

統計モデリング入門 2015 (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

北大環境科学院の講義 <http://goo.gl/76c4i>

2015-07-08

ファイル更新時刻: 2015-07-08 17:27

agenda

今日のハナシ I

① 例題: サイコロ (dice) の統計モデリング

Probability distributions as parts of model

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

③ データと確率分布の対応

probability distribution, the core of statistical model

maximum likelihood estimation of parameter λ

④ ポアソン分布のパラメーターの さいゆうすいてい 最尤推定

もっとももっともらしい推定?

the functions of statistical model

⑤ 統計モデルの要点

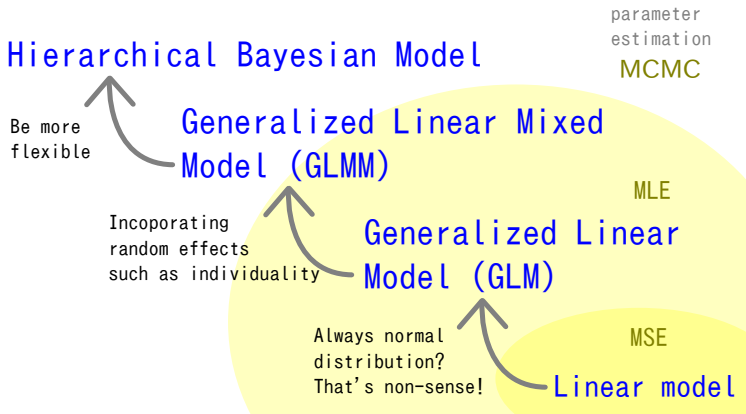
random number, estimation and prediction

乱数発生・推定・予測

本題にはいる前に
統計モデリング授業前半の
「テーマ」を
再確認しておきましょう

statistical models appeared in the class
この授業であつかう統計モデルたち

The development of linear models

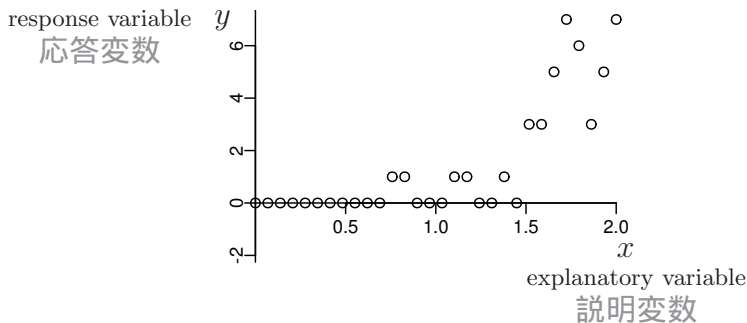


Kubo Doctrine: “Learn the evolution of linear-model family, firstly!”

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 ($y \in \{0, 1, 2, 3, \dots\}$ なデータ)

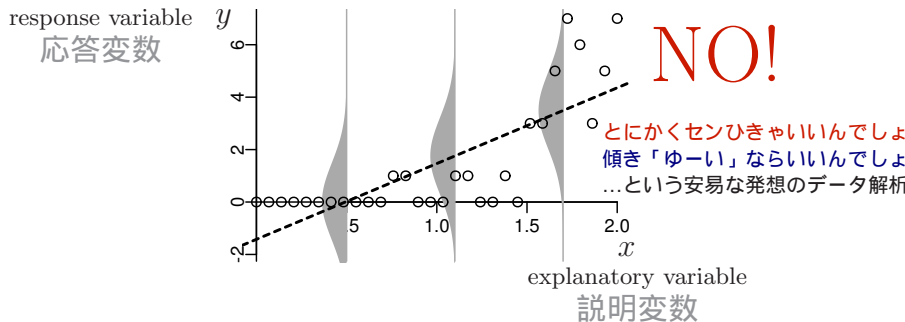


- たとえば x は植物個体の大きさ, y はその個体の花数
- 体サイズが大きくなると花数が増えるように見えるが.....
- この現象を表現する統計モデルは?

the normal distribution sucks!

正規分布を使った統計モデル ムリがある？

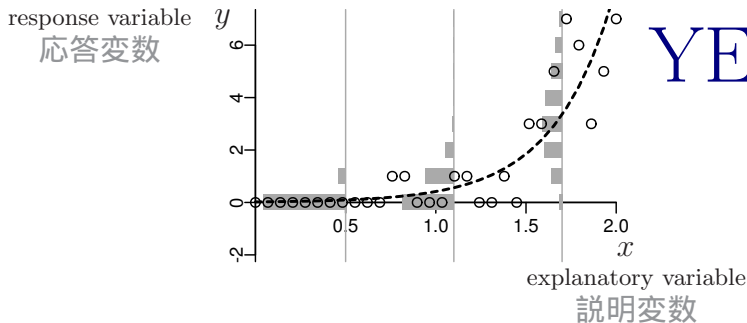
正規分布・恒等リンク関数の統計モデル



- 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 distribution 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



- 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

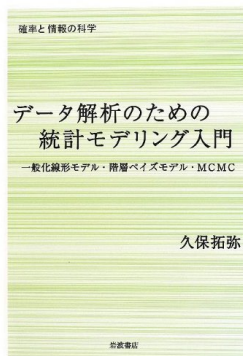
データの性質をよくみる
確率分布という**部品**を選ぶ
「ぶらっくぼっくす」にしない!

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



3. 例題: サイコロ (dice) の統計モデリング

Probability distributions as parts of model

R でデータをあつかいつつ

これ使いましょう: 統計ソフトウェア R

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

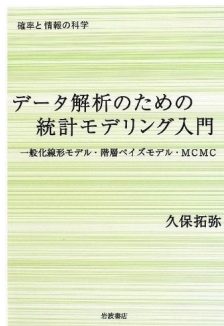


- いろいろな OS で使える **free software**
- 使いたい機能が充実している
- **作図**機能も強力
- S 言語による **プログラミング**可能
- **Rstudio** <http://www.rstudio.com/>

「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)
```

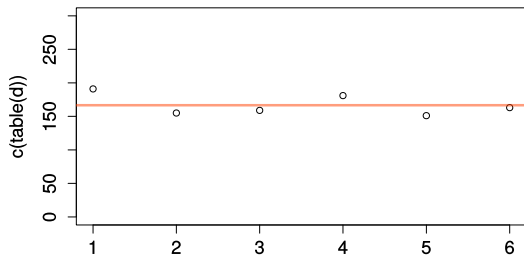
```
d
```

```
 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

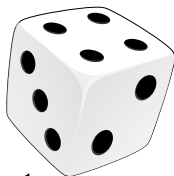
1000回サイコロふった

$1000/6 = 166.66\dots?$

「サイコロ」の確率分布は?

Categorical Distribution

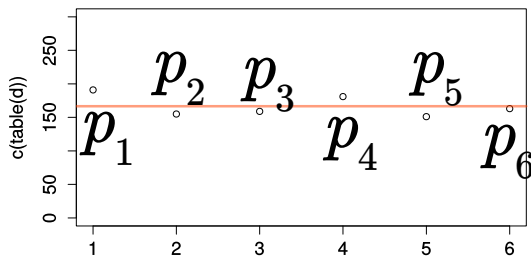
カテゴリカル分布



$$p(Y = k) = p_k$$

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

$$\sum_{k=1}^6 p_k = 1$$



架空データ

1000回サイコロふった

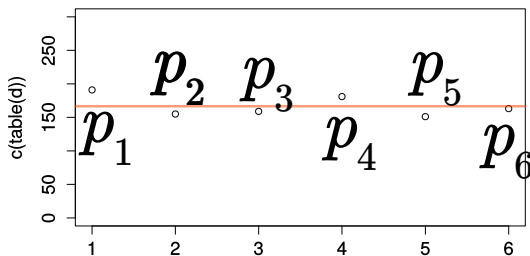
1000/6 = 166.66...?

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

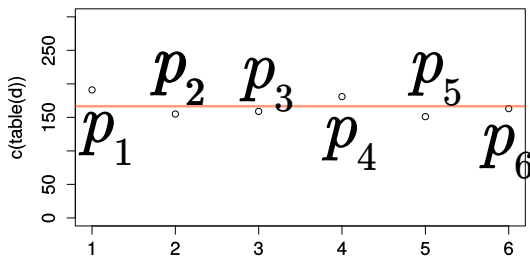
$1000/6 = 166.66\dots?$

「サイコロの統計モデル」にできること

パラメータの推定

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

```
> 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

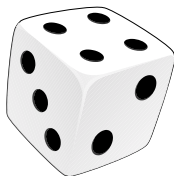
$1000/6 = 166.66\dots?$

「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)))
```

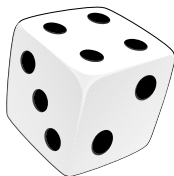
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
1	185	201	174	202	179	193	198	197
2	156	159	158	164	160	141	154	151
3	173	161	195	180	177	153	147	173
4	176	171	173	177	189	174	200	190
5	147	157	152	139	138	171	122	130
6	163	151	148	138	157	168	179	159



「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 が多く出る回数は?
> sum(sim1000[5,] > sim1000[6,])
[1] 247
> # 3 よりも 4 が多く出る回数は?
> sum(sim1000[3,] > sim1000[4,])
[1] 111
> # 5 より 6 が多いときに, 3 よりも 4 が多く出る回数は?
> sum((sim1000[5,] > sim1000[6,])
+ * (sim1000[3,] > sim1000[4,]))
[1] 35
```



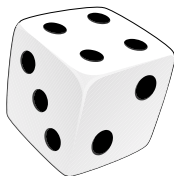
「サイコロの統計モデル」にできること モデル選択や検定

モデル選択

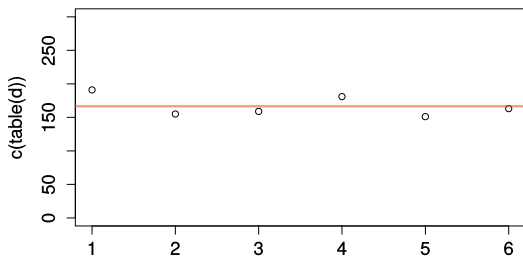
モデル1: p_k がすべて等しい

モデル2: p_k がすべて異なる

「次」の
データ



「予測力」の高い
モデルを選ぶ



1000回サイコロふった

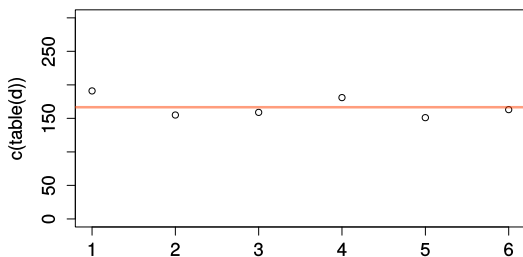
$1000/6 = 166.66\dots?$

「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

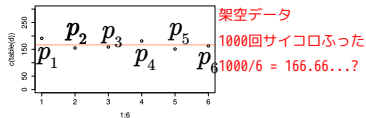
“確率分布” を部品にもつ統計モデル

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
```



「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 が多く出る回数は?
> sum(sim1000[,5] > sim1000[,6])
[1] 247
> # 3 よりも 4 が多く出る回数は?
> sum(sim1000[,3] > sim1000[,4])
[1] 111
> # 5 よりも 6 が多く出る回数に 3 よりも 4 が多く出る回数は?
```

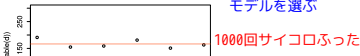


「サイコロの統計モデル」にできること
モデル選択や検定

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

「次」のデータ

「予測力」の高いモデルを選ぶ



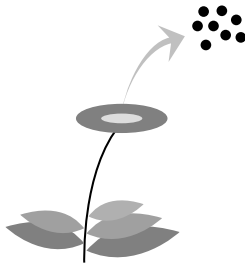
3. 例題: 種子数の統計モデリング

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

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



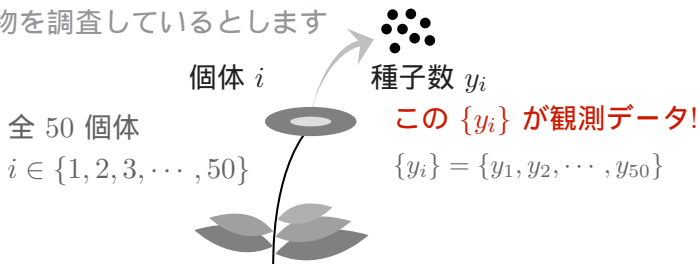
理由: よけいなことは考えなくてすむので

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



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
```


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)
```

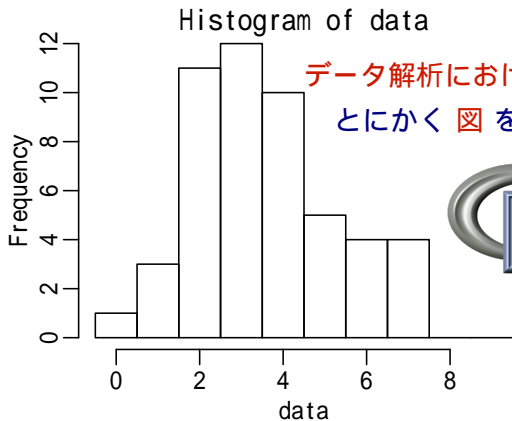
```
0  1  2  3  4  5  6  7
1  3 11 12 10  5  4  4
```


(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))
```



データ解析における最重要事項
とにかく  を描く!

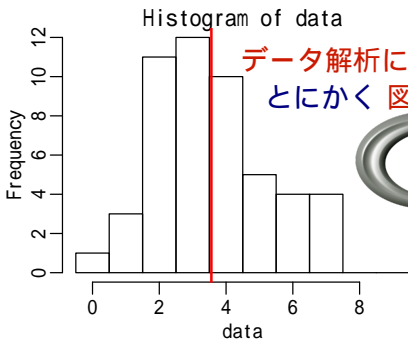


How to evaluate mean value using R?

```
> 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  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: **標本分散** sample variance

```
> var(data)
```

```
[1] 2.9861
```

sample standard deviation

標本標準偏差

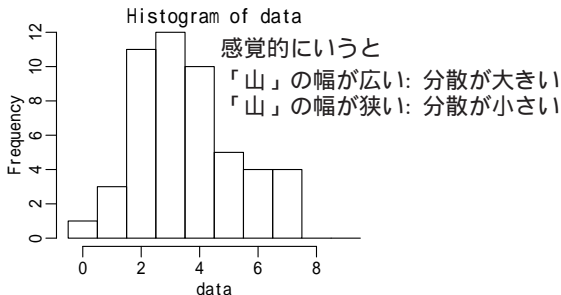
とは標本分散の平方根 ($SD = \sqrt{\text{variance}}$)

```
> sd(data)
```

```
[1] 1.7280
```

```
> sqrt(var(data))
```

```
[1] 1.7280
```



4. 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

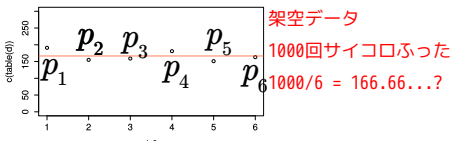
Empirical VS Theoretical Distribution

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

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

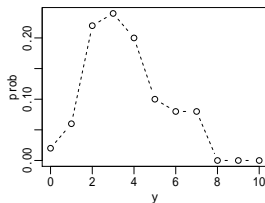
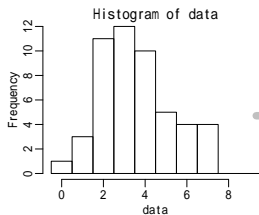
```
> table(d)
d
 1   2   3   4   5   6
191 155 159 181 151 163
```



empirical distribution

“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	0.02
1	0.06
2	0.22
3	0.24
4	0.20
5	0.10
6	0.08
7	0.08
8	0.00
9	0.00
10	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 $y = \{0, 1, 2, \dots\}$ 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
 $\{p_0, p_1, p_2, \dots, p_{99}, p_{100}, \dots\}$ 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

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

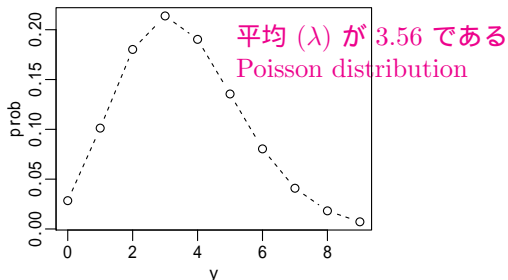
- $y!$ は y の ^{factorial} 階乗 で, たとえば $4!$ は $1 \times 2 \times 3 \times 4$ をあらわしています.
- $\exp(-\lambda) = e^{-\lambda}$ のこと ($e = 2.718 \dots$)
- ここではなぜポアソン分布の確率計算が上のようになるのかは説明しません— まあ, こういうもんだと考えて先に進みましょう

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)
```

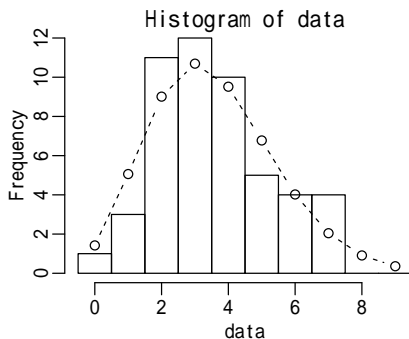


```
> # cbind で「表」作り
> cbind(y, prob)
```

	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

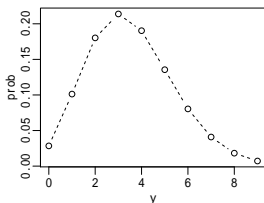
the Poisson distribution represent data?

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
```

# パラメーター $\lambda$ 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  $\lambda$  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  $\lambda$  である ( $\lambda \geq 0$ )
- 分散と平均は等しい:  $\lambda = \text{平均} = \text{分散}$
- $y \in \{0, 1, 2, \dots, \infty\}$  の値をとり, すべての  $y$  について和をとると 1 になる

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

## どういった場合にポアソン分布を使う？

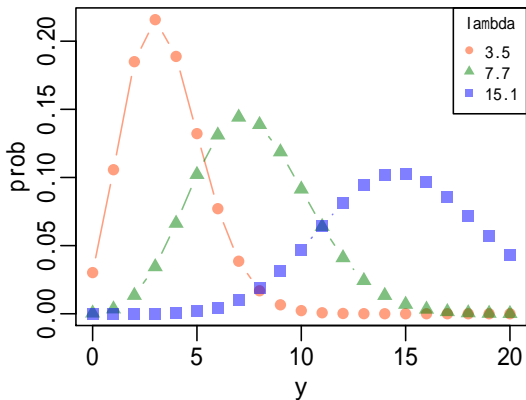
統計モデルの部品としてポアソン分布が選んだ理由:

- データに含まれている値  $y_i$  が  $\{0, 1, 2, \dots\}$  といった非負の整数である (カウントデータである)  
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  
mean  $\approx$  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

$\lambda$  changes the shape of distributionポアソン分布の  $\lambda$  を変えてみる

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

$\lambda$  は平均 <sup>mean</sup> をあらわすパラメーター



maximum likelihood estimation of parameter  $\lambda$

さいゆうすいてい

## 5. ポアソン分布のパラメーターの最尤推定

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



ゆうど

# 尤度 (likelihood) とは何か?

- maximum likelihood estimation  
最尤推定法 では、<sup>ゆうど</sup>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  $\lambda$  を変えると尤度が変わる
- goodness of fit  
もっとも「あてはまり」が良くなる  $\lambda$  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,  $\lambda$   
尤度  $L(\lambda)$  はパラメーター  $\lambda$  の関数

この例題の尤度:

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

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!)

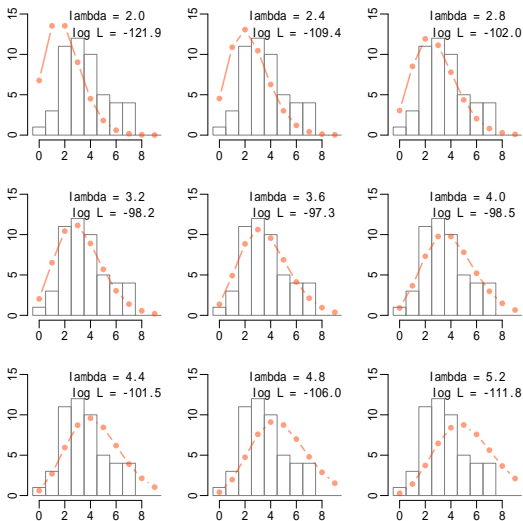
そこで, パラメーターの最尤推定では, 対数尤度関数 (log likelihood function) を使う

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

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

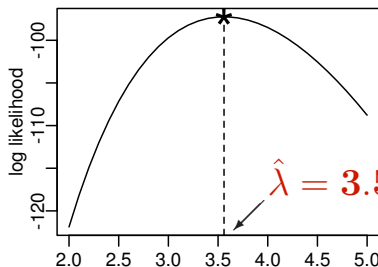
まずは, 平均をあらわすパラメーター  $\lambda$  を変化させていったときに, ポアソン分布のカタチと対数尤度がどのように変化するかを調べてみましょう

# $\lambda$ changes the log likelihood, i.e., goodness of fit $\lambda$ 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

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

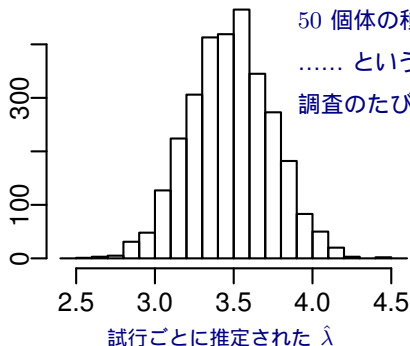


$$\frac{d \log L}{d \lambda} = 0 \text{ より}$$

- 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  $\lambda$ ” based on finite size data  
最尤推定を使っても**真の  $\lambda$** は見つからない

**真の  $\lambda$** が 3.5 の場合



データは有限なので**真の  $\lambda$** はわからない

the functions of statistical model

## 6. 統計モデルの要点

random number, estimation and prediction

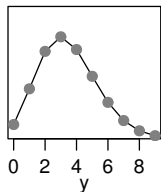
乱数発生・推定・予測

統計モデルとデータの対応づけ

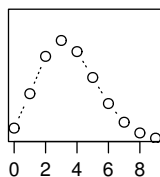
estimation

## 統計学における 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



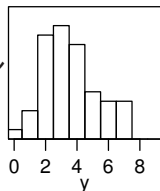
データをサンプル



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



パラメーター推定



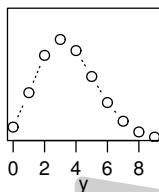
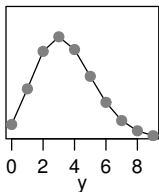
観測されたデータ



prediction

## 統計学における 予測

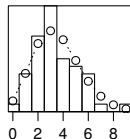
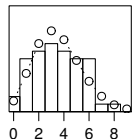
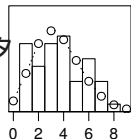
(人間には見えない)  
真の統計モデル  
 $\lambda = 3.5$  のポアソン分布



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?

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

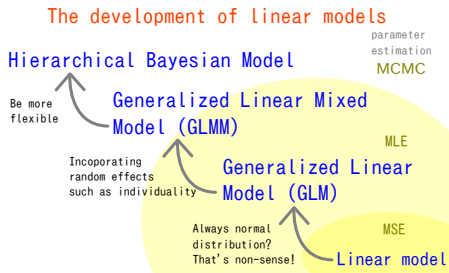
## この授業で登場する確率分布

- **ポアソン分布:**  $y \in \{0, 1, 2, 3, \dots\}$  となるデータ, 「 $y$  回なにかがおこった」
- **二項分布:**  $y \in \{0, 1, 2, \dots, N\}$  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

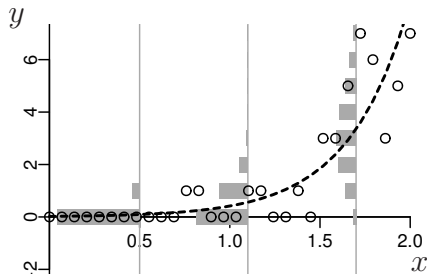
# いろいろな確率分布があるけれど.....

- 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



# 次回予告

The next topic



## 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)