ピアソンの相関係数が1または-1になりえないケース

椎名乾平　早稲田大学
久保沙織　早稲田大学
大内善広　城西国際大学
上田卓司　早稲田大学

ある条件下でr が-1または1を取りえないのを発見しました．

その条件とは

1) 変数Xは，$m \geq 2$ 個の順序カテゴリーを持ち，Yは $n \geq 2$ 個の順序カテゴリーを持つ．
2) mとnは異なる．
3) これらのm個とn個のカテゴリーはすべて，少なくとも一度は使用される．

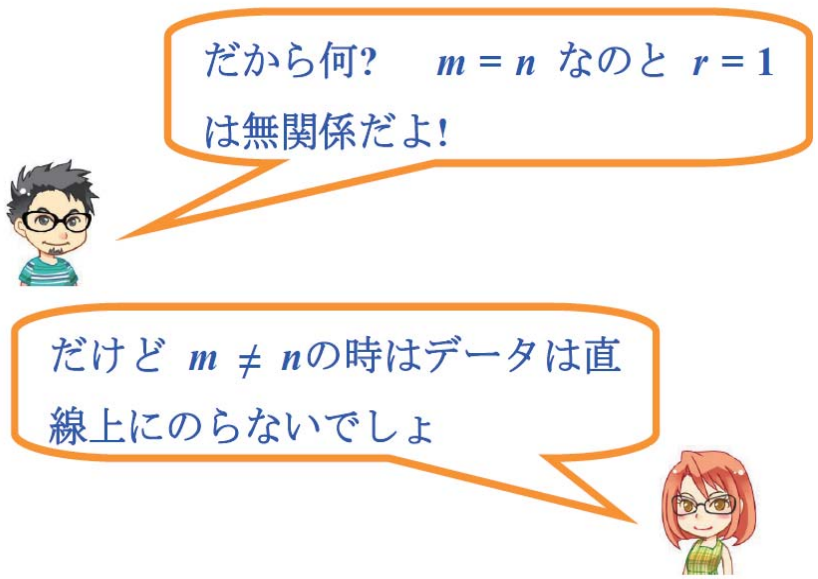
です．断固！

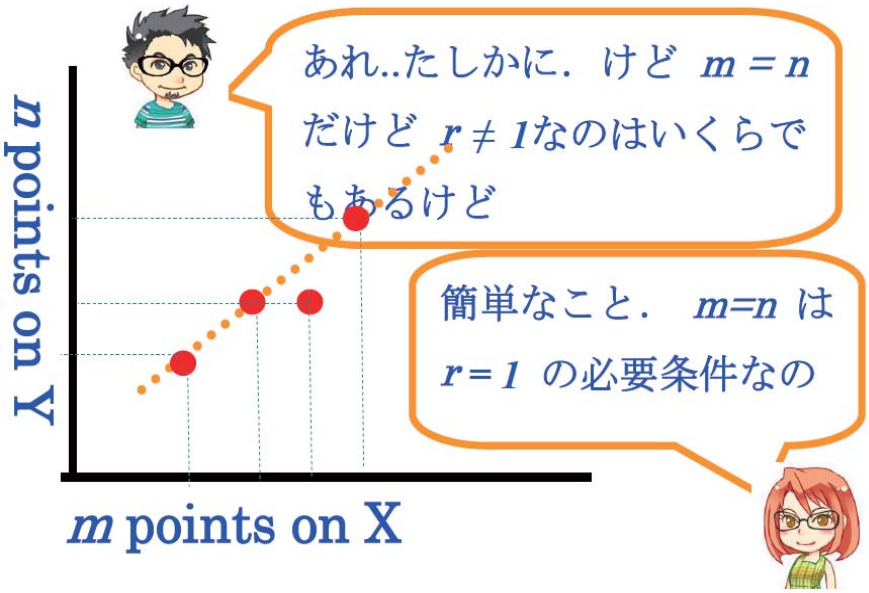
んな　アホな！

どこの本にもそんなこと書いてないし，$m \neq n$で相関を計算するのはザラだよ

でもr=1の時は，データ点は直線上に全部乗るからm=nでしょ．こんな感じ!

Esurientem sum…





証明:

- r の絶対値が1 ⇔ データ点は散布図の(傾いた)直線上にある.
- データ点が散布図の(傾いた)直線上にある. ⇒ データ点をX軸, Y軸へ正射影すると, 両軸上の点の数は等しい
- これより,
- **r の絶対値が 1 ⇒ データ点をX軸, Y軸へ正射影すると, 両軸上の点の数は等しい**
- が成り立つ, この対偶を取ると
- **データ点をX軸, Y軸へ正射影した時, 両軸上の点の数が等しくない ⇒ r の絶対値は1でない**. Q.E.D.
- r が負の場合も同様

FAQ

Q. $m \neq n$の時に相関係数にバイアスがかかるという解釈もできそうですが, 実際にはどのくらいの大きさですか?

A. 特にρ が大きい時に, r の値を押し下げます.

次の表を見てください. 斜字部分がバイアスが明らかな数字です. 斜字以外の場所にもバイアスはあるはずですが判然といたしません.

Table 1 Estimated *r*'s. Italicized numbers show clear decreasing due to the unequal category bias. D = 1024.

	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.96$	$\rho = 0.98$	$\rho = 1.0$
$m = 3, n = 3$	0.615	0.728	0.828	0.878	1.00
$m = 3, n = 4$	0.629	*0.704*	*0.748*	*0.760*	*0.764*
$m = 3, n = 5$	0.658	0.743	*0.786*	*0.792*	*0.794*
$m = 3, n = 6$	0.669	0.755	*0.808*	*0.824*	*0.832*
$m = 4, n = 4$	0.729	0.835	0.904	0.933	1.00
$m = 4, n = 5$	*0.718*	*0.810*	*0.865*	*0.881*	*0.892*
$m = 4, n = 6$	0.732	*0.827*	*0.886*	*0.909*	*0.956*
$m = 5, n = 5$	0.733	0.828	0.893	0.925	1.00
$m = 5, n = 6$	0.745	0.840	0.897	*0.916*	*0.937*
$m = 6, n = 6$	0.758	0.854	0.915	0.940	1.00

Q. 順序カテゴリーでピアソンの相関を計算するのがいけないのでは？

A. データがたとえ物理尺度（比例尺度）で測られていても $m \neq n$ならば，この現象は生起します．ですから，データの出現可能値，あるいは観察されたデータ値，の数が$m \neq n$なら*生起する*と言った方がいいかもしれません．

Q. カテゴリーの数が大きい（50とか100）の場合は？

A. それらのカテゴリーが実際に使用されるならば，この現象はほとんど探知不能になります

Q. では，この現象はどんな時に実際上問題になりますか？

A.　社会科学で評定尺度法を用いて，さらに尺度の段階数が異なる場合

テストAは0から3点まで，テストBは0から4点までの得点が与えられる時

等々

カテゴリーの数が少ない場合です．実際のデータ解析でも問題となるケースは相当にありそうです．

Q. Polychoric correlationを用いればいいのでは？

A. 一応その通りです．ただしPolychoric correlationには厳しいモデル制約があります．

Q.今までにこの現象を指摘した人はいないのですか？

A.相当数の書籍を調べ，またエキスパートにも尋ねましたが，2013年3月現在まだみつけていません．

Q とりあえず m=n にしておけば，大丈夫でしょうか？

A　そのほうがベターですが，大丈夫かどうかは保証できません．なぜなら，実際に使われたカテゴリー数が異なるならば，やはり同じ問題が起こるからです．

Q. どの統計の本にも書いてある

$-1 \leqq r \leqq 1$

は誤りということになりませんか？

A.　$m=n$ が $r=1$ の必要条件なのは数学的には 明らかです．データの取り方によっては，$m=n$が保障されず故に $-1 \leqq r \leqq 1$　にはならないという性質は，反例とも言えます．しかし，この問題は数理的問題というより，むしろ統計実践上の問題でしょう．とにかく，rを使用する人は知っているべきと考えます．

参考文献

カテゴリー数の異なる順序カテゴリー尺度同士の相関係数の性質

http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/35805/1/GakujutsuKenkyu_Jinbun_60_Shiina.pdf

$r \neq 1$の証明は本文書のものがより簡潔です．