

SF1922/SF1923: SANNOLIKHETSTEORI OCH  
STATISTIK  
FÖRELÄSNING 14-15  
PASSNING AV FÖRDELNING:  $\chi^2$ -METODER.

Tatjana Pavlenko

14 maj 2018



## PLAN FÖR DAGENS FÖRELÄSNING

- ▶ Icke-parametriska metoder. (Kap. 13.10)
- ▶ Det grundläggande  $\chi^2$ -metod.
- ▶ Test av given fördelning.
- ▶ Test av fördelning med skattade parametrar.
- ▶ Homogenitetstest.
- ▶ Oberoendetest.



- ▶ Idé: I **föregående avsnitt** studerade vi parametrisk statistisk inferens, dvs inferens om parametrar i för övrigt helt kända fördelningar, t ex  $N(\mu, \sigma)$ , eller  $Po(\lambda)$ . Nu ska vi gå igenom ett antal metoder som är av mer allmän natur, vi ska beskriva testsituationer där man är intresserad att specificera vilken fördelningen är.
- ▶ Här är hypoteserna inte bara utsagor om en parameter utan om en hel fördelning: t ex man kan vara intresserad att testa

$$H_0 : X \in Po(\lambda) \text{ för något } \lambda,$$

mot  $H_1 : X$  är inte Poissonfördelad.

- ▶ Vi kommer att behandla ett antal olika fall där så-kallad  $\chi^2$ -metod används. Vi inleder med grundläggande idé inom  $\chi^2$ -metod.



## TEST AV GIVEN FÖRDELNING.

- ▶ Antag att  $n$  st. oberoende försök utförs, av vilka vart och en kan utfalla på  $r$  olika sätt  $A_1, A_2, \dots, A_r$  med sannolikheterna respektive  $P(A_1), P(A_2), \dots, P(A_r)$ .
- ▶ Låt vidare  $x_1, x_2, \dots, x_r$  vara de absoluta frekvenserna för utfallen  $A_1, A_2, \dots, A_r$  som erhålls i en viss sådan serie,  $\sum_{i=1}^r x_i = n$ .
- ▶ Vi vill pröva

$$H_0 : P(A_1) = p_1, P(A_2) = p_2, \dots, P(A_r) = p_r,$$

där  $p_i$  är givna (kända) sannolikheter med summa  $\sum_{i=1}^r p_i = 1$ , mot den alternativa hypotesen

$$H_1 : P(A_i) \neq p_i \quad \text{för minst ett } i \text{ bland } 1, \dots, r.$$

- ▶ Observera att  $p_i$  är sannolikhetsfunktion, dvs  $H_0$  ger *hypotetisk fördelning*. Detta är *test av given fördelning*. Ibland kallas också *ett test av anpassning* (engelsk: *goodness-of-fit test*).



## DET GRUNDLÄGGANDE $\chi^2$ -METOD.

För att testa  $H_0$  med  $\chi^2$  metoden använder man följande:

1. Som testvariabel tar man

$$Q_{obs.} = \sum_{j=1}^r \frac{(x_j - np_j)^2}{np_j}.$$

Observera att  $np_j$  är förväntat antal utfall i  $A_j$  under  $H_0$ . Intuitivt: Låga  $Q_{obs.}$ -värden tyder på god överensstämmelse mellan  $x_j$  och det förväntade  $np_j$  och då finns det ingen anledning att betvivla nollhypotesen. Avvikelse från  $H_0$  visar sig genom stora  $Q_{obs.}$ -värden.

2. Om  $H_0$  är sann (dvs de sannolikheterna  $p_j$  är rätta) så kan man visa att  $Q_{obs.}$  är utfall av en s.v.  $Q$  som är *approximativt*  $\chi^2(r-1)$ -fördelad.



### 3. Belutsregel:

Förkasta  $H_0$  om  $Q_{obs.} > \chi_{\alpha}^2(r-1)$

Förkasta ej  $H_0$  om  $Q_{obs.} \leq \chi_{\alpha}^2(r-1)$

Om  $n$  är stort får detta test approximativt signifikansnivån  $\alpha$ .

Villkor:

- ▶ Approximationen med  $\chi^2$ -fördelning fungerar tillfredsställande om  $np_j \geq 5$ , dvs det förväntade antalet observationer i varje kategori får inte vara för litet. Tumregel:  $np_j \geq 5$  för alla  $j = 1, \dots, r$ .
- ▶  $n$  måste vara stor.

Exempel på tavlan.



## TEST AV FÖRDELNING MED SKATTADE PARAMETRAR

En mycket vanlig situation är att man vill testa om data kan ses som utfall från en viss parametrisk fördelning, t ex Poisson-fördelning, där man inte vill specificera parametern värde. Här används också  $\chi^2$ -metoden.

- ▶ Låt  $x_1, \dots, x_n$  vara ett stickprov/data. Med hjälp av detta vill vi testa

$$H_0 : P(A_1) = p_1(\theta), P(A_2) = p_2(\theta), \dots, P(A_r) = p_r(\theta), \quad \text{för något } \theta,$$

där  $p_j(\theta)$  i  $H_0$  är sannolikheterna att få  $A_j$  under  $H_0$ .

- ▶ Okänd parameter (parametrar) skattas med  $\theta^* = \theta^*(x_1, \dots, x_n)$  under antagande att  $H_0$  är sann.



## TEST AV FÖRDELNING MED SKATTADE PARAMETRAR (FORTS.)

- ▶ Vidare jämförs  $x_j$  med de skattade förväntade antalet  $np_j^* = np_j(\theta^*)$ . Vi får testvariabel

$$Q_{obs.} = \sum_{j=1}^r \frac{(x_j - np_j^*)^2}{np_j^*}.$$

- ▶ Under  $H_0$  är  $Q_{obs.}$  ett utfall av en s.v.  $Q$  som är *approximativt*  $\chi^2(r - k - 1)$ -fördelad där  $k$  är antal skattade parametrar.
- ▶ Obs! Man förlorar en frihetsgrad i  $Q$ 's  $\chi^2$ -fördelning för varje skattad parameter i nollhypotesens sannolikhetsfunktion.
- ▶ Exempel på tavlan.





## HOMOGENITETSTEST

Vi vill testa om  $s$  försöksserier är *homogena* eller *lika fördelade* i meningen att  $P(A_j)$  för varje  $j$  är lika stora för samtliga försöksserier. Detta kan också testas med  $\chi^2$ -metoden.

Användning: man kan t ex undersöka om  $s$  stickprov kommer från samma fördelning, se Ex. 13.19, s. 346, *Kast med två tärningar*.

- ▶ Antag att man erhåller  $s$  st. serier av  $n_i$  st. oberoende observationer i serie nr  $i$ ,  $i = 1, \dots, s$  och bilda  $r$  kategorier. Det är praktiskt att presentera data i form av tabell:

	$A_1$	$A_2$	$\dots$	$A_r$	ant. försök
Serie 1	$x_{11}$	$x_{12}$	$\dots$	$x_{1r}$	$n_1$
Serie 2	$x_{21}$	$x_{22}$	$\dots$	$x_{2r}$	$n_2$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
Serie $s$	$x_{s1}$	$x_{s2}$	$\dots$	$x_{sr}$	$n_s$
Summa	$x_{\cdot 1}$	$x_{\cdot 2}$	$\dots$	$x_{\cdot r}$	$n$



## HOMOGENITETSTEST (FORTS.)

- ▶ Vi vill undersöka om serierna i tabell (det samma som på föregående sida) kan anses vara homogena, dvs testa hypotes  $H_0$  att samma uppsättning av sannolikheterna

$$p_1 = P(A_1), p_2 = P(A_2), \dots, p_r = P(A_r)$$

förekommer i alla serier.

	$A_1$	$A_2$	$\dots$	$A_r$	ant. försök
Serie 1	$x_{11}$	$x_{12}$	$\dots$	$x_{1r}$	$n_1$
Serie 2	$x_{21}$	$x_{22}$	$\dots$	$x_{2r}$	$n_2$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
Serie $s$	$x_{s1}$	$x_{s2}$	$\dots$	$x_{sr}$	$n_s$
Summa	$x_{.1}$	$x_{.2}$	$\dots$	$x_{.r}$	$n$

- ▶ Sannolikheterna  $p_j$  är helt ospecificerade.



## HOMOGENITETSTEST (FORTS.)

- ▶ För att pröva homogeniteten använder vi  $\chi^2$ -metoden och bildar testvariabeln

$$Q_{obs.} = \sum_{i=1}^s \sum_{j=1}^r \frac{(x_{ij} - n_i p_j^*)^2}{n_i p_j^*},$$

där skattningarna  $p_j^*$  av sannolikheter  $p_j$  fås med

$$p_j^* = \frac{1}{n} \sum_{i=1}^s x_{ij} = \frac{x_{.j}}{n}.$$

- ▶ Observera att i  $Q_{obs.}$  tolkar vi  $p_j^*$  som  $p_{j_{obs.}}$ .
- ▶ Tolkning:  $p_j^*$  är den bästa skattningen av  $P(A_j)$ -värdet som kan bildas med de sammanslagna observationerna. Om  $H_0$  är sann, dvs homogenitet gäller, kan ju serierna slås samman till en enda.



## HOMOGENITETSTEST (FORTS.)

- ▶ Under  $H_0$  är  $Q_{obs.}$  utfall av testsvariabel  $Q$  som är approximativt  $\chi^2((r-1)(s-1))$  (se Anm. 13.6 s. 345 för motivering till frihetsgradantalet). Tumregel för god approximation: för alla  $i, j$  är  $n_i p_{j,obs.}^* \geq 5$ .
- ▶ Beslutsregel blir alltså:

$$\begin{aligned} \text{Förkasta } H_0 \text{ om } & Q_{obs.} > \chi_{\alpha}^2((r-1)(s-1)) \\ \text{Förkasta ej } H_0 \text{ om } & Q_{obs.} \leq \chi_{\alpha}^2((r-1)(s-1)), \end{aligned}$$

vid approximativ signifikans nivån  $\alpha$ .

- ▶ *Exempel:* För att utvärdera en ny undervisningsmetod delar man upp studenterna på en kurs i två grupper och låret en grupp undervisas traditionellt och en med den nya metoden. Vid examination erhålls följande resultat (gammal betyg):



Metod/Betyg	U	G	Vg	Totalt
Traditionell	32	46	22	100
Ny	18	52	32	102
Totalt	50	98	54	202

- ▶ Vi vill testa hypotes:

$H_0$  : Metoderna är likvärdiga (med avseende på resultat)

$H_1$  : Metoderna är inte likvärdiga.

- ▶ Vi använder  $\chi^2$ -test där vi jämför de sex *observerade* frekvenserna med de *förväntade* frekvenserna under  $H_0$ . Vi får

$$\begin{aligned}
 Q_{obs.} = & \frac{(32 - 100 \frac{50}{202})^2}{100 \frac{50}{202}} + \frac{(46 - 100 \frac{98}{202})^2}{100 \frac{98}{202}} + \frac{(22 - 100 \frac{54}{202})^2}{100 \frac{54}{202}} \\
 & + \frac{(18 - 102 \frac{50}{202})^2}{102 \frac{50}{202}} + \frac{(52 - 102 \frac{98}{202})^2}{102 \frac{98}{202}} + \frac{(32 - 102 \frac{54}{202})^2}{102 \frac{54}{202}} = 5.8263.
 \end{aligned}$$



## HOMOGENITETSTEST (FORTS.): EXEMPEL

- ▶ Observera att för att beräkna de förväntade frekvenserna, har vi använt alla marginalvärden i tabellen.
- ▶ För signifikansnivån  $\alpha = 0.05$  bör test/beslutsregel vara:

$$\text{Förkasta } H_0 \text{ om } Q_{obs.} > \chi_{0.05}^2(2) = 5.9915.$$

- ▶ Slutsats:

Eftersom  $Q_{obs.} = 5.8263 < \chi_{0.05}^2(2) = 5.9915$  så förkastas inte  $H_0$ , dvs det finns inte någon signifikant skillnad mellan metoderna vid approximativ nivå 5%.



## TEST AV OBEROENDE (AV TVÅ EGENSKAPER)

$\chi^2$ -test kan också användas för att testa om två egenskaper ( t ex  $A$  och  $B$ ) är oberoende.

- ▶ Antag igen att  $n$  st. oberoende försök utförs där varje försök kan utfalla på  $sr$  olika sätt. Utfall  $(i, j)$ , bet. med  $B_i A_j$ , där  $i = 1, \dots, s$ ,  $j = 1, \dots, r$  inträffar med (okänd) sannolikhet  $p_{ij}$ . Låt  $x_{ij}$  vara absoluta frekvenser för  $B_i A_j$ .
- ▶ Observera att  $\sum_i \sum_j x_{ij} = n$  och  $\sum_i \sum_j p_{ij} = 1$
- ▶ Marginalsannolikheten för varje  $A_j$ , dvs för något av de  $s$  sätten  $B_1 A_j, B_2 A_j, \dots, B_s A_j$  blir

$$P(A_j) = p_{.j} = \sum_{i=1}^s p_{ij}$$

och motsvarande för varje  $B_i$

$$P(B_i) = p_{i.} = \sum_{j=1}^r p_{ij}.$$



## TEST AV OBEROENDE (FORTS.)

- ▶ Idé: Vi vill testa hypotes att de två egenskaperna,  $A$  och  $B$  är oberoende, dvs

$$H_0 : p_{ij} = P(B_i A_j) = P(B_i)P(A_j) = p_{i\cdot} p_{\cdot j} \quad \text{för alla } (i, j)$$

mot

$$H_1 : p_{ij} \neq p_{i\cdot} p_{\cdot j} \quad \text{för något } (i, j)$$

- ▶ De marginalsannolikheterna  $p_{\cdot j}$  och  $p_{i\cdot}$  kan skattas från data som

$$p_{\cdot j}^* = \frac{1}{n} \sum_{i=1}^s x_{ij} \quad p_{i\cdot}^* = \frac{1}{n} \sum_{j=1}^r x_{ij}$$





- ▶ Vi bildar testvariabeln

$$Q_{obs.} = \sum_{i=1}^s \sum_{j=1}^r \frac{(x_{ij} - np_{i\cdot}^* p_{\cdot j}^*)^2}{np_{i\cdot}^* p_{\cdot j}^*},$$

vilket är utfall av s. v.  $Q$  som är  $\chi^2((s-1)(r-1))$ -fördelad under  $H_0$ .

- ▶ Samma tumregel som för god approximation:  $np_{i\cdot}^* p_{\cdot j}^* \geq 5$ .
- ▶ Skillnaden mellan homogenitetstestet och oberoendetestet är
  - ▶ nullhypotesernas formulering och
  - ▶ att vid homogenitetstestet är *radmarginalerna givna i förväg* och kolumnmarginalerna är slumpmässiga, medan vid ett oberoende test *alla marginaler är slumpmässiga*.

## TEST AV OBEROENDE: EXEMPEL

Försäkringsbolaget, *FörSäkra AB*, vill undersöka om risken att råka ut för en trafikolycka är oberoende av förarens ålder. Man valde därför slumpmässigt ut 1000 av sina bilförsäkringar och registrerade personens ålder och antalet trafikolyckor under tiden mellan 2016-01-01 och 2016-12-31. Resultatet:

Antal olyckor	18-21 år	22-30 år	31-40 år	41-50 år	51-70 år
0	162	251	179	118	30
1	49	41	22	24	4
2 eller fler	39	28	9	28	16

Undersök med ett lämpligt statistiskt test, på nivån 1% om risken att råka ut för en olycka är beroende av åldern. Ett tydligt svar bör framgå. Svaret och den tillämpade statistiska metoden bör motiveras.



## TEST AV OBEROENDE: EXEMPEL (KONT.)

- ▶ Ett oberoendetest (avsnitt 14.3 i formelsamling) med hjälp av  $\chi^2$ -metoden. Vi utför  $n$  oberoende försök (här:  $n = 1000$ ) som kan utfalla på  $sr$  olika sätt, där  $i = 1, \dots, r$  och  $j = 1, \dots, s$  motsvarar ålderskategori respektive antal olyckor.
- ▶  $H_0$  : "Risken att råka för trafikolycka och ålder är oberoende", mothypotesen  $H_1$  är att  $H_0$  inte gäller dvs att risken att råka för trafikolycka beror av åldern.
- ▶ Vi gör en kontingenstabell med observerade antal olyckor (nästa sida)



Antal olyckor	Observerade antal					Total
	18-21 år	22-30 år	31-40 år	41-50 år	51-70 år	
0	162	251	179	118	30	740
1	49	41	22	24	4	140
2 eller fler	39	28	9	28	16	120
Total	250	320	210	170	50	1000

Under  $H_0$  kan vi skatta marginalsannolikheterna för varje ålderskategori som  $250/1000$ ,  $320/1000$ ,  $210/1000$ ,  $170/1000$  och  $50/1000$ . Det förväntade antalet trafikolyckor inom varje ålderskategori ges då som  $740 \cdot 0.25 = 185$ ,  $140 \cdot 0.32 = 35$  osv. I tabellform:

Antal olyckor	Förväntade antal under $H_0$					Total
	18-21 år	22-30 år	31-40 år	41-50 år	51-70 år	
0	185	236.8	155.4	125.8	37	740
1	35	44.8	29.4	23.8	7	140
2 eller fler	30	38.4	25.2	20.4	6	120
Total	250	320	210	170	50	1000

Alla tal i inuitabellen är  $\geq 5$ , vilket gör att  $\chi^2$ -approximationen är rimlig.



## TEST AV OBEROENDE: EXEMPEL (KONT.)

- ▶ Teststorheten blir

$$Q = \frac{(162 - 185)^2}{185} + \frac{(49 - 35)^2}{35} + \frac{(39 - 30)^2}{30} + \dots \\ + \frac{(4 - 7)^2}{7} + \frac{(16 - 6)^2}{6} = 53.6043.$$

Om  $H_0$  är sann så är 53.6043 ett utfall från en stokastisk variabel som approximativt har en  $\chi^2$ -fördelning med  $(5 - 1)(3 - 1) = 8$  fr gr.

- ▶ Eftersom  $Q = 53.6043 > \chi_{0.01}^2(8) = 20.1$  så kan  $H_0$  förk. på nivån 1%. Data ger belägg för att risken för att råka ut för trafikolyckor beror av åldern.
- ▶ *Alternativt* kan vi beräkna sannolikheten att en  $\chi^2$ -variabel är större än eller lika med 53.6043 (X2cdf på en TI-räknare). Denna sannolikhet, dvs  $p$ -värdet för testet, är .2372e-09. Detta  $p$ -värde är mycket lågt så vi förkastar  $H_0$ .
- ▶ Både teststorheten och  $p$ -värdet fås direkt med funktionen 2-Test på en TI-räknare.

