Chapter 13

Design and Analysis of techniques for epidemiologic studies

**Class Activity #11 @ 4/12/2022**

**Section 13.1 Objectives:**

* Designs for epidemiologic clinical trials such as **cohort**, **case-control** and **cross-sectional** studies (section 13.2)
* **Association** of risk factor with disease outcomes (section 13.2)
* Important **risk measures** such as **risk difference**, **relative risk** and **odds ratio** (section 13.3)
* Computation and SAS implementation

**Section 13.2: Study designs**

Any study in epidemiology or clinical trials can be of the following three forms:

* cohort,
* case-control and
* cross-sectional studies

**Notations for contingency table**

For a theoretical interest, a contingency table is generally presented via a table of observed frequencies or unknown probabilities for any study design.

**Table of observed frequencies** **Table of unknown probabilities**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Factor X | Outcome Y Label 1 Label 2 | Total | | Label 1  Label 2 |  |  | | Total |  |  | | |  |  |  | | --- | --- | --- | | Factor X | Outcome Y Label 1 Label 2 | Total | | Label 1  Label 2 |  |  | | Total |  |  | |

where

1. refers to the random frequency in th row and th column;
2. is the unknown probability that a subject belongs to cell ), for which the observed frequency is
3. follows a various **discrete probability distribution** under different sampling schemes. For example, follows binomial distribution for **case-control** and **cohort study**, whereas the vector follows a distribution for a cross-sectional study

**Section 13.2.1: Association in a cohort study**

**Two samples of subjects**, one from **exposed to a factor**, and the other **unexposed** to it, are followed up for a certain period of time to see how many develop an outcome. This study design is called a **prospective or** **cohort study (c-s)**.

**Example 13.1: Does treatment prevent heart attack?**

Patients with **Myocardial Infarction (MI)** are treated with Aspirin (E=1) and Placebo (E=2, i.e., standard treatment) to see if treatment prevents heart attack (outcome heart attack is observed after the follow-up). This design is an example of a **cohort study**. See the exposure and outcome summary below:

|  |  |  |
| --- | --- | --- |
| Treatment | Heart Attack  Yes No | Total |
| ***Aspirin (E=1)***  ***Placebo (E=2)*** | ***104 10933***  ***189 10845*** | ***11034***  ***11037*** |
| Total | 293 21778 | 22071 |

|  |  |  |
| --- | --- | --- |
| Exposure status | Disease Status  D=1 D=2 | Total |
| **Exposed (E=1)**  **Unexposed (E=2)** |  |  |
| Total |  |  |

Given a cohort study, we wish to test

Test statistic to test is the **Karl** **Pearson** Chi-square statistic given by:

where is expected frequency in the ()th cell.

**Other Test Statistics and Facts:**

1. **Mantel-Haenszel Chi-Square:** Mental Haenszel chi-square is given by

where and are expected mean and variance of .

1. **Likelihood Ratio Chi-square** : The likelihood ratio chi-square test statistics to test is given by
2. **Yates “Continuity Adjusted Chi-Square”** : Yates continuity adjusted chi-square , applicable to small samples, is given by

**Analysis of Example 13.1**

Use the contingency table below for any test of association via a chi-squared test via SAS implementation.

|  |  |  |
| --- | --- | --- |
| Treatment | Heart Attack  Yes No | Total |
| ***Aspirin (E=1)***  ***Placebo (E=2)*** | ***104 10933***  ***189 10845*** | ***11034***  ***11037*** |
| Total | 293 21778 | 22071 |

**data** HeartAttack;

length Treatment $**10** HeartAttack $**30**;

input Treatment $ HeartAttack $ count;

datalines;

Aspirin Attack 104

Aspirin No-Attack 10933

Placebo Attack 189

Placebo No-Attack 10845

;

**run**;

**proc** **freq** data=HeartAttack order=data;

title 'Association in a cohort study';

tables Treatment\*HeartAttack/norow nocol nopercent chisq expected alpha=**0.05**;

weight count;

**run**;

|  |
| --- |
| Association in a cohort study |

The FREQ Procedure

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | |  | | --- | | **Frequency** | | | | **Table of Treatment by HeartAttack** | | | | | --- | --- | --- | --- | | **Treatment** | **HeartAttack** | | | | **Attack** | **No-Attack** | **Total** | | **Aspirin** | |  | | --- | | 104 | | |  | | --- | | 10933 | | |  | | --- | | 11037 | | | **Placebo** | |  | | --- | | 189 | | |  | | --- | | 10845 | | |  | | --- | | 11034 | | | **Total** | |  | | --- | | 293 | | |  | | --- | | 21778 | | |  | | --- | | 22071 | | |

|  |
| --- |
| **Statistics for Table of Treatment by HeartAttack** |

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 1 | 25.0139 | <.0001 |
| **Likelihood Ratio Chi-Square** | 1 | 25.3720 | <.0001 |
| **Continuity Adj. Chi-Square** | 1 | 24.4291 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 25.0128 | <.0001 |

**Conclusion**: Since p-value is less than 0.05, at 5% level of significance we reject the null hypothesis of no association between treatment and heart attack.

**13.2.2 Association for a case-control study**

Suppose that two samples of subjects, one from **case (diseased**) and the other from **control (without disease)**, are asked about their history of an exposure, say smoking, to determine the association. This study design is called the **retrospective** or **case-control study**.

**Example 13.2: Does smoking cause lung cancer?**

A sample of 709 subjects with **lung cancer (case)** and a sample 709 subjects **without lung cancer (control)**, are asked about their past smoking history to determine if smoking is associated with lung cancer. This design is an example of a **case-control study**. See the exposure and outcome summary below:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Smoking | Lung cancer  Yes (D=1) No (D=2) | Total | | Yes  No | 688 650  21 59 | 1338  80 | | Total | 709 709 | 1418 | | |  |  |  | | --- | --- | --- | | Exposure status | **Disease Status**  **D=1 D=2** | Total | | Exp E=1  Unex E=2 |  |  | | Total |  |  | |

What do we want to test in a case-control study?

In a case-control study, we wish to test

the rate of exposure in case and control is the same

Test statistic to test is the Pearson Chi-square statistic , defined before.

**Analysis of Example 13.2: Application of a case-control study**

Use the contingency table below under a case-control study design and test for any association via SAS implementation.

|  |  |  |
| --- | --- | --- |
| Smoking | Lung cancer  Yes (D=1) No (D=2) | Total |
| Yes  No | 688 650  21 59 | 1338  80 |
| Total | 709 709 | 1418 |

**data** LungCancer;

length Smoker $**10** LungCancer $**12**;

input Smoker $ LungCancer $ count;

datalines;

Yes Case 688

Yes Control 650

No Case 21

No Control 59

;

**run**;

**proc** **freq** data=LungCancer order=data;

title " Association in a case-control study design";

tables Smoker\*LungCancer/norow nocol nopercent chisq alpha=**0.05**;;

weight count;

**run**;

|  |
| --- |
| Association in a case-control study design |

The FREQ Procedure

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | |  | | --- | | **Frequency** | | | | **Table of Smoker by LungCancer** | | | | | --- | --- | --- | --- | | **Smoker** | **LungCancer** | | | | **Case** | **Control** | **Total** | | **Yes** | |  | | --- | | 688 | | |  | | --- | | 650 | | |  | | --- | | 1338 | | | **No** | |  | | --- | | 21 | | |  | | --- | | 59 | | |  | | --- | | 80 | | | **Total** | |  | | --- | | 709 | | |  | | --- | | 709 | | |  | | --- | | 1418 | | |

|  |
| --- |
| **Statistics for Table of Smoker by LungCancer** |

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 1 | 19.1292 | <.0001 |
| **Likelihood Ratio Chi-Square** | 1 | 19.8780 | <.0001 |
| **Continuity Adj. Chi-Square** | 1 | 18.1357 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 19.1157 | <.0001 |

**Conclusion**: We reject the null hypothesis of no association between smoking and lung cancer.

**13.2.3 Association in a cross-sectional study**

**A sample of subjects** **is cross-classified** by the presence of an ***exposure*** and an ***outcome status*** to determine if there is any association of exposure to the outcome. This studydesignis called the **cross-sectional study.**

**Example 13.3: Does the use of seat-belt prevent fatal injury?**

A sample of subjects reported involved in accident is **cross-classified** by Seatbelt use and Injury status:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Seat-Belt Use | Injury  Fatal Nonfatal | Total | | No (E=1)  Yes (E=2) | 1085 55,623  703 441,239 | 56,708  441,942 | | Total | 1788 496862 | 498,650 | | |  |  |  | | --- | --- | --- | | Exposure status | Disease Status | Total | | Exposed  Unexposed 2 |  |  | | Total |  |  | |

**We wish to test**  independence of exposure and disease status, which is carried out by implementing the Pearson Chi-square statistic , defined before.

**Remarks: The chi-squared test is invariant with respect to the study design.**

**Example 13.3**

Use the contingency table below under a cross-sectional study design, test for no association between seat-belt usage and injury via SAS implementation.

|  |  |  |
| --- | --- | --- |
| Seat-Belt Use | Injury  Fatal Nonfatal | Total |
| No (E=1)  Yes (E=2) | 1085 55,623  703 441,239 | 56,708  441,942 |
| Total | 1788 496862 | 498,650 |

**data** Injury;

length Seatbelt $**10** Injury $**30**;

input Seatbelt $ Injury $ count;

datalines;

No Fatal 1085

No Nonfatal 55623

Yes Fatal 703

Yes Nonfatal 441239

;

**run**;

**proc** **freq** data=Injury order=data;

title 'Association in a cross-sectional study';

tables Seatbelt\*Injury/norow nocol nopercent chisq alpha=**0.05**;

weight count;

**run**;

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | | --- | | Association in a cross-sectional study |   The FREQ Procedure   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | |  | | --- | | **Frequency** | | | | **Table of Seatbelt by Injury** | | | | | --- | --- | --- | --- | | **Seatbelt** | **Injury** | | | | **Fatal** | **Nonfatal** | **Total** | | **No** | |  | | --- | | 1085 | | |  | | --- | | 55623 | | |  | | --- | | 56708 | | | **Yes** | |  | | --- | | 703 | | |  | | --- | | 441239 | | |  | | --- | | 441942 | | | **Total** | |  | | --- | | 1788 | | |  | | --- | | 496862 | | |  | | --- | | 498650 | | | |

|  |
| --- |
| **Statistics for Table of Seatbelt by Injury** |

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 1 | 4328.9241 | <.0001 |
| **Likelihood Ratio Chi-Square** | 1 | 2506.4412 | <.0001 |
| **Continuity Adj. Chi-Square** | 1 | 4324.0156 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 4328.9155 | <.0001 |

**Conclusion**: We reject the null hypothesis of independence of seat-belt use and injury.

**13.3 Measures of Risk**

Given there is any association between risk factor and disease outcome, we wish to measure the extent of risk in exposed group. This can be done by three important measures:

* Risk difference
* Relative risk
* Odds ratio

**13.3.1 Risk difference**

Risk difference of a factor between exposed and unexposed status is defined by

**Properties of :**

1. RD is not estimable for a case-control study (***why?***).
3. no association of risk factor and outcome.
4. being exposed to the factor is protective.
5. being exposed to the factor is harmful.

Consider a cohort study

|  |  |  |
| --- | --- | --- |
| Exposure status | Disease Status  D=1 D=2 | Total |
| **Exposed E=1**  **Unexposed E=2** |  |  |
| Total |  |  |

For a **cohort study**, since(***why?*** what are the probability distribution ofand ***?***)**,** we have

**Point estimate of :** A point estimate of for a cohort study is given by

**Confidence Interval Estimate of :**

A CI of is given by

where and is an estimate of the variance of . Recall that, for example, for a 95% CI, .

**Theorem 1: An MLE (consistent estimator) of for a cohort study is**

**Applications**

***For simplicities of presentation, we assume that all 2 contingency tables presented are constructed by all three designs, i.e., cohort, case-control and cross-sectional study designs.***

**Example 13.4** Consider **s**moking vs outcome of lung cancer in a contingency table below

|  |  |  |
| --- | --- | --- |
| Smoking | Lung cancer  Yes (D=1) No (D=2) | Total |
| Yes  No | 688 650  21 59 | 1338  80 |
| Total | 709 709 | 1418 |

**95% CI:** A 95% CI of is

**Conclusion:** 95% CI of **does not contain 0** and the interval is **positive**. Therefore, at 5% level of significance, it appears that ***smoking is positively associated with lung cancer***.

/\*RD for lung cancer due to smoking\*/

**data** LungCancer;

length Smoker $**10** LungCancer $**12**;

input Smoker $ LungCancer $ count;

datalines;

Yes Case 688

Yes Control 650

No Case 21

No Control 59

;

**run**;

\*ods trace on;

ods select RiskDiffCol1 RiskDiffCol2;

**proc** **freq** data=LungCancer order=data;

title 'Risk Difference Measures';

tables Smoker\*LungCancer/riskdiff alpha=**0.05**;

weight count;

**run**;

\*ods trace off;

|  |
| --- |
| Risk Difference Measures |

The FREQ Procedure

Statistics for Table of Smoker by LungCancer

| **Column 1 Risk Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Risk** | **ASE** | **(Asymptotic) 95% Confidence Limits** | | **(Exact) 95% Confidence Limits** | |
| **Row 1** | 0.5142 | 0.0137 | 0.4874 | 0.5410 | 0.4870 | 0.5413 |
| **Row 2** | 0.2625 | 0.0492 | 0.1661 | 0.3589 | 0.1704 | 0.3729 |
| **Total** | 0.5000 | 0.0133 | 0.4740 | 0.5260 | 0.4736 | 0.5264 |
| **Difference** | **0.2517** | **0.0511** | **0.1516** | **0.3518** |  |  |
| **Difference is (Row 1 - Row 2)** | | | | | | |

| **Column 2 Risk Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Risk** | **ASE** | **(Asymptotic) 95% Confidence Limits** | | **(Exact) 95% Confidence Limits** | |
| **Row 1** | 0.4858 | 0.0137 | 0.4590 | 0.5126 | 0.4587 | 0.5130 |
| **Row 2** | 0.7375 | 0.0492 | 0.6411 | 0.8339 | 0.6271 | 0.8296 |
| **Total** | 0.5000 | 0.0133 | 0.4740 | 0.5260 | 0.4736 | 0.5264 |
| **Difference** | -0.2517 | 0.0511 | -0.3518 | -0.1516 |  |  |
| **Difference is (Row 1 - Row 2)** | | | | | | |

**13.3.2 Relative Risk**

Relative risk is defined as the ratio of risk of having disease in exposed to risk of having disease in unexposed:

**Properties:**

1. It is not estimable for case-control (Why?).
3. there is no association between risk factor and disease outcome.
4. being exposed to the factor is protective.
5. being exposed to the factor is harmful.

For a **cohort study**, since(***why?***)**,** we have .

**A point estimate of** For a **cohort study**, a point estimate of is given by

**Confidence Interval Estimate of**

A CI of is **,** where is an estimate of the variance of .

**Theorem 2: Let . Then, an estimate of the variance of for a cohort study using the delta method is**

|  |  |  |
| --- | --- | --- |
| Smoking | Lung cancer  Yes (D=1) No (D=2) | Total |
| Yes  No | 688 650  21 59 | 1338  80 |
| Total | 709 709 | 1418 |

**95% CI:** A CI of is

**Conclusion:** 95% CI of **does not contain 1** and the **lower limit of the interval is above 1**. Therefore, at 5% level of significance, it appears that smoking is positively associated with the development of the lung cancer.

\*ods trace on;

ods select RelativeRisks;

**proc** **freq** data=LungCancer order=data;

title 'Relative Risk & OR Measures';

tables Smoker\*LungCancer/relrisk alpha=**0.05**;

weight count;

**run**;

\*ods trace off;

|  |
| --- |
| Relative Risk & OR Measures |

The FREQ Procedure

Statistics for Table of Smoker by LungCancer

| **Odds Ratio and Relative Risks** | | | |
| --- | --- | --- | --- |
| **Statistic** | **Value** | **95% Confidence Limits** | |
| **Odds Ratio** | **2.9738** | 1.7867 | 4.9494 |
| **Relative Risk (Column 1)** | **1.9589** | **1.3517** | **2.8387** |
| **Relative Risk (Column 2)** | **0.6587** | 0.5716 | 0.7591 |

**13.3.3 Odds ratio**

OR is defined as the ratio of odds of having disease in exposed to the odds of having disease in unexposed:

It turns out that

**Properties of OR**

1. OR is estimable for all study designs.
3. means that there is no association between risk factor and disease outcome.
4. means exposure is protective.
5. means exposure is harmful.

:

:

A CI of is

were is the upper -*th* percentile of standard normal distribution and is an estimate of the variance of .

We can find in order to find CI estimate of or by the following theorem:

**Theorem 3: Let . Then, an estimate of the variance of for a cohort study using the delta method is**

|  |  |  |
| --- | --- | --- |
| Smoking | Lung cancer  Yes (D=1) No (D=2) | Total |
| Yes  No | 688 650  21 59 | 1338  80 |
| Total | 709 709 | 1418 |

*Smokers have about 3 times higher odds of having lung cancer than nonsmokers.*

Then, a CI of is

**Conclusion:** 95% CI of **does not contain 1** and the **lower limit of the interval is above 1**. Therefore, at 5% level of significance, it appears that ***smoking is positively associated with lung cancer***.

\*ods trace on;

ods select RelativeRisks;

**proc** **freq** data=LungCancer order=data;

title 'Relative Risk & OR Measures';

tables Smoker\*LungCancer/relrisk alpha=**0.05**;

weight count;

**run**;

\*ods trace off;

|  |
| --- |
| Relative Risk & OR Measures |

The FREQ Procedure

Statistics for Table of Smoker by LungCancer

| **Odds Ratio and Relative Risks** | | | |
| --- | --- | --- | --- |
| **Statistic** | **Value** | **95% Confidence Limits** | |
|  |  |  | |
| **Odds Ratio** | **2.9738** | 1.7867 | 4.9494 |
| **Relative Risk (Column 1)** | **1.9589** | 1.3517 | 2.8387 |
| **Relative Risk (Column 2)** | **0.6587** | 0.5716 | 0.7591 |