Advanced Statistics

Dr. Syed Faisal Bukhari
Associate Professor
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics**, 13th Edition, Mario F. Triola

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman

- **Probability Demystified**, Allan G. Bluman

- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce

- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson

- **Python for Probability, Statistics, and Machine Learning,** José Unpingco

References

- Probability & Statistics for Engineers & Scientists, Ninth edition, Ronald E. Walpole, Raymond H. Myer
- Elementary Statistics, Tenth Edition, Mario F. Triola
- https://libguides.library.kent.edu/spss/chisquare

These notes contain material from the above resources.

Nominal Data

Definition: Categories with no inherent order or ranking.

Examples:

- **Gender:** Male, Female, Other
- **Blood Type:** A, B, AB, O
- **Favorite Color:** Red, Blue, Green

Ordinal Data

Definition: Categories with a specific order or ranking, but intervals are not equal.

Examples:

Education Level: High School, Bachelor’s, Master’s, PhD

Customer Satisfaction: Poor, Fair, Good, Excellent

Socioeconomic Status: Low, Middle, High

What are Categorical Variables?

- Variables representing **discrete categories or groups.**
- Take on a **limited number of distinct values.**
- Often represent qualitative data.

Examples:

- **Gender** (male, female, other)
- **Blood Type** (A, B, AB, O)
- **Marital Status** (single, married, divorced)
- **Education Level** (high school, bachelor's, master's, PhD)

Types of Categorical Variables

Categorical variables can be further classified into:

1. Nominal Variables

Categories without a natural order.

Examples: blood type, gender.

2. Ordinal Variables

Categories with an inherent order or ranking.

Examples: education level (high school < bachelor's < master's < PhD).

What is Likert Data?

Likert data is collected using a Likert scale, typically used in surveys to measure attitudes, opinions, or perceptions.

It consists of ordered responses to a statement, often on a **5- or 7-point scale**.

Type: Ordinal data, as responses follow a ranked order but **intervals are not equal**.

Examples of Likert Scale

Example 1: Agreement Scale

- **Statement: "I find data science exciting."**

  1 = Strongly Disagree

  2 = Disagree

  3 = Neutral

  4 = Agree

  5 = Strongly Agree

Examples of Likert Scale

Example 2: Frequency Scale

Question: " How often do you use data analysis software? "

1 = Never

2 = Rarely

3 = Sometimes

4 = Often

5 = Always

Summary of Likert Data

- Likert data is widely used to assess opinions and attitudes.

- Considered ordinal due to ranked choices without equal intervals.

- **Commonly analyzed** with **non-parametric tests**, such as Mann-Whitney U or **Chi-Square tests.**

Chi-Square Test of Independence

- The **Chi-Square Test of Independence determines** whether there is an association between **categorical variables** (i.e., whether the variables are independent or related). It is a nonparametric test.

- **Chi-Square Test of Independence:** Tests for association between categorical variables.

- **Non-parametric:** No assumptions about data distribution or normality, ideal for categorical data.

Chi-Square Test of Independence

Contingency Table: Used to analyze relationships between two categorical variables. It is (also known as a ***cross-tabulation***, ***crosstab***, or ***two-way table***)

Structure: Rows and columns represent categories of each variable.

Requirement: Each variable must have two or more categories.

Cells: Show count of cases for each category pair

Commons Uses

The Chi-Square Test of Independence is commonly used to test the following:

Purpose: Tests association between two or more categorical variables.

Limitations:

- Only compares categorical variables.
- Does not work with continuous variables.
- Shows association, not causation

Data Requirements

Data must meet the following requirements:

- Two categorical variables with two or more categories each.
- Independent observations (no related pairs).
- Large sample size.

Expected frequencies:

- At least 1 for each cell.
- At least 5 in 80% of cells.

Contingency Tables for Testing Independence of Attributes

- Suppose a medical researcher wants to determine whether there is a **relationship** between **caffeine consumption** and **heart attack risk**. Are these variables independent or are they dependent?

- We use the **chi-square test for independence** to answer such a question.

- To perform a **chi-square test for independence**, we will use sample data that are organized in a **contingency table**.

Contingency table

- An **$r \times c$ contingency table** shows the observed frequencies for two variables.
- The observed frequencies are arranged in **r rows** and **c columns**.
- The intersection of a row and a column is called a **cell**.

Example: The contingency table shows the results of a random sample of 2200 adults classified by their favorite way to eat ice cream and gender. At a = 0.01, can you conclude that the variables favorite way to **eat ice cream** and **gender are related**?

	Favorite way to eat ice cream				
Gender	**Cup**	**Cone**	**Sundae**	**Sandwich**	**Other**
Male	592	300	204	24	80
Female	410	335	180	20	55

1. **We state our hypothesis as**:

H_0: The variables favorite way to eat ice cream and gender are independent.

H_1: The variables favorite way to eat ice cream and gender are dependent. (Claim)

2. **The level of significance is set** $\alpha = 0.01$.

3. **Test statistic to be used is**

$$\chi^2_{cal} = \sum \frac{(O_f - E_f)^2}{E_f}$$

4. Calculations

	Favorite way to eat ice cream					Total
Gender	Cup	Cone	Sundae	Sandwich	Other	
Male	592	300	204	24	80	1200
Female	410	335	180	20	55	1000
Total	1002	635	384	44	135	**2200**

Expected Frequency (E_f)

$$E_{r,c} = \frac{(\text{Sum of row } r)(\text{Sum of column } c)}{\text{Sample size}}$$

$$E_{1,1} = \frac{(1200)(1002)}{2200} = 546.55$$

	Favorite way to eat ice cream					Total
Gender	**Cup**	**Cone**	**Sundae**	**Sandwich**	**Other**	
Male	546.55	346.36	209.45	24	73.64	1200
Female	455.45	288.64	174.55	20	61.36	1000
Total	1002	635	384	44	135	**2200**

O_f	E_f	$(O_f - E_f)$	$(O_f - E_f)^2$	$\frac{(O_f - E_f)^2}{E_f}$
592	546.55	45.45	2065.7025	3.7795
300	346.36	- 46.36	2149.2496	6.2052
204	209.45	- 5.45	29.7025	0.1418
24	24	0	0	0
80	73.64	6.36	40.4496	0.5493
410	455.45	- 45.45	2065.7025	4.5355
335	288.64	46.36	2149.2496	7.4461
180	174.55	5.45	29.7025	0.1702
20	20	0	0	0
55	61.36	- 6.36	40.4496	0.6592
				χ^2_{cal} **= 23.487**

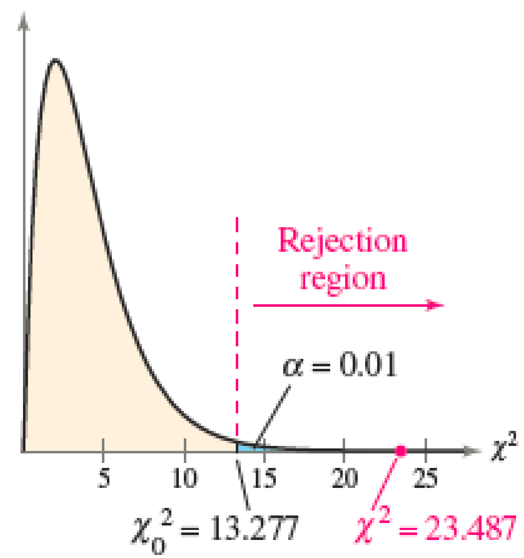
5. **Critical region:**

$\chi^2_{cal} > \chi^2_{tab}$

Where $\chi^2_{tab} = \chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.01,(2-1)(5-1)}$

$= 13.277$

$23.487 > 13.277$ (true)



6. Conclusion

There is enough evidence at the 1% level of significance to conclude that the variables *favorite way to eat ice cream* and *gender* are dependent.

Chi-Squared Table

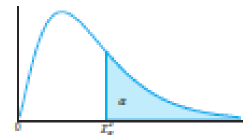


Table A.5 Critical Values of the Chi-Squared Distribution

	α									
v	**0.995**	**0.99**	**0.98**	**0.975**	**0.95**	**0.90**	**0.80**	**0.75**	**0.70**	**0.50**
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	0.102	0.148	0.455
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.647	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148	10.341
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	8.438	9.034	11.340
13	3.565	4.107	4.765	5.009	5.892	7.041	8.634	9.299	9.926	12.340
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	10.165	10.821	13.339
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	11.037	11.721	14.339
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	11.912	12.624	15.338
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	12.792	13.531	16.338
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	13.675	14.440	17.338
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	14.562	15.352	18.338
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	15.452	16.266	19.337
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	16.344	17.182	20.337
22	8.643	9.542	10.600	10.982	12.338	14.041	16.314	17.240	18.101	21.337
23	9.260	10.196	11.293	11.689	13.091	14.848	17.187	18.137	19.021	22.337
24	9.886	10.856	11.992	12.401	13.848	15.659	18.062	19.037	19.943	23.337
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	19.939	20.867	24.337
26	11.160	12.198	13.409	13.844	15.379	17.292	19.820	20.843	21.792	25.336
27	11.808	12.878	14.125	14.573	16.151	18.114	20.703	21.749	22.719	26.336
28	12.461	13.565	14.847	15.308	16.928	18.939	21.588	22.657	23.647	27.336
29	13.121	14.256	15.574	16.047	17.708	19.768	22.475	23.567	24.577	28.336
30	13.787	14.953	16.306	16.791	18.493	20.599	23.364	24.478	25.508	29.336
40	20.707	22.164	23.838	24.433	26.509	29.051	32.345	33.66	34.872	39.335
50	27.991	29.707	31.664	32.357	34.764	37.689	41.449	42.942	44.313	49.335
60	35.534	37.485	39.699	40.482	43.188	46.459	50.641	52.294	53.809	59.335

Table A.5 (continued) Critical Values of the Chi-Squared Distribution

	α									
v	**0.30**	**0.25**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.001**
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.466
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.515
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.321
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.124
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	15.119	15.984	16.985	19.812	22.362	24.736	25.471	27.688	29.819	34.527
14	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.124
15	17.322	18.245	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.698
16	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.791
18	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.819
20	22.775	23.828	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.314
21	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.796
22	24.939	26.039	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.619
26	29.246	30.435	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.051
27	30.319	31.528	32.912	36.741	40.113	43.195	44.140	46.963	49.645	55.475
28	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.994	56.892
29	32.461	33.711	35.139	39.087	42.557	45.722	46.693	49.588	52.335	58.301
30	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.702
40	44.165	45.616	47.269	51.805	55.758	59.342	60.436	63.691	66.766	73.403
50	54.723	56.334	58.164	63.167	67.505	71.420	72.613	76.154	79.490	86.660
60	65.226	66.981	68.972	74.397	79.082	83.298	84.58	88.379	91.952	99.608

Example: Survey responses (ordinal data) were collected on a 5-point Likert scale across three age groups. The contingency table and results are shown below. At a significance level of $\alpha = 0.05$, can we conclude that there is an association between **Age Group** and **5-Point Likert Scale Responses.**

Age Group	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
18-25	15	20	30	25	10
26-35	10	15	25	30	20
36-50	5	10	20	35	30

1. **We state our hypothesis as**:

H_0: The variables **Age Group** and **5-Point Likert Scale Responses** are independent.

H_1: The variables **Age Group** and **5-Point Likert Scale Responses** are dependent. (Claim)

2. **The level of significance is set** $\alpha = 0.05$.

3. **Test statistic to be used is**

$$\chi^2_{cal} = \sum \frac{(O_f - E_f)^2}{E_f}$$

4. Calculations

Observed Frequencies (O_f)

Age Group	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Row Totals
18-25	15	20	30	25	10	100
26-35	10	15	25	30	20	100
36-50	5	10	20	35	30	100
Column Totals	30	45	75	90	60	300

Expected Frequencies with Totals

Expected Frequency (E_f)

$$E_{r,c} = \frac{(\text{Sum of row } r)(\text{Sum of column } c)}{\text{Sample size}}$$

Age Group	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Row Totals
18-25	10.00	15.00	25.00	30.00	20.00	100.00
26-35	10.00	15.00	25.00	30.00	20.00	100.00
36-50	10.00	15.00	25.00	30.00	20.00	100.00
Column Totals	30.00	45.00	75.00	90.00	60.00	300.00

Detailed Chi-Square Computations (Part 1)

Observed (O_f)	Expected (E_f)	Chi-Square
15	10.0	2.5
20	15.0	1.67
30	25.0	1.0
25	30.0	0.83
10	20.0	5.0
10	10.0	0.0
15	15.0	0.0

Detailed Chi-Square Computations (Part 2)

Observed (O_f)	Expected (E_f)	Chi-Square
25	25.0	0.0
30	30.0	0.0
20	20.0	0.0
5	10.0	2.5
10	15.0	1.67
20	25.0	1.0
35	30.0	0.83
30	20.0	5.0
		χ^2_{cal} = 22.00

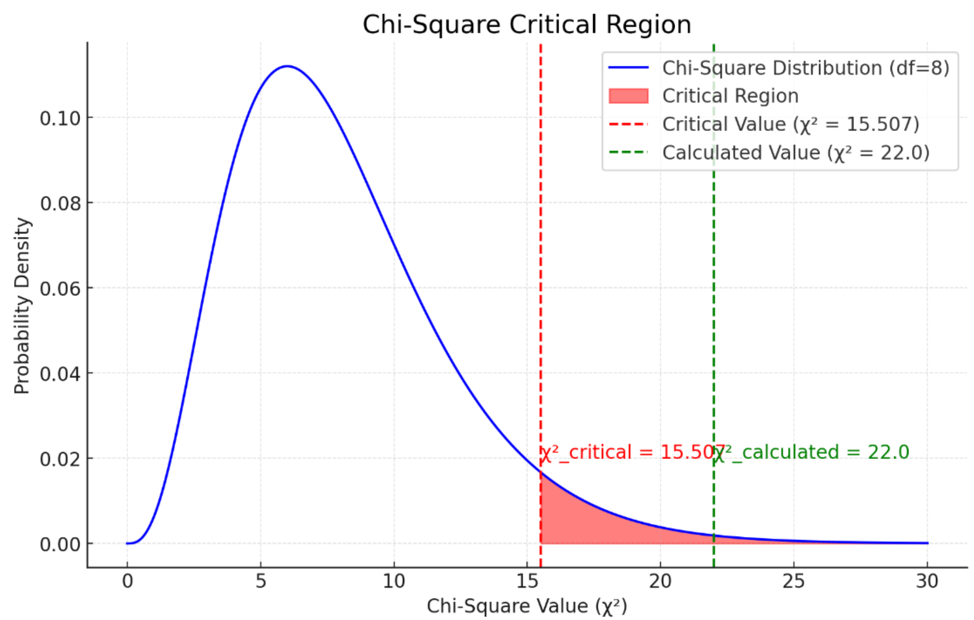
5. **Critical region:**

$\chi^2_{cal} > \chi^2_{tab}$

Where $\chi^2_{tab} = \chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,(3-1)(5-1)}$

$\chi^2_{tab} = \chi^2_{(0.05,8)} = 15.507$

22.00 > 15.507 (true)



6. Conclusion

There is enough evidence at the 5% level of significance to conclude that the variables **Age Group** and **5-Point Likert Scale Responses** are dependent.

Or

5. Critical region based on p-value

Chi-Square Statistic: 22.00

Degrees of Freedom: 8

P-Value: 0.0049

p-value is < 0.05 (TRUE)

6. Conclusion:

The p-value is < 0.05, indicating that we reject the null hypothesis and conclude there is an association.

```python
import numpy as np
from scipy.stats import chi2_contingency

# Define the contingency table
# Rows represent age groups, columns represent Likert scale responses

data = np.array([
    [15, 20, 30, 25, 10],  # Age group 18-25
    [10, 15, 25, 30, 20],  # Age group 26-35
    [5, 10, 20, 35, 30]    # Age group 36-50
])

# Perform the Chi-Square test of independence
chi2, p, dof, expected = chi2_contingency(data)
```

```python
# Print the results
print("Chi-Square Statistic:", chi2)
print("P-value:", p)
print("Degrees of Freedom:", dof)

# Decision based on significance level alpha = 0.05
alpha = 0.05
if p < alpha:
    print("Reject the null hypothesis: There is an association between age group and Likert scale responses.")
else:
    print("Fail to reject the null hypothesis: No significant association between age group and Likert scale responses.")
```

The results of the Chi-Square test are as follows:

Chi-Square Statistic: 22.00

P-value: 0.0049

Degrees of Freedom: 8

Conclusion:

Since the p-value (0.0049) is less than the significance level ($\alpha = 0.05$), we reject the null hypothesis. There is a statistically significant **association** between **Age Group** and **5-Point Likert Scale Responses**.