時間の長さを分析する

粕谷 英一

（九大・理・生物）

生態学会（2015年3月）自由集会『データ解析で出会う統計的問題: 時間の長さ・時系列データの統計モデリング』

時間

状態の持続時間

回帰的枠組みなら

目的変数

時間

状態の持続時間 duration

ある現象が起こるまでの経過時間
latency

瞬間的な現象について 時間間隔
interval

時間

状態の持続時間　duration

例.餌を食べ始めてから食べ終わるまで

例.交尾開始から交尾終了まで

例.静止してから動き始めるまで

状態→生存　　例.寿命　　例.余命

時間

ある現象が起こるまでの経過時間

latency

例.病気の潜伏期間

例.刺激を与えてから反応が現れるまでの時間

時間

時間間隔　　　interval

ある現象とある現象の間の時間間隔

例.捕食者が攻撃してくる間隔

例.ある餌場を訪れる間隔

例.バス停に人が来る間隔

時間

生存時間

文字通りの生存時間でないものも
機械やシステムの故障までの時間など

状態の持続時間　duration

ある現象が起こるまでの経過時間　latency

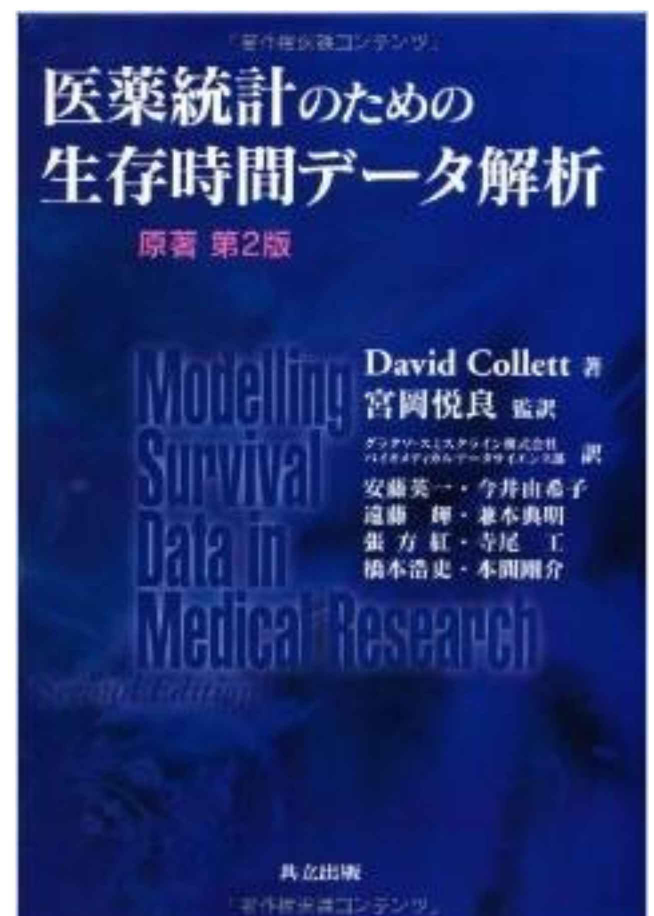
時間間隔　interval

時間　生存時間

文字通りの生存時間でないものも

機械やシステムの故障までの時間など

信頼性（reliability）



時間　データの特徴

負の値がない　正（正かゼロ）

連続量

尺度水準　比率尺度（比尺度）

重さ、長さ、体積などと同じ尺度水準

時間 データと分析の特徴

注目する特徴 時間

生存時間 生存率 ハザード

打ち切り

説明変数の値の時間的変化

よく使われる”変わった”モデル Cox比例ハザードモデル

時間

以降の内容の目次

時間依存的な説明変数

打ち切り　Kaplan-Meier推定量
logrank検定

生存関数とハザード

比例ハザードモデル

パラメトリックなモデル

時間

広義の回帰の枠組み

説明変数 ➔ 目的(応答)変数

回帰式の関数型

目的変数のばらつき具合

誤差構造と確率分布

時間の経過

説明変数→目的（応答）変数

Cox比例ハザードモデルでは共変量

説明変数の値が途中で変化しない

定数型（fixed）

観察開始時の値がとられることが多い

説明変数の値が途中で変化する　Cox比例ハザードモデルでは普通

時間依存的（time-dependent）

時間のラグがある
（時間により説明変数の係数が変化）

時間

時間依存的な説明変数

パッケージsurvivalなど

比例ハザードモデルの関数coxph

説明変数指定のオプション　tt

データ整形の関数　tmerge

データ読み取りの関数　Surv

交互作用が使えることもある

打ち切り　censor

時間の長さが正確にわからない

生存時間

ある時点では生きていた

死亡時点は不明

しばしば「観察をやめた」ことによる

ある時点では生きていた

死亡時点は不明

打ち切り censor

値はわからない
ある値以上であることはわかる
right **右側**

値はわからない
ある値以下であることはわかる
left **左側**

値はわからない
ある範囲に入っていることはわかる
interval

打ち切り　censor

値はわからない

ある値以上であることはわかる

right　**右側**

パッケージsurvivalの関数Surv

例.

2000頭

1年目生存、2年目生存	784
1年目生存、2年目死亡	16
1年目死亡	440
1年目生存、2年目打ち切り	760

1年目までの生存率

２年目までの生存率

は？

例.

2000頭

1年目生存、2年目生存	784
1年目生存、2年目死亡	16
1年目死亡	440
1年目生存、2年目打ち切り	760

1年目の生存率＝(784+16+770)/2000 =0.78

2年目の生存率＝784/(784+16) =0.98

2年目までの生存率＝1年目の生存率×2年目の生存率

=0.7644

Kaplan－Meier推定量

(product limit推定量)

ある時点までの生存率の推定値＝区間の生存率の積

区間

死亡が起こったらその時点で区間を区切る

死亡が起こった時点で区切った区間の生存率

Kaplan－Meier推定量

死亡が起こった時点で区切った区間の生存率

その積

例.

死亡あるいは打ち切り時点

$\underline{1}\ \ 3\ \ \underline{4}\ \ 5\ \ 5\ \ \underline{6}\ \ 7\ \ 7\ \ \underline{7}\ \ \underline{8}$

時点	個体数	死亡数	区間の生存率
0	10	0	1.0
3	9	1	0.89
5	7	2	0.71
7	4	2	0.5

Kaplan－Meier推定量

死亡が起こった時点で区切った区間の生存率

例.

時点	区間の生存率	最初を基準にした生存率
0	1.0	1.0
3	0.89	0.89
5	0.71	0.63
7	0.5	0.32

パッケージsurvivalの関数survfit

logrank検定

２つのサンプル（処理が異なる間）での
生存曲線の比較）

拡張

３つ以上のサンプル

ブロック（層別）あり

logrank検定

死亡が起こった時点で区切り、

両サンプルの生存個体数、死亡個体数を計算

サンプル1の生存個体数	サンプル1の死亡個体数
サンプル2の生存個体数	サンプル２の死亡個体数

複数の2×2分割表

↓

Mantel-Haentzel検定

logrank検定

区間の数だけ　　2×2分割表

サンプル1の生存個体数	サンプル1の死亡個体数
サンプル2の生存個体数	サンプル２の死亡個体数

Mantel-Haenszel検定　　ブロック内の相関

共通の傾向（共通オッズ比）

死亡の順序（生存時間の順序）

パッケージsurvivalの関数survdiff

Wilcoxon順位和検定の拡張版
(Mann-WhitneyのU検定)

２つのサンプル　生存日数の比較

時間データは順序尺度も使える

確率分布　distribution

正（正かゼロ）　連続量

指数分布

ワイブル分布

ガンマ分布　アーラン分布

カイ二乗分布

対数正規分布

ロジスティック分布

対数ロジスティック分布

ゴンペルツ分布

正規分布

確率分布

ヒストグラム ―― 確率密度関数

ある値以下のデータの割合 → 分布関数

確率密度関数の積分

ある値より大きいデータの割合

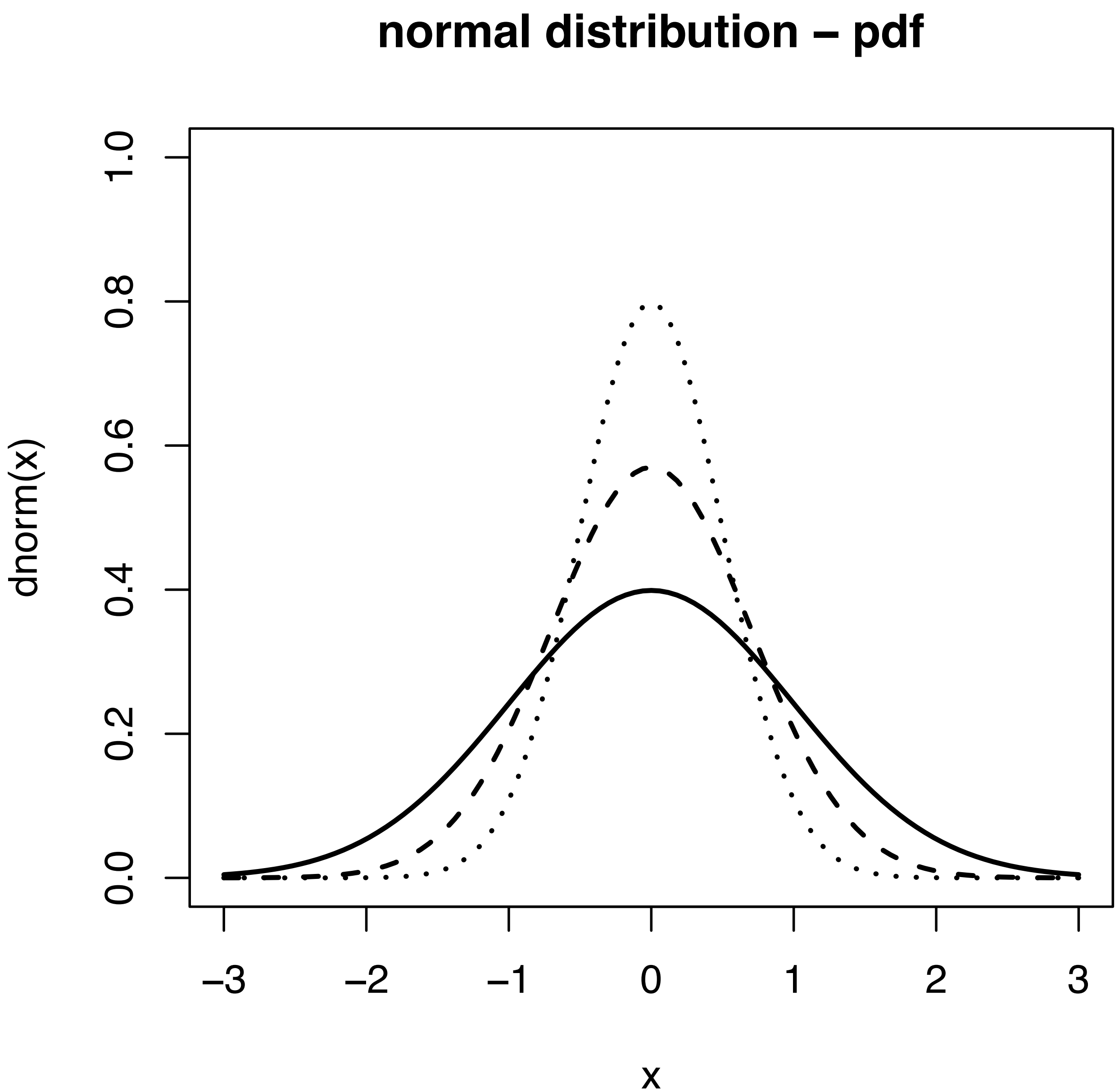
→ **生存関数** =(１−分布関数)

そのデータが生存時間なら生存率になる

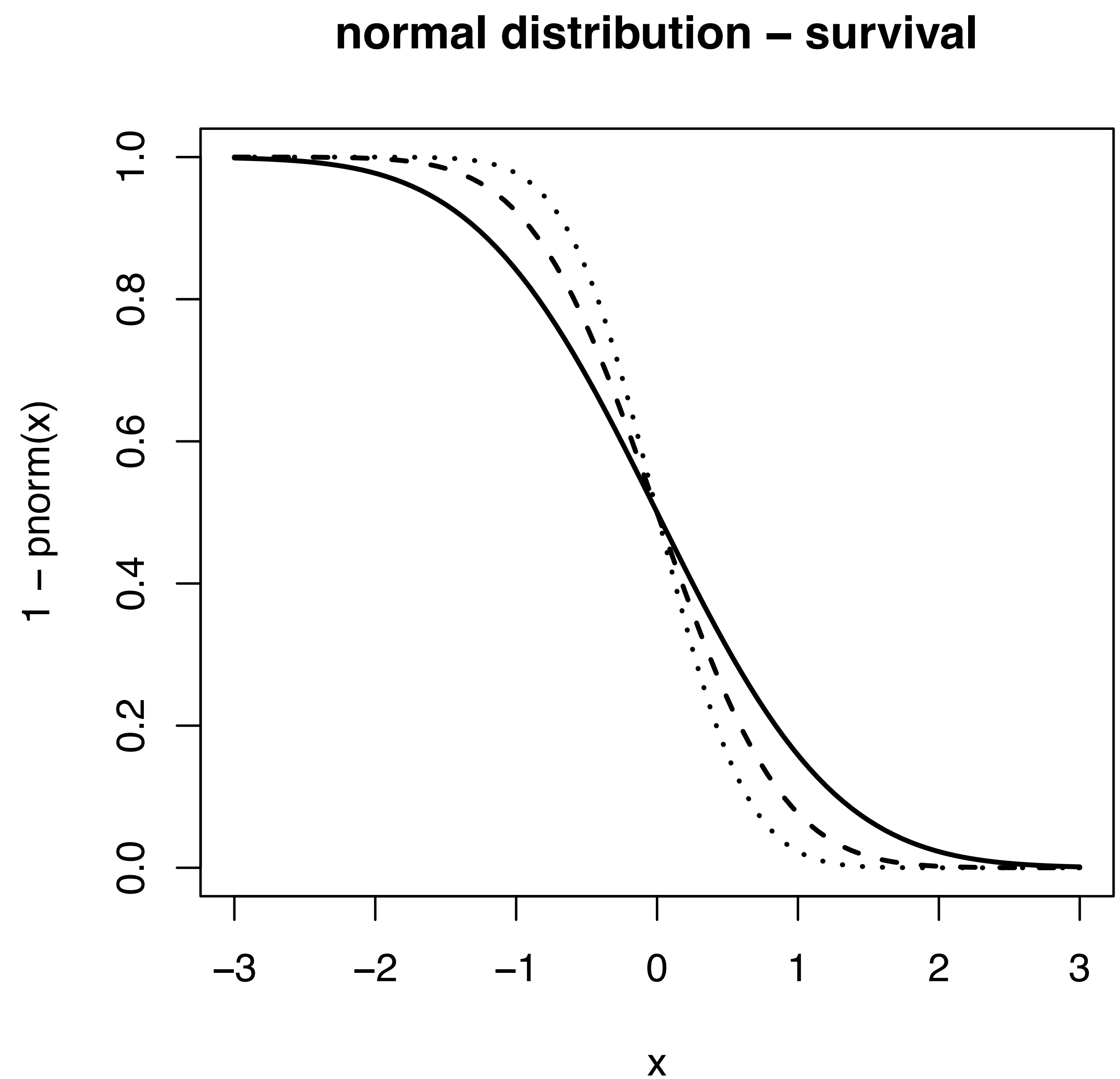
生存曲線

正規分布(平均=0)の場合

確率密度関数



生存関数



ハザード　hazard　(危険度)

✖危険率
✖検定での有意水準

時点yで生存していた個体の瞬間的死亡率

$$h(y) = \frac{f(y)}{S(y)}$$

確率密度関数

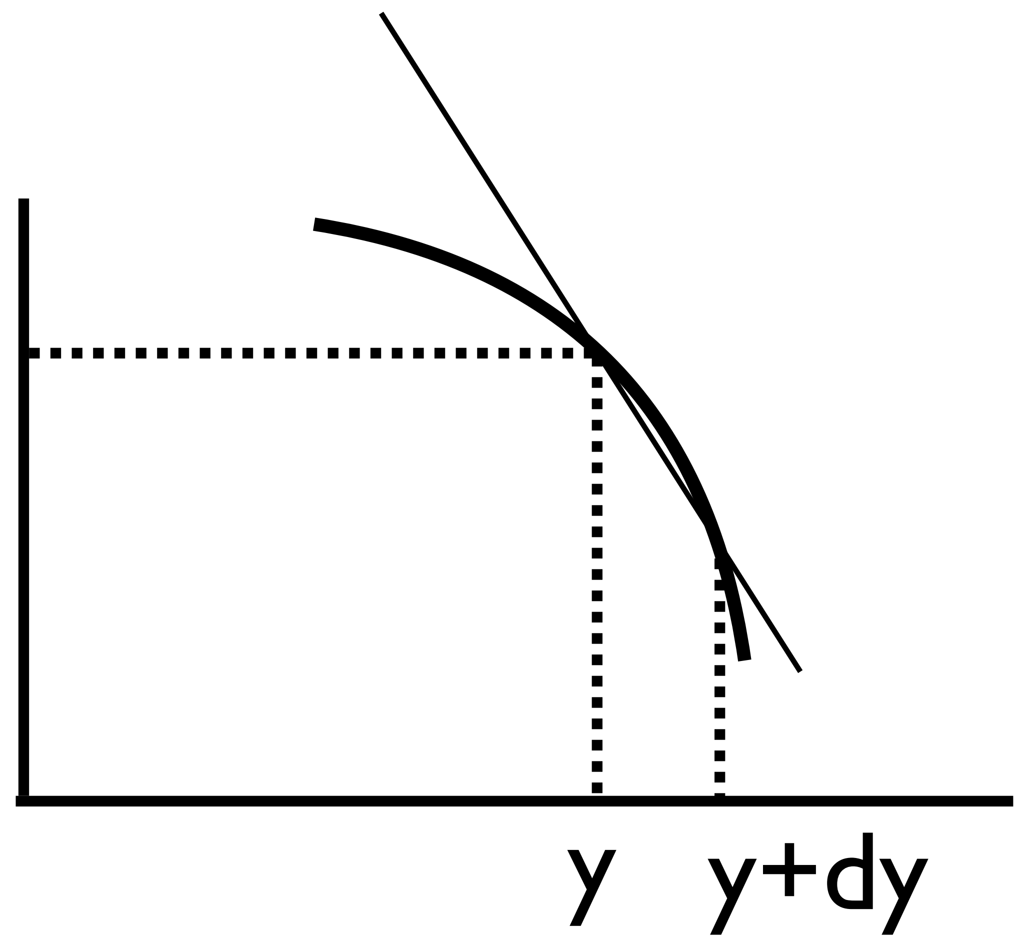
ハザード関数

生存関数

ハザード hazard (危険度)

時点yで生存していた個体の死亡率

生存関数
$S(y)$

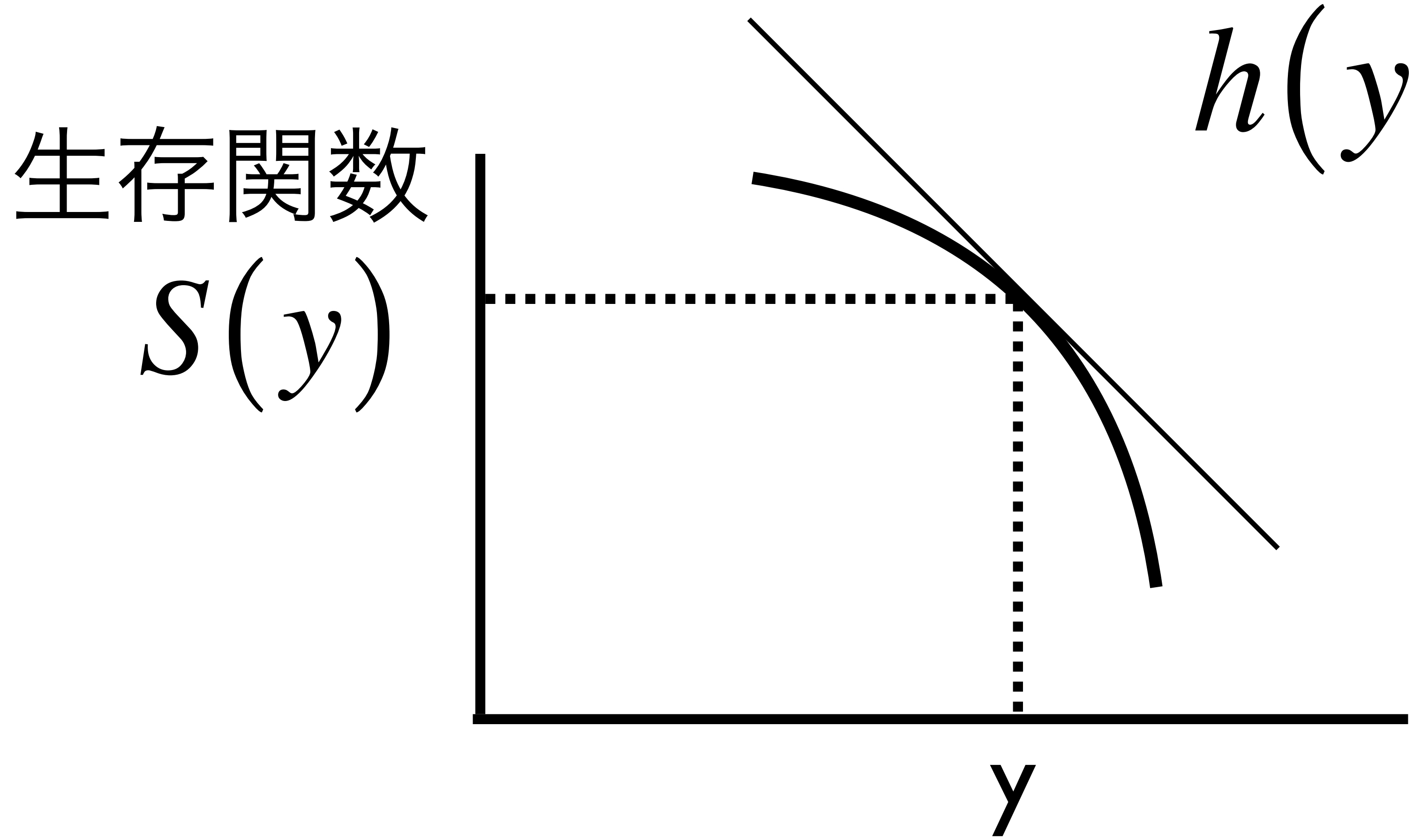


$$-\frac{\text{線の傾き}}{S(y)}$$

dy→0
時点yで生存していた個体の瞬間的死亡率

ハザード hazard (危険度)

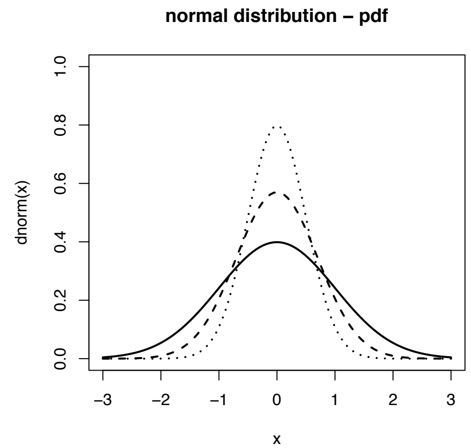
$$h(y) = -\frac{\text{接線の傾き}}{S(y)}$$



時点yで生存していた個体の瞬間的死亡率

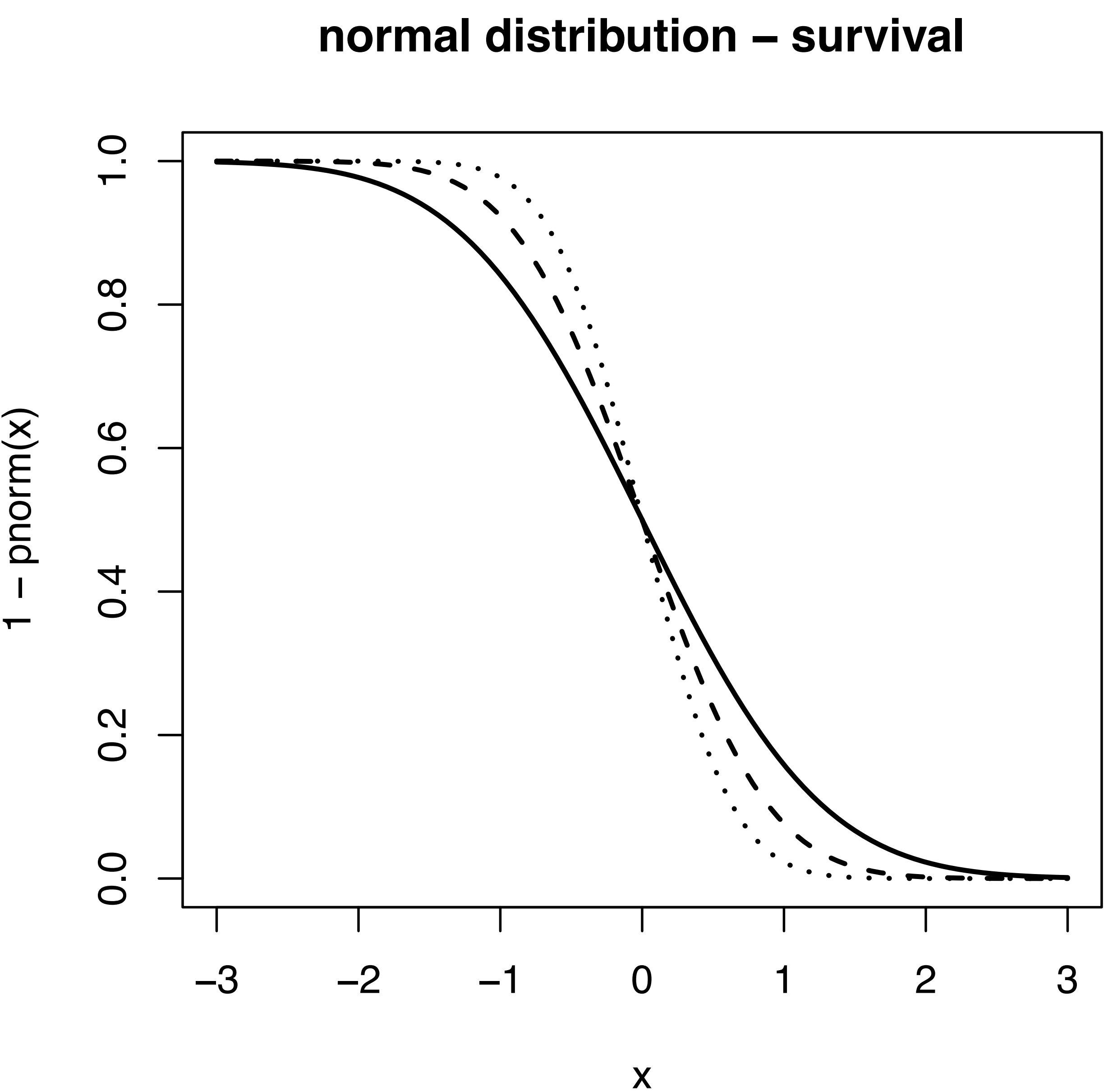
$$h(y) = \frac{f(y)}{S(y)}$$

正規分布(平均=0)の場合

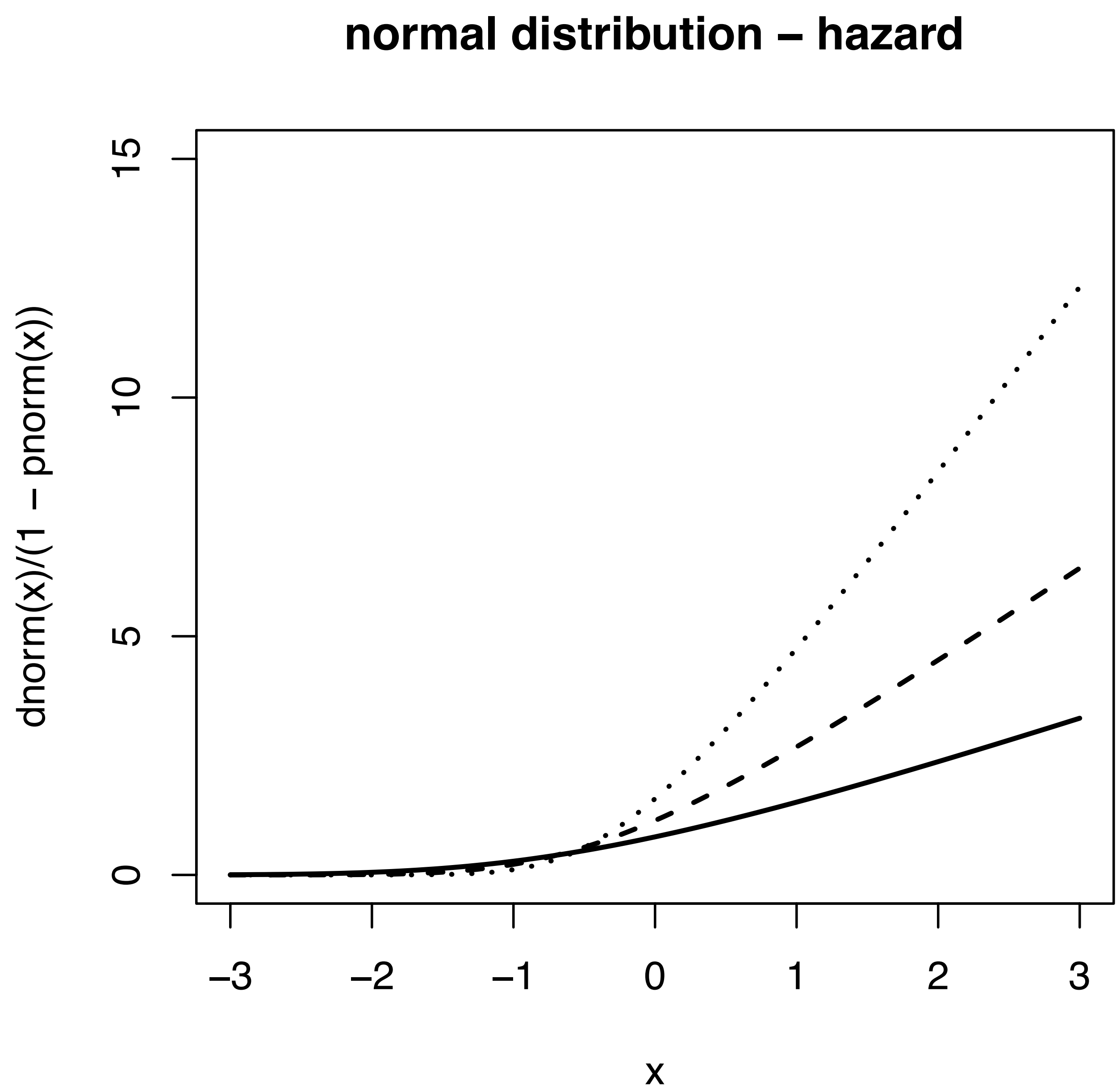


確率密度関数

生存関数



ハザード関数



指数分布

確率密度関数

$$f(y) = \lambda e^{-\lambda y}$$

生存関数

$$S(y) = e^{-\lambda y}$$

ハザード関数

$$h(y) = \lambda$$

ポアソン過程　起こる確率が一定

指数分布

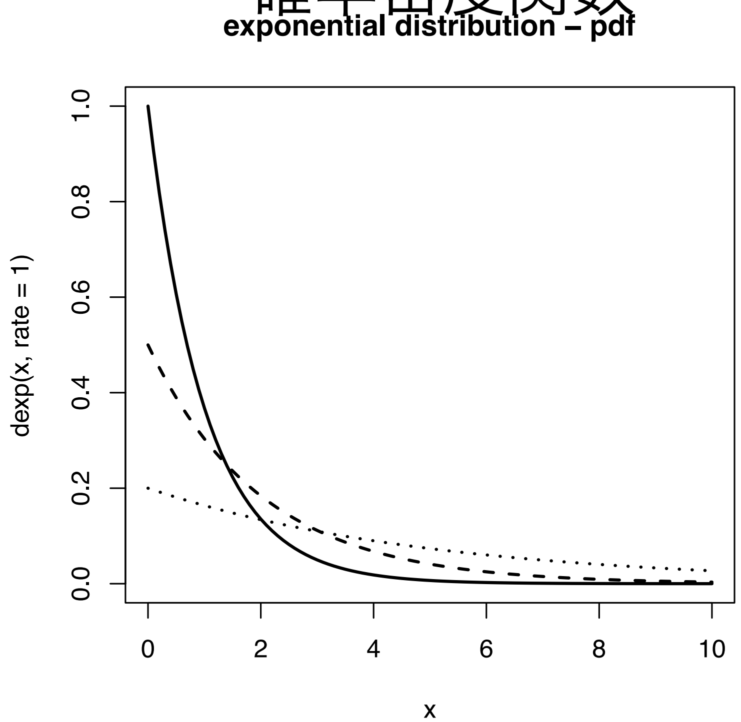
状態の終了が起こる確率が一定のときの、状態持続時間

瞬間的な事象の起こる時間間隔

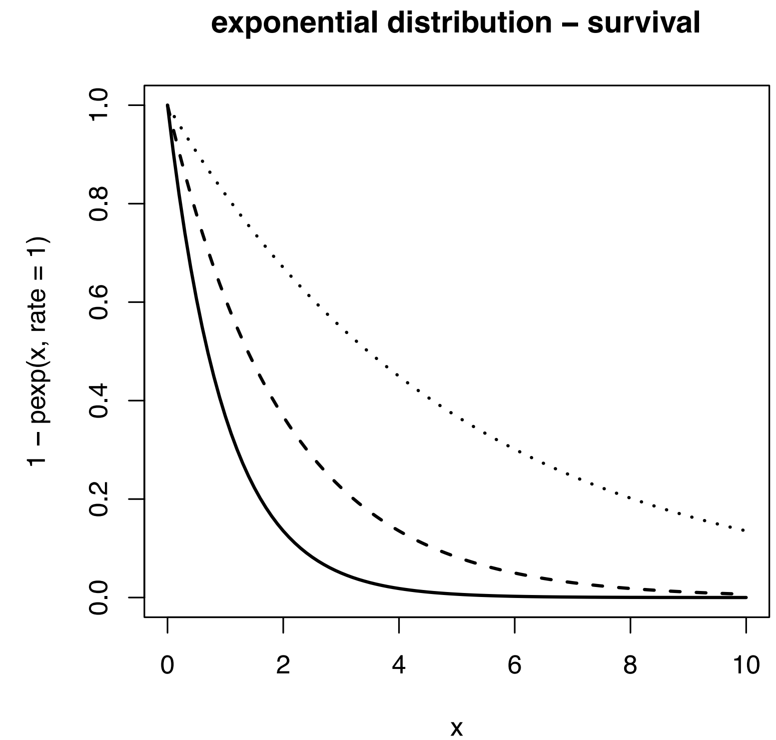
指数分布　生存率一定の生存曲線

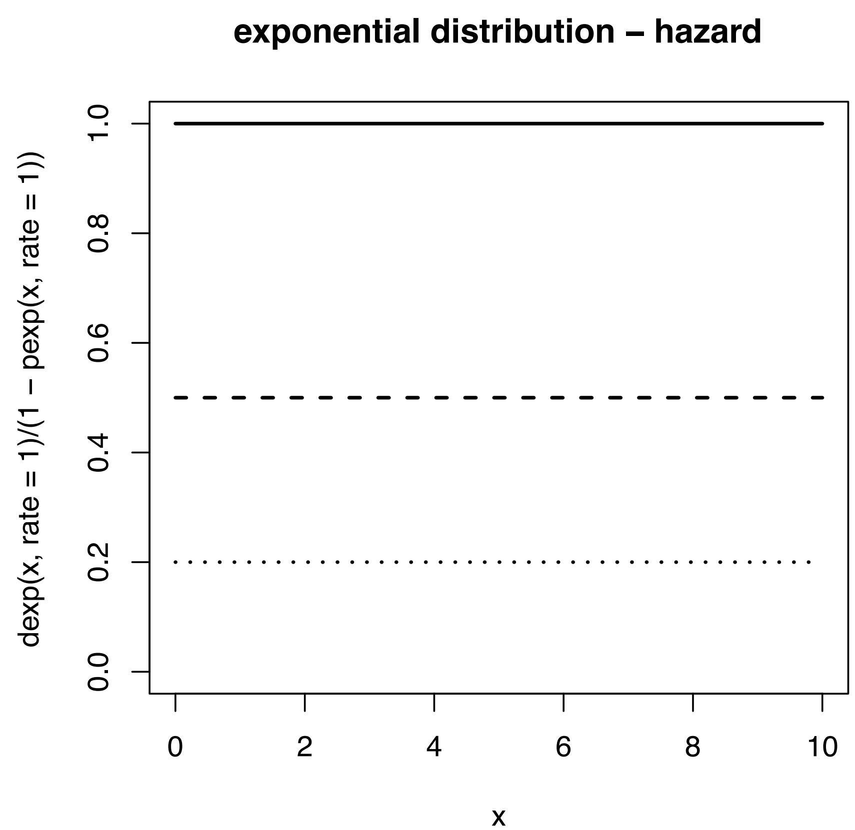
指数分布

確率密度関数



生存関数

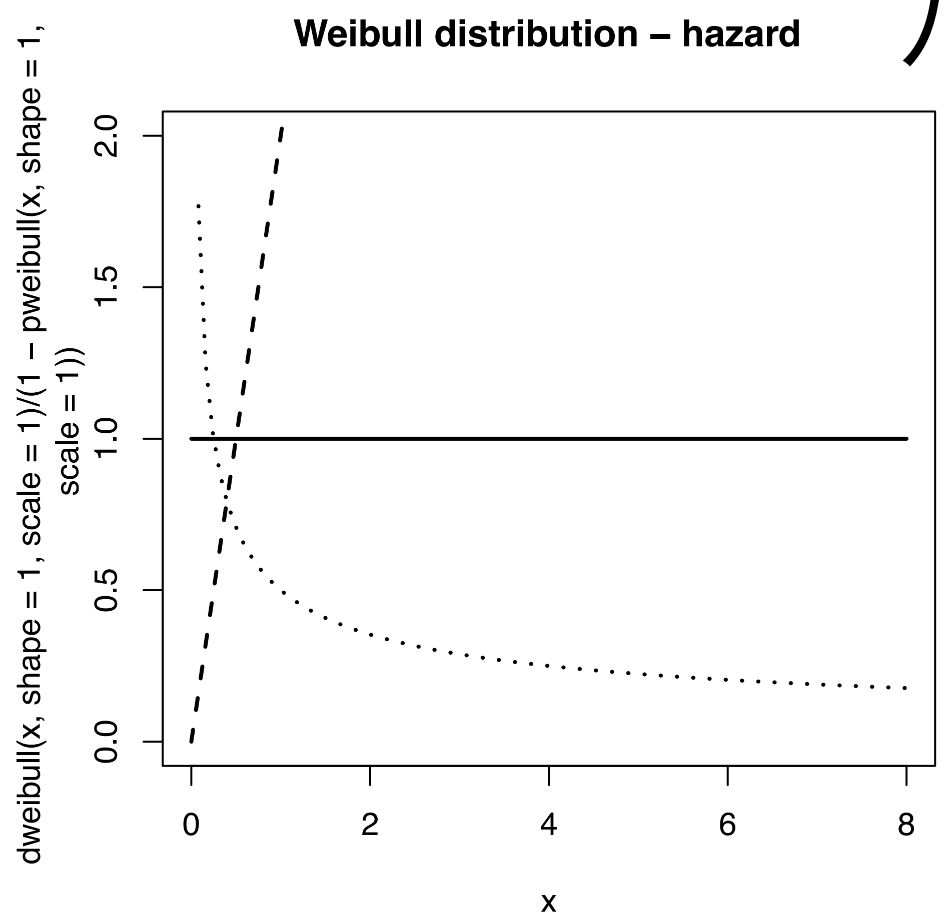




ハザード関数

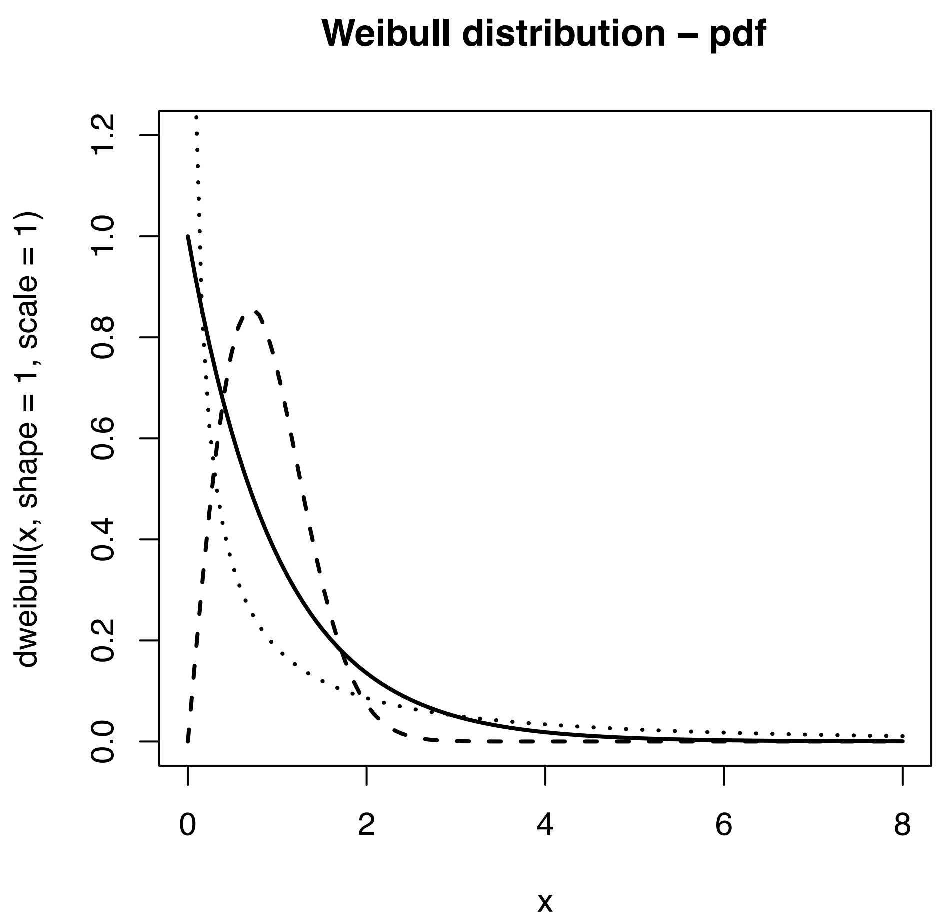
ワイブル分布

ハザード関数

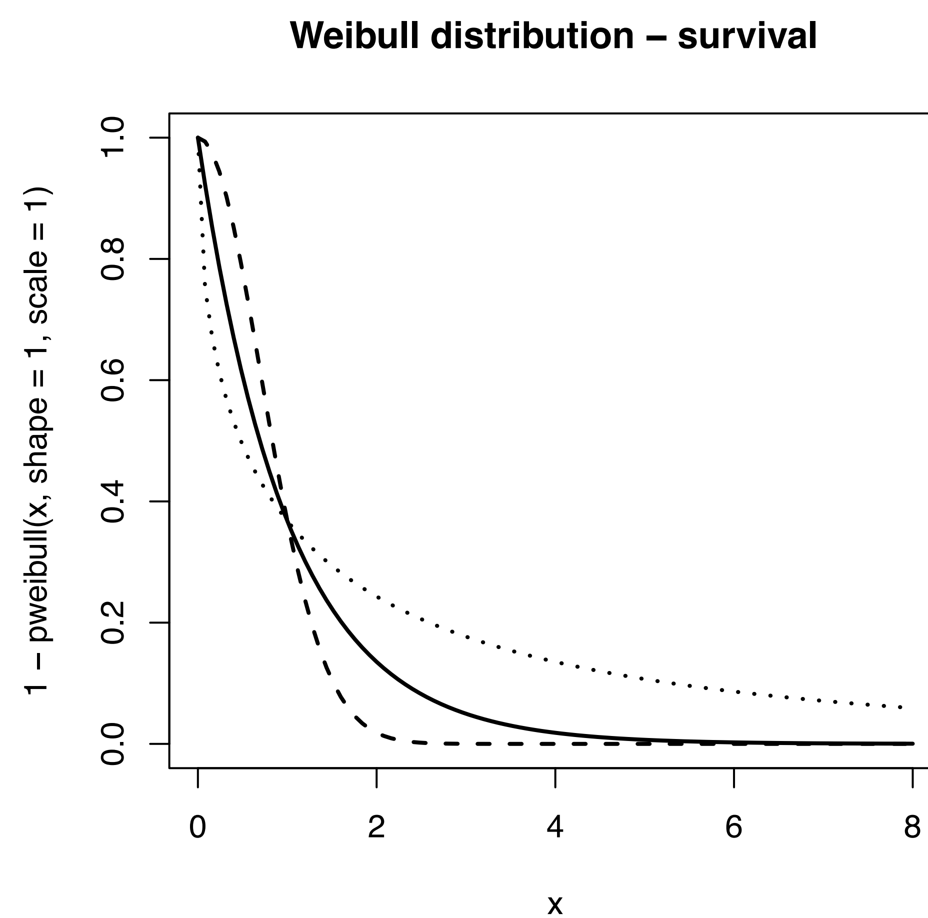


パラメーターにより、
ハザードが時間的変化

確率密度関数

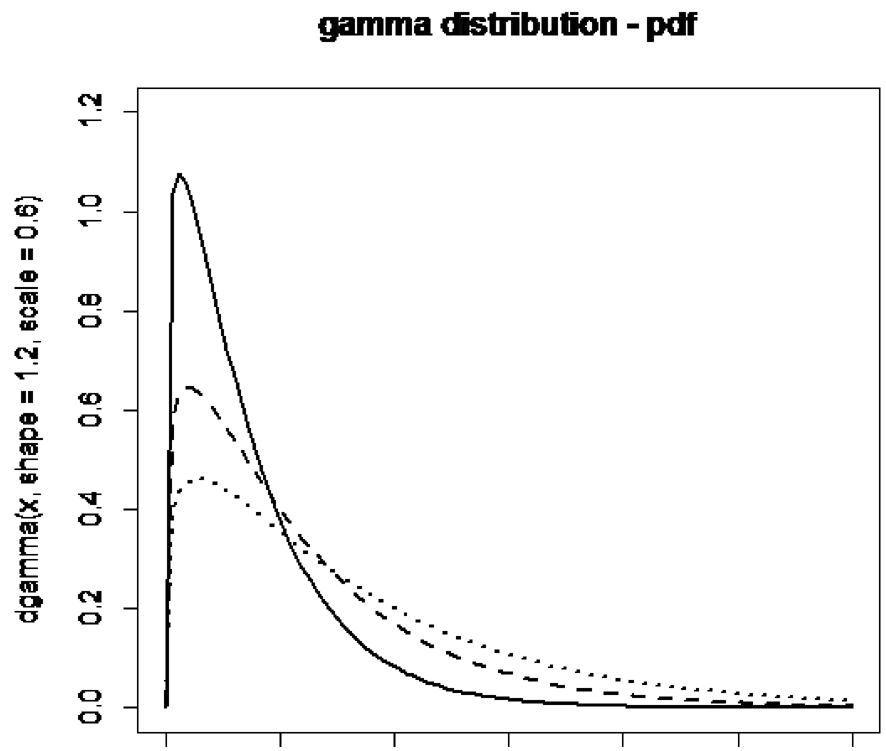


生存関数

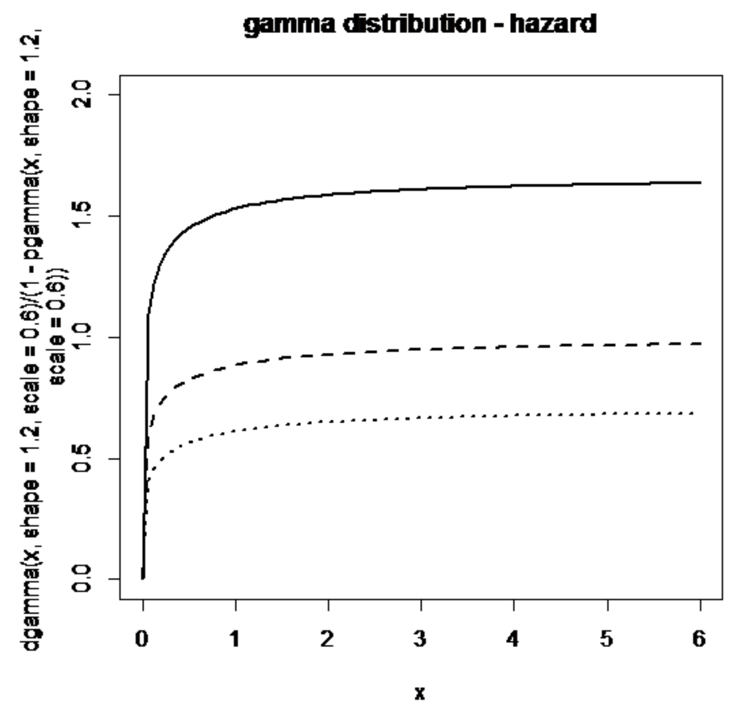


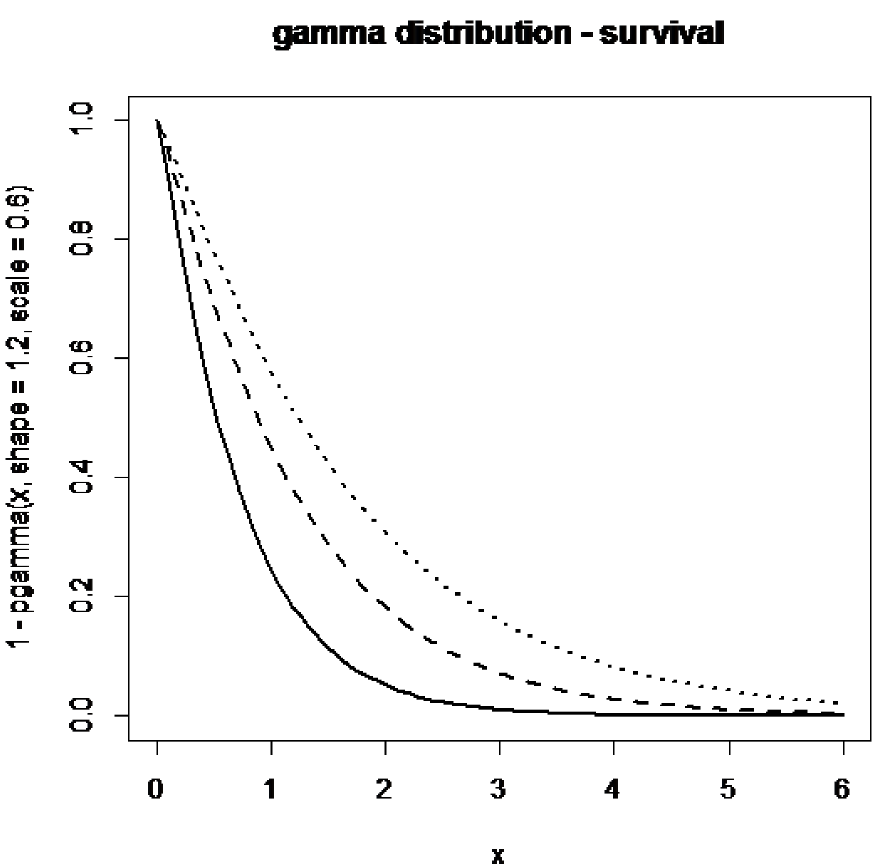
ガンマ分布

確率密度関数



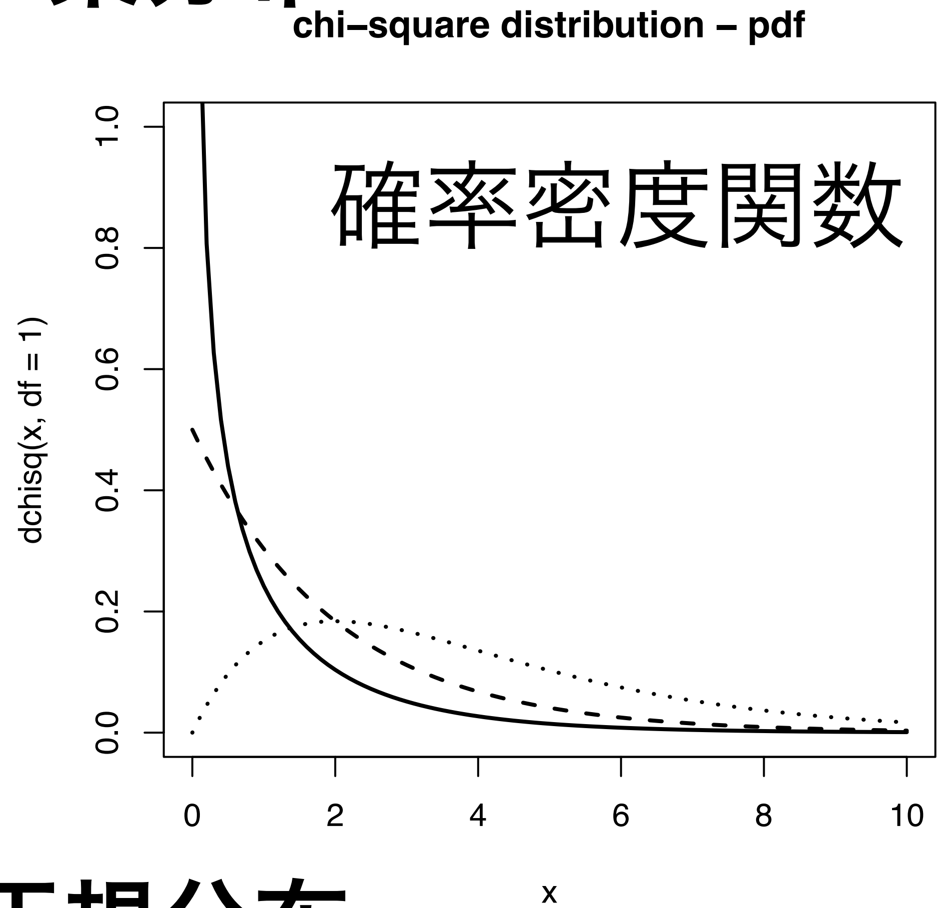
ハザード関数

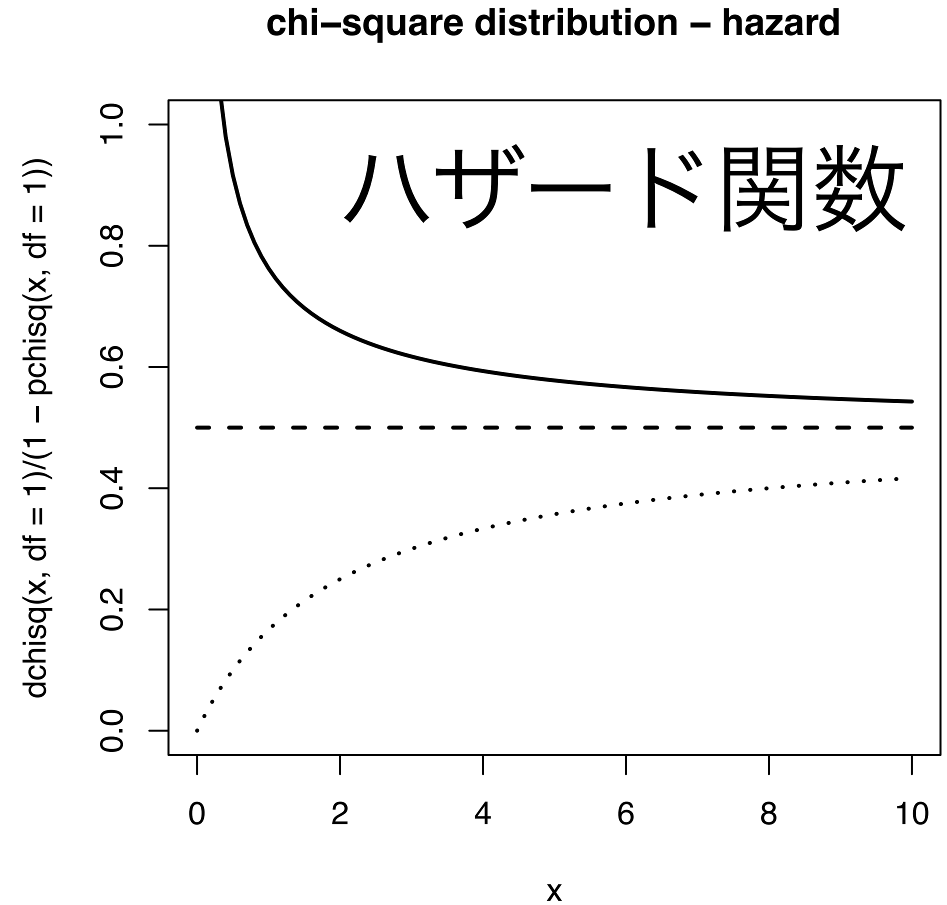




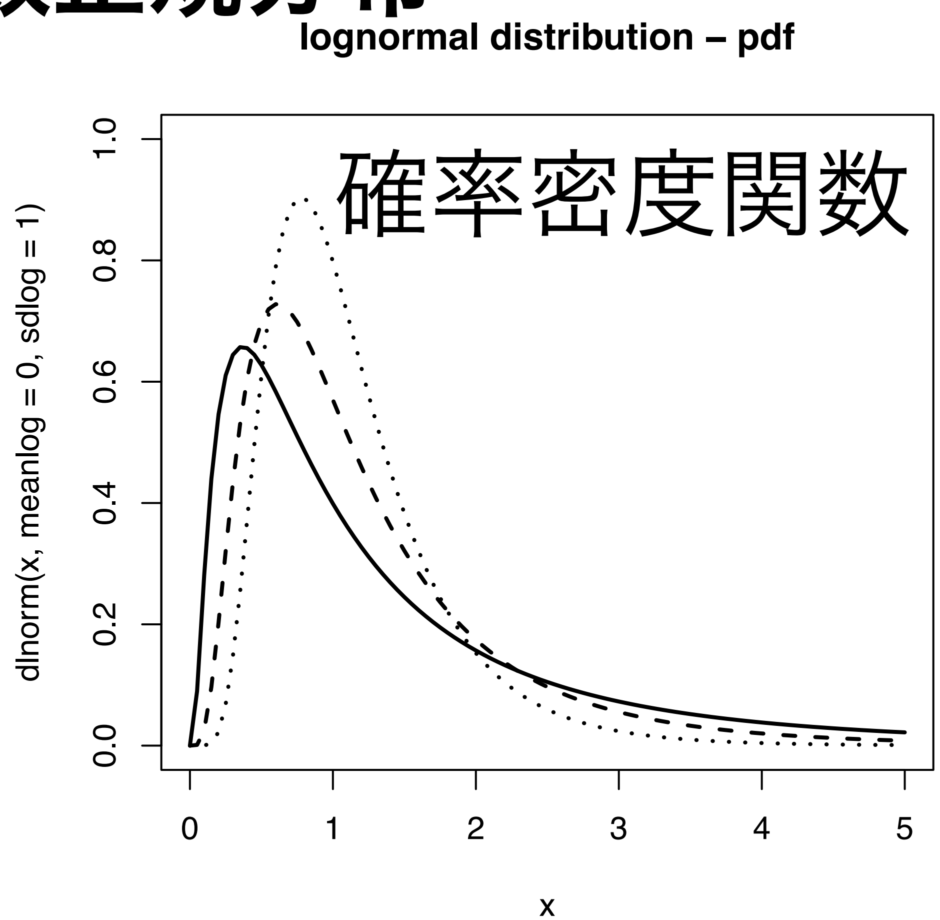
生存関数

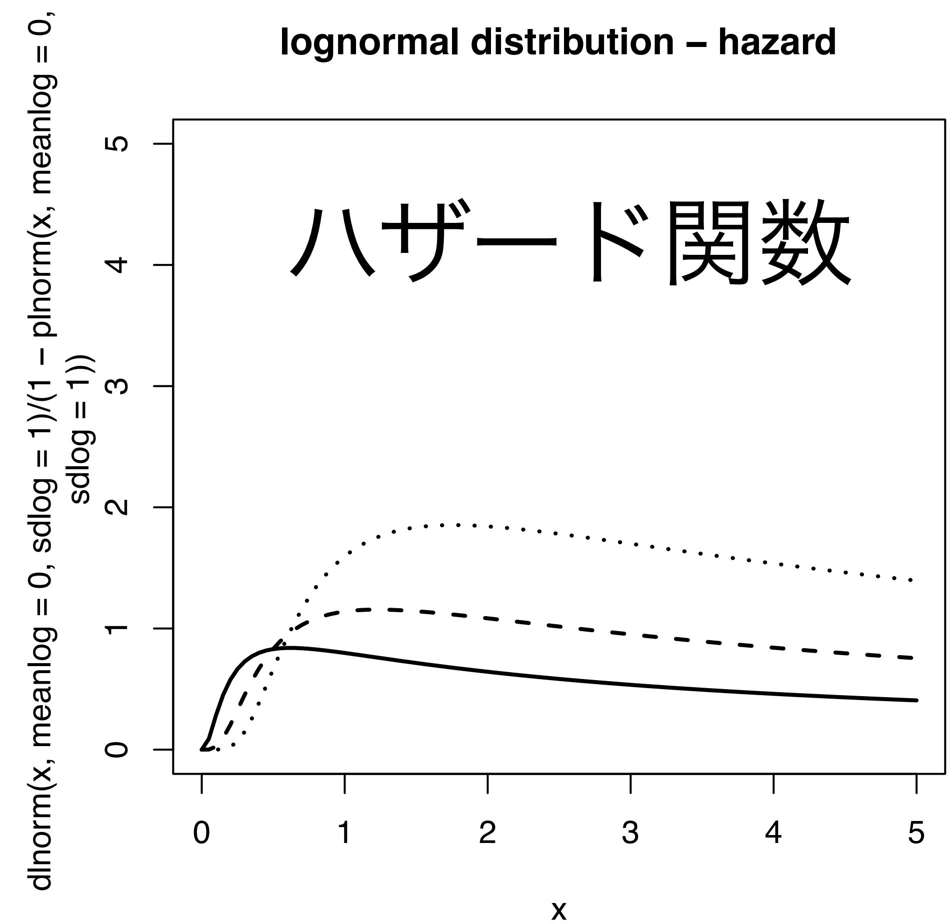
カイ二乗分布



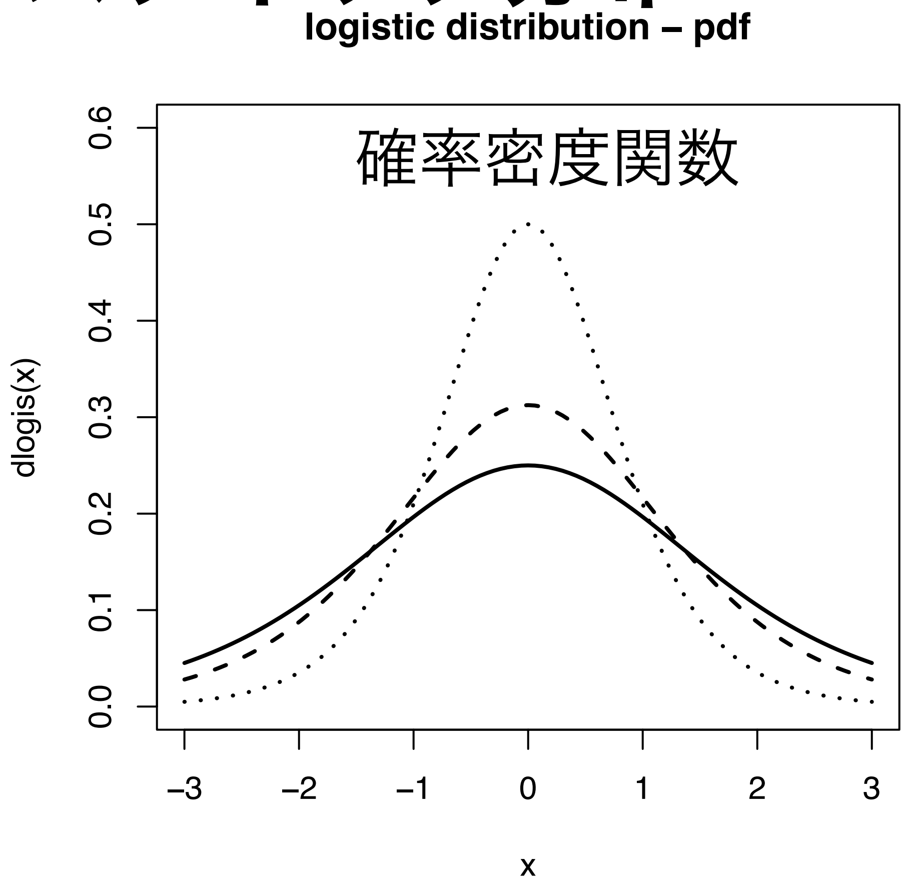


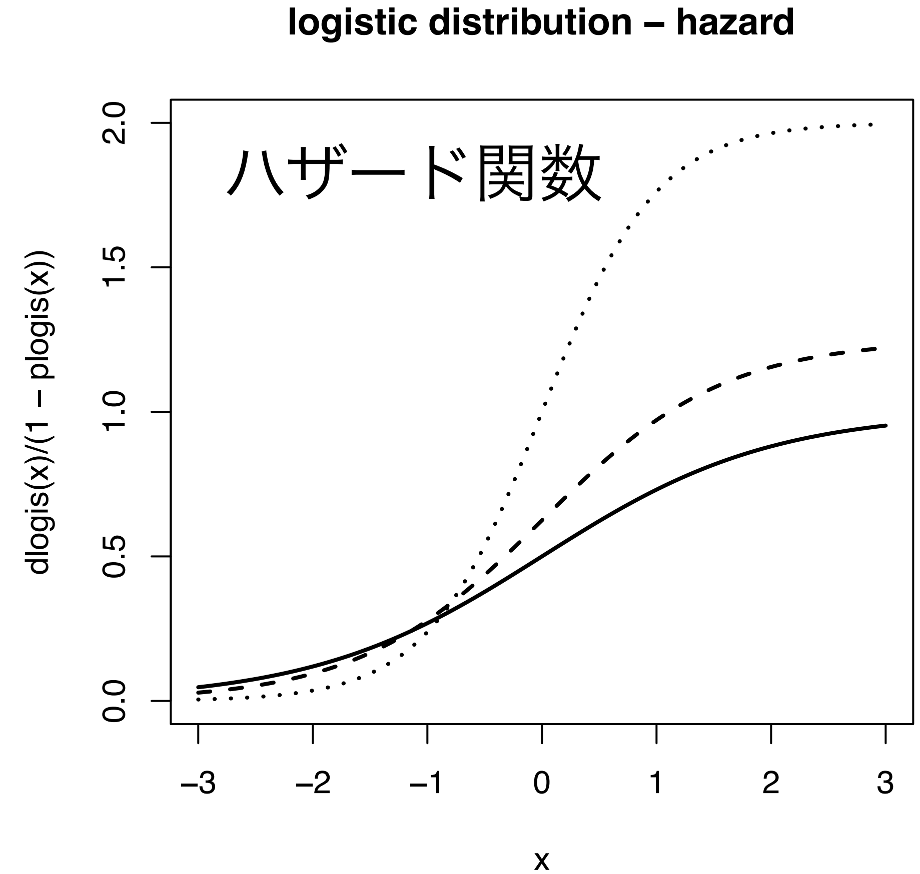
対数正規分布



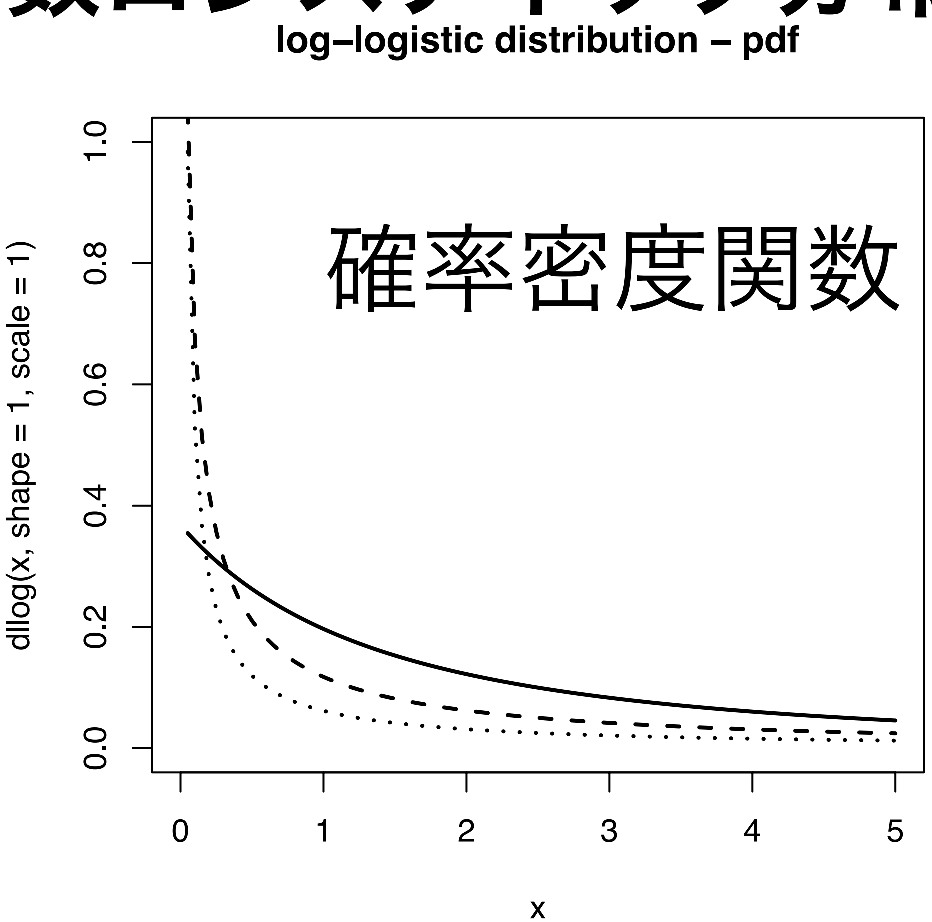


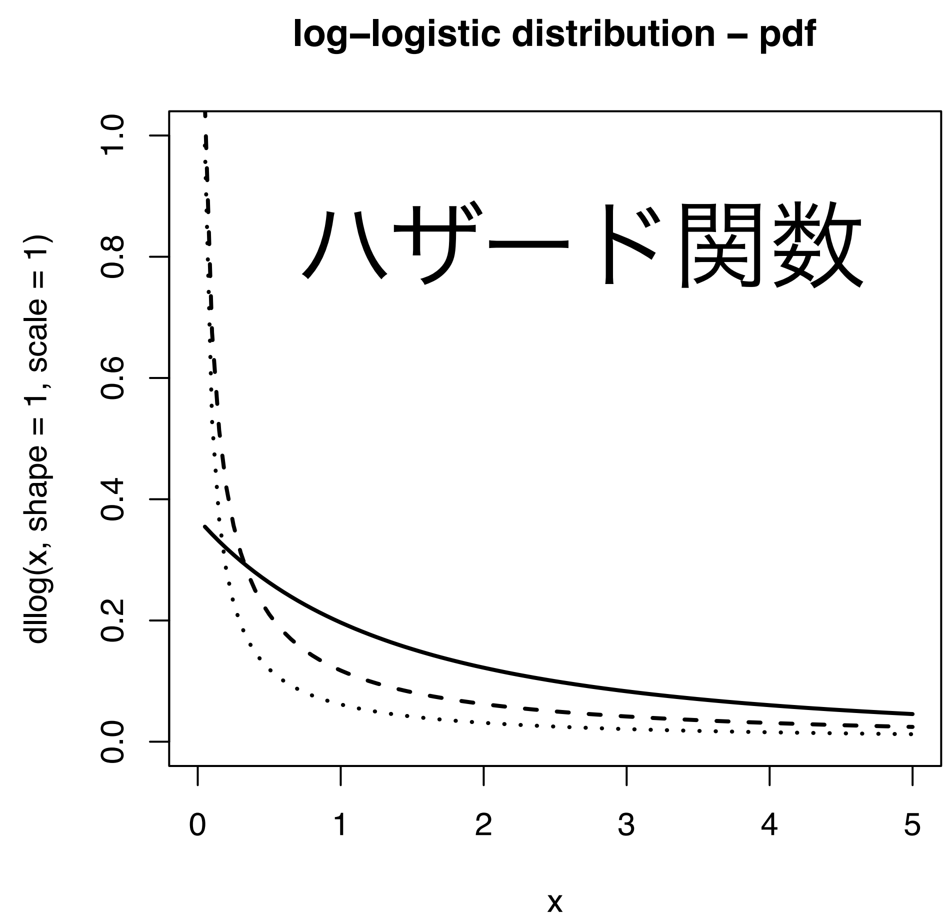
ロジスティック分布



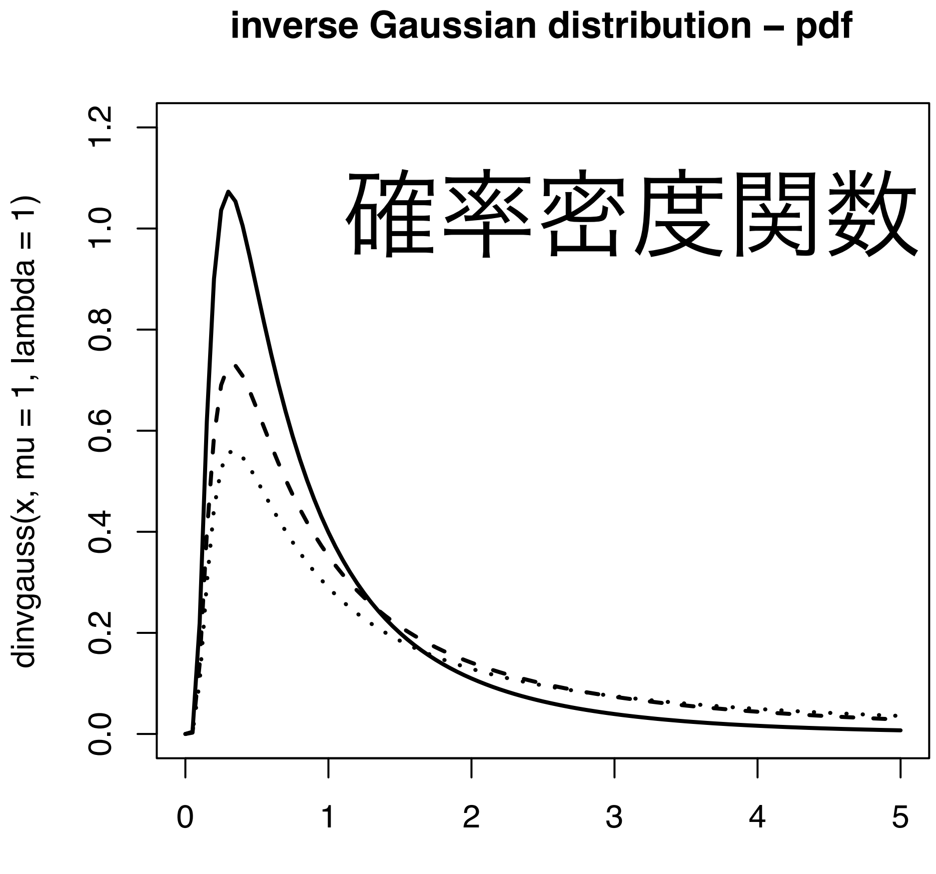


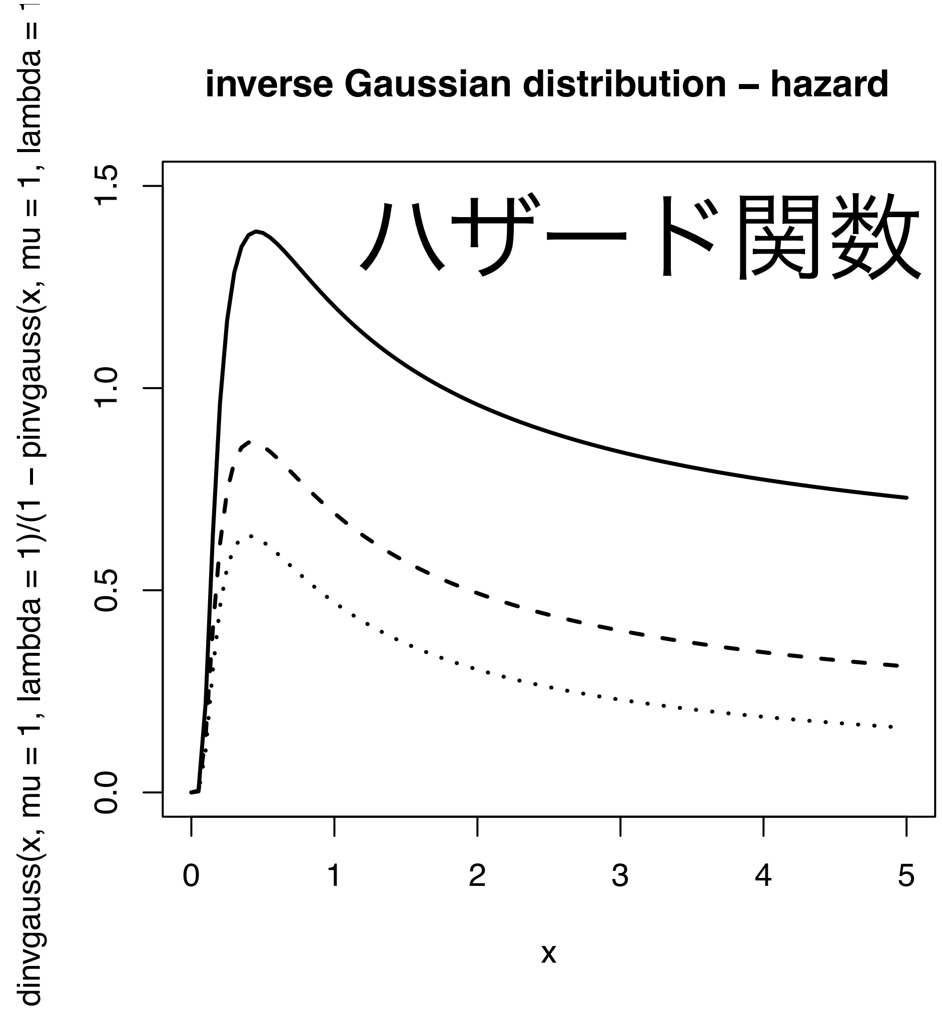
対数ロジスティック分布



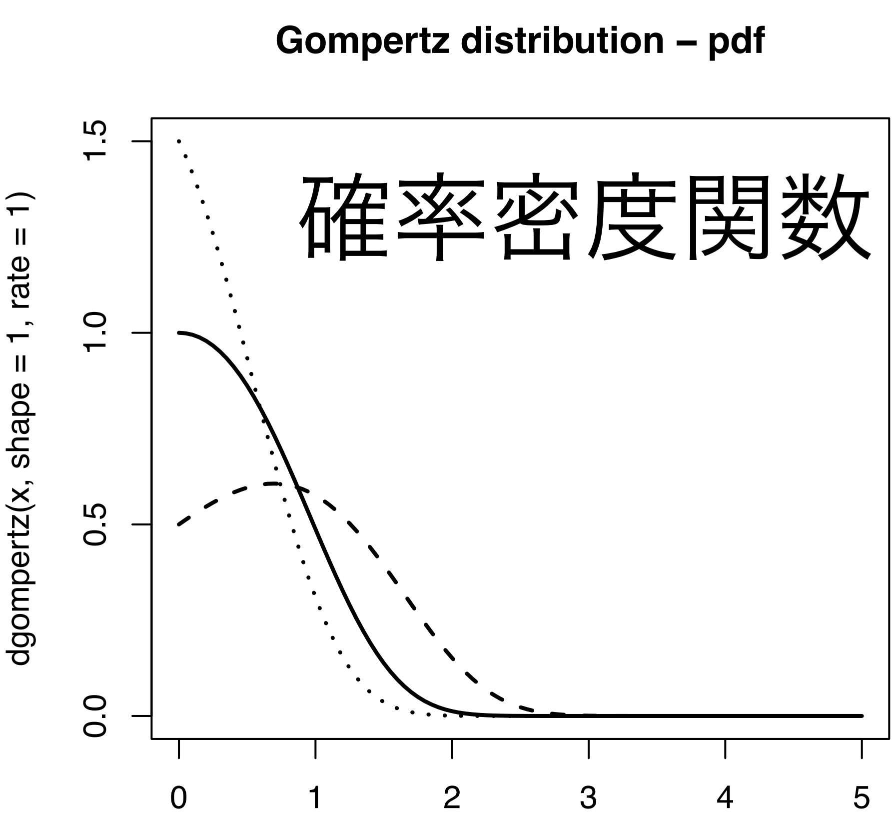


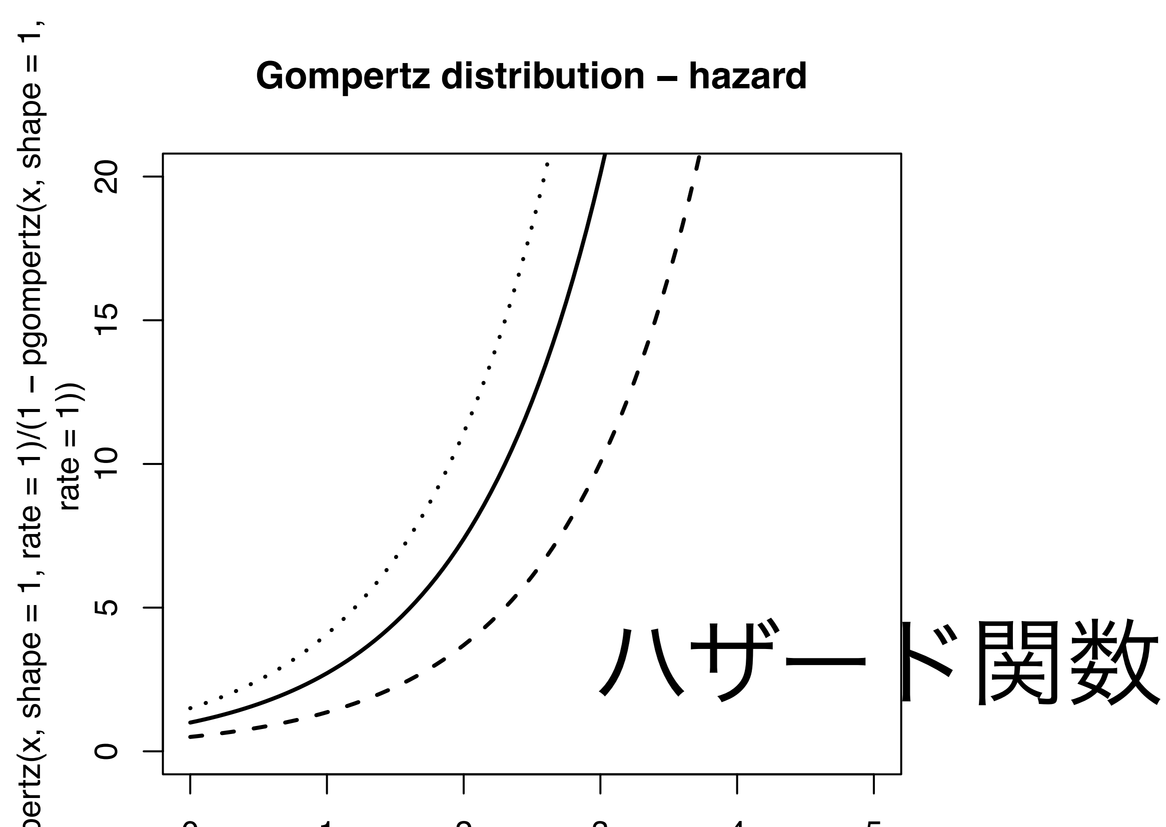
逆ガウス分布





ゴンペルツ分布



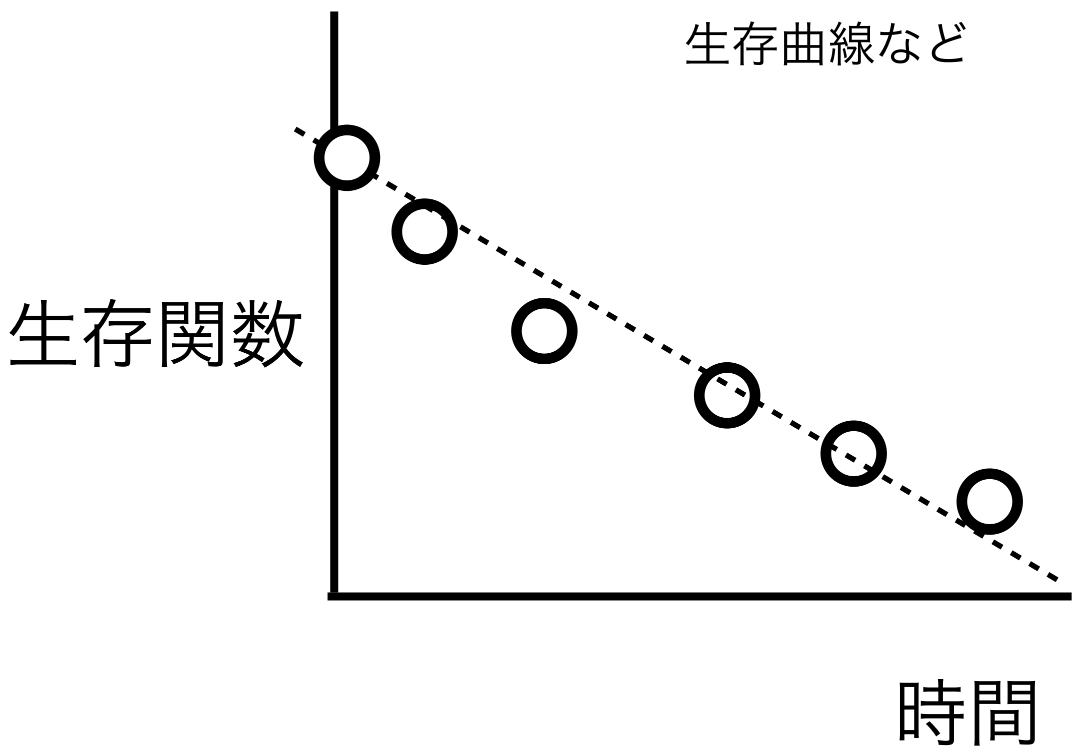


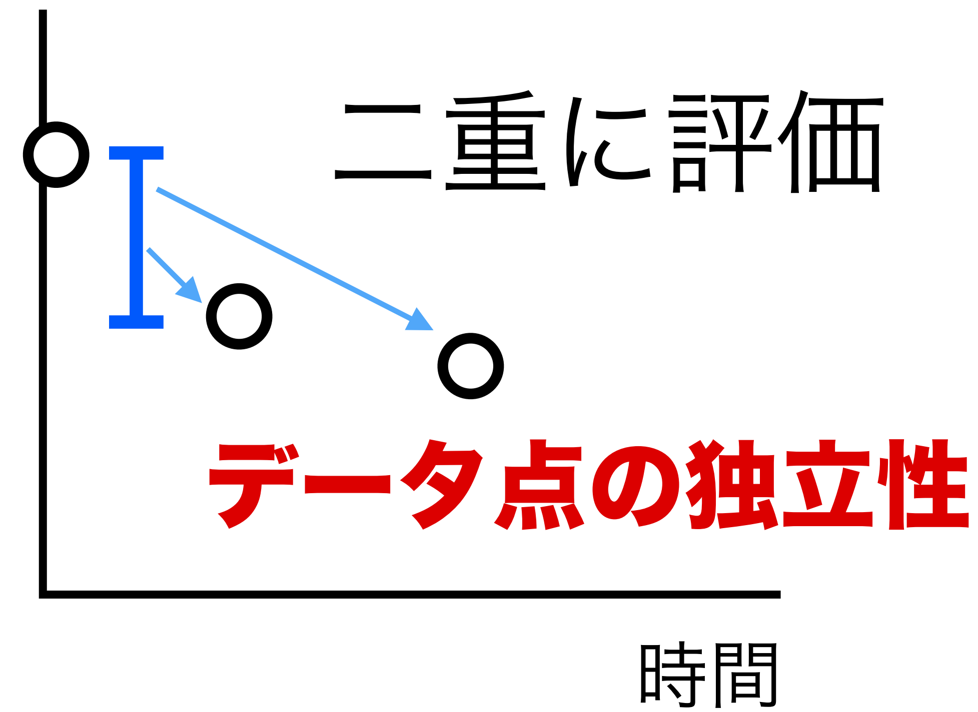
ある値より大きいデータの割合

生存関数 $S(y)=1-F(y)$

目的変数＝生存関数　説明変数＝時間など

という回帰的な分析 **要注意**





比例ハザードモデル

proportional hazard model

Cox回帰

http://www.stats.ox.ac.ukより

D.R.Cox
(1924–)

比例ハザードモデル

(1972) J.R.Statist.Soc.B34:187-220.

パラメトリック・ブートストラップ

(1961) Proc. 4th Berkeley Symposium, 105-123.

(1962) J. R. Statist. Soc. B 24, 406-424.

(2013) J. R. Statist. Soc. B75: 207-215.

http://www.stats.ox.ac.ukより

D.R.Cox
(1924–)

比例ハザードモデル

(1972) J.R.Statist.Soc.B34:187-220.

家畜などのウシ型結核菌感染とアナグマ

D. R. Cox et al. (2005) PNAS 102:17588–17593

J.R.Krebs
T.Clutton-Brock
R.Anderson
など

パラメトリック・ブートストラップ

(1961) Proc. 4th Berkeley Symposium, 105-123.
(1962) J. R. Statist. Soc. B 24, 406-424.

(2013) J. R. Statist. Soc. B75: 207-215.

比例ハザードモデル

proportional hazard model　Cox回帰

中村『(1980年のこと)初めて聞く名前であった。…どうせ重回帰の変法だろうと想像して……だが届いた論文と原論文(Cox,1972)を読んでみても、まったく理解できず困ってしまった』

比例ハザード性

説明変数　　　　　　共変量(covariate)とも呼ぶ

説明変数が $\mathbf{x}$ の個体の時刻tにおけるハザード

$$\lambda(t|\mathbf{x}) = \lambda(t|\mathbf{0}) \cdot r(\mathbf{x})$$

相対危険度

ベースラインハザード

説明変数がすべて0の個体の、時刻tにおけるハザード

比例ハザード性

説明変数　　　　　　　　　　　　共変量(covariate)とも呼ぶ

説明変数が $\mathbf{x}$ の個体の時刻tにおけるハザード

$$\lambda\left(t \middle| \mathbf{x}\right) = \lambda\left(t \middle| \mathbf{0}\right) \cdot r\left(\mathbf{x}\right)$$

対数線形性

$$r\left(\mathbf{x}\right) = \exp\left(\beta^T \mathbf{x}\right) = \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots\right)$$

比例ハザードモデル

説明変数が $\left(x_1, x_2, \cdots\right)$ の個体の時刻tにおけるハザード

$$\lambda\left(t \middle| x_1, x_2, \cdots\right) = \lambda_0\left(t\right) \cdot \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots\right)$$

$\lambda_0\left(t\right)$　ベースラインハザード

比例ハザードモデル

説明変数が $\left(x_1, x_2, \cdots\right)$ の個体の時刻tにおけるハザード

時間の要素　　説明変数の効果

$$\lambda\left(t \middle| x_1, x_2, \cdots\right) = \lambda_0(t) \cdot \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots\right)$$

$\lambda_0(t)$　ベースラインハザード

比例ハザードモデルの仮定

説明変数が異なる個体のハザード

比例ハザード性

ハザードの比　時刻tにかかわらず
いつでも同じ

比例ハザード性＋対数線形生

ハザードの比　時刻tにかかわらず
いつでも同じで、
説明変数の一次式の指数関数

比例ハザードモデルの仮定

説明変数が異なる個体（個体たち）の生存関数

$$S\left(t \middle| x_1, x_2, \cdots\right) = S_0\left(t\right)^{\exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots\right)}$$

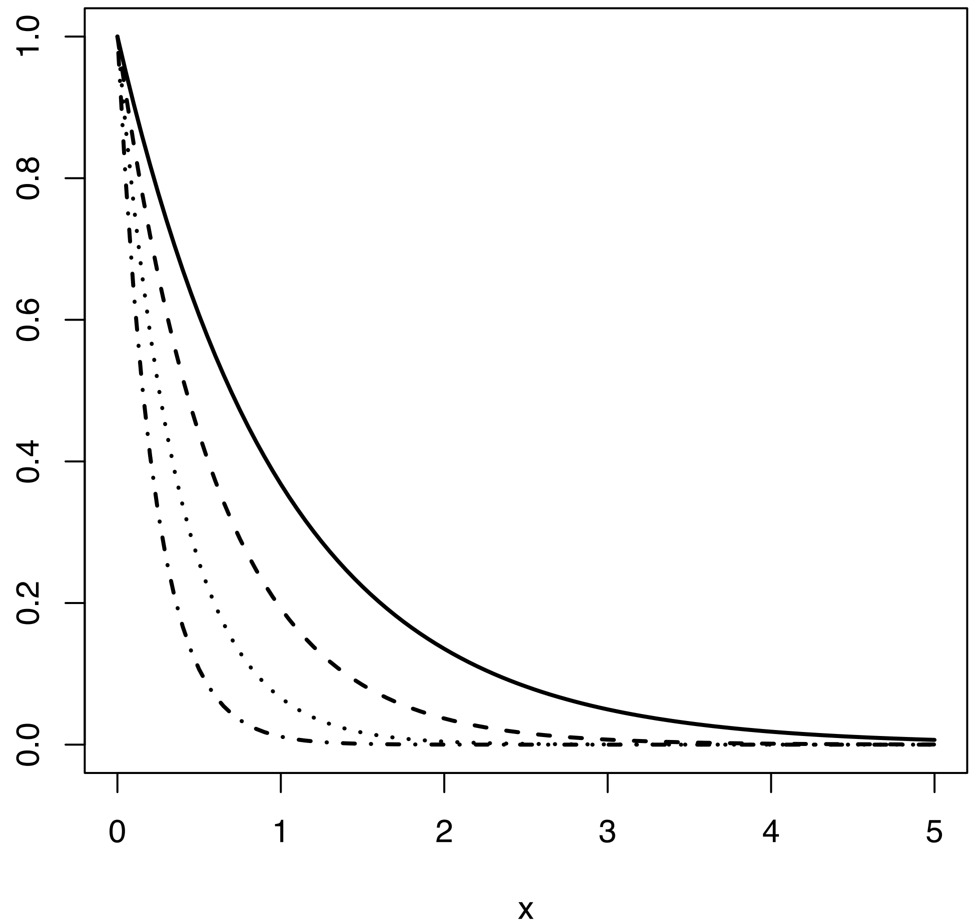
生存関数の対数の比は一定

$$\log S\left(t \middle| x_1, x_2, \cdots\right) = \left(\beta_1 x_1 + \beta_2 x_2 + \cdots\right) \cdot \log S_0\left(t\right)$$

生存曲線は交わらない

比例ハザードモデルの仮定が成り立っているとき

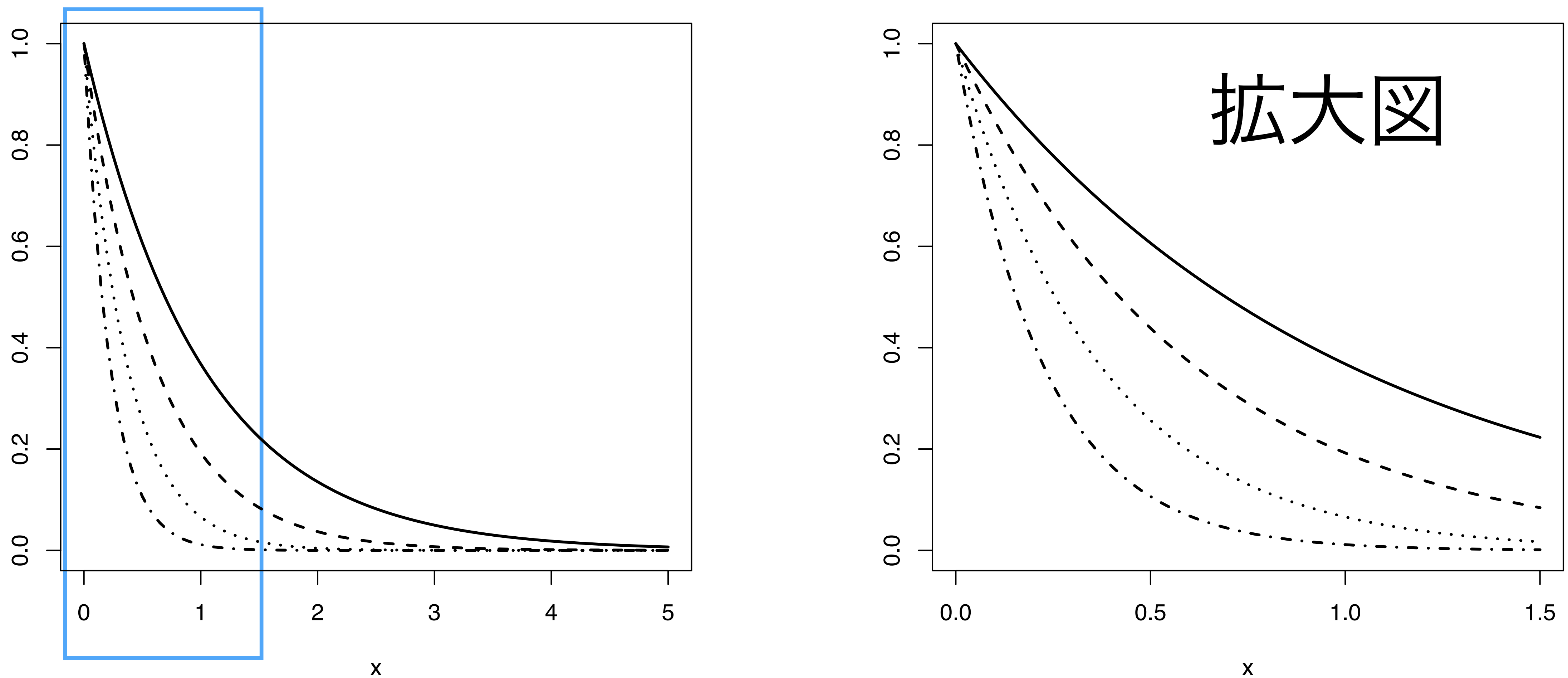
生存率



時間

比例ハザードモデルの仮定が成り立っているとき

生存率



時間

時間

比例ハザードモデルの仮定

説明変数が異なる個体（個体たち）の生存関数

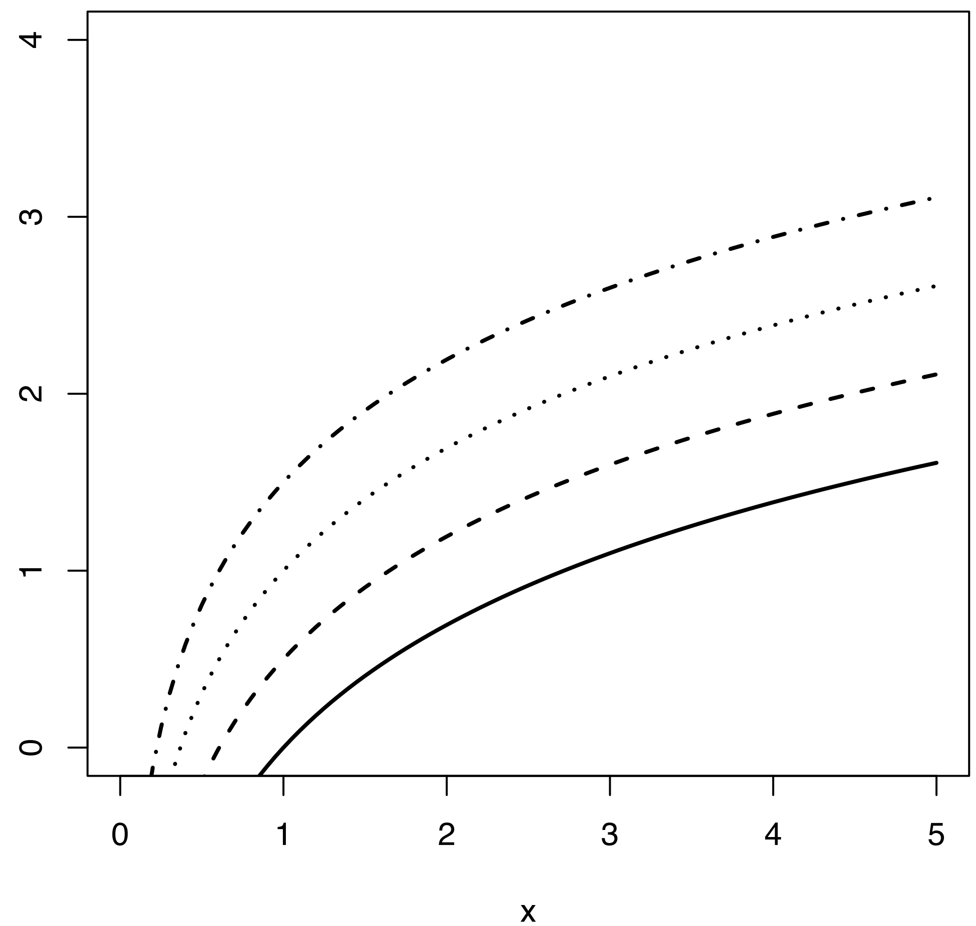
$$\log\left\{-\log S\left(t \middle| x_1, x_2, \cdots\right)\right\} = \left(\beta_1 x_1 + \beta_2 x_2 + \cdots\right) + \log\left\{-\log S_0\left(t\right)\right\}$$

生存関数の対数の符号をかえたものの対数を
時間に対してプロット

形が同じで平行移動したもの

二重対数プロット

log{-log(生存率)}



時間

比例ハザードモデル

説明変数が $(x_1, x_2, \cdots)$ の個体の時刻tにおけるハザード

$$\lambda(t|x_1, x_2, \cdots) = \boxed{\lambda_0(t)} \cdot \boxed{\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots)}$$

こちらには手をつけず　こちらだけ最尤法で推定

（部分尤度、partial likelihood）

比例ハザードモデル

生存関数や生存曲線の計算

$$\lambda\left(t \middle| x_1, x_2, \cdots\right) = \boxed{\lambda_0(t)} \cdot \boxed{\exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots\right)}$$

こちらを真の値だとして計算

比例ハザードモデル　適用例

再犯までの時間　432人

Rossi et al.(1980) Money, Work and Crime. Academic Press.
Allison(1995) Survival Analysis Using the SAS System.

目的変数

釈放後の逮捕までの時間（週単位）

説明変数

age　釈放時の年齢

prio　それ以前の収監回数

fin　釈放後に経済的援助があったか

（あり=*1*　なし=*0*）

データの一部

逮捕か打ち切り
経済的援助
就労(前)
仮釈放
収監回数
時間
年齢
結婚
学歴
個人

	week	arrest	fin	age	wexp	mar	paro	prio	educ
423	15	1	1	22	0	0	1	3	4
424	52	0	0	18	0	0	1	3	3
425	19	1	0	18	0	0	0	2	3
426	52	0	0	24	1	0	1	2	4
427	12	1	1	22	1	1	1	2	4
428	52	0	1	31	1	0	1	3	3
429	52	0	0	20	0	0	1	1	4
430	52	0	1	20	1	1	1	1	3
431	52	0	0	29	1	0	1	3	4
432	52	0	1	24	1	0	1	1	4

データの一部

目的変数

逮捕か打ち切り

時間

説明変数

経済的援助　年齢　収監回数

個人

	week	arrest	fin	age	wexp	mar	paro	prio	educ
423	15	1	1	22	0	0	1	3	4
424	52	0	0	18	0	0	1	3	3
425	19	1	0	18	0	0	0	2	3
426	52	0	0	24	1	0	1	2	4
427	12	1	1	22	1	1	1	2	4
428	52	0	1	31	1	0	1	3	3
429	52	0	0	20	0	0	1	1	4
430	52	0	1	20	1	1	1	1	3
431	52	0	0	29	1	0	1	3	4
432	52	0	1	24	1	0	1	1	4

パッケージsurvival　関数coxph

```
coxph(Surv(week,arrest)~fin+age+prio, data=Rossi)
```

目的変数部　　　　説明変数　　　データフレーム名

```
res.recidiv01<-coxph(Surv(week,arrest)~fin+age+prio, data=Rossi)
```

関数coxph　基本的な回帰式の表現が使える

関数coxphの結果

```
> res.recidiv01<-coxph(Surv(week,arrest)~fin+age+prio, data=Rossi)

> res.recidiv01
Call:
coxph(formula = Surv(week, arrest) ~ fin + age + prio, data = Rossi)
```

	coef	exp(coef)	se(coef)	z	p
fin	-0.3470	0.707	0.1902	-1.82	0.06800
age	-0.0671	0.935	0.0209	-3.22	0.00130
prio	0.0969	1.102	0.0273	3.56	0.00038

切片の項はない

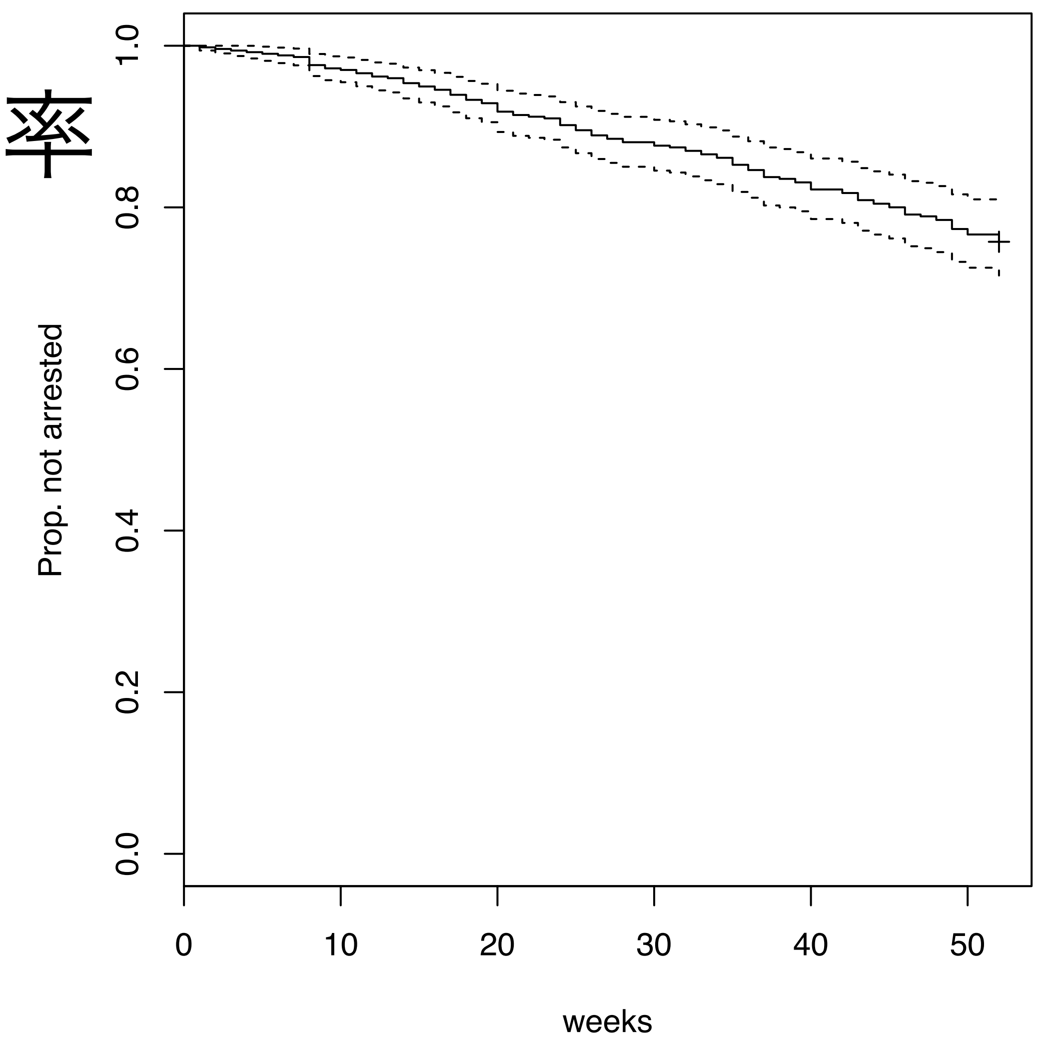
Likelihood ratio test=29.1 on 3 df, p=2.19e-06 n= 432, number of events= 114

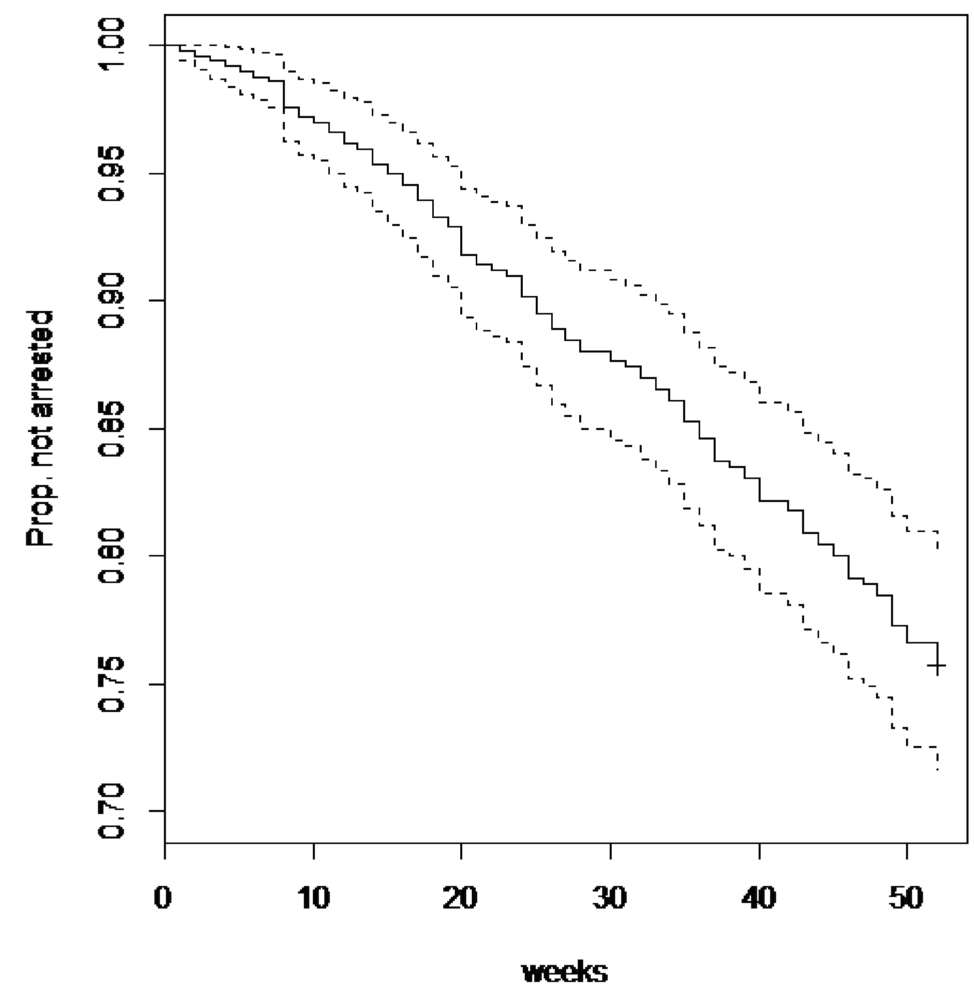
関数coxphの結果にsummary関数を使う

```
> summary(res.recidiv01)
Call:
coxph(formula = Surv(week, arrest) ~ fin + age + prio, data = Rossi)

  n= 432, number of events= 114

         coef exp(coef) se(coef)      z Pr(>|z|)
fin  -0.34695   0.70684  0.19025 -1.824 0.068197 .
age  -0.06711   0.93510  0.02085 -3.218 0.001289 **
prio  0.09689   1.10174  0.02725  3.555 0.000378 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     exp(coef) exp(-coef) lower .95 upper .95
fin     0.7068     1.4148    0.4868    1.0263
age     0.9351     1.0694    0.8977    0.9741
prio    1.1017     0.9077    1.0444    1.1622

Concordance= 0.63  (se = 0.027 )
Rsquare= 0.065   (max possible= 0.956 )
Likelihood ratio test= 29.05  on 3 df,   p=2.189e-06
Wald test            = 27.94  on 3 df,   p=3.741e-06
Score (logrank) test = 29.03  on 3 df,   p=2.203e-06
```

関数survfitによる予測値の算出とプロット

逮捕されない率
（予測値）





時間

関数cox,zphによる比例ハザード性の検定
残差の時間との相関

```
> cox.zph(res.recidiv01)
            rho     chisq      p
fin     -0.00657  0.00507  0.9433
age     -0.20976  6.54147  0.0105
prio    -0.08004  0.77288  0.3793
GLOBAL        NA  7.13046  0.0679
```

パラメトリックな回帰モデル

加速モデル

accelerated failure time model

略してAFT model

ポアソン回帰モデル

加速モデル

*i*番目の個体の生存時間　y_i

*i*番目の個体の説明変数の値　$x_{1,i}, x_{2,i}, \cdots$

$$\log_e\left(y_i\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \left(\text{error term}\right)$$

生存時間での対数線形性

説明変数と係数の部分の効果

時間の進み方が速くなる／遅くなる

加速モデル

i番目の個体の生存時間 y_i

i番目の個体の説明変数の値 $x_{1,i}, x_{2,i}, \cdots$

$$\log_e(y_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + (\text{error term})$$

生存時間での対数線形性

分布はいろいろ

ワイブル分布、指数分布、ロジスティック分布、
対数ロジスティック分布など

加速モデル

*i*番目の個体の生存時間 y_i

*i*番目の個体の説明変数の値 $x_{1,i}, x_{2,i}, \cdots$

$$\log_e(y_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + (\text{error term})$$

パッケージsurvivalの関数survreg

基本的な回帰式の表現が使える

ワイブル分布、指数分布、ロジスティック分布、
対数ロジスティック分布などの他、ユーザー定義も可能

ポアソン回帰モデル

目的変数　死亡というイベントの数

個体のグループ	説明変数の値が同じ 観察期間が同じ
個体のグループ	説明変数の値が同じ 観察期間が同じ
個体のグループ	説明変数の値が同じ 観察期間が同じ

死亡数	説明変数	個体数×観察期間
目的変数	**説明変数**	**オフセット**