Risk Assessment with Contingency Table and Simpson's Paradox

Risk factor is a variable that is thought to be related to some outcome variable, and it may be a suspected cause of some specific state of this outcome variable.

	Risk Factor		
Outcome	Exposed	Unexposed	Total
Disease	*a*	*b*	*a* + *b*
No Disease	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*n*

Relative risk, *RR*, is a standard measure of strength of the exposure effect and is defined to be

RR = *P* [disease | exposed] / *P* [disease | unexposed]

and its estimate $\hat{RR} = \dfrac{a/(a+c)}{b/(b+d)} = \dfrac{a(b+d)}{b(a+c)}$

When *a* and *b* are small relative to the values of *c* and *d* Odds Ratio is a good estimate of the relative risk.

Example: Suppose we conducted a prospective cohort study to investigate the effect of aspirin on heart disease. A group of patients who are at risk for a heart attack are randomly assigned to either a placebo or aspirin. At the end of one year, the number of patients suffering a heart attack is recorded.

	Group		
Heart Disease	Placebo	Aspirin	Total
Yes +	20	15	35
No –	80	135	215
Total	100	150	250

Relative risk = (20/100)/(15/150) = .2/.1 = 2
(The risk of a heart attack for people on placebo is twice that of people on aspirin.)

The Odds Ratio

A method for estimating the effect of the exposure effect.

	Risk Factor		
Outcome	Exposed	Unexposed	Total
Disease	a	b	$a + b$
No Disease	c	d	$c + d$
Total	$a + c$	$b + d$	n

($a + b + c + d = n$)

The odds of getting the disease, given that one has the **exposure**, are

$O_+ = P$ [disease | exposed] / P [no disease | exposed] ,
can be estimated by $[a/(a+c)]/[c/(a+c)]$ or a / c

The odds of getting the disease, given that one has **no exposure**, are

$O_- = P$ [disease | unexposed] / P [no disease | unexposed] ,
can be estimated by $[b/(b+d)]/[d/(b+d)]$ or b / d

The **odds ratio**, OR, is then defined to be $\frac{O_+}{O_-}$, and its estimate

$$\hat{OR} = \frac{[a/(a+c)]/[c/(a+c)]}{[b/(b+d)]/[d/(b+d)]} = \boxed{\frac{a/c}{b/d} = \frac{ad}{bc}} \approx \frac{a(b+d)}{b(a+c)} = \hat{RR}$$

Example: Suppose we want to determine if people with a rare brain tumor are more likely to have been exposed to benzene than people without a brain tumor. One experimental design used to answer this question. First, we start with cases, people with a disease or condition (brain tumor) and find people who are as similar as possible but who do not have brain tumors. Those people are called controls.

	Exposure		
Outcome	Yes	No	Total
Case	50	100	150
Control	20	130	150
Total	70	230	300

Odds ratio = (50/20) / (100/130) = (50x130) / (20x100) = 3.25
(Is the odds ratio different from 1?)

Is there a relation between workers from different regions in the world and making mistake in the production line?

Story

A manufacture hired a team of researchers to study the cause of certain mistake that often occurred in the production line. The researchers suspected that there is a difference between workers in their facilities in different regions of the world. They tested 1314 individuals some are from their plants in Europe and some are from Asia.

Results

The following tables summarize the results of the experiment:

Table 1: Relationship between type of workers and outcome from observing 1314 individuals.

Outcome	*Location*	
	Europe	**Asia**
Mistake	139	230
NoMistake	443	502

Table 2: Outcome for 1314 individuals categorized by age and type of workers.

Outcome	**Age Group (Years)**					
	18-44		**44-64**		**above 64**	
	Europe	**Asia**	**Europe**	**Asia**	**Europe**	**Asia**
Mistake	19	13	78	52	42	165
NoMistake	269	327	167	147	7	28

Questions for Smoking and Survival

I.

a) Use the information in Table 1 to calculate the following proportions. Of the workers who made mistake, what proportion were from Europe? Of the workers who were did not make mistake, what proportion were from Asia?

b) Calculate the following proportions using the data in Table 2. Of those workers who were 18-44 years old at the time of the study and who made mistake, what proportion were Europeans? What were the similar proportions for the 44-64 and above-64 age groups?

II.

a) From the data in Table 1, compute the overall percentage of European workers who made mistake and the percentage of Asian workers who made mistake. Also compute the relative risk of making mistake for European versus Asian workers? Do the data from this table suggest that European or Asian workers less likely to make mistake? Comment.

b) Using the information in Table 2, compute the percentage of European and Asian workers in each of the three age groups who made mistake. Also, compute the relative risk of making mistake for European versus Asian workers for all three different age groups? Do the data from this table suggest that European or Asian workers less likely to make mistake? Comment.

c) Compare your results from Part a and Part b. How can you explain any apparent contradictions?