University of New Mexico

Hypothesis Testing-4 (Fall 2017)

BIOM 505: Biostatistical Methods I (by Fares Qeadan)

---

hypothesis testing for association between two categorical variables (Chi-square Test of Independence):

The assumptions of this test are:

- The sampling method is a random sampling one.
- The variables under study are each categorical (either nominal or ordinal).
- The expected frequency for any cell should not be less than 1.
- No more than 20% of the cells can have expected values of less than 5.

The null and alternative hypotheses are:

H_0 : In the population, the two categorical variables are independent ($\chi^2 = 0$).
H_1 : In the population, two categorical variables are dependent ($\chi^2 > 0$).

Measures of association that are available in SAS:

- For two nominal variables or one ordinal, use Cramers V.
- For two ordinal variables, use gamma and Kendalls Tau B if table is square and Stuart's Tau-c if rectangular.
- For two categorical variables, use the χ^2 statistic
- For two categorical variables, use the odds ratio (OR) statistic. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

Strength of association between categorical variables (for Cramers V and Tau B and C:)

- Less than + or - 0.10: very weak
- + or -0.10 to 0.19: weak
- + or - 0.20 to 0.29: moderate
- + or - 0.30 or above: strong

Note:

1. Cramers V goes from 0 to 1 where 1 indicates strong association (for rXc tables). In 2x2 tables, the range is -1 to 1.

2. Gamma is generally larger than the Tau for the same relationship. This is due to the way Gamma is calculated. Gamma is generally not reported with a significance level. Gamma is a liberal estimate of the strength of the relationship and Tau is a conservative estimate.

(1) **Problem 1:** In a randomized experiment, one of the two categorical variables represents the treatments and the other represents the outcomes. For example, in the Physicians' Health Study subjects were treated with either aspirin or a placebo. The main recorded outcome was whether or not the subject suffered a heart attack. In this case our hypotheses can be stated as follows:

H_0 : There is no association between taking aspirin and having a heart attack.
H_1 : There is an association between taking aspirin and having a heart attack. (That is, those taking aspirin are either more likely or less likely to have a heart attack than those taking a placebo.)

Please test, at the 5% significance level, the presence of treatment effect based on the following two-way table which is representing the data from this famous study:

	Heart Attack	No Heart Attack	*Total*
Aspirin	104	10933	11937
Placebo	189	10845	11034
Total	293	21778	22071

(a) The significance level α is:

(b) The null and alternative hypotheses are:

(c) The decision rule (about H_0) is:

(d) Conduct the test in SAS:

```
data study;
   input treatment $ H_Attack $ Count;
   CARDS;
Aspirin Yes 104
Aspirin No 10933
Placebo Yes 189
Placebo No 10845
RUN;

proc freq data=study;
   tables treatment*H_Attack/chisq PLCORR out=freqcnt outexpect;
   weight Count;
run;
```

(e) Get the p-value from SAS's output:

(f) Decision:

(g) Conclusion:

(h) Please comment on the strength of association between taking aspirin and having a heart attack?

(i) Were the assumptions of this test satisfied? Explain.

(2) **problem 2:** At the 10% significance level, please examine the association between the recency of doctor visits and smoking status from the following 5x4 contingency table of counts for doctor visit recency by computed smoking status for the WASHDC data set? (this example is taken from The Oxford Handbook of Quantitative Methods, Vol. 2: Statistical Analysis: edited by Todd D. Little)

	daily current smoker	some days current smoker	former smoker	never smoked	*Total*
within past year	18	17	41	223	299
within past two years	5	5	3	44	57
within past 5 years	6	4	11	29	50
more than 5 years	2	3	6	9	20
Never	0	0	1	3	4
Total	31	29	62	308	430

(a) The significance level α is:

(b) The null and alternative hypotheses are:

(c) The decision rule (about H_0) is:

(d) Conduct the test in SAS:

```
data study2;
   input recency $ smoking $ Count;
   CARDS;
past1yr daily 18
past2yr daily 5
```

```
past5yr daily 6
more5yr daily 2
Never   daily 0
past1yr somedys 17
past2yr somedys 5
past5yr somedys 4
more5yr somedys 3
Never   somedys 0
past1yr former 41
past2yr former 3
past5yr former 11
more5yr former 6
Never   former 1
past1yr never  223
past2yr never  44
past5yr never  29
more5yr never 9
Never   never 3
run;

proc freq data=study2;
   tables recency*smoking/chisq PLCORR out=freqcnt2 outexpect;
   weight Count;
run;
```

(e) Get the p-value from SAS's output:

(f) Decision:

(g) Conclusion:

(h) Please comment on the strength of association between the recency of doctor visits and smoking status?

(i) Were the assumptions of this test satisfied? Explain

(j) How many cells and what percentage of cell has expected count less than 5?

(k) One possible solution, to having low expected counts, involves combining categories (adjacent). Could you solve the issue by combining the categories "more than 5 years" and "Never" as one category called "Never or more than 5 years" and the categories "daily current smoker" and "some days current smoker" as one category called "current smoker"?

	current smoker	former smoker	never smoked	*Total*
within past year	35	41	223	299
within past two years	10	3	44	57
within past 5 years	10	11	29	50
Never or more than 5 years	5	7	12	24
Total	60	62	308	430

Conduct the test in SAS:

```
data study3;
   input recency $ smoking $ Count;
   CARDS;
past1yr current 35
past2yr current 10
past5yr current 10
Nmore5yr current 5
past1yr former 41
past2yr former 3
past5yr former 11
Nmore5yr former 7
past1yr never 223
past2yr never 44
past5yr never 29
Nmore5yr never 12
run;

proc freq data=study3;
   tables recency*smoking/chisq PLCORR out=freqcnt3 outexpect;
   weight Count;
run;
```

(l) Are the assumptions of this test satisfied after the change you have made? Explain

(3) **problem 3:** Consider the Diabetes and obesity, cardiovascular risk factors data set we have used in Lab 1 (link: http://www.mathalpha.com/lab1/diabetesfall17.sas7bdat) to test whether diabetes status is associated with gender?

(a) The significance level α is:

(b) The null and alternative hypotheses are:

(c) The decision rule (about H_0) is:

(d) Conduct the test in SAS:

```
proc freq data=biom505.diabetesfall17 ;
table gender*diab/chisq PLCORR out=freqcnt4 outexpect;
run;
```

(e) Get the p-value from SAS's output:

(f) Decision:

(g) Conclusion:

(h) Please comment on the strength of association between diabetes and gender?

(i) Were the assumptions of this test satisfied? Explain.

(4) **problem 4:** Consider the Diabetes and obesity, cardiovascular risk factors data set we have used in Lab 1 (link: http://www.mathalpha.com/lab1/diabetesfall17.sas7bdat) to test whether body frame is associated with gender?

(a) The significance level α is:

(b) The null and alternative hypotheses are:

(c) The decision rule (about H_0) is:

(d) Conduct the test in SAS:

```
proc freq data=biom505.diabetesfall17 ;
table gender*frame/chisq PLCORR out=freqcnt5 outexpect;
run;
```

(e) Get the p-value from SAS's output:

(f) Decision:

(g) Conclusion:

(h) Please comment on the strength of association between body frame and gender?

(i) Were the assumptions of this test satisfied? Explain.