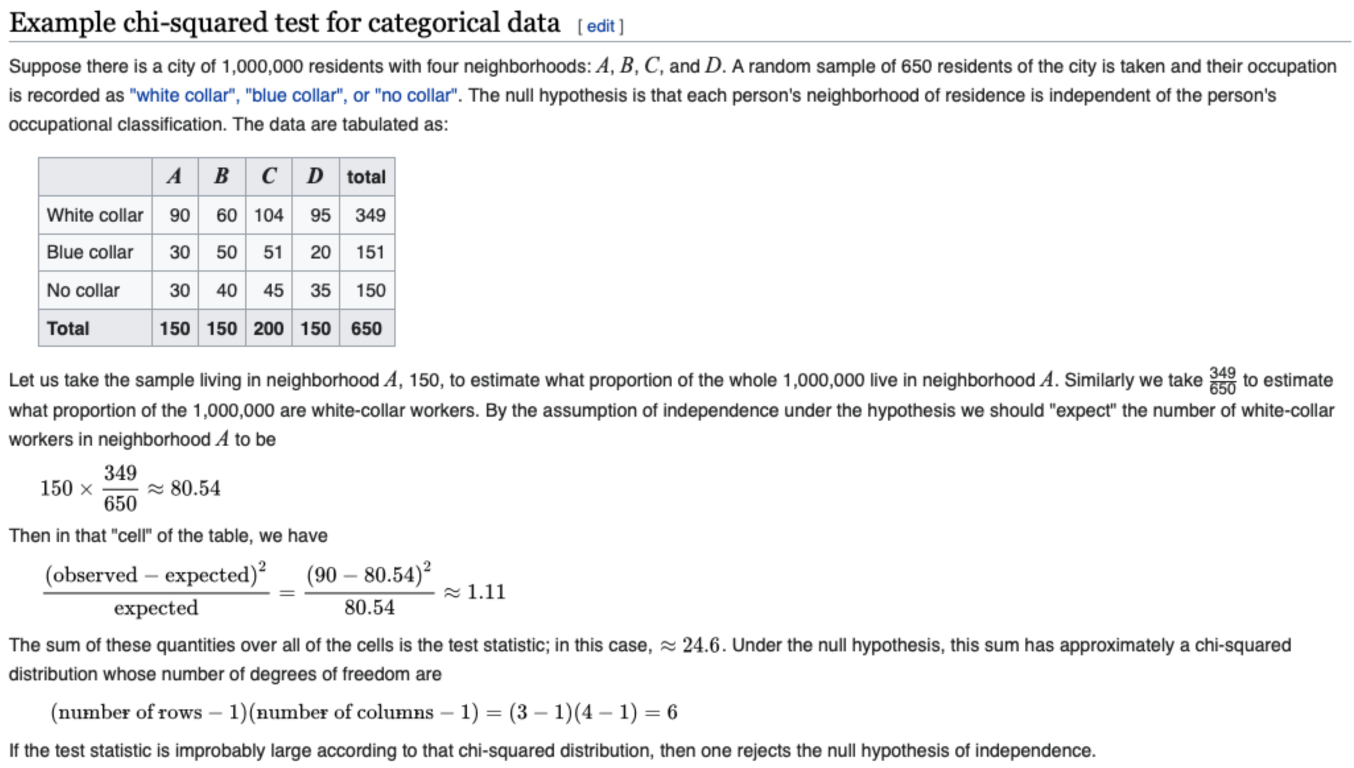
Images



Example chi-squared test for categorical data [edit]

Suppose there is a city of 1,000,000 residents with four neighborhoods: A, B, C, and D. A random sample of 650 residents of the city is taken and their occupation is recorded as "white collar", "blue collar", or "no collar". The null hypothesis is that each person's neighborhood of residence is independent of the person's occupational classification. The data are tabulated as:

	A	B	C	D	total
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	**150**	**150**	**200**	**150**	**650**

Let us take the sample living in neighborhood A, 150, to estimate what proportion of the whole 1,000,000 live in neighborhood A. Similarly we take $\frac{349}{650}$ to estimate what proportion of the 1,000,000 are white-collar workers. By the assumption of independence under the hypothesis we should "expect" the number of white-collar workers in neighborhood A to be

$$150 \times \frac{349}{650} \approx 80.54$$

Then in that "cell" of the table, we have

$$\frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{(90 - 80.54)^2}{80.54} \approx 1.11$$

The sum of these quantities over all of the cells is the test statistic; in this case, ≈ 24.6. Under the null hypothesis, this sum has approximately a chi-squared distribution whose number of degrees of freedom are

$$(\text{number of rows} - 1)(\text{number of columns} - 1) = (3 - 1)(4 - 1) = 6$$

If the test statistic is improbably large according to that chi-squared distribution, then one rejects the null hypothesis of independence.

Chi-squared Test

```
import numpy as np
import pandas as pd
import seaborn as sns
import os,sys,time
import scipy
import statsmodels

from scipy import stats
from scipy.stats import ttest_1samp
from scipy.stats import ttest_ind # independent means two samples.
from statsmodels.stats import weightstats as stests # stests.ztest

# contingency table
table = [ [10, 20, 30],
          [6,  9,  17]]

stat, p, dof, expected = stats.chi2_contingency(table)
print('dof=%d' % dof) # dof=2
print('expected values=\n',expected)
expected values=
 [[10.43478261 18.91304348 30.65217391]
 [ 5.56521739 10.08695652 16.34782609]]
print()
```

```
# interpret test-statistic
alpha = 0.05
prob = 1 - alpha # 0.95
critical = stats.chi2.ppf(1-alpha, dof)
critical = stats.chi2.ppf(prob, dof)
print('probability=%.3f, critical=%.3f, stat=%.3f' % (prob, critical, stat))
# probability=0.950, critical=5.991, stat=0.272
print()

if abs(stat) >= critical:
    print('Dependent (reject H0)')
else:
    print('Independent (fail to reject H0)')

# Independent (fail to reject H0)
print()

# interpret p-value
alpha = 1.0 - prob
print('significance=%.3f, p=%.3f' % (alpha, p)) # significance=0.050, p=0.873
if p <= alpha:
    print('Dependent (reject H0)')
else:
    print('Independent (fail to reject H0)')

# Independent (fail to reject H0)
```