健康統計学 第11回

今回は、前回に引き続き、確率分布（テキスト53～63ページ）について学習します。とくに、正規分布（標準正規分布）、カイ二乗分布、t分布、F分布について説明します。

また、母集団統計値の推定（テキスト65～76ページ）についても学習します。

テキスト

- 『やさしい保健統計学 改訂第4版』縣 俊彦著 (南江堂)

今回の内容

1. 二項分布と関係する分布
2. 正規分布と関係する分布
    a. 補足：標準化を用いた確率の計算
3. 母集団と標本
4. 点推定と区間推定

二項分布と関連する分布

二項分布 (binomial distribution)

二項分布を考える準備

- ある独立した試行を何回か連続して繰り返す場合を考える
- 例:1個のサイコロを5回連続して投げたときに、1回は1の目が出る確率について
  - 1の目がでたかどうかを確率変数 X とする
  - でた場合の事象を $\{X=1\}$ 、でない場合の事象を $\{X=0\}$ とすると、それらの確率は次のようになる

$$\mathrm{P}(X=1)=\frac{1}{6},\ \mathrm{P}(X=0)=\frac{5}{6}$$

  - また、5回サイコロを投げたとき1の目が出る組み合わせは、5通りになる

$${}_5C_1=\frac{5!}{1!\times 4!}=5$$

  - このうち1回だけ1の目が出る確率は、次のようになる

$$\left(\frac{1}{6}\right)^1\times\left(\frac{5}{6}\right)^4$$

  - したがって、サイコロを5回連続して投げたときに1回は1の目が出る確率は、次のとおりになる

$${}_5C_1\times\left(\frac{1}{6}\right)^1\times\left(\frac{5}{6}\right)^4\simeq 0.4019$$

- あらためて、サイコロを5回連続で投げて1の目がでる回数を、確率変数 X とすると、確率は次のようにして求められる

$$\mathrm{P}(X=0)={}_5C_0\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5\simeq 0.4019$$

$$\mathrm{P}(X=1)={}_5C_1\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4\simeq 0.4019$$

$$\mathrm{P}(X=2)={}_5C_2\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3\simeq 0.1608$$

$$\mathrm{P}(X=3)={}_5C_3\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2\simeq 0.0322$$

$$\mathrm{P}(X=4)={}_5C_4\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^1\simeq 0.0032$$

$$\mathrm{P}(X=5)={}_5C_5\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^0\simeq 0.0001$$

確率分布は次のようになる

X	0	1	2	3	4	5
確率	0.4019	0.4019	0.1608	0.0322	0.0032	0.0001

二項分布

- ある独立な試行で、事象 A が起こる確率を p 、起こらない確率を $q(=1-p)$ とする
- この試行を独立に n 回繰り返したときに、事象 A が起こる回数を確率変数 X としたとき、 $X=x$ (つまり x 回起こる)となる確率は次のようになる

$$
\begin{aligned}
\mathrm{P}(X=x) &= {}_nC_x p^x q^{n-x} \\
&= \frac{n!}{x!(n-x)!} p^x q^{n-x} \\
&= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}
\end{aligned}
$$

- このような確率分布を、「**二項分布**」といい、$B(n,p)$ と表す
  - 1回の試行で生じる事象が2種類しかない場合（不良品の発生など）の分布などに使われる

二項分布の特徴

二項分布 $B(n,p)$ には、次のような特徴がある。

- 確率が $p=0.5$ のとき、平均を中心とした左右対称な分布になる
  - p が小さいと、左側に寄った非対称な分布になる
  - p が大きいと、右側に寄った非対称な分布になる
- 回数 n が非常に大きくなると、左右対称な分布になる（ラプラスの定理）
- 平均は np 、分散は $npq=np(1-p)$ になる
- $np \geq 5$ で正規分布に近似できる

ポアソン分布 (poisson distribution)

ポアソン分布とは

- 二項分布の平均 np が一定の値 λ をとる場合を考えると、二項分布の式は次のように変形できる

$$
\mathrm{P}(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}
$$

  - e は自然対数の底で、$e=2.71828\cdots$
- このような分布を「**ポアソン分布**」といい、平均で分布が決まる
  - 極めてまれにしか起こらない現象を一定期間観測した場合の分析に使われる（交通事故の死亡者数、病気の死亡者数、電話の呼び出し数など）

$$
\mathrm{P}(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}
$$

  - x : 発生回数
  - λ : 平均の発生回数
  - つまり、二項分布において試行回数 n が極めて大きい場合（ $n \to \infty$ ）に用いる

ポアソン分布の特徴

- 平均 λ が大きくなると、左右対称な分布になる
- 平均も分散も、λ になる

正規分布と関連する分布

自然界や一般に観察できる多くのものについて、その分布は、平均値を中心に左右対称の**釣り鐘状**の分布になっていることがあります。

- 生物現象、毎年の雨量など
- 身長や体重、標準的なテストの成績など

正規分布 (normal distribution)

正規分布とは

- 平均値を中心とした左右対称の釣り鐘状になる分布を、「**正規分布**」という

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

  - π は円周率で、$\pi=3.14159\cdots$
  - e は自然対数の底で、$e=2.71828\cdots$
  - μ : 平均値
  - σ : 標準偏差
  - σ^2 : 分散
- 確率密度関数が上の式になるとき、『x は、平均が μ で標準偏差が σ の正規分布にしたがう』といい、$N(\mu,\sigma^2)$ と表す
  - つまり、平均値と分散 (標準偏差) から分布が決まる
- さまざまな推定や検定に使われる

正規分布の特徴

- 分布の中心は平均値 μ で、最も高い値 (極大値) をとる
  - 平均値が変化すると、分布が左右に移動する
- $\mu\pm\sigma$ で変曲点 (曲線の凹凸の変わり目) になる
  - 標準偏差が変化すると、分布の高さや広がりが変化する
- 平均値、中央値、最頻値は一致する
- 平均値と標準偏差から、分布の割合 (曲線とx軸に囲まれる面積) が決まる
  - $\mu\pm\sigma$ の範囲 : 全体の約 68.26% を含む
  - $\mu\pm2\sigma$ の範囲 : 全体の約 95.44% を含む
  - $\mu\pm3\sigma$ の範囲 : 全体の約 99.73% を含む
  - それ以外の範囲 : 全体の約 0.27% を含む

標準正規分布

標準正規分布とは

- 正規分布で、平均値が 0 、標準偏差が 1 になるように、正規分布の確率密度関数の変数 x を次のように変換する (変数変換)

$$z=\frac{x-\mu}{\sigma}$$

- すると、標準偏差の式は次のように変形される

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{\frac{z^2}{2}}$$

- 確率密度関数が上の式になるとき、『z は、標準正規分布にしたがう』という
  - 『z は、平均が 0 で標準偏差が 1 の正規分布にしたがう』ことになり、$N(0,1^2)$ と表すことができる
  - つまり、変数だけから分布が決まる
  - さまざまな推定や検定に使われる

標準化（基準化）(standardization)

- 先ほどの変数変換を、「**標準化**」または「基準化」という

$$z=\frac{x-\mu}{\sigma}$$

- (感覚としては) 分布を平均値の分だけ 0 まで移動し、分布の広がりを $\frac{1}{\sigma}$ にする
  - 単位の異なるデータや平均値･分散が異なるデータを比較するときに使う (英語と数学のテストの成績の比較)
- ちなみに**偏差値**は、標準化を応用したもので、次のような式になる

$$z=50+10\times\frac{x-\mu}{\sigma}$$

標準正規分布の特徴

- 基本的な特徴は、正規分布と同じ
- 分布の曲線とx軸で囲まれた全体の面積は1になり
  - すなわち、面積が確率 (割合) を表すことになる

カイ二乗分布

カイ二乗分布とは

- 変数 $z_1,z_2,\cdots,z_r$ が標準正規分布にしたがい、互いに独立であるとする
- 次のようになるとき、『z は、自由度 r のカイ二乗 (X^2) 分布にしたがう』という

$$\begin{aligned}X^2&=z_1{}^2+z_2{}^2+\cdots+z_r{}^2\\&=\sum_{i=1}^{r}z_i{}^2\\&=\sum_{i=1}^{r}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2\end{aligned}$$

  - 自由度とは分布の形状に影響を及ぼす値で、自由度の値が変わると分布の形状も変化する

カイ二乗分布の特徴

- 平均は r 、分散は $2r$ になる
- 適合度や独立性の検定など、分散の推定や検定に利用される
- 自由度が大きくなると、対称な分布に近づく

t分布

t分布とは

- 標準正規分布にしたがう確率変数 x と、自由度 r のカイ二乗分布 ${\chi_r}^2$ にしたがう確率変数 y があり、互いに独立であるとする
- 次のようになるとき、『x は、自由度 r のt分布にしたがう』という

$$t = \frac{x}{\sqrt{y/r}} = \frac{x}{\sqrt{\frac{{\chi_r}^2}{r}}}$$

- (感覚としては) 正規分布にしたがう統計量を標準化したものの分布を表す
- この分布を発表した William Gosset のペンネームから「スチューデント (Student) のt分布」とも呼ぶ

t分布の特徴

- 平均は 0 、分散は $\frac{r}{r-2}$ になる
- 自由度が大きくなると、標準正規分布に近づく
    - 自由度が小さいときは、正規分布よりも裾の長い分布になる
- 母平均の推定や検定、平均値の差の検定や検定など、多くの統計的推定で利用される

F分布

F分布とは

- 確率変数 x と確率変数 y が、それぞれ自由度 m と n のカイ二乗分布にしたがい、互いに独立であるとする
- 次のようになるとき、『x は、第一自由度が m で第二自由度が n のF分布にしたがう』という

$$F = \frac{x}{m} \Big/ \frac{y}{n} = \frac{{\chi_m}^2}{m} \Big/ \frac{{\chi_n}^2}{n}$$

- (感覚としては) 分散の比率についての分布を表す

F分布の特徴

- 2つの自由度を持つ
- 分散検定や分散分析 (分散比の分布を調べる) に利用される
- 平均は $\frac{n}{n-2}$ 、分散は $\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ になる

標準化を用いた確率の計算

標準得点（標準正規分布の復習）

平均が μ 、分散が σ^2 の**正規分布**から、**標準正規分布**を導くときに、次の式を用いて**標準化**を行います。

$$z = \frac{x-\mu}{\sigma}$$

このときの z を、**標準得点**（standardized score）といいます。

標準得点は、平均が0、分散が1の標準正規分布 $N(0,1)$ にしたがいます。

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}}$$

標準得点を使った確率の計算

正規分布とみなされるデータを標準化すれば、標準正規分布表を用いて、確率を計算することができます。

例題

> 高校3年生のAさんの身長は175cmである。Aさんが入学する、B大学の学生の身長について、平均は182cmで、標準偏差は8.3cmである。このとき、B大学の学生がAさんより身長が高い確率を求める。

確率の求め方

1. まず、B大学の学生の身長を x として、標準得点を計算する。

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ &= \frac{x-182}{8.3} \end{aligned}$$

2. 標準得点 z は標準正規分布にしたがうので、標準正規分布表を用いて、確率を求める。
「身長が175cmより大きい」ということは、標準得点が次のようになるということである。

$$\begin{aligned} z &> \frac{175-182}{8.3} \\ &> -0.843\cdots \\ &\simeq -0.84 \end{aligned}$$

3. したがって、「身長が175cmより大きい」確率 $P(z>-0.84)$ は、標準正規分布表から $z=-0.84$ の値を求めればよい。

$$\begin{aligned} P(x>175) &\simeq P(z>-0.84) \\ &= 1-P(z\leq 0.84) \\ &\simeq 1-0.2005 \\ &\simeq 0.80 \end{aligned}$$

つまり、「B大学の学生がAさんより身長が高い確率」は約80%（0.8）となる。

母集団と標本

母集団と標本

全数調査と標本調査

- 全数調査、または悉皆（しっかい）調査
  - 全体について調べる（例：国勢調査）
- 標本調査
  - 全体から選び出した一部分について調べる（例：選挙予測）
  - 時間的・経済的に現実的な調査

母集団と標本

- 母集団（population）
  - 標本調査のもとになる集団
  - 母集団の大きさを、 N であらわす
- 標本（sample）
  - 母集団から抽出（サンプリング）された一部分
  - 標本の大きさ（サンプルサイズ：標本に含まれる個数）を、 n であらわす

標本の抽出

- 推測統計学（inductive statistics）
  - 標本にもとづいて母集団の特性値（母数：平均、分散などのパラメータ）を推定・予測する
  - 2つ以上の母集団の特性値を比較・検討する
- 標本が「母集団の精巧なミニチュア」になるように、偏りなく標本を抽出しないといけない

無作為抽出法（random sampling）

- 母集団から無作為に標本を抽出する
- 乱数表（無規則に数字を羅列した表）等をもとに、データ（個体）を標本として選ぶ
- どの標本も全く等しい確率で選ばれる（主観や作為などが入る余地がない）
- 母集団が大きい場合は実施が困難（例えば数万人の名簿を作るのは容易ではない）

系統抽出法（systematic sampling）

- 最初の標本のみ乱数表などで選ぶ
- あとの標本は一定間隔（抽出間隔；sampling interval）で抽出する
- 実際には抽出間隔が扱いやすい数でない場合が多い
- 無作為抽出法と同じで、母集団が大きい場合は実施が困難

多段抽出法（multi-stage sampling）

- 何段階かのステップを経て標本を抽出する（例：2段階抽出法）
- 何段階かのステップで標本にする対象を絞り込み、最後に無作為抽出法で標本を選ぶ
  - 4段階の例：全国→都道府県→市町村→病院・診療所→看護師（あとは無作為抽出）
- 調査の手間ひまが少なく実践的だが、標本の偏り（バイアス）に注意しないといけない

層別抽出法（stratified sampling）

- 母集団を同じ特徴（年代、職業、都道府県など）で層に分ける（層別化）

- 層の構成比に応じて適切な数だけ、層ごとに無作為抽出法で標本を選ぶ
- 層の違いによって偏りがある状況が事前に予測される場合に適している（地域と政党の支持率、年代と好きな芸能人）
    - 母集団の特徴を少なくとも1つは事前に知っている必要がある

点推定と区間推定

点推定（point estimation）

点推定とは

- 標本の特性値1つから母集団の特性値を推定する
  - 母数（パラメータ）：母集団の特性値（統計量）
    - 母平均（母集団の平均）、母分散（母集団の分散）、母比率（母集団の比率）
    - 直接調べることはできない
  - 推定量：標本をもとに母数として推定した統計量
    - 標本平均（標本の平均）、不偏分散（標本の分散）、標本比率（標本の比率）
    - 推定値：推定量から求めた具体的な値
- 点推定では、標本の推定値が必ずしも母数と一致するとは限らない（たいてい誤差が生じる）

推定量の望ましい性質

- 不偏性
  - その期待値（平均値）が母数と一致する推定量を「不偏推定量」という
    - 標本平均について、その期待値は母平均に一致する（中心極限定理へ）
- 一致性
  - 標本数を大きくしていくと推定値が母数に近づく推定量を「一致推定量」という
    - 大数の法則から明らか（統計的確率は数学的確率に近づく）
- 有効性
  - ある母数に対して2つ以上の推定量がある場合に分散（誤差）の小さい推定量を「有効推定量」という
    - 標本の中央値に比べ、標本平均の方が、分布の分散が小さい

区間推定（interval estimation）

区間推定とは

- 標本の推定量から、「ある確率」で、母集団の特性値（母数）の範囲を示す

信頼区間と信頼係数

- 信頼区間
  - 「ある」確からしさで示される、母集団の特性値の範囲
  - 95%信頼区間
    - 標本から平均値を出したとき、母平均（母集団の平均）がその区間にあるのが**100回中95回以上の確率**で、間違える危険性が5回未満
  - 99%信頼区間
    - 標本から平均値を出したとき、母平均（母集団の平均）がその区間にあるのが**100回中99回以上の確率**で、間違える危険性が1回未満
    - 注意：「母平均の値が95（または99）%の確率でその区間のどこかにある」と解釈してはいけない
  - 信頼限界：信頼区間の上限および下限の値

- 信頼係数
  - 区間推定の確実性をあらわし、「1-α」（αは0.05または0.01）であらわす
  - 信頼度：100×(1-α)％

区間推定の考え方

1. 標本の推定量（標本平均、標本比率など）の分布を選択する
   - （標準）正規分布、t分布、F分布、カイ2乗分布
2. 推定量の分布の平均と分散（標準偏差）を求める
3. 母数にあった方法で、信頼区間を算出する
   - 一般的には次のようになる
     - 信頼下限：<推定量>－<分布における確率αの値>×<推定量の標準偏差（標準誤差）>
     - 信頼上限：<推定量>＋<分布における確率αの値>×<推定量の標準偏差（標準誤差）>

中心極限定理（central limit theorem）

- 平均が μ 、分散が σ^2 の母集団から大きさ n の標本を抽出して、その標本平均 $\overline{x}$ を調べると、その分布は平均が μ 、分散が $\frac{\sigma^2}{n}$ の正規分布に従う（n は十分大きな数とする）
  - つまり、「標本平均の平均（期待値）は、母集団の平均に一致する」
  - また、「母集団の分布に関係なく、標本平均の平均は正規分布にしたがう」
  - 母集団が正規分布にしたがうなら、標本の大きさにかかわらず、中心極限定理が成り立つ
  - 参考URL:http://www.kwansei.ac.jp/hs/z90010/sugakuc/toukei/tyuusin2/chuusin.htm