

分散分析で F 分布を利用する根拠

1元配置分散分析の手続き

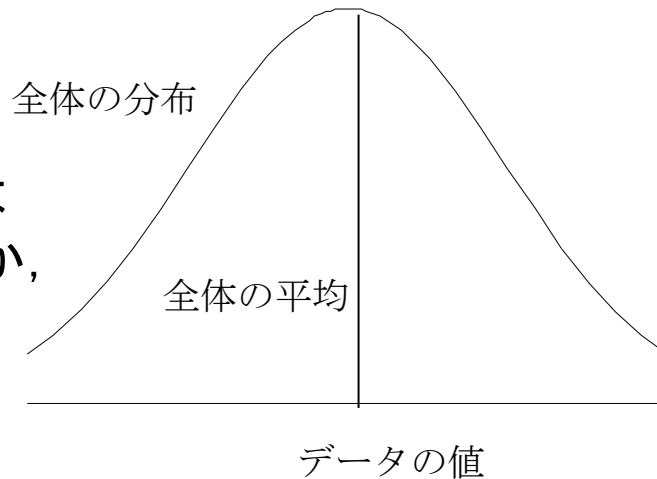
- ある1つの変数(従属変数)について、複数の群の母平均値が共通で等しいと見なせるかどうかを検定する方法.
- 例えば, ある学力テストの平均点を, 普通の公立中学, 中高一貫の公立中学, 私立の中学で比較.
- 「各群の母平均値は共通で等しい」という帰無仮説を設定.
- データから F 統計量というものの値を計算して, 統計的有意性を判断.
- F 統計量の値が基準値を超えれば, 統計的に有意であるとして, 帰無仮説を棄却.

本講の内容

- 平均値の違いを検討するのに F 分布(F 統計量)を用いる根拠を, 1元配置分散分析を例に取り解説.
- キーワード・キートピック
 - データの分布
 - 平方和の分割
 - 全体平方和, 群間平方和, 群内平方和
 - 平均値の違いと平方和の大きさとの関係
 - F 統計量の構成
- 前提 「各群のデータは分散が等質な正規分布からの無作為標本である」

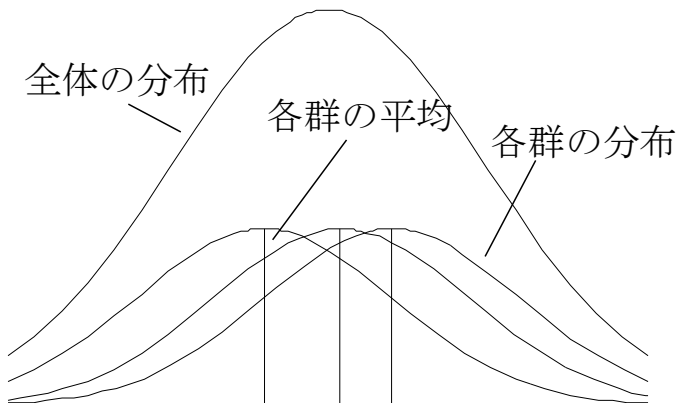
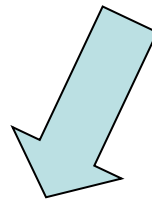
データの分布

群ごとの分布は
どうなっているか、
考えよう。

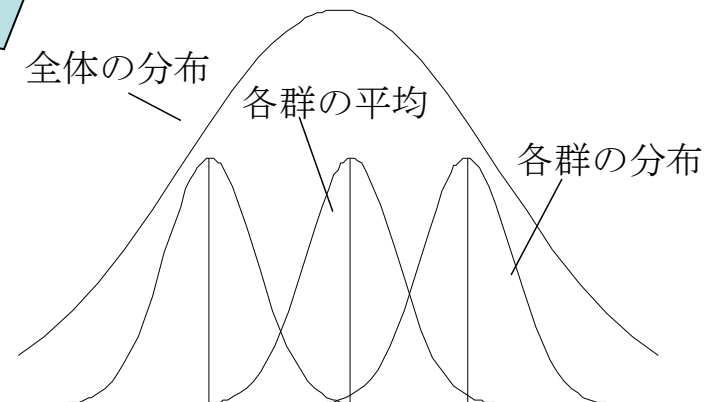
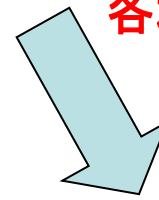


全体のデータの散らばり、
各群の標本平均の散らばり、
各群におけるデータの散ら
ばりの関係を見るため、
それらを分解してみよう。

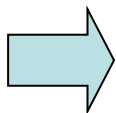
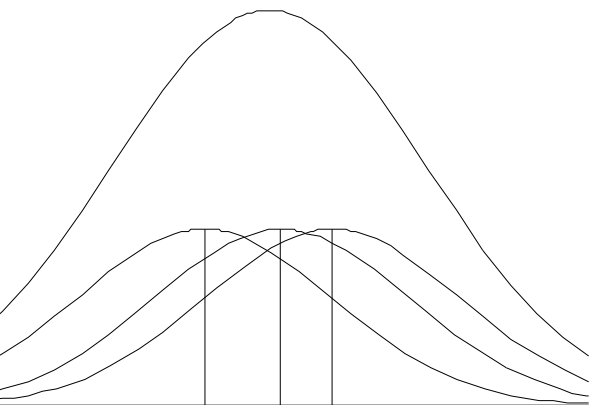
各群の標本平均が近い場合



各群の標本平均が離れている場合



各群の標本平均が近い場合 散らばりの分解



全体のデータの散らばり

各群の標本平均値の散らばり

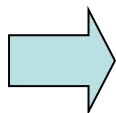
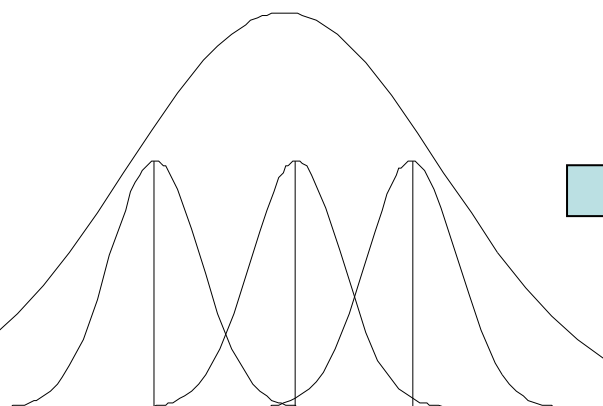
各群におけるデータの散らばり

||

^

∨ ?

統計学的に確認してみよう！



各群の標本平均が離れている場合

分散・平方和

- 散らばりを表す統計量(のひとつ)は分散

$$\text{標本分散} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

x_i : 被験者 i のデータ

$\bar{x}$: 標本平均

n : 被験者数

- $n \times$ 標本分散 = **平方和** も散らばりの程度を表す.

$$\text{平方和} = n \times \text{標本分散}$$

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

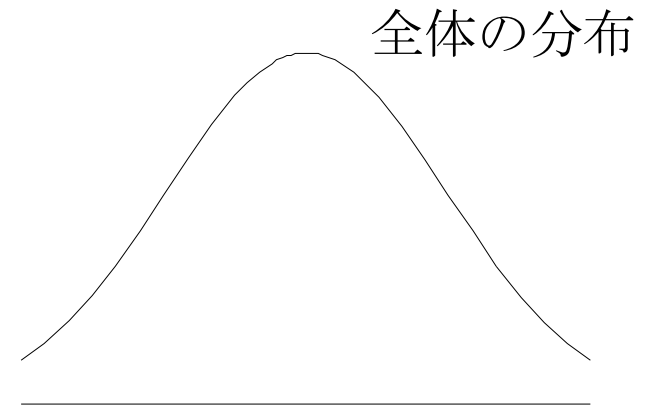
$$= \text{一辺の長さが } \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ の正方形の面積}$$

各平方和の名称

- 全体のデータの散らばり

→ 全体平方和

(Total Sum of Squares, SST)



$$\sum_{j=1}^a \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

x_{ij} : 第 j 群の被験者 i のデータ

$\bar{x}$: 全体の標本平均

$\bar{x}_j$: 第 j 群の標本平均

a : 群の数

n_j : 第 j 群の被験者数

各平方和の名称

- 各群の標本平均の散らばり

→ 群間平方和

(Between Sum of Squares, SSB)

各群の標本平均
の散らばり



$$\sum_{j=1}^a \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2$$
$$\left(= \sum_{j=1}^a n_j (\bar{x}_j - \bar{x})^2 \right)$$

x_{ij} : 第 j 群の被験者 i のデータ

$\bar{x}$: 全体の標本平均

$\bar{x}_j$: 第 j 群の標本平均

a : 群の数

n_j : 第 j 群の被験者数

各平方和の名称

- 各群におけるデータの散らばり

→ 群内平方和

(Within Sum of Squares, SS_w)

$$\sum_{j=1}^a \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

x_{ij} : 第 j 群の被験者 i のデータ

$\bar{x}$: 全体の標本平均

$\bar{x}_j$: 第 j 群の標本平均

a : 群の数

n_j : 第 j 群の被験者数

各群におけるデータの分布



平方和間の関係

全体平方和 (SS_T)

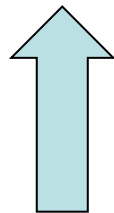
$$\sum_{j=1}^a \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

群間平方和 (SS_B)

$$\sum_{j=1}^a \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 \\ \left(= \sum_{j=1}^a n_j (\bar{x}_j - \bar{x})^2 \right)$$

群内平方和 (SS_W)

$$\sum_{j=1}^a \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$



平方和の分割

$$\begin{aligned}
SS_T &= \sum_{j=1}^a \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 \\
&= \sum_{j=1}^a \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j + \bar{x}_j - \bar{x})^2 \\
&= \sum_{j=1}^a \sum_{i=1}^{n_j} [(x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})]^2 \\
&= \sum_{j=1}^a \sum_{i=1}^{n_j} [(x_{ij} - \bar{x}_j)^2 + (\bar{x}_j - \bar{x})^2 \\
&\quad + 2(x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{x})] \\
&= \sum_{j=1}^a \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{j=1}^a \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 \\
&\quad + \sum_{j=1}^a 2(\bar{x}_j - \bar{x}) \underbrace{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)}_{\substack{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j) \\ = \sum_{i=1}^{n_j} x_{ij} - \sum_{i=1}^{n_j} \bar{x}_j \\ = \sum_{i=1}^{n_j} x_{ij} - n_j \bar{x}_j \\ = \sum_{i=1}^{n_j} x_{ij} - \sum_{i=1}^{n_j} x_{ij} \\ = 0}} \\
&= SS_W + SS_B + 0 \\
&= SS_B + SS_W
\end{aligned}$$

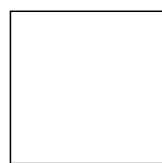
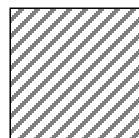
平方和間の関係

全体平方和 (SS_T) = 群間平方和 (SS_B) + 群内平方和 (SS_W)

$$\sum_{j=1}^a \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 = \sum_{j=1}^a \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 + \sum_{j=1}^a \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

- 全体平方和は、群間平方和と群内平方和に分解される。それですべて。残りなし。
- 一般的に、群間平方和も群内平方和も、正の値である。
- 全体平方和の大きさが変わらないとき、一方が大きくなれば、他方は小さくなるという関係。

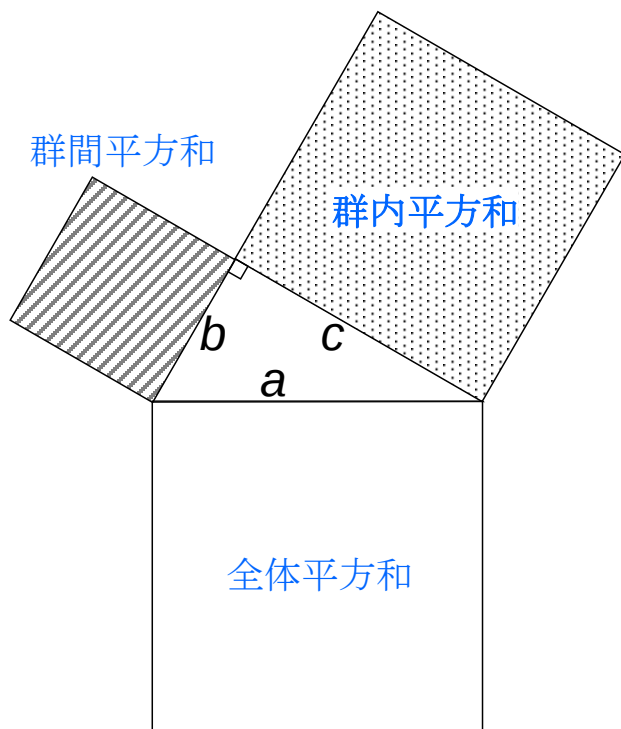
$$\text{全体平方和 (SS}_T\text{)} = \text{群間平方和 (SS}_B\text{)} + \text{群内平方和 (SS}_W\text{)}$$


 a^2
 $=$

 b^2
 $+$

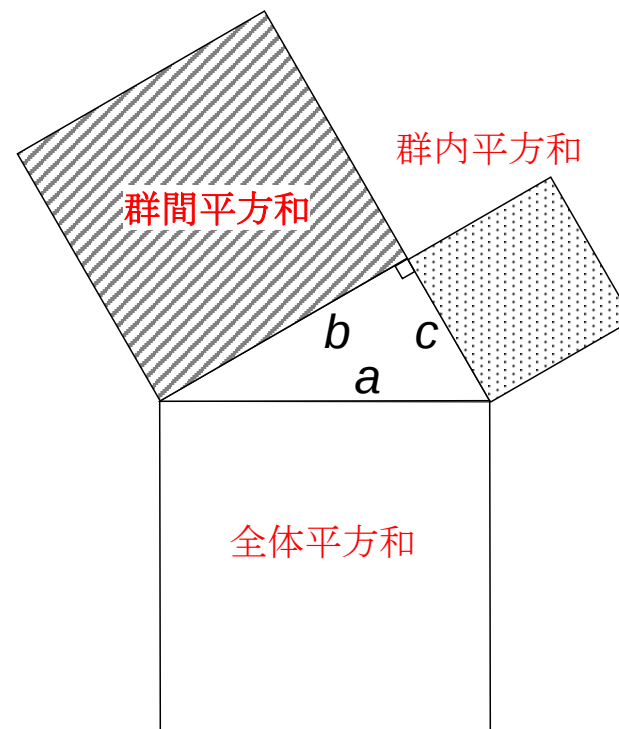
 $+$
 c^2

三平方の定理！

各群の標本平均が近い場合



各群の標本平均が離れている場合



Q.ある全体の分布(散らばり)が観測されているとき、
各群の母平均値が共通で等しいかどうかを
判断するのに利用できる指標はなにか？

- 各群の母平均値が共通で等しいと、
→各群の標本平均は近い値になるので、
群間平方和は小さく、群内平方和は大きくなる.
- 各群の母平均値が離れているほど、
→各群の標本平均も離れた値になるので、
群間平方和は大きく、群内平方和は小さくなる.

A. 群間平方和と群内平方和を利用できる.

具体的には,

- 群間平方和 小 群内平方和 大
→ 母平均は共通の値であると判断.
 - 群間平方和 大 群内平方和 小
→ 母平均は共通の値ではないと判断.
- 平均値の違いの検定には, この判断法を利用できるような統計量を用いればよい.

平均値の違いの検定に用いる 統計量に求められること

- まず、帰無仮説を「各群の母平均値は共通で等しい」として、これが真であると仮定する. $H_0: \mu_1 = \mu_2 = \dots = \mu_a = \mu$
- 群間平方和が大きいほど、群内平方和が小さいほど、値が大きくなるような統計量であって、
- 帰無仮説が真のとき、より大きな値となる確率が小さい統計量であれば、
- その統計量の値がある基準値よりも大きいとき、統計的に有意性であるとして、帰無仮説を棄却(各群の母平均は共通の値ではないと判断)することができる.
- 値が大きい方向だけを考えるので、片側検定である.

F統計量の構成

- 群間平方和が大きいほど、群内平方和が小さいほど、
値が大きくなるような統計量の構成の候補。
 - 群間平方和 — 群内平方和
 - 群間平方和 / 群内平方和
- 1元配置分散分析における F 統計量は後者を含むかたちになっている。

$$F = \frac{\text{群間平方和} / (\text{群数} - 1)}{\text{群内平方和} / (\text{被験者数} - \text{群数})} = \frac{\text{群間平均平方}}{\text{群内平均平方}}$$

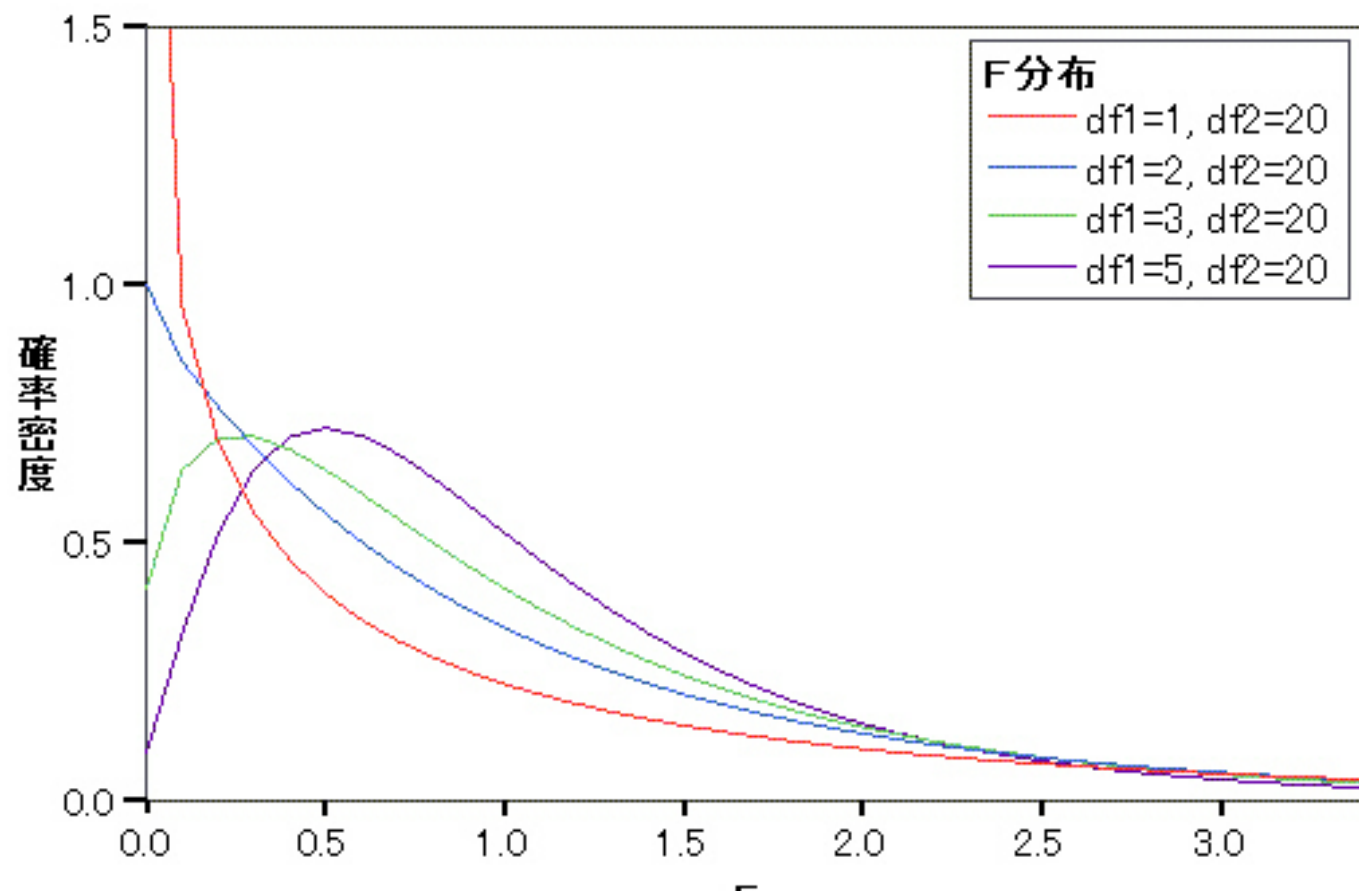
(ただし、各群のデータは、分散が等質な正規分布からの
無作為標本であることが必要)

F 分布

- F 統計量の分布. ($0 < F$)
- より大きな F 統計量の値が生起する確率は小さい.
- 2つの「自由度」の値で分布の形が決まる.
- 1元配置の場合の F 統計量の自由度は,
群間平方和の自由度「群数 - 1」 と
群内平方和の自由度「被験者数 - 群数」
の2つ.

$$F = \frac{\frac{SS_B/\sigma^2}{a-1}}{\frac{SS_W/\sigma^2}{N-a}} = \frac{SS_B/(a-1)}{SS_W/(N-a)} \quad N : \text{全被験者数}$$

F 分布



平均値の違いを検討するのに F 分布 (F 統計量) を用いる根拠

- F 統計量は、群間平方和が大きいほど、群内平方和が小さいほど、値が大きくなる統計量である。
- すなわち、 F 統計量は、標本平均の散らばりが大きいほど値が大きくなる統計量である。
- F 統計量は、帰無仮説「各群の母平均値は共通で等しい」が真のとき、より大きな値となる確率が小さい統計量である。
(← 帰無仮説が真のとき、標本平均の散らばりはそれほど大きくはない)
- よって、 F 統計量の値がある基準値を超えれば帰無仮説を棄却するという統計的判断ができる。
- それゆえ、平均値の違いを検討するのに、 F 分布 (F 統計量) を用いることができる。