統計モデリング入門 2017 (b)

probability distribution and maximum likelihood estimation
確率分布と最尤推定

久保拓弥 kubo@ees.hokudai.ac.jp

京大霊長研の講義 https://goo.gl/z9yCJY

2017–11–14

ファイル更新時刻: 2017–11–11 16:02

agenda

今日のハナシ I

1. **例題: 種子数の統計モデリング**
   An example: a distribution of seed number
2. **データと確率分布の対応**
   probability distribution, the core of statistical model

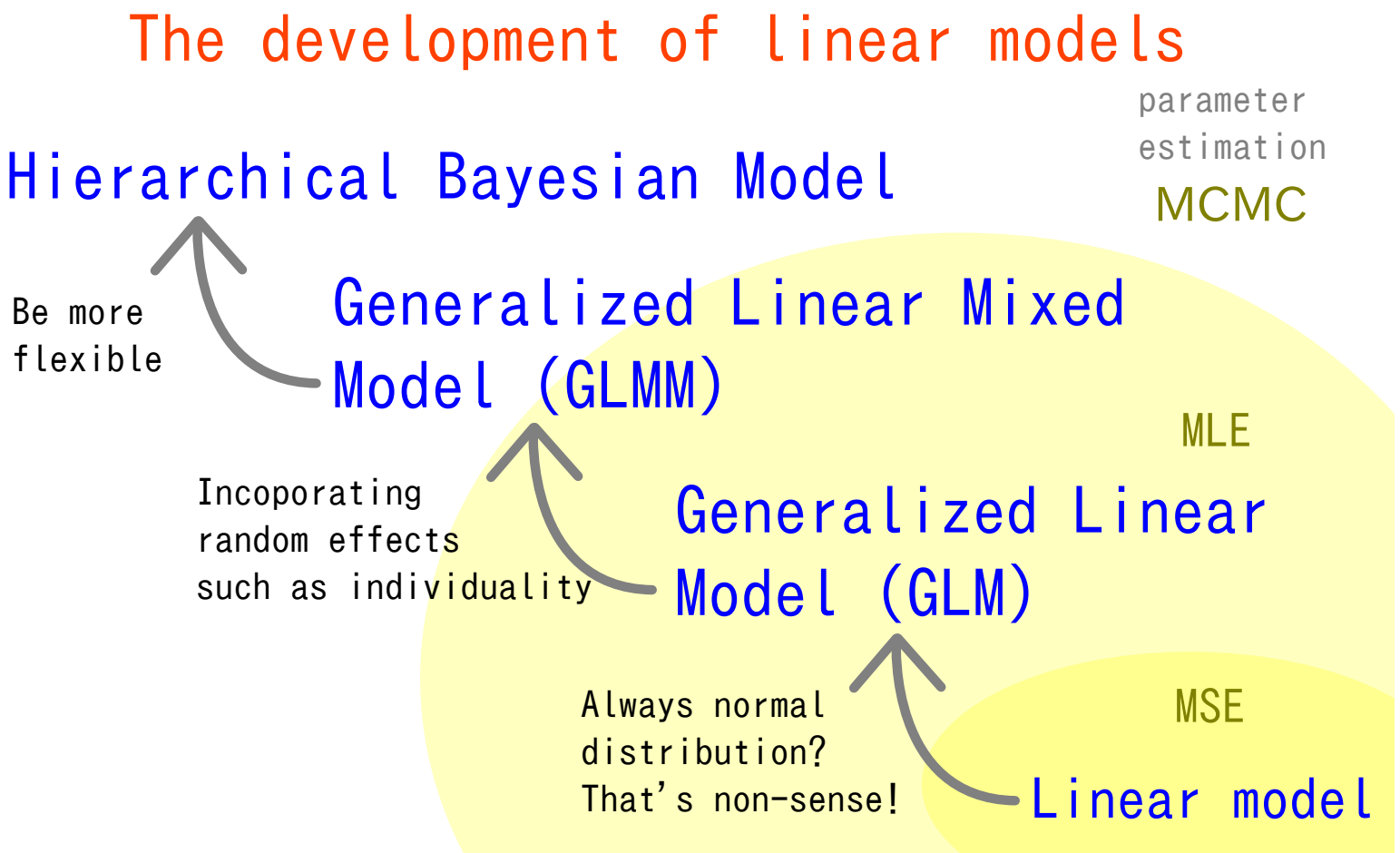
   maximum likelihood estimation of parameter λ
3. **ポアソン分布のパラメーターの 最尤推定(さいゆうすいてい)**
   もっともらしい推定?

   the functions of statistical model
4. **統計モデルの要点**
   random number, estimation and prediction
   乱数発生・推定・予測

統計モデリング授業前半の「テーマ」：データにあわせた確率分布を使ってモデル作り

statistaical models appeared in the class

この授業であつかう統計モデルたち



Kubo Doctrine: "Learn the evolution of linear-model family, firstly!"

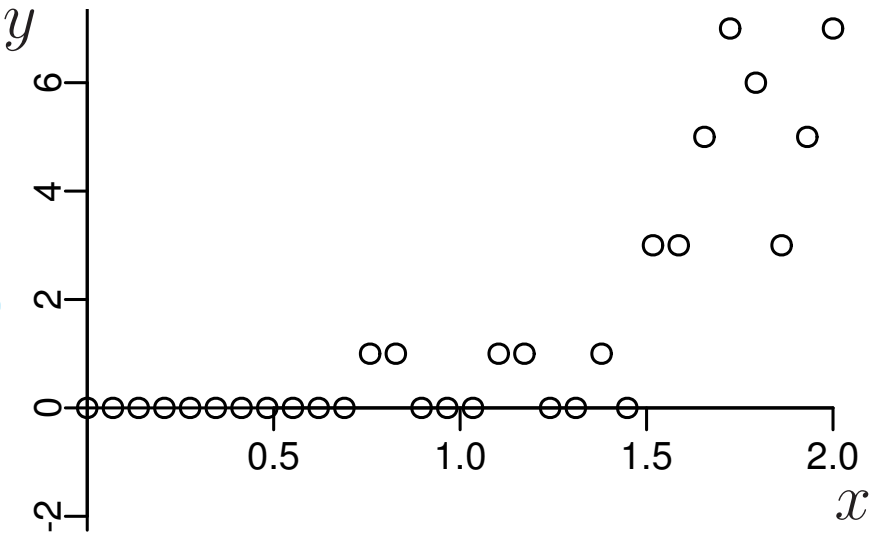
suppose that you have a “count data” set ...

0 個, 1 個, 2 個と数えられるデータ

カウントデータ $(y \in \{0, 1, 2, 3, \cdots\}$ **なデータ**$)$

response variable
応答変数

e.g. egg number
(たとえば卵数)



e.g. body size
(たとえば体重)

explanatory variable
説明変数

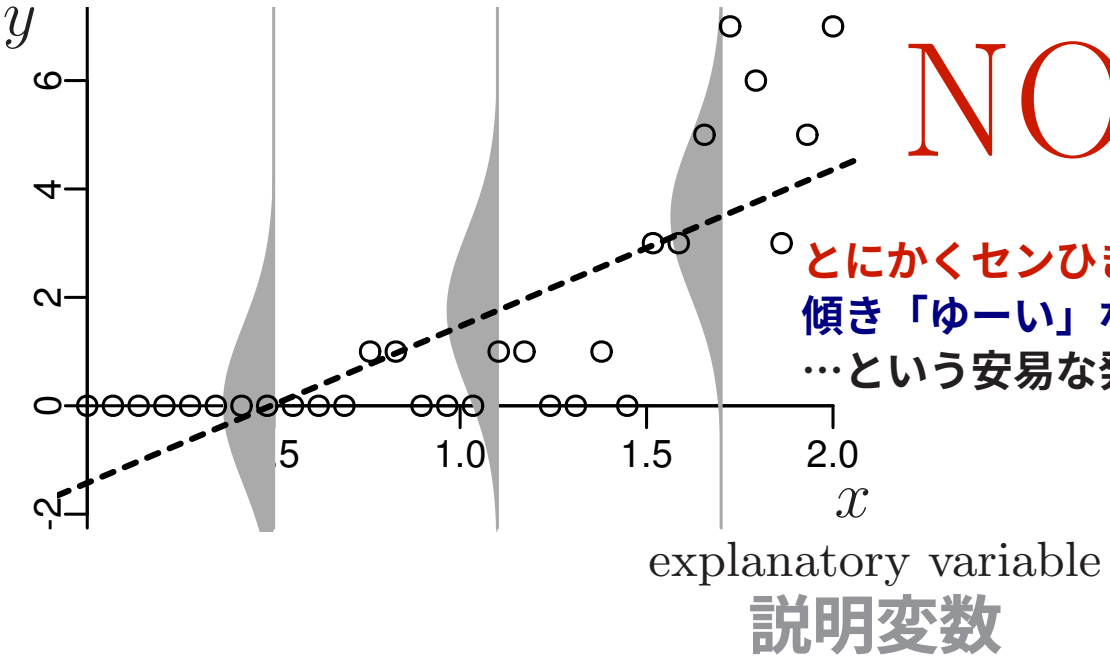
- **たとえば** x **は植物個体の大きさ，** y **はその個体の花数**
- **体サイズが大きくなると花数が増えるように見えるが……**
- **この現象を表現する統計モデルは?**

the normal distribution ... is NOT this one!

正規分布を使った統計モデル …… ムリがある?

正規分布・恒等リンク関数の統計モデル

response variable
応答変数



NO!

とにかくセンひきゃいいんでしょ
傾き「ゆーい」ならいいんでしょ
…という安易な発想のデータ解析

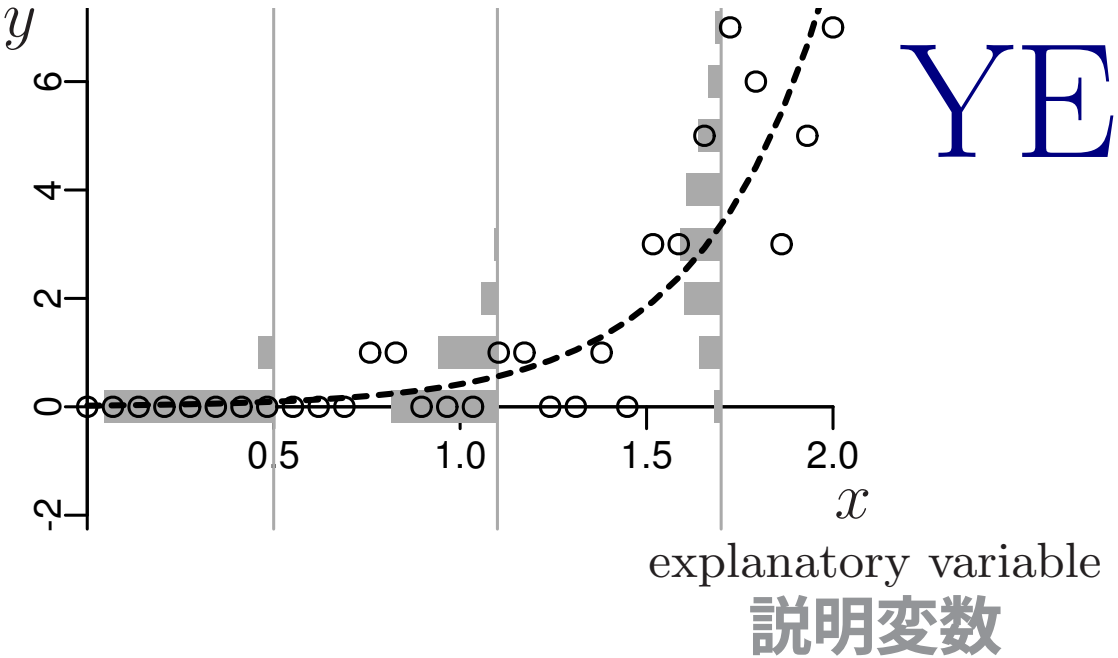
- タテ軸のばらつきは「正規分布」なのか?
- y の値は 0 以上なのに ……
- 平均値がマイナス?

the Poisson disribution approximates data

ポアソン分布を使った統計モデルなら良さそう?!

ポアソン分布・対数リンク関数の統計モデル

response variable
応答変数



YES!

- タテ軸に対応する「ばらつき」 fair distribution
- 負の値にならない「平均値」 non-negative mean
- 正規分布を使ってるモデルよりましだね bye-bye, the normal distribution

データの性質をよくみる

Plot your data and observe it

確率分布という部品を選ぶ

Choose proper distributons

「正規分布」は万能ではない!

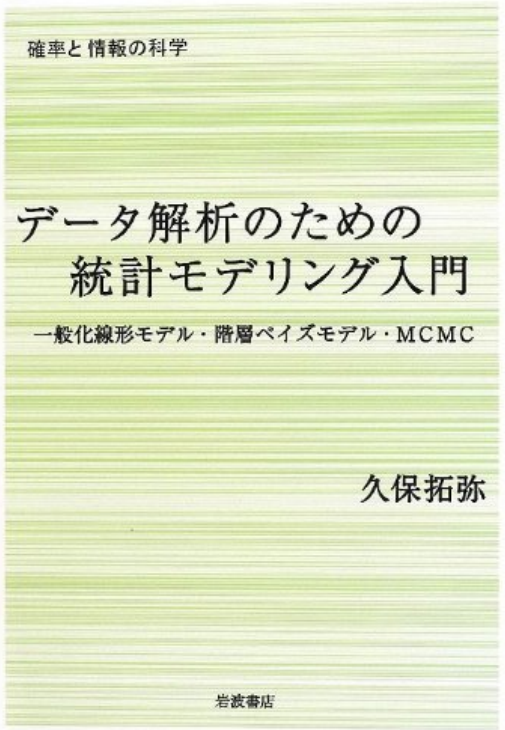
the normal distributon is NOT good at everything

今日の内容と「統計モデリング入門」との対応

http://goo.gl/Ufq2

今日はおもに「第 2 章 確率分布と統計モデルの最尤推定」の内容を説明します.

- **著者**: **久保拓弥**
- **出版社**: **岩波書店**
- 2012–05–18 **刊行**



2. 例題: 種子数の統計モデリング

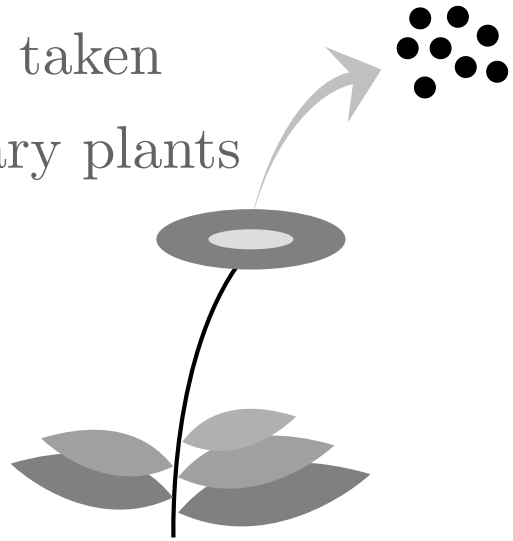
An example: a distribution of seed number

R でデータをあつかいつつ

a simplified data set, easy to understand

この授業では架空植物の架空データをあつかう

number of seeds taken
from 50 imaginary plants



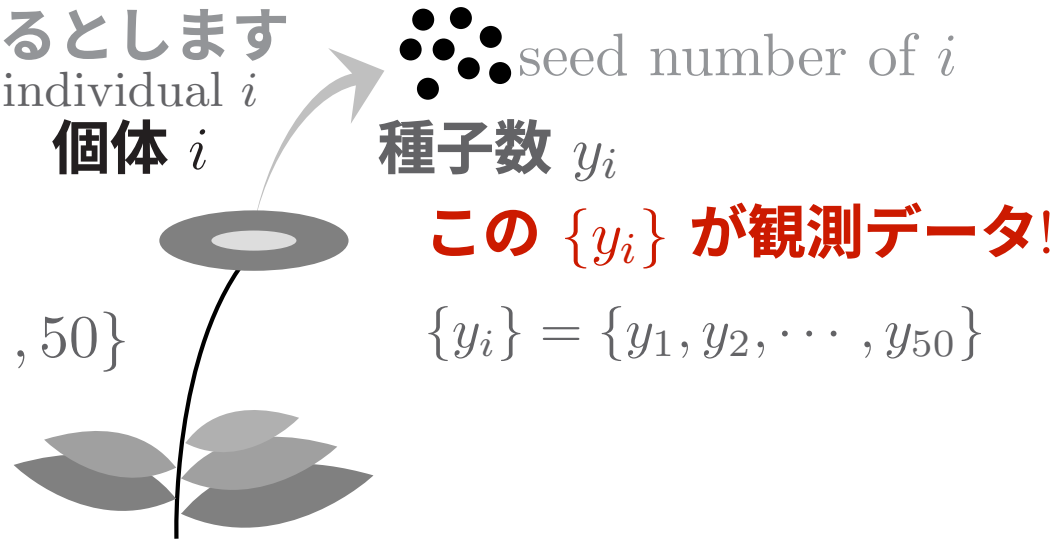
理由: よけいなことは考えなくてすむので

現実のデータはどれも授業で使うには難しすぎる……

number of seeds per plant individual

こんなデータ (架空) があったとしましょう

まあ，なんだかこういうヘンな
植物を調査しているとします



全 50 個体

$i \in \{1, 2, 3, \cdots, 50\}$

この $\{y_i\}$ が観測データ!

$\{y_i\} = \{y_1, y_2, \cdots, y_{50}\}$

このデータ $\{y_i\}$ がすでに R という統計ソフトウェアに
格納されていた，としましょう

```
> data
 [1] 2 2 4 6 4 5 2 3 1 2 0 4 3 3 3 3 4 2 7 2 4 3 3 3 4
[26] 3 7 5 3 1 7 6 4 6 5 2 4 7 2 2 6 2 4 5 4 5 1 3 2 3
```

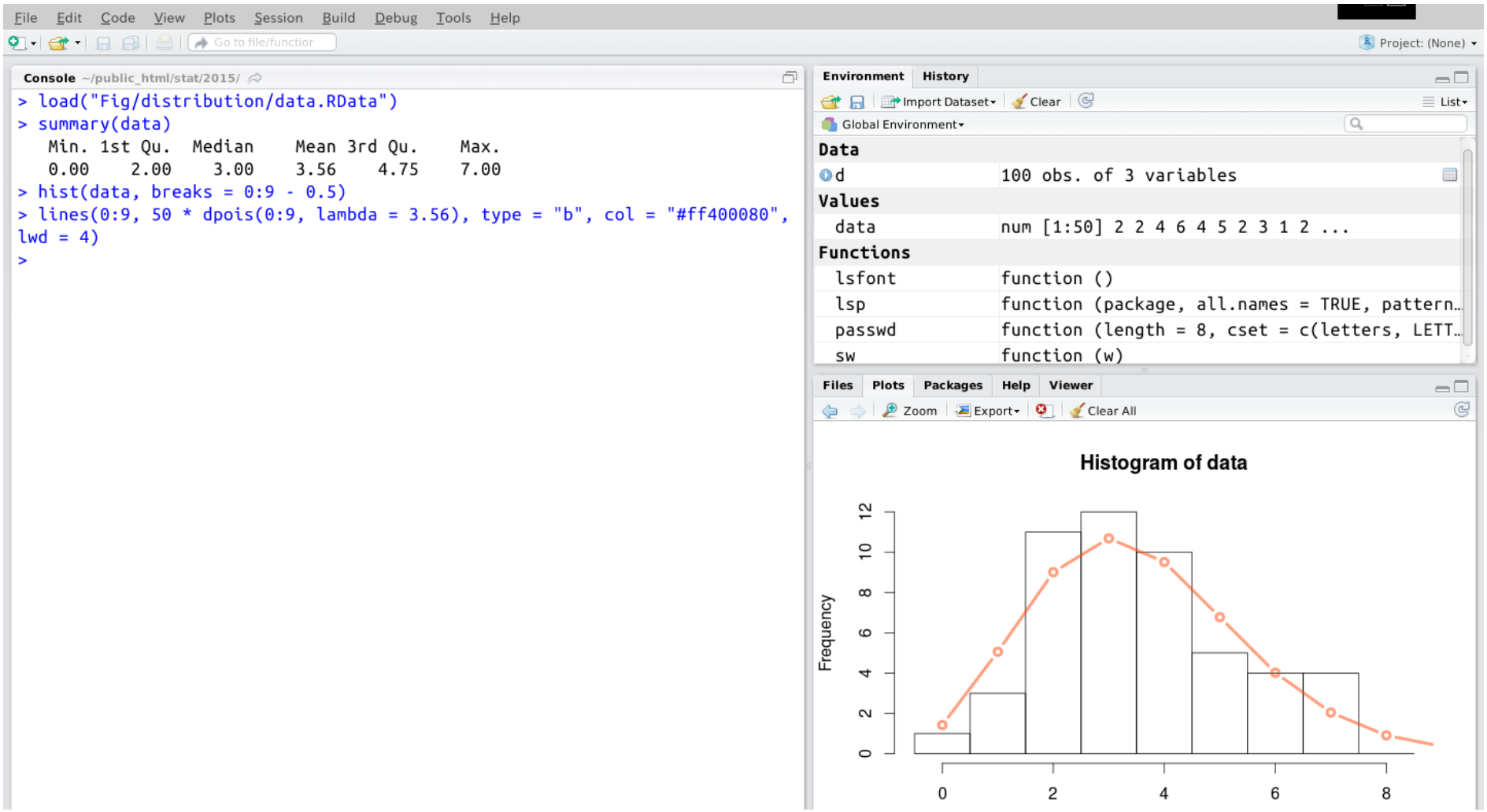
R: a free statistical software

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 free software
- 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
- RStudio http://www.rstudio.com/

RStudio



http://www.rstudio.com/

apply table() to categorize data

R でデータの様子をながめる

の table() 関数を使って種子数の頻度を調べる

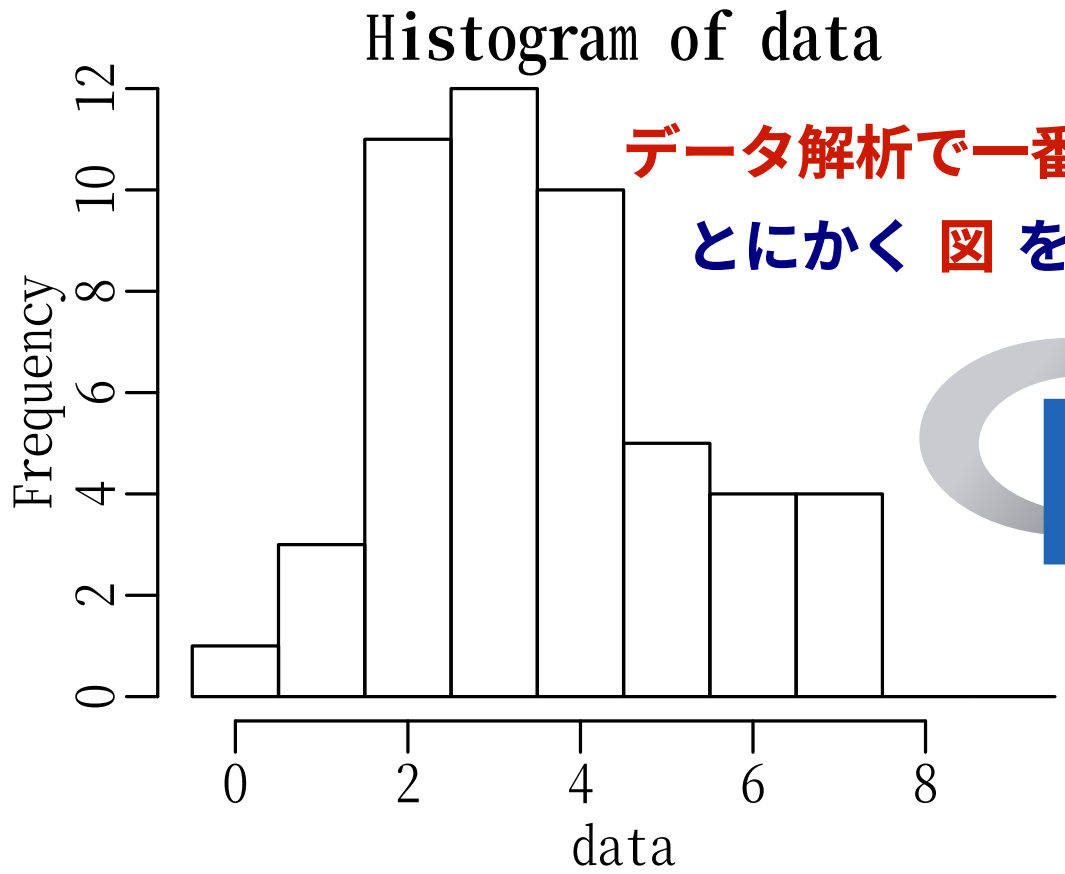
```
> table(data)
 0  1  2  3  4  5  6  7
 1  3 11 12 10  5  4  4
```

(種子数 5 は 5 個体，種子数 6 は 4 個体 ……)

start with data plotting, always

とりあえずヒストグラムを描いてみる

```
> hist(data, breaks = seq(-0.5, 9.5, 1))
```

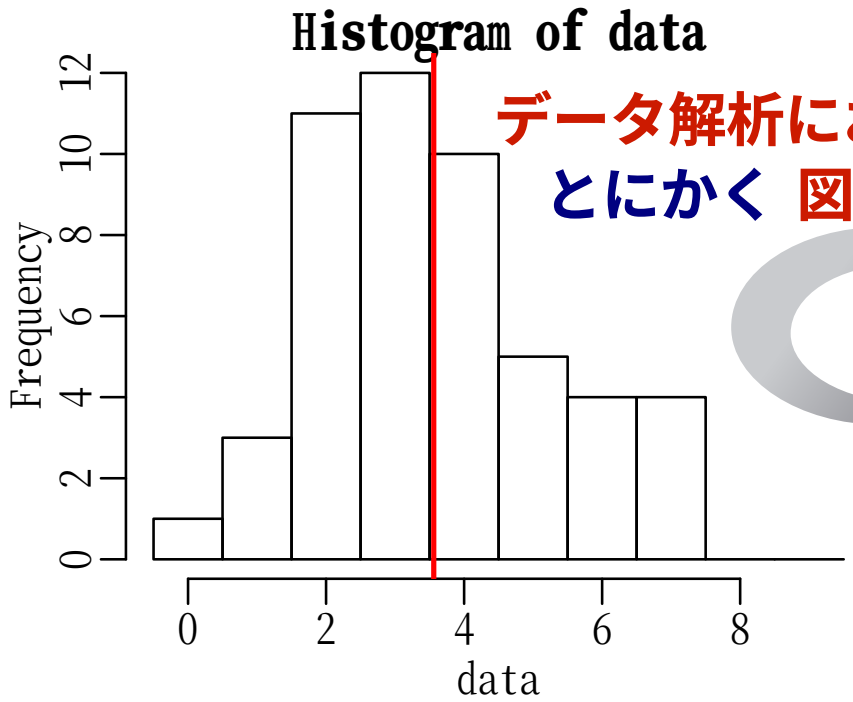


データ解析で一番たいせつなこと

とにかく 図 を描く!

How to evaluate mean value using R?

```
> mean(data)
[1] 3.56
> abline(v = mean(data))
```



データ解析における最重要事項
とにかく 図 を描く!

statistics to represent dispersion

「ばらつき」の統計量

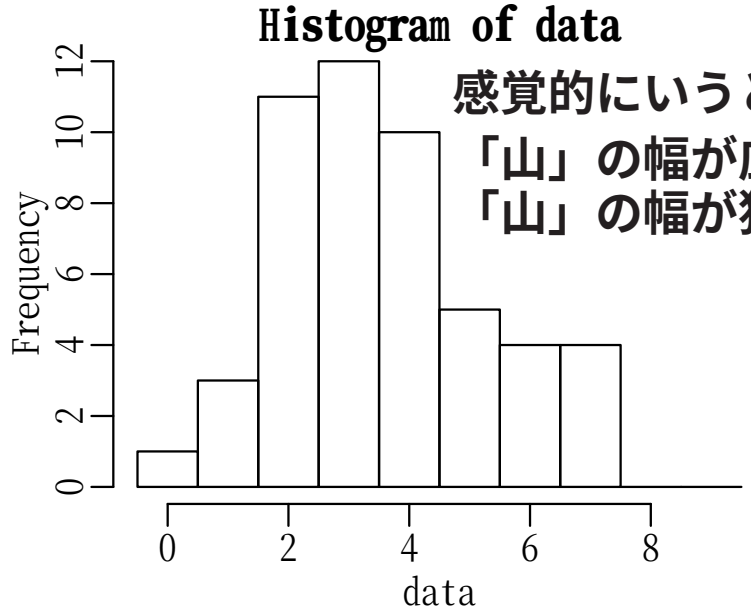
あるデータの **ばらつき** をあらわす標本統計量の例: sample variance **標本分散**

```
> var(data)
[1] 2.9861
```

sample standard deviation

標本標準偏差 とは標本分散の平方根 $(\mathrm{SD} = \sqrt{\mathrm{variance}})$

```
> sd(data)
[1] 1.7280
> sqrt(var(data))
[1] 1.7280
```



感覚的にいうと
「山」の幅が広い: 分散が大きい
「山」の幅が狭い: 分散が小さい

3. データと確率分布の対応

probability distribution, the core of statistical model

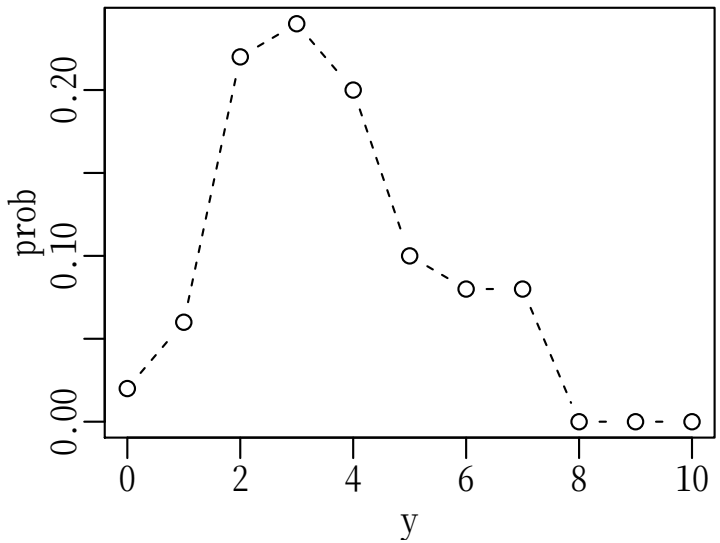
確率分布は統計モデルの重要な部品

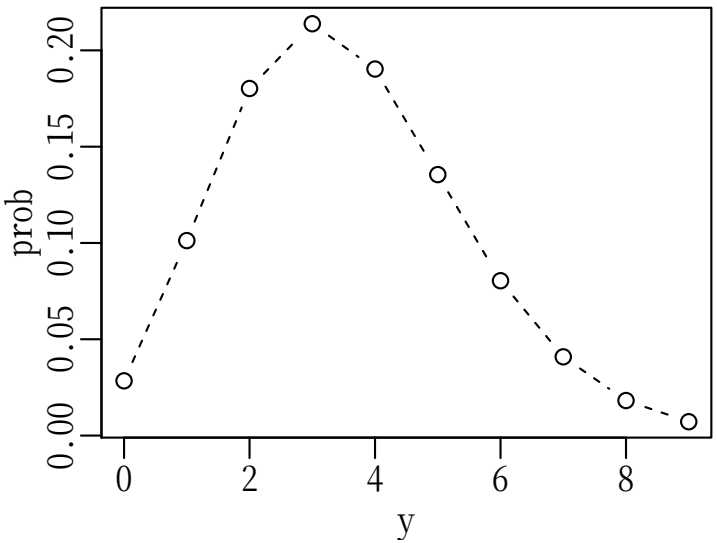
Empirical VS Theoretical Distributions

統計モデルの部品である **確率分布** には

“データそのまま” な **経験分布** と

数式で定義される**理論的な分布** がある

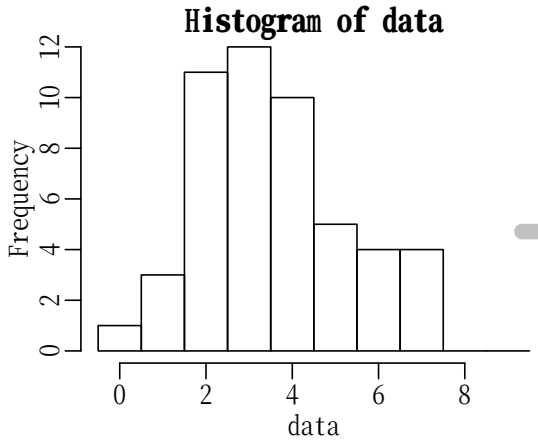


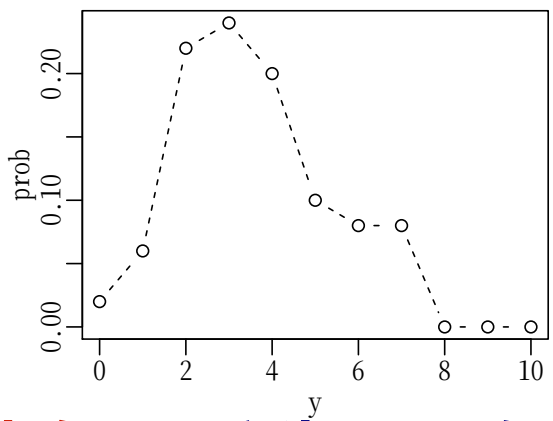


“データそのまま” な 経験分布

empirical distribution

```
> data.table <- table(factor(data, levels = 0:10))
> cbind(y = data.table, prob = data.table / 50)
```





	y	prob
0	1	0.02
1	3	0.06
2	11	0.22
3	12	0.24
4	10	0.20
5	5	0.10
6	4	0.08
7	4	0.08
8	0	0.00
9	0	0.00
10	0	0.00

- 確率分布とは **発生する事象** と **発生する確率** の対応づけ
- 確率分布のひとつである **経験分布** とは ……
  “たまたま手もとにある” データから
  “発生確率” を決める確率分布

なるほど**経験分布**は“直感的”かもしれないが……

- **データが変わると確率分布が変わる**?
- **種子数** $y = \{0, 1, 2, \cdots\}$ **となる確率が，個々におたがい無関係に決まる**?
- **パラメーターは** $\{p_0, p_1, p_2, \cdots, p_{99}, p_{100}, \cdots\}$ **無限個ある**?

道具として使うには，ちょっと不便かもしれない……

なにか理論的に導出された確率分布のほうが便利ではないか?

- 少数のパラメーターで分布の“カタチ”が決まる
- “なめらかに”確率が変化する
- いろいろと数理的な道具が準備されている (パラメーター推定方法など)

Mathematical expression of the Poisson distribution

確率分布 (ポアソン分布) を数式で決めてしまう

種子数が y である probability 確率 は以下のように決まる，と考えている

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

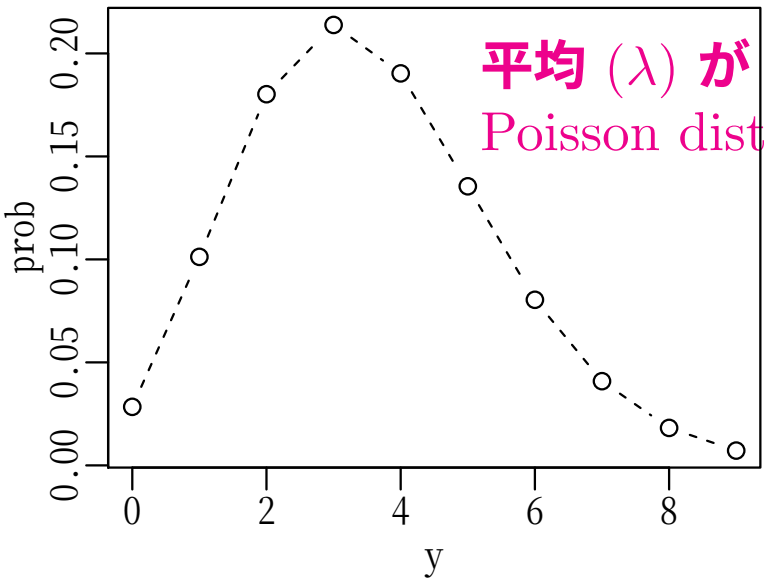
- $y!$ は y の factorial 階乗 で，たとえば $4!$ は $1 \times 2 \times 3 \times 4$ をあらわしています．
- $\exp(-\lambda) = e^{-\lambda}$ のこと $(e = 2.718\cdots)$
- ここではなぜポアソン分布の確率計算が上のようになるのかは説明しません— まあ，こういうもんだと考えて先に進みましょう

the Poisson distribution

数式で決められたポアソン分布?

とりあえず R で作図してみる

```
> y <- 0:9                                  # これは種子数（確率変数）
> prob <- dpois(y, lambda = 3.56)           # ポアソン分布の確率の計算
> plot(y, prob, type = "b", lty = 2)
```

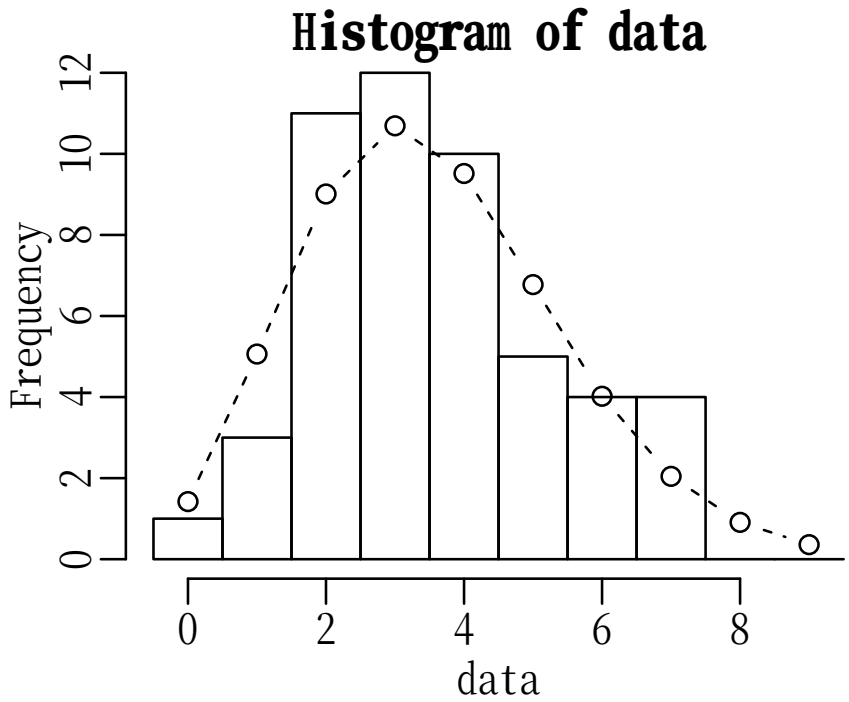


平均 (λ) が 3.56 **である**
Poisson distribution

```
> # cbind で「表」作り
> cbind(y, prob)
    y       prob
1   0 0.02843882
2   1 0.10124222
3   2 0.18021114
4   3 0.21385056
5   4 0.19032700
6   5 0.13551282
7   6 0.08040427
8   7 0.04089132
9   8 0.01819664
10  9 0.00719778
```

the Poisson distribution represent data?

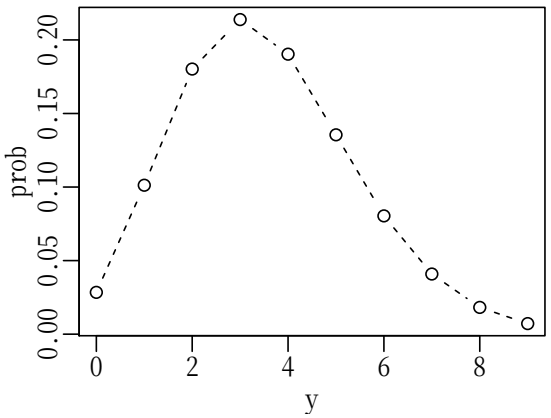
データとポアソン分布を重ね合わせる




```
> hist(data, seq(-0.5, 8.5, 0.5))      # まずヒストグラムを描き
> lines(y, prob, type = "b", lty = 2) # その「上」に折れ線を描く
```

parameter λ is the mean of the Poisson distribution

パラメーター λ はポアソン分布の平均




```
> # cbind で「表」作り
> cbind(y, prob)
    y       prob
1   0 0.02843882
2   1 0.10124222
3   2 0.18021114
4   3 0.21385056
5   4 0.19032700
6   5 0.13551282
7   6 0.08040427
8   7 0.04089132
9   8 0.01819664
10  9 0.00719778
```

- 平均 λ はポアソン分布の唯一の**パラメーター**
- 確率分布の平均は λ である（$\lambda \geq 0$）
- 分散と平均は等しい: $\lambda =$ 平均 $=$ 分散
- $y \in \{0,\ 1,\ 2,\ \cdots,\ \infty\}$ の値をとり，すべての y について和をとると 1 になる

$$\sum_{y=0}^{\infty} p(y \mid \lambda) = 1$$

The Poisson distribution is useful if ...

どういう場合にポアソン分布を使う?

統計モデルの部品としてポアソン分布が選んだ理由:

- **データに含まれている値 y_i が $\{0,\ 1,\ 2,\ \cdots\}$ といった非負の整数である（カウントデータ である）**
  y_i: count data
- **y_i に下限（ゼロ）はあるみたいだけど上限はよくわからない**
- **この観測データでは 平均と分散がだいたい 等しい**
  $\text{mean} \approx \text{variance}$
  - **このだいたい等しいがあやしいのだけど，まあ気にしないことにしましょう**

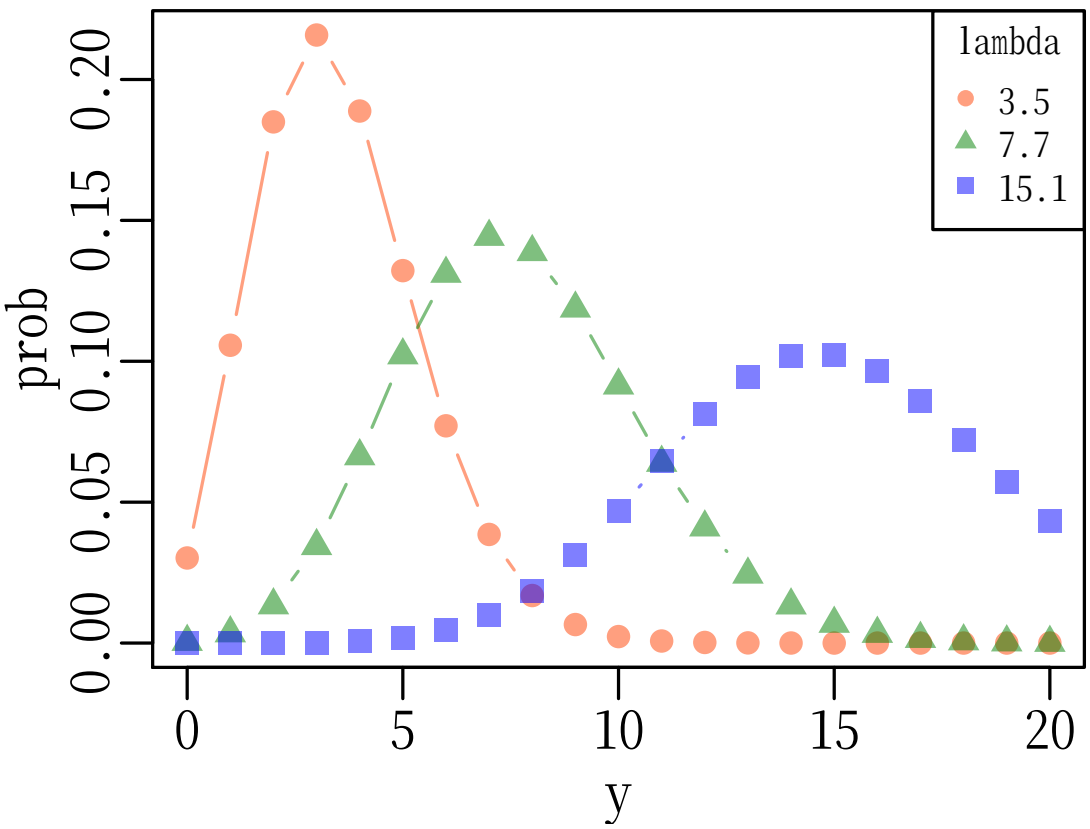
λ changes the shape of distribution

ポアソン分布の λ を変えてみる

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

mean

λ は 平均 をあらわすパラメーター



maximum likelihood estimation of parameter λ

4. ポアソン分布のパラメーターの 最尤推定（さいゆうすいてい）

もっとももっともらしい推定?

"fitting" = "parameter estimation"

「あてはめる」ことは推定すること

尤度 (likelihood) とは何か?

- 最尤推定法 (maximum likelihood estimation) では，尤度 (ゆうど) というあてはまりの良さをあらわす統計量に着目
- 尤度はデータが得られる確率をかけあわせたもの
- この例題の場合，パラメーター λ を変えると尤度が変わる
- もっとも「あてはまり」(goodness of fit) が良くなる λ を見つけたい
- たとえば，いまデータが 3 個体ぶん，たとえば，$\{y_1,\ y_2,\ y_3\} = \{2,\ 2,\ 4\}$，これだけだった場合，尤度はだいたい $0.180 \times 0.180 \times 0.19 = 0.006156$ といった値になる

likelihood $L(\lambda)$ depends on the value of mean, λ

尤度 $L(\lambda)$ はパラメーター λ の関数

この例題の尤度 (the likelihood definition for the example):

$$
\begin{aligned}
L(\lambda) &= (y_1\text{が } 2 \text{ である確率}) \times (y_2\text{が } 2 \text{ である確率}) \\
&\quad \times \cdots \times (y_{50}\text{が } 3 \text{ である確率}) \\
&= p(y_1 \mid \lambda) \times p(y_2 \mid \lambda) \times p(y_3 \mid \lambda) \times \cdots \times p(y_{50} \mid \lambda) \\
&= \prod_i p(y_i \mid \lambda) = \prod_i \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!},
\end{aligned}
$$

evaluate not likelihood, but log likelihood!

尤度はしんどいので対数尤度を使う

尤度は確率 (あるいは確率密度) の積であり，あつかいがふべん (大量のかけ算!)

そこで，パラメーターの最尤推定では，**対数尤度関数** (log likelihood function) を使う

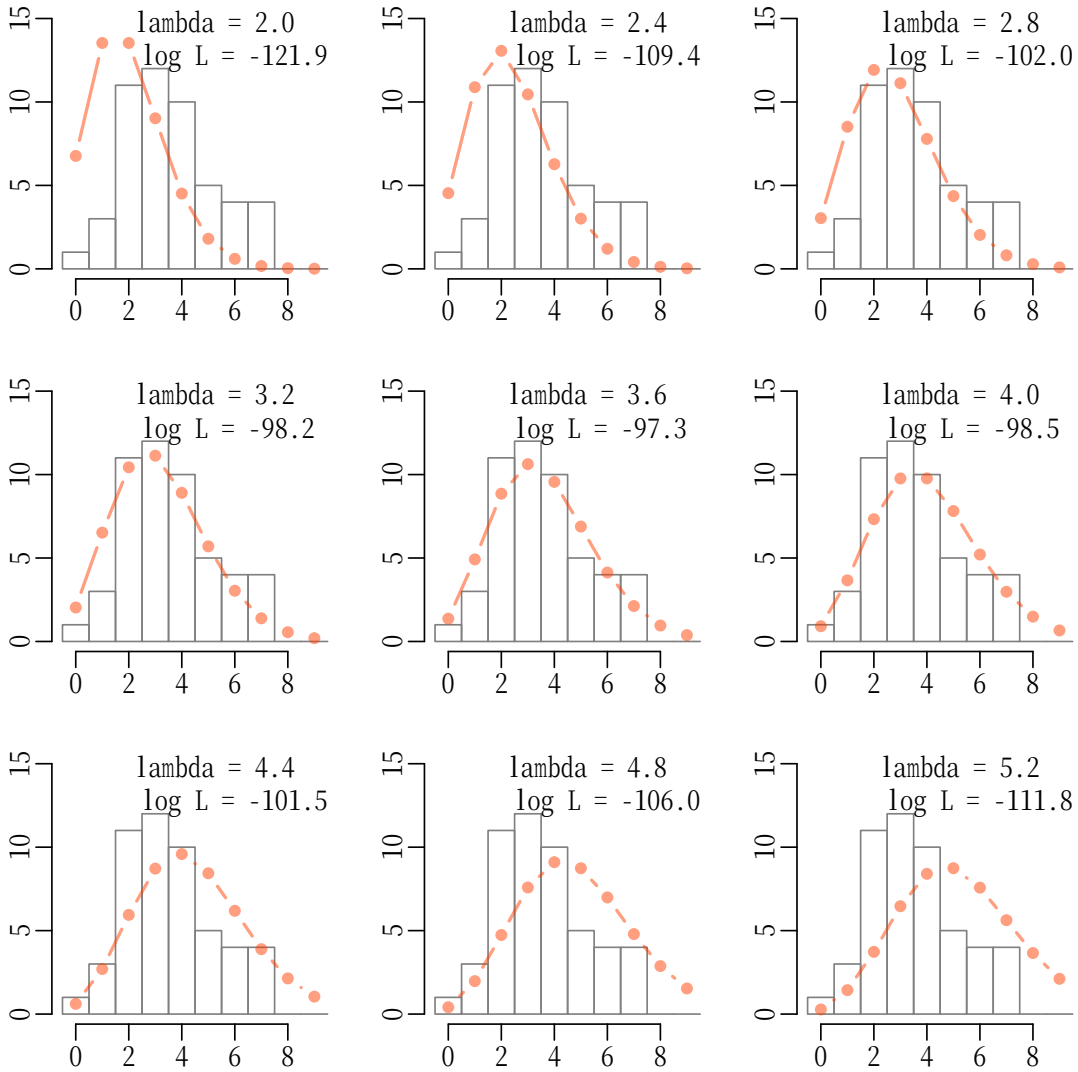
$$\log L(\lambda) = \sum_i \left(y_i \log \lambda - \lambda - \sum_k^{y_i} \log k \right)$$

対数尤度 $\log L(\lambda)$ の最大化は尤度 $L(\lambda)$ の最大化になるから

まずは，平均をあらわすパラメーター λ を変化させていったときに，ポアソン分布のカタチと対数尤度がどのように変化するのかを調べてみましょう

λ changes the log likelihood, i.e., goodness of fit

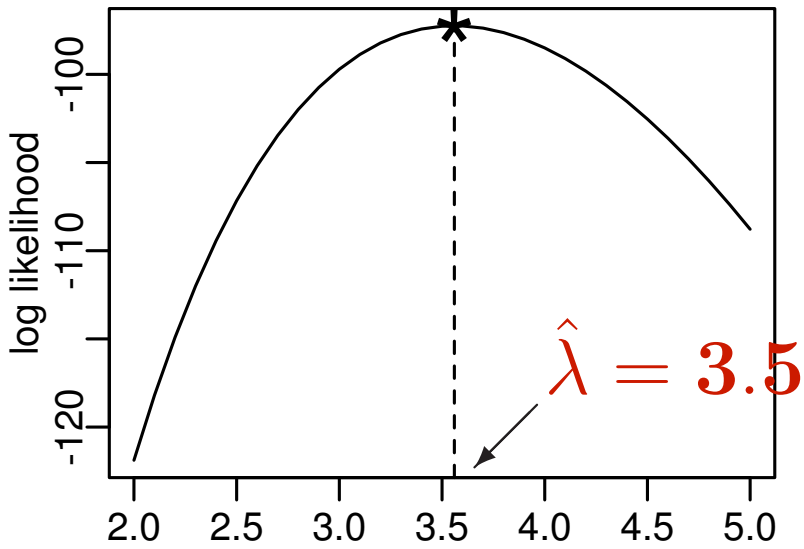
λ を変えるとあてはまりの良さが変わる



seek the maximum likelihood estimate, $\hat{\lambda}$

対数尤度を最大化する $\hat{\lambda}$ をさがす

対数尤度 $\log L(\lambda) = \sum_i \left(y_i \log \lambda - \lambda - \sum_k^{y_i} \log k \right)$



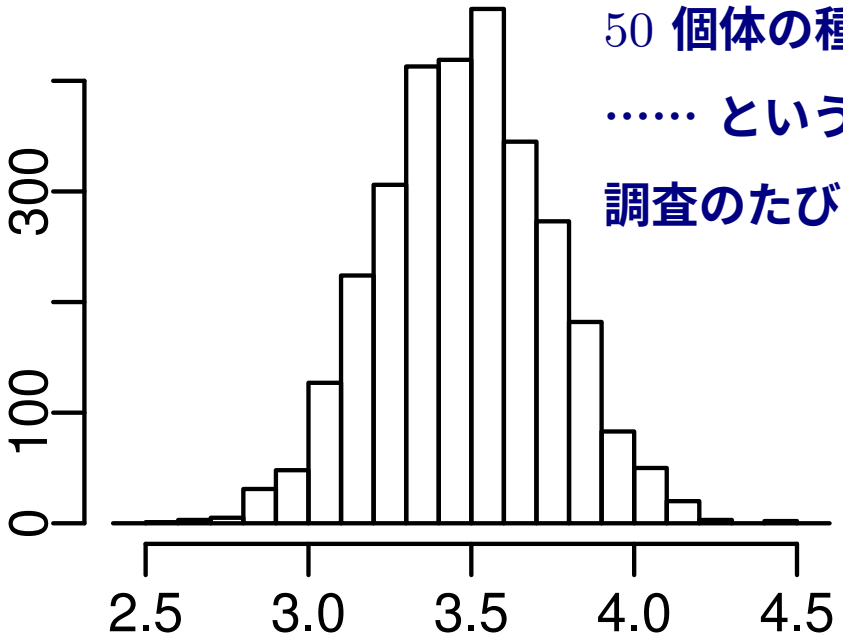
$\frac{d \log L}{d\lambda} = 0$ より

- **最尤推定量** (ML estimator): $\sum_i y_i/50$ 標本平均値!
- **最尤推定値** (ML estimate): $\hat{\lambda} = 3.56$ **ぐらい**

no one knows “the true λ” based on finite size data

最尤推定を使っても真の λ は見つからない

真の λ が 3.5 と設定して架空データを生成



50 個体の種子数を調べる

…… ということを 3000 回くりかえし

調査のたびに $\hat{\lambda}$ を最尤推定した

データは有限なので真の λ はわからない

標本サイズが 50 の場合，“平均値の推定” すらなかなかうまくできない

the functions of statistical model

5. 統計モデルの要点

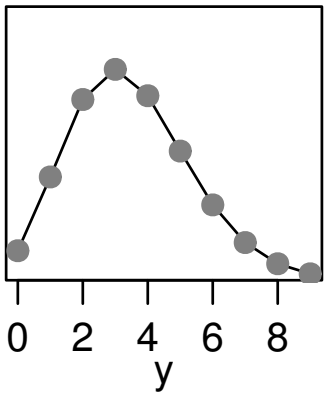
random number, estimation and prediction

乱数発生・推定・予測

統計モデルとデータの対応づけ

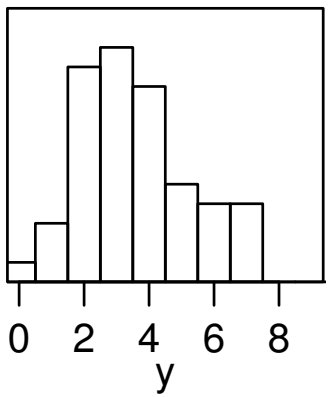
確率分布: random number generation 乱数発生 と estimation 推定

（人間には見えない）
真の統計モデル
$\lambda = 3.5$ のポアソン分布



データをサンプル

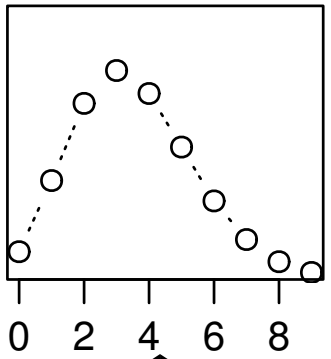
確率分布から乱数を発生



観測されたデータ

パラメーター推定

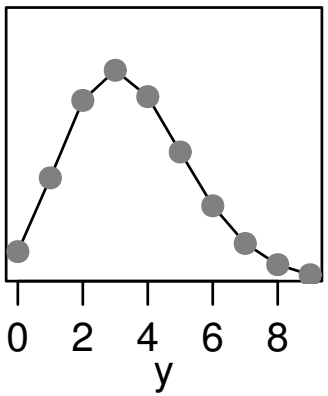
観測データから
推定された
$\hat{\lambda} = 3.56$ のポアソン分布

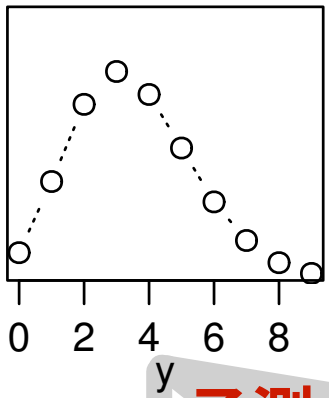


データ？…ここでは確率・統計モデルが生成していると仮定

推定されたモデルを使った prediction 予測

（人間には見えない）
真の統計モデル
$\lambda = 3.5$ のポアソン分布

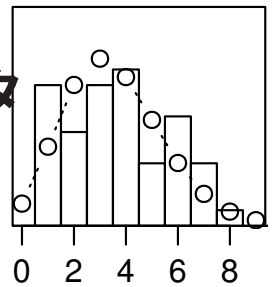


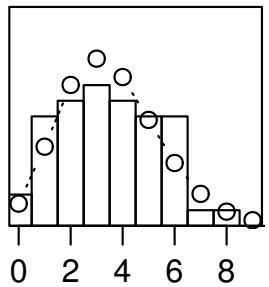


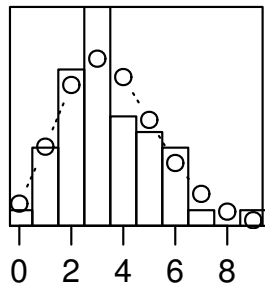
観測データから
推定された
$\hat{\lambda} = 3.56$ のポアソン分布

予測: 新しいデータに
あてはまるのか？
(**予測** の良さを調べている

新しいデータ
をサンプル







・・・

同じ調査方法で得られた新データ

probability distributions appeared in the class

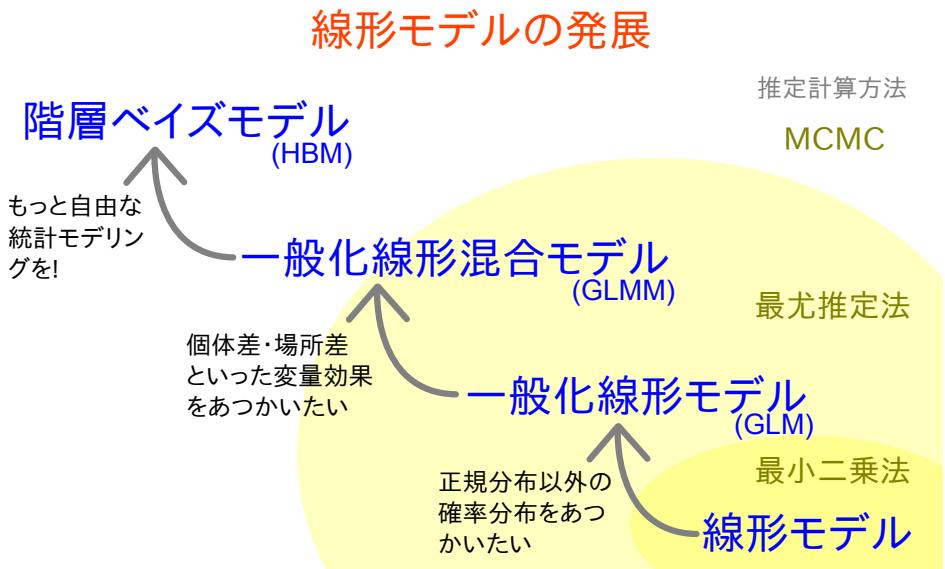
この講義で登場する確率分布

- **ポアソン分布**: $y \in \{0, 1, 2, 3, \cdots\}$ となるデータ，「y 回なにかがおこった」
- **二項分布**: $y \in \{0, 1, 2, \cdots, N\}$ となるデータ，「N 個のうち y 個で何かがおこった」
- **正規分布**: $-\infty < y < \infty$ の連続値をとるデータ
- その他あれこれ — ちょっと登場するだけ

そんなに多くの確率分布は登場しません

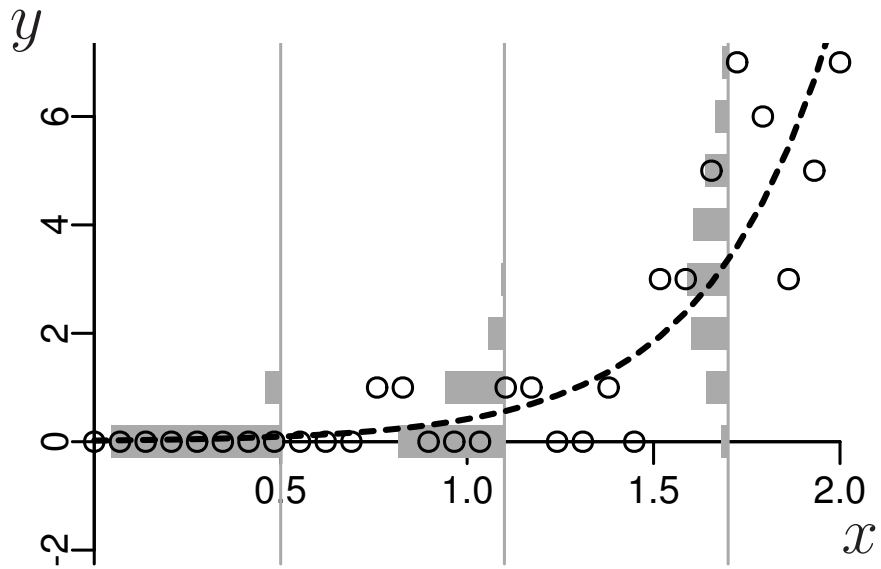
いろいろな確率分布があるけれど……

- この講義では多種多様な確率分布を**あつかいません**
- しかし **確率分布を混ぜあわせる** ことによって，自分で確率分布を作り出すことができます
- ハナシの後半に登場する GLMM や階層ベイズモデル



次回予告

The next topic



YES!

一般化線形モデルのひとつ: ポアソン回帰

Poisson Regression, a Generalized Linear Model (GLM)