Elementa om Variansanalys

för kursen sf1911, Statistik för bioteknik

Envägs variansanalys.

Kapitel tio beskrev metoder för att testa om $x_1,\ldots,x_k$ och $y_1,\ldots y_m$ kommer från fördelningar med samma väntevärde (d.v.s $\mu_1=\mu_2$, då man vet att fördelningarna är $N(\mu_1,\sigma)$ och $N(\mu_2,\sigma)$. (”Två oberoende sickprov”.)

Envägs variansanalys generaliserar detta till fler än två fördelningar. Säg att $x_1,\ldots,x_k$, $y_1,\ldots,y_m$ och $z_1,\ldots,z_n$ kommer från $N(\mu_1,\sigma)$, $N(\mu_2,\sigma)$ och $N(\mu_3,\sigma)$, respektive (observera: samma varians!). Vi skall testa nollhypotesen

H_0: $\mu_1=\mu_2=\mu_3$

Det går till så här: vi centrerar hela klabbet genom att subtrahera dess gemensamma medelvärde

$$c=\frac{1}{k+m+n}(x_1+\ldots+x_k+y_1+\ldots+y_m+z_1+\ldots+z_n)$$

och sedan bilda kvadratsumman

$$\text{sst}=(x_1-c)^2+\ldots+(z_n-c)^2.$$

Sedan beräknar vi en liknande kvadratsumma, men vi subtraherar *varje variabels* individuella medelvärde:

$$\text{ssr}=(x_1-\overline{x})^2+\ldots+(z_n-\overline{z})^2$$

Om nu sst är betydligt större än ssr så beror det förmodligen på att väntevärdena inte är lika, d.v.s. åtminstone en av likheterna i H_0 inte gäller. Det exakta testet går till så här:

Bilda testvariabeln

$$F = \frac{\text{sst} - \text{ssr}}{\text{ssr}} \; \frac{N - r}{r - 1}$$

där N är totala antalet observationer, d.v.s. $N = k + m + n$, och r är antalet väntevärden (fördelningar), d.v.s $r = 3$ i vårt fall.

Om H_0 nu är sann så är F utfallet av en $F(r - 1, N - r)$-fördelad stokastisk variabel. Vi beräknar p-vädeet för testet att H_0 är sann:

$$p\text{-värde} = \text{P}(F(r - 1, N - r) > F).$$

Resultatet brukar sammanfattas i en ANOVA-tabell likt den på sidan 412 i textboken. Det som betecknas k i boken är vårt r, SST i boken är vårt sst, SSE i boken är vårt ssr, SSTr i boken är vårt sst – ssr.

Låt oss säga att sst = 340, ssr = 112, $N = 24$ och $r = 4$ (se exempel 11.1 i textboken). ANOVA-tabellen skulle då se ut så här:

source	df	ss	mss	F	p	η^2
treatment	3	228	76	13.57	0.000047	0.6706
residual	20	112	5.6			
total	23	340				

I kolumnen ”df” är 3 = $r - 1$, 20 = $N - r$ och 23 = $N - 1$.

Jag har ändrat litet på namnen under “source”, och dessutom lagt till en kolumn η^2. Värdet 228 i kolumnen ”ss” är differensen sst – ssr (340 – 112), 76 är 228/3, 5.6 är 112/20, 13.57 är 76/5.6, p är p-värdet beräknat genom

Fcdf(13.57, E99, 3, 20)

på en TI-räknare. Värdet η^2 som jag lagt till räknas ut genom 228/340. η^2 har samma mening som R^2 i en regression. De enda värdena av intresse är egentligen p-värdet och η^2. p-värdet 0.000047 är så lågt att vi förkastar H_0.

På en TI-räknare får man ut tabellen i full ornat genom

ANOVA(L_1, L_2, L_3, L_4)

där listorna L_1, L_2, L_3, L_4 innehåller respektive datavärden. D.v.s. man får inte ut η^2 direkt, men det beräknas enkelt genom SS.Factor/(SS.Factor + SS.Error).

η^2 är ett mått på ”effect size”, och är något som starkt efterfrågas av vetenskapliga tidskrifter vid publicering numera. Många anser att *p*-värden är rätt oinformativa, och dessutom ofta (i själva verket nästan alltid) misstolkade, medan ”effecr size” är viktigare. Här en länk till en artikel om detta (Shinichi Nakagawa1, Innes C. Cuthill; “*Effect size, confidence interval and statistical significance: a practical guide for biologists,* Biol. Rev. (2007), 82.)

https://people.kth.se/~lang/Effect_size.pdf

Konfidensintervall

Konfidensintervall för skillnad i väntevärden räknas ut ungefär som i fallet ”två oberoende stickprov”:

$$\mu_x - \mu_y = \overline{x} - \overline{y} \pm t_*(N-r)\sqrt{\frac{\text{ssr}}{N-r}\left(\frac{1}{k}+\frac{1}{m}\right)}.$$

Frihetsgraderna är alltså fler än man skulle ha vid ”två oberoende stickprov” ($k + m - 2$), vilket beror på att vi använt även *z*-variablerna för att beräkna (den gemensamma) variansen. Man bör tänka på att använda Bonferroni om man gör flera konfidensintervall och vill kontrollera den totala felrisken.

Blockning: tvåvägs ANOVA

Även situationen ”observationer i par” kan generaliseras. Vi tar exemplet i boken på sidan 430, men ändrar litet i formuleringen. Antag att vi vill göra experiment på 24 råttor för att jämföra responsen med tre olika testmetoder: A, B och C. För att minska effekten av variationen mellan råttor väljer vi sex råttor ur fyra

kullar, d.v.s vi har fyra syskonskaror S1, S2, S3, S4, om vardera sex råttor. Sedan väljer vi *slumpvis* ut två ur varje kull som får behandlingen A, B och C, respektive. Vi presenterar data i en matris:

	$S1$	$S2$	$S3$	$S4$
A	9.4, 9.4	9.3, 9.2	9.8, 9.6	9.9, 9.8
B	9.2, 9.1	9.4, 9.4	9.5, 9.5	9.7, 9.7
C	9.7, 9.8	9.6, 9.7	10.0, 9.9	10.2, 10.3

Nu beräknar vi variationen i data då variationen beroende på testmetod och kull är borträknad, ssr (Sum of Squared Residuals)

$$\text{ssr} = \min_{a_i, b_j} \sum_{i,j} \sum_{k} (y_{i,j,k} - a_i - b_j)^2$$

Här betecknar t.ex. $y_{2,3,1}$ responsen hos råtta nummer 1 i kull S3 som testats med metod B, dvs första siffran i cellen på rad 2, kolumn 3, d.v.s. 9.5, o.s.v. Vi minimerar över koefficienterna a_1, a_2, a_3, b_1, b_2, b_3, b_4. Uppenbarligen blir inte dessa koefficienter entydigt bestämda; vi kan ju t.ex. addera 1 till alla a:n och subtrahera 1 från alla b:n utan att kvadratsumman påverkas. För att få entydighet kan vi t.ex. sätta någon koefficient till noll, t.ex $b_1 = 0$. Vi har alltså 6 fria koefficienter, vilket gör ”antalet frihetsfrader” i ssr till 24 – 6 = 18.

I vårt exempel blir ssr = 0.17750

Variationen ”förklarad” av skillnad mellan kullarna får vi som

$$\begin{aligned} \text{ss.kull} &= \min_{a_i} \sum_{i,j} \sum_{k} (y_{i,j,k} - a_i)^2 - \text{ssr} \\ &= 1.23875 - 0.17750 = 1.06125 \end{aligned}$$

och variationen ”förklarad” av skillnad mellan testerna som

$$\begin{aligned} \text{ss.test} &= \min_{b_j} \sum_{i,j} \sum_{k} (y_{i,j,k} - b_j)^2 - \text{ssr} \\ &= 1.10833 - 0.17750 = 0.93083 \end{aligned}$$

(ingen restriktion $b_1 = 0$ här!)

Frihetsgraderna df.kull = 6 – 3 = 3, df.test = 6 – 4 = 2: här är 6 antalet fria parametrar i ssr och 3 antalet fria parametrar i ss.kull, 4 antalet fria parametrar i ss.test. Frihetsgraderna blir helt enkelt antalet rader minus ett för test, antalet kolonner minus ett för kull. Vi är nu redo att göra vår ANOVA-tabell:

source	df	ss	mss	F	p	η^2
test	2	0.93083	0.4654	47.20	0.0000	0.8398
kull	3	1.06125	0.35375	35.87	0.0000	0.8567
resid	18	0.17750	0.00986			

Här beräknas de fyra kolumnerna längst till höger som förut: mss = ss/df, F = mss/mss.resid, p som tidigare (frihetsgrader (2,18), respektive (3, 18)) och $\eta^2 = 1 / (1+0.17750 / 0.93083)$, resp. $\eta^2 = 1 / (1+0.17750 / 1.06125)$.

Vad säger ANOVA-tabellen?

Först kontrollerar vi om det var meningsfullt att göra blockning. Skälet att göra det var att vi genom att förklara en del av variationen i responserna med kull-tillhörighet så skulle vi få ner variationen i residualen, ssr, den variation som finns kvar oförklarad. Det är bra om den är liten, för då får testet större styrka, och konfidens-intervallen (se nedan) blir kortare.

Å andra sidan är det bra med många observationer, av precis samma skäl. Det ”effektiva” antalet obsrvationer är frihetsgraderna för residualen, df.resid (alltså 18 i exemplet.) När vi inför kull-tillhörighet så minskar det (från 21 till 18) vilket är dåligt. Det som är avgörande är om kvoten ssr/df.resid minskar eller inte. Om den ökar, så har vi minskat styrkan i testet, om den minskar har vi ökat styrkan.

Det är lätt att se vad som är fallet: vi tittar på F-värdet för kull, F.kull. Om F.kull > 1 så var det lyckat att blocka, annars inte. I detta fall är F.kull = 35.87, så det var synnerligen lyckat!

Observera att *p*-värdet för kull är totalt ointeressant. Även om *p*-värdet vore, säg, 0.25, alltså inte ”signifikant” så skulle vi *ändå* ha med kull i modellen, eftersom F.kull då vore 1.49 > 1.

Skulle f.kull varit < 1, skulle vi ha bytt modell till en envägs ANOVA.

p-värdet för test är däremot intressant. Det är praktiskt taget noll, så *vi accepterar alternativet att testerna ger olika resultat*. Men tabellen ger ingen information om på vilket sätt. Vi skall därför beräkna konfidensintervall för skillnaderna mellan testmetoderna. Jag återkommer till det nedan, men först litet resonemang om vad vi åstadkommit hittills.

Författaren till textboken skriver på sidan 431, angående exemplet som liknar vårt:

“*To emphasize the benefits of blocking, we provide a nonsolution by treating this problem as a one-way ANOVA layout. This time, we fail to find any significant difference between the testing procedures (p-value 0.2196). Clearly, this approach is incorrect on other grounds: the condition of independence among the treatments, required for ANOVA, is violated.*”

Detta är totalt nonsense! När vi gör tvåvägs-ANOVA så utgår vi ifrån att alla mätobservationer är oberoende. Då kan de inte plötsligt vara beroende för att vi gör en annan analys. Beroende eller oberoende avgörs av hur data har generetats, inte på vad vi sedan gör med dessa data.

Vad han är ute efter är i stället detta: när vi gör envägs-ANOVA så antar vi att data för t.ex. $x_1,\ldots,x_k$ är observationer från samma fördelning. Men de åtta observationerna för test A, t.ex., kommer inte från samma fördelning, eftersom t.ex. de två första – 94, 94 – kommer från en fördelning och de två sista – 99, 99 – kommer från en annan fördelning, eftersom mössen tillhör olika kullar – de har

alltså inte samma väntevärden. Det är just för att eliminera variansen beroende på dessa skillnader i väntevärden som vi utför blockningen. På det sättet är förutsättningarna för envägs-ANOVA ”violated”. Det korrekta sättet att utföra en envägs-ANOVA vore att plocka ut 24 råttor på måfå, och sedan slumpvis välja ut åtta som får behandling A, o.s.v.

Antag nu att vi i vår tvåvägs-ANOVA hade fått F.kull < 1. Kan vi då använda data som de är för en envägs-ANOVA? Svaret är ”ja”. Om F.kull < 1 så är *p*-värdet för hypotesen ”samma fördelning mellan kullarna” $p = P(F(3,18) > F.kull)$ som är mer än 40%, så det bör vara ofarligt att förfara som om den hypotesen var sann.

Konfidensintervall

Om vi nu vill jämföra t.ex. test C med test B, så skall vi se detta som ”observationer i par”. Skillnaden i medelvärden för kull 1 mellan test C och test B är

$$\frac{1}{2}(9.7+9.8)-\frac{1}{2}(9.2+9.1)=0.6$$

och så vidare. Medelvärdet för dessa differenser för alla fyra kullar är 0.4624, alltså $\hat{\mu}_{C-B}=0.4625$. Men man ser ju lätt att detta även kan räknas ut som medelvärdet av C-testerna minus medelvärdet av B-testerna:

$$\hat{\mu}_{C-B}=\hat{\mu}_C-\hat{\mu}_B=\frac{1}{8}(9.7+9.8+\ldots+10.2+10.3)$$
$$-\frac{1}{8}(9.2+9.1+\ldots+9.7+9.7)=0.4625$$

Standardfelet för denna skattning beräknas så här:

$$SD(\hat{\mu}_{C-B})=\sqrt{\frac{\text{ssr}}{18}\cdot\left(\frac{1}{8}+\frac{1}{8}\right)}=0.04965$$

Här är 18 = frihetsgraderna för ssr, 8 och 8 är antalet mätdata för test C och B, respektive. Vi får nu konfidensintervall för μ_{C-B}:

$$\mu_{C-B} = \hat{\mu}_{C-B} \pm t_*(18) \cdot 0.04965$$

Notera att standardavvikelsen beräknas precis som för $\hat{\mu}_C - \hat{\mu}_B$ i en envägs-ANOVA (i det fallet är ju $N - r$ detsamma som frihetsgraderna för ssr). Här ser vi tydligt att vinsten med att göra blockning är om kvoten ssr / df.ssr är mindre med blockning. Vi får kortare konfidensintervall.

Men det innebär också att om vi gör envägs-ANOVA, så är vi bara onödigt konservativa (vi får onödigt långa konfidensintervall), men en inferens om skillnaderna är giltig, så det är inte direkt *fel* att göra envägs-ANOVA på dessa data, bara onödigt ineffektivt.

Vi beräknar tre konfidensintervall med konfidensgrad 0.9834, så att vi får en ”experimental confidence level” på åtminstone 0.9502 (Bonferroni):

$$\mu_{B-A} = (-0.2436,\ 0.0186)$$

$$\mu_{C-A} = (0.2189,\ 0.4811)$$

$$\mu_{C-B} = (0.3314,\ 0.5936)$$

Vi konstaterar att $\mu_{C-A} > 0$ och $\mu_{C-B} > 0$.