統計モデリング入門 新潟大 2015 (1)

統計モデル・確率分布・最尤推定

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

新潟大学集中講義 http://goo.gl/m8HSBM

2015–05–26

ファイル更新時刻: 2015–05–21 14:02

1. はじめに

とりあえず，全体のながれなど

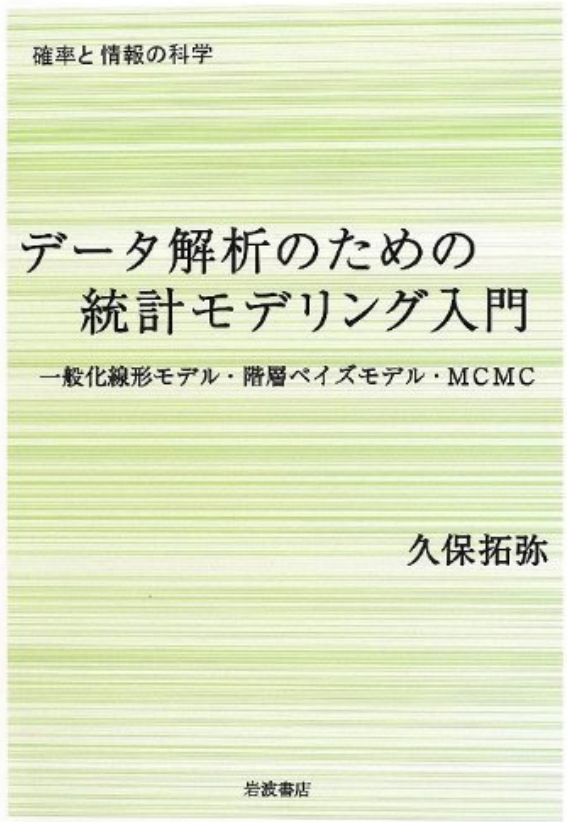
簡単な自己紹介その他あれこれ

とりあえず簡単な自己紹介: 久保拓弥 (北大・環境科学)

研究: 生態学データの統計モデリング

- 自分ではデータをとらない (野外調査・実験などをやらない) で，他のみなさんのデータ解析をすることが専門です
- これではあまりにも寄生者的なので，ときどきデータ解析に必要な統計モデリングの解説みたいなことをしております……

統計モデリングの教科書も書きました!

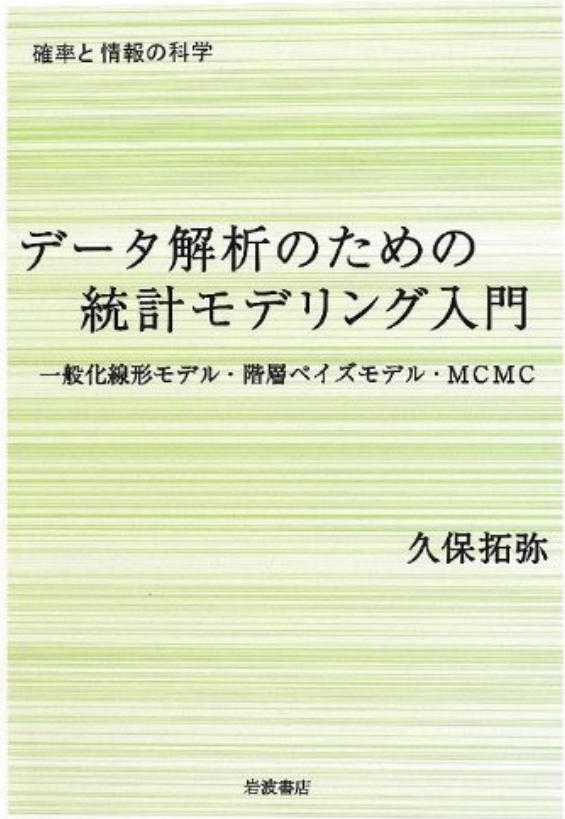


なんで，そんな本なんか書いたの?!

生態学の統計解析はあまりおもしろくなかった

- 他人の論文の method section を読んで，内容を理解しないまま同じソフトウェアを使って，$p < 0.05$ なら何でも OK といった作業になりがち
- 統計ソフトウェアが何をやっているのかわかっていないので，誤用が多い
- こういう発想は，計算環境が貧弱だった昔の遺物

この本ではブラックボックス統計学として批判



カタチ だけまねをするデータ解析

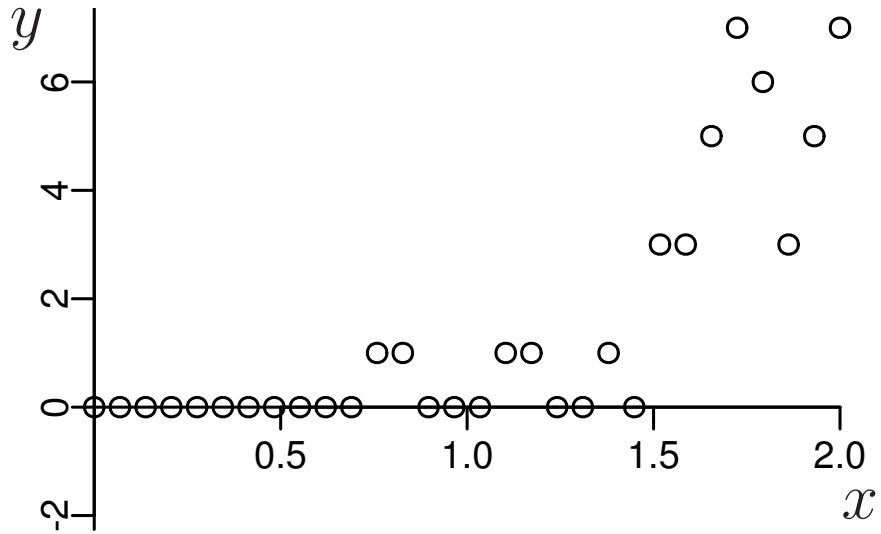
何がよくないのか?

例をあげて考えてみましょう

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**$)$

応答変数



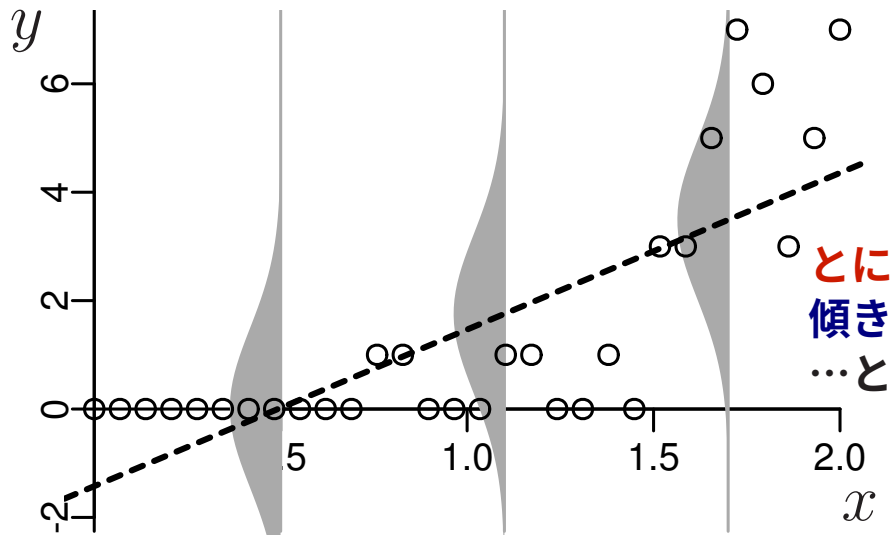
説明変数

- **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?**

“何でもかんでも**直線あてはめ**”という安易な発想……はギモン

正規分布・恒等リンク関数の統計モデル

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

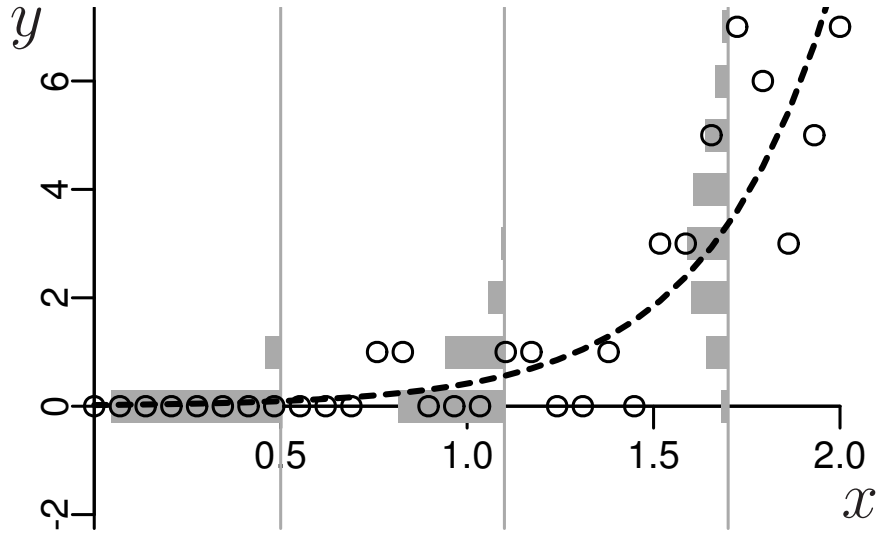
説明変数

- 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?

データにあわせた“統計モデル”つかうとマシかもね?

ポアソン分布・対数リンク関数の統計モデル

応答変数

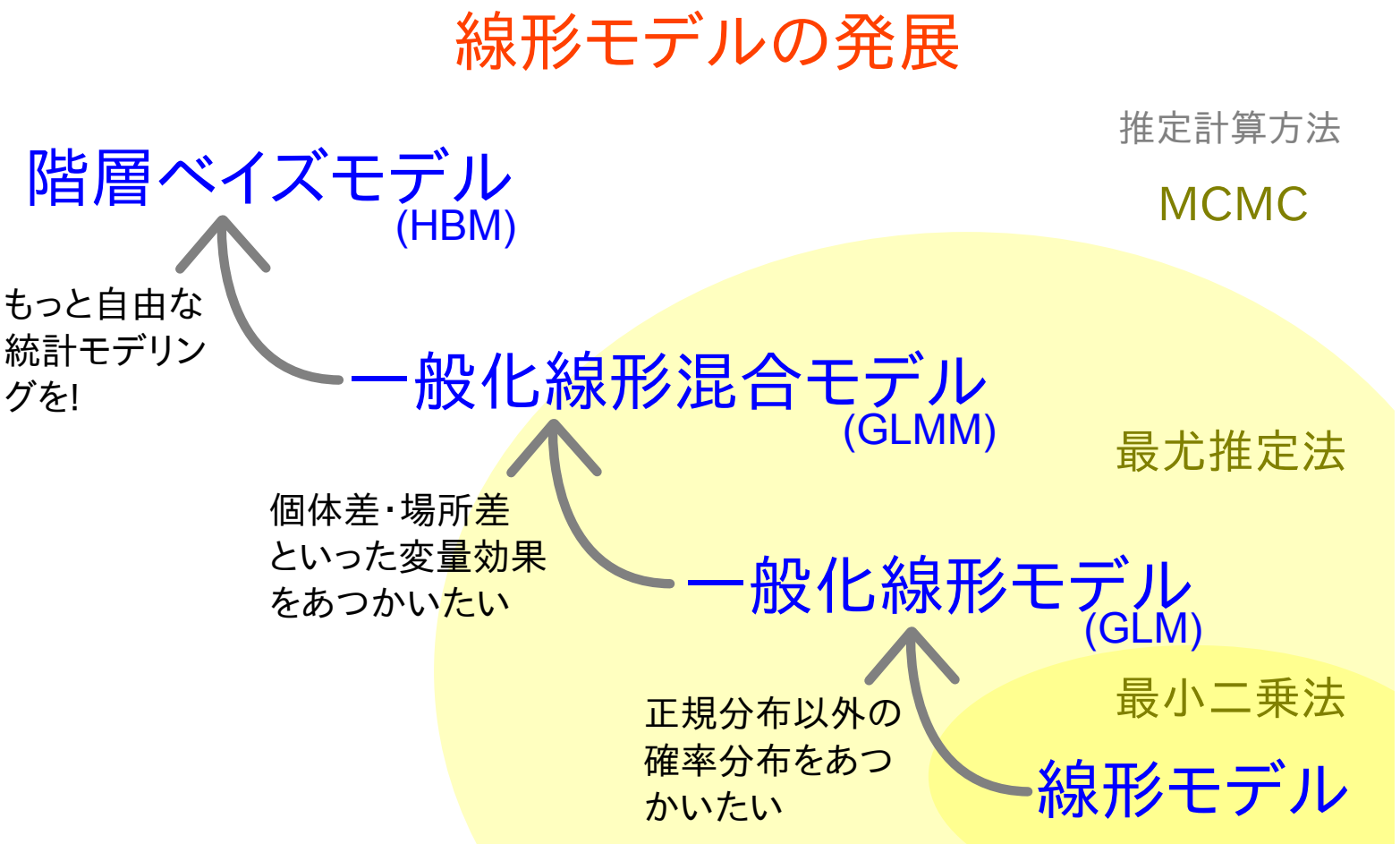


YES!

説明変数

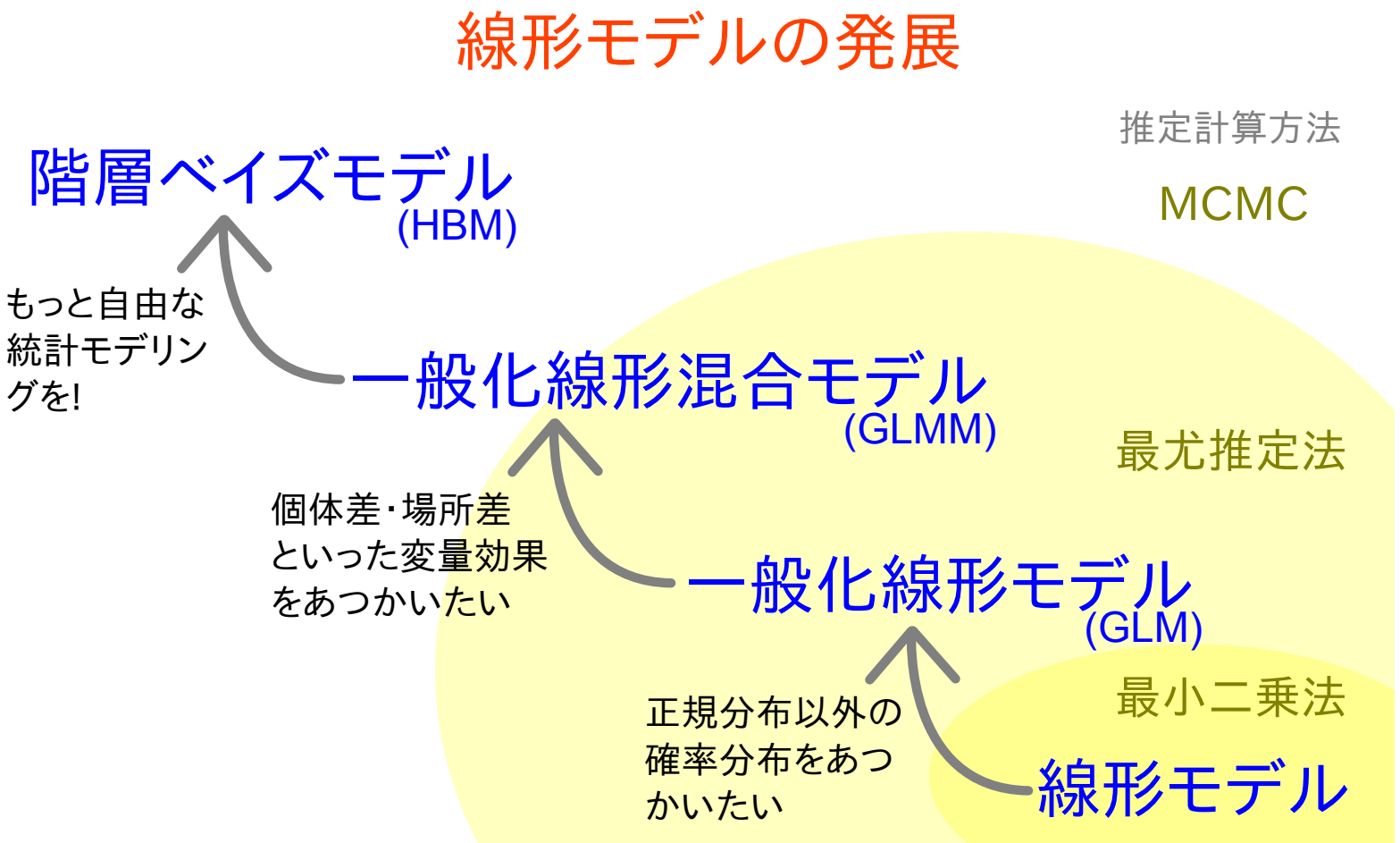
- タテ軸に対応する「ばらつき」
- 負の値にならない「平均値」
- 正規分布を使ってるモデルよりましだね

“統計モデリング入門”に登場する統計モデル



データの特徴にあわせて線形モデルを改良・発展させる

この集中講義で勉強する統計モデル



ひとことでいうと　**「直線あてはめ」をどんどん改善する**

集中講義の流れ: いずれも例題 driven なかんじで

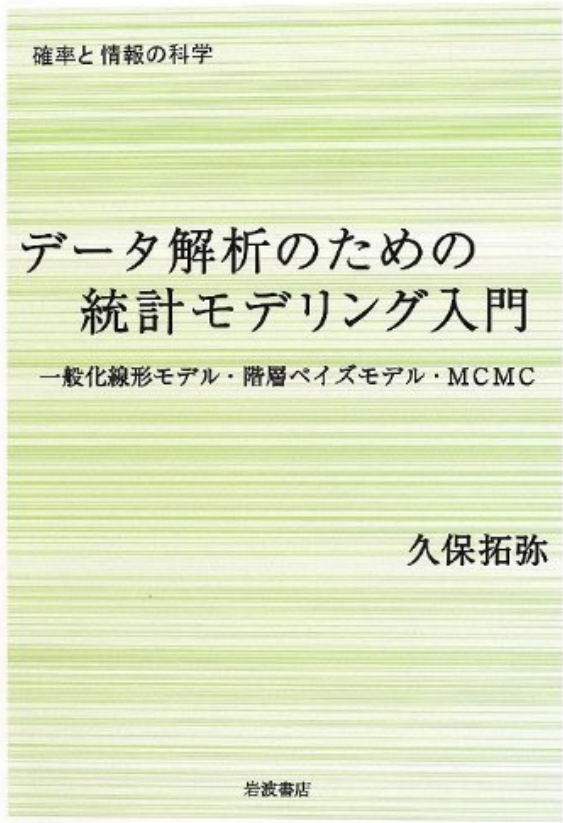
1. **統計モデル・確率分布・最尤推定**
2. **ポアソン分布の一般化線形モデル** (GLM)
3. **二項分布の** GLM **と** GLMM
4. MCMC **と階層ベイズモデル**

単純化した例題にそって統計モデルを説明

統計モデルって何?

どんな統計解析においても統計モデルが使用されている

- 観察によって**データ化された現象**を説明するために作られる
- **確率分布**が基本的な部品であり，これはデータにみられるばらつきを表現する手段である
- **データとモデルを対応づける手つづき**が準備されていて，モデルがデータにどれぐらい良くあてはまっているかを定量的に評価できる



この時間に説明したいこと

統計モデルの重要な部品: 確率分布

- データ解析をするために**統計モデル**が必要
- 統計モデルの部品として**“データにあった”確率分布**が必要
- 確率分布は**パラメーター**などを指定する必要がある
- **パラメーターの値**はデータに基づいて決めたい

1. サイコロの統計モデル

もっとも簡単な例のひとつとして

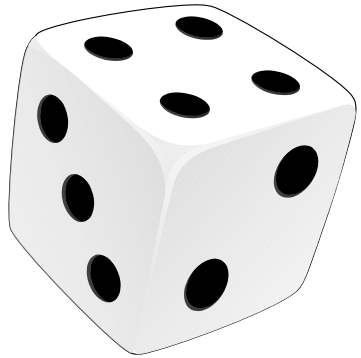
これで統計モデルの概念を考えよう

「統計モデル」とは何か?

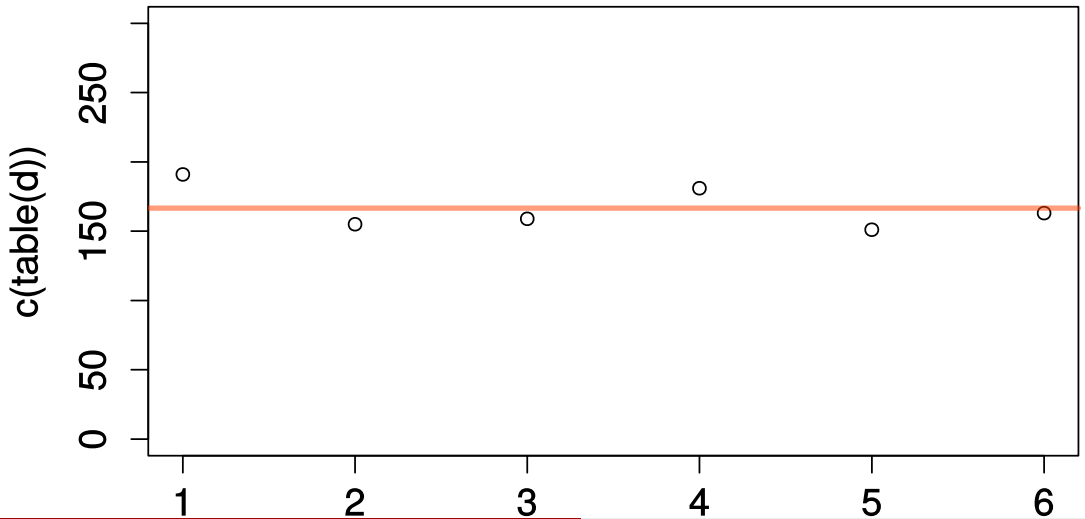
どんな統計解析においても
統計モデルが使用されている

- 観察によってデータ化された現象を説明するために作られる
- 確率分布が基本的な部品であり，これはデータにみられるばらつきを表現する手段である
- データとモデルを対応づける手つづきが準備されていて，モデルがデータにどれぐらい良くあてはまっているかを定量的に評価できる

「サイコロの統計モデル」を考えよう



```
> load("dice.RData")
> length(d)
[1] 1000
> table(d)
d
  1   2   3   4   5   6
191 155 159 181 151 163
> plot(1:6, c(table(d)), ylim = c(0, 300))
> abline(h = 1000 / 6, col = "#ff400080", lwd = 3)
```



架空データ

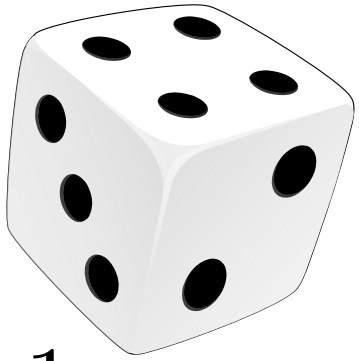
1000回サイコロふった

1000/6 = 166.66...?

「サイコロ」の確率分布は?

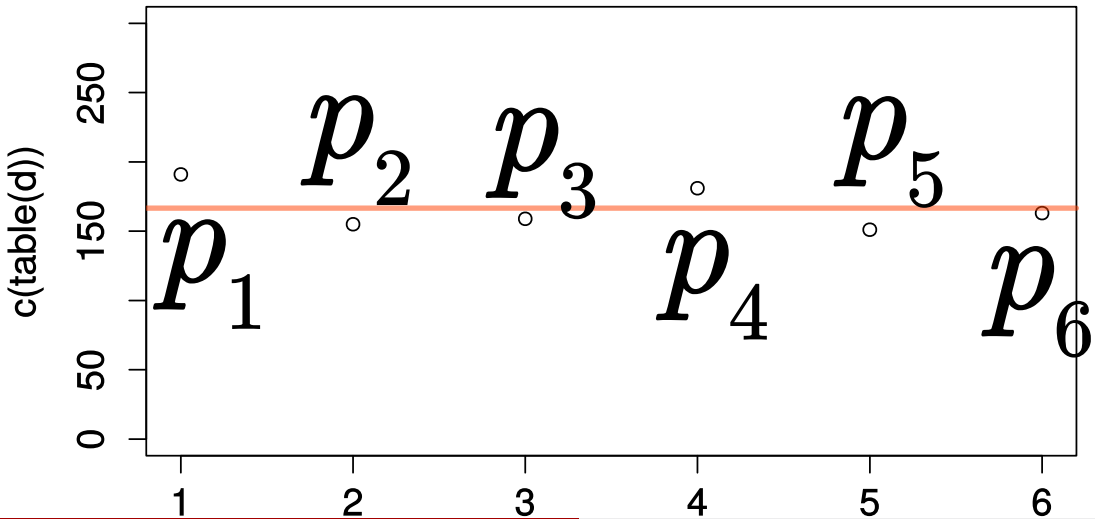
Categorical Distribution

カテゴリカル分布



$$p(Y = k) = p_k \qquad \sum_{k=1}^{6} p_k = 1$$

$$k \in \{1, 2, 3, 4, 5, 6\}$$



架空データ

1000回サイコロふった

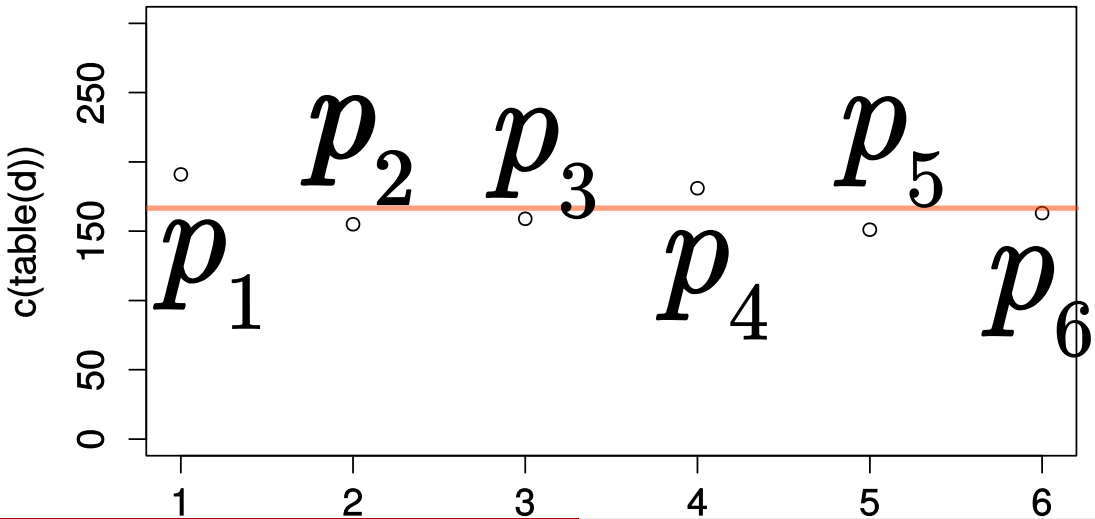
1000/6 = 166.66...?

確率分布のパラメーターは $\{p_k\}$

最尤推定量

$$\hat{p}_k = \frac{k \text{ の目が出た回数}}{1000}$$

```
> table(d)
d
  1   2   3   4   5   6
191 155 159 181 151 163
```



架空データ

1000回サイコロふった

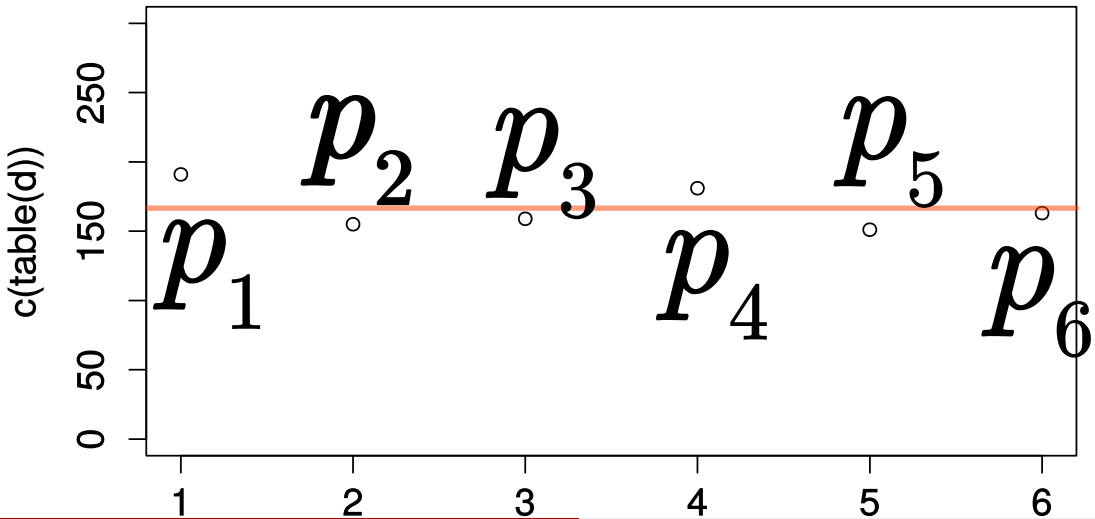
1000/6 = 166.66...?

「サイコロの統計モデル」にできること

パラメーターの推定

$$\hat{p}_k = \frac{k \text{ の目が出た回数}}{1000}$$

```
> table(d)
d
  1   2   3   4   5   6
191 155 159 181 151 163
```



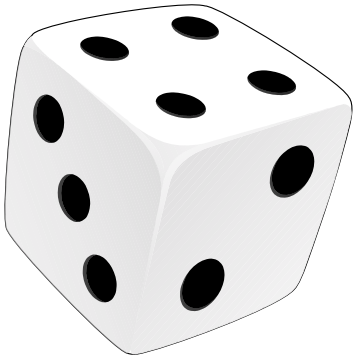
架空データ

1000回サイコロふった

1000/6 = 166.66...?

「サイコロの統計モデル」にできること

乱数発生

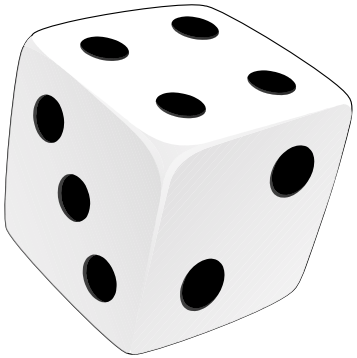


```
> v.prob <- table(d) / 1000
> replicate(8, table(sample(1:6, 1000,
+ replace = TRUE, prob = v.prob)))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
1	185	201	174	202	179	193	198	197
2	156	159	158	164	160	141	154	151
3	173	161	195	180	177	153	147	173
4	176	171	173	177	189	174	200	190
5	147	157	152	139	138	171	122	130
6	163	151	148	138	157	168	179	159

「サイコロの統計モデル」にできること

予測

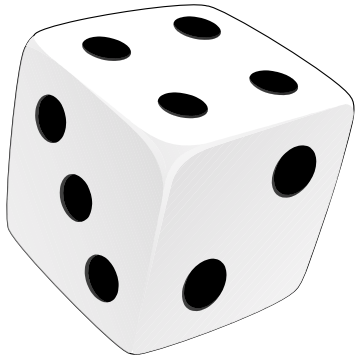


```
> # サイコロ 1000 回ふりを 1000 回やる
> sim1000 <- replicate(1000,
+ table(sample(1:6, 1000, replace = TRUE,
+ prob = v.prob)))
> # 5 よりも 6 が多くでる回数は?
> sum(sim1000[5,] > sim1000[6,])
[1] 247
> # 3 よりも 4 が多くでる回数は?
> sum(sim1000[3,] > sim1000[4,])
[1] 111
> # 5 より 6 が多いときに , 3 よりも 4 が多くでる回数は?
> sum((sim1000[5,] > sim1000[6,])
+ * (sim1000[3,] > sim1000[4,]))
[1] 35
```

「サイコロの統計モデル」にできること

モデル選択や検定

「次」の
データ

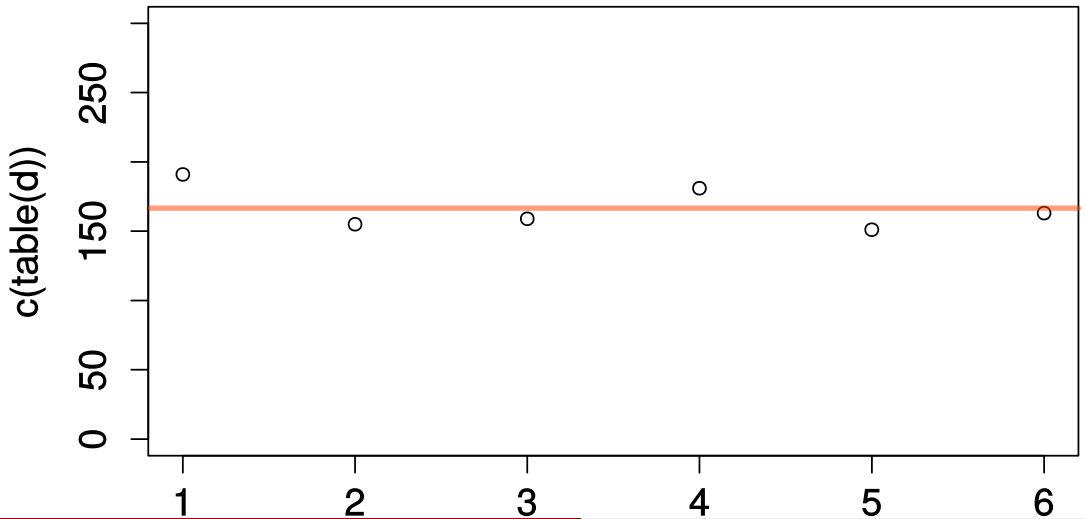


モデル選択

モデル1: p_k がすべて等しい

モデル2: p_k がすべて異なる

「予測力」の高い
モデルを選ぶ



1000回サイコロふった

1000/6 = 166.66...?

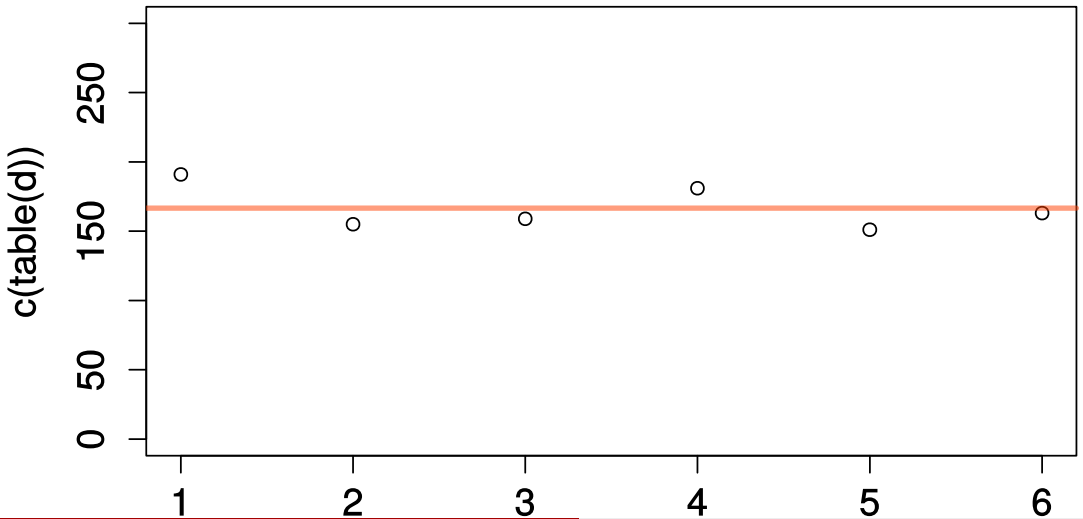
「サイコロの統計モデル」にできること

モデル選択や検定

統計学的な検定

モデル1: p_k がすべて等しい

モデル2: p_k がすべて異なる



モデル1 を安全に棄却できる確率 p だけを評価する

$p > 0.05$ なら…?

「何も言えない」と結論するのが正しい

“確率分布” を部品にもつ統計モデル

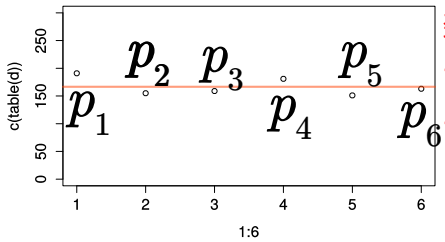
できること: 推定・乱数発生・予測・モデル選択 ……

「サイコロの統計モデル」にできること

パラメーターの推定

$$\hat{p}_k = \frac{k \text{ の目が出た回数}}{1000}$$

```
> table(d)
d
  1   2   3   4   5   6
191 155 159 181 151 163
```



架空データ

1000回サイコロふった

1000/6 = 166.66...?

「サイコロの統計モデル」にできること

乱数発生

```
> v.prob <- table(d) / 1000
> replicate(8, table(sample(1:6, 1000,
+ replace = TRUE, prob = v.prob)))
```

```
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
1  185  201  174  202  179  193  198  197
2  156  159  158  164  160  141  154  151
3  173  161  195  180  177  153  147  173
4  176  171  173  177  189  174  200  190
5  147  157  152  139  138  171  122  130
6  163  151  148  138  157  168  179  159
```

「サイコロの統計モデル」にできること

予測

```
> # サイコロ 1000 回ふりを 1000 回やる
> sim1000 <- replicate(1000,
+ table(sample(1:6, 1000, replace = TRUE,
+ prob = v.prob)))
> # 5 よりも 6 が多くでる回数は?
> sum(sim1000[5,] > sim1000[6,])
[1] 247
> # 3 よりも 4 が多くでる回数は?
> sum(sim1000[3,] > sim1000[4,])
[1] 111
> # 5 より 6 が多いときに 3 よりも 4 が多くでる回数は?
```

「サイコロの統計モデル」にできること

モデル選択や検定

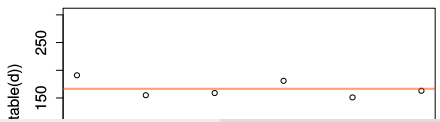
「次」のデータ

モデル選択

モデル1: p_k がすべて等しい

モデル2: p_k がすべて異なる

「予測力」の高いモデルを選ぶ



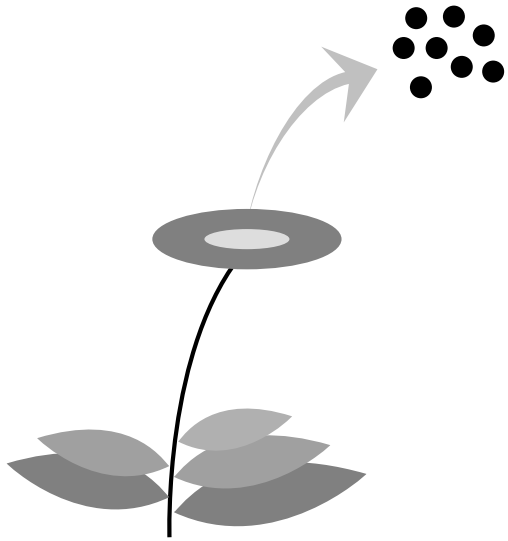
1000回サイコロふった

2. 例題: 種子数の統計モデリング

まあ，かなり単純な例から始めましょう

R でデータをあつかいつつ

この授業では架空植物の架空データをあつかう

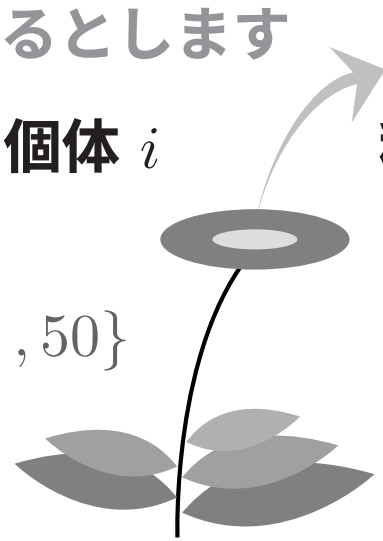


理由: よけいなことは考えなくてすむので

現実のデータはどれも授業で使うには難しすぎる……

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



個体 i

種子数 y_i

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
```

統計ソフトウェア R

統計学の勉強には良い統計ソフトウェアが必要!

- 無料で入手できる
- 内容が完全に公開されている
- 多くの研究者が使っている
- 作図機能が強力

この教科書でも R を
使って問題を解決する
方法を説明しています

R でデータの様子をながめる

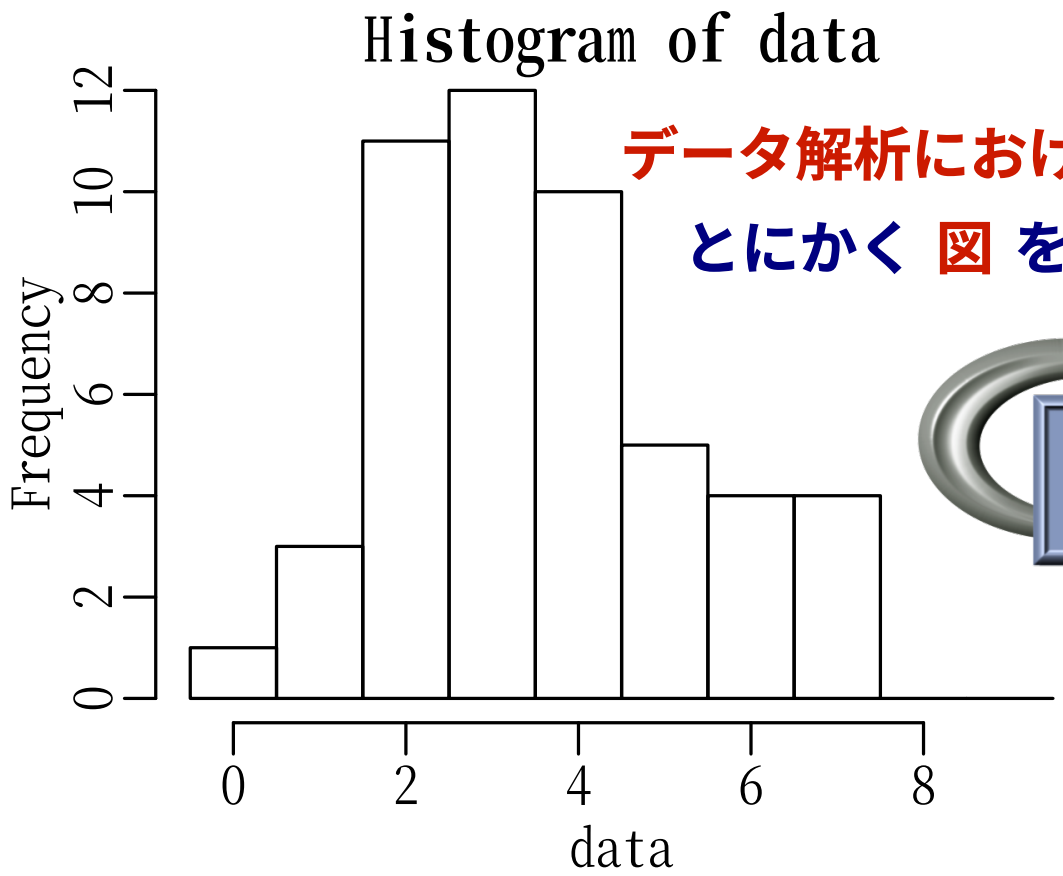
の table() 関数を使って種子数の頻度を調べる

```
> 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 ……)

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))
```

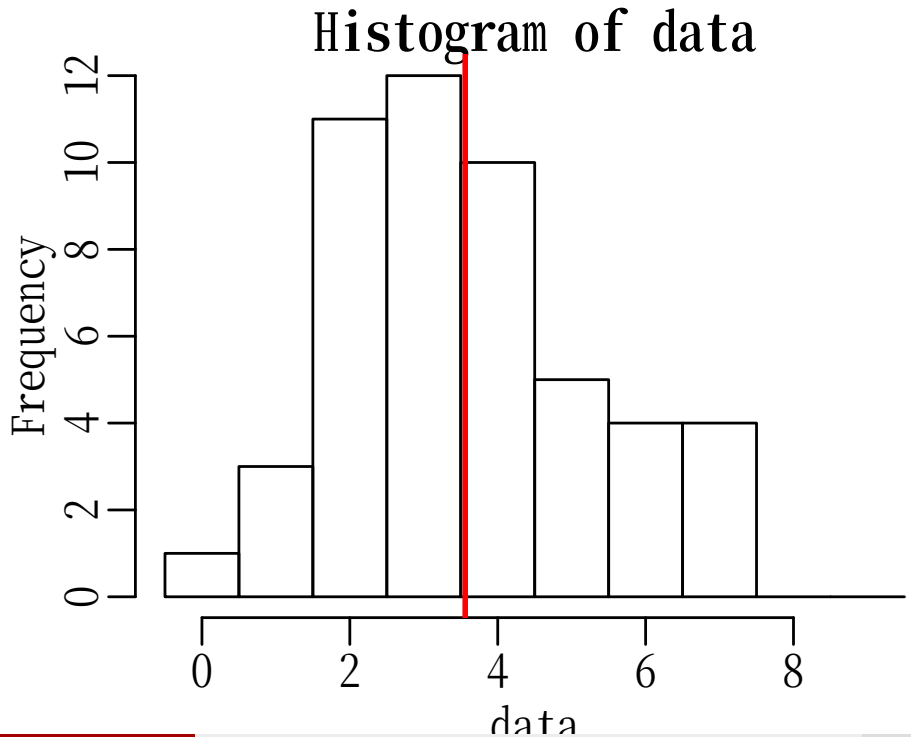


データ解析における最重要事項

とにかく 図 を描く!

標本平均という統計量

```
> mean(data)
[1] 3.56
> abline(v = mean(data), col = "red")
```



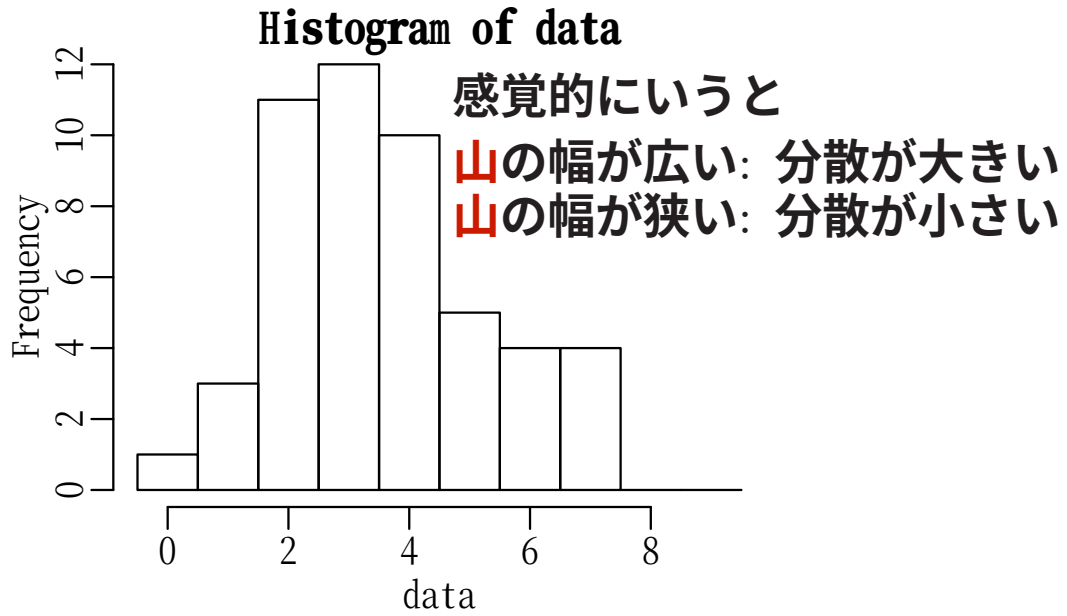
ばらつきの統計量

あるデータの ばらつき をあらわす標本統計量の例: 標本分散

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

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
```



3. 確率分布って何?

経験分布と理論分布

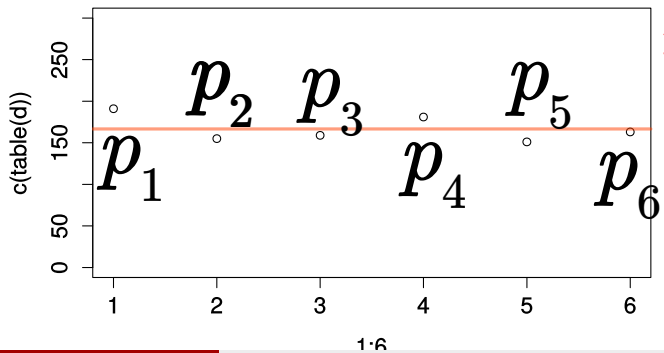
統計モデルの部品である 確率分布 には “データそのまま” な 経験分布 (cf. サイコロ) と 数式で定義される理論的な分布 がある

「サイコロの統計モデル」にできること

パラメーターの推定

$$\hat{p}_k = \frac{k \text{ の目が出た回数}}{1000}$$

```
> table(d)
d
  1   2   3   4   5   6
191 155 159 181 151 163
```



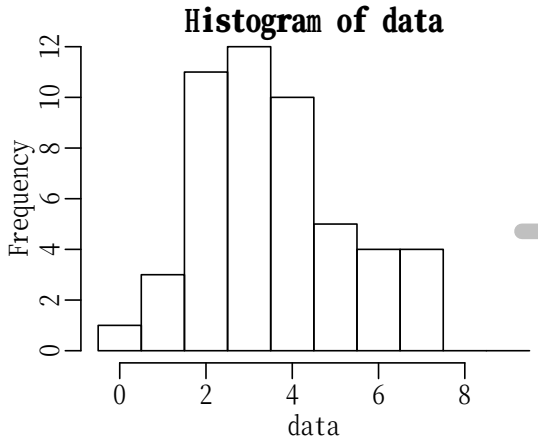
架空データ

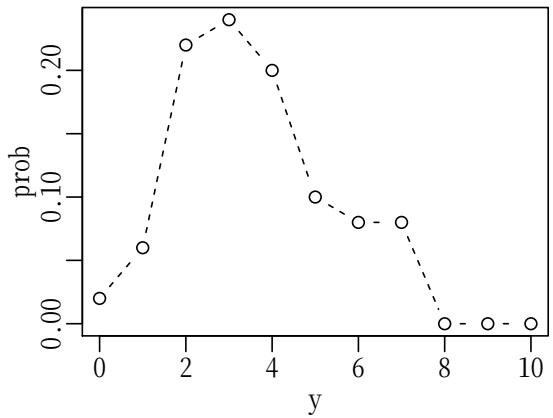
1000回サイコロふった

1000/6 = 166.66...?

“データそのまま” な経験分布

```
> 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
- “たまたま手もとにある” データから “発生確率” を決める確率分布が**経験分布**

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)

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

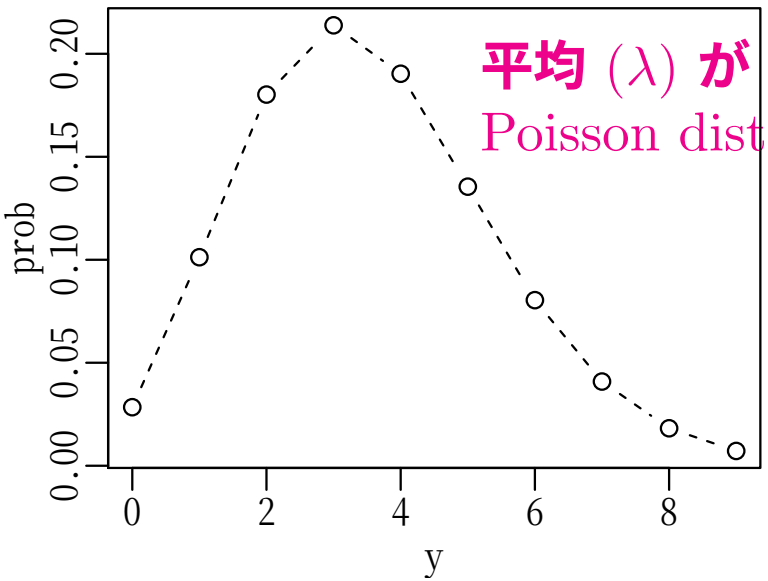
$$p(y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

- **$y!$ 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**

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

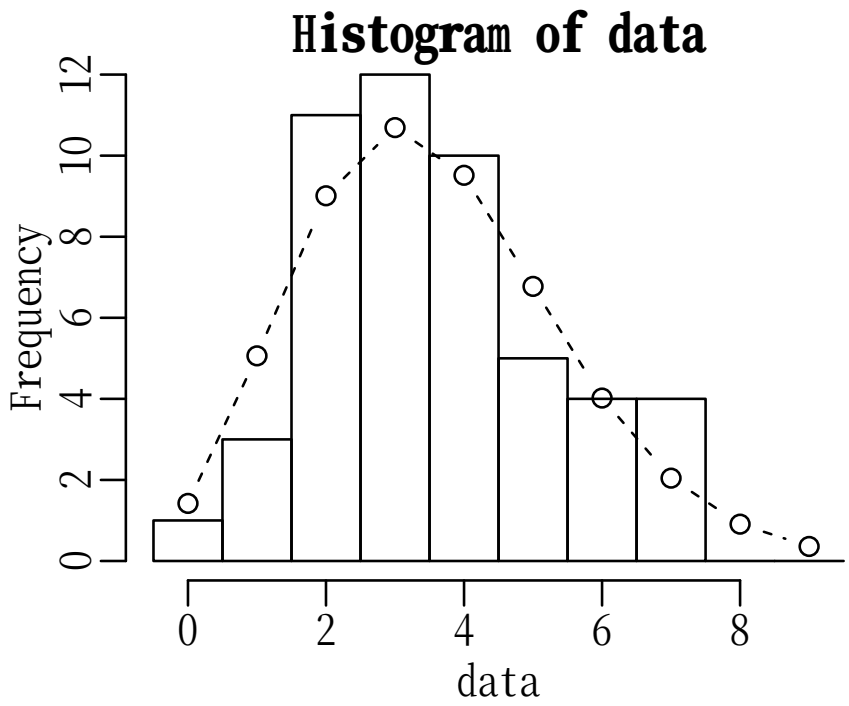
```
> y <- 0:9                                  # これは種子数（確率変数）
> prob <- dpois(y, lambda = 3.56)           # ポアソン分布の確率の計算
> plot(y, prob, type = "b", lty = 2)
```



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

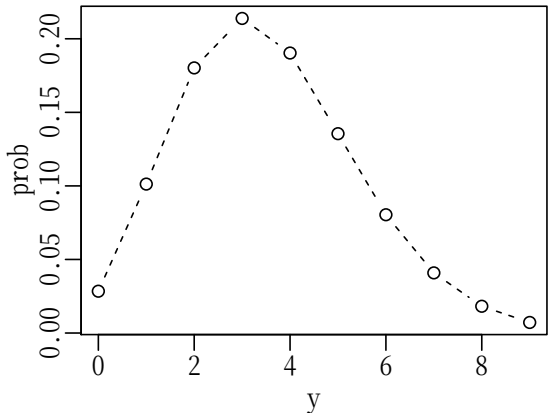
```
> # cbind で「表」作り
> cbind(y, prob)
      y       prob
 1    0 0.02843882
 2    1 0.10124222
 3    2 0.18021114
 4    3 0.21385056
 5    4 0.19032700
 6    5 0.13551282
 7    6 0.08040427
 8    7 0.04089132
 9    8 0.01819664
10    9 0.00719778
```

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
```

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

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

どういう場合にポアソン分布を使う?

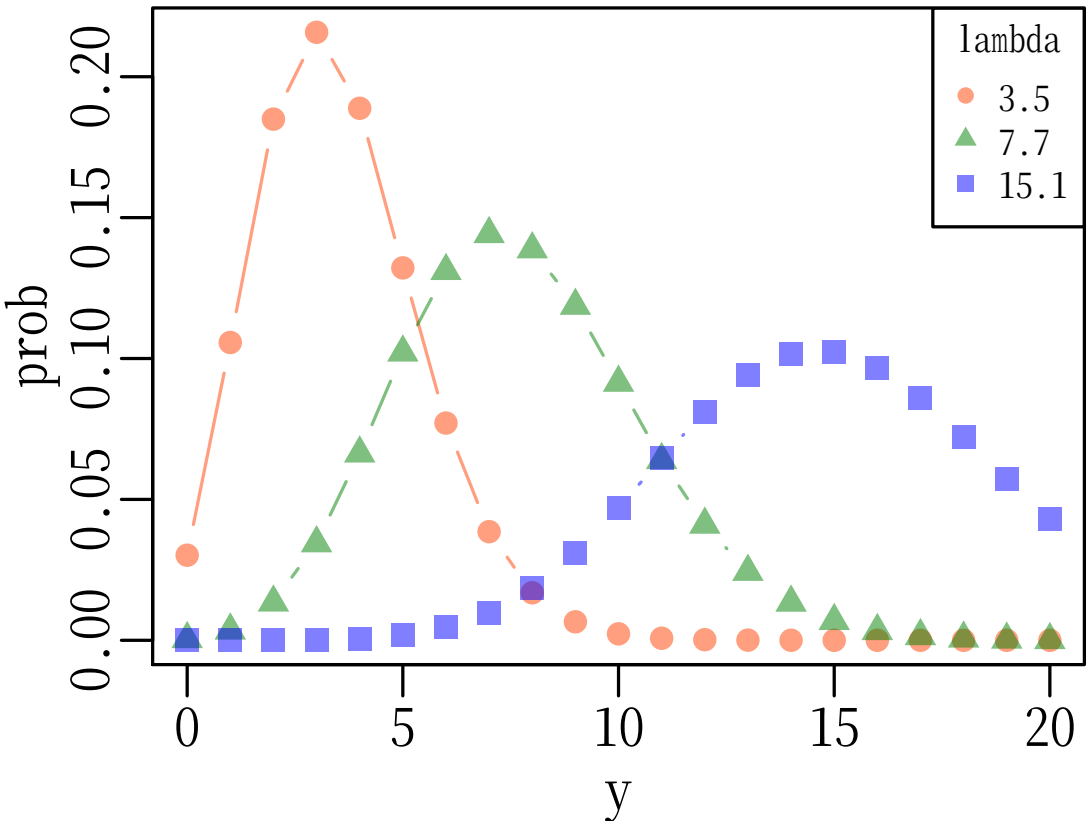
統計モデルの部品としてポアソン分布が選んだ理由:

- **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 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**
  - **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**

ポアソン分布の λ を変えてみる

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

λ は平均をあらわすパラメーター



4. ポアソン分布のパラメーターの最尤推定

もっともらしい推定?

尤度 (likelihood) とは何か?

- 最尤推定法では，尤度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」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

尤度 $L(\lambda)$ はパラメーター λ の関数

この例題の尤度:

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

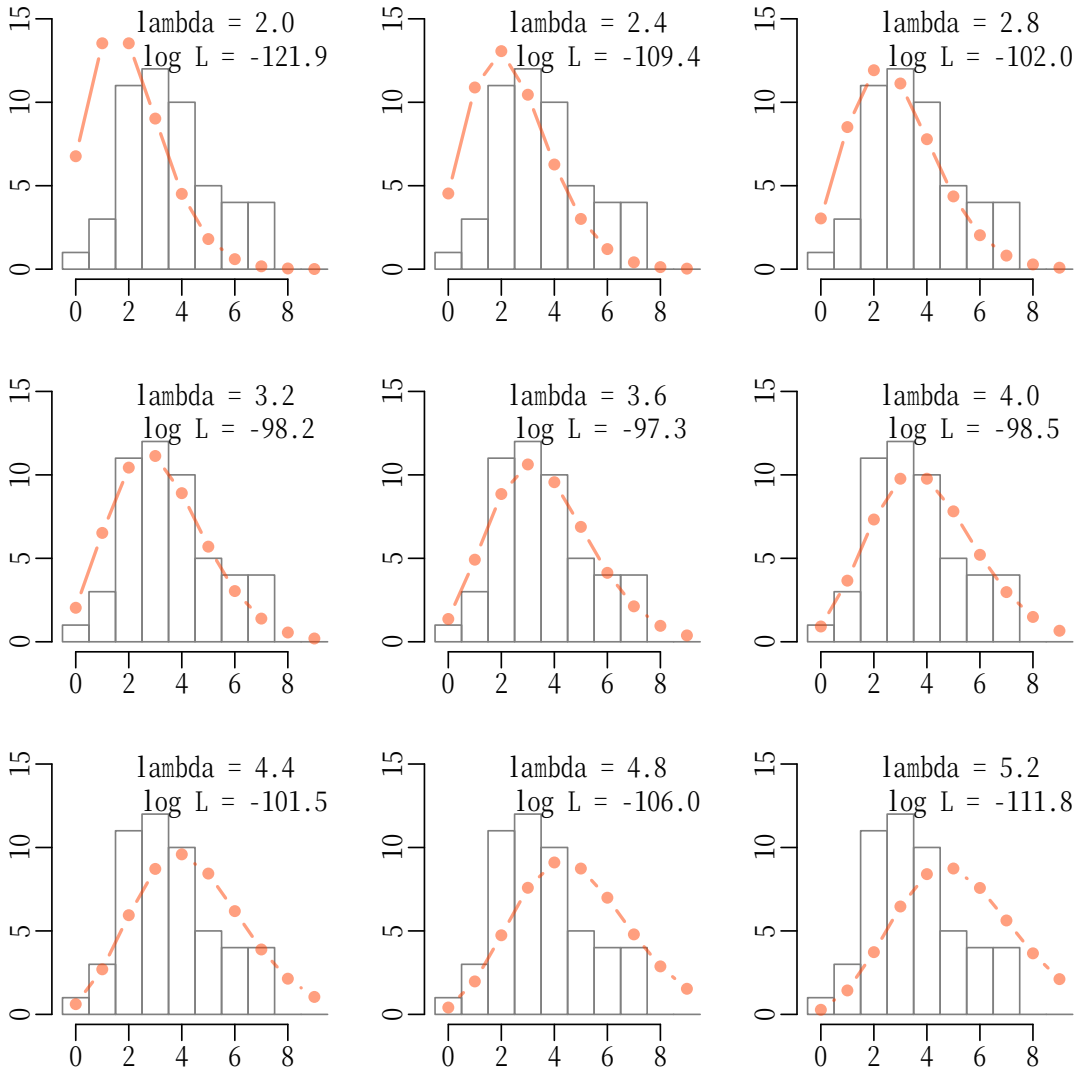
尤度はしんどいので対数尤度を使う

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

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

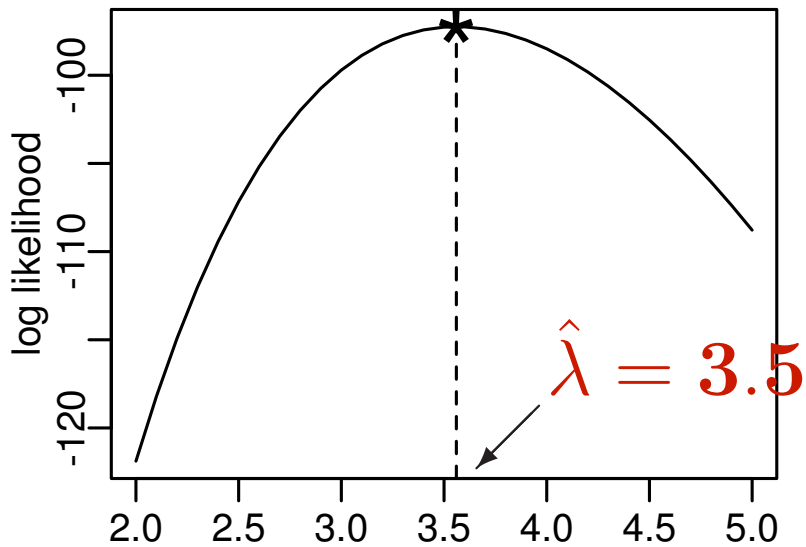
対数尤度 $\log L(\lambda)$ の最大化は尤度 $L(\lambda)$ の最大化になるから
まずは，平均をあらわすパラメーター λ を変化させていったときに，ポアソン分布のカタチと対数尤度がどのように変化するのかを調べてみましょう

λ を変えるとあてはまりの良さが変わる



対数尤度を最大化する $\hat{\lambda}$ をさがす

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

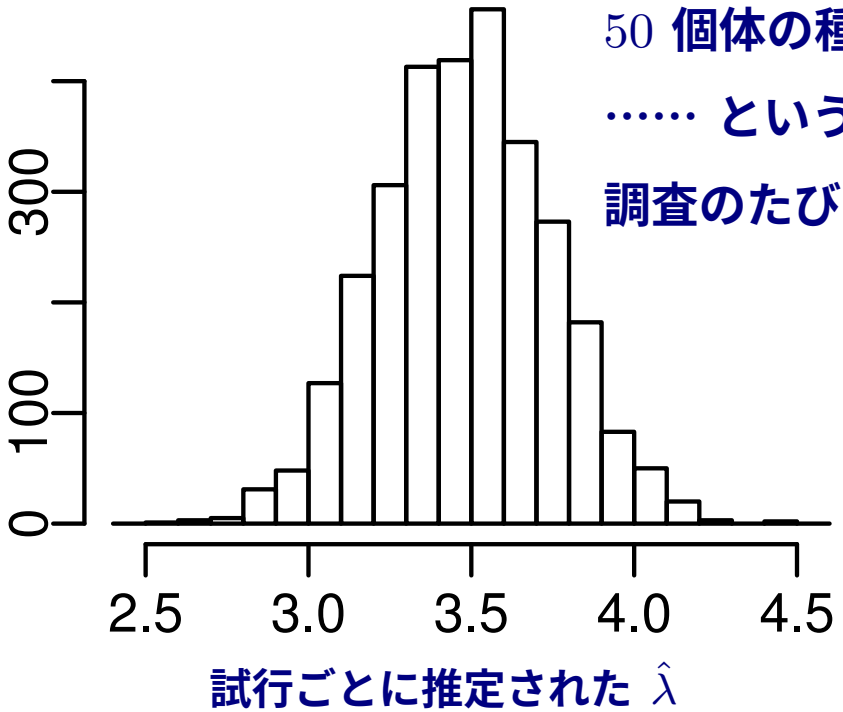


$\frac{d \log L}{d\lambda} = 0$ より

- **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**

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 の場合



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

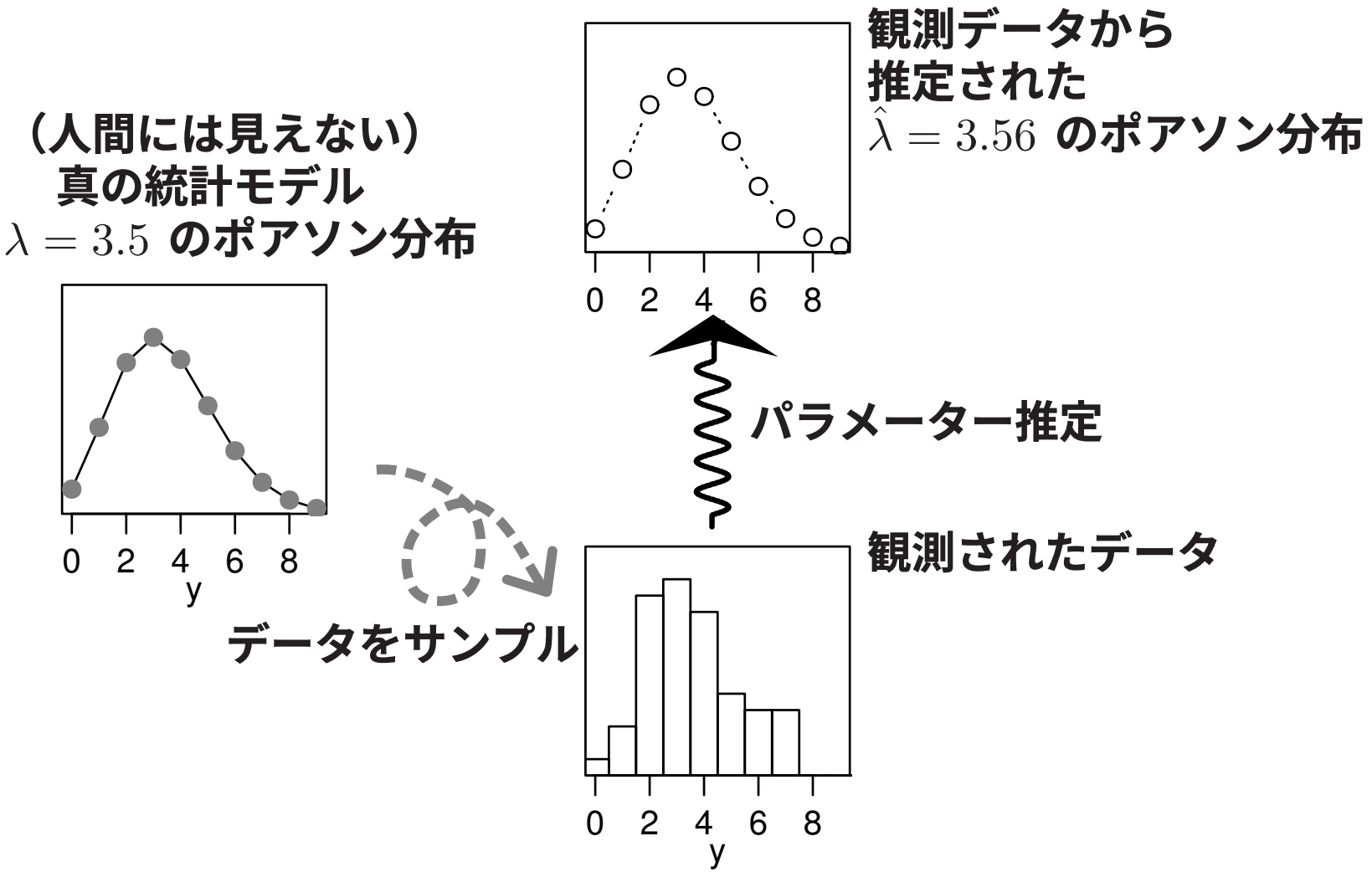
データは有限なので真の λ はわからない

5. 統計モデルの要点

乱数発生・推定・予測

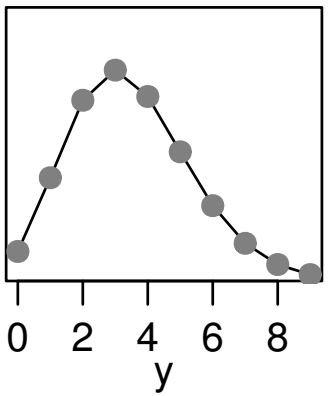
統計モデルとデータの対応づけ

統計学における推定


観測データから
推定された
$\hat{\lambda} = 3.56$ のポアソン分布
0 2 4 6 8
パラメーター推定
（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
0 2 4 6 8
y
データをサンプル
観測されたデータ
0 2 4 6 8
y

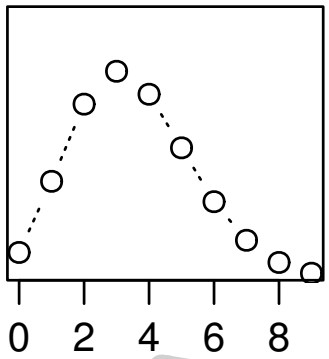
統計学における予測

（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

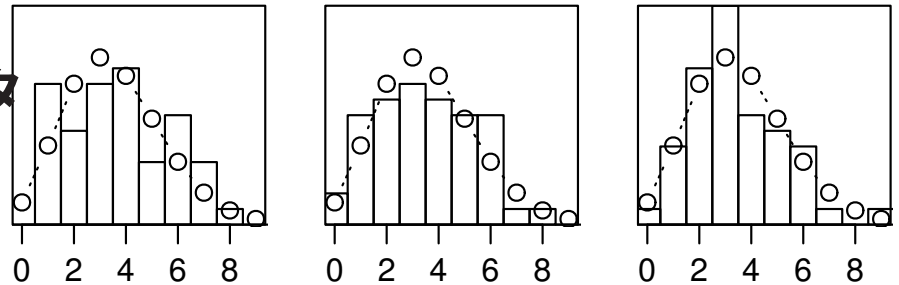


新しいデータ
をサンプル

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？



・・・

同じ調査方法で得られた新データ

この授業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……

- この集中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