8.1 One-way analysis of variance (one-way ANOVA)

Comparing three or more populations

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

The unpaired t-test determines whether a statistically significant difference exists between the means of two populations. However, many situations require that means be compared among three or more populations. Ex: A researcher wants to classify iris species based on sepal length by using a method called k-means clustering. As a first step, the researcher checks whether the mean sepal length differs among three species of iris: *setosa*, *virginica*, and *versicolor*. A possible method to compare the means is to perform three unpaired t-tests: one between *setosa* and *versicolor*, another between *setosa* and *virginica*, and finally between *versicolor* and *virginica*. Although the details are beyond the scope of the material, the probability of rejecting the null hypothesis that no significant difference in population means exists, when using multiple t-tests is 14%. Thus, a different approach is needed.

One-way analysis of variance (one-way ANOVA) controls for the errors associated with comparing multiple population means. ***One-way analysis of variance (one-way ANOVA)*** determines whether a statistically significant difference exists among the means of three or more populations. Equivalently, ANOVA tests for an association between a categorical predictor variable and a response variable. Ex: In the iris study, the predictor variable is the type of species and the response variable is sepal length. Data scientists and statisticians often refer to a categorical predictor variable as a factor and a possible value of a factor as a level. A factor can be a continuous variable partitioned into intervals commonly referred to as bins. Ex: The factor in the iris example, iris type, has three levels: *setosa*, *virginica*, and *versicolor*. If the factor only has two levels, then ANOVA is equivalent to a two-sample t-test with equal variances.

PARTICIPATION ACTIVITY 8.1.1: Comparing three or more populations.

1) Which test should be used when comparing the means of three or more populations?

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

- one-sample t-test
- unpaired t-test
- ANOVA

2) What is a level?

- categorical variables
- a possible value for a factor

- numerical variables partitioned into bins

One-way ANOVA

A model can certainly have multiple predictor variables. However, the material on ANOVA will only cover models with one predictor variable. A ***one-way ANOVA*** compares the means of three or more groups of one predictor variable.

Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

The one-way ANOVA hypothesis test follows the same process as previously discussed tests. The null hypothesis for a one-way ANOVA is that all of the group means are equal. Caution should be exercised when stating the alternative hypothesis because the negation of the null hypothesis does not say that all group means are unequal. Instead, the alternative hypothesis should state that two groups with unequal means exist. The rest of the hypothesis test involves finding the F-statistic and the p-value to make a decision based on a significance level.

The underlying assumptions for the one-way ANOVA are given below.

Assumptions for one-way ANOVA

- *Independence. Samples should be independent and drawn randomly.*
- *Normality. The underlying distribution of the populations from which the samples are drawn should be approximately normal.*
- *Homogeneity. The variances of the population distributions should be equal.*

Procedure 8.1.1: One-way ANOVA.

Given k groups taken from independent populations,

1. Set the null and alternative hypotheses

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k \quad H_a: \mu_i \neq \mu_j, \text{ for some } i \neq j$$

Completed 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

where μ_1 and μ_2 are means from independent populations.

2. Use statistical software to find the test-statistic

$$F = \frac{\text{between-group variance}}{\text{within group variance}}$$

3. Use statistical software to find the p-value that corresponds to F.

4. Make a decision given a previously selected significance level α, typically 0.05.
    - If the p-value is less than the significance level, sufficient evidence exists to reject the null hypothesis H_0 in favor of the alternative hypothesis H_a.
    - If the p-value is greater than or equal to the significance level, insufficient evidence exists to reject the null hypothesis H_0.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Python-Practice 8.1.1: ANOVA.

The `f_oneway()` function performs a one-way ANOVA. The function takes a number of lists of data as parameters, and returns the F-statistic and the p-value. The scipy.stats and statsmodels.formula.api libraries must be imported to use `f_oneway()`.

```
# The Exam Score dataset includes scores obtained in 4 exams in a class.
# Perform a hypothesis test to determine if the mean scores of the exams
# are different. Use the 5% level of significance.

import pandas as pd
import scipy.stats as st
scores = pd.read_csv('http://data-analytics.zybooks.com/ExamScores.csv')

# Statistics of each exam
exam1_score = scores[['Exam1']]
exam2_score = scores[['Exam2']]
exam3_score = scores[['Exam3']]
exam4_score = scores[['Exam4']]

print(st.f_oneway(exam1_score, exam2_score, exam3_score, exam4_score))
```

```
F_onewayResult(statistic=array([ 3.85696089]), pvalue=array([ 0.01034867]))
```

If the data is labeled according to group, `ols()` can also be used to generate an ANOVA table, which also gives the F-statistic and the corresponding p-value.

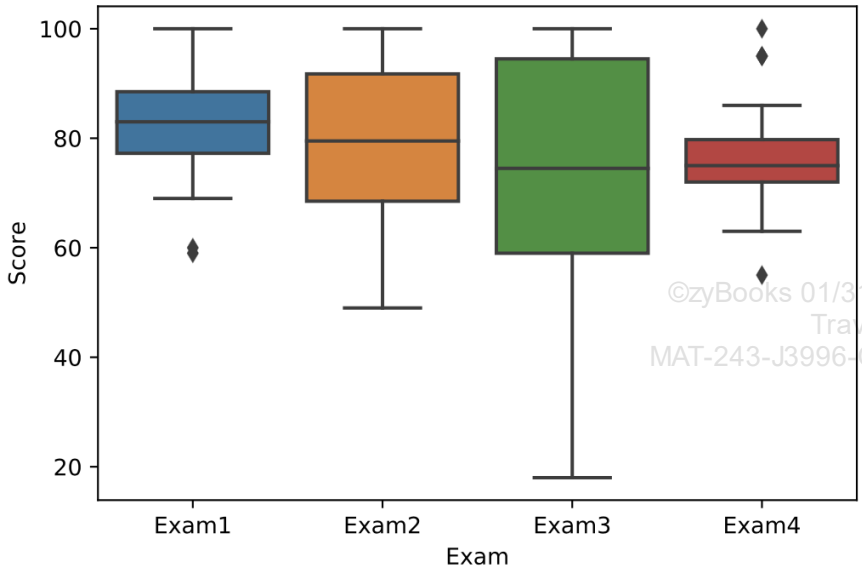
```
import statsmodels.api as sm
import pandas as pd
from statsmodels.formula.api import ols
df = pd.read_csv('http://data-
analytics.zybooks.com/ExamScoresGrouped.csv')
mod = ols('Scores ~ Exam',data=df).fit()
aov_table = sm.stats.anova_lm(mod, typ=2)
print(aov_table)
```

```
              sum_sq     df
F    PR(>F)
Exam        2400.735
3.0  3.856961  0.010349
Residual  40666.220
196.0        NaN     NaN
```

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Below is the code to generate the box plots for grouped data.

```
import pandas as pd
import seaborn as sns
df = pd.read_csv('http://data-analytics.zybooks.com/ExamScoresGrouped.csv')
sns.boxplot(x="Exam", y="Scores", data=df)
```



©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Run example

Example 8.1.1: Mean of exam scores.

A teacher believes that the exams created for the class varies in difficulty because of the differences in mean exam scores. Does sufficient evidence exist at the $\alpha = 0.01$ level to support the teacher's belief that the exams scores have different means? Use the ANOVA table below.

	sum_sq	df	F	PR(>F)
Exam	2400.735	3.0	3.856961	0.010349
Residual	40666.220	196.0	NaN	NaN

Solution

The null hypothesis is that the mean exam scores are all the same. The alternative hypothesis is that the means of at least two exam scores are different. Mathematically,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 H_a: \mu_i \neq \mu_j \text{ for some } i \neq j$$

The p-value that corresponds to $F = 3.857$ is $P(F \geq 3.857) = 0.0103$. Since the p-value is greater than the significance level $(0.0103 > 0.01)$, the null hypothesis is not rejected. That is, at the $\alpha = 0.01$ significance level, insufficient statistical evidence exists to support the claim that the mean exam scores are different.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Analysis

The alternative hypothesis to an F test is that at least two of the means are unequal. Although the means of all exam scores could be all be statistically different, the F test

cannot determine which of the exams have statistically different means.

PARTICIPATION ACTIVITY 8.1.2: Mean sepal lengths of iris species.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

A researcher wanted to check the possibility of running a classification algorithm based on iris sepal lengths. Among other things the researcher checks is whether the means of the sepal lengths for each of the factor levels are actually different.

Does sufficient evidence exist to conclude that the mean sepal length of the three iris species are not the same? Use the ANOVA table below.

	sum_sq	df	F	PR(>F)
Species	63.212133	2.0	119.264502	1.669669e-31
Residual	38.956200	147.0	NaN	NaN

Run example

1) What is the null hypothesis H_0?

- $\mu_i \neq \mu_j$ for some iris species $i \neq j$
- $\mu_{set} = \mu_{virg}$
- $\mu_{set} = \mu_{virg} = \mu_{vers}$

2) What is the alternative hypothesis H_a?

- $\mu_i \neq \mu_j$ for some iris species $i \neq j$
- $\mu_{set} \neq \mu_{vers} \neq \mu_{virg}$
- $\mu_{virg} \neq \mu_{vers}$

3) What is the conclusion of the test?

- Fail to reject H_0
- Reject H_0

Post-hoc tests

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

If the null hypothesis is not rejected, then no further work is necessary. However, if the null hypothesis is rejected, further analysis is required because the F-test does not determine which groups have different means. ***Post-hoc analysis*** determines which groups have different means, which group has the highest or lowest mean, and other relationships between the groups.

The ***Tukey Honestly Significant Difference (HSD)*** procedure gives the 95% confidence intervals for the mean difference between pairwise groups and determines which mean difference is statistically

significant.

If 0 falls in the confidence interval, then the difference between the means is not statistically significant. In other words, the null hypothesis that the means of the two groups are the same should not be rejected.

Python-Practice 8.1.2: Tukey's HSD.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

The `MultipleComparison()` function take the data frame column that contains the values and the data frame column that contains the levels as inputs and builds the models. The `tukeyhsd()` function displays the 95% Tukey's confidence intervals. The statsmodels.stats.multicomp library must be imported to use these functions.

```
import pandas as pd
from statsmodels.stats.multicomp import (pairwise_tukeyhsd, MultiComparison)
df = pd.read_csv('http://data-analytics.zybooks.com/ExamScoresGrouped.csv')
mod = MultiComparison(df['Scores'], df['Exam'])
print(mod.tukeyhsd())
```

```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
=============================================
group1 group2 meandiff  lower    upper  reject
---------------------------------------------
Exam1  Exam2    -3.3   -10.7652  4.1652 False
Exam1  Exam3   -9.36   -16.8252 -1.8948  True
Exam1  Exam4    -6.2   -13.6652  1.2652 False
Exam2  Exam3   -6.06   -13.5252  1.4052 False
Exam2  Exam4    -2.9   -10.3652  4.5652 False
Exam3  Exam4    3.16   -4.3052  10.6252 False
---------------------------------------------
```

Run example

Example 8.1.2: Determining which exams have statistically different means.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Suppose the ExamScores dataset is tested at a significance level of $\alpha = 0.05$. Thus, a post-hoc test should be performed to determine which two means are statistically equal because the null hypothesis that the means are statistically the same is rejected. Which two groups have statistically different means? Use the output below of the Tukey's HSD test.

```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
============================================
group1 group2 meandiff  lower    upper  reject
--------------------------------------------
Exam1  Exam2    -3.3   -10.7652  4.1652 False
Exam1  Exam3   -9.36   -16.8252 -1.8948  True
Exam1  Exam4    -6.2   -13.6652  1.2652 False
Exam2  Exam3   -6.06   -13.5252  1.4052 False
Exam2  Exam4    -2.9   -10.3652  4.5652 False
Exam3  Exam4    3.16   -4.3052  10.6252 False
--------------------------------------------
```

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Solution

The output above shows that the confidence interval for the comparison between the means of Exam1 and Exam3 does not contain 0. Thus, sufficient statistical evidence exist to support the claim that the means of Exam1 and Exam3 are different.

PARTICIPATION ACTIVITY 8.1.3: Post-hoc analysis on the mean sepal length of iris species.

The results of the ANOVA test on the mean sepal length of three iris species suggests that statistically significant evidence exists that the mean sepal lengths of three iris species are not the same.

Use the output below for the Tukey's HSD procedure to determine which pair of iris species have statistically different means at the $\alpha = 0.05$ significance level.

```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
==================================================
 group1     group2   meandiff lower  upper  reject
--------------------------------------------------
 setosa   versicolor   0.93   0.6862 1.1738  True
 setosa   virginica   1.582   1.3382 1.8258  True
versicolor virginica   0.652   0.4082 0.8958  True
--------------------------------------------------
```

Run example

1) The means of *setosa* and *versicolor*

- ○ Statistically the same
- ○ Statistically different

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

2) The means of *setosa* and *virginica*

- ○ Statistically the same
- ○ Statistically different

3) The means of *versicolor* and *virginica*

- Statistically the same
- Statistically different

8.2 Chi-square tests for categorical variables

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Chi-square goodness-of-fit test

The chi-square distribution can be used to test how close the distribution of a population is to a theoretical distribution. The chi-squared test statistic measures how different the observed counts are compared to the expected counts, assuming the null hypothesis is true. Specifically,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}.$$

The difference in observed and expected counts is squared to account for the fact that some of the absolute differences will be positive and some will be negative. Squaring the differences ensures that all terms will be positive. Dividing by the expected count is necessary to scale the differences as a percentage. Ex: If the expected count is 5, a squared difference of 10 is 200% of the expected count ($10/5 = 2$). A squared difference of 10 is much less relative to an expected count of 500 ($10/500 = 0.02$).

The alternate hypothesis for a chi-squared test is always two-sided, meaning that the proportions are not equal. Thus, if the null hypothesis is rejected, no conclusion can be made on which proportion is larger.

Chi-square test of goodness-of-fit

- Set the null and alternative hypothesis

  H_0: The random variable follows the expected distribution.

  H_a: The random variable does not follow the expected distribution.

  Equivalently, the null hypothesis is that the random variable does not follow the expected distribution, in which case, the alternative hypothesis is that the random variable does follow the expected distribution.
- Use to find the test statistic,

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}.$$

- Use to find the p-value. If the null hypothesis is true, the χ^2 test statistic has a chi-squared distribution with degrees of freedom equal to $k - 1$ where k is the number of categories
- Make a decision given a previously selected significance level α, typically 0.05
  - If the p-value is less than the significance level, sufficient evidence exists to reject the null hypothesis, H_0, in favor of the alternative hypothesis, H_a. Sufficient evidence exists to conclude that the distribution does not follow the expected distribution.
  - If the p-value is greater than or equal to the significance level, insufficient evidence exists to reject the null hypothesis, H_0, in favor of the alternative hypothesis, H_a. Insufficient evidence exists to conclude that the distribution does not follow the expected distribution.

©zyBooks 01/31/23 18:00 1267703 Traver Yates MAT-243-J3996-OL-TRAD-UG.23EW3

Example 8.2.1: Counting water birds.

A biologist keeps track of the composition of a population of birds on a lake. Last week, 50% of the birds were ducks, 23% were geese, 12% were cranes, 10% were swans, and 5% were coots. This week, the biologist counted 61 ducks, 17 geese, 11 cranes, 15 swans, and 6 coots. Is the composition of the population the same as last week at the $\alpha = 0.05$ significance level?

©zyBooks 01/31/23 18:00 1267703 Traver Yates MAT-243-J3996-OL-TRAD-UG.23EW3

American Coot. Source: Wikipedia. User: Baird, Mike [1]

Solution

The null hypothesis is that no difference exists between populations from last week and this week. Thus, the expected number of each type of bird is the percent from last week times the total number of birds this week, which is 110.

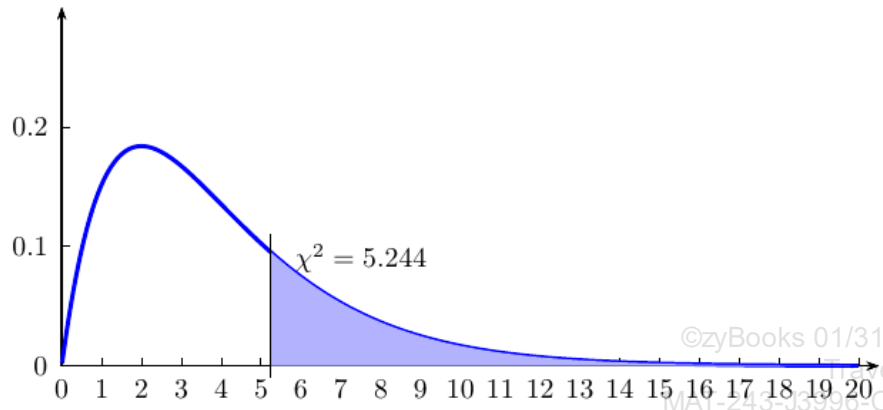
Type	Expected	Observed
Ducks	55	61
Geese	25.3	17
Cranes	13.2	11
Swans	11	15
Coots	5.5	6

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

The test statistic is

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(61-55)^2}{55} + \frac{(17-25.3)^2}{25.3} + \frac{(11-13.2)^2}{13.2} \\ &\quad + \frac{(15-11)^2}{11} + \frac{(6-5.5)^2}{5.5} \\ &= 5.244\end{aligned}$$

The distribution has $k - 1 = 4$ degrees of freedom. Using technology to calculate the p-value for $\chi^2 = 5.244$ gives $P(\chi^2 > 5.244) = 0.263$.



©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Since $0.263 > 0.05$, insufficient evidence exists to reject the null hypothesis, and the counts of birds this week are consistent with the expected distribution.

PARTICIPATION ACTIVITY 8.2.1: Daily distribution of accidents.

A police chief wants to know if traffic accidents are equally distributed across the days of the week. The chief looks at 150 randomly selected accident reports and records the data in the following table.

Day	Number of accidents
Sunday	14
Monday	32
Tuesday	20
Wednesday	22
Thursday	20
Friday	24
Saturday	18

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

1) What is the null hypothesis?

- The weekend days have a lower frequency of accidents than the weekdays.
- The counts do not differ from an equal frequency for each day.
- Mondays have a higher frequency of accidents than the other days.

2) Which table shows the correct expected counts for an even distribution across the days of the week?

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

	Expected		Observed
Day	Number of accidents		Number of accidents
Sunday	21.429		14
Monday	21.429		32
Tuesday	21.429		20
Wednesday	21.429		22
Thursday	21.429		20
Friday	21.429		24
Saturday	21.429		18

	Expected		Observed
Day	Number of accidents		Number of accidents
Sunday	150		14
Monday	150		32
Tuesday	150		20
Wednesday	150		22
Thursday	150		20
Friday	150		24
Saturday	150		18

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

	Expected		Observed
Day	Number of accidents		Number of accidents
Sunday	14.286		14
Monday	14.286		32
Tuesday	14.286		20
Wednesday	14.286		22
Thursday	14.286		20
Friday	14.286		24
Saturday	14.286		18

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

3) What is the test statistic χ^2?

- $\chi^2 = 8.582$
- $\chi^2 = -0.00014$
- $\chi^2 = 8.853$

4) How many degrees of freedom does the appropriate chi-square distribution have?

- 149
- 150
- 6

5) What is the p-value?

- 0.818
- 0.182
- 0.263

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

6) Is the distribution of accidents consistent with being evenly distributed across the days of the week at the $\alpha = 0.05$ significance level?

- Yes.

○ No.

Python-Practice 8.2.1: Using a chi-square goodness-of-fit test.

The `chisquare` object from the `scipy.stats` library can be used to perform a chi-square goodness of fit test. Ex: a biologist is counting birds on the lake and comparing the distribution to last week's counts. The observed and expected counts are given in the table below.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Type	Expected	Observed
Ducks	55	61
Geese	25.3	17
Cranes	13.2	11
Swans	11	15
Coots	5.5	6

```
from scipy.stats import chisquare

# Calculates the chi-square statistic and $p$-value for alpha =
0.05 for the given observed and expected counts
statistic, pvalue = chisquare([61, 17, 11, 15, 6], f_exp=[55,
25.3, 13.2, 11, 5.5])

print(statistic)
print(pvalue)
```

```
5.244137022397893
0.26315206062015767
```

`chisquare` thus returns the same results as calculated in the example above.

Run example

CHALLENGE ACTIVITY 8.2.1: Chi-square goodness of fit test.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

456500.2535406.qx3zqy7

.

Contingency tables

A chi-square test can also be used to determine whether two or more variables are independent by comparing the distributions of the variables over two or more categories. This test is the chi-square test for independence. The expected counts used to calculate the test statistic in this case come from a contingency table. A contingency table is constructed from the values of the variables and categories along the rows and columns. An expected cell count is calculated by multiplying the row total by the column total and dividing by the overall total.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Example 8.2.2: Calculating the expected counts for a 3x2 contingency table.

A journal article by Stouffer and Tibbits (1933) (Source: Journal of the American Statistical Association)[2] examines data on parole violations and punishment records involving 2,963 inmates at the Illinois State Reformatory. Punishment records include no prison time, 1-2 years in prison, or at least 3 years in prison. Violations are annotated as parole violation and no violation. The researchers examine the association between a punishment record while in prison with success or failure while on parole. Use the data in the table below to set up the contingency table necessary to determine if such an association exists.

	Violation		
Punishment	Parole violation	No parole violation	Total by punishment
None	405	1422	1827
1-2	240	470	710
3 or greater	151	275	426
Total by violation	796	2167	2963

Solution

To build the contingency table, each expected count is calculated by multiplying the corresponding row and column totals and then dividing by the overall total.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

$$\frac{(1827)(796)}{2963} = 491$$

$$\frac{(1827)(2167)}{2963} = 1336$$

$$\frac{(710)(796)}{2963} = 191$$

$$\frac{(710)(2167)}{2963} = 519$$

$$\frac{(426)(796)}{2963} = 114$$

$$\frac{(426)(2167)}{2963} = 312$$

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

When the process is done correctly, the totals in the Observed and Expected tables will be the same.

Observed			
	Violation		
Punishment	Parole violation	No parole violation	Total by punishment
None	405	1422	1827
1-2	240	470	710
3 or greater	151	275	426
Total by violation	796	2167	2963

Expected			
	Violation		
Punishment	Parole violation	No parole violation	Total by punishment
None	491	1336	1827
1-2	191	519	710
3 or greater	114	312	426
Total by violation	796	2167	2963

PARTICIPATION ACTIVITY 8.2.2: Calculating expected cell counts.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Observed			
	Died	Survived	Total by specialty
Infantry	374	2629	3003
Non-infantry	73	1219	1292
Total by outcome	447	3848	4295

Expected			
	Died	Survived	Total by specialty
Infantry	a	b	3003
Non-infantry	c	d	1292
Total by outcome	447	3848	4295

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Fill in the expected cell counts for the contingency table for deaths for infantry and non-infantry soldiers during a battle.

Select the definition that matches each term

1) 312.536
- d
- a
- c
- b

2) 2690.464
- d
- a
- c
- b

3) 134.464
- d
- a
- c
- b

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

4) 1157.536
- d
- a
- c

○ b

Reset

Chi-square test for independence

For a table with r rows and c columns, the test statistic for a chi-squared test follows a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom. For larger tables, the chi-square approximation requires large sample sizes. A rule of thumb is that all expected cell counts should be at least 5.

©zyBooks 01/31/23 18:00 1267703 Travel Yates MAT-243-J3996-OL-TRAD-UG.23EW3

Chi-square test of independence

- Set the null and alternative hypothesis

  H_0: The two variables are independent.

  H_a: The two variables are not independent.

  Equivalently, the null hypothesis is that no association exists between the two variables, in which case, the alternative hypothesis is that association exists between the variables.
- Use statistical software to find the test statistic,

$$\chi^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}}.$$

- Use statistical software to find the p-value. If the null hypothesis is true, the χ^2 test statistic has a chi-squared distribution with degrees of freedom equal to $(r-1)(c-1)$ where r is the number of categories of the first variable (rows in the table) and c is the number of categories of the second variable (columns in the table).
- Make a decision given a previously selected significance level α, typically 0.05
  - If the p-value is less than the significance level, sufficient evidence exists to reject the null hypothesis, H_0, in favor of the alternative hypothesis, H_a. Sufficient evidence exists to conclude that the two variables are not independent.
  - If the p-value is greater than or equal to the significance level, insufficient evidence exists to reject the null hypothesis, H_0, in favor of the alternative hypothesis, H_a. Insufficient evidence exists to conclude that the two variables are not independent.

©zyBooks 01/31/23 18:00 1267703 Travel Yates MAT-243-J3996-OL-TRAD-UG.23EW3

Example 8.2.3: Cats as part of the family.

The table below gives the ages of a sample of cat owners, and whether the cat owner thinks of the cat as part of the family. (Source: Pew Research Center)[3]

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

	Part of the family	Not part of the family
18 – 29	82	18
30 – 49	214	68
50 – 64	149	40
65 +	82	26

Are the variables age and considering a cat part of the family independent at the $\alpha = 0.05$ significance level?

Solution

The null hypothesis is that the variables are independent. The contingency tables for the observed and expected values should be constructed.

Observed		
	Part of the family	Not part of the family
18 – 29	82	18
30 – 49	214	68
50 – 64	149	40
65 +	82	26

Expected		
	Part of the family	Not part of the family
18 – 29	77.614	22.386
30 – 49	218.872	63.128
50 – 64	146.691	42.309
65 +	83.823	24.177

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

The test statistic is

$$\chi^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}}$$
$$= \frac{(82 - 77.614)^2}{77.614} + \frac{(214 - 218.872)^2}{218.872} + \frac{(149 - 146.691)^2}{146.691}$$
$$+ \frac{(82 - 83.823)^2}{83.823} + \frac{(18 - 22.386)^2}{22.386} + \frac{(68 - 63.128)^2}{63.128}$$
$$+ \frac{(40 - 42.309)^2}{42.309} + \frac{(26 - 24.177)^2}{24.177}$$
$$= 1.931$$

zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

The chi-square distribution has $(r-1)(c-1) = 3$ degrees of freedom. The p-value is 0.587. Since the p-value is greater than $\alpha = 0.05$, the null hypothesis cannot be rejected. Insufficient evidence exists to conclude that the variables are not independent. Typically, statistical software is used to calculated the test statistic and the p-value.

Python-Practice 8.2.2: Chi-squared test of independence.

The `chi2_contingency` object from the `scipy.stats` library can be used to perform a chi-square goodness of fit test. Conducting the chi-square test for independence requires first constructing a contingency table. An example using the parole data is shown below. Is there an association between a punishment record while in prison and success or failure while in prison?

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

# Construct a contingency table
parole = np.array([[405,1422], [240,470], [151,275]])
```

The command for the chi-square test and corresponding output are displayed below.

```
# Calculate the test statistic, $p$-value, degrees of
freedom, and expected counts
chi2, p, df, ex = chi2_contingency(parole)
print(chi2)
print(p)
print(df)
print(ex)
```

```
53.87860692066112
1.9971429926442894e-12
2
[[ 490.81741478
1336.18258522]
 [ 190.73911576
519.26088424]
 [ 114.44346946
311.55653054]]
```

zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Run example

PARTICIPATION ACTIVITY 8.2.3: Chi-square test of independence.

Refer to the Python practice above.

1) What are the null and alternative hypotheses?

- H_0: Punishments and parole violations are not associated.
  H_a: Punishments and parole violations are associated.
- H_0: Punishments and parole violations are associated.
  H_a: Punishments and parole violations are not associated.
- H_0: The proportion of parole violators is the same as the proportion of parole non-violators.
  H_a: The proportion of parole violators is less than the proportion of parole non-violators.
- H_0: The proportion of parole violators is the same among prisoners with records of infractions as among prisoners with no infractions.
  H_a: The proportion of parole violators is greater among prisoners with records of infractions than among prisoners with no infractions.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

2) What is the chi-square test statistic, χ^2, for the parole contingency table?

- 1.997×10^{-12}
- 53.879
- Not displayed

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

3) What conclusion can be drawn from the chi-square hypothesis test?

- Reject the null hypothesis.
- Fail to reject the null hypothesis.
- Prisoners with punishment records should not be paroled.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

4) Consider the following partial tables of observed and expected cell counts corresponding to the parole data.

Observed		
	Parole Violation	No Violation
None	405	1422
1-2	--	--
3 or greater	--	--

Expected		
	Parole Violation	No Violation
None	491	1336
1-2	--	--
3 or greater	--	--

Which cell contributes the most to the chi-square test statistic?

- All cells contribute the same to the chi-square test statistic.
- The cell corresponding to Parole Violation and None
- The cell corresponding to No Violation and None

Chi-square assumptions

For the chi-square approximation to be valid, every expected cell count must be sufficiently large. At least 5 is the rule of thumb. For tables with cell counts less than 5, one alternative is to combine categories of one or both variables to increase the cell sizes. Often, a logical way to combine categories exists.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Example 8.2.4: Combining categories to satisfy expected cell count assumption.

Suppose researchers wish to study the relationship between education level and voter preference by polling likely voters in a local election. The following results regarding which candidate voters are most likely to vote for are obtained in the table below.

	High school or less	Some college	Bachelor's degree	Master's degree	Doctoral degree
Candidate 1	75	50	3	2	1
Candidate 2	60	45	2	3	1
Candidate 3	80	60	2	2	2

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

The expected cell counts are shown below.

	High school or less	Some college	Bachelor's degree	Master's degree	Doctoral degree
Candidate 1	72.590	52.332	2.363	2.363	1.351
Candidate 2	61.508	44.343	2.003	2.003	1.144
Candidate 3	80.902	58.325	2.634	2.634	1.505

Not all expected cell counts are at least 5. Combining the last three columns of the observed table into a single column "Bachelor's degree or higher" satisfies the minimum cell count assumption. Ex: For the first candidate, combining the "Bachelor's degree," "Master's degree," and "Doctoral degree" columns of the observed table results in 3+2+1 = 6 observations. Since the row and column sums are fixed, the expected counts for the "High school" and "Some college" columns remain unchanged. All expected cell counts in the "Bachelor's degree or greater" column changed to satisfy the assumption.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Observed			
	High school	Some college	Bachelor's degree or higher
Candidate 1	75	50	6
Candidate 2	60	45	6
Candidate 3	80	60	6

Expected			
	High school	Some college	Bachelor's degree or higher
Candidate 1	72.590	52.332	6.077
Candidate 2	61.508	44.343	5.149
Candidate 3	80.902	58.325	6.773

©zyBooks 01/31/23 18:00 1267703
Jennifer Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

If a logical way to combine categories exists, categories should be combined. Ex: A survey response has five possible answers: strongly agree, agree, neutral, disagree and strongly disagree. Combining strongly agree and agree or strongly disagree and disagree makes sense. However, combining agree and disagree does not. A logical way to combine ordinal variables usually exists whereas sometimes no logical way exists for nominal variables.

Example 8.2.5: Combining categories to satisfy expected cell count assumption.

Sometimes, categories cannot be combined in a natural manner. Ex: Suppose data was collected fr breed of dog and whether they lived in an urban, suburban or rural area.

	Toy Poodle	Golden Retriever	Akita	Doberman Pinscher	German Shepherd	Great
Urban	10	30	2	2	10	1
Suburban	20	60	3	4	20	2
Rural	5	30	6	8	20	6

The expected cell counts are shown below.

	Toy Poodle	Golden Retriever	Akita	Doberman Pinscher	German Shepherd	Great
Urban	8.635	29.605	2.714	3.454	12.336	2.220
Suburban	17.155	58.816	5.391	6.862	24.507	4.411
Rural	9.211	31.579	2.895	3.684	13.158	2.368

©zyBooks 01/31/23 18:00 1267703
Travel Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Since some expected cell counts are less than 5, certain categories of the observed table should be no obvious way to combine categories exist. In this situation, the observed columns for less popula Doberman Pinscher and Great Dane) can be combined into a single column labeled as "Other" as sh below.

Observed						Expected			
	Toy Poodle	Golden Retriever	German Shepherd	Chihuahua	Other		Toy Poodle	Golden Retriever	Ger Shep
Urban	10	30	10	20	5	Urban	8.634	29.605	12.3
Suburban	20	60	20	40	9	Suburban	17.155	58.816	24.5
Rural	5	30	20	5	20	Rural	9.211	31.579	13.1

©zyBooks 01/31/23 18:00 1267703 Traver Yates MAT-243-J3996-OL-TRAD-UG.23EW3

The results from the original data and the data with combined categories should always be compar should be considered in context. A minimum expected cell count of 5 is just a rule of thumb.

PARTICIPATION ACTIVITY 8.2.4: Combining categories.

1) The table below lists the expected cell counts from a survey of visitors leaving a local aquarium. Visitors were asked to rate their experience as excellent, outstanding, good, fair, or poor. Surveyors noted whether the patrons visited with children or without. Which of the following is true about combining columns for a chi-square test?

	Excellent	Outstanding	Good	Fair	Poor
With children	2.301	7.422	12.100	2.842	7.415
No children	2.730	8.606	13.919	3.221	8.634

©zyBooks 01/31/23 18:00 1267703 Traver Yates MAT-243-J3996-OL-TRAD-UG.23EW3

- The "excellent" and "fair" columns should be combined, so that all expected counts are greater than 5.
- The "excellent" and "outstanding" columns should be combined, as well as the "fair" and "poor" columns, so that all expected counts are greater than 5.

- No appropriate way to combine categories exists in this situation.

2) People's preference for beer versus wine are recorded in three cities. Based on the expected values and chi-square test results given below, does drink preference vary among cities?

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

	City A	City B	City C
Beer	2039.122	509.7804	5.097804
Wine	1920.878	450.2196	4.902196

```
z =
np.array([[2039.122,
509.7804, 5.097804],
[1920.878, 450.2196,
4.902196]])

chi2, p, df, ex =
chi2_contingency(z)
print(chi2)
print(p)
print(df)
print(ex)
```

```
0.80409
0.66895
2
[[2051.48900768
497.33066853
5.1805278 ]
 [1908.51099232
462.66933147
4.8194722 ]]
```

- No conclusion can be made because one cell count is less than 5.
- Sufficient evidence exists to support the claim that drink preference varies between cities.
- Insufficient evidence exists to support the claim that drink preferences vary between cities.

Chi-square test of independence and test of homogeneity

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

The null hypothesis for a chi-square test of independence is that two variables are independent. A related test is the chi-square test of homogeneity. Mathematically, the two tests are equivalent and will always result in the same test statistic and p-value. However, the null hypothesis for the chi-square test of homogeneity is that the distribution of one variable is the same across all categories of the other variable.

Chi-square test of homogeneity

- Set the null hypothesis

$$\begin{aligned} H_0: p_{1,1} = p_{1,2} &= \ldots = p_{1,J} \\ p_{2,1} = p_{2,2} &= \ldots = p_{2,J} \\ \vdots &\quad \vdots \\ p_{I,1} = p_{I,2} &= \ldots = p_{I,J} \\ p_{1,1} = p_{2,1} &= \ldots = p_{I,1} \\ p_{1,2} = p_{2,2} &= \ldots = p_{I,2} \\ \vdots &\quad \vdots \\ p_{1,J} = p_{2,J} &= \ldots = p_{I,J} \end{aligned}$$

©zyBooks 01/31/23 18:00 1267703 Traver Yates MAT-243-J3996-OL-TRAD-UG.23EW3

  where $i = 1, 2, \ldots, I$ represent the categories of the first variable and $j = 1, 2, \ldots, J$ represent the categories of the second variable. Fundamentally, the null hypothesis says that the distribution of one variable is the same across all sampled populations.
- Set the alternative hypothesis

  H_a: At least one of the probability statements is false.
- Use statistical software to find the test statistic,

$$\chi^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}}.$$

  The test statistic for the chi-square test of independence and the chi-square test of homogeneity are identical.
- Use to find the p-value. If the null hypothesis is true, the χ^2 test-statistic has a chi-squared distribution with degrees of freedom equal to $(r-1)(c-1)$ where r is the number of categories of the first variable (rows in the table) and c is the number of categories of the second variable (columns in the table).
- Make a decision given a previously selected significance level α, typically 0.05
  - If the p-value is less than the significance level, sufficient evidence exists to reject the null hypothesis, H_0, in favor of the alternative hypothesis, H_a. Conclude that the distribution of one variable is not the same across categories of the other variable.
  - If the p-value is greater than or equal to the significance level, insufficient evidence exists to reject the null hypothesis, H_0, in favor of the alternative

©zyBooks 01/31/23 18:00 1267703 Traver Yates MAT-243-J3996-OL-TRAD-UG.23EW3

hypothesis, H_a. Conclude that the distribution of one variable is the same across categories of the other variable.

The chi-square test of homogeneity is appropriate when a sample is taken from two or more populations. In that situation, the null hypothesis says that the distribution of one variable is the same across all sampled populations. Ex: 5-year mortality rates of patients from 4 hospitals are recorded. The mortality rates from each hospital represent a population, so 5 populations exist in total. The null hypothesis should state that the probability of 5-year mortality is the same across all 4 hospitals. In contrast, the test of independence is used when two or more variables are compared within a single sample. The choice of which chi-square test is more appropriate in a given situation is not always clear. In practice, the test for independence and test for homogeneity are sometimes used interchangeably.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Example 8.2.6: Frequency of tea drinking.

The table below gives the frequency with which a sample of men and women drink tea (Source: Journal of Alzheimer's Disease)[4] Do men and women drink tea with the same frequency at the $\alpha = 0.05$ significance level?

	Men	Women
$<$ 5/year	551	580
5 - 10 /year	244	289
1-3/month	387	503
1-4/week	452	618
$\geq$ 5/week	443	742

Solution

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

The test for homogeneity is performed in the same way as the test for independence. The null hypothesis is that the distribution of the frequency of tea drinking is the same for men and women. The alternative hypothesis is that the distributions differ in at least one category. The contingency tables are

Observed				Expected		
	Men	Women			Men	Women
< 5/year	551	580		< 5/year	488.477	642.523
5 - 10 /year	244	289		5 - 10 /year	230.202	302.798
1-3/month	387	503		1-3/month	384.390	505.610
1-4/week	452	618		1-4/week	462.131	607.899
≥ 5/week	443	742		≥ 5/week	511.800	673.200

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

The chi-square test statistic is the same as for the test for independence.

$$\chi^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}}$$

Using Python to calculate the test statistic and p-value for a chi-square distribution with $(r-1)(c-1) = 4$ degrees of freedom, the result is

```
z = np.array([[551, 580], [244, 289], [387, 503], [452, 618], [443,742]])

chi2, p = chi2_contingency(z)
print(chi2)
print(p)
```

```
32.2443
1.70526e-6
```

Since the p-value is much smaller than $\alpha = 0.05$, the null hypothesis is rejected, and men and women do not drink tea with the same frequency distribution.

PARTICIPATION ACTIVITY 8.2.5: Test of homogeneity versus independence.

Select the appropriate test for each situation.

1) The relationship between marijuana use and party habits among college students is investigated.
   - ○ Independence
   - ○ Homogeneity

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

2) Data on deaths among infantry and non-infantry Union soldiers in the Civil War are collected. Researchers

concluded that the probability of death for infantry soldiers is different from the probability of death for non-infantry soldiers.

- Independence
- Homogeneity

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

3) Data on women's hair length and age bracket (20-29, 30-39, etc.) are collected. Researchers conclude that hair length and age are associated.

- Independence
- Homogeneity

References

(*1) Baird, Mike "File:Fulica americana3.jpg." *Wikimedia Commons, the free media repository*. 17 Nov 2018, 1, accessed 20 Jun 2019, commons.wikimedia.org/w/index.php?title=File:Fulica_americana3.jpg&oldid=328080576

(*2) Stouffer, Samuel A. "Tests of Significance in Applying Westergaard's Method of Expected Cases to Sociological Data." *Journal of the American Statistical Association*, vol. 28, no. 183, 1933, pp. 293-302. *JSTOR*, www.jstor.org/stable/2278425.

(*3) Taylor, Paul, et al. "Gauging Family Intimacy: Dogs Edge Cats (Dads Trail Both)." *Pew Research Center*, 7 March 2006, https://www.pewresearch.org/wp-content/uploads/sites/3/2010/10/Pets.pdf.

(*4) Arab, Lenore et al. "Gender differences in tea, coffee, and cognitive decline in the elderly: the Cardiovascular Health Study." *Journal of Alzheimer's disease : JAD* vol. 27,3 (2011): 553-66. doi:10.3233/JAD-2011-110431.

©zyBooks 01/31/23 18:00 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3