# Chi-Squared Test

Data Boot Camp

Lesson 7.3

# The Chi-Squared Test Formula

Math formulas with Greek symbols can look intimidating. But the idea is simple, and we'll walk through what everything here means.

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

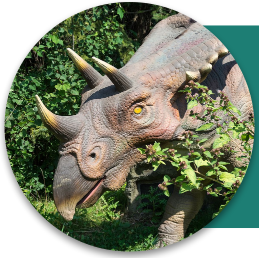
# The Chi-Squared Test Formula

## What is it used for?

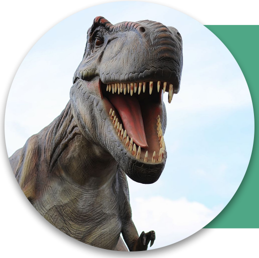
- To answer the question: Is the distribution of frequencies in the dataset meaningful?
- In other words, does the data match our expectations?
- In still other words, do we accept or reject the null hypothesis?

# The Chi-Squared Test Formula

**Example:** Out of 300 dinosaurs,



220 eat everything



55 eat only meat

25 eat only plants

**Null hypothesis:**
No statistical significance exists in the distribution of omnivores, carnivores, and herbivores. That is, this data can be explained by random distribution.

**The chi-squared test can help us accept or reject the null hypothesis.**

# The Chi-Squared Test Formula

**Σ (sigma)** the sum of

**O: observed values** 220 omnivores, 55 carnivores, 25 herbivores

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

**E: expected values (in randomly distributed data)**
100 omnivores, 100 carnivores, 100 herbivores

# The Chi-Squared Test Formula

$\chi^2 = (220-100)2/100 + (55-100)2/100 + (25-100)2/100$

$= 144 + 20.25 + 56.25$

$=$ **220.5**

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

# The Chi-Squared Test Formula

Steps for using the chi-squared test formula:

01 Determine chi-squared value.

02 Determine degree of freedom.

03 Choose a p-value.

04 Determine critical value.

05 Make a decision.

# Degree of Freedom

To determine the degree of freedom (df), take the number of rows and subtract 1:

**Omnivores:** 220
**Carnivores:** 55
**Herbivores:** 25

There are three rows, so the degree of freedom is

**3 – 1 = 2**

The degree of freedom is the number of figures required to fill out the table (like Sudoku).

If we have two of the numbers, we can figure out the value of the third.

# The Chi-Squared Test Formula

Steps for using the chi-squared test formula:

| | |
|---|---|
| 01 | Determine chi-squared value. |
| 02 | Determine degree of freedom. |
| 03 | Choose a p-value. |
| 04 | Determine critical value. |
| 05 | Make a decision. |

# P-value

The p-value is the **confidence level**, i.e., acceptable risk of a false positive.

**p = 0.05** is widely accepted in academia; may be higher in business settings.

We'll say **0.05** in this example.

# The Chi-Squared Test Formula

Steps for using the chi-squared test formula:

01 Determine chi-squared value.

02 Determine degree of freedom.

03 Choose a p-value.

04 Determine critical value.

05 Make a decision.

# Critical Value

The critical value is found by consulting a table:

| df | 0.995 | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | $\boxed{0.05}$ | 0.025 | 0.01 | 0.005 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | --- | --- | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 |
| $\boxed{2}$ | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | $\boxed{5.991}$ | 7.378 | 9.210 | 10.597 |
| 3 | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |
| 4 | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 |
| 5 | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.070 | 12.833 | 15.086 | 16.750 |
| 6 | 0.676 | 0.872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 |
| 7 | 0.989 | 1.239 | 1.690 | 2.167 | 2.833 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 |
| 8 | 1.344 | 1.646 | 2.180 | 2.733 | 3.490 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 |
| 9 | 1.735 | 2.088 | 2.700 | 3.325 | 4.168 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 |
| 10 | 2.156 | 2.558 | 3.247 | 3.940 | 4.865 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |

**df** = 2
**p** = 0.05
**critical value** = 5.991

# The Chi-Squared Test Formula

Steps for using the chi-squared test formula:

| | |
|---|---|
| 01 | Determine chi-squared value. |
| 02 | Determine degree of freedom. |
| 03 | Choose a p-value. |
| 04 | Determine critical value. |
| 05 | Make a decision. |

# Make a Decision

The calculated chi-squared value is **220.5**.

The critical value is **5.991**.

Since the chi-squared value **exceeds** the critical value, we decide that the difference between observed values and expected values are statistically significant.

# Using the Chi-Squared Test In Python

01 Import the **scipy.stats** module.

```
# The statistical module used to run chi square test
import scipy.stats as stats
```

02 Determine the **critical value.**

```
# The degree of freedom is 3-1 = 2
# With a p-value of 0.05, the confidence level is 1.00-0.05 = 0.95.
critical_value = stats.chi2.ppf(q = 0.95, df = 2)
```

```
# The critical value
critical_value
```

```
5.99146454710798
```

03 Run the **chi-squared test.**

```
# Run the chi square test with stats.chisquare()
stats.chisquare(df['observed'], df['expected'])
```

```
Power_divergenceResult(statistic=220.5, pvalue=1.3153258948574585e-48)
```

# The Chi-Squared Test Formula

Steps for using the chi-squared test formula:

| | |
|---|---|
| 01 | Determine chi square value. |
| 02 | Determine degree of freedom. |
| 03 | Choose a p-value. |
| 04 | Determine critical value. |
| 05 | Make a decision. |

# The Chi-Squared Test Formula

A few more considerations:

- The chi-squared test is used to test categorical variables; it can’t be used on continuous data.
- The categories must be mutually exclusive.
- We have covered using the chi-squared test formula to test goodness of fit.
- It can also be used to test independence. (Feel free to explore this on your own.)