HD Dataanalyse

Notesæt 2: Databeskrivelse, 2 variable

Copenhagen Business School

Emne i dette notesæt

Notesæt 1 beskæftiger sig med, hvorledes man ved hjælp af grafiske og numeriske værktøjer kan give en beskrivelse af ***ÉN variabel*** i et datamateriale.

Notesæt 2 beskæftiger sig på tilsvarende vis med, hvorledes man ved hjælp af grafiske og numeriske værktøjer kan give en beskrivelse af ***sammenhængen mellem TO variable*** i et datamateriale.

Vi er interesserede i at se på, hvordan to variable opfører sig ***i forhold til hinanden***, dvs. se på hvordan variablenes værdier varierer med hinanden.

Der er behov for forskellige metoder til analyse af sammenhænge mellem to variable, afhængig af hvilken type hver af de to variable er. Vi skal derfor se på følgende tre tilfælde:

- sammenhæng mellem to kategoriske variable
- sammenhæng mellem én kvantitativ og én kategorisk variabel
- sammenhæng mellem to kvantitative variable

Emne i dette notesæt

Eksempel: Boligpriser

Udsnit fra datafilen `Boligpriser.jmp`:

Kommune	Landsdel	Boligtype	Opførelsesår	BoligM2	Boligstørrelse	KælderM2	AntalRum	Salgspris
Kolding	Sønderjylland	Villa	1973	120	Mellem	0	4	900.000
Fredericia	Vestjylland	Villa	1948	135	Mellem	0	3	2.100.000
Fredericia	Vestjylland	Villa	1921	138	Mellem	69	5	1.850.000
Fredericia	Vestjylland	Villa	1947	146	Mellem	110	4	3.475.000
Fredericia	Vestjylland	Ejerlejlighed	1953	86	Lille	202	4	850.000

Datamaterialet i datafilen `Boligpriser.jmp` indeholder for hver af de solgte boliger blandt andet information om hvilken landsdel boligen er beliggende i, boligens type, boligens areal samt boligens salgspris.

Det virker umiddelbart rimeligt at forvente, at der er visse indbyrdes sammenhænge mellem disse variable, f.eks. at...

- visse boligtyper (f.eks. ejerlejligheder) i højere grad findes i visse landsdele (f.eks. Kbh. & Frederiksberg) end i andre (f.eks. Vestjylland)
  [dvs. en sammenhæng mellem boligtype og landsdel]
- boligens salgspris alt andet lige er højere i visse landsdele (f.eks. Kbh. & Frederiksberg) end i andre (f.eks. Bornholm)
  [dvs. en sammenhæng mellem salgspris og landsdel]
- boligens salgspris alt andet lige er højere for boliger med et større boligareal
  [dvs. en sammenhæng mellem salgspris og boligareal]

Hvordan vi mere præcist kan undersøge sådanne sammenhænge, er det vi nu skal se på.

Indholdsfortegnelse

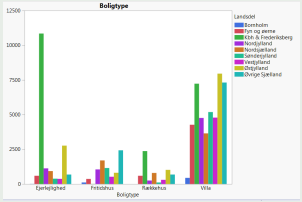
Når vi skal undersøge sammenhængen mellem ***to kategoriske variable***, kan vi gøre det...

- grafisk ved for hver værdi af den ene variabel at tegne et søjlediagram af fordelingen af den anden variabel
- numerisk ved at angive hyppigheden (absolut eller relativ) af hver mulig kombination af de to variables værdier i en tabel. En sådan tabel kaldes en **kontingenstabel** *("contingency table")* eller en **krydstabel**, fordi den "krydser" (dvs. kombinerer) de to variables værdier

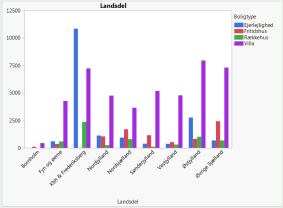
Eksempel: Boligpriser

Variablene *Landsdel* og *Boligtype* er begge kategoriske. For at undersøge en eventuel sammenhæng mellem dem kan vi for hver værdi af variablen *Boligtype* tegne et søjlediagram over fordelingen af *Landsdel*. Alternativt kan vi for hver værdi af variablen *Landsdel* tegne et søjlediagram over fordelingen af *Boligtype*.

Fordeling af *Landsdel* for hver værdi af *Boligtype*:



Fordeling af *Boligtype* for hver værdi af *Landsdel*:



Figuren til venstre viser eksempelvis, at blandt ejerlejligheder er langt størstedelen solgt i København og på Frederiksberg, mens salget af villaer er langt mere jævnt fordelt på de enkelte landsdele.

Figuren til højre viser eksempelvis, at i Vestjylland er langt størstedelen af de solgte boliger villaer, mens villaer udgør under halvdelen af det samlede antal solgte boliger i København og på Frederiksberg.

Principielt viser de to figurer præcis det samme. Men i praksis kan det være nyttigt at tegne begge figurer for at gøre det lettere at identificere eventuelle sammenhænge mellem de to variable.

JMP-video [Graph -> Graph Builder]

Eksempel: Boligpriser (fortsat)

For at få et mere præcist billede af sammenhængene mellem variablene *Landsdel* og *Boligtype* kan vi i stedet opstille en krydstabel på baggrund af datamaterialet.

Hvorvidt krydstabellen opstilles med absolutte eller relative hyppigheder er principielt underordnet. For nemheds skyld ser vi her først på tabellen baseret på *absolutte* hyppigheder.

Boligtype

Landsdel	Count	Ejerlejlighed	Fritidshus	Rækkehus	Villa	Total
	Bornholm	11	116	1	443	571
	Fyn og øerne	599	369	595	4273	5836
	Kbh & Frederiksberg	10839	14	2374	7230	20457
	Nordjylland	1132	1052	253	4758	7195
	Nordsjælland	935	1699	802	3650	7086
	Sønderjylland	384	1155	113	5186	6838
	Vestjylland	374	526	302	4782	5984
	Østjylland	2767	815	1019	7950	12551
	Øvrige Sjælland	685	2431	683	7308	11107
	Total	17726	8177	6142	45580	77625

Hver celle i tabellen viser den absolutte hyppighed (dvs. antallet af observationer) for den givne kombination af de to variable.

Eksempelvis viser tabellen, at der er solgt 10.839 ejerlejligheder i København og på Frederiksberg, og at der er solgt 4.782 villaer i Vestjylland.

Tabellens nederste række og søjlen længst til højre angiver totaler for hver af de to variable. Eksempelvis viser tabellen, at der er solgt i alt 17.726 ejerlejligheder på tværs af de forskellige landsdele, og at der er solgt i alt 20.457 boliger i København og på Frederiksberg.

JMP-video [Analyze -> Fit Y by X]

Eksempel: Boligpriser (fortsat)

Hvis krydstabellen i stedet opstilles med *relative* hyppigheder ser den således ud.

	Boligtype				
Landdel: Total %	Ejerlejlighed	Fritidshus	Rækkehus	Villa	Total
Bornholm	0,01	0,15	0,00	0,57	0,74
Fyn og øerne	0,77	0,48	0,77	5,50	7,52
Kbh & Frederiksberg	13,96	0,02	3,06	9,31	26,35
Nordjylland	1,46	1,36	0,33	6,13	9,27
Nordsjælland	1,20	2,19	1,03	4,70	9,13
Sønderjylland	0,49	1,49	0,15	6,68	8,81
Vestjylland	0,48	0,68	0,39	6,16	7,71
Østjylland	3,56	1,05	1,31	10,24	16,17
Øvrige Sjælland	0,88	3,13	0,88	9,41	14,31
Total	22,84	10,53	7,91	58,72	

Eneste forskel i forhold til tabellen på forrige side er, at tallene i hver celle her er divideret med det samlede antal observationer i datamaterialet (= 77.625).

Eksempelvis viser tabellen, at ud af det samlede boligsalg i hele landet udgøres 13,96% ($= \frac{10.839}{77.625}$) af ejerlejligheder solgt i København og på Frederiksberg, og 6,16% ($= \frac{4.782}{77.625}$) af villaer solgt i Vestjylland.

Ligeledes ses, at af det samlede boligsalg i hele landet udgøres 22,84% ($= \frac{17.726}{77.625}$) af ejerlejligheder og 26,35% ($= \frac{20.457}{77.625}$) af boliger solgt i København og på Frederiksberg.

Eksempel: Boligpriser (fortsat)

Ved opstilling af en krydstabel i JMP kan man også få tabellen opskrevet med ***søjlevist*** *relative* hyppigheder.

Boligtype

Landsdel / Col %	Ejerlejlighed	Fritidshus	Rækkehus	Villa
Bornholm	0,06	1,42	0,02	0,97
Fyn og øerne	3,38	4,51	9,69	9,37
Kbh & Frederiksberg	61,15	0,17	38,65	15,86
Nordjylland	6,39	12,87	4,12	10,44
Nordsjælland	5,27	20,78	13,06	8,01
Sønderjylland	2,17	14,12	1,84	11,38
Vestjylland	2,11	6,43	4,92	10,49
Østjylland	15,61	9,97	16,59	17,44
Øvrige Sjælland	3,86	29,73	11,12	16,03

Tabellen indeholder relative hyppigheder men på en sådan måde, at tallene i hver ***søjle*** summerer til 100%. Tabellen skal derfor læses ***hver søjle for sig***, dvs. hver boligtype for sig.

Eksempelvis viser tabellen, at af samtlige solgte ejerlejligheder blev 0,06% ($=\frac{11}{17.726}$) solgt på Bornholm, 3,38% ($=\frac{599}{17.726}$) solgt på Fyn og øerne, 61,15% ($=\frac{10.839}{17.726}$) solgt i København og på Frederiksberg osv.

De ni sandsynligheder summerer (pånær afrundingsfejl) til 100% (0,06%+3,38%+61,15%+6,39%+...+3,86% = 100,00%) og angiver dermed fordelingen på de forskellige landsdele blandt samtlige solgte ejerlejligheder.

På samme måde viser hver af tabellens øvrige søjler fordelingen af variablen *Landsdel* for hver af de øvrige boligtyper. Hver søjle svarer til én bestemt boligtype.

Bemærk: På side 7 beregnes de relative hyppigheder relativt til boligsalget ***i hele landet***, mens de på denne side beregnes relativt til boligsalget ***indenfor hver boligtype*** (dvs. hver søjle for sig).

Eksempel: Boligpriser (fortsat)

Ved opstilling af en krydstabel i JMP kan man også få tabellen opskrevet med ***rækkevist*** *relative* hyppigheder.

Boligtype

Landdel (Row %)	Ejerlejlighed	Fritidshus	Rækkehus	Villa
Bornholm	1,93	20,32	0,18	77,58
Fyn og øerne	10,26	6,32	10,20	73,22
Kbh & Frederiksberg	52,98	0,07	11,60	35,34
Nordjylland	15,73	14,62	3,52	66,13
Nordsjælland	13,20	23,98	11,32	51,51
Sønderjylland	5,62	16,89	1,65	75,84
Vestjylland	6,25	8,79	5,05	79,91
Østjylland	22,05	6,49	8,12	63,34
Øvrige Sjælland	6,17	21,89	6,15	65,80

Tabellen indeholder relative hyppigheder men på en sådan måde, at tallene i hver ***række*** summerer til 100%. Tabellen skal derfor læses ***hver række for sig***, dvs. hver landsdel for sig.

Eksempelvis viser tabellen, at af samtlige boliger solgt i København og på Frederiksberg var 52,98% ($=\frac{10.839}{20.457}$) ejerlejligheder, 0,07% ($=\frac{14}{20.457}$) var fritidshuse, 11,60% ($=\frac{2.374}{20.457}$) var rækkehuse og 35,34% ($=\frac{7.230}{20.457}$) var villaer.

De fire sandsynligheder summerer (pånær afrundingsfejl) til 100% (52,98%+0,07%+11,60%+35,34% = 100,00%) og angiver dermed fordelingen på de forskellige boligtyper blandt samtlige boliger solgt i København og på Frederiksberg.

På samme måde viser hver af tabellens øvrige rækker fordelingen af variablen *Boligtype* i hver af de øvrige landsdele. Hver række svarer til én bestemt landsdel.

Bemærk: På side 7 beregnes de relative hyppigheder relativt til boligsalget ***i hele landet***, mens de på denne side beregnes relativt til boligsalget ***i hver landsdel*** (dvs. hver række for sig).

Eksempel: Boligpriser (fortsat)

Hvis man ønsker det, kan man i JMP få opstillet alle fire typer tabeller i én og samme udskrift.

Boligtype

Isdel	Count Total % Col % Row %	Ejerlejlighed	Fritidshus	Rækkehus	Villa	Total
	Bornholm	11 0,01 0,06 1,93	116 0,15 1,42 20,32	1 0,00 0,02 0,18	443 0,57 0,97 77,58	571 0,74
	Fyn og øerne	599 0,77 3,38 10,26	369 0,48 4,51 6,32	595 0,77 9,69 10,20	4273 5,50 9,37 73,22	5836 7,52
	Kbh & Frederiksberg	10839 13,96 61,15 52,98	14 0,02 0,17 0,07	2374 3,06 38,65 11,60	7230 9,31 15,86 35,34	20457 26,35
	Nordjylland	1132 1,46 6,39 15,73	1052 1,36 12,87 14,62	253 0,33 4,12 3,52	4758 6,13 10,44 66,13	7195 9,27

Hver celle består her af tallene fra hver af de fire foregående tabeller. Øverst i hver celle står den absolutte hyppighed, herefter står den relative hyppighed, herefter den søjlevist relative hyppighed, og nederst den rækkevist relative hyppighed.

Eksempelvis viser tabellen, at for ejerlejligheder solgt i København og på Frederiksberg, er den absolutte hyppighed 10.839, den relative hyppighed 13,06%, den søjlevist relative hyppighed 61,15%, og den rækkevist relative hyppighed 52,98%.

Hvorvidt man foretrækker at have alle tal samlet i én tabel eller opdelt i fire separate tabeller er udelukkende en smagssag.

Når vi skal undersøge sammenhængen mellem ***én kvantitativ og én kategorisk variabel***, kan vi gøre det...

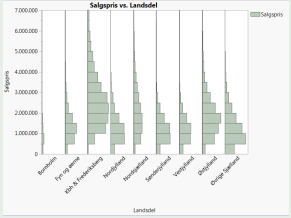
- grafisk ved for *hver* værdi af den *kategoriske* variabel at tegne et histogram og/eller et CDF plot og/eller et box plot af fordelingen af den *kvantitative* variabel
- numerisk ved for *hver* værdi af den *kategoriske* variabel at angive udvalgte nøgletal for fordelingen af den *kvantitative* variabel

Hvis den ***kategoriske*** variabel...

- kun har ganske *få mulige værdier*, kan man uden problemer sammenligne histogrammer eller CDF plots af fordelingen af den kvantitative variabel for hver værdi af den kvantitative variabel
- har *mere end blot nogle få mulige værdier*, skal man sammenligne så mange forskellige histogrammer / CDF plots, at det i praksis er umuligt at overskue. I de tilfælde nøjes man som regel med at sammenligne box plots, fordi det kan gøres langt mere overskueligt

Eksempel: Boligpriser

Variablen *Landsdel* er kategorisk og variablen *Salgspris* er kvantitativ. For at undersøge en eventuel sammenhæng mellem dem kan vi for hver værdi af variablen *Landsdel* tegne et histogram over fordelingen af *Salgspris*.



Figuren viser fordelingen af de solgte boligers priser ***indenfor hver landsdel***. Figuren kan imidlertid virke noget uoverskuelig, fordi den rummer 9 forskellige histogrammer (hver drejet 90 grader) - ét svarende til hver landsdel.

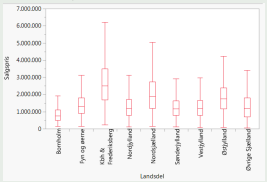
Selv om figuren er svær at overskue, kan den dog tjene det formål at vise, hvorvidt de enkelte fordelinger ser ud til at være nogenlunde klokkeformede, således at vi kan gøre brug af den empiriske regel.

Generelt ser fordelingen af boligernes salgspris ud til at være højreskæv (dvs. tendens til at enkelte boliger er solgt til langt højere priser end alle øvrige). For enkelte landsdele er skævheden stor, eksempelvis Nordsjælland, mens den for andre er noget mindre, eksempelvis Vestjylland.

JMP-video [Graph -> Graph Builder]

Eksempel: Boligpriser (fortsat)

Hvis vi for hver værdi af *Landsdel* i stedet tegner et (quantile) boxplot over fordelingen af *Salgspris*, får vi en mere illustrativ figur (NB: om de enkelte boxplots tegnes lodret - som i figuren her - eller vandret er fuldstændig ligegyldigt).



Figuren viser ligeledes fordelingen af de solgte boligers priser ***indenfor hver landsdel***, men er mindre detaljeret og dermed umiddelbart lettere at overskue.

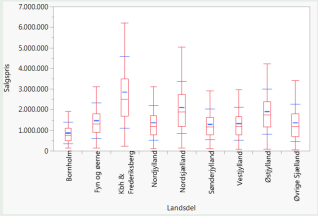
Til figuren kan vi desuden for hver landsdel yderligere tilføje følgende tre værdier (indtegnet med blåt i figuren på næste side):

- gennemsnit minus 1 standardafvigelse ($\overline{x} - s$)
- gennemsnit ($\overline{x}$)
- gennemsnit plus 1 standardafvigelse ($\overline{x} + s$)

Værdierne svarer til venstre endepunkt, midtpunktet og højre endepunkt i det interval $\left[\overline{x} - s;\ \overline{x} + s\right]$, som ifølge den empiriske regel indeholder ca. 68% af observationerne af *Salgspris* indenfor hver landsdel (forudsat at den empiriske regel finder anvendelse).

Eksempel: Boligpriser (fortsat)

Box plots af fordelingen af *Salgspris* samt 68%-interval fra den empiriske fordeling optegnet for hver værdi af *Landsdel*:



På baggrund af ovenstående figur kan vi eksempelvis se, at salgsprisen for boliger solgt i Vestjylland har...

- en 1. kvartil på ca. 800.000 ($= q_{25\%}$)
- en median på ca. 1.200.000 ($= q_{50\%}$)
- en 3. kvartil på ca. 1.700.000 ($= q_{75\%}$)
- et gennemsnit på ca. 1.300.000 ($= \overline{x}$)
- et interval der indeholder ca. 68% af observationerne på ca. [500.000; 2.100.000] ($= [\overline{x} - s;\ \overline{x} + s]$)

  (Bemærk: Vi udnytter her, at vi ovenfor har set, at fordelingen af boligernes salgspriser i Vestjylland er nogenlunde klokkeformet, således at den empiriske regel med en vis rimelighed kan benyttes.)

Eksempel: Boligpriser (fortsat)

Ved at sammenligne de forskellige box plots får man en sammenligning af boligernes salgspriser i de forskellige landsdele.

Eksempelvis kan man af figuren på forrige side se, at salgsprisen på boliger solgt i København og på Frederiksberg...

- generelt er højere end i de øvrige landsdele (median og gennemsnit ligger højere end i de øvrige box plots)
- er mere varierende end i de øvrige landsdele (interkvartilbredden er større end i de øvrige box plots)

For at få et mere præcist billede kan vi eksplicit beregne de nøgletal, der er afbilledet i figuren på forrige side (dvs. gennemsnit, median, standardafvigelse og interkvartilbredde samt evt. udvalgte fraktiler).

Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
Bornholm	157000	375000	505000	750000	1100000	1499000	5000000
Fyn og øerne	146000	618500	915000	1300000	1800000	2500000	10500000
Kbh & Frederiksberg	220000	1195000	1695000	2495000	3500000	4850000	25750000
Nordjylland	130000	550000	800000	1200000	1725000	2400000	14500000
Nordsjælland	140000	771400	1200000	1900000	2750000	3681500	12000000
Sønderjylland	125000	525000	775000	1160000	1637500	2200500	10995000
Vestjylland	84000	525000	800000	1200000	1675000	2245500	13900000
Østjylland	100000	735000	1175000	1749000	2400000	3225000	15750000
Øvrige Sjælland	70000	450000	705000	1185000	1795000	2506000	9500000

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Bornholm	571	875703	531910	22260	831981,56	919423,87
Fyn og øerne	5836	1476042	867556	11356	1453779,7	1498305,1
Kbh & Frederiksberg	20457	2857174	1734735	12129	2833400,9	2880947,1
Nordjylland	7195	1371425	848090	9998	1351825,6	1391024,8
Nordsjælland	7086	2116868	1265379	15032	2087400,2	2146335,1
Sønderjylland	6838	1292989	736711	8909	1275524,9	1310454
Vestjylland	5984	1330785	793234	10254	1310683,3	1350887,5
Østjylland	12551	1911900	1099021	9810	1892670,5	1931128,6
Øvrige Sjælland	11107	1366701	894979	8492	1350055,4	1383347,4

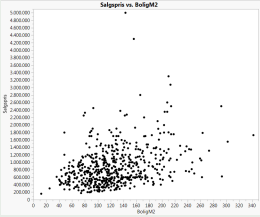
På baggrund af nøgletallene kan man foretage en mere præcis vurdering af ligheder og forskelle i salgsprisen på tværs af de enkelte landsdele.

Når vi skal undersøge sammenhængen mellem ***to kvantitative variable***, kan vi gøre det...

- grafisk ved at tegne en figur med værdierne for den ene variable ud af 1. aksen og værdierne for den anden variabel ud af 2. aksen. En sådan figur kaldes for et **scatterplot**.
- numerisk ved at beregne nøgletallet ***korrelation***.

Eksempel: Boligpriser

Variablene *Salgspris* og *BoligM2* er begge kvantitative. For at undersøge en eventuel sammenhæng mellem dem kan vi tegne deres værdier op i et scatterplot. For nemheds skyld ser vi her kun på observationer af boliger solgt på Bornholm.



Figuren viser, at der er en positiv sammenhæng mellem boligareal og salgspris, således at desto større areal en bolig har, desto dyrere er den og vice versa. Figuren viser ligeledes, at sammenhængen mellem boligareal og salgspris er nogenlunde lineær, dvs. at punkterne med en vis tilnærmelse ligger omkring en ret linje (hvad det mere konkret har af betydning, kommer vi tilbage til senere i kurset).

Endelig viser figuren ikke overraskende, at det ikke kun er boligens areal, der er afgørende for boligens salgspris. For et given boligareal ser vi, at boliger er solgt til meget forskellige priser. Eksempelvis er boliger med et boligareal på ca. 95 m^2 solgt til priser helt ned til 200.000 kr. og helt op til 2.500.000 kr. Der er således andet end boligens areal, der er bestemmende for boligens salgspris.

▶ JMP-video [Graph -> Graph Builder]

Definition (korrelation)

For kvantitative variable X og Y med observerede værdier $x_1,\ldots,x_n$ hhv. $y_1,\ldots,y_n$ beregnes variablenes indbyrdes **korrelation** *("correlation") som*

$$r_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i-\bar{x}}{s_X}\cdot\frac{y_i-\bar{y}}{s_Y}$$

hvor $\bar{x},\bar{y}$ er variablenes gennemsnit

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i$$

og s_X, s_Y er variablenes standardafvigelser

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}, \qquad s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

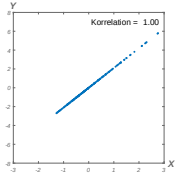
Bemærk:

Korrelationen $r_{X,Y}$ antager per konstruktion ALTID en værdi mellem -1 og $+1$.

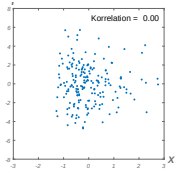
Korrelation

Der findes tre særligt vigtige tilfælde, som har fået deres egne navne:

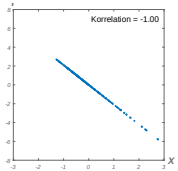
- Hvis korrelationen er 1 ligger de to variables værdier på en ret linje med ***positiv*** hældning, og de to variable siges at være **perfekt positivt korreleret**.



- Hvis korrelationen er 0 ligger de to variables værdier uden nogen synlig sammenhæng, og de to variable siges at være **ukorreleret**.

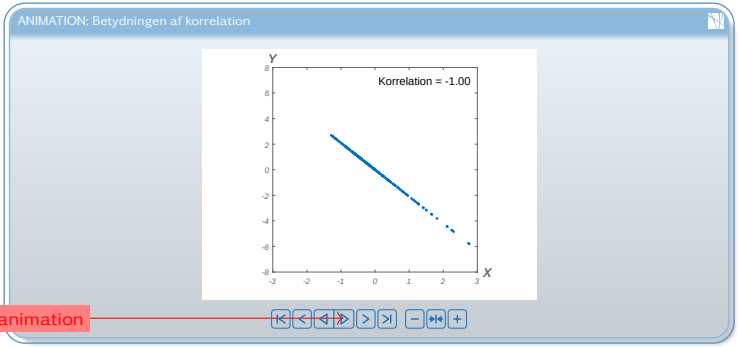


- Hvis korrelationen er -1 ligger de to variables værdier på en ret linje med ***negativ*** hældning, og de to variable siges at være **perfekt negativt korreleret**.



Korrelation

Det følgende konstruerede eksempel viser, hvorledes sammenhængen mellem to variable – kaldet X og Y – ændres i takt med at værdien af deres indbyrdes korrelation ændres.



Korrelation

Korrelationen $r_{X,Y}$ mellem to variable X og Y...

- måler hvor meget de to variable X og Y ***varierer med hinanden***
- er et mål for i hvor høj grad de to variable X og Y har en ***lineær*** sammenhæng med hinanden
  (dvs. måler i hvor høj grad værdierne af X og Y ligger på en ret linje, hvis man tegner dem op i et scatterplot)
- er per konstruktion præcis den samme som korrelationen $r_{Y,X}$ mellem Y og X
  (dvs. når man beregner korrelation er variablenes rækkefølge ligegyldig)
- måler ***kun*** graden af ***lineær sammenhæng*** mellem X og Y
  (dvs. selvom $r_{X,Y} = 0$ kan der alligevel godt være en sammenhæng mellem X og Y, blot er sammenhængen så ikke lineær, men kan f.eks. godt være kvadratisk, altså $Y = X^2$)
- siger ***kun*** noget om ***samvariationen*** mellem X og Y men ***ingenting om kausalitet*** (= årsagssammenhæng) mellem X og Y.
  (dvs. siger kun noget om, hvorvidt X og Y har en sammenhæng men ikke noget om, hvorvidt det er X, der påvirker Y eller omvendt)
- afhænger ikke af hvilken enhed variablene X og Y måles i
  (dvs. korrelationen $r_{X,Y}$ mellem f.eks. Y="boligens salgspris" og X="boligens areal" er den samme, uanset om Y måles i kr., i 1.000 kr. eller i mio. kr., og uanset om X måles i mm^2, cm^2 eller m^2)
- er følsom overfor eventuelle outliers i datamaterialet (på samme måde som gennemsnit $\bar{x}$ og standardafvigelse s)

Korrelation

Hvis korrelationen $r_{X,Y}$ er ***positiv*** (dvs. $r_{X,Y} > 0$)...

- indikerer det en ***positiv*** lineær sammenhæng mellem X og Y
  (dvs. der vil være en tendens til, at variablene vil ligge på en ret linje med *positiv* hældning, hvis man tegner dem op i et scatterplot)
- Desto ***mere positiv*** korrelationen $r_{X,Y}$ er (dvs. desto tættere den er på +1), desto ***stærkere*** er den lineære sammenhæng mellem X og Y
  (dvs. i desto *højere* grad vil variablene have en tendens til at ligge på en ret linje med positiv hældning i et scatterplot)
- Desto ***mindre positiv*** korrelationen $r_{X,Y}$ er (dvs. desto tættere den er på 0), desto ***svagere*** er den lineære sammenhæng mellem X og Y
  (dvs. i desto *mindre* grad vil variablene have en tendens til at ligge på en ret linje med positiv hældning i et scatterplot)

Hvis korrelationen $r_{X,Y}$ er ***negativ*** (dvs. $r_{X,Y} < 0$)...

- indikerer det en ***negativ*** lineær sammenhæng mellem X og Y
  (dvs. der vil være en tendens til, at variablene vil ligge på en ret linje med *negativ* hældning, hvis man tegner dem op i et scatterplot)
- Desto ***mere negativ*** korrelationen $r_{X,Y}$ er (dvs. desto tættere den er på -1), desto ***stærkere*** er den lineære sammenhæng mellem X og Y
  (dvs. i desto højere grad vil variablene have en tendens til at ligge på en ret linje med negativ hældning i et scatterplot)
- Desto ***mindre negativ*** korrelationen $r_{X,Y}$ er (dvs. desto tættere den er på 0), desto ***svagere*** er den lineære sammenhæng mellem X og Y
  (dvs. i desto mindre grad vil variablene have en tendens til at ligge på en ret linje med negativ hældning i et scatterplot)

Korrelation

Eksempel: Boligpriser

Ved at beregne korrelationen mellem *Salgspris* og *BoligM2* kan vi måle graden af lineær afhængighed mellem de to variable. Vi ser for nemheds skyld fortsat kun på boliger solgt på Bornholm.

Correlations

	Salgspris	BoligM2
Salgspris	1,0000	0,3492
BoligM2	0,3492	1,0000

Korrelationen mellem variablene *Salgspris* og *BoligM2* er 0,3492. Korrelationen er således positiv, hvilket understreger indtrykket fra scatterplottet af en positiv sammenhæng mellem de to variable. Ligeledes er korrelationen noget mindre end 1, hvilket understreger at der langt fra er en perfekt lineær sammenhæng mellem de to variable (svarende til at punkterne i scatterplottet langt fra lå helt perfekt på en ret linje).

Bemærk: Når JMP beregner korrelationer, beregner den korrelationer mellem alle mulige kombinationer af de variable, man har angivet som input. Det betyder, at man skal læse og forstå JMPs ovenstående korrelationsoutput på følgende måde:

- Det første tal angiver, at korrelationen mellem *Salgspris* og *Salgspris* er 1 (fordi de to variable er identiske)
- Det andet tal angiver, at korrelationen mellem *Salgspris* og *BoligM2* er 0,3492
- Det tredje tal angiver, at korrelationen mellem *BoligM2* og *Salgspris* er 0,3492 (husk: rækkefølgen af variablene er ligegyldig, når man beregner korrelation)
- Det fjerde tal angiver, at korrelationen mellem *BoligM2* og *BoligM2* er 1 (fordi de to variable er identiske)

JMP-video [Analyze -> Multivariate Methods -> Multivariate]

Eksempel: Indkomst

Udsnit fra datafilen Indkomst.jmp:

Indkomstår	Frederiksberg	Lemvig
2011	327	265
2012	336	266
2013	345	280
2014	353	279
2015	364	291

For variablene *Lemvig* og *Frederiksberg*, der indeholder den gennemsnitlige årlige indkomst (i 1.000 kr.) for personer bosiddende i henholdsvis Lemvig og Frederiksberg kommune, er gennemsnit og standardafvigelse givet som

$$\begin{aligned} \text{Lemvig} \quad &: \quad \overline{x} = 276,2, \ s_x = 10,85 \\ \text{Frederiksberg} \quad &: \quad \overline{y} = 345,0, \ s_y = 14,40 \end{aligned}$$

Frederiksberg

Summary Statistics

Mean	345
Std Dev	14,40486
N	5

Lemvig

Summary Statistics

Mean	276,2
Std Dev	10,848963
N	5

Korrelationen mellem de to variable kan herefter beregnes

$$\begin{aligned} r_{X,Y} &= \frac{1}{5-1} \sum_{i=1}^{5} \frac{x_i - \overline{x}}{s_x} \cdot \frac{y_i - \overline{y}}{s_y} \\ &= \frac{1}{4} \Big(\frac{265-276,2}{10,85} \cdot \frac{327-345,0}{14,40} + \frac{266-276,2}{10,85} \cdot \frac{336-345,0}{14,40} \\ &\quad + \frac{280-276,2}{10,85} \cdot \frac{345-345,0}{14,40} + \frac{279-276,2}{10,85} \cdot \frac{353-345,0}{14,40} \\ &\quad + \frac{291-276,2}{10,85} \cdot \frac{364-345,0}{14,40} \Big) = 0,9550 \end{aligned}$$

hvilket også stemmer overens med beregningen i JMP.

Correlations

	Frederiksberg	Lemvig
Frederiksberg	1,0000	0,9550
Lemvig	0,9550	1,0000

Den høje positive korrelation på 0,9550 mellem de to variable viser, at udviklingen i indkomstfordelingen har været næsten identisk i de to kommuner over den betragtede 5-årige periode.

Definition (uafhængighed)

To variable X og Y kaldes ***uafhængige*** *("independent"), hvis variablene ikke påvirker hinandens fordelinger.*

Et eksempel på variable, der er uafhængige, er i forbindelse med kast med en terning.

Hvis en terning kastes to gange og variablen X angiver antallet af øjne i 1. terningkast, og Y angiver antallet af øjne i 2. terningkast, så er X og Y uafhængige.

De to variable er uafhængige fordi værdien af X ikke påvirker fordelingen af Y (og vice versa). Der er sandsynlighed $1/6$ for hver af de seks muligheder i 2. terningkast (fordelingen af Y), uanset resultatet af 1. terningkast (værdien af X)).

Uafhængighed

Begrebet korrelation måler graden af (lineær) *afhængighed* mellem to variable, og begrebet er derfor nært knyttet til begrebet *uafhængighed*.

Resultat [uafhængighed og korrelation]

Hvis to variable X og Y er uafhængige, så er deres indbyrdes korrelation $r_{X,Y} = 0$.

(dvs. uafhængighed medfører korrelation = 0)

Selvom den indbyrdes korrelation mellem to variable X og Y er $r_{X,Y} = 0$, så er variablene ikke nødvendigvis uafhængige.

(dvs. korrelation = 0 medfører ikke uafhængighed)

Fordi korrelation kun måler graden af ***lineær*** afhængighed mellem to variable, kan man godt have at korrelationen mellem to variable er 0, men at de alligevel på den ene eller anden måde er indbyrdes afhængige (dvs. ikke er uafhængige).

At to variable er uafhængige er således et stærkere udsagn, end at de blot har en indbyrdes korrelation på 0.

1 Sammenhæng mellem to kategoriske variable

2 Sammenhæng mellem én kvantitativ og én kategorisk variabel

3 Sammenhæng mellem to kvantitative variable

4 OPSUMMERING

Kort opsummering af dette notesæt:

Sammenhængen mellem to kategoriske variable kan undersøges v.hj.a. søjlediagrammer og en krydstabel.

Sammenhængen mellem en kvantitativ og en kategorisk variabel kan undersøges v.hj.a. boxplots og beregning af nøgletal (gennemsnit, median, standardafvigelse, interkvartilbredde osv.).

Sammenhængen mellem to kvantitative variable kan undersøges v.hj.a. et scatterplot og beregning af korrelation.

Korrelationen mellem to variable X og Y...

- antager altid en værdi mellem -1 og +1
- måler graden af lineær afhængighed mellem X og Y
- siger ikke noget om kausaliteten mellem X og Y

Hvis korrelationen mellem to variable X og Y er...

- positiv, er der en tendens til en positiv lineær sammenhæng mellem X og Y, og jo mere positiv desto stærkere er sammenhængen
- negativ, er der en tendens til en negativ lineær sammenhæng mellem X og Y, og jo mere negativ desto stærkere er sammenhængen
- nul, er der ingen tendens til lineær sammenhæng mellem X og Y

Hvis X og Y er uafhængige er korrelationen mellem dem 0. Men selv om korrelationen mellem X og Y er 0, er variablene ikke nødvendigvis uafhængige.

INDEKS

Nye funktionaliteter i dette notesæt:

- *Analyze -> Fit Y by X:*
  - Krydstabel (for 2 kategoriske variable)
  - Box plots (for én kvant.var. opdelt efter en kat.var.)
- *Analyze -> Multivariate Methods -> Multivariate:*
  - Korrelation
- *Graph -> Graph Builder:*
  - Histogrammer (for én kvant.var. opdelt efter en kat.var.)
  - Scatterplot
  - Søjlediagrammer (for én kat.var. opdelt efter en anden kat.var.)

JMP-videoer: