黒木玄 **Gen Kuroki** @genkuroki on May 16

メモ：話題として思いついたもの

(1) ピアソンのカイ二乗検定量がどうして「平均 0、分散 1 になるように正規化されたポアソン分布の2乗の和」と形式的に同じ形になっているのか？

ピアソンのカイ二乗検定量は

$$\sum_{i=1}^{r}\frac{(O_i-E_i)^2}{E_i}$$

の形をしており、ポアソン分布では平均と分散が等しいので、平均が λ のポアソン分布に従う確率変数 X について

$$\frac{X-\lambda}{\sqrt{\lambda}}$$

は平均 0、分散 1 になります。ピアソンのカイ二乗統計量の各項は形式的にこの型の式の二乗の形になっている。これはなぜかという話。

これも本質的に線形代数に帰着する話とみなせます。

(2) 代数幾何における特異点解消がある種の積分の漸近挙動の解析に役に立つという話について。

確か、Atiyahさんが若い頃に指摘したことだったと思う。

黒木玄 **Gen Kuroki**
@genkuroki

以下は易しい。

(1.1) 中心極限定理より、λ が大きなとき、期待値 λ のポアソン分布に従う確率変数 X に対して、$(X-\lambda)/\sqrt{\lambda}$ は近似的に標準正規分布に従うので、その二乗 $(X-\lambda)^2/\lambda$ は近似的に自由度1のカイ二乗分布に従う。

(1.2) ゆえに、λ_k 達が大きなとき、X_k が期待値 λ_k のポアソン分布に従い、$X_1,\ldots,X_K$ が独立ならば、

$$\sum_{k=1}^{K}\frac{(X_k-\lambda_k)^2}{\lambda_k}$$

は自由度 K のカイ二乗分布に近似的に従う。

ここまでは易しい。

多項分布やその一般化である $M \times N$ の分割表の分布におけるピアソンのカイ二乗検定を理解するためには以上の易しい事実の理解だけでは足りない。

多次元正規分布に関連して、ちょっとした線形代数が必要になる。カイ二乗分布の「自由度」がどのような仕組みで下がるかという話(易しい)。

2017年05月16日 16:42 · Web · 0 · 3 · Webで開く

黒木玄 **Gen Kuroki** @genkuroki on May 16

こっちは長年数学に関わっていて色々知っている上に、大抵の事柄を直観的に理解しているから(これ大事)、統計学を学ぶときに、数学的には大して困らないのですが、学生の側は本当に大変だと思います。

ピアソンのカイ二乗検定は空気のごとく使われるのに、それについてしっかり直観的に理解するためには大学1年でやる微積分や線形代数に関する素養が色々必要になってしまう。

ぶっちゃけた話として、空気のごとくみんな使っているピアソンのカイ二乗検定について、理解せずに使っている人が多いんじゃないかと思います。

黒木玄 **Gen Kuroki** @genkuroki on May 18

Pearsonのカイ二乗検定の理解について, 私なりの解答を説明します. 長文を大放流します！

1. カイ二乗分布の理解

カイ二乗分布の定義まで戻って考え直します.

多くの教科書では

「標準正規分布に従う独立な確率変数達 $X_1, \ldots, X_n$ に対して, $X_1^2 + \cdots + X_n^2$ が従う確率分布を自由度 n のカイ二乗分布と呼ぶ」

というスタイルで, カイ二乗分布を定義します. これを定義として採用していなくても, 必ずこの特徴付けは説明されている.

しかし, この説明だけでは, カイ二乗分布の現実への応用を十分に理解できません. そのような説明は天下り的に与えられた正規直交基底だけで一般の内積について解説しているようなものです.

現実のカイ二乗検定では, n 次元の標準正規分布に従う $(X_1, \ldots, X_n)$ ではなく, 任意の多次元正規分布を扱う可能性があります.

黒木玄 **Gen Kuroki** @genkuroki on May 18

少しだけ線形代数を実行すると一般に次が言えます.

定理：$\Sigma = [\sigma_{ij}]$ は固有値がすべて正であるようなサイズ s の実対称行列であり, $\Sigma^{-1} = [b_{ij}]$ と書き, $(X_1, \ldots, X_s)$ は確率密度函数

$$\frac{\exp\left(-\frac{1}{2}\sum_{i,j=1}^{s} b_{ij}x_i x_j\right)}{\sqrt{\det(2\pi\Sigma)}}$$

で定まる確率分布に従うと仮定する. このとき,

$$Y = \sum_{i,j=1}^{s} b_{ij} X_i X_j$$

は自由度 s のカイ二乗分布に従う.

要するに台が s 次元の正規分布の密度函数の $\exp$ の中身に対応する確率変数が従う確率分布を自由度 s のカイ二乗分布の定義だと思ってよいわけです. その定義のもとで, 自由度 s のカイ二乗分布は多次元正規分布の台の次元だけで決まり, 多次元正規分布の具体形にはよりません.

黒木玄 **Gen Kuroki** @genkuroki on May 18

これを理解していると, カイ二乗分布が自然に出て来まくる理由がクリアになります. 多次元正規分布が中心極限定理によって自然に出て来ます. だから, カイ二乗分布も自然に出て来てしまうわけです. 「標準正規分布に従う独立な確率変数達」のような特殊な設定はカイ二乗分布にとって本質的ではありません.

黒木玄 **Gen Kuroki** @genkuroki on May 18

さらに一般の多次元正規分布においては, 確率変数が $X_1, \ldots, X_n$ まで n 個あるのに, それらが一次従属になっていて, 台の次元が n 未満になっていることがよくあります. そのような場合には確率密度函数は超函数になり,

$$\exp\left(-\frac{1}{2}\sum_{i,j=n}^{s} b_{ij} x_i x_j\right)$$
$$\times \left(s \text{ 次元の台を持つデルタ函数の定数倍}\right)$$

の形になります. この場合にもちょっとした線形代数で

$$Y = \sum_{i,j=1}^{n} b_{ij} X_i X_j$$

が自由度 s のカイ二乗分布に従うことがわかります. 要するに $\exp$ の中身に注意を払うだけで何が自由度幾つのカイ二乗分布に従うかがわかるわけです.

黒木玄 **Gen Kuroki** @genkuroki on May 18

2. Poisson分布の理解

K は非負の整数に値を持つ離散確率変数とします.

K がPoisson分布に従うとはある正のパラメーター λ について確率が

$$P(K=k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

を満たすことでした.

k が大きなとき, Stirlingの公式

$$\log k! \sim k\log k - k + \log\sqrt{2\pi k} + o(1)$$

より

$$\begin{aligned}&\log P(K=k)\\&=-k\log\frac{k}{\lambda}+k-\lambda-\log\sqrt{2\pi k}+o(1)\end{aligned}$$

となります. Kullback-Leiblerダイバージェンスをご存じなら, これの第一項にそれらしきものがすでに見えていることがわかります.

黒木玄 **Gen Kuroki** @genkuroki on May 18

実際, KLダイバージェンスに関するSanovの定理はこういう計算を経由して出て来ると考えることもできます. 確率・統計をちょっとかじったことがあれば, Poisson分布くらいはみんな知っているのですが, KLダイバージェンスの構成要素が隠れていることが指摘されることは少ないと思います.

Poisson分布の中心極限定理は確率変数 X を

$$X=\frac{K-\lambda}{\sqrt{\lambda}}$$

と定義すれば得られます. このとき, $K=\lambda+\sqrt{\lambda}\,X$ なので $k=\lambda+\sqrt{\lambda}\,x$ を $\log P(K=k)$ に代入すると, λ が大きなとき,

$$\log P(K=k)=-\frac{x^2}{2}-\log\sqrt{2\pi\lambda}+O\left(\frac{1}{\sqrt{\lambda}}\right)$$

これは確率変数 X が近似的に標準正規分布に従うことを意味しています.

黒木玄 **Gen Kuroki** @genkuroki on May 18

3. 多項分布と分割表分布の理解

非負の整数に値を持つ離散確率変数達 $K_1,\dots,K_r$ が多項分布に従うとは, それらが常に

$$K_1+\cdots+K_r=n=(\text{一定})\qquad(*)$$

を満たしており, $k_1+\cdots+k_r=n$ のとき

$$P(K_i=k_i\,\forall\,i)\propto\prod_{i=1}^{r}\frac{(np_i)^{k_i}}{k_i!}$$

を満たしていることです. 各 p_i は非負でそれらの総和は 1 であるとします. 比例定数は確率の総和が 1 になるという条件で決まります.

要するに, Poisson分布の直積を束縛条件$(*)$で制限して得られる条件付き確率分布が多項分布なわけです. だから, 多項分布はPoisson分布の言葉で理解可能です.

黒木玄 **Gen Kuroki** @genkuroki on May 18

もっと一般の場合を扱いましょう. 非負の整数に値を持つ離散確率変数達 K_{ij} が, 常に

$$\sum_{j=1}^{N} K_{ij} = m_i,$$

$$\sum_{i=1}^{M} K_{ij} = n_j,$$

$$\sum_{i=1}^{M}\sum_{j=1}^{N} K_{ij} = n$$

が一定という束縛条件を満たし, 同様の条件を満たす非負の整数達 k_{ij} について

$$P(K_{ij} = k_{ij}\,\forall\, i,j) \propto \prod_{i,j} \frac{(np_{ij})^{k_{ij}}}{k_{ij}!}$$

を満たす状況を考えることも結構あります. ここで p_{ij} は非負でそれらの総和は 1 で,

$$\sum_{j=1}^{N} np_{ij} = m_i,$$

$$\sum_{i=1}^{M} np_{ij} = n_j$$

を満たしているとします.

黒木玄 **Gen Kuroki** @genkuroki on May 18

このとき, K_{ij} 達が $M \times N$ の表に分割されているとみなせるので, この確率分布を「分割表分布」と呼ぶことにしましょう(標準的な呼び名があるかどうかは知らない). これもPoisson分布の直積の条件付き確率分布になっています. だから, Poisson分布の言葉で理解できるはずです. 以下では m_j, n_i, np_{ij} 達は大きいと仮定します.

Poisson分布に関する中心極限定理より, 確率変数 X_{ij} を

$$X_{ij} = \frac{K_{ij} - np_{ij}}{\sqrt{np_{ij}}}$$

と定めると, 各々の X_{ij} は近似的に標準正規分布に従います. しかし, それらは

$$\sum_{i=1}^{M} \sqrt{p_{ij}}\, X_{ij} = 0,$$

$$\sum_{j=1}^{N} \sqrt{p_{ij}}\, X_{ij} = 0$$

という束縛条件を満たしていなければいけません. この束縛条件で独立なのは $M + N - 1$ 個です.

黒木玄 **Gen Kuroki** @genkuroki on May 18

したがって, X_{ij} 達は MN 個の標準正規分布の直積を $M+N-1$ 個の線形な束縛条件で制限して得られる $(M-1)(N-1)$ 次元の台を持つ多次元正規分布に近似的に従うことになります. その多次元正規分布の $\exp$ の中身は $-1/2$ を除けば

$$\chi^2 = \sum_{i,j} X_{ij}^2 = \sum_{i,j} \frac{(K_i - np_{ij})^2}{np_{ij}}$$

になります. 束縛条件を課す前は標準正規分布の直積なので当然この形になります. この χ^2 をPearsonのカイ二乗検定量と呼びます.

s 次元の台を持つ多次元正規分布から自由度 s のカイ二乗分布が得られることより, Pearsonのカイ二乗検定量 χ^2 は自由度 $(M-1)(N-1)$ のカイ二乗分布に近似的に従うことになります.

黒木玄 **Gen Kuroki** @genkuroki on May 18

以上のように, Pearsonのカイ二乗検定量の形はPoisson分布の直積から来ており, 束縛条件によってそれが近似的に従うカイ二乗分布の自由度が下がることが, 一般の多次元正規分布とカイ二乗分布の関係から出て来るわけです.

以上の解説でごまかした部分を埋めれば, Pearsonのカイ二乗検定の基礎的理解はばっちりのはずです.

Poisson分布の漸近挙動の分析の途中でKullback-Leiblerダイバージェンスのような式が出て来ましたが, 多項分布のケースにそれを応用するとSanovの定理をPoisson分布の視点から理解できるようになります. その辺の詳しい事情については

github.com/genkuroki/Sanov

の1.8〜1.9節を見て下さい.

大放流終了！

黒木玄 **Gen Kuroki** @genkuroki on May 18

悪のりして超絶極端な大放流をしてしまったwww