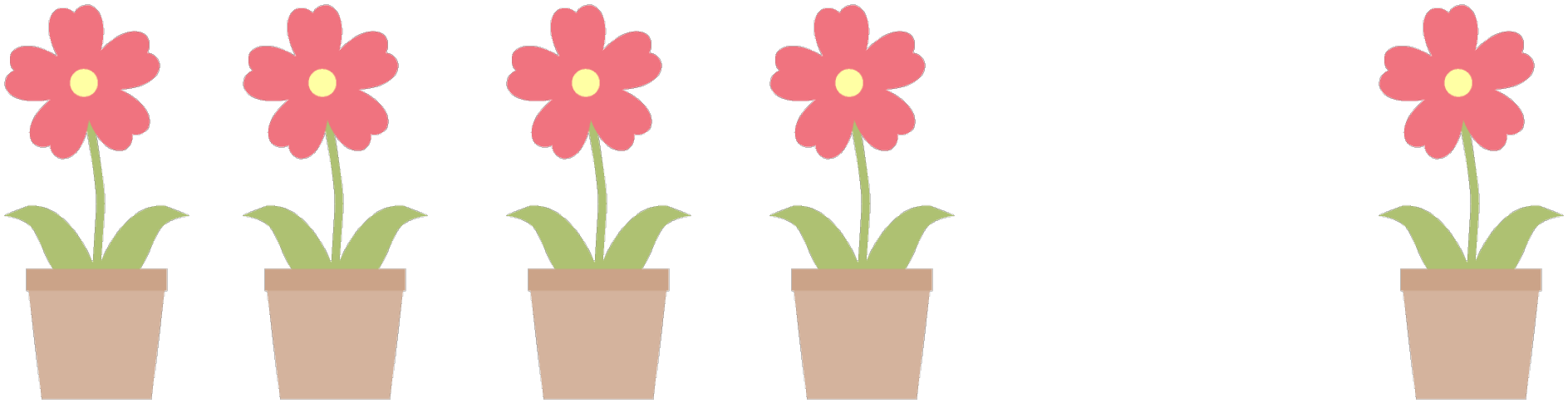
# 統計モデリング入門
# 第二章：確率分布と統計モデルの最尤推定

前半：確率分布とは何か？ポアソン分布とは？

# 今回扱う例題

- 50個体の花　そろぞれの種子数を数える
- i番目の花の種子数：$y_i$



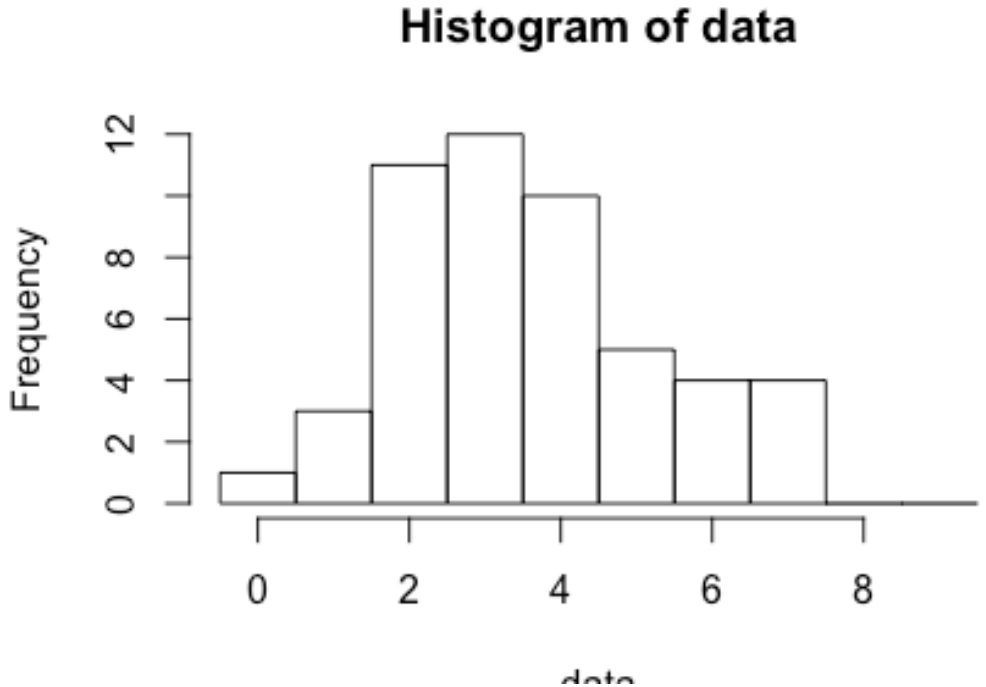
| $i=$ | 1 | 2 | 3 | 4 | …… | 5 0 |
|---|---|---|---|---|---|---|
| $y_i=$ | 2 | 2 | 4 | 6 | …… | 3 |

↑確率変数

# R操作によるデータ解析：入門

取り扱う関数

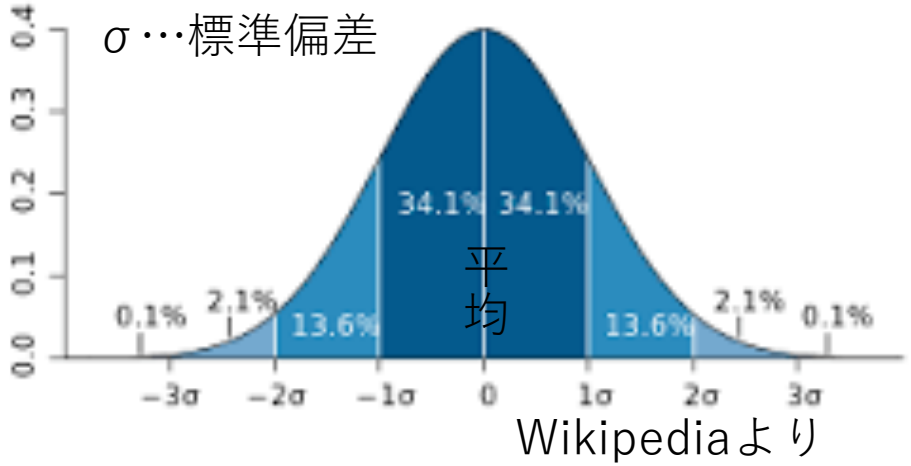
- load(“[データ名]”)
- summary()
- hist
- sd()
- length()
- var()
- sqrt(var())



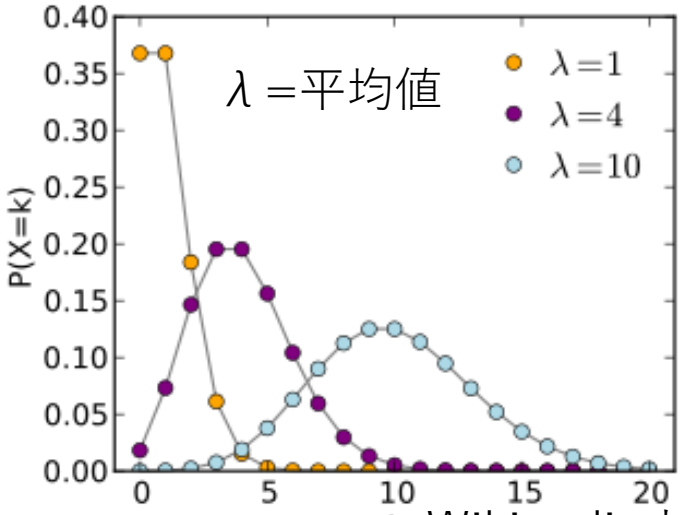

「カウントデータ」…1つ、2つ、と数えてえられるデータ。
（必然的に）非負の整数の集合になる。

# 確率分布とデータの対応関係

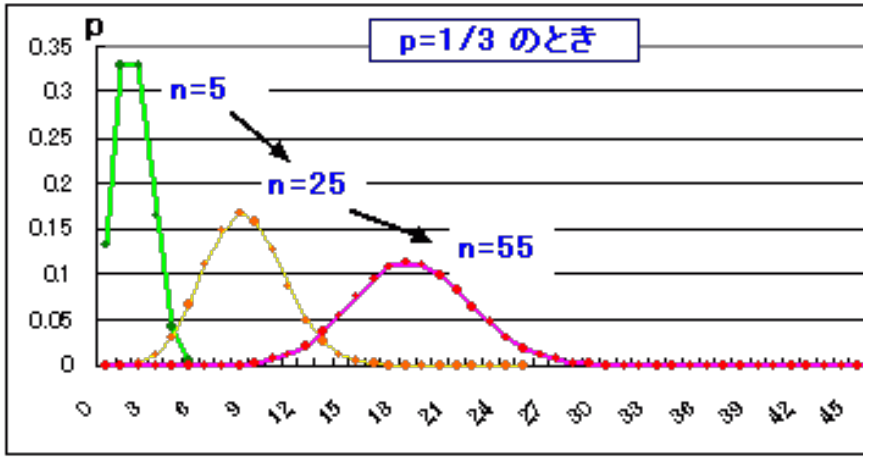
- 「確率分布」…データがどんな様態・形状にバラつくか？
  - E.g. 正規分布、ポアソン分布、二項分布　　etc
  - どの分布が最適か？　→　データの性質による。見極めは難しい。
  - 最適な分布が決定された後　→　「パラメーター」値により形状決定



Wikipediaより



Wikipediaより



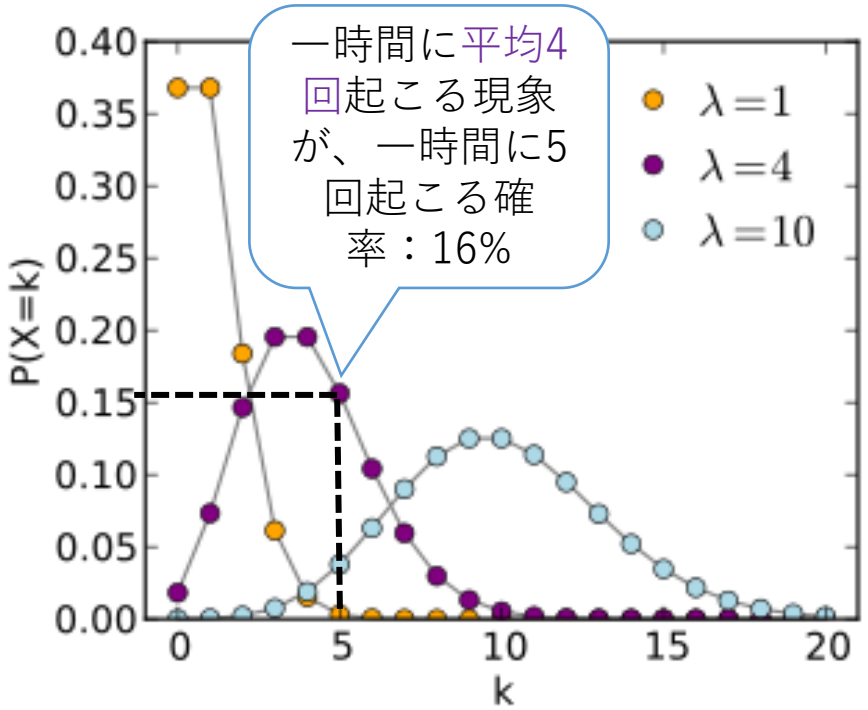
https://goo.gl/U6q20f

- 今回：ポアソン分布を採用(とりあえず)

# ポアソン分布とは？

- 1つの個体から取られる種子：平均で3.56個
- 1つの個体から2個取れる確率は？3個取れる確率は？

→それを示す方法の一つがポアソン分布




λ…平均値
λの値によって、分布の形状が違う。

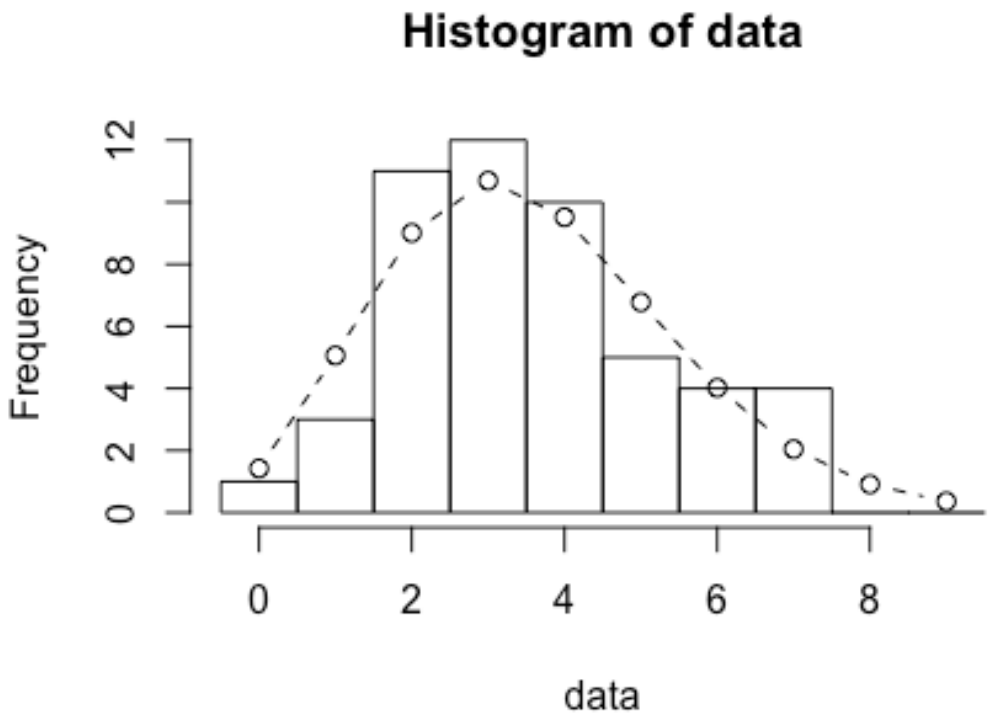
ポアソン分布では、λが唯一のパラメータ

Rでλ=3.56(平均値)のポアソン分布を出力

- y<- 0:9
- prob <- dpois(y, lambda =3.56)
- plot(y, prob, type = “b”, lty =2)

# ポアソン分布とヒストグラムの比較

Rでポアソン分布とヒストグラムを重ねる
- hist(data, breaks = seq(-0.5, 9.5, 1))
- lines(y, 50∗prob, type ="b", lty=2)




- 概ね一致しているように見える

＝観測データのばらつきが、ポアソン分布で表現できているらしい…？

→どのくらい正確に表現できているのか、定量的に評価したい！

# ポアソン分布の定義・性質

確率分布の定義　　$p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$

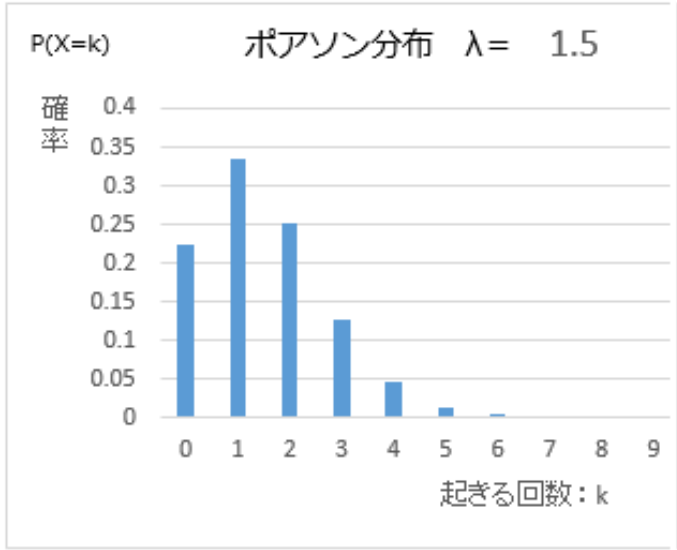
- $p(y|\lambda)$…平均が$\lambda$の時、確率変数がyになる確率

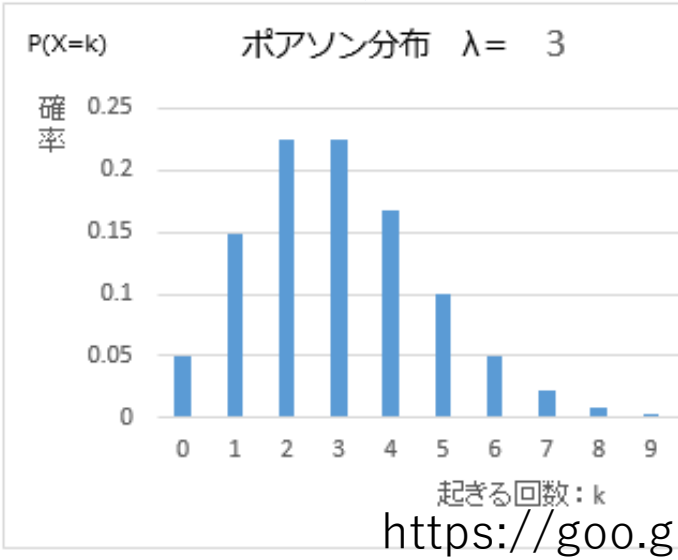
ポアソン分布の性質

- $\sum_{y=0}^{\infty} p(y|\lambda) = 1$　　（全ての$p(y|\lambda)$を足すと1になる）
- 分散と平均は等しい。

  $\lambda$ ＝平均＝分散

$\lambda$の値に関わらず、$p(y|\lambda)$を積み上げると1になる




https://goo.gl/vM7IGL

# ポアソン分布を適用しうるデータの特徴

- データの確率変数が非負の整数
- データの確率変数に下限があるが、上限がわからない
- 観測データでは、平均と分散が大体等しい

これら条件が揃う

→　平均値(仮)をパラメータとするポアソン分布が適用できる「かもしれない」とわかる

# ポアソン分布を適応する時の前提(今回の例で)

- どの植物個体も、観測にあたり条件は画一
- 全ての植物個体に関し、平均の種子数は同じ
- 個々の植物個体は独立（相関・相互作用はない）