統計検定講習会

桜美林大学 リベラルアーツ学群 森 厚

2019年 3月 24日

目 次

1　はじめに —実験科学と統計検定—

実験科学

実験科学の方法については、既に皆さんはいろいろな場所で勉強しているかもしれません。そして、教えられた状況によって、違った説明のされ方もしたかもしれません。ここで実験科学について、私なりに、私が個人的に好きな言葉を通じてお伝えしたいと思います。その言葉とは、朝永振一郎博士の言葉です[1]。

ふしぎだと思うこと
　　これが科学の芽です。
よく観察してたしかめ
そして考えること
　　これが科学の茎です。
そうして最後になぞがとける
　　これが科学の花です。

疑問提示型実験

世の中には、人間が「なぜだろう？」と思うこと (なぞ) が沢山あります。残念ながら、私たちは日常生活を当たり前のもの（常識）と考えて、常識から一歩踏み出すことが難しいために、「なぜだろう？」と思うこと自身難しいときがあります[2]。だから、「なぜだろう？」と思うような現象を明確に示すことが、ときどき必要です。実験はそのためにも使われます。ここではそのような実験を「疑問提示型実験」と呼ぶことにしましょう。

仮説検証型実験

別種の実験もあります。なぞは、考えるだけで解けるわけではありません。先ほどの言葉で、「最後になぞが解ける」の前には、考えるだけでなく「よく観察してたしかめ」とありました。この部分も実験を指していると考えられます。

例えば、次のような疑問があったとしましょう。

【疑問】　：　眼鏡に指紋がつくと白く汚れるのはなぜか。指は白くないのに。

これに対して、これが理由ではないか、と考えるような仮説を立ててみます。例えば

【仮説】　：　皮脂は白い。指についているときには、薄いので透き通って見えるのである。

といった具合です。これを確かめるためにはどうしたらいいでしょうか。そこで、実験することになります。もしも、仮説が正しかったら、どうなるだろうか、と考えます。

【検証】　：　皮脂を沢山集めてみる。その色が白く、かつ、指に載せても白いことがわかったら、仮説は支持される。

---

[1] 筑波大学 朝永記念室 http://tomonaga.tsukuba.ac.jp/kagakunome/index.htm (最終参照日 2019-03-18)

[2] 「そんなことないよ！」と思う人は、「上と下はどうやって決まっているのか」とか、「なぜ、金星の地表気温は地球よりも高いのか」とかいったことを考えながら「立方体地球」のビデオを見てみましょう。
Cubic Earth もしも地球が立方体だったら https://www.jss.or.jp/fukyu/cubicearth/ (最終参照日 2019-03-18)

といった具合です[3]。これを確かめるための実際に行うものが実験です。このような実験は、先ほどの実験と区別して「仮説検証型実験」と呼ぶことにしましょう。

どちらの実験も、科学を推し進める上では、とても重要です。

統計検定

ところが、時々、困ったことが起こります。例えば、薬の治験です。国が薬として認めるためには、その薬の効果と安全性を、実際の患者に投与して調べる必要があります。この調査が治験です。

例えば、患者 100 人を 50 人ずつに分け、一方には治験薬を投与し、他方には、何の効果もないと考えられる粉末を薬に見立てて投与し、患者に改善が見られたかを調べたとしましょう。治験薬を投与した 50 人のうち、30 人に改善が見られたとします。一方、投与しなかった方でも、50 人のうち、25 人に改善が見られたとします。これで薬の効果はあったと言えるでしょうか。

治験薬を投与した 50 人は、もともと、病気が直りやすい体質の人が、偶然、多かったのかもしれません。だから、別の 100 人で同じ実験をすると、結果が違うかもしれません。同じ人に 2 度、同じ病気になってもらって、実験するのは非人道的です。また、2 度目は直りやすい体質になっているかもしれません。

このような時に、統計検定が使えます。これから、統計検定について考えていきましょう。

2　二項検定 —テレパシーはあるか？—

2.1　はじめに—シェルドレイクの実験

皆さんは、電話が鳴ったときに、「きっと○○さんからの電話だ」と思うようなことはありませんか？電話だけではありません。洋の東西を問わず、「虫の知らせ」というものが昔から記述されています[4]。このテレパシーのようなもの (以下、テレパシーとします。) があるのかどうかを確かめるために、イギリスの生物学者ルパート・シェルドレイクは、ある実験を思いつき、実行しました[5]。

実験では、被験者 5 人を招きます。1 人は電話を受ける役です。残りの 4 人は電話をかける役です。電話をかける 4 人は、電話を受ける人の親しい人です。4 人のうち誰が電話をかけるかは、サイコロを振って決め、電話をかけます。電話を受ける人は離れた場所にいるので、誰から電話がかかってくるのか分からないようになっています。電話を受けるとき、電話が鳴ると相手が誰であるかを予想し、それを言います。言ってから通話を開始します。当たったかどうかを記録していきま

[3] これでは確かめたことにならないのではないか、もっと他の確かめ方があるのではないか、と考える人もいると思います。その考えは大切にしましょう。検証は 1 つだけでは不十分です。様々な検証を繰り替えして、初めて、仮説は正しいと認められるようになるからです。だから、1 つの検証だけで満足してはいけません。
実際、仮設が正しいとし、検証として、「黒い紙を触りつづけ、その紙が白くなるか確かめる。白くなれば仮設が支持される。」といったことも考えられます。実際どうでしょうか？

[4] 例えば、イギリスのコナン・ドイルが書いた「シャーロック・ホームズ」シリーズの「The Man with the Twisted Lip(唇の曲がった男)」では、

> There is so keen a sympathy between us that I should know if evil came upon him.
> (私たちは、深い共感があるので、悪いことがあれば分かるはずです。)

という記述があります。

[5] “Videotaped experiments on telephone telepathy”, Sheldrake, R. & Smart, P. The Journal of Parapsychology, 2003 vol: 67 (1)

す。不正がないように、ビデオカメラで撮影しながら実験します。そのような実験を 271 回行った結果、電話してきた人が当たったのは 122 回でした。詳細は論文を読んでください。

まず初めに、この実験自身についてのコメントです。いわゆる超常現象に関しては、ここで書いたテレパシーも含めて、多くの科学者が否定的です。これまでの科学的な知見とは整合性が無いからです。しかし、ふしぎだと思ったことをよく観察してたしかめることは、科学の本道です。シェルドレイクの実験は、仮説検証型の実験ではなく、科学の本道に沿った疑問提示型の実験だと、私は考えます。本質的に、なぜそうであるのかを説明できるような実験ではなく、疑問を形を整えて提示するタイプの実験だということです。皆さんもこのような実験をすることは、とてもいいことだと思います。ただ、「成果」という観点では、類似の実験を行っても成果が見込めないことも強調しておきたいです。こうした現象については、これまでの科学的な知見と、何ら接点が無いからです。そして、このような実験には、常に、何らかの考え落としがあり、人々の関心はそちらに向かってしまいます。巧妙なトリックを使ったのではないか、とか、です。そうなると、残念ながら、ますます「成果」としては乏しくなってしまいます。

話を戻しましょう。もしも、電話を受ける人が適当に答えていたら、1/4 の回数で当たることが見込まれます。271 回を 4 で割ると約 68 回です。しかし、ぴったり 68 回になることはむしろ少なく、ばらつきがあると予想されます。だから、122 回が 68 回よりも多い、というだけでは何とも言えません。実際、授業でこの話をすると、多くの学生が「この実験では、たまたま、本当に偶然で、多めに正解だったのでないでしょうか。」と反応します。皆さんはどう考えますか？

話し合ってみよう

- この実験結果は、テレパシーが無くても、たまたまなったと言えるだろうか？

- 実験結果を増やしたら、もっと確実なことは言えるのだろうか？

- この実験結果からだけで、何か言うことはできないだろうか？

2.2　検定の考え方

このような問題に答えを出してくれるのが統計学の検定という考え方 (統計検定) です。細かい用語は後回しにして、シェルドレイクの実験を題材に、考え方の概要を書きます。

> テレパシーは無いことを前提にして、つまり、当たる確率が 1/4 であること前提にして考えてみよう。これを 271 回、試して、122 回以上当たる確率を考える。1 回の確率が 1/4 だから計算できそうだ。計算の結果、271 回のうち 122 回以上当たるのは、それは滅多に起こらないとわかる。すると、確率 $1/4 = 0.25$ で起こるとした前提 (テレパシーが無いとした前提) が間違っていると考えるべきだよね。

もっと簡単に書くと、

> テレパシーが無いとすると、滅多に起きないことが起きちゃったので、テレパシーが無いと考えるのは無理があるよね。

ということです。

やや、ひねくれた論法を使っていることに気をつけてください。本当に主張したいのは「テレパシーがある」です。ところが、これを主張するために、「テレパシーが無いとして実験結果が実現する確率を計算する。その確率が低いのでテレパシーが無いと考えるのは難しい。」という論法を使っています。

このような論法を使っているために、用語が多数登場するのも統計検定の特徴です。次のような言葉が普通に使われますので、よく覚えておきましょう。

試行
: 電話をかけてきた相手を当てようとすることを試行という。「これを 10 回繰り返してみる」ことも試行という。

事象
: 試行の結果のこと。「これを 10 回繰り返して 4 回当たった」というような複合的なものも事象という。

帰無仮説
: テレパシーが無い（確率 1/4 で当たる）としても実験結果が起こるとする仮定のこと。「否定したい仮説」「将来、否定される仮説」という意味で名前がついている。

対立仮説
: 帰無仮説に対立する仮説。テレパシーがある（当たる確率は 1/4 以上である）とする仮説。

  対立仮説を

  - テレパシーがある（当たる確率は 1/4 以上である）とする仮説。
  - テレパシーがある（当たる確率は 1/4 ではない）とする仮説。

  のどちらにするかで議論が変わってくるので注意が必要である。

有意水準 (危険率)
: 「滅多にない」ことかどうかを判定する基準のこと。「帰無仮説の下では 5% でしか起きないことが起きたんだから、前提とした帰無仮説が疑われるよね」と言うための基準のこと。5% とか 1% を用いることが多い。もちろん、小さい値が推奨される。

p 値

帰無仮説の下での事象が生じる (実験結果が起こる) 確率のこと。この確率を計算して、危険率と比較する。

「テレパシーが無いんだとしたら、こんなことは滅多に起こらないので、あると考える方が自然じゃないの？」という話の中で、それでも、p 値の確率で起こりうるので、「テレパシーがある」と判断するのには危険がある。そこで、上述の有意水準のことを危険率ともいう。

統計検定

次のような議論のし方を統計検定という。

> 電話をかけてきた相手が当たる確率が 1/4 であることを帰無仮説として、271 回の試行に対して 122 回の当たった事象が発生したとする。122 回以上当たる事象が発生する p 値は、1% よりも小さいので、危険率 1% で帰無仮説は棄却された。

最後に強調しておきたいことは、結局、完全に断定することはできないのだ、ということです。しかし、確率を計算できるのであれば、滅多に起きない事象が発生したのだから、帰無仮説は否定されるべきだ、と主張することができるわけです。

次に、確率が計算できるのか、確かめてみましょう。

2.3　2 項分布

帰無仮説に基づけば、当たる確率は 1/4 です。それほど難しくなく、確率の計算方法を見つけることができそうです。

話し合ってみよう

1. 1 回の試行で、当たる事象が 0 回, 1 回発生する確率を、それぞれ求めてみよう。

2. 2 回の試行で、当たる事象が 0 回, 1 回, 2 回発生する確率を、それぞれ求めてみよう。

3. 3 回の試行で、当たる事象が 0 回, 1 回, 2 回, 3 回発生する確率を、それぞれ求めてみよう。

4. 4 回の試行で、当たる事象が 0 回, 1 回, 2 回, 3 回, 4 回発生する確率を、それぞれ求めてみよう。

このように決まる確率は、2 項分布と呼ばれています。なぜ、「2 項」という言葉をつかうかというと、皆さんが数学で学んだ (かもしれない) 次の式と関係があるからです。計算結果は、もちろん 1 ですが、途中の計算過程の各項に注目してください。

$$\begin{aligned}\left(\frac{3}{4}+\frac{1}{4}\right)^2 &= \left(\frac{3}{4}\right)^2+2\left(\frac{3}{4}\right)\left(\frac{1}{4}\right)+\left(\frac{1}{4}\right)^2\\ &= \frac{9}{16}+\frac{6}{16}+\frac{1}{16}\\ \left(\frac{3}{4}+\frac{1}{4}\right)^3 &= \left(\frac{3}{4}\right)^3+3\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)+3\left(\frac{3}{4}\right)\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^3\\ &= \frac{27}{64}+\frac{27}{64}+\frac{9}{64}+\frac{1}{64}\end{aligned}$$

どうしてこのような対応になるのでしょうか。例えば、$\left(\frac{3}{4}+\frac{1}{4}\right)^3$ を考えてみましょう。$\left(\frac{3}{4}+\frac{1}{4}\right)^3 = \left(\frac{3}{4}+\frac{1}{4}\right)\left(\frac{3}{4}+\frac{1}{4}\right)\left(\frac{3}{4}+\frac{1}{4}\right)$ です。括弧を外して展開することを考えましょう。括弧の中には 3/4 と 1/4 の 2 種類の項があります。全ての括弧の中から 3/4 を選ぶのは、1 通りの選択肢かありませんが、3/4 を 1 つ、1/4 を 2 つ選ぶのは、3 通りの選び方があります。このような計算は 3 回の試行で当たる事象が 3 回, 2 回発生する確率を計算するやり方と同じなので、対応があるわけです。

$$\begin{aligned}&\left(\frac{3}{4}+\frac{1}{4}\right)\left(\frac{3}{4}+\frac{1}{4}\right)\left(\frac{3}{4}+\frac{1}{4}\right)\\ &\left(\frac{3}{4}+\frac{1}{4}\right)\left(\frac{3}{4}+\frac{1}{4}\right)\left(\frac{3}{4}+\frac{1}{4}\right)\\ &\left(\frac{3}{4}+\frac{1}{4}\right)\left(\frac{3}{4}+\frac{1}{4}\right)\left(\frac{3}{4}+\frac{1}{4}\right)\end{aligned}$$

そうは言っても、これを 271 回の試行について計算するのは大変です。そこで計算機を使います。計算機は大変な計算にも文句を言わずに対応してくれます。ここでは、特に統計処理のために作成された「統計パッケージ R」 というソフトを使って計算してみます[6]。R の基本的な使い方については別紙を参照してください。

計算してみよう。

```
dbinom( c(0,1), 1, 0.25 )

dbinom( c(0,1,2), 2, 0.25 )

dbinom( c(0,1,2,3), 3, 0.25 )

dbinom( 0:20, 20, 0.25 )
plot( 0:20, dbinom( 0:20,20, 0.25 ), type='h' )

plot( 0:271, dbinom( 0:271, 271, 0.25 ), type='h' )
```

R の実行結果を図 1 に示しました。ここで注意しておきたいことをいくつか指摘しておきます。

[6]アメリカの ATT ベル研究所は、コンピュータのプログラミング言語 C 言語の開発で有名です。C 言語だけでなく、統計処理用にプログラミング言語である S 言語も開発されています。S 言語の処理系は有償ですが、これを模して作成された R (GNU R とか、統計パッケージ R とか、 R 言語と呼ばれるもの) は無償で利用できます。

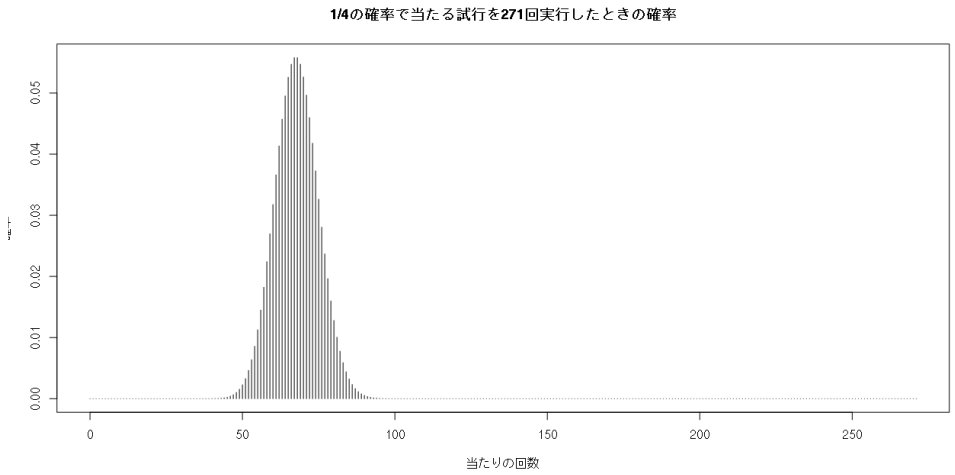

図 1: 2 項分布のグラフ

1. 当たる回数ごとの確率は小さくなる。

   271 回の試行に対して、当たる可能性がある場合の数は 0 〜 271 の 272 通りあります。場合の数が多いので、当たる回数を特定すると確率は、どうしても小さくなります。そこで、確率を考える場合には、当たりの回数については範囲 (○○回以上とか) で考えます。

2. 確率の分布はすそが広がった山の形をしている。

   特徴的な山の形になっていることが分かります。これについては、後で出てくる正規分布と関係がありますが、ここでは詳しく述べません。

3. 122 回以上当たる確率はとても小さそうである。

   図を見ると、122 回当たることはあまりなさそうに見えます。

2.4　2 項検定

1. 2 項検定とは

   帰無仮説に基づいて、当たる確率が 1/4 であるとして、271 回の試行で 122 回以上当たる確率を求め、これを合計してみましょう。

```
sum( dbinom( 122:271, 271, 0.25 ) )
```

   答えは 6.852498e-13 と表示されました。これは、6.852498×10^{-13} という意味で、もうすこしわかりやすく表現すると「1 兆分の 0.685」以下、あるいは、0.0000000000685% 以下です。

   1 兆分の 1 よりも小さい確率については、数字が小さすぎてピンとこないかもしれません。日本の人口は約 1 億人ですから、「日本人の誰か 1 人に賞金を渡す」と言ったときに、私が選ばれる確率は 1 億分の 1 です。1 兆分の 1 とは、その確率の 1 万分の 1 の確率です。世界で 1 人よりもずっと低い確率です。それが起こるとは思えません。

   まとめると、

電話をかけてきた相手が当たる確率が 1/4 であることを帰無仮説として、271 回の試行に対して 122 回の当たった事象が発生した。122 回以上当たる事象が発生する p 値は、0.0000000001% よりも小さいので、危険率 0.0000000001% で帰無仮説は棄却された。

となります。テレパシーが無いと考えて確率を計算すると、実験結果のようになる確率は 0.0000000001%以下であり、あまり起こりそうにありません。0.0000000001%程度の危険性はあるものの、テレパシーが無い、という仮説は否定できそうである、というわけです。

今回、2 項分布に基づいて統計検定を行いましたので、このような統計検定を 2 項検定といいます。

2. R を用いた計算 (片側検定と両側検定)

R には、2 項検定を行う便利な関数が備わっているので、その使い方も練習しましょう。

帰無仮説に基づいて、当たる確率が 0.25 であるとしたときに、271 回の試行で 122 回当たることについて統計検定するには、binom.test(122, 271, 0.25) と入力します。すると、次のように表示されます。

```
> binom.test( 122, 271, 0.25 )

        Exact binomial test

data:  122 and 271
number of successes = 122, number of trials = 271, p-value = 8.775e-13
alternative hypothesis: true probability of success is not equal to 0.25
95 percent confidence interval:
 0.3899386 0.5115317
sample estimates:
probability of success
             0.4501845
```

書いてあることの意味は次の通りです。

271 回の試行 (trials) で、122 回の当たり (successes) を考えると、p 値 (p-value) は、8.775e-13 である (さっきと、ちょっと値が違いますね。しかし、構わず、このまま話を進めます)。

対立仮説は「本当の当たる確率は 0.25 ではない」です。

元の確率が 0.25 ではなく、別の数だったと仮定して、同様の計算をしてみましょう。危険率を 5% として、帰無仮説が棄却されないのは、当たる確率が 0.3899386 〜 0.5115317 の間です。データに基づいて当たる確率を推定すると、0.4501845 となります。

この結果について、いくつか補足説明が必要です。

まず、最後の部分から説明しましょう。データに基づく当たる確率は 0.4501845 となりましたが、これは単に 122 を 271 で割った結果です。それが推定される確率になるのは当然ですね。

その前の部分は、重要な部分です。「本当の当たる確率」を考えて、これがいくつであるのか、データに基づいて考えることができます。これが統計学の「推定」です。考え方としては、帰無仮説に基づく当たる確率 0.25 の代わりに別の確率を設定して、同様に統計検定を行います。危険率 5% で帰無仮説が棄却されなくなれば、その確率は正しいかもしれない、と判断されます。棄却されない範囲を計算すると、0.3899386 〜 0.5115317 であった、ということです[7]。このように特に範囲を含めて推定することを「区間推定」と言い、「信頼度 95%の信頼区間は 0.3899386 〜0.5115317 である」と言います。

危険率を 1% にしたときはどうなるでしょうか。これを計算するには、

```
binom.test( 122, 271, 0.25, conf.level=0.99 )
```

と入力します。すると、値の範囲は 0.3719629 〜 0.5301958 となって広がりました。「外れる危険性を低くする」ことは、念のために、推定で求められる確率の範囲は広げざるを得ないことになるのです。

最後に、対立仮説と p 値についてです。検定を行うとき、「単に本当の確率は 0.25 ではない」とするのか「本当の確率は 0.25 よりも大きい」とするのかで、扱いが変わってきます。今回の実験では、「テレパシーがあるなら、当たる確率は高くなるはずだ」という雰囲気で話が進められています。しかし、271 回の試行で、1 回も当たらなかったとしたら、それはそれで別種のテレパシー（当てないように作用するテレパシー）があることが考えられます。

本当の確率が「0.25 より大きい」場合だけ考えて検定する場合、これを片側検定といいます。「0.25 とは異なる」として、大きい場合も小さい場合も考えて検定する場合、これを両側検定といいます。上の binom.test() では、何も指定しないと両側検定で計算することになっています。「0.25 とは異なる」という対立仮説に基づいて計算した結果なので、以前に計算した結果とは異なったのです。確認のために、片側検定で「0.25 より大きい」という対立仮説に基づいて計算してみましょう。

```
binom.test( 122, 271, 0.25, "greater" )
```

p 値は先ほど求めた結果に一致します。推定値は、下限だけが示されて上限が 1 となって示されなくなります。

3. まとめ

2 項検定についてまとめます。

- 基本的な考え方

[7]実際のこの計算は、もう少し複雑です。後ほど述べる両側検定、片側検定との話も関係してきます。

一定の確率で生じる事象であると帰無仮説を立てる。帰無仮説に基づいて、実際の試行に基づく結果が実現する確率 (p 値) を計算する。それを、設定した危険率 (5% とか 1% とか) と比べる。

p 値が危険率よりも小さかったら、滅多に起こらないはずなのに起きたと考え、帰無仮説を棄却する。

危険率よりも大きかったら、滅多に起こらないことが起きたとは言えないと考え、帰無仮説は棄却されないとする。

- 使える場面

  ポイントは、白黒判定（当たり外れ、表裏、サイコロの目が 1 かそれ以外か等など）でき、一定の確率で生じる事象を何回も繰り返すときに使える。

なお、検定にはいろいろな種類があります。今回の講習で説明するのは、そのほんの一部でしかありません。基本的な考え方を理解すれば、それぞれの状況に応じて適用することができるはずなので、基本をしっかり理解してください。

4. 練習問題

話し合いながら考えてみよう。

(a) シェルドレイクの実験で、271 回の試行で、当たった回数は 25 回だったとする。言えることは何か。

(b) 2019 年 1 月 25 日〜27 日にかけて日経新聞とテレビ東京が行った世論調査[a] では、990 件の回答に対して、安倍内閣支持率は 53% であったという。回答者の実数は示されていないので、仮に、990 名中、525 名が内閣を支持していると答えたとしよう。

このデータに基づいて、真の支持率が 50% であるとする仮説が棄却されるか検討せよ。また、真の支持率を信頼度 95% で区間推定せよ。

[a] 日経新聞 2019 年 1 月 28 日付け朝刊 1 面

3　t 検定 —真の値は？—

3.1　はじめに—長さの測定実験

勤務先の大学では、学生実験の際、図 2 のような印のついた紙を配布します。そして、数字の 1〜7 を折れ線で結んだとき、その折れ線の長さ（合計）はどれくらいになるのかを物差し（スケール）で測定してもらいます。あらかじめ、最小目盛の 1/10 まで読むように指示します。普通のスケールですので、最小目盛は 1mm ですから、0.1mm までの精度で読み取ることになります。皆さんは、その測定結果を、どう予想しますか？

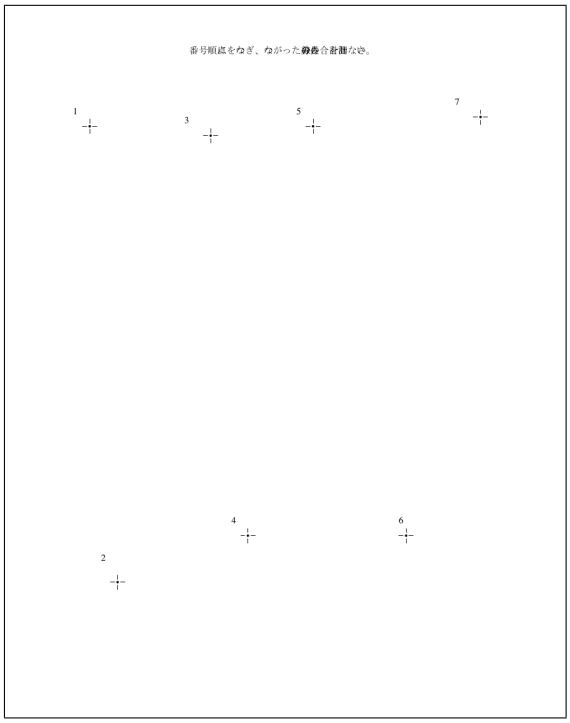

図 2: 長さを測定する用紙 (縮小版)

次の点について考えてみよう。

1. 学生が測った長さは一致するだろうか。
2. 差が生じるとすればなぜだろうか。
3. 「本当の長さ」はどのように考えたらいいだろうか。
4. 特定の長さ (例えば 100cm) より長いとか短いとかを判定するにはどうすればいいだろうか。
5. そのような議論をするためには何がわかればいいだろうか。

そのデータの一部を皆さんにも提供したので、実際にデータを扱いながら考えてみよう。手元のシステムには、既にデータが入っているので、次のようにしてデータを取り込み、頻度分布のグラフ (ヒストグラム) を描いてみましょう。ヒストグラムは、特定の測定結果の範囲に何人の学生が入っているかを示すグラフです。

```
data <- read.csv("/home/pi/Desktop/t 検定/DATA.csv")
data[[1]]

hist(data[[1]])

hist(data[[1]], seq( 91.8, 120.6, 0.4))

hist(data[[1]], seq( 91.8, 120.6, 0.2))
```

読み込んだ長さのデータは data[[1]] で参照できることがわかります。hist() はヒストグラムを作成する関数ですが、標準ではデータの区切りを自動で決めるので必ずしも良好な結果を示しません。そこで、seq(91.8, 120.6, 0.4) として、91.8 から 0.4 刻みで (あるいは 0.2 刻みで) 区間を区切ってヒストグラムを作成します。結果を図 3 に示します。

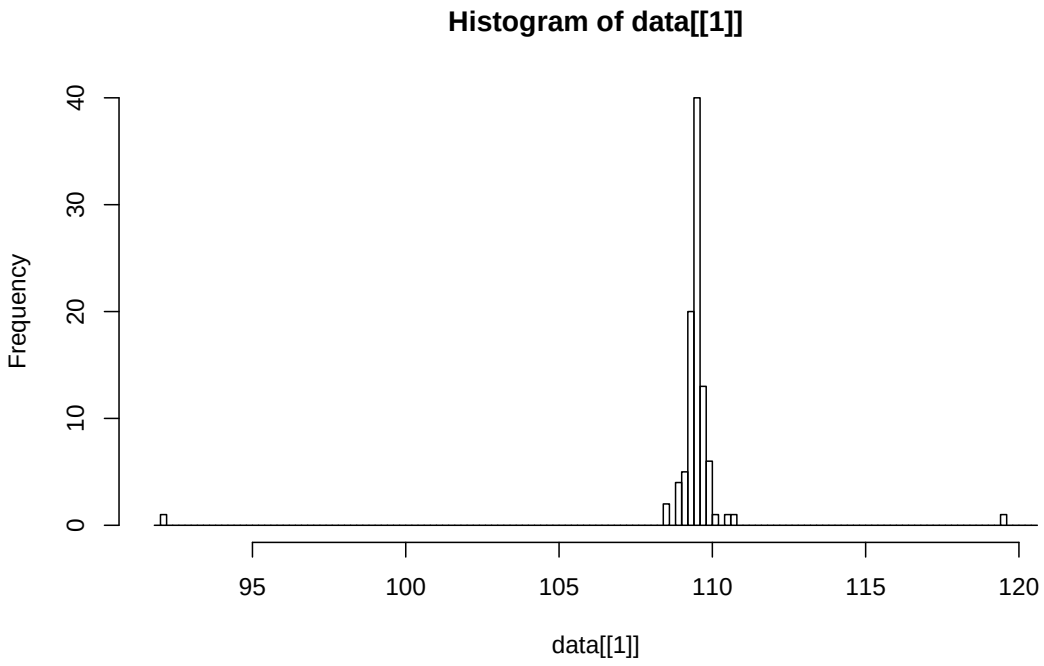

図 3: 長さの測定結果のヒストグラム

このデータから、いくつかのことに気がつきます。

1. 外れ値

   まず、極端に離れている値があることです。もちろん、これは、うっかりした学生が何かを間違ったのだと判定されます。一般にデータを扱うと、大きく外れた値が測定されることがあり、これを外れ値と呼んでいます。適切な基準は設定しなければなりませんが、その基準に基づいて外れ値は除外して考えるべきです。

2. グラフの形

   次に気がつくのは、図 3 グラフの形です。中心が高くその両脇は裾が広がっています。2 項分布のグラフ (図 1, p.9) と似ています。このような分布になっていることから、真の値は、ピークの近辺にあると予想されます。

3.2　検定の考え方

測定結果には、様々な影響があります。すぐに思いつくのは次のような影響です。

- 視差
- ものさしの不均一
- 印刷時の紙送りのずれ
- 点の大きさ
- 湿度による紙の伸縮
- …
- （読み取る人の勘違い・計算間違い）

最後の影響は、外れ値として除外できたものとします。しかし、他の影響は除き難いです。そこで、残念ながら真の値は求められないと考えるべきです。これは測定一般について言えることです。しかし、例えば、こんなことを検討することはできそうです。

> 測定結果を基にして考えると、真の値は 110cm である可能性は低い。

このようにすると、2 項検定と同じような問題になりそうだと気づきます。
そこで、発想として、

> 本当は 110cm なのに『こんな結果』になったとすると、それは滅多に起こらないことだから、本当は 110cm とした前提が間違っていると考えるべきだよね。

といったことを考え、次のような議論をしたいと思います。

> 長さが 110cm であることを帰無仮説として、測定結果のような事象が生じる p 値は、5% より小さいので、危険率 5% で帰無仮説は棄却される。

ところが、2 項検定のときと違う点もあります。2 項検定のときには、帰無仮説に基づいて「滅多に起こらない」を表現する確率を計算できました。今回の場合には、どのように計算したらいいのでしょうか。それを考えるためには、1 つの難関があります。その難関について次の節で述べます。

3.3　正規分布と中心極限定理

ここで書くことは、大学生向けの統計学の教科書でも「証明は難しいので省略する」と書いてあるくらい難しいです。そこで、ここでも簡単に説明するだけにします。証明も含めて正確なことは、大学で機会を見つけて勉強してください。

まず、測定について、もう一度思い出してください。何か特定の原因で測定値がバラつくとしましょう。できるだけ正確な測定をするためにその原因を究明して改善すれば、バラつきを減らすことができそうです。ところが、そのような努力を繰り返すと、やがてデータのバラつきを減らすことが困難になります。それは、複数の原因が同等に入り込むようになるからです。複数のデータのバラつきが重なり合うと、確率の分布は特定の形に近づくことが知られています。これを中心極限定理といい、近づく先の確率の分布を正規分布といいます。

実際にそうであるのか、確かめてみましょう。0〜1 までの実数が等確率で発生するような乱数 (一様乱数) を用意します。R では、runif() という関数で一様乱数を発生させることができます。まず、この乱数を 1000 個発生させて 0〜1 までの範囲で均等に発生することを確認してみます。

```
data <- NULL
for( i in 1:10000 ){
    data <- c(data, runif(1))
}
hist(data)
```

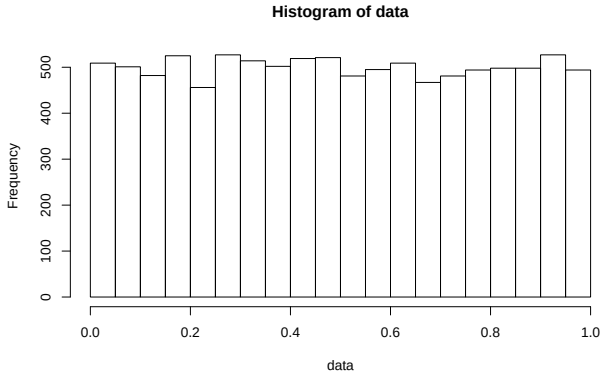

図 4: 一様乱数の確率分布

同じ乱数を 2 つ発生させて平均値を求めることを 10000 回繰り替えして同様の計算をしてみます。

```
data <- NULL
for( i in 1:10000 ){
    data <- c(data, sum(runif(2))/2 )
}
hist(data)
```

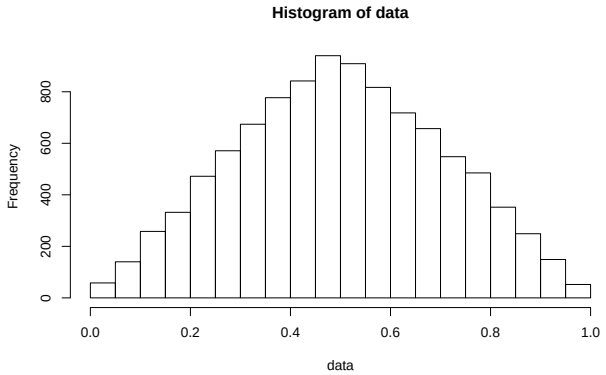

図 5: 2 つの一様乱数の平均値の確率分布

同じ乱数を 3 つ発生させて平均値を求めることを 10000 回繰り替えして同様の計算をしてみます。

```
data <- NULL
for( i in 1:10000 ){
    data <- c(data, sum(runif(3))/3 )
}
hist(data)
```

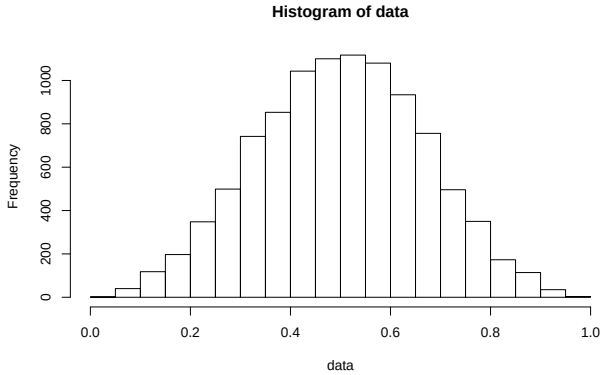

図 6: 3 つの一様乱数の平均値の確率分布

図 4 から図 6 まで観察すると、だんだん特徴的な形に近づいていくことがわかります。さらに、平均を求める個数を 10 個まで連続的に増やしながらグラフを描いてみましょう (図 7)。

```
N <- 10000
for( M in 1:10 ){
```

```
    data <- NULL
    for( i in 1:N ){
      data <- c(data, sum(runif(M)/M ))
    }
    hist(data)
}
```

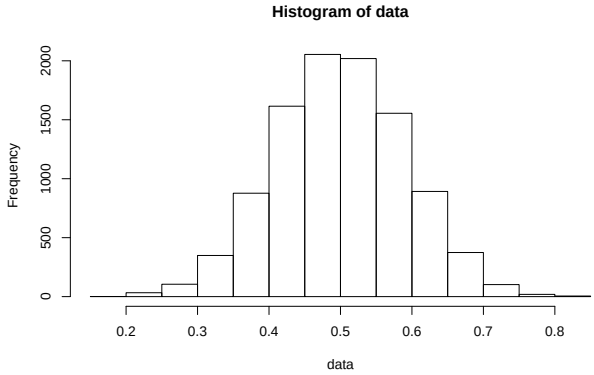

図 7: 10 個の一様乱数の平均値の確率分布

中央が高く、裾が広がる特徴的な分布に近づきました。複数のバラつきが重なり合うと、もとのバラつきがどのようであっても、データは正規分布に近づくのです。

長さのデータについても、中心極限定理に基づいて正規分布が当てはまるかどうか、検討してみた結果を図 8 に示します (プログラムは省きます。データのいくつかは、さらに外れ値として弾きました。)。長さの測定結果も、中心極限定理が予言するような正規分布に近いことが示されま

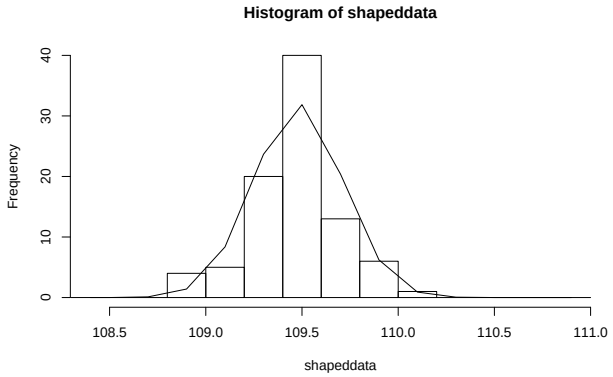

図 8: 長さの測定データ (バー) と正規分布を当てはめた値 (折れ線) の比較

した。

これらの結果から、測定値は正規分布に従って分布することが予想されます。正規分布は分布の形が理論的にわかっていますから、これに基づいて p 値を計算することができ、統計検定を行うことができそうです。

3.4　t 検定

1. t 検定とは

   t 検定にもいくつか種類があります。基本的には測定されたデータの真の値について検討する統計検定です。

   測定されたデータが、正規分布に従う沢山のデータ (母集団) から抽出されたデータであることを前提とします。このとき、

   > 測定対象の真の値が、特定の値であると仮定し、これを帰無仮説としたとき、測定結果と同様の結果が実現する p 値が、危険率を下回れば帰無仮説は棄却されるとする。

   ということをします。これが t 検定です。2 項分布の考え方と同じですが、確率の計算が 2 項分布ほど単純ではありません。そこで、R を使って計算しましょう。

2. R を用いた計算

   早速、長さのデータについて、t 検定を行ってみます。次のように入力してみましょう。ここでは、100 〜 111 の範囲のデータは外れ値ではないとして扱うことにします。

```
> data <- read.csv("/home/pi/Desktop/t 検定/DATA.csv")
> shapeddata <- data[[1]][ data[[1]] > 100 ]
> shapeddata <- shapeddata[ shapeddata < 111 ]

> t.test(shapeddata)

        One Sample t-test

data:  shapeddata
t = 3429.731, df = 92, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 109.4189 109.5457
sample estimates:
mean of x
 109.4823
```

   途中に書いてあることの意味は省略します。この関数でも区間推定を行っており、真の長さについて「信頼度 95%の信頼区間は 109.4189 〜 109.5457 である」と示しています。

   真の値が特定の値かどうか、ということを検討するためには、特定の値を設定して計算します。

```
> t.test(shapeddata, mu=110)
```

```
        One Sample t-test

data:  shapeddata
t = -16.2192, df = 92, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 110
95 percent confidence interval:
 109.4189 109.5457
sample estimates:
mean of x
 109.4823
```

真の値が 110cm であるとしたときの p 値は、2.2×10^{-16} 以下と示されています。とても低い確率です。110cm ではなく、109.6cm ではどうでしょうか。

```
> t.test(shapeddata, mu=109.6)

        One Sample t-test

data:  shapeddata
t = -3.6885, df = 92, p-value = 0.0003816
alternative hypothesis: true mean is not equal to 109.6
95 percent confidence interval:
 109.4189 109.5457
sample estimates:
mean of x
 109.4823
```

それでも p 値は 0.04 % 程度です。危険率 0.1% でも帰無仮説は棄却されます。つまり、真の値が 109.6cm であることは考えにくい、という結論です。

3. まとめ

   ここで述べた t 検定についてまとめます。

   - 基本的な考え方

     測定されたデータが特定の真の値を持ち、正規分布に従う母集団から抽出されたデータであると帰無仮説を立てる。帰無仮説に基づいて、実際の測定結果が実現する確率 (p 値) を計算する。それを、設定した危険率 (5% とか 1% とか) と比べる。

     p 値が危険率よりも小さかったら、滅多に起こらないはずなのに起きたと考え、帰無仮説を棄却する。

     危険率よりも大きかったら、滅多に起こらないことが起きたとは言えないと考え、帰無仮説は棄却されないとする。

- 使える場面

  2 項検定とは異なり、実数のデータが測定値として得られる場合の真の値を議論する場合に用いる。

なお、t 検定の中にもいろいろな種類があります。基本的な考え方を理解すれば、それぞれの状況に応じて適用することができるはずなので、基本をしっかり理解してください。

4. 練習

2 つのデータの平均値を比較して、差があると言えるのかどうかを判定する場合も t 検定といいます。R では、同じ関数 t.test() を用いますが、使い方がやや異なります。

ここでは、私の研究室で 1 秒おきに測定した 10 分間の気圧データ (データ数 600 個) について、最初の 100 個のデータと最後の 100 個のデータで平均値が異なると言えるかどうか、t 検定で調べてみましょう。

```
data0 <- read.table("/home/pi/Desktop/t 検定/30AEA44131102019030323 55.txt")

datahead <- data0$V4[1:100]
datatail <- data0$V4[501:600]

t.test(datahead)
t.test(datatail)

t.test(datahead, datatail) # データの差の区間推定値を表示する。

# 帰無仮設として、２つのデータの平均値の差を指定する。

t.test(datahead, datatail, mu=4.3, alternative="two.sided")
    # → 本当の差が 4.3 であってもこういうことは 23.97% で起こる。
t.test(datahead, datatail, mu=3.0, alternative="two.sided")
    # → 本当の差が 3.0 であってもこういうことは  0.27% で起こる。
t.test(datahead, datatail, mu=3.0, alternative="less")
t.test(datahead, datatail, mu=4.3, alternative="less")
t.test(datahead, datatail, mu=3.0, alternative="greater")
t.test(datahead, datatail, mu=4.3, alternative="greater")
```

何が言えるのか、画面表示を読み解いてみてください。

5. まとめと注意

これまで考えてきたことをまとめましょう。t 検定は、次のように表現することができます。

> 様々な要因でデータのバラつきがある測定値を、正規分布をする母集団から抽出したデータであると見なす。帰無仮説で真の値を仮定すると、どれくらいの確

率で測定結果（と同様の結果）が実現するか、p 値を計算することができる。p 値を設定した危険率 (5% とか 1% とか) と比べて滅多に起こらないことが起きた (帰無仮説は棄却される) のか、滅多に起こらないことが起きたとは言えない（帰無仮説は棄却されない）のか、判定する。

ここで注意しなければならないのは、測定データは正規分布に従うと仮定していることです。あらかじめ測定データが正規分布に従わないとわかっている場合には、このような方法は使ってはいけません。

現実として、測定データは正規分布に従わないとわかっていても、それ以外の方法が無いので、ここで述べた t 検定を使っている研究もあります。この点は十分に注意しなければならない点です。