Categorical variables

- Represent types of data which may be divided into groups.
- Examples of categorical variables are race, sex, age group, and educational level.

Three assumptions of a chi square test

1. Random samples
2. Independent observations
3. The sample size is large enough such that all expected frequencies are greater than 1 and at least 80% are greater than 5.

- If your data fails the sample size assumption, an alternative test - Fisher's Exact test is used.

Hypotheses

- H_o: The variables are independent.
- H_A: The variables are not independent (meaning they are related)

Relevant Equations

- Degrees of freedom: (number of rows – 1)*(number of columns – 1)

- Expected counts for each cell: (row total*column total)/grand total

$$X^2 = \sum \frac{(observed - expected)^2}{expected}$$

Chi Square Test

- This test is used to determine if two categorical variables are independent or related to one another.
- If two categorical variables are independent, then the value of one variable does not change the probability distribution of the other
- If two categorical variables are related, then the distribution of one depends on the level the other.
- This test measures the differences in the observed conditional distribution of one variable across levels of the other, and compares it to the marginal (overall) distribution of that variable.

A real life case study

- I designed a product called Zolpa as a safe pain killer based on herbs
- In order to have it registered as a medicine, I needed to conduct a clinical trial
- I did this for a total of 37 patients
- We then had to analyze whether this product was effective or no using Chi Square test
- The patients were divided into two groups randomly, one was given the Zolpa capsules while the other was given placebo named Aploz
- The results were recorded as relief or no relief over the period of treatment

A real life case study

	Zolpa	Aploz	Total
Relief	13	8	21
No relief	5	11	16
Total	18	19	37

Hypotheses

- H_o: The relief and the capsules are independent (meaning the relief is not because of capsules)

- H_A: The relief and capsules are not independent (meaning they are related – the relief is because of the capsules)

Relevant Equations

- Degrees of freedom: (number of rows – 1)*(number of columns – 1)

- Expected counts for each cell: (row total*column total)/grand total

$$X^2 = \sum \frac{(observed - expected)^2}{expected}$$

Expected counts for each cell: (row total*column total)/grand total

$\frac{21\ x\ 18}{37}$	$\frac{21\ x\ 19}{37}$
$\frac{16\ x\ 18}{37}$	$\frac{16\ x\ 19}{37}$

Expected counts for each cell: (row total*column total)/grand total

$\frac{21\ x\ 18}{37} = 10.22$	$\frac{21\ x\ 19}{37} = 10.78$
$\frac{16\ x\ 18}{37} = 7.78$	$\frac{16\ x\ 19}{37} = 8.21$

Is everything more than 1?
Are at least 80% values more than 5?

Now let us use the formula

$\frac{(13-10.22)^2}{10.22} = 0.76$	$\frac{(8-10.78)^2}{10.78} = 0.72$
$\frac{(5-7.78)^2}{7.78} = 1$	$\frac{(11-8.21)^2}{8.21} = 0.94$

$$\sum \frac{(observed - expected)^2}{expected}$$

For Chi Square as per this calculation, we add all four values

$\chi^2 = 0.76 + 0.72 + 1 + 0.94 = 3.42$

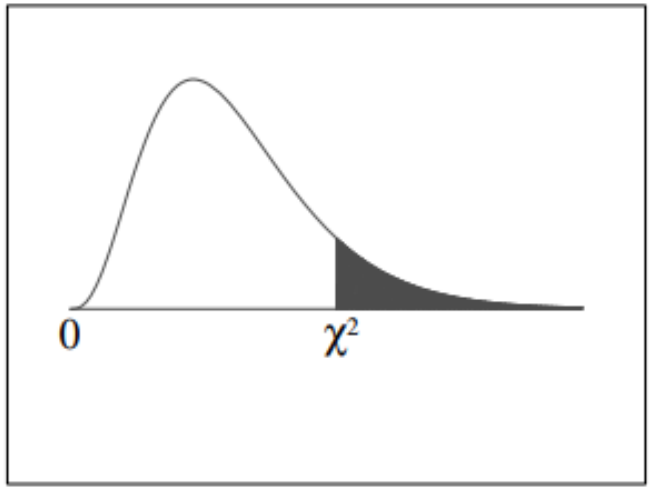
Critical Chi Square Value as per table

- First find out the degrees of freedom

- Degrees of freedom: (number of rows – 1)*(number of columns – 1)
- So, df = (2 – 1) * (2 – 1) = 1*1 = 1

We test at alpha value of 0.05

Chi-Square Distribution Table



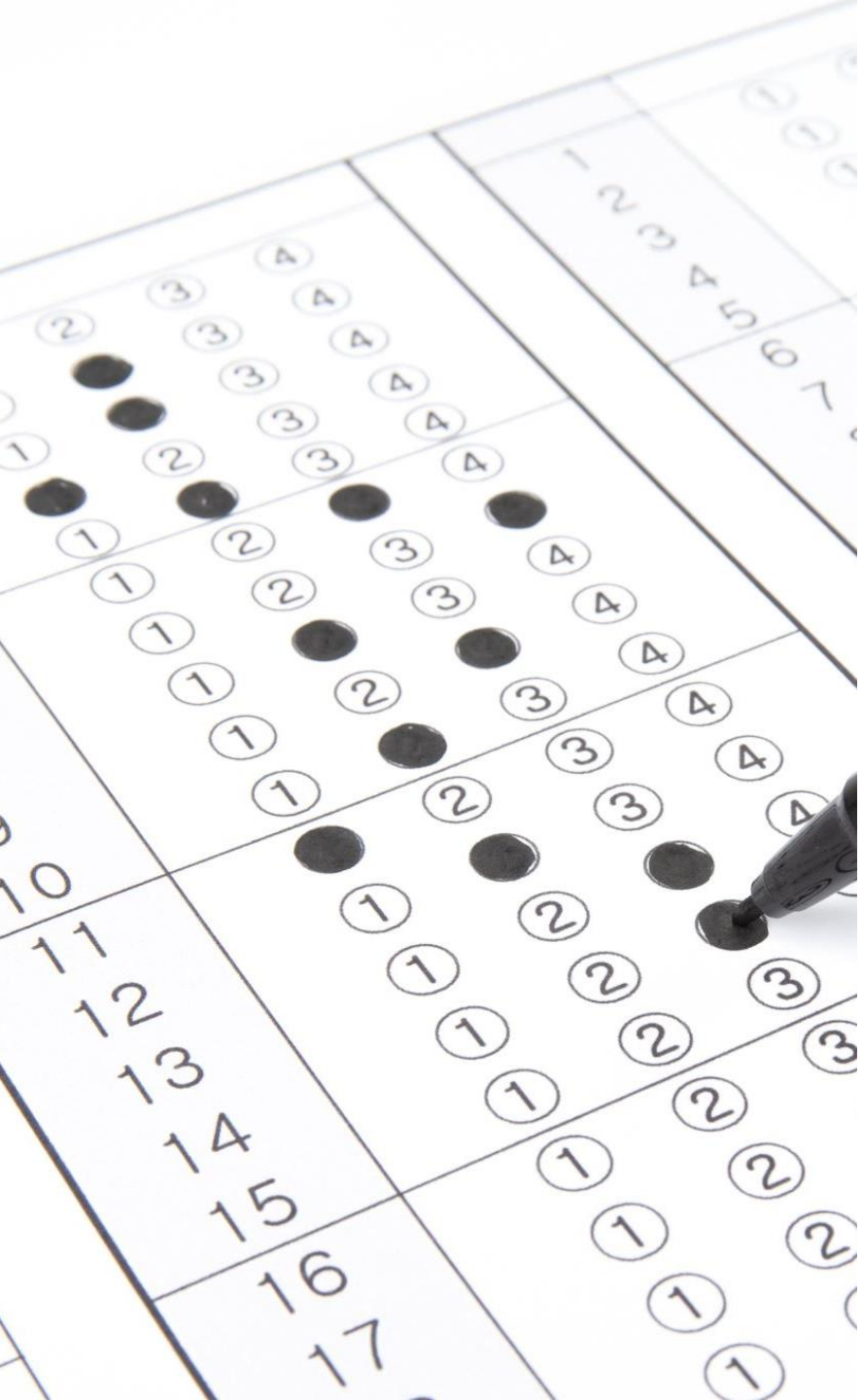
The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	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

Table value X^2 (*df*=1)=3.84, α = .05

Conclusion

- We find $3.42 < 3.84$
- Since the calculated value is less than the table value, we fail to reject the null hypothesis
- We failed to reject the null hypothesis and found evidence that treatment and relief are independent of each other



Steps in performing a chi-square test

1. State the null and alternative hypotheses.
2. Set the significance level.
3. Collect data and create a contingency table.
4. Calculate the expected frequencies.
5. Calculate the chi-square test statistic.
6. Compare the test statistic to the critical value or p-value.
7. Draw conclusions based on the results.

Characteristics of Chi-square Test and Limitations

Characteristics:

- Non-parametric test
  - (does not assume a specific distribution of data).
- Can be used with large sample sizes.
- Applicable to categorical data.

Limitations:

- Does not determine cause and effect.
- Assumes the observations are independent.
- Not suitable for small sample sizes.

Analysis of Variance and Covariance

- Analysis of Variance (ANOVA) is used to compare means across two or more groups or conditions.
- Covariance analysis (ANCOVA) is an extension of ANOVA that includes a covariate (continuous variable) in the analysis.
- ANOVA assesses whether there are significant differences between group means.
- ANCOVA allows for controlling the effects of the covariate on the dependent variable. – such as preexisting ability of an athlete and the effect of similar training given to all

Repeated Measures ANOVA

- Repeated measures ANOVA is used when you have the same measure that participants were rated on at more than two time points.
- With only two time points a paired t-test will be sufficient, but for more times a repeated measures ANOVA is required.
- For example, if you wish to track the progress of an exercise program on participants by weighing them at the beginning of the study and then every week after that for 6 weeks (a total of 7 time points) a repeated measures ANOVA would be required.

One-Way ANOVA

- A one-way (or single-factor) ANOVA can be run on sample data to determine if the mean of a numeric outcome differs across two or more independent groups.
- For example, suppose we wanted to know if the mean GPA of college students majoring in biology, chemistry, and physics differ.
- Note that we could not run a two-sample independent t-test because there are more than two groups.
- Hypotheses: (for k independent groups)
- Ho: The population means of all groups are equal, or $\mu 1 = \mu 2 = \ldots = \mu k$
  HA: At least one population mean is different, or $\mu i \neq \mu j$ for some i, j