

Chapter 2

A Few Data Sets and Research Questions

Goals of this chapter

This chapter presents a few data sets from epidemiological studies analyzed by various teams at the Bordeaux School of Public Health (Institut de Santé publique, d'Epidémiologie et de Développement—ISPED). Each data set comes with a short research question, which will help understand the context of the study. They will be used throughout this book to show how to use the functionalities of **R** for importing and manipulating data and performing appropriate statistical analyses. For each data set, we give a table with a description, the variables and the coding. The reader should refer to this chapter when the data sets are mentioned later in the book. A table at the end of the chapter indicates in which chapters each data set is used. All these data sets are available online on the website associated with the book: <http://www.biostatisticien.eu/springer>.

SECTION 2.1

Body Mass Index of Children

Presentation

A sample of 152 children (3 or 4 years old) in their first year of kindergarten in schools in Bordeaux (Gironde, SouthWest France) underwent a physical check-up in 1996–1997.

Variables and Coding

Description	Unit or coding	Variable
Gender	F for female; M for male	GENDER
School in an underprivileged area (<i>zone d'éducation prioritaire</i> , ZEP)	Y for yes; N for no	zep
Weight	Kg (to the nearest 100 g)	weight
Age at date of examination	Years	years
Age at date of examination	Months	months
Height	Cm (to the nearest 0.5 cm)	height

Data Set: BMI-CHILD

File: bmichild.xls

SECTION 2.2

Weight at Birth

Presentation

This study focused on risks associated with low weight at birth; the data were collected at the Baystate Medical Centre, Massachusetts, in 1986. Physicians have been interested in low weight at birth for several years, because underweight babies have high rates of infant mortality and infant anomalies. The behaviour of the mother-to-be during pregnancy (diet, smoking habits) can have a significant impact on the chances of having a full-term pregnancy, and thus of giving birth to a child of normal weight. The data file includes information on 189 women (identification number: ID) who came to the centre for consultation. Weight at birth is categorized as low if the child weighs less than 2,500 g.

Variables and Coding

Description	Unit or coding	Variable
Age of mother	Years	AGE
Weight of mother at last menstrual period	Pounds	LWT
Race of mother	1 = white; 2 = black; 3 = other	RACE
Smoking during pregnancy	Yes = 1; no = 0	SMOKE
Number of premature births in medical history	0 = none; 1 = one; 2 = two; etc.	PTL
Medical history of hypertension	Yes = 1; no = 0	HT
Uterine irritability	Yes = 1; no = 0	UI
Number of medical consultations during first trimester	0 = none; 1 = one; etc.	FVT
Weight at birth	Grams	BWT
Weight at birth less than 2,500 g	Yes = 1; No = 0	LOW

Data Set: WEIGHT-BIRTH

File: Birth_weight.xls

SECTION 2.3

Intima–Media Thickness**Presentation**

Atherosclerosis is the main cause of death for men above 35 and women above 45 in most developed countries. It is a thickening and hardening of internal artery walls. One of its consequences is myocardial infarction. An artery wall is made of three layers; innermost to outermost, they are called intima, media and adventitia. Intima–media thickness is a marker of atherosclerosis. It was measured by ultrasonography on a sample of 110 subjects in 1999 in Bordeaux hospitals. Information on the main risk factors was also collected.

Variables and Coding

Description	Unit or coding	Variable
Gender	1 = male; 2 = female	GENDER
Age at date of consultation	Years	AGE
Height	Cm	height
Weight	Kg	weight
Smoking status	0 = non-smoker 1 = former smoker 2 = smoker	tobacco
Estimation of tobacco consumption for smokers and former smokers	Number of packs/year	packyear
Physical activity	0 = no; 1 = yes	SPORT
Intima–media thickness	Mm	measure
Alcohol consumption	0 = non-drinker 1 = occasional drinker 2 = regular drinker	alcohol

Data Set: INTIMA–MEDIA**File:** Intima_Media_Thicness.xls

SECTION 2.4

Diet of Elderly People

Presentation

A sample of 226 elderly people living in Bordeaux (Gironde, South-West France) were interviewed in 2000 for a nutritional study.

Variables and Coding

Description	Unit or coding	Variable
Gender	2 = female; 1 = male	gender
	1 = single	
Family status	2 = living with spouse	situation
	3 = living with family	
	4 = living with someone else	
Daily consumption of tea	Number of cups	tea
Daily consumption of coffee	Number of cups	coffee
Height	Cm	height
Weight	Kg	weight
Age at date of interview	Years	age
	0 = never	
	1 = less than once a week	
Consumption of meat	2 = Once a week	meat
	3 = 2/3 times a week	
	4 = 4/6 times a week	
	5 = every day	
Consumption of fish	Idem	fish
Consumption of raw fruits	Idem	raw_fruits
Consumption of cooked fruits and vegetables	Idem	cooked_fruits_veg
Consumption of chocolate	Idem	chocol
	1 = butter	
	2 = margarine	
	3 = peanut oil	
Type of fat used for cooking	4 = sunflower oil	fat
	5 = olive oil	
	6 = mix of vegetable oils (e.g., Isio4)	
	7 = colza oil	
	8 = duck or goose fat	

Data Set: NUTRIELDERLY

File: nutrition_elderly.xls

SECTION 2.5

Study Case of Myocardial Infarction

Presentation

The study for which the following data were collected aimed at examining whether women who use or have used oral contraceptives are at a higher risk of myocardial infarction. The sample includes 149 women who had myocardial infarction (cases) and 300 women who did not (controls). The main exposure factor is usage of oral contraceptives; the data also include age, weight, height, tobacco consumption, hypertension and family history of cardiovascular diseases.

Variables and Coding

Description	Unit or coding	Variable
Myocardial infarction	0 = controls; 1 = cases	infarct
Usage of oral contraceptives	0 = never; 1 = yes	co
	0 = no	
Tobacco usage	1 = smoker	tobacco
	2 = former smoker	
Age	Years	age
Weight	Kg	weight
Height	Cm	height
Family history of cardiovascular diseases	0 = no; 1 = yes	atcd
Hypertension	0 = no; 1 = yes	hta

Data set: INFARCTION

File: Infarction.xls

SECTION 2.6

Summary Table of Use of Data Sets

	Methods				
	Import-export	Manipulation	Descriptive statistics	Tests	ANOVA Regression
Data sets	BMI-CHILD	×	×	×	
	WEIGHT-BIRTH	×	×		×
	INTIMA-MEDIA	×	×	×	×
	NUTRIELDERLY	×	×	×	
	INFARCTION	×	×		

The R Software

Fundamentals of Programming and Statistical Analysis

Lafaye de Micheaux, P.; Drouilhet, R.; Liquet, B.

2013, XXXVII, 628 p. 50 illus., 41 illus. in color.,

Hardcover

ISBN: 978-1-4614-9019-7