04/24/2023

Chapter 11 - Chi-Square and Analysis of Variance

Chi-Square: Applied to categorical data

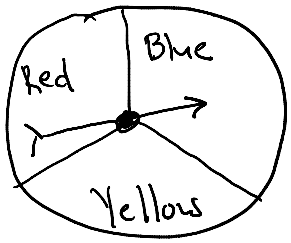
Analysis of Variance: Two or more independent random samples and compare $\mu_1, \mu_2, \mu_3, \ldots$

Categorical data — observations that are classified into categories

Ex: Smoking Status (Yes, No) } univariate categorical data
Gender (M, F) } bivariate categorical data
Eye color (brown, blue, green, hazel, other) } Multivariate Categorical data

Univariate categorical data can be summarized in a one-way frequency table.

Ex: Spinner



Spin 90 times and record color

	Color		
	Blue	Red	Yellow
Frequency	25	25	40

k = number of categories of categorical variable

In our case, $k = 3$

n = sample size

In our case, $n = 90$

Let p_1 = population proportion of time spinner lands on blue

p_2 = " " " " " " " red

p_3 = " " " " " " " yellow

Test if the spinner is fair.

i.e., $p_1 = 1/3$, $p_2 = 1/3$, $p_3 = 1/3$?

Ratios are 1:1:1

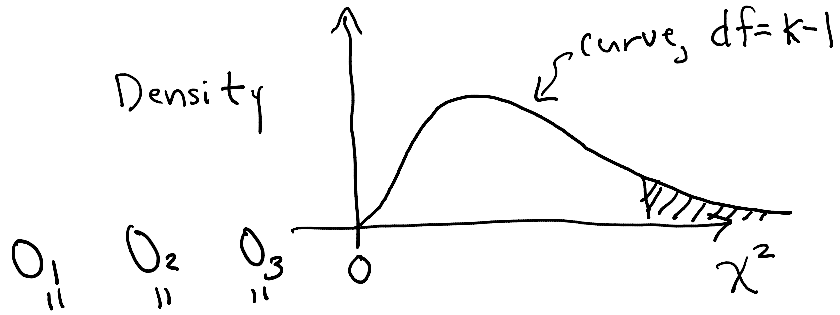
H_0: $p_1 = \frac{1}{3}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{3}$ (spinner is fair)

H_1: At least one of the p_i's differs from its corresponding hypothesized value (spinner is not fair)

This test is called a χ^2 Goodness-of-fit test.

χ^2 = "Chi-square"

(See new handout Table A-4)



	O_1	O_2	O_3
	=	=	=
Observed Frequencies	25	25	40

Expected Frequencies (under H_0)	30	30	30
	$\|\|$	$\|\|$	$\|\|$
	E_1	E_2	E_3

In general, $E_1 = np_1$, $E_2 = np_2$, $E_3 = np_3$.

A useful measure of discrepancy between observed and expected frequencies for each category is

$$\frac{(O_i - E_i)^2}{E_i}, \quad i = 1, 2, \ldots, k$$

Category 1: $\frac{(O_1 - E_1)^2}{E_1} = \frac{(25-30)^2}{30} = \frac{25}{30} = \frac{5}{6}$

Category 2: $\frac{(O_2 - E_2)^2}{E_2} = \frac{(25-30)^2}{30} = \frac{25}{30} = \frac{5}{6}$

Category 3: $\frac{(O_3 - E_3)^2}{E_3} = \frac{(40-30)^2}{30} = \frac{100}{30} = \frac{10}{3}$

Overall measure of discrepancy is

Observed test statistic: $\sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} = \frac{5}{6} + \frac{5}{6} + \frac{10}{3} = \frac{30}{6} = 5.$

Reject H_0, in our example, when $\sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i}$ is "large".

In general, H_0: p_1 = hypothesized proportion for category 1
p_2 = " " " " 2
$\vdots$
p_k = " " " " k

$\vdots$

$p_k =$ " " " " k

H_1: H_0 is not true

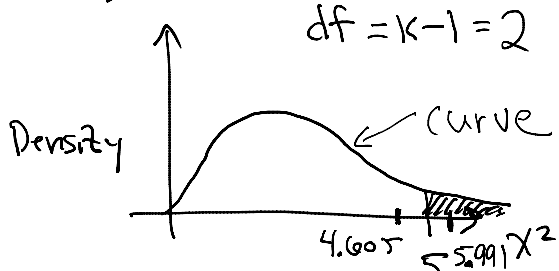
test statistic $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ where $E_i = np_i$, for $i = 1, \ldots, k$

If H_0 is true, then the test statistic follows a χ^2- distribution with $df = k-1$.

04/26/2023

P-value = $P(\chi^2 \geq$ observed test statistic$)$

Reject H_0 when P-value $\leq \alpha$.



Review: H_0: $p_1 = \frac{1}{3}, p_2 = \frac{1}{3}, p_3 = \frac{1}{3}$ (spinner example)

H_1: spinner is not fair

$\alpha = 0.10$

observed test statistic = 5

$0.05 <$ P-value $= P(\chi^2 \geq 5) < 0.10$

P-value $\leq \alpha$ so reject H_0. Based on sample, one concludes that the spinner is not fair.

Remark: χ^2 goodness-of-fit test is applicable when it is based on a random sample of frequency counts from different categories

and $E_i \geq 5$, $i = 1, 2, \ldots, k$.

Ex: (see Handout again) Ratios 4:4:1:1

$k = 4$

$H_0: p_1 = \frac{4}{10}, p_2 = \frac{4}{10}, p_3 = \frac{1}{10}, p_4 = \frac{1}{10}$

or $\begin{cases} H_0: p_1 = 0.4, p_2 = 0.4, p_3 = 0.1, p_4 = 0.1 \\ H_1: H_0 \text{ is not true} \end{cases}$

observed: 38 43 10 5

expected: 38.4 38.4 9.6 9.6

For example, $E_1 = np_1 = 96 \times 0.4 = 38.4$ $\quad \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$

Verify that the observed test statistic is 2.78.

$$P\text{-value} = P(\chi^2 \geq 2.78) > 0.10 \quad \text{where } df = 3.$$

So, P-value $> \alpha$ and we fail to reject H_0.

NOTE: Write conclusion in the context problem.