

臨床試験における対数変換と かれの仲間たち

2024.9

章立

自己紹介

1. 名前：章 立(しょう りつ/ZHANG LI)
2. 経歴：2007年10月来日、保険数理やIT技術サポートなど仕事していて、2018年EPSと縁が有ったお陰で統計解析の業界に転職、現在EPS統計解析Team大阪所属
3. 出身：中国・襄陽市
4. 大学：華中科技大学・数理統計学院
(Huazhong University of Science and Technology / HUST)



华中科技大学



本日のアジェンダ

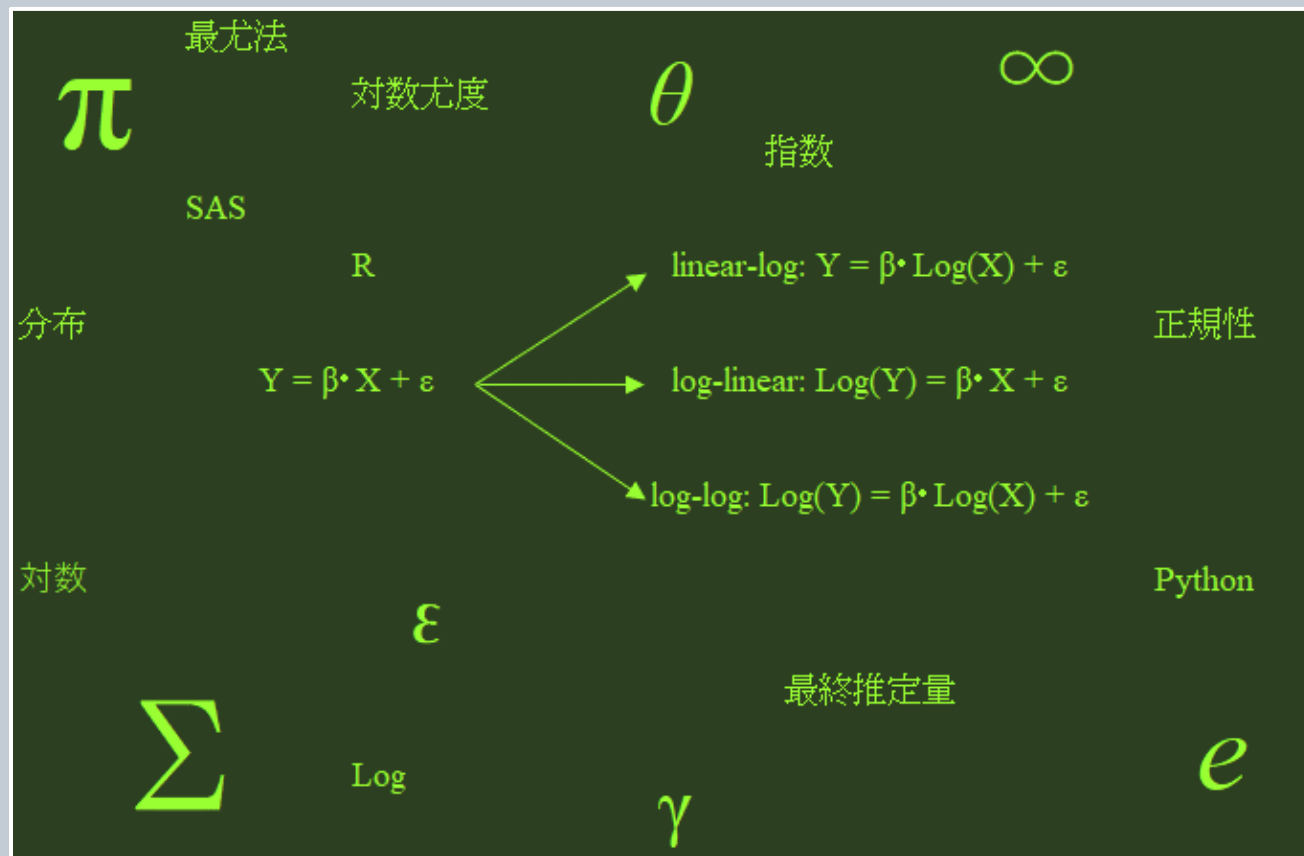
1. はじめに
2. データ変換について
3. 対数変換モデル
4. Box-Cox変換について
5. 業務実例：対数変換モデル
6. まとめ・感想
7. 課題・再考
8. おまけ・雑談
9. 参考文献

キーワード：対数変換、平方根変換、Box-Cox変換、正規分布/対数正規分布

1. 初めに

背景：業務中に対数変換で線形モデル解析を対応した時、なぜデータ変換必要か、なぜ対数変換でないといけないか、他のデータ変換方法あるのか、データ変換のメリットとデメリットは何であろうかについて思ったり、調べたりした。結果を資料にまとめ、今回の発表にたどり着いた。

なぜ仲間たちと呼ぶか：同じくデータ変換の範囲に属して、同じ目的のためにお互いに協力し合い、統計解析に力を貢献していると個人はそう思いながら、勝手に彼らは仲間たちと呼ぶ。



2. データ変換について①

› なぜデータ変換必要か

正規分布以外の分布をデータ変換で正規分布に近似したため、実データの検定よりも高い検出力で解析できる

› データ変換方法の種類

対数変換

$$Y = \log X$$

平方根変換

$$\sqrt{x}$$

その他

べき乗変換

$$Y = X^a$$

Box-Cox変換

$$Y = \begin{cases} \frac{X^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \log X & (\lambda = 0) \end{cases}$$

ロジット変換

$$\log \frac{X}{1 - X}$$

ロジスティック変換

$$\frac{1}{1 + e^{-X}}$$

2. データ変換について②

› データ変換方法の選択：

対数変換か、平方根変換か、冪乗変換か、Box-Cox変換法で最適の変換法を決める

$$Y = \begin{cases} \frac{X^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \log X & (\lambda = 0) \end{cases}$$

※補足1：最尤法いわゆる対数尤度を最大にするモデル(データ変換法)を最適モデル(最適変換法)

※補足2：尤度関数(平均が μ ，分散が σ^2

の正規分布に独立に従う乱数の値

が x_1, \dots, x_n である)

$$L(\mu, \sigma^2) = \prod_{k=1}^n \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_k - \mu)^2}{2\sigma^2}\right) \right\}$$

※補足3：対数尤度関数

$$\log L(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - \mu)^2$$

2. データ変換について③

ICH E9「臨床試験のための統計的原則」 5.4 データ変換では以下の記述がなされている。

各都道府県衛生主管部(局)長 殿

医薬審 第1047号
平成10年11月30日

厚生省医薬安全局審査管理課長

「臨床試験のための統計的原則」について

近年、優れた新医薬品の地球的規模での研究開発の促進と患者への迅速な提供を図るため、承認審査資料の国際的ハーモナイゼーション推進の必要性が指摘されている。このような要請に応えるため、日・米・EU三極医薬品規制調和国際会議(ICH)が組織され、品質、安全性及び有効性の3分野でハーモナイゼーションの促進を図るための活動が行われている。

別添の「臨床試験のための統計的原則」(以下「本ガイドライン」という。)は、ICHにおける合意に基づき、臨床試験における統計的原則について記載したものであり、臨床試験から得られる結果の偏りを最小にし、精度を最大にすることを目標としている。特に、計

4.6 独立データモニタリング委員会の役割	23
V. データ解析上で考慮すべきこと	23
5.1 解析の事前明記	23
5.2 解析対象集団	24
5.2.1 最大の解析対象集団	24
5.2.2 治験実施計画書に適合した対象集団	26
5.2.3 二つの異なる解析対象集団の役割	26
5.3 欠測値と外れ値	27
5.4 データ変換	27
5.5 推定、信頼区間及び仮説検定	28
5.6 有意水準と信頼水準の調整	28
5.7 部分集団、交互作用及び共変量	29
5.8 データの完全性の維持とコンピュータソフトウェアの妥当性	29

5.4 データ変換

重要な変数を変換するために必要な判断は解析の前に行い、先行する臨床試験での類似データに基づいて治験実施計画立案時に行うのが最善である。変換(例えば、平方根、対数)を行うことは、主要変数については特に治験実施計画書に明記すべきであり、その理論的根拠を述べるべきである。統計手法の前提を満たすことを保証するための変換の一般的な原則は、標準的な教科書に書かれている。また、特定の変数についての変換の慣例的方法は、多くの臨床領域別に開発されてきている。変数を変換するかどうか、変換するのであればどのように変換するかという判断は、臨床的な解釈を容易にする尺度を選択するという観点も含めて行われるべきである。

同様な配慮は、基準となる時点での値からの変化、基準となる時点での値からの変化割合、繰り返し測定「曲線下面積」、又は二つの異なる変数の比、といった新たな変数の導出の際にも行うべきである。新たな変数の臨床的解釈は注意深く検討されるべきであり、その正当性も治験実施計画書に述べるべきである。密接に関連した話題が2.2.2節に述べられている。

※対数変換が一般的な、伝統的な手法だが、なぜか文書に平方根が先に記述されている。

3. 対数変換モデル

対数変化に関する三つのモデルについて

・線形-対数モデル linear-log

$$Y = \beta \ln X + \varepsilon$$

このとき、 β は X が1%増加したとき、 Y がいくら（もとの単位で）増加するかを示す。

・対数-線形モデル log-linear

$$\ln Y = \beta X + \varepsilon$$

ここでは逆に、 β は X が1単位増加したとき、 Y が $100 \times \beta$ %増加するかを示す。

・対数-対数モデル log-log

$$\ln Y = \beta \ln X + \varepsilon$$

この場合は β は弾力性と呼ばれるもので、 X が1%増加したときに Y が何 β %増加するかを示す。

※ X : Independent variables/説明(数値)変数、 Y : Dependent variables/結果変数

※ 弾力性とは、経済学用語の1つで「 x の変化率に対する y の変化率の比」のことです。

(例) 所得 x が 1% 増えたら需要 y が 0.8% 増える場合、需要の所得弾力性は 0.8 となる。

※ 経済学ではlog-linearとlog-logをよく見る印象である。

4. Box-Cox変換について①

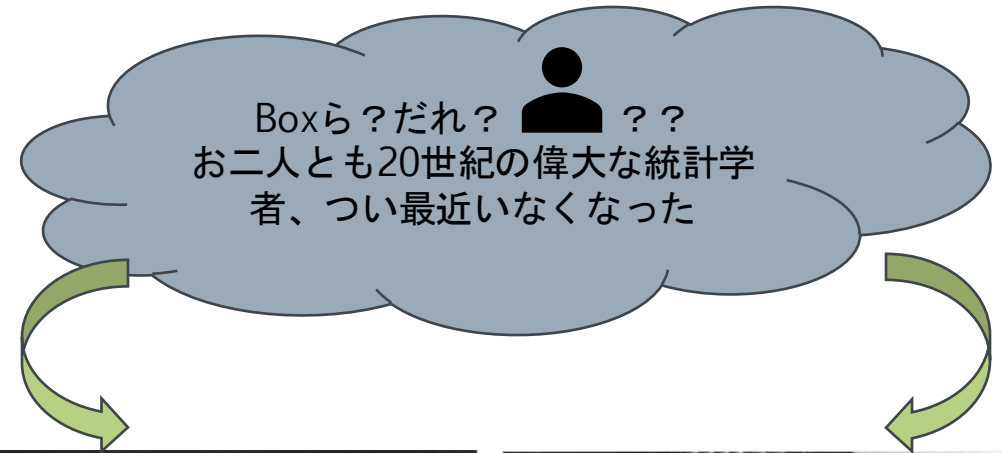
- Boxらが提唱された変換方法

Box-Cox方法のパラメータ λ を少しずつ調整し、対数尤度を最大になるかどうか探していく。
逆に言うと対数尤度最大時の λ が見つかれば、データ変換方法を決める。

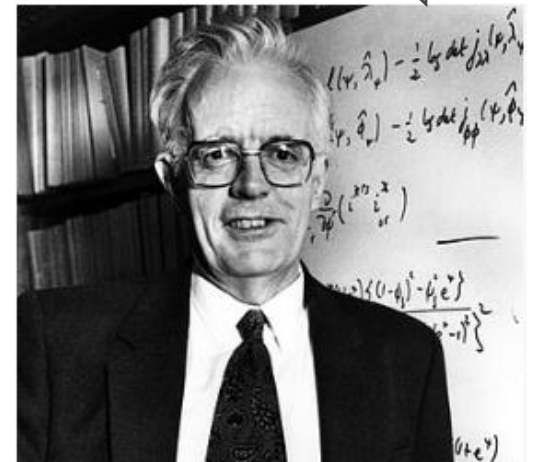
- ※補足1 : 公式↓

$$x \mapsto x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log x & \lambda = 0 \end{cases}$$

- ※補足2 : cox回帰分析も1972年David Coxが提唱した多変量解析の手法である



George Box
※Fisherの婿(むこ)さん



David Cox
※イギリス出身

4. Box-Cox変換について②

› Box-Cox変換方法の公式からわかること。

$\lambda=0 \rightarrow \log(x)$ 対数変換

$\lambda=0.5 \rightarrow 2(x^{0.5} - 1)$ 平方根変換

$\lambda=1 \rightarrow (x - 1)$ 平行移動変換

$\lambda=\dots$

$$x \mapsto x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log x & \lambda = 0 \end{cases}$$

› 変換方法を決めるための λ を探索する。

1. テストデータを作成
2. SASプロシージャtransregでテストデータを評価
3. SASプロシージャtransregの推定結果をOutput
4. 対数尤度が最大時の λ によって変換方法を決定

Box-Cox変換法を対応するSASプロシージャ

```
proc transreg data=[testdata] test;
```

```
model BoxCox(y) = identity(x);
```

```
run;
```

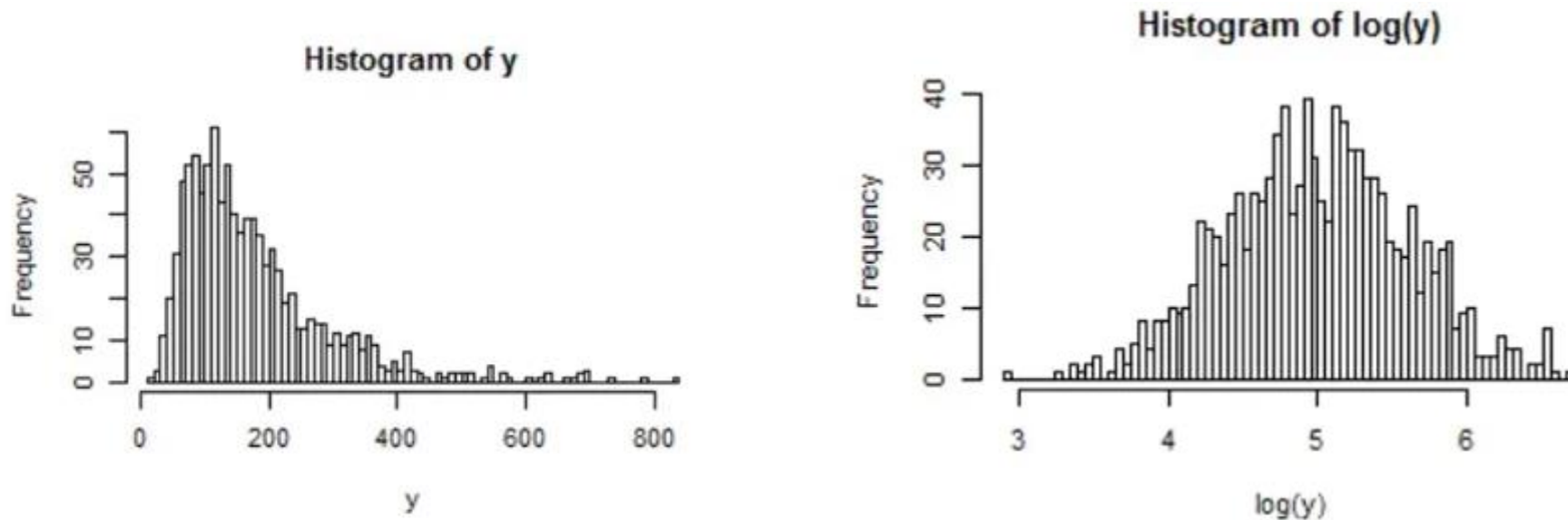
※1 y: Dependent variables/結果変数

※2 x: Independent variables/説明(数値)変数

※3 TRANSREGは、回帰分析、正準相関分析、分散分析などのモデルに対し、交互最小二乗法を使って最適変数変換を求めるためのプロシージャである。

4. Box-Cox変換について③

テストデータ準備前：どんなデータが対数変換必要か



左図：実データ、右図：対数変換データ

指数分布またガンマ分布のような右へ歪んだ分布に対数変換が良く用いられる

※1 確率変数の対数変換後の分布が正規分布である時、その確率変数は対数正規分布に従うと言う。

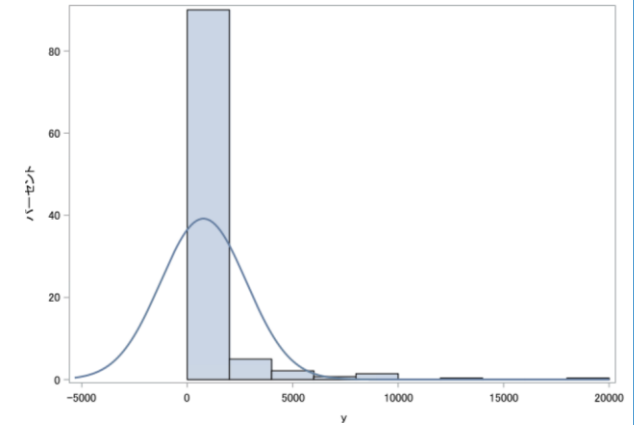
※2 対数変換できない時(データ=0存在)は、 $\text{Log}(x+C)$ で対応。例えば、実務時 $x=0$ の場合は $\text{Log}(0.5)$ で対応した。

※3 参考書：《ヘルスデータサイエンス健康科学のための統計解析》Ruth Etzioni 著/岩崎 学 監訳

4. Box-Cox変換について④

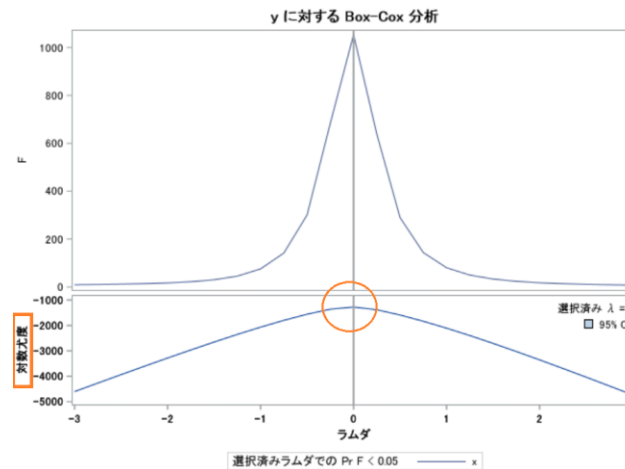
- ▶ テストデータ実験(やっとSASをいじるところに来て感動…)
- ▶ 指数分布のテストデータを事前作成
及び分布確認

```
/*SAS Code*/
data aaa;
  do x = 1 to 8 by 0.025;
    y=exp(x + normal(7));
    output;
  end;
run;
proc sgplot data=aaa;
  histogram y;
  density y;
run;
```



- ▶ proc transreg/Box-Cox変換
及び対数尤度最大時の λ を確認

```
/*SAS Code*/
ods output BoxCoxPlot=boxcoxplot;
proc transreg data=aaa test;
  model BoxCox(y) = identity(x);
run;
```



	Dependent	LogLike	Lambda
1	BoxCox(y)	-1275.314698	0.00
2	BoxCox(y)	-1360.284745	-0.25
3	BoxCox(y)	-1382.618821	0.25
4	BoxCox(y)	-1558.550552	-0.50
5	BoxCox(y)	-1589.033853	0.50
6	BoxCox(y)	-1799.246623	-0.75
7	BoxCox(y)	-1834.53237	0.75
8	BoxCox(y)	-2066.320076	-1.00
9	BoxCox(y)	-2105.884882	1.00
10	BoxCox(y)	-2352.715266	-1.25
11	BoxCox(y)	-2397.354458	1.25
12	BoxCox(y)	-2653.139503	-1.50
13	BoxCox(y)	-2704.63657	1.50
14	BoxCox(y)	-2963.742917	-1.75
15	BoxCox(y)	-3024.239463	1.75
16	BoxCox(y)	-3281.879255	-2.00
17	BoxCox(y)	-3559.881455	2.00

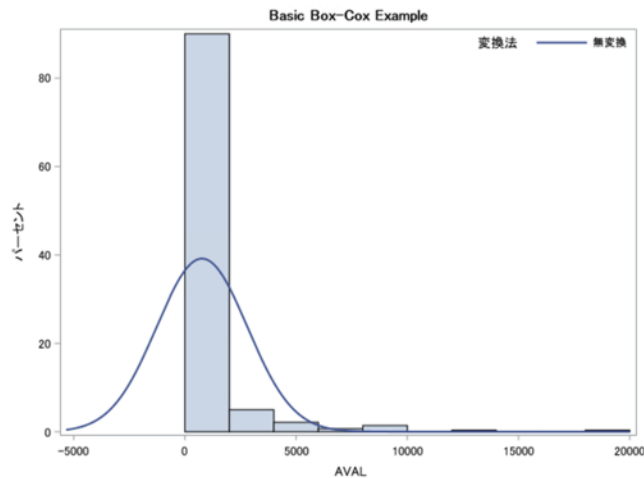
4. Box-Cox変換について⑤

- Box-Cox変換方法で $\lambda=0$ 、対数変換を決めたが、実際対数変換、平方根変換後の分布を比較しながら、Box-Cox変換方法で進められた変換方法はより正規分布に近似するかどうか、すなわち最尤モデルかどうかを確認したい。

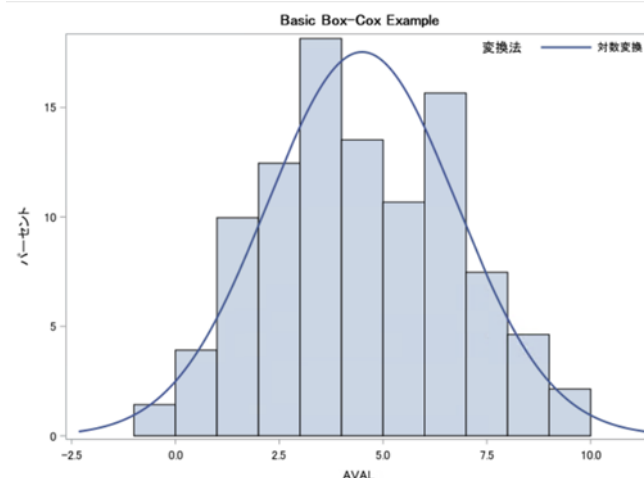
```
/*指数分布元データ*/
data aaa;
do x = 1 to 8 by 0.025;
  y=exp(x + normal(7));
  output;
end;
run;
```

```
/*元データを対数変換、平方根変換*/
data test;
set aaa;
tran_method=1; AVAL=y; output;
tran_method=2; AVAL=log(y); output;
tran_method=3; AVAL=sqrt(y); output;
run;
```

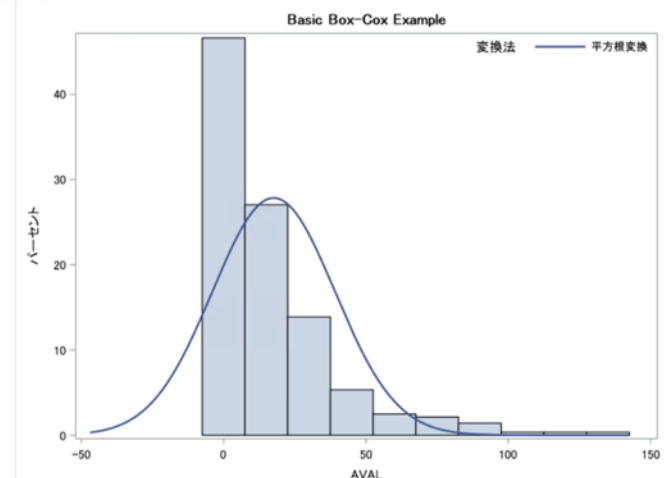
```
/*ヒストグラム、密度推定図*/
proc sgplot data=test(where=(tran_method=1));
  histogram AVAL;
  density AVAL;
run;
```



変換なし



$\lambda=0$, 対数変換



$\lambda=0.5$, 平方根変換

4. Box-Cox変換について⑥

- 正規確率点プロット(QQプロット)で正規分布に近似するかの確認

QQプロットの見方：データ分布が確率分布に一致すると、プロットの点が線形状に並びます。つまり、プロットした結果が直線に見えたら正規分布だなということになります。

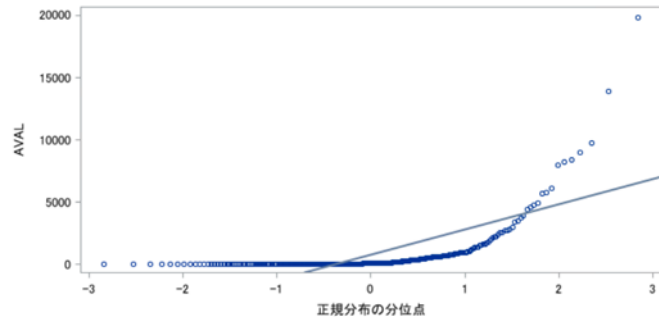
```

/*指数分布元データ*/
data aaa;
do x = 1 to 8 by 0.025;
  y=exp(x + normal(7));
  output;
end;
run;

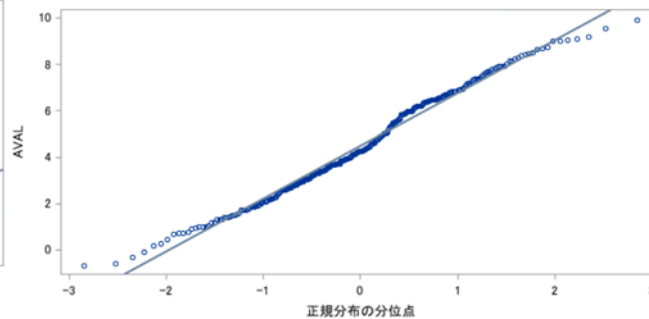
/*元データを対数変換、平方根変換*/
data test;
set aaa;
tran_method=1; AVAL=y; output;
tran_method=2; AVAL=log(y); output;
tran_method=3; AVAL=sqrt(y); output;
run;

/*QQ図*/
proc univariate data=test plot;
class tran_method;
var AVAL;
run;

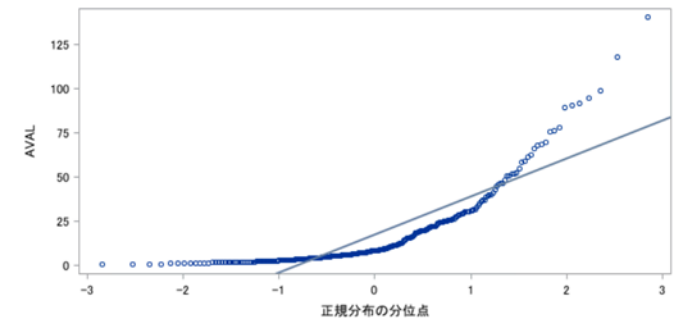
```



変換なし



$\lambda = 0$, 対数変換



$\lambda = 0.5$, 平方根変換

4. Box-Cox変換について⑦

- Box-Cox変換方法で $\lambda=0$ とピッタリなるのは偶然か、0でもなく、0.5でもないの時はどうするか
- 対策：オプションconvenientを指定し、最尤 λ の代わりに、 $\lambda=0$ が99.999%CIに入っていれば、 $\lambda=0$ を選択する

※また最尤 $\lambda=0.61$ かつ0.5が信頼区間に入っていれば、 $\lambda=0.5$ (平方根変換)を選択可能

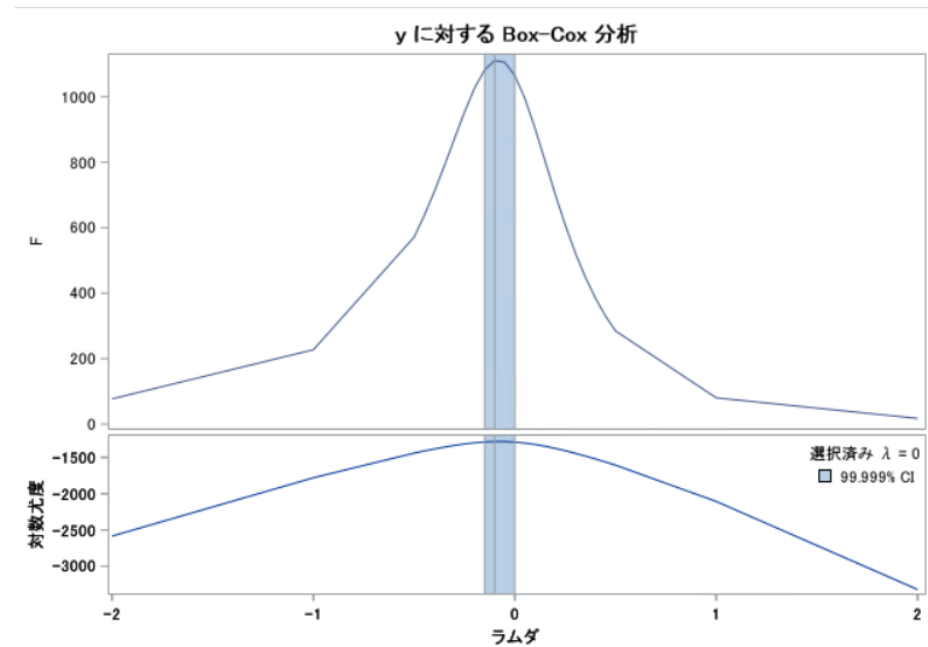
```

/*指数分布元データ*/
data aaa;
  do x = 1 to 8 by 0.025;
    y = exp(x + normal(7));
  output;
end;
run;

/*  $\lambda=0$ が最尤でないように元データを+2で少し上に平行移動させる*/
data aaa1;
  set aaa;
  y = y + 2;
run;

ods output BoxCoxPlot=boxcoxplot2;
proc transreg data=aaa1;
  model BoxCox(y / lambda=-2 -1 -0.5 to 0.5 by 0.05 1 2
               convenient alpha=0.00001)
        = identity(x);
run;

```

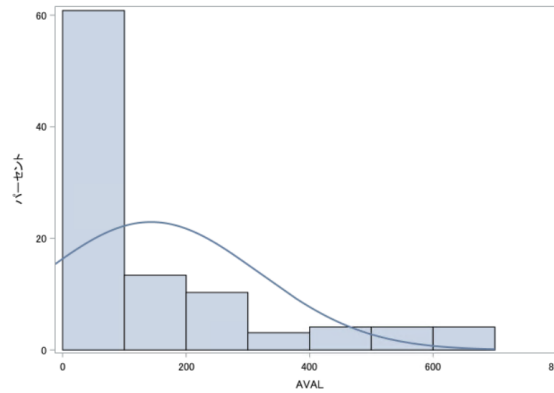


$\lambda=0.1$ 最尤けど、99.999%CIに入っている $\lambda=0$ を選択

5. 業務実例：分布状況と λ の確認

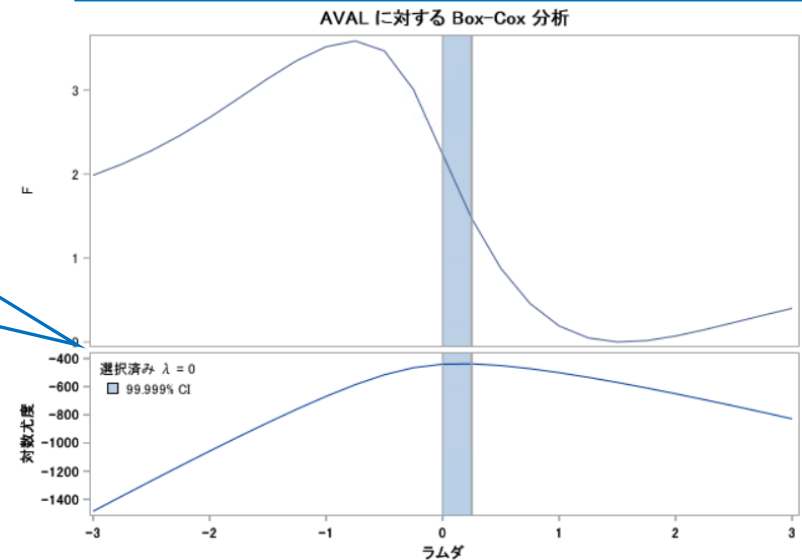
ある試験で面積で評価される某病気のデータをBox-Cox変換方法で λ を確認してみた。

	SUBJID	AVISITN	AVAL
1	01-001	30	249.76196009
2	01-001	10	205
3	01-001	20	239
4	01-001	31	318
5	01-001	90	207
6	01-001	100	351
7	01-001	110	345
8	01-001	120	243
9	01-001	130	467
10	01-003	30	230.03824191
11	01-003	10	182
12	01-003	20	245
13	01-003	31	273
14	01-003	90	212
15	01-003	100	223
16	01-003	110	138
17	01-003	120	132
18	01-003	130	99
19	01-003	140	115
20	01-003	150	46




```
/*0以上の正数データが必要*/
data adfa_test2;
set adfa_test;
if AVAL=0 then AVAL=0.5;
run;
ods output BoxCoxPlot=boxcoxplot3;
proc transreg data=adfa_test2 test;
model BoxCox(AVAL / convenient alpha=0.00001)
= identity(AVISITN);
run;
```

$\lambda = 0.25$ 最尤けど、
99.999%CIに入っ
ている $\lambda = 0$ (対数
変換)を選択



5. 業務実例：対応変換の結果と再変換

- 対数モデル: log-linear

$$\text{Log}(AVAL) = \beta \cdot \text{AVISITN} + \varepsilon$$
- SAS構文: 
- 変化率への再変換:

100*[exp(推定値)-1]

※推定値：対数変換後の差分

※推定値と再変換の結果：

●予測推定値

評価時点	対数変換値に 対するベースライン からの差	95%信頼区間	ベースラインから の変化率(%)	95%信頼区間
第4週	-0.10	(-0.31, 0.12)	-9.2	(-26.89, 12.75)
第8週	0.08	(-0.08, 0.23)	8.0	(-7.83, 26.34)
第12週	0.01	(-0.22, 0.23)	0.6	(-19.45, 25.71)
第16週	-0.41	(-0.94, 0.12)	-33.7	(-60.96, 12.67)
第20週	-0.30	(-0.93, 0.32)	-26.3	(-60.39, 37.25)
第24週	-0.61	(-1.58, 0.35)	-45.6	(-79.08, 41.42)
第28週	-0.70	(-1.38, -0.01)	-50.2	(-74.92, -1.29)
第4週～第12週	0.00	(-0.17, 0.16)	-0.4	(-15.27, 16.97)
第12週～第20週	-0.24	(-0.52, 0.05)	-21.1	(-40.45, 4.87)
第20週～第28週	-0.54	(-0.95, -0.12)	-41.6	(-81.31, -11.74)
第4週～第28週	-0.29	(-0.52, -0.07)	-25.2	(-40.32, -6.34)

```
data wk_01;
  set wk_00;
  by subjid avisitn;
  where avisitn in (0,90,100,110,120,130,140,150);
  if aval>0 then aval_log=log(aval);
  else if aval=0 then aval_log=log(0.5);
run;
```

```
ods output SolutionF =_SolutionF
  CovParms =_CovParms
  Estimates =_Est;
```

```
proc mixed data=wk_02 method=reml covtest;
  class subjid avisitn(ref="0");
  model aval_log = avisitn / solution cl ddfm=kr;
  random intercept / subject=subjid type=un solution cl;
  repeated / type=vc group=AVISITN;
```

```
estimate 'ln(W04)-lnW00' AVISITN 1 0 0 0 0 0 -1 / cl divisor=1;
estimate 'ln(W08)-lnW00' AVISITN 0 1 0 0 0 0 -1 / cl divisor=1;
estimate 'ln(W12)-lnW00' AVISITN 0 0 1 0 0 0 -1 / cl divisor=1;
estimate 'ln(W16)-lnW00' AVISITN 0 0 0 1 0 0 -1 / cl divisor=1;
estimate 'ln(W20)-lnW00' AVISITN 0 0 0 0 1 0 -1 / cl divisor=1;
estimate 'ln(W24)-lnW00' AVISITN 0 0 0 0 0 1 -1 / cl divisor=1;
estimate 'ln(W28)-lnW00' AVISITN 0 0 0 0 0 0 1 -1 / cl divisor=1;
estimate 'ln(W04-12)-lnW00' AVISITN 1 1 1 0 0 0 0 -3 / cl divisor=3;
estimate 'ln(W12-20)-lnW00' AVISITN 0 0 1 1 1 0 0 -3 / cl divisor=3;
estimate 'ln(W20-28)-lnW00' AVISITN 0 0 0 0 1 1 1 -3 / cl divisor=3;
estimate 'ln(W04-28)-lnW00' AVISITN 1 1 1 1 1 1 1 -7 / cl divisor=7;
```

```
run;
```

6. まとめ・感想

-1 はオフセット値だと言われている

› まとめ：

- データ変換の種類とデータ変換法の選択について自己学習した
- データ変換のメリット：

- ①データの分布を正規分布に近似させ、実データにt検定などの実行よりも高い検出力で解析を行うことができる
- ②対数をとると線形になることがある
- ③対数をとると分散が一定になることがある
- ④対数の差は変化率と解釈される

$$\ln(\text{week4}) - \ln(\text{week0}) = \ln\left(\frac{\text{week4}}{\text{week0}}\right) = \ln\left[1 + \frac{\text{week4} - \text{week0}}{\text{week0}}\right] \rightarrow \text{指数変換} \exp(*) - 1 \rightarrow \frac{\text{week4} - \text{week0}}{\text{week0}}$$

- データ変化のデメリット(注意点?)：モデルの推定値を元の単位(スケール)に戻す時データの解釈が難しくなることがある。

› 感想：

解析業務においては、個々のデータの意味を理解するのはもちろん重要で、データ全体の特徴(分布など)も把握すべきかと感じた

7. 課題・再考

- › 課題①：再変換の問題：推定値をそのまま指数変換で元の単位(スケール)に戻すとデータが解釈が難しくなることがある
 - 指数変換してもよい例：測定値・変化量・変化率の推定値を元の単位(スケール)に戻す時
 $100 * [\exp(\text{推定値}) - 1]$ ←業務中に変化率を指数変換で再変換した…
 - 指数変換してもよくない例：平均の推定値を元の単位(スケール)に戻す時
例えば、対数変換後の平均 $E[\text{AVAL_Log}] = \mu$ 、分散 $\text{Var}[\text{AVAL_Log}] = \sigma^2$ とすると、
元の観測値AVALの平均は $\exp(\mu)$ ではなく
 $E[\text{AVAL}] = \exp(\mu + \sigma^2/2) = \exp(\mu) * \exp(\sigma^2/2)$ となるのだから。
- › 課題②：元の単位(スケール)に戻す再変換の問題で、観測値を変換せず、観測値そのままをモデル化するため、歪んだ分布に対応した、対数リンクをもつガンマ一般化線形モデルをこれから勉強予定→proc genmod/proc logistic/proc catmod

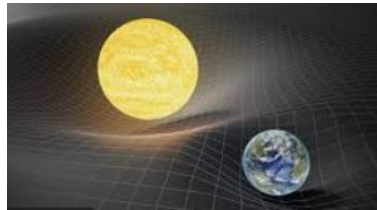
※参考《ヘルスデータサイエンス健康科学のための統計解析》Ruth Etzioni 著/岩崎 学 監訳

8. おまけ・雑談

①太陽質量の計算

万有引力からM(太陽質量)の計算式

$$M = \frac{4\pi^2}{G} \left(\frac{r^3}{T^2} \right)$$



M : 太陽質量

π : 円周率

r : 太陽から地球までの距離 ($1.496 \times 10^{11} \text{m}$)

T : 地球の公転周期 ($3.156 \times 10^7 \text{s}$)

G : 万有引力定数 ($6.674 \times 10^{-11} \text{Nm}^2/\text{kg}^2$)

※G先に定義されたかM先に計算されたかはさて置き
電卓お持ちでしたら、押すだけで結果が出るが、電卓誕生前に、そのまま手計算すると非常に手間かかるため、底数10の対数変換で手計算でもと楽…

$$\log M = 2 \log(4\pi) + 3 \log(r) - \log(G) - 2 \log(T)$$

$$= \text{対数表を調べて…}$$

$$= 30.2990$$

ですと、

$$M = 10^{30.2990} = 10^{30} * 10^{0.2990}$$

$$= 10^{30} * 1.998 \leftarrow \text{対数表を調べて…}$$

$$\doteq 2 * 10^{30} \text{ Kg}$$

【常用対数表】

数	0	1	2	3	4	5	6	7	8	9
1.0	.0000	.0043	.0086	.0128	.0170	.0212	.0253	.0294	.0334	.0374
1.1	.0414	.0453	.0492	.0531	.0569	.0607	.0645	.0682	.0719	.0755
1.2	.0792	.0828	.0864	.0899	.0934	.0969	.1004	.1038	.1072	.1106
1.3	.1139	.1173	.1206	.1239	.1271	.1303	.1335	.1367	.1399	.1430
1.4	.1461	.1492	.1523	.1553	.1584	.1614	.1644	.1673	.1703	.1732
1.5	.1761	.1790	.1818	.1847	.1875	.1903	.1931	.1959	.1987	.2014
1.6	.2041	.2068	.2095	.2122	.2148	.2175	.2201	.2227	.2253	.2279

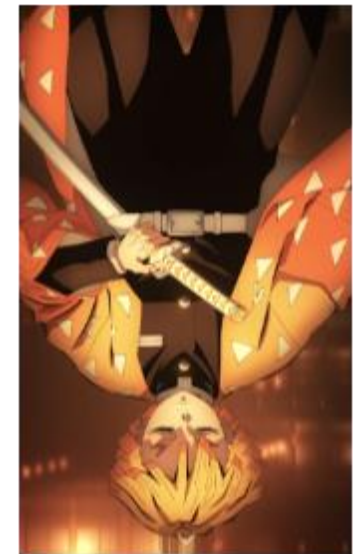
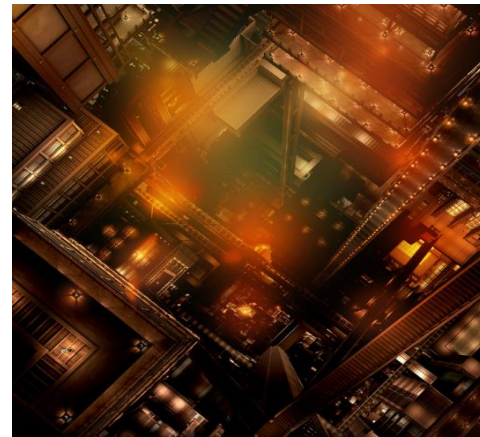
②データ変換 = 線形変換 = 空間変換

2014年11月22日上映したSF映画の「インターステラー」で人類・地球を救うため、宇宙探索中にワームホールに吸込まれ、超異次元空間に墜ち、モルルス信号…

ここでの発想：
ワームホールにも何等かのデータ変換法をもつだろう
Wormhole
 \doteq Box-Cox???



③鬼滅の刃-無限城も現実世界から何等かのデータ変換で作られている？



9. 参考文献

➤ 参考文献

1. George Box<https://en.wikipedia.org/wiki/George_E._P._Box>
2. David Cox<[https://en.wikipedia.org/wiki/David_Cox_\(statistician\)](https://en.wikipedia.org/wiki/David_Cox_(statistician))>
3. Online Statistic Book<<https://onlinestatbook.com/2/transformations/box-cox.html>>
4. 正しい分散分析結果を導くための変数変換法<https://jppa.or.jp/archive/pdf/56_10_22.pdf>
5. 臨床試験のための統計的原則<<https://www.pmda.go.jp/files/000156112.pdf>>
6. 臨床検査データの正規性/データ変換による分布の正規化
<https://www.jstage.jst.go.jp/article/kampomed1982/40/4/40_4_199/_pdf>
7. Proc TRANSREG
<https://documentation.sas.com/doc/en/statcdc/14.2/statug/statug_transreg_details02.htm>
8. 統計参考本《ヘルスデータサイエンス健康科学のための統計解析》Ruth Etzioni 著/岩崎 学 監訳

ご清聴、有難う御座いました。

