MA 214: Applied Statistics

Instructor: Ashis Gangopadhyay

Categorical Data Analysis

Where We've Been

We were discussing….

- How to model a response variable that is continuous, while the covariates (factors) are continuous or categorical.

Where We're Going

- Our objective is to explore models that allow the response variable to be categorical.
- The factors (covariates) can be continuous or categorical. Initially, we will assume that all factors are categorical.

Categorical Data Analysis (Goodness of Fit Test)

- Suppose a population can be divided into *k*-categories. We have a random sample of size *n* from the population, such that each unit in the sample must fall in one and only one category.
- We then count the observed frequency of each category and denote these frequencies by $f_1, f_2, \ldots, f_k$. Note that $\sum_{i=1}^{k} f_i = n$.

Categorical Data Analysis (Goodness of Fit Test)

- **Example:** Suppose a study was conducted to understand the level of satisfaction among BU graduating seniors with their academic experience during the last four years at BU. 100 seniors were randomly selected, and each student was asked to rate their experience at BU.

Categorical Data Analysis (Goodness of Fit Test)

Question: My level of satisfaction with the education at BU is

- Excellent
- Very Good
- Good
- Fair
- I'd rather be in hell!

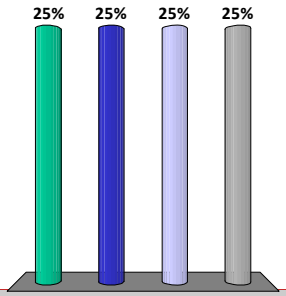
Categorical Data Analysis (Goodness of Fit Test)

- **Example (con't):** The data are used to generate the following frequency table.

Categories	*Frequency*
Excellent	60
Very good	25
Good	10
Fair	4
Hell	1
Total	100

Categorical Data Analysis is appropriate when…

1. Response variable and covariates are in general continuous.
2. Response variable is continuous, but all covariates are categorical.
3. ✓ Response variable and all covariates are categorical.
4. Response variable is categorical, but the covariates are in general continuous.



Categorical Data Analysis (Goodness of Fit Test)

- We are interested in a situation where we have a "theoretical" model in mind, and we wish to see if that model fits the observed data. Suppose we can use the theoretical model to obtain some theoretical (expected) frequencies $F_1, F_2, \ldots, F_k$, where $\sum_{i=1}^{k} F_i = n$.
- This leads to the following table:

Category	***Observed Frequency***	***Expected Frequency***
1	f_1	F_1
2	f_2	F_2
…	…	…
K	f_k	F_k
Total	n	n

Categorical Data Analysis (Goodness of Fit Test)

- Our objective is to verify that the theoretical model fits the observed data. That is, we wish to test the following hypothesis:

H_0 : The theoretical model fits the data.

H_a : The theoretical model doesn't fit the data.

Categorical Data Analysis (Goodness of Fit Test)

- If the observed frequency "matches" the expected frequency, then we can say with a reasonable degree of confidence that the null hypothesis is true.
- On the other hand, if the observed frequency is significantly different from the expected frequency, then we should conclude that the model is false, and the alternative hypothesis is true. You can see why the procedure is called Goodness of Fit.

Categorical Data Analysis (Goodness of Fit Test)

- The "closeness" between the observed and expected frequency is measured by computing the *Pearson's Chi-square statistic:*

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - F_i)^2}{F_i} = \sum_{i=1}^{k} \frac{(Obs.freq. - Exp.freq.)^2}{Exp.freq.}$$

Decision Rule: Reject H_0 if $\chi^2 > \chi^2_\alpha \quad (df = k-1)$

Categorical Data Analysis (Goodness of Fit Test)

- **Example 1:** (Mendel's Law of Genetics)

  Consider Mendel's famous Pea experiment leading to the fundamental result in Genetics. For this experiment, he considered two different traits of peas, namely color (yellow and green) and shape (round and wrinkled).

Categorical Data Analysis (Goodness of Fit Test)

- **Example 1:** (Mendel's Law of Genetics)

  In the second generation, he had four types of peas: yellow and round (both dominant characteristics), yellow and wrinkled (dominant and recessive), green and round (recessive and dominant), and green and wrinkled (both recessive).

  Mendel theorized that the peas in the four groups appear at the proportion 9:3:3:1.

Categorical Data Analysis (Goodness of Fit Test)

- **Example 1:** (Mendel's Law of Genetics)

  An animated explanation of this experiment can be found at the following link:
  http://www.mendel-museum.com/eng/1online/experiment.htm

Categorical Data Analysis (Goodness of Fit Test)

- **Example 1:** (Mendel's Law of Genetics)

  Mendel's actual experiment led to the following frequency table:

Category	*Frequency*
Round and Yellow	315
Wrinkled and Yellow	101
Round and Green	108
Wrinkled and Green	32
Total	556

Categorical Data Analysis (Goodness of Fit Test)

- We wish to test the following hypothesis:

H_0 : Mendel's hypothesis fits the data.

H_a : Mendel's hypothesis doesn't fit the data.

Categorical Data Analysis (Goodness of Fit Test)

- **Example 1:** (Mendel's Law of Genetics)

Category	f_i	F_i	$f_i - F_i$
Round and Yellow	315	312.75	2.25
Wrinkled and Yellow	101	104.25	-3.25
Round and Green	108	104.25	3.75
Wrinkled and Green	32	34.75	-2.75
Total	556	556	

Thus, $\chi^2 = \frac{(2.25)^2}{312.75} + \frac{(-3.25)^2}{104.25} + \frac{(3.75)^2}{104.25} + \frac{(-2.75)^2}{34.75} = 0.47$

Decision Rule: Reject H_0 if $\chi^2 > \chi^2_{0.05}$ (with 3 df) (= 7.81)

∴ Do not reject H_0

Categorical Data Analysis (Goodness of Fit Test)

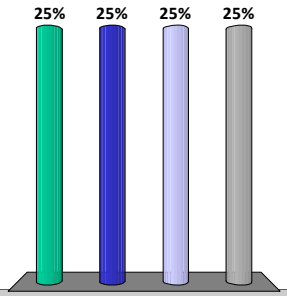
- **Example 2:**

Consider a study involving 147 industrial accidents that required medical attention. The data is listed according to the day that the accident occurred. Is there evidence to conclude that the accidents occur uniformly on all five working days of a week?

Day	M	T	W	R	F
# of accidents	31	42	18	25	31

What are the expected frequencies?

1. All are equal to 147.
2. ✔ All are equal to 29.4.
3. All we can say is that the expected frequencies add up to 147.
4. Divide observed frequency of each day of the week by 5.



Categorical Data Analysis (Goodness of Fit Test)

- To test:

H_0 : Accidents occur uniformly on all five days.

H_a : Accidents do not occur uniformly on all five days.

$$\text{Test statistic}: \chi^2 = \sum_{i=1}^{5} \frac{(f_i - F_i)^2}{F_i} = 10.65$$

Reject H_0 if $\chi^2 > \chi^2_{0.05}$ (df = 4) (Tabulated value = 9.48)

Reject H_0 and conclude that the accidents do not occur uniformly.

Two-way Classification

- This is a generalization of the one-factor model. In this case, the population is divided into categories according to the factor levels of two different factors. The main objective is to explore the relationship between the factors.

Example: In a study of 1000 deaths of males aged 45-64, the cause of death of each individual is listed along with their smoking habit. (Chartbook on smoking, tobacco and health, CDC 75-7511)

Two-way Classification

- **Example (cont'd):**

Cause of death

Smoking Habit		Cancer	Heart Disease	Other
	Smoker	135	310	205
	Nonsmoker	55	155	140

Is there evidence in the data to conclude that the smoking habit and cause of death are related factors?

Contingency Table

- A table showing the frequencies of occurrences of events characterized by the combination of factor levels of categorical covariates.
- If such a table has r rows and c columns, it is called a rxc contingency table.
- The cause of death vs. smoking habit data set is an example of a 2x3 contingency table.

Two-way Classification

- **Approach:**

  Suppose in general we have a *r* x *c* contingency table. We wish to test the hypothesis:

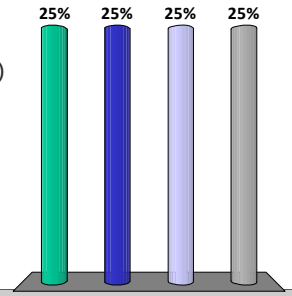
$$H_0 : \text{Two - factors are independent.}$$
$$H_a : \text{Two - factors are not independent.}$$

$$\text{Test statistic : } \chi^2 = \sum_{\text{all } r \text{ x } c \text{ cells}} \frac{(Obs.\,Freq. - Exp.\,Freq.)^2}{Exp.\,Freq.}$$

$$\text{We reject } H_0 \text{ if } \chi^2 > \chi^2_\alpha \ \ (\text{df } (r-1)(c-1))$$

Suppose two events A and B are independent. This means…

1. P(A or B) =P(A)+P(B)
2. P(A or B)=P(A)P(B)
3. P(A and B)=P(A)+P(B)
4. ✓P(A and B)=P(A)P(B)



Computation of Expected Frequencies

	Cancer	Heart Disease	Other	Total
Smoker	E_{11}	E_{12}	E_{13}	R_1
Nonsmoker	E_{21}	E_{22}	E_{23}	R_2
Total	C_1	C_2	C_3	G

Thus the objective is to figure out the expected frequencies of each cell under the model of independence when the row totals and columns totals remain the same as the observed table.

Two-way Classification

- **How to Compute Expected Frequency?**

$$\text{Expected Frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Cause of death

Smoking Habit		Cancer	Heart Disease	Other	Total
	Smoker	**135**	**310**	**205**	**650**
		(123.50)	(302.25)	(224.25)	
	Nonsmoker	**55**	**155**	**140**	**350**
		(66.50)	(162.75)	(120.75)	
		190	**465**	**345**	**1000**

Expected frequencies are in (..).

Two-way Classification

- **Thus** $\chi^2 = 8.349$

  $\text{Reject } H_0 \text{ if } \chi^2 > \chi^2_{0.05} \ \ (\text{df} = 2) \ (\text{Tabulated value} = 5.99)$

- **Reject H_0** and conclude that smoking habit and cause of death are related factors.

Two-way Classification

- Now that we have established that the cause of death and smoking habit are related factors, we can ask questions such as:

  Is there evidence to suggest that smokers are more likely to die from cancer than nonsmokers?

  Let: $p_1 = P(\text{died of cancer} \mid \text{smoker})$
  $p_2 = P(\text{died of cancer} \mid \text{nonsmoker})$

  We wish to test the hypothesis:

$$H_0 : p_1 = p_2 \text{ vs. } H_a : p_1 > p_2$$

Two-way Classification

- Test statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Note:

$$\hat{p}_1 = \frac{135}{650} = 0.21,\ \hat{p}_2 = \frac{55}{350} = 0.16,\ \hat{p} = \frac{135+55}{650+350} = 0.19$$

Thus,

$$Z = 1.92, \text{ and p - value} = P(Z > 1.92) = 0.0274$$

Two-way Classification

- Reject H_0 at 5% level of significance, suggesting the rate of cancer deaths among smokers is greater than that among non-smokers.