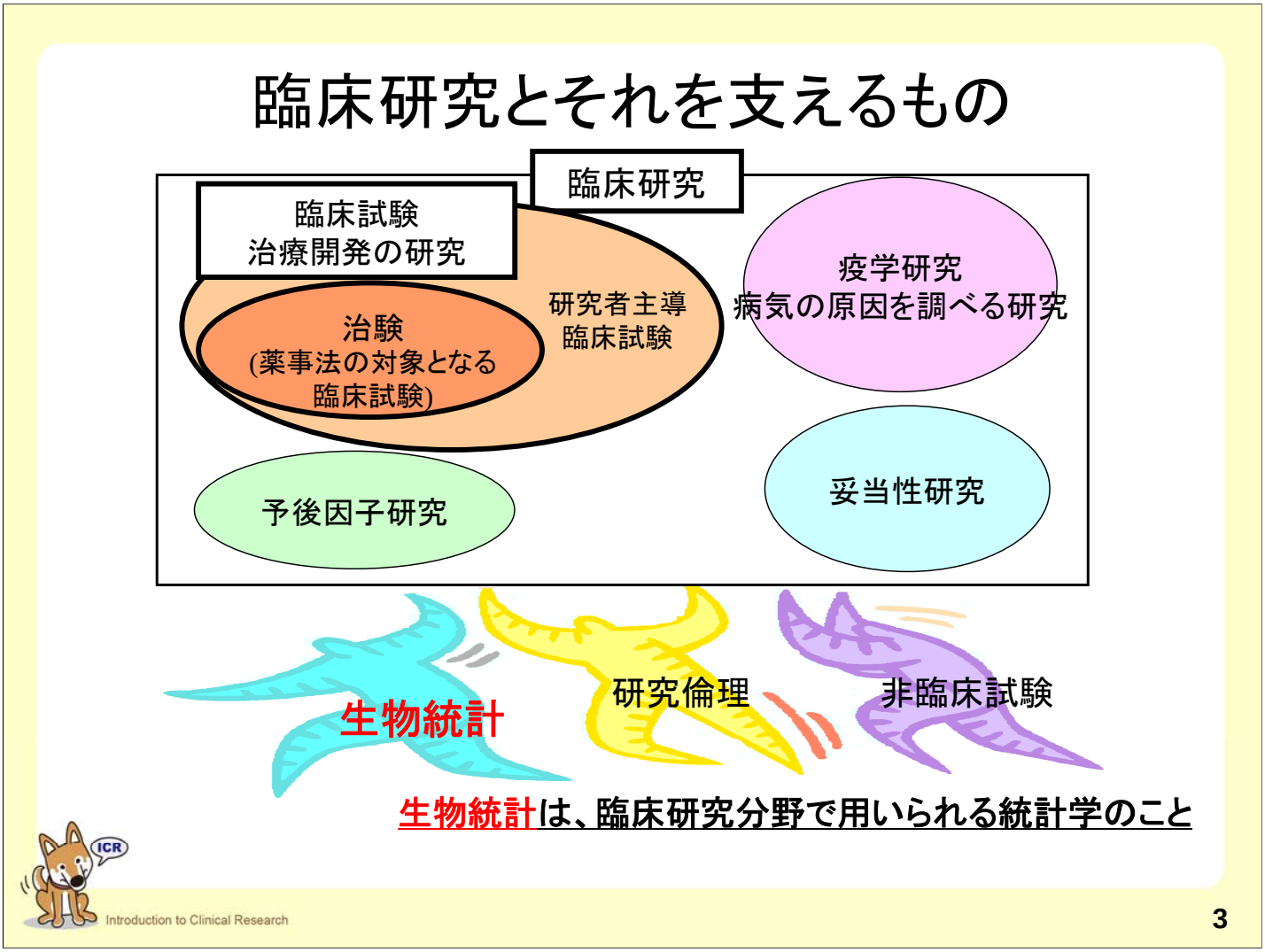
第7章
生物統計学1：仮説検定

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。



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. 

フェアなコイン？

2. 

いかさまコイン？

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

Introduction to Clinical Research

5

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
規準
いかさまと判断

- 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）

Introduction to Clinical Research

9

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！

Introduction to Clinical Research

10

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。

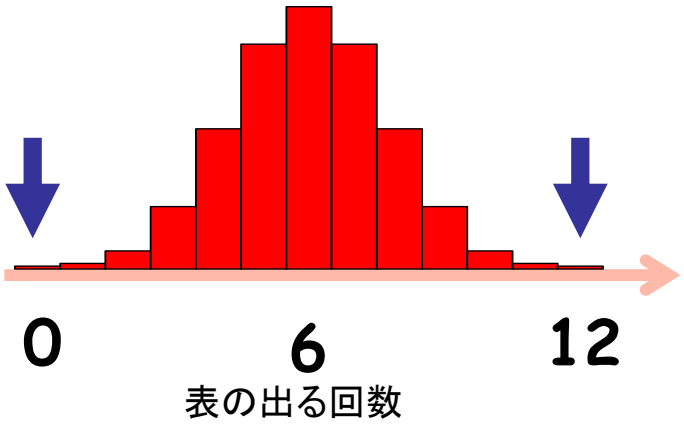
そこで、この偶然性を見積もるために統計学が必要となります。つまり、フェアなコインの場合でも表10回となることが偶然の範囲内で起こったことなのかどうかを、統計学により判断します。

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

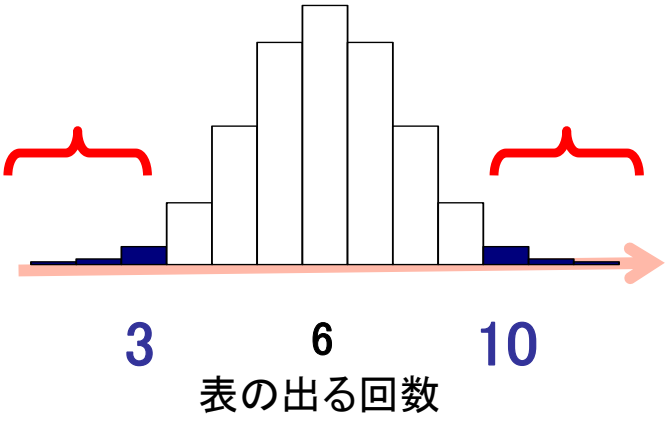
表の出る回数	確率
12	.0002
11	.0029
10	.016
9	.054
8	.12
7	.19
6	.23
5	.19
4	.12
3	.054
2	.016
1	.0029
0	.0002

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



Introduction to Clinical Research

12

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

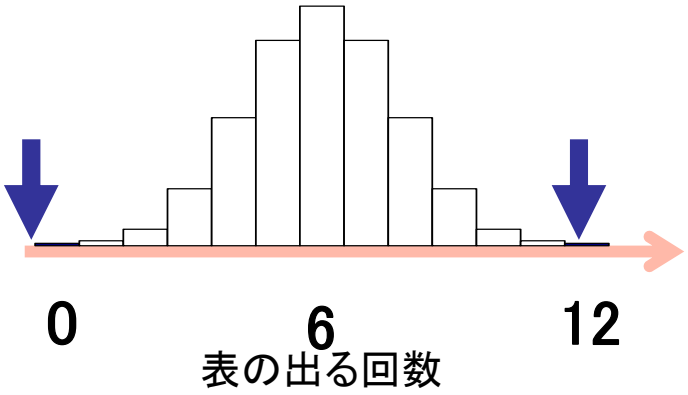
表の出る回数	確率
12	.0002
11	.0029
10	.016
9	.054
8	.12
7	.19
6	.23
5	.19
4	.12
3	.054
2	.016
1	.0029
0	.0002

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



Introduction to Clinical Research

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

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

		帰無仮説を 却下しない	帰無仮説を 却下
真実	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

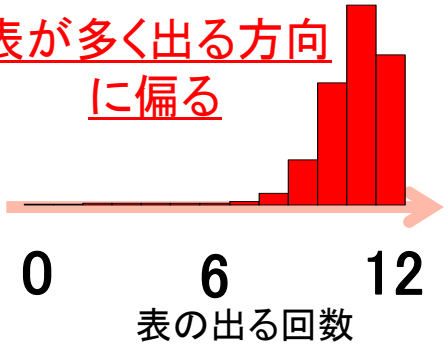
当然ながら誰もがαエラーを犯したくはないのですが、実際上αエラーを0とすることは不可能です。αエラーの大きさは、必ず事前に決めなければいけません。事後にp値を操作すると都合のよいように解釈することになるからです。

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
12	.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

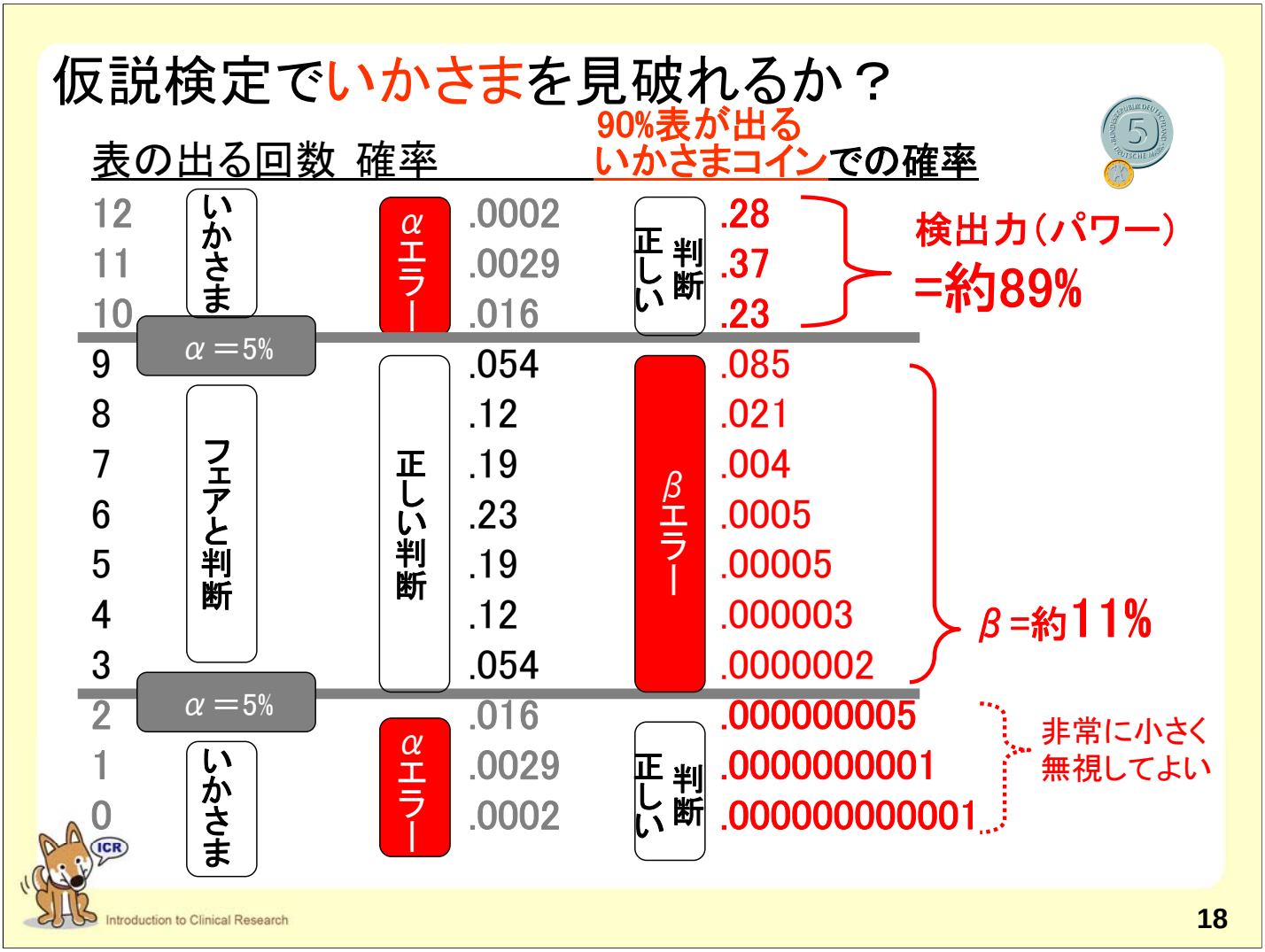


Introduction to Clinical Research

17

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

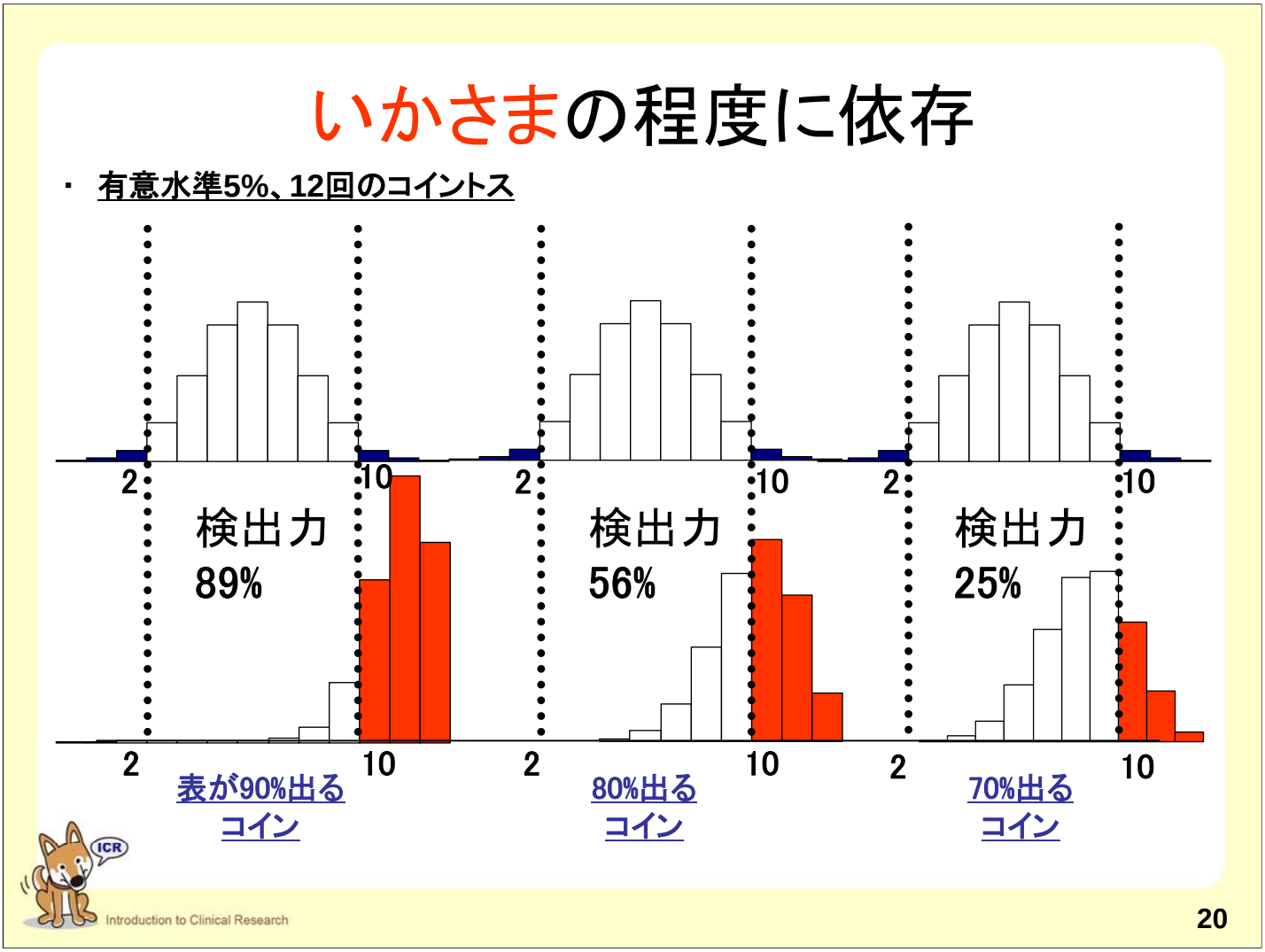
βエラー　“いかさまを誤ってフェアと判断してしまう”

- 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

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。



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

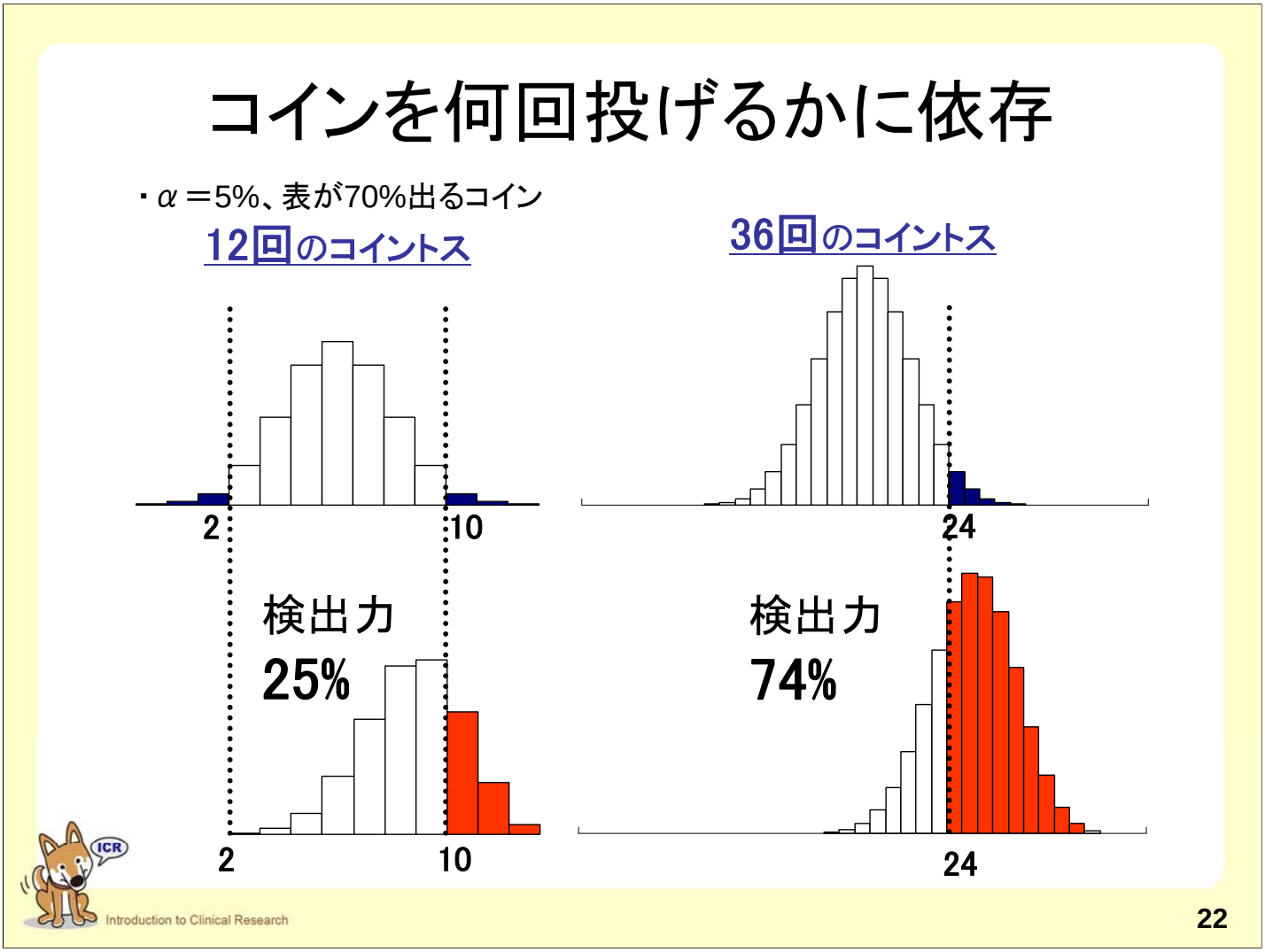
- 表が90%出る（極端にひどい）「いかさま」に対して
  中田君の実験方法は**90%の確率で見破ることができる**

- 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。



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。

検出力を踏まえて書きなおし

① 表の出る確率が1/2であるか を調べたい

② コインを 12回 投げる。この実験デザインは表が**90%**出るコインに対して検出力約**90%**である

実施する意味がある研究なのかを事前評価すべき

③ p値≦5%の場合にいかさまと判断

＝「有意水準 α =5%として仮説検定を行う」

・プロトコールにもしっかり書く！

Introduction to Clinical Research

23

検出力もふまえ、より統計的にプロトコールを書き直してみました。

また、より巧妙なコインに対して十分に高い検出力としたい場合には、
②「コインを36回投げる、この実験デザインは表が70%出るコインに対して、検出力が約75%である」
とすることもできます。

この様に、研究で用いる実験方法が適切であることを明記することは重要です。
患者さんに参加いただく臨床試験では、実施する意味がある試験かどうかを担保するため、検出力を記載します。
検出力が20～30%程度であることは実験方法が適切でないことに対応し、よって、倫理的にも適切でない可能性が強く疑われるからです。
対象者が健康な人である疫学研究においても、高い検出力を担保して研究を実施することが重要であることは同じです。

αエラーとβエラー

- **αエラー**
  - フェアなのに誤っていかさまといってしまう
  - 帰無仮説が正しいのに、対立仮説が正しいといってしまう誤り
  - 慌てもの（α-wa-te-mono）のエラー

- **βエラー**
  - いかさまなのに誤ってフェアといってしまう
  - 対立仮説が正しいのに、対立仮説が正しいといえない誤り
  - ぼんやり者（β on-ya-ri-mono）のエラー
  - 検出力は、1からβエラー（をしてしまう確率）を引き算したもの

αエラーとβエラーについてまとめると、このようになります。

本日は「中田君」を例にしましたが

- 個別の臨床研究でも基本的な考え方は同じです
- 例えば・・・
  - 「コインを投げる回数」を「サンプルサイズ」
  - 「表」を「がんが縮小した患者さん」
  - 「裏」を「がんが縮小しなかった患者さん」
    と置き換えて読めば、、、
    「奏効割合が50%と異なるか否か」を評価した臨床試験そのもの
    - 12人登録だと50%奏効する薬と90%奏効する薬なら区別できそう
    - 36人登録だと50%　　と70%　　でも区別できそう

- 各自の研究分野の言葉にも置き換えて考えましょう

ここでは、医学領域とは少し離れた中田君のコイントスを例に用いましたが、個別の臨床研究を行う上で必要となる統計的考え方は共通します。

例えば、コインを投げる回数をサンプルサイズ、コインの表をがんが縮小した患者さん、裏をがんが縮小しなかった患者さんとそれぞれ置き換えてみましょう。そうすると、奏効割合が50%と異なるか否かを評価した1群試験と考えることができます。

生物統計学の概念について、より理解を深めるためには、各自の研究分野の言葉に置き換えて考えてみることをお勧めします。