数学B　統計的推測

~高校数学のまとめ~

教科書をもとに定義や定理を独自にパネル形式でまとめています．
何度も書き直し，加筆修正を繰り返しており，完成したものではありません．
人によっては不要な部分もあるでしょう．そういうときは読み飛ばしてください．

全数調査と標本調査

1 対象とする集団の全部のものを調べる調査を **全数調査**（ぜんすうちょうさ）という.

2 対象とする集団の一部分を抜き出して調べる調査を **標本調査**（ひょうほんちょうさ）**(サンプリング)** という.

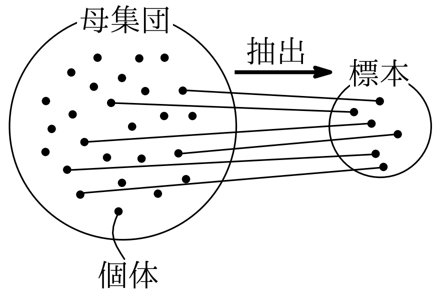
母集団と標本

標本調査の場合について,

対象とする集団全体を **母集団**（ぼしゅうだん）という.

母集団に属する個々のものを **個体**（こたい）といい,

個体の総数を **母集団の大きさ** という.



母集団から選び出された個体の一部を **標本**（ひょうほん）**(サンプル)** といい,

標本を選び出すことを **抽出**（ちゅうしゅつ）という.

標本に含まれる個体の個数を **標本の大きさ** という.

㊓ 標本調査では標本から母集団を推測する.

標本の抽出

母集団の各個体を等しい確率で抽出する方法を **無作為抽出**（むさくいちゅうしゅつ）といい,

無作為抽出によって選ばれた標本を **無作為標本**（むさくいひょうほん）という.

無作為抽出を行うには **乱数表**（らんすうひょう）や **乱数**（らんすう）**さい** などが使われる.

㊓ 乱数表は 0 から 9 までの数字を不規則に並べた表,
乱数さいは正二十面体のさいころで各面に 0 から 9 までの数字が 2 度ずつ刻まれている.
コンピュータを用いての無作為抽出もできる.

母集団の変量

母集団において，調査の対象となっている性質をその母集団の **特性** という．
とくに，
数量的に表される特性は確率変数として取り扱うことができるから **変量** という．

例 ある学校の生徒のテストの点数，身長，体重などを特性といい，数量的に取り扱えるから変量という．

母集団分布

大きさ N の母集団において，

変量 x の異なるとりうる値を x_1, x_2, $\cdots$, x_r とし，

それぞれの個体の個数を f_1, f_2, $\cdots$, f_r とする．

ここで $f_1+f_2+\cdots+f_r=N$

階級値	度数
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_r	f_r
計	N

この母集団から1個の個体を無作為抽出するとき

その個体における変量 x の値 X は偶然に決まるが

$X=x_k$ となる確率は $\dfrac{f_k}{N}$ $(k=1,\ 2,\ \cdots,\ r)$

つまり，次のような確率分布となる．

X	x_1	x_2	$\cdots$	x_k	$\cdots$	x_r	計
P	$\dfrac{f_1}{N}$	$\dfrac{f_2}{N}$	$\cdots$	$\dfrac{f_k}{N}$	$\cdots$	$\dfrac{f_r}{N}$	1

この確率分布は，母集団の相対度数の分布と一致する．

母集団における変量 x の平均を m，標準偏差を σ

確率変数 X の期待値を $E(X)$，標準偏差を $\sigma(X)$ とすると

$$E(X)=m,\ \ \sigma(X)=\sigma$$

一般に 母集団における変量 x の分布を **母集団分布**（ぼしゅうだんぶんぷ） といい，

その平均値を **母平均**（ぼへいきん），分散を**母分散**（ぼぶんさん），標準偏差を **母標準偏差**（ぼひょうじゅんへんさ） という．

大きさ1の無作為標本における変量 x の値 X は母集団分布に従う．

確率変数 X の期待値，分散，標準偏差は

それぞれ母平均，母分散，母標準偏差に一致する．

㊖ X の確率分布は，母集団において調査の対象となっている変量を特徴づけるものである．

復元抽出・非復元抽出

母集団の中から標本を抽出する場合について

1　毎回もとに戻しながら次のものを1個ずつ取り出すことを
復元抽出（ふくげんちゅうしゅつ）という．

2　取り出したものをもとに戻さないで取り出し続けて抽出することを
非復元抽出（ひふくげんちゅうしゅつ）という．

母集団の中から大きさ n の標本を無作為に抽出し，
その n 個の個体における変量の値を X_1, X_2, $\cdots$, X_n とする．

1 復元抽出のとき

母集団から大きさ1の標本を無作為に抽出する試行を n 回繰り返す
反復試行であるから，X_1, X_2, $\cdots$, X_n は
それぞれが母集団分布に従う互いに独立な確率変数となる．

2 非復元抽出のとき

標本は n 個の確率変数 X_1, X_2, $\cdots$, X_n であるが，
これらは互いに独立ではない．
しかし，母集団の大きさが標本の大きさ n に比べて十分大きいときは
母集団の構成にはほとんど変化がないから復元抽出との違いは小さくなる．
すなわち，非復元抽出で取り出した標本も近似的に復元抽出で取り出した
標本とみなすことができ，X_1, X_2, $\cdots$, X_n は
それぞれが母集団分布に従う互いに独立な確率変数となる．

補 2 母集団の大きさが標本の大きさ n に比べて十分大きいときの目安は
母集団の大きさを N として n は N の10分の1以下 $\left(n \leqq \frac{N}{10}\right)$

標本平均

母集団から無作為抽出する大きさ n の標本の変量を X_1，X_2，…，X_n とする．

これらの平均 $\overline{X}=\dfrac{X_1+X_2+\cdots+X_n}{n}$ を 標本平均（ひょうほんへいきん）という．

標本平均 $\overline{X}$ は，抽出される標本によって変化する確率変数である．

標本平均の期待値と標準偏差

母平均 m，母標準偏差 σ の母集団から大きさ n の無作為標本を復元抽出するとき，その標本平均を $\overline{X}$ とすると，次が成り立つ．

1. $\overline{X}$ の期待値は $E(\overline{X})=m$
2. $\overline{X}$ の分散は $V(\overline{X})=\dfrac{\sigma^2}{n}$
3. $\overline{X}$ の標準偏差は $\sigma(\overline{X})=\dfrac{\sigma}{\sqrt{n}}$

㋕ 母平均 m，母標準偏差 σ の母集団から大きさ n の無作為標本の変量を X_1，X_2，…，X_n とする．

1.
$$\begin{aligned} E(\overline{X}) &= E\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) \\ &= \frac{1}{n}\{E(X_1)+E(X_2)+\cdots+E(X_n)\} \\ &= \frac{1}{n}(\underbrace{m+m+\cdots+m}_{n\text{ 個}}) \\ &= \frac{1}{n}\cdot nm \\ &= m \end{aligned}$$

2.
$$\begin{aligned} V(\overline{X}) &= V\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) \\ &= \left(\frac{1}{n}\right)^2\{V(X_1)+V(X_2)+\cdots+V(X_n)\} \\ &= \frac{1}{n^2}(\underbrace{\sigma^2+\sigma^2+\cdots+\sigma^2}_{n\text{ 個}}) \\ &= \frac{1}{n^2}\cdot n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

3.
$$\sigma(\overline{X})=\sqrt{V(\overline{X})}=\sqrt{\frac{\sigma^2}{n}}=\frac{\sigma}{\sqrt{n}}$$

㋭ 非復元抽出の場合も，n に比べて母集団の大きさが十分大きいならば復元抽出と同様に扱える．

標本平均の分布の正規分布による近似

母平均 m，母分散 σ^2 の母集団から

無作為抽出された大きさ n の標本平均 $\overline{X}$ の分布は n が十分大きいとき

近似的に正規分布 $N\left(m,\ \dfrac{\sigma^2}{n}\right)$ に従う.

このことから

$$Z = \frac{\overline{X} - m}{\dfrac{\sigma}{\sqrt{n}}}$$

とおくと 確率 Z は標準正規分布 $N(0,\ 1)$ に従う.

大数の法則

母平均 m の母集団から大きさ n の無作為標本を抽出するとき

その標本平均 $\overline{X}$ は n を大きくすると母平均 m に近づく.

すなわち $\displaystyle\lim_{n\to\infty} \overline{X} = m$

㊔ 標本平均 $\overline{X}$ の分散を $V(\overline{X})$ として

$$\lim_{n\to\infty} V(\overline{X}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

すなわち，n を大きくすると $\overline{X}$ は母平均 m の近くに限りなく集中して分布する.

㊟ n を大きくすると $\overline{X}$ が m に近い値をとる確率が 1 に近づく.

母平均に対する信頼度 95 %の信頼区間

母平均 m，母標準偏差 σ の母集団から

抽出された大きさ n の無作為標本の標本平均 $\overline{X}$ を考える．

n が大きいとき，母平均 m に対する信頼度 95 %の信頼区間は

$$\overline{X}-1.96\cdot\frac{\sigma}{\sqrt{n}} \leqq m \leqq \overline{X}+1.96\cdot\frac{\sigma}{\sqrt{n}}$$

すなわち

$$\left[\overline{X}-1.96\cdot\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X}+1.96\cdot\frac{\sigma}{\sqrt{n}}\right]$$

注 「母平均 m に対して信頼度 95 %の信頼区間を求める」ことを
「母平均 m を信頼度 95 %で推定する」ということがある．

考 n が大きいとき，$\overline{X}$ の確率分布は近似的に正規分布 $N\left(m,\ \dfrac{\sigma^2}{n}\right)$ に従う．

$Z=\dfrac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}$ とおくと Z は標準正規分布 $N(0,\ 1)$ に従う．

正規分布表から $P\big(|Z|\leqq 1.96\big)=2\times P(0\leqq Z\leqq 1.96)=2\times 0.4750=0.95$

すなわち，母平均 m に対する信頼度 95 %の信頼区間は $|Z|\leqq 1.96$ として

$$\left|\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}\right|\leqq 1.96 \iff \overline{X}-1.96\cdot\frac{\sigma}{\sqrt{n}}\leqq m\leqq \overline{X}+1.96\cdot\frac{\sigma}{\sqrt{n}}$$

補 $P\big(|Z|\leqq 2.58\big)=2\times P(0\leqq Z\leqq 2.58)=2\times 0.4951=0.9902\fallingdotseq 0.99$

これより，母平均 m に対する信頼度 99 %の信頼区間は

$$\overline{X}-2.58\cdot\frac{\sigma}{\sqrt{n}}\leqq m\leqq \overline{X}+2.58\cdot\frac{\sigma}{\sqrt{n}}$$

母平均に対する信頼度 D %の信頼区間の幅

母平均 m に対する信頼度 D %の信頼区間を $A\leqq m\leqq B$ とするとき

この信頼区間の幅は $B-A$

例 母平均に対する信頼度 95 %の信頼区間 の幅は

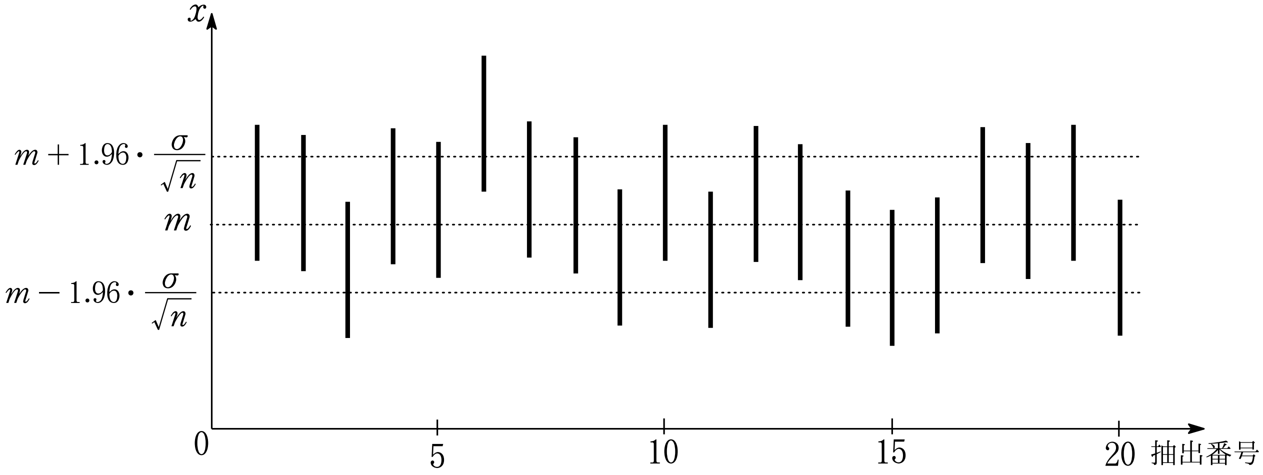
$$\overline{X}+1.96\cdot\frac{\sigma}{\sqrt{n}}-\left(\overline{X}-1.96\cdot\frac{\sigma}{\sqrt{n}}\right)=2\cdot 1.96\cdot\frac{\sigma}{\sqrt{n}}=3.92\cdot\frac{\sigma}{\sqrt{n}}$$

母平均に対する信頼度 95 %の信頼区間の意味

母平均 m に対する信頼度 95 %の信頼区間とは

無作為抽出を繰り返し，このような区間を，例えば 100 個作ると，

m を含む区間が 95 個位あることを意味する．



㊣ 上図では抽出番号 20 までの 20 個の区間を作り，抽出番号 6 の区間以外の 19 個の区間は m を含んでいる．

母比率と標本比率

母集団の中で，ある特性 A をもつ個体の割合を p とする．

この p を特性 A をもつ個体の母集団における **母比率**（ぼひりつ）という．

標本の中で特性 A をもつ個体の割合を **標本比率**（ひょうほんひりつ）という．

例 学校で，テストの点数が 80 点以上ある人，身長が 180cm 以上ある人などが特性．

標本比率の分布の正規分布による近似

特性 A をもつ個体の母比率が p である母集団から

大きさ n の標本を無作為抽出し

その標本に含まれる特性 A をもつ個体の個数を X とする．

このとき 確率変数 X は 二項分布 $B(n,\ p)$ に従い，n が大きいとき

近似的に正規分布 $N\big(np,\ np(1-p)\big)$ に従う．

また 標本比率を $\dfrac{X}{n}=R$ とすると

確率変数 R は正規分布 $N\left(p,\ \dfrac{p(1-p)}{n}\right)$ に従う．

このことから

$$Z=\frac{R-p}{\sqrt{\dfrac{p(1-p)}{n}}}$$

とおくと 確率 Z は標準正規分布 $N(0,\ 1)$ に従う．

考 確率変数 X は $B(n,\ p)$ に従うので $E(X)=np$，$V(X)=np(1-p)$

確率変数 R の平均は $E(R)=E\left(\dfrac{X}{n}\right)=\dfrac{1}{n}E(X)=\dfrac{1}{n}\cdot np=p$

確率変数 R の分散は $V(R)=V\left(\dfrac{X}{n}\right)=\left(\dfrac{1}{n}\right)^2V(X)=\dfrac{1}{n^2}\cdot np(1-p)=\dfrac{p(1-p)}{n}$

母比率に対する信頼度95 %の信頼区間

母比率 p の母集団から

抽出された大きさ n の無作為標本の標本比率 R を考える.

n が大きいとき，母平均 p に対する信頼度95 %の信頼区間は

$$R-1.96\sqrt{\frac{R(1-R)}{n}} \leqq p \leqq R+1.96\sqrt{\frac{R(1-R)}{n}}$$

すなわち

$$\left[R-1.96\sqrt{\frac{R(1-R)}{n}}\ ,\ R+1.96\sqrt{\frac{R(1-R)}{n}}\ \right]$$

注 「母比率 p に対して信頼度95 %の信頼区間を求める」ことを
「母比率 p を信頼度95 %で推定する」ということがある.

考 n が大きいとき，R の確率分布は近似的に正規分布 $N\left(p,\ \dfrac{p(1-p)}{n}\right)$ に従う.

$Z=\dfrac{R-p}{\sqrt{\dfrac{p(1-p)}{n}}}$ とおくと Z は標準正規分布 $N(0,\ 1)$ に従う.

正規分布表から $P\big(|Z|\leqq 1.96\big)=2\times P(0\leqq Z\leqq 1.96)=2\times 0.4750=0.95$

すなわち，母比率 p に対する信頼度95 %の信頼区間は $|Z|\leqq 1.96$ として

$$\left|\frac{R-p}{\sqrt{\dfrac{p(1-p)}{n}}}\right|\leqq 1.96$$

$$\iff R-1.96\sqrt{\frac{p(1-p)}{n}} \leqq p \leqq R+1.96\sqrt{\frac{p(1-p)}{n}}$$

n が十分大きいときは R を p とみなしてよいので，根号内の p を R で置き換えて

$$R-1.96\sqrt{\frac{R(1-R)}{n}} \leqq p \leqq R+1.96\sqrt{\frac{R(1-R)}{n}}$$

補 $P\big(|Z|\leqq 2.58\big)=2\times P(0\leqq Z\leqq 2.58)=2\times 0.4951=0.9902\fallingdotseq 0.99$

これより，母平均 m に対する信頼度99 %の信頼区間は

$$R-2.58\sqrt{\frac{R(1-R)}{n}} \leqq p \leqq R+2.58\sqrt{\frac{R(1-R)}{n}}$$

母比率に対する信頼度 D %の信頼区間の幅

母比率 p に対する信頼度 D %の信頼区間を $A \leqq p \leqq B$ とするとき

この信頼区間の幅は $B-A$

例 母比率に対する信頼度 95 %の信頼区間の幅は

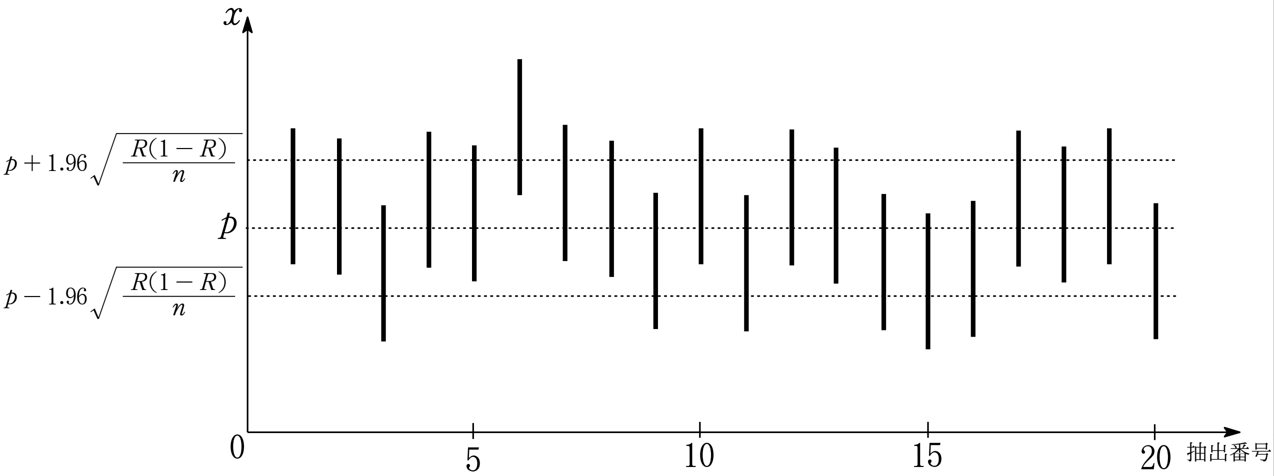
$$R+1.96\sqrt{\frac{R(1-R)}{n}}-\left(R+1.96\sqrt{\frac{R(1-R)}{n}}\right)=2\cdot 1.96\sqrt{\frac{R(1-R)}{n}}$$
$$=3.92\sqrt{\frac{R(1-R)}{n}}$$

母平均に対する信頼度 95 %の信頼区間の意味

母比率 p に対する信頼度 95 %の信頼区間とは

無作為抽出を繰り返し，このような区間を，例えば 100 個作ると，

p を含む区間が 95 個位あることを意味する．



補 上図では抽出番号 20 までの 20 個の区間を作り，抽出番号 6 の区間以外の 19 個の区間は p を含んでいる．