Report on Statistical Inferences Programming Assignment (Done using R)

The data used in this assignment is:https://docs.google.com/spreadsheets/d/1R77YoaSfXnwWAWWtIFWOkncS8vG3TeNQnLeXon_nK3Q/edit?usp=sharing

1. We had to find a distribution that would best fit the data.

Step 1: Data Loading

We first load the data from a CSV file. The data, assumed to be stored in a single column, is loaded into the variable `x`.

Step 2: Parameter Estimation

Then we estimate the parameters for four different distributions: Normal, Exponential, Log-normal, and Gamma. Each set of parameters is calculated based on the data in `x`:

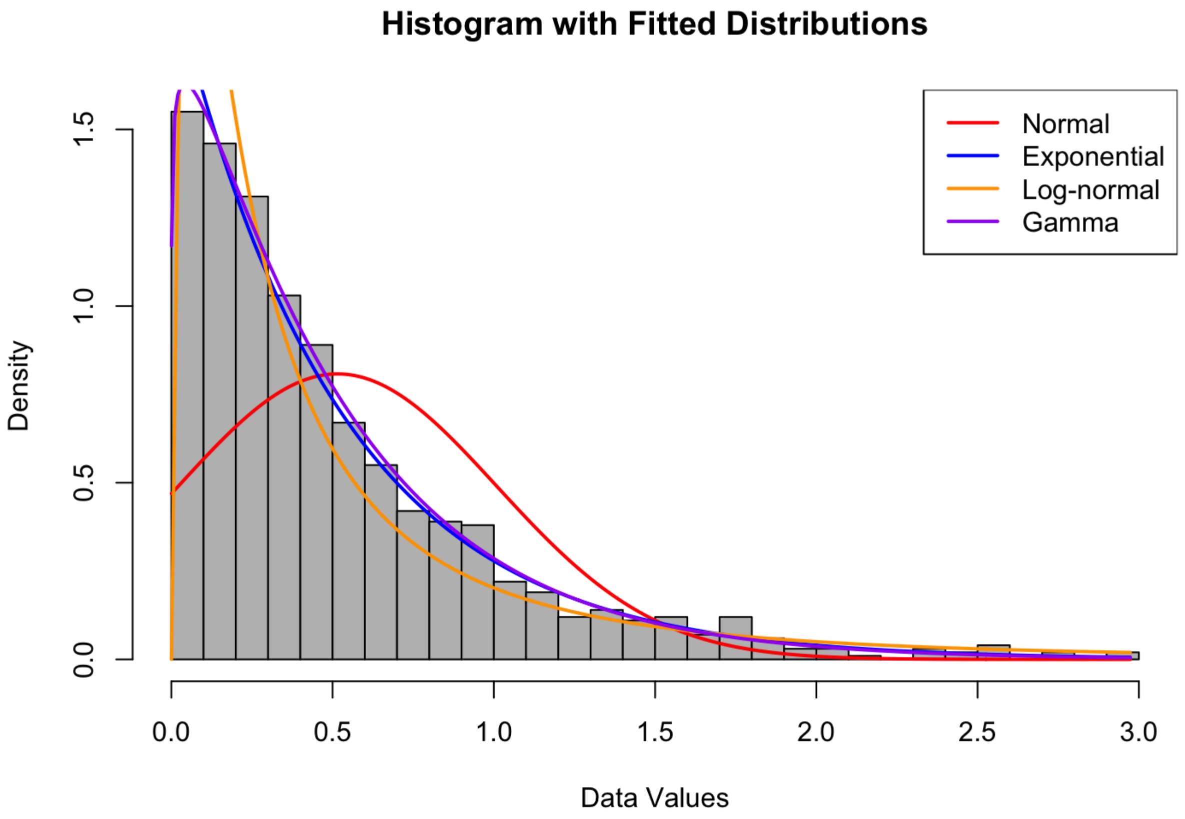
- **Normal Distribution**: The mean and standard deviation are computed.
- **Exponential Distribution**: The rate parameter is estimated as the reciprocal of the mean.

- **Log-normal Distribution**: Parameters are estimated based on the logarithm of the data, specifically calculating the mean and standard deviation of the log-transformed data.
- **Gamma Distribution**: The shape and rate parameters are estimated using the method of moments, which involves the mean and variance of the data.

Step 3: Plotting Data and Fitted Distributions

A histogram of the data is plotted with probability density functions (PDFs) for each estimated distribution overlaid:

- The histogram provides a visual representation of the data distribution.
- Lines representing the PDFs of each fitted distribution are added on top of the histogram, each in a different color to distinguish among them.



Step 4: Goodness-of-Fit Testing

Kolmogorov-Smirnov tests are performed to evaluate how well each distribution fits the data. This test compares the cumulative distribution function (CDF) of the sample against the CDF of each hypothesized model:

- The results include a statistic value (D) and a p-value for each distribution.
- A higher p-value suggests a better fit because it indicates less evidence against the hypothesis that the data follows the tested distribution.

Step 5: Determining the Best Fit

The distribution with the highest p-value from the Kolmogorov-Smirnov tests is determined to be the best fit for the data. This selection is based on statistical evidence provided by the goodness-of-fit tests.

2. List at least 2 estimators of the parameter(s) involved in the underlying distribution

1. Method of Moments (MoM) Estimation:

This method involves equating the theoretical moments of the distribution (mean and variance) with the empirical moments (sample mean and sample variance) to solve for the distribution parameters.

Gamma Distribution Parameters: The shape (α) and rate (β) of the Gamma distribution can be expressed in terms of the mean (μ) and variance (σ^2) of the distribution:

α= μ^2/σ^2

$B = \mu/\sigma$^2

- **Calculation**: The script calculates α and β by plugging in the sample mean and variance into the formulas.
- **Output**: The estimated parameters are printed using `cat()` and `sprintf()` functions, which format the output as strings.

2. Maximum Likelihood Estimation (MLE):

MLE seeks the parameter values that maximize the likelihood function, assuming that the observed data are the most probable given the parameters.

- **Log-Likelihood Function**: Defined to calculate the negative log-likelihood of the Gamma distribution given parameters α (shape) and β (rate). It uses the `dgamma()` function to compute the density of the Gamma distribution for data x and then sums the log of these densities. The sign is negative because `optim()` minimizes the function it is given.
- **Parameter Transformation**: Parameters are transformed using the exponential function within the log-likelihood function to ensure they remain positive, a requirement for Gamma distribution parameters. The `log_params` are the parameters to be optimized, and transforming them back and forth ensures that the optimization algorithm operates in the unconstrained space of real numbers while evaluating only positive parameter values.
- **Optimization**: Uses the `optim()` function with the BFGS method, a quasi-Newton method suitable for this kind of non-linear optimization. Initial guesses for α and β are provided in the log space and optimized to find the parameter values that minimize the negative log-likelihood.
- **Output**: The estimated parameters are again printed, showing the results of the MLE.

3. a. Classify the estimators in (1) into the unbiased, consistent or efficient estimators. b. Find the estimates from the data.

Method of Moments (MoM) Estimators:

- **Estimation Process**:
  - The MoM estimators for the Gamma distribution's shape (α) and rate (β) parameters are calculated using the sample's mean and variance.
  - Shape (α): α=mean^2/variance
  - Rate (β): β=mean/variance
- **Classification**:
  - **Consistency**: MoM estimators are generally consistent because, as the sample size increases, the sample moments (mean and variance) converge to their population counterparts under the law of large numbers.
  - **Unbiasedness**: MoM estimators are not necessarily unbiased, especially for complex distributions like the Gamma distribution, where the relationship between the parameters and the moments isn't linear.
  - **Efficiency**: They are usually not the most efficient, particularly when compared to MLE, as they do not minimize any form of error variance like the MLE which aims at the Cramér-Rao lower bound.

Maximum Likelihood Estimation (MLE) Estimators:

- **Estimation Process**:
  - The log-likelihood function for the Gamma distribution is formulated, where α and β are the parameters to be estimated. The parameters within the function are initially transformed using the exponential function to ensure they are positive, addressing the constraint that both shape and rate must be positive.
  - The `optim` function in R is used to find the parameter values that minimize the negative of this log-likelihood function. The optimization starts with initial guesses for α and β provided in the logarithmic space.
- **Classification**:

- **Unbiasedness**: MLE is asymptotically unbiased. This means that as the sample size increases, the bias of the MLE estimators decreases, converging to zero.
- **Consistency**: MLE estimators are consistent under regular conditions — as the sample size grows, the estimators converge in probability to the true parameter values.
- **Efficiency**: MLE achieves asymptotic efficiency, meaning that in large samples, it attains the lower bound of the variance as described by the Cramér-Rao bound, making it the most efficient estimator among all unbiased estimators.

4. Find two parameters of the distribution using Method of Moments.

1. **Method of Moments Estimation:**
   - Using the relationships derived from the properties of the Gamma distribution:
     - **Shape** (α): It's estimated as $\alpha = \text{mean}^\wedge 2/\text{variance}$. This calculation uses the square of the sample mean divided by the sample variance.
     - **Rate** (β): It's estimated as $\beta=\text{mean}/\text{variance}$, which utilizes the ratio of the sample mean to the sample variance.
   - These formulas assume that the mean of the Gamma distribution equals α/β and the variance equals $\alpha/\beta^\wedge 2$.
2. **Output the Estimates:**
   - The estimated parameters for the shape and rate of the Gamma distribution are displayed using the `cat()` and `sprintf()` functions, providing a formatted output directly in the console.

5. Find the Uniformly Minimum Variance Unbiased Estimator (UMVUE) of the parameter’s and find its esti- mate from the data.

1. Estimate Alpha Using Method of Moments (MoM)

The code first estimates the shape parameter (α) of the Gamma distribution using the Method of Moments:

```
estimated_alpha <- (mean(x)^2) / var(x)
```

- **Method of Moments for Alpha**: This formula derives α by squaring the sample mean and dividing by the sample variance, which is a standard approach in MoM for Gamma distributions.

2. Calculate the Sample Mean

The mean of the dataset x is calculated, which will be used to estimate the rate parameter (β):

```
sample_mean <- mean(x)
```

- **Sample Mean**: Provides the average value of the data points in x, representing the expected value of the distribution per unit of measure.

3. Calculate UMVUE for Beta

The UMVUE for the rate parameter (β) is calculated using the previously estimated α and the sample mean:

```
umvue_beta <- estimated_alpha / sample_mean
```

- **UMVUE for Beta**: Here, β is estimated as α divided by the sample mean. This formula stems from the relationship between the mean (μ) and the parameters of the Gamma distribution, where $\mu=\alpha/\beta$. The estimator is called UMVUE because it uses the complete sufficient statistic (sample mean) and incorporates an unbiased estimator of the parameter α, leading to the minimum variance among all unbiased estimators under this configuration.

4. Output the UMVUE for Beta

Finally, the estimated value of β is displayed:

```
cat("UMVUE for Rate (β) assuming alpha is known or estimated:\n")

cat(sprintf("UMVUE Beta = %f\n", umvue_beta))
```

- **Output**: This prints out the estimated β value, labeling it as a UMVUE under the assumption that α is either known or correctly estimated.

6.Find the interval estimator of any one parameter of the population distribution with confidence α = 0.01; 0.05; 0.1.

1. Parameter Estimation

- **Alpha Estimation**: The shape parameter (α) of the Gamma distribution is estimated using MoM based on the sample mean and variance:

  ```
  alpha_estimated <- (mean(x)^2) / var(x)
  ```

- **Sample Characteristics**: It calculates the sample size n and the sample mean.
- **Beta Estimation**: Utilizes the UMVUE formula, where β is estimated as α divided by the sample mean:

2. Standard Error Calculation

- Calculates the standard error for the reciprocal of β $(1/\beta)$, which is derived from the variance of the estimator:
- `se_one_over_beta <- sqrt(1 / (n * alpha_estimated))`

3. Confidence Interval Calculation

- **Z-values**: Retrieves the Z-values corresponding to specified confidence levels (0.01, 0.05, 0.1) using the normal distribution quantiles:

  ```
  z_values <- qnorm(1 - conf_levels / 2)
  ```

- **Confidence Intervals**: The script calculates confidence intervals using the Z-values and the standard error. The interval calculation assumes normality for the distribution of $1/\beta$. The intervals are computed inversely from margins calculated around $1/\beta$, which provides intervals for β itself:

  ```
  ci <- sapply(z_values, function(z) { error_margin <- z *
  se_one_over_beta lower_bound <- 1 / (1/beta_estimated +
  error_margin) upper_bound <- 1 / (1/beta_estimated -error_margin)
  return(c(lower_bound, upper_bound)) })
  ```

4. Output

- **Printing Results**: Displays the computed confidence intervals for β for each specified confidence level, offering insights into the uncertainty and reliability of the β estimate:

```
cat("Confidence Intervals for Beta at Different Confidence Levels:\n")

sapply(1:length(conf_levels), function(i) {

    cat(sprintf("Alpha = %.2f: (%.4f, %.4f)\n", conf_levels[i], ci[1, i],
ci[2, i]
```

7. Test the hypothesis that the mean μ is equal to μ0

1. Calculate Sample Statistics

- **Sample Mean** (sample_mean) and **Standard Deviation** (sample_sd) are computed from the dataset x.
- **Sample Size** (n) is determined using `length(x)`.

2. Set Hypothesized Mean

- μ0 is defined as 0.52, based on a reasonable assumption about the expected value of the dataset.

3. Perform Z-test

A Z-test statistic is calculated to determine how many standard deviations the sample mean deviates from the hypothesized mean, adjusted for sample size. The formula used is:

Z=(sample_mean−μ0)(sample_sd/root(n))

The **P-value** is computed for a two-tailed test, assessing both possibilities of the sample mean being either greater than or less than μ0.

4. Output Results and Conclusion

- Displays the sample mean, hypothesized mean, Z-statistic, and P-value.
- Based on the P-value:
  - **If P-value < 0.05**: Rejects the null hypothesis, indicating significant evidence that the sample mean differs from μ0

- **If P-value ≥ 0.05**: Fails to reject the null hypothesis, suggesting insufficient evidence to conclude that the sample mean is different from μ0.

8. Test the hypothesis that the variance σ^2 is equal σ0^2

1. Calculate Sample Statistics

- **Sample Variance** (sample_variance) is computed from the dataset *x*.
- **Sample Size** (*n*) is determined using `length(x)`.

2. Set Hypothesized Variance

- σ0^2 is defined as 0.25, based on a specific assumption about the expected variance of the dataset.

3. Perform Chi-squared Test for Variance

- **Chi-squared Statistic**: This statistic is calculated using the formula:

$$\chi2=(n-1)\times \text{sample\_variance}/\sigma0\text{^}2$$

This measures how much the observed variance deviates from the hypothesized variance under the null hypothesis.

- **P-value**: A two-tailed test is conducted to evaluate both possibilities (sample variance being greater or less than the hypothesized variance). The `pchisq` function is used to derive the p-value, and `min()` ensures that the lower of the two tail probabilities is doubled to reflect the two-tailed nature of the test.

4. Output Results and Conclusion

- **Output**: Displays the calculated sample variance, the hypothesized variance, the chi-squared statistic, and the p-value.
- **Conclusion**:
    - **If P-value < 0.05**: Rejects the null hypothesis, suggesting there is substantial evidence that the sample variance differs from the hypothesized value.
    - **If P-value ≥ 0.05**: Fails to reject the null hypothesis, indicating insufficient evidence to conclude that the sample variance is different from the hypothesized value.

9. Perform goodness of fit to test the hypothesis that the distribution of the collected data is same as the distribution you fitted in Qustion 1.

1. Parameter Estimation

- The shape (α) and rate (β) parameters of the Gamma distribution are estimated:
    - **Shape parameter (α)**: Calculated as the square of the sample mean divided by the sample variance.
    - **Rate parameter (β)**: Calculated as the sample mean divided by the sample variance.

2. Histogram Binning

- **Number of Bins**: The data is divided into 10 bins, a typical choice for such tests, but this might need adjustment based on the data distribution.
- **Observed Counts**: Counts of data points in each bin are calculated without plotting the histogram.
- **Break Points**: The break points that define the bins are also determined without plotting.

3. Expected Counts Calculation

- Using the Gamma cumulative distribution function (CDF), the expected counts for each bin are calculated based on the estimated parameters. The difference in CDF values between successive breaks multiplied by the total number of observations gives the expected frequency in each bin.

4. Chi-squared Test

- **Chi-squared Statistic**: Computed by summing the squared differences between observed and expected counts, normalized by the expected counts.
- **Degrees of Freedom**: Calculated as the number of bins minus one minus the number of parameters estimated from the data.
- **P-value**: Assesses whether the observed deviations from the expected distribution are significant. It is calculated from the chi-squared distribution.

5. Results and Conclusion

- **Outputs**: The chi-squared statistic and the p-value are displayed.
- **Interpretation**: If the p-value is less than 0.05, the null hypothesis that the data follows the specified Gamma distribution is rejected; otherwise, there is not enough evidence to reject the null hypothesis.