χ^2 Test and Inference for categorical data

부산대학교 정보·의생명공학대학

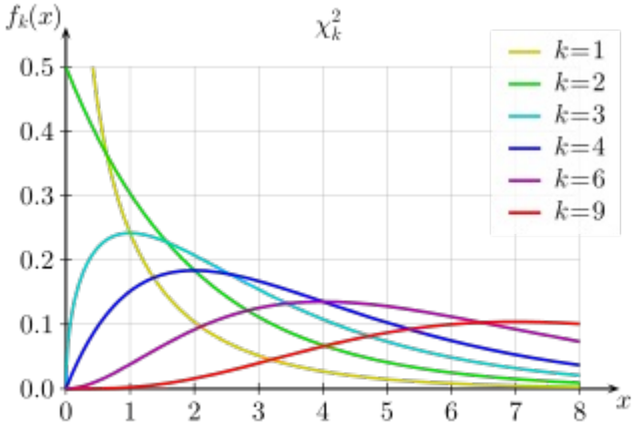
정보컴퓨터공학부

χ^2 Distribution

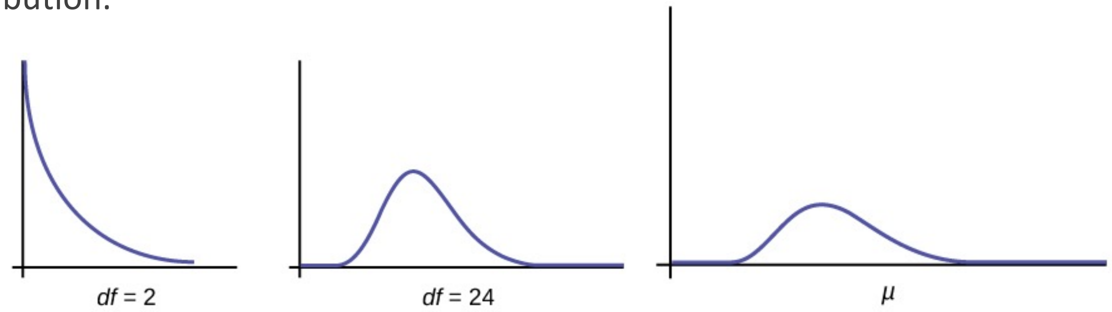
- The notation for the chi-square distribution is: χ^2_{df}
  - Where df = degrees of freedom
  - For the χ^2 distribution, the population mean is $\mu =$df and the population standard deviation is $\sigma = \sqrt{2(df)}$

- The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

$$\chi^2_k = (Z_1)^2 + (Z_2)^2 + \cdots + (Z_k)^2$$

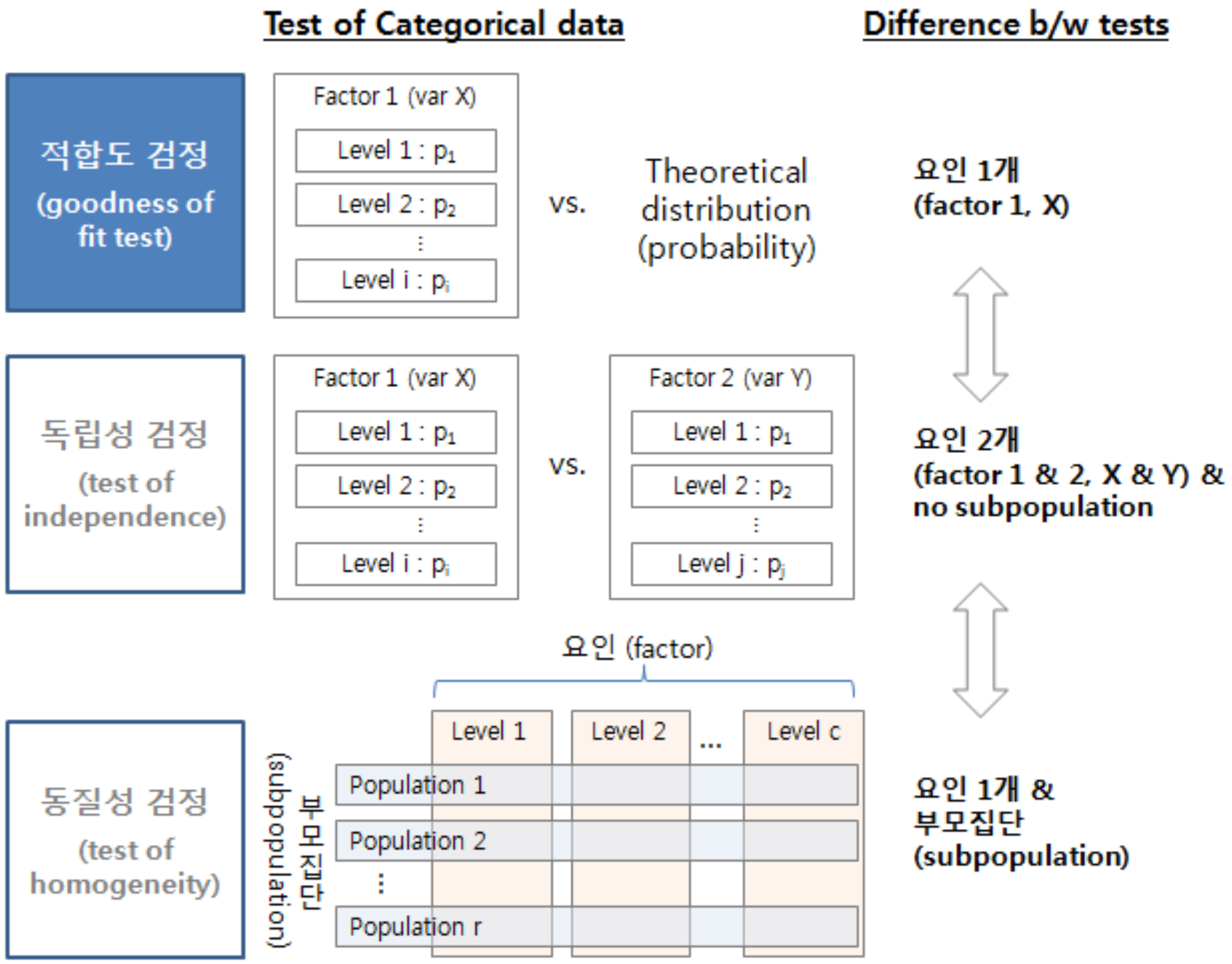


  - The curve is nonsymmetrical and skewed to the right
  - There is a different chi-square curve for each df.
  - The test statistic for any test is always greater than or equal to zero.
  - When $df > 90$, the chi-square curve approximates the normal distribution.
    - $X \sim \chi^2_{1000} \sim N\left(df, \left(\sqrt{2(df)}\right)^2\right) \sim N(1000, 44.7^2)$
  - The mean, μ, is located just to the right of the peak.



3 Major applications of the chi-square distribution

1. Goodness-of-Fit test (적합도 검정)
   - Determines if data fit a particular distribution
   - Ex) Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency?
2. Test of independence (독립성 검정)
   - Determines if events are independent
   - Example) Are the types of movies people preferred different across different age groups?
3. Test of a Homogeneity (동질성 검정)
   - Determines whether two populations follow the same unknown distribution.
   - Example) Is a coffee machine dispensing approximately the same amount of coffee each time?



Test of Categorical data

Difference b/w tests

적합도 검정 (goodness of fit test)

Factor 1 (var X)

Level 1 : p_1

Level 2 : p_2

⋮

Level i : p_i

vs.

Theoretical distribution (probability)

요인 1개 (factor 1, X)

독립성 검정 (test of independence)

Factor 1 (var X)

Level 1 : p_1

Level 2 : p_2

⋮

Level i : p_i

vs.

Factor 2 (var Y)

Level 1 : p_1

Level 2 : p_2

⋮

Level j : p_j

요인 2개 (factor 1 & 2, X & Y) & no subpopulation

동질성 검정 (test of homogeneity)

요인 (factor)

Level 1

Level 2

...

Level c

부모집단 (subpopulation)

Population 1

Population 2

⋮

Population r

요인 1개 & 부모집단 (subpopulation)

[R 분석과 프로그래밍] http://rfriend.tistory.com

Goodness of Fit test

- Test the data "fit" a particular distribution or not.

- Example) Day of the Week Employees were most Absent
  - Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe ***that employees are absent equally during the week***.
  - Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in the following table

	Monday	Tuesday	Wednesday	Thursday	Friday
# of Absence **Observed**	15	12	9	9	15
# of Absence **Expected**					

  - For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5% significance level.

- Null Hypothesis ?
  - H_0 :The absent days occur with equal frequencies
- Expected Counts and the null hypothesis
  - Use that hypothesized distribution to calculate the expected counts for each value of the variable.

- The test statistic for a goodness-of-fit test

$$\sum_k \frac{(O-E)^2}{E}$$

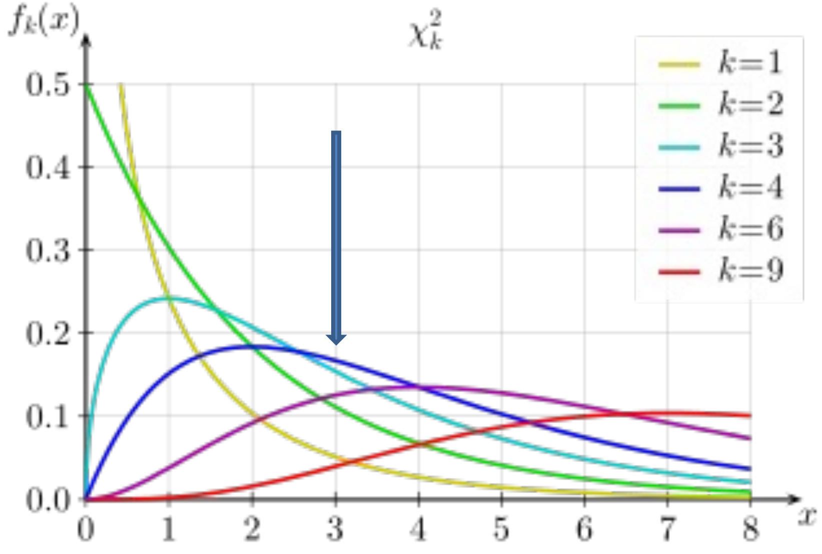
  - O = observed values
  - E = expected values
  - k = the number of different data cells or categories
  - **Degree of freedom = $k-1$**

$$\sum_k \frac{(\boldsymbol{O}-\boldsymbol{E})^2}{\boldsymbol{E}} = \frac{(15-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(15-12)^2}{12}$$

$$= \frac{9}{12} + \frac{9}{12} + \frac{9}{12} + \frac{9}{12} = \frac{36}{12} = \mathbf{3.0}$$

Goodness of Fit test

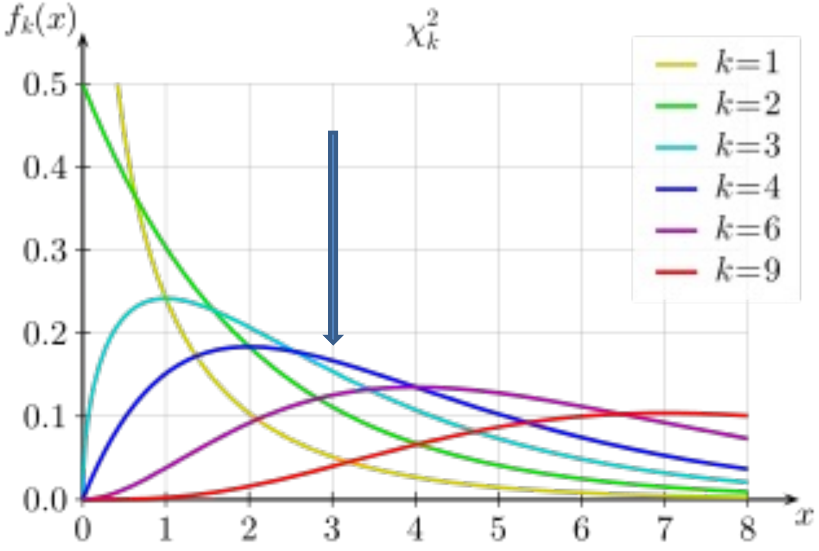
- Employee Absence Example)
  - Degree of freedom = 4, χ^2_4
  - Test statistic = 3.0
  - p-Value = $\Pr(\chi^2_4 > 3.0)$?
  - Tabulation Method
    - Determine if p-Value $< \alpha = 0.05$
    - Reject H_0 or not ?
  - Python/Colab
    - scipy.stats.chi2 : chi square distribution, cdf()
    - scipy.stats.chisquare()



Degrees of freedom (df)	χ^2 value										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
p-value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Goodness of Fit test

- Employee Absence Example)
  - Degree of freedom = 4, χ_4^2
  - Test statistic = 3.0
  - p-Value = $\Pr(\chi_4^2 > 3.0)$?
  - Tabulation Method
    - Determine if p-Value $< \alpha = 0.05$
    - Reject H_0 or not ?
  - Python/Colab
    - scipy.stats.chi2 : chi square distribution, cdf()
    - scipy.stats.chisquare()



```
import numpy as np
import scipy.stats as stats

val_ob = np.array([15, 12, 9, 9, 15])
val_ex = np.array([12, 12, 12, 12, 12])

chi_stat = sum( (val_ob-val_ex)**2/val_ex)
print('chi-sqaure statistic = ',chi_stat )
print('p-value = ', 1-stats.chi2(df=4).cdf(chi_stat))

print(stats.chisquare(val_ob,val_ex))

----------------------------------
chi-sqaure statistic = 3.0
p-value = 0.5578254003710748
Power_divergenceResult(statistic=3.0, pvalue=0.5578254003710748)
```

Conditions for a goodness-of-fit test

Random:

- The data came from a random sample from the population of interest, or a randomized experiment.
- If we sample without replacement, our sample size should be less than **10% of the population** so we can assume independence between members in the sample. We don't need to check the 10% condition in experiments that involve random assignment, since we're not sampling in those cases.

Large counts

- The large counts condition says **that all expected counts need to be at least 5**.
- There are **no conditions** attached **to the observed counts**.

Quiz

❖ Conditions for a goodness-of-fit test

- Whitney's town has 10,000 residents and three neighborhoods. These are the percentages of each neighborhood's area relative to the town's total area:

	A	B	C	Total
# of Area	55%	37%	8%	100%

- Whitney wants to test if the distribution of the neighborhoods' populations matches the distribution of the neighborhoods' areas. She plans to ask a sample of residents what neighborhood they live in. She'll carry out a χ^2 goodness-of-fit test on the resulting data.

- Which of these are conditions for carrying out this test? Choose 3.

1. She observes each neighborhood at least 5 times.
2. She expects each neighborhood to appear at least 5 times.
3. She takes a random sample of residents
4. She samples 1000 residents at most.

❖ Expected Counts

- Preston teaches at an international school. After reading an article about the distribution of the world's population by continent, he wanted to test if the distribution of students in his school was similar. He collected information about all 320 students in his school. Here are the results:

	Asia	Africa	Europe	North America	South America	Australia
Reported % of world population	61%	15%	11%	7%	5.5%	0.5%
# of students Observed	82	10	93	95	30	10
# of students Expected						

- Preston wants to perform a χ^2 goodness-of-fit test to determine if these results suggest that the distribution of students doesn't match the reported distribution.

- What is the expected count of students from Australia in Preston's group? $320 \times 0.5\% = 1.6$

Quiz

❖ Test-statistic and P-value

- Elsa is investigating rider complaints that a certain bus route is only on time 60% of the time, is early 25% of the time, and is late the remaining 15% of the time. She took a random sample of 45 times and recorded whether the bus was on time, early, or late.

- Here are her results: She wants to use these results to carry out a χ^2 goodness-of-fit test to determine if the distribution of timings for the bus route differs from the proportions the riders reported

	On Time	Early	Late
Observed	33	8	4
Expected			

- What are the values of the test statistic and p-value for Elsa's test?

  1. $\chi^2 = 2.708, 0.05 < p\text{-value} < 0.10$
  2. $\chi^2 = 2.708,\ p\text{-value} > 0.25$
  3. $\chi^2 = 3.392,\ \ 0.15 < p\text{-value} < 0.20$
  4. $\chi^2 = 3.392,\ p\text{-value} > 0.25$

❖ Conclusion in a goodness of fit test

- Salma operates a ramen restaurant. Historically, half of customers order pork ramen, and the other half of customers are split evenly between chicken and vegetarian ramen. Salma modified the chicken and vegetarian recipes, and she wonders if customers' ordering habits will change. Here are results from a sample of 32 orders after launching the new recipes along with a χ^2 goodness-of-fit test:

Type of ramen	Pork	Chicken	Vegetarian
Historical distribution	50%	25%	25%
Observed	15	4	13
Expected	16	8	8
Components	0.06	2	3.13

- $\boldsymbol{\chi^2 = 5.19, DF = 2, p}$**-value = 0.075**

- Assume that all conditions for inference were met. At the $\boldsymbol{\alpha = 0.10}$ significance level, what should Salma conclude about the proportions of each type of order? Choose 1 answer:

  1. There is sufficient evidence to conclude that the proportions of each type of order have changed.
  2. We don't have enough evidence to conclude that the proportions of each type of order have changed.
  3. There is sufficient evidence to conclude that the proportions from before the new recipes are still true.
  4. No conclusions can be made from this test, because one of the observed counts is too low for the calculations to be accurate.

χ^2 test with Contingency Table

❖ Contingency Table (분할표)

- also known as a cross tabulation or crosstab, is a type of table in a matrix format that displays the frequency distribution of the variables
- Example)
- 1) Tom and Jane wondered if they had similar tastes in music. They each took a random sample of 50 songs from their music collection and categorized the songs by genre. Here is a summary of the data
- 2) Elliot was curious if there was a relationship between a student's gender and what superpower they'd prefer to have. He obtained data from a random sample of 206 students. Here is a summary of the data

	Tom	Jane
Hip hop	22	12
Alternative	14	12
Pop	8	21
Other	6	5

	Female	Male
Fly	19	32
Freeze time	17	23
Invisibility	23	13
Telepathy	53	26

❖ Filling out frequency table for independent events

- Example)
  - One rainy Saturday morning, Adam woke up to hear his mom complaining about the house being dirty. “Mom is always grouchy when it rains,” Adam’s brother said to him.
  - So Adam decided to figure out if this statement was actually true. For the next year, he charted every time it rained and every time his mom was grouchy. What he found was very interesting – rainy days and his mom being grouchy were entirely ***independent events***. Some of his data are shown in the table below.
  - Fill in the missing values from the frequency table.

	Rainy	Not Rainy	Row Total
Grouchy	A	B	73
Not Grouchy	C	D	292
Column Total	35	330	365

- A = $35\times\left(\frac{73}{365}\right) = 7$
- B = 66, C = 28, D=264

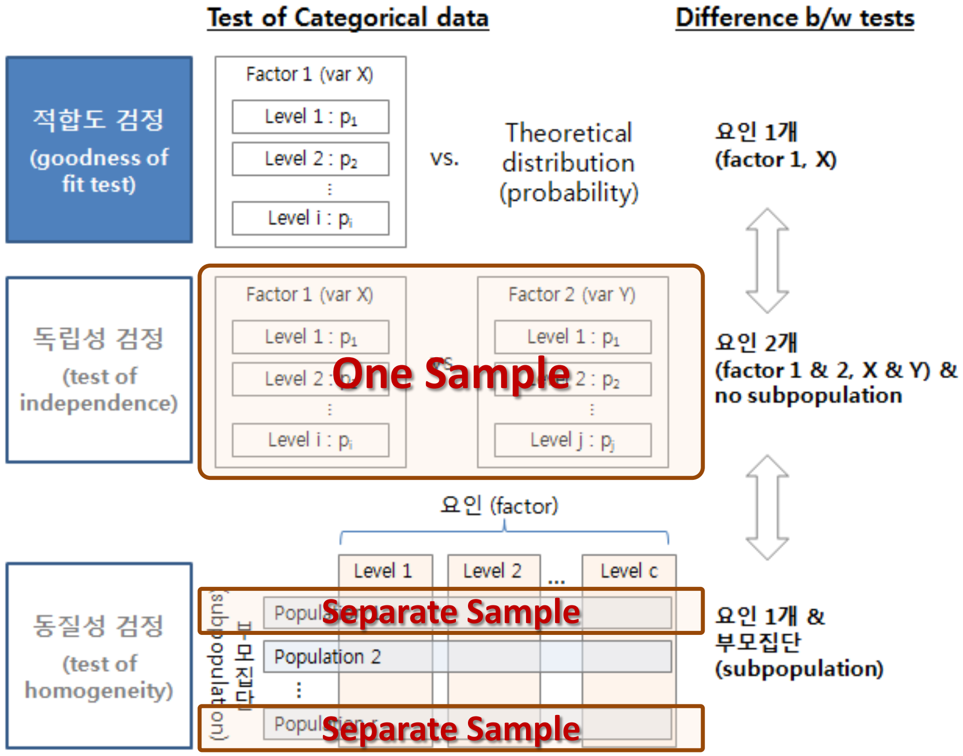
χ^2 test for Contingency Table

Separate, independent samples or groups – Test of Homogeneity

- A chi-square test can help us when we want to know whether different populations or groups are alike with regards to the distribution of a variable. Our hypotheses would look something like this:
- H_0 : The distribution of a variable is the same in each population or group.
  H_1 : The distribution of a variable differs between some of the populations or groups.
- We call this the chi-square test for **homogeneity**

One sample or group – Test of Independence

- A chi-square test can help us see whether individuals from a sample who belong to a certain category are more likely than others in the sample to also belong to another category. Our hypotheses would look something like this:
- H_0 : There is no association between the two variables (they are independent).
  H_1 : There is an association between the two variables (they are not independent).
- We call this the chi square test of **association** (or **independence**).



Test of Categorical data

Difference b/w tests

적합도 검정 (goodness of fit test)

Factor 1 (var X)
Level 1 : p_1
Level 2 : p_2
⋮
Level i : p_i

vs.

Theoretical distribution (probability)

요인 1개 (factor 1, X)

독립성 검정 (test of independence)

Factor 1 (var X)
Level 1 : p_1
Level 2 : p_2
⋮
Level i : p_i

Factor 2 (var Y)
Level 1 : p_1
Level 2 : p_2
⋮
Level j : p_j

One Sample

요인 2개 (factor 1 & 2, X & Y) & no subpopulation

동질성 검정 (test of homogeneity)

요인 (factor)

Level 1
Level 2
...
Level c

부모집단 (subpopulation)

Population 2
⋮

Separate Sample

Separate Sample

요인 1개 & 부모집단 (subpopulation)

[R 분석과 프로그래밍] http://rfriend.tistory.com

Which Test ?

Example 1)

- Tom and Jane wondered if they had similar tastes in music. They each took a random sample of 50 songs from their music collection and categorized the songs by genre. Here is a summary of the data

	Tom	Jane
Hip hop	22	12
Alternative	14	12
Pop	8	21
Other	6	5

- Test of Homogeneity vs Test of Independence ?

Example 2)

- Elliot was curious if there was a relationship between a student's gender and what superpower they'd prefer to have. He obtained data from a random sample of 206 students. Here is a summary of the data

	Female	Male
Fly	19	32
Freeze time	17	23
Invisibility	23	13
Telepathy	53	26

- Test of Homogeneity vs Test of Independence ?

Expected Counts in Contingency Table

- Example)

  - Margot surveyed a random sample of 180 people from the United States about their favorite sports to watch. Then she sent separate, similar, survey to a random sample of 180 people from the United Kingdom. Here are the results:

Favorite sport to watch	United States Observed	United States Expected	United Kingdom Observed	United Kingdom Expected	Total
Basketball	60		51		111
Football	67		14		81
Soccer	28		86		114
Tennis	25		29		54
Total	180		180		360

  - Margot wants to perform a χ^2 **test of homogeneity** on these results. What is the expected count for the cell corresponding to people from the United Kingdom whose favorite sports to watch is tennis? $= 180 \times \left(\frac{54}{360}\right) = 27$

- $$expected = \frac{row\ total \times column\_total}{table\ total}$$

Test statistic and P-value in χ^2 tests with 2-tables

❖ The test statistic for a χ^2 test with 2-tables

$$\sum_{(i,j)} \frac{(O-E)^2}{E}$$

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table
- **Degree of freedom = $(i-1)\cdot(j-1)$**

❖ Example

Favorite sport to watch	US Observed	US Expected	UK Observed	UK Expected	Total
Basketball	60	55.5	51	55.5	111
Football	67	40.5	14	40.5	81
Soccer	28	57	86	57	114
Tennis	25	27	29	27	54
Total	180	180	180	180	360

- H_0 : The distribution of "Favorite sport to watch" is the same
- $i = 4, j = 2$
- Degree of freedom = $(4-1)\cdot(2-1) = 3$
- Test Statistic

$$\sum_{(i,j)} \frac{(O-E)^2}{E} = \frac{(60-55.5)^2}{55.5} + \frac{(51-55.5)^2}{55.5} + \frac{(67-40.5)^2}{40.5} + \frac{(14-40.5)^2}{40.5} + \frac{(28-57)^2}{57} + \frac{(86-57)^2}{57} + \frac{(25-27)^2}{27} + \frac{(29-27)^2}{27} = 65.214$$

- p-Value = 4.51×10^{-14}
- Reject H_0 or not ?

χ^2 test in Colab

- Scipy.stats.chisquare(f_obs, f_exp=None, ddof=0)
  - Parameters
  - f_obs : array_like
    - Observed frequencies in each category.
  - f_exp : array_like, optional
    - Expected frequencies in each category.
    - By default the categories are assumed to be equally likely.
  - ddof : int, optional
    - "Delta degrees of freedom": adjustment to the degrees of freedom for the p-value. The p-value is computed using a chi-squared distribution with k - 1 - ddof degrees of freedom, where k is the number of observed frequencies. The default value of ddof is 0.
    - $\boldsymbol{df = k - 1 - ddof}$
    - **In χ^2 test with 2-table**
    - $\boldsymbol{df = (i-1) \cdot (j-1), k = i \cdot j}$
    - $\boldsymbol{\therefore ddof = i + j - 2}$

```
import numpy as np
import scipy.stats as stats

val_ob = np.array([60,51,67,14,28,86,25,29])
val_ex = np.array([55.5, 55.5, 40.5, 40.5, 57, 57, 27,27])

print(stats.chisquare(val_ob,val_ex, ddof=4))

---------------------------------------

Power_divergenceResult(statistic=65.2138103015296,
pvalue=4.514644867902917e-14)
```

Quiz

❖ Shopping Complex Transportation

- The owners of a large shopping complex wondered how their customers traveled to the complex. They surveyed a random sample of 100 customers. Here are the outcomes and partial results of a χ^2 test (expected counts appear below observed counts).
- Which test would they do using this data?

1. χ^2 test of independence.

2. χ^2 test of Homogeneity.

	Drive	**Public transit**	**Other**	**Total**
Made a purchase	30	25	15	70
Expected	35	21	14	
No purchase	20	5	5	30
Expected	15	9	6	
Total	50	30	20	100

- Assume that all conditions for inference were met. What are the values of the test statistic and p-value for their test? Choose 1 answer:

1. $\chi^2 = 5.159, 0.05 < P < 0.10$

2. $\chi^2 = 5.159, 0.15 < P < 0.20$

3. $\chi^2 = 6.19\ 0.025 < P < 0.05$

4. $\chi^2 = 6.19, 0.10 < P < 0.15$

- Would you reject H_0 if $\alpha = 0.05$?

Quiz

❖ Smartwatch Alert Service

- A company is testing their new smartwatch. They wonder if alerts that prompt people to exercise cause people to exercise more than they would without the alerts. They recruit 200 subjects and randomly assign each of them to either receive the alerts or not. Here are amounts of daily activity and partial results of a χ^2 test (expected counts appear below observed counts):

- Which test would they do using this data?

1. χ^2 test of independence.
2. χ^2 test of Homogeneity.

	Alert	No Alerts	Total
0-29	48	64	112
Expected	56	56	
30-59	33	27	60
Expected	30	3	
60+	19	9	28
Expected	14	14	
Total	100	100	200

- Assume that all conditions for inference were met. What are the values of the test statistic and p-value for their test? Choose 1 answer:

1. $\chi^2 = 3.229, 0.05 < P < 0.10$
2. $\chi^2 = 3.229, 0.15 < P < 0.20$
3. $\chi^2 = 6.458, 0.025 < P < 0.05$
4. $\chi^2 = 6.458, 0.05 < P < 0.10$

- Would you reject H_0 if $\alpha = 0.05$?

Quiz - Making Conclusion in chi-square test for two-way tables

❖ 2 Authors

- A linguist is studying how two authors use parts of speech in their writing. They take separate random samples of a few essays written by each author, and they tally how many times the authors use each part of speech. Here is a summary of the words in each sample and the results from a χ^2 test:

	Author A	Author B
Adverbs	60	74
Expected	61.42	72.58
Adjectives	176	199
Expected	171.9	203.1
Verbs	179	220
Expected	182.9	216.1
Nouns	582	685
Expected	580.78	686.22

- **$\chi^2 = 0.4, DF = 3, p$-value = 0.94**

- Assume that all conditions for inference were met. At the $\alpha = 0.05$ significance level, what is the most appropriate conclusion to draw from this test? Choose 1 answer:

1. This is convincing evidence of an association between the parts of speech and these authors.
2. This isn't enough evidence to say there is an association between the parts of speech and these authors.
3. This is convincing evidence that the distribution of the parts of speech differs between these authors.
4. This isn't enough evidence to say that the distribution of the parts of speech differs between these authors.