更新日：2022年3月23日

『統計的因果推論の理論と実装：潜在的結果変数と欠測データ』
（共立出版）
Q & A

『統計的因果推論の理論と実装：潜在的結果変数と欠測データ』の記述内容について，読者からの質問の一部に回答しています．なお，このQ & Aの回答は閲読を経ておらず，著者のインフォーマルな見解ですので，あくまでも参考程度の情報です．新たに追記したQ & Aは記載日時を赤字にしています．

30ページ目（**2022年3月23日記載**）

質問：2.10節の2標本t検定において，「95%信頼区間（95% confidence interval）は4.219～20.811であり，0が含まれていないことから，2つの集団の差は5%水準で有意である．」（p.30）とありますが，この記述の「0が含まれていない」とは何を意味していますか？

回答：本書では，pp.16-18 で具体的な数値例を導入してから，pp.20-24 で処置効果（因果効果）の定義をしました．p.18 では，「ID1 の学生の場合，76－68 = 8 であるから，8 点の効果があった．この数値例では，どの学生に対してもプラスの効果がある．」と述べました．正の値はプラスの因果効果を，負の値はマイナスの因果効果を，ゼロは因果効果がないことを意味しています．その流れで，p.30 では，帰無仮説が 0 であり，95%信頼区間の中に 0 が含まれていないから，5%水準で帰無仮説を棄却できており，母集団において因果効果があるといえるということです．また，p.57 では，「この数字が正の値ということは，補習授業に効果があるという意味であった．95%信頼区間は 0 を含んでいないのだから，5%の有意水準で，母集団においても効果ありと判断できる．」と述べました．この意味も，上記のとおりです．

51ページ目（**2022年3月23日記載**）

質問：「$t_\alpha(df)$は，信頼区間を構築する際に使用する信頼係数である」とありますが，信頼係数は$1-\alpha$ではないでしょうか？

回答：$\bar{X} \pm t_\alpha(df) \times s.e.(\bar{X})$の信頼区間を考えます．本書p.52の式（4.13）です．$\bar{X}$は標本平均です．αは有意水準です．dfは自由度です．$s.e.(\bar{X})$は標準誤差です．このとき，おっしゃるとおり，厳密には「$1-\alpha$」のことを信頼係数（あるいは信頼度）と呼びます．一方，$t_\alpha(df)$には，名前らしい名前はありません．本書のp.51の上から16行目では，「自由度dfのt分布の上側確率αのパーセント点を$t_\alpha(df)$」と記しました．したがって，$t_\alpha(df)$の名前は，「自由度dfのt分布の上側確率αのパーセント点」と言えなくもないですが，これを名前とするには長すぎると思います．実際には，dfは任意のデータとモデルにおいて固定です．これは

解析者が選ぶ数値ではありません．よって，信頼係数$1-\alpha$が決まると，$t_\alpha(df)$の値も一意に定まるため，本書では$t_\alpha(df)$の部分を「信頼係数」と呼んでいます．

任意の標本データにおいて，$\bar{X} \pm t_\alpha(df) \times s.e.(\bar{X})$の信頼区間を構築するとき，$\bar{X}$，$df$，$s.e.(\bar{X})$はすべて固定されています．つまり，任意の標本データから90%信頼区間，95%信頼区間，99%信頼区間を構築するとき，この3つの信頼区間の何が異なるかと言えば，$t_\alpha(df)$の値が異なります．したがって，任意の標本データから構築する信頼区間の幅は$t_\alpha(df)$によって一意に決められており，$t_\alpha(df)$は信頼区間を計算するときに「信頼度を決める係数」というぐらいの意味でご理解いただければと思います．さらに細かく見れば，$t_\alpha(df)$の値は$1-\alpha$によって一意に決められており，厳密には，$1-\alpha$のことを信頼係数（あるいは信頼度）と呼ぶのは，そのとおりです．

本書のChapter 4は，標準誤差と信頼区間の意味を簡潔かつ具体的におさらいすることを目的としており，t分布そのものに関する煩雑な議論は意図的に省いており，このあたりの区別は厳密にしていません．もしこの用語の使い方がどうしても気になるようでしたら，本書において「信頼係数」と書かれている部分は，適宜，「自由度dfのt分布の上側確率αのパーセント点」と読み替えていただいても差し支えないと思います．

90ページ目（**2022年3月23日記載**）

質問：最小二乗法の仮定1の誤差項の期待値ゼロの部分に関して，真のモデルはこの場合，何でしょうか？　p.91で，誤差項の期待値がゼロでない場合，Y切片の値に誤差項の0からのずれが吸収されるとのことでした．それを示すRのシミュレーションでは，a0の真値が1であるのに対して，推定結果は10.984でした．ここでは真値を正しく推定できていないという結論を記述しているのですが，回帰直線がその標本の関係性をどの程度正しく表しているのかという観点から考えると，むしろ推定結果の方が正しいY切片を示しているように思いました．そこで思ったのですが，そもそも正しいモデルを仮定する際に誤差項が0ではない（0から大きく外れる）ような状況を仮定するモデルは正しいのでしょうか？

回答：推定対象となる真のモデルは，p.90の証明では，一行目の$\alpha_0 + \beta_1 X_{1i}$の部分です．表7.1では，6行目のa0+b1*x1の部分です．

表7.1のRコードで考えてみましょう．変数y1の成り立ちは，「a0+b1*x1」という体系的な部分と「u1」というランダムな部分の2つの部分から構成されています．ここで，a0は1.0で，b1は1.5と設定しています．つまり，体系的には，変数y1は「1.0+1.5*x1」です．この部分が「真のモデル」です．モデル化できる部分の真の形です．ここから，さらに±u1のランダムなばらつきがあります．この部分は観測されない誤差項で，ランダムなばらつきなので，モデル化の対象外です．7行目のlm関数の中にu1が出てこないのは，u1自体はモデルに含めていないということですね．また，u1は，N(10, 1)と設定しています．結果として，実現値としての変数y1は「1.0+1.5*x1+N(10, 1)」ですが，「11.0+1.5*x1+N(0, 1)」と同じものとして実現するため，Y切片が「真のモデル」の1.0と異なるものの，傾きは1.5の

ままというのがポイントですね．ここで，あくまでも，体系的な部分は「a0+b1*x1」であり，a0は1.0と設定していますから，Y切片の真値は1.0です．

具体的な状況として，1問1点で100問出題される試験の点数を考えてみましょう．このとき，「試験の点数=a0+b1*正解数」です．ここで，a0は0（正解数が0なら0点だから）で，b1は1（1問1点だから）ですね．つまり，学生Aが50問正解したならば，体系的には，0+1*50=50点のはずです．

ところが，先生Bが採点をしたところ，この試験は難しく作り過ぎてしまったため，平均点が30点しかなかったとします．大学の授業では60点を超えないと単位が取れませんので，このままでは不合格者が続出するので，全員の点数に30点を加点したとします．なお，この先生Bは雑な性格をしており，集計するたびに学生の点数を±1点の誤差で採点ミスをするとします．（倫理的な問題は無視します．）つまり，「試験の点数=a0+b1*正解数+u1」で，u1は誤差項です．そして，u1の平均は30，分散は1です．誤差項u1が正規分布に従っているなら，N(30, 1)ということですね．7.1節のポイントは，これは，「試験の点数=(a0+30)+b1*正解数+N(0, 1)」と同じと考えてよいということです．

なぜ，(a0+30)ではなく，a0=0が切片の真値と考えているかについては，別の先生Cが採点した場合を考えてみましょう．先生Cも「試験の点数=a0+b1*正解数」（a0=0，b1=1）として採点するので，平均点は30点です．この先生Cは厳格な人で，加点を一切しないとします．また，採点ミスもしないとします．すると，「試験の点数=a0+b1*正解数+u2」で，u2は誤差項ですが，その平均はゼロで，分散もゼロです．学生Aは50問正解しており，0+1*50±0=50点です．つまり，先生Bが採点しても，先生Cが採点しても，切片a0は0ですが，先生Bが採点したときには誤差項u1がN(30, 1)なので，結果として，この場合には(a0+30)と見なせたということですね．

108ページ目（**2022年3月23日記載**）

質問：多重共線性の問題について信頼区間が広がるとありますが，データの僅かな変動が回帰係数に大きく影響することの方が問題ではないでしょうか？

回答：どちらも同じことと思います．拙著では「多重共線性の問題とは，小標本サイズの問題と同じ」（p.108）と述べました．点推定値自体は不偏ですが，小標本のときと同様に，データの僅かな変動が回帰係数に大きく影響することがあります．これは，標準誤差の大きさに反映され，信頼区間の幅に影響が出ます．計量経済学では，この問題はmicronumerosity という用語で知られています．詳しくは，Goldberger, A. S.（1991, pp.248-250). *A Course in Econometrics*, Harvard University Press もご覧ください．

次ページに続く

119ページ目（**2022年3月23日記載**）

質問：表8.13のvcovHCは何を意味していますか？

回答：vcovはvariance-covariance matrix（分散共分散行列）を意味しています．また，HCはHeteroskedasticity-Consistent Covariance Matrix Estimation（不均一分散に頑健な共分散行列推定）を意味しています．いくつかの引数（選択肢）がありますが，HCにより，Whiteの推定量が得られます．

126ページ目（**2022年3月23日記載**）

質問：重回帰分析での交互作用と多重共線性の違いがよく分かりません．その後の統計的処理の問題ですか？

回答：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 x_2) + \varepsilon$を考えます．$x_1 x_2$は交互作用項です．多重共線性は，$x_1$と$x_2$の相関が強いため，$x_1$と$x_2$の共通している部分から$y$への効果が大きい問題をいいます．このとき，$\beta_3 = 0$なら，回帰式は$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$なので，多重共線性だけなら，$x_1$から$y$への効果は$x_2$の水準に依存しません．交互作用の場合，$\beta_3 \neq 0$で，回帰式は$\hat{y} = \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2$なので，$x_1$から$y$への効果は単に$\beta_1$ではなく，$\beta_1 + \beta_3 x_2$なので，$x_2$の水準に依存します．このとき，$x_1$と$x_2$の相関は強い場合もあれば弱い場合もあります．

165ページ目（**2022年3月23日記載**）

質問：表11.12のvcovCLは何を意味していますか？

回答：vcovCLはClustered Covariance Matrix Estimation（クラスターに頑健な共分散行列推定）を意味しています．クラスターに頑健な標準誤差については，Croissant and Millo (2019, pp.109-123)を参照してください．

次ページに続く

265 ページ目（**2022 年 3 月 23 日記載**）

質問：データ内の欠測値が 9999 で表されている場合，na.strings=9999 で NA を定義できますが，同じデータ内に欠測値を表す記号が複数ある場合はどうすればよいでしょうか？

回答：左側の表のデータでは，赤字が欠測値を表しており，「9999」と「xx」はともに欠測値とします．この場合，引数 na.strings の指定をする際に，c 関数を使えば，「9999」と「xx」の両方を NA として定義してデータを読み込むことができます．

```
read.csv(file.choose(), na.strings=c(9999, "xx"))
```

	A	B	C	D
1	y	x1	x2	x3
2	185	80	9999	74.1
3	176	9999	72	72
4	171	68	68	68
5	159	59	9999	59
6	175	69	69	xx
7	171	9999	58	xx
8	170	59	59	xx
9	162	61	61	61
10	168	56	56	56
11	174	69	69	69

```
> data19a<-read.csv(file.choose())
> data19a
    y   x1   x2   x3
1 185   80 9999 74.1
2 176 9999   72   72
3 171   68   68   68
4 159   59 9999   59
5 175   69   69   xx
6 171 9999   58   xx
7 170   59   59   xx
8 162   61   61   61
9 168   56   56   56
10 174   69   69   69
>
> data19b<-read.csv(file.choose(),na.strings=c(9999,"xx"))
> data19b
    y x1 x2   x3
1 185 80 NA 74.1
2 176 NA 72 72.0
3 171 68 68 68.0
4 159 59 NA 59.0
5 175 69 69   NA
6 171 NA 58   NA
7 170 59 59   NA
8 162 61 61 61.0
9 168 56 56 56.0
10 174 69 69 69.0
```