Introductory Statistics Lectures

Tests of independence and homogeneity

Contingency tables

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

Redistribution of this material is prohibited without written permission of the author

© 2009

(Compile date: Tue May 19 14:51:09 2009)

Contents

1 Tests of independence and homogeneity

1.1 Review

The following is a partial list of statistical methods that we have discussed:

1. mean
2. median
3. mode
4. standard deviation
5. z-score
6. percentile
7. coefficient of variation
8. scatter plot
9. histogram
10. boxplot
11. normal-quantile plot
12. confidence interval for mean
13. confidence interval for difference in means
14. confidence interval for proportion

15. confidence interval for difference in proportions
16. one sample mean test
17. two independent sample mean test
18. match pair test
19. one sample proportion test
20. two sample proportion test
21. linear correlation coefficient & test
22. regression

For each situation below, which method is most applicable?

- If it's a hypothesis test, what are the null and alternative (state in words and mathematically).
- If it's a graphical method, describe what you would be looking for.

Question 1. A news reporter wants to conduct a poll to determine the proportion of support for Hillary Clinton.

Question 2. A researcher wants to determine if there is a relationship between height and income.

Question 3. A statistics professor decides it is too much work to give a final exam and decides to predict a student's final exam grade based on a student's midterm grade (using past semester data).

Question 4. A researcher wants to determine if the incidence of Down syndrome in males and females is the same.

1.2 Introduction

Question 5. We previously looked at correlation and regression which dealt with what type of data? (quantitative or categorical)

Example 1. A group of 276 healthy men and women were grouped according to their number of relationships. They were then exposed to a virus that caused colds. The data is summarized in the table below. Does the data provide sufficient evidence to indicate that susceptibility to colds is affected by the number of relationships you have?

	Number of relationships		
Contracted cold?	3 or less	4-5	6 or more
yes	49	43	34
no	31	47	62

What if we looked at a simpler version of this example:

	Number of relationships	
Contracted cold?	3 or less	4 or more
yes	49	77
no	31	109

Question 6. Do we know any statistical techniques to determine if there is a difference in the cold contraction between the two groups?

CONTINGENCY TABLE. DEFINITION 1.1

A contingency table is a table that lists the frequencies of occurrence for categories of **two variables**.

- Rows are the first variable.
- Columns are the second variable

Caution: There are only **two variables** even though there can be many rows and columns. Each row represent a levels of the first variable. Each column represents a level of the second variable.

1.3 Test of Homogeneity and Independence

USE

Often used to help answer:

1. Is the proportion of x the same in all the populations? (homogeneity)
2. Is the proportion of x different in at least one of the populations? (homogeneity)
3. Does one of many processes under evaluation have a higher proportion of x? (homogeneity)
4. Are X and Y dependent? (independence)
5. Are X and Y independent? (independence)
6. Is there a relationship between X and Y? (independence)

COMPUTATION

We will now look at methods for analyzing relationships for **categorical** data. Recall that correlation and regression were used to quantify **quantitative** data relationships.

DEFINITION 1.2 TEST OF HOMOGENEITY.

A test of homogeneity tests the null hypothesis that different populations have the same proportions of some characteristics. The key difference from the test of independence is that there are **multiple populations** that the data is drawn from.

H_0 $p_1 = p_2 = \cdots = p_n$ the proportion of X is the same in all the populations studied.

H_a At least one proportion of X is not the same.

DEFINITION 1.3 TEST OF INDEPENDENCE.

A test of independence tests the null hypothesis that there is no association between the two variables in a contingency table where the data is all drawn from **one population**.

H_0 X and Y are independent.

H_a X and Y are dependent.

For both of these tests, other than the hypothesis, all the procedures are the same!

Test of independence and Homogeneity

Requirements (1) Sample data are randomly selected, (2) For each cell in the contingency table the expected frequency $E_i \geq 5$. (3) Individual observations must be independent.[1](No distribution requirement.)

Test Statistic :

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

[1]If dependent observations, use McNemar's test.

where

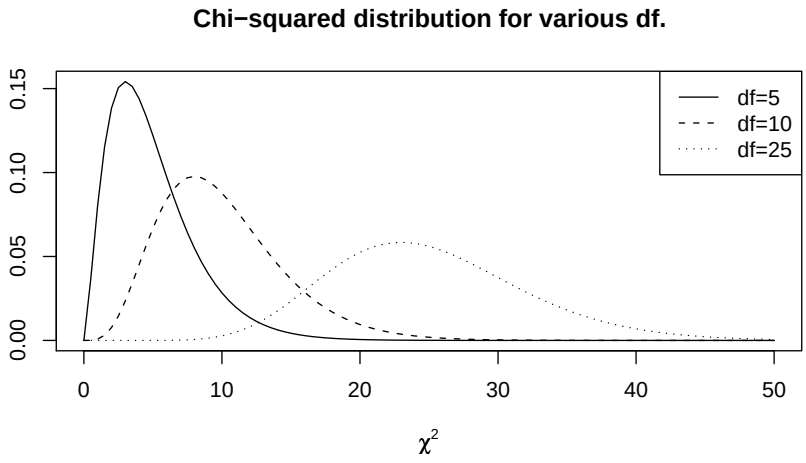
$$df = (\text{number rows - 1})(\text{number columns - 1}) \tag{2}$$

and O_i is the observed and E_i is the expected frequency for a given cell.

$$E_i = \frac{(\text{row total})(\text{column total})}{(\text{grand total})} = np_i \tag{3}$$

The p-value is always the area to the RIGHT of χ^2.

The χ^2 distribution



The χ^2 distribution

THE χ^2 CDF $F(\chi^{2\prime})$:
`pchisq(`$\chi^{2\prime}$`, df)`
Finds the are to the left of $\chi^{2\prime}$, $p = P(\chi^2 < \chi^{2\prime}) = F(\chi^{2\prime})$, with specified degrees of freedom.

R COMMAND

Example of finding expected values and the p-value.

		col 1	col 2
Observed Values:	row 1	5	10
	row 2	15	10

Question 7. Given the above table, find the expected value for row 1, column 1.

Observed Values:

	observed		expected	
	col 1	col 2	col 1	col 2
row 1	5	10	7.5	7.5
row 2	15	10	12.5	12.5

Question 8. Find χ^2 and the *df*. (Check: $\chi^2 = 2.67$)

Question 9. Find the p-value (Check 0.102)

R Command

TEST OF INDEPENDENCE AND HOMOGENEITY:
`chisq.test(D)`

D a contingency table.
R will alert you if you don't satisfy $E \geq 5$:
`"Warning message: Chi-squared approximation may be incorrect"`
See below for how to create `D` .

CONTINGENCY TABLES

Contingency tables in R

If you have a contingency table that you need to enter into R (book problems):

R Command

ENTERING A CONTINGENCY TABLE:
`D = data.frame(c1, c2, c3, ...)`

c1, c2, ... `c1` is a vector of column 1 data, `c2` is a vector of column 2 data, ...

Example 2. Given the contingency table:

	hair color		
Gender	Blond	Brown	Red
Female	1	1	2
Male	2	3	1

```
R: blond = c(1, 2)
R: brown = c(1, 3)
R: red = c(2, 1)
R: D = data.frame(blond, brown, red)
R: D
  blond brown red
1     1     1   2
2     2     3   1
```

Contingency tables in R

If you have a set of data that you need to make a contingency table from (real problems):

CREATING A CONTINGENCY TABLE:
`D = table(x1, x2)`
x1, x2 are **ordered** vectors recording two categorical variables.

R COMMAND

Example 3. Given the raw study data:

	subject									
	1	2	3	4	5	6	7	8	9	10
hair	BRN	BRN	BLOND	BRN	RED	RED	BLOND	BRN	RED	BLOND
gender	F	M	M	M	M	F	M	M	F	F

```
R: gender = c("F", "M", "M", "M", "M", "F", "M",
+     "M", "F", "F")
R: hair = c("BRN", "BRN", "BLOND", "BRN", "RED",
+     "RED", "BLOND", "BRN", "RED", "BLOND")
R: D = table(gender, hair)
R: D
      hair
gender BLOND BRN RED
     F     1   1   2
     M     2   3   1
```

1.4 A complete example

Back to our original problem

Example 4. A group of 276 healthy men and women were grouped according to their number of relationships. They were then exposed to a virus that caused colds. The data is summarized in the table below. Does the data provide sufficient evidence to indicate that susceptibility to colds is affected by the number of relationships you have?

	Number of relationships		
Contracted cold?	3 or less	4-5	6 or more
yes	49	43	34
no	31	47	62

Use our seven hypothesis testing steps.

Step 1: Test Use the test of independence. (Also possible to use test of homogeneity for this.)

Step 2: Requirements (1) Random samples, (2) $E \geq 5$ [delay until we use R].

Step 3: Hypothesis H_0 : your susceptibility to the cold and the number of relationships you have are independent. H_a : your susceptibility to the cold and the number of relationships you have are dependent.

Step 4: Significance $\alpha = 0.05$

Step 5: Find p-val Using R

Enter the data:

```
R: c1 = c(49, 31)
R: c2 = c(43, 47)
R: c3 = c(34, 62)
R: D = data.frame(c1, c2, c3)
R: D
  c1 c2 c3
1 49 43 34
2 31 47 62
```

Run the test

```
R: chisq.test(D)
        Pearson's Chi-squared test

data:  D
X-squared = 11.69, df = 2, p-value = 0.002894
```

Step 6: Decision Since p-val $\leq \alpha$, reject H_0.

Step 7: Conclusion The sample data supports the claim that the number of relationships you have and your susceptibility to the cold are dependent.

1.5 Summary

Test of Independence and Homogeneity

Requirements (1) Sample data are randomly selected, (2) For each cell in the contingency table the expected frequency $E \geq 5$. (No distribution requirement.)

Independence Hypothesis : one population

H_0 X and Y are independent.

H_a X and Y are dependent.

Homogeneity Hypothesis : more than one population

H_0 $p_1 = p_2 = \cdots = p_n$ the proportion of X is the same in all the populations studied.

H_a At least one proportion of X is not the same.

Test : `chisq.test(D)`

See page 6 for how to enter the contingency table `D` .

R will alert you if you don't satisfy $E \geq 5$.

1.6 Additional Methods

DEFINITION 1.4 YATE'S CONTINUITY CORRECTION.

Since we are using the continuous χ^2 distribution for discrete count data, we can apply Yate's continuity correction to the test statistic:

$$\chi^2 = \sum \frac{(|O_i - E_i| - 0.5)^2}{E_i} \tag{4}$$

R uses this correction.

Helpful hint
R uses Yate's continuity correction when appropriate.[2]

```
R: c1 = c(5, 15)
R: c2 = c(10, 10)
R: D = data.frame(c1, c2)
R: chisq.test(D)
        Pearson's Chi-squared test with Yates' continuity
        correction

data:  D
X-squared = 1.7067, df = 1, p-value = 0.1914
R: chisq.test(D, correct = F)
        Pearson's Chi-squared test

data:  D
X-squared = 2.6667, df = 1, p-value = 0.1025
```

Recall that the χ^2 test is only appropriate when all $E_i \geq 5$.

FISHER'S EXACT TEST. DEFINITION 1.5

Computes the **exact** p-values for **all** 2×2 contingency tables. Works the same as the χ^2 test for contingency table of form:

	class 1	class 2
sample 1	a	b
sample 2	c	d

The p-value for the test is calculated as:

$$p\text{-value} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!} \sum_i \frac{1}{a_i!b_i!c_i!d_i!} \tag{5}$$

Where $n! = a + b + c + d$.

The summation is over all possible 2×2 schemes with a cell frequency equal to or smaller than the smallest experimental frequency while keeping the row and column totals fixed as above. This gives us the tables that are "at least as unusual" as what we observed.

Example 5. Example of calculating exact fisher test value:

[2] It can have large effects on the p-value. If you want to check your answers against the book use the `correct=F` optional argument.

	class 1	class 2
sample 1	9	2
sample 2	7	6

So $a = 9, b = 2, c = 7, d = 6$:
Thus, the set of at least as unusual tables are:

9	2	10	1	11	0
7	6	6	7	5	8

$$p\text{-value} = \frac{(9+2)!(7+6)!(9+7)!(2+6)!}{24!} \times \left(\frac{1}{9!2!7!6!} + \frac{1}{10!1!6!7!} + \frac{1}{11!0!5!8!}\right) = 0.156$$

R Command

FISHER'S EXACT TEST:
`fisher.test(D)`
Conduct's the exact test for a 2×2 contingency table `D`.

Example 6. Doing the previous example:

	class 1	class 2
sample 1	9	2
sample 2	7	6

```
R: c1 = c(9, 7)
R: c2 = c(2, 6)
R: D = data.frame(c1, c2)
R: fisher.test(D, alternative = "greater")
        Fisher's Exact Test for Count Data

data:  D
p-value = 0.1557
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.59322     Inf
sample estimates:
odds ratio
    3.6427
```

We **must specify** `alternative="greater"` to get R to do it in the expected manner.[3]

[3] R can do much more advanced calculations where other alternatives would make sense. But recall, for the χ^2 we are always finding the area in the upper tail.

1.7 Additional Examples

Try the following problems in 11-3. Be sure determine which test applies, write the hypothesis, find the p-value, and write the formal conclusion.

Question 10. Do question 16 (Check: p-val: 1.19e-12)

Question 11. Do question 18 (Check: p-val: 0.00449)