STA13: Elementary Statistics

Lecture 21
Book sections 8.3-8.4

Dmitriy Izyumin

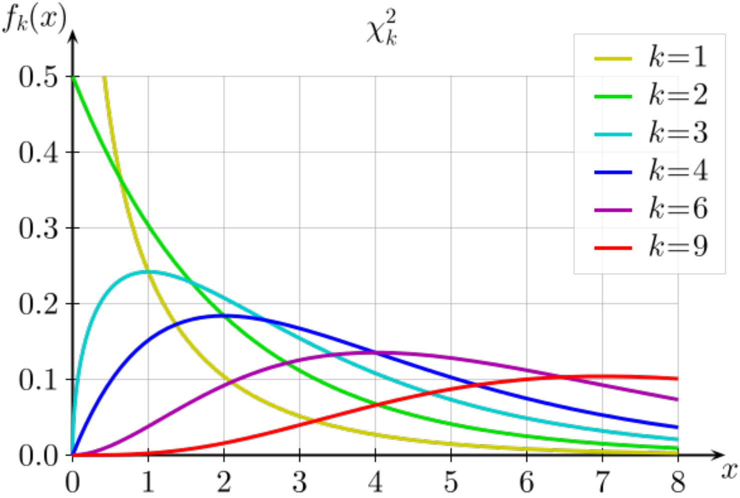
March 09 2018

A New Distribution

Suppose $Z_1, Z_2, \cdots Z_k$ are independent random variables, each distributed $N(0,1)$.

- Z_1^2 has the Chi squared distribution with 1 degree of freedom.
- $Z_1^2 + Z_2^2$ has the Chi squared distribution with 2 degrees of freedom.
- $Z_1^2 + Z_2^2 + \cdots + Z_k^2$ has the Chi squared distribution with k degrees of freedom.

Chi Squared Pdf

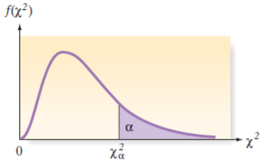


The Chi squared Distribution

- We write χ^2_k to denote a chi squared distribution with k degrees of freedom.
- Shape depends on the degrees of freedom.
- Not symmetric.
- Negative values are not possible.

Chi Squared Table

Table V Critical Values of χ^2



Degrees of Freedom	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$
1	.0000393	.0001571	.0009821	.0039321	.0157908
2	.0100251	.0201007	.0506356	.102587	.210720
3	.0717212	.114832	.215795	.351846	.584375
4	.206990	.297110	.484419	.710721	1.063623
5	.411740	.554300	.831211	1.145476	1.61031
6	.675727	.872085	1.237347	1.63539	2.20413

Table V (continued)

Degrees of Freedom	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381
4	7.77944	9.48773	11.1433	13.2767	14.8602
5	9.23635	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476

Chi Squared Table

- Find the row for the correct degrees of freedom.
- Setup is similar to the t table.
- Will need to bracket your test stastic between numbers in the table to approximate the p-value.

Multinomial Experiment

Properties of the Multinomial Experiment

1. The experiment consists of n identical trials.
2. There are k possible outcomes to each trial. These outcomes are sometimes called **classes, categories,** or **cells.**
3. The probabilities of the k outcomes, denoted by $p_1, p_2, \ldots, p_k$, where $p_1 + p_2 + \cdots + p_k = 1$, remain the same from trial to trial.
4. The trials are independent.
5. The random variables of interest are the **cell counts** $n_1, n_2, \ldots, n_k$ of the number of observations that fall into each of the k categories.

Test Scenario

- Suppose we have n observations of one qualitative variable with k categories.
  - Voters' preferred presidential candidate (with more than 2 options).
  - Fovorite color of Skittles.
  - Income / tax bracket.
- The goal is to test if our guess about the proportions of each category is correct.

Null Hypothesis

- H_0 needs to specify hypothesized values for each of k probabilities. (These have to add up to 1).
- Example:
  H_0: all probabilities are equal
  (implies that $p_{i,0} = 1/k$ for each category)
- Example:
  $H_0 : p_1 = 0.5, p_2 = 0.2, p_3 = 0.3.$

Alternative Hypothesis

- H_A states that at least one of the true proportions is not equal to the hypothesized value.
- Does not mean that ALL of the hypothesized values are wrong.
- Not ALL of the hypothesized values are right.

Test Statistic

The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{(n_i - E_i)^2}{E_i} \right)$$

1. For each category $(i = 1, \cdots, k)$
   - Find the observed count n_i
   - Find the expected count $E_i = np_{i,0}$
   - Compute $\frac{(n_i - E_i)^2}{E_i}$
2. Add up the results from each category.

Null Distribution

- If H_0 is true, then the test statistic is distributed Chi squared with $k-1$ degrees of freedom.
- Here k is the number of categories.
- Use this distribution to find p-values.

One Way Test

A Test of a Hypothesis about Multinomial Probabilities: One-Way Table

H_0: $p_1 = p_{1,0}$, $p_2 = p_{2,0}$, ..., $p_k = p_{k,0}$

where $p_{1,0}, p_{2,0}, \ldots, p_{k,0}$ represent the hypothesized values of the multinomial probabilities

H_a: At least one of the multinomial probabilities does not equal its hypothesized value

Test statistic: $\chi^2 = \sum \frac{[n_i - E_i]^2}{E_i}$

where $E_i = np_{i,0}$ is the **expected cell count**—that is, the expected number of outcomes of type i, assuming that H_0 is true. The total sample size is n.

Rejection region: $\chi^2 > \chi^2_\alpha$, where χ^2_α has $(k - 1)$ df

One Way Conditions

Conditions Required for a Valid χ^2 Test: One-Way Table

1. A multinomial experiment has been conducted. This is generally satisfied by taking a random sample from the population of interest.
2. The sample size n will be large enough so that, for every cell, the expected cell count $E(n_i)$ will be equal to 5 or more.*

Test Scenario

- Suppose we have n observations of two qualitative variables.
  - Age range and political party affiliation
  - Religious affiliation and marital status
- The goal is to see if the two variables are independent.

Test Scenario

- Suppose we have n observations of two qualitative variables.
  - Age range and political party affiliation
  - Religious affiliation and marital status
- The goal is to see if the two variables are independent.

Two-Way Table

Table 8.10 Survey Results (Observed Counts), Example 8.6

		Religious Affiliation					
		A	B	C	D	None	Totals
Marital Status	Divorced	39	19	12	28	18	116
	Married, never divorced	172	61	44	70	37	384
	Totals	211	80	56	98	55	500

Data Set: MARREL

- Also known as a contingency table.
- r rows and c columns.
- may or may not show row/column totals (this one does).

Null Hypothesis

- H_0 states that the two categorical variables are independent. (These have to add up to 1).
- Example:
  H_0: Age range and party affiliation are independent.
- Example:
  H_0: Religious affiliation and marital status are independent.

Alternative Hypothesis

- H_A states that the two categorical variables are not independent (are dependent).
- Example:
  H_0: Age range and party affiliation are dependent.
- Example:
  H_0: Religious affiliation and marital status are not independent.

Test Statistic

The test statistic is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \right)$$

1. For each cell (row i, column j)
   - R_i is the total for row i, and C_j is the total for column j
   - Find the observed count n_{ij}
   - Find the expected count $E_{ij} = R_i C_j / n$
   - Compute $\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$
2. Add up the results from each cell.

Null Distribution

- If H_0 is true, then the test statistic is distributed Chi squared with $(r-1)(c-1)$ degrees of freedom.
- Here r is the number of rows and c is the number of columns in the table.
- Use this distribution to find p-values.

Two Way Test

General Form of a Two-Way (Contingency) Table Analysis: A Test for Independence

H_0: The two classifications are independent

H_a: The two classifications are dependent

Test statistic: $\chi^2 = \sum \frac{[n_{ij} - \hat{E}_{ij}]^2}{\hat{E}_{ij}}$

where $\hat{E}_{ij} = \frac{R_i C_j}{n}$

Rejection region: $\chi^2 > \chi^2_\alpha$, where χ^2_α has $(r-1)(c-1)$ df

Two Way Conditions

Conditions Required for a Valid χ^2 Test: Contingency Tables

1. The n observed counts are a random sample from the population of interest. We may then consider this to be a multinomial experiment with $r \times c$ possible outcomes.
2. The sample size n will be large enough so that, for every cell, the expected count $\hat{E}(n_{ij})$ will be equal to 5 or more.