DATA2002

Measures of risk

Garth Tarr

THE UNIVERSITY OF
SYDNEY

Prospective and retrospective studies

Estimating population proportions

Relative risk

Odds ratio

Standard errors and confidence intervals for odds ratios

Prospective and retrospective studies

Asthma and hay fever

A large group of infants with mild respiratory problems (but asthma free) is split into those with a family history of hay fever, and those without.

Random samples of 85 from the first group and 405 from the second group are selected for special study. Of these, the number diagnosed with asthma by the age of 12 are 25 and 70 respectively.

	Asthma	No Asthma	
Hay fever	25	60	**85**
No hay fever	70	335	**405**
	95	395	

Does a family history of hay fever increase the risk of developing asthma?

Hodgkin's disease and tonsillectomies

Vianna, Greenwald, and Davies (1971) collected data on a group of **101 patients suffering from Hodgkin's disease** and a comparable control group of **107 non-Hodgkin's patients**.

They were interested in the effect of tonsil tissue as a barrier to Hodgkin's disease. They found that in the Hodgkin's disease group, there had been 67 tonsillectomies. The corresponding figure for the non-Hodgkin's patients was 43.

	Hodgkin's disease		
Tonsillectomy	Yes	No	
Yes	67	43	110
No	34	64	98
	101	**107**	

Does having a tonsillectomy increase your risk of developing Hodgkin's disease?

Notation

How a study is conducted over time will affect various conditional probabilities.

To illustrate this we will use the following symbols.

- D^+ is the event that an individual **has** a particular disease.
- D^- is the event that an individual **does not have** a particular disease.
- R^+ is the event that an individual **has** a risk factor.
- R^- is the event that an individual **does not have** a risk factor.

Prospective (or cohort study) studies

A study design where one or more samples (called cohorts) are followed prospectively and subsequent status evaluations with respect to a disease or outcome are conducted to determine which initial participants exposure characteristics (risk factors) are associated with it.

As the study is conducted, an outcome from participants in each cohort is measured and relationships with specific characteristics determined.

- A **prospective study** is based on subjects who are initially identified as *disease-free* and classified by presence or absence of a *risk factor*.
- A random sample from each group is followed in time (prospectively) until eventually classified by disease outcome.

Prospective studies -- fictitious example

- A **prospective study** was designed to assess the impact of sun exposure on skin damage in beach volleyball players.
- During a weekend tournament, players from one team wore waterproof, SPF 35 sunscreen, while players from the other team did not wear any sunscreen.
- At the end of the volleyball tournament players' skin from both teams was analyzed for texture, sun damage, and burns.
- Comparisons of skin damage were then made based on the use of sunscreen.
- The analysis showed a significant difference between the cohorts in terms of the skin damage.

Prospective study -- Asthma

A large group of infants with mild respiratory problems (but asthma free) is split into those with a family history of hay fever, R^+, and those without, R^-. Random samples of 85 from the first group and 405 from the second group are selected for special study. Of these, the number diagnosed with asthma by the age of 12, D^+, are 25 and 70 respectively.

	D^+ **Asthma by age 12**	D^- **No asthma by age 12**	**Total**
R^+ **Family history of hay fever**	25	60	**85**
R^- **No family history of hay fever**	70	335	**405**
Total	**95**	**395**	**490**

- Risk factor is **family history of hay fever**
- Disease is **asthma by age 12**

In a prospective study the numbers **highlighted in red** are fixed by design.

Retrospective (or case control) studies

A study that compares patients who have a disease or outcome of interest (cases) with patients who do not have the disease or outcome (controls), and looks back retrospectively to compare how frequently the exposure to a risk factor is present in each group to determine the relationship between the risk factor and the disease.

A **retrospective study** is based on random samples from each of the two *outcome categories* which are followed back (retrospectively) to determine the presence or absence of the *risk factor* for each individual.

Retrospective studies -- fictitious example

- There is a suspicion that zinc oxide, the white non-absorbent sunscreen traditionally worn by lifeguards is more effective at preventing sunburns that lead to skin cancer than absorbent sunscreen lotions.
- A **retrospective study** was conducted to investigate if exposure to zinc oxide is a more effective skin cancer prevention measure.
- The study involved comparing a group of former lifeguards that had developed cancer on their cheeks and noses (cases) to a group of lifeguards without this type of cancer (controls) and assess their prior exposure to zinc oxide or absorbent sunscreen lotions.
- This study would be **retrospective** in that the former lifeguards would be asked to recall which type of sunscreen they used on their face and approximately how often.

Retrospective study -- Hodgkin's disease

Vianna, Greenwald, and Davies (1971) collected data on a group of 101 patients suffering from Hodgkin's disease, D^+, and a comparable control group of 107 non-Hodgkin's patients, D^-. They were interested in the effect of tonsil tissue as a barrier to Hodgkin's disease. They found that in the D^+ group, there had been 67 tonsillectomies, R^+. The corresponding figure for the D^- group was 43.

	D^+ Hodgkin's	D^- Non-Hodgkin's	Total
R^+ Tonsillectomy	67	43	**110**
R^- No tonsillectomy	34	64	**98**
Total	**101**	**107**	**208**

- Risk factor is **tonsillectomy**
- Disease is **Hodgkin's disease**

In a retrospective study the numbers **highlighted in blue** are fixed by design.

Estimating population proportions

Estimating a population proportion

Suppose

- we have a large (but finite) population containing objects/individuals of two different types (say type 0 and type 1);
- it is desired to determine or at least estimate the overall proportion of type 1 but it is not feasible to examine every object/individual.

If we can take a random sample from the population then we can use the sample proportion of type 1 as an estimate of the population proportion of type 1.

Extending this idea, consider two events A and B,

- If we can take a random sample from the whole population, we can estimate $P(A)$ using the observed sample proportion with attribute A
- If we can take a random sample from the subpopulation defined by B, we can estimate $P(A|B)$ using the observed sample proportion (of the subpopulation) with attribute A.

Application to prospective and retrospective studies

In both kinds of study we have

- a population;
- a subpopulation/attribute determined by a risk factor R^+ (with complementary subpopulation/attribute R^-);
- an subpopulation/attribute determined by having/developing the disease D^+ (with complementary subpopulation/attribute D^-).

The labels "subpopulation" and "attribute" here are mathematically equivalent (they both mean event).

The main difference between prospective and retrospective studies are which (sub)populations we can sample from.

Prospective study

- In a prospective study we take two random samples:
  - one from the risk factor group (subpopulation) R^+;
  - another from the non-risk factor group R^-.
- We then (wait to) see how many in each group develop the disease.
- We can thus estimate $P(D^+|R^+)$ as well as $P(D^-|R^-)$.
- We cannot however estimate $P(R^+|D^+)$ or $P(R^-|D^-)$ since we did not take random samples from the disease group.

Retrospective study

- In a retrospective study we take two random samples:
  - one from the disease group (subpopulation) D^+ and
  - another from the non-disease group (subpopulation) D^-.
- We then (look back to) see how many in each group were exposed to the risk factor.
- We can thus estimate $P(R^+|D^+)$ as well as $P(R^-|D^-)$.
- We cannot however estimate $P(D^+|R^+)$ or $P(D^-|R^-)$ since we did not take random samples from the risk factor group.

Relative risk

Measures of risk

These are different ways to measure the association between a risk factor/treatment and the disease outcome.

How the data is **sampled** will greatly impact the ways in which these methods are applicable and interpretable.

Relative risk

The relative risk is defined as a ratio of two conditional probabilities,

$$RR = \frac{P(D^+|R^+)}{P(D^+|R^-)}.$$

Since probabilities are bounded between 0 and 1

$$RR = \frac{P(D^+|R^+)}{P(D^+|R^-)} \to \infty \quad \text{as} \quad P(D^+|R^-) \to 0,$$

$$RR = \frac{P(D^+|R^+)}{P(D^+|R^-)} \to 0 \quad \text{as} \quad P(D^+|R^+) \to 0,$$

and $RR \approx 1$ when $P(D^+|R^+) \approx P(D^+|R^-)$.

If D and R are **independent** then $P(D|R) = P(D)$ and so

$$RR = \frac{P(D^+|R^+)}{P(D^+|R^-)} = \frac{P(D^+)}{P(D^+)} = 1.$$

Relative risk -- interpretation

$$RR = \frac{P(D^+|R^+)}{P(D^+|R^-)}$$

The relative risk is the ratio of the probability of having the disease in the group with the risk factor to the probability of having the disease in the group without the risk factor.

- $RR = 1$ means there is **no difference** between the two groups.
- $RR < 1$ implies the disease is **less likely** to occur in the group with the risk factor.
- $RR > 1$ implies the disease is **more likely** to occur in the group with the risk factor.

Relative risk -- prospective studies

	D^+	D^-	Total
R^+	a	b	$\mathbf{a+b}$
R^-	c	d	$\mathbf{c+d}$
	$a+c$	$b+d$	$a+b+c+d$

Given data from a **prospective study** or **from a sample of completed records** we can estimate these

- $P(D^+|R^+) = \dfrac{a}{a+b}$
- $P(D^+|R^-) = \dfrac{c}{c+d}$
- **Relative risk**: $\widehat{RR} = \dfrac{P(D^+|R^+)}{P(D^+|R^-)} = \dfrac{a(c+d)}{c(a+b)}$

Recall: in a prospective study the numbers **highlighted in red** are fixed by design.

Relative risk -- retrospective studies

	D^+	D^-	Total
R^+	a	b	$a+b$
R^-	c	d	$c+d$
	$\mathbf{a+c}$	$\mathbf{b+d}$	$a+b+c+d$

In a **retrospective study** (or cohort control study) we identify two groups (D^+ and D^-) and we retrospectively assess each group them for their risk status (R^+ and R^-). We "sampled on the outcome", choosing subjects on the basis of D and then observing R.

Due to the design, we cannot extract any information about the incidence of D in the population because the proportions of cases with D^+ and D^- were decided by the investigator. I.e. we cannot estimate $P(D^+|R^+)$, $P(D^+|R^-)$, or $RR = \dfrac{P(D^+|R^+)}{P(D^+|R^-)}$.

We can estimate $P(R^+|D^+)$ and $P(R^-|D^+)$ but these are not used in the calculation of relative risk.

Recall: in a retrospective study the numbers **highlighted in blue** are fixed by design.

Aspirin (relative risk)

Steering Committee of the Physicians' Health Study Research Group (1988) provide data on a 5 year (blind) study into the effect of taking aspirin every second day on the incidence of heart attacks.

	Myocardial infarction D^+	**No myocardial infarction D^-**	
Aspirin R^+	104	10,933	11,037
Placebo R^-	189	10,845	11,034
	293	21,778	22,071

Estimates for the proportion of each group having heart attacks:

$$P(D^+|R^+) = \frac{104}{10,933+104} = 0.0094$$

$$P(D^+|R^-) = \frac{189}{10,845+189} = 0.0171$$

The estimated relative risk in the is $0.0094/0.0171 = 0.55$.

Hence, you are roughly half as likely to have myocardial infarction if you take aspirin.

Odds ratio

Odds ratio

A common alternative to the **relative risk** is the **odds ratio**, denoted OR.

Odds are a ratio of probabilities. The **odds** are used as an alternative way of measuring the likelihood of an event occurring.

If the probability of event A is $P(A)$ the **odds** of event A is defined as

$$O(A) = \frac{P(A)}{1 - P(A)}.$$

In the risk/disease setting, the probability of disease for R^+ patients is $P(D^+|R^+)$, and so the odds is,

$$O(D^+|R^+) = \frac{P(D^+|R^+)}{1 - P(D^+|R^+)} = \frac{P(D^+|R^+)}{P(D^-|R^+)}.$$

Equivalent definitions of odds ratio

The ratio of the odds of a disease for R^+ patients to the corresponding odds for R^- patients is the odds ratio, OR :

$$\text{Definition 1:} \quad OR = \frac{O(D^+|R^+)}{O(D^+|R^-)} = \frac{P(D^+|R^+)}{P(D^-|R^+)} \Big/ \frac{P(D^+|R^-)}{P(D^-|R^-)}.$$

We can show that this ratio is identical to

$$\text{Definition 2:} \quad OR = \frac{O(R^+|D^+)}{O(R^+|D^-)} = \frac{P(R^+|D^+)}{P(R^-|D^+)} \Big/ \frac{P(R^+|D^-)}{P(R^-|D^-)}.$$

This means that OR can be found from both **prospective** and **retrospective** studies, unlike RR.

Odds ratio -- invariance

Consider the table

	D^+	D^-	Total
R^+	a	b	$a+b$
R^-	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

$$\text{Def. 1: } OR = \frac{P(D^+|R^+)}{P(D^-|R^+)} \Big/ \frac{P(D^+|R^-)}{P(D^-|R^-)} = \left(\frac{\frac{a}{a+b}}{\frac{b}{a+b}}\right) \Big/ \left(\frac{\frac{c}{c+d}}{\frac{d}{c+d}}\right) = \frac{ad}{bc}$$

$$\text{Def. 2: } OR = \frac{P(R^+|D^+)}{P(R^-|D^+)} \Big/ \frac{P(R^+|D^-)}{P(R^-|D^-)} = \left(\frac{\frac{a}{a+c}}{\frac{c}{a+c}}\right) \Big/ \left(\frac{\frac{b}{b+d}}{\frac{d}{b+d}}\right) = \frac{ad}{bc}$$

Same no matter which definition is used.

W Odds ratio: Invertibility and invariance

Odds ratio -- interpretation

$$OR = \frac{P(D^+|R^+)}{P(D^-|R^+)} \Big/ \frac{P(D^+|R^-)}{P(D^-|R^-)} = \frac{ad}{bc},$$

If D and R are independent then $P(D|R) = P(D)$ and

$$OR = \frac{P(D^+|R^+)}{P(D^-|R^+)} \Big/ \frac{P(D^+|R^-)}{P(D^-|R^-)} = \frac{P(D^+)}{P(D^-)} \Big/ \frac{P(D^+)}{P(D^-)} = 1.$$

It can be shown that $OR = 1$ if and only if D and R are independent (there is no relationship between risk and disease).

Large odds ratios ($OR > 1$) implies increased risk of disease and small odd ratios ($OR < 1$) implies decreased risk of disease.

Aspirin (odds ratio)

	Myocardial infarction D^+	**No myocardial infarction D^-**	
Aspirin R^+	104	10,933	11,037
Placebo R^-	189	10,845	11,034
	293	21,778	22,071

The **odds ratio** is

$$OR = \frac{104 \times 10,845}{189 \times 10,933} = 0.55.$$

The estimated odds of heart attack for patients taking the aspirin is 0.55 times the estimated odds for those taking the placebo.

Compare this with the relative risk of 0.55. These are similar because the disease is rare.

Standard errors and confidence intervals for odds ratios

Standard errors and confidence intervals

The **odds ratio** estimator, OR, has a skewed distribution on $(0, \infty)$, with the neutral value being 1.

The **log odds** estimator, $\log(OR)$, has a more symmetric distribution centred at 0 if there is no difference between the two groups.

Note: an odds ratio of $a \in (0, 1)$ is equivalent to a value of $a^{-1} \in (1, \infty)$ just by relabeling the categories. The log transformation is such that $\log(a^{-1}) = -\log(a)$.

Standard errors and confidence intervals

The asymptotic standard error for $\log(\widehat{OR})$ is

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}.$$

A large sample 95% confidence interval for log θ is approximately

$$\log(\widehat{OR}) \pm 1.96 \times \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

from which we can approximate a confidence interval for the odds-ratio,

$$\left(\exp\left(\log(\widehat{OR}) - 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right), \exp\left(\log(\widehat{OR}) + 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)\right).$$

Note that these should only be applied if a, b, c and d are reasonably large (so that asymptotics hold).

Aspirin

	Myocardial infarction D^+	No myocardial infarction D^-	
Aspirin R^+	104	10,933	11,037
Placebo R^-	189	10,845	11,034
	293	21,778	22,071

$$\widehat{OR} = \frac{104 \times 10,845}{189 \times 10,933} = 0.55 \quad \text{and} \quad \log(\widehat{OR}) = -0.6$$

$$\text{SE}(\log(\widehat{OR})) = \sqrt{\frac{1}{104} + \frac{1}{189} + \frac{1}{10933} + \frac{1}{10845}} = 0.12.$$

A 95% confidence interval for the log odds-ratio is

$$-0.6 \pm 1.96 \times 0.12 \approx (-0.84, -0.36).$$

A 95% confidence interval for the odds-ratio is

$$(e^{-0.84}, e^{-0.36}) \approx (0.43, 0.69).$$

Returning to our initial examples

Hodgkin's disease

Vianna, Greenwald, and Davies (1971) collected data on a group of 101 patients suffering from Hodgkin's disease, and a comparable control group of 107 non-Hodgkin's patients. They were interested in the effect of tonsil tissue as a barrier to Hodgkin's disease.

	D^+ Hodgkin's	D^- Non-Hodgkin's	Total
R^+ **Tonsillectomy**	67	43	**110**
R^- **No tonsillectomy**	34	64	**98**
Total	**101**	**107**	**208**

Retrospective or prospective? Why? Can we calculate a relative risk?

Hodgkin's disease

The estimated odds-ratio and log odds-ratio are

$$\widehat{OR} = \frac{67 \times 64}{43 \times 34} = 2.93 \quad \text{and} \quad \log(\widehat{OR}) = \log(2.93) = 1.07$$

Hence, ~~tonsillectomy patients are three times as likely to have Hodgkin's disease~~ the odds of a tonsillectomy patient having Hodgkin's disease are three times the odds of a non-tonsillectomy patient having Hodgkin's. The standard error of the log odds-ratio is

$$\sqrt{\frac{1}{67} + \frac{1}{43} + \frac{1}{34} + \frac{1}{64}} = 0.29.$$

A 95% confidence interval for the log odds-ratio is

$$1.07 \pm 1.96 \times 0.29 \approx (0.51, 1.63)$$

and a 95% confidence interval for the odds-ratio is $(e^{0.51}, e^{1.63}) \approx (1.66, 5.10)$.

What can we conclude from the confidence interval? Does having a tonsillectomy increase your risk of developing Hodgkin's disease?

Asthma

A large group of infants with mild respiratory problems (but asthma free) is split into those with a family history of hay fever, R^+, and those without, R^-. Random samples of 85 from the first group and 405 from the second group are selected for special study. Of these, the number diagnosed with asthma by the age of 12, D^+, are 25 and 70 respectively.

	D^+ **Asthma by age 12**	D^- **No asthma by age 12**	**Total**
R^+ **Family history of hay fever**	25	60	**85**
R^- **No family history of hay fever**	70	335	**405**
Total	**95**	**395**	**490**

? Retrospective or prospective? Why? Can we calculate a relative risk?

Asthma

The odds ratio, log odds ratio and the standard error of the log odds ratio are,

$$\widehat{OR} = \frac{25 \times 335}{60 \times 70} = 1.99 \quad \text{and} \quad \log(\widehat{OR}) = 0.69$$

with

$$\text{SE}(\log(\widehat{OR})) = \sqrt{\frac{1}{25} + \frac{1}{60} + \frac{1}{70} + \frac{1}{335}} = 0.27.$$

A 95% confidence interval for the log odds-ratio is

$$0.69 \pm 1.96 \times 0.27 \approx (0.16, 1.22)$$

and a 95% confidence interval for the odds-ratio is $(e^{0.16}, e^{1.22}) \approx (1.17, 3.39)$.

? What can we conclude from the confidence interval? Does a family history of hay fever increase the risk of developing asthma?

References

Additional reading material for this lecture can be found in Agresti (2007) section 2.2 (starting page 25). I've added a link to this ebook in the Reading List tab on Canvas.

Agresti, A. (2007). *An introduction to categorical data analysis*. 2nd ed.. Wiley series in probability and mathematical statistics. Hoboken, NJ: Wiley-Interscience. ISBN: 9780471226185.

Steering Committee of the Physicians' Health Study Research Group (1988). "Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study". In: *New England Journal of Medicine* 318.4, pp. 262-264. DOI: 10.1056/NEJM198801283180431.

Vianna, N. J., P. Greenwald, and J. N. P. Davies (1971). "Tonsillectomy and Hodgkin's disease: the lymphoid tissue barrier". In: *The Lancet* 297.7696, pp. 431-432. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(71)92416-0.

See also:

- Principles of Epidemiology in Public Health Practice, Third Edition. An Introduction to Applied Epidemiology and Biostatistics. Lesson 3: Measures of Risk; Section 5: Measures of Association 🔗
- Cochrane Handbook for Systematic Reviews of Interventions, Version 6.2, 2021 Section 6.4 Dichotomous outcome data 🔗