Person-Time Data

CF Jeff Lin, MD., PhD.

December 14, 2005

Incidence

1. Cumulative incidence (incidence proportion)
2. Incidence density (incidence rate)

Cumulative Incidence

Cumulative Incidence is the proportion of the population will develop illness during the specified time period.

$$\text{Cumulative Incidence (C.I.)} = \frac{\text{number of NEW cases of disease during a period}}{\text{population exposed during this period}} \quad (1)$$

Cumulative Incidence: Example

Lung cancer in a community, Jan 1 – Dec 31, 1980:

Population	3,500,000	
Cases	96,250	(1250 new cases)
Cumulative incidence	0.36/1000	per year
Prevalence	2.71%	

Person-Time Data

1. In a cohort study, we identify groups of exposed and unexposed individuals at baseline, and compared the proportion of subjects who developed disease over time between two groups.
2. We referred to these proportions as **cumulative incidence (CI) rates** (i.e., the probability that a person no prior disease will develop a new case of the disease over some pre-specified time period).

Person-Time Data

3. Cumulative incidence (CI) rates are **proportions** where the person is the unit of analysis and must range between 0 and 1.
4. When we discuss the analysis of categorical data, where the person was the unit of analysis.
5. In an actual prospective study design, each subject contributes the study with different follow-up time (i.e., **person-time**).

Incidence Rate

Example: Prevalence and Incidence Rate

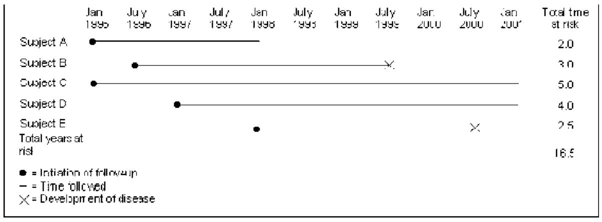

Figure 1: Incidence Rate and Follow-Up with Pearson-Time

Incidence Rate: Pearson-Time
Pearson-Years

1. Person-time is the sum of the amount of time each individual is observed while free of disease.
2. pearson-years is the sum of the amount of **years** each individual is observed while free of disease.
3. Each subject may contribute a different amount of person-years.

Incidence Rate: Pearson-Years

Person-time at risk is the denominator for incidence rates of disease

$$\begin{aligned} & \text{1000 person-years at risk} \\ = & \text{ 100,000 people for } 1/100 \text{ years} \qquad (2) \\ = & \text{ 10,000 people for } 1/10 \text{ years} \\ = & \text{ 1000 people for 1 years} \\ = & \text{ 100 people for 10 years} \\ = & \text{ 20 people for 50 years} \end{aligned}$$

Incidence Rate: Pack-Years for Smoking

$$\begin{aligned} & 1 \times 365 \text{ pack-year} \\ = & \; 0.5 \times 365 \text{ for 2 years} \\ = & \; 2 \times 365 \text{ for 0.5 years} \end{aligned}$$

Incidence Rate

An **incidence rate** (**incidence density**) is defined as the number of new cases of disease during a defined period of time, divided by the **total person-time** of observation.

$$\begin{aligned} & \text{Incidence Rate (I.R.)} \\ = & \frac{\text{number of NEW cases of disease during a period}}{\text{total person-time of observation}} \qquad (3) \end{aligned}$$

Example: Period Prevalence and Incidence Rate

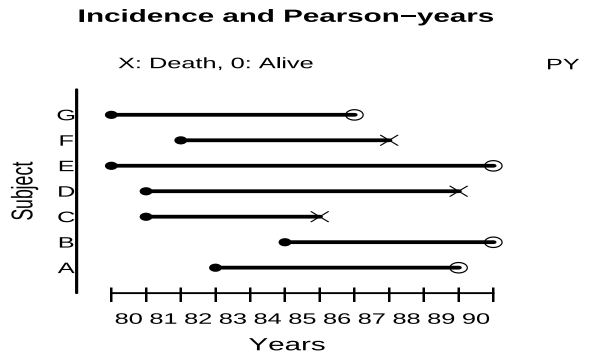

Figure 2: Incidence and Prevalence

Cumulative Incidence Rates and Incidence Rate

1. In computing cumulative incidence rates, we implicitly assume that all subjects are followed for the same period of time T.

2. This is not always the case.

3. So, the first issue to consider in a cohort study is the appropriate unit of analysis for each group.

4. If subject is used as the unit of analysis, then the problem is that different subject contribute different amounts of person-time to the analysis, and the assumption of a constant probability of an event for each subject would then be violated.

5. If a person-year is used as the unit of analysis (i.e., one person followed for one year), then since each subject can contribute more than one-person-year to the analysis, the important assumption of independence for the binomial distribution would be violated.

Breast Cancer and Oral Contrapcetive Use

1. A hypothesis of much recent interest is the possible association between the use of oral contraceptives (OC) and the development of breast cancer. To address this issue, data were collected in the Nurses' Healthy Study where disease-free women were classified in 1976 according to OC status (Current user/pat user/never user).

2. A mail questionnaire was sent out every two years in which OC status was updated and breast cancer status was ascertained over the next two years.

3. For each woman, an among of time that the woman is a current user or a never user of OC's (ignoring past use) can be calculated and this person-time can be accumulated over the entire cohort of nurses.

4. Thus, each nurse contributes a different among of person-time to the analysis.

Breast Cancer and Oral Contrapcetive Use

5. The data are presented in Table 1 for current and never users of OC's between women 40-44 years of age.
6. How should these data be used to assess any differences in the incidence rate of breast cancer by OC-use group?

Breast Cancer and Oral Contrapcetive Use

Table 1: Relationship between breast-cancer incidence and OC use between 40-44 year-old women in the Nurse's Health Study

OC-use Group	Number of cases Cases	Number of Pearson-years
Current users	13	4,761
Past Users	164	121,091
Never Users	113	98,091

Person-Time Data: Rare Event Rate

1. What is the distribution of the number of event from time 0 to T (where T is some long period of time, 1 year or 20 years) ?
2. Three assumption must be made about the incidence. Consider an general small subinterval of the time period T, denoted by ΔT.

Person-Time Data: Rare Event Rate Assumptions

1. **Rare Event Occurring Probability, Rare Event Rate**:

(a) The probability of one event occurring in a very short time period is very small.

(b) The probability of observation 1 event is directly proportional to the length of the time interval ΔT.

$$P(1 \text{ event}) = \lambda \Delta T \quad (4)$$

for some constant λ.

(c) The probability of observing 0 event over ΔT is approximately $1 - \lambda \Delta T$.

(d) The probability of observing more than 1 event over this time interval is essentially 0.

Person-Time Data: Rare Event Rate Assumptions

2. **Stationary**:

(a) Assume that the number of events per unit time is the same throughout the entire time interval T.

(b) Thus, and increase in the incidence of the event as time goes one within the time period T would violate this assumption.

(c) Note that T should not be overly long, because this assumption is less likely to hold as T increases.

(d) **Independence**: In a event occurs within time subinterval, it has no bearing on the probability of event in the next time subinterval.

(e) This assumption would be violated in some situations, (i.e., an epidemic situation or number of insurance claims in a period), because

a new event occurs, then subsequent event are likely to build up over a short period of time until after the epidemic subsides.

(f) However, in clinical situations, these assumptions are not usually valid

Person-Time Data: Rare Event Rate

(a) Given the assumptions, the Poisson probability corresponding can be derived.

(b) The probability of k events occurring in a time period T for a Poisson random variable with parameter λ is

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!}, \quad k = 1, 2, \ldots \qquad (5)$$

where $\mu = \lambda T$ and e is approximately 2.171828.

(c) In many instances we can not predict whether the assumptions for the Poisson distribution are satisfied.

(d) Fortunately, the relationship between the expected value and variance of the Poisson distribution provides an important guideline that helps identify random variables that follow this distribution.

(e) For a Poisson corresponding with parameter μ, the mean and variance

are both equal to μ.

(f) This fact is useful, because if we have a data set from a discrete corresponding where the sample mean and sample variance are about the same, then we can preliminarily identify it as a Poisson corresponding and use various tests to confirm this hypothesis.

(g) Note: Calculating Poisson Probabilities can be easily achieved by current computing environment.

Point Estimation for the Poisson Distribution

1. Suppose we assume that the number of events X over T person-years is Poisson distributed with parameter $\mu = \lambda T$.
2. An unbiased estimator of λ is given by $\hat{\lambda} = X/T$, where X is the observed number of events over T person-years.
3. If λ is the **incidence rate** per person-year, T is the number of person-years of follow-up, and we assume Poisson corresponding for the number of events X over T person-years, then the expected value of X is given by $\mathcal{E}(X) = \lambda T$.
4. Therefore,

$$\mathcal{E}(\hat{\lambda}) = \mathcal{E}(X)/T = \lambda T/T = \lambda \qquad (6)$$

Thus, $\hat{\lambda}$ is the unbiased estimator of λ.

Confidence Interval for the Poisson Distribution

1. Suppose we assume that the number of events X over T person-years is Poisson distributed with parameter $\mu = \lambda T$.
2. An unbiased estimator of λ is given by $\hat{\lambda} = X/T$, where X is the observed number of events over T person-years.
3. If λ is the **incidence rate** per person-year, T is the number of person-years of follow-up, and we assume Poisson corresponding for the number of events X over T person-years, then the expected value of X is given by $\mathcal{E}(X) = \lambda T$.

Confidence Interval for the Poisson Distribution

4. Therefore,

$$\mathcal{E}(\hat{\lambda}) = \mathcal{E}(X)/T = \lambda T/T = \lambda \qquad (7)$$

5. Thus, $\hat{\lambda}$ is the unbiased estimator of λ.

Confidence Interval for the Poisson Distribution

6. The question remains as to how to obtain an interval estimate for λ.

7. We use a similar approach as was used to obtain exact confidence limits for the binomial proportion p.

8. For this purpose, it will be easier to fist obtain a confidence interval for μ, the expected number of events over time T of the form (μ_1, μ_2) and then obtain the corresponding confidence variance for λ form $(\mu_1/T, \mu_2/T)$.

Confidence Interval for the Poisson Distribution

9. An exact $(1-\alpha)\times 100\%$ confidence interval for the Poisson parameter λ is given $(\mu_1/T, \mu_2/T)$, where μ_1 and μ_2 satisfy the equations

$$P(X \geq x|\mu=\mu_1) = \frac{\alpha}{2} = \sum_{k=x}^{\infty} \frac{e^{-\mu_1}\mu_1^k}{k!} = 1 - \sum_{k=0}^{x-1} \frac{e^{-\mu_1}\mu_1^k}{k!} \quad (8)$$

$$P(X \leq x|\mu=\mu_2) = \frac{\alpha}{2} = \sum_{k=0}^{x} \frac{e^{-\mu_1}\mu_1^k}{k!} \quad (9)$$

and x is the observed number of events, T is the number of person-years of follow-up.

Poisson Approximate to the Binomial Distribution

1. The Poisson distribution appears to fit well in some applications.

2. Another important use for the Poisson distribution is as an approximation to the binomial distribution. Consider the binomial distribution for large n and small π.

3. The mean of this distribution is given by $n\pi$ and the variance by $n\pi(1-\pi)$. note that $1-\pi$ is approximate equal to 1 for small π, thus, $n\pi(1-\pi) \approx n\pi$.

4. Therefore, the mean and variance of the binomial distribution are almost equal in this case.

5. So the binomial corresponding with large n and small π can be accurately approximated by a Poisson distribution with parameter $\mu = n\pi$.

Poisson Approximate to the Binomial Distribution

6. The rationale for using this approximation is that the Poisson corresponding is easier to work with than the binomial distribution.

7. The binomial distribution involve expression such as $\binom{n}{k}$, π^k and $(1-\pi)^{n-k}$, which are cumbersome for large n.

8. How large should n be or how small should p be before approximation is adequate?

9. A conservative rule is to use the approximation with $n \geq 100$ and $\pi \leq 0.01$.

Inference for One-Sample Poisson Distribution

1. Exact Method

2. Approximate Method

Inference for One-Sample Poisson Distribution Exact Method

Inference for One-Sample Poisson Distribution Exact Method

1. Let

$$X = \text{total observed number of events for members of the study popul}$$
$$p_i = \text{probability of event for the } i\text{th individual}$$

2. The most common event in medical studies is death for a particular disease.

Inference for One-Sample Poisson Distribution Exact Method

3. Under the null hypothesis that the event rates for the study population are the same as those for the known population, the expected number of events μ_0 is given by

$$\mu_0 = \sum_{i=1}^{n} p_i \quad (10)$$

Inference for One-Sample Poisson Distribution Exact Method

4. If the disease under study is rare, then the observed number of events may be considered approximately Poisson distributed with unknown expected value μ.

5. Let X be a Poisson random variable with expected value μ.

6. We wish to test the hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad \mu \neq \mu_0 \quad (11)$$

Inference for One-Sample Poisson Distribution Exact Method

7. Using a two-sided test with significance level α, the procedures can be followed as:

(a) We first compute

$$X = \text{observed number of events in the study population} \quad (12)$$

(b) Under H_0, the random variable X will follow a Poisson corresponding with parameter μ_0.

(c) Obtain the two-sided $(1-\alpha)\times 100\%$ confidence interval for μ based the observed x of X.

Inference for One-Sample Poisson Distribution Exact Method

(d) Denote this confidence interval (μ_1, μ_2), we

$$\text{reject } H_0, \quad \text{if } \mu_0 < \mu_1 \text{ or } \mu_0 > \mu_2; \quad (1$$
$$\text{accept } H_0, \quad \text{if } \mu_2 \leq \mu_0 \leq \mu_2. \quad (1$$

Inference for One-Sample Poisson Distribution Exact Method

(e) Thus, the exact two-sided p-value is given by

$$\min\left[2 \times \sum_{k=0}^{x} \frac{e^{-\mu_0}\mu_0^k}{k!}, 1\right], \quad \text{if } x \leq \mu_0 \qquad (1$$

$$\min\left[2 \times \left(1 - \sum_{k=0}^{x-1} \frac{e^{-\mu_0}\mu_0^k}{k!}\right), 1\right], \quad \text{if } x \leq \mu_0. \qquad (1$$

where x is the observed event for a particular data.

Approximate Method

1. If the expected number of events is large, then the following approximate method can be used.
2. Let μ be expected value of a Poisson random variable.
3. To test the hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad \mu \neq \mu_0, \qquad (17)$$

Approximate Method

1. We compute

$$X = \text{observed number of deaths in the study population} \qquad (18)$$

2. Compute the test statistic

$$X^2 = \frac{(X - \mu_0)^2}{\mu_0} \sim \chi_1^2, \quad \text{under } H_0 \qquad (19)$$

Approximate Method

3. For a two-sided test at level α, we

$$\text{reject } H_0, \quad \text{if } X^2 > \chi_{1,1-\alpha}^2; \qquad (20$$

$$\text{accept } H_0, \quad \text{if } X^2 \leq \chi_{1,1-\alpha}^2. \qquad (21$$

4. The approximate p-value is given by $P(\chi_1^2 > X^2)$.
5. This test should only be used if $\mu_0 \geq 10$.

Person-Time Data

Cumulative Incidence Rates and Incidence Rate (Density)

1. For the purpose of allowing for varying follow-up time for each individual, we define the concept of **incidence density** $(ID = \lambda)$ that a group is defined by the number of events in that group divided by the total person-year accumulated during the study group.
2. The denominator used in computing incidence density is the person-year.
3. Suppose that X events are observed over T person-years of follow-up, the incidence rate is

$$\widehat{ID} = \hat{\lambda} = \frac{Y}{T}. \qquad (22)$$

Cumulative Incidence Rates and Incidence Rate (Density)

4. Unlike cumulative incidence, incidence density may range from 0 to infinity (∞). In following a subject, the incidence density may remain the same or may vary over time (i.e., as a subject's ages over time, the incidence density generally in creases).

5. How can we relate cumulative incidence over time T to incidence density?

6. Suppose for simplicity that incidence density remains the same over some time period T.

7. If $CI(T)$ is the cumulative incidence over time T and λ is the incidence density, then it can be shown using calculus methods that

$$CI(T) = 1 - e^{\lambda T} \quad (23)$$

Cumulative Incidence Rates and Incidence Rate (Density)

8. If the cumulative incidence is lower (less than 0.1), then we can approximate $e^{-\lambda T}$ by $1 - \lambda T$ and $CI(T)$ by

$$CI(T) = 1 - e^{\lambda T} \approx 1 - (1 - \lambda T) = \lambda T \quad (24)$$

Cumulative Incidence Rates and Incidence Rate (Density)

9. Note: Incidence density has a more commonly used term **incidence rate** (λ) and distinguished it from the cumulative incidence (CI) over some time period T.

10. The former can range from 0 to infinity, while the latter is a proportion and must vary between 0 and 1.

11. As was the case in obtaining exact confidence limits for the binomial parameter p, it is difficult to compute μ_1, μ_2 exactly.

Cumulative Incidence Rates and Incidence Rate (Density)

12. In some instances, a random variable representing a rare event over time is assumed to follow a **Poisson distribution** corresponding but the actual amount of person-time is either unknown or is not reported in an article from the literature.

13. In this instance, it is still possible to obtain a confidence interval for μ, although it is impossible to obtain a confidence variance of λ.

One-Sample Inference for Incidence-Rate Data

1. Exact Method

2. Approximate Method

One-Sample Inference for Incidence-Rate Data
Approximated Method

1. Suppose that X events are observed over T person-years of follow-up and that ID is the unknown underlying incidence (rate) and is be estimated from the data.

2. We wish to test the hypothesis

$$H_0 : ID = ID_0 \quad \text{versus} H_A : ID \neq ID_0 \tag{25}$$

where ID is the unknown incidence density (rate) in the sample. and ID_0 is the known incidence density (rate) in the specific population.

One-Sample Inference for Incidence-Rate Data
Approximated Method

3. We will base out test on the observed number of which we denote Y events. we will assume that X approximately follow Poisson distribution

4. Under H_0, X has mean as $\mu = T(ID_0)$ and variance as $\mu_0 = T(ID_0)$, where T is the total number of person-years.

One-Sample Inference for Incidence-Rate Data
Approximated Method

5. If we assume that the normal approximation to the Poisson distribution is valid, then this suggests:

(a) Compute the test statistic

$$X^2 = \frac{(X - \mu_0)^2}{\mu_0} \sim \chi^2_1, \quad \text{under } H_0 \tag{26}$$

where

$$\mu_0 = T(ID) \tag{27}$$

One-Sample Inference for Incidence-Rate Data
Approximated Method

(b) For two-sided test at level α, we

reject H_0, if $X^2 > \chi^2_{1,1-\alpha}$; (2

accept H_0, if $X^2 \leq \chi^2_{1,1-\alpha}$. (2

(c) The exact p-value is $P(\chi^2_1 > X^2)$.

(d) This test should only be used if $\mu_0 = T(ID_0) \geq 10$.

One-Sample Inference for Incidence-Rate Data
Exact Method

1. Suppose that X events are observed over T person-years of follow-up.

2. Suppose that the number of events is too small to apply the large-sample test.

3. In this case, an exact test based on the Poisson distribution must be used.

4. If $\mu = T(ID)$, the we can restate the hypothesis in the form

$$H_0 : \mu = \mu_0 \quad \text{versus} H_A : \mu \neq \mu_0 \tag{30}$$

and apply the one-sample Poisson test.

One-Sample Inference for Incidence-Rate Data
Exact Method

5. Under H_0, the observed number of events (Y) will follow Poisson distribution with parameter $\mu_0 = T(ID_0)$.

6. Thus, the exact two-sided p-value is given by

$$\min\left[2 \times \sum_{k=0}^{Y} \frac{e^{-\mu_0}\mu_0^k}{k!}, 1\right], \qquad \text{if } Y < \mu_0; \tag{31}$$

$$\min\left[2 \times \left(1 - \sum_{k=0}^{Y-1} \frac{e^{-\mu_0}\mu_0^k}{k!}\right), 1\right], \qquad \text{if } Y > \mu_0 \tag{32}$$

Confidence Limits for Incidence Rates

1. Suppose that X events are observed over T person-years of follow-up.
2. To obtain confidence limits for ID, we obtain confidence limits for the expected number of events (μ) based on the Poisson distribution and then divide each confidence limit by T, the number of person-years of follow-up.

Confidence Limits for Incidence Rates

3. Specifically, we have $\hat{\mu} = X, \widehat{\mathbf{Var}}(\hat{\mu}) = X$.
4. Thus, if the normal approximation to the Poisson distribution holds (i.e., $X \geq 10$), then a approximate $(1-\alpha) \times 100\%$ confidence interval for μ is given by $X \pm Z_{1-\alpha/2}\sqrt{X}$.
5. The corresponding approximate $(1-\alpha) \times 100\%$ confidence interval for ID is given by $(X \pm Z_{1-\alpha/2}\sqrt{X})/T$.
6. Otherwise, if $X < 10$, we obtained exact confidence limits for μ $c_1, c_2)$ and divide each confidence limit by T to obtain the corresponding confidence interval for ID.

Confidence Limits for Incidence Rates

(a) A point estimate of the incidence density rate is

$$\widehat{ID} = \hat{\lambda} = X/T. \quad (33)$$

(b) To obtain a two-side $(1-\alpha) \times 100\%$ confidence interval for μ,
  i. if $X \geq 10$, then compute $X \pm Z_{1-\alpha/2}\sqrt{X} = (c_1, c_2)$,
  ii. if $X < 10$, the obtained (c_1, c_2) exact confidence interval for X.

(c) The corresponding two-sided $(1-\alpha) \times 100\%$ confidence interval for ID is given by $(c_1/T, c_2/T)$.

Two-Sample Inference for Incidence-Rate Data

1. Exact Method
2. Approximate Method

Two-Sample Inference for Incidence-Rate Data

1. How can we compare the underlying incidence rates between two different groups ?
2. One approach is to use a conditional test.
3. Specifically, suppose we consider he case of two groups and have the general table in Table 2

Two-Sample Inference for Incidence-Rate Data

Table 2: Two-Sample Inference for Incidence-Rate Data

Group	Number of Events	Person-Time
Exposed A	Y_A	T_A
Unexposed B	Y_A	T_B
Total	$Y_A + Y_A$	$T_A + T_B$

Two-Sample Inference for Incidence-Rate Data: Approximate Method

4. We wish to test the hypothesis

$$H_0 : ID_A = ID_B \quad \text{versus} \quad ID_A \neq ID_B \tag{34}$$

where ID_A and ID_B are the incidence densities (rates) for group A and B respectively.

5. Under the null hypothesis, the fraction $T_A/(T_A + T_A)$ of the total of events $(Y_A + Y_B)$ would be expected to occur in group A, and the fraction $T_B/(T_A + T_B)$ of the total number of events $(Y_A + Y_B)$ would be expected to occur in group B.

Two-Sample Inference for Incidence-Rate Data: Approximate Method

6. Furthermore, under H_0 conditional on the observed total number events $(Y_A + Y_B)$, the expected number of events in each group is given by

$$\text{Expected number of events in group } A = E_A = (Y_A + Y_B)T_A/(T_A +$$
$$\text{Expected number of events in group } B = E_B = (Y_A + Y_B)T_B/(T_A +$$

Two-Sample Inference for Incidence-Rate Data: Approximate Method

7. To assess statistical significance, the number of events in group A under H_0 is treated as a binomial random variable with parameters $n = (Y_A + Y_B)$ and $p_0 = T_A/(T_A + T_B)$.

8. Under this assumption, the hypotheses can be stated as

$$H_0 : p = p_0 \quad \text{versus} \quad H_A : p \neq p_0, \tag{35}$$

where p is the true proportion of events that are expected to occur in group A.

9. We will also assume that the normal approximation to the binomial

distribution is valid.

Two-Sample Inference for Incidence-Rate Data: Approximate Method

10. Using the normal approximation to he binomial distribution, the observed number of events in group A is Y_A is normally distributed with mean $np_0 = (Y_A + Y_B)T_A/(T_A + T_B) = E_A$, and variance is $np_0q_0 = (Y_A + Y_B)T_AT_B/(T_A + T_B)^2 = V_A$.

11. H_0 will be rejected if Y_A is much smaller or larger than E_A.

12. This is an application of the large-sample one-sample binomial test.

Two-Sample Inference for Incidence-Rate Data: Approximate Method

13. So, to test the hypothesis

$$H_0 : ID_A = ID_B \quad \text{versus} \quad ID_A \neq ID_B. \tag{36}$$

14. We use the following procedures:

Two-Sample Inference for Incidence-Rate Data: Approximate Method

(a) Compute the test statistic

$$z = \begin{cases} \frac{Y_A - E_A - 0.5}{\sqrt{V_A}}, & \text{if } Y_A > E_A; \\ \frac{Y_A - E_A + 0.5}{\sqrt{V_A}}, & \text{if } Y_A \le E_A \end{cases} \quad (37)$$

where

$$E_A = (Y_A + Y_B)T_A/(T_A + T_B) \quad (38)$$

$$V_A = (Y_A + Y_B)T_A T_B/(T_A + T_B)^2 \quad (39)$$

(b) For a two-sided test with level α

$$\text{reject } H_0, \quad \text{if } z > Z_{\alpha/2} \text{ or } z < Z_{\alpha/2}; \quad (40)$$

$$\text{accept } H_0, \quad \text{if } Z_{\alpha/2} \le z \le< Z_{1-\alpha/2}. \quad (41)$$

(42)

Two-Sample Inference for Incidence-Rate Data: Approximate Method

(c) The p-value for this test is given by

$$2 \times [1 - \Phi(z)], \quad \text{if } z \ge 0; \quad (4$$

$$2 \times \Phi(z), \quad \text{if } z \le 0; \quad (4$$

$$\text{or} \quad 2 \times [1 - \Phi(|z|)]. \quad (4$$

(d) Use this test only if $V_A \ge 5$.

Two-Sample Inference for Incidence-Rate Data: Exact Method

1. Suppose that the number of events is too small to apply th normal-theory test (i.e. $V_A < 5$). In this case, an exact test based on the binomial distribution must be used.

2. Under H_0, the number of events in group A (Y_A) will follow a binomial distribution with parameters $n = (Y_A + Y_B)$ and $p = p_0 = T_A/(T_A + T_B), q_0 = 1 - p_0$.

Two-Sample Inference for Incidence-Rate Data: Exact Method

3. We wish to test the hypothesis

$$H_0 : ID_A = ID_B \quad \text{versus} \quad ID_A \ne ID_B \quad (46)$$

or equivalently, to test

$$H_0 : p = p_0 \quad \text{versus} \quad H_A : p \ne p_0, \quad (47)$$

where p is the underlying proportion of events that occur in group A, and $p_0 = T_A/(T_A + T_B)$.

Two-Sample Inference for Incidence-Rate Data: Exact Method

4. This is an application to the exact one-sample binomial test. H_0 will be reject if the observed number of events Y_A is much smaller or much larger than the expected number of events $E_A = np_0$.

Two-Sample Inference for Incidence-Rate Data: Exact Method

(a) If $Y_A < (Y_A + Y_B)p_0$, then

$$p\text{-value} = 2 \times \sum_{k=0}^{Y_A} \binom{Y_A + Y_B}{k} p_0^k q_0^{Y_A + Y_B - k} \quad (48)$$

(b) if $Y_A > (Y_A + Y_B)p_0$, then

$$p\text{-value} = 2 \times \sum_{k=Y_A}^{Y_A + Y_B} \binom{Y_A + Y_B}{k} p_0^k q_0^{Y_A + Y_B - k} \quad (49)$$

(c) This test is valid in general for comparing two incidence densities but is particularly useful when $V_A < 5$, in which case the normal-theory estimation should not be used.

Incidence Rate Ratio (Risk Ratio)

1. Risk ratio (RR) is a measure of effect for the comparison of two proportions.
2. We applied this measure to compare cumulative incidences between two exposure groups in a prospective study, where the person was the unit of analysis.
3. A similar concept can be employed to compare two incidence rates based on the person-year data.

Incidence Rate Ratio (Risk Ratio)

4. Let λ_A, λ_B be incidence rates for an exposed and unexposed group, respectively.
5. The **rate ratio is defined as** λ_A/λ_B.
6. What is the relationship between the rate ratio based on the incidence rates and the risk ratio based on cumulative incidence ?

Incidence Rate Ratio (Risk Ratio)

7. Suppose each person in a cohort is followed for T years, with incidence rate λ_A in the exposed group A and λ_B in the unexposed group B.
8. If the cumulative incidence is low, then the cumulative incidence will be approximately $\lambda_A T$ and in the exposed group A, and $\lambda_B T$ in the unexposed group B.
9. Thus, the risk ratio will be approximately $(\lambda_A T)/(\lambda_B T) = \lambda_A/\lambda_B$, ratio.

Incidence Rate Ratio (Risk Ratio)

10. How can we estimate the rate ratio from observed data ?
11. Suppose we have the number of events in the exposed group A, and person-years shown in Table 2.
12. The estimated incidence rate in the exposed group A as Y_A/T_A and in the unexposed group B as Y_B/T_B.
13. A point estimate of the rate ratio is given by

$$RR = \frac{Y_A/T_A}{Y_B/T_B}. \quad (50)$$

Incidence Rate Ratio (Risk Ratio)

14. To obtain an interval estimate, we assume approximate normality of $\log(\widehat{OR})$.
15. The variance of $\log(\widehat{OR})$ is approximated

$$\mathbf{Var}(\log(\widehat{OR})] \approx \frac{1}{Y_A} + \frac{1}{Y_B}. \quad (51)$$

16. Therefore, a two-sided $(1-\alpha) \times 100\%$ C.I. for $\log(\widehat{OR})$ is given by

$$(c_1, c_2) = \log(\widehat{OR}) \pm Z_{1-\alpha/2}\sqrt{\frac{1}{Y_A} + \frac{1}{Y_B}}. \quad (52)$$

Incidence Rate Ratio (Risk Ratio)

17. If we take the anti-log of c_2, c_2, we obtain a two-sided $(1-\alpha) \times 100\%$ as

$$(r_1, r_2) = (e^{c_1}, e^{c_2}). \quad (53)$$

18. This interval should only be used if $V_A = [(Y_A + Y_B)T_A T_B]/[(T_A + T_B)^2] \geq 5$.

Inference for Stratified Person-Time Data

Inference for Stratified Person-Time Data

1. It is very common in the analysis of person-time data to control for confounding variables before assessing the relationship between the main exposure of interest and disease.
2. Confounding variables may include age and sex as well as other covariates that are related to exposure, disease, or both.
3. We can use methods similar to the Mantel-Haenszel test used for cumulative incidence data (or generally for count data).
4. Suppose we have k strata, where the number of events and the amount of person-time in the ith stratum are as shown in Table 3

Table 3: Stratified Two-Sample Inference for Incidence-Rate Data

Group	Number of events	Person-Time
Exposed A	Y_{iA}	T_{iA}
Unexposed B	Y_{iA}	T_{iB}
Total	$Y_{iA}+Y_{iA}$	$T_{iA}+T_{iB}$

5. Let us denote the incidence rate of disease among the exposed by p_{iA} and among the unexposed be p_{iB}.
6. Therefore, the expected number of events among the exposed is $p_{iA}T_{iA}$ and among the unexposed is $p_{iB}T_{iB}$.

7. Let p_i be the expected proportion of the total number of events over both groups that are among the exposed for stratum i.
8. We can relate p_i to p_{iA} and p_{iB} by

$$p_i = \frac{p_{iA}T_{iA}}{p_{iA}T_{iA}+p_{iB}T_{iB}} \tag{54}$$

9. **We assume that the rate ratio relating disease to expose is the same for each stratum and denote it by** RR.
10. Therefore, $RR = p_{iA}/p_{iB}$ and RR is the same for each $i = 1, \ldots, k$.
11. If we divide numerator and denominator by p_{iB}, and substitute RR for

p_{iA}/p_{iB}, we obtain

$$p_i = \frac{(p_{iA}/p_{iB})T_{iA}}{(p_{iA}/p_{iB})T_{iA}+T_{iB}} = \frac{RRT_{iA}}{RRT_{iA}+T_{iB}} \tag{55}$$

which denote by $p_{i(1)}$.

12. If $RR = 0$ then

$$p_i = \frac{RRT_{iA}}{RRT_{iA}+T_{iB}} = \frac{T_{iA}}{T_{iA}+T_{iB}} = p_{i0}, \quad \text{(under } RR = 1 \text{ assumption.)} \tag{56}$$

13. We wish to test the hypothesis

$$H_0: RR = 1 \quad \text{versus} \quad RR \neq 1 \tag{57}$$

or equivalently,

$$H_0: p_i = p_{i0} \quad \text{versus} \quad p_i = p_{i(1)}, i = 1, 2, \ldots k. \tag{58}$$

14. We will base our test on $S = \sum_{i=1}^{k} Y_{i1}$, the total observed number of events for the exposed.
15. Under H_0, we will assume that the total observed number of events for the ith stratum $(Y_{iA}+Y_{iB})$ is fixed.
16. Therefore, under H_0:

$$\mathcal{E}(Y_{iA}) = (Y_{iA}+Y_{iB})p_{i0} = (Y_{iA}+Y_{iB})T_{iA}/(T_{iA}+T_{iB}) \quad ($$

$$\mathbf{Var}(Y_{iA}) = (Y_{iA}+Y_{iB})p_{i0}(1-p_{i0}) = (Y_{iA}+Y_{iB})T_{iA}T_{iB}/(T_{iA}+T_{i} \quad ($$

and

$$\mathcal{E}(S) = \sum_{i=1}^{k} \mathcal{E}(Y_{iA}) \quad (61)$$

$$\mathbf{Var}(S) = \sum_{i=1}^{k} \mathbf{Var}(Y_{iA}) \quad (62)$$

17. Under H_A, S will be larger than $\mathcal{E}(S)$ if $RR > 1$ and will be smaller than $\mathcal{E}(S)$ if $RR < 1$.

(a) We compute the test statistic

$$X^2 = \frac{\left(|S - \mathcal{E}(S)| - 0.5\right)^2}{\mathbf{Var}(S)} \quad (63)$$

(b) which follow a chi-squared distribution with 1 df under H_0.

(c) We

$$\text{reject } H_0, \quad \text{if } X^2 > \chi^2_{1,1-\alpha}; \quad (64)$$

$$\text{accept } H_0, \quad \text{if } X^2 \leq \chi^2_{1,1-\alpha}. \quad (65)$$

(d) The p-value is $P(\chi^2_1 > X^2)$.

(e) The test should only be used if $\mathbf{Var}(S) \geq 5$.

Estimation of the Rate Ratio

Estimation of the Rate Ratio

1. We obtain estimates of the $\log(RR_i)$ in each stratum i and then compute a weighted average of the stratum-specific estimates to obtain an overall of the $\ln(RR)$.

2. Specifically, let

$$\widehat{OR}_i = (Y_{iA}/T_{iA})/(Y_{iB}/T_{iB}) \quad (66)$$

be the estimate of the rate ratio in the ith stratum.

3. We have

$$\mathbf{Var}[\log(\widehat{OR}_i)] \approx \frac{1}{Y_{iA}} + \frac{1}{Y_{iB}} \quad (67)$$

Estimation of the Rate Ratio

4. To obtained an overall estimate of $\log(\widehat{OR})$ we now compute a weighted average of $\ln(\widehat{OR}_i)$ and then take anti-log of the weighted average

$$\log[\widehat{OR}] = \frac{\sum_{i=1}^{k} w_i \log(\widehat{OR}_i)}{\sum_{i=1}^{k} w_i} \quad (68)$$

where $w_i = 1/\mathbf{Var}[\log(\widehat{OR}_i)]$.

Estimation of the Rate Ratio

5. We then obtain the variance of $\log(\widehat{OR})$ as

$$\mathbf{Var}[\log(\widehat{OR})] = \frac{1}{(\sum_{i=1}^{k} w_i)^2}\mathbf{Var}\left[\sum_{i=1}^{k} w_i \log(\widehat{OR}_i)\right] \quad (69)$$

$$= \left[(\sum_{i=1}^{k} w_i)^2\right]^{-1} \sum_{i=1}^{k} w_i^2 \mathbf{Var}[\log(\widehat{OR}_i)] \quad (70)$$

$$= \left[(\sum_{i=1}^{k} w_i)^2\right]^{-1} \sum_{i=1}^{k} w_i^2 (1/w_i) \quad (71)$$

$$= \left[(\sum_{i=1}^{k} w_i)^2\right]^{-1} \sum_{i=1}^{k} w_i \quad (72)$$

$$= \left(\sum_{i=1}^{k} w_i\right)^{-1} \quad (73)$$

Estimation of the Rate Ratio

6. Thus, a two-sided $(1-\alpha)\times 100\%$ C.I. for $\log(RR)$, (c_1, c_2) is given by $\log(\widehat{OR}) \pm Z_{1-\alpha/2} \times (\sum_{i=1}^{k} w_i)^{-1/2}$.

7. We then take the anti-log of each of the confidence limits for $\log(RR)$ to obtain confidence interval (e^{c_1}, e^{c_2}).

Testing the Assumption of Homogeneity of the Rate Ratio Across Strata

1. **An important assumption made in the estimation methods is that the underlying rate ratio is the same in all strata**.

2. If the rate ratios in different strata are all in the same direction relative to the null hypothesis (i.e., all rate ratio > 1 or all rate ratio < 1). the hypothesis-testing procedures will still be valid with only a slightly loss of power.

3. However, if the rate ratio are in different directions in different strata, or are null in some strata, then the power of the hypothesis-testing procedures will be greatly diminished.

Testing the Assumption of Homogeneity of the Rate Ratio Across Strata

4. To test this assumption, we use similar methods to those for testing the assumption of homogeneity of the odds ratio in different strata for count data.

5. Specifically, we wish to test the hypothesis

$$H_0: \quad RR_1 = RR_2 = \cdots = RR_k$$
$$\text{versus} \quad H_A: \quad \text{at least two of the } RR_i \text{ are different} \quad (74)$$

with significance level α.

Testing the Assumption of Homogeneity of the Rate Ratio Across Strata

We use the following procedures:

(a) We compute the test statistic

$$X^2_{\text{HOM}} = \sum_{i=1}^{k} w_i[\log(\widehat{OR}_i) - \log(\widehat{OR})]^2 \sim \chi^2_{k-1} \quad \text{under } H_0. \quad (75)$$

Testing the Assumption of Homogeneity of the Rate Ratio Across Strata

(b) We

reject H_0, if $xx_{\text{HOM}} > \chi^2_{k-1,1-\alpha}$; (7

accept H_0, if $xx_{\text{HOM}} \le \chi^2_{k-1,1-\alpha}$. (7

(c) The p-value is given by $P(\chi^2_{k-1} > X^2_{\text{HOM}})$.