基礎統計試験対策プリント

注１：対象は、授業に出ていない人と、教科書を持っていない人です。それ以外の人はよんでもあまり意味はないかもしれません。

注２：松原が試験に出さないと言ったありとあらゆる証明は省いてあります。教科書をもっていないひとで証明を知りたいかたは持っている人に見せてもらってください。

注３：試験は教科書、、副読本、自筆ノート、電卓持ち込み可なので、持っている人は忘れずにもっていきましょう。携帯を電卓として使ってはいけないそうです。また、ノートのコピー、シケプリは持ち込み不可ですが、抜け道はあります。シケプリを持ち込んで見つかったら「俺は TeX でノートを取ってるんだ！」と言い張りましょう。私はどうなっても責任を取りませんが。

1　統計の取り方

1.1　全数調査

母集団全体を調査。例：国勢調査、ローマの戸口調査

1.2　サンプル調査（標本調査）

母集団の一部のみを調査。（ただし、無作為にサンプルを選ばなければならない）

2　データの種類

2.1　時系列データ、クロスセクションデータ

ある対象についての異なった時点におけるデータと、異なる対象についてのデータ。

2.2　第一義的統計、第二義的統計

統計資料の作成が目的である調査の集計（国勢調査報告とか）と、もともとは統計資料の作成が目的ではない資料の集計（犯罪統計とか）。後者は業務統計ともいう。

違いがまったくわかりません。ごめんなさい。

2.3　一次統計、二次統計

データそのものと、データが加工されたもの（物価指数とか）。

3　代表値

集団を代表する値。
主なものを次に列挙します。

3.1 最頻値（モード）

集団の中で一番多い値。(そのまんま)

3.2 中央値（メディアン）

集団の中でちょうど（順位が）真ん中の値。

3.3 平均値

相加平均（算術平均ともいう）のこと。集団 $\{x_1, x_2, \ldots, x_n\}$（以下もだいたいこの集団について考える）について、平均値を $\bar{x}$ とすると、

$$\begin{aligned} \bar{x} &= \frac{x_1, x_2, \ldots, x_n}{n} \\ &= \sum_{k=1}^{n} x_k \end{aligned}$$

全部足して個数で割っただけですね。

平均値は、モード、メディアン、平均値の三つのなかで一番大きくなることが多い。

3.4 例

集団 {2,2,2,3,3,5,7,7,9,10} のモード、メディアン、平均値は？

答え；

- モード：一番多い 2。
- メディアン：集団の要素が偶数個なので、真ん中の値がない。こういうときは、3 と 5 のあいだをとって 4 がメディアン。
- 平均値：全部足したら 50。平均値は、これを個数 10 でわった 5。

3.5 例２ (過去問から)

M 教授は学期試験の 500 人の得点分布 (度数分布表) が、10 点ごとの階級区分で下から

5,5,10,10,20,20,10,10,5,5（%)

となったのに対し、$y = 10\sqrt{x}$ の変換を行った。変換前、変換後の平均値を求めなさい。ただし、計算の上で適当な仮定をおいてよい。

説明など：…この試験満点いないんでしょうか。0 から 100 点って数字 101 個あるから下から分けると 100 点のところだけ余りますが。ま、どうでもいいんですけど。100 点がいないほうが次の仮定が妥当ですし。

「適当な仮定」として、5 点が 5%、15 点が 5%、25 点が 10%、…95 点が 5%いたとして計算します。すると、変換前の平均値は、(5 × 5+15 × 5+25 × 10+…+95 × 5) ÷ 100=50

変換後の平均値を同様に $10\sqrt{5}$点が 5%、$\cdots 10\sqrt{95}$点が 5% として計算してよいかは疑問の残るところです。(実行すると 68.4 になる)

なぜなら、たとえば 10 人いて全員 5 点である場合と、0 点から 9 点までひとりづついる場合では変換後の平均値が変わるからです。（前者は 2.24、後者は 1.93）上の計算では前者であると仮定しています。しかし、現実には後者のほうを考えるのが自然であると思われます。

しかし、後者のほうで計算しようとすると、0 から 99 まですべての整数の平方根を出さなければいけないので別に 68.4 でいいんじゃないかな、と思います。

4 散らばり

データの散らばり方。

4.1 レンジ

最大値と最小値の差。レンジを R とすると、$R = \max\{x_1, x_2 \ldots, x_n\} - \min\{x_1, x_2 \ldots, x_n\}$

松原曰く、あまり重要じゃない、と。

4.2 平均偏差

平均偏差を d とすると、

$$d = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|}{n}$$

各観測値が平均からどれだけ離れているかの平均を取った。

絶対値をはずすと $d = 0$ になる。 $\ldots (\sharp)$

$(x_1 + x_2 + \ldots + x_n = \sum_{k=1}^{n} x_k \quad , \quad \bar{x} \times n = \sum_{k=1}^{n} x_k$ だから）

4.3 分散

分散を S^2 とすると、

$$S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}$$

分散を求めるとき、n ではなく、$n-1$ でわることがあります。これを不偏分散といい、s^2 で表すものとします。

自由に動ける変数が $(n-1)$ 個だからだそうです。($\sharp$ で、総和が 0 になることから、変数のうち $(n-1)$ 個が決まると残りのひとつは自動的に決まってしまう）

分散ではかっこを二乗しているために、単位（または次元）も二乗されてしまう（平均偏差は元の観測値と同じ単位を持つ）ので、単位をそろえるために、分散の平方根をとったものを考えます。これを 標準偏差 といいます。

$$S = \sqrt{S^2}$$

集団の要素 x_i について、標準得点Z を次のように定義します。

$$Z = \frac{x_i - \bar{x}}{S}$$

すると、偏差値T が次のように定義されます。

$$T = 10Z + 50$$

4.4 変動係数

（松原が、自分でやれ、と言ってたから大して重要じゃないんだろう）

平均値が大きくなると、標準偏差も大きくなるので、偏差値では散らばりが比較できなくなります。このようなとき、変動係数 $C.V. = \frac{S}{\bar{x}}$ により、平均値を考慮しながら散らばりが比較できます。

4.5 その他（四分位偏差）

レンジは端の異常値に左右されるので、異常値による影響を受けにくくするために、両端 $\frac{1}{4}$ を切り落として「半分に割った」ものが四分位偏差です。つまり、集団のうち順位が下から 25% 目の点を Q_1,75% 目の点を Q_3 とすると、四分位偏差 Q は、$Q = \frac{Q_3 - Q_1}{2}$

4.6 例 1

3.4 の集団の、分散と標準偏差、また、それぞれの要素の標準得点、偏差値は？

答え；$\bar{x} = 5$ より、

- 分散：$S^2 = \frac{(2-5)^2\times3+(3-5)^2\times2+(5-5)^2+(7-5)^2\times2+(9-5)^2+(10-5)^2}{10} = 8.4$
- 標準偏差：$S = \sqrt{S^2} = \sqrt{8.4} \sim 2.9$
- 標準得点, 偏差値：

$$\begin{aligned} 2 \longrightarrow Z &= \frac{2-5}{\sqrt{8.4}} = \frac{-3}{\sqrt{8.4}} \\ &\sim -1.0 \\ T &= 10Z + 50 \\ &\sim 40 \end{aligned}$$

2 以外は省略。

5 相関関係

2 つのデータ間の関係。2 つのデータ例の集合 $\{(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)\}(\ldots \natural)$ について、

x_i が大きいと y_i も大きいとき正の相関関係があるといい、

x_i が大きいと y_i は小さいようなとき、負の相関関係があるといいます。

たとえば、x を気温、y をクーラーの所有率とすると、気温が高いほうがクーラーの所有率が高い（はず）ので、x と y には正の相関関係があります。

5.1 相関係数

相関の程度を表す指標。♮について、相関係数を r とすると、

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

r が大きいほど正の相関が強く、小さいほど負の相関が強い。また、かならず $-1 \leq r \leq 1$ になります。(証明略)

これは積率相関係数とよばれるものですが、x_i, y_i の、x, y のなかの順位によって決まる、順位相関係数もあります。

スピアマンの定義によるものと、ケンドールの定義によるものがあり、こんなのテストに出したら大鬼だよな、と思いつつ念のため書いておきます。(そうでなくても過去問を見ると十分鬼だと思うが)

スピアマンの順位相関係数を r_s, ケンドールの順位相関係数を r_k, また、$R_i, R_i\prime$ をある観測対象 i の二つの基準による順位（つまり、♮で考えると、(x_i, y_i) について、例えば x_i が x のなかで 3 位、y_i が y のなかで 5 位だとすると、$R_i = 3, R_i\prime = 5$. また、例えば男は桜が花の中で 3 番目に好きで、女は桜が 5 番目に好きだとすると、桜を i とするとき同様の結果を得る）とすると、

$$\begin{aligned} r_s &= 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(R_i - R_i\prime)^2 \\ r_k &= \frac{G - H}{\frac{n(n-1)}{2}} \end{aligned}$$

ただし、観測対象のなかの 2 つの対 (i, j) ついて、

x, y どちらの集団でも i の順位が j の順位より勝って（負けて）いるとき、$+1$

片方の集団では i の順位が j の順位より勝っているのにもう片方の集団では逆なとき、-1

を与えるとすると、

$+1$ を与えた対の数が G、-1 を与えた対の数が H です。

分母の $\frac{n(n-1)}{2}$ は、対の選び方の総数（${}_nC_2$）。

r_s, r_k も、$-1 \leq r \leq 1$

5.2 例

6 回帰分析

この図のように、散布図にもっとも近い直線（平面などの場合もあるが）をあてはめること。

6.1 最小二乗法

この直線の式を求めます。直線：$y = bx + a \ldots (\diamondsuit)$ と、散布図上の点 (x_i, y_i) について、この直線が散布図上の点をすべて近似していると考えると、直線の式からは、x に x_i を代入して、y 座標：$bx_i + a$ が予想されますが、実際の値は y_i です。そこで、この y_i と $bx_i + a$ とのずれが、（すべての i を考慮したとき）もっとも小さくなる a, b を選びます。つまり、二乗和（ずれの二乗の和）L として、

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

これがもっとも小さくなるような a, b を選びます。これを 最小二乗法 といいます。（絶対値の和を考えずに、二乗の和を考えるのは、計算がやさしくなるから、ということにしておきましょう）

そのような a, b を $\hat{a}, \hat{b}$ とすると、L を a, b で偏微分して 0 とおいて出てくる連立方程式を解いて、

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{(\sum x_i y_i) - n\bar{x}\bar{y}}{(\sum x_i^2) - n\bar{x}^2}$$
$$\hat{a} = \bar{y} - b\bar{x}$$

これらを $\diamondsuit$ に代入して、回帰方程式：$y = \bar{b}x + \bar{a}$ が得られます。ところで、散布図で点がうまく直線っぽく並んでいるときには、回帰はそれなりに意味を持ちます。（x から y が予測できる）しかし、点がばらばらなときは、回帰はあまり意味を持ちません。（x から予測される y に信憑性がない）

そこで、「回帰がどれくらい効いたのか」を考えてみることにします。

$$\begin{aligned} SS_O &= \sum (y_i - \bar{y})^2 \\ SS_E &= \sum \{y_i - (\hat{b}x_i + \hat{a})\}^2 \end{aligned}$$

とすると、SS_O は回帰する前のばらつき（平均からのばらつき）で、SS_E は回帰した後のばらつき（直線からのばらつき）です。

すると、$SS_R = SS_O - SS_E$ は回帰によるばらつきの減少分で、$\dfrac{SS_R}{SS_O}$ を 決定係数 と定義すると、これが 1 に近いほど（つまり SS_E が小さいので）、x が y を決定する度合いが強くなります。

(x を独立変数、y を従属変数ということがある。説明変数、被説明変数ともいう)

また決定係数について、必ず次の式が成立します。

$$\frac{SS_R}{SS_O} = r^2$$

つまり、相関係数の 2 乗に等しい、と。（本当はこっちが決定係数の定義かもしれないけど）r が ± 1 に近いほど、直線の、グラフへのあてはまりが良い。

6.2 重回帰

変数が x, y だけではないとき、たとえば変数を $x_1, x_2, \ldots, x_n, y$ として、y を $x_1, x_2, \ldots, x_n$ から求めようとするなら、$y = b_1 x_1 + b_2 x_2 +, \ldots, +b_n x_n + a$ として回帰を行います。説明変数が 2 個以上あるとき、これを重回帰といいます。（2 個のときは、空間に直線を当てはめている）

6.3 多項式回帰

明らかに直線じゃあらわせないとき（直線で回帰しても無駄)、2 次式や 3 次式で回帰すること。

6.4 例（過去問から）

次の経済成長 (国内生産 y) の年次データにそのま回帰直線を当てはめることは不適切であることを示し、適切な方法を 1 通り提示しなさい。実行の必要はない。

t	1	2	3	4	5	6	7	8	9	10
y	254.0	257.0	260.2	263.1	266.2	269.5	273.2	276.9	280.4	284.1

$\bar{y} = 268.46, \bar{t} = 5.5$ より、

$$\sum (y_i - \bar{y})(t_i - \bar{t}) = (254 - 268.46)(1 - 5.5) + \cdots + (284.1 - 268.46)(10 - 5.5) = 275.9$$

$$\sqrt{\sum (y_i - \bar{y})^2} = \sqrt{(254 - 268.46)^2 + \cdots + (284.1 - 268.46)^2} = 30.4$$

$$\sqrt{\sum (t_i - \bar{t^2})} = \sqrt{(1 - 5.5)^2 + \cdots + (10 - 5.5)^2} = 9.08$$

$$\text{よって、} r(\leftarrow \text{相関係数}) = \frac{275.9}{30.4 \times 9.08} = 0.9995$$

$$\text{決定係数 } r^2 = 0.999$$

???なんで不適切なんでしょうか。すいません、素でわかりません。っていうか解答載せろよ、松原。相関係数の例を出しておきたかったのでやってみましたが。

7 確率の基礎

本当に基礎。高校でやったことは書きません。

たとえば、さいころを 1 個投げる問題で、可能な結果の全体は $\Omega = \{1,2,3,4,5,6\}$ で、このような集合を標本空間 Ω といいます。また、「偶数の目が出る」といった、標本空間の部分空間を、事象といい、それ以上分解できない事象（「5 が出る」とか）を、根源事象といいます。

あと、（A の）余事象は $\bar{A}$ だけでなくて、$A^c, \neg A$ とも書く。

7.1 確率とは何か

7.1.1 頻度説（客観確率）

600 回さいころ振って 1 が 100 回でたから、無限回振ったら確率 $\frac{1}{6}$ に収束するだろう、という考え。無限回行うことは不可能だから、理論的には厳密なものではない。

7.1.2 主観確率

研究者が、あらかじめある確率をかってに (とはいっても研究者の経験とかに基づいているが) 与えて、分析を行う。得られる確率は研究者によって違う。

7.2 数学的に厳密な確率概念は？

コルモゴロフによる確率の 3 公理に基づく概念

1. すべての事象 A に対して、$0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. 互いに排反な事象 $A_1, A_2, A_3, \ldots$ に対して、
$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$$

7.1,7.2 に松原は二重丸つけているが…いや別にいいんですけど。

7.3 条件付確率

たとえば、さいころを投げると、$P(\text{偶数}) = \dfrac{3}{6} = \dfrac{1}{2}$

しかし、4 以上の目だったときに限定して考えるとどうでしょうか。このように、事象 B のもとでの事象 A の確率を条件付確率といい、これを $P(A|B)$ とすると、

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \ldots(\spadesuit)$$

さっきのさいころの場合だと、$P(\text{偶数} \mid 4 \text{ 以上}) = \dfrac{P(\text{偶数かつ } 4 \text{ 以上})}{P(\text{偶数})} = \dfrac{2/6}{3/6} = \dfrac{2}{3}$

$P(A|B)$ は図の、A に対する黒色部分の割合と考えることもできます。

次に、トランプについて考えてみます。絵札が出る確率は、$P(\text{絵札}) = \dfrac{12}{52} = \dfrac{3}{13}$

黒いカードが出た、という事象の元での条件付確率は、$P(\text{絵札} \mid \text{黒}) = \dfrac{6/52}{26/52} = \dfrac{3}{13}$

どちらも同じ結果になりました。このように、$P(A) = P(A|B)$ であるとき、(A,B は無関係に起こり) これを、A,B は独立である といいます。

($\spadesuit$) から

$$\{A, B \text{ は独立である } \} \Leftrightarrow P(A) = P(A|B) \Leftrightarrow P(A \cap B) = P(A)P(B)$$

8 確率変数

8.1 確率変数

それぞれの値を、ある確率を持って、取る変数。たとえば、さいころの目 X(確率変数は大文字で表す) を確率変数とすると、$x = 1, 2, 3, 4, 5, 6$ なる確率はそれぞれ、$\dfrac{1}{6}$ だから、$P(X = 1) = P(X = 2) = P(X = 3) = P(X = 4) = P(X = 5) = P(X = 6) = \dfrac{1}{6}$

同様に、2 個のさいころの目の和を X とすると、$P(X=1)=\frac{1}{36}, P(X=2)=\frac{2}{36}\ldots$ 以下略。

8.2 確率分布

上の例のように、連続した値をとらない（ちょっと不正確な表現ですが…要はさいころが 3.6 とかが出る確率を考えることは無駄だと）確率変数 X を、離散型 といい、それぞれの値の確率 $P(X=x_k)=f(x_k)$ （f の方) を X の確率分布といいます。（上の例だと $\frac{1}{6}$ とか $\frac{2}{36}$ とか）

たとえば、ランダムに選んだ人間の身長の確率変数 X は連続した値をとります。（この世に 178.7cm の人間はいない、なんて事態はありえないですね。ただ 225cm がいて 224cm がいないとかは考えられますが確率 0 にすればいい）このような、連続型の確率分布の定義は、

$$P(A \leq X \leq B)=\int_a^b f(x)dx$$

ただし、$\int_{-\infty}^{+\infty} f(x)dx=1, f(x)\geq 0$ コルモゴロフがやっと役に立ったかな、という感じですね。離散型でもシグマをとれば 1 になりますし。あと、この f(x) を確率密度関数といいます。

8.3 累積分布関数

確率密度関数 f(x) に対して、ある x について図の黒い部分を表す関数 F(x) を、累積密度関数といいます。つまり、

$$F(x)=P(X\geq x)=\int_{-\infty}^{x} f(x)dx$$

また、f を積分すると F になり、F を微分すると f になります。

8.4 期待値

確率を考えた上での平均 (ていうかむしろ平均) で、たとえば、さいころの目の期待値が $1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \ldots + 6 \times 16$ であらわされるように、離散型の場合、一般に期待値 E(x)(または μ) は、

$$\begin{aligned} \mu = E(x) &= \sum_x xf(x) \\ &= \int_{-\infty}^{\infty} xf(x)dx \qquad \text{(連続型の場合)} \end{aligned}$$

期待値の性質として次のものがあげられます。

$$\begin{aligned} &(a) \qquad E(c) = c \\ &(b) \qquad E(X + c) = E(X) + c \\ &(c) \qquad E(cX) = cE(X) \\ &(d) \qquad E(X + Y) = E(X) + E(Y) \end{aligned}$$

いったん確率分布に戻ると、$f(x_k) = P(X = x_k)$　この X をいろいろいじってかんがえてみましょう。

- (a):確率変数が X=c、つまり常に定数しかとらないという意味なので、当然といっちゃあ、当然。
- (b):確率変数が X+c をとる確率が f(x) であるということ。(X+c)f(x) を積分すればいい。X 円分の宝くじと c 円の現金がいくらになるかを考えても良い。
- (c):確率変数が cX をとる確率が f(x) ということなので、cXf(x) を積分するか、または X 円分の宝くじを買った場合に比べて cX 円分の宝くじはどうなるかを考えても良い。
- (d):確率変数からアプローチするにはどうすればいいんだ？でも、2 種類の宝くじを買った場合を考えればよいでしょう。

8.5 分散

分散を V(X)（σ^2 とも書く）とすると、$V(X) = E\{(X - \mu)^2\}$ つまり、「X が期待値からどれだけ離れているか」（の二乗）の期待値です。上にあげた期待値の性質から、

$$\begin{aligned} V(X) &= E(X - \mu)^2 = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2E(\mu X) + E(\mu^2) \\ &= \cdots = E(X^2) - (E(X))^2 \end{aligned}$$

であることがわかります。最初の式よりこちらのほうが計算しやすいです。

2 個のさいころの目の分散を求めてみます。

$$E(X^2) = 2^2 \times \frac{1}{36} + 3^2 \times \frac{2}{36} + \ldots + 12^2 \times \frac{1}{36} = \frac{1974}{36} = \frac{329}{6}$$

$$\sigma^2 = E(X^2) - (E(X))^2 = \frac{329}{6} - 7^2 = \frac{35}{6} \text{(計算違うかもしれない)}$$

(…このケースは最初の式のほうが楽だったかもしれないなあ)

また、確率における分散も、平方根をとって、標準偏差D(X) が定義されます。

$$D(X) = \sqrt{V(X)} = \sigma$$

また、$V(-X) = V(X)$ です。つまり、グラフを裏返しても、ばらつきは変わらない。

ところで、E(X)、V(X) がきまっても、つまり位置とばらつきが決まっても、確率分布が一通りに決まるわけではありません。左の図が、右の図を E(X) を中心として反転したものだとすると、

左の図と右の図では E(X) も V(X) も変わりはありません。

そこで、非対称性をあらわすものとして、$\frac{E((X-\mu)^3)}{\sigma^3}$ で歪度が定義されます。左の図と右の図では、歪度が異なります。

同様に、$\frac{E((X-\mu)^4)}{\sigma^4}$ で尖度が定義されます。(歪度、尖度の式はゆがみ方、とがり方をあらわすことを知っていればよく、べつに計算できなくてもよい) また、$E((X-\mu)^r)$ とか $E(X^r)$ を r 次のモーメントという。

8.5.1 チェビシェフの不等式

連続型、離散型にかかわらず必ず次の不等式が成り立ちます。

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

つまり、期待値から離れれば離れるほど出づらくなるということです。

9 超幾何分布

9.1 二項分布

たとえば、5 枚硬貨を投げて、1 枚の硬貨の表が出る確率を 0.3 ならば、x 枚表が出る確率の確率分布は、$f(x) = {}_5C_x 0.3^x (1-0.3)^{5-x}$ 一般には、$f(x) = {}_nC_x p^x (1-p)^{n-x}$ こういう形の確率分布を二項分布といいます。二項分布では、期待値 E(X) と分散 V(X) は次のように表されます。

$$\begin{aligned} E(X) &= np \\ V(X) &= np(1-p) \end{aligned}$$

9.2 ポアソン分布

二項分布で n とか x とかが大きくなると、計算しずらくなっていきます。５０００乗とかするのは大変ですね。そのようなとき、次の定理が使えます。

$$n \longrightarrow \infty, p \longrightarrow 0, np \longrightarrow \lambda \text{のとき、}$$
$$f(x) = {}_nC_x p^x (1-p)^{n-x} \longrightarrow e^{-\lambda}\frac{\lambda^x}{x!}$$

このような確率分布を、ポアソン分布といいます。ポアソン分布では、期待値と分散は、

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x\,{}_nC_x p^x (1-p)^{n-x} = {}_nC_x p^x (1-p)^{n-x} \\ &= \cdots = np \\ V(X) &= np \end{aligned}$$

必ずこうなります。（分散は二項分布のときと形が違いますが、$p \longrightarrow 0$ の極限をとったものと考えればよいです）

9.3 幾何分布

表が出る確率 p のコインを投げる試行を考えます。x 回目で始めて表が出る確率 f(x) とすると、

$$f(x) = p(1-p)^{x-1}$$

（こんなの高校の数学ですね。）このような形で表される確率分布を、幾何分布といいます。

幾何分布における x の期待値 E(X) は、次の式で表されます。

$$E(X) = \frac{1}{p}$$

確率 p の事象は $\frac{1}{p}$ に 1 回おこることを考えると、十分に納得できるものなのではないでしょうか。ちなみに分散は $V(X) = \frac{q}{p^2}$

9.4 負の二項分布

幾何分布を一般化して、k 回成功するまでに x 回成功する確率を f(x) とすると、

$$f(x) = {}_{k+x-1}C_x p^k (1-p)^x$$

これも高校数学の範囲なので説明はしません。

一応、失敗回数の期待値は $E(X) = \frac{k(1-p)}{p}$、分散は $\frac{kq}{p^2}$ です。

幾何分布で、1 回成功するまでに失敗する回数 (成功するまでに投げる回数− 1) の期待値は、$E(X-1) = E(X) - 1 = \frac{1-p}{p}$ です。負の二項分布の期待値はこの k 倍と考えると良いでしょう。

9.5 超幾何分布

まず、次の問題を考えてみます。

8 人の男と 12 人の女からなるグループから 5 人の委員を選び出すとき、男が 2 人選ばれる確率を求めよ。

第 2 回のレポートからそのまま持ってきたわけですが、

すべての選び方が ${}_{20}C_5$ 通り、

8 人の男の中から男の委員を 2 人選ぶ選び方は、${}_8C_2$ 通り、

12 人の女の中から女の委員を 3 人選ぶ選ぶ方は、${}_{12}C_3$ 通り。よって求める確率 P は、

$$P = \frac{{}_8C_2 \times {}_{12}C_2}{{}_{20}C_5} = \frac{385}{969}$$

これを理解するのにあまり骨は折れなかったはずです。次はこれを一般化してみましょう。

つまり、男 M 人が入っている N 人の集団から n 人の委員を選ぶとき、男が x 人選ばれる確率を求めてみます。

すべての選び方は ${}_NC_n$ 通り、

男 M 人の中から x 人の男の委員を選ぶ選び方は ${}_MC_x$ 通り、

女 N-M 人の中から n − x 人の女の委員を選ぶ選び方は、${}_{N-M}C_{n-x}$ 通り。よって求める確率 f(x) は、

$$f(x) = \frac{{}_MC_X \times {}_{N-M}C_{n-x}}{{}_NC_n}$$

この確率分布を、超幾何分布といいます。

9.5.1 捕獲再捕獲法

突然ですが、湖にいる魚は何匹だ？と聞かれたら、どうやって調べるのがよいでしょうか。

魚は当然動くわけですから、マトモに数えるとしたら、全体を調べきったか、とか重複がないかを調べるのがかなり大変な作業になることでしょう。

そこで、統計学的に魚の数を数えてみることにします。そのやり方は次の通り。

M 匹の魚を釣る。

その M 匹の魚に、目印となるタグをつける。（この時点で、魚はタグのついたものと、ついていないものの 2 種類に分かれています）

放流。

n 匹の魚を、また釣る。釣られた魚もまた、タグのついたものと、ついていないものの 2 種類に分かれています。

ところで、さっきの委員の話に戻ると、数学的に、f(x) が最大となるのは、x:(n-x)=M:(N-M) であることを証明できます。

タグのついた魚を男の委員に、ついていない魚を女の委員に置き換えてみると、これと同じ議論ができます。

つまり、(2 回目に釣った魚のうちタグのついた魚):(2 回目に釣った釣った魚のうちタグのついていない魚)=(1 回目に釣ってタグをつけた魚):(1 回目に釣られていない魚) となる確率が最大となるのです。よってこの式が成り立っているものとしましょう。

最初の三つの変数は既知の変数で、わかっていないのは、1 回目に釣られていない魚の数だけです。つまり、上の式から 1 回目に釣られていない魚の数が求まります。これと、1 回目に釣った魚の数を足すことで、湖全体にいる魚の数がわかるのです。

以上のようなやり方を、捕獲再捕獲法といいます。

9.6 もう一回超幾何分布を

いったん、超幾何分布に戻ります。超幾何分布の期待値、分散は、次のように表されます。

$$\begin{aligned} E(X) &= n\frac{M}{N} \\ V(X) &= n\frac{M}{N} \times \frac{N-M}{M} \times \frac{N-n}{N-1} \end{aligned}$$

期待値はわかりやすいのではないかと思います。n 人委員を選ぶとき、全体に対する男の人数の割合 ($\frac{M}{N}$。最初の例だと $\frac{8}{20}$。二項分布だと p に相当) に n をかけた数だけ、男の委員が選ばれることを示唆しています。

分散の場合、二項分布の式に $p = \frac{M}{N}$ を突っ込むと、$np(1-p) = n\frac{M}{N} \times \frac{M-N}{N}$ となり、最後の $\frac{N-n}{N-1}$ は不要なはずです。これは、いったい、何だ？

実は、これは、たとえば魚を釣ったとき、湖にいる魚の (タグがついているかついていないかという) 割合が変わってしまうことから必要な項なのです。(有限母集団修正というらしいですが名前はどうでもいい)

N が無限と考えられるときにはこれは 1 に収束するので無視することができます。

過去問で、このことについて出題されたこともあります。ちょっと見てみましょう。

> ある沼沢地に生息する動物は 1000 匹程度と推定され、そのうち 350 匹がある種類 (A 種) とされている。今その沼沢地で 100 匹を捕獲して調査するとき、それらのうちの A 種の数の期待値、分散標準偏差を求めなさい。1000 を無限に大きいと考えるときと、有限の数と考えるときの両方を求めなさい。

珍しく代入するだけで答が出る問題なので, 解説はしませんが、(標準偏差は分散の平方根を取ると出ます。念のため) 上の考え方によって解けることがわかると思います。

10 連続型の確率分布

10.1 正規分布

代表的な連続型の確率分布です。正規分布の密度関数 f(x) は、

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{X-\mu}{2\sigma^2}} \qquad (\infty \leq x \leq \infty)$$

図 (TeX の機嫌が悪いらしく図が変なところにありますが) のように、一番高いところが期待値 $E(X) = \mu$ で、これと変曲点との距離が標準偏差 σ です。分散は $V(x) = \sigma^2$。

10.2 確率分布の記号による表記

二項分布 $\cdots Bi(n,p)$

ポアソン分布 $\cdots Po(\lambda)$
正規分布 $\cdots N(\mu, \sigma^2)$

11　最後に

以上が先週までの授業でやったところと、その補足になります。