Exercises for Applied Biostatistics I - HS 2018

1. In a random experiment two dice are thrown simultaneously. We expect they are “fair” i.e. the numbers 1 to 6 have equal probability.

a) Discribe the sample space of the elementary events.

b) What is the probability of a single elementary event?

c) Calculate the probability of the event $E_1 =$ “The sum of the spots is 7”.

d) What is the probability that event $E_2 =$ “The sum of spots is smaller then 4.” occurs.

e) Determine $P[E_3]$ of the event $E_3 =$ "Both spots are odd.".

f) Calculate $P[E_2 \cup E_3]$.

2. A, B and C are events.

a) Which of the following statements are meaningful?

(i) $P[A \cup (B \cap C)]$

(ii) $P[A] + P[B]$

(iii) $P[A^c] \cap P[B]$

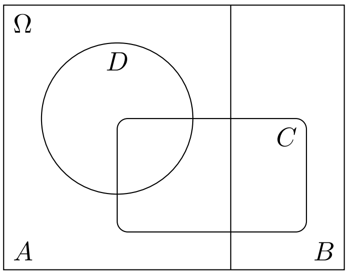
(iv) $(P[B])^c$

b) Display the following events in the given diagram.

(i) $C \cap D$

(ii) $(D \setminus C) \cup (C \cap A)$

(iii) $B \cup D$



3. *Drosophila melanogaster* has the following three genes on chromosome II: b (black body), cn (cinnabar eyes), and vg (vestigial wings); the three genes appear in this order on the chromosome. Since they are all on the same chromosome, Mendel’s law of independent assortment does not hold for their inheritance.

a) By A, B and C, we denote the events of a recombination between the gene pairs (b, cn), (cn, vg), and (b, vg), respectively. Express the event C as a set expression using the sets A and B.

b) The recombination frequency between b and cn is $0.09 = 9\%$, the recombination frequency between cn and vg is $0.095 = 9.5\%$. What is the recombination frequency between b and vg?

4. In a study, researchers asked people in Massachusets about their income and their smoking status. The result of the questionnaire is shown in the following table:

	Income			Total
	low	medium	high	
Smokers	634	332	247	1213
Non-smokers	1846	1622	1868	5336
Total	2480	1954	2115	6549

a) What is the probability someone (a randomly chosen participant) in this study smokes?

b) What is the conditional probability that a person with high income smokes?

c) Is "being a smoker" independent of "having a high income"? Why or why not?

d) Suppose that in a certain population of married couples 30% of the husbands smoke, 20% of the wives smoke, and in 8% of the households, both partners smoke. Is the smoking status (being smoker or nonsmoker) of the husband independent of that of the wive?

5. For an inspection, water samples (10 ml) are tested for contamination. As only 2% of all samples are contaminated, it's proposed to mix ten samples together. From every sample, 5 ml are mixed into a collective sample containing 50 ml. Now the collective sample is tested for contamination. If the sample isn't contaminated, the inspection for the ten samples is over. Otherwise, the 10 single sample must be tested separately.

a) What is the probability that there is no contamination in the collective sample (50 ml) (assuming the ten single samples are independent)?

b) Let the random variable Y be the number of analyses needed. Which are the possible values for Y and with which probability do they occur?

c) How many analyses have to be done in average (what is the expectation value of Y)? How many analyses can be saved by mixing the samples into the collective sample, on average?

6. Show that the equality

$$\operatorname{Var}(X) = E[X^2] - \big(E[X]\big)^2$$

holds for a discrete as well as for a continuous random variable X.

7. Let X be a Poisson distributed random variable: $X \sim \operatorname{Pois}(\lambda)$. Show that $E[X] = \lambda$.

8. The human chromosome 1 has a length of 3.05 M (3.05 Morgan). Use Haldane's model to solve the following problems.

a) What is the probability that there are no crossovers on chromosome 1 in a meiosis?

b) What is the probability that there are 2 or more crossovers?

c) What is the expected number of crossovers, m?

d) What is the probability that there are m or more crossovers?

9. (Exercise for mathematically gifted...) Two genes are located on the same chromosome, d Morgan apart (for some $d > 0$). The random variable X counts the number of crossovers happening between the genes in meiosis I during the formation of gametes. Use Haldane's model to answer the following questions.

a) When the two alleles on the chromosome of the gamete originate from *different* parental chromosomes (i.e., where located on different homolog copies of the chromosome in the diploid cell), we say that a *recombination* happened between the two gene loci. Which values of X correspond to the event of a recombination?

b) Calculate the probability of a recombination.

Hint: use the Tylor series $\sinh(x) = \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)!}$

10. Due to longtime analysis it is well-known that the lead content X in a sample of soil is normally distributed. Furthermore it is known that the expectation is 32 ppb and the standard deviation is 6 ppb.

a) Visualize the density of X in a sketch which includes the probability that the sample of soil contains between 26 and 38 ppb of lead.

b) What is the probability that a sample of soil contains at most 40 ppb of lead? Hint: Solve the problems with R. The commands pnorm and qnorm are needed.

c) Calculate the probability that a sample of soil contains at most 27 ppb of lead?

d) Below which concentration falls the lead content with probability 97.5%? That is, determine c such that $P[X \leq c] = 97.5\%$.

e) Below which concentration falls the lead content with probability 10%?

f) What is the probability of the area you draw in part a) of this exercise?

11. The expression levels of two genes have been discretized to 3 levels each: 1 = low, 2 = medium, 3 = high. The following table shows the joint probability mass function of the expression levels of the two genes (X and Y):

X/Y	1	2	3
1	0.05	0.08	0.12
2	0.14	0.19	0.09
3	0.22	0.08	0.03

a) What's the marginal distribution of X? And the marginal distribution of Y?

b) What's the probability of gene X being expressed on a low level? Does this probability change if you know that Y is being expressed on a medium level?

c) It is known from previous research that gene X regulates the expression of gene Y. Does X up- or downregulate Y? Motivate your answer.

d) In the table above, you see that $p_{X,Y}(1,2) = p_{X,Y}(3,2) = 0.08$. Does this imply that $p_{Y|X=1}(2) = p_{Y|X=3}(2)$? Comment.

12. During measurements through PET (Positron emission tomography) a detector produces "clicks". These clicks are counted for each voxel (volumetric pixel). The counts of different voxels from different PET machines are stored in the data set `count.txt`.

The goal of this exercise is to get acquainted with the basic abilities of the R statistical software. We will analyze the data in later exercises.

a) The data set is available in ILIAS. Download it and import it into R. Investigate the properties of the imported data set:

- To which class does the created R object belong?
- How many rows and columns does the dataset have?
- In which data type are the variables (columns) saved?

R hints: `class()`, `dim()`, `str()`

b) Calculate important characteristic values (mean, variance, quantile) of `a.v1` to characterize the distribution of the count data.
R-Hint: `summary()`, `mean()`, `variance()`

c) Use a histogram and a boxplot to draw the distribution of the observed count. Which of the two plots is more appropriate? How would you plot the relative frequencies in a histogram or rather which argument of the `hist` function is needed to get a scaled histogram?
R-Hint: `table()`, `boxplot()`, `hist()`

13. A test tube manufacturer wants to make sure that a large delivery contains less than 10% test tubes of minor value. He decides to take a random sample of $n = 50$ test tubes. In the sample are $x = 3$ test tubes of minor value.
The manufacturer has now the problem to decide if the whole delivery has less than 10% of minor value, or if it is plausible that in the sample are less than 10% of minor value by pure chance, although in the delivery are more than 10% of minor value.

a) Which model respectively distribution describes the number of test tubes of minor value X? Assume that the individual test tubes are independent and have the same probability to be of minor value.

b) What is the probability that the sample contains exactly 3 test tubes of minor value if the delivery contains 10%?

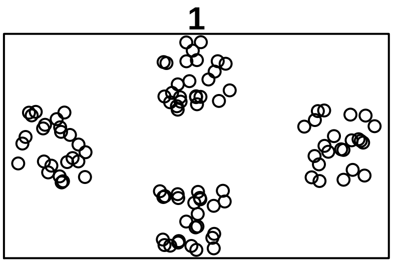
c) What is the probability that the sample contains at most 3 test tubes of minor value if the delivery contains 10%?

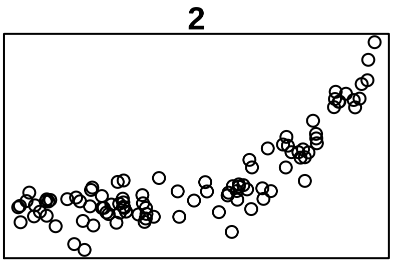
d) Calculate again the probability that $X \leq 3$. But this time use the normal approximation. What do you observe?

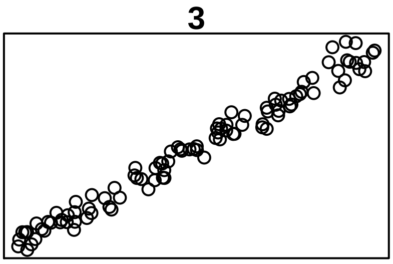
e) Explain in a few words the trade-off between to be sure (with high probability) that the delivery has the required quality and refusing the delivery even if the quality is as desired.

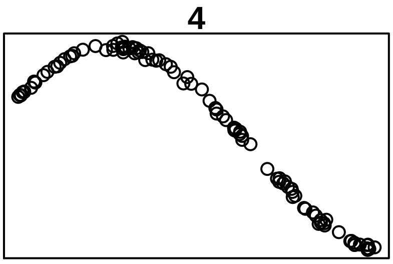
14. **a)** Match the correlations with the scatter plots.

$$a = 0.01 \qquad b = 0.99 \qquad c = 0.81 \qquad d = -0.89$$

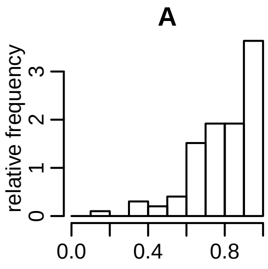


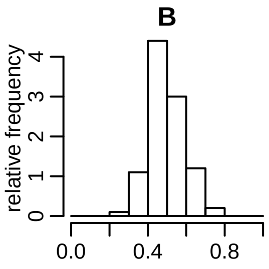


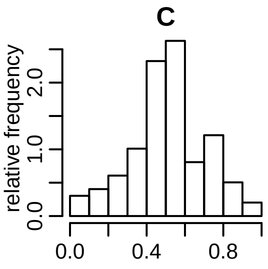


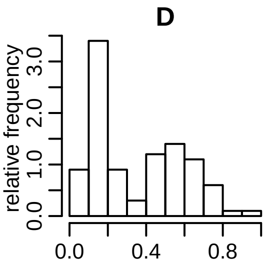


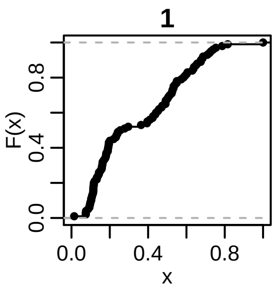
b) Match the histogram's with the CDFs.

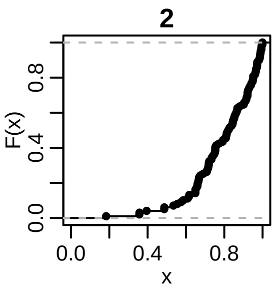


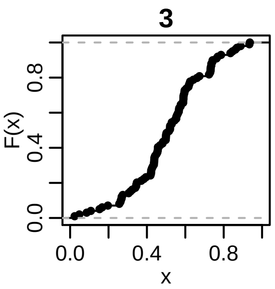


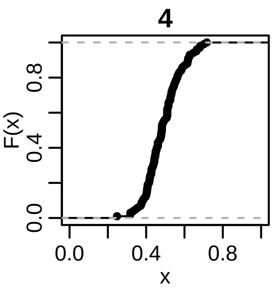




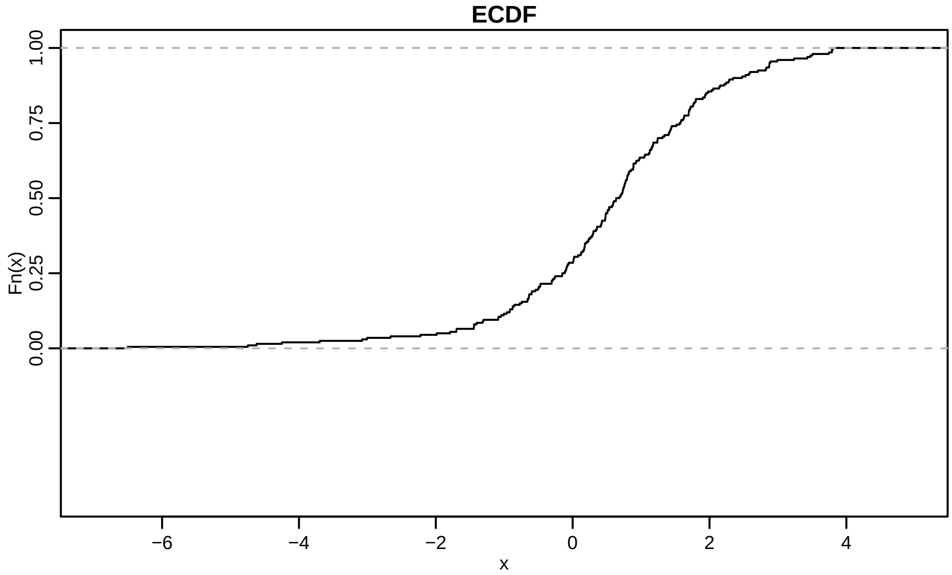








c) The following plot shows the ECDF of a data set:



Draw (by hand) a box plot of the data set.

15. An entomologist has found a new ant species with red and black individuals. She estimates that roughly one of ten ants is red.

a) What's the probability that 3 out of 5 randomly sampled ants are red?

The entomologist would like to investigate the ant species more closely. Therefore she takes a sample of 150 ants.

c) How is the number of red ants in this sample distributed? Calculate expectation value and variance of this distribution.

d) Use the normal approximation to calculate the probability that the number of red ants in the sample lies between 15 and 20.

16. A 90% confidence interval for any value should contain that value with a 90% probability—that is, in 90% of the cases you calculate a CI whose range covers the value. The aim of this exercise is to test this in a small simulation.

a) Randomly draw $B = 1000$ samples of $n = 50$ exponentially distributed random variables:

$$X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda = 2) \ .$$

For each of the 1000 samples, calculate the sample mean and an appropriate 90% confidence interval based on the assumption that the sample mean has a normal distribution (central limit theorem!). In how many cases does the expectation value $\text{Exp}(X) = \frac{1}{\lambda} = 0.5$ (i.e. the true mean) lie within the confidence interval?

[R-hints]:
`set.seed(`*<any number here will do>*`)` to make your results reproducible
`for(i in a:b){`*<expression>*`}` to run the successive execution of an expression
`rexp(`*<exponential distr parameters>*`)` to simulate the samples
`if(`*<condition>*`){`*<expression>*`}` to check a statement and set a response if statement is fulfilled (for your condition-argument you can make use of (combinations of) logical operators such as `==`, `!=`, `<`, `>`, `&`, `|`, ...)

b) Answer the same question as in a) for 1000 samples of size $n = 5$. Do you find a different number of confidence intervals containing the true expectation value? Why?

17. The data set `bone-fracture` contains 50 samples of the following four variables:

`conc`: concentration of the degradation product of a drug. All patients are taking the same dosage of the drug.

`dif`: difference between needed and available screw length. The needed screw length is determined based on X-ray images of the bone fracture. However, screws are only available in steps of 5 mm. The column `dif` lists the difference of the needed screw length to the nearest available length.

`no.cells`: number of cells counted in a proliferation assay. In this assay, the number of living cells in a petri dish are counted at a specific time.

`hit`: success (`hit` = 1) or failure (`hit` = 0) of a manual prediction whether a bone is fractured or not. The physician E. Bruch claims that he is able to determine with his own hands whether a bone is fractured or not, without looking at an X-ray image.

You find the data set under `bone-fracture.csv` on ILIAS and can read it in as follows:

```
> fracture <- read.table("bone-fracture.csv", sep = ";", header = TRUE)
```

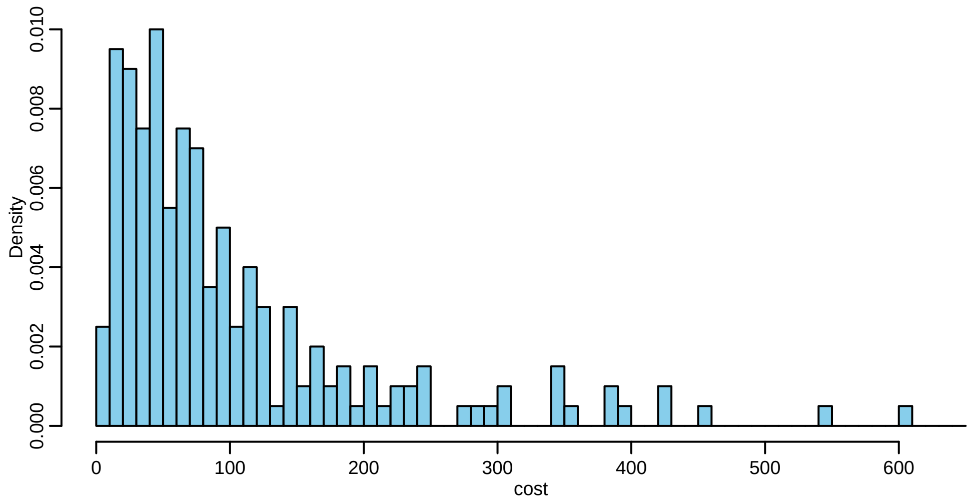
Find an appropriate distribution for each of the four random variables and estimate its parameters. If possible, indicate an approximate 95% confidence interval for each parameter. Visualize the distribution of the data and the distribution of your fitted model.
For example: if you consider a variable to have come from a binomial distribution, you estimate parameter π. Then, using the standard error of this estimated parameter and the formula, calculate the confidence interval of π. Finally, recheck whether your guess was appropriate by looking at the Q-Q plot (which should show a straight line) and adding the estimated probability mass function to the histogram (which should match one another).

[R-hints]: `fitdistr()` (from package `MASS`), `qqPlot()`, `hist()`

[R-script-hints]:

```
> qqPlot(VARIABLE-NAME, dist = "chosen-distr.", "estimated parameters of chosen distr", xlab = "Theor. quantiles (norm)", ylab = "Empirical quantiles")

> hist(VAR-NAME, breaks = 20, freq = FALSE)

> ### or "prob=TRUE"

> x.val <- seq(min(VARIABLE), max(VARIABLE), length = 50)

> ### or "by=0.1" instead of "length=50" (just make sure x.val has reasonable number of elements)

> y.val <- "d-chosen distr"(x.val, "est.parameters of distr.")

> ### so use "dnorm", "dpois", etc.

> lines(x.val, y.val)
```

18. Have a look at the following histogram of monthly costs (in CHF) for health care of 200 patients:



a) The data set is stored in the file `healt-care-cost.txt`, available on Ilias; you can read in the one-dimensional vector using the R function `scan()`. Do you know a distribution that fits the data? Try to fit this distribution to the data set, for instance by using the function `fitdistr()` from the package `MASS`, which estimates parameters for the selected distribution. Then check the model fit (e.g. Q-Q plot).

b) Perform a log-transformation on the data set. Then, draw a histogram of the transformed costs. As in task a), choose a distribution that fits the log-transformed data well, estimate its parameters and check your final model. Plot the estimated density function as a line into the histogram.

c) Explicitely indicate the values of the parameters of the fitted distribution of the log-transformed cost data.

19. Write an R-script, that contains all code for solving this exercise.
In this exercise we work again with the count dataset. Make sure that the dataset is in your workspace directory and load the dataset as follows:

$>$ d.pet $<$- read.table("count.txt", header $=$ TRUE)

a) The dataset `count.txt` contains the number of "clicks" per voxel. Therefore it seems reasonable to fit a Poisson model that is designed for count data. The Poisson model is specified by the value of the parameter λ (Lambda). Estimate the parameter λ of the variable $a.v1$ with the maximum likelihood estimator (MLE). Plot the distribution of the counts, that you would expect from your Poisson model.
R-Hint: `?dpois`

b) Simulate a time series of counts from the estimated Poisson model. The simulated series should have the same length as the observed series. Plot the simulated data.
R-Hint: `?rpois`

c) How well does the Poisson model describe the data? Give arguments, why or why not the fitted Poisson model describes the observed counts of a.v1 well. Compare the model with the data in several ways (density over a scaled histogram, quantile quantile plot) .
R-Hint: `?hist`, `?qqPlot`.
The function `qqPlot` is not a standard R-function. To use it, you should load the package "car". For this, use `install.packages(name)` and afterwards load it into your R-workspace using `library(name)`.

20. The pharmacy company Life Co. has developed a new drug against ADHS. To check the effectiveness, this drug was tested with n=10 patients. At present, the standard medication can cure 30% of the treated patients.

a) Suppose the new drug is as good as the standard one. What is the probability that it cures at most two patients?

b) The treatment with the new drug was successful with exactly four patients. Perform a one-sided hypothesis test to decide if the new drug is better than the standard one (with a level of significance of 1%). Write down explicitly all six steps.

c) How is the power of a hypothesis test defined? What is the power for the test H_0: $\pi = 0.3$ vs. H_A: $\pi = 0.6$.

21. Musculosceletal neck-and-shoulder disorders are common among office staff who perform repetitive tasks using visual display units. A study was carried out to determine whether more varied work conditions would have any impact on arm movement. The accompanying data was obtained from a sample of $n = 16$ subjects. Each observation is the amount of time, expressed as a proportion of total time observed, during which a subject's arm elevation was below 30 degrees. From each subject, two measurements were obtained 18 months apart. During this period, work conditions were changed, and subjects were allowed to engage in a wider variety of work tasks. Does the data suggest that the true average time during which elevation is below 30 degrees differs after the change from what it was before the change? Perform an appropriate test on the 10% level.

The data set can be found in the file `neckshoulder.txt` available in ILIAS

Hint: Our variable of interest (voi) constitutes the difference in time during which arm elevation is below 30 degrees before and after the change in work conditions. What would be the distribution of our voi? Estimate the necessary parameters and check your assumption before doing tests (e.g. Q-Q plot).

22. A test tube manufacturer guarantees that the percentage of tubes of lower qualitity is less than 10 %. (cf. series 5, exercise 1)

a) The number of tubes of lower quality follows a binomial distribution, where π denotes the (unknown) true probability of a tube being of lower quality. For the quality control the manufacturer takes a random sample of $n = 50$ tubes. The sample contains $x = 3$ tubes of lower quality.
The manufacturer now has the problem of deciding whether or not he can assume that his promise is true. Set up a statistical hypothesis test for the manufacturer. Use the six-step approach of hypothesis testing.
Hint: the "tricky" step of this situation is the definition of the alternative hypothesis. How do you have to set it up, and why?
Test his hypothesis that $\pi = \pi_0 = 0.1$ against the alternative hypothesis that $\pi < \pi_0$ on the significance level $\alpha = 0.05$. Use the six-step approach of hypothesis testing.

b) Calculate the probability that the test rejects the null hypothesis given that the true $\pi = 0.075$ (so the percentage of tubes of lower qualitity really is less then 10 %). In other words, calculate the power under the alternative $\pi = 0.075$.

c) What size n should the sample have to give a power of (at least) 50% for a true fraction $\pi = 0.075$ of tubes with lower quality?

23. In order to minimize the forces acting on the spine when flying a sports airplane, it is important that pilots activate certain groups of muscles in the belly and the back during the flight. To test the effectiveness of a new training programme, the muscle activation during a flight of 10 pilots was measured before and after the training; this can be done by the aid of electrodes on the skin.

The data set can be found in the file `training.txt` available in ILIAS and in the following table:

before:	46.0	34.6	37.5	43.3	32.4	25.1	40.0	33.7	31.1	85.0
after:	52.1	41.2	39.3	48.1	38.6	28.6	45.7	38.5	36.7	70.0

a) Use an appropriate plot to check whether muscle activation is normally distributed before and after the training. Do the same for the difference in muscle activation. How do you interpret these plots?

b) Perform an appropriate test to investigate whether muscle activation could be increased by using this training program. Indicate the appropriate null and alternative hypotheses. Name the test you want to use and motivate your decision. Finally, perform the test using R on a significance level of 5% and indicate the result. What is your test decision?

Hint: Consider the assumptions of the tests you know: binomial test, t test, sign test, Wilcoxon signed-rank test.

c) Assume that 200 pilots were randomly divided into 20 groups with 10 pilots each. Within each group, one of 20 different training programmes were evaluated. In those 20 tests, only one training programme showed a significant improvement of the muscle activation on a 5% level.

This one training programme was then published in an aviation magazine, together with the hint that it significantly improves muscle activation (with significance level 5%).

Comment on the outcome of the experiments and the magazine's presentation of this outcome. Motivate your answer.

Hint: definition of α (i.e. Type I error)

24. In this exercise, you will perform a statistical analysis of the distribution of palindromes in the genome of the human cytomegalovirus (HCMV) that you found in Exercise 6 of the course "Basic Programming for Non-Informaticians" of Beppo Brambilla. If you do not attend that course (or did not succeed in finding palindromes), you can nevertheless do this exercise with a provided list of palindrome positions (see ILIAS).

The HCMV is potentially life-threatening for people with a weak immune system. In order to understand how the virus replicates, it is important to find the origin of replication in the virus' genome, that is the place on the DNA onto which replication enzymes can bind. It is known that many DNA binding proteins recognize palindromic sequences. Therefore, starting the analysis of HCMV by focusing on those regions of the DNA with an accumulation of palindromes seems reasonable, as these regions would be good candidates for the origin of replication.

A **palindromic sequence** is a DNA sequence which is equal to its reverse complement. For example, the DNA sequence ACCTAGGT is palindromic because its nucleotide-by-nucleotide complement is TGGATCCA, and reversing the order of the nucleotides in the complement gives the original sequence; see also
`http://en.wikipedia.org/wiki/Palindromic_sequence`

a) In Exercise 4 of the course "Programming for Non-Informaticians", you searched for the centers of all palindromes in the DNA of HCMV. Take the output file generated there and read it into R. In the end, you should have a vector `positions` with the centers of all palindromes of length 10 or more.

If you do not have your own list of palindrome centers, you can use the the list provided in the file `palindromes.txt` available on ILIAS. It consists of only one line starting with "`10:`" (standing for the minimal palindrome length considered), followed by the comma-separated positions of the palindrome centers. It is the output of the sample solution to Exercise 4 of the course "Programming for Non-Informaticians".

Hint: you can preprocess the file manually to make it easier to read in; have a look at the R function `scan`.

b) The HCMV genome is 235727 bp long. Divide it into bins of 4000 bp length, which yields 59 bins (the last one a bit shorter than 4000 bp). Denote by X_i the number of palindrome centers in bin i ($i = 1, \ldots, 59$). If the palindromes would just occur by chance, which distribution of X_i would you expect? Why?

c) In R, calculate the actual number of palindrome centers X_i for each bin and fit the distribution you proposed in task b) by maximum-likelihood estimation. Does the distribution fit the data well?

d) Do you find an accumulation of palindromes somewhere in the HCMV genome? How would you look for that?

Hint: perform a hypothesis test adapted to the distribution proposed in b) for each of the bins. Formulate the model, as well as null- and alternative hypotheses; decide between a one-sided or a two-sided test; perform the test on the 0.05% level.

e) How do you interpret the outcome of task d)? What would you suggest to a biologist working in the lab who wants to investigate the replication mechanism of HCMV?

f) Why is it reasonable to choose such a low significance level as proposed in task d)?
Hint: since you perform the test for 59 bins, what's the expected number of type I errors (false positives) for all tests? What would it be if we chose a level of, say, 5%?

25. The data set `astro.dat` contains data from a test for space conditioning. 17 astronauts out of 26 were randomly assigned to the "salt" group (that is, they took salt before a simulated flight), the rest was assigned to the control group (not taking salt before the flight). Before and after the flight the pulse was measured.

The data set is available as `astro.dat` in ILIAS.

The columns `before` and `after` denote the pulse of the astronauts before and after the flight, resp.; the column `salt` indicates whether the astronauts have taken salt before the flight.

a) To begin, we want to know whether the pulse of the astronauts is the same before and after the flight.
Which test do you choose for that task? Why? Name the model assumptions of the test you choose, and check them with an appropriate plot if necessary.
Explicitely state the null and the alternative hypothesis. What is your test decision for a significance level of 5%?

b) Now we want to check whether the pulse after the flight is the same with and without having taking salt.
Which test do you choose this time? Why? Name the model assumptions of the test.
Explicitely state the null and the alternative hypothesis. What is your test decision for a significance level of 5%?

26. In the following table the jaw lengths of some male and female jackals are listed:

male	x_i	120	107	110	116	114	111	113	117	114	112
female	y_j	110	111	107	108	110	105	107	106	111	111

We have $\overline{x} = 113.4, \overline{y} = 108.6, s_x^2 = 13.82, s_y^2 = 5.16$.

a) Is the sample paired or unpaired? Justify!

b) Does the jaw length differ significantly between male and female jackals? Execute a t-test by **hand**. Do you use a one- or a two-sided test? Are the data paired or unpaired? Use the 6-step procedure (specifying model, null and alternative hypothesis, test statistic, significance level, range of rejection, test decision). The quantiles required for the range of rejection you can calculate in **R**.

c) Execute the t-test in **R**. Write down the p-value and the test decision. Use the **R**-function `t.test`.

27. We again consider the data set of jaw lengths of male and female jackals:

male	x_i	120	107	110	116	114	111	113	117	114	112
female	y_j	110	111	107	108	110	105	107	106	111	111

a) Implement a permutation test in R; you can take the implementation presented in the lecture as a template. Perform a one-sided test for the null hypothesis "male and female jackals have the same mean jaw length" vs. the alternative hypothesis "male jackals have on average longer jaws than females". Compute the p-value for your test.

b) Perform the same permutation test as in task a) using the R function `permTS` from the package `perm`. Do you get the same p-value as in a)? If not, what could be the reason?

c) Finally, perform a t-test for the same null and alternative hypotheses as in a). Compare the p-values to the two p-values calculated before.

d) Which test do you prefer for this data set: a permutation test (as in tasks a) and b)) or a t-test (as in task c))? Motivate your answer.

28. We consider clusters of palindromic sequences in the genome of the human cytomegalovirus. To do so, we divided the whole genome into bins of length 4000 bp and counted the number of palindromes in each bin. The resulting numbers of palindromes per bin can be found in the file `counts.txt` available on ILIAS.

a) We assume that the number of palindromes X_i in a given bin i is Poisson distributed:

$$X_i \sim \text{Pois}(\lambda_i) \; .$$

We want to perform a test to find the bins significantly enriched with palindromes. Do this by calculating the p-value for the null hypothesis "$\lambda_i = \lambda_0 = 5$" vs. the alternative hypothesis "$\lambda_i > \lambda_0 = 5$" for each bin.

For further processing, store these p-values in a vector in R.

Note: the value λ_0 of the null hypothesis is given by the average number of palindromes per bin (rounded to integer value). Hence, the null hypothesis states that all bins have on average the same number of palindromes.

b) Which bins would you consider as significantly enriched with palindromes on a *comparison-wise* significance level of 5%, based on the p-values calculated in task a)?

c) Adjust the p-values from task a) according to Bonferroni's correction. Which bins do you declare significantly enriched on an *experiment-wise* significance level of 5%?

d) Adjust the p-values from task a) according to Holm's correction. Which bins do you declare significantly enriched on an *experiment-wise* significance level of 5%?

e) Plot the adjusted p-values from tasks c) and d) against the unadjusted p-values from task a). Discuss the differences between the (adjusted) p-values calculated in tasks a) and d). Which approach do you consider most suited in the palindrome data set? Motivate your answer.

[R hint:] To distinguish the results of the different tests, plot adjusted p-values versus unadjusted p-values for the different corrections. Place all values into the same plot, and use different point types to distinguish them (argument `pch` of the `plot` or `points` function).