Chapter 9-10

Confidence Intervals and Hypothesis Testing

Goodness of Fit Tests

Statistics

Mehmet Güray Güler, PhD

Last updated 28.07.2020

Tests for Independence for Categorical Data

Chi-Squared Test for Independence

- The chi-squared test procedure could also be used to test the hypothesis of independence of two variables of classification.

- Example: Suppose we want to determine whether being a fan of a team is dependent on the gender or not.

- To do this we need data about the two variables for the residents in our sample.

Chi-Squared Test for Independence

- A random sample of residents from the region is drawn and each resident is asked two questions:

1. A question about the favorite team (Variable 1)
2. A question about their gender (Variable 2)

- For the chi-squared test of independence the data are grouped using the answers to these two questions (variables of classification).

Chi-Squared Test for Independence

- We count how many residents fall in each subgroup and place these frequencies in a contingency table as follows

	Team			
Gender	GS	FB	BJK	TOTAL
Male	182	213	203	598
Female	154	138	110	402
Total	336	351	313	1000

- The contingency table entries are the observed frequencies.

Contingency Tables

- In general a contingency table may have r rows and c columns.
- The row and column totals are called marginal frequencies.
- We form the hypotheses:
- $\boldsymbol{H_0}$: The two variables are independent (gender has no effect on the fav. team).
- $\boldsymbol{H_1}$: The two variables are NOT independent (the favorite team and the gender are somehow related).

Events and Probabilities in Contingency Tables

- To calculate expected frequencies under H_0 we define the events:
  - G: A person selected at random is a fan of GS.
  - F: A person selected at random is a fan of FB.
  - B: A person selected at random is a fan of BJK.
  - M: A person selected at random is male.
  - E: A person selected at random is female.

Chi-Squared Test for Independence

	Team			
Gender	GS	FB	BJK	TOTAL
Male	182	213	203	598
Female	154	138	110	402
Total	336	351	313	1000

- Using the marginal frequencies we can calculate the probabilities of these events as follows

- $P(G) = 336/1000, P(F) = 351/1000, P(B) = 313/1000$
- $P(M) = 598/1000, P(E) = 402/1000$

Chi-Squared Test for Independence

	Team			
Gender	GS (G)	FB (F)	BJK (B)	TOTAL
Male (M)	182	213	203	598
Female (F)	154	138	110	402
Total	336	351	313	1000

- $P(G) = 336/1000$
- $P(F) = 351/1000$
- $P(B) = 313/1000$
- $P(M) = 598/1000$
- $P(E) = 402/1000$

- Under H_0 the two variables gender and favorite team are independent, hence we calculate the joint probabilities from:

- $P(G \cap M) = (0.336)(0.598) = 0.2$
- $P(G \cap E) = (0.336)(0.402) = 0.135$

- $P(F \cap M) = (0.351)(0.598) = 0.21$
- $P(F \cap E) = (0.351)(0.402) = 0.14$
- …

Joint Probabilities in Contingency Tables

- Since expected frequency = probability x sample size,
- we can generalize the formula for calculating expected frequencies using the joint probabilities under H_0 (independence assumption) as follows:

- expected frequency = (column total) x (row total) / (grand total)

Chi-Squared Test for Independence

The expected frequencies are shown in the table within brackets.

Now, we use chi-squared test statistic X^2 as defined before and the same decision rule like the goodness-of-fit test:

Reject H_0 if $X^2 > \chi^2_\alpha$.

where χ^2_α has degrees of freedom $= v = (r - 1)\ (c - 1)$.

	Team			
Gender	GS (G)	FB (F)	BJK (B)	TOTAL
Male (M)	182	213	203	598
Female (F)	154	138	110	402
Total	336	351	313	1000

column total

row total

grand total = sample size

EXAMPLE 3. Test for Independence

- Applying this procedure, we calculate the test statistic:

$$\chi^2 = \frac{(182 - 200.9)^2}{200.9} + \frac{(213 - 209.9)^2}{209.9} + \frac{(203 - 187.2)^2}{187.2}$$

$$+ \frac{(154 - 135.1)^2}{135.1} + \frac{(138 - 141.1)^2}{141.1} + \frac{(110 - 125.8)^2}{125.8} = 7.85,$$

- DECISION RULE : Reject H_0 if $X^2 > \chi^2_{0.05} = 5.99$,
- where the chi-square has $v = (2\text{-}1)(3\text{-}1) = 2$ degrees of freedom.

EXAMPLE 3. Test for Independence

- The test statistic = $\chi^2 = 7.85$
- Critical value for $\alpha = 0.05 = \chi^2{}_{0.05} = 5.99$
- **Decision**: We reject the null hypothesis.
- ***P*-value** $= P(X^2 > 7.85) \approx 0.02$. (For 2 degrees of freedom)
- **Conclusion**: Being a fan of a team is not independent of the gender.
- Gender affects the favorite team.