Biostatistics & Epidemiological Data Analysis using R

5

Estimation

Stefan Konigorski

Health Intervention Analytics Group, HPI

November 15, 2023

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	2023.10.18
	2	First steps in data analysis using R	2023.10.25
	3	Second steps in data analysis using R	2023.11.01
Epidemiology & Statistics: concepts	4	Epidemiological study designs and study planning	2023.11.08
	5	Estimation	2023.11.15
	6	Hypothesis testing	2023.11.22
	7	Missing data	2023.11.29
Data analysis w/ regression models	8	Linear regression I	2023.12.06
	9	Linear regression II	2023.12.13
	10	Regression models for binary and count data	2024.01.10
	11	Analysis of variance & Linear mixed models I	2024.01.17
	12	Linear mixed models II & Meta analysis	2024.01.24
	13	Survival analysis	2024.01.31
	14	Causal inference & Data analysis challenge	2024.02.07

(see full schedule online)

Review of class 4

- Important epidemiological concepts and terminology.
- Epidemiological study designs.

Learning objectives

- Last week: Sample from a population (how do you do that, how do you assess and describe the sample?)
- Today (and in all further classes): Infer from your sample back onto the population (when and how can do that?)
- Understand the main concepts of estimation with examples.

Not learning objectives

Underlying statistical concepts that are not formally treated in this course:

- Measure and probability theory
- Random variables, cumulative and probability density functions, joint probability functions
- Expected value, variance, ...
- Central limit theorem ...

Intuition on basic probability theory concepts

- **Random variable** (rv): numeric variable that can take different values with different probabilities.
- Random variables can be **discrete** ("countably" many values) or **metric/continuous** (values cannot be enumerated).
- **Probability mass** function (for discrete rv)/ **probability density function** (pdf, for metric rv): Function that assigns probabilities to values (discrete)/ range of values (metric) of the rv.
- **Cumulative distribution function** (cdf): describes probability of all values of the rv smaller than a certain number (=sum/integral)
- There exist different distributions for discrete/metric rv that have different form, i.e. functions that assign different probabilities to values of the rv: **Bernoulli**, **Binomial**, **Uniform**, **Normal**, **t-distribution**.

Estimation

Motivating thought experiment

Overview

- Aim: Mimick the process of an empirical study.
- Let's consider the 17,640 children in the KiGGS dataset as the population of interest (children in Germany).

  Important note: This is just for illustration, usually the population is unknown and an abstract, infinite quantity of people.
- Study question: What is the mean BMI of children in Germany?

Exercise 1

- See `R_5_exercise_1.Rmd`.
- Load the KiGGS dataset.
- Take a random sample of 100 children.
- How can you answer the study question? $\longrightarrow$ Compute mean BMI.

What would change if we take samples of 20 kids?

Motivating thought experiment

Summary

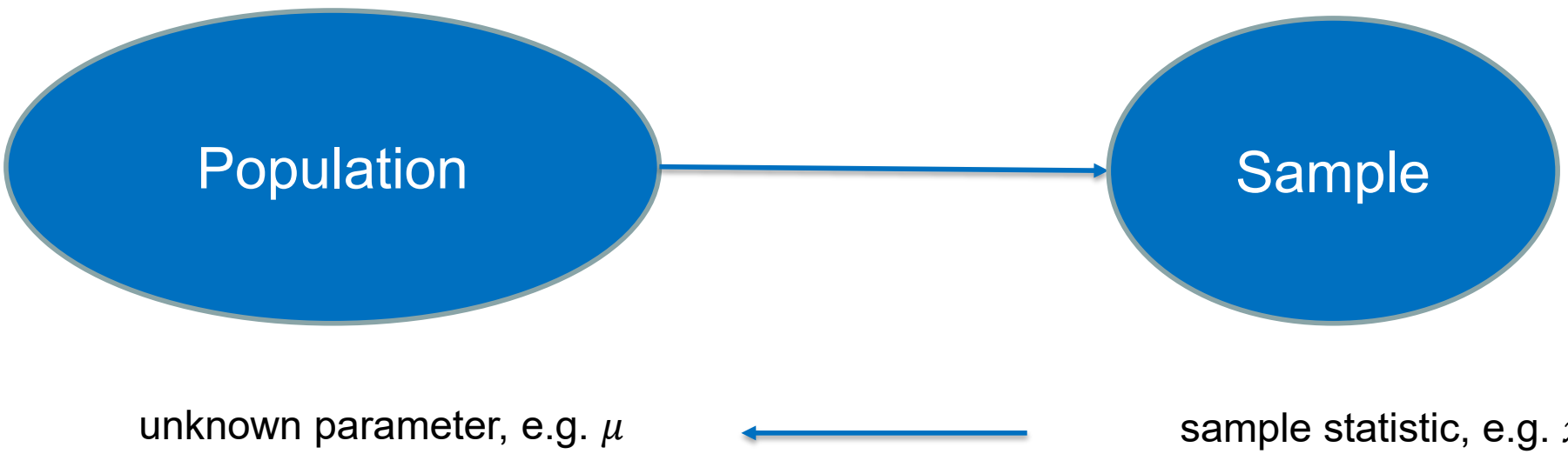
- Parameter of interest is here the expected value of BMI $E(BMI) = \mu$ (theoretical unknown parameter of the population, in our case: $\mu = 18.3$).
- To estimate the unknown population parameter μ, we can use the mean in our empirical sample, which is e.g. $\overline{BMI} = 18.9$, as a point estimate.
- As we have seen, this number is different from the true μ, and differs between our samples due to sampling error (and in general measuring error etc.).
- Hence, in addition to a single estimate, it would be helpful to compute an interval in which the true μ likely is.

Point estimation

Overview

Aim

In point estimation, the aim is to give a best guess for an unknown quantity (parameter) of interest θ.



Overview

Aim

In point estimation, the aim is to give a best guess for an unknown parameter of interest θ.

Here:

- θ is also called the estimand and is a fixed, unknown quantity.
- The best guess is called an estimate of θ: $\hat{\theta}$.
- The estimate $\hat{\theta}$ is a function of the data.
  $\longrightarrow$ estimator $=$ abstract estimator function
- Since the data is random (assume it to be a random sample), and $\hat{\theta}$ depends on it, $\hat{\theta}$ is a random variable and you can determine its distribution.
- The value of the estimate is computed by applying the function to the data.

Examples: What can you estimate?

General examples

You might be interested in estimating:

- a parameter in a statistical model,
- a cumulative distribution function F,
- a probability (density) function f,
- a regression function,
- ...

Examples: What can you estimate?

For a random variable X[1]:

Estimators $\hat{\mu}$ for expected value $E(X) = \mu$

- Mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $Median(x)$

Estimators $\hat{\sigma}^2$ for variance $Var(X) = \sigma^2$

- $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Estimators for standard deviation $SD(X) = \sqrt{Var(X)}$

- Square root of the above variance estimators

[1] x_i is the value of X of the i-th person.

Examples: What can you estimate?

Estimator for a proportion p

- Proportion p = probability/relative frequency/rate, e.g. the prevalence or incidence of a disease
- The underlying binary random variable X (has disease yes/no) has a Bernoulli distribution
- Estimator: e.g. the empirical relative frequency of the event i in n observations: $\hat{p} = \frac{i}{n}$

Examples: What can you estimate?

Odds ratio

- = Measure of association between two binary variables, e.g. height (small/tall) and disease (yes/no)
- Consider the probability of having the disease, p, which is p_{small} in small people and p_{tall} in tall people.
- The odds of having the disease are $\frac{p}{1-p}$.
- The odds ratio that a tall person has the disease compared to a small person is

$$OR = \frac{\left(\frac{p_{tall}}{1-p_{tall}}\right)}{\left(\frac{p_{small}}{1-p_{small}}\right)}$$

.

- Estimator: this equation with $\hat{p}_{small}$ and $\hat{p}_{tall}$.

Examples: What can you estimate?

More examples

- For a variable X, estimate the distribution function $F(x)$ by the empirical distribution function $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$ with

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases}$$

Estimate the regression line and coefficients in a linear regression.

Classification of estimators

Estimators can be classified as:

- parametric estimators (based on distribution assumptions)
- nonparametric estimators

Overview of estimate approaches

Approaches to derive estimators:

- Least squares (LS) method
- Maximum likelihood (ML)
- Method of moments
- Pseudo likelihood
- Estimating equations
- Generalized estimating equations (GEE)
- ...

Least squares method

Approach

Determine the estimator $\hat{\theta}$ of θ such, so that the quadratic error of the model (summed over all observations) is minimized.

Examples

- Linear regression with variables X, Y: $Y = \beta_0 + \beta_1 x + \varepsilon$
  The estimators $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = \frac{\text{Covariance}(X,Y)}{\text{Var}(X)}$ are least squares estimates of the regression coefficients β_0 and β_1 and minimize $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$.

- Multiple linear regression with variable Y and matrix $\boldsymbol{X}$ (containing multiple X) and vector $\boldsymbol{\beta}$: $Y = \boldsymbol{X}^T\boldsymbol{\beta} + \varepsilon$:
  The estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$ is LS estimator of the regression coefficients $\boldsymbol{\beta}$ and minimizes $\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2$.

Maximum likelihood method

Background: Likelihood function

Given are n observations of a random variable X with probability (density) function $f(x)$, which depends on a parameter (vector) θ. Then, the likelihood function $L(\theta)$ equals the probability function given the observations $x = (x_1, \ldots, x_n)^T$: $L(\theta) = f(\theta|x)$.

Approach

Determine the maximum likelihood estimate (MLE) $\hat{\theta}$ of θ as the value that maximizes the likelihood function, i.e. the value that is the most likely given the observations.

Maximum likelihood method

Example: simple linear regression

- Regression equation: $Y = \beta_0 + \beta_1 x + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ with n observations. Here, we want to estimate $\theta = (\beta_0, \beta_1, \sigma^2)^T$.
- Density function of the normal distribution (with $\mu = 0$): $f(Y) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{Y^2}{2\sigma^2}\right)$
- Density function of Y given x: $f(Y|x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(Y-\beta_0-\beta_1 x)^2}{2\sigma^2}\right)$
- Maximize the likelihood function: maximize $L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(Y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}\right)$
- This is equivalent to maximizing the log-likelihood function, which is easier to compute.
- The MLE is here equivalent to the LS estimate.

Properties of estimators

Why are these properties[2] relevant?

Using these properties, you can evaluate how "good" an estimator is, specify what "good" means, and e.g. if in given situation the MLE or LS estimator has better properties.

How can it be relevant your your analyses?

- For reporting the variance of a variable, there are multiple formulas, which one do you use?
- In "Table 1" of a manuscript, do you report the median or mean of a variable? Does it depend on whether a variable is normally distributed or not?

[2] German: "Gütekriterien"

Properties of estimators

- Unbiasedness: An estimator $\hat{\theta}$ is unbiased, if it is "on average correct", i.e. doesn't lead to bias. Formally: $E(\hat{\theta}) = \theta$.
- Consistency: An estimator $\hat{\theta}$ is consistent, if it converges against the true parameter θ, i.e. if it gets closer and closer to θ for large samples (formally: for $n \to \infty$).

Properties of estimators

- Efficiency[a]: an estimator $\hat{\theta}_1$ is more efficient compared to an estimator $\hat{\theta}_2$, if it varies less around the true parameter. This can be measured, for example, using the mean square error[b] $MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$.
- Sufficiency: An estimator $\hat{\theta}$ is sufficient, if it contains all information about the true parameter that can be extracted from the sample. I.e. you cannot do better (with respect to efficiency) if you use any other information from the sample.

[a] The formal definition of an efficient estimator is based on the Cramér-Rao bound.
[b] The MSE can be decomposed into the variance of $\hat{\theta}$ and the bias.

Properties of estimators

Examples for a normally-distributed random variable X

- $\bar{x}$ is an unbiased, consistent, sufficient estimator of the expected value μ.
- Median is a consistent, unbiased estimator of μ, but less efficient than $\bar{x}$ (i.e. larger variance) and not sufficient. (For distributions with long tails, the median is more efficient than the mean.)
- The sample variance $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased, consistent, sufficient estimator of the true variance σ^2.
- The variance $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is a consistent and sufficient, but not unbiased estimator of σ^2.
- $\sum_{i=1}^{n}|x_i - \bar{x}|$ is not a sufficient estimator of σ^2.

$\longrightarrow$ Examples in R: see `R_5_exercise_2_properties.Rmd`.

Standard error

Overview

In addition to obtaining a point estimate $\hat{\theta}$ of the unknown parameter θ, quantifying the precision of the estimator is another important goal.

This can be quantified by the standard error (SE):

Standard error

Standard error of an estimator $\hat{\theta}$ = standard deviation of $\hat{\theta}$

Derive/compute the standard error

How can you determine the variance & distribution of an estimator?

Derive/compute the standard error of $\hat{\theta}$

- Parametric, through theoretical derivations (e.g. through the 2^{nd} derivative of the likelihood function)
- Nonparametric, e.g. with the (nonparametric) bootstrap estimator

Bootstrap

Approach

Determine the standard error and distribution of an estimator "empirically", by doing the following:

1. From the observed sample, draw different random "bootstrap samples" with replacement ("resampling"),
2. Compute the parameter estimate in each of them,
3. Calculate the standard deviation/distribution of these parameter estimates.

Bootstrap - Example

Aim

Estimate the standard error of the estimator $\bar{x}$ of the expected value μ of a normally-distributed random variable X.

Sample

Data: $n = 100$ observations of a random variable X: $x_1, \ldots, x_{100}$.

Approach

1. Draw 1000 random samples of size $n = 100$ from $x_1, \ldots, x_{100}$ with replacement, and compute the mean in each of these bootstrap samples.
2. The bootstrap estimate of the standard error of $\bar{x}$ is the standard deviation of these 1000 means.

Bootstrap - Example

Formal description

1. Do the following k times (e.g. $k = 1000$):
   - Draw a random sample of size $n = 100$ from $x_1, \ldots, x_{100}$ with replacement $\longrightarrow$ bootstrap sample $x_1^*, \ldots, x_{100}^*$
   - Compute the mean $\bar{x}^*$ in this sample $x_1^*, \ldots, x_{100}^*$
2. Compute the bootstrap estimate of the standard error of $\bar{x}$, $SE_{BS}(\bar{x})$ as the standard deviation of the 1000 means $\bar{x}^*$:

$$SE_{BS}(\bar{x}) = \sqrt{\frac{1}{1000-1} \sum_{i=1}^{1000} (\bar{x}_i^* - \overline{\bar{x}^*})^2}$$

$\longrightarrow$ See R exercise in `R_5_exercise_3_bootstrap.Rmd`

$\longrightarrow$ Compare to the theoretical derivation of the standard error.

Confidence interval

Overview

If you know the distribution of the estimator $\hat{\theta}$ ("sampling distribution"), you can compute confidence intervals (CI) :

Confidence interval

$[a, b]$ is a $1 - \alpha$ confidence interval for a parameter θ, if the probability that this interval contains θ is at least $1 - \alpha$:

$$P(\theta \in [a, b]) \geq 1 - \alpha.$$

Notes:

- a and b depend on the data, and are therefore random variables - the true parameter θ is a fixed value.
- A specific estimate of the confidence interval, e.g. $[15, 20]$ does not have a probabilistic interpretation any more.

Interpretation of confidence intervals

Let's consider the estimation of the expected value μ of a variable in the KiGGS data. Say we have computed the 95% CI $= [15, 20]$.

Interpretation - version 1

If we repeat our procedure (take sample from the population, compute mean and confidence interval) often times, then the unknown true μ (=18.3) would be contained in the computed CI 95 out of 100 times.

$\longrightarrow$ 5% of CIs would not contain the true μ; 2.5% times it would be higher than the upper limit of the CI, 2.5% times lower than the lower limit of the CI.

Interpretation - version 2

You get a CI with 95% probability that contains the true parameter.

Derive/compute confidence intervals

Compute confidence intervals for a parameter θ

Generally necessary:

Point estimate $\hat{\theta}$, (standard error estimate of $\hat{\theta}$,) distribution of $\hat{\theta}$

The derivation of the distribution can be done analogously to the derivation of the standard error - either parametric or nonparametric.

Derivation of the 95% CI of the expected value

- Look at random variable X with mean μ and variance σ^2.
- Theory[1] tells us that (under some conditions), $\frac{\bar{X}-E(X)}{\sqrt{Var(\bar{X})}} = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
- Now, we use the knowledge that for a standard normal variable Z, $P(-1.96 \leq Z \leq 1.96) = 0.95$.
- Hence, $P(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$ and $P(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}) = 0.95$.

95% CI of the expected value μ (when the variance σ^2 is known):

$$\overline{x} \pm \left(1.96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

[1] central limit theorem

$(1-\alpha)\%$ confidence intervals

Further examples

- Of expected value μ of normally-distributed X with unknown variance σ^2: $\overline{x} \pm \left(t_{df=n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \right)$
- Proportion p: $\hat{p} \pm \left(z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$
- log-OR: $log(\widehat{OR}) \pm z_{1-\alpha/2} \cdot SE\left(log(\widehat{OR})\right)$,
  with $SE\left(log(\widehat{OR})\right) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

with sample size n and sample variance s (denominator $n-1$).

$(1-\alpha)\%$ confidence intervals

In R

- Can do all the above by using the give equations and the `qnorm` and `qt` functions.
- Or use implementations, e.g. the `binom::binom.confint` function for a proportion p, and the `questionr::odds.ratio` function for an odds ratio.

Questions?

References

Statistical fundamentals

- Knight K (1999). Mathematical statistics. CRC Press
- Rosner B (2010). Fundamentals of biostatistics. Brooks/Cole, Cengage Learning
- Wasserman L (2010). All of statistics. A concise course in statistical inference. Springer.

Bootstrap

- Efron B (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics 7, 1-26.
- Wasserman L (2010). All of statistics. A concise course in statistical inference. Springer.

Homework

Homework

See file `R_5_homework.Rmd`.