Föreläsning 5: Normalfördelningen och CGS

Matematisk statistik

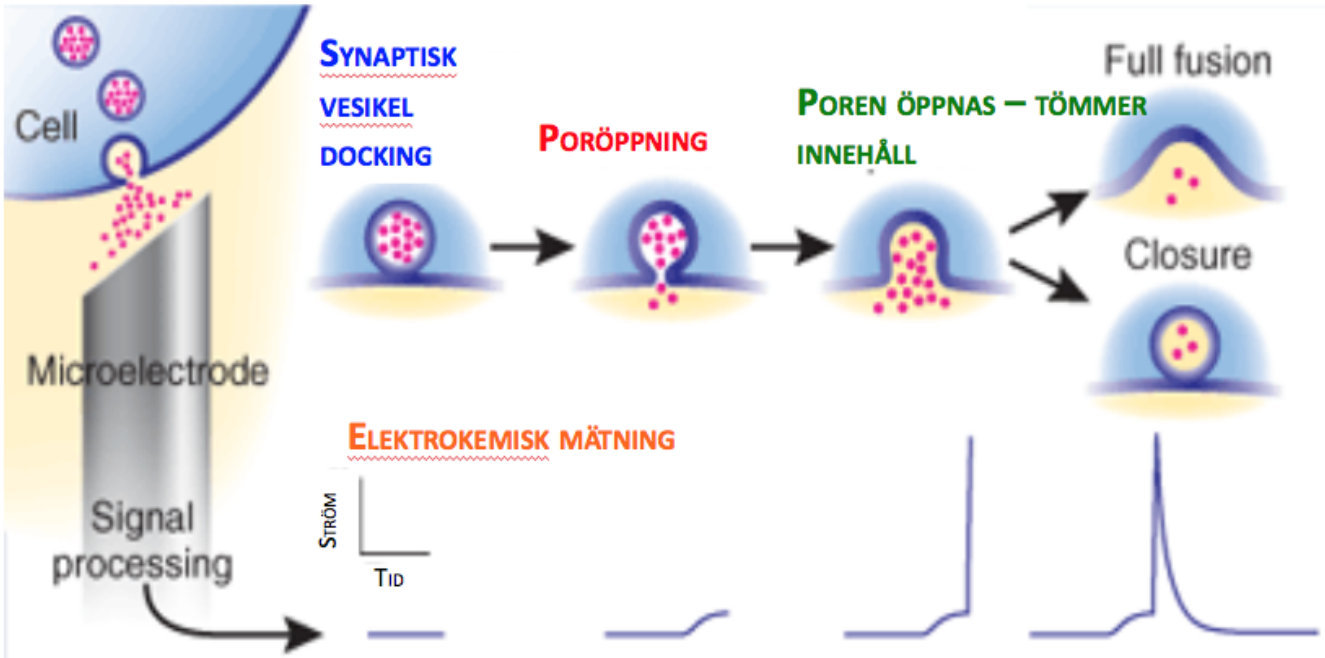
David Bolin
Chalmers University of Technology
September 11, 2017

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.
  - Anmäl er till en grupp i PING PONG.
  - Ladda ner mat-filen med data som hör till er grupp från kurshemsidan.
- 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 2017-10-11.
- 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 2017-10-27.
- 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.

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

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

kallas en täthetsfunktion.

Väntevärde

Väntevärde

Väntevärdet för en slumpvariabel definieras som

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

Väntevärdet är ”tyngdpunkten” i fördelningen.

Sats

Vi har att

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

Varians och standardavvikelse

Varians

Variansen av en stokastisk variabel definieras som $\mathsf{V}(X) = \mathsf{E}[(X-\mu)^2]$, 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

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

Ett enklare sätt är ofta att beräkna variansen som

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

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

Räkneregler för väntevärde och varians

Låt X och Y vara två slumpvariabler och låt a och b vara två tal.

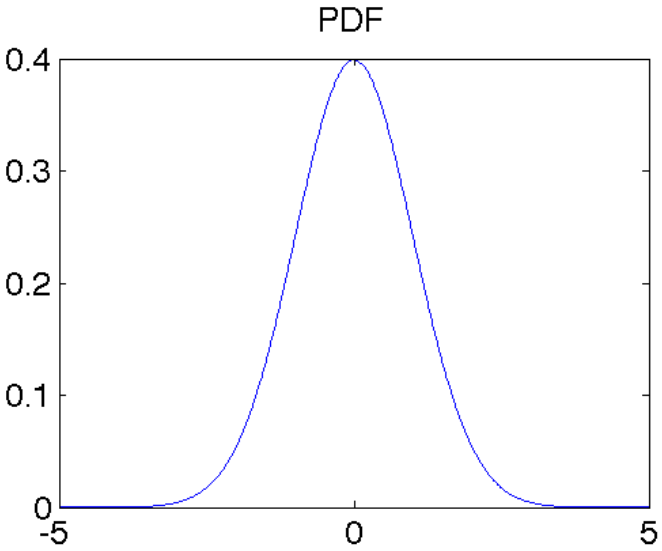
Väntevärdet uppfyller då

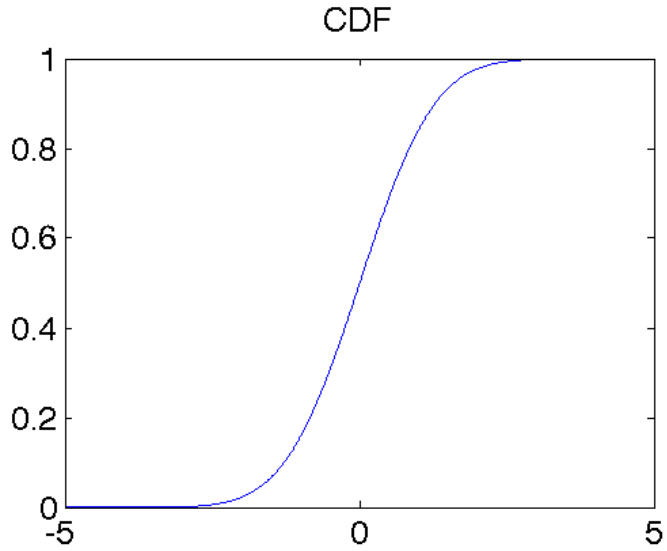
- $\mathsf{E}(a) = a$.
- $\mathsf{E}(aX) = a\mathsf{E}(X)$.
- $\mathsf{E}(aX + b) = a\mathsf{E}(X) + b$.
- $\mathsf{E}(X + Y) = \mathsf{E}(X) + \mathsf{E}(Y)$.

Variansen uppfyller istället

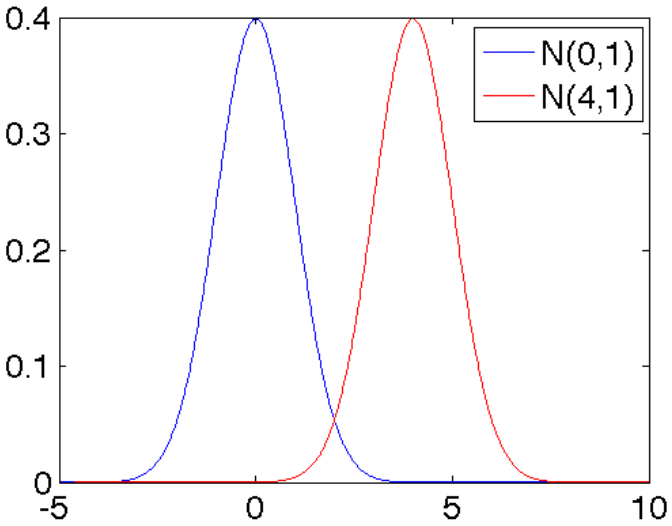
- $\mathsf{V}(a) = 0$.
- $\mathsf{V}(aX) = a^2\mathsf{V}(X)$.
- $\mathsf{V}(aX + b) = a^2\mathsf{V}(X)$.
- $\mathsf{V}(X + Y) = \mathsf{V}(X) + \mathsf{V}(Y)$ om X och Y är oberoende.

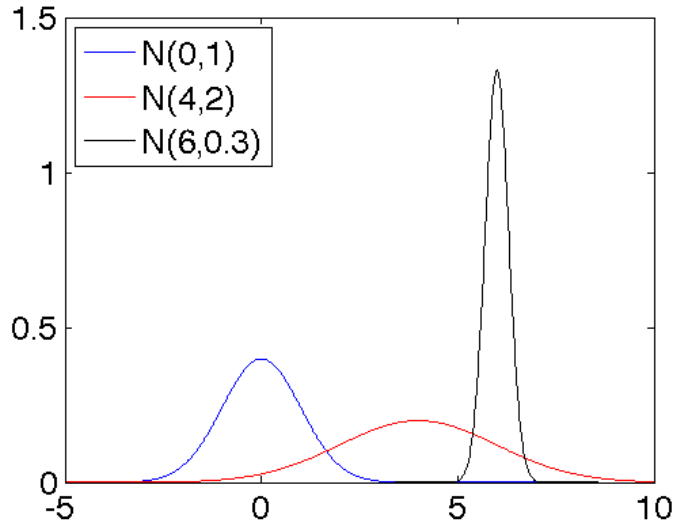
Täthetsfunktion och fördelningsfunktion till $Z \sim \mathrm{N}(0,1)$:





Normalfördelningen med olika parametrar:





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

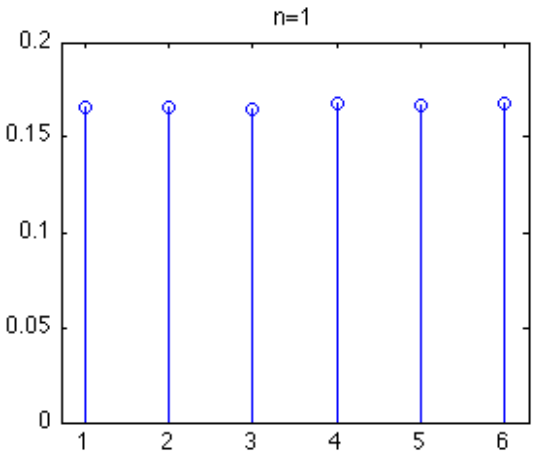
$$\mathsf{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) \to \Phi(x), \quad \text{då } n \to \infty.$$

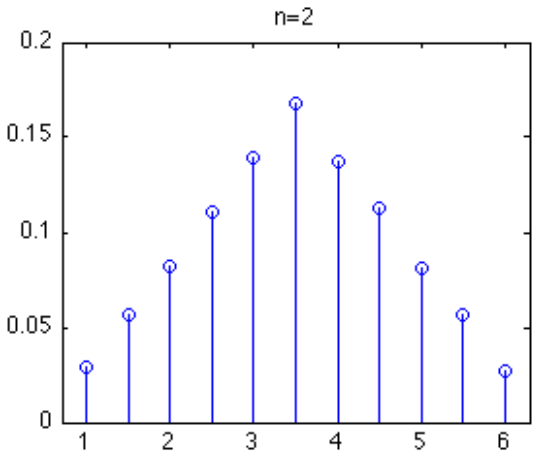
Om n är stor har vi enligt satsen att

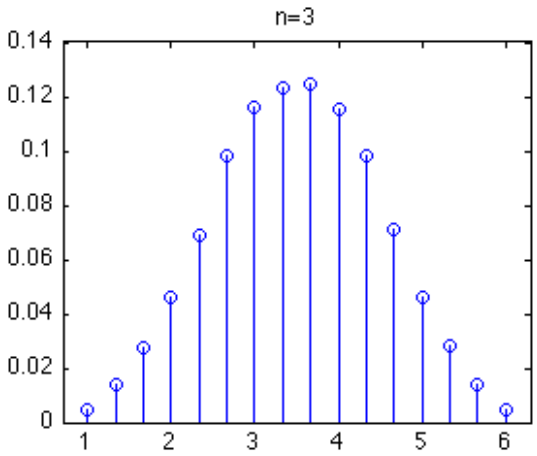
- $\sum_{i=1}^n X_i$ är approximativt $\mathsf{N}(n\mu, n\sigma^2)$-fördelad.
- $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ är approximativt $\mathsf{N}(\mu, \sigma^2/n)$-fördelad.

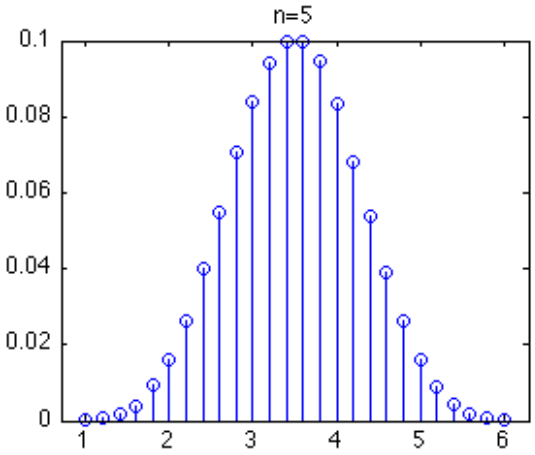
Hur stor n måste vara beror på fördelningen av X_i.

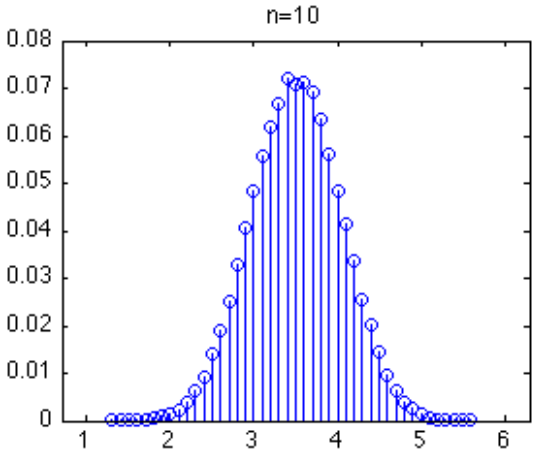
CGS: Fördelningen för medelvärdet av n tärningskast

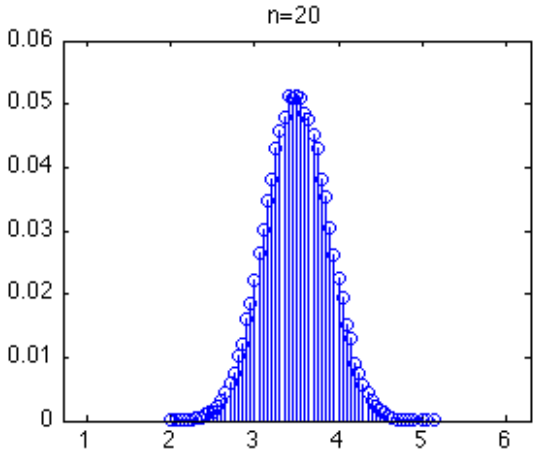












CGS: Binomialfördelningen

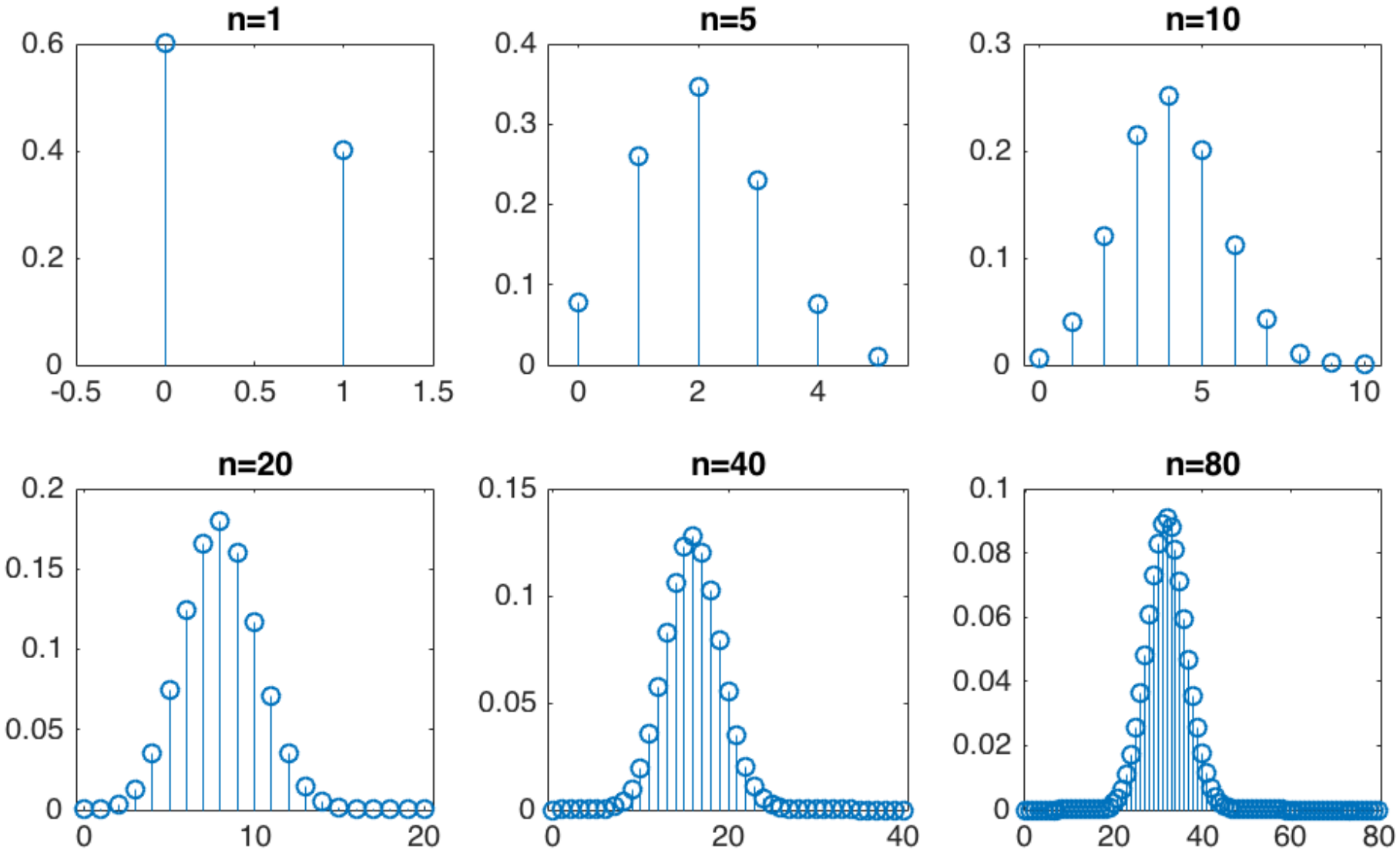
Kom ihåg att en binomialfördelad slumpvariabel $X \sim \text{Bin}(n, p)$ kan ses som en summa av n Bernoullifördelade slumpvariabler: $X = \sum_{i=1}^{n} X_i$ där $X_i \sim Be(p)$.

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 $n(1-p) > 5$.

CGS: Bin$(n, 0.4)$



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