Exam questions

2023-05-27

PROBABILITY THEORY

Mock 2023 - 1. Probability Theory (6 pts)

Consider an experiment of throwing a fair, four-sided die until get the number '4'. The outcome of the experiment is the number throws it took.

(a) (1 pt) Define a natural sample space Ω for this experiment, or in other words, a sample space Ω that allows modeling all possible outcomes of this experiment"?.

(b) (2 pts) Write down the set of outcomes corresponding to each of the following events

- A: "The number of throws is even."
- B: "It took at least 3 throws."

(c) (3pts) What is the probability of the event $A \cap B^{\mathrm{C}}$?

Mock 2022 - 1. Probability Theory (6 pts)

Consider an experiment where you flip a fair coin four times.

(a) (1 pt) Define a natural sample space Ω for this experiment.

(b) (2 pts) Write down the set of outcomes corresponding to each of the following events

- A: "We get exactly one tails"
- B: "The coin always comes with the same side up."

(c) (1 pt) Summarise in words the meaning of the event $A \cup B$.

(d) (2 pts) Compute the probability for the event $C = (A \cup B)^c$.

Mock 2021 - 1. Probability Theory (5 pts)

Assume that you pick two marbles, one at the time and without replacement, from a bag that contains many blue (B), and many red (R) marbles. Further assume that, in the beginning there are N > 2 blue marbles and M > 2 red marbles in the bag.

(a) What is the sample space of the experiment? (1pt)

(b) List all the outcomes where blue marble is obtained at the first pick. What is the probability of this event? (1pt)

(c) What is the probability of the event picking first blue then red? (1pt)

(d) List all the outcomes where a red marble is picked after picking a blue. What is the probability of picking red marble as the second conditioned to that the first one is blue? (1pt)

(e) What is the probability of the event picking the same colour for both the two marbles? (1pt)

Re-exam 2020 - 1. Probability Theory (6 pts)

A ball is drawn at random from an urn initially containing one red (R) and one green ball (G). If the green ball is drawn, it is put back into the urn. If a red ball is drawn, it is put back into the urn together with an additional red ball. One then repeats the drawing two more times in the same way without returning to the initial condition between the draws.

(a) What is the sample space of the experiment? (1pt)

(b) What is the probability of getting a green ball in all the three draws? (1pt)

(c) What is the probability of getting red ball in all the three draws? (1pt)

(d) What is the probability of getting a green ball on the second draw conditioned on that the first drawn ball is red? (1pt)

(e) What is the probability of getting two red balls and one green ball in the experiment when the order does not matter? List all the possible outcomes that yield the result. (2pts)

Exam 2020 - 1 Probability Theory (5 pts)

a. Assume that you pick two marbles, one at the time and without replacement, from a bag that contains many blue (B), and many red (R) marbles. Further assume that, in the beginning there are $N > 2$ blue marbles and $M > 2$ red marbles in the bag.

b. What is the sample space of the experiment? (1pt)

c. List all the outcomes where blue marble is obtained at the first pick. What is the probability of this event? (1pt)

d. What is the probability of the event picking first blue then red? (1pt)

e. List all the outcomes where a red marble is picked after picking a blue. What is the probability of picking red marble as the second conditioned to that the first one is blue? (1pt)

f. What is the probability of the event picking the same colour for both the two marbles? (1pt)

Mock 2020 - 1. Probability (8pt)

Consider an experiment where you twice roll a fair die with four sides. The values on the sides of the die are 1, 2, 3 and 4.

(a) (1pt) Choose a natural sample space for this experiment.

(b) (1pt) Write down the set of outcomes corresponding the each of the following events • A : "the sum of the two die rolls is greater than or equal to 6" • B : "the value of the first roll is strictly smaller than the value of the second roll" • C : "the values of the fist and second roll are the same"

(c) (1pt) Calculate the probability of the three events above, that is P(A), P(B) and P(C).

(d) (2pt) Explain in words what the probability P(A|C) means and calculate the probability P(A|C). Now consider an experiment where you roll the die 10 times. Let X be the random variable representing the number of times you roll a 2.

(e) (3pt) What is the distribution of X? What is the probability of rolling a 2 exactly 5 times in 10 rolls, i.e. P(X = 5)?

Exam 2019 - 1. Probability Theory (6 pts)

Consider an experiment where you flip a fair coin three times.

(a) (1 pt) Define a natural sample space for this experiment.

(b) (2 pts) Write down the set of outcomes corresponding to each of the following events

- A: "We get at least one tails"
- B: "The coin always comes with the same side up."

(c) (1 pt) Summarize in words the meaning of the event $A \cap B$.

(d) (2 pts) Compute the probability for the event

$$C = (A \cup B)^c$$

Exam 2019 - 2.Conditional probability (5 pts)

Your company builds metal detectors for airports, to be used at the security control before boarding airplanes. A person walking has a probability 0.1 of wearing a metal object while going through the metal detector, an event that we denote with M (with MC we denote the complement of M). Therefore we have that P[M] = 0.1. When the metal detector detects a metal object, it flashes a light, an event that we denote with T. The metal detector is built in a way such that

- P[T|M] = 0.7 • P[T|Mˆc] = 0.4

(a) (2pts) Compute the probability that a person is actually carrying a metal object if the metal detector flashes the light, that is compute P[M|T].

(b) (2pts) Given the result of the previous calculation, the metal detector you are producing is likely to form long lines at the airport, making everybody unhappy. You have a limited budget, that you can either spend to to increase P[T|M] or to decrease P[T|Mˆc]. Before deciding where to spend your money, you want to be sure that any change will have a positive effect. Suppose you set P[T|M] = 1, leaving P[T|Mˆc] = 0.4. What is now the value of P[M|T]?

(c) (1pt) Your goal is to obtain a metal detector such that P[M|T] $\geq$ 0.9. How small do you need to make P[T|Mˆc] (while maintaining P[T|M] = 0.7)?

Mock 2019 - 1 Probability Theory (8pt)

Consider an experiment where you twice roll a fair die with four sides. The values on the sides of the die are 1, 2, 3 and 4.

(a) (1pt) Choose a natural sample space Ω for this experiment.

(b) (1pt) Write down the set of outcomes corresponding the each of the following events

- A : "the sum of the two die rolls is greater than or equal to 6"
- B : "the value of the first roll is strictly smaller than the value of the second roll"
- C : "the values of the fist and second roll are the same"

(c) (1pt) Calculate the probability of the three events above, that is $P(A)$, $P(B)$ and $P(C)$.

(d) (2pt) Explain in words what the probability $P(A|C)$ means and calculate the probability $P(A|C)$.

Now consider an experiment where you roll the die 10 times. Let X be the random variable representing the number of times you roll a 2.

(e) (3pt) What is the distribution of X? What is the probability of rolling a 2 exactly 5 times in 10 rolls, i.e. $P(X = 5)$?

Exam 2017 - 1. Probability Theory (5pts)

Consider an experiment where you are flipping a coin four times.

(a) (1pt) Choose a natural sample space Ω for this experiment.

(b) (1pt) Write down the set of outcomes corresponding to each of the following events

- A : "we throw tails at least three times"
- B : "both the first and the last throw results in tails"

(c) (1pt) Assuming that the coin is fair, calculate the probability of the event $A \cap B$.

(d) (2pts) Now, suppose that the coin is biased such that heads is three times more likely than tails. What is the probability of tails?

Exam 2017 - 2.Conditional probability (6pts)

Suppose that you are testing bicycle racers for use of EPO doping in the Tour de France with a urine test. Let E be the event that "a bicyclist is doped using EPO", and let T be the event that "the urine test indicates that the bicyclist is doped with EPO". Suppose that 10% of the bicyclists in the Tour de France are doped using EPO and that the test has the properties P(T | E) = p and P(T | Ec) = 1 - p.

(a) (1pt) Describe in words the meaning of P(E | T) and P(E | Tc).

(b) (2pts) Compute P(E | T) if p = 0.92 and comment on the credibility of the test.

(c) (3pts) Find p such that P(E | T) = 0.95.

DISTRIBUTIONS, EXPECTATION, VARIANCE, INDEPENDENCE

Mock 2023 - 2. Continuous distributions, Expectation, Variance (6 pts)

Let X be a continuous random variable with the probability density function

$$f_X(x) = \begin{cases} \frac{3}{4}(1-x^2), & x \in [-1,1], \\ 0, & \text{otherwise.} \end{cases}$$

(a) (2pts) Compute the expected value of X.

(b) (2pts) Compute the variance of X.

(c) (2pts) Compute the probability $P(X > 0.5)$.

###Mock 2022 - 2. Expectation, Variance, Discrete Distributions (6 pts)

Let us consider the experiment of independently throwing two fair dice characterised by the discrete random variables X and Y, respectively. The discrete random variables X and Y take the values $a = 1, 2, 3, \ldots, 6$ and $b = 1, 2, 3, \ldots, 6$, respectively.

(a) Compute the expected value for the product $Z = XY$. (2pts)

(b) Write down the probability mass function for the the discrete random variable Z defined by the product $Z = XY$. (2pts)

(c) Compute the variance of the product $Z = XY$. (2pts)

Mock 2021 - 2. Expectation, Joint Distributions, Independence (5 pts)

Let (X, Y) be a pair of continuous random variables with the joint density function:

$$f(x,y) = \begin{cases} 1 & if \quad (x,y) \in [0,1] \times [1,2] \\ 0, & \text{otherwise.} \end{cases}$$

(a) Compute the expected value of the vector (X, Y). (1pt)

(b) Compute the covariance between the the random variables X and Y. AreX and Y correlated? (2pts)

(c) Find the the marginal densities of the random variables X and Y. Are X and Y are independent? (2pts)

Re-exam 2020 - 2.Continuous Random Variable (6 pts)

Assume that the score a student can get in an exam is a real number in the interval [0,100], and the student passes the exam if the score is at least 55. Assume that we model the score of the exam by the random variable X that has the probability density function Certainly! Here's the formula written in LaTeX:

$$f(x) = \begin{cases} x - \frac{4p}{5}, & \text{for } 80p < x \leq 80p + 10 \\ -\frac{x}{100} + \frac{4p}{5} + \frac{1}{5}, & \text{for } 80p + 10 < x < 80p + 20 \\ 0, & \text{otherwise} \end{cases}$$

where p, $0 \leq p \leq 1$ is a parameter that describes how well the student prepared for the exam: $p = 0$ refers to no preparation and $p = 1$ excellent preparation. Assume that $p = 0.5$, i.e., the student did not prepare very well. Using this model, compute

(a) the expected score the student will get, and (3pts)

(b) the probability that the student will pass the exam. (3pts)

Exam 2020 - 2 Expectation, Joint Distributions, Independence (5 pts)

Let (X, Y) be a pair of continuous random variables with the joint density function

$$f(x,y) = \begin{cases} 1 & if(x,y) \in [0,1]x[1,2] \\ 0 & otherwise \end{cases}$$

(a) Compute the expected value of the vector (X, Y). (1pt)

(b) Compute the covariance between the the random variables X and Y. Are X and Y correlated? (2pts)

c) Find the the marginal densities of the random variables X and Y. Are X and Y are independent? (2pts)

MAXIMUM LIKELIHOOD

Mock 2023 - 3. Maximum Likelihood (6 pts)

Let $x_1, x_2, x_3, \ldots, x_N$ be a dataset where all the observations are independent and identically distributed (i.i.d.) following a continuous distribution with the probability density function $f_\alpha(x)$ given by

$$f_\alpha(x) = \begin{cases} \mathrm{e}^{-(x-\alpha)} & \text{for } x \geq \alpha, \\ 0, & x < \alpha, \end{cases}$$

where α is a parameter.

(a) (2pt) Write down the likelihood function $L(\alpha)$ corresponding to the dataset above.

(b) (4pts) Determine the maximum likelihood estimate for α.

Mock 2022 - 3. Maximum likelihood (4pts)

Let x_1, x_2, ..., x_n be a dataset that is a realisation of a random sample from a $U(\alpha, \beta)$ distribution, where α and β are the unknown parameters.

(a) Write down the likelihood function of the parameters. (2pts)

(b) Determine the maximum likelihood estimates for the parameters α and β. (2pts)

Mock 2021 - 3. Maximum likelihood (6pts)

Let us consider a random process that is a sequence of independent random variables Z1, Z2, Z3, . . . so that for each i the value of Zi is either 1, 2, or 3. For all values of i the probability p that Zi = 1 is the same. Likewise, for all values of i the probability q that Zi = 2 is the same. Now, assume that you have a sequence of 100 observations from a such process.

(a) Write down the likelihood function of the data given the parameters. (2pts)

(b) Derive the maximum likelihood estimator for the parameters p and q. (4pts)

Exam 2020 - 3 Maximum likelihood (6pts)

Let us consider a random process that is a sequence of independent random variables $Z_1, Z_2, Z_3, ...$ so that for each i the value of Z_i is either 1,2, or 3. For all values of i the probability p that Z_i =1 is the same. Like wise, for all values of i the probability q that Z_i= 2 is the same. Now, assume that you have a sequence of 100 observations from a such process.

(a) Write down the likelihood function of the data given the parameters. (2pts)

(b) Derive the maximum likelihood estimator for the parameters p and q. (4pts)

SMALL R PROBLEMS

Mock 2023 - 4. Small R Problems (6 pts)

(a) (2pts) The historical data set `Langren1644` (`HistData`) contains the all the available estimates, in 1644, for the distance of Toldeo and Rome in longitude. Compute simple statistics such as mean, median, min, and max and standard deviation on the longitude. What can you conclude from these numbers?

(b) (2pts) The data set `iq` ('UsingR) contains simulated IQ scores. Find the 95% confidence interval for the mean (simulated) IQ by assuming that this dataset is a random sample from a normal distribution with *unknown* standard deviation.

Hint: you have to translate the correct formula for the confidence intervals to R code. In particular, condsider that the distribution generating the samples is normal and that its standard deviation (and hence the variance) is unknown (even though you can compute an estimate for it from the data).

(c) The data set emissions (UsingR) contains the information of CO2 emissions of different countries for 1999. Use a linear regression model to model the CO2 emission as a function of GDP per capita. Visualise the model together with the observations. What can you conclude from your results?

Mock 2022 - 4. Small R Problems (8 pts)

(a) The data set `firstchi (UsingR)` contains the age of the mother at birth of the first child. Investigate the data set by computing several simple statistics on this data set. Summarise your findings. (2 pts)

(b) The data set `rat (Using R)` contains the survival times of 20 rats exposed to radiation. Visualise the data in appropriate means and discuss in the light of the data and your knowledge, what kind of parametric model would you choose for the distribution of the survival times. (2 pts)

(c) Consider the dataset `kid.weights (UsingR)`, that reports information about a sample of 250 kids. Select the kids up until 9 years old (i.e. with an age strictly less than 108 months). Plot a scatter plot of the weight versus the height. Compute a linear regression model and add the regression line to the plot. What conclusions can you derive? (2 pts)

(d) Assume that you have implemented a scientific method that you compare to the state-of-the-art published elsewhere. You use the evaluation metric G where a bigger value refers to a better outcome. Using independent experiments you get five scores for the state-of-the-art (0.908,0.915,0.908,0.905,0.904) and five for your method (0.910,0.914,0.909,0.914,0.910). State the null and alternative hypothesis and test if there is statistical evidence that your method is better than the state-of-the-art. (2 pts)

Mock 2021 - 4. Small R Problems (8 pts)

Below the data sets are found in the packages specified in parenthesis. For instance, Langren1644 (HistData) refers to the Langren1644 data set in the package HistData.

(a) Simple statistics (2pts) The historical data set Langren1644 (HistData) contains the all the available estimates, in 1644, for the distance of Toldeo and Rome in longitude. Compute simple statistics such as mean, median, min, and max and standard deviation on the longitude. What can you conclude from these numbers?

(b) Data Exploration (2pts) The data set ozonemonthly (UsingR) contains a time series showing ozone values at Haley Bay in Antarctica. Visualize the data in appropriate means and discuss what you can learn from this data set.

(c) Linear Regression (2pts) The data set emissions (UsingR) contains the information of CO2 emissions of different countries for 1999. Use a linear model to model the CO2 emission as a function of GDP per capita. Visualise the model together with the observations. What can you conclude from your results?

(d) Confidence Intervals (2pts) The data set iq (UsingR) contains simulated IQ scores. Does the data seem normally distributed? Find the 95% confidence interval for the mean (simulated) IQ.

Re-exam 2020 - 4.Small R Problems (6 pts)

(a) Simple statistics (2pts) The data set firstchi (UsingR) contains the age of the mother at birth of the first child. Investigate the data set by computing several simple statistics on this data set. Summarise your findings.

(b) Data Exploration (2pts) The data set ewr (UsingR) contains taxi in and taxi out times for 8 different airlines at Newark airport for 1999-2001. Visualise the data set in a suitable fashion and state your conclusions about the taxi in and out times related to the airport.

(c) Confidence Intervals (2pts) The data set galton (UsingR) is a dataframe containing child's a height and midparet height. Does the data seem normally distributed? Find the mean and 95% confidence interval for the child height and midparent height.

Exam 2020 - 4 Small R Problems (8 pts)

Below the data sets are found in the packages specified in parenthesis. For instance, Langren1644 (HistData) refers to the Langren1644 data set in the package HistData.

(a) Simple statistics (2pts)

The historical data set Langren1644 (HistData) contains the all the available estimates, in 1644, for the distance of Toldeo and Rome in longitude. Compute simple statistics such as mean, median, min,and max and standard deviation on the longitude. What can you conclude from these numbers?

(b) Data Exploration (2pts)

The data set ozonemonthly (UsingR) contains a time series showing ozone values at Haley Bay in Antarctica. Visualise the data in appropriate means and discuss what you can learn from this data set.

(c) Linear Regression (2pts)

The data set emissions (UsingR) contains the information of CO2 emissions of different countries for 1999. Use a linear model to model the CO2 emission as a function of GDP per capita. Visualise the model together with the observations. What can you conclude from your results?

(d) Confidence Intervals (2pts)

The data set iq (UsingR) contains simulated IQ scores. Does the data seem normally distributed? Find the 95% confidence interval for the mean (simulated) IQ.

Mock 2020 - 4.Small R Questions (6pt)

(4a) Correlation (2pt)

The maydow (UsingR) data set contains the Dow Jones industrial average and the maximum daily temperature in New York City for May 2003. Make a scatter plot of the industrial average against the maximum temperature. Are the Dow Jones index and the temperature correlated?

(4b) Simple statistics (2pt)

The babies (Using R) data set is a collection of variables taken for each new mother in a study. The variable age contains the mothers age in years if known, otherwise the value is 99. Compute the minimum, maximum, mean, median, and standard deviation of those mothers' age which are known. What can you conclude of the age distribution on the basis of these numbers?

(4c) Linear Regression (2pt)

The data set wellbeing (UsingR) contains factors affecting people's happiness in several countries. Fit a linear model with the alcohol consumption as the explanatory variable, and estimate how many percentage the well-being is increased, in the reported units, if the alcohol consumption is decreased from 16 to 8 units.

Exam 2019 - 4. Small R Problems (8 pts)

(a) (2 pts) The dataset Orange (UsingR) contains information about the growth of orange trees. Summarise the circumference of these trees by reporting the minimum, maximum, mean, standard deviation, and median. What can you say about the distribution of the circumferences on the basis of these numbers?

(b) (2 pts) Consider the father.son (UsingR) dataset, which contains information about the height of fathers and their sons. Plot the kernel density estimate of the height of fathers and the kernel density estimate of the height of sons. What can you conclude from the plots?

(c) (2 pts) Consider the dataset kid.weights (UsingR), that reports information about a sample of 250 kids. Select the kids up until 9 years old (i.e. with an age strictly less than 108 months). Plot a scatter plot of the weight versus the height. Compute a linear regression model and add the regression line to the plot. What conclusions can you derive?

(d) (2 pts) The data set normtemp (UsingR) contains measurements of 130 healthy, randomly selected individuals in Fahrenheit. The variable temperature contains normal body temperature. Assume that the data is normally distributed. Test the hypothesis that the normal temperature has the mean at 37 °C. The Fahrenheit can be converted Celsius degrees by the formula TC = 5/9*(TF - 32).

Mock 2019 - 3 Small R Questions (8pt)

(a) Correlation (2pt)

The `maydow (UsingR)` data set contains the Dow Jones industrial average and the maximum daily temperature in New York City for May 2003. Make a scatter plot of the industrial average against the maximum temperature. Are the Dow Jones index and the temperature correlated?

(b) Simple statistics (2pt)

The `babies (Using R)` data set is a collection of variables taken for each new mother in a study. The variable `age` contains the mothers age in years if known, otherwise the value is 99. Compute the minimum, maximum, mean, median, and standard deviation of those mothers' age which are known. What can you conclude of the age distribution on the basis of these numbers?

(c) Linear Regression (2pt)

The data set `wellbeing (UsingR)` contains factors affecting people's happiness in several countries. Fit a linear model with the alcohol consumption as the explanatory variable, and estimate how many percentage the well-being is increased, in the reported units, if the alcohol consumption is decreased from 16 to 8 units.

(d) Confidence intervals (2pt)

The variable `weight` in `kid.weights (UsingR)` data set contains the weights of a random sample of children. Find a 90% confidence interval for the mean weight of 5-year-olds. You'll need to isolate just the 5-year-olds' data first.

Exam 2017 - 4. Small R Questions (8pt)

(4a) Simple statistics (2pts) In the datasets BushApproval and ObamaApproval (UsingR) the variables approval are collections of approval ratings of President Bush and President Obama. Summarise and compare the approval ratings by computing suitable simple statistics. Which one of the presidents seems more popular?

(4b) Data Exploration (2pts) The dataset cancer (UsingR) contains survival times (days) of patients with five different cancer types. What can you conclude about the survival times of each cancer types on the basis of your knowledge and this data set?

(4c) Linear Regression (2pts) In the dataset Prostitutes (HistData) the variable count indicates the number of prostitutes in Paris between the years 1812-1854. Fit a linear model to the observations, plot the number of prostitutes as a function of time together with the regression line. Predict for how many prostitutes there are in Paris today according to the model. What would you say about the validity of the prediction you made?

(4d) Hypothesis testing (2pts) The dataset carbon (UsingR) contains carbon monoxide (Monoxide) at different sites (Site). Is there significant difference in the mean of the monoxide measurements between the sites 1 and 3? Assume that both samples are normally distributed with equal variances.

CONFIDENCE INTERVALS

Exam 2019 - 3.Confidence intervals (3 pts)

You are performing an experiments, and you know that your results are a random sample taken from a normal distribution with standard deviation $\sigma = 2.2$.

(a)(2 pt) Suppose your random sample has size 30, with sample mean $\bar{x}_{30} = 12.4$. Compute a two-sided 95% confidence interval for the mean.

(b) (1 pt) Suppose you want the two-sided 95% confidence interval to have width ≤ 0.5. How many samples do you need to take at least?

Exam 2017 - 3. Confidence intervals (5pt)

Suppose that you want to examine the amount of caffeine in a cup of espresso from Café Analog. In a sample of 10 cups, the average measured caffeine contents is $\bar{x}_{10} = 62.9$ (mg) and the sample standard deviation is $s_{10} = 5.8$ (mg). You may assume that the measurements are a realization of a random sample form an $N(\mu, \sigma^2)$ distribution.

(a) (3pts) Construct a 95% confidence interval for μ.

(b) (2pts) Assuming that the standard deviation is $\sigma = 5$ (mg), how many measurements do we need, if we want a 95% confidence interval for the mean μ no wider than 1 (mg)?

HYPOTHESIS TESTING

Mock 2023 - 5. Hypothesis testing (8 pts)

Consider a practical situation that you evaluate two brands of ice cream A and B. The reported weight of each ice cream package is nominally the same (100g) but after consuming of the two brands, you feel that the ice cream A runs out earlier than the ice cream B. To investigate if your feeling is suggesting right, you want to make a statistical test evaluating if there is less ice cream in the brand A package when compared to the brand B package. To do the testing, you have 5 packages from each ice cream brand, and you weigh them to get the measurements $W_A = (99.2, 100.5, 98.9, 99.6, 97.7)$ for the brand A and $W_B = (100.5, 100.1, 99.1, 98.2, 98.3)$ for the brand B, where the measurements are in grams.

(a) (1pt) Formulate the appropriate null hypothesis and alternative hypothesis.
(b) (1pt) Why is bootstrapping a good strategy for testing the hypothesis, if compared to t-testing?
(c) (5pts) Make a program that tests the null hypothesis by bootstrapping.
(d) (1pt) Report your conclusions at significance level $\alpha = 0.05$.

Mock 2021 - 5. Bootstrapping & Hypothesis testing (8 pts)

The data set chicken (UsingR) contains the weight gain of chicken fed with three different rations. You would like to find if there is statistical evidence that the ration 2 gave more weight gain than ration1.

(a) Formulate the appropriate null hypothesis and alternative hypothesis. (1pt)

(b) Why is bootstrapping a good strategy in testing the hypothesis? (1pt)

(c) Make a program that tests the null hypothesis by bootstrapping. (5pts)

(d) Report your conclusions at significance level $\alpha = 0.05$. (1pt)

Re-exam 2020 - 3.Hypothesis testing (6pts)

The paired test is an another from of hypothesis testing that suites for a situation when two samples depend on each other is some way, for instance, the samples from twin studies. Two paired samples can be compared by a one-sample test. For paired samples x1, x2, . . . , xn and y1, y2, . . . , yn (where (xi , yi), i = 1, 2, . . . , n are the pairs) one then considers the differences x1 - y1, x2 - y2, . . . , xn - yn (instead of incorrectly using the two-sample test with the assumption of two independent sets of samples). Consider the data babies (UsingR). The variable age contains the mom's age and dage the corresponding (paired) dad's age. You would like to assess if there is statistical evidence that the child's mum is younger than child's dad.

(a) Formulate the appropriate null hypothesis and alternative hypothesis. (2pts)

(b) Which test is appropriate for testing the hypothesis? Explain why. (2pts)

(c) Do the test and report your conclusion at significance level $\alpha = 0.05$. (2pts) Hint: if the age of the mum or dad is unknown it is indicated by the age 99 in the dataset, you need to filter out these first, for instance, by filtered.babies <- subset(babies,age<99 & dage<99).

Exam 2020 - 5 Bootstrapping & Hypothesis testing (8 pts)

The data set chicken (UsingR) contains the weight gain of chicken fed with three different rations. You would like to find if there is statistical evidence that the ration 2 gave more weight gain than ration 1.

(a) Formulate the appropriate null hypothesis and alternative hypothesis. (1pt)

(b) Why is bootstrapping a good strategy in testing the hypothesis? (1pt)

(c) Make a program that tests the null hypothesis by bootstrapping. (5pts)

(d) Report your conclusions at significance level $\alpha = 0.05$. (1pt)

Mock 2020 - 3. Hypothesis Testing (6pt)

According to the Danish ministry of education, when the Danish 7-point grading scale (consisting of the grades -3, 00, 02, 4, 7, 10 and 12) is used nationally and over a long period of time, the mean pass grade should be 7. In the year 2028 you are looking at the grades for the 681 student who have passed the course in Applied Statistics at ITU in the past 10 years. The sample mean and sample standard deviation of the grades are x¯681 = 7.15 and s681 = 3.07. You observe that that the sample mean is higher than 7, so you decide to investigate if the mean grade for the Applied Statistics course may be different from 7 using hypothesis testing. With 681 observations you can assume that this is a large sample size.

(a) (1pt) Formulate an appropriate null hypothesis and alternative hypothesis.

(b) (1pt) Which test is appropriate for testing the hypothesis? Explain why.

(c) (2pt) Compute the value of the test statistic and report your conclusion at significance level $\alpha = 0.05$.

(d) (2pt) Compute the corresponding one-tailed p-value. Is the evidence against the null hypothesis strong?

Mock 2019 - 2 Hypothesis Testing (8pt)

According to the Danish ministry of education, when the Danish 7-point grading scale (consisting of the grades -3, 00, 02, 4, 7, 10 and 12) is used nationally and over a long period of time, the mean pass grade should be 7. In the year 2028 you are looking at the grades for the 681 student who have passed the course in Applied Statistics at ITU in the past 10 years. The sample mean and sample standard deviation of the grades are $\bar{x}_{681} = 7.15$ and $s_{681} = 3.07$. You observe that that the sample mean is higher than 7, so you decide to investigate if the mean grade for the Applied Statistics course may be different from 7 using hypothesis testing. With 681 observations you can assume that this is a large sample size.

(a) (1pt) Formulate an appropriate null hypothesis and alternative hypothesis.

Null hypothesis -> $H_0 : \mu = 7$

Alternate hypothesis -> $H_1 : \mu > 7$

(b) (1pt) Which test is appropriate for testing the hypothesis? Explain why.

we use a one sample t-test; as we have a large sample size, we assume it follows a standard normal distribution

(c) (3pt) Compute the value of the test statistic and report your conclusion at significance level $\alpha = 0.05$.

(d) (3pt) Compute the corresponding one-tailed p-value. Is the evidence against the null hypothesis strong?

BOOTSTRAPPING

Mock 2022 - 5. Bootstrapping and confidence intervals (8 pts)

Let us look at the datasets `female.inc (UsingR)` that contain income distribution for females in 2001. You may ignore the information about the race of the individuals. Your goal is to estimate the mean female income together with its 95% confidence intervals.

(a) Why is bootstrapping a good strategy for finding the confidence intervals for the mean female income? (1pt)

(b) Write a computer program that computes the bootstrap estimates for the 95% confidence intervals for the mean income of all the females in the dataset. (5pts)

(c) Report your numerical results for the mean and confidence interval together with a graphic illustrating the bootstrapped sample. (2pts)

Mock 2020 - 5.Bootstrapping (6pt)

The variable weight in kid.weights (UsingR) data set contains the weights of a random sample of children. Make a computer program that finds the mean weight of 2-year-olds and its 90% confidence interval by assuming non-normal weight distribution, hence, you will need to use bootstrapping. Hint: You'll need to extract the 2-year-olds' data first.

Exam 2019 - 5. Bootstrapping (8 pts)

Let us look at the datasets BushApproval and ObamaApproval (UsingR). On the basis of the datasets, you want to find if there is evidence of President Bush being more approved than President Obama.

(a) (1pt) Formulate the appropriate null hypothesis and alternative hypothesis.

(b) (1pt) Why is bootstrapping a good strategy for testing the hypothesis?

(c) (5pts) Make a program that tests the null hypothesis by bootstrapping.

(d) (1pt) Report your conclusions at significance level $\alpha = 0.05$.

Exam 2017 - 5. Bootstrapping (8pts)

Let us look the at the datasets BushApproval and ObamaApproval (UsingR) in more detail. On the basis of the data set, you want to find if there is evidence of President Bush being more approved than President Obama.

(a) (1pt) Formulate the appropriate null hypothesis and alternative hypothesis.

(b) (1pt) Why is bootstrapping a good strategy for testing the hypothesis?

(c) (5pt) Make a program that tests the null hypothesis by bootstrapping.

(d) (1pt) Report your conclusions at significance level $\alpha = 0.05$.

SIMULATION

Mock 2019 - 4 Viking Lottery (8pt)

In Viking Lottery one may select 6 numbers from 1 to 48 and an extra "Viking" number from 1 to 8. Make a program that simulates the lottery by drawing six numbers from 1 to 48 without replacement and an independent Viking number from 1 to 8. Choose your luck row of numbers and simulate the lottery by playing it 100 times. How much would you have got as a profit? Assume that playing one row costs 80 cents.

Hits (normal + Viking number)	Wins
6+1	€ 1 000 000
6+0	€ 250 000
5+1	€ 10 000
5+0	€ 1 000
4+1	€ 50
4+0	€ 25
3+1	€ 8
3+0	€ 4

Re-exam 2020 - 5. Simulation (8 pts)

The Monte Carlo method refers to a numerical method of solving mathematical problems by random sampling or by the simulation of random variables. Monte Carlo methods all share the concept of using randomly drawn samples to compute a solution to a given problem. Let us assume that you can only rely on basic arithmetic operations such as addition, subtraction, multiplication and division, but you have an access to uniformly distributed random numbers. You would like to implement a computer program that computes a Monte Carlo, numerical estimate for $\sqrt{2}$.

(a) Explain how the problem can be solved by a Monte Carlo method. (2pts)

(b) Implement a computer program that computes a Monte Carlo estimate for $\sqrt{2}$. (5pts)

(c) How many random numbers do you need to draw in order to find 4 digit accuracy for your estimate? You can evaluate against the reference value $\sqrt{2} \approx 1.4142$ here. (1pt) 3 Hint: For a non-negative real number x, $x < \sqrt{2} \iff x^2 < 2$

Mock 2020 - 2. Football Betting (8pt)

Imagine a football betting setting where there are 13 football games. Each game can have three possible outcomes: home team wins (1), the teams play even (E) or the visitor team wins (2). Model the outcome of each game as a random process where each of the three outcomes are equally probable and independent from other games. Let the random variable X characterise the number of correct guesses for the 13 outcomes in one betting.

(a) (1pt) Write the analytic forms for the probability mass function of X.

(b) (1pt) Illustrate the probability mass function by plotting it in a figure.

(c) (1pt) What is the probability that one get all the 13 outcomes right?

(d) (3pt) Simulate the betting by "playing" the betting 100 times. Present the results you got.

e) (2pt) Assume, one betting costs 0.5 EUR. How would you characterise your chances of getting profit by betting in these games by guessing the results in random? Assume that wins are distributed according to the table below (price is in EUR).

Hits	Price
13	113101
12	7761
11	373
10	63