# Applied biostatistics self-assessment

Applied Biostatistics is an advanced-level course in the Master's Programme in Biomedicine at Karolinska Institutet, not an introductory course. Although it starts with a short recapitulation of elementary statistical tools and concepts, the general expectation is that the students are already familiar with most of them. The knowledge acquired in the Biomedicine Bachelor Programme at KI suffices.

The students are expected to be familiar with the following concepts:

- Analyzing biomedical data: Descriptive statistics: graphical (bar plots, histograms, boxplots) and numerical (mean, standard deviation, median, quantiles)
- Elementary probability: definition, elementary calculations, conditional probability) and distributions (normal, binomial)
- Sampling and inference: sampling distribution and standard errors (as general concepts and for the mean), central limit theorem, confidence intervals (of means)
- Hypothesis testing: test for means, tests for proportions
- Simple linear regression (one dependent, one independent variable): definition, parameter estimation,confidence intervals for parameters, hypothesis tests of parameters and the model, prediction

## Exercises

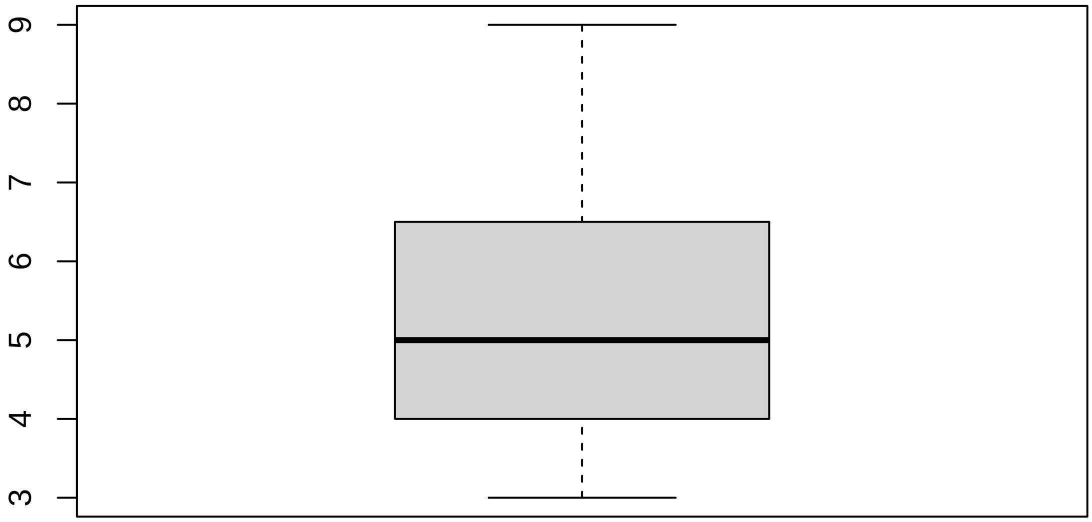
Below are some exercises that you can use to test your knowledge of the material you are expected to be familiar with. If you find any of these exercises difficult, have a look in the pdf file "Statistics for beginners" which contains the slides used in the introductory course given in the Bachelor Program in Biomedicine at KI. If you need more help, contact the course director Matteo Bottai (matteo.bottai@ki.se) as early as possible.

### Exercise 1

We have the following sample of heights: 179, 180 and 181. Give the sample mean and sample standard deviation of height. Show your calculations.

### Exercise 2

Below is a boxplot illustrating the weight of a sample of small children. Explain what the thick line in the middle is and what the grey box represents.

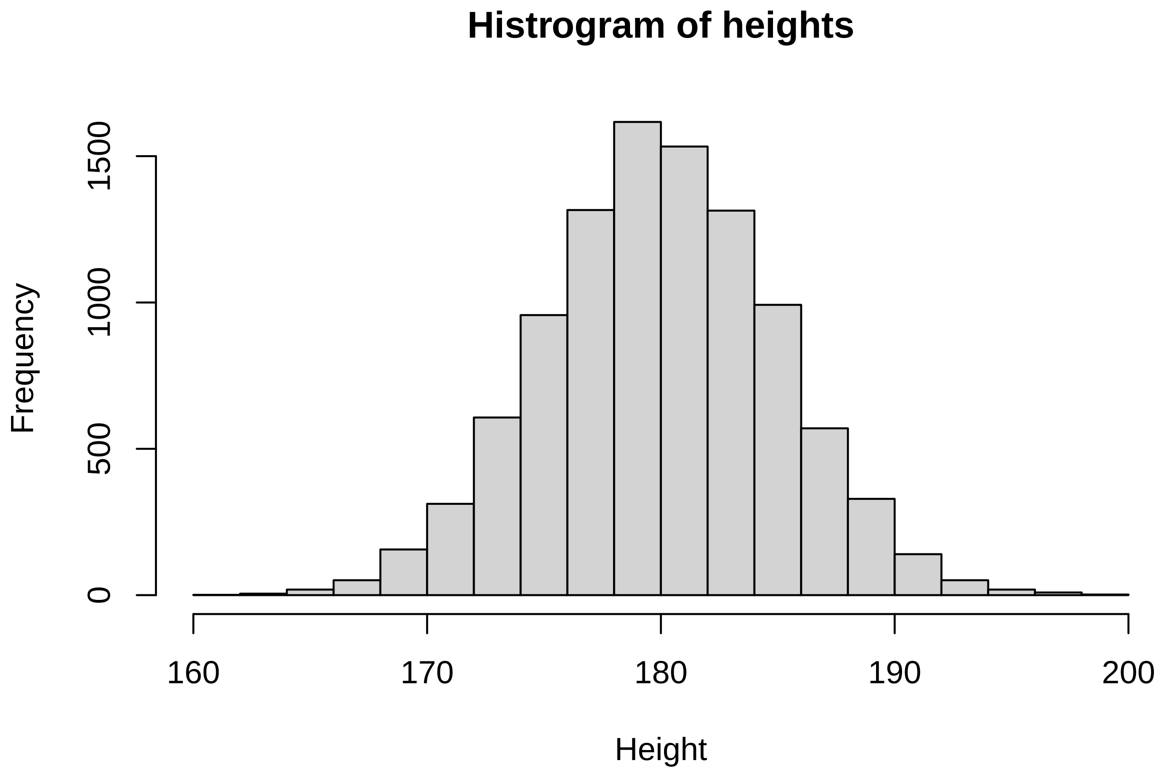


### Exercise 3

Below is a histogram illustrating the height of the participants in a random sample from a larger population.

(a) Does it look like height in the population is approximately normal? Why?

(b) Give an approximate estimate of the sample median, just by looking at the histogram.



### Exercise 4

We will perform a hypothesis test where the alternative hypothesis is that the mean height of Swedes is not 190. What is the null hypothesis?

### Exercise 5

In the context of a hypothesis test:

(a) What is Type I and Type II error?

(b) What is significance level and statistical power?

### Exercise 6

Assume that the height of Swedes follow a normal distribution with mean 180 and variance 25.

(a) What proportion of Swedes are shorter than 180?

(b) Approximately, what proportion of Swedes are taller than 190?

(c) I will select a random person from the Swedish population. What is the approximate probability that I select a person who is shorter than 170?

## Exercise 7

In no more than two sentences, explain what the central limit theorem says.

## Exercise 8

We want to know if Swedes, Danes and Norwegians have the same mean height. Based on a random sample, we perform an ANOVA. The p-value is 0.01.

(a) What null hypothesis is the ANOVA testing?

(b) Assuming that the confidence level of the test is 0.95, should we reject the null hypothesis?

## Exercise 9

I toss a fair coin five times and count the number of heads.

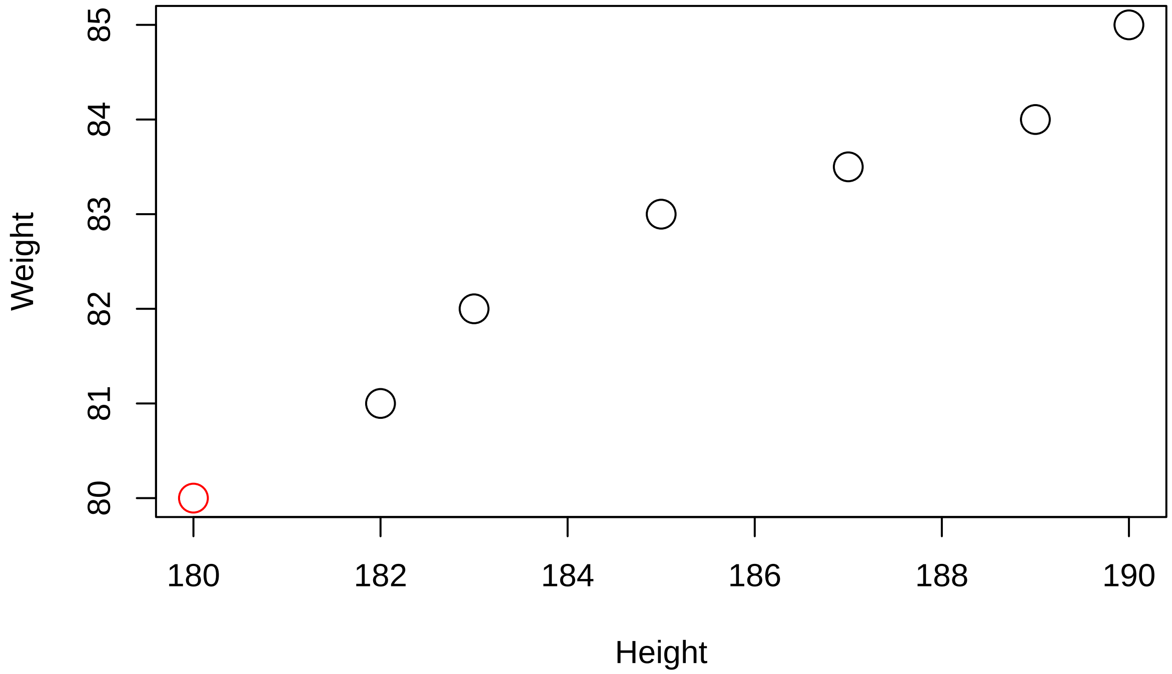
(a) What probability distribution can be used to calculate the probability of observing a certain number of heads?

(b) What is the probability that I observe five heads?

## Exercise 10

Below is a scatter plot of height and weight for participants in a study.

(a) What is the height and weight of the participant represented by the red circle?
(b) Does height and weight seem to be positively or negatively correlated?
(c) Explain what it means that height and weight are positively correlated.



### Exercise 11

Below is a contingency table describing the smoking habits in a sample of male and female participants.

(a) I select a participant at random from the sample and see that he is male. What is the probability that he is a smoker?

(b) Suggest a way of testing if sex and smoking habits are independent based on the sample.

| | non-smoker | smoker |
|---|---|---|
| female | 20 | 10 |
| male | 10 | 20 |

### Exercise 12

(a) Based on a sample, the 95% confidence interval for the mean systolic blood pressure of Swedes is [100,110]. Explain what this means.

(b) We want to test the hypothesis that the mean systolic blood pressure among Swedes is 105. If our significance level is 0.05, should we reject the null hypothesis based on the confidence interval in (a)?

### Exercise 13

According to a linear regression model, the expected weight in kg given height in cm, is

$$E(Weight|Height) = 10 + 0.5 \cdot Height.$$

What is the expected weight of a person who is 200 cm tall?

## Answers

### Exercise 1

Sample mean:

$$\frac{179 + 180 + 181}{3} = 180$$

Sample standard deviation:

$$\sqrt{\frac{(179 - 180)^2 + (180 - 180)^2 + (181 - 180)^2}{3 - 1}} = 1$$

### Exercise 2

The thick line is the sample median. The upper limit of the grey box is the 0.75 quantile and the lower limit is the 0.25 quantile, so 50% of the sample is “inside” the grey box.

### Exercise 3

(a) Yes, it is unimodal and symmetric around the mean, and has the bell-shape typical of the normal distribution.
(b) The highest bar is above 180, approximately, so due to the symmetry, the median should be around 180.

### Exercise 4

The null hypothesis must contradict the alternative hypothesis, so in this case, the null hypothesis is that the mean height of Swedes is 190.

### Exercise 5

(a) Type I error: reject a true null hypothesis. Type II error: not reject a false null hypothesis.
(b) Significance level: the probability of making a Type I error. Statistical Power: The probability of not making a type II error, ie. the probability of rejecting a false null hypothesis.

### Exercise 6

(a) 50%

(b) 2.5%

(c) 2.5%

## Exercise 7

It says that as the size of a random sample increases, the distribution of the sample mean approaches a normal distribution, no matter the distribution of the underlying variable. Note that this is *very* different from saying that the sample approaches a normal distribution (which is nonsense).

## Exercise 8

(a) The null hypothesis is that Swedes, Danes and Norwegians have the same mean height.

(b) If the confidence level is 0.95, the significance level is 1-0.95=0.05. We reject if the p-value is lower than the significance level. So we should reject he null hypothesis.

## Exercise 9

(a) The binomial distribution.

(b) $0.5^5 \approx 0.03$

## Exercise 10

(a) 180 cm and 80 kg.

(b) Looks positively correlated.

(c) If height and weight are positively correlated, then people tend to weigh more the taller they are.

## Exercise 11

(a) There are 30 males in the sample. 20 of them are smokers. So if I take a male at random, the probability of selecting a smoker is 2/3.

(b) We could perform a chi-squared test.

## Exercise 12

(a) We are 95% confident that the true mean sbp is between 100 and 110.

(b) We do not reject since 105 is inside the 95% confidence interval.

## Exercise 13

The expected weight is $10 + 0.5 \cdot 200 = 110$.