13 対立仮説の下でのランク統計量

R_N は連続な分布関数 $F_1, ..., F_N$ からの独立な確率変数 $X_1, ..., X_N$ からのランクベクトルとする．

Theorem 13.5. R_N は連続な分布関数 F に従う i.i.d サンプル $X_1, ..., X_n$ の順位ベクトルとする．スコアベクトル a_N は $a_{Ni} = E\phi(U_{N(i)})$ により生成されているとする．ただし ϕ はほとんど至る所で定数関数でない，

$$\int_0^1 \phi^2(u)du < \infty$$

をみたす可測関数とする．変数

$$T_N := \sum_{i=1}^N c_{Ni} a_{N,R_{Ni}}$$

$$\tilde{T}_N := N\bar{c}_N\bar{a}_N + \sum_{i=1}^N (c_{Ni} - \bar{c}_N)\phi(F(X_i))$$

を定義する．このとき，T_N と $\tilde{T}_N$ は漸近的に同等；$ET_N = E\tilde{T}_N$ かつ $\mathrm{var}(T_N - \tilde{T}_N)/\mathrm{var}T_N \to 0$ である．このことは，ϕ はほとんど至る所で連続で，定数関数でなく，

$$\frac{1}{N}\sum_{i=1}^N \phi^2\left(\frac{i}{N+1}\right) \to \int_0^1 \phi^2(u)du < \infty$$

を満たすとれば，$a_{Ni} = \phi\left(\frac{i}{N+1}\right)$ でスコアが定義されるときも成り立つ．

により，スコア生成関数に対する緩い条件下で，シンプル線形ランク統計量の漸近分布が得られた．しかし分布関数 F_i に対しては全て等しいという強い仮定を置いていた．これは，同一分布であるという帰無分布に対して棄却域を得るには十分であった．しかし，漸近的な有効性を議論するには，対立仮説のもとでの漸近的振る舞いを調べる必要がある．例えば，２標本問題において，$F, ..., F, G, ..., G$ という対立分布の下での漸近分布に興味がある．

帰無仮説に “十分速く”収束するような対立仮説に対する最善のアプローチは Le Cam’s third lemma というものを使うことらしい（一応 exmaple 6.7 に記載がある).

特に，もし帰無仮説 $F, ..., F, F..., F$ に対する対立仮説 $F_n, ..., F_n, G_n, ..., G_n$ の対数尤度比が漸近的に，$\sum \ell_i(X_i)$ の形で推定できるとすれば，ランク統計量の同時漸近分布と帰無仮説の下での対数尤度比は多変量の中心極限定理とスラツキーの定理から求まる．なぜなら，Theorem 13.5 はこのランク統計量に対して同様の近似が可能だからである．次に，Le Cam’s third lemma をランク統計量の極限分布を見つけるのに適用する．このアプローチは比較的簡単で多くの興味のあるケースで利用可能である．たとえば　7.5 節や 14.1.1 節を見るとよい．

より一般の対立仮説も直接扱えるはずであり，そのときはより強いスコア生成関数への条件が必要に思える．１つの方法はランク統計量を観測の経験分布関数 $\mathbb{F}_N$ と，重み付き経験分布 $\mathbb{F}_N^c(x) = \frac{1}{N}\sum_{i=1}^N c_{Ni} 1\{X_i \le x\}$ の関数として書くことである．

$R_{Ni} = N\mathbb{F}_N(X_i)$ より，

$$\frac{1}{N}\sum_{i=1}^N c_{Ni} a_{N,R_{Ni}} = \int a_{N,N\mathbb{F}_n(x)} d\mathbb{F}_N^c(x).$$

となる．次に，経験分布が Brownian bridge に収束することを示すのに von Mises 分析を行う．この方法は

Chapter 20 で一般的に説明される．

この章では Hajek's projection lemma に基づく別の方法を説明する．技術的な難しさをさけるため滑らかなスコア生成関数に制限する．$\bar{F}_N$ を $F_1, ..., F_N$ の平均とし，$\bar{F}_N^c$ を重み付き和 $\frac{1}{N}\sum_{i=1}^N c_{Ni}F_i$ とし，

$$T_N := \sum_{i=1}^N c_{Ni}\phi\left(\frac{R_{Ni}}{N+1}\right),$$
$$\hat{T}_N := \sum_{i=1}^N \left[c_{Ni}\phi(\bar{F}_N(X_i)) + \int_{X_i}^\infty \phi'(\bar{F}_N(x))d\bar{F}_N^c(x)\right].$$

我々は変数 $\hat{T}_N$ が T_N の近似の Hajek projection であることを示す．近似しない T_N 自身の Hajek projection はよりよい近似を与えるがより複雑なものになる．

Lemma 13.23. $\phi : [0,1] \mapsto \mathbb{R}$ は 2 階連続微分可能とする．このとき，普遍定数 K が存在して，

$$\mathrm{var}(T_N - \hat{T}_N) \le K\frac{1}{N}\sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2(||\phi'||_\infty^2 + ||\phi''||_\infty^2).$$

Proof. まず不等式は任意の固定された N に対して成り立つので，証明においてインデックス N は省略する．さらに主張は分散に関心があり，T_N と $\hat{T}_N$ どちらも c_{Ni} が $c_{Ni} - \bar{c}_N$ に置き換わっても定数でしか変化しないので $\bar{c}_N = 0$ と仮定する．

X_i のランクは $R_i = 1 + \sum_{k\neq i} 1[X_k \le X_i]$ と書ける．よって

$$\begin{aligned}\left|E\left(\frac{R_i}{N+1}|X_i\right) - \bar{F}(X_i)\right| &= \frac{1}{N+1}\left|1 + \sum_{k\neq i} E(1[X_k \le X_i])|X_i) - (N+1)\bar{F}(X_i)\right| \\ &= \frac{1}{N+1}\left|1 + \sum_{k\neq i} F_k(X_i) - (N+1)\bar{F}(X_i)\right| \\ &= \frac{1}{N+1}\left|1 - \bar{F}(X_i) - F_i(X_i)\right| \le \frac{1}{N}.\end{aligned}$$

さらに，Marcinkiewitz-Zygmund の不等式，独立な確率変数 $Z_1, ..., Z_n$ は $E(Z_i) = 0$, $E(|Z_i|^p) < +\infty (1 \le p < +\infty)$ とすると

$$A_p E\left(\left(\sum_{i=1}^n |Z_i|^2\right)^{p/2}\right) \le E\left(\left|\sum_{i=1}^n Z_i\right|^p\right) \le B_p E\left(\left(\sum_{i=1}^n |Z_i|^2\right)^{p/2}\right).$$

が成り立つ．ただし，A_p と B_p は p のみに依存する正の定数である．これを用いると

$$
\begin{aligned}
E\left(\frac{R_i}{N+1}-\bar{F}(X_i)\right)^4 &= \frac{1}{(N+1)^4}E\left(1+\sum_{k\neq i}1[X_k\leq X_i]-(N+1)\bar{F}(X_i)\right)^4 \\
&= \frac{1}{(N+1)^4}E\left(1+\sum_{k\neq i}(1[X_k\leq X_i]-F_k(X_i))+N\bar{F}(X_i)-F_i(X_i)-(N+1)\bar{F}(X_i)\right)^4 \\
&= \frac{1}{(N+1)^4}E\left(\sum_{k\neq i}(1[X_k\leq X_i]-F_k(X_i))+1-\bar{F}(X_i)-F_i(X_i)\right)^4 \\
&\preceq \frac{1}{N^2}EE\left(\frac{1}{N}\sum_{k\neq i}(1[X_k\leq X_i]-F_k(X_i))^4|X_i\right)+\frac{1}{N^4}\preceq\frac{1}{N^2}.
\end{aligned}
$$

最後の段の 1 つ目の不等式を説明する．まず，Marcinkiewitz-Zygmund の不等式を X_i 条件づけたものに適用すれば，

$$
\begin{aligned}
E\left(\sum_{k\neq i}(1[X_k\leq X_i]-F_k(X_i))\right)^4 &= E\left(E\left(\sum_{k\neq i}(1[X_k\leq X_i]-F_k(X_i))\right)^4|X_i\right) \\
&\leq E\left(E\left(\sum_{k\neq i}(1[X_k\leq X_i]-F_k(X_i))^2\right)^2|X_i\right) \\
&\leq E\left(E\left(\sum_{k\neq i}(1[X_k\leq X_i]-F_k(X_i))^4\right)|X_i\right)
\end{aligned}
$$

であり，先に示した，$\frac{1}{N+1}\left|1-\bar{F}(X_i)-F_i(X_i)\right|\leq\frac{1}{N}$ と合わせると，定数倍程度で不等式が成り立つことが分かる．

次に，$T=\sum_{i=1}^N c_i\phi\left(\frac{R_{Ni}}{N+1}\right)$ の ϕ を $\bar{F}(X_i)$ で 2 階テイラー展開する．

$$
\begin{aligned}
T &= \sum_{i=1}^N c_i\phi(\bar{F}(X_i))+\sum_{i=1}^N c_i\left(\frac{R_{Ni}}{N+1}-\bar{F}(X_i)\right)\phi'(\bar{F}(X_i))+\sum_{i=1}^N c_i\left(\frac{R_{Ni}}{N+1}-\bar{F}(X_i)\right)^2K_i \\
&=: T_0+T_1+T_2.
\end{aligned}
$$

ただし，確率変数 $K_i<||\phi''||_\infty$．それぞれの項についてみていく．

まず，T_0 は $\hat{T}_N$ の 1 項目と相殺される．

さらに，$T_2=\sum_{i=1}^N c_i\left(\frac{R_{Ni}}{N+1}-\bar{F}(X_i)\right)^2K_i$ はシュワルツの不等式により

$$
T_2\leq\left(\sum c_i^2K_i^2\right)\left(\sum\left(\frac{R_{Ni}}{N+1}-\bar{F}(X_i)\right)^4\right)
$$

となり，先に 4 次モーメントがバウンドされることは求めたので，lemma のように 2 次平均でバウンドされる．

最後に，$T_1=\sum_{i=1}^N c_i\left(\frac{R_{Ni}}{N+1}-\bar{F}(X_i)\right)\phi'(\bar{F}(X_i))$ が，漸近的に Hajek projection で，さらに $\hat{T}$ の 2 項目に漸近的に（定数誤差を除いて）等しいことを示す．

まず，10/13 のゼミを思い出す．独立な確率ベクトル $X_1,...,X_n$ に対して，集合

$$
\mathcal{S}=\left\{\sum_{i=1}^n g_i(X_i);g\text{ は }Eg_i^2(X_i)<\infty\text{ を満たす任意の可測関数}\right\}
$$

を定義する．有限な 2 次モーメントを持つ任意の確率変数 T のクラス $\mathcal{S}$ への射影は

$$\hat{S} = \sum_{i=1}^{n} E(T|X_i) - (n-1)ET$$

で与えられる（Lemma 11.10.）．これを Hajek projection という．これより，T_1 の Hajek projection は

$$\begin{aligned}\sum_{i=1}^{N} E(T_1|X_i) - (N-1)ET_1 &= \sum_{i=1}^{N} c_i \sum_{j=1}^{N} E\left[\frac{R_i}{N+1}\phi'(\bar{F}(X_i))|X_j\right] - \sum_{i=1}^{N} c_i \bar{F}(X_i)\phi'(\bar{F}(X_i)) \\ &= \sum_{i=1}^{N} c_i \left(\sum_{j\neq i} E\left[\frac{R_i}{N+1}\phi'(\bar{F}(X_i))|X_j\right]\right) + \sum_{i=1}^{N} c_i \left(E\left(\frac{R_i}{N+1}|X_i\right) - \bar{F}(X_i)\right)\phi'(\bar{F}(X_i)).\end{aligned}$$

2 項目は 2 次平均において有界である．さらに 1 項目は $R_i = 1 + \sum_{k\neq i} 1[X_k \leq X_i]$ を代入すると，

$$\frac{1}{N+1}\sum_{i=1}^{N} c_i \sum_{j\neq i} E(1[X_j \leq X_i]\phi'(\bar{F}(X_i))|X_j) + \text{constant}$$

と分かる．$N+1$ を N に置き換えて，定数項を消して，条件付期待値を積分表現して，対角要素（$j = i$ のところ）を加えて

$$\begin{aligned}\frac{1}{N+1}\sum_{i=1}^{N} c_i \sum_{j\neq i} E(1[X_j \leq X_i]\phi'(\bar{F}(X_i))|X_j) + \text{constant} &\approx \frac{1}{N}\sum_{i=1}^{N} c_i \sum_{j\neq i} E(1[X_j \leq X_i]\phi'(\bar{F}(X_i))|X_j) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}\int_{X_i}^{\infty} \phi'(\bar{F}(x))d\bar{F}^c(x) \\ &\approx \sum_{i=1}^{N}\int_{X_i}^{\infty} \phi'(\bar{F}(x))d\bar{F}^c(x) \\ &= \hat{T} \text{ の 2 項目}\end{aligned}$$

が得られる．2 つの表現の差は 2 次平均でバウンドされる．

あとは T_1 とその Hajek projection の差が無視できることを示せばよい．

そのためにヘフディング分解を考える．10/13 の内容を思い出すと，確率変数 T のヘフディング分解は

$$\sum_{r=0}^{n}\sum_{|A|=r} P_A T$$

で定義される．ただし，$P_A T$ は T の

$$H_A := \{g_A(X_i : i \in A) : E[g_A^2(X_i : i \in A)] < \infty, E[g_A(X_i : i \in A)|X_j : j \in B] = 0(\forall B : |B| < |A|)\}$$

への射影である．$T_1 = \sum_{i=1}^{N} c_i \left(\frac{R_{Ni}}{N+1} - \bar{F}(X_i)\right)\phi'(\bar{F}(X_i))$ に関して，$R_i\phi'(\bar{F}(X_i))$ は空間 $\sum_{|A|\leq 2} H_A$ に含まれる（分からん）．そこで，T_1 とその Hajek projection の差は T_1 の $\sum_{|A|=2} H_A$ への射影に等しいことが分かる（分からん）．

この射影は 2 次モーメント

$$\frac{1}{(N+1)^2}\sum_{|A|=2} E\left(P_A \sum_i c_i \sum_k 1\{X_k \leq X_i\}\phi'(\bar{F}(X_i))\right)^2$$

を持つ．

空間 $H_{[k,i]}$ に含まれる変数 $1\{X_k \le X_i\}\phi'(\bar{F}(X_i))$ の $H_{[a,b]}$ への射影は $\{a,b\} \subset \{k,i\}$ でなければゼロになる．よって上の表現は

$$\frac{1}{(N+1)^2}\sum_{a<b} E\left(c_b 1\{X_a \le X_b\}\phi'(\bar{F}(X_b)) + c_a 1\{X_b \le X_a\}\phi'(\bar{F}(X_a))\right)^2.$$

となり，Lemma の主張のようにバウンドされている．

□

この補題の結果から，列 $\frac{T_N - ET_N}{\mathrm{sd}T_N}$ と列 $\frac{\hat{T}_N - E\hat{T}_N}{\mathrm{sd}\hat{T}_N}$ が漸近同等となるのは

$$\frac{\sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2}{N\mathrm{var}\hat{T}_N} \to 0$$

となるときであることが分かる（Theorem 13.5. の漸近同等の定義を参照）．

この条件は，観測が同一分布であれば満たされる．よってランクベクトルは順列上で一様に分布する．そして，Lemma 13.1 によって与えられた $\mathrm{var}T_N$ の陽な表示

$$\mathrm{var}T_N = \frac{1}{N-1}\sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2 \sum_{i=1}^N (a_{Ni} - \bar{a}_N)^2.$$
$$\text{where.} \qquad a_{N,i} = \phi\left(\frac{i}{N+1}\right)$$

は左辺は $O(1/N)$ であることを示している．

lemma の条件下で，近似

$$ET_N \approx \bar{c}_N \sum_{i=1}^N \phi\left(\frac{i}{N+1}\right) + \sum_{i=1}^N (c_{Ni} - \bar{c}_N) E\phi(\bar{F}_N(X_i)).$$

が得られる．差の 2 乗は lemma の上限によりバウンドされる．

先の lemma は滑らかなスコア生成関数に制限されている．より一般のスコアへの拡張は興味のあるランク統計量とランク統計量による適切な近似の差が小さいことを示すことである．以下の補題はこの目的において有用である．（ただし同一分布のときは suboptimal であるらしい）

Lemma 13.24. Variance inequality 非減少な係数 $a_{N1} \le \cdots \le a_{NN}$ と任意のスコア $c_{N1}, ..., c_{NN}$ に対して，

$$\mathrm{var}\sum_{i=1}^N c_{Ni} a_{N,R_{Ni}} \le 21 \max_{1\le i\le N}(c_{Ni} - \bar{c}_N)^2 \sum_{i=1}^N (a_{Ni} - \bar{a}_N)^2.$$