Activity 10: Ref Chapter 10

Hypothesis Testing in Categorical Data

**10.1: Objectives**

* Test of proportions for a 2x2 tables resulting from 2 independent samples from binomial populations
* Association in RxC (page 413) contingency tables
* Kappa Statistic for reproducibility or agreement Page 431

**Section 10.2 2-Sample Proportion test**

In this section, we will test hypothesis regarding the two binomial population proportions resulting from a **2x2 contingency table for** two independent samples. We want to test the null hypothesis against any of the following alternative hypotheses:

The test statistic for this test under is given by

=

=

where . The computation of p-value for any Z-test is identical, and reject the null hypothesis at , if the -value .

It also follows that .

(1) So, in SAS, the test of is implemented via proc freq with riskdiff (equal) option. Without the equal option, it provides CI for proportion (risk) difference. With riskdiff optional **cl=wald** provides the Wald confidence interval for the difference.

(2) a proc freq with chisq option performs the test of association, which is equivalent to a Z test for proportion difference.

**Activity 10.1**

Two different tests (T1 and T2) are employed to diagnose a certain disease in 100 subjects (n1=n2=100). T1 detects 30 (x1) positive case, whereas T2 detects 45 positive cases (x2). Let be the population proportion of subjects detected as positive by two tests.

Test the null hypothesis against any of the following alternative hypotheses:

**Solution:**

**We could employ a z-test with self-coding. However, we will use proc freq chi-squared test (see code below):**

**data** chisqTest;

input Tests $ Disease $ Freq;

cards;

T1 positive 30

T1 negetive 70

T2 positive 45

T2 negetive 55

;

**run**;

\*ods trace on;

**proc** **freq** data=chisqTest order=data;

title "Test of H0: p1-p2=0 via chisq-test";

tables Tests\*Disease/norow nocol nopercent chisq riskdiff(equal)alpha=**0.05**;

weight Freq;

**run**;

\*ods trace off;

By running the above code creates, might get many outputs, not all of which you might want. Can control output via ods.

|  |
| --- |
| Test of H0: p1-p2=0 via chisq-test |

The FREQ Procedure

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | |  | | --- | | **Frequency** | | | | **Table of Tests by Disease** | | | | | --- | --- | --- | --- | | **Tests** | **Disease** | | | | **positive** | **negetive** | **Total** | | **T1** | |  | | --- | | 30 | | |  | | --- | | 70 | | |  | | --- | | 100 | | | **T2** | |  | | --- | | 45 | | |  | | --- | | 55 | | |  | | --- | | 100 | | | **Total** | |  | | --- | | 75 | | |  | | --- | | 125 | | |  | | --- | | 200 | | |

|  |
| --- |
| **Statistics for Table of Tests by Disease** |

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 1 | 4.8000 | 0.0285 |
| **Likelihood Ratio Chi-Square** | 1 | 4.8247 | 0.0281 |
| **Continuity Adj. Chi-Square** | 1 | 4.1813 | 0.0409 |
| **Mantel-Haenszel Chi-Square** | 1 | 4.7760 | 0.0289 |
| **Phi Coefficient** |  | -0.1549 |  |
| **Contingency Coefficient** |  | 0.1531 |  |
| **Cramer's V** |  | -0.1549 |  |

**Section 10.6 Tests for Association for contingency tables**

**An contingency table** is a summary of two categorical variables with R row levels and C column levels, respectively. An table takes the following form:

|  |  |  |
| --- | --- | --- |
| Rows | Columns  1 2 | Total |
| 1  2 |  |  |
| Total |  |  |

where

is the frequency in -th cell

=marginal total in ith row

=marginal total in jth column

We wish to test:

The **Pearson’s Chi-square statistic** for above test given by

follows a distribution under the , where is the expected frequency of given by

**Conclusion**

Reject null if the observed value of the test statistic is too large, i.e., the

p-value=Pr{>} at significant level .

**Activity 10.2 (Ref Example 10.38 Page 414)**

Test of association between women age at birth and development of breast cancer for data specified in the following 2X5 contingency table.

| **Table of cancer by ageatbirth** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **cancer** | **ageatbirth** | | | | | |
| **<20** | **20-24** | **25-29** | **30-34** | **>=35** | **Total** |
| **case** | |  | | --- | | 320 | | |  | | --- | | 1206 | | |  | | --- | | 1011 | | |  | | --- | | 463 | | |  | | --- | | 220 | | |  | | --- | | 3220 | |
| **control** | |  | | --- | | 1422 | | |  | | --- | | 4432 | | |  | | --- | | 2893 | | |  | | --- | | 1092 | | |  | | --- | | 406 | | |  | | --- | | 10245 | |
| **Total** | |  | | --- | | 1742 | | |  | | --- | | 5638 | | |  | | --- | | 3904 | | |  | | --- | | 1555 | | |  | | --- | | 626 | | |  | | --- | | 13465 | |

1. Read above data table in SAS.
2. Perform a test of hypothesis of no association between cancer and age at birth at 5% level of significance.
3. Report the value of the chi-squared test statistic and p-value for the test.
4. Make a conclusion about the test at 5% level of significance.

**Solution**

/\*SAS Code for Activity 10.2\*/

**data** cancer;

input cancer $ ageatbirth $ count;

cards;

case <20 320

case 20-24 1206

case 25-29 1011

case 30-34 463

case >=35 220

control <20 1422

control 20-24 4432

control 25-29 2893

control 30-34 1092

control >=35 406

;

**run**;

**proc** **freq** data=cancer order=data;

tables cancer\* ageatbirth /norow nocol nopercent expected chisq;

weight count;

title "Association of cancer with age at birth";

**run**;

|  |
| --- |
| Association of cancer with age at birth |

The FREQ Procedure

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | |  | | --- | | **Frequency** | | **Expected** | | | | **Table of cancer by ageatbirth** | | | | | | | | --- | --- | --- | --- | --- | --- | --- | | **cancer** | **ageatbirth** | | | | | | | **<20** | **20-24** | **25-29** | **30-34** | **>=35** | **Total** | | **case** | |  | | --- | | 320 | | 416.58 | | |  | | --- | | 1206 | | 1348.3 | | |  | | --- | | 1011 | | 933.6 | | |  | | --- | | 463 | | 371.86 | | |  | | --- | | 220 | | 149.7 | | |  | | --- | | 3220 | |  | | | **control** | |  | | --- | | 1422 | | 1325.4 | | |  | | --- | | 4432 | | 4289.7 | | |  | | --- | | 2893 | | 2970.4 | | |  | | --- | | 1092 | | 1183.1 | | |  | | --- | | 406 | | 476.3 | | |  | | --- | | 10245 | |  | | | **Total** | |  | | --- | | 1742 | | |  | | --- | | 5638 | | |  | | --- | | 3904 | | |  | | --- | | 1555 | | |  | | --- | | 626 | | |  | | --- | | 13465 | | |

|  |
| --- |
| **Statistics for Table of cancer by ageatbirth** |

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 4 | 130.3380 | <.0001 |
| **Likelihood Ratio Chi-Square** | 4 | 127.3850 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 129.0025 | <.0001 |
| **Phi Coefficient** |  | 0.0984 |  |
| **Contingency Coefficient** |  | 0.0979 |  |
| **Cramer's V** |  | 0.0984 |  |

|  |
| --- |
| **Sample Size = 13465** |

Note from the SAS output that and Pr{> can be computed in SAS:

**proc** **iml**;

pvalue=sdf("chisq",**130.338**, **4**);

print pvalue;

**quit**;

| **pvalue** |
| --- |
| 3.297E-27 |

So, pvalue<.

Based on p-value, we reject the null hypothesis and conclude that there is a significant relationship between age at birth and prevalence of cancer.

**Section 10.8 Kappa Statistic for reproducibility or agreement**

**Notations**

Given an contingency table from two surveys, for example, the following frequency distribution may be observed:

|  |  |  |
| --- | --- | --- |
| Survey 1 | Survey 2  1 2 | Total |
| 1  2 |  |  |
| Total |  |  |

of the contingency table

The Cohen measure of Kappa statistic is given by

where

An asymptotic estimate of variance of is given by

**Z-Confidence Interval of**

A CI of is .

**Z test statistic**

The test statistic to test vs is given by

under .

**Decision**

Reject at significant level in favor of a specified alternative if .

**Activity 10.3 (Example 10.3: Ref Example 10.50 Page 431)**

Below is a 2x2 contingency table reference to the reported beef consumption of 537 female American nurses at two different surveys

| **Survey 1** | **Survey 2** | | |
| --- | --- | --- | --- |
| **1 serving/week** | **1 serving/week** | **Total** |
| **1 serving/week** | |  | | --- | | 136 | | |  | | --- | | 92 | | |  | | --- | | 228 | |
| **>1 serving/week** | |  | | --- | | 69 | | |  | | --- | | 240 | | |  | | --- | | 309 | |
| **Total** | |  | | --- | | 205 | | |  | | --- | | 332 | | |  | | --- | | 537 | |

1. Read the data in SAS
2. Find a point estimate of Cohen’s Kappa agreement
3. Find an estimate of the SE of Kappa agreement
4. 95% CI of Cohen’s Kappa agreement
5. Test of hypothesis no Kappa agreement at 5% level of significance

/\*SAS code for activity 10.3\*/

**data** kappa;

length surv1 $**25** surv2 $**25**;

input surv1 $ surv2 $ count;

datalines;

below\_1\_serv/week below\_1\_serv/week 136

below\_1\_serv/week above\_1\_serv/week 92

above\_1\_serv/week below\_1\_serv/week 69

above\_1\_serv/week above\_1\_serv/week 240

;

**run**;

**proc** **freq** data=kappa order=data;

title 'Kappa agreement test';

tables surv1\*surv2/norow nocol nopercent agree alpha=**0.05**;

weight count;

test kappa;

**run**;

Note that the tables option “agree” provides kappa statistic. To get p-value of no kappa agreement test we have the “test kappa;” statement.

|  |
| --- |
| Kappa agreement test |

The FREQ Procedure

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | |  | | --- | | **Frequency** | | | | **Table of surv1 by surv2** | | | | | --- | --- | --- | --- | | **surv1** | **surv2** | | | | **below\_1\_serv/week** | **above\_1\_serv/week** | **Total** | | **below\_1\_serv/week** | |  | | --- | | 136 | | |  | | --- | | 92 | | |  | | --- | | 228 | | | **above\_1\_serv/week** | |  | | --- | | 69 | | |  | | --- | | 240 | | |  | | --- | | 309 | | | **Total** | |  | | --- | | 205 | | |  | | --- | | 332 | | |  | | --- | | 537 | | |

|  |
| --- |
| **Statistics for Table of surv1 by surv2** |

| **Simple Kappa Coefficient** | |
| --- | --- |
| **Kappa** | 0.3782 |
| **ASE** | 0.0404 |
| **95% Lower Conf Limit** | 0.2989 |
| **95% Upper Conf Limit** | 0.4575 |

| **Test of H0: Kappa = 0** | |
| --- | --- |
| **ASE under H0** | 0.0430 |
| **Z** | 8.7987 |
| **One-sided Pr > Z** | <.0001 |
| **Two-sided Pr > |Z|** | <.0001 |

|  |
| --- |
| **Sample Size = 537** |