Vol.2017-MPS-114 No.3
2017/7/17

# ワイブル分布による ハードディスクドライブの寿命時間の最尤推定について

小泉 大城[1,a)]

**概要**：本研究では，形状および尺度パラメータをもつワイブル分布のパラメータの最尤推定に基づくハードディスクドライブ (Hard Disk Drive, HDD) の寿命時間の推定問題を扱った．2013 年から約 4 年間，総計約 9 万件の完全データとランダム打ち切りデータを対象とし，2 つのパラメータの最尤推定値を数値計算により求めて考察したところ，次の (1)–(3) の 3 点が明らかになった．(1) ワイブル分布の形状パラメータのほとんどは 1 より大きな値となり，これは Increasing Failure Rate (IFR) を強く示唆する．(2) 完全データにより推定された平均故障時間 (Mean Time to Failure, MTTF) は，それぞれのハードディスクドライブの MTTF の経験的下界を意味する． (3) 完全データとランダム打ち切りデータのそれぞれから推定される MTTF の比により，(2) の経験的下界の効率を定量化することができる．

**キーワード**：信頼性工学，確率モデル，ワイブル分布，最尤推定，平均故障時間 (MTTF)，ハードディスクドライブ (HDD)

## 1. はじめに

### 1.1 研究背景

確率モデルによる製品や部品等の寿命時間の解析は，信頼性工学における重要な研究領域のひとつである [1], [2], [3]．その中でも形状パラメータと尺度パラメータを持つワイブル分布は頻繁に利用される確率密度関数の一種で，故障分布関数や信頼度関数，あるいは平均故障時間 (Mean Time to Failure, MTTF) 等，信頼性工学における各種指標の基盤となる [1], [2], [3]．信頼性工学では，ワイブル分布の形状パラメータが 1 未満の場合，その部品の故障率は時間と共に減少することを意味することから，Decreasing Failure Rate (DFR)[3] と分類する．もし形状パラメータが 1 であるとき，ワイブル分布は指数分布に帰着し，故障率は一定，すなわち Constant Failure Rate (CFR)[3] とみなされる．もし形状パラメータが 1 より大きい場合，故障率は時間とともに増加することから，Increasing Failure Rate (IFR)[3] となる．これら 3 つの場合を総合すると，部品の典型的な故障率は，使用開始当初は DFR であり，やがて CFR を経て IFR と変化し，いずれは寿命を迎える．横軸に時間，縦軸に故障率を取ったもとでのこの故障率の変化のプロットは，バスタブ型曲線 [3] として知られている．

一方，信頼性工学においては部品の寿命時間の観測データにも特徴がある．寿命時間の測定を目的とした信頼性試験において，もし，部品の使用を開始してから寿命を迎えるまでの正確な時間が観測できるならば，そのデータは完全データ [1], [2] と呼ばれる．完全データが観測できない場合，測定対象の部品は信頼性試験においてまだ稼働していることになるが，様々なコストの観点から寿命時間の測定を中止することがある．この時点での計測時間を打ち切りデータ [1], [2] と呼び，信頼性試験の打ち切りがランダムである場合のデータをランダム打ち切りデータ [1], [2] と呼ぶ．ランダム打ち切りデータは部品の寿命時間の下界の一種であるが，この下界のデータも活用することで，寿命時間のより精確な推定を実現できる可能性がある．このような観点から，ランダム打ち切りデータを含めた場合のワイブル分布の形状および尺度パラメータの最尤推定法 [4] がすでに提案されており，近年この手法をロケットエンジンの寿命時間推定に応用した例 [5] の報告がある．

### 1.2 研究目的

寿命時間を測定する信頼度試験においては，完全データやランダム打ち切りデータのデータ数の小さい場合がしばしば発生する．信頼性工学の分野では，このような状況下でも寿命時間の推定精度をなるべく落とさないために，確

[1] 小樽商科大学商学部社会情報学科
Otaru University of Commerce, Faculty of Commerce, Department of Information and Management Science, 3–5–1, Midori, Otaru-city, Hokkaido, 047–8501, Japan
[a)] dkoizumi@m.ieice.org

率紙や確率プロット，あるいはハザードプロットといった視覚的な方法論が古くから用いられてきた [2]．しかしこうした方法論において，推定法の基準は必ずしも明確ではない．

一方，近年の情報技術，特にネットワークサーバ領域においては，クラウド技術の普及に伴い，ハードディスクドライブ (Hard Disk Drive, HDD) の需要が劇的に増えている．このことから，HDD の寿命時間の計測が比較的大規模かつ低コストで実現可能になっている．こうした観点から，HDD の寿命時間を解析した研究 [6], [7] がいくつか報告されている．そのうちの一部では，ワイブル分布やガンマ分布が HDD の寿命時間として比較的適切で，指数分布や対数正規分布などの確率密度関数はカイ二乗検定により棄却されたことが報告されている [7]．しかし，ワイブル分布が適切であるという根拠は依然として確率プロットによる手法に留まっている [7]．

そこで本研究では，2013 年から約 4 年間，のべ約 9 万件の HDD の寿命時間の完全データとランダム打ち切りデータ [8] を対象とした．ただし，分析にあたっては外れ値は想定せず，また HDD の詳細な稼働環境 (温度，振動，ディスクアクセスエラー頻度等) も考慮していない．一連のデータをもとに 2 パラメータワイブル分布の最尤推定を行い，さらに平均故障時間 (MTTF) を推定して考察を行った．その結果，少なくとも以下の (1)–(3) の 3 点が明らかとなった．(1) 形状パラメータの最尤推定量のほとんどは 1 より大きくなり，これは IFR を強く示唆する．(2) 完全データから推定された MTTF は，それぞれの HDD の経験的下界を意味する．(3) 完全データとランダム打ち切りデータのそれぞれから推定される MTTF の比により，(2) の経験的下界の効率を定量化することが可能となる．

## 2. 準備 [1], [2]

いま，$T \geq 0$ を所望の部品の寿命時間とし，$T$ の実現値 $t \geq 0$ を連続型確率変数とする．このとき，$m, \lambda > 0$ をそれぞれ形状，尺度パラメータとする 2 パラメータワイブル分布 $f(t)$，ワイブル累積分布関数 $F(t)$，信頼整関数 $R(t)$，平均故障時間 $E(t)$ をそれぞれ以下で定義する．

**定義 2.1 (2 パラメータワイブル分布)**

$$f(t) = \frac{m}{\lambda}\left(\frac{t}{\lambda}\right)^{m-1} \exp\left[-\left(\frac{t}{\lambda}\right)^{m}\right]. \tag{1}$$

**定義 2.2 (ワイブル累積分布関数)**

$$F(t) = \int_0^t f(x)dx = 1 - \exp\left[-\left(\frac{t}{\lambda}\right)^{m}\right]. \tag{2}$$

**定義 2.3 (信頼度関数)**

$$R(t) = \int_t^\infty f(x)dx = \exp\left[-\left(\frac{t}{\lambda}\right)^{m}\right]. \tag{3}$$

**定義 2.4 (平均故障時間 (MTTF))**

$$E(t) = \int_0^\infty x\, f(x)dx = \lambda\Gamma\left(1 + \frac{1}{m}\right), \tag{4}$$

ただし，$\Gamma(\cdot)$ はガンマ関数である．

## 3. ワイブル分布のパラメータの最尤推定 [4]

いま，$t_1, t_2, \ldots, t_n$ を寿命時間に関する長さ $n$ の完全データ，$u_{n+1}, u_{n+2}, \ldots, u_{n+r}$ を長さ $r$ のランダム打ち切りデータとする．ここで，$t_{n+1}, t_{n+2}, \ldots, t_{n+r}$ は観測されていない完全データのことで，ランダム打ち切りデータはこの完全データの下界を表すので $j = 1, 2, \ldots, r$ として，$u_{n+j} < t_{n+j}$ となる．以降では簡単のため，$u_{n+j}$ のことを単に $u_j$ と表記する．長さ $(n+r)$ のランダム打ち切りデータを含む系列を観測したもとで，定義 2.1 の確率密度関数に対応する尤度関数 $L_2(m, \lambda)$，形状および尺度等パラメータの最尤推定量 $\hat{m}_2, \hat{\lambda}_2$ をそれぞれ以下で定義する．

**定義 3.1 (尤度関数)**

$$\begin{aligned} L_2(m,\lambda) &= \frac{(n+r)!}{r!}\prod_{i=1}^{n} f(t_i)\prod_{j=1}^{r} R(u_j) \\ &= \frac{(n+r)!}{r!}\prod_{i=1}^{n}\frac{m}{\lambda}\left(\frac{t_i}{\lambda}\right)^{m-1}\exp\left[-\left(\frac{t_i}{\lambda}\right)^{m}\right] \\ &\quad \prod_{j=1}^{r}\exp\left[-\left(\frac{u_j}{\lambda}\right)^{m}\right]. \end{aligned} \tag{5}$$

**定義 3.2 (最尤推定量)**

$$\begin{cases} \hat{m}_2 &= \arg\max_m L_2(m,\lambda)\,; \\ \hat{\lambda}_2 &= \arg\max_\lambda L_2(m,\lambda)\,. \end{cases} \tag{6}$$

このとき，最尤推定量の必要条件は，式 (7) のようになる [4].

**補題 3.1 (最尤推定量の必要条件)**

$$\begin{cases} \left[\frac{\sum_{i=1}^{n}(t_i)^{\hat{m}_2}\ln t_i + \sum_{j=1}^{r}(u_j)^{\hat{m}_2}\ln u_j}{\sum_{i=1}^{n}(t_i)^{\hat{m}_2} + \sum_{j=1}^{r}(u_j)^{\hat{m}_2}} - \frac{1}{\hat{m}_2}\right] \\ \qquad\qquad -\frac{1}{n}\sum_{i=1}^{n}\ln t_i = 0\,; \\ \hat{\lambda}_2 = \left[\frac{\sum_{i=1}^{n}(t_i)^{\hat{m}_2} + \sum_{j=1}^{r}(u_j)^{\hat{m}_2}}{n}\right]^{\frac{1}{\hat{m}_2}}. \end{cases} \tag{7}$$

## 4. ハードディスクドライブ (HDD) の寿命データ解析

### 4.1 データ仕様

解析対象の HDD の寿命データとして，Backblaze 社が公開しているものを利用した [8]．このデータは，2013 年から約 4 年間 (1,362 日) の完全データの総数 $n = 5,789$，ランダム打ち切りデータの総数 $r = 85,044$，データ総数 $n + r = 90,833$ であり，計 6 社の HDD 製造メーカーの

HDD モデルが対象である．データ解析にあたり考慮したのは，寿命時間の完全データ，ランダム打ち切りデータおよび HDD の状態（正常またはエラー）のみで，その他の稼働環境（温度，振動，ディスクアクセスエラー頻度等）は考慮していない．また，外れ値も一切想定していない．

### 4.2 ランダム打ち切りデータの解析結果

表 1 に主な HDD モデル毎の寿命時間のランダム打ち切りデータの解析結果を示す．表 1 では，$n$ は完全データ数，$r$ はランダム打ち切りデータ数，$\hat{m}_2$ は数値計算により求めた形状パラメータの最尤推定量，$\lambda_2$ は尺度パラメータの最尤推定量で，最右列の $\hat{E}_2(t)$ は，これらの最尤推定量を用いて，式 (8)，

$$\hat{E}_2(t) = \hat{\lambda}_2 \Gamma\left(1 + \frac{1}{\hat{m}_2}\right), \tag{8}$$

により計算した MTTF の推定値 (単位は `[hrs]`) である．

## 5. 考察

### 5.1 形状パラメータの最尤推定量

表 1 より，総計 10 の HDD 製品モデルのうち，8 モデルで形状パラメータの最尤推定量 $m_1$ は 1 より大きな値となった．これは，故障率が Increasing Failure Rate (IFR)[3] であることを強く示唆する．一方で，ある製造メーカーでは $\hat{m} = 0.55$[9] とみなしている例や，$0.71 \leq \hat{m} \leq 0.76$ [7] という結果を得た例があり，これらはいずれも Decreasing Failure Rate (DFR) である．

### 5.2 平均故障時間 (MTTF) の推定値

完全データのみによる 2 パラメータワイブル分布の形状パラメータ，尺度パラメータの最尤推定量をそれぞれ $\hat{m}_1, \hat{\lambda}_1$ とする．このとき，HDD モデル ST3000DM001 については，$n = 1{,}720$ であり，$\hat{m}_1 = 5.376, \hat{\lambda}_1 = 18{,}891$ となった．これらの最尤推定量によりこの HDD モデルの MTTF$\hat{E}_1(t)$ を計算すると，

$$\begin{aligned}\hat{E}_1(t) &= \hat{\lambda}_1 \Gamma\left(1 + \frac{1}{\hat{m}_1}\right) \\ &= 18{,}891 \cdot \Gamma\left(1 + \frac{1}{5.376}\right) = 17{,}417\texttt{[hrs]}, \end{aligned} \tag{9}$$

となり，表 1 のランダム打ち切りデータによる推定値 $\hat{E}_2(t) = 19{,}787$`[hrs]` よりわずかに小さな値となった．この HDD モデルの推定結果をさらに詳細に考察するため，図 1 のような可視化を行った．図 1 には，横軸に時間を取ったうえで寿命時間の完全データ（斜線棒）とランダム打ち切りデータ（灰色の棒）の相対ヒストグラムを作成し，さらにそれぞれのデータから最尤推定されたワイブル分布の確率密度関数も合わせてプロットした．ここで，破線は完全データ，実線はランダム打ち切りデータから最尤推定された形状・尺度パラメータによる 2 パラメータワイブル分布を意味し，破線の確率密度関数の期待値 (MTTF) が $\hat{E}_1(t) = 17{,}417$，実線の期待値が $\hat{E}_2(t) = 19{,}787$ である．図 1 より，斜線棒の完全データの山よりも横軸の左側に灰色棒のランダム打ち切りデータの山が観測されたことにより，結果としてワイブル分布の最尤推定プロットは破線から実線，すなわち横軸の右側にシフトしたことがわかる．つまり，この HDD モデルについては，完全データから最尤推定された破線のワイブル分布のプロットは，MTTF の経験的下界を表していることがわかる．

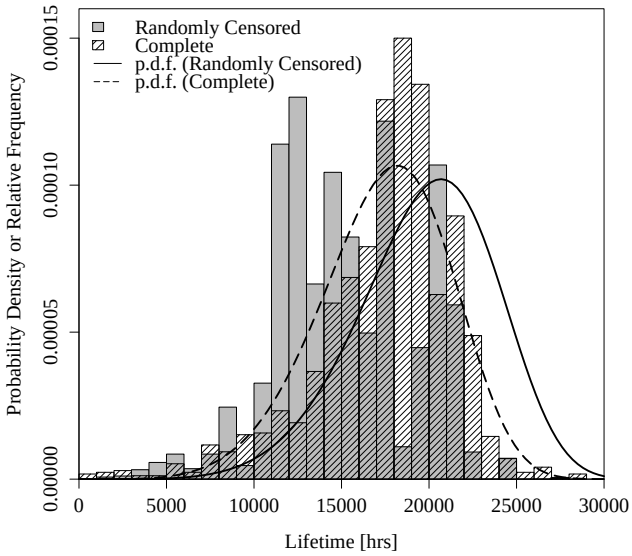


**図 1** HDD モデル ST3000DM001 の寿命時間の相対ヒストグラムおよび最尤推定されたワイブル分布の確率密度関数 ($n = 1{,}720, r = 2{,}817$)

### 5.3 MTTF の経験的下界の効率

前節の結果より，HDD モデル ST3000DM001 について，$\hat{E}_2(t)/\hat{E}_1(t)$ で表される MTTF 推定値の比を計算すると，

$$\frac{\hat{E}_2(t)}{\hat{E}_1(t)} = \frac{19{,}787}{17{,}417} = 1.14, \tag{10}$$

となる．前節で考察したように，$\hat{E}_1(t) = 17{,}417$ はこの HDD モデルの MTTF の経験的下界を表しているが，式 (10) の値は 1 よりわずかに大きい値になっている．いま，もしこの HDD モデルについて完全データから推定された $\hat{E}_3(t) = 5{,}000$ であったとしよう．このとき，式 (10) の値は，$19{,}787/5{,}000 = 3.96$ となる．$\hat{E}_1(t)$ も $\hat{E}_3(t)$ のいずれも $E_2(t)$ より小さい値のため，MTTF の経験的下界であることには変わりない．しかし下界の効率については，より値の大きい（すなわち MTTF の比が 1 に近い）$\hat{E}_1(t) = 17{,}417$ のほうが $\hat{E}_3(t)$ に比べると相対的に優れ

**表 1** ランダム打ち切りデータを対象とした推定結果

| Models | $n+r$ | $n$ | $r$ | $\hat{m}_2$ | $\hat{\lambda}_2$ | $\hat{E}_2(t)$ |
|---|---|---|---|---|---|---|
| ST4000DM000 | 36,532 | 1,807 | 34,725 | 1.122 | 227,663 | 218,260 |
| ST3000DM001 | 4,537 | 1,720 | 2,817 | 5.833 | 21,362 | 19,787 |
| ST31500541AS | 2,087 | 397 | 1,690 | 3.583 | 68,530 | 61,737 |
| Hitachi HDS722020ALA330 | 4,765 | 229 | 4,536 | 2.954 | 129,479 | 115,546 |
| ST31500341AS | 662 | 216 | 446 | 3.173 | 56,735 | 50,795 |
| WDC WD30EFRX | 1,289 | 162 | 1,127 | 0.638 | 490,229 | 683,660 |
| Hitachi HDS5C3030ALA630 | 4,661 | 134 | 4,527 | 1.460 | 458,109 | 414,993 |
| HGST HMS5C4040ALE640 | 7,162 | 103 | 7,059 | 0.666 | 12,288,903 | 16,366,251 |
| ST1500DL003 | 106 | 90 | 16 | 1.511 | 11,873 | 10,709 |
| ST320LT007 | 98 | 88 | 10 | 4.210 | 27,296 | 24,814 |

ていることがわかる．

そこですべての HDD モデルについて $\hat{E}_2(t)/\hat{E}_1(t)$ の MTTF 推定値の比を計算した結果が表 2 である．表 2 より，すべての HDD モデルにおいて式 (10) の値は 1 より大きな値になっている．これはすなわち，完全データによる MTTF の推定値 $\hat{E}_1(t)$ はランダム打ち切りデータによる MTTF の推定値 $\hat{E}_2(t)$ より小さな値になっていて，前者が各 HDD モデルの経験的下界を表していることがわかる．また，HDD モデル ST3000DM001, ST1500DL003, ST320LT007 などは，式 (10) の値が 1 に極めて近い値となっており，他の HDD モデルに比べると $\hat{E}_1(t)$ による経験的下界の効率が大きいことがわかる．逆に HGST HMS5C4040ALE640 については，式 (10) の値が極めて大きくなっており，$\hat{E}_1(t)$ による経験的下界の効率は小さいことがわかる．以上の考察から，完全データとランダム打ち切りデータのそれぞれから推定される MTTF の比により，前者の推定値による経験的下界の効率を定量化可能なことが明らかになった．

**表 2** 完全データとランダム打ち切りデータの MTTF 推定値の比

| Models | $\hat{E}_2(t)/\hat{E}_1(t)$ |
|---|---|
| ST4000DM000 | 19.25 |
| ST3000DM001 | 1.14 |
| ST31500541AS | 1.72 |
| Hitachi HDS722020ALA330 | 3.29 |
| ST31500341AS | 1.46 |
| WDC WD30EFRX | 83.10 |
| Hitachi HDS5C3030ALA630 | 16.33 |
| HGST HMS5C4040ALE640 | 2071.28 |
| ST1500DL003 | 1.16 |
| ST320LT007 | 1.03 |

## 6. 結論

本研究では，のべ約 9 万件の HDD の寿命時間の完全データとランダム打ち切りデータを対象として 2 パラメータワイブル分布の最尤推定を行った．さらに計算された形状および尺度パラメータの最尤推定量をもとに，平均故障時間 (MTTF) を推定した．その結果，以下の (1)–(3) の 3 点が明らかとなった．(1) 形状パラメータの最尤推定量のほとんどは 1 より大きくなり，これは IFR を強く示唆する．(2) 完全データから推定された MTTF は，それぞれの HDD の経験的下界を意味する．(3) 完全データとランダム打ち切りデータのそれぞれから推定される MTTF の比により，(2) の経験的下界の効率を定量化することが可能である．

## 参考文献

[1] Jerald F. Lawless, *Statistical Models and Methods for Lifetime Data,* John Wiley & Sons, 1982.
[2] Wayne B. Nelson 著, 柴田 義貞, 藤野 和建, 鎌倉 稔成 訳,「寿命データの解析」, 日科技連, 1988 年.
[3] Michael S. Hamada, Alyson G. Wilson, C. Shane Reese, and Harry F. Martz, *Bayesian Reliability,* Springer, 2008.
[4] A. Clifford Cohen, “Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples,” Technometrics, vol.7, no.4, pp. 579–588, Nov. 1965.
[5] Haibo Li, Zhengping Zhang, Yanping Hu, and Deqiang Zheng, “Maximum likelihood estimation of Weibull distribution based on random censored data and its application,” Proceeding of the 8th International Conference on Reliability, Maintainability and Safety, pp.302–304, Jul. 2009.
[6] Eduardo Pinheiro, Wolf-Dietrich Weber and Luiz Andre Barroso, “Failure Trends in a Large Disk Drive Population,” Proceeding of the 5th USENIX Conference on File and Storage Technologies (FAST’07), pp. 17–29, Feb. 2007.
[7] Bianca Schroeder and Garth A. Gibson, “Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?,” Proceeding of the 5th USENIX Conference on File and Storage Technologies (FAST’07), Article no. 1, Feb. 2007.
[8] Backblaze, Hard Drive Data and Stats [Online]. Available: `https://www.backblaze.com/b2/hard-drive-test-data.html`
[9] Gerry Cole, “Estimating Drive Reliability in Desktop Computers and Consumer Electoronics Systems,” Technology Paper from Seagate, TP–338.1, Nov. 2000.