健康統計学 第11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。

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. 母集団と標本（復習）
2. 点推定と区間推定
3. 母平均の推定
4. 母比率の推定
    a. 補足：標本の大きさの決め方
5. 母相関係数の推定
6. 確率分布に関するExcelの関数

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

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 $100(1-\alpha)\%$ 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

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

- 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

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

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

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

- 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。

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

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 $\overline{x}\pm E$ 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。

$$\begin{aligned} n &= \left(\frac{z_{(\alpha/2)}\sigma}{E}\right)^2 \\ &= \left(\frac{1.96\times 9000}{1000}\right)^2 \\ &= (17.64)^2 \\ &= 311.1696 \end{aligned}$$

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。

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

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。

$$\begin{aligned} n &= \overline{p}(1-\overline{p}) \left(\frac{z_{(\alpha/2)}}{D} \right)^2 \\ &= 57.3(1-57.3) \left(\frac{1.96}{0.02} \right)^2 \\ &= 2349.82 \end{aligned}$$

つまり、「57.3 ± 2%」で視聴率を調査するには、標本の大きさが**2350**以上である必要があります。

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 :

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

確率分布に関するExcelの関数

推定・検定で利用できる、確率分布に関するExcelの関数を紹介します。

そのほかの関数

LN

- 自然対数 ($\log_e = \ln$) の値を計算するには、**LN**関数を利用します。

LN (自然対数の値を返す)

- 書式 : LN(数値)
- 引数 : 数値 … : 自然対数を求める正の実数
- 例: $\log_e 10 = \ln 10$ の階乗を計算する

```
=LN(10)
```

EXP

- 自然対数の底 (e=2.712…) のべき乗を計算するには、**EXP**関数を利用します。

EXP (自然対数の底のべき乗の値を返す)

- 書式 : EXPT(数値)
- 引数 : 数値 … : べき乗の指数
- 例: e^2 の階乗を計算する

```
=EXP(2)
```

正規分布

NORMDIST

NORMDIST（正規分布において任意のxに対する累積確率pを返す）

- 書式 : NORMDIST(x, 平均, 標準偏差, 定数)
  - 引数 : x : 横軸 x の値
  - 引数 : 平均 : データの平均
  - 引数 : 標準偏差 : データの標準偏差
  - 引数 : 定数 : 「TRUE」なら累積確率、「FALSE」なら確率分布関数の値を返す
- 例 : 平均が1、標準偏差が2の正規分布でxが0のときの累積確率を計算する

```
=NORMDIST(0, 1, 2, TRUE)
```

NORMINV

NORMINV（正規分布において累積確率pに対するxの値を返す）

- 書式 : NORMINV(p, 平均, 標準偏差)
  - 引数 : p : 累積確率

- 引数：平均 ：データの平均
- 引数：標準偏差 ：データの標準偏差

○ 例：平均が1、標準偏差が2の正規分布で累積確率が0.975（97.5%）のときのxの値を計算する

```
=NORMINV(0.975, 1, 2)
```

標準正規分布

NORMSDIST

NORMSDIST（標準正規分布において任意のzに対する累積確率pを返す）

○ 書式：NORMSDIST(z)

- 引数：z ：横軸 z の値

○ 例：標準正規分布でzが1.95のときの累積確率を計算する

```
=NORMSDIST(1.95)
```

NORMSINV

NORMSINV（標準正規分布において累積確率pに対するzの値を返す）

○ 書式：NORMSINV(p)

- 引数：p ：累積確率

○ 例：累積確率が0.95（95%）のときのzの値を計算する

```
=NORMINV(0.95)
```

t分布

TDIST

TDIST（t分布において任意のt値に対する上側確率pを返す）

○ 書式：TDIST(t, f, 定数)

- 引数：t ：横軸 t の値
- 引数：f ：自由度
- 引数：定数 ：「1」ならpの値、「2」ならpの2倍の値を返す

○ 例：自由度が4のt分布でtが4.6のときの上側確率を計算する

```
=TDIST(4.6, 4, 1)
```

TINV

TINV（t分布において両側確率pに対するt値を返す）

○ 書式：TINV(p, f)

- 引数：p ：両側確率（上側確率を求める場合はpを2倍する）
- 引数：f ：自由度

○ 例：自由度が4のt分布で両側確率が0.05 (5%) のときのtの値を計算する

```
=TINV(0.05, 4)
```

カイ2乗分布

CHIDIST

CHIDIST（カイ2乗分布において任意のカイ2乗値 xに対する上側確率pを返す）

○ 書式 : CHIDIST(x, f)

- 引数 : x ：カイ2乗値 x の値
- 引数 : f ：自由度

○ 例：自由度が10のカイ2乗分布でカイ2乗値が18のときの上側確率を計算する

```
=CHIDIST(18, 10)
```

CHIINV

CHIINV（カイ2乗分布において上側確率pに対するカイ2乗値 x を返す）

○ 書式 : CHIINV(p, f)

- 引数 : p ：上側確率
- 引数 : f ：自由度

○ 例：自由度が10のカイ2乗分布で上側確率が0.05 (5%) のときのカイ2乗値の値を計算する

```
=CHIINV(0.05, 10)
```

F分布

FDIST

FDIST（F分布において任意の自由度とF値に対する上側確率pを返す）

○ 書式 : FDIST(F, f1, f2)

- 引数 : F ：F値
- 引数 : f1 ：第1自由度
- 引数 : f2 ：第2自由度

○ 例：第1自由度が3、第2自由度が4のF分布でF値が18のときの上側確率を計算する

```
=FDIST(18, 3, 4)
```

FINV

FINV（F分布において任意の自由度と上側確率pに対するF値を返す）

○ 書式 : CHIINV(p, f1, f2)

- 引数 : p ：上側確率
- 引数 : f1 ：第1自由度

■ 引数：f2 ：第2自由度

○ 例：第1自由度が3、第2自由度が4のF分布で上側確率が0.05（5%）のときのカイ2乗値の値を計算する

```
=FINV(0.05, 3, 4)
```