平均値の条件差のｔ検定の問題

シミュレーションによる若干の検討

○岡本安晴
日本女子大学

1　はじめに

データの平均値の比較は、最もよく行われる統計分析の１つである（Derrick et al., 2016）。２つのグループの比較において、Student の t 検定は、デフォルトになっている（Delacre et al., 2017）。Student の t 検定の結果は、p 値に集約される。p 値は、今後も使われ続けると予想されている（Benjamin et al.,2019）。p 値にはよく知られた批判がある（小杉, 2019）。p 値は約 100 年にわたる問題であるが（Ruberg, 2021）、ASA Statement (Wasserstein, 2016)では有用な統計的指標でありうるとされている（p. 131）。しかし、Student の p 値は、正規分布の仮定と等分散の仮定の下で算出される。正規分布は心理学においてデフォルトとして用いられている分布であるが（椎名、2013）、実際には多くの自然・社会現象において、正規分布とは異なる非対称な形の分布が観測されることが知られている（國仲ら、2011）。等分散の仮定が成り立たないときは Welch の p 値が用いられる。 p 値が適切に解釈されるためには、p 値は正しく算出されていなければならない。Delacre ら（2017）は、正規分布、非対称正規分布、二重指数分布、カイ二乗分布、一様分布の場合の p 値をシミュレーションによって調べ、Welch の p 値が良いことを報告している。心理学研究における分布としては、上記のものの他に、t 分布、ポアッソン分布、対数正規分布、順序カテゴリ値の分布（多項分布）などがあるが、これらの分布および正規分布について Student の p 値と Welch の p 値の比較シミュレーションを行った。また、Delacre ら（2017）のグラフは Student の p-value を横軸、Welch の p 値を縦軸に取ったものであり、２つの p 値の関係は分かるが、真の p 値との関係は不明である。p 値の定義は、「統計量がデータによって与えられる値以上の極端な値を取る確率」であるので（Wasserstein et al., 2016）、一様分布に従う。すなわち、

(1)　　$P(\text{p 値} \leq u) = u$

である。したがって、正しい p 値の累積分布曲線は、連続変数の場合、原点(0,0)から点(1,1)までの対角線となる。本報告では、シミュレーションによって得られた p 値が式（１）を満たしているかどうかを累積分布曲線により検討する。なお、ここに掲載されていないグラフは発表時に示す。

平均値の比較は、対応のあるデータの t 検定で行われることがあるので、この場合の注意点についても取り上げる。

２．独立なデータの平均値の比較

正規分布、ｔ分布、ポアッソン分布、対数正規分布、順序カテゴリ値分布の場合についてシミュレーションを行った。正規分布は、統計においてデフォルトして用いられている分布であり、ｔ分布は正規分布では外れ値として扱われる値を取り込むために用いられる。ポアッソン分布は度数データに対して用いられ、対数正規分布は自然現象および社会現象において広く見られる分布である。順序カテゴリ値は心理学において評定データで用いられる。正規分布では、平均０、標準偏差１の分布を基準にしているが、ｔ値および df がデータの原点および単位の影響を受けないので、シミュレーションの結果はかなりの一般性があると考える。データのサイズは、報告者が目にしてきた我が国における研究のデータサイズから 20 を基準にした。シミュレーション結果は、データのサイズの関係により以下のようであった。

2.1　２つのデータのサイズ M と N が等しい場合

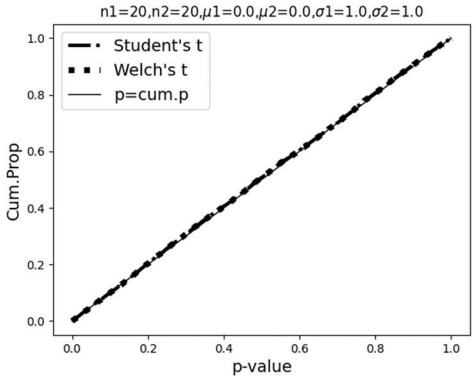

図１　正規分布、等分散、等サンプルサイズ

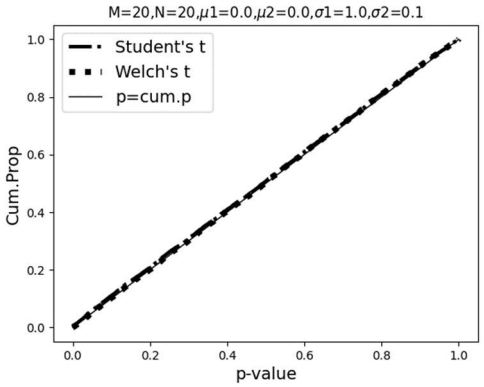

図２　正規分布、不等分散、等サンプルサイズ

等分散、不等分散のいずれの場合とも、対数正規分布を除き、Student の p 値と Welch の p 値は一様分布に従っていた（図１、図２）。対数正規分布の場合は、等分散、不等分散の場合とも Student の p 値と Welch の p 値の累積曲線は重なっており、不等分散の場合は一様分布からのずれが p 値の小さい領域で見られた（図３）。これらに対して、尤度比から求めた p 値は一様分布に良く重なっていた。ベイズ分析との比較は、p 値がモデルの下でのデータ（統計量）の確率に関わるものであるのに対して、ベイズ分析はデータに対してのモデル（パラメータ）の確率に関わるものである（Ruberg, 2021）ので直接の比較は難しい。データから情報を得るという目的にはベイズ分析の方が直接的であると思われる。

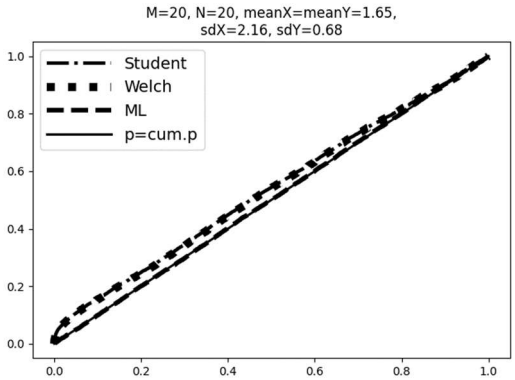

図３　対数正規分布、不等分散等サンプルサイズ

2.2　２つのデータのサイズが異なる場合

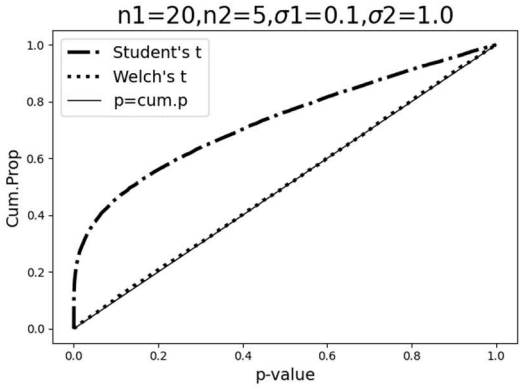

図４　正規分布、不等分散、
不等サンプルサイズ

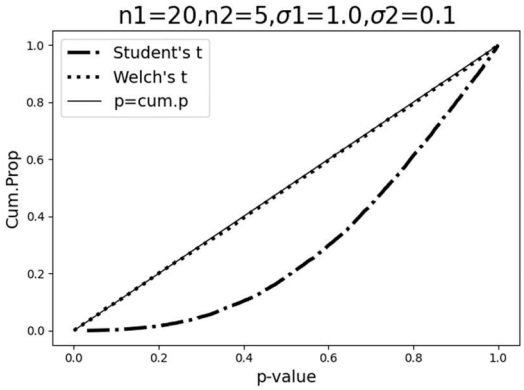

図５　正規分布、不等分散、
不等サンプルサイズ

正規分布、ｔ分布とも等分散の場合は Student の p 値、Welch の p 値は一様分布によく従っていた。不等分散のときは Welch の p 値は一様分布に良くしたっがていたが、Student の p 値は一様分布から外れていた（図４，図５参照)。ポアッソン分布では、分散は平均値に等しいので平均値が等しいとき分散も等しいモデルとなるが、Student の p 値、Welch の p 値とも一様分布の累積曲線に重なっていた。順序カテゴリ値の場合は、等分散に対しては Student の p 値、Welch の p 値とも一様分布の累積曲線に重なっていた。不等分散に対しては Student の p 値は一様分布から外れていたが Welch の p 値は一様分布の累積曲線に重なっていた。対数正規分布の場合は、等分散のときは、Student、Welch、尤度比の p 値とも概ね一様分布の累積曲線に近かった。不等分散に対しては、Student の p 値は、一様分布の累積曲線から大きく異なる分布であった。Welch の p 値と尤度比の p 値は尤度比の方が一様分布に近い分布であったがほぼ同じ分布であった。分散が大きい方のデータサイズが大きいときは一様分布の累積曲線にほぼ重なっていた。しかし、分散の大きい方のデータサイズが小さいときは一様分布の累積曲線からのずれが見られ、Welch の p 値は p 値の小さい領域で急激な増加を示していた。

３．対応のあるデータの平均値

対応のあるデータの平均値の差の検定法として t 検定がある。t 検定では抽出できなかった情報がデータに残っている場合の例を、以下にシミュレーションにより示す。

差と一方の変量との関係が１次式で表される場合

図６に対応のあるデータ例（x_pre、x_post）の差 x_post−x_preのヒストグラムを示す。このデータの t 検定の結果は p 値=0.828 となり非有意である。しかし、x_preを独立変数、x_post－x_preを従属変数としてベイズ法による単回帰分析を行うと図７の結果を得る。太い実線で描かれた直線は、係数の事後分布の中央値を係数の値とする回帰直線であり、細い実線は MCMC サンプルからランダムに選んだパラメータ値による回帰直線である。。x_post－x_preは、x_preの１次式で表されることが分かる。横軸は、x_preの値が中心化されたものである。x_preの測定後、

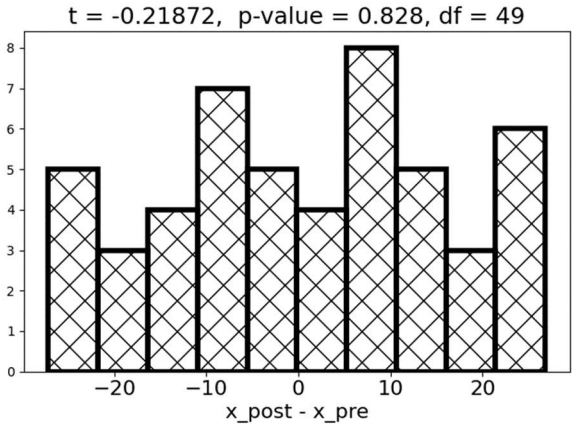

図６　データの差のヒストグラム

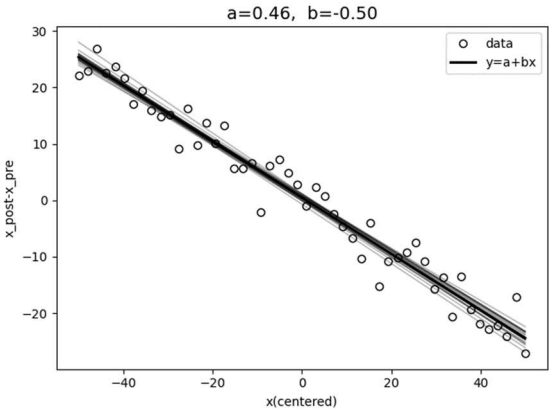

図７　回帰分析

何らかの講習会が実施され、x_postが測定されたとすると、x_preの値が平均より小さい場合は講習会は促進的効果が認められ、x_preの値が平均より大きい場合は抑制的効果が認められたと解釈できる。

逆 U 字型

データの t 検定による p 値は p 値= 0.84（両側検定）であり、有意な差は認められない。差x_post − x_presを中心化したx_preの 2 次式の回帰モデルによりベイズ分析を行うと、2 次曲線によりデータの分布が表されることがある。２次曲線が逆 U 字型であれば、講習前のスコアをx_pre、講習後のスコアをx_postとするとき、講習の効果は中ほどの受講者には正の効果が見られるが、講習前のスコアが平均より低い、あるいは高い受講者には負の効果が見られるということになる。受講前のスコアが中ほどの受講者に対する正の効果と、スコアが低いあるいは高い受講者に対する負の効果が全体として合わさって t 検定による非有意な差を示す結果となったと考えられる。

有意な t 値

t 検定により有意な差であるという結果が出ても、回帰分析を行うとより豊かな情報が得られ場合がある。例えば、p 値として0.000を得た（両側検定）データにおいて、差x_{post} − x_preをx_pre の中心化した値x_cの 1 次による回帰モデルで分析すると、差x_{post} − x_preがx_cの 1 次関数として減少しているという結果が得られる可能性がある。

４．参考文献

Derrick, B., Toher, D. and White, P. (2016) Why Welch's test is type I error robust. *The Quantitative Methods for Psychology*, 12, 30-38.

Ruberg, S. J. (2021). Détente: A practical understanding of p values and Bayesian posterior probabilities. *Clinical Pharmacology & Therapeutics*, 109, 1489-1498.

Wasserstein, R. L., and Lazar, N. A. (2016) The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70, 129-133.