

Matematisk statistik V
Föreläsningsanteckningar
Tommy Norberg
16 mars 2005

F4a: Ch 3 Bivariata datamängder och fördelningar

Nyckelord: skatter-plot, korrelationskoefficient, minsta-kvadrat-metoden, regressionslinje, dess lutning och intercept, residual- och total-kvadratsumma, standardavvikelse från regressionslinjen, mass- och täthetsfunktion för flera variabler, oberoende

Vi ska nu titta på vilka numeriska storheter som kan vara av intresse då datamängden är bivariat: (x_i, y_i) , $1 \leq i \leq n$.

3.1 Skatter-plottar (s 98-101)

I figuren nedan visas en simulering av bivariata normalfördelade data x, y med väntevärden

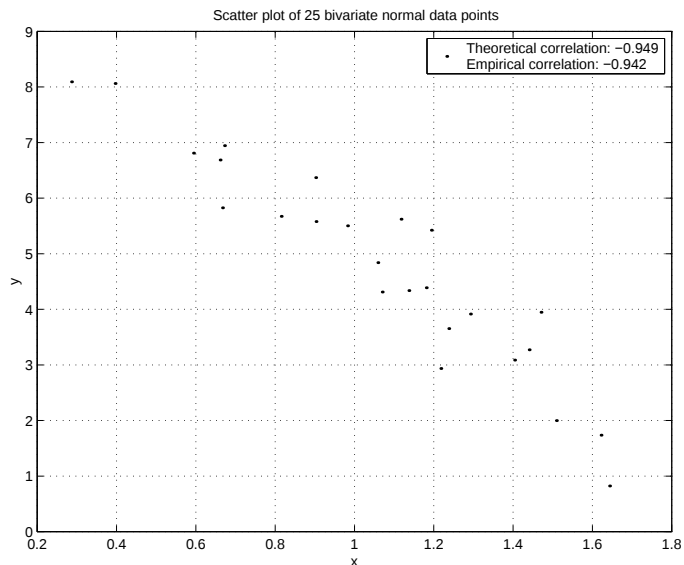
$$\mu_x = 1, \quad \mu_y = 5$$

och standardavvikelser

$$\sigma_x = 0.3, \quad \sigma_y = 1.58$$

samt kovariansen

$$\text{cov}_{xy} = -0.45$$



Figur 1: En skatter-plott med simulerade bivariata data med stark negativ korrelation.

Om man med ögat försöker att anpassa en rät linje till punkterna i figuren, så bör man få något i stil med $y = 10 - 5x$.

3.2 Korrelation (s 103-111)

Definition 1 (s 106) Pearsons *korrelationskoefficient* är

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

där

$$S_{xx} = \sum_i (x_i - \bar{x})^2$$

$$S_{yy} = \sum_i (y_i - \bar{y})^2$$

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Observera att x -observationernas varians är

$$s^2 = \frac{1}{n-1} S_{xx}$$

Vi har redan sett att

$$(n-1)s^2 = \sum_i x_i^2 - n\bar{x}^2$$

Vi har alltså att

$$S_{xx} = \sum_i x_i^2 - n\bar{x}^2$$

Analogt kan man visa att

$$S_{yy} = \sum_i y_i^2 - n\bar{y}^2$$

och att

$$S_{xy} = \sum_i x_i y_i - n\bar{x}\bar{y}$$

Observera att

1. r beror ej av sorten
2. r är symmetrisk i x och y
3. $-1 \leq r \leq 1$
4. r^2 är ett mått på graden av linjärt samband
5. $r^2 = 1 \Leftrightarrow$ punkterna ligger på en rät linje

3.6 Joint distributions (s 144-149)

Handlar mest om två-dimensionella (bivariata) fördelningar.

Definition 2a (s 144) En bivariat massfunktion $p(x, y)$ ska uppfylla

1. $p(x, y) \geq 0 \forall x, y$
2. $\sum_{x,y} p(x, y) = 1$

Definition 2b En bivariat täthet $f(x, y)$ ska uppfylla

1. $f(x, y) \geq 0 \forall x, y$

$$2. \int_R \int_R f(x, y) dy dx = 1$$

Teoretiska proportioner (sannolikheter) beräknas medelst summering i det diskreta fallet och integrering i det kontinuerliga fallet. (Analogt med den univariata (en-dimensionella) teorin).

Exempel 3.19 (s 144-145) A certain market has both an express checkout register and a superexpress register. Let x denote the number of customers queueing at the express register at a particular weekday time and let y denote the number of customers in the line at the superexpress register at that same time. The joint probability mass function $p(x, y)$ is given in the following table.

		y			
		0	1	2	3
x	0	.08	.07	.04	.00
	1	.06	.15	.05	.04
	2	.05	.04	.10	.06
	3	.00	.01	.05	.06
	4	.00	.01	.05	.06

Kalla motsvarande stokastiska variabler för X och Y . Då

$$p(x, y) = P(X = x, Y = y)$$

Beräkna

1. $P(X = Y)$
2. $P(X + Y = 2)$
3. $p_X(1) = P(X = 1)$

Räkna ut

$$P(X = Y) = p(0, 0) + p(1, 1) + p(2, 2) + p(3, 3) = 0.39$$

Räkna ut

$$P(X + Y = 2) = p(2, 0) + p(1, 1) + p(0, 2) = 0.24$$

Räkna ut

$$p_X(1) = P(X = 1) = p(1, 0) + p(1, 1) + p(1, 2) + p(1, 3) = 0.30$$

Generalisera till marginalfördelningarna

$$p_X(x) = P(X = x) = p(x, 0) + p(x, 1) + p(x, 2) + p(x, 3)$$

för $x = 0, 1, 2, 3, 4$ och

$$p_Y(y) = P(Y = y) = p(0, y) + p(1, y) + p(2, y) + p(3, y) + p(4, y)$$

för $y = 0, 1, 2, 3$.

Allmänt gäller i det diskreta fallet att de s k marginalfördelningarna (d.v.s fördelningarna för marginalerna X resp Y) definieras och beräknas enligt

$$p_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y p(x, y)$$

samt

$$p_Y(y) = P(Y = y) = \sum_x P(X = x, Y = y) = \sum_x p(x, y)$$

där $p(x, y)$ är den bivariata (två-dimensionella) massfunktionen.

I det kontinuerliga fallet gäller analogt

$$f_X(x) = \int_R f(x, y) dy$$

och

$$f_Y(y) = \int_R f(x, y) dx$$

där $f(x, y)$ är den bivariata (två-dimensionella) tätheten.

Definition 3 (s 146) Kovariansen mellan x och y är i det kontinuerliga fallet

$$\text{cov}_{xy} = \int_R \int_R (x - \mu_x)(y - \mu_y) f(x, y) dy dx$$

Den normerade kovariansen

$$\rho = \frac{\text{cov}_{xy}}{\sigma_x \sigma_y}$$

kallas för *korrelationskoefficienten*.

Korrelationskoefficienten ρ är den teoretiska motsvarigheten till Pearsons korrelationskoefficient r och den uppfyller liknande egenskaper.

Notera att

$$\begin{aligned}
 \text{cov}_{xy} &= \int_R \int_R (x - \mu_x)(y - \mu_y) f(x, y) dy dx \\
 &= \int_R \int_R (xy - \mu_x y - x \mu_y + \mu_x \mu_y) f(x, y) dy dx \\
 &= \int_R \int_R xy f(x, y) dy dx - \mu_x \mu_y \\
 &= \mu_{xy} - \mu_x \mu_y
 \end{aligned}$$

Produkten xy är ju en variabel om x och y är det. Man kan visa att

$$\mu_{xy} = \int_R \int_R xy f(x, y) dy dx$$

Rent allmänt gäller att bokens beteckningar är dåliga. Bäst är att hela tiden införa stokastiska variabler och beteckna dem med stora bokstäver X , Y , Z , etc. Istället för cov_{xy} skriver vi $\text{Cov}[X, Y]$ och vi noterar att dubbelintegralen ovan är ett väntevärde. Vi kan då skriva

$$\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])]$$

Bra att känna till, men inget som vi ska gå närmare in på i denna kurs är att väntevärdet är en linjär operator. Detta innebär att följande kalkyl är matematiskt korrekt:

$$\begin{aligned}
 \text{Cov}[X, Y] &= E[(X - E[X])(Y - E[Y])] \\
 &= E[XY - XE[Y] - YE[X] + E[X] \cdot E[Y]] \\
 &= E[XY] - E[Y] \cdot E[X] - E[X] \cdot E[Y] + E[E[X] \cdot E[Y]] \\
 &= E[XY] - E[X] \cdot E[Y]
 \end{aligned}$$

Jämför med uträkningen av cov_{xy} på OH 6.

Vi ser också att

$$\text{Cov}[X, X] = E[(X - E[X])^2] = \text{Var}[X]$$

Exempel 1 (s 146-147) Bivariata normalfördelningen är det kanske viktigaste exemplet på en flerdimensionell fördelning:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}$$

För parametern ρ gäller att $-1 < \rho < 1$, och man kan visa att om X och Y är bivariat normala med fördelning enl ovan, så är deras korrelationskoefficient lika med ρ .

Den bivariata normalfördelningen kan betecknas

$$N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$$

Korrelationskoefficienten ρ är ett mått på sambandet mellan X och Y . Notera att om $\rho = 0$, så faktoriseras $f(x, y)$ i två univariata tätheter, enligt

$$\begin{aligned} f(x, y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]} \\ &= \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2} \\ &= f_1(x)f_2(y) \end{aligned}$$

Man inser att $f_1(x)$ måste vara X :s täthet och analogt att $f_2(y)$ är Y :s täthet.

Oberoende (s 148)

Tillbaka till den allmänna teorin.

Ett viktigt specialfall uppkommer då täthets- eller massfunktionen kan faktoriseras i en faktor som bara innehåller x och en annan som bara innehåller y :

$$f(x, y) = f_1(x)f_2(y)$$

$$p(x, y) = p_1(x)p_2(y)$$

Detta är fallet då variablerna ej har något samband. Man säger då att de är *oberoende*.

Antag att vi studerar två variabler x och y med bivariat täthetsfunktion $f(x, y)$ och har fått reda på utfallet av den ena, säg x . Låt X och Y vara motsvarande stokastiska variabler. Känt är att $X = x$. Om variablerna är oberoende, så säger detta inget om Y 's utfall. Men om X och Y är beroende, så finns det information om Y 's utfall i faktumet att X fick utfallet x .

Om vi inte utnyttjar kunskapen att $X = x$ inträffat, så är Y 's marginalfördelning lika med

$$f_Y(y) = \int_R f(x, y) dx$$

enlig vad vi tidigare sagt. Se OH 5.

Om vi utnyttjar att $X = x$, så kan man räkna ut Y 's betingade täthet, givet att $X = x$, så här:

$$f_{Y|X=x}(y|x) = \frac{f(x, y)}{f_X(x)}$$

Kolla själva att oberoende innebär att den betingade tätheten ej beror av x och att då är

$$f_{Y|X=x}(y|x) = f_Y(y)$$

Exempel 2 (s 146-147) Vi ser att om X, Y är bivariat normalfördelade med korrelationskoefficient ρ , så gäller att X och Y är oberoende då $\rho = 0$.

Läs själva om fallet med fler än två variabler på sida 148-149.

Matematisk statistik V
 Föreläsningsanteckningar
 Tommy Norberg
 16 mars 2005

F4b: Ch 4 Obtaining data

Nyckelord: slumpmässigt stickprov, population, noggranhet, precision

Antag att vi i n st mätningar av variabeln x har erhållit observationerna

$$x_1, \dots, x_n$$

För att den statistiska analysen ska bli bra är det viktigt att observationerna är

- oberoende
- likafördelade

Oberoende betyder att resultatet från ett av försöken ej får lov att påverka något annat.

Likafördelade betyder helt enkelt att man observerar samma variabel x . På engelska uttrycker man ofta detta genom att säga att observationerna är från samma *population*.

Observera att n -tippeln $x_1, \dots, x_n$ i sig själv är en variabel, som kan observeras genom att vi upprepade (säg m) gånger observerar $x_1, \dots, x_n$:

försök nr	observa- tioner
1	$x_{11}, \dots, x_{1n}$
2	$x_{21}, \dots, x_{2n}$
$\vdots$	
m	$x_{m1}, \dots, x_{mn}$

Så vi kan tänka oss att $x_1, \dots, x_n$ har en (multivariat) täthet eller massfunktion $f(x_1, \dots, x_n)$. Att $x_1, \dots, x_n$ är oberoende betyder helt enkelt då att

$$f(x_1, \dots, x_n) = f_1(x_1) \cdot \dots \cdot f_n(x_n)$$

där $f_i(x_i)$ är den (marginal-) täthet eller massfunktion som beskriver x_i :s fördelning. Att de är likafördelade betyder att

$$f_1(x) = f_2(x) = \cdots = f_n(x)$$

och att de har samma fördelning som x betyder att

$$f_i(x) = f(x) \quad \text{för } i = 1, \dots, n$$

Således gäller om $x_1, \dots, x_n$ är oberoende och likafördelade observationer av variabeln x , att den multivariata tätheten som beskriver utfallen av n -tippeln $x_1, \dots, x_n$ är lika med produkten av de univariata, d.v.s.

$$f(x_1, \dots, x_n) = f(x_1) \cdot \dots \cdot f(x_n) = \prod_{i=1}^n f(x_i)$$

De statistiska analysmetoder vi ska lära oss förutsätter oberoende och likafördelade observationer. När man har sådana säger man ofta att det är ett (slumpmässigt) *stickprov* ("random sample" på engelska).

Antag nu att de n observationerna $x_1, \dots, x_n$ är oberoende och likafördelade mätningar av någon fysikalisk storhet θ . Kalla variabeln vi mäter för x och låt μ_x resp. σ_x vara x :s väntevärde och standardavvikelse.

Definition 4 (s 173) Mätmetoden säges vara mer eller mindre *noggrann* ("accurate") beroende på hur stor differensen $\theta - \mu_x$ är. Maximal noggrannhet har vi då $\theta = \mu_x$. Vi säger då att metoden är *väntevärdesriktig* ("unbiased").

Definition 5 (s 174) Mätmetoden säges vara mer eller mindre *precis* ("precise") beroende på hur stor standardavvikelsen σ_x är. Ju mindre σ_x är, desto precisare är mätmetoden.