Högskolan i Halmstad

Hypotestest (forts.)

χ^2-test

Hittills har vi betraktat test som förutsätter att vi kan förutsätta att variablerna har en viss fördelning (eller att centrala gränsvärdessatsen kan tillämpas). Ibland vill man dock kunna bevisa att en variabel inte har en given fördelning F_0 överhudtaget, dvs det är inte bara en parameter θ i fördelningen $F_0(\theta)$ som man vill visa är fel utan hela fördelningen F_0.

Antag att X har fördelningen F (som är okänd) och att vi vill testa hypotesen

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$$

och att vi observererar stickprovet $x_1, \ldots, x_n$ på X.

Vi gör då en **r-tabell** (vilket påminner om en frekvenstabell) där vi delat in värdemängden i k *klasser* (dvs delintervall: $(-\infty, a_1], (a_1, a_2], \ldots, (a_{k-2}, a_{k-1}], (a_{k-1}, \infty)$) och räknar antalet observationer O_i inom respektive klass i och *dessutom* beräknar det förväntade antalet observationer E_i inom respektive klass i under nollhypotesens antagande att $F = F_0$.

Klass	*Gränser*	*Antal obs.*	*Förv. antal*
1	$x_i \leq a_1$	O_1	E_1
2	$a_1 < x_i \leq a_2$	O_2	E_2
3	$a_2 < x_i \leq a_3$	O_3	E_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$ $\quad \vdots$
$k-1$	$a_{k-2} < x_i \leq a_{k-1}$	O_{k-1}	E_{k-1}
k	$a_{k-1} < x_i$	O_k	E_k

Obs Här görs ett avsteg ifrån den annars så konsekventa linjen att stora bokstäver betyder stokastiska variabler och små bokstäver betyder tal – O_i och E_i är ju ej stokastiska utan observerade tal.

Huvudregeln för beräkning av det förväntade antalet inom klass i, E_i är

$$E_i \;=\; n\,P(a_{i-1} < X \leq a_i) \;=\; n\,p_i$$

där n är antalet observationer i stickprovet.

Teststatistikan för att testa hypotesen ovan är

$$u \;=\; \sum_{j=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

och $U \in \chi^2(k-1)$ under H_0.

Obs Summan i teststatistikan u innehåller k stycken termer, inte n stycken!

I termer av ett arbetsschema likt de tidigare presenterade testen har vi

1. Stickprovet
2. Hypotes $\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$
3. Signifikansnivå α
4. r-tabell med klassindelning och beräkning av E_i
5. Teststatistika $U = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \in \chi^2(k-1)$ under H_0
6. Kritiskt område $C_\alpha = \{u : u > \chi^2_\alpha(k-1)\}$
7. Förkasta om $u \in C_\alpha$, inte annars!

Läs Exempel 10.7 i Vännman!

Exempel Ett fysikexperiment går ut på att belysa en skärm med ljus som passerat genom ett *gitter* (finkalibrigt galler). På skärmen kan man ana ett *interferensmönster* (ljusränder). Genom att mäta ljusstyrkan längs ett linjestycke vinkelrätt mot interferensbanden fås följande resultat.

Mätpunkt	1	2	3	4	5	6	7
Ljusstyrka	2	6	2	10	3	7	1

Kan man på 2% signifikansnivå visa att ljusspridingen ej kan vara normalfördelad?

Lösning: Betrakta ljusstyrkan som ett *antal* ljuskvanta. Bästa gissning på en normalfördelning använder vi $\bar{x}$ istället för μ och s^2 istället för σ^2. Klassindelingen är redan gjord och vi beräknar istället n, $\bar{x}$ och s^2:
$n = 2 + 6 + 2 + 10 + 3 + 7 + 1 = 31$
$\bar{x} = \frac{1}{31}(2 \cdot 1 + 6 \cdot 2 + 2 \cdot 3 + 10 \cdot 4 + 3 \cdot 5 + 7 \cdot 6 + 1 \cdot 7) = 4$
$s^2 = \frac{1}{31-1}(2 \cdot 1^2 + 6 \cdot 2^2 + 2 \cdot 3^3 + 10 \cdot 4^2 + 3 \cdot 5^2 + 7 \cdot 6^2 + 1 \cdot 7^2 - 31 \cdot 4^2) = 2.8$
Därmed är hypotesen vi vill testa $\begin{cases} H_0 : F = N(4, \sqrt{2.8}) \\ H_1 : F \neq N(4, \sqrt{2.8}) \end{cases}$.
För att beräkna de förväntade antalen inom varje klass får vi (med halvkorrektion!) under H_0 att
$p_1 = P(X \leq 1) = \Phi(\frac{1+0.5-4}{\sqrt{2.8}}) = 1 - \Phi(1.49) = 0.0681$
$p_2 = P(1 < X \leq 2) = \Phi(\frac{2+0.5-4}{\sqrt{2.8}}) - p_1 = 1 - \Phi(0.90) - p_1 = 0.116$
$p_3 = P(2 < X \leq 3) = \Phi(\frac{3+0.5-4}{\sqrt{2.8}}) - p_1 - p_2 = 1 - \Phi(0.30) - p_1 - p_2 = 0.198$
$p_4 = P(3 < X \leq 4) = \Phi(\frac{4+0.5-4}{\sqrt{2.8}}) - p_1 - p_2 - p_3 = \Phi(0.30) - p_1 - p_2 - p_3 = 0.2358$
$p_5 = P(4 < X \leq 5) = \Phi(\frac{5+0.5-4}{\sqrt{2.8}}) - p_1 - p_2 - -p_3 - p_4 = \{\text{symmetri}\} = p_3 = 0.198$
$p_6 = \{\text{symmetri}\} = p_2 = 0.116$
$p_7 = \{\text{symmetri}\} = p_1 = 0.0681$
Därmed är under de förväntade antalen E_i under H_0
$E_1 = 31p_1 = 2.11 = E_7$, $E_2 = 31p_2 = 3.6 = E_6$, $E_3 = 31p_3 = 6.14 = E_5$ och $E_4 = 31p_4 = 7.31$. Alltså får vi följande r-tabell

Klass	*Obs. antal*	*Förv. antal*
1	2	2.11
2	6	3.6
3	2	6.14
4	10	7.31
5	3	6.14
6	7	3.6
7	1	2.11

(En tumregel är att $E_i \geq 2$. Om $E_i < 2$ får man "slå samman klasser". Detta innebär att t.ex. klasserna 6 och 7 hade bildat en klass, 6–7, med det observerade antalet $O_{6-7} = 7 + 1 = 8$ och det förväntade antalet $E_{6-7} = 3.6 + 2.11 = 5.71$. I detta exempel har vi inte det problemet men mer om det i nästa exempel.)

Vi får att teststatistikans värde blir $\sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i} =$
$= \frac{(2-2.11)^2}{2.11} + \frac{(6-3.6)^2}{3.6} + \frac{(2-6.14)^2}{6.14} + \frac{(10-7.31)^2}{7.31} + \frac{(3-6.14)^2}{6.14} + \frac{(7-3.6)^2}{3.6} + \frac{(1-2.11)^2}{2.11} =$
$= 0.0057 + 1.6 + 2.7915 + 0.9899 + 1.6058 + 3.2111 + 0.5839 = 10.7879.$

Vidare är det percentilvärde som utgör gräns för det kritiska området i detta fall $\chi^2_{0.02}(7-1) = 15.0332$. Eftersom $u = 10.7879 \not> 15.0332 = \chi^2_{0.02}(6)$ kan H_0 ej förkastas, dvs man kan ej visa att ljusstyrkan hos interferensbanden ej är normalfördelad på 2% signifikansnivå. □

Exempel Efter en miljöförorening behandlades 234 drabbade individer för skador relaterade till incidenten. Av dessa fick 27% A: bestående skador, 62% B: betydande men övergående besvär, 10% C: lindriga och övergående besvär och 1% D: inga problem alls. Enligt tidigare studier vet man att, utsatt för höga doser av den miljöfarliga substansen, är $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{2}$, $P(C) = \frac{1}{9}$ och $P(D) = \frac{1}{18}$. Finns det anledning att tro att hälsan hos de drabbade 234 individerna påverkats av något annat än miljökatastrofen? Använd signifikansnivå 1%.

Lösning: Låt X vara graden av skada med värdemängd $= \Omega = \{A, B, C, D\}$. Observera att sammanfattningen av behandlingsresultaten är angivet i %, dvs i *andelar* och inte i *antal*. Vid χ^2-test hanterar man antal och därför måste dessa siffror omtolkas: $\#A : 234 \cdot 0.27 = 63.18$, $\#B : 234 \cdot 0.62 = 145.08$, $\#C : 234 \cdot 0.1 = 23.4$, $\#D : 234 \cdot 0.01 = 2.34$. P.g.a. de tidigare undersökningarna vet man vidare att $p_A = \frac{1}{3}, p_B = \frac{1}{2}, p_C = \frac{1}{9}, p_D = \frac{1}{18}$ varmed de förväntade antalen inom respektive klass blir $E_A = 234 \cdot \frac{1}{3} = 78$, $E_B = 234 \cdot \frac{1}{2} = 117$, $E_C = 234 \cdot \frac{1}{9} = 26$, $E_D = 234 \cdot \frac{1}{18} = 13$. Därmed har vi r-tabellen

Klass	*Obs. antal*	*Förv. antal*
A	63.18	78
B	145.08	117
C	23.40	26
D	2.34	13

Teststatistikans värde blir då

$$\begin{aligned} u &= \sum_{i=A,B,C,D} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(63.18-78)^2}{78} + \frac{(145.08-117)^2}{117} + \frac{(23.4-26)^2}{26} + \frac{(2.34-13)^2}{13} \\ &= 2.8158 + 6.7392 + 0.26 + 8.7412 \\ &= 18.5562 \end{aligned}$$

Eftersom $\chi^2_{0.01}(4-1) = 11.3449$ innebär detta att H_0 förkastas på 5% signifikansnivå.

EMELLERTID ser vi att vid beräkningen av teststatistikans värde utgör den sista termen en stor del. Detta är en effekt av att antalet observationer kan variera kraftigare *i förhållande* till det lilla förväntade antalet. Det är just därför man har tumregeln $E_i \geq 2$. För att undanröja eventuella misstankar om att man "fuskat" sig till ett signifikant resultat kan man bilda en ny klass, C–D, av de två sista klasserna C och D med observerat antal $O_{C-D} = 23.4 + 2.34 = 25.74$ och förväntat antal $E_{C-D} = 26 + 13 = 39$. Med obetydliga modifieringar av räkningarna ovan får vi

$$\begin{aligned} u &= \sum_{i=A,B,C-D} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(63.18-78)^2}{78} + \frac{(145.08-117)^2}{117} + \frac{(25.74-39)^2}{39} \\ &= 2.8158 + 6.7392 + 4.5084 \\ &= 14.0634 \end{aligned}$$

Nu har vi reducerat antalet klasser och får då även ett reducerat antal frihetsgrader till det värde från χ^2-fördelningen vi ska jämföra med. Eftersom $\chi^2_{0.01}(3-1) = 9.2103$ innebär det att vi även nu kan förkasta nollhypotesen, dvs ja, det finns anledning att tro att individernas hälsa påverkats av något annat än miljökatastrofen. □