

χ^2 -metoden.

χ^2 -metoden kan användas för prövning av hypoteser i flera olika slag av problem:

- om en stokastisk variabel följer en viss sannolikhetsfördelning med kända eller okända parametrar.
- om två stokastiska variabler är oberoende av varandra.

1. Test av anpassning.

Data: n observationer klassificerade i K olika kategorier:

Kategori	Antal obs.
1	O_1
2	O_2
$\vdots$	$\vdots$
K	O_K
Totalt	n

O_i = antal observationer i kategori i ($i = 1, 2, \dots, K$)

Låt P_i = sannolikheten att få en observation i kategori i ($\sum_{i=1}^K P_i = 1$).

Kategori	Antal obs.	Sannolikhet	Förväntat antal obs.
1	O_1	P_1	nP_1
2	O_2	P_2	nP_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	O_K	P_K	nP_K
Totalt	n	1	n

Sannolikheterna $P_1, P_2, \dots, P_K$ är nu okända, och vi vill pröva nollhypotesen att de har vissa bestämda numeriska värden:

$$H_0: P_1=P_{10}, P_2=P_{20}, \dots, P_K=P_{K0}$$

Nollhypotesen är att H_0 inte gäller, dvs. att *inte alla* sannolikheterna har de i H_0 givna värdena.

Vi använder ett χ^2 -test med testvariabeln: $\chi^2 = \sum_{i=1}^K \frac{(O_i - E_i)^2}{E_i}$,

där O_i = observerat antal i kategori i
 $E_i = n \cdot P_{i0}$ = förväntat antal under H_0

När H_0 är sann, så är testvariabeln approximativt χ^2 -fördelad med $K-1$ fg, ifall antalet observationer är tillräckligt stort.

Det förväntade antalet, $E_i = nP_{i0}$, måste vara ≥ 5 för alla kategorier. Om så inte är fallet, brukar man ibland slå ihop kategorier.

Stora skillnader mellan observerat och förväntat antal ger höga värden på testvariabeln. Vi förkastar därför H_0 , om (och endast om) vi får extremt *höga* värden på testvariabeln. Kritiskt värde bestäms av önskad signifikansnivå och av antal frihetsgrader.

Exempel 1: En produkt tillverkas i fyra färger. Man vill undersöka om vissa av färgerna föredras framför de övriga. Ett slumpmässigt urval av 80 presumtiva konsumenter tillfrågas om vilken färg de föredrar, och följande data erhålls:

Färg	1	2	3	4	Totalt
Antal	12	40	8	20	80

Nollhypotesen är att alla färgerna är lika populära, dvs. att vi har lika stor slh att i urvalet få en person som föredrar färg 1 som att få en person som föredrar färg 2 etc. Dvs.

Hypoteser: $H_0: P_1 = P_2 = P_3 = P_4 = 0,25$
 H_1 : Ej alla sannolikheter lika med 0,25

Sign.-nivå: 5%

Testvariabel: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ ($K-1 = 3$ fg)

Beslutsregel: H_0 förkastas om $\chi_{\text{obs}}^2 > \chi_{0,05}^2(3) = 7,81$.

Resultat:

Färg	Obs. antal	Sannolikhet enl. H_0	Förv. antal enl. H_0
1	12	0,25	$80 \cdot 0,25 = 20$
2	40	0,25	$80 \cdot 0,25 = 20$
3	8	0,25	$80 \cdot 0,25 = 20$
4	20	0,25	$80 \cdot 0,25 = 20$
Totalt	80	1	80

$$\chi_{\text{obs}}^2 = \frac{(12-20)^2}{20} + \frac{(40-20)^2}{20} + \frac{(8-20)^2}{20} + \frac{(20-20)^2}{20} = 30,400 > 7,81$$

H_0 förkastas på 5% signifikansnivå (Signifikant skillnad i popularitet mellan olika färger).

Exempel 2: A, B och C är tre konkurrerande produkter av samma typ. Under en längre tid har 30% av kunderna efterfrågat produkt A, 50% produkt B, och 20% produkt C. Nyligen har produkt C genomgått en förändring, och man vill veta om produkternas marknadsandelar därigenom har förändrats.

Vid en marknadsundersökning av 200 kunder fann man att 48 nu sade sig föredra produkt A, 98 produkt B, och 54 produkt C.

Vilken slutsats kan vi dra?

Hypoteser: $H_0: P_A = 0,3; P_B = 0,5; P_C = 0,2$
 H_1 : Det gäller inte att $P_A = 0,3; P_B = 0,5; P_C = 0,2$

Sign.-nivå: 5%

Testvariabel: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ ($K-1 = 2$ fg)

Beslutsregel: H_0 förkastas om $\chi_{\text{obs}}^2 > \chi_{0,05}^2(2) = 5,99$.

Resultat:

Prod.	Obs. antal	Sannolikhet enl. H_0	Förv. antal enl. H_0
A	48	0,3	$200 \cdot 0,3 = 60$
B	98	0,5	$200 \cdot 0,5 = 100$
C	54	0,2	$200 \cdot 0,2 = 40$
Tot.	200	1,0	200

$$\chi_{\text{obs}}^2 = \frac{(48-60)^2}{60} + \frac{(98-100)^2}{100} + \frac{(54-40)^2}{40} = 2,40 + 0,04 + 4,90 = 7,34 > 5,99$$

H_0 förkastas på 5% signifikansnivå, dvs vi tror att förändringen av produkt C medfört en ändring av marknadsandelarna.

OBS 1. Avrunda inte de förväntade antalen. Om de förväntade antalen inte blir heltal, så ta med någon eller några decimaler i beräkningarna.

OBS 2. O_i och E_i står för observerat och förväntat **antal**, inte procenttal.

2. Test av anpassning: fördelning med skattade parametrar

Säg att vi vill testa om våra data kan ses som ett stickprov från en viss typ av fördelning, t.ex. en Poissonfördelning, utan att vi i förväg har specificerat något bestämt parametervärde.

Vi måste beräkna ett **skattat parametervärde** utifrån våra data. Med hjälp av detta får vi sedan det förväntade antalet observationer (under H_0) i olika kategorier.

Antalet frihetsgrader är nu **$K-m-1$** , där

K = antalet kategorier

m = antalet skattade parametrar

Exempel 3: Vi har 262 textblock (ungefär lika långa) från ett visst dokument. För varje textblock räknar vi hur många gånger ordet "may" förekommer.

Data:

Antal förekomster	0	1	2	3 eller fler
Obs. antal	156	63	29	14

Vi vill testa om antalet förekomster av ordet "may" i ett textblock följer en Poissonfördelning. Poissonfördelningens sannolikhetsfunktion är

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad (x = 0, 1, 2, \dots)$$

och dess medelvärde är: $\mu = \lambda$. Detta medelvärde skattas nu med stickprovets medelvärde, och vi får

$$\hat{\lambda} = \bar{x} = \frac{x_1 + x_2 + \dots + x_{262}}{262} = 0,62.$$

Vi sätter sedan in detta skattade parametervärde i sannolikhetsfunktionen ovan.

Hypoteser: H_0 : Stickprovet kommer från en Poissonfördelning.

H_1 : Stickprovet kommer inte från en Poissonfördelning.

Sign.-nivå: 5%

Testvariabel: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

Frihetsgrader: $K-m-1 = 4-1-1 = 2$

Beslutsregel: H_0 förkastas om vi får $\chi_{obs}^2 > \chi_{0,05}^2(2) = 5,99$.

Resultat:

Antal före- komster	Obs. antal	$P(x \lambda=0,62)$	Förv. antal
0	156	0,5379	140,9
1	63	0,3335	87,4
2	29	0,1034	27,1
3 el. fler	14	0,0252	6,6
Totalt	262	1,0000	262,0

$$\chi_{obs}^2 = \frac{(156 - 140,9)^2}{140,9} + \frac{(63 - 87,4)^2}{87,4} + \frac{(29 - 27,1)^2}{27,1} + \frac{(14 - 6,6)^2}{6,6} = 16,86 > 5,99.$$

Nollhypotesen, att stickprovet kommer från en Poissonfördelning, förkastas på 5% signifikansnivå.

3. Test av oberoende i korstabell

Problemet i fortsättningen är följande:

- Vi har ett slumpmässigt *stickprov* av element, för vilka vi observerat två kvalitativa variabler.
- Stickprovsdata = korstabell, med *antal* observationer i varje cell.

Vi frågar: Råder det oberoende mellan de två kvalitativa variablerna i *populationen*?

Exempel 4: Disponibel inkomst och bostadskostnad för 500 enpersonshushåll.

Bostadskostnad	Inkomst			Totalt
	Låg	Mellan	Hög	
Låg	35	50	15	100
Mellan	50	120	30	200
Hög	15	80	105	200
Totalt	100	250	150	500

Finns det ett samband mellan inkomst och bostadskostnad?

Hur skulle det ha sett ut om det helt saknades samband mellan inkomst och bostadskostnad?

Dvs. om det rådde fullständigt oberoende mellan inkomst och bostadskostnad?

Vid fullständigt oberoende skulle vi ha haft följande förväntade antal enpersonshushåll:

Bostadskostnad	Inkomst			Totalt
	Låg	Mellan	Hög	
Låg	20	50	30	100
Mellan	40	100	60	200
Hög	40	100	60	200
Totalt	100	250	150	500

Information om inkomst skulle vid oberoende inte vara till hjälp, ifall vi ville försöka gissa hushållets bostadskostnad.

Stickprovdata: Korstabell med r rader och c kolumner:

	1	2	...	c	Totalt
1	O_{11}	O_{12}	...	O_{1c}	R_1
2	O_{21}	O_{22}	...	O_{2c}	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Totalt	C_1	C_2	...	C_c	n

där

$$R_i = \sum_{j=1}^c O_{ij}, \quad i=1, \dots, r;$$

$$C_j = \sum_{i=1}^r O_{ij}, \quad j=1, \dots, c.$$

H_0 : Oberoende mellan de båda indelningsgrunderna.

H_1 : Ej oberoende

Testas med testvariabeln: $\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

där

$$E_{ij} = \text{Förväntat antal under } H_0 = n \cdot \frac{R_i}{n} \cdot \frac{C_j}{n} = \frac{R_i \cdot C_j}{n}$$

	1	2	...	c	Totalt
1	E_{11}	E_{12}	...	E_{1c}	R_1
2	E_{21}	E_{22}	...	E_{2c}	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	E_{r1}	E_{r2}	...	E_{rc}	R_r
Totalt	C_1	C_2	...	C_c	n

E_{ij} -värden beräknade på detta sätt har egenskapen att:

- den relativa fördelningen i varje kolumn blir densamma som marginalfördelningen till höger,
- den relativa fördelningen i varje rad blir densamma som marginalfördelningen längst ner.

Det är ju vad vi väntar oss vid oberoende.

När H_0 är sann, så är testvariabeln approximativt χ^2 -fördelad med $(r-1)(c-1)$ fg, ifall stickprovet är tillräckligt stort.

Tumregel: E_{ij} måste vara minst lika med 5 i mer än 20% av cellerna. Ibland måste man slå ihop kategorier för att uppnå detta.

H_0 förkastas när vi får *höga* värden på testvariabeln.

Kritiskt värde bestäms på vanligt sätt från tabell med ledning av önskad signifikansnivå och antal frihetsgrader.

Ex.(forts.):

H_0 : Oberoende mellan disponibel inkomst och bostadskostnad.

H_1 : Ej oberoende.

Signifikansnivå: 5%

Testvariabel: $\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Frihetsgrader: $(3-1)(3-1) = 4$ fg

Beslutsregel: H_0 förkastas om $\chi_{\text{obs}}^2 > \chi_{0,05}^2(4) = 9,49$.

Resultat: Förväntat antal, under H_0 , står inom parentes under observerade antalet i varje cell.

Bostadskostnad	Inkomst			Totalt
	Låg	Mellan	Hög	
Låg	35 (20)	50 (50)	15 (30)	100
Mellan	50 (40)	120 (100)	30 (60)	200
Hög	15 (40)	80 (100)	105 (60)	200
Totalt	100	250	150	500

$$\chi_{\text{obs}}^2 = \frac{(35-20)^2}{20} + \dots + \frac{(105-60)^2}{60} = 93,63 > 9,49$$

H_0 förkastas på 5% sign.-nivå, dvs. att vi tror att det finns ett samband mellan disponibel inkomst och bostadskostnad.