**LABORATORY WORK 2. QUESTIONS**

|  |  |
| --- | --- |
| QUESTION | ANSWER |
| 1. Your student ID | 20222207 |
| TASK 1 and 2 | |
| 1. What is the population minimum for the price of the flat? | 5000 |
| 1. What is the average area in your sample? | 50 |
| 1. Which median of year is smaller (for population or for your sample)? Write down what numbers you compare. | Population: 15  Sample: 16  Smaller is first one |
| 1. 25 percent of flats in population are larger than …. (what number) | 36 |
| 1. 75 percent of flats (from sample) have are cheaper than … (what number) | 45606 |
| 1. Is it true that 50 percent of all flats are smaller than the average area in your sample? What numbers you compared? | Average: 49  50%: 43  It is true |
| 1. The most expensive flat in your sample is … m2? | 130 |
| 1. Where variation of area is larger (in population or in sample? What numbers you compare? | Population: 399.1119  Sample: 372.8979  Larger in population |
| TASK 3 (10-15 questions is about hypothesis about the average of price) | |
| 1. With what number you compare a price? | 4001 |
| 1. What is your null hypothesis? | 35000 |
| 1. What is your alternative hypothesis? | Two-Sided Test : true mean is not equal to 35000  Less Than Test: true mean is less than 35000  Greater Than Test: true mean is greater than 35000 |
| 1. What is your 95 percent confidence interval? | Two-Sided Test: (28761.26, 34646.99)  Less Than Test: (-Inf, 34171.03)  Greater Than Test: (29237.22, Inf) |
| 1. Which hypothesis (null or alternative) you accepted? | Two-Sided Test: alternative  Less Than Test: alternative  Greater Than Test: null |
| 1. What is your conclusion? | the true mean of price is likely not equal to 35000 and is possibly less than 35000 |
| 1. Can you say that the average area of the flat is not significantly different from 45? | The average area of flats is significantly greater than 45. |
| 1. If mu=7 in hypothesis about the average age of the flat, what conclusion you should make? | The average age of the flats is significantly different from 7 years |
| TASK 4 and 5 | |
| 1. What is the difference between end point and start point in your first interval for the price? | prob=end\_point-start\_point;prob  [1] 0.55636087 0.08157350 0.06657429 0.05433304 0.04434263 0.03618919 0.02953496 0.02410426 0.01967213  [10] 0.01605494 0.07126019 |
| 1. In which interval (for the price) you have the largest amount of flats and how many? | 0-10000$  1,5e-0.5 |
| 1. How many intervals you have for empirical density function of area? | 10 |
| 1. What is the expected rate for the area of the flat (what is the parameter of exponential distribution)? | 40 |
| 1. How many flats you have in the first interval in practice and in theory (for the area)? | 39,36 |
| 1. Can you say, that the area of the flats in your sample is distributed by exponential distribution? Why? | The area of flats in the sample does not follow an exponential distribution. |
| 1. How many intervals you have for empirical density function of year? | 6 |
| 1. What is p-value in hypothesis about the distribution of the year? | 0.2423 |
| 1. What is your conclusion in hypothesis about the distribution of the year? | The 'year' variable could possibly follow a uniform distribution based on the provided data. |
| TASK 6 and 7 | |
| 1. Write down the regression line for the price depending on year (price=\_\_\_+\_\_\_\_\*year). | price=61274-1776\*year |
| 1. Is this relationship strong? Why? | It is not strong because it also depends of area |
| 1. Calculate prognosis of the price for 35 years old flat | -886 |
| 1. Write down the regression line for the price depending on area (price=\_\_\_+\_\_\_\_\*area). | Price=−10860.26+907.29×Area |
| 1. Calculate prognosis of the price for the 160 m 2 flat | 134306.5 |
| 1. What will be the price difference (in euro) if we will compare 160 m 2 and 150 m 2 flats? | 160 m2 – 134306.5  150 m2 – 125233.6 |
| 1. How many percent of the variance between price and area are explained by regression line? | 61.63% |

**Your R code:**

**setwd("C:\\Users\\Laura\\OneDrive\\Рабочий стол\\LW2")**

**getwd()**

**dir()**

**A<-read.table("LW2\_population.txt",header = T, sep="");A**

**str(A)**

**#Task1----**

**# Calculate main characteristics for each variable;**

**summary(A)**

**var(A$year);var(A$area); var(A$price)**

**sd(A$year);sd(A$area); sd(A$price)**

**#Task2----**

**# Set seed as your student ID. Select simple random sample**

**# of 250 flats and calculate main characteristics;**

**set.seed(20222207)**

**help(sample)**

**sample(nrow(A),10) # 10 random values from 1 to 10000, because N = 1000.**

**S=A[sample(nrow(A),250,F), ];S**

**summary(S)**

**var(S$year);var(S$area); var(S$price)**

**sd(S$year);sd(S$area); sd(S$price)**

**boxplot(A$area,S$area)**

**#Task3----**

**# Using sample data test hypothesis about the mean of year, area and price.**

**# (a value you have to choose by yourself, but it can’t be x¯).**

**mean(S$year) # a = 15**

**mean(S$area) # a = 51**

**mean(S$price) # a = 4001**

**t.test(S$year, mu=7,conf.level = 0.95)**

**t.test(S$area, mu=45, alternative="greater")**

**t.test(S$price, mu=35000, alternative = "two.sided")**

**t.test(S$price, mu=35000, alternative="less")**

**t.test(S$price, mu=35000, alternative="greater")**

**#Task4----**

**# Using sample data draw empirical density functions for all variables.**

**par(mfrow=c(3,1))**

**hist(S$year,freq=FALSE,main="edf for Year")**

**hist(S$area,freq=FALSE,main="edf for area")**

**hist(S$price,freq=FALSE,main="edf for price")**

**#Task5----**

**# Using sample data test hypothesis that year is distributed by uniform**

**# distribution and area is distributed by exponential distribution.**

**y=S$year**

**t<-10 # number of intervals**

**k<-(max(y)-min(y))/t**

**intervals=seq(min(y), max(y), k)**

**intervals**

**max(y)**

**y\_gr=table(cut(y, intervals,labels=intervals[-1],**

**include.lowest=T,ordered\_result=T))**

**y\_gr**

**barplot(y\_gr/250)**

**chisq.test(y\_gr/250, p=rep(1/t,t)) # compare with uniform distribution**

**# Exponential distribution**

**# Exponential distribution**

**hist(S$area)**

**y<-S$area**

**max(y)**

**# Exponential distribution**

**lamda=1/mean(y) # parameter**

**# calculation of theoretical probabilities**

**intervals=seq(min(y),max(y)+10,10);intervals # because an exponential**

**#distribution has a long upper tail.**

**m<-length(intervals)**

**m**

**end\_point=pexp(c(intervals[2:(m-1)],Inf), lamda);end\_point**

**start\_point=pexp(c(0,intervals[2:(m-1)]), lamda);start\_point**

**prob=end\_point-start\_point;prob**

**sum(prob)**

**y\_gr=table(cut(y, intervals,**

**labels=intervals[-1], include.lowest=T,**

**ordered\_result=T));y\_gr**

**chisq.test(y\_gr/250,prob) # compare with exponential distribution**

**barplot(y\_gr/250,prob,beside=TRUE, col=c("pink", "lightblue"))**

**#Task6----**

**y=S$price**

**x=S$year**

**cor(x, y) # correlation coefficient**

**lm.out=lm(y~x)**

**lm.out**

**summary(lm.out)**

**par(mfrow = c(1, 1))**

**plot(y~x)**

**abline(lm.out, col="red")**

**lm.out$coefficients%\*%c(1,35)**

**lm.out$coefficients[1]+lm.out$coefficients[2]\*35**

**#Task7----**

**plot(S$area, S$price, main = "Flat Price vs Area", xlab = "Area (m2)", ylab = "Price ($)")**

**# Fit a linear regression model**

**model <- lm(price ~ area, data = S)**

**# Add regression line to the plot**

**abline(model, col = "green")**

**# Summary of the regression model**

**summary(model)**

**# Prediction for a 160 m2 flat**

**new\_data <- data.frame(area = 160)**

**new\_data1 <- data.frame(area = 150)**

**predicted\_price <- predict(model, newdata = new\_data)**

**predicted\_price1 <- predict(model, newdata = new\_data1)**

**predicted\_price**

**predicted\_price1**