確率変数：Random variables

Masayuki Sakai

yyyy/mm/dd

Contents

```
require(tidyverse)
require(gridExtra)
require(ggthemes)
```

1 確率変数：Random vairbales

統計学とは，データから何らかの情報を抽出する営みである．ここでいう情報とはデータに含まれるが，本質的には知り得ない情報である．確率変数とはある特定の数学的構造を持つ変数であり，その分布によって特徴付けられる．確率変数は観測するたびに値が異なる，取りうる値は確率的に決まる．未来にとる値を正確に予測することは不可能であるが，確率を用いてどの様な値を取りうるのかについて言及することはできる．

ここでは確率変数に関する数理的な基礎と有用なポイントについて概観する．

2 累積分布関数：Cumlative distribution function

確率変数（r.v.）X の累積分布関数（c.d.f.）$F(x)$ は次の様に表される．

$$F(x) = \Pr(X \leq x)$$

累積分布関数 $F(x)$ の値は X が x 以下の値をとる確率に等しい．ここでは

$$\begin{aligned} F(-\infty) &= 0 \\ F(\infty) &= 1 \end{aligned} \tag{1.01}$$

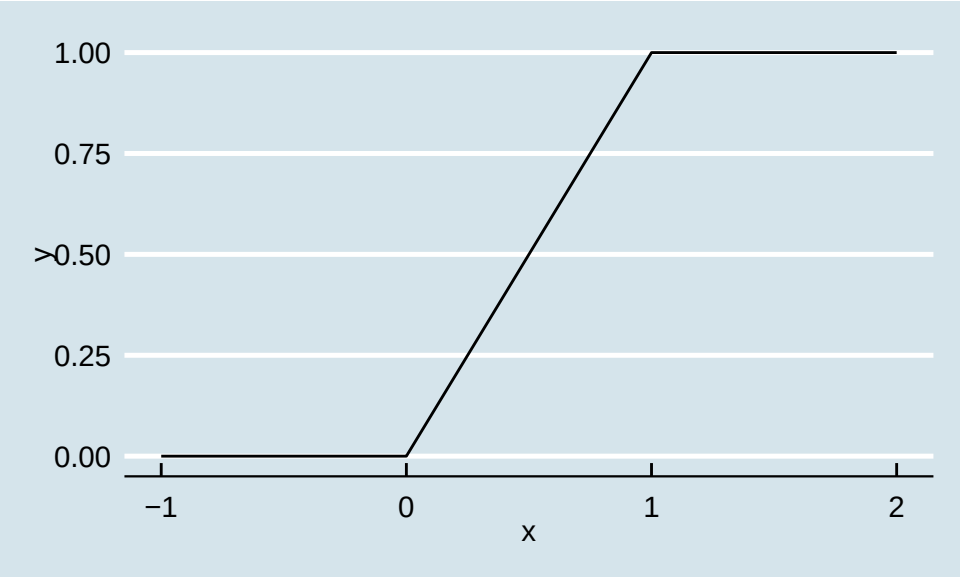
であり，$F(x)$ は単調な関数である．

ここで，$F(x)$ が連続かつ $[0,1]$ で一様分布であるとしよう．$[0,1]$ で一様な分布とは，0 から 1 の間の値全てに同じ確率が割り当てられているような分布である．この様な分布の累積分布関数は次の様に変形できる．

$$\begin{aligned} \Pr(X \leq x) = \Pr\left[F(X) \leq F(x)\right] = F(x) \\ \Rightarrow \Pr\left[F(X) \leq u\right] = u \end{aligned}$$

X よりも x の方が大きいのであれば，X 以下の値をとる確率よりも x 以下の値をとる確率の方が大きくなる．下図の通り，$[0,1]$ 上の一様分布の c.d.f の累積分布関数は $[0,1]$ で $y = x$ の直線となる．故に，$F(X)$ の値が u 以下となる確率は u と等しい．ただし $0 \leq u \leq 1$ である．

```r
data.frame(
  x = c(seq(-1,0,0.001),seq(0,1,0.001), seq(1,2,0.001)),
  y = c(rep(0,1001),qunif(p=seq(0,1,0.001), min=0, max=1), rep(1,1001))
) %>%
  ggplot(aes(x=x,y=y)) +
  geom_line() +
  ggthemes::theme_economist()
```



次に，この c.d.f. の逆関数 $F^{-}(u)$ を次の様に定義する．

$$F^{-}(u) = \min(x|F(x) \geq u)$$

この形は，F が連続である場合によく用いられる形である．F^{-} は X の分位点関数と呼ばれる．いま，U が $[0,1]$ 上の一様分布に従うとする．この時，$F^{-}(U)$ は c.d.f が F であるような X の分布と一致する．

$p \in [0,1]$ とする．このとき X の $p-$ 分位点 (p-quantile) とは，X がある値以下の値をとる確率 p を意味する．つまり $F^{-}(p)$ である．分位点は非常に有用で，あるデータ $x_1, x_2, \ldots, x_n$ が与えられた時，小さい順に並び替えて各値にたいして全体の下位何% であるかを計算することで観測データから F^{-} の形を推測することができる．これらを，理論的な分位点 $F^{-}\{(i-0.5)/n\}()$ と比較した quantile-quantile plot（QQ プロット）を見ることで理論と実測がどの程度一致しているのかを視覚的に捉えることもできる．QQ プロットでは，直線が取りうるべき値となる．

以下に例を示す．まず `x_from_unif` という変数には $[0,1]$ 上の一様分布から生成した値を，`x_from_norm` には平均 0，分散 1 の正規分布から生成した値を保存する．次に，理論的に（仮定的に）データが従う分布を平均 0，分散 1 の正規分布とする．つまり，`x_from_unif` は理論とは違う分布に従い，`x_from_norm` は理論と同じ分布に従うというケースにおける QQ プロットの違いを見る．

```r
# fix random seed
set.seed(77)

# generate random samples
x_from_unif <- runif(n=1000, min=-1, max=1)
x_from_norm <- rnorm(n=1000, mean=0, sd=1)
```

```r
# quantiles for plot
qs <- seq(0.005,0.995,0.01)
```

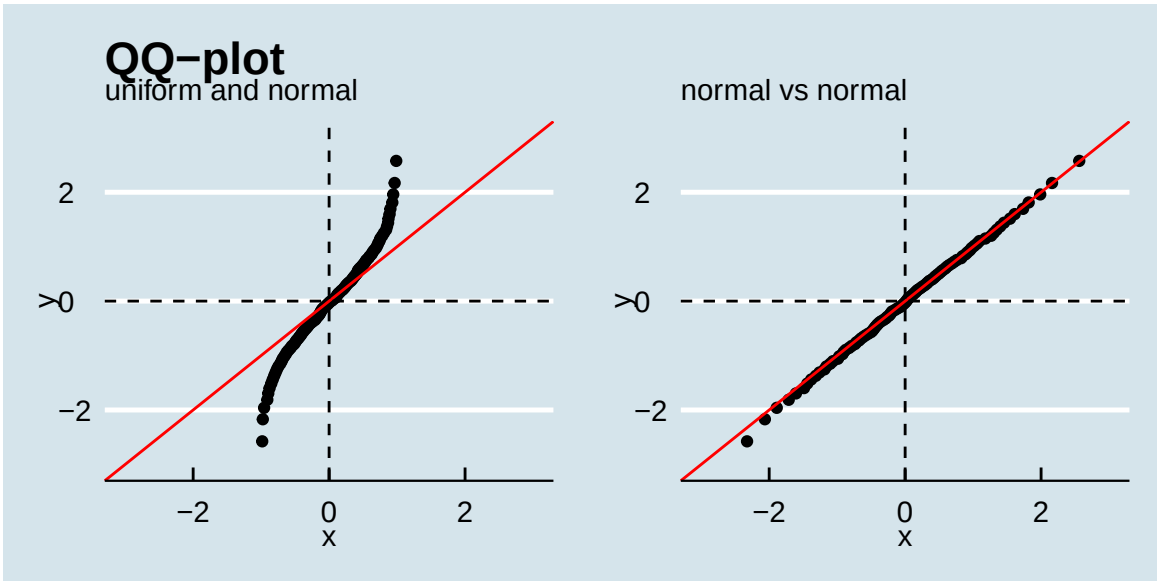
```r
# make p1 plot
# calculate observation and theoretical quantiles
x_q <- quantile(x_from_unif, probs = qs)
t_q <- qnorm(p=qs, mean = 0, sd = 1)

data.frame(x = x_q,y = t_q) %>%
  ggplot(aes(x=x, y=y)) +geom_point() +
  geom_abline(slope=1, intercept=0, color="red") +
  geom_vline(xintercept = 0, linetype=2) +
  geom_hline(yintercept = 0, linetype=2) +
  xlim(-3, 3) + ylim(-3,3) +
  ggtitle(label = "QQ-plot", subtitle = "uniform and normal") +
  ggthemes::theme_economist() -> p1

# make p2 plot
x_q <- quantile(x_from_norm, probs = qs)
t_q <- qnorm(p=qs, mean = 0, sd = 1)

df <- data.frame(x = x_q,y = t_q) %>%
  ggplot(aes(x=x, y=y)) +
  geom_point() +
  geom_abline(slope=1, intercept=0, color="red") +
  geom_vline(xintercept = 0, linetype=2) +
  geom_hline(yintercept = 0, linetype=2) +
  xlim(-3, 3) + ylim(-3,3) +
  ggtitle(label = "", subtitle = "normal vs normal") +
  ggthemes::theme_economist() -> p2

# show multiple plots
grid.arrange(p1, p2, ncol = 2)
```



確認できる様に，一様分布と正規分布との比較では直線に乗らない部分が多く，正規分布同士の比較ではほぼ直線上に乗っていることがわかる．

3 確率関数と確率密度関数：Probability (density) functions

多くの統計モデルは累積分布関数と同様に，あるいはそれ以上に「確率変数がある値をとる確率」を調べるのに役立つ．様々な統計モデルを区別するためにまず用いることができるのは確率変数が取りうる値が離散（整数など）なのか連続なのか（実数直線上の連続した区間）である．

離散型確率変数の時，X に対して確率関数 (probability function)$f(x)$ は次の様に表される．

$$f(x) = \Pr(X = x)$$

ここで，$0 \leq f(x) \leq 1$ であり，$\sum_i f(x_i) = 1$ である．$\sum_i$ は'summation' の意味であり添字に該当する変数を全て足しあげる操作を意味する．

連続型確率変数の場合，X がある特定の値をとるという考え方をすると，その確率は限りなく 0 に近く．たとえ区間が $[0, 1]$ であってもその間には無限個の実数が存在するからである．そこで，確率密度関数 (probability density function) は確率を考える対象に幅を持たせる．シンプルに考えれば連続型確率変数 X が $[0, 1]$ 区間に含まれる確率は 1（必ずそうなる）であるし，全ての値の確率が等しいのであれば $(0.5, 1]$ に含まれる値をとる確率は 1/2 程度であるべきである．すなわち $\Pr(x - \Delta/2 < X < x + \Delta/2) \simeq f(x)$ となる．ここである定数 $a \leq b$ を考えて

$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

とも表される．すなわち，確率密度関数の面積と確率を対応させた問題に置き換えるのである．確率として扱うために $f(x)$ については $f(x) \geq 0$ かつ $\int_{-\infty}^{\infty} f(x)dx = 1$ を満たす必要がある．

4 確率変数：Random vectors

通常の統計解析で扱う変数が 1 つということはほとんどなく，その場合も確率 (密度) 関数も多変数関数となる．可視化のしやすさも踏まえて 2 変数 X, Y の場合を考える．

X, Y の同時確率密度関数 (joint probability density function) を次の様に考える．Ω をある $x - y$ 平面上の空間とする．

$$\Pr\left[(X, Y) \in \Omega\right] = \int\int_{\Omega} f(x, y)dxdy$$

つまり，$f(x, y)$ は $x - y$ 平面上の空間に確率を対応させる．

次のような関数を例にして見てみよう．

$$f(x, y) = \begin{cases} x + \frac{3}{2}y^2, & 0 < x < 1 \& 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

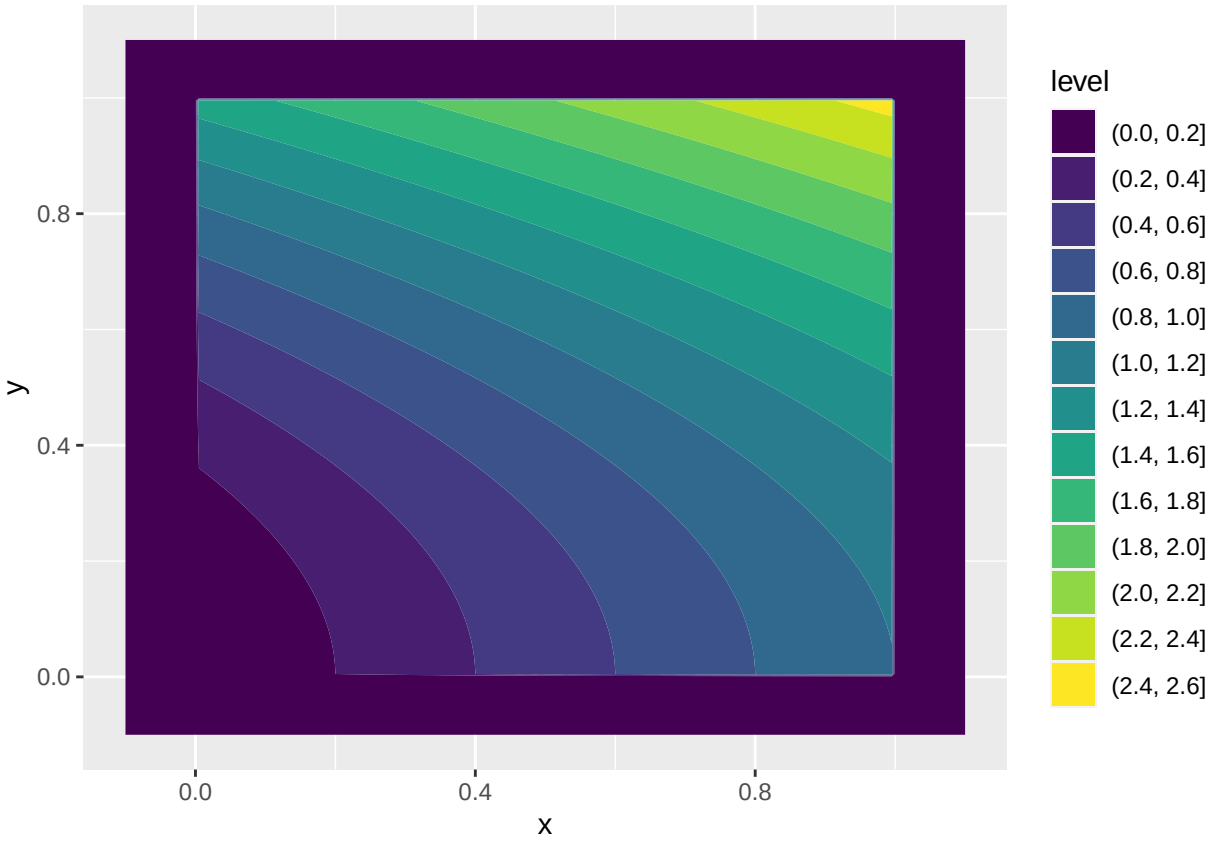
この関数を R で実装したのが下記のコードである．`if(...)` で $0 < x, y < 1$ という条件を判定している．

```
f <- function(x, y){
  if({0<x&x<1}&{0<y&y<1}){
    return(x + (3/2)*y^2)
  }else{
    return(0)
  }
}
```

x, y をそれぞれ横軸にとり，$f(x, y)$ の値で等高線を引いたのが次の図である．

```
df <- expand_grid(x=seq(-0.1, 1.1, 0.005), y=seq(-0.1, 1.1, 0.005)) %>%
  mutate(z = map2_dbl(.x=x, .y=y, f)) %>%
  ggplot(aes(x, y, z=z)) +
  geom_contour_filled() -> g

plot(g)
```



濃い紫色は 0，黄色になるにつれ $f(x, y)$ の値が大きくなっていることを示している．つまり $f(x, y)$ とは x, y ともに $0 < x, y < 1$ の範囲では入力が大きくになるにつれ出力も大きくなる様な関数であることがわかる．

5 周辺分布：Marginal distribution

前節につづき X, Y のに 2 変数のケースを考えていく．X または Y の確率密度関数を考えたい場合，$f(x, y)$ から各変数に着目した確率密度関数を得ることができる．これを X または Y の周辺分布 (Marginal

distribution) と呼ぶ．考え方としては，$-\infty < Y < \infty$ が与えられたもとでの X の確率密度関数の形を考えれば良い．すなわち，

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

となる．

6 条件付き分布：Conditional distribution

例えば，$Y = y_0$ という状況を考えよう．この時 X の分布はどの様になるだろうか．この状況を $f(x|Y = y_0)$ と表す．この時，直感的には下記の様に元の $f(x, y = y_0)$ に k 倍されたような値になると推測される．

$$f(x|Y = y_0) = kf(x, y_0)$$

$f(x|y)$ もまた確率密度関数であることから，全区間での積分は 1 にならなければいけない．すなわち，

$$k\int_{-\infty}^{\infty} f(x, y_0)dx = 1 \Rightarrow kf(y_0) = 1 \Rightarrow k = \frac{1}{f(y_0)}$$

となる．$f(y_0)$ は y の周辺分布の確率密度関数である．これより，次のことがわかる．

Definition 1. X, Y をそれぞれ連続型確率変数とし，$f(x, y)$ を X, Y の同時分布の密度関数とする．この時 $Y = y_0$ が与えられたもとでの X の条件付き分布は次の様に定義される．

$$f(x|Y = y_0) = \frac{f(x, y_0)}{y_0}$$

ここで，$f(y_0) > 0$ とする．

条件付き分布のこれまでの議論は X と Y を入れ替えても同様であることに注意されたい．数式変形をすれば，同時分布と条件付き分布の関係は $f(x, y) = f(x|y)f(y)$ と表せる．さらに，3 変数の場合，次の様に関係を表すことができる．

1. $f(x, z|y) = f(x|z, y)f(z|y)$
2. $f(x, z, y) = f(x|z, y)f(z|y)f(y)$
3. $f(x, z, y) = f(x|z, y)f(z, y)$

一般には変数の数が増えた場合には 2 変数の場合の性質が全て拡張できるわけではないことに注意されたい．

7 ベイズの定理：Bayes theorem

先に出てきた次の式

$$f(x, y) = f(x|y)f(y) = f(y|x)f(x)$$

より以下の式が成り立つ．

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}$$

これをベイズの定理と呼ぶ．ここから導出される統計モデルについては二章と六章で触れる．

8 独立と条件付き独立：Independence and conditional independence

確率変数 X, Y に対して $f(x|y)$ が与えられている時，もし $f(x|y)$ の値が y の値によらずに決定する時，x は y とは独立であるという．すなわち

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty} f(x,y)dy \\ &= \int_{-\infty}^{\infty} f(x|y)f(y)dy \\ &= f(x|y)\int_{-\infty}^{\infty} f(y)dy = f(x|y) \end{aligned}$$

を意味する．ここから，

$$f(x,y) = f(x|y)f(y) = f(x)f(y)$$

が導かれる.つまり,X, Y が互いに独立ならば $f(x,y) = f(x)f(y)$ が成り立つ.明らかに,$f(x,y) = f(x)f(y)$ が成り立てば，$f(x|y) = f(x)$ となる．確率変数の実現値として得られた各値が独立であるような場合を i.i.d と表す．これは independent identically distributed の略称である．

多くの応用においては，観測値が i.i.d ではないことも多い．しかし，ある条件のもとでの独立を考えることで問題に取り組みやすくすることができる．それが，条件付き独立という概念である．

いま，確率変数の列 $X_1, X_2, \ldots, X_n$ と $\mathbf{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{x+1}, \ldots, X_n)^{\mathrm{T}}$ を考える．$\mathbf{X}_{-i}$ が定義の通り列 $X_1, X_2, \ldots, X_n$ から i 番目の要素のみ除いた様なベクトルである．マルコフ連鎖と呼ばれる性質はこの表現を用いて次の様に表せる．

$$f(x_i|\mathbf{x}_{-i}) = f(x_i|x_{i-1})$$

すなわち，i 番目の値が $i-1$ 番目の値によって特徴付けられるような状況を表している．また，i 番目の値に影響を与えるのはあくまでも $i-1$ 番目の値のみであり，他とは無関係という仮定をおけば，

$$\begin{aligned} f(\mathbf{x}) &= f(x_n|\mathbf{x_{-n}})f(\mathbf{x}_n) \\ &= f(x_n|x_{n-1})f(\mathbf{x_{-n}}) \\ &= f(x_n|x_{n-1})f(x_{n-1}|\mathbf{x_{-(n-1)}})f(\mathbf{x_{-(n-1)}}) \\ &= f(x_n|x_{n-1})f(x_{n-1}|x_{n-2})f(\mathbf{x_{-(n-1)}}) \\ &= \ldots \\ &= \prod_{i=2}^{n} f(x_i|x_{i-1})f(x_1) \end{aligned}$$

が得られる．こちらの式の方が計算資源が少なくて済む．

9 期待値と分散：Mean and variance

確率変数の分布について完璧に把握できるに越したことはないが，多くの場合期待値と分散を調べることで分布の概要をつかむことができる．いま X を確率変数としてその確率密度関数を $f(x)$ とする．この時，期待値 $E(X)$ は次の様に定義される．

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

積分をする際にそれぞれ x を重みとしてかけている．すなわち，実現値に対してその確率で重みづけした平均である．

ある確率変数の関数 g についての期待値は次の様になる．

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

いま，$\mu = E(X)$ として，$g = (X-\mu)^2$ とする．g は平均と確率変数 X との差の二乗である．分散はまさにこの様に定義されるもので，

$$\mathrm{var}(X) = E[(X-\mu)^2]$$

となる．分散とは分布がどのように広がっているかを示す指標の 1 つである．計算上はシンプルであるがその意味の解釈性は高くない．もとの値とのスケールを合わせるよう正の平方根をとった標準偏差が用いられることも多い．

$$\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)}$$

10 平均と分散の線形変換：Mean and variance of liniear transformations

$$\begin{aligned}
\mathrm{var}(a+bX) &= E[((a+bX) - E(a+bX))^2] \\
&= E[(a+bX-a-b\mu)^2] \\
&= E[(b^2(X-\mu)^2)] \\
&= b^2E[(X-\mu)^2] \\
&= b^2\mathrm{var}(X)
\end{aligned}$$

X, Y を確率変数として，$E(X+Y) = E(X) + E(Y)$ が成り立つ．これを確かめるために X, Y の同時密度 $f(x,y)$ を考えると

$$\begin{aligned}
E(X+Y) &= \int (x+y)f(x,y)dxdy \\
&= \int xf(x,y)dxdy + \int yf(x,y)dxdy \\
&= E(X) + E(Y)
\end{aligned}$$

となる．

ここでポイントなのは，X, Y どちらにも特定の分布を仮定していないことである．つまり，この性質は期待値と分散が存在すれば，確率分布の特徴によらず成り立つ．いま，X, Y が i.i.d であるとすると，$E(XY) = E(X)E(Y)$ が成り立つ．

$$
\begin{aligned}
E(XY) &= \int xyf(x,y)dxdy \\
&= \int xf(x)yf(y)dxdy \\
&= \int xf(x)dx \int yf(y)dy \\
&= E(X)E(Y)
\end{aligned}
$$

この逆は，X, Y が共に正規分布に従う場合に成り立つ．

分散については，独立な場合のみ期待値と同様にそれぞれの確率変数の分散の和と一致する．一般の形について議論する前にまず 2 つの確率変数の共分散について紹介する．

$$
\begin{aligned}
\mathrm{cov}(X,Y) &= E[(X-\mu_x)(Y-\mu_y)] \\
&= E(XY) - E(X)E(Y)
\end{aligned}
$$

ここで，$\mathrm{cov}(X,Y)$ は X, Y の共分散を表す．また $\mu_x = E(X), \mu_y = E(Y)$ である．$X = Y$ とおくと明らかに $\mathrm{cov}(X,X) = \mathrm{var}(X)$ であるので，分散は共分散の特別な場合と考えることもできる．先ほどの期待値の議論で X, Y が独立な場合は $E(XY) = E(X)E(Y)$ であることがわかっているので，X, Y が独立な場合は $\mathrm{cov}(X,Y) = E(XY) - E(X)E(Y) = 0$ となる．

いま，$\mathbf{A}, \mathbf{b}$ をそれぞれ行列とベクトルとする．$\mathbf{b}$ は有限な定数ベクトルで，$\mathbf{A}$ と行数と同じ長さとする．次に，$\mathbf{X}$ を確率変数ベクトルとする．このとき，

$$
E(\mathbf{X}) = \boldsymbol{\mu}_x = [E(X_1), E(X_2), \ldots, E(X_n)]^T
$$

として，$E(\mathbf{AX} + \mathbf{b}) = \mathbf{A}E(\mathbf{X} + \mathbf{b})$ が成り立つ．

確率変数ベクトルに対する分散共分散行列は次の様に定義される．

$$
\begin{aligned}
\boldsymbol{\Sigma} &= E[(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^T] \\
&\text{where } \Sigma_{ij} = \mathrm{cov}(X_i, X_j)
\end{aligned}
$$

共分散は順序によらないので $\boldsymbol{\Sigma}$ は対称行列である．また，次が成り立つ．

$$
\boldsymbol{\Sigma}_{AX+b} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T
$$

この性質については次のように示される．

$$
\begin{aligned}
\boldsymbol{\Sigma}_{AX+b} &= E[(\mathbf{AX} + \mathbf{b} - \mathbf{A}\boldsymbol{\mu}_x - \mathbf{b})(\mathbf{AX} + \mathbf{b} - \mathbf{A}\boldsymbol{\mu}_x - \mathbf{b})^T] \\
&= E[(\mathbf{AX} - \mathbf{A}\boldsymbol{\mu})(\mathbf{AX} - \mathbf{A}\boldsymbol{\mu}_x)^T] \\
&= \mathbf{A}E[(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^T]\mathbf{A}^T \\
&= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T
\end{aligned}
$$

また，$\mathbf{a}$ を実数の定数ベクトルとすると，$\mathrm{var}(\mathbf{a}^T\mathbf{X}) \geq 0$ が成り立つ．

11 多次元正規分布：The multivariate normal distribution

これは正規分布の多次元拡張版であり，統計において非常に重要な分布である．

Theorem 1. n 個の i.i.d かつそれぞれ平均 *0,* 分散 *1* の正規分布に従う確率変数の集合を考える．すなわち各要素は $Z_i \sim N(0,1)$ であり，そのベクトルを $\mathbf{Z}$ と置く．この時，$\mathbf{Z}$ の分散共分散行列をは $\mathbf{I}_n$ であり，Z の期待値は $E(\mathbf{Z}) = \mathbf{0}$ である．いま，$\mathbf{B}$ を $m \times n$ の実数で有限な値を持つ行列として，$\boldsymbol{\mu}$ を長さ m の有限な実ベクトルとする．この時，ベクトル $\mathbf{X} = \mathbf{BZ} + \boldsymbol{\mu}$ は多次元正規分布に従うという．$E(\mathbf{X}) = \boldsymbol{\mu}$, $\mathrm{var}(\mathbf{X}) = \boldsymbol{\Sigma} = \mathbf{BB}^T$ となる．これを次のように表す．

$$\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

後に紹介する定理のいくつかはこの分布を仮定している。この分布の p.d.f は次の定義される．

$$f_{\mathrm{x}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}|}} e^{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \text{ for } \mathbf{x} \in \mathbb{R}^m$$

ここで $\boldsymbol{\Sigma}$ はサイズ m のフルランクの行列である．また，非負とする．

12 多次元 t 分布：A multivariate t distribution

前節で導入した多次元正規分布では n 個のそれぞれ $N(0,1)$ に従う独立な確率変数を考えたが，ここでは従う分布を t_k とおく．すなわち，$T_i \sim t_k$ とする．このとき，これらの n 個の確率変数を要素にもつ確率変数ベクトルが従う多次元の t 分布，$t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ を考えることができる．正規分布よりも裾の重い分布を扱いたい場合，多次元 t 分布はシミュレーションなどで有用である．多次元正規分布との違いの 1 つに，一般には周辺分布が t 分布にはならないことがあげられる．

13 正規分布と線形変換：Linear transformations of normal random vectors

n 次元の多変量正規分布に従う確率変数 $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ と有限な定数を要素にもつ行列 $\mathbf{A}$ を考える．このとき，多変量正規分布の定義より

$$\mathbf{AX} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T).$$

となることがわかる．$\mathbf{X} = \mathbf{BZ} + \boldsymbol{\mu}$ と表せることから，$\mathbf{AX} = \mathbf{ABZ} + \mathbf{A}\boldsymbol{\mu}$ とできることがわかる．また先の議論と同様に $\boldsymbol{\Sigma} = \mathbf{BB}^T$ とおけば，$\mathrm{var}(\mathbf{AX}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ となる．

特殊なケースとして行列ではなくベクトル $\mathbf{a}$ として，

$$\mathbf{a}^T\mathbf{X} \sim N(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}).$$

となる.これより,$\mathbf{a}$ を j 番目の要素だけ 1 で他を 0 であるようなベクトルとすると,結果は $X_j \sim N(\mu_j, \Sigma_{jj})$ と一致する．ここで $\sigma_j^2 = \Sigma_{jj}$ である．j 番目については特に制約は置いていないため，$\mathbf{X}$ が多変量正規分布に従う時，任意の X_j での周辺分布は単変量の正規分布に従うことがわかる．より一般的には，$\mathbf{X}$ の部分ベクトル $\mathbf{X}'$ も多変量正規分布に従う．

ただし，この逆は一般には成り立たない．しかし，$\mathbf{a}^T\mathbf{X}$ が多変量正規分布に従うならば，$\mathbf{X}$ は多変量正規分布に従うことは示される．

14 多変量正規分布の条件付き分布：Multivariate normal conditional distributions

まず，$\mathbf{Z}, \mathbf{X}$ をそれぞれ確率変数ベクトルとし，多変量正規分布に従うとする．この時これらの同時分布を考える．この同時分布の分散共分散行列は次のように分割して考えることができる．

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_z & \mathbf{\Sigma}_{zx} \\ \mathbf{\Sigma}_{xz} & \mathbf{\Sigma}_x \end{bmatrix},$$

そして，

$$\mathbf{X}|\mathbf{z} \sim N(\boldsymbol{\mu}_x + \mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z), \mathbf{\Sigma}_x - \mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1}\mathbf{\Sigma}_{zx}).$$

が成り立つ．この証明では，対象な分割行列の逆行列の性質を利用する．

$$\begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}\mathbf{D}^{-1}\mathbf{C}^T\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{C}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}^T\mathbf{A}^{-1} & \mathbf{D}^{-1} \end{bmatrix}$$

ここで $\mathbf{D} = \mathbf{B} - \mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}$ とする．

ここから，$\mathbf{Z}$ が与えられたもとでの $\mathbf{X}$ の条件付き分布を考える．まず $\mathbf{Q} = \mathbf{\Sigma}_x - \mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1}\mathbf{\Sigma}_{zx}, \tilde{\mathbf{z}} = \mathbf{z} - \boldsymbol{\mu}, \tilde{\mathbf{x}} = \mathbf{x} - \boldsymbol{\mu}_x$ とする．また，定数として与えられている z に関する項を z terms としてまとめると．

$$\begin{aligned}
f(\mathbf{x}|\mathbf{z}) &= f(\mathbf{x}, \mathbf{z})/f(\mathbf{z}) \\
&\propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} \tilde{\mathbf{z}} \\ \tilde{\mathbf{x}} \end{bmatrix}^T \begin{bmatrix} \mathbf{\Sigma}_z^{-1} + \mathbf{\Sigma}_z^{-1}\mathbf{\Sigma}_{zx}\mathbf{Q}^{-1}\mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1} & -\mathbf{\Sigma}_z^{-1}\mathbf{\Sigma}_{zx}\mathbf{Q}^{-1} \\ -\mathbf{Q}^{-1}\mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1} & \mathbf{Q}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{z}} \\ \tilde{\mathbf{x}} \end{bmatrix} \right\} \\
&\propto \exp\left\{ -\frac{\tilde{\mathbf{x}}\mathbf{Q}^{-1}\tilde{\mathbf{x}}}{2} + \tilde{\mathbf{x}}^T\mathbf{Q}^{-1}\mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1}\tilde{\mathbf{z}} + z \text{ term} \right\} \\
&\propto \exp\left\{ -\frac{(\tilde{\mathbf{x}} - \mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1}\tilde{\mathbf{z}})\mathbf{Q}^{-1}(\tilde{\mathbf{x}} - \mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1}\tilde{\mathbf{z}})}{2} + z \text{ terms} \right\} \\
&\propto \exp\left\{ -\frac{1}{2}\left[\mathbf{x} - \boldsymbol{\mu}_x - \mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z)\right]\mathbf{Q}^{-1}\left[\mathbf{x} - \boldsymbol{\mu}_x - \mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z)\right] + z \text{ terms} \right\}
\end{aligned}$$

よって，これは平均 $\boldsymbol{\mu}_x + \mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z)$ で分散共分散行列が $\mathbf{Q} = \mathbf{\Sigma}_z - \mathbf{\Sigma}_{xz}\mathbf{\Sigma}_z^{-1}\mathbf{\Sigma}_{zx}$ である多変量正規分布とみなすことができる．

15 確率変数の変数変換：Transformation of random variables

確率密度関数として f_z を持つような連続型の確率変数 Z を考える．また，$X = g(Z)$ であるような確率変数 X を作る．ここで g は逆関数を持つと仮定する．このとき，X の累積分布関数は Z の情報から簡単に求めることができる．

$$\begin{aligned}
F_x(x) &= \Pr(X \leq x) \\
&= \begin{cases} \Pr\left[g^{-1}(X) \leq g^{-1}(x)\right] = \Pr\left[Z \leq g^{-1}(x)\right], & g \text{ increasing} \\ \Pr\left[g^{-1}(X) > g^{-1}(x)\right] = \Pr\left[Z > g^{-1}(x)\right], & g \text{ decreasing} \end{cases} \\
&= \begin{cases} F_z[g^{-1}(x)], & g \text{ increasing} \\ 1 - F_z[g^{-1}(x)], & g \text{ decreasing} \end{cases}
\end{aligned}$$

f_x を求める際には g が単調増加か単調減少かで少し異なるので注意．これは以下のように求めることができる．

$$\begin{aligned} f_x(x) &= F'_x(x) \\ &= F'_z[g^{-1}(X)] \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| \\ &= f_z[g^{-1}(x)] \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|. \end{aligned}$$

もし，g がベクトル値関数の時，$\mathbf{Z}, \mathbf{X}$ が同じ長さのベクトルとすると，上記の式変換の結果は次のようになる．

$$f_x(\mathbf{x}) = f_z(g^{-1}(\mathbf{x}))|\mathbf{J}|$$

ここで $\mathbf{J}$ は $J_{ij} = \partial z_i / \partial x_j$ であるような行列である（ヤコビアン）．もし f_x, f_z が離散型確率変数であるとするならば，$|\mathbf{J}|$ の項は $\mathbf{1}$ となり，無視して良い．

15.1 例

まず，$Z \sim N(\mathbf{0}, \mathbf{I})$，$\mathbf{B}$ を逆行列を持つ $n \times n$ の行列とする．この時，$\mathbf{X} = \mathbf{BZ} + \boldsymbol{\mu}$ の確率密度関数を多変量正規分布の定義を元に導出してみよう．

$\mathbf{X}$ の分散共分散行列 $\boldsymbol{\Sigma}$ は $\mathbf{BB}^T$ である．また，$\mathbf{Z} = \mathbf{B}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ である．つまりこのとき，ヤコビアン $|\mathbf{J}|$ は $|\mathbf{B}^{-1}|$ と一致する．$\mathbf{Z}$ の各要素 Z_i 同士は独立であるので，$\mathbf{Z}$ の密度関数はそれぞれの要素の密度関数の積となり，

$$f(\mathbf{z}) = \frac{1}{\sqrt{2\pi}^n} \exp\left\{-\frac{\mathbf{z}^T\mathbf{z}}{2}\right\}$$

である．ここで先ほど紹介した変数変換の定理を用いれば，$\mathbf{X}$ の密度関数は

$$\begin{aligned} f(\mathbf{x}) &= \frac{|\mathbf{B}|^{-1}}{\sqrt{2\pi}^n} \exp\left\{-\frac{-(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{B}^T\mathbf{B}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right\} \\ &= \frac{1}{\sqrt{2\pi}^n|\boldsymbol{\Sigma}|} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu}^T)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right\} \end{aligned}$$

となる．ここで逆行列と行列式の関係 $|A^{-1}| = \frac{1}{|A|}$ を用いた．

16 積率母関数：Monent generating functions

p.d.f や c.d.f. は確率変数が従う分布を特徴付けるものであったが，ここでもう 1 つの分布を特徴付ける概念を紹介する．それが積率母関数 (moment generating function：m.g.f.) である. これは

$$M_X(s) = E(e^{sX})$$

と表わされる．ここで s は実数である．m.g.f. の k 次の導関数の $s = 0$ での値を $k - th$ モーメントと呼ぶ. また，モーメントは次の関係を持ち

$$\left.\frac{\mathrm{d}^k M_X}{\mathrm{d}s^k}\right|_{s=0} = E(X^k).$$

$M_X(0)=1, M_X'(0)=E(X), M_X''(0)=E(X^2)$ が成り立つ．以下の性質は有用である．

1. もし 2 つの確率変数 X,Y にたいしてその積率母関数 M_X, M_Y が一致する時，すなわち $M_X(s)=M_Y(s)$ であれば，X,Y の分布は一致する．
2. 2 つの確率変数 X,Y が互いに独立ならば，

$$\begin{aligned} M_{X+Y}(s) &= E\left[\exp(s(X+Y))\right] \\ &= E[e^{sX}e^{sY}] \\ &= E(e^{sX})E(e^{sY}) \\ &= M_X(s)M_Y(s) \end{aligned}$$

   が成り立つ．
3. $M_{a+bX}(s)=E(e^{as+bXs})=e^{as}M_X(bs)$.

17 中心極限定理：The central limit theorem

$X_i \overset{\text{i.i.d}}{\sim} N(\mu,\sigma^2),\ i=1,2,\ldots,n$ を考える．また $\bar{X}_n=\sum_{i=1}^n X_i/n$ とする．中心極限定理とは，この式において $\lim n\to\infty$ とした時の $\bar{X}_n$ の振る舞いについて述べたものである．すなわち

$$\bar{X}_n \sim N(\mu,\sigma^2/n)$$

が成り立つという定理である．直感的には，X_n の p.d.f.，f を未知として，この f の対数をとった $l(\bar{x}_n)=\log f(\bar{x}_n)$ についてのテイラー展開を考えたものである．$\hat{x}_n$ 周りでのテイラー展開

$$f(\bar{x}) \simeq \exp\left\{l(\hat{x}_n)+l''(\bar{x}_n-\hat{x}_n)^2/2+l'''(\bar{x}_n-\hat{x}_n)^3/6+\cdots\right\}$$

について $n\to\infty$ のとき，$\bar{x}_n-\hat{x}_n\to 0$ となる．

より厳密には，積率母関数の性質を利用する．まず以下を定義する．

$$\begin{aligned} Y_i &= \frac{X_i=\mu}{\sigma}, \\ Z_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i = \frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}. \end{aligned}$$

Z_n の積率母関数を Y_i の積率母関数を用いて

$$\begin{aligned} M_{Z_n}(s) &= \left\{M_Y(s/\sqrt{n})\right\}^n \\ &= \left\{M_Y(0)+M_Y'(0)\frac{s}{\sqrt{n}}+M_Y''(0)\frac{s^2}{2n}+o(n^{-1})\right\}^n \\ &= \left\{1+\frac{s^2}{2n}+o(n^{-1})\right\}^n \\ &= \exp\left[n\log\left\{1+\frac{s^2}{2n}+o(n^{-1})\right\}\right] \\ &\to \exp\left(\frac{s^2}{2}\right) \ \text{ as } \ n\to\infty. \end{aligned}$$

この最後の形が $N(0,1)$ の積率母関数の形と一致する．

18 大数の法則：Chebyshev, Jensen and the law of large numbers

18.1 チェビシェフの不等式：Chebyshev's inequality

確率変数 X と 2 次のモーメント $E(X^2)$ が有限 $(< \infty)$ であるとする．このとき

$$\Pr(|X| \geq a) \leq \frac{E(X^2)}{a^2}.$$

(証明)　　期待値の定義より

$$E(X^2) = E(X^2|a \leq |X|)\Pr(a \leq |X|) + E(X^2|a > |X|)\Pr(a > |X|)$$

とできる．右辺の 2 つの項は非負であるので，$E(X^2) \geq E(X^2|a \leq |X|)\Pr(a \leq |X|)$ が成り立つ．しかし，もし $a \leq |X|$ であれば，明らかに $a^2 \leq E(X^2|a \leq |X|)$ であり，$E(X^2) \geq a^2\Pr(|X| \geq a)$ となる．これより上式は示された

18.2 大数の法則：The law of large numbers

平均が μ で，$E(|X_i|) < \infty$ であるような分布に従う互いに独立な n 個のの確率変数 $X_1, X_2, \ldots, X_n$ を考える．$\bar{X}_n = \sum_{i=1}^n X_i/n$ としたとき．任意の ϵ に対して，

$$\Pr\left(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon\right) = 1$$

が成り立つ．このとき，$\bar{X}_n$ は μ に概収束するという．これを大数の強法則という．
また，仮定として $\mathrm{var}(X_i) = \sigma^2 < \infty$ を加えた時，次が成り立つ．

$$\lim_{n\to\infty} \Pr\left(|\bar{X}_n - \mu| \geq \epsilon\right) = 0,$$

これを大数の弱法則と呼ぶ．この時，X_n は μ に確率的に収束する．この証明は以下のような形で示される．

$$\begin{aligned}\Pr(|\bar{X}_n - \mu| \geq \epsilon) &\leq \frac{E(\bar{X}_n - \mu)^2}{\epsilon^2} \\ &= \frac{\mathrm{var}(\bar{X}_n)}{\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2}\end{aligned}$$

最後の式を見ると，これは $n \to \infty$ で 0 に収束する．この不等式をチェビシェフの不等式という．ちなみに互いに独立という条件は $\mathrm{var}(\bar{X}_n) = \sigma^2/n$ という等式が成り立つことを保証している．

18.3 イェンセンの不等式：Jensen's inequality

イェンセンの不等式とは，任意の確率変数 X と凸な関数 c について，

$$c\{E(X)\} \geq E\{c(X)\}$$

が成り立つという定理である．

この証明をするために，まずは確率変数を離散型と仮定する．ある下に凸な関数 c があって，

$$c(w_1x_1 + w_2x_2) \geq w_1c(x_1) + w_2c(x_2) \tag{18.1}$$

を満たすような任意の非負の数 w_1, w_2 を考える．ただし，$w_1 + w_2 = 1$ を満たす．この時，

$$c\left(\sum_{i=1}^{n-1} w_i'x_i\right) \geq \sum_{i=1}^{n-1} w_i'c(x_i) \tag{18.2}$$

ここで非負の定数 w_0' について $\sum_{i=1}^{n-1} w_i' = 1$ として，上式が成り立つとする．仮定を満たすような w_i' の任意の組を考えると，

$$\begin{aligned} c\left(\sum_{i=1}^{n} w_ix_i\right) &= c\left((1-w_n)\sum_{i=1}^{n-1}\frac{w_ix_i}{1-w_n} + w_nx_n\right) \\ &\geq (1-w_n)c\left(\sum_{i=1}^{n-1}\frac{w_ix_i}{1-w_n} + w_nc(x_n)\right) \end{aligned} \tag{18.3}$$

が成り立つ．ここで最後の不等式は式 (18.1) を用いた．また $\sum_{i=1}^{n} w_i = 1$ より，$\sum_{i=1}^{n-1} w_i/(1-w_n) = 1$ が成り立つ．式 (18.2), (18.3) の結果を用いて

$$c\left(\sum_{i=1}^{n-1}\frac{w_ix_i}{1-w_n}\right) \geq \sum_{i=1}^{n-1}\frac{w_ic(x_i)}{1-w_n}$$

より

$$c\left(\sum_{i=1}^{n} w_ix_i\right) \geq \sum_{i=1}^{n} w_ic(x_i)$$

を得る．

$w_i = f(x_i)$ とし，$f(x_i)$ は離散型の確率変数 X の p.d.f とすると、$c[E(X)] \geq E[c(X)]$ は直ちに従う．連続型の場合は積分の形式にする必要があるが，同様に成り立つ．

19 統計量

一般に統計量とは確率変数の関数である．また，統計量そのものも確率変数である．よく用いられる標本平均と標本分散を紹介する．まず，データセットとして $x_1, x_2, \ldots, x_n$ が与えられたとする．この時

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\ s^2 &= \frac{1}{n-1}\sum_{n-1}^{1}(x_i - \bar{x})^2 \end{aligned}$$

をそれぞれ標本平均，標本分散という．データは実現値として与えられ，標本平均，標本分散も実現値として得られるが，これらの統計量は確率変数同士の演算によって得られていることから，明らかに確率的な振る舞いをすることがわかる．

$t(\mathbf{x})$ という統計量が与えられた時，$\mathbf{x}$ の p.d.f は次のように表せる．

$$f_\theta(\mathbf{x}) = h(\mathbf{x})g_\theta\{t(\mathbf{x})\}$$

ここで h は $\boldsymbol{\theta}$ に依存しない．また g は $t(\mathbf{x})$ を通して $\mathbf{x}$ にのみ依存する．このような統計量を $\boldsymbol{\theta}$ に対する十分統計量と呼ぶ．その意図するところは，推定したいパラメーター $\boldsymbol{\theta}$ に対する全ての情報が $t(\mathbf{x})$ を通して $\mathbf{x}$ から得られているという意味である．ここで使う’情報’という言葉のより詳しい定義については４章で扱う．また，十分という点については，$t(\mathbf{x})$ が与えられた下での $\mathbf{x}$ の分布が $\boldsymbol{\theta}$ に依存しないことを表している．

20 Exercises