『ベイズ推論による機械学習入門』のノート Chapter 3 ベイズ推論による学習と予測

*@anemptyarchive**

2020/02/11-2020/03/06

Contents

*https://www.anarchive-beta.com/

Chapter 3 ベイズ推論による学習と予測

3.1 学習と予測

この章では、確率推論を用いたパラメータの学習と未観測の値の予測について説明する。確率計算によってデータを観測した後のパラメータの事後分布を求めることを学習と呼ぶ。また、未観測値の予測分布についても確率推論を使って求める。

3.1.1 パラメータの事後分布

訓練データ集合を $\mathcal{D}$ としたとき、ベイズ学習では未知のパラメータ θ との同時分布 $p(\mathcal{D}, \theta)$ を考えることによってデータを表現するモデルを構築する。

$$p(\mathcal{D}, \theta) = p(\mathcal{D}|\theta)p(\theta) \tag{3.1}$$

パラメータ θ に関する事前の不確実性を、事前分布 $p(\theta)$ を設定することによってモデルに反映させている。また、特定のパラメータ θ からどのようにしてデータ $\mathcal{D}$ が発生しているのかを尤度 $p(\mathcal{D}|\theta)$ で示している。これを θ の関数と見たとき尤度関数とも呼ばれる。また観測データ $\mathcal{D}$ の発生過程を表しているため、観測モデルとも呼ぶ。

観測したデータ $\mathcal{D}$ を用いて、パラメータの不確実性 (θ の分布) をベイズの定理を用いて更新する。

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{3.2}$$

更新後の観測データ $\mathcal{D}$ を条件とした分布 $p(\theta|\mathcal{D})$ を事後分布と呼ぶ。事後分布 $p(\theta|\mathcal{D})$ は事前分布 $p(\theta)$ と比べて、尤度関数 $p(\mathcal{D}|\theta)$ を通すことによって、観測データ $\mathcal{D}$ に関する特徴を捉えていることが規定できる。

このように事後分布を求めることがベイズ学習の枠組みにおける「学習」に当たる。

3.1.2 予測分布

更に、学習されたパラメータの分布 (事後分布) を使って、未観測のデータ x_* の予測分布 $p(x_*|\mathcal{D})$ を求める。

パラメータ θ について周辺化 (積分消去) することで、x_* の予測分布となる。

$$p(x_*|\mathcal{D}) = \int p(x_*|\theta)p(\theta|\mathcal{D})d\theta \tag{3.3}$$

$\mathcal{D},\ x_*$ は、共にパラメータ θ に従って生成されるとモデル化している。そのため、$\mathcal{D}$ と x_* は、パラメータ θ を与えられた下での条件付き独立と言える。

$$p(\mathcal{D}, x_*|\theta) = p(\mathcal{D}|\theta)p(x_*|\theta)$$

また、パラメータ θ も含めた同時分布は

$$
\begin{aligned}
p(\mathcal{D}, x_*, \theta) &= p(\mathcal{D}, x_*|\theta)p(\theta) \\
&= p(\mathcal{D}|\theta)p(x_*|\theta)p(\theta)
\end{aligned} \tag{3.4}
$$

と構成される。

データ $\mathcal{D}$ だけが手元にあるとき、残りの変数の事後分布は

$$
\begin{aligned}
p(x_*, \theta|\mathcal{D}) &= \frac{p(\mathcal{D}, x_*, \theta)}{p(\mathcal{D})} \\
&= \frac{p(\mathcal{D}|\theta)p(x_*|\theta)p(\theta)}{p(\mathcal{D})} \\
&= p(x_*|\theta)p(\theta|\mathcal{D})
\end{aligned} \tag{3.5}
$$

【途中式の途中式】

0. 同時分布 $p(\mathcal{D}, x_*, \theta)$ に対してベイズの定理を用いることで、事後分布 (左辺) が求まる。
1. 式 (3.4) より同時分布 $p(\mathcal{D}, x_*, \theta)$ を分解する。
2. 式 (3.2) より分布 $p(\theta|\mathcal{D})$ となる。

となる。

この事後分布 (3.5) から θ を積分消去した周辺分布が予測分布 (3.3) であることが分かる。

3.1.3 共役事前分布

特になし。

3.1.4 共役でない事前分布の利用

特になし。

3.2 離散確率分布の学習と予測

パラメータの事後分布と予測分布を解析的に計算する方法を解説する。

3.2.1 ベルヌーイ分布の学習と予測

・事後分布の計算

ベルヌーイ分布に従うと仮定する N 個の観測データ $\boldsymbol{X} = \{x_1, x_2, \cdots, x_N\}$ を用いて、パラメータ μ の事後分布を求めていく。

まずは観測モデル $p(\boldsymbol{X}|\mu)$ について確認する。離散値データ $\boldsymbol{X} = \{x_1, x_2, \cdots, x_N\}$ はそれぞれ独立に発生しているとの仮定の下で、観測モデルは

$$
\begin{aligned}
p(\boldsymbol{X}|\mu) &= p(x_1, x_2, \cdots, x_N|\mu) \\
&= p(x_1|\mu)p(x_2|\mu)\cdots p(x_N|\mu) \\
&= \prod_{n=1}^{N} p(x_n|\mu)
\end{aligned}
$$

と分解できる。

これを用いて、μ の事後分布 $p(\mu|\boldsymbol{X})$ はベイズの定理より

$$
\begin{aligned}
p(\mu|\boldsymbol{X}) &= \frac{p(\boldsymbol{X}|\mu)p(\mu)}{p(\boldsymbol{X})} \\
&= \frac{\left\{\prod_{n=1}^{N} p(x_n|\mu)\right\} p(\mu)}{p(\boldsymbol{X})} \\
&= \frac{\left\{\prod_{n=1}^{N} \mathrm{Bern}(x_n|\mu)\right\} \mathrm{Beta}(\mu|a,b)}{p(\boldsymbol{X})} \\
&\propto \left\{\prod_{n=1}^{N} \mathrm{Bern}(x_n|\mu)\right\} \mathrm{Beta}(\mu|a,b) \qquad (3.12)
\end{aligned}
$$

となる。分母の $p(\boldsymbol{X})$ は μ に影響しないため省略して、比例関係にのみ注目する。省略した部分については、最後に正規化することで対応できる。

次にこの分布の具体的な形状を明らかにしていく。対数をとって指数部分の計算を分かりやすくして進めると

$$
\begin{aligned}
\ln p(\mu|\boldsymbol{X}) &= \ln \prod_{n=1}^{N} \mathrm{Bern}(x_n|\mu) + \ln \mathrm{Beta}(\mu|a,b) - \ln p(\boldsymbol{X}) \\
&= \sum_{n=1}^{N} \ln \mathrm{Bern}(x_n|\mu) + \ln \mathrm{Beta}(\mu|a,b) + \mathrm{const.} \\
&= \sum_{n=1}^{N} \ln \left\{\mu^{x_n}(1-\mu)^{1-x_n}\right\} + \ln \left\{C_B(a,b)\mu^{a-1}(1-\mu)^{b-1}\right\} + \mathrm{const.} \\
&= \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\} + \ln C_B(a,b) + (a-1)\ln \mu + (b-1)\ln(1-\mu) + \mathrm{const.} \\
&= \sum_{n=1}^{N} x_n \ln \mu + \left(N - \sum_{n=1}^{N} x_n\right) \ln(1-\mu) + (a-1)\ln \mu + (b-1)\ln(1-\mu) + \mathrm{const.} \\
&= \left(\sum_{n=1}^{N} x_n + a - 1\right) \ln \mu + \left(N - \sum_{n=1}^{N} x_n + b - 1\right) \ln(1-\mu) + \mathrm{const.} \qquad (3.13)
\end{aligned}
$$

【途中式の途中式】

0. 式 (3.12) の下から 2 行目の式を用いている。
1. μ に影響しない項 (ここでは $-\ln p(\boldsymbol{X})$) を const. とおく。

総乗部分の対数をとると

$$\begin{aligned}\ln p(\boldsymbol{X}|\mu) &= \ln \prod_{n=1}^{N} p(x_n|\mu) \\ &= \ln\Big\{p(x_1|\mu) * p(x_2|\mu) * \cdots * p(x_N|\mu)\Big\} \\ &= \ln p(x_1|\mu) + \ln p(x_2|\mu) + \cdots + \ln p(x_N|\mu) \\ &= \sum_{n=1}^{N} \ln p(x_n|\mu)\end{aligned}$$

のように総和になる。

2. $\mathrm{Bern}(x_n|\mu)$, $\mathrm{Beta}(\mu|a,b)$ に、それぞれベルヌーイ分布の定義式 (2.16)、ベータ分布の定義式 (2.41) を用いて具体的な確率分布を代入する。
3. それぞれ対数をとる。$\ln(x^a y^b) = a \ln x + b \ln y$ である。
4. $\sum_{n=1}^{N}$ に関する括弧を展開し、また μ に影響しない項 ($\ln C_B(a,b)$) を const. にまとめる。
5. μ と $(1-\mu)$ の項をまとめて式を整理する。

となる。

式 (3.13) について

$$\begin{aligned}\hat{a} &= \sum_{n=1}^{N} x_n + a \\ \hat{b} &= N - \sum_{n=1}^{N} x_n + b \end{aligned} \tag{3.15}$$

とおき

$$\ln p(\mu|\boldsymbol{X}) = (\hat{a} - 1)\ln \mu + (\hat{b} - 1)\ln(1-\mu) + \text{const.}$$

更に ln を外して、const. を正規化項に置き換える (正規化する) と

$$p(\mu|\boldsymbol{X}) = C_B(\hat{a}, \hat{b})\mu^{\hat{a}-1}(1-\mu)^{\hat{b}-1} = \mathrm{Beta}(\mu|\hat{a}, \hat{b}) \tag{3.14}$$

となる。式の形状から、事後分布 $p(\mu|\boldsymbol{X})$ がパラメータ $\hat{a}$, $\hat{b}$ を持つベータ分布であることを確認できる。
　また事後分布のパラメータの計算式 (3.15) によって、ハイパーパラメータ a, b が更新される。

・予測分布の計算

続いてベルヌーイ分布に従う未観測のデータ x_* に対する予測分布を求めていく。

(観測データによる学習を行っていない) 事前分布 $p(\mu)$ を用いて、パラメータ μ を周辺化することで予測分布 $p(x_*)$ となる。

$$
\begin{aligned}
p(x_*) &= \int p(x_*|\mu)p(\mu)d\mu \\
&= \int \mathrm{Bern}(x_*|\mu)\mathrm{Beta}(\mu|a,b)d\mu \\
&= \int \mu^{x_*}(1-\mu)^{1-x_*}C_B(a,b)\mu^{a-1}(1-\mu)^{b-1}d\mu \\
&= C_B(a,b)\int \mu^{x_*+a-1}(1-\mu)^{1-x_*+b-1}d\mu
\end{aligned} \tag{3.16}
$$

積分部分に注目すると、パラメータ $x_*+a,\ 1-x_*+b$ を持つ正規化項のないベータ分布の形をしていることが分かる。よって、ベータ分布の定義式 (2.41) を用いて

$$
\int \mu^{x_*+a-1}(1-\mu)^{1-x_*}+b-1d\mu = \frac{1}{C_B(x_*+a, 1-x_*+b)} \tag{3.17}
$$

正規化項の逆数に変形できる (そもそもこの部分の逆数が正規化項である)。これを式 (3.16) に代入すると、予測分布 $p(x_*)$ は

$$
p(x_*) = \frac{C_B(a,b)}{C_B(x_*+a, x_*+b)}
$$

となる。更にベータ分布の正規化項 (2.42) を用いて、式を整理すると

$$
\begin{aligned}
p(x_*) &= C_B(a,b)\frac{1}{C_B(x_*+a, x_*+b)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(x_*+a)\Gamma(1-x_*+b)}{\Gamma(x_*+a+1-x_*+b)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(x_*+a)\Gamma(1-x_*+b)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(x_*+a)\Gamma(1-x_*+b)}{(a+b)\Gamma(a+b)} \\
&= \frac{\Gamma(x_*+a)\Gamma(1-x_*+b)}{\Gamma(a)\Gamma(b)(a+b)}
\end{aligned} \tag{3.18}
$$

となる。ガンマ関数の性質 $\Gamma(x+1) = x\Gamma(x)$ を用いている。

x_* は 0 か 1 しかとらないため、場合分けしてこの式を整理する。$x_* = 1$ のとき

$$
\begin{aligned}
p(x_* = 1) &= \frac{\Gamma(1+a)\Gamma(1-1+b)}{\Gamma(a)\Gamma(b)(a+b)} && (3.18) \\
&= \frac{a\Gamma(a)\Gamma(b)}{\Gamma(a)\Gamma(b)(a+b)} \\
&= \frac{a}{a+b} && (3.19)
\end{aligned}
$$

となる。また $x_* = 0$ のとき

$$
\begin{aligned}
p(x_* = 0) &= \frac{\Gamma(0+a)\Gamma(1-0+b)}{\Gamma(a)\Gamma(b)(a+b)} && (3.18) \\
&= \frac{\Gamma(a)b\Gamma(b)}{\Gamma(a)\Gamma(b)(a+b)} \\
&= \frac{b}{a+b} && (3.20)
\end{aligned}
$$

となる。これは

$$
\frac{b}{a+b} = 1 - \frac{a}{a+b}
$$

と置き換えることができる。

従って、$x^0 = 1$ であることを利用して式 (3.19) と式 (3.20) をまとめると、式 (3.18) は

$$
\begin{aligned}
p(x_*) &= \left(\frac{a}{a+b}\right)^{x_*} \left(\frac{b}{a+b}\right)^{1-x_*} \\
&= \left(\frac{a}{a+b}\right)^{x_*} \left(1 - \frac{a}{a+b}\right)^{1-x_*}
\end{aligned}
$$

と書き換えることができる。($x_* = 0$ のとき前の項が 1 となり、$x_* = 1$ のとき後ろの項が 1 となる。)

この式について

$$
\mu_* = \frac{a}{a+b}
$$

とおくと、予測分布 $p(x_*)$ は

$$
p(x_*) = \mu_*^{x_*}(1-\mu_*)^{1-x_*} = \mathrm{Bern}(x_*|\mu_*) \tag{3.21}
$$

となる。式の形状から、$p(x_*)$ はパラメータ μ_* を持つベルヌーイ分布であることが分かる。
ちなみに $\frac{a}{a+b}$ は、ベータ分布 $\mathrm{Beta}(\mu|a,b)$ の期待値 $\mathbb{E}[\mu]$ である。

予測分布のパラメータ μ_* を構成する $a,\ b$ について、それぞれ事後分布のパラメータ (3.15) を用いると

$$
\begin{aligned}
\hat{\mu}_* &= \frac{\hat{a}}{\hat{a} + \hat{b}} \\
&= \frac{\sum_{n=1}^N x_n + a}{\sum_{n=1}^N x_n + a + N - \sum_{n=1}^N x_n + b} \\
&= \frac{\sum_{n=1}^N x_n + a}{N + a + b}
\end{aligned}
$$

となり、(観測データ $\boldsymbol{X}$ によって学習した) 事後分布 $p(\mu|\boldsymbol{X})$ を用いた予測分布 $p(x_*|\boldsymbol{X})$

$$
p(x_*|\boldsymbol{X}) = \hat{\mu}_*^{x_*}(1 - \hat{\mu}_*)^{1-x_*} = \mathrm{Bern}(x_*|\hat{\mu}_*) \tag{3.22}
$$

が得られる。

・R でやってみよう

各確率分布に従いランダムに生成したデータを用いて、パラメータを推定してみましょう。

```r
# 利用パッケージ
library(tidyverse)
```

利用するパッケージを読み込みます。

・事後分布

```r
## パラメータの初期値を指定
# 観測モデルのパラメータ
mu_truth <- 0.25

# 事前分布のパラメータ
a <- 1
b <- 1

# 試行回数
N <- 50
```

$x_n = 1$ となる確率 μ を `mu_truth` とします。この値を推定するのが目的です。

事前分布のパラメータ (ハイパーパラメータの初期値)a, b をそれぞれ `a`, `b` とします。

生成するデータ数 N を `N` とします。

```r
# ベルヌーイ分布に従うデータを生成
x_n <- rbinom(n = N, size = 1, prob = mu_truth)
```

二項分布に従う乱数を発生させる関数 `rbinom()` を使って、ランダムにデータを生成します。これをベルヌーイ分布とするには、`size` 引数に 1 を指定します。

試行回数の引数 `n` には `N`、確率の引数 `prob` には `mu_truth` を指定します。

サンプルを確認してみましょう。

```
# 観測データを確認
table(x_n)
```

```
## x_n
##  0  1
## 37 13
```

このデータを用いて事後分布のパラメータを計算します。

```
# 事後分布のパラメータを計算
a_hat <- sum(x_n) + a
b_hat <- N - sum(x_n) + b
```

事後分布のパラメータを計算式 (3.15) の計算を行い、$\hat{a}, \hat{b}$ をそれぞれ `a_hat`, `b_hat` とします。

```
# 事後分布を計算
posterior_df <- tibble(
  mu = seq(0, 1, by = 0.001),  # 作図用の値
  C_B = lgamma(a_hat + b_hat) - lgamma(a_hat) - lgamma(b_hat),  # 正規化項（対数）
  density = exp(C_B + (a_hat - 1) * log(mu) + (b_hat - 1) * log(1 - mu)) # 確率密度
)
```

μ が取り得る 0 から 1 までの値を `seq(0, 1)` を使って用意します。`by` 引数で刻み幅を指定できるので、グラフが粗かったり処理が重くなったりする場合はこの値を調整してください。

`mu` の各値に対してベータ分布の定義式 (2.41) の計算を行い、確率密度を求めます。ただし値が大きくなると `gamma()` で計算できなくなるため、対数をとって計算することにします。なので最後に `exp()` で値を戻します。

計算結果は作図用にデータフレームにまとめておきます。推定結果を確認してみましょう。

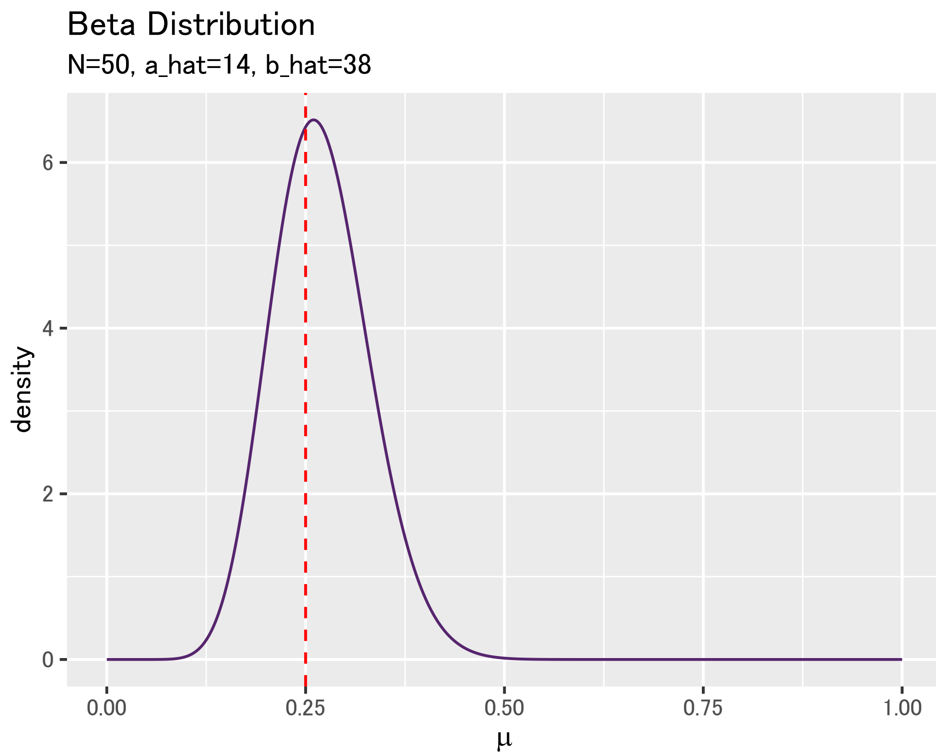
```
head(posterior_df)
```

```
## # A tibble: 6 x 3
##      mu   C_B  density
##   <dbl> <dbl>    <dbl>
## 1 0      30.5 0.
## 2 0.001  30.5 1.74e-26
## 3 0.002  30.5 1.38e-22
## 4 0.003  30.5 2.58e-20
## 5 0.004  30.5 1.05e-18
## 6 0.005  30.5 1.84e-17
```

このデータフレームを用いてグラフを描きます。

```
# 作図
ggplot(posterior_df, aes(mu, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = mu_truth),
```

```
              color = "red", linetype = "dashed") + # 垂直線
  labs(title = "Beta Distribution",
       subtitle = paste0("N=", N, ", a_hat=", a_hat, ", b_hat=", b_hat),
       x = expression(mu)) # ラベル
```



ggplot2 パッケージを利用してグラフを作成します。

推定値は折れ線グラフ geom_line() を用いて描きます。
パラメータ μ の真の値を垂直線 geom_vline() で示します。

・予測分布

```
# 予測分布のパラメーターを計算
mu_hat <- a_hat / (a_hat + b_hat)
```

予測分布のパラメータを計算式の計算を行い、$\hat{\mu}$ を mu_hat とします。

```
mu_hat <- (sum(x_n) + a) / (N + a + b)
```

このように、観測データ x_n と事前分布のパラメータ a, b を使って計算することもできます。

```
# 予測分布を計算
predict_df <- tibble(
  x = c(0, 1),  # 作図用の値
  prob = mu_hat^x * (1 - mu_hat)^(1 - x) # 確率
)
```

x_* が取り得る 0 と 1 の値を用意します。
x の各値となる確率をベルヌーイ分布の定義式 (2.16) で計算します。

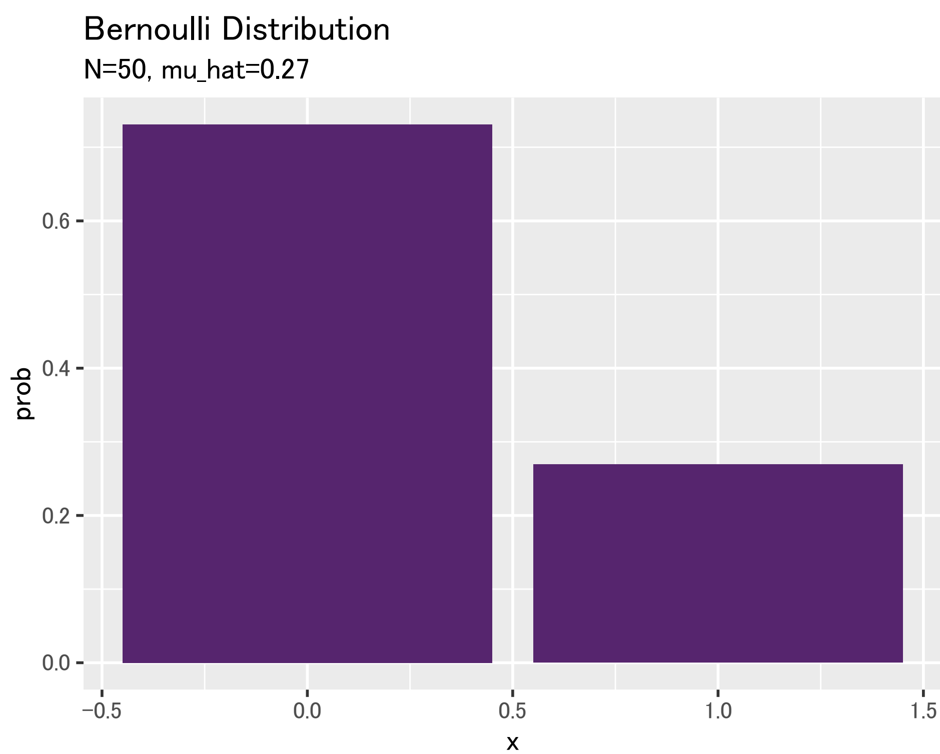
推定結果は次のようになります。

```r
head(predict_df)
```

```
## # A tibble: 2 x 2
##       x  prob
##   <dbl> <dbl>
## 1     0 0.731
## 2     1 0.269
```

こちらも ggplot2 パッケージを利用して作図します。

```r
# 作図
ggplot(predict_df, aes(x, prob)) +
  geom_bar(stat = "identity", position = "dodge", fill = "#56256E") + # 棒グラフ
  labs(title = "Bernoulli Distribution",
       subtitle = paste0("N=", N, ", mu_hat=", round(mu_hat, 2))) # ラベル
```



棒グラフは geom_bar() を使います。

・おまけ

gganimate パッケージを利用して、パラメータの推定値の推移の gif 画像を作成するためのコードです。

```r
# 利用パッケージ
library(tidyverse)
library(gganimate)

## パラメータの初期値を指定
# 観測モデルのパラメータ
mu_truth <- 0.25

# 事前分布のパラメータ
a <- 1
b <- 1

# 試行回数
N <- 50

# 事前分布を計算
posterior_df <- tibble(
  mu = seq(0, 1, by = 0.001),  # 作図用の値
  C_B = lgamma(a + b) - lgamma(a) - lgamma(b),  # 正規化項（対数）
  density = exp(C_B + (a - 1) * log(mu) + (b - 1) * log(1 - mu)), # 確率密度
  N = 0  # 試行回数
)

# 初期値による予測分布のパラメーターを計算
mu_star <- a / (a + b)

# 初期値による予測分布を計算
predict_df <- tibble(
  x = c(0, 1),  # 作図用の値
  prob = mu_star^x * (1 - mu_star)^(1 - x),  # 確率
  N = 0 # 試行回数
)

# パラメータを推定
x_n <- rep(0, N) # 受け皿
for(n in 1:N){

  # ベルヌーイ分布に従うデータを生成
  x_n[n] <- rbinom(n = 1, size = 1, prob = mu_truth)

  # ハイパーパラメータを更新
  a <- x_n[n] + a
  b <- 1 - x_n[n] + b

  # 事後分布を計算
  tmp_posterior_df <- tibble(
    mu = seq(0, 1, by = 0.001),  # 作図用の mu
    C_B = lgamma(a + b) - lgamma(a) - lgamma(b),  # 正規化項（対数）
    density = exp(C_B + (a - 1) * log(mu) + (b - 1) * log(1 - mu)), # 確率密度
    N = n  # 試行回数
  )

  # 予測分布のパラメーターを更新
```

```
  mu_star <- a / (a + b)

  # 予測分布を計算
  tmp_predict_df <- tibble(
    x = c(0, 1),  # 作図用の値
    prob = mu_star^x * (1 - mu_star)^(1 - x),  # 確率
    N = n # 試行回数
  )

  # 推定結果を結合
  posterior_df <- rbind(posterior_df, tmp_posterior_df)
  predict_df <- rbind(predict_df, tmp_predict_df)
}

# 観測データを確認
table(x_n)

## 事後分布
# 作図
posterior_graph <- ggplot(posterior_df, aes(mu, density)) +
  geom_line(color = "#56256E") +  # 折れ線グラフ
  geom_vline(aes(xintercept = mu_truth),
             color = "red", linetype = "dashed") + # 垂直線
  transition_manual(N) + # フレーム
  labs(title = "Beta Distribution",
       subtitle = "N= {current_frame}",
       x = expression(mu)) # ラベル

# 描画
animate(posterior_graph)

## 予測分布
# 作図
predict_graph <- ggplot(predict_df, aes(x, prob)) +
  geom_bar(stat = "identity", position = "dodge", fill = "#56256E") + # 棒グラフ
  transition_manual(N) + # フレーム
  labs(title = "Bernoulli Distribution",
       subtitle = "N= {current_frame}") # ラベル

# 描画
animate(predict_graph)
```

異なる点のみを簡単に解説します。

観測された各データによってどのように学習する (推定値 (の分布) が変化する) のかを確認するため、こちらは for ループで 1 データずつ処理します。

よって生成するデータ数として設定した `N` がイタレーション数になります。

パラメータの推定値について、$\hat{a}, \hat{b}$ に対応する `a_hat`, `b_hat` を新たに作るのではなく、`a`, `b` をイタレーションごとに更新 (上書き) していきます。

それに伴い、事後分布のパラメータの計算式 (3.15) の $\sum_{n=1}^{N}$ の計算は、for ループによって N 回繰り返し `x_n[n]` を加えることで行います。n 回目のループ処理のときには、n-1 回分の `x_n[n]`(`b` の場合は 1) が

既に a と b に加えられているわけです。

3.2.2 カテゴリ分布の学習と予測

・事後分布の計算

カテゴリ分布に従うと仮定する N 個の観測データ $\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N\}$ を用いて、パラメータ $\boldsymbol{\pi}$ の事後分布を求めていく。

まずは観測モデル $p(\boldsymbol{S}|\boldsymbol{\pi})$ について確認する。離散値データ $\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N\}$ はそれぞれ独立に生成されているとの仮定の下で、観測モデルは

$$\begin{aligned}
p(\boldsymbol{S}|\boldsymbol{\pi}) &= p(\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N|\boldsymbol{\pi}) \\
&= p(\boldsymbol{s}_1|\boldsymbol{\pi})p(\boldsymbol{s}_2|\boldsymbol{\pi})\cdots p(\boldsymbol{s}_N|\boldsymbol{\pi}) \\
&= \prod_{n=1}^{N} p(\boldsymbol{s}_n|\boldsymbol{\pi})
\end{aligned}$$

と分解できる。更に $\boldsymbol{s}_n$, $\boldsymbol{\pi}$ は K 次元ベクトルであるため

$$\begin{aligned}
p(\boldsymbol{s}_n|\boldsymbol{\pi}) &= p(s_{n,1}, s_{n,2}, \cdots, s_{n,K}|\boldsymbol{\pi}) \\
&= p(s_{n,1}|\pi_1)p(s_{n,2}|\pi_2)\cdots p(s_{n,K}|\pi_K) \\
&= \prod_{k=1}^{K} p(s_{n,k}|\pi_k)
\end{aligned}$$

と分解できる。従って観測モデルは

$$p(\boldsymbol{S}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(s_{n,k}|\pi_k)$$

である。

これを用いて、$\boldsymbol{\pi}$ の事後分布 $p(\boldsymbol{\pi}|\boldsymbol{S})$ はベイズの定理より

$$\begin{aligned}
p(\boldsymbol{\pi}|\boldsymbol{S}) &= \frac{p(\boldsymbol{S}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{\alpha})}{p(\boldsymbol{S})} \\
&= \frac{\left\{\prod_{n=1}^{N} p(\boldsymbol{s}_n|\boldsymbol{\pi})\right\} p(\boldsymbol{\pi}|\boldsymbol{\alpha})}{p(\boldsymbol{S})} \\
&= \frac{\left\{\prod_{n=1}^{N} \mathrm{Cat}(\boldsymbol{s}_n|\boldsymbol{\pi})\right\} \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})}{p(\boldsymbol{S})} \\
&\propto \left\{\prod_{n=1}^{N} \mathrm{Cat}(\boldsymbol{s}_n|\boldsymbol{\pi})\right\} \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})
\end{aligned} \tag{3.25}$$

となる。分母の $p(\boldsymbol{S})$ は $\boldsymbol{\pi}$ に影響しないため省略して、比例関係にのみ注目する。省略した部分については、最後に正規化することで対応できる。

次にこの分布の具体的な形状を明らかにしていく。対数をとって指数部分の計算を分かりやすくして進めると

$$
\begin{aligned}
\ln p(\boldsymbol{\pi}|\boldsymbol{S}) &= \ln \prod_{n=1}^{N} \mathrm{Cat}(\boldsymbol{s}_n|\boldsymbol{\pi}) + \ln \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) - \ln p(\boldsymbol{S}) \\
&= \sum_{n=1}^{N} \ln \mathrm{Cat}(\boldsymbol{s}_n|\boldsymbol{\pi}) + \ln \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) + \mathrm{const.} \\
&= \sum_{n=1}^{N} \ln \left(\prod_{k=1}^{K} \pi_k^{s_{n,k}} \right) + \ln \left\{ C_D(\boldsymbol{\alpha}) \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \right\} + \mathrm{const.} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} s_{n,k} \ln \pi_k + \ln C_D(\boldsymbol{\alpha}) + \sum_{k=1}^{K} (\alpha_k - 1) \ln \pi_k + \mathrm{const.} \\
&= \sum_{k=1}^{K} \left(\sum_{n=1}^{N} s_{n,k} + \alpha_k - 1 \right) \ln \pi_k + \mathrm{const.}
\end{aligned}
\tag{3.26}
$$

【途中式の途中式】

0. 式 (3.25) の下から 2 行目の式を用いている。
1. 対数をとると積が和に変わる。また、$\boldsymbol{\pi}$ に影響しない項 $(-\ln p(\boldsymbol{S}))$ を const. とおく。
2. $\mathrm{Cat}(\boldsymbol{s}_n|\boldsymbol{\pi})$, $\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$ に、それぞれカテゴリ分布の定義式 (2.29)、ディリクレ分布の定義式 (2.48) を用いて具体的な確率分布を代入する。
3. それぞれ対数をとる。$\ln(x^a y^b) = a \ln x + b \ln y$ である。
4. $\boldsymbol{\pi}$ に影響しない項 $(\ln C_D(\boldsymbol{\alpha}))$ を const. にまとめる。
5. $\ln \pi_k$ の項をまとめて式を整理する。

となる。

式 (3.26) について

$$
\hat{\alpha}_k = \sum_{n=1}^{N} s_{n,k} + \alpha_k \tag{3.28}
$$

とおき

$$
\ln p(\boldsymbol{\pi}|\boldsymbol{S}) = \sum_{k=1}^{K} (\hat{\alpha}_k - 1) \ln \pi_k + \mathrm{const.}
$$

更に ln を外して、const. を正規化項に置き換える (正規化する) と

$$
p(\boldsymbol{\pi}|\boldsymbol{S}) = C_D(\hat{\boldsymbol{\alpha}}) \prod_{k=1}^{K} \pi_k^{\hat{\alpha}_k - 1} = \mathrm{Dir}(\boldsymbol{\pi}|\hat{\boldsymbol{\alpha}}) \tag{3.27}
$$

となる。式の形状から、事後分布 $p(\boldsymbol{\pi}|\boldsymbol{S})$ がパラメータ $\hat{\boldsymbol{\alpha}}$ を持つディリクレ分布であることが確認できる。

また事後分布のパラメータの計算式 (3.28) によって、ハイパーパラメータ $\boldsymbol{\alpha}$ が更新される。

・予測分布の計算

続いて未観測のデータ $\boldsymbol{s}_* = \{s_{*,1}, s_{*,2}, \cdots, s_{*,K}\}$ に対する予測分布を求めていく。

(観測データによる学習を行っていない) 事前分布 $p(\boldsymbol{\pi}|\boldsymbol{\alpha})$ を用いて、パラメータ $\boldsymbol{\pi}$ を周辺化することで予測分布 $p(\boldsymbol{s}_*)$ となる。

$$
\begin{aligned}
p(\boldsymbol{s}_*) &= \int p(\boldsymbol{s}_*|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{\alpha})d\boldsymbol{\pi} \\
&= \int \mathrm{Cat}(\boldsymbol{s}_*|\boldsymbol{\pi})\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})d\boldsymbol{\pi} \\
&= \int \prod_{k=1}^{K} \pi_k^{s_{*,k}} C_D(\boldsymbol{\alpha})\pi_k^{\alpha_k-1}d\boldsymbol{\pi} \\
&= C_D(\boldsymbol{\alpha}) \int \prod_{k=1}^{K} \pi_k^{s_{*,k}+\alpha_k-1}d\boldsymbol{\pi}
\end{aligned}
$$

積分部分に注目すると、パラメータ $s_{*,k} + \alpha_k$ を持つ正規化項のないディリクレ分布の形をしている。よって、ディリクレ分布の定義式 (2.48) を用いて

$$
\int \prod_{k=1}^{K} \pi_k^{s_{*,k}+\alpha_k-1}d\boldsymbol{\pi} = \frac{1}{C_D((s_{*,k}+\alpha_k)_{k=1}^{K})}
$$

正規化項の逆数に変形できる (そもそもこの部分の逆数が正規化項である)。これを先ほどの式に代入すると、予測分布 $p(\boldsymbol{s}_*)$ は

$$
p(\boldsymbol{s}_*) = \frac{C_D(\boldsymbol{\alpha})}{C_D\Big((s_{*,k}+\alpha_k)_{k=1}^{K}\Big)} \tag{3.29}
$$

となる。更にディリクレ分布の正規化項 (2.49) を用いると

$$
\begin{aligned}
p(\boldsymbol{s}_*) &= C_D(\boldsymbol{\alpha})\frac{1}{C_D\Big((s_{*,k}+\alpha_k)_{k=1}^{K}\Big)} \\
&= \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma(s_{*,k}+\alpha_k)}{\Gamma(\sum_{k=1}^{K}s_{*,k}+\alpha_k)}
\end{aligned} \tag{3.30}
$$

となる。

ある k' に対して $s_{*,k'} = 1$ となる場合のみを取り出して考えてみると、$\sum_{k=1}^{K} s_{*,k} = 1$ より

$$
\begin{aligned}
p(s_{*,k'}=1) &= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\alpha_{k'})} \frac{\Gamma(s_{*,k'}+\alpha_{k'})}{\Gamma(\sum_{k=1}^K s_{*,k}+\alpha_k)} \\
&= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\alpha_{k'})} \frac{\Gamma(1+\alpha_{k'})}{\Gamma(1+\sum_{k=1}^K \alpha_k)} \\
&= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\alpha_{k'})} \frac{\alpha_{k'}\Gamma(\alpha_{k'})}{(\sum_{k=1}^K \alpha_k)\Gamma(\sum_{k=1}^K \alpha_k)} \\
&= \frac{\alpha_{k'}}{\sum_{k=1}^K \alpha_k} \qquad (3.31)
\end{aligned}
$$

となる。

従って、$s_{*,k}=1$ のときそれ以外の $s_{*,1},\cdots,s_{*,k'-1},s_{*,k'+1},\cdots,s_{*,K}$ は全て 0 であるため $x^0=1$ であることを利用して、$s_{*,1}$ から $s_{*,K}$ までをまとめると式 (3.30) は

$$
p(s_{*,k}) = \prod_{k=1}^K \Big(\frac{\alpha_k}{\sum_{k'=1}^K \alpha_{k'}}\Big)^{s_{*,k}}
$$

と書き換えられる。(分母の k' はこれまでのある k' とは別物。あくまで分子の k と区別するためのもの。本では i と表記することで区別している。)

この式について

$$
\boldsymbol{\pi}_* = (\pi_{*,1},\pi_{*,2},\cdots,\pi_{*,K}),\ \pi_{*,k} = \frac{\alpha_k}{\sum_{k'=1}^K \alpha_{k'}}
$$

とおくと、予測分布 $p(s_{*,k})$ は

$$
p(s_{*,k}) = \prod_{k=1}^K \pi_{*,k}^{s_{*,k}} = \mathrm{Cat}(\boldsymbol{s}_*|\boldsymbol{\pi}_*) \qquad (3.32)
$$

となる。式の形状から、$p(x_*)$ はパラメータ $\boldsymbol{\pi}_*$ を持つカテゴリ分布であることが分かる。
ちなみに $\frac{\alpha_k}{\sum_{k'=1}^K \alpha_{k'}}$ は、ディリクレ分布 $\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$ の期待値 $\mathbb{E}[\pi_k]$ である。

予測分布のパラメータ $\boldsymbol{\pi}_*$ を構成する α_k について、事後分布のパラメータ (3.28) を用いると

$$
\begin{aligned}
\hat{\pi}_{*,k} &= \frac{\hat{\alpha}_k}{\sum_{k'=1}^K \hat{\alpha}_{k'}} \\
&= \frac{\sum_{n=1}^N s_{n,k}+\alpha_k}{\sum_{k'=1}^K \sum_{n=1}^N s_{n,k'}+\alpha_{k'}}
\end{aligned}
$$

となり、(観測データ $\boldsymbol{S}$ によって学習した) 事後分布 $p(\boldsymbol{\pi}|\boldsymbol{S})$ を用いた予測分布 $p(s_{*,k}|\boldsymbol{S})$

$$
p(s_{*,k}|\boldsymbol{S}) = \prod_{k=1}^K \hat{\pi}_{*,k}^{s_{*,k}} = \mathrm{Cat}(\boldsymbol{s}_*|\hat{\boldsymbol{\pi}}_*) \qquad (3.32)
$$

が得られる。

・**R でやってみよう**

各確率分布に従いランダムに生成したデータを用いて、パラメータを推定してみましょう。

```r
# 利用パッケージ
library(tidyverse)
```

利用するパッケージを読み込みます。

・**事後分布**

```r
## パラメーターの初期値を指定
# 観測モデルのパラメータ
K <- 3 # 次元数（固定）
pi_k_truth <- c(1/6, 3/6, 2/6)

# 事前分布のパラメータ
alpha_k <- c(2, 2, 2)

# 試行回数
N <- 50
```

このプログラムでは三角座標で作図するため、各パラメータの次元数 (K) は 3 で固定です。

データの発生確率 $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ を `pi_k_truth` とします。この値を推定するのが目的です。

事前分布のパラメータ (ハイパーパラメータの初期値)$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ を `alpha_k` とします。

生成するデータ数 N を `N` とします。

```r
# 作図用の pi の値を満遍なく生成
pi <- tibble(
  pi_1 = rep(rep(seq(0, 1, by = 0.02), times = 51), times = 51),
  pi_2 = rep(rep(seq(0, 1, by = 0.02), each = 51), times = 51),
  pi_3 = rep(seq(0, 1, by = 0.02), each = 2601)
)

# 正規化
pi <- pi / apply(pi, 1, sum)

# 重複した組み合わせを除去（ハイスペ機なら不要…）
pi <- pi %>%
  mutate(pi_1 = round(pi_1, 3), pi_2 = round(pi_2, 3), pi_3 = round(pi_3, 3)) %>%
  count(pi_1, pi_2, pi_3) %>%
  select(-n) %>%
  as.matrix()
```

π が取り得る 0 から 1 までの値を 3 次元分用意します。ただし重複した組み合わせなく値を作るには少し工夫がいります。

まずは各引数の作用を確認しましょう。seq() の by は最小値 (第 1 引数) から最大値 (第 2 引数) までを刻む幅です。rep() の第 1 引数にベクトル (複数の値) を渡した場合、times はベクトルを繰り返す回数、each はベクトルの各要素を複製する個数です。

また by に指定した値によって他の引数に指定すべき値が変わります。by に指定した値を n とすると、seq() によって、1/n+1 個 (+1 は 0 の分) の要素のベクトルが返ってきます。この 1/n+1 を他の引数に指定します。

ただし、データフレームでいう 3 列目 (サンプルコードでいうと 5 行目) の each は他で指定した値の 2 乗の値 ((1/n+1)^2) になります。

入れ子関係や使う引数自体にも注意しましょう。

簡単な例で確認すると次のようにしたいわけです。

```r
tibble(
  v1 = rep(rep(seq(1, 2), times = 2), times = 2),
  v2 = rep(rep(seq(1, 2), each = 2), times = 2),
  v3 = rep(seq(1, 2), each = 4)
)
```

```
## # A tibble: 8 x 3
##      v1    v2    v3
##   <int> <int> <int>
## 1     1     1     1
## 2     2     1     1
## 3     1     2     1
## 4     2     2     1
## 5     1     1     2
## 6     2     1     2
## 7     1     2     2
## 8     2     2     2
```

何言ってんだ ?? となった場合は、次のランダムに点を打つ方法を使いましょう。

用意した値は、列ごとの総和が 1 となるように総和で割って正規化します。

私の環境だと、データが多くて作図時に固まってしまったので、重複した箇所を間引くことにしました…。

この処理をしない場合は、正規化の前後どちらかで as.matrix(pi) の処理を加えてマトリクスに変換しておく必要があります。

```r
# 作図用の pi の値をランダムに生成
pi <- matrix(
  sample(seq(0, 1, 0.01), size = 90000, replace = TRUE),
  nrow = 3
)

# 正規化
pi <- pi / apply(pi, 1, sum)
```

満遍なく点で埋めつくす設定は少しややこしいため、ランダムに点を生成してもいいです。

点の数が少なすぎると疎らになり、多すぎると処理が重くなります。点の数は size 引数に指定する値で調整できます。ただしマトリクスに欠損値が出ないように、値を 3 の倍数にする必要があるので注意してください。

こちらのやり方でも、各行の値の和が 1 となるように正規化します。

```r
# カテゴリ分布に従うデータを生成
s_nk <- rmultinom(n = N, size = 1, prob = pi_k_truth) %>%
  t()
```

多項分布に従う乱数を発生させる関数 rmultinom() を使って、ランダムにデータを生成します。これをカテゴリ分布とするためには、size 引数に 1 を指定します。
試行回数の引数 n には N、確率の引数 prob には pi_k_truth を指定します。
試行ごとの結果を列としたマトリクスが返ってくるので、t() で転置して式と合わせます。

サンプルを確認してみましょう。

```r
# 観測データを確認
apply(s_nk, 2, sum)
```

```
## [1]  7 25 18
```

このデータを用いて事後分布のパラメータを計算します。

```r
# 事後分布のパラメータを計算
alpha_k_hat <- apply(s_nk, 2, sum) + alpha_k
```

事後分布のパラメータの計算式 (3.28) の計算を行い、$\hat{\alpha}_k$ を alpha_k とします。

```r
# 事後分布を計算
posterior_df <- tibble(
  x = pi[, 2] + (pi[, 3] / 2),  # 三角座標への変換
  y = sqrt(3) * (pi[, 3] / 2),  # 三角座標への変換
  C_D = lgamma(sum(alpha_k_hat)) - sum(lgamma(alpha_k_hat)),  # 正規化項（対数）
  density = exp(C_D + apply((alpha_k_hat - 1) * log(t(pi)), 2, sum)) # 確率密度
)
```

最初に用意した pi を使って各値の確率密度を計算します。ただし 3 次元の情報を 2 次元の図で表現するために、三角座標に変換します。(よく解ってないゆえ解説なし)
pi の各値に対してディリクレ分布定義式 (2.48) の計算を行い、確率密度を求めます。ただし値が大きくなると gamma() で計算できなくなるため、対数をとって計算することにします。なので最後に exp() で値を戻します。

計算結果は作図用にデータフレームにまとめておきます。推定結果を確認してみましょう。

```r
head(posterior_df)
```

```
## # A tibble: 6 x 4
##       x     y   C_D density
##   <dbl> <dbl> <dbl>   <dbl>
## 1 0.5   0.866  57.1       0
## 2 0.51  0.849  57.1       0
## 3 0.510 0.848  57.1       0
```

```
## 4 0.511 0.847  57.1       0
## 5 0.511 0.846  57.1       0
## 6 0.512 0.845  57.1       0
```

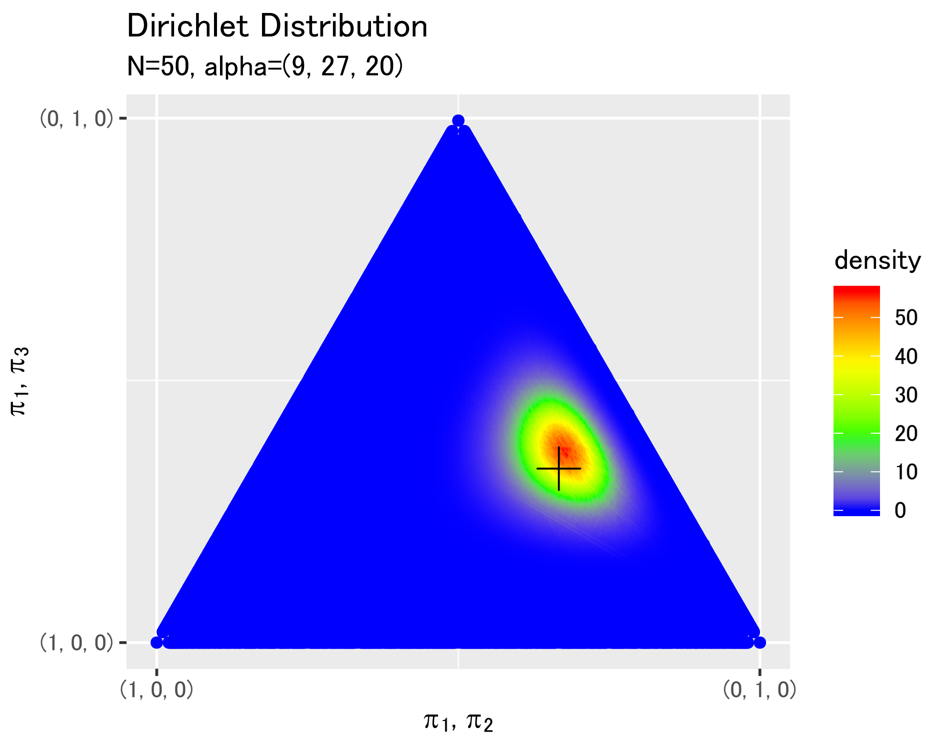
このデータフレームを用いてグラフを描きます。

```r
# pi の真の値のプロット用データフレームを作成
pi_truth_df <- tibble(
  x = pi_k_truth[2] + (pi_k_truth[3] / 2),  # 三角座標への変換
  y = sqrt(3) * (pi_k_truth[3] / 2),        # 三角座標への変換
)
```

最初に設定した $\boldsymbol{\pi}$ の値を推定できているのか確認するために、真の値の位置もプロットしましょう。

`pi_k_truth` の値も三角座標に変換します。`ggplot()` にはデータフレームで渡す必要があるため、データフレームで保存しておきます。

```r
# 描画
ggplot() +
  geom_point(data = posterior_df, aes(x, y, color = density)) + # 散布図
  geom_point(data = pi_truth_df, aes(x, y), shape = 3, size = 5) + # pi の真の値
  scale_color_gradientn(colors = c("blue", "green", "yellow", "red")) + # プロットの色
  scale_x_continuous(breaks = c(0, 1),
                     labels = c("(1, 0, 0)", "(0, 1, 0)")) + # x 軸目盛
  scale_y_continuous(breaks = c(0, 0.87),
                     labels = c("(1, 0, 0)", "(0, 1, 0)")) + # y 軸目盛
  coord_fixed(ratio = 1) + # 縦横比
  labs(title = "Dirichlet Distribution",
       subtitle = paste0("N=", N, ", alpha=(", paste(alpha_k_hat, collapse = ", "), ")"),
       x = expression(paste(pi[1], ", ", pi[2], sep = "")),
       y = expression(paste(pi[1], ", ", pi[3], sep = ""))) # ラベル
```



ggplot2 パッケージを利用してグラフを作成します。

推定値は散布図 geom_point() を用いて描きます。
パラメータ π の真の値も geom_point() で示します。

・予測分布

```
# 予測分布のパラメータを計算
pi_k_hat <- alpha_k_hat / sum(alpha_k_hat)
```

予測分布のパラメータの計算式の計算を行い、$\hat{\pi}_k$ を pi_k_hat とします。

```
pi_k_hat <- (apply(s_nk, 2, sum) + alpha_k) / sum(apply(s_nk, 2, sum) + alpha_k)
```

このように、観測データ s_nk と事前分布のパラメータ alpha_k を使って計算することもできます。

```
# 作図用の s の値
s_sk <- matrix(c(1, 0, 0, 0, 1, 0, 0, 0, 1), ncol = 3)
```

$s_{*,k}$ が取り得る値を用意します。

```
# 予測分布を計算
predict_df <- tibble(
  k = seq(1, 3),  # 作図用の値
  prob = apply(pi_k_hat^s_sk, 1, prod) # 確率
)
```

`k` の各値となる確率は、カテゴリ分布の定義式 (2.29) で計算します。

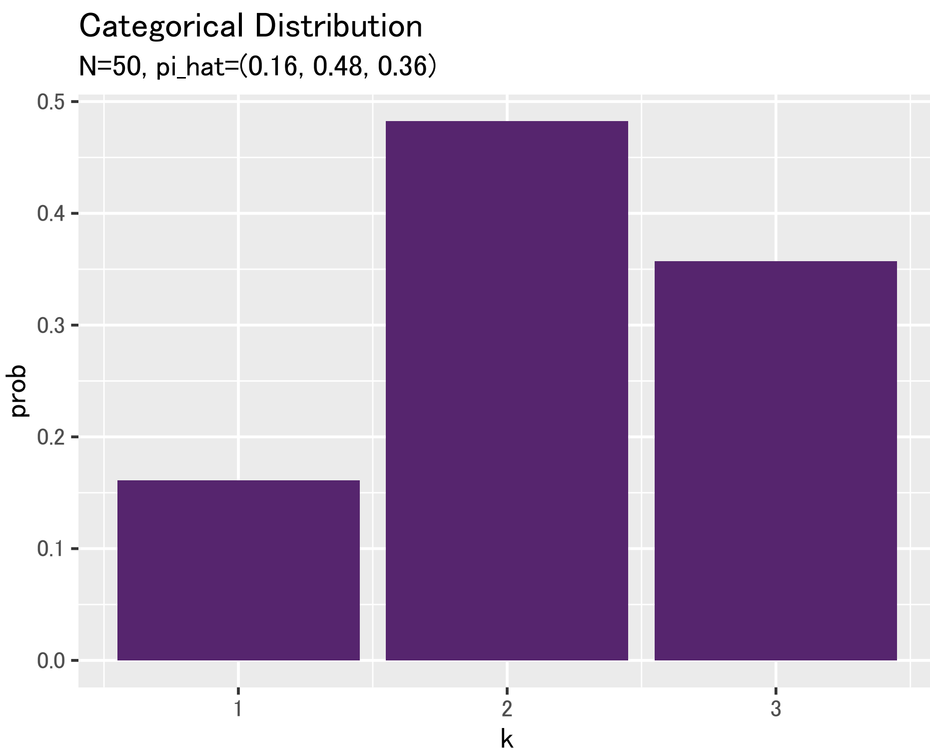
推定結果を確認してみましょう。

```
head(predict_df)
```

```
## # A tibble: 3 x 2
##       k  prob
##   <int> <dbl>
## 1     1 0.161
## 2     2 0.482
## 3     3 0.357
```

こちらも `ggplot2` パッケージを利用して作図します。

```
# 作図
ggplot(predict_df, aes(k, prob)) +
  geom_bar(stat = "identity", position = "dodge", fill = "#56256E") + # 棒グラフ
  labs(title = "Categorical Distribution",
       subtitle = paste0("N=", N, ", pi_hat=(", paste(round(pi_k_hat, 2), collapse = ", "), ")")) # ラベ
```



棒グラフは `geom_bar()` を使います。

・おまけ

`gganimate` パッケージを利用して、パラメータの推定値の推移の gif 画像を作成するためのコードです。

```r
# 利用パッケージ
library(tidyverse)
library(gganimate)

## パラメーターの初期値を指定
# 観測モデルのパラメータ
K <- 3 # 次元数（固定）
pi_k_truth <- c(1/6, 3/6, 2/6)

# 事前分布のパラメータ
alpha_k <- c(2, 2, 2)

# 試行回数
N <- 50

# 作図用の pi の値
pi <- tibble(
  pi_1 = rep(rep(seq(0, 1, by = 0.025), times = 41), times = 41),
  pi_2 = rep(rep(seq(0, 1, by = 0.025), each = 41), times = 41),
  pi_3 = rep(seq(0, 1, by = 0.025), each = 1681)
)

# 正規化
pi <- pi / apply(pi, 1, sum)

# 重複した組み合わせを除去（ハイスペ機なら不要…）
pi <- pi %>%
  mutate(pi_1 = round(pi_1, 3), pi_2 = round(pi_2, 3), pi_3 = round(pi_3, 3)) %>%
  count(pi_1, pi_2, pi_3) %>%
  select(-n) %>%
  as.matrix()

# 事前分布を計算
posterior_df <- tibble(
  x = pi[, 2] + (pi[, 3] / 2),  # 三角座標への変換
  y = sqrt(3) * (pi[, 3] / 2),  # 三角座標への変換
  C_D = lgamma(sum(alpha_k)) - sum(lgamma(alpha_k)),  # 正規化項（対数）
  density = exp(C_D + apply((alpha_k - 1) * log(t(pi)), 2, sum)), # 確率密度
  N = 0 # 試行回数
)

# 初期値による予測分布のパラメーターを計算
pi_k_hat <- alpha_k / sum(alpha_k)

# 作図用の s の値
s_sk <- matrix(c(1, 0, 0, 0, 1, 0, 0, 0, 1), ncol = 3)

# 初期値による予測分布を計算
predict_df <- tibble(
  k = seq(1, 3),  # 作図用の値
  prob = apply(pi_k_hat^s_sk, 1, prod),  # 確率
  N = 0 # 試行回数
)
```

```r

# パラメーターを推定
s_nk <- matrix(0, nrow = N, ncol = K) # 受け皿
for(n in 1:N){

  # カテゴリ分布に従うデータを生成
  s_nk[n, ] <- rmultinom(n = 1, size = 1, prob = pi_k_truth) %>%
    as.vector()

  # ハイパーパラメータを更新
  alpha_k <- s_nk[n, ] + alpha_k

  # 事後分布を計算
  tmp_posterior_df <- tibble(
    x = pi[, 2] + (pi[, 3] / 2),  # 三角座標への変換
    y = sqrt(3) * (pi[, 3] / 2),  # 三角座標への変換
    C_D = lgamma(sum(alpha_k)) - sum(lgamma(alpha_k)),  # 正規化項 (対数)
    density = exp(C_D + apply((alpha_k - 1) * log(t(pi)), 2, sum)), # 確率密度
    N = n # 試行回数
  )

  # 予測分布のパラメーターを計算
  pi_k_hat <- alpha_k / sum(alpha_k)

  # 予測分布を計算
  tmp_predict_df <- tibble(
    k = seq(1, 3),  # 作図用の値
    prob = apply(pi_k_hat^s_sk, 1, prod),  # 確率
    N = n # 試行回数
  )

  # 結果を結合
  posterior_df <- rbind(posterior_df, tmp_posterior_df)
  predict_df <- rbind(predict_df, tmp_predict_df)
}

# 観測データを確認
apply(s_nk, 2, sum)

# pi の真の値のプロット用データフレームを作成
pi_truth_df <- tibble(
  x = pi_k_truth[2] + (pi_k_truth[3] / 2),  # 三角座標への変換
  y = sqrt(3) * (pi_k_truth[3] / 2),  # 三角座標への変換
  N = seq(0, N)
)

## 事後分布
# 作図
posterior_graph <- ggplot() +
  geom_point(data = posterior_df, aes(x, y, color = density)) + # 散布図
  geom_point(data = pi_truth_df, aes(x, y), shape = 3, size = 5) + # pi の真の値
  scale_color_gradientn(colors = c("blue", "green", "yellow", "red")) + # プロットの色
  scale_x_continuous(breaks = c(0, 1),
                     labels = c("(1, 0, 0)", "(0, 1, 0)")) + # x 軸目盛
```

```r
  scale_y_continuous(breaks = c(0, 0.87),
                     labels = c("(1, 0, 0)", "(0, 1, 0)")) + # y 軸目盛
  coord_fixed(ratio = 1) + # 縦横比
  transition_manual(N) + # フレーム
  labs(title = "Dirichlet Distribution",
       subtitle = "N= {current_frame}",
       x = expression(paste(pi[1], ", ", pi[2], sep = "")),
       y = expression(paste(pi[1], ", ", pi[3], sep = ""))) # ラベル

# 描画
animate(posterior_graph)

## 予測分布
# 作図
predict_graph <- ggplot(predict_df, aes(k, prob)) +
  geom_bar(stat = "identity", position = "dodge", fill = "#56256E") + # 棒グラフ
  transition_manual(N) + # フレーム
  labs(title = "Categorical Distribution",
       subtitle = "N= {current_frame}") # ラベル

# 描画
animate(predict_graph)
```

異なる点のみを簡単に解説します。

各データによってどのように学習する (推定値が変化する) のかを確認するため、こちらは for ループで 1 データずつ処理します。
よって生成するデータ数として設定した `N` がイタレーション数になります。

パラメータの推定値について、$\hat{\alpha}_k$ に対応する `alpha_k_hat` を新たに作るのではなく、`alpha_k` をイタレーションごとに更新 (上書き) していきます。
それに伴い、事後分布のパラメータの計算式 (3.28) の $\sum_{n=1}^N$ の計算は、for ループによって N 回繰り返し `s_nk[n, ]` を加えることで行います。n 回目のループ処理のときには、n-1 回分の `s_nk[n, ]` が既に `alpha_k` に加えられているわけです。

3.2.3 ポアソン分布の学習と予測

・事後分布の計算

ポアソン分布に従うと仮定する N 個の観測データ $\boldsymbol{X} = \{x_1, x_2, \cdots, x_N\}$ を用いて、λ の事後分布を求めていく。

まずは観測モデル $p(\boldsymbol{X}|\lambda)$ について確認する。離散値データ $\boldsymbol{X} = \{x_1, x_2, \cdots, x_N\}$ はそれぞれ独立に生成されているとの仮定の下で、観測モデルは

$$
\begin{aligned}
p(\boldsymbol{X}|\lambda) &= p(x_1, x_2, \cdots, x_N|\lambda) \\
&= p(x_1|\lambda)p(x_2|\lambda)\cdots p(x_N|\lambda) \\
&= \prod_{n=1}^N p(x_n|\lambda)
\end{aligned}
$$

と分解できる。

これを用いて、λ の事後分布 $p(\lambda|\boldsymbol{X})$ はベイズの定理より

$$
\begin{aligned}
p(\lambda|\boldsymbol{X}) &= \frac{p(\boldsymbol{X}|\lambda)p(\lambda|a,b)}{p(\boldsymbol{X})} \\
&= \frac{\left\{\prod_{n=1}^{N} p(x_n|\lambda)\right\} p(\lambda|a,b)}{p(\boldsymbol{X})} \\
&= \frac{\left\{\prod_{n=1}^{N} \mathrm{Poi}(x_n|\lambda)\right\} \mathrm{Gam}(\lambda|a,b)}{p(\boldsymbol{X})} \\
&\propto \left\{\prod_{n=1}^{N} \mathrm{Poi}(x_n|\lambda)\right\} \mathrm{Gam}(\lambda|a,b) \qquad (3.35)
\end{aligned}
$$

となる。分母の $p(\boldsymbol{X})$ は λ に影響しないため省略して、比例関係にのみ注目する。省略した部分については、最後に正規化することで対応できる。

次にこの分布の具体的な形状を明らかにしていく。対数をとって指数部分の計算を分かりやすくして進めると

$$
\begin{aligned}
\ln p(\lambda|\boldsymbol{X}) &= \ln \prod_{n=1}^{N} \mathrm{Poi}(x_n|\lambda) + \ln \mathrm{Gam}(\lambda|a,b) - \ln p(\boldsymbol{X}) \\
&= \sum_{n=1}^{N} \ln \mathrm{Poi}(x_n|\lambda) + \ln \mathrm{Gam}(\lambda|a,b) + \mathrm{const.} \\
&= \sum_{n=1}^{N} \ln \left\{ \frac{\lambda^{x_n}}{x_n!} \exp(-\lambda) \right\} + \ln \Big\{ C_G(a,b) \lambda^{a-1} \exp(-b\lambda) \Big\} + \mathrm{const.} \\
&= \sum_{n=1}^{N} \{ x_n \ln \lambda - \ln x_n! - \lambda \} + \ln C_G(a,b) + (a-1) \ln \lambda - b\lambda + \mathrm{const.} \\
&= \sum_{n=1}^{N} x_n \ln \lambda - N\lambda + (a-1) \ln \lambda - b\lambda + \mathrm{const.} \\
&= \left(\sum_{n=1}^{N} x_n + a - 1 \right) \ln \lambda - (N+b)\lambda + \mathrm{const.} \qquad (3.36)
\end{aligned}
$$

【途中式の途中式】

0. 式 (3.36) の下から 2 行目の式を用いている。
1. 対数をとると積が和に変わる。また、λ に影響しない項 $(-\ln p(\boldsymbol{X}))$ を const. とおく。
2. $\mathrm{Poi}(x_n|\lambda)$, $\mathrm{Gam}(\lambda|a,b)$ に、それぞれポアソン分布の定義式 (2.37)、ガンマ分布の定義式 (2.56) を用いて具体的な確率分布を代入する。
3. それぞれ対数をとる。$\ln(x^a y^b) = a \ln x + b \ln y$ である。
4. λ と関係のない $-\sum_{n=1}^{N} \ln x_n!$, $\ln C_G(a,b)$ を const. にまとめる。
5. $\ln \lambda$ と λ の項をまとめて式を整理する。

となる。

　式 (3.36) について

$$
\begin{aligned}
\hat{a} &= \sum_{n=1}^{N} x_n + a \\
\hat{b} &= N + b
\end{aligned}
\tag{3.38}
$$

とおき

$$
\ln p(\lambda|\boldsymbol{X}) = \hat{a}\ln\lambda - \hat{b}\lambda + \mathrm{const.} \tag{3.36}
$$

更に ln を外し、const. を正規化項に置き換えると

$$
p(\lambda|\boldsymbol{X}) = C_G(\hat{a}, \hat{b})\lambda^{\hat{a}}\exp(-\hat{b}\lambda) = \mathrm{Gam}(\lambda|\hat{a}, \hat{b}) \tag{3.37}
$$

となる。式の形状から、事後分布 $p(\lambda|\boldsymbol{X})$ がパラメータ $\hat{a},\ \hat{b}$ を持つガンマ分布であることを確認できる。
　また事後分布の計算式 (3.38) によって、ハイパーパラメータ $a,\ b$ が更新式される。

・予測分布の計算

　続いて未観測のデータ x_* に対する予測分布を求めていく。

　(観測データによる学習を行っていない) 事前分布 $p(\lambda)$ を用いて、パラメータ λ を周辺化することで予測分布 $p(x_*)$ となる。

$$
\begin{aligned}
p(x_*) &= \int p(x_*|\lambda)p(\lambda)d\lambda \\
&= \int \mathrm{Poi}(x_*|\lambda)\mathrm{Gam}(\lambda|a, b)d\lambda && (3.39) \\
&= \int \frac{\lambda^{x_*}}{x_*!}\exp(-\lambda)C_G(a, b)\lambda^{a-1}\exp(-b\lambda)d\lambda \\
&= \frac{C_G(a, b)}{x_*!}\int \lambda^{x_*+a-1}\exp\Big(-(1+b)\lambda\Big)d\lambda && (3.40)
\end{aligned}
$$

　積分部分に注目すると、パラメータ $x_* + a,\ 1 + b$ を持つ正規化項のないガンマ分布の形をしている。よって、ガンマ分布の定義式 (2.56) を用いて

$$
\int \lambda^{x_*+a-1}\exp\Big(-(1+b)\lambda\Big)d\lambda = \frac{1}{C_G(x_* + a, 1 + b)}
$$

正規化項の逆数に変形できる (そもそもこの部分の逆数が正規化項である)。これを式 (3.40) に代入すると、予測分布 $p(x_*)$ は

$$p(x_*) = \frac{C_G(a, b)}{x_*! C_G(x_* + a, 1 + b)} \tag{3.42}$$

となる。更にガンマ分布の正規化項 (2.57) を用いて、式を整理すると

$$\begin{aligned}
p(x_*) &= \frac{C_G(a, b)}{x_*!} \frac{1}{C_G(x_* + a, 1 + b)} \\
&= \frac{b^a}{x_*! \Gamma(a)} \frac{\Gamma(x_* + a)}{(1 + b)^{x_* + a}} \\
&= \frac{\Gamma(x_* + a)}{x_*! \Gamma(a)} \frac{b^a}{(1 + b)^a} \frac{1}{(1 + b)^{x_*}} \\
&= \frac{\Gamma(x_* + a)}{x_*! \Gamma(a)} \Big(\frac{b}{1 + b}\Big)^a \Big(\frac{1}{1 + b}\Big)^{x_*} \\
&= \frac{\Gamma(x_* + a)}{x_*! \Gamma(a)} \Big(1 - \frac{1}{1 + b}\Big)^a \Big(\frac{1}{1 + b}\Big)^{x_*}
\end{aligned}$$

となる。

この式の形状の確率分布は負の二項分布と呼ばれる。

この式について

$$\begin{aligned}
r &= a \\
p &= \frac{1}{1 + b}
\end{aligned} \tag{3.44}$$

とおくと、予測分布 $p(x_*)$ は

$$p(x_*) = \frac{\Gamma(x_* + r)}{x_*! \Gamma(r)} (1 - p)^a p^{x_*} = \mathrm{NB}(x_* | r, p) \tag{3.43}$$

パラメータ $r,\ p$ を持つ負の二項分布となることが分かる。

予測分布のパラメータ $r,\ p$ を構成する $a,\ b$ について、事後分布のパラメータ (3.38) を用いると

$$\begin{aligned}
\hat{r} &= \hat{a} \\
&= \sum_{n=1}^N x_n + a
\end{aligned}$$

また

$$\begin{aligned}
\hat{p} &= \frac{1}{1 + \hat{b}} \\
&= \frac{1}{N + 1 + b}
\end{aligned}$$

となり、(観測データ $\boldsymbol{X}$ によって学習した) 事後分布 $p(\lambda | \boldsymbol{X})$ を用いた予測分布 $p(x_* | \boldsymbol{X})$

$$p(x_*|\boldsymbol{X}) = \frac{\Gamma(x_* + \hat{r})}{x_*!\Gamma(\hat{r})}(1-\hat{p})^a \hat{p}^{x_*} = \mathrm{NB}(\hat{x}_*|\hat{r}, \hat{p})$$

が得られる。

・負の二項分布

$$\mathrm{NB}(x|r,p) = \frac{\Gamma(x+r)}{x!\Gamma(r)}(1-p)^r p^x \tag{3.43}$$

・R でやってみよう

各確率分布に従いランダムに生成したデータを用いて、パラメータを推定してみましょう。

```
# 利用パッケージ
library(tidyverse)
```

利用するパッケージを読み込みます。

・事後分布

```
## パラメータの初期値を設定
# 観測モデルのパラメータ
lambda_truth <- 4

# 事前分布のパラメータ
a <- 1
b <- 1

# 試行回数
N <- 50
```

パラメータ λ を `lambda_truth` とします。この値を推定するのが目的です。

事前分布のパラメータ (ハイパーパラメータの初期値)$a,\ b$ をそれぞれ `a` と `b` とします。

生成するデータ数 N を `N` とします。

```
# ポアソン分布に従うデータを生成
x_n <- rpois(n = N ,lambda = lambda_truth)
```

ポアソン分布に従う乱数を発生させる関数 `rpois()` を使って、ランダムにデータを生成します。
試行回数の引数 `n` には `N`、パラメータ引数 `lambda` には `lambda_truth` を指定します。

サンプルを確認してみましょう。

```r
# 観測データを確認
x_n_df <- x_n %>%
  table() %>%
  as_tibble()
colnames(x_n_df) <- c("n", "x")
x_n_df$n <- as.numeric(x_n_df$n)
head(x_n_df)
```

```
## # A tibble: 6 x 2
##       n     x
##   <dbl> <int>
## 1     0     1
## 2     1     9
## 3     2    10
## 4     3     7
## 5     4     8
## 6     5     9
```

このデータを用いて事後分布のパラメータを計算します。

```r
# 事後分布のパラメータを計算
a_hat <- sum(x_n) + a
b_hat <- N + b
```

事後分布のパラメータの計算式 (3.38) の計算を行い、$\hat{a}, \hat{b}$ をそれぞれ `a_hat`, `b_hat` とします。

```r
# 事後分布を計算
posterior_df <- tibble(
  lambda = seq(0, 2 * lambda_truth, by = 0.001),  # 作図用の値
  C_G = a_hat * log(b_hat) - lgamma(a_hat),    # 正規化項（対数）
  density = exp(C_G + (a_hat - 1) * log(lambda) - b_hat * lambda) # 確率密度
)
```

λ が取り得る値を `seq()` を使って値を用意します。ただし `lambda_truth` の値によって必要な範囲が変わります。ここでは 0 から `lambda_truth` の 2 倍までとします。そこで第 1 引数 (最小値) には 0、第 2 引数 (最大値) には `2 * lambda_truth` を指定します。

`lambda` の各値に対してガンマ分布の定義式 (2.56) の計算を行い、確率密度を求めます。ただし値が大きくなると `gamma()` で計算できなくなるため、対数をとって計算することにします。なので最後に `exp()` で値を戻します。

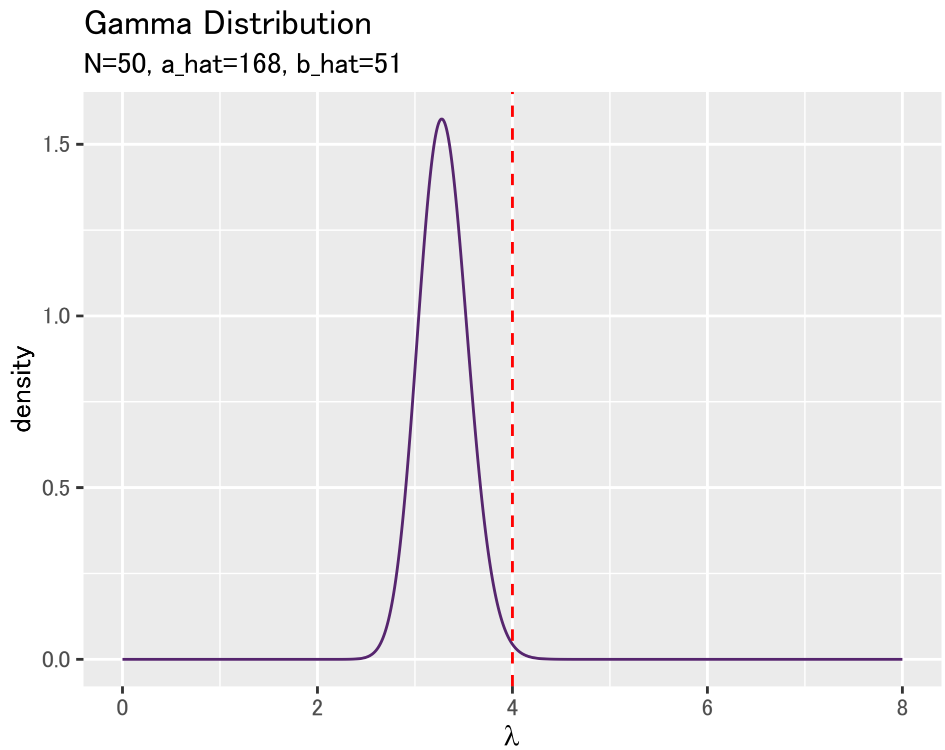
計算結果は作図用にデータフレームにまとめておきます。推定結果を確認してみましょう。

```r
head(posterior_df)
```

```
## # A tibble: 6 x 3
##   lambda   C_G density
##    <dbl> <dbl>   <dbl>
## 1  0     -30.6       0
## 2  0.001 -30.6       0
## 3  0.002 -30.6       0
## 4  0.003 -30.6       0
## 5  0.004 -30.6       0
## 6  0.005 -30.6       0
```

このデータフレームを用いてグラフを描きます。

```r
# 作図
ggplot(posterior_df, aes(lambda, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = lambda_truth),
             color = "red", linetype = "dashed") + # 垂直線
  labs(title = "Gamma Distribution",
       subtitle = paste0("N=", N, ", a_hat=", a_hat, ", b_hat=", b_hat),
       x = expression(lambda)) # ラベル
```



ggplot2 パッケージを利用してグラフを作成します。

推定値は折れ線グラフ geom_line() を用いて描きます。
パラメータ λ の真の値を垂直線 geom_vline() で示します。

・予測分布

```r
# 予測分布のパラメータを計算
r_hat <- a_hat
p_hat <- 1 / (b_hat + 1)
```

予測分布のパラメータの計算式の計算を行い、$\hat{r}$, $\hat{p}$ をそれぞれ r_hat, p_hat とします。

```r
r_hat <- sum(x_n) + a
p_hat <- 1 / (N + 1 + b)
```

このように、観測データ x_n と事前分布のパラメータ a, b を使って計算することもできます。

```
# 予測分布を計算
predict_df <- tibble(
  x = seq(0, 4 * lambda_truth),  # 作図用の値
  C_NB = lgamma(x + r_hat) - lgamma(x + 1) - lgamma(r_hat), # 正規化項（対数）
  prob = exp(C_NB + r_hat * log(1 - p_hat) + x * log(p_hat)) # 確率
)
```

事後分布と同様に、x_* が取り得る値を seq() を使って用意します。
x の各値となる確率は、負の二項分布の定義式 (3.43) で計算します。

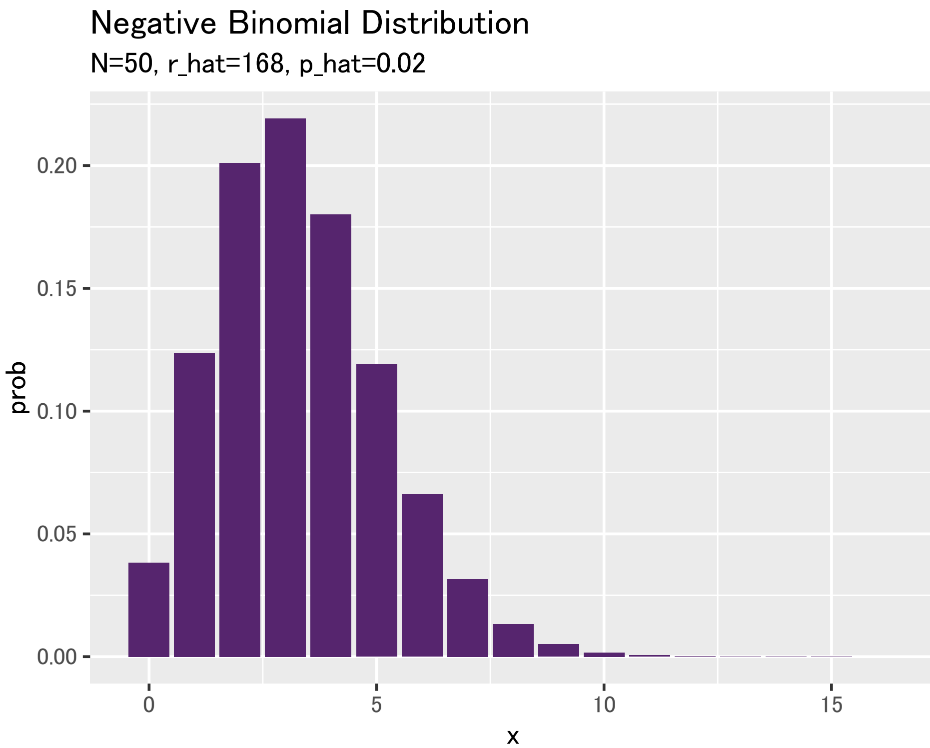
推定結果を確認してみましょう。

```
head(predict_df)
```

```
## # A tibble: 6 x 3
##       x  C_NB   prob
##   <int> <dbl>  <dbl>
## 1     0  0    0.0383
## 2     1  5.12 0.124
## 3     2  9.56 0.201
## 4     3 13.6  0.219
## 5     4 17.4  0.180
## 6     5 20.9  0.119
```

こちらも ggplot2 パッケージを利用して作図します。

```
# 作図
ggplot(predict_df, aes(x, prob)) +
  geom_bar(stat = "identity", position = "dodge", fill = "#56256E") + # 棒グラフ
  labs(title = "Negative Binomial Distribution",
       subtitle = paste0("N=", N, ", r_hat=", r_hat, ", p_hat=", round(p_hat, 2))) # ラベル
```



棒グラフは geom_bar() を使います。

・おまけ

gganimate パッケージを利用して、パラメータの推定値の推移の gif 画像を作成するためのコードです。

```
# 利用パッケージ
library(tidyverse)
library(gganimate)

## パラメータの初期値を設定
# 観測モデルのパラメータ
lambda_truth <- 4

# 事前分布のパラメータ
a <- 1
b <- 1

# 試行回数
N <- 50

# 事前分布を計算
posterior_df <- tibble(
  lambda = seq(0, 2 * lambda_truth, by = 0.001),  # 作図用の値
  C_G = a * log(b) - lgamma(a),    # 正規化項（対数）
  density = exp(C_G + (a - 1) * log(lambda) - b * lambda), # 確率密度
```

```r
  N = 0 # 試行回数
)

# 初期値による予測分布のパラメータを計算
r <- a
p <- 1 / (b + 1)

# 初期値による予測分布を計算
predict_df <- tibble(
  x = seq(0, 4 * lambda_truth),  # 作図用の値
  C_NB = lgamma(x + r) - lgamma(x + 1) - lgamma(r), # 正規化項（対数）
  prob = exp(C_NB + r * log(1 - p) + x * log(p)), # 確率
  N = 0 # 試行回数
)

# パラメータを推定
x_n <- rep(0, N) # 受け皿
for(n in 1:N){

  # ポアソン分布に従うデータを生成
  x_n[n] <- rpois(n = 1 ,lambda = lambda_truth)

  # ハイパーパラメータを更新
  a <- sum(x_n[n] * 1) + a
  b <- 1 + b

  # 事後分布を推定
  tmp_posterior_df <- tibble(
    lambda = seq(0, 2 * lambda_truth, by = 0.001),  # 作図用の lamnda の値
    C_G = a * log(b) - lgamma(a),   # 正規化項（対数）
    density = exp(C_G + (a - 1) * log(lambda) - b * lambda), # 確率密度
    N = n # 試行回数
  )

  # 予測分布のパラメータを計算
  r <- a
  p <- 1 / (b + 1)

  # 予測分布を計算
  tmp_predict_df <- tibble(
    x = seq(0, 4 * lambda_truth),  # 作図用の値
    C_NB = lgamma(x + r) - lgamma(x + 1) - lgamma(r), # 正規化項（対数）
    prob = exp(C_NB + r * log(1 - p) + x * log(p)), # 確率
    N = n # 試行回数
  )

  # 結果を結合
  posterior_df <- rbind(posterior_df, tmp_posterior_df)
  predict_df <- rbind(predict_df, tmp_predict_df)
}

# 観測データを確認
table(x_n)
```

```
## 事後分布
# 作図
posterior_graph <- ggplot(posterior_df, aes(lambda, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = lambda_truth),
             color = "red", linetype = "dashed") + # 垂直線
  transition_manual(N) + # フレーム
  labs(title = "Gamma Distribution",
       subtitle = "N= {current_frame}",
       x = expression(lambda)) # ラベル

# 描画
animate(posterior_graph)

## 予測分布
# 作図
predict_graph <- ggplot(predict_df, aes(x, prob)) +
  geom_bar(stat = "identity", position = "dodge", fill = "#56256E") + # 棒グラフ
  transition_manual(N) + # フレーム
  labs(title = "Negative Binomial Distribution",
       subtitle = "N= {current_frame}") # ラベル

# 描画
animate(predict_graph)
```

異なる点のみを簡単に解説します。

各データによってどのように学習する (推定値が変化する) のかを確認するため、こちらは for ループで 1 データずつ処理します。

よって生成するデータ数として設定した `N` がイタレーション数になります。

パラメータの推定値について、$\hat{a}$, $\hat{b}$ に対応する `a_hat`, `b_hat` を新たに作るのではなく、`a`, `b` をイタレーションごとに更新 (上書き) していきます。

それに伴い、事後分布のパラメータの計算式 (3.38) の $\sum_{n=1}^N$ の計算は、for ループによって N 回繰り返し `x_n[n]` を加えることで行います。n 回目のループ処理のときには、n-1 回分の `x_n[n]`(`b` の場合は 1) が既に `a` と `b` に加えられているわけです。

3.3 1 次元ガウス分布の学習と予測

ここでは、ガウス分布のパラメータの 1 つである分散 σ^2 の代わりに、その逆数である精度パラメータ $\lambda = \sigma^{-2}$ を用いる。

$$\lambda = \sigma^{-2} = \frac{1}{\sigma^2}$$
$$\sigma^2 = \frac{1}{\lambda} = \lambda^{-1}$$

3.3.1 平均が未知の場合

・事後分布の計算

ガウス分布に従うと仮定する N 個の観測データ $\boldsymbol{X} = \{x_1, x_2, \cdots, x_N\}$ を用いて、平均パラメータ μ の事後分布を求めていく。

観測モデル $p(\boldsymbol{X}|\mu)$ を用いて、観測データ $\boldsymbol{X}$ によって学習した μ の事後分布 $p(\mu|\boldsymbol{X})$ はベイズの定理より

$$
\begin{aligned}
p(\mu|\boldsymbol{X}) &= \frac{p(\boldsymbol{X}|\mu)p(\mu)}{p(\boldsymbol{X})} \\
&= \frac{\left\{\prod_{n=1}^{N} p(x_n|\mu)\right\} p(\mu)}{p(\boldsymbol{X})} \\
&= \frac{\left\{\prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1})\right\} \mathcal{N}(\mu|m, \lambda_\mu^{-1})}{p(\boldsymbol{X})} \\
&\propto \left\{\prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1})\right\} \mathcal{N}(\mu|m, \lambda_\mu^{-1})
\end{aligned}
\tag{3.49}
$$

となる。分母の $p(\boldsymbol{X})$ は μ に影響しないため省略して、比例関係にのみ注目する。省略した部分については、最後に正規化することで対応できる。

次にこの分布の具体的な形状を明らかにしていく。対数をとって指数部分の計算を分かりやすくして進めると

$$
\begin{aligned}
\ln p(\mu|\boldsymbol{X}) &= \ln \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) + \ln \mathcal{N}(\mu|m, \lambda_\mu^{-1}) - \ln p(\boldsymbol{X}) \\
&= \sum_{n=1}^{N} \ln \mathcal{N}(x_n|\mu, \lambda^{-1}) + \ln \mathcal{N}(\mu|m, \lambda_\mu^{-1}) + \text{const.} \\
&= \sum_{n=1}^{N} \ln \left\{ \frac{1}{\sqrt{2\pi\lambda^{-1}}} \exp\left(-\frac{(x_n-\mu)^2}{2\lambda^{-1}} \right) \right\} + \ln \left\{ \frac{1}{\sqrt{2\pi\lambda_\mu^{-1}}} \exp\left(-\frac{(\mu-m)^2}{2\lambda_\mu^{-1}} \right) \right\} + \text{const.} \\
&= \sum_{n=1}^{N} \left\{ -\frac{1}{2} \ln 2\pi\lambda^{-1} - \frac{(x_n-\mu)^2}{2\lambda^{-1}} \right\} - \frac{1}{2} \ln 2\pi\lambda_\mu^{-1} - \frac{(\mu-m)^2}{2\lambda_\mu^{-1}} + \text{const.} \\
&= -\frac{1}{2} \left\{ \sum_{n=1}^{N} (x_n-\mu)^2\lambda + (\mu-m)^2\lambda_\mu \right\} + \text{const.} \\
&= -\frac{1}{2} \left\{ \sum_{n=1}^{N} (x_n^2\lambda - 2x_n\lambda\mu + \lambda\mu^2) + \lambda_\mu\mu^2 - 2m\lambda_\mu\mu + m^2\lambda_\mu \right\} + \text{const.} \\
&= -\frac{1}{2} \left(-2\sum_{n=1}^{N} x_n\lambda\mu + N\lambda\mu^2 + \lambda_\mu\mu^2 - 2m\lambda_\mu\mu \right) + \text{const.} \\
&= -\frac{1}{2} \left\{ (N\lambda+\lambda_\mu)\mu^2 - 2\left(\sum_{n=1}^{N} x_n\lambda + m\lambda_\mu \right)\mu \right\} + \text{const.}
\end{aligned}
\tag{3.50}
$$

【途中式の途中式】

0. 式 (3.49) の下から 2 行目の式を用いている。
1. 対数をとると積が和に変わる。また μ と関係のない項 $(-\ln p(\boldsymbol{X}))$ を const. とおく。
2. $\mathcal{N}(x_n|\mu,\lambda^{-1})$, $\mathcal{N}(\mu|m,\lambda_\mu^{-1})$ に、1 次元ガウス分布の定義式 (2.64) を用いて具体的な確率分布を代入する。
3. それぞれ対数をとる。$\ln\sqrt{\frac{1}{x}}=\ln x^{-\frac{1}{2}}=-\frac{1}{2}\ln x$、また $\ln\exp(x)=x$ である。
4. $-\frac{1}{2}$ を括り出し、μ と関係のない項を const. に含める。
5. 2 乗部分を展開する。
6. $\sum_{n=1}^{N}$ に関する括弧を展開し、μ と関係のない項を const. に含める。
7. μ^2, μ の項をそれぞれまとめて式を整理する。

となる。

これまでと同様に、この式が (対数をとった)1 次元ガウス分布の形 (2.65) をしていることを確認したい。そこで

$$\begin{aligned}\hat{\lambda}_\mu &= N\lambda+\lambda_\mu \\ \tilde{m} &= \lambda\sum_{n=1}^{N}x_n+\lambda_\mu m\end{aligned} \tag{3.53}$$

とおき、平方完成を行う。

$$\begin{aligned}\ln p(\mu|\boldsymbol{X}) &= -\frac{1}{2}\left\{\hat{\lambda}_\mu\mu^2-2\tilde{m}\mu\right\}+\text{const.} \\ &= -\frac{1}{2}\hat{\lambda}_\mu\left\{\mu^2-2\frac{\tilde{m}}{\hat{\lambda}_\mu}\mu\right\}+\text{const.} \\ &= -\frac{1}{2}\hat{\lambda}_\mu\left\{\mu^2-2\frac{\tilde{m}}{\hat{\lambda}_\mu}\mu+\left(\frac{\tilde{m}}{\hat{\lambda}_\mu}\right)^2-\left(\frac{\tilde{m}}{\hat{\lambda}_\mu}\right)^2\right\}+\text{const.} \\ &= -\frac{1}{2}\hat{\lambda}_\mu\left\{\left(\mu-\frac{\tilde{m}}{\hat{\lambda}_\mu}\right)^2-\left(\frac{\tilde{m}}{\hat{\lambda}_\mu}\right)^2\right\}+\text{const.} \\ &= -\frac{1}{2}\left(\mu-\frac{\tilde{m}}{\hat{\lambda}_\mu}\right)^2\hat{\lambda}_\mu+\frac{\tilde{m}^2}{2\hat{\lambda}_\mu}+\text{const.}\end{aligned}$$

後の項は μ と無関係であるため const. に含めて

$$\begin{aligned}\hat{m} &= \frac{\tilde{m}}{\hat{\lambda}_\mu} \\ &= \frac{\lambda\sum_{n=1}^{N}x_n+\lambda_\mu m}{\hat{\lambda}_\mu}\end{aligned} \tag{3.54}$$

とおくと、事後分布 $p(\mu|\boldsymbol{X})$ は

$$\ln p(\mu|\boldsymbol{X}) = -\frac{1}{2}(\mu-\hat{m})^2\hat{\lambda}_\mu+\text{const.}=\ln\mathcal{N}(\mu|\hat{m},\hat{\lambda}_\mu^{-1}) \tag{3.51}$$

平均 $\hat{\mu}$、精度 $\hat{\lambda}_{\mu}^{-1}$ の 1 次元ガウス分布であることが確認できる。
　また事後分布のパラメータの計算式 (3.53)、(3.54) によって、ハイパーパラメータ $m,\ \lambda_{\mu}$ が更新される。

・予測分布の計算

　続いて未観測のデータ x_* に対する予測分布を求めていく。

　(観測データによる学習を行っていない) 事前分布 $p(\mu)$ を用いて、パラメータ μ を周辺化することで予測分布 $p(x_*)$ となる。

$$
\begin{aligned}
p(x_*) &= \int p(x_*|\mu)p(\mu)d\mu \\
&= \int \mathcal{N}(x_*|\mu,\lambda^{-1})\mathcal{N}(\mu|m,\lambda_{\mu}^{-1})d\mu
\end{aligned}
\tag{3.55}
$$

　この計算を直接するのではなく、より簡単に計算したい。そこで予測分布 $p(x_*)$ と事前分布 $p(\mu)$ の関係を考えてみる。ベイズの定理より

$$
p(\mu|x_*) = \frac{p(x_*|\mu)p(\mu)}{p(x_*)} \tag{3.56}
$$

という関係が成り立つことが分かる。両辺で対数をとり、予測分布に関して整理すると

$$
\begin{aligned}
\ln p(\mu|x_*) &= \ln p(x_*|\mu) + \ln p(\mu) - \ln p(x_*) \\
\ln p(x_*) &= \ln p(x_*|\mu) - \ln p(\mu|x_*) + \text{const.}
\end{aligned}
\tag{3.57}
$$

となる。x_* と関係のない $\ln p(\mu)$ を const. とおいている。
　つまり式 (3.57) の計算によっても予測分布を求められることが分かる。

　$p(\mu|x_*)$ は、1 つのデータ x_* が与えられた下での μ の条件付き分布である。つまり $p(\mu|x_*)$ は、N 個の観測データが与えられた下での事後分布 $p(\mu|\boldsymbol{X})$ の導出と同様の手順で求められる。
　従って事後分布のパラメータ (3.53)、(3.54) を用いると、$p(\mu|x_*)$ は $N=1$ より

$$
\ln p(\mu|x_*) = -\frac{1}{2}\left(\mu - \frac{\lambda x_* + \lambda_{\mu} m}{\lambda + \lambda_{\mu}}\right)^2 (\lambda + \lambda_{\mu}) + \text{const.} = \ln \mathcal{N}(\mu|m(x_*),(\lambda+\lambda_{\mu})^{-1}) \tag{3.58}
$$

平均 $m(x_*)$、精度 $\lambda + \lambda_{\mu}$ を持つ 1 次元ガウス分布となる。ただし

$$
m(x_*) = \frac{\lambda x_* + \lambda_{\mu} m}{\lambda + \lambda_{\mu}} \tag{3.59}
$$

とおく。

　$p(\mu|x_*)$ が求まったので、予測分布 $p(x_*)$ の計算に戻る。
　式 (3.58) を式 (3.57) に代入して、$x_*^2,\ x_*$ について整理すると

$$
\begin{aligned}
\ln p(x_*) &= \ln p(x_*|\mu) - \ln p(\mu|x_*) + \text{const.} && (3.57)\\
&= \ln \mathcal{N}(x_*|\mu, \lambda^{-1}) - \ln \mathcal{N}\left(\mu|m(x_*), (\lambda+\lambda_\mu)^{-1}\right) + \text{const.}\\
&= -\frac{1}{2}\left\{(x_*-\mu)^2\lambda + \ln 2\pi\lambda^{-1}\right\} + \frac{1}{2}\left\{\left(\mu - \frac{\lambda x_* + \lambda_\mu m}{\lambda+\lambda_\mu}\right)^2(\lambda+\lambda_\mu) + \ln 2\pi(\lambda+\lambda_\mu)^{-1}\right\} + \text{const.}\\
&= -\frac{1}{2}\left\{(x_*-\mu)^2\lambda - \left(\mu - \frac{\lambda x_* + \lambda_\mu m}{\lambda+\lambda_\mu}\right)^2(\lambda+\lambda_\mu)\right\} + \text{const.}\\
&= -\frac{1}{2}\left\{\lambda x_*^2 - 2\mu\lambda x_* + \mu^2\lambda - \mu^2(\lambda+\lambda_\mu) + 2\mu(\lambda x_* + \lambda_\mu m) - \frac{(\lambda x_* + \lambda_\mu m)^2}{\lambda+\lambda_\mu}\right\} + \text{const.}\\
&= -\frac{1}{2}\left\{\lambda x_*^2 - 2\mu\lambda x_* + 2\mu\lambda x_* - \frac{\lambda^2 x_*^2 + 2m\lambda\lambda_\mu x_*}{\lambda+\lambda_\mu}\right\} + \text{const.}\\
&= -\frac{1}{2}\left\{\frac{\lambda(\lambda+\lambda_\mu) - \lambda^2}{\lambda+\lambda_\mu}x_*^2 - \frac{2m\lambda\lambda_\mu}{\lambda+\lambda_\mu}x_*\right\} + \text{const.}\\
&= -\frac{1}{2}\left\{\frac{\lambda\lambda_\mu}{\lambda+\lambda_\mu}x_*^2 - \frac{2m\lambda\lambda_\mu}{\lambda+\lambda_\mu}x_*\right\} + \text{const.} && (3.60)
\end{aligned}
$$

となる。適宜 x_* と関係のない項を const. に含めている。

この式について

$$
\begin{aligned}
\lambda_* &= \frac{\lambda\lambda_\mu}{\lambda+\lambda_\mu} && (3.62\text{a})\\
\tilde{\mu}_* &= \frac{2m\lambda\lambda_\mu}{\lambda+\lambda_\mu}
\end{aligned}
$$

とおき、平方完成を行うと

$$
\begin{aligned}
\ln p(x_*) &= -\frac{1}{2}\{\lambda_* x_*^2 - \tilde{\mu}_* x_*\} + \text{const.}\\
&= -\frac{1}{2}\lambda_*\left\{x_*^2 - \frac{\tilde{\mu}_*}{\lambda_*}x_*\right\} + \text{const.}\\
&= -\frac{1}{2}\lambda_*\left\{x_*^2 - \frac{\tilde{\mu}_*}{\lambda_*}x_* + \left(\frac{\tilde{\mu}_*}{2\lambda_*}\right)^2 - \left(\frac{\tilde{\mu}_*}{2\lambda_*}\right)^2\right\} + \text{const.}\\
&= -\frac{1}{2}\lambda_*\left\{\left(x_*^2 - \frac{\tilde{\mu}_*}{2\lambda_*}\right)^2 - \left(\frac{\tilde{\mu}_*}{2\lambda_*}\right)^2\right\} + \text{const.}\\
&= -\frac{1}{2}\left(x_*^2 - \frac{\tilde{\mu}_*}{2\lambda_*}\right)^2\lambda_* + \frac{\tilde{\mu}_*^2}{4\lambda_*} + \text{const.}
\end{aligned}
$$

となる。後の項は x_* と無関係であるため const. に含めて

$$
\begin{aligned}
\mu_* &= \frac{\tilde{\mu}_*}{2\lambda_*}\\
&= \frac{1}{2}\frac{\lambda+\lambda_\mu}{\lambda\lambda_\mu}\frac{2m\lambda\lambda_\mu}{\lambda+\lambda_\mu}\\
&= m && (3.62\text{b})
\end{aligned}
$$

とおくと、予測分布 $p(x_*)$ は

$$\ln p(x_*) = -\frac{1}{2}(x_*^2 - \mu_*)^2\lambda_* + \text{const.} = \mathcal{N}(x_*|\mu_*, \lambda_*) \tag{3.61}$$

平均 μ_*、精度 λ_* の 1 次元ガウス分布となることが分かる。

ちなみに精度パラメータ λ_* の逆数である分散 $\sigma^2 = \lambda_*^{-1}$ に注目すると

$$\begin{aligned}
\lambda_*^{-1} &= \frac{\lambda + \lambda_\mu}{\lambda\lambda_\mu} \\
&= \frac{\lambda_\mu}{\lambda\lambda_\mu} + \frac{\lambda}{\lambda\lambda_\mu} \\
&= \frac{1}{\lambda} + \frac{1}{\lambda_\mu} \\
&= \lambda^{-1} + \lambda_\mu^{-1}
\end{aligned} \tag{3.63}$$

である。

精度パラメータ λ_* を構成するパラメータ $\lambda_\mu,\ m$ について、事後分布パラメータ (3.53)、(3.54) を用いると

$$\begin{aligned}
\hat{\lambda}_* &= \frac{\lambda\hat{\lambda}_\mu}{\lambda + \hat{\lambda}_\mu} \qquad (3.62\text{a}) \\
&= \frac{\lambda(N\lambda + \lambda_\mu)}{\lambda + N\lambda + \lambda_\mu} \\
&= \frac{(N\lambda + \lambda_\mu)\lambda}{(N+1)\lambda + \lambda_\mu}
\end{aligned}$$

また

$$\begin{aligned}
\hat{\mu}_* &= \hat{m} \qquad (3.62\text{b}) \\
&= \frac{\lambda\sum_{n=1}^N x_n + \lambda_\mu m}{\hat{\lambda}_\mu} \\
&= \frac{\lambda\sum_{n=1}^N x_n + \lambda_\mu m}{N\lambda + \lambda_\mu}
\end{aligned}$$

となり、(観測データ $\boldsymbol{X}$ によって学習した) 事後分布 $p(\mu|\boldsymbol{X})$ を用いた予測分布 $p(x_*|\boldsymbol{X})$

$$p(x_*|\boldsymbol{X}) = \mathcal{N}(x_*|\hat{\mu}_*, \hat{\lambda}_*)$$

が得られる。

・**R** でやってみよう

各確率分布に従いランダムに生成したデータを用いて、パラメータを推定してみましょう。

```r
# 利用パッケージ
library(tidyverse)
```

利用するパッケージを読み込みます。

・事後分布

```r
## パラメータの初期値を指定
# 観測モデルのパラメータ
mu_truth <- 25
lambda <- 0.01

# 事前分布のパラメータ
m <- 20
lambda_mu <- 0.001

# 試行回数
N <- 50
```

観測モデルの平均パラメータ μ を mu_truth とします。この値を推定するのが目的です。
また観測モデルの精度パラメータ λ を lambda とします。この値は推定の中で更新されません。

続いて事前分布のパラメータ (ハイパーパラメータの初期値) を指定します。平均 m を m、精度 λ_μ を lambda_mu とします。

生成するデータ数 N を N とします。

λ, λ_μ はどちらも分散の逆数であるため、値が大きくなるほど散らばり具合が小さくなります。

```r
# ガウス分布に従うデータを生成
x_n <- rnorm(n = N, mean = mu_truth, sd = sqrt(lambda^(-1)))
```

ガウス分布に従う乱数を発生させる関数 rnorm() を使って、ランダムにデータを生成します。
試行回数の引数 n には N、平均の引数 mean には mu_truth を指定します。
しかし標準偏差の引数 sd には lambda のままでは指定できません。そこで標準偏差 $\sigma = \sqrt{\lambda_\mu^{-1}}$ に置き換えるため、sqrt(lambda^(-1)) とします。

サンプルを確認してみましょう。

```r
# 観測データを確認
summary(x_n)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5338 17.6790 22.7886 23.5921 30.4795 45.8550
```

このデータを用いて事後分布のパラメータを計算します。

```r
# 事後分布のパラメータを計算
lambda_mu_hat <- N * lambda + lambda_mu
m_hat <- (lambda * sum(x_n) + lambda_mu * m) / lambda_mu_hat
```

事後分布のパラメータの計算式 (3.53)、(3.54) の計算を行い、$\hat{\lambda}_\mu$, $\hat{m}$ をそれぞれ `lambda_mu_hat`, `m_hat` とします。

`m_hat` の計算には、事前分布のパラメータ `lambda_mu` と事後分布のパラメータ `lambda_mu_hat` の両方を使うので注意しましょう。

```r
# 事後分布を計算
posterior_df <- tibble(
  mu = seq(
    m_hat - 3 * sqrt(lambda_mu_hat^(-1)),
    m_hat + 3 * sqrt(lambda_mu_hat^(-1)),
    by = 0.01
  ),  # 作図用の値
  density = dnorm(mu, mean = m_hat, sd = sqrt(lambda_mu_hat^(-1))) # 確率密度
)
```

μ が取り得る値を `seq()` を使って値を用意します。

ただし `mu_truth` の値によって作図に必要な範囲が変わります。ここでは μ を中心に標準偏差 σ の 3 倍をプロットすることにします。そこで `seq()` の第 1 引数 (最小値)、第 2 引数 (最大値) に、`mu_truth` から `3 * sqrt(lambda^(-1))` を引いた値と足した値をそれぞれ指定します。

`mu` の各値の確率密度は、ガウス分布の定義式 (2.64) で計算できます。しかし計算式が複雑なので、ここでは手を抜いてみましょう。

ガウス分布に従う確率密度を返す関数 `dnorm()` を使います。引数については `rnorm()` と同様です。ただしここでは事前分布のパラメータ `m` と `lambda_mu` を用います。

前節で扱った確率分布についてもこのような関数が用意されています。

計算結果は作図用にデータフレームにまとめておきます。推定結果を確認してみましょう。

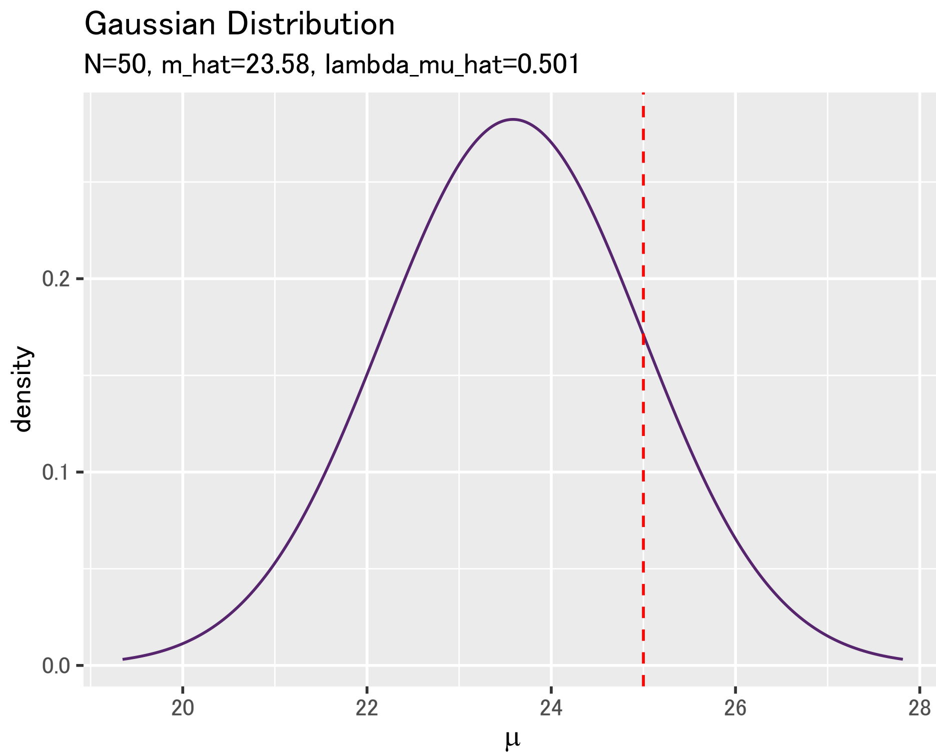
```r
head(posterior_df)
```

```
## # A tibble: 6 x 2
##      mu density
##   <dbl>   <dbl>
## 1  19.3 0.00314
## 2  19.4 0.00320
## 3  19.4 0.00327
## 4  19.4 0.00334
## 5  19.4 0.00341
## 6  19.4 0.00349
```

このデータフレームを用いてグラフを描きます。

```r
# 作図
ggplot(posterior_df, aes(mu, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = mu_truth),
             color = "red", linetype = "dashed") + # 垂直線
  labs(title = "Gaussian Distribution",
       subtitle = paste0("N=", N, ", m_hat=", round(m_hat, 2),
```

```r
                       ", lambda_mu_hat=", lambda_mu_hat),
       x = expression(mu)) # ラベル
```



ggplot2 パッケージを利用してグラフを作成します。

推定値は折れ線グラフ geom_line() を用いて描きます。
パラメータ μ の真の値を垂直線 geom_vline() で示します。
m_hat は 23.5849644 のように細かい値となるので、round() で丸めておきます。

・予測分布

```r
# 予測分布のパラメータを計算
lambda_hat <- lambda * lambda_mu_hat / (lambda + lambda_mu_hat)
mu_hat <- m_hat
```

予測分布のパラメータの計算式の計算を行い、$\hat{\lambda}_*$, $\hat{\mu}_*$ をそれぞれ lambda_hat, mu_hat とします。

```r
lambda_hat <- (N * lambda + lambda_mu) * lambda / ((N + 1) * lambda + lambda_mu)
mu_hat <- (lambda * sum(x_n) + lambda_mu * m) / (N * lambda + lambda_mu)
```

このように、観測データ x_n と観測モデルと事前分布のパラメータ lambda, lambda_mu を使って計算することもできます。

```r
# 予測分布の計算
predict_df <- tibble(
  x = seq(
```

```
    mu_hat - 3 * sqrt(lambda_hat^(-1)),
    mu_hat + 3 * sqrt(lambda_hat^(-1)),
    by = 0.01
  ),  # 作図用の値
  density = dnorm(x, mean = mu_hat, sd = sqrt(lambda_hat^(-1))) # 確率密度
)
```

x_* の分布も事後分布のときと同様にして計算します。

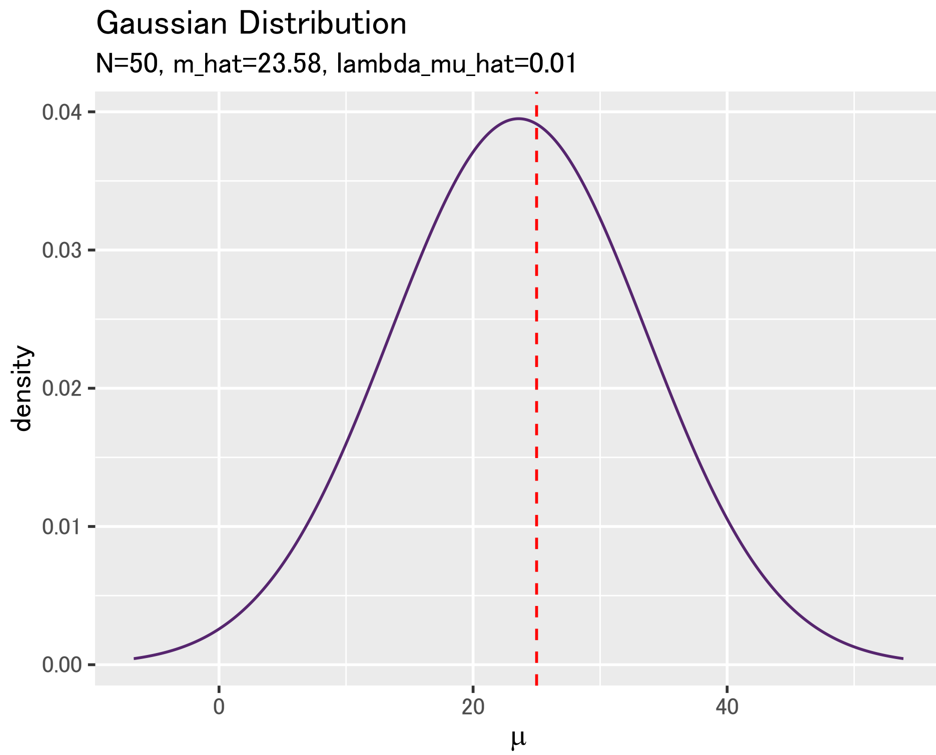
推定結果を確認してみましょう。

```
head(predict_df)
```

```
## # A tibble: 6 x 2
##       x  density
##   <dbl>    <dbl>
## 1 -6.71 0.000439
## 2 -6.70 0.000440
## 3 -6.69 0.000441
## 4 -6.68 0.000443
## 5 -6.67 0.000444
## 6 -6.66 0.000445
```

こちらも ggplot2 パッケージを利用して作図します。

```
# 作図
ggplot(predict_df, aes(x, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = mu_truth),
             color = "red", linetype = "dashed") + # 垂直線
  labs(title = "Gaussian Distribution",
       subtitle = paste0("N=", N, ", m_hat=", round(mu_hat, 2), ", lambda_mu_hat=", round(lambda_hat, 2
       x = expression(mu)) # ラベル
```



・おまけ

`gganimate` パッケージを利用して、パラメータの推定値の推移の gif 画像を作成するためのコードです。

```r
# 利用パッケージ
library(tidyverse)
library(gganimate)

## パラメータの初期値を指定
# 観測モデルのパラメータ
mu_truth <- 25
lambda <- 0.01

# 事前分布のパラメータ
m <- 20
lambda_mu <- 0.001

# 試行回数
N <- 50

# 事前分布を計算
posterior_df <- tibble(
  mu = seq(
    mu_truth - 2 * sqrt(lambda^(-1)),
    mu_truth + 2 * sqrt(lambda^(-1)),
    by = 0.01
  ),  # 作図用の値
```

```r
  density = dnorm(mu, mean = m, sd = sqrt(lambda_mu^(-1))), # 確率密度
  N = 0   # 試行回数
)

# 初期値による予測分布のパラメータを計算
lambda_star <- lambda * lambda_mu / (lambda + lambda_mu)
mu_star <- m

# 初期値による予測分布を計算
predict_df <- tibble(
  x = seq(
    mu_truth - 3 * sqrt(lambda^(-1)),
    mu_truth + 3 * sqrt(lambda^(-1)),
    by = 0.01
  ),  # 作図用の値
  density = dnorm(x, mean = mu_star, sd = sqrt(lambda_star^(-1))),  # 確率密度
  N = 0 # 試行回数
)

# パラメータを推定
x_n <- rep(0, N) # 受け皿
for(n in 1:N){

  # ガンマ分布に従うデータを生成
  x_n[n] <- rnorm(n = 1, mean = mu_truth, sd = sqrt(lambda^(-1)))

  # パラメータを更新
  lambda_mu_old <- lambda_mu
  lambda_mu <- 1 * lambda + lambda_mu
  m <- (lambda * x_n[n] + lambda_mu_old * m) / lambda_mu

  # 事後分布を計算
  tmp_posterior_df <- tibble(
    mu = seq(
      mu_truth - 2 * sqrt(lambda^(-1)),
      mu_truth + 2 * sqrt(lambda^(-1)),
      by = 0.01
    ),  # 作図用の値
    density = dnorm(mu, mean = m, sd = sqrt(lambda_mu^(-1))),  # 確率密度
    N = n   # 試行回数
  )

  # 予測分布のパラメータを計算
  lambda_star <- lambda * lambda_mu / (lambda + lambda_mu)
  mu_star <- m

  # 予測分布を計算
  tmp_predict_df <- tibble(
    x = seq(
      mu_truth - 3 * sqrt(lambda^(-1)),
      mu_truth + 3 * sqrt(lambda^(-1)),
      by = 0.01
    ),  # 作図用の値
```

```
    density = dnorm(x, mean = mu_star, sd = sqrt(lambda_star^(-1))),  # 確率密度
    N = n # 試行回数
  )

  # 推定結果を結合
  posterior_df <- rbind(posterior_df, tmp_posterior_df)
  predict_df <- rbind(predict_df, tmp_predict_df)
}

# 観測データを確認
summary(x_n)

## 事後分布
# 作図
posterior_graph <- ggplot(posterior_df, aes(mu, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = mu_truth),
             color = "red", linetype = "dashed") + # 垂直線
  transition_manual(N) + # フレーム
  labs(title = "Gaussian Distribution",
       subtitle = "N= {current_frame}",
       x = expression(mu)) # ラベル

# 描画
animate(posterior_graph)

## 予測分布
predict_graph <- ggplot(predict_df, aes(x, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = mu_truth),
             color = "red", linetype = "dashed") + # 垂直線
  transition_manual(N) + # フレーム
  labs(title = "Gaussian Distribution",
       subtitle = "N= {current_frame}",
       x = expression(x)) # ラベル

# 描画
animate(predict_graph)
```

異なる点のみを簡単に解説します。

各データによってどのように学習する (推定値が変化する) のかを確認するため、こちらは for ループで 1 データずつ処理します。
よって生成するデータ数として設定した `N` がイタレーション数になります。

パラメータの推定値について、$\hat{m}, \hat{\lambda}_\mu$ に対応する `m_hat`, `lambda_mu_hat` を新たに作るのではなく、`m`, `lambda_mu` をイタレーションごとに更新 (上書き) していきます。
それに伴い、事後分布のパラメータの計算式 (3.54)、(3.53) の $\sum_{n=1}^N$ の計算は、for ループによって N 回繰り返し `x_n[n]` を加えることで行います。n 回目のループ処理のときには、n-1 回分の `x_n[n]`(`lambda_mu` の場合は `lambda`) が既に `m` と `lambda_mu` に加えられているわけです。

$\hat{m}$ の計算には、更新前の λ_μ と更新後の $\hat{\lambda}_\mu$ の両方が必要です。そこで更新前の `lambda_mu` の値を

`lambda_mu_old`として残しておきます。そして`lambda_mu`を更新してから、`lambda_mu_old`と`lambda_mu`を使って`m`を更新します。

作図範囲を固定するため (更新されない) 観測モデルのパラメータを使います。

3.3.2 精度が未知の場合

・事後分布の計算

ガウス分布に従うと仮定する N 個の観測データ $\boldsymbol{X} = \{x_1, x_2, \cdots, x_N\}$ を用いて、精度パラメータ λ の事後分布を求めていく。

観測モデル $p(\boldsymbol{X}|\lambda)$ を用いて、観測データ $\boldsymbol{X}$ によって学習した λ の事後分布 $p(\lambda|\boldsymbol{X})$ はベイズの定理より

$$
\begin{aligned}
p(\lambda|\boldsymbol{X}) &= \frac{p(\boldsymbol{X}|\lambda)p(\lambda)}{p(\boldsymbol{X})} \\
&= \frac{\left\{\prod_{n=1}^N p(x_n|\lambda)\right\} p(\lambda)}{p(\boldsymbol{X})} \\
&= \frac{\left\{\prod_{n=1}^N \mathcal{N}(x_n|\mu, \lambda^{-1})\right\} \mathrm{Gam}(\lambda|a, b)}{p(\boldsymbol{X})} \\
&\propto \left\{\prod_{n=1}^N \mathcal{N}(x_n|\mu, \lambda^{-1})\right\} \mathrm{Gam}(\lambda|a, b)
\end{aligned} \tag{3.66}
$$

となる。分母の $p(\boldsymbol{X})$ は λ に影響しないため省略して、比例関係にのみ注目する。省略した部分については、最後に正規化することで対応できる。

次にこの分布の具体的な形状を明らかにしていく。対数をとって指数部分の計算を分かりやすくして進めると

$$
\begin{aligned}
\ln p(\lambda|\boldsymbol{X}) &= \ln \prod_{n=1}^N \mathcal{N}(x_n|\mu, \lambda^{-1}) + \ln \mathrm{Gam}(\lambda|a, b) - \ln p(\boldsymbol{X}) \\
&= \sum_{n=1}^N \ln \mathcal{N}(x_n|\mu, \lambda^{-1}) + \ln \mathrm{Gam}(\lambda|a, b) + \mathrm{const.} \\
&= \sum_{n=1}^N -\frac{1}{2}\left\{(x_n - \mu)^2\lambda + \ln \lambda^{-1} + \ln 2\pi\right\} + (a-1)\ln\lambda - b\lambda + \ln C_G(a, b) + \mathrm{const.} \\
&= -\frac{1}{2}\sum_{n=1}^N (x_n - \mu)^2\lambda + \frac{1}{2}N\ln\lambda + (a-1)\ln\lambda - b\lambda + \mathrm{const.} \\
&= \left(\frac{N}{2} + a - 1\right)\ln\lambda - \left\{\frac{1}{2}\sum_{n=1}^N (x_n - \mu)^2 + b\right\}\lambda + \mathrm{const.}
\end{aligned} \tag{3.67}
$$

【途中式の途中式】

0. 式 (3.66) の下から 2 行目の式を用いている。
1. 対数をとると積が和に変わる。また μ に影響しない項 $(-\ln p(\boldsymbol{X}))$ を const. とおく。
2. $\mathcal{N}(x_n|\mu, \lambda^{-1})$, $\mathrm{Gam}(\lambda|a, b)$ に、それぞれ対数をとった 1 次元ガウス分布の定義式 (2.65)、対数をとったガンマ分布の定義式 (2.58) を用いて具体的な確率分布を代入する。
3. $\sum_{n=1}^{N}$ に関する括弧を展開し、λ と関係のない項を const. に含める。$-\ln x^{-1} = -(-1)\ln x = \ln x$ である。
4. $\ln \lambda$, λ の項をそれぞれまとめて式を整理する。

となる。

式 (3.67) について

$$\begin{aligned}\hat{a} &= \frac{N}{2} + a \\ \hat{b} &= \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2 + b\end{aligned} \tag{3.69}$$

とおくと、事後分布 $p(\lambda|\boldsymbol{X})$ は

$$\ln p(\lambda|\boldsymbol{X}) = (\hat{a} - 1)\ln \lambda - \hat{b}\lambda + \ln C_G(\hat{a}, \hat{b}) = \ln \mathrm{Gam}(\lambda|\hat{a}, \hat{b}) \tag{3.68}$$

パラメータ $\hat{a}$, $\hat{b}$ を持つガンマ分布であることが分かる。

また事後分布のパラメータの計算式 (3.69) によって、ハイパーパラメータ a, b が更新される。

・予測分布の計算

続いて未観測のデータ x_* に対する予測分布を求めていく。

(観測データによる学習を行っていない) 事前分布 $p(\lambda)$ を用いて、パラメータ λ を周辺化することで予測分布 $p(x_*)$ となる。

$$\begin{aligned}p(x_*) &= \int p(x_*|\lambda)p(\lambda)d\lambda \\ &= \int \mathcal{N}(x_*|\mu, \lambda^{-1})\mathrm{Gam}(\lambda|a, b)d\lambda\end{aligned} \tag{3.70}$$

3.3.1 項のときと同様に、この計算を直接するのではなく予測分布 $p(x_*)$ と事前分布 $p(\lambda)$ の関係を考えてみる。ベイズの定理より

$$p(\lambda|x_*) = \frac{p(x_*|\lambda)p(\lambda)}{p(x_*)}$$

という関係が成り立つことが分かる。両辺で対数をとり、予測分布に関して整理すると

$$
\begin{aligned}
\ln p(\lambda|x_*) &= \ln p(x_*|\lambda) + \ln p(\lambda) - \ln p(x_*) \\
\ln p(x_*) &= \ln p(x_*|\lambda) - \ln p(\lambda|x_*) + \text{const.}
\end{aligned} \tag{3.72}
$$

となる。x_* と関係のない $\ln p(\lambda)$ を const. とおいている。
つまり式 (3.72) の計算によっても予測分布を求められることが分かる。

$p(\lambda|x_*)$ は、1 つのデータ x_* が与えられた下での λ の条件付き分布である。つまり $p(\lambda|x_*)$ は、N 個の観測データが与えられた下での事後分布 $p(\lambda|\boldsymbol{X})$ の導出と同様の手順で求められる。
従って事後分布のパラメータ (3.69) を用いると、$p(\lambda|x_*)$ は $N = 1$ より

$$
\ln p(\lambda|x_*) = \left(\frac{1}{2} + a - 1\right)\ln\lambda - \left\{\frac{1}{2}(x_* - \mu)^2 + b\right\}\lambda + \ln C_G\left(\frac{1}{2} + a, b(x_*)\right) = \ln \mathrm{Gam}\left(\lambda\middle|\frac{1}{2} + a, b(x_*)\right) \tag{3.73}
$$

パラメータ $\frac{1}{2} + a,\ b(x_*)$ を持つガンマ分布となる。ただし

$$
b(x_*) = \frac{1}{2}(x_* - \mu)^2 + b \tag{3.74}
$$

とおく。

$p(\lambda|x_*)$ が求まったので、予測分布 $p(x_*)$ の計算に戻る。
式 (3.73) を式 (3.72) に代入すると

$$
\begin{aligned}
\ln p(x_*) &= \ln p(x_*|\lambda) - \ln p(\lambda|x_*) + \text{const.} && (3.72)\\
&= \ln \mathcal{N}(x_*|\mu, \lambda^{-1}) - \ln \mathrm{Gam}\left(\lambda\middle|\frac{1}{2} + a, b(x_*)\right) + \text{const.}\\
&= -\frac{1}{2}\{(x_* - \mu)^2\lambda + \ln\lambda^{-1} + \ln 2\pi\}\\
&\qquad - \left(\frac{1}{2} + a - 1\right)\ln\lambda + \left\{\frac{1}{2}(x_* - \mu)^2 + b\right\}\lambda - \ln C_G\left(\frac{1}{2} + a, b(x_*)\right) + \text{const.}\\
&= -\frac{1}{2}(x_* - \mu)^2\lambda + \frac{1}{2}(x_* - \mu)^2\lambda\\
&\qquad - \left(\frac{1}{2} + a\right)\ln\left\{\frac{1}{2}(x_* - \mu)^2 + b\right\} + \ln\Gamma\left(\frac{1}{2} + a\right) + \text{const.}\\
&= -\left(\frac{1}{2} + a\right)\ln\left\{\frac{1}{2b}(x_* - \mu)^2 + 1\right\}b + \text{const.}\\
&= -\left(\frac{1}{2} + a\right)\ln\left\{\frac{1}{2b}(x_* - \mu)^2 + 1\right\} - \left(\frac{1}{2} + a\right)\ln b + \text{const.}\\
&= -\frac{2a + 1}{2}\ln\left\{1 + \frac{1}{2b}(x - \mu)^2\right\} + \text{const.} && (3.75)
\end{aligned}
$$

となる。5 行目から 6 行目では、$a\ln xy = a(\ln x + \ln y) = a\ln x + a\ln y$ の式変形を行っている。また適宜 x_* と関係のない項を const. にまとめている。
この式の形状の確率分布はスチューデントの t 分布と呼ばれる。

この式について、$\frac{1}{2b} = \frac{a}{2ab}$ として

$$\begin{aligned} \mu_s &= \mu \\ \lambda_s &= \frac{a}{b} \\ \nu_s &= 2a \end{aligned} \tag{3.79}$$

とおくと、予測分布 $p(x_*)$ は

$$\ln p(x_*) = -\frac{\nu_s + 1}{2} \ln \left\{ 1 + \frac{\lambda_s}{\nu_s}(x - \mu_s)^2 \right\} + \mathrm{const.} = \ln \mathrm{St}(x_* | \mu_s, \lambda_s, \nu_s) \tag{3.78}$$

パラメータ $\mu_s,\ \lambda_s,\ \nu_s$ を持つスチューデントの t 分布となることが分かる。

予測分布のパラメータ $\lambda_s,\ \nu_s$ を構成する事前分布のパラメータ $a,\ b$ について、事後分布のパラメータ (3.69) を用いると

$$\begin{aligned} \hat{\lambda}_s &= \frac{\hat{a}}{\hat{b}} \\ &= \frac{\frac{N}{2} + a}{\frac{1}{2}\sum_{n=1}^N (x_n - \mu)^2 + b} \\ &= \frac{N + 2a}{\sum_{n=1}^N (x_n - \mu)^2 + 2b} \end{aligned}$$

また

$$\begin{aligned} \hat{\nu}_s &= 2\hat{a} \\ &= N + 2a \end{aligned}$$

となり、(観測データ $\boldsymbol{X}$ によって学習した) 事後分布 $p(\lambda | \boldsymbol{X})$ を用いた予測分布 $p(x_* | \boldsymbol{X})$

$$\ln p(x_* | \boldsymbol{X}) = -\frac{\hat{\nu}_s + 1}{2} \ln \left\{ 1 + \frac{\hat{\lambda}_s}{\hat{\nu}_s}(x - \mu_s)^2 \right\} + \mathrm{const.} = \ln \mathrm{St}(x_* | \mu_s, \hat{\lambda}_s, \hat{\nu}_s)$$

が得られる。

・スチューデントの t 分布

$$\mathrm{St}(x | \mu_s, \lambda_s, \nu_s) = \frac{\Gamma(\frac{\nu_s + 1}{2})}{\Gamma(\frac{\nu_s}{2})} \left(\frac{\lambda_s}{\pi \nu_s} \right)^{\frac{1}{2}} \left\{ 1 + \frac{\lambda_s}{\nu_s}(x - \mu_s)^2 \right\}^{-\frac{\nu_s + 1}{2}} \tag{3.76}$$

対数をとって、確率変数 x と関わらない項を const. とおくと

$$\ln \mathrm{St}(x | \mu_s, \lambda_s, \nu_s) = -\frac{\nu_s + 1}{2} \ln \left\{ 1 + \frac{\lambda_s}{\nu_s}(x - \nu_s)^2 \right\} + \mathrm{const.} \tag{3.77}$$

である。

・R でやってみよう

各確率分布に従いランダムに生成したデータを用いて、パラメータを推定してみましょう。

```r
# 利用パッケージ
library(tidyverse)
```

利用するパッケージを読み込みます。

・事後分布

```r
## パラメータの初期値を指定
# 観測モデルのパラメータ
lambda_truth <- 0.01
mu <- 25

# 事前分布のパラメータ
a <- 1
b <- 1

# 試行回数
N <- 50
```

観測モデルの精度パラメータ λ を `lambda_truth` とします。この値を推定するのが目的です。
観測モデルの平均パラメータ μ を `mu` とします。この値は推定の中で更新されません。

続いて事前分布のパラメータ (ハイパーパラメータの初期値) を指定します。a を `a`、b を `b` とします。

生成するデータ数 N を `N` とします。

λ は分散の逆数であるため、値が大きくなるほど散らばり具合が小さくなります。

```r
# ガウス分布に従うデータを生成
x_n <- rnorm(n = N, mean = mu, sd = sqrt(lambda_truth^(-1)))
```

ガウス分布に従う乱数を発生させる関数 `rnorm()` を使って、ランダムにデータを生成します。
試行回数の引数 `n` には `N`、平均の引数 `mean` には `mu` を指定します。
しかし標準偏差の引数 `sd` には `lambda_truth` のままでは指定できません。そこで標準偏差 $\sigma = \sqrt{\lambda^{-1}}$ に置き換えるため、`sqrt(lambda_truth^(-1))` とします。

サンプルを確認してみましょう。

```r
# 観測データを確認
summary(x_n)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.807  19.867  25.009  25.369  31.711  41.597
```

このデータを用いて事後分布のパラメータを計算します。

```
# 事後分布のパラメータを計算
a_hat <- N / 2 + a
b_hat <- sum((x_n - mu)^2) / 2 + b
```

事後分布のパラメータの計算式 (3.69) の計算を行い、$\hat{a}, \hat{b}$ をそれぞれ a_hat, b_hat とします。

```
# 事後分布を計算
posterior_df <- tibble(
  lambda = seq(0, 3 * lambda_truth, by = 0.0001),  # 作図用の値
  C_G = a_hat * log(b_hat) - lgamma(a_hat),  # 正規化項（対数）
  density = exp(C_G + (a_hat - 1) * log(lambda) - b_hat * lambda) # 確率密度
)
```

λ が取り得る値を seq() を使って用意します。

ただし lambda_truth の値によって必要な範囲が変わるので、ここでは 0 から lambda_truth の 3 倍までとします。そこで seq() の第 1 引数 (最小値) に 0、第 2 引数 (最大値) に 3 * mu_truth を指定します。

lambda の各値に対してガンマ分布の定義式 (2.57) の計算を行い、確率密度を求めます。ただし値が大きくなると gamma() で計算できなくなるため、対数をとって計算することにします。なので最後に exp() で値を戻します。

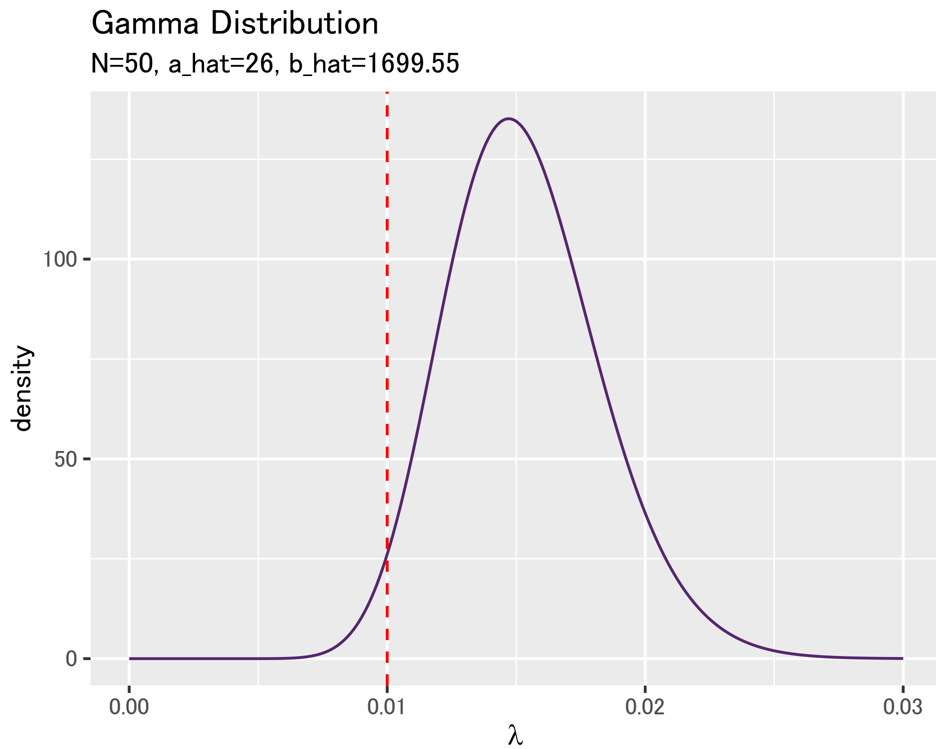
計算結果は作図用にデータフレームにまとめておきます。推定結果を確認してみましょう。

```
head(posterior_df)
```

```
## # A tibble: 6 x 3
##   lambda   C_G  density
##    <dbl> <dbl>    <dbl>
## 1 0       135. 0.
## 2 0.0001  135. 5.30e-42
## 3 0.0002  135. 1.50e-34
## 4 0.0003  135. 3.20e-30
## 5 0.0004  135. 3.58e-27
## 6 0.0005  135. 8.00e-25
```

このデータフレームを用いてグラフを描きます。

```
# 作図
ggplot(posterior_df, aes(lambda, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = lambda_truth),
             color = "red", linetype = "dashed") + # 垂直線
  labs(title = "Gamma Distribution",
       subtitle = paste0("N=", N, ", a_hat=", a_hat, ", b_hat=", round(b_hat, 2)),
       x = expression(lambda)) # ラベル
```



ggplot2 パッケージを利用してグラフを作成します。

推定値は折れ線グラフ geom_line() を用いて描きます。
パラメータ λ の真の値を垂直線 geom_vline() で示します。
事前分布のパラメータ $\hat{b}$ は 1699.5537105 のように細かい値となるので、round() で丸めておきます。

・予測分布

```
# 予測分布のパラメータを計算
mu_s <- mu
lambda_s_hat <- a_hat / b_hat
nu_s_hat <- 2 * a_hat
```

予測分布のパラメータの計算式の計算を行い、μ_s, $\hat{\lambda}_s$, $\hat{\mu}_s$ をそれぞれ mu_s, lambda_s_hat, nu_s_hat とします。

```
lambda_s_hat <- (N + 2 * a) / (sum((x_n - mu)^2) + 2 * b)
nu_s_hat <- N + 2 * a
```

このように、観測データ x_n と事前分布のパラメータ a, b を使って計算することもできます。

```
# 予測分布を計算
predict_df <- tibble(
  x = seq(-mu_s, 3 * mu_s, by = 0.01),  # 作図用の値
  C_St = lgamma((nu_s_hat + 1) / 2) - lgamma(nu_s_hat / 2),  # 正規化項（対数）
  term1 = log(lambda_s_hat / pi / nu_s_hat) / 2,
```

```
  term2 = - (nu_s_hat + 1) / 2 * log(1 + lambda_s_hat / nu_s_hat * (x - mu_s)^2),
  density = exp(C_St + term1 + term2) # 確率密度
)
```

x_* が取り得る値を seq() を使って用意します。

ただし mu_s の値によって作図に必要な範囲が変わります。ここでは $-\mu_s$ から $3\mu_s$ までとします。必要に応じて調整してください。特にこのままですと、mu_s が負の値をとるとエラーとなります。

x の各値に対してスチューデントの t 分布の定義式 (3.76) の計算を行い、確率密度を求めます。ちなみにこの pi は、前節で出てきたパラメータではなく円周率のことです。ご注意ください。

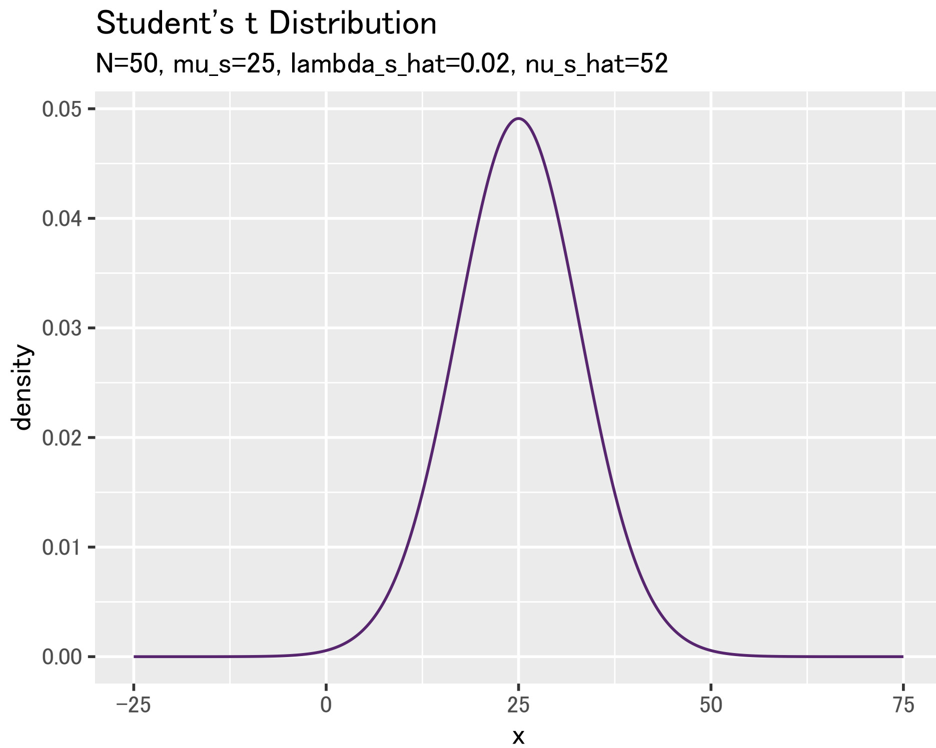
推定結果を確認してみましょう。

```
head(predict_df)
```

```
## # A tibble: 6 x 5
##       x  C_St term1 term2      density
##   <dbl> <dbl> <dbl> <dbl>        <dbl>
## 1 -25    1.62 -4.64 -14.6 0.0000000222
## 2 -25.0  1.62 -4.64 -14.6 0.0000000223
## 3 -25.0  1.62 -4.64 -14.6 0.0000000224
## 4 -25.0  1.62 -4.64 -14.6 0.0000000225
## 5 -25.0  1.62 -4.64 -14.6 0.0000000226
## 6 -25.0  1.62 -4.64 -14.6 0.0000000227
```

こちらも ggplot2 パッケージを利用して作図します。

```
# 作図
ggplot(predict_df, aes(x, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  labs(title = "Student's t Distribution",
       subtitle = paste0("N=", N, ", mu_s=", mu_s,
                         ", lambda_s_hat=", round(lambda_s_hat, 2), ", nu_s_hat=", nu_s_hat)) # ラベル
```



・おまけ

`gganimate` パッケージを利用して、パラメータの推定値の推移の gif 画像を作成するためのコードです。

```
# 利用パッケージ
library(tidyverse)
library(gganimate)

## パラメータの初期値を指定
# 観測モデルのパラメータ
lambda_truth <- 0.01
mu <- 25

# 事前分布のパラメータ
a <- 1
b <- 1

# 試行回数
N <- 50

# 事前分布を計算
posterior_df <- tibble(
  lambda = seq(0, 3 * lambda_truth, by = 0.0001),  # 作図用の値
  C_G = a * log(b) - lgamma(a),  # 正規化項（対数）
  density = exp(C_G + (a - 1) * log(lambda) - b * lambda), # 確率密度
  N = 0  # 試行回数
)
```

```r

# 初期値による予測分布のパラメータを計算
mu_s <- mu
lambda_s <- a / b
nu_s <- 2 * a

# 初期値による予測分布を計算
predict_df <- tibble(
  x = seq(-mu_s, 3 * mu_s, by = 0.01),  # 作図用の値
  C_St = lgamma((nu_s + 1) / 2) - lgamma(nu_s / 2),  # 正規化項（対数）
  term1 = log(lambda_s / pi / nu_s) / 2,
  term2 = - (nu_s + 1) / 2 * log(1 + lambda_s / nu_s * (x - mu_s)^2),
  density = exp(C_St + term1 + term2), # 確率密度
  N = 0 # 試行回数
)

# パラメータを推定
x_n <- rep(0, N) # 受け皿
for(n in 1:N){

  # ガウス分布に従うデータを生成
  x_n[n] <- rnorm(n = 1, mean = mu, sd = sqrt(lambda_truth^(-1)))

  # ハイパーパラメータを更新
  a <- 1 / 2 + a
  b <- (x_n[n] - mu)^2 / 2 + b

  # 事後分布を計算
  tmp_posterior_df <- tibble(
    lambda = seq(0, 3 * lambda_truth, by = 0.0001),  # 作図用の lambda
    C_G = a * log(b) - lgamma(a),  # 正規化項（対数）
    density = exp(C_G + (a - 1) * log(lambda) - b * lambda), # 確率密度
    N = n  # 試行回数
  )

  # 予測分布のパラメータを更新
  mu_s <- mu
  lambda_s <- a / b
  nu_s <- 2 * a

  # 予測分布を計算
  tmp_predict_df <- tibble(
    x = seq(-mu_s, 3 * mu_s, by = 0.01),  # 作図用の値
    C_St = lgamma((nu_s + 1) / 2) - lgamma(nu_s / 2),  # 正規化項（対数）
    term1 = log(lambda_s / pi / nu_s) / 2,
    term2 = - (nu_s + 1) / 2 * log(1 + lambda_s / nu_s * (x - mu_s)^2),
    density = exp(C_St + term1 + term2), # 確率密度
    N = n # 試行回数
  )

  # 結果を結合
  posterior_df <- rbind(posterior_df, tmp_posterior_df)
  predict_df <- rbind(predict_df, tmp_predict_df)
}
```

```r

# 観測データを確認
summary(x_n)

## 事後分布
# 作図
posterior_graph <- ggplot(posterior_df, aes(lambda, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = lambda_truth),
             color = "red", linetype = "dashed") + # 垂直線
  transition_manual(N) +  # フレーム
  labs(title = "Gamma Distribution",
       subtitle = "N= {current_frame}",
       x = expression(lambda)) # ラベル

# 描画
animate(posterior_graph)

## 予測分布
# 作図
predict_graph <- ggplot(predict_df, aes(x, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  transition_manual(N) + # フレーム
  labs(title = "Student's t Distribution",
       subtitle = "N= {current_frame}") # ラベル

# 描画
animate(predict_graph)
```

異なる点のみを簡単に解説します。

各データによってどのように学習する (推定値が変化する) のかを確認するため、こちらは for ループで 1 データずつ処理します。
よって生成するデータ数として設定した `N` がイタレーション数になります。

パラメータの推定値について、$\hat{a}, \hat{b}$ に対応する `a_hat`, `b_hat` 新たに作るのではなく、`a`, `b` をイタレーションごとに更新 (上書き) していきます。
それに伴い、事後分布のパラメータの計算式 (3.69) の $\sum_{n=1}^N$ の計算は、for ループによって N 回繰り返し `(x_n[n] - mu)^2` を加えることで行います。n 回目のループ処理のときには、n-1 回分の `(x_n[n] - mu)^2`(a の場合は `1 / 2`) が既に `a` と `b` に加えられているわけです。

3.3.3 平均・精度が未知の場合

ガウス分布に従うと仮定する N 個の観測データ $\boldsymbol{X} = \{x_1, x_2, \cdots, x_N\}$ を用いて、平均パラメータ μ と精度パラメータ λ の事後分布を求めていく。

観測データ $\boldsymbol{X}$、平均パラメータ μ、精度パラメータ λ の同時分布は次のように分解できる。

$$p(\boldsymbol{X}, \mu, \lambda) = p(\mu|\lambda, \boldsymbol{X})p(\lambda|\boldsymbol{X})p(\boldsymbol{X}) \tag{3.84}$$

また事前分布 (パラメータの同時分布)$p(\mu,\lambda)$ をガウス・ガンマ分布と仮定すると、同じ形式の事後分布となることが知られている。

$$
\begin{aligned}
p(\mu,\lambda) &= \mathrm{NB}(\mu,\lambda|m,\beta,a,b)\\
&= \mathcal{N}(\mu|m,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b)\\
&= \frac{1}{\sqrt{2\pi(\beta\lambda)^{-1}}}\exp\left\{-\frac{(\mu-m)^2}{2(\beta\lambda)^{-1}}\right\}C_G(a,b)\lambda^{a-1}\exp(-b\lambda)
\end{aligned}
\tag{3.81}
$$

対数をとると

$$
\begin{aligned}
\ln p(\mu,\lambda) &= \ln\mathrm{NB}(\mu,\lambda|m,\beta,a,b)\\
&= \ln\mathcal{N}(\mu|m,(\beta\lambda)^{-1}) + \ln\mathrm{Gam}(\lambda|a,b)\\
&= -\frac{1}{2}\left\{(\mu-m)^2\beta\lambda + \ln(\beta\lambda)^{-1} + \ln 2\pi\right\} + (a-1)\ln\lambda - b\lambda + \ln C_G(a,b)
\end{aligned}
$$

となる。

・ガウス・ガンマ分布

$$
\begin{aligned}
\mathrm{NB}(\mu,\lambda|m,\beta,a,b) &= \mathcal{N}(\mu|m,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b)\\
&= \frac{1}{\sqrt{2\pi(\beta\lambda)^{-1}}}\exp\left\{-\frac{(\mu-m)^2}{2(\beta\lambda)^{-1}}\right\}C_G(a,b)\lambda^{a-1}\exp(-b\lambda)
\end{aligned}
\tag{3.81}
$$

対数をとると

$$
\begin{aligned}
\ln\mathrm{NB}(\mu,\lambda|m,\beta,a,b) &= \ln\mathcal{N}(\mu|m,(\beta\lambda)^{-1}) + \ln\mathrm{Gam}(\lambda|a,b)\\
&= -\frac{1}{2}\left\{(\mu-m)^2\beta\lambda + \ln(\beta\lambda)^{-1} + \ln 2\pi\right\} + (a-1)\ln\lambda - b\lambda + \ln C_G(a,b)
\end{aligned}
$$

となる。

・平均 μ の事後分布の計算

まずは μ の事後分布 $p(\mu|\lambda,\boldsymbol{X})$ を導出する。

図 3.5 のグラフィカルモデルでパラメータの依存関係を確認すると、3.3.1 項と同様の手順で求められることが分かる。よって μ の事後分布のパラメータ (3.53)、(3.54) について、$\beta\lambda=\lambda_\mu$ で置き換えると

$$
\hat{\lambda}_\mu = N\lambda + \lambda_\mu \tag{3.53}
$$

$$
\hat{\beta}\lambda = N\lambda + \beta\lambda
$$

$$
\hat{\beta} = N + \beta \tag{3.83a}
$$

また

$$
\begin{aligned}
\hat{m} &= \frac{\lambda \sum_{n=1}^{N} x_n + \lambda_\mu m}{\hat{\lambda}_\mu} \qquad (3.54) \\
&= \frac{\lambda \sum_{n=1}^{N} x_n + \beta\lambda m}{N\lambda + \beta\lambda} \\
&= \frac{1}{\hat{\beta}} \left(\sum_{n=1}^{N} x_n + \beta m \right) \qquad (3.83b)
\end{aligned}
$$

となる。従って μ の事後分布 $p(\mu|\lambda, \boldsymbol{X})$ は

$$
p(\mu|\lambda, \boldsymbol{X}) = \mathcal{N}(\mu|\hat{m}, (\hat{\beta}\lambda)^{-1}) \qquad (3.82)
$$

平均パラメータ μ、精度パラメータ $(\hat{\beta}\lambda)^{-1}$ を持つ 1 次元ガウス分布となる。

・精度 λ の事後分布の計算

次に λ の事後分布 $p(\lambda|\boldsymbol{X})$ を導出する。

λ の事後分布 $p(\lambda|\boldsymbol{X})$ は式 (3.84) より

$$
\begin{aligned}
p(\lambda|\boldsymbol{X}) &= \frac{p(\boldsymbol{X}, \mu, \lambda)}{p(\mu|\lambda, \boldsymbol{X})p(\boldsymbol{X})} \\
&= \frac{p(\boldsymbol{X}|\mu, \lambda)p(\mu, \lambda)}{p(\mu|\lambda, \boldsymbol{X})p(\boldsymbol{X})} \\
&= \frac{\prod_{n=1}^{N} p(x_n|\mu, \lambda)p(\mu, \lambda)}{p(\mu|\lambda, \boldsymbol{X})p(\boldsymbol{X})} \\
&\propto \frac{\prod_{n=1}^{N} p(x_n|\mu, \lambda)p(\mu, \lambda)}{p(\mu|\lambda, \boldsymbol{X})} \qquad (3.85)
\end{aligned}
$$

となる。分母の $p(\boldsymbol{X})$ は λ に影響しないため省略して、比例関係にのみ注目する。省略した部分については、最後に正規化することで対応できる。

次にこの分布の具体的な形状を明らかにしていく。対数をとって指数部分の計算を分かりやすくして進めると

$$
\begin{aligned}
\ln p(\lambda|\boldsymbol{X}) &= \sum_{n=1}^{N} \ln p(x_n|\mu,\lambda) + \ln p(\mu,\lambda) - \ln p(\mu|\lambda,\boldsymbol{X}) + \text{const.} \\
&= \sum_{n=1}^{N} \ln \mathcal{N}(x_n|\mu,\lambda^{-1}) + \ln \mathcal{N}(\mu|m,(\beta\lambda)^{-1}) + \ln \mathrm{Gam}(\lambda|a,b) - \ln \mathcal{N}(\mu|\hat{m},(\hat{\beta}\lambda)^{-1}) + \text{const.} \\
&= -\frac{1}{2}\sum_{n=1}^{N} \left\{ (x_n-\mu)^2\lambda + \ln\lambda^{-1} + \ln 2\pi \right\} \\
&\quad - \frac{1}{2}\left\{ (\mu-m)^2\beta\lambda + \ln(\beta\lambda)^{-1} + \ln 2\pi \right\} \\
&\quad + (a-1)\ln\lambda - b\lambda + \ln C_G(a,b) \\
&\quad + \frac{1}{2}\left\{ (\mu-\hat{m})^2\hat{\beta}\lambda + \ln(\hat{\beta}\lambda)^{-1} + \ln 2\pi \right\} + \text{const.} \\
&= -\frac{1}{2}\sum_{n=1}^{N} x_n^2\lambda + \sum_{n=1}^{N} x_n\mu\lambda - \frac{N}{2}\mu^2\lambda \\
&\quad - \frac{1}{2}\mu^2\beta\lambda + \mu m\beta\lambda - \frac{1}{2}m^2\beta\lambda \\
&\quad + (a-1)\ln\lambda - b\lambda + a\ln b - \ln\Gamma(a) \\
&\quad + \frac{1}{2}\mu^2(N+\beta)\lambda - \mu\frac{1}{N+\beta}\left(\sum_{n=1}^{N} x_n + \beta m\right)(N+\beta)\lambda + \frac{1}{2}\hat{m}^2\hat{\beta}\lambda + \text{const.} \\
&= -\frac{1}{2}\sum_{n=1}^{N} x_n^2\lambda - \frac{1}{2}m^2\beta\lambda + (a-1)\ln\lambda - b\lambda + \frac{1}{2}\hat{m}^2\hat{\beta}\lambda + \text{const.} \\
&= \left(\frac{N}{2} + a - 1\right)\ln\lambda - \left\{ \frac{1}{2}\left(\sum_{n=1}^{N} x_n^2 + \beta m^2 - \hat{\beta}\hat{m}^2\right) + b \right\}\lambda + \text{const.}
\end{aligned}
\tag{3.86}
$$

となる。適宜 λ と関係のない項を const. に含めている。

この式について

$$
\begin{aligned}
\hat{a} &= \frac{N}{2} + a \\
\hat{b} &= \frac{1}{2}\left(\sum_{n=1}^{N} x_n^2 + \beta m^2 - \hat{\beta}\hat{m}^2\right) + b
\end{aligned}
\tag{3.88}
$$

とおくと、λ の事後分布 $p(\lambda|\boldsymbol{X})$ は

$$
\ln p(\lambda|\boldsymbol{X}) = (\hat{a}-1)\ln\lambda - \hat{b}\lambda + \text{const.} = \ln \mathrm{Gam}(\lambda|\hat{a},\hat{b}) \tag{3.87}
$$

パラメータ $\hat{a}$, $\hat{b}$ を持つガンマ分布となる。
また事後分布のパラメータの計算式 (3.88) によって、ハイパーパラメータが更新される。

・予測分布の計算

続いて未観測のデータ x_* に対する予測分布を求めていく。

(観測データによる学習を行っていない) 事前分布 $p(\mu, \lambda)$ を用いて、パラメータ μ, λ を周辺化することで予測分布 $p(x_*)$ となる。

$$p(x_*) = \iint p(x_*|\mu, \lambda)p(\mu, \lambda)d\mu d\lambda \tag{3.89}$$

ここでもこの計算を直接するのではなく、予測分布 $p(x_*)$ と事前分布 $p(\mu, \lambda)$ の関係を考えてみる。ベイズの定理より

$$p(\mu, \lambda|x_*) = \frac{p(x_*|\mu, \lambda)p(\mu, \lambda)}{p(x_*)}$$

という関係が成り立つことが分かる。両辺で対数をとり、予測分布に関して整理すると

$$\begin{aligned} \ln p(\mu, \lambda|x_*) &= \ln p(x_*|\mu, \lambda) + \ln p(\mu, \lambda) - \ln p(x_*) \\ \ln p(x_*) &= \ln p(x_*|\mu, \lambda) - p(\mu, \lambda|x_*) + \text{const.} \end{aligned} \tag{3.90}$$

となる。x_* と関係のない $\ln p(\mu, \lambda)$ を const. とおいている。

つまり式 (3.90) の計算によっても予測分布を求められることが分かる。

$p(\mu, \lambda|x_*)$ は、1 つのデータ x_* が与えられた下での μ, λ の条件付き同時分布である。つまり $p(\mu, \lambda|x_*)$ は、N 個の観測データが与えられた下での事後分布 $p(\mu|\boldsymbol{X})$, $p(\lambda|\boldsymbol{X})$ の導出と同様の手順で求められる。

従って、事後分布のパラメータ (3.82)、(3.87) を用いると、$p(\mu, \lambda|x_*)$ は $N = 1$ より

$$\begin{aligned} p(\mu, \lambda|x_*) &= p(\mu|\lambda, x_*)p(\lambda|x_*) \\ &= \mathcal{N}(\mu|m(x_*), \{(1+\beta)\lambda\}^{-1})\text{Gam}\left(\lambda\middle|\frac{1}{2} + a, b(x_*)\right) \end{aligned} \tag{3.91}$$

となる。ただし

$$\begin{aligned} m(x_*) &= \frac{x_* + \beta m}{1 + \beta} \\ b(x_*) &= \frac{\beta}{2(1+\beta)}(x_* - m)^2 + b \end{aligned} \tag{3.92}$$

とおく。

$p(\mu, \lambda|x_*)$ が求まったので、予測分布 $p(x_*)$ の計算に戻る。

式 (3.92) を式 (3.90) に代入して x_* について整理すると

$$
\begin{aligned}
\ln p(x_*) &= \ln p(x_*|\mu,\lambda) - p(\mu,\lambda|x_*) + \text{const.} \qquad (3.90)\\
&= \ln \mathcal{N}(x_*|\mu,\lambda^{-1}) - \ln \mathcal{N}(\mu|m(x_*),\{(1+\beta)\lambda\}^{-1}) - \ln \mathrm{Gam}\left(\lambda\middle|\frac{1}{2}+a, b(x_*)\right) + \text{const.}\\
&= -\frac{1}{2}\Big\{(x_*-\mu)^2\lambda + \ln\lambda^{-1} + \ln 2\pi\Big\}\\
&\quad + \frac{1}{2}\left[\left(\mu - \frac{x_*+\beta m}{1+\beta}\right)^2(1+\beta)\lambda + \ln\{(1+\beta)\lambda\}^{-1} + \ln 2\pi\right]\\
&\quad - \left(\frac{1}{2}+a-1\right)\ln\lambda + \frac{\beta}{2(1+\beta)}\{(x_*-m)^2+b\}\lambda\\
&\quad - \left(\frac{1}{2}+a\right)\ln\left\{\frac{\beta}{2(1+\beta)}(x_*-m)^2+b\right\} + \ln\Gamma\left(\frac{1}{2}+a\right) + \text{const.}\\
&= -\frac{1}{2}\lambda x_*^2 + \mu\lambda x_* - \frac{1}{2}\mu^2\lambda\\
&\quad + \frac{1}{2}(1+\beta)\mu^2\lambda - (x_*+\beta m)\mu\lambda + \frac{(x_*+\beta m)^2\lambda}{2(1+\beta)}\\
&\quad + \frac{\beta(x_*-m)^2\lambda}{2(1+\beta)} + \frac{b\beta\lambda}{2(1+\beta)} - \left(\frac{1}{2}+a\right)\ln\left\{\frac{\beta}{2(1+\beta)b}(x_*-m)^2+1\right\}b + \text{const.}\\
&= -\frac{(1+\beta)\lambda x_*^2}{2(1+\beta)} + \mu\lambda x_*\\
&\quad - \mu\lambda x_* - \beta m\mu\lambda + \frac{\lambda x_*^2 + 2\beta m\lambda x_* + \beta^2m^2\lambda}{2(1+\beta)}\\
&\quad + \frac{\beta\lambda x_*^2 - 2\beta m\lambda x_* + \beta m^2\lambda}{2(1+\beta)} - \left(\frac{1}{2}+a\right)\ln\left\{\frac{\beta}{2(1+\beta)b}(x_*-m)^2+1\right\} - \left(\frac{1}{2}+a\right)\ln b + \text{const.}\\
&= -\frac{\lambda x_*^2+\beta\lambda x_*^2}{2(1+\beta)} + \frac{\lambda x_*^2}{2(1+\beta)} + \frac{\beta\lambda x_*^2}{2(1+\beta)} - \left(\frac{1}{2}+a\right)\ln\left\{\frac{\beta}{2(1+\beta)b}(x_*-m)^2+1\right\} + \text{const.}\\
&= -\frac{1+2a}{2}\ln\left\{1+\frac{\beta}{2(1+\beta)b}(x_*-m)^2\right\} + \text{const.}
\end{aligned}
$$

となる。適宜 x_* と関係のない項を const. にまとめている。

この式について、$\frac{\beta}{2(1+\beta)b} = \frac{\beta a}{2a(1+\beta)b}$ として

$$
\begin{aligned}
\mu_s &= m\\
\lambda_s &= \frac{\beta a}{(1+\beta)b}\\
\nu_s &= 2a
\end{aligned}
\qquad (3.95)
$$

とおくと、予測分布 $p(x_*)$ は

$$
\ln p(x_*) = -\frac{1+\nu_s}{2}\ln\left\{1+\frac{\lambda_s}{\nu_s}(x_*-\mu_s)^2\right\} + \text{const.} = \ln \mathrm{St}(x_*|\mu_s,\lambda_s,\nu_s) \qquad (3.94)
$$

パラメータ μ_s, λ_s, ν_s を持つスチューデントの t 分布となることが分かる。

予測分布のパラメータ $\mu_s,\ \lambda_s,\ \nu_s$ を構成する事前分布のパラメータ $\hat{\beta},\ \hat{m},\ a,\ b$ について、事後分布のパラメータ (3.83)、(3.88) を用いると

$$
\begin{aligned}
\hat{\mu}_s &= \hat{m} \\
&= \frac{1}{N+\beta}\left(\sum_{n=1}^N x_n + \beta m\right)
\end{aligned}
$$

また

$$
\begin{aligned}
\hat{\lambda}_s &= \frac{\hat{\beta}\hat{a}}{(1+\hat{\beta})\hat{b}} \\
&= \frac{(N+\beta)\left(\frac{N}{2}+a\right)}{(N+1+\beta)\left\{\frac{1}{2}\left(\sum_{n=1}^N x_n^2 + \beta m^2 - \hat{\beta}\hat{m}^2\right) + b\right\}}
\end{aligned}
$$

また

$$
\begin{aligned}
\hat{\nu}_s &= 2\hat{a} \\
&= N + 2a
\end{aligned}
$$

となり、(観測データ $\boldsymbol{X}$ によって学習した) 事後分布 $p(\mu,\lambda|\boldsymbol{X})$ を用いた予測分布 $p(x_*|\boldsymbol{X})$

$$
\ln p(x_*|\boldsymbol{X}) = -\frac{1+\hat{\nu}_s}{2}\ln\left\{1+\frac{\hat{\lambda}_s}{\hat{\nu}_s}(x_*-\hat{\mu}_s)^2\right\} + \mathrm{const.} = \ln \mathrm{St}(x_*|\hat{\mu}_s,\hat{\lambda}_s,\hat{\nu}_s) \tag{3.94}
$$

が得られる。

・**R でやってみよう**

各確率分布に従いランダムに生成したデータを用いて、パラメータを推定してみましょう。

```r
# 利用パッケージ
library(tidyverse)
```

利用するパッケージを読み込みます。

・**事後分布**

```r
## パラメータの初期値を設定
# 観測モデルのパラメータ
mu_truth <- 25
lambda_truth <- 0.01

# 平均 mu の事前分布のパラメータ
m <- 20
beta <- 1

```

```r
# 精度 lambda の事前分布のパラメータ
a <- 1
b <- 1

# 試行回数
N <- 50
```

観測モデルの精度パラメータ λ を lambda_truth、平均パラメータ μ を mu とします。この値を推定するのが目的です。

続いて事前分布のパラメータ (ハイパーパラメータの初期値) を指定します。m を m、β を beta、a を a、b を b とします。

生成するデータ数 N を N とします。

λ は分散の逆数であるため、値が大きくなるほど散らばり具合が小さくなります。

```r
# ガウス分布に従うデータを生成
x_n <- rnorm(n = N, mean = mu_truth, sd = sqrt(lambda_truth^(-1)))
```

ガウス分布に従う乱数を発生させる関数 rnorm() を使って、ランダムにデータを生成します。
試行回数の引数 n には N、平均の引数 mean には mu_truth を指定します。
しかし標準偏差の引数 sd には lambda_truth のままでは指定できません。そこで標準偏差 $\sigma = \sqrt{\lambda^{-1}}$ に置き換えるため、sqrt(lambda_truth^(-1)) とします。

サンプルを確認してみましょう。

```r
# 観測データを確認
summary(x_n)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.329  17.078  22.135  24.424  30.100  52.518
```

このデータを用いてハイパーパラメータを更新します。

```r
# 事後分布のパラメータを計算
beta_hat <- N + beta
m_hat <- (sum(x_n) + beta * m) / beta_hat
a_hat <- N / 2 + a
b_hat <- (sum(x_n^2) + beta * m^2 - beta_hat * m_hat^2) / 2 + b
lambda_bar <- a_hat / b_hat
```

μ の事後分布のパラメータ (3.83) の計算を行い、$\hat{\beta}$, $\hat{m}$ をそれぞれ beta_hat, m_hat とします。
λ の事後分布のパラメータ (3.88) の計算を行い、$\hat{a}$, $\hat{b}$ をそれぞれ a_hat, b_hat とします。
また λ については分布推定しているため、ガンマ分布の期待値 $\mathbb{E}[\lambda] = \frac{\hat{a}}{\hat{b}}$ を用いることにします。計算結果は lambda_bar とします。

```r
# mu の事後分布を計算
posterior_mu_df <- tibble(
  mu = seq(
    mu_truth - 3 * sqrt((beta_hat * lambda_bar)^(-1)),
```

```r
    mu_truth + 3 * sqrt((beta_hat * lambda_bar)^(-1)),
    by = 0.01
  ),  # 作図用の値
  density = dnorm(mu, mean = m_hat, sd = sqrt((beta_hat * lambda_bar)^(-1))) # 確率密度
)
```

μ が取り得る値を seq() を使って値を用意します。

ただし mu_truth の値によって作図に必要な範囲が変わります。ここでは μ を中心に標準偏差 $\sigma = \sqrt{(\beta\lambda)^{-1}}$ の 3 倍をプロットすることにします。そこで seq() の第 1 引数 (最小値)、第 2 引数 (最大値) に、mu_truth から 3 * sqrt((beta_hat * lambda_bar)^(-1)) を引いた値と足した値をそれぞれ指定します。

mu の各値の確率密度は、ガウス分布の定義式 (2.64) で計算できます。しかし計算式が複雑なので、ここでは手を抜いてみましょう。

ガウス分布に従い確率密度を返す関数 dnorm() を使います。引数については rnorm() と同様です。ただしここでは事前分布のパラメータ m_hat と beta_hat * lambda_bar を用います。

次のガンマ分布についてもこのような関数が用意されていますが、そちらは計算式に従って計算しましょう。

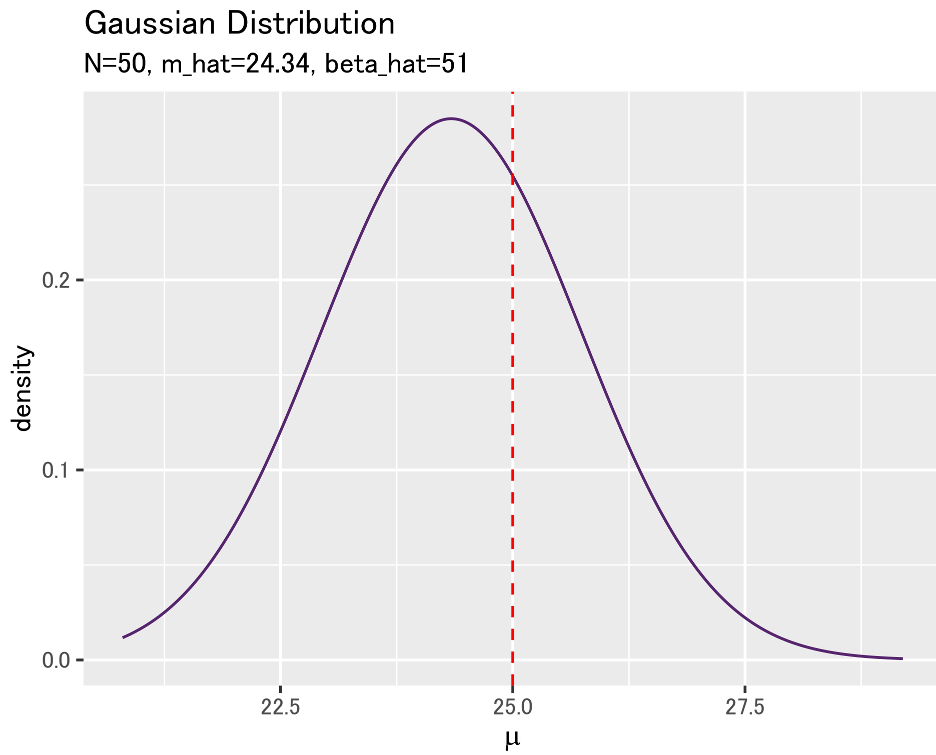
計算結果は作図用にデータフレームにまとめておきます。推定結果を確認してみましょう。

```r
head(posterior_mu_df)
```

```
## # A tibble: 6 x 2
##      mu density
##   <dbl>   <dbl>
## 1  20.8  0.0117
## 2  20.8  0.0119
## 3  20.8  0.0121
## 4  20.8  0.0124
## 5  20.8  0.0126
## 6  20.8  0.0128
```

このデータフレームを用いてグラフを描きます。

```r
# 作図
ggplot(posterior_mu_df, aes(mu, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = mu_truth),
             color = "red", linetype = "dashed") + # 垂直線
  labs(title = "Gaussian Distribution",
       subtitle = paste0("N=", N, ", m_hat=", round(m_hat, 2), ", beta_hat=", beta_hat),
       x = expression(mu)) # ラベル
```



ggplot2 パッケージを利用してグラフを作成します。

推定値は折れ線グラフ geom_line() を用いて描きます。
パラメータ μ の真の値を垂直線 geom_vline() で示します。
事前分布のパラメータ m_hat は 24.3372754 のように細かい値となるので、round() で丸めておきます。

```r
# lambda の事後分布を計算
posterior_lambda_df <- tibble(
  lambda = seq(0, 3 * lambda_truth, by = 0.00001),  # 作図用の値
  C_G = a_hat * log(b_hat) - lgamma(a_hat),  # 正規化項 (対数)
  density = exp(C_G + (a_hat - 1) * log(lambda) - b_hat * lambda) # 確率密度
)
```

λ が取り得る値を seq() を使って用意します。
ただし lambda_truth の値によって必要な範囲が変わるので、ここでは 0 から lambda_truth の 3 倍までとします。そこで seq() の第 1 引数 (最小値) に 0、第 2 引数 (最大値) に 3 * mu_truth を指定します。

lambda の各値に対してガンマ分布の定義式 (2.57) の計算を行い、確率密度を求めます。ただし値が大きくなると gamma() で計算できなくなるため、対数をとって計算することにします。なので最後に exp() で値を戻します。

こちらも同様に、計算結果は作図用にデータフレームにまとめておきます。推定結果を確認してみましょう。

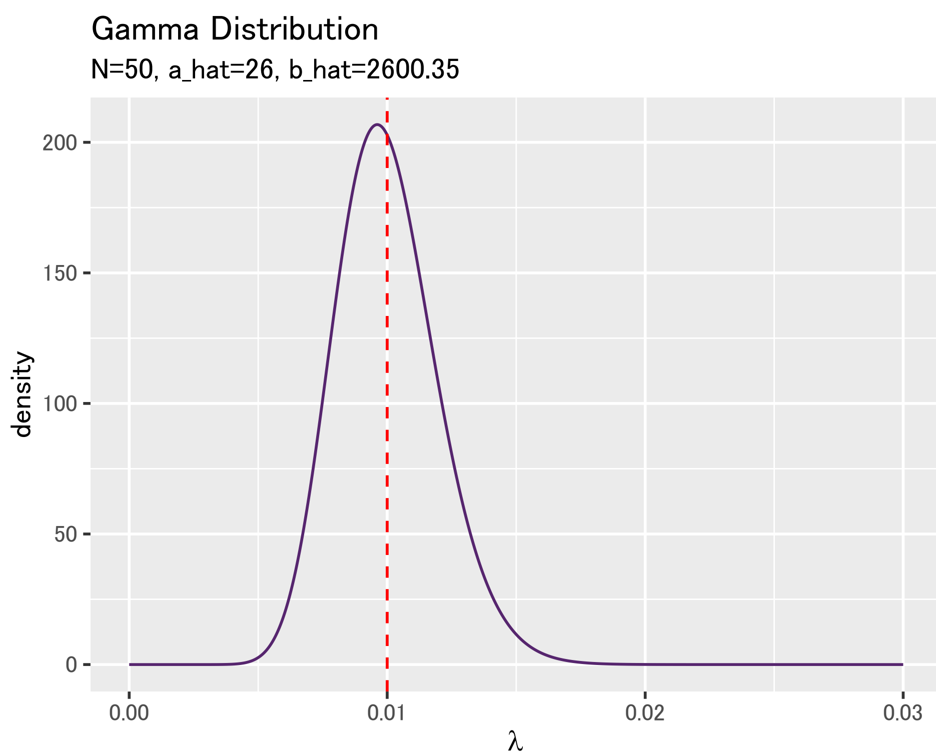
```r
head(posterior_lambda_df)
```

```
## # A tibble: 6 x 3
##    lambda   C_G  density
##     <dbl> <dbl>    <dbl>
```

```
## 1 0         146. 0.
## 2 0.00001  146. 3.88e-62
## 3 0.00002  146. 1.27e-54
## 4 0.00003  146. 3.12e-50
## 5 0.00004  146. 4.04e-47
## 6 0.00005  146. 1.04e-44
```

これも同様に ggplot2 パッケージを利用してグラフを描きます。

```r
# 作図
ggplot(posterior_lambda_df, aes(lambda, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = lambda_truth),
             color = "red", linetype = "dashed") + # 垂直線
  labs(title = "Gamma Distribution",
       subtitle = paste0("N=", N, ", a_hat=", a_hat, ", b_hat=", round(b_hat, 2)),
       x = expression(lambda)) # ラベル
```



・予測分布

```r
# 予測分布のパラメータを計算
mu_s_hat <- m_hat
lambda_s_hat <- beta_hat * a_hat / (1 + beta_hat) / b_hat
nu_s_hat <- 2 * a_hat
```

予測分布のパラメータの計算式の計算を行い、$\hat{\mu}_s$, $\hat{\lambda}_s$, $\hat{\mu}_s$ をそれぞれ mu_s_hat, lambda_s_hat, nu_s_hat とします。

```
mu_s_hat <- (sum(x_n) + beta * m) / (N + beta)
tmp_lambda_numer <- (N + beta) * (N / 2 + a)
tmp_lambda_denom <- (N + 1 + beta) * ((sum(x_n^2) + beta * m^2 - beta_hat * m_hat^2) / 2 + b)
lambda_s_hat <- tmp_lambda_numer / tmp_lambda_denom
nu_s_hat <- N + 2 * a
```

このように、観測データ `x_n` と事前分布のパラメータを使って計算することもできます (本当は `beta_hat`, `m_hat` も `beta`, `m` に置き換える必要があります。)。

```
# 予測分布を計算
predict_df <- tibble(
  x = seq(-mu_s_hat, 3 * mu_s_hat, by = 0.001),  # 作図用の値
  C_St = lgamma((nu_s_hat + 1) / 2) - lgamma(nu_s_hat / 2),  # 正規化項（対数）
  term1 = log(lambda_s_hat / pi / nu_s_hat) / 2,
  term2 = - (nu_s_hat + 1) / 2 * log(1 + lambda_s_hat / nu_s_hat * (x - mu_s_hat)^2),
  density = exp(C_St + term1 + term2) # 確率密度
)
```

x_* が取り得る値を `seq()` を使って用意します。
ここでは、$-\mu_s$ から $3\mu_s$ までとしました。必要に応じて調整してください。特にこのままですと、`mu_s` が負の値をとるとエラーとなります。

`x` の各値に対してスチューデントの t 分布の定義式 (3.76) の計算を行い、確率密度を求めます。
ちなみにこの `pi` は、前節で出てきたパラメータではなく円周率のことです。ご注意ください。

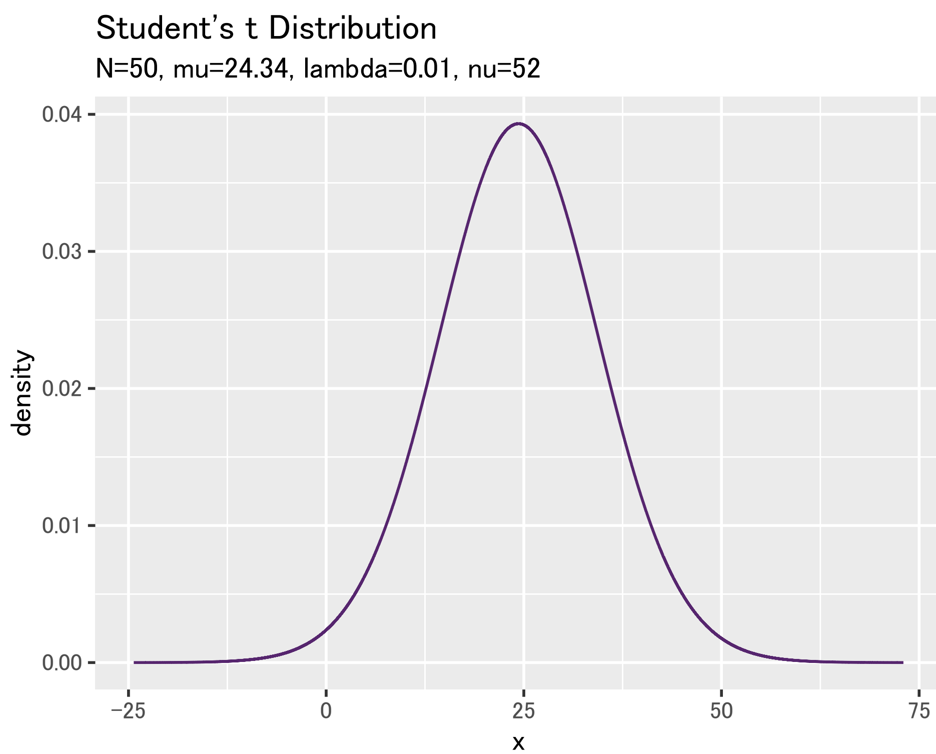
推定結果を確認してみましょう。

```
head(predict_df)
```

```
## # A tibble: 6 x 5
##       x  C_St term1 term2    density
##   <dbl> <dbl> <dbl> <dbl>      <dbl>
## 1 -24.3  1.62 -4.86 -9.79 0.00000221
## 2 -24.3  1.62 -4.86 -9.79 0.00000221
## 3 -24.3  1.62 -4.86 -9.79 0.00000221
## 4 -24.3  1.62 -4.86 -9.79 0.00000221
## 5 -24.3  1.62 -4.86 -9.79 0.00000221
## 6 -24.3  1.62 -4.86 -9.79 0.00000221
```

こちらも ggplot2 パッケージを利用して作図します。

```
# 作図
ggplot(predict_df, aes(x, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  labs(title = "Student's t Distribution",
       subtitle = paste0("N=", N, ", mu=", round(mu_s_hat, 2),
                         ", lambda=", round(lambda_s_hat, 2), ", nu=", nu_s_hat)) # ラベル
```



・おまけ

gganimate パッケージを利用して、パラメータの推定値の推移の gif 画像を作成するためのコードです。

```
# 利用パッケージ
library(tidyverse)
library(gganimate)

## パラメータの初期値を設定
# 観測モデルのパラメータ
mu_truth <- 25
lambda_truth <- 0.01

# 平均 mu の事前分布のパラメータ
m <- 20
beta <- 1

# 精度 lambda の事前分布のパラメータ
a <- 1
b <- 1
lambda_bar <- a / b

# 試行回数
N <- 50

# mu の事前分布を計算
posterior_mu_df <- tibble(
```

```r
  mu = seq(
    mu_truth - 2 * sqrt(lambda_truth^(-1)),
    mu_truth + 2 * sqrt(lambda_truth^(-1)),
    by = 0.01
  ),  # 作図用の値
  density = dnorm(mu, mean = m, sd = sqrt(beta * lambda_bar)^(-1)), # 確率密度
  N = 0 # 試行回数
)

# lambda の事前分布を計算
posterior_lambda_df <- tibble(
  lambda = seq(0, 4 * lambda_truth, by = 0.00001),  # 作図用の値
  C_G = a * log(b) - lgamma(a),  # 正規化項（対数）
  density = exp(C_G + (a - 1) * log(lambda) - b * lambda), # 確率密度
  N = 0 # 試行回数
)

# 初期値による予測分布のパラメータを計算
mu_s <- m
lambda_s <- beta * a / (1 + beta) / b
nu_s <- 2 * a

# 初期値による予測分布を計算
predict_df <- tibble(
  x = seq(-mu_truth, 3 * mu_truth, by = 0.001),  # 作図用の値
  C_St = lgamma((nu_s + 1) / 2) - lgamma(nu_s / 2),  # 正規化項（対数）
  term1 = log(lambda_s / pi / nu_s) / 2,
  term2 = - (nu_s + 1) / 2 * log(1 + lambda_s / nu_s * (x - mu_s)^2),
  density = exp(C_St + term1 + term2), # 確率密度
  N = 0 # 試行回数
)

# パラメータを推定
x_n <- rep(0, N)
for(n in 1:N){

  # ガウス分布に従うデータを生成
  x_n[n] <- rnorm(n = 1, mean = mu_truth, sd = sqrt(lambda_truth^(-1)))

  # パラメータを更新
  beta_old <- beta
  beta <- 1 + beta
  m_old <- m
  m <- (x_n[n] + beta_old * m) / beta
  a <- 1 / 2 + a
  b <- (x_n[n]^2 + beta_old * m_old^2 - beta * m^2) / 2 + b
  lambda_bar <- a / b

  # mu の事後分布を計算
  tmp_posterior_mu_df <- tibble(
    mu = seq(
      mu_truth - 2 * sqrt(lambda_truth^(-1)),
      mu_truth + 2 * sqrt(lambda_truth^(-1)),
```

```
      by = 0.01
    ),  # 作図用の mu
    density = dnorm(mu, mean = m, sd = sqrt(beta * lambda_bar)^(-1)), # 確率密度
    N = n # 試行回数
  )

  # lambda の事後分布を計算
  tmp_posterior_lambda_df <- tibble(
    lambda = seq(0, 4 * lambda_truth, by = 0.00001),  # 作図用の値
    C_G = a * log(b) - lgamma(a),  # 正規化項（対数）
    density = exp(C_G + (a - 1) * log(lambda) - b * lambda), # 確率密度
    N = n # 試行回数
  )

  # 予測分布のパラメータを更新
  mu_s <- m
  lambda_s <- beta * a / (1 + beta) / b
  nu_s <- 2 * a

  # 予測分布を計算
  tmp_predict_df <- tibble(
    x = seq(-mu_truth, 3 * mu_truth, by = 0.001),  # 作図用の値
    C_St = lgamma((nu_s + 1) / 2) - lgamma(nu_s / 2),  # 正規化項（対数）
    term1 = log(lambda_s / pi / nu_s) / 2,
    term2 = - (nu_s + 1) / 2 * log(1 + lambda_s / nu_s * (x - mu_s)^2),
    density = exp(C_St + term1 + term2), # 確率密度
    N = n # 試行回数
  )

  # 推定結果を結合
  posterior_mu_df <- rbind(posterior_mu_df, tmp_posterior_mu_df)
  posterior_lambda_df <- rbind(posterior_lambda_df, tmp_posterior_lambda_df)
  predict_df <- rbind(predict_df, tmp_predict_df)
}

# 観測データを確認
summary(x_n)

## mu の事後分布
# 作図
posterior_mu_graph <- ggplot(posterior_mu_df, aes(mu, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = mu_truth),
             color = "red", linetype = "dashed") + # 垂直線
  transition_manual(N) + # フレーム
  labs(title = "Gaussian Distribution",
       subtitle = "N= {current_frame}",
       x = expression(mu)) # ラベル

# 描画
animate(posterior_mu_graph)

## lambda の事後分布
```

```r
# 作図
posterior_lambda_graph <- ggplot(posterior_lambda_df, aes(lambda, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  geom_vline(aes(xintercept = lambda_truth),
             color = "red", linetype = "dashed") + # 垂直線
  transition_manual(N) + # フレーム
  labs(title = "Gamma Distribution",
       subtitle = "N= {current_frame}",
       x = expression(lambda)) # ラベル

# 描画
animate(posterior_lambda_graph)

## 予測分布
# 作図
predict_graph <- ggplot(predict_df, aes(x, density)) +
  geom_line(color = "#56256E") + # 折れ線グラフ
  transition_manual(N) + # フレーム
  labs(title = "Student's t Distribution",
       subtitle = "N= {current_frame}") # ラベル

# 描画
animate(predict_graph)
```

異なる点のみを簡単に解説します。

各データによってどのように学習する (推定値が変化する) のかを確認するため、こちらは for ループで 1 つずつ処理します。

よって生成するデータ数として設定した `N` がイタレーション数になります。

パラメータの推定値について、$\hat{\beta}$ や $\hat{a}$ に対応する `beta_hat` や `a_hat` などの新たなオブジェクトを作るのではなく、`beta, a` をイタレーションごとに更新 (上書き) していきます。

それに伴い、事後分布のパラメータの計算式 (3.83)、(3.88) の $\sum_{n=1}^N$ の計算は、for ループによって N 回繰り返し `x_n[n]`(や `1`) を加えることで行います。n 回目のループ処理のときには、n-1 回分の `x_n[n]`(や `1`) が既に加えられているわけです。

作図範囲を固定するため (更新されない) 観測モデルのパラメータを使います。

3.4 多次元ガウス分布の学習と予測

3.5 線形回帰の例