum of ranks of positive and negative difference
6. The smaller of the two sums is the Wilcoxon Signed Rank Statis
7. Reject the null hypothesis if the value of the test statistic is less t
equal to the value in Appendix table 3.

2.3.2. The Wilcoxon signed test for paired samples (large sample size)

When the number of n nonzero differences in the sample is large (n >2
normal distribution provides a good approximation to the distribution
Wilcoxon statistic T under the null hypothesis that the population diffe
are centered on 0.
Let T denote the smaller of the rank sums.
With increasing sample size of n (n>20) nonzero differences, th
hypothesis is that the population differences are centered on 0, Wil
Signed Rank test has mean and variance given by

$$E(T) = \mu_T = \frac{n(n+1)}{4}$$

and

$$Var(T) = \sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$$

For large n, the distribution of the random variable, Z, is approxi
standard normal where

$$Z = \frac{T - \mu_T}{\sigma_T}$$

ie number of nonzero differences is large and T is the observed value of
Wilcoxon Signed test statistic, then the following tests have significance
:l α,
the alternative hypothesis is one sided, reject the null hypothesis if

$$\frac{T - \mu_T}{\sigma_T} < -z_\alpha$$

f the alternative hypothesis is two sided, reject the null hypothesis if

$$\frac{T - \mu_T}{\sigma_T} < -z_{\alpha/2}$$

mple:
andom sample of 38 students who had just completed courses in statistics
accounting was asked to rate each in terms of level of interest, on a scale
n one (very uninteresting) to ten (very interesting). The 38 differences in
pairs of ratings were calculated and the absolute differences ranked. The
ller of the rank sums, which was for those finding accounting the more
resting, was 278. Test at 5 % significance level the null hypothesis that
population of students would rate these courses equally against the
rnative that the statistics course is viewed as the more interesting.
o find the p-value.
ution:
m the given information

$$n = 38; \qquad T = 278$$

mean and variance of the Wilcoxon statistic are

$$\mu_T = \frac{n(n+1)}{4} = \frac{38 \cdot (38+1)}{4} = 370.5$$

$$\sigma_T^2 = \frac{n(n+1)(2n+1)}{24} = \frac{38 \cdot 39 \cdot 77}{24} = 4754.75$$

he standard deviation is

$$\sigma_T = 68.95$$

ording to the condition, the null and alternative hypothesis can be written

H_0 : both courses rated equally interesting
H_1 : statistics course rated more interesting
is the observed value of the test statistic, the null hypothesis is rejected
inst one sided alternative if

$$\frac{T-\mu_T}{\sigma_T} < -z_\alpha$$

Here, the value of T is $T = 278$ and the value of test statistic is

$$\frac{T-\mu_T}{\sigma_T} = \frac{278-370.5}{68.95} = -1.34$$

$\alpha = 0.05$; $F_z(z_{0.05}) = 0.95$; $z_{0.05} = 1.65$; $-z_{0.05} = -1.6$

Since -1.34 is not less than -1.65 we fail to reject H_0, and accept it.

The value of α corresponding to $z_\alpha = -1.34$ is, from Table 1 Appendix, $(1 - 0.9099) = 0.0901$. Then the null hypothesis can be reje all significance levels greater than 9.01%. The data contain modest evi suggesting that statistics course is more interesting.

Exercises

1. Two critics rate the service at six award winning restaurants continuous 0 to 10 scale. Apply Wilcoxon signed rank test with $\alpha =$ there is no difference between the critics' ratings?

Restaurant	Critic 1	Critic 2
1	6.2	8.4
2	5.3	5.8
3	7.5	7.1
4	7.4	7.0
5	4.3	5.1
6	9.8	9.9

2. Two computer specialists estimated the amount of computer memory (in gigabytes) required by five different offices

Office	Specialist A	Specialist B
1	5.7	6.1
2	6.4	6.8
3	3.2	3.1
4	2.0	2.9
5	8.1	12.3

Apply Wilcoxon signed rank test with $\alpha = 0.05$ to test the null hypo that there is no difference between estimations against a two alternative.

. Twelve customers were asked to estimate the selling price of two models f refrigerators. The estimates of selling price provided by the customers are own below:

Customer	Model A	Model B
1	$650	$900
2	760	720
3	740	690
4	700	850
5	590	920
6	620	800
7	700	890
8	690	920
9	900	1000
10	500	690
11	610	700
12	720	700

se these data and test at the 0.05 level of significance to determine if there no difference in the customers' perception of selling price of the two odels.

. A certain brand of microwave oven was priced at 12 stores in two fferent cities.

hese data are presented below:

District A	District B
18 500	16 700
16 000	20 500
12 000	23 000
20 000	17 500
19 000	22000
17 000	21 000
16 500	21 500
19 000	19 500
15 500	17 000
16 000	23 000
17 500	21 000
18 000	22 000

se a 0.05 level of significance and apply the Wicoxon signed rank test to st whether o not prices for the microwave oven are the same in the two ties.

5. The company is interested in the impact of the newly introduced management program on job satisfaction of workers. A random samp workers was asked to assess level of satisfaction on a scale from 1 to month before the program. These same sample members were asked t this assessment again two month after the introduction of the progra 34 differences in the pairs of ratings were calculated and absolute diff ranked. The smaller of the rank sums, which was for those more s before the introduction of the program, was 178. What can be co from these findings?

6. A random sample of 90 members was taken. Each sample mem asked to assess the amounts of time in a month spent watching TV amounts of time in a month spent reading. The 90 differences in time were then calculated and their absolute differences ranked. The sm the of the rank sums, which was for watching TV, was 1680. Test t hypothesis that the population amounts of time spent on watching reading divides equally against the alternative that watching TV take amounts of time.

7. Suppose you wish to test hypothesis that two treatments, A and equivalent against the alternative that the responses for A tend to b than those of B. If the number of pairs equals 25, and smaller of the the absolute differences is 273, then what would you decide? Use α then find p-value for the test and interpret it.

8. An experiment was conducted to compare two print types, A an determine whether type A is easier to read. A sample of 22 perso given the same material to read. First they read the material printed w A, then read the same material printed with type B. The times neces each person to read the materials (in seconds) were

Type A: 95;122;101;99;108;122;135;127;119;127;99;98;97;96;112;9 116; 111;117;102;103

Type B: 110;102;115;112;120;117;119;127;137;119;99;100;102;103; 99;89;97;112;116; 178; 94.

Do the data provide sufficient evidence to indicate that print type A an type B are the same for reading against the alternative that print typ easier to read? Test using $\alpha = 0.05$.

Answers

$S. = 7$; accept H_0; **2.** $T.S. = 1$; reject H_0; **3.** $T.S. = 6$; reject H_0; $S. = 3$; reject H_0 virtually at any levels; **5.** $T.S. = -2.04$; p-value = 4.12%; $T.S. = -1.48$; reject H_0 at levels higher than 6.94%; **7.** $T.S. = 2.97$; accept at any levels; **8.** $T.S. = -0.71$; reject H_0.

2.4. The Mann-Whitney test

Suppose two independent random samples are to be used to compare populations. We may be unwilling to make assumptions about the form ne underlying population probability distributions or we may be unable to in exact values of the sample measurements. If the data can be ranked in r of magnitude for either of these situations, the *Mann-Whitney* test hetimes called *Mann-Whitney U test)* can be used to test the hypothesis the probability distributions associated with the two populations are ivalent.

ume that apart from any possible differences in central location, that the population distributions are identical. Suppose that n_1 observations are lable from the first population and n_2 observations from the second ulation. The two samples are pooled and the observations are ranked in nding order, with ties assigned the average of the next available ranks. R_1 denote the sum of the ranks from the first population. The Mann- itney statistic is

$$U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$$

sting the null hypothesis that the central locations of the two population ibutions are the same, we assume that the two population distributions identical. It can be shown that if the null hypothesis is true, the random able U has mean

$$E(U) = \mu_U = \frac{n_1 \cdot n_2}{2}$$

variance

$$Var(U) = \sigma_U^2 = \frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}$$

Then for large sample sizes (both at least 10), the distribution of th variable,

$$Z = \frac{U - \mu_U}{\sigma_U}$$

is well approximated by the standard normal distribution.

Decision rules for the Mann-Whitney test

Suppose that two population distributions are identical, a any possible differences in central location. In testing the null hypo two population distributions have the same central location, the test have significance level α :

H_0 : Two population distributions have the same central location

1. If the alternative hypothesis is one sided hypothesis that the lo population 1 is higher than the location of population 2, the decision

$$\text{Reject } H_0 \text{ if } \frac{U - \mu_U}{\sigma_U} < -z_\alpha$$

2. If the alternative hypothesis is one sided hypothesis that the lo population 1 is lower than the location of population 2, the decision

$$\text{Reject } H_0 \text{ if } \frac{U - \mu_U}{\sigma_U} > z_\alpha$$

3. If the alternative hypothesis is two sided hypothesis that population distributions differ, the decision rule is

$$\text{Reject } H_0 \text{ if } \frac{U - \mu_U}{\sigma_U} < -z_{\alpha/2} \text{ or } \frac{U - \mu_U}{\sigma_U} > z_{\alpha/2}$$

Example:

Let us demonstrate the methodology of the Mann-Whitney test b conduct a test on the population of account balances at two branche Bank. Data collected from two independent simple random sam from each branch, are shown in Table 2.2.

Table2.2

Branch 1		Branch 2	
Sampled Account	Account balance	Sampled account	Account balance
1	1 095	1	885
2	955	2	850
3	1 200	3	915
4	1 195	4	950
5	925	5	800
6	950	6	750
7	805	7	865
8	945	8	1 000
9	875	9	1 050
10	1 055	10	935
11	1 025		
12	975		

first step in the Mann- Whitney test is to rank the *combined* (pooled) from the two samples from low to high. Using the combined set of 22 ervations shown in Table 2.2, the lowest value of \$750(item 6 of ple2) is ranked number 1. Continuing the ranking, we have

Account balance	Item	Rank
750	6 of sample 2	1
800	5 of sample 2	2
805	7 of sample 1	3
.....	.....	...
1 195	4 of sample1	21
1 200	3 of sample 1	22

1 6 of sample 1 and item 4 of sample 2 both have the same account nce, \$950. We could give one of these items a rank 12 and the other a k 13, but this could lead to an erroneous conclusion. In order to avoid this iculty the usual treatment for tied data values is to assign each value the equal to the average of the ranks associated with the tied items. Thus tied observations of \$950 are both assigned ranks of 12.5. Table 2.3 ws the entire data set with the rank of each observation.

Table

Branch 1			Branch 2	
Sampled Account	Account balance	Rank	Sampled account	Account balance
1	1 095	20	1	885
2	955	14	2	850
3	1 200	22	3	915
4	1 195	21	4	950
5	925	9	5	800
6	950	12.5	6	750
7	805	3	7	865
8	945	11	8	1 000
9	875	6	9	1 050
10	1 055	19	10	935
11	1 025	17		
12	975	15		
Sum of ranks		169.5		

The next step in the Mann-Whitney test is to sum the ranks for each s
These sums are shown in Table 2.3. The test procedure can be base
the sum of the ranks for either sample. In the following discussion
the sum of the ranks for the sample from branch 1. We will denote th
by R_1. Thus, in our example $R_1 = 169.5$.

The value observed for the Mann-Whitney test is

$$U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1 = 12 \cdot 10 + \frac{12 \cdot 13}{2} - 169.5 = 28.5$$

Since two samples are selected from identical populations and n_1 a
each is 10 or greater, the sampling distribution of U can be approxima
a normal distribution with mean

$$E(U) = \mu_U = \frac{n_1 \cdot n_2}{2} = \frac{12 \cdot 10}{2} = 60$$

and variance

$$Var(U) = \sigma_U^2 = \frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12} = \frac{12 \cdot 10 \cdot 23}{12} = 230$$

ose that we want to test the null hypothesis that the central locations of distributions of account balance are identical against the two-sided native for $\alpha = 0.05$. The decision rule is to reject the null hypothesis if

$$\frac{U - \mu_U}{\sigma_U} < -z_{\alpha/2} \text{ or } \frac{U - \mu_U}{\sigma_U} > z_{\alpha/2}$$

$$\frac{U - \mu_U}{\sigma_U} = \frac{28.5 - 60}{\sqrt{230}} = -2.08$$

$= z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$

e -2.08 is less than -1.96, we reject the null hypothesis that two ilation account balances are identical. Thus we conclude that two ilations are not identical. The probability distribution of account nces at branch 1 is not the same as that at branch 2.

, from Table1 of the Appendix, the value of $\alpha/2$ corresponding to a (-2.08) is 0.0188, so the corresponding α is 0.0376

$$p - \text{value} = 2 \cdot (1 - F_z(\text{test statistics})) = 2(1 - 0.9812) = 0.0376$$

null hypothesis will be rejected for any significance level higher than %. Thus, these data do not contain strong evidence against the thesis that the central locations of accounts at two branches are the . There is very strong support that two branches account balances are identical.

Exercises

1. Starting salaries were recorded for ten recent business administ graduates at each of two well-known universities. Use $\alpha = 0.1$ and t the difference in the starting salaries from the two universities is zero the alternative that starting salaries are higher for the university A.

University A		University B	
Student	Monthly salary ($)	Student	Monthly salar
1	890	1	1 000
2	950	2	1 020
3	1 200	3	1 140
4	1 150	4	1 000
5	1 300	5	975
6	1 350	6	925
7	990	7	900
8	1 050	8	1 025
9	1 400	9	1 075
10	1 450	10	930

2. The following data show product weights for items produced o production lines

Line 1: 13.6; 13.8; 14.0; 13.9; 13.4; 13.2; 13.3; 13.6; 12.9; 14.4

Line 2: 13.7; 14.1; 14.2; 14.0; 14.6; 13.5; 14.4; 14.8; 14.5; 14.3; 15.0; 1

Test that the difference between the product weights for the two lines i against the alternative that product weights of second line is higher.

Use $\alpha = 0.10$. Also find p-value.

3. A random sample of 14 male students and an independent random sa of 16 female students were asked to write essays at the conclusion writing course. Their grades were recorded below:

Male: 75; 80; 60; 80; 95; 100; 65; 70; 75; 60; 50; 55; 90; 95

Female: 85; 70; 90; 100; 95; 67; 50; 50; 67; 83; 78; 62; 43; 97; 89; 73

Test the 5% significance level null hypothesis that, in the aggregate the and female students are equally ranked, against a two-sided alternative. find p-value.

4. For a random sample of 12 management department gradates an economics department graduates were asked their starting salaries. T salaries were then ranked from 1 to 26. The following rankings resulted

Management: 2; 6; 7; 1; 11; 20; 8; 14; 21; 12; 4; 26

Economics: 13; 3; 17; 25; 5; 9; 10; 24; 15; 23; 16; 22; 18; 1

Analyze the data using the Mann-Whitney test, and comment on the resu

Starting salaries of graduates from two leading universities were pared. Independent random samples of 40 from each university were n, and the 80 starting salaries were pooled and ranked. The sum of the s for students from one of these universities was 1450. Test the null othesis that the central locations of the population distributions are tical against two sided alternative.

A stock market analyst produced at the beginning of the year a list of ks to buy and another list of stocks to sell. For a random sample of ten ks from the "buy list", percentage returns over the year were as follows:

.6; 5.2; 12.8; 16.2; 10.6; 4.3; 3.1; 11.7; 13.9; 11.3

an independent random sample of ten stocks from the "sell list", entage returns over the year were as follows:

.6; 6.1; 9.9; 11.3; 2.3; 3.9; -2.3; 1.3; 7.9; 10.8

$\alpha = 0.05$ use the Mann-Whitney test to interpret these data. Also find interpret p-value.

Answers

.S. = ; reject H_0; **2.** $T.S.$ = ; reject H_0; p-value = 0.3%;

T.S. = ; accept H_0; **4.** $T.S.$ = ; p- value =12.36%; H_0 will be rejected ll levels higher than 12.36%; **5.** $T.S.$ = ; p-value = 0.101; H_0 will be ted at any level higher than 10.1%; **6.** $T.S.$ = ; reject H_0 at 5%; alue = 2.58%.

Chapter 3

Simple linear regression

3.1. Introduction

In day-to-day decisions-making situations, businessperson economists frequently draw conclusions and make recommendations on the relationship between two variables. For example, a mai manager may project sales volume based upon observed relations be advertising expenditures and sales volume. Although in some instanc manager will rely on his or her intuition as to how the variables are r the safest approach, by far, is to collect data on the two variables an evaluate their relationship statistically. These relationships are exp mathematically as

$$y = f(x)$$

where the function may follow linear and nonlinear forms.

3.2. The scatter diagram

As a first step in determining if a relationship exists between two var we could plot or graph the available data for the two variables. Suppos a sales manager has recorded containing data on annual sales and ye experience. The information is given in the following table:

Salesperson	1	2	3	4	5	6	7	8	9
Years of experience	1	3	4	4	6	8	10	10	11
Annual sales ($1000's)	80	97	92	102	103	111	119	123	117

Let us plot these data on a graph with years of selling experience o horizontal axis and annual sales on the vertical axis. We now have a sc **diagram.** It is given this name because the plotted points are "scatt over the graph or diagram. The scatter diagram for these data is sho Figure 3.1.

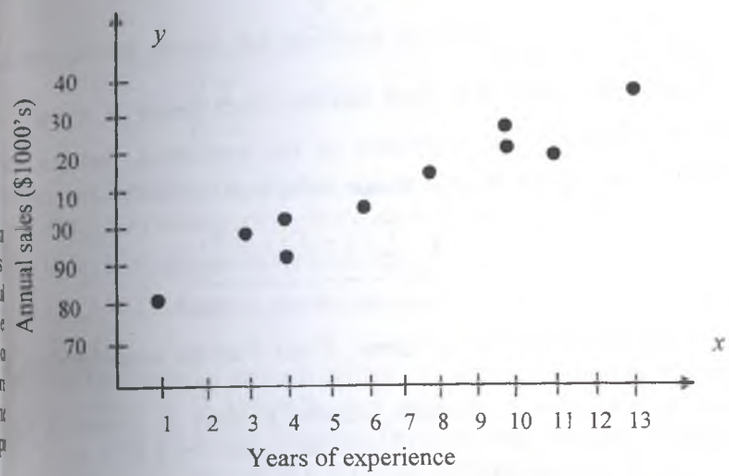

Fig.3.1. Scatter diagram of annual sales and years of experience

regression analysis statisticians commonly will classify a variable as an **lependent** or a **dependent** variable. The classification is used to indicate ich variable is doing the predicting or explaining (independent variable) which variable is being predicted or explained (dependent variable). In example, the years of selling experience is referred to as the independent iable. It is used to predict the sales volume, or dependent variable.

es the scatter diagram in Fig. 3.1 allow us to draw conclusions?

gives us an overview of the data. It indicates that in this case there is a d chance that the variables are related. In fact, it appears that the tionship between these two variables may be approximated by a straight or linear function.

3.3. Correlation analysis

We will introduce some statistical measures that provide greater p
for describing relationships.

Let X and Y be a pair of random variables, with means μ_x and μ
variances σ_x^2 and σ_y^2. As a measure of the association betwee
variables, we introduced ***the covariance***, defined as

$$Cov(x,y) = S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

where x_i and y_i are the observed values, $\bar{X}$ and $\bar{Y}$ are the sample mea
n is the sample size.

A positive value of the covariance indicates a direct or increasing
relationships and a negative value of covariance indicates a decreasin
relationship. Positive association indicates that the high values of X
be associated with high values of Y and low X with low Y. When th
negative association, so that high values of X are associated with low
of Y and low X with high Y, the covariance is negative. If there is n
association between X and Y, their covariance is 0.

Another measure of the relationship between two variables is the **corr
coefficient**. In this section we will consider the simple linear correlat
short linear correlation, which measures the strength of the linear ass
between two variables.

Definition:

The simple linear correlation, denoted by, r_{xy}, measures the strength
linear relationship between two variables for a sample and is calculate

$$r_{xy} = \frac{Cov(x,y)}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3.1)$$

An equivalent expression is

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x}\,\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n \cdot (\bar{x})^2\right) \cdot \left(\sum_{i=1}^{n} y_i^2 - n \cdot (\bar{y})^2\right)}} \quad (3.2)$$

e sample correlation coefficient ranges from -1 to $+1$ with,

a) $r_{xy} = +1$ indicates a perfect positive linear relationship;

b) $r_{xy} = 0$ indicates no relationships between X and Y

c) $r_{xy} = -1$ indicates a perfect decreasing linear relationship between X and Y.

ositive correlations indicate positive or increasing linear relationship
values closer to +1, indicating data points closer to a straight line, and
r to 0, indicating greater deviations from a straight line.

egative correlations indicate negative or decreasing linear relationship
values closer to -1, indicating data points closer to a straight line, and
r to 0, indicating greater deviations from a straight line.

=1, it is said to be a case of perfect positive linear correlation. In such
, all points in the scatter diagram lie on a straight line that slopes
rd from left to right, if $r_{xy} = -1$, the correlation is said to be a perfect
tive linear correlation. In this case, all points in a scatter diagram fall on
ight line that slopes downward from left to right.

If the correlation between two variables is positive and close to 1,
ay that the variables have a *strong positive linear correlation*. If the
lation between two variables is positive but close to 0, then the
bles have a *weak positive linear correlation*. On the other hand, if the
lation between two variables is negative and close to 1, then the
bles are said to have a *strong negative linear correlation*. Also, if the
lation between two variables is negative and close to 0, there exists a
negative linear correlation between the variables.

ple:

is interested in the relationship between food expenditure and
late the sample correlation coefficient for the data recorded on
ly incomes and food expenditure of seven households.

Household	1	2	3	4	5	6
Income (100's of $)	35	49	21	39	15	28
Food expenditure (100's of $)	9	15	7	11	5	8

Solution: The sample means are

$$\bar{x} = \frac{\sum x_i}{n} = \frac{212}{7} = 30.29; \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{64}{7} = 9.14$$

The sample correlation coefficient can be calculated either by (3.1) or
It is more convenient to use (3.2) to calculate correlation coefficient:
Necessary calculations of the sample correlation for the data are set
the following table 3.1

Table 3.1

Household	Income (x_i)	Food expenditure (y_i)	$x_i \cdot y_i$	x_i^2	y_i^2
1	35	9	315	1225	81
2	49	15	735	2401	225
3	21	7	147	441	49
4	39	11	429	1521	121
5	15	5	75	225	25
6	28	8	224	784	64
7	25	9	225	625	81
Sums	212	64	2150	7222	646

Hence, the sample correlation is:

$$r_{xy} = \frac{\sum_{I=1}^{n} x_i y_i - n \cdot \bar{x}\,\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n \cdot (\bar{x})^2\right) \cdot \left(\sum_{i=1}^{n} y_i^2 - n \cdot (\bar{y})^2\right)}} =$$

$$= \frac{2150 - 7 \cdot (30.29) \cdot (9.14)}{\sqrt{(7222 - 7 \cdot (30.29)^2) \cdot (646 - 7 \cdot (9.14)^2)}} = \frac{212.05}{221.25} = 0.96$$

sample correlation, 0.96, indicates very strong positive relationships
een monthly income and food expenditure. The high value of monthly
ne tends to be associated with the higher value of food expenditure.

3.3.1. Hypothesis test for correlation

sample correlation coefficient r_{xy} is useful as a descriptive measure of
trength of linear association in a sample. We can also use the correlation
ficient to test the null hypothesis that there is no linear association in the
lation between a pair of random variables; that is

$$H_0 : \rho = 0$$

can show that when the null hypothesis is true and the random variable
a joint normal distribution then the random variable

$$t = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}$$

ws a Student's t distribution with (n-2) degrees of freedom. The
wing tests of the null hypothesis

$$H_0 : \rho = 0$$

a significance level of α :

test H_0 against the alternative

$$H_1 : \rho > 0$$

ecision rule is

reject H_0 if $T.S. > t_{n-2,\alpha}$

test H_0 against the alternative

$$H_1 : \rho < 0$$

ecision rule is

reject H_0 if $T.S. < -t_{n-2,\alpha}$

test H_0 against the two sided alternative

$$H_1 : \rho \neq 0$$

ecision rule is

reject H_0 if $T.S. > t_{n-2,\alpha/2}$ or $T.S. < -t_{n-2,\alpha/2}$

where $T.S. = \dfrac{r_{xy}\cdot\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$, and $t_{n-2,\alpha}$ is the number for which

$$P(t_{n-2} > t_{n-2,\alpha}) = 2$$

where the random variable t_{n-2} follows a Student's t distribution wit degrees of freedom.

Example:

A sample data set produced the following information

$n = 10$; $\sum x_i = 66$; $\sum y_i = 588$; $\sum x_i y_i = 22$

$\sum x_i^2 = 396$; and $\sum y_i^2 = 58734$

Find the sample correlation, and test against a two sided alternative t hypothesis that the population correlation is 0. Take $\alpha = 0.05$.

Solution: Denoting by ρ the population correlation, we want t

$$H_0 : \rho = 0$$

against the two sided alternative

$$H_1 : \rho \neq 0$$

the decision rule is

reject H_0 if $T.S. > t_{n-2,\alpha/2}$ or $T.S. < -t_{n-2,\alpha/2}$

Firstly, let us find the value of sample correlation coefficient

$$r_{xy} = \frac{2244 - 12\cdot 5.5\cdot 49}{\sqrt{(396 - 12\cdot(5.5)^2)\cdot(58734 - 12\cdot(49)^2)}} = \frac{990}{993.69} = -0.996$$

The value of the test statistic is

$$T.S. = \frac{r_{xy}\cdot\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{-0.996\cdot\sqrt{10-2}}{\sqrt{1-(-0.996)^2}} = \frac{2.817}{0.0894} = -31.5$$

$$t_{n-2,\alpha/2} = t_{8,0.025} = 2.306 \text{ and } -t_{8,0.025} = -2.306$$

Since $-31.5 < -2.306$ we reject H_0. Virtually for any level of α we re hypothesis that there is no association between x and y. These data con very strong evidence of positive (linear) association between x and y.

Exercises

For the data set

x	0	1	6	3	5
y	4	3	0	2	1

Construct a scatter diagram

Guess the sign and value of the correlation coefficient

Calculate the correlation coefficient.

A sample data set produced the following information.

$= 460$; $\sum x_i = 9880$; $\sum y_i = 1456$; $\sum x_i y_i = 85080$;

$\sum x_i^2 = 485870$; and $\sum y_i^2 = 135675$

culate the linear correlation coefficient r_{xy}.

Calculations from a data set of $n = 48$ pairs of (x, y) values have provided following results

$(x_i - \bar{x})^2 = 260.2$; $\sum (y_i - \bar{y})^2 = 403.7$; $\sum (x_i - \bar{x})(y_i - \bar{y}) = 298.8$

culate the linear correlation coefficient.

The following table gives the experience (in years) and monthly salaries hundred of dollars) of nine randomly selected secretaries

Experience	14	3	5	6	4	9	18	5	16
Monthly salary	22	12	15	17	15	19	24	13	27

Do you expect the experience and monthly salaries to be positive or atively related?

Compute the correlation coefficient.

Test at the 10% significance level, against a two sided alternative, the null othesis that the population correlation coefficient is zero.

The following data were collected regarding the starting monthly salary the grade point average (GPA) for students who had obtained a degree in iness administration and economics:

GPA	2.6	3.4	3.6	3.2	3.5	2.9
Monthly salary ($)	900	1200	600	1100	1400	1000

a) Develop a scatter diagram for the above data.
b) Compute the sample correlation coefficient between grade point a and salary.
c) Test at the 5% significance level the null hypothesis that the popu correlation coefficient is zero against the alternative that it is positive.

6. The management of a supermarket wanted to check the effect number of broadcast on TV on the gross sales at the store. The manag experimented for eight weeks by broadcasting a different numb commercials each week on TV. The following table gives the numb commercials during each week and the gross sales (in 1000's of dollars

Number of commercials	22	16	28	12	30	19	24	1
Gross sales per week	3.64	3.12	4.08	2.84	3.98	3.55	4.02	

a) Compute the sample correlation coefficient between number of broa and gross sales.
b) Test the null hypothesis that number of broadcasts and gross sal uncorrelated in the population against the alternative that popu correlation is positive.

Answers

1. b) high negative correlation: c) -0.992; **2.** 8.99; **3.** 0.92; **4.** a) positive b) 0.95; c) $T.S. = 8.051$; fail to reject H_0; **5.** b) 0.114; c) $T.S. = 0.23$; H_0; **6.** a) 0.96; b) $T.S. = 8.40$; reject H_0 at virtually any level.

3.4. Spearman rank correlation

ppose that a random sample $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ of n pairs of servations is taken. If x_i and y_i are each ranked in ascending order and the mple correlation of these ranks is calculated, the resulting coefficient is lled the Spearman rank correlation coefficient. If there are no tied nks, an equivalent formula for computing this coefficient is

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

here the d_i are the differences of the ranked pairs:

$$d_i = rank(x_i) - rank(y_i)$$

emark: Spearman rank correlation shares the properties of r_{xy} that $-1 \le r_s \le 1$ and that values near +1 indicate a tendency for the larger values X to be paired with the larger values of Y.

he following test of the null hypothesis H_0 of no association in the pulation

H_0 : x and y are independent

ve significance level α:

To test against the alternative of positive association, the decision rule is

Reject H_0 if $r_s > r_{s,\alpha}$

To test against the alternative of negative association, the decision rule is

Reject H_0 if $r_s < -r_{s,\alpha}$

To test against the two sided alternative of some association, the decision le is

Reject H_0 if $r_s > r_{s,\alpha/2}$ or $r_s < -r_{s,\alpha/2}$

he table of critical values of the Spearman rank correlation coefficient is ven in the Table 5 of the Appendix.

xample:

ores of 8 salesmen made on a test that measures their aggressiveness (x), d their sales in thousands of dollars for their second year with a certain ompany (y).

iveness)	30	17	35	28	42	25	19	34
	35	31	40	46	50	32	33	42

a) Find and interpret Spearman rank correlation

b) Test the null hypothesis that aggressiveness and sales are inde again the alternative that they are positively correlated. Take $\alpha = 0.05$.

Solution:

a) First of all, let us rank **separately** x and y in ascending order. The rank appear in third and fourth columns of the following table 3.2

Table 3.2

x	y	Rank x_i	Rank y_i	$d_i = x_i - y_i$	d_i^2
30	35	4	5	1	1
17	31	8	8	0	0
35	40	2	4	-2	4
28	46	5	2	3	9
42	50	1	1	0	0
25	32	6	7	-1	1
19	33	7	6	1	1
34	42	3	3	0	0
sum					16

The differences between ranks and squared differences between ran shown in the last two columns of the table. Substituting the values $n=$ $\sum d_i^2 = 16$ into formula for Spearman rank correlation, we obtain

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 16}{8 \cdot 63} = 1 - 0.19 = 0.81$$

It means that there exists strong positive correlation between aggressi and sales volume.

b) The null and alternative hypotheses are

H_0 : x and y are independent

H_1 : x and y are positively correlated

The decision rule is

Reject H_0 if $r_s > r_{s,\alpha}$

For a sample of size n=8, and $\alpha = 0.05$,

$$r_{s,\alpha} = r_{8,0.05} = 0.643$$

Since 0.81>0.643 we reject H_0, and accept the alternative hypothesis and y are positively correlated.

Exercises

Specify the rejection region for Spearman's nonparametric test for rank relation in each of the following cases

$H_0 : \rho = 0;\ H_1 : \rho \neq 0;\ n = 10;\ \alpha = 0.05$

$H_0 : \rho = 0;\ H_1 : \rho > 0;\ n = 20;\ \alpha = 0.025$

$H_0 : \rho = 0;\ H_1 : \rho < 0;\ n = 30;\ \alpha = 0.01$

Compute Spearman's rank correlation coefficient for each of the lowing pairs of sample observations

b)

	33	61	20	19	40
	26	36	65	25	35

x	5	20	15	10	3
y	80	83	91	82	87

A random sample of nine pairs of observations are recorded on two iables, x and y.

x	19	27	15	35	13	29	16	22	17
y	12	19	7	25	11	10	16	13	18

the data provide sufficient evidence to indicate that ρ, the rank relation between x and y, differs from zero? Test using $\alpha = 0.05$.

Two expert wine testers were asked to rank six brands of wine. Their kings are shown in the table.

Brand	Expert 1	Expert 2
A	6	5
B	5	6
C	1	2
D	3	1
E	2	4
F	4	3

the data present sufficient evidence to indicate a positive correlation in rankings of the two experts?

Refer to the data of Exercise 6 of the previous section. Find Spearman's k correlation coefficient, and use it to test, against two sided alternative,

the null hypothesis of no association in the population between this pa
random variables.

6. Refer to the data of Exercise 5 of the previous section. Find Spearm
rank correlation coefficient, and use it to test the null hypothesis that t
quantities are uncorrelated in the population against the alternative
population correlation is negative.

Answers

1. a) $r_s > 0.648$ or $r_s < -0.648$; b) $r_s > 0.450$; c) $r_s < -0.432$; **2.** a) $r_s = 0$
b) $r_s = 0.2$; **3.** $r_s = 0.48$; accept H_0; **4.** $r_s = 0.66$; reject H_0 virtually a
level; **5.** $r_s = 0.93$; reject H_0 virtually at any level; **6.** $r_s = 0.143$; can
reject H_0 at 5% level.

3.5. The linear regression model

Let us return to the example of an economist investigating
relationship between food expenditure and income. What factors or varia
does a household consider when deciding how much money should be sp
on food every week or every month? Certainly, income of household is
factor. Many other factors, say, the size of household, the preferences
tests of household members, are some of the variables that will influen
household's decision about food expenditure. These variables are c
independent variables because they are all vary independently and
explain the variation in food expenditure among different households
other words, these variables explain why different households sp
different amounts of money on food. Food expenditure is called
dependent variable because it depends on the independent variab
Studying the effect of two or more independent variables on a depen
variable using regression analysis is called **multiple regression.** I
choose only one (usually the most important) independent variable and s
the effect of that single variable on a dependent variable, it is called a sim
regression. Thus, simple regression includes only two variables:
independent and one dependent.

Definition: A regression model is a mathematical equation that desc
relationship between two or more variables. A simple regression m
includes only two variables: one independent and one dependent.

dependent variable is the one being explained and the independent variable is the one used to explain the variation in the dependent variable.

The relationship between two variables in a regression analysis is expressed by a mathematical equation called a **regression equation** or **model.**

A regression equation that gives a straight line relationship between two variables is called a **linear regression model;** otherwise, it is called a **nonlinear regression model.** In this chapter we will consider only linear regression model.

In a regression model, the independent variable is usually denoted by x and the dependent variable is usually denoted by y. Simple linear regression model is written as

$$y = \alpha + \beta x \qquad (1)$$

In model (1), α gives the value of y for $x = 0$, and β gives the change in y due to a change of one unit in x. This model simply states that y is determined exactly by x and for a given value of x there is one and only one value of y. For example, if y is food expenditure and x is income, then model (1) would state that food expenditure is determined by income only and that all households with the same income will spend the same amount on food. But as mentioned above, food expenditure is determined by many variables, only one of which is included in model (1). In reality, different households with the same income spend different amounts of money on food because of the differences in size of the household, their preferences and tastes. Hence, to take these variables into consideration and make model complete, we add another term to the right side of model (1). This term is called the **error term.** It is denoted by ε (Greek letter *epsilon*). The complete regression model is written as

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \qquad (2)$$

Equation (2) is called the **population (or true) regression line** of y on x.

In equation (2) α and β are the population model coefficients and ε is a random error term .

Population data are difficult to obtain. As a result, we almost always use sample data to estimate model (2). The estimated regression model is given by the equation

$$y_i = a + b \cdot x_i + e_i$$

where a and b are estimated values of the coefficients and e is the differe between the predicted value of y on the regression line, defined as

$$\hat{y}_i = a + b \cdot x_i$$

and the observed value y_i. The difference between y_i and $\hat{y}_i$ for each va of x is defined as the residual

$$e_i = y_i - \hat{y}_i = y_i - (a + b \cdot x_i)$$

Thus for each observed value of x there is a predicted value of y from t estimated model and an observed value. The difference between t observed and predicted values of y is defined as the residual. The resid e_i, is not the model error, ε, but is the combined measure of the mo error and errors in estimating, a and b, and in turn the errors in estimating predicted value.

3.5. 1. Least squares coefficient estimators

The population regression line is useful theoretical construct, but applications we need to determine an estimate of the model using avail data. Suppose that we have n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots$ $\ldots (x_n, y_n)$. We would like to find the straight line that best fits these poi To do this we need estimators of unknown coefficients α and β of population regression line.

We obtain the coefficient estimators, a and b using equations derived using the least squares procedure. As shown in Figure 3.2 there is deviation, e_i between the observed, y_i and the predicted value, $\hat{y}_i$, on estimated regression equation for each value of x, where $e_i = y_i - \hat{y}_i$.

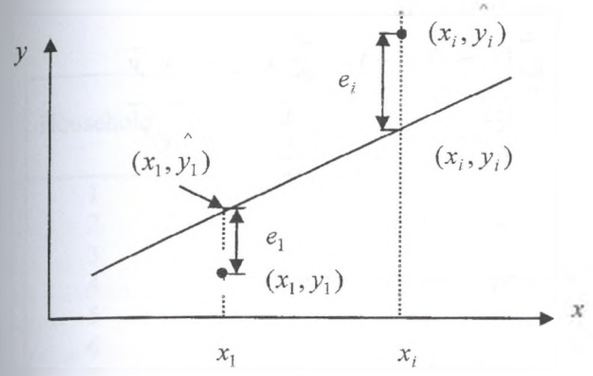

Fig. 3.2

Some of the e_i will be positive and some negative. We then compute a mathematical function that represents the effect of squaring all of the residuals and computing the sum of the squared residuals. This function- whose left side is labeled *SSE* –includes the coefficients, a and b. The quantity *SSE* is defined as the "Error Sum of Squares". The coefficient estimators a and b are selected as the estimators that minimize the Error Sum of Squares.

3.5.2. Least square procedure

The least square procedure obtains estimates of the linear equation coefficients, a and b, in the model

$$\hat{y}_i = a + b \cdot x_i$$

by minimizing the sum of the squared residuals e_i

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

The coefficients a and b are chosen so that the quantity

$$SSE = \sum e_i^2 = \sum (y_i - (a + bx_i))^2$$

is minimized. It can be shown that the resulting estimates are

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}x_iy_i-n\cdot\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n}x_i^2-n\cdot(\bar{x})^2}$$

and $$a=\bar{y}-b\cdot\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the respective sample means.

The line

$$\hat{y}=a+b\cdot x$$

is called the **sample regression line** or **the least squares regression line** y on x.

Example:

Find the least squares regression line for the data on incomes (in hundreds dollars) and food expenditures of seven households given in the table below

Household	1	2	3	4	5	6	7
Income x	35	49	21	39	15	28	25
Food expenditure y	9	15	7	11	5	8	9

Use income as an independent variable and food expenditure as a dependent variable.

Solution:

We are to find the values of a and b for the regression model $\hat{y}_i=a+b\cdot x$

The following table shows the calculations required for the computations of a and b.

Using data from the table 3.3 we find

$$\bar{x}=\frac{212}{7}=30.2857;\qquad \bar{y}=\frac{64}{7}=9.1429$$

Table3.3

Household	Income (x_i)	Food expenditure (y_i)	$x_i\cdot y_i$	x_i^2
1	35	9	315	1225
2	49	15	735	2401
3	21	7	147	441
4	39	11	429	1521
5	15	5	75	225
6	28	8	224	784
7	25	9	225	625
Sums	212	64	2150	7222

$$b=\frac{\sum_{i=1}^{n}x_iy_i-n\cdot\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n}x_i^2-n\cdot(\bar{x})^2}=\frac{2150-7\cdot(30.2857)\cdot(9.1429)}{7222-7\cdot(30.2857)^2}=0.2642$$

$$a=\bar{y}-b\cdot\bar{x}=9.1429-(0.2642)\cdot(30.2857)=1.1414$$

Thus, our estimated regression model $\hat{y}=a+b\cdot x$ is

$$\hat{y}=1.1414+0.2642\cdot x$$

This regression line is called the least squares regression line. It gives the *regression of food expenditure on income.*

Using this estimated model, we can find the predicted value of y for a specific value of x. For example, suppose that we randomly select a household whose monthly income is \$3500 so that $x=35$ (x denotes income in hundred of dollar in our example). The predicted value of food expenditure for this household is

$$\hat{y}=1.1414+0.2642\cdot 35=\$10.3884 \text{ hundred}$$

In other words, based on our regression line, we predict that a household with a monthly income of \$3500 is expected to spend \$1038.84 per month on food.

3.5.3. Interpretation of a and b

a) Interpretation of a

Consider a household with zero income. Using the estimated regressi
obtained above, the predicted value of y for $x = 0$ is

$$\hat{y} = 1.1414 + 0.2642 \cdot 0 = \$1.1414 \text{ hundred}$$

Thus, we can state that a household with no income is expected to
$114.4 per month on food. We should be very careful while maki
interpretation of a. In example of seven households, the incomes vary f
minimum of $1500 to a maximum of $4900. Hence, our regression
valid only for the values of x between 15 and 49. If we predict y for a
of x outside this range, the prediction usually will not hold true. Thus,
$x = 0$ is outside the range of household incomes that we have in the s
data, the prediction that a household with zero income spends $114.
month on food does not carry much credibility.

b) Interpretation of b

The value of b in a regression model gives the change in y (depe
variable) due to a change of one unit in x (independent variable).

For example, by using the regression line $\hat{y} = 1.1414 + 0.2642 \cdot x$

when $x = 30$; $\hat{y} = 1.1414 + 0.2642 \cdot 30 = 9.0674$

when $x = 31$; $\hat{y} = 1.1414 + 0.2642 \cdot 31 = 9.3316$

Hence, when x increased by one unit, from 30 to 31, $\hat{y}$ inc
by $9.3316 - 9.0674 = 0.2642$, which is the value of b. Because of u
measurement in hundred of dollars, we can state that, on average, a
increase in income will cause a $26.42 increase in food expenditure. W
also state that, on average, a $1 increase in income of household
increase the food expenditure by $0.2642.

Note that when b is positive, an increase in x will lead to an increase in
decrease in x will lead to a decrease in y. Such a relationship between x
is called a positive linear relationship. On the other hand, if the value o
negative, an increase in x will cause a decrease in y and a decrease in
cause an increase in y. Such a relationship between x and y is ca
negative linear relationship.

The values of y- intercept and slope calculated from sample data on x and y are called estimated values of α and β and denoted by a and b. Using a and b we can write estimated model as

$$\hat{y} = a + b \cdot x$$

where $\hat{y}$ (read as y *hat*) is the **estimated** or **predicted** value of y for a given value of x.

3.5.4. Assumptions of the regression model

Like any other theory, the linear regression analysis is also based on certain assumptions. Consider the population regression model

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

There are four assumptions made about this model.

Assumption1: The random error term ε has a mean equal to zero for each x. In other words, among all households with the same income, some spend more than predicted food expenditure; others spend less than predicted food expenditure. Some of positive errors equal to the sum of negative errors so that the mean of errors for all households with the same income is zero.

Assumption 2: The errors associated with different observations are independent. According to this assumption, the errors for any two households are independent. All households decide independently how much spend on food.

Assumption 3: For any given x, the distribution of errors is normal. In other words, food expenditure for all households with the same income are normally distributed.

Assumption 4: The distribution of population errors for each x has the same (constant) standard deviation, which is denoted by σ_ε. This assumption indicates that the spread of points around the regression line is similar for all x values.

Exercises

1. Plot the following straight lines. Give the values of the y-intercept slope for each of these lines and interpret them. Indicate whether each of lines gives a positive or negative relationships between x and y.

a) $y = 53 + 7x$; b) $y = 75 - 6x$

2. The following information is obtained from a sample data

$n = 10$; $\sum_{i=1}^{10} x_i = 100$; $\sum_{i=1}^{10} y_i = 220$; $\sum_{i=1}^{10} x_i y_i = 3680$; $\sum_{i=1}^{10} x_i^2 = 1140$

Find the estimated regression line.

3. Computing from a data set of (x, y) values we obtained the follo summary statistics

$n = 14$; $\bar{x} = 3.5$; $\bar{y} = 5.1$; $\sum_{i=1}^{14} (x_i - \bar{x})^2 = 10.82$;

$\sum_{i=1}^{14} (x_i - \bar{x})(y_i - \bar{y}) = 2.677$; $\sum_{i=1}^{14} (y_i - \bar{y})^2 = 2.01$

Obtain the equation of the estimated regression line.

4. Given the five pairs of (x, y) values,

x	0	1	6	3	5
y	4	3	0	2	1

a) Construct a scatter diagram

b) Calculate the least squares estimates a and b.

c) Determine the fitted line and draw the line on the scatter diagram.

5. A researcher took a sample of 36 electronic companies and fou following relationship between x and y where x is the amount of mon thousands of dollars) spent on advertising by a company during a year represents the total gross sales (in thousands of dollars) of that comp that year.

$\hat{y} = 5.6 + 22.5 \cdot x$

a) An electronic company spent 2000$ on advertising during a year. are its expected gross sales for that year?

b) Suppose five electronic companies spent 2000$ each on advertising during that year. Do you expect these five companies to have the same actual gross sales for that year? Explain.

6. An economist wanted to determine whether or not the amount of phone bills and income of households are related. The following table gives information on the monthly incomes (in hundreds of dollars) and monthly telephone bills (in dollars) for a random sample of 10 households

Income	16	45	36	32	30	13	41	15	36	40
Phone bill	35	78	102	56	75	26	130	42	59	85

a) Find the regression line with income as an independent variable and the amount of the phone bill as a dependent variable.

b) Give an interpretation of the values of a and b calculated in part a.

c) Estimate the amount of the monthly phone bill for a household with a monthly income of $2500.

7. An auto manufacturing company wanted to investigate how the price of one of its car models depreciates with age. The research department at the company took a sample of 9 cars of this model and collected the following information on the ages (in years) and prices (in hundreds of dollars) of these cars.

Age	8	3	7	10	3	5	6	9
Price	16	74	38	21	98	56	49	30

a) Construct a scatter diagram for these data. Interpret your results.

b) Find the regression line with price as a dependent variable and age as an independent variable.

c) Give a brief interpretation of the values of a and b calculated in part b.

d) Predict the price of a 4 year-old car of this model.

e) Estimate the price of a 19-year-old car of this model. Comment on this finding.

8. Construct a scatter diagram for the data in the following table

x	0.5	1	1.5
y	2	1	3

a) Plot the following two lines on your scatter diagram

1) $y = 3 - x$ and 2) $y = 1 + x$

b) Which of these lines would you choose to characterize the relati between x and y? Explain
c) Show that the sum of errors for both of these lines equals 0.
d) Which of these lines has smaller *SSE* ?
e) Find the least squares regression line for the data and compare it to lines described in part a.

Answers

2. $\hat{y} = -83.714 + 10.571 \cdot x$; **3.** $\hat{y} = 4.225 + 0.247 \cdot x$; **4.**c) $\hat{y} = 3.845 - 0.61$

5. a)\$50.6 thousand; b)different amounts; **6.**a) $\hat{y} = 2.3173 + 2.1869 \cdot x$;

c) \$56.99; **7.** $\hat{y} = 111 - 9.84 \cdot x$; **8.** b) The second line; d) The second

e) $\hat{y} = 1 + x$.

3.6. The explanatory power of a linear regression equation

In Figure 3.3 it is shown that the deviation of an individual y value fro mean can be

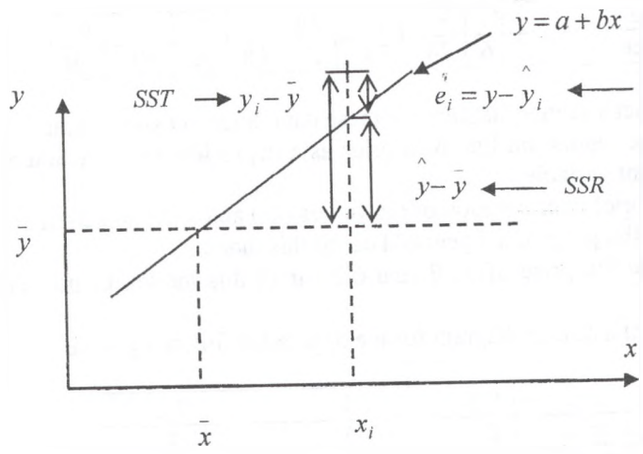

Fig.3.3

partitioned into deviation of the predicted value from the mean and the deviation of the observed value from the predicted value

$$y_i - \bar{y} = (y_i - \hat{y}) + (\hat{y}_i - \bar{y})$$

We square each side of the equation-because the sum of deviations about the mean is equal to zero-and sum the results over all n points

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Some of you may note the squaring of the right- hand side should include the cross product of the two terms in addition to their squared quantities. It can be shown that the cross predicted term goes to zero. This equation is expressed as

$$SST = SSR + SSE$$

We see that the total variability-*SST*- consists of two components-*SSR*-the amount of variability explained by the regression equation- named "Regression Sum of Squares" and –*SSE*-random or unexplained deviation of points from the regression line-named "Error Sum of Squares". Thus

Total sum of squares: $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Regression Sum of Squares: $SSR = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2 = b^2 \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$

Error Sum of Squares: $SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 = \sum_{i=1}^{n} (e_i)^2$

For a given set of observed values of the dependent variables, y, the *SST* is fixed as the total variability of all observations from the mean. We see that in the partitioning larger values of *SSR* and hence smaller value of *SSE* indicate a regression equation that "fits" or comes closer to the observed data. This partitioning is shown graphically in Figure 3.3.

Example:

Let us find *SST*, *SSR* and *SSE* for the data on incomes and food expenditure. Using calculation given in the table 3.3 we find the value of total sum of squares as

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{7} y_i^2 - \frac{\left(\sum_{i=1}^{7} y_i\right)^2}{n} = 646 - \frac{64^2}{7} = 60.8571$$

Table3.4

x	y	$\hat{y}$	y^2	e_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	e_i^2
35	9	10.3884	81	-1.3884	4.7143	22.2246	1.9277
49	15	14.0872	225	0.9128	18.7143	350.225	0.8332
21	7	6.6896	49	0.3104	-9.2857	86.2242	0.0963
39	11	11.4452	121	-0.4452	8.7143	75.9390	0.1982
15	5	5.1044	25	-0.1044	-15.286	233.653	0.0109
28	8	8.5390	64	-0.5390	-2.2857	5.2244	0.2905
25	9	7.7464	81	1.2536	-5.2857	27.9386	1.5715
			646			801.429	4.9283

The error sum of squares *SSE* is given in the sum of the eights column in Table 3.4. Thus,

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(e_i)^2 = 4.9283$$

The regression sum of squares can be found from $SST = SSR + SSE$. Thus

$$SSR = SST - SSE = 60.8571 - 4.9283 = 55.9288.$$

The value of *SSR* can also be computed by using the formula.(Check!!)

$$SSR = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 = b^2 \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

The total sum of squares *SST* is a measure of the total variation in food expenditures, *SSR* is the portion of total variation explained by the regression model (or by income), and the error sum of squares *SSE* is the portion of total variation not explained by the regression model.

3.6.1. Coefficient of determination R^2

If we divide both side of the equation

$$SST = SSR + SSE$$

by *SST*, we obtain

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

We have seen that the fit of the regression equation to the data is improved as *SSR* increases and *SSE* decreases. The ratio $\frac{SSR}{SST}$ provides a descriptive measure of the proportion or percent of the total variability that is explained by the regression model. This measure is called the *coefficient of determination*-or more generally R^2.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The coefficient of determination is often interpreted as the percent of variability in y that is explained by the regression equation. We see that R^2 increases directly with the spread of the independent variable.

R^2 can vary from 0 to 1 since *SST* is fixed and $0 < SSE < SST$. A larger R^2 implies a better regression, everything else being equal.

Interpretation of R^2: About $(100 \cdot R^2)\%$ of the sample variation in y (measured by the total sum of squares of deviations of the sample y values about their mean $\bar{y}$) can be explained by using x to predict y in the straight line model.

Example:

Calculate the coefficient of determination for the data on monthly incomes and food expenditures of seven households.

Solution:

From earlier calculations

$$SSR = 55.9288 \quad \text{and} \quad SST = 60.8571$$

Hence,

$$R^2 = \frac{SSR}{SST} = \frac{55.9288}{60.8571} = 0.92$$

We can state that 92% of the variability in y is explained by linear regression, and the linear model seems very satisfactory in this respect. In other words, we can state that 92% of the total variation in food expenditures of households occurs because of the variation in their incomes, and the remaining 8% is due to other variables, like differences in size of the household, preferences and tastes and so on.

3.6.2. Estimation of model error variance

When we consider income and food expenditures, all households with the same income are expected to spend different amounts on food. Consequently, the random error ε_i will have different values for these households. The variance $\hat{\sigma}_i^2$ measures the spread of these errors around the population regression line. Note that $\hat{\sigma}_i^2$ denotes the variance of errors for the population. However, usually $\hat{\sigma}_i^2$ is unknown. In such cases, it is estimated by s_e^2, which is the standard deviation of errors for the sample data.

An estimator for the variance of the population model error is

$$\hat{\sigma}_e^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$$

Division by $(n-2)$ instead of $(n-1)$ results because the simple regression model uses two estimated parameters, a and b, instead of one.

The formula for *SSE* is

$$SSE = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

If we introduce the following notations

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

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

where SS stands for "sums of squares", then

$$SSE = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}.$$

Exercises

1. The following information is obtained from a sample data set

$$n = 12; \quad \sum_{i=1}^{12} x_i = 66; \quad \sum_{i=1}^{12} y_i = 588; \quad \sum_{i=1}^{12} x_i y_i = 2244;$$

$$\sum_{i=1}^{12} x_i^2 = 396; \quad \text{and} \quad \sum_{i=1}^{12} y_i^2 = 58734$$

Find the values of s_e^2 and R^2.

2. A sample data set produced the following information

$$n = 460; \quad \sum_{i=1}^{460} x_i = 3920; \quad \sum_{i=1}^{460} y_i = 2650; \quad \sum_{i=1}^{460} x_i y_i = 26570;$$

$$\sum_{i=1}^{460} x_i^2 = 48530; \quad \text{and} \quad \sum_{i=1}^{460} y_i^2 = 39347$$

Find the values of s_e^2 and R^2.

3. Computing from a data set of (x, y) values produced the following summary statistics

$$n = 14; \quad \bar{x} = 3.5; \quad \bar{y} = 2.32;$$

$$SS_{xx} = 10.82; \quad SS_{xy} = 2.677; \quad SS_{yy} = 1.035$$

a) Obtain equation of the best fitting straight line.

b) Estimate $\hat{\sigma}_e^2$.

4. Computing from a data set of (x, y) values produced the following summary statistics

$$n=14; \quad \bar{x}=1.2; \quad \bar{y}=5.1;$$
$$SS_{xx}=14.10; \quad SS_{xy}=2.31; \quad SS_{yy}=2.01$$

Determine the proportion of variation in y that is explained by linear regression.

5. A calculation shows that $SS_{xx}=10.1, SS_{yy}=16.5$, and $SS_{xy}=9.3$, determine the proportion of variation in y that is explained by linear regression.

6. The following table lists the sizes of offices (in hundreds of square meters) and the rents (in dollars) paid for those offices.

Size of offices	22	17	19	28	35	24
Monthly rent	710	590	730	880	1080	820

a) Find the regression line $y=a+bx$ with the size of an office as an independent variable and monthly rent as a dependent variable.
b) Give a brief interpretation of the values of a and b.
c) Predict the monthly rent for the office with 2400 square meters.
d) One of the offices is 2600 square meters and its rent is \$850. What is the predicted rent for this office? Find the error for this office.
e) Compute the standard deviation of errors.
f) Calculate the coefficient of determination. What percentage of the variation in monthly rents explained by the sizes of the offices? What percentage of this variation is not explained?

7. Refer to exercise 7 of previous chapter. The following table which gives the ages (in years) and prices (in hundred of dollars) of eight cars of specific model, is reproduced from that exercise.

Age	8	3	7	10	3	5	6	9
Price	16	74	38	21	98	56	49	30

a) Calculate the standard deviation of errors.
b) Compute the coefficient of determination and give a brief interpretation of it.

Answers

1. 22.2; 0.99; 2. 50.06; 0.04; 3. a) $\hat{y} = 1.454 + 0.247 \cdot x$; b) $s_e^2 = 0.031$;

4. 0.188; 5. 0.5190; 6. a) $\hat{y} = 194 + 25.1 \cdot x$ e)40.2; f) 0.953;

; 7. a) 11.39; b) 0.856

3.7. Statistical inference: Hypothesis tests and confidence intervals

One of the main purposes for determining a regression line is to find the true value of the slope β of the population regression line. However, in almost all cases, the regression line is estimated using sample data. Then based on the sample regression line, inferences are made about the population regression line. The slope b of a sample regression line is a point estimator of the slope β of the population regression line. The different sample regression lines estimated for different samples taken from the same population will give different values of b. If only one sample is selected, then the value of b will depend on which elements are included in the sample. Thus, b is a random variable and it possesses a probability distribution called a sampling distribution.

Assume that assumptions 3.5.4 are hold. Then b is an unbiased estimator of β and has a population variance

$$\sigma_b^2 = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} x_i^2 - n \cdot (\bar{x})^2}$$

and unbiased estimator of σ_b^2 is provided by

$$s_b^2 = \frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_e^2}{\sum_{i=1}^{n} x_i^2 - n \cdot (\bar{x})^2} = \frac{s_e^2}{SS_{xx}}$$

In applied regression analysis we first would like to know if there is a relationship. We see that if β is zero then there is no relationship- y would not continuously increase or decrease with increase in x.

3.7.1. Hypothesis testing about β

Let β be a population regression slope and b its least square estimate based on n pairs of sample observations. Assume that assumptions 3.5.4 hold and also assume that the errors ε_i are normally distributed. Then the random variable

$$t = \frac{b - \beta}{s_b}$$

is distributed as Student's t distribution with $(n-2)$ degree of freedom. If we use notation

$$T.S. = t = \frac{b - \beta}{s_b}$$

for the test statistic then the following tests have a significance level α

1. To test either null hypothesis

$$H_0 : \beta = \beta_0 \text{ or } H_0 : \beta \leq \beta_0$$

against the alternative

$$H_1 : \beta > \beta_0$$

the decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > t_{n-2,\alpha}$$

2. To test either null hypothesis

$$H_0 : \beta = \beta_0 \text{ or } H_0 : \beta \geq \beta_0$$

against the alternative

$$H_1 : \beta < \beta_0$$

the decision rule is

$$\text{Reject } H_0 \text{ if } T.S. < -t_{n-2,\alpha}$$

3. To test null hypothesis

$$H_0 : \beta = \beta_0$$

against the two sided alternative

$$H_1 : \beta \neq \beta_0$$

the decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > t_{n-2,\alpha/2} \text{ or } T.S. < -t_{n-2,\alpha/2}$$

Remark1: To test the hypothesis that x does not determine y linearly and there is no linear relationship, we will test the null hypothesis that the slope of the regression line is zero, that is $H_0 : \beta = \beta_0 = 0$; the alternative hypothesis that $H_1 : \beta \neq \beta_0 \neq 0$ means x determines y linearly; $H_1 : \beta > \beta_0 = 0$ means x determines y positively; $H_1 : \beta < \beta_0 = 0$ means x determines y negatively.

Remark2: The null hypothesis does not always have to be $\beta = 0$. We may test the null hypothesis that β is equal to a value different from zero.

Example:

Test at the 5% significance level if the slope of the population regression line for the example on incomes and food expenditure of seven households is positive.

Solution:

From earlier calculations we have

$n = 7$; $b = 0.2642$ and $s_e = 0.9922$

$$s_b^2 = \frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{0.9856}{801.429} = 0.001229; \text{ and } s_b = 0.0350.$$

We are to test whether or not slope β of the population regression line is positive. The two hypotheses are

$H_0 : \beta = \beta_0$ (Slope is zero)

$H_1 : \beta > \beta_0$ (Slope is positive)

The decision rule is

reject H_0 if $T.S. > t_{n-2,\alpha}$.

The value of the test statistic is

$$T.S. = t = \frac{b - \beta}{s_b} = \frac{0.2642 - 0}{0.0350} = 7.549$$

The significance level is 0.05. Therefore,

$$t_{n-2,\alpha} = t_{5,0.05} = 2.015$$

The value of the test statistic $T.S. = 7.549$ is greater than the critical value of $t = 2.015$ and it falls in the rejection region. Hence, we reject the null hypothesis and conclude that x (income) determines y (food expenditure)

positively. That is, food expenditure increases with an increase in income and it decreases with a decrease in income.

3.7.2. Confidence intervals for the population regression slope β

We can derive confidence intervals for the slope β of the population regression line by using coefficient b and variance estimators we have developed.

If the assumptions 3.5.4 hold, and if the regression errors, ε_i, are normally distributed, then $100(1-\alpha)\%$ confidence interval for the population regression slope β is given by

$$b - t_{n-2,\alpha/2} \cdot s_b < \beta < b + t_{n-2,\alpha/2} \cdot s_b$$

where $t_{n-2,\alpha/2}$ is the number for which $P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$ and the random variable t_{n-2} follows Student's t distribution with $(n-2)$ degrees of freedom.

Example:

Construct a 95% confidence interval for β for the data on incomes and food expenditures of seven households.

Solution:

From earlier calculations we have

$n = 7$; $b = 0.2642$; $s_e = 0.9922$ and $s_b = 0.0350$

The confidence level is 95%. So

$$100(1-\alpha)\% = 95\%$$
$$1-\alpha = 0.95$$
$$\alpha = 0.05$$
$$\alpha/2 = 0.025$$
$$t_{n-2,\alpha/2} = t_{5,0.025} = 2.571$$

The 95% confidence interval for β is

$$b - t_{n-2,\alpha/2} \cdot s_b < \beta < b + t_{n-2,\alpha/2} \cdot s_b$$
$$0.2642 - 2.571 \cdot 0.0350 < \beta < 0.2642 + 2.571 \cdot 0.0350$$
$$0.17 < \beta < 0.35$$

Thus, we are 95% confident that slope β for the population regression line is between 0.17 and 0.35.

Exercises

1. The following information is obtained for a sample of 16 observations taken from a population

$$SS_{xx} = 340.700; \qquad s_e = 1.951; \qquad \text{and } \hat{y} = 12.45 + 6.32 \cdot x$$

a) Make a 99% confidence interval for β.

b) Using a significance level of 0.025, test the null hypothesis that β is zero against the alternative that β is positive.

c) Using a significance level of 0.01, can you conclude that β is zero against the alternative that it is different from zero?

d) Using a significance level of 0.02, test whether β is different from 4.50.

2. The following information is obtained for a sample of 100 observations taken from a population. (Note that because $n > 30$, we can use the normal distribution to make a confidence interval and test a hypothesis about β)

$$SS_{xx} = 524.884; \qquad s_e = 1.464; \qquad \text{and } \hat{y} = 5.48 + 2.50 \cdot x$$

a) Make a 98% confidence interval for β

b) Test at the 2% significance level whether β is zero against the alternative that it is positive.

c) Can you conclude that β is zero? Use $\alpha = 0.01$.

d) Using a significance level of 0.01, test whether β is 1.75 against the alternative that it is greater than 1.75.

3. Refer to exercise 7 of previous chapter. The following table which gives the ages (in years) and prices (in hundred of dollars) of eight cars of specific model, is reproduced from that exercise.

Age	8	3	7	10	3	5	6	9
Price	16	74	38	21	98	56	49	30

a) Construct a 95% confidence interval for β.

b) Test at the 5% significance level if β is zero against the alternative that it is negative.

4. The following table gives the experience (in years) and monthly salaries (in thousands of tenge) of nine randomly selected secretaries

Experience	14	3	5	6	4	9	18	5	16
Monthly salary	22	12	15	17	15	19	24	13	27

a) Find the least squares regression line with experience as an independent and monthly salary as dependent variables.
b) Construct a 95% confidence interval for β .
c) Test at the 2.5% significance level if β is zero against the alternative that it is positive.

5. The data on the size of six offices (in hundreds of square meters) and the monthly rents (in dollars) paid by firms for those offices are reproduced below from exercise 6 of the previous section.

Size of offices	22	17	19	28	35	24
Monthly rent	710	590	730	880	1080	820

a) Construct a 99% confidence interval for β . You can use the calculations made in exercise 6 of previous section here.
b) Test at the 5% significance level the null hypothesis that β is zero against the alternative that it is different from zero.

6. The following data give information on the ages (in years) and the number of breakdowns during the past year for a sample of six machines at a large company.

Age	9	14	18	15	10	11
Number of breakdowns	34	46	52	64	42	44

a) Find the least squares regression line $y = a + b \cdot x$
b) Give a brief interpretation of the values a and b.
c) Compute and interpret R^2 .
d) Compute the standard deviation of errors.
e) Construct a 98% confidence interval for β .
f) Test at the 2.5% significance level the null hypothesis that β is zero against the alternative that it is positive.

7. The following table gives information on the temperature in a city and volume of the ice cream (in thousands) sold at the supermarket for a random sample of eight days during the summer.

Temperature	22	16	28	12	30	19	24	32
Ice cream sold	3.64	3.12	4.08	2.84	3.98	3.55	4.02	4.38

a) Find the least squares regression line $\hat{y} = a + b \cdot x$. Take temperature as an independent variable and volume of ice cream sold as a dependent variable.
b) Give a brief interpretation of the values a and b.
c) Compute and interpret R^2.
d) Compute the standard deviation of errors.
e) Construct a 95% confidence interval for β.
f) Test at the 1% significance level the null hypothesis that β is zero against the alternative that it is positive.

Answers

1.a) 6.01 to 6.63;b) $T.S. = t = 59.792$; reject H_0;c) $T.S. = t = 59.792$; reject H_0 d) $T.S. = t = 17.219$; reject H_0; **2.**a)2.35 to 2.65;b) $T.S. = z = 39.12$; reject H_0; c) $T.S. = z = 39.12$; reject H_0; d) $T.S. = z = 11.74$; reject H_0; **3.** a) -15.53 to -7.17; b) $T.S. = t = -6.645$; reject H_0; **4.** a) $\hat{y} = 10.4986 + 0.8689 \cdot x$; b) 0.5559 to 1.1819; c) $T.S. = t = 8.323$; reject H_0; **5.** a) 12.34 to 37.92 ; b) $T.S. = t = 4.604$; reject H_0 **6.** a) $\hat{y} = -1.4337 + 0.8916 \cdot x$; c) $R^2 = 0.94$; d) $s_e = 0.9285$; e) 0.5708 to 1.2124; f) $T.S. = t = 9.356$; reject H_0;

7. a) $\hat{y} = 2.0680 + 0.0714 \cdot x$; c) $R^2 = 0.92$;d) $s_e = 0.1537$; e) 0.0511 to 0.0917; f) $T.S. = t = 8.602$ reject H_0.

3.8. Using the regression model for prediction a particular value of y

The second major use of a regression model is to predict a particular value of y for a given value of x, say x_0. For example, we may want to predict the food expenditure of a randomly selected household with a monthly income of \$3000. In this case, we are not interested in the mean food expenditure of all households with a monthly income of \$3000 but in the food expenditure of one particular household with a monthly income of \$3000. This predicted value of y is denoted by y_p. To predict a single value of y for $x = x_0$ from estimated sample regression line, we use the value of $\hat{y}$ as *a point estimate of* y_p. Using the estimated regression line, we find $\hat{y}$ for x =30 as

$$\hat{y} = 1.1414 + 0.2642 \cdot (30) = 9.0674$$

Thus, based on our regression line, the point estimate for the food expenditure of a given household with a monthly income of \$3000 is \$906.74 per month.

Different regression lines estimated by using different samples of seven households each taken from the same population will give different values of the point estimator for the predicted value of y for x =30.Hence , a confidence interval constructed for y_p based on one sample will give a more reliable estimate of y_p than will a point estimate. The confidence interval constructed for y_p is more commonly called a prediction interval.

Suppose that the population regression model is

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \qquad (i = 1,2,\ldots\ldots,n+1)$$

and that the standard regression assumptions hold, and that the ε_i are normally distributed. Let a and b be the least squares estimates of α and β.

It can be shown that the following are $100(1-\alpha)\%$ intervals:

1. For the forecast of the single value resulting for y_{n+1} at a given x_{n+1}, the prediction interval is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} \cdot s_e$$

2. For the forecast of the conditional expectation, $E(y_{n+1} / x_{n+1})$, the confidence interval is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} \cdot s_e$$

where

$$\bar{x} = \frac{\sum_{i+1}^{n} x_i}{n} \qquad \text{and} \qquad \hat{y}_{n+1} = a + b \cdot x_{n+1}.$$

Example:

For the data on incomes and food expenditures of seven households, find

a) 99% prediction interval for the predicted food expenditure for a single household with a monthly income of \$3500;

b) Obtain a 99% confidence interval for the expected food expenditure for all households with a monthly income of \$3000.

Solution:

a) The point estimate of the predicted food expenditure for $x = 35$ is given by

$$\hat{y} = 1.1414 + 0.2642 \cdot (35) = 10.3884$$

$$100(1-\alpha)\% = 99\%$$

$$\alpha = 0.01$$

$$\alpha / 2 = 0.005$$

$$t_{n-2,\alpha/2} = t_{5,0.005} = 4.032$$

Using data from the previous chapters

$s_e = 0.9922$; $\bar{x} = 30.2857$; and $SS_{xx} = 801.4286$

Hence, the 99% prediction interval for y_p for $x = 35$ is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} \cdot s_e$$

$$10.3884 \pm 4.032 \cdot \sqrt{\left[1 + \frac{1}{7} + \frac{(35 - 30.2857)^2}{801.4286}\right]} \cdot 0.9922 =$$

$$= 10.3884 \pm 4.3284 = 6.0600 \text{ to } 14.7168$$

Thus, with 99% confidence we can state that the predicted food expenditure of a household with a monthly income of \$3500 is between \$606.00 and \$1471.68.

b) Once again, the point estimate of the expected food expenditure for $x = 35$ is

$$\hat{y} = 1.1414 + 0.2642 \cdot (35) = 10.3884$$

Hence, the 99% confidence interval for $E(y_{n+1}/35)$ is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} \cdot s_e$$

$$10.3884 \pm 4.032 \cdot \sqrt{\left[\frac{1}{7} + \frac{(35 - 30.2857)^2}{801.4286}\right]} \cdot 0.9922 =$$

$$= 10.3884 \pm 1.6523 = 8.7361 \text{ to } 12.0407$$

Thus, with 99% confidence we can state that the mean food expenditure for all households with monthly income of \$3500 is between \$873.61 and \$1204.07.

As we can observe, the interval in part a) $[606.00 \text{ to } 1471.68]$ is much wider than the one for the mean value of y for $x = 35$ calculated in part b) $[873.61 \text{ to } 1204.04]$. This is always true. The prediction interval for predicting a single value of y is always larger than the confidence interval for estimating the mean value of y for a certain value of x.

Exercises

1. Construct a 99% confidence interval for the mean value of y and a 99% prediction interval for the predicted value of y for the following

a) $\hat{y} = 3.25 + 80 \cdot x$ for $x = 15$ given $s_e = 0.954$;

$\bar{x} = 18.52$; $SS_{xx} = 144.65$; and $n = 10$

b) $\hat{y} = -27 + 7.67 \cdot x$ for $x = 12$ given $s_e = 2.46$; $\bar{x} = 13.43$;

$SS_{xx} = 369.77$; and $n = 10$

2. Refer to Exercise 4 of the previous section. Construct a 90% confidence interval for the mean monthly salary of secretaries with 10 years of experience. Construct a 90% prediction interval for the monthly salary of a randomly selected secretary with 10 years of experience.

3. Refer to Exercise 6 of the previous section. Construct a 95% confidence interval for the mean number of breakdowns for all cars which are 16 years old. Determine a 95% prediction interval for y_p for $x = 16$.

4. The following data give information on the lowest cost price (in dollars) and the average attendance (thousand) for the past year for eight football teams

Ticket price	3.6	3.3	2.8	2.6	2.7	2.9	2.0	2.6
Attendance	24	21	22	22	18	13	9	6

a) Taking ticket price as an independent variable and attendance as a dependent variable, estimate the regression of attendance on the ticket price.
b) Interpret the slope of the estimated regression line.
c) Find and interpret the coefficient of determination.
d) Find and interpret a 90% confidence interval for the slope of the population regression line.
e) Find a 90% confidence interval for expected number of attendance for which the price of ticket is 20.

5. A sample of 25 employees at a production plant was taken. Each employee was asked to assess his or her own job satisfaction (x), on scale from 1 to 10. In addition, the number of days absent (y) from work during the last year were found for these employees. The sample regression line

$$\hat{y} = 13.6 - 1.2 \cdot x$$

was estimated by least squares for these data. Also found that

$$\bar{x} = 6.0; \qquad \sum_{i=1}^{25} (x_i - \bar{x})^2 = 130; \qquad SSE = 80.6$$

a) Test at the 1% significance level against the appropriate one sided alternative the null hypothesis that job satisfaction has no linear effect on absence.

b) A particular employee has job satisfaction level 4. Find a 90% confidence interval for the number of days this employee would be absent from work in a year.

Answers

1. a) 13.871 to 16.629; b) 11.765 to 18.735; **2.** 18.108 to 20.267; 15.838 to 22.537; **3.** **4.** a) $\hat{y} = 2.029 + 0.0464 \cdot x$; c) $R^2 = 0.4194$; d) $0.003 < \beta < 0.0928$; e) 2.6525 to 3.2615; **5.** a) $T.S. = t = -7.303$; reject H_0; b) 5.4798 to 12.1202.

Chapter 4

Multiple regression analysis

4.1. Introduction

In Chapter 3 we showed how regression analysis could be used to develop an equation that would estimate the relationship between two variables. Recall that we limited our discussion to the development of a linear relationship between the two variables, or what is commonly referred to as simple linear regression. There are many important situations, however, where the underlying relationship between two variables can not be explained adequately with a straight-line relationship. In addition, the most of real world problems require the consideration of more than one independent variable in order to predict the dependent variable. In this chapter we discuss how multiple regression analysis can be used to handle such situations.

4.2. Multiple regression model

Usually a dependent variable is affected by more than one independent variable. When we include two or more independent variables in a regression model, it is called a **multiple regression model.**

A multiple regression model with y as a dependent variable and x_1, x_2 as independent variables is written as

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \varepsilon$$

where the numbers $\alpha, \beta_1,$ and β_2 must be estimated from sample data.

More generally, a multiple regression model with y as a dependent variable and $x_1, x_2, x_3, \ldots\ldots$ and x_k as independent variables is written as

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \ldots + \beta_k \cdot x_k + \varepsilon \qquad (1)$$

where the numbers α represents the constant term and $\beta_1, \beta_2, \beta_3, \ldots\ldots \beta_k$ are the regression coefficients of an independent variables $x_1, x_2, x_3, \ldots\ldots$ and x_k, respectively.

In (1) if each of the independent variables is set to 0, it follows that

$$E(y / x_1 = 0; x_2 = 0; \ldots\ldots x_k = 0) = \alpha$$

Thus, α is expected value of the dependent variable when every independent variable takes value 0. Frequently this interpretation does not carry practical interest and often leads to meaningless.

The interpretation of the coefficients $\beta_1, \beta_2, \beta_3, \ldots\ldots \beta_k$ is extremely important. For example, β_1 is expected increase in y resulting from 1 unit increase in x_1 when the values of the other independent variables remain constant. In general, β_i is expected increase in the dependent variable resulting from a 1-unit increase in the independent variable x_i when the values of the other independent variables remain constant.

If model (1) is estimated using sample data, which is usually the case, the estimated regression model is written as

$$\hat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \ldots\ldots + b_k \cdot x_k \qquad (2)$$

In model (2) $a, b_1, b_2, b_3, \ldots\ldots$ and b_k are the sample statistics, which are the point estimators of α, $\beta_1, \beta_2, \beta_3, \ldots\ldots$ and β_k, respectively.

In model (1) y denotes the actual values of the dependent variable. In model (2), $\hat{y}$ denote the predicted or estimated values of the dependent variable

The difference between y and $\hat{y}$ gives the error of prediction.

The method of fitting multiple regression of least squares model is similar to that of fitting the linear regression model: method of least squares. That is, we choose the estimated model

$$\hat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \ldots\ldots + b_k \cdot x_k$$

that minimizes

$$SSE = \sum (y - \hat{y})^2 .$$

4.3. Standard assumptions for the multiple regression models

Like the simple linear regression model, the multiple regression model is also based on certain assumptions.

Consider the multiple regression model

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \ldots. + \beta_k \cdot x_k + \varepsilon$$

The following assumptions are often made:

Assumption 1: For any given set of values of $x_1, x_2, x_3, \ldots\ldots$ and x_k, the random error ε has a normal probability distribution with mean equal to 0 and variance equal to σ^2.

Assumption 2: The errors associated with different sets of values of independent variables are independent.

Assumption 3: The independent variables are not linearly related. If any of them is linearly related, then we can eliminate one of the variables by making substitution and reduce the number of independent variables.

Assumption 4: It is not possible to find a set of numbers $c_0, c_1, c_2, \ldots\ldots c_k$, such that

$$c_0 + c_1 \cdot x_1 + c_2 \cdot x_2 + c_3 \cdot x_3 + \ldots. + c_k \cdot x_k = 0$$

4.4. The explanatory power of a multiple regression equation

4.4.1. Estimation of error variance

The variance of errors (also called the variance of the estimate) for the multiple regression model

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \ldots. + \beta_k \cdot x_k + \varepsilon$$

is denoted by σ_e^2. However, when sample data are used to estimate multiple regression model (1), the variance of errors, denoted by s_e^2, is an unbiased estimate of the σ_e^2. The formula for calculating s_e^2 is as follows

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - K - 1} = \frac{SSE}{n - K - 1}$$

where

n-is the sample size

K-is the number of independent variables included in the model.

The positive square root of the variance s_e is also called the standard error of the estimate.

4.4.2 The coefficient of determination

Multiple regression uses independent variables to explain the behaviour of the dependent variable. Part of the variability in the dependent variable can be explained by its linear association with the independent variables. We will develop a measure of the proportion of the variability in the dependent variable that can be explained by the multiple regression.

Let the multiple regression model fitted by least squares be

$$y_i = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + b_3 \cdot x_{3i} + \ldots\ldots + b_k \cdot x_{ki} + e_i = \hat{y}_i + e_i$$

where $a, b_1, b_2, b_3, \ldots\ldots$ and b_k are the least squares estimates of the population regression model and e_i's are the residuals from the estimated regression model.

The model variability can be partitioned into the components

$$SST = SSR + SSE$$

where

Total sum of squares: $$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Error sum of squares: $$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Regression sum of squares: $$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

The coefficient of determination for a multiple regression model, usually called the **coefficient of determination,** is denoted by R^2 and is defined as the proportion of the total sample sum of squares SST that is explained by the multiple regression model.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

It tells us how good the multiple regression model is and how well the independent variables included in the model explain the dependent variable.

The value of the coefficient of determination R^2 always lies in the range 0 to 1, that is

$$0 \leq R^2 \leq 1$$

4.4.3 Adjusted coefficient of determination

There is a potential problem with using R^2 as an overall measure of the quality of a fitted equation. The value of R^2 generally increases as we add more and more explanatory variables to the regression model. Therefore, by adding a large number of variables to our regression model (even if they are not included in the model) we can make the value of R^2 very close to 1. Such a value of R^2 will be misleading, and it will not represent the true explanatory power of the regression model. To eliminate this shortcoming of R^2, it is preferable to use the **adjusted coefficient of determination,** which is denoted by $\bar{R}^2$. The value of $\bar{R}^2$ may increase, decrease, or stay the same as we add more explanatory variables to our regression model. If a new variable added to the regression model contributes significantly to explain the variation in y, then $\bar{R}^2$ increases; otherwise it decreases. The value of $\bar{R}^2$ is calculated as follows

$$\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-K-1} \text{ or}$$

$$\bar{R}^2 = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)}$$

4.4.4 Predictions from the multiple regression models

Given that the population regression model

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + ... + \beta_k \cdot x_k + \varepsilon$$

holds and assume that the standard regression assumptions are valid. L
$a, b_1, b_2, ...b_k$ be the least squares estimates of the model coefficie
$\alpha, \beta_1, \beta_2, ...\beta_k$, correspondingly. Then given a new observation of a d
point, $x_{1,n+1}$; $x_{2,n+1}$;.... $x_{k,n+1}$ the best linear unbiased forecast of y_{n+1} is

$$\hat{y}_{n+1} = a + b_1 \cdot x_{1,n+1} + b_2 \cdot x_{2,n+1} + ... + b_k \cdot x_{k,n+1}$$

Remark: It is very risky to calculate forecasts that are based on x_i valu
outside the range of the data used to estimate the model coefficients, becau
we have not included these points to the linear model.

Exercises

1. The regression model $y_i = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \varepsilon_i$ was fitted to a d
set obtained from 20 runs of an experiment in which two predictors x_{1i}
x_{2i} were observed along with the response y_i. The least squares estim
were

$$a = 4.21; \qquad b_1 = 11.37; \qquad b_2 = -0.513$$

Predict the response for

a) $x_1 = 8; \qquad x_2 = 30$

b) $x_1 = 8; \qquad x_2 = 50$

2. The following model was fitted to a sample of 25 families in ord
explain household milk consumption

$$y_i = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \varepsilon_i$$

where

y_i -milk consumption, in liters per week

x_{1i} -weekly income, in hundreds of dollars

x_{2i} -family size

The least squares estimates of the regression parameters were

$$a = -0.30; \qquad b_1 = 2.32; \qquad b_2 = 1.41$$

a) the estimates b_1 and b_2

b) Is it possible to provide a meaningful interpretation of the estimate a ?

3. The following model was fitted to a sample of 20 students using data obtained at the end of the education year. The aim was to explain students' weight gains.

$$y_i = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \beta_3 \cdot x_{3i} + \varepsilon_i$$

where

y_i -weight gained, in kilograms, during the academic year

x_{1i} -average number of meals eaten per week

x_{2i} -average number of exercise per week, (in hours)

x_{3i} -average number of beers consumed per week

The least squares estimates of the regression parameters were

$$a = 12.9; \qquad b_1 = 4.5; \qquad b_2 = -6.3; \qquad b_3 = 3.14$$

a) Interpret the estimates b_1, b_2 and b_3

b) Is it possible to provide a meaningful interpretation of the estimate a ?

4. In the study of exercise 2, where the least squares estimates were based on 25 sets of sample observations, the following data were found

$$SST = 160.6 \quad \text{and} \quad SSR = 80.3$$

a) Find and interpret the coefficient of determination.

b) Find the adjusted coefficient of determination.

5. In the study of exercise 3, sample of 20 observations were used to calculate the least squares estimate. The regression sum of squares and error sum of squares were found to be

$$SST = 82.6 \quad \text{and} \quad SSE = 49..3$$

a) Find and interpret the coefficient of determination.

b) Find the adjusted coefficient of determination.

6. A multiple linear regression was fitted to a data set obtained from 27 runs of an experiment, in which four predictors x_1, x_2, x_3, and x_4 were observed along with the response y. The following results were obtained:

$$a = -5.46; \qquad b_1 = 2.35; \qquad b_2 = 18.4; \qquad b_3 = -0.91; \qquad b_4 = 6.2;$$

$$SSR = 920.60; \qquad SSE = 78.92$$

a) Predict response for $x_1 = 14; x_2 = 0.6; x_3 = 5; x_4 = 5.2$

b) Estimate the error standard deviation σ

c) What proportion of the y variability I explained by the fitted regression?

Answers

1. a) 79.78; b) 69.52; **4.** a) 0.5; b) 0.45; **5.** a) 0.4; b) 0.52; **6.** a) 66.17; b) 1.89; c) 0.92.

4.5 Computer solution of multiple regressions

Usually the calculations for a multiple regression model are made by using statistical software package for computers, such as MINITAB, instead of using the formula manually. In this chapter we will analyze the multiple regression models using MINITAB statistical software. The solutions obtained using other packages can be interpreted the same way.

Remark:

To use MINITAB menu follow the following instructions

1. Select Stat>Regression
2. Select Response column
3. Select Predictors columns
4. Click OK.

Example1:

Suppose that we want to find the effect of driving experience and the number of driving violations on auto insurance premiums. A random sample of 10 drivers insured with a company and having similar auto insurance policies was selected. Table 4.1 lists the yearly auto insurance premiums (in dollars) paid by these drivers, y, their driving experience (x_1, in years), and the number of driving violations that each of them has committed during the past five years.

Table 4.1

y	x_1	x_2
74	5	2
50	6	1
97	4	6
57	11	3
99	3	1
35	19	0
40	15	1
49	13	2
101	2	8
42	10	3

Use a computer package to perform a regression analysis using model

$$y_i = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \varepsilon_i$$

and answer the following questions:

a) Write the estimated regression equation;

b) Interpret the meaning of the estimated regression coefficients;

c) What are the values of the variance and standard deviation of errors, the coefficient of determination, and the adjusted coefficient of determination?

d) What is the predicted auto insurance premium paid per month by a driver with seven years of experience and four driving violations?

Solution:

Using MINITAB, we first enter the data of y, x_1, x_2 in three different columns and then use the regression command. The computer executes a multiple regression analysis. We focus our attention on the principal aspects of the output as shown in Figure 4.1

Figure 4.1

```
The regression equation is
Y= 87.9 - 3.39 X1 + 2.33 X2

Predictor      Coef      St. dev.      T        P
Constant      87.92        13.96      6.30    0.000
X1           -3.3869       0.9930    -3.41    0.011
X2            2.327        2.264      1.03     0.338

S = 13.77      R-SQ = 78.4%      R-SQ(adj) = 72.3%

Analysis of Variance

Source           DF       SS          MS          F        P
Regression        2      4824.5      2412.3      12.72    0.005
Residual Error    7      1327.9       189.7
Total             9      6152.4

Source    DF      SEQ. SS
X1         1       4624.1
X2         1        200.4
```

We now proceed to interpret the results in Figure 4.1 and use them to make further statistical inferences.

a) The equation of the fitted linear regression is

$$\hat{y} = 87.9 - 3.39 \cdot x_1 + 2.33 \cdot x_2$$

From this equation,

$a = 87.9$; $b_1 = -3.39$; $b_2 = 2.33$

We can also read the values of these coefficients from the column labeled COEF in the MINITAB solution of Figure 4.1.

Notice that in this column the coefficients appear with more digits after the decimal point. With these coefficient values, we can write the estimated regression equation as

$$\hat{y} = 87.92 - 3.3869 \cdot x_1 + 2.327 \cdot x_2$$

b) The value of $a = 87.92$ is in the estimated regression equation gives the value of $\hat{y}$ for $x_1 = 0$ and $x_2 = 0$. It means that a driver with no experience and no driving violations is expected to pay an auto insurance premium of $87.92 per year. This is the technical interpretation of a.

The value of $b_1 = -3.3869$ in the estimated regression model gives the change in $\hat{y}$ for a one unit change in x_1 when x_2 is held constant. Thus, we can state that a driver with one extra year of experience but with the same number of violations is expected to pay $3.3869 less for the auto insurance premium per year.

The value of $b_2 = 2.327$ in the estimated regression model gives change in for a one unit change in x_2 when x_1 is held constant. Thus, we can state that a driver with one extra driving violation but with the same years of driving experience is expected to pay $82.327 more per year for the auto insurance premium.

b) σ_e^2 is estimated by $s_e^2 = \dfrac{SSE}{n-K-1} = \dfrac{1327.9}{7} = 189.7$, so $s_e = 13.77$.

The values of the standard deviation of errors, the coefficient of determination and the adjusted coefficient of determination are also given in MINITAB solution. From Figure 4.1 we obtain

$s = s_e = 13.77$; R-SQ $= R^2 = 78.4\%$; R-SQ (adj) $= \bar{R}^2 = 72.3\%$

The value of $R^2 = 78.4\%$ tells us that the two independent variables included in our model explain 78.4% of the variation in the dependent variable.

The value of $\bar{R}^2 = 72.3\%$ is the value of the coefficient of determination adjusted for degrees of freedom. It states that when adjusted for degrees of freedom, the two independent variables explain 72.3% of the variation in the dependent variable.

c) To predict auto premium paid per year by a driver with seven years of experience and four driving violations, we substitute $x_1 = 7$ and $x_2 = 4$ in the estimated regression model

$$\hat{y} = 87.92 - 3.3869 \cdot x_1 + 2.327 \cdot x_2 = 87.92 - 3.3869 \cdot 7 + 2.327 \cdot 4 = 73.5197$$

Note that this value of $\hat{y}$ is a point estimate of the predicted value of y, which is denoted by y_p.

Remark:

In figure 4.1 there is portion of solution in the end reproduced below,

Source	DF	SEQ. SS
X1	1	4624.1
X2	1	200.4

which we have not used in any of the examples. From figure 4.1 we have

$$SSR = 4824.5$$

If we estimate the simple linear regression of y on x_1,

$$y = \alpha + \beta_1 \cdot x_1 + \varepsilon$$

the value of *SSR* will be 4624.1, which is the value in the row of **X1** and the column labeled SEQ.SS. That is, x_1 alone will reduce *SST* by 4624.1

Then, if we add x_2 to model above, the *SST* will further be reduced by 200.4, which is the value in the row of **X2** and the column labeled SEQ.SS.

The sum of the two numbers in the column of SEQ.SS is

$$4624.1 + 200.4 = 4824.5$$

which is the value of *SSR* in the Figure 4.1.

Example2:

We are interested in studying the blood pressure y of males in relation to weight x_1 and age x_2. Sample of 10 male was selected. The data set listed below:

y	x_1	x_2
120	76	60
160	84	45
134	95	37
149	99	46
153	74	49
164	83	70
130	92	38
170	110	54
148	80	28
125	79	19

Use a computer package to perform a regression analysis using model

$$y_i = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \varepsilon_i$$

Solution:

Using MINITAB, we first enter the data of y, x_1, x_2 in three different columns and then use the regression command. The computer executes a multiple regression analysis. We focus our attention on the principal aspects of the output as shown in Table 4.2

Table 4.2

The regression equation is

Y = 81.2 + 0.493 X1 + 0.474 X2

Predictor	Coef	St. dev.	T	P
Constant	81.17	43.50	1.87	0.104
X1	0.4929	0.4749	1.04	0.334
X2	0.4741	0.3641	1.30	0.234

S = 16.32 R-SQ = 30.2%

Analysis of Variance

Source	DF	SS	MS	F	
Regression	2	805.2	402.6	1.51	0.25
Residual Error	7	1864.9	266.4		
Total	9	2670.1			

Source	DF	SEQ. SS
X1	1	353.4
X2	1	451.8

We now proceed to interpret the results in table 4.2 and use them to make further statistical inferences.

a) The equation of the fitted linear regression is

$$\hat{y} = 81.2 + 0.493 \cdot x_1 + 0.474 \cdot x_2$$

This means that the mean blood pressure increases by 0.493 if weight x_1 increases by 1 kilogram and age x_2 remains fixed.

Similarly, a 1-year increase in age with the weight held fixed will increase the mean blood pressure by 0.474.

b) The estimated regression coefficients and the corresponding estimated standard errors are

$a = 81.17$ estimated standard error $S.E.(a) = 43.50$

$b_1 = 0.4929$ estimated standard error $S.E.(b_1) = 0.4749$

$b_2 = 0.4741$ estimated standard error $S.E.(b_2) = 0.3641$

Further, the error standard deviation σ estimated by $s = 16.32$ with degrees of freedom $= n -$ (number of variables) - 1 = 10 - 2 - 1 = 7.

These results are useful in interval estimation and hypothesis tests about the regression coefficients.

c) In Table 4.2, the result "$R - SQ = 30.2\%$" or $R^2 = 0.302$ tells us that 30.2% of the variability of y is explained by the fitted multiple regression of y on x_1 and x_2. The **analysis of variance** shows the decomposition of the total variability $\sum (y - \bar{y})^2 = 2670.1$ into the two components

2670.1	= 805.2	+	1864.9
Total variablity of y	Variability explained by the regression		Residual or unexplained variability

Thus,

$$R^2 = \frac{805.2}{2670.1} = 0.302$$

and σ^2 is estimated by $s^2 = \frac{1864.9}{7} = 266.41$, so $s = 16.32$.

4.6. Confidence interval for individual coefficients

The values of $a, b_1, b_2, \ldots\ldots$ and b_k obtained by estimating model (1) using sample data give the point estimates of $\alpha, \beta_1, \beta_2, \ldots\ldots$ and β_k, respectively, which are the population parameters. Using the values of the sample statistics $a, b_1, b_2, \ldots\ldots$ and b_k, we can make $100 \cdot (1-\alpha)\%$ confidence intervals for the corresponding population parameters $\alpha, \beta_1, \beta_2, \ldots\ldots$ and β_k, respectively.

If the population errors, ε_i, are normally distributed, then $100 \cdot (1-\alpha)\%$ confidence interval for the regression coefficients β_i, are given by

$$b_i - t_{n-K-1,\alpha/2} \cdot s_{b_i} < \beta_i < b_i + t_{n-K-1,\alpha/2} \cdot s_{b_i}$$

where $t_{n-K-1,\alpha/2}$ is the number for which

$$P(t_{n-K-1} > t_{n-k-1,\alpha/2}) = \frac{\alpha}{2}$$

and the random variable t_{n-K-1} follows a Student's t distribution with $(n-K-1)$ degrees of freedom.

Example:

Determine a 90% confidence interval for β_1 (the coefficient of experience) for the multiple regression of auto insurance premium on driving experience and the number of driving violations. Use the MINITAB solution of Fig. 4.1.

Solution:

The portion of the solution is shown below

Predictor	Coef	St. dev.	T	P
Constant	87.92	13.96	6.30	0.000
X1	-3.3869	0.9930	-3.41	0.011
X2	2.327	2.264	1.03	0.338

From the given information we obtain

$n = 10$; $b_1 = -3.3869$; and $s_{b_1} = 0.9930$

The confidence level is 90%. So,

$$t_{n-K-1,\alpha/2} = t_{10-2-1,0.05} = t_{7,0.05} = 1.895$$

The 90% confidence interval for β_1 is

$$b_1 - t_{n-K-1,\alpha/2} \cdot s_{b_1} < \beta_1 < b_1 + t_{n-K-1,\alpha/2} \cdot s_{b_1}$$

$$-3.3869 - 1.895 \cdot 0.9930 < \beta_1 < -3.3869 + 1.895 \cdot 0.9930$$

$$-5.269 < \beta_1 < -1.505$$

Thus, the 90% confidence interval for β_1 is -5.269 to -1.505. That is, we can state with 90% confidence that for one extra year of driving experience, the yearly auto insurance premium decreases by an amount between \$1.505 and \$5.269.

Exercises

1. In the study of exercise 2 of the previous chapter, where the sample regression was based on 25 observations, the estimated standard errors were $s_{b_1} = 0.089$; $s_{b_2} = 0.45$;

a) Find 90% and 95% confidence intervals for β_1

b) Find 95% and 99% confidence intervals for β_2

2. In the study of exercise 3 of the previous chapter, where the sample regression was based on 20 observations, the estimated standard errors were $s_{b_1} = 0.18$; $s_{b_2} = 0.64$;

Find 90%, 95% and 99% confidence intervals for β_1

3. Following is the MINITAB solution for a regression of y on x_1 and x_2.

The regression equation is

Y = 12.4 + 0.24 X1 + 0.036 X2

Predictor	Coef	St. dev.	T	P
Constant	12.410	5.234	2.12	0.060
X1	0.2415	0.0345	7.47	0.000
X2	0.0362	0.024	-1.60	0.140

S = 1.76 R-SQ = 97.8% R-SQ(adj) = 96.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	849.65	424.83	130.10	0.000
Residual Error	10	32.65	3.27		
Total	12	882.30			

Source	DF	SEQ. SS
X1	1	841.25
X2	1	8.40

Using the MINITAB solution, answer the following questions for the population regression model $y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \varepsilon$

a) Write the estimated regression equation.

b) Write the values of a, b_1, b_2 and explain the meaning of these estimated regression coefficients.

c) Write the values of the standard deviation of the coefficients of a, b_1, b_2.

d) What are the values of the variance and standard deviations of errors, the coefficient of determination, the adjusted coefficient of determination, *SST*, *SSR*, *SSE*, *MSR*, and *MSE*?

e) What is the predicted value of y for $x_1 = 74$ and $x_2 = 140$?

f) Construct a 99% confidence interval for the coefficient of x_1 in the population regression model.

g) Make a 95% confidence interval for the coefficient of x_2 in the population regression model.

h) Determine a 90% confidence interval for α, the constant term in the population regression model.

4. The following is the MINITAB solution for a regression of y on x_1, x_2 and x_3.

Regression Analysis: Y versus X1, X2, X3

The regression equation is

Y = 22.2 + 0.203 X1 - 0.0499 X2 - 0.216 X3

Predictor	Coef	SE Coef	T	P
Constant	2.212	4.602	4.83	0.003
X1	0.20276	0.06171	3.29	0.017
X2	-0.04991	0.04166	-1.20	0.276

X3	-0.21648	0.04836	-4.48	0.004

S = 1.016 R-SQ = 98.6% R-SQ(adj) = 97.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	422.21	140.74	136.44	0.000
Residual Error	6	6.19	1.03		
Total	9	428.40			

Source	**DF**	**SEQ SS**
X1	1	401.42
X2	1	0.12
X3	1	20.67

Using the MINITAB solution, answer the following questions for the population regression model $y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \varepsilon$

a) Write the estimated regression equation.

b) Write the values of $a, b_1, b_2,$ b_3 and explain the meaning of these estimated regression coefficients.

c) Write the values of the standard deviation of the coefficients of $a, b_1, b_2,$ b_3.

d) What are the values of the variance and standard deviations of errors, the coefficient of determination, the adjusted coefficient of determination, *SST*, *SSR*, *SSE*, *MSR*, and *MSE*?

e) What is the predicted value of y for $x_1 = 33$, $x_2 = 50$ and $x_3 = 60$?

f) Construct a 95% confidence interval for the coefficient of x_1 in the population regression model.

g) Make a 90% confidence interval for the coefficient of x_2 in the population regression model.

h) Determine a 99% confidence interval for x_3, in the population regression model.

5. In a study of revenue generated by national lotteries, the following regression equation was fitted to a data from 26 countries with lotteries:

$y = -30.29 + 0.0354 \cdot x_1 + 0.9734 \cdot x_2 - 340.9524 \cdot x_3$

(0.00652) (0.3210) (225.78)

$R^2 = 0.56\%$

where

y – dollars of net revenue per capita per year generated by lottery;

x_1 – mean per capita personal income of the country

x_2 – number of hotel, motel, resort rooms per thousand of people

x_3 – spendable revenue per capita per year generated by legalized gambling

The numbers in parentheses below the coefficient estimates are the corresponding estimated standard errors.

a) Interpret the estimated coefficient on x_1, x_2 and x_3.

b) Find and interpret a 90% confidence interval for the coefficient on x_2, in the population regression.

c) Find and interpret a 99% confidence interval for the coefficient on x_3, in the population regression.

Answers

1. a) 2.167 to 2.473; 2.135 to 2.504; b) 0.48 to 2.34; 0.142 to 2.678; **2.** 4.19 to 4.814; 4.12 to 4.88; 3.98 to 5.02; **3.** a) $\hat{y} = 12.4 + 0.24 \cdot x_1 + 0.036 \cdot x_2$; b) $a = 12.410$; $b_1 = 0.2415$; $b_2 = 0.0362$; c) $s_a = 5.234$; $s_{b_1} = 0.0345$; $s_{b_2} = 0.024$; d) $s_e^2 = 3.0976$; $s_e = 1.76$; $R^2 = 97.8\%$; $\bar{R}^2 = 96.6\%$; $SST = 882.30$; $SSR = 849.65$; $SSE = 32.65$; $MSR = 424.83$; $MSE = 3.27$; e) $y_p = 35.2$; f) 0.132 to 0.348; g) -0.017 to 0.089; h) 2.93 to 21.89;

4. a) $\hat{y} = 22.2 + 0.203 \cdot x_1 - 0.0499x_2 - 0.216x_3$; b) $a = 2.212$; $b_1 = 0.20276$; $b_2 = -0.04991$; $b_3 = -0.21648$; c) $s_a = 4.602$; $s_{b_1} = 0.06171$; $s_{b_2} = 0.04166$; $s_{b_3} = 0.04836$; d) $s_e^2 = 1.032$; $s_e = 1.016$; $R^2 = 98.6\%$; $\bar{R}^2 = 97.8\%$; $SST = 428.40$; $SSR = 422.21$; $SSE = 6.19$; $MSR = 140.74$; $MSE = 1.03$; e) $y_p = 13.444$; f) 0.051 to 0.355; g) -0.132 to 0.031; h) -0.394 to -0.038;

5. b) 0.422 to 1.524 c) -977.42 to 295.52.

4.7. Test of hypothesis about individual coefficients

We can make a test of hypothesis about any of the β_i coefficients of model

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + ... + \beta_k \cdot x_k + \varepsilon$$

Using the same procedure that we used to make a test of hypothesis about β for a simple regression model in previous chapter. The only difference is the degrees of freedom, which are equal to $(n - K - 1)$ for a multiple regression. In this case the value of the test statistic t for b_i is calculated as

$$T.S. = t = \frac{b_i - \beta_i}{s_{b_i}}$$

The value of β_i is substituted from the null hypothesis.

If the regression errors ε_i are normally distributed and the standard regression assumptions hold, then the following hypothesis tests have significance level α

1. To test either null hypothesis

$$H_0 : \beta_i = \beta_0 \text{ or } H_0 : \beta_i \leq \beta_0$$

against the alternative

$$H_1 : \beta_i > \beta_0$$

the decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > t_{n-K-1,\alpha}$$

2. To test either null hypothesis

$$H_0 : \beta_i = \beta_0 \text{ or } H_0 : \beta_i \geq \beta_0$$

against the alternative

$$H_1 : \beta_i < \beta_0$$

the decision rule is

$$\text{Reject } H_0 \text{ if } T.S. < -t_{n-K-1,\alpha}$$

3. To test null hypothesis

$$H_0 : \beta_i = \beta_0$$

against the two sided alternative

$$H_1 : \beta_i \neq \beta_0$$

the decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > t_{n-K-1,\alpha/2} \text{ or } T.S. < -t_{n-K-1,\alpha/2}$$

Remark: In most cases we are interested in the null hypothesis $H_0 : \beta_i = 0$.

Example:

For example 1 of the section 4.5, using 1% significance level, can you conclude that the slope of the number of driving violations in regression model is 0 against the alternative that it is positive? Use the MINITAB solution given in Figure 4.1.

Solution:

x_2 is the number of driving violations committed during the past five years.

The portion of the solution is reproduced below

Predictor	Coef	St. dev.	T	P
Constant	87.92	13.96	6.30	0.000
X1	-3.3869	0.9930	-3.41	0.011
X2	2.327	2.264	1.03	0.338

We are to test the following null and alternative hypotheses

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 > 0$$

The decision rule is

reject H_0 if $T.S. > t_{n-K-1,\alpha}$

From solution we obtain that $t = T.S. = 1.03$. It also can be found as

$$T.S. = t = \frac{b_i - \beta_i}{s_{b_i}} = \frac{2.327 - 0}{2.264} = 1.03$$

$$d.f. = n - K - 1 = 10 - 2 - 1 = 7$$

$$t_{n-K-1,\alpha} = t_{7,0.01} = 2.998$$

Since $1.03 < 2.998$, we accept the null hypothesis. Consequently, we conclude that the slope of x_2 in regression model is zero. That is, the number of driving violations is not significant and an increase (or decrease) in the number of driving violations does not affect the auto insurance premium.

Remark:

Note that the observed value of test statistic T (test statistic t) is obtained from the MINITAB solution only if the null hypothesis is $H_0 : \beta_2 = 0$. However, if the null hypothesis is that β_2 is equal to a number other than zero, then the t value obtained from the MINITAB solution is no longer valid. In this case observed value of the test statistic will be calculated as

$$T.S. = t = \frac{b_2 - \beta_2}{s_{b_2}}$$

4.8. Tests on sets of regression parameters

In the previous section we developed a hypothesis test for individual regression parameters. There are situations, where we are interested in the effect of the combination of several variables. Now we will perform a test of hypothesis with the null hypothesis that the coefficients of all independent variables in the regression model are equal to zero and the alternative hypothesis that the coefficients of all independent variables are not zero. For the multiple regression model

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + ... + \beta_k \cdot x_k + \varepsilon$$

the two hypotheses for such test are written as

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0$$

$$H_1 : \text{at least one } \beta_t \neq 0$$

A test of hypothesis for this case is performed by using the F distribution. The F distribution has 2 degrees of freedom- df for numerator and - df for denominator. Table 6 in Appendix lists the values of F for F distribution. The value of the test statistic $T.S. = F$ can be obtained from the computer solution, or it can also be calculated by using formula

$$T.S. = F = \frac{SSR / K}{SSE /(n - K - 1)} \quad \text{or}$$

$$T.S. = F = \frac{MSR}{MSE}$$

where MSR stands for the mean square regression and MSE for the mean square error.

$$MSR = \frac{SSR}{K}; \qquad MSE = \frac{SSE}{(n - K - 1)}$$

In the end, to test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0$$

against the alternative hypothesis

$$H_1 : \text{at least one } \beta_t \neq 0$$

the decision rule for a significance level α is

reject H_0 if $T.S. = F > F_{K,n-K-1,\alpha}$

where $F_{K,n-K-1,\alpha}$ is the number for which $P(F_{K,n-K-1} > F_{K,n-K-1,\alpha}) = \alpha$ and $F_{K,n-K-1}$ follows an F distribution with numerator degrees of freedom K and denominator degrees of freedom $(n-K-1)$.

Example:

Using 5% significance level, can you conclude that the coefficients of all independent variables in the example 4.1 are equal to zero? Use the MINITAB solution shown in Figure 4.1

Solution:

The two hypotheses are

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

The portion of the solution is reproduced below

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	4824.5	2412.3	12.72	0.005
Residual Error	7	1327.9	189.7		
Total	9	6152.4			

From the portion of MINTAB solution we obtain

$MSR = 2412.3$; $MSE = 189.7$

and the value of the test statistic is $T.S. = F = 12.07$

$$F_{K,n-K-1,\alpha} = F_{2,10-2-1,0.05} = F_{2,7,0.05} = 4.74$$

Because the value of the test statistic $T.S. = F = 12.07$ greater than 4.74, it falls in the rejection region. Consequently, we reject the null hypothesis and conclude that at least one of the two β 's is different from zero.

Exercises

1. Refer to the study on milk consumption, described in exercise 2 and 4 after the section 4.4.4. Test the null hypothesis that $\beta_1 = \beta_2 = 0$

2. Refer to the study on weight gains, described in exercises 3 and 5 after the section 4.4.4. Test the null hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$

3. The following is the MINITAB solution for a regression of y on x_1, x_2 and x_3.

The regression equation is

Y = 51.6 - 0.0599 X1 + 0.0850 X2 - 0.00477 X3

Predictor	Coef	SE Coef	T	P
Constant	51.61	11.25	4.59	0.000
X1	-0.05993	0.01526	-3.93	0.003
X2	0.08497	0.02875	2.96	0.014
X3	-0.004773	0.008707	-0.55	0.596

S = 0.8995 R-SQ = 99.7% R-SQ(adj) = 99.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	2909.91	969.97	1198.86	0.000
Residual Error	10	8.09	0.81		
Total	13	2918.00			

Source	DF	SEQ SS
X1	1	2901.82
X2	1	7.85
X3	1	0.24

Using the MINITAB solution, answer the following questions for the population regression model $y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \varepsilon$

a) Write the estimated regression equation.

b) Write the values of a, b_1, b_2, b_3 and explain the meaning of these estimated regression coefficients.

c) What are the values of the standard deviation of errors, the coefficient of determination, the adjusted coefficient of determination, *SST*, *SSR*, *SSE*, *MSR*, and *MSE*?

d) Write the values of the standard deviation, the value of test statistic, and the *p-value* for each of the coefficients of a, b_1, b_2, b_3.

e) What is the predicted value of y for $x_1 = 310$, $x_2 = 260$ and $x_3 = 180$?

f) Construct a 95% confidence interval for the coefficient of x_1 in the population regression model.

g) Make a 99% confidence interval for the coefficient of x_2 in the population regression model.

h) Make a 98% confidence interval for the coefficient of x_3 in the population regression model.

i) Determine a 95% confidence interval for α, the constant term in the population regression model.

j) Using the 5% significance level, test the null hypothesis that the coefficient of x_1 in the population regression model is zero against the alternative that it is negative.

k) Using the 1% significance level, can you conclude that the coefficient of x_2 in the population regression model is zero against the alternative that it is positive?

l) At the 2.5% significance level, test if the coefficient of x_3 in the population regression model is zero against the alternative that it is negative.

m) Using the 5% significance level, can you conclude that the coefficients of all independent variables in the population regression model are equal to zero?

4. The Corporation has a large number of restaurants through the country. The research department wanted to find if the sales of the restaurants depend on the size of the population within a certain area surrounding the restaurants and the mean income of households in those areas. They collected information on these variables for 10 restaurants. The following table gives information on the monthly sales (in thousands of dollars) of these restaurants, the population (in thousands) within 10 kilometers of the restaurants, and means monthly income (in hundreds of dollars) of the households of those areas.

Sales	18	28	16	20	13	29	34	23	19	28
Population	22	16	33	19	47	70	30	45	77	41
Income	39	51	28	32	28	37	42	27	20	18

Using MINITAB (or any other statistical software package), find the regression of sales on a population and income. Using solution, answer the following questions.

a) Write the estimated regression equation.

b) Explain the meaning of the estimates of the constant term and regression coefficients of the population and income.

c) What are the values of the standard deviation of errors, the coefficient of determination, the adjusted coefficient of determination?

d) What are the value of the total sum of squares? What portion of *SST* is explained by our regression model? What portion of *SST* is not explained by our regression model?

e) What is the predicted sales for a restaurant with 52 thousand people living within 10 km surrounding it and $3600 mean monthly income of households living in those areas?

f) Construct a 95% confidence interval for the coefficient of ***income***.

g) Using the 5% significance level, test the null hypothesis that the coefficient of ***population*** in regression model is zero against the two-sided alternative.

h) Using the 1% significance level, can you conclude that the coefficients of both independent variables in the population regression model are equal to zero?

Answers

1. $T.S.=11$; reject H_0 virtually at any level; 2. $T.S.=3.06$; reject H_0 at 5% level; 3. a) $\hat{y}=51.6-0.0599\cdot x_1+0.0850\cdot x_2-0.0048\cdot x_3$; b) $a=51.61$; $b_1=-0.05993$; $b_2=0.08497$; $b_3=-0.004773$; c) $s_e=0.8995$; $R^2=99.7\%$; $\bar{R}^2=99.6\%$; $SST=2918.00$; $SSR=2909.91$; $SSE=8.09$; $MSR=969.97$; $MSE=0.81$; d) $s_a=11.25$; $t_a=4.59$; $p_a=0.000$; $s_{b_1}=0.01526$; $t_{b_1}=-3.93$; $p_{b_1}=0.003$; $s_{b_2}=0.02875$; $t_{b_2}=2.96$; $p_{b_2}=0.014$; $s_{b_3}=0.008707$; $t_{b_3}=-0.55$; $p_{b_3}=0.596$; e) $y_p=54.2724$; f) -0.09393 to -0.02593; g) -0.00614 to 0.17608; h) -0.028839 to 0.019293; i) 26.545 to 76.675; j) $T.S.=-3.927$; reject H_0; k) $T.S.=2.955$; reject H_0; l) $T.S.=-0.548$; accept H_0; m) $T.S.=F=1198.86$; reject H_0

4.9. Dummy variables in the regression models

In the discussion of multiple regression we have assumed that the independent variables, x_i, have existed over a range and contained many different values. All independent variables we have considered were quantitative. We may include in regression model a variable that is qualitative. Such a variable contains different categories instead of numerical values. We will introduce independent variable that will take only two values: 0 and 1. This structure is commonly defined as a "dummy variable", and we will see that it provides a valuable tool for applying multiple regression to situations involving categorical variables.

Let us consider a simple regression equation

$$y = \alpha + \beta_1 \cdot x_1$$

Now suppose that we introduce a dummy variable, x_2, that has values 0 and 1 and the resulting equation becomes

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$$

When $x_2 = 0$ in this equation the constant is α, but when $x_2 = 1$ the constant is $\alpha + \beta_2$. Thus we see that the dummy variable shift the linear relationship between y and x_1 by the value of the coefficient β_2.

The number of dummy variables in a regression model is equal to the number of categories minus 1. For instance, if a variable contains two categories, then we introduce one dummy variable in the regression model for this variable. If a qualitative variable contains three categories, we will introduce two dummy variables and so on.

The following example shows how a dummy variable is used in regression model.

Example:

Refer to example 1. Following table reproduces the data from that example with additional column that contains information for each of the 10 drivers.

Yearly premium y	Driving experience x_1	Number of violations (past 5 years) x_2	Gender
74	5	2	Male
50	6	1	Female
97	4	6	Female
57	11	3	Female
99	3	1	Female
35	19	0	Male
40	15	1	Female
49	13	2	Female
101	2	8	Male
42	10	3	Male

Using MINITAB, find the regression of yearly auto insurance premium on the years of experience, the number of driving violations, and the gender of drivers. Answer the following questions

a) Write the estimated regression equation.

b) Explain the meaning of the estimated regression coefficient of the independent variable *gender*.

c) What is the predicted auto insurance premium paid per year by a male driver with 14 years of driving experience and 3 driving violations?

d) What is the predicted auto insurance premium paid per year by a female driver with 14 years of driving experience and 3 driving violations?

e) Construct a 99% confidence interval for the coefficient of *gender*.

f) Using 1% significance level, test the null hypothesis that the coefficient of *gender* is zero.

Solution:

Gender is not a quantitative variable, it is a qualitative variable. So, we will use a dummy variable for it in regression model. Let

x_1 – driving experience (in years)

x_2 – number of driving violations (during past 5 years)

We can denote dummy variable by x_3. Also we can denote it by letter D.

Suppose

$$D = \begin{cases} 0 & \text{if a driver is a male} \\ 1 & \text{if a driver is a female} \end{cases}$$

In this case, our population regression model becomes

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot D + \varepsilon$$

Assuming values of 0 and 1 to male and female respectively, we rewrite the data

Yearly premium y	Driving experience x_1	Number of violations (past 5 years) x_2	Gender
74	5	2	0
50	6	1	1
97	4	6	1
57	11	3	1
99	3	1	1
35	19	0	0
40	15	1	1
49	13	2	1
101	2	8	0
42	10	3	0

The following figure shows the MINITAB solution

Regression Analysis: y versus X1, X2, D

The regression equation is

y = 84.5 - 3.32 X1 + 2.56 X2 + 3.57 D

Predictor	**Coef**	**St.dev.**	**T**	**P**
Constant	84.54	17.58	4.81	0.003
X1	-3.318	1.078	-3.08	0.022
X2	2.559	2.502	1.02	0.346
D	3.573	9.829	0.36	0.729

S = 14.72 R-Sq = 78.9% R-Sq(adj) = 68.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	4853.1	1617.7	7.47	0.019
Residual Error	6	1299.3	216.5		
Total	9	6152.4			

Source	**DF**	**Seq SS**
X1	1	4624.1
X2	1	200.4
D	1	28.6

a) The estimated regression equation is

$$\hat{y} = 84.5 - 3.32 \cdot x_1 + 2.56 \cdot x_2 + 3.57 \cdot D \qquad (1)$$

We also can use column labeled COEF and write the estimated regression equation as

$$\hat{y} = 84.54 - 3.318 \cdot x_1 + 2.559 \cdot x_2 + 3.573 \cdot D \qquad (2)$$

b) The coefficient of the variable *gender* is $b_3 = 3.57$. It indicates that the female drivers pay, an average, \$3.573 more than male drivers with similar driving experiences and the same number of driving violations.

In fact, by using dummy variable D in our regression model, we have estimated two regression models: one for male drivers, and another for the female drivers. Since for male drivers $D = 0$, after substituting it into the estimated regression model we find the estimated model for male drivers as

$$\hat{y} = 84.54 - 3.318 \cdot x_1 + 2.559 \cdot x_2$$

For female drivers $D = 1$. Substituting it in the regression model, we obtain the estimated regression model for the female drivers

$$\hat{y} = 84.54 - 3.318 \cdot x_1 + 2.559 \cdot x_2 + 3.573 \cdot (1) =$$

$$= 88.113 - 3.318 \cdot x_1 + 2.559 \cdot x_2$$

We see that, the constant term for female drivers is 3.573 greater than that for the male drivers' model. Thus, on average, female drivers pay a yearly auto insurance premium that is \$3.573 more than the yearly auto insurance premium paid by male drivers with similar driving experiences and the same number of driving violations.

c) To find the predicted auto insurance premium for a male driver with 14 years of driving experience and 3 driving violations, we substitute $x_1 = 14$, $x_2 = 3$, and $D = 0$ in the estimated regression model (2),

$$\hat{y} = 84.54 - 3.318 \cdot 14 + 2.559 \cdot 3 + 3.573 \cdot 0 = 45.765 = \$45.765$$

Thus, a male driver with 14 years of driving experience and 3 driving violations is expected to pay a yearly auto insurance premium of \$45.765.

d) To find the predicted auto insurance premium for a female driver with 14 years of driving experience and 3 driving violations, we substitute $x_1 = 14$, $x_2 = 3$, and $D = 1$ in the estimated regression model (2),

$$\hat{y} = 84.54 - 3.318 \cdot 14 + 2.559 \cdot 3 + 3.573 \cdot 1 = 49.338 = \$49.338$$

Thus, a female driver with 14 years of driving experience and 3 driving violations is expected to pay a yearly auto insurance premium of \$49.338.

e) We are to make a 99% confidence interval for β_3. From the given information and from the MINITAB solution we obtain

$n = 10$; $b_3 = 3.573$; and $s_{b_3} = 9.829$

$$t_{n-K-1,\alpha/2} = t_{10-3-1,0.005} = t_{6,0.005} = 3.707$$

So, from

$$b_3 - t_{n-K-1,\alpha/2} \cdot s_{b_3} < \beta_3 < b_3 + t_{n-K-1,\alpha/2} \cdot s_{b_3}$$

a 99% confidence interval for β_3 is

$$3.573 - 3.707 \cdot 9.829 < \beta_3 < 3.573 + 3.707 \cdot 9.829$$

$$-32.863 < \beta_3 < 40.009$$

Thus, the 99% confidence interval for β_3 is -\$32.863 to \$40.009. We can state with 99% confidence that female drivers pay somewhere between \$32.863 less than to \$40.009more than male drivers with similar values for the x_1 and x_2 variables.

f) We are to test whether or not the coefficient β_3 of *gender* in model (1) is zero. The two hypotheses are

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

The decision rule is

Reject H_0 if $T.S. = t > t_{n-K-1,\alpha/2}$ or $T.S. = t < -t_{n-K-1,\alpha/2}$

From MINITAB solution we find that the value of test statistic is

$$T.S. = t = 0.36$$

$t_{n-K-1,\alpha/2} = t_{10-3-1,0.005} = t_{6,0.005} = 3.707$ and

$$-t_{n-K-1,\alpha/2} = -3.707$$

Since 0.36 is not greater than 3.707, the value of test statistic falls in the non rejection region. Consequently, we fail to reject the null hypothesis and conclude that β_3 in regression model is not different from zero. That is, the variable *gender* has no effect on the auto insurance premiums paid by drivers.

Remark:

The number of dummy variables used for qualitative variable in a regression model is one less than the number of categories for that variable. For example, we may want to investigate influence of quarters. Because the variable *quarter* is a qualitative variable, we will use dummy variables to represent it in our regression model. Since there are 4 quarters in a year, we will use 3 dummy variables. Let D_1 – be the dummy variable for the first quarter, D_2 – be the dummy variable for the second quarter and D_3 – be the dummy variable for the third quarter. Then

$D_1 = 1$ for the first quarter, and zero for other quarters

$D_2 = 1$ for the second quarter, and zero for other quarters

$D_3 = 1$ for the third quarter, and zero for other quarters

If our regression model consists of two independent variables x_1 and x_2, then we will estimate regression model as

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot D_1 + \beta_4 \cdot D_2 + \beta_5 \cdot D_3 + \varepsilon$$

Exercises

1. The following model was fitted to data on 28 insurance companies.

$$y = 6.85 - 0.12 \cdot x_1 + 2.13 \cdot D$$
$$\qquad\qquad (0.006) \quad (0.532)$$

where

y – price-earning ratio

x_1 – size of insurance company assets, in millions of dollars

D – dummy variable, taking the value 1 for foreign companies, and 0 for national companies

a) Interpret the estimated coefficient on the dummy variable

b) Test against a two-sided alternative the null hypothesis that the true coefficient on the dummy variable is zero. Take $\alpha = 0.05$

2. The following model was fitted, to explain the selling prices of home, to a sample of 815 sales.

$$y = -1264 + 48.18 \cdot x_1 + 3382 \cdot x_2 + 3219 \cdot x_3 + 2005 \cdot x_4 \qquad \bar{R}^2 = 0.86$$
$$(0.91) \quad (515) \quad (947) \quad (768)$$

where

y – selling price of home, in thousands of dollars

x_1 – square meters of living area

x_2 – size of garage, in square of meters

x_3 – dummy variable taking the value 1 if the house has a fireplace, and 0 otherwise

x_4 – dummy variable taking the value 1 if the house has a wood floors, and 0 otherwise

a) Interpret the estimated coefficient of x_3.

b) Interpret the estimated coefficient of x_4.

c) Find a 95% confidence interval for the impact of fireplace on a selling price, all other being equal.

d) Test the null hypothesis that type of flooring has no impact on selling price, against the alternative that, all other things equal, house with wood floors have a higher selling price than other flooring.

3. The following MINITAB solution was obtained for the regression model

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot D + \varepsilon$$

for a sample data set.

The regression equation is

y = 22.3 + 0.190 X1 - 0.0533 X2 - 0.190 X3 - 1.93 D

Predictor	Coef	SE Coef	T	P
Constant	22.315	2.802	7.96	0.001
X1	0.19019	0.03776	5.04	0.004
X2	-0.05325	0.02538	-2.10	0.090
X3	-0.19033	0.03046	-6.25	0.002
D	-1.9349	0.5785	-3.34	0.020

```
S = 0.6183        R-Sq = 99.6%      R-Sq(adj) = 99.2%

Analysis of Variance
Source           DF      SS        MS         F       P
Regression        4    426.49    106.62    278.87   0.000
Residual Error    5      1.91      0.38
Total             9    428.40

Source        DF       Seq SS
X1             1       401.42
X2             1         0.12
X3             1        20.67
D              1         4.28
```

Using MINITAB solution, answer the following questions.

a) Write the estimated regression equation.

b) Explain the meaning of b_4 obtained by estimating the given regression model.

c) What is the predicted value of y for $x_1 = 45$; $x_2 = 40$; $x_3 = 60$; and $D = 0$?

d) Construct a 99% confidence interval for the coefficient of D.

e) Using the 5% significance level, test the null hypothesis that the coefficient of D is zero against two-sided alternative.

4. The salaries of workers are expected to be depending on the number of years they have spent in school and their work experience. The following table gives information on annual salaries (in thousands of dollars), number of years of studying, and years of experience for 10 persons. It also includes information on gender. In the table, M represents males and F refers to females.

Salary	30	22	21	45	36	39	17	22	18	19
Studying	18	16	15	22	20	20	14	16	12	14
Experience	8	7	6	15	14	16	2	4	3	4
Gender	F	F	M	M	F	F	M	M	F	M

Using MINITAB (or any other statistical software package), find the regression of salary on studying, experience, and gender. Then answer the following questions.

a) Write the estimated regression equation.
b) Explain the meaning of the estimated regression coefficient of the dummy variables.
c) By estimating the regression model with gender as a dummy variable, you have actually estimated two regression models-one for males and the other for females. Write these two regression equations.
d) How much salary is a male worker with 18 years of studying and 7 years of work experience expected to earn?
e) How much salary is a female worker with 18 years of studying and 7 years of work experience expected to earn?
f) Determine a 95% confidence interval for the coefficient of dummy variable.
g) Using the 5% significance level, can you conclude that female workers are paid lower salaries than male workers?

Answers

1. a) All else being equal, expected price-earning ratio is higher by 2.13 million of dollars for foreign companies; b) $T.S. = t = 4$; reject H_0;

2. a) All else being equal, expected selling price is higher by \$3219 if house has a fireplace; b) All else being equal, expected selling price is higher by \$2005 if house has a wood floor; c) $1363 < \beta_3 < 5075$; d) $T.S. = 2.611$; can reject H_0 at 0.5%; **3.** a) $y = 22.3 + 0.190 \cdot x_1 - 0.0533 \cdot x_2 - 0.190 \cdot x_3 - -1.93 \cdot D$; c) 17.32375; d) -0.4025 to 4.2625; e) $T.S. = t = -3.34$; reject H_0.

Chapter 5

Analysis of variance (ANOVA)

5.1. Introduction

In Chapter 1 we discussed how to test whether or not the means of two populations are equal. Recall that the test involved the selection of an independent random sample from each of the populations. In this chapter we will discuss a statistical procedure for determining whether or not the means of more than two populations are equal. The technique that we will be introduced is called the analysis of variance (ANOVA) procedure.

5.2. One-way analysis of variance

This section discusses the **one-way analysis of variance** procedure to make tests comparing means often is called one-way analysis, because we will analyze only one factor. Sometimes we may analyze the effects of two factors. This is called two-way analysis of variance.

Suppose that we have independent samples of sizes $n_1, n_2, n_3, ...n_K$ observations selected randomly from K populations. Assume that population means are $\mu_1, \mu_2, \mu_3, ...\mu_K$. The **one-way analysis of variance** procedure designated to test the null hypothesis

$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_K$ (All population means are equal)

against the alternative hypothesis

H_1 : At least one of the population has different mean

The application of one way analysis of variance requires that the following assumptions hold:

1. The populations from which the samples are selected are (approximately) normal.
2. The populations from which the samples are selected have the same variance (or standard deviation).
3. The samples selected from different populations are independent.

Suppose that from K populations samples of sizes $n_1, n_2, n_3, ...n_K$ are selected (Figure 5.1)

Figure 5.1

Block	POPULATION (GROUP)		
1	2	....	K
x_{11}	x_{21}	....	x_{K1}
x_{12}	x_{22}	....	x_{K2}
....	....	....	....
....	....	....	.....
x_{1n_1}	x_{2n_2}	....	x_{Kn_K}

1) The first step is to calculate the sample mean for the K groups of observations. These sample means will be denoted as $\bar{x}_1, \bar{x}_2, \ldots\ldots \bar{x}_K$. In general

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$$

where n_i denotes the number of observations in i^{th} group.

2) The second step is to find overall mean of the all sample observations, denoted $\bar{x}$, and defined as

$$\bar{x} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} x_{ij}}{n}$$

where n denotes the total number of sample observations

$$n = \sum_{i=1}^{K} n_i$$

An equivalent expression for overall mean is

$$\bar{x} = \frac{\sum_{i=1}^{K} n_i \bar{x}_i}{n}$$

3) In third step, we consider *variability within-groups*. To measure variability in the any group, we calculate the sum of squared deviations of the observations about their sample means. Within-groups variability will be denoted by SS. For example, for the first group the sum of squared deviations of the observations about their sample mean $\bar{x}_1$ is

$$SS_1 = \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2$$

For the second group, whose sample mean is $\bar{x}_2$, we calculate

$$SS_2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2$$

and so on.

4) In fourth step, we find *total within-groups* variability, denoted SSW. That is

$$SSW = SS_1 + SS_2 + \ldots\ldots + SS_K$$

or

$$SSW = \sum_{i=1}^{K}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

5) Now we need a measure of *variability between groups*. It is based on the discrepancies between the individual group means and the overall mean. Total between-groups sum of squares denoted SSG, and defined as

$$SSG = \sum_{i=1}^{K} n_i \cdot (\bar{x}_i - \bar{x})^2$$

6) As a last step, we calculate the sum of squared discrepancies of all the sample observations about their overall mean. This is called the *total sum of squares*, denoted SST, expressed as

$$SST = \sum_{i=1}^{K}\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

It can be shown, that the total sum of squares is the sum of the within-groups and between-groups sum of squares, that is

$$SST = SSW + SSG$$

Testing the equality of population means is based on the assumption that K populations have equal variances (or standard deviations).

If $H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_K$

is true, each of the *SSW* and *SSG* can be used as the basis for estimate of the common population variance. To obtain these estimates, the sums of squares must be divided by the corresponding numbers of degrees of freedom.

SSW divided by $(n-K)$ results estimate called the *within-groups mean square*, denoted *MSW*, so that

$$MSW = \frac{SSW}{n-K}$$

SSG divided by $(K-1)$ results estimate called the *between-groups mean square*, denoted *MSG*, so that

$$MSG = \frac{SSG}{K-1}$$

The test of null hypothesis is based on the ratio of the mean squares

$$F = \frac{MSG}{MSW}$$

If this ratio is close to 1, there would be little cause to doubt the null hypothesis of equality of population means.

Summary

We define the following **sums of squares:**

Within-groups: $SSW = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$

Between groups: $SSG = \sum_{i=1}^{K} n_i \cdot (\bar{x}_i - \bar{x})^2$

Total: $SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$

We define the **mean squares** as follows:

Within-groups: $MSW = \frac{SSW}{n-K}$

Between groups: $MSG = \frac{SSG}{K-1}$

The null hypothesis is

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

The decision rule is

$$\text{Reject } H_0 \text{ if } T.S. = F = \frac{MSG}{MSW} > F_{K-1,n-K,\alpha}$$

where $F_{K-1,n-K,\alpha}$ is the number for which $P(F_{K-1,n-K} > F_{K-1,n-K,\alpha}) = \alpha$ and $F_{K-1,n-K}$ follows an F distribution with numerator degrees of freedom (K-1) and denominator degrees of freedom $(n-K)$. (Table 6 of Appendix).

For convenience, these calculations are often recorded in a table called a *one-way analysis of variance table* or ANOVA table, shown below (Table5.1):

Table 5.1

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
Between groups	SSG	$(K-1)$	MSG	$\frac{MSG}{MSW}$
Within groups	SSW	$(n-K)$	MSW	
Total	SST	$(n-1)$		

Example:

A company buys thousands of light bulbs every year. The company is considering three brands of light bulbs to choose from. Before the company decides which light bulbs to buy, it wants to investigate if the mean life of the three types of light bulbs is the same. The research department selects randomly a few bulbs of each type and tested them. Table lists number of hours (in thousands) that each of the bulbs in each brand survived before being burned out.

Brand I	Brand II	Brand III
22	18	27
23	23	24
26	22	20
27	21	21
22		23

At the 5% significance level, test the null hypothesis that the mean life of bulbs for each of these three brands is the same.

Solution:

There are 3 groups: Brand I, Brand II, and Brand III:

$$n_1 = 5;\ n_2 = 4;\ n_3 = 5;\ n = 14$$

1) Let us calculate mean of each group

$$\bar{x}_1 = \frac{22+23+26+27+22}{5} = \frac{120}{5} = 24$$

$$\bar{x}_2 = \frac{18+23+22+21}{4} = \frac{84}{4} = 21$$

$$\bar{x}_3 = \frac{27+24+20+21+23}{5} = \frac{115}{5} = 23$$

2) Overall mean is

$$\bar{x} = \frac{\sum_{i=1}^{K} n_i \bar{x}_i}{n} = \frac{5 \cdot 24 + 4 \cdot 21 + 5 \cdot 23}{14} = \frac{319}{14} = 22.79$$

3) In the first group, sum of squared deviations is

$$SS_1 = \sum_{j=1}^{5} (x_{1j} - \bar{x}_1)^2 = (22-24)^2 + (23-24)^2 + (26-24)^2 +$$

$$+ (27-24)^2 + (22-24)^2 = 4+1+4+9+4 = 22$$

Similarly,

$$SS_2 = \sum_{j=1}^{4} (x_{2j} - \bar{x}_2)^2 = (18-21)^2 + (23-21)^2 + (22-21)^2 +$$

$$+ (21-21)^2 = 9+4+1+0 = 14$$

and

$$SS_3 = \sum_{j=1}^{5} (x_{3j} - \bar{x}_3)^2 = (27-23)^2 + (24-23)^2 + (20-23)^2 +$$

$$+ (21-23)^2 + (23-23)^2 = 16+1+9+4+0 = 30$$

4) $SSW = SS_1 + SS_2 + \ldots\ldots + SS_K = 22+14+30 = 66$

5) Now, let us calculate between group variability

$$SSG = \sum_{i=1}^{K} n_i \cdot (\bar{x}_i - \bar{x})^2 = 5 \cdot (24 - 22.79)^2 +$$

$$+ 4 \cdot (21 - 22.79)^2 + 5 \cdot (23 - 22.79)^2 = 7.32 + 12.82 + 0.22 = 20.36 .$$

6) $SST = SSW + SSG = 66 + 20.36 = 86.36$

Within-groups mean square is obtained as

$$MSW = \frac{SSW}{n - K} = \frac{66}{14 - 3} = 6$$

Between-groups mean square is obtained as

$$MSG = \frac{SSG}{K - 1} = \frac{20.36}{3 - 1} = 10.18$$

The null hypothesis is

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

The decision rule is reject H_0 if

$$T.S. = F = \frac{MSG}{MSW} > F_{K-1,n-K,\alpha}$$

The value of test statistic is $F = \frac{MSG}{MSW} = \frac{10.18}{6} = 1.70$

$$F_{K-1,n-K,\alpha} = F_{2,11,0.05} = 3.98 .$$

Since 1.70 is not greater than 3.98 we fail to reject H_0. We accept hypothesis that mean of all three populations are equal, in other words, there is no difference for company which brand to choose.

In the end, substituting the values of various quantities in Table 5.1, we write an ANOVA table for our example as

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
Between groups	20.36	2	10.18	$\frac{10.18}{6} = 1.70$
Within groups	66	11	6	
Total	86.36	13		

Remark1: An alternative formula for *SSB* and *SSW* are

$$SSB = \left(\frac{T_1^{2\cdot}}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + ... \right) - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}$$

$$SSW = \sum_{i=1}^{n} x_i^2 - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + ... \right)$$

where

T_i – the sum of the values in sample i

$\sum_{i=1}^{n} x_i$ – the sum of the values in all samples $= T_1 + T_2 + T_3 + ...$

$\sum_{i=1}^{n} x_i^2$ – the sum of the squares of the values in all samples.

Example:

Consider the following data obtained for two samples selected from two populations

Sample I	Sample II
9	4
3	1
7	1
8	6
	8

Set out the analysis of variance table for these data.

Solution:

$T_1 = 9+3+7+8 = 27$

$T_2 = 4+1+1+6+8 = 20$

$\sum_{i=1}^{9} x_i = T_1 + T_2 = 27 + 20 = 47$

$n_1 = 4;$ $n_2 = 5;$ $n = n_1 + n_2 = 9$

$$\sum_{i=1}^{n} x_i^2 = 9^2 + 3^2 + 7^2 + 8^2 + 4^2 + 1^2 + 1^2 + 6^2 + 8^2 = 321$$

Substituting all the values in the formula for *SSG* and *SSW*, we obtain

$$SSG = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + ... \right) - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n} =$$

$$= \left(\frac{27^2}{4} + \frac{20^2}{5} \right) - \frac{47^2}{9} = 16.81$$

$$SSW = \sum_{i=1}^{n} x_i^2 - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + ... \right) =$$

$$= 321 - \left(\frac{27^2}{4} + \frac{20^2}{5} \right) = 58.75$$

Hence, the variance between samples *MSG* and the variance within samples *MSW* are

$$MSG = \frac{SSG}{K-1} = \frac{16.81}{2-1} = 16.81$$

$$MSW = \frac{SSW}{n-K} = \frac{58.75}{9-2} = 8.39$$

We write an ANOVA table for our example as

Source of variation	Sum of squares	Degrees of freedom	Mean squares	*F* ratio
Between groups	16.81	1	16.81	2.00
Within groups	58.75	7	8.39	
Total	75.56	8		

Remark2:

To use MINITAB menu follow the following instructions:

1. Select Stat>ANOVA>One-way (Unstacked)
2. Select data columns
3. Click OK

Exercises

1. The following ANOVA table, based on information obtained for four samples selected from four independent populations that are normally distributed with equal variances, has a few missing values

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
Between groups				$F = \frac{?}{?} = 5.67$
Within groups		16	8.245	
Total		19		

a) Complete the analysis of variance table.

b) Using $\alpha = 0.05$, test the null hypothesis that the means of the four populations are equal against the alternative hypothesis that the means of the four populations are not equal.

2. Respond to each of the following questions using this partially completed one-way ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
Between groups		3		$F = ?$
Within groups	405			
Total	888	31		

a) How many different populations are being considered in this analysis?

b) Fill in the ANOVA table with missing values.

c) State the appropriate null and alternative hypothesis.

d) What conclusions should be reached regarding the null hypothesis? Test using an $\alpha = 0.05$.

3. Three samples randomly selected from three independent populations that are normally distributed with equal variances produced the following data.

Sample I	Sample II	Sample III
32	45	47
28	43	32
40	38	43
36	44	37
39	33	41

a) Set out the analysis of variance table for these data.
b) Test at a 1% significance level, the null hypothesis that the means of these three populations are equal.

4. Consider the following data obtained for two samples selected at random from two populations that are independent and normally distributed with equal variances

Sample I	Sample II
29	37
31	27
27	36
28	20
25	

a) Set out the analysis of variance table for these data.
b) Test at a 5% significance level, the null hypothesis that the means of these three populations are equal.

5. A company hired three new salespersons with degrees in mathematics, economics and marketing. The company wants to check if the fields of study have any effect on the mean number of sales made by these salespersons. The following table lists the number of items sold by these three salespersons during certain randomly selected days.

Person with Math. degree	Person with economics degree	Person with marketing degree
7	2	3
8	5	1
6	3	1
11	1	2
9	2	6
13	5	5

a) Set out the analysis of variance table for these data.
b) Using the 5% significance level, can you reject the null hypothesis that the mean number of items sold per day by all salespersons with degrees in each of these three areas is the same?

6. A consumer agency that wanted to compare drying times for paints made by three companies tested a few samples of paints from each of these companies. The following table lists the drying times (in minutes) for these samples of paints

Company A	Company B	Company C
43	58	44
52	64	49
42	62	50
46	54	57
41	52	43
50	62	40
55		46

a) Set out the analysis of variance table for these data.
b) Using the 1% significance level, test the null hypothesis that the mean drying times for paints of these companies are equal.

Answers

1. b) $T.S. = F = 5.67$; reject H_0; **2.** a) 4; d) $T.S. = F = 11.13$; reject H_0;

3. a)

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
Between groups	94.5	2	47.3	
Within groups	333.2	12	27.8	$F = 1.70$
Total	427.7	14		

b) accept H_0;

4. a)

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
Between groups	8.9	1	8.9	
Within groups	214.0	7	30.6	$F = 0.29$
Total	222.9	8		

b) accept H_0;

5. a)

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
Between groups	144.00	2	72.00	
Within groups	70.00	15	4.67	$F = 15.43$
Total	214.00	17		

b) reject H_0;

6. a)

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
Between groups	571.7	2	285.8	
Within groups	481.3	17	28.3	$F = 10.10$
Total	1053.0	19		

b) reject H_0.

5.3. The Kruskal-Wallis test

One-way analysis of variance test is based on an assumption that the underlying population has a normal distribution. If the population distribution is not normal, it is possible to develop a nonparametric alternative to the one-way analysis of variance. This nonparametric test is known as the **Kruskal-Wallis test.** Like the most of nonparametric the Kruskal-Wallis test is based on the ranks of the sample observations.

Suppose that we have independent random samples of sizes $n_1, n_2, n_3, \ldots\ldots, n_K$ observations, selected from K populations. Let

$$n = n_1 + n_2 + n_3 + \ldots\ldots + n_K$$

Denote the total number of observations. The null hypothesis is

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

To apply Kruskal-Wallis test it is necessary to take following steps:

1. Pool together all sample observations.
2. Rank all of pooled sample observations in ascending order.
3. Denote by $R_1, R_2, \ldots\ldots R_K$ the sums of ranks for the K samples
4. Calculate the value of the test statistic

$$W = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1)$$

A test of significance level α is given by the decision rule

Reject H_0 if $W > \chi^2_{K-1,\alpha}$

where $\chi^2_{K-1,\alpha}$ is the number that is exceed with probability α by a χ^2 random variable with $(K-1)$ degrees of freedom.

Example:

The following table gives the response time (in minutes) of three fire companies in a city for certain randomly selected incidents after a fire was reported.

Company A	Company B	Company C
1.6	1.4	0.8
0.8	2.6	1.3
2.7	0.9	1.7
1.2	3.5	0.9
3.4	1.2	1.1
1.9	1.5	0.7
4.3		2.1

Perform at the 5% significance level the Kruskal-Wallis test to test the null hypothesis that the mean response time for each of these fire companies for all fire incidents are the same.

Solution:

First of all we pool all sample observation together and rank them in ascending order. The following table illustrates this procedure

Company A	Rank	Company B	Rank	Company C	Rank
1.5	11	1.4	10	0.8	2.5
0.8	2.5	2.6	16	1.3	9
2.7	17	0.9	4.5	1.7	13
1.2	7.5	3.5	19	0.9	4.5
3.4	18	1.2	7.5	1.1	6
1.9	14	1.6	12	0.7	1
4.3	20			2.1	15
Rank sums	90		69		51

The null hypothesis is

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

The decision rule is

Reject H_0 if $W > \chi^2_{K-1,\alpha}$

The value of the test statistic is

$$W = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1) =$$

$$= \frac{12}{20 \cdot (20+1)} \left(\frac{90^2}{7} + \frac{69^2}{6} + \frac{51^2}{7} \right) - 3 \cdot (20+1) = 66.35 - 63 = 3.35$$

$$\chi^2_{K-1,\alpha} = \chi^2_{2,0.05} = 5.99$$

Since, 3.35 is not greater than 5.99, we fail to reject the null hypothesis. And we accept that the mean response time for each of these fire companies for all fire incidents is the same.

Exercises

1. A manufacturer of wall papers is considering three alternative colors: red, green and blue. To check whether such consideration has an effect on sales, 21 stores are chosen. Red color papers are sent 7 of them, green color papers are sent 9 of them, and blue color papers are sent to the remaining 5. After a few days, a check is made on the number of sales in each store. The results are shown below

Red	Green	Blue
44	53	84
38	77	29
58	29	46
77	65	85
60	89	35
29	34	
58	43	
	62	
	35	

At the 5% significance level perform a Kruskal-Wallis test of the null hypothesis that the population mean sales levels are identical for three wall paper colors. Also find the *p*-value.

2. A study was conducted in which samples were selected independently from four populations. The sample size from each population was 21. The data were converted to ranks. The sum of the ranks for the data from each sample is

	Sample A	Sample B	Sample C	Sample D
Sum of ranks	642	784	458	1361

a) State the appropriate null and alternative hypothesis if you wish to determine whether the populations have equal means.
b) At the 2.5% significance level perform a Kruskal-Wallis test.

3. The following summary data have been collected from three samples selected from three populations

$n_1 = 20$	$n_2 = 25$	$n_3 = 35$
$\sum R_1 = 1660$	$\sum R_2 = 1150$	$\sum R_3 = 1350$

Based on these data, what can be concluded about the means for three populations? Apply Kruskal-Wallis test at an $\alpha = 0.01$.

4. Given the following data:

Group 1	Group 2	Group 3	Group 4
21	28	16	22
27	26	15	23
26	21	18	19
22	29	20	17
25	30	15	20
30	25		
23			

Use the Kruskal-Wallis procedure to test the null hypothesis that the mean values for all four populations are the same. What conclusion should be reached using a significance level of 0.10? Also find the p-value.

5. Suppose as a part of your job you are responsible for installing emergency lighting in a series of buildings. Bids have been received from four manufacturers of battery-operated emergency lights. The costs are about equal, so the decision will be based on the length of time the lights last before failing. A sample of five lights from each manufacturer has been tested, and values (time in hours) recorded for each manufacturer

Type I	Type 2	Type 3	Type 4
1025	1222	1121	989
1122	1250	1201	987
1250	1390	1190	1087
1023	1426	1122	1121
1130	1322	1390	1200

Using $\alpha = 0.01$, what conclusion should you reach about the mean length of time the lights last before failing for the four manufacturers? Explain.

Answers

1. $T.S. = 0.0376$; accept H_0; **2.** $T.S. = -7.74$; accept H_0; **3.** $T.S. = 206.58$; reject H_0; **4.** b) $T.S. = 14.95$; reject H_0; **5.** $T.S. = 2.22$; fail to reject H_0.

5.4. Two-way analysis of variance

In section 5.2 we introduced one-way ANOVA for testing hypothesis involving three or more population means. This ANOVA method is appropriate as long as we are interested in analyzing one factor at a time and we select independent random samples from the populations. There are situations in which another factor affects the observed response in a one-way design.

Suppose that we have K groups and H blocks. We will use x_{ij} to denote the sample observations corresponding to the i^{th} group and the j^{th} block, as it shown in Figure 5.2. There are total

$$n = K \cdot H$$

number of observations.

Our aim is to test the null hypothesis that all group means are equal, and the null hypothesis that all block means are equal.
To develop these tests we need to set two-way ANOVA table.

BLOCK	GROUP			
	1	2	.....	K
1	x_{11}	x_{21}	.....	x_{K1}
2	x_{12}	x_{22}	.....	x_{K2}
.	...	...	.....	.....
.	....	....	.....	...
H	x_{1H}	x_{2H}	.....	x_{KH}

Figure 5.2

To set this table we will use following steps:
1) Find sample mean for each group. For the mean of the i^{th} group we use notation $\bar{x}_{i\bullet}$, defined as

$$\bar{x}_{i\bullet} = \frac{\sum_{j=1}^{H} x_{ij}}{H}; \qquad (i = 1,2,3.....,K)$$

2) Find sample mean for each block. The mean of the j^{th} block we use notation $\bar{x}_{\bullet j}$, defined as

$$\bar{x}_{\bullet j} = \frac{\sum_{i=1}^{K} x_{ij}}{K}; \qquad (j = 1,2,3.....,H)$$

3) Find the overall mean of the sample observations. The overall mean denoted $\bar{x}$, defined as

$$\bar{x} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{H} x_{ij}}{n} = \frac{\sum_{i=1}^{K} \bar{x}_{i\bullet}}{K} = \frac{\sum_{j=1}^{H} \bar{x}_{\bullet j}}{H}$$

4) Find between groups sum of squares, denoted SSG, defined as

$$SSG = H \cdot \sum_{i=1}^{K} (\bar{x}_{i\bullet} - \bar{x})^2$$

5) Find between blocks sum of squares, denoted SSB, defined as

$$SSB = K \cdot \sum_{j=1}^{H} (\bar{x}_{\bullet j} - \bar{x})^2$$

6) Find the error sum of squares, denoted SSE, defined as

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{j\bullet} + \bar{x})^2$$

7) Find the total sum of squares, denoted SST, defined as

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \bar{x})^2$$

It can be shown that

$$SST = SSG + SSB + SSE$$

8) We define the following mean squares

Between-groups: $MSG = \dfrac{SSG}{K-1}$

Between-blocks: $MSB = \dfrac{SSB}{H-1}$

Error: $MSE = \dfrac{SSE}{(K-1)(H-1)}$

9) We define two F ratios

$$\frac{MSG}{MSE} \quad \text{and} \quad \frac{MSB}{MSE}$$

We will use ratios above to test the null hypothesis about equality of population blocks and population groups.

10) The null hypothesis H_0 that the K population group means are the same is provided by the decision rule

Reject H_0 if $\dfrac{MSG}{MSE} > F_{K-1,(K-1)(H-1),\alpha}$

2) The null hypothesis H_0 that the H population block means are the same is provided by the decision rule

$$\text{Reject } H_0 \text{ if } \frac{MSB}{MSE} > F_{H-1,(K-1)(H-1),\alpha}$$

where, $F_{\nu_1,\nu_2,\alpha}$ is the number exceeded with probability α by a random variable following an F distribution with numerator degrees of freedom ν_1 and denominator degrees of freedom ν_2.

It is very convenient to summarize the calculations in tabular form, called a **two-way analysis of variance table** or ANOVA table, shown below (Table5.2):

Table 5.2

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratios
Between groups	SSG	$(K-1)$	$MSG=\frac{SSG}{K-1}$	$\frac{MSG}{MSE}$
Within blocks	SSB	$(H-1)$	$MSB=\frac{SSB}{H-1}$	$\frac{MSB}{MSE}$
Error	SSE	$(K-1)(H-1)$	$MSE=\frac{SSB}{(K-1)(H-1)}$	
Total	SST	$(n-1)$		

Exercise:

Four drivers tested three types of cars for fuel consumptions. The accompanying table shows fuel consumptions of cars

Block (Drivers)	Group (Cars)		
	A	B	C
1	22	24	26
2	21	25	22
3	19	20	23
4	18	19	21

a) Set out the two-way analysis of variance table.

b) Test the null hypothesis that the population mean fuel consumption is the same for all three types of cars. Take $\alpha = 0.05$.

c) Test the null hypothesis that population values of mean fuel consumption are the same for each driver. Take $\alpha = 0.05$.

Solution:

a)

1) Let us find sample mean for each group

$$\bar{x}_{i\bullet} = \frac{\sum_{j=1}^{H} x_{ij}}{H} \qquad (i=1,2,3)$$

$$\bar{x}_{1\bullet} = \frac{22+21+19+18}{4} = \frac{80}{4} = 20$$

$$\bar{x}_{2\bullet} = \frac{24+25+20+19}{4} = \frac{88}{4} = 22$$

$$\bar{x}_{3\bullet} = \frac{26+22+23+21}{4} = \frac{92}{4} = 23$$

2) Find sample mean for each block.

$$\bar{x}_{\bullet j} = \frac{\sum_{i=1}^{K} x_{ij}}{K}; \qquad (j=1,2,3,4)$$

$$\bar{x}_{\bullet 1} = \frac{22+24+26}{3} = \frac{72}{3} = 24; \qquad \bar{x}_{\bullet 2} = \frac{21+25+22}{3} = \frac{68}{3} = 22.67$$

$$\bar{x}_{\bullet 3} = \frac{19+20+23}{3} = \frac{62}{3} = 20.67; \qquad \bar{x}_{\bullet 4} = \frac{18+19+21}{3} = \frac{58}{3} = 19.33$$

3) Find the overall mean of the sample observations.

$$\bar{x} = \frac{\sum_{j=1}^{K}\sum_{i=1}^{H} x_{ij}}{n} = \frac{\sum_{i=1}^{K} \bar{x}_{i\bullet}}{K} = \frac{\sum_{j=1}^{H} \bar{x}_{\bullet j}}{H}$$

$$\bar{x} = \frac{\sum_{i=1}^{K} \bar{x}_{i\bullet}}{K} = \frac{20+22+23}{3} = 21.67$$

4) Find between groups sum of squares

$$SSG = H \cdot \sum_{i=1}^{K} (\bar{x}_{i\bullet} - \bar{x})^2 =$$

$$= 4 \cdot \left((20-21.67)^2 + (22-21.67)^2 + (23-21.67)^2\right) = 18.67$$

5) Find between blocks sum of squares

$$SSB = K \cdot \sum_{j=1}^{H} (\bar{x}_{\bullet j} - \bar{x})^2 = 3 \cdot ((24 - 21.67)^2 + (22.67 - 21.67)^2 +$$

$$+ (20.67 - 21.67)^2 + (19.33 - 21.67)^2) = 38.71$$

6) Let us find total sum of squares

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \bar{x})^2 = (22 - 21.67)^2 + (24 - 21.67)^2 +$$

$$+ (26 - 21.67)^2 + (21 - 21.67)^2 + (25 - 21.67)^2 + (22 - 21.67)^2 +$$

$$+ (19 - 21.67)^2 + (20 - 21.67)^2 + (23 - 21.67)^2 + (18 - 21.67)^2 +$$

$$+ (19 - 21.67)^2 + (21 - 21.67)^2 = 68.67$$

7) By subtraction we obtain error sum of squares

$$SSE = SST - SSG - SSB = 68.67 - 18.67 - 38.71 = 11.29$$

8) For the fuel consumption data, the mean squares are

$$MSG = \frac{SSG}{K-1} = \frac{18.67}{3-1} = 9.34$$

$$MSB = \frac{SSB}{H-1} = \frac{38.71}{4-1} = 12.90$$

$$MSE = \frac{SSE}{(K-1)(H-1)} = \frac{11.29}{(3-1)(4-1)} = 1.88$$

9) We define two F ratios

$$\frac{MSG}{MSE} = \frac{9.34}{1.88} = 4.97$$

and

$$\frac{MSB}{MSE} = \frac{12.90}{1.88} = 6.86$$

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratios
Between groups	18.67	2	9.34	4.97
Within blocks	38.71	3	12.90	6.86
Error	11.29	6	1.88	
Total	68.67	11		

b) We can write the null hypothesis that the population means fuel consumption is the same for all three types of cars as

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

The decision rule is

$$\text{Reject } H_0 \text{ if } \frac{MSG}{MSE} > F_{K-1,(K-1)(H-1),\alpha}$$

$$F_{K-1,(K-1)(H-1),\alpha} = F_{2,6,0.05} = 5.14$$

Since the value of test statistic 4.97 is not greater than 5.14, we fail to reject the null hypothesis. Therefore, we accept the hypothesis that the fuel consumptions are the same for all types of cars.

c) We write the null hypothesis of equality of the population values of mean fuel consumption for all four drivers as

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

The decision rule is

$$\text{Reject } H_0 \text{ if } \frac{MSB}{MSE} > F_{H-1,(K-1)(H-1),\alpha}$$

$$F_{H-1,(K-1)(H-1),\alpha} = F_{3,6,0.05} = 4.76$$

Since the value of test statistic 6.86 is greater than 4.76, we reject the null hypothesis. Therefore, we accept the hypothesis that the fuel consumptions are not the same for each driver age class. In other words, fuel consumption of car depends on driver's habit.

Remark1:

To use MINITAB menu follow the following instructions:

1. Select Stat>ANOVA>Two-way
2. Enter Response variable
3. Enter row factor
4. Enter column factor
5. Click OK.

Remark2:
For example above the MINITAB instruction is shown below

C1	C2	C3
Driver	Car	Fuel consumption
1	1	22
1	2	24
1	3	26
2	1	21
2	2	25
2	3	22
3	1	19
3	2	20
3	3	23
4	1	18
4	2	19
4	3	21

In this case the row factor is "Car" and the column factor is "driver"

Exercises

1. An ANOVA was performed and the following partially completed ANOVA table is available:

Source of variation	Sum of squares	Degrees of freedom	Mean squares	*F* ratios
Between groups		7	26	
Within blocks		14		
Error	57.900			
Total	402.100			

a) Complete the analysis of variance table.
b) How many populations are being tested?
c) How many blocks were used in this analysis of variance?
d) Using an alpha equal to 0.05, test the null hypothesis that the population means are equal for all groups.
e) Using an alpha equal to 0.05, test the null hypothesis that the population means are equal for all blocks.

2. Three analysts were asked to predict earnings growth over the coming year for four companies producing cars. Their forecasts (in percentage increase in earnings) are given below

Analysts	Cars companies			
	A	B	C	D
1	7	11	10	13
2	8	9	13	12
3	9	13	16	15

a) Set out analysis of variance table
b) For $\alpha = 0.01$ test the null hypothesis that the population mean growth forecasts are the same for all car companies.
c) For $\alpha = 0.05$ test the null hypothesis that the population mean growth forecasts are the same for all analysts.

3. A soft drink producer wants to compare the effects on sales of can colors: red, yellow, and blue. He selects four regions and in three stores in each region, each to sell one color cans. The following table shows number of sales (in hundreds of cans) at the end of the experiment.

Regions	Can colors		
	Red	Yellow	Blue
Region I	37	28	25
Region II	29	33	22
Region III	31	29	27
Region IV	23	26	26

a) Set out analysis of variance table
b) Test the null hypothesis that the population mean sales are the same for each can color. Use p-value approach.
c) Test the null hypothesis that the population mean sales are the same for all four regions. Use p-value approach.

4. Three real estate agents were each asked to assess the values of five houses. The results, in thousands of dollars, are shown below.

House	Agents		
	A	B	C
I	200	210	220
II	190	192	196
III	180	195	205
IV	160	182	194
V	170	171	185

a) Set out analysis of variance table.

b) Test at 5% significance level the null hypothesis that population mean valuations are the same for the three real estate agents.

Answers

1. b) 8; c) 15; d) $\frac{MSG}{MSE} = 44.06$; reject H_0; e) $\frac{MSB}{MSE} = 19.63$; reject H_0;

2. a)

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratios
Between car comp.	54.00	3	18.00	10.29
Within analysts	22.17	2	11.08	6.33
Error	10.50	6	1.75	
Total	86.67	11		

b) $T.S. = 10.29$, reject H_0; c) accept H_0;

3. a)

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratios
Between colors	56.0	2	28.0	1.71
Within regions	42.0	3	14.0	0.86
Error	98.0	6	16.3	
Total	196.0	11		

4. a)

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratios
Between agents	1000.0	2	500.0	14.79
Within houses	2285.3	4	571.3	16.90
Error	270.7	8	33.8	
Total	3556.0	14		

Chapter 6

Statistical quality control

6.1. Introduction

In manufacturing process, there is always some variation in the items manufactured. For example, even though a machine is designed to cut a piece of pipe a certain length, the lengths of the pipes cut will not all be exactly equal. There are two causes of variation: normal or chance of variation; and variation that is due to human or mechanical reasons in the manufacturing process. If the variation is due to human or mechanical reasons in the process, this causes defective parts, it must be detected at an early stage and corrected.

A real life example concerns the weights of canned food. In this case, a food producer wishes to provide assurance that the minimum weight of its canned product is being met. Food producers have a tendency to fill cans with more food than the can's label indicates, out of fear of being caught with under weight cans by the state inspectors.

The question is, "how does one determine (with the variability in the filling process) with assurance that the minimum weight of all cans is being achieved and at the same time, not lose money by overfilling each can? This chapter will explain a technique called *statistical quality control*, which can be used to answer this question.

6.2. Variation

One of the fundamental principles of the statistical thinking is that variation exists in all process. It is important to understand variation in order to predict the future performance of the process. There are two causes of variation:

1) common causes
2) assignable causes

Common causes of variation (also called uncontrollable causes) are those causes that are random in occurrence and are happens during all process. Management, (not workers), are responsible for these causes.

Assignable causes of variation (also called special cases) are the results of external sources, that is, sources that are outside of the system. These causes

can and must be detected, and corrective actions must be taken to remove them from the process. Not taking actions will increase variation and lower the quality.

Definition:

A production process is called **stable** (in-control) if all assignable causes are removed; thus, variation results only from common causes.

6.3. Control charts

Figure 6.1 illustrates the general format of a process control chart. The upper and lower control limits define the normal operating region for the process.

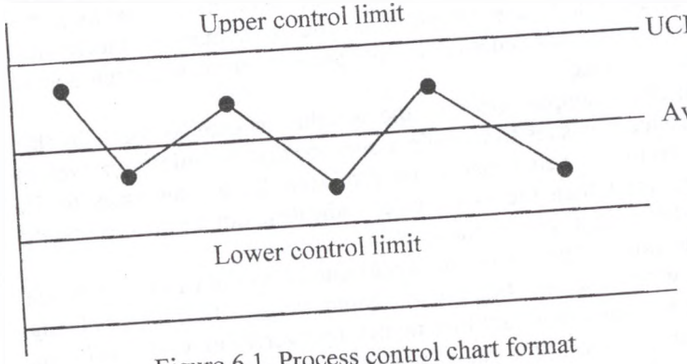

Figure 6.1. Process control chart format

The horizontal axis reflects the passage of time, or order of production. The vertical axis corresponds to the variable of interest. There are number of different types of process control charts. We will consider the most commonly used process control charts

$\bar{x}$– chart

$\bar{s}$– chart

p – chart

c – chart

6.3.1 Control charts for means and standard deviations

Let us consider a production process that yields an output whose characteristic of interest can be measured on a continuum. It is necessary to set up a quality control scheme for that process. It can be done by taking, over time, a sequence of small samples of output. Often samples of four or five are taken. The frequency of sample observations depends on the characteristics of production process. Management is often interested in both average performance of production process and the variability in process. Sample means and standard deviations are used to track process performance. Three measures used in the development of control charts for means and standard deviations. They are: overall mean, the average sample standard deviation, and the process standard deviation.

Assume that a sequence of K samples, each of n observations, is selected over time from the production process. The sample means denoted $\bar{x}_i$ for $i = 1,2,3,.....K$ can be graphed on an $\bar{x}$– chart. The **average of these sample means** is called the **overall mean** of all the sample observations

$$\bar{\bar{x}} = \sum_{i=1}^{K} \bar{x}_i / K$$

The sample standard deviations denoted s_i for $i = 1,2,3,.....K$ can be graphed on an s – chart. The **average sample standard deviation** is

$$\bar{s} = \sum_{i=1}^{K} \bar{s}_i / K$$

The **process standard deviation,** σ is the standard deviation of the population from which the samples were drawn, and it must be estimated from the sample data.

Since the sample standard deviation s_i is based on n observations, it can be shown that

$$E(s_i) = c_4 \cdot \sigma$$

where c_4 is the number that can be computed as a function of the sample size it follows that

$$E(\bar{s}) = c_4 \cdot \sigma$$

and hence that an **unbiased estimate of the process standard deviation** is given by

$$\hat{\sigma} = \bar{s}/c_4$$

The value of c_4-control chart factor, can be found in Table 6.1. Table 6.1 lists values of c_4, corresponding sample sizes from two to ten. It also contains factors for other control charts that will be discussed throughout this chapter.

Table6.1 Factors for control charts

N	C_4	A_3	B_3	B_4
2	0.7879	2.659	0	3.267
3	0.8862	1.954	0	2.568
4	0.9213	1.628	0	2.266
5	0.9400	1.427	0	2.089
6	0.9515	1.287	0.030	1.970
7	0.9594	1.182	0.118	1.882
8	0.9650	1.099	0.185	1.815
9	0.9690	1.032	0.239	1.761
10	0.9727	0.975	0.284	1.716

To determine control limits for $\bar{x}$ -charts, we assume that the process has been operating at a constant level of performance over the whole observation period and, assume that all sample observations have been drawn from the same normal distribution.

The sampling distribution is centered on the overall mean, and the value of the overall mean determines the central line, called **center line.** Then, if three-standard error limits are to be used, the control limits are

$$\bar{\bar{x}} \pm 3 \cdot \hat{\sigma}/\sqrt{n} = \bar{\bar{x}} \pm 3 \cdot \bar{s}/(c_4 \cdot \sqrt{n}) = \bar{\bar{x}} \pm A_3 \cdot \bar{s}$$

where

$$A_3 = 3/(c_4 \cdot \sqrt{n})$$

Control chart for $\bar{x}$– means

The $\bar{x}$– **chart** is a time plot of the sequence of sample means.
The **center line** is

$$CL_{\bar{x}} = \bar{\bar{x}}$$

In addition, there are three-standard error control limits.
The **lower control limit** is

$$LCL_{\bar{x}} = \bar{\bar{x}} - A_3 \cdot \bar{s}$$

The **upper control limit** is

$$UCL_{\bar{x}} = \bar{\bar{x}} + A_3 \cdot \bar{s}$$

where the values of A_3 are given in Table 6.1.

Example:
The accompanying table shows sample means and sample standard deviations for a sequence of 10 samples of seven observations on a quality characteristic of a product

Sample	$\bar{x}$	s
1	145.2	2.3
2	139.2	3.1
3	146.3	2.1
4	138.2	1.9
5	141.2	2.4
6	144.3	2.2
7	140.1	3.1
8	139.9	2.3
9	145.5	2.7
10	143.3	2.8

a) Find the center line and lower and upper control limits for an $\bar{x}$– chart.

b) Draw the $\bar{x}$– chart.

Solution:

a) First of all, let us find overall mean and average of the sample standard deviations are

$$\bar{\bar{x}} = \frac{145.2 + 139.2 + ... + 143.3}{10} = 142.32$$

$$\bar{s} = \frac{2.3 + 3.1 + ... + 2.8}{10} = 2.49$$

The sample size is seven. So from table 6.1 we obtain $A_3 = 1.182$.

The central line is

$$CL_{\bar{x}} = \bar{\bar{x}} = 142.32$$

The lower control limit is

$$LCL_{\bar{x}} = \bar{\bar{x}} - A_3 \cdot \bar{s} = 142.32 - 1.182 \cdot 2.49 = 139.38$$

The upper control limit is

$$UCL_{\bar{x}} = \bar{\bar{x}} + A_3 \cdot \bar{s} = 142.32 + 1.182 \cdot 2.49 = 145.26$$

b) Each of the individual sample means are plotted on $\bar{x}$– chart in Figure 6.1. Three of the values fall out side of the control limits and there seems a great cause for concern. It is necessary to take action to correct the production process.

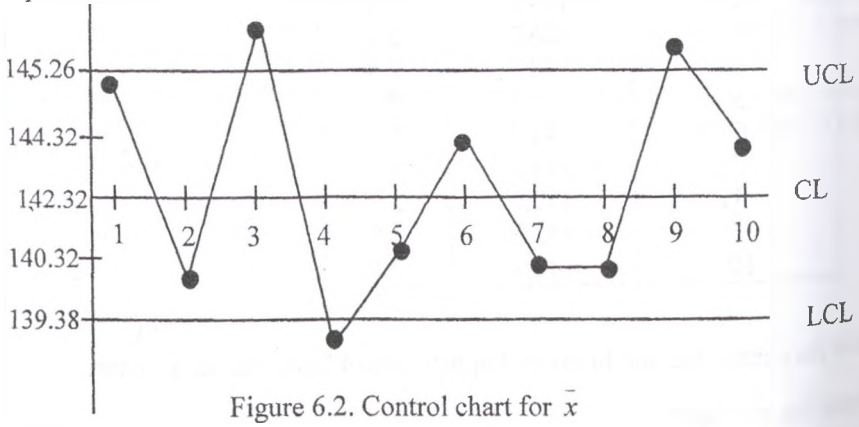

Figure 6.2. Control chart for $\bar{x}$

Control chart for s – standard deviations

The s – **chart** is a time plot of the sequence of sample standard deviations.
The **center line** is

$$CL_s = \bar{s}$$

In addition, there are three-standard error control limits.
The **lower control limit** is

$$LCL_s = B_3 \cdot \bar{s}$$

The **upper control limit** is

$$UCL_s = B_4 \cdot \bar{s}$$

where the values of B_3 and B_4 are given in Table 6.1.

Example: Refer to the data of example above.
a) Find the center line and lower and upper control limits for an s – chart.
b) Draw the s – chart and discuss its features
Solution:

a) $$\bar{s} = \frac{2.3 + 3.1 + ... + 2.8}{10} = 2.49$$

The sample size is seven. So from table 6 of the appendix we obtain $B_3 = 0.118$ and $B_4 = 1.882$

The center line is

$$CL_s = \bar{s} = 2.49$$

In addition, there are three-standard error control limits.
The lower control limit is

$$LCL_s = B_3 \cdot \bar{s} = 0.118 \cdot 2.49 = 0.29$$

The upper control limit is

$$UCL_s = B_4 \cdot \bar{s} = 1.882 \cdot 2.49 = 4.69$$

b) We plot each of the individual standard deviations on a control chart with center line $CL_s = 2.49$, lower control limit $LCL_s = 0.29$, and upper control limit, $UCL_s = 4.69$. The s-chart will look like Figure 6.3.

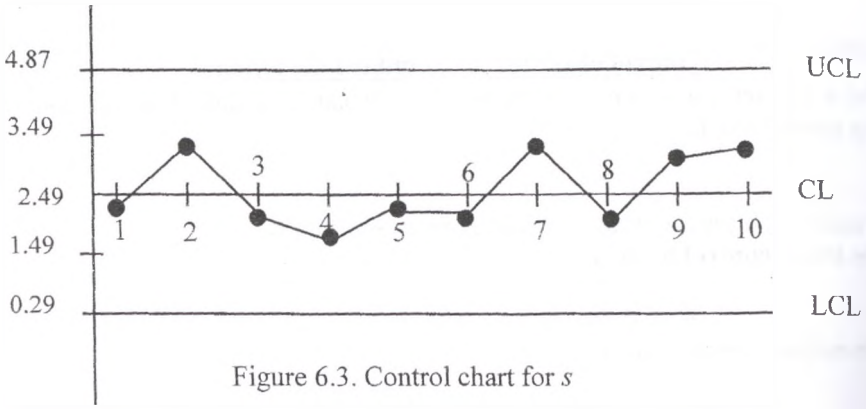

Figure 6.3. Control chart for *s*

The observed sample standard deviations are fall between lower and upper control limits. From this point of view, the process is under control. But since, the mean production process is out of control, the general production process must be corrected.

6.4. Interpretation of control charts

Once control charts have been developed, they can be used to determine whether the process is in control or out of control. The control charts are used to provide signals that something has changed. There are four primary signals indicating that a process might be out of control:

1. One or more points outside the upper or lower control limits

2. Nine or more points in a row above (or below) the center line

3. Six or more consecutive points moving in the same direction (increasing or decreasing).

These signals reduce to variation called **assignable-cause variation.** This variation is not random and is due to defects or problems in the manufacturing process, such as operators using the machine incorrectly, raw materials changing, etc. Assignable cause variations must be corrected in order to maintain quality in the production process.

Remark:

To use MINITAB menu follow the following instructions

1. Select Stat>Control charts>Select Xbar S
2. Enter variable location (for example, C1)
3. Enter subgroup size
4. Click OK.

Exercises

1. Data were collected on a quantitative measure with a sample of 6 observations for a sequence of thirty samples. 30 samples were collected, and following results were found

$$\bar{\bar{x}} = 42.3; \qquad \bar{s} = 4.2$$

a) Find the center line and lower and upper control limits for an $\bar{x}$– chart.
b) Find the center line and lower and upper control limits for an s – chart.

2. Weights of samples of canned fruit were measured. Results were available for a sequence of thirty samples, each of seven observations. The overall mean of the sample observations was 192.6 grams, and the average sample standard deviation was 5.42.

a) Use an unbiased estimator to find an estimate of the process standard deviation.

b) Find the center line and lower and upper control limits for an $\bar{x}$– chart.
c) Find the center line and lower and upper control limits for an s – chart.

3. The accompanying table shows sample means and sample standard deviations for a sequence of 14 samples of eight observations on a quality characteristic of a product

Sample	$\bar{x}$	s
1	146.4	4.37
2	152.8	6.79
3	150.6	3.17
4	149.2	4.71
5	150.6	4.98
6	150.4	6.28
7	151.1	6.20
8	152.9	6.97
9	147.2	4.28
10	154.3	7.29
11	151.8	3.1
12	149.9	5.31
13	146.7	4.73
14	152.1	6.12

a) Find the overall mean of the sample observations.
b) Find the average sample standard deviation.
c) Use an unbiased estimator to find an estimate of the process standard deviation.

d) Find the center line and lower and upper control limits for an $\bar{x}$– chart.

e) Draw the $\bar{x}$– chart and discuss its features.
f) Find the center line and lower and upper control limits for an s – chart.
g) Draw the s – chart and discuss its features.

4. Ten samples, each consisting of five automobile batteries, are tested for strength. The means and sample standard deviations are given here.

Sample	Mean	Standard deviation
1	12.2	2.1
2	12.5	1.3
3	12.3	1.5
4	11.9	2.1
5	11.8	1.2
6	11.2	1.1
7	12.1	1.3
8	12.0	1.3
9	12.2	1.8
10	11.8	2.0

a) Find the overall mean of the sample observations.
b) Find the average sample standard deviation.
c) Use an unbiased estimator to find an estimate of the process standard deviation.

d) Find the center line and lower and upper control limits for an $\bar{x}$– chart.

e) Draw the $\bar{x}$– chart and discuss its features.
f) Find the center line and lower and upper control limits for an s – chart.
g) Draw the s – chart and discuss its features.

Answers

1.a) $CL_{\bar{x}} = 42.3$; $LCL_{\bar{x}} = 36.89$; $UCL_{\bar{x}} = 47.71$; b) $CL_s = 4.2$; $LCL_s = 0.13$; $UCL_s = 8.27$; 2.a) $\hat{\sigma} = 5.65$; b) $CL_{\bar{x}} = 192.6$; $LCL_{\bar{x}} = 186.20$; $UCL_{\bar{x}} = 199.00$; c) $CL_s = 5.42$; $LCL_s = 0.64$; $UCL_s = 10.19$; 3. a) 150.43; b) 5.307; c) $\hat{\sigma} = 5.50$; d) $CL_{\bar{x}} = 150.43$; $LCL_{\bar{x}} = 144.60$; $UCL_{\bar{x}} = 156.26$; f) $CL_s = 5.307$; $LCL_s = 0.98$; $UCL_s = 9.63$; 4. a) 12.00; b) 1.57; c) 1.67; d) $CL_{\bar{x}} = 12.00$; $LCL_{\bar{x}} = 9.76$; $UCL_{\bar{x}} = 14.24$; f) $CL_s = 1.57$; $LCL_s = 0.00$; $UCL_s = 3.28$;

6.5. Control charts for proportions

Now let us consider situations, where individual product items will be checked to have conformed or not to have conformed to specifications. Again, a sequence of samples is taken over time to control product quality. Our interest is the proportion of **nonconforming**, or **defective,** items in each sample. It is clear that, it is desirable that this proportion be as small as possible, and any increasing trend over time should cause concern.

The $p-$ chart is used to monitor the proportion of defective items. One important difference between $p-$ chart and charts of previous section is that here much larger sizes are necessary. Any competently engineered production process is not going to generate a large proportion of nonconforming items. Therefore, to get a reasonable result, a relatively large sample size is necessary.

Suppose that a sequence of K samples, each of n observations, is taken from production process. The proportions of sample members **not conforming** to conforming can be determined as

$$\hat{p}_i = \frac{x}{n_i} \qquad i = 1,2,3,\ldots,K$$

where x – number of **not conforming** items;

n – sample size

If the samples are the same size, the **average of the sample proportions** is the **overall proportion of nonconforming** items. This is

$$\bar{p} = \sum_{i=1}^{n} \frac{\hat{p}_i}{K}$$

If sample sizes are not equal, then **overall proportion of nonconforming items** is defined as

$$\bar{p} = \frac{\sum_{i=1}^{n} n_i \cdot \hat{p}_i}{\sum_{i=1}^{n} n_i}$$

We know that individual sample proportions $\hat{p}_i$ have sampling distribution with mean estimated by $\bar{p}$ and standard deviation (standard error) given by

$$\hat{\sigma}_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Similar to the $\bar{x}-$chart and $s-$chart, three-standard error limits will be used for $p-$chart.

Control chart for $p-$proportions

The **$p-$chart** is a time plot of the sequence of sample proportions of nonconforming items.

The **central line** is

$$CL_p = \bar{p}$$

The **lower control limit** is

$$LCL_p = \bar{p} - 3 \cdot \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

The **upper control limit** is

$$UCL_p = \bar{p} + 3 \cdot \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Remark1:

If the lower control limit gets a negative value, which is impossible, we set the lower control limit at zero.

Remark2:

Interpretation of $p-$chart is similar to the interpretation of $\bar{x}-$chart and $s-$chart.

Remark3:

To use MINITAB menu follow the following instructions:

1. Select Stat>Control charts>Select P
2. Enter variable location (for example, C1)
3. Enter subgroup size (200, 300 etc.). If sample sizes vary, corresponding sample sizes should be in column (for example, C2)
4. Click OK.

Exercises

1. In the study of production process, 25 samples, each of 300 observations were taken. The average of the sample proportions of nonconforming items was 0.064. Find the center line and upper and lower control limits for p-chart.
2. Data were collected from a process in which the factor of interest was whether the finished item contained a particular attribute. A total of 30 samples were selected. The common sample size was 100 items. The total number of nonconforming items was 270. Based on these data, compute the central line, the upper and lower control limits for p-chart.
3. Samples of bowling balls are selected and screened to see whether there are any defects on their surfaces. The number of defective balls recorded for each sample is given here

Sample	Size	Number of defective balls
1	200	10
2	200	8
3	200	9
4	200	32
5	200	15
6	200	35
7	200	38
8	200	10

a) Find the center line and upper and lower control limits for p-chart.
b) Draw the p-chart and discuss its features.

4. Baseball caps are manufactured and checked to see whether the logos are properly printed on them. The number of defective logos per sample is shown below

Sample	Size	Number of defective balls
1	100	8
2	100	6
3	100	5
4	100	27
5	100	9
6	100	10

a) Find the center line and upper and lower control limits for p-chart.
b) Draw the p-chart and discuss its features.

5. Proportions of nonconforming items in a sequence of 12 samples, each 400 observations, are given below

Sample	$\hat{p}$	Sample	$\hat{p}$
1	0.061	7	0.068
2	0.060	8	0.036
3	0.043	9	0.064
4	0.051	10	0.056
5	0.037	11	0.046
6	0.042	12	0.039

a) Find the center line and upper and lower control limits for p-chart.
b) Draw the p-chart and discuss its features.

Answers

1. $CL_p = 0.064$; $LCL_p = 0.022$; $UCL_p = 0.106$; **2.** $CL_p = 0.090$;
$LCL_p = 0.004$; $UCL_p = 0.176$; **3.** a) $CL_p = 0.098$; $LCL_p = 0.035$;
$UCL_p = 0.161$; **4.** a) $CL_p = 0.108$; $LCL_p = 0.015$; $UCL_p = 0.202$;
5. a) $CL_p = 0.050$; $LCL_p = 0.017$; $UCL_p = 0.083$;

6.6. Control charts for number of occurrences: *c*-chart

The *p*-chart just discussed is used when you select sample of items and you determine the number of the sample items that possesses a specific attribute of interest. Each item either has or does not have that attribute. In practice often we meet the situations that involve attribute data but differ from the *p*-chart. Each sampling unit could have one or more of the attributes of interest. Number of attributes of interest, called the *number of occurrences* , counts number of imperfections per item over time. This is called a *c*-chart.

As with the other control charts studied in this chapter, some general notations used for control charts for number of occurrences are necessary.

Assume that, a sequence of K items is inspected over time, and for each item, the number of occurrences of some event, such as imperfections, is recorded. These numbers of occurrences denoted c_i for $i = 1,2,....K$.

The sample mean of occurrences is

$$\bar{c} = \sum_{i=1}^{K} \frac{c_i}{K}$$

A 3-sigma (3 standard deviation) control chart for the number of occurrences (*c*-chart) can be constructed in the usual way:

Control chart for *c*- number of occurrences

The c – **chart** is a time plot of the number of occurrences over the time.
The **central line** is

$$CL_c = \bar{c}$$

The **lower control limit** is

$$LCL_c = \bar{c} - 3 \cdot \sqrt{\bar{c}} \qquad \text{if } \bar{c} > 9$$

$$LCL_c = 0 \qquad \text{if } \bar{c} \le 9$$

The **upper control limit** is

$$UCL_c = \bar{c} + 3 \cdot \sqrt{\bar{c}}$$

Example:

Handheld calculators are manufactured and checked for defects. If a calculator is not defective, it is packaged and shipped to a retail store. Any defective calculators are repaired before they are shipped. Twelve of the defective calculators are checked for the number of defects per calculator. The numbers of defects per calculator are:

6, 3, 2, 5, 6, 7, 4, 3, 7, 8, 9 and 5

a) Find the central line and lower and upper limits for *c*-chart.

b) Draw the *c*-chart and discuss its features.

Solution:

a) First of all, let us find the mean number of defects per calculator, $\bar{c}$.

$$\bar{c} = \sum_{i=1}^{K} \frac{c_i}{K} = \frac{6+3+2+5+6+7+4+3+7+8+9+5}{12} = \frac{65}{12} = 5.42$$

$$CL_c = \bar{c} = 5.42$$

The lower and upper control limits are

$$LCL_c = 0 \qquad \text{since } \bar{c} \le 9$$

$$UCL_c = \bar{c} + 3 \cdot \sqrt{\bar{c}} = 5.42 + 3 \cdot \sqrt{5.42} = 12.40$$

b) Figure 6.4 represents *c*-chart

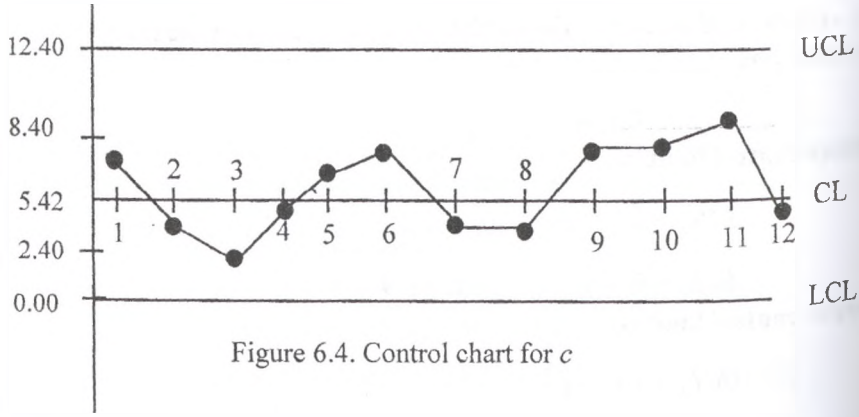

Figure 6.4. Control chart for *c*

Since, all points fall within the upper and lower control limits, the process is in control. That is, in the defective calculators, the number of defects per calculator is not excessive.

Remark:

To use MINITAB menu follow the following instructions:

1. Select Stat>Control charts>Select C
2. Enter variable location (C1; C2; etc.)
3. Enter subgroup size (200, 300 etc.)
4. Click OK.

Exercises

1. Sixty minute cassette tapes are checked for defects. The number of defects in each of 8 tapes is shown below

Tape	1	2	3	4	5	6	7	8
Number of defects	1	2	1	1	1	3	6	2

a) Find the central line and lower and upper limits for *c*-chart.
b) Draw the *c*-chart and discuss its features.

2. Workers are painting a large apartment building. An inspector checks several walls for paint defects. The number of defects per wall is shown below

Sample	1	2	3	4	5	6	7	8	9	10
Number of defects	6	12	9	8	8	6	12	10	11	8

a) Find the central line and lower and upper limits for *c*-chart.
b) Draw the *c*-chart and discuss its features.

3. A reader has very carefully read local paper for 15 weeks. For each Sunday's edition he has counted the number of typographical and spelling errors. The results are shown below

Week	Errors	Week	Errors
1	13	9	12
2	15	10	13
3	14	11	21
4	17	12	9
5	12	13	17
6	20	14	19
7	7	15	22
8	14		

a) Find the sample mean number of errors for these 15 weeks.
b) Find the central line and lower and upper limits for *c*-chart.
c) Draw the *c*-chart and discuss its features.

4. In a large market, the number of complaints in a week was recorded over 16 weeks. The results are shown in accompanying table

Week	Errors	Week	Errors
1	17	9	23
2	11	10	15
3	15	11	10
4	22	12	16
5	21	13	17
6	29	14	22
7	8	15	23
8	18	16	20

a) Find the sample mean number of complaints per week.
b) Find the central line and lower and upper limits for *c*-chart.
c) Draw the *c*-chart and discuss its features.

Answers

1. a) $CL_c = 2.125; LCL_c = 0; UCL_c = 6.498$; **2.** a) $CL_c = 9.00;\ LCL_c = 0;$ $UCL_c = 18$; **3.** a)15; b) $CL_c = 15;\ LCL_c = 3.381; UCL_c = 26.62;$ **4.** a) 17.94; b) $CL_c = 17.94;\ LCL_c = 5.232;\ UCL_c = 30.64$.

Chapter 7

Time series analysis and forecasting

7.1. Introduction to index numbers

7.1.1 Price index for a single item (Simple index number)

While analyzing time series data, decision maker often compares one value measured at one point in time with other values measured at different points in time. For example, a student may wish to compare book price in 2005 with prices in previous years. A common procedure for making relative comparisons is to begin by determining a base period index to which all other data values can be compared. This kind of index is called **a price index for a single item.** To form a price index, one time period is chosen as a base and the price for all other periods are expressed as a percentage of the base period price. If we denote the price in the base period by p_0, and the price in another period by p_1, then, the price index for the another period is

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

For example, the price of the house in 2001 was \$73.000 and in 2004 the price was \$121.000. If we take 2001 as a base period, then $p_0 = \$73.000$ and $p_1 = \$121.000$. The price index is

$$100 \cdot \left(\frac{p_1}{p_0} \right) = 100 \cdot \frac{121.000}{73.000} = 165.75\%$$

It means that the price of house increased by 65.75% in 2004 compared with 2001.

7.1.2 Unweighted aggregate price index

While price index for a single item can be used to identify price changes of a single item, we often are interested in the general price change for a group of items taken as a whole. An unweighted aggregate price index can be developed by simply summing the unit

prices in the time of interest and dividing this sum by the sum of the unit prices in the base year. Suppose that we have K number of items. Let

p_{0i} – denote the price of the i^{th} item in base period

p_{1i} – denote the price of the i^{th} item in period t

The unweighted aggregate price index in period t, denoted I_t, is given by

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

Example:

The following table lists prices of three different types of cars in different years

Year	Car types		
	A	B	C
2000	15.000	32.000	25.000
2001	16.000	34.000	26.000
2002	18.000	35.000	31.000
2003	21.000	37.000	32.000
2004	23.000	39.000	35.000

Use 2000^{th} year as a base year, find and interpret unweighted aggregate price index.

Solution:

Year Car types

	A p_1	B p_2	C p_3	Sum	$100 \cdot \frac{\sum p_1}{\sum p_0}$
2000	15.000	32.000	25.000	72.000	100%
2001	16.000	34.000	26.000	76.000	105.56%
2002	18.000	35.000	31.000	84.000	116.67 %
2003	21.000	37.000	32.000	90.000	125 %
2004	23.000	39.000	35.000	97.000	134.47%

For example, 116.67 % indicates that as a group, prices for the three types of cars in 2002 year have increased by 16.67 % since 2000 year.

7.1.3 A weighted aggregate price index

The philosophy behind the weighted aggregate index is that each item in the group should be weighted according to its importance. In most cases the **quantity** of usage provides the best measure of importance. Suppose that we have a group of K items. Let

q_{oi} – be the quantity of i^{th} item in the base period;

p_{0i} – be the price of i^{th} item in the base period;

p_{1i} – be the price of i^{th} item in the period of t

The Laspeyres price index for the period t is given by

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

Example:

The company sells beer, wine, and soft drinks. Prices and quantity are shown below

Item	Quantity (bottles)	Unit price ($) 2000	2004
Beer	35 000	5.1	5.7
Wine	5 000	35.0	37.0
Soft drink	50 000	2.80	3.5

Compute a weighted aggregate price index for the company sales in 2004, with 2000 as the base period.

Solution:

From the information

$$q_{01} = 35000; \quad q_{02} = 5000; \quad q_{03} = 50000$$
$$p_{01} = 5.1; \quad p_{02} = 35.0; \quad p_{03} = 2.80$$
$$p_{11} = 5.7; \quad p_{12} = 37.0; \quad p_{13} = 3.5$$

$$I_p = \frac{5.7 \cdot 35000 + 37.0 \cdot 5000 + 3.5 \cdot 50000}{5.1 \cdot 35000 + 35.0 \cdot 5000 + 2.80 \cdot 50000} \cdot 100 = \frac{559500}{493500} \cdot 100 = 113.37\%$$

It means in year 2004, at the prices prevailing, the total income of company from sales in the base period would have been increased by 13.37%.

7.1.4 A weighted aggregate quantity index

Suppose that we have a group of K items. Let

p_{oi} – be the price of i^{th} item in the base period;

p_{1i} – be the price of i^{th} item in the base period;

q_{0i} – be the price of i^{th} item in the period of t

The Laspeyres quantity index for the period t is given by

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

7.2. Commonly used index numbers

a) Consumer Price Index

To most of us, inflation has come to mean increased prices and less purchasing power of dollar. The Consumer Price Index (CPI) attempts to measure the overall changes in retail prices of goods and services. The CPI uses a "market basket" of goods and services used by a typical wage earner living in a certain area. The CPI based on items grouped into seven categories, including food, housing, clothing, transportation, medical care, entertainment, and miscellaneous items.

b) Producer Price Index

The Producer Price Index (PPI), measures the monthly changes in prices in primary markets in the US. The index is based on prices for the transaction of each product in nonretail markets. All commodities sold in commercial transactions in these markets are represented, including those imported for sale. One of the common uses of this index is as leading indicator of the future trend of consumer prices and the cost of living An increase in the PPI reflects producer price increases that will eventually be passed on to the consumer through higher retail prices.

c) Dow Jones Average

The Dow Jones averages are indexes which are designed to show price trends and movements on the New York Stock Exchange. The best known of the Dow Jones indexes is the Dow Jones Industrial Average, which is based on common-stock prices of 30 industrial stocks. It is a weighted average of these stock prices where the weighs are revised from time to time to adjust for stock splits and switching of companies in the index.

7.3. Deflating a series by price indexes

Many business and economic series reported over time, such as company sales, industry sales, and inventories, are measured in dollar amounts. These time series often show an increasing growth pattern over time which is generally interpreted as showing an increase in the physical volume associated with these activities. In periods where price changes are significant the changes in the dollar amounts may be very misleading unless we are able to adjust the time series to eliminate the price-change effect. Whenever we remove the price-increase effect from the time series, we say we are **deflating the series.**

In the area of personal income and salaries we often hear discussions concerning issues such as "real salaries" or the "purchasing power". These concepts are based on the notion of deflating salary index. For example, the following table shows monthly salary income of factory workers for the past 5 years.

Year	Salary ($)	CPI (2000 base)
2000	490	100
2001	540	105
2002	585	113
2003	640	122
2004	700	138

At first glance, we see sharply increasing trend in monthly salaries, with excellent growth from $490 to $700. Should the factory workers be pleased with this growth in salary? Perhaps yes, but on the other hand, if the cost of living has increased just as fast as salary, maybe the answer is no. If we can compare purchasing power of the $490 in 2000 with the purchasing power of the $700 in 2004, we will have a better idea of the relative improvement in salaries. Table above also shows the Consumer Price Index (CPI) for the period 2000-2004. Here we use 2000 as the base for CPI. With these data we will see how the CPI can be used to deflate the index of monthly salaries. In effect we will be removing the consumer price increases from power of salaries in an attempt to measure the change in purchasing power of the wages.

The calculations used to deflate the salaries are not difficult. The deflated series is found by dividing the monthly salary in each year by the corresponding value of CPI

Year	Deflated salary
2000	($490/100) ·100=$490
2001	($540/105) ·100=$514.3
2002	($585/113) ·100=$517.7
2003	($640/122) ·100=$524.6
2004	($700/138) ·100=$507.2

What does deflated series of salaries tells us about the "real salary" or "purchasing power" of workers during 2000-2004? In terms of 2000 dollars, the monthly salary has risen from $490 to $507.2 or approximately 3.5% In fact, after we remove the price increase effect we see that factory workers are doing little more than keeping even with the inflationary price increases of the period. Thus, we see that the advantage of using price indexes to deflate a series is that we have a clearer picture of the real dollar changes that are occurring.

Exercises

1. Consider the following revenue data for the past 8 years

Year	Revenue (in millions of $)
1	18.2
2	22.6
3	25.2
4	32.6
5	38.6
6	39.7
7	40.2
8	41.3

Use year 1 as a base year to construct a relative index showing how revenues have increased.

2. The accompanying table shows monthly salaries over 5 years for three types of employees in a small company

Year	Economist	Manager	Clerk
1	450	420	320
2	490	510	350
3	520	550	390
4	560	600	400
5	600	650	420

Take year 1 as a base. In that year there were 20 economists, 15 managers and 5 clerks.

a) Find the unweighted index of monthly salary rates.

b) Find the Laspeyres index for monthly salary rates.

3. Company produces three types of items: A, B, and C. The beginning-year cost per item, the ending year cost per item, and the number of items sold in the beginning-year period are shown below:

Items	Beginning year cost	Ending year cost	Number of items sold
A	2.50	3.95	25
B	8.75	9.90	15
C	0.99	0.95	60

a) Compute the price index for single items.

b) Compute a weighted aggregate Laspeyres price index. What is your interpretation of this index value?

4. Total personal income for the 5 years 1995 to 1999 as follows

Year	Total personal income (In millions of dollars)
1995	1200
1996	1450
1997	1650
1998	1800
1999	2050

Use following Consumer Price Index below

Year	CPI
1995	100
1996	110
1997	118
1998	122
1999	130

to deflate the personal income series. What has been the percent increase in "real personal income" from 1995 to 1999? Sketch actual and real personal incomes and interpret your results.

5. The following table reports total inventories of all companies for the 5 years 1990 to 1994 as follows

Year	Total inventories (In billions of dollars)
1990	155
1991	163
1992	178
1993	198
1994	227

Use following Producers Price Index below

Year	CPI
1990	100
1991	105
1992	113
1993	122
1994	138

to deflate this series. Sketch the real and actual total inventories and interpret your findings.

Answers

1. 100%; 124.18%; 138.46%; 179.12%; 212.09%; 218.13%; 220.88%; 226.92%; **2.** a) 100%; 113.45%; 122.69%; 131.09; 140.34%; b) 100%; 113.61%; 121.89%; 131.36%; 141.12%; **3.** a) 158%; 113%; 96%; b) 120%; **4.** 1200; 1318.2; 1398.3; 1475.4; 1576.9; **5.** 155; 155; 158; 162; 164.

7.4 A nonparametric test for randomness

In the process of analyzing time series the first step is to consider a test for randomness. We will consider the **runs test,** which is a nonparametric test.

7.4.1. The runs test for the small sample sizes

To demonstrate this test, we will consider series with even number of observations. Let us consider a series of 14 observations. The data are shown below

N	Value	N	Value
1	99	8	90
2	92	9	109
3	100	10	121
4	76	11	119
5	123	12	89
6	27	13	84
7	78	14	129

First of all we write data in ascending order and find the median.

$$Median = \left[\frac{n+1}{2}\right]^{th} = \left[\frac{15}{2}\right]^{th} = \frac{7^{th}+8^{th}}{2} = \frac{92+99}{2} = 95.5$$

The run test developed here separates the observations into a subgroup above the median and a subgroup below the median. Then letting a "+"denote observations above the median and a "–" denote observations below the median we find the following pattern over the sequence

+ − + − + − − − + + + − − +

This sequence consists of a run of one "+", followed by a one run of one "-", a run of one " +", a run of one "-", a run of one "-", a run of three "-", a run of three "+", a run of two "-" , and one run of "+". In total there are $R=9$ runs.

The null hypothesis is that the series is a set of random variables. The table 7 in the Appendix gives the smallest significance level against which this null hypothesis can be rejected against the alternative of positive association between adjacent observations, as a function of R and n.

If the alternative hypothesis is two-sided hypothesis on randomness, the significance level must be doubled if it is less than 0.5. Alternatively, if the significance level, α, read from table is greater than 0.5, the corresponding significance level for the test against the two sided alternative is $2(1-\alpha)$.

In our case, $n=14$, and $R=9$. From table in the appendix we see that for $n=14$ observations, the probability under the null hypothesis of finding 9 or fewer runs is 0.791. Therefore, the null hypothesis of randomness can only be rejected against the alternative hypothesis of positive association between adjacent observations at the 79.1% significance level. We have not found strong evidence to reject the null hypothesis that series are randomness.

7.4.2 The run test for the large sample sizes

If the sample size is large $(n>20)$, the distribution of the runs under the null hypothesis can be approximated by a normal distribution. Under the null hypothesis the random variable

$$Z=\frac{R-\frac{n}{2}-1}{\sqrt{\frac{n^2-2n}{4(n-1)}}}$$

has a standard normal distribution. In formula above, R, defines the number of runs, as the number of sequences above or below the median.

We want to test the null hypothesis

H_0 :The series is random

1) If the alternative hypothesis is positive association between adjacent observations, the decision rule is

$$\text{Reject } H_0 \text{ if } Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2 - 2n}{4(n-1)}}} < -z_\alpha$$

2) If the alternative hypothesis is that series are nonrandom, then the decision rule is

$$\text{Reject } H_0 \text{ if } Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2 - 2n}{4(n-1)}}} < -z_{\alpha/2} \text{ or } Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2 - 2n}{4(n-1)}}} > z_{\alpha/2}$$

Remark:

To use MINITAB menu follow the following instructions

1. Select Stat>Nonparametrics>Runs test
2. Enter time series variable (for example, C1)
3. Select • Above and below
4. Insert value of the median
5. Select "generate forecasts"
6. Click OK.

Exercises

1. The following table shows country's industrial production index over 14 years.

Year	Index	Year	Index
1	65	8	83
2	74	9	88
3	80	10	98
4	86	11	100
5	89	12	99
6	87	13	104
7	89	14	112

Test this series for randomness using the run test.

2. The following table shows 24 annual observations on sale of certain brand of product

Year	Sales	Year	Index
1	857	13	995
2	690	14	1234
3	710	15	987
4	839	16	653
5	858	17	345
6	791	18	674
7	768	19	980
8	478	20	945
9	658	21	783
10	751	22	342
11	723	23	456
12	567	24	610

Use the large- sample variant of the runs test to test this series for randomness.

3. The following table shows annual return on a stock market index over 14 years.

Year	Return %	Year	Return %
1	-8.9	8	3.5
2	5.6	9	32.6
3	18.4	10	19.5
4	34.5	11	3.4
5	-7.8	12	16.6
6	23.4	13	32.1
7	19.6	14	0.3

Test for randomness using runs test.

4. The table shows earnings per share of a company over a period of 28 years.

Year	Earnings	Year	Earnings	Year	Earnings
1	49.3	11	19.6	21	9.5
2	34.5	12	25.7	22	19.5
3	18.3	13	29.1	23	45.2
4	23.6	14	37.5	24	76.6
5	37.7	15	48.3	25	72.1
6	31.3	16	42.1	26	67.7
7	17.9	17	36.6	27	87.3
8	12.2	18	30.1	28	46.5
9	18.0	19	21.3		
10	21.9	20	19.3		

Use the large-sample variant of the runs test to test this series for randomness.

Answers

1. Median=88.5; $R=6$; $p-\text{value}=2\cdot(0.209)=0.481$; Fail to reject H_0 at 10% level; **2.** Median = 737; $R=10$; $Z=-1.2523$; p-value $=1-0.8944=0.1056$; Reject H_0 at level above 10.56%; **3.** Median = 17.5; $R=9$; p- value = 0.791; Fail to reject H_0 at 10% level; **4.** Median = 30.7; $R=7$; $Z=-3.0813$; p- value = $=(1-0.999)=0.001$; Reject H_0 at 0.01%.

7.5. Components of time series

A critical aspect of managing any company is planning for future. In fact, the long-run success of any company is closely related to how well management is able to foresee the future and develop appropriate strategies.

Let us suppose for a moment that we have been asked to provide quarterly estimates of the sales volume for a particular product during the coming 1-year period. We will certainly want to review the actual sales data for the product in past periods. From these historical data we can identify the general level of sales and determine whether or not there is a long-term trend such as an increase or decrease in sales volume over time. The historical sales data referred to here is called a **time series.**

The classical time series model has been developed in an attempt to explain the pattern or behavior of the data in a time series. The classical model is based on the assumption that four separate components

-Trend (T)
-Cyclical (C)
-Seasonal (S)
-Irregular (I)

taken together cause the time series to assume specific values. By analyzing each of these four components separately, we hope to identify the effect each of these components has had on the time series in the past.

a) Trend component

Trend is defined as the long-term movement in a time series. In other words, an increase or decrease in the values of a variable occurring over a period of several years gives a trend.

b) Cyclical component

It is fact that most of time series show alternating sequences of points below and above the trend line. The regular pattern of sequences of points above and below the trend line is attributable to the **cyclical component** of the time.

c) Seasonal component

Seasonal component in a time series are defined as the movement that occur in a time series within one-year period. Many business activities,

such as production and sales, exhibit seasonal patterns over different time periods of a year. For example, a manufacturer of snow removal equipment and heavy clothing expects low sales activity in the spring and summer months, with peak sales occurring in the fall and winter months.

d) Irregular component

The random or chance variations in a time series are referred to as the **irregular component.** For example, strikes and natural disasters such as storms and earthquakes can cause unpredicted irregular movement in the time series. Since this component accounts for the random variability in the time series, it is unpredictable. Thus we can not attempt to predict its impact on the time series in advance.

7.6. Moving averages

Sometimes the irregular component in time series may be so large that it creates difficulties in interpretation of the time plot series. In such cases we reduce this problem by using moving averages. We can smooth any irregularities using moving averages, based on the idea that any large irregular component at any point will have a smaller effect if we average the point with its immediate neighbors. This procedure is called a simple centered $(2m+1)-$ point moving average.

Let $x_1, x_2, ..., x_n$ be n observations in a time series. A smoothed series can be obtained by using a simple centered $(2m+1)-$point moving averages

$$X_t^* = \frac{1}{2m+1}\sum_{j=-m}^{m} x_{t+j} \qquad (t = m+1, m+2, ..., n-m)$$

For instance, if we want to find 3-point moving averages, then solve

$$2m+1=3$$

and find $m=1$. If $m=1$, then the first available data will be X_2^*.

General X_t^* in this case is

$$X_t^* = \frac{x_{t-1} + x_t + x_{t+1}}{3}$$

If we set $m=2$, then a 5-point moving averages will be formed as

$$X_t^* = \frac{x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2}}{5}.$$

Example:

The following data show the sales over the past six years. Compute a simple centered 3-point moving averages to smooth data

Year	Sales
1999	2169
2000	3678
2001	2789
2002	4783
2003	1280
2004	2379

Solution:

Since we need to find 3-point moving averages then $(2m+1)=3$, and $m=1$.

Then $$X_t^* = \frac{x_{t-1} + x_t + x_{t+1}}{3}$$

Using formula above, we obtain

$$X_2^* = \frac{x_1 + x_2 + x_3}{3} = \frac{2169 + 3678 + 2789}{3} = 2878.67$$

$$X_3^* = \frac{x_2 + x_3 + x_4}{3} = \frac{3678 + 2789 + 4783}{3} = 3750$$

$$X_4^* = \frac{x_3 + x_4 + x_5}{3} = \frac{2789 + 4783 + 1280}{3} = 2950.67$$

$$X_5^* = \frac{x_4 + x_5 + x_6}{3} = \frac{4783 + 1280 + 2379}{3} = 2814$$

The original data and smoothed data are given below:

Year	Sales	X_t^*
1999	2169	--
2000	3678	2878.67
2001	2789	3750
2002	4783	2950.67
2003	1280	2814
2004	2379	--

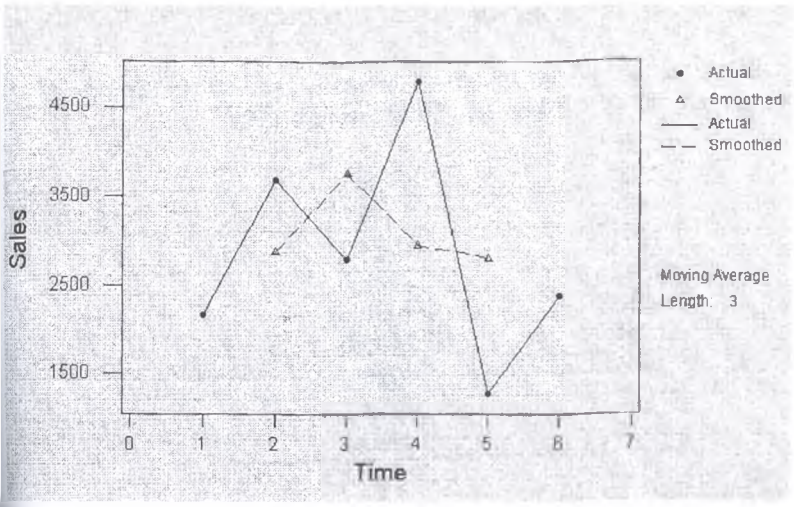

Figure 7.1

The original data and smoothed data are graphed in Figure 7.1.

Remark:

To use MINITAB menu follow the following instructions:

1. Select Stat>Time series>Moving averages
2. Enter time series variable (for example, C1)
3. Enter the moving average length
4. Click results and select summary table and results table
5. Click OK.

Exercises

1. The following table gives the gross domestic product (in billions of dollars) of the country for the years 1990 through 1997

Year	GDP
1990	1768.4
1991	1974.1
1992	2488.6
1993	3030.6
1994	3405.0
1995	4038.7
1996	4268.6
1997	4900.4

Compute a simple centered 3-point moving average for the GDP. Plot the smoothed series and comment on your results.

2. The following table shows the year-end price of gold (in dollars) over 10 consecutive years.

Year	Price ($)
1	120
2	135
3	147
4	220
5	256
6	289
7	312
8	350
9	430
10	535

Compute a simple centered 5-point moving averages for the gold price data.

Draw a time plot of the smoothed series and comment on your results.

where α is a smoothing constant whose value is fixed between 0 and 1. And standing at time n, we obtain forecasts of future values, x_{t+h} of the series by

$$\hat{x}_{t+h} = \hat{x}_h ; \qquad h = 1,2,3,.....$$

Example:

The following table shows the price of a share of common stock for a well-known computer firm over the past 8 weeks. The price shown is the closing price on the same day of the week for 8 consecutive weeks.

Week	Stock price
1	50
2	53
3	49
4	50
5	42
6	57
7	52
8	57

Use the method of simple exponential smoothing to obtain forecasts of stock price over the next three weeks. Use a smoothing constant of $\alpha = 0.4$. Graph the observed time series and the forecasts.

Solution:

As mentioned above, we let the smoothed value of the time series for the first period equal the actual first value of the time series. So,

$$\hat{x}_1 = x_1 = 50$$

To illustrate the nature of the computation, we will use a smoothing constant of $\alpha = 0.4$. Thus the smoothed value for period two becomes

$$\hat{x}_2 = \alpha \cdot \hat{x}_1 + (1-\alpha) \cdot x_2 = 0.4 \cdot 50 + 0.6 \cdot 53 = 51.8$$

The smoothed value for period 3 becomes

$$\hat{x}_3 = \alpha \cdot \hat{x}_2 + (1-\alpha) \cdot x_3 = 0.4 \cdot 51.8 + 0.6 \cdot 49 = 50.12$$

Continuing this process in the end we obtain the following complete set of smooth values shown in the table:

Week	Stock price	Smoothed time series value
1	50	50.00
2	53	51.80
3	49	50.12
4	50	50.05
5	42	45.22
6	57	52.29
7	52	52.12
8	57	55.05

Now, let us use the results of exponential smoothing to develop a forecast of the stock price for the 9^{th},10^{th}, and 11^{th} weeks. The assumption of the simple exponential smoothing is that the smoothed value of the time series at one period provides the best estimate of the time series for the next periods. Thus, the simple exponential smoothing model forecast of the stock price for the following 3 weeks is \$55.05.

$\hat{x}_9 = \$55.05$; $\hat{x}_{10} = \$55.05$; $\hat{x}_{11} = \$55.05$

Figure 7.2 shows the plot of smoothed values for the time series.

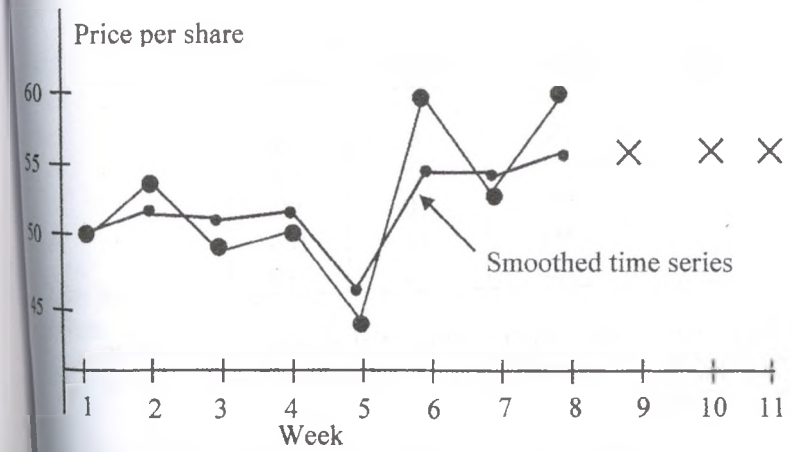

Figure 7.2. Exponential smoothing of stock price time series with smoothing constant $\alpha = 0.4$

Remark:

To use MINITAB menu follow the following instructions

1. Select Stat>Time series>Single exponential smoothing
2. Enter time series variable (for example, C1)
3. Select • *Use*
4. Insert $(1-\alpha)$
5. Select "generate forecasts"
6. **Number of forecasts:** Insert an integer to indicate number of forecasts you want.
7. **Starting from origin:** Enter a positive integer to specify a starting point for the forecasts. For example, if you specify 4 forecasts and 10 for the origin, Minitab computes forecasts for periods 11, 12, 13, and 14, based on the level and trend components at period 10. If you leave this space blank, Minitab generates forecasts from the end of the data.
8. Select Options
9. Select graphics, outputs
10. Enter 1 to the window "Use average of __ observations"
11. Click OK.

Exercises

1. The following time series shows the sales of a particular product over the past 12 month.

Month	Sales	Month	Sales
1	105	7	145
2	135	8	140
3	120	9	100
4	105	10	80
5	90	11	100
6	120	12	110

Use $\alpha = 0.3$ to compute the simple exponential smoothing to obtain forecasts of sales over the next three months. Graph the observed time series and the forecasts.

2. The following table gives the gross domestic product (in billions of dollars) of the country for the years 1990 through 1997.

Year	GDP
1990	1768.4
1991	1974.1
1992	2488.6
1993	3030.6
1994	3405.0
1995	4038.7
1996	4268.6
1997	4900.4

Use the method of simple exponential smoothing, with a smoothing constant of $\alpha = 0.7$, to obtain forecasts of GDP for the next two years. Graph the observed time series and the forecasts.

3. The following table shows the year-end price of gold (in dollars) over 10 consecutive years.

Year	Price ($)
1	120
2	135
3	147
4	220
5	256
6	289
7	312
8	350
9	430
10	535

Use the method of simple exponential smoothing, with a smoothing constant of $\alpha = 0.6$, to obtain forecasts of the price of gold in the next four years. Graph the observed time series and the forecasts.

Answers

1. $\hat{x}_{13} = \hat{x}_{14} = \hat{x}_{15} = 106.053$; **2.** $\hat{x}_9 = \hat{x}_{10} = 3869.67$; **3.** $\hat{x}_{11} = \hat{x}_{12} = \hat{x}_{13} = \hat{x}_{14} = 425.369$.

7.8. Double exponential smoothing (Holt-Winters exponential smoothing forecasting model)

Many business forecasting procedures are based on extensions of simple exponential smoothing. Double exponential smoothing (also called the Holt-Winters exponential smoothing procedure) allows for trend (seasonality also) in time series. In double exponential smoothing, a second smoothing constant, β, is included to account for the trend. We will consider a nonseasonal time series.

We denote the observed value as x_t and $\hat{x}_t$ as the estimate of the level. The trend estimate is represented as T_t. In Holt-Winters exponential smoothing forecasting model these two variables will be estimated as

$$\hat{x}_t = \alpha \cdot (\hat{x}_{t-1} + T_{t-1}) + (1-\alpha) \cdot x_t ; \qquad (0 < \alpha < 1)$$

$$T_t = \beta \cdot T_{t-1} + (1-\beta) \cdot (\hat{x}_t - \hat{x}_{t-1}) ; \qquad (0 < \beta < 1)$$

where α and β are smoothing constants whose values lie between 0 and 1.

To apply double exponential smoothing, we begin the computations by setting

$$T_2 = x_2 - x_1 \qquad \text{and} \qquad \hat{x}_2 = x_2$$

Then the above equations are applied for $t = 3,4,.....,n$. Standing at time n,

We obtain forecasts of future values, x_{n+h}, of the series by

$$\hat{x}_{n+h} = \hat{x}_n + h \cdot T_n ; \qquad (h = 1,2,3,....)$$

Example:

The sales manager needs to determine a monthly forecast for the number of men's sweaters that will be sold so he can order an appropriate amount of packing boxes. Data for the past 8 months are given below:

Month	Sales
1	145
2	165
3	175
4	149
5	167
6	156
7	176
8	195

Develop Holt-Winters double exponential smoothing model using $\alpha = 0.2$ and $\beta = 0.3$ as smoothing constants to forecast sales for the next three months.

Solution:

The initial estimates of level and trend in month 2, are

$$\hat{x}_2 = x_2 = 165 \quad \text{and} \quad T_2 = x_2 - x_1 = 165 - 145 = 20$$

This smoothing application will use $\alpha = 0.2$ and $\beta = 0.3$ and the equations

$$\hat{x}_t = 0.2 \cdot (\hat{x}_{t-1} + T_{t-1}) + 0.8 \cdot x_t;$$

$$T_t = 0.3 \cdot T_{t-1} + 0.7 \cdot (\hat{x}_t - \hat{x}_{t-1})$$

Then for $t = 3$:

$$\hat{x}_3 = 0.2 \cdot (\hat{x}_2 + T_2) + 0.8 \cdot x_3 = 0.2 \cdot (165 + 20) + 0.8 \cdot 175 = 177$$

and in addition

$$T_3 = 0.3 \cdot T_2 + 0.7 \cdot (\hat{x}_3 - \hat{x}_2) = 0.3 \cdot 20 + 0.7 \cdot (177 - 165) = 14.4$$

Then for $t = 4$:

$$\hat{x}_4 = 0.2 \cdot (\hat{x}_3 + T_3) + 0.8 \cdot x_4 =$$

$$= 0.2 \cdot (177 + 14.4) + 0.8 \cdot 149 = 157.48$$

nd in addition

$$= 0.3 \cdot T_3 + 0.7 \cdot (\hat{x}_4 - \hat{x}_3) = 0.3 \cdot 14.4 + 0.7 \cdot (157.48 - 177) = -9.344$$

For $t=5$:

$$\hat{x}_5 = 0.2 \cdot (\hat{x}_4 + T_4) + 0.8 \cdot x_5 =$$
$$= 0.2 \cdot (157.48 - 9.344) + 0.8 \cdot 167 = 163.23$$

$$T_5 = 0.3 \cdot T_4 + 0.7 \cdot (\hat{x}_5 - \hat{x}_4) = 0.3 \cdot (-9.344) + 0.7 \cdot (163.23 - 157.48) = 1.22$$

For $t=6$:

$$\hat{x}_6 = 0.2 \cdot (\hat{x}_5 + T_5) + 0.8 \cdot x_6 = 0.2 \cdot (163.23 + 1.22) + 0.8 \cdot 156 = 157.7$$

$$T_6 = 0.3 \cdot T_5 + 0.7 \cdot (\hat{x}_6 - \hat{x}_5) = 0.3 \cdot 1.22 + 0.7 \cdot (157.7 - 163.23) = -3.51$$

For $t=7$:

$$\hat{x}_7 = 0.2 \cdot (\hat{x}_6 + T_6) + 0.8 \cdot x_7 = 0.2 \cdot (157.7 - 3.51) + 0.8 \cdot 176 = 171.64$$

$$T_7 = 0.3 \cdot T_6 + 0.7 \cdot (\hat{x}_7 - \hat{x}_6) = 0.3 \cdot (-1.04) + 0.7 \cdot (171.64 - 157.7) = 8.7$$

For $t=8$:

$$\hat{x}_8 = 0.2 \cdot (\hat{x}_7 + T_7) + 0.8 \cdot x_8 = 0.2 \cdot (171.64 + 8.7) + 0.8 \cdot 195 = 192.2$$

$$T_8 = 0.3 \cdot T_7 + 0.7 \cdot (\hat{x}_8 - \hat{x}_7) = 0.3 \cdot 8.7 + 0.7 \cdot (192.2 - 171.64) = 17.00$$

In general for h periods forecasting is

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

The most recent level and trend estimates are

$$\hat{x}_8 = 192.2; \qquad T_8 = 17.00$$

Then the forecasts for the next three months are

$$\hat{x}_9 = 192.2 + 1 \cdot 17.00 = 209.2$$
$$\hat{x}_{10} = 192.2 + 2 \cdot 17.00 = 226.2$$
$$\hat{x}_{11} = 192.2 + 3 \cdot 17.00 = 243.2$$

The results of these calculations are shown below:

Month	Sales	$\hat{x}_t$	MINITAB solution
1	145	--	146.150
2	165	165	161.457
3	175	177	174.503
4	149	157.48	156.590
5	167	163.23	163.157
6	156	157.7	157.823
7	176	171.64	171.735
8	195	192.2	192.106

The last column of the table above contains MINITAB solution. According to MINITAB solution the predictions are

$\hat{x}_9 = 209.004$; $\hat{x}_{10} = 225.902$; $\hat{x}_{11} = 242.800$

The values calculated by the MINITAB program differ slightly from those in the third column of the table above. The MINITAB procedures will generally provide slightly better forecasts compared to the more simplified procedure we have shown. The observed time series and forecasts are shown in Figure 7.3.

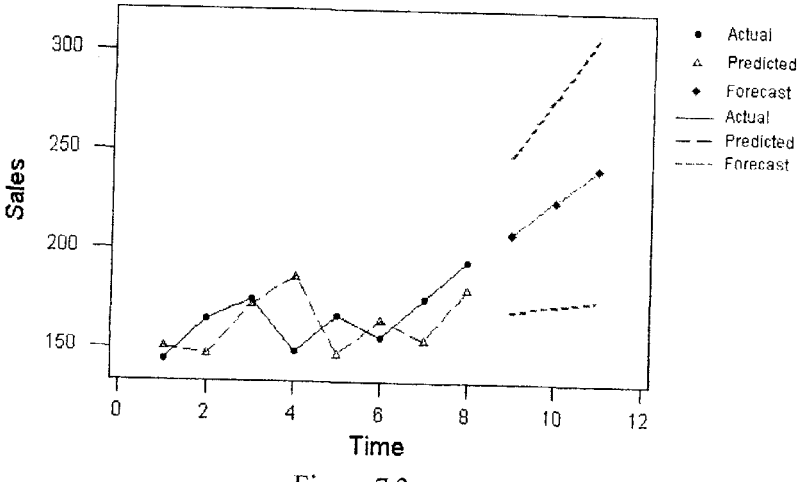

Figure 7.3

Remark:

To use MINITAB menu follow the following instructions:

1. Select Stat>Time series>Double exponential smoothing
2. Enter time series variable (for example, C1)
3. Select • *Use*
4. Enter $1-\alpha$ -for level
5. Enter $1-\beta$ - for trend
6. Select "generate forecasts"
7. Enter number of forecasting
8. Enter number of starting point for forecasting
9. Select Options
10. Select graphics, outputs
11. Click OK.

Exercises

1. The table shows the sales of a particular brand of automatic dishwasher detergent in a store over a period of 7 weeks.

Week	Sale
1	22
2	23
3	19
4	16
5	21
6	28
7	19

Use the Holt-Winters procedure, with smoothing constants $\alpha = 0.4$ and $\beta = 0.5$ to obtain forecasts over the next 5 weeks.

2. The following table shows manufacturing monthly earnings over 12 months.

Month	Earnings	Month	Earnings
1	350	7	420
2	370	8	400
3	340	9	350
4	370	10	420
5	390	11	360
6	410	12	450

Use the Holt-Winters procedure, with smoothing constants $\alpha = 0.3$ and $\beta = 0.4$ to obtain forecasts for the next 3 months. Graph the data and forecasts.

3. The following table shows percentage profit of a company over a period of 8 years.

Year	Profit margin
1	7.6
2	6.4
3	6.9
4	7.9
5	8.1
6	7.0
7	8.2
8	6.2

Find forecasts for the next three years, using the Holt-Winters procedure, with smoothing constants $\alpha = 0.7$ and $\beta = 0.6$. Graph the data and forecasts.

Answers

1. 22.31; 22.50; 22.68; 22.87; 23.06; **2.** 449.01; 474.56; 500.12; **3.** 7.29; 7.23; 7.17.

APPENDIX

Table 1: Cumulative distribution function of the standard normal distribution

z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$
.00	.5000										
.01	.5040	.21	.5832	.41	.6591	.61	.7291	.81	.7910	1.01	.8438
.02	.5080	.22	.5871	.42	.6628	.62	.7324	.82	.7939	1.02	.8461
.03	.5120	.23	.5910	.43	.6664	.63	.7357	.83	.7967	1.03	.8485
.04	.5160	.24	.5948	.44	.6700	.64	.7389	.84	.7995	1.04	.8508
.05	.5199	.25	.5987	.45	.6736	.65	.7422	.85	.8023	1.05	.8531
.06	.5239	.26	.6026	.46	.6772	.66	.7454	.86	.8051	1.06	.8554
.07	.5279	.27	.6064	.47	.6803	.67	.7486	.87	.8078	1.07	.8577
.08	.5319	.28	.6103	.48	.6844	.68	.7517	.88	.8106	1.08	.8599
.09	.5359	.29	.6141	.49	.6879	.69	.7549	.89	.8133	1.09	.8621
.10	.5398	.30	.6179	.50	.6915	.70	.7580	.90	.8159	1.10	.8643
.11	.5438	.31	.6217	.51	.6950	.71	.7611	.91	.8186	1.11	.8665
.12	.5478	.32	.6255	.52	.6985	.72	.7642	.92	.8212	1.12	.8686
.13	.5517	.33	.6293	.53	.7019	.73	.7673	.93	.8238	1.13	.8708
.14	.5557	.34	.6331	.54	.7054	.74	.7704	.94	.8264	1.14	.8729
.15	.5596	.35	.6368	.55	.7088	.75	.7734	.95	.8289	1.15	.8749
.16	.5636	.36	.6406	.56	.7123	.76	.7764	.96	.8315	1.16	.8770
.17	.5675	.37	.6443	.57	.7157	.77	.7794	.97	.8340	1.17	.8790
.18	.5714	.38	.6480	.58	.7190	.78	.7823	.98	.8365	1.18	.8810
.19	.5753	.39	.6517	.59	.7224	.79	.7852	.99	.8389	1.19	.8830
.20	.5793	.40	.6554	.60	.7257	.80	.7881	1.00	.8413	1.20	.8849

1.21	.8869	1.46	.9279	1.71	.9564	1.96	.9750	2.21	.9864	2.46	.9931
1.22	.8888	1.47	.9292	1.72	.9573	1.97	.9756	2.22	.9868	2.47	.9932
1.23	.8907	1.48	.9306	1.73	.9582	1.98	.9761	2.23	.9871	2.48	.9934
1.24	.8925	1.49	.9319	1.74	.9591	1.99	.9767	2.24	.9875	2.49	.9936
1.25	.8944	1.50	.9332	1.75	.9599	2.00	.9772	2.25	.9878	2.50	.9938
1.26	.8962	1.51	.9345	1.76	.9608	2.01	.9778	2.26	.9881	2.51	.9940
1.27	.8980	1.52	.9357	1.77	.9616	2.02	.9783	2.27	.9884	2.52	.9941
1.28	.8997	1.53	.9370	1.78	.9615	2.03	.9788	2.28	.9887	2.53	.9943
1.29	.9015	1.54	.9382	1.79	.9633	2.04	.9793	2.29	.9890	2.54	.9945
1.30	.9032	1.55	.9394	1.80	.9641	2.05	.9798	2.30	.9893	2.55	.9946
1.31	.9049	1.56	.9406	1.81	.9649	2.06	.9803	2.31	.9896	2.56	.9948
1.32	.9066	1.57	.9418	1.82	.9656	2.07	.9808	2.32	.9898	2.57	.9949
1.33	.9082	1.58	.9429	1.83	.9664	2.08	.9812	2.33	.9901	2.58	.9951
1.34	.9099	1.59	.9441	1.84	.9671	2.09	.9817	2.34	.9904	2.59	.9952
1.35	.9115	1.60	.9452	1.85	.9678	2.10	.9821	2.35	.9906	2.60	.9953
1.36	.9131	1.61	.9463	1.86	.9686	2.11	.9826	2.36	.9909	2.61	.9955
1.37	.9147	1.62	.9474	1.87	.9693	2.12	.9830	2.37	.9911	2.62	.9956
1.38	.9162	1.63	.9484	1.88	.9699	2.13	.9834	2.38	.9913	2.63	.9957
1.39	.9177	1.64	.9495	1.89	.9706	2.14	.9838	2.39	.9916	2.64	.9959
1.40	.9192	1.65	.9505	1.90	.9713	2.15	.9842	2.40	.9918	2.65	.9960
1.41	.9207	1.66	.9515	1.91	.9719	2.16	.9846	2.41	.9920	2.66	.9961
1.42	.9222	1.67	.9525	1.92	.9726	2.17	.9850	2.42	.9922	2.67	.9962
1.43	.9236	1.68	.9535	1.93	.9732	2.18	.9854	2.43	.9925	2.68	.9963
1.44	.9251	1.69	.9545	1.94	.9738	2.19	.9857	2.44	.9927	2.69	.9964
1.45	.9265	1.70	.9554	1.95	.9744	2.20	.9861	2.45	.9929	2.70	.9965

Table 1: Cumulative distribution function of the standard normal distribution (continue)

z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$
2.71	.9966	2.91	.9982	3.11	.9991	3.31	.9995	3.51	.9998	3.71	.9999	3.91	1.0000
2.72	.9967	2.92	.9982	3.12	.9991	3.32	.9996	3.52	.9998	3.72	.9999	3.92	1.0000
2.73	.9968	2.93	.9983	3.13	.9991	3.33	.9996	3.53	.9998	3.73	.9999	3.93	1.0000
2.74	.9969	2.94	.9984	3.14	.9992	3.34	.9996	3.54	.9998	3.74	.9999	3.94	1.0000
2.75	.9970	2.95	.9984	3.15	.9992	3.35	.9996	3.55	.9998	3.75	.9999	3.95	1.0000
2.76	.9971	2.96	.9985	3.16	.9992	3.36	.9996	3.56	.9998	3.76	.9999	3.96	1.0000
2.77	.9972	2.97	.9985	3.17	.9992	3.37	.9996	3.57	.9998	3.77	.9999	3.97	1.0000
2.78	.9973	2.98	.9986	3.18	.9993	3.38	.9996	3.58	.9998	3.78	.9999	3.98	1.0000
2.79	.9974	2.99	.9986	3.19	.9993	3.39	.9997	3.59	.9998	3.79	.9999	3.99	1.0000
2.80	.9974	3.00	.9986	3.20	.9993	3.40	.9997	3.60	.9998	3.80	.9999		
2.81	.9975	3.01	.9987	3.21	.9993	3.41	.9997	3.61	.9998	3.81	.9999		
2.82	.9976	3.02	.9987	3.22	.9994	3.42	.9997	3.62	.9999	3.82	.9999		
2.83	.9977	3.03	.9988	3.23	.9994	3.43	.9997	3.63	.9999	3.83	.9999		
2.84	.9977	3.04	.9988	3.24	.9994	3.44	.9997	3.64	.9999	3.84	.9999		
2.85	.9978	3.05	.9989	3.25	.9994	3.45	.9997	3.65	.9999	3.85	.9999		
2.86	.9979	3.06	.9989	3.26	.9994	3.46	.9997	3.66	.9999	3.86	.9999		
2.87	.9979	3.07	.9989	3.27	.9995	3.47	.9997	3.67	.9999	3.87	.9999		
2.88	.9980	3.08	.9990	3.28	.9995	3.48	.9997	3.68	.9999	3.88	.9999		
2.89	.9981	3.09	.9990	3.29	.9995	3.49	.9998	3.69	.9999	3.89	1.0000		
2.90	.9981	3.10	.9990	3.30	.9995	3.50	.9998	3.70	.9999	3.90	1.0000		

Table 2: Cut-off point of Student's *t* distribution

ν	α				
	0.100	**0.050**	**0.025**	**0.010**	**0.005**
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.160
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
∞	1.282	1.645	1.960	2.326	2.576

Table 3: Chi- square distribution

α									
.995	**.990**	**.975**	**.950**	**.900**	**.100**	**.050**	**.025**	**.010**	**.005**
0.0^4393	0.0^3157	0.0^3982	0.0^2393	0.0158	2.71	3.84	5.02	6.63	7.88
0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
59.20	61.75	65.65	69.16	73.29	107.6	113.1	118.1	124.1	128.3
67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

Table 4: Cutoff point of the distribution of the Wicoxon test statistic

n	α				
	0.005	**0.01**	**0.025**	**0.05**	**0.10**
4	0	0	0	0	1
5	0	0	0	1	3
6	0	0	1	3	4
7	0	1	3	4	6
8	1	2	4	6	9
9	2	4	6	9	11
10	4	6	9	11	15
11	6	8	11	14	18
12	8	10	14	18	22
13	10	13	18	22	27
14	13	16	22	26	32
15	16	20	26	31	37
16	20	24	30	36	43
17	24	28	35	42	49
18	28	33	41	48	56
19	33	38	47	54	63
20	38	44	53	61	70

Table 5: Cutoff point of the distribution of the Spearman's rank correlation coefficient

n	α			
	0.05	**0.025**	**0.01**	**0.005**
5	.900	-	-	-
6	.829	.886	.943	-
7	.714	.786	.893	-
8	.643	.738	.833	.881
9	.600	.683	.783	.833
10	.564	.648	.745	.794
11	.523	.623	.736	.818
12	.497	.591	.703	.780
13	.475	.566	.673	.745
14	.457	.545	.646	.716
15	.441	.525	.623	.689
16	.425	.507	.601	.666
17	.412	.490	.582	.645
18	.399	.476	.564	.625
19	.388	.462	.549	.608
20	.377	.450	.534	.591
21	.368	.438	.521	.576
22	.359	.428	.508	.562
23	.351	.418	.496	.549
24	.343	.409	.485	.537
25	.336	.400	.475	.526
26	.329	.392	.465	.515
27	.323	.385	.456	.505
28	.317	.377	.448	.496
29	.311	.370	.440	.487
30	.305	.364	.432	.478

Table 2: Cutoff points for the F distribution

$\alpha = 0.05$

ν_2 – denominator

ν_1 – numerator

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

$\alpha = 0.01$

ν_2 – denominator

ν_1 – numerator

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.48	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.38	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Table 7. Cumulative distribution function of the runs test statistic

n	***R***																		
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
6	.100	.300	.700	.900	1.00														
8	.029	.114	.371	.629	.886	.971	1.00												
10	.008	.040	.167	.357	.643	.833	.960	.992	1.00										
12	.002	.013	.067	.175	.392	.608	.825	.933	.987	.998	1.00								
14	.001	.004	.025	.078	.209	.383	.617	.791	.922	.975	.996	.999	1.00						
16	.000	.001	.009	.032	.100	.214	.405	.595	.786	.900	.968	.991	.999	1.00	1.00				
18	.000	.000	.003	.012	.044	.109	.238	.399	.601	.762	.891	.956	.988	.997	1.00	1.00	1.00		
20	.000	.000	.001	.004	.019	.051	.128	.242	.414	.586	.758	.872	.949	.981	.996	.999	1.00	1.00	1.00

References

1.Paul Newbold, William L. Carlson, Betty M. Thorne, "Statistics for business and economics", Pearson education, Inc., upper saddle river, New jersey, 2003.
2.David R. Anderson, Dennis J. Sweeney, Thomas A.Williams, " Introduction to statistics" West Publishing company, St. Paul, Minnesota., 1981.
3. H. T. Hayslett, MS, : Statistics", British library cataloguing in publication data, 1981.
4. Paul Newbold, "Statistics for business and economics" Prentice – Hall Inc., 1995.
5. James T.McClave, P.George Benson; Therry Sincich, " Statistics for business and economics", Prentice-Hall, Inc., 1998.
6. David F.Groebner, Patric W. Shannon, Phillip C.Fry, Kent D.Smith, Business statistics, Pearson education, Inc., upper saddle river, New jersey, 2005.

Гумбат Наби оглы Алиев

СТАТИСТИКА
Задачи и упражнения. Часть 1

Humbat Nabi oglu Aliyev

STATISTICS
SOLVED PROBLEMS AND EXERCISES. PART 2

Подписано в печать 16.02.2009

Формат 60 x 84 1/16. Бумага тип. Ризограф.
Усл.печ.15,1 л. Уч.-изд. 15,3 л. Тираж 500.

Заказ № 700

Цена договорная

Издание Казахской головной архитектурно-строительной академии
Издательский дом КазГАСА «Строительство и архитектура»
480043, г.Алматы, ул. К.Рыскулбекова, 28