University of New Mexico

Hypothesis Testing-4 (Fall 2016)

PH 538: Public Health 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 STATA:

- 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 Kendalls Tau C if rectangular.

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 STATA:

```
tabi 104 10933 \ 189 10845, all expected
```

(e) Get the p-value from STATA'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 STATA:

```
tabi 18 17 41 223\5 5 3 44\6 4 11 29\2 3 6 9\0 0 1 3, all expected
```

(e) Get the p-value from STATA'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 STATA:

```
tabi 35 41 223\10 3 44\10 11 29\5 7 12 , all expected
```

(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/diabetesfall16.dta) 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 STATA:
```
tab gender diab,row all expected
```

(e) Get the p-value from STATA'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/diabetesfall16.dta) 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 STATA:
```
tab gender frame, row all expected
```

(e) Get the p-value from STATA'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.