第5章

病気の原因を調べるための疫学研究1：ケース・コントロール研究

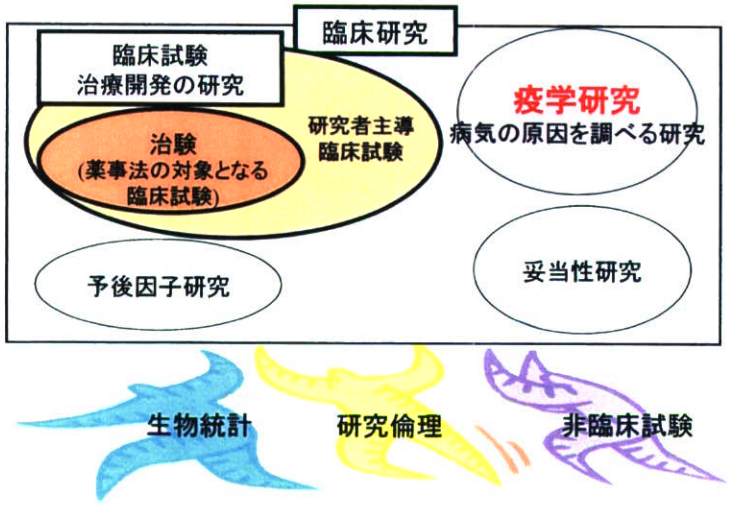
本講義の内容

- ケース・コントロール研究について、その利点と欠点を知る。
- この研究デザインがエビデンスレベルの中でどこに位置付けられているのかということを理解する。

第1章の講義「臨床研究概論」にて、疫学研究とは病気の原因を調べる研究であると説明しました。
本講義では、疫学研究デザインの1つであるケース・コントロール研究について、上記の内容を理解することを目的として説明していきます。

http://www.icrweb.jp/icr/

臨床研究とそれを支えるもの



疫学研究とは病気の原因を調べる研究

本講義では臨床研究を医学研究と同じ意味で用いることにしています。
そのように定義すると、病気の原因を調べる疫学研究も、臨床研究の1つといえます。

第5章 病気の原因を調べるための疫学研究1：ケース・コントロール研究

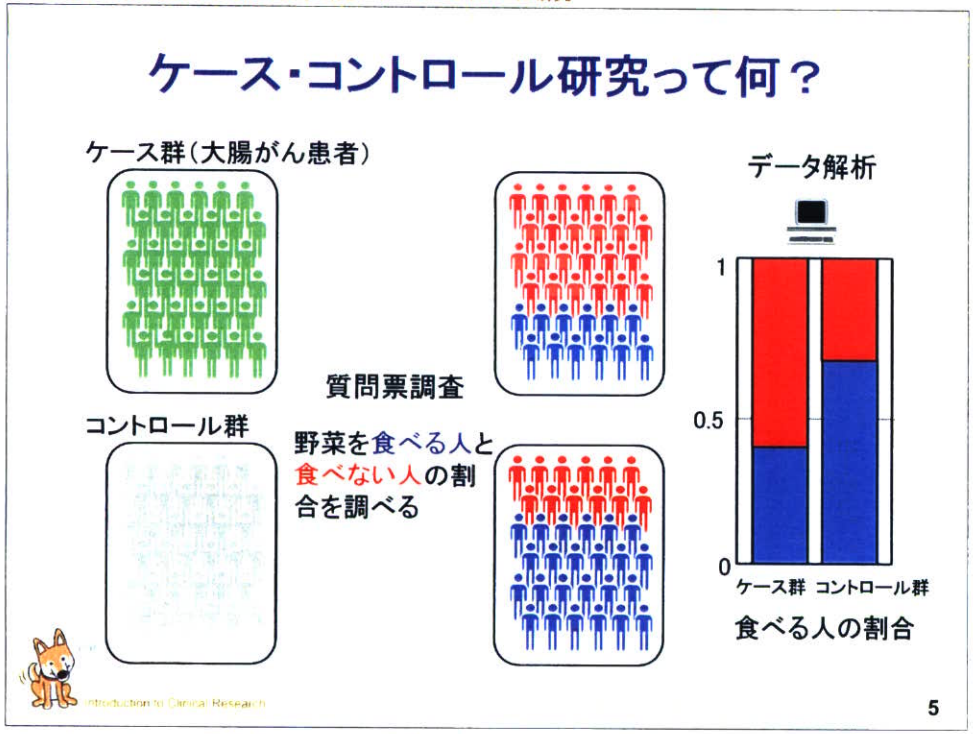
関連を見る研究

- 観察研究
  - ケース・コントロール研究（症例対照研究）
  - コホート研究
- 「野菜摂取と大腸がんの関連」を例に
  研究デザインを考える。

ケース・コントロール研究は、日本語では「症例対照研究」あるいは「患者対照研究」といいます。

では、野菜摂取と大腸がんの関連を調べる研究を実例として、ケース・コントロール研究について説明していきましょう。

http://www.icrweb.jp/icr/



実際に国立がんセンターが中心となって行った研究を例に取ります。

研究の目的は食事等の生活習慣と大腸がんの関係を疫学的に調べることでした。長野県の4つの病院の協力を得て、1998年から2001年にかけてケース・コントロール研究を行いました。

この研究では、4つの病院で新たに大腸がんと診断された患者さんのうち、入院患者さんを『ケース』としました。

一方『コントロール』は、ケースをリクルートした病院の人間ドック受診者の中から、「性別」・「年齢」が一致し、かつ、がんを持たない健康な人を選び、リクルートしました。このように、ケースと条件が同じ人を選ぶことを「マッチング」と呼びます。

マッチングは、性別、年齢等のように、がんのリスクと関係あるかもしれない要因の影響を取り除くとともに、統計解析時の効率を上げるために行います。マッチングした場合には、マッチングを考慮した統計解析をしなければならないのですが、本講義では、簡便のためにマッチングをしなかったことにして話を進めます。数字についても若干改変しています。

ケース・コントロール研究では、ケース（患者さん）とコントロール（患者さんでない人）で関心のある要因を比較することが基本です。

まずケース群とコントロール群を集めます。

次にそれぞれの対象者に対して質問票等によって調査を行い、それぞれの群での野菜を多く食べる人とあまり食べない人を同定します。

野菜を多く食べる人の割合をケース群とコントロール群で比較することで、野菜と大腸がんの関連を調べます。

簡単にいうと、コントロール群でケース群より野菜を多く食べる人の割合が多ければ、野菜が予防的に働くのではないかと推論することになります。

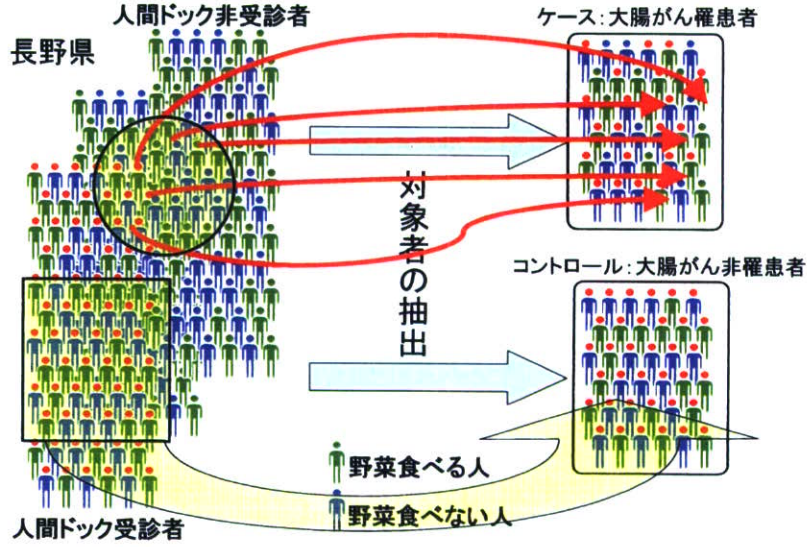
コントロールの選び方（1）

- ケースと同じ母集団からサンプリング
  - –コントロールは大腸がんになったら調査対象医療機関を受診する人から選ぶ。
- 人間ドック受診者は理想的なコントロールか？

ケース・コントロール研究で最も重要なことは、コントロールの選び方です。
コントロールはケースが発生してくる集団と同じ集団から選ばなければなりません。
いい換えると、もしコントロールに選んだ人が大腸がんであった場合には、ケースとして選ばれる可能性があるということです。

この研究では、ケースと同じ病院の人間ドック受診者をコントロールと決定しました。
しかしながら、このコントロールは理想的なコントロールといえるでしょうか。

コントロールの選び方（2）



この研究は長野県で行われていますので、想定母集団として長野県の住民を考えます。
長野県住民を単純に人間ドック受診者（赤い頭の人）と非受診者（赤くない人）に分けます。
この中からケースとコントロールをリクルートします。

ケースは人間ドック受診者、非受診者ともに、大腸がんになれば調査対象医療機関を受診するので、長野県全体から選ばれます。
一方、コントロールは人間ドック受診者を対象にしていますので、人間ドック受診者しかいません。つまり、がんになって調査対象医療機関を受診するであろう人のうち、一部の人しか対象にしていないということになります。そう考えると、人間ドッグ受診者というのは必ずしも理想的なコントロールとはいえません。

理想的なコントロールを選ぶためには、ケースをリクルートした病院の医療圏を正確に定義し、そこに住む住民からランダムに抽出する必要があります。この2つのステップは実際にはどちらも非常に難しい作業です。

理想的なコントロールを選ぶことが難しい場合は、できるだけ理想に近いコントロールを選ぶことが重要です。人間ドック受診者は、もしがんが見つかったらその受診していた病院を受診する可能性がほかの人よりは高いだろうと考えられるので、コントロールのいい候補といえるかもしれません。また、同じ病院の受診者でほかの疾患にかかった人をコントロールとして選ぶことも次善の策としてよく行われる方法です。

セレクションバイアス

- 研究対象を選択するときに生じるバイアス（偏り）
  - コントロールの選択に注意
  - ケースとコントロールで調査協力者の割合が異なったら？
    - ケースは一般的に協力的
    - コントロールは？
      - 人間ドック受診者は一般住民より協力的

Introduction to Clinical Research

9

この例のように、研究対象者を選択するときに生じる結果の偏り（バイアス）を「セレクション（選択）バイアス」と呼びます。

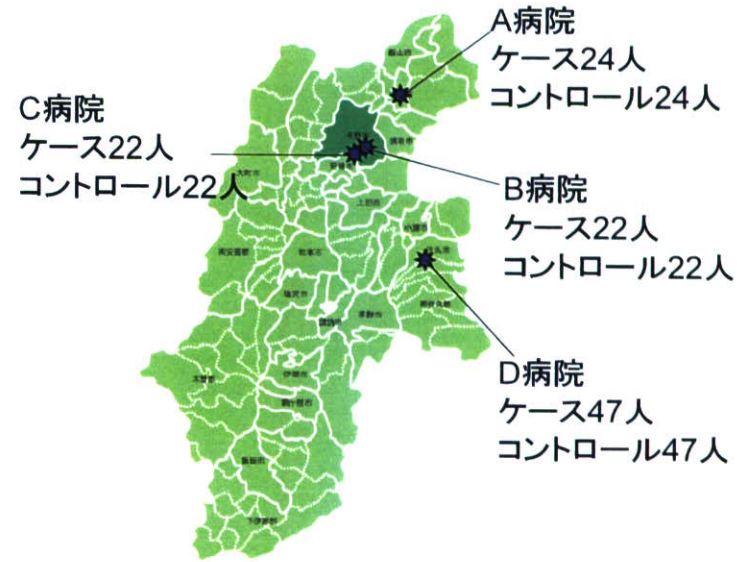
セレクションバイアスは、コントロールをどの集団から選ぶかを決める際に、特に問題になります。

しかし、コントロールを選ぶ集団を決めたあとでも、ケースとコントロールで調査協力者の割合が異なった場合には、さらにセレクションバイアスが生じる可能性があります。

一般的にケースは調査に協力的であり、高い協力割合が期待できますが、コントロールである健康な人は調査になかなか協力してくれないことが多く、ケースと同じだけの協力割合を期待することは困難です。

人間ドック受診者というのは、少なくとも一般住民よりは協力的であると考えられ、その点は人間ドック対象者をコントロールとした利点といえます。

ケースとコントロールの数



Introduction to Clinical Research

10

このような方法でケースとコントロールを選び、全体で大腸がんのケース115人、コントロール115人の方の協力を得ることができました。

ケースのリクルートはコントロールより困難？

- 調査期間と目標ケース数
  - 2年間で150人以上
- 現実は？
  - 約3年半で115人

実際には、ケースをリクルートすることも大変です。

この研究では、研究計画の中で調査期間を2年間、目標ケース数を150人以上と計画を立てましたが、現実には3年半で115人の協力にとどまりました。

2年間で150人以上というのは理論的には可能な数字だったのですが、ケースのリクルートを忙しい臨床医の先生方にお願いしたため、115人が限界だったといえます。

ケース・コントロール研究だけでなく、どんな研究を行う際にもいえることですが、人的なサポートも含めて研究期間内に研究が完遂できるような体制を整えることが重要です。

調査項目

- 曝露を測定する
  - 自記式の質問票
    - 調査担当者による記入漏れのチェック
  - 採血
    - 治療開始前に研究用として14mL採血
- 疾患の情報収集
  - 臨床情報
    - ケースのみ病理組織型等の情報

この研究は、野菜と大腸がんの関連を見ることが目的ですが、疾患の原因かどうかを知りたい『野菜』のような要因のことを疫学では「曝露」と呼びます。対象者から研究協力の了解が得られると、まず曝露を測定することから始めます。

疫学研究は対象者の数が多いので、実施可能性を重視し、簡便な方法として自記式の質問票（アンケート調査票）等を用いて曝露情報の収集を行うことが多いといえます。質問票には関心のある野菜だけでなく、交絡要因となる可能性のもの（第8章 生物統計学2で詳述します）、つまり、大腸がんのリスクとなるようなものを全て質問項目として含める必要があります。質問票は単に作ればいいものではなく、十分に妥当性や再現性が検証されたものを用いなければなりません。妥当性研究については別の機会に説明します。

この研究では、ケースについては入院が決まって予約をしたときに質問票を配布し、入院時に回収して調査担当者による記入漏れのチェックを行いました。コントロールは1泊2日の人間ドック参加者を対象としましたので、ドックの期間内に記入していただき回収しました。

また、質問票による調査以外に、治療開始前に14mLの採血を行っています。採血すると、野菜と大腸がんに加えて、血中のベータカロテン濃度と大腸がんリスクの関連や、ある特定の遺伝子と大腸がんの関連等も調べることができます。

この質問票調査と採血については、ケース、コントロールともに行いますが、もう1つ集めなければならない重要な情報として疾患の情報があります。ケースについては臨床情報として病理組織型等の詳しい情報を集めることも重要です。

質問票に含まれる項目

- 21ページのマークシート
  - 身長、体重
  - 職歴
  - 家族歴、既往歴 、服薬歴
  - 喫煙、飲酒習慣
  - サプリメント使用状況
  - 食物摂取頻度調査（141項目）

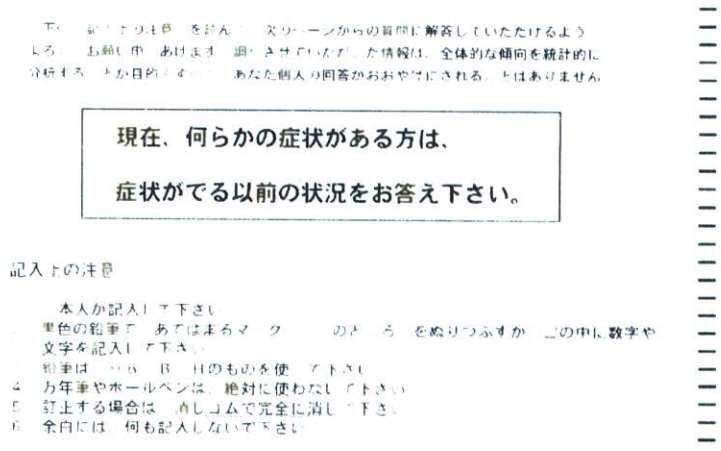
この研究で使用した質問票は、21ページからなるマークシートで、上のような項目が含まれています。

野菜と大腸がんの関係を調べる場合には、食物摂取頻度調査票というものが用いられることが一般的です。
先ほど、関連を見たい野菜を「曝露」と呼ぶと説明しましたが、質問票に含まれている要因はそれぞれ疾患との関連が疑われる要因ですから、野菜だけでなく、それぞれを「曝露」として取り扱い、疾患との関連を調べることが可能です。

このようにケース・コントロール研究では、いろいろな要因と（研究対象となった）疾患の関連をみることができます。

ケース・コントロール研究での注意点

今の食生活を聞いてもそれは病気になってからのことかもしれない。



ケース・コントロール研究で食物摂取頻度調査票を用いて食事の状況や野菜の摂取量を聞く場合には、特に注意しなければならない点があります。

ケースについて質問することを考えてみましょう。
ケースは患者さんが対象ですので、調査時点で食生活を尋ねても、それは病気になってからのことかもしれません。調査票には「現在何らかの症状がある方は、症状が出る前の状況をお答えください」という注意書きを示し、病気になる前の食生活について答えていただくことが大事なポイントになります。

リコールバイアス

- 「思い出す」という行為に伴って生じるバイアス（偏り）
- 過去のことを聞いても正しいかどうかわからない
- ケースは、過去の行動と現在の結果を考え、過大または過小に評価するかもしれない
  - ケースとコントロールで思い出し方が異なる場合にバイアスとなり、結果に影響を与える

同じ過去のことを尋ねたとしても、ケースとコントロールで答え方に差が出る場合があります。

ケースは自分がなぜ病気になったのだろうと考え、その原因となったものを探す傾向にあります。その結果、コントロールの方に比べてよりたくさん思い出す傾向にあります。本当は疾患に関連がなくても、ケースではよりたくさん思い出すことによって、見かけ上の関連が生じてしまうことがあります。

これはケースとコントロールの間での思い出し方に差があることで生じる偏り（バイアス）ですので、思い出しバイアス（リコールバイアス）と呼びます。

関連の指標

	ケース	コントロール
食べる	57	71
食べない	58	44
合計	115	115

$$\text{オッズ比} = \frac{\frac{57}{58}}{\frac{71}{44}} = 0.61$$

$$\text{オッズ比} = \frac{\text{ケースで野菜を食べる人の人数} \,/\, \text{ケースで野菜を食べない人の人数}}{\text{コントロールで野菜を食べる人の人数} \,/\, \text{コントロールで野菜を食べない人の人数}}$$

実際に曝露のデータが得られた場合、野菜摂取と大腸がんの関連をケース・コントロール研究で検討するにはどのようにすればよいのでしょうか？

曝露有り無し、疾患有り無しで2x2分割表、あるいは4つ目表と呼ばれるものを作ることが基本となります。今回はケースとコントロールでそれぞれ「野菜を食べる群」「食べない群」に分け、集計表を作ります。

この研究の場合、ケース115人のうち、野菜をよく食べる人は57人、あまり食べない人は58人でした。コントロール115人のうち、野菜をよく食べる人は71人、あまり食べない人は44人でした。

野菜を食べる人の食べない人に対する比のことをオッズといいます。つまり、ケースでのオッズは57/58、コントロールでのオッズは71/44です。また、ケースとコントロールのオッズの比をオッズ比と呼びます。ケースとコントロールで野菜を食べる人の割合が変わらない場合、ケースとコントロールのオッズは等しくなる（オッズ比が1）ことが期待され、ケースよりコントロールで野菜を食べる人の割合が多ければ（野菜が予防的な場合）は、オッズ比は1より小さく、ケースよりコントロールで野菜を食べる人の割合が少なければ（野菜がリスクの場合）、オッズ比は1より大きくなります。

この研究では、オッズ比は0.61となり、野菜が予防的に働いていると解釈することができます。オッズ比は、疾患の発生率がそれほど高くない場合、野菜をたくさん食べる場合の食べない場合に対する疾患発生リスクの比と解釈することができます。つまりこの例の場合、野菜を食べる人は食べない人に比べて0.61倍大腸がんになりやすい、あるいは1/0.61=1.6倍大腸がんになりにくいと解釈できます。

結果へのバイアスの影響は？

- セレクションバイアス
  - コントロールは人間ドック受診者なので一般住民より野菜をたくさん食べる人が多いかも知れない。
    - 見かけ上、野菜を多く取る群のリスクが低く出る可能性あり
- リコールバイアス
  - 野菜が大腸がんの予防になるという知識が普及していた場合、ケースは「野菜を食べなかったからがんになってしまった」と考え、摂取量を低く申告するかも知れない。
    - 見かけ上、野菜を多く取る群のリスクが低く出る可能性あり

今回得られた結果にバイアスがどういう影響を与えたかということを考察してみます。

まずセレクションバイアスについて考えます。
今回は人間ドック受診者をコントロールとしました。人間ドックや検診の受診者は、一般的に健康的な生活習慣の方が多く、一般住民より野菜をたくさん食べる人が多い可能性があります。もしこのような理由で単にコントロールで野菜をたくさん食べる方が多かった場合、見かけ上野菜を多くとる群でのリスクが低く出る可能性があります。

次にリコールバイアスについて考えます。
野菜が大腸がんの予防になるという知識が普及しており、ケースが野菜を食べなかったからがんになってしまった、と考え、摂取量を低く申告したということがあった場合には、これも見かけ上リスクが低く出る可能性があります。

今回のケース・コントロール研究では、セレクションバイアスやリコールバイアスが起こっていない保証はありません。今回、リスクが低いという結果が出ましたが、この結果はこれらのバイアスの影響のせいかもしれません。

Take Home Messages

ケース・コントロール研究の特徴

- ケースとコントロールで調べたい要因保持の差を比較する研究。オッズ比を計算し、要因有り無しによる疾患のなりやすさを調べる
- 欠点
  - バイアスの影響を受けやすい
    - セレクションバイアス、リコールバイアス
  - 1度に1つの疾患しか扱えない
- 利点
  - まれな疾患に有効
  - 費用・時間が少ない
    - 費用は年間数百万円、期間は数年

最後に、ケース・コントロール研究について、その特徴を簡単にまとめます。

ケース・コントロール研究とは、ケースとコントロールで調べたい要因保持の差を比較してオッズ比を計算することにより、要因有り無しによる疾患のなりやすさを調べる研究方法です。

欠点は、セレクションバイアスやリコールバイアスという2つの代表的なバイアスの影響を受けやすいことです。
また、1度に1つの疾患しか扱えないということも欠点に挙げられます。

一方、利点としては、疫学研究のもう1つの典型的な方法であるコホート研究では実施の難しい、発生率の低いまれな疾患に対する研究でも有効であり、また、コホート研究に比べて費用・時間が少なくて済むということが挙げられます。

コホート研究については次の章（第6章）で説明します。

第7章
生物統計学1：仮説検定

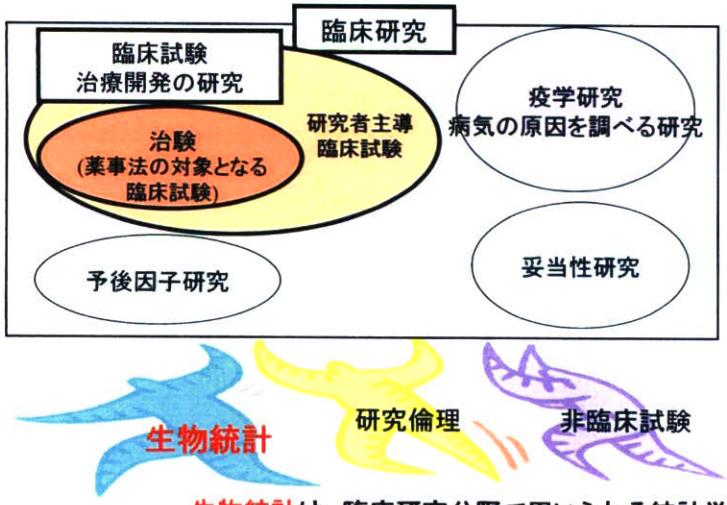
本講義の内容

- 生物統計学とは
- 仮説検定
  - p値
  - αエラーとβエラー
  - サンプルサイズと検出力

生物統計学とは、臨床研究分野で用いられる統計学のことです。

ICR初級編では、臨床研究にかかわる上で知っておく必要のある生物統計学の概念について、本講義（第7章）と次の講義（第8章）の2回に分けて説明していきます。

臨床研究とそれを支えるもの



生物統計は、臨床研究分野で用いられる統計学のこと

本講義で説明する生物統計学は、研究倫理や非臨床試験とともに、臨床研究を支える重要な要素といえます。

生物統計学とは

- 統計学を医学など生物分野で応用する学問
  - Bio-statistics 生物統計学、Bio-metrics 計量生物学ともいう
- いかなる臨床研究でも
  データを測定し、データから結果を解釈する上で
  生物統計学的な考え方は非常に重要
- 実際の応用には専門家が必要
  - 「エクセル、Stat View、SAS、、、で計算すること」と
    「臨床研究で生物統計学を利用すること」は全く別のこと

ここでは、生物統計学について、もう少し詳しく説明します。

物理や化学の実験であれば、起こる現象についていろいろな公式を使って説明しようと試みます。
しかし、人間の体で起こることを理論計算するためには、いろいろな条件が複雑すぎるといえます。そこで、ある程度の個体数を観察することによって、知りたい仮説を実証的に検証していこうという学問が生物統計学です。
いかなる臨床研究でもデータを測定してデータから結果を解釈する上で、生物統計学的な考え方は非常に重要となります。
実際にこの学問を応用するのは、生物統計家と呼ばれる専門家の仕事ですが、生物統計学について基本的な考え方を理解することは臨床研究を理解する上で必須といえます。

なお、公式に基づいて必要対象者数を計算したり、解析ソフトを使って解析するということ自体は統計学ではありません。

仮説検定

もしも  中田君が科学者ならば

- よい方法の1つは**実験** experiment
  - 実際に審判にコインを何度か宙へと投げてもらって
    ほぼ1/2の割合で表が出るかを試すこと

1. 

2. 

2つ目の結果の方がコインがいかさまである可能性がより高そう

仮説検定について、「中田君とコイントス」という例を使って説明をしていきます。

中田君は、サッカーの試合開始前に行われるコイントスで表が出ることが多いという印象を持っています。

そこで、「審判がいかさまコインを使っているのではないか」、つまり、コインの表が出る確率が1/2であるかどうかについて疑っています。

この疑いが事実であるかを簡単に調べる方法としては実験があります。審判にコインを何度か投げてもらい、ほぼ1/2の割合で表が出るかどうかを試すということです。

例えば、上の1.のように6回投げて表・裏が共に3回出るような実験結果が得られれば、フェアなコインと判定してもよいでしょう。また2.のように6回中5回も表が出る結果ならば、いかさまコインというべきかもしれません。

では、実際にコインを投げる実験を行い、中田君の疑いを調べてみましょう。

仮説の設定

① 表の出る確率が1/2であるか を調べたい
仮説

② コインを 12回 投げる
実験の方法

③ 表が4回以下、或いは8回以上の場合に
いかさまと判断
規準

- 実験デザインを事前に決めることは科学的な結果を導くために大事
  - プロトコールをしっかり書く（臨床研究でも同様）

実験を科学的に行うには、
①（この実験で確かめたい）仮説
②実験方法
③規準
を事前に決めなければなりません。

この実験において「仮説」とは、表の出る確率が1/2であるか否かです。

また、「実験方法」として、ここで中田君は審判にコインを12回投げてもらうことに決めたとします。

最後に、「規準」として、「表が4回以下、あるいは8回以上の場合にいかさまとする」と決めました。例えば表が4回の場合、裏が8回となり表の2倍の頻度で裏が出ることになります。中田君はこの規準を直感的に決めました。

実験から科学的な結論を導くためには、この①～③を事前に決めなければなりません。当然ながら臨床研究を行おうとする場合にも、研究実施計画書（プロトコール）にこの①～③全てを、研究実施前に書いておかなければなりません。

仮説検定で用いる仮説には呼び名がある

専門用語

仮説：表の出る確率が1/2であること

- 事前に決めた規準により却下するもの
- 帰無仮説（null hypothesis、直訳すればゼロ仮説）

対立する仮説：表の出る確率が1/2でないこと

- 規準により1つ目の仮説を却下した際に支持するもの
- 対立仮説（alternative hypothesis）
  - 仮説検定で用いる仮説はこの2つ

仮説検定で用いる仮説には、統計上の専門用語が2つあります。
事前に決めた規準によって却下したい仮説を「帰無仮説」、帰無仮説を却下した際に支持する（それだと判断する）ものを「対立仮説」と呼びます。

それでは実験をしよう！

事前に決めた規準

表の回数4回以下か8回以上であればいかさまと判断

- 目の前で審判に12回コインを投げてもらうと

- 12回中、表10回（裏2回）という結果を観察
- 規準にしたがって いかさま と判断

実際にコインを投げる実験を行ったところ、12回中表10回、裏2回という結果が得られたとします。この「表10回」の結果は、事前に決めた規準を満たしますので、中田君はいかさまと判断できます。中田君は審判に抗議をしなければなりません。

これが仮説検定 hypothesis tests
（仮説の評価）

- **仮説**を明確にして
  - 表の出る確率が1/2であること（ここでは帰無仮説という）
- **実験デザイン、規準**を定めて
  - 12回投げて、表が4回以下か8回以上であればいかさま
- **実験によりデータ**を測定し
- **定めた規準**に従って、その**仮説を評価**する
  - 表が10回出たので、いかさまと判断
    （仮に表が7回以下だったら、いかさまと判断しない）

一般に、どんな仮説を評価するかを明確にし、実験デザインと判断規準を定めた後に実験でデータを測定し、事前に決められた規準に従ってその仮説を評価するという一連の流れを仮説検定と呼んでいます。
これまでに中田君が行ったことはまさに、仮説検定といえます。

ここで統計学の出番

- 表の出た回数が、偶然の範囲内なのか、或いは偶然を超えたものかが問題
  - 偶然でないといえるならば、  君の主張が正しい
- 統計学を用いて偶然性を見積もる
  - フェアなコインが正しい場合にも
    表10回となることが偶然でどの程度起こるか
    を統計学を用いて見積もる
- 実際に統計学を使って偶然さを見積もってみよう！

しかし、中田君の仮説検定の問題として、事前に決めた判断規準が直感に頼ったものであることが挙げられます、直感的であるがゆえに恣意的である可能性が否定できません。判断規準は、誰もが納得できる明確なものでなければいけません。

この場面では、「いかさま」と主張したい中田君に対して審判は「フェア」と主張したいわけですから、当然ながら双方の納得のいく規準であることが求められます。例え規準を事前に決めた上で科学的な実験を行ったとしても、その規準が妥当なものでないのであれば、審判の「私のコインはフェア」との主張を覆すことはできないでしょう。

もう一度整理してみると、この場面では、表の出た回数が偶然の範囲内ではないことを示すことによって、審判の主張を覆すことができます。つまり、ここでの問題は表の出た回数が、偶然の範囲内なのか偶然を超えたものであるのかが問題となっているといえます。

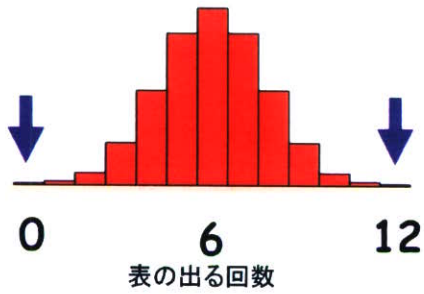
そこで、この偶然性を見積もるために統計学が必要となります。つまり、フェアなコインの場合でも表10回となることが偶然の範囲内で起こったことなのかどうかを、統計学により判断します。

仮説検定

全確率 ＝ 100%

表の出る回数	確率
12	.0002
11	.0029
10	.016
9	.054
8	.12
7	.19
6	.23
5	.19
4	.12
3	.054
2	.016
1	.0029
0	.0002

- 表は0回から12回のどれかなので、全部の確率を足すと必ず100%になる
- フェアでも12回全て表、全て裏という極端な結果が生じうる（＝確率は0でない）



コインを12回投げた場合、そのうち表の出る回数は0から12回のいずれかになります。左表はコインがフェアであった場合にそれぞれが生じる確率を算出したものです。例えば、フェアなコインを12回投げた場合に表と裏が同数6回となる確率は23%であることが分かります。

表・裏共に同数6回となる確率が最も高くなり、反対に、12回とも表あるいは裏が出る確率は非常に小さくなります。また、フェアなコインであっても非常に稀な確率では、極端な結果が生じてしまうことが分かります。ただし、一方でこのような極端な結果が出た場合には偶然でない可能性も高いといってよいかもしれません。少なくとも直感的にはそう思われることでしょう。

中田君は偶然であるかを知りたい

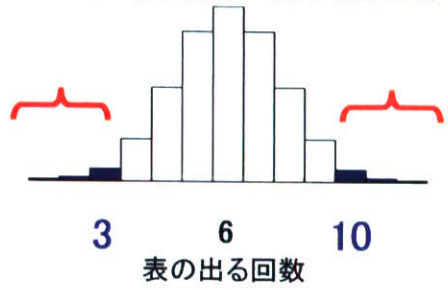
p値 表の出る回数	確率
12	.0002
11	.0029
10	.016
9	.054
8	.12
7	.19
6	.23
5	.19
4	.12
3	.054
2	.016
1	.0029
0	.0002

p値＝4%

偶然さを表す指標：p値

偶然かどうかを判断する規準として使える

p値はデータと等しいかより極端な結果の確率を足し算



さて、中田君が知りたかったのは、表が10回出ることが偶然かどうかでした。

帰無仮説が正しいとした下での偶然の程度を表す指標として「p値」があります。

この「p値」は、観察された実験結果と等しいか、より極端な結果の確率を足し合わせたものとして定義されます。

中田君のように表が10回という実験結果を得た場合、この定義に従って赤枠で囲った確率を全て足し合わせると、p値＝4%が求まります。

p値を計算してみよう

p値

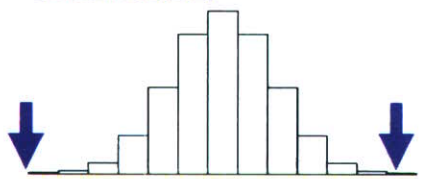
表の出る回数	確率
12	.0002
11	.0029
10	.016
9	.054
8	.12
7	.19
6	.23
5	.19
4	.12
3	.054
2	.016
1	.0029
0	.0002

もしも全て表（表12回）ならば

p値＝0.04%

小さなp値
実際にも偶然とは考えにくい

大きなp値
実際にも偶然の範囲内
といってよさそう



13

参考として、他の例でもp値を計算してみましょう。

もし最も極端な結果が得られて12回とも全て表だった場合、p値は0.04%となります。
直感的にも偶然とは考えにくい実験結果であるほど、p値も小さな値になることが分かります。

表が5回の場合、同様に求めるとp値は77%となります。
直感的にも偶然であろうと考えられる実験結果である場合、p値は大きな値になります。

p値を使って判断（統計的仮説検定）

- 偶然さを表す指標である **p値** を規準にして判断したい
- 医学領域では慣習的にp値≦5%の場合に「統計的に有意である」と判断して帰無仮説を却下する規準が一般的によく用いられる
- 中田君の実験で用いると「p値＝4% ＜ 5%」となるため

「フェア」を却下し「いかさま」と判断
帰無仮説　対立仮説

14

このように、p値という統計的な規準を用いて行う仮説検定を特に「統計的仮説検定」と呼ぶことがあります。医学領域では、慣習的にp値が5%より小さい場合、統計的に意味があると判断をして、帰無仮説を却下するという規準が一般的に用いられています。

中田君が実験前に、「p値が5%以下」を規準とするとプロトコールに書き、プロトコールに従って実験を行い、表10回裏2回という結果が得られれば、帰無仮説を却下して対立仮説が正しいと「統計的仮説検定」により判断することができていたといえます。

αエラー

判断する以上、エラーがある

仮説検定の結果

真実		フェアと判断	いかさまと判断
	フェアなコイン（表の確率1/2）	正しい判断	誤り

αエラー"フェアを誤っていかさまと言ってしまう"

- 事前に決めた規準が「p≦5%」である実験の場合、（もしも真実がフェアなのであれば）αエラーは5%
  - 算出したp値が4%でも、0.00001%であっても αエラーは事前に定めた規準に等しく5% になる

ここでは、実験結果から仮説検定を用いて、審判はいかさまをしていると判断しましたが、実はフェアである可能性を否定できたわけではありません。ここで用いた論理は「もしもフェアだったとしたら、このような実験結果が得られる可能性は非常に低い。よって、いかさまと判断する」ということでした。仮にフェアだったとしても、10,000回に2回の確率では12回連続で表が出ることがあるのです。

「統計的仮説検定」を使って判断しても、必ずこの判断自体にエラーがありえます。本当は「フェア」であるのに「いかさま」と判断したら誤りです。この誤りを統計学では「αエラー」と呼びます。

αエラー

仮説検定の結果

真実		帰無仮説を却下しない	帰無仮説を却下
	帰無仮説が正しい	正しい判断	誤り

- 実際上、0にすることは不可能
  - αエラーが問題であれば、状況によって変える
    - 仮説を探索する研究では大きめ、20%や30%だって構わない
    - 仮説を検証する研究では小さめ、5%を用いることが多い
  - αエラーは必ず事前に決める約束事
    - 事前（実験前）に決めないことは データを見てから都合の良いように解釈することに等しい 後出しじゃんけん、当然適切でない！

ここでは「p≦5%」と事前に決めた

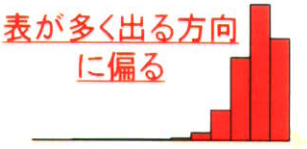
当然ながら誰もがαエラーを犯したくはないのですが、実際上αエラーを0とすることは不可能です。αエラーの大きさは、必ず事前に決めなければいけません。事後にp値を操作すると都合のよいように解釈することになるからです。

もしも90%表が出るいかさまだったら

表の出る回数	確率	90%表が出るいかさまコインでの確率
2	.0002	.28
11	.0029	.37
10	.016	.23
9	.054	.085
8	.12	.021
7	.19	.004
6	.23	.0005
5	.19	.00005
4	.12	.000003
3	.054	.0000002
2	.016	.000000005
1	.0029	.0000000001
0	.0002	.000000000001

$= (0.9)^{12}$

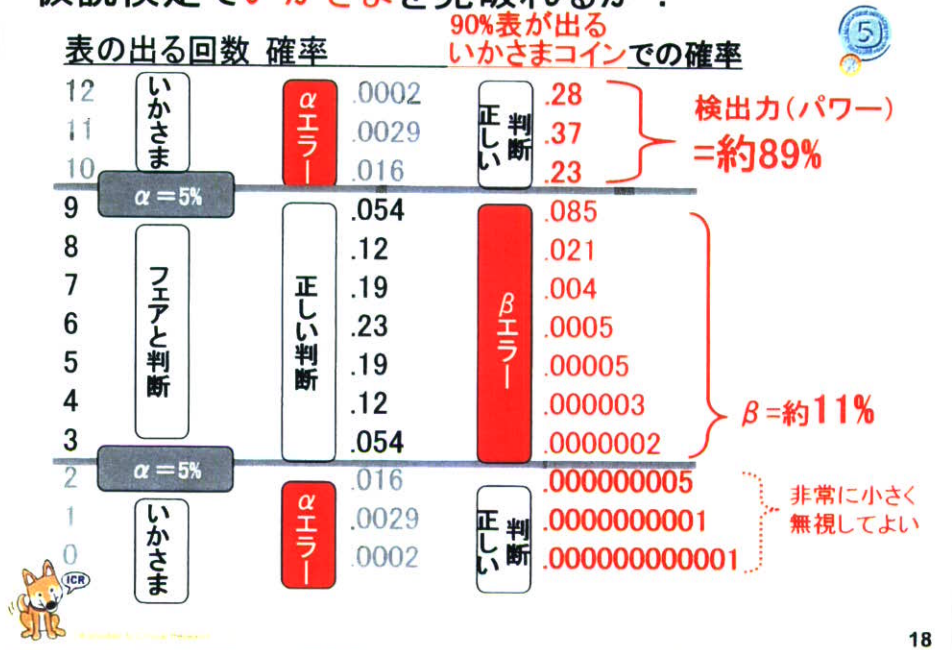
表が多く出る方向に偏る



ここまでの仮説検定では、帰無仮説つまり「フェア」が正しい場合のみを考えています。では、本当は「いかさま」が正しい場合、この実験はどういうものであったでしょうか。ここで表が90%の確率で出る「いかさまコイン」を考えます。実験をしていかさまを見破れる否かが問題となります。

表が90%出る「いかさまコイン」で同様に表の出る回数に対応して求めた確率を表に示しています。90%表が出るコインであるため、この表からも極端に表が多く出る方向に偏った確率となっていることが分かります。

仮説検定でいかさまを見破れるか？



仮説検定で用いる規準を、先ほどと同様にαエラーが5%になるように決めます。「フェア」が正しい場合の確率を、真ん中に示しています。

「いかさまコイン」の場合、正しく「いかさま」といえれば正しい判断です。反対に、「いかさまコイン」なのに「フェア」と判断してしまうことは誤りです。この誤りを「βエラー」と呼びます。

90%表が出る「いかさまコイン」に対して、この実験のβエラーを犯す確率は約11%と求められます。一方、このコインに対して正しく「いかさま」と判断できる確率は、その余事象から約89%と求められます。この確率のことを「検出力（パワー）」と呼びます。検出力とは対立仮説が正しい場合に、帰無仮説を正しく却下できる確率に対応します。

仮説検定でいかさまコインを見破れるか？

仮説検定の結果

		フェアと判断	いかさまと判断
真実	フェアなコイン（表の確率1/2）	正しい判断	誤り
真実	いかさまコイン	誤り	正しい判断

βエラー “いかさまを誤ってフェアと判断してしまう”

- 検出力：いかさまをいかさまと正しくいえる確率　1－β
  - ・中田君の実験は表が90%出るいかさまに対する検出力：89%

{ いかさまの程度
{ コインを何回投げるか　によって検出力は変わる

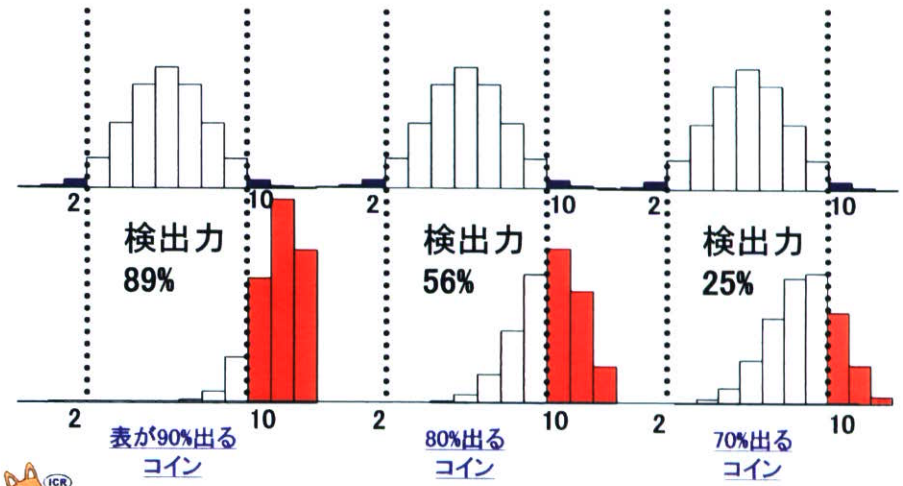
19

βエラーは「いかさま」を「フェア」と判断してしまう誤りで、検出力はいかさまを正しく「いかさま」と判断するものなので、(1－β)となります。中田君の実験では、表が90%出るいかさまコインに対する検出力は89%でした。

この例では、表が90%出るいかさまコインという設定で検出力を計算しましたが、検出力は「いかさま」の程度とコインを投げる回数によって決まります。

いかさまの程度に依存

- 有意水準5%、12回のコイントス



20

検出力は「いかさま」の程度に依存します。

これをαが5%、12回コイントスをした実験で確かめてみましょう。

いずれの場合も12回投げることから規準は同一とし、2回以下と10回以上の場合に「いかさま」と判断します。これまでと同様に考えると表が80%出るコインの場合、先ほどの表が90%出るコインに比べて偶然10回未満となることも多くなることから、中田君の行った実験の検出力は56%に下がってしまいます。つまり、いかさまの程度が少し巧妙になると、正しくいかさまといえる確率が下がるのです。同様に表が70%出るコインでは、中田君の行った実験の検出力は25%となります

臨床試験や臨床研究の場合、いかさまの程度は治療効果の大きさに該当します。同一の実験方法（患者数）であれば、効果の高い治療に対してはその治療効果をより証明をしやすく、治療がマイルドであればより証明しにくいという傾向があります。

疫学研究の場合には、真の曝露効果の大きさに比例します。

今回の実験について

- 表が90%出る（極端にひどい）「いかさま」に対して
  中田君の実験方法は90%の確率で見破ることができる

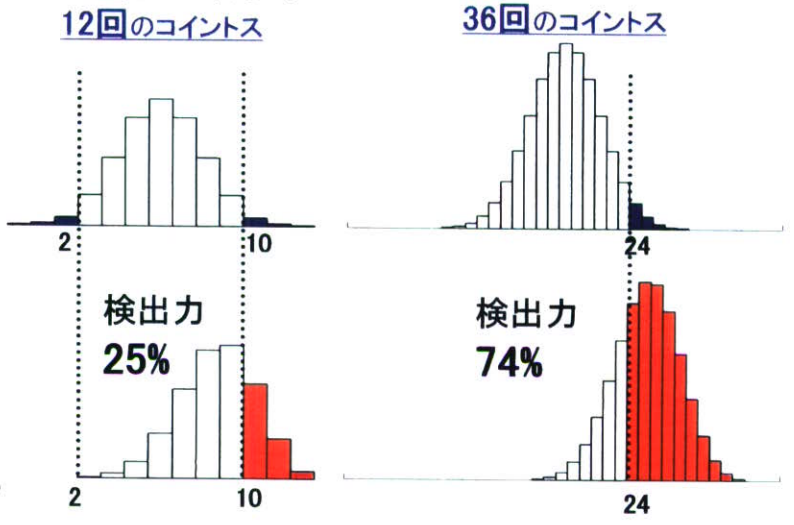
- ただし、表が70%出るような「いかさま」に対しては
  わずか25%の確率でしか見破れない
  - ・70%＝3回中2回以上も表が出る「いかさま」なのに
    見破れる可能性がこんなにも低い実験では当然困る
  - ・さて、中田君はどうすればよいか・・・

今回の実験についてまとめてみました。

表が70%出るような「いかさま」でも見破りたいと考える場合、どうすればよいでしょうか。

コインを何回投げるかに依存

・α＝5%、表が70%出るコイン



答えはコイントスの回数を増やすことです。

検出力は、先ほどの「いかさま」具合以外にも、コイントスの回数に依存します。そこで、投げる回数の違いから検出力をみてみましょう。

表が70%出るコインに対して中田君の行った実験の検出力は25%でしたが、コイントスを36回に増やすと検出力は74%に増加します。

臨床試験や疫学研究の場合、検出力は患者数や研究対象者数（サンプルサイズ）に比例します。