母集団と標本

□母集団　指標の値の分布あるいは法則

□標本　　母集団からの抽出（母集団の分布に従う）により得たデータ

注意：いままでやってきた平均，分散，共分散，標準偏差，相関係数は標本に関するもので，区別のため「標本」をつけて標本平均，標本分散，標本標準偏差，標本共分散，標本相関係数と呼ぶ

例：出生の性別　母集団の分布＝法則（男：105，女：100）

ある病院で1ヶ月に生まれた新生児の性別を調べる

→標本の抽出，観測，データ

標本（データ）の例　205人中女性101人男性104人

しばらく母集団について考える

母集団の分布を表現する

□ 母集団のデータは飛び飛びの値をとる

→母集団のデータは離散確率変数

確率変数＝確率的に決まるデータ指標

例：新生児の性別： 男に1という値を割り振り，女には0を割り振る

ある夫婦の子供の数：0，1，2，3，4，…

□離散確率変数の分布の表現：確率関数

例：新生児の性別

X＝1の確率（$\Pr[X=1]$と書く） $\frac{105}{205}$

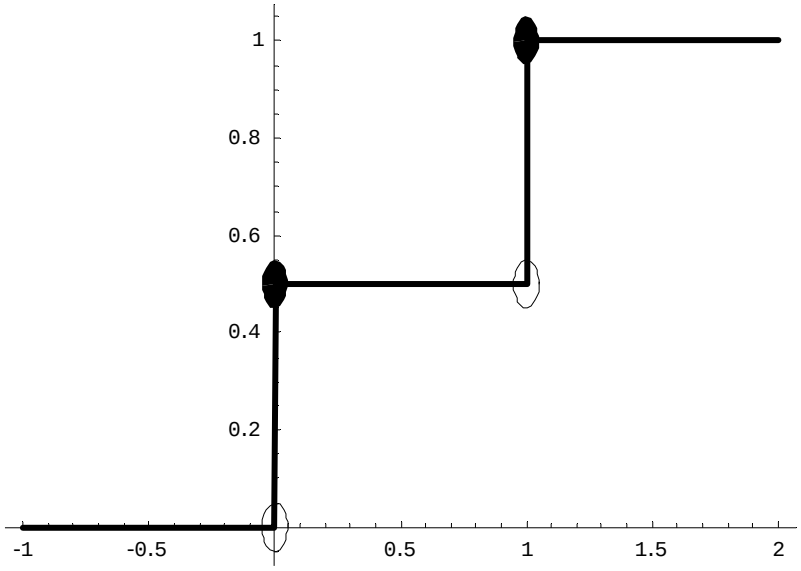
X＝0の確率（$\Pr[X=0]$と書く） $\frac{100}{205}$

□もうひとつの分布の表現：分布関数

$F(t)=\Pr[X \le t]$　(確率変数Xがt以下の確率)

例：新生児の性別

$$F(t)=\Pr[X \le t]=\begin{cases}\Pr[X<0]=0 & t<0\\ \Pr[X<1]=100/205 & 0\le t<1\\ \Pr[X\le 1]=1 & t\ge 1\end{cases}$$



後で出てくる連続確率変数(連続的データ)と統一的な扱いができるので分布関数を考える.

分布関数から確率を求める

$$\Pr[a < X \le b] = \Pr[X \le b] - \Pr[X \le a] = F(b) - F(a)$$

$$\begin{aligned}\Pr[X = a] &= \lim_{c \uparrow a} \Pr[c < X \le a] = \Pr[X \le a] - \lim_{c \uparrow a} \Pr[X \le c] \\ &= F(a) - \lim_{c \uparrow a} F(c)\end{aligned}$$

$$\begin{aligned}\Pr[a \le X \le b] &= \Pr[X \le b] - \Pr[X < a] = \Pr[X \le b] - (\Pr[X \le a] - \Pr[X = a]) \\ &= F(b) - \left\{F(a) - \left(F(a) - \lim_{c \uparrow a} F(c)\right)\right\} \\ &= F(b) - \lim_{c \uparrow a} F(c)\end{aligned}$$

分布の特性を示す指標

□ ⟨母集団⟩ 平均または分布の平均, 期待値　　注意：標本平均と区別せよ

⟨値×その値をとる確率⟩ の合計　$E[X]$と書く

→データの中心を代表

例：新生児の性別

$$E[X] = 0 \times \frac{100}{205} + 1 \times \frac{105}{205} = \frac{105}{205}$$

□ ⟨母集団⟩ 分散または分布の分散　注意：標本分散と区別せよ

データの散らばり方を示す

$\{(\text{値} - \mathrm{E}[\mathrm{X}])^2 \times \text{その値をとる確率}\}$ の合計　$\mathrm{V}[\mathrm{X}]$

実は $(\text{値}^2 \times \text{その値の確率})$ の合計 $-$ 平均^2 **で計算することもできる**

例：新生児の性別

$$\mathrm{V}[\mathrm{X}] = \left(0 - \frac{105}{205}\right)^2 \times \frac{100}{205} + \left(1 - \frac{105}{205}\right)^2 \times \frac{105}{200} = 0.249851$$

□ ⟨母集団⟩ **標準偏差または分布の標準偏差**

$\sigma_{\mathrm{X}} = \sqrt{\mathrm{V}[\mathrm{X}]}$

二つのデータの間の相互関係を表す分布の特性を示す指標

□ 二つの変数の場合も二つのデータの相互関係を示す分布＝法則

⟨同時分布，または，結合分布のこと⟩

□ **分布の共分散（データ X と Y の）　注意：標本共分散と区別せよ**

$$\mathrm{Cov}[X,Y]=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]$$

$\mathrm{Cov}[X,Y]\neq 0$　　分布として相関がある

□**分布の相関係数**　　**注意：標本相関係数と区別せよ**

$$\mathrm{Cor}[X,Y]=\frac{\mathrm{Cov}[X,Y]}{\sigma_X\sigma_Y}$$

□分布の相関

$\mathrm{Cor}[X,Y]\neq 0$　　分布として相関がある

$-1\le \mathrm{Cor}[X,Y]\le 1$　　$|\mathrm{Cor}[X]|$**が大きい　分布として相関が強い**

$|\mathrm{Cor}[X]|$**の正負が相関の正負を決める**

$|\mathrm{Cor}[X,Y]|=1$ならば$Y=\alpha+\beta X$の関係がある．(完全相関)

標本相関係数の場合は観測値が散布図上で一直線に並ぶ

分布の場合はどんな標本をとっても必ず$Y=\alpha+\beta X$になる

連続確率変数

□母集団のデータは連続の値をとる→母集団のデータは連続確率変数

例：身長，体重，実質 GDP（デフレーションの関係で），名目値は近似的に連続とみなせる

□連続確率変数の例

区間$[0,1]$のどの値も同様の確からしさでとる確率変数

物差しの上に錐を投げて，そこの目盛りをcm単位で読む．cm以下だけをとると上の確率変数を得る

一様確率変数と呼ぶ

□ 連続確率変数には確率関数は存在しない

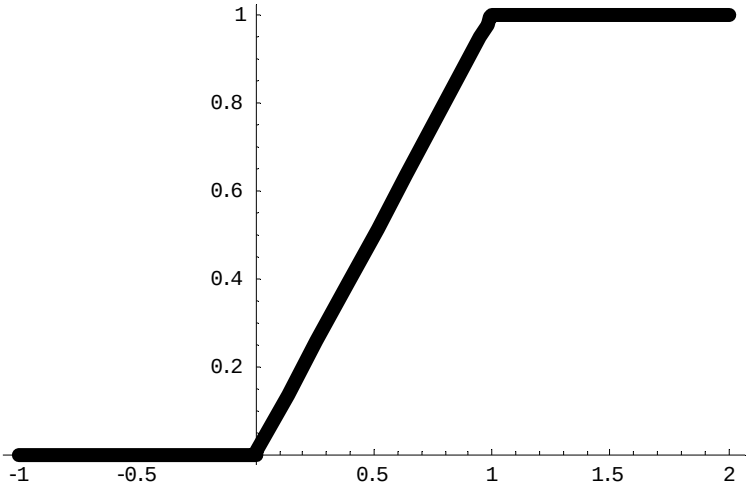
連続確率変数の分布関数

□ 分布関数　定義は離散確率変数の場合と同じ

$F(t) = \Pr[X \le t]$　(確率変数Xがt以下の確率)

例：標準一様確率変数の場合 (この分布を標準一様分布と呼ぶ)

$$F(t) = \begin{cases} 0 & x \le 0 \\ t & 0 < x < 1 \\ 1 & x \ge 1 \end{cases}$$



注意：連続確率変数の分布関数は必ず連続関数である.

□分布関数から確率を求める（連続確率変数の場合）

$$\Pr[a < X \le b] = \Pr[X \le b] - \Pr[X \le a] = F(b) - F(a)$$

$$\Pr[X = a] = \Pr[X \le a] - \lim_{c \uparrow a} \Pr[X \le c] = F(a) - \lim_{c \uparrow a} F(c) = F(a) - F(a) = 0$$

分布関数$F(t)$が連続関数なので

$\lim_{c \uparrow a} \Pr[X \le c] = \lim_{c \uparrow a} F(c) = F(a)$だから

＜つまり連続確率変数では，ある特定の値になる確率は

0である＞

$$\begin{aligned}\Pr[a \le X \le b] &= \Pr[X \le b] - \Pr[X < a] \\ &= \Pr[X \le b] - (\Pr[X \le a] - \Pr[X = a]) \\ &= F(b) - \lim_{c \uparrow a} F(c) = F(b) - F(a)\end{aligned}$$

確率密度関数（連続の場合の確率関数）

$$\Pr[X = t] = \lim_{h \to 0} \Pr[t \le X \le t + h] = \lim_{h \to 0} \{F(t + h) - F(t)\} = F(t) - F(t)$$
$$= 0$$

$\Pr[t \le X \le t + h] \approx p\,h$ と近似

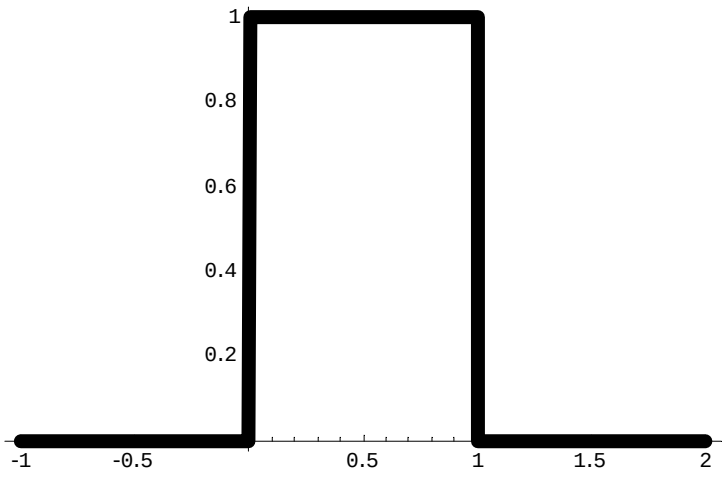
hの増加に対する確率の増加率pで X がtの近辺の値をとる

容易さを示す

確率変数 X の密度関数：$f(t) = \lim_{h \to 0} \frac{F(t + h) - F(h)}{h} = F'(t)$

密度関数の例：標準一様確率変数（標準一様分布）の場合

$$f(t) = \begin{cases} 1 & 0 \le t \le 1 \\ 0 & \text{それ以外} \end{cases}$$



分布の特性を示す指標（連続確率変数版）

□ （母集団）平均または分布の平均，期待値　　注意：標本平均と区別せよ

$\int_{-\infty}^{\infty} f(t)dt$　$E[X]$と書く

例：標準一様分布

$$E[X] = \int_{-\infty}^{0} t \cdot 0dt + \int_{0}^{1} t \cdot 1dt + \int_{1}^{\infty} t \cdot 0dt = \int_{0}^{1} tdt = \left[\frac{t^2}{2}\right]_0^1 = \frac{1}{2}$$

□ （母集団）分散または分布の分散　注意：標本分散と区別せよ

$\int_{-\infty}^{\infty} (t - E[X])^2 f(t)dt$　　$V[X]$

実は$\int_{-\infty}^{\infty} t^2 f(t)dt - E[X]^2$ **で計算することもできる**

例：標準一様分布

$$V[X] = \int_0^1 \left(t - \frac{1}{2}\right)^2 dt = \int_0^1 t^2 dt - \left(\frac{1}{2}\right)^2 = \left[\frac{t^3}{3}\right]_0^1 - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

□ （母集団）**標準偏差または分布の標準偏差**　　**離散の場合と同じ**

平均･分散と線形演算

□ **期待値（平均）の性質**

$E[X + Y] = E[X] + E[Y]$

$E[cX] = cE[X]$（cが確率変動しない定数の場合）

□ **分散の性質**

$$V[X+Y]=V[X]+2Cov[X,Y]+V[Y]$$

X, Y **が無相関なら** $V[X+Y]=V[X]+V[Y]$

$$V[cX]=c^2V[X]$$

確率変数の関数の分布

□単純な3段論法

確率変数の関数は確率変数．

確率変数は分布を持つ．よって，確率変数の関数も分布を持つ．

確率変数の関数の例：標本平均

よって，標本平均も分布を持つ

標本は分布をもつ母集団からの抽出で得られる．

標本の度数分布はこの母集団の分布とどう関わりがあるのか？

新生児の男女の比率はほぼ半々

しかし，世の中には3姉妹（3人子供がいる家庭の約$\frac{1}{8}$），

4姉妹（4人子供がいる家庭の約$\frac{1}{16}$）がいる．

たかだか3，4つの標本を観測しその度数分布をみただけでは，

母集団の分布はわからないものである．

結論：母集団の分布と標本の度数分布は別，標本の度数分布は母集団の分布を隠してしまうこともある．

標本平均と母集団平均との関係はどうなのか？

新生児の性別： 男に1という値を割り振り，女には0を割り振るという例

X の標本平均は$\frac{\text{男の合計}}{\text{標本新生児数}}$である．一方，母平均は約$\frac{1}{2}$．

ところが，3姉妹，4姉妹の場合，標本平均は0．

やはり，標本平均も母平均を裏切るのか？

今後の展開

標本平均の分布の解析が必要→標本平均は回帰モデル

$y_t = \hat{\alpha} + 残差$ の最小二乗法による $\hat{\alpha}$

だから最小二乗推定量と母平均の関係がわかればよい．最小二乗推定量の性質を見ていく．

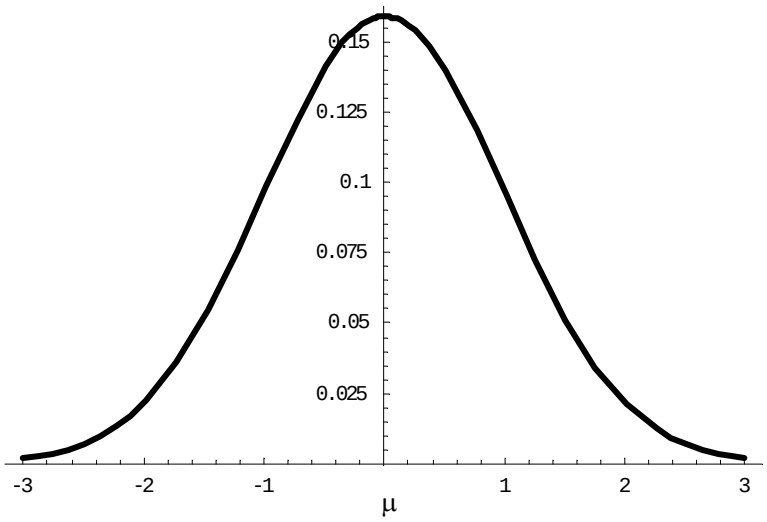
中心極限定理～正規分布での近似

□ 正規分布

平均 μ，分散 σ^2 の正規分布：$N(\mu, \sigma)$

確率密度関数：

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2} \right]$$



□ 中心極限定理

$X_1, \cdots, X_n$ が独立，$V[X_i] = \sigma^2$ のとき，