対数正規分布（Lognormal Distribution）のあてはめについて

日野寛三

1．はじめに

最近，R.F.Mould:Introductory Medical Statistics 2nd Edition（Adam Hilger, 1989）を訳出する機会をもった。本書は，チェルノブイリ事故,癌,エイズなど，こんにち社会的に最も注目を浴びている諸問題をはじめ,統計学上古くからよく知られているデータなどをとりあげて，医学統計学の基礎やその適用を論じたものである。内容は多彩で，医学関係の学術雑誌からもいくつか賛辞がよせられている。

著者は，英国の国際的にも著名な医学物理学者で，病院勤務の経験も長く，癌統計に関して多くの著書や論文がある。世界保健機関（ＷＨＯ）や国際原子力機関（ＩＡＥＡ）のコンサルタントとしても活躍し，チェルノブイリ事故後の英国派遣団の一員として訪れた現地の詳細な報告書：Chernobyl —— The Real Story には邦訳もある。*)

それにもかかわらず，著者多忙のため推敲や校正が十分でなかったのか，ケアレスミスその他の不備が目立って多く，正直に言って翻訳は難渋した。単純なミスは一々断らずに，あるいは，著者が引用した論文に直接目を通したうえで訂正したが，たんにケアレスとは言い難いいくつかの本質的な疑問点については著者にそのことを婉曲に指摘した。著者からは，半年近くを経て，一つを除き，指摘の通りであるとの回答が寄せられたので，必要な訂正を行うことができた。その一つ“対数正規曲線の計算”については訂正への同意が得られなかったので，不本意ではあったが，原文のまま訳出せざるを得なかった。しかし，その部分の計算方法には明らかに誤りがあり，これが二か所にわたっている。このような事情から，やむを得ず「訳者あとがき」のなかで訳者としての見解を述べておいたが，できれば，なんらかの形でそ

の部分を訂正し結果を明らかにしておきたい。

2．対数正規分布について

和洋を問わず手近かな統計学の書で，対数正規分布について詳しくとりあげたものは，筆者の知る限り皆無に近いので，ここで必要なものについて簡単に述べておこう。

正規曲線は，周知のように，

$$f(x)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],\quad -\infty<x<\infty$$

で定義されている。ここに，μは平均，σは標準偏差である。ここで，確率変数に変換 $x=\log y$ を行うと，$\log y$ が平均μ，標準偏差σの正規分布に従うことになる。このとき，新しい確率変数 y は対数正規分布に従うという。その確率密度関数を求めてみよう。

$$\int_{-\infty}^{\infty} f(x)\,dx=1$$

であるから，$x=\log y$，$-\infty<x<\infty$，$0<y<\infty$ とおけば

$$\int_0^{\infty}\frac{1}{y}f(x)\,dy=1.$$

したがって，yの確率密度関数を $F(y)$ とすれば，これは次のように表される。

$$F(y)=\begin{cases} y>0 \text{ のとき} & \dfrac{1}{y\sqrt{2\pi}\,\sigma}\exp\left[-\dfrac{(\log y-\mu)^2}{2\sigma^2}\right], \\ y\leqq 0 \text{ のとき} & 0. \end{cases}$$

こんご上記の関数を次のように書き表すことにする：

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left[-\frac{(\log x-\mu)^2}{2\sigma^2}\right].$$

なお，変換に用いた対数はeを底とする自然対数である。このことを明確にしておく必要がある。こんごも底eは省略する。

ここで，確率変数xの平均，最頻値および中央値を求めておこう。

（1）平均は定義により

$$\int_{-\infty}^{\infty} x f(x)\,dx = \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\log x-\mu)^2}{2\sigma^2}\right] dx$$

で与えられる。

$$z = \frac{\log x-\mu}{\sqrt{2}\sigma}\ ,\ 0<x<\infty,\ -\infty<z<\infty \quad \text{とおくと}$$

$$\int_{-\infty}^{\infty} x f(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} x \cdot \exp(-z^2)\,dz$$

$$= \frac{\exp(\mu)}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-z^2+\sqrt{2}\sigma z)\,dz.$$

$$t = z - \frac{\sigma}{\sqrt{2}},\ -\infty<z<\infty,\quad -\infty<t<\infty \quad \text{とおけば}$$

$$\int_{-\infty}^{\infty} x f(x)\,dx = \frac{\exp(\mu+\sigma^2/2)}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-t^2)\,dt,$$

$$\int_{-\infty}^{\infty} \exp(-t^2)\,dt = \sqrt{\pi} \quad \text{を考慮して，けっきょく}$$

$$\int_{-\infty}^{\infty} x f(x)dx = \exp\left(\mu+\frac{\sigma^2}{2}\right)$$

が得られる。

（2）最頻値は$f(x)$が最大値をとるxの値であるから，$f'(x)=0$ より

$$-\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(\log x-\mu)^2}{2\sigma^2}\right]\left(\frac{1}{x^2}+\frac{\log x-\mu}{x^2\sigma^2}\right)=0$$

をxについて解き　　$\dfrac{\log x-\mu}{\sigma^2}=-1,$

すなわち，　　$x=\exp(\mu-\sigma^2).$

（3）中央値は

$$\int_{-\infty}^{x} f(t)\,dt=\frac{1}{2}$$

を満たすxの値として求められる。したがって

$$\int_0^x \frac{1}{t\sqrt{2\pi}\sigma}\exp\left[-\frac{(\log t-\mu^2)}{2\sigma^2}\right]dt=\frac{1}{2}$$

において，（1）と同様に置換

$$z=\frac{\log t-\mu}{\sqrt{2}\sigma},\quad 0<t<x,\quad -\infty<z<\frac{\log x-\mu}{\sqrt{2}\sigma}$$

を行うと，　$t=\exp(\sqrt{2}\sigma z+\mu)$，$dt=\sqrt{2}\sigma t\,dz.$

ゆえに，　$z_0=(\log x-\mu)/\sqrt{2}\sigma$とおけば，

$$\frac{1}{\sqrt{\pi}}\int_{-\infty}^{z_0}\exp(-z^2)\,dz=\frac{1}{2},$$

$$\int_{-\infty}^{z_0}\exp(-z^2)\,dz=\frac{\sqrt{\pi}}{2},$$

これから　　$z_0=(\log x-\mu)/\sqrt{2}\sigma=0,$

すなわち，　　$x=\exp(\mu).$

以上をまとめると，対数正規分布の平均値，最頻値および中央値はそれぞれ

$$\text{平均値} = \exp\left(\mu + \frac{\sigma^2}{2}\right),\ \text{最頻値} = \exp(\mu - \sigma^2),\ \text{中央値} = \exp(\mu)$$

で与えられる。

3．胃癌患者783人のデータへの対数正規分布のあてはめ

著者は，ウェストミンスター病院で，１９４５～１９７０の期間に治療をうけた　783人の患者の症状期間**) に関するデータを度数分布表で次のように与えている。

Table 3.6 Data requirements prior to a graphical demonstration of lognormality, see figure 3.8.

Symptom time range (months)	Frequency	Cumulative frequency	Percentage cumulative frequency
0–1	138	17.6	17.6
1.1–2	111	14.2	31.8
2.1–3	85	10.9	42.7
3.1–4	63	8.0	50.7
4.1–5	42	5.4	56.1
5.1–6	51	6.5	62.6
6.1–9	67	8.6	71.2
9.1–12	66	8.4	79.6
12.1–18	56	7.2	86.8
18.1–24	36	4.6	91.4
24.1–36	23	2.9	94.3
36.1–48	7	0.9	95.2
48.1–60	11	1.4	96.6
60.1 and above	27	3.4	100
	Total = 783	Total = 100 %	

この表3.6のデータを対数正規確率紙にプロットしたものが次の図3.8である。この図からもわかるように，データはほとんど直線上に並んでいる。このことから，表3.6のデータは対数正規分布で近似できると思われる。著者はこのグラフから，累積百分率50%および95%に対応する症状期間Tの値としてM＝4.1および39を読み取り，

$$\log_{10}4.1=0.6128,\quad \log_{10}39=1.5911,$$

したがって，

$$S=(1.5911-0.6128)/1.645=0.595$$

とした。なお，著者は原著を通じて母数と統計量（推定量）とを厳密には区別していないが，Tは上記の確率変数xに対応し，MおよびSはそれぞれ中央値 $\exp(\mu)$ および σ の推定値である。これらの推定値 M＝4.1 および S＝0.6 を用いてこのデータにあてはめる対数正規曲線として図3.7を与えた。そのための計算が表3.5である。

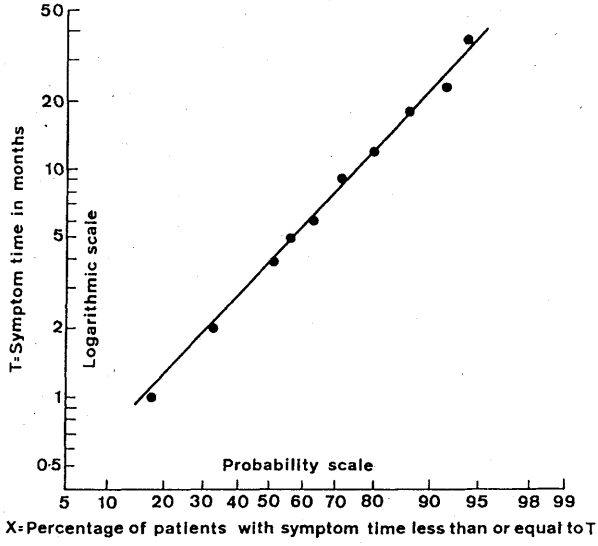

Figure 3.8 Graphical demonstration of lognormality using the data in table 3.6.

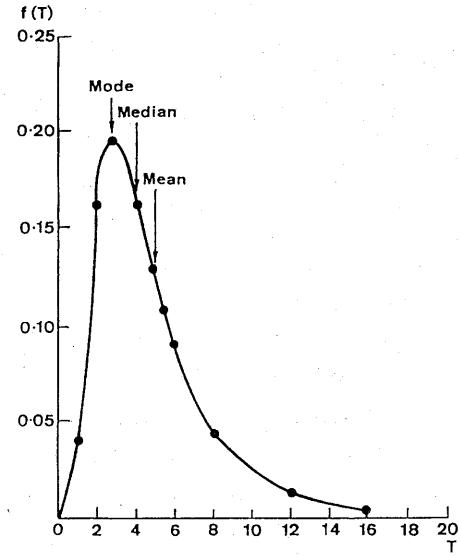

Figure 3.7 Lognormal curve with $M = 4.1$ and $S = 0.6$.

Table 3.5 Calculations required to be able to draw the lognormal curve with $M = 4.1$ and $S = 0.6$. The constants are: $2S^2 = 0.72$ and $1/S\sqrt{2\pi} = 0.665$. The mode, median and mean are indicated in the footnote.

T	$T/4.1$	$\log_e(T/M)$	$\log_e(T/M)^2/0.72 = A$	e^{-A}	$0.665/T$	$0.665\, e^{-A}/T = f(T)$
1.0	0.244	−1.411	2.765	0.063	0.665	0.042
2.0	0.488	−0.718	0.716	0.489	0.332	0.163
2.86†	0.698	−0.360	0.180	0.835	0.233	0.194
4.10‡	1.000	0	0	1.000	0.162	0.162
4.91§	1.198	0.180	0.180	0.045	0.956	0.135
5.5	1.342	0.294	0.120	0.887	0.121	0.107
6.0	1.463	0.381	0.201	0.818	0.111	0.091
8.0	1.951	0.668	0.621	0.538	0.083	0.045
12.0	2.927	1.074	1.602	0.202	0.055	0.011
16.0	3.902	1.362	2.575	0.076	0.042	0.003

†Mode.
‡Median.
§Mean.

さらに，これらの値を用いて

Tの平均値$=4.1 \cdot \exp(S^2/2)$
$=4.1 \cdot \exp(0.18)$
$=4.91,$

Tの中央値$=4.1,$

Tの最頻値$=4.1/\exp(S^2)$
$=4.1 \cdot \exp(0.36)$
$=2$

を求めた。これらは表3.5および図3.7にそれぞれ表示されている。

しかし，以上の計算で常用対数$\log_{10}$を用いたのは誤りである。対数正規分布の定義のあとで強調しておいたように，そこでの対数変換には自然対数 log を用いている。したがって，自然対数 log を用いて計算するべきであった。これが訳者としての指摘である。M＝4.1, S＝0.6 の対数正規曲線を計算するための表3.5では自然対数を用いているので，この計算とこれにもとずく図3.7のグラフそれ自体はたしかに間違いではない。しかし，この計算とグラフの作成とは，表3.6のデータに M＝4.1, S＝0.6 の対数正規分布があてはめられるとして行われたものであるから，S＝0.6 に誤りがあれば意味がない。ちなみに，図3.7からもわかるように，M＝4.1, S＝0.6 の対数正規曲線の最頻値は区間 2－4 にふくまれるが，表3.6のデータの最頻値は明らかに区間 0－1 にふくまれている。これだけでも誤りであることがわかろう。

そこで，数値の読み取りが妥当に行われたものとして，表3.6のデータにあてはめるための対数正規曲線を正しく求めてみよう

4．表3.6 のデータにあてはめられる対数正規曲線

まず，S を次のようにして求める。

$\log 4.1 = 1.4110,\quad \log 39 = 3.6636,\quad S = (3.6636 - 1.4110)/1.645 = 1.370$

これから 平均値，最頻値 および中央値 は次のようになる。

Tの平均値$=4.1 \cdot \exp(1.37^2/2) = 10.48,$
Tの最頻値$=4.1/\exp(1.37^2) = 0.63,$
Tの中央値$=4.1.$

さて，M＝4.1, S ＝1.37 の対数正規曲線のグラフをえがくための（表3.5

に代わる）計算を次に示そう。これを表1とし，手順は表3.5に対応させた。

表1　M＝4.1, S＝1.37 の対数正規曲線のグラフをえがくための計算

$2S^2=3.75$,　$1/S\sqrt{2\pi}=0.291$

T	T/4.1	log (T/M)	A = [log (T/M)]²/3.75	e^{-A}	0.291/T	f(T)
0.5	0.122	−2.104	1.180	0.307	0.582	0.179
0.63†	0.154	−1.873	0.936	0.392	0.462	0.181
1.0	0.244	−1.411	0.531	0.588	0.291	0.171
2.0	0.488	−0.718	0.137	0.872	0.146	0.127
3.0	0.732	−0.312	0.026	0.974	0.097	0.094
4.1‡	1	0	0	1	0.071	0.071
5.0	1.220	0.198	0.010	0.990	0.058	0.058
6.0	1.463	0.381	0.039	0.962	0.049	0.047
8.0	1.951	0.668	0.119	0.888	0.036	0.032
10.48§	2.561	0.938	0.235	0.791	0.028	0.022
12.0	2.927	1.074	0.308	0.735	0.024	0.018
16.0	3.902	1.361	0.494	0.610	0.018	0.011

† 最頻値
‡ 中央値
§ 平均値

上の計算にもとずく対数正規曲線のグラフが次の図1である。

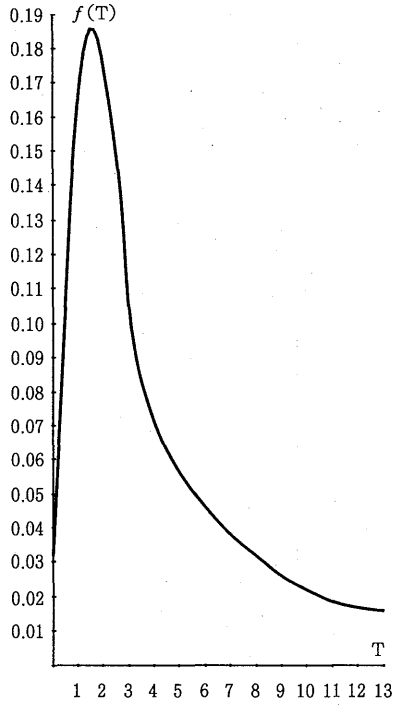

図1　M=4.1, S=1.37 の対数正規曲線

5．データへの対数正規分布の適合度検定

図3.8を見るかぎり点はほぼ直線上に並ぶので，著者は，当然与えられたデータは対数正規分布に従うものと考えたであろう。しかし，その適合度検定は行われていない。S=0.6としたのでは，検定にたえ得る数値は得られない。そこでここでは，M=4.1, S=1.37として，χ^2検定による対数正規分布の適合度検定を行ってみた。仮説 H_0 および有意水準 α を次のようにおいた。その計算が表2である。

H_0 : 表3.6のデータは M＝4.1, S＝1.37 の対数正規分布に従う。
$\alpha = 0.05$

表2　χ^2 検定のための計算

				ζ_i	$-\infty$からζ_iまでの面積			
				‖	‖		期待値	
i	期間(T_i)		度数(O_i)	$\log(T_i/4.1)/1.37$	$P(\zeta_i)$	$P(\zeta_i)-P(\zeta_{i-1})$	E_i	$\chi^2=(O_i-E_i)^2/E_i$
1	0	− 1	138	-1.030	0.1515	0.1515	118.6	3.173
2	1.1	− 2	111	-0.524	0.3001	0.1486	116.4	0.251
3	2.1	− 3	85	-0.228	0.4098	0.1097	85.9	0.009
4	3.1	− 4	63	-0.018	0.4928	0.0830	65.0	0.062
5	4.1	− 5	42	0.145	0.5577	0.0649	50.8	1.524
6	5.1	− 6	51	0.278	0.6095	0.0518	40.6	2.664
7	6.1	− 9	67	0.574	0.7170	0.1075	84.2	3.514
8	9.1	−12	66	0.784	0.7835	0.0665	52.1	3.708
9	12.1	−18	56	1.080	0.8599	0.0764	59.8	0.241
10	18.1	−24	36	1.290	0.9015	0.0416	32.6	0.355
11	24.1	−36	23	1.586	0.9436	0.0421	33.0	3.030
12	36.1	−48	7	1.796	0.9638	0.0202	15.8	4.901
13	48.1	−60	11	1.959	0.9749	0.0111	8.7	0.608
14	60	～	27		1.0000	0.0251	19.7	2.705
合計			783				783.2	26.745

この検定では d.f.＝14−3＝11, χ^2 分布表から $\alpha = 0.05$ に対応する値は19.68であるから，仮説 H_0 は有意水準5％で棄却される。

6．おわりに

以上で，表3.6のデータへの対数正規分布あてはめの計算とグラフの概形を示し，あわせてその適合度検定を行ってみた。著者が，M および S を求めるのに常用対数を，あとの計算に自然対数を用いた理由ははっきりしない。対数正規確率紙にデータをプロットして対数正規性を検証するとき，対数目盛りを自然対数とみなしても，常用対数とみなしても尺度の違いにすぎないから，直線による検証それ自体には関係ない。誤用は，あるいは，そこらに起因しているのかもしれない。このあとの“2期の子宮頸部癌のため死亡した患者 583人の生存期間” に関するデータの分析においても，これと同じ混用が見られる

なお，図3.8ではデータはほぼ直線上に並んでいるが，上記の仮説H_0は棄却された。この場合仮説そのものが妥当でないことも考えられる。直線がどのように引かれたか，母数の推定値がどこまで精確に読み取られたか，原著からは判然としない。そこで，計算は省略するが，データから最小二乗直線を求め，これにもとずいて得た推定値 M＝3.85，S＝1.445 を用いて改めてχ^2検定の計算を行ってみた。 その結果 χ^2＝23.2 とやや小さな値が得られたが依然として棄却域に含まれる。 そこでさらに，表3.6の 2つの区間，6.1－9，9.1－12 を1つにまとめて検定を行ってみた。 次がその結果である。

H_0 :表3.6のデータは M=3.85, S=1.445 の対数正規分布に従う。
$\alpha=0.05$

表3　χ^2 検定のための計算

			ζ_i		−∞からζ_iまでの面積		
			‖		‖	期待値	
i	期間(T_i)	度数(O_i)	$\log(T_i/3.85)/1.445$	$P(\zeta_i)$	$P(\zeta_i)-P(\zeta_{i-1})$	E_i	$\chi^2=(O_i-E_i)^2/E_i$
1	0 − 1	138	−0.933	0.1754	0.1754	137.3	0.004
2	1.1− 2	111	−0.453	0.3253	0.1499	117.4	0.349
3	2.1− 3	85	−0.173	0.4313	0.1060	83.0	0.048
4	3.1− 4	63	0.026	0.5104	0.0791	61.9	0.020
5	4.1− 5	42	0.181	0.5718	0.0614	48.1	0.774
6	5.1− 6	51	0.307	0.6206	0.0488	38.2	4.289
7	6.1−12	133	0.787	0.7843	0.1637	128.2	0.180
8	12.1−18	56	1.067	0.8570	0.0727	56.9	0.014
9	18.1−24	36	1.266	0.8973	0.0403	31.6	0.613
10	24.1−36	23	1.547	0.9390	0.0417	32.7	2.877
11	36.1−48	7	1.746	0.9596	0.0206	16.1	5.143
12	48.1−60	11	1.901	0.9714	0.0118	9.2	0.352
13	60〜	27			0.0286	22.4	0.945
計		783				783	15.608

この検定では d.f.=13−3=10, χ^2分布表から $\alpha=0.05$に対応する値は18.31であり，この場合仮説H_0は有意水準 5%では棄却できない。

以上のことから，表3.6のデータは，精密とは言えないにしても，おおまかには対数正規分布で近似できると判断してもよかろう。ただこのデータはかなり古く，医学が進歩し，予防医学や健康への関心も一般に高まっているこんにち，癌の症状期間や生存期間がどのような分布を示すのか大いに興味がもたれるところである。

＊）R.F.モールド　小林定喜訳：目で見る――チェルノブイリの真実（西村書店，1992)。

＊＊）ここでの症状期間（Symptom time）とは，患者が症状を自覚してから，はじめて受診するまでの遅延期間を言う。

＊＊＊）Percentage relative frequency の誤り。