統計モデルの基礎: 統計モデルって何だろう?

確率分布，最尤推定，ポアソン回帰

久保拓弥 kubo@ees.hokudai.ac.jp

生態学基礎論 (生物多様性論 II) http://goo.gl/Omkhqm

2014–01–20

北海道大学・環境科学院の授業

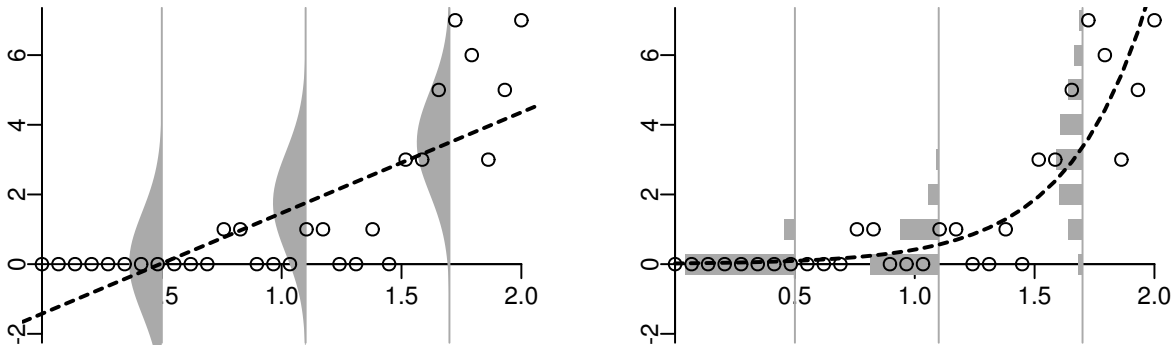
ファイル更新時刻: 2014–02–03 22:01

この時間のハナシ I

1. 例題: 種子数の統計モデリング
   R でデータ操作実演などしつつ
2. データと確率分布の対応
   確率分布は統計モデルの重要な部品
3. ポアソン分布って何?
   平均を変えると分布のカタチが変わる
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
   もっともっともらしい推定?
5. 統計モデルの要点
   乱数発生・推定・予測
6. ポアソン回帰の統計モデル
   一般化線形モデルにとりくんでみる
7. 例題: いろいろな要因に影響される種子数
   植物個体の属性，あるいは実験処理が種子数に影響?

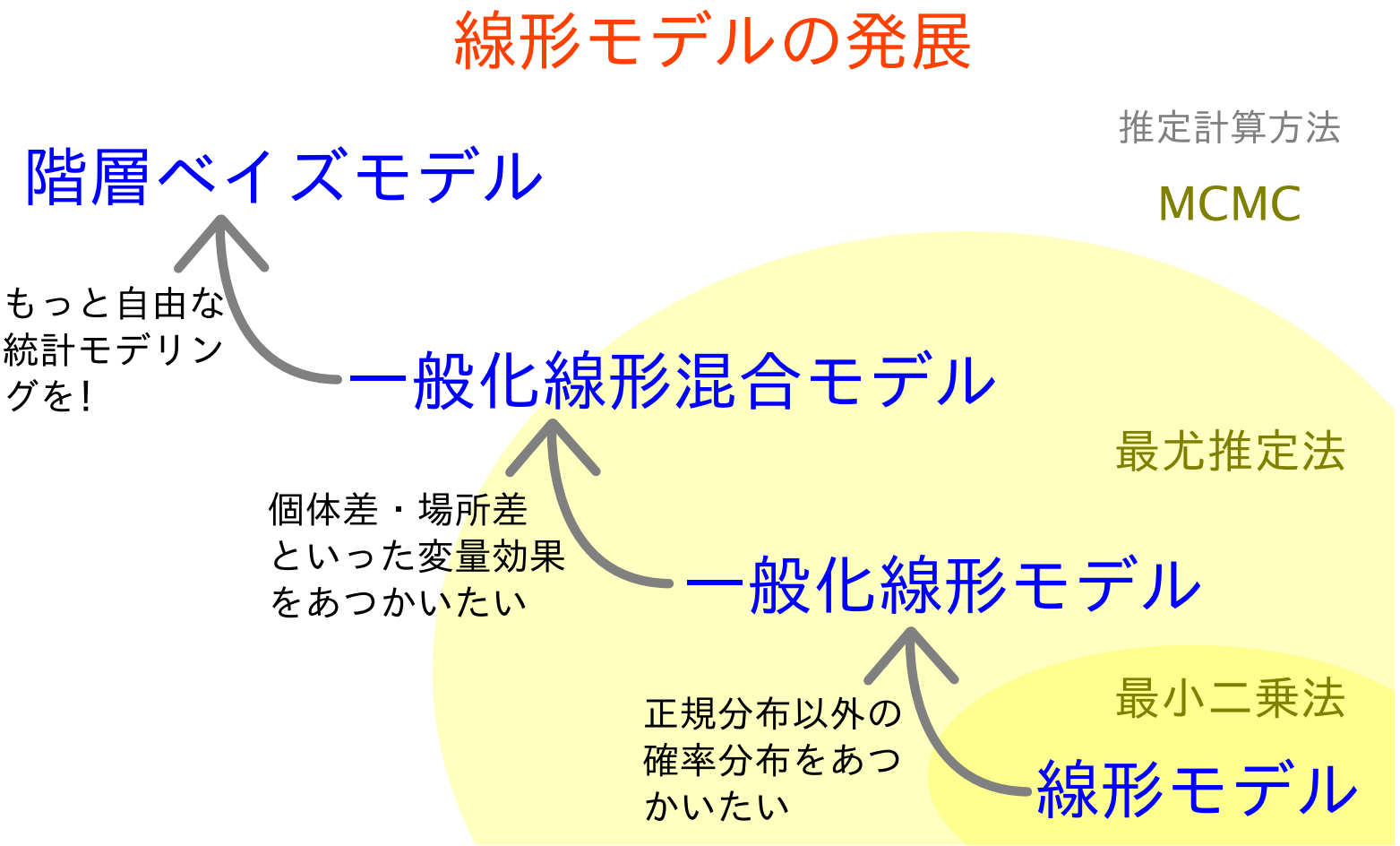
この時間のハナシ II

8. GLM の詳細を指定する
   確率分布・線形予測子・リンク関数
9. R でパラメーターを推定
   あてはまりの良さは対数尤度関数で評価
10. 推定されたモデルを使って予測
    推定された結果とデータを比較する
11. 「処理をした・しなかった」効果も統計モデルに入れる
    GLM の因子型説明変数



くわしい説明の前に
統計モデリング授業前半で
「説明したいこと」
をとりあえず要約してみます

この授業であつかう統計モデルたち

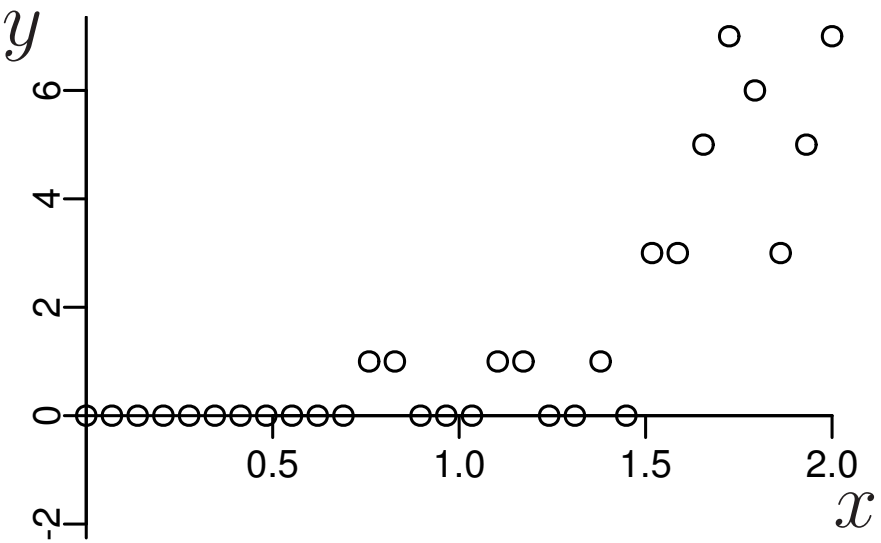


データの特徴にあわせて線形モデルを改良・発展させる

0 個, 1 個, 2 個と数えられるデータ

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

応答変数



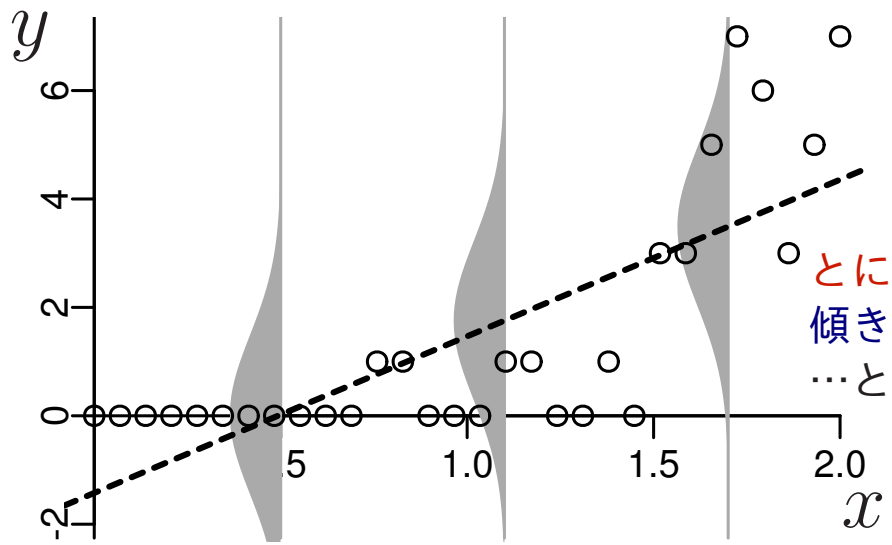
説明変数

- 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?

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

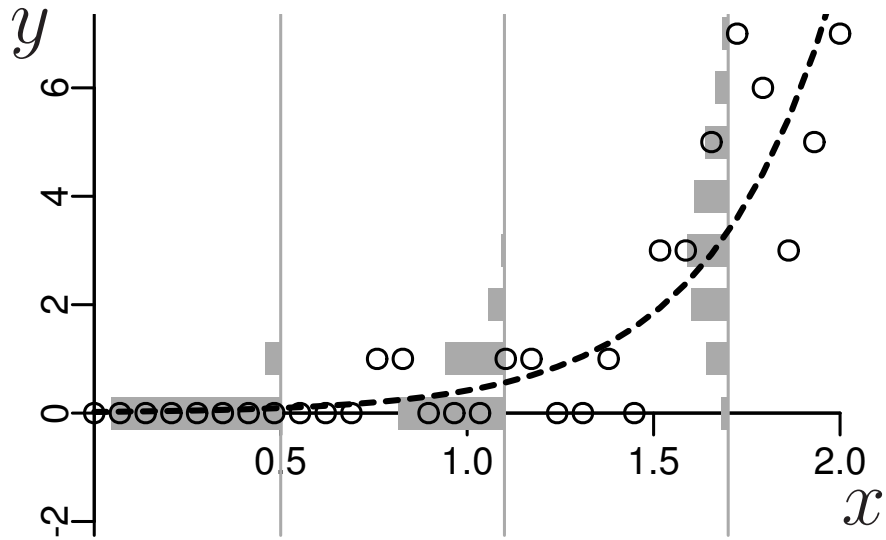
説明変数

- 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

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

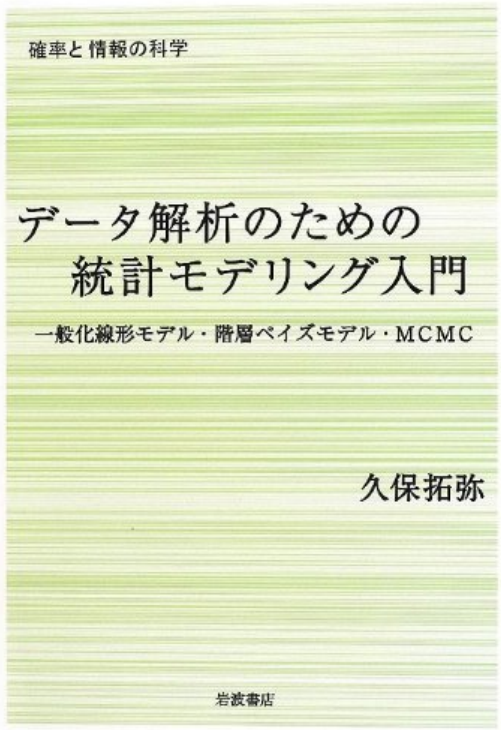
データの構造や性質をよく見て
統計モデルの部品である
確率分布などを選んでいく

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 2–3 章」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

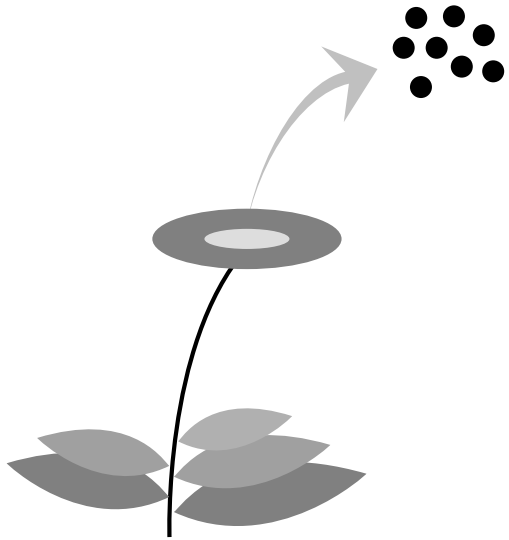


2. 例題: 種子数の統計モデリング

R でデータ操作実演などしつつ

まあ，かなり単純な例から始めましょう

この授業では架空植物の架空データをあつかう

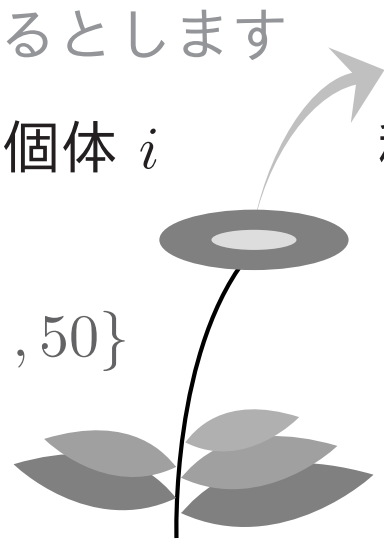


理由: よけいなことは考えなくてすむので

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
```

R でデータの様子をながめる

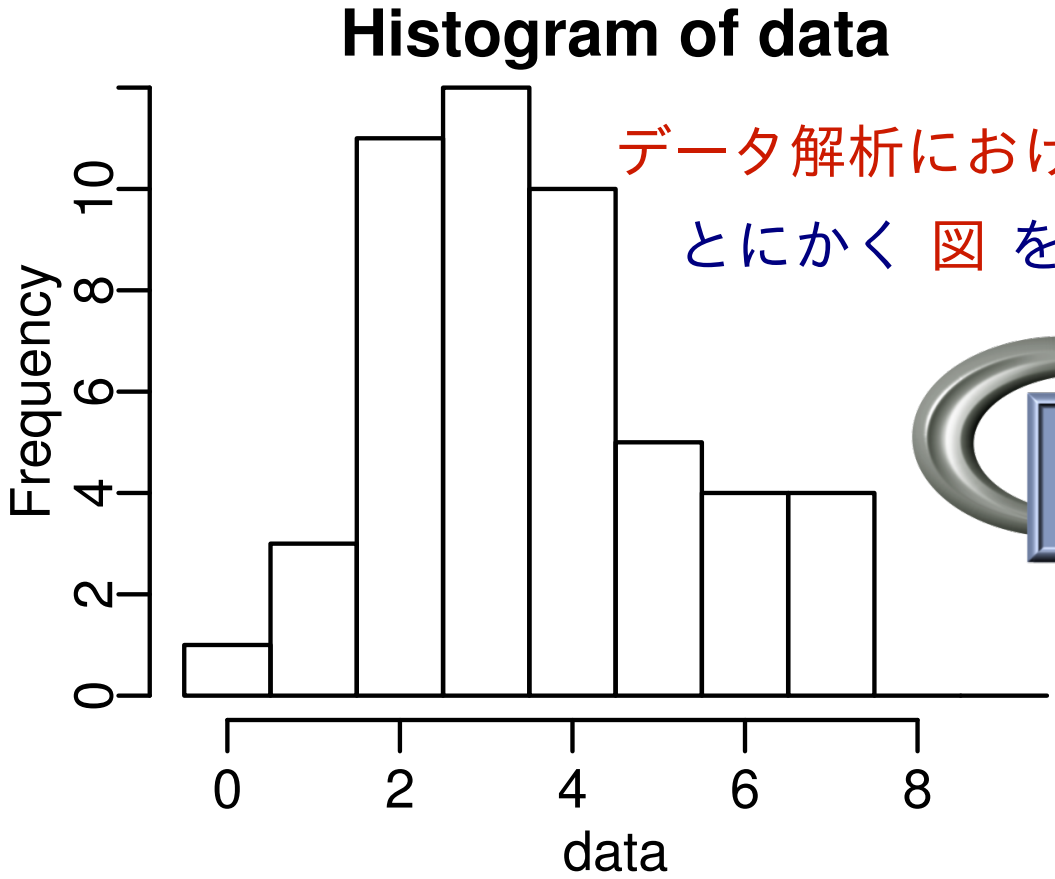
の table() 関数を使って種子数の頻度を調べる

```
> 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!

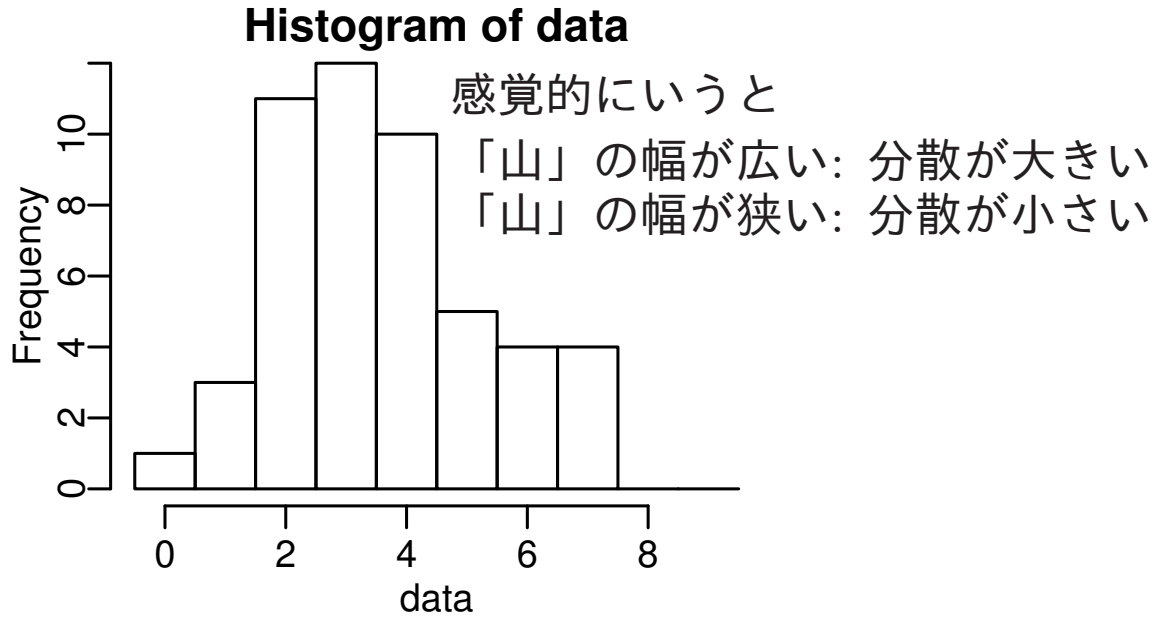
「ばらつき」の統計量

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. データと確率分布の対応

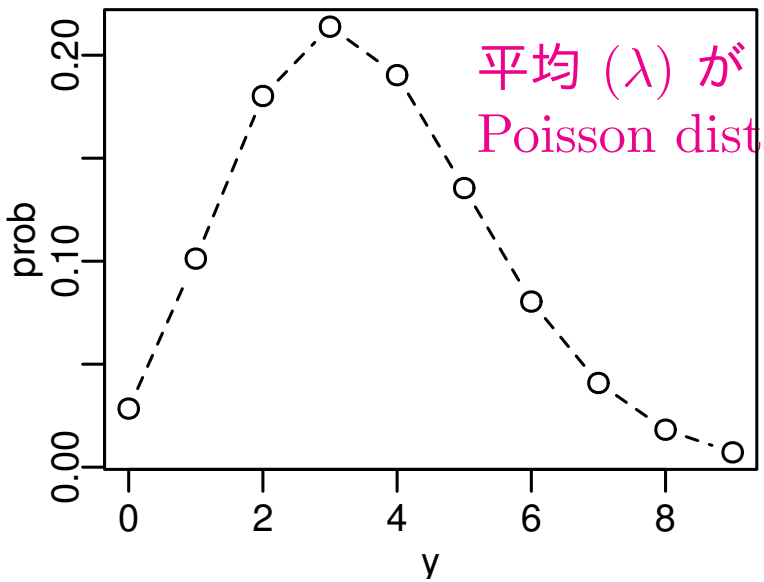
確率分布は統計モデルの重要な部品

ばらついてる「データ」を近似する道具

ポアソン分布とは何か?

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)
```

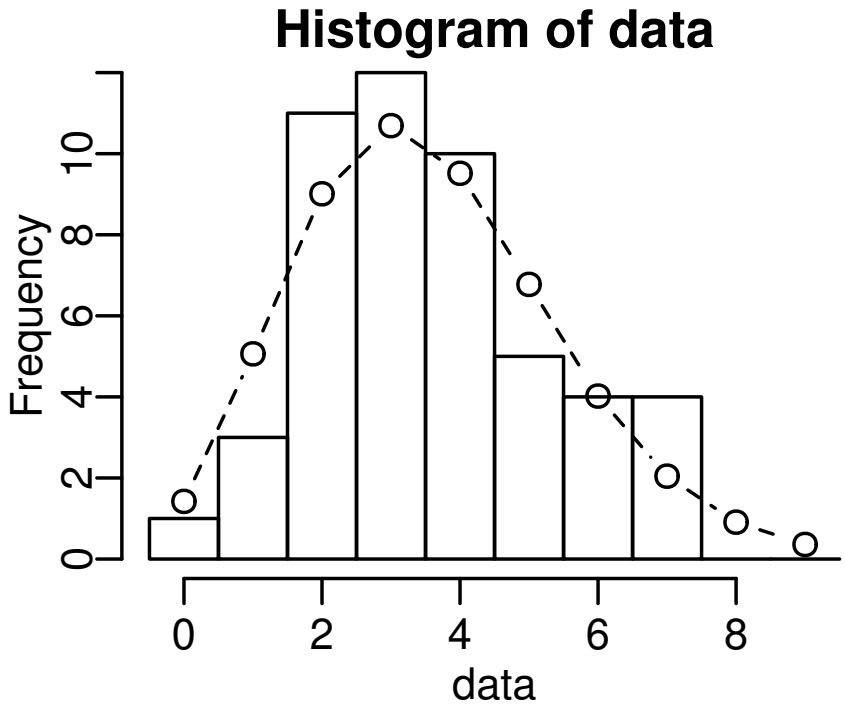


平均 (λ) が 3.56 である
Poisson distribution

```
> # 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
```

4. ポアソン分布って何?

平均を変えると分布のカタチが変わる

確率分布のカタチをきめるパラメーター

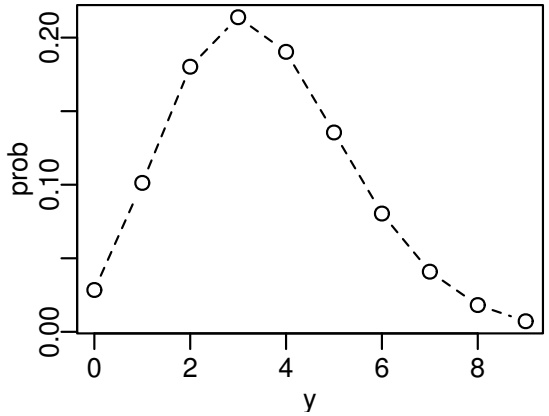
ポアソン分布を数式で表現してみる

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

Paramter λ, the mean of 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
```

- 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?

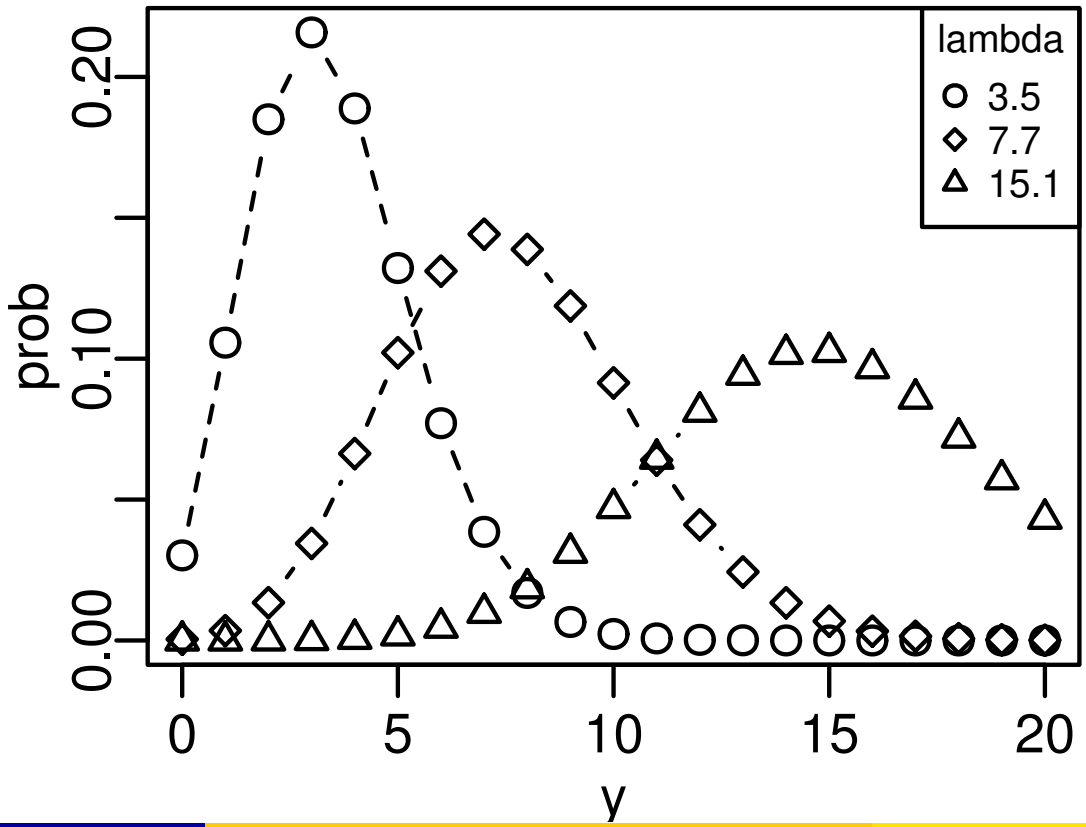
統計モデルの部品としてポアソン分布が選んだ理由:

- 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



各個体の y_i が独立にポアソン分布にしたがう

……ってどういう意味?

- 個体 1 の種子数は平均 λ のポアソン分布にしたがうと仮定する
  $\rightarrow$ 観測された種子数は $y_1 = 2$ だった
- 個体 2 の種子数は平均 λ のポアソン分布にしたがうと仮定する
  $\rightarrow$ 観測された種子数は $y_2 = 2$ だった
- 個体 3 の種子数は平均 λ のポアソン分布にしたがうと仮定する
  $\rightarrow$ 観測された種子数は $y_3 = 4$ だった
- —(以下,同様)—

といった意味 (他個体とは無関係,ということ)

このように仮定すると,全 50 個体のデータから全個体に共通する λ は 3.56 ぐらいではないかなぁといった憶測が可能になる

— (つづく)

5. 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?

「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 (ML) estimation 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

尤度 $L(\lambda)$ はパラメーター λ の関数

この例題の尤度:

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

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

尤度はしんどい → 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

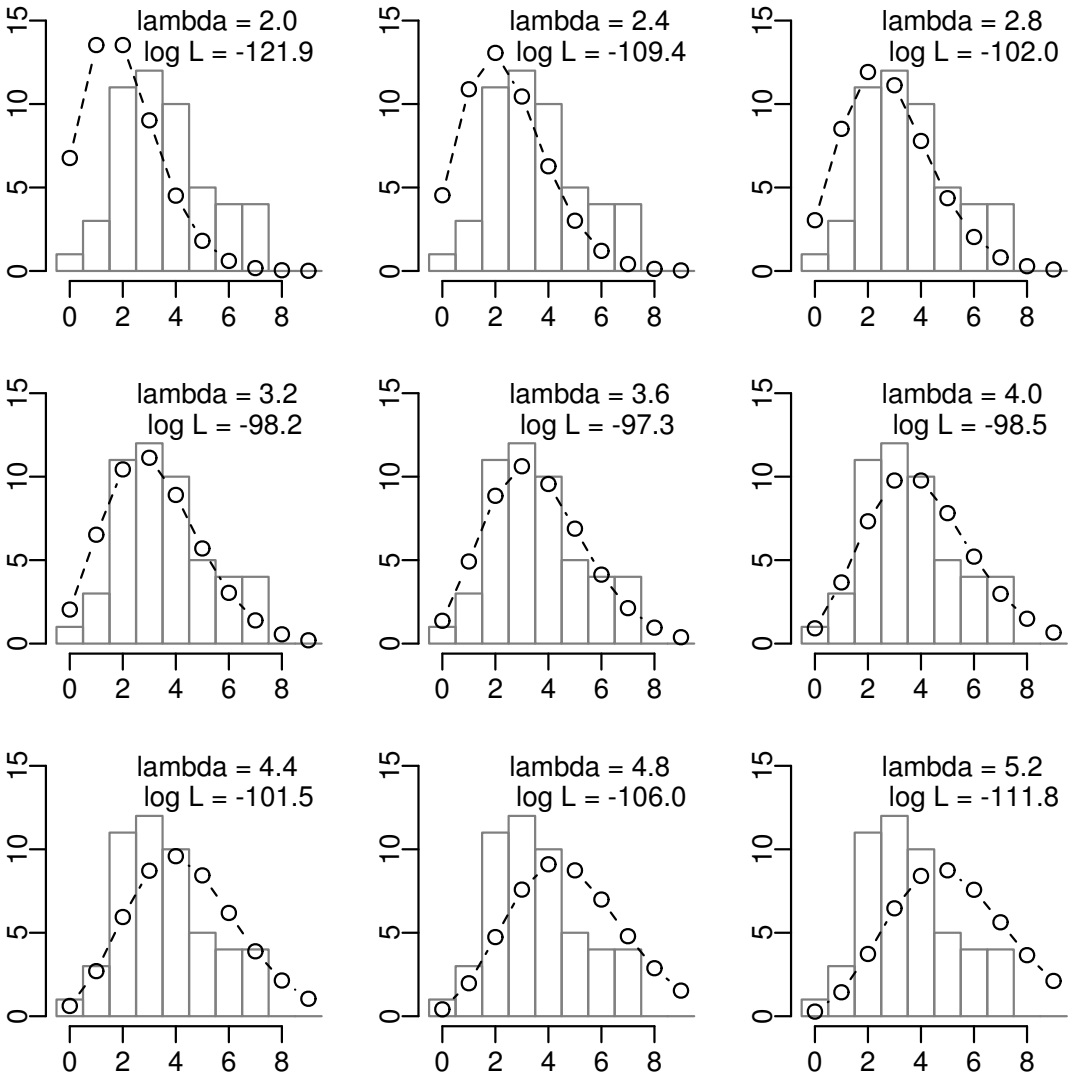
$$\log L(\lambda) = \sum_i \left(y_i \log \lambda - \lambda - \sum_k^{y_i} \log k \right)$$

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

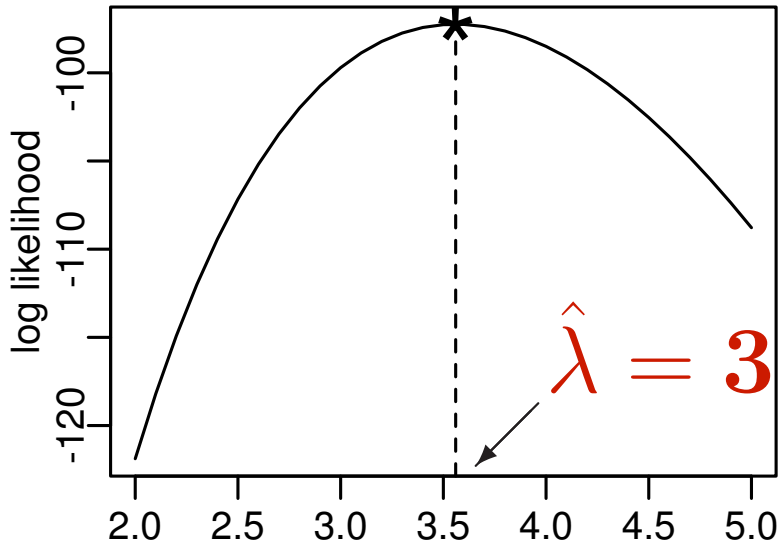
対数尤度 $\log L$

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

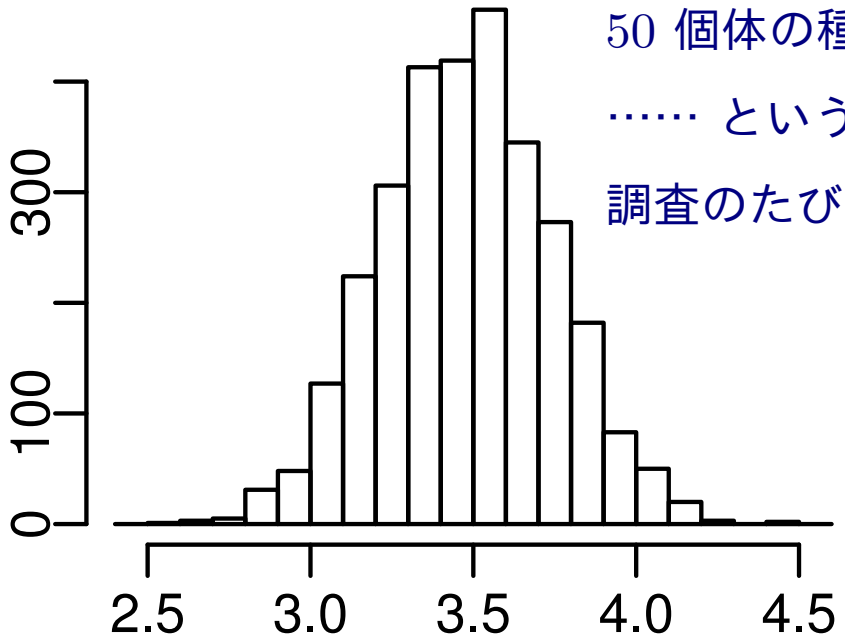


$\frac{d \log L}{d\lambda} = 0$ より

- 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

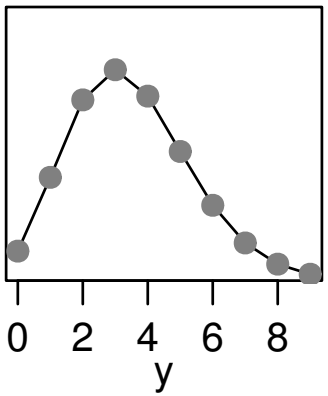
6. 統計モデルの要点

乱数発生・推定・予測

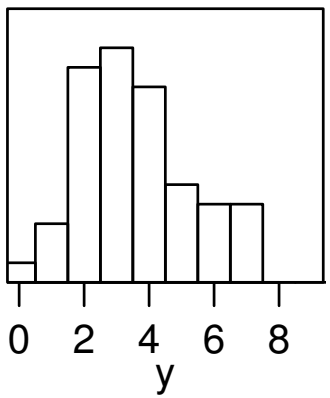
統計モデルとデータの対応づけ

統計学における 推定 (estimation) – 統計モデルのあてはめ

（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



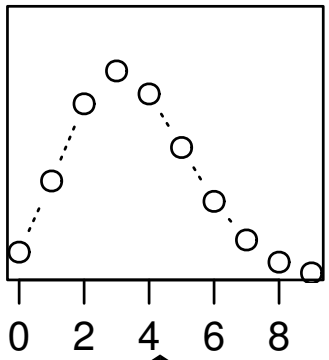
データをサンプル



観測されたデータ

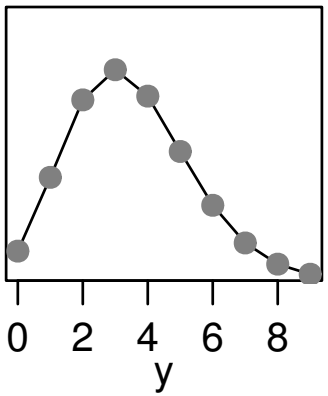
パラメーター推定

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



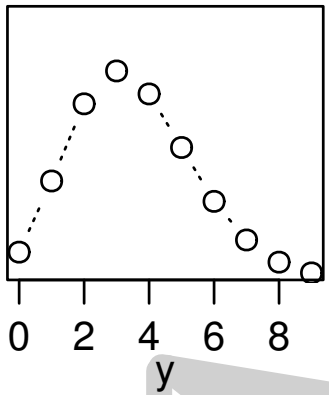
統計モデルを使った prediction 予測

（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

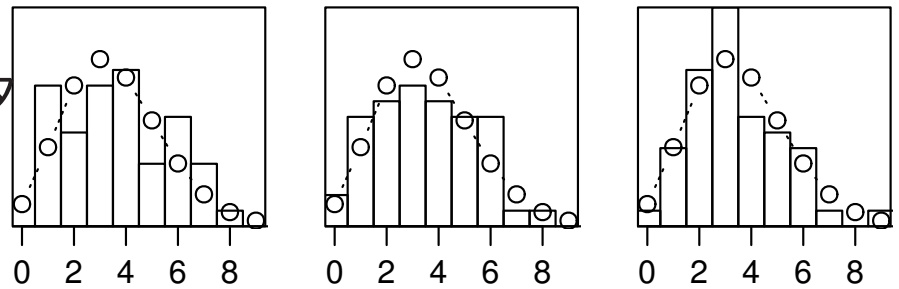


新しいデータをサンプル

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？



• • •

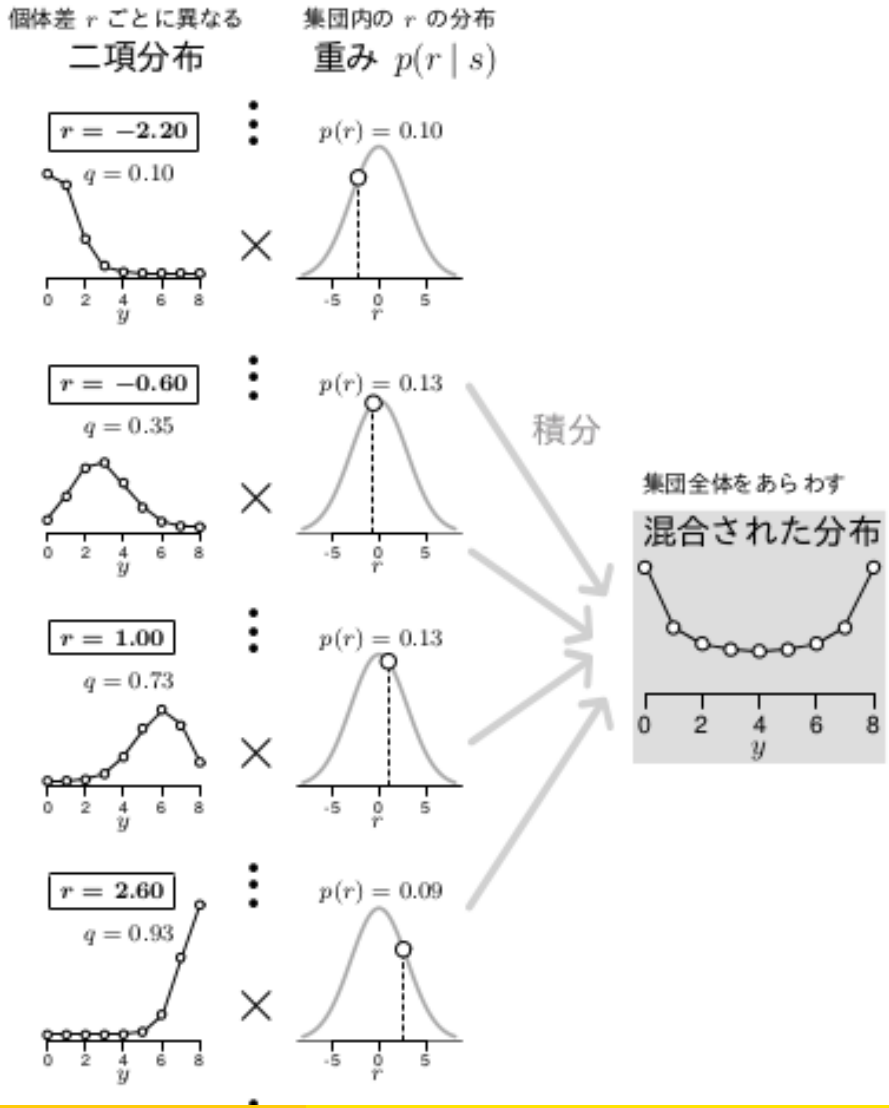
同じ調査方法で得られた新データ

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
- 一様分布，ガンマ分布 — ちょっと登場するだけ

いろいろな確率分布があるけれど……

- この授業では多種多様な確率分布をあつかいません
- この授業後半では，「ポアソン分布と正規分布を混ぜる」「二項分布と正規分布を混ぜる」といった，確率分布まぜワザを使って，現実にみられる複雑な分布を再現してみます



7. ポアソン回帰の統計モデル

一般化線形モデルにとりくんでみる

「なんでも直線回帰」ではない!

一般化線形モデルって何だろう?

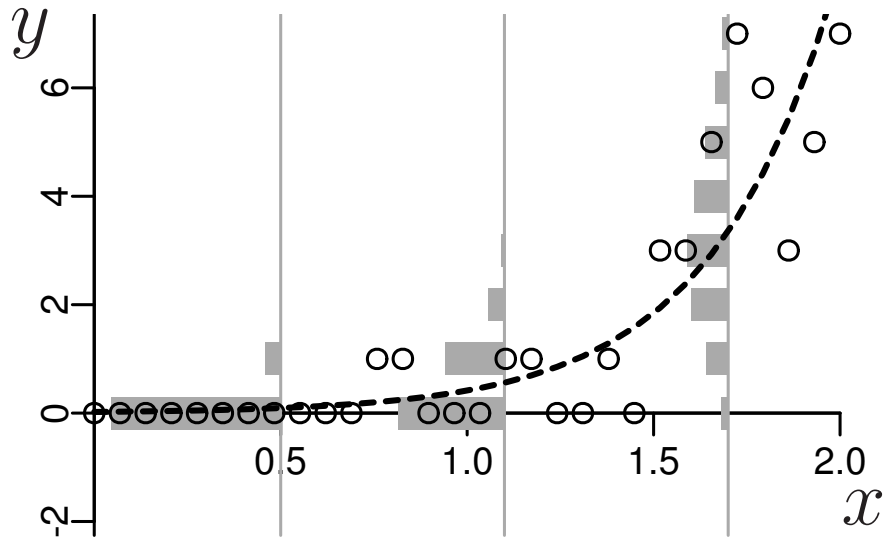
一般化線形モデル (GLM)

- ポアソン回帰 (Poisson regression)
- ロジスティック回帰 (logistic regression)
- 直線回帰 (linear regression)
- ……

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

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

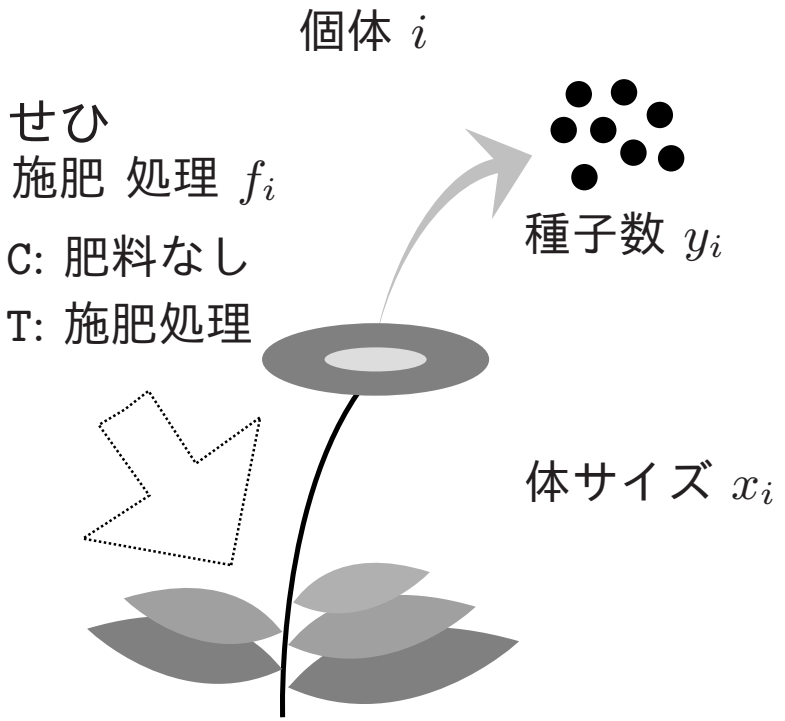
8. 例題: いろいろな要因に影響される種子数

植物個体の属性，あるいは実験処理が種子数に影響?

まずはデータの概要を調べる

個体サイズと実験処理の効果を調べる例題

- 応答変数: 種子数 $\{y_i\}$
- 説明変数:
  - 体サイズ $\{x_i\}$
  - 施肥処理 $\{f_i\}$



標本数

- 無処理 ($f_i = \mathtt{C}$): 50 sample ($i \in \{1, 2, \cdots 50\}$)
- 施肥処理 ($f_i = \mathtt{T}$): 50 sample ($i \in \{51, 52, \cdots 100\}$)

データファイルを読みこむ

data3a.csv は CSV (comma separated value) format file なので，R で読みこむには以下のようにする:

```
> d <- read.csv("data3a.csv")
```

データは d と名付けられた data frame (「表」みたいなもの) に格納される

とりあえず

data frame d を表示

```
> d
    y     x  f
1   6  8.31  C
2   6  9.44  C
3   6  9.50  C
... (中略) ...
99  7 10.86  T
100 9  9.97  T
```

data frame d を調べる: dx, dy

```
> d$x
  [1]  8.31  9.44  9.50  9.07 10.16  8.32 10.61 10.06
  [9]  9.93 10.43 10.36 10.15 10.92  8.85  9.42 11.11
... (中略) ...
 [97]  8.52 10.24 10.86  9.97

> d$y
  [1]  6  6  6 12 10  4  9  9  9 11  6 10  6 10 11  8
 [17]  3  8  5  5  4 11  5 10  6  6  7  9  3 10  2  9
... (中略) ...
 [97]  6  8  7  9
```

data frame d を調べる: d$f — factor type!

施肥処理の有無をあらわす f 列はちょっと様子がちがう

```
> d$f
  [1] C C C C C C C C C C C C C C C C C C C C C C C C C
 [26] C C C C C C C C C C C C C C C C C C C C C C C C C
 [51] T T T T T T T T T T T T T T T T T T T T T T T T T
 [76] T T T T T T T T T T T T T T T T T T T T T T T T T
Levels: C T
```

因子型データ: いくつかの水準をもつデータ
ここでは C と T の 2 水準

Rのデータのクラスとタイプ

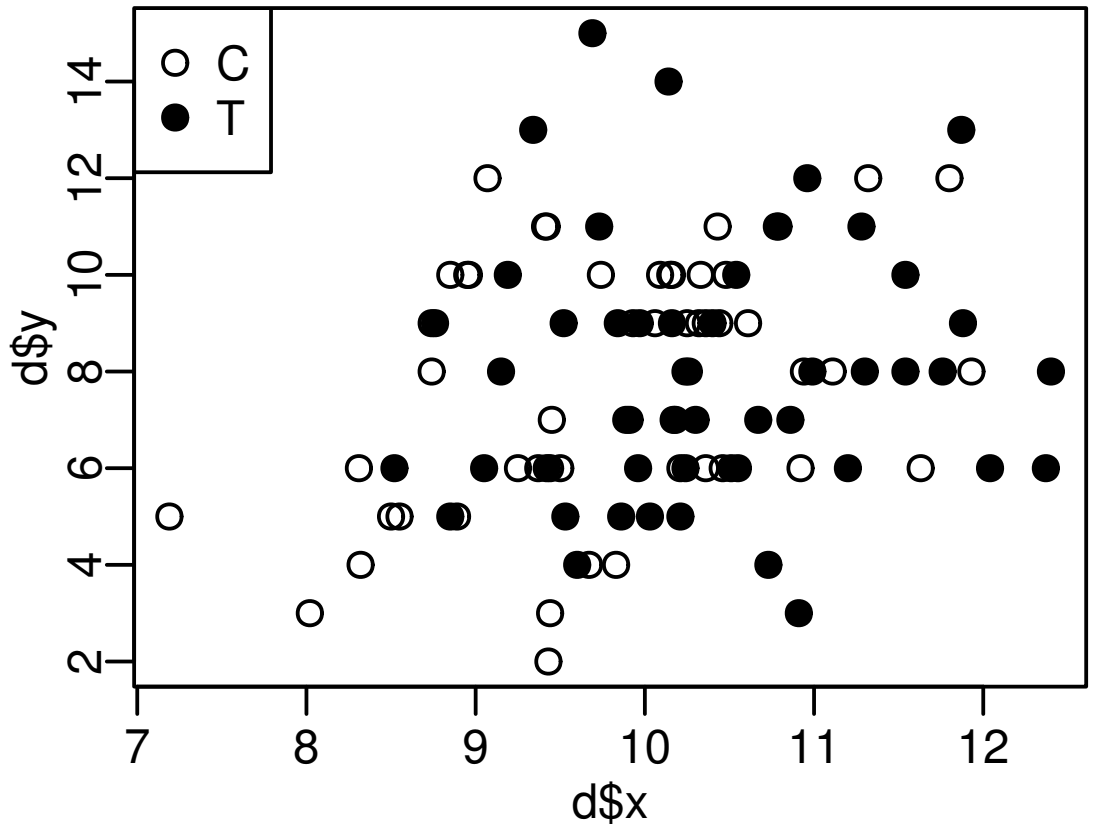
```
> class(d) # d は data.frame クラス
[1] "data.frame"
> class(d$y) # y 列は整数だけの integer クラス
[1] "integer"
> class(d$x) # x 列は実数も含むので numeric クラス
[1] "numeric"
> class(d$f) # そして f 列は factor クラス
[1] "factor"
```

data frame の summary()

```
> summary(d)
       y                x               f
 Min.   : 2.00   Min.   : 7.190   C:50
 1st Qu.: 6.00   1st Qu.: 9.428   T:50
 Median : 8.00   Median :10.155
 Mean   : 7.83   Mean   :10.089
 3rd Qu.:10.00   3rd Qu.:10.685
 Max.   :15.00   Max.   :12.400
```

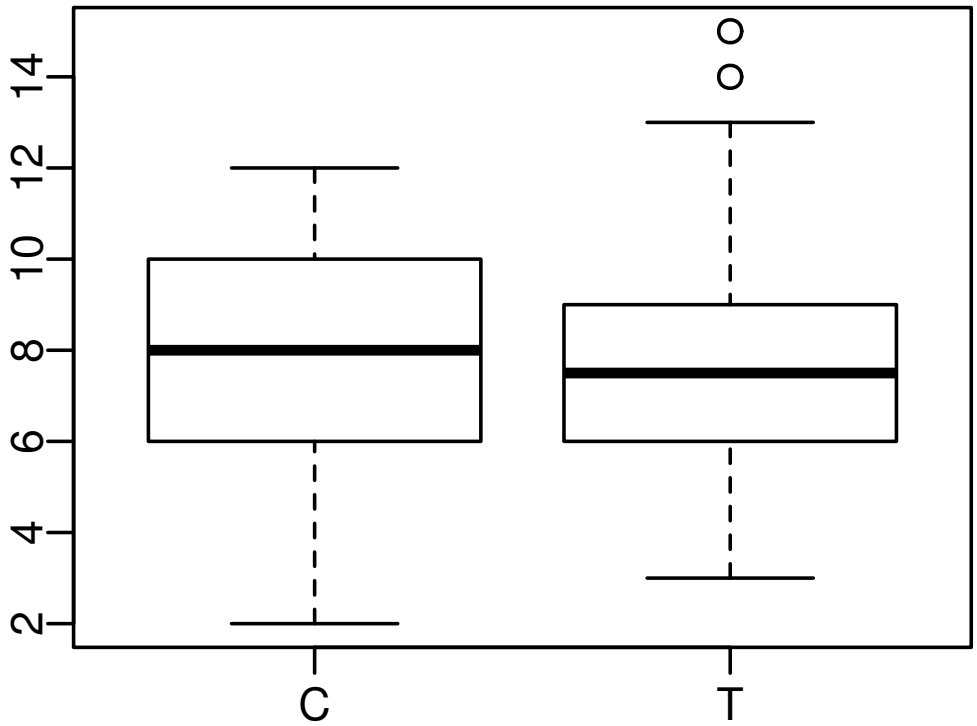
データはとにかく図示する!

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> legend("topleft", legend = c("C", "T"), pch = c(21, 19))
```



施肥処理 f を横軸とした図

```
> plot(d$f, d$y)
```



9. GLM の詳細を指定する

確率分布・線形予測子・リンク関数

ポアソン回帰では log link 関数を使うのが便利

一般化線形モデルを作る

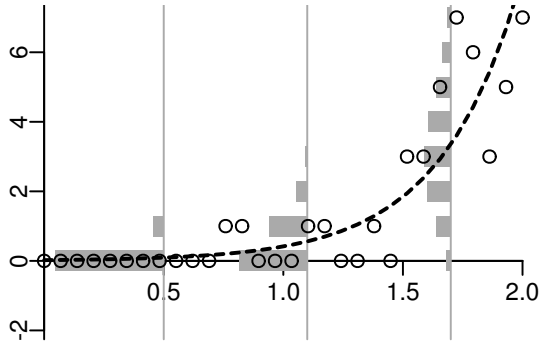
一般化線形モデル (GLM)

- 確率分布は?
- 線形予測子は?
- リンク関数は?

GLM のひとつであるポアソン回帰モデルを指定する

ポアソン回帰のモデル

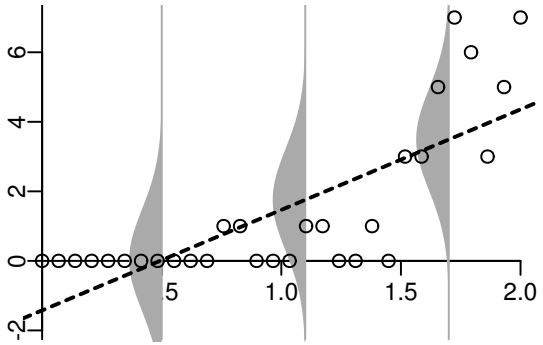
- 確率分布: ポアソン分布
- 線形予測子: $\mathrm{e.g.}, \beta_1 + \beta_2 x_i$
- リンク関数: 対数リンク関数



GLM のひとつである直線回帰モデルを指定する

直線回帰のモデル

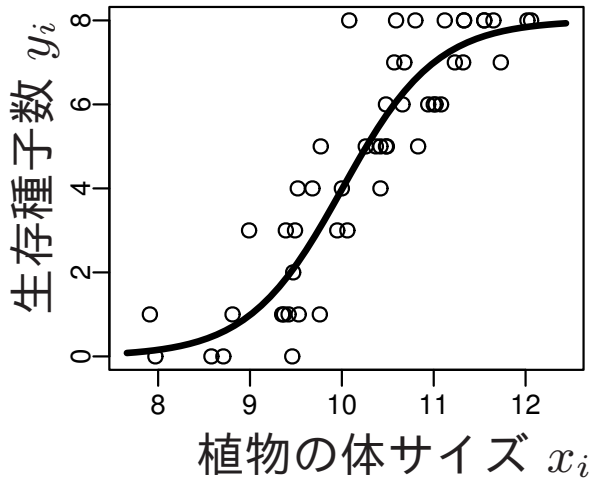
- 確率分布: 正規分布
- 線形予測子: $\mathrm{e.g.,}\ \beta_1 + \beta_2 x_i$
- リンク関数: 恒等リンク関数



GLM のひとつである logistic 回帰モデルを指定する

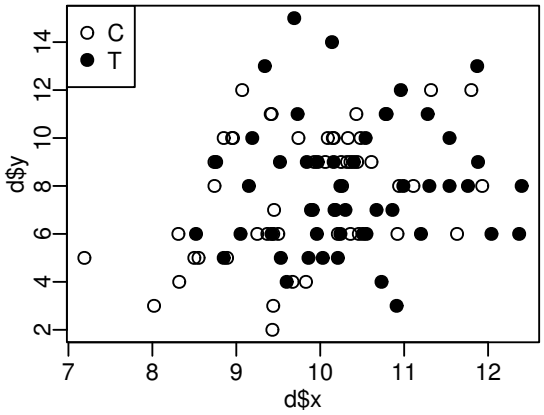
ロジスティック回帰のモデル

- 確率分布: 二項分布
- 線形予測子: e.g., $\beta_1 + \beta_2 x_i$
- リンク関数: logit リンク関数



これは次の時間に説明します

さてさて，この例題にもどって



種子数 y_i は平均 λ_i のポアソン分布にしたがうとしましょう

$$p(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

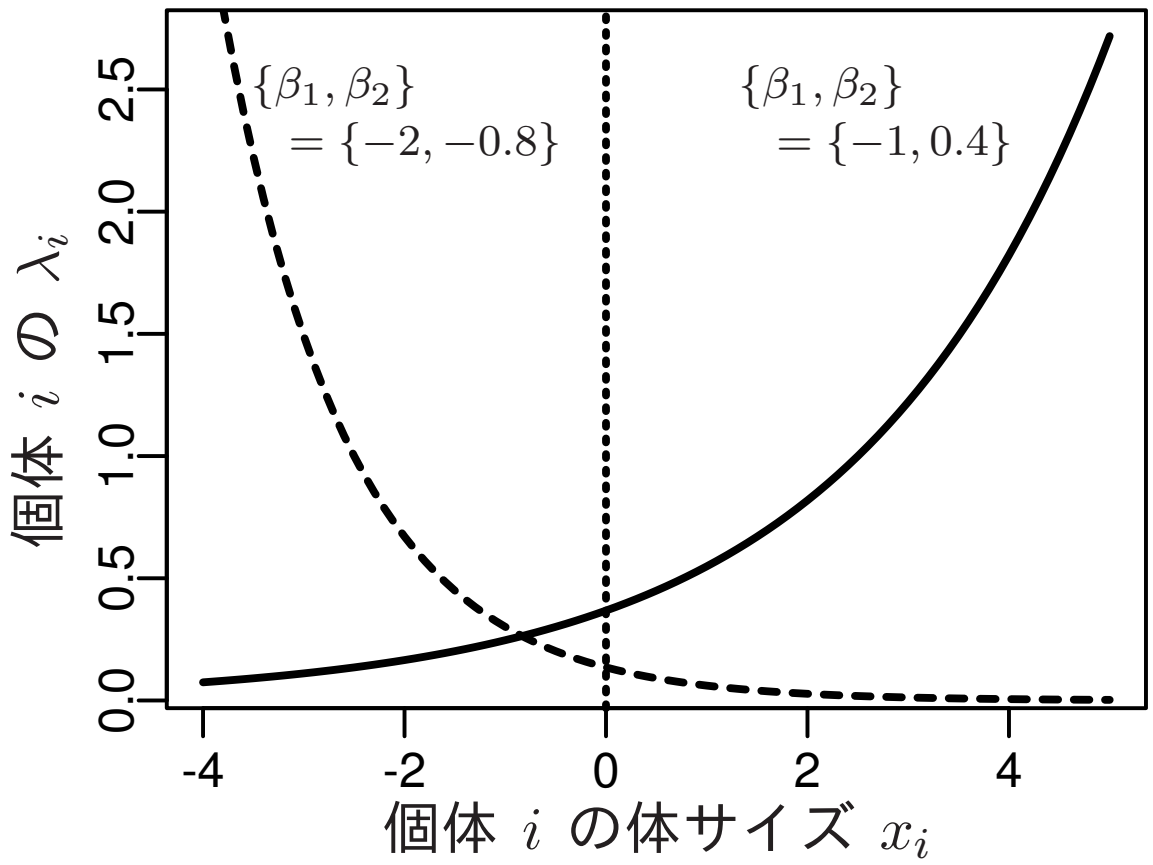
個体 i の平均 λ_i を以下のようにおいてみたらどうだろう……?

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$$

- β_1 と β_2 は係数 (パラメーター)
- x_i は個体 i の体サイズ
- f_i はとりあえず無視

指数関数ってなんだっけ?

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$$



ポアソン回帰でよく使う log link 関数

個体 i の平均 λ_i

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$$

$$\Updownarrow$$

$$\log(\lambda_i) = \beta_1 + \beta_2 x_i$$

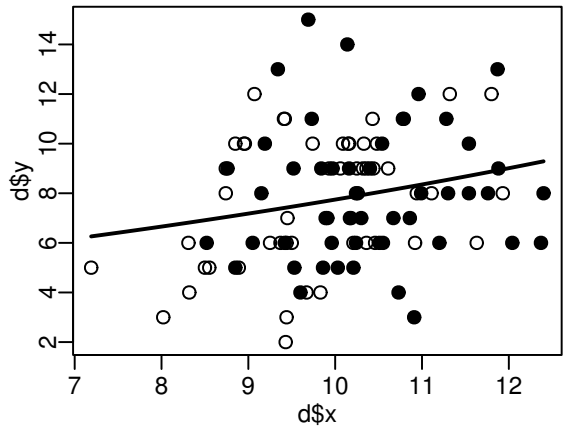
このように $\log$(平均) $=$ 線形予測子

とする連結関数を log link function という

この例題のための統計モデル

ポアソン回帰のモデル

- 確率分布: ポアソン分布
- 線形予測子: $\beta_1 + \beta_2 x_i$
- リンク関数: 対数リンク関数



10. R でパラメーターを推定

あてはまりの良さは対数尤度関数で評価

推定計算はコンピューターにおまかせ

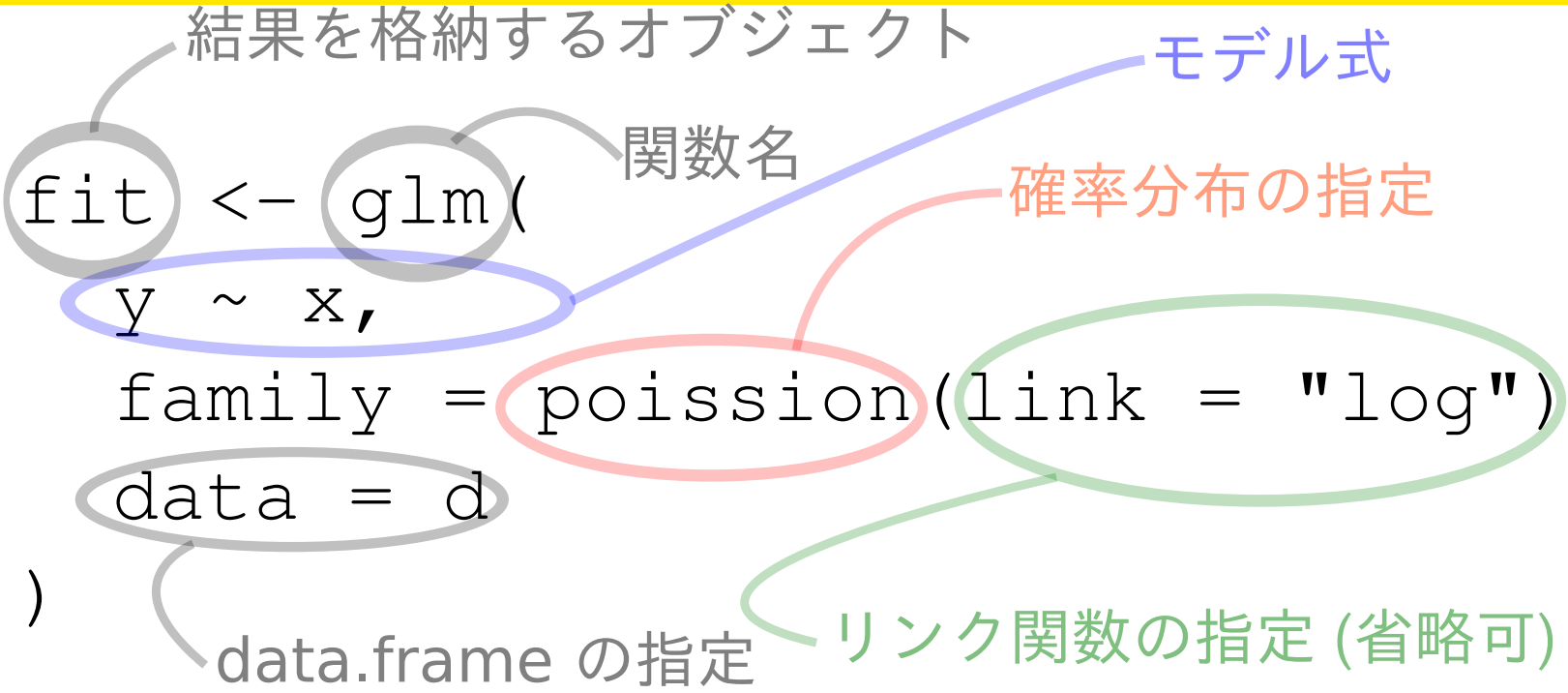
glm() 関数の指定

```
> d
    y     x  f
1   6  8.31  C
2   6  9.44  C
3   6  9.50  C
... (中略) ...
99  7 10.86  T
100 9  9.97  T
```

これだけで OK!

```
> fit <- glm(y ~ x, data = d, family = poisson)
```

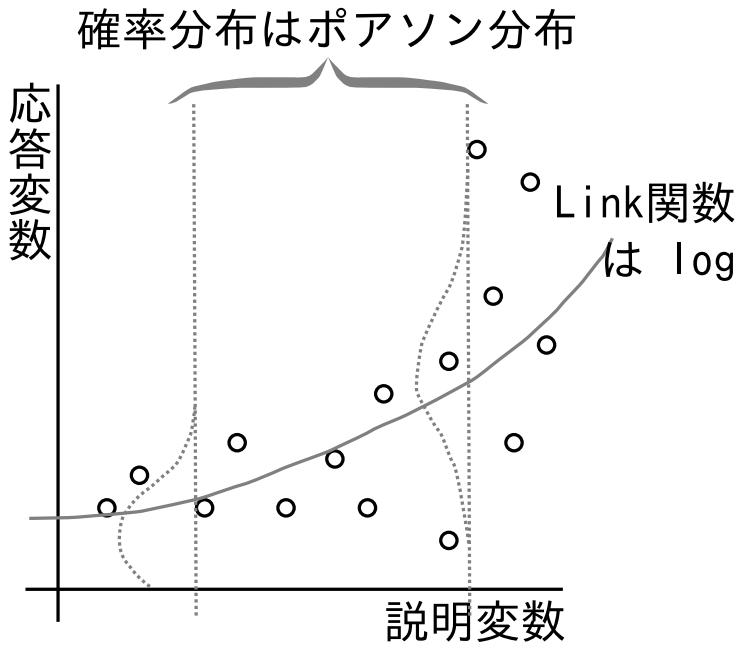
glm() 関数の指定の意味



- モデル式 (線形予測子 z): どの説明変数を使うか?
- link 関数: z と応答変数 (y) 平均値 の関係は?
- family: どの確率分布を使うか?

glm() 関数の指定を再確認

- family: poisson, ポアソン分布
- link 関数: "log"
- モデル式 (線形予測子 z): たとえば y ~ x と指定したとする



- 線形予測子 $z = \beta_1 + \beta_2 x$
  β_1, β_2 は推定すべきパラメーター
- 応答変数の平均値を λ とすると $\log(\lambda) = z$
  つまり $\lambda = \exp(z) = \exp(\beta_1 + \beta_2 x)$
- 応答変数 は平均 λ のポアソン分布に従う: $y \sim \mathrm{Pois}(\lambda)$

glm() 関数の出力 — $\{\beta_1, \beta_2\}$ の数値的最尤推定

```
> fit <- glm(y ~ x, data = d, family = poisson)

all:  glm(formula = y ~ x, family = poisson, data = d)

Coefficients:
(Intercept)           x
     1.2917       0.0757

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:      89.5
Residual Deviance: 85  AIC: 475
```

glm() 関数のくわしい出力

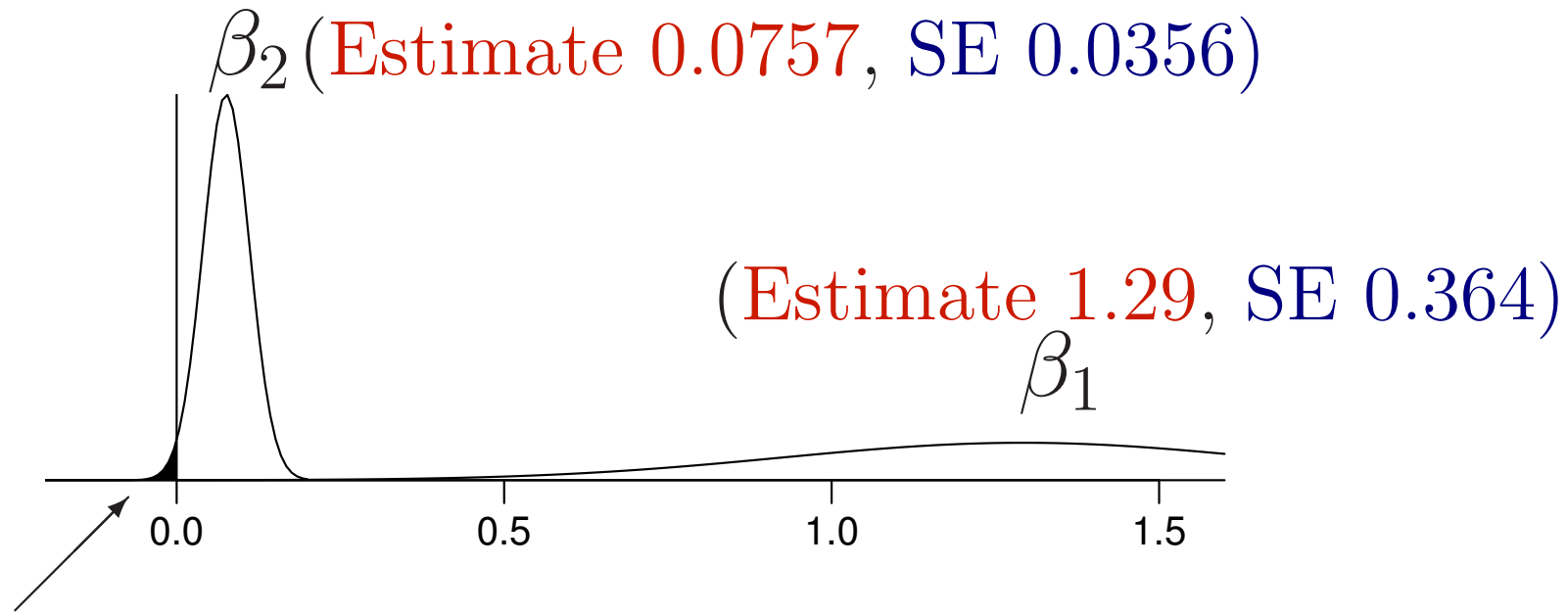
```
> summary(fit)
Call:
glm(formula = y ~ x, family = poisson, data = d)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.368  -0.735  -0.177   0.699   2.376

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2917     0.3637    3.55  0.00038
x             0.0757     0.0356    2.13  0.03358

…… (以下, 省略) ……
```

推定値と標準誤差 と Pr(>|z|)



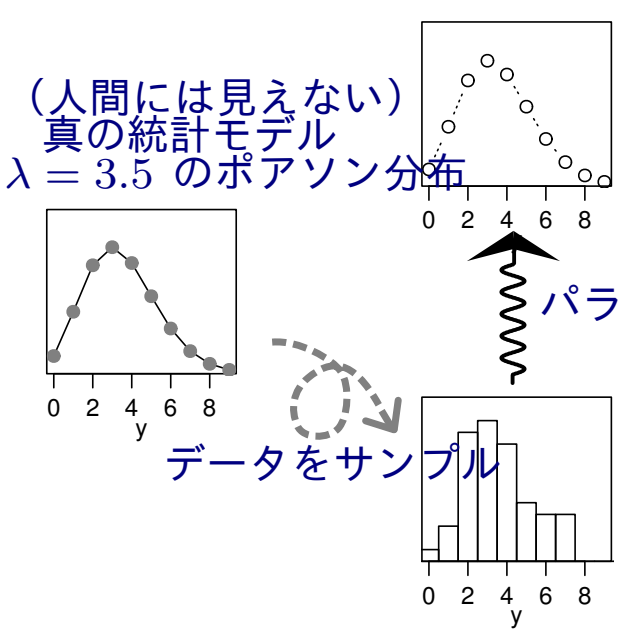
この面積を 2 倍したものが summary(glm()) 出力の Pr(>|z|)

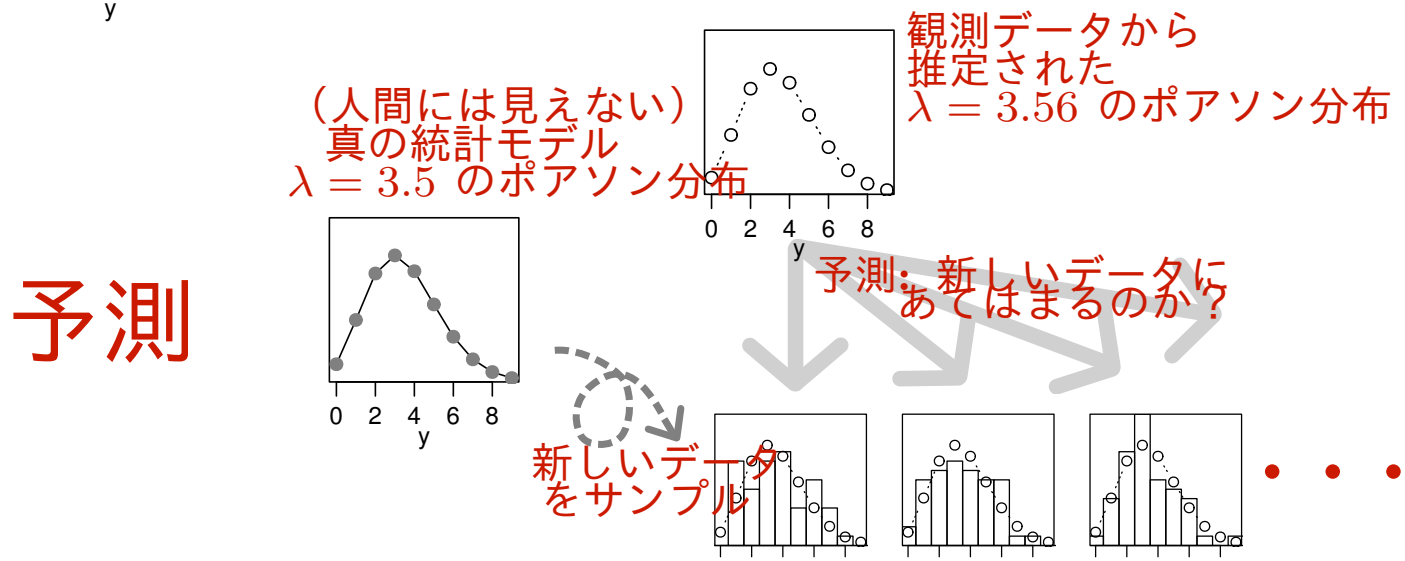
11. 推定されたモデルを使って予測

推定された結果とデータを比較する

ここでも作図が重要!

統計学における推定と予測



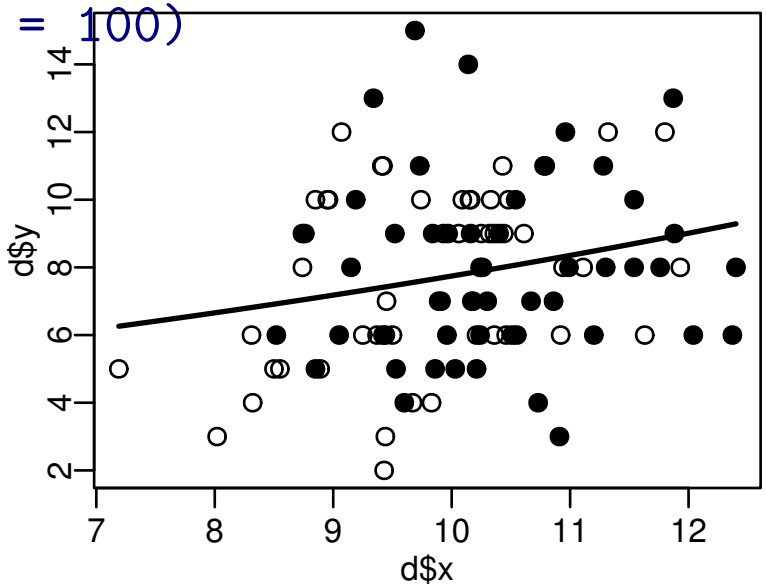


モデルの予測

```
> fit <- glm(y ~ x, data = d, family = poisson)
...
Coefficients:
(Intercept)            x
     1.2917       0.0757

> plot(d$x, d$y, pch = c(21, 19)[d$f]) # data
> xp <- seq(min(d$x), max(d$x), length = 100)
> lines(xp, exp(1.2917 + 0.0757 * xp))
```

ここでは観測データと予測の関係
を見ているだけ，なのだが

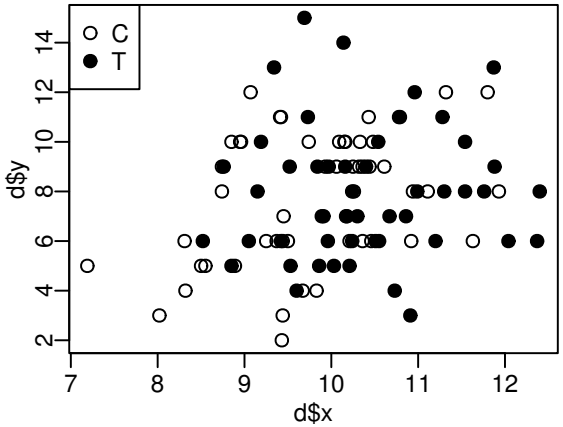


12. 「処理をした・しなかった」効果も統計モデルに入れる

GLM の因子型説明変数

「数量型 + 因子型」という組み合わせで

肥料の効果 f_i もいれましょう



種子数 y_i は平均 λ_i のポアソン分布にしたがう としましょう

$$p(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

個体 i の平均 λ_i を次のようにする

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i + \beta_3 d_i)$$

- β_3 は施肥処理の効果 の 係数
- f_i のダミー変数

$$d_i = \begin{cases} 0 & (f_i = \texttt{C} \text{ の場合}) \\ 1 & (f_i = \texttt{T} \text{ の場合}) \end{cases}$$

glm(y ~ x + f, ...) の出力

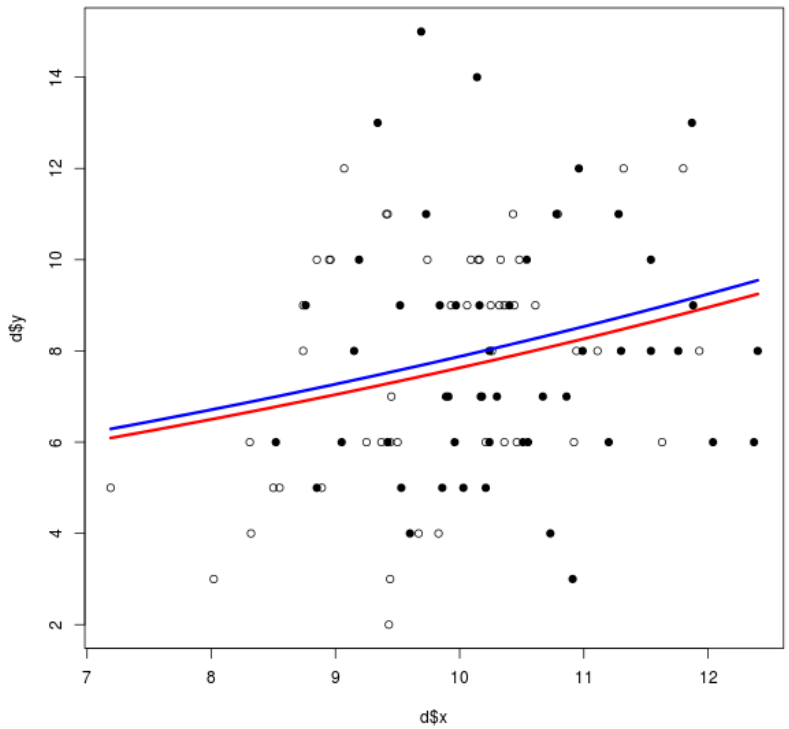
```
> summary(glm(y ~ x + f, data = d, family = poisson))
...(略)...

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2631     0.3696    3.42  0.00063
x             0.0801     0.0370    2.16  0.03062
fT           -0.0320     0.0744   -0.43  0.66703

…… (以下, 省略) ……
```

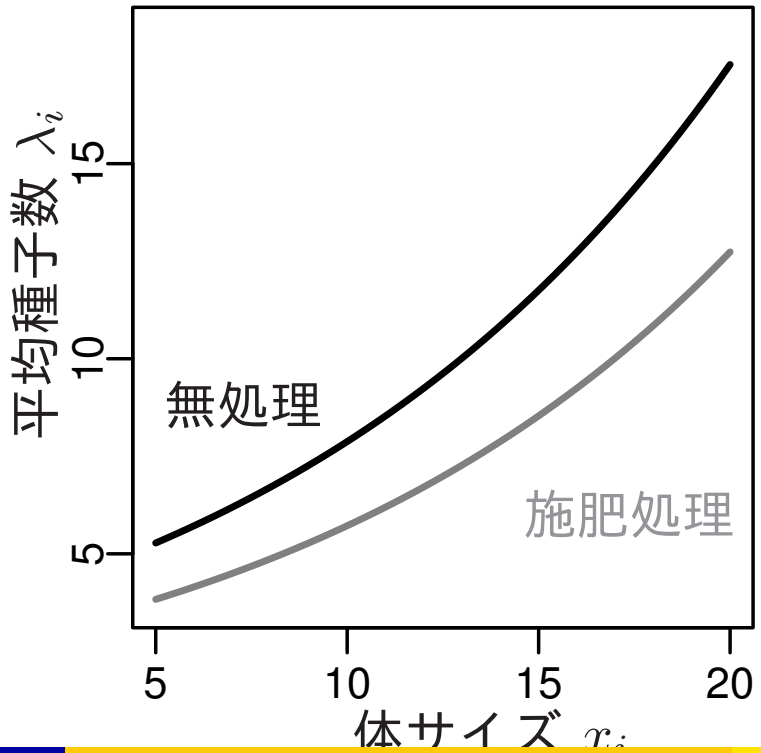
x + f モデルの予測

```
> plot(d$x, d$y, pch = c(21, 19)[d$f]) # data
> xp <- seq(min(d$x), max(d$x), length = 100)
> lines(xp, exp(1.2631 + 0.0801 * xp), col = "blue", lwd = 3) # C
> lines(xp, exp(1.2631 + 0.0801 * xp - 0.032), col = "red", lwd = 3) # T
```

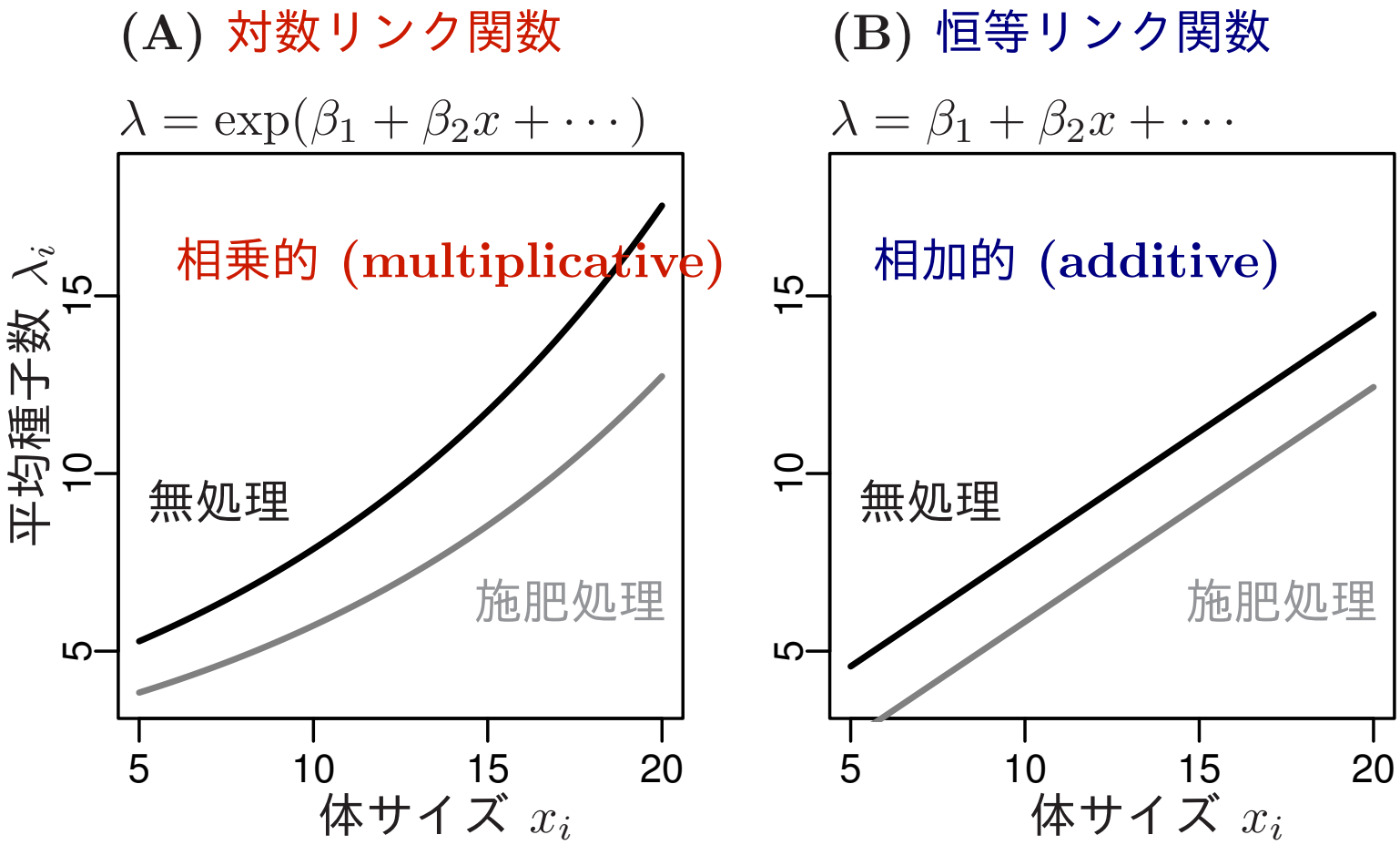


複数の説明変数をいれた場合の統計モデル

- $f_i = \mathtt{C}$: $\lambda_i = \exp(1.26 + 0.0801 x_i)$
- $f_i = \mathtt{T}$: $\lambda_i = \exp(1.26 + 0.0801 x_i - 0.032)$
  $= \exp(1.26 + 0.0801 x_i) \times \exp(-0.032)$

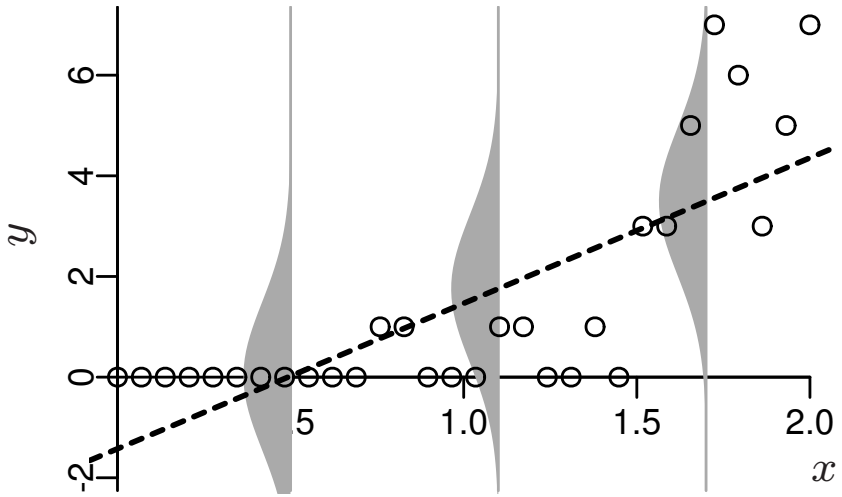


リンク関数が違うとモデルの解釈が異なる

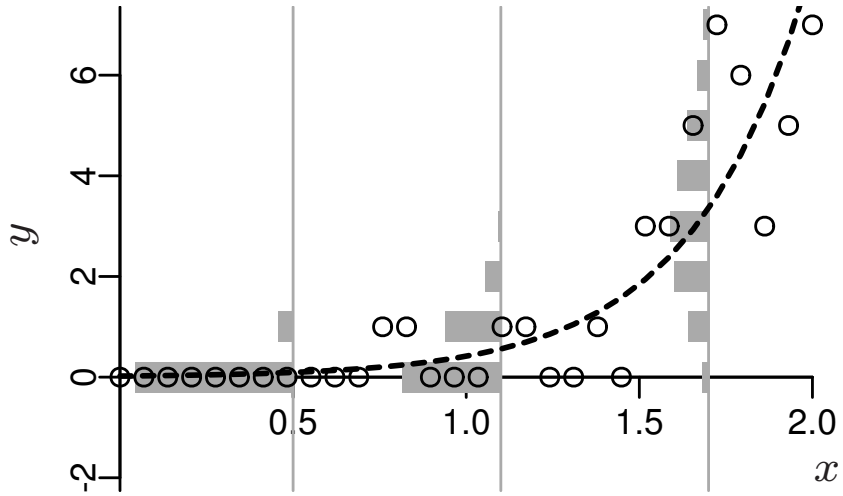


GLM: 適切な確率分布 とリンク関数を選ぶ (とりあえずここまで)

正規分布・恒等リンク関数の統計モデル



ポアソン分布・log リンク関数の統計モデル



しつこいですが ……

「とにかくセンをひけば，それでいい」

というデータ解析はありえない!

念のための注意!

ここでやっていることは，いわゆる

「○○変数変換をして……

正規性がどーのこーの……」

といったこととはまったくちがいます

そもそもポアソン乱数とか対数変換しても正規分布になるわけないでしょ