Màster universitari en Estadística i Investigació Operativa

Epidemiology, Course 2021/22

Final exam

1st June, 2022, 15h–18h

Post-exam review:
Friday, 10th June, 2022, 12h–13h.
Room 214, C5 building, North Campus (UPC).

Note: Exercises can be answered in either English, Spanish, or Catalan.

Exercise 1 (3 Points)

(a) (0.4 + 0.4 + 0.2 = 1 point)

Which type of study design would you use to

(i) estimate the prevalence of COVID-19 in a population of interest,

(ii) estimate the incidence of COVID-19 in a population of interest?

Justify your answer explaining the characteristics of the corresponding study designs.

(iii) What inconveniences do both study designs have?

(b) (0.4 + 0.4 + 0.2 = 1 point)

A confounder can introduce confounding (bias) in the estimation of the exposure-disease association measure of interest.

(i) Which is the definition of a confounder?

(ii) Give a fictitious example of confounding in a cohort study.

(iii) In your example, how could confounding be avoided

- when designing the study,
- when analyzing the data?

(c) (0.2 + 0.2 + 0.3 + 0.3 = 1 point)

The crude mortality rate (MR) is the number of deaths divided by population size per unit of time (most often, 1 year). In order to compare the mortality rates of two populations, say A and B, we can use the (crude) mortality rate ratio: $\text{MRR} = \text{MR}_A/\text{MR}_B$.

(i) In which situation would you use this measure?

An alternative measure is the standardized mortality ratio: $\text{SMR} = O/E$.

(ii) What do O and E stand for?

(iii) In which situation would you use this measure to compare mortality in populations A and B?

(iv) If you wanted to compare the mortality rates of more than two populations, which measure would you use and why?

Exercise 2 (4 Points)

(a) (0.6 points)
Which are the following epidemiological measures?

(i) The number of people with HIV worldwide in 2021 divided by the world population in 2021.

(ii) The number of new adult hepatitis C cases in Spain in 2021 divided by the midyear population in Spain in 2021.

(iii) The number of newborn babies with congenital malformation in Europe in 2020 divided by the total number of newborn babies in Europe in 2020.

(iv) The proportion of women that suffer from postpartum depression during the first eight months after giving birth to their babies.

(b) (0.7 points)
In their scientific paper *Prediction of Psychosis in Adolescents and Young Adults at High Risk*[1], Ruhrmann *et al.* present results from the prospective European Prediction of Psychosis Study (EPOS). The measure of interest is the incidence of psychosis among young adults at high risk. In the abstract, the authors say:

> *"At 18-month follow-up, the incidence rate for transition to psychosis was 19%."*

Moreover, in the paragraph on the statistical analyses, they authors explain that the incidence rate after 18 months has been estimated as the probability of having suffered psychosis during the 18 months of follow-up.
Do you consider that the term *incidence rate* has been correctly employed? If not, which measure do you think the authors actually present. Justify your answer.

[1] Ruhrmann *et al.* (2010). Prediction of Psychosis in Adolescents and Young Adults at High Risk. *Archives of General Psychiatry*, 67(5), 241–251.

(c) (0.3 + 0.3 + 0.3 + 0.3 = 1.2 points)

The following table shows the (fictitious) data of a cohort study on the relation between a disease and an exposure of interest.

Table 1: Fictitious data of a cohort study (Exercise 2c).

	Disease		
Exposure	Yes	No	**Total**
Yes	250	99750	100000
No	125	199875	200000
Total	375	299625	300000

(i) Estimate the relative risk and interpret its value.

(ii) Which would be the value of the population attributable risk (PAR) if the exposure prevalence in the population was 25%?

(iii) Given the data of Table 1, which is the possible maximum of the PAR? Is that value meaningful?

(iv) The attributable risk of the exposed (EAR) is the proportion of disease cases among exposed people that could be avoided if nobody was exposed. Which is the value of the EAR?

Note: The PAR can be calculated as follows:

$$\mathrm{PAR} = \frac{\mathrm{P}(E)(\mathrm{RR}-1)}{1+\mathrm{P}(E)(\mathrm{RR}-1)}.$$

(d) (0.4 + 0.4 = 0.8 points)

In a study on the possible interaction between an exposure of interest (E) and a certain characteristic (C) with respect to the risk of disease D, the following values are obtained:

$$\text{RERI} = 0, \qquad \frac{\text{RR}_{E,D|C}}{\text{RR}_{E,D|\bar{C}}} = 2. \tag{1}$$

(i) What do you conclude from both values about the interaction between E and C?

(ii) Given the values in (1), fill in the missing values in the table corresponding to $\bar{C}$.

	C			$\bar{C}$		
	D	$\bar{D}$	**Total**	D	$\bar{D}$	**Total**
E	20	80	100			100
$\bar{E}$	5	95	100			100

(e) **(0.5 + 0.2 = 0.7 points)**

In a matched case-control study on a certain type of cancer, 1:3-matching was used, that is, for each case, 3 controls of the same sex were chosen for the study. The values concerning exposure (E) of the 70 case-controls groups are shown in Table 2.

Table 2: Data of a matched case-control study using 1:3-matching (Exercise 2e).

Type	**Exposure pattern**	**Number**
A	Case is exposed & 3 controls are exposed	10
B	Case is exposed & 2 controls are exposed	12
C	Case is exposed & 1 control is exposed	12
D	Case is exposed & No control is exposed	14
E	Case is not exposed & 3 controls are exposed	2
F	Case is not exposed & 2 controls are exposed	7
G	Case is not exposed & 1 control is exposed	6
H	Case is not exposed & No control is exposed	7
Total		70

To study whether E is a risk factor for the cancer under study, the odds ratio is used.

(i) Use the Mantel-Haenszel estimator to estimate the odds ratio with the data of Table 2.

(ii) Which value of $\widehat{\text{OR}}$ would be obtained if matching was ignored?

Exercise 3 (3 Points)

For this exercise, we use the data set of the detoxication unit of the Hospital Germans Trias i Pujol in Badalona. It contains the (cross-sectional) data of 387 hepatitis C-infected injection drug users that were admitted to the unit between 1994 and 2004. The outcome of interest is liver inflammation (determined by the levels of the enzyme alanine transaminase) and the variables considered in the following models are sex (Female/Male), HIV infection (No/Yes), and cholesterol level (measured in mg/dl).

The following tables show the parameter estimates of two logistic regression models obtained with R: the first model includes variables sex and HIV infection as well as their interaction (Table 3(a)), and the second model does also contain the cholesterol level (Table 3(b)).

Table 3: Logistic regression models for liver inflammation.

(a) Model including sex and HIV infection.

	$\hat{\beta}$	$s.e.(\hat{\beta})$	Z	p	$\exp(\hat{\beta})$
Intercept	-1.145	0.434	-2.64	0.008	0.318
Males	1.727	0.467	3.7	< 0.001	5.624
HIV+	0.817	0.533	1.53	0.125	2.264
Males×VIH+	-1.398	0.581	-2.41	0.016	0.247

(b) Model including sex, HIV, and cholesterol level.

	$\hat{\beta}$	$s.e.(\hat{\beta})$	Z	p	$\exp(\hat{\beta})$
Intercept	-0.379	0.706	-0.54	0.591	0.685
Males	1.671	0.469	3.56	< 0.001	5.317
HIV+	0.742	0.536	1.38	0.166	2.1
Cholesterol	-0.004	0.003	-1.37	0.172	0.996
Males×HIV+	-1.394	0.582	-2.39	0.017	0.248

(a) (0.7 points)

Give the (theoretical) expression of the logistic regression model in Table 3(b). Introduce all the notation needed. Which of the model parameters does have a unit and which one is it?

(b) **(0.6 points)**

Give an interpretation of the following values in the last column in Table 3(a): $\exp(-1.145) = 0.318$, $\exp(0.817) = 2.264$, and $\exp(-1.398) = 0.247$.

(c) **(0.6 points)**

Which are the estimated (prevalence) odds ratios associated with the comparison of a female HIV-positive injection drug user and a male HIV-negative injection drug user according to both models in Table 3? Give an interpretation of both estimates.

(d) **(0.4 points)**

Interpret the output of function `HLtest` explaining first which is the hypothesis under study. What can you conclude?

```
> library(vcdExtra)
> HLtest(lrmod2)

Hosmer and Lemeshow Goodness-of-Fit Test

Call:
glm(formula = altHigh ~ sex * hiv + colester, family = "binomial", 
    data = exam22)
 ChiSquare df   P_value
  4.122709  8 0.8458859
```

(e) **(0.2 + 0.3 + 0.2 = 0.7 points)**

In their paper in the Journal of the American Medical Association, Zhang and Ju[2] propose the following formula to estimate the relative risk given an estimation of the odds ratio:

$$\mathrm{RR} = \frac{\mathrm{OR}}{(1 - \mathrm{P}_0) + (\mathrm{P}_0 \times \mathrm{OR})},$$

where P_0 is the disease incidence (prevalence) among nonexposed people in a cohort (cross-sectional) study. According to the authors, replacing OR by $\widehat{\mathrm{OR}}$, yields and estimation of the adjusted relative risk (adjusted prevalence ratio).

In the model in Table 3(b), the estimation of the odds ratio associated to HIV infection among women is $\exp(0.742) = 2.1$.

(i) Apply the formula of Zhang and Ju to estimate the adjusted prevalence ratio knowing that 7 out of 29 HIV-negative women had a liver inflammation.

(ii) What is wrong with the interpretation that the value obtained is the adjusted prevalence ratio.

(iii) What would you do to estimate the adjusted prevalence ratios associated with the variables of the second model?

[2] Zhang and Ju (1998). What's the Relative Risk? *Journal of the American Medical Association*, 280(19), 1690–1691.