Comparing the distributions fitted to time to kidney infection after catheter replacement using different distribution

ETIKUDO JOEL EDIDION

Contents

List of Figures

Abstract

This study investigates the suitability of four probability distributions Exponential, Half-Normal, Weibull and Gamma in fitting a given dataset. Each distribution is assessed through histogram and probability plot analysis, accompanied by rigorous goodness-of-fit tests, including Pearson Chi-square, Anderson-Darling, and Kolmogorov-Smirnov tests. Results indicate that while the Exponential, Weibull, and Gamma distributions exhibit favorable fits overall, with notable exceptions in the Pearson Chi-square test for the Exponential and Weibull distributions, the Half-Normal distribution emerges as the optimal fit. Goodness of fit tests consistently support the suitability of the Half-Normal distribution for representing the dataset, thus affirming its selection as the most appropriate model. This study underscores the importance of comprehensive model evaluation techniques in identifying the best-fitting probability distribution for empirical data analysis.

CHAPTER 1

Introduction

1.1 Background Of the Study

Catheter replacement procedures are common medical interventions performed to manage various urinary conditions, including urinary retention, urinary incontinence, and bladder dysfunction. Urinary catheters are flexible tubes inserted into the bladder through the urethra or a surgical opening in the abdomen to drain urine from the bladder. Catheter replacement involves the removal of an existing catheter and the insertion of a new one, typically to prevent complications such as catheter-associated urinary tract infections (CAUTIs) or to address issues with catheter function. Importance in Healthcare include: Urinary Management: Catheter replacement procedures play a crucial role in managing urinary dysfunction in patients who are unable to urinate independently due to medical conditions such as spinal cord injury, neurological disorders, or post-surgical recovery. Prevention of Complications: Regular catheter replacement is essential for preventing complications such as catheter blockage, leakage, or infection. Catheters that are left in place for extended periods can become colonized with bacteria, leading to CAUTIs, which can result in serious complications such as kidney infections or bloodstream infections. Patient Comfort and Quality of Life: Timely catheter replacement helps maintain patient comfort and improves the quality of life for individuals requiring long-term catheterization. Proper catheter care and maintenance, including regular replacement, reduce the risk of discomfort, pain, and complications associated with urinary catheter use. Infection Control: Catheter replacement procedures are a critical component of infection control protocols in healthcare settings. By adhering to recommended catheter replacement schedules and sterile insertion techniques, healthcare providers can minimize the risk of catheter-related infections and promote patient safety. Optimizing Healthcare Resource Utilization: Efficient catheter replacement practices contribute to the effective utilization of healthcare resources by reducing the incidence of catheter-related complications that may require additional medical interventions, hospitalizations, or antibiotic treatments.

1.1.1 Risk of kidney infections associated with catheters

Urinary catheters are essential medical devices used to manage urinary dysfunction in patients who are unable to urinate independently. However, prolonged catheter use increases the risk of urinary tract infections (UTIs), including kidney infections, which can have serious consequences. Here are some key points highlighting the risk of kidney infections associated with catheters: Catheter-Associated Urinary Tract Infections (CAUTIs): Urinary catheters provide a direct pathway for bacteria to enter the urinary tract, leading to catheter-associated urinary tract infections (CAUTIs). CAUTIs are one of the most common healthcare-associated infections, and they account for a significant proportion of all hospital-acquired infections. Bacterial Colonization: Catheters left in place for extended periods can become colonized with bacteria, which form biofilms on the catheter surface. These biofilms provide a reservoir for bacteria to multiply and thrive, increasing the risk of infection. Ascending Infections: Bacteria can ascend from the urethra or around the catheter into the bladder, causing bladder infections (cystitis). If left untreated or inadequately managed, these bladder infections can progress to involve the kidneys, leading to pyelonephritis, or kidney infection. Serious Complications: Kidney infections, or pyelonephritis, are serious bacterial infections that affect the kidneys. They can cause symptoms such as fever, chills, flank pain, nausea, vomiting, and malaise. If not promptly diagnosed and treated with appropriate antibiotics, kidney infections can lead to sepsis, kidney damage, or even life-threatening complications. Increased Healthcare Costs: Catheter-associated kidney infections contribute to increased healthcare costs due to prolonged hospitalizations, additional diagnostic tests, antibiotic treatments, and potential complications requiring intensive care management. Impact on Patient Well-being: Kidney infections associated with catheters can significantly impact the well-being and quality of life of affected individuals. Patients may experience pain, discomfort, and distressing symptoms, leading to decreased mobility, functional impairment, and prolonged recovery periods. Preventive Measures: Preventive measures to reduce the risk of kidney infections associated with catheters include proper catheter care and maintenance, adherence to sterile insertion techniques, regular catheter replacement, and minimizing catheter use whenever possible.

1.1.2 Significance of predicting infection time for early intervent

Predicting the time to infection following catheter replacement is crucial for enabling early intervention and improving patient outcomes. Here's why it's significant:

Early Detection and Treatment: Predicting the time to infection allows healthcare providers to monitor patients closely and detect signs of infection at an early stage. Early detection enables prompt initiation of appropriate treatment, such as antibiotic therapy, which can help prevent the infection from worsening and spreading to other parts of the urinary tract, including the kidneys. Reduced Morbidity and Complications: Early intervention can help reduce the morbidity associated with catheter-associated infections. By treating infections promptly, healthcare providers can minimize the severity of symptoms, prevent complications such as kidney damage or sepsis, and improve patient outcomes. Prevention of Catheter-Related Complications: Catheter-associated infections, including kidney infections, are a

significant cause of morbidity and mortality in healthcare settings. Predicting the time to infection allows healthcare providers to implement preventive measures, such as timely catheter replacement or removal, to reduce the risk of infection and associated complications. Optimization of Antibiotic Use: Early prediction of infection allows for targeted antibiotic therapy, optimizing antibiotic use and reducing the risk of antibiotic resistance. By tailoring antibiotic treatment to the specific pathogen causing the infection, healthcare providers can improve treatment efficacy and minimize the development of antibiotic-resistant bacteria. Improved Resource Allocation: Predicting the time to infection helps healthcare facilities allocate resources more efficiently. By identifying patients at higher risk of infection, healthcare providers can prioritize monitoring and intervention efforts, ensuring that resources such as diagnostic tests, antibiotics, and healthcare personnel are utilized effectively.

1.2 Statement of problem

Exponential distribution, gamma distribution, and Weibull distribution do not fit well to the time to kidney infection after catheter replacement data.

1.3 Aim and Objectives

1.3.1 Aim

To compare the effectiveness of different probability distributions in modeling time to kidney infection after catheter replacement

1.3.2 Objectives

- Fit exponential, Weibull, half-normal, and gamma distributions to the data
- Evaluate the goodness-of-fit for the best distribution
- Identify the distribution that best predicts infection time

1.4 Significance of the Study

This study will provide valuable insights into choosing the most accurate distribution for predicting infection time, ultimately improving patient care through earlier detection and intervention. The significance of the study lies in its potential to contribute valuable insights into the distribution of time to kidney infection following catheter replacement. Here are some key aspects highlighting the significance of the study:

- Clinical Relevance: Understanding the distribution of time to kidney infection is crucial for clinicians and healthcare providers involved in the management of patients with urinary catheters. By characterizing the distribution of infection onset, clinicians can better anticipate and monitor patients at risk, leading to improved clinical decision-making and patient care.

- Patient Outcomes: The study has the potential to directly impact patient outcomes by providing clinicians with valuable information to guide early intervention and treatment strategies. Early detection and management of kidney infections can prevent complications, reduce morbidity, and improve overall patient outcomes.

- Healthcare Resource Allocation: Insights from the study can inform healthcare resource allocation and management strategies. By identifying high-risk patient populations or specific timeframes associated with increased infection risk, healthcare facilities can allocate resources more effectively, prioritize interventions, and optimize patient care delivery.

- Preventive Measures: Knowledge of the distribution of time to kidney infection can guide the development and implementation of preventive measures aimed at reducing catheter-associated infections. This may include strategies such as optimizing catheter maintenance protocols, enhancing infection prevention practices, and promoting antimicrobial stewardship.

- Research and Innovation: The study contributes to the existing body of scientific knowledge in the field of catheter-associated infections and adds to the evidence base for future research and innovation. Findings from the study may stimulate further investigations into risk factors, pathogenesis, etc.

1.5 Scope of the Study

- Scope of Data Collection: The project will focus specifically on collecting data related to the time to kidney infection following catheter replacement. Data collection will be limited to patients who have undergone catheter replacement procedures in a defined healthcare setting or cohort.

- Types of Catheter Replacement: The project will specify the types of catheter replacement procedures included in the study. This may include both indwelling urinary catheters and other types of catheters commonly used in clinical practice.

- Infection Definition: Clear criteria will be established for defining kidney infections in the context of the project. This may include specific diagnostic criteria based on clinical signs, symptoms, laboratory findings, or microbiological culture results.

- Selection of Distributions: The project will focus on comparing the fit of four specific probability distributions to the time to kidney infection data: exponential distribution, Weibull distribution, half-normal distribution, and gamma distribution. Other distributions will not be considered within the scope of this project.

- Statistical Analysis Methods: The project will delineate the statistical methods used for fitting the distributions to the data and evaluating their goodness-of-fit. This may include maximum likelihood estimation, method of moments, or other appropriate techniques.

- Interpretation of Results: The project will specify the criteria for interpreting the results of the distribution fitting process. This may involve assessing goodness-of-fit statistics, visual inspection of probability plots, or comparison of model parameters.
- Limitations: Clear acknowledgment of the limitations of the study will be provided, including potential biases in the data, assumptions of the selected distributions, and generalizability of the findings to broader populations or clinical settings.

1.6 Literature review

1.6.1 Relevant research on time to kidney infection after catheter replacement

Research on the time to kidney infection after catheter replacement is crucial for understanding the risk factors, predictors, and potential interventions for this medical complication. Here are some relevant studies in this area: "Catheter-Associated Urinary Tract Infections: Epidemiology, Pathogenesis, and Prevention" by Saint S, Chenoweth CE. This review article provides an overview of catheter-associated urinary tract infections (CAUTIs), including risk factors, pathogenesis, and preventive strategies. It discusses the importance of catheter management in reducing the risk of infections, including kidney infections. "Risk Factors for Catheter-Associated Urinary Tract Infection in a Pediatric Institution" by Moore KN, Fader M, Getliffe K. This study explores the risk factors associated with catheter-associated urinary tract infections in pediatric patients. While the focus is on urinary tract infections, findings related to catheter duration and other factors may be relevant to kidney infections Safdar et al. (2002). "Preventing Catheter-Associated Urinary Tract Infections in Acute Care: The Bundle Approach" by Meddings J, Rogers MAM, Krein SL, et al. This research evaluates the effectiveness of a bundle approach in preventing catheter-associated urinary tract infections in acute care settings. While not specific to kidney infections, it provides insights into strategies for reducing overall catheter-related complications. "Urinary Catheter Use and the Risk of Mortality in Hospitalized Patients" . This retrospective cohort study investigates the association between urinary catheter use and mortality in hospitalized patients. Understanding the risks associated with catheterization can inform decisions about catheter replacement and management to prevent complications such as kidney infections Chen et al. (2024). "Long-term Catheterization of the Bladder: Risks and Benefits" by Saint S, Lipsky BA, Goold SD. This review article discusses the risks and benefits of long-term bladder catheterization, including the potential for urinary tract infections and other complications. While not specific to kidney infections, it highlights the importance of careful catheter management to minimize adverse outcomes.

1.6.2 Existing methods for predicting infection time

Predicting Time to Catheter-Related Bloodstream Infections Using Survival Analysis Methods. This study focuses on utilizing survival analysis techniques to predict the time to catheter-related bloodstream infections (CRBSIs). The researchers employed Kaplan-Meier curves and Cox proportional hazards regression to analyze patient data and identify risk factors associated with CRBSIs, aiding in the development

of predictive models Bond et al. (2020). Time-to-Event Analysis of Catheter-Associated Infections in Intensive Care Patients. This retrospective cohort study employs time-to-event analysis techniques to predict the occurrence of catheter-associated infections in intensive care unit (ICU) patients. By analyzing a large dataset of ICU admissions, the researchers identify temporal patterns and risk factors associated with catheter-related infections, facilitating early detection and intervention strategies Bae et al. (2022).

1.6.3 Studies that compare different probability distributions for similar applications

Comparison of Probability Distributions for Modeling Wind Speed Data. This study compares various probability distributions, including the Weibull, Gamma, Lognormal, and Exponential distributions, for modeling wind speed data. The researchers evaluate the goodness-of-fit of each distribution using statistical tests and assess their performance in wind energy applications Wang et al. (2018) . A Comparative Study of Probability Distributions for Flood Frequency Analysis. This research compares different probability distributions, such as the Generalized Extreme Value (GEV), Log-Pearson Type III, and Generalized Pareto distributions, for flood frequency analysis. By analyzing historical flood data, the study assesses the suitability of each distribution for estimating flood probabilities and designing flood control measures Totaro et al. (2024). Comparative Analysis of Probability Distributions for Modeling Rainfall Data. In this studyl, various probability distributions, including the Gamma, Lognormal, and Weibull distributions, are compared for modeling rainfall data. The researchers utilize statistical techniques to evaluate distributional characteristics and select the most appropriate distribution for hydrological applications Kumar et al. (2021). Evaluation of Probability Distributions for Modeling Stock Returns . This research compares different probability distributions, such as the Normal, Student's t, and Generalized Hyperbolic distributions, for modeling stock returns. By analyzing historical stock market data, the study investigates the goodness-of-fit and tail behavior of each distribution to improve risk assessment in financial markets Chen et al. (2019). Comparative Study of Probability Distributions for Modeling Health Data. This study compares several probability distributions, including the Exponential, Weibull, and Lognormal distributions, for modeling health-related data, such as disease onset times or patient survival times. The researchers employ survival analysis techniques to evaluate the performance of each distribution in healthcare applications Goligher et al. (2018).

CHAPTER 2

Data collection and methodology

2.1 DATA DESCRIPTION

The data used in this work is the time to kidney infection after catheter replacement gotten from.

2.2 DATA PRESENTATION

Presenting the data allows for the assessment of data quality, including completeness, accuracy, and consistency. Researchers can identify any missing or erroneous values, assess data distributional properties, and address data pre-processing tasks as needed to ensure the reliability of the analysis results. Below is the data.

2.3 METHODOLOGY

2.3.1 SELECTION OF CANDIDATE DISTRIBUTION

The candidate distributions below Will be fitted with the data to see which distribution is best fit for it.

- Exponential distribution
- Weibull distribution
- Half-normal distribution
- Gamma distribution
- Normal distribution

Table 2.1: The time to kidney infection after catheter replacement.

1.5	3.5	4.5	4.5	5.5	8.5	8.5	9.5
10.5	11.5	15.5	16.5	18.5	23.5	26.5	2.5
2.5	3.5	3.5	3.5	3.5	4.5	5.5	6.5
6.5	7.5	7.5	7.5	7.5	8.5	9.5	10.5
11.5	12.5	12.5	13.5	14.5	14.5	21.5	21.5
22.5	22.5	25.5	27.7	0.5	0.5	0.5	0.5
0.5	2.5	2.5	3.5	6.5	15.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	1.5
1.5	1.5	1.5	1.5	2.5	2.5	2.5	2.5
2.5	3.5	3.5	3.5	3.5	3.5	4.5	4.5
4.5	5.5	5.5	5.5	5.5	5.5	6.5	7.5
7.5	7.5	8.5	8.5	8.5	9.5	9.5	10.5
10.5	10.5	11.5	11.5	12.5	12.5	12.5	12.5
14.5	14.5	16.5	16.5	18.5	19.5	19.5	19.5
20.5	22.5	24.5	25.5	26.5	26.5	28.5	

2.3.2 FITTING DISTRIBUTIONS TO THE DATA

Each of these candidate distributions above will be fitted to the dataset using the method âDistributionFitTest[]â which is available in Mathematica, a statistical package. The parameter of each distribution that best describes the data will then be estimated.

2.3.3 ASSESSING THE GOODNESS OF FIT

The goodness of fit tests will be performed to evaluate how well each fitted distribution fits the data. Common tests such as the Kolmogorov-Smirnov test, Pearson Chi-square test, Anderson Darling test, Cramer-VonMises test, and $WatsonU^2$ test will be employed. Graphical methods such as the Q-Q plot and histograms will be observed as well. The p-value and test statistic from each test will be compared to determine the adequacy of fit for each distribution. The distribution with the best fit to the data based on the goodness of fit results will be chosen.

CHAPTER 3

DATA ANALYSIS AND DISCUSSION OF RESULTS

3.1 DATA ANALYSIS

3.2 SELECTION OF CANDIDATE DISTRIBUTIONS

In this context, we will make use of the following distributions for the comparison; Exponential distribution with the probability density function (PDF) is given below as:

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

Where x = the time to event. (In this text, time to kidney infection after catheter replacement). Lambda = the rate parameter, the rate at which an event occur per unit time. The half-normal distribution with probability density function stated below as :

$$f(x|\sigma) = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma}e^{-\frac{x^2}{2\sigma^2}}$$

Where x is the random variable Sigma is the scale parameter representing the standard deviation of the corresponding normal distribution

The normal distribution with probability density function given below as

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where: X is the random variable Miu is the mean of the distribution Sigma is the standard deviation of the distribution

The Weibull distribution with probability density function written below as:

$$f(x|k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$$

Where: X is the random variable K is the shape parameter controlling the shape of the distribution Lambda is the scale parameter controlling the scale of the distribution.

The gamma distribution with probability density function given below as

$$f(x|k,\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

Where: X is the random variable K is the shape parameter and must be a positive real number Theta is the scale parameter and must be a positive real number Gamma(k) denotes the gamma function defined as the integral from 0 to infinity of

$$\int_0^t t^{k-1} e^{-t} dt$$

In the next section, each of these distributions will be fitted to the data set after which a goodness of fit test is conducted.

3.3 FITTING DISTRIBUTIONS TO THE DATA AND OBTAINING THE BEST FIT

Case 1: Exponential distribution **Test of hypothesis(Goodness of fit):** Anderson-Darling Test H0: The data follows exponential distribution H1: The data does not follow exponential distribution Level of significance: 0.05 Test statistic:

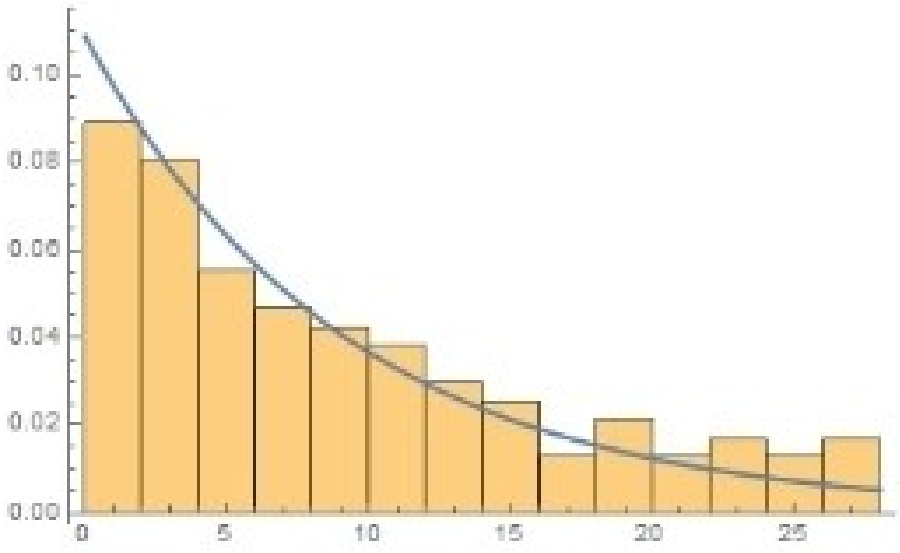
$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} [(2i-1) \cdot (\ln(F(X_i)) + \ln(1 - F(X_{n+1-i})))]$$

Where: (n) is the sample size $F(x_i)$ is the empirical cumulative distribution function (ECDF) evaluated at the ith ordered data point (x_i) (x_i) are the ordered data point Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
1.17	0.27892

Table 3.1:

Conclusion: Since P-value = 0.27892 is greater than level of significance (0.05), we do not have enough evidence to reject the null hypothesis. Therefore the exponential distribution is fit for the data.



0.89

Figure 3.1: The histogram of exponential distribution fitted to the data

0.89

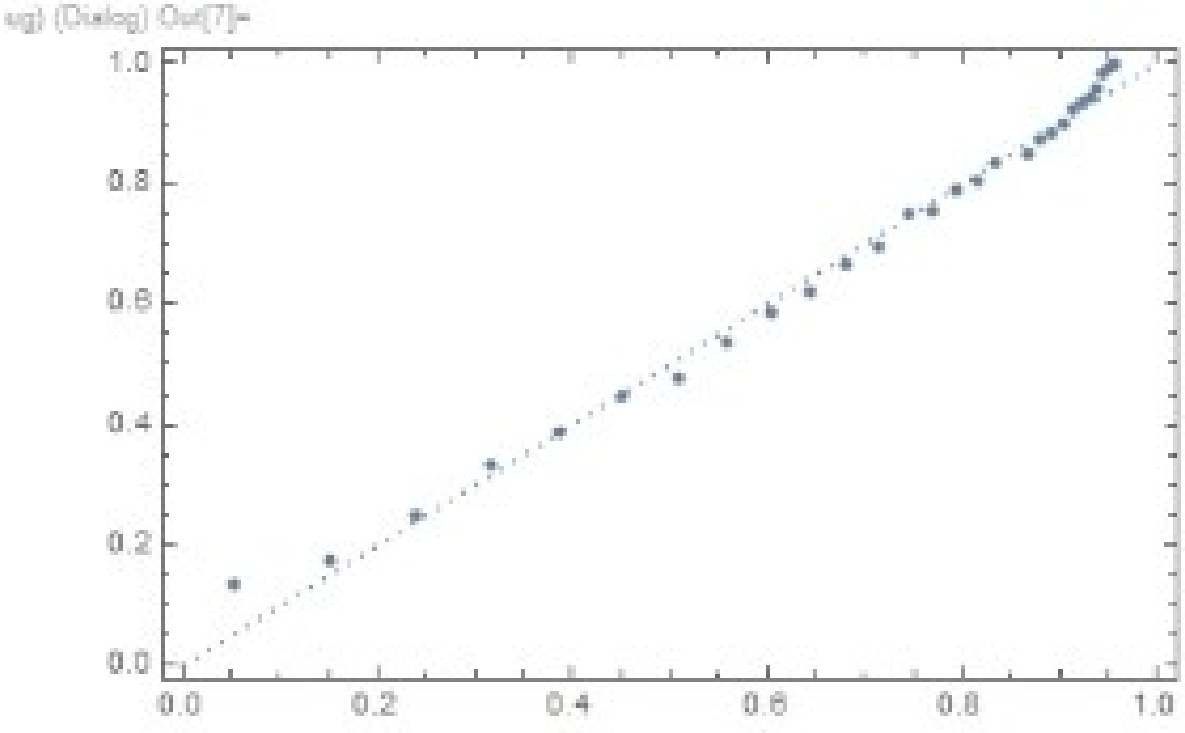


Figure 3.2: The probability plot (Q-Q plot) of exponential distribution fitted to the data

Figure 3.3: histigram and probability plot of exponential fitted to the data

Test of hypothesis(Goodness of fit): CramÃ©r-von Mises Test H0: The data follows exponential distribution H1: The data does not follow exponential distribution Level of significance: 0.05 Test statistic:

$$W^2 = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \, dF(x)$$

Where: $F_n(x)$ is the empirical cumulative distribution function (ECDF) of the observed data F(x) is the cumulative distribution function (CDF) of the exponential distribution Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation: Conclusion: Since P-value

Statistic	P-value
0.140193	0.420894

Table 3.2:

= 0.420894 is greater than level of significance (0.05), we do not have enough evidence to reject the null hypothesis. Therefore the exponential distribution is fit for the data.

Test of hypothesis(Goodness of fit): $WatsonU^2$ Test

$H0 : U^2 = 0$

$H1 : U^2 > 0$ Level of significance: 0.05 Test statistic:

$$U^2 = \frac{\sum_{t=2}^{T}(X_t - X_{t-1})^2}{\sum_{t=1}^{T} X_t^2}$$

Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
0.102519	0.264267

Table 3.3:

Conclusion: Since P-value = 0.264267 is greater than level of significance (0.05), we do not have enough evidence to reject the null hypothesis. Therefore the data may be adequately described by the exponential distribution, in other words, exponential distribution fits for modeling the data.

Test of hypothesis(Goodness of fit): Kolmogorov-Smirnov Test H0: The data is drawn from an exponential distribution

H1: The data is not drawn from an exponential distribution Level of significance: 0.05 Test statistic:

$$D = \max |F_n(x) - F(x)|$$

Where: $(F_n(x))$ is the empirical cumulative distribution function of the observed data defined as the proportion of data points less than or equal to (x) (F(x)) is the cumulative distribution function of the theoretical distribution being tested (max) denotes the maximum absolute difference taken over all values of (x) in the dataset. Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
0.081459	0.387844

Table 3.4:

Conclusion: Since P-value = 0.387844 is greater than level of significance (0.05), we do not have enough evidence to reject the null hypothesis. Therefore, we conclude that the data is drawn from an exponential distribution.

Test of hypothesis(Goodness of fit): Pearson Chi-square Test

H0: There is no significant difference between the observed and the expected frequencies

H1: There is a significant difference between the observed and the expected frequencies Level of significance: 0.05 Test statistic:

$$[\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}]$$

Where: (k) is the number of categories or cells in the contingency table (O_i) is the observed frequency in the (i)-th category (E_i) is the expected frequency in the (i)-th category under the null hypothesis. Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation: Conclusion: Since P-value = 0.0363638 is less than the level of significance (0.05), we reject the null

Statistic	P-value
23.4706	0.0363638

Table 3.5:

hypothesis and conclude that there is a significant difference between the observed and the expected frequencies. In other words, the observed data does not fit an exponential distribution well.

Case 2: Half-normal distribution

Test of hypothesis(Goodness of fit): Anderson-Darling Test H0: The data follows half-normal distribution H1: The data does not follow half-normal distribution Level of significance: 0.05 Test statistic:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}[(2i-1)\cdot(\ln(F(X_i)) + \ln(1-F(X_{n+1-i})))]$$

Where: (n) is the sample size $F(x_i)$ is the empirical cumulative distribution function (ECDF) evaluated at the ith ordered data point (x_i) (x_i) are the ordered data point Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
2.25688	0.0667498

Table 3.6:

Conclusion: Since P-value = 0.0667498 is greater than level of significance (0.05), we do not have enough evidence to reject the null hypothesis. Therefore the half-normal distribution is fit for the data.

Test of hypothesis(Goodness of fit): CramÃ©r-von Mises Test H0: The data follows half-normal distribution

H1: The data does not follow half-normal distribution

Level of significance: 0.05 Test statistic:

$$W^2 = \int_{-\infty}^{\infty}[F_n(x) - F(x)]^2\, dF(x)$$

Where: $F_n(x)$ is the empirical cumulative distribution function (ECDF) of the observed data F(x) is the cumulative distribution function (CDF) of the exponential distribution Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
0.281274	0.15278

Table 3.7:

Conclusion: Since P-value = 0.15278 is greater than level of significance (0.05), we do not have enough evidence to reject the null hypothesis. Therefore the half-normal distribution is fit for the data.

Test of hypothesis(Goodness of fit): $WatsonU^2$ Test

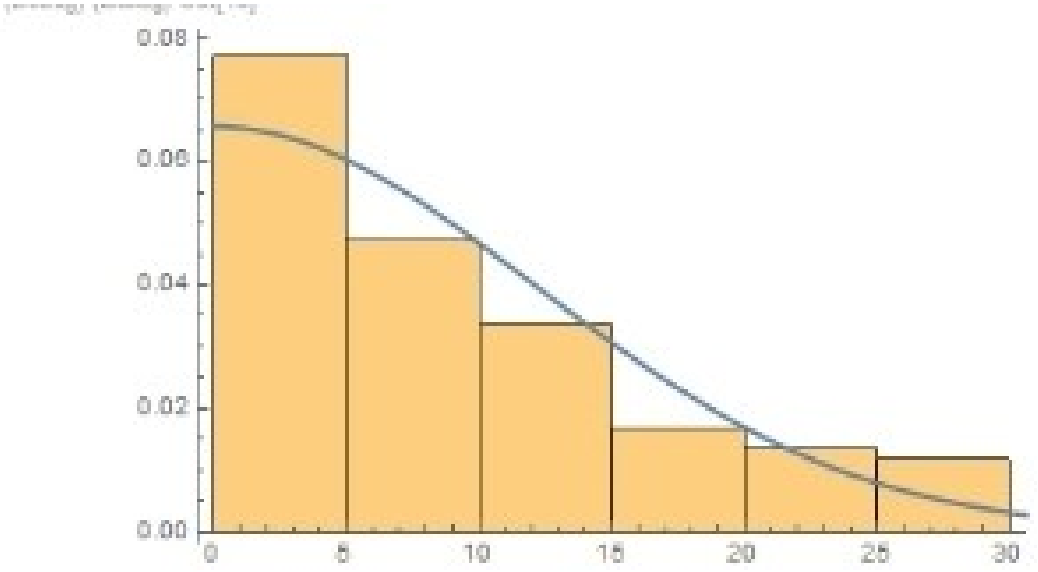
$H0 : U^2 = 0$

$H1 : U^2 > 0$ Level of significance: 0.05 Test statistic:

$$U^2 = \frac{\sum_{t=2}^{T}(X_t - X_{t-1})^2}{\sum_{t=1}^{T}X_t^2}$$

Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Conclusion: Since P-value = 0.097652 is greater than level of significance (0.05), we do not have enough evidence to reject the null hypothesis. Therefore the data may be adequately described by the half-normal distribution, in other words, half-normal distribution fits for modelling the data.



0.89

Figure 3.4: The histogram of half normal distribution fitted to the data

0.89

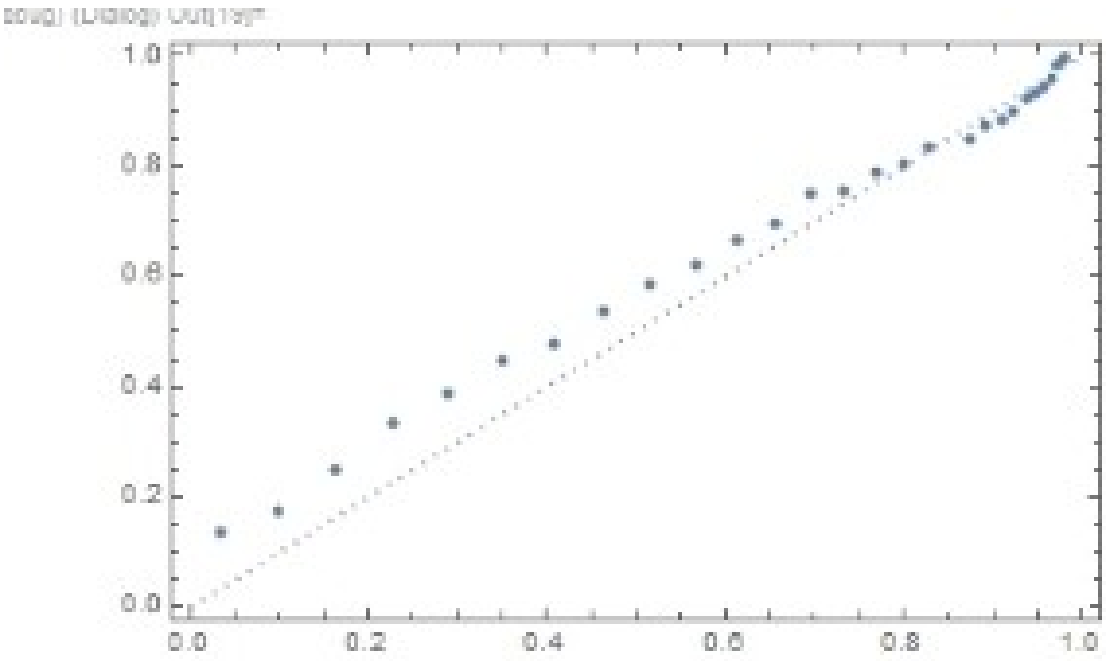


Figure 3.5: The probability plot (Q-Q plot) of half normal distribution fitted to the data

Figure 3.6: histigram and probability plot of half normal distribution fitted to the data

Statistic	P-value
0.152767	0.097652

Table 3.8:

Test of hypothesis(Goodness of fit): Kolmogorov-Smirnov Test H0: The data is drawn from a half-normal distribution H1: The data is not drawn from a half-normal distribution Level of significance: 0.05 Test statistic:

$$D = \max |F_n(x) - F(x)|$$

Where: $(F_n(x))$ is the empirical cumulative distribution function of the observed data defined as the proportion of data points less than or equal to (x) (F(x)) is the cumulative distribution function of the theoretical distribution being tested (max) denotes the maximum absolute difference taken over all values of (x) in the dataset. Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
0.10937	0.107606

Table 3.9:

Conclusion: Since P-value = 0.107606 is greater than level of significance (0.05), we do not have enough evidence to reject the null hypothesis. Therefore, we conclude that the data is drawn from a half-normal distribution.

Test of hypothesis(Goodness of fit): Pearson Chi-square Test H0: There is no significant difference between the observed and the expected frequencies H1: There is a significant difference between the observed and the expected frequencies Level of significance: 0.05 Test statistic:

$$[\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}]$$

Where: (k) is the number of categories or cells in the contingency table (O_i) is the observed frequency in the (i)-th category (E_i) is the expected frequency in the (i)-th category under the null hypothesis. Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

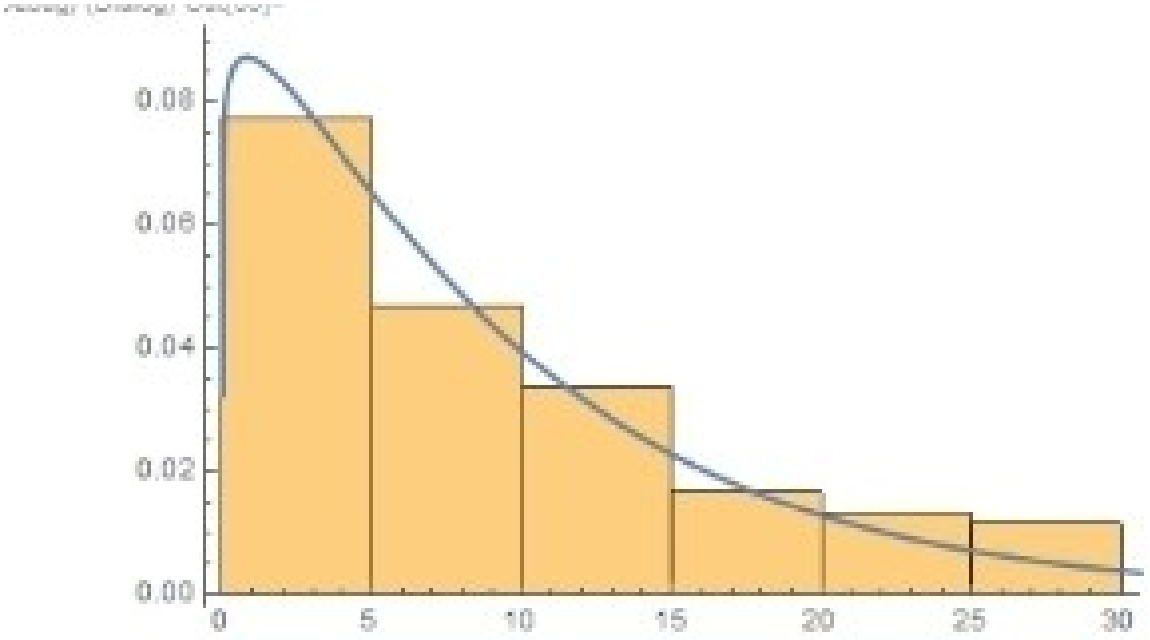
Statistic	P-value
19.7059	0.102791

Table 3.10:

Conclusion: Since P-value = 0.102791 is greater than the level of significance (0.05), we do not have enough evidence to reject the null hypothesis, therefore we conclude that there is no significant difference between the observed and the expected frequencies. In other words, the observed data does fit a half-normal distribution well.

Case 3: Weibull distribution **Test of hypothesis(Goodness of fit):** Anderson-Darling Test H0: The data follows Weibull distribution H1: The data does not follow Weibull distribution Level of significance: 0.05 Test statistic:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} [(2i-1) \cdot (\ln(F(X_i)) + \ln(1 - F(X_{n+1-i})))]$$



0.89

Figure 3.7: The histogram of Weibull distribution fitted to the data

0.89

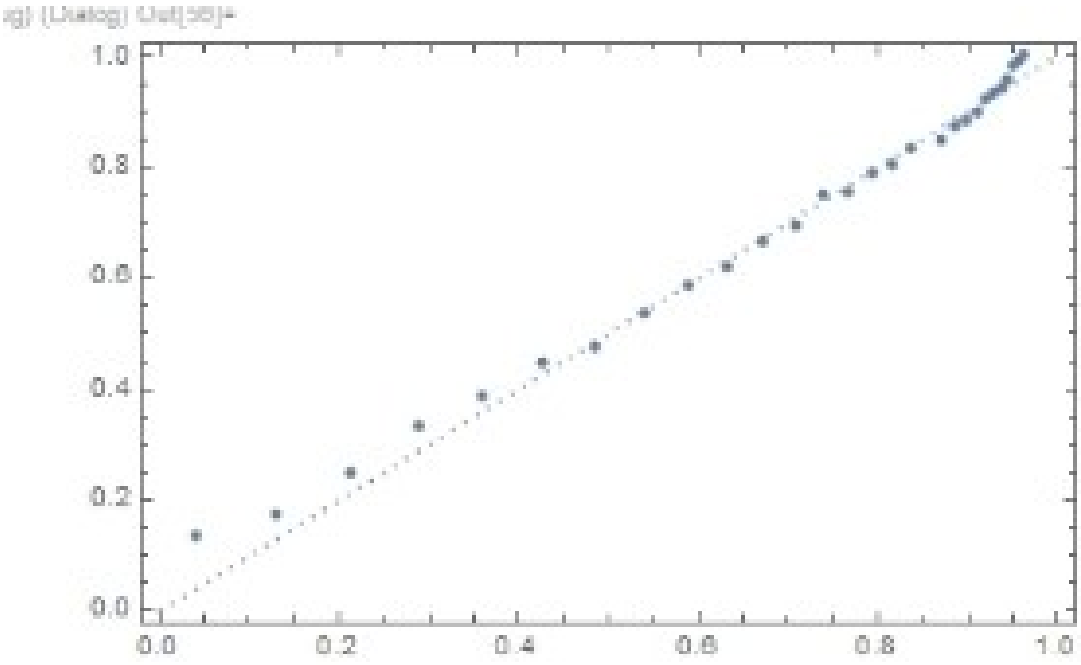


Figure 3.8: The probability plot (Q-Q plot) of Weibull distribution fitted to the data

Figure 3.9: histigram and probability plot of Weibul distribution fitted to the data

Where: (n) is the sample size $F(x_i)$ is the empirical cumulative distribution function (ECDF) evaluated at the ith ordered data point (x_i) (x_i) are the ordered data point Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
1.10606	0.305774

Table 3.11:

Conclusion: Since P-value = 0.305774 is greater than level of significance (0.05), we do not have enough evidence to reject the null hypothesis. Therefore Weibull distribution is fit for the data.

Test of hypothesis(Goodness of fit): CramÃ©r-von Mises Test H0: The data follows Weibull distribution

H1: The data does not follow Weibull distribution

Level of significance: 0.05

Test statistic:

$$W^2 = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \, dF(x)$$

Where: $F_n(x)$ is the empirical cumulative distribution function (ECDF) of the observed data F(x) is the cumulative distribution function (CDF) of the exponential distribution Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
0.0957118	0.606309

Table 3.12:

Conclusion: Since P-value = 0.606309 is greater than level of significance (0.05), we do not have enough evidence to reject the null hypothesis. Therefore Weibull distribution is fit for the data.

Test of hypothesis(Goodness of fit): $WatsonU^2$ Test H0: U^2 = 0 H1: U^2 > 0 Level of significance: 0.05 Test statistic:

$$U^2 = \frac{\sum_{t=2}^{T} (X_t - X_{t-1})^2}{\sum_{t=1}^{T} X_t^2}$$

Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
0.0930546	0.318549

Table 3.13:

Test of hypothesis(Goodness of fit): Kolmogorov-Smirnov Test H0: The data is drawn from Weibull distribution

H1: The data is not drawn from Weibull distribution

Level of significance: 0.05 Test statistic:

$$D = \max |F_n(x) - F(x)|$$

Where: $(F_n(x))$ is the empirical cumulative distribution function of the observed data defined as the proportion of data points less than or equal to (x) (F(x)) is the cumulative distribution function of the theoretical distribution being tested (max) denotes the maximum absolute difference taken over all values of (x) in the dataset. Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation: Conclusion: Since P-value = 0.235802 is greater than level of significance

Statistic	P-value
0.093347	0.235802

Table 3.14:

(0.05), we do not have enough evidence to reject the null hypothesis. Therefore, we conclude that the data is drawn from a Weibull distribution.

Test of hypothesis(Goodness of fit): Pearson Chi-square Test H0: There is no significant difference between the observed and the expected frequencies

H1: There is a significant difference between the observed and the expected frequencies

Level of significance: 0.05 Test statistic:

$$[\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}]$$

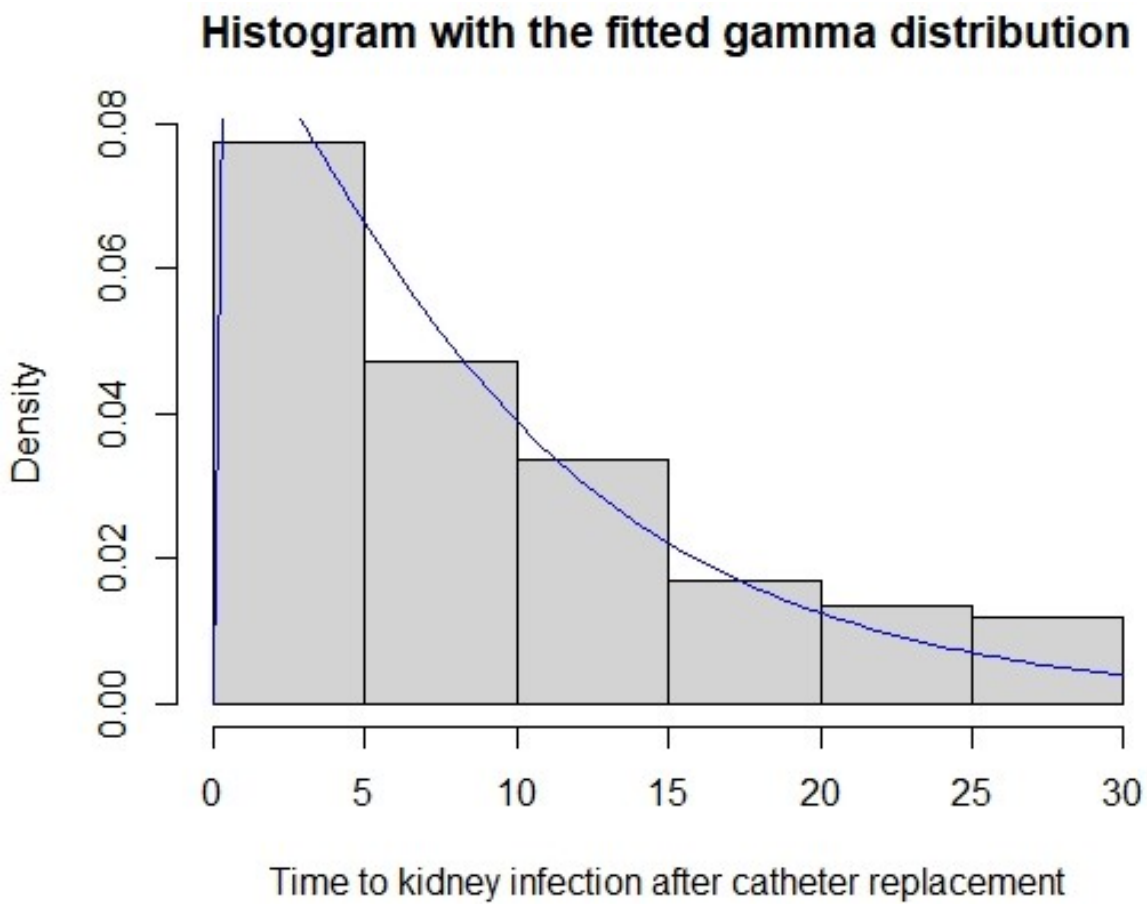
Where: (k) is the number of categories or cells in the contingency table (O_i) is the observed frequency in the (i)-th category (E_i) is the expected frequency in the (i)-th category under the null hypothesis. Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
23.2353	0.0389378

Table 3.15:

Conclusion: Since P-value = 0.0389378 is less than the level of significance (0.05), we reject the null hypothesis and conclude that there is a significant difference between the observed and the expected frequencies. In other words, the observed data does not fit Weibull distribution well

Case 4: Gamma distribution



0.89

Figure 3.10: The histogram of Gamma distribution fitted to the data

0.89

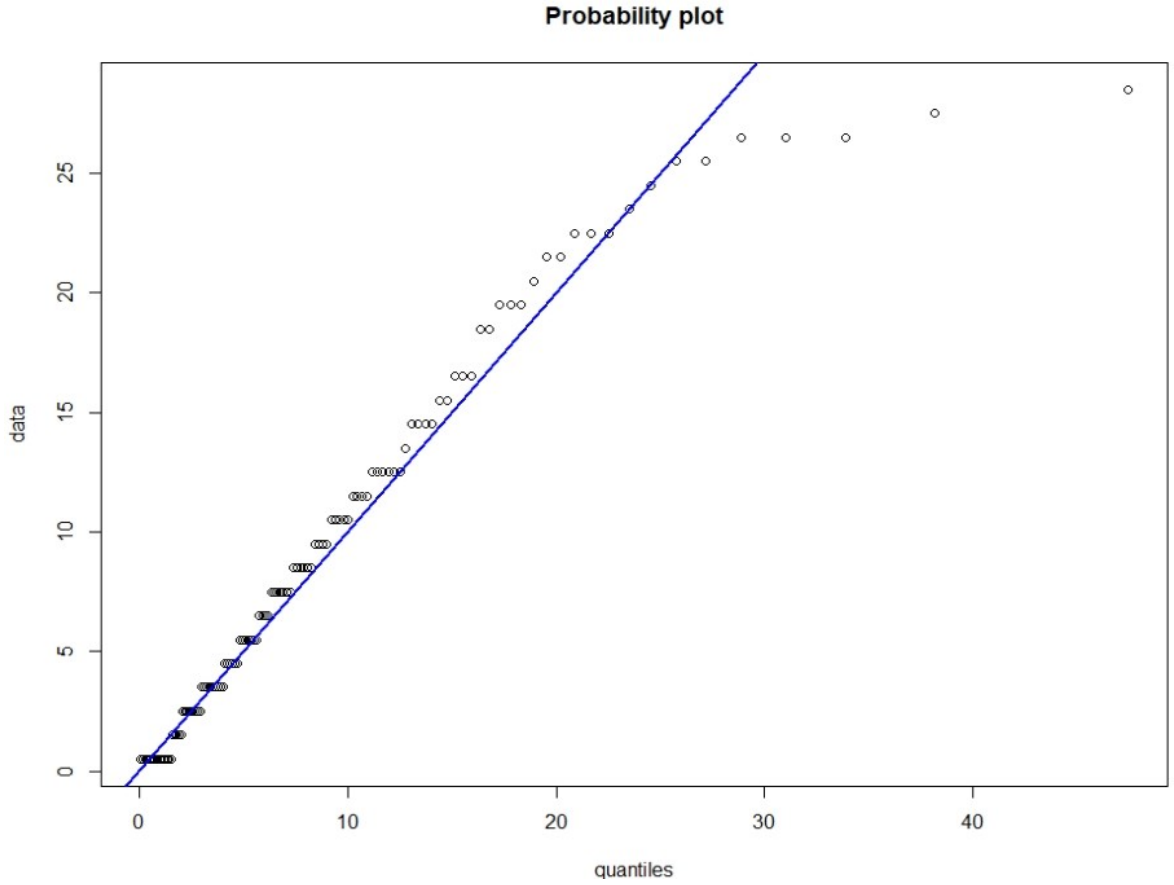


Figure 3.11: The probability plot (Q-Q plot) of Gamma distribution fitted to the data

Test of hypothesis(Goodness of fit): Anderson-Darling Test H0: The data follows gamma distribution
H1: The data does not follow gamma distribution
Level of significance: 0.05 Test statistic

Where: (n) is the sample size $F(x_i)$ is the empirical cumulative distribution function (ECDF) evaluated at the ith ordered data point (x_i) (x_i) are the ordered data point Decision rule: Reject the null hypothesis if critical-value is less than the calculated value (Statistic) Computation:

Statistic	P-value
1.194133	0.75017868

Table 3.16:

Conclusion: Since critical-value = 0.75017878 is less than the calculated value, we reject the null hypothesis and conclude that the gamma distribution is not fit for the data at 0.05 level of significance.

Test of hypothesis(Goodness of fit): CramÃ©r-von Mises Test H0: The data follows gamma distribution
H1: The data does not follow gamma distribution
Level of significance: 0.05
Test statistic:

$$W^2 = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \, dF(x)$$

Where: $F_n(x)$ is the empirical cumulative distribution function (ECDF) of the observed data F(x) is the cumulative distribution function (CDF) of the exponential distribution Decision rule: Reject the null hypothesis if critical-value is less than the calculated value (Statistic) Computation: Conclusion: Since

Statistic	P-value
0.1146312	0.1334904

Table 3.17:

critical-value = 0.1334904 is greater than the calculated value, we do not have enough evidence to reject the null hypothesis. Therefore the gamma distribution is fit for the data.

Test of hypothesis(Goodness of fit): $WatsonU^2$ Test H0: U^2 = 0
H1: U^2 > 0
Level of significance: 0.05
Test statistic:

$$U^2 = \frac{\sum_{t=2}^{T}(X_t - X_{t-1})^2}{\sum_{t=1}^{T} X_t^2}$$

Decision rule: Reject the null hypothesis if critical-value is less than the calculated value (Statistic) Computation:

Statistic	P-value
0.1043107	0.122977

Table 3.18:

Conclusion: Since critical-value = 0.122977 is greater than the calculated value, we do not have enough evidence to reject the null hypothesis. Therefore the data may be adequately described by gamma distribution, in other words, gamma distribution fits for modelling the data.

Test of hypothesis(Goodness of fit): Kolmogorov-Smirnov Test H0: The data is drawn from a gamma distribution

H1: The data is not drawn from a gamma distribution

Level of significance: 0.05

Test statistic:

$$D = \max |F_n(x) - F(x)|$$

Where: $(F_n(x))$ is the empirical cumulative distribution function of the observed data defined as the proportion of data points less than or equal to (x) (F(x)) is the cumulative distribution function of the theoretical distribution being tested (max) denotes the maximum absolute difference taken over all values of (x) in the dataset. Decision rule: Reject the null hypothesis if critical-value is less than the calculated value (Statistic) Computation:

Statistic	P-value
0.09334785	0.08507054

Table 3.19:

Conclusion: Since critical-value = 0.08507054 is less than the calculated value, we reject the null hypothesis and conclude that the data is drawn from a gamma distribution.

Test of hypothesis(Goodness of fit): Pearson Chi-square Test H0: There is no significant difference between the

observed and the expected frequencies H1: There is a significant difference between the

observed and the expected frequencies Level of significance: 0.05

Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Where: (k) is the number of categories or cells in the contingency table (O_i) is the observed frequency in the (i)-th category (E_i) is the expected frequency in the (i)-th category under the null hypothesis. Decision rule: Reject the null hypothesis if p-value is less than the level of significance (0.05) Computation:

Statistic	P-value
0.4285714	0.5126908

Table 3.20:

Conclusion: Since P-value = 0.5126908 is greater than the level of significance (0.05), we do not have enough evidence to reject the null hypothesis, therefore we conclude that there is no significant difference between the observed and the expected frequencies. In other words, the observed data does fit a gamma distribution well.

Test of hypothesis	**Exponential-D**	**Half-N**	**Weibull**	**Gamma**
Anderson	0.27892	0.06674	0.305	0.75017868
Crammer	0.42089	0.15278	0.6063	0.1334904
Watson	0.26426	0.09765	0.31854	0.122977
KMJV	0.38784	0.10760	0.23580	0.08507054
Pearson	0.03636	0.10279	0.03893	0.5126908

Table 3.21: P-values for different test statistics

3.4 DISCUSSION OF RESULTS

From case 1 which is the case of the exponential distribution, the plotted histogram fig 1, shows that the data approximately followed an exponential distribution. The probability plot (Q-Q plot) fig 2 shows that the data approximately follows an exponential distribution also because the dots are lying closer to the diagonal line. Further test is carried out to see if exponential distribution fits the data, the Anderson-Darling, CramÃ©r-von Mises, $WatsonU^2$, Kolmogorov-Smirnov and Pearson Chi-square goodness of fit test. All other tests besides the Pearson Chi-square test indicates good fit, Pearson Chi-square opposed the motion with P-value = 0.27892 which is less than the level of significance (0.05). Case 2 which is the case of half-normal distribution, case 3 which is Weibull distribution, case 4 which is the gamma distribution and case 5 which is the normal distribution are considered as well to see which is best for the data. For Case 2, which is the half-normal distribution, all the test including the Pearson Chi-square test shows that the half-normal distribution is best for the data. The histogram plotted exhibits the characteristics of half-normal distribution, the probability plot as well indicates good fit. The last case which is the normal distribution shows that the data does not follow normal distribution, the histogram is not exhibiting the characteristics of the normal distribution and the dots are dispersed obviously from the probability line therefore, there is no need for further test. Case 3 which is the Weibull distribution just like case 1, shows evidence of good fit to the data from the histogram and Q-Q plot and supported by all other goodness of fit test accept the Pearson Chi-square test. Case 4 which is the gamma distribution, shows evidence of good fit to the data from the histogram and Q-Q plot and supported by all other goodness of fit test accept the Anderson-Darling test and Kolmogorov-Smirnov test. Besides the case of normal distribution, all other distributions considered in this text shows a good fit to the data. The best fit is the half normal distribution which is not opposed by any goodness of fit test which was employed in this text.

CHAPTER 4

SUMMARY, CONCLUSION AND RECOMMENDATION

4.1 SUMMARY AND CONCLUSION

The Exponential distribution, half normal distribution, Weibull distribution, gamma distribution, and normal distribution were fitted to the data, their histogram and probability plot were obtained, and all other distributions besides the normal distribution depicted a good fit. Several goodness of fit tests were conducted for further confirmation. For the exponential, the Weibull distribution, and the gamma distribution, the goodness of fit tests showed a good fit besides the Pearson Chi-square test in the two former cases, the Anderson-Darling test and Kolmogorov-Smirnov test in the latter case. In the case of half-normal distribution, all the goodness of fit test including the Pearson Chi-square test indicates that it is a good fit for the data. The half-normal distribution was then deemed the best fit for the data.

4.2 RECOMMENDATION

- The half-normal distribution is identified as the best fit for the data based on various goodness-of-fit tests, including the Pearson chi-square test. This distribution should be used to model and describe the underlying probability distribution of the data in future analyses.
- Since the half-normal distribution has been identified as the best fit, further analysis should be conducted using this distribution to explore relationships, make predictions, or perform other statistical tasks relevant to the research objectives.
- While the half-normal distribution appears to be the best fit based on the conducted tests, it's essential to validate the chosen distribution through additional analyses and sensitivity checks. This may involve assessing the robustness of the findings under different assumptions or investigating potential sources of bias or uncertainty.

Bibliography

Safdar, N., Kluger, D. M., & Maki, D. G. (2002). A review of risk factors for catheter-related bloodstream infection caused by percutaneously inserted, noncuffed central venous catheters: implications for preventive strategies. *81*(6), 466-479

Chen, W., Wang, Z., Wang, G., Cao, C., Hong, B., Liu, J., ... & Wang, R.(2024) A meta-analysis of risk factors for a Dacron-cuffed catheter related infection in hemodialysis. BMC nephrology, *25*(1),126

Guo, H., Zhang, L., He, H., & Wang, L. (2024). isk factors for catheter-associated bloodstream infection in hemodialysis patients: A meta-analysis. *Plos one*, *19*(3), e0299715.

Wang et al.(2018)Wang, J., Huang, X., Li, Q., & Ma, X. (2018) Comparison of seven methods for determining the optimal statistical distribution parameters: A case study of wind energy assessment in the large-scale wind farms of China. *Energy*, 164, 432-448.

Totaro, V., Gioia, A., Kuczera, G., & Iacobellis, V. Modelling multidecadal variability in flood frequency using the Two-Component Extreme Value distribution. *Stochastic Environmental Research and Risk Assessment*, 1-18

Kumar, R., Arora, H. C., Prabhakar, V., Ram, S., & Singh, D. K. A Study on Comparison of Six Probability Distributions for Extreme Value Analysis of Rainfall Data. *i-Manager's Journal on Civil Engineering*, *11*(3),12.

Chen, R. T., Behrmann, J., Duvenaud, D. K., & Jacobsen, J. H Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*,32

Goligher, E. C., Tomlinson, G., Hajage, D., Wijeysundera, D. N., Fan, E., JÃ¼ni, P., ... & Combes, A. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome and posterior probability of mortality benefit in a post hoc Bayesian analysis of a randomized clinical trial. *Jama*, *320*(21), 2251-2259.

Bond, A., Chadwick, P., Smith, T. R., Nightingale, J. M., & Lal, S.(2020) Diagnosis and management of catheter-related bloodstream infections in patients on home parenteral nutrition. *Frontline Gastroenterology*,*11*(1),48-54.

Bae, S., Kim, Y., Chang, H. H., Kim, S., Kim, H. J., Jeon, H., ... & Kim, S. W. (2022). The effect of the automatic notification of catheter days on reducing central line-related bloodstream infection.