基礎統計シケプリ（２００５年夏)

2005 夏 火 1(倉田博史)
S1-22 高橋 講平

1　シケプリの概要

基礎統計（倉田先生）の過去問を見ればすぐわかりますが、基礎統計の試験の内容は主に基礎事項の定義と基本定理の使い方の確認に終始しています。それに則してこのシケプリは主に統計分析における基礎事項の定義と基本定理を要約したものになっています。ですから授業に出席し、ノートをとっている人にとっては価値が無いと思いますがご了承ください。

2　データの扱い

2.1　一次元データの扱い

ある一次元データの集まり x_1,x_2,,,,x_n に対して、平均、分散、標準偏差を

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad S = \sqrt{S^2}$$

と定義する。また、あるデータ x_i に対して

$$x_i = \bar{x} + z_i S$$

とあらわすことを x_i を標準化すると言う。

$$X = 50 + 10 \times z_i$$

を偏差値と言う。さらに $y_i = ax_i + b$ となる y_i に対して、

$$\bar{y} = a\bar{x} + b \quad {S_y}^2 = a{S_x}^2$$

が成り立つ。

2.2　二次元データの扱い

ある二次元データの集まり $(x_1, y_1), (x_2, y_2), , , , (x_n, y_n)$ に対して,

$$C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

を共分散と言い、$C_{xy} > 0$ を正の相関、$C_{xy} < 0$ を負の相関、$C_{xy} = 0$ を無相関 と言う。また、

$$r_{xy} = \frac{C_{xy}}{S_x S_y}$$

相関係数という。

このあたりのことはとりあえず事実を飲み込んでおけば何とかなります。これらの意味など知りたい人は教科書でもみてください。（ほんとは数値の意味を知らないとまったく無意味です。）

2.3 回帰分析

ある二次元データの集まあり $(x_1, y_1), (x_2, y_2), , , , (x_n, y_n)$ に対して x と y の関数形を一次関数の形で予想したい。そこで $y = ax + b$ とおいて、各データ x_i に対して y_i と $\hat{y} = ax_i + b$ の誤差の二乗の和を最小にするような定数 a,b 求める。すなわち、

$$L = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

を最小にする a,b を求めればよい。これを解くと一般に $b = \bar{y} - a\bar{x} \quad a = \frac{C_{xy}}{S_x^2}$ ともとまる。

定理

$$\sum (y - \bar{y})^2 = \sum (\hat{y_i} - \bar{y})^2 + \sum d_i$$

で A=$\sum (y - \bar{y})^2$ B=$\sum (\hat{y_i} - \bar{y})^2$ と置く。(決定係数) $= \frac{B}{A}$ と定義すると

$0 \leq$ (決定係数) ≤ 1

(決定係数) $= 1 \Leftrightarrow (x_1, y_1)(x_2, y_2), , , , (x_n, y_n)$ を座標平面上にプロットすると同一直線上

(決定係数) $= r_{xy}^2$

が成り立つ。

このあたりのことは図を見たほうがよくわかるのでよくわかりたい人は教科書を参照して下さい。

3 確率

3.1 確率の定義と諸公式

(1) 定義　P は Ω 上の確率である

$\overset{\text{def}}{\Leftrightarrow} 0 \leq \mathrm{P}(x) \leq 1, \forall x \quad \mathrm{P}(\Omega) = 1, \mathrm{P}(\phi) = 0$

$A_1, A_2, , , , A_n$が排反なら $\mathrm{P}(A_1 \cup A_2 \cup, , , \cup A_n) = P(A_1) + P(A_2) +, , , , ,$

(2) 諸公式

$A \cap B = \phi \Rightarrow P(A \cup B) = P(A) + P(B) \quad P(A^c) = 1 - P(A) \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$

3.2 条件付確率

(1) 定義　$P(B) \geq 0$ とすると $P(A/B) \overset{\text{def}}{=} \frac{P(A \cap B)}{P(B)}$ (B を条件とする A の条件付確率)

(2) 定理　$\Omega = H_1 \cup H_2 \cup, , , , H_n$のとき　$P(A) = \sum_{i=1}^{n} \frac{P(A/H_i)}{P(H_i)}$(全確率の公式)

3.3 事象の独立

A と B が互いに独立な事象ならば $P(A/B) = P(A) \Leftrightarrow P(A \cap B) = P(A)P(B)$ このあたりのことは普通な確率の話なのでかるく流しておきます。

4　確率変数

とりうる値ごとに確率があたえられる変数を確率変数といい、X、Y などとあらわす。

4.1　離散型確率変数

ある確率変数が $X = X_1, X_2, , , X_n$ といったようにとびとびの値しか取らない場合これをを離散型確率変数といい、その確率を確率を与える関数を f(x) とすれば

$$P(X = x_i) = f(x_i) \quad \sum_{k=1}^{n} f(x_i) = 1$$

となる。（というか確率の定義である。）また、離散型確率変数に対して期待値 E(X)、分散 V(X) を以下のように定義する。（$\phi(x_i) = x_i$ とすれば X の期待値すなわち平均が出る）

$$E(X) = \sum_{k=1}^{\infty} \phi(x_i) f(x_k) \quad V(X) = E(x_i - E(X))^2$$

またさらに次の定理がなりたつ。

$$E(ax + b) = aE(x) + b \quad V(ax + b) = a^2 V(x)$$

4.2　連続型確率変数

ある確率変数が $a \leq X \leq b$ といったようにある区間で連続的にすべての値をとりうる場合これを連続型確率変数といい、その確率を与える関数を g(x) とすれば

$$P(a \leq X \leq b) = \int_a^b f(x)dx \quad \int_{\text{全変域}} f(x)dx = 1$$

となる。また連続型確率変数に対して期待値 E(X)、分散 V(x) を次のように定義する。

$$E(x) = \int_{-\infty}^{+\infty} \phi(x) f(x)dx \quad V(X) = E(x_i - E(x))^2$$

4.3　標準化と標準偏差

確率変数 X に対して $\sigma = \sqrt{V(X)}$を標準偏差といい、$X = \mu + \sigma z_i$とあらわすことを X を標準化すると言う。また、定理 $E(X) = 0, V(x) = 1$ が成り立つ。

5　確率分布

5.1　Belnoull（ベルヌーイ）試行

$X = 0, 1$ のとき　$P(0) = 1 - p$　$P(1) = p$ なる試行を独立に n 回繰り返す試行を「長さ n」の Belnoulli 試行という。

5.2 二項分布

ある離散型確率変数 X に対して、$P(X = x_i) = nCx_i p^{x_i}(1-p)^{n-x_i}$ となる確率分布を二項分布といい、$X \sim Bi(n,p)$ とあらわす。このとき、

$$E(x) = np \quad V(x) = np(1-p)$$

5.3 Poisson（ポアソン）分布

ある離散型確率変数に対して $P(X = x_i) = e^{\lambda}\frac{\lambda^{x_i}}{x_i!}$ となる確率分布を Poisson 分布といい、$X \sim Po(\lambda)$ とあらわす。このとき、

$$E(X) = \lambda \quad V(X) = \lambda$$

5.4 幾何分布

ある離散型確率変数に対して $P(X = x_i) = p(1-p)^{x_i - 1}$ となる確率分布を幾何分布といい、$X \sim Ge(p)$ とあらわす。このとき、

$$E(X) = \frac{1}{p} \quad V(X) = \frac{1-p}{p\hat{}\ 2}$$

幾何分布に従う実例は「確率 p で表が出るコインを投げて、x_i 回目にはじめて重手が出る確率」などである。また、次の定理が成り立つ。

$$P(X = a + b/X > b) = P(X = a)$$

これを実例で説明すると、「すでに b 回投げていて表が出ていない状態でそのまま続けて a+b 回目に初めて表が出る確率は、はじめから a 回目に表が出る確率に等しい」ということを言っているだけであり、当然と言えば当然である。これを無記憶性といい、幾何分布は無記憶性を満たし、無記憶性を満たす分布は幾何分布である。

5.5 正規分布

ある連続型確率変数に対して $P(X = x_i) = \frac{1}{\sqrt{2\pi}}exp(-\frac{(x_i-\mu)^2}{2\sigma^2})$ となる分布を正規分布といい、$X \sim N(\mu\sigma^2)$ とあらわす。このとき

$$E(X) = \mu \quad V(X) = \sigma^2$$

（このことは証明は困難だそうです。）また、次の重要な定理が成り立つ。

$$X \sim N(\mu, \sigma^2) \Rightarrow aX + b \sim N(a\mu + b, a^2\sigma^2)$$

要するに $aX + b$ も正規分布に従うと言うのである。これを利用することにより、さまざまな確率がより計算しやすくなる。なぜこのようなことが役に立つかを説明しよう。一般に正規分布の確率密度関数に x を代入して計算（しかも連続型確率分布なので積分！）をすることは人間にはほぼ不可能である。だから積分を

計算した結果の表を与えるわけだが、μ, σ の値はさまざまな値をとるから、あらゆる μ, σ に対応する結果を与えなくてはいけないことになる。これもまた不可能である。ところが、上の定理を利用すればどんな μ, σ の値が与えられようとも、あるひとつの μ, σ の値に帰着させることができ、その表を与えればよい。ここで $\mu = 0, \sigma = 1$ の場合を標準正規分布といい、この場合に帰着させて考える。具体的には以下のとおりである。

$X \sim N(\mu, \sigma)$ のとき $Z = \dfrac{X - \mu}{\sigma}$ とすれば $Z \sim N(0, 1)$ となるので、

$$P(a \leq X \leq b) = P(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma})$$

と $N(0, 1)$ に帰着できる。

これの結果と表から確率を計算するわけですが、表の見方についてはここでは説明を省きます！表と長たらしい文章を掲載しなくてはいけないからです。しかしここは試験で確実にでる部分であり、表を読めないとまったく意味が無いので表の見方については教科書、六回目の講義プリントあるいはその他の書籍を参照してください。ごめんなさい。面倒くさいだけです。

5.6 指数分布

ある連続型確率変数に対して

$$P(X = x_i) = \begin{cases} \lambda e^{-\lambda x_i} & (x_i \geq 0) \\ 0 & (x_i < 0) \end{cases}$$

となる確率分布を指数分布といい、$X \sim Ex(\lambda)$ とあらわす。このとき、

$$E(X) = \frac{1}{\lambda} \quad V(X) = \frac{1}{\lambda^2}$$

6 二次元の確率変数

6.1 同時分布、周辺分布

ある二つの確率変数 (X, Y) に対して確率が $P(X = x_i, Y = y_j) = f(x_i, y_j)$ とあらわされるときこれを (X, Y) の同時確率分布といい、このとき$\phi(X, Y)$ の期待値を

$$E(\phi(X, Y)) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \phi(x_i, y_i) f(x_i, y_i)$$

と定義する。E(X) を求めようと思ったら、$\phi(X, Y) = X$ とすれば $E(X)$ がもとまる。さらに、$P(X = x_i) = \sum_{j=1}^{\infty} f(x_i, y_j)$ を X の周辺分布という。$E(Y)$、$V(X)$、$V(Y)$ も同様に定義どおりに求めればよい。

6.2 和の分布

X,Y を確率変数として、次の定理が成り立つ。

$E(X+Y) = E(X)+E(Y) \quad V(X+Y) = V(X)+V(Y)+2Cov(X, Y) \quad (Cov(X, Y) = E(X - E(X))(Y - E(Y))$ とする)

また、$Cov(X,Y) > 0$ を正の相関、$Cov(X,Y) < 0$ を負の相関、$Cov(X,Y) = 0$ を無相関という。さらに、$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$と定義すれば　$-1 \leq \rho(X,Y) \leq 1$　が成り立つ。特に$\rho(X,Y) = \pm 1$　ならば X と Y は直線的関係がある。

6.3　独立性

独立性の定義は、

$$P(X = x_i, Y = y_j) = P(x_i)P(y_i) \quad \forall(i,j)$$

であり、このとき定理として

$$E(XY) = E(X)E(Y)、Cov(X,Y) = 0、V(X+Y) = 0$$

が成り立つ。

6.4　n 個の独立変数に対する和の分布

$$* \begin{cases} X_1, X_2, , , X_n \text{は互いに独立} \\ X_1, X_2, , , X_n \sim \text{同一分布} \end{cases} \tag{1}$$

のとき、$E(X_1) = E(X_2) =, , , = E(X_n) = \mu$　$V(X_1) = V(X_2) =, , , = V(X_n) = \sigma$となり、その条件の下,以下の定理が成り立つ。

$$E(\bar{X}) = \mu \quad V(\bar{X}) = \frac{\sigma^2}{n}$$

6.5　再生性

$X_1, X_2, , , X_N$は互いに独立として
(1)$X_i \sim Bi(n_i, p) \Rightarrow \sum X_i \sim Bi(\sum n_i, p)$
(2)$X_i \sim Po(\lambda_i) \Rightarrow \sum X_i \sim Po(\sum \lambda_i)$
(3)$X_i \sim N(\mu, \sigma^2) \Rightarrow \sum X_i \sim N(\sum \mu_i, \sum \sigma_i^2)$
が成り立つ。これらを再生性という。(X_iは同一の種類の分布であってn, λ, μとσが異なってもよいことに注意。)

7　大数法則と中心極限定理

7.1　大数法則

＊の条件の下、

$$\forall\epsilon \quad \lim_{n\to\infty}(P(|\bar{X} - \mu|) \geq \epsilon) = 0$$

これはつまり、n が十分大きいときに（確率変数の数が十分多いとき）n 個の確率変数の値の平均は正確に期待値 μ を言い当てることを言っている。

7.2　中心極限定理

＊の条件の下、ｎが十分大きければ、

$$\bar{X} \simeq N(\mu, \frac{\sigma^2}{n})$$

8　標本分布

￥subsection 母集団と標本分布母集団から無作為抽出した取り出した $X_1, X_2, , , X_n$ から母集団がどのような確率にしたがうのか推測しようというのが統計分析の試みである。

8.1　統計量

$X_1, X_2, , , X_n$の関数 $f(X_1, X_2, , , X_n)$ を統計量といい、その分布を標本分布と言う。そのうち、$\bar{X}$：標本平均　S^2：標本分散　s^2：標本普不偏分散 $= \frac{1}{n-1}\sum(X_i - \bar{X})^2$ とすると＊の下で定理　$E(\bar{X} = \mu)$　$V(\bar{X} = \frac{\sigma^2}{n})$　$E(s^2) = \sigma^2$　$\bar{X}$と s^2は独立　が成り立つ。

8.2　正規母集団からの標本

(1)χ^2分布 $Z_1, Z_2, , , Z_n \sim N(0, 1)$ のとき

$$Y = \sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$$　（自由度ｋのχ^2分布）

と定義する。このとき次の定理が成り立つ

$X_1, X_2, , , X_n \sim N(\mu, \sigma^2)$ のとき　$\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

(2)t 分布 $X \sim N(0, 1)$、$Y \sim \chi^2(k)$、X と Y は独立、のとき

$$\frac{X}{\sqrt{\frac{Y}{n}}} = t(k)$$　（自由度ｋの ｔ 分布）

と定義する。このとき次の定理が成り立つ。

$X_1, X_2, , , X_n$のとき　$\dfrac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t(n-1)$

8.3　二標本問題

統計データを用いてある二つの母集団の比較を行うことを二標本問題と言う。ここで、ある二つの標本、$X_1, X_2, , , X_m \sim N(\mu_1, \sigma_1^2)$、$Y_1, Y_2, , , Y_n \sim N(\mu_2, \sigma_2^2$、$X_1, X_2, , , Y_n$はすべて独立　にたいして前述のとおり

$$\bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{m} \quad \bar{Y} \sim N(\mu_1, \frac{\sigma_2^2}{n}$$

が成り立つ。また、正規分布の性質より

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$$

が成り立つで $\bar{Z} = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$ と定義すれば、$Z \sim N(0,1)$ と標準化できる。母分散が未知であるが等しい場合は合併の分散

$$s^2 = \frac{1}{(m-1)+(n-1)}((m-1)s_1^2 + (n-1)s_2^2)$$

（ただし　$s_1^2 = \frac{1}{m-1}\sum(X_i - \bar{X})^2$　$s_2^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$）

を考えれば、

$$E(s^2) = \sigma^2、\frac{(m+n-2)s^2}{\sigma^2} \sim \chi^2(m+n-2)、\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{s^2(\frac{1}{m}+\frac{1}{n})}} \sim t(m+n-2)$$

となる。

9　最後に

基礎統計のシケプリはここまでです。いくつか不足している点があるのであげておきます。一つ目は、このシケプリは最後の三回ほどの授業の内容が含まれていないことです。つまり、６／２８現在までの内容と言うことになります。二つ目は演習問題および諸定理の証明を一切省いていることです。これはシケプリの内容が膨大な量になり手に負えなくなることと、一部の定理は授業でも証明せずに天下り的に与えられたからです。三つ目は致命的なのですが、このシケプリはいくつかの重要事項を欠いています。それはグラフを見ないと理解できない内容と、信頼区間に関する議論です。前者は文章に盛り込むのに手間がかかるためで、後者は文章での説明が困難だからです。そして前者者はプリントで、後者は教科書でカバーしてもらうしかありません。申し訳ありませんがそうしてください。あとは過去問を見ながらがんばってください。以上です。役に立ちそうに無いシケプリですいません。最後に一言だけ声を大にして言わせてください。「ぶっちゃけ過去問があればシケプリはいらねー。」