統計について

高校２年１組　大道修

1　はじめに

本日は部誌を手にとっていただきありがとうございます。
統計ということについて、私の部誌を読んで、興味を持っていただければ有難く思います。

2　使用する語句

(1) 分布 $\cdots$ ある現象がさまざまの大きさで起こること。

例) 成績は分布する。

(2) 確率 $\cdots$ ある現象が起こる確からしさの程度を割合であらわしたもの。

例) さいころを振って 1 が出る確率は $\frac{1}{6}$ である。

(3) 母集団 $\cdots$ あらゆる現象のあらゆる観測値の集合を母集団と言う。

例) 灘校生 60 回生の数学の成績の母集団。
注) 大きさが無限のものを無限母集団、有限のものを有限母集団と呼ぶ (上の例は有限母集団)。

(4) 標本 $\cdots$ 母集団から一部分を取り出したもの。

注) ひとつの標本に含まれる観測データの数のことを標本サイズと言う。

3　分布の特性値

統計で大切なのは、分布の特徴を知ることである。ここで分布の特徴を知るための基本的特性値を見て行きたい。

(1) 分布の位置の特性値

a. 平均　(記号　$\overline{x}$)
これは良く知られているであろうが一応計算式を書く。

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

b. 中央値　(記号　$\widetilde{x}$)
分布の真ん中の点。例えば $(30, 50, 70, 80, 90)$ という分布があるとすると中央値は 70 となる。
また、$(30, 50, 70, 80)$ のときの中央値は $\frac{1}{2}(50 + 70) = 60$ と求めると良い。

c. 最頻値　(記号　$\widehat{x}$)
観測データを階級に分けたときに最も度数 (階級に含まれる個数) が大きい階級の階級値を最頻値と言う。

(2) 分布の広がりの特性値

a. 分散　(記号　S^2)

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

の形で表される。
個々の i に対し、$x_i - \overline{x}$ を偏差と言い、分布全体としての散らばりの指標とするために偏差の 2 乗を合計したものを自由度 $n-1$ で割ったものである。

b. 標準偏差　(記号　S)
記号からもわかるように

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

偏差値を求めるときはそのものを x_1 とすると

$$(x_1 - \overline{x}) \div S \times 10 + 50$$

という値である。

c. 変動係数　(記号　C)

$$C = \frac{S}{\overline{x}}$$

これは異なる標本の変動を比較するのに有効に使える。
例) 象の体重の分布と蝶の体重の分布を比較すれば、標準偏差の場合明らかに象の分布が大きくなり、分布の比較はうまくいかない。しかし変動係数であれば平均からの乖離率であり割合なので、うまく比較できる。

4　正規分布

(1) 正規分布の特徴

- Bell Shape(釣鐘型) である
- 左右対称である
- 平均 μ は中央にある
- 平均と中央値と最頻値は一致する

(2) 正規分布の確率密度関数

a. 確率密度関数とは
確率密度関数とは、確率分布の形を現す方程式のことをいう。またこのとき $a \sim b$ の値をとる確率は

$$\int_a^b f(x)\ dx$$

の形で表せる。

b. 正規分布の確率密度関数は?

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} (\text{ただし}\sigma\text{は標準偏差}\quad \mu\text{は平均})$$

しかしながらこの関数の原始関数は既知の形で表せないので、普通は表を引くか、コンピューターを使う。

c. 正規分布の平均、分散、標準偏差
平均 μ、分散 σ^2、標準偏差 σ は確率密度関数に出てくるので、これより求めることはできず、何らかの形で他から教えてもらわなければならない。

(3) 正規分布の標準化

$z = \dfrac{x-\mu}{\sigma}$ とすると

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}z^2} \end{aligned}$$

となり、標準正規分布 (平均 0、標準偏差 1 の正規分布) となり、このようにすると、計算が楽になるのと、標準正規分布が使えるようになるので良い。

(4) 中心極限定理

$x_1, x_2, \ldots, x_n$ を独立で同分布な確率変数とする。m と v でこれらの確率の期待値と分散を表すことにすると

$$\lim_{n\to\infty} P\{a < \frac{\sum x_i - mv}{\sqrt{mv}} < b\} = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad dx$$

となる。このままでは意味が分からないと思うので意味を書くと、いかなる分布でもその標本平均を標準化した $z = \dfrac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ は n が大きくなるにつれて標準正規分布になる、ということである。

5 t 分布

(1) t 分布の確率密度

$$f(t) = \frac{P(\frac{m+1}{2})}{P(\frac{m}{2})\sqrt{m\pi}} (1 + \frac{t^2}{m})^{-\frac{1}{2}(m+1)} \quad (\text{ただし } m \text{ は標本の自由度で } m = n-1)$$

正規分布と同じで、この式を計算する必要はなく、表かコンピューターを使えばよい。

(2) t 分布する変数

t 分布する変数の代表的なものを一つ挙げる。

$$t = \frac{\overline{x}-\mu}{S/\sqrt{n}} \quad (\text{ただし} \quad \mu = \text{母集団} \quad S = \text{標本標準偏差} \quad \overline{x} = \text{標本平均} \quad n = \text{標本サイズ})$$

6　x^2 分布

(1) x^2 分布の作り方

1. 標本サイズ n の標本を抽出する。
2. 標本データの一つ一つを標準化。(x_i を $z_i = \dfrac{x_i - \mu}{\sigma}$ とする)
3. z_i を 2 乗する
4. $x^2 = \sum_{i=1}^{n} z_i$ を求める。

このようにしてできた x^2 の分布を x^2 分布という。

(2) x^2 分布の確率関数

$$f(x^2) = \frac{1}{2^{\frac{m}{2}} P(\frac{m}{2})} (x^2)^{\frac{m}{2}-1} \cdot e^{(-\frac{1}{2}x^2)}$$

やはり、これもこのまま計算せずに表を引くかコンピュータを使って求める。

7　母平均の推定

(中心極限定理より正規分布を用いる)

(1) 母標準偏差が分かっている時

a. 標本平均を標準化すると

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

b. $\overline{x}$ の平均 $\mu\overline{x}$ は母平均と等しく $\sigma\overline{x} = \dfrac{\sigma}{\sqrt{n}}$ より

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

c. 未知数 μ について解くと

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

d. ここで z を求める。

z の求め方は、求める信頼係数を $x\%$ とすると

$$50 - \frac{1}{2}x \leq 100 \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a^2} \quad da \leq 50 + \frac{1}{2}x$$

となる z を求めればよい。実際には表かコンピューターを用いて求める。

例)$x = 95$ の時

$$-1.96 \leq z \leq 1.96$$

e. これより $x = \alpha$ のとき $-\beta \leq z \leq \beta$ となるとすると $(\beta \geq 0)$

$$\overline{x} - \beta\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + \beta\frac{\sigma}{\sqrt{n}}$$

(2) 母標準偏差が分からず標本数も少ない時

a. $t = \dfrac{\overline{x} - \mu}{S/\sqrt{n}}$ が t 分布するので

$$t = \frac{\overline{x} - \mu}{S/\sqrt{n}} \quad \text{となる。}$$

b. μ について解くと

$$\mu = \overline{x} - t\frac{S}{\sqrt{n}}$$

c. t を求める。

t の求め方は信頼区間を $x\%$ とすると

$$50 - \frac{1}{2}x \leq \int_{-\infty}^{t} \frac{P(\frac{m+1}{2})}{P(\frac{m}{2})\sqrt{m\pi}}(1 + \frac{a^2}{m})^{-\frac{1}{2}m+1} \quad da \leq 50 + \frac{1}{2}x$$

d. これより $x = \alpha$ のときに $-\beta \leq t \leq \beta$ となるとすると $(\beta \geq 0)$

$$\overline{x} - \beta\frac{S}{\sqrt{n}} \leq \mu \leq \overline{x} + \beta\frac{S}{\sqrt{n}}$$

(3) 母標準偏差が分かっており、標本数も多い時

このとき、t 分布は標準正規分布に非常に近く、また 30 以上になると、t 分布表は 30 より大きくなると詳細な表になっていないので正規分布で代用すると

$$\overline{x} - z\frac{S}{\sqrt{n}} \leq \mu \leq \overline{x} + z\frac{S}{\sqrt{n}} \quad \text{となるようになる。}$$

z の求め方は (1) と同じ。

8　母標準偏差の推定

(1) 標本分散が S^2 とすると

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

これより

$$\begin{aligned}
(n-1)S^2 &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \sum\{(x_i - \mu) + (x_i + \mu)\}^2 \\
&= \sum(x_i - \mu)^2 - n(\overline{x} - \mu)^2 \\
\iff \frac{(n-1)S^2}{\sigma^2} &= \sum(\frac{x_i - \mu}{\sigma})^2 - (\frac{\overline{x} - \mu}{\sigma/\sqrt{n}})^2 \\
\therefore \quad \sum(\frac{\overline{x_i} - \mu}{\sigma})^2 &= (\frac{\overline{x} - \mu}{\sigma/\sqrt{n}})^2 + \frac{(n-1)S^2}{\sigma^2}
\end{aligned}$$

左辺は自由度 n の x^2 分布に従い、右辺第一項は 1 個の標準正規変数の二乗であるから、右辺第二項は $(n-1)$ 個の標準正規変数の二乗和に他ならない。したがって $\dfrac{(n-1)S^2}{\sigma^2}$ は自由度 $n-1$ の x^2 分布に従う。

(2) 求め方

a. 信頼係数を設定する。
　信頼係数を $\alpha\%$ とする。

b. x^2 分布表から $x^2_{\alpha}(m)$ と $x^2_{100-\alpha}(m)$ を読み取る。
　注)$m = n-1$(n は標本サイズ)

c. 母分散 σ^2 の範囲を求める

$$(n-1)S^2/x^2_{100-\alpha}(m) \leq \sigma^2 \leq (n-1)S^2/x^2_{\alpha}(m)$$

d. 母標準偏差を求める

$$\sqrt{(n-1)S^2/x^2_{100-\alpha}(m)} \leq \sigma \leq \sqrt{(n-1)S^2/x^2_{\alpha}(m)}$$

9　仮説検定

(1) 仮説の立て方
仮説検定において、仮説は次の二つをおく。

帰無仮説　　(記号 H_0)
対立仮説　　(記号 H_1)

ここにおいて示したいのは H_1 である。
つまり、H_1 に反する仮説をたてて H_0 を否定することで、H_1 を示すのである。

(2) なぜこのような方法をとるか
このような方法をとる理由は、統計には

a) H_0 が正しいのに H_0 を棄却して正しくない H_1 を採択する
という誤りと

b) H_1 が正しいのに H_1 を採択しない
という誤り

があり、有意水準は a) の確率に相当するが、b) の確率を調節することはできないので、b) が起こることがないようにし、また、過誤の確率を調節するためである。

(3) 仮説検定の方法

a. 仮説を立てる。
(1) の方法で、示したいこと (H_1) と、示したいことと反すること (H_0) を立て

b. H_0 が棄却域に入ることを示す。
注) 棄却域とは、信頼係数に対応する有意水準の範囲のこと。

c. H_0 が棄却されることより、H_1 が棄却できないので採択されることをいう。

d. 結論を書く。

10　相関分析

(1) 相関関係とは？

2 変数 X, Y が、X が変化するにつれて Y が変化するとき、X, Y との間に相関関係があると言う。

(2) 相関の程度

X が増えるにつき Y が増える関係を正の相関、
X が増えるにつき Y が減る関係を負の相関と言う。
また、X と Y の相関の程度を示す係数を相関係数 (記号 r) という。

正相関のとき　　$0 < r \leq 1$
負相関のとき　　$-1 \leq r < 0$

となる。

(3) r の求め方

$$\begin{aligned} r^2 &= \frac{\{\sum (X_i - \overline{X})(Y_i - \overline{Y})\}^2}{\sum (X_i - \overline{X})^2 \cdot \sum (Y_i - \overline{Y})^2} \\ &= \frac{(\sum X_i Y_i - n\overline{X}\ \overline{Y})^2}{(\sum X_i^2 - n\overline{X}^2)(\sum Y_i^2 - n\overline{Y}^2)} \end{aligned}$$

となるので、
正相関のときは $r = \sqrt{r^2}$ となり、
負相関のときは $r = -\sqrt{r^2}$ となり。

(4) 相関係数の有意性

相関係数は分布し、この分布は自由度によって変わる。自由度は、標本サイズ − 変数サイズの式で表される。
ここで、相関係数の有意水準がどのようになるかは表を引く。
例えば、自由度 10、5% 有意水準では、0.576 である。

11　おわりに

最後まで読んでいただき有難うございます。初めての部誌なのでいろいろと至らないところがあると思いますが、そこのところは大目に見てください。ご意見、ご感想などがありましたら

omichiosamu@yahoo.co.jp

までメールをお願いします。

参考文献

[1] 鳥居泰彦著『はじめての統計学』(日本経済新聞社、1994 年)