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

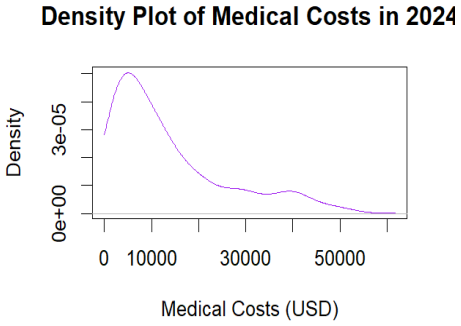


Figure 1: Density plot of the medical charges (in USD) of 149 individuals charged by private hospitals in the USA in 2024

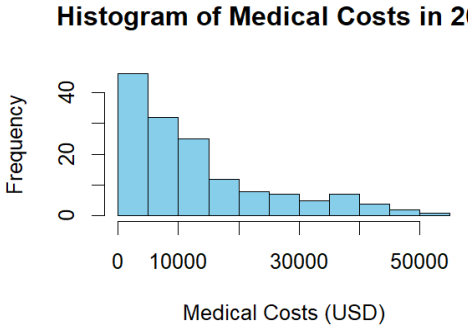


Figure 2: Scatter plot of the medical charges (in USD) of 149 individuals charged by private hospitals in the USA in

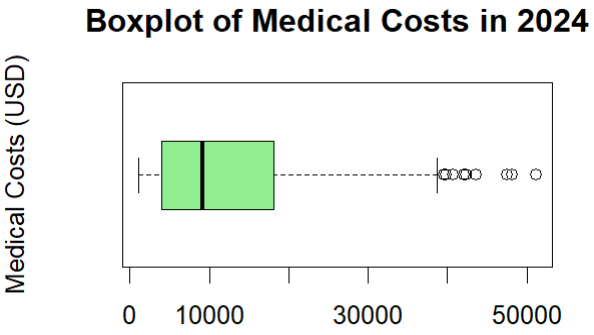


Figure 3: Boxplot of the medical charges (in USD) of 149 individuals charged by private hospitals in the USA in 2024

The histogram (Figure 2) reveals a unimodal and highly right-skewed (positive) distribution, indicating that most patients incur relatively low medical charges, while a few face substantially higher costs. This skewness is supported by the numerical summary, where the mean considerably exceeds the median. The maximum medical cost observed is $51,195, and the minimum is $1,137. The standard deviation, at $12,225, is relatively large, reflecting the wide variability in charges. The presence of several outliers above $40,000, clearly visible in the boxplot (Figure 3), contributes to this high standard deviation and further emphasizes the skewed nature of the data.

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]**

Based on the MLE (maximum likelihood) estimate for the exponential distribution, where

$\hat{\lambda} = \frac{1}{mean(charges)}$ = 7.456 x 10^{-5}, the original data does not appear to closely follow an exponential distribution. While the QQ plot (Figure 4) indicates that most points in the lower tail align well with the line Y=X, several points in the middle deviate slightly above it, and approximately 6 to 7 points in

the upper tail fall below the line. Notably, from around $25,000 onward, the data starts to show increasing deviations from the expected exponential trend, suggesting a heavier tail than predicted by the model.

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. The QQ plot of the log-transformed data (Figure 5) shows a closer fit to a normal distribution with mean $\mu = 9$ and variance $\sigma^2 = 1$. While a few points in the lower tail fall slightly above the line Y=X, and some in the upper tail fall below it, the overall distribution appears reasonably normal. This is further supported by the histogram and boxplot in Figure 6, which show little skewness and no major deviations from normality in the log-transformed data.

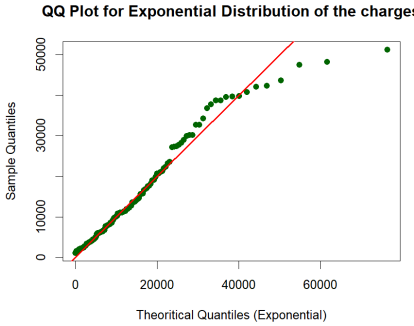


Figure 4: (left) Exponential QQ plot of the charges

Figure 5: (right) Normal QQ plot of the log-transformed charges

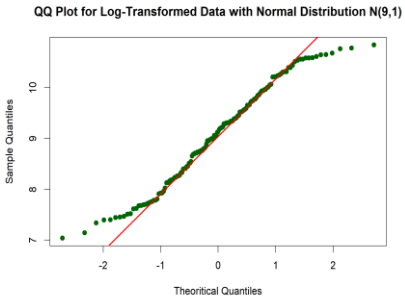
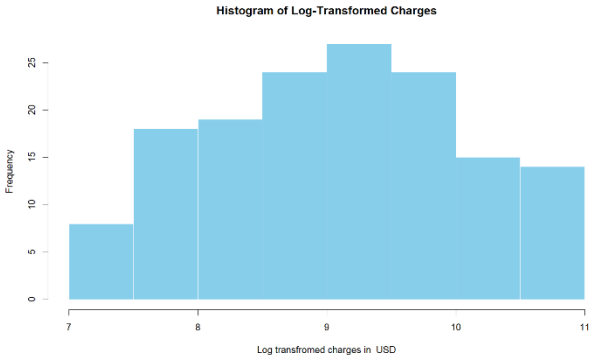
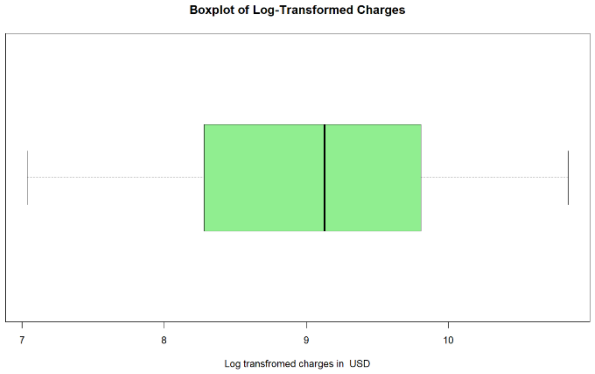


Figure 6: Histogram and boxplot for log transformed data





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 $\mu = 9$ and variance $\sigma 2 = 1$. 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
- H_1: the original data cannot be modelled using an exponential distribution

Let the original data be $x_1, \ldots, x_n$. Since the null hypothesis does not specify the parameter $\hat{\lambda}$ of the exponential distribution, we estimate it from the data using the standard MLE, as computed:

$\lambda = \frac{n}{\Sigma_{i=1}^{n} x_i} = 7.456\ X\ 10^{-5}$. A χ^2 test with 9 equal probability cells results in expected frequencies of $e_i = \frac{149}{9} = 16{,}556$ for $i = 1, 2, 3, 4, 5, 6, 7, 8, 9$. The equal-probability intervals (calculated in R), along with the corresponding observed frequencies, are as follows:

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

Intervals (costs)	Observed frequencies
[0; 1.58 X 10^3)	3
[1.58 X 10^3; 3.37 X 10^3)	24
[3.37 X 10^3; 5.44 X 10^3)	21
[5.44 X 10^3; 7.88 X 10^3)	18
[7.88 X 10^3; 1.09 X 10^4)	16
[1.09 X 10^4; 1.47 X 10^4)	21
[1.47 X 10^4; 2.02 X 10^4)	12
[2.02 X 10^4; 2.95 X 10^4)	14
[2.95 X 10^4; +∞)	20

The test static value: $\chi^2 = \sum_{i=1}^{9} \frac{(f_i - e_i)^2}{e_i} = 19.342$

Degree of freedom: d = k – p – 1

d = 9 – 1 – 1

d = 7

Thus, the p-value is: $P(\chi_7^2 > 19.342) = 0.00718 < 0.05$

There is sufficient evidence to reject H_0. Overall, the original data does not appear to be well modelled by an Exponential distribution. Notably, the QQ plot (Figure 4) indicates some deviations from the Exponential model. While the majority of the data, especially in the lower tail, aligns reasonably well, noticeable discrepancies remain.

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

- H_0 : the log transformed data can be reasonably modelled using a Normal distribution $N(9,1)$
- H_1 : the log transformed data do cannot be modelled using a Normal distribution $N(9,1)$

A χ^2 test with 9 equal probability cells, assuming a mean μ = 9 and variance σ² = 1, gives excepted frequencies $e_i = \frac{149}{9} = 16{,}56$ for $i = 1, 2, 3, 4, 5, 6, 7, 8, 9$. The equal-probability intervals (calculated in R), along with the corresponding observed frequencies, are as follows:

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

Intervals (log transformed costs in USD)	Intervals (costs in USD)	Observed frequencies
(-∞; 7.78)	[0; 2.392 X 10^3)	19
[7.78; 8.24)	[2.392 X 10^3; 3.790 X 10^3)	16
[8.24; 8.57)	[3.790 X 10^3; 5.271 X 10^3)	13
[8.57; 8.86)	[5.271 X 10^3; 7.044 X 10^3)	14
[8.86; 9.14)	[7.044 X 10^3; 9.321 X 10^3)	13
[9.14; 9.43)	[9.321 X 10^3; 1.246 X 10^4)	19
[9.43; 9.76)	[1.246 X 10^4; 1.733 X 10^4)	15
[9.76; 10.2)	[1.733 X 10^4; 2.690 X 10^4)	17
[10.2; +∞)	[2.690 X 10^4; +∞)	23

The test static value: $\chi^2 = \sum_{i=1}^{9} \frac{(f_i - e_i)^2}{e_i} = 5.28859$

Degree of freedom: d = k – p – 1

d = 9 – 1 = 8

Thus, the p-value is: $P(\chi_8^2 > 5.28859) = 0.721918 > 0.05$

Since the p-value is greater than 0.05, we have no evidence to reject the null hypothesis. The log-transformed data seems to be well approximated by a Normal distribution with a mean of 9 and a variance of 1. Nevertheless, the QQ plot (Figure 5) shows that a few observations in the lower and upper extremes deviate from the expected linear trend.

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

The Chi-squared goodness-of-fit test indicates that the exponential distribution does not provide an appropriate model for the medical cost data. This finding is reinforced by the QQ plot in Figure 4, which shows significant departures from the expected exponential pattern, especially in the upper range of the data.

In comparison, the log-transformed data aligns more closely with a normal distribution characterized by a mean of 9 and a variance of 1. Evidence from the QQ plot, histogram, and boxplot (Figures 5 and 6), along with the Chi-squared test, supports this conclusion. Nevertheless, some minor discrepancies are observed in the tails of the QQ plot, and the histogram does not completely conform to the ideal shape of a normal distribution.

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]**

If the parameters of the hypothesized distribution are known in advance, the Kolmogorov–Smirnov (KS) test can be applied. This allows for a valid comparison between the empirical distribution function (EDF) of the sample data and the specified cumulative distribution function (CDF).

In part (i), the parameter λ of the exponential distribution is not specified, so the KS test cannot be applied. However, in part (ii), the parameters of the normal distribution are given as μ = 9 and σ = 1, making the KS test appropriate in this case.

When the test is conducted in R, the test statistic is calculated as D = 0.076052, with a p-value of 0.3548. This suggests that the log-transformed data reasonably fits the proposed Normal(9, 1) distribution. This result aligns with the findings of the chi-squared test from part (3).

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 observations significantly affects both the KS test and the chi-squared goodness-of-fit test.

The KS test remains applicable with smaller samples, as it does not rely on binning. It directly compares the EDF of the sample with the CDF of the target distribution. However, its statistical power decreases with smaller samples, making it less sensitive to subtle deviations from the hypothesized distribution.

In contrast, the chi-squared test is more sensitive to reductions in sample size. It requires adequate expected counts (typically at least 5) in each bin to yield valid results. With only 30 data points, meeting this condition becomes difficult unless the number of bins is reduced. Fewer bins reduce the

test's resolution and may obscure meaningful differences between the observed and expected distributions. In some cases, the test may no longer be applicable.

Overall, the KS test is generally more robust for small sample sizes, as it avoids the loss of information caused by binning. The chi-squared test, on the other hand, is more reliable for larger datasets with well-structured binning schemes.

Question 5: Use a non-parametric bootstrap methodology to obtain the empirical sampling distribution of the sample median, $\hat{m}$, 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]**

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 $7,281.51 to $11,163.57.

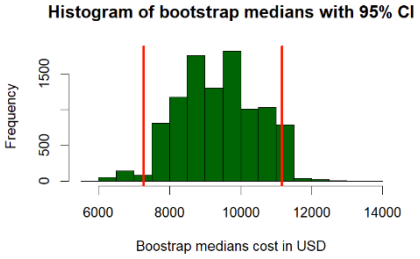
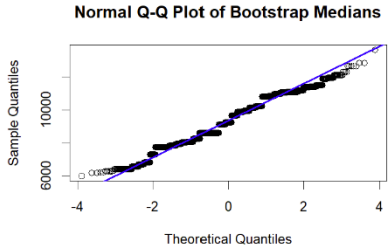


Figure 9: Histogram of the median cost of the 10,000 bootstraps

Figure 10: Normal QQ plot of boostrap medians



Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
5,990	8,601	9,229	9,322	10,115	13,770

Figure 9 and 10 displays the empirical (bootstrap) sampling distribution of the sample median, $\hat{m}$. Both the histogram and the QQ plot suggest a slight positive skew in the distribution, which aligns with the observation that the mean is slightly greater than the median, indicating the presence of higher cost outliers.

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]**

Assuming that the natural logarithm of medical costs follows a normal distribution with mean μ = 9 and variance $\sigma^2 = 1$, we can apply asymptotic theory to estimate a 95% confidence interval (CI) for the population median. According to asymptotic results, the sample median $\hat{m}$ is approximately normally distributed:

$$\hat{m} \sim N\left(m, \left(\frac{1}{2f(m)\sqrt{n}}\right)^2\right)$$

where $f(\hat{m})$ is the value of the probability density function of $N(9,1)$ evaluated at the sample median and $\hat{m}$ is the log transformed sample median.

Given:

- Sample median (log-transformed): $\hat{m} = 9.130089$
- Sample size: n = 149
- $f(\hat{m})$ = 0.3955808 (from R)

The standard error of the sample median is estimated as: $se(\hat{m}) = \frac{1}{2f(\hat{m})\sqrt{n}} = 0.103548.$

Using this, the 95% CI for the log-median is: $\hat{m} \pm 1.959964 \times se(\hat{m}) = (8.927, 9.333)$

Exponentiating the endpoints gives the 95% CI for the median of the original (unlogged) medical costs: ((exp(8.927), exp(9.333)) = (7533.681, 11305.455).

Both the asymptotic and non-parametric bootstrap confidence intervals are quite close, despite small differences (Bootstrap CI: [7,282 $; 11,083 $] and Asymptotic CI: [7,533.68 $; 11,305.46 $]). The similarity is due to the fact that the sampling distribution of the bootstrap medians (Figure 9) is only mildly right-skewed and approximately normal. Given the relatively large sample size (n = 149), the assumptions of asymptotic theory hold reasonably well. Moreover, the histogram and boxplot of the log-transformed data (Figure 6) do not indicate significant deviations from normality.

While both methods yield valid results, the CI based on asymptotic theory is narrower (range ≈ 3,772 $ vs. 3,801 $) and computationally more efficient. Therefore, in this context, it is a preferred method due to its simplicity and reliability.

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

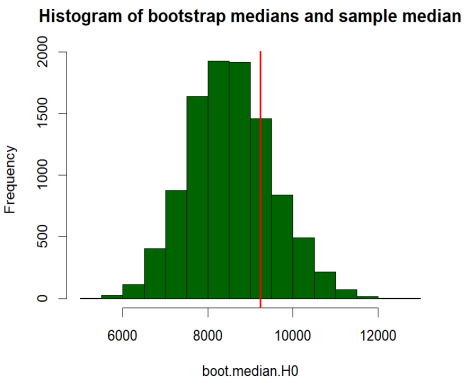


Figure 11: Sampling distribution of bootstrap medians of sample from Exponential (8.15467×10−5) distribution, with red line showing sample median.

The cumulative distribution function (CDF) of an exponential distribution with rate parameter λ is given by $F_x(x) = 1-e^{-\lambda x}$. Assuming the null hypothesis H_0, which posits a population median of $8,500, we determine the value of λ using the formula $\lambda = \frac{\log(2)}{8{,}500} = 8.15467 \times 10^{-5}$

Under the H_0 hypothesis, Figure 11 displays the sampling distribution of the sample median, assuming the data follow an exponential distribution with. The observed sample median is $9,228.847. The resulting p-value is 0.7609, indicating that there is no significant evidence to reject H_0. Therefore, we cannot conclude that the true median medical cost is less than $8,500.

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

To sum up, in 2024, private hospital medical costs (in USD) across different U.S. regions show a strongly right-skewed distribution, based on a sample of 149 observations. The mean cost is $13,412, the median is $9,229, with values ranging from $1,137 to $51,195, and a standard deviation of $12,225.63.

Log-transforming the data gives a distribution that fits well with a normal distribution $N(9, 1)$, as supported by QQ plots, chi-squared goodness-of-fit tests, and the KS test. Note: the KS test is valid in part 3(ii), where parameters are specified (normal case), but not in part 3(i) for the exponential case with unspecified parameters.

When constructing 95% confidence intervals for the population median using both non-parametric and asymptotic methods, the results are fairly close. The asymptotic interval is preferred for its simplicity, stronger theoretical basis, and narrower range, especially given the sufficiently large sample size (n = 149).

Finally, a parametric bootstrap hypothesis test assuming an exponential distribution shows no evidence that the population median of medical costs is below $8,500.

[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
```

```
# 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
# Histogram
hist(log_charges,
    breaks = 8,
    main = "Histogram of Log-Transformed Charges",
    xlab = "Log transfromed charges in  USD",
    ylab = "Frequency",
    col = "skyblue",
    border = "white")
# Boxplot
boxplot(log_charges,
       main = "Boxplot of Log-Transformed Charges",
       col = "lightgreen",
       xlab = "Log transfromed charges in  USD",
       horizontal = TRUE)

###
#3# Chi squared tests
###
#i)
```

```
# Given parameters
lambda_hat <- 7.456e-5
n <- length(charges)
# Generate cumulative probabilities for 9 equal-probability intervals
probs <- seq(0, 1, length.out = 10)
# Calculate the quantile breakpoints for an exponential distribution
charges.breaks <- qexp(probs, rate = lambda_hat)
# Ensure strictly increasing breaks (by removing duplicates if necessary)
charges.breaks <- unique(charges.breaks)
# Replace the last value with Inf to capture the final interval properly
charges.breaks[length(charges.breaks)] <- Inf
# Check if we still have 10 breakpoints (i.e., 9 intervals)
if (length(charges.breaks) != 10) {
  stop("Breakpoints are not unique enough to create 9 intervals. Try adjusting lambda_hat or check the data distribution.")
}
# Now bin the data
charges.cut <- cut(charges, breaks = charges.breaks, right = FALSE, include.lowest = TRUE)
obs.f.exp <- as.numeric(table(charges.cut))  # Observed frequencies
# Expected frequency per bin under H0
exp.f.exp <- rep(n / 9, 9)
# Chi-squared test statistic
x2.exp <- sum((obs.f.exp - exp.f.exp)^2 / exp.f.exp)
# Degrees of freedom: 9 intervals - 1 constraint - 1 estimated parameter
df <- 9 - 1 - 1
# p-value
pval.exp <- 1 - pchisq(x2.exp, df)
# Output results
list(
  "Observed Frequencies" = obs.f.exp,
```

```
  "Expected Frequencies" = exp.f.exp,
  "Chi-squared Statistic" = x2.exp,
  "Degrees of Freedom" = df,
  "p-value" = pval.exp
)
#ii)
# Log-transform the charges
log_charges <- log(charges)
# Parameters under H0
mu <- 9
sigma <- 1
n <- length(log_charges)
# Equal-probability quantiles for 9 bins
probs <- seq(0, 1, length.out = 10)
log_breaks <- qnorm(probs, mean = mu, sd = sigma)
# Ensure the last interval goes to infinity
log_breaks[1] <- -Inf
log_breaks[10] <- Inf
# Bin the log-transformed data
log_charges_cut <- cut(log_charges, breaks = log_breaks, right = FALSE, include.lowest = TRUE)
obs_freq <- as.numeric(table(log_charges_cut))  # Observed frequencies
# Expected frequencies: equal under H0
exp_freq <- rep(n / 9, 9)
# Chi-squared statistic
x2_stat <- sum((obs_freq - exp_freq)^2 / exp_freq)
# Degrees of freedom: 9 intervals - 1 constraint (no parameters estimated here)
df <- 8
# p-value
p_val <- 1 - pchisq(x2_stat, df)
# Output the results
```

```
list(
  "Observed Frequencies" = obs_freq,
  "Expected Frequencies" = exp_freq,
  "Chi-squared Statistic" = x2_stat,
  "Degrees of Freedom" = df,
  "p-value" = p_val
)

###
#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 <- 10000 # 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
```

```
# Histogram of the bootstrap medians
hist(median.bst,
    main = "Histogram of bootstrap medians with 95% CI",
    xlab = "Boostrap medians cost in USD",
    col  = "darkgreen")
abline(v = bootstrap_CI, lwd = 3, col = 'red') # Add vertical lines for the 95% CI
# QQ plot
qqnorm(median.bst,
      main = "Normal Q-Q Plot of Bootstrap Medians")
qqline(median.bst, col = "blue", lwd = 2)

###
#6#
###
# Sample median (on log scale) and sample size
m_hat <- 9.130089
n <- 149
# Density of N(9,1) at m_hat
f_mhat <- dnorm(m_hat, mean = 9, sd = 1) # gives 0.3955808
f_mhat
# Standard error based on asymptotic theory
se_mhat <- 1 / (2 * f_mhat * sqrt(n)) # gives 0.103548
se_mhat
# 95% CI on log scale
z <- qnorm(0.975) # 1.959964
lower_log <- m_hat - z * se_mhat
upper_log <- m_hat + z * se_mhat
c(lower_log, upper_log) # (8.927, 9.333)
# 95% CI on original scale
lower_original <- exp(lower_log)
```

```
upper_original <- exp(upper_log)
c(lower_original, upper_original) # (7533.678, 11305.440)

###
#7#
###
B <- 10000 # Number of bootstraps
median.hyp <- 8500 # Hypothesized median 8,500 USD
lambda.hyp <- log(2) / median.hyp # Hypothesized population median under H0
boot.median.H0 = numeric(B) # Initialize the vector to store the medians of the bootstrap samples
# Generate bootstrap medians under the null hypothesis
for (i in 1:B){
  y.boot = rexp(length(charges), rate = lambda.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 + sum(boot.median.H0 <= m.hat)) / (B + 1)
boot.pval
```