2003.11.19

生物多様性論 I: 「生態学における統計学的手法の基礎」 (と勝手に改題)

全部で 3 回講義の 2

統計モデリングと推定を重視してみる

— 理解できる統計学めざして —

http://hosho.ees.hokudai.ac.jp/~kubo/stat/2003/

講釈: 久保拓弥　kubo@ees.hokudai.ac.jp

この 3 回だけの統計学授業でやること

統計学が 何もわかってない 修士課程大学院生が対象．この善男善女どもがデータ解析に際して「統計学って，まるっきり理解不能」と遭難する確率を減少せしめるべく，この世界の おおまかな地図を提供し解説する．

- **第 1 回**: 2003.11.17 (**月**)
  「検定」の使われかたを観察しよう
  — **「検定」ってそんなにエラいのか**?
- **第 2 回**: 2003.11.19 (**水**)
  統計モデリングと推定を重視してみる
  — **理解できる統計学めざして**
- **第 3 回**: 2003.11.26 (**水**)
  データにあわせる統計モデリングの例
  — **なんでも割算すんな**!

個別的なワザより全体に共通する考えかたを— ただし内容は偏ってるよ

今日のハナシ: ばらつきのモデリング

1. **なぜ「ばらつきモデル」を?**

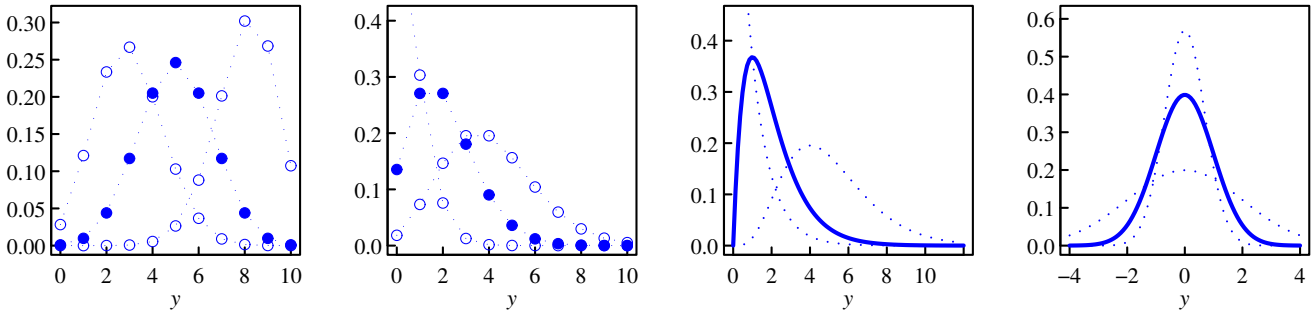
   基本を押えて一般化線形モデル (GLM) につなぐ

2. **乱数 (標本) と推定**

   統計学的データ解析とは何か，を説明してみる

3. **確率分布あれこれ**

   役に立ちそうな分布をざっとながめる



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

いろいろな確率分布をまとめて使いたおすワザ
(おトク感は統計学的技法の中では最高!)

- **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，$\cdots$) で説明されるばらつきのデータに適用できる**
- link **関数を指定できる**
- **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)**

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

- **難点: 日本語で書かれたよい教科書が少ない**

確率分布を推定する方法たちの階層性

なぜ一般化線形モデルを? — より広い世界を統一的に扱えるから

[**最尤推定法**で扱えるモデル]
何でもいいから確率分布があるモデル
一般化線形混合モデルなどなど

[**一般化線形モデル** (GLM)]
指数関数族の確率分布 + 線形モデル
ロジスティック回帰，ポアソン回帰などなど

[**最小二乗法**的に考えるモデル]
等分散正規分布 + 線形モデル
直線回帰，いわゆる「分散分析」などなど

一般化線形モデルさえ使えれば……

事後的なモデリング，探索的なデータ解析を主体とする研究において，
以下の方法などを個別に勉強して使わなくてよい (と勝手に断定):

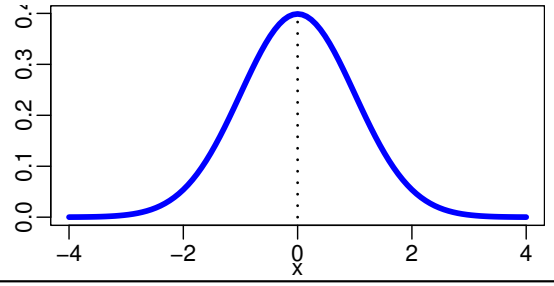
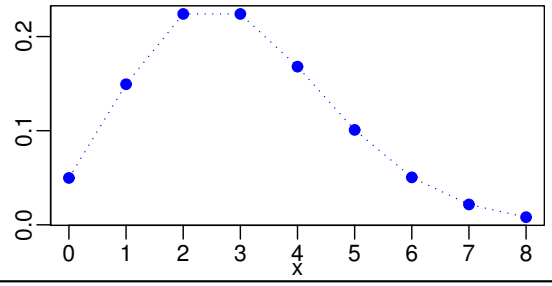
- 直線回帰・重回帰
- いわゆる分散分析
- 母平均の検定
- 比率の検定
- 「傾き」の検定
- いわゆる共分散分析
- 分割表の χ^2 検定
- (その他あれこれ)

ただし

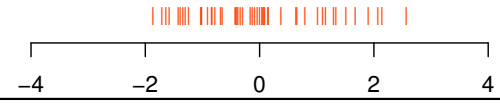
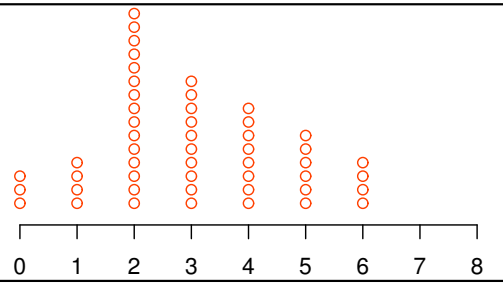
- 指導教官が「○○を使え」と命令しなければ
- 実験計画法にもとづく研究の場合，最初にきめた統計解析法を使わなければならない (途中で変更できない)

GLM 使うには確率分布と推定の理解が必要

といってもムズかしいことではなく，下の図を説明できること，をめざしてハナシを進める

(よびかた)	[連続確率密度分布]	[離散確率密度分布]
• モデル • 確率分布 • 母集団		

サンプリング ⇓ ⇑ (パラメーター) 推定

• データ • 乱数 • 標本集団		

これ使いましょう: 統計ソフトウェア R

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

- いろいろな OS で使える freeware
- 使いたい機能が充実している
- 作図機能も強力
- S 言語によるプログラミング可能
- よい教科書が出版されつつある
  - 「R による統計解析の基礎」 中澤港 (2003)
  - “Introductory Statistics with R” P. Dalgaard (2002)
  - “Computational Statistics” M. Crawley (2002)
  - ネット上のあちこち

乱数とは何か?

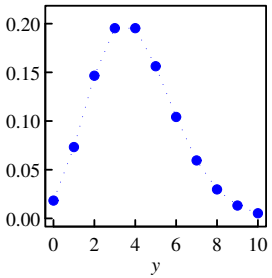
統計学の中核概念

ある 確率分布 (母集団・モデル) から
無作為に得られた値 (標本・データ)

ポアソン分布

R の関数:

dpois(y, $\lambda = 3$)



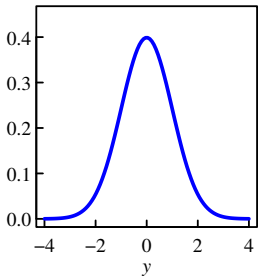
→

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

正規分布

R の関数:

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



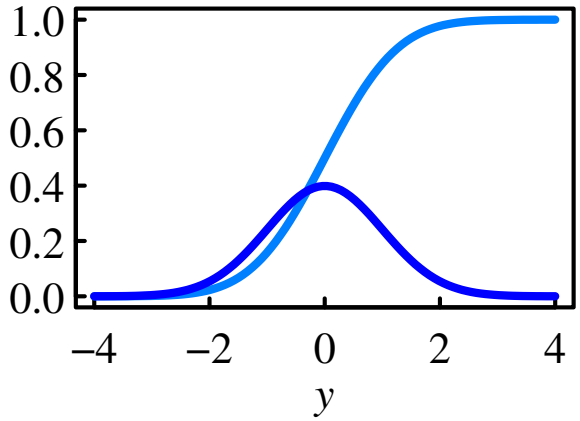
→

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

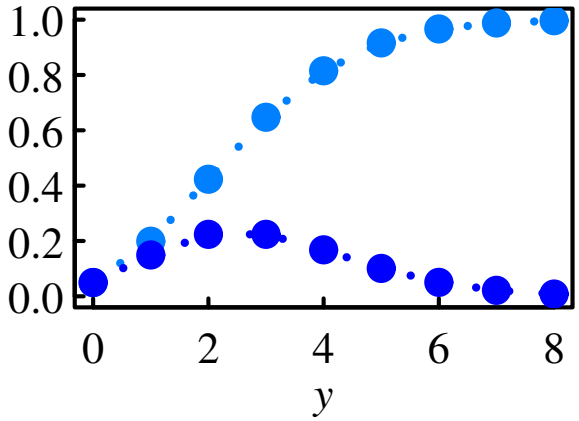
確率分布: 確率分布 (関数) と確率密度分布 (関数)

確率分布関数 $F(y)$ と確率分布密度関数 $f(y)$ の関係

連続関数の例: 正規分布



離散関数の例: ポアソン分布



カタチを決めるパラメーター

平均: 重心

$$m = \int_{-\infty}^{\infty} y \, \mathrm{d}f(y) \qquad m = \sum_{0}^{\infty} y f(y)$$

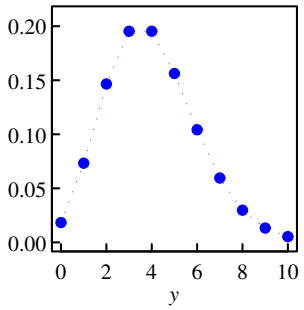
分散: ばらつき

$$\mathrm{Var} = \int_{-\infty}^{\infty} (y-m)^2 \, \mathrm{d}f(y) \qquad \mathrm{Var} = \sum_{0}^{\infty} (y-m)^2 f(y)$$

じゃあ推定ってのは何なの?

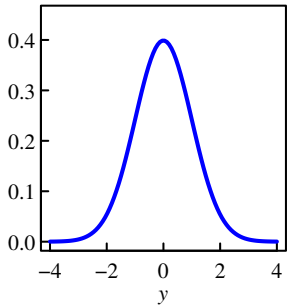
ポアソン分布の推定

5 4 3 2 4 2 4 1 7 1 →



正規分布の推定

```
 1.4851004 -0.9912880 -0.1092131 →
-2.1752314 -0.3779424  1.1360432
 1.2493592 -1.2405408 -0.4425550
```

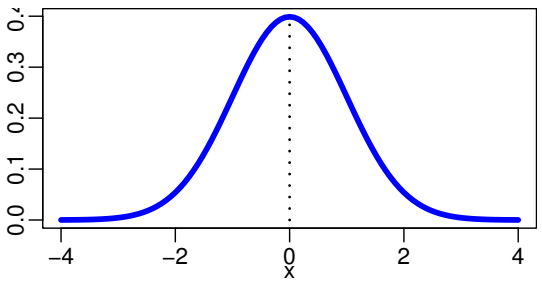
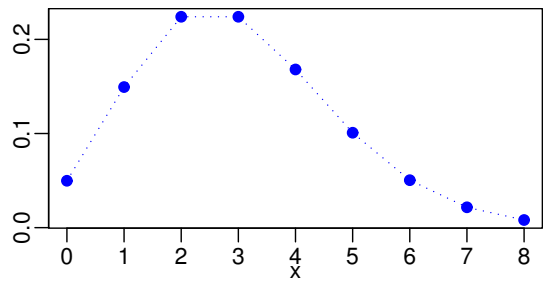


乱数とみなされる標本集団
→ 母集団すなわち確率分布を憶測する技法

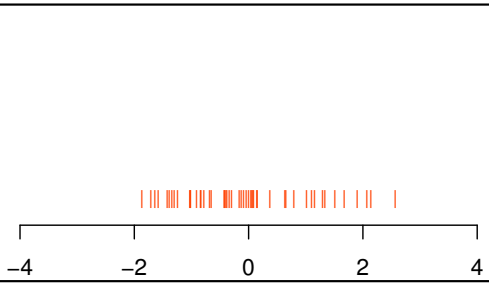
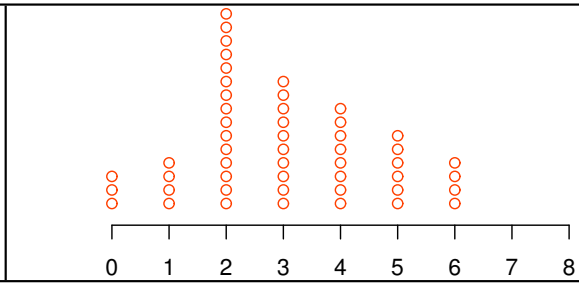
＊ **推定の代表的なワザである最尤推定法は次回に説明する**

統計学とは結局これ: 確率分布，乱数と推定

今日はこの関係さえ理解してもらえればそれで OK!

(よびかた)	[連続確率密度分布]	[離散確率密度分布]
• モデル • 確率分布 • 母集団		

サンプリング ⇓⇑ **(パラメーター) 推定**

• データ • 乱数 • 標本集団		

検定やモデル選択は推定のついでにやってるよーなことで

「ノンパラメトリック」に関する独断的補足

	ぱらめとりっく	のんぱらめとりっく
確率分布のカタチ	重視	あまり重視されない
推定で計算する統計量	カタチの統計量つまり「パラメーター推定量」	順序統計量
検定で計算するもの	帰無仮説の前提における統計量の確率分布	(左に同じ)
モデル選択	あてはまりの良さとモデルの自由度の両方を勘案する	(よくわからん — 不勉強)

つまり，まあ乱暴にいえば，同じようなものでは，と

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
```

統計学勉強における乱数利用のススめ

(;_;) 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

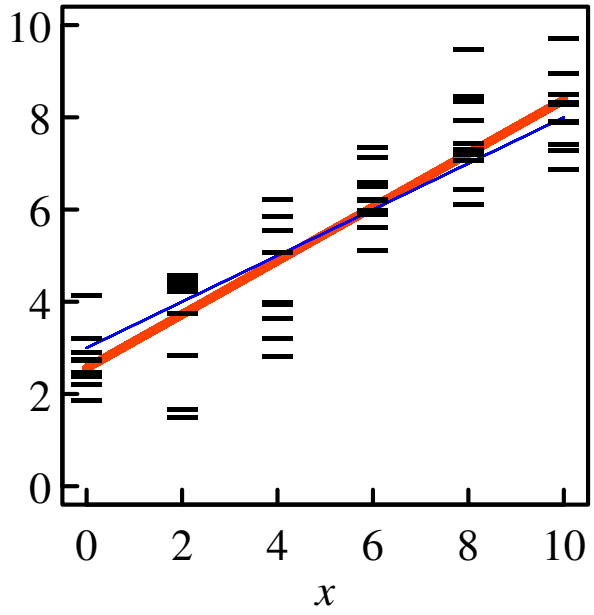
(^o^) その「実験」結果を考える・利用する

「実験」 = 「乱数」生成 + よくわからん統計学的手法

R での乱数の飼いかた: 乱数に芸を教える

線形モデル: $\mu(x) = \beta_0 + \beta_1 x$

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



- **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 10 個ずつ
mu <- beta0 + beta1 * x
```

- **正規乱数を rnorm で**

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

- **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

統計モデルの世界を「実感」できる

あとは気楽に，乱数の「もと」見物でも……

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

あなたのデータの「ばらつき」はどのタイプ?

R と乱数と 一般化線形モデル (glm())

	確率分布	乱数生成	パラメーター推定
(離散)	ベルヌーイ分布	rbinom()	glm(family = binomial)
	二項分布	rbinom()	glm(family = binomial)
	ポアソン分布	rpois()	glm(family = poisson)
	負の二項分布	rnbinom()	glm.nb()
(連続)	ガンマ分布	rgamma()	glm(family = gamma)
	正規分布	rnorm()	glm(family = gaussian)

- glm() で使える確率分布は上記以外もある
- glm.nb() は MASS library 中，またここには rnegbin() なども含まれる

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}, $\cdots$

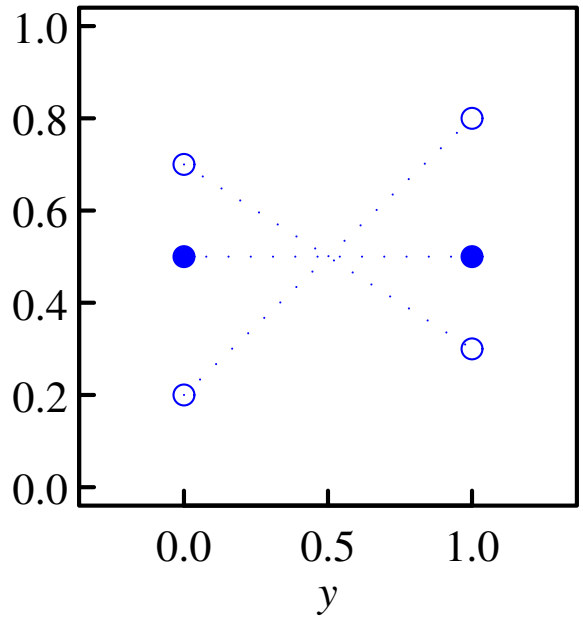
ベルヌーイ分布 (Bernoulli distribution)

- **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**

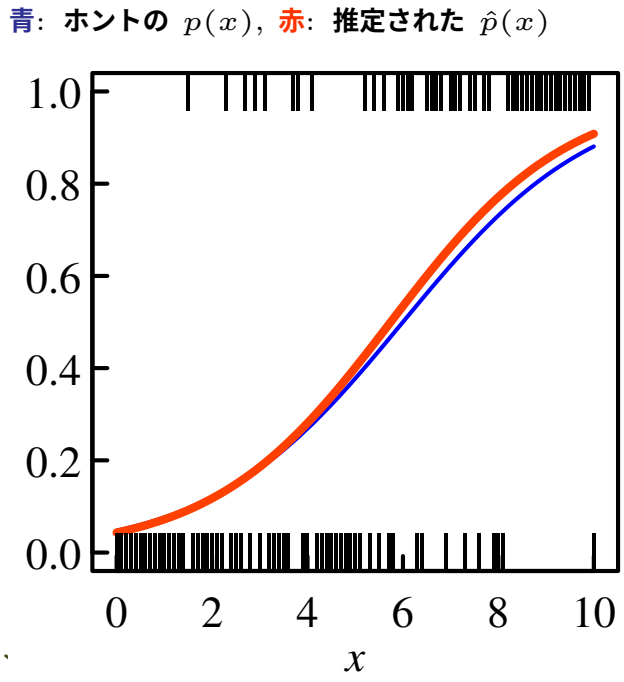
R の関数: `dbinom`$(y, 1, p)$



- **「割合」「比率」の計算なんぞヤメて，**logistic **回帰するのが当世ふう (これは次回に説明)**

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))
```

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
...
```

＊ も少し詳しい説明はまた次回に

二項分布 (binomial distribution)

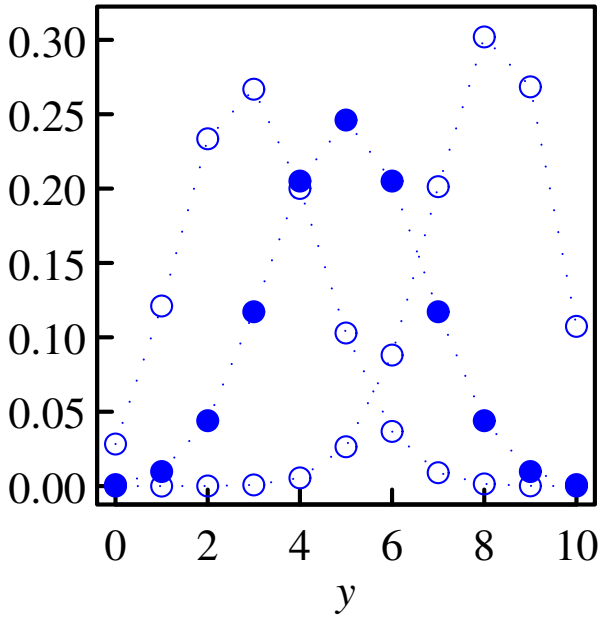
先ほどのベルヌーイ試行を N 回やるとできる分布

- **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** $Np(1-p)$
- **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**

R の関数: `dbinom`(y, N, p)



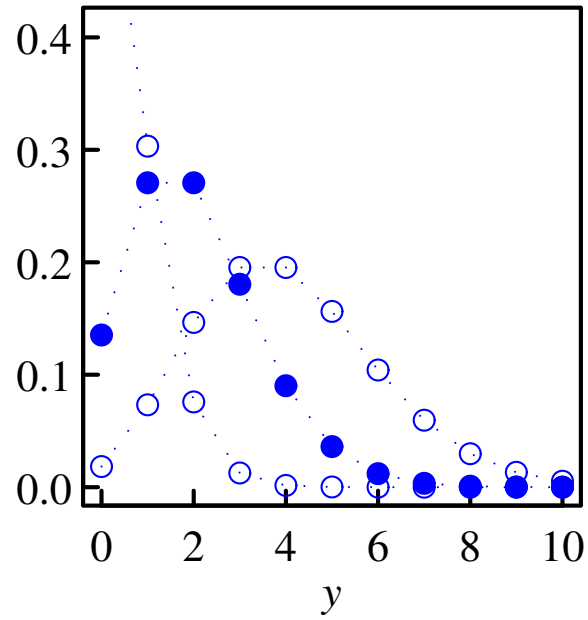
ポアソン分布 (Poisson distribution)

- **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)

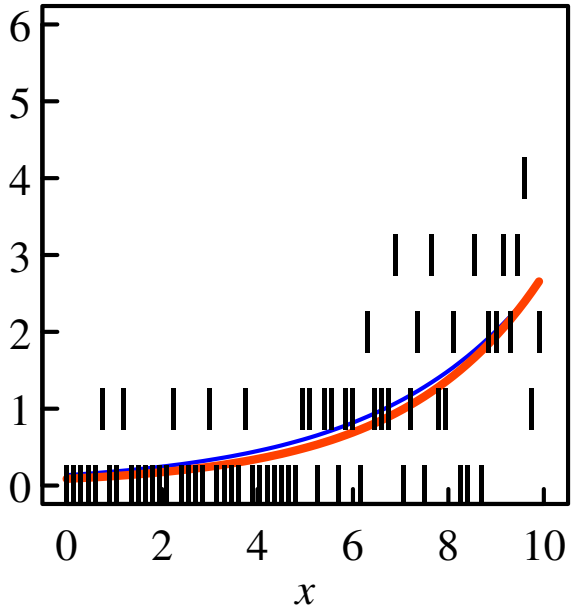


- **個数のデータが得られたら，まずは「ポアソン分布で説明できないか?」と考えてみる**

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(x.min, x.max, 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"))
```

負の二項分布 (negative binomial distribution)

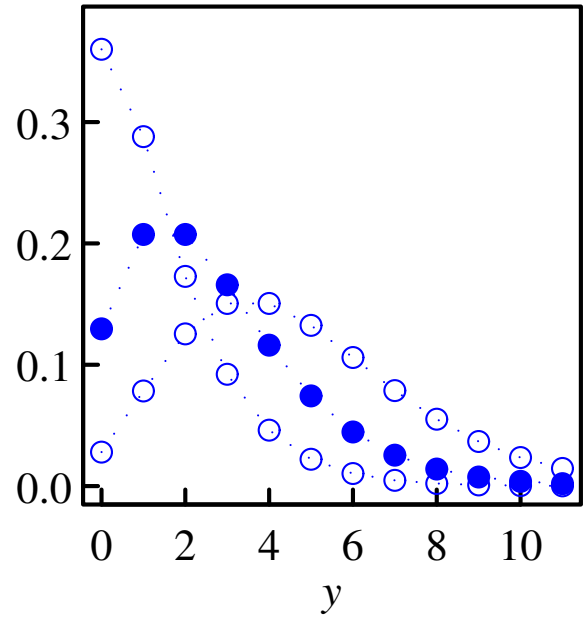
先ほどのポアソン分布では説明しきれんときに

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

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

- **期待値** $k(1-p)/p$**，分散** $k(1-p)/p^2$
- **使いどころ: ポアソン分布と同じ**
  - **ポアソン分布よりばらつきが大きいとき (つまり平均 $<$ 分散)**

R の関数: `dnbinom`(y, k, p)



- **個体の集中分布，なにか集中的に発生する現象の回数の説明に**

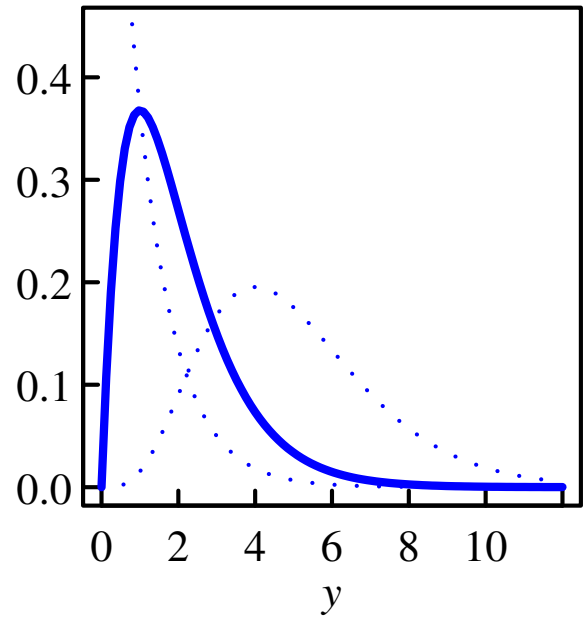
ガンマ分布 (Γ distribution)

- **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**

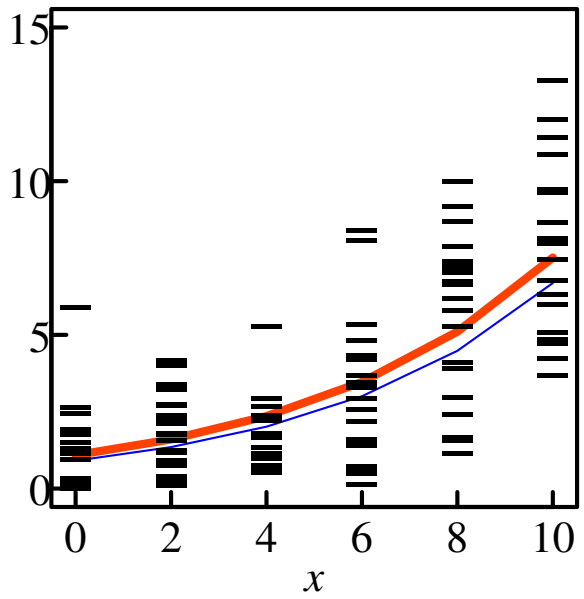
R の関数: `dgamma`(y, α, β)



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 =
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))
```

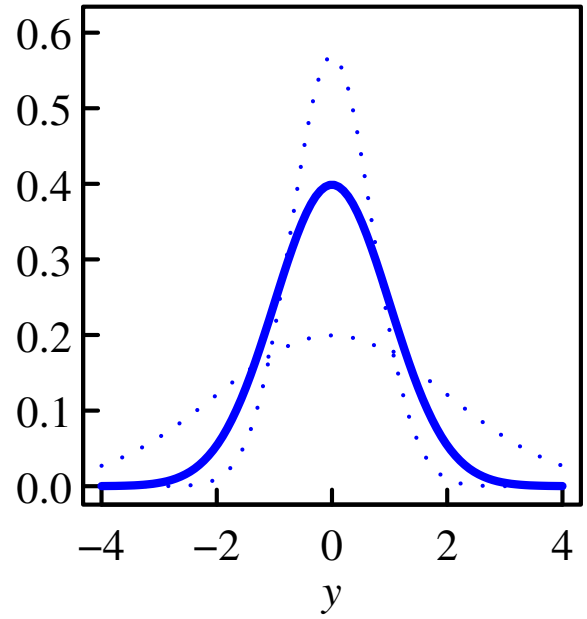
正規分布 (normal or Gaussian distribution)

- **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
- **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 の関数: `dnorm`(y, μ, σ)



- `R` **の** `nls` **を使うと「等分散性」がない場合でも** OK (**分散関数を指定可能**)

本日のまとめ

1. **確率分布と乱数の関係**

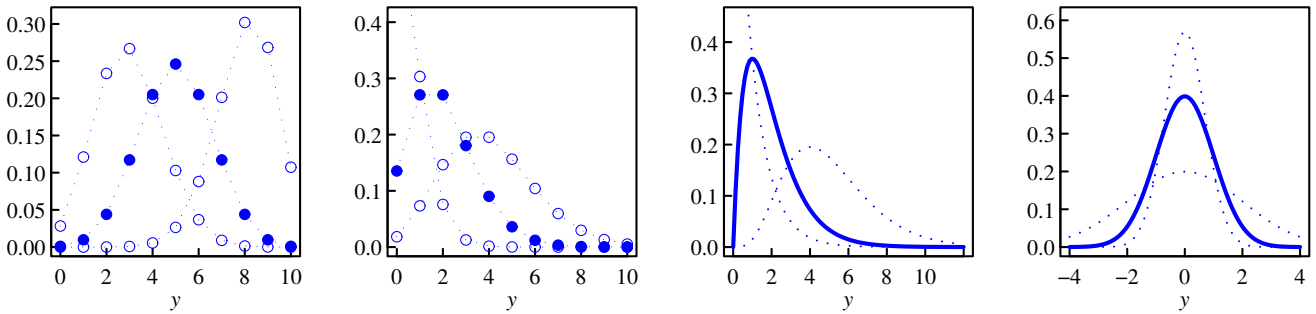
   あるいは母集団 vs 標本，モデル vs データ — 統計学のキモ

2. **わからんコトは乱数で実験・実感しつつ理解**

   統計学でわからないことがあったら「実験」してみる

3. **データよく見て，分布を憶測する**

   ひとつ気楽に `glm()` でも使ってみよう



次回予告

2003.11.26

生物多様性論 I: 「生態学における統計学的手法の基礎」 (と勝手に改題)

全部で 3 回講義の 3

データにあわせる統計モデリングの例

— なんでも割算すんな! —

“割算ぬきの「率」の推定，最尤推定”

http://hosho.ees.hokudai.ac.jp/~kubo/stat/2003/