Biostatistics & Epidemiological Data Analysis using R

3

Descriptive statistics

Stefan Konigorski

Health Intervention Analytics Group, HPI

November 2, 2022

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2022.10.19
	2	First steps in data analysis using R	2022.10.26
	3	Second steps in data analysis using R	2022.11.02
Epidemiology & Statistics: concepts	4	Epidemiological study designs and study planning	2022.11.09
	5	Estimation	2022.11.16
	6	Hypothesis testing	2022.11.23
	7	Missing data	2022.11.30
Data analysis w/ regression models	8	Linear regression I	2022.12.07
	9	Linear regression II	2022.12.14
	10	Regression models for binary and count data	2023.01.11
	11	Analysis of variance & Linear mixed models I	2023.01.18
	12	Linear mixed models II & Meta analysis	2023.01.25
	13	Survival analysis	2023.02.01
	14	Causal inference & Data analysis challenge	2023.02.08

(see full schedule online)

1. Review and introduction
   - Review
   - Overview: Descriptive statistics

2. Descriptive statistics
   - Nominal variables
   - Ordinal & metric variables

Assignments

Reminder

You have to hand in at least 9 of 11 ("accepted") homework assignments, in order to be able to take the final open book take home exam and get credits for the class

Review of class 2

- Getting to know R Markdown
  $\longrightarrow$ Assignment 2, Exercise 1.
- Learned which data checks can be important and how to do them in R $\longrightarrow$ Assignment 2, Exercise 2.
- Learned and practiced how to manipulate objects in R in order to create new variables, transform variables, and select subsets of variables or observations $\longrightarrow$ Assignment 2, Exercises 2-4.

Review: data check

Examples

- Does anyone have age smaller than 0: `table(age < 0)`
- How many missing values does the variable age have: `table(is.na(age))`
- Are those people with BMI 0 the same people with insulin 0: `table((BMI == 0) & (insulin == 0))`

Review: data manipulation of variables

Examples

- Change variable type:
  `dat$lunchtime <- as.Date(dat$lunchtime)`
- Create new variable through mathematical operation:
  `dat$BMI <- dat$weight/(dat$height^2)`
- Remove/add/replace variable values:
  `dat$BMI[1] <- 20`

Review: data manipulation of variables

Assignment 2, Exercise 2

- Create new variable 'BMIOutlier', which has value 0 if a women has BMI $\leq$ 50, and 1 if she has BMI higher than 50.
  - First create the variable and set its values to 0 for everyone: `dat$BMIOutlier <- 0`
  - then set it to 1 for those with BMI higher than 50: `dat$BMIOutlier[dat$BMI > 50] <- 1`

  Alternatives:
  - `dat$BMIOutlier2 <- ifelse(dat$BMI > 50, 1, 0)`
  - Use `replace()` function or `as.numeric(dat$BMI > 50)`.
  - ...

Why might you want to create such a variable?
For example, to later filter out those people.

Review: data manipulation of data frames

Examples

- `dat[!dat$Age == 0, ]`
- `dat_female <- dat[dat$Gender == "F", ]`
- `dat_final <- data.frame(ID = dat_female$PatientId, Age = dat_female$Age, NoShow = dat_female$No-show)`

Note: `attach()` allows to call variables of data frame directly, `detach()` closes connection.
My suggestion: don't use it.

Main steps of a data analysis

1. Import dataset from an external file (e.g. xls, txt, SPSS file).
2. Import check: check if dataset has been read correctly.
3. Save dataset as R dataset (.Rdata), e.g. as `dat_raw.Rdata`.
4. Data check: check if data is correct/missing, and e.g. remove probands/variables or decide for imputation. Save corrected dataset as new dataset, e.g. `dat_corrected.Rdata`.
5. Transform variables, compute new variables, and/or select subset for final analysis. Save this again as new dataset, e.g. as `dat_final.Rdata`, and use in all further steps.
6. Descriptives to describe main characteristics of study sample.
7. Main analyses.
8. Secondary analyses.
9. Sensitivity analyses.

Descriptive statistics

Learning objectives

Leading questions

- How do you best describe the variables of a study?
- How do you present these descriptions?
- How do you do this in R/RStudio?

Learning goals

- Compute descriptive statistics.
- Create tables and plots to visualize them.
- Export tables and figures.

Descriptive statistics

Goal

- Describe your study sample regarding its main characteristics - i.e. regarding main personal/sociodemographic variables, covariates, and outcome/exposure variables.
- (Also think about: number and pattern of missing values.)

Descriptive statistics

Goal 1: for you

- Compute plots (and tables) to get an understanding of the characteristics of the variables, their distribution, and association between variables.

Goal 2: for presenting to others

- Describe the main characteristics of your study sample.
- Present these descriptive statistics in an easily accessible minimal table (or graphic).

Example 1

Study of change in wellbeing following work exit in 8037 persons:

Table 2. Descriptive statistics of individual-level variables for the analytic sample ($n = 8037$)

		SHARE		ELSA		Combined	
Variable	Categories	n	%	n	%	n	%
Total sample		6031	100	2006	100	8037	100.0
				0			
Route of exit from work	Old-age pension	2952	49.0	601	30.0	3553	44.2
	Disability pension	268	4.4	123	6.1	391	4.9
	Unemployment benefit	314	5.2	25	1.3	339	4.2
	Sickness benefit	106	1.8	6	0.3	112	1.4
	Social Assistance	34	0.6	6	0.3	40	0.5
	Early-retirement pension	590	9.8	0	0.0	590	7.3
	None	1767	29.3	1245	62.0	3012	37.5
Age at exit from work	>1 year before pensionable age	2631	43.6	1332	66.4	3963	49.3
	Pensionable age ±1 year	1799	29.8	347	17.3	2146	26.7
	>1 year after pensionable age	1601	26.6	327	16.3	1928	24.0
Country-specific quartile of household wealth	1 (poorest)	1090	18.0	228	11.4	1318	16.4
	2	1374	22.8	438	21.8	1812	22.6
	3	1742	28.9	618	30.8	2360	29.4
	4 (wealthiest)	1825	30.3	722	36.0	2547	31.7

(from Richardson et al., `https://doi.org/10.1093/ije/dyy205`)

Informative? Easy to understand? Has all relevant information?

Example 2

Association of adverse employment history and health in 31,718 persons:

Table 1. Sample description: observations (No.) and percentage (Col. %) or mean and standard deviation (SD), by sex ($n = 31\,718$)

		Men		Women	
	Categories or range	No.	Col. % or mean (SD)	No.	Col. % or mean (SD)
Age	45–60	15 134	52.7 (4.5)	16 584	52.6 (4.5)
Partnership[a]	Living with partner	11 910	80.5	11 761	72.4
	Living as single	2889	19.5	4492	27.6
Education[b]	Low	1470	9.9	1620	9.9
	Medium	6213	41.9	5948	36.5
	High	7129	48.1	8736	53.6
Current employment situation	In paid work	12 526	82.8	13 497	81.4
	Not in paid work	2608	17.2	3087	18.6
Current income[c]	Low income	4714	32.3	5321	33.4
	Medium income	4867	33.3	5106	32.1
	High income	4441	30.4	4677	29.4
	Answer refused	587	4.0	815	5.1

(from Wahrendorf et al., `https://doi.org/10.1093/ije/dyy235`)

Informative? Easy to understand? Has all relevant information?

Example 3

Association of diet and breast cancer in 571 postmenopausal UK women:

Informative? Easy to understand? Has all relevant information?

Table 1. Characteristics of 691 571 women at baseline, and details of follow-up

Characteristics	Mean (SD)
Personal	
Age at first dietary assessment, years	59.9 (4.9)
Height, m	1.62 (0.07)
BMI, kg/m^2	25.9 (4.4)
Number of full-term pregnancies	2.1 (1.2)
Foods and alcohol	
Meat g/d	55.7 (34.0)
⋮	
Macronutrients	
Energy kJ/d	6772 (1802)
Protein % energy	16.4 (2.7)
⋮	
Follow-up for breast cancer	
Person-years of follow-up per woman, mean (SD)	11.9 (3.0)
Incident breast cancers, *n*	29 005
ER+ve breast cancers, *n*	10 838
ER-ve breast cancers, *n*	1658

ER+ve, estrogen receptor-positive breast cancers; ER-ve, estrogen receptor-negative breast cancers; d, day.

(from Key et al., `https://doi.org/10.1093/ije/dyy238`)

Example 4

Metformin use and risk of cancer in patients with type 2 diabetes in 55,629 type 2 diabetes patients:

Table 1. Demographics of included patients from the CPRD at study entry

	No medication $N = 49\,524$	Metformin $N = 6105$	Total $N = 55\,629$
Mean (SD) median, 25th percentile-75th percentile)			
Age at diagnosis (years)	62.2 (12) 63, 54-71	57.6 (11.8) 57, 49-66	61.7 (12) 62, 53 -71
HbA1c (%) at study entry	7.2 (1.6) 6.8, 6.2-7.7	9.4 (2.3) 9, 7.4-11	7.5 (1.8) 6.9, 6.3-8
BMI (kg/m^2) at study entry	31.6 (6.3) 30.7, 27.3-34.9	33.4 (6.9) 32.3, 28.6-37.1	31.8 (6.3) 30.9, 27.5-35.2
N (%)			
Gender			
Male	27 763 (56.1)	3594 (58.9)	31 357 (56.4)
Female	21 761 (43.9)	2511 (41.1)	24 272 (43.6)
History of chronic kidney disease			
No	46 463 (93.8)	5866 (96.1)	52 329 (94.1)
Yes	3061 (6.2)	239 (3.9)	3300 (5.9)
History of cardiovascular disease			
No	41 868 (84.5)	5479 (89.7)	47 347 (85.1)
Yes	7656 (15.5)	626 (10.3)	8282 (14.9)
Use of statins in previous year			
No	25 035 (50.6)	2739 (44.9)	27 774 (49.9)
Yes	24 489 (49.4)	3366 (55.1)	27 855 (50.1)

(from Farmer et al., `https://doi.org/10.1093/ije/dyz005`)

Informative? Easy to understand? Has all relevant information?

Example

Main question (not directly investigated today)

- Does sending a reminder SMS have an effect on whether people come to their doctor appointment?

Data

- NoShow dataset.

Descriptive statistics - what do you do?

- Work with the final dataset for analysis (i.e. all missing values have been identified previously, dataset has been cleaned).
- Focus on the relevant variables (exposure, outcome, covariates) that you have specified at the beginning of your study.
- Explore them to get a feeling of the data.
- Depending on the scale level of each variable (nominal, ordinal, metric), choose the best metric to describe them.

- On which scales where the variables measured in the NoShow dataset?
- On which scales where the relevant variables measured in the NoShow dataset?

Descriptive statistics of nominal variables

Which descriptive statistics to use for nominal variables?

All about frequencies[1]

- absolute, relative, cumulative
- of each category, across categories
- of 1 variable or of multiple variables together (stratified)
- ...
- mode of a variable = value that occurs with the highest frequency.

[1] Note that for binary variables, mean = relative frequency of 1s.

Formal details on frequencies

- Let's look at a sample of n observations, and a nominal variable X.
- The observations $x_i, i = 1 \dots n$, hence all take either the value 0 or 1.

With this, we can formally describe frequencies:

- Absolute frequencies n_0, n_1: count the number of observations in each of the two categories 0 and 1. This can be formally written as:
  $n_0 = \sum_{i=1}^{n} \mathbf{1}_{x_i=0}$, $n_1 = \sum_{i=1}^{n} \mathbf{1}_{x_i=1}$,
  where the indicator function $\mathbf{1}$ takes the value 1 if the condition is satisfied and 0 otherwise.
- Relative frequencies of each category: $\frac{n_0}{n}, \frac{n_1}{n}$
- Cumulative frequencies: sum of the relative frequencies.

How to compute those in R?

In general

- Absolute frequencies: `table(var1)`
- Relative frequencies: `table(var1) / length(var1)`
- Further functions to create frequency tables: `prop.table()`, `janitor::tabyl()`, `summarytools::freq()`
- Frequencies of 2 variables: `table(var1, var2)`
- Alternative: `expss::cro(var1, var2)`

How to save the computed statistics/tables?

- Draft table in e.g. Word, copy/paste values from R manually.
- Create your table directly in R, export to e.g. csv or excel file (or by knitting directly to word/pdf file).
  - E.g. use `writexl::write_xlsx()`, `write.table()`, `write.csv()` functions, or functions in the `openxlsx` package.
- Create your report through R Markdown, and generate the tables/figures there directly.
  - E.g. use `summary_table()` functions in `qwraps2` package to generate formatted tables.
  - Note: if you knit to a Word file, then you have a formatted and editable table in Word! :-)
  - See `https://cran.r-project.org/web/packages/qwraps2/vignettes/summary-statistics.html`.
  - See later in `R_3b_tables_plots_in_R.pdf`.

Exercise 1

See exercise 1 in `R_3a_exercises.Rmd` and `R_3a_exercises_solutions.Rmd`.

Frequency plots

- Bar plot: `barplot(table(var1))`
- Pie chart: `pie(table(var1))`
- Stratified barplots with
  `barplot(table(var1, var2))`
  or
  `mosaicplot(table(var1, var2))`
- (Histogram: `hist()` for metric variables)

How to save the computed plots?

- Use `pdf()` and `jpeg()` to open the connection to a pdf or jpeg file.
- Generate the plot normally, which saves the plot in the file.
- Close the connection to the file with `dev.off()`.

There are also other ways how to do it.

Exercise 2

See exercise 2 in `R_3a_exercises.Rmd` and `R_3a_exercises_solutions.Rmd`.

Descriptive statistics for ordinal & metric variables

Overview of descriptive statistics

Descriptive statistics of nominal variables

- Frequencies

Descriptive statistics of ordinal variables

- Frequencies (if not many categories)
- Minimum, maximum, median
- Range, quantiles, IQR, median absolute deviation (MAD)
- Used but not fully appropriate: mean, standard deviation (SD)

Descriptive statistics of metric variables

- Mean, median, min, max
- Range, quantiles, IQR, MAD, SD, variance
- (Skewness, kurtosis)

Overview of descriptive statistics

These different measures can be used to describe:

- a central value of the variable (mean, median)
- its variation (range, IQR, MAD, SD, variance) and
- its distribution more generally (quantiles, skewness, kurtosis).

Formal details on descriptive statistics

- Minimum and maximum of a variable X are often denoted as $min(X), max(X)$.
- The range of X denotes the width of values and is either described as the interval between $min(X)$ and $max(X)$ or as $max(X) - min(X)$.
- The α% quantile ($0 \leq \alpha \leq 1$) of a variable X is defined as follows: it is the smallest value x of X so that $P(X \geq x) \geq 1 - \alpha$.
- The median of X is the 50% quantile, i.e. the value so that (generally) at least 50% of the values are smaller and at least 50% of the values are larger than it.
- First quartile of a variable = 25% quantile, second quartile = median, third quartile = 75% quantile.
- IQR = interquartile range = third quartile - first quartile
- MAD = median absolute deviation, is the equivalent of the variance for ordinal variables and defined as $MAD(X) = median(|x_i - median(X)|)$

Formal details on descriptive statistics

- The mean of a variable X, $\bar{x}$, is its arithmetic mean (average): $mean(X) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- The variance[1] of a variable X, $var(X)$, is the sum of the squared distances of all observations to the mean: $var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- The standard deviation SD of a variable is the square root of the variance: $SD(X) = \sqrt{var(X)}$
- The skewness of a variable describes whether the distribution of a variable is symmetric (skewness of 0) or skewed (positive or negative skewness depending if left-skewed or right-skewed) and is defined based on the third moment[2] of X.
- The kurtosis of a variable is based on the fourth moment of X and describes where its values are concentrated: highly concentrated at the center (positive kurtosis), "standard" (kurtosis = 0) or flattened and uniform (negative kurtosis).

[1] see lecture 5 for estimating the population variance, with denominator $n - 1$

[2] which is the expected value of X^3

Overview of descriptive statistics

Which statistics do you report?

- Check if variable has nominal/ordinal/metric scale.
- Also check if variable has many of few categories/unique values.
- Also important consideration for choice of appropriate statistics: distribution of variable!

Relevant plots for ordinal & metric variables

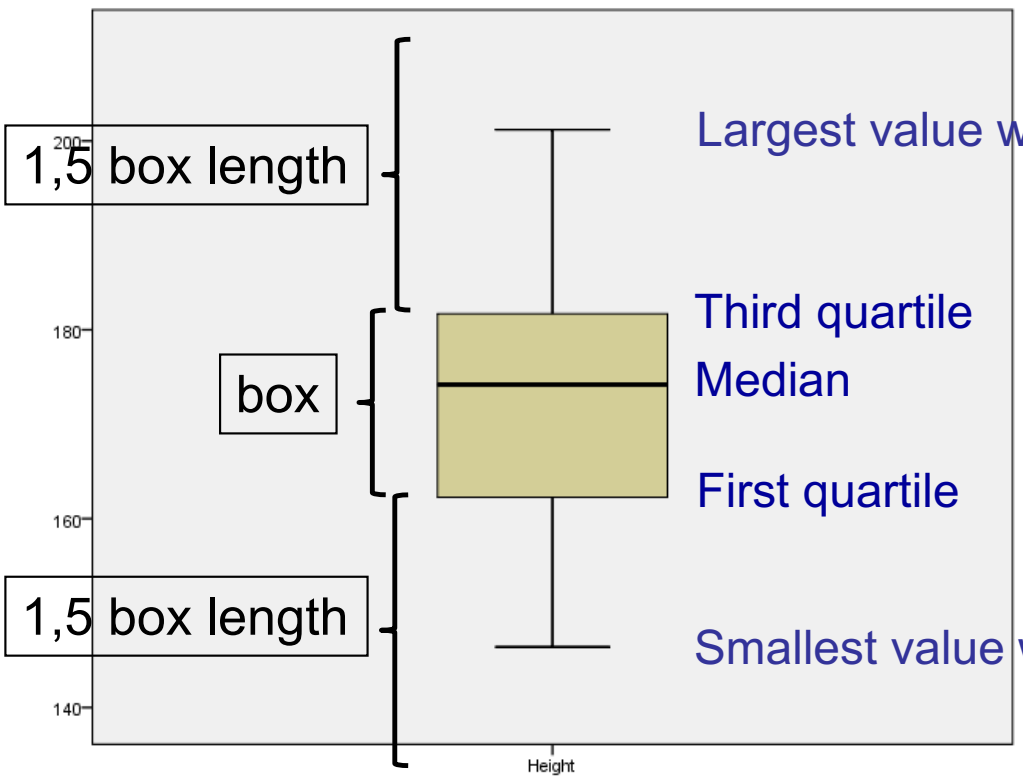
Ordinal variables

- Bar plot, mosaic plot (if not many categories)
- Histogram (if there are many categories; might be not entirely "correct")
- Boxplot
- Scatterplots

Metric variables

- Histogram, boxplot
- Scatterplots
- Quantile-quantile plots

Details on boxplot



Points further away 1,5·box length are depicted as single points

Details on QQ-plot

A QQ-plot plots the quantiles of two variables against each other, e.g. of an observed variable vs. the quantiles of a theoretical distribution, and can be therefore used to compare the distribution of two variables.

How to compute descriptive statistics

- Minimum, maximum: `min(x)`, `max(x)`
- Range: `range(x)`,
- Median, MAD: `median(x)`, `mad(x)`
- Quantiles: `quantile(x, probs = seq(0, 1, 0.25))`
- Mean, SD, variance: `mean(x)`, `sd(x)`, `var(x)`
- E.g. `skewness()`, `kurtosis()` functions in `fBasics` package.

Things to remember:

- `na.rm = TRUE` option to remove missing values!
- SD and variance are based on denominator n-1.
- `quantile()` function has 9 types how to compute quantiles!

$\longrightarrow$ see exercise 3 in `R_3a_exercises.Rmd`.

More on descriptive statistics

- To compute stratified tables, use e.g. `tapply()` function, see exercise 3 in `R_3a_exercises.Rmd`, or also using the `summary_table()` function (see `R_3b_tables_plots_in_R.pdf`).

How to compute plots

- Histogram: `hist(x)`
- Boxplot: `boxplot(x)`
- Scatterplot: `plot(x)`
- Quantile-quantile plot: `qqplot(x)`, `qqnorm(x)`

- The same plots and many more with fancier formatting and many more options can be computed using the functions in the `ggplot2` package, see `R_3b_tables_plots_in_R.pdf`.

$\longrightarrow$ see exercise 4 in `R_3a_exercises.Rmd`.

Homework

Homework

See file `R_3_homework.Rmd`.

Questions?