基礎統計（倉田博史教官・2005 夏）

○はじめに（このシケプリの構成について）

まず、シケプリという性質上、

① 試験の１週間前程度に読んで、十分に消化できること

② 単位が取れるほどの内容であること

の２点が必要であると考えました。

さらに、この基礎統計の講義が

① 講義、補足プリントが丁寧であり、教科書に忠実であること

② 非常に大人数の講義であること（＝他クラスのシケプリが手に入りやすい）

③ 定義の暗記では乗り切れないこと（＝問題が解けなくてはならない）

から、詳しい板書や定義の確認は教科書や、他クラスのシケプリを参考にしてもらうのが適当だという結論にいたりました。このシケプリではテストに出そうなところを絞って解説し、過去問の解答例を載せるという構成にしたいと思います。

○定義や基本公式

・**標本平均**　$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

・**標本分散**　$S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$

・**期待値**　$\mu = E(X) = \sum_{k=1}^{\infty} x_k f(x_k)$　**ただし、**$f(x_k) = P(X = x_k)$

・**分散**　$\sigma^2 = V(X) = E\{(X-\mu)^2\}$

・**標準偏差**　$\sigma = \sqrt{V(X)}$

・**標準化**　$X = \mu + z\sigma \quad \Leftrightarrow \quad z = \frac{X-\mu}{\sigma}$

によって z **を定めることを** X **の標準化という。**

※標準化は後に出てくる正規分布の計算において、非常に重要になってきます。この公式は必ず暗記しましょう。

・**単位の変換**

$E(aX+b) = aE(X) + b$

$V(aX+b) = a^2 V(X)$

○相関係数、回帰分析

用語と公式を覚えましょう。証明は意外と煩雑なものが多く、点数を狙うだけなら覚える必要のないものが多いです。暗記で乗り切りましょう。

・共分散 $C_{xy} = C_{yx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ **（定義）**

・相関係数 $r_{xy} = \frac{C_{xy}}{S_x S_y}$ **（定義）**

・回帰直線 $y = a + bx$ **は** $b = \frac{C_{xy}}{S_x^{\,2}}, a = \bar{y} - b\bar{x}$ **を満たす（定理）**

・決定係数 $R^2 = r_{xy}^{\,2}$ **を満たす（定理）**

＜２００４年度過去問より＞

50 **組の父子の身長を計測したところ、**$(x_1, y_1)...(x_{50}, y_{50})$ **なるデータが得られたものとする。**X **は父の身長、**y **は子供の身長とする。単位は** cm **である。このデータから以下のような数値が得られたとする。**

$$\bar{x} = \frac{1}{50}\sum_{i=1}^{50} x_i = 167.2\,,\ \bar{y} = \frac{1}{50}\sum_{i=1}^{50} y_i = 172.4$$

$$S_x^{\,2} = \frac{1}{50}\sum_{i=1}^{50}(x_i - \bar{x})^2 = 38.7\,,\ S_y^{\,2} = \frac{1}{50}\sum_{i=1}^{50}(y_i - \bar{y})^2 = 31.8$$

$$C_{xy} = \frac{1}{50}\sum_{i=1}^{50}(x_i - \bar{x})(y_i - \bar{y}) = 12.0$$

（１）回帰直線 y=a+bx **を計算せよ。（２）回帰係数** b **の値からどのようなことが分かるか。（３）決定係数を計算せよ。**

＜解＞公式に当てはめましょう。

（１） $b = \frac{C_{xy}}{S_x^{\,2}} \approx 0.310$ $a = \bar{y} - b\bar{x} \approx 120.6$ **より、** $y = 120.6 + 0.310x$

（２）父の身長差は子の身長差より大きい。（３） $R^2 = r_{xy}^{\,2} \approx 0.117$

○二項分布、Poisson 分布、幾何分布

Bernoulli 試行において現れる分布です。過去問を見る限り、公式に当てはめるだけの大問が１題出ると予想されます。

＜定義＞

難しい説明をすれば「n 個の独立な事象が～」といった風になりますが、要はコイン投げに帰着される試行が従う分布です。

基本は二項分布。すなわち　$P(X=x)={}_nC_x\,p^x(1-p)^{n-x}$ で表される分布。確率変数 X が二項分布（n 回の試行）に従うとき、$X \sim Bi(n, p)$ と記します。n が大きいときは Poisson 分布　${}_nC_x\,p^x(1-p)^{n-x} \to e^{-\lambda}\dfrac{\lambda^x}{x!}$　（ただし,$np=\lambda$=一定）

「x 回目ではじめて～」と出てきたときは幾何分布　$P(X=x)=p(1-p)^{x-1}$　を用います。

定理としては二項分布 $X \sim Bi(n, p)$ において成り立つ

$$E(X)=np, V(X)=np(1-p)$$

だけ覚えておけばよいでしょう。後に使うことになります。

＜２００４年度過去問より＞

（１）日本人の３０％はなんらかの宗教を信仰している。６人の日本人に宗教を進行しているか否かを尋ね、信仰していると答える人数を X とするとき、X=２である確率（P=2）を求めよ。

（２）日本人の０．２パーセントは自分を上流階級と考えている。１０００人の日本人に自分が上流階級か否かを尋ね、上流階級と答える人の人数を X とするとき、X=3 である確率（P=3）を求めよ。

（３）日本人の６０％は北枕を嫌う。団体旅行の添乗員が宿泊先の都合上、客の誰かに北枕で寝てもらうことを順々に頼まなければならない。X 人目ではじめて北枕 OK の返事をもらえるとするとき、確率 P（X≦3）を求めよ。

（１）全ての事象が互いに独立な試行（Bernoulli 試行）なので、二項分布に従います。

すなわち、$P(X=x)={}_nC_x\,p^x(1-p)^{n-x}$　が成り立ちます。これを用います。

＜解＞

$$(P=2)={}_6C_2(0.3)^2(1-0.3)^{6-2}=0.324135\approx 0.324$$

（２）Bernoulli 試行において $np=\lambda$（一定）としたとき、n が非常に大きくなり、 p が非常に小さくなると、

${}_nC_x\,p^x(1-p)^{n-x}\to e^{-\lambda}\dfrac{\lambda^x}{x!}$　が成り立ちます。これを用います。

＜解＞今、$n=1000,\ p=0.002$ なので、$\lambda=np=2$

$$(P=3)=(2.7)^{-2}\frac{2^3}{3!}\approx 0.183$$

（３）「x 回目ではじめて～」と出てきたときは、幾何分布の公式

$P(X=x)=p(1-p)^{x-1}$　を用いましょう。

＜解＞「北枕を嫌わない人」は４０％であることに注意して、

$$P(X\leqq 3)=P(X=1)+P(X=2)+P(X=3)$$
$$=0.4+(0.4)(1-0.4)+(0.4)(1-0.4)^2=0.784$$

○Chebyshev（チェビシェフ）の不等式

過去問で出ているのでやっておきます。確率変数がある範囲にどの程度の確率で入っているかを評価するのに使えます。ちなみにチェビシェフはＰｏｃｃｉｙａの数学者です。

$E(X)=\mu, V(X)=\sigma^2$ **であるとき、任意の** $k \geqq 0$ **に対して**

$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$　　$P(|X-\mu| \leq k\sigma) \geq 1-\frac{1}{k^2}$　**が成り立つ。**

＜２００４年度過去問より＞

コインを 10000 **回投げるとき、表の出る回数が** 4850 **回以上、**5150 **回以下である事実をチェビシェフの不等式を用いて評価せよ。**

$P(4850 \leq X \leq 5150)$ **をチェビシェフの不等式を用いて評価します。**

$P(|X-\mu| \leq k\sigma) = P(-k\sigma \leq |X-\mu| \leq k\sigma)$　**であることに注意しましょう。**

＜解＞表の出る回数を X **とおけば、**

$P(X=x) = {}_{10000}C_x \left(\frac{1}{2}\right)^{10000}$　　$(x=0,1,\cdots,10000)$

二項分布において、

$\mu = np = 5000$　$\sigma^2 = np(1-p) = 2500$　$\sigma = 50$　**だから**

$$P(4850 \leq X \leq 5150) = P(5000-3\times 50 \leq X \leq 5000+3\times 50)$$

$$= P(\mu-3\sigma \leq X \leq \mu+3\sigma) \geq 1-\frac{1}{3^2} = \frac{8}{9} \approx 0.889$$

○正規分布の計算

正規分布の計算は確実に出るそうです。必要なことは標準正規分布に変形して、数表の値を当てはめるだけなので、公式化して覚えてしまいましょう。

X が期待値μ、分散σ^2 の正規分布に従うことを $X \sim N(\mu,\sigma^2)$ **と書く。**

標準正規分布 $Z \sim N(0,1)$ **において、** $P(Z \le z) = \Phi(z)$**、と置きます。**
さらに、 $Q(z) = 1-\Phi(z)$ **とする。すなわち** $Q(z) = P(Z \ge z)$ **です。これを正規分布の上側確率といいます。この値が表で与えられるので、** $P(Z \le z)$ **を求めることができます。**

一般の場合、つまり $X \sim N(\mu,\sigma^2)$ **であるとき、** $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$ **と標準化し、**

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

$$= \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$ **とすれば計算することができます。**

＜２００４年度過去問から＞

確率変数 X **は正規分布** N(50,100)**に従うものとする。確率（ｉ）** $P(70 \le X)$ **（ii）** $P(40 \le X \le 60)$ **（iii）コインを** 10000 **回投げるとき、表の出る回数が** 4850 **回以上、** 5150 **回以下である事実をチェビシェフの不等式を用いて評価せよ。**

をそれぞれ求めよ。

＜解＞ $Z = \dfrac{X-50}{10} \sim N(0,1)$ **であるから、**

（ｉ） $P(70 \le X) = P(\dfrac{70-50}{10} \le \dfrac{X-50}{10}) = P(2 \le Z) = 0.22750 \approx 0.228$

（ii）標準正規分布密度関数の対象性から $\Phi(-1) = 1-\Phi(1)$ **が成り立ちます。**

$$P(40 \le X \le 60) = P(\frac{40-50}{10} \le \frac{X-50}{10} \le \frac{60-50}{10}) = P(-1 \le Z \le 1)$$

$$= \Phi(1) - \Phi(-1) = 0.84134 - 0.15866 = 0.68268 \approx 0.683$$

（iii） $P(X \le 55) = P(\dfrac{X-50}{10} \le \dfrac{55-50}{10}) = P(Z \le 0.5)$

$$= 1 - 0.30854 = 0.69146 \approx 0.691$$

○大数法則（LLN)、中心極限定理（CLT）

まず大数法則。教科書を読むと、難しい式が並んでいますが、つまりは標本（つまりは n）**が大きいとき、観察された標本平均を母集団の真の平均（母平均）とみなしてよい。ということです。未知の母平均を近似するのに使います。**

次に中心極限定理. $X_1, X_2, \ldots, X_n$ **がすべて互いに独立で、同一の分布に従うとき、**

$\overline{X}$ **は近似的に正規分布** $N(\mu, \frac{\sigma^2}{n})$ **に従うとしてよい。という定理です。**

＜２００４年度過去問より＞LLN **を使う問題は後でやりましょう。まずは** CLT。

コインを 10000 **回投げるとき、表の出る回数が** 4850 **回以上、**5150 **回以下である事実を中心極限定理を用いて評価せよ。**

表が出ることを X=1,**裏が出ることを** X=0 **で表したときの、**$P(0.485 \le \overline{X} \le 0.515)$ **を評価すればよいわけです。** $\overline{X}$ **が正確にはわからないので、中心極限定理を用いて正規分布と近似的にみなして計算します。**

また、 $\overline{X}$ **については** $E(\overline{X}) = p, V(\overline{X}) = \frac{p(1-p)}{n}$ **という定理が成り立ちます。**

＜解＞ $X_1 \ldots X_{10000}$ **は互いに独立に同一の二項分布に従う。コインを１回投げて、表が出ることを** $X_i = 1$**、裏が出ることを** $X_i = 0$ **で表すと、**

$P(X_i = 0) = 1 - p = 0.5 \quad P(X_i = 1) = p = 0.5$ **また、二項分布なので**

$E(X_i) = p = 0.5 \qquad V(X_i) = p(1-p) = 0.25$

このとき定理より、 $E(\overline{X}) = p = 0.5 \qquad V(\overline{X}) = \frac{p(1-p)}{n} = 0.005^2$

ここで、中心極限定理により $\overline{X} \sim N(0.5, 0.005^2)$ **が近似的に成り立つ。**

従って、 $Z = \frac{\overline{X} - 0.5}{\sqrt{0.005^2}} \sim N(0,1)$ **が近似的に成り立つ。**

よって $P(0.485 \le \overline{X} \le 0.515) = P(0.5 - 3 \times 0.005 \le \overline{X} \le 0.5 + 3 \times 0.005)$

$$= P(-3 \le Z \le 3) \approx \Phi(3) - \Phi(-3)$$
$$= (1 - 0.0013499) - 0.0013499 \approx 0.997$$

○CLT を用いた区間推定

区間推定とは、事前にα（0＜α＜1）を定め、
$P(L \le \theta \le U) = 1-\alpha$ **となるような（L,U）を定めることをいいます。通常α は** 0.05 **や** 0.01 **などで、**$1-\alpha$ **を信頼係数、（L,U）を信頼区間といいます。**

＜２００４年度過去問より＞
ある種子の発芽率を p **とする。**100 **個の種子を観察したところ、発芽したものは** 70 **個であった。中心極限定理を用いて** p **に関する信頼係数** 0.95 **の信頼区間を作れ。**

＜解＞
$X_1 \dots X_{100} \sim Bi(1, p)$ **であり、** $E(X_i) = p \quad V(X_i) = p(1-p)$

中心極限定理により、 $\overline{X}$ **は近似的に** $N(p, \frac{p(1-p)}{n})$ **に従うとしてよい。これを標準化すると、** $Z = \frac{\overline{X} - p}{\sqrt{p(1-p)/n}}$ **は近似的に** $N(0,1)$ **に従うとしてよい。これより、**

$$P(-1.96 \le \frac{\overline{X} - p}{\sqrt{p(1-p)/n}} \le 1.96) = 0.95$$

$$\Leftrightarrow P(\overline{X} - 1.96\sqrt{p(1-p)/n} \le p \le \overline{X} + 1.96\sqrt{p(1-p)/n}) = 0.95$$

ここで大数法則より、 $p(1-p) \quad \overline{X}(1-\overline{X})$ **とできる。**

$\overline{X} = 0.7, n = 100$ **を代入すると、求める信頼区間は** [0.610,790]

○χ^2分布、ｔ分布による区間推定

まず、必要なのは

標本普遍分散　$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$　です。

・χ^2分布

定義は教科書や、板書で確認してください。寧ろ、次の定理が重要だと思います。

＜定理＞

$X_1, X_2, \ldots, X_n$が、すべて互いに独立で$N(\mu,\sigma^2)$に従うとき、

$$\frac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1)$$

式の形から分かるように、標本普遍分散から、母分散σ^2を検定するのに使えます。

・ｔ分布

こっちも定理だけ。

＜定理＞

$$\frac{\overline{X}-\mu}{\sqrt{s^2/n}}\sim t(n-1)$$

こっちも式の形からわかるように、母分散が未知のときに母平均μを検定するのに使えます。

＜２００４年度過去問より（一部改題）＞

ある企業は、生産工程で必要となる工業原料をＡ社から購入しているとする。Ａ社を調べたところ、Ａ社の標本平均　$\overline{X}=\frac{1}{10}\sum_{i=1}^{10}X_i=14.4\%$,標本普遍分散

$s^2=\frac{1}{10-1}\sum_{i=1}^{10}(X_i-\overline{X})^2=4.90$であった。Ａ社の標本は正規母集団$N(\mu,\sigma^2)$からの無作為標本であると仮定できるものとする。

（１）Ａ社の母平均μ関する信頼係数0.95の区間を作れ。

（２）Ａ社の母分散σ^2に関する信頼係数0.95の区間を作れ。

（1）は母分散が未知の状態において母平均に関する推定なのでｔ分布を、（2）は母分散に関する推定なのでχ^2分布を利用しましょう。

＜解＞（1） $\dfrac{\overline{X}-\mu}{\sqrt{s^2/n}} \sim t(n-1)$ より

$$P(-t_{\alpha/2}(n-1) \le \frac{\overline{X}-\mu}{\sqrt{s^2/n}} \le t_{\alpha/2}(n-1)) = 1-\alpha$$

$$\Leftrightarrow P(\overline{X}-t_{\alpha/2}(n-1)\sqrt{s^2/n} \le \mu \le \overline{X}+t_{\alpha/2}(n-1)\sqrt{s^2/n}) = 1-\alpha$$

$\alpha = 0.05, \overline{X} = 14.4, s^2 = 4.90, n = 10$ **を代入して、求める信頼区間は** [12.8, 16.0]

（2） $\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$ より

$$P(\chi^2{}_{1-\alpha/2}(n-1) \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2{}_{\alpha/2}(n-1)) = 1-\alpha$$

$$\Leftrightarrow P(\frac{(n-1)s^2}{\chi^2{}_{\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{P(\chi^2{}_{1-\alpha/2}(n-1)}) = 1-\alpha$$

$\alpha = 0.05, s^2 = 4.90, n = 10$ **を代入して、求める信頼区間は** [2.61,13.3]

○検定

有意水準α（第一種の誤りを犯す確率）を定め、それより大きいか、それ以下かを考えます。例えば母分散が既知で、正規分布 $N(\mu,\sigma^2)$ に従うとき、帰無仮説 $H_0:\mu_1=\mu_2$,対立仮説 $H_1:\mu_1\neq\mu_2$ とすれば、

$\dfrac{\left|\overline{X}-\mu\right|}{\sqrt{\sigma^2/n}}>Z_{\alpha/2}$ ならば、H_0 を棄却し、$\dfrac{\left|\overline{X}-\mu\right|}{\sqrt{\sigma^2/n}}\leq Z_{\alpha/2}$ ならば、H_0 を棄却しません。

母分散が未知で母平均を検定するときにはｔ分布、母分散を検定するときには χ^2 分布を使います。やり方は同様です。

○２標本問題

（ｉ）２つの正規母集団 $N(\mu_1,{\sigma_1}^2),N(\mu_2,{\sigma_2}^2)$ があり、それぞれから、大きさ m,n の標本 $X_1,X_2,\ldots,X_m$ $Y_1,Y_2,\ldots,Y_n$ を取り出す。このとき、$H_0:\mu_1=\mu_2$, $H_1:\mu_1\neq\mu_2$ とする。２つの母分散が ${\sigma_1}^2={\sigma_2}^2=\sigma^2$ であるならば、$s^2=\dfrac{(m-1){s_1}^2+(n-1){s_2}^2}{m+n-2}$ とすると、H_0 のもとで、$t=\dfrac{\overline{X}-\overline{Y}}{s\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}}\sim t(m+n-2)$

よって $|t|\leq t_{\alpha/2}(m+n-2)$ のとき、H_0 を棄却せず、、そうでないときには棄却する。

（ii）$H_0:{\sigma_1}^2={\sigma_2}^2, H_1:{\sigma_1}^2>{\sigma_2}^2$ で、$F={s_1}^2/{s_2}^2$ とすると、$F\sim F(m-1,n-1)$

よって、$F\leq F_\alpha(m-1,n-1)$ のとき、H_0 を棄却せず、そうでないとき棄却する。

＜２００４年度過去問より＞

ある企業は、生産工程で必要となる工業原料をＡ社とＢ社から購入しているとする。Ａ社とＢ社を調べたところ、Ａ社の標本平均　$\overline{X}=\dfrac{1}{10}\sum_{i=1}^{10}X_i=14.4\%$,標本普遍分散 ${s_1}^2=\dfrac{1}{10-1}\sum_{i=1}^{10}(X_i-\overline{X})^2=4.90$,Ｂ社の標本平均 $\overline{Y}=\dfrac{1}{10}\sum_{i=1}^{10}Y_i=17.9(\%)$,標本普遍分散 $s^2=\dfrac{1}{10-1}\sum_{i=1}^{10}(Y_i-\overline{Y})^2=2.10$ であった。Ａ社とＢ社ののの標本はそれぞれ正規母集

団 $N(\mu_1,\sigma_1^{\,2})$, $N(\mu_2,\sigma_2^{\,2})$ からの無作為標本であると仮定できるものとする。

（1）母分散に関して $\sigma_1^{\,2}=\sigma_2^{\,2}$ が成立するものとして、帰無仮説 $H_0:\mu_1=\mu_2$ を対立仮説 $H_1:\mu_1\neq\mu_2$ に対して有意水準 0.05 で検定せよ。

（2）帰無仮説 $H_0:\sigma_1^{\,2}=\sigma_2^{\,2}$ を対立仮説 $H_1:\sigma_1^{\,2}>\sigma_2^{\,2}$ に対して有意水準 0.05 で検定せよ。

＜解＞（1） $s^2=\dfrac{(10-1)s_1^{\,2}+(10-1)s_2^{\,2}}{10+10-2}=3.50$

$$|t|=\frac{\left|\overline{X}-\overline{Y}\right|}{s\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}}\approx 4.47\ ,\ t_{0.025}(10+10-2)=2.101$$

$|t|>t_{\alpha/2}(m+n-2)$ なので、H_0 は棄却される。

（2） $F=s_1^{\,2}/s_2^{\,2}\approx 2.33 \quad F_{0.05}(10-1,10-1)=3.179$

$F\le F_{\alpha}(m-1,n-1)$ なので、H_0 は棄却されない。