Relative Risk and Odds Ratio

The **relative risk** (RR) is the probability that a member of an exposed group will develop a disease relative to the probability that a member of an unexposed group will develop that same disease:

\[
RR = \frac{P(\text{disease} | \text{exposed})}{P(\text{disease} | \text{unexposed})}
\]

If an event takes place with probability \(p\), the odds in favor of the event are \(\frac{1}{1-p}\) to 1. \(p = \frac{1}{2}\) implies 1 to 1 odds; \(p = \frac{2}{3}\) implies 2 to 1 odds.

In this class, the **odds ratio** (OR) is the odds of disease among exposed individuals divided by the odds of disease among unexposed.

\[
OR = \frac{P(\text{disease} | \text{exposed})}{P(\text{disease} | \text{unexposed})} = \frac{\frac{A}{A+B}}{\frac{C}{C+D}} = \frac{31/1628}{65/4540} = 1.33
\]

The collection of women who gave birth at a later age (\(\geq 25\)) are at increased risk for developing BC. Is this increase significant, or just chance? Would we expect to see this increase again if another sample of women was taken?

Suppose that among 100,000 women with negative mammograms, 20 will have BC diagnosed within 2 years; and among 100 women with positive mammograms, 10 will have BC diagnosed within 2 years.

<table>
<thead>
<tr>
<th>exposure or factor</th>
<th>disease or condition</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BC+</td>
<td>BC−</td>
</tr>
<tr>
<td>Mammogram+</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>Mammogram−</td>
<td>20</td>
<td>99,980</td>
</tr>
<tr>
<td>total</td>
<td>30</td>
<td>100,070</td>
</tr>
</tbody>
</table>

\[
RR = \frac{P(\text{disease} | \text{exposed})}{P(\text{disease} | \text{unexposed})} = \frac{\frac{A}{A+B}}{\frac{C}{C+D}} = \frac{31/1628}{65/4540} = 1.33
\]

Women with positive mammograms are at increased risk of developing BC. The RR here is very far from 1! We might conclude this is significant...how do we know for sure?

Disease is breast cancer (BC). A woman is considered to be exposed if she gave birth at or after the age of 25.

<table>
<thead>
<tr>
<th>exposure or factor</th>
<th>disease or condition</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BC+</td>
<td>BC−</td>
</tr>
<tr>
<td>Birth (\geq 25)</td>
<td>31</td>
<td>1597</td>
</tr>
<tr>
<td>Birth &lt; 25</td>
<td>65</td>
<td>4475</td>
</tr>
<tr>
<td>total</td>
<td>96</td>
<td>6072</td>
</tr>
</tbody>
</table>

\[
RR = \frac{P(\text{disease} | \text{exposed})}{P(\text{disease} | \text{unexposed})} = \frac{\frac{A}{A+B}}{\frac{C}{C+D}} = \frac{31/1628}{65/4540} = 1.33
\]

Odds and Odds Ratio

If an event takes place with probability \(p\), the odds in favor of the event are \(\frac{1}{1-p}\) to 1. \(p = \frac{1}{2}\) implies 1 to 1 odds; \(p = \frac{2}{3}\) implies 2 to 1 odds.

In this class, the **odds ratio** (OR) is the odds of disease among exposed individuals divided by the odds of disease among unexposed.

\[
OR = \frac{P(\text{disease} | \text{exposed})}{P(\text{disease} | \text{unexposed})} = \frac{(A)/(A+B)}{(C)/(C+D)} = \frac{31/1628}{65/4540} = 1.33
\]

Note that the OR is sometimes defined alternatively as

\[
OR_{\text{alt}} = \frac{P(\text{exposure} | \text{disease})}{P(\text{exposure} | \text{nondiseased})} = \frac{(A)/(A+B)}{(C)/(C+D)} = \frac{31/1628}{65/4540} = 1.33
\]

Note that these definitions are equivalent.
A random variable $X$ is a variable whose value is determined by the result of a random experiment. It can be thought of as a function, $X(\cdot)$ which assigns some real number to each outcome in $\Omega$.

Example: Consider the experiment of taking a single blood test to determine HIV status. Then outcomes are $\{HIV^+\}$ and $\{HIV^-\}$. Let the random variable $X$ denote the number of positive tests: $X(\omega) = 1$ if $\omega = HIV^+$ and $X(\omega) = 0$ if $\omega = HIV^-$. The random variable associates a real number with each outcome of the experiment.

A discrete random variable can assume only a finite or countable number of outcomes.

A continuous random variable can take on any value in some specified range.

Random variables are also often denoted by $X$, $Y$ and $Z$.

Examples of discrete Random Variables:

1. Experiment is surgery on two people. Outcomes are each is a success ($s$), the first is a success and the second is a failure ($f$), the second is a success and the first is a failure, each is a failure.
   
   $$X = \begin{cases} 
   2 & \text{if } (s,s) \\
   1 & \text{if } (s,f) \\
   1 & \text{if } (f,s) \\
   0 & \text{if } (f,f) 
   \end{cases}$$

2. Experiment is to observe the number of people that get tested for HIV in one week at a given clinic. Suppose 500 is the maximum possible number of tests given in a week. Then any non-negative integer less than or equal to 500 is a conceivable outcome.
   $X = \text{number of tests in a given week}$

3. Experiment is to record the number of places that a person has lived in his or her lifetime. Possible outcomes are $1, 2, 3, \ldots$
   $X = \text{number of places a person has lived}$

4. Experiment is to record the sex of a person. Outcomes are female or male.
   $$X = \begin{cases} 
   1 & \text{if f} \\
   0 & \text{if m} 
   \end{cases}$$
Experiment is to record a person’s sex. Possible outcomes are female (f) and male (m).

\[ X = \begin{cases} 
1 & \text{if f} \\
0 & \text{if m}
\end{cases} \]

Experiment is to record the number of places that a person has lived in his or her lifetime. Possible outcomes are 1, 2, 3, …, 10

\[ X = \text{number of places a person has lived.} \]

Probability Distributions

For a discrete random variable \( X \), a *probability distribution* is a function that assigns to any possible value \( x \) of \( X \) the probability \( P(X = x) \). Note that the random variable is denoted by uppercase \( X \) and lowercase \( x \) is used to represent possible values that \( X \) could take on.

Your book uses the terminology *probability distribution* a bit loosely; *probability mass function* is also often used to emphasize the discrete case discussed here.

Example: Consider again the experiment of taking a single blood test to determine HIV status. Let the random variable \( X \) denote the number of positive tests.

\[ X = \begin{cases} 
1 & \text{if } \text{HIV}^+ \\
0 & \text{if } \text{HIV}^-
\end{cases} \]

If we knew that the prevalence of HIV was 0.11, then

\[ P(X = 1) = 0.11 \text{ and } P(X = 0) = 0.89 \]

These two equations completely describe the probability distribution of the discrete (dichotomous) random variable \( X \).

Bernoulli Distribution

A random experiment with outcomes that can be classified into two categories (disease positive or negative, success or failure, absent or present, …) is called a *Bernoulli trial*. Oftentimes, in a Bernoulli trial, a random variable \( X \) (called a Bernoulli random variable) is defined to be 1 if the Bernoulli trial results in success and 0 if the same trial results in failure.
Example of a Bernoulli Trial

Let $Y$ be a random variable that represents smoking status. $Y = 1$ if person is a smoker and $Y = 0$ if person is a non-smoker. Suppose we know that 29% of adults in the U.S. are smokers. Then $P(Y = 1) = 0.29$ and $P(Y = 0) = 0.71$.

Again, these two equations completely describe the probability distribution function of the Bernoulli random variable $Y$.

Suppose the experiment is to select two individuals and record their smoking status. Let $X$ denote the number of smokers in the pair.

Then $X = 0; 1; 2$ are possible outcomes.

What is the probability distribution function of $X$?

Suppose the experiment is to select three individuals and record their smoking status. $Y_i$ denotes the smoking status of the $i^{th}$ person ($i = 1, 2, 3$). As before, $Y_i$ are independent.

Let $X$ denote the total number of smokers. Then $X = 0, 1, 2, 3$ are possible outcomes.

What is the pdf of $X$?

<table>
<thead>
<tr>
<th>Number of Smokers ($x$)</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>$P(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$(1 - p) \cdot (1 - p) \cdot (1 - p)$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$p \cdot (1 - p) \cdot (1 - p)$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$(1 - p) \cdot p \cdot (1 - p)$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>$(1 - p) \cdot (1 - p) \cdot p$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>$p \cdot p \cdot (1 - p)$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>$p \cdot (1 - p) \cdot p$</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>$(1 - p) \cdot p \cdot p$</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$p \cdot p \cdot p$</td>
</tr>
</tbody>
</table>

Recall that the probability that an individual is a smoker is 0.29 ($P(Y_i = 1) = 0.29$).

$P(X = 0) = (1 - p)^3 = (0.71)^3 = 0.358$

$P(X = 1) = 3 \cdot p \cdot (1 - p)^2 = 3 \cdot (0.29) \cdot (0.71)^2 = 0.439$

$P(X = 2) = 3 \cdot p^2 \cdot (1 - p) = 3 \cdot (0.29)^2 \cdot (0.71) = 0.179$

$P(X = 3) = p^3 = (0.29)^3 = 0.024$

Suppose the experiment is to select four individuals and record their smoking status. $Y_i$ denotes the smoking status of the $i^{th}$ person ($i = 1, 2, 3, 4$). As before, $Y_i$ are independent.

Let $X$ denote the total number of smokers. Then $X = 0, 1, 2, 3, 4$ are possible outcomes.

What is the pdf of $X$?
Discrete Probability Distributions

Suppose $X$ is a discrete random variable taking on the values $x_1, x_2, \ldots, x_n$. Then the function $f_X(x)$ defined by

$$f_X(x) = \begin{cases} P(X = x) & \text{if } x = x_1, x_2, \ldots, x_n \\ 0 & \text{if } x \neq x_1, x_2, \ldots, x_n \end{cases}$$

Two properties of discrete probability distribution functions

For $i = 1, 2, \ldots, n$,

$$0 \leq f_X(x_i) \leq 1,$$

$$\sum_{i=1}^{n} f_X(x_i) = 1,$$

Note that the random variable is denoted by uppercase $X$ and lowercase $x$ is used to represent possible values that $X$ could take on.

Your book uses the terminology probability distribution a bit loosely; probability mass function is also often used to emphasize the discrete case discussed here.

Binomial Distribution

The probability distributions of $X$ in the last three examples are special cases of the Binomial distribution. Specifically, if $X$ represents the number of successes in $n$ independent Bernoulli trials (each with probability $p$ of success), then the probability distribution function of $X$ is the Binomial distribution function with parameters $p$ and $n$.

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where

$$\binom{n}{x} = \frac{n!}{(n-x)!x!}$$

is the binomial coefficient

Note that $x! = x \cdot (x - 1) \cdot (x - 2) \cdots 1$ and $0! = 1$.

$P(X = x)$ is also written as $f_X(x)$ or $f_X(x; n, p)$. 