

# Föreläsning 10: Pearsons $\chi^2$ -test

Johan Thim (johan.thim@liu.se)

14 oktober 2018

*"Oh, you suffer beautifully."*  
–Pinhead

## 1 Konvergens

För att kunna få lite precision i argumenten i detta område behöver vi lite begrepp angående konvergens av stokastiska variabler. Eftersom vi har introducerat sannolikhet så uppstår nu en hel rad spännande möjligheter till olika typer av konvergens. Vissa av dessa har vi redan (mer eller mindre) implicit stött på. Tänk på de stora talens lag eller konsistens hos skattningar (detta brukar vara konvergens i sannolikhet) respektive centrala gränsvärdesatsen (konvergens i fördelning).

Låt oss börja med en annan typ av konvergens som kan vara värd att ha sett om inte annat än för att kunna säga saker som att något är "nästan säkert" och faktiskt mena något väldigt specifikt...



**Definition.** Låt  $X_n, n = 1, 2, \dots$ , vara en följd stokastiska variabler. Vi säger att  $X_n$  konvergerar till  $X$  nästan säkert (almost surely) om

$$P(\{\omega \in \Omega : X_n(\omega) \rightarrow X(\omega)\}) = 1.$$

Vi skriver i detta fall att  $X_n \xrightarrow{\text{a.s.}} X$ .

Tänk på att en stokastisk variabel  $X$  avbildar ett utfallsrum  $\Omega$  in i  $\mathbf{R}$  (eller  $\mathbf{R}^n$ ). Vad definitionen ovan säger är att denna *funktion* konvergerar punktvis  $X_n(\omega) \rightarrow X(\omega)$  för alla  $\omega$  förutom en delmängd som har sannolikhet noll.



### Konvergens i sannolikhet

**Definition.** Låt  $X_n, n = 1, 2, \dots$ , vara en följd stokastiska variabler. Vi säger att  $X_n$  konvergerar till en stokastisk variabel  $X$  i sannolikhet om för alla  $\epsilon > 0$  så gäller att

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \epsilon) = 0$$

och vi skriver i detta fall att  $X_n \xrightarrow{P} X$ . Generaliserar naturligt till högre dimensioner.

Vi kan notera att

$$X_n \xrightarrow{\text{a.s.}} X \quad \Rightarrow \quad X_n \xrightarrow{P} X,$$

men inte omvänt. Detta är inte självklart utan hänger i princip på att vi kan byta ut ordningen på att beräkna sannolikhet och ta ett gränsvärde. Den intresserade kan slå upp Fatous lemma. Den sista konvergenstypen vi betraktar är konvergens i fördelning. Vad detta innebär informellt är att fördelningsfunktionerna för  $X_n$  konvergerar punktvis mot fördelningsfunktionen för  $X$ .



**Definition.** Låt  $X_n$ ,  $n = 1, 2, \dots$ , vara en följd stokastiska variabler. Vi säger att  $X_n$  konvergerar till en stokastisk variabel  $X$  i fördelning om

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$$

för alla  $x$  (där  $F$  är kontinuerlig). Här är  $F_{X_n}$  och  $F_X$  respektive fördelningsfunktion, och vi skriver att  $X_n \xrightarrow{D} X$ . I högre dimensioner formuleras ofta kraven direkt i termer av sannolikhet enligt

$$\mathbf{X}_n \xrightarrow{D} \mathbf{X} \quad \Leftrightarrow \quad \lim_{n \rightarrow \infty} P(\mathbf{X}_n \in E) = P(\mathbf{X} \in E)$$

för alla rimliga mängder  $E$  sådana att  $P(\partial E) = 0$ .

Egenskapen att mängden  $E$  uppfyller att  $P(\partial E) = 0$  brukar kallas för att  $E$  är en kontinuitetsmängd (kommer från mått-teorin) för måttet  $P$ . Alternativt kan man betrakta fördelningsfunktionen, så låt

$$E(\mathbf{x}) = \{\mathbf{y} \in \mathbf{R}^k : y_1 \leq x_1, y_2 \leq x_2, \dots, y_k \leq x_k\}$$

så att fördelningsfunktionen  $F$  ges av  $F(x_1, x_2, \dots, x_k) = P(\mathbf{X} \in E(\mathbf{x}))$ . Detta följer direkt från att den flerdimensionella fördelningsfunktionen ges av

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_k \leq x_k)$$

Randen  $\partial E$  till  $E$  kan vi uttrycka som

$$E(\mathbf{x}) \cap \{\mathbf{y} \in \mathbf{R}^k : y_i = x_i \text{ för något } i, 1 \leq i \leq k\},$$

så om  $P(\mathbf{X} \in \partial E(\mathbf{x})) = 0$  är helt enkelt  $F$  kontinuerlig i punkten  $\mathbf{x}$ .

Värt även att notera att

$$X_n \xrightarrow{P} X \quad \Rightarrow \quad X_n \xrightarrow{D} X.$$

Även här är beviset ganska tekniskt, men det är värt att komma ihåg att konvergens i fördelning är den svagaste typen av konvergens vi tagit upp.



**Definition.** Om  $X_n \xrightarrow{D} X$  kallas fördelningen för  $X$  för den **asymptotiska** fördelningen för sekvensen  $X_n$ ,  $n = 1, 2, \dots$

Ibland krävs att denna fördelning inte är allt för degenererad för att kallas för en asymptotisk fördelning.

## 1.1 Ordo i sannolikhet

Redo för något riktigt skoj? För att beskriva hastigheten hos konvergens används ibland varianter av ordo-notationen ni har stött på tidigare (envariabel del 2).

Vi skriver att  $X_n = o_p(a_n)$  om  $\frac{X_n}{a_n} = o_p(1)$ , där

$$Y_n = o_p(1) \Leftrightarrow Y_n \rightarrow 0 \text{ i sannolikhet.}$$

Vi säger att  $X_n = O_p(a_n)$  om det för varje  $\epsilon > 0$  finns ett ändligt  $M > 0$  och ett ändligt  $N > 0$  så att

$$P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \epsilon \text{ för alla } n > N.$$

## 1.2 Ett par användbara resultat (utan bevis)

Vi har nu introducerat några begrepp kring konvergens av följder av stokastiska variabler. Ofta är man intresserad av funktioner av stokastiska variabler på olika sätt, så vad kan man säga om konvergensen efter att ha gjort någon form av sammansättning?

Till exempel kan vi notera att  $X_n \xrightarrow{D} X$  och  $Y_n \xrightarrow{D} Y$  *inte* medför att  $X_n + Y_n \xrightarrow{D} X + Y$  eller  $X_n Y_n \xrightarrow{D} XY$  i det generella fallet. Men under vissa förutsättningar har vi resultat som ofta duger när vi vill åt ovanstående.



### The Continuous Mapping Theorem

**Sats.** Låt  $g$  vara kontinuerlig. Då gäller att

$$(i) X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X);$$

$$(ii) X_n \xrightarrow{D} X \Rightarrow g(X_n) \xrightarrow{D} g(X).$$

Generaliserar till vektorvärda stokastiska variabler.

Kontinuerliga operationer fungerar alltså precis som vi förväntar oss. Konvergens i fördelning bevaras av kontinuerliga avbildningar. Beviset är inte jättekomplicerat, men bygger på argument och definitioner vi inte har tillgång till i nuläget (åter igen denna mått- och integrationsteori). Man kan ju tro att summor av följder borde fungera lika enkelt, men så är inte fallet om vi endast har konvergens i fördelning.



### Slutskys sats

**Sats.** Om  $X_n \xrightarrow{D} X$  och  $Y_n \xrightarrow{P} c$ , där  $c$  är en konstant, så gäller att

$$(i) X_n + Y_n \xrightarrow{D} X + c \quad (ii) X_n Y_n \xrightarrow{D} X c \quad (iii) \frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{c} \text{ under förutsättning att } c \neq 0.$$

Även detta generaliserar till vektorvärda stokastiska variabler.

### 1.3 Flerdimensionella centrala gränsvärdessatsen

Bara för att påminna så såg vi i grundkursen i sannolikhetslära att summan av oberoende likafördelade variabler (med ändlig varians) alltid går mot en normalfördelning (konvergens i fördelning). Kompakt uttryckt gäller alltså att  $\sqrt{n}(\bar{X} - \mu) \xrightarrow{D} N(0, \sigma^2)$  om  $E(X_i) = \mu$  och  $V(X_i) = \sigma^2$ . Motsvarande gäller i högre dimensioner, men i vanlig ordning har vi nu en hel kovariansmatris att hålla ordning på istället för endast variansen. En variant av den multivariata CGS kan formuleras enligt följande.



#### Flerdimensionella centrala gränsvärdessatsen

**Sats.** Låt  $\mathbf{X} = (X_1, \dots, X_k)$  vara en vektorvärd stokastisk variabel med kovariansmatris  $C$ . Implicit här är att  $V(X_j) < \infty$  för  $j = 1, 2, \dots, k$ . Låt  $\mathbf{X}_n$  vara en följd av oberoende vektorer med samma fördelning som  $\mathbf{X}$ . Då gäller att

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{X}_i - E(\mathbf{X})) \xrightarrow{D} N(0, C).$$

Notera även att

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{X}_i - E(\mathbf{X})) = \sqrt{n} (\bar{\mathbf{X}}_n - E(\mathbf{X})),$$

så vi kan mer kompakt skriva  $\sqrt{n} (\bar{\mathbf{X}}_n - E(\mathbf{X})) \rightarrow N(0, C)$ .

### 1.4 Delta-metoden

En naturlig fråga är följande: om vi vet att  $X_n$  har en asymptotisk fördelning, vad kan man säga om  $g(X_n)$ ? Ett naturligt angreppssätt är givetvis att helt enkelt ta en Taylorutveckling av  $g$  och visa att resttermen beter sig som  $o_p(1)$  så att den inte stör. Det generella fallet blir lite bökiigt, så vi koncentrerar oss på normalfördelningen.



#### Delta-metoden för asymptotisk normalfördelning

**Sats.** Antag att  $X_n \xrightarrow{P} \theta$  och  $\sqrt{n}(X_n - \theta) \xrightarrow{D} X \sim N(0, \sigma^2)$  då  $n \rightarrow \infty$  (i princip resultatet av centrala gränsvärdessatsen) och låt  $g \in C^1$  i en omgivning av  $\theta$  samt antag att  $g'(\theta) \neq 0$ . Då gäller att

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{D} X \sim N(0, (g'(\theta))^2 \sigma^2),$$

då  $n \rightarrow \infty$ .

**Bevis.** Enligt medelvärdessatsen så gäller att

$$g(X_n) - g(\theta) = g'(\xi) (X_n - \theta),$$

där  $\xi$  ligger mellan  $X_n$  och  $\theta$ . Eftersom  $X_n \xrightarrow{P} \theta$  så måste även  $\xi \xrightarrow{P} \theta$ . Satsen ovan om kontinuerliga avbildningar medför då att  $g'(\xi) \xrightarrow{P} g'(\theta)$ . Alltså måste

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{D} Z \sim N(0, (g'(\theta))^2 \sigma^2),$$

allt enligt Slutskys sats! Vi kan även formulera det hela asymptotiskt enligt

$$\sqrt{n}(g(X_n) - g(\theta)) = g'(\theta)\sqrt{n}(X_n - \theta) + o_p(1),$$

genom att helt enkelt skriva

$$\begin{aligned}\sqrt{n}(g(X_n) - g(\theta)) &= \sqrt{n}g'(\xi)(X_n - \theta) \\ &= \underbrace{\sqrt{n}(X_n - \theta)g'(\theta)}_{=O_p(1)} + \underbrace{\sqrt{n}(X_n - \theta)(g'(\xi) - g'(\theta))}_{=o_p(1)},\end{aligned}$$

där vi nyttjar att  $\sqrt{n}(X_n - \theta)$  konvergerar i fördelning – vilket betyder stokastiskt begränsad så  $O_p(1)$  – och att  $g'(\xi) \xrightarrow{P} g'(\theta)$ .  $\square$

Vad händer i flera dimensioner? I princip är det helt analogt. Låt  $\hat{\theta}$  vara en konsistent skattning av  $\theta$  så att

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} Z \sim N(\mathbf{0}, C).$$

Om vi för enkelhetens skull antar att  $g \in C^2$  i en omgivning av  $\theta$ , så är som bekant

$$g(\hat{\theta}) = g(\theta) + \nabla g(\theta)^T \cdot (\hat{\theta} - \theta) + R(\hat{\theta}).$$

Om vi betraktar variansen för vår approximation (där vi bortser från resttermen) så ser vi att

$$\begin{aligned}V\left(g(\theta) + \nabla g(\theta)^T \cdot (\hat{\theta} - \theta)\right) &= V\left(\nabla g(\theta)^T \cdot \hat{\theta}\right) = \text{Cov}\left(\nabla g(\theta)^T \cdot \hat{\theta}\right) \\ &= \nabla g(\theta)^T \text{Cov}(\hat{\theta}) \nabla g(\theta) = \nabla g(\theta)^T \frac{1}{n} C \nabla g(\theta).\end{aligned}$$

Genom att likt i envariabelfallet använda medelvärdessatsen kan vi nu visa att

$$\sqrt{n}\left(g(\hat{\theta}) - g(\theta)\right) \xrightarrow{D} Z \sim N\left(\mathbf{0}, \nabla g(\theta)^T C \nabla g(\theta)\right).$$

## 2 Det grundläggande $\chi^2$ -testet

Antag att vi har följande situation

- (i) Vi har  $n$  stycken oberoende stokastiska variabler  $X_j$  med samma fördelning, där  $X_j$  har precis  $k$  möjliga utfall.
- (ii) Numrera utfallen enligt  $A_1, \dots, A_k$  och låt  $p_j = P(A_j)$  vara respektive sannolikhet. Då är  $p_1 + p_2 + \dots + p_k = 1$ .
- (iii) Låt  $Y_i$ ,  $i = 1, 2, \dots, k$ , vara antalet gånger händelsen  $A_i$  inträffar.

För att konkretisera en aning, tänk att vi har  $k$  stycken lådor  $A_j$  vi kastar bollar i. Experimentet är uppställt så att en kastad boll alltid hamnar i en låda. Vi låter  $p_j$  vara sannolikheten att en boll hamnar i låda  $A_j$ . Vi kastar  $n$  bollar (oberoende) och räknar sedan hur många bollar  $Y_j$  som det finns i varje låda. Givetvis kommer  $Y_j \sim \text{Bin}(n, p_j)$ , men variablerna  $Y_j$  är *inte* oberoende av varandra (antalet bollar i alla lådorna summerar till  $n$ ).

Vad vi kommer göra är att betrakta uppdelningar av denna typ och ställa upp hypotestest där vi låter nollhypotesen  $H_0$  ges av

$$H_0 : P(A_1) = p_1, P(A_2) = p_2, \dots, P(A_k) = p_k,$$

där  $p_1, p_2, \dots, p_k$  är sannolikheter så att  $p_1 + \dots + p_k = 1$ , och testar mot hypotesen

$$H_1 : \text{det finns något } j \text{ så att } P(A_j) \neq p_j.$$

Om  $H_0$  är sann, så blir de förväntade frekvenserna  $E(Y_j) = n \cdot p_j$ ,  $j = 1, 2, \dots, k$ . Låt oss definiera

$$q = \sum_{j=1}^k \frac{(y_j - np_j)^2}{np_j},$$

där  $y_j$  är observationen av  $Y_j$ . Ett stort värde på  $q$  borde rimligen indikera att  $H_0$  inte gäller (åtminstone något  $p_j$  måste skilja sig markant från det förväntade värdet  $np_j$ ).

Storheten  $q$  är en observation av den stokastiska variabeln

$$Q = \sum_{j=1}^k \frac{(Y_j - np_j)^2}{np_j} \stackrel{\text{appr.}}{\sim} \chi^2(k-1).$$

Att detta blir approximativt  $\chi^2$ -fördelat följer av följande sats.



**Sats.** Med beteckningarna ovan gäller att  $\sum_{j=1}^k \frac{(Y_j - np_j)^2}{np_j} \xrightarrow{D} X$ , där  $X \sim \chi^2(k-1)$ . Konvergensten är alltså i fördelning.

**Bevis.** Eftersom  $Y_j$  är binomialfördelat vet vi att  $E(Y_j) = np_j$  och  $V(Y_j) = np_j(1 - p_j)$ , så de standardiserade variablerna

$$\frac{Y_j - np_j}{\sqrt{np_j(1 - p_j)}} \xrightarrow{D} \widetilde{Z}_j \sim N(0, 1),$$

för något  $\widetilde{Z}_j$  enligt centrala gränsvärdessatsen (CGS). Konvergensten är i meningen att fördelningsfunktionen  $F_{n,j}(y) \rightarrow \Phi(y)$  för alla  $y \in \mathbf{R}$ . En följd av detta är att

$$\frac{Y_j - np_j}{\sqrt{np_j}} \xrightarrow{D} Z_j \sim N(0, 1 - p_j),$$

eftersom om  $U_n \xrightarrow{D} U$  så gäller att  $h(U_n) \xrightarrow{D} h(U)$  för alla kontinuerliga funktioner  $h$  (brukar kallas sannolikhetsteorins open mapping theorem). Anledningen till den sista manövern är att vi ska få det lite lättare att analysera beroendestrukturen hos  $Z_j$ ,  $j = 1, 2, \dots, k$ . Eftersom väntevärdet är  $E(Y_j) = np_j$  kommer

$$\begin{aligned} C \left( \frac{Y_i - np_i}{\sqrt{np_i}}, \frac{Y_j - np_j}{\sqrt{np_j}} \right) &= E \left( \frac{Y_i - np_i}{\sqrt{np_i}} \frac{Y_j - np_j}{\sqrt{np_j}} \right) = \frac{1}{n\sqrt{p_i p_j}} (E(Y_i Y_j) - 2n^2 p_i p_j + n^2 p_i p_j) \\ &= \frac{1}{n\sqrt{p_i p_j}} (E(Y_i Y_j) - n^2 p_i p_j) \end{aligned}$$

För att beräkna  $E(Y_i Y_j)$  går vi tillbaka till variablerna  $X_i$ ,  $i = 1, 2, \dots, n$ . Låt  $I_A$  beteckna indikatorfunktionen för mängden  $A$ . Detta innebär att

$$I_{A_j}(X_i) = \begin{cases} 1 & \text{om } X_i \in A_j, \\ 0 & \text{om } X_i \notin A_j. \end{cases}$$

Vi kan då skriva  $Y_j = \sum_{i=1}^n I_{A_j}(X_i)$  och eftersom  $X_i$  är Bernoullifördelade (2-punktsfördelade) följer det att  $E(I_{A_j}(X_i)) = p_j$ . Vi har nu, för  $i \neq j$ ,

$$\begin{aligned} E(Y_i Y_j) &= E\left(\left(\sum_{l=1}^n I_{A_i}(X_l)\right)\left(\sum_{m=1}^n I_{A_j}(X_m)\right)\right) = E\left(\sum_{l=1}^n \sum_{m=1}^n I_{A_i}(X_l) I_{A_j}(X_m)\right) \\ &= E\left(\sum_{l=1}^n I_{A_i}(X_l) I_{A_j}(X_l)\right) + E\left(\sum_{l=1}^n \sum_{\substack{m=1 \\ m \neq l}}^n I_{A_i}(X_l) I_{A_j}(X_m)\right) \\ &= 0 + \sum_{l=1}^n \sum_{\substack{m=1 \\ m \neq l}}^n E(I_{A_i}(X_l)) E(I_{A_j}(X_m)) = \sum_{l=1}^n \sum_{\substack{m=1 \\ m \neq l}}^n p_i p_j = n(n-1)p_i p_j, \end{aligned}$$

eftersom  $I_{A_i}(X_l) I_{A_j}(X_l) = 0$  (samma boll kan inte hamna i två lådor) samt att  $I_{A_i}(X_l)$  och  $I_{A_j}(X_m)$  är oberoende om  $l \neq m$ . Således blir

$$C\left(\frac{Y_i - np_i}{\sqrt{np_i}}, \frac{Y_j - np_j}{\sqrt{np_j}}\right) = -\sqrt{p_i p_j},$$

för  $i \neq j$ . Följaktligen måste således kovariansmatrisen för  $\mathbf{Z} = (Z_1 \ Z_2 \ \cdots \ Z_k)^T$  ha utseendet

$$C_{\mathbf{Z}} = \begin{pmatrix} 1 - p_1^2 & -\sqrt{p_1 p_2} & -\sqrt{p_1 p_3} & \cdots & -\sqrt{p_1 p_k} \\ -\sqrt{p_2 p_1} & 1 - p_2^2 & -\sqrt{p_2 p_3} & \cdots & -\sqrt{p_2 p_k} \\ -\sqrt{p_3 p_1} & -\sqrt{p_3 p_2} & 1 - p_3^2 & \cdots & -\sqrt{p_3 p_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\sqrt{p_k p_1} & -\sqrt{p_k p_2} & -\sqrt{p_k p_3} & \cdots & 1 - p_k^2 \end{pmatrix}.$$

vilket kan skrivas lite mer kompakt som  $C_{\mathbf{Z}} = I - \mathbf{p}\mathbf{p}^T$ , där  $\mathbf{p} = (\sqrt{p_1} \ \sqrt{p_2} \ \cdots \ \sqrt{p_k})^T$ . Denna omskrivning gör att vi enkelt kan se att

$$(I - \mathbf{p}\mathbf{p}^T)^2 = I - \mathbf{p}\mathbf{p}^T \quad \text{och} \quad (I - \mathbf{p}\mathbf{p}^T)^T = I - \mathbf{p}\mathbf{p}^T,$$

så  $I - \mathbf{p}\mathbf{p}^T$  är en projektionsmatris och har därför egenvärdena  $\lambda = 0$  och  $\lambda = 1$ . Vi har nu att  $\mathbf{Z} \sim N(\mathbf{0}, C_{\mathbf{Z}})$ . På samma sätt som i beviset av regressionsanalysens huvudsats ser vi att

$$\text{rank}(I - \mathbf{p}\mathbf{p}^T) = \text{tr}(I - \mathbf{p}\mathbf{p}^T) = k - 1,$$

så  $\lambda = 0$  är ett enkelt egenvärde. Matrisen är symmetrisk och positivt semidefinit, så det finns en ON-matris  $C$  så att  $C^T C_{\mathbf{Z}} C = \text{diag}(1, 1, \dots, 1, 0)$  blir en diagonalmatris. Om vi låter  $\mathbf{W} = C\mathbf{Z}$  ser vi att  $\mathbf{W} \sim N(\mathbf{0}, \text{diag}(1, 1, \dots, 1, 0))$  och att

$$\mathbf{Z}^T \mathbf{Z} = \mathbf{W}^T \mathbf{W} = \sum_{j=1}^{k-1} W_j^2,$$

där  $W_j \sim N(0, 1)$  är oberoende. Denna summa är som bekant  $\chi^2(k-1)$ -fördelad! □



### När duger approximationen?

Föregående sats gäller alltså *asymptotiskt* (då  $n \rightarrow \infty$ ) och säger inget direkt om vad som gäller i det enskilda fallet. En tumregel är att vi vill ha  $np_j \geq 5$  för  $j = 1, 2, \dots, k$  för att vara ganska säkra på att approximationen är bra. Har vi lådor med väldigt få "bollar" i kan det hända att testet inte blir bra.

## 3 Test av given diskret fördelning

Låt  $X_1, X_2, \dots, X_n$  vara oberoende diskreta stokastiska variabler med  $X_j \in A$  för någon diskret mängd  $A$ . Vi är intresserade av att testa om  $X_j \sim F$  för någon given diskret fördelning med sannolikhetsfunktion  $p(j)$ ,  $j \in A$ . Vi kommer använda nollhypotesen

$$H_0 : P(X = j) = p(j), \quad j \in A,$$

och testar den med mothypotesen

$$H_1 : P(X = j) \neq p(j) \text{ för något } j \in A.$$



### Exempel

Den stokastiska variabeln  $X$  antar värden i mängden  $\{0, 1, 2\}$ . Vid 1250 observationer fann man att  $X = 0$  783 gånger,  $X = 1$  425 gånger samt  $X = 2$  42 gånger. Testa med signifikansnivån 1% om  $X \sim \text{Bin}(2, 1/5)$ .

**Lösning.** Vi låter  $H_0 : X \sim \text{Bin}(2, 1/5)$ . Om vi antar att  $H_0$  är sann så gäller att

$$P(X = 0) = \binom{2}{0} \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^2 = \frac{16}{25},$$

$$P(X = 1) = \binom{2}{1} \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^1 = \frac{8}{25},$$

$$P(X = 2) = \binom{2}{2} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^0 = \frac{1}{25}.$$

Kom ihåg att kontrollera att dessa summerar till 1, det är en billig kontroll på tentan. Utifrån detta kan vi beräkna de förväntade frekvenserna vid 1250 försök (om  $H_0$  är sann):

$$np_j = \begin{cases} 800, & j = 0, \\ 400, & j = 1, \\ 50, & j = 2. \end{cases}$$

Testvariabeln  $q$  ges nu av

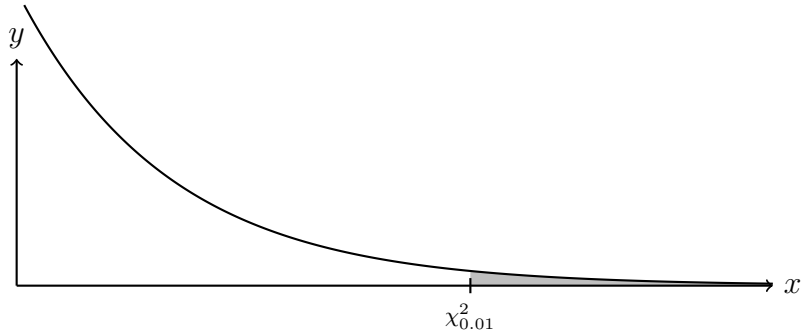
$$q = \sum_{j=0}^2 \frac{(x_j - np_j)^2}{np_j} = \frac{(783 - 800)^2}{800} + \frac{(425 - 400)^2}{400} + \frac{(42 - 50)^2}{50} \approx 3.2038.$$

Eftersom  $k = 3$  är  $q$  en observation av  $Q \stackrel{\text{appr.}}{\sim} \chi^2(2)$  om  $H_0$  är sann. Vi finner att

$$0.01 = P(Q > \chi_{0.01}^2(2)) \quad \Leftrightarrow \quad \chi_{0.01}^2(2) = 9.21$$

ur tabell.





Eftersom  $q = 3.2038 < 9.21$  kan vi inte förkasta  $H_0$ . Fördelningen kan mycket riktigt vara binomialfördelning med  $p = 1/5$ .

## 4 Test för kontinuerlig fördelning

Om vi istället har en kontinuerlig situation där vi vill testa om mätdata följer en given fördelning  $F$  måste vi agera lite annorlunda. Vi skulle önska att ställa upp

$$H_0 : X \sim F$$

mot

$$H_1 : X \text{ har ej fördelningen } F.$$

Men detta blir lite för komplicerat i det generella fallet.

Istället gör vi så att vi diskretiserar det hela på något sätt. Vi gör oftast detta genom att skapa lådor i form av intervall och sedan undersöka hur många observationer som hamnar i varje delintervall. Detta gör att vi inte exakt testar om nollhypotesen ovan utan vi testar en svagare nollhypotes.

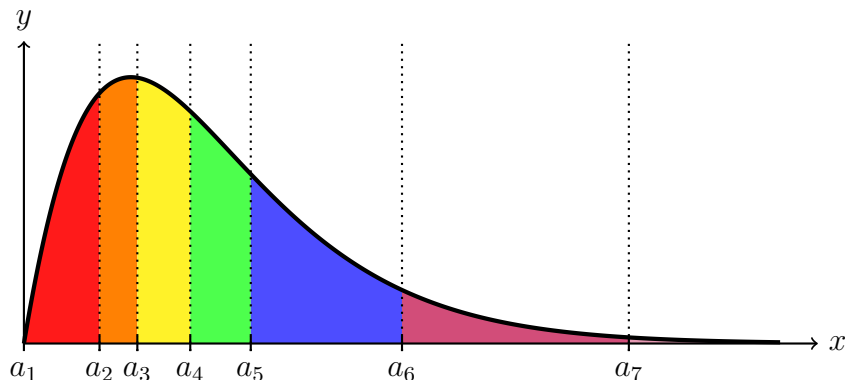
Låt  $X_i, i = 1, 2, \dots, n$  vara oberoende och likafördelade variabler med täthetsfunktion  $f(x)$ . Vi väljer  $a_j, j = 1, 2, \dots, k + 1$ , så att

$$-\infty \leq a_1 < a_2 < \dots < a_k < a_{k+1} \leq \infty$$

och definierar  $A_j = [a_j, a_{j+1}[$  för  $j = 2, 3, \dots, k$  och låter typiskt  $A_1 = ]-\infty, a_2[$ . Vi definierar sedan

$$p_j = P(X_i \in A_j) = \int_{a_j}^{a_{j+1}} f(x) dx.$$

Om  $f$  är en täthetsfunktion så blir nu  $p_1 + p_2 + \dots + p_k = 1$  och vi har täckt alla möjligheter. Om stödet för  $f$  inte är hela  $\mathbf{R}$  modifierar vi naturligt definitionen (eller låter  $f(x) = 0$  utanför sin definition). En tumregel för valet är att vi låter  $k \approx n/10$ . En annan tumregel är att välja intervallen så stora att alla  $p_j$  är ungefär lika stora.



Hypotesen vi kommer testa är

$$H_0 : P(X \in A_j) = p_j, \quad j = 1, 2, \dots, k,$$

mot

$$H_1 : P(X \in A_j) \neq p_j \text{ för något } j.$$

Skulle  $X$  ha rätt fördelning kommer  $H_0$  att vara sann med stor sannolikhet, men om vi styrker  $H_0$  innebär det inte nödvändigtvis att det är just den fördelning vi utgick från när vi ställde upp  $A_j$  som är den sanna (bara någon med motsvarande sannolikheter i uppdelningen). Vill man ha ett starkare resultat krävs andra metoder.



### Exempel

Säljaren på ELFA hävdar bestämt att livslängden på en komponent är exponentialfördelad med väntevärde 2 år. Uttråkade pensionären Sture tror inte på det utan köper 50 stycken komponenter för att testa. Sture kopplar upp komponenterna och kikar till var 6:e månad för att se hur många som gått sönder.

Tid (mån)	< 6	< 12	< 18	< 24	< 30	< 36	< 42	< 48	< 54	< 60
Antal:	11	19	25	31	36	39	39	40	42	43

Undersök om antagandet är rimligt på approximativt 1% nivån.

**Lösning.** Vi kan organisera om datan mer användbart enligt hur många enheter som gick sönder under en viss tidsenhet. För att få ungefär jämnstora klasser så buntar vi ihop enligt följande.

Tid	Hur många dog
$I_1 = [0, 6)$	11
$I_2 = [6, 12)$	8
$I_3 = [12, 24)$	12
$I_4 = [24, 36)$	8
$I_5 = [36, \infty)$	11

Om vi antar  $H_0$  så gäller att täthetsfunktionen för livslängden hos en komponent  $X$  ges av  $f(x) = \mu^{-1} \exp(-\mu^{-1} x)$ , så

$$P(a \leq X < b) = \int_a^b \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) dx = \exp\left(-\frac{a}{\mu}\right) - \exp\left(-\frac{b}{\mu}\right).$$

Med siffrorna ovan ser vi att

$$P(X \in I_k) = \begin{cases} p_1 = 0.2212, & k = 1, \\ p_2 = 0.1723, & k = 2, \\ p_3 = 0.2387, & k = 3, \\ p_4 = 0.1447, & k = 4, \\ p_5 = 0.2231, & k = 5. \end{cases}$$

Teststorheten vi använder kommer nu ges av

$$q = \sum_{j=1}^5 \frac{(x_j - np_j)^2}{np_j} = \frac{(11 - 50 \cdot 0.2212)^2}{50 \cdot 0.2212} + \dots + \frac{(11 - 50 \cdot 0.2231)^2}{50 \cdot 0.2231} = 0.1276.$$

Om  $H_0$  är sann så kommer  $q$  vara en observation av  $Q \stackrel{\text{appr.}}{\sim} \chi^2(5-1) = \chi^2(4)$ , så med det kritiska området  $C = (0, c)$  där  $c = 13.28$ , ser vi att vi inte kan förkasta  $H_0$ . Säljaren kan mycket väl ha rätt.

## 5 Skattade storheter

Normalt sätt kanske vi inte får exakt väntevärde (eller andra parametrar i fördelningen) utan dessa måste skattas innan vi kan utföra testet. Hur påverkar det fördelningen för teststorheten  $Q$ ? Svaret är enkelt: för varje skattning vi gör tappar vi en frihetsgrad, under förutsättningen att skattningen är vettig (ML-skattningar brukar bete sig bra). Bevis är däremot lite bökgigare (å andra sidan får vi det första  $\chi^2$ -testet mer eller mindre på köpet). Får jag tid över kommer jag skriva ned det och uppdatera anteckningarna. Om vi antar att sannolikheterna  $p_j$  beror på okända  $\boldsymbol{\theta} = (\theta_1 \theta_2 \dots \theta_r)^T$ , så gäller alltså att

$$Q = \sum_{j=1}^k \frac{(Y_j - n\hat{p}_j(\boldsymbol{\theta}))^2}{n\hat{p}_j(\boldsymbol{\theta})} \stackrel{\text{appf.}}{\approx} \chi^2(k - r - 1),$$

under förutsättning att skattningarna som används beter sig tillräckligt bra.



### Exempel

Linnea gör en signalbehandlingslaboration i matlab men hennes algoritm fungerar inte som planerat. Givetvis tycker Linnea att felet måste ligga i matlabs sätt att generera normalfördelade slumpetal. För att testa hypotesen att slumpetalen inte är normalfördelade genererar Linnea 1000 slumpetal och sorterar dessa i storleksordning följt av en klassindelning så det är precis 100 element i varje klass. Gränserna kan ses nedan.

Under gräns	Övre gräns
1.57	12.46
12.47	14.98
15.00	17.03
17.04	18.77
18.80	20.32
20.33	21.75
21.76	23.25
23.26	24.99
25.00	27.42
27.43	40.80

Det beräknade medelvärdet är  $\bar{x} = 20.14$  och stickprovsvariansen är  $s^2 = 35.25$ . Testa på nivån 5% om värdena är normalfördelade.

**Lösning.** Låt  $H_0$  : datan kommer från  $N(\mu, \sigma^2)$  och  $H_1$  : datan är inte normalfördelad. Om vi använder  $\bar{x} = 20.14$  som skattning för väntevärdet och  $s = \sqrt{35.25} = 5.94$  som skattning för standardavvikelsen, så kan vi (om vi antar att  $H_0$  är sann) beräkna sannolikheterna för en normalfördelad variabel  $Z$  att hamna i de olika klasserna enligt

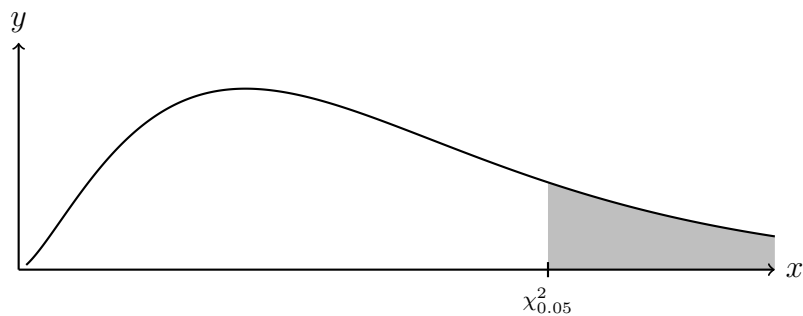
$$\begin{aligned} P(a \leq Z < b) &= P\left(\frac{a - \mu}{\sigma} \leq \frac{Z - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \\ &\approx \Phi\left(\frac{b - 20.14}{5.94}\right) - \Phi\left(\frac{a - 20.14}{5.94}\right). \end{aligned}$$

Resultatet kan beskådas nedan.

Intervall	Sannolikhet
$I_1 = (-\infty, 12.46)$	0.10
$I_2 = [12.47, 14.99)$	0.09
$I_3 = [15.00, 17.03)$	0.11
$I_4 = [17.04, 18.77)$	0.11
$I_5 = [18.78, 20.32)$	0.10
$I_6 = [20.33, 21.75)$	0.09
$I_7 = [21.76, 23.25)$	0.09
$I_8 = [23.26, 24.99)$	0.09
$I_9 = [25.00, 27.42)$	0.10
$I_{10} = [27.43, \infty)$	0.11

Vi ser redan nu att sannolikheterna väldigt nära hamnar runt 10-delar (vilket borde ske om normalfördelning gäller med tanke på konstruktionen). Men låt oss ställa upp teststorheten och se:

$$q = \sum_{j=1}^{10} \frac{(x_j - n\hat{p}_j)^2}{n\hat{p}_j} = \frac{(100 - 1000 \cdot 0.10)^2}{1000 \cdot 0.10} + \dots + \frac{(100 - 1000 \cdot 0.11)^2}{1000 \cdot 0.11} = 4.34.$$



Om  $H_0$  är sann så är  $q$  en observation av  $\chi^2(10 - 2 - 1) = \chi^2(7)$  eftersom vi skattar två parametrar. På nivån 0.1% så gäller att  $P(Q > 14.07) = 0.05$ , och då  $4.34 < 14.07$  så kan vi inte förkasta nollhypotesen. Linnea har antagligen implementerat sin algoritm fel.

Att testa normalfördelning på detta sätt är inte helt lämpligt. Det finns betydligt bättre metoder som till exempel Kolmogorov-Smirnovs metod som istället baserar sig på den empiriska fördelningsfunktionen. Test av denna typ ger bättre resultat i allmänhet.

## 6 Homogenitetstest

Det kan ofta vara intressant att avgöra om egenskaper skiljer sig åt mellan olika grupper. Låt oss sätta upp följande scenario. Vi har  $s$  stycken grupper eller serier av försök som vi är nyfikna på om de uppvisar samma sorts fördelning med avseende på en mängd egenskaper  $A_1, A_2, \dots, A_r$ . Vi kan då ställa upp datan enligt följande där siffrorna är absoluta frekvenser.

	Egenskap 1	Egenskap 2	...	Egenskap $r$	Summa
Grupp 1	$N_{11}$	$N_{12}$	...	$N_{1r}$	$G_1$
Grupp 2	$N_{21}$	$N_{22}$	...	$N_{2r}$	$G_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Grupp $s$	$N_{s1}$	$N_{s2}$	...	$N_{sr}$	$G_s$
Summa	$E_1$	$E_2$	...	$E_r$	$N$

Om vi antar att grupperna är *homogena*, dvs att de uppvisar samma fördelning för egenskaperna, så är en bra skattning för sannolikheten  $p_j$  att ett objekt har egenskap  $j$  helt enkelt

$$\hat{p}_j = E_j/N.$$

Vi formar samma sorts teststorhet som vi gjort innan

$$Q = \sum_{i=1}^s \sum_{j=1}^r \frac{(N_{ij} - G_i \cdot \hat{P}_j)^2}{G_i \hat{P}_j} \stackrel{\text{appr.}}{\approx} \chi^2((r-1)(s-1)).$$

Att det blir just  $(r-1)(s-1)$  kommer från de linjära restriktioner som trillar ut ur tabellen ovan. Vi kan se att

$$\sum_{j=1}^r \frac{(N_{ij} - G_i \cdot p_j)^2}{G_i p_j} \stackrel{\text{appr.}}{\approx} \chi^2(r-1)$$

enligt tidigare argument, men då antar vi att  $p_j$  är kända. Sedan summerar vi  $s$  sådana oberoende variabler, så resultatet blir  $\chi^2(s(r-1))$ -fördelat. Men nu vill vi skatta  $p_j$  och för varje skattning tappar vi en frihetsgrad. Märk dock att vi inte skattar alla  $r$  stycken  $p_j$ , utan bara  $r-1$  stycken då den sista ges av att summan måste bli ett. Vi tappar alltså  $r-1$  frihetsgrader. Totalt sett har vi alltså  $s(r-1) - (r-1) = (s-1)(r-1)$  frihetsgrader.

Så när gäller approximationen? Den gäller under förutsättning att  $n_i \hat{p}_j$  är stora. En rimlig tumregel är att  $n_i \hat{p}_j \geq 5$ .



### Exempel

Ifrån en stor population frågar vi två grupper om de tycker det borde vara lagligt att kasta tallkottar på hundägare som inte håller sina hundar kopplade.

Grupp	Kottkastning är OK!	Nej man får inte kasta tallkottar på folk.
$G_1$	59	41
$G_2$	145	55

Testa på signifikansnivån 1% (approximativt) om det finns någon skillnad mellan vad grupperna tycker.

**Lösning.** Vi utför ett homogenitetstest.

Grupp	Kottkastning är OK!	Nej man får inte kasta tallkottar på folk.	Summa
$G_1$	59	41	100
$G_2$	145	55	200
$G_1 + G_2$	204	96	300
$\hat{p}_j$	0.68	0.32	1.0

Låt  $H_0$  : Grupperna tycker likadant mot  $H_1$  : Grupperna tycker olika. Vår observation av teststorheten ges av

$$q = \frac{(59 - 68)^2}{68} + \frac{(41 - 32)^2}{32} + \frac{(145 - 136)^2}{136} + \frac{(55 - 64)^2}{64} \approx 5.58.$$

Detta är en observation av  $Q \stackrel{\text{appr.}}{\approx} \chi^2(1 \cdot 1)$ . Ur tabell finner vi att  $P(Q > 6.6349) = 0.01$ . Eftersom  $q < 6.6349$  så kan vi inte förkasta hypotesen att grupperna tycker lika.



## Exempel

Alla som lyssnar på hårdrock i någon form har säkert funderat över vilken av Slayer-låtarna *Angel of Death* och *Raining Blood* som är bäst<sup>a</sup>. Examinator funderade över om resultaten är homogena över några olika grupper och samlade in följande siffror på internet:

	<i>Angel of Death</i>	<i>Raining Blood</i>
Returntothepit.com	199	173
MetalStorm.net	47	43
RockBand.com	21	16
MetalRules.com	23	3

Utför ett homogenitetstest på nivån 5% för att se om man kan förkasta hypotesen att åsikterna är likafördelade i de fyra olika grupperna.

<sup>a</sup>Självklart är *Angel of Death* den bästa av dessa två, men det är inte poängen!

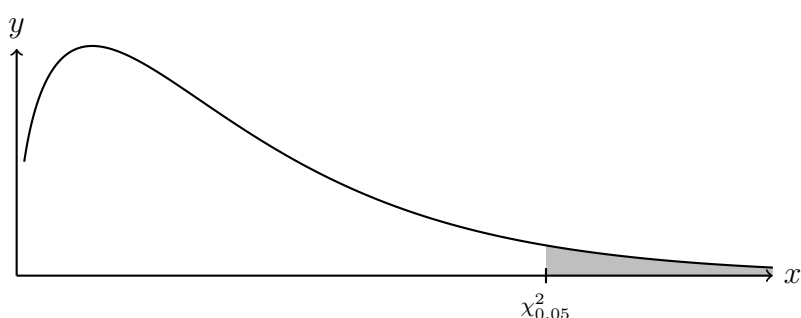
**Lösning.** Först kompletterar vi tabellen med all information som behövs:

	<i>Angel of Death</i>	<i>Raining Blood</i>	Summa ( $n_i$ )
Returntothepit.com	199	173	372
MetalStorm.net	47	43	90
RockBand.com	21	16	37
MetalRules.com	23	3	26
Summa	290	235	525
$\hat{p}_j$ (skattat $p_j$ )	$\hat{p}_1 = 0.552$	$\hat{p}_2 = 0.448$	1.00

Vi beräknar observationen  $q$

$$\begin{aligned}
 q &= \sum_{i=1}^4 \sum_{j=1}^2 \frac{(x_{ij} - n_i \hat{p}_j)^2}{n_i \hat{p}_j} \\
 &= \frac{(199 - 372 \hat{p}_1)^2}{372 \hat{p}_1} + \frac{(173 - 372 \hat{p}_2)^2}{372 \hat{p}_2} + \frac{(47 - 90 \hat{p}_1)^2}{90 \hat{p}_1} + \frac{(43 - 90 \hat{p}_2)^2}{90 \hat{p}_2} \\
 &\quad + \frac{(21 - 37 \hat{p}_1)^2}{37 \hat{p}_1} + \frac{(16 - 37 \hat{p}_2)^2}{37 \hat{p}_2} + \frac{(23 - 26 \hat{p}_1)^2}{26 \hat{p}_1} + \frac{(3 - 26 \hat{p}_2)^2}{26 \hat{p}_2} \\
 &= 12.43
 \end{aligned}$$

Låt  $H_0$  vara utsagan att favoriten bland de två låtarna är likadant fördelad i alla fyra serier. Det vill säga, att  $P(\text{AoD favorit}) = p_1$  och  $P(\text{RB favorit}) = p_2$  gäller i alla fyra serierna med samma sannolikheter  $p_j$ . Antag att  $H_0$  är sann.



Vi förkastar  $H_0$  om  $Q > \chi_\alpha^2(3)$ , d v s om den observerade testvariabeln hamnar utanför det skuggade området i figuren ovan. Med  $\alpha = 0.05$  finner vi att  $\chi_{0.05}^2(3) = 7.81$  (ur en tabell eller med matlab), så  $Q > \chi_\alpha^2(3)$ . Vi kan alltså förkasta hypotesen att alla grupperna tycker likadant (ganska tydligt från siffrorna att den fjärde raden skiljer sig markant från de andra).

**Svar:** Vi kan förkasta hypotesen om homogenitet på nivån 5%.