健康統計学 第14回

今回は、母集団統計値の推定（テキスト65～76ページ）について学習します。また、仮説検定の考え方（77～87ページ）についても説明します。

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

母平均の区間推定

母集団から抽出した標本をもとに母集団の平均（母平均）を区間推定する

- ○大卒100人の初任給のデータからすべての大卒の初任給の平均を推定
- ○あるクラスの男子の身長のデータから日本全体の同年代の男子の身長の平均を推定

母平均の区間推定の準備

標準得点

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**を行う。

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

このときの z を、**標準得点**（standardized score）という。標準得点は、平均が0で分散が1の標準正規分布 $N(0,1)$ にしたがう。

中心極限定理と標本平均の分布

中心極限定理では、「平均が μ で分散が σ^2 の母集団について、母集団の分布が正規分布でなくても、標本の大きさ n が十分大きい標本を抽出すれば、標本平均 $\bar{x}$ の分布は平均が μ で分散が $\frac{\sigma^2}{n}$ の正規分布にしたがう」ことが成り立つ。

ある標本の標本平均 $\bar{x}_i$ を標準化した分布を考える。標本平均を標準得点 z_i に変換すると、次の式になる。

$$z_i=\frac{\bar{x}_i-\mu}{\sqrt{\frac{\sigma^2}{n}}}$$

標本平均の分布と信頼区間

標準正規分布にしたがう標本平均の信頼区間について考える。

95%信頼区間は、標本平均を標準化した z が、-1.96～1.96の区間を示す。つまり、信頼度を95%とした95%信頼区間では、標本平均の存在する範囲は次の式のようになる。

$$-1.96<\frac{\bar{x}_i-\mu}{\sqrt{\frac{\sigma^2}{n}}}<1.96$$

ここで、信頼度を 100(1-α)% とすると、次のように書き換えることができる。

$$-z_{(\alpha/2)}<\frac{\bar{x}_i-\mu}{\sqrt{\frac{\sigma^2}{n}}}<z_{(\alpha/2)}$$

区間推定では、調べたいのは母平均 μ の範囲になるので、上の式を μ について解くと、次の式が得られる。これは「母平均が標本平均±z値×標準誤差の範囲にある」ことを示している。

$$\bar{x}-z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}}$$

母分散が既知の場合

- 分散が σ^2 母集団から抽出した大きさ n の標本の平均 (標本平均) が $\bar{x}$ であるとき
- 母平均 (母集団の平均) μ の信頼度 100(1-α)% の信頼区間は次のとおり

$$\bar{x}-z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}}$$

  - なお z は次のように標準化した統計量で、標準正規分布にしたがう

$$z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

  - 推定量 (この場合は標本平均) の分散の平方根を**標準誤差** (SE : Standard Error) といい、次のように表す

$$SE=\sqrt{\frac{\sigma^2}{n}}=\frac{\sigma}{\sqrt{n}}$$

- 標本平均 $\bar{x}$ の分布は正規分布にしたがい (中心極限定理より) 、平均は μ 、分散は $\frac{\sigma^2}{n}$ となる
- 標本数が多い場合にも使う

母分散が未知の場合（t推定）

- 母標準偏差 σ のかわりに、標本標準偏差 s を用いる
  - 分散は不偏分散になる

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

- 母集団から抽出した大きさ n の標本の平均 (標本平均) が $\bar{x}$ 、分散 (不偏分散) s^2 がであるとき
- 母平均 μ の信頼度 100(1-α)% の信頼区間は次のとおり

$$\bar{x}-t_{(\alpha/2)}(n-1)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x}+t_{(\alpha/2)}(n-1)\frac{s}{\sqrt{n}}$$

  - t は次のように標準化した統計量で、t分布にしたがい、平均は μ 、分散は $\frac{s^2}{n}$ となる

$$t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

  - 標準誤差の推定値は $\sqrt{\frac{s^2}{n}}$ となる
  - $t_{(\alpha/2)}(n-1)$ は、自由度 n-1、確率 α/2 のtの値
- 標本数が少ない場合にも用いる
  - 自由度 (すなわち標本数) が増えれば、t分布が標準正規分布 N(0, 1) に近づくので、母分散が既知の場合と同じになる

母比率の推定

母集団から抽出した標本をもとに母集団の比率（母比率）を区間推定する

- 選挙前の候補者の支持率 (支持する／支持しない) を推定できる
- 好き嫌いのようなアンケート調査から全体の傾向を推定できる

正規分布による近似（標本数の多い場合）

- 二項分布 (ある事象が起こるか起こらないかの確率の分布) は、試行回数 n が十分大きい場合、正規分布に近似できることを利用
- 母集団のある事象について、 n 回の試行 (標本の大きさが n) の標本の比率 (標本比率) を $\overline{p}$ とするとき
- 母比率 p の信頼度 100(1-α)% の信頼区間は次のとおり

$$\overline{p}-z_{(\alpha/2)}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \le p \le \overline{p}+z_{(\alpha/2)}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

  - なお、標本比率を $\overline{p}=\frac{X}{n}$ とすると、その平均 $E(\overline{p})$ と分散 $V(\overline{p})$ は、次のようになる

$$\begin{aligned} E(\overline{p}) &= E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)=\frac{1}{n}np \\ &= p \\ V(\overline{p}) &= V\left(\frac{X}{n}\right)=\frac{1}{n^2}V(X)=\frac{1}{n^2}np(1-p) \\ &= \frac{p(1-p)}{n} \end{aligned}$$

正規分布による近似（標本数の少ない場合）

- 大きさが N 母集団のある事象について、大きさが n) の標本の比率 (標本比率) を $\overline{p}$ とするとき
- 母比率 p の信頼度 100(1-α)% の信頼区間は次のとおり

$$\overline{p}-z_{(\alpha/2)}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\sqrt{\frac{N-n}{N-1}} \le p \le \overline{p}+z_{(\alpha/2)}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\sqrt{\frac{N-n}{N-1}}$$

- 最初の式 (標本数が多い場合の式) より近似の精度が良い (母比率に近い値になる)

F分布から算出（標本数の少ない場合）

- 母集団のある事象について、 n 回の試行 (標本の大きさが n) の標本の比率 (標本比率) を $\frac{x}{n}$ とするとき
- 母比率 p の信頼度 100(1-α)% の信頼限度は次のとおり
  - 信頼上限：

$$\frac{m_1F_U}{m_1F_U+m_2}, \quad m_1=2(x+1), \quad m_2=2(n-x)$$

    - 第1自由度 m_1 、第2自由度 m_2 に対応するF分布の値を F_U とする
  - 信頼下限：

$$\frac{n_2 F_L}{n_1 F_L + n_2}, \quad n_1 = 2(n-x+1), \quad n_2 = 2x$$

- **第1自由度** n_1 **、第2自由度** n_2 **に対応するF分布の値を** F_L **とする**

標本の大きさの決め方

区間推定の考え方を使えば、95%または99%信頼区間において**ある決まった**標準誤差（標準偏差）になるような、標準サイズを計算して求めることができます。

母平均の区間推定から標本の大きさを決める

母平均の区間推定の場合で、標本の大きさの決め方を考えます。

標本の大きさの求め方の考え

母分散が既知の場合は、正規分布を利用して、母平均 μ を推定できます。信頼度 100(1-α)% の信頼区間は次のとおりです。

$$\bar{x}-z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}}$$

ここで、母平均と標本平均との誤差の限度を E とします。

$$E=z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}}$$

この式を、標本サイズ n について解くと、次のようになります。

$$n=\left(\frac{z_{(\alpha/2)}\sigma}{E}\right)^2$$

これで、標本平均が $\bar{x}\pm E$ になる、標本サイズを n を求めることができます。

例題

ある県の大学生全体の1か月の平均の生活費について調査して推定するとき、全体の標準偏差が9,000円であれば、、誤差を1,000円以内にして95%信頼区間で推定するためには、何人に調査すればよいか。

1. 標本の大きさを n とする。
2. 誤差の限度 $E=1000$ とすると、標本の大きさ n は、次のようになると考えられる。

$$\begin{aligned} n &= \left(\frac{z_{(\alpha/2)}\sigma}{E}\right)^2 \\ &= \left(\frac{1.96\times9000}{1000}\right)^2 \\ &= (17.64)^2 \\ &= 311.1696 \end{aligned}$$

つまり、**312人**の大学生に調査をすればよいことになります。

母比率の区間推定から標本の大きさを決める

母比率の区間推定の場合で、標本の大きさの決め方を考えます。

標本の大きさの求め方の考え

標本の大きさを n が十分大きい場合、正規分布による近似で母比率 p を推定できます。信頼度 100(1-α)% の信頼区間は次のとおりです。

$$p \le \overline{p} \pm z_{(\alpha/2)} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

ここで、母比率と標本比率との誤差の限度を D とします。

$$D = z_{(\alpha/2)} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

この式を、標本サイズ n について解くと、次のようになります。

$$n = \overline{p}(1-\overline{p}) \left(\frac{z_{(\alpha/2)}}{D} \right)^2$$

これで、標本比率が $\overline{p} \pm D$ になる、標本サイズを n を求めることができます。

例題

> 2010年6月29日に生中継された、サッカーのワールドカップの南アフリカ大会決勝トーナメント1回戦、日本対パラグアイ戦の平均視聴率は関東地区57.3%だった。この視聴率が、95%信頼区間において「57.3 ± 2%」となるような、標本サイズを求めよ。

1. 標本の大きさを n として、ある程度大きいと仮定する。
2. 標本数が大きい場合、正規分布による近似で母比率を推定できることを利用すると、標本サイズ n は、次のようになると考えられる。

$$\begin{aligned} n &= \overline{p}(1-\overline{p}) \left(\frac{z_{(\alpha/2)}}{D} \right)^2 \\ &= 0.573(1-0.573) \left(\frac{1.96}{0.02} \right)^2 \\ &= 2349.82 \end{aligned}$$

つまり、「57.3 ± 2%」で視聴率を調査するには、**2350台**以上のテレビを調査する必要があります。

ちなみに、某視聴率調査会社の場合、関東地区で視聴率を調査する機械を取り付けてあるテレビの台数は、600台といわれています。

母相関係数の推定

母相関係数の推定の手順

おおまかに、次のような手順で母相関係数の推定を行う。

1. 母相関係数の有意性の検定（無相関の検定）
2. 母相関係数の推定

母相関係数の有意性の検定

- 母集団において無相関かどうか（母相関係数 $\rho=0$ かどうか）を調べる
  - 標本において相関があっても、母集団では相関が（ほとんど）ない場合がある

次のような手順でチェックをする。

1. 仮説を立てる
   - 「母相関係数は0である」という仮説を考える（帰無仮説という）
2. 有意水準を設定する
   - 仮説が成り立たない（棄却するという）確率を有意水準という
   - よくα=0.05かα=0.01が使われる
3. tの値を算出する

$$t_0=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

   - 標本相関係数を r 、標本数を n とする
4. t分布表から有意水準に対応するtの値（自由度 n-2 ）をもとめる
   - t分布表から、確率 1-α、自由度 n-2 のtの値を算出する
5. 2つのtの値をもとに判定する
   - $t_0 \geq t_{(\alpha/2)}(n-2)$ の場合
     - 帰無仮説を棄却する、すなわち、 $\rho \neq 0$ で母相関係数は0ではない（続けて、区間を推定する）
   - $t_0 < t_{(\alpha/2)}(n-2)$ の場合
     - 帰無仮説を棄却しない、すなわち、 $\rho=0$ で母相関係数は無相関（ここで終わり）

母相関係数の推定

次のような手順で推定する。

1. 標本相関係数 r をz変換する
   - r を正規分布で近似させるために、フィッシャー(Fisher)のz変換で変換する

$$z_r=\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

   - $\ln$ は自然対数で $\log_e$ をあらわす
   - z は標準偏差 $s_z=\frac{1}{\sqrt{n-3}}$ の正規分布で近似される

2. 母相関係数をz変換した、z_ρ の信頼限界を算出する

- 信頼上限：

$$z_U = z_r + z_{(\alpha/2)}\frac{1}{\sqrt{n-3}}$$

- 信頼下限：

$$z_L = z_r - z_{(\alpha/2)}\frac{1}{\sqrt{n-3}}$$

3. z_U と z_L を r に逆変換して、ρ の信頼限界を求める

- 信頼上限：

$$\rho_U = \frac{e^{2z_U}-1}{e^{2z_U}+1}$$

- 信頼下限：

$$\rho_L = \frac{e^{2z_L}-1}{e^{2z_L}+1}$$

仮説検定の考え方

仮説検定の考え方について、簡単な例を用いて考えてみましょう。

コインを投げて表が出た回数を数える

例として、コインを10回投げて、表が出た回数を数えることを考えてみましょう。

ゆがみがない、いわゆる「かたよりのない」コインであれば、コインを1回投げた結果、表が出る確率も裏が出る確率も、ちょうど半分の $\frac{1}{2}=0.5$ と考えて問題ないでしょう。つまり、10回投げた結果として表が出る回数は5回ぐらいが最も多いと考えられます。

そこで、「**あるコイン**」を10回投げたところ、表が9回も出たとします。この「あるコイン」は「かたよりがない」コインでしょうか？ それとも「かたよりがある」コインでしょうか？

「コインにはかたよりがない」という仮説を立てる

仮説検定では、母集団に対するある**仮説**を立てます。そして、母集団から取り出した一部分、つまり標本を使って、その結果が偶然のものなのか必然なのかを確率的に調べて、仮説が正しいかどうかを判断する方法です。

今、「コインを10回投げたうち9回表がでた」ことについて考えています。コインにかたよりがある可能性がありそうです。そのことを仮説検定で確かめてみましょう。

仮説検定では、どちらかというと主張したいことに反対の仮説をまず考えます。これが、「母集団に対するある仮説」になります。この仮説を「無に帰することを予定した」という意味で、「**帰無仮説**」といいます。そして、帰無仮説に対立する仮説、つまり、どちらかといえば主張したい仮説を「**対立仮説**」といいます。

もしコインにかたよりがあるとしても、どの程度かたよっているかまではわかりません。そこで、帰無仮説として「**コインにかたよりはない**」という仮説を立てることにします。

まとめると、帰無仮説と対立仮説は次のようになります。

- **帰無仮説：「コインにはかたよりがない」**
- **対立仮説：「コインにはかたよりがある」**

コインの表が出る回数の確率を求める

次に、帰無仮説として立てた仮説のもとでの確率を求めて、その確率をもとに、今考えている事象（コインを10回投げたら9回表が出た）が偶然起こったことか必然的に起こったことかを判断してみましょう。

コインのように表または裏の2種類の結果を考えるには、第9回で学習した、二項分布の考え方を利用します。二項分布は、ある独立な試行について事象 A が起こる確率を p 、起こらない確率を $q(=1-p)$ とすると、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ります。

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

今回は、10回のうち表がでる回数を確率変数 X として考えます。表が出る確率も表が出ない（裏が出る）確率も同じで $p=q=\frac{1}{2}$ となりますから、表が x 回でる確率は次のようになります。

$$\begin{aligned}\mathrm{P}(X=x) &= {}_{10}\mathrm{C}_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{10-x}\\ &= {}_{10}\mathrm{C}_x\left(\frac{1}{2}\right)^{10}\\ &= \frac{{}_{10}\mathrm{C}_x}{1024}\end{aligned}$$

この式を計算した結果を、確率分布として次のようにまとめておきます。

X	0	1	2	3	4	5	6	7	8	9	10
確率	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

仮説から求めた確率をもとに判断でする

今考えている仮説は「コインはかたよりがない」です。では、10回中9回表が出るコインにかたよりがないのかあるのかを判断するには、どうすればよいでしょう。

ここで、「コインにかたよりがないという仮説のもとで、まれなこと（ある一定の確率以下の出来事）が起きた場合は、そのコインはかたよりがないとは見なせない」としましょう。これが仮説検定では重要な考え方です。

この「まれなこと」が起きたと判断する基準を、**有意水準**といいます。有意水準はあらかじめ決めておきます。一般には5%（0.05）か1%（0.01）が使われます。今回は有意水準を5%としておきましょう。

> 有意水準は、「100回のうち5回以下しか起こらない」というどのくらい稀（まれ）なことが起こるかの判断基準となります。
> しかし、正しい仮説であっても、100回のうち5回以下しか起こらないことが起こってしまう場合も考えられます。つまり、「100回のうち5回未満は間違った判断をして、正しい仮説を捨ててしまう」可能性があることになります。このような誤った判断をする危険があるため、有意水準を「**危険率**」とも呼びます。そして、「正しい帰無仮説を捨ててしまい、対立仮説を採択してしまう」ことを**第一種の過誤**といいます。

表が9回はでる確率は、「表が9回でた」場合と「表が10回でた」場合の確率を足し合わせたものになります。少なくとも9回は出た、と考えます。確率分布の表から、表が9回は出る確率は、0.010+0.001=0.011となります。有意水準を5%(0.05)と考えると、それより小さい確率です。

有意水準より小さい確率で起きてしまったことを、仮説検定では「**仮説では起こるはずのないことが起こった**」と見なします。このことを「**帰無仮説を棄却する**」といいます。もし、有意水準より大きい確率だった場合は、「仮説で起こるはずのないことが起こらなかった」とみなして、「帰無仮説を棄却できなかった」といいます。今回の場合、表が9回でるのはめったに起こらないことが起こったので、帰無仮説を棄却し、「**コインにはかたよりがある**」という判断になります。

ちなみに、表が8回でたコインが別にあったとしましょう。そのコインにかたよりがあるかどうかを考えると、8回以上表が出る確率は、確率分布から求めると、0.044+0.010+0.001=0.055となり、有意水準5%を超えることになります。つまり「コインにかたよりがない**とはいえない**」という判断になります。