χ^2-fordelte variable

Definition af χ^2-fordelingen

Kvadratsummen V_n af n uafhængige standardiserede normalfordelte stokastiske variable siges at være χ^2-fordelt med n frihedsgrader. V_n fremkommer altså som

$$V_n = X_1^2 + X_2^2 + \cdots + X_n^2,$$

hvor $X_i \sim N(0,1)$, $i = 1, \ldots, n$, X_i'erne uafhængige. Vi skriver kort $V_n \sim \chi^2(n)$.

Bestemmelse af $V_1 = X_i^2 \sim \chi^2(1)$

Tæthedsfunktionen for X_i^2, hvor $X_i \sim N(0,1)$, bliver

$$f_{X_i^2}(x) = 2\frac{1}{\sqrt{2\pi}}e^{-\frac{(\sqrt{x})^2}{2}}\left|\frac{1}{2\sqrt{x}}\right| = \frac{1}{\sqrt{2\pi x}}e^{-\frac{1}{2}x}, \qquad 0 < x < \infty.$$

Bemærk, at $f_{X_i^2}(x)$ kan skrives

$$f_{X_i^2}(x) = \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)}x^{\frac{1}{2}-1}e^{-\frac{1}{2}x}, \qquad 0 < x < \infty,$$

som er tæthedsfunktion for en gammafordelt variabel med formparameter $\frac{1}{2}$ og skalaparameter 2, dvs.

$$\chi^2(1) = \Gamma\left(\tfrac{1}{2}, \tfrac{1}{2}\right).$$

Øvelse. Bestem først fordelingsfunktionen for X_i^2 (facit: $2\Phi(\sqrt{x}) - 1$), og dernæst tæthedsfunktionen for X_i^2 ved differentiation. □

Gammafordelingens foldningsegenskab

Inspireret af ovenstående resultat vil vi se på foldningen af to gammafordelte variable med ens intensitet λ (ækvivalent med ens skalaparameter $\frac{1}{\lambda}$).

Lad $X \sim \Gamma(\alpha, \lambda)$ og $Y \sim \Gamma(\beta, \lambda)$, X og Y uafhængige. Tæthedsfunktionerne er

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, \qquad 0 < x < \infty,$$

$$f_Y(y) = \frac{\lambda^\beta}{\Gamma(\beta)}x^{\beta-1}e^{-\lambda x}, \qquad 0 < y < \infty,$$

og dermed

$$\begin{aligned}
f_{X+Y}(z) &= f_X * f_Y(z) \\
&= \int_0^z \frac{\lambda^\alpha}{\Gamma(\alpha)}(z-u)^{\alpha-1}e^{-\lambda(z-u)}\frac{\lambda^\beta}{\Gamma(\beta)}u^{\beta-1}e^{-\lambda u}du \\
&= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)}e^{-\lambda z}\int_0^z (z-u)^{\alpha-1}u^{\beta-1}du, \qquad \begin{matrix} u = zw \\ du = zdw \end{matrix} \\
&= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)}e^{-\lambda z}\int_0^1 z^{\alpha+\beta-2}(1-w)^{1-\alpha}w^{1-\beta}zdw \\
&= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)}z^{\alpha+\beta-1}e^{-\lambda z}\int_0^1 (1-w)^{1-\alpha}w^{1-\beta}dw \\
&= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)}z^{\alpha+\beta-1}e^{-\lambda z}B(\alpha,\beta),
\end{aligned}$$

jf. baggrundsnote til sandsynlighedsregning side 7 linie 3 fra neden. Idet der gælder, at $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ [1], får vi

$$f_{X+Y}(z) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha+\beta)}z^{\alpha+\beta-1}e^{-\lambda z}, \qquad 0 < z < \infty,$$

dvs. $X + Y \sim \Gamma(\alpha+\beta, \lambda)$.

Bestemmelse af $V_n \sim \chi^2(n)$

Ved at benytte af gammafordelingens foldningsegenskab får vi umiddelbart for $i \neq j$

$$V_2 = X_i^2 + X_j^2 \;\sim\; \Gamma(\tfrac{1}{2}+\tfrac{1}{2}, \tfrac{1}{2}) = \Gamma(1, \tfrac{1}{2}),$$

og ved fortsat anvendelse

$$V_n = X_1^2 + X_2^2 + \cdots + X_n^2 \;\sim\; \Gamma(\tfrac{1}{2}+\tfrac{1}{2}+\cdots+\tfrac{1}{2}, \tfrac{1}{2}) = \Gamma(\tfrac{n}{2}, \tfrac{1}{2}).$$

Der gælder altså

$$\chi^2(n) = \Gamma(\tfrac{n}{2}, \tfrac{1}{2}).$$

Herefter kan vi let udregne middelværdi og varians i χ^2-fordelingen:

$$\mathrm{E}\,V_n = \frac{\frac{n}{2}}{\frac{1}{2}} = n \qquad\qquad \mathrm{Var}\,V_n = \frac{\frac{n}{2}}{\left(\frac{1}{2}\right)^2} = 2n$$

---

[1] $\Gamma(\alpha)\Gamma(\beta) = \int_0^\infty s^{\alpha-1}e^{-s}ds\int_0^\infty t^{\beta-1}e^{-t}dt = \int_0^\infty\int_0^\infty s^{\alpha-1}t^{\beta-1}e^{-(s+t)}dsdt$

$$\begin{matrix} x = s+t & s = xy \\ y = \dfrac{s}{s+t} & t = x(1-y) \end{matrix} \qquad J(x,y) = \begin{vmatrix} y & x \\ 1-y & -x \end{vmatrix} = -xy - x + xy = -x$$

$$= \int_0^\infty\int_0^1 (xy)^{\alpha-1}(x(1-y))^{\beta-1}e^{-x}|-x|\,dydx = \int_0^\infty x^{\alpha+\beta-1}e^{-x}dx\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy$$

$$= \Gamma(\alpha+\beta)B(\alpha,\beta) \quad\Rightarrow\quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Tæthedsfunktionen for V_n bliver

$$f_{V_n}(x) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x} = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x}, \qquad 0 < x < \infty.$$

Øvelse. Eftervis, at χ^2-fordelingen med 2 frihedsgrader er identisk med eksponentialfordelingen med intensitet $\lambda = \frac{1}{2}$. □

Sum af uafhængige χ^2-fordelte variable

Lad $V_m \sim \chi^2(m)$ og $V_n \sim \chi^2(n)$, V_m og V_n uafhængige. Ved igen at benytte af gammafordelingens foldningsegenskab får vi

$$V_m + V_n \sim \Gamma(\tfrac{m}{2} + \tfrac{n}{2}, \tfrac{1}{2}) = \Gamma(\tfrac{m+n}{2}, \tfrac{1}{2}) = \chi^2(m+n),$$

som umiddelbart kan generaliseres til en sum af flere uafhængige χ^2-fordelinger.

Modifikationer

Med $\boldsymbol{X} \sim N_n(\boldsymbol{0}, \sigma^2 I)$ er $\frac{X_i}{\sigma} \sim N(0,1)$, $i = 1, \ldots, n$, uafhængige, og dermed $\frac{1}{\sigma^2}\sum X_i^2 \sim \chi^2(n)$. Dette skrives ofte $\sum X_i^2 \sim \sigma^2\chi^2(n)$.

Med $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, I)$ bliver $\sum X_i^2$ såkaldt ikke-centralt χ^2-fordelt med n frihedsgrader og ikke-centralitetsparameter $\boldsymbol{\mu}^\top\boldsymbol{\mu} = \sum \mu_i^2$. Dette skrives $\sum X_i^2 \sim \chi^2(n; \boldsymbol{\mu}^\top\boldsymbol{\mu})$.

Øvelse. Vis, at middelværdien af $\sum X_i^2$ afhænger af $\|\boldsymbol{\mu}\|$ og ikke af de enkelte μ_i'er. □

Med $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \sigma^2 I)$ bliver $\sum X_i^2 \sim \sigma^2\chi^2(n; \frac{\boldsymbol{\mu}^\top\boldsymbol{\mu}}{\sigma^2})$.

For mere om den ikke-centrale χ^2-fordeling, se fx `http://en.wikipedia.org/wiki/Noncentral_chi-square_distribution`.

En stokastisk kvadratisk form

Betragt en stokastisk kvadratisk form, hvor $\boldsymbol{X} \sim N_n(\boldsymbol{0}, I)$, og A er en $n \times n$ symmetrisk og idempotent matrix[2)] med $\operatorname{rang} A = q \leq n$. Vi vil vise, at $\boldsymbol{X}^\top A\boldsymbol{X} \sim \chi^2(q)$.

Når A er symmetrisk, findes en ortogonal matrix C, som diagonaliserer A, dvs. $C^\top AC = \Lambda = \lceil\lambda_1\ \lambda_2\ \cdots\ \lambda_n\rfloor$, hvor λ_i'erne er A's egenværdier.

Med A idempotent er også Λ idempotent. Det gælder derfor om alle A's egenværdier, at $\lambda^2 = \lambda$, dvs. alle egenværdierne er enten 1 eller 0. Af $\operatorname{rang} A = \operatorname{rang} \Lambda = \operatorname{tr} \Lambda$ får vi $\operatorname{tr} \Lambda = q$, dvs. A har netop q egenværdier med værdien 1 og $n - q$ egenværdier med værdien 0.

Den ortogonale transformation $\boldsymbol{Y} = C^\top\boldsymbol{X}$ har samme fordeling som $\boldsymbol{X}$, idet

$$\boldsymbol{Y} = C^\top\boldsymbol{X} \sim N_n(C^\top\boldsymbol{0}, C^\top IC) = N_n(\boldsymbol{0}, I),$$

[2)] En symmetrisk og idempotent matrix kaldes også en projektionsmatrix.

hvorefter

$$\boldsymbol{X}^\top A\boldsymbol{X} = \boldsymbol{X}^\top C\Lambda C^\top \boldsymbol{X} = (C^\top \boldsymbol{X})^\top \Lambda C^\top \boldsymbol{X} = \boldsymbol{Y}^\top \Lambda \boldsymbol{Y} = \sum_{i:\lambda_i=1} Y_i^2$$

viser, at $\boldsymbol{X}^\top \Lambda \boldsymbol{X} \sim \chi^2(q)$.

Øvelse. Vis, at hvis $\boldsymbol{X} \sim N_n(\boldsymbol{0}, \sigma^2 I)$, så er $\boldsymbol{X}^\top \Lambda \boldsymbol{X} \sim \sigma^2\chi^2(q)$, og hvis $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \sigma^2 I)$, så er $\boldsymbol{X}^\top \Lambda \boldsymbol{X} \sim \sigma^2\chi^2(q; \frac{\boldsymbol{\mu}^\top\boldsymbol{\mu}}{\sigma^2})$. □

Antager vi omvendt, at $\boldsymbol{X}^\top A\boldsymbol{X} \sim \chi^2(q)$ vil vi vise, at A er idempotent med rang $A = q$. Da A er symmetrisk, kan vi diagonalisere A til $\Lambda = \lceil \lambda_1\ \lambda_2\ \cdots\ \lambda_n \rfloor$ og ved tilsvarende regninger som ovenfor får vi

$$\boldsymbol{X}^\top A\boldsymbol{X} = \boldsymbol{Y}^\top \Lambda \boldsymbol{Y} = \sum_{i=1}^n \lambda_i Y_i^2,$$

hvor $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, I)$.

For at $\sum_{i=1}^n \lambda_i Y_i^2$ kan modsvare kvadratsummen af q uafhængige standardiserede normalfordelte variable må der gælde, at q af λ_i'erne er 1 og $n-q$ er 0. Heraf følger, at A er idempotent med rang $A = \operatorname{rang}\Lambda = \operatorname{tr}\Lambda = q$.

For helt at udelukke at en linearkombination $\sum_{i=1}^n \lambda_i X_i^2$ med andre koefficienter end 0 og 1 kan være $\chi^2(q)$-fordelt, betragter vi momentfrembringende funktioner[3)]. Et krav om $\sum_{i=1}^n \lambda_i X_i^2 \sim \chi^2(q)$ udmøntes til

$$\prod_{i=1}^n (1-2\lambda_i t)^{-\frac{1}{2}} = (1-2t)^{-\frac{q}{2}} \quad \Rightarrow \quad \prod_{i=1}^n (1-2\lambda_i t) = (1-2t)^q,$$

men dette viser igen, at λ_i'erne må vælges som ovenfor anført.

Uafhængighed mellem stokastiske kvadratiske former

Betragt to stokastiske kvadratiske former $\boldsymbol{X}^\top A_1\boldsymbol{X}$ og $\boldsymbol{X}^\top A_2\boldsymbol{X}$, hvor $\boldsymbol{X} \sim N_n(\boldsymbol{0}, I)$ og hvor både A_1 og A_2 er $n\times n$ symmetriske og idempotente matricer med rang hhv. q_1 og q_2. Vi vil vise, at

$$\boldsymbol{X}^\top A_1\boldsymbol{X} \text{ og } \boldsymbol{X}^\top A_2\boldsymbol{X} \text{ uafhængige} \quad \Leftrightarrow \quad A_1A_2 = O.$$

Antag først, at $\boldsymbol{X}^\top A_1\boldsymbol{X}$ og $\boldsymbol{X}^\top A_2\boldsymbol{X}$ er uafhængige. Summen af to uafhængige χ^2-fordelte variable er igen χ^2-fordelt, dvs. $\boldsymbol{X}^\top(A_1+A_2)\boldsymbol{X} \sim \chi^2(q_1+q_2)$. Heraf følger, at A_1+A_2 er idempotent, dvs.

$$(A_1+A_2)(A_1+A_2) = A_1 + A_1A_2 + A_2A_1 + A_2 = A_1 + A_2,$$

hvilket kræver, at $A_1A_2 + A_2A_1 = O$. Multiplicerer vi denne ligning med A_1 hhv. forfra og bagfra får vi

$$A_1A_2 + A_1A_2A_1 = O \quad \text{og} \quad A_1A_2A_1 + A_2A_1 = O.$$

3) $M_{V_1}(t) = \mathrm{E}[e^{tV_1}] = \int_{-\infty}^{\infty} e^{tx^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx = (1-2t)^{-\frac{1}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(1-2t)x^2}{2}}\,d\big((1-2t)^{\frac{1}{2}}x\big)$
$= (1-2t)^{-\frac{1}{2}} \quad \Rightarrow \quad M_{V_n}(t) = (1-2t)^{-\frac{n}{2}}$

Subtraktion giver $A_1A_2 - A_2A_1 = O$, som sammenholdt med $A_1A_2 + A_2A_1 = O$ viser, at $A_1A_2 = O$.

Omvendt, når $A_1A_2 = O$, bliver $A_1\boldsymbol{X}$ og $A_2\boldsymbol{X}$ er uafhængige, idet

$$\mathrm{Cov}(A_1\boldsymbol{X}, A_2\boldsymbol{X}) = A_1 \mathrm{Var}(\boldsymbol{X})A_2^\top = A_1IA_2 = A_1A_2 = O.$$

Uafhængigheden mellem $A_1\boldsymbol{X}$ og $A_2\boldsymbol{X}$ medfører uafhængighed mellem $(A_1\boldsymbol{X})^\top A_1\boldsymbol{X} = \boldsymbol{X}^\top A_1\boldsymbol{X}$ og $(A_2\boldsymbol{X})^\top A_2\boldsymbol{X} = \boldsymbol{X}^\top A_2\boldsymbol{X}$.

Dekomposition af en χ^2-fordelt stokastisk variabel

Vi vil undersøge, om det er muligt at dekomponere $V_n = \boldsymbol{X}^\top\boldsymbol{X} \sim \chi^2(n)$, $\boldsymbol{X} \sim N_n(\boldsymbol{0}, I)$, i en sum af parvis uafhængige χ^2-fordelte variable.

Antag, at $\boldsymbol{X}^\top\boldsymbol{X}$ kan skrives som en sum af stokastiske kvadratiske former,

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{X}^\top A_1\boldsymbol{X} + \boldsymbol{X}^\top A_2\boldsymbol{X} + \cdots + \boldsymbol{X}^\top A_k\boldsymbol{X},$$

hvor A_i er symmetrisk og idempotent med rang $A_i = q_i$, $i = 1, \ldots, k$.

Bemærk, at

$$\boldsymbol{X}^\top A_1\boldsymbol{X} + \boldsymbol{X}^\top A_2\boldsymbol{X} + \cdots + \boldsymbol{X}^\top A_k\boldsymbol{X} = \boldsymbol{X}^\top(A_1 + A_2 + \cdots + A_k)\boldsymbol{X},$$

dvs. $A_1 + A_2 + \cdots + A_k = I$.

Der gælder umiddelbart, at $\boldsymbol{X}^\top A_i\boldsymbol{X} \sim \chi^2(q_i)$, $i = 1, \ldots, k$, men spørgsmålet er uafhængighed skal også afklares.

Lad A_j repræsentere en vilkårlig af de kvadratiske former. Da A_j er symmetrisk og idempotent findes en ortogonal matrix C, som diagonaliserer A_j til Λ med alle diagonalelementer 1 eller 0.

Multiplicerer vi på begge sider af $A_1 + A_2 + \cdots + A_k = I$ forfra med $C^\top$ og bagfra med C, får vi

$$\sum_{i\neq j} C^\top A_iC + \Lambda = I.$$

Matricerne $C^\top A_iC$, $i \neq j$, er som A_i'erne idempotente. Alle diagonalelementer er derfor ikke-negative. Dette betyder, at hvor Λ har 1 i diagonalen, er de tilsvarende diagonalelementer i $C^\top A_iC$, $i \neq j$, nødvendigvis alle 0.

Endvidere gælder, at når der i $C^\top A_iC$ optræder 0 i hoveddiagonalen, så vil alle elementerne i den aktuelle række og i den aktuelle søjle også have værdien 0 [4)].

Multiplikation af $C^\top A_iC$, $i \neq j$ med Λ giver derfor O. Ved at indføre $\Lambda = C^\top A_jC$ får vi

$$C^\top A_iC\Lambda = C^\top A_iCC^\top A_jC = C^\top A_iA_jC = O,$$

hvoraf $A_iA_j = O$, $i \neq j$.

[4)] Lad C_i diagonalisere $C^\top A_iC$ til Λ_i. Så er $C^\top A_iC = C_i\Lambda_iC_i^\top = C_i\Lambda_i^{\frac{1}{2}}\Lambda_i^{\frac{1}{2}}C_i^\top = C_i\Lambda_i^{\frac{1}{2}}(C_i\Lambda_i^{\frac{1}{2}})^\top$. Et 0 i hoveddiagonalen i $C^\top A_iC$ kræver, at den tilsvarende række i $C_iA^{\frac{1}{2}}$ er en nulrække, hvoraf påstanden følger.

Da A_j var arbitrært valgt, har vi $A_iA_j = O$ for alle $i,j = 1,\dots,k$, $i \neq j$, hvilket viser, at $\boldsymbol{X}^\top A_i \boldsymbol{X}$, $i = 1,\dots,k$, er parvis uafhængige. Heraf følger, at

$$\boldsymbol{X}^\top A_1 \boldsymbol{X} + \boldsymbol{X}^\top A_2 \boldsymbol{X} + \cdots + \boldsymbol{X}^\top A_k \boldsymbol{X} \sim \chi^2\big(\textstyle\sum_{i=1}^k q_i\big),$$

som sammenholdt med $\boldsymbol{X}^\top \boldsymbol{X} \sim \chi^2(n)$ viser, at $\sum_{i=1}^k q_i = n$ er en nødvendig og tilstrækkelig betingelse for gyldigheden af den betragtede dekomposition.

Mere udførligt kan vi slutte, at følgende udsagn er ækvivalente:

- $\boldsymbol{X}^\top A_i \boldsymbol{X} \sim \chi^2(q_i)$, $i = 1,\dots,k$, parvis uafhængige.
- $A_iA_j = O$, $i,j = 1,\dots,k$, $i \neq j$.
- $\sum_{i=1}^k q_i = n$.

Dette resultat, som findes i flere varianter, kaldes Cochrans sætning.

Øvelse. Lad $\boldsymbol{X} \sim N_n(\boldsymbol{0}, I)$. Betragt $V_n = \boldsymbol{X}^\top \boldsymbol{X}$ og $V_q = \boldsymbol{X}^\top A \boldsymbol{X}$, A symmetrisk og idempotent, rang $A = q$. Sæt $U = V_n - V_q$, og vis, at $U \sim \chi^2(n-q)$, samt at V_q og U er uafhængige. □

Hvis i stedet $\boldsymbol{X} \sim N_n(\boldsymbol{0}, \sigma^2 I)$, kan udledningen opretholdes, når blot $\chi^2(\cdot)$ alle steder erstattes af $\sigma^2\chi^2(\cdot)$. Hvis $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \sigma^2 I)$, kan udledningen stadig opretholdes, når yderligere $\boldsymbol{X}$ erstattes af $\boldsymbol{X} - \boldsymbol{\mu}$.

Øvelse. Betragt en normalfordelt stikprøve $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \sigma I)$, $\boldsymbol{\mu} = \mu\boldsymbol{1} = \mu(1,1,\dots,1)$. Vis, at $|\overline{X} - \mu|$ og S^2 er uafhængige. Vink: Der gælder

$$\sum_{i=1}^n (X_i - \mu)^2 = (n-1)S^2 + n(\overline{X} - \mu)^2$$

samt

$$(n-1)S^2 = \sum_{i=1}^n (X_i - \overline{X})^2 = \sigma^2(\boldsymbol{X} - \mu\boldsymbol{1})^\top (I - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top)(\boldsymbol{X} - \mu\boldsymbol{1})$$

og

$$(\overline{X} - \mu)^2 = \sigma^2(\boldsymbol{X} - \mu\boldsymbol{1})^\top (\frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top)(\boldsymbol{X} - \mu\boldsymbol{1}).$$

(Vi kan ikke ad denne vej nå helt frem til, at $\overline{X}$ og S^2 er uafhængige.) □

26.1.2012/BR