

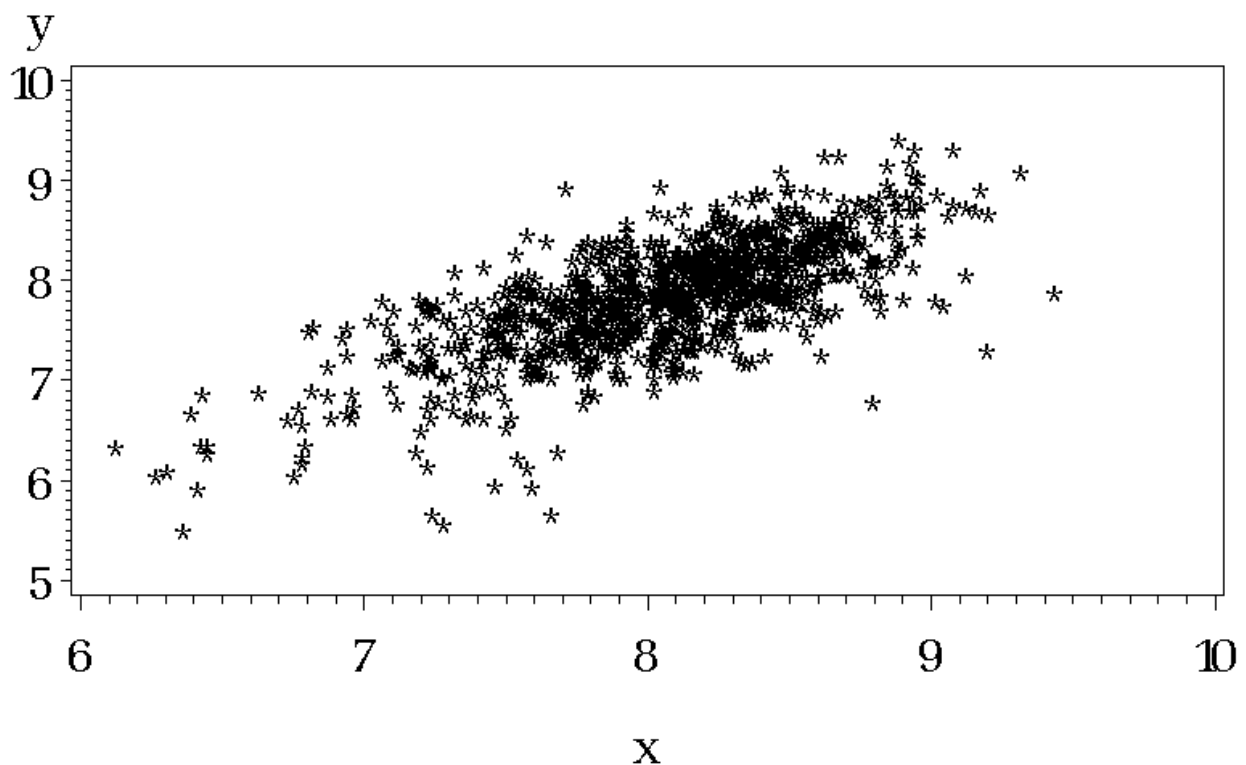
Et eksempel på en todimensional normalfordeling

Anders Milhøj

September 2006

I dette notat gennemgås et eksempel, der illustrerer den todimensionale normalfordelings egenskaber. Notatet lægger sig op af gennemgangen i Teoretisk Statistik for Økonomer. Det bygger desuden især på gennemgangen i Anders Halds klassiske bog "Statistiske Metoder" eller den engelske udgave, til hvilken der henvises for en mere udførlig behandling. I notatet anvendes uden videre beregnede talstørrelser i stedet for teoretiske parameterværdier.

Figuren viser gennemsnittet af årskaracteren i skriftlig matematik sammenholdt med gennemsnitskaracteren ved den tilsvarende skriftlige eksamen. Data er fra 9-ende klasse sommeren 2003 for ca. 1000 "normale" kommunale folkeskoler med flere end 5 elever ved 9-ende klasses afgangseksamen. Data kan hentes fra undervisningsministeriets hjemmeside, hvor mange tilsvarende datasæt om karakterer i skolerne let kan hentes. I det følgende betegnes årskaractererne med x og eksamenskaractererne med y .



De marginale normalfordelinger

En simpel beregning af gennemsnit og varians for de ca. 1000 observationer giver

$$E[x] = 8.06,$$

$$E[y] = 7.84,$$

$$\text{var}(x) = 0.32 = 0.57^2,$$

$$\text{var}(y) = 0.25 = 0.50^2$$

Tilsvarende beregnes covariansen til

$$\text{cov}(x,y) = 0.20.$$

Kovariansmatricen er derfor

$$\Sigma = \begin{pmatrix} 0.32 & 0.20 \\ 0.20 & 0.25 \end{pmatrix}$$

Korrelationskoefficienten bliver

$$\text{corr}(x,y) = 0.71.$$

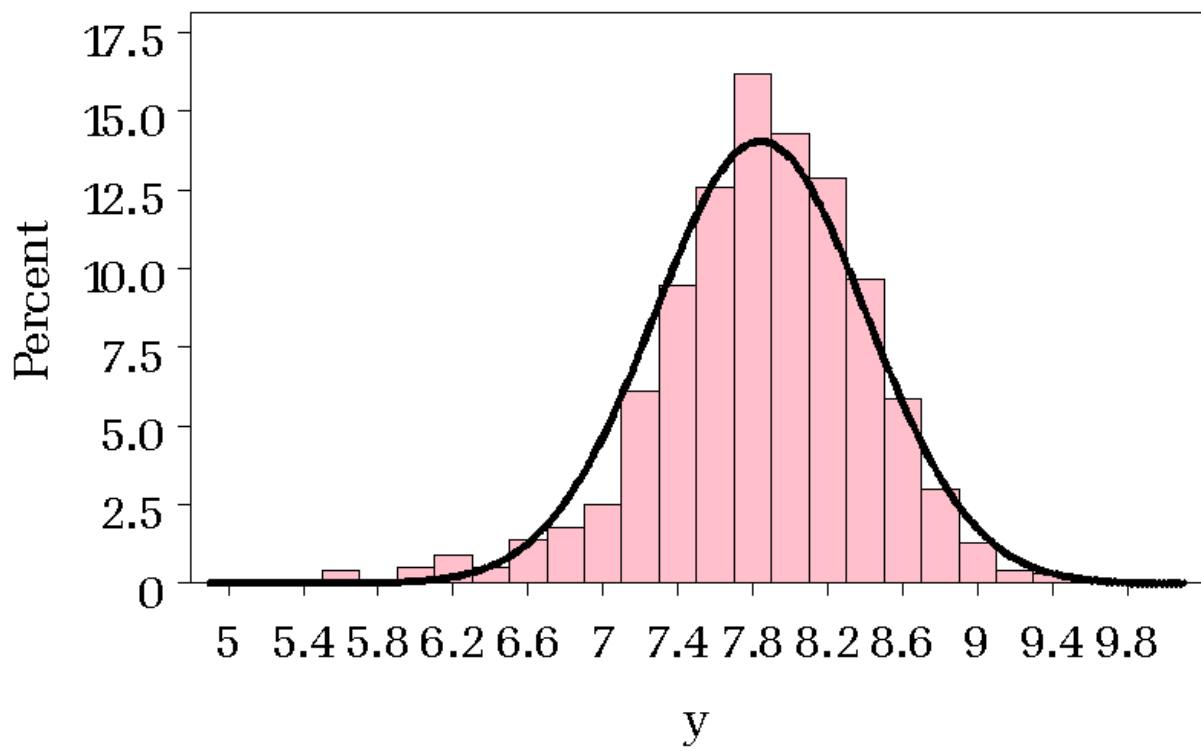
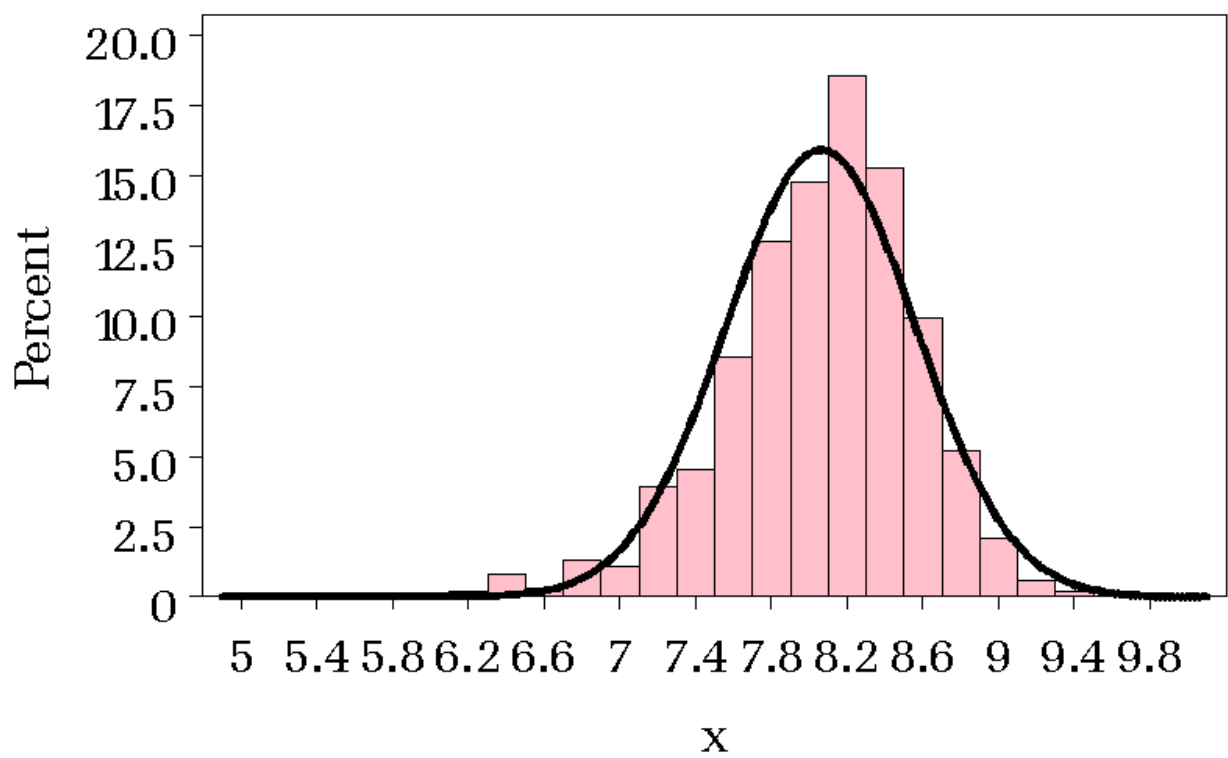
Korrelationsmatricen er derfor

$$\begin{pmatrix} 1 & 0.71 \\ 0.71 & 1 \end{pmatrix}$$

De to marginale fordelinger kan begge beskrives ved normalfordeling, hvad histogrammerne med indlagte normalfordelingstæthedsfunktioner også viser. Den endimensionale tæthedsfunktion for x er givet ved

$$\frac{1}{\sqrt{2\pi \text{var}(x)}} \exp \left\{ -\frac{(x - E[x])^2}{2 \text{var}(x)} \right\}$$

og tilsvarende for y .



Tabellerne viser værdierne af den marginale tæthedsfunktion for årskarakteren x lig 5, 6, 7 og 8 samt de tilsvarende værdier af den marginale tæthedsfunktion for eksamenskarakteren, y.

x	Tæthed	y	Tæthed
5	0.00653	5	0.01414
6	0.08450	6	0.14681
7	0.40211	7	0.56069
8	0.70397	8	0.78774

Den todimensionale normalfordeling

Det er naturligt, at beskrive den todimensionale fordeling af de to karaktergennemsnit ved en todimensional normalfordeling. Denne todimensionale normalfordeling har middelværdivektor μ og varians-kovariansmatrix Σ . For disse parametre indsættes de tilsvarende empiriske ækvivalenter.

Det giver tæthedsfunktionen

$$f(x,y) = \frac{1}{2\pi\sqrt{\det\begin{pmatrix}\sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2\end{pmatrix}}} \exp\left\{-\frac{1}{2} \begin{pmatrix}(x-\mu_1) & (y-\mu_2)\end{pmatrix} \begin{pmatrix}\sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2\end{pmatrix}^{-1} \begin{pmatrix}(x-\mu_1) \\ (y-\mu_2)\end{pmatrix}\right\}$$

$$= \frac{1}{2\pi\sqrt{\text{var}(x) \text{var}(y)(1 - \text{corr}(x, y)^2)}} \times$$

$$\exp\left\{-\frac{1}{2} \frac{(x - E[x])^2}{\text{var}(x)(1 - \text{corr}(x, y)^2)} + \frac{(y - E[y])^2}{\text{var}(y)(1 - \text{corr}(x, y)^2)} - 2(x - E[x])(y - E[y]) \frac{\text{corr}(x, y)}{\sqrt{\text{var}(x) \text{var}(y)(1 - \text{corr}(x, y)^2)}}\right\}$$

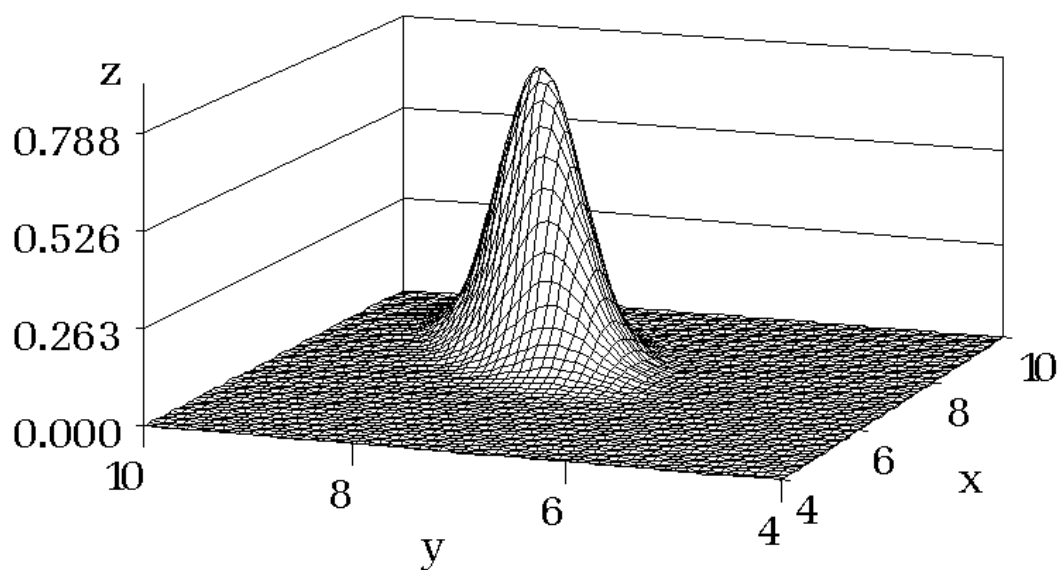
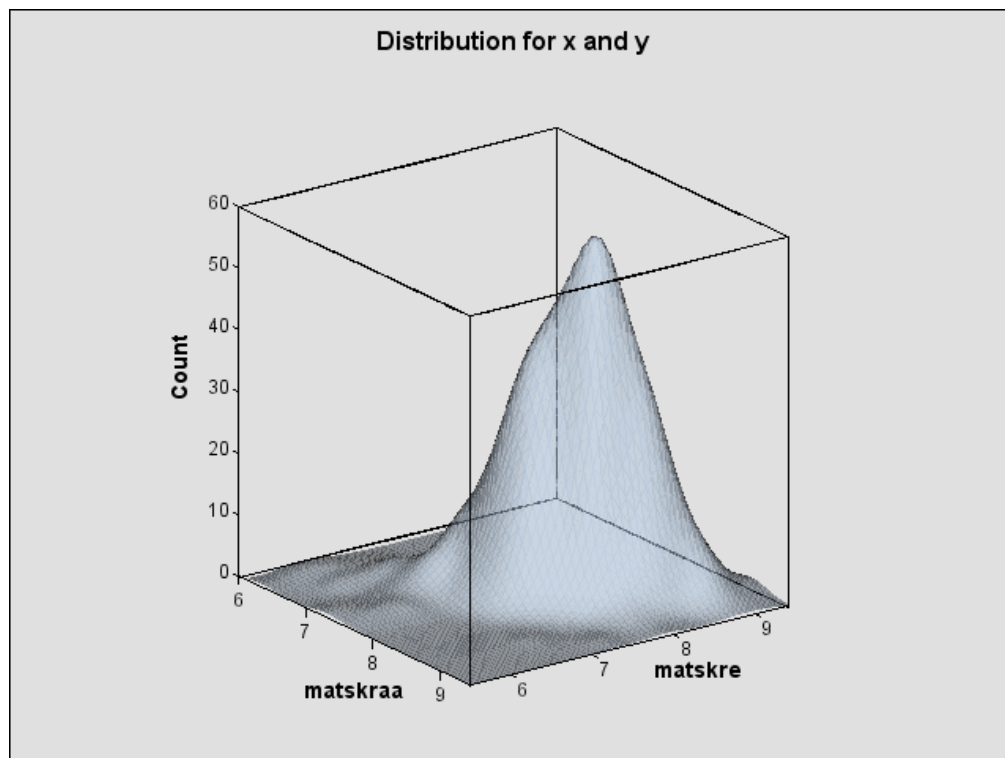
Den todimensionale simultane normalfordelingstæthed for x, y = 5, 6, 7, 8 er angivet i tabellen.

Bemærk at den simultane fordeling ikke er lig med produktet af de to marginale tæthedsfunktioners værdier, da de to karakterer jo er kraftigt korrelerede.

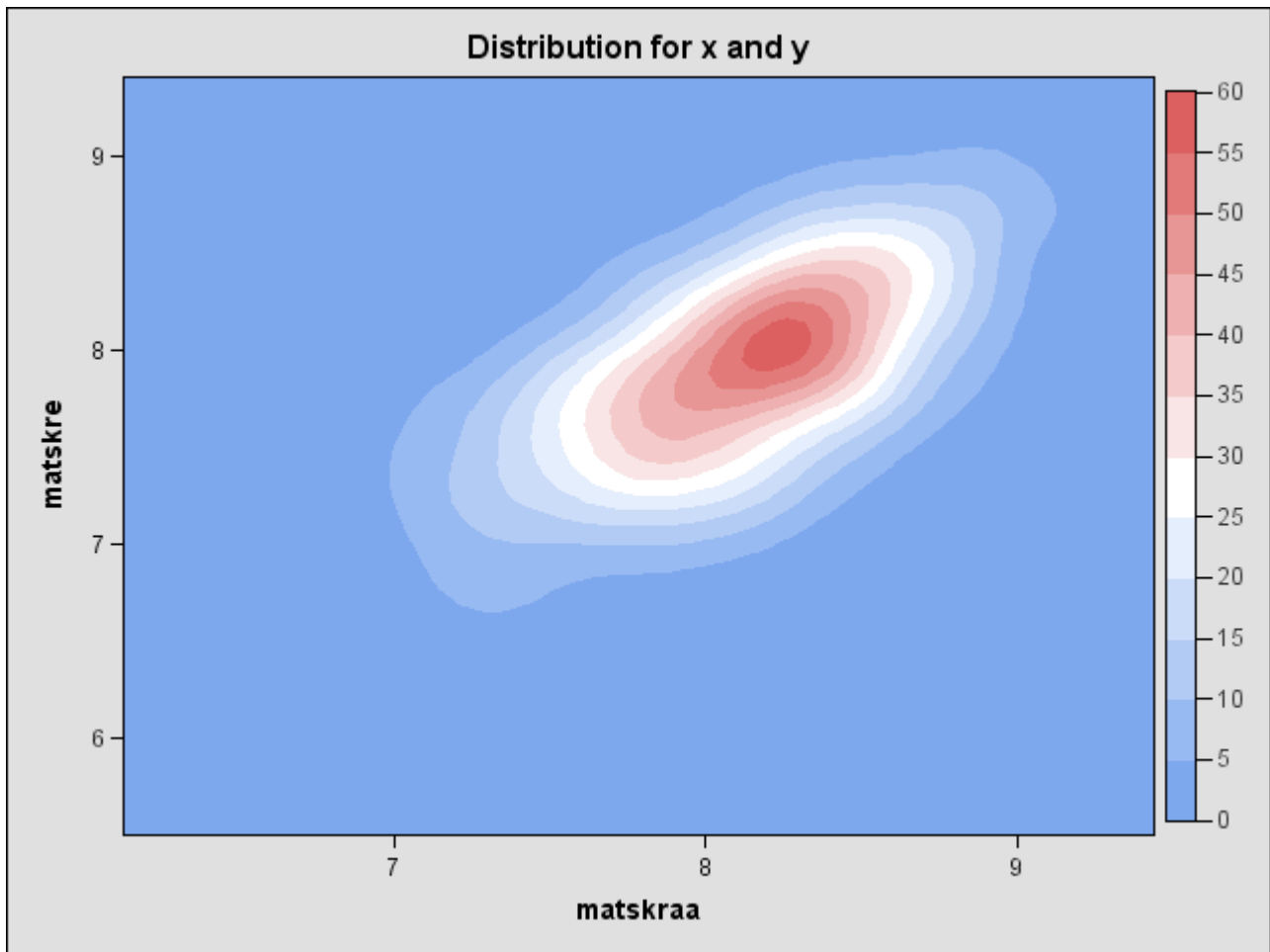
.

x	y = 5	y = 6	y = 7	y = 8
5	1.16714E-8	.000000345	0.00000	0.00000
6	6.75887E-8	.000315074	0.00046	0.00000
7	7.1754E-10	.000528161	0.12207	0.00886
8	1.3965E-14	.000001623	0.05923	0.67875

Den næste figur, der gør sig bedst i farver, viser et udglattet todimensionalt histogram for karaktererne, mens den følgende viser tæthedsfunktionen for den tilpassede todimensionale normalfordeling.



Niveaukurverne i den todimensionale tæthedsfunktion er de værdier af (x,y) , der giver samme værdi af tæthedsfunktionen. Den næste figur, der også bedst ses i farver, viser disse niveaukurver i datamaterialet.



I den todimensionale normalfordeling er niveaukurverne givet ved ligningerne

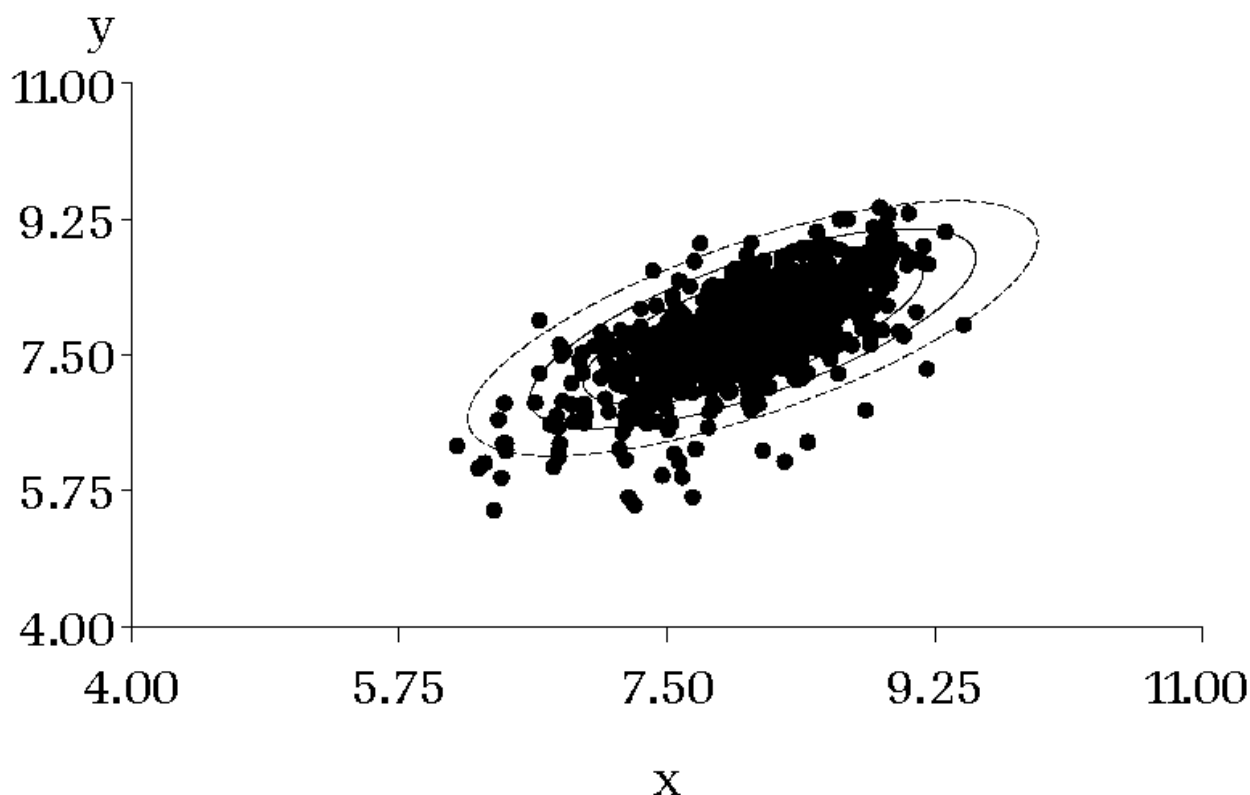
$$((x - \mu_1), (y - \mu_2)) \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{pmatrix}^{-1} \begin{pmatrix} (x - \mu_1) \\ (y - \mu_2) \end{pmatrix} = c$$

eller

$$\frac{(x - E[x])^2}{\text{var}(x)(1 - \text{corr}(x, y)^2)} + \frac{(y - E[y])^2}{\text{var}(y)(1 - \text{corr}(x, y)^2)} - 2(x - E[x])(y - E[y]) \frac{\text{corr}(x, y)}{\sqrt{\text{var}(x) \text{var}(y)(1 - \text{corr}(x, y)^2)}} = c$$

Dette er matematisk set ligningen for en ellipse, da kovariansmatricen er positivt definit. Ellipsen afgrænser områder omkring ellipsens centrum, der er punktet $(E[x], E[y])$, med sandsynligheder, der

afhænger af ellipsens størrelse. På figuren er disse ellipser tegnet svarende til 50% og 95% sandsynlighed, idet også de observerede punkter er indtegnet. Det betyder på nær en vis usikkerhed, at 50% af punkterne ligger indenfor den mindste ellipse, der dog er svær at se, og 95% indenfor den yderste.



På figuren er ellipsernes storakse tegnet. Teorien for ellipser (som ikke længere spiller en fremtrædende rolle i gymnasiets matematikundervisning) viser, at denne linie har hældningskoefficienten

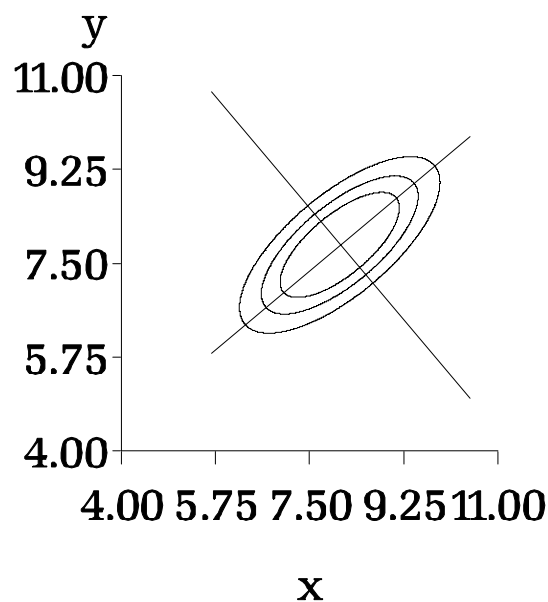
$$\beta = \frac{1}{2 \text{cov}(x, y)} \left[\text{var}(y) - \text{var}(x) \pm \sqrt{4 \text{cov}(x, y)^2 + (\text{var}(y) - \text{var}(x))^2} \right]$$

idet de to mulige fortegn angiver hældningen for storaksen og lilleaksen alt efter hvilken af de to varianser, der er størst. Ellipsernes akser går gennem midtpunktet, der er middelværdien - punktet $(E[x], E[y])$. I datamaterialet beregnes de to hældningskoefficienter til

$$\beta_1 = 0.84$$

og

$$\beta_2 = -1.19$$



De betingede fordelinger

I den todimensionale normalfordeling, der beskriver den todimensionale simultane fordeling af de to karakterer, kan man beregne de betingede fordelinger

Den betingede fordeling af eksamenskarakteren givet, at årskarakteren er lig med x , er normalfordelingen med middelværdi

$$(1) E[y|x] = E[y] + \text{corr}(x, y) \frac{\sqrt{\text{var}(y)}}{\sqrt{\text{var}(x)}} (x - E[x])$$

og varians

$$(2) \text{var}(y|x) = \text{var}(y)(1 - \text{corr}(x, y)^2)$$

Tabellen angiver værdier af den betingede tæthedsfunktion for y , når der betinges med $x = 5, 6, 7$ og 8 .

x	<i>dens5</i>	<i>dens6</i>	<i>dens7</i>	<i>dens8</i>
5	0.033542	1.11343	0.01161	0.00000
6	0.000064	0.32419	0.51349	0.00026
7	0.000000	0.00413	0.99450	0.07516
8	0.000000	0.00000	0.08431	0.96831

Den generelle formel, der binder betinget og ubetinget middelværdi sammen

$$E[y] = E[E[y|x]]$$

eftervises i det konkrete tilfælde, da sidste led i (1) har middelværdi nul.

Det ses umiddelbart af (2), at den betingede varians er mindre end den ubetingede varians af y , der jo er konstant lig med $\text{var}(y)$. Jo større den numeriske værdi af korrelationskoefficienten er, jo mindre er den betingede varians i forhold til n ubetingede.

Den generelle formel, der binder betinget og ubetinget varians sammen

$$\text{var}(y) = E[\text{var}[y|x]] + \text{var}(E[y|x])$$

eftervises i det konkrete tilfælde ved at

$$\text{var}(E[y | x]) = \text{corr}(x, y)^2 \frac{\text{var}(y)}{\text{var}(x)} \text{var}(x - E[x]) = \text{corr}(x, y)^2 \text{var}(y)$$

Som en kontrol af den fundamentale regneregul, at simultan tæthed er lig med betinget tæthed gange marginal tæthed kan tæthedsfunktionerne ganges sammen. F_x er den simultane tæthedsfunktion for $x = 8$ og $y = 7$ givet ved den marginale tæthedsfunktion af x for $x = 8$ ganget med den betingede tæthedsfunktion givet $x = 8$ for $y = 7$. I tal er det

$$0.05923 = 0.70397 \cdot 0.08431 .$$

Det ses, at den betingede middelværdi $E[y|x]$ som funktion af x er en ret linie, der går gennem punktet $(E[x], E[y])$ og har hældningskoefficienten

$$\text{corr}(x, y) \frac{\sqrt{\text{var}(y)}}{\sqrt{\text{var}(x)}}$$

Forlænges denne brøk med

$$\frac{\sqrt{\text{var}(x)}}{\sqrt{\text{var}(x)}}$$

ses, at hældningen er lig med regressionskoefficienten i en lineær regression med y som responsvariabel og x som forklarende variabel

$$\text{cov}(x, y) \frac{\text{var}(y)}{\text{var}(x)}$$

Linien, der svarer til en betingede middelværdi $E[x|y]$ af x givet y , er også et lineært udtryk

$$E[x|y] = E[x] + \text{corr}(x, y) \frac{\sqrt{\text{var}(x)}}{\sqrt{\text{var}(y)}} (y - E[y]).$$

Koefficienten til $(y - E[y])$

$$\text{corr}(x, y) \frac{\sqrt{\text{var}(x)}}{\sqrt{\text{var}(y)}}$$

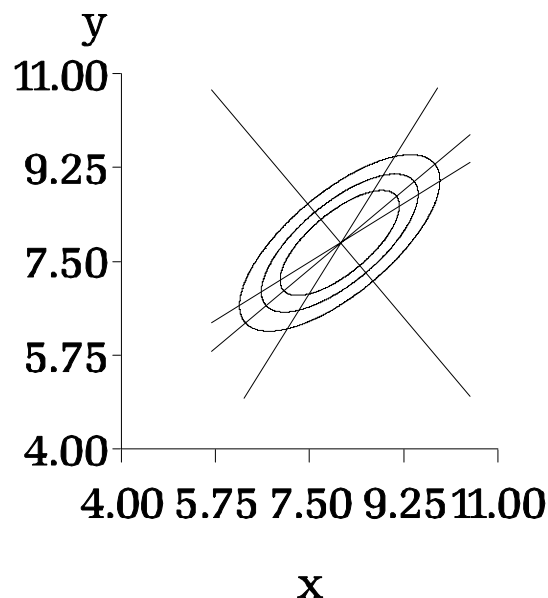
er regressionskoefficienten, når x regresseres på y .

Udtrykket for $E[x|y]$ angiver en linie, der går igennem punktet $(E[x], E[y])$ og som indtegnet i et koordinatsystem, hvor x som sædvanligt afsættes ud af den vandrette akse, har hældningen

$$\frac{1}{\text{corr}(x, y)} \frac{\sqrt{\text{var}(y)}}{\sqrt{\text{var}(x)}}.$$

Produktet af de to liniers hældningskoefficienter, når de betragtes i samme koordinatsystem er 1. Men de to umiddelbart estimerede regressionskoefficienter er IKKE hinandens inverse, hvad man ellers kunne forledes til at tro, idet de har produktet $\text{corr}(x, y)^2$.

På figuren er de to linier indtegnet sammen med ellipsernes storakser.



Principalkomponentanalyse

For en kovarians- og korrelationsmatrix kan egenverdier og egenvektorer beregnes. I matematiske fremstillinger står teorien ofte beskrevet som spektraldekomponering eller lignende. I dette tilfælde er egenverdierne af kovariansmatricen givet ved

0.49 og 0.08

og egenvektorerne er

$$\begin{pmatrix} 0.77 \\ 0.65 \end{pmatrix} \text{ og } \begin{pmatrix} -0.65 \\ 0.77 \end{pmatrix}$$

Matricen

$$\mathbf{A} = \begin{pmatrix} 0.77 & -0.65 \\ 0.65 & 0.77 \end{pmatrix}$$

er ortonormal, således at dens inverse blot er den transponerede.

$$\begin{pmatrix} 0.77 & -0.65 \\ 0.65 & 0.77 \end{pmatrix}^{-1} = \begin{pmatrix} 0.77 & -0.65 \\ 0.65 & 0.77 \end{pmatrix}^T = \begin{pmatrix} 0.77 & 0.65 \\ -0.65 & 0.77 \end{pmatrix}$$

Det betyder, at kovariansmatricen er faktoriseret ved

$$\mathbf{\Sigma} = \begin{pmatrix} 0.32 & 0.20 \\ 0.20 & 0.25 \end{pmatrix} = \begin{pmatrix} 0.77 & -0.65 \\ 0.65 & 0.77 \end{pmatrix} \begin{pmatrix} 0.49 & 0 \\ 0 & 0.08 \end{pmatrix} \begin{pmatrix} 0.77 & 0.65 \\ -0.65 & 0.77 \end{pmatrix}$$

og omvendt

$$\begin{pmatrix} 0.77 & -0.65 \\ -0.65 & 0.77 \end{pmatrix} \begin{pmatrix} 0.32 & 0.20 \\ 0.20 & 0.25 \end{pmatrix} \begin{pmatrix} 0.77 & -0.65 \\ 0.65 & 0.77 \end{pmatrix} = \begin{pmatrix} 0.49 & 0 \\ 0 & 0.08 \end{pmatrix}$$

For variablene x og y betyder dette, at

$$z_1 = 0.77x + 0.65y$$

og

$$z_2 = -0.65x + 0.77y$$

er ukorrelerede, da

$$\text{cov}(z_1, z_2) = -0.77 \cdot 0.65 \cdot 0.32 + 0.65 \cdot 0.77 \cdot 0.25 + (0.77^2 - 0.65^2) \cdot 0.20 = 0$$

på nær afrundingsfejl.

Varianserne af de transformerede variable er

$$\text{var}(z_1) = 0.77^2 \cdot 0.32 + 0.65^2 \cdot 0.25 + 2 \cdot 0.65 \cdot 0.77 \cdot 0.20 = 0.49$$

og

$$\text{var}(z_2) = 0.65^2 \cdot 0.32 + 0.77^2 \cdot 0.25 - 2 \cdot 0.65 \cdot 0.77 \cdot 0.20 = 0.08$$

på nær afrundingsfejl.

I vektorer udtrykkes disse resultater ved at

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 0.77 & -0.65 \\ 0.65 & 0.77 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{A}^{-1} \mathbf{x}$$

har kovariansmatricen

$$\mathbf{A}^{-1} \mathbf{\Sigma} \mathbf{A} = \begin{pmatrix} 0.49 & 0 \\ 0 & 0.08 \end{pmatrix}$$

Teorien for egenverdier for positivt definitte matricer siger bl.a., at de stokastiske variabel z_1 og z_2 defineret på denne måde har den størst respektive mindste varians blandt alle linearkombinationer af de stokastiske variable x og y , hvor kvadratsummen af koefficienter er lig med 1. Som en formel udtrykkes det ved

$$\max_{b_1^2 + b_2^2 = 1} \text{var}(b_1 x + b_2 y) = \text{største egenverdi} \quad \text{og} \quad \min_{b_1^2 + b_2^2 = 1} \text{var}(b_1 x + b_2 y) = \text{mindste egenverdi}$$

hvor koefficienterne b_1 og b_2 udgøre de respektive egenvektorer.

Det skal bemærkes, at en ene egenvektor z_1 stort set er proportional med gennemsnittet af x og y . Den er altså et udtryk for gennemsnittet af de to karakterer. Egenvektoren z_2 er tilsvarende proportional med differensen mellem x og y . Det ses, at $\text{var}(z_1)$ er stor, da der er stor forskel på skolernes matematiske niveau, mens $\text{var}(z_2)$ er lille, da de to karakterer er meget ens på alle skoler.

Der er en umiddelbar forbindelse mellem egenvektorerne og konturellipserne, idet egenvektorerne er parallelle med ellipsernes akser. Det ses ved, at forholdet mellem koordinaterne i egenvektorerne

$$\begin{pmatrix} 0.77 \\ 0.65 \end{pmatrix} \text{ og } \begin{pmatrix} -0.65 \\ 0.77 \end{pmatrix}$$

er hhv.

$$\beta_1 = 0.65/0.77 = 0.84$$

og

$$\beta_2 = 0.77/(-0.65) = -1.19$$

på nær afrundingsfejl.

Transformationen fra (x, y) til (z_1, z_2) udgør altså en bijektiv (enentydig) transformation af de korrelerede variable til et sæt af uafhængige variable. Ved denne transformation indeholder z_1 mest muligt af datamaterialets variation mens z_2 indeholder mindst, men da der kun er to variable, indeholder z_2 i dette tilfælde alt som z_1 ikke kan forklare.