Feature Selection Techniques using Statistics [Habiba Shera]

ANOVA for feature selection in Machine Learning

> 💡 The ***biggest challenge*** in machine learning is ***selecting the best features*** to train the model. We need only the features which are highly dependent on the response variable (variable which will be predicted later). **But what if the response variable is continuous and the predictor is categorical ?. Now we need to use ANOVA**

ANOVA (**An**alysis **of Va**riance) helps us to complete our job of selecting the best features

> 💡 **ANOVA tests the relationship between categorical predictor vs continuous response.**

Steps

- we will check ***if there is equal variance between groups of categorical feature and continuous response***.
- ***If there is equal variance between groups, it means this feature has no impact*** on response and it ***can not be considered for model training.***

ANOVA Types

- **One way ANOVA**

    - we can check ***only single predictor vs response and determine the relationship***
- **Two way ANOVA**
    - if we have two features (predictors)
- **multi-factor ANOVA**
    - if there are ***more than two features***

Chi-Square Test for Feature Selection in Machine learning

The **chi-square** test helps you to **solve the problem in feature selection** by testing the relationship between the features.

- **Example**
    - why customers are leaving the bank, ***Gender*** of a customer with values as ***Male/Female*** and ***Exited*** describes whether a customer is leaving the bank with values Yes/No as the response.

Steps to perform the Chi-Square Test:

1. Define Hypothesis.
2. Build a Contingency table.
3. Find the expected values.
4. Calculate the Chi-Square statistic.
5. Accept or Reject the Null Hypothesis

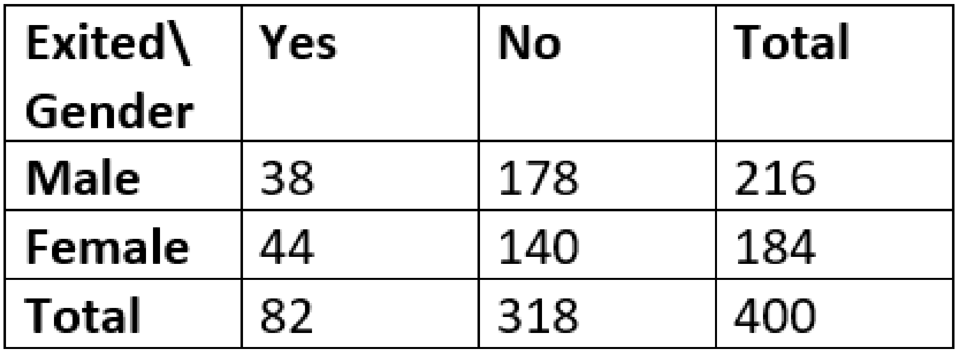
Lets work on the last example which is the relationship between gender and exited the bank

1. **Define Hypothesis**

   Null Hypothesis (H0) : Two variables are independent.

   Alternate Hypothesis (H1) : Two variables are not independent.

2. **Contingency table**



Exited\ Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

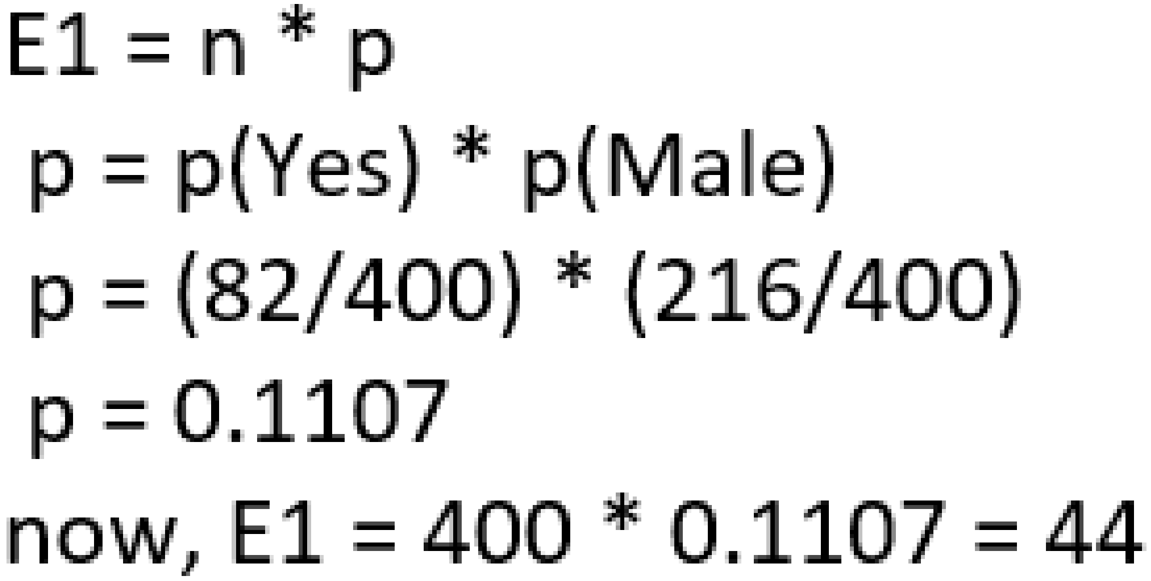
Then calculate **degree of freedom :** `(rows-1) * (columns-1)` **= (2-1) * (2-1) = 1**

3. **Find the Expected Value**

   - **Based on the null hypothesis** that the ***two variables are independent.*** We can say if A, B are two independent events when

$$P(A \cap B) = P(A) * P(B)$$

Then lets the expected value :



$$E1 = n * p$$

$$p = p(Yes) * p(Male)$$

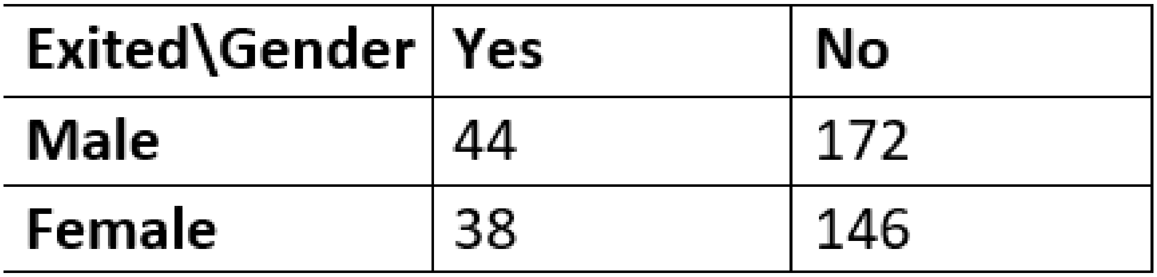
$$p = (82/400) * (216/400)$$

$$p = 0.1107$$

$$\text{now, } E1 = 400 * 0.1107 = 44$$

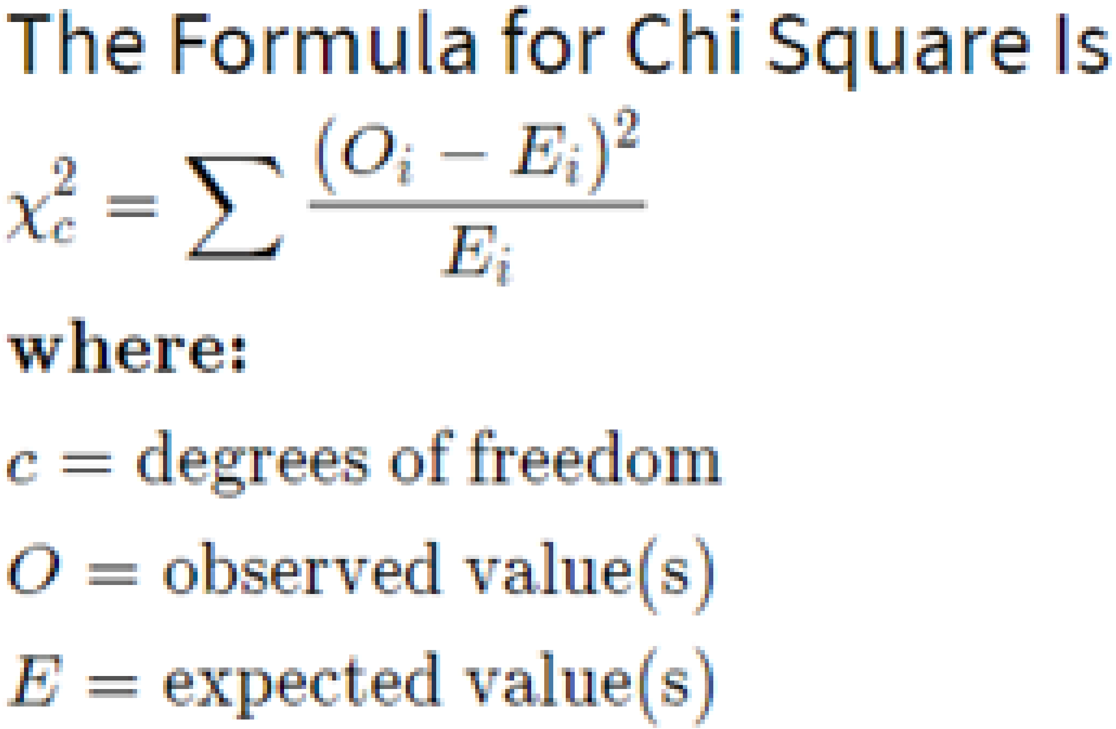
The calculation for the expected value

In similar, we calculate E2, E3, E4



Exited\Gender	**Yes**	**No**
Male	44	172
Female	38	146

4. **Calculate Chi-Square value**



The Formula for Chi Square Is

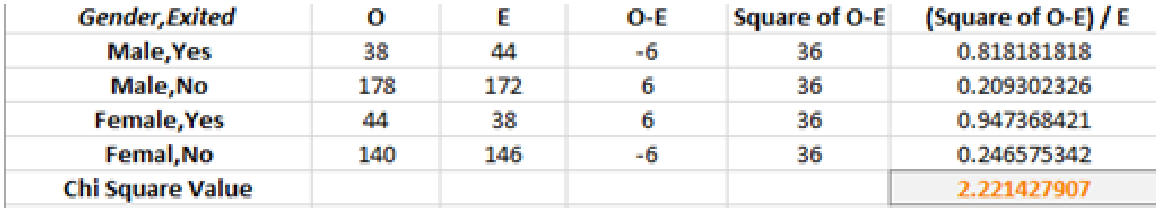
$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

$c =$ degrees of freedom

$O =$ observed value(s)

$E =$ expected value(s)



Gender,Exited	O	E	O-E	Square of O-E	(Square of O-E) / E
Male,Yes	38	44	-6	36	0.818181818
Male,No	178	172	6	36	0.209302326
Female,Yes	44	38	6	36	0.947368421
Femal,No	140	146	-6	36	0.246575342
Chi Square Value					2.221427907

We can see Chi-Square is calculated as 2.22 by using the Chi-Square statistic formula.

5. **Accept or Reject the Null Hypothesis**

- **With 95% confidence that is alpha = 0.05**, we will check the calculated Chi-Square value falls in the acceptance or rejection region.
- Going to Chi-Square Table (mathsisfun.com) we can find that chi-square value is 3.93

So here we are accepting the null hypothesis **since the Chi-Square value is less than the critical Chi-Square value.**

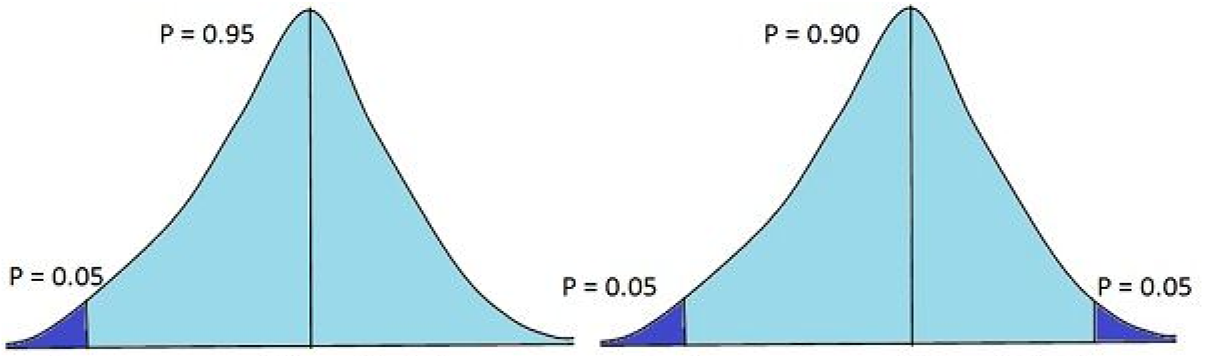
Chi-Square in python

Difference Between One-tailed and Two-tailed Test

💡 **In a test**, there are **two divisions of probability** density curve, region of **acceptance** and region of **rejection**. the region of rejection is called as a **critical region**.

💡 There are two types of test which are **one-tailed test** and **two-tailed test**

	One-Tailed Test	**Two-Tailed Test**
meaning	statistical hypothesis test in which hypothesis has **only one end**	hypothesis **has two ends**
rejection region	Either **left or right**	**Both left and right**
result	Greater or less **than certain value.**	Greater or less than c**ertain range of values.**



One-tailed Test Vs Two-tailed Test

Chi-Square Test vs ANOVA

- there are two different types of **Chi-Square tests**
    - The **Chi-Square Goodness of Fit Test** : Used to determine ***whether or not a categorical variable follows a hypothesized distribution***.
        - **Example :**
            - We want to know if a die is fair, so we roll it 50 times and record the number of times it lands on each number.
    - The **Chi-Square Test of Independence** : Used to determine ***whether or not there is a relationship between two categorical variables***.
        - **Example:**
            - We want to know if a person's favorite color is associated with their favorite sport so we survey 100 people and ask them about their preferences for both.

- **In ANOVA :**
    - **we calculate the relationship between numeric values and categorical values**
    - **Example**
        - We want to know if three different studying techniques lead to different mean exam scores.

> 💡 **To use an ANOVA when there is at least one categorical variable and one continuous dependent variable.**

When to Use Chi-Square Tests vs ANOVA

- **Use Chi-Square Tests** when every variable you're working with is categorical

- **Use ANOVA** when you have at least one categorical variable and one continuous dependent variable.

Practice Examples

- Suppose a researcher want to know if education level and marital status are associated or not (`Chi-square`)
- basketball trainer wants to know if three different training techniques lead to different mean jump height among his players. (`ANOVA`)

F-Test

> 💡 **F-Test is used when** you want to know whether there is a statistical difference **between two continuous variables** (height and weight). or **to test to see if two samples come from populations with the same variance**

> 💡 **F-Test assumes** that data are **normally distributed** and that **samples are independent** from one another.

> 🏁 **ANOVA assumes a linear relationship between the feature and the target and that the variables follow a Gaussian distribution**. If this is not true, the result of this test may not be useful.

References

- ANOVA for Feature Selection in Machine Learning | by sampath kumar gajawada | Towards Data Science
- Chi-Square Test for Feature Selection in Machine learning | by sampath kumar gajawada | Towards Data Science
- Difference Between One-tailed and Two-tailed Test (with Comparison Chart) - Key Differences
- Chi-Square Test vs. ANOVA: What's the Difference? - Statology
- Chi-Square Test vs. F Test | Quality Gurus
- Chi Square and Anova — Feature Selection for ML | by ML2021dsb | Medium