Example, continued...

In population,

- Risk difference (RD) = 5% - 1% = 4%
- Risk Ratio (RR) = 5%/1% = 5
- Odds Ratio (OR) = 5%/95% × 99%/1% = 5.2

If event $E$ has probability $P$ the the odds of $E$ is $\frac{P}{1-P}$. E.g., if $P = 1/3$ then odds are 1:2 or 1/2.

For two events, $E_1$ and $E_0$, the odds ratio is

$$\frac{P_1}{1-P_1} \div \frac{P_0}{1-P_0} = \frac{P_1}{1-P_1} \cdot \frac{1-P_0}{P_0}$$

Because apparent risk is unchanged in studies A and B, all three of these measures are the same as in A and B as in the whole population.

In study C,

- Apparent RD = 74% - 36% = 39%
- Risk Ratio (RR) = 74%/36% = 2.1
- Odds Ratio (OR) = 20/7 × 29/16 = 5.2

while row and column percents vary depending on sampling scheme, the invariant feature common to all tables is the Odds Ratio.

Properties of OR:

- If there is no association between exposure and disease, then $P_1 = P_0$ so that OR = 1.
- If $P_1 > P_0$ then OR > 1
- If $P_1, P_0 \approx 0$ then $1 - P_1, 1 - P_0 \approx 1$ so

$$\frac{P_1}{1-P_1} \cdot \frac{1-P_0}{P_0} \approx \frac{P_1}{P_0},$$

so OR is an approximation to Risk Ratio (RR).

- While is is not possible to recover RR ($\frac{P_1}{P_0}$) when only the OR is known, the OR is always more extreme than RR I.e., if $P_1 < P_0$, then OR < RR < 1 and if $P_1 > P_0$, then OR > RR > 1.

Example:

<table>
<thead>
<tr>
<th></th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td>E+</td>
<td>5</td>
<td>1000</td>
</tr>
<tr>
<td>E-</td>
<td>1</td>
<td>1000</td>
</tr>
</tbody>
</table>

OR = 5, R.R = 5/1005 × 1001/1 = 4.98
If the roles of disease/non-disease or exposed/non-exposed are interchanged, the OR is inverted. This is *NOT* true of RR. Risk Ratio depends on the reference category and choice of marginal totals.

- Odds ratios arise naturally from logistic regression models and conditional analyses.

If there is a single concept in this course that is the most critical, it would be the concept of Odds Ratios.

**Underlying Probability Model**

Suppose that \( t = 0 \) is the start of observation for an arbitrary disease-free subject and let \( P(t) \) be the probability that the subjects acquires disease prior to time \( t \). (If \( T \) is the random time of onset of disease, then \( P(t) \) is the cumulative distribution function (CDF) of \( T \) and is an increasing function of \( t \).) If \( t_1 < t_2 \) then the probability of acquiring disease during the interval \((t_1, t_2)\) is \( P(t_2) - P(t_1) \). The *conditional* probability that a subject acquires disease in the interval \((t_1, t_2)\) given that they were disease-free prior to \( t_1 \) is

\[
P(t_2) - P(t_1) \quad \frac{1}{1 - P(t_1)}.
\]

If \( \Delta t = t_2 - t_1 \) is small, then

\[
\lambda(t_1) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t_1 + \Delta t) - P(t_1)}{1 - P(t_1)} = \frac{1}{1 - P(t_1)} \frac{dP(t_1)}{dt}
\]

is the *Instantaneous Risk* or *Hazard* of acquiring disease at time \( t \). Note that the units are cases/number of subjects at risk/unit of time. (*E.g.* cases per 100,000 per year)
Solving the above equation for $P(t)$, we get

$$P(t) = 1 - \exp(-\int_0^t \lambda(u)du) = 1 - \exp(-\Lambda(t))$$

where

$$\Lambda(t) = \int_0^t \lambda(u)du$$

is the *Cumulative Incidence* or *Cumulative Hazard*.

Also, the probability density function (PDF) of $T$ is

$$f(t) = \frac{dP(t)}{dt} = \lambda(t) \exp(-\Lambda(t)).$$

Certainly $P(t)$ is a probability (between 0 and 1), however, $\Lambda(t)$ may be greater than one. So what is the significance of $\Lambda(t)$? Usually we don’t attribute significance to $\Lambda(t)$ itself, however, if subjects were allowed to re-acquire disease (say, for example, the “common cold”) and future episodes are independent of past episodes, then the expected number of episodes between time 0 and time $t$ is $\Lambda(t)$.

**Important Special Case:** Now, consider the special case where $\lambda(t) = \lambda$ is constant. Then $\Lambda(t) = \lambda t$, and

$$P(t) = 1 - \exp(-\lambda t).$$

This is the *exponential* distribution.

Suppose that we have a group of subjects $i = 1, 2, \ldots, n$ who are followed for varying lengths of time $T_1, T_2, \ldots, T_n$ at which time each acquires the disease in question. How can we estimate $\lambda$?

**Maximum Likelihood:**

The *likelihood* for $\lambda$ is simply

$$L = \prod_{i=1}^n f(T_i) = \prod_{i=1}^n \lambda \exp(-\lambda T_i).$$

The *log-likelihood* is

$$\log(L) = \sum_{i=1}^n \log(\lambda) - \lambda T_i.$$  

The *Maximum Likelihood Estimate* of $\lambda$, $\hat{\lambda}$ is the maximizer of $L$ or $\log(L)$ which we obtain by solving

$$\frac{\partial}{\partial \lambda} \log(L) = \sum_{i=0}^n \frac{1}{\lambda} - T_i = 0.$$  

Hence, if $T = \sum T_i$ is the total observation time,

$$\hat{\lambda} = \frac{n}{T}$$
which is simply the total number of subjects or events divided by the total observation time.

If we do not observe all subjects to fail (acquire disease) then the likelihood for those subjects is simply the probability of reaching time $T_i$ without failing which is $1 - P(T_i)$. If we let $\delta_i$ be the indicator of disease for subject $i$, i.e.

$$\delta_i = \begin{cases} 1 & \text{if subject } i \text{ acquires disease} \\ 0 & \text{if not.} \end{cases}$$

Then the likelihood can be written as:

$$L = \prod_{i=1}^{n} f(T_i)^{\delta_i} (1 - P(t))^{1-\delta_i}$$

$$= \prod_{i=1}^{n} (\lambda \exp(-\lambda T_i))^{\delta_i} \exp(-\lambda T_i)^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \lambda^{\delta_i} \exp(-\lambda T_i)$$

[To be mathematically precise, $L$ does not correspond to the product of density functions in the usual sense (i.e. with respect to Lebesgue measure). However, if $U_i$ represent censoring times, (which may be random variables, provided that they are independent of the $T_i$), then $L$ is the product of Radon-Nykodym derivatives of the CDF’s of the $T_i$ with respect to Lebesgue measure plus point masses at the $U_i$. One sometimes calls this a generalized density function]

If $d = \sum \delta_i$ is the total number of observed events, then we may write the log-likelihood as

$$\log(L) = d \log(\lambda) - \lambda T.$$

Again, this is maximized by

$$\hat{\lambda} = \frac{d}{T} = \frac{\text{events}}{\text{person years exposure}}$$

How reliable is the estimate $\hat{\lambda}$?

MLE theory (we’ll talk about this in more detail later) says that

$$\text{var}(\hat{\lambda}) \approx \frac{1}{E[\partial^2 \log(L)/\partial \lambda^2]}$$

And,

$$-E \left[ \frac{\partial^2 \log(L)}{\partial \lambda^2} \right] = E \left[ \frac{d}{\lambda^2} \right] = \frac{E[d]}{\lambda^2}$$
so we can estimate \( \var(\hat{\lambda}) \) by
\[
\var(\hat{\lambda}) \approx \frac{\hat{\lambda}^2}{d} = \frac{d}{T^2}
\]

**Rate Standardization:**

Now suppose that we have a sample of \( n \) subjects followed for varying lengths of time, starting at a variety of ages. Since hazard is typically a function of age, we can’t expect that the hazard will be the same for all subjects, nor constant across the followup period.

A simple procedure is to divide the range of observations into subintervals so that hazard is approximately constant within each interval. Within each interval, we may count the total number of events and the total observation time and estimate the hazard in each interval separately.

For example, in the adjacent figure, start of observation for each subject is denoted by the \( \bullet \), end of observation without an event by “W” (withdrawn) and events are denoted by “D”. In the 4\(^{th}\) interval, there is one event and one withdrawal and the total observation time is 4.7. We estimate hazard by \( \lambda_4 = 1/4.7 = .21 \). The remaining \( \lambda_i \) can be computed similarly.

It would also be useful to be able to summarize the overall hazard in some simple, meaningful way. Since different populations may differ in the proportion of subjects in various age categories, a useful summary measure should depend only on the underlying hazard function and not on the population distribution.

Suppose that we have two populations: a standard population, A, and a population under study, B. If the adjacent figure, population A is generally younger than population B. Thus if there both populations share a common increasing hazard function, the observed overall incidence of disease will be greater in population B than in population A. Thus, the overall incidence rate is not an adequate summary measure.

For each subinterval, let \( w_i \) be the proportion (or number) of subjects in the standard population in the \( i^{th} \) subinterval. For convenience, assume that
\[
\sum_i w_i = 1
\]
If $d_i$ and $T_i$ are the total number of events and total observation time in the $i^{th}$ interval, let

$$\hat{\lambda}_i = \frac{d_i}{T_i}$$

**Direct Standardization:** The quantity

$$\hat{\lambda} = \sum w_i \hat{\lambda}_i$$

is the expected *cumulative incidence* rate if population B had the same population distribution as population A. Note that $\hat{\lambda}$ depends on the population distribution in the standard population A, not in population B.

Example:

<table>
<thead>
<tr>
<th>age</th>
<th>$w_i$</th>
<th>$\hat{\lambda}_i(B)$ (per 10$^6$)</th>
<th>$w_i(A)\hat{\lambda}_i$</th>
<th>$w_i(B)\hat{\lambda}_i$</th>
<th>$\lambda_i^*$ (per 10$^6$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-20</td>
<td>.40</td>
<td>.400</td>
<td>2</td>
<td>.80</td>
<td>.80</td>
</tr>
<tr>
<td>20-40</td>
<td>.25</td>
<td>.400</td>
<td>5</td>
<td>1.25</td>
<td>2.0</td>
</tr>
<tr>
<td>40-60</td>
<td>.22</td>
<td>.150</td>
<td>10</td>
<td>2.20</td>
<td>1.5</td>
</tr>
<tr>
<td>60-80</td>
<td>.11</td>
<td>.045</td>
<td>20</td>
<td>2.20</td>
<td>.90</td>
</tr>
<tr>
<td>80+</td>
<td>.02</td>
<td>.005</td>
<td>40</td>
<td>.80</td>
<td>.20</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>7.25</td>
<td>5.4</td>
<td>5.9*</td>
<td></td>
</tr>
</tbody>
</table>

(* $\sum w_i(A)\lambda_i^* = 5.9$)

Here $\lambda_i^*$ is the hazard rate in the standard population. (Note that the same hazard rates in different populations yield different cumulative hazards.)

It was common historically to standardize rates to specific predefined populations. Tables 2.4 and 2.5 in Breslow and Day, Vol II, give distributions for a number of standard populations. The choice of standard population may be determined by the nature and purpose of the data under study.

We can easily compute variances:

$$\text{var}(\hat{\lambda}) = \sum w_i^2 \text{var}(\lambda_i) = \sum w_i^2 \frac{d_i}{T_i^2}$$

**Comparative Mortality Figure:** If, in addition, we have standard rates $\lambda_i^*$, we can compute $\Lambda^*$ and the ratio

$$\frac{\hat{\Lambda}}{\Lambda^*} = \frac{\sum w_i \hat{\lambda}_i}{\sum w_i \lambda_i^*}$$

This ratio is called the “Comparative Mortality Figure” (CMF).

In the example, $\text{CMF} = 7.25/5.9 = 1.23$. 

Last modified 1/26/05