List of typical exam questions

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 the questions from which the exam questions are chosen!

Basics – Data Analysis and R (Introduction)

(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?

Simple Linear Regression (descriptive)

(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? Which regression model is used? Which are the unknown parameters that are determined through the analysis?
- What does the resulting model look like, i.e. what formula for the line that best describes the sample data is suggested by the result? Interpret the model?
- According to the R output, how well does the model describe the variation observed in the dependent variable, or in other words, how well does it fit to the sample data it is based on?
- Visualize (graph) the resulting model?

Probability and Combinatorics

(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 4^{th} 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.

Discrete Probability Distributions

(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.

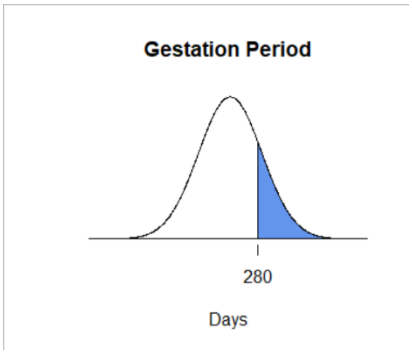
(21) A shipment of 120 fasteners that contains 4 defective items was sent to a manufacturing plant. The quality control manager at the plant randomly selects 5 fasteners. What is the probability that exactly one of the selected ones is defective?

(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?

Continuous Probability Distributions

(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.



(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.

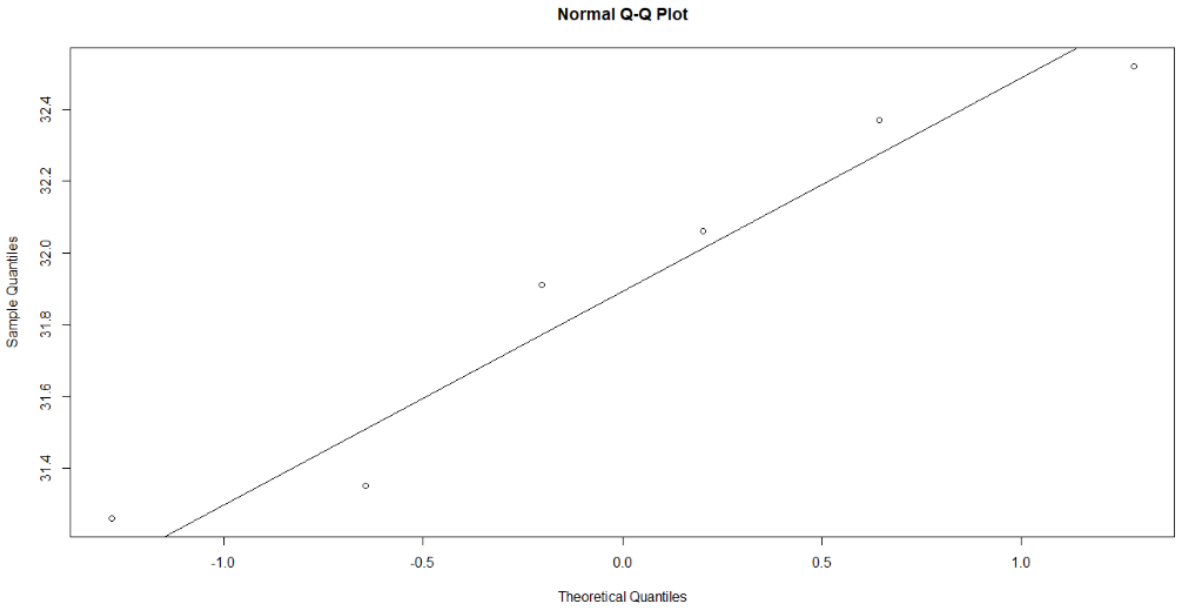
(26) The data in the following table represent the finishing times (in seconds) for six randomly selected races of a greyhound named Barbies Bomber. Is there evidence to support the belief that the variable "finishing time" is normally distributed?

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:



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

Sampling Distributions

(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 options and why both give at least approx. the same result.

Estimation

(32) Every Monday, the Energy Information Admission (EIA) determines the national average gasoline price by collecting retail prices for gasoline from a sample of 900 retail gasoline outlets from across the nation. On July 14, 2008, the EIA reported the national average retail price for regular grade gasoline to be $ 4.113 per gallon.

a) Assuming that the population is normally distributed and the population standard deviation is σ=$ 0.110 per gallon, construct and interpret a 95% confidence interval for the national mean price per gallon for regular-grade gasoline on July 14, 2008.

b) An administrator of a rural town finds that the mean price of gasoline for all five gasoline stations in her town is $ 4.263 per gallon. Should she conclude that the interval obtained from the EIA is inaccurate? Why?

(33) A researcher asked a random sample of 32 adults, "How many days per week do you participate in exercise activities?" The sample did not have outliers and a normal-q-q-plot indicated that the number of days per week in which adults engage in exercise activities is normally distributed. The sample resulted in a sample mean of $\bar{x}$ = 3 and a sample standard deviation of s = 2.24. Construct and interpret a 95% confidence interval for the mean number of days per week in which adults engage in exercise activities.

(34) Consider the sample data in the following table. The data represent diameters of randomly selected white oak trees in a forest reserve.

Oak Tree Diameters in cm in a forest preserve, simple random sample			
64.0	33.4	45.8	56.0
51.5	29.2	63.7	

Construct a 95% confidence interval for the mean diameter of mature white oak trees in the forest reserve. Interpret this interval.

Statistical Tests

(35) According to a survey in 2016 the mean travel time in a city had been 25.3 min with a standard deviation of 8.4 min. The city's department of transportation just reprogrammed all traffic lights in order to reduce travel time. In order to determine if the reprogramming of the traffic lights had the desired effect the administration then collected travel time data from a sample of 2500 commuters (n=2500). The mean travel time of the sample: 24.9 min.

a) Did the reprogramming of the traffic lights have the desired effect on the travel time in the city? Do a statistical test of the effect at the $\alpha=0.10$ level of significance. Did the applied measure decrease the travel time in a statistically significant way? Do not forget to state the null and alternative hypotheses. Decide which hypothesis test is appropriate and reason your decision.

b) Interpret the result of subtask a. Is the decrease of any practical significance?

c) Test the hypothesis at the $\alpha=0.10$ level of significance with n=50. Leave all other parameters at the same level. Is a sample mean of 24.9 min significantly lower than former mean of 25.3 min? Compare this result with your findings in a and b. What do you conclude?

(36) A label on a bag with potato chips states that the bag contains 12.5 g potato chips with a standard deviation of 0.12g. A consumer advocate believes that the manufacturer of the potato chips is underfilling the bags. A random sample of size n = 36 results in a sample mean = 12.45g. Does the sample data support the believes of the consumer advocate?

(37) 15% of all Americans have hearing trouble. 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?

(38) According to the Centers for Disease Control and Prevention, in 2005, 15.2% of tenth grade students had tried marijuana. The Drug Abuse and Resistance Education (DARE) program underwent several major changes to keep up with technology and issues facing students in the 21st century. After the changes, a school resource officer(SRO) thinks that the proportion of tenth-grade students who have tried marijuana has decreased from the 2005 level.

a) Determine the null and alternative hypotheses.

b) If sample data indicate that the null hypothesis should not be rejected, state the conclusion of the SRO.

c) Suppose, in fact, that the proportion of tenth-grade students who have tried marijuana is 14.7%. Was a Type I or Type II error committed?

(39) To test $H_0: \mu = 50$ vs $H_1: \mu < 50$, a random sample of size n=24 is obtained from a population that is known to be normally distributed with $\sigma=12$.

a) If the sample mean is determined to be $\bar{x} = 47.1$, compute the test statistic.

b) If the researcher decides to test this hypothesis at the $\alpha=0.05$ level of significance, determine the critical value.

c) Draw a normal curve that depicts the critical region.

d) Will the researcher reject the null hypothesis? Why?

(40) To test $H_0: \mu = 50$ vs $H_1: \mu < 50$, a random sample of size n=24 is obtained from a population that is known to be normally distributed.

a) If $\bar{x} = 47.1$ and s=10.3, compute the test statistic.

b) If the researcher decides to test this hypothesis at the α=0.05 level of significance, determine the critical value.

c) Draw a t-distribution that depicts the critical region.

d) Will the researcher reject the null hypothesis? Why?

(41) 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 ...


> 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? Which regression model is used? Which are the unknown parameters that are determined through the analysis?
- What does the resulting model look like, i.e. what formula for the line that best describes the sample data is suggested by the result? Interpret the model?
- The resulting model is an estimate for the true regression line for the population. The coefficients of the resulting model from the sample data are estimates for the unknown parameters of the true regression line.

  ➔ Determine the 95% confidence intervals for the coefficients. Interpret the results.

➔ The lm-call involves performing hypothesis tests on the coefficients of the model. Which hypothesis tests have been performed on the coefficients in the given example? Interpret the results.