## 6. Applications of Probability in Epidemiology

<table>
<thead>
<tr>
<th>Topics</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Probability in Diagnostic Testing</td>
<td>3</td>
</tr>
<tr>
<td>a. Prevalence</td>
<td>3</td>
</tr>
<tr>
<td>b. Incidence</td>
<td>3</td>
</tr>
<tr>
<td>c. Sensitivity, Specificity</td>
<td>4</td>
</tr>
<tr>
<td>d. Predictive Value Positive, Negative Test</td>
<td>7</td>
</tr>
<tr>
<td>2. Probability and Measures of Association for the 2x2 Table</td>
<td>9</td>
</tr>
<tr>
<td>a. Risk</td>
<td>9</td>
</tr>
<tr>
<td>b. Odds</td>
<td>11</td>
</tr>
<tr>
<td>c. Relative Risk</td>
<td>13</td>
</tr>
<tr>
<td>d. Odds Ratio</td>
<td>15</td>
</tr>
</tbody>
</table>
Please Quiet Cell Phones and Pagers

Thank you.
1. Probability in Diagnostic Testing

a. Prevalence ("existing")

The point prevalence of disease is the proportion of individuals in a population that has disease at a single point in time (point), regardless of the duration of time that the individual might have had the disease.

Prevalence is NOT a probability.

Example -

A study of sex and drug behaviors among gay men is being conducted in Boston, Massachusetts. At the time of enrollment, 30% of the study cohort were sero-positive for the HIV antibody. Rephrased, the point prevalence of HIV sero-positivity was 0.30 at baseline.

b. Incidence ("new")

The incidence of disease is the probability an individual who did not previously have disease will develop the disease over a specified time period.

Example -

Consider again Example 1, the study of gay men and HIV sero-positivity. Suppose that, in the two years subsequent to enrollment, 8 of the 240 study subjects sero-converted. This represents a two year cumulative incidence rate of 8/240 or 3.33%.
Sensitivity, Specificity

The ideas of sensitivity, specificity, predictive value of a positive test, and predictive value of a negative test are most easily understood using data in the form of a 2x2 table:

<table>
<thead>
<tr>
<th>Disease Status</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Positive</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Test Negative</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+b+c+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

In this table, each of a total of (a+b+c+d) individuals are cross-classified according to their values on two variables: disease (present or absent) and test result (positive or negative). It is assumed that a positive test result is suggestive of the presence of disease. The counts have the following meanings:

\[ a = \text{number of individuals who test positive AND have disease} \]
\[ b = \text{number of individuals who test positive AND do NOT have disease} \]
\[ c = \text{number of individuals who test negative AND have disease} \]
\[ d = \text{number of individuals who test negative AND do NOT have disease} \]
\[ (a+b+c+d) = \text{total number of individuals, regardless of test results or disease status} \]
\[ (b+d) = \text{total number of individuals who do NOT have disease, regardless of their test outcomes} \]
\[ (a+c) = \text{total number of individuals who DO have disease, regardless of their test outcomes} \]
\[ (a+b) = \text{total number of individuals who have a POSITIVE test result, regardless of their disease status.} \]
\[ (c+d) = \text{total number of individuals who have a NEGATIVE test result, regardless of their disease status.} \]
Sensitivity

Among those persons who are known to have disease, what are the chances that the diagnostic test will yield a positive result?

To answer this question requires restricting attention to the subset of \((a+c)\) persons who actually have disease. The number of persons in this subset is \((a+c)\). Among this "restricted total" of \((a+c)\), it is observed that \(\text{“a” test positive.}\)

\[
\text{sensitivity} = \frac{a}{a+c}
\]

Sensitivity is a conditional probability. It is the conditional probability that the test suggests disease given that the individual has the disease. For \(E_1\)=event that individual has disease and \(E_2\)=event that test suggests disease:

\[
\text{sensitivity} = P(E_2 \mid E_1)
\]

To see that this is equal to what we think it should be, \(\left( \frac{a}{a+c} \right)\), use the definition of conditional probability:

\[
P(E_2 \mid E_1) = \frac{P(E_2 \text{ and } E_1)}{P(E_1)}
\]

\[
= \frac{[a / (a + b + c + d)]}{[(a + c) / (a + b + c + d)]}
\]

\[
= \frac{a}{a+c}, \text{ which matches.}
\]

Different References have Other names for "Sensitivity":
* positivity in disease
* true positive rate
**Specificity**

Specificity pertains to:

Among those persons who do NOT have disease, what is the likelihood that the diagnostic test indicates this?

Specificity is a conditional probability. It is the conditional probability that the test suggests absence of disease given that the individual is without disease. For $E_3=$event that individual is disease free and $E_4=$event that test suggests absence of disease:

$$\text{sensitivity} = P(E_4 \mid E_3)$$

To see that this is equal to what we think it should be, $(d / [b+d])$, use the definition of conditional probability:

$$P(E_4 \mid E_3) = \frac{P(E_4 \text{ and } E_3)}{P(E_3)}$$

$$= \frac{[b / (a + b + c + d)]}{[(b + d) / (a + b + c + d)]}$$

$$= \frac{b}{b + d} \frac{1}{P_4} \text{, which matches.}$$

Different References have Other names for "Specificity":

* negativity in health
* true negative rate
d. Predictive Value Positive, Negative

Sensitivity and specificity are not very helpful in the clinical setting.

- We don’t know if the patient has disease.
- This is what we are wanting to learn.
- Thus, sensitivity and specificity are not the calculations performed in the clinical setting.

“For the person who is known to test positive, what are the chances that he or she truly has disease?".

- This is the idea of “predictive value positive test”

"For the person who is known to test negative, what are the chances that he or she is truly disease free?".

- This is the idea of “predictive value negative test”

Predictive Value Positive Test

Among those persons who test positive for disease, how many will actually have the disease?

Predictive value positive test is also a conditional probability. It is the conditional probability that an individual with a test indicative of disease actually has disease. Attention is restricted to the subset of the (a+b) persons who test positive. Among this "restricted total" of (a+b),

\[
\text{Predictive value positive} = \frac{a}{a + b}
\]
Other Names for "Predictive Value Positive Test":

* posttest probability of disease given a positive test
* posterior probability of disease given a positive test

Finally, will unnecessary care be given to a person who does not have the disease?

**Predictive Value Negative Test**

Among those persons who test negative for disease, how many are actually disease free?

Predictive value negative test is also a conditional probability. It is the conditional probability that an individual with a test indicative of NO disease is actually disease free. Attention is restricted to the subset of the (c+d) persons who test negative. Among this "restricted total" of (a+b),

\[
\text{Predictive value negative} = \frac{d}{c + d}
\]

Other Names for "Predictive Value Negative Test":

* posttest probability of NO disease given a negative test
* posterior probability of NO disease given a negative test
2. Probability and Measures of Association for the 2x2 Table

Consider next the question of the relationship between a dichotomous exposure variable and a dichotomous disease outcome variable. A 2x2 summary table is again useful.

\[
\begin{array}{ccc}
\text{Disease Status} & \text{Present} & \text{Absent} \\
\text{Exposed} & a & b & a + b \\
\text{Not} & c & d & c + d \\
\text{a} + c & b + d & a + b + c + d \\
\end{array}
\]

\[
(a+b+c+d) = \text{total number of individuals, regardless of exposure or disease status}
\]

\[
(b + d) = \text{total number of individuals who do NOT have disease, regardless of their exposure status}
\]

\[
(a + c) = \text{total number of individuals who DO have disease, regardless of their exposure status}
\]

\[
(a + b) = \text{total number of individuals who have a POSITIVE exposure, regardless of their disease status.}
\]

\[
(c + d) = \text{total number of individuals who have a NO exposure, regardless of their disease status.}
\]

a. Risk ("simple probability")

Risk of disease, without referring to any additional information, is simply the probability of disease. An estimate of the probability or risk of disease is provided by the relative frequency:

\[
\frac{a + c}{a + b + c + d}
\]
Typically, however, conditional risks are reported. For example, if it were of interest to estimate the risk of disease for persons with a positive exposure status, then attention would be restricted to the (a+b) persons positive on exposure. For these persons only, it seems reasonable to estimate the risk of disease by the relative frequency:

The straightforward calculation of the risk of disease for the persons known to have a positive exposure status is:

\[ P(\text{Disease among Exposed}) = \frac{a}{a + b} \]

Repeating the calculation using the definition of conditional probability yields the same answer. Let \( E_1 = \text{event of positive exposure} \) and \( E_2 = \text{event of disease} \). Then:

Risk (disease given POSITIVE exposure) =

\[
P(E_2|E_1) = \frac{P(E_2 \text{ and } E_1)}{P(E_1)}
\]

\[
= \frac{a / (a + b + c + d)}{(a + b) / (a + b + c + d)}
\]

\[
= \frac{a}{(a + b)c} \text{ which matches.}
\]
b. Odds("comparison of two complementary (opposite) outcomes"):

In words, the odds of an event "E" is the chances of the event occurring in comparison to the chances of the same event NOT occurring.

\[
\text{Odds} = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}
\]

Example -

Perhaps the most familiar example of odds is reflected in the expression "the odds of a fair coin landing heads is 50-50". This is nothing more than:

\[
\text{Odds(heads)} = \frac{P(\text{heads})}{P(\text{heads}^c)} = \frac{P(\text{heads})}{P(\text{tails})} = \frac{.50}{.50}
\]

Similarly, for the exposure-disease data in the 2x2 table,

\[
\text{Odds(\text{disease})} = \frac{P(\text{disease})}{P(\text{disease}^c)} = \frac{P(\text{disease})}{P(\text{NO disease})} = \frac{(a + c)}{(b + d)}
\]

\[
\text{Odds(\text{exposed})} = \frac{P(\text{exposed})}{P(\text{exposed}^c)} = \frac{P(\text{exposed})}{P(\text{NOT exposed})} = \frac{(a + b)}{(c + d)}
\]
What if it is suspected that exposure has something to do with disease? In this setting, it might be more meaningful to report the odds of disease separately for persons who are exposed and persons who are not exposed. The mathematical term for such odds is conditional odds.

\[
\text{Odds(disease | exposed)} = \frac{\Pr(\text{disease|exposed})}{\Pr(\text{NO disease|exposed})} = \frac{a}{b} = \frac{a}{(a+b)} \quad \text{and} \quad \frac{b}{(a+b)}
\]

\[
\text{Odds(disease | NOT exposed)} = \frac{\Pr(\text{disease|not exposed})}{\Pr(\text{NO disease|not exposed})} = \frac{c}{d} = \frac{c}{(c+d)} \quad \text{and} \quad \frac{d}{(c+d)}
\]

Similarly, one might calculate the odds of exposure separately for diseased persons and NON-diseased persons:

\[
\text{Odds(exposed | disease)} = \frac{\Pr(\text{exposed|disease})}{\Pr(\text{NOT exposed|disease})} = \frac{a}{c} = \frac{a}{(a+c)} \quad \text{and} \quad \frac{c}{(a+c)}
\]

\[
\text{Odds(exposed | NO disease)} = \frac{\Pr(\text{exposed|NO disease})}{\Pr(\text{NOT exposed|NO disease})} = \frac{b}{d} = \frac{b}{(b+d)} \quad \text{and} \quad \frac{d}{(b+d)}
\]
c. Relative Risk("comparison of two conditional probabilities")

Recall that various epidemiological studies (prevalence, cohort, case-control) give rise to data in the form of counts in a 2x2 table.

We consider now the goal of assessing the association between exposure and disease.

Recall our 2x2 table.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c</td>
</tr>
<tr>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

Let’s consider some actual counts:

<table>
<thead>
<tr>
<th>Disease</th>
<th>Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>2</td>
</tr>
<tr>
<td>Not exposed</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>298</td>
</tr>
</tbody>
</table>

We might have more than one 2x2 table if the population of interest is partitioned into subgroups or strata.

Example: Stratification by gender would yield a separate 2x2 table for men and women.
Relative Risk

The relative risk is the ratio of the conditional probability of disease among the exposed to the conditional probability of disease among the non-exposed.

\[
RR = \frac{a}{a+b} \div \frac{c}{c+d}
\]

**Example:** In our 2x2 table, we have \(\frac{a}{a+b} = \frac{2}{10} = .20\), \(\frac{c}{c+d} = \frac{10}{300} = .0333\)

Thus, \(RR = .20 / .0333 = 6.006\)

- It has been found empirically that many exposure-disease relationships vary with age in such a way that the log linear model is a good description. Specifically, the change with age in the relative risk of disease with exposure is reasonably stable. In such instances, the model is preferable to an additive risk model.
d. Odds Ratio

The odds ratio measure of association has some wonderful advantages, both biological and analytical. Recall first the meaning of an “odds”:

Recall that if \( p = \text{Probability[event]} \) then \( \text{Odds[Event]} = \frac{p}{1-p} \)

Let’s look at the odds that are possible in our 2x2 table:

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>Healthy</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td></td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

Odds of disease among exposed = \( \frac{a}{b} = \frac{2}{8} = .25 \) ("cohort” study)

Odds of disease among non exposed = \( \frac{c}{d} = \frac{10}{290} = .0345 \) ("cohort study”)

Odds of exposure among diseased = \( \frac{a}{c} = \frac{2}{10} = .20 \) ("case-control study”)

Odds of exposure among healthy = \( \frac{b}{d} = \frac{8}{290} = .0276 \) ("case-control study”)

Odds ratio

In a **cohort study:**

\[
\text{OR} = \frac{\text{Odds disease among exposed}}{\text{Odds disease among non-exposed}} = \frac{a/b}{c/d} = \frac{ad}{bc}
\]

In a **case-control study:**

\[
\text{OR} = \frac{\text{Odds exposure among diseased}}{\text{Odds exposure among healthy}} = \frac{a/c}{b/d} = \frac{ad}{bc}
\]

Terrific!

The OR is the same, regardless of the study design, cohort (prospective) or case-control (retrospective)

**Example:** In our 2x2 table, a =2, b=8, c=10, and d=290 so the OR = 7.25. This is slightly larger than the value of the RR = 6.006.
Thus, there are advantages of the Odds Ratio, OR.

1. Many exposure disease relationships are described better using ratio measures of association rather than difference measures of association.

2. $\text{OR}_{\text{cohort study}} = \text{OR}_{\text{case-control study}}$

3. The OR is the appropriate measure of association in a case-control study.
   - Note that it is not possible to estimate an incidence of disease in a retrospective study. This is because we select our study persons based on their disease status.

4. When the disease is rare, $\text{OR}_{\text{case-control}} = \text{RR}$