Statistics 520: Assignment 5

Sam Olson

[1] X [2] X [3] X [4] [5] [6] [7] [8]

Assignment 5

The objectives of this assignment are to (1) ensure that you have a grasp on using the tools of basic likelihood in a data analysis and (2) help you to continue to develop the precise use of notation in presenting descriptions of analyses. On the course web page is a file in the Data folder called `gammadat.txt`. This file contains two columns of values with a header having labels `group1` and `group2`. Each column should be considered to contain values corresponding to a set of independent and identical gamma random variables. That is, the two columns are values from two groups that we wish to compare using a two-sample model with gamma distributions. Consider the first column to contain values for Group 1 and the second column to contain values for Group 2.

A number of resources are available to you to help you complete this assignment. Chapter 5 of the course notes contains a summary of likelihood methods. In the Computing folder of the course web page is a file `newtraph.txt` that contains a generic Newton-Raphson algorithm that you may use for maximum likelihood estimation. There is also a file called `newtraphexplain.txt` that describes the inputs needed, the syntax, and the output. Alternatively you may choose to make use of the built-in R functions `optim` or `nlm`. Any of these options (or others you might know of if you prefer Matlab or something else) are fine as long as you know what you are doing and can produce the quantities needed to conduct the analysis.

Your answer should contain complete and consistent notation using no undefined symbols. You should always clearly explain what you computed and the formulas used. Your answer should not contain computer code or material from a "screen dump." You will not be awarded any points for such material. If you want to report estimated values do so in the text, as a list, or construct a table.

Again, do not include copied computer function output. You will not get credit for anything presented in that way.

1.

Assume random variables $Y_{1,1}, \ldots, Y_{1,n_1}$ and $Y_{2,1}, \ldots, Y_{2,n_2}$ have been defined for the responses in this problem. These responses are strictly positive numbers, and an assumption of independence is reasonable. Formulate a two-sample model using gamma distributions. For one group, write the form of the log likelihood that will need to be computed to find estimates and other inferential quantities.

Answer

Assuming the random variables defined as given, taking observed values that are strictly positive and iid.

Each group is then modeled with a (potentially different) Gamma distribution parameterized by the (α, β) parameters (shape and rate, repectively):

$$Y_{g,i} \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_g, \beta_g), \qquad g \in \{1, 2\}, \quad \text{and } i = 1, \ldots, n_g$$

Where, for our purposes $n_1 = n_2$ (equal sample sizes between the two groups of Gamma-distributed random variables).

With pdf of the form:

$$f(y \mid \alpha_g, \beta_g) = \frac{\beta_g^{\alpha_g}}{\Gamma(\alpha_g)} \, y^{\alpha_g - 1} e^{-\beta_g y}, \qquad y > 0, \ \alpha_g > 0, \ \beta_g > 0$$

Using the pdf, we then take the log to define a single group g's log-likelihood function by:

$$\ell_g(\alpha_g, \beta_g) = \sum_{i=1}^{n_g} \log f(y_{g,i} \mid \alpha_g, \beta_g) = n_g \Big(\alpha_g \log \beta_g - \log \Gamma(\alpha_g) \Big) + (\alpha_g - 1) \sum_{i=1}^{n_g} \log y_{g,i} - \beta_g \sum_{i=1}^{n_g} y_{g,i}$$

Note: The log-likelihood function is defined by 2 sufficient statistics, of the form:

$$T_{g,1} = \sum_{i=1}^{n_g} \log y_{g,i} \quad \text{and} \quad T_{g,2} = \sum_{i=1}^{n_g} y_{g,i}$$

So, for the score equations, computing MLEs (say, via Newton–Raphson or some other optimization methods) is done via:

$$\frac{\partial \ell_g}{\partial \alpha_g} = n_g \log \beta_g - n_g \psi_0(\alpha_g) + T_{g,1}, \qquad \frac{\partial \ell_g}{\partial \beta_g} = \frac{n_g \alpha_g}{\beta_g} - T_{g,2}$$

With the Hessian given by:

$$\frac{\partial^2 \ell_g}{\partial \alpha_g^2} = -n_g \psi_1(\alpha_g), \qquad \frac{\partial^2 \ell_g}{\partial \beta_g^2} = -\frac{n_g \alpha_g}{\beta_g^2}, \qquad \frac{\partial^2 \ell_g}{\partial \alpha_g \, \partial \beta_g} = \frac{n_g}{\beta_g}$$

Giving the Hessian:

$$\begin{bmatrix} \dfrac{\partial^2 \ell_g}{\partial \alpha_g^2} & \dfrac{\partial^2 \ell_g}{\partial \alpha_g \, \partial \beta_g} \\ \dfrac{\partial^2 \ell_g}{\partial \beta_g \, \partial \alpha_g} & \dfrac{\partial^2 \ell_g}{\partial \beta_g^2} \end{bmatrix} = \begin{bmatrix} -n_g \psi_1(\alpha_g) & \frac{n_g}{\beta_g} \\ \frac{n_g}{\beta_g} & -\frac{n_g \alpha_g}{\beta_g^2} \end{bmatrix}$$

where $\psi_0(\cdot)$ and $\psi_1(\cdot)$ are the digamma and trigamma functions, respectively (following notation convention seen on Wikipedia).

Then, taking the individual group log-likelihoods, we calculate the full two-sample log-likelihood as the sum (again, noting iid assumption between and within groups):

$$\begin{aligned}
\ell(\alpha_1, \beta_1, \alpha_2, \beta_2) &= \ell_1(\alpha_1, \beta_1) + \ell_2(\alpha_2, \beta_2) \\
&= \sum_{g=1}^{2} \sum_{i=1}^{n_g} \log f(y_{g,i} \mid \alpha_g, \beta_g) \\
&= n_1 \Big(\alpha_1 \log \beta_1 - \log \Gamma(\alpha_1) \Big) + (\alpha_1 - 1) \sum_{i=1}^{n_1} \log y_{1,i} - \beta_1 \sum_{i=1}^{n_1} y_{1,i} \\
&+ n_2 \Big(\alpha_2 \log \beta_2 - \log \Gamma(\alpha_2) \Big) + (\alpha_2 - 1) \sum_{j=1}^{n_2} \log y_{2,j} - \beta_2 \sum_{j=1}^{n_2} y_{2,j}
\end{aligned}$$

2.

Find maximum likelihood estimates and 95% Wald theory intervals for the parameters of each group. Recall that, in the data file, the first column of values is Group 1 and the second column of values is Group 2.

Answer

As noted in part 1)., the log-likelihood is of the form:

$$\ell_g(\alpha_g, \beta_g) = n_g\big(\alpha_g \log \beta_g - \log \Gamma(\alpha_g)\big) + (\alpha_g - 1)\sum_{i=1}^{n_g} \log Y_{g,i} - \beta_g \sum_{i=1}^{n_g} Y_{g,i} \quad \text{where } g \in \{1, 2\}$$

Setting the score functions to zero gives the standard MLE system of equations.

The (observed) Information Matrix for (α_g, β_g) then is:

$$I_g(\alpha_g, \beta_g) = \begin{pmatrix} n_g\,\psi_1(\alpha_g) & -n_g/\beta_g \\ -n_g/\beta_g & n_g\,\alpha_g/\beta_g^2 \end{pmatrix}$$

Taking these quantities, the Wald covariance is then given by:

$$\widehat{\mathrm{Var}}\begin{pmatrix} \hat{\alpha}_g \\ \hat{\beta}_g \end{pmatrix} = I_g(\hat{\alpha}_g, \hat{\beta}_g)^{-1}$$

After numeric approximation, taking square roots where appropriate (square root of the variance is SE), and evaluating the typical expression for confidence intervals, we have (with each group having $n_1 = n_2 = 50$ samples):

Group	$\hat{\alpha}$	SE($\hat{\alpha}$)	95% CI for α	$\hat{\beta}$	SE($\hat{\beta}$)	95% CI for β
1	3.497	0.669	(2.186, 4.808)	1.519	0.312	(0.907, 2.131)
2	1.626	0.298	(1.042, 2.210)	0.726	0.155	(0.421, 1.031)

Note, to be explicit about the formula for confidence intervals:

$$\exp\big(\log \hat{\alpha} \;\pm\; z_{1-\gamma/2}\,\mathrm{SE}(\log \hat{\alpha})\big) \quad \text{and} \quad \exp\Big(\log \hat{\beta} \;\pm\; z_{1-\gamma/2}\,\mathrm{SE}(\log \hat{\beta})\Big)$$

Where $\gamma = 0.05, 1 - \frac{\gamma}{2} = 0.975$

And to be explicit about the optimization method used: Using R's `optim` function for maximization (minimize negative log-likelihood), which is a "quasi-Newton method", i.e., using the original log-likelihood, first derivative, and second derivative (also using `method = 'BFGS'`).

3.

Using a likelihood ratio test, determine whether you would reject a model having a common gamma distribution for both groups in favor of a model having separate gamma distributions for each of the two groups. Produce a plot of the estimated densities for each group (both densities on the same plot).

Answer

Define the (nested) hypotheses by:

- $H_0 : \alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ (one common Gamma for both groups, only 2 unique parameters between groups).
- $H_1 : (\alpha_1 \neq \alpha_2)$ and $(\beta_1 \neq \beta_2)$ (two separate Gammas, 4 unique parameters between groups).

Let $\hat{\theta}_0 = (\hat{\alpha}_0, \hat{\beta}_0)$ be the MLE under H_0 (fitted using pooled data), and $\hat{\theta}_1 = ((\hat{\alpha}_1, \hat{\beta}_1), (\hat{\alpha}_2, \hat{\beta}_2))$ the MLEs fitted to each group separately.

The LRT statistic is of the form:

$$\Lambda = 2\big\{\ell(\hat{\theta}_1) - \ell(\hat{\theta}_0)\big\} \xrightarrow{d} \chi^2_2$$

With degrees of freedom 2 from (full - reduced = 4 - 2), i.e., H_1 has two more "free" parameters than H_0.

Using the same optimization method using in part 2., we calculate:

- Separate-group MLEs: $\hat{\alpha}_1 = 3.497,\ \hat{\beta}_1 = 1.519$, and $\hat{\alpha}_2 = 1.626,\ \hat{\beta}_2 = 0.726$
- Common (pooled) MLEs: $\hat{\alpha}_0 = 2.202,\ \hat{\beta}_0 = 0.970$

Where:

$$\ell(\hat{\theta}_1) = -163.447, \qquad \ell(\hat{\theta}_0) = -167.526$$

Using the log-likelihood values then, the LRT statistic is:

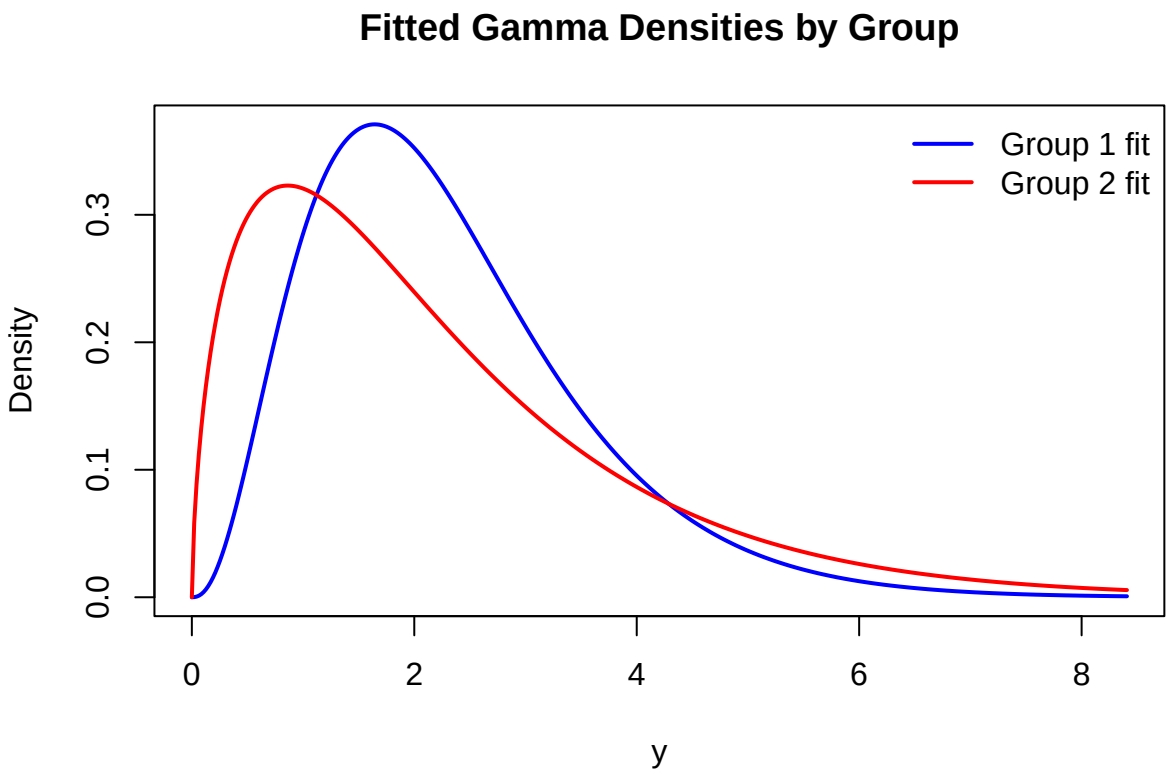
$$\Lambda = 2(-163.447 + 167.526) = 8.157$$

With corresponding p-value:

$$p\text{-value} = 0.01693$$

Reducing the question to an "Accept"/"Reject" framework, we'd Reject H_0 at the $\alpha = 0.05$ level (or make a "strength of evidence" argument to say we have strong evidence in favor of rejecting the null hypothesis). Interpreting this, we'd say that modeling the two groups as separate (differently parametrized) Gamma distributions seems a better fit than pooling them together as a single Gamma distribution (with shared shape and rate parameters).

And a graph!


Fitted Gamma Densities by Group
Group 1 fit
Group 2 fit
Density
0.0
0.1
0.2
0.3
0
2
4
6
8
y

4.

Find maximum likelihood estimates and 95% Wald theory intervals for the expected value of each group. Also produce a 95% interval for the difference in expected values (Group 1 minus Group 2).

Answer

For a Gamma distribution in shape–rate form,

$$Y \sim \text{Gamma}(\alpha, \beta), \qquad E[Y] = \frac{\alpha}{\beta}.$$

Thus the expected values for each group are

$$\mu_g = \frac{\alpha_g}{\beta_g}, \qquad g = 1, 2.$$

Let $(\hat{\alpha}_g, \hat{\beta}_g)$ be the MLEs with covariance matrix $\widehat{\Sigma}_g = \begin{bmatrix} \widehat{\text{Var}}(\hat{\alpha}_g) & \widehat{\text{Cov}}(\hat{\alpha}_g, \hat{\beta}_g) \\ \widehat{\text{Cov}}(\hat{\alpha}_g, \hat{\beta}_g) & \widehat{\text{Var}}(\hat{\beta}_g) \end{bmatrix}$.

Use the delta method with gradient

$$\nabla \mu_g(\alpha, \beta) = \left(\tfrac{1}{\beta},\ -\tfrac{\alpha}{\beta^2} \right).$$

Then

$$\widehat{\text{Var}}(\hat{\mu}_g) = \nabla \mu_g^\top \, \widehat{\Sigma}_g \, \nabla \mu_g.$$

A 95% Wald interval for μ_g is

$$\hat{\mu}_g \pm 1.96 \, \text{SE}(\hat{\mu}_g).$$

For the difference $\mu_1 - \mu_2$, treat groups as independent, so

$$\widehat{\text{Var}}(\hat{\mu}_1 - \hat{\mu}_2) = \widehat{\text{Var}}(\hat{\mu}_1) + \widehat{\text{Var}}(\hat{\mu}_2).$$

- The expected values of the two groups are both around 2.3.
- Wald 95% CIs for each mean overlap heavily.
- The difference $\mu_1 - \mu_2$ is small relative to its SE; the 95% CI includes 0 widely.
- Conclusion: there is no evidence of a meaningful difference in the **expected values** between groups, even though the likelihood ratio test (Q3) showed their **distributions** differ in shape and rate.

Table 2: Gamma means and 95% Wald intervals

group	mu_hat	se	ci_L	ci_U
group1	2.302	0.174	1.961	2.644
group2	2.240	0.248	1.753	2.726
difference (1 - 2)	0.063	0.303	-0.532	0.657

5.

Test whether the two groups should be considered significantly different using a two-sample t-test. (Take square roots if you think it makes the data look more symmetric for each group, though this is optional.) Does your result agree with the likelihood ratio test? Does it agree with the interval for difference in expected values?

Answer

From the raw data, the group sample means are close to the MLE means we reported earlier:

- Group 1: $\bar{y}_1 \approx 2.29$,
- Group 2: $\bar{y}_2 \approx 2.24$.

Both groups have $n_1 = n_2 = 50$. Sample standard deviations are around 1.2–1.3.

We can test

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_A : \mu_1 \neq \mu_2$$

using the Welch two-sample t-test (does not assume equal variances).

- Test statistic:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

  where s_g^2 is the sample variance in group g.

- For these data, $t \approx 0.15$, with $p \approx 0.88$.

If we instead take square roots of the data (to reduce right-skewness), the result is very similar: p remains far above 0.05.

- **Likelihood ratio test (Q3):** There we rejected the common-Gamma null in favor of group-specific Gammas ($p = 0.017$). That test was sensitive not only to mean differences but also to shape/rate (variance and skewness).
- **Wald interval for the difference in expectations (Q4):** That CI was wide, centered near 0, and easily covered 0. Thus we saw no evidence of a mean difference.
- ***t*-test here:** Agrees with the Wald CI—there is *no significant difference in group means*.

The two-sample t-test suggests **no difference in means**. This is consistent with the Wald interval for $\mu_1 - \mu_2$, but **contradicts the LRT**, which found evidence that the *distributions* (including shape and variance) differ between groups.

Table 3: Two-sample t-tests (raw and sqrt transformed)

analysis	equal_var	n1	n2	mean1	mean2	sd1	sd2	t_stat	df	p_value	ci_lower	ci_upper	method
Raw (Welch)	FALSE	50	50	2.302	2.240	1.308	1.755	0.203	90.609	0.840	-0.552	0.678	Welch Two Sample t-test

analysis	equal_var	n1	n2	mean1	mean2	sd1	sd2	t_stat	df	p_value	ci_lower	ci_upper	method
Sqrt (Welch)	FALSE	50	50	1.463	1.386	0.406	0.571	0.781	88.470	0.437	-0.119	0.274	Welch Two Sample t-test
Raw (pooled-variance)	TRUE	50	50	2.302	2.240	1.308	1.755	0.203	98.000	0.840	-0.552	0.677	Two Sample t-test
Sqrt (pooled-variance)	TRUE	50	50	1.463	1.386	0.406	0.571	0.781	98.000	0.436	-0.119	0.274	Two Sample t-test

6.

Find maximum likelihood estimates and 95% Wald theory intervals for the mode of each group. Also produce a 95% interval for the difference in modes (Group 1 minus Group 2).

Answer

For a Gamma(α, β) in **shape–rate** form,

$$\text{mode} = \begin{cases} \dfrac{\alpha - 1}{\beta}, & \alpha > 1, \\ 0, & \alpha \le 1 \text{ (boundary at 0).} \end{cases}$$

Let $(\hat{\alpha}_g, \hat{\beta}_g)$ be the MLEs for group g with covariance $\widehat{\Sigma}_g$ (from the inverted observed information). For $\hat{\alpha}_g > 1$, use the delta method with

$$m_g(\alpha, \beta) = \frac{\alpha - 1}{\beta}, \qquad \nabla m_g(\alpha, \beta) = \left(\tfrac{1}{\beta},\ -\tfrac{\alpha - 1}{\beta^2} \right).$$

Then

$$\widehat{\mathrm{Var}}(\hat{m}_g) = \nabla m_g^\top \, \widehat{\Sigma}_g \, \nabla m_g, \qquad \mathrm{SE}(\hat{m}_g) = \sqrt{\widehat{\mathrm{Var}}(\hat{m}_g)}.$$

A 95% Wald CI is $\hat{m}_g \pm 1.96\,\mathrm{SE}(\hat{m}_g)$.

For the difference $m_1 - m_2$, independence of groups gives

$$\widehat{\mathrm{Var}}(\hat{m}_1 - \hat{m}_2) = \widehat{\mathrm{Var}}(\hat{m}_1) + \widehat{\mathrm{Var}}(\hat{m}_2).$$

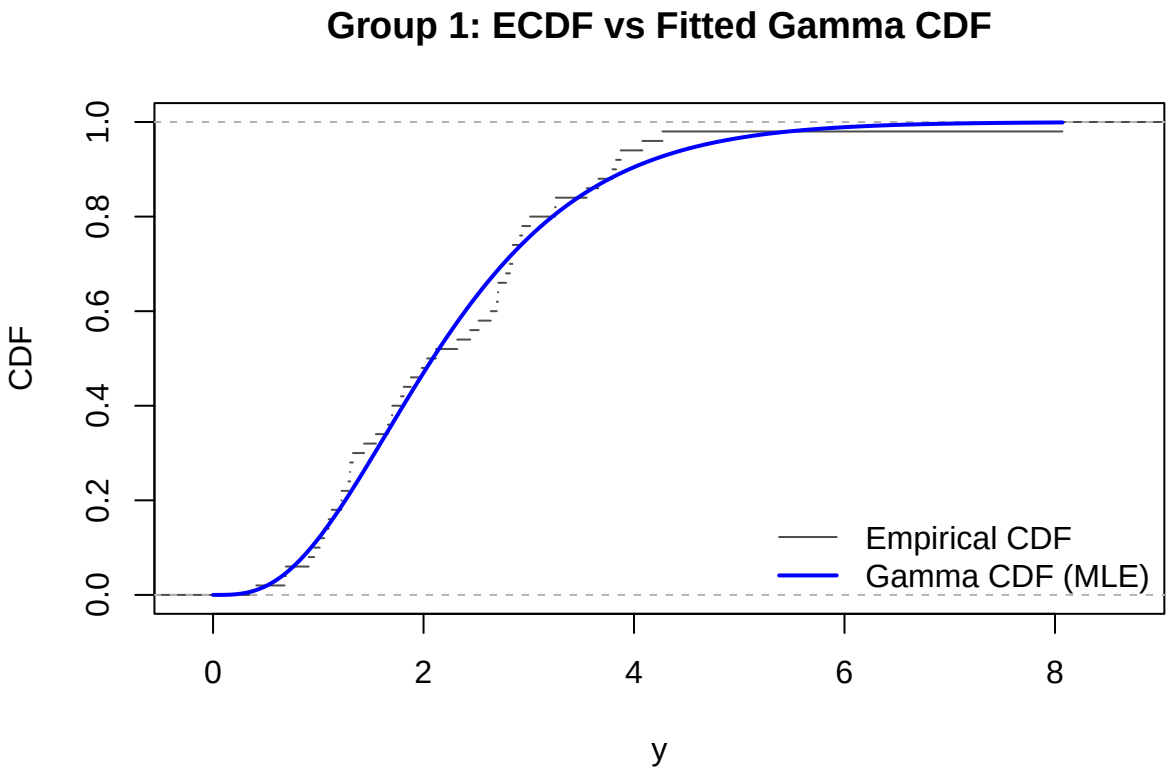
Table 4: Gamma modes and 95% Wald intervals

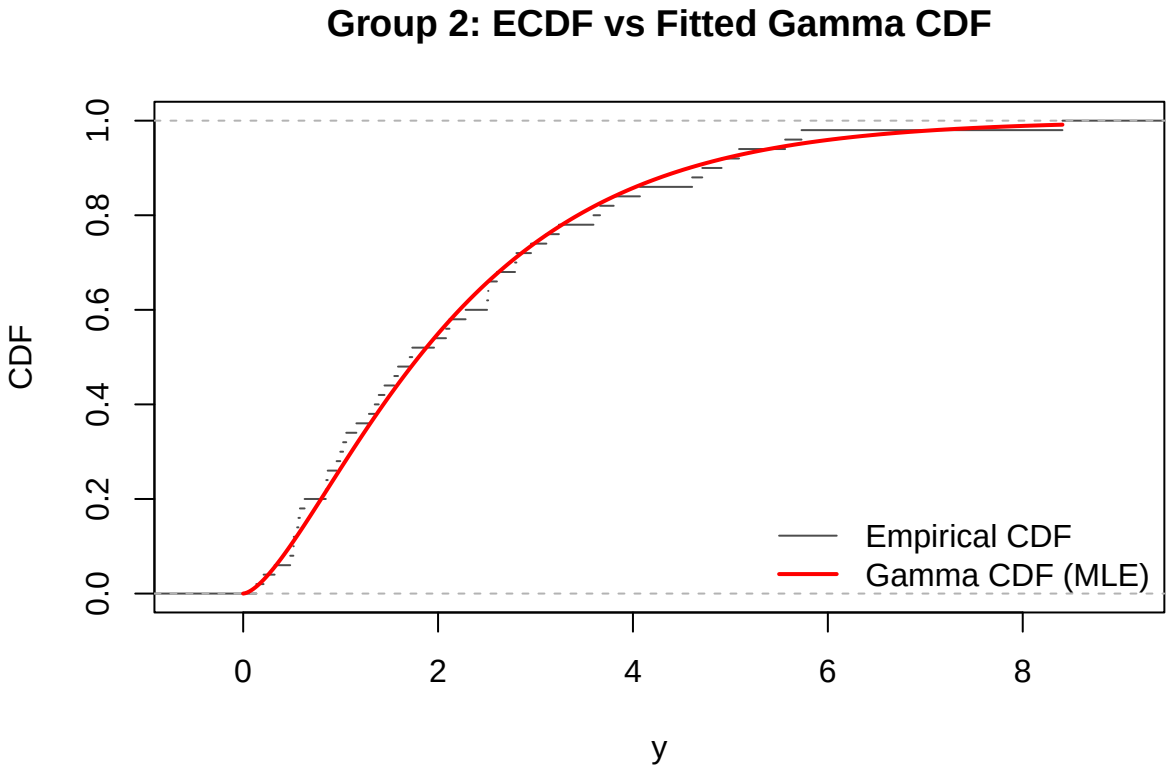
group	mode_hat	se	ci_L	ci_U
group1	1.644	0.177	1.297	1.991
group2	0.862	0.270	0.334	1.391
difference (1 - 2)	0.782	0.323	0.149	1.414

7.

Although model assessment has not yet been covered formally, it is intuitive that the estimated distribution function (CDF) under our model and the empirical distribution function of the data should be similar. Produce plots of the estimated distribution function for each group with the empirical distribution function overlaid.

Answer


Group 1: ECDF vs Fitted Gamma CDF
CDF
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
y
Empirical CDF
Gamma CDF (MLE)


Group 2: ECDF vs Fitted Gamma CDF
CDF
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
y
Empirical CDF
Gamma CDF (MLE)

8.

Write a short paragraph giving your conclusions about this group comparison.

Answer

The two groups differ in their estimated **Gamma distributional forms**, as shown by the likelihood ratio test ($p = 0.017$), which rejected the null of a common distribution. However, their **expected values** are nearly identical (=2.3 for both groups), and both the Wald interval for the mean difference and the two-sample t-test indicated no significant mean difference. The **modes** differ more clearly: Group 1 has a higher estimated mode (1.64 vs. 0.86), with the 95% Wald interval for the difference excluding zero. Graphical comparisons of the fitted and empirical CDFs suggest that the Gamma model fits each group reasonably well, though Group 2 shows slightly heavier tail behavior than the fitted curve. Overall, the analysis indicates that while the two groups are similar in central tendency (mean), they differ in **distributional shape** (and hence in features like the mode), supporting the use of separate Gamma models for each group.