Transformacje stabilizujące wariancję

Przypuśćmy, że mamy k niezależnych zmiennych losowych $X_1,...,X_k$ z rozkładów $N(m_i,\sigma_i^2)$ i=1,...,k, przy czym zakładamy, że $\sigma_i=f(m_i)$ i f jest znaną funkcją. W praktyce możemy znać tę funkcję, gdyż znając sposób pomiaru, możemy np. stwierdzić, że błąd pomiaru wyrażony odchyleniem standardowym jest proporcjonalny do poziomu mierzonej wielkości. W innych przypadkach możemy na podstawie wykresu empirycznej zależności odchylenia standardowego od wartości średniej rozpoznać charakter tej zależności .

Oznaczmy przez m_X i σ_X wartość oczekiwaną i odchylenie standardowe zmiennej losowej X.

Niech

$$\sigma_X = f(m_X),$$

gdzie f jest znaną nieujemną funkcją.

Problem. Chcemy znaleźć taką transformację t, aby ciąg niezależnych zmiennych losowych $Y_1=t(X_1),...,Y_k=t(X_k)$ był ciągiem o stałej (choćby w przybliżeniu) wariancji $V(Y_i)$=const , i=1,...,k.

Stosując rozwinięcie Taylora wokół wartości oczekiwanej możemy napisać

$$Y_i=t(X_i)\approx t(m_i)+ t'(m_i)(X_i- m_i),$$ skąd otrzymujemy

$$V[Y_i] \approx [t'(m_i)]^2\ V[X_i]= [t'(m_i)]^2 f^2(m_i).$$

Warunek $V[Y_i] \approx$ const prowadzi do równania różniczkowego $t'(x)\ f\ (x)=c$, a skąd uzyskujemy już poszukiwaną transformację w postaci

$$t(x) = c_1 \int \frac{dx}{f(x)} + c_2,$$

gdzie stałe c_1 i c_2 można wybrać tak, aby transformacja miała najwygodniejszą postać. Zauważmy, że jeśli $f(x)=x$, to (z dokładnością do stałych) poszukiwaną transformacją jest transformacja logarytmiczna $t(x)=\ln x$. Jeśli $f(x)=x^\alpha$, $\alpha \neq 1$, to poszukiwaną transformacją (z dokładnością do stałych) jest transformacja $t(x) = \frac{1}{1-\alpha} x^{1-\alpha}$. Wykorzystując znany fakt $\lim_{x\to 0} \frac{a^x-1}{x} = \ln a$, możemy określić szeroką klasę transformacji potęgowych znanych jako transformacje Boxa i Coxa w następujący sposób

$$t_\lambda(x) = \begin{cases} \frac{x^\lambda-1}{\lambda}, & \lambda \neq 0 \\ \ln x\ , & \lambda = 0 \end{cases}, \text{ dla } x>0.$$

Warunek $x>0$ nie jest zbytnio krępujący, gdyż zawsze można wstępnie przesunąć zakres obserwowanych wielkości w obszar wartości dodatnich, co odpowiada zastąpieniu argumentu x transformacji $t_\lambda(x)$ przesuniętą wartością $x+c$ tego argumentu. Transformacja $t_\lambda(x)$ odpowiada zależności $f(x)=x^{1-\lambda}$ pomiędzy odchyleniami standardowymi i wartościami oczekiwanymi ciągu zmiennych losowych . Jeżeli rzeczywista zależność jest postaci $f(x)=x^{1-\lambda}$ dla pewnego λ, to parametr ten wyznacza konkretną transformację z rodziny transformacji Boxa-Coxa.

Normalizacyjna transformacja Boxa-Coxa

Transformacji Boxa-Coxa można również użyć jako transformacji normalizującej, tzn. transformacji przekształcającej ciągły rozkład danej zmiennej losowej w rozkład normalny, gdyż jeśli zmienna losowa X ma rozkład z absolutnie ciągłą i ściśle rosnącą dystrybuantą F_X, to zmienna losowa $F_X(X)$ ma rozkład jednostajny na przedziale (0,1), więc zmienna losowa $F^{-1}_{N(m,\sigma^2)}[F_X(X)]$, gdzie $F_{N(m,\sigma^2)}$ oznacza dystrybuantę rozkładu normalnego $N(m,\sigma^2)$, ma rozkład normalny. Zbadajmy jaką rodzinę rozkładów na półprostej $x>0$ można transformować do normalności za pomocą transformacji Boxa-Coxa. Oznaczmy przez F_X i f_X odpowiednio dystrybuantę i funkcję gęstości zmiennej losowej X. Przypuśćmy, że istnieje taka wartość parametru λ, że zmienna losowa $Y= t_\lambda(X)$ ma rozkład $N(m,\sigma^2)$. Z monotoniczności transformacji t_λ wynika, że

$$F_X(x)=P(X\le x)=P(t_\lambda(X)\le t_\lambda(x))=P(Y\le t_\lambda(x))= F_{N(m,\sigma)}(t_\lambda(x)).$$

Wykorzystując postać dystrybuanty rozkładu $N(m,\sigma^2)$ poprzez różniczkowanie powyższej tożsamości uzyskujemy 3-parametrową rodzinę funkcji gęstości transformowalnych do 2-parametrowej rodziny rozkładów normalnych za pomocą transformacji t_λ

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}x^{\lambda-1}e^{-\frac{(t_\lambda(x)-m)^2}{2\sigma^2}},\ x>0.$$

Jeśli zmienna losowa X ma rozkład o funkcji gęstości nie należącej do powyższej klasy funkcji gęstości, to możemy próbować aproksymować nieznaną funkcję gęstości funkcją z powyższej klasy. Innymi słowy transformacja t_λ może być użyta jako przybliżona transformacja do normalności. Nieznany parametr λ można wyznaczyć **metodą największej wiarygodności**, zakładając (choćby w przybliżeniu), że zmienne losowe $Y_1= t_\lambda(X_1),\ldots,Y_n= t_\lambda(X_n)$ są niezależne i mają ten sam rozkład $N(m,\sigma^2)$ o nieznanych parametrach m i σ.

Funkcja wiarygodności ma postać

$$L(m,\sigma,\lambda)=(\tfrac{1}{2\pi})^{\frac{n}{2}}\frac{1}{\sigma^n}(\prod_{i=1}^{n}X_i)^{\lambda-1}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t_\lambda(X_i)-m)^2},$$

a jej logarytm ma postać

$$l(m,\sigma,\lambda)=\ln L(m,\sigma,\lambda)=-\tfrac{n}{2}\ln 2\pi-n\ln\sigma+(\lambda-1)\sum_{i=1}^{n}\ln X_i-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t_\lambda(X_i)-m)^2.$$

Warunek konieczny istnienia ekstremum

$$\frac{\partial l}{\partial m}=0,\ \frac{\partial l}{\partial \sigma}=0,\ \frac{\partial l}{\partial \lambda}=0$$

prowadzi do układu równań

$$\begin{cases} m(\lambda) = \frac{1}{n}\sum_{i=1}^{n} t_\lambda(X_i) \\ \sigma^2(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(t_\lambda(X_i) - m(\lambda))^2 \\ \sigma^2(\lambda)\sum_{i=1}^{n}\ln X_i = \sum_{i=1}^{n}(t_\lambda(X_i) - m(\lambda))\frac{\partial t_\lambda(X_i)}{\partial \lambda} \end{cases}$$

Zamiast rozwiązywać powyższy układ równań można, wykorzystując możliwość dekompozycji zadania maksymalizacji funkcji wiarygodności, wykorzystując dwa pierwsze równania znaleźć rozwiązanie problemu postaci :

$$\hat{\lambda} = \arg\max_{\lambda} l(m(\lambda), \sigma(\lambda), \lambda) = \arg\max_{\lambda} ((\lambda - 1)\sum_{i=1}^{n}\ln X_i - \tfrac{n}{2}\ln\sigma^2(\lambda))$$

Uwaga. Jeżeli wstępnie przeskalujemy zmienne dzieląc każdą z nich przez średnią geometryczną, to optymalne λ wyznaczamy z nieco prostszego warunku

$$\hat{\lambda} = \arg\min_{\lambda} \sigma^2(\lambda).$$

Krótki przegląd konserwatywnych testów post hoc.

Wszystkie rozważane testy konserwatywne prowadzą do następującej reguły

Średnie m_i i m_j są istotnie różne jeżeli $|\overline{X}_i - \overline{X}_j| >$ wartość progowa (właściwa dla danego testu)

Rozważane testy kontrolują prawdopodobieństwa różnych błędów

Test Fishera *NIR* (najmniejsza istotna różnica) (inaczej *LSD* least significant diffrence).

Algorytm

1. Przeprowadzić ANOVA w celu przetestowania H_0: m_1=…= m_k przeciwko alternatywie H_1: co najmniej dwie średnie różnią się między sobą
2. Jeżeli nie ma podstaw do odrzucenia H_0 kończymy analizę.
3. Jeżeli H_0 została odrzucona, to definiujemy najmniejszą istotną różnicę *NIR* (*LSD*) pomiędzy średnimi próbkowymi, którą należy zaobserwować, aby uznać odpowiadające im średnie w odpowiednich populacjach za istotnie rożne.
4. Dla wyspecyfikowanej wartości α w celu porównania m_i z m_j obliczamy *NIR* wg wzoru $NIR = t_{1-\frac{\alpha}{2},\sum_{i=1}^{k} n_i}\sqrt{S_w^2(\frac{1}{n_i}+\frac{1}{n_j})}$, gdzie $t_{1-\frac{\alpha}{2},\sum_{i=1}^{k} n_i}$ jest kwantylem rzędu $1-\frac{\alpha}{2}$ rozkładu *t*-Studenta o $\sum_{i=1}^{k} n_i$ stopniach swobody a S_w^2 jest sumą kwadratów wewnątrz grup (z ANOVA)
5. Porównujemy wszystkie pary średnich próbkowych. Jeżeli $|\overline{X}_i - \overline{X}_j| \geq NIR$, to uznajemy, że m_i i m_j są istotnie różne.
6. Dla wszystkich porównań par prawdopodobieństwo błędu I rodzaju jest ustalone na poziomie α

Komputerowe symulacje wykazały, że jeśli stosujemy test Fishera w połączeniu z ANOVA (tak jak opisano powyżej), to poziom istotności złożonego testu porównań wielokrotnych jest w przybliżeniu równy poziomowi istotności testu *F* (procedura Fisher's protected LSD).

Jeżeli stosujemy test *NIR* samodzielnie (bez ANOVA), to kontrolujemy jedynie błąd przy pojedynczych porównaniach (**per comparison**). Odpowiada to wielokrotnemu stosowaniu testu istotności różnicy dwóch średnich opartego na statystyce *t*-Studenta, przy czym estymator wariancji opieramy na całej próbie (z powodu jednorodności wariancji w grupach) a nie tylko na obserwacjach z aktualnie porównywanych grup.

Test W Tukey'a oparty jest na studentyzowanym rozstępie $\sqrt{n}\,\frac{\overline{X}_{\max}-\overline{X}_{\min}}{S}$ pomiędzy średnimi próbkowymi.

Algorytm (dla jednakowo licznych grup)

1. Uporządkować średnie próbkowe $\overline{X}_i$, i=1,...,k
2. m_i i m_j są istotnie różne jeżeli $|\overline{X}_i-\overline{X}_j|\geq W$, gdzie $W=q_\alpha(k,df)\sqrt{\frac{S_W^2}{n}}$, df jest liczbą stopni swobody w S_w^2, n ilość obserwacji w każdej grupie, $q_\alpha(k,df)$ prawostronna wartość krytyczna studentyzowanego rozstępu (tablice).
3. Prawdopodobieństwo zaobserwowania fałszywie istotnej różnicy dla porównań parami jest równe α.

Jeżeli grupy nie są równoliczne, to zamiast W należy użyć $W_{ij}=q_\alpha(k,df)\sqrt{\frac{S_W^2}{2}(\frac{1}{n_i}+\frac{1}{n_j})}$

Test Tukey'a kontroluje błąd I rodzaju dla wszystkich porównań parami, tzn. prawdopodobieństwo (przy H_0: m_1=…= m_k) zaobserwowania takiego układu średnich próbkowych $\overline{X}_i$, i=1,...,k, dla którego przynajmniej jedna różnica pomiędzy średnimi m_i jest fałszywie uznana za istotną jest równe α. Jeżeli jest k grup, to test Tukey'a kontroluje łączny błąd I rodzaju dla $\binom{k}{2}$ porównań jednocześnie (per experiment)

Test Newmana-Keulsa jest modyfikacją testu Tukey'a wykorzystującą informację o ilości miejsc pomiędzy badanymi średnimi w uporządkowanym ciągu średnich .

Algorytm (dla jednakowo licznych grup)

1. Uporządkować średnie próbkowe $\overline{X}_i$, i=1,...,k
2. Dla dwóch średnich $\overline{X}_i$ i $\overline{X}_j$ odległych o r miejsc odpowiadające im średnie m_i i m_j są istotnie różne jeżeli $|\overline{X}_i-\overline{X}_j|\geq W_r$, gdzie $W_r=q_\alpha(r,df)\sqrt{\frac{S_W^2}{n}}$, df jest liczbą stopni swobody w S_w^2, n ilość obserwacji w każdej grupie, $q_\alpha(r,df)$ prawostronna wartość krytyczna studentyzowanego rozstępu (tablice). **Uwaga**. Przyjmujemy, że sąsiednie średnie odległe są o 2 a skrajne o n czyli r_{ij}=|ranga($\overline{X}_i$) - ranga($\overline{X}_j$)| +1

Jeżeli grupy nie są równoliczne, to zamiast W_r należy użyć $W_{rij}=q_\alpha(r,df)\sqrt{\frac{S_W^2}{2}(\frac{1}{n_i}+\frac{1}{n_j})}$.

Test wielokrotnych rozstępów Duncana jest podobny do dwóch poprzednich, gdyż jest oparty na studentyzowanym rozstępie, lecz różni się od poprzednich tym, że zmienny jest poziom istotności przy poszczególnych porównaniach. Gdy średnie próbkowe zostaną uporządkowane i są odległe o r

miejsc, to istotność różnicy odpowiadających im średnich w populacjach jest testowana na poziomie $1-(1-\alpha)^{r-1}$

Algorytm (dla jednakowo licznych grup)

1. Uporządkować średnie próbkowe $\overline{X}_i$, i=1,...,k
2. Średnie m_i i m_j są istotnie różne jeżeli odpowiadające im średnie próbkowe odległe o r miejsc spełniają warunek $|\overline{X}_i - \overline{X}_j| \geq W_r'$, gdzie $W_r' = q'_\alpha(r, df)\sqrt{\frac{S_W^2}{n}}$, df jest liczbą stopni swobody w S_w^2, n ilość obserwacji w każdej grupie, $q'_\alpha(r, df)$ prawostronna wartość krytyczna testu Duncana (tablice).

Jeżeli grupy są w przybliżeniu równoliczne, to zamiast n należy użyć $\tilde{n} = \frac{k}{\frac{1}{n_1}+\frac{1}{n_2}+...+\frac{1}{n_k}}$.

Test Scheffe'go

1. Rozważmy dowolny kontrast $I_{\mathbf{a}} = \sum_{i=1}^{k} a_i m_i$, $\sum_{i=1}^{k} a_i = 0$. Chcemy zweryfikować hipotezę

   H_0: $\forall \mathbf{a} \quad I_{\mathbf{a}} = \sum_{i=1}^{k} a_i m_i = 0$ wobec alternatywy H_1: $\exists \mathbf{a} \; I_{\mathbf{a}} \neq 0$.

2. Rozważmy dowolny kontrast próbkowy $\hat{I}_{\mathbf{a}} = \sum_{i=1}^{k} a_i \overline{X}_i$ dla którego nieobciążonym estymatorem wariancji jest $\hat{V}(\hat{I}) = \frac{1}{n-k} S_w^2 \sum_{i=1}^{k} \frac{1}{n_i} a_i^2$

3. Kontrast $I_{\mathbf{a}}$ uznamy za istotny (istotnie różny od 0), gdy

   $$\left| \frac{(n-k)\hat{I}_{\mathbf{a}}^2}{(k-1)S_w^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}} \right| > F_{k-1,n-k,1-\alpha},$$

   gdzie, $F_{k-1,n-k,1-\alpha}$ jest kwantylem rzędu 1-α rozkładu centralnego Snedecora Fishera F.

4. Prawdopodobieństwo zaobserwowania fałszywie istotnego kontrastu jest równe α.

Test Scheffe'go kontroluje łączny błąd większej liczby porównań niż test Tukey'a, więc jest bardziej konserwatywny (trudniej odrzucić H_0). W teście Scheffego na poziomie α prawdopodobieństwo zaobserwowania fałszywie istotnego kontrastu nie przekracza α.

Porównania testów. W poniższej tabeli zestawiono wartości progowe po przekroczeniu których różnice uznajemy za istotne. Rozważono porównanie 6 grup, przy czym próbka dla każdej grupy liczy n=5 elementów a S_w^2=2451.

Test	Liczba miejsc pomiędzy średnimi r				
	2	3	4	5	6
Fisher *NIR*	64.63	64.63	64.63	64.63	64.63
Tukey	96.75	96.75	96.75	96.75	96.75
Newman-Keuls	64.65	78.16	86.35	92.33	96.75
Duncan	64.65	67.97	69.74	71.29	72.62
Scheffe	196.31	196.31	196.31	196.31	196.31

Widać, że najbardziej konserwatywny jest test Scheffe'go a najmniej konserwatywny (czyli najbardziej czuły) test *NIR*. Z uwagi na kontrolę błędu godny polecenia jest test Tukey'a. Symulacje preferują raczej test Newmana-Keulusa. Z uwagi na czułość duże uznanie wśród praktyków wzbudził test Duncana.

Na uwagą zasługuje jeszcze test **Dunetta** wielokrotnych porównań z **wyróżnioną grupą kontrolną**