QUESTIONS WITH MODELS ANSWERS

Session 6: Tutorial 1 - Tools of the trade: understanding and interpreting the findings commonly reported in papers

Dr Alex Bottle, Dr Mireille Toledano

Instructions:

This tutorial is designed to help you to understand the commonly reported findings you see in papers published in medical journals.

Based on feedback from previous years' students, we have changed the way we teach medical and epidemiological statistics. The focus is now on the interpretation of the statistics rather than their calculation and teaching will be done via a tutorial session rather than in a lecture theatre. Please note that some of the material in this tutorial will build on what you will learn in lectures 5, 6, 7, 8, 9, and 10, but some of it will not be covered anywhere else in the course. **All material taught in this tutorial will be included in your examinations.**

The introductory text describes two worked examples. These examples have been provided to teach you core concepts and to help put into context what you have already learnt; there is also a glossary at the end of the tutorial to define the key terms you will need to know, these terms are italicised in the text. You should read these worked examples before the tutorial, so that you have sufficient time to work through the questions provided during the timetabled session with your tutor.

Learning outcomes:

- Be able to understand the concept of sampling and sampling variation
- Be able to understand that from a *sample*, estimates of the true underlying *risk* in a population can be calculated.
- Be able to define and interpret a *P value* and a *confidence interval*
- Be able to explain the role of statistical hypothesis testing and *confidence intervals* when dealing with chance
- To know the difference between probability and odds and be able to interpret and explain measures of association (*relative risk*, *attributable risk*, *odds ratio*) from simple examples
- Define *confounding* and understand the problems associated with it. Be able to list some methods for dealing with *confounding* (including *stratification*, *standardisation* and *regression*).

Suggested further reading:

Martin Bland (2000) ***An introduction to medical statistics.*** Oxford University Press.

Worked example 1 (sampling, P values and confidence intervals)

What is the role of statistics in medicine? Discuss!

1.1 Sampling – estimating prevalence of disease or risk factors

A Primary Care Trust (PCT) wants to estimate the *prevalence* of smoking among their 100,000 residents. What does prevalence mean? How would they do this?
Suppose they surveyed a random *sample* of people – why take a random sample?
Suppose they asked 100 people if they smoked and found that 28 did. If they then asked another 100, would they also find that 28 of them smoked? Why might they not?

If they kept sampling sets of 100 people and plotted the percentage of smokers (prevalence of smoking) in each sample, we would expect to see a *normal distribution* (see glossary), with most sample estimates centred around the true population percentage.

1.2 Confidence intervals and P values – assessing the role of chance

The PCT's estimate of their population's smoking prevalence is 28% from their sample, but there will be some uncertainty around this estimate. We express this uncertainty using a 95% *confidence interval* (95% CI) around the estimate, e.g. 19% to 37%. This means that if we repeated the sampling 100 times, we would expect the true prevalence of smoking in the PCT to fall within the CI in 95 of the 100 samples.

Suppose the PCT wanted to lower this prevalence; they could implement a smoking reduction campaign and then see if it worked by comparing their first estimated prevalence with an estimate after the campaign. They took two random samples, the first finding that 28% smoked as above, and the second finding that 21% smoked. Can we therefore say for certain that the campaign has worked and cut the prevalence by 28-21=7%?
Why not?
We want to know whether the difference of 7% could simply be due to chance (sampling error) or is a real difference in prevalence. This is done statistically by setting up a *null hypothesis* of no difference and looking for evidence to disprove it: what is the likelihood that our two samples were 28% and 21% if the two true underlying prevalences were the same? We then choose the appropriate statistical test (e.g. chi-squared test to compare the two proportions) to get this likelihood, which is the *P value*. The lower the P value, the less likely that our estimated difference is a chance finding. Suppose the P value was 0.014. Convention has it that if $P<0.05$ (and this is an arbitrary cut-off!) then we can reject the null hypothesis and conclude that the smoking prevalence fell after the campaign. Such a result is called statistically significant.

THE PROCEDURE:

1. Set up a null hypothesis (e.g. difference in prevalence between the two groups is zero)
2. Choose an appropriate statistical test
3. Inspect the results (estimated measure of association – or, in this case, estimated difference in prevalences – plus its CI and P value) for evidence of real difference: can we reject the null hypothesis?

Are statistically significant results more or less likely with small sample size than with large sample sizes? Why (the answer is to do with the nature of the sample rather than statistics)?

Worked example 2 (measures of association)

The main aim of epidemiological research is to investigate the association between exposure to a risk factor (e.g. smoking) and the occurrence of disease (e.g. lung cancer). We compare the incidence in a group of people exposed to the risk factor with a group who were not exposed. Suppose the incidence in one group is higher than in the other – what are the two different ways of stating this? If Joe is 36 and John is 18, how could we say by how much Joe is the elder?

2.1 Ratio measures: relative risk and odds ratio

Two key concepts: risk and odds. What is the difference?

Suppose you wanted to look at possible risk factors for lung cancer: smoking and occupational exposure. How might you select your population sample to do this?

COHORT STUDY: Malarcher et al (2000) took a large group of US males, some smoke and some never have done, and followed them up over time. They measured the *rates* of lung cancer in the two groups. They set up the null hypothesis of equal rates and calculated the *relative risk*: 27 for smokers compared with those who have never smoked (95% CI 19 to 38). Can we reject the null hypothesis?

The interpretation of a relative risk is straightforward: if you smoke, you are 27 times more likely to die from lung cancer than if you don't smoke.

CASE-CONTROL STUDY: Richiardi 2005 took a group of people with lung cancer (the *cases*) and another without lung cancer (the *controls*) and asked each about their occupation (whether they were dockers or freight handlers). Their occupation is the exposure here. Richiardi measured the odds of exposure (odds of working as a docker or freight handler) in the cases and then in the controls. They set up the null hypothesis of equal odds and calculated the *odds ratio*: 1.5 for those with lung cancer compared with those without (95% CI 1.1 to 2.1). Can we reject the null hypothesis?

This odds ratio means that someone with lung cancer is 1.5 times more likely to have worked as a docker or freight handler than someone who doesn't have lung cancer. Notice that it compares the exposure in the two groups – it does not compare the disease rates in the two groups, which the relative risk does. The odds ratio is an estimate of the relative risk, and it is usually more useful (and easier!) to interpret an odds ratio to mean that if you work as a docker or freight handler you are 1.5 times more likely to get lung cancer than if you work in a different occupation. See the glossary for an explanation of the relationship between relative risk and odds ratio and on why case-control studies can only provide us with the latter.

2.2 Difference measure: attributable risk (or attributable fraction)

The attributable risk for lung cancer in smokers is the rate of lung cancer amongst smokers minus the rate of lung cancer amongst non-smokers (i.e. the risk difference). It gives an indication of how many extra cases for which the exposure is responsible, making the important assumption that the relation between the exposure and the disease is causal (i.e. not explained by other confounding factors – see below). The attributable risk and related measures are typically used to help guide policymakers in planning public health interventions.

2.3 Confounding – and controlling for it

How can we prove that an exposure causes a disease, rather than is merely associated with higher rates of that disease? We try to eliminate (i.e. control or adjust for) the effects of confounders. Confounders are associated with both the exposure of interest and the outcome of interest (e.g. developing a disease or dying).

Confounding can be dealt with at the design stage of a study by *randomisation* (in a randomised controlled trial), *restriction*, or *matching* (in a case-control study). Alternatively, confounding variables can be controlled for at the analysis stage, by *stratification* (splitting the analysis e.g. by age group), *standardisation*, or *regression* (building a statistical model).

In Richiardi's case-control study, regression was used to control for the effect of smoking on lung cancer risk. The lung cancer risk associated with working as a docker or freight handler after controlling for the effect of smoking was reduced to 1.3 (95% confidence interval 0.9 to 1.9). Although the odds ratio is still higher (by 30%) for dockers or freight handlers, the confidence interval now spans 1 and so we can accept the null hypothesis that working as a docker or freight handler has no effect on lung cancer risk. This is because the higher odds reported for dockers or freight handlers could just have been found by chance. Smoking is therefore a confounder here, as it's associated with both the exposure (being a docker or freight handler) and the disease (lung cancer).

We can only adjust (control) for confounding factors if we have measured them. How often have you watched a TV news piece about an association between some potential risk factor and a disease and wondered, 'but could that be due to X instead?' Journalists rarely bother to talk about confounders.

Tutorial questions
The following questions will be undertaken in small groups, facilitated by a tutor. All the questions are designed to test your understanding of, and help you apply, the knowledge you will have learnt by reading the above worked examples, from listening to your tutor briefly explain the core concepts in the worked examples, and from the material covered in your lectures on the course so far. The questions should be worked though in groups; if you get stuck at any point please refer to the glossary at the end of this tutorial and ask your tutor for help.

Question 1 – Sampling distribution and confidence intervals
A study was conducted to assess whether hormone replacement therapy (HRT) conferred a protective effect on acute myocardial infarction *risk*. 1013 women with an acute myocardial infarction and 5000 women of a similar age range without acute myocardial infarction were asked whether or not they currently used HRT. 13.1% of the women who had had an MI used HRT, whereas 17.1% of the women who had not had an MI had used HRT. This study reported an *odds ratio* of 0.72 (95% *confidence interval* 0.59-0.88) for current or recent HRT use on acute myocardial infarction *risk* (Varas Lorenzo, 2000).

a) What type of study is this?
Case-control – participants were selected on the basis of whether they were cases or not, and then their exposure status was assessed.

b) Why were 1013 women with an MI recruited instead of, say, 50? Why not 50,000?
50 would give an unreliable estimation of the proportion taking HRT, whilst 50,000 would cost a fortune and might be overkill in terms of reliability of estimates.

c) Why were the 5000 "*controls*" (women without the *outcome* of interest, i.e. MI) chosen to have a similar age range as the "*cases*" (women with MI)?
Age is clearly related to a woman's chances of taking HRT. Imagine if the controls had been taken from women aged 65+ or 18-44: would this have been a fair comparison?

d) What is the *null hypothesis* that this study is trying to disprove? Always be specific – don't just say "that there is no difference".
That the odds of taking HRT in women who had had an MI are the same as the odds of taking HRT in women who had not had an MI, i.e. the odds ratio equals 1. This would mean that taking HRT does not affect your chances of getting an MI (at least in the age range of those studied here)

e) The 95% *confidence interval* for the *odds ratio* was 0.59-0.88. What does this mean?
In general, if we repeated this study at a 100 other hospitals with 100 other sets of controls, each time generating an odds ratio and a 95% confidence interval for that odds ratio, 95% of these confidence intervals would contain the true (population-level) odds ratio. Although not strictly correct, it is best to think of the 95% CI as meaning that the real odds ratio for the effect of HRT on MI is somewhere between 0.59 and 0.88.

f) For us to accept the *null hypothesis*, what would the 95% *confidence interval* look like? Give an example of its values.
It would include the value expected under the null hypothesis, i.e. 1.00, e.g. 0.35-1.09.

g) What does the *odds ratio* of 0.72 mean in words, and how would you explain this *odds ratio* to someone taking HRT?
The odds of taking HRT in women who have had an MI is 0.72 times the odds of taking HRT in women without an MI.
The risk of acute myocardial infarction is reduced by ~30% in women who currently or recently used HRT.

Question 2 – Dealing with confounding (in study design and analysis)
In a randomised controlled trial of patient self-monitoring of blood pressure in Birmingham general practices (McManus et al, 2005), 441 hypertensives were randomly allocated to either the usual monitoring by the practice (*control* group) or self-monitoring (intervention group). After six months, the intervention group reduced their systolic BP by an average of 4.3 mmHg (95% CI 0.8-7.9) more than the *control* group.

a) What is an appropriate distribution for a group of patients' BP?
The normal distribution, defined by its mean and SD.

b) What was the main *null hypothesis* for this study? Be specific, rather than just saying that "there is no difference".
That the mean systolic BP of the two groups would be the same after six months of follow up: i.e. after six months, the mean BP of the intervention group minus the mean BP of the control group would be 0.

c) Do we have evidence to reject the *null hypothesis*? What does this mean?
Yes – the CI does not include 0, the value expected under the null hypothesis. The mean difference of 4.3 mmHg is unlikely to be due to chance, suggesting a genuine difference between the two groups. However, this difference might be explained by other factors (bias and confounding).

d) Why did the investigators randomly allocate patients to the two groups?
To try to ensure that possible confounders were equally distributed across the groups. This randomisation is one method for controlling for unknown confounding variables at the design stage of a study.

e) *Randomisation* was "stratified by diabetic status". What does this mean and why was it done?
All the people with diabetes were divided evenly (with each person allocated at random to a particular group) between the groups so that any treatment effect would not be due to a disproportionate number of people with diabetes, who have a higher BP on average than people without, in one group or the other. This is an example of controlling for confounding in the design stage.

f) Other than diabetes, what other confounders might we want to control for?
You would want to control for any factors known to be associated with both the exposure (monitoring blood pressure) and outcome (blood pressure), and also for any factors thought to be potentially associated with exposure and outcome. The authors also adjusted for GP practice, sex and deprivation.

g) The authors found that the intervention group had lost more weight and cut down their alcohol at the six-month follow-up stage. What is the relevance of this finding?
It gives a mechanism for the observed fall in mean BP for the intervention group, i.e. they improved their lifestyle and BMI and therefore reduced their BP. A useful point when assessing whether the study's finding is real or due to chance, bias and confounding is whether there is a plausible biological explanation for it, as here. Recall the Bradford-Hill criteria for causation.

Question 3 – Understanding measures of association (and confounding)
To estimate the *incidence* of breast cancer in the UK population, records from the NHS breast screening programme (which screens women aged between 50 and 70) were explored. These data indicated that the *incidence* of breast cancer was 289 per 100,000 population.

a) What does "*incidence*" mean?
The number of new cases of breast cancer over a defined period of time divided by the number of people in the population over the same period (usually we would restrict the population to those at risk of breast cancer).

b) What can the *incidence* in this *sample* of the population tell us about the *incidence* in the whole UK female population?
Because breast cancer risk varies with age, these data cannot tell us anything about the incidence of breast cancer in the whole female population of the UK. However, assuming that a fairly random selection of women in this age group attend the screening, these data can provide a pretty good estimate of the incidence in women aged between 50 and 70.

c) A *null hypothesis* that the *incidence* of breast cancer in the UK female population aged 50-70 (289 per 100,000) is that same as the *incidence* in the UK female population aged 30-50 (90 per 100,000) gives a *p-value* of <0.0001. How would you interpret this *p-value*?
This p-value suggests that the null hypothesis can be rejected. This observed difference in incidence of breast cancer in these two populations of different ages would only be expected to occur by chance less than once every 10,000 times, meaning we can be quite happy in accepting that the incidences by age group are different.

d) The *risk* of getting breast cancer if you are a woman aged 50-70 relative to the *risk* of getting breast cancer if you are a woman aged 30-50 is 3.20 (95% *confidence interval* 3.11-3.29). How would you interpret this *relative risk*?
Women aged 50-70 are 3.2 times more likely to get breast cancer than women aged 30-50. The confidence intervals give us an indication as where the true (population-level) relative risk is likely to be, and in this case women aged 50-70 are most likely to be between 3.11 and 3.29 times more likely to get breast cancer than women aged 30-50.

e) The *odds* of being aged 50-70 if you have breast cancer compared with the *odds* of being aged 30-50 if you have breast cancer is also 3.2 (95% *confidence interval* 3.11-3.29). When the *odds ratio* and *relative risk* are calculated differently, why is it that they are the same in this study?
The odds ratio is an estimate of the relative risk, and for rare outcomes (such as cancers), these measures of effect will give similar estimates of risk. However, this is not the case for common outcomes, or for rare outcome studied over very long periods of time.

f) The crude *relative risk* of breast cancer in women who are current users of HRT is 1.83 (95% CI 1.72-1.93), compared with the age-adjusted *relative risk* of 2.00 (1.91-2.09) (Beral, 2003). Which of these *risk* estimates would you consider to best reflect the *risk* of breast cancer associated with HRT use?
The incidence of breast cancer is very dependant on the age of the women (as shown above); additionally HRT use is likely to be related to age (pre versus post menopausal). Here age is likely to act as a confounder in the relationship between HRT use and breast cancer, so the age adjusted risk is likely to be most informative.
Because many disease incidences vary with age, and because most populations do not have identical age structures it is common to see 'age standardised' relative risks reported. This means that if the cases and controls, exposed and non exposed, or study and comparison populations have different age structures, a useful risk measure can still be derived.

Question 4 – Relative risk vs attributable risk
An occupational study was carried out to investigate the effect of *exposure* to aromatic amines on bladder cancer *risk*. 6667 workers with potential *exposure* to aromatic amines were followed over 30 years to see what effect this *exposure* had on bladder cancer *risk*.

a) What type of study is this?
A cohort study. A group of people were selected for the study on the basis of their exposure status, and were followed over time to see who developed bladder cancer.

b) One quarter of the study population were exposed to aromatic amines, and the *risk* associated with this *exposure* on bladder cancer was found to be 296.94 (95% CI 41.45-2127.34). What does this *risk* measure tells us?
That exposure to aromatic amines is strongly (and statistically significantly) associated with bladder cancer risk.

c) How would you explain this *risk* to someone with occupational *exposure* to aromatic amines?
If you are exposed to aromatic amines at work your risk of getting bladder cancer is 297 times higher than if you were not exposed to aromatic amines.

d) The *population excess fraction* (excess fraction of bladder cancer due to aromatic amine *exposure* in the whole study population) is 98.7 percent. How would you interpret this figure?
Assuming causality, 98.7 percent of the bladder cancer cases in the study population can be attributed to occupational exposure to aromatic amines.

e) One quarter of this study population are cigarette smokers. Cigarettes contain low doses of aromatic amines and have also been found to be associated with an excess *risk* of bladder cancer, with a *relative risk* of 5.11 (95% CI 3.42-7.64), and a *population excess fraction* of 50.7%. To reduce bladder cancer *incidence* in this cohort, would it be better to reduce work place *exposure* to aromatic amines, or to encourage the workers to stop smoking?
Reducing work place exposures to aromatic amines would lead to the greatest reduction in bladder cancer risk. Whilst the prevalence of occupational and smoking related exposure in this cohort is the same (25%), the risk associated with occupational exposure is much greater that that associated with smoking.

f) Assume the same *risks* associated with occupational aromatic amine *exposure* (*relative risk* of ~297) and smoking (*relative risk* of ~5) in the occupational cohort apply to the whole population of England. In this England 'cohort', only 0.001% of the population has occupational exposure to aromatic amines, whilst 25% smoke. The *population excess fraction* is now 22.8% for aromatic amines, but remains at ~50% for smoking. Which *exposure* should be minimised to reduce *incidence* of bladder cancer in this population?
Although the relative risk associated with smoking is much lower than the relative risk associated with exposure to aromatic amines, the prevalence of exposure to aromatic amines is very low in this population. Assuming these exposures are causally related to bladder cancer, then of the ~10,000 new cases of bladder cancer in England each year, half of them (~5,000) could be attributed to smoking, and 22.8% of them (~2,280) could be attributed to occupational exposure to aromatic amines. It would therefore be more effective to reduce the number of smokers in the population than to minimise aromatic amine exposure (although obviously it would be best to reduce both exposures!).

g) Assuming the *relative risk* of smoking on coronary heart disease mortality is ~2 (*population excess fraction* ~20%), and again taking the *relative risk* of smoking on bladder cancer in the population of England to be 5 (*population excess fraction* ~50%), and how is it that more deaths from coronary heart disease are attributed to smoking than bladder cancer *cases*?
Coronary heart disease mortality is a much more common in the population than bladder cancer incidence, so although 50% of bladder cancer cases can be attributed to smoking (~5000 cases

per year in England), 20% of coronary heart disease deaths in England amounts to the much larger figure of ~20,000 per year.
The population excess fraction is not just influenced by the relative risk associated with exposure, it is also dependent on the prevalence of exposure in the population being studied, as well as on the underlying incidence of the disease in the population.

h) What is the most useful measure of *risk* – the relative or the absolute (excess fraction) *risk*?
It depends on what you want to know. If you want to identify the risk factors for a disease, the relative measure tells you what you need to know (i.e. how many times more likely are the exposed compared with the non exposed people to develop the outcome of interest). If you can establish that an exposure is causally associated with a disease, you can then think about the impact of exposure on the incidence of the disease in the population (attributable or excess risk).

There are several limitations to reporting population attributable risks and excess fractions. You need to know that the exposure is causally related to the outcome of interest; you need to know that there is no bias or confounding that might influence the risk measure; and if you want to extrapolate the findings from a specific cohort, you also need to be sure that the study population is generalisable to the wider population. It is difficult to satisfy all these criteria, and as a result attributable risks should be looked upon as a best guess of the impact of exposure. You should also be aware that by calculating population excess fractions individually (e.g. for smoking, occupational exposure, diet), and ignoring the fact that many risk factors interact with each other, the percentages can add up to more than 100%. In the occupational cohort above, 98.7% of the bladder cancer cases were attributed to aromatic amine exposure; however 50.7% of the same cases were attributed to smoking!

References

Beral, V. & Million Women, S. C. 2003, "Breast cancer and hormone-replacement therapy in the Million Women Study.[see comment][erratum appears in Lancet. 2003 Oct 4;362(9390):1160]", *Lancet*, vol. 362, no. 9382, pp. 419-427.

Malarcher, A. M., Schulman, J., Epstein, L. A., Thun, M. J., Mowery, P., Pierce, B., Escobedo, L., & Giovino, G. A. 2000, "Methodological issues in estimating smoking-attributable mortality in the United States", *American Journal of Epidemiology*, vol. 152, no. 6, pp. 573-584.

McManus RJ, Mant J, Roalfe A, Oakes RA, Bryan S, Pattison HM, Hobbs FDR. Targets and self monitoring in hypertension: randomised controlled trial and cost effectiveness analysis. Br Med J 2005 (Sep); 331: 493-498

Richiardi, L., Forastiere, F., Boffetta, P., Simonato, L., & Merletti, F. 2005, "Effect of different approaches to treatment of smoking as a potential confounder in a case-control study on occupational exposures", *Occupational & Environmental Medicine*, vol. 62, no. 2, pp. 101-104.

Varas-Lorenzo, C., Garcia-Rodriguez, L. A., Perez-Gutthann, S., & Duque-Oliart, A. 2000, "Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case-control study", *Circulation*, vol. 101, no. 22, pp. 2572-2578.