Comparing Proportions in two-by-two tables:

This section presents measures for comparing groups on binary responses.

The following table is from a report on the relationship between aspirin use and heart attacks.

A five-year randomized study was conduct to test "whether regular intake of aspirin reduces mortality from heart attack disease". Every other day, the participants took either one aspirin tablet or a placebo. The study was "blind" (the participants did not know which type of pill they were taking).

	Heart attack		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Here we want to compare the success probabilities (Heart attack decease) of two groups. For this purpose we have three methods:

(1) Difference of proportion:

We we treat groups as independent binomial sample.

In which $n_1 = 11,034$, 189 participants who took placebo suffered from Heart attack

and their proportion is $\frac{189}{11,034} = 0.0171$.

And $n_2 = 11,037$, 104 participants who took aspirin are suffered from heart attack and their proportion is $\frac{104}{11,037}$ $= 0.0094$.

The sample difference of proportions is

$$P_1 - P_2$$
$$= 0.0171 - 0.0094$$
$$= 0.0077$$ (taking aspirin appears to result in a diminished risk of heart attack).

The difference $P_1 - P_2$ always fall between -1 and $+1$.

we compare it with zero.

If $P_1 - P_2 > 0$ (P_1 has more risk or P_2 has less risk)

If $P_1 - P_2 < 0$ (P_1 has less risk or P_2 has more risk)

If $P_1 - P_2 = 0$ (P_1 & P_2 has risk effect)

⇓

It means the response is independent of the group classification.

$$C.I = (P_1 - P_2) - Z_{\alpha/2}(SE) < \pi_1 - \pi_2 < (P_1 - P_2) + Z_{\alpha/2}(SE)$$

$$S.E = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

(2) Relative Risk / Risk Ratio

Another way to compare proportions is

$$\text{Relative Risk} = P_1/P_2 \quad \text{where } P_2 \neq 0$$

compare it by 1

If $P_1/P_2 > 1$ (P_1 has more risk)

$P_1/P_2 < 1$ (P_2 has more risk)

$P_1/P_2 = 1$ (P_1 & P_2 has equal risk)

$$\text{Relative Risk} = \frac{0.0171}{0.0094} = 1.82$$

(The sample proportion of heart attack cases was 82% higher for the group taking placebo.)

The sample difference of proportions of 0.0077 makes it seem as if the two groups differ by a trivial amount. By contrast, the relative risk of 1.82 shows that the difference may have important implications. {Using the difference of proportions alone to compare two groups can be misleading when the proportions are both close to zero.}

The Sampling distribution of the sample relative risk is highly skewed unless the sample sizes are quite large. Because of this, the C.I is rather complex

A large sample confidence interval for the log of the relative risk is

$$\log(p_1/p_2) \pm Z_{\alpha/2}\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}$$

Antilogs of the endpoints yeild an interval for the true relative risk.

The 95% C.I for the true relative risk π_1/π_2 is (1.43, 2.30).

odd+1

3-Relative Odds / Odds Ratio

Odds = $P_1/1-P_1$ or $P_2/1-P_2$

Odds of having heart attack for those persons who take placebo

$$odds_1 = \frac{P_1}{1-P_1} = \frac{0.0171}{1-0.0171} = \frac{0.0174}{0.89}$$

$= 0.0174$ (1.74% yes outcome for every 100 no outcomes)

odds of heart attack for those persons who take aspirine.

$$Odds_2 = \frac{P_2}{1-P_2} = \frac{0.0094}{1-0.0094} = \frac{0.0094}{0.9906}$$

$= 0.0094$ (0.95 yes outcome for every 100 no outcomes).

$$\text{Odds Ratio} = \theta = \frac{odd_1}{odd_2} = \frac{P_1/1-P_1}{P_2/1-P_2} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}}$$

$$= \pi_{11}\pi_{22}/\pi_{12}\pi_{21}$$

$$= \frac{0.0174}{0.0094} = 1.85$$

The odds value of placebo is 1.85 times or 85% higher than the aspirine.

odd Ratio = 1 (odds1 is same as odd2)

odd Ratio < 1 (odds2 is higher than odd1)

odd Ratio > 1 (odds1 is higher than odds2)

Sample odd ratio

$$\hat{\theta} = \frac{p_1/1-p_1}{p_2/1-p_2} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}\,n_{22}}{n_{12}\,n_{21}}.$$

⇒ The odds ratio does not change value when the table oriented reverses so that the rows become the columns and columns become the rows. The same value occurs when we treat the columns as the response variable and rows as the explanatory variable & vise visa.

Thus it is unnecessary to identify one classification as a response variable in order to estimate odds ratio. By contrast, the relative risk requires this.

So the odds ratio can be defined using joint probabilities as

$$\text{Odds Ratio} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\,\pi_{22}}{\pi_{12}\,\pi_{21}}$$

The odd ration is also called the cross-product ratio. because it equals the ratio of the products $\pi_{11}\,\pi_{22}$ and $\pi_{12}\,\pi_{21}$ of cell probabilities from diagonally opposite cells.

⇒ The sample odds ratio equal 0 or ∞ if any $n_{ij} = 0$

Inference for odds Ratios and log odds Ratios.

Unless the sample size is extremely large, the sampling distribution of the odds ratio is highly skewed.

Because of this skewness statistical inference for the odds ration uses an alternative but equivalent measure ⇒ Its natural logarithm (log odd ratio).

$$\log \text{odd ratio} = 0 \quad (\text{independence})$$
$$< 0 \quad (\text{less})$$
$$> 0 \quad (\text{high})$$

The log odds ration is symmetric about zero, in the sense that reversing rows or reversing columns changes its sign.

The sample log odd ratio ($\log \hat{\theta}$) has a less skewed sampling distribution that is bell-shaped. Its approximating normal distribution has a mean of $\log \theta$ and a standard error of

$$S.E = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

The S.E decreases as the cell counts increase.

Because sampling distribution is closer to normality for $\log(\hat{\theta})$ than $\hat{\theta}$ so it is better to construct confidence intervals for $\log \theta$. Transform back (that is, take antilogs, using the exponential function) to form a confidence interval for θ.

A large-sample confidence interval for $\log\theta$ is

$$CI(\log\theta) = \log\hat{\theta} \pm Z_{\alpha/2}(S.E)$$

Exponentiating endpoints of this confidence interval yields one for θ.

Relationship b/w odd Ratio and Relative Risk:

Relative Risk = 1.83 means that p_1 is 1.83 time higher than p_2 or p_1 is 83% higher than p_2

Where the odds Ration = 1.83 means that odds value $p_1/1-p_1$ is 1.83 times higher than the odds value $p_2/1-p_2$ or $p_1/1-p_1$ is 83% higher than $p_2/1-p_2$

$$\text{odds ratio} = \frac{p_1/1-p_1}{p_2/1-p_2} = \text{Relative Risk} \times \left(\frac{1-p_2}{1-p_1}\right)$$

When p_1 and p_2 are both close to zero, the fraction in the last term of this expression equals approximately 1. The odds ratio and Relative Risk then take similar values.

This relationship between the odd ratio and the relative risk is useful for some data sets direct

estimation of the relative risk is not possible, yet one can estimate the odd ratio and use it to approximate the relative risk.

Independence:

Two variables are said to be statistically independent if the population conditional distribution of Y are identical at each level of X.

e.g.: The probability of having cancer is equal for both females and males.

When both variables are response variables, we can discribe their relationship using their joint distribution or the conditional distribution of Y given X, or the conditional distribution of X given Y. Statistically independence is equivalently the property that all joint probabilities equal the product of their marginal probabilities.

$$\pi_{ij} = \pi_{i+} \cdot \pi_{+j} \quad \text{for} \quad i = 1, \cdots I \quad j = 1, \cdots, J$$

Thus the probability that X falls in row i and Y falls in column j is the product of the probability that X falls in row i with the probability that Y falls in column j.