IV-3. 統計検定

IV-3-1. Student の t 検定

t 検定とは、２組のデータがあるときに、そのデータの平均値に差があると言えるか言えないかという検定です。確率論的に言えば、平均値が同じ母集団からとられたデータである確率を検定します。

この検定が想定しているモデルを理解するために、ある値（x）があり、その値があらかじめ予測される平均値の候補として妥当であるかを検討してみましょう。
データーから計算された平均値をM、母集団の平均値をμ、母集団の標準偏差をσとすると、それぞれの値の関係は以下のように表せます。

$$x = \mu \pm \alpha\sigma$$

αは、σを単位とした時のどのくらい離れているかを示す値で、具体的なデーターがあれば、

$$\alpha = \frac{x - \mu}{\sigma}$$

という計算式で求められます。
この値を標準正規分布に当てはめて、確率水準によって決まる値より大きいか小さいかを検討し，大きければ、その値は予測される値の候補としては不適切であると結論します。
たとえば、95%の信頼限界（危険率 p=0.05 と表すほうが一般的）で検討するのであれば、αの値が 1.96 より大きければ、その値を候補として採用することは不適切と判断します。
確かにそうですが、私たちが知っているのはデーターの平均（M）と標本集団から計算された母集団の分散の推測値σ'^2であって真の平均値μ、母集団の分散σ^2ではありません。μもσも知りようがないから計算できません。標本サイズを大きくすれば、標本集団の平均値と分散は母集団の平均値と分散に近づくから、問題ないだろうという考え方はあります。実際､記述統計学の大成者である Karl Pearson はそう考えていました｡実務家にとっては、実験の繰り返し数を増やすには、場所や費用の制約があります。小標本で統計学的な確率を論じたいというのは、実務家として当然の願望です。ギネスの醸造所で働いていた William Gosset も小標本による統計的な判定に強い興味を持ちました。ギネスでは技術者が論文を投稿することを禁止していましたから、この問題についての論文を、ピアソンの主催している Biometrika という雑誌に投稿しました。この時に使っていたペンネームが Student です。これが分散分析の創始者である Ronald Fisher に認められて、Fisher によって Student の t 分布として整理されます。
Gosset の考え方を整理します。
まず分散についてです。分散の値も確率的に変動するのだから、正規分布ではなくて、分散を確率変数としているχ^2分布と正規分布を組み合わせた t 分布を確率モデルとして使え

ば良い。これが一つです。
次にμですが。t 検定では、平均値に差があるかないかを検定しています。つまり、平均値の差が 0 であることが帰無仮説ですから、$\mu = 0$です。
もう一つ重要なことは、平均値であるということです。データから予測された平均値は、母集団の期待値（平均値）の周りに標準誤差を偏差として分散しているはずです。ですから、標準偏差σではなくて標準誤差$\frac{\sigma}{\sqrt{n}}$を差の平均値Mから 0 の距離の単位にしなければなりません。

$$M = \mu \pm t\frac{\sigma}{\sqrt{n}}$$

t は、$\frac{\sigma}{\sqrt{n}}$を単位とした時のどのくらい離れているかを示す値。
これを変形して

$$\mathrm{t} = \frac{\mu - M}{\frac{\sigma}{\sqrt{n}}}$$

$\mu = 0$を仮定しているから

$$\mathrm{t} = \frac{M}{\frac{\sigma}{\sqrt{n}}}$$

基本的な考え方は以上です。Gosset のオリジナルはもう少し違う形で書かれています。この形に整理したのは Fisher です。シンプルですが、これでは具体的にどうすればよいのかわかりませんね。
実際の検定では、σやnをどのように計算すればよいかを説明します。２つの場合が考えられます。一つは対になったデータというもので、いくつもの鉢があるときに、同じ鉢に２つの異なる植物を植えて成長を比較するというように、比較する相手が１対１に決まっているデータです。もう一つは、鉢を２つのグループに分けて、それぞれのグループの鉢には同じ種類の植物を植えて、グループ間で比較するというものです。初めの例を対になったデータの t 検定と呼び、普通の t 検定と区別します。分散の計算の考え方は、「4-1. データの扱い（情報の取り出し）」を読むと、理解できます。

対になった t 検定（paired t test）

この検定では、対になったデータのどちらかが相手に比べ大きいか、データに差があるといってよいかということを検定します。
対になったデータを比べるとは、たとえば、右手と左手とどちらが長いかとか、手と足とどちらが長いかとか、一つの鉢に違う植物をうえて、これを繰り返してどちらの植物の成長が早いかなどをくらべることをイメージすればよいでしょう。対になっているのだから、１番目の人の右手にはそれと対になった左手があります。このデータをA_1とB_1と表現すると、以下A_2，B_2，A_3，B_3のように表現できます。必ず一組になっているのだから、C_kとして$A_k - B_k$を求めることができます。C_kも標本から得られるデータで、確率的に変

動します。ところで、ＡとＢに差がないということはＣの平均が 0 ということです。つまり、帰無仮説は$\bar{C}$=0、差がないということです。
この場合は、データ対の数をデータ数nをとして、差の値の平均M、分散σを計算し、次のようにｔの値を求めて、危険率 p を適切に決めて、自由度$n-1$としてｔ分布表を使って検定します。

$$\mathrm{t}=\frac{M}{\frac{\sigma}{\sqrt{n}}}$$

式 42

対になっていないデータのｔ検定

上記の知識を使って、対になっていないデータのｔ検定の方法を考えます。ｔ検定の考え方の基本は、対になったｔ検定の場合と同じです。技術的に克服しなければならないポイントは差の分散をどのようにして求めるかです。
これが求まれば、式 51 を使って

$$x=\mu\pm t\frac{\sigma_{A-B}}{\sqrt{N}}$$

として

$$x=M_A\quad \mu=M_B$$

を代入して

$$t=\frac{M_A-M_B}{\frac{\sigma_{A-B}}{\sqrt{N}}}$$

を求め、この大小を検討すればよいでしょう。M_A　M_B は、それぞれ、Ａ、Ｂ，の平均値です。ここで問題が 2 つあります。σ_{A-B}と Nです。N は式 42 では n とあらわされていますが、A 群のデーターの数ともＢ群のデーターの数とも違います。新たに Nを決めなくてはなりません。A 群のデーター数をm、B 群のデーター数をn、$\mathrm{N}=\bar{n}$とします。

$$\frac{\sigma_{A-B}}{\sqrt{\bar{n}}}=S$$

と表します。$\sigma_{A-B}{}^2$はＡ、Ｂ共通の分散ですが、これをＡ、Ｂの分散$(\sigma_A{}^2,\sigma_B{}^2)$から求めなければなりません。
　ここでわれわれが証明しようとしている帰無仮説は､A,B 両群が同じ母集団からとられたということです。つまり、

$$\sigma_A{}^2=\sigma_B{}^2$$

もし､これが成り立つならば､重みをつけようが軽みをつけようがその平均である $\sigma_{A\text{-}B}$ は、

$$\sigma_{A-B}{}^2=\sigma_A{}^2=\sigma_B{}^2$$

です。
すでに説明したように

$$\mathrm{E}\left({M_A}^2\right)=\frac{1}{m}{\sigma_A}^2$$

つまり、A群の母集団の平均値の周りに、A群のデータによって推定された平均値が

$\frac{1}{m}$ σ²A の広がりで分布しています。

$$\mathrm{E}\left({M_A}^2\right)=\frac{1}{m}{\sigma_A}^2=\frac{1}{m}{\sigma_{A-B}}^2$$

$$\mathrm{E}\left({M_B}^2\right)=\frac{1}{n}{\sigma_B}^2=\frac{1}{n}{\sigma_{A+B}}^2$$

$$\mathrm{SS}=\mathrm{E}\left({M_{A-B}}^2\right)=\mathrm{E}\left({M_A}^2\right)+\mathrm{E}\left({M_B}^2\right)=\frac{1}{m}{\sigma_A}^2+\frac{1}{n}{\sigma_B}^2$$

だから

$$\mathrm{SS}=\frac{1}{m}{\sigma_{A-B}}^2+\frac{1}{n}{\sigma_{A-B}}^2=\left(\frac{1}{m}+\frac{1}{n}\right){\sigma_{A-B}}^2$$

$$\mathrm{S}=\sigma_{A-B}\sqrt{\frac{1}{m}+\frac{1}{n}}$$

これを$\frac{\sigma_{A-B}}{\sqrt{\bar{n}}}=S$に入れると

$$\frac{\sigma_{A-B}}{\sqrt{\bar{n}}}=\sigma_{A-B}\sqrt{\frac{1}{m}+\frac{1}{n}}$$

$$\bar{n}=\frac{1}{\frac{1}{m}+\frac{1}{n}}=\frac{mn}{m+n}$$

$$\bar{n}=\left(\frac{mn}{m+n}\right)$$

求める t 値は

$$\mathrm{t}=\frac{M_A-M_B}{\sigma_{A-B}\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

式 43

となります。

また、IV-2-2. 和の分散、差の分散」で説明したように、A-B の分散は、A+B の分散と同じで、A，Bそれぞれの分散を自由度で重みをつけた平均（下式）になります。

$${\sigma_{A-B}}^2={\sigma_{A+B}}^2=\frac{(m-1){\sigma_A}^2+(n-1){\sigma_B}^2}{m+n-2}$$

以上で対になっていないデータのｔ検定ができます。IV-2-2 の差の分散の説明でつかった表 13 の具体的なデータを使って実際の計算例を示します。

表 13.具体的なデータ(再掲)

	A	B
	1	1
	5	5
	6	6
		8
平均	4	5
SS	14	26

A：データー数　m=3、平均 M_A＝4　SS_A ＝14

B：データー数　n=4、　平均 M_B＝5　SS_B ＝26　分散 σ^2＝8,66667

２つをあわせた自由度は７－２＝５

$SS_{A\text{-}B}=SS_A+SS_B$＝14+26＝40

$\sigma_{A\text{-}B}^2$=40/5=8

$\sigma_{A\text{-}B}$＝$2\sqrt{2}$ =2.8284

$\sqrt{\frac{1}{m}+\frac{1}{n}}$ $=\sqrt{\frac{1}{3}+\frac{1}{4}}$ =0.763763

M_A-M_B の絶対値１

ｚ＝1/(2,2824×0.763763)=0.46291

ｔ表によればｐ＝０．０５（５％の危険率）でのｔの臨海値は 2.571 ですから、Ａの母集団とＢの母集団の平均値が異なっていると結論することはできません。