# Exercises day 5

Basic Statistics for health researchers 2022

14 November 2022

## Warming up

Before starting the exercise below, learn from the R-demo of Lecture 5 (available from the course webpage):

1. Read and run the code.
2. Check that the output matches the results presented on the slides.
3. Do not hesitate to add your own comments into the script.

## Exercise A (inference in 2 by 2 tables, subgroup analyses)

For this exercise we will work with the "Smoking" data (available from the course webpage).

### Part I

In this first part we assume that we would like to (naively) study the (overall) association between smoking and 20-year survival.

#### Question 1

Read the description of the data and load the data into R. Visualize the first lines of the data and a summary of the data. How many women were still alive 20 years after the initial survey? How many women were smoking at the time of the initial survey?

#### Question 2

Use the `table()` function to compute a "2 by 2" table and investigate the association between smoking and survival status.

1. How many smokers were alive and dead 20 years after the initial survey?
2. How many non smokers were alive and dead 20 years after the initial survey?

3. Estimate the 20-year risk of death for smokers and non smokers.

4. Are the results surprising?

**Question 3**

1. Compute confidence intervals for the 20-year risk of death for smokers and non smokers.

2. Do these first results "suggest" that the direction of the results can likely be explained by small sample random variation (i.e. "chance")?

**Question 4**

1. Use a statistical test to investigate whether the smoking status seems significantly associated with the 20-year survival outcome. Conclude.

2. Use the function `table2x2()` from the `Publish` package to estimate each of the following association measure, with confidence intervals:

   - Risk difference
   - Risk ratio
   - Odds ratio

3. Write a conclusion sentence using the risk ratio to describe the association between smoking and survival.

**Question 5**

Produce a barplot to compare the number of included women in each age group between smokers and non smokers. What do you observe? **Hint:** you can use the `beside=TRUE` option in the function `barplot()`.

**Question 6**

Produce a barplot to compare the 20-year survival probability in each age group. What do you observe?

**Question 7**

Produce a barplot to compare the 20-year survival probability in each age group between smokers and non smokers. What do you observe?
**Hint:** you could first create an appropriate 2 by 4 table (one row per smoking group, one column per age group), and then use the `barplot()` function with this table. In the R-demo there is an example of how to create a table "from scratch". This might help you to put all the relevant numbers together in the same table.

### Question 8

Can you now provide a plausible explanation for the results obtained at question 3?

## Part II (if time allows)

In this second, part we assume that we would like to study the association between smoking and 20-year survival separately within each age group.

### Question 9

Compute the risk ratio and the corresponding confidence intervals to describe the association in each age group. Conclude.

### Question 10

Why do the results of Part II seem more interesting that those of Part I?

# Exercise B (Sample size and power calculation)

Assume that we aim to study a new fertility treatment. We plan to compare the chances of pregnancy within 6 months. Assume that, according to previous studies, we can reasonably think that without fertility treatment the chance of pregnancy within 6 months is approximately 40%, for the (specific) population of women that we study.

## Question 1

How many women should we enroll in our study to show a significant results if we want a statistical power of 85% and if we expect that the treatment results in 60% chance of pregnancy. As usual, we plan to control the risk of false positive finding at 5%.

## Question 2

Using the sample size suggested by the result to the previous question:

1. What does the power become if the treatment is actually less effective that we thought and results in only 50% or 55% chance of pregnancy?
2. What is the smallest chance difference that we can hope to show with a "decent" power of 75%?

# Exercise C (Case-control study)

## Part I

In a Case-control study by Carpenter et al.[1], the following results were observed, when comparing cases (i.e. Sudden unexplained infant deaths) to controls (i.e. live infants of the same age, living in the same survey area at the time) between infants born from multiple birth versus singleton.

| | Case | Control | total |
|---|---|---|---|
| Multiple birth | 47 | 36 | 83 |
| Singleton | 687 | 2364 | 3051 |
| total | 734 | 2400 | 3134 |

### Question 1

Estimate the odds ratio for sudden infant death of multiple birth versus singleton with 95% confidence interval using the data of the above Table.

## Part II

The study presented in Carpenter et al. recruited roughly 3 controls for each case. The results section states that in Nordrhein-Westfalen (Germany) the incidence of sudden unexplained infant death is roughly 1 case per 1000 births. Assuming this incidence, a cohort study would have to include approximately 734,000 subjects to observe a comparable number of cases. Consider the following hypothetical 2x2 contingency table, that we could expect from such a cohort study.

| | Case | Control | total |
|---|---|---|---|
| Multiple birth | 47 | 10,999 | 11,046 |
| Singleton | 687 | 722,267 | 722,954 |
| total | 734 | 733,266 | 734,000 |

### Question 2

Estimate the odds ratio and compute a 95% confidence interval using the data of the above Table. Compare the results to those obtained from the first table.

### Question 3

For each of the above two tables, compute a risk ratio to (naively) investigate the association between Sudden unexplained infant deaths and type of birth. Compare the results and explain what you see.

[1] Carpenter, R. G., et al. "Sudden unexplained infant death in 20 regions in Europe: case control study." The Lancet 363.9404 (2004): 185-191.