Student: Arthur Grossmann—Le Mauguen H00494101

F79MB: Statistical Models B

Assessment Project 1 2025 – Edinburgh

Academic Year: 2024-2025

Tasks

The charges.txt file available on Canvas under Modules > Assessed Project 1, contains data on a sample of 149 individual medical costs (in USD) billed by private hospitals across various regions of the United States for the year 2024.

Question 1 : Present appropriate numerical and graphical summaries, and comment on the distribution of the data. **[2 marks]**

The data represents medical charges in USD charged by private hospitals across different regions of the United States in 2024, as summarized in the statistics and graphical summaries:

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard deviation	Interquartile range (IQR)	Mode
1 137	3 947	9 229	13 412	18 158	51 195	12 225.63	14 210.46	16 884.92

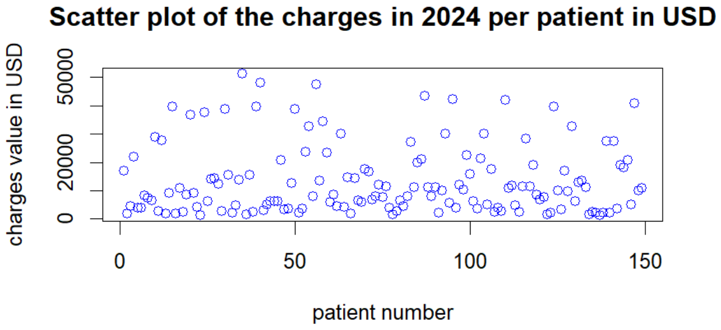


Figure 1: Scatter plot of the medical charges (in USD) of 149 individuals charged by private hospitals in the USA in 2024

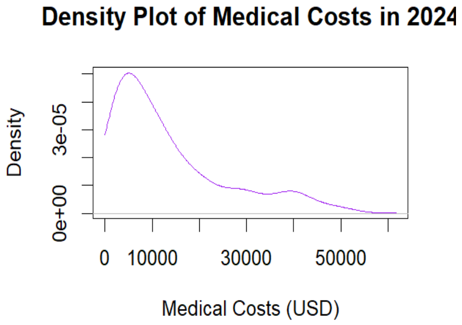


Figure 2: Density plot of the medical charges (in USD) of 149 individuals charged by private hospitals in the USA in 2024

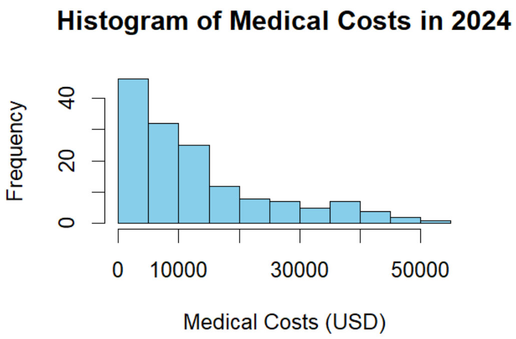


Figure 3: Scatter plot of the medical charges (in USD) of 149 individuals charged by private hospitals in the USA in

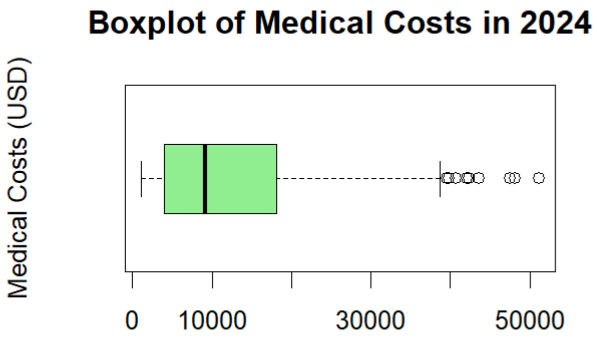


Figure 4: Boxplot of the medical charges (in USD) of 149 individuals charged by private hospitals in the USA in 2024

- The histogram indicates a right-skewed (positive) distribution, where most patients have relatively low medical charges, while a few incurring substantially higher costs. This is confirmed by the numerical summary, as the mean exceeds the median. Additionally, the standard deviation is relatively large, reflecting the wide spread of medical charges.

- The data includes some outliers with charges above $40,000, which contribute to the high standard deviation. The maximum is well above the third quartile at $51,195.
- The scatterplot shows no specific pattern, indicating that charges are broadly distributed across patients without a strong trend or correlation with the patient number.

Question 2: Use QQ plots to explore whether (i) the original data can be adequately modelled by an exponential distribution with parameter λ estimated from MLE, or (ii) the (natural) logarithm of the data can be reasonably modelled by a normal distribution with mean μ = 9 and variance σ2 = 1. **[4 marks]**

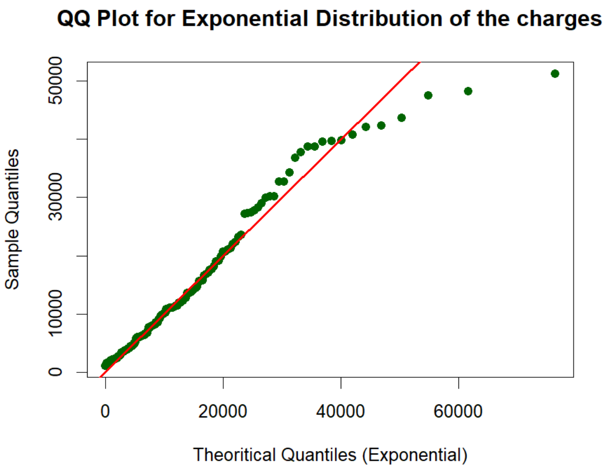
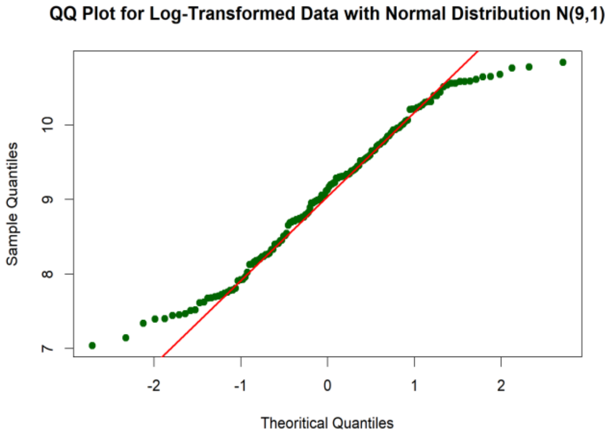


Figure 5: Exponential QQ plot of the charges

Figure 6: Normal QQ plot of the log-transformed charges



I estimated the rate parameter λ using MLE (maximum likelihood), where $\hat{\lambda} = \frac{1}{mean(charges)}$. This provided the required parameter for generating the theoretical quantiles of the exponential distribution.

I created a qqplot to compare the quantiles of the sample data with the theoretical quantiles of the exponential distribution. The red reference line in the plot represented a perfect fit. The data points until 25,000$ follow closely this line, this indicate that this part of the dataset fits the exponential distribution well. However, the data points beyond 25,000$ are very fare from the reference line. We can conclude that this model does not fit well our data especially for the extreme values.

Since data that follows an exponential distribution often becomes approximately normal when log transformed, I applied a natural logarithm to the charges dataset to prepare for a normality check.

Then plotted a QQ plot for the log-transformed data to check for normality. The log-transformed data points are aligned with the reference line, it suggets that the transformed data follows a normal distribution with mean μ = 9 and variance σ^2 = 1. But there seems to be some deviations at the tails, most likely caused by extreme values.

Question 3.A: Perform a chi-squared goodness-of-fit test to formally assess whether (i) the original data can be reasonably modelled using an Exponential distribution, and (ii) the (natural) logarithm of the data can be reasonably modelled by a normal distribution with mean μ = 9 and variance σ2 = 1. You should use 9 cells for your testing procedure, beginning from zero, ensuring minimal arbitrariness in the process. **[6 marks]**

i) A χ^2 goodness-of-fit test is conducted to evaluate the following hypothesis:

- H_0 : the original data can be reasonably modelled using an exponential distribution $(\lambda = \frac{1}{\mathrm{x}})$

- H_1: the original data cannot be modelled using an exponential distribution ($\lambda = \frac{1}{\bar{x}}$)

A χ^2 test with 9 cells and $\lambda = \frac{1}{\bar{x}}$ = 7,46e-5 gives excepted frequencies $e_i = \frac{149}{9} = 16{,}56$, for $i = 1, 2, 3, 4, 5, 6, 7, 8, 9$.

Figure 7: Table of interval of charges and the observed frequencies

Intervals (costs in USD)	Observed frequencies
0 – 5,688	48
5,688 – 11,376	40
11,377 – 17,065	20
17,065 – 22,753	13
22,753 – 28,441	7
28,441 – 34,130	6
34,130 – 39,818	8
39,818 – 45,506	4
45,506 – 51,195	2

The test static value: $\chi^2 = \sum_{i=1}^{10} \frac{(f_i - e_i)^2}{e_i} = 7$

Degree of freedom: d = k – p – 1

d = 9 – 1 – 1

d = 7

Thus, the p-value is: $P(\chi_7^2 > 7.00) = 0.429$ > 0.05

Since this p-value is relatively high, we cannot reject the null hypothesis. However, this does not confirm that the data is perfectly modelled by the exponential distribution either, but it remains a possibility.

ii) A χ^2 goodness-of-fit test is conducted to evaluate the following hypotheses:

- H_0 : the logarithm of the data can be reasonably modelled using a $N(9,1)$ distribution
- H_1 : the logarithm of the data cannot be modelled using a $N(9,1)$ distribution

A χ^2 test with 9 cells, assuming a mean μ = 9 and variance σ² = 1, gives excepted frequencies $e_i = \frac{149}{9} = 16{,}56$ for $i = 1, 2, 3, 4, 5, 6, 7, 8, 9$.

Figure 8: Table of interval of logarithmic charges and the observed frequencies

Intervals (costs in USD)	Observed frequencies
7.0362 – 7.4592	7
7.4592 – 7.8822	15
7.8822 – 8.3052	16
8.3052 – 8.7283	15
8.7283 – 9.1513	22
9.1513 – 9.5743	26
9.5743 – 9.9973	18
9.9973 – 10.4204	15
10.4204 – 10.8434	14

The test static value: $\chi^2 \;= \sum_{i=1}^{10} \frac{(f_i - e_i)^2}{e_i} = 13.96$

Degree of freedom: d = k – p – 1

d = 9 – 2 – 1

d = 6

Thus, the p-value is: $P(\chi_6^2 > 13.96) = 0.03 < 0.05$

Since the p-value is less than 0.05, we reject the null hypothesis. This suggests that the logarithm of the data does not follow a N(9,1) distribution.

Question 3.B: Overall, what would your recommendation regarding of an appropriate model for the observed data? **[2 marks]**

The log-normal distribution was rejected based on the χ^2 test, meaning it does not provide a good fit for the data. On the other hand, the exponential distribution was not rejected, indicating it could be a reasonable model. However, the qqplot suggests a good fit for lower values but noticeable deviations at the extremes, highlighting potential issues with extreme values. To confirm whether the exponential distribution is truly appropriate, additional tests, such as the Kolmogorov-Smirnov test, should be conducted.

Question 4.A: For both (i) and (ii) in question (3), determine whether the Kolmogorov-Smirnov (KS) test can be applied, and justify your reasoning. If the KS test is applicable, compare its results with those from the chi-squared goodness-of-fit test conducted in part (3). **[2 marks]**

A Kolmogorov-Smirnov test is conducted to determine which distribution provides the best fit. This test is particularly useful for small datasets, as it is more precise than the χ^2 test in such cases. However, the Kolmogorov-Smirnov test can only be applied to continuous and independent distributions.

Figure 9: Table of p values using the different model of distributions

Distribution tested	Chi square test (p-value)	Kolmogorov-Smirnov test (p-value)
Charges modelled by an exponential distribution	0.42945	0.1393
Log(charges) modelled by a normal distribution	0.03004	0.3548

The pvalues obtained with the Kolmogorov-Smirnov test do not clearly indicate which distribution is the best fit, as the results are somewhat contradictory. This inconsistency arises because the Kolmogorov-Smirnov test requires that distribution parameters are not estimated from the data. In the case of the exponential distribution, the parameter is estimated using the sample mean, which leads to an invalid and low p-value for this distribution. With the Kolmogorov-Smirnov test, the log-normal distribution is not rejected. This may be due to the test's low sensitivity to the tails of the distribution. QQ plots revealed that the primary issue with the log-normal distribution lies in its tails.

Question 4.B: Assume the sample size is reduced from 149 to 30 data points. Discuss the impact of this reduced sample size on both the KS test and the chi-squared goodness-of-fit test. **[2 marks]**

Reducing the sample size from 149 to 30 significantly affects both the **Kolmogorov-Smirnov (KS) test** and the **chi-squared goodness-of-fit test**, but in different ways.

1) **Kolmogorov-Smirnov (KS) Test**

- The **KS test** measures the difference between the **empirical cumulative distribution function (ECDF)** of the sample and a given **theoretical cumulative distribution function (CDF)**.
- **Reduced sample size increases the variability** of the ECDF, making the test less reliable.
- **Power reduction:** The KS test may fail to detect small differences between the sample and the theoretical distribution, leading to a higher chance of Type II errors (failing to reject a false null hypothesis).
- With a smaller sample size, extreme values become less frequent, potentially mitigating the KS test's weak sensitivity to tails. As a result, the log-normal distribution could be rejected more easily, as the issue with the tails would be less pronounced. Conversely, the exponential distribution might perform even better with fewer data points, as the Kolmogorov-Smirnov test is more reliable for small datasets. A smaller sample size could confirm that the exponential model is a good fit, but further testing would still be necessary.
- **However, the test remains applicable** and does not rely on binning data, unlike the chi-squared test, making it preferable for small samples compared to the chi-squared test.

2) Chi-Squared Goodness-of-Fit Test

- This test compares **observed** and **expected** frequencies in **predefined bins (cells)**.
- The chi-squared test **requires a sufficient number of observations in each cell** to provide reliable results. With only 30 data points, it becomes difficult to ensure that each of the 9 cells has the recommended minimum frequency (usually at least 5 observations per cell).
- **Increased risk of Type I (false positives) and Type II (false negatives) errors:** If expected frequencies are too low, the test statistic may be inaccurate, and the p-value may be inaccurate.
- **Not recommended for very small samples** due to its dependence on cell counts and potential loss of statistical power.

Conclusion

- **KS Test**: Still applicable for small samples but less powerful.
- **Chi-Squared Test**: Less reliable and potentially invalid due to low expected frequencies in each cell.
- For a sample size of 30, it is **better to rely on the KS test** or other alternatives than the chi-squared test.

Question 5: Use a non-parametric bootstrap methodology to obtain the empirical sampling distribution of the sample median, ˆm, and present the distribution graphically with comments. Compute a non-parametric bootstrap 95% confidence interval for the population median medical costs across the United States for the year 2024. **[3 marks]**

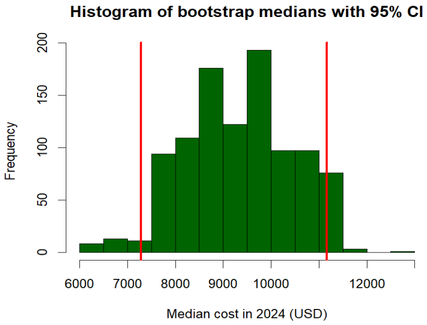
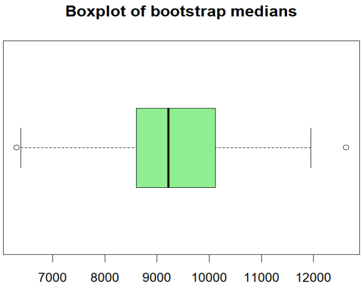


Figure 10: Histogram of the median cost of the 1,000 bootstraps

Figure 11: Boxplot of the median cost of the 1,000 bootstraps



The objective of this analysis is to estimate the population's median medical cost for 2024 in the United States, using a non-parametric bootstrap method. The dataset consists of 149 patients, and from this original sample, 10,000 bootstrap samples are generated. We generate those multiple bootstrap samples by randomly resampling the original data with replacement. This simulates what would happen if we were to collect many new samples from the population. The sample size for each bootstrap sample is equal to the original sample size, and each sample is drawn independently with replacement. The median is calculated for each of these samples, and a 95% confidence interval is derived from the range that includes 95% of the medians, which is from $7,281.50 to $11,082.58. The histogram (Figure 10) and boxplot (Figure 11) reveal a slight right skew in the bootstrap distribution, indicating the presence of higher cost outliers.

Graphical interpretations :

- The histogram of the bootstrap medians shows how the median varies across the bootstrap samples.
- A **symmetric distribution** of the bootstrap medians would suggest that the sample median is approximately unbiased.
- The **95% confidence interval** indicates that we are 95% confident that the true population median lies within this range.

Question 6: Under the assumption that the (natural) logarithm of the medical costs data can be reasonably modelled by a normal distribution with mean μ = 9 and variance σ2 = 1, use asymptotic theory to compute 95% confidence interval for the population median medical costs. Comment on the validity of the two confidence intervals obtained using asymptotic theory and non-parametric bootstrap in part (5). **[4 marks]**

Under the assumption that the (natural) logarithm of the medical costs data can be modelled by a normal distribution N(9,1), I want to determine a 95% confidence interval for the population median medical costs. I used the sample median $\hat{m}$ to estimate the median and asymptomatically, I know that $\hat{m} \sim N\left(m, \left(\frac{1}{2f(m)\sqrt{n}}\right)^2\right)$ with $f(m)$ the density of the probability at the median m.

The standard error for a normal distribution: $se(\hat{m}) = \sigma\sqrt{\frac{\pi}{2n}}$

I can conclude that the confidence interval is: $m \pm Z_{0.025}\ x\ se(\hat{m}) = \mu \pm 1.96\ x\ se(\hat{m})$

Because in a normal distribution the mean equals the median. Since this confidence interval is expressed in the logarithmic scale, applying the exponential function is required to convert it into the original scale and obtain the 95% confidence interval for the population median.

With 95% confidence, the median medical cost of this 149 person dataset is estimated between 6,626$ and 9,909$, the samples median at 9,229$ lls near the upper end of this range, indicating that extreme values observed in the sample may be less common in the overall population.

Nonparametric bootstrap 95% confidence interval for the population median: [7,282$; 11,083$]

Asymptotic theory 95 % confidence interval for the population median: [6,626$; 9,909$]

Using asymptotic theory (range = $3,283), the confidence interval is slightly narrower than the non-parametric bootstrap interval (range = $3,801). The non-parametric bootstrap interval is also slightly higher. The difference between these intervals might be due to the approximation using the normal distribution in the asymptotic theory method. Indeed, the qqplot shows that the normal distribution does not perfectly fit the higher values.

Question 7: Under the assumption that the (original) medical costs data can be reasonably modelled by an Exponential distribution, perform a parametric bootstrap hypothesis test to assess whether there is evidence that the median medical costs is less than $8,500. **[3 marks]**

Using a parametric bootstrap method, under the assumption that the original medial costs data follow the exponential distribution I will test the following:

- H_0 : the median medical cost is equal to or greater than $8,500
- H_1 : the median medical cost is less than $8,500

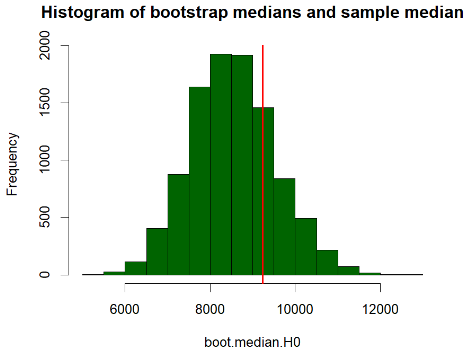


Figure 12: Histogram of the median cost of the 10,000 bootstraps

I generate 10,000 bootstrap samples, and the distribution of bootstrap medians under H_0 is analyzed to determine whether H_0 can be rejected. The obtained p-value is 0.236, meaning 23% the bootstrap medians under H0H_0H0 are less than or equal to the observed sample median. Since the pvalue is greater than 0.05, the null hypothesis cannot be rejected at the 5% significance level. Consequently, no conclusion can be drawn that the median medical cost is significantly less than $8,500.

Question 8: Summarize your overall conclusions. **[4 marks]**

The analysis of medical charges in the United States for 2024 reveals a highly skewed distribution, with most patients incurring relatively low costs while a few faces significantly higher expenses. Statistical tests indicate that an exponential distribution is a reasonable model, though deviations in the tail suggest further investigation is needed. The log-normal distribution was rejected based on the chi-squared test but not by the Kolmogorov-Smirnov test, highlighting inconsistencies in goodness-of-fit assessments. A reduction in sample size would diminish the reliability of the chi-squared test while making the KS test less powerful. Finally, a bootstrap analysis estimates the population median medical cost with a 95% confidence interval ranging from 7,282$ to 11,083$.

[Overall exposition/presentation: 8 marks]

[Project total: 40 marks]

Appendices

Sources:

- Maxime PROD'HOMME's ISEN Nantes statistics and probability 2023-2024 course and R code
- Heriot Watt Edinburgh F79MB Statistical Model B course and R code
- R all in one for dummies 2023 by Joseph Schmuller
- https://en.wikipedia.org/wiki/68–95–99.7_rule
- https://physics.stackexchange.com/questions/107682/kolmogorov-smirnov-test-vs-chi-squared-test
- https://en.wikipedia.org/wiki/Kolmogorov–Smirnov_test

- https://en.wikipedia.org/wiki/Chi-squared_test
- https://www.reddit.com/r/rstats/comments/1875p6e/help_with_chisquare/
- https://stats.libretexts.org/Bookshelves/Applied_Statistics/Biological_Statistics_(McDonald)/02%3A_Tests_for_Nominal_Variables/2.08%3A_Small_Numbers_in_Chi-Square_and_GTests#:~:text=The%20conventional%20rule%20of%20thumb,Fisher%27s%20exact%20test%20of%20independence.
- https://www.bibmath.net/dico/index.php?action=affiche&quoi=c/chideuxtest.html

R Code:

```
getwd() # Print the current working directory


###
#1# Summary statistics and graphical summaries of the data
###
# Read the data from the txt file into R
charges <- scan("charges.txt")
#Numerical summaries
summary(charges) # Summary statistics
sd(charges) # Standard deviation
IQR(charges) # Interquartile range
get_mode <- function(x) { # Function to calculate the mode
  uniq_vals <- unique(x)
  uniq_vals[which.max(tabulate(match(x, uniq_vals)))]}
mode_charges <- get_mode(charges) # Calculate and print the mode
cat("Mode:", mode_charges, "\n")
# Graphical summaries
#Graph
plot(charges,
     main = "Scatter plot of the charges in 2024 per patient in USD",
     xlab = "patient number",
     ylab = "charges value in USD",
     col = "blue")
# Histogram
```

```
hist(charges,
    main = "Histogram of Medical Costs in 2024",
    xlab = "Medical Costs (USD)",
    col = "skyblue")
# Boxplot
boxplot(charges,
      main = "Boxplot of Medical Costs in 2024",
      ylab = "Medical Costs (USD)",
      col = "lightgreen",
      horizontal = TRUE)
# Density plot
plot(density(charges, from = 0), # Density plot of the charges starting from 0
    main = "Density Plot of Medical Costs in 2024",
    xlab = "Medical Costs (USD)",
    col = "purple")


###
#2#
###
#i)
#MLE
lambda_hat <- 1 / mean(charges)
lambda_hat # Print the estimated λ using MLE
# QQ Plot for Exponential distribution
qqplot(qexp(ppoints(length(charges)), rate = lambda_hat), sort(charges), # sort the charges for the qqplot
     main = "QQ Plot for Exponential Distribution of the charges",
     xlab = "Theoritical Quantiles (Exponential)",
     ylab = "Sample Quantiles",
     pch=19, col="darkgreen")
```

```
abline(0,1, col="red", lwd=2) # qqline, reference line
#ii) Log-transform the data and QQ Plot for Normal distribution
log_charges <- log(charges)
# QQ Plot for Normal distribution with mean = 9 and variance = 1
qqnorm(log_charges,
      main = "QQ Plot for Log-Transformed Data with Normal Distribution N(9,1)",
      xlab = "Theoritical Quantiles",
      ylab = "Sample Quantiles",
      pch = 19, col = "darkgreen")
qqline(log_charges, col = "red", lwd = 2) # qqline, reference line


###
#3# Chi squared tests
###
#i)
charges.breaks = seq(0, max(charges), length.out = 10) # Define 9 cells to bin the data into
charges.breaks # Print the 9 cells
charges.cut = cut(charges, breaks = charges.breaks, right = F) # Bin the data into the 9 cells
charges.table <- table(charges.cut) # Count the data in each bin
prob.exp <- numeric(9) # Initialize the vector to store the probabilities
exp.f.exp <- numeric(9) # Initialize the vector to store the expected frequencies
for (i in 1:(length(charges.breaks) - 1)) {
  prob.exp[i] <- pexp(charges.breaks[i+1], rate = lambda_hat) - pexp(charges.breaks[i], rate = lambda_hat)
  exp.f.exp[i] <- prob.exp[i] * length(charges)}
exp.f.exp # Print expected frequencies
obs.f.exp <- as.numeric(charges.table) # Observed frequencies
obs.f.exp # Print observed frequencies
x2.exp <- sum((obs.f.exp - exp.f.exp)^2 / exp.f.exp) # Calculate the chi-squared statistic
x2.exp # Print the chi-squared statistic
```

```
pval.exp <- 1 - pchisq(x2.exp, df = 7) # Calculate the p-value using the chi-squared distribution and the degrees of freedom

pval.exp  # Print the p-value

#ii)

log_charges.breaks = seq(min(log_charges), max(log_charges), length.out = 10)  # Define 9 cells to bin the data into

log_charges.breaks # Print the 9 cells

log_charges.cut = cut(log_charges, breaks = log_charges.breaks, right = F) # Bin the data into the 9 cells

log_charges.table <- table(log_charges.cut) # Count the data in each bin

prob.norm <- numeric(9) # Initialize the vector to store the probabilities

exp.f.norm <- numeric(9) # Initialize the vector to store the expected frequencies

for (i in 1:(length(log_charges.breaks) - 1)) {

  prob.norm[i] <- pnorm(log_charges.breaks[i+1], mean = 9, sd = 1) - pnorm(log_charges.breaks[i], mean = 9, sd = 1)

  exp.f.norm[i] <- prob.norm[i] * length(log_charges)}

exp.f.norm # Print expected frequencies

obs.f.norm <- as.numeric(log_charges.table) # Observed frequencies

obs.f.norm # Print observed frequencies

x2.norm <- sum((obs.f.norm - exp.f.norm)^2 / exp.f.norm) # Calculate the chi-squared statistic

x2.norm  # Print the chi-squared statistic

pval.norm <- 1 - pchisq(x2.norm, df = 6) # Calculate the p-value using the chi-squared distribution and the degrees of freedom

pval.norm  # Print the p-value


###

#4# Kolmogorov-Smirnov tests

###

#i)

sort(charges) # Sort the charges

ks.test(charges, pexp, rate = lambda_hat) # Perform the KS test
```

```
#ii)
sort(log_charges) # Sort the log-transformed charges
ks.test(log_charges, pnorm, mean = 9, sd = 1) # Perform the KS test


###
#5#
###
B <- 1000 # set the number of bootstrap samples
median.bst <- numeric(B) # Initialize the vector to store the medians of the bootstrap samples
for(i in 1:B){ # Loop over the number of bootstrap samples
  boot.sample <- sample(charges, size = 149, replace = TRUE) # Generate a bootstrap sample with replacement and the same size as the original sample
  median.bst[i] = median(boot.sample)} # Store the median of each bootstrap sample
summary(median.bst) # Summary statistics of the bootstrap medians
bootstrap_CI = quantile(median.bst, c(0.025, 0.975)) # 95% bootstrap CI, 2.5% and 97.5% quantiles
bootstrap_CI
par(mfrow=c(1,1)) # Histogram of the bootstrap medians
hist(median.bst,
     main = "Histogram of bootstrap medians with 95% CI",
     xlab = "Median cost in 2024 (USD)",
     col  = "darkgreen")
abline(v = bootstrap_CI, lwd = 3, col = 'red') # Add vertical lines for the 95% CI
boxplot(median.bst, # Boxplot of the bootstrap medians
        main = "Boxplot of bootstrap medians",
        col  = "lightgreen",
        horizontal = TRUE)


###
#6#
###
```

```
#Normal distribution parameters
mu_log <- 9 # mean
sigma2_log <- 1 # variance
sigma_log <- sqrt(sigma2_log)  # standard deviation
median_log <- mu_log # median = mean in a normal distribution
n <- 149 # sample size
#m_hat <- exp(mu_log); # Median of the sample in original scale
#m_hat # Print the median of the sample in the original scale
se <- sigma_log * sqrt(pi / (2 * n)); # Standard error of the sample median
se # Print the standard error of the sample median
#95% CI for the population median on the logarithmic scale
median_log - 1.96 * se; # Lower
median_log + 1.96 * se # Upper
#95% CI for the population median on the original scale
exp(median_log - 1.96 * se); # Lower
exp(median_log + 1.96 * se) # Upper

###
#7#
###
B <- 10000 # Number of bootstraps
median.hyp = 8500 # Hypothesized median 8,500 USD
boot.median.H0 = numeric(B) # Initialize the vector to store the medians of the bootstrap samples
for (i in 1:B){
  y.boot = rexp(length(charges), rate = log(2) / median.hyp)
  boot.median.H0[i] = median(y.boot)} # Store the median of each bootstrap sample
# Histogram of the bootstrap medians and the sample median
hist(boot.median.H0,
    col="darkgreen",
    main="Histogram of bootstrap medians and sample median")
```

```
m.hat = median(charges) # Sample median

abline(v = m.hat, col="red", lwd=2.0) # Add a vertical line for the sample median

boot.pval = (1 + length(boot.median.H0[boot.median.H0 >= m.hat])) / (B + 1)

boot.pval # Print the p-value

t.test(charges, mu=median.hyp, alternative="greater") # One-sample t-test
```