計算機実験用ランダム標本集団の飼い方・しつけ方

— 試行錯誤しながら使いこなす統計学 —

久保拓弥　kubo@ees.hokudai.ac.jp

2003.03.22

生態学会つくば大会自由集会

「データ解析の落とし穴，穴から脱出する計算機ワザ」

今日のハナシの要点

(;_;) **統計学がわからない → ひたすら考える・わからぬまま使う**

(^_^) **統計学がわからない → とりあえず「実験」してみる**

(^o^) **その「実験」結果を考える・利用する**

「実験」 = 「乱数」生成 + 統計学的手法 (検定とか回帰とか ···)

まったく蛇足ではありますが — 私の出自 —

- モデル屋属
- 現実の観測データと対応のつかない数式モデルが**イヤ**になった
- 観測データを解析してモデルを構築するようになった
- データ解析・統計学に興味をもち，ぎょーかいの常識に
  ほんの少しばかりケチをつけるようになった
  - 「ゑくせるでデータ解析なんてダメですよ」

  とか
  - あるいは……

素描: 多くの生態学者にとっての統計学

- 基本こそがよくわからない
  - 「統計ソフトウェアにできることだけをやる」
  - わからない → 「手を止めて」考える → やはりわからない
- 確率分布なんぞは いっさい考えず に平和に人生を終えたい
  - (ほぼ無自覚に) 等分散の正規分布使う
  - 「そりゃ正規分布じゃないでしょ」と指摘されたりしたら

→ { 変数変換 → 「それはだめ」(独立変数が複数のとき by 粕谷さん)
  { ノンパラメトリックス → 「それもだめ」(母集団に等分散性ないとき by 粕谷さん)

- すぐに割算して比率とかを計算したがる……ゑくせるとかで (今日は省略)
- 「ゆーい差」決戦主義 (今日は省略)

本日ハナしてみようとすることども

- **乱数** (random number) **の復習**

  “R” **の紹介を試みつつ**

- **データ解析に使えそうな確率分布あれこれ**

  一般化線形モデル (glm) **と連関させつつ**

- **今回は** (**残念ながら**) **説明しない内容**— **またの機会に** —
  - overdispersion (**過剰なばらつき**)
  - resampling & randomization (bootstrap **法・並び換えなどなど**)
  - ...

乱数 (random number) の復習

“R” の紹介を試みつつ

“R”: これからの統計ソフトウェア

http://www.r-project.org/

- **いろいろな** OS **で使える** freeware (≈ S-Plus **の** free **版**)
- **どんどん発展している**
- **機能が充実している** (glm **とか**)
- S **言語によるプログラミング可能**
- **作図機能も強力**
- **よい教科書が出版されつつある**
  - “Introductory Statistics with R” P. Dalgaard (2002)
  - “Computational Statistics” M. Crawley (2002)

乱数とは何だったか?

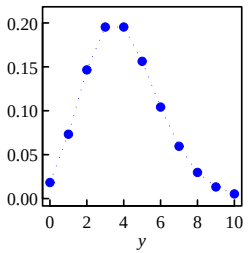
統計学の中核概念: **ある 確率分布 で表現される母集団から無作為に得られた数**

(今日はパラメトリックな確率分布のハナシ)

ポアソン分布

R の関数:

dpois$(y,\ \lambda = 3)$



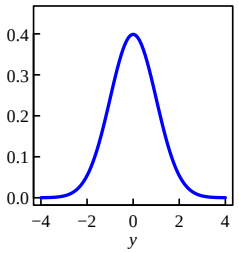
→

```
> rpois(10, lambda = 3)
5 4 3 2 4 2 4 1 7 1
```

正規分布

R の関数:

dnorm$(y,\ \mu = 0,$
$\sigma = 1)$



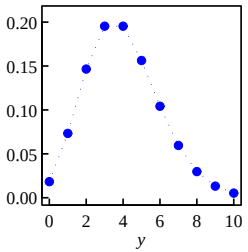
→

```
> rnorm(9, mean = 0, sd = 1)
 1.4851004 -0.9912880 -0.1092131
-2.1752314 -0.3779424  1.1360432
 1.2493592 -1.2405408 -0.4425550
```

推定とは何か?

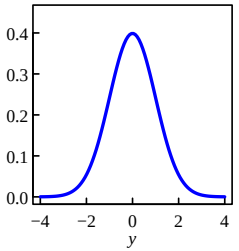
ポアソン分布の推定

5 4 3 2 4 2 4 1 7 1 $\longrightarrow$



正規分布の推定

1.4851004 -0.9912880 -0.1092131 $\longrightarrow$
-2.1752314 -0.3779424 1.1360432
1.2493592 -1.2405408 -0.4425550



乱数とみなされる標本集団 $\longrightarrow$ 母集団を特徴づける確率分布を特定

しかし「ゆーい差」決戦主義者たちは「推定」という過程にすら無自覚なんですよね……

乱数つかうと何ができるか?

- よくわからない統計学的手法 (推定・検定・モデル選択) を「実験的」に理解できる
- データから推定した確率分布，これを母集団として乱数生成してみる → サンプリング・推定の偏りをチェックできる
- 数式を使わずに検定 (危険率の計算)・検出力の計算ができる → 自分のデータ専用の検定が作れる
- 自分の作った統計学的手法がまっとうかどうか検査できる

確率の世界を「実感」できる

R と乱数と glm

		乱数生成	推定・検定
離散分布	二項分布	rbinom	glm(family = binom)
	ベルヌーイ分布	rbinom	glm(family = binom)
	多項分布	(無し?)	multinom
	ポアソン分布	rpois	glm(family = poisson)
	幾何分布	rgeom	(最尤推定?)
連続分布	指数分布	rexp	glm(family = gamma)
	ガンマ分布	rgamma	glm(family = gamma)
	正規分布	rnorm	glm(family = gaussian)
	対数正規分布	rlnorm	glm(family = gaussian)
	ベータ分布	rbeta	(最尤推定?)
	一様分布	runif	(推定したくない)

一般化線形モデル (generalized linear model, glm)

正規分布ではないパラメトリックな方法を使いたおす

- **指数関数族に属する確率分布あれこれ (正規分布，二項分布，ポアソン分布，$\cdots$) で説明されるバラつきのデータに適用できる**
- link **関数を指定できる**
- **擬似尤度法が使える (今日は省略)**
- **独立変数は何でもよい: 連続変数，名義変数，順序変数**
- **パラメーターは線形に結合していなくてはならない (線形モデル)**

$$\mathrm{link}(\mu(\mathbf{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots = \sum_i \beta_i x_i$$

確率分布を推定する方法たちの階層性

- **パラメトリックな確率分布**
  - **最尤推定法** (maximum likelihood method)
    - **一般化線形モデル** (glm): **指数関数族の確率分布 + 線形モデル**
      - **最小二乗法**: **等分散正規分布 + 線形モデル**
  - **モーメント法** (moment method)
- **ノンパラメトリックな確率分布**
  - **一般加法モデル**
  - ・・・

蛇足: R と Perl (http://www.perl.org/)

それぞれに使いどころアリ

- R **の強み: 統計学的な処理が「3 行プログラム」で片づいたり** (Perl **だと長くなる**)

  例: ファイル読みこんで logistic **回帰**

```
xy <- scan("data.txt", list(x = 0, y = 0)) # data.txt 読め
logistic <- glm(y ~ 1 + x, family = binomial(logit), data = xy)
summary(logistic)# 結果を出せ
```

- Perl **の強み: テキスト処理に格段にすぐれる**
  - R **に読ませるファイルの前処理に**
    (**例: 呪われゑくせるファイルの「解毒」**)

R での乱数の飼いかた: 乱数を産ませる

簡単すぎて説明しようがない……いろいろな乱数がこんなに簡単に得られるとは画期的です

```
> rnorm(1)
[1] 1.225641
> rnorm(3)
[1]  0.06946610 -0.77775513  0.09740263
> rnorm(3, mean = 10, sd = 3)
[1]  4.140088 10.766689  9.179323  9.711892  8.932404

> rnorm(20, mean = 100, sd = 1)
 [1] 100.07112 100.76318  98.50378  99.95469  98.53718  98.53915 100.33856
 [8] 101.26489 101.29528  99.33203 100.55881  99.15316  99.40310 101.15178
[15] 100.05952 100.04701  99.61895  99.18690 101.65745  99.70014
```

R での乱数の飼いかた: 乱数を verctor にしまう

たくさんの「数」をまとめて操作できる

```
> rnumbers <- rnorm(3)
> rnumbers
[1] -0.8029520  0.3082426  1.6835991
> rnumbers + 1
[1] 0.1970480 1.3082426 2.6835991
> rnumbers2 <- rnumbers * 2
> rnumbers2
[1] -1.6059040  0.6164853  3.3671982
```

R での乱数の飼いかた: 乱数に芸を教える

線形モデル: $\mu(x) = \beta_0 + \beta_1 x$

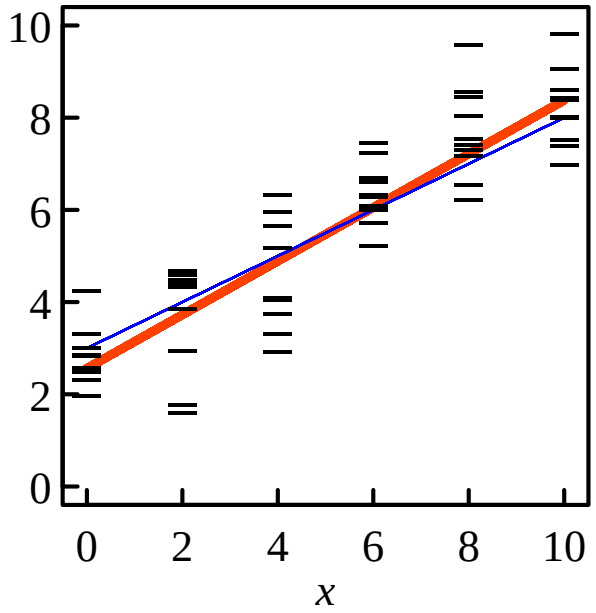
- **まず平均値 $\mu(x)$ の** vector **をつくる**

```
beta0 <- 3
beta1 <- 0.5
x <- seq(x.min, x.max, by = 2.0)
x <- rep(x, 10) # 各 x について 10 個ずつ
mu <- beta0 + beta1 * x
```

- **正規乱数を** rnorm **で**

```
sample <- rnorm(length(y),
                mean = mu, sd = 1)
```

青: ホントの $\mu(x)$, 赤: 推定された $\hat{\mu}(x)$



- **R の** glm **関数による推定** (**この場合は** glm **でなくて** lm **でいいんだが**)

```
result <- glm(sample ~ 1 + x, family = gaussian)
```

乱数の「もと」

データ解析に使えそうな確率分布あれこれ

一般化線形モデル (glm) と連関させつつ

「ばらつき」をしっかり見る

確率分布あれこれふたたび

		乱数生成	推定・検定
離散分布	二項分布	rbinom	glm(family = binom)
	ベルヌーイ分布	rbinom	glm(family = binom)
	多項分布	(無し?)	multinom
	ポアソン分布	rpois	glm(family = poisson)
	幾何分布	rgeom	(最尤推定?)
連続分布	指数分布	rexp	glm(family = gamma)
	ガンマ分布	rgamma	glm(family = gamma)
	正規分布	rnorm	glm(family = gaussian)
	対数正規分布	rlnorm	glm(family = gaussian)
	ベータ分布	rbeta	(最尤推定?)
	一様分布	runif	(推定したくない)

あなたのデータにぴったりの確率分布はコレ!

何でもかんでも変数変換しない・データにあわせて分布を選んで推定

— 選びかたの三つのポイント —

1. **説明したい量は離散か連続か?**
   - **離散**: { **生きてる**, **死んでる** }, **カウントデータ**, $\cdots$
   - **連続**: $\{0.56,\ 1.33,\ 12.4,\ 9.84,\ \cdots\}$, $\cdots$
2. **説明した量の範囲は?**
   - $\{0, 1, \cdots, N\}$, $\{0, 1, \cdots, \infty\}$, $[y_{\min}, y_{\max}]$, $[-\infty, \infty]$, $\cdots$
3. **説明したい量の分散 (ばらつき) と平均の関係は?**
   - **分散** $\approx$ **定数**, **分散** $\approx$ **平均**, **分散** $\propto$ **平均**, **分散** $\propto$ **平均** ^{n}, $\cdots$

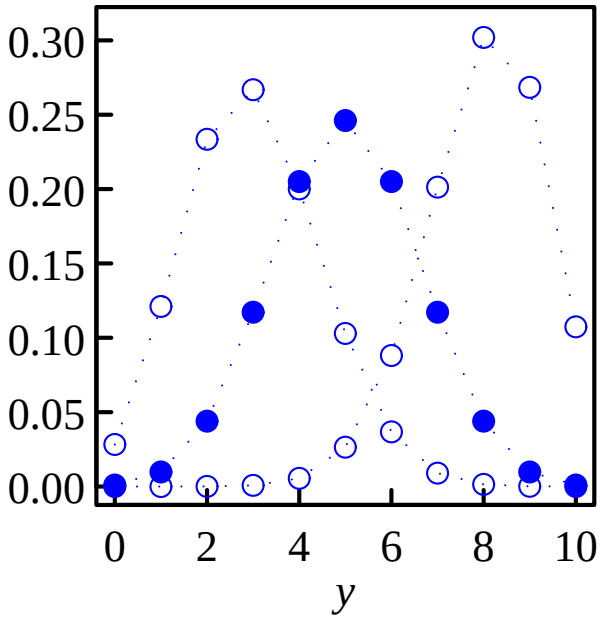
二項分布 (binomial distribution)

- **離散分布** $y \in \{0, 1, 2, \cdots, N\}$
- **確率密度関数** (paramter: N, p)

$$\binom{N}{y} p^y (1-p)^{N-y}$$

- **期待値** Np**，分散** $Np(1-p)$

R の関数: `dbinom`(y, N, p)



- **使いどころ**: **個体を区別しない (属性ごとにグループ化した) カウントデータ**
  - **個体の状態** $\in$ { **生きてる**, **死んでる** }
  - **処理に応答した・しなかった**
- **「割合」「比率」の計算なんぞヤメて，**logistic **回帰するのが当節の流行**

ベルヌーイ分布 (Bernoulli distribution)

- **離散分布** $y \in \{0, 1\}$
- **確率密度関数** (paramter: p)

$$p^y(1-p)^{1-y}$$

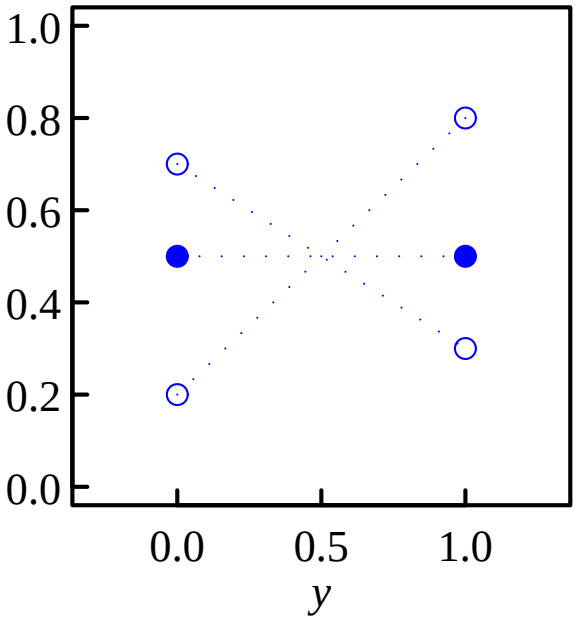
- **期待値** p**，分散** $p(1-p)$

○ **使いどころ: 個体を区別するカウントデータ**

・ **個体サイズが生き死にに与える影響**

○ $N = 1$ **の二項分布として計算すればよい**

R の関数: `dbinom`$(y, 1, p)$



○ **ということで，これも** logistic **とか** probit **モデルによる推定**

glm: 二値データと logistic モデル

logistic **モデル**: $p(x) = \frac{1}{1+\exp[-(\beta_0+\beta_1 x)]}$

- **まず** $p(x)$ **の** vector **をつくる**

```
beta0 <- -3
beta1 <- 0.5
x <- seq(x.min, x.max, 0.1)
p <- 1 / (1 + exp(-beta0 - beta1 * x))
```

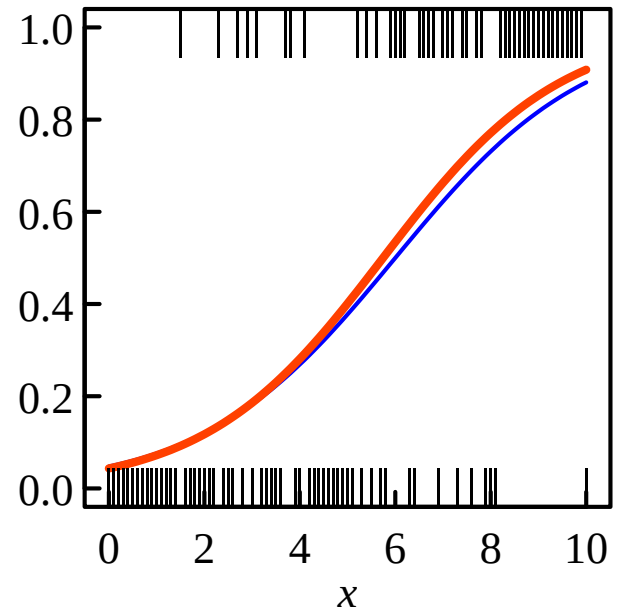
- $p(x)$ **から二項乱数を** rbinom **で**

```
sample <- rbinom(length(p), 1, prob = p)
```

- R **の** glm **関数による推定**

```
logistic <- glm(sample ~ 1 + x, family = binomial(logit))
```

青: ホントの $p(x)$, 赤: 推定された $\hat{p}(x)$



さらに summary(logistic) と命じると……

```
Call:
glm(formula = sample ~ 1 + x, family = binomial(logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0481  -0.8413  -0.4843   0.8688   1.9038

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.12486    0.51363  -4.137 3.52e-05 ***
x            0.40912    0.09029   4.531 5.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

...
```

多項分布 (multinomial distribution)

- **離散分布** $y_i \in \{0, 1, 2, \cdots, N\}$
- **確率密度関数** (paramter: $N, \{p_i\}$)

$$\frac{N!}{y_1 y_2 \cdots y_M} \prod_{i=1}^{M} p_i^{y_i}$$

- **期待値** Np_i**，分散** $Np_i(1-p_i)$
  - **使いどころ**: 3 **状態以上の** (**かつ上限ありそうな**) **カウントデータ**
    - **個体の状態** $\in$ { **健康**, **不健康**, **死** }
    - **一本の枝から何本新しい枝が出るか**

(**うまく図に描けん……**)

多項分布のカタチをデータから推定するのもさまざまな方法がある．R **で使える面白い関数の例**: multinom (multinomial logistic regression)**．**

これは logistic **回帰の多値版で，一般化線形モデルのひとつである** (**しかし** glm **関数では取り扱えない**)**．**

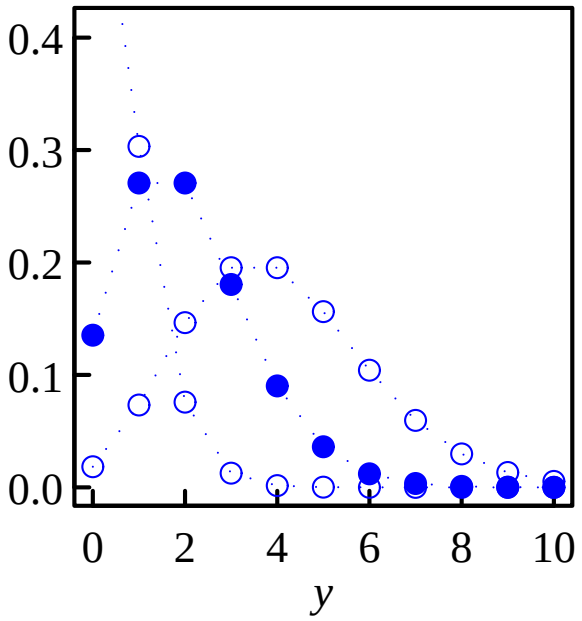
ポアソン分布 (Poisson distribution)

- **離散分布** $y_i \in \{0, 1, 2, \cdots, \infty\}$
- **確率密度関数** (paramter: λ)

$$\frac{\lambda^y \exp(-\lambda)}{y!}$$

- **期待値** λ**，分散** λ
  - **使いどころ: 「一定時間にかかってくる電話の回数」……上限を設定できないカウントデータ**
    - **産卵数・種子数**

R の関数: `dpois`(y, λ)

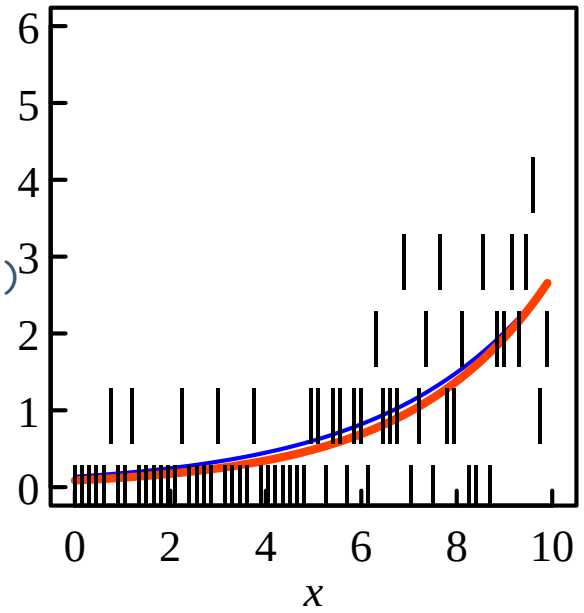


- **個数のデータが得られたら，まずは「ポアソン分布で説明できないか?」と考えてみる**

glm: ポアソン分布 ($\lambda(x)$ = 平均 = 分散) の推定

"log link": $\lambda(x) = \exp(\beta_0 + \beta_1 x)$

青: ホントの $\lambda(x)$, 赤: 推定された $\hat{\lambda}(x)$



- まず $\lambda(x)$ の vector をつくる

```
beta0 <- -2
beta1 <- 0.3
x <- seq(from = x.min, to = x.max, by = 0.1)
y <- exp(beta0 + beta1 * x)
```

- $\lambda(x)$ からガンマ乱数を rpois で

```
sample <- rpois(length(y), lambda = y)
```

- R の glm 関数による推定

```
pois <- glm(sample ~ 1 + x, family = poisson(link = "log"))
```

さらに summary(pois) と命じると……

```
Call:
glm(formula = sample ~ 1 + x, family = poisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0548  -0.8627  -0.3371   0.5896   1.8564

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.53556    0.26984  -5.691 1.27e-08 ***
x            0.23779    0.03698   6.430 1.28e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

...
```

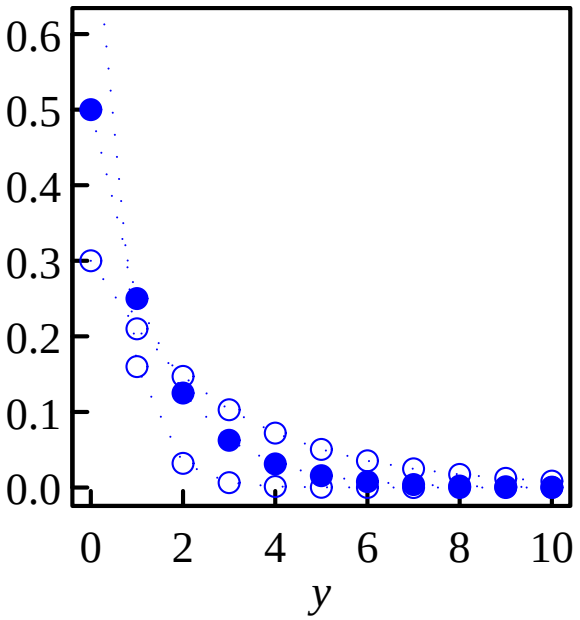
幾何分布 (geometric distribution)

- **離散分布** $y_i \in \{0, 1, 2, \cdots, \infty\}$
- **確率密度関数** (paramter: p)

$$p(1-p)^y$$

- **期待値** $1/p$**，分散** $(1-p)/p^2$
  - **使いどころ**: **「コイン投げして表がでるまでの回数」というのがふつーの説明なんだが……**
  - **少し邪道な使いかたかもしれないけど，ポアソン分布ではうまく説明できないデータ**
    - **分散** $\approx$ **平均** $\rightarrow$ **ポアソン分布をつかう**
    - **分散** $\propto$ **平均** ^{2} $\rightarrow$ **幾何分布をつかう**

R の関数: `dgeom`(y, p)



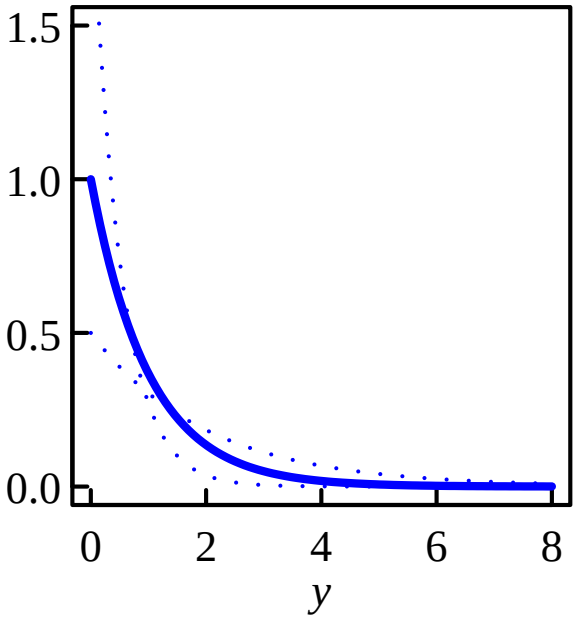
指数分布 (exponential distribution)

- **連続分布** $y \in [0, \infty]$
- **確率密度関数** (paramter: λ)

$$\lambda \exp(-\lambda y)$$

- **期待値** $1/\lambda$**，分散** $1/\lambda^2$
  - **幾何分布の連続版.**
  - **使いどころ**:
    - 1 **個の突然変移が生じるまでの時間分布**
    - **植物個体群のサイズ分布**?
  - **次に説明するガンマ分布の特殊なもの，とみなせばよい**
  - R **の** glm **の** gamma **の分散関数はデフォルトで指数分布**

R **の関数**: dexp(y, λ)



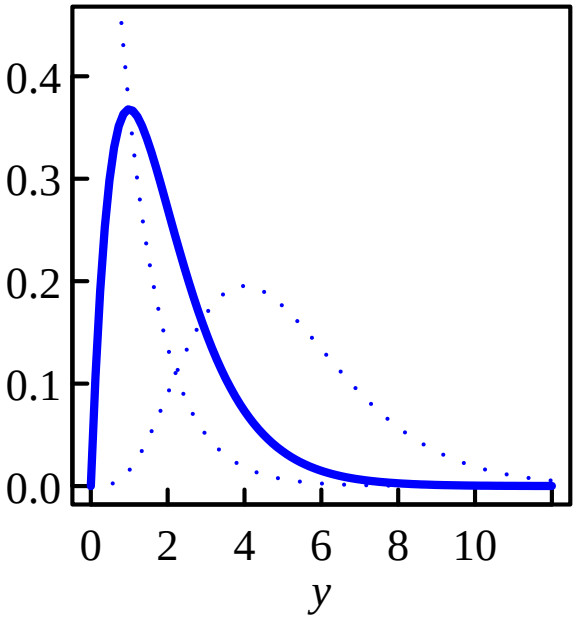
ガンマ分布 (Γ distribution)

- **連続分布** $y \in [0, \infty]$
- **確率密度関数** (paramter: α, β)

$$\frac{y^{\alpha-1} \exp(\frac{-y}{\beta})}{\beta^{\alpha} \int_0^{\infty} u^{\alpha-1} \exp(-u) du}$$

- **期待値** $\alpha\beta$**，分散** $\alpha\beta^2$
  - **使いどころ**: **「負の値をとったらイヤ」な連続値**
  - **「左右非対称」で正規分布ではダメっぽいとき**
  - **分散** $\propto$ **平均，から 分散** $\propto$ **平均** ^{2}**，ぐらい**
    - **身長・体重・サイズ成長量などなど**

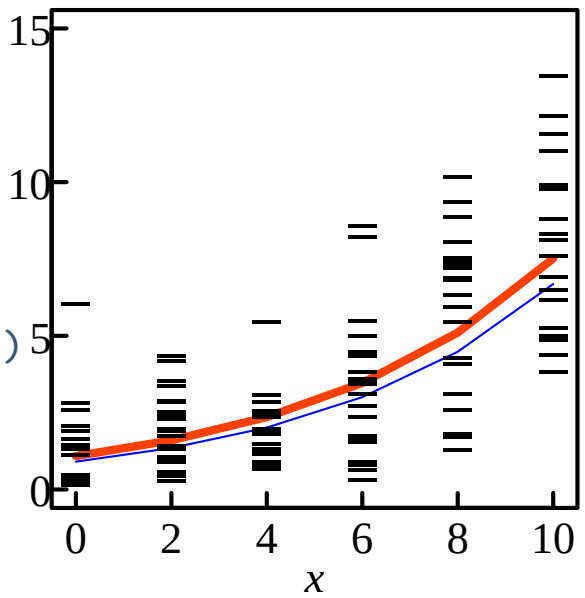
R の関数: `dgamma`(y, α, β)



glm: ガンマ分布 (分散 $\propto$ 平均) の推定

"log link": $\mu(x) = \alpha\beta = \exp(\beta_0 + \beta_1 x)$

青: ホントの $\mu(x)$, 赤: 推定された $\hat{\mu}(x)$



- まず $\mu(x)$ の vector をつくる

```
beta0 <- -0.1
beta1 <- 0.2
scale <- 1.5
x <- seq(from = x.min, to = x.max, by = 2.0)
x <- rep(x, 20)
y <- exp(beta0 + beta1 * x)
```

- $\mu(x)$ からガンマ乱数を rgamma で

```
sample <- rgamma(length(y), shape = y / scale, scale = scale)
```

- R の glm 関数による推定

```
gam <- glm(sample ~ 1 + x, family = Gamma(link = log))
```

さらに summary(gam) と命じると……

```
Call:
glm(formula = sample ~ 1 + x, family = Gamma(link = log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3726  -0.8527  -0.1934   0.3070   2.2978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.43582    0.14525  -3.001  0.00329 **
x            0.22761    0.02399   9.489 3.25e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

...
```

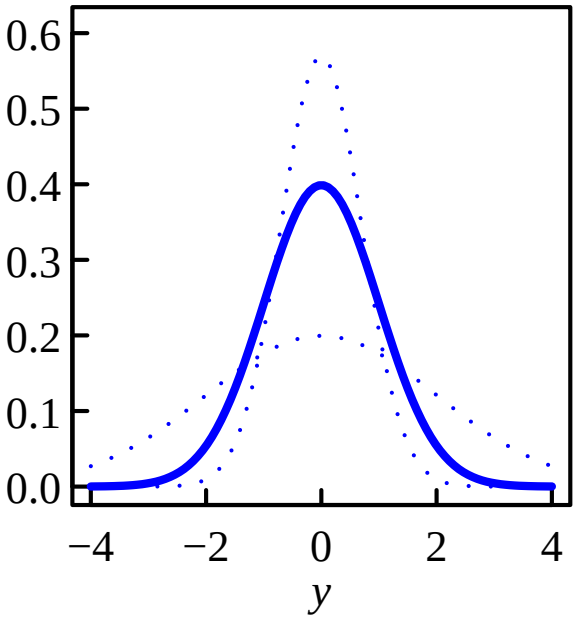
正規分布 (normal or Gaussian distribution)

- **連続分布** $y \in [-\infty, \infty]$
- **確率密度関数** (paramter: μ, σ)

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

- **期待値** μ**，分散** σ^2

R の関数: `dnorm`(y, μ, σ)



- 使いどころ: 分布を考えるのが面倒くさいとき
  - 何でもてきとーに
  - 人間の計測誤差の推定
- R の nls を使うと「等分散性」がない場合でも OK (分散関数を明示的に指定する)

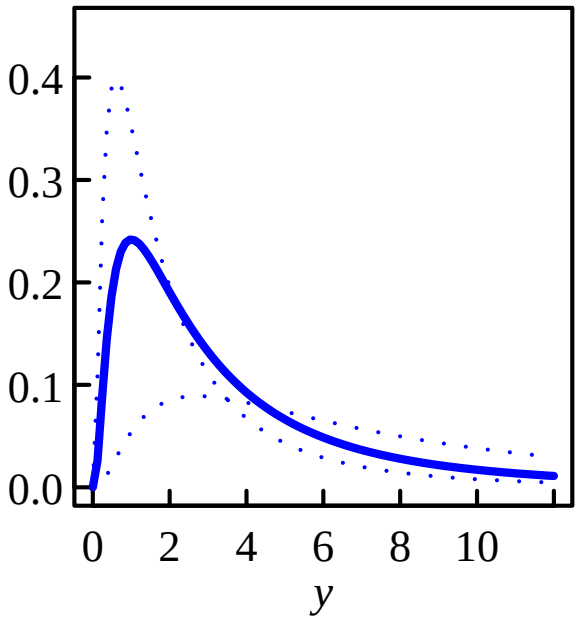
対数正規分布 (log normal distribution)

- **連続分布** $y \in [0, \infty]$
- **確率密度関数** (paramter: μ, σ)

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

- **期待値** $\exp(\mu + \sigma^2/2)$**,**
  分散 $\exp(2\mu+\sigma^2)[\exp(\sigma^2)-1]$
  - **「貧乏-金持ちの分布」**
  - **使いどころ: 値が非負で分散が (平均にくらべて) とてつもなくでかい場合**
    - **重量・バイオマス分布?**

R の関数: `dlnorm`(y, μ', σ')



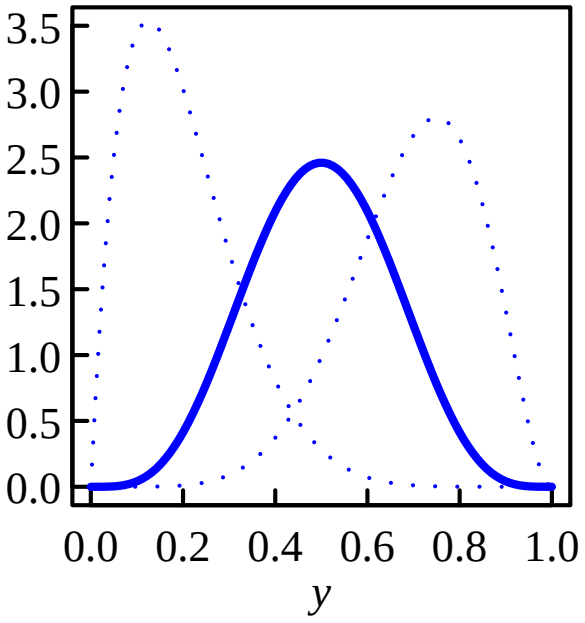
ベータ分布 (β distribution)

- **連続分布** $y \in [0, 1]$
- **確率密度関数** (paramter: α, β)

$$\frac{y^{\alpha-1}(1-y)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du}$$

- **期待値** $\frac{\alpha}{\alpha+\beta}$**，分散** $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
  - **使いどころ**: $[0, 1]$ **の連続値をとるデータ**
    - **成分含量分析器の結果出力の解析**?
  - **これは一般化線形モデル** (glm) **では扱えない**
  - **パラメーターは最尤法で推定する**

R の関数: dbeta(y, α, β)



そろそろ時間切れなので本日はここまで

おススめしたいこと:

わからんコトは乱数で実験・実感しつつ理解
データよく見て，分布を考え，glm しよう

- 統計学でわからないことがあったら「実験」してみる

「実験」 ＝ 「乱数」生成 + 統計学的手法

- R で乱数グラフを作図して観賞する
- 自分のデータも glm してみる
- glm を疑る — glm にも落し穴? — 実験してみる
- いろいろな乱数利用法を考えてみる