Envägs variansanalys (ANOVA) för test av olika väntevärde i flera grupper

Tobias Abenius <Tobias.Abenius@Chalmers.se>

February 21, 2012

Envägs variansanalys (ANOVA)

I envägs variansanalys utnyttjas att täljaren i s^2 kan separeras

$$N = \sum_{i=1}^{k} n_i \tag{1}$$

$$(N-1)s^2 = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\cdot\cdot})^2}_{S_{\text{Total}}} = \tag{2}$$

$$= \underbrace{\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2}_{S_B \quad \text{mellan grupper}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2}_{S_W \quad \text{inom grupper}}. \tag{3}$$

Från detta kan man sen testa hur förhållandet mellan spridningen i data mellan grupper och spridningen inom grupper ser ut.

Storheterna S_B och S_W har olika namn i olika litteratur. Det kortare S skrivs också SS och står för *sum of squares* vilket är lätt att komma ihåg. Bokstäverna B och W står för *between* respektive *within*. Det som skiljer mellan olika grupper kallas också för behandlingar eller *treatments* – SS_T, medan inom grupper kallas för oförklarade fel eller *error* – SS_E. Om man sedan delar med antalet frihetsgrader så kallas det ett medelkvadratsumma, *mean square, mean sum of squares* – MS eller motsvarande.

Antaganden

1. Respons är normalfördelad (eller approximativt normalfördelad)
2. Observationerna är oberoende
3. Populationernas varians är lika
4. Respons för en given grupp är oberoende och likafördelade normalfördelade stokastiska variabler

Hypoteser

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_K$$
$$H_A: \ \text{alla ej lika, dvs}: \ \exists i, j : \mu_i \neq \mu_j, i \neq j$$

För K st grupper.

Test

Om man kan förkasta hypotes H_0 eller ej på nivå α kan testas med ett *F-test*. Ett F-test testar om sannolikheten att ett värde inte tillhör en F-fördelning är mindre än α eller ej. Om vi bildar kvoten mellan två χ^2-fördelade storheter är den F-fördelad. Vi bildar kvoten

$$F = \frac{\text{Spridning mellan grupper}}{\text{Spridning inom grupper}}$$

Spridning mellan grupper är då

$$MS_B = \sum_i \frac{n_i(\bar{y}_{i\cdot} - \bar{y})^2}{df_b},$$
$$df_b = K - 1$$

där n_i är antalet observationer i den i:te gruppen och punkten i $\bar{y}_{i\cdot}$ betyder att man tar genomsnittet över den i:te gruppen, och man summerar över alla de K st grupperna. Frihetsgraderna är här antalet grupper $K - 1$ (för det skattade medelvärdet för alla observationer i alla grupper).

Detta är alltså avståndet mellan varje gruppmedelvärde och medelvärdet för alla K gruppers alla observationer.

Spridning inom grupper är

$$MS_W = \sum_{i,j} \frac{(y_{ij} - \bar{y}_{i\cdot})^2}{df_w},$$
$$df_w = N - K,$$

dvs avståndet mellan varje enstaka observation och denna observations gruppmedelvärde summerat över alla observationer i alla grupper. Frihetsgraderna är här antalet observationer minus ett skattat medelvärde för varje grupp.

N är här det totala antalet observationer och y_{ij} är den j:te observationen i den i:te gruppen.

Om kvoten F når ut i svansen på F-fördelningen kan man förkasta H_0 med sannolikheten

$$p = 1 - F_{df_b, df_w}(F),$$

dvs 1− fördelningsfunktionen för F-fördelningen med df_b respektive df_w frihetsgrader, alltså svansens yta. Se exempel nedan.

Exempel

Vi har observerat en storhet som tros vara normalfördelad med samma varians i tre olika grupper, A, B och C. Värdena på storheten i fråga för de olika grupperna mättes till

$A : 2.3,\ 2.2,\ 2.1,\ 2.3,\ 2.0,\ 2.5,\ 2.3,\ 2.4,\ 2.2,\ 2.4$
$B : 2.1,\ 2.3,\ 2.0,\ 2.2,\ 2.0,\ 2.1,\ 2,\ 2.2,\ 2.3,\ 2.2,\ 2.2,\ 2.1,\ 2.2,\ 2.2,\ 2.3$
$C : 2.2,\ 2.4,\ 2.2,\ 2.3,\ 2.2,\ 2.3,\ 2.2$

Vi numrerar grupperna så att A är 1, B är 2 och C är grupp 3. Antalet observationer och medelvärden för grupperna är då

$n_1 = 10 \quad \bar{y}_{1\cdot} = 2.27$
$n_2 = 15 \quad \bar{y}_{2\cdot} = 2.16$
$n_3 = 7 \quad \bar{y}_{3\cdot} = 2.257143$

Medelvärdet över alla observationer kan man då få genom att se att

$$\begin{aligned}\bar{y}_{\cdot\cdot} = \frac{\sum_{i,j} y_{ij}}{N} &= \frac{\sum_{i,j} y_{ij}}{n_1 + n_2 + n_3} = \\ &= \frac{\sum_i n_i \bar{y}_{i\cdot}}{n_1 + n_2 + n_3} = \\ &= \frac{10 \cdot 2.27 + 15 \cdot 2.16 + 7 \cdot 2.257143}{10 + 15 + 7} = \\ &= 2.215625\end{aligned}$$

Vi räknar också ut

$$s^2 = 0.01555444 = \frac{SS_{\text{Total}}}{N - 1}$$

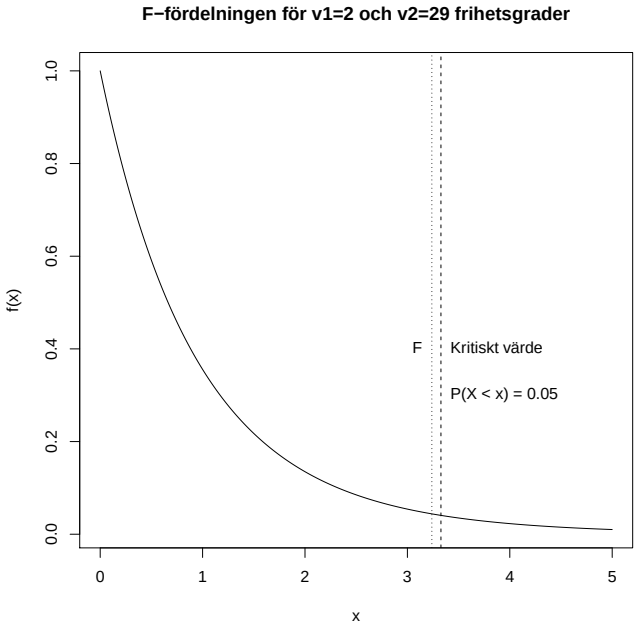

Figure 1: F-fördelningen för detta exempel

Kvadratsumman S_B får vi genom att

$$\begin{aligned} S_B &= \sum_i n_i(\bar{y}_{i\cdot} - \bar{y})^2 = \\ &= 10 \cdot (2.27 - 2.215625)^2 + 15 \cdot (2.16 - 2.215625)^2 + \\ &\quad + 7 \cdot (2.257143 - 2.215625)^2 = \\ &= 0.08804473 \end{aligned}$$

Vi delar med antalet frihetsgrader $df_b = K - 1$ och får medelkvadratsumman

$$MS_B = S_B/(K-1) = 0.08804473/(3-1) = 0.04402232$$

som mäter variationen mellan grupperna. Nu ska vi hitta variationen inom grupperna enligt, vilket är omständigt att göra för hand,

$$\begin{aligned} S_W &= \sum_{i,j} (y_{ij} - \bar{y}_{i\cdot})^2 = \\ &= \underbrace{0.201}_{A} + \underbrace{0.156}_{B} + \underbrace{0.03714286}_{C} = 0.3941429. \end{aligned}$$

Istället utnyttjar vi att

$$
\begin{aligned}
(N-1)s^2 &= S_{\text{Total}} = S_W + S_B \\
S_W &= S_{\text{Total}} - S_B = \underbrace{(10+15+7-1)}_{N-1}\underbrace{0.01555444}_{s^2} - \underbrace{0.08804473}_{SS_B} = \\
&= 0.3941429
\end{aligned}
$$

Vi delar med antalet frihetsgrader $df_w = N - K$ och får medelkvadratsumman

$$MS_W = S_W/(N-K) = 0.3941429/(32-3) = 0.01359113$$

som mäter variationen inom grupperna. Vi bildar kvoten F som är vår teststatistika

$$F = \frac{MS_B}{MS_W} = 3.239047.$$

Detta kan vi jämföra med det kritiska värdet,

$$F < F_{\text{crit}} = 3.327654,$$

dvs detta är ej ett signifikant resultat. Vi kan ej förkasta H_0 om lika väntevärde i alla grupper. Om det finns en skillnad mellan grupperna kan vi inte se den. Kanske måste vi ha fler mätningar för att se den eller så finns den inte alls.