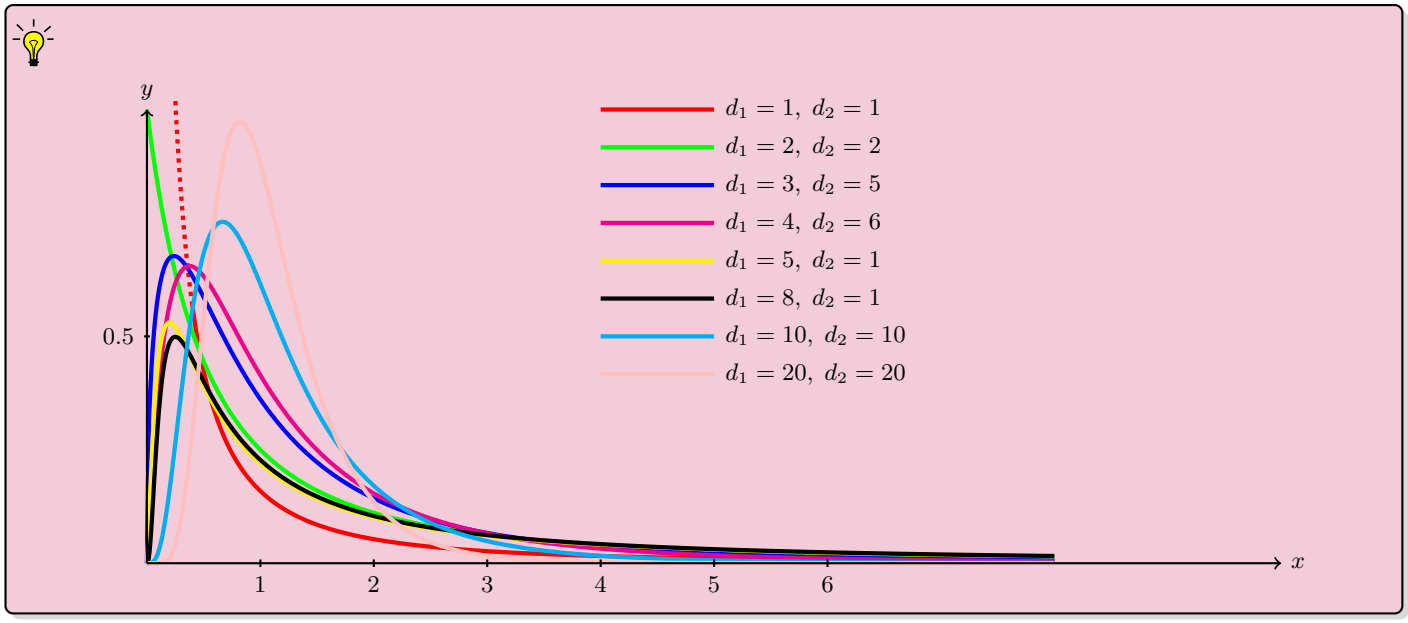
TAMS24: Statistisk inferens ht 2018

Föreläsningsanteckningar

Johan Thim, MAI



Föreläsning 1: Repetition och punktskattningar

Johan Thim (johan.thim@liu.se)

29 augusti 2018

1 Repetition

Vi repeterar lite kort de viktigaste begreppen från sannolikhetsläran. Materialet kan återfinnas i mer fullständig form (med exempel och mer diskussion) i föreläsningsanteckningarna för kursen TAMS79.

1.1 Sannolikhetsteorins grunder

Ett **slumpförsök** är ett försök där resultatet ej kan förutsägas deterministiskt. Slumpförsöket har olika möjliga **utfall**. Vi låter **Utfallsrummet** Ω vara mängden av alla möjliga utfall. En **händelse** är en delmängd av Ω, dvs en mängd av utfall. Men, alla möjliga delmängder av Ω behöver inte vara *tillåtna* händelser. För att precisera detta kräver vi att mängden av alla händelser (detta är alltså en mängd av mängder) är en så kallad σ-algebra.

σ-algebra

Definition. $\mathcal{F}$ är en σ-algebra på Ω om $\mathcal{F}$ består av delmängder av Ω så att

(i) $\Omega \in \mathcal{F}$.

(ii) om $A \in \mathcal{F}$ så kommer komplementet $A^* \in \mathcal{F}$.

(iii) om $A_1, A_2, \ldots \in \mathcal{F}$ så är unionen $A_1 \cup A_2 \cup \cdots \in \mathcal{F}$.

Det enklaste exemplet på en σ-algebra är $\mathcal{F} = \{\Omega, \emptyset\}$, dvs endast hela utfallsrummet och den tomma mängden. Av förklarliga skäl kommer vi inte så långt med detta. Ett annat vanligt exempel är att $\mathcal{F}$ består av *alla* möjliga delmängder till Ω; skrivs ibland $\mathcal{F} = 2^\Omega$, och kallas potensmängden av Ω. Denna konstruktion är lämplig när vi har diskreta utfall. Om Ω består av ett kontinuum så visar det sig dock att 2^Ω blir alldeles för stor för många tillämpningar.

Kolmogorovs Axiom: Sannolikhetsmått

Definition. Ett **sannolikhetsmått** på en σ-algebra $\mathcal{F}$ över ett utfallsrum Ω tilldelar ett tal mellan noll och ett, en sannolikhet, för varje händelse som är definierad (dvs tillhör $\mathcal{F}$). Formellt är P en mängdfunktion; $P\colon \mathcal{F} \to [0,1]$. Sannolikhetsmåttet P *måste* uppfylla Kolmogorovs axiom:

(i) $0 \leq P(A) \leq 1$ för varje $A \in \mathcal{F}$.

(ii) $P(\Omega) = 1$.

(iii) Om $A \cap B = 0$ så gäller att $P(A \cup B) = P(A) + P(B)$.

Oberoende

Definition. Två händelser A och B kallas **oberoende** om $P(A \cap B) = P(A)P(B)$.

För att kunna precisera vad för slags funktion (för det är en funktion) en stokastisk variabel är, behöver vi diskutera öppna mängder på den reella axeln $\mathbf{R}$.

Definition. Den minsta (minst antal element) σ-algebran på $\mathbf{R}$ som innehåller *alla* öppna intervall betecknar vi med $\mathscr{B}$. Denna algebra brukar kallas för **Borel-σ-algebran** på $\mathbf{R}$.

Algebran $\mathscr{B}$ innehåller alltså alla mängder av typen $(a, b) \subset \mathbf{R}$, $(-\infty, c) \cup (d, \infty) \subset \mathbf{R}$, komplement av sådana mängder, samt alla uppräkneliga unioner av mängder av föregående typ. Detta är ganska tekniskt, och inget vi kommer att arbeta med direkt. Men för att få en korrekt definition behövs begreppet.

Stokastisk variabel

Definition. En **stokastisk variabel** är en reellvärd funktion definierad på ett utfallsrum Ω. Funktionen X avbildar alltså olika utfall på reella tal; $X : \Omega \to \mathbf{R}$.
Mer precist så kräver vi att $X^{-1}(B) \in \mathcal{F}$ för alla $B \in \mathscr{B}$. Mängden $X^{-1}(B)$ definieras som $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ och kallas för **urbilden** av B. Mängden består alltså av alla $\omega \in \Omega$ som avbildas in i B. **Bilden** av en delmängd A av Ω betecknas med $X(A)$, och

$$X(A) = \{x \in \mathbf{R} : X(\omega) = x \text{ för något } \omega \in A\}.$$

Mängden $X(A)$ är alltså värdemängden för X på mängden A.
Om $X(\Omega)$ är ändlig, eller bara har uppräkneligt många värden, så kallar vi X för en **diskret stokastisk variabel**. Annars kallas vi X för **kontinuerlig**.

Uttrycket $X(\omega)$ är alltså det siffervärde vi sätter på ett visst utfall $\omega \in \Omega$. Varför kravet att urbilden $X^{-1}(B)$ skall tillhöra de tillåtna händelserna? Det faller sig ganska naturligt, då $X^{-1}(B)$ är precis de utfall i Ω som avbildas in i mängden B. Således vill vi gärna att denna samling utfall verkligen utgör en händelse, annars kan vi inte prata om någon sannolikhet för denna samling utfall.
För variabler i högre dimension fokuserar vi på två-dimensionella variabler. Det är steget från en dimension till två som är det svåraste. Generaliseringar till högre dimensioner följer utan problem i de flesta fall. I $\mathbf{R}^2$ är Borelfamiljen $\mathscr{B}$ den minsta σ-algebra som innehåller alla öppna rektanglar $(a, b) \times (c, d)$. Generaliserar naturligt till högre dimensioner.

Flerdimensionell stokastisk variabel

Definition. En tvådimensionell stokastisk variabel är en reell-vektorvärd funktion (X, Y) definierad på ett utfallsrum Ω. (X, Y) avbildar alltså olika utfall på reella vektorer; $(X, Y) : \Omega \to \mathbf{R}^2$. Vi kräver att $(X, Y)^{-1}(B) \in \mathcal{F}$ för alla $B \in \mathscr{B}$. Algebran $\mathcal{F}$ är mängden av alla tillåtna händelser. Om (X, Y) bara antar ändligt eller uppräkneligt många värden så kallar vi (X, Y) för en diskret stokastisk variabel. Om varken X eller Y är diskret kallar vi (X, Y) för kontinuerlig.

Definitionen är analog med envariabelfallet. Observera dock följande: en situation som kan uppstå är att vi får ”halvdiskreta” variabler med ena variabeln diskret och den andra kontinuerlig!

Vad måste jag förstå av all matematiska?

- En stokastisk variabel är en *snäll* funktion från Ω till $\mathbf{R}^n$.
- Utfallsrummet Ω kan vara abstrakt, e.g., $\Omega = \{\text{Krona, Klave}\}$.
- Det är mängden $X(\Omega)$ som består av siffror (vektorer av siffror).
- Ibland finns en naturlig koppling mellan Ω och $X(\Omega)$, säg om vi kastar en tärning och räknar antalet ögon vi får.
- En händelse är en *snäll* delmängd av Ω.
- Om A är en händelse så är $X(A)$ *värdemängden* för funktionen X med A som definitionsmängd. Speciellt så är $X(\Omega)$ alla möjliga värden vi kan få från variabeln X.
- Urbilden $X^{-1}(B)$ av en delmängd $B \subset \mathbf{R}^n$ består av *alla* utfall $\omega \in \Omega$ så att siffran (vektorn) $X(\omega)$ ligger i mängden B.

1.2 Beskrivningar av stokastiska variabler

Om $X(\Omega)$ är ändlig eller uppräkneligt oändlig så kallade vi X för diskret. En sådan variabel kan vi karaktärisera med en så kallad **sannolikhetsfunktion**.

Sannolikhetsfunktion

Definition. Sannolikhetsfunktionen $p_X \colon X(\Omega) \to [0,1]$ för en diskret stokastisk variabel definieras av $p_X(k) = P(X = k)$ för alla $k \in X(\Omega)$.

Den vanligaste situationen vi stöter på är att utfallsrummet är numrerat med heltal på något sätt så att p_X är en funktion definierad för (en delmängd av) heltal (när det finns en naturlig koppling mellan Ω och $X(\Omega)$). Ibland är vi slarviga och tänker oss att $p_X(k) = 0$ för siffror k som ej är möjliga ($p_X(-1) = 0$ om X är antal ögon vi ett tärningskast till exempel).
Vissa egenskaper gäller för alla alla sannolikhetsfunktioner:

Egenskaper hos sannolikhetsfunktionen

(i) $p_X(k) \geq 0$ för alla $k \in X(\Omega)$.

(ii) $\displaystyle\sum_{k \in X(\Omega)} p_X(k) = 1.$

(iii) Om $A \subset X(\Omega)$ så är $P(X \in A) = \displaystyle\sum_{k \in A} p_X(k).$

En sannolikhetsfunktion är alltså *aldrig* negativ, om vi summerar över alla möjliga värden (alla $k \in X(\Omega)$) så måste summan bli ett, och om vi är ute efter sannolikheten att få vissa värden på X så summerar vi sannolikheten för vart och ett av dessa värden!

Fördelningsfunktion

Definition. Fördelningsfunktionen $F_X(x)$ för en stokastisk variabel X definieras enligt sambandet $F_X(x) = P(X \leq x)$ för alla $x \in \mathbf{R}$.

Det följer från definitionen att följande påståenden gäller.

Egenskaper hos fördelningsfunktionen

(i) $F_X(x) \to \begin{cases} 0, & x \to -\infty, \\ 1, & x \to +\infty. \end{cases}$

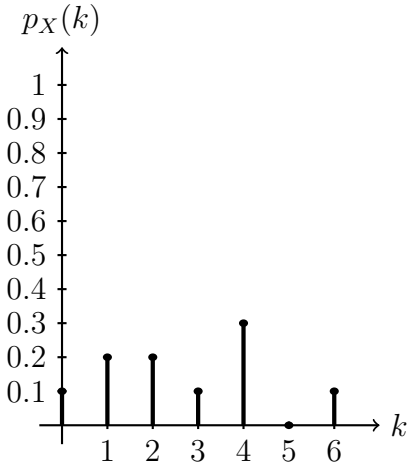
(ii) $F_X(x)$ är icke-avtagande och högerkontinuerlig.

(iii) $F_X(x) = \sum_{\{k \in X(\Omega): k \leq x\}} p_X(k)$.

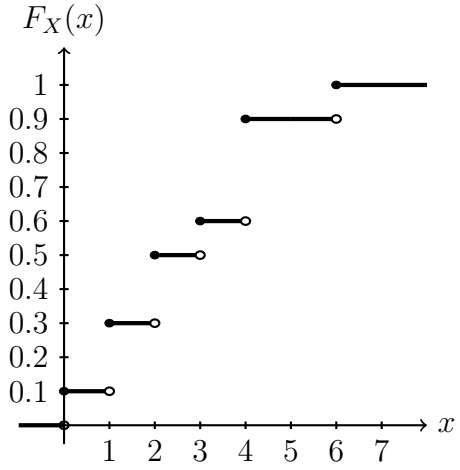
(iv) $P(X > x) = 1 - F_X(x)$.

(v) $F_X(k) - F_X(k-1) = p_X(k)$ för $k \in X(\Omega)$.

Exempel på hur en sannolikhetsfunktion och motsvarande fördelningsfunktion kan se ut:



Sannolikhetsfunktion $p_X(k) = P(X = k)$.



Fördelningsfunktion $F_X(x) = P(X \leq x)$.

1.3 Kontinuerliga stokastiska variabler

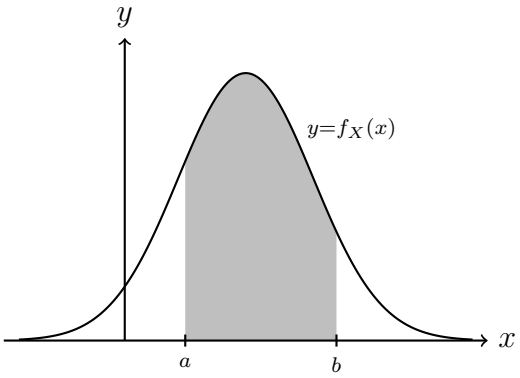
Tähetsfunktion

Definition. Om det finns en icke-negativ *integrerbar* funktion f_X så att

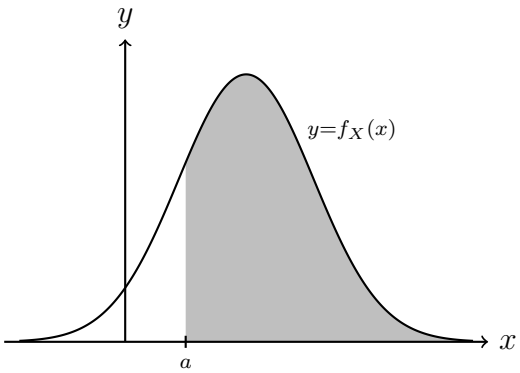
$$P(a < X < b) = \int_a^b f_X(x)\,dx$$

för alla intervall $(a, b) \subset \mathbf{R}$, kallar vi f_X för variabelns **täthetsfunktion**.

Exempel:



Skuggad area: $P(a \leq X \leq b)$.



Skuggad area: $P(X > a) = \int_a^\infty f_X(x)\,dx$.

Egenskaper hos täthetsfunktionen

(i) $f_X(x) \geq 0$ för alla $x \in \mathbf{R}$.

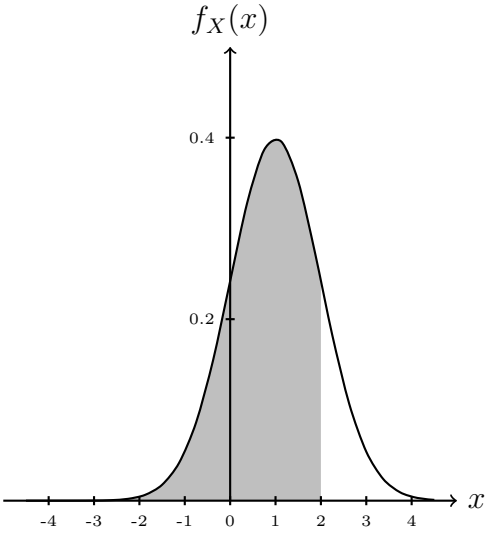
(ii) $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

(iii) $f_X(x)$ anger hur mycket sannolikhetsmassa det finns per längdenhet i punkten x.

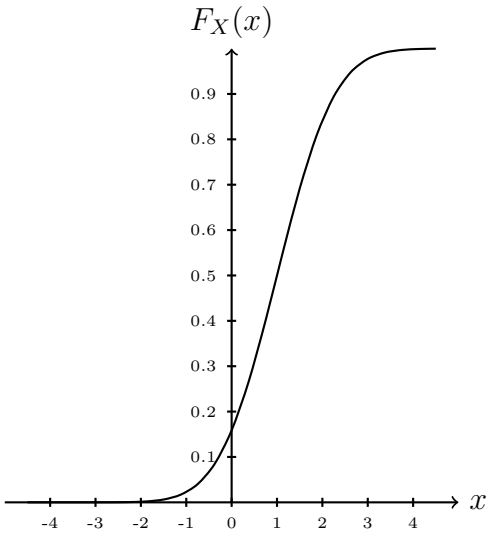
Vi definierar **fördelningsfunktionen** $F_X(x)$ på samma sätt som i det diskreta fallet, och finner att

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt, \qquad x \in \mathbf{R}.$$

Fördelningsfunktionen uppfyller (i)–(iii) från det diskreta fallet, och i alla punkter där $f_X(x)$ är kontinuerlig gäller dessutom att $F_X'(x) = f_X(x)$.
Exempel på hur en täthetsfunktion och motsvarande fördelningsfunktion kan se ut:



Täthet: Hur ”sannolikhetsmassan” är fördelad. Skuggad area är $P(X \leq 2) = F_X(2)$.



Fördelningsfunktionen är växande och gränsvärderna mot $\pm\infty$ verkar stämma!

Väntevärde

Definition. Väntevärdet $E(X)$ av en stokastisk variabel X definieras som

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx \qquad \text{respektive} \qquad E(X) = \sum_k k p_X(k)$$

för kontinuerliga och diskreta variabler.

Andra vanliga beteckningar: μ eller μ_X. Väntevärdet är ett lägesmått som anger vart sannolikhetsmassan har sin tyngdpunkt (jämför med mekanikens beräkningar av tyngdpunkt).

1.4 Högre dimensioner

Sannolikhetsfunktionen för en diskret 2D-variabel ges av $p_{X,Y}(j,k) = P(X = j, Y = k)$.

Egenskaper hos sannolikhetsfunktionen

(i) $p_{X,Y}(j,k) \geq 0$ för alla (j,k).

(ii) $\sum_j \sum_k p_{X,Y}(j,k) = 1$.

(iii) Om $A \subset \mathbf{R}^2$ så är $P(X \in A) = \sum_{(j,k)\in A} \sum p_{X,Y}(j,k)$.

Analogt med en dimension kan man introducera begreppet täthetsfunktion för en kontinuerlig 2D-variabel.

Egenskaper hos den simultana täthetsfunktionen

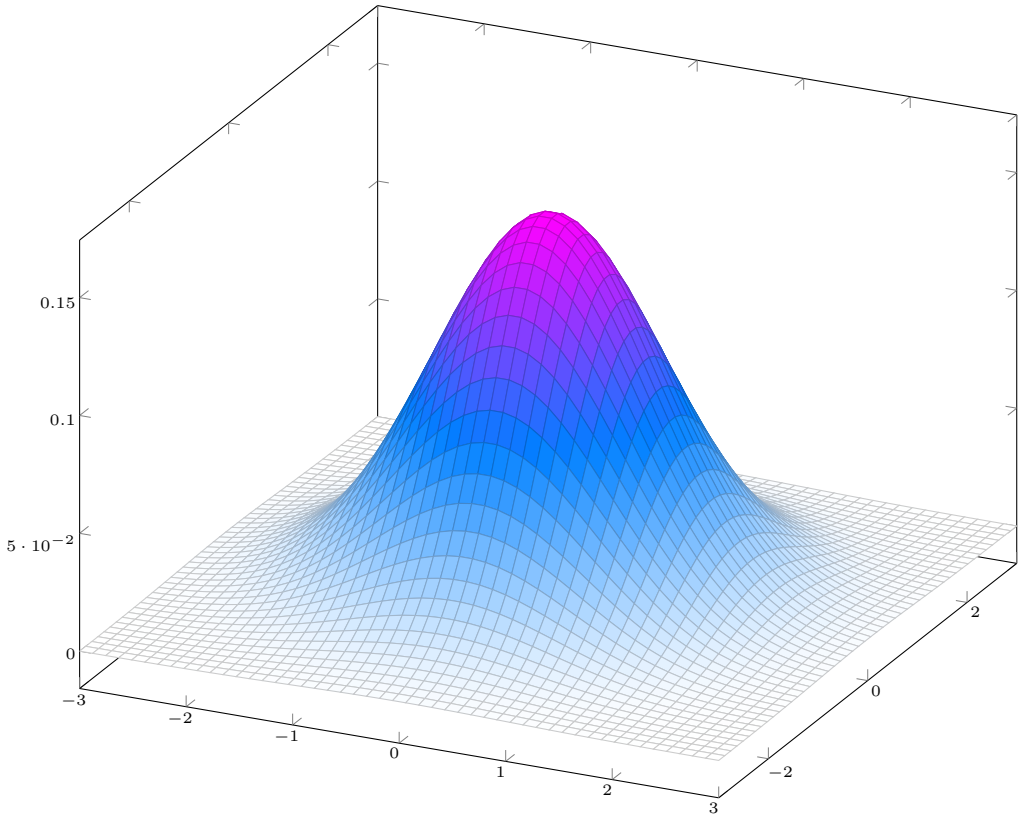
(i) $f_{X,Y}(x,y) \geq 0$ för alla $(x,y) \in \mathbf{R}^2$.

(ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dxdy = 1$.

(iii) Om $A \in \mathscr{B}$ (så $A \subset \mathbf{R}^2$ är *snäll*) så är $P((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,dxdy$.

(iv) Talet $f_{X,Y}(x,y)$ anger hur mycket sannolikhetsmassa det finns per areaenhet i punkten (x,y).

Exempel på hur en tvådimensionell täthetsfunktion kan se ut. Det är nu volymen, inte arean, som ska vara ett.



Väntevärde och funktioner av stokastiska variabler

Sats. Låt $Y = g(X)$ och $W = h(X_1, X_2, \ldots, X_n)$. I de kontinuerliga fallen blir

$$E(Y) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

och

$$E(W) = \int \cdots \int_{\mathbf{R}^n} h(x_1, x_2, \ldots, x_n) f_{(x_1,\ldots,x_n)}(x_1, x_2, \ldots, x_n)\, dx_1 \cdots dx_n.$$

Det diskreta fallet är analogt.

Varians och standardavvikelse

Definition. Låt X vara en stokastisk variabel med $|E(X)| < \infty$. **Variansen** $V(X)$ definieras som $V(X) = E((X - E(X))^2)$. **Standardavvikelsen** $D(X)$ definieras som $D(X) = \sqrt{V(X)}$.

Steiners sats

Sats. $V(X) = E(X^2) - E(X)^2$.

För väntevärdet gäller bland annat följande regler.

Linjäritet och oberoende produkt

Låt $X_1, X_2, \ldots, X_n$ vara stokastiska variabler. Då gäller

(i) $E\left(\sum_{k=1}^{n} c_k X_k\right) = \sum_{k=1}^{n} c_k E(X_k)$ för alla $c_1, c_2, \ldots, c_n \in \mathbf{R}$;

(ii) Om X_i och X_j är oberoende gäller $E(X_i X_j) = E(X_i)E(X_j)$.

(iii) $V(aX_i + b) = a^2 V(X_i)$ för alla $a, b \in \mathbf{R}$;

(iv) $V(aX_i \pm bX_j) = a^2 V(X_i) + b^2 V(X_j) + 2ab(E(X_i X_j) - E(X_i)E(X_j))$ för alla $a, b \in \mathbf{R}$;

(v) Om $X_1, X_2, \ldots, X_n$ är *oberoende* stokastiska variabler är $V\left(\sum_{k=1}^{n} c_k X_k\right) = \sum_{k=1}^{n} c_k^2 V(X_k)$ för alla $c_1, c_2, \ldots, c_n \in \mathbf{R}$;

Varianser adderas alltid!

Observera att det *alltid* blir ett plustecken mellan varianserna: $V(aX \pm bY) = a^2 V(X) + b^2 V(Y)$. Vi kommer *aldrig* att bilda skillnader mellan varianser!

Det följer att standardavvikelsen för en linjärkombination $aX+bY$ av två oberoende stokastiska variabler ges av $\sigma_{aX+bY} = \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2}$.

Definition. De **marginella** täthetsfunktionerna f_X och f_Y för X och Y i en kontinuerlig stokastisk variabel (X, Y) ges av

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy \qquad \text{och} \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx.$$

Motsvarande gäller om (X, Y) är diskret:

$$p_X(j) = \sum_{k} p_{X,Y}(j, k) \qquad \text{och} \qquad p_Y(k) = \sum_{j} p_{X,Y}(j, k).$$

Oberoende variabler

Sats. Om (X, Y) är en stokastisk variabel med simultan täthetsfunktion $f_{X,Y}$ gäller att X och Y är oberoende om och endast om $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. För en diskret variabel är motsvarande villkor $p_{X,Y}(j, k) = p_X(j) p_Y(k)$.

Faltningssatsen

Sats. Om X och Y är oberoende kontinuerliga stokastiska variabler så ges täthetsfunktionen f_Z för $Z = X + Y$ av

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx, \quad z \in \mathbf{R}.$$

Motsvarande gäller för diskreta variabler:

$$p_Z(k) = \sum_j p_X(j) p_Y(k - j).$$

2 Normalfördelning

Normalfördelningen är så viktig att den får ett eget avsnitt. Se till att ni verkligen kommer ihåg hur man hanterar normalfördelning, mycket är vunnet senare om detta maskineri sitter bra.

Normalfördelning

Variabeln X kallas **normalfördelad** med parametrarna μ och σ, $X \sim \mathrm{N}(\mu, \sigma^2)$, om

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbf{R}.$$

Om $\mu = 0$ och $\sigma = 1$ kallar vi X för **standardiserad**, och i det fallet betecknar vi täthetsfunktionen med

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbf{R}.$$

Fördelningsfunktionen för en normalfördelad variabel ges av

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du, \quad x \in \mathbf{R},$$

och även här döper vi speciellt den standardiserade fördelningsfunktionen till

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) du, \quad x \in \mathbf{R}.$$

Standardisering av variabel

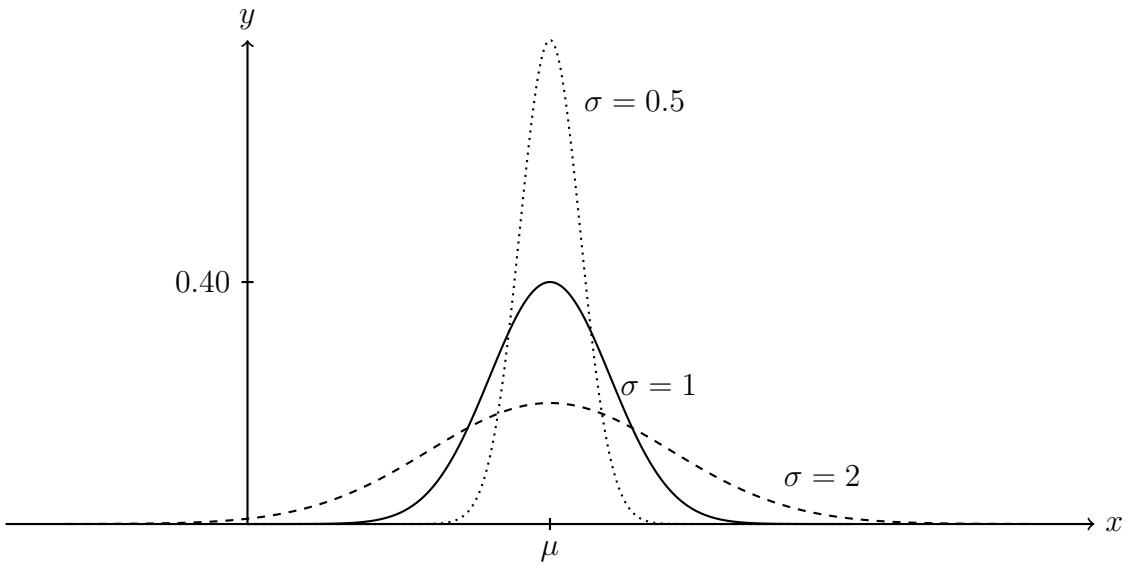
Om X är en stokastisk variabel med $E(X) = \mu$ och $V(X) = \sigma^2$, så är $Z = (X - \mu)/\sigma$ en stokastisk variabel med $E(Z) = 0$ och $V(Z) = 1$. Vi kallar Z för **standardiserad**.

$$X \sim \mathbf{N}(\mu, \sigma^2)$$

Om $X \sim \mathrm{N}(\mu, \sigma^2)$ så är $E(X) = \mu$ och $V(X) = \sigma^2$.

Standardavvikelse eller varians?

I kursboken (Blom et al) används beteckningen $X \sim N(\mu, \sigma)$, så den andra parametern är alltså standardavvikelsen σ, inte variansen σ^2 som vi använt ovan.



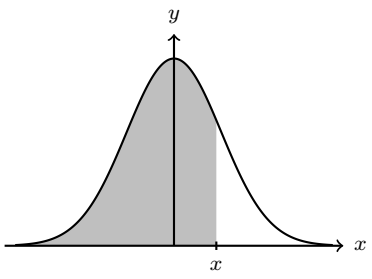
Bruk av tabell för $\Phi(x)$

Låt $X \sim \mathrm{N}(0,1)$. Då gäller

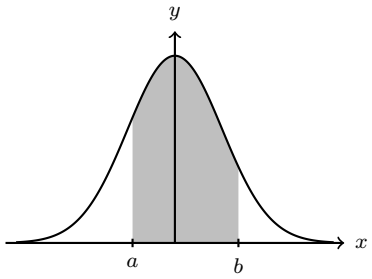
(i) $P(X \leq x) = \Phi(x)$ för alla $x \in \mathbf{R}$;

(ii) $P(a \leq X \leq b) = \Phi(b) - \Phi(a)$ för alla $a, b \in \mathbf{R}$ med $a \leq b$;

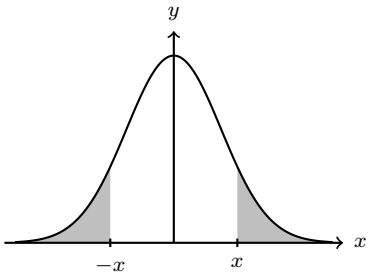
(iii) $\Phi(-x) = 1 - \Phi(x)$ för alla $x \in \mathbf{R}$.



$\Phi(x)$



$\Phi(b) - \Phi(a)$



$\Phi(-x) = 1 - \Phi(x)$

Exempel

Låt $X \sim \mathrm{N}(0,1)$. Bestäm $P(X \leq 1)$, $P(X < 1)$, $P(X \leq -1)$, samt $P(0 < X \leq 1)$.

Direkt ur tabell, $P(X \leq 1) = \Phi(1) \approx 0.8413$. Eftersom X är kontinuerlig kvittar det om olikheterna är strikta eller inte, så $P(X < 1) = P(X \leq 1) = \Phi(1)$ igen. Vidare har vi

$$P(X \leq -1) = \Phi(-1) = 1 - \Phi(1) = 0.1587$$

och $P(0 < X \leq 1) = \Phi(1) - \Phi(0) = 0.8413 - 0.5 = 0.3413$.

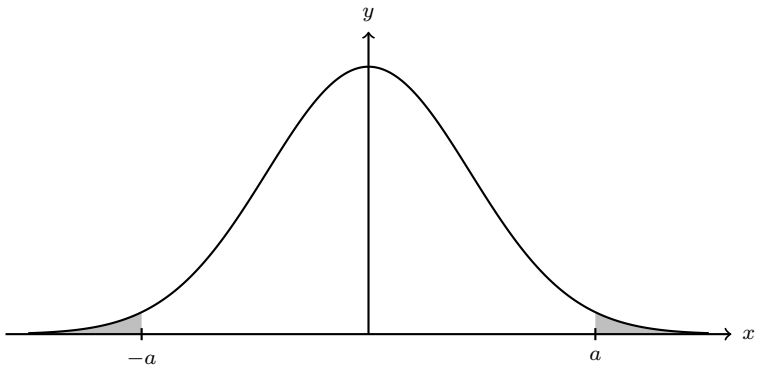
Standardisering av normalfördelning

Sats. $X \sim \mathrm{N}(\mu, \sigma^2) \quad \Leftrightarrow \quad Z = \dfrac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1)$.

Exempel

Låt $X \sim \mathrm{N}(0, 1)$. Hitta ett tal a så att $P(|X| > a) = 0.05$.

Situationen ser ut som i bilden nedan. De skuggade områderna utgör tillsammans 5% av sannolikhetsmassan, och på grund av symmetri måste det vara 2.5% i varje ”svans”.



Om vi söker talet a, och vill använda funktionen $\Phi(x) = P(X \leq x)$, måste vi söka det tal som ger $\Phi(a) = 0.975$ (dvs de 2.5% i vänstra svansen tillsammans med de 95% som ligger i den stora kroppen). Detta gör vi genom att helt enkelt leta efter talet 0.975 i tabellen över $\Phi(x)$ värden. Där finner vi att $a = 1.96$ uppfyller kravet att $P(X \leq a) = 0.975$.

Summor och medelvärde

Sats. Låt $X_1, X_2, \ldots, X_n$ vara oberoende och $X_k \sim \mathrm{N}(\mu, \sigma^2)$ för $k = 1, 2, \ldots, n$. Då gäller följande:

$$X := \sum_{k=1}^{n} X_k \sim \mathrm{N}(n\mu, n\sigma^2) \qquad \text{och} \qquad \overline{X} := \frac{1}{n}\sum_{k=1}^{n} X_k \sim \mathrm{N}(\mu, \sigma^2/n).$$

Mer generellt, om $X_k \sim \mathrm{N}(\mu_k, \sigma_k^2)$ och $c_0, c_1, \ldots, c_n \in \mathbf{R}$ är

$$c_0 + \sum_{k=1}^{n} c_k X_k \sim \mathrm{N}\left(c_0 + \sum_{k=1}^{n} c_k \mu_k, \sum_{k=1}^{n} c_k^2 \sigma_k^2\right).$$

Likheten för medelvärdet är intressant då det innebär att ju fler ”likadana” variabler vi tar med i ett medelvärde, desto *mindre* blir variansen. Till exempel får vi alltså säkrare resultat ju fler mätningar vi gör (något som känns intuitivt korrekt). Det är dock mycket viktigt att variablerna är *oberoende*. Annars gäller inte satsen! Vi bildar aldrig heller några skillnader mellan varianser, utan det som gör att variansen minskar med antalet termer är faktorn $1/n$ i medelvärdet:

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}V\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum_{k=1}^{n} V(X_k) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

eftersom variablerna är oberoende och $V(X_k) = \sigma^2$ för alla k.
Nottera även att satsen faktiskt säger att summan av normalfördelade variabler fortfarande är normalfördelad, något som inte gäller vilken fördelning som helst (se faltningssatsen).

Statistisk inferens

"We have such sights to show you"
–Pinhead

3 Begrepp

Vi är nu redo för att dyka ned i statistisk inferensteori! I sedvanlig ordning börjar vi med att definiera lite begrepp så vi är överens om vad vi diskuterar.

Stickprov

Definition. Låt de stokastiska variablerna X_1, X_2,..., X_n vara oberoende och ha samma fördelningsfunktion F. Följden $X_1, X_2, \ldots, X_n$ kallas ett **slumpmässigt stickprov** (av F). Ett **stickprov** x_1, x_2,..., x_n består av **observationer** av variablerna X_1, X_2,..., X_n. Samtliga möjliga observationer brukar kallas **populationen**. Vi säger att **stickprovsstorleken** är n.

Exempel

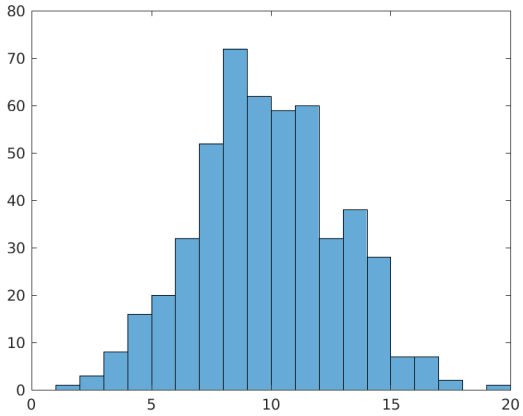
Antag att vi kastar en perfekt tärning 5 gånger och att dessa kast är oberoende. Före kasten representerar den stokastiska variabeln X_k resultatet vid kast k där alla X_k har samma fördelning; vi vet ännu inte vad resultatet blir, men känner sannolikhetsfördelningen. Följden X_k, $k = 1, 2, \ldots, 5$ är det *slumpmässiga stickprovet*.
Efter kasten har vi erhållit observationer $x_1, x_2, \ldots, x_5$ av det slumpmässiga stickprovet. Dett är vårt *stickprov* och består alltså av utfallen vid kasten. Dessa observationer tillhör *populationen*. *Stickprovsstorleken* är 5.

Värt att notera är att språkbruket ibland är slarvigt där både stickprov och slumpmässigt stickprov används för att beskriva både följden av stokastiska variabler och följden av observationer (utfall). Det viktiga är att hålla koll på vad ni själva menar när ni genomför analyser.

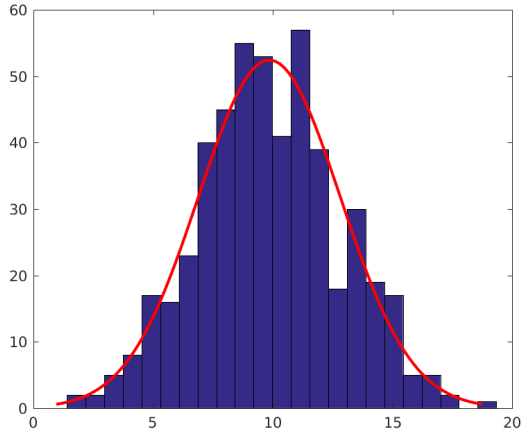
4 Representation av stickprov

Man kan representera statistiska data på en hel drös olika sätt med allt från tabeller till stolpdiagram till histogram till lådplottar. Läs avsnittet i boken om detta. Vi nöjer oss med att titta lite närmare på de verktyg vi kommer använda oss av i kursen. Ett mycket vanligt sätt att visualisera fördelningen för en mängd data är med hjälp av histogram. Vi genererar lite normalfördelad slumpdata i MATLAB och renderar ett histogram.

```
>> U = normrnd(10,3,500,1);
>> histogram(U);
```

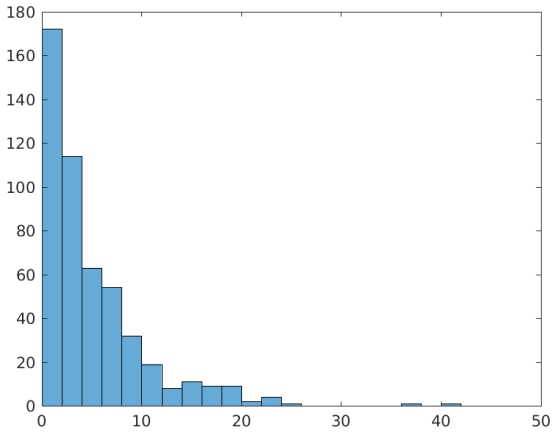



```
>> U = normrnd(10,3,500,1);
>> histfit(U)
```



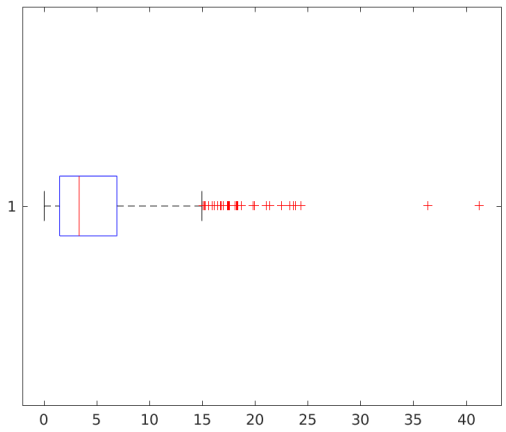
Om vi testar med exponentialfördelning istället blir resultatet enligt nedan.

```
>> U = exprnd(10,500,1);
>> histogram(U);
```

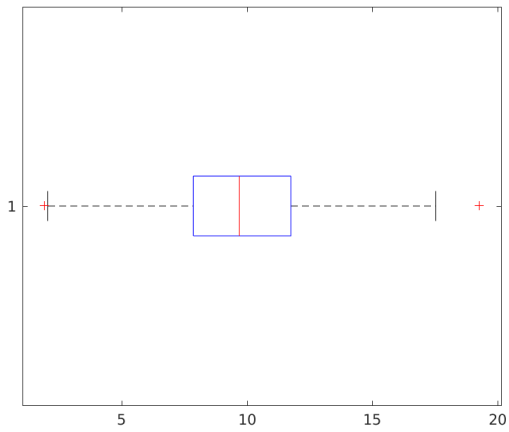


Ett lådagram (boxplot) representerar också materialet, kanske på ett enklare sätt för den oinsatte i sannolikhetsfördelningar. Lådan innehåller 50% av resultaten och den vänstra lådkanten är den undre kvartilen (25% till vänster om den) och den högra är den övre kvartilen (med 25% till höger om den). Medianen markeras med ett streck i lådan. Maximum och minimum markeras med små vertikala streck i slutet på en horisontell linje genom mitten på lådan. Värden som bedöms vara uteliggare markeras med kryss längs samma centrumlinje.

```
>> U = exprnd(5,500,1);
>> boxplot(U, 'orientation', 'horizontal');
```




```
>> U = normrnd(10,3,500,1);
>> boxplot(U, 'orientation', 'horizontal');
```

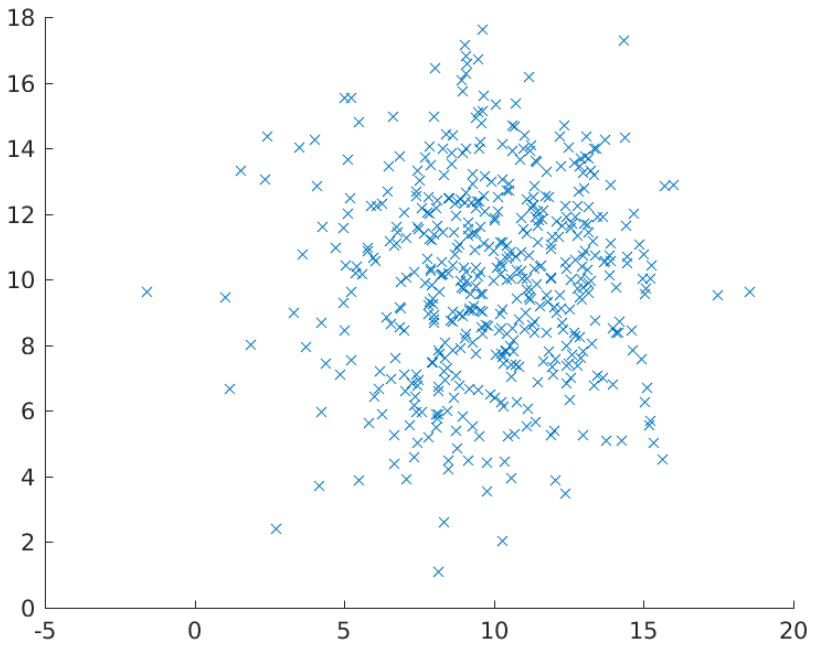


Vi kan tydligt se skillnad på hur mätvärden är spridda. Jämför även med motsvarande histogram ovan.

4.1 Två-dimensionell data

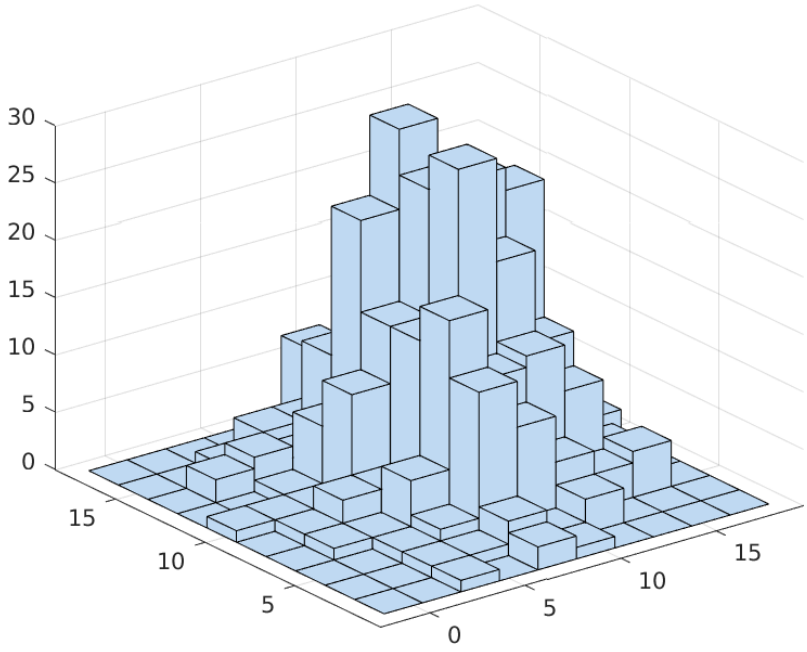
Vi kan även ha mätvärden i form av punkter (x, y) och den vanligaste figuren i dessa sammahang är ett **spridningsdiagram** (scatter plot) där man helt enkelt plottar ut punkter vid varje koordinat (x_i, y_i).

```
>> U = normrnd(10,3,500,2);
>> scatter(U(:,1), U(:,2), 'x');
```



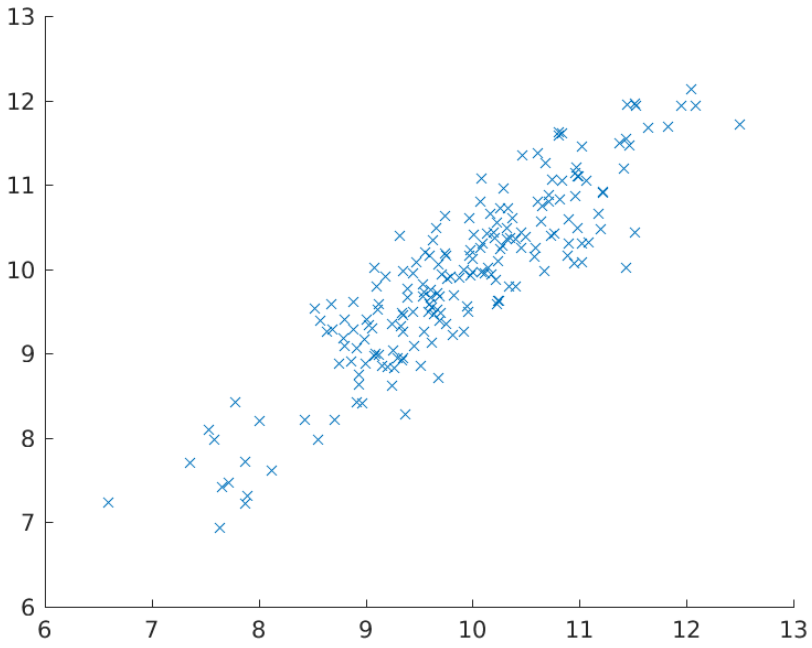
Vi ser från figuren att värdena verkar vara centrerade kring $(10, 10)$ och att det verkar föreligga någon form av cirkulär symmetri. Stämmer det för den bivarata normalfördelningen när komponenterna är oberoende? Vi kan även rendera ett två-dimensionellt histogram.

```
>> hist3(U);
```



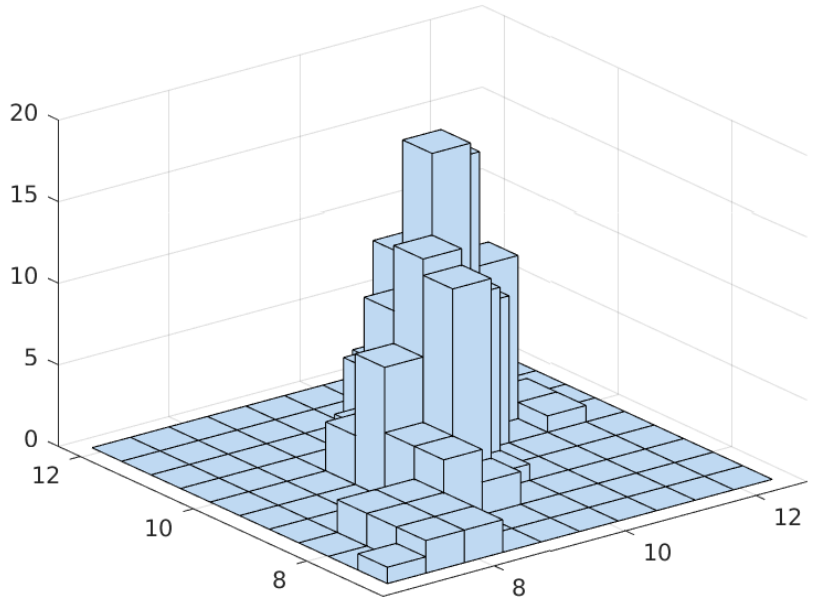
Vad gäller om variablerna i den bivariata normalfördelningen inte är oberoende?

```
>> mu = [10 10]; rho = 0.90; s1 = 1; s2 = 1;
>> Sigma = [s1*s1 s1*s2*rho; s1*s2*rho s2*s2]
>> R = chol(Sigma);
>> z = repmat(mu, 200, 1) + randn(200,2)*R;
>> scatter(z(:,1),z(:,2), 'x');
```



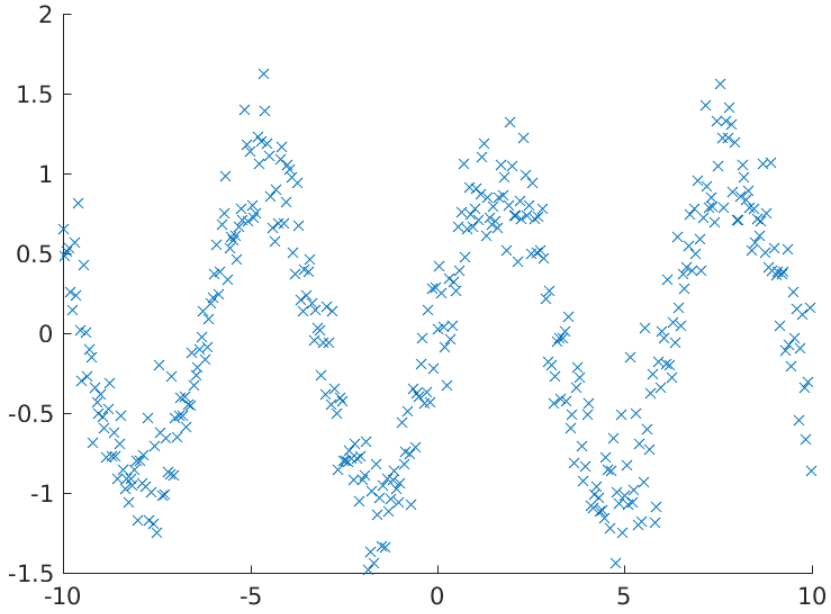
Värdena verkar fortfarande vara centrerade kring $(10, 10)$ (i någon mening) men symmetrin verkar nu utdragen diagonalt. Stämmer det för en bivarat normalfördelningen med korrelationen 0.90? Ett histogram kan genereras som ovan.

```
>> hist3(z);
```



Något konstigare? Visst.

```
>> x = (-10:0.05:10); y = sin(x) + normrnd(0,0.25,size(x));
>> scatter(x,y,'x');
```



Vi kan tydligt urskilja sinus-termen och något slags brus som gör att det inte blir en perfekt linje. Kan man få bort bruset?

5 Punktskattningar

Antag att en fördelning beror på en okänd parameter θ. Med detta menar vi att fördelningens täthetsfunktion (eller sannolikhetsfunktion) beror på ett okänt tal θ, och skriver $f(x\,;\,\theta)$ respektive $p(k\,;\,\theta)$ för att markera detta. Om vi har ett stickprov från en fördelning med en okänd parameter, kan vi *skatta* den okända parametern? Med andra ord, kan vi göra en ”gissning” på det verkliga värdet på parametern θ?

Punktskattning

Definition. En **punktskattning** $\widehat{\theta}$ av parametern θ är en funktion (ibland kallad **stickprovsfunktion**) av de observerade värdena $x_1, x_2, \ldots, x_n$:

$$\widehat{\theta} = g(x_1, x_2, \ldots, x_n).$$

Vi definierar motsvarande **stickprovsvariabel** $\widehat{\Theta}$ enligt

$$\widehat{\Theta} = g(X_1, X_2, \ldots, X_n).$$

Det är viktigt att tänka på att $\widehat{\theta}$ är en siffra, beräknad från de observerade värdena, medan $\widehat{\Theta}$ är en stokastisk variabel. Som vanligt använder vi stora bokstäver för att markera att vi syftar på en stokastisk variabel. Sambandet mellan $\widehat{\theta}$ och $\widehat{\Theta}$ är alltså att $\widehat{\theta}$ är en observation av den stokastiska variabeln $\widehat{\Theta}$. Förutom detta dras vi fortfarande med det okända talet θ, som inte är stokastiskt, utan endast en okänd konstant.

Exponentialfördelning

Betrakta en exponentialfördelning med okänt väntevärde. Formeln är välkänd: för alla $x \geq 0$ gäller att $f(x\,;\,\mu) = \mu^{-1}\exp(\mu^{-1}x)$. Parametern θ är alltså väntevärdet μ i detta fall. Ibland använder man exponentialfördelningen för att beskriva elektriska komponenters livslängd, och genom att betrakta ett stickprov kan man då uppskatta livslängden för en hel tillverkningsomgång.

Stokastiskt eller ej?

Var noggran med att tydligt visa och göra skillnad på vad som är stokastiskt eller inte i din redovisning! Vi har tre storheter:

(i) θ – verkligt värde. Okänt. Deterministiskt.

(ii) $\widehat{\theta}$ – skattat värde. Känt (beräknat från stickprovet). Deterministiskt.

(iii) $\widehat{\Theta}$ – stickprovsvariabeln. Denna är stokastisk!

Sannolikheter som beräknas bör använda sig av $\widehat{\Theta}$ då $\widehat{\Theta}$ beskriver variationen hos $\widehat{\theta}$ för olika stickprov. Om bara $\widehat{\theta}$ och θ ingår är sannolikheten alltid noll eller ett (varför?).

Så om vi har ett stickprov från en fördelning som beror på en okänd parameter, hur hittar vi skattningsfunktionen g? Fungerar vad som helst?

Exempel

Vid fyra dagar på en festival gjordes ljudnivåmätningar vid lunchtid. Följande mätdata erhölls: 107dB, 110dB, 117dB, 101dB. Vi antar att mätningarna är observationer av oberoende och likafördelade variabler med okänt väntevärde μ. Hur hittar vi en skattning $\widehat{\mu}$?

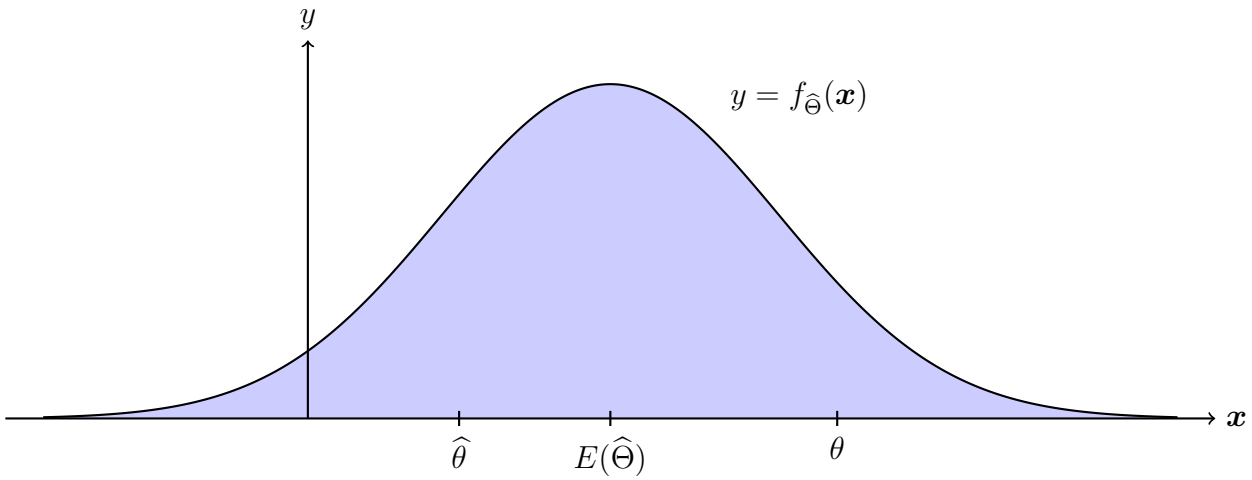
(i) $\widehat{\mu} = 100$dB är en skattning.

(ii) $\widehat{\mu} = 107$dB (den första dagen) är en skattning.

(iii) $\widehat{\mu} = (107 + 110 + 117 + 101)/4 = 108.75$dB (medelvärdet) är en skattning.

(iv) $\widehat{\mu} = \min\{107, 110, 117, 101\} = 101$dB är en skattning.

Så svaret är i princip ”ja,” alla värden $\widehat{\theta}$ som är tillåtna i modellen vi betraktar är punktskattningar. Hur väljer vi då den bästa, eller åtminstone en bra, punktskattning? Stickprovsvariabeln $\widehat{\Theta}$ är en stokastisk variabel, så normalt sett har den en täthetsfunktion (alternativt sannolikhetsfunktion). Vi skisserar en tänkbar täthetsfunktion för $\widehat{\Theta}$ (tänk dock på att $\widehat{\Theta}$ typisk är en flerdimensionell stokastisk variabel då $\widehat{\Theta} = g(X_1, X_2, \ldots, X_n)$).



Vi vet att $\widehat{\theta}$ beräknas från observerade siffror, så $\widehat{\theta}$ kan hamna lite vart som helst. Dessutom vet vi inte om väntevärdet $E(\widehat{\Theta})$ sammanfaller med det okända värdet θ. Så hur kan vi då avgöra om en punktskattning är bra eller inte? Det finns två viktiga kriterier: *väntevärdesriktighet* och *konsistens*. Vi återkommer till dessa nästa föreläsning.

Väntevärdesriktig skattning

Definition. Stickprovsvariabeln $\widehat{\Theta}$ kallas **väntevärdesriktig** (vvr) om $E(\widehat{\Theta}) = \theta$.

Om en punktskattning inte är väntevärdesriktig pratar man ibland om ett systematiskt fel. Vi definierar detta som skillnaden $E(\widehat{\Theta}) - \theta$. En väntevärdesriktig skattning $\widehat{\Theta}$ har alltså inget systematiskt fel; i ”medel” kommer den att hamna rätt (tänk på de stora talens lag).

Systematiskt fel; bias

Definition. Om $E(\widehat{\Theta}) - \theta \neq 0$ så säger vi att $\widehat{\Theta}$ har ett systematiskt fel (ett bias).

5.1 Vanliga punktskattningar

Vissa punktskattningar är så vanliga att de ha fått egna namn. Vi vill ofta skatta medelvärdet som positionsmått och stickprovsstandardavvikelsen är ett vanligt mått på spridningen.

Stickprovsmedelvärde

Definition. Stickprovsmedelvärdet $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ är en skattning av stickprovsväntevärdet $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.

Stickprovsvarians och stickprovsstandardavvikelse

Definition. Stickprovsvariansen $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ skattar $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
Stickprovsstandardavvikelsen skattar vi med med $s = \sqrt{s^2}$, dvs s är en skattning av S.

Varför $n-1$? Vi återkommer till det nästa föreläsning.

6 Vilka skattningar är bra?

När vi har ett stickprov från en fördelning som beror på en okänd parameter så fungerar alltså i princip vad som helst som skattning på parametern. Funktionen g är således godtycklig. Vi vet att $\widehat{\theta}$ beräknas från observerade siffror, så $\widehat{\theta}$ kan hamna lite vart som helst. Dessutom vet vi inte om väntevärdet $E(\widehat{\Theta})$ sammanfaller med det okända värdet θ. Så hur kan vi då avgöra om en punktskattning är bra eller inte? Det finns två viktiga kriterier: *väntevärdesriktighet* som vi såg ovan och *konsistens*.
Om en punktskattning inte är väntevärdesriktig pratar man ibland om ett systematiskt fel. Vi definierar detta som skillnaden $E(\widehat{\Theta}) - \theta$. En väntevärdesriktig skattning $\widehat{\Theta}$ har alltså inget systematiskt fel; i "medel" kommer den att hamna rätt (tänk på de stora talens lag). Vi vill också gärna ha egenskapen att en punktskattning blir bättre ju större stickprov vi använder.

Konsistent skattning

Definition. Antag att vi har en punktskattning $\widehat{\Theta}_n$ för varje stickprovsstorlek n. Om det för varje $\epsilon > 0$ gäller att

$$\lim_{n\to\infty} P(|\widehat{\Theta}_n - \theta| > \epsilon) = 0,$$

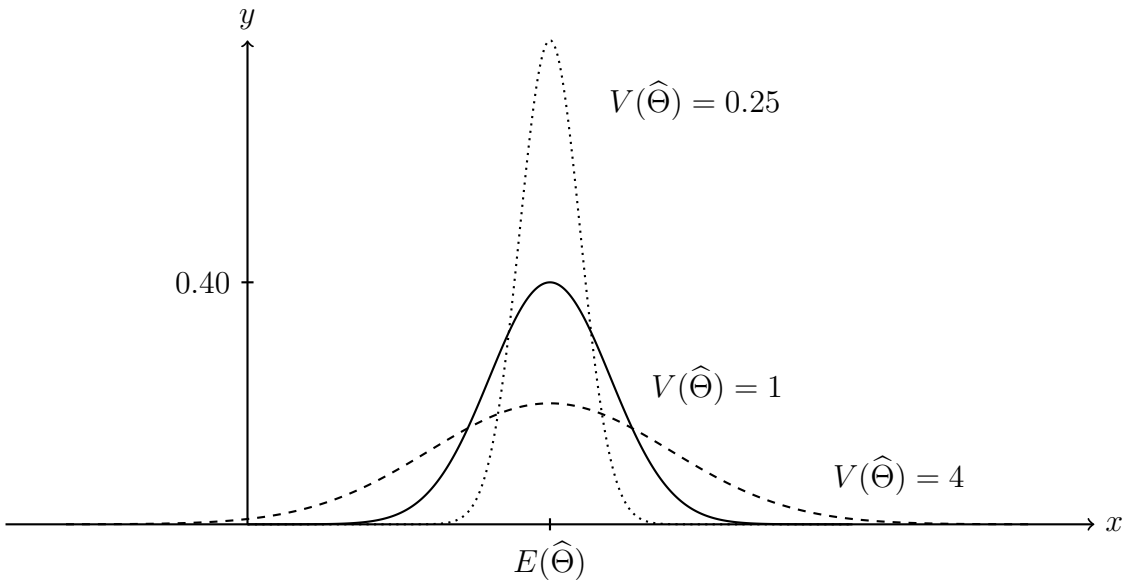
så kallar vi denna punktskattning för *konsistent*.

Teknisk definition, men innebörden bör vara klar. När stickprovsstorleken går mot oändligheten så är sannolikheten att skattningen befinner sig nära det okända värdet stor. Villkoret för konsistens kan vara lite jobbigt att arbeta med så följande sats är ofta användbar för att kontrollera konsistens.

Ett kriterium för konsistens

Om $E(\widehat{\Theta}_n) = \theta$ för alla n och $\lim_{n\to\infty} V(\widehat{\Theta}_n) = 0$ så är skattningen konsistent.

Bevisskiss: Här använder vi Tjebysjovs olikhet: om $a > 0$ och X är en stokastisk variabel så att $E(X) = \mu$ och $V(X) = \sigma^2 < \infty$, så gäller $P(|X - \mu| > a\sigma) \leq \frac{1}{a^2}$. Om vi låter $a = \epsilon/\sigma_n$ för fixt $\epsilon > 0$ så erhåller vi $P(|\widehat{\Theta} - \theta| > \epsilon) \leq \frac{\sigma_n^2}{\epsilon^2} \to 0$ då $n \to \infty$, eftersom $\sigma_n^2 = V(\widehat{\Theta}_n) \to 0$ då $n \to \infty$. □



Mindre varians för $\widehat{\Theta}$ medför att sannolikhetsmassan är mer centrerad kring väntevärdet (vid symmetrisk fördelning).

Exempel

Betrakta exemplet med ljudnivåerna igen, vi hade följande mätdata: 107dB, 110dB, 117dB, 101dB. Vi undersöker skattningarna lite närmare.

(i) $\widehat{\mu} = 100$dB är en fix siffra och kan varken vara väntevärdesriktig eller konsistent. Dålig skattning.

(ii) Den första siffran är en observation av den första variabeln X_1 i stickprovet. Alltså är $\widehat{M} = X_1$. Eftersom $E(\widehat{M}) = E(X_1) = \mu$ så är skattningen väntevärdesriktig. Med konstant varians oavsett stickprovsstorlek kan den dock inte vara konsistent.

(iii) Medelvärdet är både väntevärdesriktigt och konsistent; se nästa avsnitt!

(iv) Här blir det lite klurigare när vi bildar minimum av observationerna. Vi undersöker ett specialfall där variablerna är exponentialfördelade, säg $X_i \sim \text{Exp}(\mu)$. Då gäller att (se avsnittet med kö-teori i TAMS79)

$$\widehat{M} = \min\{X_1, X_2, X_3, X_4\} \sim \text{Exp}(\mu/4).$$

Således erhåller vi att $E(\widehat{M}) = \mu/4 \neq \mu$. Det finns alltså gott om fall då detta inte är en väntevärdesriktig skattning! Går det att korrigera skattningen?

6.1 Effektivitet – jämförelse mellan skattningar

Så om vi har två olika stickprovsvariabler $\widehat{\Theta}$ och Θ^*, hur avgör vi vilken som är ”bäst”? Om båda är väntevärdesriktiga och konsistenta, kan man säga att en är bättre?

Effektivitet

Definition. En skattning $\widehat{\Theta}$ kallas *effektivare* än en skattning Θ^* om $V(\widehat{\Theta}) \leq V(\Theta^*)$.

Den stickprovsvariabel med minst varians kallas alltså mer effektiv, och med mindre varians känns det rimligt att kalla den skattningen bättre (om den är någorlunda väntevärdesriktig).

7 Momentmetoden

Så kan man systematiskt finna lämpliga skattningar på något sätt om man känner till viss information om fördelningen? Svaret är ja, det finns många sådana metoder. Bland annat momentmetoden, MK-metoden (minsta kvadrat), och kanske den vanligaste, ML-skattningar (maximum likelihood). Vi börjar med att betrakta momentmetoden.

Momentmetoden (för en parameter)

Definition. Låt $E(X_i) = \mu(\theta)$ för alla i. Momentskattningen $\widehat{\theta}$ av θ fås genom att lösa ekvationen $\mu(\widehat{\theta}) = \overline{x}$.

Exempel

Låt $x_1, x_2, \ldots, x_n$ vara ett stickprov från en fördelning med täthetsfunktionen $f(x\,;\,\theta) = \theta e^{-\theta x}$ för $x \geq 0$. Använd momentmetoden för att punktskatta θ.

Lösning: Vi börjar med att beräkna väntevärdet, det vill säga funktionen $\mu(\theta)$. Alltså,

$$\mu(\theta) = \int_0^\infty x\theta e^{-\theta x}\,dx = \left[x\theta\frac{e^{-\theta x}}{-\theta}\right]_0^\infty + \int_0^\infty e^{-\theta x} = \theta^{-1}.$$

Vi löser nu ekvationen $\mu(\widehat{\theta}) = \overline{x}$, och erhåller då att

$$\widehat{\theta}^{-1} = \overline{x} \quad \Leftrightarrow \quad \widehat{\theta} = \frac{1}{\overline{x}},$$

så länge $\overline{x} \neq 0$. Momentskattningen av θ ges alltså av $\widehat{\theta} = (\overline{x})^{-1}$. Vad händer om $\overline{x} = 0$?

Om man har flera parametrar då? Här visar det sig varför metoden ovan kallas *momentmetoden*.

Moment

Definition. Låt X vara en stokastisk variabel X. För $k = 1, 2, \ldots$ definierar vi momenten m_k för X enligt $m_k = E(X^k)$.

Det första momentet m_1 är alltså inget annat än väntevärdet för X.

Momentskattning med flera parametrar

Definition. Låt $X \sim F(x\,;\theta_1,\theta_2,\ldots,\theta_j)$ bero på j okända parametrar $\theta_1,\theta_2,\ldots,\theta_j$ och definiera $m_i(\theta_1,\theta_2,\ldots,\theta_j) := E(X^i)$, $i = 1,2,\ldots$ Momentskattningarna för θ_k, $k = 1,2,\ldots,j$, ges av lösningen till ekvationssystemet

$$m_i(\widehat{\theta_1},\widehat{\theta_2},\ldots,\widehat{\theta_j}) = \frac{1}{n}\sum_{k=1}^{n} x_k^i, \quad i = 1,2,\ldots,j.$$

Observera att det inte är säkert att en lösning finns eller att lösningen är entydig i de fall den existerar. Vidare kan det även inträffa att lösningen hamnar utanför det område som är tillåtet för parametern (i vilket fall vi givetvis inte kan använda den).

Exempel

Låt $X_k \sim N(\mu,\sigma^2)$, $k = 1,2,\ldots,n$ vara ett stickprov. Hitta momentskattningarna för μ och σ^2.

Lösning. Vi vet att $E(X) = \mu$ och $E(X^2) = V(X) + E(X)^2 = \sigma^2 + \mu^2$, så

$$\begin{cases} \widehat{\mu} = \overline{x}, \\ \widehat{\sigma}^2 + \widehat{\mu}^2 = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k^2. \end{cases}$$

Således erhåller vi direkt att $\widehat{\mu} = \overline{x}$. För $\widehat{\sigma}^2$ är

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \overline{x}^2 = \cdots = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2.$$

Nästan stickprovsvariansen alltså.

Vektornotation för parametrar

Definition. När vi har en fördelning som beror på flera parametrar, säg $\theta_1,\theta_2,\ldots,\theta_j$, så skriver vi ibland $\boldsymbol{\theta} = (\theta_1,\theta_2,\ldots,\theta_j) \in \mathbf{R}^j$ som en j-dimensionell vektor. Notationen blir då mer kompakt. Bokstäver typsatta i fet stil indikerar oftast en vektor i denna kurs.

Föreläsning 2: Punktskattningar

Johan Thim (johan.thim@liu.se)

27 augusti 2018

1 Repetition

Stickprov

Definition. Låt de stokastiska variablerna X_1, X_2,..., X_n vara oberoende och ha samma fördelningsfunktion F. Ett stickprov x_1, x_2,..., x_n består av observationer av variablerna X_1, X_2,..., X_n. Vi säger att *stickprovsstorleken* är n.

Punktskattning

Definition. En *punktskattning* $\widehat{\theta}$ av parametern θ är en funktion av de observerade värdena $x_1, x_2, \ldots, x_n$:

$$\widehat{\theta} = g(x_1, x_2, \ldots, x_n).$$

Vi definierar motsvarande *stickprovsvariabel* $\widehat{\Theta}$ enligt

$$\widehat{\Theta} = g(X_1, X_2, \ldots, X_n).$$

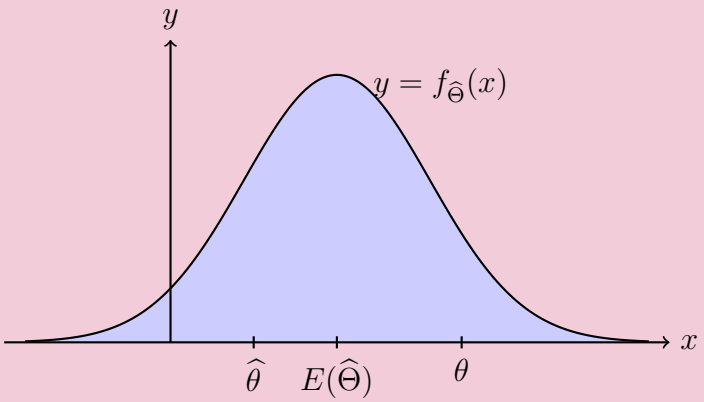
Stokastiskt eller ej?

Var noggran med att tydligt visa och göra skillnad på vad som är stokastiskt eller inte i din redovisning! Vi har tre storheter:

(i) θ – verkligt värde. Okänt. Deterministiskt.

(ii) $\widehat{\theta}$ – skattat värde. Känt (beräknat från stickprovet). Deterministiskt.

(iii) $\widehat{\Theta}$ – stickprovsvariabeln. Denna är stokastisk!



Sannolikheter som beräknas bör använda sig av $\widehat{\Theta}$ då $\widehat{\Theta}$ beskriver variationen hos $\widehat{\theta}$ för olika stickprov. Om bara $\widehat{\theta}$ och θ ingår är sannolikheten alltid noll eller ett (varför?).

Egenskaper för skattningar

Definition. Stickprovsvariabeln $\widehat{\Theta}$ kallas

(i) *väntevärdesriktig* (vvr) om $E(\widehat{\Theta}) = \theta$;

(ii) *konsistent* om det för varje $\epsilon > 0$ gäller att

$$\lim_{n\to\infty} P(|\widehat{\Theta}_n - \theta| > \epsilon) = 0,$$

där $\widehat{\Theta}_n$ är punktskattningen för varje stickprovsstorlek n;

(iii) *effektivare* än en skattning Θ^* om $V(\widehat{\Theta}) \leq V(\Theta^*)$.

Om en punktskattning inte är väntevärdesriktig pratar man ibland om ett systematiskt fel. Vi definierar detta som skillnaden $E(\widehat{\Theta}) - \theta$. En väntevärdesriktig skattning $\widehat{\Theta}$ har alltså inget systematiskt fel; i ”medel” kommer den att hamna rätt (tänk på de stora talens lag). Vi vill också gärna ha egenskapen att en punktskattning blir bättre ju större stickprov vi använder.

2 Vanliga punktskattningar

Vi stötte på medelvärdet och stickprovsvariansen på föregående föreläsning. Dessa skattningar är vettiga skattningar av väntevärdet och variansen i meningen att de är väntevärdesriktiga och konsistenta.

Medelvärde

Medelvärdet $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ är en väntevärdesriktig och konsistent skattning av väntevärdet.

Bevis: Variablerna X_k är oberoende och likafördelade. Låt $E(X_i) = \mu$ och $V(X_i) = \sigma^2$ för alla i. Eftersom väntevärdesoperatorn är linjär så gäller att

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n\mu}{n} = \mu.$$

Alltså är $\overline{X}$ en väntevärdesriktig skattning av μ.
Då variablerna är oberoende kan vi göra en liknande kalkyl för variansen:

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Här ser vi att $V(\overline{X}) \to 0$ då $n \to \infty$, så enligt satsen ovan är skattningen konsistent.

Stickprovsvarians

Stickprovsvariansen $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ är en väntevärdesriktig skattning av variansen.

Bevis: Detta bevis är lite bökigare, men följer samma princip.

$$E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2 - 2X_i\overline{X} + \overline{X}^2\right)$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\big(E(X_i^2) - 2E(X_i\overline{X}) + E(\overline{X}^2)\big).$$

Vi vet att $E(\overline{X}) = \mu$ och att $V(\overline{X}) = \sigma^2/n$. Steiners formel säger att $E(Y^2) = V(Y) + E(Y)^2$ för en stokastisk variabel Y, vilket vi kan utnyttja för att skriva

$$E(X_i^2) = V(X_i) + E(X_i)^2 = \sigma^2 + \mu^2 \quad \text{samt} \quad E(\overline{X}^2) = \sigma^2/n + \mu^2.$$

Vidare så ser vi att

$$E(X_i\overline{X}) = E\left(X_i\frac{1}{n}\sum_{k=1}^{n}X_k\right) = \frac{1}{n}\sum_{k=1}^{n}E(X_iX_k)$$

och eftersom $E(X_iX_k) = E(X_i)E(X_k) = \mu^2$ om $i \neq k$ (eftersom dessa variabler är oberoende) och $E(X_i^2) = \sigma^2 + \mu^2$ (då $i = k$) kan vi skriva

$$E(X_i\overline{X}) = ((n-1)\mu^2 + \sigma^2 + \mu^2)/n = \mu^2 + \sigma^2/n.$$

Vi återgår till det sökta väntevärdet:

$$E(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}\big(\sigma^2 + \mu^2 - 2(\mu^2 + \sigma^2/n) + \sigma^2/n + \mu^2\big) = \frac{n\sigma^2 - n\sigma^2/n}{n-1} = \sigma^2.$$

Alltså är S^2 en väntevärdesriktig skattning (av σ^2). Värt att notera är att $S = \sqrt{S^2}$ **inte** är en väntevärdesriktig skattning av σ (men den används oftast ändå!).

3 Metoder för att hitta punktskattningar

Vi har slarvat lite i definitionen av punktskattningar när det gäller vilka värden på den okända parametern θ som är tillåtna. Vi inför begreppet **parameterrum**.

Parameterrum

Definition. Vi låter Ω_θ beteckna **parameterrummet** av alla tillåtna värden på parametern θ.

Parameterrummet är alltså en delmängd av $\mathbf{R}^p$ där p är antalet parametrar (tänk på att $\boldsymbol{\theta}$ kan vara en vektor $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$).

Exempel

(i) Om $X \sim N(\mu, \sigma^2)$ kan vi tänka oss $\boldsymbol{\theta} = (\mu, \sigma^2)$, i vilket fall parameterrummet kan representeras som $\mathbf{R} \times (0, \infty)$.

(ii) Om $X \sim \text{Bin}(n, p)$ där n är fixerad är parameterrummet $\Omega_p = [0, 1]$.

Skulle vi med någon metod hitta en skattning som faller utanför parameterrummet måste den förkastas. Så åter till frågan hur vi hittar skattningar mer systematiskt.

3.1 Momentmetoden

Vi såg momentmetoden i förra föreläsningen. Låt oss endast repetera vad den gick ut på.

Momentskattning med flera parametrar

Definition. Låt $X \sim F(x\,;\theta_1, \theta_2, \ldots, \theta_j)$ bero på j okända parametrar $\theta_1, \theta_2, \ldots, \theta_j$ och definiera $m_i(\theta_1, \theta_2, \ldots, \theta_j) := E(X^i)$, $i = 1, 2, \ldots$ Momentskattningarna för θ_k, $k = 1, 2, \ldots, j$, ges av lösningen till ekvationssystemet

$$m_i(\widehat{\theta_1}, \widehat{\theta_2}, \ldots, \widehat{\theta_j}) = \frac{1}{n}\sum_{k=1}^{n} x_k^i, \quad i = 1, 2, \ldots, j.$$

3.2 MK-skattning

Minsta kvadrat-metoden har vi egentligen stött på i tidigare kurser, mer specifikt när vi hittade approximativa lösningar till överbestämda ekvationssystem. Faktum är att vi kommer att upprepa den proceduren senare i denna kurs i sammband med linjär regression.
Låt $x_1, x_2, \ldots, x_n$ vara observationer av oberoende stokastiska variabler $X_1, X_2, \ldots, X_n$ sådana att $E(X_k) = \mu_k(\boldsymbol{\theta})$ och $V(X_k) = \sigma^2$ för $k = 1, 2, \ldots, n$ (alltså samma varians men potentiellt olika väntevärden).

Minsta kvadrat-skattning

Definition. Minsta kvadrat-skattningen för $\boldsymbol{\theta}$ ges av den vektor $\widehat{\boldsymbol{\theta}}$ som minimerar

$$Q(\widehat{\boldsymbol{\theta}}) = \sum_{k=1}^{n} \Big(x_k - \mu_k(\widehat{\boldsymbol{\theta}})\Big)^2.$$

Exempel

Låt $X_1, \ldots, X_n$ vara ett slumpmässigt stickprov från en fördelning F. Hitta MK-skattningen för väntevärdet μ.

Lösning. Vi ställer upp funktionen

$$Q(\mu) = \sum_{k=1}^{n}(x_k - \mu)^2, \mu \in \mathbf{R}.$$

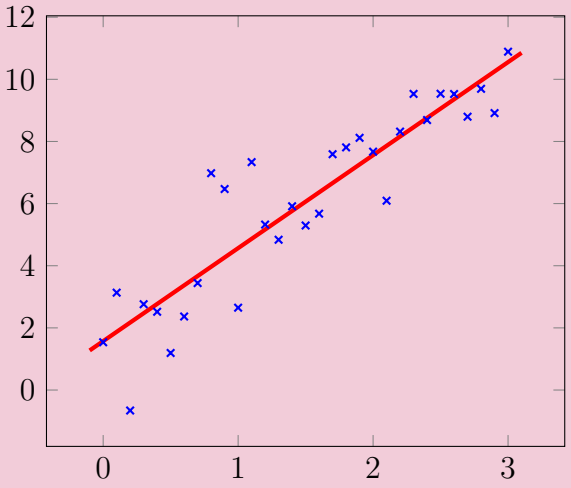
Vi söker nu det värde $\widehat{\mu}$ som minimerar Q. Enklast är att ta till envariabelanalysen och derivera och söka efter stationära punkter:

$$0 = Q'(\mu) = -2\sum_{k=1}^{n}(x_k - \mu) \quad \Leftrightarrow \quad n\mu = \sum_{k=1}^{n} x_k \quad \Leftrightarrow \quad \mu = \frac{1}{n}\sum_{k=1}^{n} x_k = \overline{x}.$$

Är detta ett minimum? Eftersom $Q''(\overline{x}) = 2n > 0$ är det mycket riktigt ett minimum. Den eftersökta MK-skattningen av väntevärdet är alltså $\widehat{\mu} = \overline{x}$.

Enkel linjär regression

Antag att vi gjort mätningar y_k på något vid vissa värden x_k, $k = 1, 2, \ldots, n$ och att ett spridningsdiagram visar något i stil med figuren till höger. Det förefaller rimligt att det föreligger ett approximativt linjärt samband. Kan vi hitta en linje som passar in i mätserien? vi söker alltså en linje $y = \beta_0 + \beta_1 x$ som i någon mening approximerar mätresultaten. I vilken mening? Där finns flera sätt, men det vanligaste är nog att minimera kvadraten i felen.



Lösning. Vi betraktar varje punkt (x_k, y_k) som att x_k är fixerad och y_k är en observation av en stokastisk variabel $Y = \beta_0 + \beta_1 x_k + \epsilon_k$ där ϵ_k är oberoende stokastiska variabler med $E(\epsilon_k) = 0$ och $V(\epsilon_k) = \sigma^2$. Detta är den typiska modellen vid linjär regression. Konstanterna β_0 och β_1 är okända och det är dessa vi vill bestämma. Eftersom

$$E(Y_k) = \beta_0 + \beta_1 x_k \qquad \text{och} \qquad V(Y_k) = \sigma^2$$

så blir

$$Q(\beta_0, \beta_1) = \sum_{k=1}^{n}\big(y_k - E(Y_k)\big)^2 = \sum_{k=1}^{n}\big(y_k - \beta_0 - \beta_1 x_k\big)^2.$$

Minimering av denna funktion med avseende på β_0 och β_1 ger skattningarna $\widehat{\beta}_0$ och $\widehat{\beta}_1$. Jakten på minimum sker nog enklast med lite flervariabelanalys:

$$\mathbf{0} = \nabla Q = (Q'_{\beta_0}, Q'_{\beta_1}) = -2\sum_{j=1}^{n}(y_j - \beta_0 - \beta_1 x_j, x_j(y_j - \beta_0 - \beta_1 x_j))$$

så

$$n\beta_0 + \beta_1 \sum_{j=1}^{n} x_j = \sum_{j=0}^{n} y_j \quad \Leftrightarrow \quad \beta_0 + \beta_1 \overline{x} = \overline{y}$$

och

$$\beta_0 \sum_{j=0}^{n} x_j + \beta_1 \sum_{j=0}^{n} x_j^2 = \sum_{j=1}^{n} x_j y_j \quad \Leftrightarrow \quad n\beta_0 \overline{x} + \beta_1 \sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} x_j y_j.$$

Första ekvationen ger att $\beta_0 = \overline{y} - \beta_1 \overline{x}$, så

$$n\overline{x}\,\overline{y} - \beta_1 n\overline{x}^2 + \beta_1 \sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} x_j y_j$$

vilket om vi löser ut β_1 leder till

$$\beta_1 = \frac{\sum_{j=1}^{n} x_j y_j - n\overline{x}\,\overline{y}}{\sum_{j=1}^{n} x_j^2 - n\overline{x}^2} = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

3.3 ML-skattning

Låt $X_1, X_2, \ldots, X_n$ vara oberoende stokastiska variabler med täthets- eller sannolikhetsfunktioner $f_i(x; \boldsymbol{\theta})$ respektive $p_i(k; \boldsymbol{\theta})$. Vi antar att samtliga endera är kontinuerliga eller diskreta. Det typiska är att alla variablerna har samma fördelning, men det är inget nödvändigt krav för metoden (däremot förenklar det så klart). Samtliga fördelningar beror dock på en och samma parameter $\boldsymbol{\theta}$ som kan vara vektorvärd.

ML-skattning

Definition. ML-skattningen för $\boldsymbol{\theta}$ är det värde som gör att **likelihood-funktionen** $L(\boldsymbol{\theta})$ maximeras, där

$$L(\boldsymbol{\theta}) = \prod_{k=1}^{n} f_k(x_k; \boldsymbol{\theta}) = f_1(x_1; \boldsymbol{\theta}) \cdot f_2(x_2; \boldsymbol{\theta}) \cdots f_n(x_n; \boldsymbol{\theta})$$

i det kontinuerliga fallet och

$$L(\boldsymbol{\theta}) = \prod_{k=1}^{n} p_k(x_k; \boldsymbol{\theta}) = p_1(x_1; \boldsymbol{\theta}) \cdot p_2(x_2; \boldsymbol{\theta}) \cdots p_n(x_n; \boldsymbol{\theta})$$

i det diskreta fallet.

Så vad är då ML-skattningen? Ganska enkelt är det den skattning som gör att det stickprov vi observerat är det mest troliga. Eftersom vi antar att variablerna som stickprovet är observationer av är oberoende ges den simultana täthets- eller sannolikhetsfunktionen av produkten av de marginella, så vi väljer helt enkelt den skattning som maximerar den simultana tätheten/sannolikheten.
Ofta när man arbetar med ML-skattningar nyttjar man den så kallade log-likelihood-funktionen:

$$l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}).$$

Denna funktion bevarar de flesta av de egenskaper vi är intresserade av eftersom ln är strängt växande och $L(\boldsymbol{\theta}) \in [0, 1]$. Specifikt så har $L(\boldsymbol{\theta})$ och $l(\boldsymbol{\theta})$ samma extrempunkter.

Exempel

Låt $x_1, x_2, \ldots, x_n$ vara ett stickprov av en exponentialfördelning med okänd intensitet θ. Hitta ML-skattningen för θ.

Lösning. Täthetsfunktionen ges av $f(x) = \theta e^{-\theta x}$, $x \geq 0$, så

$$L(\theta) = \prod_{k=1}^{n} \theta e^{-\theta x_k} = \theta^n \exp\left(-\theta \sum_{k=1}^{n} x_k\right) \quad \Rightarrow \quad l(\theta) = \ln L(\theta) = n \ln \theta - \theta \sum_{k=1}^{n} x_k.$$

Vi undersöker vart det finns extrempunkter och finner att

$$0 = l'(\theta) = \frac{n}{\theta} - \sum_{k=1}^{n} x_k \quad \Leftrightarrow \quad \theta = \frac{k}{\sum_{k=1}^{n} x_k} = \frac{1}{\overline{x}},$$

under förutsättning att $\overline{x} \neq 0$. Är detta ett maximum? Använd det ni lärt er i envariabelanalysen! Till exempel ser vi att

$$l''(\theta) = -\frac{n}{\theta^2},$$

så $l''(\theta) < 0$ för alla $\theta > 0$. Således är det ett maximum vi funnit.

Exempel

Låt $X \sim \text{Bin}(n, p)$ med p okänd och låt x vara en observation av X. Hitta ML-skattningen för p.

Lösning. Sannolikhetsfunktionen ges av $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, så

$$\begin{aligned} L(p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &\Rightarrow \quad l(p) = C(n,x) + x \ln p + (n-x)\ln(1-p), \end{aligned}$$

där $C(n,x)$ är en konstant (med avseende på p). Parameterrummet ges av $\Omega_p = (0,1)$. Vi deriverar och erhåller att

$$0 = l'(p) = \frac{x}{p} - \frac{n-x}{1-p} = \frac{(1-p)x - (n-x)p}{p(1-p)} = \frac{x - np}{p(1-p)} \quad \Leftrightarrow \quad x = n\,p \quad \Leftrightarrow \quad p = \frac{x}{n}.$$

ML-skattningen är således $\widehat{p} = \dfrac{x}{n}$ om detta är ett maximum. Vi kontrollerar:

		$\widehat{p}$	
$l'(p)$	$+$	0	$-$
$l(p)$	$\nearrow$	max	$\searrow$

Vad skulle hända om observationen blev $x = 0$ (eller $x = n$)?

Exempel

Låt $x_1, x_2, \ldots, x_n$ vara ett stickprov från $N(\mu, \sigma^2)$ där både μ och σ^2 är okända. Hitta ML-skattningarna för μ och σ^2.

Lösning. Vi har nu två okända parametrar och likelihoodfunktionen ges av

$$L(\mu, v) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(x_k-\mu)^2}{2v}\right) = \frac{1}{(2\pi v)^{n/2}} \exp\left(-\frac{1}{2v}\sum_{k=1}^{n}(x_k-\mu)^2\right),$$

där $v = \sigma^2$, så

$$l(\mu, v) = \text{konstant} - \frac{n}{2}\ln v - \frac{1}{2v}\sum_{k=1}^{n}(x_k-\mu)^2.$$

Parameterrummet ges av $\Omega_{\mu,v} = \mathbf{R} \times (0, \infty)$ och vi vill maximera $l(\mu, v)$. Stationära punkter finner vi där $\nabla l(\mu, v) = (0, 0)$, så vi beräknar de partiella derivatorerna:

$$l'_\mu(\mu, v) = \frac{1}{v}\sum_{k=1}^{n}(x_k-\mu) = \frac{n}{v}(\overline{x}-\mu)$$

och

$$l'_v(\mu, v) = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{k=1}^{n}(x_k-\mu)^2.$$

Det är tydligt att $\mu = \overline{x}$ och

$$\frac{n}{2v} = \frac{1}{2v^2}\sum_{k=1}^{n}(x_k-\mu)^2 \quad \Leftrightarrow \quad v = \frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)^2,$$

så $\nabla l = 0$ precis då

$$\mu = \overline{x} \quad \text{och} \quad v = \frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2.$$

Är detta ett maximum? Vi undersöker närmare:

$$H(\mu, v) = \begin{pmatrix} l''_{\mu\mu} & l''_{\mu v} \\ l''_{v\mu} & l''_{vv} \end{pmatrix} = \begin{pmatrix} -\frac{n}{v} & -\frac{n}{v^2}(\overline{x}-\mu) \\ -\frac{n}{v^2}(\overline{x}-\mu) & \frac{n}{2v^2} - \frac{1}{v^3}\sum_{k=1}^{n}(x_k-\mu)^2 \end{pmatrix},$$

där vi låter $\text{SS} = \sum_{k=1}^{n}(x_k-\mu)^2$ och i punkten $(\mu, v) = \left(\overline{x}, \frac{1}{n}\text{SS}\right)$ blir

$$H\left(\overline{x}, \frac{1}{n}\text{SS}\right) = \begin{pmatrix} -\frac{n^2}{SS} & 0 \\ 0 & \frac{n^3}{2SS^2} - \frac{n^3}{SS^3}SS \end{pmatrix} = \begin{pmatrix} -\frac{n^2}{SS} & 0 \\ 0 & -\frac{n^3}{2SS^2} \end{pmatrix},$$

vilket är en negativt definit matris, så detta är ett maximum.
Vi vet sedan tidigare att skattningen för v behöver ha faktorn $1/(n-1)$ för att vara väntevärdesriktig, så ML-skattningen av σ^2 är således inte väntevärdesriktig.

4 Flera stickprov; sammanvägd variansskattning

Antag att vi har två stickprov $x_1, x_2, \ldots, x_m$ och $y_1, y_2, \ldots, y_n$ från normalfördelningar med olika väntevärde men samma varians. ML-skattningarna för respektive väntevärde blir $\widehat{\mu}_1 = \overline{x}$ respektive $\widehat{\mu}_2 = \overline{y}$. För standardavvikelsen kan man visa att den **sammanvägda variansskattningen** (*pooled variance*) blir

$$s^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{n+m-2},$$

där s_1^2 och s_2^2 är stickprovsvarianserna för respektive stickprov. Formeln generaliserar naturligt till fler stickprov. Vi kan även direkt se att

$$E(S^2) = \frac{1}{m+n-2}\left((m-1)E(S_1^2) + (n-1)E(S_2^2)\right) = \frac{1}{m+n-2}\left((m+n-2)\sigma^2\right) = \sigma^2,$$

så skattningen är väntevärdesriktig.

5 Medelfel

Vi har använt variansen $V(\widehat{\Theta})$ (eller standardavvikelsen $D(\widehat{\Theta})$) för att jämföra olika skattningar (effektivitet och konsistens). Mindre varians betyder helt enkelt att skattningen i någon mening är bättre. Detta är ett problem då dessa storheter i allmänhet inte är kända. Vad vi gör är att vi helt enkelt skattar de okända storheterna i $D(\widehat{\Theta})$ och kallar resultatet för medelfelet.

Medelfel

Definition. En skattning $d = d(\widehat{\Theta})$ av standardavvikelsen $D(\widehat{\Theta})$ kallas för skattningens **medelfel**.

Vi ersätter alltså helt enkelt okända storheter i $V(\widehat{\Theta})$ med skattningar. Givetvis påverkar detta precisionen och sättet vi väljer att ersätt de okända storheterna har inverkan på resultatet.

Exempel

Om $X_1, \ldots, X_n$ är ett slumpmässigt stickprov av en $N(\mu, \sigma^2)$-fördelning där både μ och σ^2 är okända kan vi uppskatta μ med medelvärdet $\widehat{M} = \overline{X}$. Således är $D(M) = \dfrac{\sigma}{\sqrt{n}}$, men då σ är okänd behöver vi skatta σ med något. Förslagsvis med stickprovsstandardavvikelsen s, vilket ger medelfelet

$$d(\widehat{M}) = \frac{s}{\sqrt{n}}.$$

Detta är inte på något sätt unikt. En annan skattning av σ ger ett annat medelfel. Med det sagt är detta ett ganska naturligt val för medelfelet.

Ett annat vanligt exempel är vid skattningar av andel. Ofta gör vi som i följande exempel.

Exempel

Ett annat vanligt exempel är när p ska skattas i binomialfördelning. Låt $X \sim \text{Bin}(n, p)$. Vi vet att $V(X) = np(1-p)$ så om vi skattar p med $\widehat{P} = \dfrac{X}{n}$ erhåller vi att $D(\widehat{P}) = \sqrt{\dfrac{p(1-p)}{n}}$. Eftersom p är okänd känner vi inte denna storhet exakt, men medelfelet skulle bli

$$d(\widehat{P}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Föreläsning 3: Konfidensintervall

Johan Thim (johan.thim@liu.se)

5 september 2018

"[we are] Explorers in the further regions of experience. Demons to some. Angels to others."
–Pinhead

1 Intervallskattningar

Vi har nu studerat hur man mer eller mindre systematiskt kan hitta skattningar för okända parametrerar när vi har stickprov från en fördelning som beror på parametern. Den naturliga följdfrågan är givetvis hur "bra" skattningen är. Vi har vissa mått i form av väntevärdesriktighet, konsistens och effektivitet, men går det att säga något med en given sannolikhet? Kan vi hitta ett intervall som med en viss given sannolikhet måste innehålla den okända parametern?

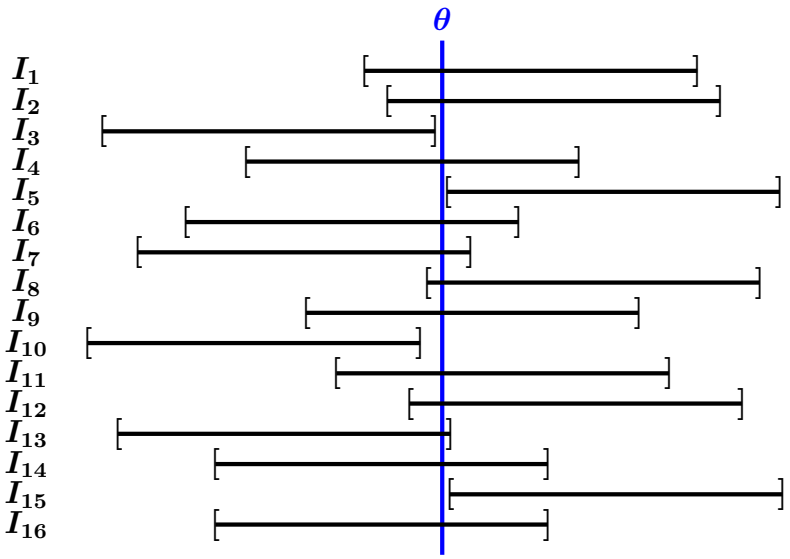
Konfidensintervall

Definition. Låt $x_1, \ldots, x_n$ vara ett stickprov av en fördelning som beror på en okänd parameter θ och låt $\alpha \in [0,1]$. Ett intervall $I_\theta^{1-\alpha} = (\widehat{\theta}_L, \widehat{\theta}_U)$ kallas för ett **konfidensintervall** för θ med **konfidensgrad** $1-\alpha$ om

$$P(\widehat{\Theta}_L < \theta < \widehat{\Theta}_U) = 1-\alpha.$$

Gränserna $\widehat{\theta}_L = a(x_1, \ldots, x_n)$ och $\widehat{\theta}_U = b(x_1, \ldots, x_n)$ är skattningar som beräknas från stickprovet. Dessa ändpunkter kallas **konfidensgränser**.

Så hur fungerar detta i praktiken? Säg att vi har tillgång till 100 olika stickprov $\boldsymbol{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$ från en och samma fördelning som beror på samma okända parameter θ. Vi hittar konfidensintervall för alla 100 stickproven med konfidensgrad $1-\alpha$. Då kommer $100\cdot(1-\alpha)$ av dessa intervall att innehålla θ (i snitt).



Av 16 intervall är det 4 som inte innehåller det verkliga värdet på θ. Så med andra ord verkar det som att ungefär $12/16 = 3/4$ av intervallen innehåller det verkliga värdet på θ. Detta innebär att konfidensgraden vid skattningen är ungefär 75%.

Notera att det är gränserna i konfidensintervallet som är stokastiska variabler (eller skattningar därav). Storheten θ är okänd (och behöver inte ens ligga i intervallet).

Inga intervall är mer värda

Vi kan inte säga att till exempel I_9 är ett ”bättre” intervall än I_{13}, utan det är en binär fråga: gäller det att $\theta \in I_k$ eller inte.

Så då kommer vi till nästa rimliga fråga: hur hittar vi systematiskt konfidensintervall med given konfidensgrad?

Konstruktion av konfidensintervall

1. Ställ upp en lämplig skattningsvariabel $\widehat{\Theta}$ för θ. Här kan vi använda de metoder vi tagit fram tidigare (moment-, MK- och ML-skattningar till exempel).

2. Konstruera en hjälpvariabel H (teststorhet) utifrån $\widehat{\Theta}$. Hjälpvariabeln får endast innehålla kända storheter utöver θ (och om θ förekommer flera gånger kan vi behöva skatta bort en del instanser för att få något användbart).

3. Stäng in hjälpvariabeln i ett intervall $I = (c, d)$ så att $P(c < H < d) = 1 - \alpha$.

4. Lös ut θ ur olikheten $c < H < d$:

$$c < H < d \quad \Leftrightarrow \quad a(X_1, \ldots, X_n) < \theta < b(X_1, \ldots, X_n)$$

   vilket ger att $P(a(X_1, \ldots, X_n) < \theta < b(X_1, \ldots, X_n)) = 1 - \alpha$.

5. Ersätt de stokastiska storheterna $X_1, \ldots, X_n$ med observationerna $x_1, \ldots, x_n$ vilket ger intervallet

$$I_\theta^{1-\alpha} = (a(x_1, \ldots, x_n), b(x_1, \ldots, x_n)).$$

Något som kommer bli viktigt är följande definition från sannolikhetsteorin.

Kvantil

Definition. En α-kvantil λ_α för en stokastisk variabel X är ett tal λ_α sådant att

$$P(X > \lambda_\alpha) = \alpha.$$

Vi finner ofta kvantiler i tabell endera genom en explicit kvantiltabell eller genom att söka upp sannolikheten $1 - \alpha$ och identifiera (approximativt) vilket värde på x som gör att vi erhåller $F(x) = 1 - \alpha$, där F är fördelningsfunktionen. Saknar vi tabell får vi istället lösa ekvationen

$$1 - \alpha = \int_{-\infty}^{\lambda_\alpha} f_X(x)\, dx.$$

Observera att svaret inte nödvändigtvis är entydigt.

2 χ^2-fördelningen

En situation som dyker upp frekvent i statistik inferens är summor av kvadrater av normalfördelade variabler, så en naturlig fråga är så klart vilken fördelning en sådan summa får (åtminstone då variablerna antas vara oberoende). Svaret fås i form av χ^2-fördelningen.

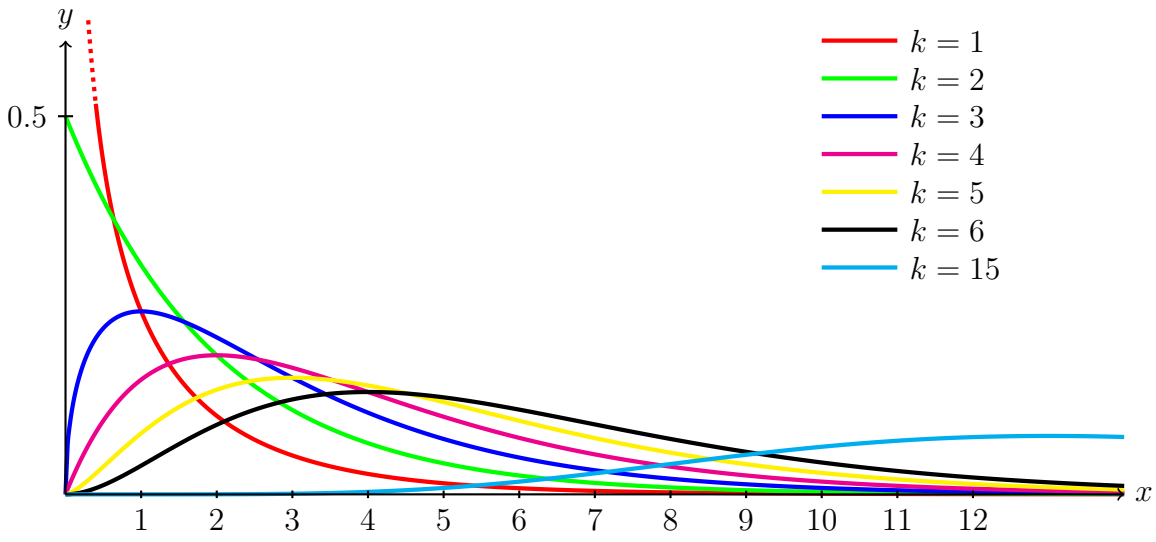
χ^2-fördelning

Definition. Om X är en stokastisk variabel med täthetsfunktionen

$$f_X(x) = \frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1}e^{-x/2}, \quad x \geq 0 \text{ om } k > 1,$$

kallar vi X för $\chi^2(k)$-fördelad med k frihetsgrader, där $k = 1, 2, \ldots$

Här är Γ gamma-funktionen[1] och $\Gamma(n) = (n-1)!$ och $\Gamma(n+1/2) = \dfrac{(2n)!}{4^n\, n!}\sqrt{\pi}$ om $n \in \mathbf{N}$.



Om $X \sim \chi^2(k)$ är $E(X) = k$ och $V(X) = 2k$.

Bevis. Låt täthetsfunktionen skrivas $f(x) = cx^{k/2-1}e^{-x/2}$. Då gäller att

$$\begin{aligned} E(X) &= c\int_0^\infty x^{k/2}e^{-x/2}dx = c\left(\left[-2x^{k/2}e^{-x/2}\right]_0^\infty + 2\int_0^\infty \frac{k}{2}x^{k/2-1}e^{-x/2}\,dx\right) \\ &= k\int_0^\infty f(x)\,dx = k. \end{aligned}$$

På samma sätt följer att

$$E(X^2) = c\int_0^\infty x^{k/2+1}e^{-x/2}\,dx = 2\left(\frac{k}{2}+1\right)c\int_0^\infty x^{k/2}e^{-x/2}\,dx = (k+2)E(X) = k^2+2k,$$

så $V(X) = E(X^2) - E(X)^2 = k^2 + 2k - k^2 = 2k$. □

[1] Se avsnitt 10 nedan för mer detaljer.

Sats. Om $X \sim \chi^2(\nu_1)$ och $Y \sim \chi^2(\nu_2)$ är oberoende så är $X + Y \sim \chi^2(\nu_1 + \nu_2)$.

Bevis. Enklast är att betrakta Fouriertransformen för täthetsfunktionen (alternativt den närbesläktade **karakteristiska funktionen** definierad enligt $E(e^{itX})$). Det är nämligen så att

$$\mathcal{F}(f_X)(t) = (1+2it)^{-\nu_1/2}, \quad \mathcal{F}(f_Y)(t) = (1+2it)^{-\nu_2/2}$$

och

$$\mathcal{F}(f_X * f_Y) = \mathcal{F}(f_X)\mathcal{F}(f_Y) = (1+2it)^{-(\nu_1+\nu_2)/2},$$

så $f_{X+Y} \sim \chi^2(\nu_1 + \nu_2)$. □

Sats. Om $X_1, X_2, \ldots, X_n$ är oberoende och $X_k \sim N(0,1)$ så är

$$\sum_{k=1}^{n} X_k^2 \sim \chi^2(n).$$

Bevis. Eftersom variablerna är oberoende ges den simultana täthetsfunktionen av

$$f(x_1, \ldots, x_n) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} = \frac{1}{2^{n/2}\pi^{n/2}} \exp\left(-\frac{1}{2}\left(x_1^2 + \cdots + x_n^2\right)\right).$$

Vi söker fördelningen för $Z = X_1^2 + \cdots X_n^2$, så låt oss ställa upp fördelningsfunktionen:

$$\begin{aligned} F_Z(z) = P(Z \le z) &= \int_{x_1^2+\cdots x_n^2 \le z} f(x_1, \ldots, x_n) dx_1\, dx_2 \cdots dx_n \\ &= \frac{1}{2^{n/2}\pi^{n/2}} \int_{S^{n-1}} \int_0^{\sqrt{z}} r^{n-1} e^{-r^2/2}\, dr\, dS \\ &= \frac{1}{2^{n/2}\pi^{n/2}} \frac{2\pi^{n/2}}{\Gamma(n/2)} \int_0^{\sqrt{z}} r^{n-1} e^{-r^2/2}\, dr \\ &= \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^z t^{n/2-1} e^{-t/2}\, dt, \end{aligned}$$

där S^{n-1} är enhetssfären i $\mathbf{R}^n$ och dS är ytmåttet på S^{n-1}. Då enhetssfären har ytmåttet $|S^{n-1}| = \dfrac{2\pi^{n/2}}{\Gamma(n/2)}$ följer likheten ovan efter ett variabelbyte i sista integralen (låt $t = r^2$). Analysens huvudsats medför nu att (för $z > 0$) att

$$f_Z(z) = F_Z'(z) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}.$$

För $z < 0$ är givetvis $f_Z(z) = 0$ (varför?). □

3 t-fördelningen

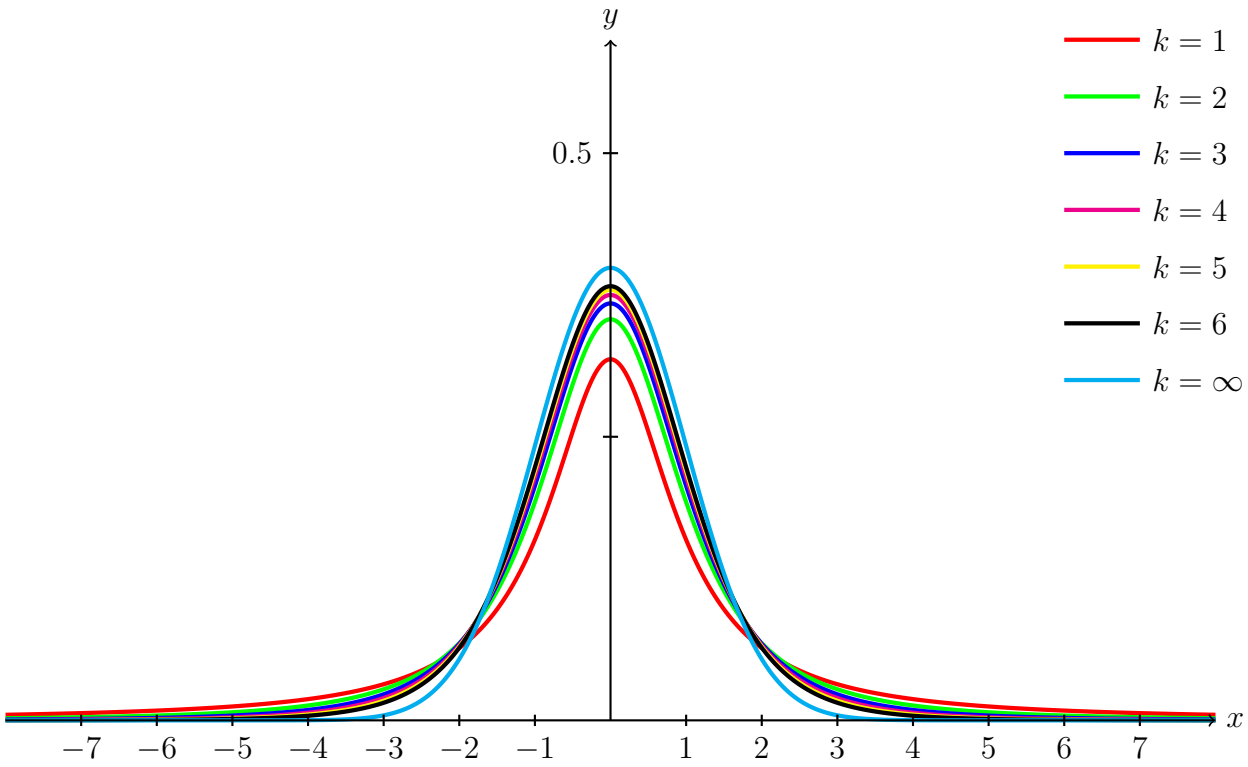
t-fördelning

Definition. Om X är en stokastisk variabel med täthetsfunktionen

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbf{R} \text{ och } \nu > 0,$$

kallar vi X för $t(\nu)$-fördelad med ν frihetsgrader.

Denna fördelning är symmetrisk och om antalet frihetsgrader går mot oändligheten konvergerar täthetsfunktionen mot täthetsfunktionen för normalfördelning.



Sats. Om $X \sim t(\nu)$ är $E(X) = 0$ (om $\nu > 1$) och $V(X) = \nu/(\nu-2)$ (om $\nu > 2$).

Bevis. Om $\nu > 1$ är integralen $E(X)$ absolutkonvergent (visa det) och då integranden är udda blir således $E(X) = 0$. För att beräkna $E(X^2)$ låter vi $c_\nu = \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}$.

Om $\nu > 2$ ser vi genom partialintegration att

$$\begin{aligned} E(X^2) &= c_\nu \int_{-\infty}^{\infty} x \cdot x \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx = c_\nu \int_{-\infty}^{\infty} \frac{\nu}{\nu-1}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu-1}{2}} dx \\ &= c_\nu \frac{\nu}{\nu-1}\int_{-\infty}^{\infty}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu-1}{2}} dx \\ &= \frac{c_\nu}{c_{\nu-2}}\,\frac{\nu^{3/2}}{(\nu-1)\sqrt{\nu-2}}\, c_{\nu-2}\int_{-\infty}^{\infty}\left(1+\frac{u^2}{\nu-2}\right)^{-\frac{\nu-1}{2}} du = \frac{c_\nu}{c_{\nu-2}}\,\frac{\nu^{3/2}}{(\nu-1)\sqrt{\nu-2}} \end{aligned}$$

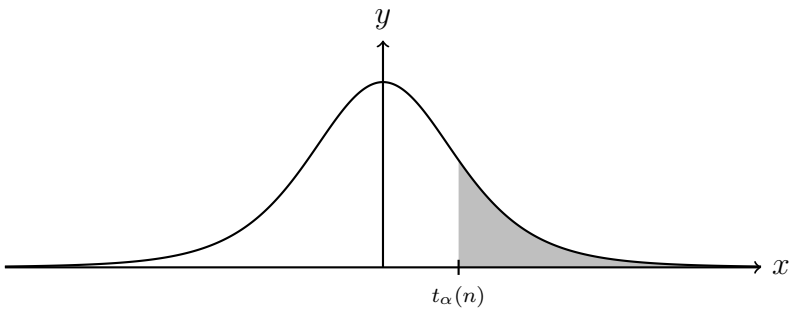
där vi bytte variabel så $x\sqrt{\nu-2} = u\sqrt{\nu}$ och utnyttjade att integralen som dök upp är precis integralen av täthetsfunktionen för en $t(\nu-2)$-fördelad variabel (om $\nu > 2$). Vi förenklar uttrycket och finner att

$$\begin{aligned}\frac{c_\nu}{c_{\nu-2}}\,\frac{\nu^{3/2}}{(\nu-1)\sqrt{\nu-2}} &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)\Gamma\left(\frac{\nu-2}{2}\right)\sqrt{(\nu-2)\pi}}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{\nu-1}{2}\right)}\,\frac{\nu^{3/2}}{(\nu-1)\sqrt{\nu-2}} \\ &= \frac{\frac{\nu-1}{2}\Gamma\left(\frac{\nu-1}{2}\right)\Gamma\left(\frac{\nu-2}{2}\right)}{\frac{\nu-2}{2}\Gamma\left(\frac{\nu-2}{2}\right)\Gamma\left(\frac{\nu-1}{2}\right)}\,\frac{\nu}{\nu-1} = \frac{\nu}{\nu-2},\end{aligned}$$

där vi nyttjat att $\Gamma(z+1) = z\Gamma(z)$. Eftersom $E(X) = 0$ följer det nu att $V(X) = E(X^2)$. $\square$

3.1 *t*-fördelningens kvantiler

Kvantilerna för t-fördelningen är de tal $t_\alpha(n)$ sådana att $P(T > t_\alpha(n)) = 1 - \alpha$. Det vill säga gränser $t_\alpha(n)$ sådana att för $T \sim t(n)$ gäller att andelen α av sannolikhetsmassan ligger till höger om $t_\alpha(n)$. Eftersom gränserna är jobbiga att räkna fram för hand brukar vi använda tabellverk enligt nedan (studera även formelsamlingen).



n \ α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	1.292	1.664	1.990	2.374	2.639	3.195	3.416
90	1.291	1.662	1.987	2.368	2.632	3.183	3.402
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

4 Vektorer med stokastiska variabler

Låt $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ vara en vektor vars komponenter är stokastiska variabler. Vi strävar efter att skriva vektorer som kolonnvektorer. Det faller sig naturligt att definera väntevärdet av $\boldsymbol{X}$ genom

$$E(\boldsymbol{X}) = (E(X_1), E(X_2), \ldots, E(X_n)).$$

På samma sätt definerar vi väntevärdet av en matris av stokastiska variabler. Variansen blir lite konstigare så vi introducerar den så kallade kovariansmatrisen mellan två vektorer (av samma dimension). Låt $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ och definiera $C(\boldsymbol{X}, \boldsymbol{Y})$ enligt

$$C(\boldsymbol{X}, \boldsymbol{Y}) = \begin{pmatrix} C(X_1, Y_1) & C(X_1, Y_2) & \cdots & C(X_1, Y_n) \\ C(X_2, Y_1) & C(X_2, Y_2) & \cdots & C(X_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(X_n, Y_1) & C(X_n, Y_2) & \cdots & C(X_n, Y_n) \end{pmatrix}$$

där $C(X_i, Y_j) = E(X_i Y_j) - E(X_i)E(Y_j)$ är kovariansen mellan X_i och Y_j.
En stor anledning att blanda in vektorer och matriser är givetvis att få tillgång till maskineriet från linjär algebra. Kovariansen (en matris) mellan två vektorer $\boldsymbol{X}$ och $\boldsymbol{Y}$ kan då lite mer kompakt skrivas

$$C(\boldsymbol{X}, \boldsymbol{Y}) = E(\boldsymbol{X}\boldsymbol{Y}^T) - E(\boldsymbol{X})E(\boldsymbol{Y})^T,$$

där $(\,\cdot\,)^T$ innebär transponering. En produkt $A = \boldsymbol{x}\boldsymbol{y}^T$ brukar kallas för den yttre produkten och består av element $(a)_{ij} = x_i y_j$, $i, j = 1, 2, \ldots, n$. Detta är alltså *inte* skalärprodukten ($\boldsymbol{X}^T\boldsymbol{Y}$).
Låt $A, B \in \mathbf{R}^{n\times n}$ vara matriser. Då är $A\boldsymbol{X}$ en linjärkombination av $X_1, X_2, \ldots, X_n$ och $B\boldsymbol{Y}$ en linjärkombination av $Y_1, Y_2, \ldots, Y_n$. Dessutom kan *alla* linjärkombinationer skrivas på detta sätt. Vidare gäller nu tack varje linjäriteten att

$$E(A\boldsymbol{X}) = AE(\boldsymbol{X}) \quad \text{och} \quad C(A\boldsymbol{X}, B\boldsymbol{Y}) = A\boldsymbol{X}(B\boldsymbol{Y})^T = A\boldsymbol{X}\boldsymbol{Y}^T B^T.$$

5 Cochrans sats

Vi ska nu betrakta ett speciallfall av en ganska generell sats (Cochrans sats) ..

Sats. Låt $X_1, X_2, \ldots, X_n$ vara oberoende likafördelade stokastiska variabler där $X_k \sim N(\mu, \sigma^2)$ för $k = 1, 2, \ldots, n$. Då gäller att

$$\frac{1}{\sigma^2}\sum_{k=1}^{n}(X - \overline{X})^2 \sim \chi^2(n-1).$$

Bevis. Låt $Y_k = X_k - \mu$ så att $Y_k \sim N(0, \sigma^2)$. Vi ser att

$$\sum_{k=1}^{n}(X_k - \overline{X})^2 = \sum_{k=1}^{n}(Y_k - \overline{Y})^2.$$

Låt J vara $n \times n$-matrisen vars samtliga element är 1 och låt $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. Då kan vi skriva

$$\begin{pmatrix} Y_1 - \overline{Y} \\ Y_2 - \overline{Y} \\ \vdots \\ Y_n - \overline{Y} \end{pmatrix} = I\boldsymbol{Y} - \frac{1}{n}J\boldsymbol{Y} = Q\boldsymbol{Y},$$

där $Q = I - (1/n)J$. Låt $P = I - Q = (1/n)J$. Då är

$$P + Q = I, \qquad P^2 = P^T = P, \qquad Q^2 = Q^T = Q \qquad \text{samt } PQ = QP = 0.$$

Matriserna P och Q representerar alltså ortogonala projektioner på $\mathbf{R}^n$ och av naturliga skäl är $\text{rank}(P) = 1$ så $\text{rank}(Q) = n - 1$ (eftersom $P + Q = I$).
Vidare gäller att

$$C(P\boldsymbol{Y}, Q\boldsymbol{Y}) = PC(\boldsymbol{Y}, \boldsymbol{Y})Q^T = P\sigma^2 I Q^T = PQ^T = PQ = 0$$

då $E(\boldsymbol{Y}) = 0$ och $C(Y_i, Y_j) = \sigma^2$ om $i = j$ och $C(Y_i, Y_j) = 0$ då $i \neq j$ eftersom olika Y_k är oberoende. Således är $Y_i - \overline{Y}$ och $\overline{Y}$ oberoende stokastiska variabler (eftersom kovariansen noll mellan normalfördelade variabler är ekvivalent med oberoende). Eftersom $\text{rank}(Q) = n - 1$ så kan vi representera $Q\boldsymbol{Y}$ i en ortogonal bas så att

$$\frac{1}{\sigma^2}\sum_{k=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{\sigma^2}(Q\boldsymbol{Y})^T Q\boldsymbol{Y} = \frac{1}{\sigma^2}\boldsymbol{Y}^T Q\boldsymbol{Y} = Z_1^2 + Z_2^2 + \cdots + Z_{n-1}^2,$$

där $Z_k \sim N(0,1)$ och dessa variabler är oberoende. Vi kan nu nyttja den tidigare satsen om att summan av n stycken kvadrater av $N(0,1)$-fördelade variabler är $\chi^2(n)$-fördelad för att dra slutsatsen att $\dfrac{1}{\sigma^2}\boldsymbol{X}^T Q\boldsymbol{X} \sim \chi^2(n-1)$. □

6 t- och χ^2-fördelning; Gossets sats

Det finns givetvis en anledning till att vi studerar just dessa två fördelningar. William Gosset bevisade nämligen följande sats.

Sats. Låt $Z \sim N(0,1)$ och $V \sim \chi^2(\nu)$ vara oberoende. Då är $\dfrac{Z}{\sqrt{V/\nu}} \sim t(\nu)$.

Bevis. Eftersom Z och V är oberoende ges den simultana täthetsfunktionen av

$$f(z,v) = \frac{1}{\sqrt{2\pi}\,2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} e^{-z^2/2} v^{\nu/2-1} e^{-v/2}, \quad z \in \mathbf{R},\ v \geq 0.$$

Låt $T = \dfrac{Z}{\sqrt{V/\nu}}$ och $c = \dfrac{1}{\sqrt{2\pi}\,2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}$. Vi söker täthetsfunktionen f_T för T. Betrakta

$$P(T \leq t) = \iint_{z/\sqrt{v/\nu} \leq t} f(z,v)\,dz\,dv \quad = c \iint_{z/\sqrt{v/\nu} \leq t} e^{-z^2/2} v^{\nu/2-1} e^{-v/2}\,dz\,dv.$$

Vi gör ett variabelbyte,

$$\begin{cases} u\sqrt{\dfrac{v}{\nu}}, \\ w = v \end{cases} \qquad \Rightarrow \qquad \frac{d(z,v)}{d(u,w)} = \begin{vmatrix} \sqrt{\frac{w}{\nu}} & \frac{u}{2\nu\sqrt{\frac{w}{\nu}}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{w}{\nu}},$$

så integralen blir

$$c\iint_{u\le t}\sqrt{\frac{w}{\nu}}e^{-u^2w/(2\nu)}w^{\nu/2-1}e^{-v/2}\,dz\,dv = \frac{c}{\sqrt{\nu}}\int_{-\infty}^{t}\int_0^{\infty} w^{\frac{\nu+1}{2}-1}\exp\left(-\frac{w}{2}\left(1+\frac{u^2}{\nu}\right)\right)dw\,du$$
$$= \frac{c}{\sqrt{\nu}}\int_{-\infty}^{t}\int_0^{\infty}\frac{r^{\frac{\nu+1}{2}-1}}{\left(1+\frac{u^2}{\nu}\right)^{\frac{\nu+1}{2}}}e^{-\frac{r}{2}}dr\,du,$$
$$= \frac{c}{\sqrt{\nu}}\int_{-\infty}^{t}\left(1+\frac{u^2}{\nu}\right)^{-\frac{\nu+1}{2}}\int_0^{\infty} r^{\frac{\nu+1}{2}-1}e^{-\frac{r}{2}}dr\,du,$$

där vi gjorde ett variabelbyte $r = w\left(1+\frac{u^2}{\nu}\right)$ i den innersta integralen och bröt ut den faktor som inte beror på u. Den innersta integralen är nu nästan (upp till normeringskonstanten) integralen av täthetsfunktionen för en $\chi^2(\nu+1)$-variabel, så

$$\int_0^{\infty} r^{\frac{\nu+3}{2}-1}e^{-\frac{r}{2}}dr = 2^{(\nu+1)/2}\Gamma\left(\frac{\nu+1}{2}\right).$$

Således ges fördelningsfunktionen

$$F_T(t) = \frac{2^{(\nu+1)/2}\,\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\sqrt{2\pi}\,2^{\nu/2}\,\Gamma\left(\frac{\nu}{2}\right)}\int_{-\infty}^{t}\left(1+\frac{u^2}{\nu}\right)^{-\frac{\nu+1}{2}}du = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\int_{-\infty}^{t}\left(1+\frac{u^2}{\nu}\right)^{-\frac{\nu+1}{2}}du$$

vilket efter derivering ger täthetsfunktionen

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

vilket är precis täthetsfunktionen för en $t(\nu)$-fördelad variabel. □

Följdsats. Om $X_1, X_2, \ldots, X_n$ är oberoende och likafördelade med fördelningen $N(\mu,\sigma^2)$ så är $T = \dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$, där S^2 är stickprovsvariansen.

Bevis. Detta följer direkt från föregående resultat och Cochrans sats. Vi kan formulera T enligt

$$T = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\cdot\frac{1}{\frac{1}{\sigma}S} = \frac{Z}{\sqrt{\frac{V}{n-1}}},$$

där $Z \sim N(0,1)$ och $V = \dfrac{1}{\sigma^2}\sum_{k=1}^{n}(X_k-\overline{X})^2 \sim \chi^2(n-1)$ (med $S^2 = \dfrac{1}{n-1}V$). □

7 Konfidensintervall för μ och σ i normalfördelning

7.1 Konfidensintervall för μ när σ är känd

Låt $x_1, x_2, \ldots, x_n$ vara ett stickprov från en $N(\mu,\sigma^2)$-fördelning där vi känner σ och vill hitta ett konfidensintervall för μ. En punktskattning för väntevärdet ges av

$$\widehat{M} = \bar{X} = \frac{1}{n}\sum_{k=1}^{n}X_k \sim N(\mu,\sigma^2/n).$$

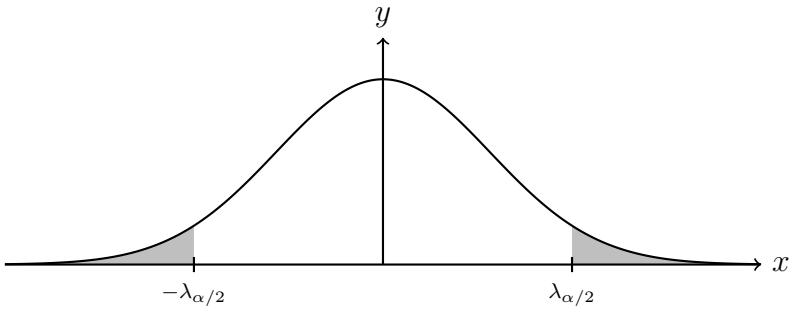
Vi skapar testvariabeln

$$Z = \frac{\widehat{M} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Det följer då att vi kan välja ett tal $\lambda_{\alpha/2}$ så att

$$P(-\lambda_{\alpha/2} < Z < \lambda_{\alpha/2}) = 1 - \alpha. \tag{1}$$

Talet $\lambda_{\alpha/2}$ är $\alpha/2$-kvantilen för en $N(0,1)$-fördelning och ges av $\lambda_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. Eftersom vi saknar explicit uttryck för denna invers är det enklast (utan dator åtminstone) att slå i tabell. Standardtabell som finns i formelsamlingen enligt nedan (vi får uttnyttja symmetri för att finna sannolikheter mindre än 0.5).



Det skuggade områdena är sannolikheten att $P(Z < -\lambda_{\alpha/2}) + P(Z > \lambda_{\alpha/2})$.

Vi löser ut μ ur olikheten i sannolikhetsmåttet i ekvation (1) ovan:

$$\begin{aligned} -\lambda_{\alpha/2} < Z < \lambda_{\alpha/2} \quad &\Leftrightarrow \quad -\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \widehat{M} - \mu < \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &\Leftrightarrow \quad \widehat{M} - \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \widehat{M} + \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

Om vi ersätter $\widehat{M}$ med den observerade punktskattningen $\widehat{\mu} = \overline{x}$ (medelvärdet av observationerna) så får vi ett konfidensintervall

$$I_\mu = \left(\overline{x} - \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{x} + \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right)$$

med konfidensgrad $1-\alpha$.

Exempel

Vid en mätning av en process fick man följande mätdata:

6.04 4.96 4.93 3.40 7.04 4.73 3.57 7.70 4.55 3.82

Antag att mätningarna är ett stickprov på en normalfördelad variabel $X \sim N(\mu, 2^2)$ (man tycker sig veta så pass mycket om processen att standardavvikelsen anses vara känd). Beräkna ett 99% konfidensintervall för väntevärdet μ.

Lösning. Vi betraktar siffrorna som ett stickprov från oberoende s. v. $X_j \sim N(\mu, \sigma^2 = 4)$. Vi punktskattar väntevärdet μ med

$$\widehat{M} = \bar{X} = \frac{1}{10}\sum_{j=1}^{10} X_j \sim N(\mu, 4/10).$$

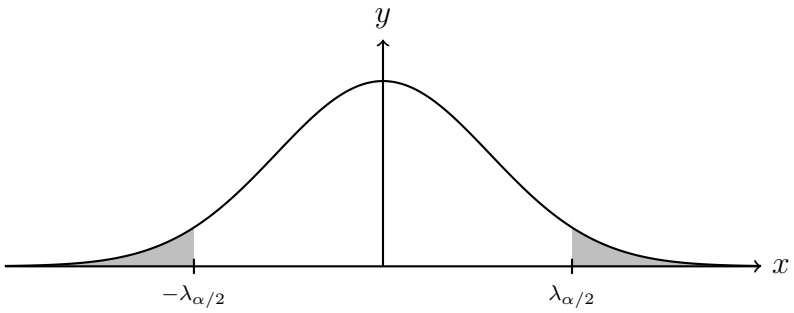
Vi skapar testvariabeln

$$Z = \frac{\widehat{M} - \mu}{\sigma/\sqrt{10}} \sim N(0,1).$$

Det följer då att

$$P(-\lambda_{\alpha/2} < Z < \lambda_{\alpha/2}) = 1 - \alpha, \tag{2}$$

och då vi söker ett 99% konfidensintervall så är $\alpha = 0.01$ och $\lambda_{0.005} \approx 2.575$ (det sista ur tabell).

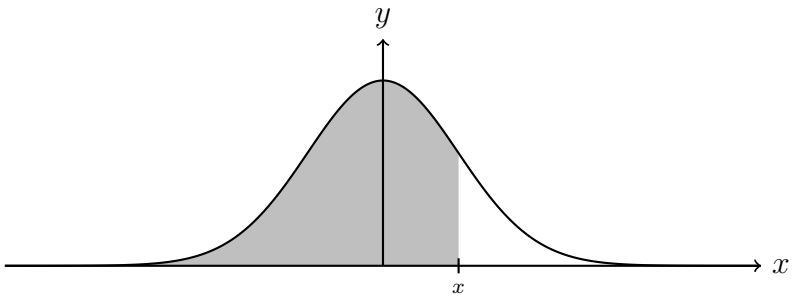


Det skuggade området är $\alpha \cdot 100\%$ av sannolikhetsmassan jämt fördelad på svansarna.
Vi löser ut μ ur olikheten i sannolikhetsmåttet i ekvation (2) ovan och erhåller att

$$\widehat{M} - \frac{2.575 \cdot 2}{\sqrt{10}} < \mu < \widehat{M} + \frac{2.575 \cdot 2}{\sqrt{10}}.$$

Om vi ersätter $\widehat{M}$ med den observerade punktskattningen $\widehat{\mu} = 5.074$ (medelvärdet av observationerna) så får vi ett konfidensintervall $I_\mu = (3.45, 6.70)$ med konfidensgrad 99%.

7.1.1 Normalfördelningstabell



Det skuggade området är sannolikheten att $P(Z \leq x)$, där $Z \sim N(0,1)$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

7.2 Okänd varians

Låt $x_1, x_2, \ldots, x_n$ vara ett stickprov från en $N(\mu, \sigma^2)$-fördelning där vi *inte* vet vad σ är och vi vill hitta ett konfidensintervall för μ. En punktskattning för väntevärdet ges av

$$\widehat{M} = \bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \sim N(\mu, \sigma^2/n).$$

Eftersom σ är okänd behöver vi en skattning och förslagsvis väljer vi stickprovsstandardavvikelsen. Vi skapar sedan testvariabeln

$$T = \frac{\widehat{M} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

där faktumet att T är t-fördelad följer från Gossets sats. Det följer då att vi kan välja ett tal $t_{\alpha/2}$ så att

$$P(-t_{\alpha/2}(n-1) < T < t_{\alpha/2}(n-1)) = 1 - \alpha. \tag{3}$$

Talet $t_{\alpha/2}(n-1)$ är $\alpha/2$-kvantilen för en $t(n-1)$-fördelning (vi finner denna i tabell). Vi löser ut μ ur olikheten i sannolikhetsmåttet i ekvation (3) ovan:

$$\begin{aligned} -t_{\alpha/2}(n-1) < T < t_{\alpha/2}(n-1) \quad &\Leftrightarrow \quad -t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} < \widehat{M} - \mu < t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \\ &\Leftrightarrow \quad \widehat{M} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} < \mu < \widehat{M} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \end{aligned}$$

Om vi ersätter $\widehat{M}$ med den observerade punktskattningen $\widehat{\mu} = \overline{x}$ (medelvärdet av observationerna) och S med stickprovsstandardavvikelsen så får vi ett konfidensintervall

$$I_\mu = \left(\overline{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} \right)$$

med konfidensgrad $1 - \alpha$.

Exempel

Samma exempel som tidigare där man vid en mätning av en process fick man följande mätdata:

6.04 4.96 4.93 3.40 7.04 4.73 3.57 7.70 4.55 3.82

En som arbetar med processen håller inte med om att standardavvikelsen kan antas vara given, utan tycker att man måste skatta den utifrån datan. Hjälp personen i fråga med att ställa upp ett 99% konfidensintervall för väntevärdet μ då mätningarna är ett stickprov på en normalfördelad variabel $X \sim N(\mu, \sigma^2)$ och σ är okänd.

Lösning. Vi betraktar siffrorna som ett stickprov från oberoende s. v. $X_j \sim N(\mu, \sigma^2)$. Vi punktskattar med $\widehat{M} = \bar{X} \sim N(\mu, \sigma^2/10)$ som tidigare och skattar σ med s, där

$$s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2 \approx 2.0842$$

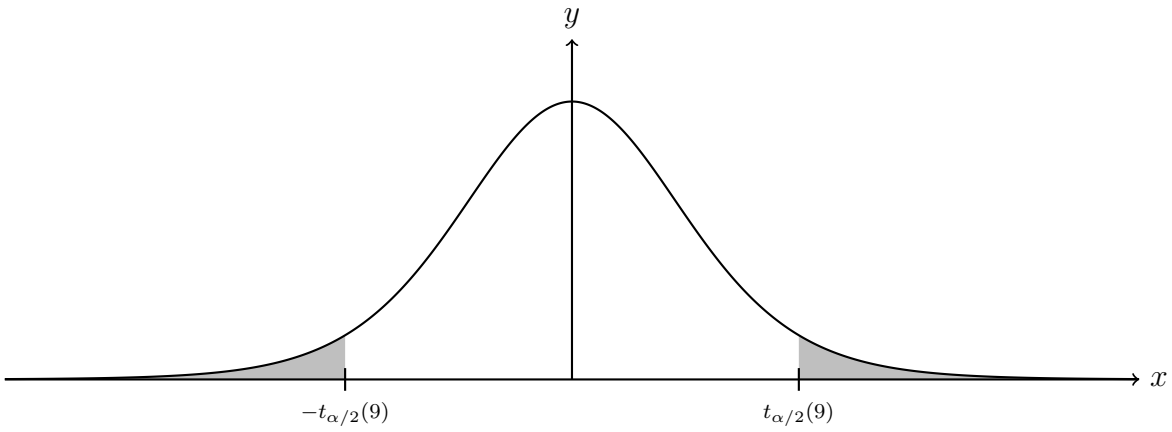
är stickprovsvariansen. Vi skapar testvariabeln

$$T = \frac{\widehat{M} - \mu}{S/\sqrt{10}} \sim t(9).$$

Som i förra deluppgiften följer det att

$$P\big(-t_{\alpha/2}(9) < T < t_{\alpha/2}(9)\big) = 1 - \alpha,$$

där $t_\beta(9)$ är kvantilerna till $t(9)$-fördelningen, $\beta \in [0, 1]$.



Ur tabell finner vi $t_{0.005}(9) = 3.25$. Genom att lösa ut μ ur olikheten i sannolikhetsmåttet får vi

$$\widehat{M} - \frac{3.25 \cdot S}{\sqrt{10}} < \mu < \widehat{M} + \frac{3.25 \cdot S}{\sqrt{10}}.$$

Om vi ersätter $\widehat{M}$ med de observerade punktskattningarna $\widehat{\mu} = 5.074$ (medelvärdet av observationerna) och $s = \sqrt{2.0842} = 1.444$ (stickprovsstandardavvikelsen) så får vi ett konfidensintervall $I_\mu = (3.59, 6.56)$ med konfidensgrad 99%.

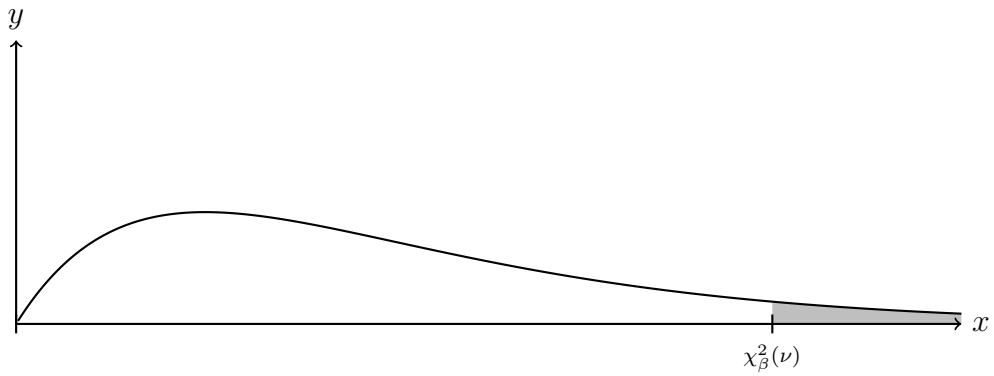
7.3 Konfidensintervall för varians

Kan man avgöra om en gissning på variansen är rimlig? Vi behöver en testvariabel där vi känner fördelningen. Enligt Cochrans sats är $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, så vi kan stänga in denna variabel. Således är

$$\begin{aligned} \chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1) \quad &\Leftrightarrow \quad \frac{1}{\chi^2_{1-\alpha/2}(n-1)} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi^2_{\alpha/2}(n-1)} \\ &\Leftrightarrow \quad \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}, \end{aligned}$$

där $\chi^2_\beta(\nu)$ är β-kvantilen för $\chi^2(\nu)$-fördelningen:

$$P(V > \chi^2_\beta(\nu)) = \beta, \quad \text{då } V \sim \chi^2(\nu).$$



k \ α	0.0500	0.0250	0.0100	0.0010	0.9500	0.9750	0.9900	0.9990
1	0.0039	0.0010	0.0002	0.0000	3.8415	5.0239	6.6349	10.8276
2	0.1026	0.0506	0.0201	0.0020	5.9915	7.3778	9.2103	13.8155
3	0.3518	0.2158	0.1148	0.0243	7.8147	9.3484	11.3449	16.2662
4	0.7107	0.4844	0.2971	0.0908	9.4877	11.1433	13.2767	18.4668
5	1.1455	0.8312	0.5543	0.2102	11.0705	12.8325	15.0863	20.5150
6	1.6354	1.2373	0.8721	0.3811	12.5916	14.4494	16.8119	22.4577
7	2.1673	1.6899	1.2390	0.5985	14.0671	16.0128	18.4753	24.3219
8	2.7326	2.1797	1.6465	0.8571	15.5073	17.5345	20.0902	26.1245
9	3.3251	2.7004	2.0879	1.1519	16.9190	19.0228	21.6660	27.8772
10	3.9403	3.2470	2.5582	1.4787	18.3070	20.4832	23.2093	29.5883
11	4.5748	3.8157	3.0535	1.8339	19.6751	21.9200	24.7250	31.2641
12	5.2260	4.4038	3.5706	2.2142	21.0261	23.3367	26.2170	32.9095
13	5.8919	5.0088	4.1069	2.6172	22.3620	24.7356	27.6882	34.5282
14	6.5706	5.6287	4.6604	3.0407	23.6848	26.1189	29.1412	36.1233
15	7.2609	6.2621	5.2293	3.4827	24.9958	27.4884	30.5779	37.6973
16	7.9616	6.9077	5.8122	3.9416	26.2962	28.8454	31.9999	39.2524
17	8.6718	7.5642	6.4078	4.4161	27.5871	30.1910	33.4087	40.7902
18	9.3905	8.2307	7.0149	4.9048	28.8693	31.5264	34.8053	42.3124
19	10.1170	8.9065	7.6327	5.4068	30.1435	32.8523	36.1909	43.8202
20	10.8508	9.5908	8.2604	5.9210	31.4104	34.1696	37.5662	45.3147
21	11.5913	10.2829	8.8972	6.4467	32.6706	35.4789	38.9322	46.7970
22	12.3380	10.9823	9.5425	6.9830	33.9244	36.7807	40.2894	48.2679
23	13.0905	11.6886	10.1957	7.5292	35.1725	38.0756	41.6384	49.7282
24	13.8484	12.4012	10.8564	8.0849	36.4150	39.3641	42.9798	51.1786
25	14.6114	13.1197	11.5240	8.6493	37.6525	40.6465	44.3141	52.6197
26	15.3792	13.8439	12.1981	9.2221	38.8851	41.9232	45.6417	54.0520
27	16.1514	14.5734	12.8785	9.8028	40.1133	43.1945	46.9629	55.4760
28	16.9279	15.3079	13.5647	10.3909	41.3371	44.4608	48.2782	56.8923
29	17.7084	16.0471	14.2565	10.9861	42.5570	45.7223	49.5879	58.3012
30	18.4927	16.7908	14.9535	11.5880	43.7730	46.9792	50.8922	59.7031
40	26.5093	24.4330	22.1643	17.9164	55.7585	59.3417	63.6907	73.4020
50	34.7643	32.3574	29.7067	24.6739	67.5048	71.4202	76.1539	86.6608
60	43.1880	40.4817	37.4849	31.7383	79.0819	83.2977	88.3794	99.6072
70	51.7393	48.7576	45.4417	39.0364	90.5312	95.0232	100.4252	112.3169
80	60.3915	57.1532	53.5401	46.5199	101.8795	106.6286	112.3288	124.8392
90	69.1260	65.6466	61.7541	54.1552	113.1453	118.1359	124.1163	137.2084
100	77.9295	74.2219	70.0649	61.9179	124.3421	129.5612	135.8067	149.4493

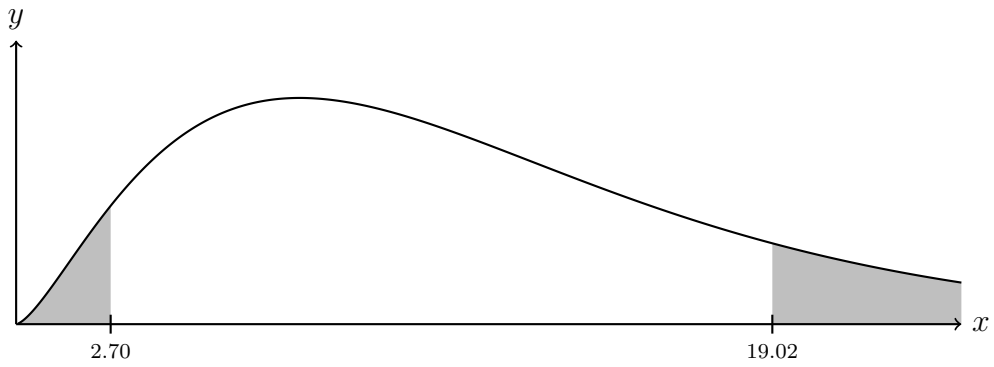
Exempel

Samma exempel (igen!) som tidigare där man vid en mätning av en process fick man följande mätdata:

6.04 4.96 4.93 3.40 7.04 4.73 3.57 7.70 4.55 3.82

Ställ upp ett 95%-igt konfidensintervall för variansen. Var antagandet att $\sigma = 2$ rimligt?

Lösning. Stickprovsstorleken är $n = 10$ och vi låter $V = \dfrac{9\,S^2}{\sigma^2}$. Under antagande att det är ett stickprov från en $N(\mu, \sigma^2)$-fördelning kommer $V \sim \chi^2(9)$. Ur tabell hittar vi gränser a och b så att $P(a < V < b) = 0.95$ genom att välja $a = \chi^2_{0.975}(9) = 2.70$ och $b = \chi^2_{0.025}(9) = 19.02$.



Vi löser nu ut σ^2 ur olikheten:

$$a < \frac{9\,S^2}{\sigma^2} < b \quad \Leftrightarrow \quad \frac{9\,S^2}{b} < \sigma^2 < \frac{9\,S^2}{a}$$

och skattar S^2 med stickprovsvariansen $s^2 = 2.08$ och erhåller då intervallet

$$I_{\sigma^2} = (0.98,\, 6.93)\,.$$

Vi kan utifrån detta även skatta ett konfidensintervall för standardavvikelsen enligt

$$I_\sigma = (0.99,\, 2.63)\,.$$

Eftersom $2 \in I_\sigma$ kan vi inte säga att $\sigma = 2$ är orimligt.

8 Enkelsidiga konfidensintervall

De konfidensintervall vi arbetat med har varit tvåsidiga i den meningen att båda gränserna har varit observationer av stokastiska variabler. Det innebär att vi lagt ut den osäkerhet vi tillåter på båda ”svansarna” i fördelningen. Men det är givetvis inte nödvändigt. Kanske är vi bara intresserade av gränsen åt ena hållet?
Typexemplet är konfidensintervall för variansen. Att variansen är liten brukar inte vara något större bekymmer, så vi lägger allt krut på att hålla koll på gränsen uppåt. Men det kan även handla om väntevärdet (eller ett predikterat värde; se nästa avsnitt). Kanske mäter vi något där vi inte får överstiga en viss nivå. Kanske en situation där det inga problem är om koncentrationen av något skadligt ämne är låg, men ett betydligt större problem om koncentrationen är hög?

Så hur åstadkommer vi detta? Vi betraktar ett par exempel.

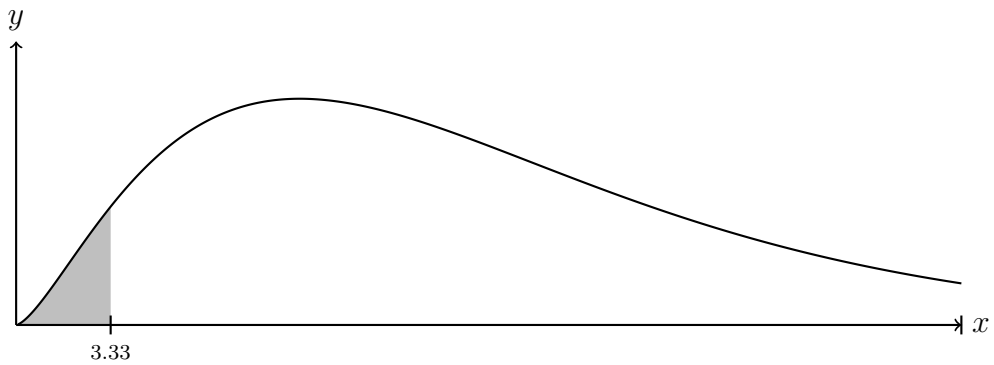
Exempel

Samma exempel (igen igen!) som tidigare där man vid en mätning av en process fick man följande mätdata:

6.04 4.96 4.93 3.40 7.04 4.73 3.57 7.70 4.55 3.82

Ställ upp ett 95%-igt konfidensintervall för variansen där vi endast är intresserade av hur stor variansen är. Skulle antagandet att $\sigma = 2.5$ vara rimligt? Jämför med föregående exempel.

Lösning. Stickprovsstorleken är $n = 10$ och vi låter även nu $V = \dfrac{9\,S^2}{\sigma^2} \sim \chi^2(9)$. Ur tabell hittar vi en gräns c så att $P(c < V) = 0.95$ genom att välja $c = \chi^2_{0.95}(9) = 3.33$.



Vi löser nu ut σ^2 ur olikheten:

$$c < \frac{9\,S^2}{\sigma^2} \quad \Leftrightarrow \quad \sigma^2 < \frac{9\,S^2}{c}$$

och skattar S^2 med stickprovsvariansen $s^2 = 2.08$ och erhåller då intervallet

$$I_{\sigma^2} = (0,\, 5.62)\,.$$

Vi kan utifrån detta även skatta ett konfidensintervall för standardavvikelsen enligt

$$I_{\sigma} = (0,\, 2.37)\,.$$

Eftersom $2 \in I_\sigma$ kan vi inte säga att $\sigma = 2$ är orimligt. Däremot kan vi säga att $\sigma = 2.5$ är orimligt (vilket vi *inte* kunde göra i föregående exempel!).

9 Prediktionsintervall

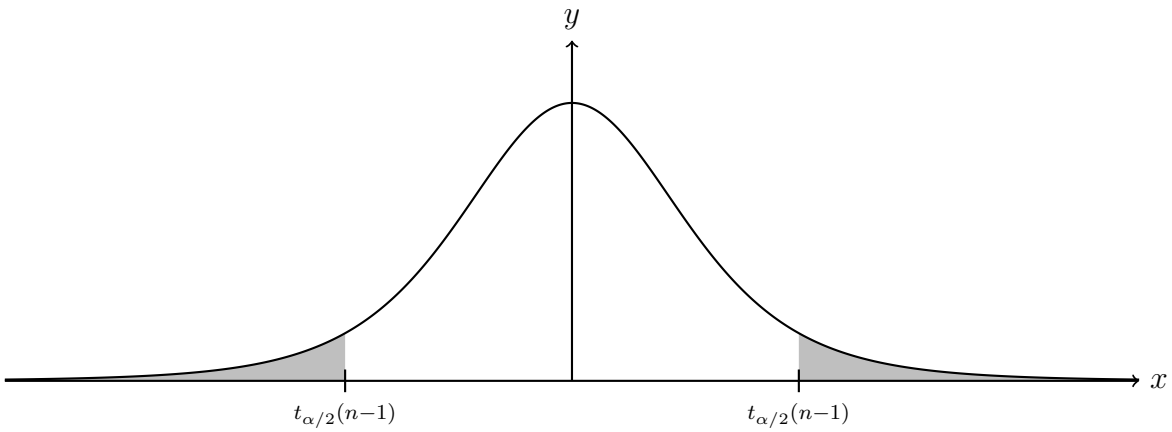
Vi har hittat konfidensintervall för både väntevärde och varians (och därigenom skattat intervall för standardavvikelsen), men kan man säga något om vart en enskild observation hamnar? Det är ju en stokastisk variabel, så det måste gå. Det vanliga är följande manöver.
Låt $X_1, X_2, \ldots, X_n$ vara ett slumpmässigt stickprov från $N(\mu, \sigma^2)$. Vi vill stänga in en enskild observation av en variabel X_0 (som antags vara oberoende) från denna fördelning. Givetvis vill vi utnyttja stickprovet, så vi betraktar variabeln $X_0 - \overline{X}$ som är normalfördelad med

$$E(X_0 - \overline{X}) = 0 \quad \text{och} \quad V(X_0 - \overline{X}) = \sigma^2 + \frac{\sigma^2}{n}.$$

Alltså kommer

$$T = \frac{X_0 - \overline{X}}{S\sqrt{1+\frac{1}{n}}} \sim t(n-1),$$

eftersom S^2 fortfarande är $\chi^2(n-1)$-fördelad. Vi kan på samma sätt som tidigare stänga in denna variabel med sannolikhet $1-\alpha$,



och sedan lösa ut X_0:

$$\begin{aligned} -t_{\alpha/2}(n-1) < \frac{X_0 - \overline{X}}{S\sqrt{1+\frac{1}{n}}} &< t_{\alpha/2}(n-1) \\ \Leftrightarrow \quad \overline{X} - t_{\alpha/2}(n-1)\, S\sqrt{1+\frac{1}{n}} < X_0 &< \overline{X} + t_{\alpha/2}(n-1)\, S\sqrt{1+\frac{1}{n}}. \end{aligned}$$

Vi ersätter nu $\overline{X}$ med det observerade medelvärdet $\overline{x}$ och S med stickprovsstandardavvikelsen s och får då intervallet

$$I_{X_0} = \left(\overline{x} - t_{\alpha/2}(n-1)\, s\sqrt{1+\frac{1}{n}},\ \overline{x} + t_{\alpha/2}(n-1)\, s\sqrt{1+\frac{1}{n}} \right).$$

Se upp med om en fråga ställs angående konfidensintervall för väntevärde eller ett prediktionsintervall. Det är helt olika frågor! Svarar du med ett konfidensintervall för väntevärdet när det efterfrågas ett prediktionsintervall blir det noll poäng. Om vi jämför intervallen för väntevärde respektive predikterat värde ser vi att vi alltid har $I_{X_o} \subset I_\mu$ med den metod vi använt ovan (och aldrig likhet).

10 Bonus: Gammafunktionen

För $z \in \mathbf{C}$ med $\operatorname{Re} z > 0$ är integralen

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$$

absolutkonvergent. Detta är ett sätt att definiera gammafunktionen på. Funktionen ovan går även att analytiskt utvidga till Re $z \leq 0$ förutom för negativa heltal. Ifrån definitionen ovan kan vi medelst partialintegration erhålla att

$$\Gamma(z+1) = \int_0^\infty x^z e^{-x}\,dx = \left[-x^z e^{-x}\right]_0^\infty + z\int_0^\infty x^{z-1}e^{-x}\,dx = z\Gamma(z).$$

Eftersom $\Gamma(1) = 1$ visar denna likhet att

$$\Gamma(n) = (n-1)!$$

för alla positiva heltal. Gamma-funktionen utvidgar således fakultetet till alla komplexa z förutom negativa heltal.
Kopplingen till normaliseringen av χ^2-fördelningen är ganska naturlig. Vi ser att om $X \sim \chi^2(k)$ så är

$$\begin{aligned}\int_0^\infty f_X(x)\,dx &= \frac{1}{2^{k/2}\Gamma(k/2)}\int_0^\infty x^{k/2-1}e^{-x/2}\,dx \\ &= \Big/ \text{variabelbyte: } \begin{matrix} u = x/2 \\ dx = 2du \end{matrix} \Big/ = \frac{1}{2^{k/2}\Gamma(k/2)}\int_0^\infty 2^{k/2}u^{k/2-1}e^{-u}\,du \\ &= \frac{1}{\Gamma(k/2)}\Gamma(k/2) = 1.\end{aligned}$$

Även identiteten $\Gamma(z+1) = z\,\Gamma(z)$ är det vi använde när vi beräknade $E(X)$ och $V(X)$ (gå tillbaka och studera partialintegrationen!).
För att identifera $\Gamma(n+1/2)$ kan vi till exempel göra variabelbytet $u = \sqrt{x}$ och partialintegrera n gånger:

$$\begin{aligned}\Gamma(n+1/2) &= \int_0^\infty x^{n+1/2}e^{-x}\,dx = 2\int_0^\infty u^{2n}e^{-u^2}\,du = 2\int_0^\infty u^{2n-1}\cdot(ue^{-u^2})\,du \\ &= -\left[u^{2n-1}e^{-u^2}\right]_0^\infty + (2n-1)\int_0^\infty u^{2n-3}\cdot(ue^{-u^2})\,du \\ &= (2n-1)\left(-\frac{1}{2}\left[u^{2n-3}e^{-u^2}\right]_0^\infty + \frac{2n-3}{2}\int_0^\infty u^{2n-5}\cdot(ue^{-u^2})\,du\right) \\ &= \frac{(2n-1)(2n-3)}{2}\left(-\frac{1}{2}\left[u^{2n-5}e^{-u^2}\right]_0^\infty + \frac{2n-5}{2}\int_0^\infty u^{2n-7}\cdot(ue^{-u^2})\,du\right) \\ &= \cdots = \frac{(2n-1)(2n-3)\cdots(2n-(2n-1))}{2^{n-1}}\int_0^\infty u^{2n-2n}e^{-u^2}\,du \\ &= \frac{(2n-1)(2n-3)\cdots(2n-(2n-1))}{2^{n-1}}\,\frac{\sqrt{\pi}}{2} = \frac{(2n)!}{4^n\,n!}\sqrt{\pi},\end{aligned}$$

där vi i sista steget använde den välkända identiteten $\displaystyle\int_0^\infty e^{-u^2}\,du = \frac{\sqrt{\pi}}{2}$.
Det finns mer eleganta sätt att ta fram identiteten för $\Gamma(n+1/2)$ genom exempelvis Eulers reflektionsformel:

$$\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin(\pi z)}, \quad z \notin \mathbf{Z},$$

men den likheten är lite mer komplicerad att bevisa, så vi nöjer oss med ovanstående.

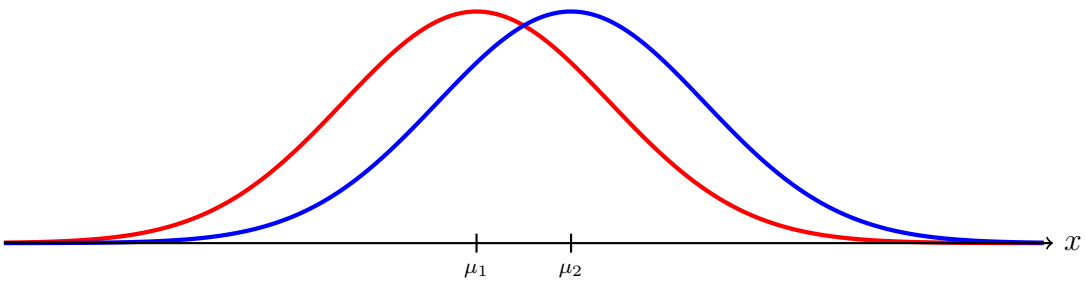
Föreläsning 4: Konfidensintervall (forts.)

Johan Thim (johan.thim@liu.se)

13 september 2018

1 Skillnad mellan parametrar

Vi kommer nu fortsätta med att konstruera konfidensintervall och vi kommer betrakta lite olika situationer där vi börjar med att titta på framförallt skillnader mellan olika mätningar. En rimlig fråga är om det föreligger någon skillnad mellan till exempel väntevärden för två stycken stickprov. Antag att vi har två slumpmässiga stickprov från två normalfördelningar. Vi vet inte direkt om fördelningarna har samma parametrar, så situationen skulle kunna se ut enligt följande.



Hur avgör vi om till exempel $\mu_1 = \mu_2$? Eller snarare om det är så att $\mu_1 \neq \mu_2$? Eller kanske om $\mu_2 > \mu_1$? Går det att avgöra om varianserna skiljer sig åt? Vad gör vi om inte stickprovet är från en normalfördelning?

2 Linjärkombinationer av normalfördelningar

Låt $X_1, \ldots, X_m$ och $Y_1, \ldots, Y_n$ vara oberoende slumpmässiga stickprov från $N(\mu_1, \sigma_1^2)$ respektive $N(\mu_2, \sigma_2^2)$. Om c_1 och c_2 är konstanter, kan vi hitta ett konfidensintervall för linjärkombinationen $c_1\mu_1 + c_2\mu_2$? Svaret beror på vilka antaganden vi gör. Vi börjar med att hitta en lämplig stokastisk storhet. Vi ser att

$$E(c_1\overline{X} + c_2\overline{Y}) = c_1\mu_1 + c_2\mu_2 \quad \text{och} \quad V(c_1\overline{X} + c_2\overline{Y}) = c_1^2\,\frac{\sigma_1^2}{m} + c_2\,\frac{\sigma_2^2}{n},$$

så eftersom vi har oberoende normalfördelade variabler gäller att

$$Z = \frac{c_1\overline{X} + c_2\overline{Y} - (c_1\mu_1 + c_2\mu_2)}{\sqrt{c_1^2\,\frac{\sigma_1^2}{m} + c_2\,\frac{\sigma_2^2}{n}}} \sim N(0,1). \tag{1}$$

Om vi känner σ_1 och σ_2 räcker detta för att ställa upp ett resultat.

2.1 Känd varians

Kända varianser

Antag att följande värden är uppmätta.

x_i	47.7	55.6	51.3	46.1	54.9			
y_i	29.2	47.8	30.9	37.7	27.9	40.1	41.5	40.9

Låt x_i vara observationer av stokastiska variabler $X_i \sim \mathrm{N}(\mu_1, 4^2)$ och y_i observationer av stokastiska variabler $Y_i \sim \mathrm{N}(\mu_2, 9^2)$, där samtliga variabler är oberoende. Ange ett 95% konfidensintervall för $\mu_1 - 2\mu_2$.

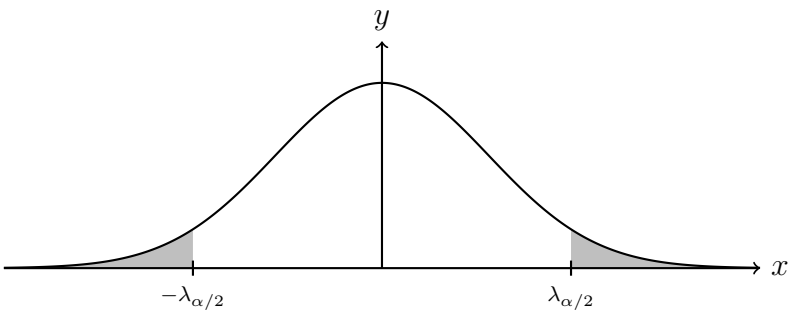
Lösning:
Låt $W = \overline{X} - 2\overline{Y}$. Varför? Denna storhet har egenskapen att $E(W) = E(\overline{X}) - 2E(\overline{Y}) = \mu_1 - 2\mu_2$, vilket är precis vad vi är intresserade av. Vidare är

$$V(W) = V(\overline{X}) + (-2)^2 V(\overline{Y}) = \frac{4^2}{5} + 4\frac{9^2}{8} = 43.7.$$

En lämplig teststorhet ges av

$$Z = \frac{W - (\mu_1 - 2\mu_2)}{\sqrt{V(W)}} \sim \mathrm{N}(0, 1).$$

Obligatorisk principfigur!



Eftersom

$$P(-\lambda_{\alpha/2} < Z < \lambda_{\alpha/2}) = 1 - \alpha$$

kan vi ur olikheten lösa ut sambandet

$$W - \lambda_{\alpha/2}\sqrt{V(W)} < \mu_1 - 2\mu_2 < W + \lambda_{\alpha/2}\sqrt{V(W)}.$$

Vi skattar W med $w = \overline{x} - 2\overline{y} = 51.12 - 2 \cdot 37 = -22.88$. Ur tabell finner vi att $\Phi(1.96) = 0.95 + 0.025 = 0.975$, så $\lambda_{\alpha/2} = 1.96$. Alltså blir intervallet

$$\begin{aligned} I_{\mu_1 - 2\mu_2} &= \left(-22.88 - 1.96 \cdot \sqrt{43.7},\ -22.88 + 1.96 \cdot \sqrt{43.7}\right) \\ &= (-35.84,\ -9.92). \end{aligned}$$

Vad säger detta oss? Jo, att med 95% säkert så ligger det verkliga värdet för $\mu_1 - 2\mu_2$ i intervallet $(-35.84,\ -9.92)$. Till exempel ser vi att noll inte finns med i intervallet, så det måste vara så att $2\mu_2 > \mu_1$ med hög säkerhet!

2.2 Okända men likadana varianser ($\sigma_1 = \sigma_2$)

Så om vi inte känner till vad varianserna är behöver vi skatta dessa. Om vi dessutom antar att $\sigma_1 = \sigma_2$ får vi ett enklare resultat, så vi börjar med det. Om vi nyttjar att $\sigma_1 = \sigma_2 = \sigma$ i ekvation (1) erhåller vi att

$$Z = \frac{c_1\overline{X} + c_2\overline{Y} - (c_1\mu_1 + c_2\mu_2)}{\sigma\sqrt{\frac{c_1^2}{m} + \frac{c_2^2}{n}}} \sim N(0,1).$$

Men vi vet fortfarande inte vad σ är, så vi ersätter σ med stickprovsstandardavvikelsen s. Eftersom vi har två stickprov viktar vi ihop dessa på sedvanligt sätt:

$$s^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}.$$

Motsvarande stickprovsvariabel S^2 uppfyller som bekant att $\dfrac{(m+n-2)S^2}{\sigma^2} \sim \chi^2(m+n-2)$ och enligt Gossets sats blir

$$T = \frac{c_1\overline{X} + c_2\overline{Y} - (c_1\mu_1 + c_2\mu_2)}{S\sqrt{\frac{c_1^2}{m} + \frac{c_2^2}{n}}} \sim t(m+n-2).$$

Okänd varians

Samma siffror som i exemplet ovan, men nu vet vi inte vad standardavvikelserna är. Antag att de är lika, dvs att $\sigma_1 = \sigma_2 = \sigma$. Finn ett 95% K.I. för $\mu_1 - \mu_2$ (inte samma uttryck som sist!). Kan du säga något om påståendet att $\mu_1 > \mu_2$?

Lösning:
Vi antar alltså här att $X_i \sim \mathrm{N}(\mu_1, \sigma^2)$ och $Y_i \sim \mathrm{N}(\mu_2, \sigma^2)$.
Vi kan skatta varianserna för varje serie med de vanliga stickprovsvarianserna, så

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \text{och} \qquad s_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \overline{y})^2$$

är kända storheter. Dessa viktas ihop enligt

$$s^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}.$$

Det följer nu att

$$T = \frac{c_1\overline{X} + c_2\overline{Y} - (c_1\mu_1 + c_2\mu_2)}{S\sqrt{\frac{c_1^2}{n} + \frac{c_2^2}{m}}} \sim t(n+m-2).$$

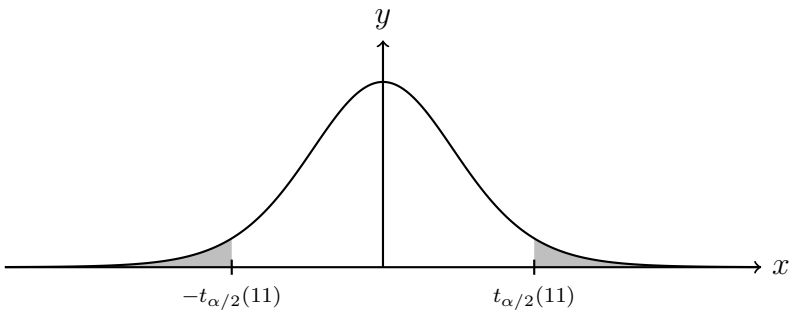
Låt $k := \sqrt{\frac{c_1^2}{n} + \frac{c_2^2}{m}}$. Snarlikt med fallet där vi kände varianserna kan vi stänga in T:

$$P(-t_{\alpha/2}(n+m-2) < T < t_{\alpha/2}(n+m-2)) = 1-\alpha$$

där vi ur olikheten kan lösa ut sambandet

$$T - t_{\alpha/2}(n+m-2)\cdot S\cdot k < c_1\mu_1 + c_2\mu_2 < T + t_{\alpha/2}(n+m-2)\cdot S\cdot k.$$

Vi har $n = 5$ och $m = 8$, så $m+n-2 = 11$ frihetsgrader. Ur tabell finner vi att $t_{0.025}(11) = 2.20$.



Vi kan räkna ut stickprovsvarianserna för x_i och y_i separat (med formel eller miniräknare). Vi erhåller $s_1^2 = 17.822$ och $s_2^2 = 49.009$ (små bokstäver, ej stokastiskt!). Den sammanvägda standardavvikelsen blir då

$$s = \sqrt{\frac{4s_1^2 + 7s_2^2}{11}} = 6.1374.$$

Vidare är $c_1 = 1$ och $c_2 = -1$, så

$$k = \sqrt{\frac{c_1^2}{n} + \frac{c_2^2}{m}} = \sqrt{\frac{1}{5} + \frac{1}{8}} = 0.5701.$$

Alltså blir

$$t_{0.025}(11)s\sqrt{\frac{c_1^2}{n} + \frac{c_2^2}{m}} = 2.20 \cdot 6.1374 \cdot 0.5701 = 7.6976.$$

Vi kan också räkna ut att $\overline{x} - \overline{y} = 14.12$, så det sökta intervallet ges av

$$\begin{aligned} I_{\mu_1-\mu_2} &= (14.12 - 7.70,\ 14.12 + 7.70) \\ &= (6.42,\ 21.82). \end{aligned}$$

Vi ser att noll ej ingår i intervallet, så det förligger troligt att $\mu_1 > \mu_2$.

2.3 Okända varianser ($\sigma_1 \neq \sigma_2$)

Ha ha. Well.. vi har inget användbart exakt samband, men det finns metoder för att hantera även denna situation. Dessa metoder ligger utanför denna kurs, men det kanske kan vara intressant att ha hört talas om dem. Problemet ligger i att uppskatta frihetsgraden ν för $t(\nu)$-fördelningen. Man kan visa (Welch-Satterthwaite-ekvationen) att

$$\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \overset{\text{appr.}}{\sim} \chi^2(\nu), \quad \text{där } \nu = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Bigg/ \left(\frac{1}{n_1 - 1}\,\frac{s_1^4}{n_1^2} + \frac{1}{n_2 - 1}\,\frac{s_2^4}{n_2^2}\right).$$

Därifrån kan vi till exempel använda att

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\text{appr.}}{\sim} t(\nu)$$

för att ställa upp ett konfidensintervall för $\mu_1 - \mu_2$.

3 Stickprov i par

Om stickproven $X_1, \ldots, X_m$ och $Y_1, \ldots, Y_n$ inte är oberoende får vi problem. Åtminstone om inte beroendet är känt. Låt oss betrakta ett vanligt förekommande exempel, nämligen stickprov i par. Av nödvändighet är då $m = n$ så stickproven har samma storlek. Vi tänker oss att x_k är observationer från $X_k \sim \mathrm{N}(\mu_k, \sigma_1^2)$ och $Y_k \sim \mathrm{N}(\mu_k + \Delta, \sigma_2^2)$. Typexemplet är när vi mäter något före och efter en förändring.
Bilda nu ett "nytt" stickprov Z_k av oberoende variabler:

$$Z_k = Y_k - X_k \sim \mathrm{N}(\Delta, \sigma^2),$$

för något σ. Vi är nu tillbaka där vi var föregående föreläsning, så de tekniker vi utvecklade där fungerar även nu.

Exempel

Preparat mot (h)järnbrist. Mätningar (i lämplig enhet) före och efter behandling hos nio patienter.

Person	1	2	3	4	5	6	7	8	9
Före	15.8	12.1	18.2	9.4	11.8	16.6	13.7	13.5	17.5
Efter	14.8	12.4	18.3	9.5	12.2	15.6	13.4	14.4	16.0

Bestäm ett 99% KI av den genomsnittliga effekten hos preparatet. Kan du styrka att det fungerar?

Lösning:
Låt x_i vara värde före behandling för person i och y_i motsvarande efter. Vi antar att olika personer är oberoende och att x_i är observationer från $X_i \sim \mathrm{N}(\mu_i, \sigma_1^2)$ och $Y_i \sim \mathrm{N}(\mu_i + \Delta, \sigma_2^2)$. Bilda $Z_i = Y_i - X_i \sim \mathrm{N}(\Delta, \sigma^2)$. Vi har nu en enda serie $z_i = y_i - x_i$ som ges enligt

z_i	-1.0	0.3	0.1	0.1	0.4	-1.0	-0.3	0.9	0.5

Vi räknar ut $s = 0.7886$ och $\overline{z} = 0.2222$. Vidare är $n - 1 = 8$ och $\alpha = 0.01$, så $t_{\alpha/2}(8) = t_{0.005}(8) = 3.36$. Alltså,

$$I_\Delta = (0.222 - 3.36 \cdot 0.7886/\sqrt{9},\ 0.222 + 3.36 \cdot 0.7886/\sqrt{9}) = (-0.66,\ 1.11).$$

Eftersom nollan finns med kan vi inte förkasta att $\Delta = 0$ (med 99% säkerhet). Preparatet kan alltså vara verkningslöst.

4 Jämförelse av varianser

"The box. You opened it. We came."
–Pinhead

Vi antog tidigare att stickproven hade samma varians (för att kunna ställa upp en lämplig teststorhet). Hur vet vi det? Kan vi på något sätt avgöra om det antagandet är rimligt? Vi vill alltså jämföra varianserna för två stickprov och för att göra det behöver vi introducera en ny fördelning (ljuva lycka!).

4.1 F-fördelningen

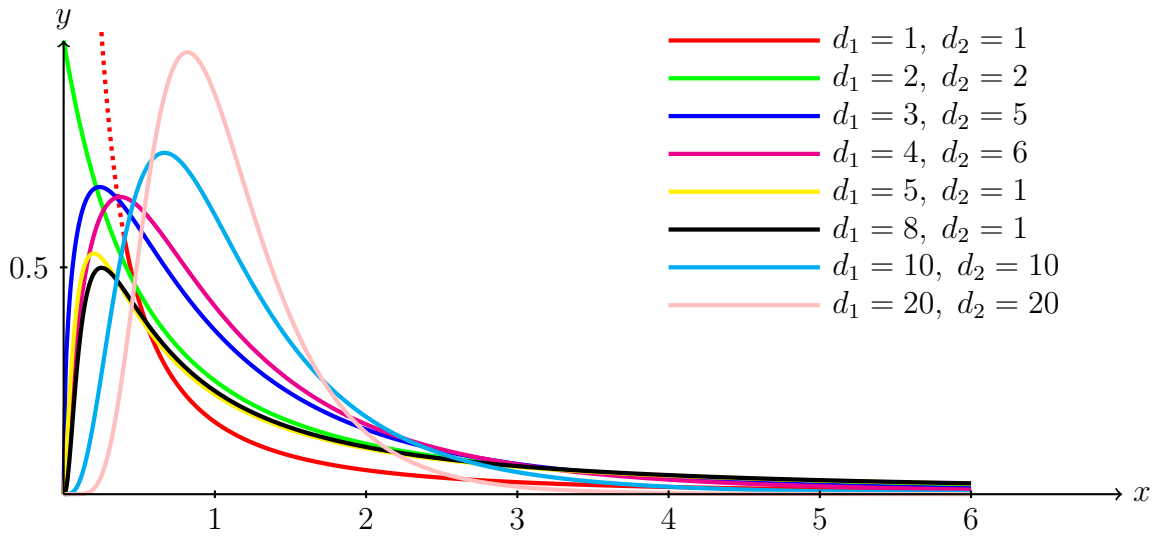
F-fördelning

Definition. Vi kallar $X \sim \mathrm{F}(d_1, d_2)$ **F-fördelad** med frihetsgraderna $d_1 > 0$ och $d_2 > 0$ om

$$f_X(x) = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}\, x\right)^{-\frac{d_1+d_2}{2}}, \quad x \geq 0,$$

där $B(a,b) = \dfrac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}$ är beta-funktionen.

Notera att $X \sim \mathrm{F}(d_1, 1) \;\Leftrightarrow\; X \sim \chi^2(d_1)$.



Sats. Om $V_1 \sim \chi^2(d_1)$ och $V_2 \sim \chi^2(d_2)$ är oberoende så gäller att

$$\frac{V_1/d_1}{V_2/d_2} \sim F(d_1, d_2).$$

Bevis. Vi börjar med att betrakta hur man kan hitta täthetsfunktionen för kvoten $Z = X/Y$ av två oberoende stokastiska variabler X och Y. Vi antar att respektive täthetsfunktion är kontinuerlig. Det gäller att $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ och

$$\begin{aligned} F_Z(z) &= P\left(\frac{X}{Y} \leq z\right) = P(X \leq Yz,\, Y > 0) + P(X \geq Yz,\, Y < 0) \\ &= \int_0^\infty \int_{-\infty}^{yz} f_{X,Y}(x,y)dx\,dy + \int_{-\infty}^0 \int_{yz}^\infty f_{X,Y}(x,y)dx\,dy \\ &= \int_0^\infty f_Y(y) F_X(yz)\,dy + \int_{-\infty}^0 f_Y(y)(1 - F_X(yz))\,dy, \end{aligned}$$

från vilket det följer att

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \int_0^\infty y f_Y(y) f_X(yz)\,dy + \int_{-\infty}^0 -y f_Y(y) f_X(yz)\,dy$$
$$= \int_{-\infty}^\infty |y| f_Y(y) f_X(yz)\,dy.$$

Vi noterar även att för $r > 0$ gäller att

$$P\left(\frac{X}{r} \le x\right) = P(X \le rx) \quad \Rightarrow \quad f_{X/r}(x) = r f_X(rx).$$

Således ges täthetsfunktionerna för V_1/d_1 och V_2/d_2 av

$$f_{V_1/d_1}(x) = \frac{d_1^{d_1/2}}{2^{d_1/2}\Gamma\left(\frac{d_1}{2}\right)} x^{d_1/2-1} e^{-d_1 x/2}, \quad x \ge 0$$

och

$$f_{V_2/d_2}(y) = \frac{d_2^{d_2/2}}{2^{d_2/2}\Gamma\left(\frac{d_2}{2}\right)} y^{d_2/2-1} e^{-d_2 y/2}, \quad y \ge 0,$$

så enligt resultatet ovan för kvoten $\dfrac{V_1/d_1}{V_2/d_2}$ erhåller vi att

$$\begin{aligned}
f_Z(z) &= \int_0^\infty y f_{V_2/d_2}(y) f_{V_1/d_1}(yz)\,dy \\
&= \frac{d_2^{d_2/2} d_1^{d_1/2}\, z^{d_1/2-1}}{2^{(d_1+d_2)/2}\,\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)} \int_0^\infty y^{d_1/2+d_2/2-1} e^{-(y(d_2+d_1 z))/2}\,dy \\
&= \Bigg/ \text{variabelbyte: } \begin{matrix} u = y(d_2 + d_1 z) \\ dy = (d_2 + d_1 z)^{-1} du \end{matrix} \Bigg/ \\
&= \frac{d_2^{d_2/2} d_1^{d_1/2}\, z^{d_1/2-1} (d_2 + d_1 z)^{-(d_1+d_2)/2}}{2^{(d_1+d_2)/2}\,\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)} \int_0^\infty u^{(d_1+d_2)/2-1} e^{-u/2}\,du \\
&= \frac{d_2^{d_2/2-1} d_1^{d_1/2-1}\Gamma\left(\frac{d_1+d_2}{2}\right)\, z^{d_1/2-1} (d_2 + d_1 z)^{-(d_1+d_2)/2}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)}
\end{aligned}$$

eftersom

$$\frac{1}{2^{(d_1+d_2)/2}\,\Gamma\left(\frac{d_1+d_2}{2}\right)} \int_0^\infty u^{(d_1+d_2)/2-1} e^{-u/2}\,du = 1$$

då detta är integralen av täthetsfunktionen för en stokastisk variabel $U \sim \chi^2(d_1 + d_2)$. Vi kan hyffsa till slutresultatet för $f_Z(z)$ genom att bryta ut d_2 ur $(d_2 + d_1 z)^{-(d_1+d_2)/2}$ och använda beta-funktionen:

$$\begin{aligned}
f_Z(z) &= \frac{d_2^{d_2/2}\, z^{d_1/2-1} d_1^{-d_2/2} \left(1 + \frac{d_1}{d_2} z\right)^{-(d_1+d_2)/2}}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \\
&= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_2}{d_1}\right)^{-d_2/2} z^{d_1/2-1} \left(1 + \frac{d_1}{d_2} z\right)^{-(d_1+d_2)/2},
\end{aligned}$$

vilket är precis vad vi ville visa. □

Sats. Om $X \sim \mathrm{F}(d_1, d_2)$ så är

$$E(X) = \frac{d_2}{d_2 - 2},\ d_2 > 2, \quad \text{och} \quad V(X) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)},\ d_2 > 4.$$

Bevis. Välj två oberoende stokastiska variabler $V_1 \sim \chi^2(d_1)$ och $V_2 \sim \chi^2(d_2)$. Eftersom vi visade ovan att $\dfrac{V_1/d_1}{V_2/d_2} \sim F(d_1, d_2)$ följer det att

$$E(X) = E\left(\frac{V_1/d_1}{V_2/d_2}\right) = \Big/ \frac{V_1}{d_1} \text{ och } \frac{V_2}{d_2} \text{ oberoende } \Big/ = E\left(\frac{V_1}{d_1}\right) E\left(\frac{d_2}{V_2}\right) = \frac{d_1\, d_2}{d_1} E\left(\frac{1}{V_2}\right),$$

där vi nyttjat att $E(V_1) = d_1$. Vi beräknar $E(1/V_2)$:

$$\begin{aligned} E\left(\frac{1}{V_2}\right) &= c\int_0^\infty x^{d_2/2-2} e^{-x/2} dx = c\left(\left[\frac{1}{d_2/2-1}\, x^{d_2/2-1} e^{-x/2}\right]_0^\infty + \frac{1}{d_2-2}\int_0^\infty x^{d_2/2-1} e^{-x/2}\, dx\right) \\ &= \frac{1}{d_2-2}\int_0^\infty f_{V_2}(x)\, dx = \frac{1}{d_2-2}, \end{aligned}$$

under förutsättning att $d_2 > 2$. Således blir

$$E(X) = \frac{d_2}{d_2 - 2} \text{ om } d_2 > 2.$$

När det gäller variansen använder vi ett analogt resonemang:

$$\begin{aligned} V(X) &= V\left(\frac{V_1/d_1}{V_2/d_2}\right) = \frac{d_2^2}{d_1^2}\left(E\left(\frac{V_1^2}{V_2^2}\right) - E\left(\frac{V_1}{V_2}\right)^2\right) \\ &= \Big/ V_1 \text{ och } V_2 \text{ oberoende } \Big/ = \frac{d_2^2}{d_1^2}\left(E\left(V_1^2\right) E\left(\frac{1}{V_2^2}\right) - E\left(V_1\right)^2 E\left(\frac{1}{V_2}\right)^2\right) \\ &= \frac{d_2^2}{d_1^2}\left(\left(V\left(V_1\right) + E(V_1)^2\right) E\left(\frac{1}{V_2^2}\right) - \frac{d_1^2}{(d_2-2)^2}\right) \\ &= \frac{d_2^2}{d_1^2}\left(\left(2d_1 + d_1^2\right) E\left(\frac{1}{V_2^2}\right) - \frac{d_1^2}{(d_2-2)^2}\right) \end{aligned}$$

eftersom $V(V_1) = 2d_1$ och vi använt resultatet för $E(1/V_2)$ ovan. Vi partialintegrerar nu för att beräkna $E(1/V_2^2)$:

$$\begin{aligned} E\left(\frac{1}{V_2^2}\right) &= c\int_0^\infty x^{k/2-3} e^{-x/2}\, dx \\ &= c\left(\left[\frac{1}{d_2/2-2}\, x^{d_2/2-2} e^{-x/2}\right]_0^\infty + \frac{1}{d_2-4}\int_0^\infty x^{d_2/2-2} e^{-x/2}\, dx\right) \\ &= \frac{1}{d_2-4} E\left(\frac{1}{V_2}\right) = \frac{1}{(d_2-4)(d_2-2)}, \end{aligned}$$

om $d_2 > 4$ och vi nyttjat kalkylen för $E(1/V_2)$ ovan.
Alltså blir

$$V(X) = \frac{d_2^2}{d_1^2}\left(\frac{2d_1 + d_1^2}{(d_2-4)(d_2-2)} - \frac{d_1^2}{(d_2-2)^2}\right) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2-2)^2(d_2-4)},$$

vilket var precis vad vi ville visa. □

Sats. Om $X \sim F(d_1, d_2)$ så är $1/X \sim F(d_2, d_1)$.

Bevis. Låt $V = 1/X$ och antag att $v > 0$. Då gäller att

$$F_V(v) = P(1/X \le v) = P(X \ge 1/v) = 1 - F_X(1/v) \quad \Rightarrow \quad f_V(v) = \frac{1}{v^2} f_X(1/v),$$

så

$$\begin{aligned} f_V(v) &= \frac{1}{v^2} \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_2}{d_1}\right)^{-d_2/2} \left(\frac{1}{v}\right)^{d_1/2-1} \left(1 + \frac{d_1}{d_2}\frac{1}{v}\right)^{-(d_1+d_2)/2} \\ &= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_2}{d_1}\right)^{-d_2/2} \left(\frac{1}{v}\right)^{d_1/2+1} \left(\frac{d_1}{d_2 v}\right)^{-(d_1+d_2)/2} \left(\frac{d_2}{d_1}v + 1\right)^{-(d_1+d_2)/2} \\ &= \frac{1}{B\left(\frac{d_2}{2}, \frac{d_1}{2}\right)} \left(\frac{d_1}{d_2}\right)^{-d_1/2} v^{d_2/2-1} \left(1 + \frac{d_2}{d_1}v\right)^{-(d_2+d_1)/2} \end{aligned}$$

eftersom $B(a, b) = B(b, a)$. Således är $V \sim F(d_2, d_1)$. □

Sats. Om $T \sim t(n)$ så är $T^2 \sim F(1, n)$.

Bevis. Låt $V = T^2$ och antag att $v \ge 0$. Då gäller att

$$F_V(v) = P(T \le v) = P(-\sqrt{v} \le T \le \sqrt{v}) = F_T(\sqrt{v}) - F_T(-\sqrt{v}),$$

så

$$\begin{aligned} f_V(v) = F_V'(v) &= \frac{1}{2\sqrt{v}} f_T(\sqrt{v}) - \frac{-1}{2\sqrt{v}} f_T(-\sqrt{v}) \\ &= \frac{1}{\sqrt{v}} f_T(\sqrt{v}) = \frac{1}{\sqrt{v}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} v^{-1/2} \left(1 + \frac{v}{n}\right)^{-(n+1)/2} \\ &= \frac{\Gamma\left(\frac{n}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{n}\right)^{1/2} v^{-1/2} \left(1 + \frac{v}{n}\right)^{-(n+1)/2} \\ &= \frac{1}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \left(\frac{1}{n}\right)^{1/2} v^{-1/2} \left(1 + \frac{1}{n}v\right)^{-(n+1)/2}, \end{aligned}$$

eftersom $f_T(-t) = f_T(t)$ och $\Gamma(1/2) = \sqrt{\pi}$. Således är $V \sim F(1, n)$. □

4.2 Jämförelse av två varianser

Låt $X_1, \ldots, X_{n_1}$ och $Y_1, \ldots, Y_{n_2}$ vara oberoende slumpmässiga stickprov från $N(\mu_1, \sigma_1^2)$ respektive $N(\mu_2, \sigma_2^2)$. Då vet vi att

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \quad \text{och} \quad \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1).$$

Det följer då enligt ovan att

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

Exempel

Betrakta det tidigare exempel igen, där vi hade

x_i	47.7	55.6	51.3	46.1	54.9			
y_i	29.2	47.8	30.9	37.7	27.9	40.1	41.5	40.9

Antag att x_i är oberoende observationer av $N(\mu_1, \sigma_1^2)$ och att y_i är oberoende observationer av $N(\mu_2, \sigma_2^2)$. Ange ett 95% konfidensintervall för σ_1/σ_2.

Lösning. Låt $F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$. På grund av antagandet följer det att $F \sim F(4,7)$. Vi söker ett konfidensintervall med konfidensgrad 95% så vi behöver gränser a och b så att

$$P(F < a) = 0.025 \quad \text{och} \quad P(F > b) = 0.025.$$

Ur tabell finner vi att $a = 0.1102$ och $b = 5.5226$ (i MATLAB `finv([0.025 0.975], 4, 7)`). Notera att tabeller oftast endast innehåller värden för sannolikheter ≥ 0.5. Anledning till det är att vi kan använda att

$$F \sim F(m,n) \quad \Rightarrow \quad \frac{1}{F} \sim F(n,m).$$

Konkret för oss just nu blir det således

$$0.025 = P(F < a) = P\left(\frac{1}{a} < \frac{1}{F}\right) = 1 - P\left(\frac{1}{F} \leq \frac{1}{a}\right) \quad \Leftrightarrow \quad P\left(\frac{1}{F} \leq \frac{1}{a}\right) = 0.975.$$

Vi försöker nu lösa ut σ_1/σ_2:

$$a < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} < b \quad \Leftrightarrow \quad \frac{1}{b}\frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{a}\frac{S_1^2}{S_2^2}.$$

Vi skattar nu S_1^2 och S_2^2 med respektive stickprovsvarians:

$$s_1^2 = 17.822 \quad \text{och} \quad s_2^2 = 49.0086.$$

Ett konfidensintervall för σ_1^2/σ_2^2 ges alltså av

$$I = \left(\frac{1}{5.5226}\frac{17.822}{49.0086}, \ \frac{1}{0.1102}\frac{17.822}{49.0086}\right) = (0.0658, \ 3.2999).$$

Vill vi ha ett konfidensintervall för σ_1/σ_2 tar vi helt enkelt roten ur gränserna:

$$I_{\sigma_1/\sigma_2} = (\sqrt{0.0658}, \ \sqrt{3.2999}) = (0.2566, \ 1.8165).$$

I MATLAB kan man använda funktionen `vartest2` för att skapa konfidensintervallet.

```
>> x = [47.7  55.6  51.3  46.1  54.9 ];
>> y = [29.2  47.8  30.9  37.7  27.9  40.1  41.5  40.9];
>> [H P CI] = vartest2(x,y,0.05,'both')
H =
     0
P =
    0.3452
CI =
    0.0658    3.2998
```

Vad H och P representerar kommer vi till på nästa föreläsning.

5 Konfidensintervall via CGS

Så vad gör vi om stickprovet inte är från en normalfördelning?

6 Stickprov för andel

Exempel

Ett företag som sysslar med opinionsanalys väljer slumpmässigt ut 400 vuxna i Sverige och frågar om de har åsikt A. Av dessa svarar 80 ja (alla svarar). Bestäm ett approximativt 95% konfidensintervall för andelen av den stora populationen som håller åsikt A.

Lösning. Vi låter X vara antalet som svarar ja. Då är egentligen $X \sim \text{Hyp}(N, 400, p)$, där N är antalet vuxna i Sverige (rimligen ca 8 miljoner). Då $400 \ll 8000000$ är det helt rimligt att anta att $X \overset{\text{appr.}}{\sim} \text{Bin}(400, p)$. Vi vill skatta den okända andelen p och väljer som skattningsvariabel

$$\widehat{P} = \frac{X}{400}$$

Vi har observerat att $\widehat{p} = 80/400 = 0.2$.
Binomialfördelningen är lite jobbig eftersom den är diskret, så vi försöker oss på en approximation. Eftersom

$$400 \cdot \widehat{p} \cdot (1 - \widehat{p}) = 400 \cdot 0.2 \cdot 0.8 = 64$$

är ordentligt större än 10 är det rimligt att approximera binomialfördelningen med normalfördelning. Alltså,

$$\widehat{P} \overset{\text{appr.}}{\sim} \text{N}(p, p(1-p)/400).$$

Låt oss bilda

$$Z = \frac{\widehat{P} - p}{\sqrt{\widehat{p}(1-\widehat{p})/400}} \overset{\text{appr.}}{\sim} N(0, 1).$$

Observera att vi ersatt med det skattade värdet på p i kvadratroten (men **inte** i täljaren). Vi nyttjar här alltså medelfelet d, dvs

$$d(\widehat{P}) = \sqrt{\widehat{p}(1-\widehat{p})/400} = 0.02.$$

Vi kan nu räkna precis som om vi känner standardavvikelsen exakt, så om vi söker ett approximativt 95% K.I. erhåller vi

$$I_p = (0.2 - 1.96 \cdot 0.02,\ 0.2 + 1.96 \cdot 0.02) = (0.16, 0.24).$$

7 Jämförelse av två andelar

Antag att vi har två maskiner. Vid uppmätning fann man att Maskin 1 producerade 20 defekta enheter av 400, och att Maskin 2 producerade 60 defekta enheter av 600.
Modell: Låt X vara antal defekta enheter från Maskin 1 och Y antal defekta enheter från Maskin 2. Under lämpligt oberoendeantagande vet vi att $X \sim \text{Bin}(400, p_1)$ och $Y \sim \text{Bin}(600, p_2)$ där p_1 och p_2 är de verkliga felsannolikheterna. Vi skattar lämpligen med

$$\widehat{P_1} = \frac{X}{400} \qquad \text{och} \qquad \widehat{P_2} = \frac{Y}{600}.$$

Vi har observerat att $\widehat{p_1} = 20/400 = 0.05$ och $\widehat{p_2} = 60/600 = 0.10$. Alltså är $\widehat{p_1} - \widehat{p_2} = -0.05$. Är detta signifikant? För att svara på frågan behöver vi räkna lite sannolikheter. Eftersom både $n_1\widehat{p_1}(1-\widehat{p_1})$ och $n_2\widehat{p_2}(1-\widehat{p_2})$ är mycket större än 10 är det rimligt att approximera binomialfördelningen med normalfördelning. Alltså,

$$\widehat{P_1} \overset{\text{appr.}}{\sim} \mathrm{N}\left(p_1, \frac{p_1(1-p_1)}{400}\right) \qquad \text{och} \qquad \widehat{P_2} \overset{\text{appr.}}{\sim} \mathrm{N}\left(p_2, \frac{p_2(1-p_2)}{600}\right).$$

Då följer det att

$$\widehat{P_1} - \widehat{P_2} \overset{\text{appr.}}{\sim} \mathrm{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{400} + \frac{p_2(1-p_2)}{600}\right).$$

Vi bildar nu

$$Z = \frac{\widehat{P_1} - \widehat{P_2} - (p_1 - p_2)}{\sqrt{\widehat{p_1}(1-\widehat{p_1})/400 + \widehat{p_2}(1-\widehat{p_2})/600}} \overset{\text{appr.}}{\sim} N(0,1).$$

Observera att vi ersatt med skattade värden på p_1 och p_2 i kvadratroten (men **inte** i täljaren). Det blir fortfarande approximativt (men lite sämre så klart) normalfördelat, men underlättar mycket för beräkningar. Vi har

$$\sqrt{\widehat{p_1}(1-\widehat{p_1})/400 + \widehat{p_2}(1-\widehat{p_2})/600} = 0.0164.$$

Vi kan nu räkna precis som om vi känner standardavvikelsen exakt, så om vi söker ett approximativt 95% K.I. erhåller vi

$$I_{p_1-p_2} = (-0.05 - 1.96 \cdot 0.0164,\ -0.05 + 1.96 \cdot 0.0164) = (-0.08, -0.02).$$

Endast negativa värden, så $p_1 < p_2$ med hög sannolikhet! Maskin 2 är antagligen sämre.

Föreläsning 5: Hypotesprövningar

Johan Thim (johan.thim@liu.se)

24 november 2018

Vi har nu studerat metoder för hur man hittar lämpliga skattningar av okända parametrar och även stängt in dessa skattningar i konfidensintervall för att ha kontroll på vad som är rimligt eller ej. Den sista frågan kan man närma sig på lite annorlunda (men egentligen mer naturligt sätt) genom så kallade hypotestester (ibland kallade signifikanstester).
Ett hypotesttest i detta sammanhang består av en **nollhypotes** H_0 och en **mothypotes** H_1. Typiskt är att nollhypotesen är något vi vill motbevisa (och därmed styrka att mothypotesen antagligen gäller). I denna kurs kommer vi oftast begränsa oss till så kallade enkla nollhypoteser och oftast av typen

$$H_0 : \theta = \theta_0.$$

Mothypotesen kan väljas på olika sätt beroende på vad vi vill visa. De vanligaste är av typerna

$$H_1 : \theta \neq \theta_0 \quad \text{eller} \quad H_1 : \theta > \theta_0 \quad \text{eller} \quad H_1 : \theta < \theta_0.$$

För att testa hypotesen behöver vi en teststorhet t som avgör hur ett stickprov ska behandlas. Denna storhet har analog funktion med de som användes när vi ställde upp konfidensintervall. Vi låter $x_1, x_2, \ldots, x_n$ vara ett stickprov från en fördelning F som beror på en okänd parameter θ. Motsvarande slumpmässiga stickprov betecknas $X_1, X_2, \ldots, X_n$ i vanlig ordning.

Teststorhet/Testvariabel

Definition. En funktion $t : \mathbf{R}^n \to \mathbf{R}$ given av $t(x_1, x_2, \ldots, x_n)$ kallas **teststorhet** eller **testvariabel** och är en observation av den stokastiska variabeln $t(X_1, X_2, \ldots, X_n)$.

För att avgöra om vi ska förkasta H_0 väljer vi en **signifikansnivå** α och bestämmer sedan ett kritiskt område C som är en delmängd av det område funktionen t varierar över (en del av värdemängden). Detta område beror på fördelningen F och den signifikansnivå vi vill utföra hypotestestet på.

Kristiskt område, signifikansnivå

Definition. Det **kritiska området** C är ett område så att H_0 förkastas om

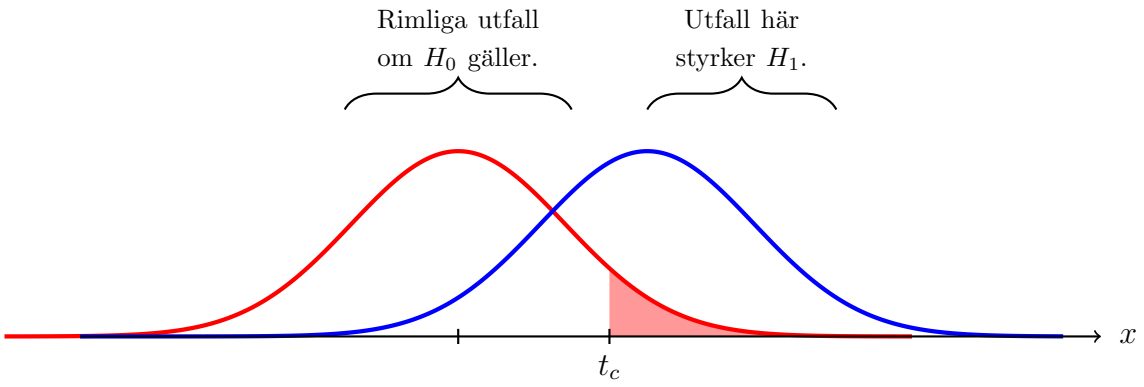
$$t(x_1, \ldots, x_n) \in C.$$

Om H_0 förkastas säger vi att H_1 är styrkt och drar slutsatsen att H_1 gäller. Sannolikheten

$$\alpha = P(t(X_1, \ldots, X_n) \in C \,|\, H_0 \text{ är sann})$$

kallas för testets **signifikansnivå**.

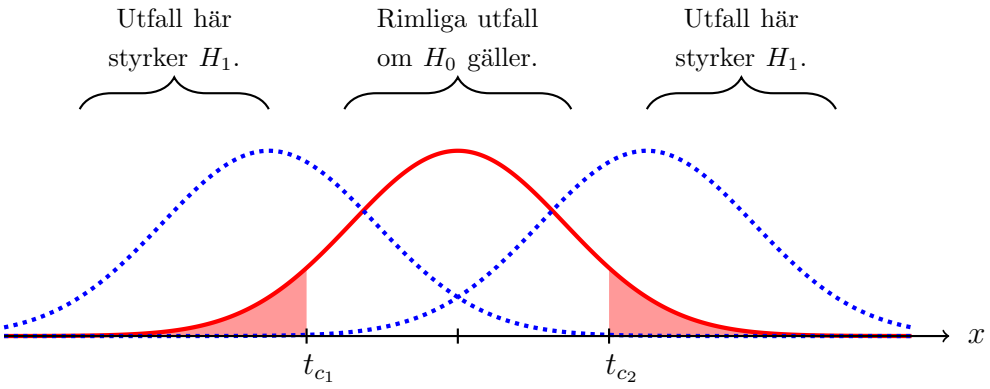
Det kritiska området består alltså av värden som är för extrema för att vara troliga under förutsättningen att nollhypotesen gäller.
Låt oss ställa upp ett hypotestest för väntevärdet för fördelningen F enligt $H_0 : \mu = \mu_0$ mot $H_1 : \mu > \mu_0$. Vi vill således styrka att det verkliga väntevärdet är större än μ_0.



Den röda kurvan är täthetsfunktionen för $t(X_1, \ldots, X_n)$ om H_0 skulle vara sann medan den blå är den verkliga täthetsfunktionen. Vi ser att observerade värden är betydligt rimligare i det kritiska området om den blå fördelningen gäller. Det kritiska området blir således

$$C = \{x \in \mathbf{R} : x > t_c\}.$$

Om $t > t_c$ så förkastar vi H_0.
Om vi istället skulle testa $H_0 : \mu = \mu_0$ mot $H_1 : \mu \neq \mu_0$, vad blir skillnaden? Vi vill således i detta läge styrka att det verkliga väntevärdet är något annat än μ_0 (inte nödvändigtvis att det verkliga väntevärdet är större).



De blå kurvorna är potentiella verkliga fördelningar för $t(X_1, \ldots, X_n)$ medan den röda fortfarande är fördelningen om H_0 skulle vara sann. Det kritiska området blir således

$$C = \{x \in \mathbf{R} : x > t_{c_2} \text{ eller } x < t_{c_1}\}.$$

Om $t > t_{c_2}$ eller om $t < t_{c_1}$ så förkastar vi H_0. När vi vet mer om fördelningen för $t(X_1, \ldots, X_n)$ kan vi under antagandet att H_0 stämmer hitta gränserna explicit.

Att ställa upp H_0 och H_1 ska göras *innan* stickprov observerats. Utgår man från mätdatan för att hitta på sina hypoteser beter man sig bedrägligt.

Styrka

Definition. Vi definierar **styrkefunktionen** $h(\theta)$ enligt

$$h(\theta) = P(H_0 \text{ förkastas} \,|\, \theta \text{ är det riktiga värdet}).$$

Sannolikheten $h(\theta)$ kallas för testets styrka i θ.

För ett bra hypotestest bör $h(\theta)$ vara stor för $\theta \in H_1$ och $h(\theta)$ liten för $\theta \in H_0$. Notera även att $h(\theta_0) = \alpha$.

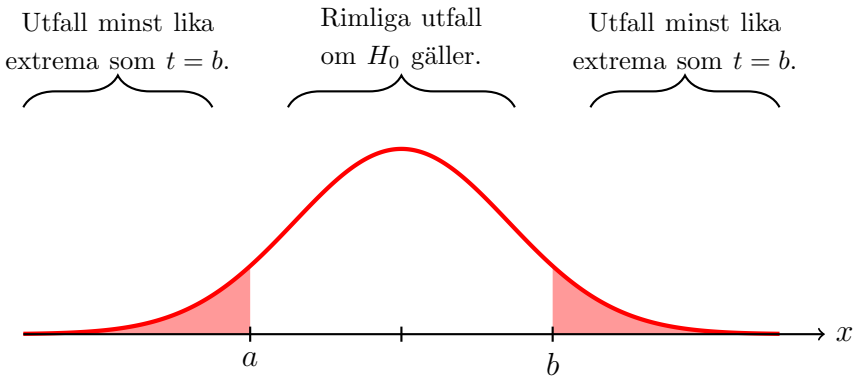
Fel av typ I och II

Definition. Att förkasta H_0 då H_0 är sann kallas **fel av typ I** och har sannolikheten α. Risken för ett fel av typ I är således signifikansnivån.
Att *inte* förkasta H_0 då H_0 är falsk kallas för **fel av typ II**.

p-värde

Definition. För ett givet stickprov kan man för ett signifikanstest beräkna ett p-värde. Denna sannolikhet är den lägsta signifikansnivån på vilken vi skulle förkasta H_0. Med andra ord är p sannolikheten att vi får ett minst lika extremt utfall som det givna stickprovet med antagandet att H_0 är sann.

Låt oss testa $H_0 : \theta = \theta_0$ mot $H_1 : \theta \neq \theta_0$. Om vi utifrån stickprovet beräknar teststorheten $t(x_1, \ldots, x_n) = b$ så behöver vi alltså karakterisera alla utfall som är minst lika extrema om H_0 gäller. Nu blir vi beroende av hur fördelningen ser ut. Låt oss anta något symmetriskt.



Så p-värdet kan om fördelningen ser symmetrisk ut enligt ovan beräknas enligt

$$p = P(t(X_1, \ldots, X_n) \leq a) + P(t(X_1, \ldots, X_n) \geq b) = 2P(t(X_1, \ldots, X_n) \geq b),$$

där a måste väljas så vi har samma sannolikhetsmassa i båda ”svansarna.” Om fördelningen har en riktig skum uppsyn då? Ja, då blir det svårt. En variation vi kan hantera är om mothypotesen är av typen $H_1 : \theta > \theta_0$ (till exempel) då vi endast har

$$p = P(t(X_1, \ldots, X_n) \geq b)$$

eftersom utfall i vänstra svansen nu inte längre räknas som extrema. Utseendet på mothypotesen är alltså fundamentalt.

Märk väl att p-värdet inte säger någonting om huruvida H_0 är sann eller ej givet observationen av t. Det vi har är sannolikheten för ett lika extremt utfall *givet* att H_0 gäller. *Inte* tvärtom!

Alla principfigurer ovan har varit små söta symmetriska och kontinuerliga historier. Hur blir det vid andra typer av fördelningar?

2 Hypotestest för Binomialfördelning

Vi undersöker situationen med ett belysande exempel.

Exempel

Ett mynt kastas (oberoende) 30 gånger och vid 10 av dessa blir det en krona. Kan vi förkasta hypotesen att myntet är ärligt med signifikansnivå 5%? Vad är styrkan om sannolikheten för krona är 3/10?

Lösning. Vi vill testa om myntet är ärligt, så vi börjar med att ställa upp en modell. Låt X vara antalet krona vid 30 kast. Då är $X \sim \text{Bin}(n, p)$ där $n = 30$ och $p =$ sannolikheten för krona är okänd. så en rimlig nollhypotes ges av

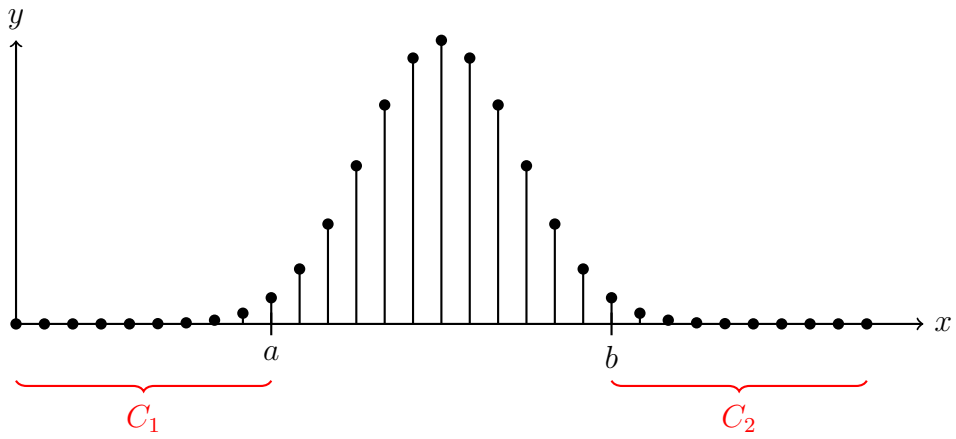
$$H_0 : p = \frac{1}{2}$$

och innan experimentet vet vi inte om mothypotesen bör vara $p < 1/2$ eller $p > 1/2$, så vi tar det säkra före det osäkra och väljer att testa mot

$$H_1 : p \neq \frac{1}{2}.$$

Givet att H_0 är sann så förväntar vi oss frekvensen $30 \cdot 0.5 = 15$ utfall som är krona. Är 10 signifikant mindre? Vi ställer upp det kritiska området:

$$C = \{x \in \mathbf{Z} : 0 \leq x \leq a \text{ eller } b \leq x \leq n\}$$



Hur hittar vi a och b? Vi får helt enkelt testa oss fram (och använda tabeller). Eftersom

$$p(x) = \binom{30}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{30-x}$$

kan vi beräkna att

$$\sum_{x=0}^{9} p(x) = 0.0214 \quad \text{och} \quad \sum_{x=0}^{10} p(x) = 0.0494$$

samt (känt redan pga symmetri då $p = 0.5$ men för fullständighetens skull):

$$\sum_{x=21}^{30} p(x) = 0.0214 \quad \text{och} \quad \sum_{x=20}^{30} p(x) = 0.0494.$$

Vi väljer $a = 9$ och $b = 21$. Då gäller att

$$\begin{aligned} P(X \in C \,|\, H_0) &= P(X \in C_1 \,|\, H_0) + P(X \in C_2 \,|\, H_0) \\ &= P(X \leq a) + P(X \geq b) = 0.0214 + 0.0214 = 0.0428 < 0.05. \end{aligned}$$

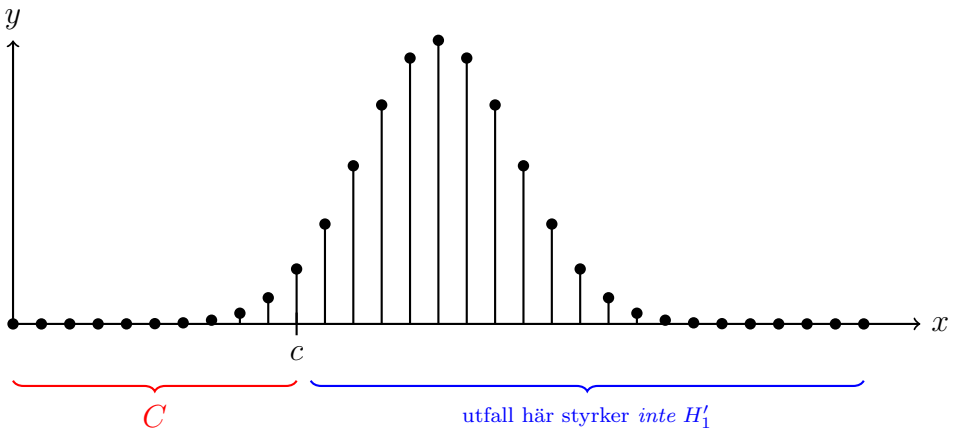
Detta är det största kritiska område vi kan få för att hålla signifikansnivån. Observera att vi alltså inte kan träffa $\alpha = 0.05$ exakt. Detta är typiskt vid diskreta fördelningar.
Eftersom $x = 10 \notin C$ kan vi inte dra någon slutsats, utan myntet kan mycket väl vara ärligt. Vi kan således *inte* förkasta H_0 (vilket inte på något sätt betyder att H_0 är sann).
Styrkan vid $p = 0.3$ blir

$$\begin{aligned} h(0.3) &= P(H_0 \text{ förkastas} \,|\, p = 0.3) = P(X \in C \,|\, p = 0.3) \\ &= \sum_{x=0}^{9} \binom{30}{x} 0.3^x 0.7^{30-x} + \sum_{x=21}^{30} \binom{30}{x} 0.3^x 0.7^{30-x} = 0.5888 + 7.28 \cdot 10^{-6} = 0.5888. \end{aligned}$$

Antag att vi istället vill testa mothypotesen H_1' att myntet ger färre krona än klave. Vi har då

$$H_1' : p < \frac{1}{2}.$$

Hur ser det kritiska området C ut?



Eftersom

$$\sum_{x=0}^{10} p(x) = 0.0494 \quad \text{och} \quad \sum_{x=0}^{11} p(x) = 0.1002$$

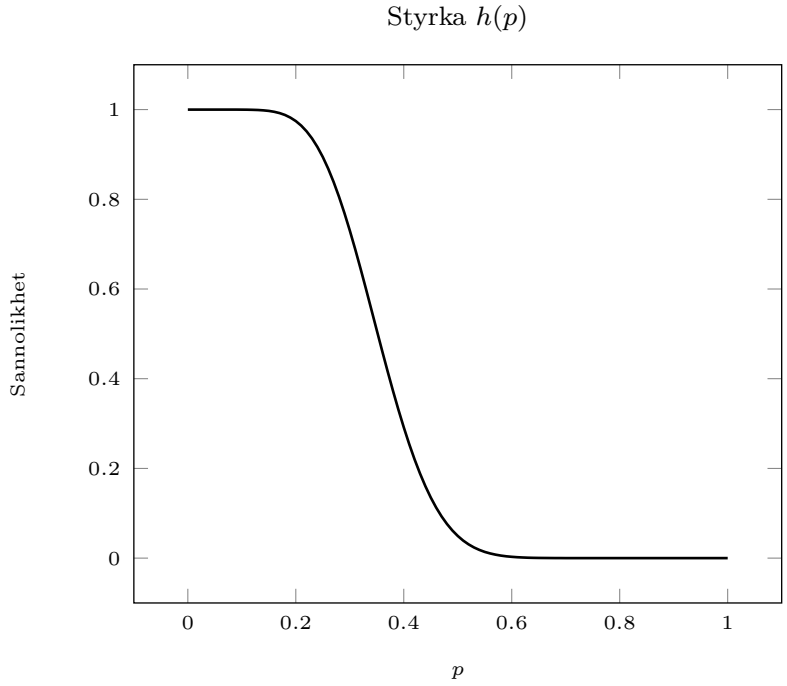
så ser vi att $c = 10$ är nödvändigt. Därmed blir

$$C = \{x \in \mathbf{Z} : 0 \leq x \leq 10\}$$

och vår observation $x = 10 \in C$. Alltså kan vi förkasta H_0 och anse att H_1' är styrkt. Styrkan vid $p = 0.3$ blir

$$\begin{aligned} h(0.3) &= P(H_0 \text{ förkastas} \,|\, p = 0.3) = P(X \in C \,|\, p = 0.3) \\ &= \sum_{x=0}^{10} \binom{30}{x} 0.3^x 0.7^{30-x} = 0.7304. \end{aligned}$$

Notera alltså att styrkan *beror* på mothypotesen! Ganska naturligt när man tänker efter, men det är lätt att tro att styrkan för ett test bara har med nollhypotesen att göra. Det är alltså helt fel. Vi kan även låta MATLAB räkna ut styrkefunktionen för alla $p \in [0, 1]$ för att se hur det ser ut.



3 Hypotestest för Poissonfördelning

Övriga diskreta fördelningar kan givetvis hanteras analogt med binomialexemplet i föregående avsnitt och med datorkraft är det inte större problem att räkna exakt i väldigt många fall. Men som vi kommer ihåg från tidigare kurser går det även att approximera flera diskreta fördelningar med normalfördelning om vissa förutsättningar är uppfyllda. Låt oss studera ett exempel med Poissonfördelning på två sätt.

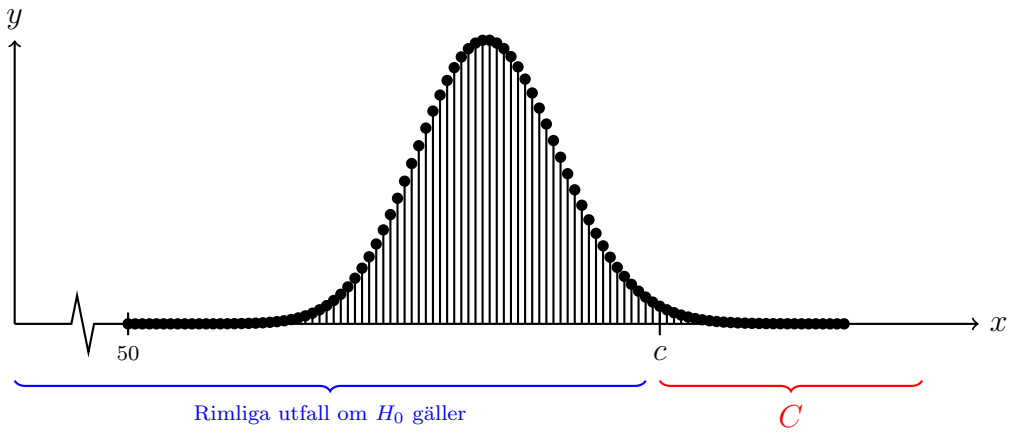
Exempel

Antalet datapaket till en server kan betraktas som en Poissonprocess $X(t)$ med en okänd intensitet λ. För att kunna hantera överbelastning har man ett varningssystem som varnar om antalet paket överstiger en gräns N på två tidsenheter. Varningen sker alltså om intensiteten är större än väntat. Antag att $\lambda = 50$ (enhet: tusen paket). Det är dyrt att avbryta servicen så man vill högst tillåta felaktig varning med 1% risk.
Hitta gränsen N och avgör om man bör varna om $x = 120$ vid en mätning. Vad skulle p-värdet bli om $x = 130$?

Lösning. Det förväntade antalet paket är $\mu = E(X(t)) = \lambda t$, så om $\lambda = 50$ förväntar vi oss $\mu = 50 \cdot 2 = 100$ (tusen) paket. Låt

$$H_0 : \mu = 100 \quad \text{och} \quad H_1 : \mu > 100.$$

Vi söker det kritiska området C. En figur kan vara bra.



Låt $p(k)$, $k = 0, 1, 2, \ldots$, vara sannolikhetsfunktionen för en Po(100)-fördelad variabel. Ur tabell (eller med hjälp av matlab och funktionerna poisspdf eller poisscdf) kan vi finna att

$$\sum_{k=124}^{\infty} p(k) = 1 - \sum_{k=0}^{123} p(k) = 0.0112 \quad \text{och} \quad \sum_{k=125}^{\infty} p(k) = 0.0088.$$

Således blir det kritiska området

$$C = \{k \in \mathbf{Z} : k \geq 125\}.$$

Eftersom observationen $x = 120 \notin C$ så kan vi inte förkasta H_0. Vi bör inte varna.

Vi beräknar p-värdet vid observationen $x = 130$ genom

$$p = P(X \geq 130 \,|\, H_0) = \sum_{k=130}^{\infty} p(k) = \Big/ \text{tabell} \Big/ = 0.0023.$$

Vi summerar alltså sannolikheterna för alla utfall som är minst lika extrema som $x = 130$.

Om vi stirrar lite på plotten ovan så ser den tämligen normalfördelad ut, eller hur? Det är ingen slump. Om $X \sim \text{Po}(\mu)$ med $\mu \geq 15$ så är $X \overset{\text{appr.}}{\sim} N(\mu, \mu)$ (variansen är μ). Vi kan använda detta för att hitta en approximativ gräns N. Låt $X \sim \text{Po}(100)$. Då gäller att

$$0.01 = P(X \geq N) = 1 - P(X < N) = 1 - P\left(\frac{X - 100}{\sqrt{100}} < \frac{N - 100}{\sqrt{100}}\right) = 1 - \Phi\left(\frac{N - 100}{10}\right).$$

Således är

$$\begin{aligned} 0.01 = 1 - \Phi\left(\frac{N - 100}{10}\right) \quad &\Leftrightarrow \quad 0.99 = \Phi\left(\frac{N - 100}{10}\right) \\ &\Leftrightarrow \quad 2.3263 = \frac{N - 100}{10} \\ &\Leftrightarrow \quad N = 23.263 + 100 = 123.263. \end{aligned}$$

Eftersom N måste vara ett heltal väljer vi $N = 124$. Även med halvstegskorrigering hamnar vi inte på det exakta värdet, men det är tillräckligt nära för de flesta ändamål. Vi kan även återskapa kalkylen för p-värdet vid $x = 130$ enligt

$$p \approx 1 - \Phi\left(\frac{130 - 100}{10}\right) = 0.0013.$$

4 Normalapproximation – Generellt

När vi approximerar med normalfördelningen är tillvägagångssättet nästan alltid det samma. Vi har en punktskattning $\widehat{\theta}$ där $\widehat{\Theta} \overset{\text{appr.}}{\sim} N(\theta, D^2)$ och vi vill testa nollhypotesen $H_0 : \theta = \theta_0$. Som teststorhet använder vi då oftast

$$Z = \frac{\widehat{\Theta} - \theta_0}{D} \quad \text{eller} \quad Z = \frac{\widehat{\Theta} - \theta_0}{d}.$$

Den senare teststorheten då vi inte känner D exakt utan skattar med d. Vi förutsätter att d är en vettig skattning av D då H_0 är sann. Notera att i båda fallen kommer $Z \overset{\text{appr.}}{\sim} N(0, 1)$ om H_0 är sann. Vi använder alltså ingen t-fördelning här (det finns inget som säger att det skulle bli bättre i det generella fallet).
Hur det kritiska området ser ut beror på hur vi ställer upp mothypotesen. Om $H_1 : \theta \neq \theta_0$ får C utseendet $]-\infty, -a[\cup]a, \infty[$. Är mothypotesen enkelsidig blir det bara ett av intervallen (med annan parameter a). Talet a hittar vi i normalfördelningstabell.

5 Test för skillnad i andel

En mycket vanlig situation är att vi vill undersöka om det föreligger någon skillnad i andel mellan två grupper. Antag att vi har x_1 som observation av $X_1 \sim \text{Bin}(n_1, p_1)$ och x_2 som observation av $X_2 \sim \text{Bin}(n_2, p_2)$ (vi antar oberoende).

Vi är intresserade av att testa hypotesen $H_0 : p_1 = p_2$ mot till exempel $H_1 : p_1 \neq p_2$. Om H_0 är sann så är en lämplig skattning av $p = p_1 = p_2$

$$\widehat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

Faktum är att detta är ML-skattningen (om H_0 är sann) och därmed har den bra egenskaper såsom konsistens. Vad gäller fördelningen för $\widehat{P}$ blir den värre (vad händer om man summerar binomialfördelningar?). Men, om n_1 och n_2 är ganska stora och p inte är allt för nära ändpunkterna i $[0, 1]$, så kanske vi kan normalapproximera? Vi har redan gjort detta (se konfidensintervall för $p_1 - p_2$), men för fullständighetens skull låt oss repetera. Om H_0 är sann gäller att

$$E(\widehat{P}) = \frac{n_1 p + n_2 p}{n_1 + n_2} = p$$

och

$$V(\widehat{P}) = \frac{n_1 p(1-p) + n_2 p(1-p)}{(n_1 + n_2)^2} \to 0,$$

då $n_1 + n_2 \to \infty$, så skattningen av p är väntevärdesriktig och konsistent. För att testa H_0 använder vi $\widehat{P_1} - \widehat{P_2}$, och om H_0 är sann så gäller att

$$\widehat{P_1} - \widehat{P_2} \overset{\text{appr.}}{\sim} N\left(0, \widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

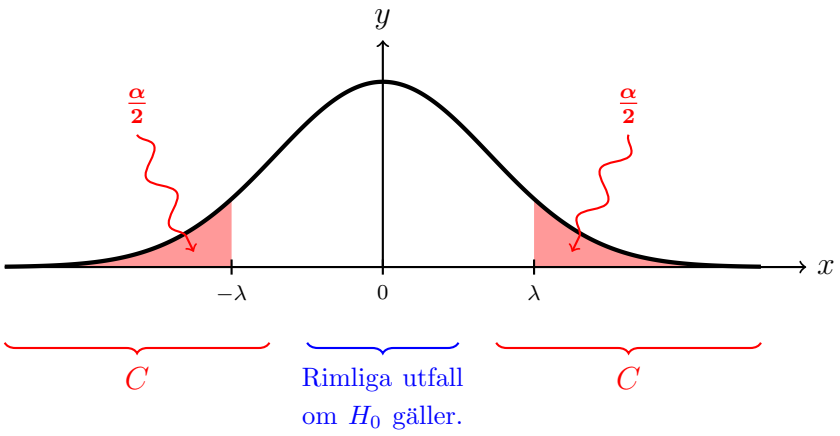
Eftersom vi inte känner p exakt använder vi skattningen $\widehat{p}$ ovan i uttrycket för variansen (eller vi ersätter standardavvikelsen med medelfelet). Vi kan även gå över i standardiserad form så vi känner igen oss:

$$Z = \frac{\widehat{P_1} - \widehat{P_2}}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\text{appr.}}{\sim} N(0, 1).$$

Om $H_1 : p_1 \neq p_2$ så ges det kritiska området av

$$C = \{z \in \mathbf{R} : |z| > \lambda\}$$

för något lämpligt $\lambda = \Phi^{-1}(1 - \alpha/2)$ vi finner ur tabell (eller MATLAB).



Exempel

Två opinionsinstitut Analysera Mera AB och StickProvarna AB undersöker om befolkningen tycker att sommaren varit för varm. AM frågar 500 personer och andelen $p_1 = 0.7$ (350 st) håller med. SP frågar 400 personer och $p_2 = 0.8$ (320 stycken) håller med. Undersök om det finns någon signifikant skillnad mellan resultaten på signifikansnivån 5% (approximativt).

Lösning. Låt $H_0 : p_1 = p_2 = p$ och $H_1 : p_1 \neq p_2$. Om H_0 är sann väljer vi skattningen

$$\widehat{p} = (350 + 320)/(500 + 400) = 0.744.$$

Med beteckningarna ovan gäller då (om H_0 är sann) att

$$Z = \frac{\widehat{P_1} - \widehat{P_2}}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{500} + \frac{1}{400}\right)}} = \frac{\widehat{P_1} - \widehat{P_2}}{0.0293} \overset{\text{appr.}}{\sim} N(0,1).$$

Det är rimligt att approximera både $\widehat{P_1}$ och $\widehat{P_2}$ med normalfördelning eftersom både $500 \cdot 0.7 \cdot 0.3 \geq 10$ och $400 \cdot 0.8 \cdot 0.2 \geq 10$. Vi hittar det kritiska området

$$C = \{z \in \mathbf{R} : |z| > \lambda\}$$

där $\lambda = \Phi^{-1}(0.975) = 1.96$. Således ska – om H_0 är sann –

$$\left|\frac{\widehat{p_1} - \widehat{p_2}}{0.0293}\right| > 1.96 \quad \Leftrightarrow \quad |\widehat{p_1} - \widehat{p_2}| > 1.96 \cdot 0.0293 = 0.0573$$

för att vi ska förkasta H_0. Med $\widehat{p_1} = 0.7$ och $\widehat{p_2} = 0.8$ ser vi att $0.1 > 0.0573$, så vi förkastar H_0. Det är troligen en skillnad i resultaten.

Ett alternativ är att ställa upp konfidensintervallet $I_{p_1-p_2}$ för $p_1 - p_2$ och sedan testa hypotesen genom att undersöka om $0 \in I_{p_1-p_2}$. Skulle det vara så att 0:an ingår kan vi inte förkasta H_0. Ligger intervallet helt på ena sidan 0 däremot så förkastar vi H_0. Detta test är helt ekvivalent eftersom vi nyttjar samma testvariabel.

6 Poissonapproximation

Som bekant kan man även approximera binomialfördelning med Poissonfördelning om $n \geq 10$ och $p \leq 0.1$. Detta kan vara nödvändigt då p ligger nära 0 eller 1 så normalapproximation inte fungerar bra. Vi betraktar ett exempel.

Exempel

En leverantör av laboratorieutrustning hävdar att deras pipetter bara behöver kalibreras en gång per år och att risken för att en pipett faller utanför toleransnivån innan dess är 0.5% (vid normal användning). Laboratorieansvarig Laura (för ett stort laboratorie) tycker inte att det stämmer och har ett år efter inköpet och kontinuerligt användande av 1000 stycken behövt kalibrera om 11 st. Testa hypotesen att felrisken är 0.5% mot att den är högre på signifikansnivån 1% (approximativt).

Lösning. Den stokastiska variabeln X är antalet av de 1000 pipetterna som behövs kalibreras i förtid. Om vi antar att händelserna är oberoende (är det rimligt?) så är $X \sim \text{Bin}(1000, p)$ där p är felrisken. Låt $H_0 : p = 0.005$ och $H_1 : p > 0.005$. Vi kan använda $\widehat{P} = \dfrac{X}{1000}$, men enklare är att direkt nyttja X. Om H_0 är sann så gäller att

$$X \overset{\text{appr.}}{\sim} \text{Po}(1000 \cdot 0.005) = \text{Po}(5).$$

Det kritiska området väljs som

$$C = \{z \in \mathbf{Z} : z > k\}$$

för något $k \in \mathbf{Z}$. Vi vill att

$$P(X \in C \,|\, H_0) \leq 0.01$$

och i tabell (eller med `k` = poissinv(0.99, 5) i MATLAB, vilket ger det minsta heltalet k så att $P(X \leq k) \geq 0.99$) finner vi att $k = 11$. Alltså gäller

$$P(X > 11 \,|\, H_0) < 0.01 \quad \text{(exakt värde: 0.0055)},$$

och Lauras observation $x = 11$ är alltså *inte* signifikant. Vi kan inte förkasta H_0 och säga att leverantören har fel.

Exempel

Laura är inte nöjd och kräver att examensarbetaren Audrey ska göra om hypotestestet och använda normalapproximation *som folk*. Motivera varför det inte är bra men utför testet. Undersök också hur hypotestestet blir om man inte approximerar för att hjälpa den stackars examensarbetaren att motivera.

Lösning. Vid normalapproximation kräver vi att $np(1-p) \geq 10$ och om vi väljer att skatta p med $\widehat{p} = 10/1000 = 0.01$ hamnar vi precis kring den gränsen så osäkerheten är stor. Använder vi leverantörens $p = 0.005$ blir det betydligt under. Alltså inget att rekomendera. Men om vi envisas så skulle

$$X \overset{\text{appr.}}{\sim} N(1000p, 1000p(1-p)) \quad \text{som dålig approximation.}$$

Om vi antar att H_0 är sann skulle då

$$Z = \frac{X - 1000 \cdot 0.005}{\sqrt{1000 \cdot 0.005 \cdot (1 - 0.005)}} \overset{\text{appr.}}{\sim} N(0, 1),$$

återigen som en tveksam approximation. Kritiskt område ges av

$$0.01 = P(Z > \lambda) = 1 - \Phi(\lambda) \quad \Leftrightarrow \quad \lambda = \Phi^{-1}(0.99) = 2.3263$$

så

$$\frac{X - 5}{\sqrt{4.975}} > 2.3263 \quad \Leftrightarrow \quad X > 10.1888$$

och vi skulle därför låta C ges av $X \geq 11$, varvid resultatet $x = 11$ skulle verka signifikant.

Vi kan ställa upp ett exakt test genom att låta $H_1 : p > 0.005$ och välja

$$C = \{x \in \mathbf{Z} : x > k\}$$

för något $k \in \mathbf{Z}$. Precis som med Poissonapproximationen hittar vi k genom att i MATLAB använda `k` = binoinv(0.99, 1000, 0.005) vilket resulterar i $k = 11$. Alltså samma gräns som vi fick med Poissonapproximationen. Exakt värde här blir $P(X > k) = 0.0053$.

Föreläsning 6: Hypotestester (forts.)

Johan Thim (johan.thim@liu.se)

24 november 2018

Vi fortsätter nu exkursionen i hypotesernas förlovade land. Fokus kommer vara på den vanligaste typen av hypotestester, nämligen när datan antas vara normalfördelad. Vi kommer nu åter stifta bekantskap med t- och χ^2-fördelningar.
Låt oss börja med ett enklare exempel.

Exempel

I en fabrik med mångårig erfarenhet tillverkar man material av en viss tjocklek. Man mäter med jämna mellanrum tjockleken på 9 nytillverkade material och testar om medelvärdet uppfyller $|\overline{x} - 5.0| > 0.05$. Om så är fallet stoppas tillverkningen och tekniker får gå igenom maskineriet. Ansvarig för metodutvecklingen vet av erfarenhet att $\sigma = 0.1$. Om vi antar normalfördelning, vad är bästa signifikansnivån för testet om $H_0 : \mu = 5$ testas mot $H_1 : \mu \neq 5$?

Lösning. Vi antar att $X_i \sim N(\mu, \sigma^2) = N(\mu, 0.1^2)$, $i = 1, 2, \ldots, 9$, är oberoende. För att testa H_0 ställer vi upp teststorheten

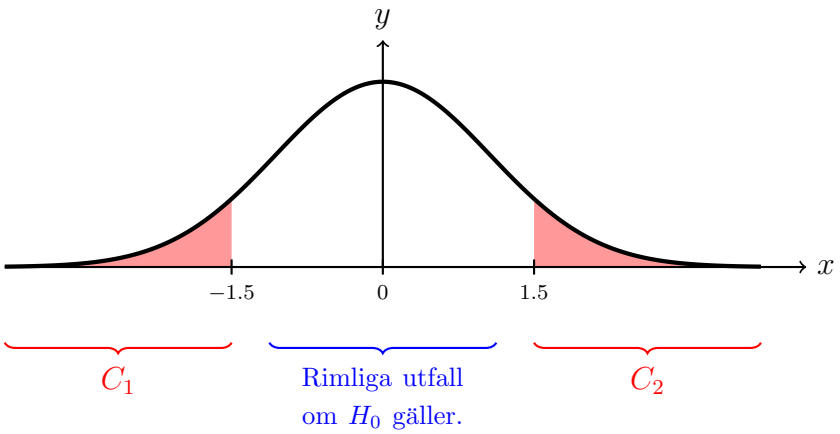
$$Z = \frac{\overline{X} - 5.0}{0.1/\sqrt{9}}.$$

Om H_0 är sann så är $Z \sim N(0, 1)$. Om vi jämför med fabrikens test så ser vi att

$$\left|\frac{0.1}{3}\,Z\right| > 0.05 \quad \Leftrightarrow \quad |Z| > 1.5$$

ger det kritiska området

$$C = \{z \in \mathbf{R} : |z| > 1.5\}.$$



Så signifikansnivån α kan om fördelningen ser symmetrisk ut enligt ovan beräknas enligt

$$\begin{aligned} p &= P(Z \leq -1.5) + P(Z \geq 1.5) = 2P(Z \leq -1.5) \\ &= 2\Phi(-1.5) = 2(1 - \Phi(1.5)) = 0.1336. \end{aligned}$$

Den bästa signifikansnivån vi kan välja är alltså $\alpha = 0.1336$.

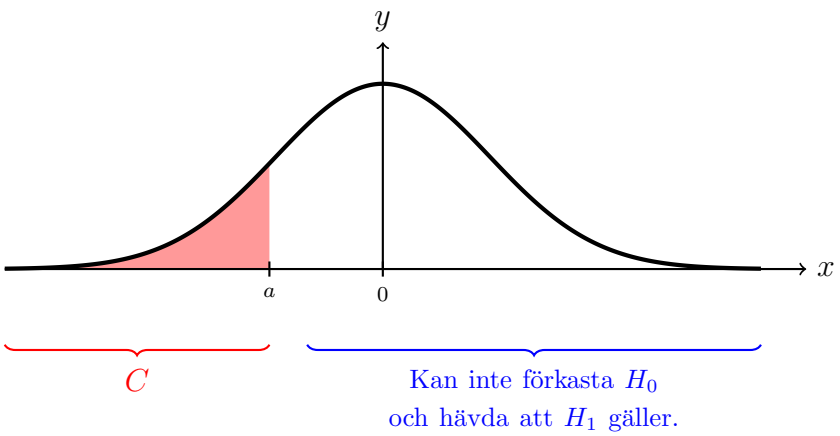
Exempel

Fabriken har fått en ny beställare som inte har något problem om materialet blir tjockare. Ansvarig tänker lite snabbt och ställer upp ett test med samma signifikansnivå för att endast testa att materialet inte blir för tunt. Hur ser testet ut nu och varför är detta antagligen inte vad man vill göra?

Lösning. Vi har fortfarande $H_0 : \mu = 5$ men mothypotesen ges nu av $H_1 : \mu < 5$. Vi kan använda samma teststorhet och om H_0 är sann så är

$$Z = \frac{\overline{X} - 5.0}{0.1/\sqrt{9}} \sim N(0, 1).$$

Vi söker en gräns a så att $\overline{X} < a$ med sannolikheten $\alpha = 0.1336$ om H_0 är sann.



Det är tydligt att

$$\overline{X} < a \quad \Leftrightarrow \quad Z = \frac{\overline{X} - 5.0}{0.1/3} < \frac{a - 5.0}{0.1/3},$$

så

$$\begin{aligned} 0.1336 = P\left(Z < \frac{a - 5.0}{0.1/3}\right) \quad &\Leftrightarrow \quad \frac{a - 5.0}{0.1/3} = \Phi^{-1}(0.1336) = -1.1095 \\ &\Leftrightarrow \quad a - 5.0 = -0.0370. \end{aligned}$$

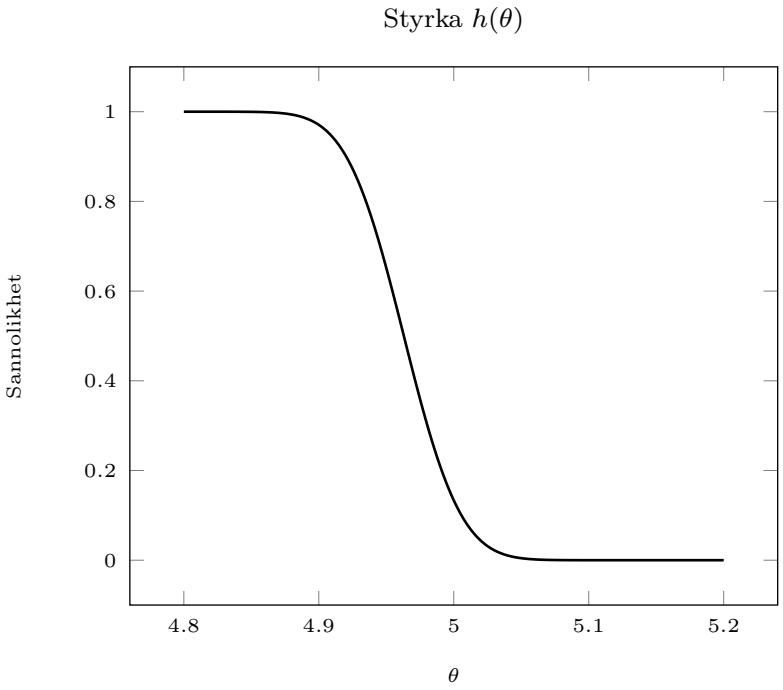
Det sökta värdet blir alltså $a = 4.9630$. Detta test blir alltså mer känsligt för att materialet är för tunnt än det föregående. Om den nya beställaren har samma tolerans för fel som de tidigare är det kanske mer strategiskt att istället sänka signifikansnivån till hälften.

Exempel

Föregående hypotestest har en lite udda signifikansnivå. Hur ser styrkefunktionen ut?

Lösning. Styrkefunktionen definieras enligt

$$h(\theta) = P(H_0 \text{ förkastas} \mid \ \mu = \theta) = P\left(\overline{X} < 4.9630 \,\middle|\, \overline{X} \sim N\left(\theta, \frac{0.1^2}{9}\right)\right)$$
$$= \Phi\left(\frac{4.9630 - \theta}{0.1/3}\right) = \Phi(148.89 - 30\theta).$$



Exempel

En nyanställd i fabriken (med en kurs i statistisk inferens i bagaget) påtalar att det kanske är olämpligt att anta att variansen är känd och att man borde skatta den från mätningen. Vid en mätning fick man stickprovsvariansen 0.0144, vad ger testet $|\overline{x} - 5.0| > 0.05$ för signifikansnivå i denna situation?

Lösning. Vi testar således $H_0 : \mu = 5.0$ mot $H_1 : \mu \neq 5.0$ och som testvariabel blir

$$T = \frac{\overline{X} - 5.0}{S/\sqrt{9}} \sim t(8)$$

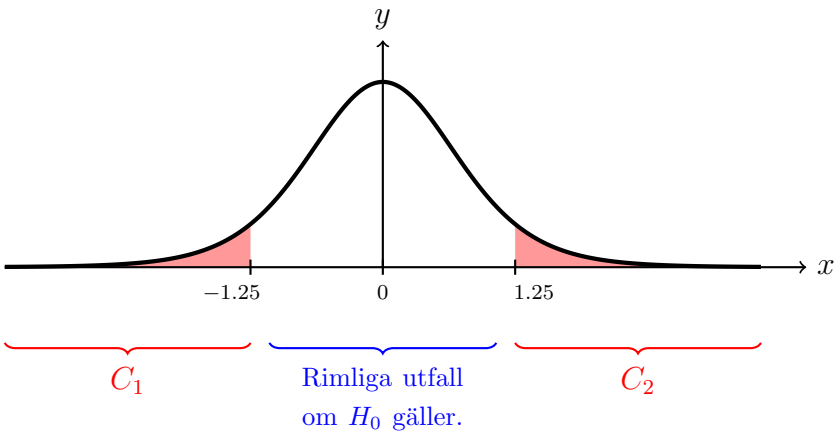
om H_0 är sann. Analogt med första exemplet måste då

$$\left|\frac{s}{3}\, t\right| > 0.05 \quad \Leftrightarrow \quad |t| > \frac{0.15}{s},$$

vilket ger det kritiska området

$$C = \left\{t \in \mathbf{R} : |t| > \frac{0.15}{s}\right\} = \{t \in \mathbf{R} : |t| > 1.25\}$$

i vårt fall. Fördelningen för T är symmetrisk lik normalfördelningen, så situationen är snarlik.



Så p-värdet kan om fördelningen ser symmetrisk ut enligt ovan beräknas enligt

$$\begin{aligned} p &= P(T \leq -1.25) + P(T \geq 1.25) = 2P(T \leq -1.25) \\ &= 2F_T(-1.25) = 2 \cdot 0.1233 = 0.2466. \end{aligned}$$

Här använde vi `tcdf(-1.25,8)` i MATLAB (vi har inga tabeller i formelsamlingen för att slå på t-fördelningar i "den riktningen").
Den bästa signifikansnivån vi kan välja är alltså i princip $\alpha = 0.25$. Mindre lyckat! Testet kanske behöver ändras.

2 Väntevärde för ett stickprov

I föregående samling exempel såg vi vad som hände när vi kände till variansen exakt och vad som hände när vi behövde uppskatta den från mätdatan. Osäkerheten ökar vid varje skattning, men känner vi inte exakta värden är skattningarna nödvändiga.
Låt oss undersöka den generella situationen. Vi har ett stickprov $X_1, X_2, \ldots, X_n$ från en normalfördelning $N(\mu, \sigma^2)$. Vi kan därför tänka oss att

$$X_i = \mu + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),\ i = 1, 2, \ldots, n,$$

där ϵ_i är obereoende. Notera att samtliga variabler har samma varians. Som föregående avsnitt visade är det skillnad på när vi känner variansen exakt och när den behöver skattas.
Vi börjar med att testa

$$H_0 : \mu = \mu_0 \qquad \text{mot} \qquad H_1 : \mu \neq \mu_0.$$

Givetvis kan man vilja testa mot $H_1' : \mu > \mu_0$ eller $H_1'' : \mu < \mu_0$ också, vi kommer ta upp något sådant exempel också.

2.1 Känd varians

Eftersom

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

kan vi när σ är känd direkt använda $\overline{X}$ som teststorhet. Men för att göra det hela systematiskt och analogt med fallet då σ inte är känd skapar vi en testvariabel

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

under förutsättning att H_0 är sann. Vad vi egentligen gör är att vi utnyttjar att $\overline{X}$ är en skattning (konsistent och väntevärdesriktig) av (det okända) väntevärdet μ. Testet går ut på att se om det uppmätta värdet på skattningen sticker ut så mycket från vad som är förväntat att det gör H_0 orimlig.
Det kritiska området C ges av

$$P(Z \in C \,|\, H_0) = \alpha$$

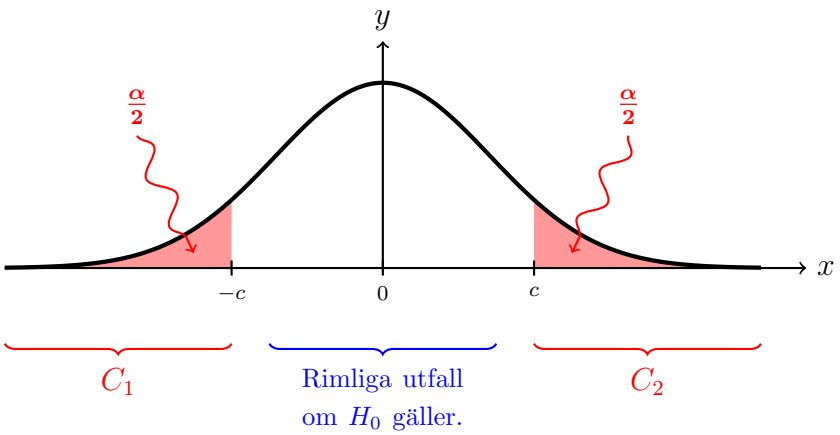
där vi av symmetriskäl (eftersom $Z \sim N(0,1)$) kan – för något $c > 0$ – uttrycka C enligt

$$C = \{z \in \mathbf{R} : |z| > c\} = \{z \in \mathbf{R} : z > c \text{ eller } z < -c\}$$

Vi noterar att C består av två delar C_1 och C_2 där talen i C_1 är negativa och talen i C_2 är positiva. Återigen, av symmetriskäl måste

$$P(Z \in C_1) = P(Z \in C_2) = \frac{\alpha}{2}.$$

Gränsen hittar vi i tabell genom att leta reda på ett tal $c = \Phi^{-1}(1 - \alpha/2)$ (sitter du med MATLAB kan du använda `c = -norminv(alpha/2)`).



2.1.1 Approximativt test via CGS

Som vi såg på förra föreläsningen kan man använda approximationer för att utföra hypotestest. Om vi i vårt fall inte vet att X_i är normalfördelad kan vi ändå via centrala gränsvärdessatsen säga att

$$\overline{X} \overset{\text{appr.}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

om $n \geq 30$ (lite beroende på hur skev fördelningen för X_i är). Som teststorhet använder vi sedan

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \overset{\text{appr.}}{\sim} N(0,1).$$

Notera att vi ersätter σ med s utan att förändra fördelningen (eftersom vi redan håller på med approximationer vet vi inte om det blir bättre med t-fördelningen). Faktum är att vi kan göra detta även om variablerna är lite beroende. Det finns flera varianter av CGS som kan hantera lite olika situationer.

2.2 Okänd varians

Om vi inte känner till σ så kan vi inte direkt använda $\overline{X}$ som teststorhet och inte heller Z från föregående stycka fungera bra (vad ska vi göra med den okända storheten σ?). Vad vi brukar göra är att ersätta σ^2 med stickprovsvariansen s^2, vilket vi tidigare visat leder till t-fördelningen. Så, då gäller att

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

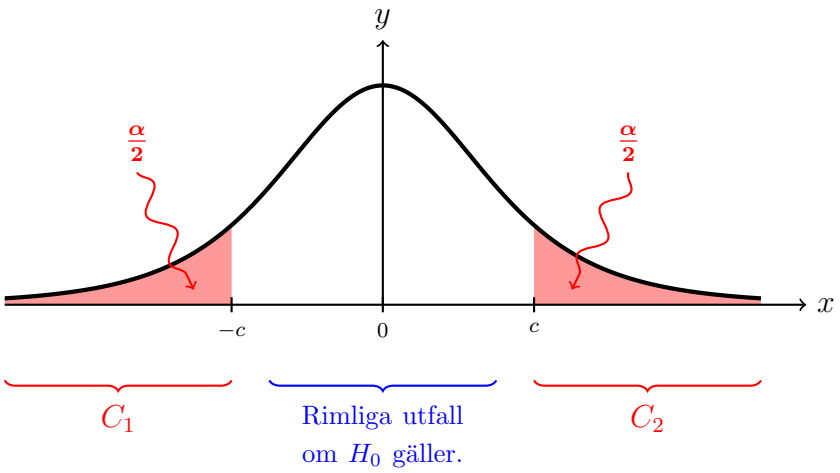
om H_0 är sann. Helt analogt med föregående situation erhåller vi nu

$$C = \{t \in \mathbf{R} : t > c \text{ eller } t < -c\}$$

där

$$P(T > c) = P(T < -c) = \frac{\alpha}{2}.$$

Gränsen hittar vi i tabell genom att leta reda på ett tal $c = F_T^{-1}(1 - \alpha/2)$ (sitter du med MATLAB kan du använda `c = -tinv(alpha/2)`).



2.3 Varianstest

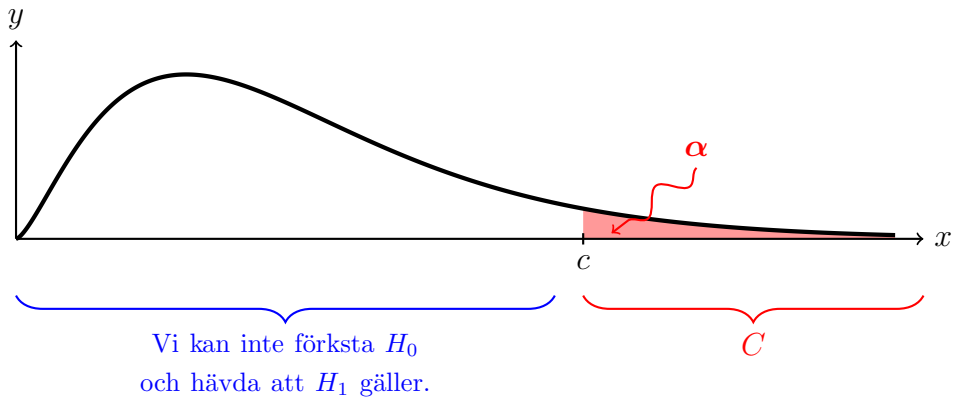
Om någon ställer upp en hypotesen $H_0 : \sigma^2 = \sigma_0^2$, kan vi utföra ett hypotestest? Givetvis kan vi det. Vi vet att

$$V = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

om H_0 är sann. Så detta är en lämplig teststorhet. Alternativt kan man även använda S^2 som teststorhet. Hur ska mothypotesen se ut? Det vanligaste brukar vara $H_1 : \sigma^2 > \sigma_0^2$ för att vi oftast endast är intresserade av att se till att variansen inte blir för stor. Men givetvis kan man även testa mot $H_1 : \sigma^2 \neq \sigma_0^2$ (eller $\sigma^2 < \sigma_0^2$ för den delen). Vi söker nu en gräns c så att

$$\alpha = P(V > c) = P\left(S^2 > \frac{c\sigma_0^2}{n-1}\right)$$

och definierar det kritiska området som $C =]c, \infty[$. Vi använder tabell för att hitta c (eller i MATLAB funktionen `chi2inv(alpha,n-1)`).



Exempel

Tillbaka till fabriken. Efter kritiken så vill man testa om $s = 0.12$ indikerar att det tidigare antagandet $\sigma = 0.1$ är för lågt. Utför testet med signifikansnivå 5%.

Lösning. Låt nollhypotesen vara $H_0 : \sigma = 0.1$ och mothypotesen $H_1 : \sigma > 0.1$. Vi har $n = 9$ och därmed blir

$$V = \frac{8\,S^2}{0.1^2} \sim \chi^2(8).$$

Vi hittar $c = 15.5073$ (chi2inv(0.95,8) eller ur tabell). Vi ser nu att

$$v = \frac{8 \cdot 0.12^2}{0.1^2} = 11.52 < c.$$

Vi kan alltså *inte* förkasta H_0. Betyder det att man hade rätt när fabriken sade att $\sigma = 0.1$?

3 Hypotestester och konfidensintervall

Den observante läsaren har nog redan reflekterat över att det vi ägnat oss åt är ganska snarlikt de föregående föreläsningarna om konfidensintervall. Vi ställer upp liknande storheter (ja, identiska för det mesta) men istället för att stänga in något okänt i ett intervall så testar vi skattningen mot ett kritiskt område.
Ett annat sätt att testa hypoteserna på är att ställa upp konfidensintervall och sedan testa om intervallet täcker nollhypotesen eller ej.

Exempel

I ett husvagnsdrivet laboratorie kokar Janne ihop den suspekta kemikalen $C_{10}H_{15}N$. Varje vecka startar han med samma mängd utgångsmaterial och följer samma procedur. Janne har fått en ny köpare och har hävdat att han kan producera 500 gram i veckan. För att inte riskera problem med hälsan vill Janne testa hypotesen $H_0 : \mu = 500$ mot $H_1 : \mu > 500$ på signifikansnivån 1%. Under 16 veckor producerar han i snitt 525 gram med stickprovsstandardavvikelsen 30 gram.

Lösning. Vi antar normalfördelning och ställer upp ett enkelsidigt konfidensintervall för väntevärdet μ. Låt

$$T = \frac{\overline{X} - \mu}{S/\sqrt{16}} \sim t(15).$$

Då gäller att

$$P(T < t) = 0.99$$

om $t = 2.6025$ (ur tabell). Eftersom

$$T < t \quad \Leftrightarrow \quad \frac{\overline{X} - \mu}{S/4} < t \quad \Leftrightarrow \quad \frac{S\,t}{4} > \overline{X} - \mu \quad \Leftrightarrow \quad \mu > \overline{X} - \frac{S\,t}{4}.$$

så erhåller vi konfidensintervallet

$$I_\mu = \left(\overline{x} - \frac{s\,t}{4},\ \infty\right) = (505.48,\ \infty)\,.$$

Om H_0 är sann så kommer $\mu = 500 \in I_\mu$ med sannolikheten 99%. Eftersom detta inte är sant kan vi förkasta H_0. Ska Janne sitta lugnt i båten att han inte lovat för mycket? Mäter Jannes test rätt sak?

4 Generellt om hypotestester

Innan vi avslutar med en diskussion och tester vid flera stickprov tar vi och summerar lite att tänka på.

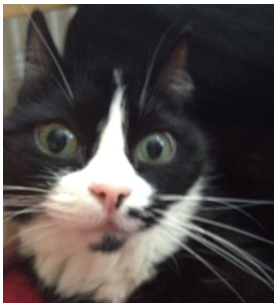
(i) Formulera hypoteser *innan* du tittar på datan. Att du vill göra ett enkelsidigt test ska inte bero på hur datan ser ut. Av den anledningen är den vanligaste typen av tester två-sidiga.

(ii) Men vissa situationer är alltid enkelsida på grund av konstruktion. Vi kommer se det i samband med regressionsanalysen där vi till exempel vet att varians minskar med fler förklaringsvariabler.

(iii) Kom ihåg när hypotestestet ställs upp att det är *mothypotesen* vi vill styrka.

(iv) Var *mycket* försiktig med tolkning av resultaten.

Att man inte förkastar H_0 betyder inte att H_0 gäller. Ett klassiskt exempel handlar om fyrbenta djur: låt H_0 : djuret har fyra ben och H_1 : djuret har inte fyra ben. Vi vill undersöka om en observation är en häst och testar H_0 mot H_1. Bara för att vi inte kan förkasta H_0 när djuret är en katt betyder det inte att det är en häst, eller hur? Kanske ett urartat exempel, det kan vara betydligt mer diffust att läsa av resultaten rätt i andra fall.



(v) Signifikansnivån kan också vara missvisande. Vid stora stickprov kan man ofta se en skillnad och förkasta H_0 även om skillnaden kanske inte spelar någon större roll i praktiska fall.

5 Flera stickprov

Vi kan givetvis betrakta flera stickprov samtidigt. Ofta är man intresserade av att testa om de har samma väntevärde och/eller samma varians. Men vi kan ställa upp test för linjärkombinationer av väntevärdena direkt.
Låt $X_1, X_2, \ldots, X_m$ och $Y_1, Y_2, \ldots, Y_n$ vara stickprov från $N(\mu_X, \sigma_X^2)$ respektive $N(\mu_Y, \sigma_Y^2)$. Vi kan ställa upp hypotestester för linjärkombinationen $c_1\mu_X + c_2\mu_Y$. Om varianserna är kända kan vi direkt använda att

$$Z = \frac{c_1\overline{X} + c_2\overline{Y} - (c_1\mu_X + c_2\mu_Y)}{\sqrt{c_1^2\sigma_X^2/m + c_2^2\sigma_Y^2/n}} \sim N(0,1),$$

alltså precis samma variabel vi såg när vi tog fram konfidensintervall för $c_1\mu_X + c_2\mu_Y$.

5.1 $\sigma_X = \sigma_Y = \sigma$ okänd

Helt analogt med motsvarande situation när vi tog fram konfidensintervall använder vi att

$$T = \frac{c_1\overline{X} + c_2\overline{Y} - (c_1\mu_X + c_2\mu_Y)}{S\sqrt{c_1^2/m + c_2^2/n}} \sim t(m+n-2),$$

där S^2 är den sammanvägda variansskattningen. Vi betraktar ett exempel.

Exempel

Janne har fått konkurens av den före detta lärlingen Rossana som använder samma metod. Under 9 veckor producerar hon i snitt 600 gram med en stickprovsstandardavvikelse på 50 gram. Testa på signifikansnivån 1% hypotesen $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 < \mu_2$ med antagandet att variansen är densamma, där μ_1 är Jannes förväntade värde och μ_2 är Rossanas. Borde köparen byta leverantör?

Lösning. Vi formulerar om enligt $H_0 : \mu_2 - \mu_1 = 0$ mot $H_1 : \mu_2 - \mu_1 > 0$. Om H_0 är sann så gäller att

$$T = \frac{\overline{Y} - \overline{X}}{S\sqrt{1/9 + 1/16}} = \frac{\overline{Y} - \overline{X}}{0.4167\,S} \sim t(23),$$

och det kritiska området blir

$$C = \{t \in \mathbf{R} : t > 2.4999\}$$

eftersom $P(T < 2.4999) = 0.99$. Med uppmätta siffrorna blir

$$t = \frac{600 - 525}{0.4167 \cdot s_p} = \frac{75}{0.4167 \cdot 38.16} = 4.7161,$$

där

$$s_p^2 = \frac{15\,s_1^2 + 8\,s_2^2}{23} = 38.16^2$$

är den sammanvägda variansskattningen. Eftersom $4.7161 \in C$ förkastar vi H_0. Blir det samma resultat om vi testar mot $H_1 : \mu_1 \neq \mu_2$? (svar: ja, förkasta H_0. Vad ändras?)

Vi kan givetvis ta fram ett konfidensintervall $I_{\mu_2-\mu_1}$ och testa nollhypotesen genom att undersöka om $0 \in I_{\mu_2-\mu_1}$ också (förkasta H_0 om $0 \notin I_{\mu_2-\mu_1}$).

5.2 Test för variansskillnad

Så med föregående avsnitt i tankarna är en rimlig fråga om vi kanske kan testa huruvida varianserna är lika eller inte. Vi gör detta med ett så kallat F-test. Vi testar hypotesen

$$H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2$$

mot

$$H_1 : \sigma_X^2 \neq \sigma_Y^2.$$

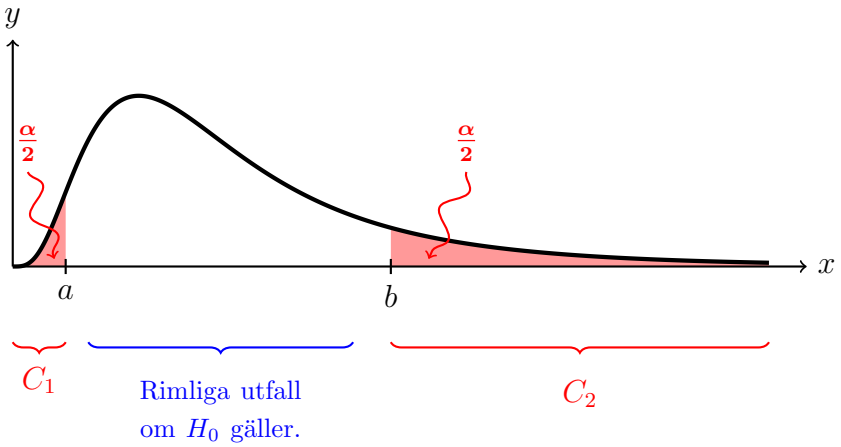
Om H_0 är sann, så gäller att $(m-1)S_X^2/\sigma^2 \sim \chi^2(m-1)$ och $(n-1)S_Y^2/\sigma^2 \sim \chi^2(n-1)$. Enligt tidigare resultat vet vi att följande då gäller:

$$V = \frac{\dfrac{(m-1)S_X^2}{\sigma^2}/(m-1)}{\dfrac{(n-1)S_Y^2}{\sigma^2}/(n-1)} = \frac{S_X^2}{S_Y^2} \sim F(m-1, n-1).$$

Vi söker nu ett kritiskt område C så att

$$\alpha = P(V \in C \,|\, H_0)$$

och som vi anser styrker H_1 om vi får utfall där. En figur kan vara lämplig för att se hur fördelningen ser ut.



Vi hittar gränser a och b ur tabell (eller med `finv(p, n-1, m-1)` i MATLAB) så att

$$P(V < a) = P(V > b) = \frac{\alpha}{2}.$$

Exempel

Var det rimligt att anta att Jannes och Rossanas tillvägagångssätt hade samma varians? Testa med signifikansnivån 5%.

Lösning. Med $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ och $H_1 : \sigma_1^2 \neq \sigma_2^2$. Med

$$V = \frac{S_1^2}{S_2^2} \sim F(15, 8)$$

hittar vi det kritiska området

$$C = \{v \in [0, \infty[: v < a \text{ eller } v > b\}$$

med $a = 0.3126$ (`finv`(0.025, 15, 8)) och $b = 4.1012$ (`finv`(0.975, 15, 8)). Eftersom

$$v = \frac{30^2}{50^2} = 0.36 \notin C$$

kan vi inte förkasta H_0. Varianserna kan vara lika (men är de det?). Vad händer på signifikansnivån 1%? (samma sak, kan man säga det utan att räkna?)

Föreläsning 7: Stokastiska vektorer

Johan Thim (johan.thim@liu.se)

27 september 2018

1 Repetition

Definition. Låt X och Y vara stokastiska variabler med $E(X) = \mu_X$, $V(X) = \sigma_X^2$, $E(Y) = \mu_Y$ samt $V(Y) = \sigma_Y^2$. Kovariansen $C(X,Y)$ definieras enligt

$$C(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big)$$

och korrelationen mellan X och Y enligt

$$\rho(X,Y) = \frac{C(X,Y)}{\sigma_X\,\sigma_Y}.$$

Både kovarians och korrelation är ett mått på linjärt beroende mellan X och Y där korrelationen är normerad så det går att jämföra olika fall. Vi listar lite kända egenskaper.

(i) Om $C(X,Y) = 0$ kallas X och Y för okorrelerade.

(ii) $C(X,Y) = E(XY) - E(X)E(Y)$.

(iii) Om X och Y är oberoende så är $C(X,Y) = 0$.

(iv) $C(X,X) = V(X)$.

(v) $C\left(a_0 + \sum_{i=1}^{m} a_i X_i,\, b_0 + \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j C(X_i, Y_j)$.

(vi) $|\rho(X,Y)| \leq 1$ med likhet om och endast om det finns ett linjärt samband mellan X och Y.

Observera att $C(X,Y) = 0$ *inte* nödvändigtvis innebär oberoende. Låt till exempel $X \sim \mathrm{Re}(-1,1)$ och $Y = X^2$. Uppenbarligen beroende variabler, men

$$C(X,Y) = E(XY) - E(X)E(Y) = E(X^3) - 0\cdot E(Y) = E(X^3) = \int_{-1}^{1} x^3\cdot\frac{1}{2}\,dx = 0,$$

så X och Y är okorrelerade.

2 Vektorer av stokastiska variabler

Låt $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ vara en vektor vars komponenter är stokastiska variabler. Vi strävar efter att skriva vektorer som kolonnvektorer. Det faller sig naturligt att definiera väntevärdet av $\boldsymbol{X}$ genom **väntevärdesvektorn**

$$E(\boldsymbol{X}) = (E(X_1), E(X_2), \ldots, E(X_n)).$$

På samma sätt definierar vi väntevärdet av en matris av stokastiska variabler. Variansen blir lite konstigare så vi introducerar **kovariansmatrisen** mellan två vektorer (av samma dimension). Låt $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ och definiera $C(\boldsymbol{X}, \boldsymbol{Y})$ enligt

$$C(\boldsymbol{X}, \boldsymbol{Y}) = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} = \begin{pmatrix} C(X_1, Y_1) & C(X_1, Y_2) & \cdots & C(X_1, Y_n) \\ C(X_2, Y_1) & C(X_2, Y_2) & \cdots & C(X_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(X_n, Y_1) & C(X_n, Y_2) & \cdots & C(X_n, Y_n) \end{pmatrix}$$

där c_{ij} är kovariansen mellan X_i och Y_j.
En stor anledning att blanda in vektorer och matriser är givetvis att få tillgång till maskineriet från linjär algebra. Kovariansen mellan två vektorer $\boldsymbol{X}$ och $\boldsymbol{Y}$ kan då lite mer kompakt skrivas

$$C(\boldsymbol{X}, \boldsymbol{Y}) = E(\boldsymbol{X}\boldsymbol{Y}^T) - E(\boldsymbol{X})E(\boldsymbol{Y})^T,$$

där $(\,\cdot\,)^T$ innebär transponering. En produkt $A = \boldsymbol{x}\boldsymbol{y}^T$ brukar kallas för den yttre produkten och består av element $(a)_{ij} = x_i y_j$, $i, j = 1, 2, \ldots, n$. Detta är alltså *inte* skalärprodukten $(\boldsymbol{X}^T\boldsymbol{Y})$.
Låt $A, B \in \mathbf{R}^{n\times n}$ vara matriser. Då är $A\boldsymbol{X}$ en linjärkombination av $X_1, X_2, \ldots, X_n$ och $B\boldsymbol{Y}$ en linjärkombination av $Y_1, Y_2, \ldots, Y_n$. Dessutom kan *alla* linjärkombinationer skrivas på detta sätt. Vidare gäller nu tack varje linjäriteten att

$$E(A\boldsymbol{X}) = AE(\boldsymbol{X}) \quad \text{och} \quad C(A\boldsymbol{X}, B\boldsymbol{Y}) = A\boldsymbol{X}(B\boldsymbol{Y})^T = A\boldsymbol{X}\boldsymbol{Y}^T B^T.$$

Exempel

Låt $\boldsymbol{X} = (X_1\ X_2)^T$ vara en stokastisk variabel med $E(\boldsymbol{X}) = (1\ 2)^T$ och $C_{\boldsymbol{X}} = \begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix}$.
Hitta en prediktor $X_2 = aX_1 + b$ så att $E(aX_1 + b) = E(X_2)$ och $V(X_2 - (aX_1 + b))$ är minimal.

Lösning. Vi ser direkt att

$$E(aX_1 + b) = aE(X_1) + b = a + b \quad \text{och} \quad E(X_2) = 2,$$

så $a + b = 2$. Vidare gäller att

$$X_2 - (aX_1 + b) = \begin{pmatrix} -a & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} - b,$$

så

$$\begin{aligned} V(X_2 - aX_1 - b) &= V\left(\begin{pmatrix} -a & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \right) = \begin{pmatrix} -a & 1 \end{pmatrix} C_{\boldsymbol{X}} \begin{pmatrix} -a \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} -a & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix} \begin{pmatrix} -a \\ 1 \end{pmatrix} = a^2 + 4a + 4 = (a+2)^2. \end{aligned}$$

Minimum sker uppenbarligen när $a = -2$, vilket ger att $b = 4$.

3 Skattningar för kovarians och korrelation

Om vi har ett stickprov (x_k, y_k), $k = 1, 2, \ldots, n$, där (X_k, Y_k) är stokastiska variabler med samma fördelning, så skattar vi kovariansen C med

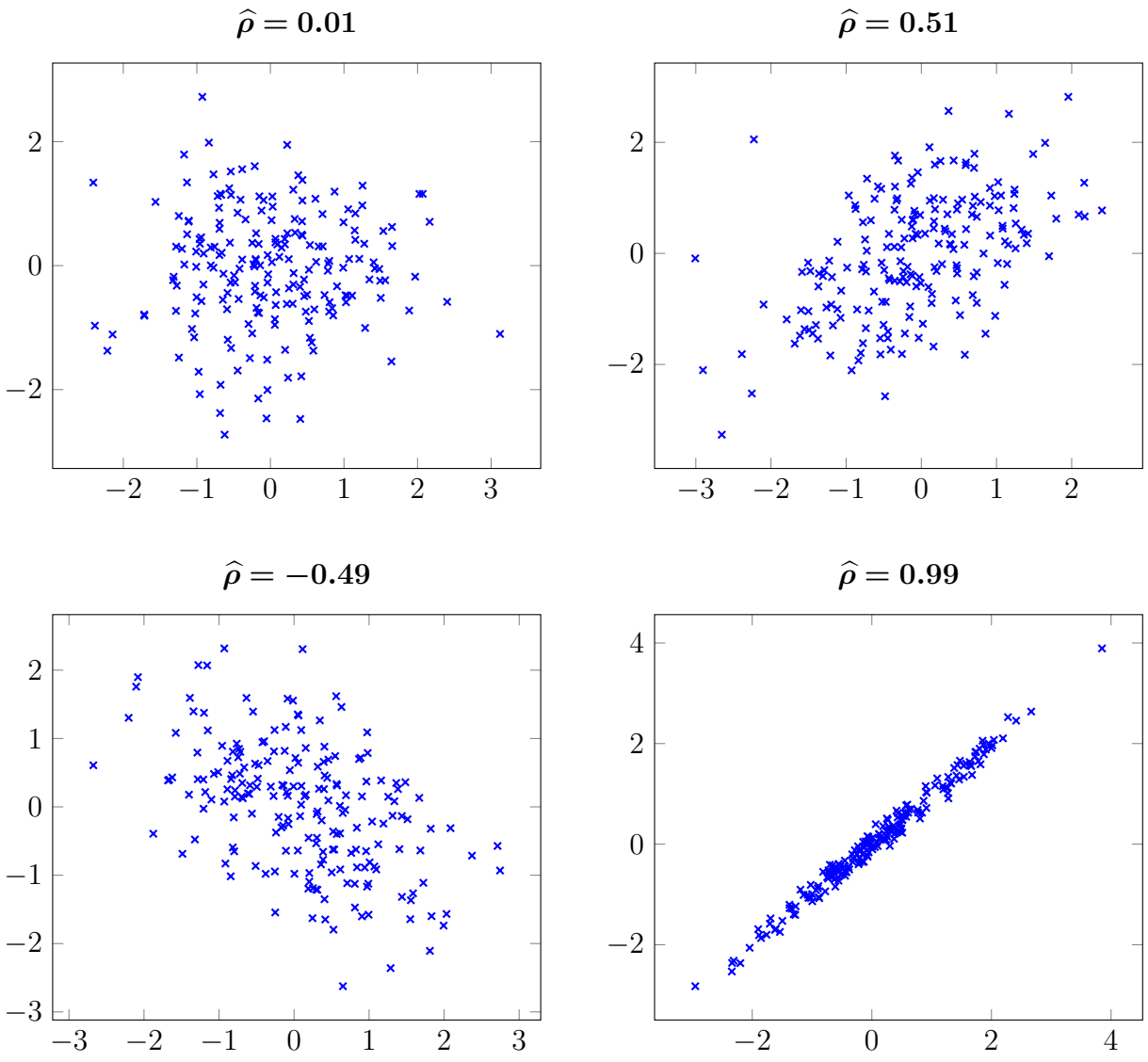
$$\widehat{c} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

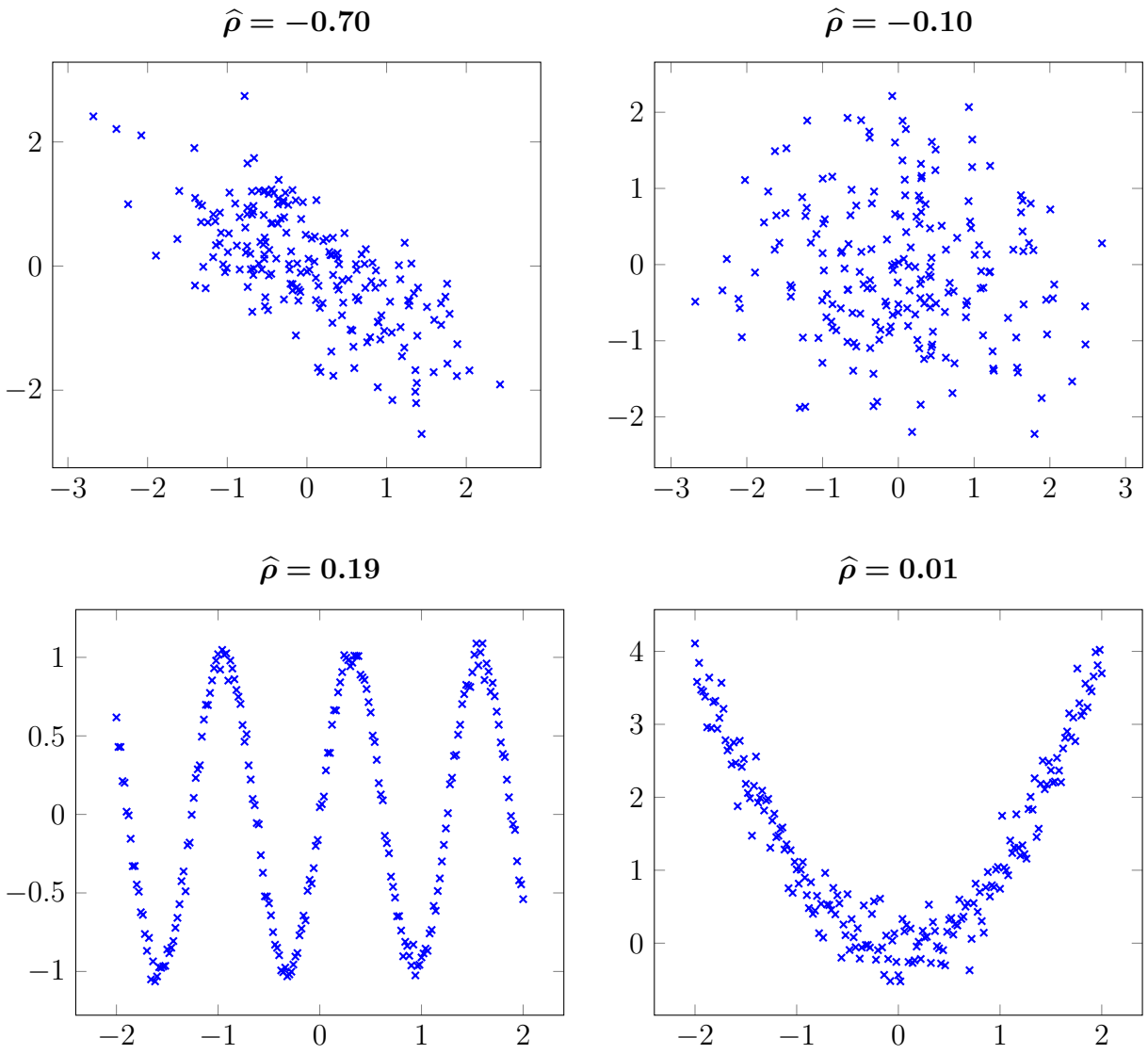
och korrelationen med

$$\widehat{\rho} = \frac{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})}{\left(\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2 \right)^{1/2} \left(\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2 \right)^{1/2}}.$$

Av tradition betecknar man ofta $\widehat{\rho} = r$. En naturlig fråga i detta skede är om vi kan säga något om fördelningen för den skatta korrelationen under något lämpligt antagande om det slumpmässiga stickprovet. Vi återkommer i fallet med normalfördelning i nästa avsnitt.

3.1 Vad innebär korrelationen grafiskt?





4 Multivariat normalfördelning

"Pain has a face. Allow me to show it to you."
–Pinhead

Vi har stött på den flerdimensionella normalfördelningen tidigare, men vi kan formulera det hela lite mer kompakt på följande sätt.

Multivariat normalfördelning

Definition. Vi säger att $\boldsymbol{Y}$ har en **multivariat normalfördelning** om det finns en konstant vektor $\boldsymbol{\mu} \in \mathbf{R}^n$ och en konstant matris $A \in \mathbf{R}^{n \times m}$ så att $\boldsymbol{Y} = \boldsymbol{\mu} + A\boldsymbol{X}$, där $\boldsymbol{X}$ är en vektor med stokastiska variabler, $\boldsymbol{X} = (X_1\ X_2\ \cdots\ X_n)^T$, och $X_i \sim N(0,1)$ är oberoende.

Är definitionen vettig? Ja, den reducerar åtminstone till det förväntade resultatet om $n = 1$: $Y = \mu + \sigma^2 X$ där $X \sim N(0,1)$. Vidare gäller så klart att

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} + AE(\boldsymbol{X}) = \boldsymbol{\mu}$$

och

$$C_{\boldsymbol{Y}} = AC_{\boldsymbol{X}}A^T = AA^T$$

eftersom $C_{\boldsymbol{X}}$ är identitetsmatrisen (variablerna är oberoende om har varians 1).

Exempel

Låt $\boldsymbol{X} \sim \mathrm{N}\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$. Bestäm fördelningen för $Y = X_1 + X_2$.

Lösning. Vi skriver $Y = (1\ 1)(X_1\ X_2)^T = A\boldsymbol{X}$. Då blir

$$E(Y) = AE(X) = (1\ 1)(1\ \ -1) = 0$$

och

$$C_Y = AC_XA^T = (1\ 1)\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}(1\ 1)^T = 5.$$

Sats. Om $\boldsymbol{Y}$ har väntevärdesvektorn $\boldsymbol{\mu}$ och en kovariansmatris C som uppfyller att $\det C \neq 0$ så gäller att $\boldsymbol{Y}$ har multivariat normalfördelning om och endast om $\boldsymbol{Y}$ har den simultana täthetsfunktionen

$$f_{\boldsymbol{Y}}(y_1, y_2, \ldots, y_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \quad \boldsymbol{y} \in \mathbf{R}^n.$$

Bevis. Eftersom kovariansmatrisen C alltid är positivt semidefinit (varför?) och vi antar att determinanten $|C| := \det C \neq 0$, så är C positivt semidefinit och då finns alltid en inverterbar matris $A \in \mathbf{R}^{n\times n}$ sådan att $C = AA^T$. Definiera $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{\mu}$, där $\boldsymbol{X} = (X_1\ X_2\ \cdots\ X_n)^T$ och $X_k \sim N(0,1)$ är oberoende. Täthetsfunktionen för $\boldsymbol{X}$ ges då av

$$f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right), \quad \boldsymbol{x} \in \mathbf{R}^n.$$

Enligt transformationssatsen för flerdimensionella stokastiska variabler så kommer

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left(\left(A^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)\left|\frac{d(x_1, x_2, \ldots, x_n)}{d(y_1, y_2, \ldots, y_n)}\right|\right.$$

eftersom $\boldsymbol{X} = A^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$. Vi ser att jacobianen ges av

$$\frac{d(x_1, x_2, \ldots, x_n)}{d(y_1, y_2, \ldots, y_n)} = |A^{-1}| = |A|^{-1},$$

så

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \frac{1}{(2\pi)^{n/2}}|A|^{-1}\exp\left(-\frac{1}{2}\left(A^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)^T\left(A^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)\right) \\ &= \frac{1}{(2\pi)^{n/2}}|AA^T|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T(A^{-1})^TA^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \\ &= \frac{1}{(2\pi)^{n/2}}|AA^T|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^TC^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \end{aligned}$$

där vi utnyttjat att $|A|^{-1} = |A|^{-1/2}|A^T|^{-1/2} = |AA^T|^{-1/2}$.
Omvänt, om $\boldsymbol{Y}$ är normalfördelad så säger definitionen att det finns en matris $A \in \mathbf{R}^{n\times m}$ och en vektor $\boldsymbol{\mu} \in \mathbf{R}^n$ så att $Y = A\boldsymbol{X} + \boldsymbol{\mu}$ för $\boldsymbol{X} = (X_1\ X_2\ \cdots\ X_m)^T$ där $X_k \sim N(0,1)$ är oberoende. Faktum är att $m = n$ är nödvändigt då $C = AA^T$ antas vara inverterbar, eftersom $n = \operatorname{rank}(AA^T) \leq \min\{\operatorname{rank}(A), \operatorname{rank}(A^T)\}$ (ty vid produkter av matriser vinner alltid den med lägst rank) och $\operatorname{rank}(A^T) = \operatorname{rank}(A)$, så $\operatorname{rank}(A) = n$ eftersom vi har n kolonner. Samma argument som ovan visar nu att täthetsfunktionen ges av uttrycket i satsen. □

Exempel

Låt $X_1,\ X_2 \sim \mathrm{N}(0,1)$ vara oberoende och definiera $\boldsymbol{Y} = (X_1 - X_2,\ 2X_1 + X_2)$. Bestäm täthetsfunktionen för $\boldsymbol{Y}$.

Lösning. Vi skriver

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = A \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \text{där } A = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix}.$$

Då blir

$$\boldsymbol{\mu_Y} = E\left(A \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right) = A \cdot \mathbf{0} = \mathbf{0}$$

och

$$C_{\boldsymbol{Y}} = AC_{\boldsymbol{X}}A^T = A \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} A^T = AA^T = \begin{pmatrix} 2 & 1 \\ 1 & 5 \end{pmatrix}.$$

Således blir $\det C_{\boldsymbol{Y}} = 9$ och

$$C_{\boldsymbol{Y}}^{-1} = \frac{1}{9} \begin{pmatrix} 5 & -1 \\ -1 & 2 \end{pmatrix}.$$

Alltså blir

$$f_{\boldsymbol{Y}}(y_1, y_2) = \frac{1}{6\pi} \exp\left(-\frac{1}{18}\left(5y_1^2 - 2y_1y_2 + 2y_2^2\right)\right)$$

ty

$$(y_1\ y_2)\, \frac{1}{9} \begin{pmatrix} 5 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \frac{1}{9}\left(5y_1^2 - 2y_1y_2 + 2y_2^2\right).$$

Sats. Låt $\boldsymbol{Z} = \boldsymbol{d} + B\boldsymbol{Y}$, där $\boldsymbol{Y}$ är multivariat normalfördelad. Då är även $\boldsymbol{Z}$ multivariat normalfördelad.

Bevis. Följer direkt från definitionen.

Sats. För $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, C)$ gäller att komponenterna i $\boldsymbol{Y}$ är oberoende om och endast om C är en diagonalmatris (under förutsättning att A är inverterbar).

Bevis. Kravet på att A ska vara inverterbar följer av att om så icke är fallet så är fördelningen degenererad eftersom $A\boldsymbol{x} = 0$ har oändligt många lösningar. Det är alltså självklart i detta läge att komponenterna i $\boldsymbol{Y}$ inte kan vara oberoende. Så antag nu att det $A \neq 0$.
Den ena riktningen är mer eller mindre självklar eftersom om komponenterna i $\boldsymbol{Y}$ är oberoende kommer $C(Y_i, Y_j) = 0$ för $i \neq j$ och $C(Y_i, Y_i) = \sigma_i^2$, så $C_{\boldsymbol{Y}}$ blir en diagonalmatris.
Antag nu att $C_{\boldsymbol{Y}}$ är en diagonalmatris, säg

$$\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}.$$

Eftersom $C_{\boldsymbol{Y}} = AA^T$ kommer $C_{\boldsymbol{Y}}$ att vara inverterbar, vilket innebär att samtliga $\sigma_i^2 \neq 0$. Inversen $C_{\boldsymbol{Y}}^{-1}$ är även den en diagonalmatrisen med diagonalelementen σ_i^{-2}. Således blir den simultana täthetsfunktionen

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det C_{\boldsymbol{Y}}}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma_1\sigma_2\cdots\sigma_n} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}(y_j-\mu_j)\sigma_j^{-2}(y_j-\mu_j)\right) \\ &= \prod_{j=1}^{n} \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(y_j-\mu_j)^2}{2\sigma_j^2}\right) = \prod_{j=1}^{n} f_{Y_j}(y_j). \end{aligned}$$

Eftersom den simultana täthetsfunktionen ges av produkten av täthetsfunktionerna för Y_j följer det att variablerna är oberoende. □

4.1 Bivariat normalfördelning

Specialfallet när $n = 2$ förtjänar lite kommentarer eftersom den situationen frekvent dyker upp. Låt (X, Y) vara normalfördelad med väntevärdesvektor och kovariansmatris enligt

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \quad \text{och} \quad \begin{pmatrix} \sigma_X^2 & C(X,Y) \\ C(Y,X) & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

Täthetsfunktionen ges enligt ovan av

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\,\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right),$$

för $(x, y) \in \mathbf{R}^2$.
Vi ser direkt att om $\rho = 0$ blir det produkten av täthetsfunktionerna för två oberoende variabler, precis som satsen i föregående avsnitt påstod. Men vad händer om variablerna inte är oberoende, dvs om $\rho \neq 0$ (oberoende och okorrelerade är ekvivalent i *normalfördelningsfallet*)? Låt oss beräkna den marginella tätheten $f_X(x)$ bara för kul (fast vi har nytta av den snart..). För att underlätta notationen låter vi

$$u = \frac{x-\mu_X}{\sigma_X} \quad \text{och} \quad v = \frac{y-\mu_Y}{\sigma_Y}.$$

Vi har nu

$$\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 = u^2 - 2\rho uv + v^2 = (v-\rho u)^2 + (1-\rho^2)u^2,$$

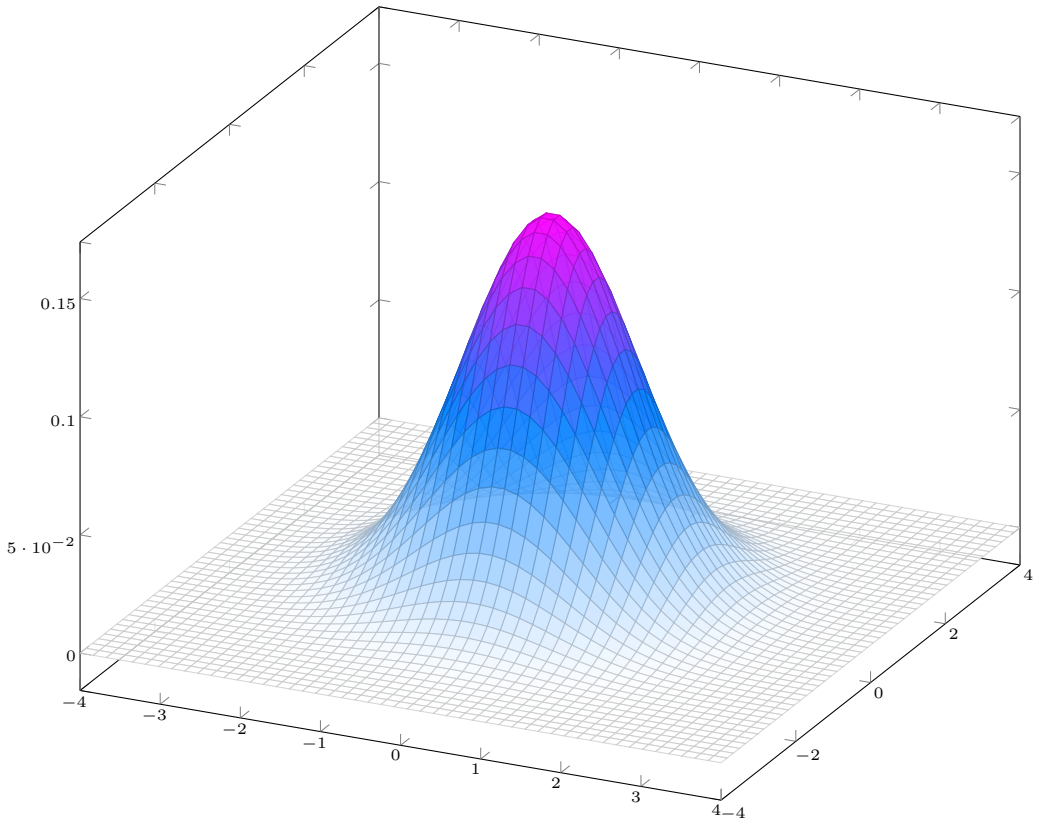
så

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}u^2\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}(v-\rho u)^2\right) dy \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{1}{2}u^2\right) \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}(v-\rho u)^2\right) dv \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{1}{2}u^2\right), \end{aligned}$$

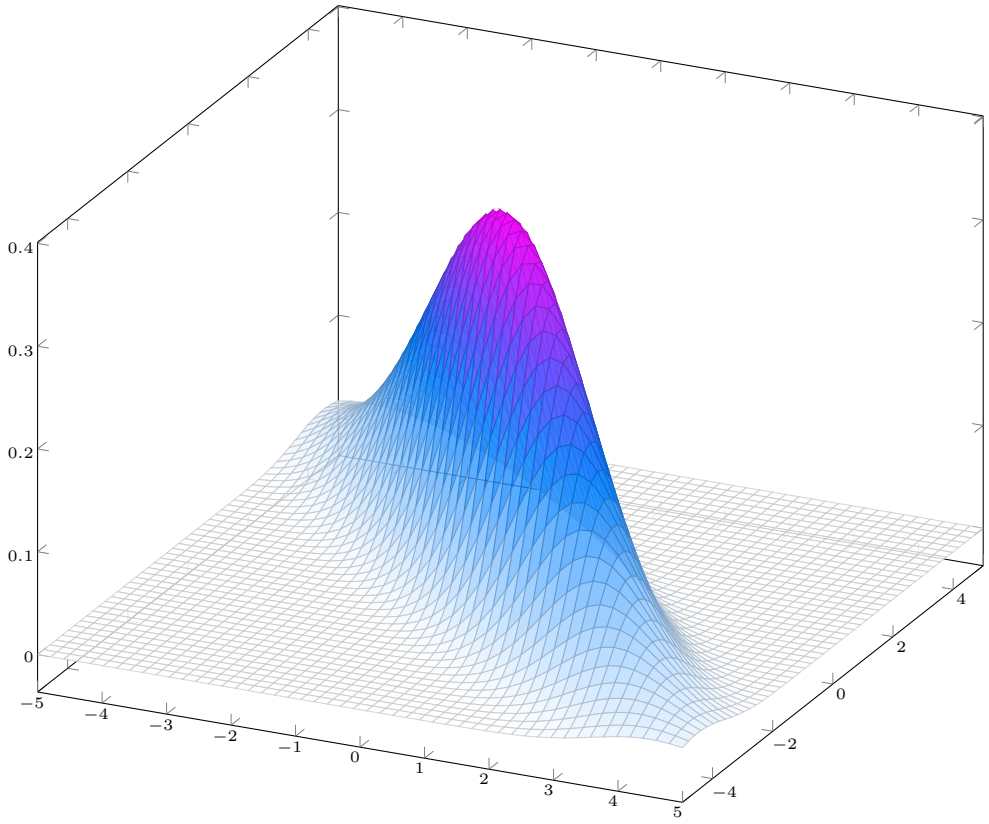
ty

$$\frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}(v-\rho u)^2\right) dv = 1.$$

Hur ser bivariata normalfördelningar ut? Om $\sigma_X = \sigma_Y = 1$ och $\rho = 0$ får vi följande figur:



och med $\sigma_X = \sigma_Y = 1$ och $\rho = 0.9$ erhåller vi



4.2 Test för $\rho = 0$

Att direkt ge sig på uttrycket för $\widehat{\rho}$ är komplicerat, så vi börjar lite annorlunda. Låt (X, Y) vara bivariat normalfördelad. Då har (X, Y) en simultan täthetsfunktion $f(x, y)$ och den betingade (på $X = x$) täthetsfunktionen blir

$$f_{Y \mid X=x}(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(v-\rho u\right)^2\right),$$

vilket är tätheten för en normalfördelad variabel $Y \mid X = x$ med

$$E(Y \mid X = x) = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X + \rho\frac{\sigma_Y}{\sigma_X}x = \beta_0 + \beta_1 x$$

och

$$V(Y \mid X = x) = \sigma_Y^2(1-\rho^2).$$

Det betingade (för givet X) väntevärdet är alltså en rät linje $y = \beta_0 + \beta_1 x$. Intressant! Åter igen något som är halvmagiskt för normalfördelningen (det finns ingen fördelning ni kan misshandla lika mycket). Den observante läsaren funderar nog även om detta har med regressionsanalysen att göra, vilket vi kommer till nästa föreläsning. För nuvarande situation, notera specifikt att

$$\beta_1 = \rho\,\frac{\sigma_Y}{\sigma_X}.$$

Anledningen till denna manöver är att vi hellre betraktar tester för β_1 än direkt för ρ. Varför? Det har med ovanstående att göra (linjär regression). Tänk tillbaka till anda föreläsningen. Där visade vi att MK-skattningen $\widehat{\beta_1}$ av β_1 ges av

$$\widehat{\beta_1} = \frac{\sum_{j=1}^n (x_j - \overline{x})(Y_j - \overline{Y})}{\sum_{j=1}^n (x_j - \overline{x})^2}$$

och att $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\,\overline{x}$. Anledning till att blanda in detta är att vi på nästa föreläsning kommer att visa att

$$\widehat{\beta}_1 \sim N\left(\beta_1,\ \frac{\sigma^2}{\sum_{j=1}^n (x_j-\overline{x})^2}\right).$$

Vi introducerar lite förenklande beteckningar. Låt

$$S_x^2 = \frac{1}{n-1}\sum_{j=1}^n (x_j-\overline{x})^2, \quad S_y^2 = \frac{1}{n-1}\sum_{j=1}^n (Y_j-\overline{Y})^2 \quad \text{samt} \quad S_{xy} = \frac{1}{n-1}\sum_{j=1}^n (x_j-\overline{x})(Y_j-\overline{Y}).$$

Notera nu att vi kan skriva R (den stokastiska motsvarigheten till $\widehat{\rho}$) som

$$R = \frac{S_{xy}}{S_x\,S_y}$$

och därmed blir

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_x^2} = R\frac{S_y}{S_x}.$$

Vi introducerar det totala kvadratfelet, dvs summan av kvadraterna på skillnaden mellan mätvärden Y_j och de skattade värdena $\widehat{\beta}_0 + \widehat{\beta}_1 x_j$:

$$\mathrm{SS_E} = \sum_{j=1}^n (Y_j - \widehat{\beta}_0 - \widehat{\beta}_1 x_j)^2.$$

Vi kommer (även detta nästa föreläsning) att visa att $\mathrm{SS_E}$ är oberoende av $\widehat{\beta}_1$ och att $\frac{1}{\sigma^2}\mathrm{SS_E} \sim \chi^2(n-2)$. Vidare har denna storhet egenskapen att

$$\mathrm{SS_E} = (n-1)S_Y^2(1-R^2)$$

vilket kan ses genom att expandera kvadraten i summan som definierar $\mathrm{SS_E}$:

$$\begin{aligned}\mathrm{SS_E} &= (n-1)\left(S_y^2 - 2\widehat{\beta}_1 S_{xy} + \widehat{\beta}_1{}^2 S_x^2\right)\\ &= (n-1)\left(S_y^2 - 2R\frac{S_y}{S_x}S_{xy} + R^2\frac{S_y^2}{S_x^2}S_x^2\right) = (n-1)S_y^2(1-R^2).\end{aligned}$$

Det följer då (Gossets sats) att

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{n-2}\,\mathrm{SS_E}/(n-1)S_x^2}} \sim t(n-2).$$

Om nu $\rho = 0$ (vilket innebär att $\beta_1 = 0$ enligt ovan) så gäller att identiteten

$$\frac{\widehat{\beta}_1}{\sqrt{\frac{1}{n-2}\mathrm{SS_E}/((n-1)S_X^2)}} = \frac{rS_Y/S_X}{\sqrt{\frac{(n-1)S_Y^2(1-r^2)}{(n-2)(n-1)S_X^2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

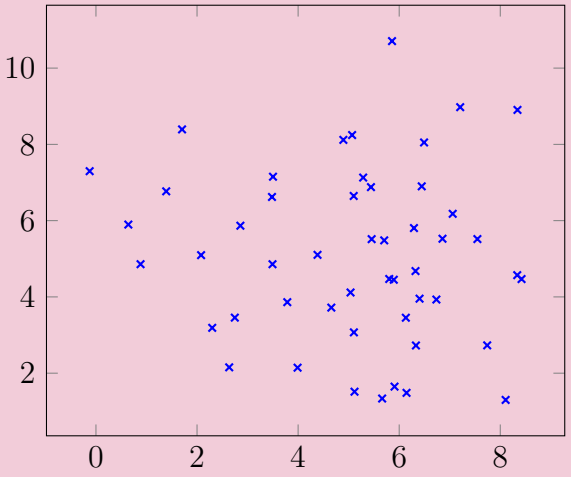
medför att

$$\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t(n-2).$$

Vi kan alltså använda denna storhet för att testa hypotesen $H_0 : \rho = 0$.

Exempel

Astrid och Åsa grälar om två variabler är okorrelerade eller inte. Vid 50 mätningar av två variabler X och Y erhöll de diagrammet till höger som spridningsplott. Den empiriska korrelationen beräknades till $\widehat{\rho} = -0.0838$. Astrid hävdar att det tyder på att $\rho = 0$ medan Åsa anser att det absolut är signifikant (om än lågt pga slumpen). Om vi antar att X och Y är normalfördelade, testa hypotesen att $H_0 : \rho = 0$ mot $H_1 : \rho \neq 0$ med signifikansnivån 5%.



Lösning. Med 50 mätningar av (X, Y) blir

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t(48)$$

om H_0 är sann. Kritiskt område erhålls därmed som

$$C = \{t \in \mathbf{R} : |t| > 2.0106\}$$

ty $P(T \leq 2.0106) = 0.975$. Med det uppmätta $r = -0.0838$ blir

$$t = \frac{-0.0838\sqrt{48}}{\sqrt{1-(-0.0838)^2}} = -0.5826.$$

Eftersom $t \notin C$ så kan vi inte förkasta H_0. Variablerna kan mycket väl vara okorrelerade (men vi vet inte det!).

5 Bonus: fördelningen för R

Sats. Om $\rho = 0$ ges fördelningen för R av täthetsfunktionen

$$f_R(r) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)\sqrt{\pi}}(1-r^2)^{(n-4)/2}, \quad -1 < r < 1.$$

Bevis. Eftersom $g(s) = \dfrac{s}{\sqrt{1-s^2}}$ är en strängt växande funktion för $-1 < s < 1$ så gäller att

$$F_R(r) = P(R \leq r) = P\left(\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \leq \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right) = F_T\left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right),$$

där F_T är fördelningsfunktionen för en $t(n-2)$-fördelad variabel. Denna funktion är en integral av en kontinuerlig täthet, så vi kan derivera fram

$$\begin{aligned} f_R(r) =& \frac{d}{dr}\left(F_T\left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)\right) = f_T\left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)\frac{\sqrt{n-2}}{(1-r^2)^{3/2}} \\ =& \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}\left(1+\frac{r^2}{1-r^2}\right)^{-(n-1)/4}(1-r^2)^{-3/2} \\ =& \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}\left(1-r^2\right)^{(n-1)/4}(1-r^2)^{-3/2} = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}\left(1-r^2\right)^{(n-4)/4}, \end{aligned}$$

vilket är täthetsfunktionen given i satsen. □

Vad händer om $\rho \neq 0$? En fullt rimlig fråga, men fördelningen har inget trevligt utseende då (inkluderar hypergeometriska funktioner).

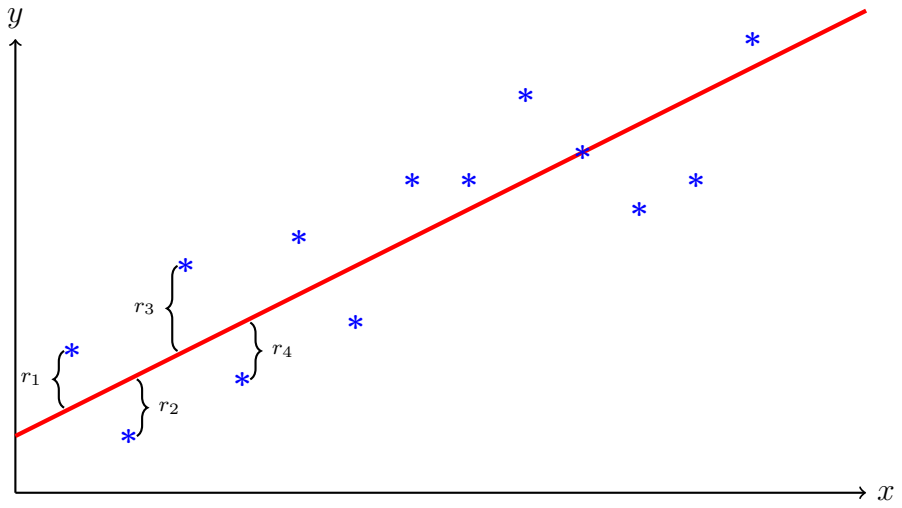
Föreläsning 8: Linjär regression – del I

Johan Thim (johan.thim@liu.se)

29 september 2018

"Your suffering will be legendary, even in hell."
–Pinhead

Vi återgår nu till ett exempel vi stött på redan vid ett flertal tillfällen (föregående föreläsning och föreläsning 2 för att inte tala om tidigare kurser som linjär algebra), nämligen att anpassa en rät linje $y = \beta_0 + \beta_1 x$ efter mätdata (x_i, y_i), $i = 1, 2, \ldots, n$. Grafiskt illustrerat enligt nedan.



Målsättningen är att – givet en mätserie – hitta den linje som approximerar denna serie på lämpligt sätt. Det resulterar i ett par naturliga funderingar.

(i) Hur hittar man en approximativ linje systematiskt?

(ii) Om man upprepar försöket, får man samma linje?

(iii) I vilken mening är linjen optimal? På vilket sätt mäter vi avvikelserna mellan linjen och mätserien?

Vi ska försöka svara på dessa frågor och för att göra det behöver vi ställa upp en modell.

Enkel linjär regression

Definition. Vi kommer betrakta följande modell: givet (x_j, y_j), $j = 1, 2, \ldots, n$, där vi betraktar x_j som fixerade och y_j som observationer av stokastiska variabler

$$Y_i = \beta_0 + \beta_1 x_j + \epsilon_j, \quad j = 1, 2, \ldots, n,$$

där $\epsilon_j \sim N(0, \sigma^2)$ antas oberoende (och likafördelade). Den räta linjen $y = \beta_0 + \beta_1 x$ kallas **regressionslinjen**.
Vi använder beteckningen $\mu_j = \beta_0 + \beta_1 x_j = E(Y_j)$.

Är det givet att denna modell är sann? Nej, det är inte självklart utan hänger på vilka förutsättningar datan kommer från. Däremot tenderar modellen att fungera bra i de flesta fall om vi har en rimlig mängd observationer.
Med notationen från figuren ser vi att

$$r_j = y_j - \mu_j$$

om vi tar hänsyn till tecknet (positivt tecken om y_j ligger ovanför regressionslinjen). Ett mått på hur väl linjen approximerar mätserien ges av kvadratsumman

$$\sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (y_j - \mu_j)^2.$$

Vi skulle kunna minimera denna summa med avseende på (β_0, β_1), vilket ger MK-skattningar för β_0 och β_1. Detta var det som hände i exemplet från föreläsning 2. Istället för att upprepa argumentet går vi över till den generella modellen för linjär regression.

Linjär regression

Definition. Givet $(x_{j1}, x_{j2}, \ldots, x_{jk}, y_j)$, $j = 1, 2, \ldots, n$, för något positivt heltal k, där vi betraktar x_{ji} som fixerade och y_j som observationer av stokastiska variabler

$$Y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk} + \epsilon_j, \quad j = 1, 2, \ldots, n,$$

där $\epsilon_j \sim N(0, \sigma^2)$ antas oberoende (och likafördelade) och $\beta_0, \beta_1, \ldots, \beta_k$ är okända parameterar. Den räta linjen

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

kallas **regressionslinjen**.
Vi låter

$$\mu_j = E(Y_j) = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk}.$$

Dubbelindexeringen är lite jobbig att arbeta med, så låt oss gå över till matrisnotation:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 x_{11} & \cdots & \beta_k x_{1k} \\ \beta_0 & \beta_1 x_{21} & \cdots & \beta_k x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0 & \beta_1 x_{n1} & \cdots & \beta_k x_{nk} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$
$$= \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Vi låter

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \text{samt } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

vilket leder till sambandet

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Således gäller att

$$E(\boldsymbol{Y}) = X\boldsymbol{\beta} \quad \text{och} \quad C_{\boldsymbol{Y}} = \sigma^2 I_n,$$

där I_n är den n-dimensionella enhetsmatrisen. Vi söker MK-skattningen $\widehat{\boldsymbol{\beta}}$ för $\boldsymbol{\beta}$, vilket vi kan erhålla genom att minimera den kvadratiska formen

$$Q(\beta_0, \ldots, \beta_k) = \sum_{j=1}^{n} (y_j - \mu_j)^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}),$$

där $\boldsymbol{y} = (y_1\ y_2\ \cdots\ y_n)^T$. En variant är att sätta igång och derivera, men lite mer elegant gäller följande sats.

Normalekvationerna

Sats. Om det $X^T X \neq 0$ så ges MK-skattningen av $\boldsymbol{\beta}$ enligt

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Bevis. Vi kan skriva vektorn $\boldsymbol{Y}$ av uppmätta värden som

$$\boldsymbol{Y} = \beta_0 \boldsymbol{x_0} + \beta_1 \boldsymbol{x_1} + \cdots + \beta_k \boldsymbol{x_k} + \boldsymbol{\epsilon},$$

där $\boldsymbol{x_i}$, $i = 1, 2, \ldots, k$, är kolonn $i+1$ i matrisen X. Från linjär algebra vet vi att avståndet mellan observationen $\boldsymbol{y}$ och linjärkombinationer av vektorerna $\boldsymbol{x_j}$, dvs

$$\left|\boldsymbol{y} - (\widehat{\beta}_0 \boldsymbol{x_0} + \widehat{\beta}_1 \boldsymbol{x_1} + \cdots \widehat{\beta}_k \boldsymbol{x_k})\right|,$$

blir minimalt precis då

$$X\widehat{\boldsymbol{\beta}} = \sum_{j=1}^{k} \widehat{\beta}_j \boldsymbol{x_j}$$

är projektionen av $\boldsymbol{y}$ på det linjära höljet span$\{\boldsymbol{x_0},\ \boldsymbol{x_1},\ \cdots,\boldsymbol{x_k}\}$. Således måste $\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}$ vara vinkelrät mot $\boldsymbol{x_j}$ för alla $j = 0, 1, \ldots, k$. Detta medför att

$$X^T(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) = \boldsymbol{0} \quad \Leftrightarrow \quad X^T X\widehat{\boldsymbol{\beta}} = X^T \boldsymbol{y} \quad \Leftrightarrow \quad \widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y},$$

vilket är precis vad vi ville visa! □

När det gäller $\widehat{\boldsymbol{\beta}}$ och dess komponenter kommer vi använda samma beteckningar för observationer av punktskattningen och den stokastiska variabeln. Skulle vi vara konsekventa borde den stokastiska variabeln betecknas $\widehat{\boldsymbol{B}}$, men av tradition görs inte så. Var observant!

Sats. Med förutsättningarna ovan gäller att

$$\widehat{\boldsymbol{\beta}} \sim N_{k+1}\left(\boldsymbol{\beta},\ \sigma^2(X^TX)^{-1}\right).$$

Bevis. Det faktum att $\widehat{\boldsymbol{\beta}}$ är normalfördelad följer direkt från faktumet att elementen i $\widehat{\boldsymbol{\beta}}$ är linjärkombinationer av normalfördelade variabler. Återstår att visa väntevärde och kovarians:

$$E(\widehat{\boldsymbol{\beta}}) = (X^TX)^{-1}X^TE(\boldsymbol{Y}) = (X^TX)^{-1}X^TX\boldsymbol{\beta} = \boldsymbol{\beta}$$

och

$$\begin{aligned} C_{\widehat{\boldsymbol{\beta}}} &= (X^TX)^{-1}X^TC_{\boldsymbol{Y}}((X^TX)^{-1}X^T)^T = (X^TX)^{-1}X^T\sigma^2I_n((X^TX)^{-1}X^T)^T \\ &= (X^TX)^{-1}X^T\sigma^2I_n(X^T)^T((X^TX)^{-1})^T = \sigma^2(X^TX)^{-1}X^TX((X^TX)^T)^{-1} = \sigma^2(X^TX)^{-1}. \end{aligned}$$

Således blir fördelningen precis som beskriven i satsen. □

2 Variansanalys

Vi låter

$$\widehat{\mu}_j = \widehat{\beta}_0 + \widehat{\beta}_1x_{j1} + \widehat{\beta}_2x_{j2} + \cdots + \widehat{\beta}_kx_{jk}, \quad j = 1, 2, \ldots, n.$$

Ibland betecknas vektorn $\widehat{\boldsymbol{\mu}}$ som $\widehat{\boldsymbol{y}}$ eftersom det i någon mening är en skattningen av $\boldsymbol{y}$, men det blir lite olyckligt för det är inte $\boldsymbol{y}$ vi skattar. Vi ska nu studera hur bra den skattning vi tagit fram är.

Regressionsanalysens kvadratsummor

Definition. Vi definierar tre kvadratsummor som beskriver variationen hos y-värdena:

(i) Den **totala variationen** definieras enligt $\mathrm{SS_{TOT}} = \sum_{j=1}^{n}(y_j - \overline{y})^2$.

(ii) Variationen som förklaras av $x_1, x_2, \ldots, x_k$, definieras enligt $\mathrm{SS_R} = \sum_{j=1}^{n}(\widehat{\mu}_j - \overline{y})^2$.

(iii) Variationen som inte förklaras av regressionsmodellen är $\mathrm{SS_E} = \sum_{j=1}^{n}(y_j - \widehat{\mu}_j)^2$.

Ibland används beteckningarna Q_{TOT} för $\mathrm{SS_{TOT}}$, Q_{REGR} för $\mathrm{SS_R}$ och Q_{RES} för $\mathrm{SS_E}$.
Hur hänger dessa summor ihop? Som tur är finns ett enkelt svar.

Sats. Den totala variationen $\mathrm{SS_{TOT}}$ kan delas upp enligt

$$\mathrm{SS_{TOT}} = \mathrm{SS_R} + \mathrm{SS_E}.$$

Bevis. Vi vill uttrycka $\mathrm{SS}_{\mathrm{TOT}}$ i termer av SS_{R} och SS_{E}, så

$$\begin{aligned}\mathrm{SS}_{\mathrm{TOT}} &= \sum_{j=1}^{n}(y_j-\overline{y})^2 = \sum_{j=1}^{n}(y_j-\widehat{\mu}_j+\widehat{\mu}_j-\overline{y})^2 \\ &= \sum_{j=1}^{n}(y_j-\widehat{\mu}_j)^2 + 2\sum_{j=1}^{n}(y_j-\widehat{\mu}_j)(\widehat{\mu}_j-\overline{y}) + \sum_{j=1}^{n}(\widehat{\mu}_j-\overline{y})^2 \\ &= \mathrm{SS}_{\mathrm{R}} + 2\sum_{j=1}^{n}(y_j-\widehat{\mu}_j)\widehat{\mu}_j - 2\overline{y}\sum_{j=1}^{n}(y_j-\widehat{\mu}_j) + \mathrm{SS}_{\mathrm{E}}.\end{aligned}$$

Vi visar att de båda summorna i mitten summerar till noll. Eftersom $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$ gäller det att

$$\begin{aligned}\sum_{j=1}^{n}(y_j-\widehat{\mu}_j)\widehat{\mu}_j &= \widehat{\boldsymbol{\mu}}^T(\boldsymbol{y}-\widehat{\boldsymbol{\mu}}) = (X\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\beta}}^T X^T(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}) \\ &= \widehat{\boldsymbol{\beta}}^T(X^T\boldsymbol{y} - X^TX(X^TX)^{-1}X^T\boldsymbol{y}) = \widehat{\boldsymbol{\beta}}^T(X^T\boldsymbol{y}-X^T\boldsymbol{y}) = 0.\end{aligned}$$

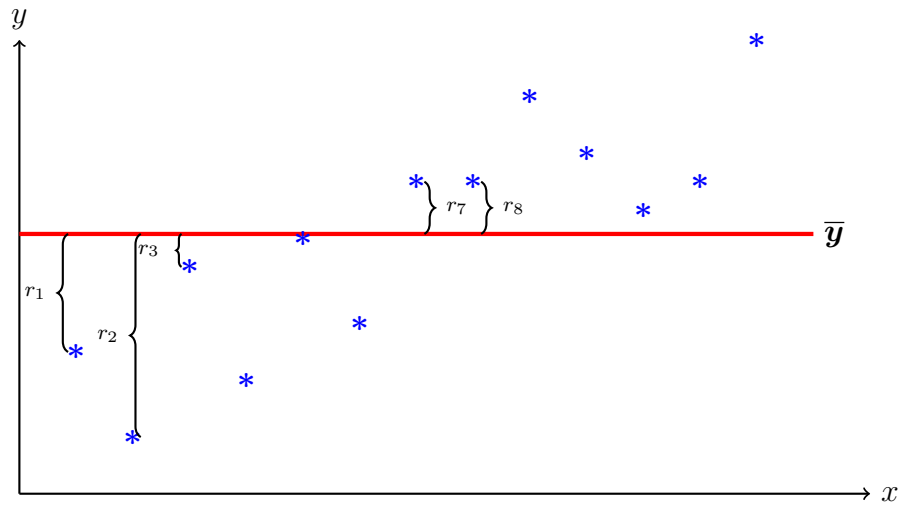
Med andra ord är $\boldsymbol{y}-\widehat{\boldsymbol{\mu}}$ vinkelrät mot $\widehat{\boldsymbol{\mu}}$.
Vidare vet vi att $\widehat{\boldsymbol{\beta}}$ minimerar $Q(\boldsymbol{\beta}) = \sum_{j=1}^{n}(y_j-\mu_j)^2$, så

$$0 = \frac{\partial}{\partial_{\beta_0}}Q(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = -2\sum_{j=1}^{n}(y_j-\widehat{\mu}_j) \quad \Leftrightarrow \quad \sum_{j=1}^{n}y_j = \sum_{j=1}^{n}\widehat{\mu}_j,$$

vilket var precis det vi behövde. □

2.1 $\mathrm{SS}_{\mathrm{TOT}}$

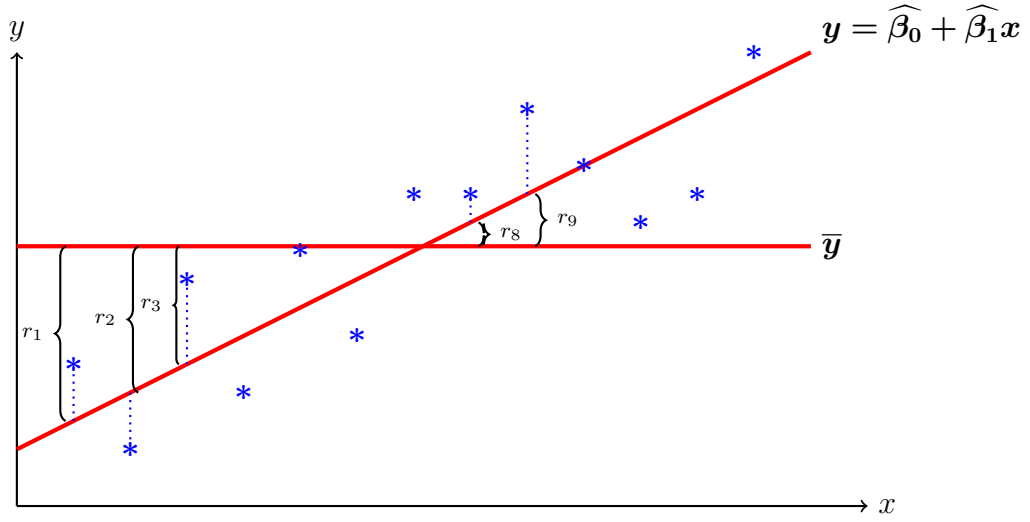
Storheten $\mathrm{SS}_{\mathrm{TOT}}$ mäter den totala variation av mätvärden jämfört med mätvärdenas medelvärde. I fallet då vi använder modellen $Y = \beta_0 + \epsilon$ ger således $\mathrm{SS}_{\mathrm{TOT}}$ hela felet i regressionen.



Vi ser att $\mathrm{SS}_{\mathrm{TOT}} = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n}(y_j-\overline{y})^2$.

2.2 $\mathrm{SS_R}$ och $\mathrm{SS_E}$

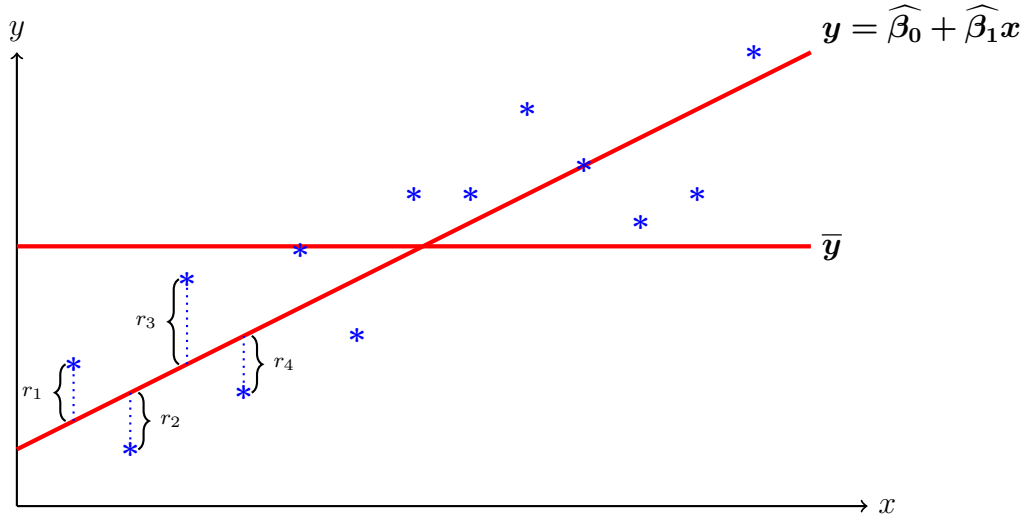
Om vi istället använder modellen $y = \beta_0 + \beta_1 x + \epsilon$ blir summorna $\mathrm{SS_E}$ och $\mathrm{SS_R}$ relevanta ($\mathrm{SS_{TOT}}$ ser fortfarande ut som ovan). Låt oss först illustrera $\mathrm{SS_R}$.



Vi ser att

$$\mathrm{SS_R} = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (\widehat{\beta}_0 + \widehat{\beta}_1 x_j - \overline{y})^2$$

och $\mathrm{SS_R}$ mäter alltså hur mycket den skattade regressionslinjen (i mätpunkterna x_j) skiljer sig från medelvärdet av y-värdena.
När det gäller $\mathrm{SS_E}$ är det istället skillnaden mellan den skattade regressionslinjen och mätvärdena vi betraktar.



Kvadratsumman av dessa avvikelser blir

$$\mathrm{SS_E} = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (y_j - (\widehat{\beta}_0 + \widehat{\beta}_1 x_j))^2.$$

3 Projektionsmatriser

Eftersom vi arbetar med matriser och MK-lösningen i princip är projektionen på ett underrum av $\mathbf{R}^n$ så är följande resultat inte allt för förvånande.

Hatt-matrisen

Definition. Vi definierar $H = X(X^TX)^{-1}X^T$. Vi definierar även J som en matris vars samtliga element är 1. Vi skriver J_{nm} om vi vill markera dimensionen.

Varför kallar vi H för hatt-matrisen? Ganska enkelt:

$$H\boldsymbol{y} = X(X^TX)^{-1}X^T\boldsymbol{y} = X\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{y}}.$$

Avbildningen sätter alltså hatten på!

Sats. Matriserna H, $I-H$ och $H-\frac{1}{n}J$ är projektionsmatriser P som uppfyller $P^2 = P^T = P$.

Bevis. Direkt från definitionen av H blir

$$H^2 = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) = X(X^TX)^{-1}X^T = H$$

och

$$H^T = (X(X^TX)^{-1}X^T)^T = X(X^TX)^{-T}X^T = X((X^TX)^T)^{-1}X^T = H.$$

Därav följer det att $(I-H)^2 = I^2 - 2H + H^2 = I - H$ och att $(I-H)^T = I^T - H^T = I - H$. För den sista operatorn skriver vi

$$\left(H - \frac{1}{n}J\right)^2 = H^2 - \frac{1}{n}HJ - \frac{1}{n}JH + \frac{1}{n^2}J^2 = H - \frac{1}{n}HJ - \frac{1}{n}JH + \frac{1}{n}J$$

ty J^2 är matrisen med samtliga element lika med n. Vidare gäller att J kommuterar med alla kvadratiska matriser, så $JH = HJ$. Här kan vi se att $H\mathbf{1} = \mathbf{1}$ (där $\mathbf{1}$ är en vektor med samtliga element lika med 1) eftersom lösningen till regressionsproblemet om $\boldsymbol{y}$ är konstant är just den konstanten (lösningen är exakt). Alltså blir

$$\left(H - \frac{1}{n}J\right)^2 = H - \frac{1}{n}J.$$

Givetvis gäller även att $\left(H - \frac{1}{n}J\right)^T = H - \frac{1}{n}J.$ □

En följdsats av detta är att vi kan skriva kvadratsummorna som kvadratiska former.

Sats. $\mathrm{SS_{TOT}} = \boldsymbol{y}^T\left(I - \frac{1}{n}J\right)\boldsymbol{y}$, $\mathrm{SS_R} = \boldsymbol{y}^T\left(H - \frac{1}{n}J\right)\boldsymbol{y}$, samt $\mathrm{SS_E} = \boldsymbol{y}^T(I-H)\,\boldsymbol{y}$.

4 Förklaringsgrad

När vi utför regressionsanalys vill vi ofta ha ett mått som preciserar hur bra modellen passar de uppmätta värdena. Ett sådant mått är förklaringsgraden.

Definition. Förklaringsgraden R^2 definieras som

$$R^2 = \frac{\mathrm{SS_R}}{\mathrm{SS_{TOT}}} = \frac{\mathrm{SS_{TOT}} - \mathrm{SS_E}}{\mathrm{SS_{TOT}}} = 1 - \frac{\mathrm{SS_E}}{\mathrm{SS_{TOT}}}.$$

Ibland använder man lite andra varianter av R^2 och en mycket vanlig är den så kallade **justerade förkaringsgraden**

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SS_E}/(n-k-1)}{\mathrm{SS_{TOT}}/(n-1)}.$$

Om inte n är förhållandevis stor i jämförelse med k (vilket är nödvändigt för att regressionen ska vara vettig) så blir den justerade graden sämre.

5 Regressionsanalysens huvudsats

Innan vi ger oss in på huvudsatsen visar vi en hjälpsats från linjär algebra.

Sats. Låt $\boldsymbol{x} = (x_1\ x_2\ \cdots\ x_n)$ vara sådan att

$$\boldsymbol{x}^T\boldsymbol{x} = \sum_{j=1}^{n} x_j^2 = \boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{x}^T B\boldsymbol{x}$$

för $A, B \in \mathbf{R}^{n\times n}$ positivt semi-definita och symmetriska med $\mathrm{rank}(A) = r$ och $\mathrm{rank}(B) = n-r$ för något heltal r så att $0 < r < n$. Då finns en ON-matris C så att med $\boldsymbol{x} = C\boldsymbol{y}$ gäller att

$$\boldsymbol{x}^T A\boldsymbol{x} = \sum_{j=1}^{r} y_j^2 \quad \text{och} \quad \boldsymbol{x}^T B\boldsymbol{x} = \sum_{j=r+1}^{n} y_j^2.$$

Bevis. Eftersom A är positivt semi-definit har A endast icke-negativa egenvärden som vi ordnar enligt $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Vi vet även att $\mathrm{rank}(A) = r$, så $\lambda_{r+1} = \cdots = \lambda_n = 0$. Vidare är A diagonaliserbar eftersom A är symmetrisk, så det finns en ON-matris C så att

$$C^T AC = D = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \vdots & \ddots & \vdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_r & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Med $\boldsymbol{x} = C\boldsymbol{y}$ gäller då att

$$\boldsymbol{x}^T\boldsymbol{x} = (C\boldsymbol{y})^T(C\boldsymbol{y}) = \boldsymbol{y}^T C^T C\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{y}$$

och

$$\boldsymbol{x}^T A\boldsymbol{x} = (C\boldsymbol{y})^T AC\boldsymbol{y} = \boldsymbol{y}^T C^T AC\boldsymbol{y} = \boldsymbol{y}^T D\boldsymbol{y} = \sum_{j=1}^{r} \lambda_j y_j^2.$$

Vi har även

$$\boldsymbol{x}^T B\boldsymbol{x} = (C\boldsymbol{y})^T BC\boldsymbol{y} = \boldsymbol{y}^T C^T BC\boldsymbol{y}$$

och då blir

$$\sum_{j=1}^{n} y_j^2 = \boldsymbol{y}^T\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{x}^T B\boldsymbol{x} = \sum_{j=1}^{r} \lambda_j y_j^2 + \boldsymbol{y}^T C^T BC\boldsymbol{y}$$

så

$$\boldsymbol{y}^T C^T BC\boldsymbol{y} = \sum_{j=1}^{r} (1-\lambda_j) y_j^2 + \sum_{j=r+1}^{n} y_j^2.$$

Det faktum att $\operatorname{rank}(B) = n - r$ visar att $\lambda_1 = \lambda_2 = \cdots \lambda_r = 1$ och vi erhåller identiteten

$$\boldsymbol{y}^T C^T BC\boldsymbol{y} = \sum_{j=r}^{n} y_j^2.$$

Alla beteckningar är nu ur vägen och vi är framme vid huvudresultatet.

Regressionsanalysens huvudsats

Sats. Med förutsättningarna ovan gäller följande för $\mathrm{SS_E}$ och $\mathrm{SS_R}$ sedda som stokastiska variabler.

(i) $\dfrac{\mathrm{SS_E}}{\sigma^2} = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{n} (Y_j - \widehat{\mu}_j)^2 \sim \chi^2(n-k-1).$

(ii) Givet att $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ är $\dfrac{\mathrm{SS_R}}{\sigma^2} = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{n} (\widehat{\mu}_j - \overline{Y})^2 \sim \chi^2(k).$

(iii) Både $\mathrm{SS_R}$ och $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$ är oberoende av $\mathrm{SS_E}$.

Bevis. Eftersom H är en projektionsmatris har H endast egenvärdena $\lambda = 0$ och $\lambda = 1$. Således gäller det finurliga att matrisens rang är lika med summan av egenvärdena, vilka kan beräknas genom att ta spåret[1] av matrisen:

$$\operatorname{rank}(H) = \operatorname{tr}(H) = \operatorname{tr}(X(X^T X)^{-1} X^T) = \operatorname{tr}(X^T X(X^T X)^{-1}) = \operatorname{tr}(I_{k+1}) = k+1,$$

Det följer sedan att

$$\operatorname{rank}(I - H) = n - k - 1.$$

[1] se sista (bonus)avsnittet för lite detaljer kring spår av matriser.

Eftersom $HX = X$ så gäller att

$$\boldsymbol{Y}^T(I-H)\boldsymbol{Y} = \boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon}.$$

Detta är användbart eftersom $E(\boldsymbol{\epsilon}) = \mathbf{0}$. Enligt föregående hjälpsats gäller att sambandet

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T H\boldsymbol{\epsilon}$$

medför att det finns en ON-matris C så att $\boldsymbol{Z} = C\boldsymbol{\epsilon}$ reducerar likheten till

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sum_{j=1}^{n-k-1} Z_j^2 + \sum_{j=n-k}^{n} Z_j^2$$

där

$$\boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon} = \boldsymbol{Y}^T(I-H)\boldsymbol{Y} = \mathrm{SS_E} = \sum_{j=1}^{n-k-1} Z_j^2.$$

Eftersom komponenterna i $\boldsymbol{\epsilon}$ är oberoende och $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$ följer det att

$$E(\boldsymbol{Z}) = C\mathbf{0} = \mathbf{0} \quad \text{och} \quad C_{\boldsymbol{Z}} = C_{C\boldsymbol{\epsilon}} = \sigma^2 C^T I C = \sigma^2 I$$

så $\boldsymbol{Z} \sim N\left(\mathbf{0}, \sigma^2 I\right)$. Komponenterna i $\boldsymbol{Z}$ är alltså oberoende och $Z_j \sim N(0, \sigma^2)$. Detta medför att

$$\frac{\mathrm{SS_E}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{n-k-1} Z_j^2 \sim \chi^2(n-k-1).$$

Vi erhåller även att $\mathrm{SS_E}$ och $H\boldsymbol{\epsilon}$ är oberoende (de delar inga variabler Z_j). Detta medför att $\mathrm{SS_E}$ och $\widehat{\boldsymbol{\beta}}$ är oberoende. Det faktum att $\mathrm{SS_E}$ och $\mathrm{SS_R}$ är oberoende följer av att kovariansen

$$\begin{aligned} C\left(\left(H-\frac{1}{n}J\right)\boldsymbol{Y}, (I-H)\boldsymbol{Y}\right) &= \left(H-\frac{1}{n}J\right)C(\boldsymbol{Y},\boldsymbol{Y})(I-H)^T = \sigma^2\left(H-\frac{1}{n}J\right)(I-H) \\ &= \sigma^2\left(H - H^2 - \frac{1}{n}J + \frac{1}{n}JH\right) = \sigma^2\left(-\frac{1}{n}J + \frac{1}{n}HJ\right) \\ &= \sigma^2\left(-\frac{1}{n}J + \frac{1}{n}J\right) = 0, \end{aligned}$$

så dessa vektorer är okorrelerade och normalfördelade, så oberoende. Det följer direkt att $\mathrm{SS_E}$ och $\mathrm{SS_R}$ är oberoende (som funktioner av oberoende variabler).
Om vi fokuserar på fördelningen för $\mathrm{SS_R}$ så kan vi på samma sätt som ovan utnyttja att $\left(H - \frac{1}{n}J\right)$ är en projektionsmatris, så

$$\operatorname{rank}\left(H-\frac{1}{n}J\right) = \operatorname{tr}\left(H-\frac{1}{n}J\right) = \operatorname{tr}(H) - \operatorname{tr}\left(\frac{1}{n}J\right) = k+1-1 = k.$$

Matrisen J är synnerligen rangdefekt med $\operatorname{rank}(J) = 1$ (eftersom alla kolonner är lika spänner vi bara upp ett en-dimensionellt rum).
Nu stöter vi på lite problem. Vi ser att

$$\left(H-\frac{1}{n}J\right)\boldsymbol{Y} = \left(I-\frac{1}{n}J\right)X\boldsymbol{\beta} + \left(H-\frac{1}{n}J\right)\boldsymbol{\epsilon}$$

där den första termen *inte* försvinner såvida inte $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. Men under detta antagande har vi åter igen en situation där vi kan "byta ut" $\boldsymbol{Y}$ mot $\boldsymbol{\epsilon}$. Föregående hjälpsats visar – analogt med föregående argument – att

$$\boldsymbol{Y}^T \left(H - \frac{1}{n} J \right) \boldsymbol{Y} = \sum_{j=1}^{k} Z_j^2,$$

där Z_j är oberoende och $Z_j \sim N(0, \sigma^2)$, vilket medför att $\frac{1}{\sigma^2} \mathrm{SS_R} \sim \chi^2(k)$, under förutsättningen att $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. □

Kommentar: utan villkoret $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ kan man fortfarande genomföra argumentet, men resultatet blir en icke-centrerad $\chi^2(k)$-fördelning (något som inte ingår i kursen).

6 Hypotestester och kofidensintervall

Vi har nu samlat på oss en ordentlig verktygslåda, så det kanske är dags att se hur vi använder de olika delarna.

6.1 Skattning av $\boldsymbol{\sigma^2}$

Variansen σ^2 skattar vi med

$$s^2 = \frac{\mathrm{SS_E}}{n - k - 1}.$$

Denna skattning är väntevärdesriktig:

$$E(S^2) = \frac{\sigma^2}{n-k-1} E\left(\frac{\mathrm{SS_E}}{\sigma^2} \right) = \sigma^2 \frac{n-k-1}{n-k-1} = 1$$

ty $\frac{1}{\sigma^2} \mathrm{SS_E} \sim \chi^2(n-k-1)$.

6.2 Finns det något vettigt i modellen?

En rimlig fråga efter utförd regression är om modellen faktiskt säger något. Vi brukar testa denna hypotes enligt följande. Vi testar nollhypotesen

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

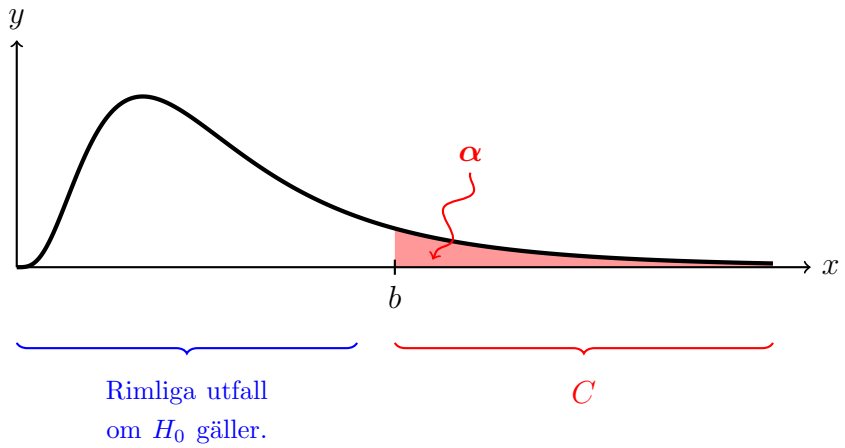
mot

$$H_1 : \text{minst något } \beta_i \neq 0.$$

En naturlig testvariabel ges av

$$v = \frac{\mathrm{SS_R}/k}{\mathrm{SS_E}/(n-k-1)}.$$

Huvudsatsen medför att $V \sim F(k, n-k-1)$ (om H_0 är sann) och vi förkastar H_0 om v är stor. Hur stor avgörs av signifikansnivån och lite slagning i F-tabeller.



Vi hittar gränsen b ur tabell (eller med `finv(1-alpha, k, n-k-1)` i MATLAB) så att

$$P(V > b) = \alpha.$$

6.3 Enskilda koefficienter

Eftersom $\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1}\right)$ introducerar vi beteckningen

$$(X^TX)^{-1} = \begin{pmatrix} h_{00} & h_{01} & \cdots & h_{0k} \\ h_{10} & h_{11} & \cdots & h_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ h_{k0} & h_{k1} & \cdots & h_{kk} \end{pmatrix}.$$

Då är $\widehat{\beta}_i \sim N(\beta_i, \sigma^2 h_{ii})$. Om vi känner σ^2 kan vi använda att

$$Z = \frac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{h_{ii}}} \sim N(0,1)$$

för att testa hypotesen $H_0 : \beta_i = 0$ eller för att ställa upp konfidensintervall I_{β_i}.
Nu är det extremt sällan vi känner σ^2 exakt, men vi vet att $\widehat{\boldsymbol{\beta}}$ är oberoende av $\mathrm{SS_E}$ enligt huvudsatsen, så $\widehat{\beta}_i$ är oberoende av $\mathrm{SS_E}$. Vidare är $s^2 = \mathrm{SS_E}/(n-k-1)$ en skattning av σ^2 vi känner väl, så

$$\frac{(\widehat{\beta}_i - \beta_i)/(\sigma\sqrt{h_{ii}})}{\sqrt{((n-k-1)S^2/\sigma^2)/(n-k-1)}} = \frac{\widehat{\beta}_i - \beta_i}{S\sqrt{h_{ii}}} \sim t(n-k-1)$$

enligt Gossets sats.

6.4 Hypotestest: $H_0 : \beta_i = 0$

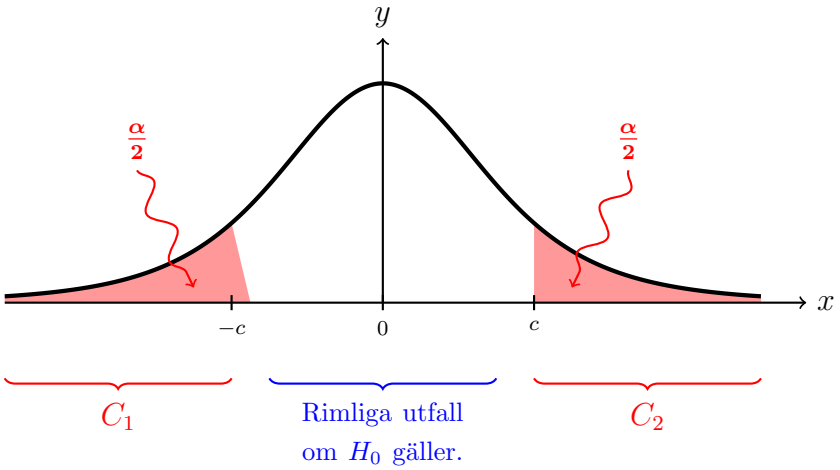
Vi kan även utföra hypotestester för en ensklid koefficient β_i för att se om den förklaringsvariabeln tillför något signifikant (givetvis går det att ställa upp konfidensintervall också för mer kvantitativt innehåll).
Låt $H_0 : \beta_i = 0$ och $H_1 : \beta_i \neq 0$. Vi vet enligt ovan att

$$T = \frac{\widehat{\beta}_i - \beta_i}{S\sqrt{h_{ii}}} \sim t(n-k-1),$$

så det kritiska området finner vi enligt

$$C = \{t \in \mathbf{R} : |t| > c\}$$

för lämpligt tal $c > 0$ beroende på signifikansnivån och antalet frihetsgrader.



Gränsen hittar vi i tabell genom att leta reda på ett tal $c = F_T^{-1}(1 - \alpha/2)$ (sitter du med MATLAB kan du använda `c = -tinv(alpha/2)`).

7 Exempel: enkel linjär regression

Vi diskuterade enkel linjär regression tidigare i samband med MK-skattningar (föreläsning 2). Låt oss visa att vi får samma resultat med den metod vi nu tagit fram (och fördelningar för ingående storheter på ett enkelt sätt).
Vi låter $x_1, x_2, \ldots, x_n$ vara fixerade tal och $y_1, y_2, \ldots, y_n$ vara observationer från

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

där $\epsilon_i \sim N(0, \sigma^2)$ är oberoende. Alltså precis den modell vi använt tidigare. Syntesmatrisen X ges av

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

så

$$X^T X = X = \begin{pmatrix} n & n\overline{x} \\ n\overline{x} & \sum_{i=1}^n x_i^2 \end{pmatrix} \quad \Rightarrow \quad (X^T X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (n\overline{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix}$$

om $\det(X^T X) \neq 0$. Således blir

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} &= (X^T X)^{-1} X^T \boldsymbol{y} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (n\overline{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix} \begin{pmatrix} n\overline{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (n\overline{x})^2} \begin{pmatrix} n\overline{y}\sum_{i=1}^n x_i^2 - n\overline{x}\sum_{i=1}^n x_i y_i \\ -n^2\overline{x}\overline{y} + n\sum_{i=1}^n x_i y_i \end{pmatrix}, \end{aligned}$$

vilket ger att

$$\widehat{\beta_1} = \frac{\sum_{j=1}^n x_j y_j - n\overline{x}\,\overline{y}}{\sum_{j=1}^n x_j^2 - n\overline{x}^2} = \frac{\sum_{j=1}^n y_j(x_j - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2} = \frac{\sum_{j=1}^n (y_j - \overline{y})(x_j - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2} \quad \text{och} \quad \widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}.$$

Det följer nu att

$$\widehat{\beta_0} \sim N\left(\beta_0,\, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2}\right) \quad \text{och} \quad \widehat{\beta_1} \sim N\left(\beta_1,\, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right).$$

Någonstans nu inser vi vilket kraftig syntaktiskt verktyg matriser och linjär algebra är..
Eftersom σ är okänd skattar vi σ^2 med

$$s^2 = \frac{\mathrm{SS_E}}{n-2} = \frac{\sum_{j=1}^n (y_j - \widehat{\beta_0} - \widehat{\beta_1} x_j)^2}{n-2}$$

där $S^2 \sim \chi^2(n-2)$. Vi kan skriva om $\mathrm{SS_E}$ genom att utnyttja att

$$\begin{aligned}\mathrm{SS_R} &= \sum_{j=1}^n (\widehat{\mu_j} - \overline{y})^2 = \widehat{\beta_1}^2 \sum_{j=1}^n (x_j - \overline{x})^2 = \left(\frac{\sum_{j=1}^n (y_j - \overline{y})(x_j - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2}\right)^2 \sum_{j=1}^n (x_j - \overline{x})^2 \\ &= r^2 \sum_{j=1}^n (y_i - \overline{y})^2\end{aligned}$$

så

$$\mathrm{SS_E} = \mathrm{SS_{TOT}} - \mathrm{SS_R} = (1 - r^2)\sum_{j=1}^n (y_i - \overline{y})^2.$$

Vi kan här se att $r = 1$ ger perfekt matching så alla (x_j, y_j) ligger på en rät linje.

8 Bonus: determinanter och spår

För en kvadratisk matris A med dimension $n \times n$ definierar vi som bekant det **karakteristiska polynomet** enligt

$$p(\lambda) = \det(A - \lambda I).$$

Bekant från linjär algebra är att $p(\lambda)$ kan representeras enligt

$$\begin{aligned} p(\lambda) &= (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n) \\ &= (-1)^n\big(\lambda^n - \lambda^{n-1}(\lambda_1 + \lambda_2 + \cdots + \lambda_n) + \cdots + (-1)^n \lambda_1\lambda_2\cdots\lambda_n\big), \end{aligned} \tag{1}$$

där λ_i, $i = 1, 2, \ldots, n$, är matrisens egenvärden.
Spåret för A betecknar $\operatorname{tr} A$ och definieras som summan av diagonalelementen:

$$\operatorname{tr} A = \sum_{i=1}^n a_{ii}.$$

Spåret är linjär och uppfyller alltså att

$$\operatorname{tr}(cA) = c \operatorname{tr} A \quad \text{och} \quad \operatorname{tr}(A + B) = \operatorname{tr} A + \operatorname{tr} B.$$

Dessutom gäller att

$$\operatorname{tr}(AB) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ji} = \sum_{j=1}^{m}\sum_{i=1}^{n} b_{ji}a_{ij} = \operatorname{tr}(BA)$$

om A är $n \times m$ och B är $m \times n$. Denna likhet leder direkt till att

$$\operatorname{tr}(PAP^{-1}) = \operatorname{tr}(P^{-1}(PA)) = \operatorname{tr} A,$$

vilket innebär att spåret (likt determinanten) är en invariant.
Vi vet även att $\det A = \lambda_1 \lambda_2 \cdots \lambda_n$, dvs determinanten ges av produkten av alla egenvärden. Finns något liknande för spåret? Svaret är vad man kanske skulle kunna gissa, nämligen att

$$\operatorname{tr} A = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

Enklaste beviset är att uttnyttja att alla matriser A kan skrivas på Jordans normalform, dvs att det finns en basbytesmatris så att $J = PAP^{-1}$, där J har $\lambda_1, \lambda_2, \ldots, \lambda_n$ på diagonalen och är nästan en diagonalmatris. Uttnyttjar vi att spåret är ivariant är vi klara. Kanske ligger detta argument lite utanför grundkursen i linjär algebra, så låt oss studera det karakteristiska polynomet lite närmare som alternativ. Genom att utveckla efter exempelvis rad 1 kan determinanten skrivas

$$p(\lambda) = (a_{11} - \lambda)\begin{vmatrix} a_{22} - \lambda & a_{23} & \cdots & a_{2n} \\ a_{32} & a_{33} - \lambda & \cdots & a_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n2} & a_{n3} & \cdots & a_{nn} - \lambda \end{vmatrix} + p_1(t),$$

där $\operatorname{grad} p_1(\lambda) \leq n-2$ eftersom vi tagit bort en λ-term. Om vi utför samma operation på denna determinant ser vi att

$$p(\lambda) = (a_{11} - \lambda)(a_{22} - \lambda)\begin{vmatrix} a_{33} - \lambda & a_{34} & \cdots & a_{3n} \\ a_{43} & a_{44} - \lambda & \cdots & a_{4n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n3} & a_{n4} & \cdots & a_{nn} - \lambda \end{vmatrix} + p_2(t),$$

där $\operatorname{grad} p_2(\lambda) \leq n - 2$ med något (nytt) polynom $p_2(\lambda)$.
Upprepa proceduren n gånger och vi erhållet att

$$\begin{aligned} p(\lambda) &= (a_{11} - \lambda)(a_{22} - \lambda)\cdots(a_{nn} - \lambda) + p_n(\lambda) \\ &= (-1)^n\lambda^n + (-1)^{n-1}\lambda^{n-1}(a_{11} + a_{22} + \cdots + a_{nn}) + p_{n+1}(\lambda) \\ &= (-1)^n\big(\lambda^n - \lambda^{n-1}\operatorname{tr} A\big) + p_{n+1}(\lambda), \end{aligned} \tag{2}$$

där $p_{n+1}(\lambda)$ är något polynom med grad högst $n-2$. Om vi jämför (1) med (2) ser vi att

$$\operatorname{tr} A = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

Föreläsning 9: Linjär regression – del II

Johan Thim (johan.thim@liu.se)

29 september 2018

"No tears, please. It's a waste of good suffering."
–Pinhead

Vi fixerar en vektor $\boldsymbol{u}^T = (1 \ u_1 \ u_2 \ \cdots \ u_k)$, där u_i kommer vara värdet på x_j i den punkt vi kommer betrakta. Vi är alltså intresserade av vad modellen har att säga vid en fixerad punkt där vi inte gjort någon mätning. Vi betraktar Y_0 definierad av

$$Y_0 = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \cdots + \beta_k u_k + \epsilon_0 = \boldsymbol{u}^T\boldsymbol{\beta} + \epsilon_0.$$

Vi antar att $\epsilon_0 \sim N(0,\sigma^2)$ är oberoende av $\boldsymbol{\epsilon}$. Vi definierar

$$\mu_0 = E(Y_0) = \boldsymbol{u}^T\boldsymbol{\beta}.$$

2 Konfidensintervall för $E(Y_0)$

En naturlig skattning av μ_0 ges av $\boldsymbol{u}^T\widehat{\boldsymbol{\beta}}$, så vi sätter

$$\widehat{\mu_0} = \boldsymbol{u}^T\widehat{\boldsymbol{\beta}}.$$

Eftersom $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$ blir

$$E(\widehat{\mu_0}) = \boldsymbol{u}^T E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{u}^T\boldsymbol{\beta}$$

och

$$V(\widehat{\mu_0}) = \sigma^2\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}.$$

Då $\widehat{\mu_0}$ är en linjärkombination av normalfördelade variabler gäller att

$$\widehat{\mu_0} \sim N(\boldsymbol{u}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}).$$

Således gäller att

$$\frac{\widehat{\mu_0} - \boldsymbol{u}^T\boldsymbol{\beta}}{\sigma\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} \sim N(0,1).$$

I vanlig ordningen brukar vi behöva skatta σ^2 och gör det med

$$s^2 = \frac{\mathrm{SS_E}}{n-k-1}, \quad \text{där } S^2 \sim \chi^2(n-k-1).$$

Då gäller (enligt Gossets sats) att

$$\frac{\widehat{\mu_0} - \boldsymbol{u}^T\boldsymbol{\beta}}{S\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} \sim t(n-k-1).$$

Genom att nyttja denna variabel kan vi ställa upp ett tvåsidigt konfidensintervall för $E(Y_0)$:

$$I_{\mu_0} = \left(\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} - t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}},\ \boldsymbol{u}^T\widehat{\boldsymbol{\beta}} + t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}\right).$$

Intervallet I_{μ_0} beskriver vart uppmätta värden vid $\boldsymbol{u}$ hamnar i snitt, dvs vid många upprepningar med samma $\boldsymbol{u}$ så hamnar vi i intervallet. Det säger inget om vart en enskild mätning hamnar, för det behöver vi prediktionsintervall!

3 Prediktionsintervall för Y_0

Vill vi uppskatta (förutsäga) vad mätvärdet y_0 blir i en viss punkt $\boldsymbol{u}$ ställer vi upp ett **prediktionsintervall**. Eftersom $Y_0 \sim N(\mu_0, \sigma^2)$ och $\widehat{\mu_0} = \boldsymbol{u}^T\widehat{\boldsymbol{\beta}} \sim N(\mu_0, \sigma^2\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u})$ är oberoende gäller det att

$$V(Y_0 - \widehat{\mu_0}) = \sigma^2\left(1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}\right)$$

så

$$Y_0 - \widehat{\mu_0} \sim N\left(0,\ \sigma^2\left(1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}\right)\right).$$

Vi skattar σ^2 med s^2 och nyttjar Gossets sats:

$$\frac{Y_0 - \widehat{\mu_0}}{S\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} \sim t(n-k-1).$$

Vi kan stänga in denna variabel och lösa ut Y_0:

$$I_{Y_0} = \Bigg(\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} - t_{\alpha/2}(n-k-1)s\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}},$$
$$\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} + t_{\alpha/2}(n-k-1)s\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}\Bigg).$$

4 Konfidens- och prediktionsband

Vid grafisk representation av enkel linjär regression ser man ofta så kallade konfidens- och prediktionsband inritade. Dessa definieras enligt följande.

Konfidensband

Definition. Ett **konfidensband** ges av en funktion g sådan att för varje x gäller att

$$P(|\mu_0(x) - \widehat{\mu_0}(x)| < g(x)) = 1 - \alpha.$$

Ett **simultant konfidensband** uppfyller att

$$P(|\mu_0(x) - \widehat{\mu_0}(x)| < g(x) \text{ för alla } x) = 1 - \alpha.$$

Skillnaden mellan ett simultant band och dess icke-simultana motsvarighet kanske är svår att se, men det simultana bandet uppfyller alltså instängningen med sannolikheten $1-\alpha$ för *alla* x på en gång medan den icke-simultana uppfyller denna sannolikhet för varje x *en i taget*! Likformighet är något det simultana bandet erbjuder. Om vi endast har ett icke-simultant konfidensband och vill titta i två punkter x_1 och x_2 samtidigt är det inte säkert att dessa intervall *samtidigt* uppfyller konfidensgraden $1 - \alpha$. Det är precis samma problem vi sett vi beräkningar av flera konfidensintervall samtidigt tidigare.

Prediktionsband

Definition. Ett **prediktionsband** ges av en funktion h sådan att för varje x gäller att

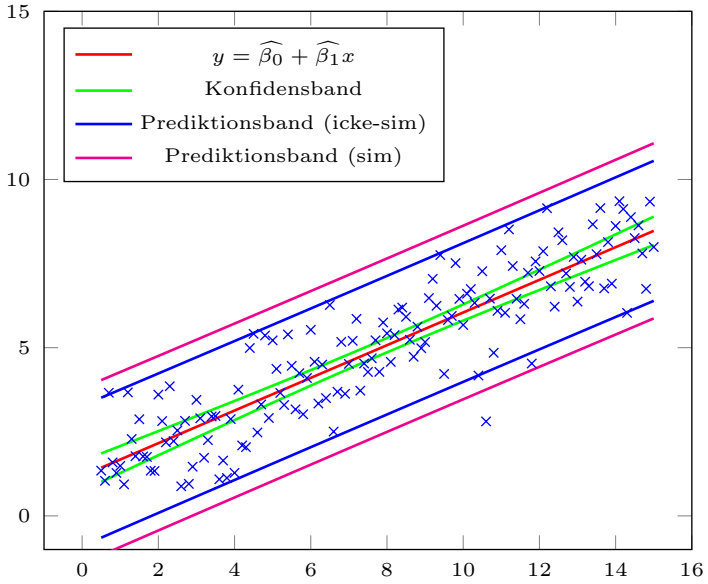
$$P(|y(x) - \widehat{y}(x)| < h(x)) = 1 - \alpha.$$

Ett **simultant prediktionsband** uppfyller att

$$P(|y(x) - \widehat{y}(x)| < h(x) \text{ för alla } x) = 1 - \alpha.$$

Grafiskt kan det se ut enligt nedan. Man ritar ofta i både konfidens- och prediktionsband samtidigt. Notera att konfidensbandet är betydligt smalare än prediktionsbandet.

Konfidensband; $\alpha = 0.05$



5 Residualanalys

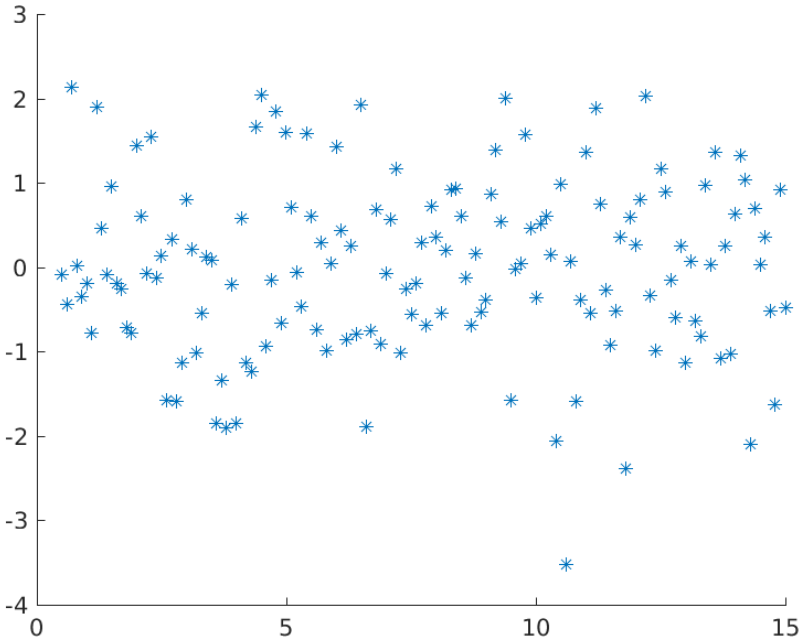
Efter utförd regression har vi skattade y-värden $\widehat{\mu}$ (eller $\widehat{y}$), som används för att beräkna kvadratsumman $\mathrm{SS_E}$ för felen som modellen inte förklarar. Antagandet vi gjort på **residualerna** $e_j = y_j - \widehat{y_j}$ är att dessa är oberoende och normalfördelade med samma varians och väntevärde 0. Detta är något som bör undersökas efter regressionen för att motivera antagandet. I matlab kan vi ta fram residualerna vid regression genom kommandot

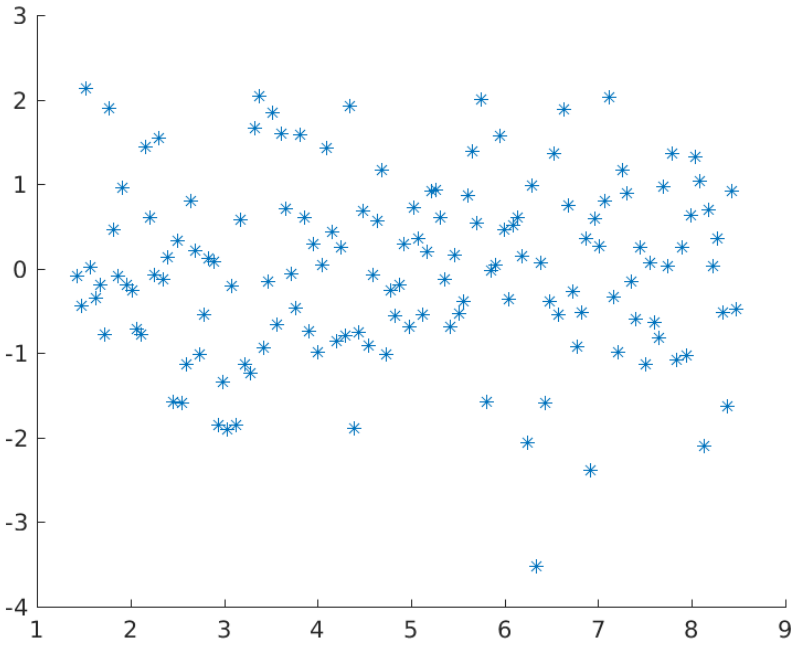
```
>> r = regstats(y, x, 'linear', 'all');
>> res = r.r;
>> yhat = r.yhat;
```

5.1 Residualer vs x eller $\widehat{y}$

Vi kan plotta residualer mot x-värden eller skattade y-värden ($\widehat{y} = \widehat{\mu}$):

```
>> figure; scatter(x, res, '*');
>> figure; scatter(yhat, res, '*');
```



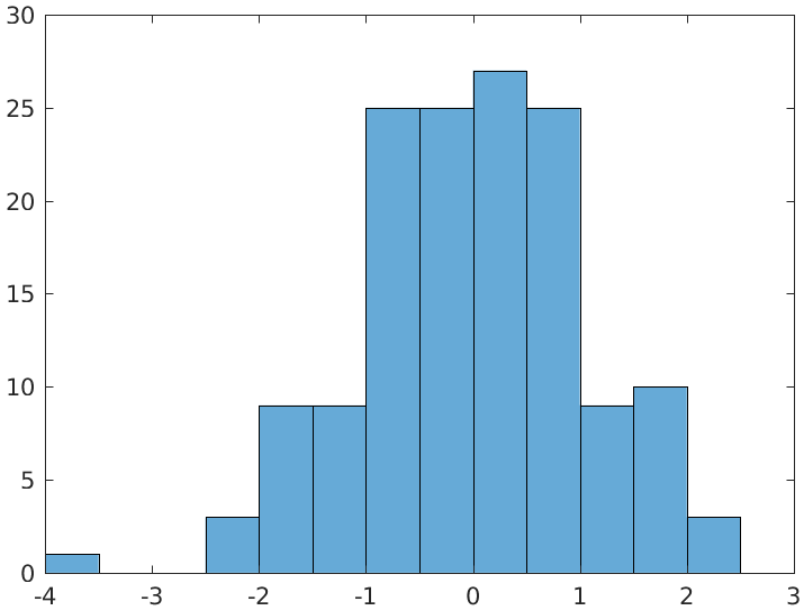


Det är svårt att se något direkt samband. Vilket är bra. Hade vi sett ett tydligt samband hade vi haft problem med modellen. Men mycket mer än så kan vi inte säga från dessa figurer.

5.2 Histogram

Vi kan plotta ett histogram för residualerna:

```
>> figure; histogram(res);
```

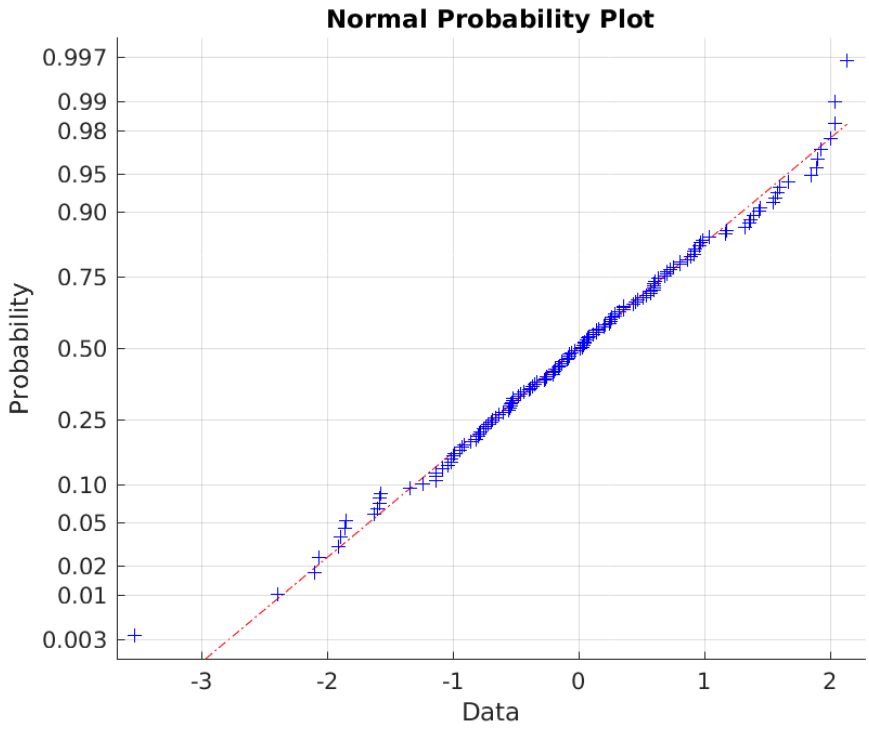


Det ser hyfsat Gaussiskt ut och masscentrum är runt nollan. Inte helt orimligt med normalfördelning.

5.3 Normalplot

Matlab kan även enkelt generera en så kallad normalplot:

```
>> figure; normplot(res);
```



I figuren så skalar alltså y-axeln mot sannolikheter som gäller för normalfördelning (tänk på exempelvis log-skala fungerar). Idealiskt skulle vi endast ha punkter som ligger *på* en linje. Nu finns kanske lite tillstymmelse till så kallad **S-form** på kurvan, men absolut inte på den nivå att vi borde ifrågasätta antagandet kring normalfördelning. Betydligt mer S-lika kurvor skulle accepteras som rimligt normalfördelade.

6 Variabeltransformation

Det vi håller på med kallas linjär regression, men det är inget som hindrar oss att ändå använda linjär struktur för att anpassa ett polynom eller mer generella funktioner till mätdata istället[1]

6.1 Polynomiell regression

Antag att vi vill bestämma ett polynom av grad k som minimerar kvadratfelet. Modellen är att x_j är fixerade tal och att y_j är observationer av

$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \cdots + \beta_k x_j^k + \epsilon_j,$$

där $\epsilon_j \sim N(0, \sigma^2)$ är oberoende. Vi löser detta problem med linjär regression genom att låta

$$x_{j1} = x_j, \quad x_{j2} = x_j^2, \quad x_{j3} = x_j^3, \quad \cdots \quad x_{jk} = x_j^k,$$

och sedan betrakta modellen

$$Y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk} + \epsilon_j,$$

där $\epsilon_j \sim N(0, \sigma^2)$ är oberoende.

6.2 Exponentiell regression

Antag att vi har data som verkar vara följa en exponentialkurva. Modellen är att x_j är fixerade tal och att y_j är observationer av

$$Y_j = a \exp(b x_j) \cdot E_j \tag{1}$$

där E_j är **lognormal-fördelade** och oberoende.

Lognormal-fördelning

Definition. Slumpvariabeln X kallas **lognormal-fördelad** med parametrarna μ och σ om

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

Vi skriver $X \sim \text{Lognormal}(\mu, \sigma^2)$.

[1] Helt analogt med vad som gjorts i linjär algebra; tänk polynombaser. Eller Fourieranalys för den delen.

Det följer att $E(X) = \exp(\mu + \sigma^2/2)$ och $V(X) = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$ eftersom

$$E(h(X)) = \int_0^\infty \frac{h(x)}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) dx = \left/ \begin{matrix} y = \ln x \\ dy = \dfrac{dx}{x} \end{matrix} \right/$$
$$= \int_{-\infty}^\infty \frac{h(e^y)}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dx$$

så

$$E(X) = \int_{-\infty}^\infty \frac{e^y}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dx = \left/ \begin{matrix} u\sigma = y - \mu \\ dy = \sigma\, du \end{matrix} \right/$$
$$= \int_{-\infty}^\infty \frac{e^{\mu+\sigma^2/2}}{\sqrt{2\pi}} \exp\left(-\frac{(u-\sigma)^2}{2}\right) dx = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

och på samma sätt blir

$$E(X^2) = \exp\left(2\mu + 2\sigma^2\right).$$

vilket ger

$$V(X) = E(X^2) - E(X)^2 = \left(\exp(\sigma^2) - 1\right)\exp\left(2\mu + \sigma^2\right).$$

Sats. Om $X \sim \text{Lognormal}(\mu, \sigma^2)$ så är $\ln X \sim N(\mu, \sigma^2)$.

Bevis. Låt X vara lognormalfördelad och låt $Y = \ln X$. Eftersom exp är strängt växande gäller att

$$F_Y(y) = P(Y \le y) = P(\ln X \le y) = P(X \le e^y)$$

vilket medför att

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X(e^y)\, e^y = \frac{e^y}{e^y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln e^y - \mu)^2}{2\sigma^2}\right)$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

Således är $Y = \ln X \sim N(\mu, \sigma^2)$. □

Vi löser nu problemet i (1) med linjär regression genom att logaritmera sambandet:

$$\ln Y_j = \ln a + bx_j + \ln E_j = \beta_0 + \beta_1 x_j + \epsilon_j,$$

där $\epsilon_j \sim N(0, \sigma^2)$ är oberoende. Sen använder vi tekniker vi tagit fram!

7 Val av modell

Så låt oss säga att vi har en mängd mätdata i form av y-värden för en mängd olika värden på variabler $x_1, x_2, \ldots, x_k$. Hur ska vi välja modell? Tillför alla variabler något användbart? Hur jämför vi två olika modeller? Frågorna hopar sig.
Vad man alltid kan göra är att studera skattningen för σ^2. Denna skattning kommer i allmänhet från residualerna och idealiskt skulle dessa i princip vara lika med noll (perfekt lösning). Ett mindre värde på s^2 innebär alltså att modellen förklarar lite mer. Nu kan tilläggas att om man lägger till variabler kommer *alltid* s^2 att bli mindre (varför?), så vi behöver avgöra om skillnaden är signifikant.

(i) Val av variabler. Vilka har vi tillgång till? Vilka kan vi utesluta på grunden att de inte bör ingå i modellen? Är vissa variabler väldigt starkt korrelerade (i så fall kan det vara bättre att bara ta med en)?

(ii) Är sambandet linjärt? Kan det genom någon lämplig transformation skrivas som ett linjärt problem? Om det inte går kommer linjär regression fungera dåligt.

(iii) Vid flera möjliga modellval, hur testar vi om skillnaden mellan modellerna är signifikant? Vi vill inte ta med variabler i onödan.

Vi börjar med att diskutera begreppet **inkapslade modeller** (eller **nästlade**). Modeller där vi i någon mening kan säga den ena är en del av den andra.

8 Inkapslade modeller

Om vi har två modeller att välja mellan med syntesmatriserna X_1 respektive X_2. Vi låter H_1 och H_2 vara respektive hattmatriser, så blir

$$H_1 = X_1(X_1^T X_1)^{-1} X_1^T \quad \text{och} \quad H_2 = X_2(X_2^T X_2)^{-1} X_2^T.$$

Vi låter $\boldsymbol{\beta} \in \mathbf{R}^{k_1+1}$ respektive $\boldsymbol{\beta} \in \mathbf{R}^{k_2+1}$ för de olika modellerna. Dimensionerna för X_1 och X_2 är $n \times (k_1+1)$ respektive $n \times (k_2+1)$.

Inkapslade modeller

Definition. Vi kallar modell 1 för **inkapslad** i modell 2 om

$$V_1 = \{X_1\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbf{R}^{k_1+1}\} \subset \{X_2\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbf{R}^{k_2+1}\} = V_2.$$

Definitionen är lite abstrakt, men vad som säges är att kolonnrummet som spänns upp av X_1 ska vara ett underrum till kolonnrummet som spänns upp av X_2. Exempelvis gäller det att modellen

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

är inkapslad i modellen

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k + \beta_{k+1} x_{k+1} + \cdots + \beta_{k+p} x_{k+p} + \epsilon.$$

Detta följer eftersom de första $k+1$ kolonnerna i X_1 och X_2 är identiska. Detta är i huvudsak vad vi ska använda inkapslade modeller till: att undersöka om det blir signifikant bättre av att lägga till förklaringsvariabler (alternativ att det inte skadar att ta bort förklaringsvariabler).

Sats. Om $V_1 \subset V_2$ gäller att

$$H_1H_2 = H_2H_1 = H_1 \quad \text{och} \quad (I - H_1)(I - H_2) = (I - H_2)(I - H_1) = I - H_2.$$

Vidare gäller att $H_2 - H_1$ är en projektionsmatris med $\text{rank}(H_2 - H_1) = \text{rank}(H_2) - \text{rank}(H_1)$.

Bevis. Eftersom

$$H_1\boldsymbol{y} \in V_1 \subset V_2$$

för alla $\boldsymbol{y}$ följer det att $H_2H_1\boldsymbol{y} = H_1\boldsymbol{y}$. Eftersom H_1 och H_2 är symmetriska så medföljer även att $H_1H_2 = H_1$.
För den andra likheten noterar vi att $V_1 \subset V_2$ implicerar att ortogonalkomplementen uppfyller $V_2^\perp \subset V_1^\perp$. Således blir

$$(I - H_2)\boldsymbol{y} \in V_2^\perp \subset V_1^\perp$$

och

$$(I - H_1)(I - H_2)\boldsymbol{y} = (I - H_2)\boldsymbol{y}.$$

Analogt med ovan följer även att $(I - H_2)(I - H_1) = I - H_2$. Det faktum att $H_2 - H_1$ är en projektionsmatris följer av att den uppenbarligen är symmetrisk och

$$(H_2 - H_1)^2 = H_2^2 - H_2H_1 - H_1H_2 + H_1^2 = H_2 - 2H_1 + H_1 = H_2 - H_1.$$

Således är samtliga egenvärden 0 eller 1 och

$$\operatorname{rank}(H_2 - H_1) = \operatorname{tr}(H_2 - H_1) = \operatorname{tr}(H_2) - \operatorname{tr}(H_1) = \operatorname{rank}(H_2) - \operatorname{rank}(H_1).$$

Den sista likheten på grund av att H_1 och H_2 också är projektionsmatriser. □
Vi kan nu formulera (och bevisa) en variant på regressionsanalysens 2:a huvudsats. Den går att formulera mer generellt, men detta är mer än tillräckligt för våra ändamål.

Regressionsanalysens 2:a huvudsats

Sats. Låt H_1 och H_2 ha rang $k_1 + 1$ respektive $k_2 + 1$. Om $V_1 \subset V_2$ så gäller att:

(i) $\mathrm{SS}_\mathrm{E}^{(2)}$ och $\mathrm{SS}_\mathrm{E}^{(1)} - \mathrm{SS}_\mathrm{E}^{(2)}$ är oberoende;

(ii) $\dfrac{\mathrm{SS}_\mathrm{E}^{(2)}}{\sigma^2} \sim \chi^2(n - k_2 - 1)$;

(iii) samt om $E(\boldsymbol{Y}) = \boldsymbol{\mu_1} = X_1\boldsymbol{\beta_1}$ så är $\dfrac{\mathrm{SS}_\mathrm{E}^{(1)} - \mathrm{SS}_\mathrm{E}^{(2)}}{\sigma^2} \sim \chi^2(k_2 - k_1)$.

Bevis. Vi ser att

$$\mathrm{SS}_\mathrm{E}^{(2)} = \boldsymbol{Y}^T(I - H_2)\boldsymbol{Y}$$

och

$$\mathrm{SS}_\mathrm{E}^{(1)} - \mathrm{SS}_\mathrm{E}^{(2)} = \boldsymbol{Y}^T(I - H_1 - (I - H_2))\boldsymbol{Y} = \boldsymbol{Y}^T(H_2 - H_1)\boldsymbol{Y}.$$

Eftersom

$$(I - H_2)(H_2 - H_1) = H_2 - H_1 - H_2^2 + H_2H_1 = -H_1 + H_1 = 0$$

så kommer $(I - H_2)\boldsymbol{Y}$ och $(H_2 - H_1)\boldsymbol{Y}$ att vara okorrelerade och normalfördelade. Således är dessa variabler oberoende vilket medför punkt (i). Punkt (ii) är identisk med resultatet från regressionsanalysens första huvudsats (se förra föreläsningen). Den sista punkten följer av ett liknande argument som på förra föreläsningen. Först, eftersom $V_1 \subset V_2$, så finns ett $\boldsymbol{\alpha} \in \mathbf{R}^{k_2+1}$ så att $X_2\boldsymbol{\alpha} = X_1\boldsymbol{\beta_1}$. Detta medför att

$$(H_2 - H_1)X_1\boldsymbol{\beta_1} = H_2X_2\boldsymbol{\alpha} - X_1\boldsymbol{\beta_1} = X_2\boldsymbol{\alpha} - X_1\boldsymbol{\beta_1} = X_1\boldsymbol{\beta_1} - X_1\boldsymbol{\beta_1} = 0,$$

så $E((H_2 - H_1)\boldsymbol{Y}) = 0$. Därav följer det att

$$\mathrm{SS}_{\mathrm{E}}^{(1)} - \mathrm{SS}_{\mathrm{E}}^{(2)} = \boldsymbol{\epsilon}(H_2 - H_1)\boldsymbol{\epsilon}^T.$$

Eftersom $H_2 - H_1$ är en projektionsmatris med rang $k_2 - k_1$ och $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$ finns en ON-matris C så att med $\epsilon = C\boldsymbol{Z}$ blir

$$\boldsymbol{\epsilon}(H_2 - H_1)\boldsymbol{\epsilon}^T = \sum_{j=1}^{k_2-k_1} Z_j^2,$$

där $Z_j \sim N(0, \sigma^2)$ är oberoende. Alltså stämmer fördelningen i punkt (iii) eftersom kvadratsumman av oberoende $N(0,1)$-variabler blir χ^2-fördelad med frihetsgraden lika med antalet termer. □

Anmärkning. Om vi inte skulle anta att $E(\boldsymbol{Y}) = \boldsymbol{\mu_1}$ så skulle vi fortfarande erhålla en χ^2-fördelningen, men den blir inte centrerad. Överkurs.

Exempel

Vi har två modeller:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

och

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + \cdots + \beta_{k+p} x_{k+p} + \epsilon.$$

Hur kan man testa om $\beta_{k+1} = \beta_{k+2} = \cdots = \beta_{k+p} = 0$ (dvs om de tillförda variablerna hjälper på en signifikant nivå)?

Lösning. Enligt föregående diskussion är modell 1 inkapslad i modell 2. Låt nollhypotesen ges av

$$H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_{k+p} = 0,$$

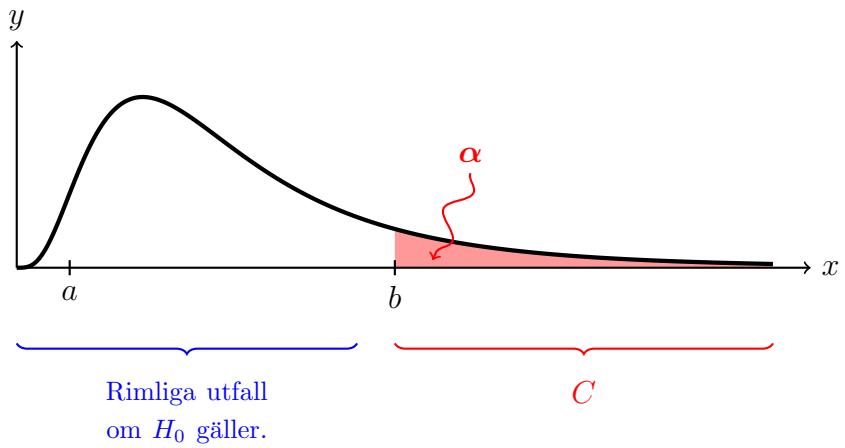
med mothypotesen

$$H_1 : \text{något } \beta_j,\ j = k+1, k+2, \ldots, k+p, \text{ är inte } = 0.$$

Om H_0 är sann så gäller att $\boldsymbol{Y} \sim N(X_1\boldsymbol{\beta_1}, \sigma^2 I)$, så satsen ovan medför direkt att

$$W = \frac{(\mathrm{SS}_{\mathrm{E}}^{(1)} - \mathrm{SS}_{\mathrm{E}}^{(2)})/p}{\mathrm{SS}_{\mathrm{E}}^{(2)}/(n-k-p-1)} \sim F(p, n-k-p-1) \quad \text{om } H_0 \text{ är sann}$$

eftersom det är en kvot av oberoende χ^2-fördelade variabler. Om H_0 inte är sann kommer det att göra att W tenderar att bli stor, så vårt kritiska område kommer ges av $C =]c, \infty[$ för något $c > 0$.



9 Stegvis regression

En tänkbar lösning på problemet att hitta en modell som tar med precis de variabler som är signifikanta är givetvis att helt enkelt testa alla kombinationer. Med k möjliga förklaringsvariabler ger det 2^k olika modeller. Vi kan utföra regression för var och en och sedan undersöka vilka variabler som förefaller vara relevanta. Otympligt? Jo, kanske det Så en annan variant är att lägga till en variabel i taget till vi inte ser någon signifikant skillnad längre när vi lägger till fler variabler. Så hur börjar vi?
Den bästa förklaringsvariabeln är alltid den som är starkast korrelerade med y. Detta fenomen följer av exemplet från förra föreläsningen angående enkel linjär regression där vi visade att $\mathrm{SS_E} = (1 - r^2)\sum_{j=1}^{n}(y_j - \overline{y})^2$. Däremot kan vi inte direkt se vilken den näst bästa är utan att utföra en regression. Så processen kommer att se ut enligt följande.

(i) Jämför korrelationen mellan $\boldsymbol{y}$ och de olika x-kolonnerna i X och välj den där r^2 är störst som första förklaringsvariabel.

(ii) Testa och lägg till var och en av resterande variabler en i taget och beräkna $\mathrm{SS_E}$ för varje modell. Välj den variabel som minimerar $\mathrm{SS_E}$. Detta är den nästa bästa förklaringsvariabeln. Lägg till den.

(iii) Testa den nya modellen genom att endera göra ett F-test för att se om den är signifikant bättre eller gör ett t-test för att se om hypotesen $H_0 : \beta_i = 0$ för den tillagda β_i kan förkastas. Om variabeln inte tillför något är vi färdiga. Annars lägg till variabeln i modellen.

(iv) Upprepa steg 2 tills dess att vi inte får någon signifikant skillnad när vi lägger till en ny variabel.


Vi kan *endast* hitta den bästa förklaringsvariabeln genom att studera korrelation mellan y och de olika x_i-variablerna. Eventuell övrig information från exempelvis kovariansmatrisen ger inte nödvändigvis någon information om vad som blir bäst när man väl tagit med den bästa variabeln. Ny analys krävs efter regressionssteget!

10 Kategorier och ”dummy”-variabler

Ibland har man data som är beroende av någon storhet som är binär (eller åtminstone har diskreta nivåer). Till exempel skulle det kunna handla om en modell för åtgång av förbrukningsvaror hos ett café vid stranden. Beroende på om det är sommar eller vinter kanske saker och ting ser helt annorlunda ut. Vi kan då lägga till en variabel i modellen som har värdet 1 vid sommar och 0 när det är vinter. På det sättet kan vi ta med all data i en och samma modell.

11 Problem och fallgropar

Det finns en uppsjö med problem förknippade med regression.

11.1 Stark korrelation

Om två variabler är starkt korrelerade innebär det att matrisen X nästan blir singulär (den blir dåligt **konditionerad**), vilket ställer till det rent numeriskt då avrundningsfel och dylikt nu kan förändra svar drastiskt. Systemet blir helt enkelt väldigt störningskänsligt.
Man brukar undvika starkt korrelerade variabler.
Ett specialfall är när matrisen $X^T X$ inte är inverterbar. Då behöver någon/några variabler tas bort.

11.2 Extrapolation

När vi har våra uppmätta data så får vi direkt ett rätblock i $\mathbf{R}^k$ där

$$x_i^- \leq x_i \leq x_i^+, \quad i = 1, 2, \ldots, k.$$

Talen $x_i^\pm$ är helt enkelt max och min vid mätningen för den uppmätta variabeln x_i. Mellan dessa gränser undersöker vi en linjär regressionsmodell. Denna modell bör inte okvalificerat användas för att uttala sig (prediktera) något utanför rätblocket.

11.3 Residualfördelning

Se till att göra några undersökningar om residualerna. Om de uppvisar ett mönster är det ett tecken på att felen inte uppfyller de krav vi ställt. Om inte felen är normalfördelade (med samma varians) så leder detta till att *samtliga* tester (F-test, varianstest, test för $\beta_i = 0$ etc) inte är tillförlitliga.

Föreläsning 10: Pearsons χ^2-test

Johan Thim (johan.thim@liu.se)

31 augusti 2018

"Oh, you suffer beautifully."
–Pinhead

Antag att vi har följande sitaution.

(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 $A_1, \ldots, A_k$ och låt $p_j = P(A_j)$ vara respektive sannolikhet. Då är $p_1 + p_2 \cdots + p_k = 1$.

(iii) Låt Y_i, $i = 1, 2, \ldots, k$, 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 gör är att betrakta uppdelningar av denna typ och ställa upp hypotestest där vi låter nollhypotesen H_0 ges av

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

där $p_1, p_2, \ldots, p_k$ är sannolikheter så att $p_1 + \cdots + p_k = 1$, 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 $E(Y_j) = n \cdot p_j$, $j = 1, 2, \ldots, k$. 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

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

Att detta blir approximativt χ^2-fördelat följer av följande sats.

Sats. Med beteckningarna ovan gäller att

$$\sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \xrightarrow{D} Q,$$

där $Q \sim \chi^2(k-1)$. Konvergensen är i meningen av distributionen (på samma sätt som i CGS).

Bevis. Eftersom Y_j är binomialfördelad 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). Konvergensen är i meningen att fördelningsfunktionen $F_{n,j}(y) \to \Phi(y)$ 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, $j = 1, 2, \ldots, k$. Eftersom väntevärdet är $E(Y_j) = np_j$ kommer

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

För att beräkna $E(Y_i Y_j)$ går vi tillbaka till variablerna X_i, $i = 1, 2, \ldots, n$. 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 $i \neq j$,

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

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 $\boldsymbol{Z} = (Z_1\ Z_2\ \cdots\ Z_k)^T$ ha utseendet

$$C_{\boldsymbol{Z}} = \begin{pmatrix} 1-p_1^2 & -\sqrt{p_1p_2} & -\sqrt{p_1p_3} & \cdots & -\sqrt{p_1p_k} \\ -\sqrt{p_2p_1} & 1-p_2^2 & -\sqrt{p_2p_3} & \cdots & -\sqrt{p_2p_k} \\ -\sqrt{p_3p_1} & -\sqrt{p_3p_2} & 1-p_3^2 & \cdots & -\sqrt{p_3p_k} \\ \vdots & \vdots & & \ddots & \vdots \\ -\sqrt{p_kp_1} & -\sqrt{p_kp_2} & -\sqrt{p_kp_3} & \cdots & 1-p_k^2 \end{pmatrix}.$$

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

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

så $I - \boldsymbol{p}\boldsymbol{p}^T$ är en projektionsmatris och har därför egenvärdena $\lambda = 0$ och λ_1. Vi har nu att $\boldsymbol{Z} \sim N(\boldsymbol{0}, C_{\boldsymbol{Z}})$. På samma sätt som i beviset av regressionsanalysens huvudsats ser vi att

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

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

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

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 $n \to \infty$) och säger inget direkt om vad som gäller i det enskilda fallet. En tummregel är att vi vill ha $np_j \geq 5$ för $j = 1, 2, \ldots, k$ 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.

2 Test av given diskret fördelning

Låt $X_1, X_2, \ldots, X_n$ 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

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

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

$$P(X = 0) = \binom{2}{0} \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^2 = \frac{16}{25},$$
$$P(X = 0) = \binom{2}{1} \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^1 = \frac{8}{25},$$
$$P(X = 0) = \binom{2}{2} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^0 = \frac{1}{25}.$$

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 förök (om H_0 är sann):

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

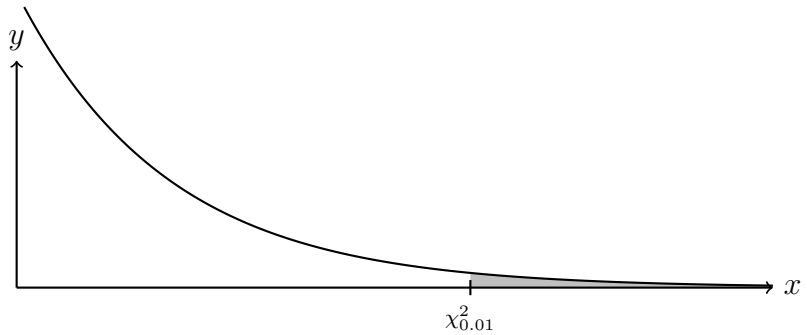
Testvariabeln q ges nu av

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

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

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

ur tabell.



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

3 Test för kontinuerlig fördelning

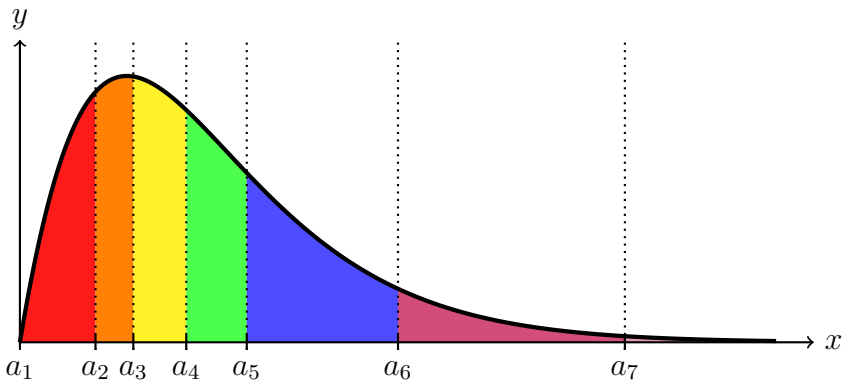
Om vi istället har en kontinuerlig situation där vi vill testa om mätdata följer en given fördelning F måste vi agera lite annorlunda. Vi skulle önska att ställa upp $H_0 : X \sim F$ mot $H_1 : X$ har ej fördelningen F. Men detta blir lite för komplicerat i det generella fallet.
Istället gör vi så att 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, $i = 1, 2, \ldots, n$ vara oberoende och likafördelade variabler med täthetsfunktion $f(x)$. Vi väljer a_j, $j = 1, 2, \ldots, k+1$, så att

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

och definierar $A_j = [a_j,\ a_{j+1}[$ för $j = 2, 3, \ldots, k$ 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 $p_1 + p_2 + \cdots + p_k = 1$ 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

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

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.

4 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 beter sig bra). Beviset..

5 Homogenitetstest

Det kan ofta vara intressant att avgöra om till exempel åsikter skiljer sig åt mellan olika grupper.
varför $Q \overset{\text{appr.}}{\sim} \chi^2((r-1)(k-1))$
om $n_i\widehat{p_j}$ är stora. En rimlig tummregel är att $n_i\widehat{p_j} \geq 5$.

Exempel

Ifrån en stor population frågar vi två grupper om de tycker A eller B.

Grupp	A	B
G_1	59	41
G_2	145	55

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

Lösning.

Grupp	A	B	
G_1	59	41	100
G_2	145	55	200
$G_1 + G_2$	204	96	300
$\widehat{p_j}$	0.68	0.32	1.0

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

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[a]. Examinator funderade över om resultaten är homogena över några olika grupper och samlade in följande siffror på internet:

	Angel of Death	*Raining Blood*
`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.

[a]Självklart är *Angel of Death* den bästa av dessa två, men det är inte poängen!

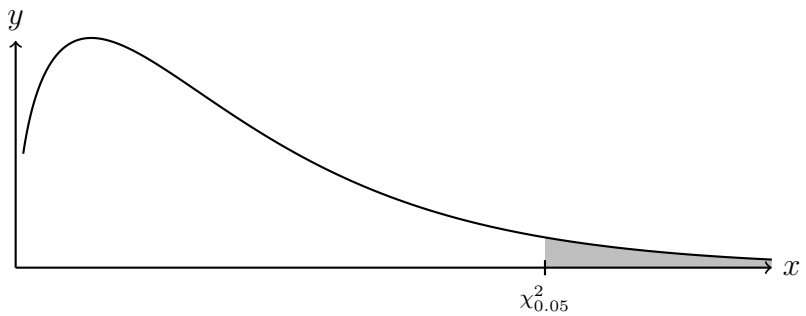
Lösning. Först kompletterar vi tabellen med all information som behövs:

	Angel of Death	*Raining Blood*	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
$\widehat{p}_j$ (skattat p_j)	$\widehat{p}_1 = 0.552$	$\widehat{p}_2 = 0.448$	1.00

Vi beräknar observationen q

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

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 $Q > \chi^2_\alpha(3)$, d v s om den observerade testvariabeln hamnar utanför det skuggade området i figuren ovan. Med $\alpha = 0.05$ finner vi att $\chi^2_{0.05}(3) = 7.81$ ur tabell, så $Q > \chi^2_\alpha(3)$. 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%.

6 Bonus: Kolmogorov-Smirnoff

Empiriska fördelningsfunktionen..

A short glossary of terms (with Swedish translations)

Johan Thim (johan.thim@liu.se)

18 september 2018

Confidence interval (Konfidensintervall): An interval $I = I_\theta^{1-\alpha} = (a, b)$ such that $\theta \in I$ with probability $1 - \alpha$. The probability α is referred to as the level of significance, which is the probability for the interval to *not* cover the unknown θ.

Consistent (Konsistent): An estimator $\widehat{\Theta}$ is called consistent if for every $\epsilon > 0$

$$\lim_{n\to\infty} P(|\widehat{\Theta} - \theta| > \epsilon) = 0.$$

The estimator $\widehat{\Theta}_n$ *converges in probability* to θ.

Distribution (Fördelning): The distribution of a random variable. Typically characterized by either the density function (täthetsfunktion) in the continuous case or the probability function in the discrete case.

Estimate (Skattning): The same as point estimate.

Estimator (Skattningsvariabel): The random variable $\widehat{\Theta}$ of which the estimate $\widehat{\theta}$ is an observation of; $\widehat{\theta}$ is an observation $\widehat{\Theta}$.

Expectation (Väntevärde): The expectation of a random variable: $E(X)$.

Hypothesis testing (Hypotesprövning): We test, e.g., $H_0 : \mu = \mu_0$ against some alternative hypothesis, e.g., $H_1 : \mu \neq \mu_0$.

Null hypothesis (Nollhypotes): The hypothesis H_0 that we are trying to disprove.

Alternate hypothesis (Mothypotes): The hypothesis H_1 (for example).

Critical region (Kritiskt område): A region C such that if the *statistic* $t(x_1, \ldots, x_n) \in C$, we reject H_0.

Reject (Förkasta): We reject H_0 if the statistic $t = t(x_1, \ldots, x_n)$ ends up in the critical region: $t \in C$.

Point estimate (Punktskattning): *Any* estimate $\widehat{\theta}$ of an unknown parameter θ.

Power (Styrka): The power at θ of a hypothesis test is the probability of rejecting H_0 when θ is the real parameter value. Defined by $h(\theta) = P(\text{reject } H_0 \,|\, \theta \text{ is the real value})$.

***p*-value (*p*-värde)**: For a given sample $x_1, \ldots, x_n$, the p-value is the lowest probability where we can reject H_0. The lowest possible significance level is p.

Sample (Stickprov): A sequence of observations $x_1, x_2, \ldots, x_n$, typically assumed to be independent observations of random variables $X_1, X_2, \ldots, X_n$ of some distribution(s) that depend on something unknown (a parameter θ).

Significance (Signifikans): The significance of a hypothesis test is the probability of rejecting H_0 when H_0 is true.

Random sample (Slumpmässigt stickprov): The sequence $X_1, X_2, \ldots, X_n$ of random variables that corresponds to the sample $x_1, x_2, \ldots, x_n$.

Unbiased (Väntevärdesriktig): An estimator $\widehat{\Theta}$ is called unbiased if $E(\widehat{\Theta}) = \theta$.

Recap: distributions, point estimates and confidence

Johan Thim (johan.thim@liu.se)

September 23, 2018

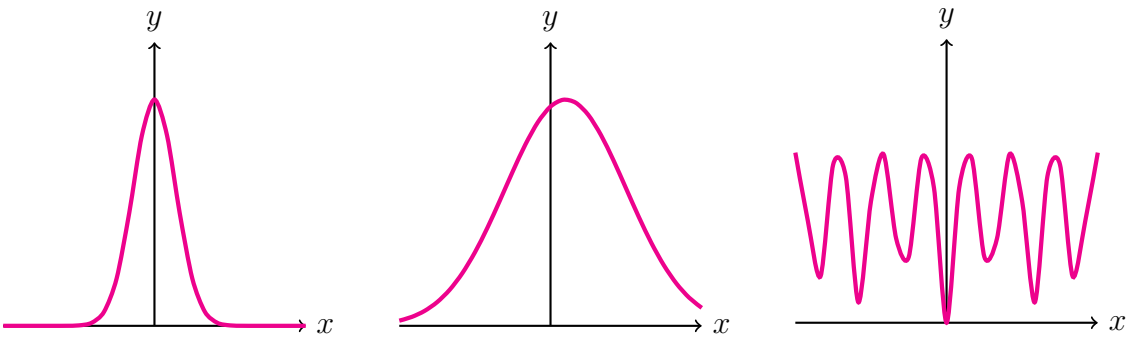
So what's the deal with all the terminology? Distributions, samples, random samples, estimators and confidence intervals. How does it all fit together? That's what I've been trying to show you guys, but maybe the idea gets lost in all the details. Trees, forests and all... So let us recap a bit and collect what we know.
The starting point is pretty much the same $x_1, x_2, \ldots, x_n$. This is typically measured data. What can we do with this? To do any form of reasonable analysis, we need a model. What we usually do is assume that the sample consists of observations of random variables, so:

Assumption 1: $X_1, X_2, \ldots, X_n$ are random variables and $x_1, x_2, \ldots, x_n$ are observations.

Next up is independence. We almost always assume that the random sample $X_1, X_2, \ldots, X_n$ is independent. Analysis gets much harder without this assumptions, so:

Assumption 2: $X_1, X_2, \ldots, X_n$ are independent.

So now we have independent random variables. What's next? Well, how are they distributed? Assuming a continuous distribution for simplicity, how does the probability density look?



Like this? Or this? Or maybe this?

We obviously have to assume something about the distributions to obtain something useful. So:

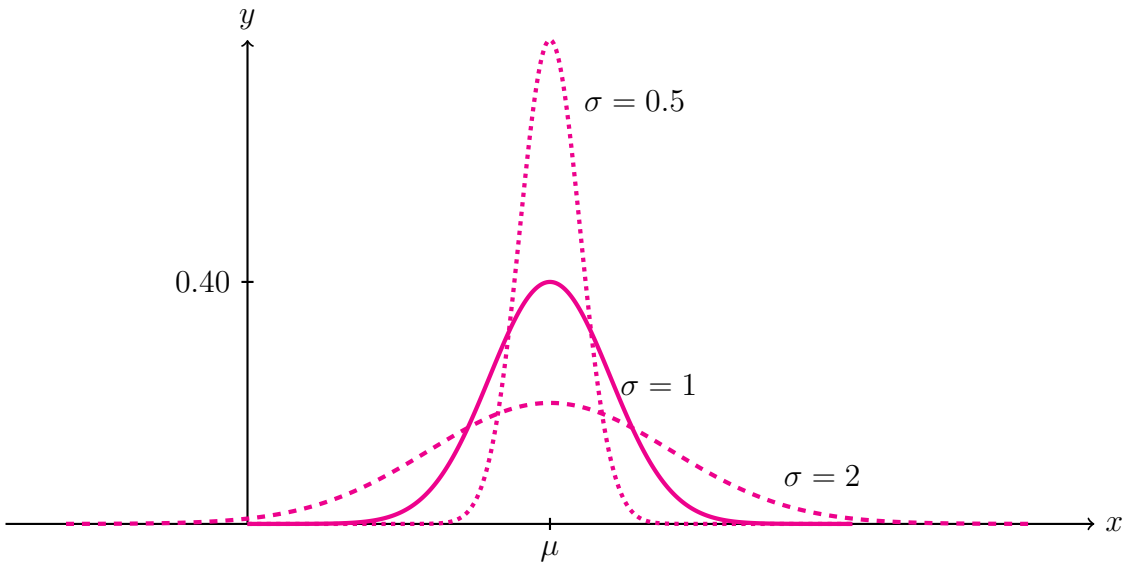
Assumption 3: $X_1, X_2, \ldots, X_n$ have known *types* of distribution, that is, we know for example that they are normally distributed. However, the exact distribution depends on something unknown: the parameter θ (which might be a vector of unknown parameters). It is not necessary that they all share the exact same distribution, but usually we will assume this. However, it is important that if the have different distributions, they depend on the same unknown parameter.

Normal distribution

Consider the normal distribution. Its shape is well-known (the Bell curve or Gaussian), but it might be moved around and it might be stretched (or contracted). The parameters μ and σ does this. How? Well, the density function for a normal distribution looks as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbf{R},$$

where $\mu \in \mathbf{R}$ and $\sigma > 0$ are parameters. It turns out that they happen to be the expectation and the standard deviation, but that's not clear from the definition above but follows when doing the necessary calculations. Right now they are just parameters for the distribution.



We see the same type of shape, but different values on the parameters yield different curves. Our aim now would be — using a sample $x_1, x_2, \ldots, x_n$ from $N(\mu, \sigma^2)$ — to estimate the unknown parameters μ and/or σ.
To make calculations easier, we usually make the following assumption:

Assumption 4: $X_1, X_2, \ldots, X_n$ are independent.

The Goal

If $x_1, x_2, \ldots, x_n$ are a sample from a distribution F that depends on an unknown parameter θ, we wish to find a *good* way of using the known quantities $x_1, x_2, \ldots, x_n$ to **estimate** θ by some function

$$\widehat{\theta} = g(x_1, x_2, \ldots, x_n).$$

What about θ, $\widehat{\theta}$, $\widehat{\Theta}$...

What about the big ball with the H inside and the hat ontop[1]? There are three types of θ:s involved here. Let's see how this works.

[1] Thank you dear student, for the subtle hint about my handwriting on the board..

So the situation is as follows. We have a method of finding an estimate $\widehat{\theta}$ for the unknown parameter θ by using the sample $x_1, x_2, \ldots, x_n$, namely by the so called **statistic**

$$\widehat{\theta} = g(x_1, x_2, \ldots, x_n).$$

Question 1. If we repeat the "experiment" and thus obtain a new sample $y_1, y_2, \ldots, y_n$, what can we say about the estimate $\widehat{\theta} = g(y_1, y_2, \ldots, y_n)$?

Obviously we won't get the same estimate (in general) since the values will likely have changed, so what's going on? This is where the probability comes in to play. We know (by assumption) what type of distribution we're working with. We know how this distribution depend on the unknown parameter θ. So we can – in theory at least – do calculations with regards to the estimator viewed as a random variable as long as we allow our answers to contain the unknown θ. How do we move to something random? Considering that we view $x_1, x_2, \ldots, x_n$ as observations of random variables $X_1, X_2, \ldots, X_n$, we just exchange all instances of the former by the latter. Thus we obtain the estimator

$$\widehat{\Theta} = g(X_1, X_2, \ldots, X_n)$$

as an n-dimensional random variable. Note that this is the *same* function $g\colon \mathbf{R}^n \to \mathbf{R}^p$ that was used when defining $\widehat{\theta}$ (p is the number of parameters we are estimating).

Question 2. Can we use $\widehat{\Theta}$ to obtain some type of bounds for the unknown θ with a given probability?

The answer is yes, at least if we can transform $\widehat{\Theta}$ to something with a known distribution. This is the procedure used to find **confidence intervals**. What is a confidence interval? We'll get back to this down below.

What would we like to happen?

Good Property 1. We do *not* want the estimator to be biased. By that we mean that we want the estimator $\widehat{\theta}$ to *be* the unknown parameter θ on *average*. Well, that's rather unspecific, so instead we mathematically define an unbiased estimator as an estimator such that $E(\widehat{\Theta}) = \theta$.

Good Property 2. We would like for our estimator to have the property that as the sample grows larger, the probability of $\widehat{\Theta}$ being off from θ is tending to zero. This is **consistency**. In mathematical terms, we want the following to hold. Let $\widehat{\Theta}_n$ be the estimator for a sample of size n. Then for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\widehat{\Theta}_n - \theta| > \epsilon) = 0.$$

Phrased differently, this means that the sequence $\widehat{\Theta}_n$ of random variables converges to θ in probability. It is difficult to work with this directly, so we normally use a result that follows from Chebyshev's inequality. Namely that if $V(\widehat{\Theta}_n) \to 0$ as $n \to \infty$ and $\widehat{\Theta}_n$ is unbiased (note that this is needed for the theorem to hold, not for the estimator to be consistent), then the estimator is consistent.

Expectation and variance

We've used (implicitly or explicitly) some results concerning the expectation and the variance above. Let us recapture what's allowed. Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed with $E(X_k) = \mu$ and $V(X_k) = \sigma^2$ for $k = 1, 2, \ldots, n$. Then

$$E\left(a_0 + \sum_{k=1}^{n} a_k X_k\right) = a_0 + \sum_{k=1}^{n} a_k E(X_k) = a_0 + \mu \sum_{k=1}^{n} a_k,$$

since the expectation is a *linear* operator (remember that it is either a sum or an integral, both of which are linear). For the variance, it is true that

$$V\left(a_0 + \sum_{k=1}^{n} a_k X_k\right) = /X_k \text{ are independent } / = \sum_{k=1}^{n} a_k^2 V(X_k) = \sigma^2 \sum_{k=1}^{n} a_k^2.$$

The assumption about independence is crucial here. Without it, the sum above will get very messy with covariances all over the place. Note though that for the expectation, independence is not required.

Something neat with the normal distribution is that linear combinations of normally distributed variables are still normally distributed (the mean and variance might change, but we've seen above how that works). So assuming for a minute that $X_k \sim \mathrm{N}(\mu, \sigma^2)$ for $k = 1, 2, \ldots, n$, we have

$$X := \sum_{k=1}^{n} X_k \sim \mathrm{N}(n\mu, n\sigma^2) \qquad \text{and} \qquad \overline{X} := \frac{1}{n} \sum_{k=1}^{n} X_k \sim \mathrm{N}(\mu, \sigma^2/n).$$

This is clear from the formulas above. Note in particular that

$$V(\overline{X}) = V\left(\frac{1}{n} \sum_{k=1}^{n} X_k\right) = \frac{1}{n^2} V\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2} \sum_{k=1}^{n} V(X_k) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

where we used the independence of the variables. This equality shows that as n grows larger, the variance of $\overline{X}$ tends to zero. Larger sample size means less variance for $\overline{X}$.

Random or not?

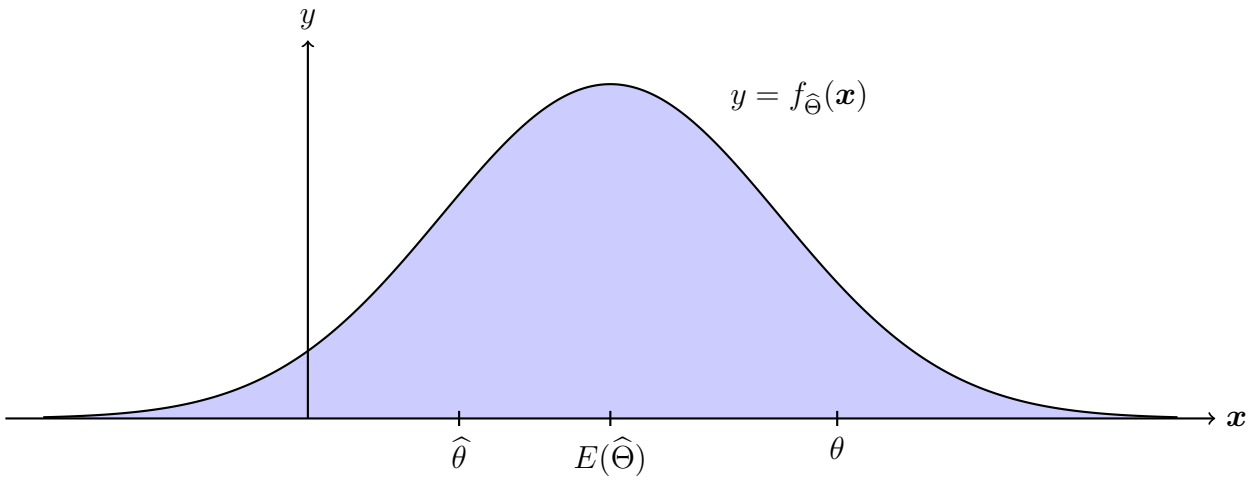
Be careful with explicitly showing the difference between what is random and what is not. We have three quantities:

(i) θ – real value. Unknown but not random.

(ii) $\widehat{\theta}$ – estimate for θ. Known value calculated from the sample $x_1, x_2, \ldots, x_n$.

(iii) $\widehat{\Theta}$ – The estimator. This is a random variable!

If you want to calculate probabilities (so expectations and variances and such), you have to use $\widehat{\Theta}$.



Note the following:

(i) $\widehat{\theta}$ is calculated from observations, so it might end up anywhere basically.

(ii) We do not now if the expectation $E(\widehat{\Theta})$ coincides with the unknown θ, there might be a bias.

(iii) The random variable $\widehat{\Theta}$ is n-dimensional (since it depends on $X_1, X_2, \ldots, X_n$, hence the bold font for x in the graph above) and might also be vector-valued in the case that $\theta \in \mathbf{R}^p$ with $p > 1$. In the case when $p > 1$, the density gets difficult to render on paper though...

A non-biased estimator $\widehat{\Theta}$ of θ will — on average — hit the unknown θ. This is a direct consequence of the law of large numbers. What this means more exactly is that if we were to form the average of estimates (repeating the experiment yielding $x_1, x_2, \ldots, x_n$ and find an estimate $\widehat{\theta}_k$ for each $k = 1, 2, 3, \ldots$), this average will converge to θ with probability one (that's the strong law of large numbers) with reasonable assumptions on the distributions.

Certainty

So we have found and estimator $\widehat{\Theta}$ of some unknown θ. Can we use the information contained in the distribution of $\widehat{\Theta}$ to say something about which values for estimates of θ that are reasonable? I mean, if we look at the graph of the density function (or probability function). we can see where it is likely that observations end up (around points where large amounts of probability mass is accumulated). In other words, can we find a set I such that $\theta \in I$ with some given probability? That was basically question 2 above.
Let's assume that $\theta \in \mathbf{R}$ (so we only have one dimension). A **confidence interval** I with **confidence degree** $1-\alpha$ $(0 < \alpha < 1)$ is *any* interval such that θ belongs to it with probability $1-\alpha$. The end points of such an interval typically need to be observations of random variables (transformations of $\widehat{\Theta}$). So the systematic question now is how to go from the estimator $\widehat{\Theta}$ — which depends on the unknown parameter unless something unusual occurs — to something with a completely known distribution. Let's look at a couple of examples.

Assume that $x_1, x_2, \ldots, x_n$ is a sample of observations of $X_1, X_2, \ldots, X_n$ of independent identically distributed random variables. The type of distribution is known but not some unknown parameter.

Estimator for the variance. Since we introduced the sample variance, that seems to be a reasonable place to start. Indeed, we have seen that $E(S^2) = \sigma^2$, so it is an unbiased estimator of σ^2. It is also a consistent estimator. To say something more specific, we need to now the type of distribution we're dealing with. Let's assume that we have samples from $N(\mu, \sigma^2)$. Then

$$V := \frac{(n-1)}{\sigma^2} S^2 \sim \chi^2(n-1)$$

according to Cochran's theorem. This is nice, since we now have a distribution that is completely known. To find a confidence interval, we find numbers a and b such that

$$P(a < V < b) = 1 - \alpha.$$

Note that a and b depend on both α and n and that we need to use a table or computer software. We then solve $a < V < b$ for σ^2, obtaining that

$$\frac{(n-1)S^2}{b} < \sigma^2 < \frac{(n-1)S^2}{a}.$$

To obtain a confidence interval (the above expression is not a fixed interval since the limits are stochastic), we need to estimate all involved random variables. The natural thing to do is to use the sample variance s^2 to estimate S^2. Thus we get the confidence interval

$$I_{\sigma^2} = \left(\frac{(n-1)s^2}{b}, \ \frac{(n-1)s^2}{a} \right).$$

Estimator for the expectation. The expectation is where X_k ends up "on average," so a reasonable starting point would be to consider the mean value $\overline{X} = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} X_k$. We know that $E(\overline{X}) = \mu$ and $V(\overline{X}) = \sigma^2/n$ if $E(X_k) = \mu$ and $V(X_k) = \sigma^2$. To transform our estimator into something with a known distribution, we standardize the variable by removing the mean and dividing by the standard deviation. If we again assume that we have samples from $N(\mu, \sigma^2)$, we see that

$$Z := \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

So if σ is known, this works great. If σ is unknown, we estimate it by s. This means that we use

$$T := \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

instead, a result that follows from Gosset's theorem. Proceeding as above, we find a number t (both distributions are symmetric with respect to the y-axis) such that

$$P(-t < T < t) = 1 - \alpha$$

in the case when σ is unknown. We find the number t in a table or by using computer software. Note that it depends on both α and n. If σ is known, we use Z and the normal distribution instead. Solving for μ in the inequality, we see that

$$-t < T < t \quad \Leftrightarrow \quad \overline{X} - t\frac{S}{\sqrt{n}} < \mu < \overline{X} + t\frac{S}{\sqrt{n}}.$$

Estimating S by s and $\overline{X}$ by $\overline{x}$, we obtain the confidence interval

$$I_\mu = \left(\overline{x} - t\frac{s}{\sqrt{n}}, \ \overline{x} + t\frac{s}{\sqrt{n}} \right).$$

To be continued...!

Linear Regression: a couple of worked examples in MATLAB

Johan Thim (johan.thim@liu.se)

October 1, 2018

1 Introduction

The idea of this document is just to – briefly – show how we can use MATLAB to carry out linear regression and get quantities necessary to perform the usual tests.

2 Example 1: simple linear regression

Exempel

Suppse that we have the following data:

x	1.0000	1.1000	2.0000	2.5000	3.5000	4.0000	4.2000	7.7000	9.0000	10.0000	13.0000	13.5000	15.5000	17.0000
y	0.3002	9.2698	6.4508	5.8739	9.4295	8.5901	9.1517	19.3794	21.8181	23.8344	28.3430	25.5850	33.4345	38.2605

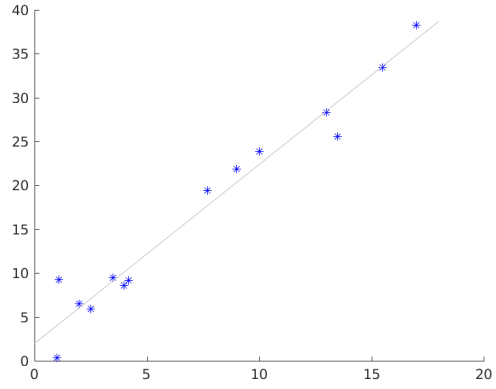
Find the empirical correlation between x and y. Is a linear relationship reasonable? Using the model

$$Y_j = \beta_0 + \beta_1 x_j + \epsilon_j,$$

where $\epsilon_j \sim N(0, \sigma^2)$ are independent, find estimates for β_0 and β_1 and perform the hypothesis test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at the significance level 1%. What's the R^2-value? Also find a prediction interval for a future observation y_0 at $x = 5$ with confidence level 95%.

Solution. We start by looking at a scatter plot (a good practice). The command `lsline` draws in a regression line matching the data.

```
>> scatter(x,y,'b*');
>> hold on
>> lsline
```



Using `corr` we can find the empirical correlation:

```
>> corr(x,y)
ans =
    0.9798
```

so yes, a linear relationship is reasonable. To find $\widehat{\boldsymbol{\beta}}$, we perform the regression:

```
>> rs = regstats(y,x,'linear','all');
>> betahat = rs.tstat.beta
betahat =
    1.9659
    2.0404
```

So our point estimates are $\widehat{\beta}_0 = 1.9659$ and $\widehat{\beta}_1 = 2.0404$. Let's assume that H_0 is true. Then

$$T = \frac{\widehat{\beta}_1 - 0}{S\sqrt{h_{11}}} \sim t(n-2),$$

where $s^2 = \dfrac{\mathrm{SS_E}}{n-2}$ is used to estimate σ^2 and h_{11} is the second element on the diagonal of $(X^TX)^{-1}$. We can find s^2 from the regression stats:

```
>> s2 = rs.mse
s2 =
    5.8841
```

and we can find $(X^TX)^{-1}$ by using the covariance matrix for $\widehat{\beta}$ since $C_{\widehat{\beta}} = \sigma^2(X^TX)^{-1}$,

```
>> Cbetahat = rs.covb;
>> XtXinv = Cbetahat/s2
XtXinv =
    0.2072   -0.0183
   -0.0183    0.0025
```

so it is clear that our test statistic is

$$t = \frac{2.0404 - 0}{\sqrt{5.8841 \cdot 0.0025}} = 16.8231.$$

From table (or using `tinv(0.995,12)`) we find $t_{0.005}(12) = 3.0545$ so we reject H_0 since $t > 3.0545$. Note that we can find the t-statistics by (where we see that we have introduced rounding errors)

```
>> t_betahat = rs.tstat.t
t_betahat =
    1.7806
   16.9604
```

The R^2-value is easily found by $R^2 = 1 - \dfrac{\mathrm{SS_E}}{\mathrm{SS_{TOT}}}$ but MATLABhas handidly done the calculation for us

```
>> R2 = 1 - rs.fstat.sse/rs.fstat.ssr
R2 =
    0.9583
>> R2 = rs.rsquare
R2 =
    0.9600
```

Note also that we have the p-values for testing if $\beta_i = 0$ already calculated:

```
>> format long
>> P_betahat = rs.tstat.pval
P_betahat =
   0.100280455850894
   0.000000000946812
```

To find a prediction interval, we'll employ that if $\boldsymbol{u} = (1\ 4.5)$ and Y_0 is a future measurment at $x = 4.5$, then

$$T = \frac{Y_0 - \widehat{\mu}_0}{S\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} \sim t(12),$$

where $\widehat{\mu}_0 = \boldsymbol{u}^T\widehat{\boldsymbol{\beta}}$. Letting $t_{0.025}(12) = 2.1788$ we find that

$$P(|T| < t_{0.025}(12)) = 0.95.$$

Solving the inequality in the probability measure and using the proper point estimates, we obtain that

$$I_{Y_0} = \Big(\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} - t_{\alpha/2}(n-k-1)s\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}},$$
$$\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} + t_{\alpha/2}(n-k-1)s\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}\Big).$$

With numbers:

```
>> u = [1 4.5]';
>> k = tinv(0.975,12) * sqrt(s2 * (1 + u' * XtXinv * u));
>> CI = [ u' * betahat - k   u'*betahat + k]
CI =
    5.6233   16.6718
```

3 Example 2: multiple linear regression

Exempel

The following data has been measured.

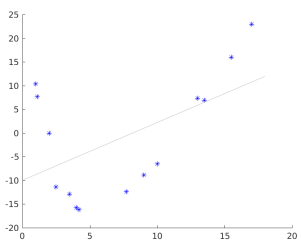
x_1	x_2	x_3	y
1.0000	4.0000	1.0000	10.3331
1.1000	5.0000	2.0000	7.6873
2.0000	4.5000	5.0000	-0.0825
2.5000	6.0000	8.0000	-11.4330
3.5000	6.1000	10.0000	-12.9390
4.0000	6.2000	11.0000	-15.7947
4.2000	6.5000	12.0000	-16.1596
7.7000	9.0000	15.0000	-12.3880
9.0000	10.0000	16.0000	-8.8999
10.0000	10.5000	16.5000	-6.5445
13.0000	11.0000	17.0000	7.3035
13.5000	14.0000	17.5000	6.8997
15.5000	16.0000	18.0000	15.9900
17.0000	18.0000	18.2000	22.9394

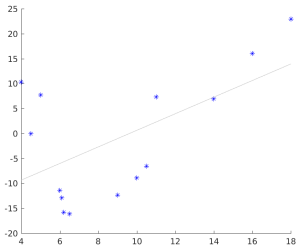
Assuming the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$, where $\epsilon \sim \mathrm{N}(0, \sigma^2)$, and that y are independent observations, perform the regression. With significance level 5%, can we claim that $\beta_i \neq 0$? Remove those variables where $\beta_i = 0$ is reasonable and perform a new regression. Is this model significantly worse (at 5%) than the full model?

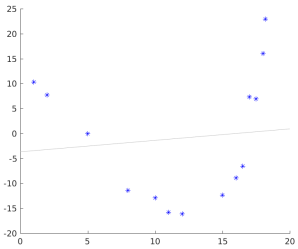
Solution. Let's start by looking at the correlations:

```
>> corr([x1 x2 x3 y])
ans =
    1.0000    0.9793    0.8977    0.5341
    0.9793    1.0000    0.8507    0.5765
    0.8977    0.8507    1.0000    0.1078
    0.5341    0.5765    0.1078    1.0000
```

We see that x_1 and x_2 are correlated with y but that x_3 has a lower value. It wouldn't be unreasonable to expect that x_3 might be removed. We also see that the variables are strongly correlated, so we might need to be careful. A few scatterplots of x_i versus y.







The full regression is performed by

```
>> rs = regstats(y,[x1 x2 x3],'linear','all');
>> betahat = rs.tstat.beta
```

```
betahat =
    9.9060
    5.2090
   -0.0735
   -4.0920
>> Stdev_betahat = rs.tstat.se
Stdev_betahat =
    1.1914
    0.2589
    0.2741
    0.0930
>> t_betahat = rs.tstat.t
t_betahat =
    8.3148
   20.1186
   -0.2683
  -43.9810
>> P_betahat = rs.tstat.pval
P_betahat =
    0.0000
    0.0000
    0.7939
    0.0000
```

We can proceed by testing $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$ with $\alpha = 5\%$ using the information above. The easiest way is the p-value method. We see that only the p-value for β_2 is higher than the significante level 0.05. Hence only in the case of β_2 can we *not* reject H_0. Let's try the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon.$$

Note that these are *not* the same β_i as before.

```
>> rs2 = regstats(y,[x1 x3],'linear','all');
>> betahat2 = rs2.tstat.beta
betahat2 =
    9.6253
    5.1443
   -4.0840
>> P_betahat2 = rs2.tstat.pval
P_betahat2 =
   1.0e-08 *
    0.2046
    0.0000
    0.0000
```

All the p-values are below the significance level, so we reject $H_0 : \beta_i = 0$ in all cases. Is this model significantly worse than the previous full model? We can perform an f-test using the fact that this reduced model is nested inside the full model, so (using the full model),

$$H_0 : \beta_2 = 0,$$

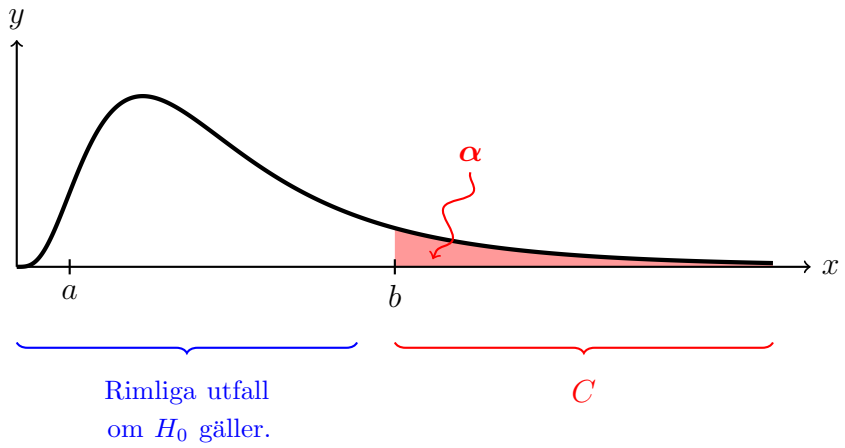
with the alternative hypothesis

$$H_1 : \beta_2 \neq 0.$$

If H_0 is true, then

$$W = \frac{(\mathrm{SS}_{\mathrm{E}}^{(2)} - \mathrm{SS}_{\mathrm{E}}^{(1)})/1}{\mathrm{SS}_{\mathrm{E}}^{(1)}/10} \sim F(1, 10) \quad \text{if } H_0 \text{ is true,}$$

where $\mathrm{SS}_{\mathrm{E}}^{(1)}$ is for the full model and $\mathrm{SS}_{\mathrm{E}}^{(2)}$ is for the reduced model. If H_0 is not true, when W will tend to be big, so our critical domain is given by $C =]c, \infty[$ for some $c > 0$. At 5%, we obtain that $c = 4.9646$.



From our numbers we obtain that

```
>> (rs2.fstat.sse - rs.fstat.sse)/(rs.fstat.sse/rs.fstat.dfe)
ans =
    0.0720
```

which is *not* in the critical domain, so we can't reject H_0. The reduced model might be as powerful as the full model.

4 Example 3: stepwise linear regression