

Starting Off

We can use an ANOVA test to see if a continuous variable is independent of a categorical variable.

What statistical measure would we use to measure if there is a relationship between two continuous variables?

X Chi-squared test

Goal -> implement chi-squared

- > Use cases
- > Review key concepts
- > When to use the chi-squared test
- > What is chi-squared
- > Calculate!

Let's review

1. The fundamentals of the sampling distributions for the sample mean and the sample proportion.
2. We illustrated how these sampling distributions form the basis for estimation (confidence intervals) and testing for one mean or one proportion.
3. Then we extended the discussion to analyzing situations for two variables; one a response and the other an explanatory. When both variables were categorical we compared two proportions; when the explanatory was categorical, and the response was quantitative, we compared two means.
4. Next, we will take a look at other methods and discuss how they apply to situations where:
 - both variables are categorical with at least one variable with more than two levels (Chi-square Test of Independence)
 - the explanatory variable is categorical with more than two levels, and the response is quantitative (Analysis of Variance or ANOVA)
 - both variables are quantitative (Linear Regression)

Let's review

Parametric tests:

- Require assumptions about population characteristics: normality of the underlying distribution, homogeneity of variance, known mean / variance.
- Examples: F, z, t tests

Nonparametric tests:

- Do not require assumptions about population characteristics.
- Can be used with very skewed distributions or when the population variance is not homogeneous.
- Can be used with ordinal or nominal data.
- Examples: Chi-square, Wilcoxon, and Kruskal-Wallis tests

Nonparametric tests are less powerful than parametric tests, so we don't use them when parametric tests are appropriate. But if the assumptions of parametric tests are violated, we use nonparametric tests.

What is chi-squared used for?

Statistical evidence of **association or relationship** between **categorical variables**

Goodness of fit - does observed frequency distribution differ from a theoretical distribution.

Homogeneity - is two populations distributions for a categorical variable are the same.

Independence - determine if two variables are independent of each other.

<https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/chi-square-tests-review/>

What is Chi-Squared Test of Independence

Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a so-called contingency table.

Null: no association or relationship between the distributions -> they are independent

Alternative: there is an association or relationship between the distributions -> they are not independent

Chi-squared test is like the z-test for two independent proportions, but when you have more than two groups that you are comparing against.

Is the rate of survival different for different types of cancer (colon, breast, throat, etc.)?

Check for understanding

Should we use the chi-squared tests in the following situations?

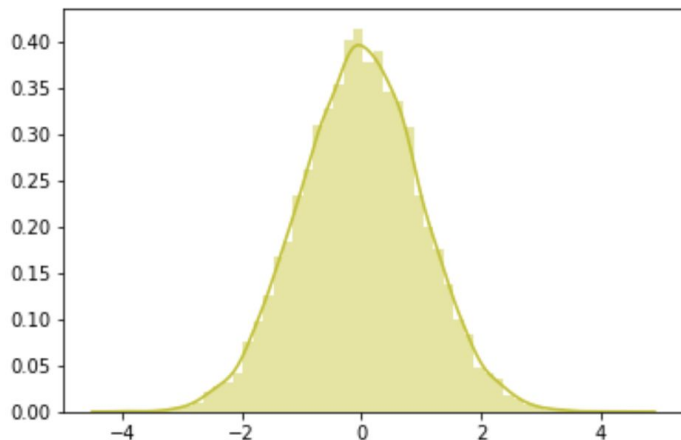
1. Do first generation college students study the same majors (Business, Art History, Computer Science, etc.) as legacy students?
2. Do children who receive vaccinations have a higher incidence of autism than children who don't?
3. Is the average highway gas mileage the same for 3 different types of cars?
4. 6 months after graduation, do students from different data science programs (Flatiron, General Assembly, Metis) have different job placement rates?

What is chi-squared part 2

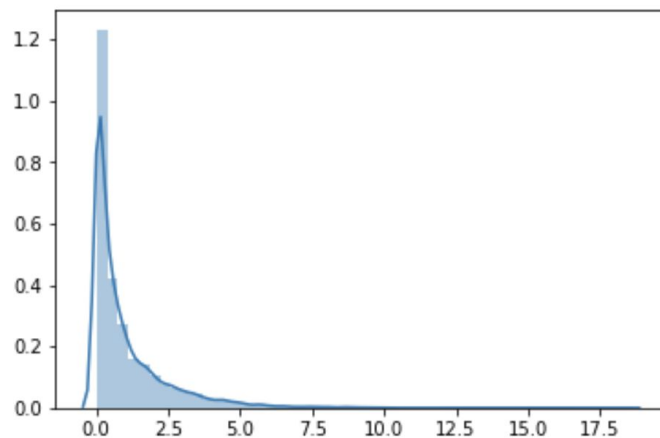
Cool graphs

Normal distribution \rightarrow square the random variable \rightarrow **chi-squared**

pdf



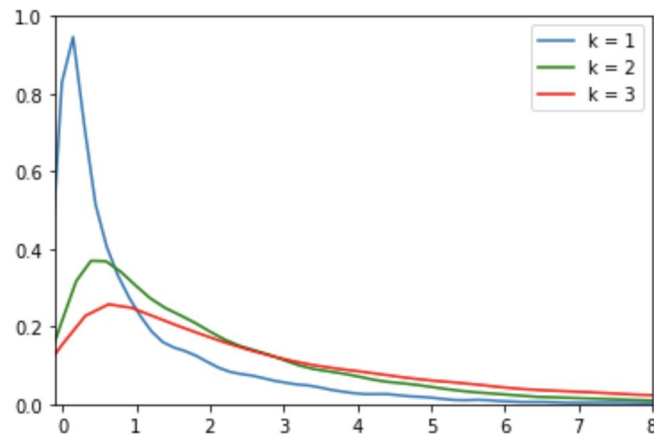
pdf



To use chi-squared

Every statistical test has assumptions, for chi-squared

- Data should be frequencies, not percentages
- Categories are mutually exclusive
- The observations are assumed be independent of each other - you randomly sampled your observations.
- More than 80% of cells must have a >5 count



Calculating

First, are we comparing categorical variables?

1. Null and alternative hypotheses
2. Chi-squared test statistic
3. Degrees of freedom
4. Confidence level
5. Compare all three ^
6. Reject or fail to reject the null

The primary method for displaying the summarization of categorical variables is called a contingency table. When we have two measurements on our subjects that are both the categorical, the contingency table is sometimes referred to as a two-way table.

What is chi-squared?

Start with a random variable Y , make a normal distribution $Z = \sum \frac{(Y - \mu)}{\sigma}$

Square the random variable $Z^2 = \sum \left(\frac{Y - \mu}{\sigma} \right)^2$

Summing the random variables gives the distribution $Q = \sum_{i=1}^k Z_i^2$

with one parameter -> **degrees of freedom** $Q \sim \chi^2(k)$

1. define the question

A sample of students have the option of ice cream or cake after school. Is there a relationship between grade level and snack? Test the hypothesis with a significance of 5%

	Ice cream	Cake	total
3rd grade	18	11	29
4th grade	15	16	31
5th grade	9	15	24
total	42	42	84

$\alpha =$

This is a 3x2

Contingency Table

2a. find chi-squared

The chi-squared test statistic is

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

α is 0.05

	Ice cream	Cake	total
3rd grade	O = 18 E =	O = 11 E =	29
4th grade	O = 15 E =	O = 16 E =	31
5th grade	O = 9 E =	O = 15 E =	24
total	42	42	84

$\alpha = 0.05$

2b. find expected values

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

O is observed values (that are given)

E is expected values (row total x column total)/sample size

	Ice cream	Cake	total
3rd grade	O = 18 E =	O = 11 E =	29
4th grade	O = 15 E =	O = 16 E =	31
5th grade	O = 9 E =	O = 15 E =	24
total	42	42	84

$$\alpha = 0.05$$

2c. find chi-squared test statistic

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

Now calculate χ^2

	Ice cream	Cake	total
3rd grade	O = 18 E = 14.5	O = 11 E = 14.5	29
4th grade	O = 15 E = 15.5	O = 16 E = 15.5	31
5th grade	O = 9 E = 12	O = 15 E = 12	24
total	42	42	84

$$\alpha = 0.05$$

$$\chi^2 = 3.222$$

3. degrees of freedom

degrees of freedom = (number of rows - 1)(number of columns - 1)

	Ice cream	Cake	total
3rd grade	O = 18 E = 14.5	O = 11 E = 14.5	29
4th grade	O = 15 E = 15.5	O = 16 E = 15.5	31
5th grade	O = 9 E = 12	O = 15 E = 12	24
total	42	42	84

$$\alpha = 0.05$$

$$\chi^2 = 3.222$$

$$df =$$

4. + 5. compare confidence interval

Use a [table](#) of **degrees of freedom** and **chi-squared statistic** to compare

	Ice cream	Cake	total
3rd grade	O = 18 E = 14.5	O = 11 E = 14.5	29
4th grade	O = 15 E = 15.5	O = 16 E = 15.5	31
5th grade	O = 9 E = 12	O = 15 E = 12	24
total	42	42	84

$$\alpha = 0.05$$

$$\chi^2 = 3.222$$

$$df = 2$$

$$CV = 5.9915$$

6. Reject the null?

If the chi-squared statistic > critical value, reject the null hypothesis $Q_k \geq \chi^2_{k,\alpha}$

Else, fail to reject the null hypothesis

	Ice cream	Cake	total
3rd grade	O = 18 E = 14.5	O = 11 E = 14.5	29
4th grade	O = 15 E = 15.5	O = 16 E = 15.5	31
5th grade	O = 9 E = 12	O = 15 E = 12	24
total	42	42	84

$$\alpha = 0.05$$

$$\chi^2 = 3.222$$

$$\text{df} = 2$$

$$\text{CV} = 5.9915$$

chi-squared in action

Are floaties and paddleboards distributed equally among new york beaches?

	Brighton	Coney island	Rockaway	Fort Tilden	total
floaties		3	31	13	
paddle boards	15	28		5	71
total	17		54	18	120

$\alpha =$

$\chi^2 =$

df =

critical value =

chi-squared in code

χ^2 Test with scipy

```
In [8]: 1 # chi-squared test with similar proportions
        2 from scipy.stats import chi2_contingency
        3 from scipy.stats import chi2
```

```
In [9]: 1 # contingency table
        2 table = [10, 20, 30],
        3           [6, 9, 17]]
        4 print(table)
```

```
[[10, 20, 30], [6, 9, 17]]
```

```
In [10]: 1 stat, p, dof, expected = chi2_contingency(table)
        2 print('dof=%d' % dof)
        3 print(expected)
```

```
dof=2
```

```
[[10.43478261 18.91304348 30.65217391]
 [ 5.56521739 10.08695652 16.34782609]]
```

chi-squared in code

```
In [11]: 1 # interpret test-statistic
          2 prob = 0.95
          3 critical = chi2.ppf(prob, dof)
```

```
In [12]: 1 print('probability=%.3f, critical=%.3f, stat=%.3f' % (prob, critical, stat))
          2 if abs(stat) >= critical:
          3     → print('Dependent (reject H0)')
          4 else:
          5     → print('Independent (fail to reject H0)')
```

probability=0.950, critical=5.991, stat=0.272
Independent (fail to reject H0)

```
In [13]: 1 # interpret p-value
          2 alpha = 1.0 - prob
          3 print('significance=%.3f, p=%.3f' % (alpha, p))
          4 if p <= alpha:
          5     → print('Dependent (reject H0)')
          6 else:
          7     → print('Independent (fail to reject H0)')
```

significance=0.050, p=0.873
Independent (fail to reject H0)

Closing Question