

茨城大集中講義 2016 (b)

確率分布と最尤推定

久保拓弥 kubo@ees.hokudai.ac.jp

筑波大の講義 <http://goo.gl/aFLLHZ>

2016-10-06

ファイル更新時刻: 2016-09-29 13:32

今日のハナシ I

① 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

② データと確率分布の対応

probability distribution, the core of statistical model

③ 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?

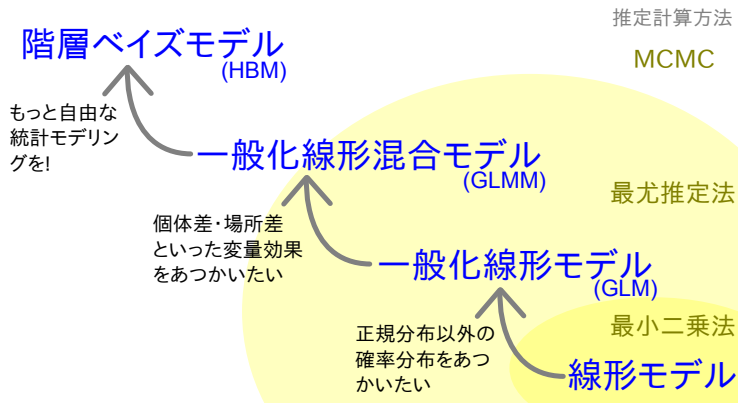
④ 統計モデルの要点

乱数発生・推定・予測

統計モデリング授業前半の
主題は
「線形モデルを発展させる」
こと

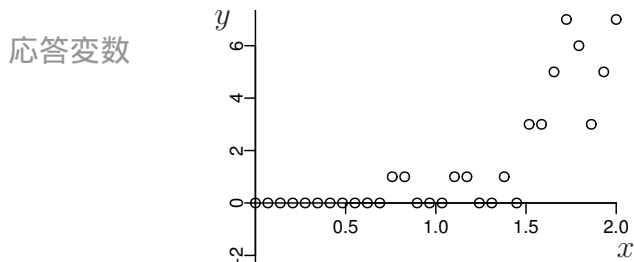
この授業であつかう統計モデルたち

線形モデルの発展



データの特徴にあわせて線形モデルを改良・発展させる

0 個, 1 個, 2 個と数えられるデータ

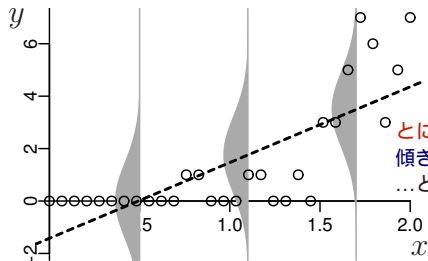
カウントデータ ($y \in \{0, 1, 2, 3, \dots\}$ なデータ)

- 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?

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

応答変数



NO!

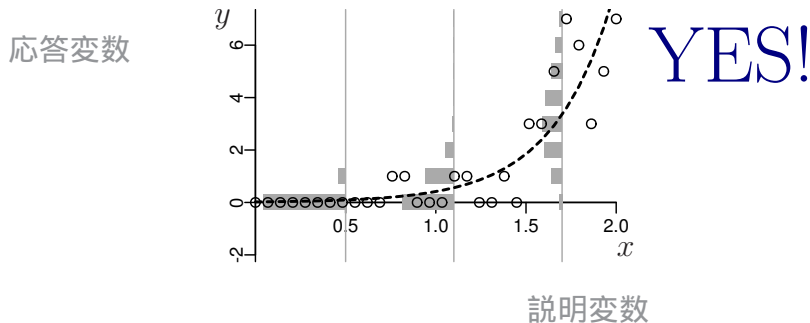
とにかくセンひきゃいいんでしょ
傾き「ゆーい」ならいいんでしょ
...という安易な発想のデータ解析

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？

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

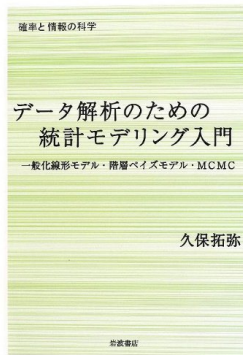
データの性質をよくみる
確率分布という**部品**を選ぶ
「ぶらっくぼっくす」にしない!

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

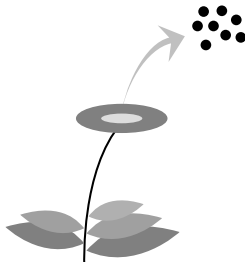


2. 例題: 種子数の統計モデリング

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

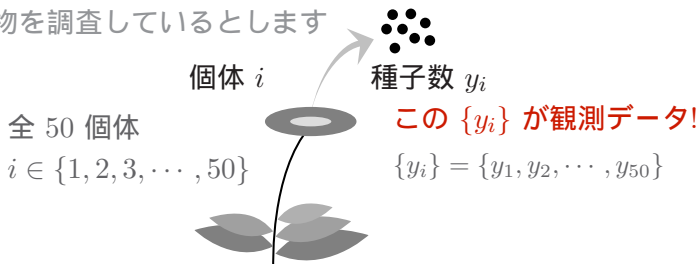


理由: よけいなことは考えなくてすむので

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



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
```

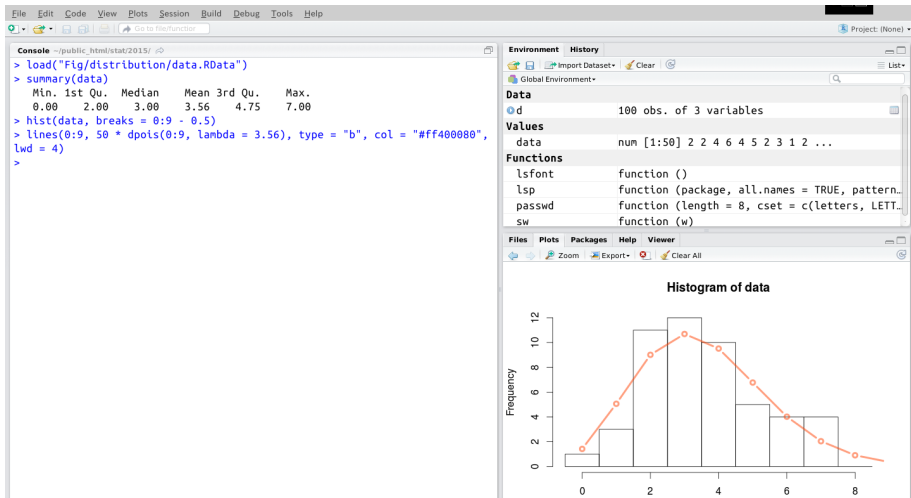
これ使いましょう: 統計ソフトウェア R

<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/>

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)
```

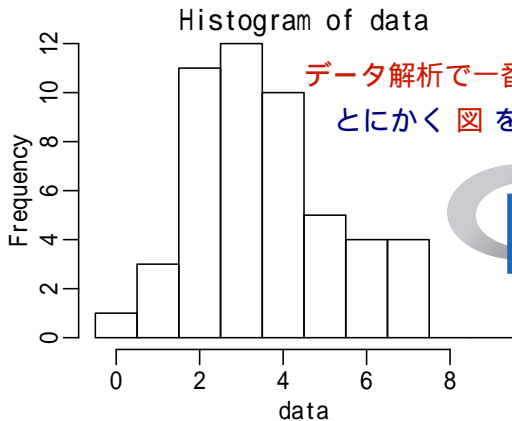
```
0  1  2  3  4  5  6  7
```

```
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))
```



データ解析で一番たいせつなこと
とにかく  を描く!

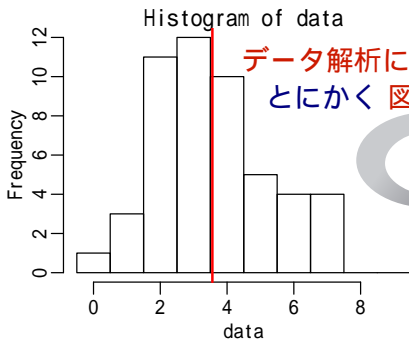


How to evaluate mean value using R?

```
> 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!



「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
```

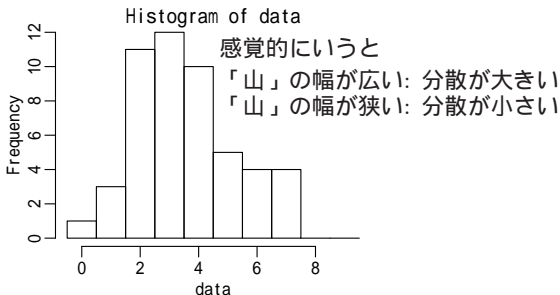
標本標準偏差 とは標本分散の平方根 ($SD = \sqrt{\text{variance}}$)

```
> sd(data)
```

```
[1] 1.7280
```

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

```
[1] 1.7280
```



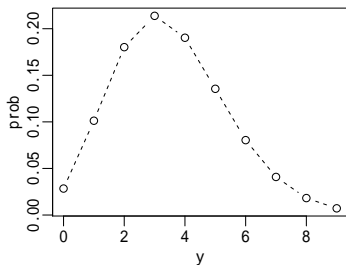
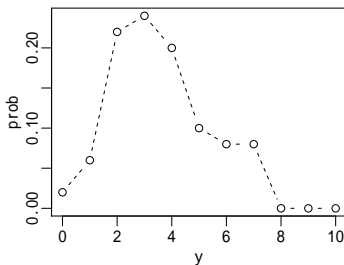
3. データと確率分布の対応

probability distribution, the core of statistical model

確率分布は統計モデルの重要な部品

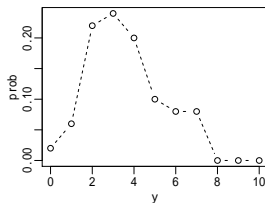
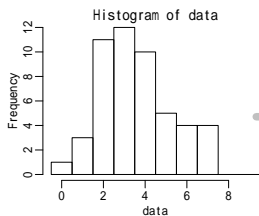
Empirical VS Theoretical Distributions

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** (cf. サイコロ) 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

```
> 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)

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

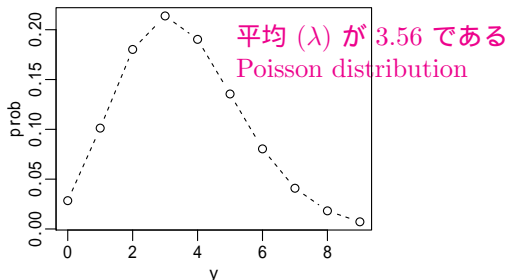
$$p(y | \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

- $y!$ 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 ($e = 2.718 \dots$)
- 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

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

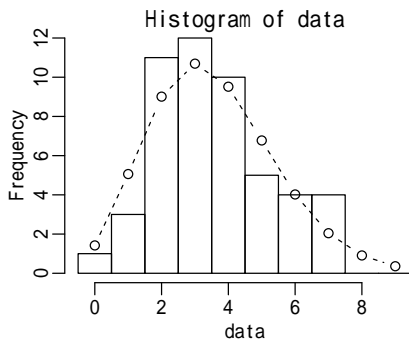
```
> y <- 0:9 # これは種子数 (確率変数)
> prob <- dpois(y, lambda = 3.56) # ポアソン分布の確率の計算
> plot(y, prob, type = "b", lty = 2)
```



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

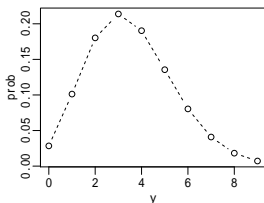
	y	prob
1	0	0.02843882
2	1	0.10124222
3	2	0.18021114
4	3	0.21385056
5	4	0.19032700
6	5	0.13551282
7	6	0.08040427
8	7	0.04089132
9	8	0.01819664
10	9	0.00719778

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: $\lambda = \text{平均} = \text{分散}$
- $y \in \{0, 1, 2, \dots, \infty\}$ 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？

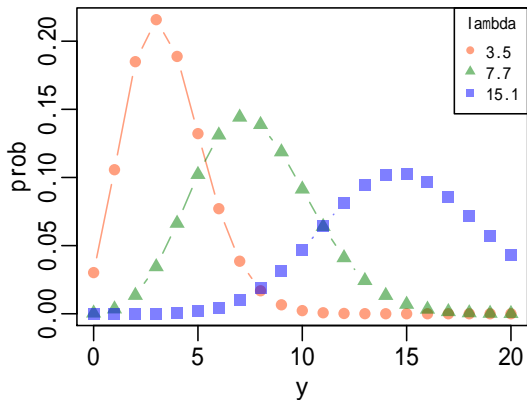
統計モデルの部品としてポアソン分布が選んだ理由:

- 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)
- 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

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

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^{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?

「あてはめる」ことは推定すること

ゆうど

尤度 (likelihood) とは何か?

- 最尤推定法では、^{ゆうど}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:

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

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!)

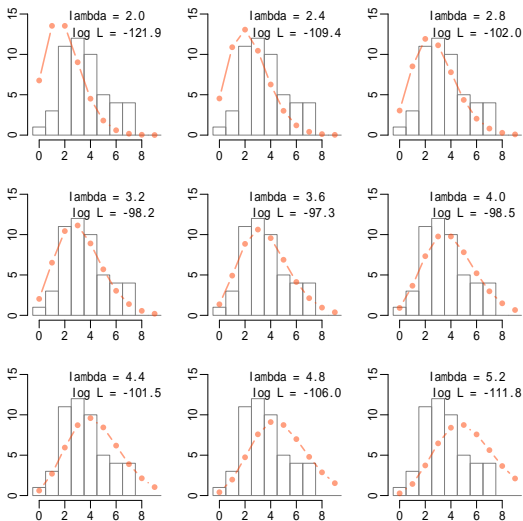
そこで, パラメータの最尤推定では, **対数尤度関数** (log likelihood function) を使う

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

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

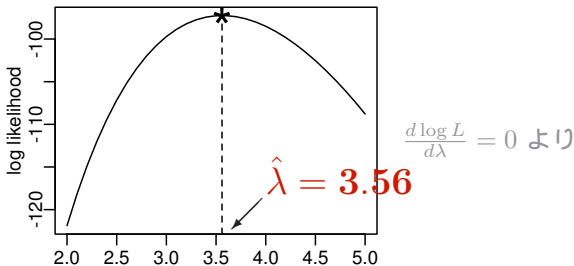
まずは, 平均をあらわすパラメータ λ を変化させていったときに, ポアソン分布のカタチと対数尤度がどのように変化するかを調べてみましょう

λ を変えるとあてはまりの良さが変わる

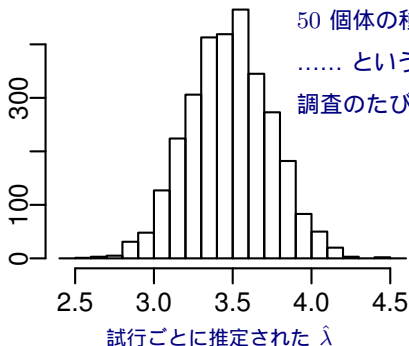


対数尤度を最大化する $\hat{\lambda}$ をさがす

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



- 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

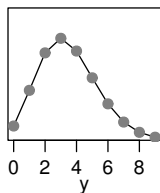
5. 統計モデルの要点

乱数発生・推定・予測

統計モデルとデータの対応づけ

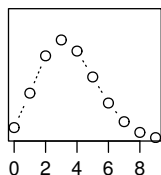
確率分布: 乱数発生 と 推定

(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



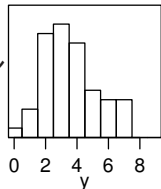
データをサンプル

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

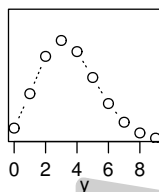
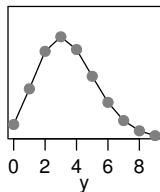


観測されたデータ

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 予測

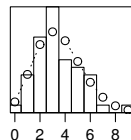
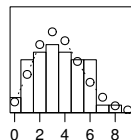
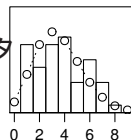
(人間には見えない)
真の統計モデル
 $\lambda = 3.5$ のポアソン分布



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

この講義で登場する確率分布

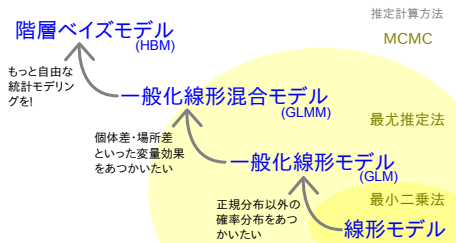
- **ポアソン分布:** $y \in \{0, 1, 2, 3, \dots\}$ となるデータ, 「 y 回なにかがおこった」
- **二項分布:** $y \in \{0, 1, 2, \dots, N\}$ となるデータ, 「 N 個のうち y 個で何かがおこった」
- **正規分布:** $-\infty < y < \infty$ の連続値をとるデータ
- その他あれこれ — ちょっと登場するだけ

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.....

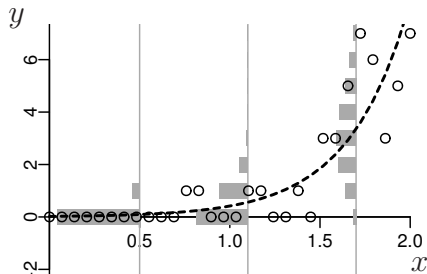
- この講義では多種多様な確率分布を**あつかいません**
- しかし **確率分布を混ぜあわせる** ことによって、自分で確率分布を作り出すことができます
- ハナシの後半に登場する GLMM や階層ベイズモデル

線形モデルの発展



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)