気象情報解析特論第11回

相関係数と回帰係数の検定と推定

神山 翼, @t_kohyama,
tsubasa@is.ocha.ac.jp,
理3-703

今日は，二つの変数の関係に意味があるかを統計的に検出する方法を学びます

相関係数と回帰係数の検定と推定

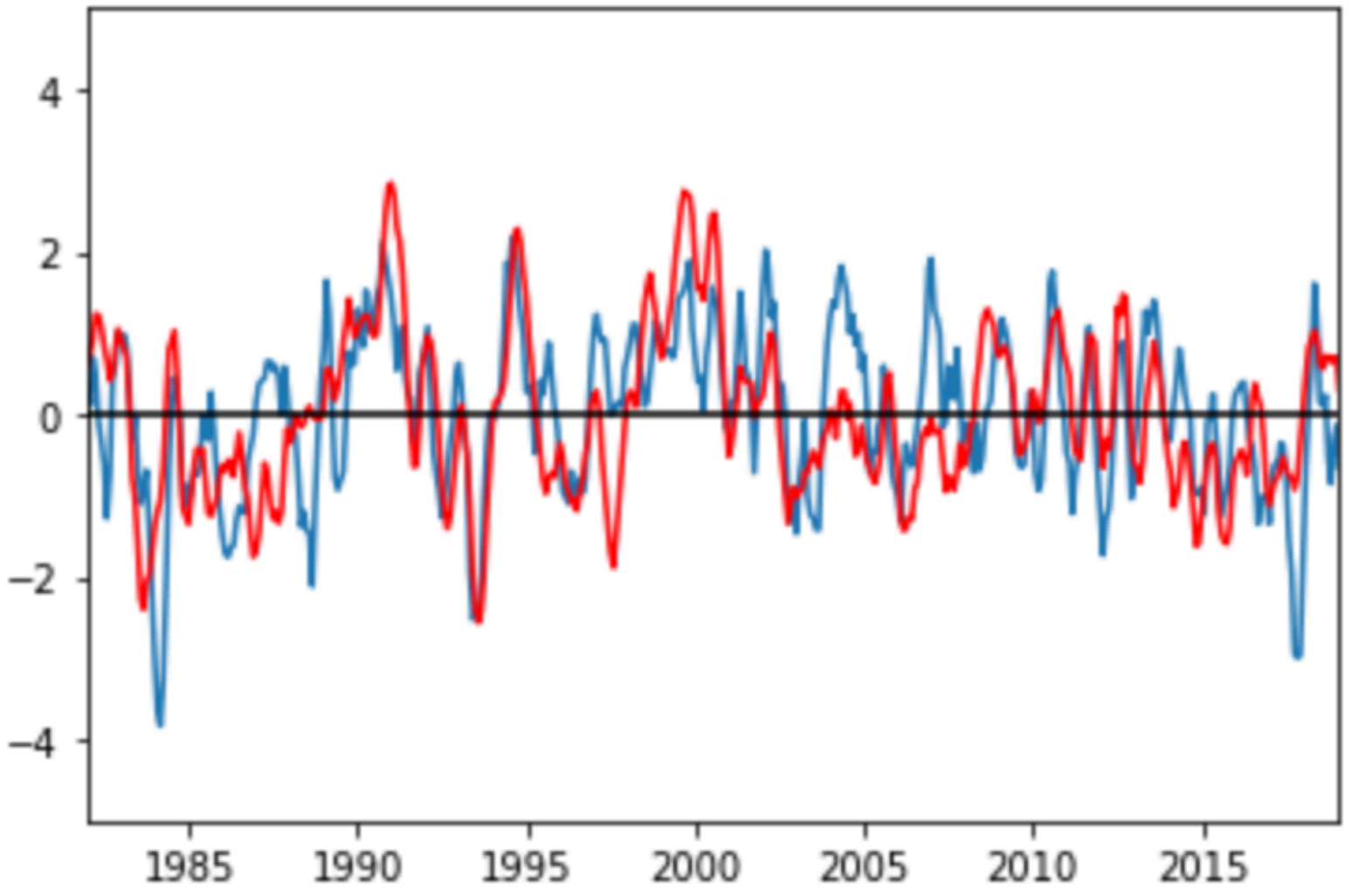
二つの気候変数のサンプルから見積もられた関係が
本当は関係なくてもランダムな変動で生じうるのか
本当に関係があることを意味するのかを調べる

相関係数の検定結果がポジティブであっても
やみくもに検定結果を信じてはいけない

統計検定の考え方がわかれば色んな統計量に応用できる

気温と海面水温のビミョーな関係

東京の気温(青)と日本東方沖の海面水温（赤）



0.586という非常に微妙な相関係数

この相関には意味がある？

相関係数についての統計検定

今日もt検定を利用します

もし東京の気温と日本東方沖の海面水温が
何も関係ない（母相関がゼロ）と仮定

サンプリング変動でこの程度の
サンプル相関は生じうるのか
（どのくらい生じづらいのか）をテスト

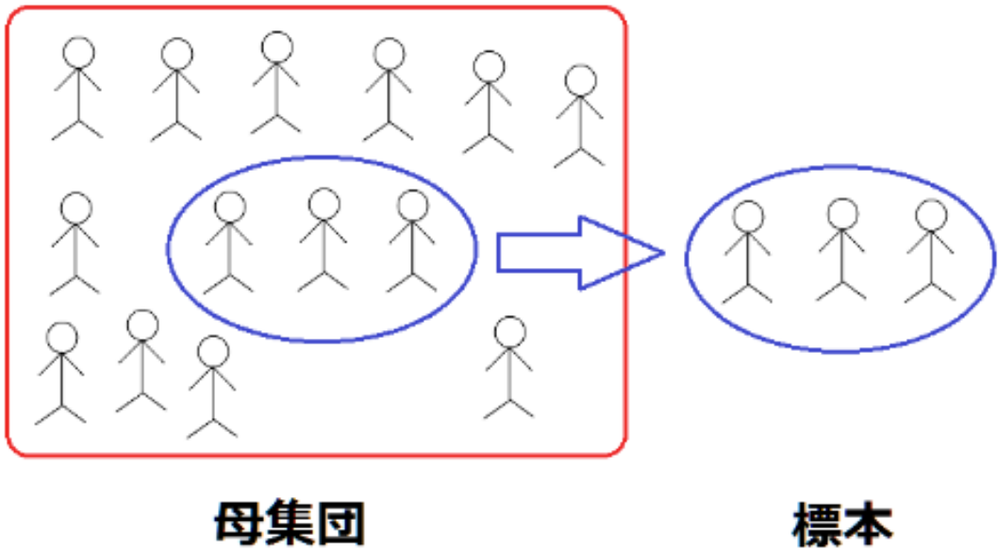
復習：気候解析における母集団と標本

母集団

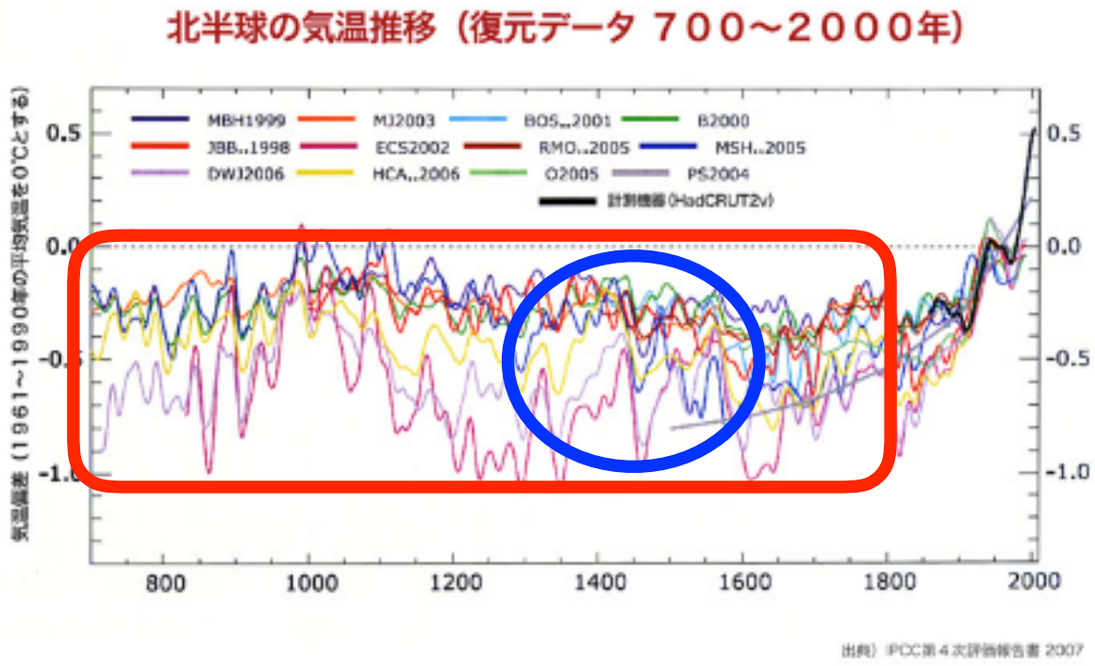
外部強制が同一の気候の時代が無限に続くと
考えたときの気象データ全体の分布

サンプル（標本）

実際に利用できる独立なデータ



https://ai-trend.jp/basic-study/basic/sample-space/



https://www.jccca.org/download/13313

相関係数のt検定

t検定の手順

1. 要求する有意水準を決める

どの程度の確率で起こることを「十分確率が低い」と判定するかの基準（気象学では普通0.05= 信頼度95%）

2. 帰無仮説をつくる

棄却するためにわざと作る仮説

「東京の気温と日本東方沖の海面水温の間の母相関係数はゼロである」

3. 対立仮説をつくる

帰無仮説を棄却することによって，示したい仮説

「東京の気温と日本東方沖の海面水温の間の母相関係数はゼロでない」

t検定の手順

4. 検定統計量を計算

帰無仮説が正しいと仮定するとt分布に従う量

$$t = \frac{r}{\sqrt{\frac{1-r^2}{N^*-2}}}$$

（適宜統計の教科書を参照すること）

5. 検定統計量と棄却域を比較し，帰無仮説を棄却するかどうかを判断

「もし帰無仮説が正しいなら，これ以上検定統計量tがt分布の中心から外れる確率は低いので，帰無仮説は正しくないのだろう」と判断する領域 = 棄却域

レッドノイズの実効的サンプルサイズを求める公式（Bretherton et al. 1999）

$$N^* = N\frac{1 - r_1 r_2}{1 + r_1 r_2}$$

前回と微妙に違う！

N^*：実効的サンプル数

N：元のサンプル数

r_1, r_2：時間ステップを1つずらした時系列と元の時系列の相関係数（「ラグ1自己相関」）

相関係数の検定する際の注意点

アポステリオリな恋

「スキー場で恋に落ちた」はロマンチックか？

たまたまなんとなく気に入っていた相手に
スキー場でバッタリ会った
＝アプリオリな期待があったのでロマンチック

スキー場でナンパし続けた
＝アポステリオリに「スキー場で出会う運命」
を作り上げただけ

アポステリオリな議論

「有意な相関係数の大きさ」を知ってから，その条件に合うような時系列を見つけて来るのは簡単

(例) 火星との距離と降水量の相関が有意になるような場所を見つけたい

➡ 100地点の降水を95%信頼度で検定すれば5地点くらいは有意な相関が見つかる

「大手町と火星との距離に有意な相関がありまぁす！」

= **a posteriori reasoning**

アプリオリな期待

検定結果を信じるためには
「統計以外に基づく期待」が必要

(例)「火星から大手町に降り注ぐ未知の物質が存在することが物理学の法則で予言された」
(= **a priori expectation**)

➡物理学によって選ばれし観測地点が
統計学的にも100地点から偶然に選ばれた5地点のうちの一つと一致した

相関関係は因果関係ではない
(Correlation is not causation)

AとBが相関していても

AとBに因果関係があるとは限らない

共通の原因Cによるだけかもしれない

(例1) 警察官が増えると110番通報が増える

（共通の原因C ＝ 人口）

(例2) 体重が重いほど年収は高い

（共通の原因C ＝ 年齢）

相関係数の推定

フィッシャーのZ変換

サンプル相関係数rから正規分布に従う量Zに変換

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad \sigma_z = \frac{1}{\sqrt{N^* - 3}}$$

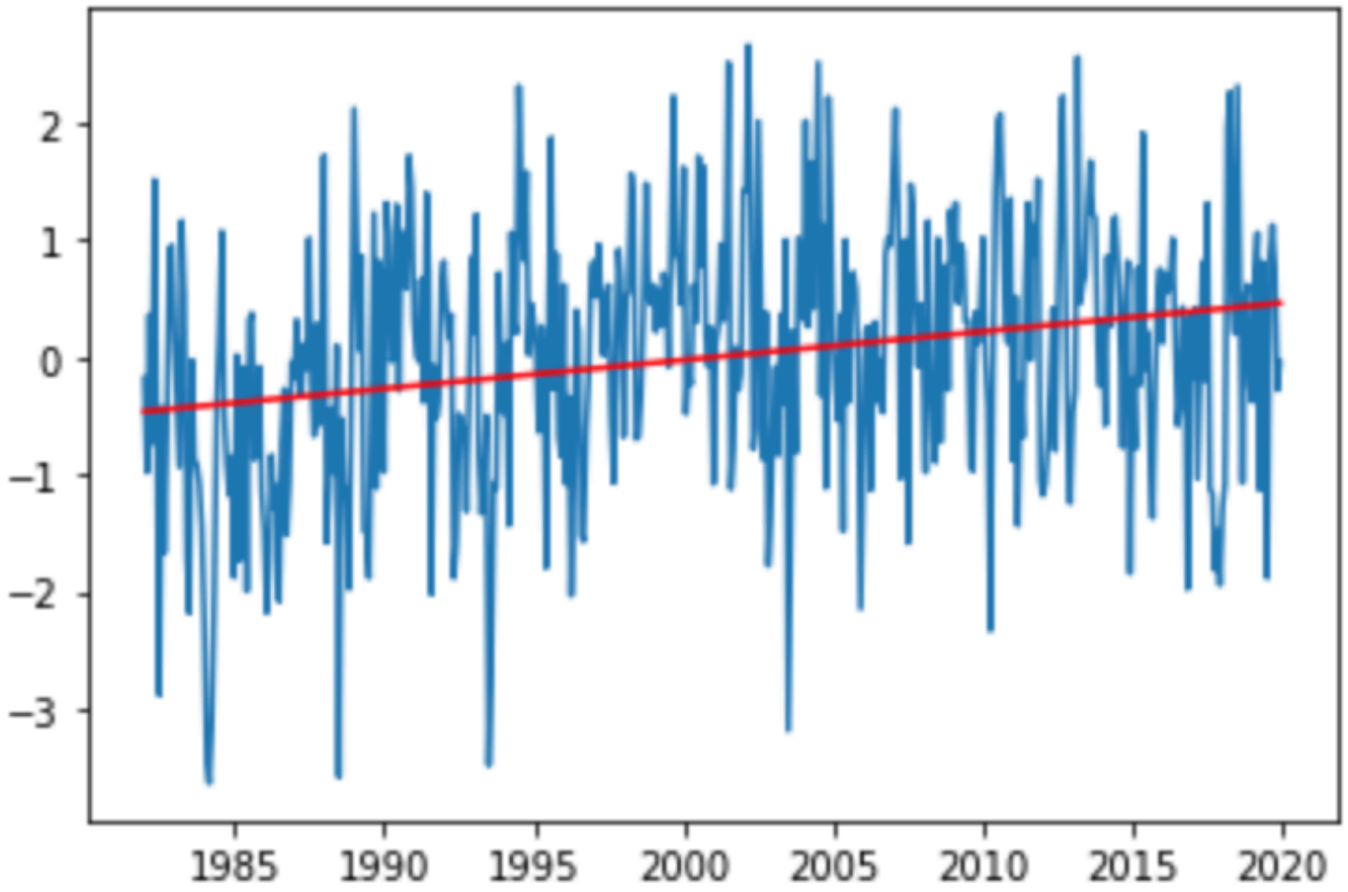
逆変換して母相関係数の95%信頼区間を求める

$$\tanh(Z - 1.96\sigma_Z) < \rho < \tanh(Z + 1.96\sigma_Z)$$

回帰係数の検定と推定

この上昇トレンドは偶然？

1982年から2019年の東京の気温



もう慣れてきましたか？

t検定の手順

1. 要求する有意水準を決める

どの程度の確率で起こることを「十分確率が低い」と判定するかの基準（気象学では普通0.05= 信頼度95%）

2. 帰無仮説をつくる

棄却するためにわざと作る仮説

「1982年から2019年の東京の気温のトレンドはゼロである」

3. 対立仮説をつくる

帰無仮説を棄却することによって，示したい仮説

「1982年から2019年の東京の気温のトレンドはゼロでない」

t検定の手順

4. 検定統計量を計算

帰無仮説が正しいと仮定するとt分布に従う量

$$t = \frac{a}{\left(\frac{\sigma_e}{\sqrt{N}\sigma_x} \right)}$$

ただし

$$\sigma_e = \sigma_y \sqrt{\frac{N}{N^* - 2}(1 - r^2)}$$

$$N^* = N \frac{1 - r_{1y}}{1 + r_{1y}}$$

5. 検定統計量と棄却域を比較し，帰無仮説を棄却するかどうかを判断

「もし帰無仮説が正しいなら，これ以上検定統計量tがt分布の中心から外れる確率は低いので，帰無仮説は正しくないのだろう」と判断する領域 = 棄却域

回帰係数の推定

t分布を仮定し母回帰係数の95%信頼区間を求める

$$a - t_{975,N^*-2}\frac{\sigma_e}{\sqrt{N}\sigma_x} < \alpha < a + t_{975,N^*-2}\frac{\sigma_e}{\sqrt{N}\sigma_x}$$

下端が0より大きい　　または　　上端が0より小さい

$\Leftrightarrow$　　棄却域　　$|t| > t_{975,N^*-2}$　　（示してみよ）

実は推定すれば検定したことにもなる

今日は，二つの変数の関係に意味があるかを統計的に検出する方法を学びます

相関係数と回帰係数の検定と推定

二つの気候変数のサンプルから見積もられた関係が
本当は関係なくてもランダムな変動で生じうるのか
本当に関係があることを意味するのかを調べる

相関係数の検定結果がポジティブであっても
やみくもに検定結果を信じてはいけない

統計検定の考え方がわかれば色んな統計量に応用できる

本日の導入パートは以上です。
何でも良いので渡した紙に
授業に関係のあるコメントを
してください（出席代わり）。

コメント拾いが終わったら，
早速今日のプログラミングに進みましょう。