stcp-rothwell-med

Medical Statistics definitions

Medical statistics is a branch of statistics which focuses on medical applications. It introduces new methods for analysing proportions of events, which can be defined as $\boldsymbol{p = \frac{a}{n} = \frac{number\ of\ events\ occuring}{number\ in\ group}}$ where $\boldsymbol{a}$ is a subset of $\boldsymbol{n}$, in terms of risk. For example, the proportion of people who ate toast for breakfast would have $\boldsymbol{a}$ as the number of people who ate toast from the sample questioned and $\boldsymbol{n}$ as the total number of people in the sample.

This sheet will briefly explain various terms which arise in medical statistics regularly.

Risk/ Prevalence (P): The prevalence (or risk) of a disease is calculated as

$$\boldsymbol{P_{disease} = \frac{number\ of\ people\ with\ disease}{total\ number\ of\ people\ within\ group}}$$

Risks are not always negative e.g. risk of surviving or probability of winning the lottery are calculated in the same way.

We are often interested in comparing risks from different groups and there are several ways of doing this. The following table will be used to demonstrate the formulae involved.

	Event occurs (Develop disease/ Died)	Event does not occur (Does not develop disease/ Survived)	Total	Risk of disease (death) by group
Exposed/ Treated	a	b	a+b	$\boldsymbol{p_{exp} = \frac{a}{(a+b)}}$.
Not Exposed/ Not treated	c	d	c+d	$\boldsymbol{p_{unexp} = \frac{c}{(c+d)}}$

Relative risk: This measures how much more likely the event is to occur in one group compared to another.

The risk of developing the disease for the exposed population is $\boldsymbol{P_{exp} = \frac{a}{(a+b)}}$.

The risk of developing the disease for the unexposed population is $\boldsymbol{P_{unexp} = \frac{c}{(c+d)}}$.

Relative risk = $\boldsymbol{RR = \frac{risk\ of\ developing\ disease\ in\ exposed\ group}{risk\ of\ developing\ disease\ in\ unexposed\ group} = \frac{P_{exp}}{P_{unexp}} = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} = \frac{a(c+d)}{c(a+b)}}$

This is sometimes called a Risk Ratio. If $\boldsymbol{RR > 1}$ then the risk of disease for the exposed group is larger than the risk of disease for the unexposed group.

© Joanne Rothwell
University of Sheffield

Reviewer: Chris Knox
University of Sheffield

Example: A randomised controlled trial investigated mortality rates within a year for 300 patients with lung cancer. The first group received a new chemotherapy treatment for lung cancer (New treatment) and the other received the standard chemotherapy treatment (Control treatment).

	Died	Survived	Total	Risk of dying
Control treatment	50	150	200	50/200 = 0.25
New treatment	10	90	100	10/100 = 0.1

The relative risk is $\frac{risk\ of\ dying\ in\ control\ group}{risk\ of\ dying\ in\ new\ treatment\ group} = \frac{0.25}{0.1} = 2.5$. This means that those in the control group were 2.5 times more likely to die than those in the treatment group. When calculating relative risks, it is easier to use the group with the highest risk in the numerator.

Sometimes a confidence interval is reported with the Relative Risk calculated from a sample. A confidence interval gives a range of likely values for the population relative risk.

95% Confidence Interval (CI) for a RR: For large samples this can be calculated using the natural logarithm (ln) because the confidence interval is not symmetrical.

First the variance of $\mathbf{ln}(\boldsymbol{RR})$ needs to be calculated:

$$\boldsymbol{Var(lnRR)} = \frac{\mathbf{1}}{\boldsymbol{a}} - \frac{\mathbf{1}}{\boldsymbol{a+c}} + \frac{\mathbf{1}}{\boldsymbol{b}} - \frac{\mathbf{1}}{\boldsymbol{b+d}} = \frac{\mathbf{1}}{\mathbf{50}} + \frac{\mathbf{1}}{\mathbf{50+10}} + \frac{\mathbf{1}}{\mathbf{150}} + \frac{\mathbf{1}}{\mathbf{150+90}} = \mathbf{0.0475}$$

95% Confidence interval for the ln(RR)

$$\mathbf{ln}(\boldsymbol{RR}) \pm \mathbf{1.96} \times \sqrt{\boldsymbol{Var(lnRR)}} = \mathbf{ln}(\mathbf{2.5}) \pm \mathbf{1.96} \times \sqrt{\mathbf{0.0475}} = \mathbf{0.916} \pm \mathbf{0.427} = (\mathbf{0.489}, \mathbf{1.343})$$

To get the confidence interval for the actual relative risk, take the exponential of the upper and lower value, so the 95% confidence interval will be

$$\left(\mathbf{e}^{\mathbf{ln}(\boldsymbol{RR})-\mathbf{1.96}\times\sqrt{\boldsymbol{Var(lnRR)}}}, \mathbf{e}^{\mathbf{ln}(\boldsymbol{OR})+\mathbf{1.96}\times\sqrt{\boldsymbol{Var(lnRR)}}}\right) = (\boldsymbol{e}^{\mathbf{0.489}}, \boldsymbol{e}^{\mathbf{1.343}}) = (\mathbf{1.63}, \mathbf{3.83}).$$

The relative risk for the whole population is likely to be between 1.63 and 3.83. If the confidence interval includes 1, the risk in one group is not significantly higher than the risk in the second group. Here, both values are above 1 so the risk in the control group is significantly higher (RR 2.5, 95% CI: 1.63 to 3.83).

The **Relative Risk Difference** (RRD) is given by $\boldsymbol{RRD} = \mathbf{1} - \boldsymbol{RR}$ when $(\boldsymbol{RR} < \mathbf{1})$.

Put the smallest risk (treatment group) on top to get a RR under 1: $\frac{\mathbf{0.1}}{\mathbf{0.25}} = \mathbf{0.4}$. Therefore the RRD is 1 – 0.4 = 0.6. The risk of dying is reduced by 60% in the treatment group.

The **Absolute Risk Difference** (ARD) is given by

$\boldsymbol{ARD} = |\boldsymbol{P_{exp}} - \boldsymbol{P_{unexp}}| = |\mathbf{0.25} - \mathbf{0.1}| = \mathbf{0.15}$. The absolute risk has decreased by 15%.

The **Number Needed to Treat (**NNT) **is the additional number of people you would need to give a new treatment to in order to cure one extra person compared to the**

old treatment and is given by $\boldsymbol{NNT = \frac{1}{ARD} = \frac{1}{0.15} = 6.67}$. So 7 people would need to receive the new treatment for one extra person to survive compared to the old treatment.

Another common measure used in medical statistics is the **Odds Ratio** (OR). First, odds are calculated by $\boldsymbol{odds = \frac{p}{1-p}}$, where $\boldsymbol{p}$ is the probability of an event occurring.

Therefore, the odds of disease in the exposed group would be $\boldsymbol{odds_{exp} = \frac{P_{exp}}{1-P_{exp}}}$ and similarly the odds of disease in the unexposed group would be $\boldsymbol{odds_{unexp} = \frac{P_{unexp}}{1-P_{unexp}}}$.

Then the odds ratio is $\boldsymbol{OR = \frac{odds_{exp}}{odds_{unexp}} = \frac{ad}{bc}}$, which would be the odds of disease in the exposed group compared to the unexposed group. If the $\boldsymbol{OR > 1}$ then the odds of disease occurring in the exposed group are larger than the odds of disease in the unexposed group, so exposure to the factor has increased the risk of contracting the disease. For our example the odds ratio is $\boldsymbol{OR = \frac{ad}{bc} = \frac{50\times90}{10\times150} = 3}$, The odds ratio looks at the odds of being in a particular treatment group given that you had the disease. So, in this example, those who improved were 3 times more likely to have received the new treatment than the control treatment.

Note: The odds ratio and relative risk are similar if the total sample is large and the disease is rare.

95% Confidence Interval (CI) for an OR: This can be calculated for large samples and must be carried out using the natural logarithm (ln) because the confidence interval is not symmetrical.

First the variance of $\boldsymbol{\ln(OR)}$ needs to be calculated.

$$\boldsymbol{Var(lnOR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} = \frac{1}{50} + \frac{1}{150} + \frac{1}{10} + \frac{1}{90} = 0.138}$$

95% CI: $\boldsymbol{\ln(OR) \pm 1.96 \times \sqrt{Var(lnOR)} = \ln(3) \pm 1.96 \times \sqrt{0.138} = (0.362, 1.818)}$ then take the exponential of the upper and lower value, so the 95% confidence interval will be $\boldsymbol{\left(e^{\ln(OR)-1.96\times\sqrt{Var(lnOR)}}, e^{\ln(OR)+1.96\times\sqrt{Var(lnOR)}}\right) = (e^{0.362}, e^{1.818}) = (1.43, 6.16)}$

The odds ratio comparing death rates after the standard treatment to the new treatment was 3 (95% CI: 1.43, 6.16) with those on the standard treatment being more likely to die.

Diagnostic Tests

		True Diagnosis		
		Disease +ve	**Disease -ve**	**Total**
Test Results	**+ve**	a	b	a+b
	-ve	c	d	c+d
		a+c	b+d	N

Sometimes there is a need to establish how good a diagnostic test is in detecting disease. One would have a table similar to the one above. A number of different measures can be gained from this information.

Sensitivity: $\boldsymbol{sens = \frac{a}{(a+c)}}$ This is the probability of getting a positive test result *given that the person has the disease.* $\boldsymbol{P(+ve|D)}$

Specificity: $\boldsymbol{spec = \frac{d}{(b+d)}}$ This is the probability of getting a negative test result *given that the person does not have the disease.* $\boldsymbol{P(-ve|ND)}$

Positive Predictive Value*: $\boldsymbol{PPV = \frac{a}{(a+b)}}$ This is the probability of the person having the disease *given they get a positive test result.* $\boldsymbol{P(D|+ve)}$

Negative Predictive Value*: $\boldsymbol{NPV = \frac{d}{(c+d)}}$ This is the probability of the person not having the disease *given they get a negative test result.* $\boldsymbol{P(ND|-ve)}$

Positive Likelihood Ratio: $\boldsymbol{LR^{+} = \frac{sens}{1-spec}}$ This gives a ratio of the test being positive for patients with disease compared with those without disease. Aim to be much greater than 1 for a good test.

Negative Likelihood Ratio: $\boldsymbol{LR^{-} = \frac{1-sens}{spec}}$ This gives a ratio of the test being negative for patients with disease compared with those without disease. Aim to be much less than 1 for a good test.

General rule – A screening test needs high sensitivity, a diagnostic test needs high specificity.

*these tests must have a **random** sample of the whole population; they depend on the prevalence of the disease which cannot be calculated if the sample is not random.

For tests with a continuous outcome, such as a blood biomarker measurement, one can determine a good cut-off point for the test using an ROC curve. This plots sensitivity against (1-specificity). A good diagnostic test will be the point closest to the top left corner of the plot.

There will be a trade-off between high sensitivity and high specificity. Ideally both of these values should be high.