統計学の基礎理論②
多重検定問題と分散分析

東京科学大学　ILA国府台
中林　潤

JH人材育成課　バイオインフォマティシャン育成講座　③

内容

- 多重検定問題
- p値の補正方法
- 分散分析

多重検定

- 検定を複数回繰り返すこと。

- 第1種の過誤が発生する。

		事実	
		帰無仮説は正しい	帰無仮説は誤り
検定	帰無仮説を棄却する	False Positive (FP)	True Positive (TP)
	帰無仮説を採用する	True Negative (TN)	False Negative (FN)

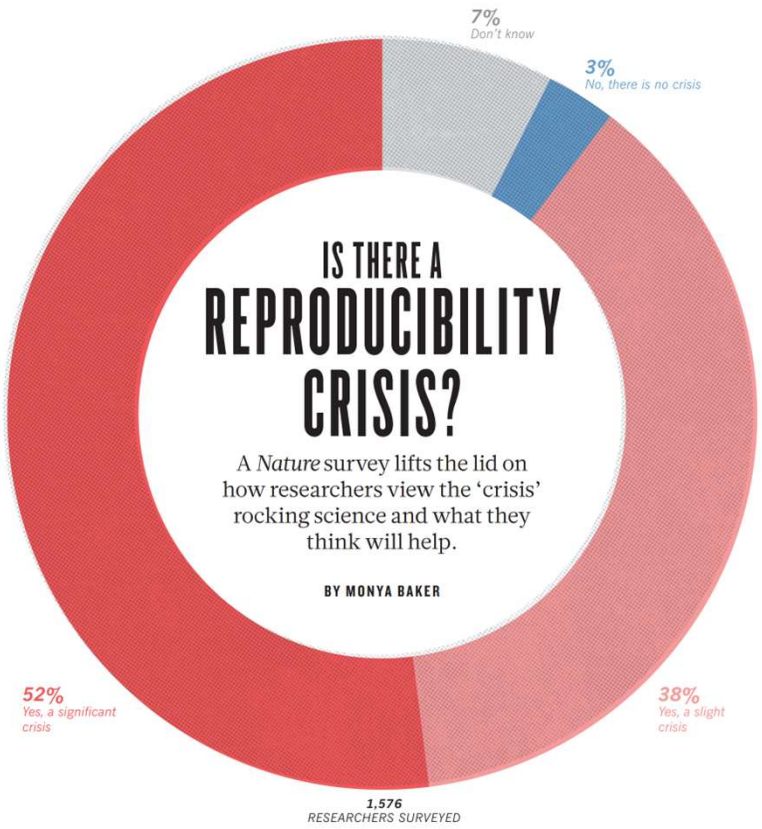
有意水準5%の検定を20回繰り返す→1回は有意差ありと判定される。

再現性とQRPs

- 再現性の問題
  ある研究者によって統計的に有意の差があるとされた結果が、他の研究者には同程度の有意の差と判定されないこと。

- QRPs（Questionable Research Practice）
  正しい結論にたどり着く確率を下げるような研究手法

再現性の危機



Natureが実施したネットワークアンケート調査

- 1576人の科学者が回答
- 52% significant crisis
- 38% slight crisis
- 3% no crisis
- 7% don't know

Nature 2016 May Vol.533

ある研究者が統計的に有意と判定した結果が、他の研究者には同程度の有意と判定されない

バイオインフォマティクスと多重検定問題

- バイオインフォマティクスでは大容量のデータを扱うことが多い。
- 多重検定になりがち。
- 有意水準の補正が必要なケースが生じる。

有意水準の補正方法

- Bonferroni補正
- Benjamini-Hochberg補正（BH法）

Bonferroni補正

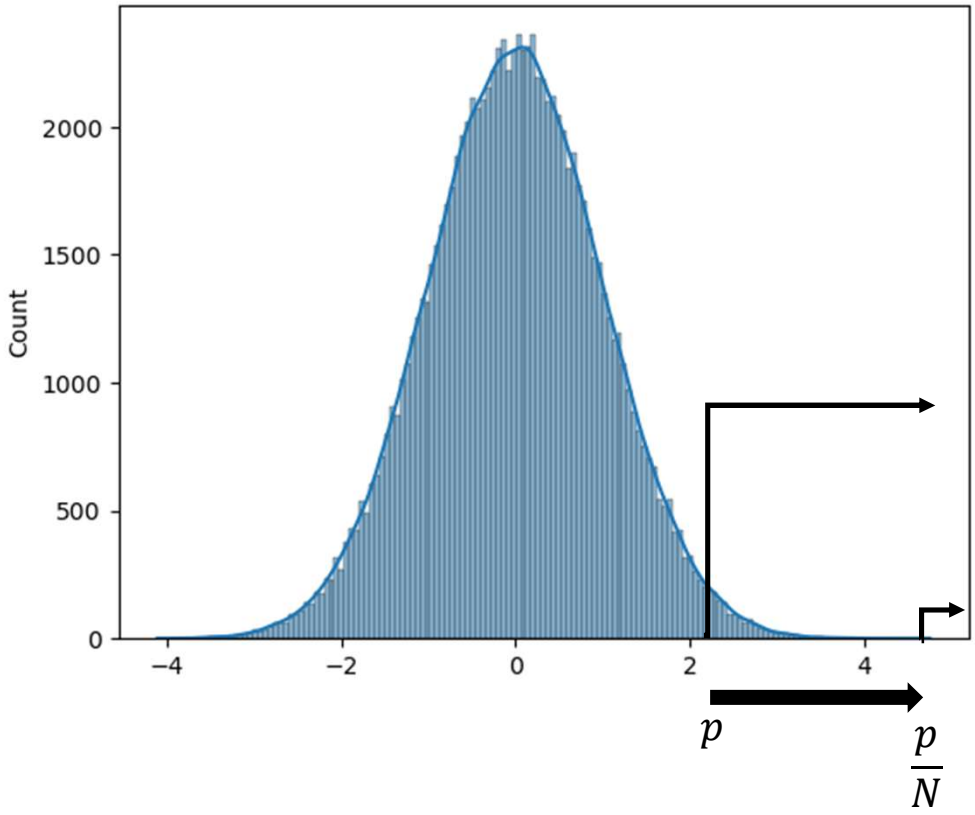
- 有意水準を検定回数で割る。

  $P = \frac{p}{N}$

- 簡便に実行できる。

- 検出力が落ちる。

- 保守的な補正方法



Benjamini-Hochberg補正（BH法）

- False Positiveを許容する考え方
- False Discovery Rate（FDR）：棄却した帰無仮説の中の正しい帰無仮説の割合

$$FDR = \frac{FP}{TP+FP}$$

事実

検定		帰無仮説は正しい	帰無仮説は誤り
	帰無仮説を棄却する	False Positive（FP）	True Positive（TP）
	帰無仮説を採用する	True Negative（TN）	False Negative（FN）

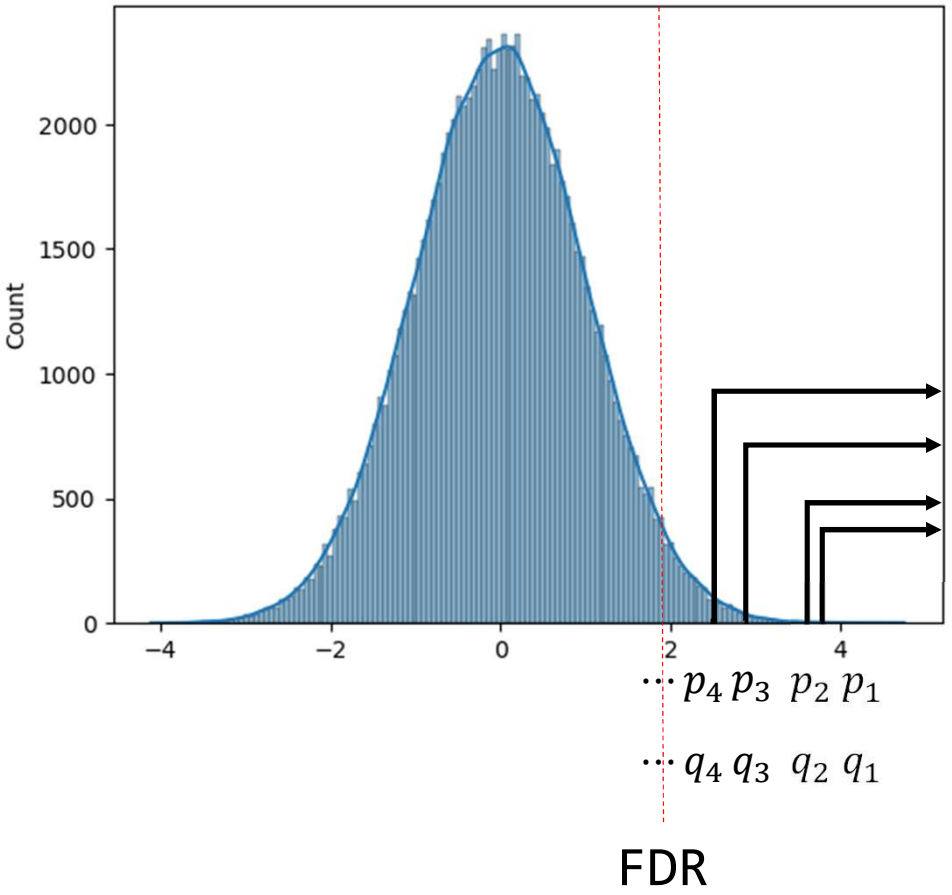
- FDRをある値以下に抑える。

Benjamini-Hochberg補正（BH法）

- N回検定を繰り返す。
- P値を小さい順に並べる。$(p_1, p_2, \cdots, p_N)$
- Q値を計算する。

$$q_i = \frac{Np_i}{i}$$

- $q_i < FDR$閾値となるところまでを有意とみなす。
- バイオインフォマティクスではむしろQ値をよく使う



多重検定問題

- 検定を複数回繰り返すことで、第1種の過誤が生じる可能性が増加する。

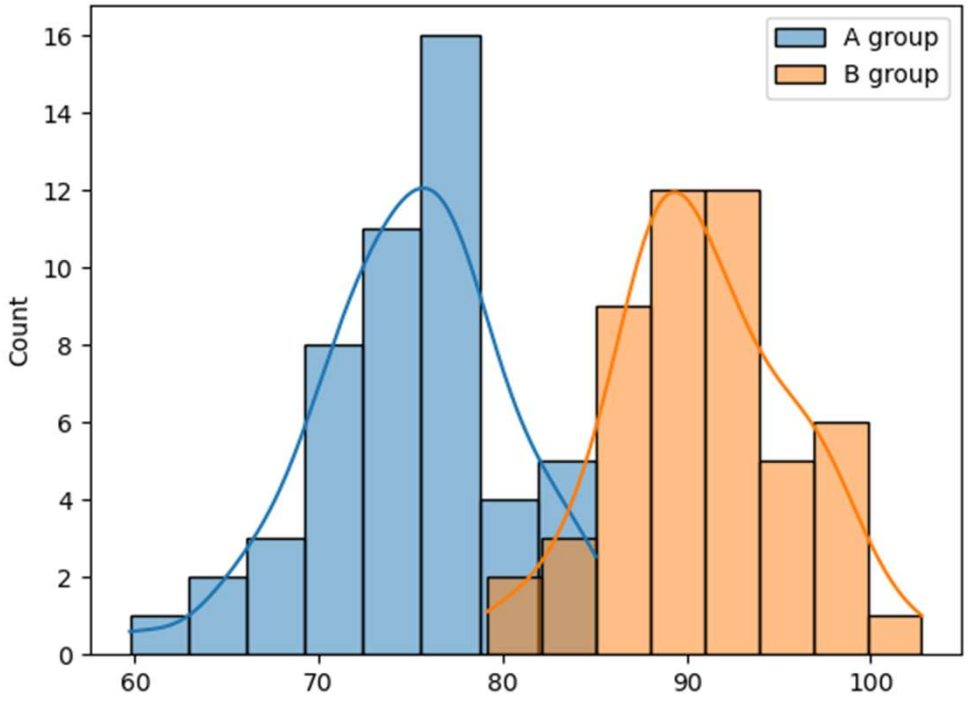
- 有意水準を適切に補正する必要がある。
  Bonferroni補正、Benjamini-Hochberg補正
  バイオインフォマティクスではQ値を用いることが多い

- 見逃されやすい多重検定
  データの追加、複数種類の統計解析

- 不必要な検定は避けるべき。

グループ間比較

- 2群間で平均値を比較する。
  t検定



帰無仮説：A群、B群は平均の等しい母集団から得られた
対立仮説：　　異なる
検定統計量

$$T = \frac{(\overline{x_1} - \overline{x_2})}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

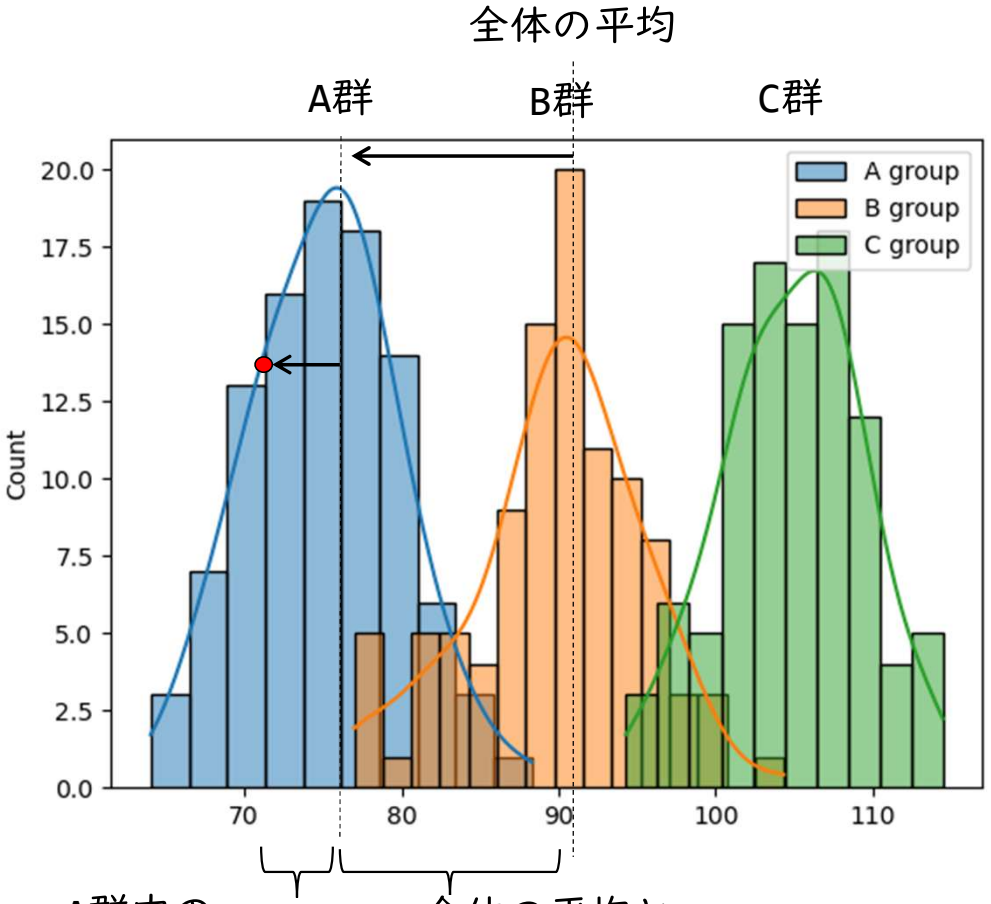
$$df = N - 1$$

Tは自由度dfのt分布に従う。
A群の平均−B群の平均=0が得られる確率はp以下
↓
pが十分小さければ、帰無仮説を棄却し二群間に差があると判定する。

3群以上のグループ間比較

- 分散分析（Analysis of Variance ANOVA）
  帰無仮説：グループ間に平均値の差は無い。
  対立仮説：少なくとも一つのグループ間で平均値に差がある。
  どのグループ間に差があるのかは解らない。
  2群間比較を繰り返してはいけない（多重検定）。
  分散を使って平均値の差を検定するものである。

分散分析の考え方



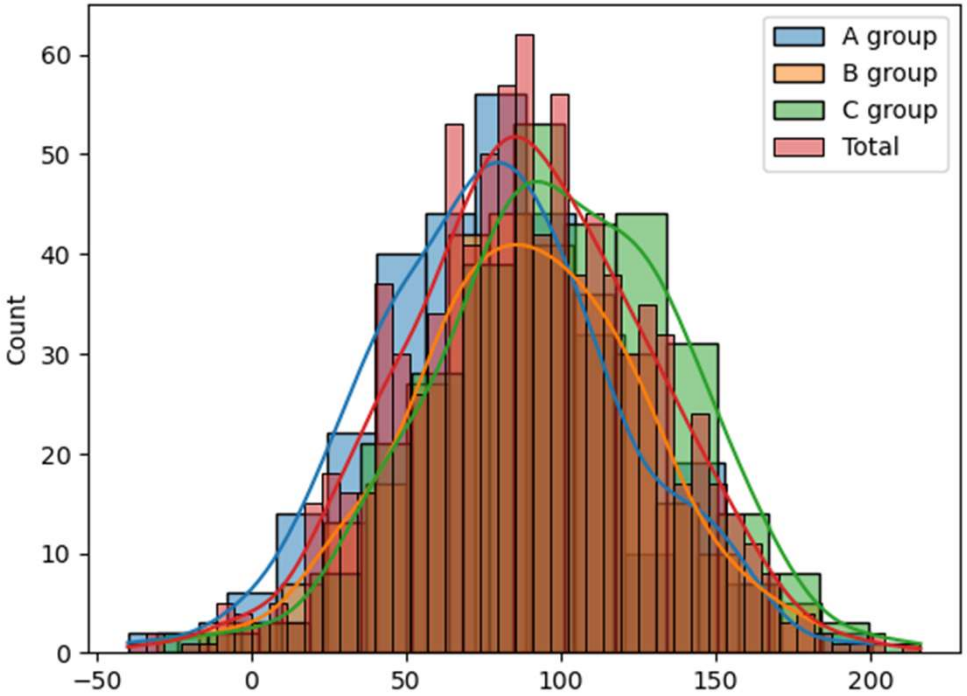
データ値の分散はグループ間の分散と
グループ内の分散の和

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

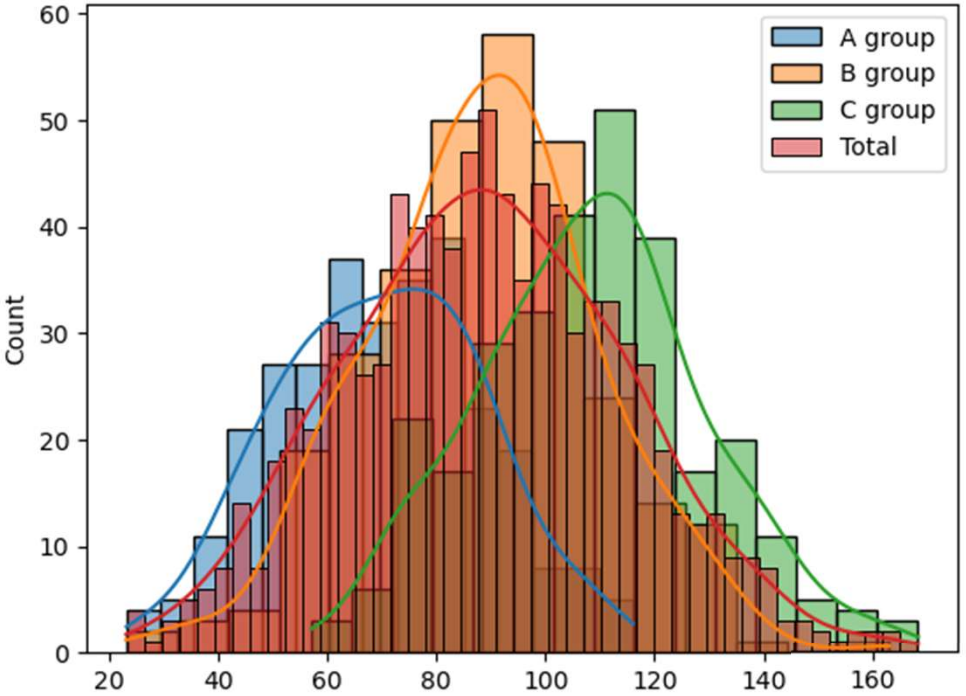
$$= \sum_{j=1}^{a} \sum_{i=1}^{n_j} \left(\bar{x_j} - \bar{x}\right)^2 + \sum_{j=1}^{a} \sum_{i=1}^{n_j} \left(x_i - \bar{x_j}\right)^2$$

$$= S_B + S_W$$

ばらつきの関係



グループ間のばらつき＜グループ内のばらつき



グループ間のばらつき＞グループ内のばらつき

分散分析表

	平方和	自由度	平均平方
群間	$\sum_{j=1}^{a} n_j\left(\bar{x_j}-\bar{x}\right)^2$	$a-1$	$\frac{\sum\left(x_j-\bar{x}\right)^2}{a-1}$
群内	$\sum_{j=1}^{a}\sum_{i=1}^{n_j}\left(x_i-\bar{x_j}\right)^2$	$\sum\left(n_j-1\right)=N-a$	$\frac{\sum\left(x_i-\bar{x_j}\right)^2}{N-a}$
全体	$\sum_{i=1}^{N}(x_i-\bar{x})^2$	$N-1$	$\frac{\sum(x_i-\bar{x})^2}{N-1}$

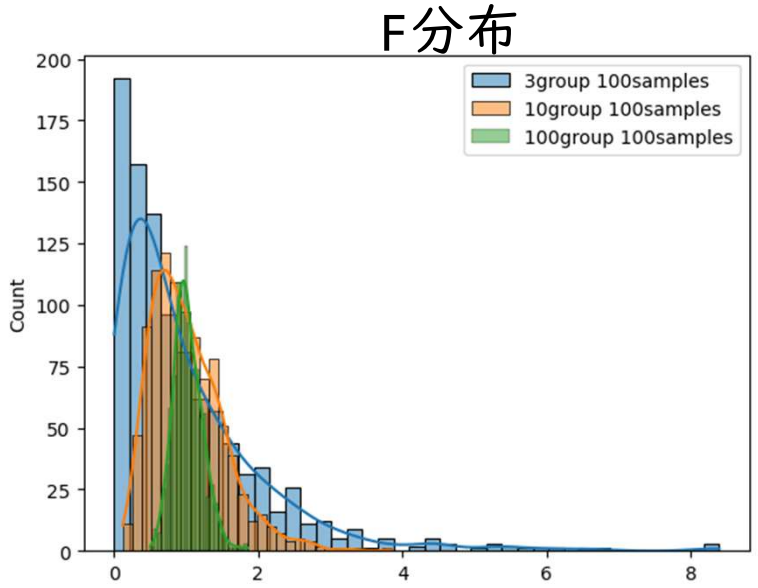
$\bar{x_j}$：各グループの平均値
$\bar{x}$：全体の平均値
a：グループ数
n_j：各グループのサンプル数
N：総サンプル数

F値

- F値＝群間の平均平方/群内の平均平方＝$\dfrac{\frac{\sum_{j=1}^{a} n_j\left(\overline{x_j}-\overline{x}\right)^2}{a-1}}{\frac{\sum_{j=1}^{a}\sum_{i=1}^{n_j}\left(x_i-\overline{x}_j\right)^2}{N-a}}$

- F値が大きいほどグループ間の平均に差がある可能性が高い

- F分布に従う。



$F(m,n)=\dfrac{\frac{\chi^2}{m}}{\frac{\chi^2}{n}}$

χ^2はカイ二乗分布

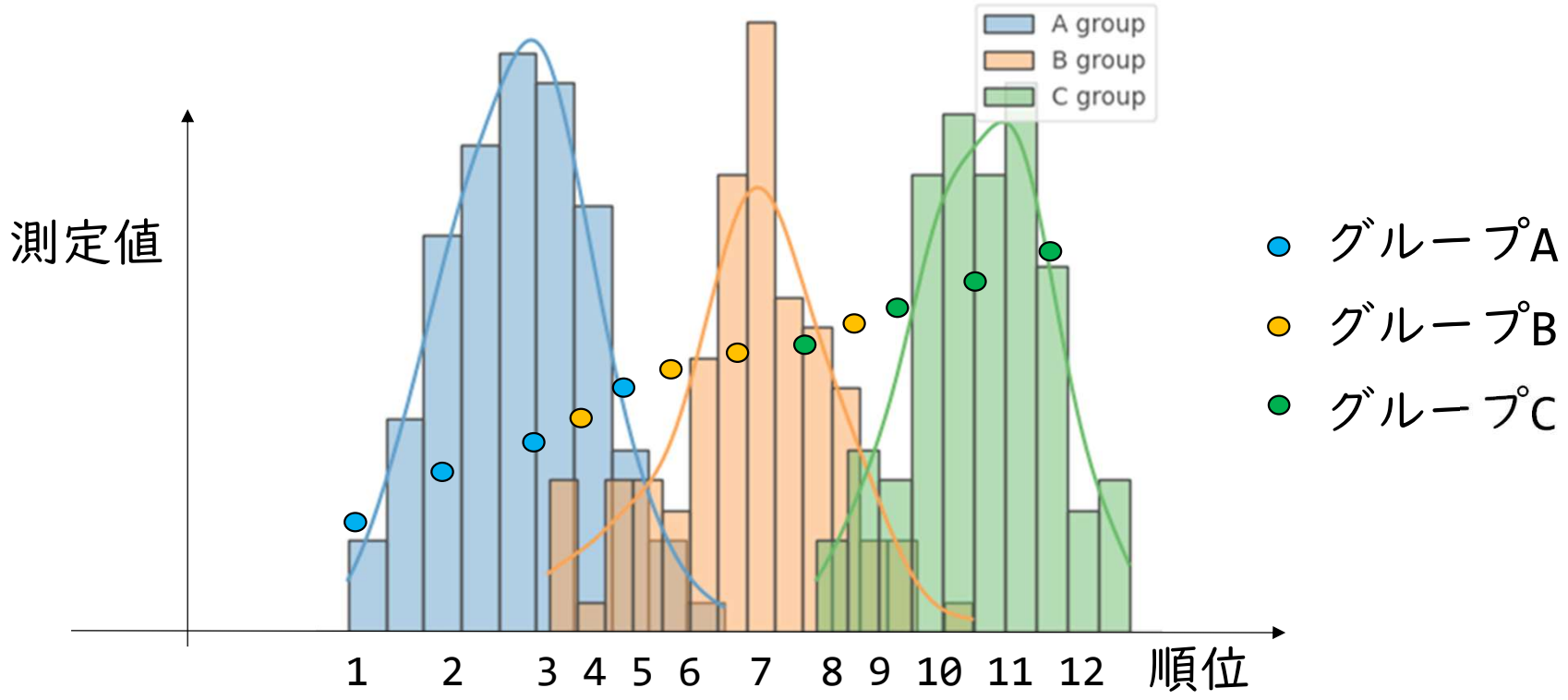
m,nは自由度

分散分析（ANOVA）

- 帰無仮説：グループ間の平均値に差がない。
- 対立仮説：少なくとも一つの組み合わせでグループ間の平均値に差がある。
- 検定統計量：F値＝群間の平均平方/郡内の平均平方＝$\frac{\frac{\sum_{j=1}^{a} n_j\left(\overline{x_j}-\overline{x}\right)^2}{a-1}}{\frac{\sum_{j=1}^{a}\sum_{i=1}^{n_j}\left(x_i-\overline{x}_j\right)^2}{N-a}}$

Kruskal-Wallis検定

- ノンパラメトリックな検定方法
- 測定値を小さい順に並べて、順位の偏りを調べる



Kruskal-Wallis検定

- 検定統計量

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{j} \frac{{R_i}^2}{n_i} - 3(N+1)$$

N：総サンプル数

n_i：グループiのサンプル数

R_i：グループiの順位和

j：グループ数

- Hは自由度$j-1$のχ二乗分布に従う。

まとめ

- 多重検定問題
  検定を複数回繰り返すこと。
  第1種の過誤が生じる可能性が増加する。
  不必要な検定は避けるべき。

- 有意水準の補正
  Bonferroni補正
  Benjamini-Hochberg補正

- 分散分析
  3群以上の平均値を比較する。
  得られ結論は少なくとも一つの組み合わせでグループ間の平均値に差がある。