Biostatistics (MATH11230)
Study designs

Vanda Inácio

University of Edinburgh

Semester 1, 2022/2023

Study designs

General context

$\hookrightarrow$ In this part we follow closely Chapter 5 of Jewell (2003).

$\hookrightarrow$ The two measures of disease-exposure association we learned (RR and OR) applied to properties of a population.

$\hookrightarrow$ However, in almost all situations, we do not have access to the entire population of interest and, as such, we are restricted to deal with samples with limited and varying information.

$\hookrightarrow$ The question of how to estimate a measure of association from a **random sample** of the population of interest then arises.

$\hookrightarrow$ The answer to this question depends on how the random sample was drawn from the population of interest.

Study designs
Random sample (...reminder...)

↪ Suppose we are interested in sampling *n* individuals from a given population of interest composed of $N(> n)$ subjects. If the *n* individuals are selected *at random*, then the sample is designated a *random sample*.

↪ *At random* implies that although the investigator specifies the sample size *n*, they do not predetermine which *n* individuals will be selected.

Study designs
Nested components of a population

↪ Before we learn about study designs, let us briefly discuss the nested components of the population of interest.

↪ The **target population** is the population to which we wish to apply our estimates and inferences regarding the relationship between disease (or, more generally, the health outcome of interest) and exposure to some risk factor.

↪ Sometimes, it may be (extremely) difficult to sample directly from the target population.

↪ In such cases, the specific population from which data are collected is the **study population**.

↪ The **sample** comprises the actual sampled individuals from the study population for whom data are collected on disease status, exposure, and other factors.

↪ Often, the study population is a very large fraction of the target population, while the sample is substantial smaller than the study and target populations.

Study designs
Nested components of a population

$\hookrightarrow$ Carefully choice of the study population will allow investigating the association of exposure with the disease.

$\hookrightarrow$ If the study population is not representative of the target popuation with regard to the exposure-disease relationship of concern, then the estimate of the association measure between E and D may be biased. This bias is known as **selection bias**.

$\hookrightarrow$ If the sample we end up working with is randomly selected from the study population, then we can apply our inferences to the study population.

$\hookrightarrow$ But we want more, we want to apply our inferences from the sample to the target population. This is only possible if the study population is representative, with regard to the $E - D$ relationship, of the target population (in addition to the sample being randomly selected from the study population).

Study designs

Selection bias

Study designs

Nested components of a population: exercise (adapted from Kleinbaum et al. , 2006, Chapter 7)

Consider an epidemiological study carried out in New York city (NYC) to assess whether obesity is associated with hypertension in young adults. The investigator decided that it was not feasible to consider taking a sample from among all young adults in the city. It was decided that fitness centres would provide a large source of young NYC adults. A sample of subjects is taken from several randomly selected fitness centres throughout the city and then blood pressure is measured to determine hypertension status and the body mass index of the individuals was also recorded.

1. What is a possible target population for this study?

2. What is the study population in this study?

3. Does the sample represent the study population?

4. Does the study population represent the target population?

Study designs

Nested components of a population: exercise (adapted from Kleinbaum et al. , 2006, Chapter 7)

1. Target population: all young adults in NYC.

2. Study population: all young adults who attended fitness centres in NYC.

3. Yes, because the sample is randomly selected from the study population.

4. Maybe not. It is not to hard to imagine that the group of all young adults attending fitness centres in NYC may be very different (probably healthier) from the group of all young adults in NYC.

Study designs

$\hookrightarrow$ We are now confronted with the task of how to obtain a random sample from the study population.

$\hookrightarrow$ The three most commonly used sampling schemes are:

   $\hookrightarrow$ Population-based studies.

   $\hookrightarrow$ Cohort studies

   $\hookrightarrow$ Case-control studies.

$\hookrightarrow$ As we did before, attention will be restricted to the association between the presence and absence of two binary factors: the outcome D and the exposure E.

$\hookrightarrow$ As a result, the sample data from any of the three study designs can be summarised by a 2×2 contingency table of the following form

	D	not D
E	-	-
not E	-	-

Study designs
Population-based studies

↪ The two main steps of a **population-based** design are:

   1 Take a random sample from the study population.

   2 Subsequently, measure the presence and absence of both D and E for all sampled individuals.

↪ Note that in a population-based design the participants are selected without regard to exposure or disease status.

↪ Data arising from a population-based study allow estimation of all probabilities of interest, namely:

   ↪ Joint probabilities (e.g., $\Pr(D\&E)$, $\Pr(D\&\text{ not } E)$, etc.).

   ↪ Marginal probabilities ($\Pr(D)$ and $\Pr(E)$).

   ↪ Conditional probabilities ($\Pr(D \mid E)$, $\Pr(D \mid \text{not E})$, $\Pr(E \mid D)$, $\Pr(E \mid \text{not } D)$).

↪ All these population probabilities are simply estimated by the corresponding observed proportion of the random sample.

Study designs
Population-based studies

$\hookrightarrow$ The following example is from Jewell (2003, p 48).

$\hookrightarrow$ Remember the running example from week 1, about infant mortality in 1991 in the USA. The risk factors considered (separately) were the mother's marital status (unmarried vs married) and the birth weight of the baby (low vs normal weight).

$\hookrightarrow$ A follow-up question is the extent to which the impact of marital status on infant mortality might be explained by birth weight.

$\hookrightarrow$ A possible hypothesis is that unmarried women may receive less good nutrition and pre natal care than married mothers-to-be and therefore deliver, on average, lower birth weight babies which, in turn, leads to a considerable increase in the risk of infant mortality.

Study designs
Population-based studies

$\hookrightarrow$ In order to examine the hypothesised relationship between marital status and birth weight, data needs to be collected on these two factors in the population of interest.

$\hookrightarrow$ Here low birth weight would be the outcome of interest and the marital status of the mother at the time of birth would act as the exposure.

$\hookrightarrow$ Let us suppose that a sample size of 200 has been chosen for a population-based study, that is, a random sample of 200 births was selected from the study population.

$\hookrightarrow$ The weight of the baby at birth and the marital status of the mother (at the time of birth) were then ascertained for all 200 babies in the sample and the information obtained could then by summarised by a 2×2 contingency table as the one in slide 7.

$\hookrightarrow$ The measures of association (RR and OR) could then be estimated and conclusions about a possible association drawn.

Study designs
Exposure-based sampling: cohort studies

↪ The main feature of a **cohort** study is that sampling is conducted separately for the exposed and unexposed subpopulations, thus leading to two distinct cohorts.

↪ The three main steps of a cohort design are:

1. Identify two subgroups of the population on the basis of the presence or absence of E.

2. Take a random sample from the exposed and unexposed groups separately.

3. Measure subsequently the presence and absence of D for the individuals in both random samples.

↪ The key statistical property of a cohort design is the separate identification and sampling of the two exposure groups.

↪ Out of curiosity: in Roman times, a cohort was a military unit, typically comprising 480 soldiers.

Study designs
Exposure-based sampling: cohort studies

↪ Note that the investigator has to prespecify the sizes of the samples from the exposed and unexposed groups (and they do not need to be equal).

↪ This is important in determining the amount of information that a cohort study yields on the disease-exposure relationship.

↪ As an extreme example, think that if one exposure group is allocated a very small sample size, then there will be little information available on the disease-exposure relationship.

Study designs
Exposure-based sampling: cohort studies

$\hookrightarrow$ Let us consider again the example aiming to study the possible association between a mother's marital status and birth weight.

$\hookrightarrow$ We now illustrate a possible cohort design to study the hypothesised relationship.

$\hookrightarrow$ To that end, the investigator has selected two random samples, one from the subpopulation of unmarried mothers and another from the subpopulation of married mothers (at the time of birth).

$\hookrightarrow$ This is by opposition to the population-based design where the investigator has selected the sample of births from the study population without regard to the baby's weight or marital status of the mother at the time of birth.

Study designs
Exposure-based sampling: cohort studies

↪ Data arising from a cohort study cannot be used to estimate population joint and marginal probabilities.

↪ For instance, in the context of our running example, the frequency of unmarried mothers (at birth time) with low birth weight babies is artificially influenced by the preselected number of unmarried mothers and so it cannot be used to estimate the corresponding population probability $\Pr(D\&E)$. A similar reason applies to all other joint ptobabilities.

↪ For the same reason, the population marginal probability of unmarried mothers and the population marginal probability of married mothers cannot also be estimated, that is, we cannot estimate $\Pr(E)$ and $\Pr(\text{not } E)$.

↪ If the investigator selected, e.g., 100 married mothers and 100 unmarried mothers, then the proportion of unmarried mothers in our sample will be 50%, not because 50% of the mothers-to-be in our study population were unmarried but because the investigator specifically has chosen a sample in which 50% of the mothers were unmarried.

Study designs
Exposure-based sampling: cohort studies

$\hookrightarrow$ It may be less obvious why we cannot estimate the population marginal probabilities of low and normal birth weight, that is, why we cannot estimate $\Pr(D)$ and $\Pr(\text{not } D)$.

$\hookrightarrow$ The reason is that if there is as an association (positive or negative) between marital status and birth weight, then the preselected number of married/unmarried women will also artificially influence the number of low and normal birth weight babies.

$\hookrightarrow$ For instance, if unmarried mothers are more likely to deliver low birth weight babies than those who are married then, the proportion of low birth weight babies in the combined sample will change according to the way the sampling scheme fixes the proportions of married and unmarried mothers.

Study designs
Exposure-based sampling: cohort studies

↪ In a cohort study, only conditional probabilities that condition on the exposure status can be estimated, i.e., $\Pr(D \mid E)$ and $\Pr(D \mid \text{not } E)$.

↪ For instance, we can estimate the probability of low birth weight given that the mother is unmarried.

↪ The fact that we can estimate the conditional probabilities that condition on the exposure status E, allow us to estimate the two association measures we have learned: the relative risk and the odds ratio.

Study designs
Further considerations

$\hookrightarrow$ In cohort studies, the two groups of individuals (sampled from a population at risk for the disease of interest) are usually followed over time and the occurrences of the disease outcome D in each group are identified. This allows investigating etiological factors of disease.

$\hookrightarrow$ For cohort studies with long periods of follow-up, there are issues concerning due to losses to follow-up and also the exposure status of some individuals may change as well (without the investigator setting the study noticing it).

$\hookrightarrow$ Also, for rare conditions, very large cohorts or very long monitoring periods are required to observe any cases. Case-control studies (see next slide) are to be preferred in this case.

Study designs
Diseased-based sampling: case-control studies

↪ By opposition to a cohort study, in a **case-control** study separate samples are selected from the cases (*D*) and the nondiseased individuals or controls (not *D*).

↪ The essential steps of a case-control study are:

1. Identify two subgroups of the population on the basis of the presence or absence of *D*.

2. Take a random sample from the cases and control groups separately.

3. Measure subsequently the presence and absence of *E* for the individuals in both random samples.

↪ As it was the case for cohort studies, in a case-control study the investigator must prespecify the number of cases and controls in each random sample.

Study designs
Diseased-based sampling: case-control studies

↪ In the context of our example, implementing a case-control design would involve taking one random sample of low birth weight babies and another sample of normal weight babies.

↪ For similar reasons as in cohort designs, joint and marginal probabilities cannot be estimated, as these would be artificially influenced, by the prespecified number of low birth and normal birth weight babies.

↪ In case control-studies only conditional probabilities that condition on the disease outcome status D, here infant birth weight, can be estimated. That is, we can estimate, for instance, $\Pr(E \mid D)$ and $\Pr(E \mid \text{not } D)$.

↪ As a result, the RR cannot be estimated from a case-control study because it involves $\Pr(D \mid E)$ and $\Pr(D \mid \text{not } E)$, which cannot be estimated in such a design.

Study designs
Diseased-based sampling: case-control studies

$\hookrightarrow$ For the same reason, we cannot also estimate the odds ratio for D associated with E, that is, we cannot estimate $OR_{D|E}$.

$\hookrightarrow$ However, as we have learned, the OR is symmetric with respect to the roles of D and E, i.e., $OR_{D|E} = OR_{E|D}$.

$\hookrightarrow$ Because all conditional probabilities involved in $OR_{E|D}$ condition on D, which are estimable from a case-control study, we can thus estimate $OR_{D|E}$.

$\hookrightarrow$ Further, when the outcome D is rare in both the exposed and unexposed populations, we have seen that $OR \approx RR$, so that the case-control estimate of the OR can be used as an approximate estimate of the RR (if the disease is rare).

Study designs
Diseased-based sampling: case-control studies

$\hookrightarrow$ The selection of the controls is a tricky issue in case-control studies.

$\hookrightarrow$ The controls should be individuals who do not have the disease in question but who are otherwise comparable to the cases. Two possible sources for controls are the study population and patients with other diseases.

$\hookrightarrow$ The latter is usually more practical because of savings costs and accessibility. However while it is easier to use patients with a different disease as controls, there may be bias introduced because prevalence of the exposure in such a group of individuals is likely to be different from the exposure in the general population of controls.

Study designs
Further considerations

↪ There are two main types of epidemiologic studies: **experimental** and **observational**.

↪ Population-based, cohort, and case-control studies, all form part of observational studies.

↪ An experimental study uses randomisation to allocate subjects to the exposure groups. In an experiment, we make some intervention and observe the result.

↪ Ethical and cost reasons restrict most epidemiologic research to observational studies.

↪ By opposition, in an observational study, we *observe* the existing situation and try to understand what is happening.