IV-3-4. χ^2 test

Homoscedasticity and normal distribution are necessary condition of analysis of variance including t test and F test. There are many datasets which do not have homoscedasticity. For example, when we detect the impact of medication by the ratio of people who shows the effect of the treatment, there is no impact in low dosage and data of the ratio will reach a saturated level in higher dosage. In such cases, data fluctuation is smaller in low and high dosage and wider in middle dosage. When we detect the relation of dosage level and the impact by proportion of people who show the effect of the medicine in the people who are administered same level of dose, the impact can be detect upper level than border. The border has range. The data fluctuate in the range, and the fluctuation is stable at 0% or 100% out of the range.   Such phenomena are explained by figure 41.

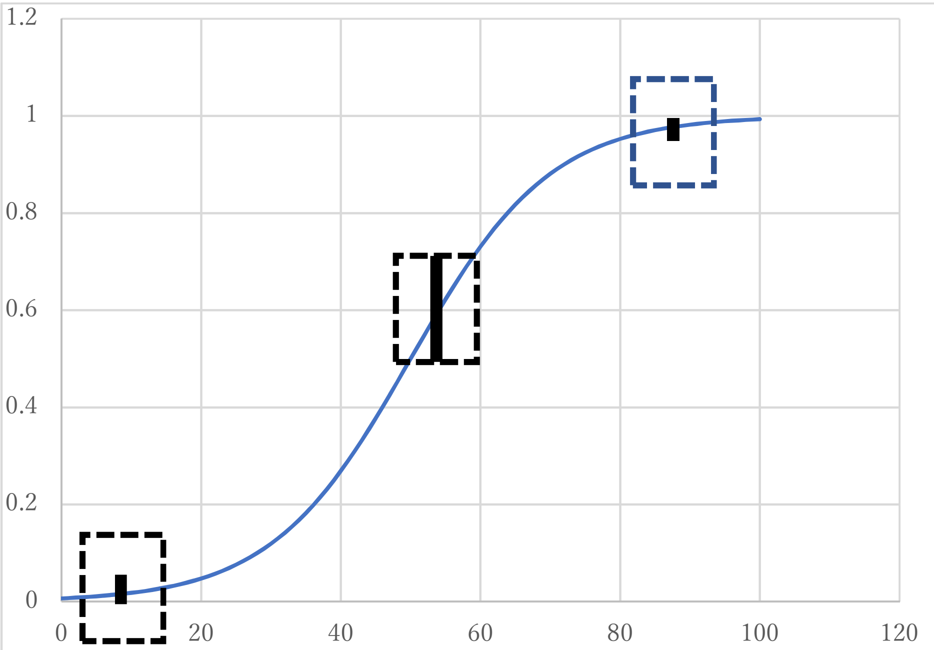

Fig. 41. Nature of rate data

Blue line in figure 41 is sigmoid curve, though we do not need to consider definition of sigmoid function. Here it is enough to accept the line is S shape to understand the explanation.   The slope of the curve is gentle in lower and higher part and the slope is steep in middle part. Effect of dose is different among people depending on body size, health condition and so on. This means that the curve in the figure fluctuates horizontally. There are three rectangles in the figure. The shape and size of three rectangles are the same and width of the rectangle shows range of horizontal fluctuation (fluctuation of coordinate can be replaced with fluctuation of line). Black bars in the rectangles show vertical fluctuation of the data in the range. Vertical fluctuation is highest around inflection point of S shape curve. As the result, we cannot expect homoscedasticity among levels of explanatory variable, when the dataset shows S shape curve totally. There are many phenomena which show S shape relation. All cumulative distribution functions are S shape or S shape like. In the case of experiment of medication, the ratio of affected people is a kind of cumulative value. Text books of statistics are suggesting avoidance of use of analysis of variance for analysis of proportional data, because proportional data has no homoscedasticity in many cases. It is not enough, because there many datasets which have no homoscedasticity other than proportional data. The author suggests confirmation of homoscedasticity before analysis by checking variances of each sub sample group or by consideration of nature of the dataset.

Chai square test is applicable to dataset of which homoscedasticity is debatable. Many text books recommend various analytical methods other than χ^2 test. Most of them are so called "nonparametric analysis". Chai square is parametric, and it includes concept of ratio of variance. In this meaning, the author is considering that χ^2 test is a kind of analysis of variance. Requirement of χ^2 test is existence of expectation value, because χ^2 distribution is distribution of ratio of variance of difference between observed value and expectation value and variance of expectation value (See III-2-5. Chai square distribution)

Observed χ^2 value is as follow.

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

f_i: observed data in sub sample population i

e_i: expectation value of f_i

Formula 47

This formula is commonly introduced in many text books. In the case if proportional data, probability distribution of each proportional data is binomial distribution. In binomial distribution $e_i = \frac{\sigma_i^2}{2}$

And the formula can be expressed as follow.

$$\chi^2 = \sum \frac{2(f_i - \mu_i)^2}{\sigma_i^2}$$

From this transformation we can understand that χ^2 test is a comparison of variances.

Often used example is analysis of distortion of dice. For the analysis we set null hypothesis that the dice has no distortion. From this null hypothesis, we can calculate expectation values. Following is the calculation table when we roll 600 times. Expectation value of times of each pip is 100.

pip	observed	expectation	difference	difference^2	
	A	B	C=A-B	D=C^2	E=D/B
1	120	100	20	400	4
2	85	100	-15	225	2.25
3	113	100	13	169	1.69
4	115	100	15	225	2.25
5	80	100	-20	400	4
6	87	100	-13	169	1.69
sum	600	600			15.88

Observed value of χ^2 is 15.88 and degree of freedom is 6-1=5.
The threshold of χ^2 at degree of freedom 5 and risk rate $p \leq 0.01$ is 15.0863
From this, we can reject the null hypothesis that the dice has no distortion.
In the case of dice, we do not need to consider what should be expectation value. However, we should consider what should be expectation value, for application of this method for significance test of difference.
Table 38 is an example of such case. There are 2 medicines. When we administered medicine A to 116 people, the medicine was effective to 66 persons. When we administered medicine B to 97 people, the medicine is effective to 56 persons.

Table 38. Result of administration of two drugs expressed in actual counts.

	A	B	sum
effective	66	56	122
ineffective	50	41	91
sum	116	97	213

Table 39. Result of administration of two drugs expressed in ratio.

	A	B	Sum
effective	0.568966	0.57732	0.57276995
ineffective	0.431034	0.42268	0.42723005
sum	1	1	1

Here, we assume the effects of A and B are the same. This is null hypothesis that the ratio in each medicine group should be similar to the ratio of total group (0.57276995 and 0.42723005).

Following is an example of calculation of Chai square. From this we can make expectation values multiplying the ration in total to number of each group.

$116 \times 0.57276995, \quad 116 \times 0.42723005, \quad 97 \times 0.57276995, \quad 97 \times 0.42723005$

Expectation values

	A	B	sum
effective	66.44131	55.55869	122
ineffective	49.55869	41.44131	91
sum	116	97	213

Differences of observed value and expectation value

	A	B	Sum
effective	-0.44131	0.44131455	0
inefective	0.441315	-0.4413146	0
sum	0	0	0

Square of differences

	A	B
effecitive	0.194759	0.194759
ineffective	0.194759	0.194759

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$$

	A	B	Sum
effective	0.002931	0.00350546	0.00643674
inefective	0.00393	0.00469962	0.00862948
sum	0.006861	0.00820508	0.01506622

Observed value of χ^2 is 0.01506622

Degree of freedom is $(2-1)(2-1) = 1$

This value is too small to compare the threshold value of χ^2, and we cannot reject the null hypothesis that the effect of medicine A and B is similar. In other word, we cannot say that effects of medicine A and B is different.

We can apply χ^2 test to more complicated analysis of dataset.

Social status is not richness. Even poorest people may think that they are belonging high society, because of their cultural refinement. It is relating to the culture of the area. Following of the data of interview survey asking social level of people. Which do you think you are belonging, high society, middle society or low society. The survey is implemented in A, B, C and D city. We want to know whether there are differences among cities.

Table 40. Comparison of mental social level among cities.

	A	B	C	D	Sum	ratio
high	10	20	5	15	50	0.208333
middle	35	30	30	50	145	0.604167
low	15	0	20	10	45	0.1875
sum	60	50	55	75	240	1

Calculation of chai square

Expectation value

	A	B	C	D	Sum
high	12.5	10.41667	11.45833	15.625	50
middle	36.25	30.20833	33.22917	45.3125	145
low	11.25	9.375	10.3125	14.0625	45
sum	60	50	55	75	240

Difference between observed value and expectation value

	A	B	C	D
high	-2.5	9.583333	-6.45833	-0.625
middle	-1.25	-0.20833	-3.22917	4.6875
low	3.75	-9.375	9.6875	-4.0625

Square of difference

	A	B	C	D
high	6.25	91.84028	41.71007	0.390625
middle	1.5625	0.043403	10.42752	21.97266
low	14.0625	87.89063	93.84766	16.50391

Calculation of χ^2 value $\sum\frac{(f_i-e_i)^2}{e_i}$

	A	B	C	D	Sum
high	0.5	8.816667	3.640152	0.025	12.98182
middle	0.043103	0.001437	0.313806	0.484914	0.84326
low	1.25	9.375	9.100379	1.173611	20.89899
sum	1.793103	18.1931	13.05434	1.683525	34.72407

Observed χ^2 is 34.72407

Degree of freedom of city is 4-1=3

Degree of freedom of social class is 3-1=2

So, the freedom of this dataset is (4-1)(3-1)=6

Threshold of χ^2 value at df=6 $p \leq 0.005$ is 18.5476

Form this result of the analysis, we can say that the social mind of people who lives in 4 cities are not same.