Chi-Squared Tests: Takeaways

Syntax

- Calculate the chi-squared test statistic and create a histogram of all the chi-squared values:

```
chi_squared_values = []
from numpy.random import random
import matplotlib.pyplot as plt
for i in range(1000):
    sequence = random((32561,))
    sequence[sequence < .5] = 0
    sequence[sequence >= .5] = 1
    male_count = len(sequence[sequence == 0])
    female_count = len(sequence[sequence == 1])
    male_diff = (male_count - 16280.5) ** 2 / 16280.5
    female_diff = (female_count - 16280.5) ** 2 / 16280.5
    chi_squared = male_diff + female_diff
    chi_squared_values.append(chi_squared)
plt.hist(chi_squared_values)
```

- Calculate a chi-squared sampling distribution with two degrees of freedom:

```
import numpy as np
from scipy.stats import chisquare
observed = np.array([5, 10, 15])
expected = np.array([7, 11, 12])
chisquare_value, pvalue = chisquare(observed, expected) # returns a list
```

Concepts

- The chi-squared test enables us to quantify the difference between sets of observed and expected categorical values to determine statistical significance.
- To calculate the chi-squared test statistic, we use the following formula: $\frac{\text{observed} - \text{expected}^{\,2}}{\text{expected}}$.
- A p-value allows us to determine whether the difference between 2 values is due to chance, or due to an underlying difference.
- Chi-squared values increase as sample size increases, but the chance of getting a high chi-squared value decreases as the sample gets larger.
- A degree of freedom is the number of values that can vary without the other values being "locked in."

Resources

- [Chi-Square Test]
- [Degrees of Freedom]
- [Scipy Chi-Square documentation]