

Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology-Kharagpur

Lecture-43
Tutorial on Relation Analysis

Hi guys, so today is the tutorial session, we have done a couple of lectures.

(Refer Slide Time: 00:33)



The screenshot shows a presentation slide with a dark blue header containing the text "Concepts Covered". Below the header, there is a list of topics:

- Solving objective type questions
 - To test the level of understanding about correlation and regression analysis
- Problems to ponder
 - To build problem solving aptitude

In the bottom right corner of the slide, there is a small video inset showing Prof. Monalisa Sarma. At the bottom of the slide, there are logos for IIT Kharagpur and NPTEL, along with the text "Monalisa Sarma" and "IIT Kharagpur".

So, we have learned like correlation analysis, different types of correlation analysis we have done regression analysis, so now it is time for tutorial. So, in this tutorial as like my previous tutorial first we will be doing few objective questions and then we will be solving a bit bigger problems. So, now coming to the objective type question, my suggestion is that please do not see the answer immediately.

(Refer Slide Time: 00:55)

Question-9.1

T 9.1: Which of the following statement(s) is(are) NOT true?

- a) Pearson's correlation analysis is applicable to only numeric data.
- b) Spearman's correlation analysis is applicable to only cardinal data.**
- c) χ^2 correlation analysis is applicable to only categorical data.
- d) Any non-parametric statistical learning approach is applicable when the entire population is known.

Manalisa Sarma
IIT KHARAGPUR

First try to think and then if you cannot then only you go to the answer, that is a very easy questions. First is which of the following statements are not true? So, Pearson correlation analysis is applicable to only numeric data, yeah, that is correct right, Pearson coordination analysis applicable to numeric data. Spearman correlation analysis applicable to only cardinal data, no, what is cardinal data?

Cardinal data is the numbers which we use to count numbers 2, 4, 5 those are where we used to count some numbers we call it as a cardinal number. Then Chi square test is applicable to categorical data, yeah of course chi square test is used for nominal data that is categorical data. And any non-parametric statistical learning approach is applicable when the entire population is known, very true and the entire population is known or we have a very big sample size unlike the parametric case.

So, here only one statement is wrong that is b, Spearman correlation is analysis is applicable to only cardinal data, Spearman correlation lines is applicable to what sort of data? It is data which where there is an intrinsic ranking among the data, so that is ordinal data. Spearman of course we can use for numeric data also like for ratio, scale data for interface scale data also you can use it but then better is that we use for the ordinal scale data. And the numeric data when we try using Spearman concept we lose some of the information basically.

(Refer Slide Time: 02:24)

Question-9.2

T 9.2: The value of correlation coefficient (r) lies between

- a) 0 to 1
- b) -1 to 1**
- c) $-\alpha$ to $+\alpha$
- d) 1 to 5

The slide features a background with a tree diagram and various icons. A small video inset in the bottom right corner shows a woman in a pink shirt. Logos for NPTEL and the presenter, Manalisa Sarma, are visible at the bottom.

So, next question the value of the correlation coefficient lies between, the values of the correlation coefficient lies between what? Correlation coefficient, so it lies between -1 to +1, is not it. So, because correlation there can be negative type of correlations, there can be positive correlation, is not it, so the value lies from -1 to +1.

(Refer Slide Time: 02:45)

Question-9.3

T 9.3: If there is a very strong correlation between two variables then the correlation coefficient must be

- a) any value larger than 1
- b) any value close to 0
- c) values closer to -1 or +1, depending upon whether the correlation is negative or positive.**
- d) none of the above

The slide features a background with a tree diagram and various icons. A small video inset in the bottom right corner shows a woman in a pink shirt. Logos for NPTEL and the presenter, Manalisa Sarma, are visible at the bottom.

If there is a strong correlation between 2 variables then the correlation coefficient must be strong correlation, when there is a strong correlation then it will be closer to either -1 or +1. If it is a negative correlation then it will be closer to -1, if it is a positive correlation then it will be closer to +1. So, for strong correlation when it is closer to -1 or +1, weak correlations closer to 0.

(Refer Slide Time: 03:12)

Question-9.4

T 9.4: To measure ranked variables the following correlation coefficient is used

- a) Pearson's
- b) Spearman's**
- c) Fisher's
- d) Chi-square

The slide features a background with a stylized tree and various icons. A small video inset in the bottom right shows a woman in a pink shirt. The footer includes the NPTEL logo and the name 'Monalisa Sarma'.

To measure rank variables the following correlation coefficient is used, when we have to find a correlation between the rank variables. So, rank variable is the ordinal data, ordinal data is a Spearman.

(Refer Slide Time: 03:27)

Question-9.5

T 9.5: If the sample data in a χ^2 test contains m rows and n columns, then the degree of freedom will be

- a) $m \times n$
- b) m
- c) $(m-1) \times (n-1)$**
- d) $(m \times n - 2)$

The slide features a background with a stylized tree and various icons. A small video inset in the bottom right shows a woman in a pink shirt. The footer includes the NPTEL logo and the name 'Monalisa Sarma'.

If the sample data in a Chi square test contains m rows and n columns, remember Chi square test we form a matrix where there are the independent variable we put it in the column and the dependent variable we put it in the row. And so it is we form a m cross n matrix, so then the degree of freedom what will be the degree of freedom? Degree of freedom is $m - 1$ cross $n - 1$. So, this is the degree of freedom.

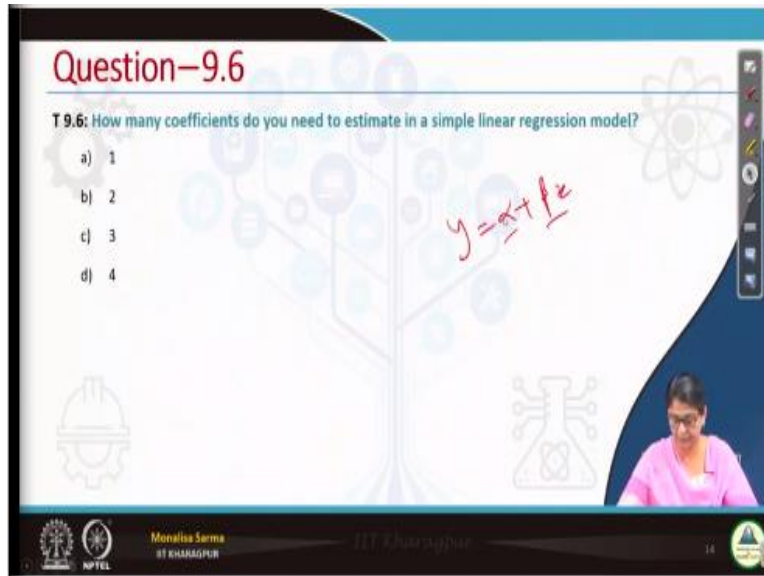
(Refer Slide Time: 04:02)

Question-9.6

T 9.6: How many coefficients do you need to estimate in a simple linear regression model?

- a) 1
- b) 2
- c) 3
- d) 4

$y = \alpha + \beta x$



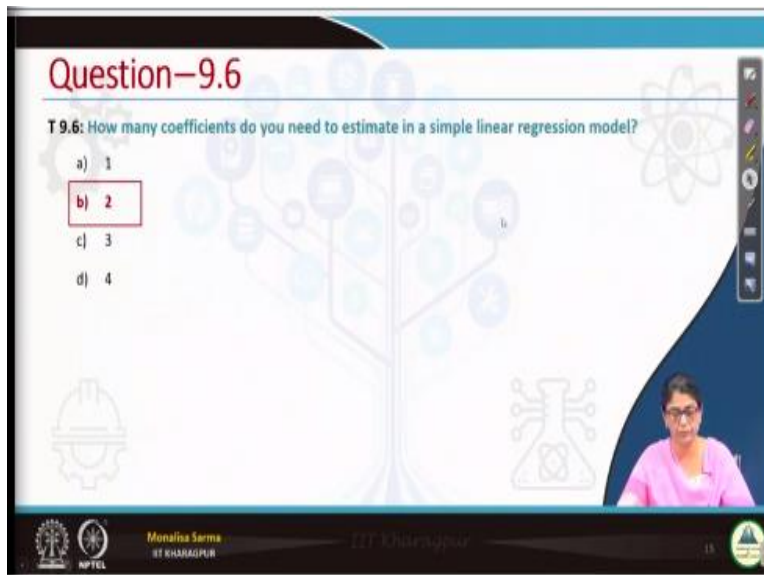
So, how many coefficients do you need to estimate an simple linear regression model? Simple linear regression model what is the expression for simple linear regression model? $y = \alpha + \beta x$, so what are the coefficient? Coefficient is α and β , for simple linear regression we need to estimate 2 coefficients.

(Refer Slide Time: 04:22)

Question-9.6

T 9.6: How many coefficients do you need to estimate in a simple linear regression model?

- a) 1
- b) 2
- c) 3
- d) 4



(Refer Slide Time: 04:24)

Question-9.7

T 9.7: The square of the correlation coefficient r , that is, r^2 will always be positive and is called

- a) regression
- b) coefficient of determination**
- c) KNN
- d) association

So, the square of the correlation coefficient r , that is r^2 will always be positive, square of the correlation coefficient r , we have done an r^2 test. Remember, what r^2 test the way we do, and what is r^2 called? R^2 is called coefficient of determination, is not it. So, now we will be doing some problems.

(Refer Slide Time: 04:53)

Problem-9.8

T 9.8: The owner of a shoe store wants to know if shoe size and weight are correlated in adult males. She measures the shoe size and asks the weight of 14 consecutive customers. The shoe sizes are given in the table. Find Pearson correlation in this data.

Size	9	7.5	10	12	9.5	10	10	10.5	13	8	8.5	9.5	9	11
Weight	176	141	185	202	174	150	193	237	248	159	136	174	172	183

So very easy one just simple if you know the technique, you know the formula, you will be able to do it, nothing much to understand here as such just you have to understand what this and some problem it is the direct straight forward, it is directly given use this and some you will have to find out what we have to use and accordingly you should know the formula for that. Now here the owner of a shoe store wants to know if the shoe size and the weight are correlated in adult males.

If the size of the shoes and weights if there is there any correlation between them the owner wants to know. So, she measures the shoe size and asks the weight of 14 consecutive customers, for the 14 consecutive customers she took the weight as well as the weight of the customers as well as the size of the shoe. The shoe sizes are given in the table; find Pearson correlation in this data.

So, it is just given you have to find the Pearson correlation. So, just if you can remember what is the formula for the Pearson correlation you do not have to do anything else just put it in a formula simple formula.

(Refer Slide Time: 05:55)

Problem-9.8 : Solution

The Karl Pearson's coefficient of correlation is defined as

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}$$

T 9.8: The owner of a shoe store wants to know if shoe size and weight are correlated in adult males. She measures the shoe size and asks the weight of 14 consecutive customers. The shoe sizes are given in the table. Find Pearson correlation in this data.

Monalisa Sarma
IIT KHARAGPUR

So, this is the Pearson correlation coefficient what is the formula? Covariance of X and Y divided by the standard deviation of X into standard deviation of Y, covariance of X and Y. When we write covariance of X and Y, this is the formula this numerator what we see and will be divided by the degrees of freedom. What is the degrees of freedom? Degrees of freedom is n - 1, similarly standard deviation of X into standard deviation of Y.

So, standard deviation is this again under $\sqrt{\quad}$ I will be having n - 1 and here also I have $\sqrt{n - 1}$, here also I will have $\sqrt{n - 1}$, here I just have add n - 1. So, $\sqrt{n - 1}$ into n - 1 is n - 1, this n - 1 and n - 1 gets cut, so what remains is this. So, this is the Carl Pearson coefficient is just this is the formula. If you cannot remember this formula very easy just remember covariance of X and Y divided by the

standard deviation of X into standard deviation of Y. So, now we what we need? We need X_i , from the table will be getting X_i value and Y_i value.

(Refer Slide Time: 07:05)

Problem-9.8 : Solution

The Karl Pearson's coefficient of correlation is defined as

$$r^2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \sqrt{\sum(Y_i - \bar{Y})^2}}$$

For given data,

$$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = 504.78$$

$$\sqrt{\sum(X_i - \bar{X})^2} = \sqrt{28.00} = 5.366$$

$$\sqrt{\sum(Y_i - \bar{Y})^2} = \sqrt{13502.86} = 116.20$$

$$r^2 = \frac{504.78}{5.366 \times 116.20} = 0.809$$

Size	Weight	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
9	176	-0.82143	-4.7143	0.67	22.22	3.87
7.5	141	-2.32143	-39.7143	5.39	1577.23	92.19
10	185	0.178571	4.2857	0.032	18.37	0.76
12	202	2.178571	21.2857	4.77	453.08	46.37
9.5	174	-0.32143	-6.7143	0.10317	45.08	2.15
10	150	0.178571	-30.7143	0.031888	943.37	-5.48
10	193	0.178571	12.2857	0.031888	150.94	2.19
10.5	237	0.678571	56.2857	0.460459	3168.08	38.14
13	248	3.178571	67.2857	10.10331	4527.37	213.87
8	138	-1.82143	-21.7143	3.317604	471.51	39.55
8.5	136	-1.32143	-44.7143	1.746175	1999.37	59.09
9.5	174	-0.32143	-6.7143	0.10317	45.08	2.16
9	172	-0.82143	-8.7143	0.674746	75.94	7.16
11	183	1.178571	2.2857	1.38903	5.22	2.69

T 9.8: The owner of a shoe store wants to know if shoe size and weight are correlated in adult males. She measures the shoe size and asks the weight of 14 consecutive customers. The shoe sizes are given in the table. Find Pearson correlation in this data.

Monalisa Serna
IIT KHARAGPUR

And also we need \bar{X} and \bar{Y} , that is the mean of X and mean of Y, just put it in the formula you will get the r^2 value. So, this is the r^2 value, r value are this correlation coefficient that is 0.809. Now see here, from this value what can you tell whether the correlation is a strong correlation or weak correlation? This value is closer to 1, so that means it is a strong correlation.

Now again do you also remember one thing like this is the value, this correlation coefficient we got it, okay, fine but again we have to do the significance test also whether what this correlation coefficient what we got is from a sample. This correlation coefficient what we got from the sample, is it also applicable to the population or it is just by coincidence we got that.

We have taken a random sample because samples are random; we have taken a random sample. In the random sample we got this correlation coefficient but actually it is not so in the case of a population. So, whether it is so whether this correlation coefficient what we got in the sample is it also applicable to the population, so we do the significance test, if you can remember that.

Significant test means we have find out the null hypothesis, we find that the alternate hypothesis but as a null hypothesis there is no correlation coefficient and alternate hypothesis is there is a

correlation coefficient. Now here for Carl Pearson we use the t statistics, so we compute the value of the t and for a particular significance level we see if the value falls in the critical regions then we reject the null hypothesis, remember.

So, anyway here we are not doing it just asking for the correlation coefficient, so with this, this is our answer, that is all. If it has asked to find out whether this is, see it passed the significance test or not then you will have to do that.

(Refer Slide Time: 08:56)

Problem-9.9

T 9.9: The grades for 10 students on Class 9 exam and class 10 examinations in the English subject are given in the table below. Calculate the rank correlation coefficient.

Class 9	84	98	91	72	86	93	80	0	92	87
Class 10	73	63	87	66	78	78	91	0	88	77

NPTEL
Monalisa Sarma
IIT Madras

Now the grades of 10th students on class 9 exam and class 10 examination in the English subjects are given in the table below, calculate the rank correlation coefficient. Which is the rank correlation coefficient? Rank correlation means Spearman, Spearman we do it rank wise. So, different student's marks are given 10th student marks are given marks for class 9 exam and for class 10 exam, same set of students.

Same set of students they are given marks for class 9 and class 10. So, now we have to find out the rank correlation coefficient. So, actually this question looks sounds a bit weird, the same set of students for class 9 exams and class 10 exams. Take 2 different types of exam or the same subject instead of 9 and 10 they say for maths and science that way.

(Refer Slide Time: 09:51)

Problem-9.9 : Solution

From the formula of Rank correlation coefficient

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

T 9.9: The grades for 10 students on Class 9 exam and class 10 examinations in the English subject are given in Table below. Calculate the rank correlation coefficient.

Original Marks		Assigned Ranks		Difference between ranks (d_i)	d_i^2
Class 9	Class 10	Class 9 ranks	Class 10 ranks		
84	73	7	7	0	0
98	63	1	9	-8	64
91	87	4	3	1	1
72	66	9	8	1	1
86	78	6	4.5	1.5	2.25
93	78	2	4.5	-2.5	6.25
80	91	8	1	7	49
0	0	10	10	0	0
92	88	3	2	1	1
87	77	5	6	-1	1

Monalisa Sarma
IIT KHARAGPUR

Then how to find out the rank correlation coefficient? Rank correlation coefficient is this is the formula. So, how do I find that d_i^2 ? d_i is the difference in the rank of 2 numbers, how do we give the ranking? Ranking highest number gets the rank 1 and accordingly we come down, is not it. So, first we will have to find out which is the highest number and accordingly we give the rank 1, then rank 2, rank 3 and gradually.

And then we will find out the difference of the rank of the 2 variables what we are trying to correlate. Is there any correlation between the class 9 marks and the class 10 mark? So, we will find out the difference of the ranks. So, once we find out that difference of the rank is d_i and what is n ? n is the total sample size, so that is how just putting it in the formula we will get the r_s value. Now again for the significance days what I will do?

Here we will be using this Spearman correlation graph is there and for a significance level that I remember I told you in the class last in the one of the lecture when we have discussing Spearman correlation. It is available in the standard textbooks, the graph for Spearman correlation. So, from the graph if you see if it falls into a rejection region, then there is no correlation between the 2 variables.

So, see here, say the class 9 this is 98 is the highest mark I have given the rank 1, after 98 it is 93 I have given rank 2. Similarly here 91 is the highest marks I have given rank 1, so likewise and

then we find d_i , then I find d_i^2 . Then accordingly put it in this formula and get the value of r_s . Here it starts to find out the rank correlation coefficient, so we will not do the significance test if it asks. So, find out the significances then accordingly left to it.

(Refer Slide Time: 11:38)

Problem-9.9 : Solution

From the formula of Rank correlation coefficient

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

$$= 1 - \frac{6 \times 125.5}{10 \times 90}$$

$$= 1 - \frac{753}{990}$$

$$= 0.24$$

Original Marks		Assigned Ranks		Difference between ranks (d_i)	d_i^2
Class 9	Class 10	Class 9 ranks	Class 10 ranks		
84	73	7	7	0	0
98	61	1	9	-8	64
91	87	4	3	1	1
72	66	9	8	-1	1
86	78	6	4.5	1.5	2.25
93	78	2	4.5	-2.5	6.25
80	91	8	1	7	49
0	0	10	10	0	0
92	88	3	2	1	1
87	77	5	6	-1	1

T 9.9: The grades for 10 students on Class 9 exam and class 10 examinations in the English subject are given in Table below. Calculate the rank correlation coefficient.

Monalisa Sarma
IIT KHARAGPUR

(Refer Slide Time: 11:40)

Problem-9.10

T 9.10: The owner of a company provides pneumococcal pneumonia vaccine in order to keep the sick leave count as low as possible. Due to vaccine shortages, at a certain point of time, only some of the employees received vaccine and the others did not. Then the company kept track of the number of employees who contracted pneumonia and which type of pneumonia each had. The data were organized as given below. The company wanted to know if providing the vaccine made a difference.

Health Outcome	Unvaccinated	Vaccinated
Sick with pneumococcal pneumonia	24	5
Sick with non-pneumococcal pneumonia	8	10
No pneumonia	61	76

Monalisa Sarma
IIT KHARAGPUR

Say here the correlation coefficient 0.24 it is a weak correlation basically. So, this is a very interesting question, see here. The owner of a company provides pneumococcal, pneumococcal is a type of pneumonia basically, pneumonia vaccine in order to keep the sick leave count as low as possible. Due to vaccine shortage at a certain point of time only some of the employees using vaccine and others did not.

Find as a vaccine like provide a vaccine sort of initially, so all did not get vaccine few got and few did not get. Then the company kept track of the number of employees who contracted pneumonia and which type of pneumonia each had. There are different types of pneumonia the vaccine is for pneumococcal pneumonia but there are different types of pneumonia. So, company is trying to keep track of the people who are contracting pneumonia and for what type?

Are they contracting the type for which the vaccine they have given the vaccine or they are contracting some other type of pneumonia. The data were organized as given below. So, the company wanted to know if providing the vaccine made a difference, what the company wanted to know? That providing a vaccine does it made a difference? That means by taking the vaccine does it have any effect on the occurrence of the pneumococcal pneumonia, does it have?

By taking vaccine that means we are having less number of pneumococcal pneumonia or we are having same having vaccine or not having vaccine we are having the same number of pneumococcal pneumonia? That is what we need to find out. So, this is very much equation of Chi square distribution because these are the different categories and we have to find out the correlation of these 2 categories, what are the 2 categories here?

You have to find out the different categories are the for the independent variable if you see the different there is the vaccination status that is the unvaccinated and a vaccinated, does this vaccination status have any effect on the health outcome? Health outcome we have 3 different categories, fine, one is sick with pneumococcal pneumonia, sick with non pneumococcal pneumonia and no pneumonia.

So, these are the 3 health in 3 different categories for the dependent variable and for the independent variable there are 2 different categories. So, this is very much equation of Chi square distribution that means we will have to find a Chi square correlation. So, we will have to find out the Chi square value of this whole table and then we will be able to tell it whether this providing the vaccine made a difference or it does not make a difference?

So, if that Chi square value that we get if it falls in the rejection region then providing a vaccine really makes a difference if it falls into acceptance regions that means providing vaccine does not make any difference. Basically there is no correlation between people contracting pneumococcal pneumonia with whether they are vaccinated or not vaccinated; there is no correlation between these 2.

So, first thing is that we will have to find out the Chi square value. Remember in this table I have already told you in the class the first thing is will fill in the table with observed value, these are the observed value 24, 5. 24 people caught pneumococcal pneumonia which are unvaccinated, 5 people caught pneumococcal pneumonia were vaccinated, similarly 8 non-pneumococcal, 10 here for vaccinated and this is no pneumonia 61 and no pneumonia for a vaccination is 76.

So, we need to find out the Chi-square value. So, for finding out the Chi-square value what is the first step? The first step is we will have to find out the row total and the column total if you can remember. If you do not remember I just say please go through the lecture once again.

(Refer Slide Time: 15:41)

Problem-9.10 : Solution

Calculation of marginals:

Health Outcome	Unvaccinated Col 1	Vaccinated Col 2	Row marginal (Row sum)
Sick with pneumococcal pneumonia	24	5	29
Sick with non-pneumococcal pneumonia	8	10	18
No pneumonia	61	76	137
Column marginal (Sum of the column)	93	91	N=184

T 9.10: The owner of a company provides pneumococcal pneumonia vaccine in order to keep the sick leave count as low as possible. Due to vaccine shortages, at a certain point of time, only some of the employees received vaccine and the others did not. Then, the company kept track of the number of employees who contracted pneumonia and which type of pneumonia each had.

So, first thing is that we will be finding the row total and the column total. This row total and the column total we call it a marginal, so row marginal and column marginal, so this 2 put together is 29, this is the row margin of this row and then the row marginal of this row is 18, row marginal of

this row is 137. And similarly the column marginal of this is 93, column marginal of this is 91 this is sum and the total data is 184.

So, now first in the table we have the observed data and then from our task is to find out the Chi square statistics. To finding out the Chi square statistics first thing what we will have to do? Our first step is calculating the row marginal and column marginal. Then after we calculate the row marginal and column marginal then next step is to find out the expected value. Why we will find out the expected value? Expected value how we will find?

Well, the expected value is that value which will get considering that there is no effect on the vaccine. So, how do we find out the expected value? If the expected value is something meaning that there is no effect on the vaccine, vaccine or no vaccine what value we are expecting that people will be contacting a particular type of pneumonia or people will not be contacting.

(Refer Slide Time: 17:00)

Problem-9.10 : Solution

Calculation of Expected values (E):

The formula for computing the Chi-square expected values

$$E = \frac{M_r \times M_c}{n}$$

E = represents the cell expected value,
 M_r = represents the row marginal for that cell,
 M_c = represents the column marginal for that cell,
n = represents the total sample size

The cell expected values are shown in the table.

T 9.10: The owner of a company provides pneumococcal pneumonia vaccine in order to keep the sick leave count as low as possible. Due to vaccine shortages, at a certain point in time, only some of the employees received vaccine and the others did not. Then the company kept track of the number of employees who contracted pneumonia and which type of pneumonia each had.

Health Outcome	Unvaccinated Col 1	Vaccinated Col 2
Sick with pneumococcal pneumonia	$\frac{23 \times 93}{184} = 11.66$	14.34
Sick with non-pneumococcal pneumonia	9.10	8.90
No pneumonia	69.24	67.76

So, to find out the expected value for each cell this is the formula, row marginal into column marginal divided by the total number of values. So, now here for this case, this is 1 cell, this is the second cell, this is the third, this is the fourth, this is the fifth, this is the sixth cell, fine. So, now to find out the expected value for this cell how do I find out the expected value for this?

Row marginal into column marginal divided by the total number of values, you see what is the row marginal here 29 and column marginal here 93, 29 into 93 divided by 184, that is what we got it 29 into 93 divided by 184 that is how we got the expected value. Similarly we will have to calculate expected value for each column, let us do and see because.

(Refer Slide Time: 17:51)

	NV	V	RH	
1	24	5	29	29
2	8	10	18	18
3	61	76	137	137
CH	93	91	184	

Handwritten notes on the table include: $\frac{29}{93}$ above the first column, $\frac{5}{91}$ above the second column, $\frac{29}{184}$ above the first row, and $\frac{(O-E)^2}{E}$ on the right side.

See here this is the vaccination status that is I think this is the first one is non-vaccinated and second was is vaccinated and this side we have row we have the health status, another one is first let me just call it as a first category, second category and third category. First category is sick with that is pneumococcal pneumonia and second is sick with non pneumococcal pneumonia and third with it stayed healthy, that means they will did not have any pneumonia.

So, for that what was our observed value? Initial I will make the whole table that will be better actually. What was the observed value? Our observed value was initially this was 24 and this was 5, this was 8 and this was 10, this was 61 and this was 76 this is our observed value. First I will write the observed value, what is the observed value? 24 and then we have here 8, we have here 61 and here we have 5, then 10, then we have 76.

So, this total is 93 and this total is 91 and what we have this total here is 29 and here is how much it is 18 and here how much it is 137 and total we have is 184. So, this is the observed value, this is the row marginal and this is the column marginal. So, now once we get row marginal and column

marginal, next we will find out the expected value, this is the observed value, next we will find out the expected value.

So, what is the expected value for this? Expected value for this is 29 into 93 divided by 184, 184 is the 4 total data. This if you calculate it you will get this 14.66 I am just putting the data you can calculate it afterwards, this we will get 9.10, this we will get 69.24 for this 5 you will get observed value is 14.34 this will get 8.90, this will get 67.76, fine, 67.76 this is the expected value.

Now once we calculate the expected value our next step is to calculate the Chi square value, how do you find out the Chi square value? $(O - E)^2$ by E, so this is the formula for the Chi square value $O - E^2$ by E, O is the observed value, E is the expected value. So, when you can calculate the X^2 value we get a X^2 value is 5.95 for this, this is the X^2 5.95 this we got 6.08, this we get 0.13, this we got 0.14, 0.98 and 1.00. So, these are the X^2 value.

So, now we have to find out the X^2 status, the X^2 status means sum of all this X^2 value. If we do the sum of all this X^2 value, so what we will get? Let us come back I will come back just to this table again. So, see here.

(Refer Slide Time: 22:03)

Problem-9.10 : Solution

Cell Chi-square Values:
 The cell χ^2 values are computed with this formula

$$\chi^2 = \frac{(O - E)^2}{E}$$

Degree of freedom in this case $(3 - 1)(2 - 1) = 2$

The cell χ^2 values are provided in below table. Summing all the entries, we obtain the χ^2 statistics which is equal to 14.28.

Health Outcome	Unvaccinated Col 1	Vaccinated Col 2
Sick with pneumococcal pneumonia	$\frac{(24 - 14.66)^2}{14.66} = 5.95$	$\frac{(5 - 14.34)^2}{14.34} = 6.08$
Sick with non-pneumococcal pneumonia	$\frac{(0 - 9.10)^2}{9.10} = 0.13$	$\frac{(10 - 8.9)^2}{8.9} = 0.14$
No pneumonia	$\frac{(61 - 69.24)^2}{69.24} = 0.98$	$\frac{(76 - 67.76)^2}{67.76} = 1.00$

T 9.10: The owner of a company provides pneumococcal pneumonia vaccine in order to keep the sick leave count as low as possible. Due to vaccine shortages, at a certain point of time, only some of the employees received vaccine and the others did not. Thus, the company kept track of the number of employees who contracted pneumonia and which type of pneumonia each had.

Mamata Serna
 IIT BHARATPUR

So, this is how we calculate the X^2 value. Once we get the X^2 value these are the X^2 values given here. So, the sum of all this X^2 value is 14.28, we will have to find out the sum of all X^2 squared

that is my total χ^2 value. So, total χ^2 value I got is 14.28, now I will have to see it in the whether it is 14.28 values is a significant value or not, for that we will have to consult the χ^2 table.

We will consult the χ^2 table what is the degrees of freedom for that? How many rows and how many columns? Let us say $m - 1$ into $n - 1$. So, $3 - 1$ into $2 - 1$, so the degree of freedom in this case is 2. So, for 2 degrees of freedom for value 14.28 if you see the χ^2 table what you will see? (Refer Slide Time: 22:50)

Problem-9.10 : Solution

Cell Chi-square Values:

- The cell χ^2 values are computed with this formula
$$\chi^2 = \frac{(O - E)^2}{E}$$
- Degree of freedom in this case $(3 - 1)(2 - 1) = 2$
- The cell χ^2 values are provided in the table. Summing all the entries, we obtain the χ^2 statistics which is equal to 14.28.
- Using a χ^2 table, the significance of a Chi-square value of 14.28 with 2 degree of freedom, equals $P < 0.001$.
- As the P-value of the table is less than $P < 0.05$, the researcher rejects the null hypothesis and accepts the alternate hypothesis: "There is a difference in occurrence of pneumococcal pneumonia between the vaccinated and unvaccinated groups."

T 9.10: The owner of a company provides pneumococcal pneumonia vaccine in order to keep the sick leave count as low as possible. Due to vaccine shortage, at a certain point of time, only some of the employees received vaccine and the others did not. Then the company kept track of the number of employees who contracted pneumonia and which type of pneumonia each had.

Manalika Sarma
IIT Kharagpur

The P value, the probability for that is basically less than **0.0001** 0.001 if you see the χ^2 , I am not showing the table here, I have showed it many, many times and I am sure by now all of you are in comfortable with that. So, if you consult the χ^2 table you will see for this 14.28 with 2 degrees of freedom it has a P value of 0.001, very less. So, and significance value usually we consider around 0.05, so 0.05 and this is less than that 0.001.

That means if we draw something here this is my significance level, if it falls in this area then I reject the null hypothesis and my value is falling here much lesser than the significance value. So, that means I reject the null hypothesis, that means the researcher rejects the null hypothesis and accept the alternate hypothesis, there is difference in occurrence of pneumococcal pneumonia between the vaccinated and unvaccinated group.

That means there is the vaccination, what was my earlier question, this vaccination does it have any effect on the occurrence of pneumococcal pneumonia? Yes, vaccination has an effect on the pneumococcal pneumonia in the people. Because null hypothesis does not have effect, alternative hypothesis it has an effect. Now this story does not end here, but that is one of the beauty of Chi square distribution I should say actually.

So, here from this okay we found out that, we rejected null hypothesis but we got a X^2 value by taking the sum of which cell? Actually which cell has contributed to this rejection? We can find it out, in earlier test there is no specific which data has contributed to this, now we can find out which cell has actually contributed to the rejection of the null hypothesis.

Now if we go back to our table, see whichever column we get a X^2 value high very we got a X^2 value > 1 is in this and this, less dollar this is very much less than 1, very much less than 1, this is less than 1 and this is just 1. For this 2 value we got a X^2 value which is very much bigger than 1. What does the highest value of X^2 is 6.08, why did I get a highest value of such a highest value?

Because you see, here my expected value considering that the vaccine has no effect, my expected value was 14.34. I expected if vaccine did not have any effect then I expected that the people will have this type of pneumonia around 14.3 but how many people had? Only 5 people had, so by chance this cannot be a random event, there has to be is a significant value, where expected is 14.34 and people who got is very 5.

That means vaccine had an effect that is why we got such a less value, whereas if the vaccine did not have an effect the expected value is 14.34. Similarly next highest value is here X^2 value 5.95. Here see our expected value is 14.66 and how many observed value is 24, this also it cannot occur by chance. So, if the vaccine did not have any effect the expected value is we are expecting people around 14.66 people will be having the pneumonia that particular type of pneumonia before which the vaccine has been developed.

But the number of people who actually had that pneumonia is 24, so this cannot be a value by chance, we cannot just consider it a random event. So, that means the vaccine has an effect that is

why people who has taken vaccine, very less people got the disease whereas expected is so high. And similarly here because vaccine have an effect say these people since they did not take any vaccine, so though the expected value was 14 but 24 people had the disease.

And rest all the values it is very less, lesser than 1, just 1 value is just 1, so if this other cells has did not contribute to draw a decision, the only factor which contributed to our decision is this cell only the cell 1 and cell 2. So, interesting, right, this is how we find a Chi square distribution. We can not only be able to tell whether there are some sort of correlation or not, we can also specifically point out this presence of this correlation is for which category of data basically. So, this is by correlation using Chi square distribution.

(Refer Slide Time: 27:31)

Problem-9.11

T 9.11: The thrust of an engine (y) is a function of exhaust temperature (x) in °F when other important variables are held constant. Consider the following data.

Fit a simple linear regression for the data and find the equation of the regression line.

x	y
1760	4300
1652	4650
1485	3200
1390	3150
1820	4950
1665	4010
1550	3810
1700	4500
1270	3008

$y = \alpha + \beta x$

α β

So, quickly we will go to the other problem. Now this the thrust of an engine is a function of exhaust temperature Y is a function of X basically in degree Fahrenheit, when other important variables are held constant, consider the following data. Fit a simple linear regression for the data and find the equation of the regression line. So, we have just asked to fit a simple linear regression for the data and find the equation of the regression line.

So, fitting the data to a simple linear regression that we do not have to do anything, just we know simple linear equation expression is $y = \alpha + \beta x$. So, we know what is α , we know what is β , we have already calculated using the least square estimation method, is not it.

(Refer Slide Time: 28:12)

Problem-9.11 : Solution

x	y
1760	4300
1652	4650
1485	3200
1390	3150
1820	4950
1665	4010
1550	3810
1700	4500
1270	3008

SCATTER PLOT OF THE DATASET

T 9.11: The thrust of an engine (y) is a function of exhaust temperature (x) in $^{\circ}\text{F}$ when other important variables are held constant. Consider the following data. Fit a single linear regression for the data and find the equation of the regression line.

Monalisa Sarma
IIT KHARAGPUR

So, let us set the data we can just use the scatter plot just to see.

(Refer Slide Time: 28:16)

Problem-9.11 : Solution

From the formula of the simple linear regression line coefficients, we know,

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

X	y	x - \bar{x}	y - \bar{y}	(x - \bar{x})(y - \bar{y})	(x - \bar{x}) ²
1760	4300	172	346.89	59665.08	29584
1652	4650	64	696.89	44600.96	4096
1485	3200	-103	-753.11	77570.33	10609
1390	3150	-198	-803.11	159015.8	39204
1820	4950	232	996.89	231278.5	53824
1665	4010	77	56.89	4380.53	5929
1550	3810	-38	-143.11	5438.18	1444
1700	4500	112	546.89	61251.68	12544
1270	3008	-318	-945.11	300545	101124
$\bar{x} = 1588$	$\bar{y} = 3953$			$\sum (x - \bar{x})(y - \bar{y}) = 943746$	$\sum (x - \bar{x})^2 = 258358$

T 9.11: The thrust of an engine (y) is a function of exhaust temperature (x) in $^{\circ}\text{F}$ when other important variables are held constant. Consider the following data. Fit a single linear regression for the data and find the equation of the regression line.

Monalisa Sarma
IIT KHARAGPUR

And what α and β that value which we have assumed, for β we have assumed b , for α we have assumed a , these are the values we have calculated using least square estimation method. We have done it I think in the last class only, we have last to last lecture maybe. So, this is the parameter b , this is the parameter a , so we have here what we need? We need x_i , \bar{x} , y_i and \bar{y} , from the data we can find out x_i , \bar{x} , y_i , \bar{y} is nothing so great. So, we found out the data, we found out the a value, we found out the b value.

(Refer Slide Time: 28:46)

Problem-9.11 : Solution

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{943716}{250350} = 3.653$$

$$a = \bar{y} - b\bar{x}$$

$$= 3953 - 3.653 \times 1588$$

$$= -1847.964$$

So the equation of the regression line is

$$\hat{y} = -1847.964 + 3.653x$$

T 9.11: The thrust of an engine (y) is a function of exhaust temperature (x), in °F, when other important variables are held constant. Consider the following data. Fit a single linear regression for the data and find the equation of the regression line.

SCATTER PLOT OF THE DATASET

$y = 3.6529x - 1847.96$

Monalisa Sarma
IIT KHARAGPUR

Once we found out a value and the b below our equation is nothing but this whatever this is c, our this is our a value, a value + b into x, this is my expression. So, this is my linear equation regression line, cool. Just we need to remember what is b, what is a, if you do not want to do the whole process of least square estimation then the minimizing the least square estimation how do you minimize it by doing first order derivation, partial half first order derivation with respect to each parameter 1.

Then if you equalizing it to 0 and then finding out the value instead of if you want to do it, do it great, if you do not want to do it just still have to remember this formula. From there only we got this a and b value, there is no rocket science here.

(Refer Slide Time: 29:36)

Problem-9.12

T 9.12: In a particular village, the production of rice is primarily dependent on the rain-water only. In a survey, the amount of average daily rainfall and the total amount of rice production is recorded for ten years. In one particular year, 4.8 mm rainfall is recorded. Can you predict what would be the rice produced?

X (mm)	Y (Ton)
4.3	126
4.5	121
5.9	116
5.6	118
6.1	114
5.2	118
3.8	132
2.1	141
7.5	108

Monalisa Sarma
IIT KHARAGPUR

So, next question in a particular village the production of rice is primarily dependent on the rain water only. In a survey the amount of average daily rainfall and the total amount of rice production is written, recorded. So, this is the rainfall and is the rice production. In one particular year 4.8 mm rainfall is recorded. In a particular year 4.8 mm rainfall is recorded, can you predict what would be the rice produced?

Simple question of regression analysis is not it. Given the value of the independent variable we will have to find out the value of the response. So, how we will do that? For that we will have to need to know the function, that means we will have to find out the regression line that is $y = \alpha + \beta x$. That means we need to know the value of the parameter, that is α and β , how we will know the value of a parameter?

That the way we have done in the last example, same way find out the value of the parameter. Once we know the value of the parameter we got the equations, once we got the equation in the equation put the value of the x and you will get the value of y, cool.

(Refer Slide Time: 30:39)

Problem-9.12 : Solution

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{-121.8}{19.26} = -6.324$$

$$a = \bar{y} - b\bar{x}$$

$$= 121.5556 + 6.324 \times 5$$

$$= 153.18$$

So, the equation of the regression line becomes, $\hat{y} = 153.18 - 6.324x$

For $x = 4.8$, the predicted rice output = $153.18 - 6.324 \times 4.8 = 123 \text{ ton}$

x	y	x - \bar{x}	y - \bar{y}	(x - \bar{x})(y - \bar{y})	(x - \bar{x}) ²
4.3	126	-0.7	4.4444	-3.11108	0.49
4.5	121	-0.5	-0.5556	0.2778	0.25
5.9	116	0.9	-5.5556	-5.00004	0.81
5.6	118	0.6	-5.5556	-3.33336	0.36
6.1	114	1.1	-7.5556	-8.31116	1.21
5.2	118	0.2	-5.5556	-1.11112	0.04
3.8	132	-1.2	10.4444	-12.53328	1.44
2.1	141	-2.9	19.4444	-56.38876	8.41
7.5	108	2.5	-13.5556	-33.889	6.25
$\sum (x - \bar{x})(y - \bar{y}) = -121.8$				$\sum (x - \bar{x})^2 = 19.26$	

T 9.12 In a particular village, the production of rice is primarily dependent on the rain water only. In a survey, the amount of average daily rainfall and the total amount of rice production is recorded for ten years. In one particular year, 4.8 mm rainfall is recorded. Can you predict what would be the rice produced?

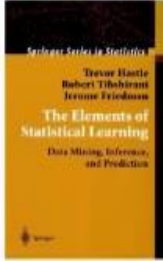
Monalisa Serna
11 KHAMRUI/PLN

So, this is b, this is a, we calculated b and a, so this is our equation, expression this is the linear regression line. Now here given our the x value is 4.8 mm, so put 4.8 mm in x then what value you get for y you get 123 ton.

(Refer Slide Time: 31:03)

REFERENCES

- ① The detail material related to this lecture can be found in
The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.



The image shows a slide titled 'REFERENCES' with a blue header. The main content area is white. On the right side, there is a vertical toolbar with various icons. At the bottom, there is a footer with logos for NPTEL and IIT Kharagpur, and the name 'Monalisa Sarma'.

So, that is all, it is a very simple tutorial, is not it, so this is the reference and thank you guys.