

Föreläsning 5: Normalfördelning och CGS

Matematisk statistik

David Bolin
Chalmers University of Technology
March 31, 2014



Diskreta fördelningar

En diskret slumpvariabel antar ett ändligt eller uppräkneligt oändligt antal värden, i allmänhet någon delmängd av heltalen.

Exempel

Kasta en tärning en gång och låt X mäta antalet prickar. Vi har då att $P(X = k) = 1/6$ för $k = 1, 2, 3, 4, 5, 6$ och $P(X = k) = 0$ för alla andra värden på k . Denna fördelning brukar kallas för en diskret likformig fördelning.

Diskreta slumpvariabler brukar beskrivas med den så kallade sannolikhetsfunktionen.

Sannolikhetsfunktion

Till en diskret stokastisk variabel X definierar vi sannolikhetsfunktionen $p(k)$ genom $p(k) = P(X = k)$.

Kontinuerliga fördelningar

I praktiken stöter vi ofta på situationer då de möjliga värdena är alla värden i ett intervall, tex $[0, 1]$, eller är alla positiva reella tal.

Kontinuerlig slumpvariabel

En kontinuerlig slumpvariabel kan anta alla värden i något (eller några) intervall av reella tal och sannolikheten för att den antar varje specifikt värde är noll.

Kontinuerliga slumpvariabler kan beskrivas med täthetsfunktionen.

Täthetsfunktion

Låt X vara en kontinuerlig slumpvariabel, en funktion $f(x)$ så att

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)dx = 1, \quad P(a \leq X \leq b) = \int_a^b f(x)dx,$$

kallas en täthetsfunktion.

Fördelningsfunktioner

Ett annat vanligt sätt att beskriva en stokastisk variabel är att använda den så kallade fördelningsfunktionen.

Fördelningsfunktionen

Låt X vara en slumpvariabel. Dess fördelningsfunktion ges då av $F(x) = P(X \leq x)$. Om X är en diskret variabel har vi

$$F(x) = \sum_{k \leq x} p_X(k),$$

och om X är kontinuerlig har vi

$$F(x) = \int_{-\infty}^x f(y) dy.$$

Väntevärde

I beskrivande statistik används ofta sampelmedelvärdet och sampelvariansen för att beskriva datamängder. Motsvarande termer för stokastiska variabler är väntevärdet och variansen.

Väntevärde

Väntevärdet för en diskret fördelning definieras som

$$E(X) = \sum_{k=-\infty}^{\infty} kp(k).$$

På motsvarande sätt definieras väntevärdet för en kontinuerlig slumpvariabel som

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Väntevärdet är "tyngdpunkten" i fördelningen.

Väntevärdet av en funktion

Antag att vi känner fördelningen för en slumpvariabel X och vill veta väntevärdet av en funktion $g(X)$, till exempel väntevärdet av X^2 . Vi kan då använda följande sats

Sats

Vi har att

$$E(g(X)) = \begin{cases} \sum_{k=-\infty}^{\infty} g(k)p(k), & \text{om } X \text{ är diskret} \\ \int_{-\infty}^{\infty} g(x)f(x)dx, & \text{om } X \text{ är kontinuerlig.} \end{cases} \quad (1)$$

Varians och standardavvikelse

Varians

Variansen av en stokastisk variabel definieras som

$V(X) = E[(X - \mu)^2]$, där μ är väntevärdet av X .

Vi ser alltså att variansen definieras som väntevärdet av den kvadratavvikelsen av X från dess väntevärde. Vi beräknar variansen som

$$V(X) = \begin{cases} \sum_{k=-\infty}^{\infty} (k - \mu)^2 p_X(k), & \text{om } X \text{ är diskret} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx, & \text{om } X \text{ är kontinuerlig.} \end{cases}$$

Ett enklare sätt är ofta att beräkna variansen som

$$V(X) = E(X^2) - \mu^2$$

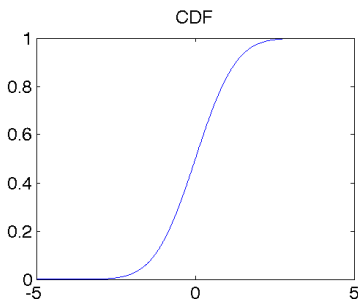
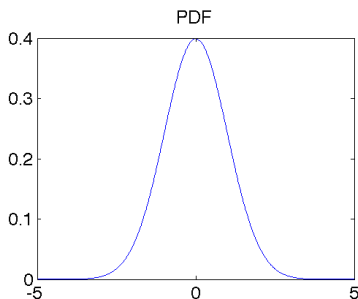
Standardavvikelsen av en stokastisk variabel ges av $\sigma = \sqrt{V(X)}$.

Räkneregler för väntevärde och varians

Låt X och Y vara två slumpvariabler och låt a och b vara två tal, vi har då att

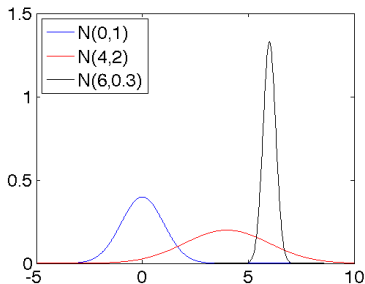
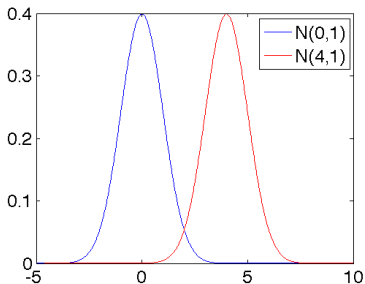
- $E(a) = a$.
- $E(aX) = aE(X)$.
- $E(aX + b) = aE(X) + b$.
- $V(a) = 0$.
- $V(aX) = a^2V(X)$.
- $V(aX + b) = a^2V(X)$.
- $E(X + Y) = E(X) + E(Y)$.
- $V(X + Y) = V(X) + V(Y)$ om X och Y är oberoende.

Standardiserad normalfördelning

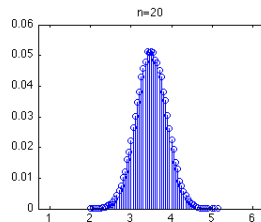
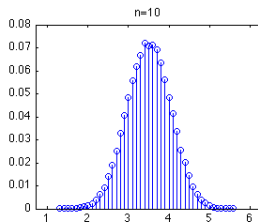
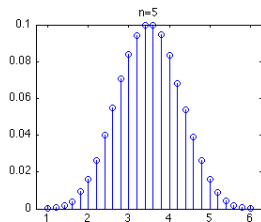
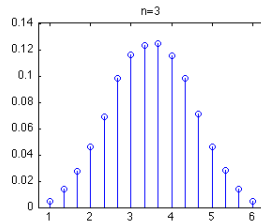
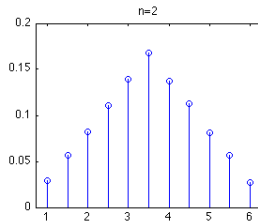
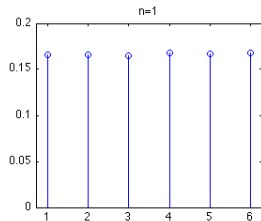


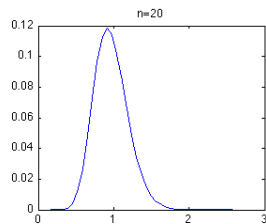
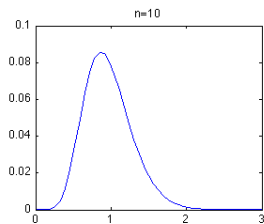
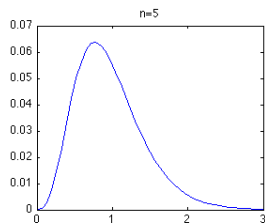
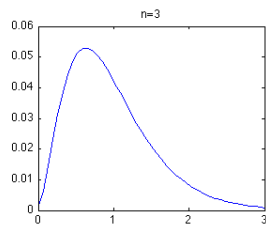
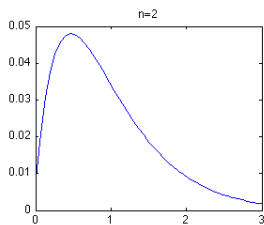
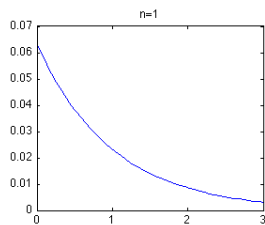
Täthetsfunktion (vänster) och fördelningsfunktion (höger) till den standardiserade normalfördelningen.

Normalfördelningen med olika parametrar

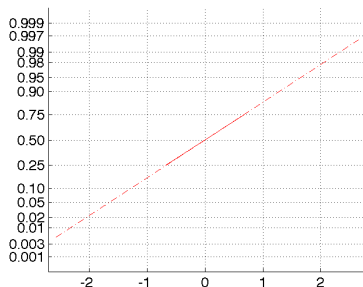
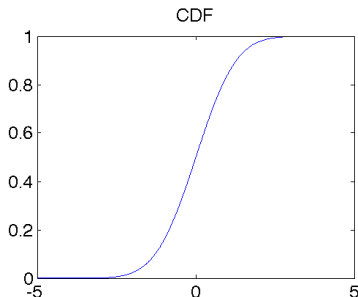


Täthetsfunktioner för normalfördelningar med olika parametrar.

CGS: Fördelningen för medelvärdet av n tärningskast

CGS: Fördelningen för medelvärdet av $n \exp(1)$ -variabler

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

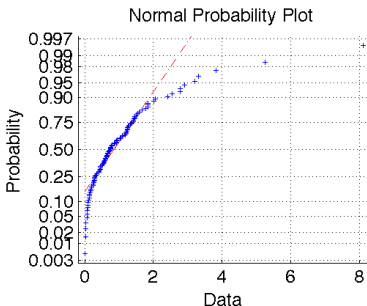
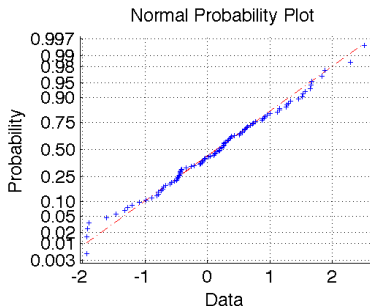
Antag att vi har data $x_1, \dots, x_n$ och vill veta om en normalfördelning är en rimlig modell för datan. Vi kan använda normalfördelningsdiagrammet för detta.

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`.