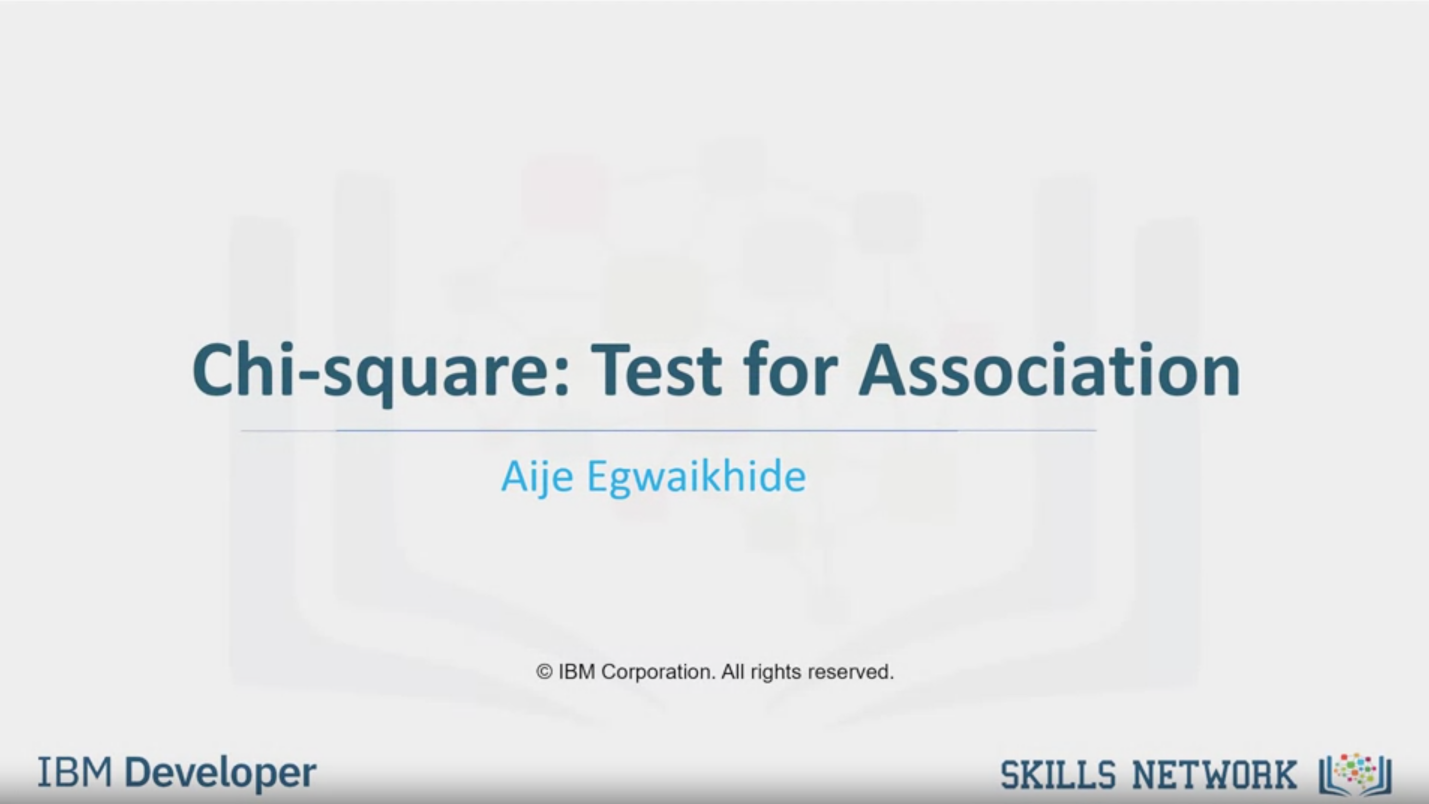


IBM Developer

SKILLS NETWORK



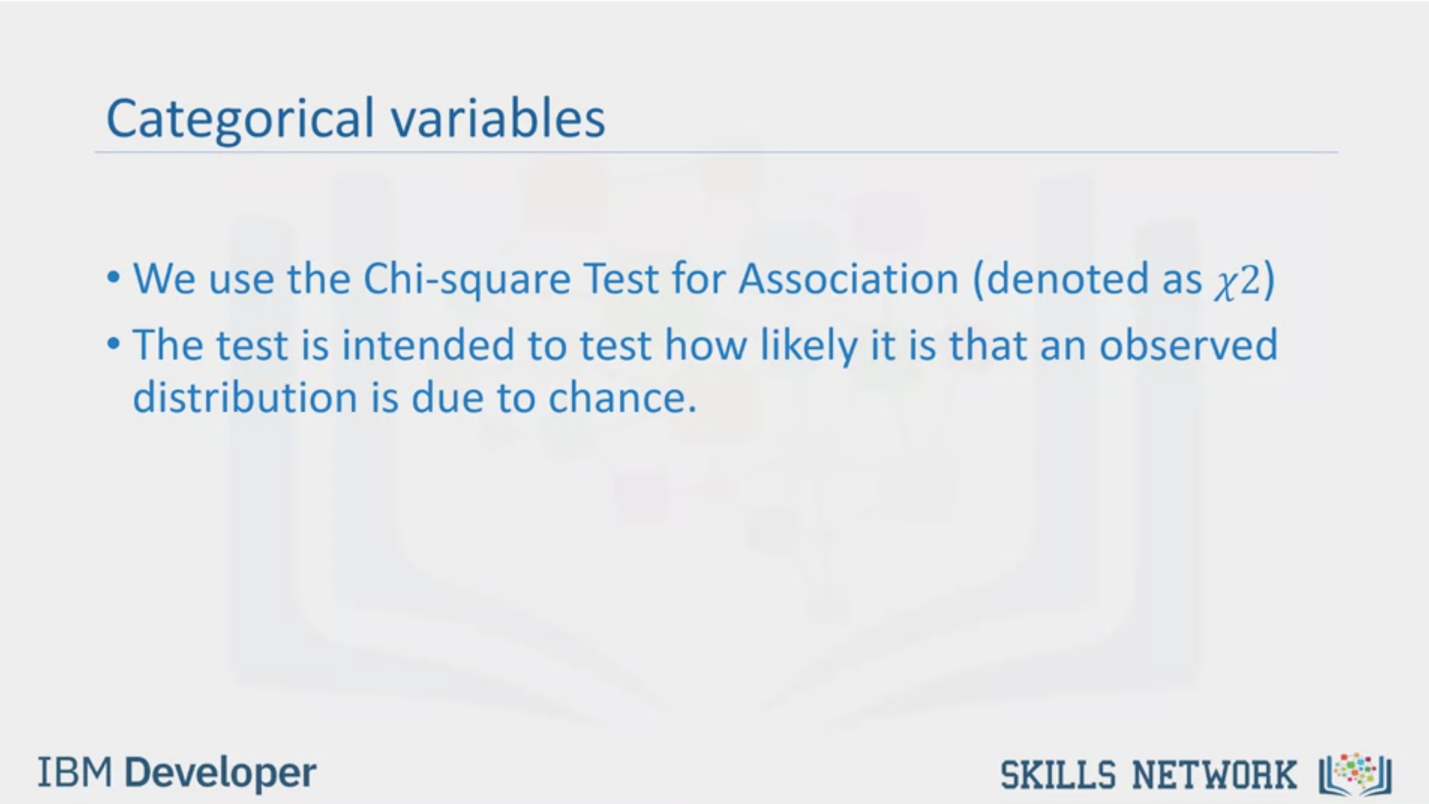
Chi-square: Test for Association

Aije Egwaikhide

© IBM Corporation. All rights reserved.

IBM Developer

SKILLS NETWORK



Categorical variables

• We use the Chi-square Test for Association (denoted as χ2)

• The test is intended to test how likely it is that an observed

distribution is due to chance.

IBM Developer

SKILLS NETWORK



Chi-Square Test for association

• The Chi-square tests a null hypothesis that the variables are

independent.

• The Chi-square does not tell you the type of relationship that exists

between both variables; but only that a relationship exists.

IBM Developer

SKILLS NETWORK



Categorical variables

• Is there an association between fuel-type and aspiration?

Observed value

Standard

Turbo

Total

diesel

gas

Total

7

13

20

161

24

185

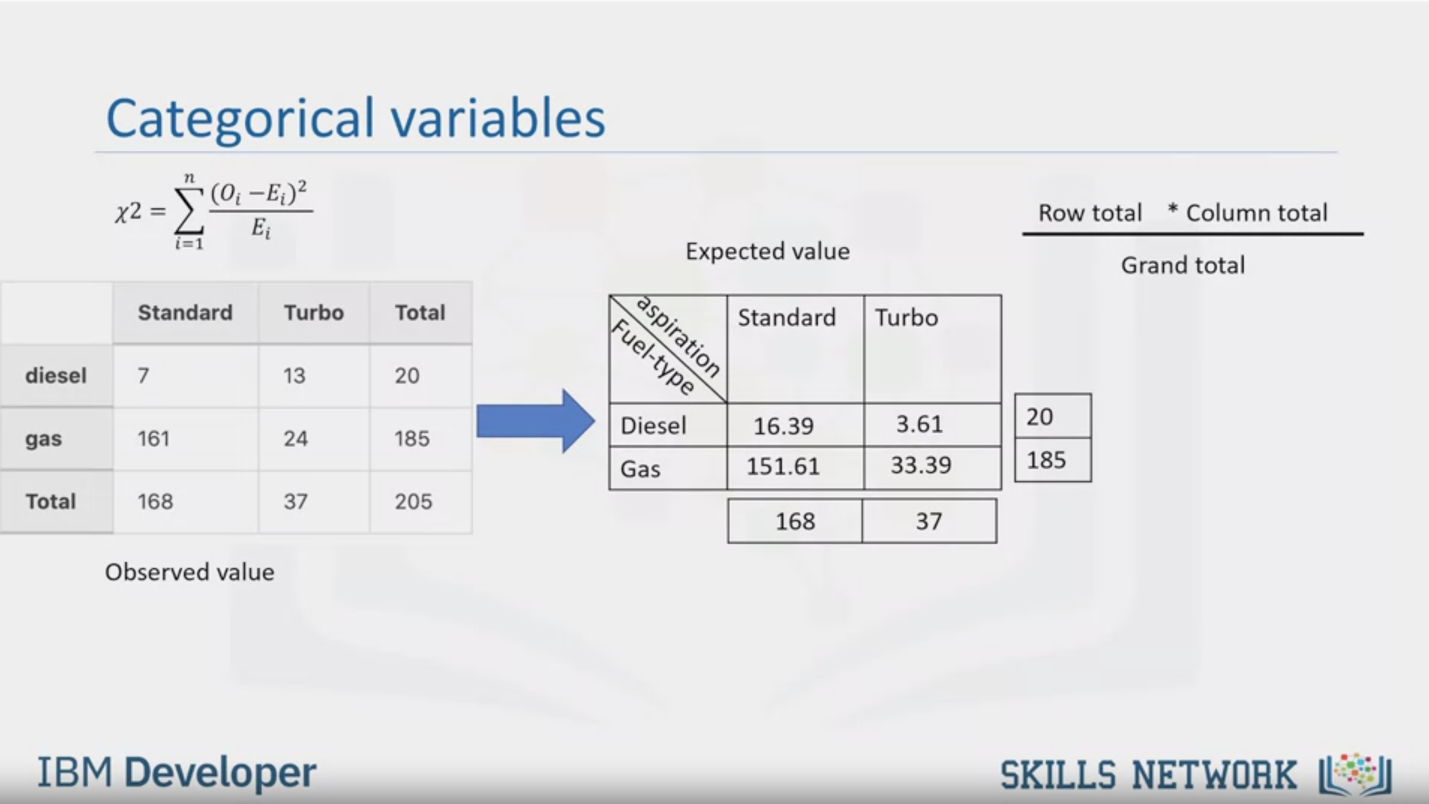
168

37

205

IBM Developer

SKILLS NETWORK



Categorical variables

Standard

Turbo

Total

diesel

gas

Total

7

161

168

13

24

37

20

185

205

Observed value

Expected value

aspiration

Fuel-type

Diesel

Gas

Standard

Turbo

16.39

151.61

3.61

33.39

20

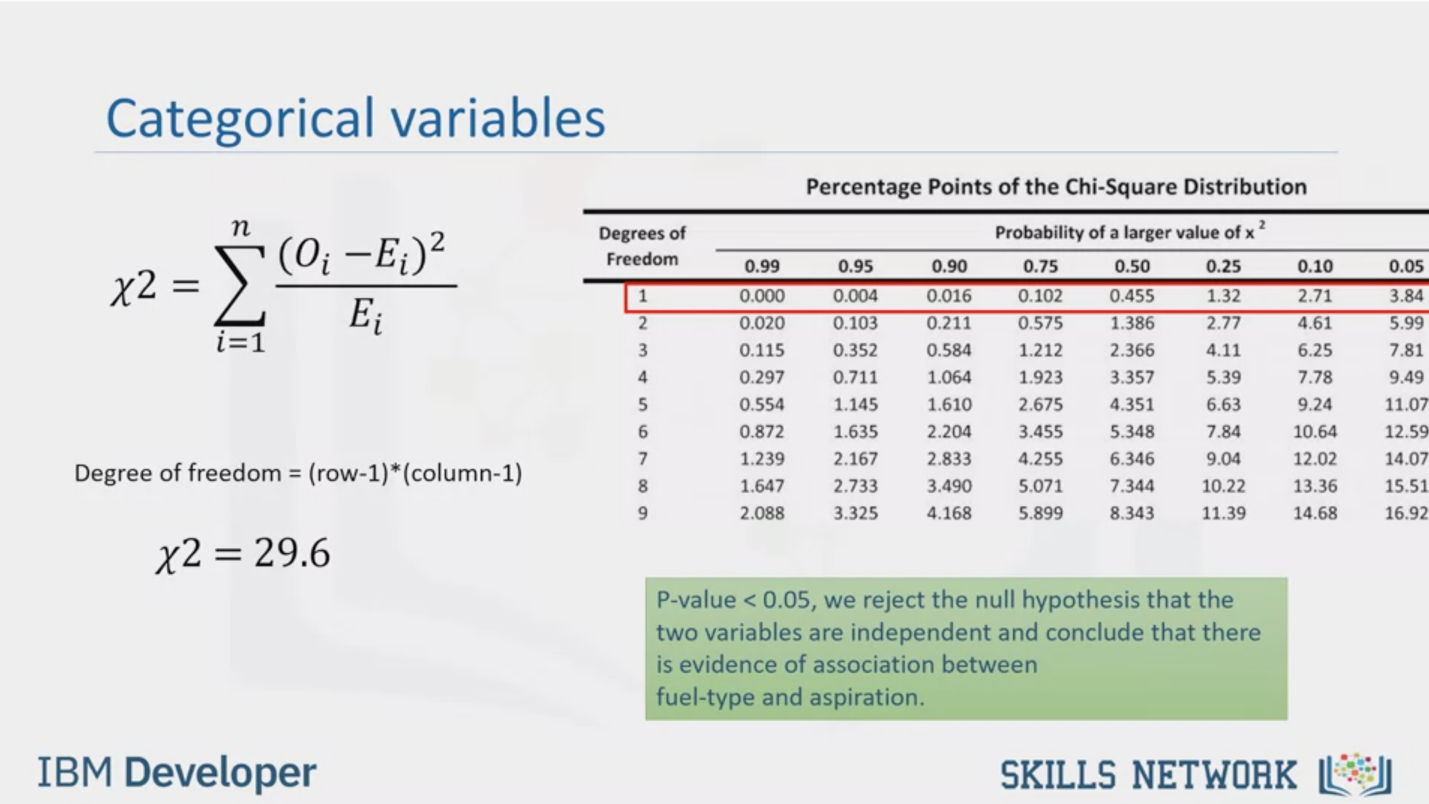
185

168

37

IBM Developer

SKILLS NETWORK



Categorical variables

Degree of freedom = (row-1)\* (column-1)

29.6

Degrees of

Freedom

0.99

0.000

0.020

0.115

0.297

0.554

0.872

1.239

1.647

2.088

Percentage Points of the Chi-Square Distribution

0.95

0.004

0.103

0.352

0.711

1.145

1.635

2.167

2.733

3.325

0.90

Probability of a larger value of x?

0.75

0.50

0.25

0.016

0.211

0.584

1.064

1.610

2.204

2.833

3.490

4.168

0.102

0.575

1.212

1.923

2.675

3.455

4.255

5.071

5.899

0.455

1.386

2.366

3.357

4.351

5.348

6.346

7.344

8.343

1.32

2.77

4.11

5.39

6.63

7.84

9.04

10.22

11.39

0.10

2.71

4.61

6.25

7.78

9.24

10.64

12.02

13.36

14.68

0.05

3.84

5.99

7.81

9.49

11.07

12.59

14.01

15.51

16.92

P-value < 0.05, we reject the null hypothesis that the

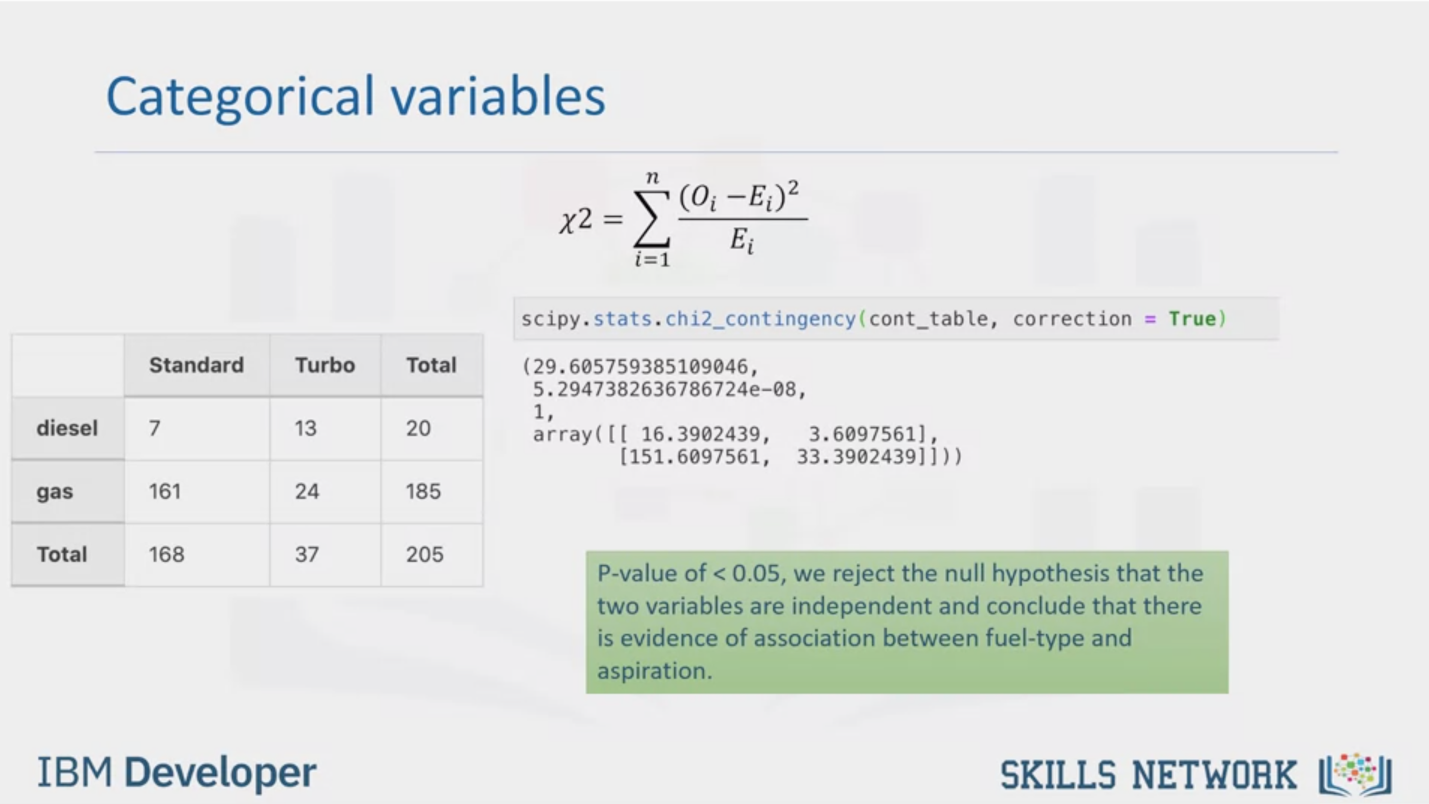
two variables are independent and conclude that there

is evidence of association between

fuel-type and aspiration.

IBM Developer

SKILLS NETWORK



Categorical variables

diesel

Standard

7

gas

Total

161

168

Turbo

13

24

37

Total

20

185

205

scipy.stats.chi2\_contingency(cont\_table, correction = True)

(29.605759385109046,

5.2947382636786724e-08,

1,

array([[ 16.3902439, 3.60975611],

[151.6097561, 33.390243911]]))

P-value of <0.05, we reject the null hypothesis that the

two variables are independent and conclude that there

is evidence of association between fuel-type and

aspiration.

IBM Developer

SKILLS NETWORK



IBM Developer

SKILLS NETWORK

In this video, we will learn how to find out if there is a relationship between two categorical

variables. When dealing with the relationships between

two categorical variables, we can’t use the same correlation method for continuous

variables, we will have to employ the use of chi square

test for the association.

The Chi-square test is intended to test how likely it is that an observed distribution

is due to chance. It measures how well the observed distribution

of data fits with the distribution that is expected if the variables are independent.

Before we go into an example, let's look through some important points.

The Chi-square tests a null hypothesis that the variables are independent. The test compares

the observed data to the values that the model expects if the data was distributed in different

categories by chance. Anytime the observed data doesn't fit within the model of the expected

values, the probability that the variables are dependent becomes stronger, thus proving

the null hypothesis incorrect. The Chi-square does not tell you the type

of relationship that exists between both variables, but only that a relationship exists.

We will use the cars dataset. Assuming we want to test the relationship between fuel-type

and aspiration; these are categorical variables. It is either the fuel-type of the car, gas

or diesel, and the aspiration is that either the car is standard or turbo. To do this we

will find the observed counts of cars in each category. This can be done by creating a crosstab

using the pandas library. A crosstab is a table showing the relationship between two

or more variables. When the table only shows the relationship between two categorical variables,

the crosstab is also known as a contingency table. In our case the crosstab or contingency

table shows us the counts in each category:

A standard car with diesel fuel,

a standard car with gas fuel, a turbo car with diesel fuel,

or a turbo car with gas fuel. The formula for chi-square is given as follows:

The summation of the observed value i.e., the counts in each group, minus the expected

value, all squared, divided by the expected value.

Expected values are based on the given totals, that is, what can we say individual cells would

be if we did not know the observed values? To calculate the expected value of a standard

car with diesel, we take the row total which is twenty multiplied

by the column total, one hundred and sixty-eight,

divided by the grand total of two hundred and five.

This will give you sixteen point three nine.

If we do the same thing for turbo cars with gas fuel, we will take the

row total, one hundred and eighty-five, multiplied by

column total, thirty-seven, and we divide by

the grand total, two hundred and five, to get

thirty-three point three nine. If we repeat the same procedure for all of

them we get these values.

If we took the row totals, column totals, and grand total

we will get the same values as the totals as the observed values.

Now going back to this formula, if we took a summation of all the observed, minus the

expected values, all squared, divided by the expected value, we will get a chi-square value

of twenty-nine point six.

On the chi-square table

we check on the degree of freedom equals one row and find the value closest to twenty-nine point six.

Here we can see that twenty-nine point six will fall in between a p-value less

than zero point zero five. Therefore, we can say the p-value is less than

zero point zero five. Since the p-value is less than point zero five, we reject the

null hypothesis that the two variables are independent and therefore we conclude that

there is an association between fuel type and aspiration.

To do this in Python we will use the chi square contingency function in the SciPy dot statistics

package. The function will print out the chi-square

test value twenty-nine point six, the second value is the p-value which is very close to

zero and a degree of freedom of one. If you remember, the chi-square table did not give an exact

p-value but a range in which it falls. Python will give the exact p-value. We can see the

same results as in our previous slides. It also prints out the expected values which

we also calculated by hand. Since the p-value is close to zero, we reject

the null hypothesis that the two variables are independent and conclude that there is

evidence of association between fuel-type and aspiration.