2020年度春学期　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

A. Asano, Kansai Univ.

A. Asano, Kansai Univ.

「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？

A. Asano, Kansai Univ.

「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。

A. Asano, Kansai Univ.

「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

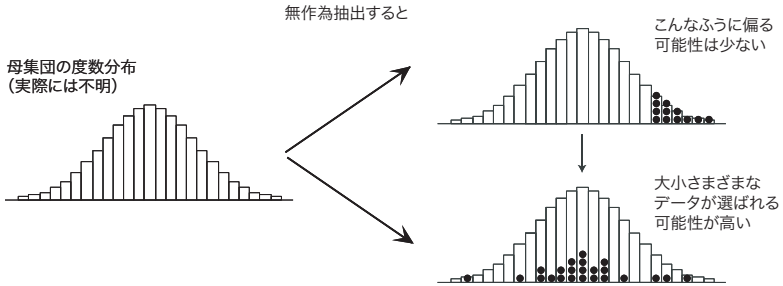
選んでしまって，おかしな結果になる可能性がないわけではないけれど，

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

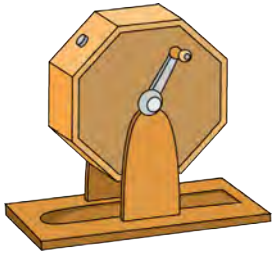
度数分布と確率分布

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

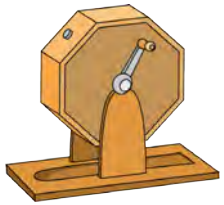


この中に入っている
当たりくじの割合が
20%とする

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

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

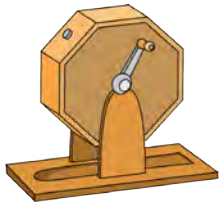


当たりくじの割合が
20%なら，
当たる確率も
20%

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？

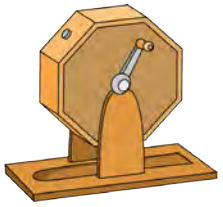
くじを1本だけひいても，当たり確率はわからない

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，



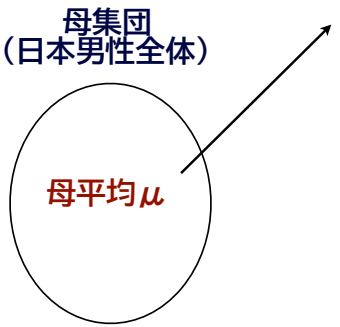
出たくじをすぐに箱に戻す[復元抽出]

実際にはあまりやらない…

標本平均と母平均

A. Asano, Kansai Univ.

母平均の推定



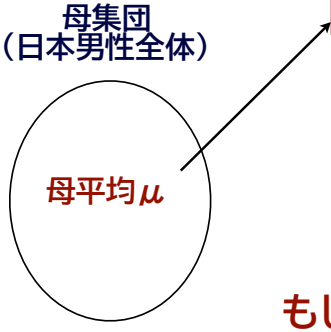
標本として数値を
いくつか取り出して，
それらの平均
[標本平均]

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

A. Asano, Kansai Univ.

母平均の推定



[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

A. Asano, Kansai Univ.

母平均の推定

母集団

サイズnの標本1セット

X1 X2 … Xn

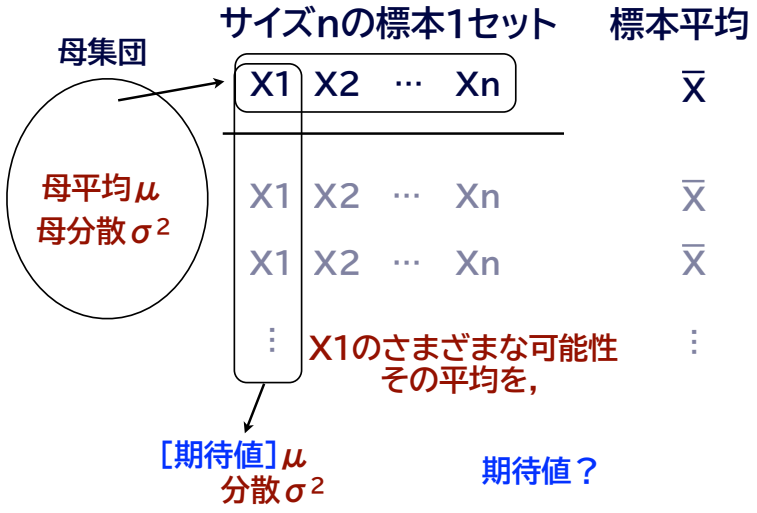
標本平均

$\bar{X}$

仮に，何度も標本を
抽出したとしたら？

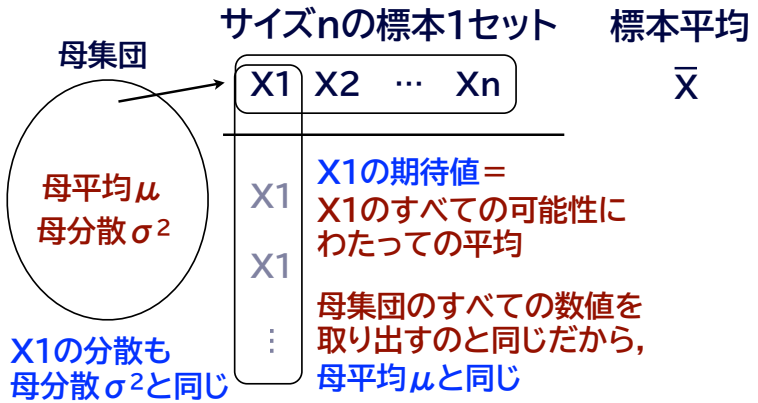
A. Asano, Kansai Univ.

X1の期待値と分散は

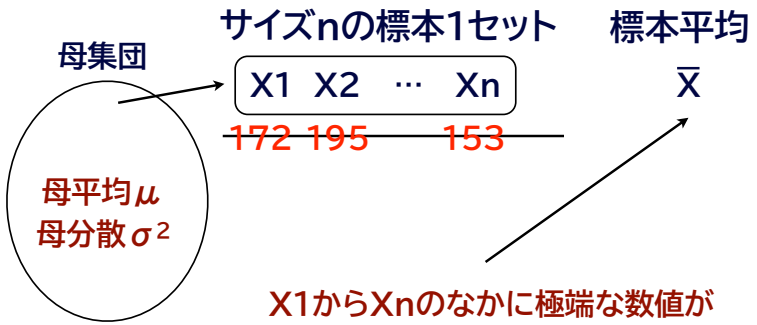


期待値とは？

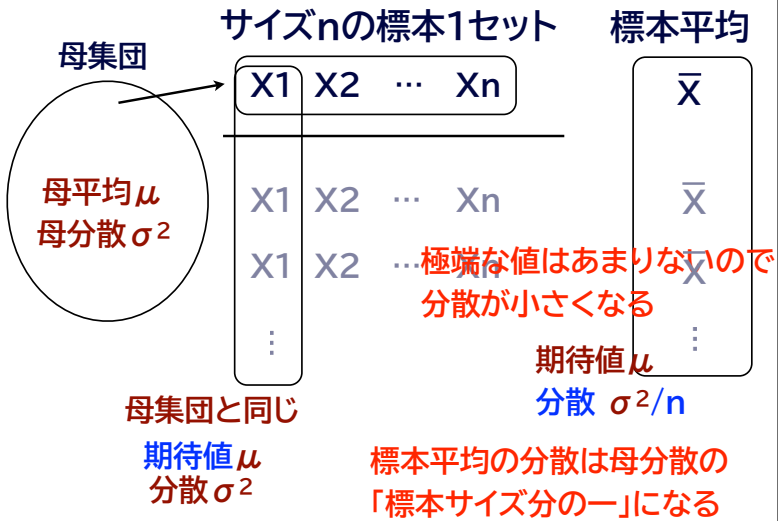
平均の一種
「すべての可能性にわたっての平均」



標本平均の期待値と分散は



標本平均の期待値と分散は

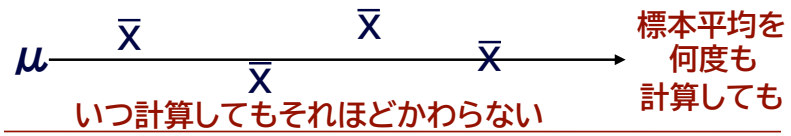


母平均の推定

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



A. Asano, Kansai Univ.

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

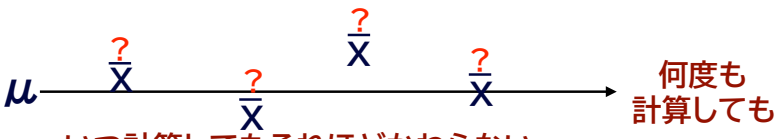
分散が小さくなっているので，
「たいてい，ほぼ」母平均に近い値になる

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

A. Asano, Kansai Univ.

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



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

A. Asano, Kansai Univ.

ところで

母分散がσ^2のとき，標本平均の分散がσ^2/n

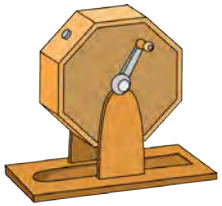
標本平均の分散に関係しているのは
標本の大きさであって，
母集団の大きさは関係ない

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

A. Asano, Kansai Univ.

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
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。