§３　統計調査

3.1　統計調査

統計調査とは？

○**統計調査**には，国勢調査のように対象となる集団の全部について調べる**全数調査**と，アンケート調査のように集団の中の一部を調べて全体を推測する**標本調査**があります。

（※日本では５年おきに人口，性別，年齢，世帯構成や就業状況などのさまざまな情報を調べる**国勢調査**が行われています。）

統計調査
- 全数調査[対象となる集団の全部を調査]……国勢調査など
- 標本調査[集団の一部を調べて全体を推測]…アンケート調査など

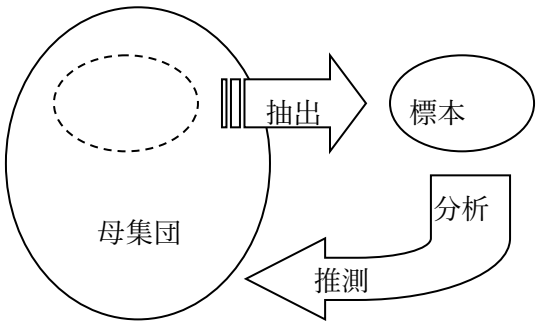
3.2　抽出

抽出とは？

○統計調査において，調査の対象となる集団を**母集団**といい，母集団の構成要素の個数を**母集団の大きさ**といいます

○統計調査のために母集団から取り出された要素の集まりを**標本**といい，標本に含まれる要素の個数を**標本の大きさ**といいます。

○母集団から標本を取り出すことを**抽出**といいます。



統計的推定

1）母集団から標本を抽出する。
2）標本を分析する
3）分析された結果を用いて，母集団についての推測をする

抽出方法を紹介します

○母集団からランダムに標本を抽出する方法を**無作為抽出**といいます。
○母集団を年齢などの層別に分け，各層から標本を抽出する方法を
層別抽出といいます。
○母集団をいくつかのグループに分け，
その中からランダムにグループを抽出します。
更に，選んだグループをいくつかの小グループに分け
その中からランダムに標本を抽出する方法を
多段抽出といいます。
例）全国を都道府県別に分け，選ばれ都道府県から
対象となる地区をランダムに選び，その地区の
住民の何人かにアンケート調査を行う。
[※集団の大きさによりグループの細分化を何回か行います]

§４　仮説検定

4.1　仮説検定

仮説検定とは？

○調べたい(統計的)仮説を**帰無仮説**(記号：H_0)といい，
H_0 との比較のために設けられる仮説を**対立仮説**(記号：H_1)といいます。
○帰無仮説 H_0 が正しくないと判断することを，H_0 を**棄却する**といい，
そうでない場合を**受容する**といいます。
○帰無仮説 H_0 を棄却するか受容するかの判断をすることを，
仮説検定といいます。

過誤とは？

○仮説検定を行ったとき，
次のような２種類の**過誤**(error)を犯すことがあります。
(1) 帰無仮説 H_0 を正しくないとし棄却したが，実際は H_0 は正しかった
場合の過誤を，**第１種の過誤**と言います。
(2) 帰無仮説 H_0 を正しいと容認したが，実際は H_0 は正しくなかった
場合の過誤を，**第２種の過誤**と言います。

棄却域，有意水準とは？

○帰無仮説 H_0 を棄却すべき統計量の満たす範囲を**棄却域**という。
○仮説検定においては，あらかじめ α $(0<\alpha<1)$ を定めて，
統計量が棄却域に入る確率が α となるようにする。
このときの α を**有意水準**または**危険率**という。
（有意水準 α は 0.05 又は 0.01 にすることが多い）
○有意水準 α で帰無仮説 H_0 が棄却されるとき，
検定結果は有意水準 α で**有意である**といいます。

次に仮説検定で大切な考え方を述べます

帰無仮説 H_0 が受容する場合，他のいろいろなことから帰無仮説が正しいと考えられるときには，帰無仮説を**採択**するといいます。
つまり，「受容」と「採択」の間には開きがあるということです。
何度もデータを取り直して検定しても H_0 が棄却できなければ，
いよいよ H_0 を採用することになります。

＝＝＝

【まとめ：仮説検定の手順】

1） H_0 に基づく統計量 T を，標本（X_1，X_2 ,X_3,…,X_n）を用いて求める。
2）確率変数 T が従う確率分布を考える。
3）確率変数 T の実現値 t がある値以上または以下となるのかを観察する。
4）帰無仮説 H_0 の取捨を判断する。

注意(大切な考え方)　仮説検定とは
　　棄却される　⇒　仮説が正しくない
　　受容される　⇒　仮説が正しい
という二値論理的主張をしているのではなく，
　　棄却される　⇒　仮説は非常に疑わしい
　　受容される　⇒　仮説が正しくないというには
　　証拠が不十分である
という判断がなされるということです。

＝＝＝

4.2　独立性の仮説検定

独立事象とは？

まずは，復習からです。

２つの事象 A, B が**独立事象**であるとは，**積の法則**において

［独立事象］　$P(A\cap B)=P(A)P(B)$

が成り立つときを言います。

（※通常の積の法則は，**条件付き確率**を用いて表現されます。

［積の法則］　$P(A\cap B)=P(A)P_A(B)$ ）

この内容を利用して，**独立性の仮説検定**を行います。

次の例を用いて，その手順を説明していきます

例）ＴＶゲーム機使用と近視であることの間には，なんらかの関係があるかどうかを調べたい。そこで，ある中学１年生 500 人について，ゲーム機の使用者と近視者の数を実際に調べたら，下の表のようになった。

観測度数表

	近視者	近視でない者	計
ゲーム機　使用者	69	163	232
ゲーム機未使用者	58	268	268
計	127	373	500

[※このような表を**クロス集計表（分割表）**と言います]

計算の説明のために，一般的表記も同時に掲載していきます。

観測度数表

	B	$\overline{B}$	計
A	a	b	$a+b$
$\overline{A}$	c	d	$c+d$
計	$a+c$	$b+d$	N

[※記号 $\overline{A}$（$\overline{B}$）は A（B）の**補集合**を意味する]

今回の**帰無仮説**と**対立仮説**は，次のようになります。

帰無仮説　H_0 ：ゲーム機使用と近視との間には関係がない

［A, B は独立である］

対立仮説　H_1 ：ゲーム機使用と近視との間には関係がある

［A, B は独立でない］

相対度数表を確率とみなせば，ある中学 1 年生から無作為に 1 人を選ぶとき，

ゲーム機使用者(A)である確率は　　$P(A)=\dfrac{232}{500}\left(=\dfrac{a+b}{N}\right)$

近視者(B)である確率は　　$P(B)=\dfrac{127}{500}\left(=\dfrac{a+c}{N}\right)$

よって，ゲーム機使用者(A)で近視者(B)である確率は

$P(A\cap B)=P(A)P(B)=\dfrac{232}{500}\times\dfrac{127}{500}\left(=\dfrac{a+b}{N}\times\dfrac{a+c}{N}\right)$　[理由：独立事象より]

故に，500 人(総度数 N)のうち，ゲーム機使用者(A)で近視者(B)である度数は

$N\times P(A\cap B)=500\times\left(\dfrac{232}{500}\times\dfrac{127}{500}\right)\left(=\dfrac{(a+b)(a+c)}{N}\right)=58.9$ 人

であると期待される。それぞれの度数を計算すると，次の表のようになる。

期待度数表

	近視者	近視でない者	計
ゲーム機　使用者	58.9	173.1	232
ゲーム機未使用者	68.1	199.9	268
計	127	373	500

観測結果から得られる確率　$P(A)=\dfrac{a+b}{N}$，$P(B)=\dfrac{a+c}{N}$　を

用いて得られる度数表を**期待度数表**と言います。

つまり，それぞれの度数が「理論値」となります。

期待度数表

	B	$\overline{B}$	計
A	$\dfrac{(a+b)(a+c)}{N}$	$\dfrac{(a+b)(b+d)}{N}$	$a+b$
$\overline{A}$	$\dfrac{(c+d)(a+c)}{N}$	$\dfrac{(c+d)(b+d)}{N}$	$c+d$
計	$a+c$	$b+d$	N

[※補集合の確率　$P(\overline{A})=1-P(A)$]

このとき，統計量 $\boxed{T=\sum\dfrac{(\text{観測度数}-\text{期待度数})^2}{\text{期待度数}}}$ を考えます。

今回の統計量は

$$T=\frac{(69-58.9)^2}{58.9}+\frac{(163-173.1)^2}{173.1}+\frac{(58-68.1)^2}{68.1}+\frac{(210-199.9)^2}{199.9}=4.305$$

一般式　$T=\dfrac{\left\{a-\dfrac{(a+b)(a+c)}{N}\right\}^2}{\dfrac{(a+b)(a+c)}{N}}+\dfrac{\left\{b-\dfrac{(a+b)(b+d)}{N}\right\}^2}{\dfrac{(a+b)(b+d)}{N}}$

$$+\dfrac{\left\{c-\dfrac{(c+d)(a+c)}{N}\right\}^2}{\dfrac{(c+d)(a+c)}{N}}+\dfrac{\left\{d-\dfrac{(c+d)(b+d)}{N}\right\}^2}{\dfrac{(c+d)(b+d)}{N}}$$

$$=\frac{(ad-bc)^2}{N(a+b)(a+c)}+\frac{(ad-bc)^2}{N(a+b)(b+d)}+\frac{(ad-bc)^2}{N(c+d)(a+c)}+\frac{(ad-bc)^2}{N(c+d)(b+d)}$$

$$=\frac{(ad-bc)^2N}{(a+b)(c+d)(a+c)(b+d)}$$　（但し　$N=a+b+c+d$ ）

［統計量］　$T=\dfrac{(ad-bc)^2N}{(a+b)(c+d)(a+c)(b+d)}$

観測度数表

	B	$\overline{B}$	計
A	a	b	$a+b$
$\overline{A}$	c	d	$c+d$
計	$a+c$	$b+d$	N

実際は，この式に，観測度数　$a=69,\ b=163,\ c=58,\ d=268$　を代入して

$$T=\frac{(69\times210-163\times58)^2\times500}{232\times268\times127\times373}=4.305$$

と計算します。

この統計量Tは，標本の大きさが大きければ近似的に「自由度１のχ^2分布」に従うことが知られています。[※χ^2の読み：カイ　２乗]

よって，**有意水準**５％で**棄却域**は　$T\geqq\chi_1^2(0.05)=3.84$　となります。

[※次頁の説明も読んでください]

よって，今回の統計量$T=4.305$は棄却域に入っているので，

帰無仮説H_0は**棄却**されます。（⇔対立仮説H_1が**受容**される）

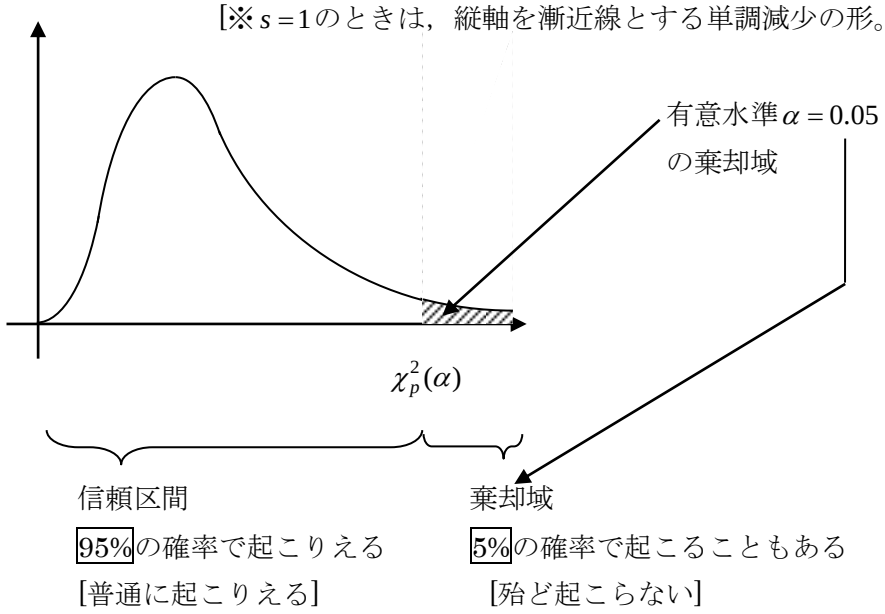
結論：「ゲーム機使用と近視との間には関係がある」ことが受容される。

＝＝

【研究：自由度 s の χ^2 分布】

自由度 s の χ^2 分布は，概ね下のような分布の形をしています。

[※ $s=1$ のときは，縦軸を漸近線とする単調減少の形。]



95%の確率で起こりえる
[普通に起こりえる]

5%の確率で起こることもある
[殆ど起こらない]

つまり，統計量 T が棄却域に入ると，「起こりえないことが起こった」
(理論的に矛盾する)ことになるので，帰無仮説が棄却されます。

※自由度 s はクロス集計表が $m\times n$ 型(今回は 2×2 型)のとき

［自由度］　$s=(m-1)(n-1)$

により，計算されます。(今回は $s=1\times 1=1$)

※棄却域の境になる値 $\chi_s(\alpha)$ の値は，下のような表が準備されています。

自由度 s	$\alpha=0.05$	$\alpha=0.01$
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09

注意：今回の統計量 $T=4.305$ は，有意水準 $\alpha=0.01$ では
棄却域 $T\geqq\chi_1^2(0.01)=6.63$ に入らないので，
帰無仮説 H_0 が受容されます。

＝＝

例題　ある都市で，有権者のＡ内閣に対する支持率を調べた。有権者から男性 150 人，女性 100 人を抽出し，支持する者，支持しない者を調べたらその人数は右の表のようであった。

	男	女	計
支持する	60	35	95
支持しない	90	65	155
計	150	100	250

Ａ内閣の支持率は男性と女性とで，違いがあるとみてよいか。有意水準５％で検定せよ。

［解答］　帰無仮説 H_0 ：男性と女性で支持率に差がない

対立仮説 H_1 ：男性と女性で支持率に差がある

（※帰無仮説には，必ず，関連性がない(独立な)ものを取り上げます。）

統計量　$T=\dfrac{(60\times65-35\times90)^2\times250}{150\times100\times95\times155}=0.637$

よって，棄却域 $T\geqq\chi_1^2(0.05)=3.84$ に入っていないので，

帰無仮説 H_0 は棄却できない。

問 3.10　ある食品の好みが，東京都と熊本県で差があるのかを調査した結果が，次の表のようになった。東京都と熊本県で好みの差があるとみてよいか。有意水準５％で検定せよ。（表は演習プリント参照）

□□□□【参考】□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

疑似相関の話をします。

○散布図で２つの変数の間に強い相関関係(因果関係)があるように見える，見せかけの相関を**疑似相関**と言います。

例)「ポストの数」と「交通事故数」には，ちょっとした比例関係があります。常識的には疑わしいですが，実はどちらも「人口数」と深く密接な関係があるために疑似相関が起きます。

○疑似相関が疑われるとき，両方の変数と相関があるもう一つの変数を**交絡変数**と言います。

○疑似相関がある２つの変数 x，y に対して，交絡変数を z とする。

それぞれの相関係数 r_{xy}，r_{xz}，r_{yz} に対して，次の値を**偏相関係数**といいます。

［偏相関係数］　$r=\dfrac{r_{xy}-(r_{xz}\times r_{yz})}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}}$

○偏相関係数は，交絡変数 z の影響を取り除いた２つの変数 x，y の相関係数を意味します。

□□