List of typical exam questions

!!! Notice this document is a draft. It is still under construction and will be extended!!

The exam of the course Data Analysis / Statistics of the summer semester 2021 will include different types of questions. The questions will cover different topics, resp. refer to the different parts of the course. The following list is supposed to allow the examinees to get an idea of what types of questions to expect. It is a sample of the questions that could be used in the exam. The following list is not a comprehensive catalogue of all possible exam questions!

Part1

(1) Explain the meaning of the following terms in Data Analysis:

- – Population
- – Sample
- – Data
- – Variable

(2) In statistics one can distinguish between descriptive statistics and inferential statistics. Explain what these two fields of statistics are about.

(3) A classical problem driven data analysis can be considered a process that can be divided into 4 major steps. Which are those steps.

(4) Consider the following R code:

```
# Create a vector with integer values
valVec <- c(12, 16, 22, 7, 32)

# Operate on the vector with the integer values.
valVecConv <- valVec *2/1+1

# Print the results to the console.
valVecConv
```

Which value or values will be printed to the console (format does not matter)?

(5) In R NA is a reserved token. What does it stand for? What is used to represent?

(6) Why is the default missing value, NA, a logical vector? What's special about logical vectors? (Hint: think about c(FALSE, NA_character_).)

(7) Name the four different systems/groups of objects in R.

(8) What happens to a factor in R when you modify its levels? Consider the following code:

```
f1 < - factor(letters)
levels(f1) < - rev(levels(f1))
```

(9) Assign the names of the R data structures listed below the following table to the correct empty cells of the table according to the dimensionality of the data structures and their ability to represent a set of different data types (heterogeneous) or not (homogenous).

Dimension	Homogeneous	Heterogeneous
1d		
2d		
nd		

Atomic Vector, Matrix, Array, List, Data frame

(10) Which R data structure can one consider the equivalent of a data matrix of a multivariate data set, with variables of different levels of measurement?

(11) Consider the following R code snippet:

```
vecA <- c(1,2,3)
vecB <- c(3,4,5)
data <- data.frame(vecA,vecB)
data[2,]
```

Which values are printed to the console after the last line of code?

Part2

(12) What is Regression Analysis? What is the difference between a simple and a multiple regression?

(13) How does the method of the least squares in regression analysis work?

(14) What does the measure that is called the coefficient of determination (R^2) measure?

(15) Consider the following R code (line starts with ">") and the respective output on the R console:

```
> data<-read.csv("compensation.csv")
> str(data)

'data.frame':    40 obs. of  3 variables:
 $ Root   : num  6.22 6.49 4.92 5.13 5.42 ...
 $ Fruit  : num  59.8 61 14.7 19.3 34.2 ...
 $ Grazing: Factor w/ 2 levels "Grazed","Ungrazed": 2 2 2 2 2 2 2 2 2 2 ...



   Continued on next page.
```

```
> summary(data)

      Root              Fruit              Grazing
 Min.   : 4.426   Min.   : 14.73   Grazed  :20
 1st Qu.: 6.083   1st Qu.: 41.15   Ungrazed:20
 Median : 7.123   Median : 60.88
 Mean   : 7.181   Mean   : 59.41
 3rd Qu.: 8.510   3rd Qu.: 76.19
 Max.   :10.253   Max.   :116.05


> model <- lm(data[data$Grazing=="Ungrazed", ]$Fruit~data[data$Grazing=="Ungrazed",
]$Root)
> summary(model)

Call:
lm(formula = data[data$Grazing == "Ungrazed", ]$Fruit ~ data[data$Grazing ==
    "Ungrazed", ]$Root)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4542 -4.2430 -0.4643  4.8578  8.8639

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  -94.367      9.211  -10.24 6.14e-09 ***
data[data$Grazing == "Ungrazed", ]$Root       23.996      1.507   15.93 4.72e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.755 on 18 degrees of freedom
Multiple R-squared:  0.9337,     Adjusted R-squared:   0.93
F-statistic: 253.6 on 1 and 18 DF,  p-value: 4.715e-12
```

- Which particular analysis has been carried out with the lm(…) call?

- Which variables are modeled?

- What does the resulting model look like?

- Interpret the result of the lm(…) call!

- Visualize (graph) the resulting model?

Part3

(16) For a parallel structure of identical components, the system can succeed if at least one of the components succeeds. Assume that components fail independently of each other and that each component has a 0.15 probability of failure.

- Would it be unusual to observe one component fail? Two components?

- What is the probability that a parallel structure with 2 identical components will succeed?

- How many components would be needed in the structure so that the probability the system will succeed is greater than 0.9999?

(17) A box contains 50 red balls and 6 blue balls. What is the probability in a random selection of 5 balls that exactly 3 of the selected ones are blue?

(18) 1% of all people have cancer. 90% of people who have cancer test positive when given a cancer-detecting blood test, meaning the test detects cancer 90% of the time. 5% of people will have

false positives, meaning that 5% of the time this test will produce a positive result when people do not have cancer.

Given the above data, what is the probability that a person has cancer if they have a positive cancer-test result? (Note: answers are rounded to the nearest 4th decimal place)

(19) Data from car accidents all over the country have shown that most accidents occur within 25 miles from the place of residence of the involved drivers. Based on this statistic people might erroneously draw the conclusion that drivers are less likely to have an accident when they are further away from home. Use the given example to explain what the "Confusion of the Inverse" phenomenon is about.

Part 4

(20) A term life insurance policy will pay a beneficiary a certain sum of money upon the death of the policyholder. These policies have premiums that must be paid annually. Suppose a life insurance company offers such a term life insurance policy for 18-year-old males that would pay €150,000 to the beneficiary and would cost €850 annually. Consider that the probability for an 18-year-old male to survive the year is 0.9998 and compute the expected value of this policy to the insurance company. Interpret the result.

(21) A shipment of 120 fasteners that contains 4 defective items was sent to a manufacturing plant. The quality control manager at the plant randomly selects 5 fasteners. What is the probability that exactly one of the selected ones is defective?

(22) According to a study 86% of the countries households own a cellularphone (one or more). In a simple random sample of 300 households, 275 owned at least one cellphone. Is this result unusual?

Part 5

(23) Imagine a friend is usually late. Suppose that she could be equally likely on time or up to 30 minutes late. Given a random appointment with that friend, what is the probability for the friend to be between 10 and 20 minutes late?

(24) The Empirical Rule states that about 68% of the data in a bell-shaped distribution lies within 1 standard deviation of the mean. For the standard normal distribution, this means about 68% of the data lies between z =-1 and z=1. Verify this result. Verify that about 95% of the data lies within 2 standard deviations of the mean. Finally, verify that about 99.7% of the data lies within 3 standard deviations of the mean.

Gestation Period

280
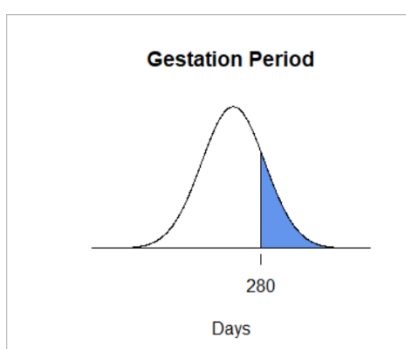
Days

(25)   The lengths of human pregnancies are normally distributed with μ= 266 days and σ=16 days. The figure on the left represents the normal curve with μ= 266 days and σ=16 days. The area to the right of x=280 is 0.1908. Provide two interpretations of this area.

(26) The data in the following table represent the finishing times (in seconds) for six randomly selected races of a greyhound named Barbies Bomber. Is there evidence to support the belief that the variable "finishing time" is normally distributed?

Barbies Bombers's finishing times	
31.35	32.52
32.06	31.26
31.91	32.37

(27) The following lines of code are used in R in order to solve problem (26):

```
barbVec <- c(31.35,32.52,31.26,32.06,31.91,32.37)
qqBarb <- qqnorm(barbVec)
qqline(barbVec)

As a result R plots the following normal q-q plot:
```


Normal Q-Q Plot
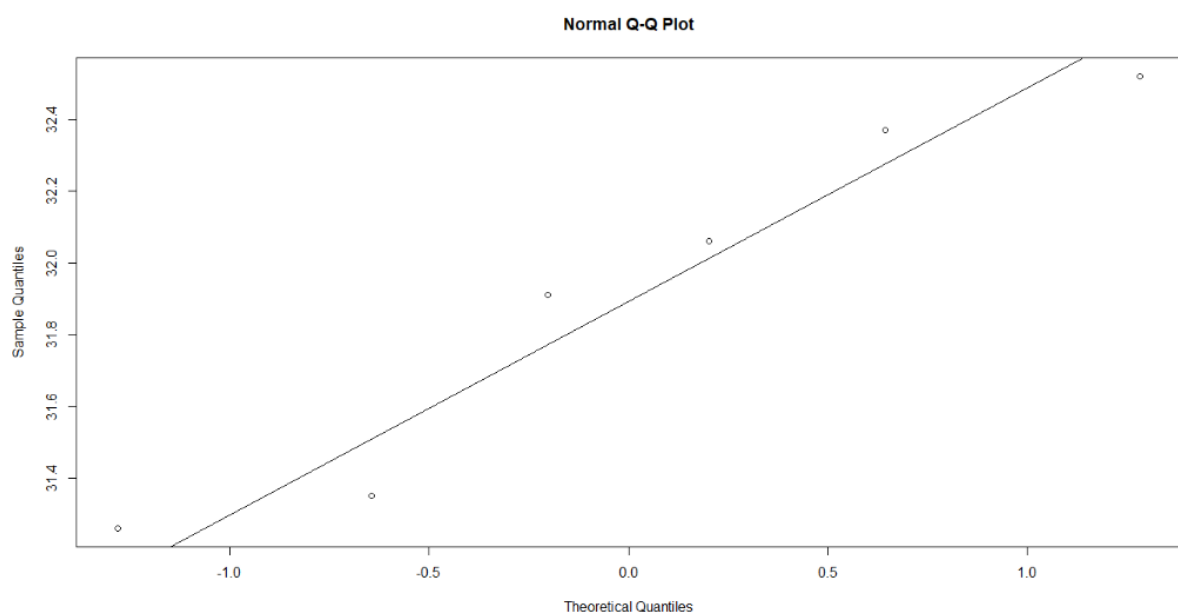
In brief, explain what this plot shows and which conclusion can be drawn with regard to the problem described in task (26).

Part 6

(28) The length of human pregnancies is approximately normally distributed with mean μ= 266 days and standard deviation σ = 16 days.

   –   What is the probability a randomly selected pregnancy lasts less than 260 days?

- Suppose a random sample of 20 pregnancies is obtained. Describe the sampling distribution of the sample mean length of human pregnancies.

- What is the probability that a random sample of 20 pregnancies has a mean gestation period of 260 days or less?

- What is the probability that a random sample of 50 pregnancies has a mean gestation period of 260 days or less?

- What might you conclude if a random sample of 50 pregnancies resulted in a mean gestation period of 260 days or less?

(29) The mean weight gain during pregnancy is 30 pounds, with a standard deviation of 12.9 pounds. Weight gain during pregnancy is skewed right. In a certain neighborhood the mean of a sample with 35 persons is 36.2. Is this result unusual?

(30) 15% of all Americans have hearing trouble.

a) In a random sample of 120 Americans, what is the probability at most 12 % have hearing troubles?

b) Suppose that a random sample of 120 Americans who regularly listen to music using headphones results in 26 having hearing trouble. What might you conclude?

(31) According to the American Red Cross, 7% of people in the United States have blood type 0-negative. If you would want to determine the probability of a single random sample of 500 people in the US with fewer than 30 to have blood type 0-negative, you could get the solution with the following two different code snippets in R.

Code 1:

```
n <- 500
p <- 0.07
pbinom(29,n,p)
```

Code 2 (with n*p*(1-p)>9):

```
n <- 500
p <- 0.07
pnorm(29.5, n*p, sqrt(n*p*(1-p)))
```

Explain the difference between the two and why both give at least approx. the same result.

To be continued …