

Föreläsning 6: Sannolikhetsteori för flera variabler

Matematisk statistik

David Bolin
Chalmers University of Technology
April 3, 2014



Normalfördelningen

Normalfördelningen

En kontinuerlig slumpvariabel X är normalfördelad, $N(\mu, \sigma^2)$, med parametrar $\mu \in \mathbb{R}$ och $\sigma > 0$, om den har täthetsfunktionen

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

Fördelningsfunktionen ges av

$$F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right) dy$$

Parametrar

Om $X \sim N(\mu, \sigma^2)$ har vi att $E(X) = \mu$ och $V(X) = \sigma^2$.

Normalfördelningen

Ett viktigt specialfall är den standardiserade normalfördelningen.

Standardiserad normalfördelning

En slumpvariabel Z sägs ha en standardiserad normalfördelning om $Z \sim N(0, 1)$. Denna fördelning har täthetsfunktion

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

och fördelningsfunktion

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy.$$

Beräkning av sannolikheter

Sats

Om $X \sim N(\mu, \sigma^2)$ så gäller att $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

Detta betyder att om $X \sim N(\mu, \sigma^2)$

- kan vi skriva $X = \mu + \sigma Z$ där $Z \sim N(0, 1)$.
- har vi att $Z = (X - \mu)/\sigma \sim N(0, 1)$.

Vi kan använda detta för att beräkna sannolikheter:

$$P(X < x) = P\left(\frac{X - \mu}{\sigma} < \frac{x - \mu}{\sigma}\right) = P\left(Z < \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Centrala gränsvärdessatsen

Den viktigaste satsen inom statistikteorin, och en av anledningarna till varför normalfördelningen är så viktig.

CGS

Låt $X_1, \dots, X_n$ vara oberoende och likafördelade slumpvariabler med väntevärde μ och varians $\sigma^2 < \infty$. Då gäller att

$$P\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) \rightarrow \Phi(x), \quad \text{då } n \rightarrow \infty.$$

Om n är stort har vi enligt satsen att

- $\sum_{i=1}^n X_i$ är approximativt $N(n\mu, n\sigma^2)$ -fördelad.
- $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ är approximativt $N(\mu, \sigma^2/n)$ -fördelad.

Hur stor n måste vara beror på fördelningen av X_i .

Sannolikhetsteori för flera variabler

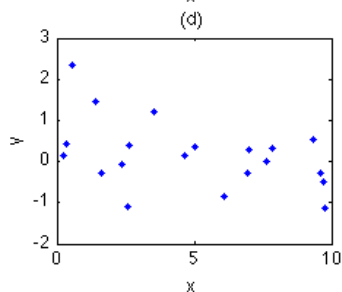
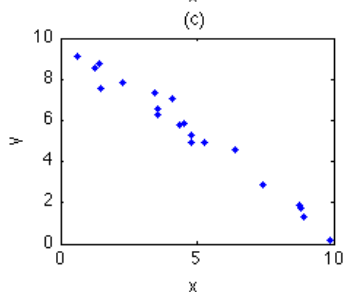
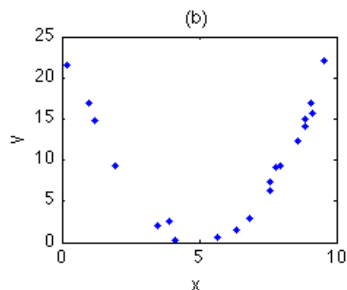
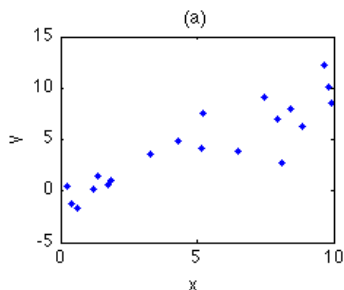
Ibland vill man studera två slumpvariabler X och Y samtidigt. Ett exempel på två situationer av detta slag kan vara att vi

- ① Mäter funktionsavvikelse X på n enheter och färgavvikelse Y på n andra enheter.
- ② Mäter funktionsavvikelse X och färgavvikelse Y simultant på n enheter.

I det första fallet får vi information om de två slumpvariablerna separat men i det andra fallet får man också kunskap om samvariationen mellan de två variablerna eftersom vi kan undersöka om avvikelserna tenderar att uppkomma samtidigt.

I båda fallen kan vi skriva upp den tvådimensionella slumpvariabeln (X, Y) .

Exempel på spridningsdiagram



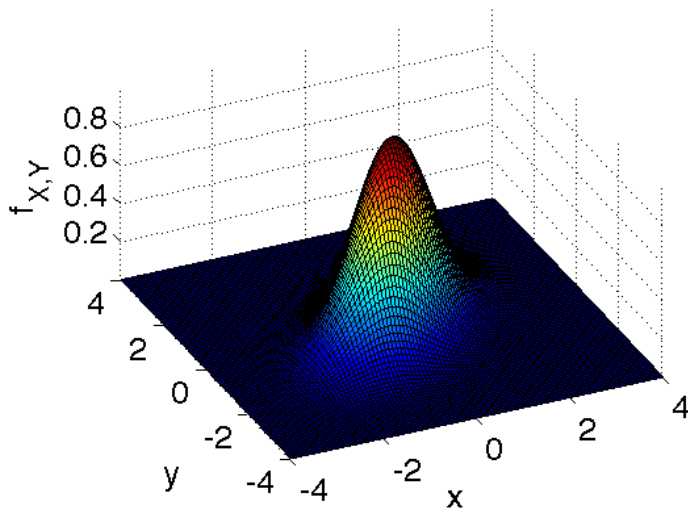
Den tvådimensionella normalfördelningen

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

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

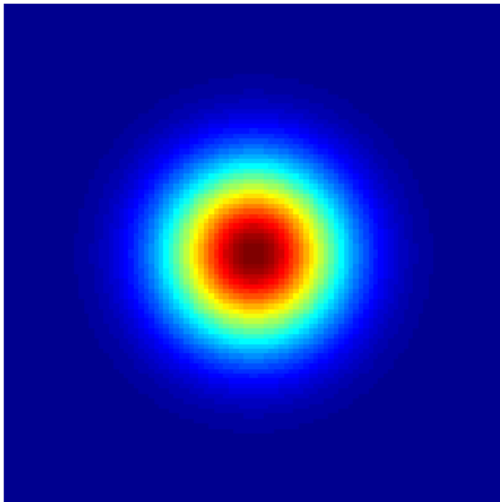
$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right)$$

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

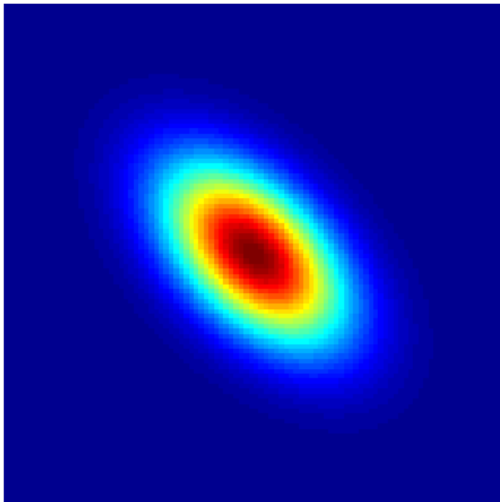


Ett exempel av fördelningen med $\rho = 0.5$.

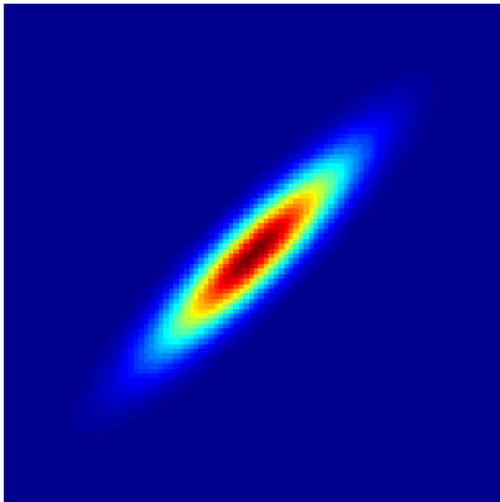
$$\rho = 0$$



$$\rho = -0.5$$



$$\rho = 0.9$$



Vi har alltså att om (X, Y) har en tvådimensionell normalfördelning så är

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

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

En alternativ formulering

Ett alternativt sätt att skriva täthetsfunktionen är att införa en väntevärdesvektor $\boldsymbol{\mu}$ en *kovariansmatris* Σ och en vektor som innehåller x och y som

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

Vi kan då skriva täthetsfunktionen som

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

där $d = 2$ och $|\Sigma|$ är determinanten av Σ . Om vi har d variabler kan vi använda denna formulering för att beskriva en d -dimensionell normalfördelning.