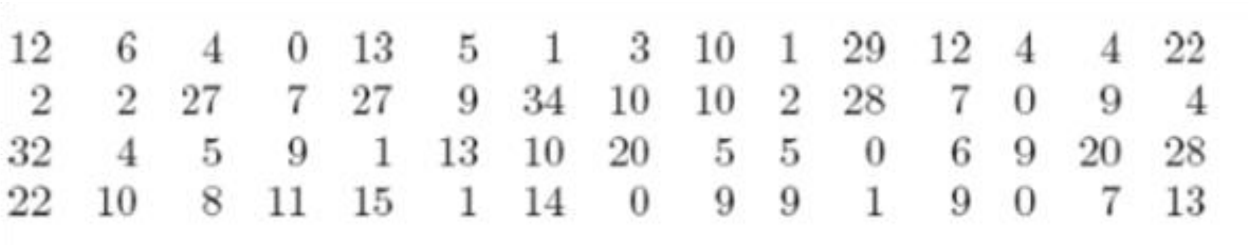
Chapter 10, Problem 1E
(0)

Problem
Does the number of unsolicited (spam) e-mails follow a Poisson distribution? Here is the record of the number of spam e-mails received during 60 consecutive days (Spam). Choose suitable bins and conduct a goodness-of-fit test at the 1% level of significance.



12	6	4	0	13	5	1	3	10	1	29	12	4	4	22
2	2	27	7	27	9	34	10	10	2	28	7	0	9	4
32	4	5	9	1	13	10	20	5	5	0	6	9	20	28
22	10	8	11	15	1	14	0	9	9	1	9	0	7	13

Step-by-step solution

Show all steps

Step 1/6
The null hypothesis and alternative hypotheses are,

$H_0:$ The number of mails follows a Poisson distribution.

$H_1:$ The number of mails does not follow a Poisson distribution.

Let the level of significance be $\alpha = 0.01$.

The mean is,

$$\begin{aligned}\lambda &= \overline{x}\\ &= \frac{\sum x}{N}\\ &= \frac{0+0+\cdots+34}{60}\\ &= \frac{600}{60}\\ &= 10\end{aligned}$$

Step 2/6
Now we will form following table using the mean value. This value of the sample mean is used as the estimate of ? for the purposes of finding the probabilities from the tables of the Poisson distribution. From Table, for $\lambda = 10$, the frequency of *X* successes (X = 0, 1, 2, 3, 4, 5, 6, 7, 8 ...) can be determined. The theoretical frequency for each is obtained by multiplying the appropriate Poisson probability by the sample size *n*.

The probability mass function of a Poisson distribution is,

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

The formula for the expected frequency is,

$$E = N \times P(x)$$

The following table shows the Poisson probabilities, observed and expected frequencies.

x	Observed frequency, *O*	Poisson Probabilities, $P(x)$	Expected frequency, *E*
0	5	0.000045	0
1	5	0.000454	0
2	3	0.002270	0
3	1	0.007567	0
4	5	0.018917	1
5	4	0.037833	2
6	2	0.063055	4
7	3	0.090079	5
8	1	0.112599	7
9	7	0.125110	8
10	5	0.125110	8
11 or more	19	0.417000	25
Total	60		60

Step 3/6

Since, some of the expected frequencies less than 5, so the revised contingency table is,

frequency, *O*	frequency, *E*
25	7
3	5
1	7
7	8
5	8
19	25
$\sum E = 60$	$\sum O = 60$

Step 4/6

The chi square statistic is,

$$
\begin{aligned}
\chi^2 &= \sum \frac{\text{(expcetd-observed)}}{\text{expected}} \\
&= \frac{(25-7)^2}{7} + \frac{(3-5)^2}{5} + \frac{(1-7)^2}{7} + \frac{(7-8)^2}{8} + \frac{(5-8)^2}{8} + \frac{(19-25)^2}{25} \\
&= 54.919
\end{aligned}
$$

Step 5/6

One degrees of freedom is lost because of the linear constraint $\sum E = \sum O$, I degrees of freedom is lost because the parameter $\overline{x} = \lambda$ has been estimated from the known data and is then used for computing the expected frequencies, 6 degrees of freedom are lost because of pooling the first 7 frequencies which are less than 5.

The degrees of freedom is,

$$
\begin{aligned}
df &= n-1-1-6 \\
&= 12-8 \\
&= 4
\end{aligned}
$$

Step 6/6

Using the excel function formula, the *P*-value is,

$$
\begin{aligned}
P\text{-value} &= \left(\text{=CHIDIST(54.919,4)}\right) \\
&\approx 0
\end{aligned}
$$

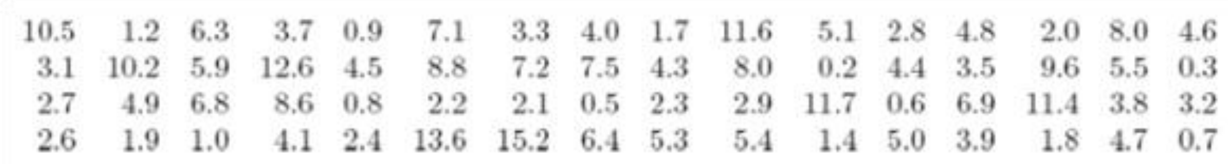
The *P*-value is less than the significance level of 0.01. Reject the null hypothesis. Therefore, it can be concluded that the number of mails does not follow a Poisson distribution.

Chapter 10, Problem 2E

(0)

Problem

Applying the theory of M/M/1 queuing systems, we assume that the service times follow Exponential distribution. The following service times, in minutes, have been observed during 24 hours of operation (ServiceTimes):



10.5	1.2	6.3	3.7	0.9	7.1	3.3	4.0	1.7	11.6	5.1	2.8	4.8	2.0	8.0	4.6
3.1	10.2	5.9	12.6	4.5	8.8	7.2	7.5	4.3	8.0	0.2	4.4	3.5	9.6	5.5	0.3
2.7	4.9	6.8	8.6	0.8	2.2	2.1	0.5	2.3	2.9	11.7	0.6	6.9	11.4	3.8	3.2
2.6	1.9	1.0	4.1	2.4	13.6	15.2	6.4	5.3	5.4	1.4	5.0	3.9	1.8	4.7	0.7

Is the assumption of Exponentiality supported by these data?

Step-by-step solution

Show all steps

100% (1 rating) for this solution

Step 1/2

The data of the service time during 24 hours of operation is given as follows :

10.5	1.2	6.3	3.7	0.9	7.1	3.3	4.0	1.7	11.6	5.1	2.8	4.8	2.0	8.0	4.6
3.1	10.2	5.9	12.6	4.5	8.8	7.2	7.5	4.3	8.0	0.2	4.4	3.5	9.6	5.5	0.3
2.7	4.9	6.8	8.6	0.8	2.2	2.1	0.5	2.3	2.9	11.7	0.6	6.9	11.4	3.8	3.2
2.6	1.9	1.0	4.1	2.4	13.6	15.2	6.4	5.3	5.4	1.4	5.0	3.9	1.8	4.7	0.7

The mean of the above values is $\overline{X} = 5.0$ and total observations *N* = 64.

Now after arranging data into increasing order, we form a continuous type frequency table of the above data which is given below :

Class Interval	Frequency
0 - 2	13
2 - 4	16
4 - 6	15
6 - 8	7
8 - 10	5
10 - 12	5
12 - 14	2
14 - 16	1
Total	64

The cumulative distribution function *F*(*x*) of the exponential distribution is given by,

$$
\begin{aligned}
&F(x) \\
&= P(X \le x) \\
&= 1 - e^{-\lambda x} \quad 0 \le x < \infty
\end{aligned}
$$

Where the parameter

$$
\begin{aligned}
\lambda &= \frac{1}{\overline{X}} \\
&= \frac{1}{5.0} \\
&= 0.2
\end{aligned}
$$

So,

$$F(x)=1-e^{-0.2x} \quad 0 \le x < \infty$$

Step 2/2

Using the above function we can calculate the expected frequencies for *i*th class $(a_i - b_i)$ by using

$$e_i = F(b_i) - F(a_i)$$

In this way the table of expected, observed frequencies and other required columns comes out to be,

Class Interval	o_i	e_i	$(o_i - e_i)^2 / e_i$
0 - 2	13	21.10	3.11
2 - 4	16	14.14	0.24
4 - 6	15	9.48	3.21
6 - 8	7	6.36	0.07
8 - 10	5	4.26	0.13
10 - 12	5	2.86	1.61
12 - 14	2	1.91	0.00
14 - 16	1	3.89	2.15
Total	**64**	**64**	**10.52**

Now to test the hypothesis at 5% significance level whether the given data follows an exponential distribution, the test statistic is

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$
$$= 10.52$$

Degrees of freedom in this case is

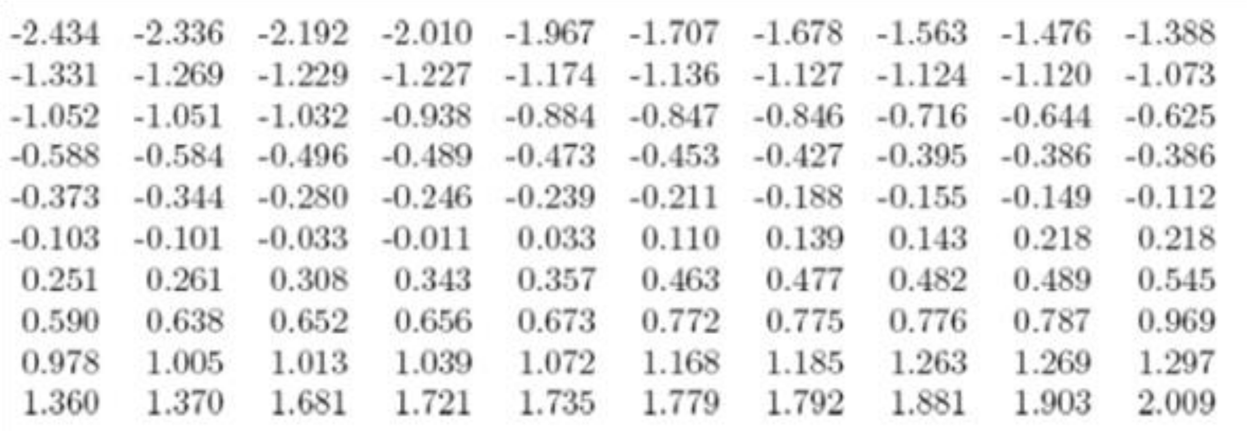
$$df = n - 1$$
$$= 8 - 1$$
$$= 7$$

The *p* – value for the above value of test statistic at 7 *df* can be found using MS-Excel or some other statistical software and founded to be 0.161. Which is more than the significance level 0.05.

We conclude that there is no evidence against an exponential distribution of the service time.

Chapter 10, Problem 3E
(0)

Problem
The following sample, RandomNumbers is collected to verify the accuracy of a new random number generator (it is already ordered for your convenience).



-2.434	-2.336	-2.192	-2.010	-1.967	-1.707	-1.678	-1.563	-1.476	-1.388
-1.331	-1.269	-1.229	-1.227	-1.174	-1.136	-1.127	-1.124	-1.120	-1.073
-1.052	-1.051	-1.032	-0.938	-0.884	-0.847	-0.846	-0.716	-0.644	-0.625
-0.588	-0.584	-0.496	-0.489	-0.473	-0.453	-0.427	-0.395	-0.386	-0.386
-0.373	-0.344	-0.280	-0.246	-0.239	-0.211	-0.188	-0.155	-0.149	-0.112
-0.103	-0.101	-0.033	-0.011	0.033	0.110	0.139	0.143	0.218	0.218
0.251	0.261	0.308	0.343	0.357	0.463	0.477	0.482	0.489	0.545
0.590	0.638	0.652	0.656	0.673	0.772	0.775	0.776	0.787	0.969
0.978	1.005	1.013	1.039	1.072	1.168	1.185	1.263	1.269	1.297
1.360	1.370	1.681	1.721	1.735	1.779	1.792	1.881	1.903	2.009

(a) Apply the χ^2 goodness-of-fit test to check if this sample comes from the Standard Normal distribution.

(b) Test if this sample comes from the Uniform(-3,3) distribution.

(c) Is it theoretically possible to accept both null hypotheses in (a) and (b) although they are contradicting to each other? Why does it make sense?

Step-by-step solution

Show all steps

Step 1/5
The data of the random number generator is given as follows :

-2.434	-2.336	-2.192	-2.010	-1.967	-1.707	-1.678	-1.563	-1.476	-1.388
-1.331	-1.269	-1.229	-1.227	-1.174	-1.136	-1.127	-1.124	-1.120	-1.073
-1.052	-1.051	-1.032	-0.938	-0.884	-0.847	-0.846	-0.716	-0.644	-0.625
-0.588	-0.584	-0.496	-0.489	-0.473	-0.453	-0.427	-0.395	-0.386	-0.386
-0.373	-0.344	-0.280	-0.246	-0.239	-0.211	-0.188	-0.155	-0.149	-0.112
-0.103	-0.101	-0.033	-0.011	0.033	0.110	0.139	0.143	0.218	0.218
0.251	0.261	0.308	0.343	0.357	0.463	0.477	0.482	0.489	0.545
0.590	0.638	0.652	0.656	0.673	0.772	0.775	0.776	0.787	0.969
0.978	1.005	1.013	1.039	1.072	1.168	1.185	1.263	1.269	1.297
1.360	1.370	1.681	1.721	1.735	1.779	1.792	1.881	1.903	2.009

The mean of the above values is $\overline{X} = -0.058$, standard deviation $\sigma = 1.058$and total observations N = 100.

Now after arranging data into increasing order, we form a continuous type frequency table of the above data which is given below :

Class Interval	Frequency
below -2.0	4
-2.0 to -1.5	4
-1.5 to -1.0	15
-1.0 to -0.5	9
-0.5 to 0.0	22
0.0 to 0.5	15
0.5 to 1.0	12
1.0 to 1.5	11
1.5 to 2.0	7
2.0 & Above	1
Total	100

Step 2/5

(a)

Using the table of normal distribution the table of expected, observed frequencies and other required columns comes out to be,

Class Interval	o_i	e_i	$(o_i - e_i)^2 / e_i$
below -2.0	4	3.32	0.14
-2.0 to -1.5	4	5.32	0.33
-1.5 to -1.0	15	10.02	2.48
-1.0 to -0.5	9	15.14	2.49
-0.5 to 0.0	22	18.38	0.71
0.0 to 0.5	15	17.92	0.48
0.5 to 1.0	12	14.03	0.29
1.0 to 1.5	11	8.82	0.54
1.5 to 2.0	7	4.46	1.45
2.0 & Above	1	2.59	0.97
Total	100	100	9.88332

Now to test the hypothesis at 5% significance level whether the given data follows a normal distribution, the test statistic is

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$
$$= 9.88$$

Degrees of freedom in this case is

$$\begin{aligned} df &= n-1 \\ &= 10-1 \\ &= 9 \end{aligned}$$

The *p* – value for the above value of test statistic at 9 *df* can be found using MS-Excel or some other statistical software and founded to be 0.360. Which is more than the significance level 0.05.

We conclude that there is no evidence against the data follows a normal distribution.

Step 3/5
(b)

The pdf of Uniform distribution is given by,

$$f(x)$$
$$= \frac{1}{b-a} \; a \le x \le b$$

In this problem, $a = -3$, $b = +3$

Using the above formula the table of expected, observed frequencies and other required columns comes out to be,

Class Interval	o_i	e_i	$(o_i - e_i)^2 / e_i$
below -2.0	4	16.67	9.63
-2.0 to -1.5	4	8.33	2.25
-1.5 to -1.0	15	8.33	5.33
-1.0 to -0.5	9	8.33	0.05
-0.5 to 0.0	22	8.33	22.41
0.0 to 0.5	15	8.33	5.33
0.5 to 1.0	12	8.33	1.61
1.0 to 1.5	11	8.33	0.85
1.5 to 2.0	7	8.33	0.21

Step 4/5
2.0 & Above

1

16.67

14.73

Total

100

100

62.42

Now to test the hypothesis at 5% significance level whether the given data follows a Uniform distribution, the test statistic is

$$\begin{aligned}\chi^2 &= \sum \frac{(o_i - e_i)^2}{e_i} \\ &= 62.42\end{aligned}$$

Degrees of freedom in this case is

$$\begin{aligned}df &= n-1 \\ &= 10-1 \\ &= 9\end{aligned}$$

The *p* – value for the above value of test statistic at 9 *df* can be found using MS-Excel or some other statistical software and founded to be < 0.001. Which is much lower than the significance level 0.05.

We conclude that there is strong evidence against the data follows a Uniform distribution.

Step 5/5
(c)

Theoretically it is possible that a data follows Normal and Uniform distributions simultaneously for a large sample according to the central limit theorem.

Chapter 10, Problem 4E
(0)

Problem
In Example 10.3 on p. 319, we tested whether the number of concurrent users is approximately Normal. How does the result of the Chi-square test depend on our choice of bins? For the same data, test the assumption of a Normal distribution using a different set of bins $B_1, \ldots, B_N$.

Step-by-step solution

Show all steps

Step 1/2

The basic assumption about the applicability of the χ^2 test is that the expected frequency of every bin must be more than 5 which makes χ^2 distribution continuous. So choosing a larger number of bins may violate this condition and so the test will generate erroneous results. Further the degrees of freedom used in χ^2 test also depend on number of bins. More the number of bins means more the degrees of freedom, that will increase the probability of non significance of null hypothesis. In other words, for making larger number of bins there are more chances of acceptance of hypothesis that data follows the normal distribution and vice versa.

Step 2/2

Let we make only four bins and will test the hypothesis at 5% significance level whether the given data follows a Normal distribution.

As calculated previously,

The mean of the data values is $\bar{X} = 17.95$, standard deviation $\sigma = 3.13$ and total observations N = 100. The frequency table is given as follows :

k	Class Interval	Frequency
1	below 14.5	6
2	14.5 - 18.0	21
3	18.0 - 21.5	13
4	above 21.5	10
Total	50	

Using the table of normal distribution the table of expected, observed frequencies and other required columns comes out to be,

k	Class Interval	o_i	e_i	$(o_i - e_i)^2 / e_i$
1	below 14.5	6	6.76	0.09
2	14.5 - 18.0	21	18.56	0.32
3	18.0 - 21.5	13	18.26	1.52
4	above 21.5	10	6.42	2.00
	Total	50.00	50.00	3.92

The *p* – value for the above value of test statistic at 1 *df* can be found using MS-Excel or some other statistical software and founded to be 0.048. Which is less than the significance level 0.05.

We conclude that there is significant evidence against the data follows a normal distribution.

So we see that after altering the no. of bins, the inference found just reversed the previous one.

Chapter 10, Problem 5E
(0)

Problem
Show that the sample size is too small in Example 10.9 on p. 327 to conduct a χ^2 goodness-of-fit test of Normal distribution that involves estimation of its both parameters.

Step-by-step solution

Show all steps

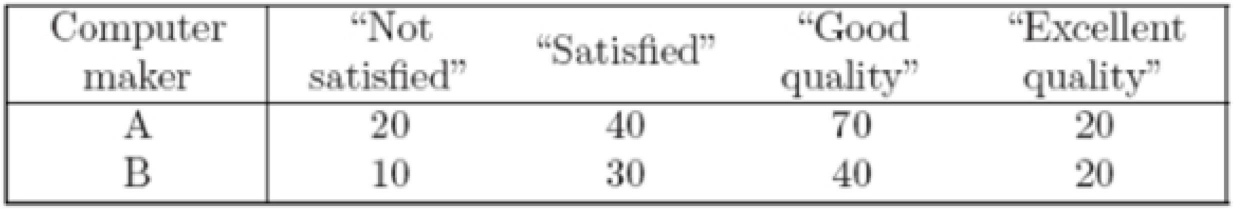
Step 1/1
Here the sample size is 18. For testing the goodness of fit of Normal Distribution, it requires at least 4 bins for making the least *df* 1 of χ^2 distribution.

$$\begin{aligned}\because df &= n - \text{no. of parameters estimated} - 1\\ &= 4-2-1 \ (\text{for 4 bins})\\ &= 1\end{aligned}$$

So if we divide the sample size of 18 into 4 bins equally, the observed frequencies will arrive to be very small (less than 5) in at least one bin for which the χ^2 test failed.

Chapter 10, Problem 6E
(0)

Problem
Two computer makers, A and B, compete for a certain market. Their users rank the quality of computers on a 4-point scale as “Not satisfied”, “Satisfied”, “Good quality”, and “Excellent quality, will recommend to others”. The following counts were observed,



Computer maker	“Not satisfied”	“Satisfied”	“Good quality”	“Excellent quality”
A	20	40	70	20
B	10	30	40	20

Is there a significant difference in customer satisfaction of the computers produced by A and by B?

Step-by-step solution

Show all steps

Step 1/3

Null hypothesis, $H_0:$ there is no significant difference between A and $B.$

Alternative hypothesis, $H_1:$ there is a significant difference between A and $B.$

Level of significance, $\alpha = 0.01$

Use the following MS Excel steps to perform the Chi-Square test.

Step-1: Enter the data into excel sheet

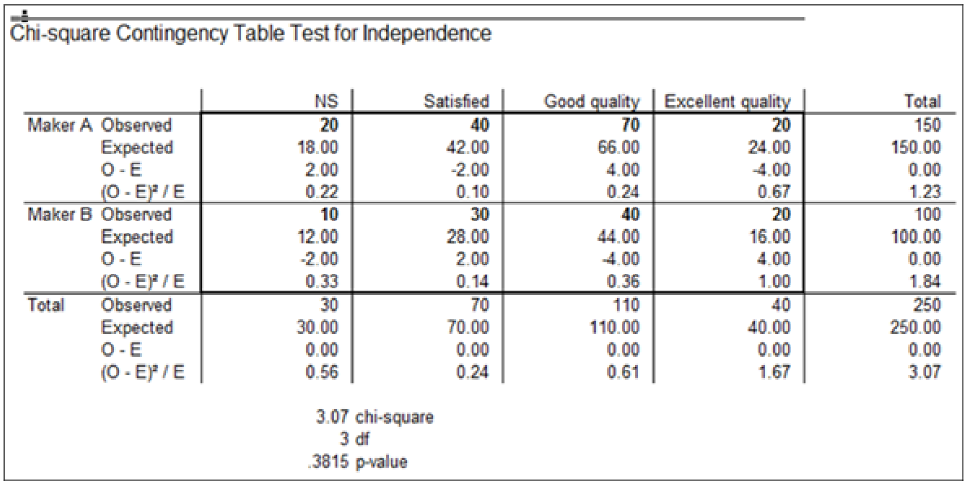
Step-2: Add Ins → Mega Stat →Chi-Square/Crosstab

Step-3: Contingency table →Enter Input range

Step-4: Click OK.

Step 2/3

The obtained output is as follows:



Chi-square Contingency Table Test for Independence

		NS	Satisfied	Good quality	Excellent quality	Total
Maker A	Observed	**20**	**40**	**70**	**20**	150
	Expected	18.00	42.00	66.00	24.00	150.00
	O - E	2.00	-2.00	4.00	-4.00	0.00
	$(O - E)^2 / E$	0.22	0.10	0.24	0.67	1.23
Maker B	Observed	**10**	**30**	**40**	**20**	100
	Expected	12.00	28.00	44.00	16.00	100.00
	O - E	-2.00	2.00	-4.00	4.00	0.00
	$(O - E)^2 / E$	0.33	0.14	0.36	1.00	1.84
Total	Observed	30	70	110	40	250
	Expected	30.00	70.00	110.00	40.00	250.00
	O - E	0.00	0.00	0.00	0.00	0.00
	$(O - E)^2 / E$	0.56	0.24	0.61	1.67	3.07

3.07 chi-square
3 df
.3815 p-value

Hence, the chi-square test statistic is, $\boxed{3.07}$.

Step 3/3

Conclusion:

Since, the *p* value is greater than the given significance level 0.01, so it is fail to reject the null hypothesis and conclude that there is no significant difference between A and $B.$

Chapter 10, Problem 7E

(0)

Problem

An AP test has been given in two schools. In the first school, 162 girls and 567 boys passed it whereas 69 girls and 378 boys failed. In the second school, 462 girls and 57 boys passed the test whereas 693 girls and 132 boys failed it.

(a) In the first school, are the results significantly different for girls and boys?

(b) In the second school, are the results significantly different for girls and boys?

(c) In both schools together, are the results significantly different for girls and boys?

For each school, construct a contingency table and apply the Chisquare test.

Remark 1. This data set is an example of a strange phenomenon known as **Simpson's paradox**. Look, the girls performed better than the boys in *each* school; however, in both schools together, the boys did better!!!

Check for yourself… In the first school, 70% of girls and only 60% of boys passed the test. In the second school, 40% of girls and only 30% of boys passed. But in both schools together, 55% of boys and only 45% of girls passed the test. Wow!

Step-by-step solution

Show all steps

Step 1/5

a)

Null hypothesis, $H_0:$ the results are not significantly different for boys and girls.

Alternative hypothesis, $H_1:$ the results are significantly different for boys and girls.

Use the following MS Excel steps to perform the Chi-Square test.

Step-1: Enter the data into excel sheet

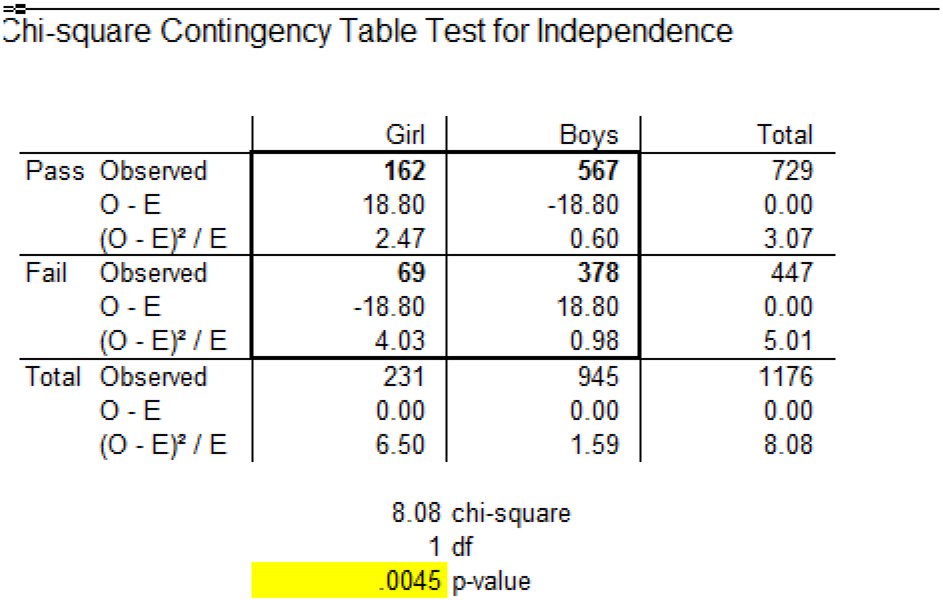
Step-2: Add INS → Mega Stat → Chi-Square/Crosstab

Step-3: Contingency table → Enter Input range

Step-4: Click OK.

Step 2/5

The obtained output is as follows:



Chi-square Contingency Table Test for Independence

		Girl	Boys	Total
Pass	Observed	**162**	**567**	729
	O - E	18.80	-18.80	0.00
	(O - E)² / E	2.47	0.60	3.07
Fail	Observed	**69**	**378**	447
	O - E	-18.80	18.80	0.00
	(O - E)² / E	4.03	0.98	5.01
Total	Observed	231	945	1176
	O - E	0.00	0.00	0.00
	(O - E)² / E	6.50	1.59	8.08

8.08 chi-square
1 df
.0045 p-value

From above output, the $p-$ value is lies between $(0.001 < p < 0.005)$, so it is rejecting null hypothesis and conclude that the results are significantly different for boys and girls.

Step 3/5
b)

Null hypothesis, $H_0:$ the results are not significantly different for boys and girls.

Alternative hypothesis, $H_1:$ the results are significantly different for boys and girls.

Use the following MS Excel steps to perform the Chi-Square test.

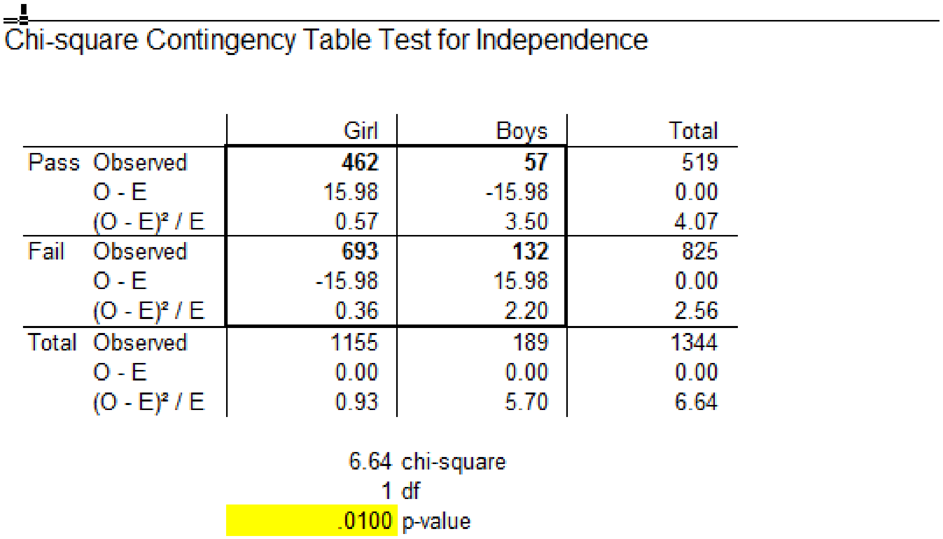
Step-1: Enter the data into excel sheet

Step-2: Add INS → Mega Stat → Chi-Square/Crosstab

Step-3: Contingency table → Enter Input range

Step-4: Click OK.

Step 4/5
The obtained output is as follows:



Chi-square Contingency Table Test for Independence

		Girl	Boys	Total
Pass	Observed	**462**	**57**	519
	O - E	15.98	-15.98	0.00
	(O - E)² / E	0.57	3.50	4.07
Fail	Observed	**693**	**132**	825
	O - E	-15.98	15.98	0.00
	(O - E)² / E	0.36	2.20	2.56
Total	Observed	1155	189	1344
	O - E	0.00	0.00	0.00
	(O - E)² / E	0.93	5.70	6.64

6.64 chi-square
1 df
.0100 p-value

From above output, the $p-$ value is lies between $(0.005 < p < 0.01)$, so it is rejecting null hypothesis and conclude that the results are significantly different for boys and girls.

c)

Null hypothesis, $H_0:$ the both schools results are not significantly different for boys and girls.

Alternative hypothesis, $H_1:$ the both schools results are significantly different for boys and girls.

Use the following MS Excel steps to perform the Chi-Square test.

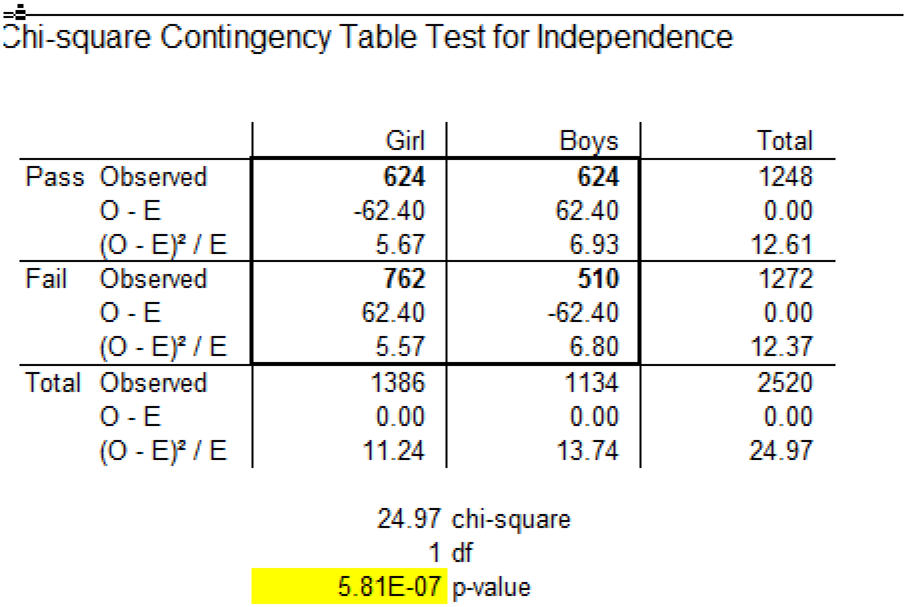
Step-1: Enter the data into excel sheet

Step-2: Add INS → Mega Stat →Chi-Square/Crosstab

Step-3: Contingency table →Enter Input range

Step-4: Click OK.

Step 5/5
The obtained output is as follows:



Chi-square Contingency Table Test for Independence

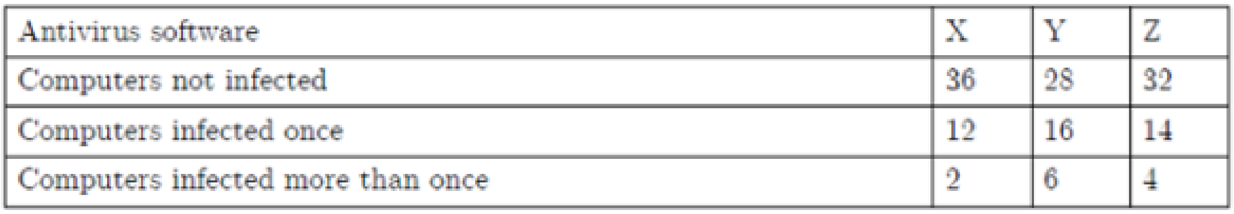
		Girl	Boys	Total
Pass	Observed	**624**	**624**	1248
	O - E	-62.40	62.40	0.00
	(O - E)² / E	5.67	6.93	12.61
Fail	Observed	**762**	**510**	1272
	O - E	62.40	-62.40	0.00
	(O - E)² / E	5.57	6.80	12.37
Total	Observed	1386	1134	2520
	O - E	0.00	0.00	0.00
	(O - E)² / E	11.24	13.74	24.97

24.97	chi-square
1	df
5.81E-07	p-value

From above output, the $p-$ value is less than the 0.01 level of significance, so it is rejecting null hypothesis conclude that the both schools results are significantly different for boys and girls.

Problem

A computer manager decides to install the new antivirus software on all the company’s computers. Three competing antivirus solutions (X, Y, and Z) are offered to her for a free 30-day trial. She installs each solution on 50 computers and records infections during the following 30 days. Results of her study are in the table.



Antivirus software	X	Y	Z
Computers not infected	36	28	32
Computers infected once	12	16	14
Computers infected more than once	2	6	4

Does the computer manager have significant evidence that the three antivirus solutions are *not* of the same quality?

Step-by-step solution

Show all steps

Step 1/3

Null hypothesis, $H_0:$ the 3 antivirus solutions are same quality.

Alternative hypothesis, $H_1:$ at least one antivirus solution is different quality.

Level of significance, $\alpha = 0.05$

Use the following MS Excel steps to perform the Chi-Square test.

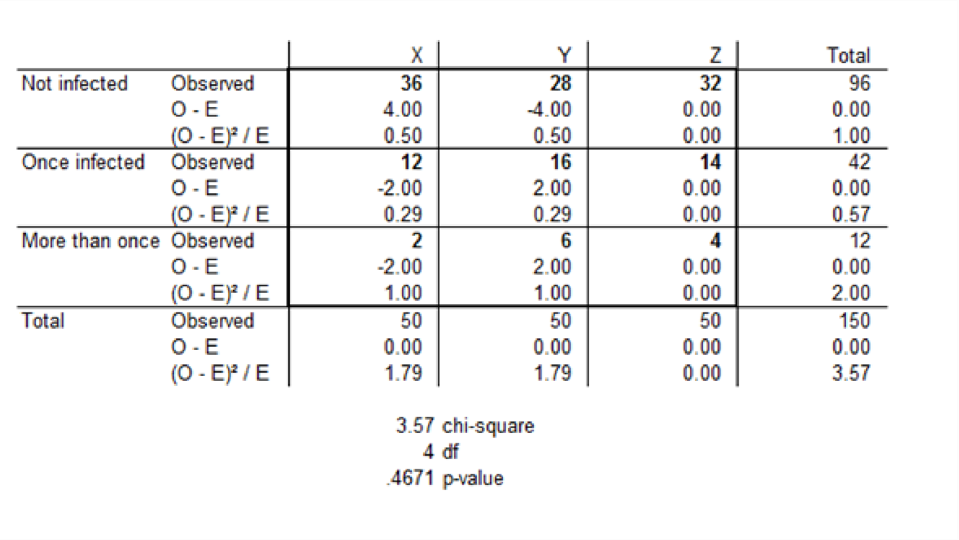
Step-1: Enter the data into excel sheet

Step-2: Add INS → Mega Stat → Chi-Square/Crosstab

Step-3: Contingency table → Enter Input range

Step-4: Click OK.

Step 2/3
The obtained output is as follows:



		X	Y	Z	Total
Not infected	Observed	**36**	**28**	**32**	96
	O - E	4.00	-4.00	0.00	0.00
	$(O - E)^2 / E$	0.50	0.50	0.00	1.00
Once infected	Observed	**12**	**16**	**14**	42
	O - E	-2.00	2.00	0.00	0.00
	$(O - E)^2 / E$	0.29	0.29	0.00	0.57
More than once	Observed	**2**	**6**	**4**	12
	O - E	-2.00	2.00	0.00	0.00
	$(O - E)^2 / E$	1.00	1.00	0.00	2.00
Total	Observed	50	50	50	150
	O - E	0.00	0.00	0.00	0.00
	$(O - E)^2 / E$	1.79	1.79	0.00	3.57

3.57 chi-square
4 df
.4671 p-value

From the above output, the chi-square test statistic is, $\boxed{3.57}$.

Step 3/3
Conclusion:

Since, the *p* value is greater than the given level of significance 0.05, so it is fail to reject the null hypothesis and conclude that the three antivirus solutions are same quality.

Chapter 10, Problem 9E
(0)

Problem

The Probability and Statistics course has three sections - S01, S02, and S03. Among 120 students in section S01, 40 got an A in the course, 50 got a B, 20 got a C, 2 got a D, and 8 got an F. Among 100 students in section S02, 20 got an A, 40 got a B, 25 got a C, 5 got a D, and 10 got an F. Finally, among 60 students in section S03, 20 got an A, 20 got a B, 15 got a C, 2 got a D, and 3 got an F. Do the three sections differ in their students' performance?

Step-by-step solution

Show all steps

Step 1/3

Null hypothesis, $H_0:$ all three sections performance is equal.

Alternative hypothesis, $H_1:$ all three sections performance is not equal.

Level of significance, $\alpha = 0.05$

Use the following MS Excel steps to perform the Chi-Square test.

Step-1: Enter the data into excel sheet

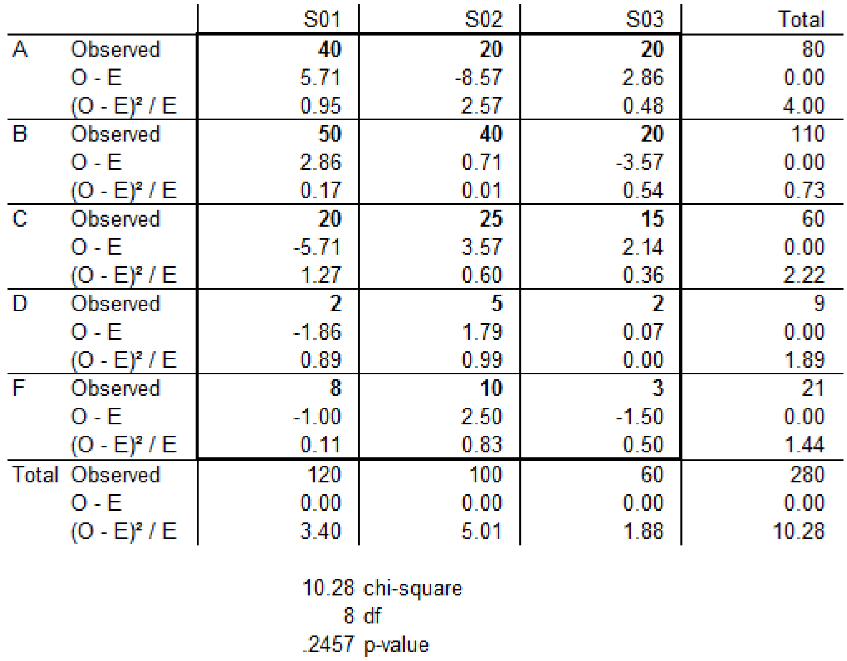
Step-2: Add INS → Mega Stat → Chi-Square/Crosstab

Step-3: Contingency table → Enter Input range

Step-4: Click OK.

Step 2/3

The obtained output is as follows:



		S01	S02	S03	Total
A	Observed	**40**	**20**	**20**	80
	O - E	5.71	-8.57	2.86	0.00
	(O - E)² / E	0.95	2.57	0.48	4.00
B	Observed	**50**	**40**	**20**	110
	O - E	2.86	0.71	-3.57	0.00
	(O - E)² / E	0.17	0.01	0.54	0.73
C	Observed	**20**	**25**	**15**	60
	O - E	-5.71	3.57	2.14	0.00
	(O - E)² / E	1.27	0.60	0.36	2.22
D	Observed	**2**	**5**	**2**	9
	O - E	-1.86	1.79	0.07	0.00
	(O - E)² / E	0.89	0.99	0.00	1.89
F	Observed	**8**	**10**	**3**	21
	O - E	-1.00	2.50	-1.50	0.00
	(O - E)² / E	0.11	0.83	0.50	1.44
Total	Observed	120	100	60	280
	O - E	0.00	0.00	0.00	0.00
	(O - E)² / E	3.40	5.01	1.88	10.28

10.28 chi-square
8 df
.2457 p-value

From the above output, the chi-square test statistic value is, $\boxed{10.28}$.

Step 3/3
Conclusion:

Since, the *p* value greater than the given significance level 0.05, so it fail to reject the null hypothesis and conclude that the three sections performance is equal.

Chapter 10, Problem 10E
(0)

Problem
Among 25 jobs sent to the printer at random times, 6 jobs were printed in less than 20 seconds each, and 19 jobs took more than 20 seconds each. Is there evidence that the median response time for this printer exceeds 20 sec? Apply the sign test.

Step-by-step solution

Show all steps

Step 1/1
534884-10-10E AID: 4342 | 22/9/2016

To perform sign test for given data. This is a one sided right tail test. Here

Null hypothesis, $H_0 : M = 20$

Alternative hypothesis, $H1 : M > 20$

The rejection of null hypothesis should imply more than 50% of printer exceeding time. The sign statistic is $S = 19$. The sample size is 25. The null distribution is approximately normal with mean 12.5 and standard deviation is 2.5.

The P value for this is 0.0073 from normal distribution table. This value is less than 0.05 which means that there is significant evidence that median speed is above 20 sec.

Problem

At a computer factory, the median of certain measurements of some computer parts should equal *m*. If it is found to be less than *m* or greater than *m*, the process is recalibrated. Every day, a quality control technician takes measurements from a sample of 20 parts. According to a 5% level sign test, how many measurements on either side of *m* justify recalibration? (In other words, what is the rejection region?)

Step-by-step solution

Show all steps

Step 1/1

534884-10-11E AID: 4342 | 22/9/2016

RID1: 6224 | 01/01/2017

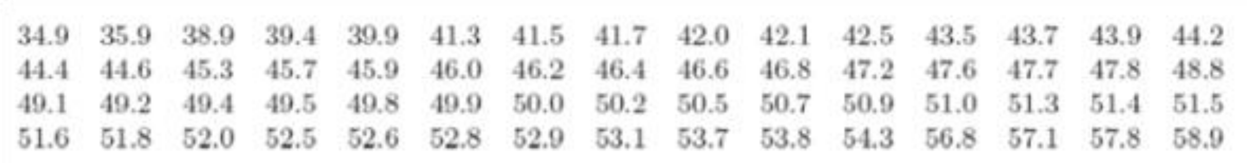
From given information we can infer that the rejection region is between (0, 5) or (15, 20) or both. This means that 15 or more measurements on either side of *m* justify recalibration.

Chapter 10, Problem 12E

(0)

Problem

When a computer chip is manufactured, its certain crucial layer should have the median thickness of 45 nm (nanometers; one nanometer is one billionth of a metre). Measurements are made on a sample of 60 produced chips (Chips), and the measured thickness is recorded as



34.9	35.9	38.9	39.4	39.9	41.3	41.5	41.7	42.0	42.1	42.5	43.5	43.7	43.9	44.2
44.4	44.6	45.3	45.7	45.9	46.0	46.2	46.4	46.6	46.8	47.2	47.6	47.7	47.8	48.8
49.1	49.2	49.4	49.5	49.8	49.9	50.0	50.2	50.5	50.7	50.9	51.0	51.3	51.4	51.5
51.6	51.8	52.0	52.5	52.6	52.8	52.9	53.1	53.7	53.8	54.3	56.8	57.1	57.8	58.9

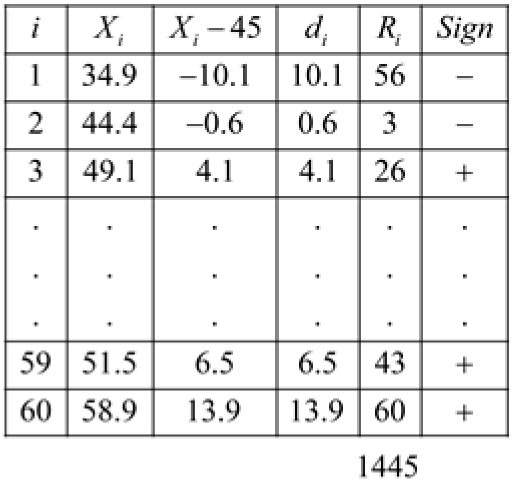
(This data set is already ordered, for your convenience.)

Will a 1% level sign test conclude that the median thickness slipped from 45 nm?

Step-by-step solution

Show all steps

Step 1/3



i	X_i	$X_i - 45$	d_i	R_i	$Sign$
1	34.9	−10.1	10.1	56	−
2	44.4	−0.6	0.6	3	−
3	49.1	4.1	4.1	26	+
. . .	. . .	. . .	. . .	. . .	. . .
59	51.5	6.5	6.5	43	+
60	58.9	13.9	13.9	60	+
				1445	

Step 2/3

Calculate the test statistic of positive ranks only is,

$$
\begin{aligned}
Z &= \frac{w_s - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \\
&= \frac{1445 - \frac{60(60+1)}{4}}{\sqrt{\frac{60(60+1)(2\times 60+1)}{24}}} \\
&= \frac{530}{135.83997} \\
&= 3.90
\end{aligned}
$$

Step 3/3

Find the $p-$ value.

$$\begin{aligned} p-\text{value} &= 2P(Z > Z_0) \\ &= 2(1-P(Z \leq 3.90)) \\ &= 2(1-0.999952) \\ &= 0.0000961 \end{aligned}$$

Since, the $p-$ value is less than the given significance level 0.05, so it is rejecting null hypothesis and conclude that the median thickness is skipping from 45 nm.

Chapter 10, Problem 13E

(0)

Problem

Refer to Exercise 10.12. It is also important to verify that the first quartile Q_1 of the layer thickness does not exceed 43 nm. Apply the idea of the sign test to the first quartile instead of the median and test

$$H_0 : Q_1 = 43 \quad \text{vs} \quad H_A : Q_1 > 43.$$

The test statistic will be the number of measurements that exceed 43 nm. Find the null distribution of this test statistic, compute the Pvalue of the test, and state the conclusion about the first quartile.

Step-by-step solution

Show all steps

Step 1/1

534884-10-13E AID: 4342 | 22/9/2016

RID1: 6224 | 01/01/2017

We have to perform sign test for given data for quartile. This is a one tail test. Here the rejection of null hypothesis should imply more than 75% of thickness exceeds 43nm.

The sign statistic is $S = 49$.

The sample size is 60.

The null distribution is (60, .75).

The *P* value for this is 0.1492 from sign table values.

This value is more than 0.05 which means that there is no significant evidence that median thickness is exceeding 43 nm.

Chapter 10, Problem 14E
(0)

Problem
The median of certain measurements on the produced computer parts should never exceed 5 inches. To verify that the process is conforming, engineers decided to measure 100 parts, one by one. To everyone's surprise, after 63 of the first 75 measurements exceeded 5 in, one engineer suggests to halt measurements and fix the process. She claims that whatever the remaining 25 measurements are, the median will be inevitably found significantly greater than 5 inches after all 100 measurements are made.

Do you agree with this statement? Why or why not? Certainly, if the remaining 25 measurements make no impact on the test then it should be a good resource-saving decision to stop testing early and fix the process.

Step-by-step solution

Show all steps

Step 1/1
534884-10-14E AID: 4342 | 22/9/2016

The given statement is correct,

Since, it is more than 50% of parts have been found to exceed the specified measurement; we should halt the process and fix the problem. Here the rejection of null hypothesis should imply more than 50% of measurement exceeds 5 inches.

Chapter 10, Problem 15E

Problem
The teacher states that the median score on the last test was 84. Masha asks 12 of her classmates and records their scores as

$$76,\ 96,\ 74,\ 88,\ 79,\ 95,\ 75,\ 82,\ 90,\ 60,\ 77,\ 56.$$

Assuming that she picks classmates at random, can she treat these data as evidence that the class median was less than 84? Can she get stronger evidence by using the sign test or the Wilcoxon signed rank test?

Step-by-step solution

Show all steps

Step 1/1
534884-10-15E AID: 4342|22/9/2016

Since the data has been chosen randomly and the values are continuous and symmetric, we can treat these data as evidence. Stronger evidence can be obtained using signed rank test as we can split the data in two groups of less than 84 and more than 84.

Problem
The starting salaries of eleven software developers are 47, 52, 68, 72, 55, 44, 58, 63, 54, 59, 77 thousand of dollars.

Does the 5% level Wilcoxon signed rank test provide significant evidence that the median starting salary of software developers is above $50,000?

Step-by-step solution

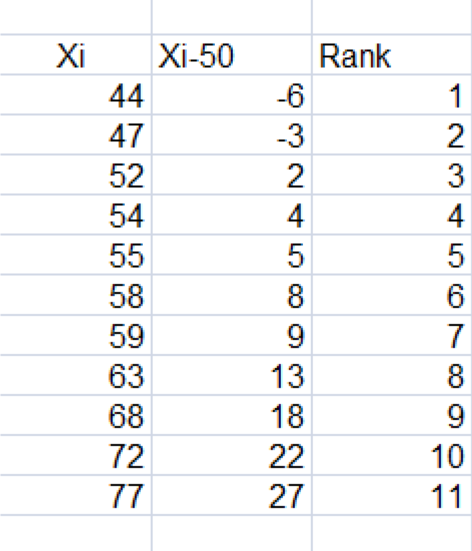
Show all steps

Step 1/1
534884-10-16E AID: 4342|22/9/2016

We have to perform Wilcoxon signed rank test for given data. This is a one tailed test. Here

$$H0 : M = 50$$
$$H1 : M > 50$$

Now we will find ranks as below:



Xi	Xi-50	Rank
44	-6	1
47	-3	2
52	2	3
54	4	4
55	5	5
58	8	6
59	9	7
63	13	8
68	18	9
72	22	10
77	27	11

Sum of positive ranks is 63. As per A8 table, for n equal to 11, we will reject null hypothesis if when sum of positive ranks is more than 53. Hence we will reject null hypothesis. We can say that these data provide significant evidence that median salary is above 50,000$.

Chapter 10, Problem 17E
(0)

Problem
Refer to Exercise 10.12. Does the Wilcoxon signed rank test confirm that the median thickness no longer equals 45 nm?

Step-by-step solution

Show all steps

Step 1/2
The data of the said problem is given as follows :

34.9	35.9	38.9	39.4	39.9	41.3	41.5	41.7
44.4	44.6	45.3	45.7	45.9	46.0	46.2	46.4
49.1	49.2	49.4	49.5	49.8	49.9	50.0	50.2
51.6	51.8	52.0	52.5	52.6	52.8	52.9	53.1
42.0	42.1	42.5	43.5	43.7	43.9	44.2	
46.6	46.8	47.2	47.6	47.7	47.8	48.8	
50.5	50.7	50.9	51.0	51.3	51.4	51.5	
53.7	53.8	54.3	56.8	57.1	57.8	58.9	

We want to test the hypothesis, $H_0 : M = 45$ vs H_a : M 45

Step 2/2

For testing the above values, let we calculate $d_i = |X_i - 45|$, ranks R_i of d_i and signs. All the figures are shown in the following table :

X_i	X_i-45	$d_i = \|X_i$	R_i	sign		X_i	X_i-45	$d_i = \|X_i$	R_i	sign
34.9	-10.1	10.1	56	-		49.1	4.1	4.1	26	+
35.9	-9.1	9.1	54	-		49.2	4.2	4.2	27	+
38.9	-6.1	6.1	40	-		49.4	4.4	4.4	28	+
39.4	-5.6	5.6	36	-		49.5	4.5	4.5	29	+
39.9	-5.1	5.1	33	-		49.8	4.8	4.8	30	+
41.3	-3.7	3.7	24	-		49.9	4.9	4.9	31	+
41.5	-3.5	3.5	23	-		50.0	5.0	5	32	+
41.7	-3.3	3.3	22	-		50.2	5.2	5.2	34	+
42.0	-3.0	3	21	-		50.5	5.5	5.5	35	+
42.1	-2.9	2.9	20	-		50.7	5.7	5.7	37	+
42.5	-2.5	2.5	16	-		50.9	5.9	5.9	38	+
43.5	-1.5	1.5	12	-		51.0	6.0	6	39	+

43.7	-1.3	1.3	10	-		51.3	6.3	6.3	41	+
43.9	-1.1	1.1	8	-		51.4	6.4	6.4	42	+
44.2	-0.8	0.8	5	-		51.5	6.5	6.5	43	+
44.4	-0.6	0.6	3	-		51.6	6.6	6.6	44	+
44.6	-0.4	0.4	2	-		51.8	6.8	6.8	45	+
45.3	0.3	0.3	1	+		52.0	7.0	7	46	+
45.7	0.7	0.7	4	+		52.5	7.5	7.5	47	+
45.9	0.9	0.9	6	+		52.6	7.6	7.6	48	+
46.0	1.0	1	7	+		52.8	7.8	7.8	49	+
46.2	1.2	1.2	9	+		52.9	7.9	7.9	50	+
46.4	1.4	1.4	11	+		53.1	8.1	8.1	51	+
46.6	1.6	1.6	13	+		53.7	8.7	8.7	52	+
46.8	1.8	1.8	14	+		53.8	8.8	8.8	53	+
47.2	2.2	2.2	15	+		54.3	9.3	9.3	55	+
47.6	2.6	2.6	17	+		56.8	11.8	11.8	57	+
47.7	2.7	2.7	18	+		57.1	12.1	12.1	58	+
47.8	2.8	2.8	19	+		57.8	12.8	12.8	59	+
48.8	3.8	3.8	25	+		58.9	13.9	13.9	60	+

The test statistic (*W*) of the Wilcoxon signed rank test is the sum of positive signed ranks and is given by,

$$\begin{aligned} W &= \sum_{X_i > m} R_i \\ &= 1445 \end{aligned}$$

The critical value of the Wilcoxon signed rank test at 5% level of significance is less than the calculated value. So the Wilcoxon test confirms that the median thickness no longer equals 45 nm.

Chapter 10, Problem 18E

(0)

Problem

Refer to Exercise 10.2. Do these data provide significant evidence that the median service time is less than 5 min 40 sec? Conduct the Wilcoxon signed rank test at the 5% level of significance. What assumption of this test may not be fully satisfied by these data?

Step-by-step solution

Show all steps

Step 1/1

The data of the said problem is given as follows :

10.5	1.2	6.3	3.7	0.9	7.1	3.3	4.0	1.7	11.6	5.1	2.8	4.8	2.0	8.0	4.6
3.1	10.2	5.9	12.6	4.5	8.8	7.2	7.5	4.3	8.0	0.2	4.4	3.5	9.6	5.5	0.3
2.7	4.9	6.8	8.6	0.8	2.2	2.1	0.5	2.3	2.9	11.7	0.6	6.9	11.4	3.8	3.2
2.6	1.9	1.0	4.1	2.4	13.6	15.2	6.4	5.3	5.4	1.4	5.0	3.9	1.8	4.7	0.7

We want to test the hypothesis, $H_0 : M = 5.67$ (5 min 40 sec)

For testing the above values, let we calculate $d_i = |X_i - 5.67|$, ranks *R*ᵢ of *d*ᵢ and signs. All the figures are shown in the following table :

X_i	X_i- 5.67	$d_i = \lvert X_i$	R_i	sign		X_i	X_i- 5.67	$d_i = \lvert X_i$	R_i	sign
0.2	-5.5	5.47	58	-		4.4	-1.3	1.27	16	-
0.3	-5.4	5.37	57	-		4.5	-1.2	1.17	14	-
0.5	-5.2	5.17	56	-		4.6	-1.1	1.07	12	-
0.6	-5.1	5.07	55	-		4.7	-1.0	0.97	11	-
0.7	-5.0	4.97	54	-		4.8	-0.9	0.87	10	-
0.8	-4.9	4.87	53	-		4.9	-0.8	0.77	9	-
0.9	-4.8	4.77	51	-		5.0	-0.7	0.67	7	-
1.0	-4.7	4.67	50	-		5.1	-0.6	0.57	5	-
1.2	-4.5	4.47	48	-		5.3	-0.4	0.37	4	-
1.4	-4.3	4.27	47	-		5.4	-0.3	0.27	3	-
1.7	-4.0	3.97	46	-		5.5	-0.2	0.17	1	-
1.8	-3.9	3.87	44	-		5.9	0.2	0.23	2	+
1.9	-3.8	3.77	43	-		6.3	0.6	0.63	6	+
2.0	-3.7	3.67	42	-		6.4	0.7	0.73	8	+
2.1	-3.6	3.57	41	-		6.8	1.1	1.13	13	+
2.2	-3.5	3.47	40	-		6.9	1.2	1.23	15	+
2.3	-3.4	3.37	39	-		7.1	1.4	1.43	18	+
2.4	-3.3	3.27	38	-		7.2	1.5	1.53	19	+
2.6	-3.1	3.07	36	-		7.5	1.8	1.83	23	+
2.7	-3.0	2.97	35	-		8.0	2.3	2.33	27.5	+
2.8	-2.9	2.87	33	-		8.0	2.3	2.33	27.5	+
2.9	-2.8	2.77	32	-		8.6	2.9	2.93	34	+
3.1	-2.6	2.57	31	-		8.8	3.1	3.13	37	+
3.2	-2.5	2.47	30	-		9.6	3.9	3.93	45	+
3.3	-2.4	2.37	29	-		10.2	4.5	4.53	49	+

3.5	-2.2	2.17	26	-		10.5	4.8	4.83	52	+
3.7	-2.0	1.97	25	-		11.4	5.7	5.73	59	+
3.8	-1.9	1.87	24	-		11.6	5.9	5.93	60	+
3.9	-1.8	1.77	22	-		11.7	6.0	6.03	61	+
4.0	-1.7	1.67	21	-		12.6	6.9	6.93	62	+
4.1	-1.6	1.57	20	-		13.6	7.9	7.93	63	+
4.3	-1.4	1.37	17	-		15.2	9.5	9.53	64	+

The test statistic (*W*) of the Wilcoxon signed rank test is the sum of negative signed ranks and is given by,

$$\begin{aligned} W &= \sum_{X_i > m} R_i \\ &= 1335 \end{aligned}$$

The critical value of the Wilcoxon signed rank test at 5% level of significance is less than the calculated value. So the Wilcoxon test confirms that the median service time is less than 5 min 40 sec.

The data of the service time generally seems to be positively skewed. Which violates the data symmetry assumption of Wilcoxon signed rank test.

Chapter 10, Problem 19E
(0)

Problem
Use the recursive formula (10.6) to calculate the null distribution of Wilcoxon test statistic *W* for sample sizes *n* = 2, *n* = 3, and *n* = 4.

Step-by-step solution

Show all steps

Step 1/4
The recursive formula of Wilcoxon test statistic *W* is given by :

$$p_n(w) = 0.5p_{n-1}(w) + 0.5p_{n-1}(w-n)$$

Where *w* are the possible values of ranks for the sample size *n* whose range is 0 to $\frac{n(n+1)}{2}$

Hence for the sample size *n* = 2, the distribution will be

$p_2(w) = 0.5p_1(w) + 0.5p_1(w-2)$, where $w = 0,1,2,3$

That gives

$$\begin{aligned} p_2(0) &= 0.5p_1(0) \\ &= 0.5\times 0.5 \\ &= 0.25 \end{aligned}$$

$$\begin{aligned} p_2(1) &= 0.5p_1(1) \\ &= 0.5\times 0.5 \\ &= 0.25 \end{aligned}$$

$$\begin{aligned} p_2(2) &= 0.5p_1(2)+0.5p_1(0) \\ &= 0.5\times 0+0.5\times 0.5 \\ &= 0.25 \end{aligned}$$

$$\begin{aligned} p_2(3) &= 0.5p_1(3)+0.5p_1(1) \\ &= 0.5\times 0+0.5\times 0.5 \\ &= 0.25 \end{aligned}$$

Step 2/4

For the sample size *n* = 3, the distribution will be

$p_3(w)=0.5p_2(w)+0.5p_2(w-3)$, where $w=0,1,2,3,4,5,6$

That gives

$$\begin{aligned} p_3(0) &= 0.5p_2(0) \\ &= 0.5\times 0.25 \\ &= 0.125 \end{aligned}$$

$$\begin{aligned} p_3(1) &= 0.5p_2(1) \\ &= 0.5\times 0.25 \\ &= 0.125 \end{aligned}$$

$$\begin{aligned} p_3(2) &= 0.5p_2(2) \\ &= 0.5\times 0.25 \\ &= 0.125 \end{aligned}$$

$$\begin{aligned} p_3(3) &= 0.5p_2(3)+0.5p_2(0) \\ &= 0.5\times 0.25+0.5\times 0.25 \\ &= 0.250 \end{aligned}$$

$$\begin{aligned} p_3(4) &= 0.5p_2(4)+0.5p_2(1) \\ &= 0.5\times 0+0.5\times 0.25 \\ &= 0.125 \end{aligned}$$

$$\begin{aligned} p_3(5) &= 0.5p_2(5) + 0.5p_2(2) \\ &= 0.5 \times 0 + 0.5 \times 0.25 \\ &= 0.125 \end{aligned}$$

$$\begin{aligned} p_3(6) &= 0.5p_2(6) + 0.5p_2(3) \\ &= 0.5 \times 0 + 0.5 \times 0.25 \\ &= 0.125 \end{aligned}$$

Step 3/4

For the sample size *n* = 4, the distribution will be

$p_3(w) = 0.5p_2(w) + 0.5p_2(w-4)$, where $w = 0,1,2,3,4,5,6,7,8,9,10$

That gives,

$$\begin{aligned} p_4(0) &= 0.5p_3(0) \\ &= 0.5 \times 0.125 \\ &= 0.0625 \end{aligned}$$

$$\begin{aligned} p_4(1) &= 0.5p_3(1) \\ &= 0.5 \times 0.125 \\ &= 0.0625 \end{aligned}$$

$$\begin{aligned} p_4(2) &= 0.5p_3(2) \\ &= 0.5 \times 0.125 \\ &= 0.0625 \end{aligned}$$

$$\begin{aligned} p_4(3) &= 0.5p_3(3) \\ &= 0.5 \times 0.250 \\ &= 0.1250 \end{aligned}$$

$$\begin{aligned} p_4(4) &= 0.5p_3(4) + 0.5p_3(0) \\ &= 0.5 \times 0.125 + 0.5 \times 0.125 \\ &= 0.1250 \end{aligned}$$

$$\begin{aligned} p_4(5) &= 0.5p_3(5) + 0.5p_3(1) \\ &= 0.5 \times 0.125 + 0.5 \times 0.125 \\ &= 0.1250 \end{aligned}$$

$$\begin{aligned} p_4(6) &= 0.5p_3(6) + 0.5p_3(2) \\ &= 0.5 \times 0.125 + 0.5 \times 0.125 \\ &= 0.1250 \end{aligned}$$

$$\begin{aligned} p_4(7) &= 0.5p_3(7)+0.5p_3(3) \\ &= 0.5\times 0+0.5\times 0.250 \\ &= 0.1250 \end{aligned}$$

$$\begin{aligned} p_4(8) &= 0.5p_3(8)+0.5p_3(4) \\ &= 0.5\times 0+0.5\times 0.125 \\ &= 0.0625 \end{aligned}$$

$$\begin{aligned} p_4(9) &= 0.5p_3(9)+0.5p_3(5) \\ &= 0.5\times 0+0.5\times 0.125 \\ &= 0.0625 \end{aligned}$$

$$\begin{aligned} p_4(10) &= 0.5p_3(10)+0.5p_3(6) \\ &= 0.5\times 0+0.5\times 0.125 \\ &= 0.0625 \end{aligned}$$

Step 4/4

The above results are summarized in the following table :

Sample size *n* =	2	3	4			
Distribution	*w*	*p*(*w*)	*w*	*p*(*w*)	*w*	*p*(*w*)
	0	0.25	0	0.125	0	0.0625
	1	0.25	1	0.125	1	0.0625
	2	0.25	2	0.125	2	0.0625
	3	0.25	3	0.250	3	0.1250
			4	0.125	4	0.1250
			5	0.125	5	0.1250
			6	0.125	6	0.1250
					7	0.1250
					8	0.0625
					9	0.0625
					10	0.0625

Chapter 10, Problem 20E

(0)

Problem

Apply the Mann–Whitney–Wilcoxon test to the quiz grades in Exercise 9.23 on p. 312 to see if Anthony's median grade is significantly higher than Eric's. What is the P-value?

Step-by-step solution

Show all steps

Step 1/2

The data of the said problem is given as follows :

Name	Quiz 1	Quiz 2	Quiz 3	Quiz 4	Quiz 5	Quiz 6
Anthony	85	92	97	65	75	96
Eric	81	79	76	84	83	77

Here no. of observations for Anthony grades, *n* = 6

No. of observations for Eric grades, *m* = 6

Let us denote Anthony's median grade by M_A and Eric's median grade by M_E . We want to test the hypothesis

$H_0 : M_A = M_E$ vs $H_A : M_A > M_E$

Step 2/2

Now combine all the values into one sample and rank them we get

Name	Grade	Rank
Anthony	65	1
Anthony	75	2
Eric	76	3
Eric	77	4
Eric	79	5
Eric	81	6
Eric	83	7
Eric	84	8
Anthony	85	9
Anthony	92	10
Anthony	96	11
Anthony	97	12

The sum of all ranks corresponding to Anthony grades is $U = 45$

Under null hypothesis, the theoretical value of *U* for *n* = = 6 and at 5% level of significance is $U_{H_0} = 50$

The critical value of the Mann-Whitney Wilcoxon two sample rank-sum test at 5% level of significance is more than the calculated value. So there is no sufficient evidence in favour that the Anthony’s median grade is greater than the Eric’s grade. The actual *p*-value of this comparison is 0.200

Chapter 10, Problem 21E
(0)

Problem
Two internet service providers claim that they offer the fastest internet in the area. A local company requires the download speed of at least 20 Megabytes per second (Mbps), for its normal operation. It decides to conduct a fair contest by sending 10 packets of equal size through each network and recording their download speed.

For the 1st internet service provider, the download speed is recorded as 26.7, 19.0, 26.5, 29.1, 26.2, 27.6, 26.8, 24.2, 25.7, 23.0 Mbps. For the 2nd provider, the download speed is recorded as 19.3, 22.1, 23.4, 24.8, 25.9, 22.2, 18.3, 20.1, 19.2, 27.9 Mbps.

(a) According to the sign test, is there significant evidence that the median download speed for the 1st provider is at least 20 Mbps? What about the 2nd provider? Calculate each Pvalue.

(b) Repeat (a) using the Wilcoxon signed rank test. Do the Pvalues show that this test is more sensitive than the sign test?

(c) At the 1% level, is there significant evidence that the median download speed for the 1st provider exceeds the median download speed for the 2nd provider? Use the suitable test.

These data are also available in Internet.

Step-by-step solution

Show all steps

Step 1/5
The data of the said problem is given as follows :

1st Service Provider	26.7	19.0	26.5	29.1	26.2	27.6	26.8	24.2	25.7	23.0
2nd Service Provider	19.3	22.1	23.4	24.8	25.9	22.2	18.3	20.1	19.2	27.9

Step 2/5
(a)

To test whether the 1st service provider median downloaded speed is at least 20 mbps, the hypothesis formed is

$H_0 : M_1 = 20$ vs $H_A : M_1 > 20$

Where M_1 is the median downloaded speed of the first service provider.

To conduct the sign test we count the no. of observations above 20, that comes out to be

$S_{obs} = 9$

The corresponding P-value is given by,

$$\begin{aligned} P\left(S \ge S_{obs}\right) &= P\left(S \ge 9\right) \\ &= 0.011 \text{ (using binomial law)} \end{aligned}$$

So there is sufficient evidence in favour that 1st service provider median downloaded speed is at least 20 mbps at 5% level of significance.

To test whether the 2nd service provider median downloaded speed is at least 20 mbps, the hypothesis formed is

$H_0 : M_2 = 20$ vs $H_A : M_2 > 20$

Where M_2 is the median downloaded speed of the first service provider.

To conduct the sign test we count the no. of observations above 20, that comes out to be

$S_{obs} = 7$

The corresponding P-value is given by,

$$\begin{aligned} P\left(S \ge S_{obs}\right) &= P\left(S \ge 7\right) \\ &= 0.172 \text{ (using binomial law)} \end{aligned}$$

So there is not sufficient evidence in favour that 2nd service provider median downloaded speed is at least 20 mbps at 5% level of significance.

Step 3/5
(b)

Now we will test the same hypothesis using Wilcoxon signed rank test.

For the 1st service provider let we calculate $d_i = |X_i - 20|$, ranks R_i of d_i and signs. All the figures are shown in the following table :

Xi	Xi-20	di	Ri	sign
19.0	-1.0	1	1	-
23.0	3.0	3	2	+
24.2	4.2	4.2	3	+
25.7	5.7	5.7	4	+
26.2	6.2	6.2	5	+
26.5	6.5	6.5	6	+
26.7	6.7	6.7	7	+
26.8	6.8	6.8	8	+
27.6	7.6	7.6	9	+
29.1	9.1	9.1	10	+

The test statistic (*W*) of the Wilcoxon signed rank test is the sum of positive signed ranks and is given by,

$$\begin{aligned} W &= \sum_{X_i > m} R_i \\ &= 54 \end{aligned}$$

While the critical value at 5% level of significance is 45.

The critical value of the Wilcoxon signed rank test at 5% level of significance is less than the calculated value. So the Wilcoxon test confirms that the 1st service provider median downloaded speed is at least 20 mbps at 5% level of significance.

Step 4/5

For the 2nd service provider let we calculate $d_i = |X_i - 20|$, ranks R_i of d_i and signs. All the figures are shown in the following table :

Xi	Xi-20	di	Ri	sign
18.3	-1.7	1.7	1	-
19.2	-0.8	0.8	2	-
19.3	-0.7	0.7	3	-
20.1	0.1	0.1	4	+
22.1	2.1	2.1	5	+
22.2	2.2	2.2	6	+
23.4	3.4	3.4	7	+
24.8	4.8	4.8	8	+
25.9	5.9	5.9	9	+
27.9	7.9	7.9	10	+

The test statistic (*W*) of the Wilcoxon signed rank test is the sum of positive signed ranks and is given by,

$$W = \sum_{X_i > m} R_i$$
$$= 49$$

While the critical value at 5% level of significance is 45.

The critical value of the Wilcoxon signed rank test at 5% level of significance is less than the calculated value. So the Wilcoxon test confirms that the 2nd service provider median downloaded speed is at least 20 mbps at 5% level of significance.

The Wilcoxon test appears to be more sensitive.

Step 5/5
(c)

To test the hypothesis whether the median download speed of 1st provider is more than the 2nd provider, the suitable test is Mann-Whitney Wilcoxon two sample rank-sum test.

Now combine all the values into one sample and rank them we get

Provider	Xi	Rank
1st Provider	19.0	2
1st Provider	23.0	8
1st Provider	24.2	10
1st Provider	25.7	12
1st Provider	26.2	14
1st Provider	26.5	15
1st Provider	26.7	16
1st Provider	26.8	17
1st Provider	27.6	18
1st Provider	29.1	20
2nd Provider	18.3	1
2nd Provider	19.2	3
2nd Provider	19.3	4
2nd Provider	20.1	5
2nd Provider	22.1	6
2nd Provider	22.2	7
2nd Provider	23.4	9
2nd Provider	24.8	11
2nd Provider	25.9	13
2nd Provider	27.9	19

The sum of all ranks corresponding to 1st provider is $U = 132$

Under null hypothesis, the theoretical value of *U* for *n* = = 10 and at 1% level of significance is $U_{H_0} = 136$

The critical value of the Mann-Whitney Wilcoxon two sample rank-sum test at 1% level of significance is not less than the calculated value. So there is no sufficient evidence in favour that the median speed of 1st service provider exceeds the 2nd provider.

Chapter 10, Problem 22E
(0)

Problem
Fifteen e-mail attachments were classified as benign and malicious. Seven benign attachments were 0.4, 2.1, 3.6, 0.6, 0.8, 2.4, and 4.0 Mbytes in size. Eight malicious attachments had sizes 1.2, 0.2, 0.3, 3.3, 2.0, 0.9, 1.1, and 1.5 Mbytes. Does the Mann–Whitney– Wilcoxon test detect a significant difference in the distribution of sizes of benign and malicious attachments? (If so, the size could help classify e-mail attachments and warn about possible malicious codes.)

Step-by-step solution

Show all steps

Step 1/2
The data about the file sizes of benign and malicious attachments is given as follows :

benign	0.4	2.1	3.6	0.6	0.8	2.4	4.0	
malicious	1.2	0.2	0.3	3.3	2.0	0.9	1.1	1.5

Let us denote benign median size by M_b and malicious median size by M_m . We want to test the hypothesis

$H_0 : M_b = M_m$ vs $H_A : M_b \neq M_m$

Step 2/2
We will apply Mann-Whitney Wilcoxon two sample rank-sum test.

Combine all the values into one sample and rank them we get the table

Type	Size	Rank
benign	0.4	3
benign	0.6	4
benign	0.8	5
benign	2.1	11
benign	2.4	12
benign	3.6	14
benign	4.0	15
malicious	0.2	1

malicious	0.3	2
malicious	0.9	6
malicious	1.1	7
malicious	1.2	8
malicious	1.5	9
malicious	2.0	10
malicious	3.3	13

The sum of all ranks corresponding to benign type $U = 64$

The P-value for the two sided test is

$$\begin{aligned} P &= 2\min\left(P(U \le 64), P(U \ge 64)\right) \\ &= 2 \times 0.200 \\ &= 0.400 \end{aligned}$$

There is no significant difference between the distributions of sizes of benign and malicious.

Chapter 10, Problem 23E
(0)

Problem
During freshman year, Eric’s textbooks cost $89, $99, $119, $139, $189, $199, and $229. During his senior year, he had to pay $109, $159, $179, $209, $219, $259, $279, $299, and $309 for his textbooks. Is this significant evidence that the median cost of textbooks is rising, according to the Mann–Whitney–Wilcoxon test?

Step-by-step solution

Show all steps

Step 1/2
The data of costs of Eric’s textbooks during freshman year and senior year is given as follows :

freshman year cost (in $)	89.0	99.0	119.0	139.0	189.0	199.0	229.0		
senior year cost (in $)	109.0	159.0	179.0	209.0	219.0	259.0	279.0	299.0	309.0

Let us denote freshman year median cost by M_f and senior year median cost by M_s. We want to test the hypothesis

$H_0 : M_f = M_s$ vs $H_A : M_f < M_s$

Step 2/2

We will apply Mann-Whitney Wilcoxon two sample rank-sum test.

Combine all the values into one sample and rank them we get the table

Year	cost	Rank
freshman year	89.0	1
freshman year	99.0	2
freshman year	119.0	4
freshman year	139.0	5
freshman year	189.0	8
freshman year	199.0	9
freshman year	229.0	12
senior year	109.0	3
senior year	159.0	6
senior year	179.0	7
senior year	209.0	10
senior year	219.0	11
senior year	259.0	13
senior year	279.0	14
senior year	299.0	15
senior year	309.0	16

The sum of all ranks corresponding to freshman year $U = 41$

Under null hypothesis, the theoretical value of *U* for *n* = = 9 and at 5% level of significance is $U_{H_0} = 43$ which is greater than the calculated value.

There is significant evidence in favour that the median cost of senior year exceeds the median cost of freshman year.

Chapter 10, Problem 24E

(0)

Problem

Two teams, six students each, competed at a programming contest. The judges gave the overall 1st, 3rd, 6th, 7th, 9th, and 10th places to members of Team A. Now the captain of Team A claims the overall victory over Team B, according to a one-sided Mann– Whitney–Wilcoxon test? Do you concur with his conclusion? What hypothesis are you testing?

Step-by-step solution

Show all steps

Step 1/1
Here we will test the hypothesis

$H_0 : M_A = M_B$ vs $H_A : M_A < M_B$

Where M_A & M_B are the median positions of team A and team B respectively.

The sum of all ranks (positions) corresponding to team A, $U = 36$

Under null hypothesis, the theoretical value of U for n = = 6 and at 5% level of significance is $U_{H_0} = 28$ which is less than the calculated value.

There is no sufficient evidence in favour that the team A have the overall victory over team B.

Chapter 10, Problem 25E
(0)

Problem
Refer to Exercise 10.2 and ServiceTimes. After the first 32 service times were recorded (the first two rows of data), the server was substantially modified. Conduct the Mann–Whitney–Wilcoxon test at the 10% level to see if this modification led to a reduction of the median service time.

Step-by-step solution

Show all steps

Step 1/4
The data on service time before and after modification of the server is given as follows :

Before Modification		After Modification		
10.5	3.1		2.7	2.6
1.2	10.2		4.9	1.9
6.3	5.9		6.8	1.0
3.7	12.6		8.6	4.1
0.9	4.5		0.8	2.4
7.1	8.8		2.2	13.6
3.3	7.2		2.1	15.2
4.0	7.5		0.5	6.4
1.7	4.3		2.3	5.3
11.6	8.0		2.9	5.4
5.1	0.2		11.7	1.4

2.8	4.4		0.6	5.0
4.8	3.5		6.9	3.9
2.0	9.6		11.4	1.8
8.0	5.5		3.8	4.7
4.6	0.3		3.2	0.7

Here we want to test the hypothesis

$H_0 : M_B = M_A$ vs $H_A : M_B < M_A$

Where M_B & M_A are the median service time of server before and after modification respectively.

Step 2/4

We will apply Mann-Whitney Wilcoxon two sample rank-sum test.

Combine all the values into one sample and rank them we get the table :

Status	time	Rank		Status	time	Rank
After Modification	0.5	3		Before Modification	0.2	1
After Modification	0.6	4		Before Modification	0.3	2
After Modification	0.7	5		Before Modification	0.9	7
After Modification	0.8	6		Before Modification	1.2	9
After Modification	1.0	8		Before Modification	1.7	11
After Modification	1.4	10		Before Modification	2.0	14
After Modification	1.8	12		Before Modification	2.8	21
After Modification	1.9	13		Before Modification	3.1	23
After Modification	2.1	15		Before Modification	3.3	25

After Modification	2.2	16		Before Modification	3.5	26
After Modification	2.3	17		Before Modification	3.7	27
After Modification	2.4	18		Before Modification	4.0	30
After Modification	2.6	19		Before Modification	4.3	32
After Modification	2.7	20		Before Modification	4.4	33
After Modification	2.9	22		Before Modification	4.5	34
After Modification	3.2	24		Before Modification	4.6	35
After Modification	3.8	28		Before Modification	4.8	37
After Modification	3.9	29		Before Modification	5.1	40
After Modification	4.1	31		Before Modification	5.5	43
After Modification	4.7	36		Before Modification	5.9	44
After Modification	4.9	38		Before Modification	6.3	45
After Modification	5.0	39		Before Modification	7.1	49
After Modification	5.3	41		Before Modification	7.2	50
After Modification	5.4	42		Before Modification	7.5	51

After Modification	6.4	46		Before Modification	8.0	52.5
After Modification	6.8	47		Before Modification	8.0	52.5
After Modification	6.9	48		Before Modification	8.8	55
After Modification	8.6	54		Before Modification	9.6	56
After Modification	11.4	59		Before Modification	10.2	57
After Modification	11.7	61		Before Modification	10.5	58
After Modification	13.6	63		Before Modification	11.6	60
After Modification	15.2	64		Before Modification	12.6	62

The sum of all ranks corresponding to Before Modification $U = 1142$

Under null hypothesis, for large sample size the Wilcoxon statistic U follows the normal distribution with

mean

$$\mu = \frac{n(n+m+1)}{2}$$
$$= 1040$$

And SD

$$\sigma = \sqrt{\frac{nm(n+m+1)}{12}}$$
$$= 74.48$$

Step 3/4

So the P-value for testing the above hypothesis is given by

$$\begin{aligned} P &= P(U \geq 1142) \\ &= P\left(z \geq \frac{1142-1040}{74.48}\right) \\ &= P(z \geq 1.37) \\ &= 0.085 \end{aligned}$$

Step 4/4
The *P*-value comes out to be less that the significance level (10% or 0.10)

We can say that at 10% level of significance, there is significant reduction in service time after the modification in server.

Chapter 10, Problem 26E
(0)

Problem
On five days of the week, Natasha spends 2, 2, 3, 3, and 5 hours doing her homework.

(a) List all possible bootstrap samples and find the probability of each of them.

(b) Use your list to find the bootstrap distribution of the sample median.

(c) Use this bootstrap distribution to estimate the standard error and the bias of a sample median.

Step-by-step solution

Show all steps

Step 1/3
(a)

Here the given data set is {2, 2, 3, 3, 5}.

Some of the bootstrap samples of the above data set are

Sample	Median				
2	2	2	2	2	2
2	2	2	2	3	2
2	2	2	3	3	2
2	2	3	3	3	3
2	3	3	3	3	3
3	3	3	3	3	3
2	2	2	2	5	2
2	2	2	5	5	2

2	2	5	5	5	5
2	5	5	5	5	5
5	5	5	5	5	5
2	2	2	3	5	2
2	2	3	3	5	3
2	3	3	3	5	3
3	3	3	3	5	3
3	3	3	5	5	3
3	3	5	5	5	5
3	5	5	5	5	5
2	2	3	5	5	3
2	3	3	5	5	3
2	3	5	5	5	5

And so on.

Step 2/3

(b)

We will find the bootstrap distribution of sample median. Based on the given sample of size *n* = 5, the sample median can be equal to 2, 3 or 5. Let us compute the probability of each value.

Values 2 & 3 appear with probability 2/5 and 5 appear with probability 1/5.

The sample median in any bootstrap sample (B_i) is the central or the 3rd smallest observation. Thus it equals 2 if at least 3 of 5 values in (B_i) equal 2. The probability of that is

$$\begin{aligned} P(2) &= P(Y \geq 3) \\ &= \sum_{y=3}^{5} \binom{5}{y} \left(\frac{2}{5}\right)^{y} \left(\frac{3}{5}\right)^{5-y} \\ &= \frac{992}{5^5} \\ &= 0.3174 \end{aligned}$$

It equals 3 if at least 3 of 5 values in (B_i) equal 2 or 3. The probability of that is

$$\begin{aligned} P(3) &= P(Y \geq 3) \\ &= \sum_{y=3}^{5} \binom{5}{y} \left(\frac{4}{5}\right)^{y} \left(\frac{1}{5}\right)^{5-y} \\ &= \frac{2944}{5^5} \\ &= 0.9421 \end{aligned}$$

It equals 5 if at least 3 of 5 values in (B_i) equal 2, 3 or 5. The probability of that is

$$P(5)=1$$

From this cdf, the bootstrap probability mass function is found to be

M	2	3	5
p(*M*)	0.3174	0.62464	0.058

Step 3/3

(c)

The expected value of the median is calculated as follows

M	2	3	5	E(M)
p(*M*)	0.317	0.625	0.058	
Mp(*M*)	0.635	1.874	0.290	2.799

The actual median is 3. So the bias is

$$\begin{aligned} bias(M) &= E(M)-M \\ &= 2.799-3 \\ &= -0.201 \end{aligned}$$

The standard error is calculated as follows :

M	2	3	5	Total
p(*M*)	0.317	0.625	0.058	
M×*p*(*M*)	0.635	1.874	0.290	2.799
M ^{2}×*p*(*M*)	1.270	5.622	1.449	8.340

Now

$$\begin{aligned} Var(M) &= E(M^2)-\{E(M)\}^2 \\ &= 8.340-2.799^2 \\ &= 0.509 \end{aligned}$$

So the standard error is found to be

$$\begin{aligned} & SE(M) \\ &= \sqrt{\frac{n}{n-1}Var(M)} \\ &= \boxed{0.797} \end{aligned}$$

Chapter 10, Problem 27E
(0)

Problem
In Exercise 10.16, we tested the median starting salary of software developers. We can actually estimate this starting salary by the sample median, which for these data equals $\hat{M} = \$58,000$.

(a) How many different bootstrap samples can be generated? Find the number of all possible ordered and unordered samples.

(b) Find the bootstrap distribution of the sample median. Do not list all the bootstrap samples!

(c) Use this distribution to estimate the standard error of $\hat{M}$.

(d) Construct an 88% bootstrap confidence interval for the population median salary *M*.

(e) Use the bootstrap distribution to estimate the probability that 11 randomly selected software developers have their median starting salary above $50,000.

Exercises 10.28–10.30 require the use of a computer.

Step-by-step solution

Show all steps

Step 1/5
(a)

There are 11 software developers whose starting salaries (in thousands of dollars) in increasing order are given as below :

44	47	52	54	55	58	59	63	68	72	77

The number of possible unordered samples = $11^{11} = \boxed{285311670611}$

The number of possible ordered samples can be obtained by computing all the arrangements of smaller no. inclusion in the above order number chosen and then removing from the above figure of unordered samples

This figure is calculated to be 352716.

Step 2/5
(b)

We will find the bootstrap distribution of sample median. Based on the given sample of size *n* = 11, the sample median can be equal to 44, 47, 52 etc. Let us compute the probability of each value.

All values appear with probability 1/11.

The sample median in any bootstrap sample (B_i) is the central or the 6th smallest observation. Thus it equals 44 if at least 6 of 11 values in (B_i) equal 44. The probability of that is

$$\begin{aligned}P(44) &= P(Y \geq 6)\\ &= \sum_{y=6}^{11} \binom{11}{y} \left(\frac{1}{11}\right)^{y} \left(\frac{10}{11}\right)^{11-y}\\ &= 0.0002\end{aligned}$$

It equals 47 if at least 6 of 11 values in (B_i) equal 44 or 47. The probability of that is

$$\begin{aligned}P(47) &= P(Y \geq 6)\\ &= \sum_{y=6}^{11} \binom{11}{y} \left(\frac{2}{11}\right)^{y} \left(\frac{8}{11}\right)^{11-y}\\ &= 0.0072\end{aligned}$$

Proceeding in this way we get the cdf of the sample median. Which is given in the following table :

Median	Cumulative Probability
44	0.0002
47	0.0072
52	0.0512
54	0.1727
55	0.3786
58	0.6214
59	0.8273
63	0.9488
68	0.9928
72	0.9998
77	1.0000

Using the above table we obtain the probability distribution of the sample median by subtracting the preceding cumulative probability from the corresponding cumulative probability of the median value. In this way we get the following probability distribution of the median.

Median	Probability
44	0.0002
47	0.0070
52	0.0440
54	0.1215
55	0.2059
58	0.2427

59	0.2059
63	0.1215
68	0.0440
72	0.0070
77	0.0002

Step 3/5

(c)

The standard error is calculated as follows :

M	44	47	52	54	55	58	59	63	68	72	77	Total
p(M)	0.0002	0.0070	0.0440	0.1215	0.2059	0.2427	0.2059	0.1215	0.0440	0.0070	0.0002	
Mp(M)	0.008	0.331	2.290	6.561	11.324	14.079	12.147	7.655	2.995	0.506	0.013	57.908
M2p(M)	0.34	15.54	119.08	354.29	622.80	816.57	716.68	482.23	203.64	36.46	1.03	3368.668

Now

$$\begin{aligned} Var(M) &= E(M^2)-\{E(M)\}^2 \\ &= 3368.7-57.9^2 \\ &= 15.35 \end{aligned}$$

So the standard error was found to be

$$SE(M)$$
$$=\sqrt{\frac{n}{n-1}Var(M)}$$
$$=\boxed{4.12}$$

Step 4/5

(d)

An 88% bootstrap confidence interval for the population median salary is

$$\left(\hat{M}-z_{12/2}s, \hat{M}+z_{12/2}s\right)$$

Here $\hat{M}=57.91$

So,

$$\left(\hat{M}-z_{.12/2}s,\hat{M}+z_{.12/2}s\right)$$
$$=\left(57.91-z_{.06}\times 4.12,57.91+z_{.06}\times 4.12\right)$$
$$=\left(51.5,64.3\right)$$

Step 5/5
(e)

The required probability will be obtained by :

$$\begin{aligned}P\left(M>50\right)&=1-P\left(M=44\right)-P\left(M=47\right)\\&=1-0.0072\\&=\boxed{0.993}\end{aligned}$$

Chapter 10, Problem 31E
(0)

Problem
A new section of a highway is opened, and *X* = 4 accidents occurred there during one month. The number of accidents has Poisson(*θ*) distribution, where *θ* is the expected number of accidents during one month. Experience from the other sections of this highway suggests that the prior distribution of *θ* is Gamma(5,1). Find the Bayes estimator of *θ* under the squared-error loss and find its posterior risk.

Step-by-step solution

Show all steps

Step 1/2

We know that if the prior distribution of θ is $Gamma\left(\alpha,\lambda\right)$, and the model $f\left(x/\theta\right)$ follows a Poisson distribution $P\left(\theta\right)$. Then the posterior distribution of θ follows a $Gamma\left(\alpha+n\bar{X},\lambda+n\right)$

In this problem the prior distribution of θ is $Gamma\left(5,1\right)$ and $\bar{X}=4$ with Poisson pmf with *n* = 1. So the posterior distribution of θ follows a $Gamma\left(5+1\times 4,1+1\right)$

That is θ follows a $Gamma\left(9,2\right)$

Step 2/2
So the Bayes estimator of θ,

$$\begin{aligned}\hat{\theta} &= E(\theta \mid x)\\ &= \frac{\alpha + n\overline{X}}{\lambda + n}\\ &= \frac{9}{2}\\ &= \boxed{4.5}\end{aligned}$$

and the posterior risk or posterior variance is equal to

$$\begin{aligned}p(\hat{\theta}) &= Var(\theta \mid x)\\ &= \frac{\alpha + n\overline{X}}{(\lambda + n)^2}\\ &= \frac{9}{4}\\ &= \boxed{2.25}\end{aligned}$$

Chapter 10, Problem 32E
(0)

Problem
The data set consists of a sample ***X*** = (2, 3, 5, 8, 2) from the Geometric distribution with unknown parameter *θ*.

(a) Show that the Beta family of prior distributions is conjugate.

(b) Taking the Beta(3,3) distribution as the prior, compute the Bayes estimator of *θ*.

Step-by-step solution

Show all steps

Step 1/2
(a)

Let the prior distribution of θ is $Beta(\alpha, \beta)$,

And *f*(*x*/) *Geometric()*

Given that sample $\vec{x} = (x_1, x_2, \ldots, x_n)$*, the likelihood function for* θ *for making its posterior distribution is*

$$
\begin{aligned}
f\left(\bar{x}/\theta\right) &= \prod_{i=1}^{n} \theta\left(1-\theta\right)^{x_i-1} \\
&= \theta^{n}\left(1-\theta\right)^{\sum_{i=1}^{n} x_i - n} \\
&= \theta^{n}\left(1-\theta\right)^{n\bar{x}-n}
\end{aligned}
$$

and the prior density function of θ *is*

$$\pi\left(\theta\right) = \theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}$$

Therefore the posterior distribution of θ*,*

$$
\begin{aligned}
&\pi\left(\theta/\bar{x}\right) = f\left(\bar{x}/\theta\right)\pi\left(\theta\right) \\
&= \theta^{n}\left(1-\theta\right)^{n\bar{x}-n}\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1} \\
&= \theta^{\alpha+n-1}\left(1-\theta\right)^{\beta+n\bar{x}-n-1}
\end{aligned}
$$

We see that the posterior distribution of θ *is again Beta distribution* $Beta\left(\alpha+n, \beta+n\bar{x}-n\right)$*. This implies that the Beta family of prior distributions is conjugate.*

Step 2/2

(b)

The given sample of size *n* = 5 has values {2, 3, 5, 8, 2}

So, $\bar{x} = 4$

The prior distribution of θ is $Beta\left(3,3\right)$

Therefore the posterior distribution of θ

$$
\begin{aligned}
&= Beta\left(3+5, 3+5\times 4-5\right) \\
&= Beta\left(8,18\right)
\end{aligned}
$$

Now the Bayes estimate of θ is,

$$
\begin{aligned}
\hat{\theta} &= E\left(\theta \mid x\right) \\
&= \frac{\left(\alpha+n\right)}{\left(\alpha+n\right)+\left(\beta+n\bar{x}-n\right)} \\
&= \frac{8}{8+18} \\
&= \boxed{0.31}
\end{aligned}
$$

Chapter 10, Problem 33E
(0)

Step-by-step solution

Show all steps

Step 1/4
(a)

Let the prior distribution of θ is $Gamma(\alpha, \lambda)$,

And $f(x/\theta) \sim Exponential(\theta)$

Given that sample $\vec{x} = (x_1, x_2, ..., x_n)$, the likelihood function for θ for making its posterior distribution is

$$
\begin{aligned}
f(\vec{x}/\theta) &= \prod_{i=1}^{n} \theta e^{-\theta x_i} \\
&= \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \\
&= \theta^n e^{-n\theta \bar{x}}
\end{aligned}
$$

and the prior density function of θ is

$$\pi(\theta) = \theta^{\alpha-1} e^{-\lambda\theta}$$

Therefore the posterior distribution of θ,

$$
\begin{aligned}
&\pi(\theta/\vec{x}) = f(\vec{x}/\theta)\pi(\theta) \\
&= \theta^n e^{-\theta n\bar{x}} \theta^{\alpha-1} e^{-\lambda\theta} \\
&= \theta^{\alpha+n-1} e^{-(n\bar{x}+\lambda)\theta}
\end{aligned}
$$

We see that the posterior distribution of θ is again Gamma distribution $Gamma(\alpha + n, \lambda + n\bar{x})$. This implies that the Gamma family of prior distributions is conjugate.

Step 2/4
(b)

The given sample of size *n* = 5 has values {4, 3, 2, 5, 5}

So, $\bar{x} = 3.8$

The prior distribution of θ is $Gamma(3,1)$

Therefore the posterior distribution of θ

$$\begin{aligned}&= Gamma\left(3+5, 1+5\times 3.8\right)\\ &= Gamma\left(8, 20\right)\end{aligned}$$

Now the Bayes estimate of θ is,

$$\begin{aligned}\hat{\theta} &= E\left(\theta \mid x\right)\\ &= \frac{\left(\alpha+n\right)}{\left(\lambda+n\bar{x}\right)}\\ &= \frac{8}{20}\\ &= \boxed{0.40}\end{aligned}$$

And the posterior risk is given by

$$\begin{aligned}p\left(\hat{\theta}\right) &= Var\left(\theta \mid x\right)\\ &= \frac{\alpha+n}{\left(\lambda+n\bar{x}\right)^2}\\ &= \frac{8}{20^2}\\ &= \boxed{0.02}\end{aligned}$$

Step 3/4
(c)

The posterior probability that $\theta \geq 0.5$ is calculated as follows :

$$\begin{aligned}P\left(\theta \geq 0.5\right) &= \int_{0.5}^{\infty} \theta^{8-1} e^{-20\theta} d\theta\\ &= 0.22 \quad \left(\text{found using Excel Gammadist() function}\right)\end{aligned}$$

Step 4/4
(d)

To test the hypothesis $H_0 : \theta \geq 0.5$ vs $H_A : \theta < 0.5$, we calculate

$$\begin{aligned}&\rho\left(\text{reject } H_0\right)\\ &= \omega P\left(\theta < 0.5 / \vec{x}\right)\\ &= \omega\left(1-0.22\right)\\ &=0.88\omega\end{aligned}$$

where ω is the loss of error under accepting or rejecting H_0 (which is same in this problem).

and

$$\rho(\text{accept } H_0)$$
$$=\omega P(\theta \geq 0.5/\vec{x})$$
$$=0.22\omega$$

So the decision is : Reject H_0.

Chapter 10, Problem 34E
(0)

Problem
An internet service provider studies the distribution of the number of concurrent users of the network. This number has Normal distribution with mean *θ* and standard deviation 4,000 people. The prior distribution of *θ* is Normal with mean 14,000 and standard deviation 2,000.

The data on the number of concurrent users are collected; see ConcurrentUsers and Exercise 8.2 on p. 240.

(a) Give the Bayes estimator for the mean number of concurrent users *θ*.

(b) Construct the highest posterior density 90% credible set for *θ* and interpret it.

(c) Is there significant evidence that the mean number of concurrent users exceeds 16,000?

Step-by-step solution

Show all steps

Step 1/4
The data on the number of concurrent users (in thousands) is shown below :

17.2	22.1	18.5	17.2	18.6	14.8	21.7	15.8	16.3	22.8
24.1	13.3	16.2	17.5	19.0	23.9	14.8	22.2	21.7	20.7
13.5	15.8	13.1	16.1	21.9	23.9	19.3	12.0	19.9	19.4
15.4	16.7	19.5	16.2	16.9	17.1	20.2	13.4	19.8	17.7
19.7	18.7	17.6	15.9	15.2	17.1	15.0	18.8	21.6	11.9

For the above data sample mean $\bar{x}=18000$

The prior distribution of θ is $N\left(14000, 2000\right)$,

$$f\left(\vec{x}/\theta\right) \sim N\left(\theta, 4000\right)$$

Now we know that if

Prior distribution of θ, $\pi\left(\theta\right) \sim N\left(\mu, \tau\right)$

$$f\left(\vec{x}/\theta\right) \sim N\left(\theta, \sigma\right)$$

then posterior distribution of θ

$$\pi\left(\theta/\vec{x}\right) \sim N\left(\frac{n\overline{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{\sqrt{n/\sigma^2 + 1/\tau^2}}\right)$$

Step 2/4
(a)

So in this case

$$\begin{aligned}\pi\left(\theta/\vec{x}\right) &\sim N\left(\frac{50\times 18000/4000^2 + 14000/2000^2}{50/4000^2 + 1/2000^2}, \frac{1}{\sqrt{50/4000^2 + 1/2000^2}}\right)\\ &\sim N\left(\frac{0.05975}{3.375\times 10^{-6}}, \frac{1}{\sqrt{3.375\times 10^{-6}}}\right)\\ &\sim N\left(17704, 544.33\right)\end{aligned}$$

So the Bayes estimator of θ,

$$\begin{aligned}\hat{\theta} &= E\left(\theta \mid x\right)\\ &= \boxed{17704}\end{aligned}$$

Step 3/4
(b)

The 90% HPD credible set for $\theta$is given by

$$\begin{aligned}90\%\ HPD &= \left[\mu_x - z_{.05}\tau_x, \mu_x + z_{.05}\tau_x\right]\\ &= \left[17704 - 1.645\times 544, 17704 + 1.645\times 544\right]\\ &= \boxed{\left[16808, 18600\right]}\end{aligned}$$

It means we are 90% confident that the number of concurrent users on an average lies between (16808, 18600).

Step 4/4
(c)

As 16000 lies below the credible set found in part (b), we say that at 90% significance level there is significant evidence that mean number of concurrent users exceeds 16000.

Chapter 10, Problem 35E
(0)

Problem
Continue Exercise 10.34. Another statistician conducts a non- Bayesian analysis of the data in Exercise 8.2 on p. 240 about concurrent users.

(a) Give the non-Bayesian estimator for the mean number of concurrent users *θ*.

(b) Construct a 90% confidence interval for *θ* and interpret it.

(c) Is there significant evidence that the mean number of concurrent users exceeds 16,000?

(d) How do your results differ from the previous exercise?

Step-by-step solution

Show all steps

Step 1/4
(a)

The data on the number of concurrent users (in thousands) is shown below :

17.2	22.1	18.5	17.2	18.6	14.8	21.7	15.8	16.3	22.8
24.1	13.3	16.2	17.5	19.0	23.9	14.8	22.2	21.7	20.7
13.5	15.8	13.1	16.1	21.9	23.9	19.3	12.0	19.9	19.4
15.4	16.7	19.5	16.2	16.9	17.1	20.2	13.4	19.8	17.7
19.7	18.7	17.6	15.9	15.2	17.1	15.0	18.8	21.6	11.9

For the above data sample mean $\overline{x} = 18000$ and sample standard deviation $\sigma = 3.16$

In the non Bayesian analysis we do not use the prior information available for θ and use only the information provided by the sample.

So the non-Bayesian estimate of θ is,

$$\hat{\theta} = \overline{x}$$
$$= \boxed{18000 \text{ users}}$$

Step 2/4
(b)

A 90% confidence interval for θ is given by

$$\begin{aligned} 90\%\ CI &= \left[\bar{x} - z_{.05}\frac{s}{\sqrt{n}}, \bar{x} - z_{.05}\frac{s}{\sqrt{n}}\right] \\ &= \left[18000 - 1.645 \times \frac{3157}{\sqrt{50}}, 18000 + 1.645 \times \frac{3157}{\sqrt{50}}\right] \\ &= \boxed{[17266, 18735]} \end{aligned}$$

It means without any prior information based on a sample of size 50, we are 90% confident that the number of concurrent users on an average lies between (17266, 18735).

Step 3/4
(c)

As 16000 lies below the credible set found in part (b), we say that at 90% significance level there is significant evidence that mean number of concurrent users exceeds 16000.

Step 4/4
(d)

In this exercise the credible set is broader than the previous one. It means adding prior information increases the credibility of the estimation.

Chapter 10, Problem 36E
(1)

Problem
In Example 9.13 on p. 258, we constructed a confidence interval for the population mean μ based on the observed Normally distributed measurements. Suppose that prior to the experiment we thought this mean should be between 5.0 and 6.0 with probability 0.95.

(a) Find a conjugate prior distribution that fully reflects your prior beliefs.

(b) Derive the posterior distribution and find the Bayes estimator of μ. Compute its posterior risk.

(c) Compute a 95% HPD credible set for μ. Is it different from the 95% confidence interval? What causes the differences?

Step-by-step solution

Show all steps

100% (1 rating) for this solution

Step 1/3
(a)

It is given that prior to the experiment, the population mean should be between 5.0 and 6.0 with probability 0.95. Comparing it with the confidence limit of the normal distribution

$$\begin{aligned}&\left[\mu - z_{.025}\tau, \mu + z_{.025}\tau\right]\\&=\left[\mu - 1.96\tau, \mu + 1.96\tau\right]\end{aligned}$$

We get

$\mu - 1.96\tau = 5.0$,

$\mu + 1.96\tau = 6.0$

Solving the above two equations, we get

$\hat{\mu} = 5.5$, $\tau = 0.255$

So the conjugate prior distribution was $\pi(\mu) = N(5.5, 0.255)$

Step 2/3
(b)

According to the referred example, sample size *n* = 6, sample mean $\overline{x} = 6.50$, population standard deviation $\sigma = 2.2$

i.e. $f(\vec{x}/\mu) \sim N(\mu, 2.2)$

So posterior distribution of μ

$$\pi(\mu/\vec{x}) \sim N\left(\frac{n\overline{x}/\sigma^2 + \hat{\mu}/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{\sqrt{n/\sigma^2 + 1/\tau^2}}\right)$$

So in this case

$$\begin{aligned}\pi(\mu/\vec{x}) &\sim N\left(\frac{6\times 6.5/2.2^2 + 5.5/0.255^2}{6/2.2^2 + 1/0.255^2}, \frac{1}{\sqrt{6/2.2^2 + 1/0.255^2}}\right)\\&\sim N\left(\frac{92.641}{16.618}, \frac{1}{\sqrt{16.618}}\right)\\&\sim N(5.575, 0.245)\end{aligned}$$

So the Bayes estimator of μ ,

$$\begin{aligned} \bar{\mu} &= E(\mu \mid x) \\ &= \boxed{5.575} \end{aligned}$$

and its posterior risk is

$$\begin{aligned} p(\hat{\mu}) &= Var(\mu \mid x) \\ &= \frac{1}{16.618} \\ &= \boxed{0.060} \end{aligned}$$

Step 3/3
(c)

The 95% HPD credible set for μ is

$$\begin{aligned} 95\%\ HPD &= \left[\mu_x - z_{.025}\tau_x, \mu_x + z_{.025}\tau_x\right] \\ &= \left[5.575 - 1.96 \times 0.245, 5.575 + 1.96 \times 0.245\right] \\ &= \boxed{[5.094, 6.055]} \end{aligned}$$

While the 95% confidence interval calculated previously was [4.74, 8.26]

We see that HPD credible set is narrower than the confidence interval because it uses the additional prior information.

Chapter 10, Problem 37E
(0)

Problem
If ten coin tosses result in ten straight heads, can this coin still be fair and unbiased?

By looking at a coin, you believe that it is fair (a 50-50 chance of each side) with probability 0.99. This is your prior probability. With probability 0.01, you allow the coin to be biased, one way or another, so its probability of heads is Uniformly distributed between 0 and 1. Then you toss the coin ten times, and each time it turns up heads. Compute the posterior probability that it is a fair coin.

Step-by-step solution

Show all steps

Step 1/2

The prior distribution of θ (probability of heads) is $Uniform(0,1)$, And $f(x/\theta) \sim Binomial(n, \theta)$

Then based on the sample having *k* successes, the likelihood function is given by

$$f(k/\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

and

$$\pi(\theta) \sim 1$$

So the posterior distribution of θ comes out to be as follows:

$$\begin{aligned} f(\theta/k) &\propto \theta^k (1-\theta)^{n-k} \\ &\propto \theta^k (1-\theta)^{n-k} \end{aligned}$$

this is a beta distribution $\beta(k+1, n-k+1)$.

Step 2/2

The posterior probability that it is a fair coin is computed as follows:

$$\begin{aligned} \text{Prob}(\text{fair coin}) &= 1 - P(\theta > 0.5/\bar{x}) \\ &= 0.0005 \end{aligned}$$

The posterior probability that it is a fair coin is 0.0005.

Chapter 10, Problem 38E

(0)

Problem

Observed is a sample from Uniform(0, *θ*) distribution.

(a) Find a conjugate family of prior distributions (you can find it in our inventory in Section A.2).

(b) Assuming a prior distribution from this family, derive a form of the Bayes estimator and its posterior risk.

Step-by-step solution

Show all steps

Step 1/2

(a)

Given that the sample follows a $Uniform(0,\theta)$ distribution, whose *pdf* is given by

$$f(x/\theta)=\frac{1}{\theta}\ 0<x<\theta$$

We want a conjugate family of prior distributions; it means we want a distribution such that after finding posterior, it should the same *pdf* as the prior one.

Let the sample be $\vec{x}=(x_1,x_2,...,x_n)$, the likelihood function for θ for making its posterior distribution is

$$\begin{aligned}f(\vec{x}/\theta)&=\prod_{i=1}^{n}\frac{1}{\theta}\\&=\frac{1}{\theta^n}\end{aligned}$$

After a thorough inspection of various density functions we see that $Pareto(\alpha,\sigma)$ distribution fulfills the criterion of conjugate prior.

The *pdf* of $Pareto(\alpha,\sigma)$ is given by $\pi(\theta)=\alpha\sigma^{\alpha}\theta^{-\alpha-1}$

Based on the sample $\vec{x}$ the posterior distribution of θ is given as shown below:

$$\begin{aligned}f(\theta/\vec{x})&\sim f(\vec{x}/\theta)\pi(\theta)\\&\sim\frac{1}{\theta^n}\alpha\sigma^{\alpha}\theta^{-\alpha-1}\\&\sim\alpha\sigma^{\alpha}\theta^{-\alpha-n-1}\end{aligned}$$

Step 2/2
(b)

The Bayes estimator of θ is the expected value of θ in Pareto distribution which is given by,

$$\begin{aligned}\hat{\theta}&=\frac{\alpha'\sigma}{\alpha'-1}\\&=\frac{(\alpha+n)\sigma}{\alpha+n-1}\end{aligned}$$

Note that in the posterior distribution the parameter α is replaced by $\alpha+n$

And the posterior risk is shown below:

$$\begin{aligned} p\left(\hat{\theta}\right) &= Var\left(\theta \mid x\right) \\ &= \frac{\alpha'\sigma^2}{\left(\alpha'-1\right)^2\left(\alpha'-2\right)} \\ &= \frac{\left(\alpha+n\right)\sigma^2}{\left(\alpha+n-1\right)^2\left(\alpha+n-2\right)} \end{aligned}$$

Chapter 10, Problem 40E
(0)

Problem
Anton played five chess games and won all of them. Let us estimate *θ*, his probability of winning the next game. Suppose that parameter *θ* has Beta(4,1) prior distribution and that the game results are independent of each other.

(a) Compute the Bayes estimator of *θ* and its posterior risk.

(b) Construct a 90% HPD credible set for *θ*.

(c) Is there significant evidence that Anton's winning probability is more than 0.7? To answer this question, find the posterior probabilities of $\theta \leq 0.7$ and $\theta > 0.7$ and use them to test H_0: $\theta \leq 0.7$ vs H_A: $\theta > 0.7$.

Notice that the standard frequentist estimator of *θ* is the sample proportion $\hat{p} = 1,$, which is rather unrealistic because it gives Anton no chance to lose!

Step-by-step solution

Show all steps

Step 1/4
(a)

Let *k* be a binomial random variable with parameters *n* & θ. Where n is the no. of trials and θ is the probability of success. Let the prior distribution of θ follows a $Beta\left(\alpha,\beta\right)$ distribution.

Then based on the sample having *k* successes, the likelihood function is given by

$$f\left(k/\theta\right) = \binom{n}{k}\theta^k\left(1-\theta\right)^{n-k}$$

and $\pi\left(\theta\right) \sim \theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}$

So the posterior distribution of θ comes out to be as shown below:

$$\begin{aligned} f(\theta/k) &\propto \theta^{k}(1-\theta)^{n-k}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{k+\alpha-1}(1-\theta)^{n+\beta-k-1} \end{aligned}$$

Step 2/4

The posterior distribution of θ follows a $Beta(k+\alpha, n+\beta-k)$

In the given problem n = 5, k = 5, α = 4, β =1.

Here the posterior distribution of θ follows a $Beta(9,1)$

The Bayes estimator $\hat{\theta}$is the expected value of θ by $Beta(9,1)$i.e, as shown below:

$$\begin{aligned} \hat{\theta} &= \frac{9}{9+1} \\ &= \boxed{0.9} \end{aligned}$$

and its posterior risk is computed as follows:

$$\begin{aligned} \rho(\hat{\theta}) &= \frac{9\cdot 1}{(9+1)^2(9+1+1)} \\ &= \boxed{0.0082} \end{aligned}$$

Step 3/4
(b)

Let L &U be the lower and upper limits of 90% HPD credible set for θ. Then we can calculate these values by several methods. One of this is shown below:

$$\int_0^L f(\theta/k)\,d\theta = 0.05$$

And $\int_0^U f(\theta/k)\,d\theta = 0.95$

Above integrations can be obtained by the excel function $=BETAINV(prob,\alpha,\beta)$and gives $L = 0.717$, $U = 0.999$

Step 4/4
(c)

To test the hypothesis $H_0 : \theta \le 0.7$ vs $H_A : \theta > 0.7$,

$$\rho(\text{reject } H_0) = \omega P(\theta > 0.7 / \vec{x})$$
$$= 0.96\omega$$

Here ω is the loss of error under accepting or rejecting H_0 (which is same in this problem).

and

$$\rho(\text{accept } H_0) = \omega P(\theta \le 0.7 / \vec{x})$$
$$= 0.04\omega$$

Hence, there is significant evidence to reject H_0.

Chapter 11, Problem 1E
(0)

Problem
The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average size of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmittance time was 0.04 seconds with the standard

deviation of 0.01 seconds. The correlation coefficient between the time and the size was 0.86.

Based on this data, fit a linear regression model and predict the time it will take to transmit a 400 Kbyte file.

Step-by-step solution

Show all steps

100% (2 ratings) for this solution

Step 1/4
Let *S* represent the size of the file.

From the given information,

Hence, the Average file size,

$\overline{S} = 126\text{Kbytes}$

The standard deviation of file size,

$s_S = 35$ Kbytes

Step 2/4
Let *T* denotes the time will take to transmit a file.

Average time of file transmission,

$\overline{T} = 0.04$ Seconds,

Standard deviation of file transmission,

$s_T = 0.01$ Seconds

The correlation coefficient between the time and size $r = 0.86$

Step 3/4
The fitted regression model between the time and file size is given by,

$$T - \overline{T} = r\frac{s_T}{s_s}\left(S - \overline{S}\right)$$

Substitute the given values in the above regression model,

$$T-0.04=0.86\frac{0.01}{35}(S-126)$$

$$T=0.86\frac{0.01}{35}(S-126)+0.04$$

$$T=2.457\times10^{-4}S+9.04\times10^{-3}$$

Therefore, the required regression equation is,

$$\boxed{T=2.457\times10^{-4}S+9.04\times10^{-3}}$$

Step 4/4
By using the above equation we can predict the time it will take to transmit a 400 Kbyte file.

Here *S* = 400.

Substituting this value in the above equation,

$$\begin{aligned}T&=\left(2.457\times10^{-4}\times400\right)+\left(9.04\times10^{-3}\right)\\&=\left(9.828\times10^{-2}\right)+\left(9.04\times10^{-3}\right)\\&=\boxed{0.107\text{ seconds}}\end{aligned}$$

Chapter 11, Problem 2E
(0)

Problem
The following statistics were obtained from a sample of size *n* = 75:

– the predictor variable *X* has mean 32.2, variance 6.4;

–the response variable *Y* has mean 8.4, variance 2.8; and

– the sample covariance between *X* and *Y* is 3.6.

(a) Estimate the linear regression equation predicting *Y* based on *X*.

(b) Complete the ANOVA table. What portion of the total variation of *Y* is explained by variable *X*?

(c) Construct a 99% confidence interval for the regression slope. Is the slope significant?

Step-by-step solution

Show all steps

Step 1/7
From the given information,

The mean and variance of the predictor variable *X* is,

$$\bar{X} = 32.2$$

$$\sigma_x{}^2 = 6.4$$

The mean and variance of the response variable *Y* is,

$$\bar{Y} = 8.4$$

$$\sigma_Y{}^2 = 2.8$$

The sample covariance between the variables X and *Y* is,

$$\operatorname{cov}(X,Y) = 3.6$$

Step 2/7
(a)

The linear regression equation predicting *Y* based on *X* is expressed as follows:

$$\begin{aligned} Y-\bar{Y} &= \frac{\operatorname{cov}(X,Y)}{\sigma_x{}^2}\left(X-\bar{X}\right) \\ Y-8.4 &= \frac{3.6}{6.4}\left(X-32.2\right) \\ Y &= 0.56X-18.11+8.4 \\ Y &= 0.56X-9.71 \end{aligned}$$

Therefore, the required regression equation is $\boxed{Y=0.56X-9.71}$

Step 3/7
(b)

Following parameters are to be calculated for making ANOVA table:

$$\begin{aligned} SS_{TOT} &= (n-1)s_Y{}^2 \\ &= (75-1)\times 2.8 \\ &= 207.2 \end{aligned}$$

$$\begin{aligned} SS_{REG} &= (n-1)b_1{}^2 s_X{}^2 \\ &= (75-1)\times 0.5625^2\times 6.4 \\ &= 149.85 \end{aligned}$$

$$\begin{aligned} SS_{ERR} &= SS_{TOT} - SS_{REG} \\ &= 207.2 - 149.85 \\ &= 57.35 \end{aligned}$$

Step 4/7

ANOVA table for the above model is given by,

Source	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Model	149.85	1	149.85	190.742
Error	57.35	73	0.7856	
Total	207.2	74	2.8	

The explained portion of total variation will be estimated by *R*-square using the following formula:

$$\begin{aligned} R^2 &= \frac{SS_{REG}}{SS_{TOT}} \\ &= \frac{149.85}{207.2} \\ &= 0.723 \end{aligned}$$

It means by the regression model the variable *X* explains 72.3% of the total variation of *Y*.

Step 5/7

(c)

The $(1-\alpha)100\%$ confidence interval for the slope b_1 is given by,

$$\left(b_1 - t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}},\ b_1 + t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}} \right)$$

Here,

$$\begin{aligned} s &= \sqrt{0.7856} \\ &= 0.886 \end{aligned}$$

For 99% confidence interval,

$$t_{0.005,74} = 2.894$$

Step 6/7

$$\begin{aligned} S_{xx} &= \sum (x_i - \overline{x})^2 \\ &= ns_x^{\,2} \\ &= 75 \times 6.4 \\ &= 480 \end{aligned}$$

Hence, the 99% CI for the regression slope is calculated as follows:

$$\begin{aligned} CI_{99\%} &= \left(b_1 - t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}},\ b_1 + t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}} \right) \\ &= \left(0.5625 - 2.894 \times \frac{0.886}{\sqrt{480}},\ 0.5625 - 2.894 \times \frac{0.886}{\sqrt{480}} \right) \\ &= \boxed{(0.455, 0.670)} \end{aligned}$$

Step 7/7

The calculated value of *t* is

$$\begin{aligned} t &= \frac{b_1 \sqrt{S_{xx}}}{s} \\ &= \frac{0.5625 \times \sqrt{480}}{0.886} \\ &= 13.91 \end{aligned}$$

Therefore, the calculated value of *t* is more than the theoretical value of *t*-distribution at 99% level of significance and 74 degrees of freedom.

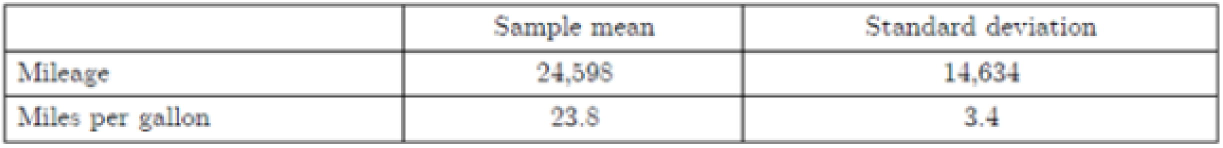
Therefore, there is a sufficient evidence to conclude that the regression slope is significant at 99% level.

Chapter 11, Problem 3E

(0)

Problem

At a gas station, 180 drivers were asked to record the mileage of their cars and the number of miles per gallon. The results are summarized in the table.



	Sample mean	Standard deviation
Mileage	24,598	14,634
Miles per gallon	23.8	3.4

The sample correlation coefficient is $r = -0.17$.

(a) Compute the least squares regression line which describes how the number of miles per gallon depends on the mileage. What do the obtained slope and intercept mean in this situation?

(b) Use R^2 to evaluate its goodness of fit. Is this a good model?

(c) You purchase a used car with 35,000 miles on it. Predict the number of miles per gallon. Give a 95% prediction interval for your car and a 95% confidence interval for the average number of miles per gallon of all cars with such a mileage.

Step-by-step solution

Show all steps

Step 1/7

Let the random variable *X* represent mileage.

Let the random variable *Y* represent number of miles per gallon.

From the given information,

The mean and standard deviation of the random variable *X* is,

$$\bar{X} = 24598$$

$$\sigma_x = 14634$$

The mean and standard deviation of the random variable *Y* is,

$$\bar{Y} = 23.8$$

$$\sigma_Y = 3.4$$

The correlation coefficient between the random variables *X* and *Y* is $r = -0.17$

Step 2/7

(a)

The linear regression equation predicting *Y* based on *X* is expressed as follows:

$$\begin{aligned} Y - \bar{Y} &= r\frac{\sigma_y}{\sigma_x}\left(X - \bar{X}\right) \\ Y - 23.8 &= -0.17 \times \frac{3.4}{14634}\left(X - 24598\right) \\ Y - 23.8 &= \left(-0.17\right)\left(0.000232336\right)\left(X - 24598\right) \\ Y &= 24.771 - 3.95 \times 10^{-5} X \end{aligned}$$

Therefore, the required regression equation is $\boxed{Y = 24.771 - 3.95 \times 10^{-5} X}$

Here, the slope is $b_1 = 3.95 \times 10^{-5}$

And intercept mean $b_0 = 24.771$

Step 3/7
(b)

Compute the value of R^2 as follows:

$$\begin{aligned} SS_{REG} &= (n-1) b_1^2 s_X^2 \\ &= (180-1) \times \left(3.95 \times 10^{-5}\right)^2 \times 14634^2 \\ &= 59.8 \end{aligned}$$

$$\begin{aligned} SS_{TOT} &= (n-1) s_Y^2 \\ &= (180-1) \times 3.4^2 \\ &= 2069.2 \end{aligned}$$

$$\begin{aligned} R^2 &= \frac{SS_{REG}}{SS_{TOT}} \\ &= \frac{59.8}{2069.2} \\ &= 0.029 \end{aligned}$$

Since, the regression model the variable *X* explains only 2.9% of the total variation of *Y*

Hence, this is not a good model.

Step 4/7
(c)

To predict number of miles per gallon with the mileage of , substitute the value of *X* = the following equation.

$$\begin{aligned} Y &= 24.771 - 3.95 \times 10^{-5} X \\ &= 24.771 - 3.95 \times 10^{-5} (35000) \\ &= 24.771 - 1.382 \\ &= \boxed{23.39 \text{ miles per gallon}} \end{aligned}$$

Following parameters are required to calculate a 95% prediction interval for a given value of *X* = *x*.

$$\begin{aligned} SS_{ERR} &= SS_{TOT} - SS_{REG} \\ &= 2069.2 - 59.8 \\ &= 2009.4 \end{aligned}$$

$$\begin{aligned} s &= \sqrt{\frac{SS_{ERR}}{(n-2)}} \\ &= \sqrt{\frac{2009.4}{178}} \\ &= 3.36 \end{aligned}$$

Step 5/7

$$\begin{aligned} S_{xx} &= \sum (x_i - \overline{x})^2 \\ &= n{s_x}^2 \\ &= 180 \times 14634^2 \\ &= 3.85 \times 10^{10} \end{aligned}$$

Step 6/7

The 95% prediction interval for a given value of *X* = *x* is calculated by using the following formula:

$$CI_{95\%} = b_0 + b_1 x \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$

Substitute the values in the above formula:

$$\begin{aligned} CI_{95\%} &= 24.771 - 3.95 \times 10^{-5} \times 35000 \pm 2.26 \times 3.36 \times \sqrt{1 + \frac{1}{180} + \frac{(35000-24598)^2}{3.85 \times 10^{10}}} \\ &= 23.39 \pm 7.59 \times \sqrt{1 + \frac{1}{180} + \frac{(35000-24598)^2}{3.85 \times 10^{10}}} \\ &= 23.39 \pm 7.59 \times 1.004 \\ &= (15.77,\ 31.01) \end{aligned}$$

Step 7/7

The 95% confidence interval for the average number of miles per gallon is calculated by using the following formula:

$$CI_{95\%} = b_0 + b_1 x \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$

Substitute the values in the above formula:

$$
\begin{aligned}
CI_{95\%} &= 24.771 - 3.95\times10^{-5}\times35000 \pm 2.26\times3.36\times\sqrt{\frac{1}{180}+\frac{(35000-24598)^2}{3.85\times10^{10}}} \\
&= 23.39 \pm 7.59\times\sqrt{\frac{1}{180}+\frac{(35000-24598)^2}{3.85\times10^{10}}} \\
&= 23.39 \pm 7.59\times0.091 \\
&= (22.7,\ 24.1)
\end{aligned}
$$

Chapter 11, Problem 6E

(0)

Problem

For a univariate linear regression, show that

$$SS_{\mathrm{TOT}} = SS_{\mathrm{REG}} + SS_{\mathrm{ERR}}.$$

Hint: Write $SS_{\mathrm{TOT}} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}((y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}))^2.$

Step-by-step solution

Show all steps

Step 1/2

The objective is to show that for a univariate linear regression,

$$SS_{TOT} = SS_{REG} + SS_{ERR}$$

$$
\begin{aligned}
SS_{TOT} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})
\end{aligned}
$$

Step 2/2

Since,

$$SS_{ERR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SS_{REG} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

Hence, substitute the notations in the above equation as follows:

$$\begin{aligned}SS_{TOT} &= \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2+\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\overline{y}\right)\\ &= SS_{ERR}+SS_{REG}+\sum_{i=1}^{n}\left(y_i-b_0-b_1x_i\right)\left(b_0+b_1x_i-\overline{y}\right)\\ &= SS_{ERR}+SS_{REG}+\sum_{i=1}^{n}\varepsilon_i\left(b_0+b_1x_i-b_0-b_1\overline{x}\right)\end{aligned}$$

$$\begin{aligned}&= SS_{ERR}+SS_{REG}+\sum_{i=1}^{n}\varepsilon_i\left(b_1x_i-b_1\overline{x}\right)\\ &= SS_{ERR}+SS_{REG}+0 \quad \text{(by the assumption of least squares)}\\ &= SS_{ERR}+SS_{REG}\end{aligned}$$

Therefore, the above proof show‘s that $SS_{TOT}=SS_{REG}+SS_{ERR}$.

Chapter 11, Problem 7E
(0)

Problem
For a univariate linear regression, show that R-square is the squared sample correlation coefficient,

$$R^2=r^2.$$

Hint: Write the regression sum of squares as

$$SS_{\text{REG}}=\sum_{i=1}^{n}\left(b_0+b_1x_i-\bar{y}\right)^2$$

and substitute our derived expressions for the regression intercept b_0 and slope b_1.

Step-by-step solution

Show all steps

Step 1/3
Since, *R*-square is expressed as follows:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}}$$

$$= \frac{\sum_{i=1}^{n}\left(b_0 + b_1x_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

$$= \frac{\sum_{i=1}^{n}\left(b_0 + b_1x_i - b_0 - b_1\overline{x}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

$$= \frac{b_1{}^2\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

Step 2/3

From the regression line of *Y* on *X*, the regression coefficient is, expressed as follows:

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

Here, s_x represents the sample standard deviation of the random variable *X.*

Here, s_Y represents the sample standard deviation of the random variable *Y.*

Hence, the sample variances of the random variables *X* and *Y* are expressed as follows:

$$s_x{}^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

$$s_y{}^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

Step 3/3

Substituting all these values in the expression of R^2,

$$\begin{aligned} R^2 &= \frac{b_1^2 \sum_{i=1}^{n}(x_i-\overline{x})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \\ &= \left(r\frac{s_y}{s_x}\right)^2 \frac{s_x^2}{s_y^2} \\ &= r^2 \end{aligned}$$

Therefore, from the above result there is a sufficient evidence that $R^2 = r^2$

Chapter 11, Problem 8E

(0)

Problem

Anton wants to know if there is a relation between the number of hours he spends preparing for his weekly quiz and the grade he receives on it. He keeps records for 10 weeks. It turns out that on the average, he spends 3.6 hours a week preparing for the quiz, with the standard deviation of 0.5 hours. His average grade is (out of 100), with the standard deviation of 14. The correlation between the two variables is *r* = 0.62.

(a) Find the equation of the regression line predicting the quiz grade based on the time spent on preparation.

(b) This week, Anton studied for 4 hours. Predict his grade.

(c) Does this linear model explain most of the variation? Is it a good fit? Why?

Step-by-step solution

Show all steps

Step 1/4

(a)

Let the random variable *X* represents the number of hours that Anton spends for his weekly quiz.

Let the random variable *Y* represents the grade received by Anton.

From the given information,

n = 10

The mean and standard deviation of the random variable *X* is,

$\overline{X} = 3.6$

$\sigma_x = 0.5$

The mean and standard deviation of the random variable *Y* is,

$\overline{Y} = 82$

$\sigma_Y = 14$

The correlation coefficient between the random variables *X* and *Y* is,

$r = 0.62$

Step 2/4

The linear regression equation predicting *Y* based on *X* is expressed as follows:

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$$

Substitute the given values in the above equation for predicting *Y* based on *X* is as follows:

$$\begin{aligned} Y - 82 &= 0.62 \times \frac{14}{0.5}\left(X - 3.6\right) \\ Y - 82 &= 17.36\left(X - 3.6\right) \\ Y - 82 &= 17.36X - 62.496 \\ Y &= 17.36X + 19.50 \end{aligned}$$

Therefore, the required regression equation is $\boxed{Y = 19.50 + 17.36X}$

Step 3/4

(b)

If Anton studied for 4 hours, then the predicted grade can be obtained by substituting *x* = the above equation.

$$\begin{aligned} Y &= 19.50 + 17.36 \times 4 \\ &= 19.50 + 69.44 \\ &= 88.94 \end{aligned}$$

Therefore, the predicted grade is 88.94.

Step 4/4

(c)

The measure of goodness of fit is calculated as follows:

$$\begin{aligned} R^2 &= r^2 \\ &= 0.62^2 \\ &= 0.38 \end{aligned}$$

The model explains only 38% of the variation in *Y*.

Therefore, it is not a very good fit model.

Chapter 11, Problem 9E
(0)

Problem
For the data in Example 11.10 and data set Efficiency, fit a linear regression model predicting the program efficiency (number of processed requests) based on the database structure x_2 only. Complete the ANOVA table, compute R-square and adjusted R-square. Is this reduced model significant?

Step-by-step solution

Show all steps

Step 1/7

The data in example (11.10) including program efficiency and x_2 is shown below:

No. of tables (x_2)	4	20	20	10	10	2	1
y	40	55	50	41	17	26	16

From the given information,

$$n = 7$$

Hence, mean of the variables x_2 and *Y* is,

$\bar{x}_2 = 9.57$, $\bar{y} = 35.0$

Step 2/7
Therefore, further calculations are shown in the following table:

No. of tables (x2)	4	20	20	10	10	2	1	67
y	40	55	50	41	17	26	16	245
S_{XX}	31	109	109	0.18	0.18	57.33	73.5	379.71
S_{XY}	-28	209	156	2.57	-7.71	68.14	163	563.00

S_{YY}	25	400	225	36	324	81	361	1452.00

The last column indicates the row totals:

$$\begin{aligned} S_{xx} &= \sum (x_i - \bar{x})^2 \\ &= 379.71 \end{aligned}$$

$$\begin{aligned} S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) \\ &= 563.00 \end{aligned}$$

$$\begin{aligned} S_{yy} &= \sum (y_i - \bar{y})^2 \\ &= 1452.00 \end{aligned}$$

Step 3/7

The slope of the regression equation is calculated as follows:

$$\begin{aligned} b_1 &= \frac{S_{xy}}{S_{xx}} \\ &= \frac{563.00}{379.71} \\ &= 1.483 \end{aligned}$$

Step 4/7

Sum of squares:

$$\begin{aligned} SS_{REG} &= b_1^2 s_{xx} \\ &= 1.483^2 \times 379.71 \\ &= 834.76 \end{aligned}$$

$$\begin{aligned} SS_{TOT} &= S_{yy} \\ &= 1452.0 \end{aligned}$$

$$\begin{aligned} SS_{ERR} &= SS_{TOT} - SS_{REG} \\ &= 1452.00 - 834.76 \\ &= 617.24 \end{aligned}$$

Step 5/7

The ANOVA table for the given data is as follows:

Source	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Model	834.76	1	834.76	6.76
Error	617.24	5	123.45	
Total	1452.0	6	242.00	

Step 6/7

$$\begin{aligned} R^2 &= \frac{SS_{REG}}{SS_{TOT}} \\ &= \frac{834.76}{1452.0} \\ &= 0.575 \end{aligned}$$

The adjusted R^2 is given by,

$$\begin{aligned} R_{adj}{}^2 &= 1 - \frac{MSS_{ERR}}{MSS_{TOT}} \\ &= 1 - \frac{123.45}{242.0} \\ &= 0.490 \end{aligned}$$

Step 7/7
Conclusion:

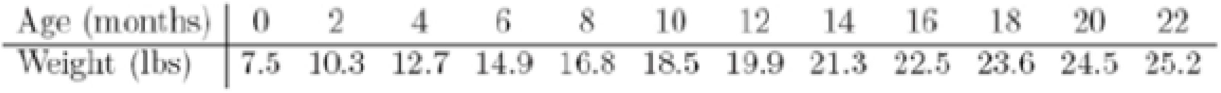
The theoretical value of *F* with df (1, 5) at 0.05 level of significance is 6.61 and at 0.025 level of significance is 10.

Since, the calculated value 6.76 lies between these two figures.

It means the model is significant at 0.05 level of significance but not significant at 0.025 level of significance.

Chapter 11, Problem 12E
(0)

Problem
Masha weighed 7.5 lbs when she was born. Then her weight increased according to the table.



Age (months)	0	2	4	6	8	10	12	14	16	18	20	22
Weight (lbs)	7.5	10.3	12.7	14.9	16.8	18.5	19.9	21.3	22.5	23.6	24.5	25.2

(a) Construct a time plot of Masha’s weight. What regression model seems appropriate for these data?

(b) Fit a quadratic model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + ?$. Why does the estimate of β_2 appear negative?

(c) What additional portion of the total variation is explained by the quadratic term? Would you keep it in the model?

(d) Predict Masha's weight at 24 months using the quadratic model and using the linear model. Which prediction is more reasonable?

(e) Can one test significance of the quadratic term by an F-test? Why or why not?

Step-by-step solution

Show all steps

Step 1/7

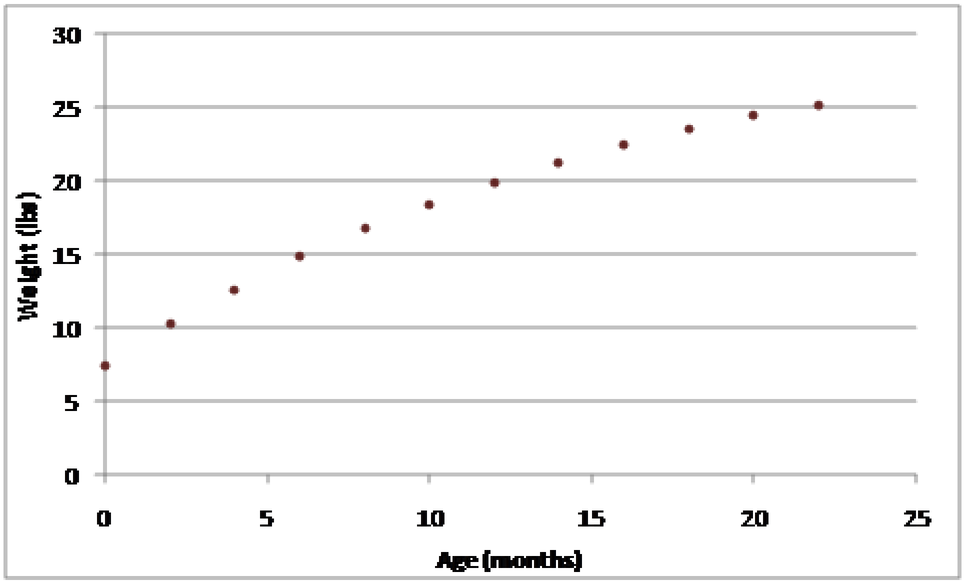
The data of Masha weights with age is shown below :

Age (*x*)	Weight (*y*)
0	7.5
2	10.3
4	12.7
6	14.9
8	16.8
10	18.5
12	19.9
14	21.3
16	22.5
18	23.6
20	24.5
22	25.2

Step 2/7

(a)

A time plot of the above data is given as follows :



The above figure indicates a non linear curved relationship. So a quadratic parabolic regression model seems appropriate for this data.

Step 3/7
(b)

We can fit this model by multivariate regression analysis.

The predictor matrix *X* and the response vector *Y* are :

X =

1	0	0
1	2	4
1	4	16
1	6	36
1	8	64
1	10	100
1	12	144
1	14	196
1	16	256
1	18	324
1	20	400
1	22	484

and

Y^T =

7.5	10.3	12.7	14.9	16.8	18.5	19.9	21.3	22.5	23.6	24.5	25.2

We then compute

$$X^T X = \begin{pmatrix} 12 & 132 & 2024 \\ 132 & 2024 & 34848 \\ 2024 & 34848 & 639584 \end{pmatrix}$$

and

$$X^T Y = \begin{pmatrix} 217.7 \\ 2846.4 \\ 46149.6 \end{pmatrix}$$

To obtain the estimated vector of slopes

$$b = \left(X^T X\right)^{-1}\left(X^T Y\right)$$
$$= \begin{pmatrix} 7.72 \\ 1.31 \\ -0.02 \end{pmatrix}$$

Thus the regression equation is

$$\boxed{y = 7.72 + 1.31x - 0.02x^2}$$

The negative quadratic term indicates that as age increases, the increase in weight slows.

Step 4/7
(c)

By using a statistical software (MINITAB or SPSS) we generate the following results:

$$s^2 = MS_{ERR}$$
$$= 0.019$$

Step 5/7
Excluding the quadratic term, the linear model depicts

$$R^2 = 0.967$$

And after including the quadratic term, the quadratic model depicts

$$R^2 = 0.999$$

The difference of the above two values is 0.0325

So the quadratic term explains 3.25% of the total variation.

For finding the significance of the quadratic term we find

$$Var(b) = s^2\left(X^T X\right)^{-1}$$
$$= \begin{pmatrix} 0.010 & -0.002 & 6.5\times10^{-5} \\ -0.002 & 0.000 & -2.0\times10^{-5} \\ 6.5\times10^{-5} & -2.0\times10^{-5} & 8.9\times10^{-7} \end{pmatrix}$$

From the above dispersion matrix

$$s(b_2) = \sqrt{8.9\times10^{-7}}$$
$$= 9.4\times10^{-4}$$

The *t*-statistic is then

$$\begin{aligned} t &= \frac{b_2}{s(b_2)} \\ &= \frac{0.019}{9.4\times10^{-4}} \\ &= 20.2 \end{aligned}$$

Which shows a high significance of the coefficient b_2

It is necessary to keep it in the model.

Step 6/7
(d)

Using the quadratic model the predicted weight at 24 months is :

$$\begin{aligned} y &= 7.72+1.31\times24-0.02\times24^2 \\ &= 25.53 \end{aligned}$$

The predicted weight at 24 months is .

Now the linear model using a statistical software (MINITAB or SPSS) is found to be

$$y = 9.455+0.790x$$

Using the linear model the predicted weight at 24 months is :

$$\begin{aligned} y &= 9.455+0.790\times24 \\ &= 28.42 \end{aligned}$$

The predicted weight at 24 months is .

Obviously the prediction using quadratic model is more reasonable.

Step 7/7
(e)

The significance of the quadratic term cannot be tested by the F test. Because that term is not independent of other term. This also shows from dispersion matrix calculated in part (c).

Chapter 11, Problem 13E
(0)

Problem
Refer to Exercise 8.6 on p. 241 and data set PopulationUSA. Does a linear regression model provide an adequate fit? Estimate regression parameters, make a plot of 10-year

increments in the U.S. population, along with your estimated regression line. What can you infer about the U.S. population growth?

Step-by-step solution

Show all steps

Step 1/4

The data of population increment with time is shown below :

Year	x = year - 1800	increment (y)
1800	0	1.4
1810	10	1.9
1820	20	2.4
1830	30	3.3
1840	40	4.2
1850	50	6.1
1860	60	8.2
1870	70	7.2
1880	80	11.6
1890	90	12.8
1900	100	13.2
1910	110	16
1920	120	13.8
1930	130	17.2
1940	140	9
1950	150	19.1
1960	160	28
1970	170	24
1980	180	23.2
1990	190	22.2
2000	200	32.7
2010	210	27.3

We already know the fact that population grows in quadratic way. So for the increment which is the difference in populations in two successive time points, a linear model will be fitted suitably.

Step 2/4

Now we will estimate the regression parameters using the following table :

Year	*x*	*y*	S_{xx}	S_{xy}	S_{yy}
1800	0	1.4	11025	1307.73	155.12
1810	10	1.9	9025	1135.68	142.91
1820	20	2.4	7225	973.64	131.21
1830	30	3.3	5625	791.59	111.40

1840	40	4.2	4225	627.55	93.21
1850	50	6.1	3025	426.50	60.13
1860	60	8.2	2025	254.45	31.97
1870	70	7.2	1225	232.91	44.28
1880	80	11.6	625	56.36	5.08
1890	90	12.8	225	15.82	1.11
1900	100	13.2	25	3.27	0.43
1910	110	16	25	10.73	4.60
1920	120	13.8	225	-0.82	0.00
1930	130	17.2	625	83.64	11.19
1940	140	9	1225	-169.91	23.57
1950	150	19.1	2025	236.05	27.51
1960	160	28	3025	778.00	200.09
1970	170	24	4225	659.45	102.93
1980	180	23.2	5625	700.91	87.34
1990	190	22.2	7225	709.36	69.65
2000	200	32.7	9025	1790.32	355.15
2010	210	27.3	11025	1411.77	180.78
Total	**2310**	**304.8**	**88550**	**12035**	**1839.67**

Here $n = 22$, $\overline{x} = 105$, $\overline{y} = 13.85$

The last row of the table indicates the column totals.

$$\begin{aligned} S_{xx} &= \sum \left(x_i - \overline{x}\right)^2 \\ &= 88550 \end{aligned}$$

$$\begin{aligned} S_{xy} &= \sum \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) \\ &= 12035 \end{aligned}$$

$$\begin{aligned} S_{yy} &= \sum \left(y_i - \overline{y}\right)^2 \\ &= 1840 \end{aligned}$$

Step 3/4

The regression slope is calculated as

$$\begin{aligned} b_1 &= \frac{S_{xy}}{S_{xx}} \\ &= \frac{12035}{88550} \\ &= 0.136 \end{aligned}$$

and the constant term

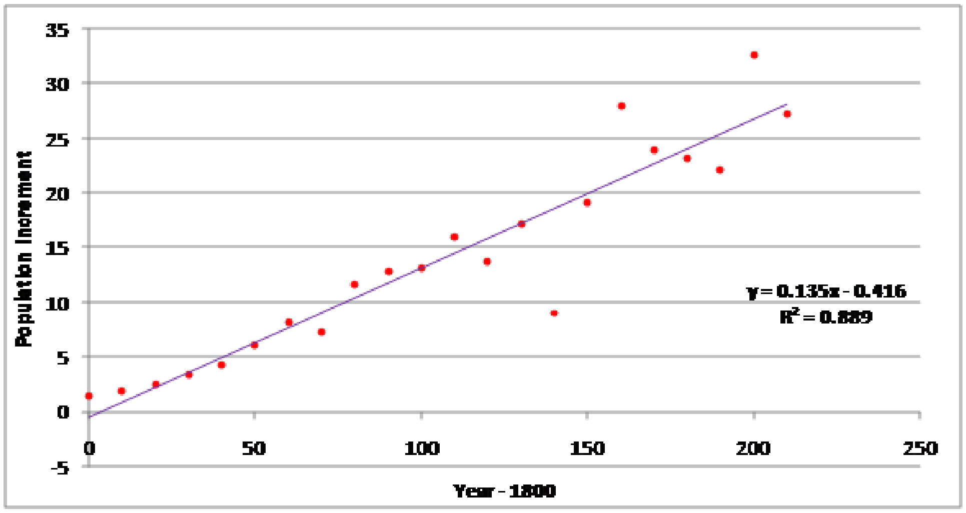
$$
\begin{aligned}
b_0 &= \bar{y} - b_1\bar{x} \\
&= 13.85 - 0.136 \times 105 \\
&= -0.416
\end{aligned}
$$

So the linear regression equation represents the relationship between population increment in millions (*y*) and year – 1800 (*x*) is given by

$$\boxed{Y = -0.416 + 0.136X}$$

Step 4/4

A plot of 10 year increments in the population along with the estimated regression line is shown below :



The 88.9% of the variation in U.S. population increment is explained by a linear model.

Chapter 11, Problem 14E

(0)

Problem

Refer to Exercise 8.7 on p. 241 and data set PopulationUSA.

(a) Fit a linear regression model to the 10-year relative change of the U.S. population. Estimate the regression parameters.

(b) Complete the ANOVA table and compute R-square.

(c) Conduct ANOVA F-test and comment on the significance of the

fitted model.

(d) Compute a 95% confidence interval for the regression slope.

(e) Compute 95% prediction intervals for the relative change of the population between years 2010 and 2020, and for the relative change between years 2020 and 2030.

(f) Construct a histogram of regression residuals. Does it support our assumption of the normal distribution?

Step-by-step solution

Show all steps

Step 1/9

The data of population relative changes with time is shown below :

Year	x = year - 1800	increment (y)
1800	0	0.359
1810	10	0.358
1820	20	0.333
1830	30	0.344
1840	40	0.326
1850	50	0.357
1860	60	0.353
1870	70	0.229
1880	80	0.301
1890	90	0.255
1900	100	0.210
1910	110	0.210
1920	120	0.150
1930	130	0.162
1940	140	0.073
1950	150	0.144
1960	160	0.185
1970	170	0.134
1980	180	0.114
1990	190	0.098
2000	200	0.131
2010	210	0.097

(a)

Now we will estimate the linear regression parameters and model using the following table :

Year	x	y	S_{xx}	S_{xy}	S_{yy}
1800	0	0.359	11025	-14.19	0.0183
1810	10	0.358	9025	-12.80	0.0181
1820	20	0.333	7225	-9.31	0.0120
1830	30	0.344	5625	-9.00	0.0144

1840	40	0.326	4225	-6.62	0.0104
1850	50	0.357	3025	-7.31	0.0177
1860	60	0.353	2025	-5.83	0.0168
1870	70	0.229	1225	-0.19	0.0000
1880	80	0.301	625	-1.92	0.0059
1890	90	0.255	225	-0.47	0.0010
1900	100	0.210	25	0.07	0.0002
1910	110	0.210	25	-0.07	0.0002
1920	120	0.150	225	-1.11	0.0055
1930	130	0.162	625	-1.54	0.0038
1940	140	0.073	1225	-5.28	0.0227
1950	150	0.144	2025	-3.57	0.0063
1960	160	0.185	3025	-2.13	0.0015
1970	170	0.134	4225	-5.85	0.0081
1980	180	0.114	5625	-8.23	0.0120
1990	190	0.098	7225	-10.69	0.0158
2000	200	0.131	9025	-8.77	0.0085
2010	210	0.097	11025	-13.31	0.0161
Total	**2310**	**4.92361**	**88550**	**-128.11**	**0.2153**

Here $n = 22$, $\overline{x} = 105$, $\overline{y} = 0.224$

The last row of the table indicates the column totals.

$$\begin{aligned} S_{xx} &= \sum \left(x_i - \overline{x}\right)^2 \\ &= 88550 \end{aligned}$$

$$\begin{aligned} S_{xy} &= \sum \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) \\ &= -128.11 \end{aligned}$$

$$\begin{aligned} S_{yy} &= \sum \left(y_i - \overline{y}\right)^2 \\ &= 0.2153 \end{aligned}$$

Step 2/9

The regression slope is calculated as

$$\begin{aligned} b_1 &= \frac{S_{xy}}{S_{xx}} \\ &= \frac{-128.11}{88550} \\ &= \boxed{-0.0015} \end{aligned}$$

and the constant term

$$\begin{aligned} b_0 &= \overline{y} - b_1\overline{x} \\ &= 0.224 - 0.0015 \times 105 \\ &= \boxed{0.375} \end{aligned}$$

So the linear regression equation represents the relationship between population increment in millions (*y*) and year – 1800 (*x*) is given by

$$\boxed{y = 0.375 - 0.0015x}$$

Step 3/9
(b)

We can make ANOVA table in the following way :

$$\begin{aligned} SS_{REG} &= b_1^2 S_{xx} \\ &= (-0.0015)^2 \times 88550 \\ &= 0.185 \end{aligned}$$

$$\begin{aligned} SS_{TOT} &= S_{yy} \\ &= 0.2153 \end{aligned}$$

$$\begin{aligned} SS_{ERR} &= SS_{TOT} - SS_{REG} \\ &= 0.2153 - 0.1853 \\ &= 0.0299 \end{aligned}$$

The ANOVA table is then completed as

Source	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Model	0.185	1	0.185	185
Error	0.0299	20	0.001	
Total	0.2153	21	0.010	

Now

$$\begin{aligned} R^2 &= \frac{SS_{REG}}{SS_{TOT}} \\ &= \frac{0.185}{0.2153} \\ &= \boxed{0.861} \end{aligned}$$

Step 4/9
(c)

For the above F-value, p-value comes out to be <0.001. It means the fitted model is highly significant.

Step 5/9

(d)

Now,

$$\begin{aligned} s &= \sqrt{MSS_{ERR}} \\ &= \sqrt{0.001} \\ &= 0.0387 \end{aligned}$$

A 95% confidence interval for the regression slope is given by,

$$\begin{aligned} &\left(b_1 - t_{.05/2} \frac{s}{\sqrt{S_{xx}}},\ b_1 + t_{.05/2} \frac{s}{\sqrt{S_{xx}}} \right) \\ &= \left(-0.0015 - 2.086 \times \frac{0.0387}{\sqrt{88550}},\ -0.0015 - 2.086 \times \frac{0.0387}{\sqrt{88550}} \right) \\ &= \boxed{(-0.00172, -0.00118)} \end{aligned}$$

Step 6/9

(e)

A 95% prediction interval for a given value of *X* = *x* is

$$\begin{aligned} &b_0 + b_1 x \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}} \\ &= 0.375 - 0.0015x \pm 2.086 \times 0.0387 \times \sqrt{1 + \frac{1}{22} + \frac{(x - 105)^2}{88550}} \end{aligned}$$

A 95% prediction interval for the relative change of the population between years 2010 and 2020 is given by putting x = the above equation and we get,

$$\begin{aligned} &b_0 + b_1 x \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}} \\ &= 0.375 - 0.0015x \pm 2.086 \times 0.0387 \times \sqrt{1 + \frac{1}{22} + \frac{(x - 105)^2}{88550}} \\ &= 0.375 - 0.0015 \times 220 \pm 2.086 \times 0.0387 \times \sqrt{1 + \frac{1}{22} + \frac{(220 - 105)^2}{88550}} \\ &= (-0.025, 0.140) \end{aligned}$$

Step 7/9

A 95% prediction interval for the relative change of the population between years 2020 and 2030 is given by putting x = the above equation and we get,

$$b_0 + b_1 x \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$

$$= 0.375 - 0.0015x \pm 2.086 \times 0.0387 \times \sqrt{1 + \frac{1}{22} + \frac{(x-105)^2}{88550}}$$

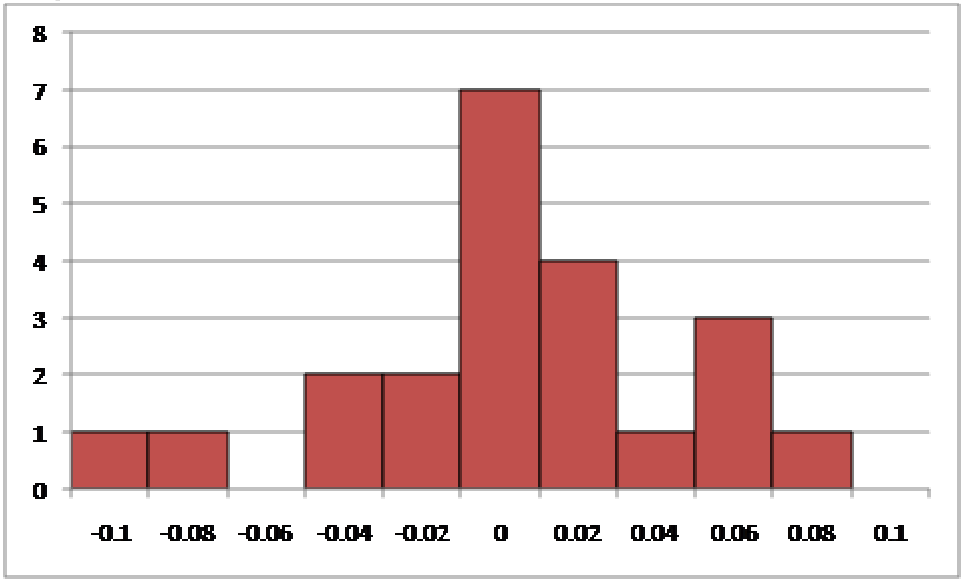
$$= 0.375 - 0.0015 \times 230 \pm 2.086 \times 0.0387 \times \sqrt{1 + \frac{1}{22} + \frac{(230-105)^2}{88550}}$$

$$= (-0.039, 0.126)$$

Step 8/9
(f)

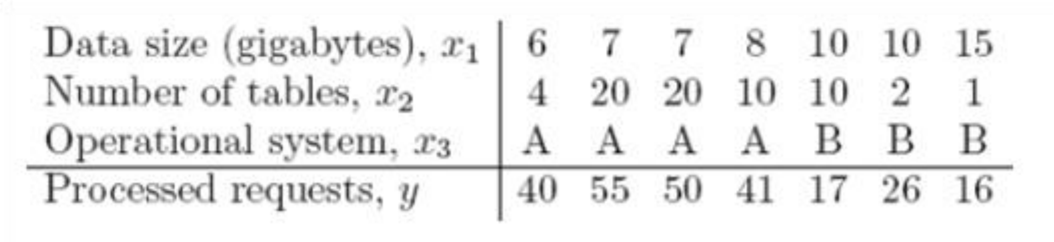
The histogram for the regression residuals is shown below :

Step 9/9



It shows a skewed distribution and violated assumption of normality.

Problem
Consider the program efficiency study in Examples 11.6–11.7 and 11.10–11.11 and data set Efficiency. The computer manager makes another attempt to improve the prediction power. This time she would like to consider the fact that the first four times the program worked under the operational system A and then switched to the operational system B.



Data size (gigabytes), x_1	6	7	7	8	10	10	15
Number of tables, x_2	4	20	20	10	10	2	1
Operational system, x_3	A	A	A	A	B	B	B
Processed requests, y	40	55	50	41	17	26	16

(a) Introduce a dummy variable responsible for the operational system and include it into the regression analysis. Estimate the new regression equation.

(b) Does the new variable improve the goodness of fit? What is the new R-square?

(c) Is the new variable significant?

(d) What is the final regression equation that you would recommend to the computer manager any time when she needs to predict the number of processed requests given the size of data sets, the number of tables, and the operational system? Select the best regression equation using different model selection criteria.

Step-by-step solution

Show all steps

Step 1/6

(a)

Create the dummy variable corresponding to the operational system A and B in the following way:

$$z_i = \begin{cases} 0 \text{ if the program worked under OS A} \\ 1 \text{ if the program worked under OS B} \end{cases}$$

Now the data of program efficiency is shown below:

x_1	x_2	z	y
6	4	0	40
7	20	0	55
7	20	0	50
8	10	0	41
10	10	1	17
10	2	1	26
15	1	1	16

Step 2/6

Fit the model by multivariate regression analysis.

The predictor matrix *X* and the response vector *Y* are shown below:

1	6	4	0

1	7	20	0
1	7	20	0
1	8	10	0
1	10	10	1
1	10	2	1
1	15	1	1

And $Y^T = \boxed{40 \quad 55 \quad 50 \quad 41 \quad 17 \quad 26 \quad 16}$

Step 3/6

We then compute

$$X^T X = \begin{pmatrix} 7 & 63 & 67 & 3 \\ 63 & 623 & 519 & 35 \\ 67 & 519 & 1021 & 13 \\ 3 & 35 & 13 & 3 \end{pmatrix}$$

and

$$X^T Y = \begin{pmatrix} 245 \\ 1973 \\ 2908 \\ 59 \end{pmatrix}$$

To obtain the estimated vector of slopes

$$b = \left(X^T X\right)^{-1}\left(X^T Y\right)$$
$$= \begin{pmatrix} 43 \\ -0.6 \\ 0.57 \\ -19 \end{pmatrix}$$

Thus the regression equation is $\boxed{y = 43 - 0.6x_1 + 0.57x_2 - 19z}$

Step 4/6

(b)

For this model we get

$$\hat{Y} = Xb$$

which is calculated as shown below:

$$\hat{Y}^T = \boxed{41.7 \quad 50.2 \quad 50.2 \quad 43.9 \quad 23.9 \quad 19.3 \quad 15.8}$$

Calculate for the new model:

$$\begin{aligned} SS_{REG} &= \left(\hat{y}-\overline{y}\right)^{T}\left(\hat{y}-\overline{y}\right) \\ &= \boxed{1325.67} \end{aligned}$$

$$\begin{aligned} SS_{ERR} &= \left(y-\hat{y}\right)^{T}\left(y-\hat{y}\right) \\ &= \boxed{126.33} \end{aligned}$$

$$df_{ERR} = 3$$

$$\begin{aligned} MSS_{ERR} &= 126.33/3 \\ &= \boxed{42.1} \end{aligned}$$

Whether the new variable improves the goodness of fit, apply F-test for this.

For the old model (before including the new variable)

$SS_{REG} = 1143.3$ (As given in example 11.10)

The extra sum of squares is computed as follows:

$$\begin{aligned} SS_{EX} &= SS_{REG}\left(\text{Full}\right) - SS_{REG}\left(\text{Reduced}\right) \\ &= 1325.67 - 1143.3 \\ &= 182.4 \end{aligned}$$

Compute *F*-Statistic:

$$\begin{aligned} F &= \frac{SS_{EX}/df_{EX}}{MSS_{ERR}} \\ &= \frac{182.4/1}{42.1} \\ &= \boxed{4.33} \end{aligned}$$

While $F_{1,3} = 10.1$.The calculated *F*-value is less than the theoretical *F* value at *df* (1,3) and 5% significance level. So there is no significant evidence that the new variable improves the goodness of fit.

$$\begin{aligned} R^{2} &= \frac{SS_{REG}}{SS_{TOT}} \\ &= \frac{1325.67}{1452} \\ &= \boxed{0.913} \end{aligned}$$

Step 5/6
(c)

According to the above result, the new variable is not significant.

Step 6/6
(d)

To select the best regression equation we will use the adjusted R^2 criterion adjusted R^2 for the reduced model is as shown below:

$$\begin{aligned} R_{adj}{}^{2}(reduced) &= 1 - \frac{MSS_{ERR}}{MSS_{TOT}} \\ &= 1 - \frac{77.2}{1452/6} \\ &= \boxed{0.681} \end{aligned}$$

The adjusted R^2 for the full model is as shown below:

$$\begin{aligned} R_{adj}{}^{2}(full) &= 1 - \frac{MSS_{ERR}}{MSS_{TOT}} \\ &= 1 - \frac{42.1}{1452/6} \\ &= \boxed{0.826} \end{aligned}$$

According to the adjusted R^2 criterion, inclusion of the new variable is beneficial to improve the predictive power.