Bivariate data analysis

Association between pair of variables

We consider here how the association between pair of variables can be measured using appropriate statistics.

For more than two variables it is necessary to resort to more elaborate techniques that often imply to adopt more articulated mathematical models that are beyond the scope of this course.

We distinguish between two cases:

- The two variables are numerical (in this case we talk of "correlation" measures)
- The two variables are categorical (in this case we talk of "association" measures)

Note: association does not mean causation. Although when things are related by a casual relationship, we might expect some (numerically evident) association, one variable being a cause of another one would require more substantive reasoning and rationale.

Correlation

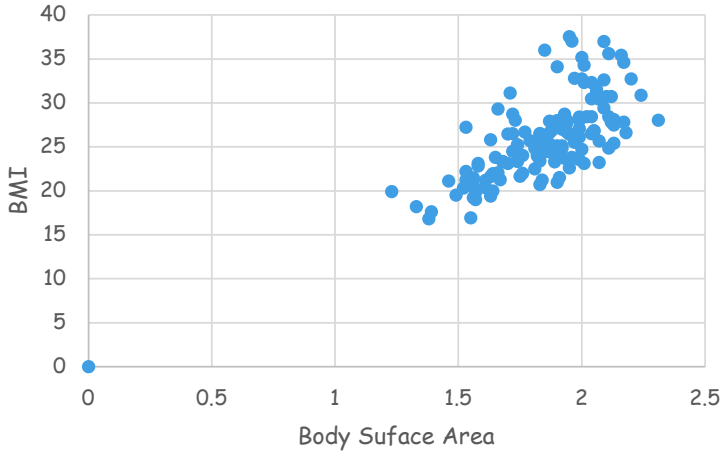
In statistics, the meaning of the term “correlation” does not differ from that given to the word in everyday life: it indicates a mutual relationship between phenomena.

Specifically, the term refers to the relationship between two quantitative variables. Thus, it makes sense to talk about the correlation between Age and Body Surface Area or between Body Surface Area and BMI and so on.

Scatter plots

Let X and Y be two numerical variables and (x_j, y_i) be the values that X and Y take on the i –th sample unit simultaneously.

A scatter plot is a graphical representation that displays the relationship between X and Y. It involves plotting individual data points on a two-dimensional Cartesian coordinate system, with one variable represented on the horizontal axis and the other variable represented on the vertical axis.



Each point represents the value of the two variables for a single observation.

If some causal dependency can be assumed between X and Y, the x-axis typically represents the independent (*explanatory*) variable, while the y-axis represents the dependent (*response*) variable.

Covariance and correlation

Two variables are *positively correlated* if they tend to increase or decrease jointly.

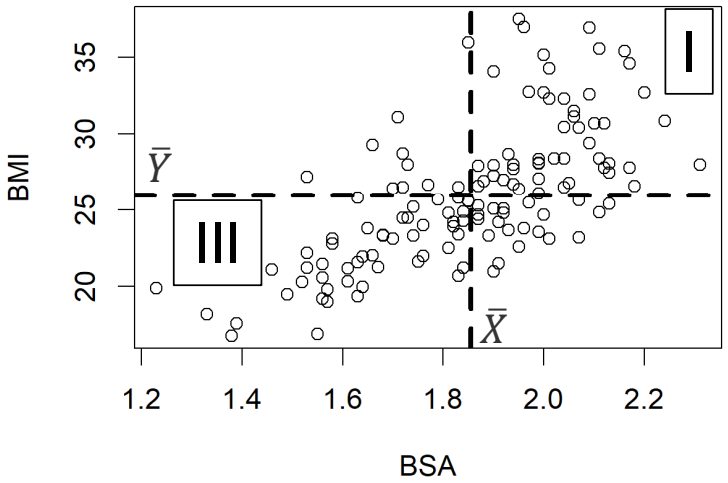
Two variables are *negatively correlated* if one increases when the other decreases.

In quadrant I we get $x_i - \bar{X} > 0$ and $y_i - \bar{Y} > 0$

In quadrant III we get $x_i - \bar{X} < 0$ and $y_i - \bar{Y} < 0$

Hence $(x_i - \bar{X})(y_i - \bar{Y}) > 0$

In quadrants II and IV we get $(x_i - \bar{X})(y_i - \bar{Y}) < 0$



Therefore, if X and Y are positively correlated, positive terms exceed negative ones and

$$\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) > 0$$

Covariance and correlation

The *Pearson's correlation coefficient* is defined by

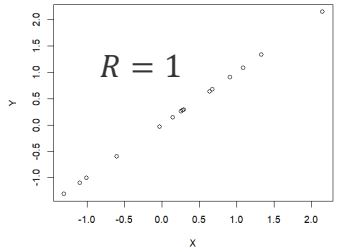
$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2 \ \sum_{i=1}^{n}(y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{\tilde{S}_X^2 \tilde{S}_Y^2}}$$

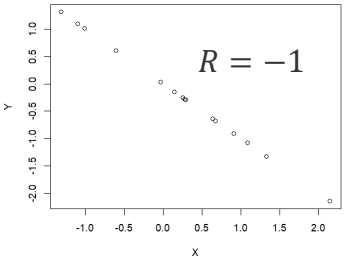
where $S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$ is known as the *covariance* between X and Y.

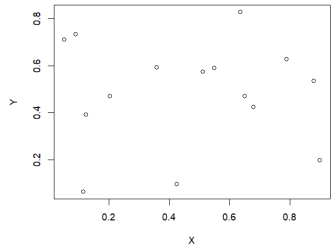
Property $-1 \leq R \leq 1$

- $R = 1$ exact positive linear correlation
- $R = -1$ exact negative linear correlation
- $R = 0$ no linear correlation

Intermediate R values can be interpreted accordingly.







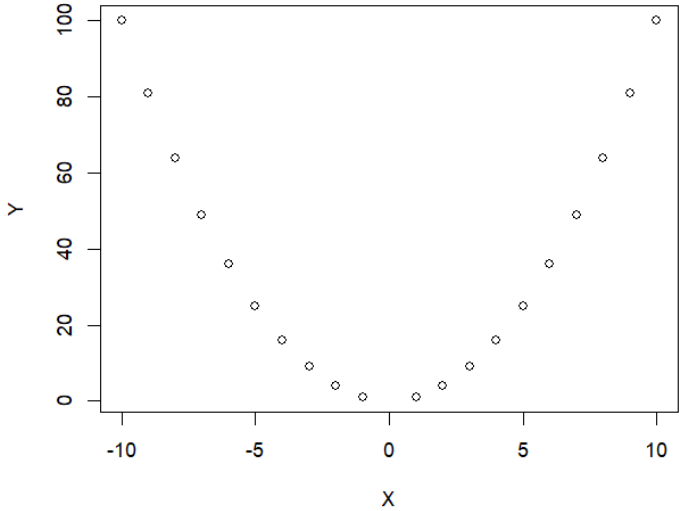
$R \approx 0$

Covariance and correlation cont'd

Note that the values of R do not change if the observed values of X and Y are multiplied by a constant (i.e. the unit of measurement is changed) or if a constant is added to them.

The Pearson's correlation coefficient measures only the linear relationship between X and Y and may not be sensitive to other kind of dependencies.

In this case the two variables are perfectly associated although $R = 0$. The relationship is not linear.



The interpretation of a correlation coefficient depends on the context. A correlation of 0.7 may be low if one is verifying a physical law using high-quality instruments, but may be regarded as very high in social sciences, where there may be a greater contribution from complicating factors.

Correlation between Ranks

Let $x_{(1)}, \dots, x_{(n)}$ be the values of X arranged in ascending order and $r_{x1}, \dots, r_{xn}$ their ranks.
Let $y_{(1)}, \dots, y_{(n)}$ be the values of Y arranged in ascending order and $r_{y1}, \dots, r_{yn}$ their ranks.

The *Spearman's rank correlation coefficient*, R_S is the Pearson's coefficient calculated using the ranks of the data of the two variables.

It can be proved that it holds true $R_s = 1 - \frac{6 \sum_{i=1}^{n} (r_{xi} - r_{yi})^2}{n(n^2-1)}$.

A perfect Spearman correlation of +1 or –1 occurs when each of the variables is a perfect monotone function of the other.

Correlation between Ranks cont’d

Spearman's correlation evaluates whether the relationship between two variables increases or decreases together, without necessarily being linear. Therefore, it is particularly useful when the scatter plots point out non-linear relationships or when the data is ordinal rather than continuous.

A tie occur when two (or more) data points have the same value.

When there are tied ranks a correction factor is applied to R_s to adjust for the effect of ties.

A common approach is to assign each tied value the average of the ranks it would have received had there been no ties. Then, the standard formula for Spearman's rank correlation coefficient is applied using these adjusted ranks.

Association between categorical variables

Let X and Y be two discrete or categorical variables.

Assume that X takes s different values $\tilde{x}_1, \ldots, \tilde{x}_s$, and Y takes t different values $\tilde{y}_1, \ldots, \tilde{y}_t$

Let $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, be the values that X and Y take on each unit of the data set.

Assume that n_{ij} is the number of units of the sample for which $X = \tilde{x}_i$ and $Y = \tilde{y}_j$

n_{ij} is called the **joint frequency** of the pair $(\tilde{x}_i, \tilde{y}_j)$.

	$\tilde{y}_1$		$\tilde{y}_j$		$\tilde{y}_t$	Marginal X
$\tilde{x}_1$	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{x}_i$	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{it}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{x}_s$	n_{s1}	$\cdots$	n_{sj}	$\cdots$	n_{st}	$n_{s.}$
Marginal Y	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.t}$	n

The joint frequency distribution is a rectangular $s \times t$ table that lists all the pairs of values/classes of X and Y along with their corresponding frequencies.

This table is known as **Contingency Table**

Bivariate frequency distributions

Frequency Distribution

	ECOG PS			
Treatment	0	1	2	Randomisation distribution
A	47	24	3	74
B	48	10	6	64
ECOG PS distribution	95	34	9	138

Relative Frequency Distribution

	ECOG PS			
Treatment	0	1	2	Distribution of Randomisation
A	0.34	0.17	0.02	0.54
B	0.35	0.07	0.04	0.46
Distribution of ECOG PS	0.69	0.25	0.07	**1**

Variable Description

ECOG PS=0:

ECOG PS=1:

ECOG PS=2:

ECOG PS is an ordered variables

Randomisation is measured on a nominal scale

Comparing frequency distributions

Consider the relative frequency distribution (rfd) of a given variable (say Y) within a particular subgroup (say $X = \tilde{x}_i$)

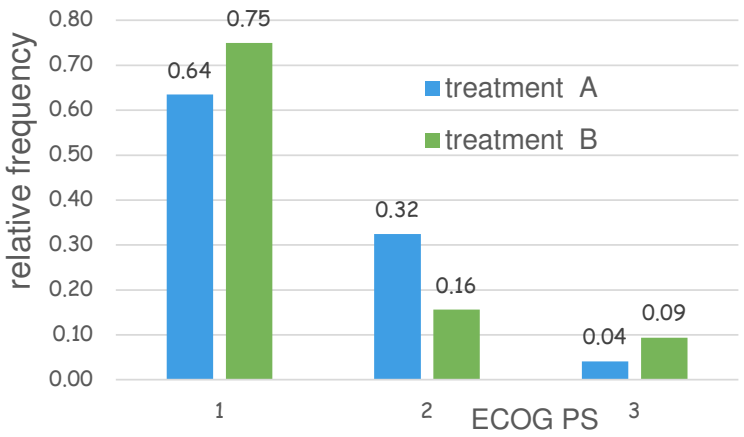
$$p_{j|i} = \frac{n_{ij}}{n_{i.}}$$

(*conditional frequency distribution*)

where $n_{i.} = \sum_{j=1}^{t} n_{ij}$ is the overall number of units in category i.

Bar charts corresponding to different rfd allow one to compare different distribution shapes across different groups.

	$\tilde{y}_1$	⋯	$\tilde{y}_j$	⋯	$\tilde{y}_t$	
$\tilde{x}_1$	$p_{1\|1}$	⋯	$p_{j\|1}$	⋯	$p_{t\|1}$	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\tilde{x}_i$	$p_{1\|i}$	⋯	$p_{j\|i}$	⋯	$p_{t\|i}$	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\tilde{x}_s$	$p_{1\|s}$	⋯	$p_{j\|s}$	⋯	$p_{t\|s}$	1



Comparing frequency distributions

Consider the conditional frequency distribution of Y given $X = \tilde{x}_i$

$p_{j|i} = \frac{n_{ij}}{n_{i\cdot}} \quad j = 1, \cdots, t$

$n_{i\cdot} = \sum_{j=1}^{t} n_{ij}$: overall number of units in category i

	$\tilde{y}_1$		$\tilde{y}_j$		$\tilde{y}_t$	
$\tilde{x}_1$	$p_{1\|1}$	$\cdots$	$p_{j\|1}$	$\cdots$	$p_{t\|1}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{x}_i$	$p_{1\|i}$	$\cdots$	$p_{j\|i}$	$\cdots$	$p_{t\|i}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{x}_s$	$p_{1\|s}$	$\cdots$	$p_{j\|s}$	$\cdots$	$p_{t\|s}$	1

The variables of the contingency table are statistically independent if the conditional relative frequency distributions of one of them are all equal.

In formulas $p_{j|i} = p_{j|k}$ for all $j = 1, \cdots, t$ and any choice of i and k.

The two variables are associate otherwise.

Comparing frequency distributions

Consider the conditional frequency distribution of Y given $X = \tilde{x}_i$

$p_{j|i} = \frac{n_{ij}}{n_{i.}} \quad j = 1, \cdots, t$

$n_{i.} = \sum_{j=1}^{t} n_{ij}$: overall number of units in category i

	$\tilde{y}_1$		$\tilde{y}_j$		$\tilde{y}_t$	
$\tilde{x}_1$	$p_{1\|1}$	$\cdots$	$p_{j\|1}$	$\cdots$	$p_{t\|1}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{x}_i$	$p_{1\|i}$	$\cdots$	$p_{j\|i}$	$\cdots$	$p_{t\|i}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{x}_s$	$p_{1\|s}$	$\cdots$	$p_{j\|s}$	$\cdots$	$p_{t\|s}$	1

If the two variables are statistically independent then $p_{j|i} = p_{j|k} = g(j)$ for all $j = 1, \cdots, t$ i.e. the conditional frequency distribution does not depend on the conditioning variable

If $n_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ then $p_{j|i} = \frac{n_{ij}}{n_{i.}} = \frac{n_{i.} \times n_{.j}}{n\, n_{i.}} = \frac{n_{j.}}{n} = f_{j.} \quad \forall j$ i.e. the c.f.d. does not depend on the conditioning variable.

Vice versa if $p_{j|i} = \frac{n_{j.}}{n} = f_{j.}$ for all $j = 1, \cdots, t$ then $n_{ij} = \frac{n_{i.} \times n_{.j}}{n}$

Equivalent condition for independence

$$n_{ij} = \frac{n_{i.} \times n_{.j}}{n} \; \forall i, j$$

$$p_{j|i} = f_{j.} \quad \forall i, j$$

Measures of association for qualitative variables

Indicate by $\hat{n}_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ the joint frequency of cell i, j under independence of Y and X.

A measure of how far the observed data deviate from the condition of independence can be obtained by comparing the observed joint frequency n_{ij} to $\hat{n}_{ij}$.

	$\tilde{y}_1$	⋯	$\tilde{y}_j$	⋯	$\tilde{y}_t$	Marginal X
$\tilde{x}_1$	n_{11}	⋯	n_{1j}	⋯	n_{1t}	$n_{1.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\tilde{x}_i$	n_{i1}	⋯	n_{ij}	⋯	n_{it}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\tilde{x}_s$	n_{s1}	⋯	n_{sj}	⋯	n_{st}	$n_{s.}$
Marginal Y	$n_{.1}$	⋯	$n_{.j}$	⋯	$n_{.t}$	n

The usual statistics adopted for this is the X^2 Pearson's statistics

$$X^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(\hat{n}_{ij} - n_{ij})^2}{\hat{n}_{ij}}$$

Note X^2 here is not the square of X ...unfortunate notation!!

X^2 is symmetric i.e. its value does not depend on which variable is defined as Y and X;

$X^2 = 0$ if the independence assumption holds true;

$X^2 > 0$ and it is the greater the more Y and X tend to be associated.

Further insights in comparing proportions: difference in proportions

Assume that X and Y take two values or modality say 1 (success) 0 (failure).

In this case we have a 2×2 contingency table.

$X \backslash Y$	0	1	
0	n_{11}	n_{12}	n_{2*}
1	n_{21}	n_{22}	n_{2*}

We can compare the difference in the frequency of success of Y across the X values using differences between conditional probs

$$p_{2|1} - p_{2|2}$$

$X \backslash Y$	0	1	
0	$p_{1\|1}$	$p_{2\|1}$	1
1	$p_{1\|2}$	$p_{2\|2}$	1

i.e. the greater (in absolute value) the difference, the higher the frequency of successes of the Y variable under the X =0 regime

i.e. the difference measures the degree of association between Y and X.

If an asymmetric dependency can be assumed between X and Y it measures the degree of dependency of Y on X.

In this case Y is called the response variable and X the explanatory variable

Further insights in comparing proportions: difference in proportions

Assume that X and Y take two values or modality say 1 (success) 0 (failure) and use the difference $p_{2|1} - p_{2|2}$ to compare proportions.

$X \backslash Y$	0	1	
0	$p_{1\|1}$	$p_{2\|1}$	1
1	$p_{1\|2}$	$p_{2\|2}$	1

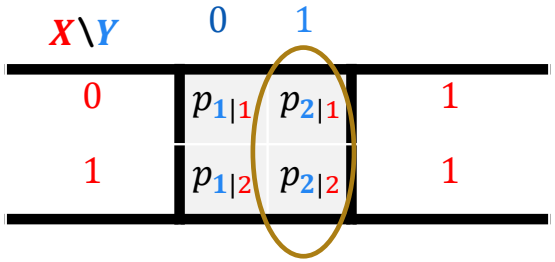
A value of the difference of fixed size may have greater importance when both, $p_{2|1}$ and $p_{2|2}$ are close to 0 or 1 than when they are not.

Example For a study comparing two treatments on the proportion of subjects who die, the difference between 0.010 and 0.001 may be more noteworthy than the difference between 0.410 and 0.401, even though both are 0.009

In this circumstances it is more appropriate to use *ratios* instead.

Further insights in comparing proportions: relative risk

Assume that X and Y take two values or modality say 1 (success) 0 (failure) and use the difference $p_{2|1} - p_{2|2}$ to compare proportions.



Relative risk: $p_{2|1}/p_{2|2}$

A relative risk of 1.0 corresponds to independence.

Example. For a study comparing two treatments on the proportion of subjects who die, the relative risk between 0.010 and 0.001 is 0.01/0.001 = 10.0 (i.e. huge difference in risk of death) whereas the relative risk between 0.410 and 0.401 il 0.410/0.401 = 1.02 (the difference is negligible).

Note: Comparing the rows on the second response category gives a different relative risk

Further insights in comparing proportions: odd and odds ratio

Let assume that π is the relative frequency of success.

Definition of ***odds*** $\Omega = \pi/(1-\pi)$

Odds is nonnegative and $\Omega > 1$ when a success is more frequent than a failure

Example. When $\pi = 0.75$ then $\Omega = 0.75/0.25$ i.e. a success is three times as likely as a failure, and we expect about three successes for every one failure. Vice versa when $\Omega = 1/3$, a failure is three times as likely as a success.

Consider again the case of a 2×2 contingency table that cross classifies X and Y.

$X \backslash Y$	0	1	
0	$p_{1\|1}$	$p_{2\|1}$	1
1	$p_{1\|2}$	$p_{2\|2}$	1

The odds (of $Y = 1$) for $X = 0$ is $\Omega_1 = p_{2|1}/(1-p_{2|1}) = p_{2|1}/p_{1|1}$

The odds (of $Y = 1$) for $X = 1$ is $\Omega_2 = p_{2|2}/(1-p_{2|2}) = p_{2|2}/p_{1|2}$

Further insights in comparing proportions: odd and odds ratio

2×2 contingency table that cross classifies X and Y.

The odds for $X = 0$ is $\Omega_1 = p_{2|1}/(1 - p_{2|1}) = p_{2|1}/p_{1|1}$

The odds for $X = 1$ is $\Omega_2 = p_{2|2}/(1 - p_{2|2}) = p_{2|2}/p_{1|2}$

where $p_{j|i} = \frac{n_{ij}}{n_{i\cdot}} = \frac{f_{ij}}{f_{i\cdot}}$

$X \backslash Y$	0	1	
0	$p_{1\|1}$	$p_{2\|1}$	1
1	$p_{1\|2}$	$p_{2\|2}$	1

Definition

The **odds ratio** is given by $\theta = \frac{\Omega_1}{\Omega_2} = \frac{p_{2|1}/(1-p_{2|1})}{p_{2|2}/(1-p_{2|2})} = \frac{f_{11}f_{22}}{f_{12}f_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$

$X \backslash Y$	0	1	
0	n_{11}	n_{12}	n_{2*}
1	n_{21}	n_{22}	n_{2*}

- The odds ratio is always nonnegative
- If one cell is empty $\theta = 0$ or $\theta = +\infty$
- When the order of the rows (columns) is reversed, the new θ is the inverse of the original value.
- The odds ratio does not change value when the orientation of the table reverses i.e. the rows become the columns and the columns become the rows ➔ It is unnecessary to identify one classification as the response variable in order to use θ

Further insights in comparing proportions: odd and odds ratio

The odds for $X = 0$ is $\Omega_1 = p_{2|1}/(1 - p_{2|1}) = p_{2|1}/p_{1|1}$

The odds for $X = 1$ is $\Omega_2 = p_{2|2}/(1 - p_{2|2}) = p_{2|2}/p_{1|2}$

The odds ratio is given by $\theta = \frac{\Omega_1}{\Omega_2} = \frac{p_{2|1}/(1-p_{2|1})}{p_{2|2}/(1-p_{2|2})} = \frac{f_{11}f_{22}}{f_{12}f_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$

$X \backslash Y$	0	1	
0	$p_{1\|1}$	$p_{2\|1}$	1
1	$p_{1\|2}$	$p_{2\|2}$	1

- If $\theta = 1$ ➔ $\Omega_1 = \Omega_2$ i.e. $\frac{p_{2|1}}{p_{1|1}} = \frac{p_{2|2}}{p_{1|2}}$. This occurs if $p_{2|1} = p_{2|2}$ (hence $p_{1|1} = p_{1|2}$) therefore X and Y are independent
- When $1 < \theta < +\infty$ ➔ $\Omega_1 > \Omega_2$ ➔ $p_{2|1} > p_{2|2}$ i.e. subjects in row 1 ($X = 0$) are more likely to have a success than are subjects in row 2 ($X = 1$). For instance, when $\theta = 4$ the odds of success in row 1 are four times the odds in row 2. (This does not mean that $p_{2|1} = 4p_{2|2}$, this is the relative risk =4!!)
- When $0 < \theta < 1$ ➔ $\Omega_1 < \Omega_2$ ➔ $p_{2|1} < p_{2|2}$ i.e. subjects in row 1 ($X = 0$) are less likely to have a success than are subjects in row 2 ($X = 1$).

Further insights in comparing proportions: odd and odds ratio

The odds for $X = 0$ is $\Omega_1 = p_{2|1}/(1 - p_{2|1}) = p_{2|1}/p_{1|1}$

The odds for $X = 1$ is $\Omega_2 = p_{2|2}/(1 - p_{2|2}) = p_{2|2}/p_{1|2}$

The odds ratio is given by $\theta = \frac{\Omega_1}{\Omega_2} = \frac{p_{2|1}/(1-p_{2|1})}{p_{2|2}/(1-p_{2|2})}$

$X \backslash Y$	0	1	
0	$p_{1\|1}$	$p_{2\|1}$	1
1	$p_{1\|2}$	$p_{2\|2}$	1

- Values of θ farther from 1.0 in a given direction represent stronger association.
- Two values represent the same association, but in opposite directions, when one is the inverse of the other. For instance, when $\theta = 0.25$ the odds of success in row 1 are 1/4 times the odds in row 2, or equivalently, the odds of success in row 2 are 1/0.25=4.0 times the odds in row 1.
- It is sometimes more convenient to consider the odd-ratio on a log-scale
  - $-\infty < \log\theta < +\infty$
  - $\log\theta = 0$ (i.e. $\theta = 1$) means independence of the two variables
  - two values for $\log\theta$ that are the same except for sign represent the same strength of association.

Odd and odds ratio

Odds ratio can be extended to the case of tables larger than a 2×2 table

Odds ratio can be extended to the case of tables obtained by classifying more than two variables

Reference

Agresti A. (2002) Categorical Data Analysis John Wiley & Sons, Inc. Book Series: Wiley Series in Probability and Statistics