原稿作成日：　2017 年 3 月 31 日

# 多重性の問題：研究計画の重要性

〈教材提供〉
AMED 支援「国際誌プロジェクト」　提供



**草案**
新谷歩　大阪市立大学医学研究科医療統計学講座教授
加葉田大志朗　大阪市立大学医学研究科医療統計学講座特任助教

**査読**
大門貴志　兵庫医科大学医療統計学教授
山中竹春　横浜市立大学医学部臨床統計学教授
市川家國　信州大学特任教授
山本紘司　大阪市立大学大学院医学研究科医療統計学講座准教授
石原拓磨　大阪市立大学大学院医学研究科医療統計学講座特任助教

## 目次

はじめに

多重検定の対処法

- ボンフェローニ法
- ダネット法
- ホルム法
- グローバル（総括的な）検定を用いた場合

まとめ

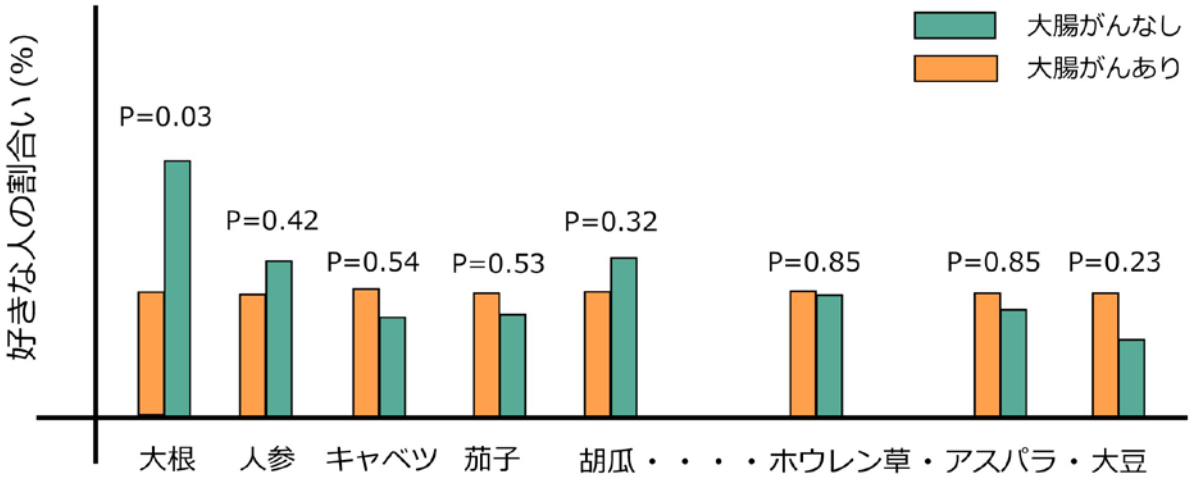
## はじめに

### 仮説「大根好きは大腸がんになりにくい」

研究を始めるときに最も重要なことは、何について研究をするのか、「データを集める前に研究仮説を立てる」ことです。「とりあえずデータを集めてみて、統計的に有意な差が出たものについて後付けで仮説を立てればよい」という姿勢は極めて不適切です。それでは以下の例でなぜ後付けの仮説がいけないのか考えてみましょう。


大腸がんの人とそうでない人にそれぞれ 100 人にアンケートを取って、好きな野菜を聞いてみました。

大根・人参・キャベツ・茄子・胡瓜・ジャガイモ・山芋・里芋・さつま芋・ズッキーニ・ピーマン・キャベツ・白菜・モヤシ・オクラ・筍・レタス・ケール・ネギ・玉ねぎ・セロリ・ホウレン草・アスパラ・大豆




データを解析すると、大根が好きな人の割合は、大腸がんの人で 30%、そうでない人で 60%でした。P 値は 0.03 で統計的に有意な差がありました。ニンジンが好きな人の割合は、大腸がんの人で 30%、そうでない人で 37%であり、P 値は 0.42 で統計的に有意な差はありませんでした。他の 22 個の野菜でも同様に比較すると、統計的に有意差が確認された野菜は大根だけでした。


論文の標題：「大根好きは大腸がんになりにくい」
方法：大腸がんになった人となっていない人それぞれ 100 人に大根好きかどうかアンケートで調査。

これは明らかに後付けの仮説ですね。仮説もなしに様々な野菜の好き嫌いの回答データを収集して野菜ごとに仮説検定を繰り返し、「大根」で統計的に有意な差があることを確認した上で、あたかも「大根好きは大腸がんになりにくい」という仮説を当初から設定し、この仮説を証明すべく「大根」のデータだけ収集したかのように論文を書いて投稿してしまいました。他の 23 個の野菜を調べたことは論文には記載されていません。

なぜ、後付けの仮説がいけないのか、ここに統計学の落とし穴が潜んでいます。仮説検定の単元でも学習しましたが、P 値とは「データをもとに差がないのに間違って差があると示してしまう確率」ともいえます。つまり、大腸がんのあり・なしで、24 個の野菜のそれぞれについて好きと回答する割合を比較し、それぞれの比較で P 値を計算した場合、24 個の P 値が得られます。それぞれの P 値が 5%より小さければ統計的に有意な差があると判断してしまうと、24 個の野菜のうち仮にどの一つの野菜も本当は関連がない場合でも少なくとも一つの野菜で統計的に有意な差が生じる確率は「$1-0.95^{24}$」となり、この確率は 71%にもなります。仮説もなく 24 個の野菜のそれぞれについて検定を手あたり次第行うと、(24 個のどの野菜も大腸がんとは関連がなくても)、71%の確率でどれか一つの野菜が「間違って統計的に有意な差があると判断してしまう」のです。医学研究の論文にこれをなぞらえると、一つの論文で 50 個も P 値が計算されているような研究で、間違って少なくとも一つの検定で統計的に有意な差があると判断してしまう確率は 92%にもなります。仮に検定の数が 10 個に減っても間違って統計的に有意な差があると判断してしまう確率は 40%です。このように複数の P 値を計算してしまうと、少なくとも一つの P 値について間違って統計的に有意な差があると判断してしまう確率が高くなる。そのことを統計の専門用語で「多重検定の問題」と呼んでいます。

| 手あたり次第計算した P 値の数 | 少なくともどれか一つの P 値に対して統計的に有意な差があると判断してしまうる確率 |
|---|---|
| 1 | 5% |
| 5 | 23% |
| 10 | 40% |
| 30 | 79% |
| 50 | 92% |
| 100 | 99% |

このように仮説を立てず手当たり次第に検定を繰り返し、統計的に有意な差があることを確認して後、あたかも最初からその項目だけをピンポイントで検定したように論文にまとめてしまうことを、P ハッキング、フィッシング、チェーリーピッキング、データドレッジングなどと呼び、このような行き当たりばったりの研究には再現性がないとして、国際誌では厳しく取り締まっています。

通常、研究を始めるときには、日常診療から得られた臨床的疑問を解消するためであったり、先行研究から得られた知見をさらに発展させるためであったりと、何らかの目

的がなくてはなりません。この目的に対応する仮説を設定し、綿密に計画を立て、当該仮説を証明するために必要と考えられるデータを収集し、そのデータに適切な統計解析を行った上で成果を報告することが望まれます。しかし、今日でもこのようなプロセスを踏まず、手当たり次第にデータを収集して解析を行い、結果を報告していると思われるものも少なくありません。研究計画は、研究全体の中でも極めて重要な部分を占め、また統計解析を行う際にも注意が必要です。本単元では、統計解析を行う際に研究計画がなぜ重要であるかについて勉強していきます。

**学習目標**

本単元を通じてあなたが修得を目指すものは：

- 研究計画の重要性を学ぶ
- 多重性の問題を考慮せねばならない場面を学ぶ
- 多重性の問題への基本的な対処法を学ぶ



## 多重検定の対処法

多重検定の問題への対処法は、大きくは「事前の対処法」と「事後の対処法」に分けることができます。

【事前の対処法】 検定する仮説に優先順位をつける
【事後の対処法】 実施した検定の数を考慮し、その中で適切な結論を得るために、統計的有意差が出にくい方向にP値または有意水準を補正する

【事前の対処法】
これは、研究計画時に、実施しようとしている検定に優先順位をつけるというやり方です。　例えば、大腸がんの予防に食べた食物の繊維質の量が関わっているという仮説を証明したいとしましょう。このとき、繊維質の多いと思われる野菜から優先的に検定を行います。具体的には、優先順位を①ごぼう、②ほうれん草、③ブロッコリー、④ピーマン、⑤大豆として、この順に検定を行います。まずごぼうから検定を行って、次の順位の野菜については、その前の野菜で統計的に有意な差があると判断されない限り、検定を行わないといったやり方です。このように検定の優先順位をつけるやり方は医薬品開発などでもよく用いられています。例えば、米国 FDA（Food and Drug Administration）は医薬品開発時のアウトカムの順位を決めるよう推奨しており、以下のように記載しています。

- 主要アウトカム
    - 薬剤の効果を定義するために必要（または充分）なアウトカム
- 副次的アウトカム
    - **2次的**
        - 薬効の証明に必要ではないが（主要アウトカムで効果のない場合はこれ自体では薬効を示すのに十分でないが）、有効性のラベルに

表示される可能性がある

- **3 次的（探索的）**
  - 将来的な仮説を創生するための探索的なアウトカム

研究結果は、統計的に有意な差の有無にかかわらず研究計画時に決定した順序で結果をまとめ、主要アウトカムで有意差がでなかった場合は、「主要アウトカムでは有意差は認められなかった（示唆できなかった）が、副次的アウトカムでは有意差が認められた」というように発表することが大切です。このようにアウトカム及びその検定の重要度を研究計画時に決めておき、その順番で「検定を行うことにより、多重性の問題に対処することが可能です。

【事後の対処法】

これは、研究終了後に、実施した検定の数を考慮して統計的に有意差がでにくい方向に P 値または有意水準を補正する方法です。具体的には、実施した検定の数をそれぞれの P 値に掛け算して、その実施した検定の数が大きければ大きいほど有意差を出にくくする方法です。例えば、実施した検定の数が 3 つの場合はそれぞれの P 値に 3 を掛けます。大腸がんと大根の例では 24 個の検定を実施しているので、それぞれの P 値に 24 を掛けます。例えば、大根の P 値は 0.03 だったので、0.03×24＝0.72 となります。人参の P 値は 0.42 だったので、計算上は 0.42×24＝10.08 となります。これは 1 を超えています。P 値は確率なので当然 1 を超えることはできないので、この場合は 1 と置き換えます。この方法を採用すると、大腸がんと大根の例では、大根以外の野菜の P 値がほとんどすべて「1」に近くなり、統計的に有意な差が得られにくい方向に補正が行われています。もちろん大根についても、補正後の P 値は 0.72 ですから、統計的に有意な差は認められません。

一方、P 値を補正するのではなく、有意水準を補正する方法もあります。具体的には、有意水準を 5%としたとき、この 5%を実施した検定の数で除することで補正し、個々の検定の P 値をこの補正後の有意水準と比較します。先ほどの大腸がんと大根の例で考えると、有意水準を 5%としたとき、0.05÷24＝0.00208 であり、大根についての P 値をこの有意水準と比較することになります。大根の P 値（0.03）は、この有意水準（0.00208）よりも大きいので、統計的に有意な差が認められないということになります。P 値を補正するのではなく、有意水準を補正する方法でも、実施した検定の数が増えるほど統計的に有意な差があると判断されにくくなることが分かります。

P 値の補正は、以下のような場面で行われます。

（1）比較群が 3 つ以上存在する
（2）アウトカムが 2 つ以上存在する
（3）中間解析など研究実施中にデータの比較が繰り返し行われている
（4）回帰分析などでリスクファクターなど暴露因子が 2 つ以上存在する
（5）データが経時的に繰り返し収集され、それぞれの時点において比較が行われている

以下では、このうちの（1）の場面を例示します。

## ボンフェローニ法

3 つ以上の群で比較を行うときに、そのうちの 2 つの群を取り出して比較することを対比較といいます。ボンフェローニ法は、考えられるすべての対比較を行うことを前提に P 値の補正を行います。新薬 A、新薬 B、既存薬 C を投薬された患者さん 3 群で血圧を比較する場面を考えましょう。考えられるすべての対比較は新薬 A 対新薬 B、新薬 B 対既存薬 C、新薬 A 対既存薬 C の 3 つであるので、各対比較の P 値に 3 を掛けて補正後の P 値が得られます。

## ダネット法

ボンフェローニ法とは異なり、最初に参照群を設定しておいて、参照群と他の群の対比較のみを行うことで対比較を減らす方法です。ボンフェローニ法で想定したものと同じ場面を考えましょう。既存薬 C を参照群とすると、新薬 A と新薬 B を比較することはせず、ダネット法では、新薬 A と既存薬 C、新薬 B と既存薬 C の 2 通りの対比較を行い、その対比較の数と各群の患者数から計算される補正係数を用いて補正後の P 値が得られます。ボンフェローニ法と比べて、比較の対を事前に絞り込むことになるので、P 値の補正が緩くなります。

## ホルム法

ホルム法もまた、ボンフェローニ法よりも P 値の補正が少し緩い方法になります。ホルム法では、ボンフェローニ法と同様、すべての対比較を扱うことができますが、P 値の小さなものほど補正を強くしていくという方法です。具体的には、三つの対比較を行った場合、P 値が 1 番小さい対比較の P 値に、ボンフェローニ法と同じように、対比較の数を掛け、P 値が 2 番目に小さい対比較の P 値に、対比較の数よりも「1」小さい数（対比較の数－1）を掛け、P 値が 3 番目に小さい対比較の P 値に対比較の数よりも更に「2」小さい数（対比較の数－2）を掛ける、といった補正を行います。

---

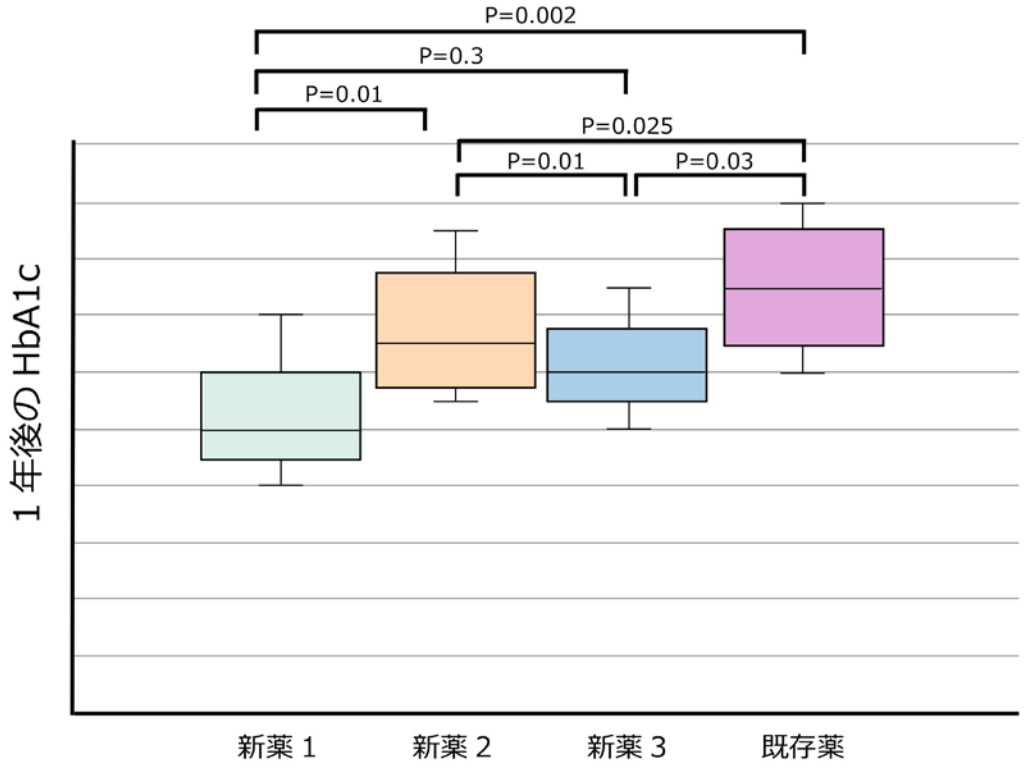
以下の例で上の 3 つの方法を経験してみましょう。

3 種類の新薬（新薬１、新薬２、新薬３）と既存薬を糖尿病患者に 1 年間投与し、比較的長期間の血糖値の指標である HbA1c を比較する場面を想定します。この場面では、最大で 6 通りの対比較を考えることができます。図には、この 6 つの各対比較に対して補正を行わない P 値を付与しています。

上述の三つの方法を用いて補正された P 値は表のとおりになります。

| 対比較群 | 修正なし | ボンフェリーニ法 | ダネット法 | ホルム法 |
|---|---|---|---|---|
| 新薬 1 vs 既存薬 | 0.002 | 0.002 x 6 | 0.002 x 3 | 0.002 x 6 |
| 新薬 1 vs 新薬 1 | 0.01 | 0.01 x 6 | – | 0.01 x 5 |
| 新薬 2 vs 既存薬 | 0.025 | 0.025 x 6 | 0.025 x 3 | 0.025 x 4 |
| 新薬 3 vs 既存薬 | 0.03 | 0.03 x 6 | 0.03 x 3 | 0.03 x 3* |
| 新薬 3 vs 新薬 2 | 0.1 | 0.1 x 6 | – | 0.1 x 2 |
| 新薬 1 vs 新薬 3 | 0.3 | 0.3 x 6 | – | 0.3 x 1 |

＊補正後に順序が変わるので，順序が入れ替わる相手の値で置き換える．




## グローバル（総括的な）検定を用いた場合

比較群が 3 つ以上ある場合に、どの群でも構わないので 1 群でも他の群と違いがあるかどうかを検定する方法を「グローバル検定」と呼ぶことがあります。例えば、4 群の平均を比較するような場面では、この 4 群の中のどれか 1 つの群でも構わないので平均が他の群と違いがあるか、について仮説検定を行うというものです。これは後の単元で説明する分散分析のところで詳しく学びます。

グローバル検定の帰無仮説：　すべての群でアウトカムの平均が等しい

グローバル検定において P 値が 0.05 未満になれば、少なくともどこか一つの群のアウトカムの平均がほかの一つの群のものと統計的に有意に異なるということになります。この検定は比較群の数が増えても P 値は一つしか計算されないため、多重性の問題は起こりません。グローバル検定で統計的に有意な差が認められなかった場合には、どの群とどの群のアウトカムの平均に違いかあるかを確認する対比較へ進むことはできません。

A群、B群、C群で血圧を比べる場面を考えましょう。
このとき、帰無仮説を「A群、B群、C群すべてで平均血圧が等しい」と設定することも可能です。このとき、データを収集後、グローバル検定で帰無仮説を棄却できた場合、「A群とB群の平均血圧が等しい」、「B群とC群の平均血圧が等しい」、「A群とC群の平均血圧が等しい」という3つの帰無仮説に対してそれぞれ、A群とB群、B群とC群、A群とC群の対比較をスチューデントのt検定などを用いて行います。
グローバル検定で帰無仮説を棄却できなかった場合、このような対比較は行わず検定を終了します。

## まとめ

研究計画時に仮説を設定せず研究を開始し、そこで収集されたデータの検定結果（統計的に有意な差かどうか）に応じて後付けの仮説をあたかも研究計画時から存在していたかのように事後に設定し、その正しいとはいえない結果を発表してしまうことを防ぐために、最近では研究を開始する前に研究計画を公示することを多くの国際誌で義務付けています。研究実施計画書には主要評価項目（主となるアウトカム）、副次評価項目が何かをはっきりと表記し、検定の結果、主要評価項目で統計的に有意な差が認められなくとも、主要評価項目、副次評価項目と研究実施計画書に記載された順に、それらの結果を論文に表記することが求められます。このような優先順位がつけられない場合は、有意差の出にくい方向に P 値や有意水準の補正を行う必要があります。優先順位をつけない手当たり次第の検定では、研究結果の再現性はおぼつき難く、再現性のある研究を行うためにも、先ず研究計画をしっかりと立てることが大変重要です。

**この単元に関係するビデオ教材**

多重検定問題

本単元は日本医療研究開発機構:研究公正高度化モデルである「医系国際誌が規範とする研究の信頼性にかかる倫理教育プログラム」（略称：AMED 国際誌プロジェクト）によって作成された教材です。作成および査読等に参加した専門家の方々の氏名は、冒頭に掲載されています。

この単元に関係する国際誌におけるチェックポイントをいくつか紹介します。
（内容は解釈を助けるために一部意訳している部分もあります）

①Nature
（http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/12/2010128212513557501.pdf; visited on 2018.02.11）
②New England Journal of Medicine （http://www.nejm.org/page/author-center/manuscript-submission#electronic; visited on 2018.02.11）
③Science （http://www.sciencemag.org/authors/science-editorial-policies; visited 2018.02.11）
④The EMBO Journal （http://emboj.embopress.org/authorguide#embargopolicy; visited on 2018.02.11）
⑤JAMA （http://jamanetwork.com/journals/jama/pages/instructions-for-authors; visited on 2018.02.11）

**① Nature**

■ 多重性によるＰ値の補正を行なっている場合にはその旨を説明すること

**③ Science**

■ 多重比較により偽陽性の確率を調整するために有意水準の補正法（Bonferroni の調整法やその他）を利用した場合は、その手法についても記載すること。

**④ The EMBO Journal**

■ ひとつのデータセットを利用して複数の統計検定を行う場合、第Ｉ種の過誤が上昇することについてどのような対策をとったのかを記載すべきである。

**⑤ JAMA**

■ ２個以上の主評価項目が用いられた場合は、そのＰ値について多重性による補正を行うこと。副次的な評価項目においては多重性によるＰ値の補正を行うか、その結果を探索的なもの（あくまでも仮説を生み出すものであって仮説を検証するものではない）として報告すること。
■ 副次的な解析やサブグループ解析の場合、検定が多重になるため関連がないのにあるとしてしまう第１の過誤をどう回避したか（有意水準の補正等）を記載すること。そのような記載がない場合の解析結果は探索的または後付けの解析として扱うこと。