

Föreläsning 5: Normalfördelning och CGS

Matematisk statistik

David Bolin
Chalmers University of Technology
September 14, 2015



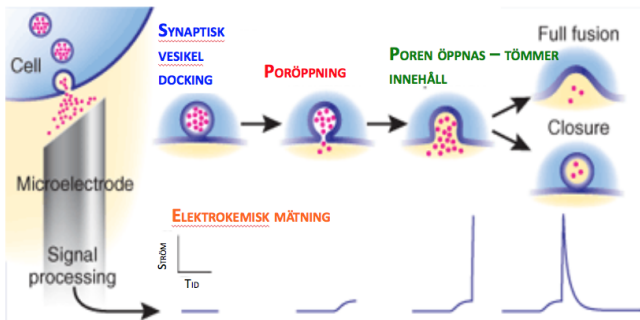
Projektuppgift



Projektet går ut på att studera frisättningen av dopamin hos nervceller och de två huvudsakliga frågeställningarna är hur processen påverkas av

- osmotiskt tryck och
- kolesterolhalt i cellmembranet.

Data och syfte



Syftet med uppgiften är att bland annat att ni ska träna på att

- använda Matlab för att tillämpa statistiska metoder på ett kemiskt forskningsproblem.
- konstruera och analysera statistiska modeller samt utföra kritisk granskning av modellerna.
- skriftligt redovisa statistiska undersökningar.

Projektarbete: praktiska saker (I)

- Ett godkänt projekt krävs för att bli godkänd på kursen men projektet påverkar inte graderingen 3,4 eller 5.
- Arbete kommer genomföras grupper om två till tre studenter.
- Indelning av grupper:
 - Skicka ett email till david.bolin@chalmers.se med namn och CID på medlemmarna i er grupp.
 - Jag svarar på mailet med ert gruppnummer. I projektet kommer ni sedan att arbeta med data specifik för er grupp.
- Projektet är indelat i fyra delar som motsvarar varsin datorövning.
- Varje datorövning kommer ha en allmän del när ni tränar på något moment i kursen och sedan tid för arbete med projektet.
- Huvuddelen av projektarbetet görs under datorövningarna, men troligen kommer mer tid krävas för att slutföra arbetet.

Projektarbete: praktiska saker (II)

- Projektet ska redovisas i form av en skriftlig rapport som lämnas in senast 2015-10-14.
- Den rättade rapporten lämnas tillbaks på Datorövning 5.
 - Eventuella mindre korrigeringar görs på plats under övningen.
 - Vid större korrigeringar som inte hinns med under övningen måste en uppdaterad rapport lämnas in senast 2015-10-30.
- Några riktlinjer kring rapporten:
 - Rapporten ska innehålla klara och tydliga formuleringar av frågeställningarna, modeller och antaganden.
 - Texten ska vara väl strukturerad och ska kunna läsas utan tillgång till vare sig kod eller projektbeskrivning.
- Mer detaljer och tips kring rapporteringen av projektet finns i Appendix till projektbeskrivningen.

Diskreta fördelningar

Diskreta slumpvariabler

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

Sannolikhetsfunktion

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

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.

Kontinuerliga fördelningar

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.

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.

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

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

Fördelningsfunktionen för vanliga fördelningar finns ofta i tabeller.

Väntevärde

Väntevärdet för en slumpvariabel definieras som

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

Väntevärdet är "tyngdpunkten" i fördelningen.

Sats

Vi har att

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

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

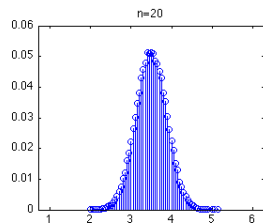
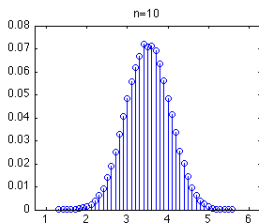
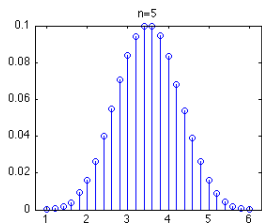
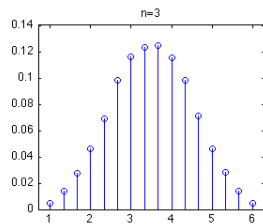
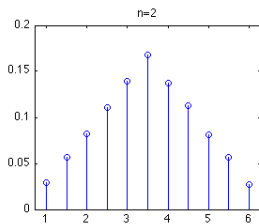
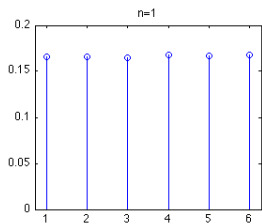
$$V(X) = E(X^2) - \mu^2$$

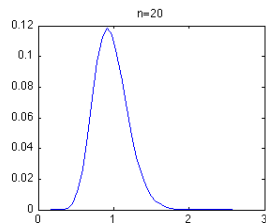
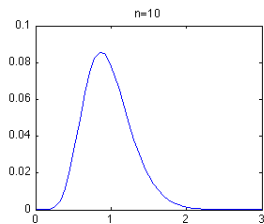
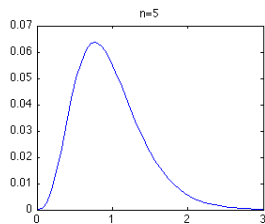
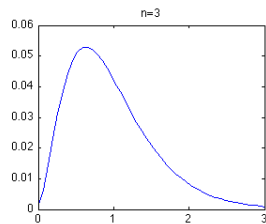
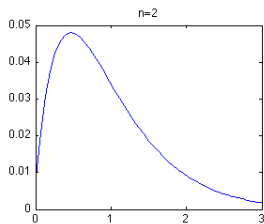
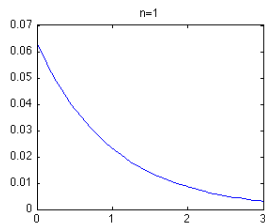
Standardavvikelsen av en stokastisk variabel ges av $\sigma = \sqrt{V(X)}$.

Räkningregler för väntevärde och varians

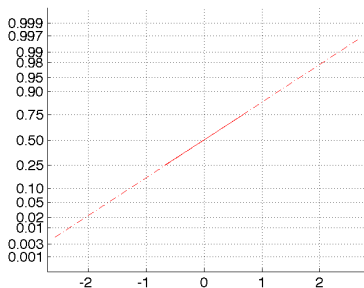
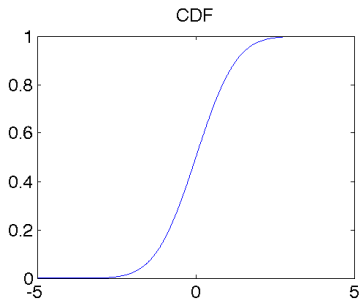
Låt X och Y vara två slumpvariabler och låt a och b vara två tal, vi har då att

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

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

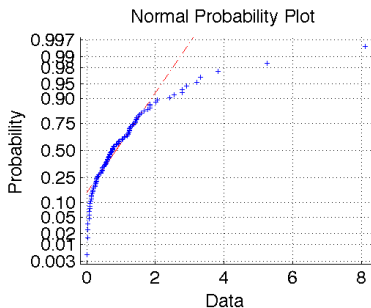
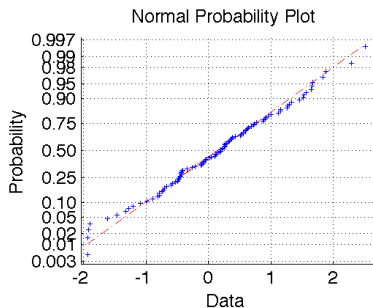
Antag att vi har data x_1, \dots, x_n och vill veta om en normalfördelning är en rimlig modell för datan. Vi kan använda normalfördelningsdiagrammet för detta.

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`.