

You are to write a report (using \TeX), due Tuesday November 9, entitled “Estimating the median of the one parameter Weibull distribution.” Be sure to write the report using good English (complete sentences, correct spelling, etc.). The report must have the title, your name, the course number, and the date at the top. The report must have the following sections:

1. The One Parameter Weibull Distribution

In this section you must state f , F , Q , expectation (μ), median (M), and variance (σ^2) corresponding to the Weibull pdf

$$f(x; \alpha) = \alpha x^{\alpha-1} e^{-x^\alpha}, \quad x > 0.$$

Make sure to write these quantities as functions of α , particularly the median, since the main thrust of the project is to estimate the median. Thus be sure to include the function g so that $M = g(\alpha)$.

You must have a half-page graph in the report that superimposes the Weibull pdf from the 0.001 to 0.999 quantile for $\alpha = 1, 2, 3, 4$ (the graph must be produced in Splus and imported into your report using `psfig`).

You must use the Splus function `text` to label the values of α and M on the graph for each of the four pdfs.

2. Estimation of the Median

This section must have the three subsections given below.

2.1 Two Estimators of the Median

Describe two estimators of M based on a random sample X_1, \dots, X_n from the Weibull distribution. The first is just the sample median, \hat{M} . The other estimator first estimates α by maximum likelihood estimation (MLE) and then substitutes this into the formula for M , thus giving an estimator called $\tilde{M} = g(\hat{\alpha})$ where $\hat{\alpha}$ is the MLE of α .

Note that in general the MLE of a parameter given data X_1, \dots, X_n is obtained by maximizing the likelihood function

$$L(\alpha) = \prod_{i=1}^n f(X_i; \alpha),$$

or equivalently

$$l(\alpha) = \log L(\alpha) = \sum_{i=1}^n \log f(X_i; \alpha).$$

Newton’s method can be used to find $\hat{\alpha}$ by finding the zero of $l'(\alpha)$.

2.2 Using Newton’s Method to Get $\hat{\alpha}$

There are two problems in blindly applying Newton’s method

$$\alpha_{i+1} = \alpha_i - \frac{l'(\alpha_i)}{l''(\alpha_i)}$$

to get $\hat{\alpha}$. Both problems are caused by the fact that the true value of α must be positive so we want to be sure that our starting value and the successive values of α_i are positive. Since $M = g(\alpha) = (\log(2))^{1/\alpha}$, it would be tempting to use $\alpha_0 = g^{-1}(\hat{M}) = \log(\log(2))/\log(\hat{M})$ as the starting value. The numerator of this is negative (it is -0.3665129), so for α_0 to be positive, we must have $\hat{M} < 1$. Unfortunately, even though M itself is always less than 1, it is possible for the median of a sample of size n to be greater than 1 (finding

the probability of this is a great STAT 610 problem), in which case α_0 would be negative which would cause all kinds of trouble in Newton's method. Thus we need a different starting value.

Here's one possible method. We have $Q(u) = (-\log(1-u))^{1/\alpha}$ which means that $\log(-\log(1-u)) = \alpha \log(Q(u))$ which from our data we can approximate by

$$\log(-\log(1-u_k)) = \alpha X_{(k)}, \quad k = 1, \dots, n,$$

where $u_k = (k - .5)/n$ and $X_{(k)}$ is the k th smallest element of the sample. Thus

$$\alpha_0 = \frac{\sum_{k=1}^n x_k y_k}{\sum_{k=1}^n x_k^2}$$

would be a possible starting value, where $x_k = \log(X_{(k)})$ and $y_k = \log(-\log(1-u_k))$. Note that in your Fortran routine, you will already have the X 's sorted to get the sample median. There is still no guarantee that this starting value will be positive, but I have tried it many times with good success.

The second problem arises because there is no guarantee in

$$\alpha_{i+1} = \alpha_i - l'(\alpha_i)/l''(\alpha_i),$$

that the α_i 's will remain positive. The standard way to handle this problem is to think of α as being e^β where β can be any real number (this is called a *reparametrization* of the problem). This forces α to be positive. We can get $\hat{\alpha} = e^{\hat{\beta}}$, where $\beta_0 = \log(\alpha_0)$ and

$$\beta_{i+1} = \beta_i - l'(\beta_i)/l''(\beta_i),$$

where $l(\beta)$ means $l(\alpha)$ with α written as e^β , and $l'(\beta)$ means the derivative of $l(\beta)$ with respect to β .

In this section of your report, you must give the formulas for l , l' , and l'' .

2.3 Comparing the Two Estimators of the Median

The standard method for comparing two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of some parameter θ is to first verify that they are unbiased (at least asymptotically as $n \rightarrow \infty$), that is, their average value over many samples is θ , and then to compare $\hat{\theta}_1$ and $\hat{\theta}_2$ in terms of how far they are from θ on the average for many samples, that is, compare their standard errors

$$se(\hat{\theta}_i) = \sqrt{\text{Var}(\hat{\theta}_i)} = \sqrt{E(\hat{\theta}_i - \theta)^2}.$$

In our present case, both \hat{M} and \tilde{M} are asymptotically unbiased. Unfortunately, $se(\hat{M})$ and $se(\tilde{M})$ are too difficult to determine analytically. The usual thing to do in a case where $se(\hat{\theta})$ cannot be evaluated is to estimate it by the root mean square error

$$rmse(\hat{\theta}) = \sqrt{\frac{1}{r} \sum_{i=1}^r (\hat{\theta}_i - \theta)^2},$$

where $\hat{\theta}_1, \dots, \hat{\theta}_r$ are the values of the estimator $\hat{\theta}$ for r replications.

To get an idea of $se(\hat{\theta})$ at least for large sample size n , one can usually find analytically a quantity called the asymptotic standard error of $\hat{\theta}$ denoted by $ase(\hat{\theta})$ satisfying

$$\lim_{n \rightarrow \infty} \frac{ase(\hat{\theta})}{se(\hat{\theta})} = 1.$$

You will see in STAT 611 and 613 that for the sample median from a sample of size n from a distribution having pdf f ,

$$\text{ase}(\hat{M}) = \frac{1}{\sqrt{4nf^2(M)}},$$

and for an MLE $\hat{\theta}$, $\text{ase}(\hat{\theta}) = \sqrt{\text{avar}(\hat{\theta})}$, where

$$\text{avar}(\hat{\theta}) = \frac{1}{-n\text{E}\left(\frac{\partial^2 \log f(X; \theta)}{\partial \theta^2}\right)}.$$

To get $\text{ase}(\tilde{M})$ we use the general result that

$$\text{avar}(g(\hat{\theta})) = (g'(\theta))^2 \text{avar}(\hat{\theta}).$$

In this section, you must briefly derive $\text{ase}(\hat{M})$ and $\text{ase}(\tilde{M})$.

3. A Simulation Study

In this section describe the following simulation (which must be done using Fortran that is called from Splus). For each of the four α 's given above and for each of $n = 10, 20, 50, 100$:

- Generate 1000 samples of size n (you must do this by writing an Splus function called `rweib` which generates random Weibull random numbers and then calling that routine to get 1000 n Weibulls which get passed to Fortran in an $n \times 1000$ matrix. For each sample, find \hat{M} and \tilde{M} . The calculating of the 1000 \hat{M} 's and \tilde{M} must be done in the Fortran routine. You can call the `qsor` subroutine in `6041.o` and must use Newton's method to get $\hat{\alpha}$ and then $\tilde{M} = g(\hat{\alpha})$.
 - Vectors of 1000 \hat{M} 's and \tilde{M} 's must be passed back from Fortran to Splus.

4. Results

This section must have a table presenting the following information for each α and n (the table should have α 's across the top and n 's down the side):

1. The true value of M .
2. $\text{ase}(\hat{M})$ and $\text{ase}(\tilde{M})$.
3. The average values of the 1000 \hat{M} 's and \tilde{M} 's.
4. $\text{rmse}(\hat{M})$ and $\text{rmse}(\tilde{M})$.

Also, for each α and n you must create an Splus graph which superimposes kernel density estimates of the distribution of \hat{M} and \tilde{M} . Make sure these estimates are not too smooth.

5. Summary

This section is a brief review of the first four sections and presents overall conclusions about the results. For example, you must discuss:

1. Does the average value of \hat{M} and \tilde{M} tend to get closer to M as n increases?
2. Which ase is smaller, that of \hat{M} or \tilde{M} ?
3. Do the rmse's agree well with the ase's?
4. Do the density estimates look Gaussian? Do the widths of the two density estimates agree with the relative values of the ase's?

5. Which estimator is better, \hat{M} or \tilde{M} ?

6. Computer Code

In this section give the code you used to do the project. You must use the `listing` macro in the `csbook.sty` file in the 604 directory (make sure to `include` this file at the top of your TeX file; you just put `\listing{filename}` to list a file).

Make sure to use comments in your code and to explain in this section what code does what.