Train In Data

Chi-square test

Chi-square test

Chi-square is a statistic metric, used to determine if 2 samples of categorical features were extracted from the same population.

Chi-square test

Chi-square is a statistic metric, used to determine if 2 samples of categorical features were extracted from the same population.

Compares the distributions of the categories.

Chi-square test – main uses

Chi-square goodness of fit

Chi square test of independence

Chi-square test – main uses

Chi-square goodness of fit

Chi square test of independence

Chi-square goodness of fit

Determines if a categorical variable follows a hypothesized distribution.

Chi-square test

Chi-square test compares the distributions of the categories across the variables.

	Female x died	Female x survived	Male x survived	Male x died
Expected	100	100	50	50

Data consists of 200 women and 100 man

Chi-square test

Chi-square test compares the distributions of the categories across the variables.

	Female x died	Female x survived	Male x survived	Male x died
Expected	100	100	50	50
Random sample 1	100	100	50	50

Data consists of 200 women and 100 man

Chi-square test

Chi-square test compares the distributions of the categories across the variables.

	Female x died	Female x survived	Male x survived	Male x died
Expected	100	100	50	50
Random sample 1	100	100	50	50
Random sample 2	95	105	45	55

Data consists of 200 women and 100 man

Chi-square test

Chi-square test compares the distributions of the categories across the variables.

	Female x died	Female x survived	Male x survived	Male x died
Expected	100	100	50	50
Random sample 1	100	100	50	50
Random sample 2	95	105	45	55
Random sample 3	120	90	30	60
Random sample 4	150	100	20	30

Data consists of 200 women and 100 man

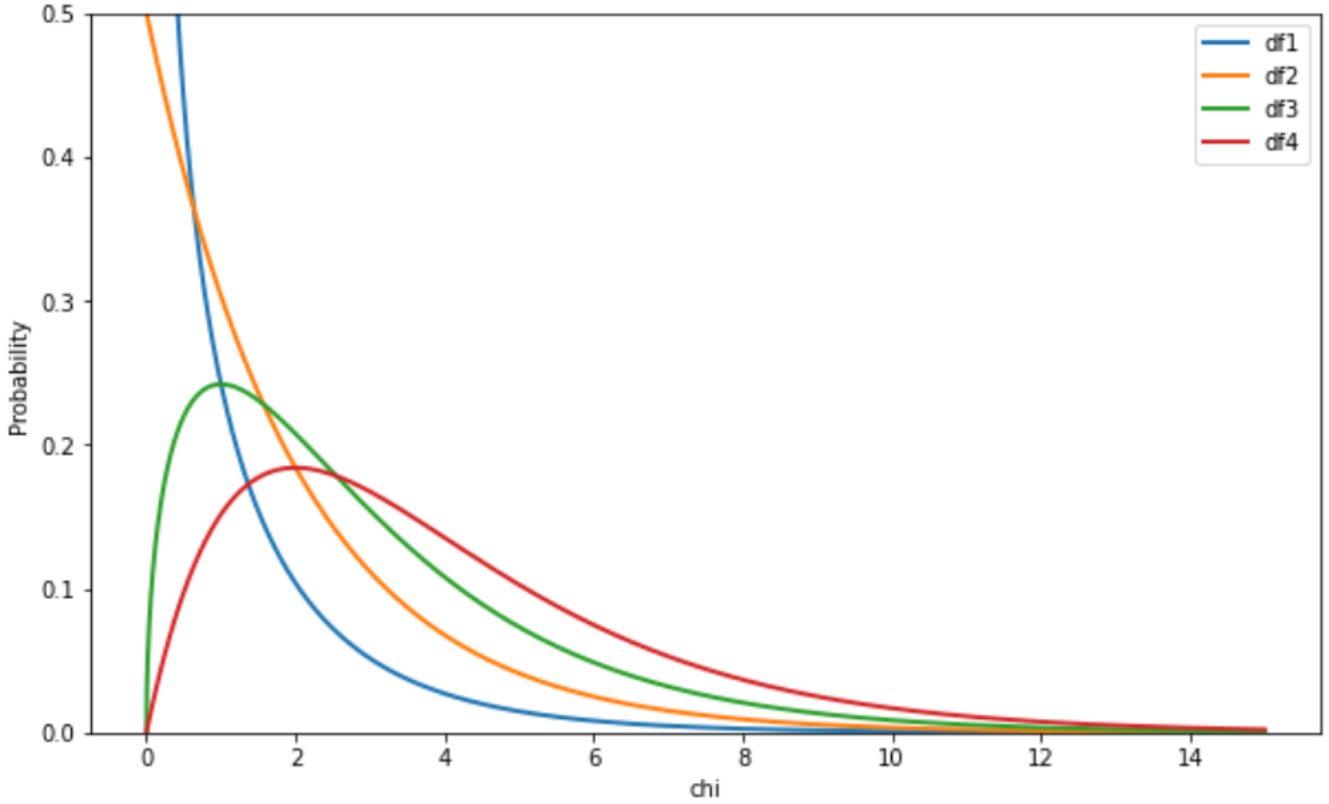
Chi-square test

Chi-square test compares the distributions of the categories across the variables.

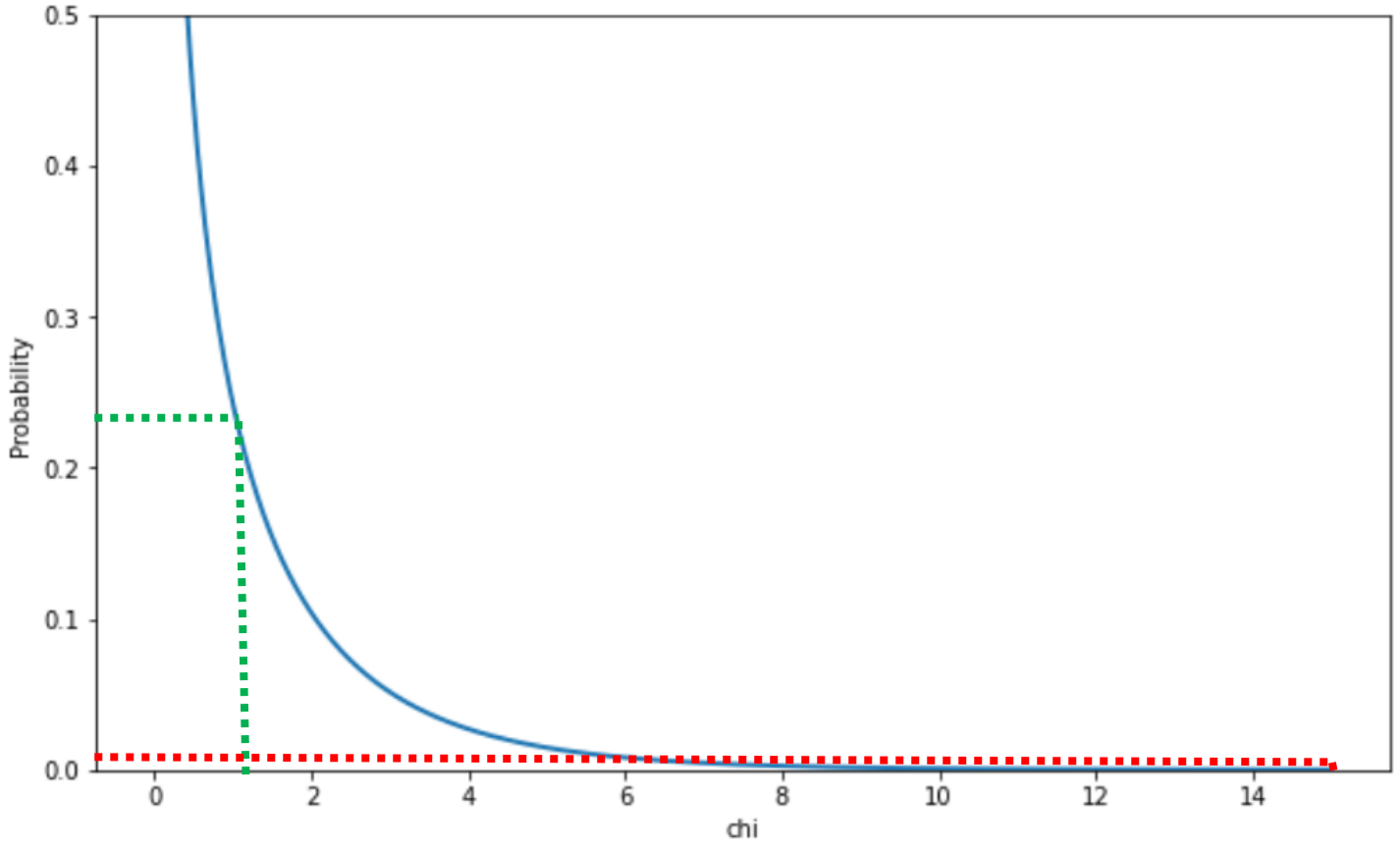
	Female x died	Female x survived	Male x survived	Male x died	χ^2
Expected	100	100	50	50	
Random sample 1	100	100	50	50	0
Random sample 2	95	105	45	55	1.5
Random sample 3	120	90	30	60	15
Random sample 4	150	100	20	30	51

Data consists of 200 women and 100 man

Chi-square distribution



Chi-square distribution



With $\chi 2$, we can obtain an estimate of the probability based on the chi-squared distribution.

Chi-square calculation

$\chi 2$ = sum (Observed – expected)^2 / expected

	Female x died	Female x survived	Male x survived	Male x died
Expected	100	100	50	50
Random sample 3	120	90	30	60

$$\frac{(120-100)^2}{100} + \frac{(90-100)^2}{100} + \frac{(30-50)^2}{50} + \frac{(60-50)^2}{50}$$

Data consists of 200 women and 100 man

Chi-square calculation

$$\chi 2 = \text{sum } (\text{Observed} - \text{expected})^2 / \text{expected}$$

	Female x died	Female x survived	Male x survived	Male x died
Expected	100	100	50	50
Random sample 2	120	90	30	60

$$4 + 1 + 8 + 2 = 15$$

Data consists of 200 women and 100 man

Contingency table

$$\chi 2 = \text{sum } (\text{Observed} - \text{expected})^2 / \text{expected}$$

Observed

	Female	Male
Died	120	60
Surived	92	30

Expected

	Female	Male
Died	100	50
Surived	100	50

Data consists of 200 women and 100 man

Contingency table

$$\chi 2 = \text{sum } (\text{Observed} - \text{expected})^2 / \text{expected}$$

Observed

	Cats	Dogs
Brown	200	60
Ginger	100	10

Expected

	Cats	Dogs
Brown	?	?
Ginger	?	?

We want to predict fur color based on the animal species.

Data consists of 300 cats and 70 dogs

Contingency table

$$\chi 2 = \text{sum } (\text{Observed} - \text{expected})^2 / \text{expected}$$

Observed

	Cats	Dogs	
Brown	200	60	260
Ginger	100	10	110
	300	70	370

Expected

	Cats	Dogs
Brown	?	?
Ginger	?	?

We calculate the marginals.

Data consists of 300 cats and 70 dogs

Contingency table

$$\chi 2 = \text{sum } (\text{Observed} - \text{expected})^2 / \text{expected}$$

Observed

	Cats	Dogs	
Brown	200	60	260
Ginger	100	10	110
	300	70	370

Expected

	Cats	Dogs
Brown	260 x 300 / 370	260 x 70 / 370
Ginger	110 x 300 / 370	110 x 70 / 370

$$E = (\text{Row x Column}) / \text{Total}$$

With the marginal, we obtain the expected frequency.

Data consists of 300 cats and 70 dogs

Contingency table

$$\chi 2 = \text{sum } (\text{Observed} - \text{expected})^2 / \text{expected}$$

Observed

	Cats	Dogs	
Brown	200	60	260
Ginger	100	10	110
	300	70	370

Expected

	Cats	Dogs
Brown	210.8	49.19
Ginger	89.19	20.81

$$E = (\text{Row} \times \text{Column}) / \text{Total}$$

With the marginal, we obtain the expected frequency.

Data consists of 300 cats and 70 dogs

Contingency table

$\chi 2$ = sum (Observed – expected)^2 / expected

Observed

	Cats	Dogs	
Brown	200	60	260
Ginger	100	10	110
	300	70	370

Expected

	Cats	Dogs
Brown	210.8	49.19
Ginger	89.19	20.81

$$\frac{(200-210.8)^2}{210.8} + \frac{(60-49.19)^2}{49.19} + \frac{(100-89.19)^2}{89.19} + \frac{(10-20.8)^2}{20}$$

Data consists of 300 cats and 70 dogs

Contingency table

$$\chi 2 = \text{sum } (\text{Observed} - \text{expected})^2 / \text{expected}$$

Observed

	Cats	Dogs	
Brown	200	60	260
Ginger	100	10	110
	300	70	370

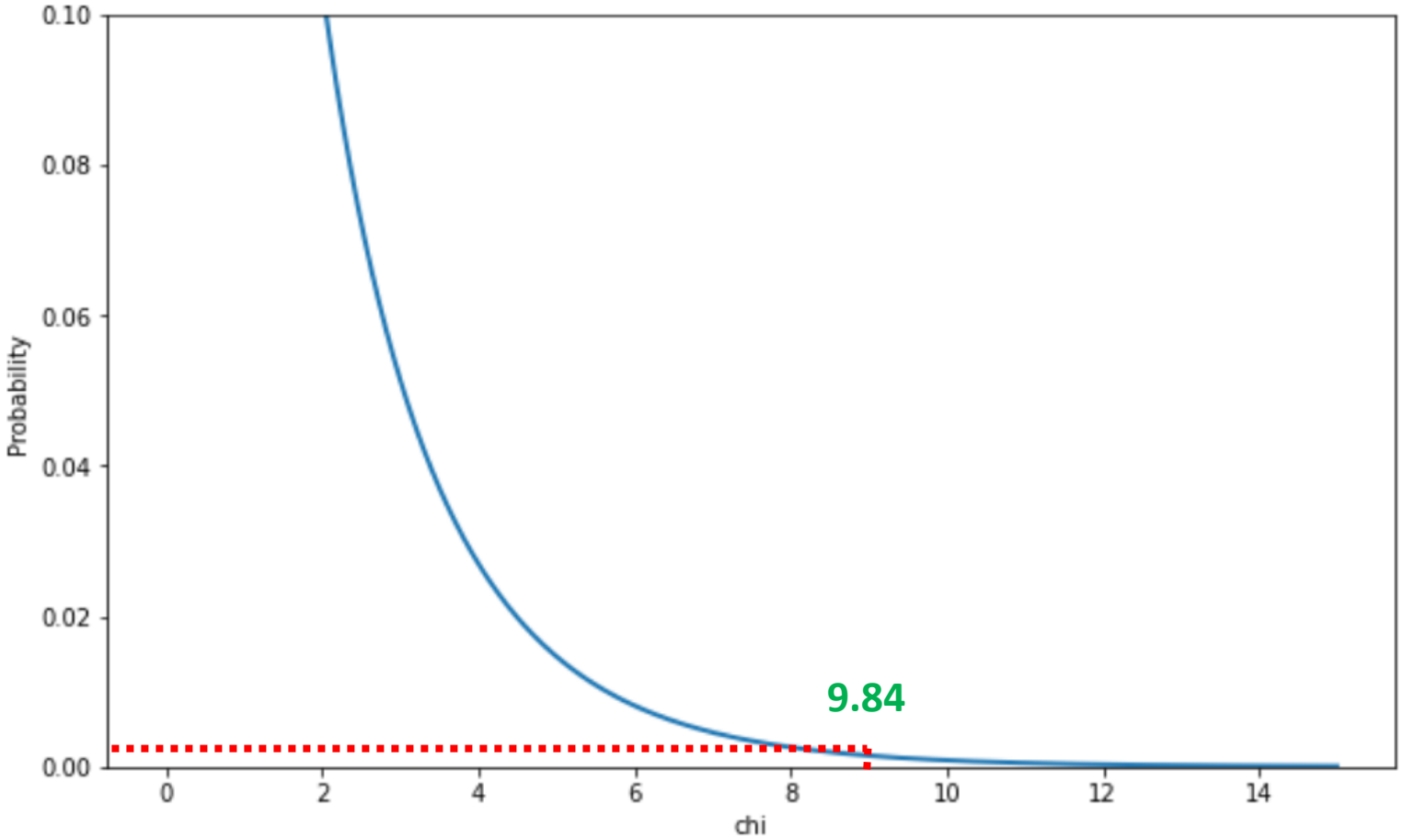
Expected

	Cats	Dogs
Brown	210.8	49.19
Ginger	89.19	20.81

$$0.55 + 2.37 + 1.31 + 5.61 = 9.84$$

Data consists of 300 cats and 70 dogs

Chi-square distribution



The probability of cats and dogs having the same distribution of brown and ginger is very low.

There is an association between color and animal.

Degrees of freedom (dof)

Observed

	Cats	Dogs
Brown	200	60
Ginger	100	10

dof = (Row-1) x (Column-1)

dof = (2-1) x (2-1) = 1

Chi-square for categorical data

If data contains:

- Categorical variables.
- Binary or multi-class target.

We can assess the association of the categorical variable with the target, using the chi-squared test.

Chi-square ranking process

1. Create a contingency table between the categorical variable and the target (observed)
2. Find the expected distribution
3. Calculate the chi-square statistic
4. Obtain the p-value

Chi-square ranking process

scipy.stats.contingency.chi2_contingency

`scipy.stats.contingency.chi2_contingency(observed, correction=True, lambda_=None)`

Chi-square test of independence of variables in a contingency table. [source]

This function computes the chi-square statistic and p-value for the hypothesis test of independence of the observed frequencies in the contingency table [1] *observed*. The expected frequencies are computed based on the marginal sums under the assumption of independence; see `scipy.stats.contingency.expected_freq`. The number of degrees of freedom is (expressed using numpy functions and attributes):

```
dof = observed.size - sum(observed.shape) + observed.ndim - 1
```

Selection based on Chi-square

1. Rank the features based on the p-value or chi-square
    1. The higher the chi-square or the lower the p-value the more predictive the feature
2. Select the top ranking features
    1. Cut-off for top ranking features is arbitary

Train In Data

THANK YOU

www.trainindata.com