Qualitative Issues of a non parametric test

Khagendra kumar
www.politicindia.com

Types of Data

- Basically two types of random variables and they yield two types of data: numerical and categorical.
- A chi square (X^2) statistic is used to investigate whether distributions of categorical variables differ from one another.
- Basically categorical variable yield data in the categories and numerical variables yield data in numerical form.
- Responses to such questions as "What is your elective?" or Do you own a bike?" are categorical because they yield data such as “Women education" or “yes.“
- In contrast, responses to such questions as "How tall are you?" or "What is your G.P.A.?" are numerical. Numerical data can be either discrete or continuous.

Data Type	Question Type	Possible Responses
Categorical	What is your gender?	male or female
Numerical	Disrete- How many cars do you own?	two or three
Numerical	Continuous - How tall are you?	72 inches

Discrete data arise from a counting process, while continuous data arise from a measuring process.

The Chi Square statistic compares the tallies or counts of categorical responses between two (or more) independent groups. (note: Chi square tests can only be used on actual numbers and not on percentages, proportions, means, etc.)

Chi Square Goodness of Fit (One Sample Test)

- This test allows us to compare a collection of categorical data with some theoretical expected distribution.
- This test is often used in genetics to compare the results of a cross with the theoretical distribution based on genetic theory.
- Suppose you preformed a simple monohybrid cross between two individuals that were heterozygous for the trait of interest.
- Tt x Tt

- A test based upon the Chi-squared distribution is a nonparametric test.
- Nonparametric tests determine the probability that an observed distribution of data, based upon rankings or distribution into categories of a qualitative nature, is due to chance (sampling error) alone.
- If you have numbers that appear to follow a normal or t-distribution, then you would want to use a parametric test such as 'Student's' t test to address your question.
- The chi-square test is very useful, especially when data are not quantitative. It will probably be most effective to explain the process

Results of a monohybrid cross between two heterozygotes for the ‘t' gene

	T	t	Totals
T	10	42	52
t	33	15	48
Totals	43	57	100

- The penotypic ratio 85 of the T type and 15 of the t-type (homozygous recessive). In a monohybrid cross between two heterozygotes, however, we would have predicted a 3:1 ratio of phenotypes.
- In other words, we would have expected to get 75 T-type and 25 t-type. Are our results different?

Another examples for goodness of fit

- **Is the distribution of pine trees related to soil type?**
- Is smoking habit related to skin colours/stress/ reading habits
- Goodness of fit although involves some statistical basis but it forms the basis for qualitative study
- It has major practical significance in social-behavioual and natural sciences

Calculate the chi square statistic x^2 by completing the following steps:

1.For each *observed* number in the table subtract the corresponding *expected* number (*O* — *E*).

2.Square the difference [(O —E)^2].

3.Divide the squares obtained for each cell in the table by the *expected* number for that cell [(O - E)^2 / E].

4.Sum all the values for (O - E)^2 / E. This is the chi square statistic.

For our example, the calculation would be:

$x^2 = 5.33$

	Observed	Expected	(O — E)	(O — E)2	(O — E)2/ E
T-type	85	75	10	100	1.33
t-type	15	25	10	100	4.0
Total	100	100			chi square 5.33

Chi Square distribution tab

probability level (alpha)

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	**3.841**	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

Discussion on goodness of fit

- Frequently, however, there are research problems in which one wants to make directinferences about two or more distributions, either by asking if a population distribution has some particular specifiable form, or by asking if two or more population distributions are identical.
- These questions occur most often when variables are qualitative in nature, making it impossible to carry out the usual inferences in terms of means or variances. For such problems, we usenonparametric methods.
- Nonparametric methods (1) do not depend on any assumptions about the parameters of the parent population (2) generally assume data are only measured at the nominal or ordinal level

There are two common types of hypothesis-testing problems that are addressed with nonparametric methods:

(1) How well does a sample distribution correspond with a hypothetical population distribution? As you might guess, the best evidence one has about a population distribution is

the sample distribution.

The greater the discrepancy between the sample and theoretical distributions, the more we question the "goodness" of the theory.

EX: Suppose we wanted to see whether the distribution of educational achievement had changed over the last 25 years. We might take as our null hypothesis that the distribution of educational achievement had not changed, and see how well our modern-day sample supported that theory.

(2) We often wish to find evidence for association between two qualitative variables - hence we analyze cross-classifications of two discrete distributions.

EX: What is the relationship between sex and party vote - are women more likely than men to support Democratic party candidates?

Chi Square Test of Independence

- For a contingency table that has *r* rows and *c* columns, the chi square test can be thought of as a test of independence. In a test of independence the null and alternative hypotheses are:
- Ho: The two categorical variables are independent.
- Ha: The two categorical variables are related.
- We can use the equation Chi Square = the sum of all the $(fo - fe)^2 / fe$
- Here *f*o denotes the frequency of the observed data and *f*e is the frequency of the expected values. The general table would look something like in the text slide:

	Category I	Category II	Category III	Row Totals
Sample A	a	b	c	a+b+c
Sample B	d	e	f	d+e+f
Sample C	g	h	i	g+h+i
Column Totals	a+d+g	b+e+h	c+f+i	a+b+c+d+e+f+g+h+i=N

Table : Incidence of three types of malaria in three tropical regions.

We could now set up the following table:

	Asia	Africa	South America	Totals
Malaria A	31	14	45	90
Malaria B	2	5	53	60
Malaria C	53	45	2	100
Totals	86	64	100	250

Chi Square = 125.516

Observed	Expected	|O -E|	(O — E)2	(O — E)2/ E
31	30.96	0.04	0.0016	0.0000516
14	23.04	9.04	81.72	3.546
45	36.00	9.00	81.00	2.25
2	20.64	18.64	347.45	16.83
5	15.36	10.36	107.33	6.99
53	24.00	29.00	841.00	35.04
53	34.40	18.60	345.96	10.06
45	25.60	19.40	376.36	14.70
2	40.00	38.00	1444.00	36.10

What result says?

- Thus, we would reject the null hypothesis that there is no relationship between location and type of malaria.
- Our data tell us there is a relationship between type of malaria and location, but that's all it says.
- Can’t explain Kind of relationship or any further relationship
- Test of independence prepares ground for further exploration.

CONCLUDING REMARKS

- **The chi-square distribution is easy to work with, but there are some important differences between it and the Normal distribution or the T distribution. Note that**
- **The chi-square distribution is NOT symmetric**
- **T All chi-square values are positive**
- **As with the T distribution, the shape of the chi-square distribution depends on the degrees of freedom.**
- **Hypothesis tests involving chi-square are usually one-tailed. We are only interested in whether the observed sample distribution significantly differs from the hypothesized distribution. We therefore look at values that occur in the upper tail of the chi-square distribution. That is, low values of chi-square indicate that the sample distribution and the hypothetical distribution are similar to each other, high values indicate that the distributions are dissimilar.**
- **A random variable has a chi-square distribution with N degrees of freedom if it has the same distribution as the sum of the squares of N independent variables, each normally distributed, and each having** expectation 0 and variance 1

THANK U