Tema 2: Stokastiske variabler og sannsynlighetsfordelinger Kapittel 3

ST1101

2019-01-13 12:44 (Gunnar Taraldsen)

Det antas i notatet at S er et utfallsrom utstyrt med en sannsynlighet $P(A)$ for enhver hendelse $A \in \mathcal{F}$. $\mathcal{F}$ er en σ-algebra av delmengder. Det antas med andre ord at $(S, \mathcal{F}, P)$ er et underliggende sannsynlighetsrom som alle andre definisjoner baseres på. I forelesningene benyttes noen ganger symbolene $(\Omega, \mathcal{E}, P)$ for det underliggende sannsynlighetsrommet. Dette er et vanlig valg i mange lærebøker.

Stokastiske variabler

Definisjon: En stokastisk variabel X er en funksjon $X : S \to \mathbb{R}$ slik at $(X \leq x) = \{s | X(s) \leq x\}$ alltid er en hendelse. Den kumulative fordelingsfunksjonen F_X til X er definert ved $F_X(x) = P(X \leq x)$.

Diskret stokastisk variabel

- $X : S \to \mathbb{R}$
- Verdimengden $\mathcal{X} = X(S)$ er tellbar

Kontinuerlig stokastisk variabel

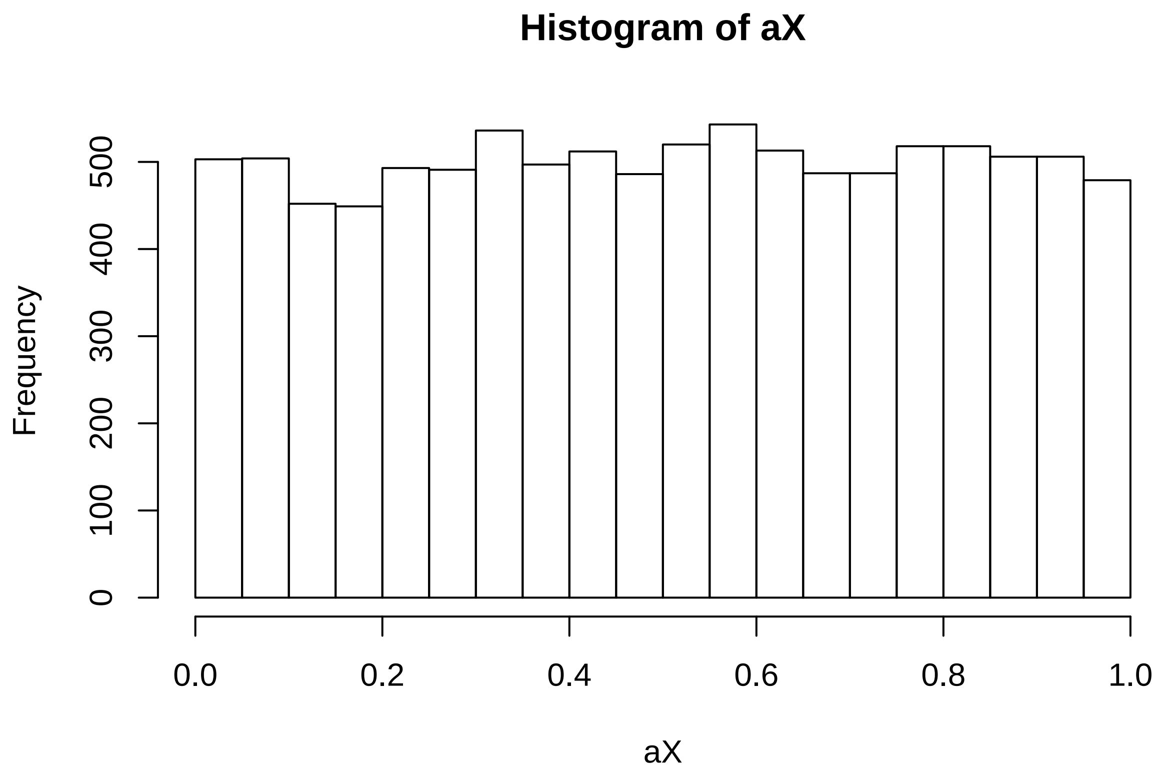
- $Y : S \to \mathbb{R}$
- Sannsynlighetstettheten $f_Y = F'_Y$ eksisterer. Verdimengden $\mathcal{Y} = Y(S)$ er da ikke-tellbar uendelig

Simulering av stokastiske variabler i R

```r
iN=10000 # Number of trials
aX = runif(iN) # experiment! Use rnorm(iN, 0, 1) in Ex1 :-)
summary(aX) # Show results
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0002495 0.2599578 0.5067085 0.5034681 0.7522271 0.9998878
```

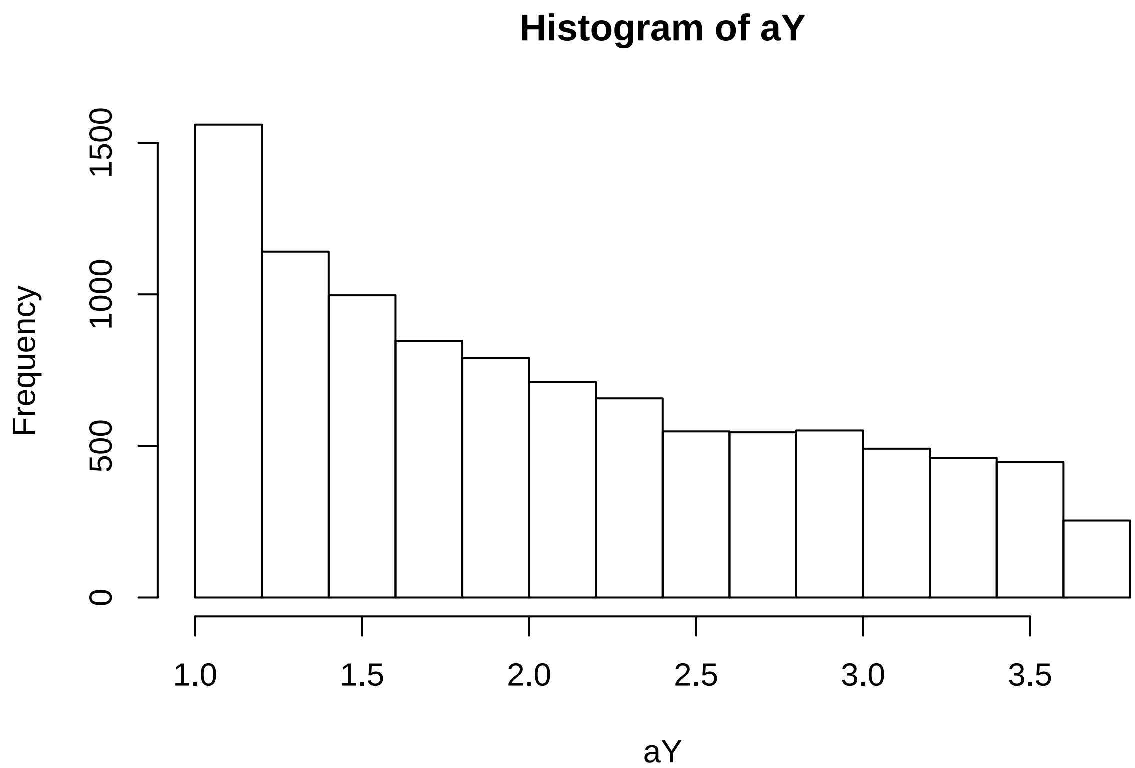
```r
hist(aX)
```




```
g = function(x) x^2 + exp(x)
aY = g(aX)
summary(aY) # Show results
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.364   1.917   2.059   2.688   3.718
```

```
hist(aY)
```



Diskret stokastisk variabel

$X : S \to \mathbb{R}$, $\mathcal{X}$ (verdimengen) endelig eller tellbar uendelig

Definisjon D.1:

$p_X(x)$ er en **punktsannsynlighet** og $p_X : \mathcal{X} \to \mathbb{R}$ er **sannsynlighetsfordelingen** til X dersom

1. $p_X(x) = P(X = x) = P(\{e \in S : X(e) = x\})$
2. $p_X(x) \geq 0$ for alle $x \in \mathcal{X}$
3. $\sum_{x \in \mathcal{X}} p_X(x) = 1$

Definisjon D.2:

Den kumulative fordelingsfunksjonen til X, med sannsynlighetsfordeling p_X er

$$F_X(x) = P(X \leq x) = \sum_{k \leq x} p_X(k), \quad k \in \mathcal{X}$$

Kumulativ fordelingsfunksjon

Regneregler

1. $F_X(x) = P(X \leq x) = 1 - P(X > x)$
2. $P(x_1 < X \leq x_2) = \sum_{k=x_1+1}^{x_2} p_X(k) = F_X(x_2) - F_X(x_1)$

Egenskaper $F_X(x)$

- $0 \leq F_X(x) \leq 1$
- $F_X(x)$ er voksende
- $F_X(x)$ er en høyrekontinuerlig trappefunksjon

Kontinuerlig stokastisk variabel

$Y : S \to \mathbb{R}$, verdimengde $\mathcal{Y}$ ikke-tellbar uendelig

Definisjon K.1:

Funksjonen $f_Y(y)$ er en sannsynlighetstetthet for Y dersom

1. $P(a \leq Y \leq b) = P(\{e \in S : a \leq Y(e) \leq b\}) = \int_a^b f_Y(y)dy$
2. $f_Y(y) \geq 0$ for alle $y \in \mathbb{R}$
3. $\int_{-\infty}^{\infty} f_Y(y)dy = 1$

Definisjon K.2:

Den kumulative fordelingsfunksjonen til Y, med sannsynlighetstetthet $f_Y(y)$, er

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(t)dt$$

Teorem 3.4.1

$$f_Y(y) = \frac{d}{dy} F_Y(y)$$

Forventningsverdi

Forventningsverdien til en stokastisk variabel er gjennomsnittet vi vil få dersom vi repeterer det stokastiske forsøket uendelig mange ganger: Gjennomsnittet $\overline{X}$ til et tilfeldig utvalg fra fordelingen til X konvergerer med sannsynlighet 1 mot forventningsverdien $E(X)$ når utvalgets størrelse går mot uendelig. Dette er **store talls lov** og det gir spesielt tolkningen av sannsynligheten til en hendelse som en grense av relativ hyppighet.

Definisjon: La X være en diskret stokastisk variabel med punktsannsynligheter gitt av $p_X(x)$ og anta at verdimengden $\mathcal{X} = X(S)$ er endelig. Da er X en **enkel stokastisk variabel** og forventningsverdien til X er definert ved

$$E(X) = \mu = \sum_{x \in \mathcal{X}} x \cdot p_X(x)$$

Forventningsverdien $E(X)$ til en generell stokastisk variabel defineres som en grense $E(X) = \lim E(X_n)$ hvor X_n er enkle stokastiske variable som konvergerer mot X. Dette definerer samtidig integralet

$$E(X) = \int X(s)\, P(ds)$$

Definisjon 3.5.1 Forventningsverdi

La X være en diskret stokastisk variabel med punktsannsynligheter gitt av $p_X(x)$. Forventningsverdien til X er definert ved

$$E(X) = \mu = \sum_{x \in \mathcal{X}} x \cdot p_X(x)$$

La Y være en kontinuerlig stokastisk variabel med sannsynlighetstetthet $f_Y(y)$. Forventningsverdien til Y er gitt ved

$$E(Y) = \mu = \int_{-\infty}^{\infty} y \cdot f_Y(y) dy$$

Teorem 3.5.1 For en tilfeldig variabel $X \sim \text{binomisk}(n, p)$, så er $E(X) = np$.

Definisjon 3.5.2 Median

La X være en diskret stokastisk variabel. Medianen er den verdien m som oppfyller $P(X < m) = P(X > m)$. Dersom $P(X \leq m_1) = P(X \geq m_2) = 0.5$ så er $m = \frac{m_1+m_2}{2}$.

La Y være en kontinuerlig stokastisk variabel. Medianen er løsningen på likningen $\int_{-\infty}^{m} f_Y(y)dy = 0.5$. Se eksempel 3.5.8.

Funksjoner av stokastiske variable

Teorem: La $Y = g(X)$. Da gjelder

$$E(Y) = E(g(X)) = \int g(X(s))\, P(ds) = \int g(x)\, P_X(dx)$$

hvor $P_X(A) = P(X \in A) = P\{s | X(s) \in A\}$.

Teorem 3.5.3 a)

X diskret, verdimengde $\mathcal{X}$, punktsannsynlighet $p_X(x)$. Da er

$$E(g(X)) = \sum_{x \in \mathcal{X}} g(x) p_X(x)$$

hvis $\sum_{x \in \mathcal{X}} |g(x)| p_X(x) < \infty$

Teorem 3.5.3 b)

Y kontinuerlig, sannsynlighetstetthet $f_Y(y)$. Da er

$$E(g(Y)) = \int_{-\infty}^{\infty} g(y) f_Y(y) dy$$

hvis $\int_{-\infty}^{\infty} |g(y)| f_Y(y) dy < \infty$.

Korrolar 3.5.1

W er en stokastisk variabel med forventningsverdi $E(W)$. For konstanter a og b så er

$$E(aW + b) = aE(W) + b$$

Varians

Definisjon 3.6.1 a)

X diskret, punktsannsynlighet $p_X(x)$, forventningsverdi $E(X) = \mu$. Da er

$$\text{Var}(X) = \sigma^2 = E((X-\mu)^2) = \sum_{x \in \mathcal{X}} (x-\mu)^2 p_X(x)$$

Definisjon 3.6.1 b)

Y kontinuerlig, sannsynlighetstetthet $f_Y(y)$, forventningsverdi $E(Y) = \mu$. Da er

$$\text{Var}(Y) = \sigma^2 = E((Y-\mu)^2) = \int_{-\infty}^{\infty} (y-\mu)^2 f_Y(y) dy$$

Teorem 3.6.1

W stokastisk variabel, forventningsverdi $E(W) = \mu$, og $|E(W^2)| < \infty$.

$$\text{Var}(W) = \sigma^2 = E(W^2) - \mu^2$$

Teorem 3.6.2

W stokastisk variabel, forventningsverdi $E(W) = \mu$, og $|E(W^2)| < \infty$. For konstanter a og b så er

$$\text{Var}(aW + b) = a^2 \text{Var}(W)$$

Momenter og momentgenererende funksjoner

Definisjon 3.6.2 (1)

Moment r for en stokastisk variabel W er

$$\mu_r = E(W^r)$$

Definisjon 3.12.1

Den **momentgenererende funksjonen** $M_W(t)$ for W er

$$M_W(t) = E(e^{tW})$$

for alle t der $E(e^{tW})$ eksisterer.

Teorem 3.12.1

Dersom moment r eksisterer så er

$$E(W^r) = \frac{d^r}{dt^r} M_W(t)|_{t=0} = M_W^{(r)}(0)$$

Teorem 3.12.2

For to stokastiske variable W_1 og W_2 der $M_{W_1}(t) = M_{W_2}(t)$ så er $f_{W_1}(w) = f_{W_2}(w)$. Med andre ord har W_1 og W_2 samme sannsynlighetsfordeling.

Teorem 3.12.3 (a):

La W være en stokastisk variabel med momentgenererende funksjon $M_W(t)$. La $V = aW + b$. Da er

$$M_V(t) = e^{bt} M_W(at).$$

Simultanfordeling

Definisjon 3.7.1 - Diskret simulatanfordeling

$X : S \to \mathbb{R}$, $Y : S \to \mathbb{R}$ er **diskrete** stokastiske variable. Simultan punktsannsynlighet er definert som

$$p_{X,Y}(x,y) = P(X = x, Y = y) = P(\{e \in S : X(e) = x \text{ og } Y(e) = y\})$$

Merk:

- $p_{X,Y}(x,y) \geq 0$
- $\sum_x \sum_y p_{X,Y}(x,y) = 1$

Teorem 3.7.1 - Diskret marginalfordeling

La $p_{X,Y}(x,y)$ være simultan punktsannsynlighet for diskrete variable X og Y. Da er

$$p_X(x) = \sum_y p_{X,Y}(x,y), \quad \text{og} \quad p_Y(y) = \sum_x p_{X,Y}(x,y)$$

Definisjon 3.7.3 - Kontinuerlig simultanfordeling

$X : S \to \mathbb{R}$, $Y : S \to \mathbb{R}$ er **kontinuerlige** stokastiske variable. Simultan sannsynlighetstetthet $f_{X,Y}(x,y)$ for et areal $A \subset \mathbb{R}^2$ er definert ved

$$P(\{e \in S : (X(e), Y(e)) \in A\}) = P((X,Y) \in A) = \int \int_A f_{X,Y}(x,y) dx dy$$

Merk:

- $f_{X,Y}(x,y) \geq 0$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$

Teorem 3.7.2 - Kontinuerlig marginalfordeling

La $f_{X,Y}(x,y)$ være simultan sannsynlighetstetthet for kontinuerlige variable X og Y. Da er

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy, \quad \text{og} \quad f_y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx$$

Definisjon 3.7.4 - Kumulativ simultanfordeling

La U og V være to stokastiske variable. Da er den *kumulative simultanfordelingen*

$$F_{U,V}(u,v) = P(U \leq u \text{ og } V \leq v)$$

Teorem 3.7.3

La X og Y være to **kontinuerlige** stokastiske variable med kumulativ simultanfordeling $F_{X,Y}(x,y)$. Da er

$$f_{x,y} = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$$

Simultanfordelinger for mer enn to variable

Diskret:

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

Kontinuerlig: For en region $R \subset \mathbb{R}^n$:

$$P(Y_1, \ldots, Y_n \in R) = \int \cdots \int_R f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) dy_1 \cdots dy_n$$

Betinget sannsynlighet og uavhengige stokastiske variable

Definisjon 3.11.1 (a) Betinget punktsannsynlighet

La X og Y være diskrete, med simultan punktsannsynlighet $p_{X,Y}(x,y)$. Betinget punktsannsynlighet for Y, *gitt at* $X = x$ er

$$p_{Y|x}(y) = P(Y = y|\ X = x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

for $p_X(x) > 0$

Definisjon 3.11.1 (b) Betinget sannsynlighetstetthet

La X og Y være kontinuerlige, med simultan sannsynlighetstethet $f_{X,Y}(x,y)$. Betinget sannsynlighetstetthet for Y, *gitt at* $X = x$ er

$$f_{Y|x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

for $f_X(x) > 0$. Dermed er

$$P(a \leq Y \leq b|\ X = x) = \int_a^b f_{Y|x}(y)dy$$

Definisjon 3.7.5 - Uavhengige stokastiske variable La X og Y være diskrete, med simultan punktsannsynlighet $p_{X,Y}(x,y)$. Da er X og Y uavhengige dersom

$$p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

La X og Y være kontinuerlige, med simultan sannsynlighetstethet $f_{X,Y}(x,y)$ Da er X og Y uavgengige dersom

$$\int_a^b \int_c^d f_{X,Y}(x,y)dydx = \int_a^b f_X(x)dx \int_c^d f_Y(y)dy$$

Teorem 3.7.4 La X og Y være kontinuerlige, med simultan sannsynlighetstethet $f_{X,Y}(x,y)$ Da er X og Y uavgengige hvis og bare hvis

$$f_{X,Y}(x,y) = g(x)h(y)$$

for funksjoner $g(x)$ og $h(y)$. Dersom dette er sant, så finnes en konstant k slik at $f_X(x) = kg(x)$ og $f_Y(y) = \frac{1}{k}h(y)$, altså er $g(x)h(y) = f_X(x)f_Y(y)$.

Lineærtransformasjoner - sannsynlighetsfordeling

$Y = aX + b$, $E(Y) = aE(X) + b$, $\mathrm{Var}(Y) = a^2\mathrm{Var}(X)$

Teorem 3.8.1

X diskret, $Y = aX + b$,

$$p_Y(y) = p_X\left(\frac{y-b}{a}\right)$$

Teorem 3.8.2

X kont, $Y = aX + b$,

$$f_Y(y) = \frac{1}{|a|}f_X\left(\frac{y-b}{a}\right)$$

Lineærtransformasjoner

Teorem 3.8.1

X diskret, $Y = aX + b$

$$p_Y(y) = p_X\left(\frac{y-b}{a}\right)$$

Teorem 3.8.2

X kont, $Y = aX + b$

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

Kombinasjoner av stokastiske variabler

$$Z = g(X, Y), \quad g : \mathbb{R}^2 \to \mathbb{R}$$

Teorem 3.9.1

X, Y diskrete med simultan punktsannsynlighet $p_{X,Y}(x, y)$. Da er

$$E(g(X, Y)) = \sum_x \sum_y g(x, y) p_{X,Y}(x, y)$$

X, Y kontinuerlige med simultantetthet $f_{X,Y}(x, y)$. Da er

$$E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy$$

Sum av to stokastiske variable

Teorem 3.9.2

$$E(aX + bY) = aE(X) + bE(Y)$$

for

- X og Y diskrete, eller X og Y kontinuerlige,
- X og Y avhengige eller X og Y uavhengige

Teorem 3.9.5

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

Definisjon 3.9.1 - kovarians

Kovariansen til to stokastiske variabler X og Y er

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y)))E(XY) - E(X)E(Y)$$

Teorem 3.9.4

Dersom X og Y er uavhengige så er $\text{Cov}(X, Y) = 0$.

Korollar 3.9.4

Dersom X og Y er uavhengige så er

$$\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$$

Korrelasjon

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

$$-1 \leq \mathrm{Corr}(X,Y) \leq 1$$

Sum av to uavhengige stokastiske variabler

$Z = X + Y$, X og Y uavhengige.

- $E(Z) = E(X) + E(Y)$
- $\mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
- $p_Z(z) = ?$, $f_Z(z) = ?$

Teorem 3.12.3 (b)

La $Z = X + Y$. For uavhengige stokastiske variable X og Y med momentgenererende funksjoner $M_X(t)$ og $M_Y(t)$ så er

$$M_Z(t) = M_X(t) \cdot M_Y(t)$$

Teorem 3.8.3 (1)

X og Y er diskrete stokastiske variable og $Z = X + Y$. Da er

$$p_Z(z) = \sum_{x \in \mathcal{X}} p_X(x) p_Y(z-x)$$

Teorem 3.8.3 (2)

X og Y er kontinuerlige stokastiske variable og $Z = X + Y$. Da er

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx$$

Tilfeldig utvalg og sentralgrenseteoremet

Definisjon 3.7.7

La $X_1, \ldots, X_n$ være n uavhengige stokatiske variable fra den samme sannsynlighetsfordelingen ($p_X(x)$ eller $f_X(x)$) - altså er de identisk fordelte. Da er $X_1, \ldots, X_n$ et *tilfeldig utvalg* fra $p_X(x)$ eller $f_X(x)$.

Standardisering av stokastiske variable

La X være en stokastisk variabel med forventningsverdi $E(X) = \mu$ og varians $\mathrm{Var}(X) = \sigma^2$. Da er $\frac{X-\mu}{\sigma}$ en stokastisk variabel med forventningsverdi 0 og varians 1.

La $X_1, \ldots, X_n$ være identisk fordelte, uavhengige stokastiske variabler, slik at $E(X_i) = \mu$ og $\mathrm{Var}(X_i) = \sigma^2$ for alle $i = 1, \ldots, n$. Da er $E(\sum_{i=1}^n X_i) = n\mu$ og $\mathrm{Var}(\sum_{i=1}^n X_i) = n\sigma^2$, slik at $\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}}$ er en stokastisk variabel med forventningsverdi 0 og varians 1.

La $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Da er $E(\bar{X}) = \mu$ og $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$. Da er $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ en stokastisk variabel med forventningsverdi 0 og varians 1.

Normaltilnærming til binomisk fordeling

La $X \sim \text{binom}(n, p)$, da er

$$\lim_{n\to\infty} P\left(a < \frac{X - np}{\sqrt{np(1-p)}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Det følger at $P(a \leq X \leq b) \approx P(a \leq Y \leq b)$ der $Y \sim N(np, np(1-p))$. Når n ikke er særlig stor så kan vi gjøre en kontinutetskorreksjon. Vi har sett at $P(a \leq X \leq b) \approx P(a - 0.5 < Y \leq b + 0.5)$ gir en bedre tilnærming når n er liten. Dersom vi endrer ulikhetene får vi tilsvarende $P(a < X \leq b) \approx P(a + 0.5 < Y \leq b + 0.5)$, osv. Dette er enklest å se ved å tegne et sannsynlighetshistogram for den binomiske fordelingen og sannsynlighetstettheten for normalfordelingen.

Teorem 4.3.2 - Sentralgrenseteoremet

La $X_1, \ldots, X_n$ være identisk fordelte, uavhengige stokastiske variabler, slik at $E(X_i) = \mu$ og $\text{Var}(X_i) = \sigma^2$ for alle $i = 1, \ldots, n$. Da er

$$\lim_{n\to\infty} P\left(a < \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Alternativ formulering

$$\lim_{n\to\infty} P\left(a < \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$