検定

京都大学大学院情報学研究科知能情報学専攻
中澤 敏明
nakazawa@nlp.kuee.kyoto-u.ac.jp

2010/9/27 勉強会

0 はじめに

本稿の目標は、統計学の入門書に出てくるぐらいの (検定に最低限必要な) 基礎的な統計学の知識を身につけ、検定の基礎を学ぶことにある。あわよくば、実際に実験結果などに適用して論文に書けるぐらいになれれば十分である。なお検定の種類、方法、使える場面などは多種多様であり、全てを網羅することは不可能であるので、ここでは初歩的なものや研究上役立ちそうなものに限定する。

筆者は統計学の専門家ではないし、統計学をものすごく勉強したわけでもないので、内容に誤りなどがあったらすみません。もし誤りなどがあれば上記メールアドレスまで連絡いただければと思います。また本稿の内容や例題などは参考書をそのまま、もしくは一部改変して使用しています。

1 検定体験

1.1 問題

O 友君が M 月君にあるゲームで 100 番勝負を挑みましたが、結果は 39 勝 61 敗でした。

O:	いやー、今回は**偶然**俺の方が負け数が多かったな。
M:	偶然? **実力の差**だろ。
O:	はっはっは、冗談だろ？偶然に決まってる。
M:	偶然でこんなに偏るわけがないだろ。そんな**確率はほとんどない**。
O:	ほとんど? 具体的に言ってみろよ。
M:	今から計算してやる! そのかわり**確率が 5%より小さかったら**素直に実力差を認めろよ!

さて、二人の実力差を**統計的に**証明してください。

1.2 検定とは

ある仮説 H_1(**対立仮説**) が正しいと主張したいとする。しかしこれを直接証明するのは難しいので、これと比較する仮説 H_0(**帰無仮説**) を立てて、検証により H_0 を**棄却**することで、間接的に H_1 が正しいことを示す。この棄却できるかどうかを決定することを**検定**という。またこのような論法は論理学でいう**背理法**の考え方である。

上の例では

- 帰無仮説 H_0: 二人の実力は同じ ($\Leftrightarrow$ M が勝つ確率 $= 1/2$)
- 対立仮説 H_1: 二人には実力差がある ($\Leftrightarrow$ M が勝つ確率 $> 1/2$)

であり、H_0 を仮定したときに 39 勝 61 敗という結果をふくめて、これよりも極端な結果が出る確率 (p **値**) が閾値 α (**有意水準**、上の問題では 5%) より小さければ、H_0 は棄却され、H_1 が正しいことが言える。このとき、α の設定が小さければ小さいほど、より強く H_1 が正しいと主張することができる。よく論文などで "statistically significant ($p < 0.05$)"などという記述を目にすることがあるが、これは "有意水準 0.05 で統計的に有意な差があった" という意味である。

ちなみに、p 値 $\geq \alpha$ になったからといって H_0 **が正しいとは言えない**ので注意。この場合は H_0 が観測結果と「矛盾しない」ことが言えただけである (今回の観測データからだけでは判断できないという結論になる)。詳しくは 2.2 章を参照。

1.3 解答例 1

二人の実力差はない (M が勝つ確率は 1/2) と仮定する。

第 i 戦の結果を**確率変数**

$$X_i = \begin{cases} 1 & (\text{M の勝ち}) \\ 0 & (\text{M の負け}) \end{cases} \tag{1.1}$$

で表す ($i = 1, ..., 100$)。また各 X_i は**独立同分布** (i.i.d.)[1]であるとする。このとき確率変数 $S \equiv \sum_{i=1}^{100} X_i$ が M の合計勝ち数を表すとすると、S の**確率分布**は**二項分布**[2]$\mathrm{B}(100, 1/2)$ に従う。S が 61 以上となる確率 $\mathrm{P}(S \geq 61)$ は

$$\mathrm{P}(S \geq 61) = \sum_{k=61}^{100} {}_{100}C_k (1/2)^k (1/2)^{(100-k)} \approx 0.0176 \tag{1.2}$$

である。これは有意水準 5%より小さいため H_0 は棄却され、二人には実力差があることが統計的に示された。

ちなみに Excel で二項分布の確率を計算するときは、BINOMDIST(成功数, 試行回数, 成功率, 関数形式) という関数で行える。関数形式に TRUE を指定した場合の戻り値は累積分布関数の値 ($i = 0$ から成功数までの確率の累積和)、FALSE の場合は確率密度関数の値となる。よって式 1.2 を計算するときは 1-BINOMDIST(60,100,0.5,TRUE) とすればよい。

1.4 大数の法則

確率変数 $X_i (i = 1, ..., n)$ は i.i.d. であり、平均 μ、分散 σ^2 に従うとする。これら確率変数の平均 Z は、

$$Z = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1.3}$$

となる。ここで注意が必要なのは、Z **は揺らぐ量=確率変数**であるということである。平均を取ったからといって、ある一つの値に決まるわけではない。

Z の期待値 $E[Z]$ と分散 $V[Z]$ を求めると、

$$E[Z] = \frac{\sum_{i=1}^{n} E[X_i]}{n} = \mu, \quad V[Z] = \frac{\sum_{i=1}^{n} V[X_i]}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \tag{1.4}$$

となる。この結果からわかることは、

- 分散が元の分散の $1/n$ ということは**標準偏差**は $1/\sqrt{n}$ であり、精度を 10 倍 (典型的な振れ幅を 1/10) にするには n を 100 倍にする必要がある

[1]全ての確率変数が独立であり、かつ全て同じ分布に従う

[2]$\mathrm{B}(n, p) = {}_nC_x p^x (1-p)^{n-x}$

- $n \to \infty$ とすると、$V[Z] \to 0$ となり、$Z \to \mu$ となる

実はこの二つめが**大数の法則**である。つまり**母集団**がどのような分布をしているのかわからなくても、そこからランダムに**抽出**された**標本**を数多く集めると、標本の平均は真の平均 (母集団の平均[3]) と一致するということである。

1.5　中心極限定理

中心極限定理とは、大雑把にいうと以下のことである。

> 母集団の分布がどんな分布であっても、そこから抽出された標本 $X_i(i = 1, ..., n)$ の和 $S_n = \sum_{i=1}^{n} X_i$ の確率分布は n **を大きくすると正規分布に近づく**。母集団分布の平均を μ、分散を σ^2 とすると、S_n は $\mathrm{N}(n\mu, n\sigma^2)$ に従い、$Z_n = S_n/n$ は $\mathrm{N}(\mu, \sigma^2/n)$ に従う。

例えばサイコロを何度も振り、n 回目までに出た目の平均 Z_n を考える。各目の出る確率分布は一様分布 (全て 1/6) であるので、Z_1 も一様分布となり正規分布ではない。しかし 2 回振って出た目の平均 Z_2 は、$Z_2 = 3.5$ が最も大きくなり、$Z_2 = 3$ や 4 の分布が次いで大きく、逆に $Z_2 = 1$ や 6 の分布は小さくなる。これを何度も繰り返すと、Z_n の分布は正規分布に近づいていくというのが中心極限定理である。

中心極限定理は**母集団の分布がどんなときでも成り立つ**ため、逆に言うと**多くの標本を抽出すれば、その分布は正規分布と近似してもよい**となる。なお、一般的に「多くの標本」は $n \geq 100$ 程度とされているが、元の分布が対称的であるか偏っているかによって、正規分布に近づく早さは大きく異なるので注意が必要である。

1.6　解答例 2

先の問題を、中心極限定理により S を正規分布に近似することで解く。$S \sim \mathrm{B}(100, 1/2)$ の期待値は $E(S) = 50$、分散は $V(S) = 25$ である[4]ので、S は近似的に $S \sim \mathrm{N}(50, 25)$ とすることができる。

S が 61 以上となる確率 $\mathrm{P}(S \geq 61)$ は

$$\mathrm{P}(S \geq 61) = \int_{61}^{\infty} \frac{1}{\sqrt{2\pi \cdot 25}} e^{\frac{-(x-50)^2}{2 \cdot 25}} dx \tag{1.5}$$

であり、これを計算すれば良い。

正規分布の積分は頻出するが、毎回異なる平均・分散で計算するのは大変であるし、そもそもこの積分はちょっとやそっとでは解けない (ガウス積分)。そこで代表的な正規分布 (**標準正規分布、Z 分布**) についてあらかじめ計算しておき、数値表として持っておけば、それを参照するだけで良くなる。

式 A.16 を利用すると、任意の正規分布に従う確率変数を標準正規分布に従うように変換 (**標準化**) することができる。$X \sim \mathrm{N}(\mu, \sigma^2)$ とすると、

$$Z = \frac{X - \mu}{\sigma} \tag{1.6}$$

と変換することにより $Z \sim \mathrm{N}(0, 1)$ とすることができる。

標準化を用いて式 1.5 の積分値を実際に求める。$S \sim \mathrm{N}(50, 25)$ であったので、

$$Z = \frac{S - 50}{5} \tag{1.7}$$

と標準化することにより Z は標準正規分布に従う。

$$\mathrm{P}(S \geq 61) = \mathrm{P}(Z \geq \frac{61 - 50}{5}) = \mathrm{P}(Z \geq 2.2) \tag{1.8}$$

[3]母集団の平均や分散などのように、母集団の分布を特徴付けるパラメータを**母数**という

[4]確率変数 $X \sim \mathrm{B}(n, p)$ の期待値は $E(X) = np$、分散は $V(X) = np(1 - p)$

であるので、標準正規分布において $Z \geq 2.2$ となる確率を求めれば良いことになる。正規分布表によりこの確率は 0.013903 であることがわかり、この場合も帰無仮説は棄却される。なお解答例 1 の結果との差は約 0.37%であり、$n = 100$ 程度でも精度良く近似できることがわかる。もっと大きい n ならばさらに近似は高精度となる。

このように、正規分布は非常に便利なものであり、様々な場面で広く利用される。なお Excel で正規分布を扱う場合には、NORMDIST(x, 平均, 標準偏差, 関数形式) という関数で行える。関数形式に TRUE を指定した場合の戻り値は累積分布関数の値 ($x = -\infty$ から指定した値までの定積分値)、FALSE の場合は確率密度関数の値が返される。よって式 1.5 の値は 1-NORMDIST(61, 50, 5, TRUE)、もしくは標準化して 1-NORMDIST(2.2, 0, 1, TRUE) とすればよい。

2　検定の基礎

検定は

1. 帰無仮説と対立仮説を設定
2. 検定統計量を計算
3. 棄却域を設定
4. 帰無仮説が棄却できるかを判定

という手順で行われる。検定に必要な**検定統計量**は

- パラメトリックかノンパラメトリックか
- 検定の目的
- サンプルサイズ
- 分散が既知かどうか
- 対応のあるデータか対応のないデータか

などによって変わってくるため、行う検定ごとに適切なものを使用する必要がある。また**棄却域**は帰無仮説が棄却される統計量の値の範囲のことであり、逆に帰無仮説が採択される範囲を**採択域**と呼ぶ。棄却域は

- 有意水準
- 両側検定か片側検定か
- 変数の自由度

などによって決まる。

2.1　用語説明

2.1.1　パラメトリックとノンパラメトリック

母集団分布の形が与えられており (二項分布、ポアソン分布、正規分布など)、いくつかの定数 (**母数**) さえわかれば母集団分布が定まる場合を**パラメトリック**の場合と呼ぶ。現実世界におけるパラメトリックな現象は数多く知られている。

これに対し、いくつかの母数だけでは母集団分布が決定できない場合を**ノンパラメトリック**の場合と呼ぶ。例えば世界各国の人口や面積などの分布は具体的な分布を当てはめるのは困難であるため、ノンパラメトリックの場合であると考えられる。

2.1.2　分散が既知かどうか

多くの場合、母分散の分散は未知である。このような場合には母分散に変わる標本分散などを用いる必要がある。ただし、サンプルサイズが十分大きい場合 ($n \geq 100$) には標本分散を近似的に母分散として用いることができる。なお分散が必要となるのは平均値に関する検定だけである。

2.1.3　対応のあるデータと対応のないデータ

対応のあるデータとは、例えば同一調査対象者に対して 1 年前に測定したデータと現在のデータのように、同一標本から得られたデータのことである。これに対して**対応のないデータ**とは、例えば首都圏と京阪神圏、20 代と 30 代のように、2 つの異なる標本から得られたデータのことである。

2.1.4　両側検定と片側検定

両側検定とは棄却域が採択域の両側に分けられて設定されている検定で、**片側検定**とは棄却域が採択域の左右どちらか片側に設定されている検定であり、左側にある場合を左片側検定、右側にある場合を右片側検定という。

両側検定は以下のような場合に用いる。

- 分析者が両側の偏りに関心がある
- 調査結果の方向性が想定できない
- 具体的に「A と B に違いがあるか」と問われている

一方、片側検定は以下のような場合に用いる。

- 分析者が特定の方向に関心がある
- 具体的に「A は B より大きい (小さい) か」と問われている

片側検定は例えば「新薬の効果は現行薬より高いか (右側検定)」や「あるダイエット法によって本当に体重は減少したか (左側検定)」といったように、方向が決まっている場合などに用いられる。

注意点としては、有意水準を α に設定した場合、片側検定では累積確率が α の範囲を棄却域と設定するが、両側検定では棄却域が左右に分散されているため、それぞれの範囲で累積確率が $\alpha/2$ の範囲を棄却域と設定する。このため、両側検定の方が帰無仮説が棄却される可能性が低くなる。逆にいうと、両側検定で帰無仮説が棄却することができれば、より強く対立仮説が正しいことが主張できることになる。このような理由から、特別の理由がない限りは両側検定をする方が良い (ただし、片側検定でも元の有意水準を半分に設定すれば同じ精度で検定できる)。

2.2　仮説検定の誤り

仮説検定においては表 2.2 に示すように 4 通りの場合が存在し、誤った判定をしてしまうこともある。このうち**第 1 種の誤り** (本当は帰無仮説は正しいのに誤って棄却してしまう) については、有意水準 α を低く設定しているため、比較的起こりにくい誤りであるといえる。問題となるのは**第 2 種の誤り** (帰無仮説は誤っているのに採択してしまう) である。つまり帰無仮説を採択する場合にはある確率 (危険率 β) で判断が誤っている可能性があるということになり、このため帰無仮説を採択する場合には「帰無仮説は正しい」とはいわずに、「対立仮説が正しいとはいえない」と結論付け、判断を保留する。

2.3　(検定) 統計量

標本を要約し、母集団の母数のいろいろな推測に使われるものを**統計量**という。重要なことは、統計量は標本を要約しただけ (観測された値を使って求めるもの) であるので**未知のパラメータを含まない**ということである。以下、代表的な統計量を示す。

表 2.1: 仮説検定の誤り

	帰無仮説を採択	帰無仮説を棄却
帰無仮説は真	正しい判定	**第 1 種の誤り**
帰無仮説は偽	**第 2 種の誤り**	正しい判定

2.3.1 標本平均

母平均 μ、母分散 σ^2 の母集団分布からの標本 $X_i (i = 1, ..., n)$ に対して、**標本平均** $\overline{X}$ は

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2.1}$$

であり、 その期待値は $E(\overline{X}) = \mu$、分散は $V(\overline{X}) = \frac{\sigma^2}{n}$ である。

2.3.2 Z 分布

$\overline{X}$ の標準化変数

$$Z = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \tag{2.2}$$

の分布を Z **分布**といい、$Z \sim \mathrm{N}(0, 1)$ である。Z 分布は母集団分布が正規分布でしかも**分散** σ^2 **が既知**の場合、もしくは**標本数が十分大きく**、中心極限定理により標本分布が正規分布と仮定できる場合の検定に利用される。(例題 4.1-1)

2.3.3 標本分散

母平均 μ、母分散 σ^2 の母集団分布からの標本 $X_i (i = 1, ..., n)$ に対して、**標本分散** s^2 は

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1} \tag{2.3}$$

と定義される。分母は n ではなく $n - 1$ であることに注意が必要である。これは $E(s^2) = \sigma^2$ となり、母分散と一致し、母分散を正しく推定できるためである。この標本分散 s^2 を特に**不偏分散**という。また、この $n - 1$ を**自由度**という。自由度とは「自由に動ける変数の数」という意味である。標本分散においては $\frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$ であるので、$X_1, ..., X_{n-1}$ と $\overline{X}$ が決まると、X_n は自由に動くことができないため、自由度は $n - 1$ となる。

2.3.4 χ^2 分布

$Z_1, ..., Z_n$ をそれぞれ $\mathrm{N}(0, 1)$ に従う確率変数とする。いま

$$\chi^2 = \sum_{i=1}^{n} Z_i^2 \tag{2.4}$$

とするとき、確率変数 χ^2 が従う確率分布を**自由度** n **の** χ^2 **分布**と呼び、$\chi^2(n)$ と書く。χ^2 分布は確率密度関数 $f(x)$ が $x > 0$ の部分でのみ正の値を取り、ガンマ分布 $Ga(n/2, 1/2)$ と同じである。

この χ^2 分布を用いると、標本分散を s^2 として

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2} \tag{2.5}$$

は $\chi^2(n - 1)$ に従う。χ^2 分布は主に**母分散の検定、適合度や独立性の検定**に用いられる。

なお Excel で χ^2 分布の片側確率 ($x > 0$ から ∞ までの定積分) を返す関数は CHIDIST(x, 自由度) である。(例題 4.2-1)

2.3.5　t 分布

標本数が小さく、しかも**分散 σ^2 が未知**の場合、標準化変数 Z の σ^2 の代わりに標本分散 s^2 を用いて

$$t = \frac{\overline{X} - \mu}{\sqrt{\frac{s^2}{n}}} \tag{2.6}$$

という分布を考える。これを t **統計量 (スチューデントの** t**)** と呼ぶ。注意すべきは、この分布は**正規分布には従わない**ということである。これを変形すると、

$$\begin{aligned} t &= \frac{\overline{X} - \mu}{\sqrt{\frac{s^2}{n}}} && (2.7) \\ &= \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \bigg/ \sqrt{\frac{s^2}{\sigma^2}} && (2.8) \\ &= \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \bigg/ \sqrt{\frac{(n-1)s^2}{\sigma^2} \bigg/ (n-1)} && (2.9) \end{aligned}$$

となり、分子 $(\overline{X} - \mu)/\sqrt{\sigma^2/n}$ は標準正規分布、分母 $(n-1)s^2/\sigma^2$ は自由度 $n-1$ の χ^2 分布に従うような分布であることがわかる。これを新たに**自由度** $n-1$ **の** t **分布**と呼び、$t(n-1)$ と書く。

一般に、二つの確率変数 Y と Z が

- $Z \sim \mathrm{N}(0,1)$
- $Y \sim \chi^2(n)$
- Z と Y は独立

を満たすとき、確率変数

$$t = \frac{Z}{\sqrt{Y/k}} \tag{2.10}$$

が従う確率分布を**自由度** n **の** t **分布 (スチューデントの** t **分布)** と呼ぶ。

t 分布は正規分布と非常に良く似た分布をしており、正規分布を少し横に引っ張ったような形をしている。このため、同じ有意水準を設定した場合、Z 分布よりも t 分布の方が棄却域が原点より遠くなる。t 分布は $n \geq 30$ 程度でほとんど標準正規分布と変わらない形になり、$n \to \infty$ とすると t 分布は標準正規分布に一致する。(例題 4.1-2)

なお Excel で t 分布の確率を返す関数は TDIST($x > 0$, 自由度, 尾部) である。尾部に 1 を指定した場合は片側確率 (x から ∞ までの定積分)、2 を指定した場合は両側確率 (片側確率の 2 倍) となる。

3　2 標本問題

2 つの異なる母集団から抽出された 2 種類の標本の比較をする問題を **2 標本問題**という。ここでは 1 つ目の標本 $X_1, ..., X_m$ を母集団分布 $\mathrm{N}(\mu_1, \sigma_1^2)$ から抽出し、2 つ目の標本 $Y_1, ..., Y_n$ を母集団分布 $\mathrm{N}(\mu_2, \sigma_2^2)$ から抽出することを考える。各標本の標本平均は

$$\overline{X} = \frac{\sum_{i=1}^m X_i}{m}, \quad \overline{Y} = \frac{\sum_{j=1}^n Y_j}{n} \tag{3.1}$$

となる。二つの標本間の母平均の差を分析することが重要となることが多いが、これは標本平均の差を分析すればよい。またこの際、標本平均の分散を考慮するべきである。

3.1　分散が既知、もしくは大標本のとき

2 つの母集団の分散をそれぞれ σ_1^2、σ_2^2 とすると、$\overline{X} \sim \mathrm{N}(\mu_1, \sigma_1^2/m)$、$\overline{Y} \sim \mathrm{N}(\mu_2, \sigma_2^2/n)$ となる。$\overline{X}$ と $\overline{Y}$ は独立であるため、

$$\overline{X} - \overline{Y} \sim \mathrm{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}) \tag{3.2}$$

となる (確率変数の引き算の場合、平均は引き算になるが分散は足し算になるので注意)。これをさらに標準化すると、

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \tag{3.3}$$

となり、$Z \sim \mathrm{N}(0, 1)$ となる。

この分布を用いて、2 つの母集団の平均の差の検定を行うことを考える。帰無仮説 H_0 は「$\mu_1 = \mu_2$」、両側検定の場合の対立仮説 H_1 は「$\mu_1 \neq \mu_2$」、片側検定の場合は「$\mu_1 > \mu_2$」または「$\mu_1 < \mu_2$」となる。検定統計量は

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \tag{3.4}$$

であり、この値が棄却域に入るかどうかで検定すればよい。(例題 4.1-3)

さらに各標本が対応のあるデータのときを考える。このとき、$m = n$ となることに注意する。$\overline{D} = \overline{X} - \overline{Y}$ とすると、$\overline{D}$ の標本分散 $s_d^2 = \sigma_1^2 + \sigma_2^2$ であるので、検定統計量は

$$Z = \frac{\overline{D}}{\sqrt{s_d^2/n}} \tag{3.5}$$

となる。(例題 4.1-5)

3.2　分散が未知だが両標本で等しいとき

2 つの母分散は未知だが、$\sigma_1^2 = \sigma_2^2$ であることがわかっている場合、$\overline{X} - \overline{Y}$ の標本分散 s^2 を 2.3.3 章で説明した標本分散を用いて以下のように定義する。

$$s^2 = \frac{\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2}{m + n - 2} \tag{3.6}$$

$$= \frac{(m-1)s_1^2 + (n-1)s_2^2}{m + n - 2} \tag{3.7}$$

s_1^2 と s_2^2 はそれぞれの標本の標本分散である。このように定義すると、$(m+n-2)s^2/\sigma^2 \sim \chi^2(m+n-2)$ であり、s^2 と $\overline{X} - \overline{Y}$ は独立となる。これを用いると、$\overline{X} - \overline{Y}$ の標準化変数は

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{s\sqrt{1/m + 1/n}} \tag{3.8}$$

となり、自由度 $m + n - 2$ の t 分布 $t(m + n - 2)$ に従う。

この分布を用いて検定を行う場合には、検定統計量

$$t = \frac{\overline{X} - \overline{Y}}{s\sqrt{1/m + 1/n}} \tag{3.9}$$

を用いればよい。(例題 4.1-4)

また対応のあるデータの場合は、

$$t = \frac{\overline{D}}{\sqrt{s_d^2/n}} \tag{3.10}$$

となる。式自体は式 3.5 と同一であるが、それぞれの統計量が従う分布が異なることに注意が必要である。(例題 4.1-6)

表 4.1: 対応のあるデータ (大標本)

n	x_i	y_i	$d(=x_i-y_i)$	d^2
1	8	6	+2	4
2	6	7	-1	1
3	9	5	+4	16
…	…	…	…	…
220	7	5	+2	4
計			+308	1012

表 4.2: 対応のあるデータ (小標本)

n	x_i	y_i	$d(=x_i-y_i)$	d^2
1	4,310	3,860	+450	202,500
2	2,050	2,660	-610	372,100
3	3,700	4,110	-410	168,100
…	…	…	…	…
30	4,890	3,220	+1,670	2,788,900
計			+6,960	21,211,200

3.3　分散が未知で両標本で等しくないとき

この場合はどのように工夫をしても母分散によらない統計量を作ることができないが、近似的に分布を求める**ウェルチの近似法**が知られている。詳細は省略するが、

$$t = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{s_1^2/m+s_2^2/n}} \tag{3.11}$$

が、近似的に自由度が

$$\nu = \frac{(s_1^2/m+s_2^2/n)^2}{\frac{(s_1^2/m)^2}{m-1}+\frac{(s_2^2/n)^2}{n-1}} \tag{3.12}$$

に最も近い整数 ν^* の t 分布 $t(\nu^*)$ に従うというものである。

4　パラメトリック検定の例題

ここまでで説明した方法で、以下の問題を有意水準 $\alpha = 0.05$ で検定せよ。

4.1　母平均の検定

1. ある地域で、乗用車を購入した 250 人に購入価格をたずねたところ、平均購入価格は 1,183,000 円、標準偏差は 101,000 円であった。この平均購入価格は 1,150,000 より高いといえるか。

2. あるレストランで食事をした 30 人に、支払った金額をたずねたところ、平均支払額は 3,280 円、標準偏差は 950 円であった。この平均支払額は 3,000 円より高いといえるか。

3. 携帯電話の購入金額について、18-34 歳の 240 人は平均 16,720 円、標準偏差 5,210 円であり、35-49 歳の 230 人は平均 12,830 円、標準偏差 4,170 円であった。両年代の平均購入額に違いはあるか。

4. 風邪薬の購入金額について、A 都市の 41 人は平均 1,240 円、標準偏差 80 円であり、B 都市の 35 人は平均 1,060 円、標準偏差 70 円であった。両都市での平均購入額に差はあるか。

5. 220 人に 2 種類のラーメン x と y の味について、10 点満点で評価してもらった結果を表 4.1 に示す。ラーメン x と y の評価に差はあるか。

6. 30 人にある文具の購入金額について、「最近 3ヶ月」と「1 年前の 3ヶ月」を比較してもらったところ、表 4.2 に示す。両時点での平均購入額に差はあるか。

4.2　母分散の検定

1. 新しい製造方法で生産した 30 個の製品の重量を測定した結果、分散は 0.17 であった。これは従来の製造方法での分散 0.39 より小さくなったと判断してよいか。

表 5.1: 12 人の被験者による睡眠薬の効果 (http://oku.edu.mie-u.ac.jp/ okumura/stat/signtest.html 一部改変)

被験者	1	2	3	4	5	6	7	8	9	10	11	12
睡眠時間	1.9	0.8	1.1	0.1	0	-0.1	4.4	5.5	1.6	0	4.6	3.4
増減の絶対値	1.9	0.8	1.1	0.1	0	0.1	4.4	5.5	1.6	0	4.6	3.4
順位	6	3	4	1.5	-	1.5	8	10	5	-	9	7

5 ノンパラメトリック検定

ここまではパラメトリックな場合 (母集団の分布がいくつかの母数で決定できる場合) について述べたが、ここからはそうでない場合について考える。

5.1 符号検定とウィルコクソンの符合付順位和検定

問題 10 人の患者にある睡眠薬を飲ませたところ、睡眠時間がそれぞれ表 5.1 に示す時間だけ増えた。このとき、この睡眠薬は睡眠時間に効果があったか。帰無仮説 H_0 は「睡眠薬に効果はない」であり、対立仮説 H_1 は「睡眠薬に効果はある」である。

このような問題に使えるのが**符号検定**や**ウィルコクソンの符合付順位和検定**である。前者は評価値が二値の場合 (+/-や◯/×など) に用いられ、後者は評価値の差が定義でき、かつ差の順位付けができるときに用いられる。なおどちらの場合でも、**評価に差のないデータは除外**して考える。

まず符号検定による検定を行う。表 5.1 の結果を「睡眠時間が増えたかどうか」という結果だと思うと、増えたと答えたのが 9 人、減ったと答えたのが 1 人、変わらないと答えたのが 2 人である。変わらないと答えた 2 人はデータから除かれるので、この場合はデータ数 10 で、増えたのが 9 人、減ったのが 1 人ということになる。さて、睡眠薬に効果がないならば、各符号 (増えたか減ったか) が出る確率は 1/2 ずつであり、二項分布 $\mathrm{B}(10, 1/2)$ に従う。すると、実はこれは 1 章で紹介した例と全く同じ方法で検定が行えることになる。

次にウィルコクソンの符合付順位和検定による検定を行う。まず睡眠時間の増減の絶対値に基づいて、評価結果の順位付けを行う。このとき値が同じものがある場合は、順位の平均値をそれぞれに割り当てる。次に、評価が+(睡眠時間が増加) だったものと-(睡眠時間が減少) だったものの順位をそれぞれ足す。

- 睡眠時間が増加: $6 + 3 + 4 + 1.5 + 8 + 10 + 5 + 9 + 7 = 53.5$
- 睡眠時間が減少: 1.5

このうち、**値が小さいものが検定統計量**となる。この場合は 1.5 である。有効標本数 10 と有意水準から棄却限界値を求め、**検定統計量が棄却限界値よりも小さい場合に帰無仮説が棄却**される。有意水準を 5%とすると、標本数 10 の場合の両側検定の棄却限界値は 8 であり、1.5 はこれより小さいため、帰無仮説は棄却される。

なお標本数が大きい場合には、この統計量の平均が $\frac{n(n+1)}{4}$、分散が $\frac{n(n+1)(2n+1)}{24}$ になることから、

$$Z = \frac{\left|T - \frac{n(n+1)}{4}\right|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{5.1}$$

のように標準化を行い、Z 分布による検定を行う。式中の T は先に求めた検定統計量 (ここでは 1.5) である。

5.2 マクネマー検定

問題 あるタスクに対する既存手法と提案手法による結果を表 5.2 に示す。このとき、二つの手法の間で精度に違いはあるか。帰無仮説 H_0 は「精度に差はない (=母比率に差はない)」であり、対立仮説 H_1 は「精度に差がある (= 母比率に差がある)」である。

表 5.2: 既存手法と提案手法における正解の分布

		提案手法		
		正解	不正解	計
既存手法	正解	$a = 44$	$b = 25$	69
	不正解	$c = 33$	$d = 48$	81
	計	77	73	150

このように **2×2 分割表**で表現されたデータのうち、$b + c$ **の値が大きい場合**に用いられるのが**マクネマー検定**である。これは二つの要素 (ここでは既存手法と提案手法) 間で母比率 (全体のうち、一方の事象が占める割り合い。ここでは正解率) に違いがあるかを検定するものである。これにより、例えば精度向上したかど うかの検定が行える。

マクネマー検定の検定統計量は

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{5.2}$$

であり、自由度 1 の χ^2 分布に従う。問題文の検定統計量を計算すると、

$$\chi^2 = \frac{(25-33)^2}{25+33} = 1.103 \tag{5.3}$$

であり、有意水準 1%で検定を行うと、棄却限界値は 6.6349 である。よって帰無仮説は棄却されない (精度が向上したとは言えない)。

6 まとめ

本稿では統計学に入門し、超基礎的な統計学の知識を身につけ、検定の基礎を学んだ。ただし、この内容で十分というわけでは全くないので、この先は各自必要なことを学習すること。

A 公式集

以下に様々な公式を示す。これらは証明なしに用いるが、証明自体は易しいので各自確認すること。なお以下では特に断りのない限り、c は定数、X や Y は確率変数、$f(x)$ は確率密度関数とする。

A.1 期待値

$$E[X] = \sum_x x f(x) \quad (X \text{ が離散値を取る場合}) \tag{A.1}$$

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx \quad (X \text{ が連続値を取る場合}) \tag{A.2}$$

$$E[c] = c \tag{A.3}$$

$$E[X + c] = E[X] + c \tag{A.4}$$

$$E[cX] = cE[X] \tag{A.5}$$

$$E[X + Y] = E[X] + E[Y] \tag{A.6}$$

A.2　分散

$$V[X] = \sum_x (x - E[X])^2 f(x) \quad (X \text{ が離散値を取る場合}) \tag{A.7}$$

$$V[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx \quad (X \text{ が連続値を取る場合}) \tag{A.8}$$

$$V[X] = E[X^2] - (E[X])^2 \tag{A.9}$$

$$V[c] = 0 \tag{A.10}$$

$$V[X + c] = V[X] \tag{A.11}$$

$$V[cX] = c^2 V[X] \tag{A.12}$$

$$V[X + Y] = V[X] + V[Y] \quad (\text{ただし、} X \text{ と } Y \text{ が}\textbf{独立の場合のみ}) \tag{A.13}$$

A.3　正規分布 (ガウス分布)

ある確率変数 X が正規分布に従うとき、その確率密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{A.14}$$

となり、平均 $E(X) = \mu$、分散 $V(X) = \sigma^2$ となる。正規分布はこの平均と分散さえわかればただ 1 つに形が決まるのが特徴である。またこの正規分布を $\mathrm{N}(\mu, \sigma^2)$ と書き、$X \sim \mathrm{N}(\mu, \sigma^2)$ で X が正規分布に従うことを表現する。

$\mu = 0$、$\sigma^2 = 1$ の正規分布を特に**標準正規分布**と呼び、$\mathrm{N}(0, 1)$ で表す。

$$\mathrm{N}(0, 1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \tag{A.15}$$

以下に正規分布に関する公式を示す。

$$X \sim \mathrm{N}(\mu, \sigma^2) \rightarrow aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2) \tag{A.16}$$

$$X \sim \mathrm{N}(\mu_1, \sigma_1^2), Y \sim \mathrm{N}(\mu_2, \sigma_2^2) \rightarrow aX + bY \sim \mathrm{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2) \quad (\text{ただし } X \text{ と } Y \text{ は独立}) \tag{A.17}$$

参考文献

[1] 奥村学, 高村大也. 言語処理のための機械学習入門. コロナ社, 2010.

[2] 上田拓治. 44 の例題で学ぶ統計的検定と推定の解き方. オーム社, 2010.

[3] 東京大学教養学部統計学教室編. 統計学入門. 東京大学出版会, 2005.

[4] 平岡和幸, 堀玄. プログラミングのための確率統計. オーム社, 2009.