

Chapter 13.

Association for Nominal

Pearson Independent Chi-square Test

2. A researcher hypothesizes that people who believe in "life after death" are, as a result, more likely to support the death penalty for serious criminal offenses. But, the researcher thinks that this might not be equally true for male people and female people. Data are collected from a nationally representative random sample (note, these results are fictitious), with the following results.

Among male persons, 450 respondents believed in life after death. Of these 400 favored the death penalty. 75 males did not believe in life after death. Among these people, 50 supported the death penalty.

Among female persons, 450 respondents believed in life after death. Of these, 300 favored the death penalty. 125 women reported that they did not believe in life after death. Among these women, 100 supported the death penalty.

Make Crosstabulations

Men	Believe	Don't Believe
Favor	400	50
Don't favor	50	25

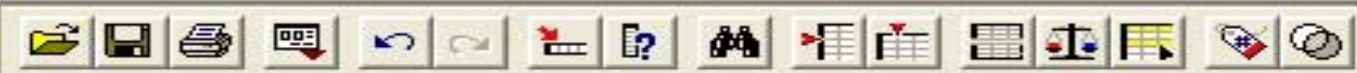
Women	Believe	Don't Believe
Favor	300	100
Don't favor	150	25

Input data in SPSS.

- Recall how to input data in SPSS.
- Set the rule of coding.
- Command SPSS to regard one variable about frequency as frequency. (Weighted cases)
- Chi-square test for each sex.
- Interpret.
- Subquestions A and B are about chi-square test.

Cramer's V and Phi

- Chi-square is sensitive to sample size. Its minimum value is zero, but there is no upper limit. Chi-square is not that good to show the strength of relationship.
- Phi is used for 2X2 table.
- Cramer's V is used for tables more than 2 categories. (Esp. when the number of rows is not the same with the number of columns)
- Unlike Chi-square, the value both measures can have ranges from 0 to 1.
- We can say about strong, moderate, weak association, but can't say that, e.g. the association of the second distribution is two times that of the first when one Phi score is 0.3 and another is 0.6.



18 : degree 1

1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19	1978	339	3	0	0				31

Crosstabs [X]

Row(s): [OK]

GSS YEAR FOR T [^]

Crosstabs: Statistics [X]

- Chi-square
- Correlations
- Nominal**
 - Contingency coefficient
 - Phi and Cramer's V
 - Lambda
 - Uncertainty coefficient
- Ordinal**
 - Gamma
 - Somers' d
 - Kendall's tau-b
 - Kendall's tau-c
- Nominal by Interval**
 - Eta
- Kappa
- Risk
- McNemar
- Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals: [1]

[Continue] [Cancel] [Help]

Proportional Reduction of Error (PRE)

- Recall that if two variables are not independent, you can predict dependent variable with independent variable with some amount of error.
- PRE is defined as follows: $(b-a)/b$
- B is the original number of errors (before using independent variable as a predictor).
- A is the new number of errors (after using independent variable).
- This concept is very important in social statistics although your homework does not ask about this.

The Relationship between PRE and Association

- If a is 0, you will get 1 pre. What did a mean? The new number of errors. If the new number of errors is 0, what does that mean? If you predict dependent variable with independent variable, there is no error any more, that is, perfect reduction of prediction error, which implies perfect association between two variables.
- If a and b are the same, you will get 0 pre. The original number of errors is the same with the new number of errors. What does this mean? Using independent variable as a predictor is not at all helpful to reduce the number of errors. In other words, the two variables are independent (not associated), so you can't predict the distribution of dependent variable even if you know information about the distribution of independent variable.

Let's think about B

Women	Believe	Don't	Total
Favor	300	100	400
Don't	150	<u>125</u>	<u>275</u>
total	450	225	675

- I replaced 25 by 125.
- If you don't know about independent variable, your best guess would be the most frequent category.
- So, if you want to minimize the error in your guess, it is better to choose favor(400) than don't favor(275).
- Therefore, 275 will be the original number of errors in your guess.

Let's think about a

not concentrated

Women	Believe	Don't	Total
Favor	300	100	400
Don't	150	125	275
total	450	225	675

- Now, you know about information of independent variable in addition to that about dependent variable as follows. Your best guess is again the most frequent category.
- To minimize the number of errors in your guess, it is better to choose favor(300) than don't favor(150) for 'believe', and don't favor(125) than favor(100) for 'don't believe.'
- The number of errors is 150 and 100 respectively. The total number of errors is 250.

Interpretation of PRE

- B was 275 and A was 250.
- Oh! Employing the information about independent variable is so useful to reduce the number of errors by 25.
- PRE is $(b-a)/b=(275-250)/275=0.09$. You can say the number of errors in your guess with the introduction of independent variable has been reduced by about 10%.
- In terms of association, you also can say that the association between two variables is not that strong because employing IV is a little bit helpful for your guess.

What are any other measures?: Lambda

- Lambda (Guttman coefficient) is none other than what you already calculated.
- Gamma and Kendall's tau b (for ordinal variables) and Pearson's correlation coefficient (for interval/ratio variables) are based on this principle of PRE.
- However, Lambda has critical limitations in some situations where all frequencies in independent variables are concentrated in only one category.

Let's think about b

Men	Believe	Don't	Total
Favor	400 (89%)	50 (67%)	450
Don't	50 (11%)	25 (33%)	75
total	450 (100%)	75 (100%)	525

- If you don't know about independent variable, what would be the original number of errors?
- Your best guess will be the most frequent category. So, if you chose favor(450), the number of errors in your guess is 75.

concentrated in
one category, FAVOR!

Let's think about A

Men	Believe	Don't	Total
→ Favor	400 (89%)	50 (67%)	450
Don't	<u>50</u> (11%)	<u>25</u> (33%)	75
total	450 (100%)	75 (100%)	525

- After employing independent variable, you got this table.
- Your best guess is again the most frequent category. So, for 'believe' you want to choose favor(400), and for 'don't believe' you might as well choose favor(50) to minimize the errors in your guess.
- The number of errors is 50 and 25 respectively. The total is 75.

- A and B are the same (75). Therefore, PRE is 0(%)
- This implies that employing independent variable is not at all useful to reduce the original number of errors. In other words, the result of your guess with the help of independent variable is actually the same as the result of your guess without predictors.
- This result “seems” to indicate that two variables are completely independent. However, as I said earlier, Lambda has its own limitation because the highest frequencies for IV is concentrated in only one category, “Believe.” Comparing percentages across IV (89% vs 67%) and (11% and 33%), two variables actually have association!

Social status is reproduced over the generation?

Parent \ Offspring	Lower	mid	Upper	Total
Lower	93	27	6	126
Mid	15	48	15	78
upper	0	15	81	96
total	108	90	102	300

- What is b? Your best guess before using IV is 126. B is $(78+96)=174$.
- What is a? Your best guess after introducing IV, 93, 48, 81, and the number of errors $(27+6)$, $(15+15)$, $(0+15)$. A is $33+30+15=78$.
- $\Lambda = (174-78)/174 = 0.55$

Odds Ratio and Relative Risk

What is Odds?

- Odds of an event is [the probability that the event might happen divided by the probability that it might not.]
- Let's think about the Titanic example again.

Observed	Female	Male	Total
Death →	123(a)	694(b)	817(a+b)
Survival →	324(c)	175(d)	499(c+d)
Total	447(a+c)	869(b+d)	1316(N)

Odds calculation

- (1) The probability of female's death out of the total death: $123/817(=a/(a+b))$
- (2) The probability of male's death out of the total death: $694/817(=b/(a+b))$
- 'Death' is an event here. The odds is the probability that female was dead over the probability that female was not. Therefore, the odds is $(1)/(2)=123/694=a/b$.
- In the same way, for 'survival' the odds is $324/175=c/d$.

What is Odds ratio?

- Odds ratio is literally the ratio of odds.
- As you know, OR is ad/bc .
- Recall you learned just before. This originates from $(a/b)/(c/d)=ad/bc$! (i.e., the odds for death divided by the odds for survival)
- This approach is good, but let's interpret the same example in a different way.

What is Odds?

Observed	Female ↓	Male ↓	Total
Death	123(a)	694(b)	817(a+b)
Survival	324(c)	175(d)	499(c+d)
Total	447(a+c)	869(b+d)	1316(N)

- (1) For 'female', what is the probability of death?
 $123/447=a/(a+c)$
- (2) For 'female', what is the probability of survival?
 $324/447=c/(a+c)$
- What is the odds then? The probability that female was dead over the probability that female was not.
 $(1)/(2)=123/324=a/c$

Odds ratio, again

- In the same way, for another group, 'male,' you can calculate the odds like this: b/d .
- What is the odds ratio, then? $(a/c)/(b/d)=ad/bc$. This result is the same with what you got first.
- Both of approaches are acceptable, but I like the latter because the latter is more useful for comparison of the probability of an event for two groups. (Intuitively, more understandable than when comparing something for two events!)

Calculate OR and natural log of OR

- Odds ratio is 0.096, which indicates the chance that female was dead is really smaller than the chance that male was dead in the Titanic disaster.
- Natural log of OR is $\log(0.096)=-2.343$

Margin of Error from SE of OR

- Standard error is expressed by square root of $[(1/a)+(1/b)+(1/c)+(1/d)]$.
- Hence, SE is the square root of $[(1/123)+(1/694)+(1/324)+(1/175)]=\sqrt{0.0184}=0.136$
- What is the margin of error, then? (z score) × (SE).
At the confidence level of 95%, z is 1.96. Therefore, it is $(1.96)(0.136)=0.266$

Don't forget to exponentiate confidence interval

- The confidence interval for "log OR" is as follows: Lower limit is log(OR)-margin of error. Therefore, $-2.343-0.266=-2.609$. Upper limit is log(OR)+margin of error. Therefore, $-2.343+0.266=-2.077$
- This is not your final answer. Exponentiate this to get the confidence level for "OR".
- The final answer: Lower limit is $e(-2.609)=0.0736$, while upper limit is $e(-2.077)=0.125$.

Why so complicated?

- Some of you might think, for example, why we should calculate 'natural log' and 'exponential'... I don't want you to memorize all of these things.
- This might be difficult, but let me explain as easily as possible.
- Can you remember how to get the confidence level of the population mean? You used this equation.

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- As you know, \bar{X} is sample mean, and σ/\sqrt{n} is standard error. Z score is the critical value, for example, 1.96 at the confidence level of 95%. The population mean is located within that range.
- Similarly, it could be said that the true value of $\log(\text{OR})$ is located between lower limit and higher limit. Just as you used sample mean as unbiased estimator, so you can use $\log(\text{OR})$ in the same way. Standard error is calculated in a different way: square root of $[(1/a)+(1/b)+(1/c)+(1/d)]$. But, the overall logic is the same: Unbiased estimator $\pm (z_{\alpha/2})(\text{SE})$

- In other words, it could be said that $\log(\text{OR})$ follows the distribution of z .
- What you have to keep in mind is that this confidence interval is about 'log(OR)' not 'Odds ratio'. (This is why you have to exponentiate the confidence interval for 'log(OR)' at the last moment of calculation process.)
- In a nutshell, the confidence interval can be calculated as follows more easily at the confidence level of 95% : $[\text{OR} \times e^{(-1.96 \times \text{SE})}, \text{OR} \times e^{(+1.96 \times \text{SE})}]$

Odds ratio vs Relative risk

- As you know, the equation for 'relative risk' $[a/(a+b)]/[c/(c+d)]$ is different from the equation for odds ratio.
- Both odds ratio and relative risk basically compare the likelihood of an event between two categories, but there are some other differences.