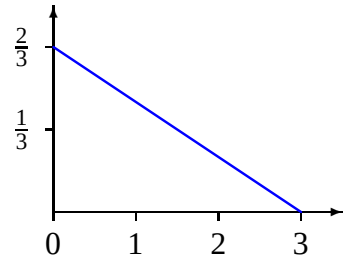
Kontinuerliga stokastiska variabler

- En stokastisk variabel är *kontinuerlig* om den kan anta vilka värden som helst i ett intervall, *men sannolikheten för varje enskilt utfall är noll:* $P(X = x) = 0$.
- Typexemplet är tid, eller vilken annan kvantitet som helst som man kan uppmäta kontinuerligt: Sannolikheten att kvantiteten skall vara *exakt* lika med x och inte $x + 0.0001$ eller $x + 0.0000001$ t.ex., är noll.

Exempel



Figur: Tätheten $f(x) = \frac{2}{3} - \frac{2}{9}x$ för $0 \leq x \leq 3$, $f(x) = 0$ för $x < 0, x > 3$.

Integraler i stället för summor

- Kom ihåg: För diskreta stokastiska variabler är $P(a \leq X \leq b) = \sum_{x=a}^{b} f(x)$.
- För kontinuerliga stokastiska variabler gäller i stället $P(a \leq X \leq b) = \int_a^b f(x)dx$, för en funktion f som också kallas täthetsfunktionen.
- $f(x) \geq 0$ och $\int_{-\infty}^{\infty} f(x)dx = 1$.
- För kontinuerliga stokastiska variabler är $f(x)$ INTE en sannolikhet, man kan ha $f(x) > 1$ vilket är omöjligt för diskreta stokastiska variabler ($P(X = x)$ kan ju inte vara större än ett!).

Integraler i stället för summor

- Fördelningsfunktionen har dock samma definition som för diskreta stokastiska variabler: $F(x) = P(X \leq x)$.
- Detta betyder att $F(x) = \int_{-\infty}^{x} f(u)du$ och $f(x) = F'(x)$.
- Obs: Det är ofta lättare analytiskt att integrera och derivera än att beräkna summor.
- Väntevärde: $\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$, $E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx$.
- Obs: $f(x)$ kan vara lika med noll utanför ett intervall $[a, b]$, så att de faktiska gränserna i integralerna ovan blir $\int_a^b$.
- $\sigma^2 = \mathrm{Var}(X) = E[X^2] - (E[X])^2$ som för diskreta variabler.

Exempel

- $f(x) = \frac{2}{3} - \frac{2}{9}x$ för $0 \leq x \leq 3$, $f(x) = 0$ för $x < 0, x > 3$.
- $F(x) = \int_{-\infty}^{x} f(u)du = 0$ om $x < 0$.
- För $0 \leq x \leq 3$:

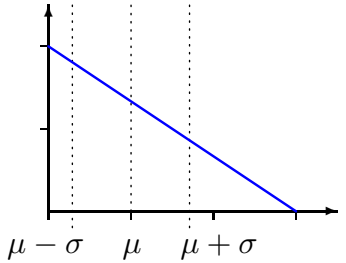
$$F(x) = \int_{-\infty}^{x} f(u)du = \int_{0}^{x} f(u)du = \int_{0}^{x} \left(\tfrac{2}{3} - \tfrac{2}{9}u\right) du$$
$$= \left[\tfrac{2}{3}u - \tfrac{1}{9}u^2\right]_0^x = \tfrac{2}{3}x - \tfrac{1}{9}x^2.$$

- För $x > 3$ så är

$$F(x) = \int_{-\infty}^{x} f(u)du = \int_{-\infty}^{3} f(u)du + \int_{3}^{x} f(u)du = F(3) + 0 = 1.$$

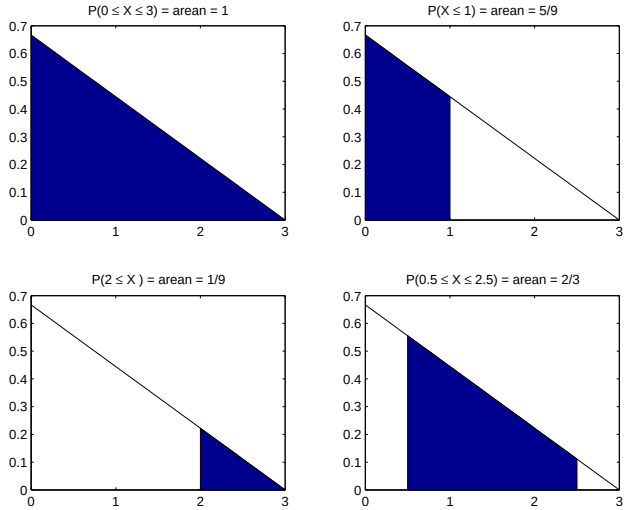
Exempel

- $E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{3} xf(x)dx = \int_{0}^{3} \left(\frac{2}{3}x - \frac{2}{9}x^2\right) dx = 1.$
- $E[X^2] = \int_{0}^{3} x^2 f(x)dx = \int_{0}^{3} \left(\frac{2}{3}x^2 - \frac{2}{9}x^3\right) dx = 1.5.$
- $\mathrm{Var}(X) = 1.5 - 1^2 = 0.5$, $\mathrm{SD}(X) = \sqrt{0.5} \doteq 0.71.$



Figur: $E[X] = 1$, $\mathrm{Var}(X) = \frac{1}{2}$, $\mathrm{SD}(X) \doteq 0.71$

Exempel

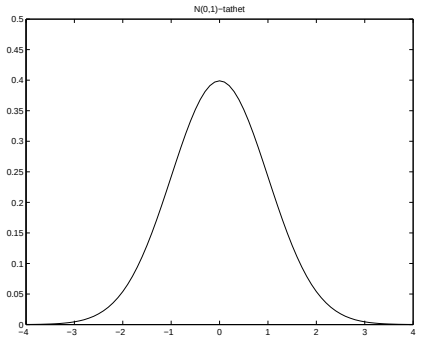


Normalfördelningen

Den viktigaste fördelningen är den så kallade *normalfördelningen* med täthetsfunktion

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty.$$

Man brukar skriva $X \sim N(\mu, \sigma^2)$ eller $X \sim N(\mu, \sigma)$.



Normalfördelningen

- Normalfördelningen är viktig pga den ”centrala gränsvärdessatsen” (kap. 7.3) som vi kommer att gå igenom senare och som förklarar varför denna fördelning är så vanligt förekommande överallt.
- Parametrarna μ och σ^2 är precis väntevärdet och variansen för den normalfördelade stokastiska variabeln.
- Obs: En normalfördelad stokastisk variabel kan anta hur små värden som helst eftersom $f(x) > 0$ för alla x. T.ex. betyder det att man ofta menar att något är approximativt normalfördelat då man t.ex. säger ”Längden på en godtyckligt vald person i Sverige är normalfördelad”. Längd kan ju aldrig vara negativ, men sannolikheten för detta är ofta mikroskopisk, så det gör inget.

Kvantiler

- Med hjälp av Table V. kan vi svara på frågor av typen ”$P(X \leq x) = ?$”, där x är ett givet tal och X är normalfördelad.
- Ibland vill man kunna svara på den ”omvända” frågan: $P(X > ?) = \alpha$, där α är en given sannolikhet.
- Lösningen x_α, $P(X > x_\alpha) = \alpha$, kallas för α-kvantilen.
- Vi skriver z_α för $N(0,1)$-fördelningens α-kvantil:

α	0.1	0.05	0.025	0.01	0.005
z_α	1.282	1.645	1.960	2.326	2.576

Standardisering

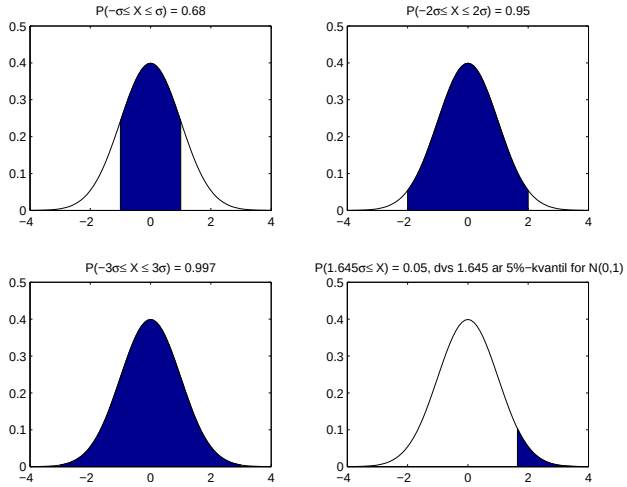
- $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$, X och Y ober. $\Rightarrow$
  - $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$,
  - $aX + b \sim N(a\mu_X + b, a^2\sigma_X^2)$.
- Så med $a = \frac{1}{\sigma_X}$ och $b = -\frac{\mu_X}{\sigma_X}$ är alltså $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.
- Detta gör att man kan omvandla alla sannolikheter som har med X att göra till sannolikheter som har med Z att göra, dvs behöver inte bekymra oss om olika specialfall för olika värden på μ och σ^2.
- Fördelningsfunktionen F för Z finns tabulerad (Table V.).
- Ex: $X \sim N(3, 2^2)$.

$$\begin{aligned} P(4 \leq X \leq 6) &= P\left(\frac{4-3}{2} \leq \frac{X-3}{2} \leq \frac{6-3}{2}\right) \\ &= P(0.5 \leq Z \leq 1.5) \\ &= F(1.5) - F(0.5) \\ &\doteq 0.9332 - 0.6915 = 0.2417 \end{aligned}$$

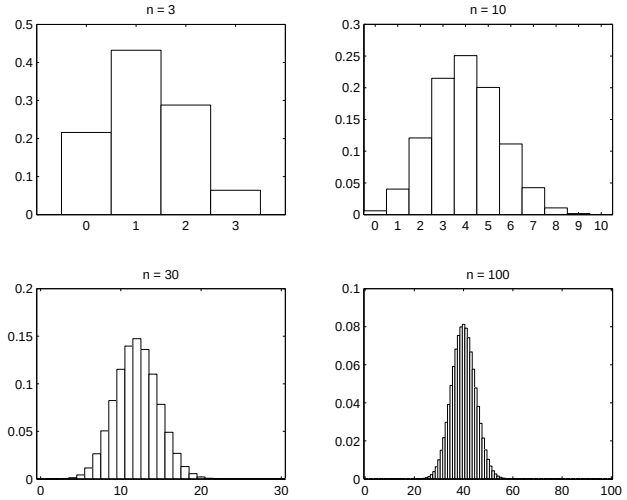
Ex. Kvalitet. Antag att du producerar en vara som inte får väga för mycket. Produktionsprocessen är sådan att vikten X för en produkt kan anses vara normalfördelad med väntevärde $\mu = 100$g, och $\sigma = 5$g. Om vi ska ge en garanti: ”Varan väger högst x g”, som ska hålla för 95% av alla varor så ska vi hitta 5%-kvantilen x.

$$\begin{aligned} 0.05 = P(X > x) &= P\left(\tfrac{X-100}{5} > \tfrac{x-100}{5}\right) = P(Z > \tfrac{x-100}{5}) \\ &\Rightarrow \\ \tfrac{x-100}{5} &= z_{0.05} \doteq 1.645 \\ &\Rightarrow \\ x &= 5z_{0.05} + 100 \doteq 108.23 \end{aligned}$$

Normalfördelade variabler avviker oftast inte så mycket, mätt i standardavvikelser, från sitt väntevärde



Binomialfördelningen liknar normalfördelningen då n ökar. Exempel: Bin(n,0.4), $n = 3, 10, 30, 100$.



Standardavvikelser och Tjebyshevs olikhet

- För normalfördelade variabler har man (förra bilden):

$$P(|X-\mu| > \sigma) \doteq 0.3174$$
$$P(|X-\mu| > 2\sigma) \doteq 0.0456$$
$$P(|X-\mu| > 3\sigma) \doteq 0.0026$$

- För alla stokastiska variabler gäller Tjebyshevs olikhet:

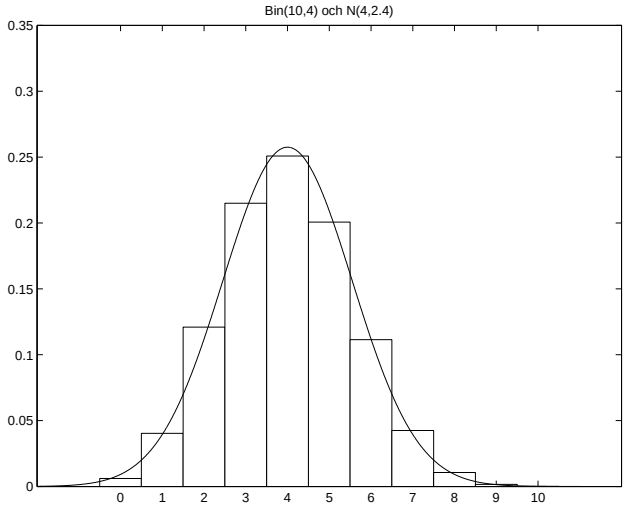
$$P(|X-\mu| > k\sigma) \leq \frac{1}{k^2}.$$

- Bevisskiss: $KP(Y > K) = K\int_K^\infty f(y)dy \leq \int_K^\infty yf(y)dy = E[Y]$. Låt $Y = (X-\mu)^2$ och $K = k^2\sigma^2$.

$$k^2\sigma^2 P((X-\mu)^2 > k^2\sigma^2) = k^2\sigma^2 P(|X-\mu| > k\sigma)$$
$$\leq E[(X-\mu)^2] = \sigma^2,$$

vilket ger Tjebyshevs olikhet.

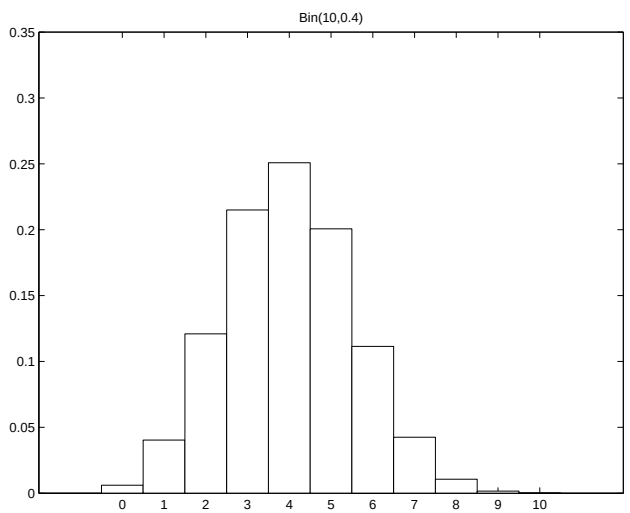
Bin(10,0.4) och N(4,2.4) i samma figur



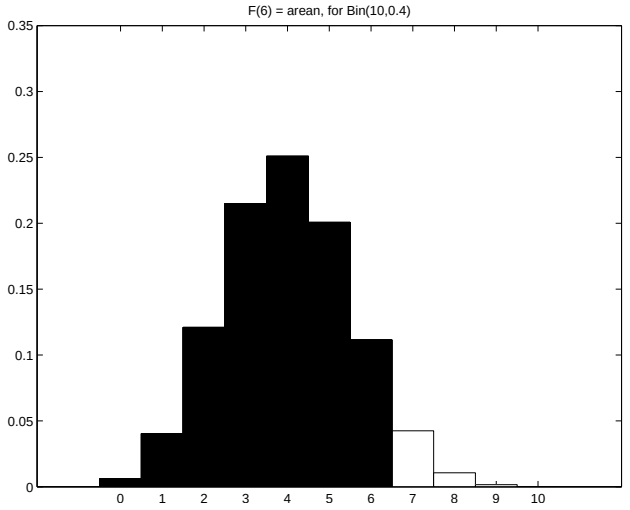
Normalapproximation av binomialfördelningen

- Låt $X \sim \text{Bin}(n, p)$, dvs $E[X] = np$, $\text{Var}(X) = np(1-p)$.
- *Normalapproximation*: $X \approx Y \sim N(np, np(1-p))$ då n är stort.
- Tumregel: n är ”stort” om $n \cdot \min\{p, 1-p\} > 5$.
- Obs. X är diskret och Y är kontinuerlig så om x är ett heltal:

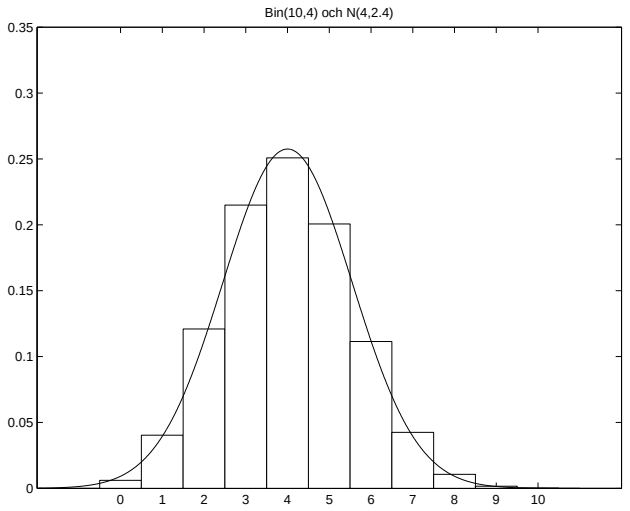
$$P(X \leq x) = P(X \leq x + 0.5) \approx P(Y \leq x + 0.5).$$



Figur: Låt $X \sim \text{Bin}(10, 0.4)$.

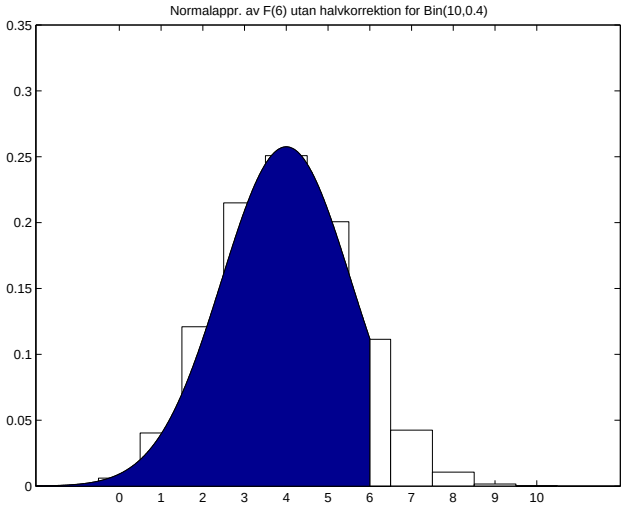


Figur: $P(X \leq 6) = 0.9452$.



Figur: $n \cdot \min\{p, 1-p\} = 10 \cdot 0.4 = 4 \leq 5$ så normalapproximationen med $Y \sim N(4, 2.4)$ kanske inte funkar så bra. . . ?

Utan halvkorrektion



Figur: $P(Y \leq 6) = 0.9017$, ($P(X \leq 6) = 0.9452$).

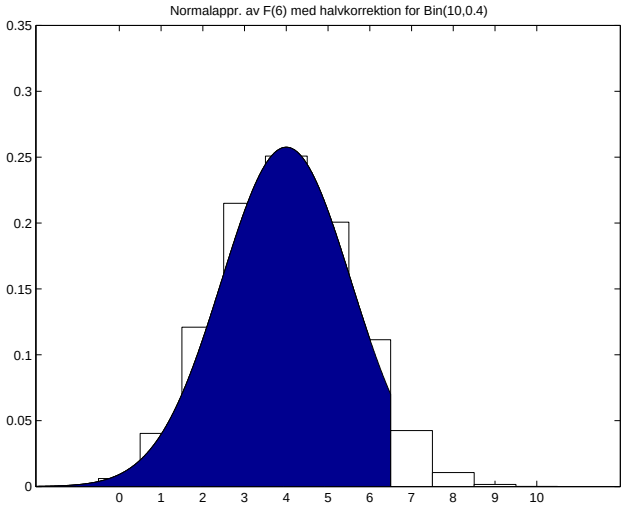
Hur får man fram siffrorna?

Låt F vara fördelningsfunktionen för $Z \sim N(0,1)$.

$$P(Y \leq 6.5) = P\Big(\underbrace{\tfrac{Y-4}{\sqrt{2.4}}}_{\sim Z} \leq \underbrace{\tfrac{6.5-4}{\sqrt{2.4}}}_{\doteq 1.61}\Big) = F(1.61) = 0.9463, \text{ enl. Table V.}$$

$P(X \leq 6) = 0.9452$, enl. Table I.

Med halvkorrektion



Figur: $P(Y \leq 6.5) = 0.9463 \approx 0.9452 = P(X \leq 6)$!

Simultana diskreta fördelningar

- Hittills har vi talat om stokastiska variabler en och en.
- Det är enkelt att utvidga de definitioner vi har till så kallade *simultana fördelningar* med flera stokastiska variabler på en och samma gång.
- Ex. Diskret simultan täthet $f_{XY}(x,y) = P(X = x, Y = y)$ som kan sammanfattas i en tabell.

$f_{XY}(x,y)$				
$x \backslash y$	0	1	2	3
0	0.840	0.030	0.020	0.010
1	0.060	0.010	0.008	0.002
2	0.010	0.005	0.004	0.001

- Eftersom tätheterna är sannolikheter måste vi ha $f_{XY}(x,y) \geq 0$ och $\sum_x \sum_y f_{XY}(x,y) = 1$.

Marginell fördelning

- Obs: $f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f_{XY}(x,y)$, och på samma sätt $f_Y(y) = \sum_x f_{XY}(x,y)$. Detta kallas för den *marginella tätheten*.
- Ex. forts.

$f_{XY}(x,y)$					
$x \backslash y$	0	1	2	3	Σ_y
0	0.840	0.030	0.020	0.010	0.900
1	0.060	0.010	0.008	0.002	0.080
2	0.010	0.005	0.004	0.001	0.020
Σ_x	0.910	0.045	0.032	0.013	1.000

Väntevärden, etc.

- Allmänt: $E[h(X,Y)] = \sum_x \sum_y h(x,y) f_{XY}(x,y)$.
- $$E[X] = \sum_x \sum_y x f_{XY}(x,y) = \sum_x x \sum_y f_{XY}(x,y) = \sum_x x f_X(x),$$

  som sig bör!
- $$E[X+Y] = \sum_x \sum_y (x+y) f_{XY}(x,y) dxdy$$
  $$= \sum_x \sum_y x f_{XY}(x,y) + \sum_x \sum_y y f_{XY}(x,y) = E[X] + E[Y],$$

  vilket vi påstod tidigare.

Simultana kontinuerliga fördelningar

- För kontinuerliga stokastiska variabler finns det också en simultan täthet $f_{XY}(x,y)$ som uppfyller:

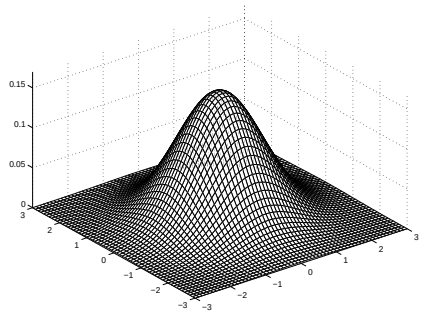
$$f_{XY}(x,y) \geq 0$$
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) dxdy = 1$$
$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \left(\int_c^d f_{XY}(x,y) dy \right) dx$$

- Marginella tätheter:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y) dy$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y) dx$$

Oberoende

- Kom ihåg: Händelserna A och B är oberoende om $P(A \cap B) = P(A)P(B)$.
- Två stokastiska variabler X och Y är *oberoende* om $f_{XY}(x,y) = f_X(x) f_Y(y)$ *för alla* x *och* y.

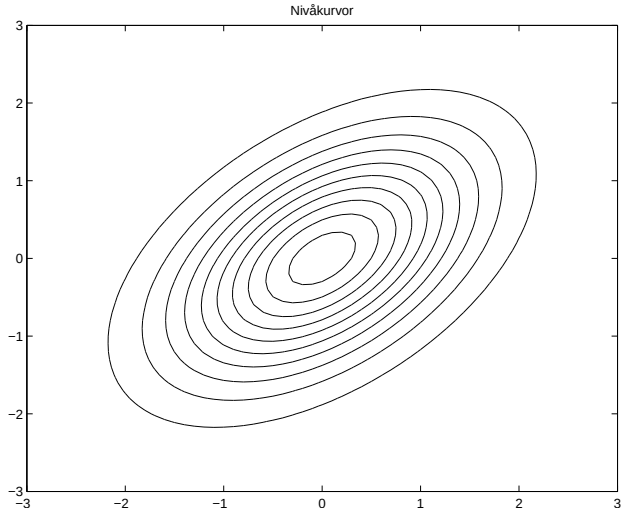


Figur: Täthet för två oberoende normalfördelade stokastiska variabler: $f_{XY}(x,y) = f_X(x) f_Y(y) = \exp\{-0.5x^2\} \exp\{-0.5y^2\}/2\pi$.

Kovarians

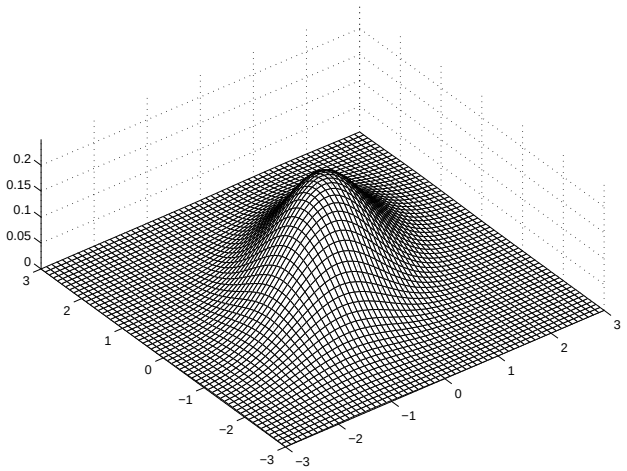
- Det som gör det intressant att tala om simultana fördelningar är det faktum att variablerna kan vara beroende.
- *Kovariansen* är ett mått på beroendet:
  $\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- Om X och Y tenderar att röra sig ”åt samma håll” så är $\mathrm{Cov}(X,Y) > 0$, och omvänt.
- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
- X och Y oberoende $\Rightarrow \mathrm{Cov}(X,Y) = 0$, ty
  $E[XY] = \sum_x \sum_y xyf_{XY}(x,y) = \sum_x \sum_y xyf_X(x)f_Y(y) =$
  $\sum_x xf_X(x) \sum_y yf_Y(y) = E[X]E[Y]$.
- Det omvända är inte alltid sant: $\mathrm{Cov} = 0 \not\Rightarrow$ oberoende.

Exempel, forts.



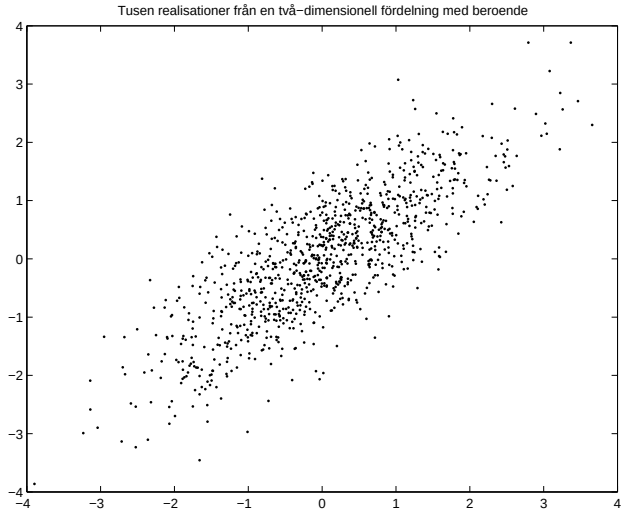
Figur: Nivåkurvor för föregående täthet.

Exempel



Figur: Täthet med positivt beroende.

Exempel, forts.

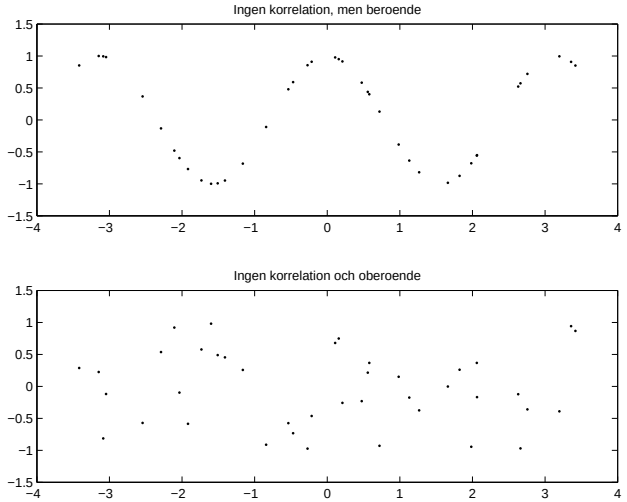


Figur: Tusen realisationer från fördelningen med föregående täthet.
$\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$, $\mathrm{Cov}(X,Y) = 0.5$.

Korrelation

- Definition av *korrelation*: $\rho = \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}$
- Det gäller alltid att $-1 \leq \rho \leq 1$, så korrelationen är enhetslös.
- $\rho = 0 \iff \mathrm{Cov}(X,Y) = 0$.
- $\rho = \pm 1 \iff Y = \pm aX + b, a > 0$, dvs X och Y är helt *linjärt* beroende.
- Men: $\rho = 0 \not\Rightarrow X$ och Y oberoende.

Korrelation säger bara något om *linjärt* beroende:



Figur: Överst: $\rho = 0$, men vi har ett tydligt beroende ($Y = \cos 2X$).
Nederst: $\rho = 0$ och X och Y är oberoende.

Korrelation, exempel