ベイズ統計

古谷知之

講義概要

- 自然共役事前分布
- 一変量正規分布のベイズ推定
  - 正規分布の平均のベイズ推定
  - 正規分布の分散のベイズ推定
- 経験ベイズ推定と階層ベイズ推定
- 尤度原理

自然共役事前分布

- 事前分布と事後分布が、互いに似たような確率分布を持つような場合に、その事前分布を自然共役事前分布という
- 主な自然共役事前分布は以下のとおり

事前分布	尤度関数	事後分布
二項分布	ベータ分布	ベータ分布
ポアソン分布	ガンマ分布	ガンマ分布
正規分布(平均)	正規分布	正規分布
正規分布(分散)	正規分布	逆ガンマ分布

試合の得失点差のベイズ推定

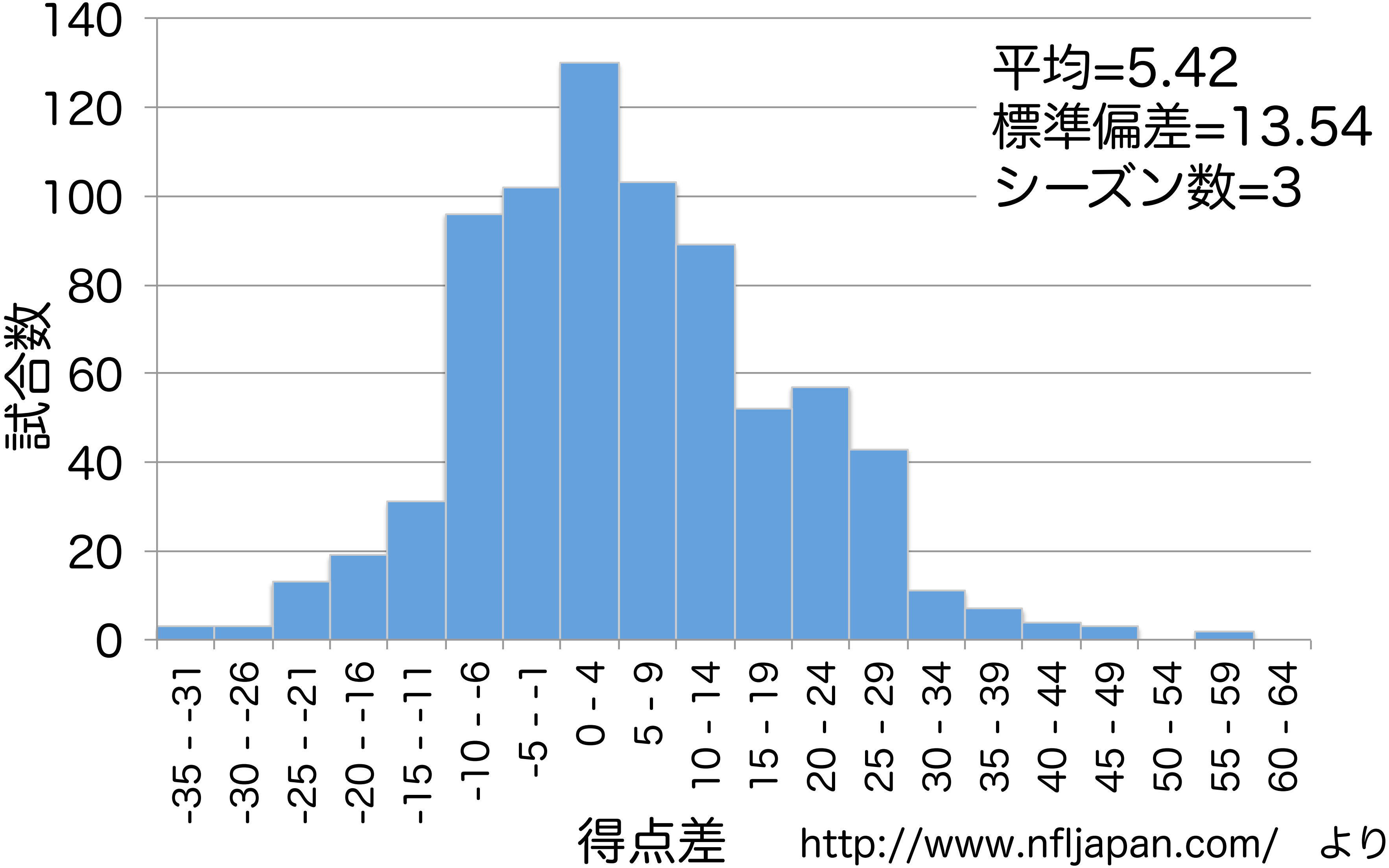
- あなたなあるプロスポーツチームのアナリスト兼ストラテジストだとします。
- チームの攻撃力と守備力を示す統計量(stats)に試合の(ホームチームとアウェイチームの)得失点差(H-A)を用いています。
- 今シーズンのあなたのチームの攻撃力と守備力を試すために、プレシーズンマッチ５試合を戦いました。

http://www.nfljapan.com/

試合の得失点差のベイズ推定

- プレシーズンマッチでは、得失点差の平均と標準偏差は以下のとおりでした。
  – 平均＝6.00、標準偏差＝13.62、5試合分
- 他方、過去3年間のリーグ全体の得失点差の平均と標準偏差は以下のとおりでした。
  – 平均＝5.42、標準偏差＝13.54、768試合分
- あなたのチームの得点差の平均について、少ないプレシーズンマッチの結果から何か言えるのだろうか？

NFL2011-2013シーズンの試合別得点差分布(H-A)

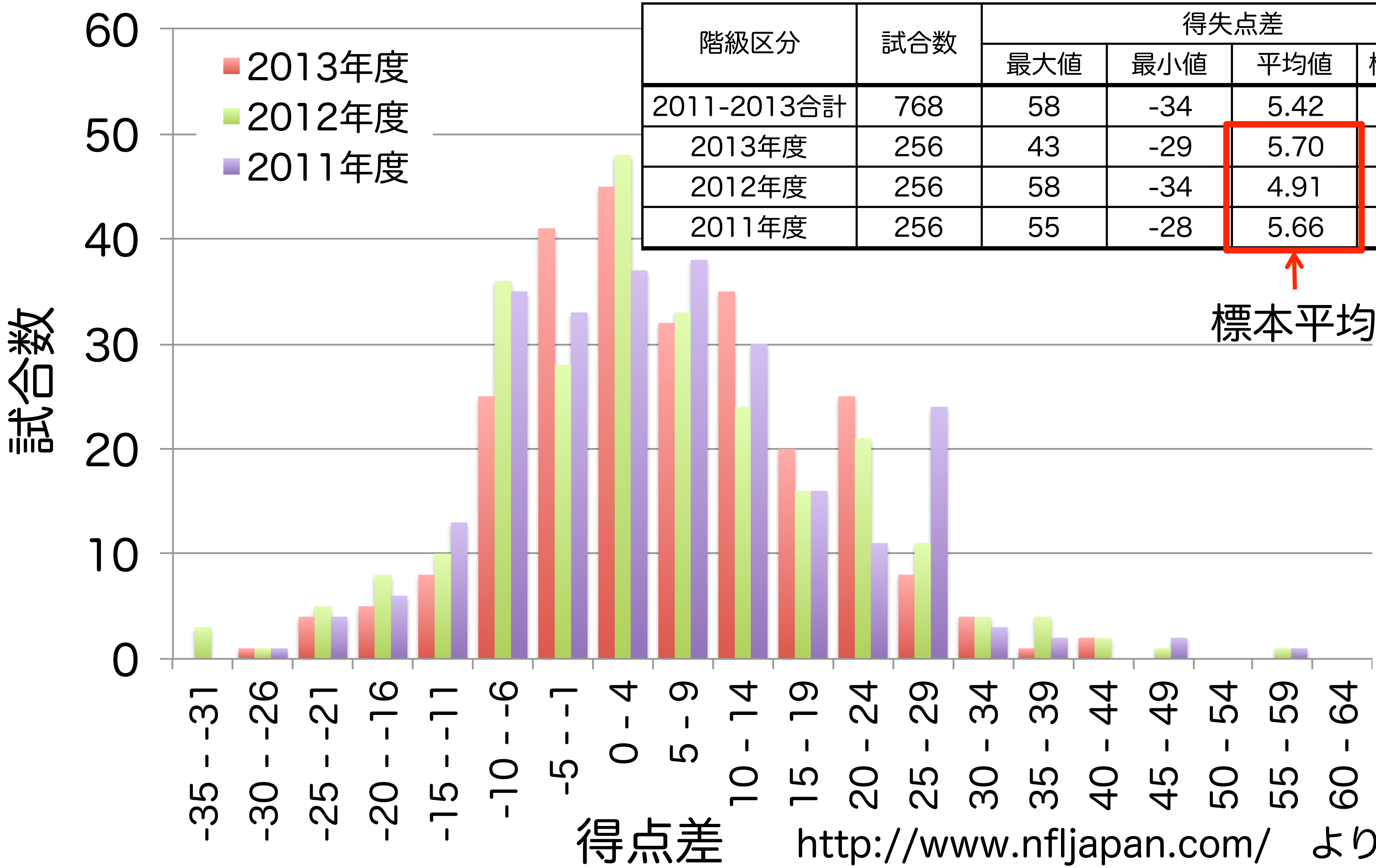


http://www.nfljapan.com/　より作成

NFL2011-2013シーズンの試合別得点差分布(H-A)

階級区分	試合数	得失点差			
		最大値	最小値	平均値	標準偏差
2011-2013合計	768	58	-34	5.42	13.54
2013年度	256	43	-29	5.70	12.40
2012年度	256	58	-34	4.91	14.44
2011年度	256	55	-28	5.66	13.74

標本平均



http://www.nfljapan.com/　より作成

試合の得失点差のベイズ推定

- 得失点差の分布はある値（平均値）を中心に左右対称の釣鐘型の分布をしているように見えます。
- このような形をする確率分布に「正規分布」があります。

正規分布

- 平均μ、分散σ^2(標準偏差σ)の正規分布N(μ, σ^2)は

$$N\left(\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right)$$

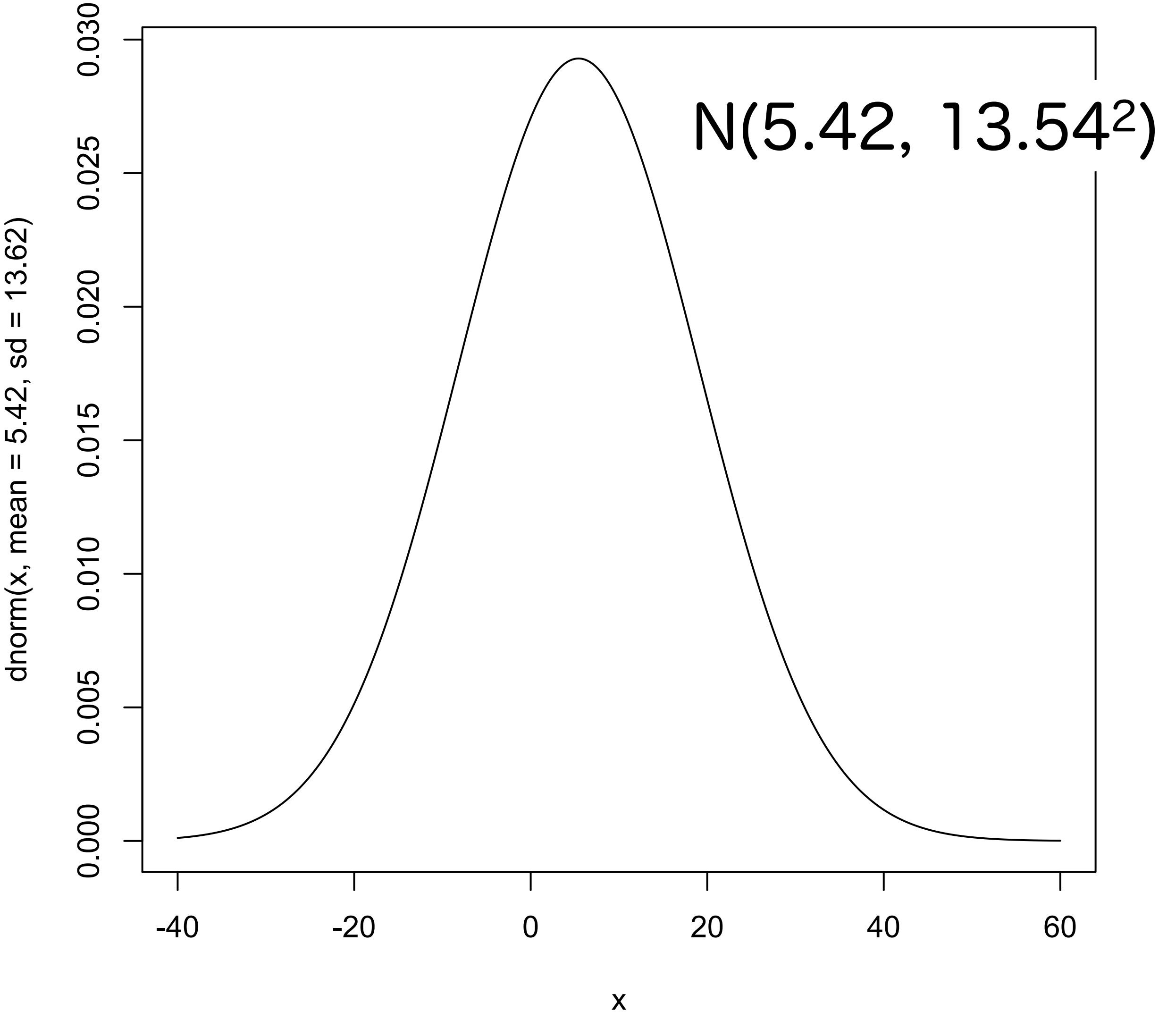
得失点差の確率分布

- 2011-2013年の３シーズン(全768試合)のデータから、平均5.42、標準偏差13.54の正規分布N(5.42, 13.54^2)は

$$N\left(5.42, 13.54^2\right)$$

$$= \frac{1}{\sqrt{2\pi \times 13.54^2}} \exp\left(-\frac{\left(x-5.42\right)^2}{2\times 13.54^2}\right)$$

得失点差の確率分布



N(5.42, 13.54²)
dnorm(x, mean = 5.42, sd = 13.62)
0.000
0.005
0.010
0.015
0.020
0.025
0.030
-40
-20
0
20
40
60
x

試合の得失点差のベイズ推定

- 得失点差に関する事前分布と尤度関数を正規分布で与える
  - 事前分布：プレシーズンマッチ5試合分
  - 尤度関数：2011-2013年の3シーズン分
- 事後分布として得失点差の平均をベイズ推定したい
  - プレシーズンマッチと2011-2013シーズンのデータから、今シーズンの攻撃力・守備力を占いたい。
  - プレシーズンマッチの試合結果から判断するのは拙速ではないか？

試合の得失点差のベイズ推定

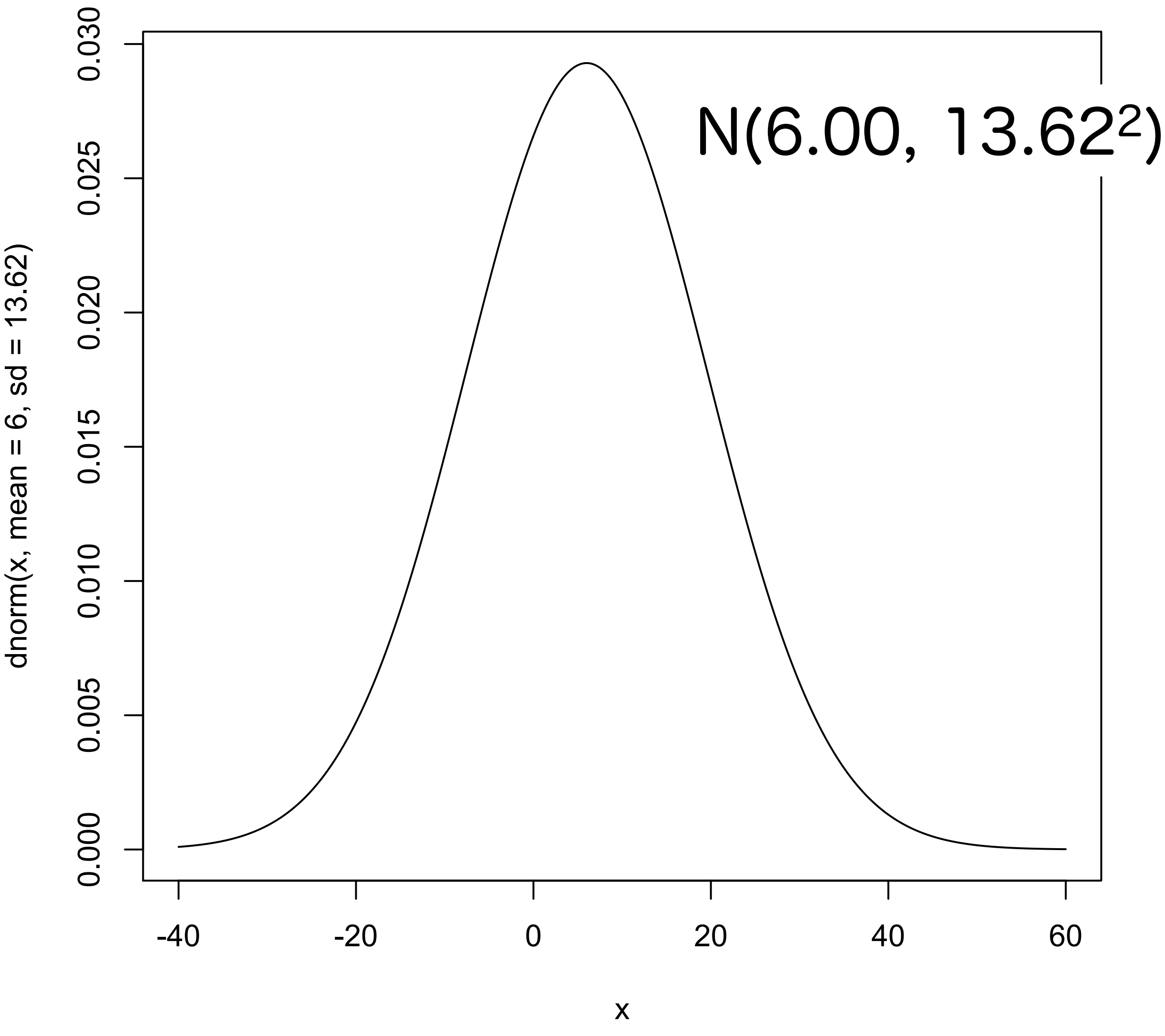
- 事前分布

$$\mu \sim N\left(6.00, 13.62^2\right)$$

$$\pi\left(\mu\right) = \frac{1}{\sqrt{2\pi} \cdot 13.62} \exp\left(-\frac{1}{2} \cdot \frac{\left(\mu - 6.00\right)^2}{3.62^2}\right)$$

$$\propto \exp\left(-\frac{1}{2} \cdot \frac{\left(\mu - 6.00\right)^2}{3.62^2}\right)$$

得失点差の確率分布



N(6.00, 13.62²)
dnorm(x, mean = 6, sd = 13.62)
0.000
0.005
0.010
0.015
0.020
0.025
0.030
-40
-20
0
20
40
60
x

試合の得失点差のベイズ推定

- 尤度関数
- 過去の全試合(母集団)から3シーズン分を標本分布として切り出したデータを使う
- 各シーズンのデータをXi (i=1～3)とする

$$X_1, X_2, X_3 \sim N\left(5.42, 13.54^2\right)$$

$$f\left(x_i \mid \mu,\right) = \frac{1}{\sqrt{2\pi}\cdot 13.54}\exp\left(-\frac{1}{2}\cdot\frac{\left(x_i - \mu\right)^2}{13.54^2}\right)$$

試合の得失点差のベイズ推定

- 尤度関数
- 過去の全試合(母集団)から3シーズン分を標本分布として切り出したデータを使う
- 各シーズンのデータをXi (i=1〜3)とする

$$X_1, X_2, X_3 \sim N\left(5.42, 13.54^2\right)$$

$$f\left(x_i \mid \mu, 13.54\right) \propto \exp\left(-\frac{1}{2} \cdot \frac{\left(x_i - \mu\right)^2}{13.54^2}\right)$$

試合の得失点差のベイズ推定

- 尤度関数
- ３シーズン分の合計⇒正規分布の積

$$\prod_{i=1}^{3} f\left(x_i \mid \mu, 13.54\right) \propto \prod_{i=1}^{3} \exp\left(-\frac{1}{2}\cdot\frac{\left(x_i-\mu\right)^2}{13.54^2}\right)$$

$$= \exp\left(-\frac{1}{2}\cdot\frac{\sum_{i=1}^{3}\left(x_i-\mu\right)^2}{13.54^2}\right)$$

試合の得失点差のベイズ推定

- 事後分布

$$\pi(\mu \mid X) \propto \exp\left(-\frac{1}{2}\cdot\frac{\sum_{i=1}^{3}\left(x_i-\mu\right)^2}{13.54^2}\right)$$

$$\cdot\exp\left(-\frac{1}{2}\cdot\frac{\left(\mu-6.00\right)^2}{13.62^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\cdot\frac{\mu'}{\sigma'^2}\right)$$

試合の得失点差のベイズ推定

- 事後分布

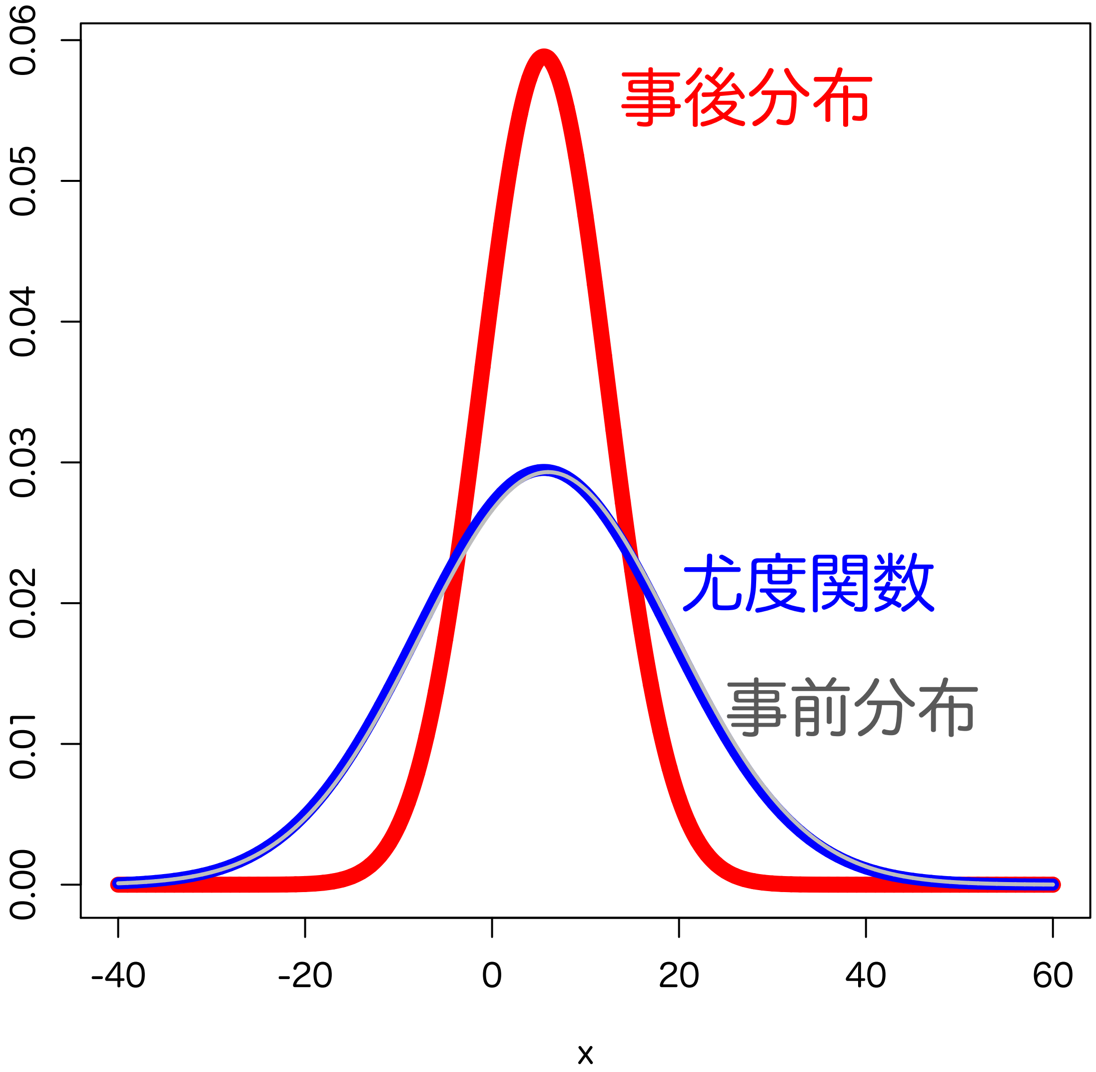
$$\pi(\mu \mid X) \propto \exp\left(-\frac{1}{2}\cdot\frac{\mu'}{\sigma'^2}\right)$$

$$\mu' = \frac{6.00/13.62^2 + 3\cdot 5.42/13.54^2}{1/13.62^2 + 3/13.54^2}$$

$$\approx 5.56$$

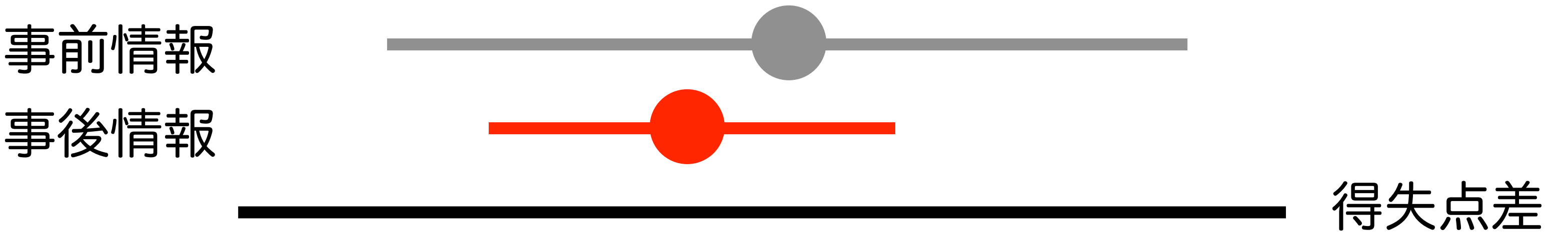
$$\frac{1}{\sigma'^2} = \frac{1}{13.62^2} + \frac{3}{13.54^2} \approx 6.77^2$$

試合の得失点差のベイズ推定



事前情報だけで判断しない理由

- 平均と分散(標準偏差)だけでなく、95%信頼区間も異なる
- 事後情報を採用するのがベイジアン



	平均	標準偏差	95%信頼区間
事前情報	6.00	13.62	[-20.7, 32.7]
事後情報	5.56	6.78	[-7.7, 18.9]

正規分布の平均のベイズ推定

- 事前分布(分散既知の正規分布) $\mu \sim N\left(\mu_0, \sigma_0^2\right)$

$$\pi\left(\mu\right) \propto \exp\left(-\frac{\left(\mu-\mu_0\right)^2}{2\sigma_0^2}\right)$$

- 尤度関数(標本分布が正規分布) $X_i \sim N\left(\mu, \sigma^2\right)$

$$\prod_{i=1}^{n} f\left(x_i \mid \mu, \sigma^2\right) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)$$

$$\propto \exp\left(-\frac{\sum_{i=1}^{n}\left(x_i-\mu\right)^2}{2\sigma^2}\right)$$

正規分布の平均のベイズ推定

- 事後分布

$$f(\mu \mid X) \propto \prod_{i=1}^{n} f\left(x_i \mid \mu, \sigma^2\right) \cdot \pi(\mu)$$

$$\propto \exp\left(-\frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{\left(\mu - \mu_0\right)^2}{2\sigma_0^2}\right)$$

$$= \exp\left(-\frac{\left(\mu - \mu_1\right)^2}{2\sigma_1^2}\right)$$

$$\mu_1 = \frac{\mu_0/\sigma_0^2 + n\bar{x}/\sigma^2}{1/\sigma_0^2 + n/\sigma^2}, \frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

正規分布の平均のベイズ推定

- 事後分布導出の過程で

$$
\begin{aligned}
\sum_{i=1}^{n}\left(x_i-\mu\right)^2 &= \sum_{i=1}^{n}\left(x_i-\bar{x}+\bar{x}-\mu\right)^2 \\
&= \sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+n\left(\bar{x}-\mu\right)^2 \\
&= \left(n-1\right)s^2+n\left(\bar{x}-\mu\right)^2
\end{aligned}
$$

↑

標本分散
＝定数

正規分布の平均のベイズ推定

クイズ

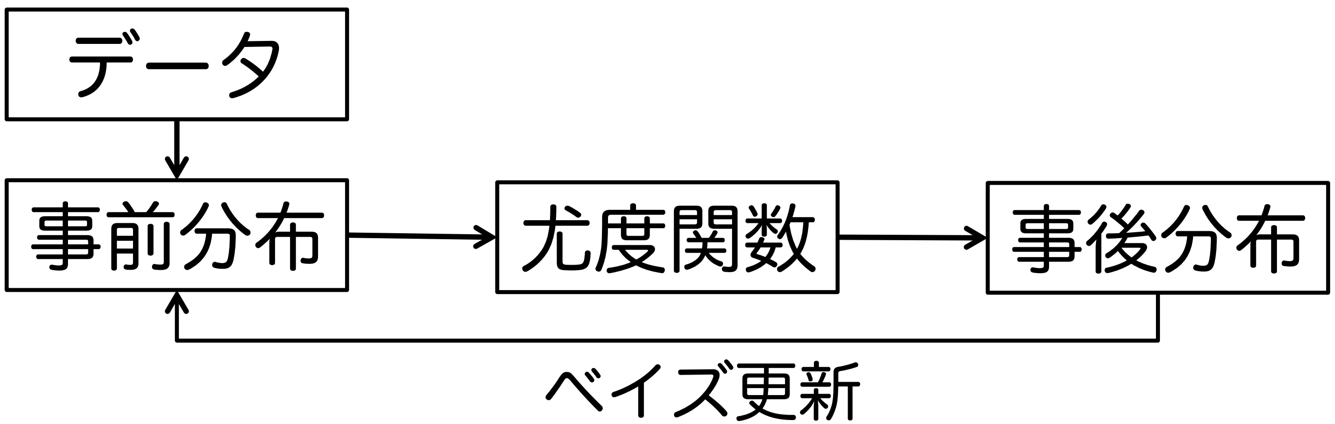
環境汚染が問題となっているある地点で、環境汚染物質の測定を5回の行ったところ、平均=3.2、分散=4.8^2であった。この環境汚染物質は、同じ時点・場所・測定器で測定しても値がばらつくことが知られている。

これまで200箇所での測定を通じて、その分布が正規分布$N(2.5, 4.5^2)$に従うことが知られている。

上述の測定地点における環境汚染物質の事後分布平均値(事後平均)と95%信頼区間を求めなさい。

経験ベイズ推定

- 事前分布をデータから与える方法を経験ベイズ(empirical Bayes)推定という
- 事前分布のパラメータを超パラメータ(hyper parameter)という
  - 事前分布の分散(精度)など



精度(precision)

- 正規分布をもちいてベイズ推定する際に、分散の逆数を精度として表すことがある

$$\tau = \frac{1}{\sigma^2}$$

- ベイズ統計では、精度がより重要である
- 分散が小さい⇔精度が高い
- 事後正規分布の精度は、事前正規分布の精度とデータの精度の和
- 事後正規分布の平均は、事前正規分布の精度とデータの精度の平均

正規分布の分散のベイズ推定

- 正規分布のベイズ推定における尤度関数

$$\prod_{i=1}^{n} f\left(x_i \mid \mu, \sigma^2\right) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}\left(x_i-\mu\right)^2}{2\sigma^2}\right)$$

- ここで $\alpha = n/2 - 1,\ \lambda = \sum_{i=1}^{n}\left(x_i-\mu\right)^2 \big/ 2$ と置き換えると、

$$\prod_{i=1}^{n} f\left(x_i \mid \mu, \sigma^2\right) \propto \left(\frac{1}{\sigma^2}\right)^{\alpha-1} \exp\left(-\lambda \cdot \frac{1}{\sigma^2}\right)$$

正規分布の分散のベイズ推定

- さらに、$1/\sigma^2$を精度τで置き換えると

$$\prod_{i=1}^{n} f\left(x_i \mid \mu, \sigma^2\right) \propto \left(\tau\right)^{\alpha-1} \exp\left(-\lambda \cdot \tau\right)$$

- これはガンマ関数Ga(α, λ)に比例する

$$Ga\left(\alpha, \lambda\right) \propto \left(\tau\right)^{\alpha-1} \exp\left(-\lambda \cdot \tau\right)$$

- すなわち事前分布の精度τに対して、自然共役事前分布としてガンマ分布を用いることと同じ意味を持つ

ガンマ関数

- ガンマ関数は次式で表される

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\,dx$$

$$\alpha > 0$$

- ガンマ関数は以下のような性質を持つ

$$\Gamma(\alpha) = (\alpha - 1)!$$

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$$

$$\Gamma(1/2) = \sqrt{\pi}$$

ガンマ分布

- ガンマ分布の確率密度関数とその性質は次式のようになる

$$f(x) = \Gamma(\alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},\ x \geq 0$$

$$E(X) = \frac{\alpha}{\lambda},\ V(X) = \frac{\alpha}{\lambda^2}$$

正規分布の分散のベイズ推定

- さらにいえば、事前分布の分散σ^2に対する自然共役事前分布として、逆ガンマ分布を用いることと同じ意味である。
- 逆ガンマ分布は、正規分布のベイズ推定にはよく用いられる

- もっとも、事前分布の分散の事前分布として逆ガンマ分布を用いず、単に分散の逆数などとして与える場合もある=無情報事前分布

逆ガンマ分布

- 逆ガンマ分布

$$\Gamma^{-1}(\alpha, \lambda) = \left(x^{\alpha-1} e^{-\lambda x}\right)^{-1}$$

- 性質
  - 平均

    $\lambda/(\alpha - 1),\ \alpha > 1$
  - 分散

    $$\frac{\lambda^2}{(\alpha-1)^2(\lambda-1)},$$

    $\alpha > 2,\ \lambda > 1$