Föreläsning 6: Sannolikhetsteori för flera variabler

Matematisk statistik

David Bolin
Chalmers University of Technology
September 16, 2015

Normalfördelningen

Normalfördelningen

En kontinuerlig slumpvariabel X är normalfördelad, $\mathsf{N}(\mu, \sigma^2)$, 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 $X \sim \mathsf{N}(\mu, \sigma^2)$ har vi att $\mathsf{E}(X) = \mu$ och $\mathsf{V}(X) = \sigma^2$.

Normalfördelningen

Ett viktigt specialfall är den standardiserade normalfördelningen.

Standardiserad normalfördelning

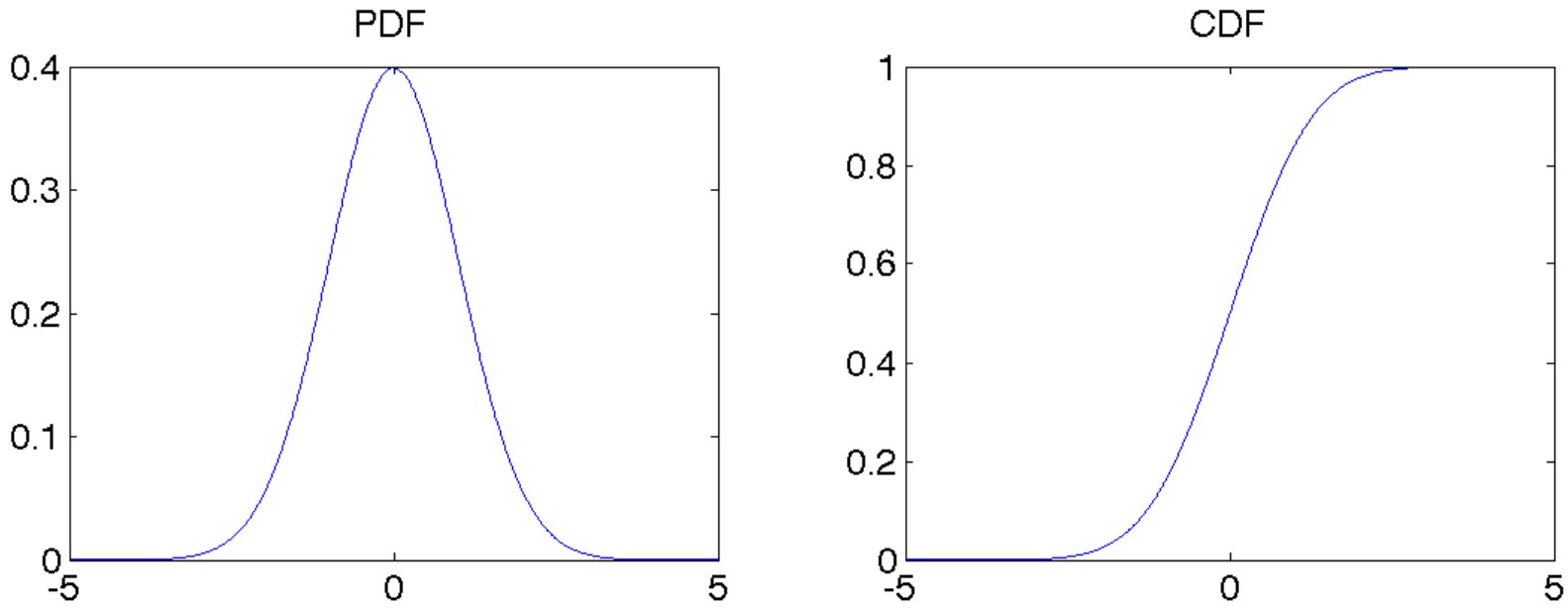
En slumpvariabel Z sägs ha en standardiserad normalfördelning om $Z \sim \mathrm{N}(0,1)$. Denna fördelning har täthetsfunktion

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

och fördelningsfunktion

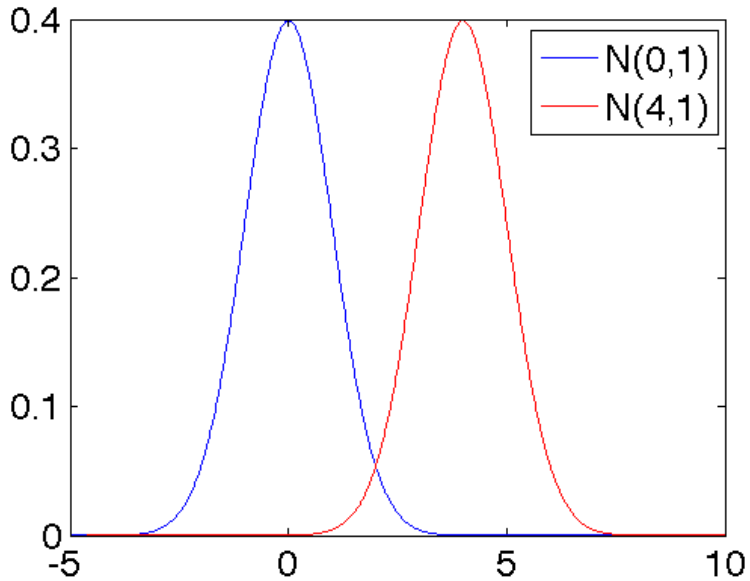
$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy.$$

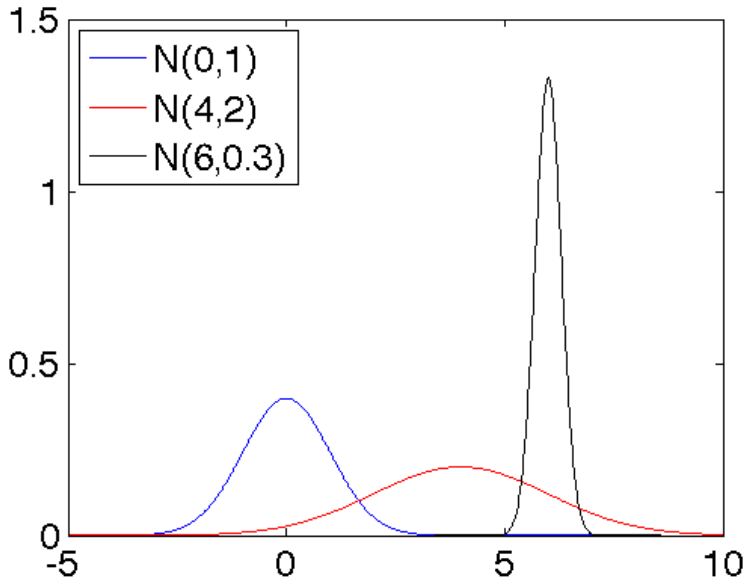
Standardiserad normalfördelning



Täthetsfunktion (vänster) och fördelningsfunktion (höger) till den standardiserade normalfördelningen.

Normalfördelningen med olika parametrar





Täthetsfunktioner för normalfördelningar med olika parametrar.

Beräkning av sannolikheter

Sats

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

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

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

Vi kan använda detta för att beräkna sannolikheter:

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

Centrala gränsvärdessatsen

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

På grund av egenskaperna hos väntevärde och varians har vi då

$$\mathsf{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{E}(X_i) = n\mu$$

$$\mathsf{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{V}(X_i) = n\sigma^2$$

Centrala gränsvärdessatsen säger dessutom att fördelningen för summan kommer vara approximativt normalfördelad om n är stor!

Centrala gränsvärdessatsen

CGS

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

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

Om n är stor har vi enligt satsen att

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

Hur stor n måste vara beror på fördelningen av X_i.

CGS: Exempel

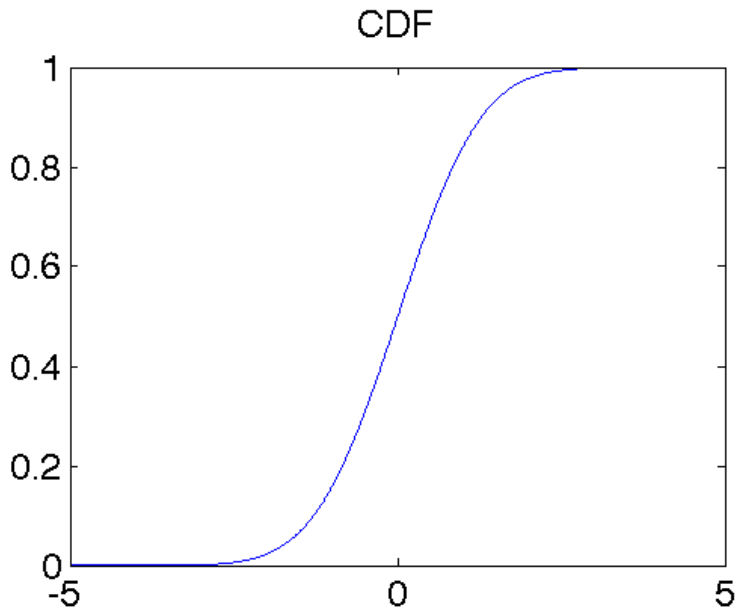
Kom ihåg att en binomialfördelad slumpvariabel $X \sim \text{Bin}(n, p)$ kan ses som en summa av n Bernoullifördelade slumpvariabler: $X = \sum_{i=1}^{n} X_i$ där $X_i \sim Be(p)$.

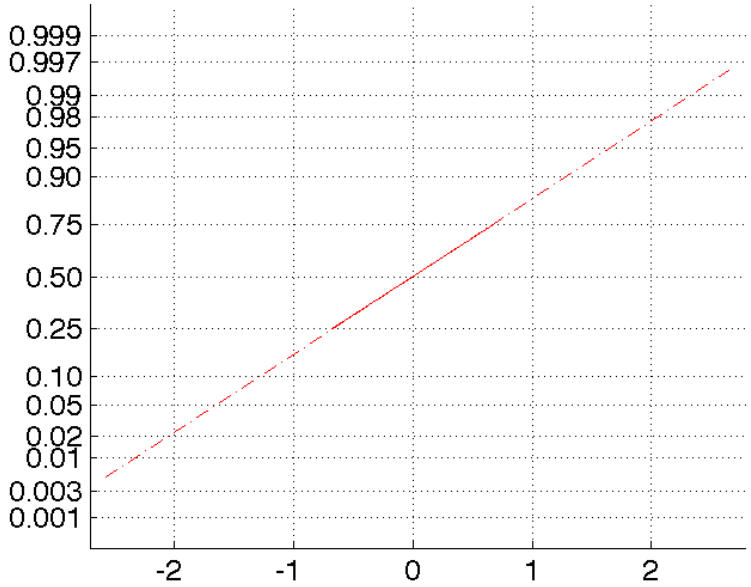
Enligt CGS får vi därför att X för stora n är approximativt normalfördelad med väntevärde np och varians $np(1-p)$.

Man brukar säga att normalapproximation är lämplig om

- $p \leq 0.5$ och $np > 5$ eller
- $p > 0.5$ och $n(1-p) > 5$.

Normalfördelningsdiagram





Som bekant kan fördelningsfunktionen för en normalfördelning skrivas som

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

Om vi plottar $F(x)$ är det möjligt att transformera skalan på y-axeln så funktionen blir en rät linje. Detta illustras i Figuren ovan.

Normalfördelningsdiagram

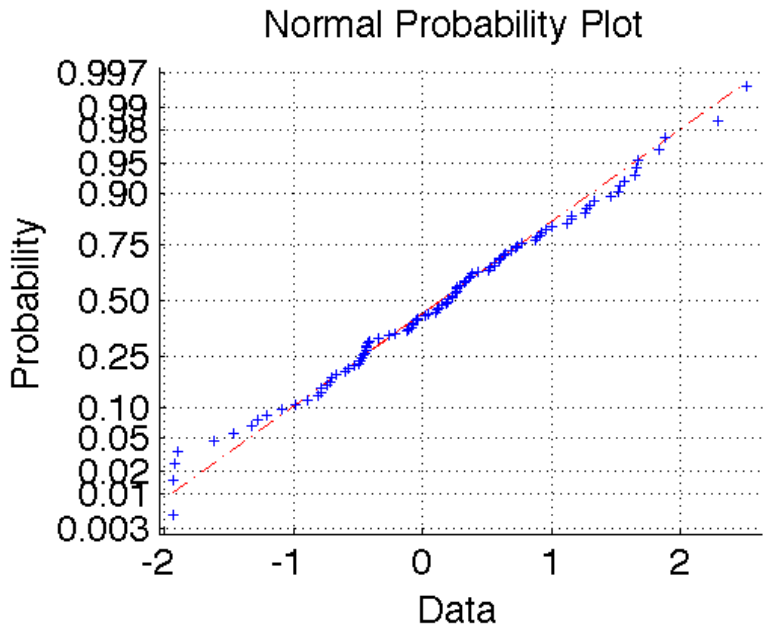
Antag att vi har data $x_1, \ldots, x_n$ och vill veta om en normalfördelning är en rimlig modell för datan. Vi kan använda normalfördelningsdiagrammet för detta.

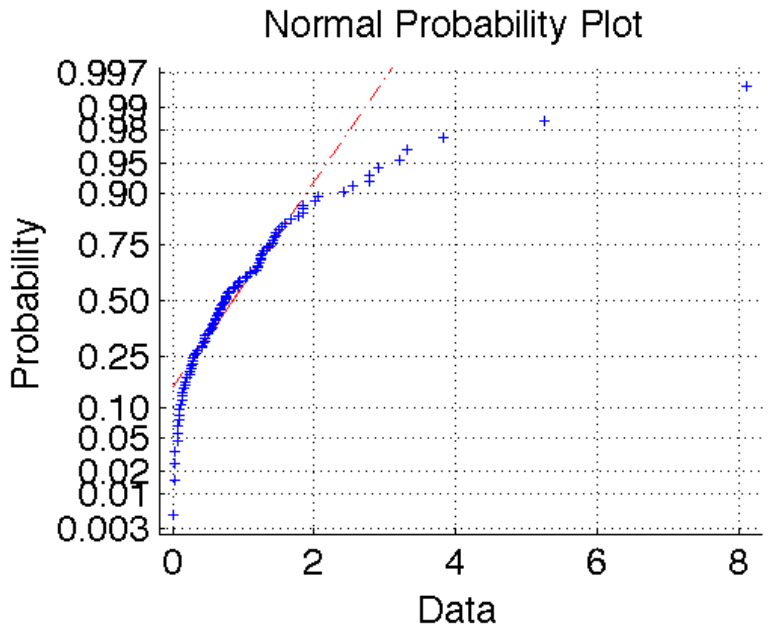
Vi börjar med att beräkna den *empiriska fördelningsfunktionen*

$$F^*(x) = \frac{\text{antal värden} \leq x}{n} = \sum_{i=1}^{n} \mathbb{I}(x_i \leq x)$$

Vi plottar sedan punkterna $F^*(x_j)$ i ett normalfördelningsdiagram, och om datan är normalfördelad ska dessa punkter ligga längs en rät linje.

Normalfördelningsdiagram





Två exempel där vi plottar normalfördelad och exponentialfördelad data i normalfördelningsdiagram. Detta görs enkelt i Matlab med kommandot `normplot`.

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

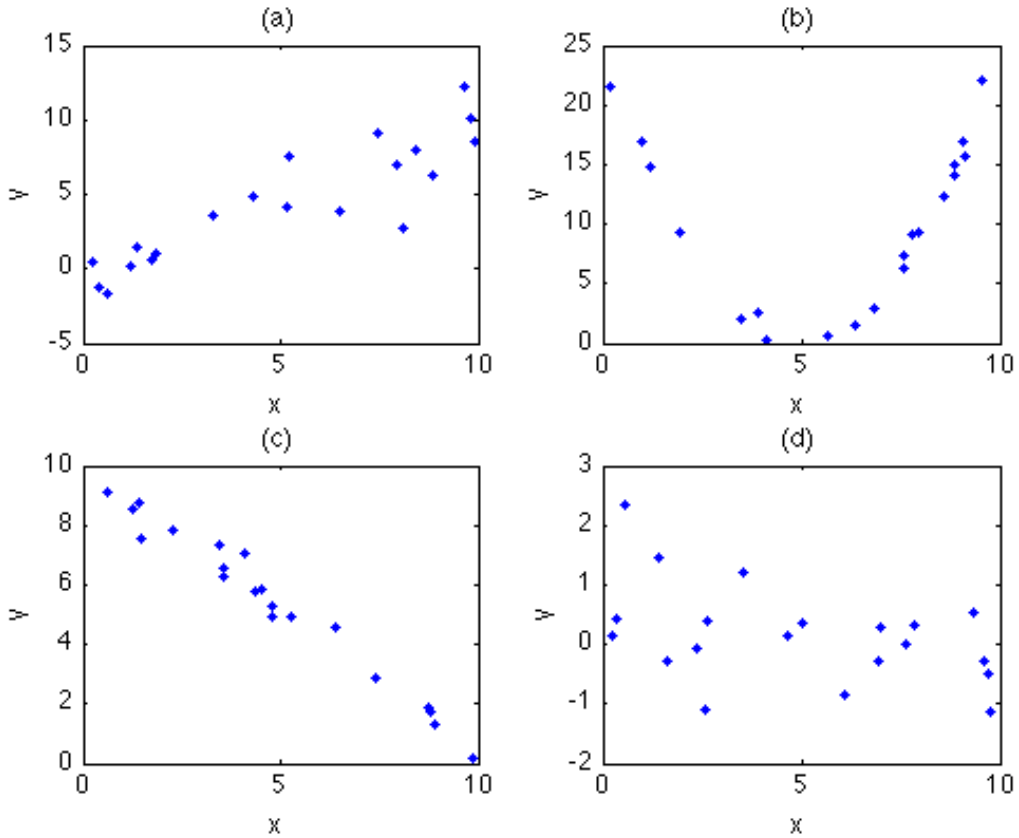
1. Mäter funktionsavvikelse X på n enheter och färgavvikelse Y på n andra enheter.
2. Mäter funktionsavvikelse X och färgavvikelse Y simultant på n enheter.

I det första fallet får vi information om de två slumpvariablerna separat

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).

Beskrivning av tvådimensionell data



Antag att vi har mätningar (x_i, y_i). Ett spridningsdiagram är ett tvådimensionellt punktdiagram där varje mätning (x_i, y_i) ritas ut som en punkt i xy-planet.

Numerisk beskrivning av tvådimensionell data

För att ge numeriska mått på samvariation kan vi beräkna stickprovskovariansen, som definieras som

$$c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

eller stickprovskorrelationen som definieras som

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Stickprovskorrelationen ett mått på linjärt beroende.

I bilden ovan har vi $r_{xy} = 0.8067$ i (a), $r_{xy} = 0.2912$ i (b), $r_{xy} = -0.9884$ i (c), och $r_{xy} = 0.3640$ i (d).