Föreläsning 7: Punktskattningar

Matematisk statistik

David Bolin
Chalmers University of Technology
September 21, 2015

Tvådimensionella fördelningar

Definition

En två dimensionell slumpvariabel (X, Y) tillordnar två numeriska värden till varje utfall i Ω.

För diskreta variabler har vi sannolikhetsfunktionen

$$p_{X,Y}(i, j) = \mathsf{P}(X = i, Y = j).$$

Där $p_{X,Y}(i, j) \geq 0$ och $\sum_{i,j} p_{X,Y}(i, j) = 1$. För kontinuerliga variabler har vi en täthetsfunktion $f_{X,Y}(x, y)$ som är sådan att

1. $f_{X,Y}(x, y) \geq 0$,
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$, och
3. $\mathsf{P}(a \leq X \leq b \text{ och } c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y) dx dy$.

Marginalfördelningar

Givet en diskret tvådimensionell slumpvariabel (X, Y) definieras marginalfördelningarna för X och Y som

$$p_X(i) = \sum_{j=-\infty}^{\infty} p_{X,Y}(i,j)$$

$$p_Y(j) = \sum_{i=-\infty}^{\infty} p_{X,Y}(i,j)$$

I det kontinuerliga fallet definieras marginalfördelningarna för X och Y på motsvarande sätt som

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Väntevärden

Väntevärde

Givet en tvådimensionell slumpvariabel (X, Y) definieras väntevärdena för X och Y som

$$\mathsf{E}(X) = \begin{cases} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} i p_{X,Y}(i,j), & \text{i det diskreta fallet} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dx dy, & \text{i det kontinuerliga fallet} \end{cases}$$

och

$$\mathsf{E}(Y) = \begin{cases} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} j p_{X,Y}(i,j), & \text{i det diskreta fallet} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x,y) dx dy, & \text{i det kontinuerliga fallet} \end{cases}$$

Väntevärden

Notera att vi också kan formulera väntevärdena med marginalfördelningarna för X och Y om vi har beräknat dem först. Till exempel för det diskreta fallet har vi

$$\begin{aligned}\mathsf{E}(X) &= \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} i p_{X,Y}(i,j) = \sum_{i=-\infty}^{\infty} i \sum_{j=-\infty}^{\infty} p_{X,Y}(i,j) \\ &= \sum_{i=-\infty}^{\infty} i p_X(i)\end{aligned}$$

Väntevärden av funktioner

Väntevärdet av en funktion $H(X,Y)$ definieras som

$$\mathsf{E}(H(X,Y)) = \begin{cases} \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} H(i,j)p_{X,Y}(i,j), & \text{diskret} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)f_{X,Y}(x,y)dxdy, & \text{kontinuerlig} \end{cases}$$

Beroende

Oberoende slumpvariabler

Två slumpvariabler X och Y sägs vara oberoende om deras täthetsfunktion (eller sannolikhetsfunktion) kan skrivas som produkten av marginalfördelningarna. Det vill säga, för det diskreta fallet

$$p_{X,Y}(i,j) = p_X(i)p_Y(j)$$

och för det kontinuerliga fallet

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

Betingade fördelningar

Den betingade fördelningen för X givet $Y = y$ definieras för det kontinuerliga fallet som

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

givet att $f_Y(y) > 0$. På motsvarande sätt har vi den betingade fördelningen för Y givet $X = x$:

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

givet att $f_X(x) > 0$.

Samma formler gäller för diskreta variabler om vi byter f mot p.

Kovarians

Kovarians

Kovariansen mellan X och Y definieras som
$\mathsf{C}(X,Y) = \mathsf{E}[(X-\mu_X)(Y-\mu_Y)]$ där $\mu_X = \mathsf{E}(X)$ och $\mu_Y = \mathsf{E}(Y)$.

Enligt definitionen kan vi skriva kovariansen som

$$\mathsf{C}(X,Y) = \begin{cases} \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty}(i-\mu_X)(j-\mu_Y)p_{X,Y}(i,i), & \text{diskret,} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f_{X,Y}(x,y), & \text{kontinuerlig,} \end{cases}$$

Notera att vi har att $\mathsf{C}(X,X) = \mathsf{V}(X)$.

Korrelation och oberoende

Sats

Kovariansen kan beräknas som

$$\mathsf{C}(X, Y) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y)$$

Sats

Om X och Y är oberoende så är $\mathsf{C}(X, Y) = 0$ och $\mathsf{E}(XY) = \mathsf{E}(X)\mathsf{E}(Y)$.

Varians av summor

För två slumpvariabler X och Y, och två tal a och b har vi

$$\mathsf{V}(aX + bY) = a^2\mathsf{V}(X) + b^2\mathsf{V}(Y) + 2ab\mathsf{C}(X, Y)$$

Korrelation och oberoende

Korrelation

Den så kallade korrelationskoefficienten definieras som

$$\rho(X,Y) = \frac{\mathsf{C}(X,Y)}{\sqrt{\mathsf{V}(X)\mathsf{V}(Y)}}.$$

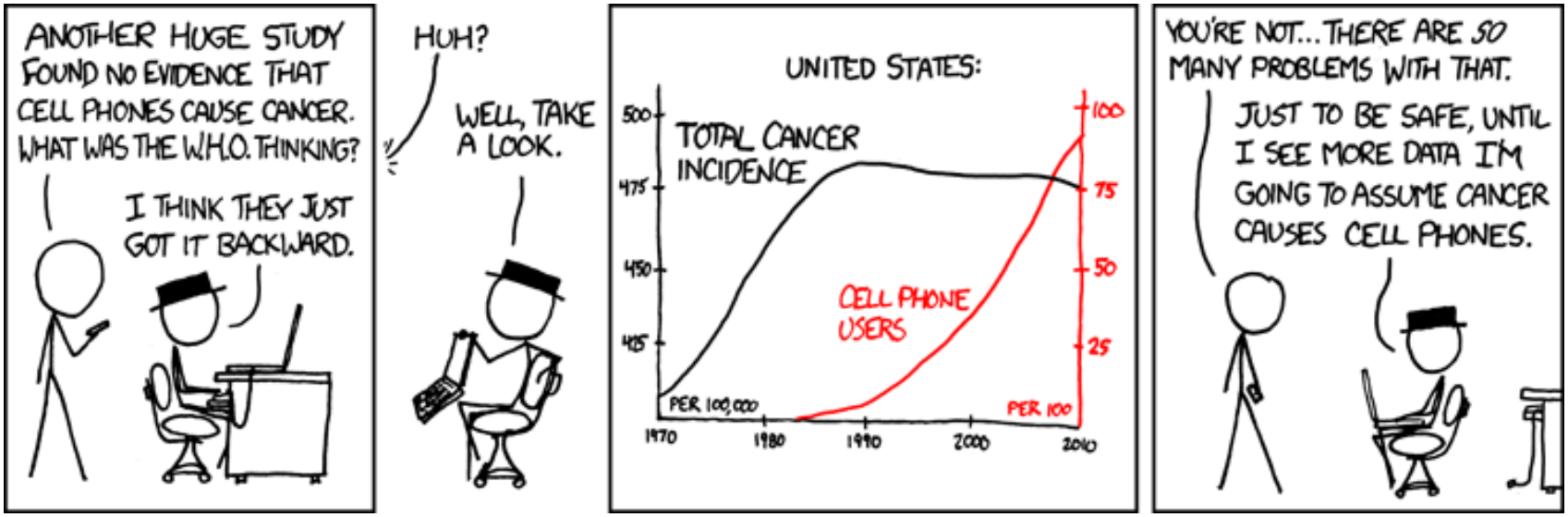
- Detta mått beskriver linjär samvariation mellan X och Y.
- Vi har att $-1 \leq \rho \leq 1$.
- Vi säger att X och Y är okorrelerade om $\rho(X,Y) = 0$.

Vi har följande samband mellan beroende och korrelation:

- Om X och Y är oberoende så är de okorrelerade.
- Om X och Y är okorrelerade så behöver de inte vara oberoende.

Dessa samband är naturliga eftersom två slumpvariabler är oberoende om det inte finns någon samvariation alls mellan dem, medan de är okorrelerade om det saknas *linjär* samvariation.

Korrelation och kausatilitet (I)



Korrelation beskriver linjärt beroende mellan två variabler.[1]

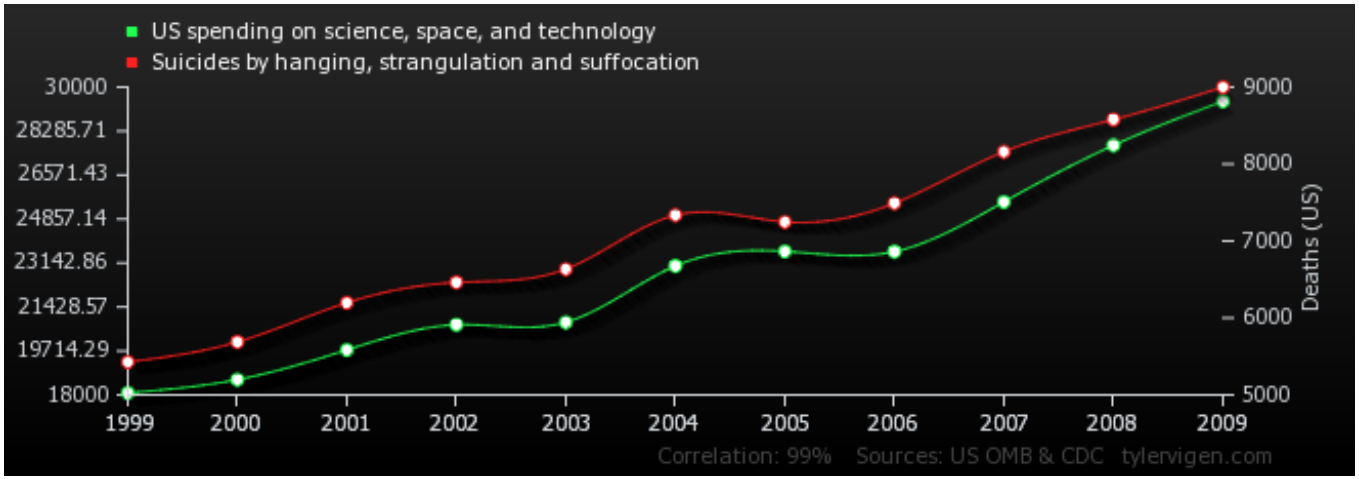
Korrelation säger inget om orsakssamband (kausalitet)!

(men telefoner $\Rightarrow$ cancer kanske är troligare än cancer $\Rightarrow$ telefoner)

Vi kan också ha att all samvariation kan förklaras av en tredje icke uppmätt variabel.

[1] Serier från `xkcd.com`

Korrelation och kausatilitet (II)



US spending on science, space, and technology

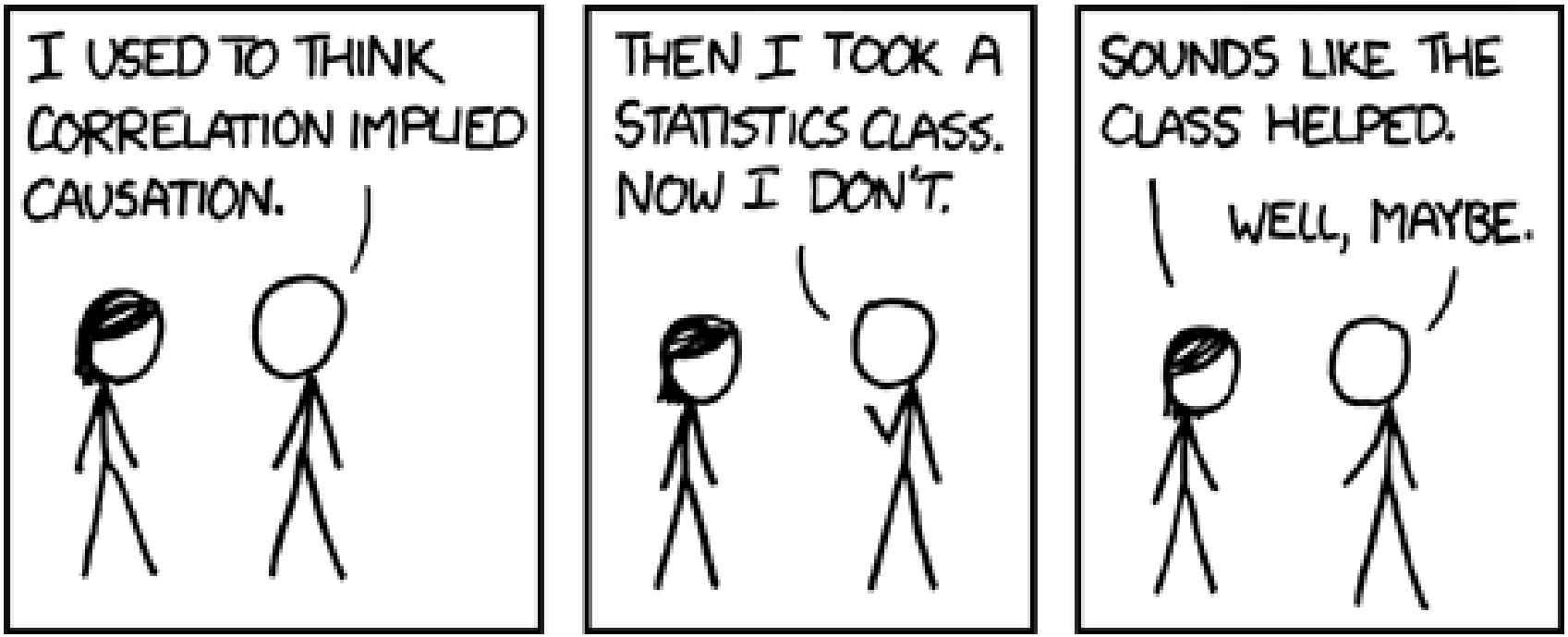
correlates with

Suicides by hanging, strangulation and suffocation[2]

Correlation: 0.992082

[2]Källa: `http://www.tylervigen.com/view_correlation?id=1597`

Korrelation och kausatilitet (III)



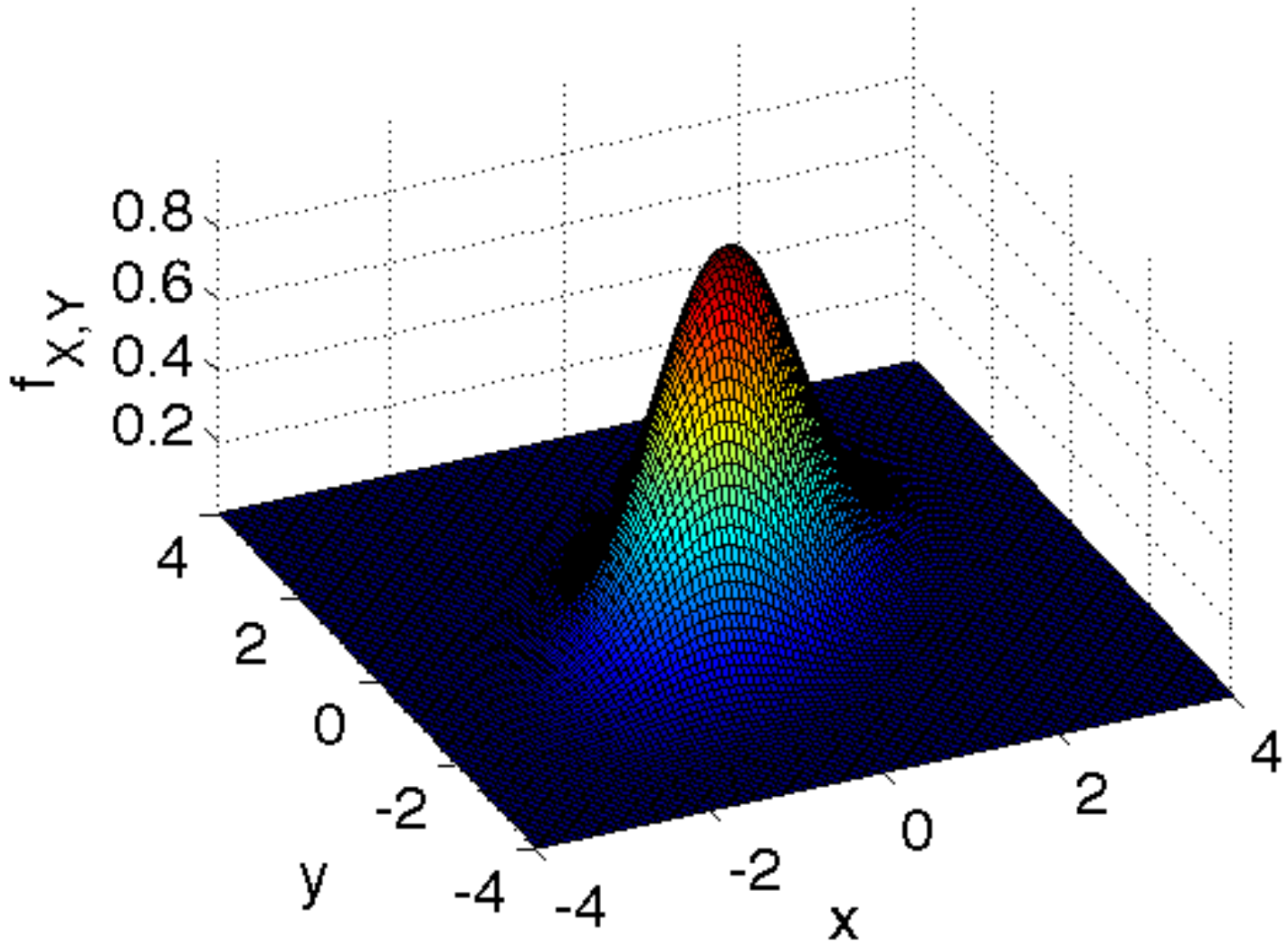
Den tvådimensionella normalfördelningen

En vanlig modell för en kontinuerlig tvådimensionell slumpvariabel är den tvådimensionella normalfördelningen.

(X, Y) sägs ha en tvådimensionell normalfördelning om den simultana täthetsfunktionen är

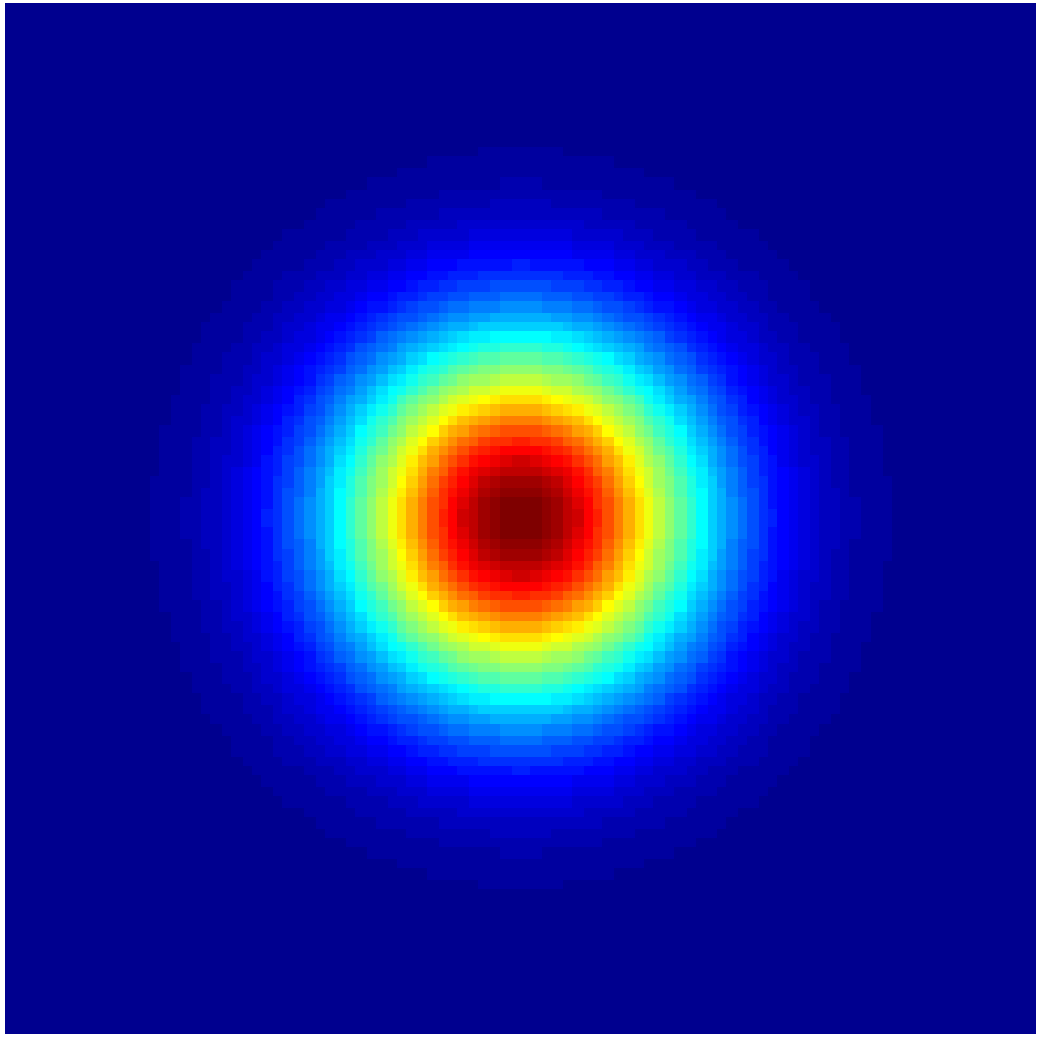
$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot$$
$$\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right)$$

här är (μ_x, μ_y) och (σ_x, σ_y) väntevärdena och standardavvikelserna för X och Y, och ρ är korrelationskoefficienten.

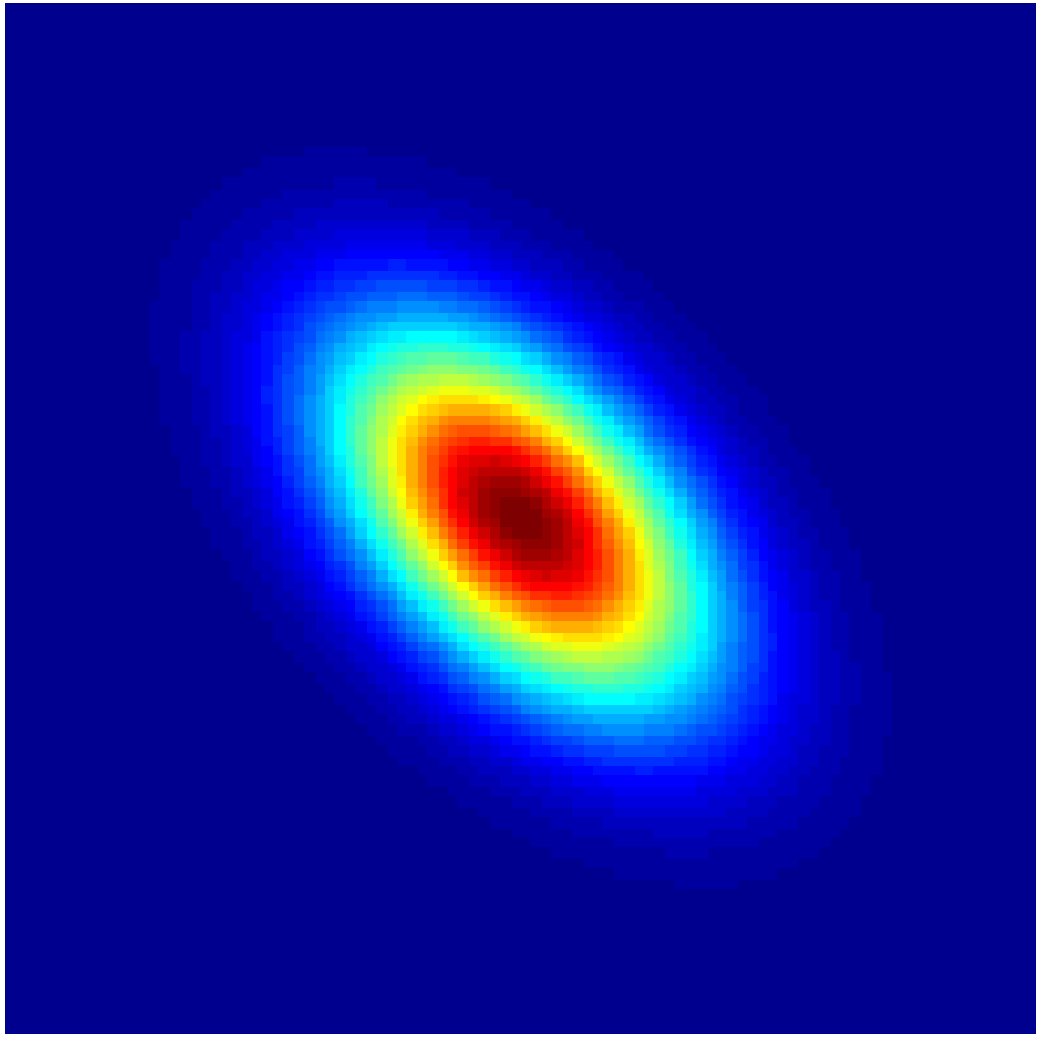


Ett exempel av fördelningen med $\rho = 0.5$.

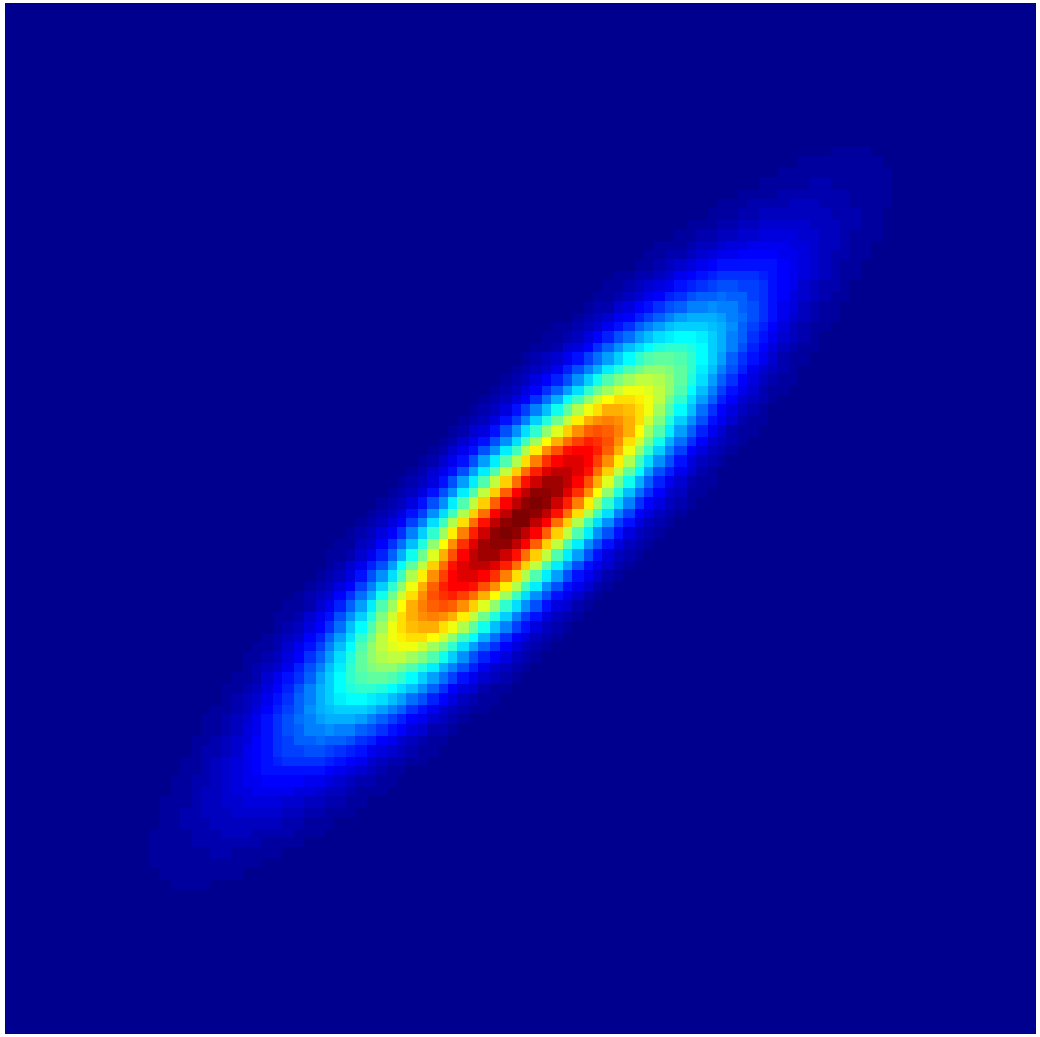
$\rho = 0$



$\rho = -0.5$



$\rho = 0.9$



Egenskaper

Om (X, Y) har en tvådimensionell normalfördelning är

- $X \sim \mathsf{N}(\mu_x, \sigma_x^2)$
- $Y \sim \mathsf{N}(\mu_y, \sigma_y^2)$.

Vidare kan den tvådimensionella normalfördelningen endast beskriva linjär samvariation: Om $\rho = 0$ så är X och Y oberoende.

Ett alternativt sätt att skriva täthetsfunktionen är att införa en väntevärdesvektor $\boldsymbol{\mu}$ en *kovariansmatris* Σ

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_y\rho \\ \sigma_x\sigma_y\rho & \sigma_y^2 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

Vi kan då skriva täthetsfunktionen som

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right),$$

Statistikteori

Statistikteori

Antag att vi har fått observationer $x_1, \ldots, x_n$ från en viss fördelning. Givet dessa observationer, vad kan vi säga om parametrarna i fördelningen?

Vi kommer titta på tre olika frågeställningar:

punktskattning Vad är ett troligt värde för parametrarna?

konfidensintervall Hur säker är denna skattning? För att karakterisera osäkerheten vill vi bilda ett intervall som med hög sannolikhet ska täcka det sanna värdet på parametern.

hypotesprövning Undersök hypoteser kring parametrana, till exempel kan vi vilja veta om $\mu > 0$ i en normalfördelning.

Populationer

Vi kommer utgå från att vi har ett stickprov av observationer $x_1, \ldots, x_n$ som vi vill använda för att få kunskap om en population.

Begreppet population kan innebära lite olika saker, vi kan skilja på två typfall:

- Populationen kan vara en väl avgränsad mängd saker, där vi i teorin räkna upp alla ingående element i en lista.
- Populationen kan också vara något lite mer abstrakt. Till exempel kan vi ha en process av vilken vi kan göra upprepade mätningar. Populationen här kan tänkas som alla möjliga slumpmässiga försök av denna typ, vilket inte har något konkret avgränsning.

Stickprov

Det är viktigt att vårt stickprov är *representativt* för populationen, vilket betyder att det är någorlunda typiskt för populationen. Om detta inte är fallet kan vi få systematiska fel i våra skattningar.

En matematisk definition av stickprov

Ett stickprov av storlek n är n oberoende observationer av en slumpvariabel X. Vi kan alltså skriva stickprovet som observationer av n slumpvariabler $X_1, \ldots, X_n$ där alla X_i är oberoende och likafördelade.

Denna definition täcker fallet med upprepade experiment ovan, och kan oftast ses som en bra approximation i fallet med ändlig population om n är litet i förhållande till storleken på populationen.