Chi-square test, Fisher's Exact test, McNemar's Test

Mikhail Dozmorov

Spring 2021

Tests for enrichment

- Fisher's exact
- Hypergeometric
- Binomial
- Chi-squared
- Z
- Kolmogorov-Smirnov
- Permutation
- ...

Chi-square test for comparisons between 2 categorical variables

Test for independence between two variables

- The null hypothesis for this test is that the variables are independent (i.e. that there is no statistical association).
- The alternative hypothesis is that there is a statistical relationship or association between the two variables.

Test for equality of proportions between two or more groups

- The null hypothesis for this test is that the 2 proportions are equal.
- The alternative hypothesis is that the proportions are not equal (test for a difference in either direction)

Contingency tables

- Let X_1 and X_2 denote categorical variables
- X_1 having I levels and X_2 having J levels. There are IJ possible combinations of classifications.

	Level 1	Level 2	...	Level J
Level 1				
Level 2				
...				
Level I				

- When the cells contain frequencies of outcomes, the table is called a contingency table.

Chi-square Test: Testing for Independence

Step 1: Hypothesis (always two-sided):

- H_0: Independent
- H_A: Not independent

Step 2: Calculate the test statistic:

$$X^2 = \sum \frac{(x_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2 \text{ with } df = (I-1)(J-1)$$

Chi-square Test: Testing for Independence

Step 3: Calculate the p-value

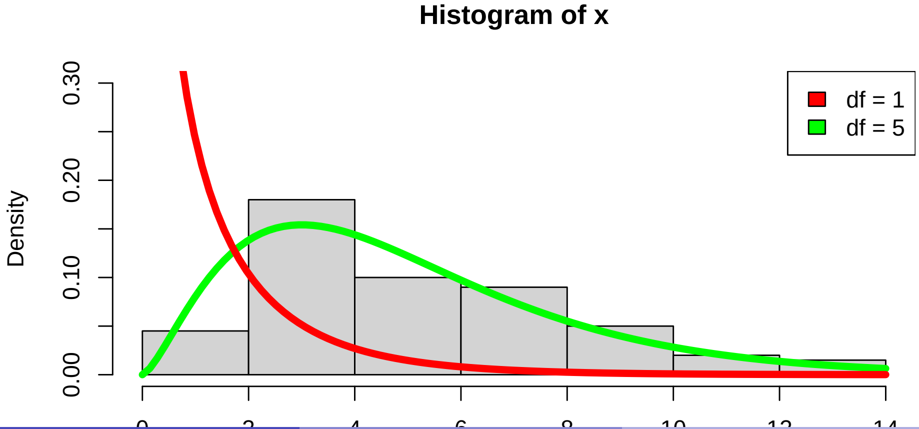
- $p - value = P(\chi^2 > X^2)$ - 2-sided p-value

Step 4: Draw a conclusion

- $p - value < \alpha$ - reject independence
- $p - value > \alpha$ - do not reject indepencence

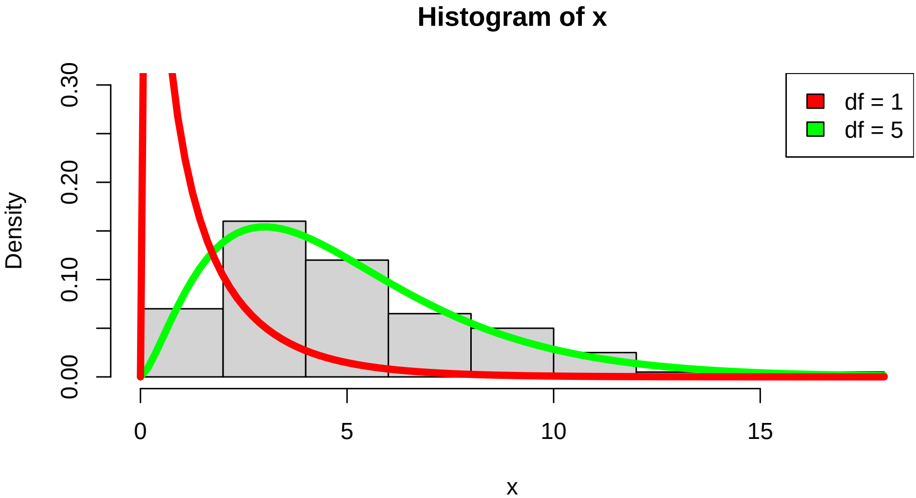
χ^2 Distribution

- The chi-square distribution with 1 df = the square of the Z distribution.
- Since the distribution only has positive values all the probability is in the right-tail.



χ^2 Distribution

- Critical value for $\alpha = 0.05$ and Chi-square with 1 df is 3.8414588
- $qchisq(1 - 0.05, 5)$



Chi-square Test: Testing for Independence

- Expected frequencies are calculated under the null hypothesis of independence (no association) and compared to observed frequencies.
- Recall: A and B are independent if: $P(A\ and\ B) = P(A) * P(B)$
- Use the Chi-square (X^2) test statistic to observe the difference between the observed and expected frequencies.

Chi-square Test: Testing for Independence

- Calculating expected frequencies for the observed counts

	Diff. exp. genes	Not Diff. exp. genes	Total
In gene set	84	3132	3216
Not in gene set	24	2886	2910
Total	108	6018	6126

- Under the assumption of independence:

$$P(In\ gene\ set\ AND\ Diff.\ exp.\ genes) = P(In\ gene\ set) * P(Diff.\ exp.\ genes)$$
$$= (108/6126) * (3216/6126) = 0.009256$$

- Expected cell count
  $E_{In\ gene\ set\ and\ Diff.\ exp.\ genes} = 0.009256 * 6126 = 56.70$
- ToDo: Calculate other expected cell counts

Chi-square Test: Testing for Independence

- Calculating expected frequencies for the observed counts

	Diff. exp. genes	Not Diff. exp. genes	Total
In gene set	56.70	3159.30	3216
Not in gene set	51.30	2858.70	2910
Total	108	6018	6126

- Expected Cell Counts = (Marginal Row total * Marginal Column Total) / n
- Check to see if expected frequencies are > 2
- No more than 20% of cells with expected frequencies < 5

Chi-square Test: Testing for Independence

- Calculate the test statistics $X^2 = \sum \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$

$$X^2 = \frac{(84-56.70)^2}{56.70} + \frac{(3132-3159.30)^2}{3159.30} + \frac{(24-51.30)^2}{51.30} + \frac{(2886-2858.70)^2}{2858.70}$$
$$= 13.24 + 0.26 + 14.59 + 0.34 = 27.152$$

- Calculate the p-value $p - value = P(X^2 > 27.152) = pchisq(27.152, df = 1, lower.tail = FALSE) = 0.000000188$
- Draw a conclusion:
  - If $p - value < \alpha$ - reject independence.
  - A significant association exists between differentially expressed genes and the selected gene set
  - Differentially expressed genes may affect functions of this gene set

Other applications of chi-square test: Equality or Homogeneity of Proportions

- Testing for equality or homogeneity of proportions – examines differences between proportions drawn from two or more independent populations.
- Example of two populations:
  - 100 differentially expressed genes, classified as within/outside of a pathway
  - 100 randomly selected genes, classified as within/outside of a pathway

Chi-Square Testing for independence

Hypotheses

- H_0: two classification criteria are independent
- H_A: two classification criteria are not independent.

Requirements

- One sample selected randomly from a defined population.
- Observations cross-classified into two nominal criteria.
- Conclusions phrased in terms of independence of the two classifications.

Chi-Square Testing for Equality

Hypotheses

- H_0: populations are homogeneous with regard to one classification criterion.
- H_A: populations are not homogeneous with regard to one classification criterion.

Requirements

- Two or more samples are selected from two or more populations.
- Observations are classified on one nominal criterion.
- Conclusions phrased with regard to homogeneity or equality of treatment

https://www.graphpad.com/quickcalcs/contingency1.cfm

Small Expected Frequencies

- Chi-square test is an approximate method.
- The chi-square distribution is an idealized mathematical model.
- In reality, the statistics used in the chi-square test are qualitative (have discrete values and not continuous).
- For 2 X 2 tables, use Fisher's Exact Test (i.e. $P(x = k) \sim B(n, p)$) if your expected frequencies are less than 2.

Fisher's Exact Test: Description

The Fisher's exact test calculates the exact probability of the table of observed cell frequencies given the following assumptions:

- The null hypothesis of independence is true
- The marginal totals of the observed table are fixed

Fisher's Exact Test: Description

Calculation of the probability of the observed cell frequencies uses the factorial mathematical operation.

- Factorial is notated by ! which means multiply the number by all integers smaller than the number
- Example: $7! = 7 * 6 * 5 * 4 * 3 * 2 * 1 = 5040$.

Fisher's exact test: Calculations

a	b	a+b
c	d	c+d
a+c	b+d	n

If margins of a table are fixed, the exact probability of a table with cells a, b, c, d and marginal totals $(a+b), (c+d), (a+c), (b+d) = \frac{(a+b)! * (c+d)! * (a+c)! * (b+d)!}{n! * a! * b! * c! * d!}$

Fisher's Exact Test: Calculation Example

1	8	9
4	5	9
5	13	18

The exact probability of this table is $\frac{9! * 9! * 13! * 5!}{18! * 1! * 8! * 4! * 5!} = \frac{136080}{1028160} = 0.132$

Probability for all possible tables with the same marginal totals

The 6 possible tables for the observed marginal totals: 9, 9, 5, 13.

0	9	9
5	4	9
5	13	18

Pr = 0.0147

1	8	9
4	5	9
5	13	18

Pr = 0.132, this is for the observed table

Additional possible tables with marginal totals: 9,9,5,13

2	7	9
3	6	9
5	13	18

$\Pr = 0.353$

3	6	9
2	7	9
5	13	18

$\Pr = 0.353$

Additional possible tables with marginal totals: 9,9,5,13

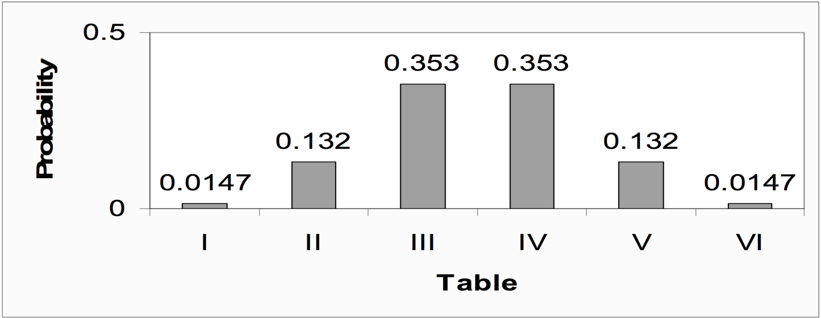
4	5	9
1	8	9
5	13	18

$\Pr = 0.132$

5	4	9
0	9	9
5	13	18

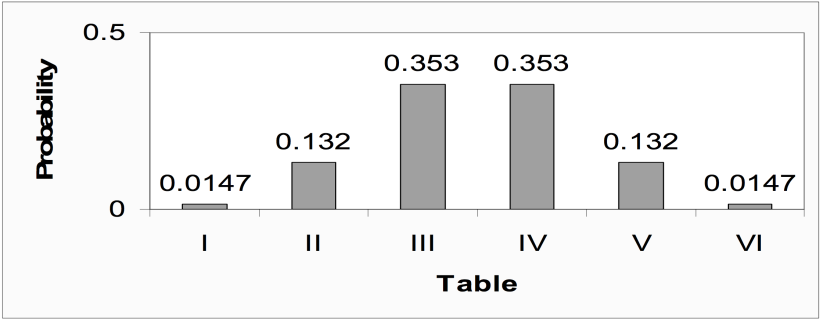
$\Pr = 0.0147$

Fisher's Exact Test: p-value



- The p-value for the Fisher's exact test is calculated by summing all probabilities less than or equal to the probability of the observed table.
- The probability is smallest for the tables (tables I and VI) that are least likely to occur by chance if the null hypothesis of independence is true.

Fisher's Exact Test: p-value



- The observed table (Table II) has probability $= 0.132$
- P-value for the Fisher's exact test $=$ Pr (Table II) $+$ Pr (Table V) $+$ Pr (Table I) $+$ Pr (Table VI) $= 0.132 + 0.132 + 0.0147 + 0.0147 = 0.293$

Conclusion of Fisher's Exact test

- At significance level 0.05, the null hypothesis of independence is not rejected because the p-value of $0.294 > 0.05$.
- Looking back at the probabilities for each of the 6 tables, only Tables I and VI would result in a significant Fisher's exact test result: $p = 2 * 0.0147 = 0.0294$ for either of these tables.
- This makes sense, intuitively, because these tables are least likely to occur by chance if the null hypothesis is true.

Fisher's exact test p-value formula

$$p(r) = \sum_{k=0}^{min(c,d)} \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+k)!(b-k)!(c-k)!(d+k)!n!}$$

Can Chi-squared or Fisher's tests be used if your data is categorical?

- When data are paired and the outcome of interest is a proportion, the **McNemar Test** is used to evaluate hypotheses about the data.
- Developed by Quinn McNemar in 1947
- Sometimes called the McNemar Chi-square test because the test statistic has a Chi-square distribution
- The McNemar test is only used for paired nominal data.
- Use the Chi-square test for independence when nominal data are collected from independent groups.

Examples of Paired Data for Proportions

Pair-Matched data can come from

- Case-control studies where each case has a matching control (matched on age, gender, race, etc.)
- Twins studies – the matched pairs are twins.

Before - After data

- The outcome is presence (+) or absence (-) of some characteristic measured on the same individual at two time points.

Summarizing the Data

- Like the Chi-square test, data need to be arranged in a contingency table before calculating the McNemar statistic
- The table will always be 2 X 2 but the cell frequencies are numbers of 'pairs' not numbers of individuals

Pair-Matched Data for Case-Control Study: outcome is exposure to some risk factor

		Control	
		Exposed	Unexposed
Case	Exposed	a	b
	Unexposed	c	d

- a - number of case-control pairs where both are exposed
- b - number of case-control pairs where the case is exposed and the control is unexposed
- c - number of case-control pairs where the case is unexposed and the control is exposed
- d - number of case-control pairs where both are unexposed

Paired Data for Before-After counts

- The data set-up is slightly different when we are looking at 'Before-After' counts of some characteristic of interest.
- For this data, each subject is measured twice for the presence or absence of the characteristic: before and after an intervention.
- The 'pairs' are not two paired individuals but two measurements on the same individual.
- The outcome is binary: each subject is classified as + (characteristic present) or – (characteristic absent) at each time point.

Null hypotheses for Paired Nominal data

- The null hypothesis for case-control pair matched data is that the proportion of subjects exposed to the risk factor is equal for cases and controls.
- The null hypothesis for twin paired data is that the proportions with the event are equal for exposed and unexposed twins
- The null hypothesis for before-after data is that the proportion of subjects with the characteristic (or event) is the same before and after treatment.

McNemar's test

- For any of the paired data, the following are true if the null hypothesis is true:

$$b = c$$

$$b/(b + c) = 0.5$$

- Since cells b and c are the cells that identify a difference, only cells b and c are used to calculate the test statistic.
- Cells b and c are called the discordant cells because they represent pairs with a difference
- Cells a and d are the concordant cells. These cells do not contribute any information about a difference between pairs or over time so they aren't used to calculate the test statistic.

McNemar Statistic

- The McNemar's Chi-square statistic is calculated using the counts in the b and c cells of the table:

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

- Square the difference of $(b - c)$ and divide by $b + c$.
- If the null hypothesis is true the McNemar Chi-square statistic $= 0$.

McNemar statistic distribution

- The sampling distribution of the McNemar statistic is a Chi-square distribution.
- Since the McNemar test is always done on data in a 2 X 2 table, the degrees of freedom for this statistic $= 1$
- For a test with $\alpha = 0.05$, the critical value for the McNemar statistic $= 3.84$.
  - The null hypothesis is not rejected if the McNemar statistic < 3.84.
  - The null hypothesis is rejected if the McNemar statistic > 3.84.