**2019 年度秋学期　統計学　第１４回**

# 分布についての仮説を検証する — 仮説検定

「統計学」の最後は，**仮説検定**（あるいは**検定**）という考え方について説明します。これは，例えば「母平均は 0 でない」「母平均は 0 より大きい」といった，母集団分布についての仮説が，適切かどうかを推測する方法です。

---

## 検定の考え方

次の例を考えてみましょう。

> 店員が「確率 50%で当たる」と宣伝しているくじがあるとします。ところが，あなたがこのくじを 10 回引いてみたところ，1 回も当たりませんでした。
>
> 店員は「運が悪かったねー」と言っていますが，あなたはどうも納得がいきません。「『確率 50%で当たる』という宣伝はウソじゃないの？」と思います。さて，店員かあなたか，どちらが正しいでしょうか？

店員の言っていることが正しいかどうかは，くじ箱を開けて中のくじを全部調べれば，確実にわかります。もちろん，そんなことはふつうはできません。しかし，そのようにして調べない限り，店員がウソをついているのか，それともあなたの運がものすごく悪いのか，結論は出せません。そこで，次のように考えてみます。

店員の宣伝では，1 回のくじ引きで，当たりもはずれも確率は $\frac{1}{2}$ だと言っています。そこで，1 回目のくじ引きと 2 回目のくじ引きが独立とみなせるのであれば，2 回続けてはずれる確率は，それぞれではずれる確率の積で，$\frac{1}{2} \times \frac{1}{2}$ となります。

同じように考えると，「くじを 10 回引いて 1 回も当たらない」確率は，$\left(\frac{1}{2}\right)^{10}$ すなわち $\frac{1}{1024}$ ということになります。つまり，店員の「確率 50%で当たる」という宣伝を信じるのであれば，「くじを 10 回引いて 1 回も当たらない」という結果になる確率は $\frac{1}{1024}$ ということになります。

確率とは，「すべての可能性のうち，どの結果になりやすいか」の度合いを表すものです。ということは，「店員の宣伝を正しいと信じる」ことは，「10 回のくじびきの結果のすべての可能性のうち，$\frac{1}{1024}$ という小さな確率でしか起きないことが，たまたま今，目の前で起きている」という考えを受け入れることになります。そんな無理のある考えを受け入れるよりも，「『確率 50%で当たる』という宣伝のほうが間違っている」と考えるほうが自然ではないでしょうか？

こういう論理で，「『確率 50%で当たる』という宣伝は間違っている」という判断を下すのが，検定の考え方です。検定の論理は基本的にはこれだけで，問題によって異なるのは，「小さな確率でしか起きないことが，今たまたま目の前で起きているなどという考えは，受け入れられない」という考えを導くための，確率の計算のしかたです。次節では，もうすこし実際的な例で，検定の使い方と確率の計算のしかたをみてみましょう。

## $t$ 分布と検定

次のような問題を考えてみます。

> 10 人の被験者に，薬 A を与えた場合と薬 B を与えた場合とで，それぞれある検査を行うと，その結果の数値は次の表の通りとなりました。このとき，薬 B は，薬 A よりも，検査の数値を高くする働きがあるといえるでしょうか。

| 被験者番号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 薬 A | 60 | 65 | 50 | 70 | 80 | 40 | 30 | 80 | 50 | 60 |
| 薬 B | 64 | 63 | 48 | 75 | 83 | 38 | 32 | 83 | 53 | 66 |

検査結果の数値自体は，人によって大きく違います。ここでいう「薬 B のほうが高くなる」というのは，それぞれの被験者において，数値がどう変化しているかを問題にしています。そこで，各被験者について，数値の差（薬 B での数値引く薬 A での数値）を求めてみます。

| 被験者番号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 薬 A | 60 | 65 | 50 | 70 | 80 | 40 | 30 | 80 | 50 | 60 |
| 薬 B | 64 | 63 | 48 | 75 | 83 | 38 | 32 | 83 | 53 | 66 |
| 差 | 4 | −2 | −2 | 5 | 3 | −2 | 2 | 3 | 3 | 6 |

差を見ても，被験者によって正だったり負だったり，ばらつきがあります。つまり，薬 B での数値のほうが高くなっている被験者もいれば，逆に低くなっている被験者もいます。ここで問題にしているのは，差の平均値をもって「薬 B のほうが高くなる」といえるかどうかです。

差の平均値は +2 で，ゼロではありませんから，差の平均値をもって「薬 B のほうが高くなっている」かといえば，なっていることはなっています。問題は，その差が，偶然のために生じたものではなく，本質的な差であるかどうかです。本質的かどうかなど，どうすればわかるのでしょうか。

この疑問に対して，次のように考えます。

- 「母集団（ここでは，世界のすべての患者）について『薬 A と薬 B での差』を求めると，平均は 0 になる」と仮説を設定する。つまり，「本質的な差はない」という仮説を設定する。
- 被験者は，母集団から無作為抽出された，10 人からなる標本と考える。
- このとき，被験者 10 人での「薬 A と薬 B での差」の平均値が，わずかな確率でしか生じないほどの大きな差であるなら，この差は「偶然によって生じたものではない」と考える。
- すなわち，「本質的な差はない」という当初の仮説は誤り，と結論する。

このような考えで，「母集団に関する仮説が正しいとすると，いま標本について得られている結果は，偶然とは考えにくい」→「母集団に関する仮説は間違っている，と考える」という推論を，**仮説検定（検定）**といいます。次節以降で，最初にあげた例題について，どのように検定を行うかを考えます。

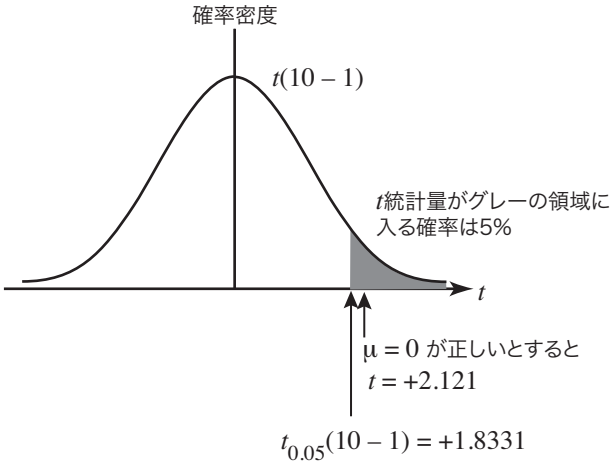


図 1: $t$ 統計量.

**片側検定**

では，例題の，薬 B での結果のほうが，薬 A での結果よりも「本質的に」高いといえるか，を考えます。ここで，母集団全体での「薬 A と薬 B での差」は，平均 $\mu$ の正規分布にしたがうと考えます。そうすると，前回の「$t$ 分布にもとづく区間推定」で説明したように，

- 標本サイズを $n$（例題では 10）
- 標本平均を $\bar{X}$（例題では，10 人の被験者における差の平均値で，$+2$）
- 不偏分散を $s^2$（例題では，10 人の被験者についての不偏分散で，計算すると 8.89）

とすると，$t$ 統計量

$$t = \frac{\overline{X} - \mu}{\sqrt{\frac{s^2}{n}}} \tag{1}$$

は，自由度 $n-1$ の $t$ 分布にしたがいます。

さて，前節の「検定の考え方」の第 1 項にある「母集団について『薬 A と薬 B での差』を求めると，平均は 0 になる」という仮説を考えます。この仮説は，つまり「$\mu = 0$」ということになります。

この仮説が正しいとすると $\mu = 0$ で，また例題では $n = 10,\ \bar{X} = +2,\ s^2 = 8.89$ です。これらの数値を (1) 式に代入すると，$t$ 統計量の値は $t = +2.121$ となります。

一方，例題では，$t$ 統計量は自由度 $(10-1) = 9$ の $t$ 分布にしたがいます。このとき，$t$ 統計量が上側 5%点よりも大きくなる確率は 5%で，自由度 9 のときの上側 5%点，すなわち $t_{0.05}(9)$ は，数表から $+1.8331$ となります。これらを，$t$ 分布の確率密度関数のグラフを使って図示したのが，図 1 です。グラフのグレーの部分の面積が 5%で，$t$ 統計量が上側 5%点よりも大きくなる確率が 5%であることを示しています。

この図からわかることは，「$\mu = 0$」という仮説が正しいとき，$t$ 統計量は「確率 5%でしか起きないほど大きな値」になる，ということです。つまり，「母集団について『薬 A と薬 B での差』を求めると，平均は 0 になる」という仮説が正しいとすると，$t$ 統計量が偶然このような大きな値になる確率は 5%しか

なく，偶然とは考えにくい，ということになります。したがって、この仮説は正しくなくて，いまの標本についての差の平均 (+2) は「本質的な差」と考える方が自然，ということです。

では，この仮説が間違いであれば，かわりにどんな結論が得られるのでしょうか。(1) 式を見ると，$\mu$ が 0 でなくもっと大きければ，$t$ 統計量はもっと小さくなり，図 1 のグレーの領域から外れます。つまり，「確率 5%でしか起きないほど大きな値」ではなくなります。$\mu$ が 0 でなくもっと大きい，すなわち「$\mu > 0$」とは，「薬 B での結果のほうが数値が高い」ということです。つまり，「薬 B での結果のほうが数値が高い，と考える方が自然だ」というのが，この例題に対する答えとなります。

### 検定の言葉

以上のように，この例題に検定を使って答えました。ただ，検定には独特の用語があり，統計学の教科書ではその用語が使われています。ここでは，例題を使って，用語について説明します。

例題で，「$\mu = 0$」という仮説は「間違っている」と判断されました。このときの「$\mu = 0$」という仮説を**帰無仮説**といい，$H_0 : \mu = 0$ と表します[1]。また，帰無仮説を「間違っている」とした判断を，**帰無仮説を棄却する**といいます。さらに，帰無仮説を棄却した結果，正しいと判断した「$\mu > 0$」という仮説を**対立仮説**といい，$H_1 : \mu > 0$ と表します。この判断を，**対立仮説を採択する**といいます。

上の推論では，「確率 5%でしか起きないことが，偶然起きていると考えるのは不合理」と考えています。つまり，5%の確率でしか起きないことが起きたということを説明する時，「偶然起きた」という説明ではなく，帰無仮説が間違っているという「必然」によって起きた，という説明のほうが合理的だ，と考えているのです。偶然ではなく必然的に何かが起きることを「**有意**である」といい，この「5%」を**有意水準**といいます。

例題では，帰無仮説が正しいとするとき，「$t$ 統計量が $-t_{0.05}(10-1)$ 以上である」ならば帰無仮説を棄却する，という推論をしました。つまり，図 1 のように，「帰無仮説が正しいとするとき，$t$ 統計量がここに入ったら，帰無仮説を棄却する」という区間（グレーの部分）が，$t$ 分布の確率密度関数で片側（右側）にあります。その意味で，今回のやりかたの検定を**片側検定**といいます。

なお，上の「帰無仮説が正しいとするとき，$t$ 統計量がここに入ったら，帰無仮説を棄却する」という区間のことを**棄却域**といい，棄却域を表すのに用いる統計量（ここでは $t$ 統計量）を**検定統計量**といいます。また，検定統計量の値が棄却域に入ることを，**棄却域に落ちる**という表現をします。

### 両側検定

では，この例題で，設問を「薬 A を与えた場合と薬 B を与えた場合で，検査の数値に違いがあると言えるでしょうか」に変えてみます。最初の設問との違いは，最初の設問が「薬 B での数値のほうが高くなる」かどうかを知りたいのに対して，今度は「薬 B での数値のほうが，薬 A での数値よりも，高いか低いか，いずれにしても本質的な差がある」かどうかを知りたいわけです。

この場合，帰無仮説はさきほどと同じ「$\mu = 0$」ですが，それが棄却されて得られる対立仮説は，「$\mu$ はもっと大きい」だけではなく，「$\mu$ はもっと小さい」という場合も考える必要があります。つまり，対立仮説は「$\mu \neq 0$」となります。

帰無仮説「$\mu = 0$」が棄却された時に採択される対立仮説が「$\mu$ はもっと大きい」と「$\mu$ はもっと小さい」の両方ですから，この検定では，帰無仮説が正しいとしたときに，$t$ 統計量が「大きすぎる」ときも「小さすぎる」ときも，帰無仮説を棄却します。したがって，この検定の棄却域は，図 2 のように，確率密

[1] $H$ は，hypothesis（仮説）という英語の頭文字です。

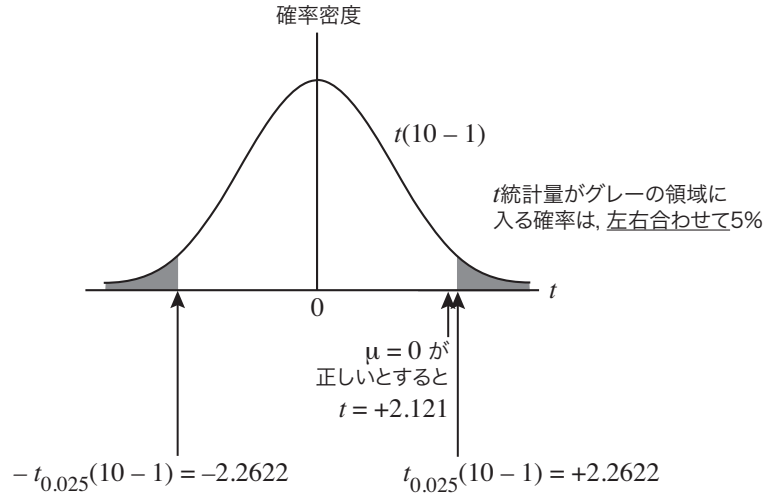


図 2: 両側検定.

度関数のグラフの両側にあります。有意水準を，前の例と同じく 5%とすると，$t$ 統計量が棄却域に入る確率は両側合わせて 5%で，左右それぞれ 2.5%ずつとなります。つまり，帰無仮説が正しいとするとき，

> $t$ 統計量が，上側 2.5%点より大きいか，または，下側 2.5%点より小さい

ときに，帰無仮説を棄却します。このやり方の検定を，**両側検定**といいます。

帰無仮説「$\mu = 0$」が正しいとするとき，$t$ 統計量の値は，前の片側検定の例と同じく，+2.121 です。一方，上側 2.5%点 $t_{0.025}(9) = +2.2622$，下側 2.5%点は $-2.2622$ ですから，$t$ 統計量は上側 2.5%点と下側 2.5%点の間にあり，帰無仮説「$\mu = 0$」は棄却されません。したがって，対立仮説「$\mu \neq 0$」は採択されず，「薬 A を与えた場合と薬 B を与えた場合で，検査の数値に違いがある」とはいえない，ということになります。

#### 帰無仮説が棄却されないときは

「帰無仮説が棄却されない」ときは，その理由は「帰無仮説が正しい ($\mu = 0$) とするとき，いま得られているような $t$ 統計量が得られる確率は，非常に小さい（5%）とまではいえない」ということになります。したがって，

> 「帰無仮説が間違っているかどうかはわからない」「対立仮説が採択できるかどうかはわからない」

という結論を導かなくてはなりません。今回の例でいえば，帰無仮説が棄却されなかった場合は，「$\mu = 0$ でないとはいえない」「$\mu = 0$ でないとまで断言する自信はない」という結論になります。

注意しなければならないのは，あくまで，「いま得られているような $t$ 統計量が得られる確率は，非常に小さいとまではいえない」のであって，「確率が大きい」のではない，ということです。したがって，帰無仮説が棄却されなかったときに，「帰無仮説が正しい」「対立仮説は間違っている」という結論が得られるわけではありません。今回の例でも，「$\mu = 0$ である」「薬 A を与えた場合と薬 B を与えた場合で，検査の数値に違いはない」などと答えてはいけません。つまり，

帰無仮説を棄却しない
＝　× 帰無仮説を採択する
　　◯ 対立仮説を採択するべきかどうか断言できない

ということです。なお，「帰無仮説を棄却すべきなのに棄却しない」という誤りを**第２種の誤り**といいます[2]。

### 有意水準について

ここまでの例で，有意水準は 5%としていました。有意水準の値は，5%あるいは 1%がよく用いられます。

有意水準の値は，検定をする人の「大胆さ・慎重さ」の程度を表しています。

有意水準が大きい (5%) ときは，帰無仮説が仮に正しいとしても，いま起きている現実（$t$ 統計量の値が +2.121）が起きる確率が 5% と「そこそこ小さな確率」であれば，「そんなことが起きるはずがない，帰無仮説は間違っている」と結論します。はっきり物を言う態度ではありますが，帰無仮説が実は正しいときでも「間違っている」と断言してしまう可能性があります．大胆ですが，勇み足も多い，というわけです。

有意水準が小さい (1%) ときは，いま起きているような現実が起きる確率が，1%と相当小さくないと，「まあそんなことも起きるかもしれない，帰無仮説は間違っているとは言い切れない」となり，結論を出しません。慎重ですが，煮え切らない態度ということになります。

---

### 片側検定と両側検定

#### どうもおかしい…

ところで，ここまで見てきた「片側検定」と「両側検定」，おかしくないでしょうか？ 有意水準が 5%なのは同じなのに，

- 「薬 B での数値のほうが高い」かと聞かれたら，「高いと言える」と答える（片側検定）
- 「薬 B での数値と薬 A での数値に違いがある」かと聞かれたら，「違いがあるとは言えない」と答える（両側検定）

という結論になりました。「高いと言える」のに「違いがあるとは言えない」というのも変な話です。

#### 実はおかしくない

これは，おかしくないのです。片側検定と両側検定では，「聞いている」こと，つまり検定している内容が違うのです。

検定とは，帰無仮説での想定（例題では，$\mu = 0$，すなわち「薬 B での数値と薬 A の数値に本質的な差はない」）が，現実にデータを調べた結果（つまり標本，あるいは標本から求めた標本平均などの値）と食い違っているかどうかを検査しています。そしてそのような食い違いが，確率 5%でしか起こらな

[2]第２種の誤りを，俗に「ぼんやり者の誤り」といいます。第２種の誤りの確率をしばしば $\beta$ で表すことにかけています。

いような，つまり偶然とは言えない（有意な）食い違いのとき，帰無仮説での想定は誤りとして，帰無仮説を棄却します。

**両側検定の場合**

両側検定は，帰無仮説が標本と食い違っているかどうかだけを検査しています。ですから，帰無仮説での想定が，標本に比べて，大きい方に食い違っていても，小さい方に食い違っていても，帰無仮説を棄却します。今回の例でいえば，帰無仮説でいう「$\mu$=0」が，標本平均に比べて大きすぎても小さすぎても，帰無仮説を棄却します。すなわち，「標本について，薬 B での数値が，薬 A での数値よりも著しく大きい（例えば，標本平均が +10）」場合でも，「著しく小さい（例えば，標本平均が −10）」場合でも，いずれも「薬 B での数値と薬 A の数値に本質的な差はない」という帰無仮説を棄却します。

**片側検定の場合**

これに対して, 片側検定は, 帰無仮説が標本に比べて, 大きすぎるか, または小さすぎるかのどちらか一方を検査します。ですから，帰無仮説での想定が，標本に比べて「ある一方向に」食い違っているときだけ帰無仮説を棄却します。例題では，対立仮説が「$\mu > 0$」，すなわち「薬 B での数値のほうが，薬 A での数値より高い」という片側検定をしています。つまり，帰無仮説の「$\mu = 0$」が，標本平均（例題では +2）に比べて小さすぎると言えるかどうかだけを検査していますから，帰無仮説の「$\mu = 0$」が標本平均に比べて小さすぎるときだけ，帰無仮説を棄却します。

**片側検定は，調べていないことは「見逃す」**

では，もし例題で，標本平均が例えば「−10」，つまり標本について「薬 B での数値が，薬 A の数値よりも著しく小さ」くて，帰無仮説の「$\mu = 0$」という想定が標本平均に比べて大きすぎる時はどうなるのでしょうか？

両側検定では，この場合も帰無仮説を棄却します。しかし，対立仮説が「$\mu > 0$」という片側検定では，帰無仮説を棄却しません。この場合も，帰無仮説の「$\mu = 0$」が標本平均と大きく食い違っているにかかわらず，片側検定はそれを見逃し，「帰無仮説を棄却できない」と答えてしまいます。それは，「$\mu = 0$ は標本平均に比べて大きすぎるかどうか」は，この片側検定では検査の対象ではないからです。たとえこれが，「$\mu = 1000$」つまり「薬 B での数値が，薬 A の数値よりも，本質的に 1000 高い」などというとんでもない帰無仮説でも，「そんなことは今検査していることではない」といって棄却しないのです。

**同じ「有意水準 5%の検定」でも**

片側検定では，棄却域は図 1 のように「片側で確率 5%」に対応する部分です。一方，両側検定では，棄却域は図 2 のように「両側それぞれに確率 2.5%」に対応する部分です。同じ「有意水準 5%の検定」でも，片側だけをみれば，片側検定のほうが棄却域が広くなります。そのために，今回の例題のような，片側検定と両側検定の結果が一見矛盾するようなことがおきます。

片側検定の例題で，「薬 B での数値は，薬 A での数値よりも高い」かどうかを検定しました。この検定をするのは，あらかじめ「薬 B での数値は，薬 A での数値よりも高いだろう」という目論見があるからです。仮に，逆の「薬 B での数値が，薬 A での数値よりも低い」という結果になっても，それは見逃してもかまわないので，そのぶん「大胆」な検定を行います。

一方，両側検定の場合は，「薬 B での数値は，薬 A での数値よりも高い」「薬 B での数値が，薬 A での数値よりも低い」のどちらの場合も見逃してはいけないので，同じ有意水準でも「慎重」な検定とな

ります。

### くじびきを例にして考えてみる

片側検定と両側検定の違いを，「くじびき」を例にして考えてみましょう。くじをひくほうの立場からすると，「当たり確率は 50%」と称するくじが「10 回ひいて全部はずれ」れば不満です。しかし，「10 回ひいて全部当たり」の時は，「当たり確率は 50%」というのは正しくないような気はしますが，得をしたのですから，別に不満は持ちません。

一方で，賞品を出すほうの立場に立てば，逆に「10 回ひいて全部当たり」の時は賞品を皆持っていかれて不満ですが，「10 回ひいて全部はずれ」でも，客に「残念でしたね」というだけで，とくに不満は持ちません。

こういうふうに，「当たる確率は 50%」という帰無仮説と現実の当たり数を比べて，現実の当たりが「少なすぎる」という不満，あるいは「多すぎる」という不満の，どちらかだけを検査するのが片側検定です。

ところが，このくじびきを主催している商店街の商店会長からすると，「あそこのくじびきは何かおかしい」という噂が流れると困ります。ですから，現実の当たりが「少なすぎる」ときも「多すぎる」ときも不満です。この両方の不満をとりあげるのが両側検定で，つまり「くじびきが双方にとって公正かどうか」を問題にすることになります。

大事なのは，「どちらの検定をするかは，検定の目的に沿って，データを調べる前に決める」ことです。データを見てから，帰無仮説が棄却されそうな検定を選んではいけません。それは，アンフェアなやりかたです。

---

### 検定はどんなときにするものなのか

有意水準 5%の検定では，帰無仮説が仮に正しいとするとき，確率 5%でしか起きないはずのことが起きていることになってしまうのなら，帰無仮説を棄却します。

しかし，「確率 5%でしか起きないはずのこと」は，言い換えれば確率 5%で起きるのであって，確率ゼロではありませんから，それが偶然起きることはあるはずです。ですから，例えばここまでの例題で，母平均 $\mu$ が本当に 0 である，つまり「母集団について，薬 A と薬 B での検査結果の差が平均 0 である」という帰無仮説が正しいときでも，得られた標本が偶然母平均から大きくはずれていて，その結果帰無仮説を偶然棄却してしまうことが，確率 5%で起きます。これは間違った判断ですが，このような間違いをする確率が 5%であるわけです。このような間違いを**第 1 種の誤り**といいます[3]。つまり，

> 帰無仮説が本当に正しいとしても，有意水準 5%の仮説検定を何度も行うと，そのうち 5%の場合では第 1 種の誤りを犯して棄却し，採択すべきでない対立仮説を採択してしまう

ことになります。

ですから，同じ現象について何度もデータを集めて，同じ帰無仮説について検定を繰り返し，たまに対立仮説が採択されても，直ちに「帰無仮説は間違っている」とはいえません。例えば，「血液型と性格

[3]第 1 種の誤りを，俗に「あわて者の誤り」といいます。第 1 種の誤りの確率（＝有意水準）をしばしば $\alpha$ で表すことにかけています。

に関係はない」という帰無仮説について何度もデータを集めて検定を行い，たまに「血液型と性格に関係がある」という結論が出ても，直ちに「やっぱり血液型と性格に関係がある」ということにはなりません。何度も検定を行うと，帰無仮説が間違っていない場合でも，たまに対立仮説が採択されるのは，むしろ自然なことです。血液型と性格の問題でいえば，ごくたまに「血液型と性格に関係がある」という結論が出る程度であれば，「血液型と性格に関係があるとは今のところ言えない」というのが，科学的態度です。

では，検定の結論は結局何を言っているのでしょうか？ それは，

> 私は，帰無仮説は間違いだ，と判断する。
> ただし，私は 100 回中 5 回はウソを言う（第 1 種の誤りを犯す）人間である。
> 私が今回，本当のことを言っているのか，ウソを言っているのか，それは誰にもわからない。

というのと同じことです。

この程度のことしか言っていないのに，検定にはどういう意味があるのでしょうか？ それは，検定とは，少ない数のデータしか調べられず，しかもそれを 1 度だけしか調べられないときに，「それだけのデータからでも十分な確信をもって述べられる疑いだけを述べる」方法ということなのです。何度も検定できるほどデータを集められるのなら，検定を用いるのは不適切です。

---

### 今日の演習

下の各質問に，各々 100 文字以内で簡潔に答えてください。

1. 仮説検定とは，どういう場合に何をすることですか。
2. 片側検定と両側検定の 2 種類の検定がありますが，どういう問題のときにどちらを選べばよいのでしょうか。