2020年度秋学期 統計学　第10回

分布の推測とは ― 標本調査，度数分布と確率分布

浅野 晃
関西大学総合情報学部

「統計学」の後半は 統計的推測

「統計的推測」とは

ここまでは

データを度数分布で整理する
度数分布を要約する（平均・分散）

記述統計学

調べたいデータ全体を調べられるか？

「統計的推測」とは

調べたいデータ全体を調べられるか？
日本男性全員の身長を調べられるか？

データ全体の数値をすべて調べるのは，
費用や時間がかかる

最近はそうでもないのでは…（ぼそ）

その通りで，「ビッグデータ」という言葉もよく聞くようになりました。

「統計的推測」とは

調べたいデータ全体を調べられるか？
日本男性全員の身長を調べられるか？

データを集める手間は劇的に減ったけれど
測定作業の手間や費用は変わらない

それに, 調べると, 壊れてしまうものもある

料理をすべて味見してしまったら, 食べるものがなくなってしまう

「統計的推測」とは

調べたいデータ全体を調べられるか？
日本男性全員の身長を調べられるか？

データの一部を調べて度数分布を推測する

いや, せめて平均や分散を推測する

統計的推測

統計的推測の基本は
「くじびき」🎯

無作為抽出

統計的推測は,

集団のデータ全体を調べていないのに,
集団全体のようすを調べようとする

結果が間違っている可能性がある

バレーボール🏐やバスケットボール🏀の選手ばかり選んでしまったら

「日本人はすごく背が高い？」

無作為抽出

わざわざ背の高い人ばかり選ぶことはない

高低まんべんなく選べば,
その平均は集団の平均とだいたい同じ

それはそうだけど

集団にどんな人👽がいるか何も知らないのに
選ばれた人が, 集団のなかで背が高いか低いかなどわからない

無作為抽出

集団にどんな人がいるか何も知らないのに

選ばれた人が, 集団のなかで背が高いか低いかなどわからない

「まんべんなく選ぶ」のは無理

なので

公平なくじ引きで選ぶ

無作為抽出

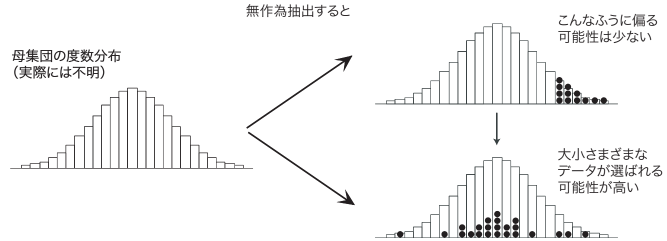
集団からくじびきで選ぶと

偶然, 🏐🏀選手のような人ばかりを
選んでしまって, おかしな結果になる可能性が
ないわけではないけれど,

そうなる確率は小さい　　その確率も計算できる。

無作為抽出

集団からくじびきで選ぶと



★たくさんの人を抽出すると, 偏らないか？
無作為抽出なら, そう期待できる。(今日の後半)
無作為抽出でなければ, 必ずしもそうではない。
(ツイッターのTLは「鏡に映った自分の意見」)

無作為抽出

データ全体から，いくつかの数値を
公平なくじびきで選ぶ

［無作為標本抽出］という

調べたい（が全部を調べるのは無理な）集団［母集団］

調べられる程度のデータ［標本（サンプル）］

標本「サイズ」

「母集団」や「標本」という言葉は，
「データ」と同様，数値の集まりをさす
（1つ1つの数値ではない）

母集団も標本も，その中に含まれる
数値の個数を大きさ（サイズ）という
（［標本サイズ］とはいうが，標本数とはいわない）

家族（family）という言葉に似ている

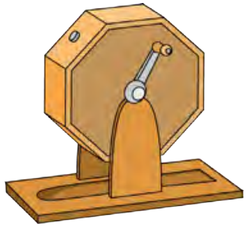
度数分布と確率分布

度数分布と確率分布

標本を無作為抽出するとき
ある数値が出てくる確率がどのくらいになるか

さっきの
「偏った数値ばかり選んでしまう」確率を求めるのにも必要

「公平なくじびき」と当たり確率

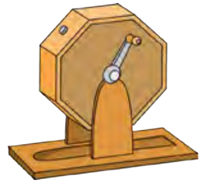


https://illpop.com/png_season/dec01_a07.htm

この中に入っている当たりくじの割合が
20%とする

くじを1回ひいて，当たる確率は？　20% 本当？

「公平なくじびき」と当たり確率

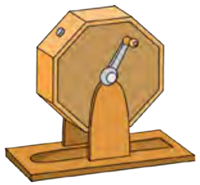


当たりくじの割合が20%なら，
当たる確率も20%

これが本当であるためには，

- どのくじも同じ確率で選ばれる
- ある回のくじびきの結果が，他の回に影響しない（独立）

「公平なくじびき」と当たり確率



どのくじも同じ確率で選ばれるのなら，

くじの総数のうち20%が当たり
→当たりが出る確率は20%

（ラプラスの確率の定義）

度数分布で考えると

母集団の度数分布

階級値	相対度数
…	
162.5	15%
167.5	20%
172.5	20%
177.5	10%
…	

無作為抽出 →

階級値172.5の人が
選ばれる確率は

20%

度数分布で考えると

どの階級についても同じだから

母集団の度数分布

階級値	相対度数
162.5	15%
167.5	20%
172.5	20%
177.5	10%

無作為抽出 →

標本の[確率分布]

階級値	選ばれる確率
162.5	15%
167.5	20%
172.5	20%
177.5	10%

確率分布と確率変数

つまり 母集団の度数分布（母集団分布） = 標本の確率分布

階級値	選ばれる確率
162.5	15%
167.5	20%
172.5	20%
177.5	10%

標本の値がいくらになるかは決まっていない

しかし確率分布が決まっている

標本は[確率変数]である，という（中国語では随機変数）

何が知りたいのか

母集団の度数分布が知りたい

標本の確率分布を推定すればよい

標本の確率分布，推定できる？

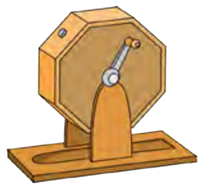
くじを1本だけひいても，当たり確率はわからない

どうする？

復元抽出と非復元抽出

・どのくじも同じ確率で選ばれる
・ある回のくじびきの結果が，他の回に影響しない（独立）

こうであるためには，



出たくじをすぐに箱に戻す
[復元抽出]

実際にはあまりやらない…

標本平均と母平均

母平均の推定

母集団
（日本男性全体）

母平均μ

標本として数値をいくつか取り出して，
それらの平均［標本平均］

標本平均は母平均に近い値になるか？

母平均が知りたい が，日本男性全員は調べられない

母平均の推定

母集団
（日本男性全体）

母平均μ

［標本平均］

標本平均は母平均に近い値になるか？

もし偏った標本が得られていたら，
標本平均は母平均と大きく食い違うことに

母平均の推定

母集団

母平均μ
母分散σ^2

サイズnの標本1セット

X1 X2 … Xn

標本平均

$\bar{X}$

仮に，何度も標本を抽出したとしたら？

X1の期待値と分散は

サイズnの標本1セット　標本平均

母集団

X1 X2 … Xn　$\bar{X}$

母平均μ
母分散σ^2

X1 X2 … Xn　$\bar{X}$

X1 X2 … Xn　$\bar{X}$

⋮　⋮

X1のさまざまな可能性
その平均を，

期待値？

[期待値] μ
分散σ^2

期待値とは？

期待値は平均の一種で「すべての可能性にわたっての平均」

サイズnの標本1セット　標本平均

母集団

X1 X2 … Xn　$\bar{X}$

母平均μ
母分散σ^2

X1
X1
⋮

X1の期待値＝
X1のすべての可能性にわたっての平均

母集団のすべての数値を取り出すのと同じだから，
母平均μと同じ

X1の分散も
母分散σ^2と同じ

標本平均の期待値と分散は

サイズnの標本1セット　標本平均

母集団

X1 X2 … Xn　$\bar{X}$

~~172 195 153~~

母平均μ
母分散σ^2

X1からXnのなかに極端な数値があっても，

n個の数値を平均すれば，
そんなに極端な数値にはまずならない

標本平均の期待値と分散は

サイズnの標本1セット　標本平均

母集団

X1 X2 … Xn　$\bar{X}$

母平均μ
母分散σ^2

X1 X2 … Xn　$\bar{X}$

X1 X2 … Xn　$\bar{X}$

⋮　⋮

極端な値はあまりないので
分散が小さくなる

母集団と同じ
期待値μ
分散σ^2

期待値μ
分散 σ^2/n

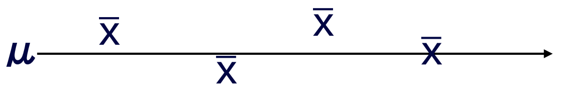
標本平均の分散は，母分散の「標本サイズ分の一」になる

母平均の推定

母平均がμ
母分散がσ^2
のとき，
標本平均の期待値がμ
標本平均の分散がσ^2/n

仮に，何度も標本を抽出して，
何度も標本平均を計算したとすると

分散が小さくなっているので，
「たいてい，ほぼ」母平均に近い値になる



標本平均を
何度も計算しても

いつ計算してもそれほどかわらない

母平均の推定

母平均がμ
母分散がσ^2
のとき，
標本平均の期待値がμ
標本平均の分散がσ^2/n

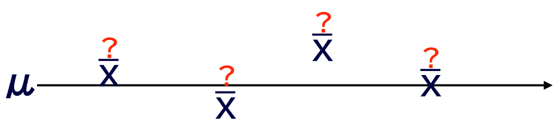
仮に，何度も標本を抽出して，
何度も標本平均を計算したとすると

分散が小さくなっているので，
「たいてい，ほぼ」母平均に近い値になる

いま1回だけ計算した標本平均も，
たいてい，ほぼ母平均に近い値だろう

母平均の推定

いま1回だけ計算した標本平均も，
「たいてい，ほぼ」母平均に近い値だろう



何度も計算しても

いつ計算してもそれほどかわらない

いま1回だけ計算した標本平均は
上のどれにあたるかはわからないが，
いずれにせよあまりかわらない

ところで

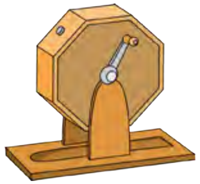
母分散がσ^2のとき，標本平均の分散がσ^2/n

標本平均の分散に関係しているのは
標本の大きさであって，母集団の大きさは関係ない

推測の確かさに影響するのは
標本の大きさであって，
標本の大きさの，母集団の大きさに対する割合 ではない

標本の大きさとは

「10人からなる標本」の意味は，
1,000人からなる母集団でも
100,000人からなる母集団でも同じ 🤔…



理想的な無作為抽出であれば，標本サイズは，
「数値の個数」というよりも
「同一の母集団から数値ひとつひとつを
取り出す回数」

母平均の推定

いま1回だけ計算した標本平均も，
「たいてい，ほぼ」母平均に近い値だろう

どのくらい近い？

どのくらいの確率で？
はずれる確率は？

このあたりは次回。