HD Dataanalyse

Notesæt 1: Databeskrivelse, 1 variabel

Copenhagen Business School

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Emne i dette notesæt

Første skridt i en analyse af et datamateriale er at skabe sig et overblik over indholdet af datamaterialet.

I praksis indeholder datamaterialer ofte mange variable og et stort antal observationer. Det gør det svært at overskue, hvilken information der kan uddrages af et datamateriale.

Derfor er der behov for værktøjer, der kort og præcist kan give et overblik over indholdet af et stort og omfattende datamateriale.

Notesæt 1 beskæftiger sig med, hvorledes man ved hjælp af grafiske og numeriske (dvs. talmæssige) værktøjer kan opsummere indholdet af et datamateriale, uanset dets størrelse.

Emne i dette notesæt

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

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 information om salget af 77.625 boliger (adresse, opførelsesår, størrelse, salgstidspunkt, pris m.m.).

Datamaterialet er så stort, at det er umuligt at overskue med det blotte øje.

Derfor er der behov for værktøjer, der kort og præcist kan opsummere de væsentligste karakteristika i datamaterialet.

Indholdsfortegnelse

1. Variabeltyper
2. Kategorisk data – grafisk og numerisk beskrivelse
3. Kvantitativt data – grafisk beskrivelse
4. Kvantitativt data – numerisk beskrivelse
5. OPSUMMERING

1 Variabeltyper

Kvantitativ/kategorisk • Kontinuert/diskret • Ordinal/nominal • Beskrivelse af data

2 Kategorisk data – grafisk og numerisk beskrivelse

3 Kvantitativt data – grafisk beskrivelse

4 Kvantitativt data – numerisk beskrivelse

5 OPSUMMERING

KVANTITATIV/KATEGORISK

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Når vi skal beskrive en variabel i et datamateriale, skelner vi mellem to forskellige typer af variable, afhængig af om variablens værdier har en naturlig talmæssig fortolkning eller ej.

En variabel kaldes...

- **kvantitativ** *("quantitative"/"numerical")* hvis dens værdier mest naturligt angives som ***tal***, dvs. hvis variablen kan måles i en eller anden form for meningsfuld enhed: meter, kilo, kr., dage o.lign.

  (variable som f.eks. *højde, vægt, indkomst, ventetid*)

- **kategorisk** *("categorical"/"qualitative")* hvis dens værdier mest naturligt angives v.hj.a. ***bogstaver/ord***

  (variable som f.eks. *køn, øjenfarve, kundetilfredshed, uddannelsesmæssig baggrund*)

Kort sagt:
Alle de variable vi enten kan måle eller tælle er ***kvantitative*** variable, mens alle øvrige variable er ***kategoriske*** variable.

KVANTITATIV/KATEGORISK

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

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

Variablen *BoligM2* er en kvantitativ variabel, der angiver boligens areal i m^2. *AntalRum* og *Salgspris* er andre eksempler på kvantitative variable.

Variablen *Kommune* er en kategorisk variabel, der angiver hvilken kommune boligen er beliggende i. *Landsdel* og *Boligtype* er andre eksempler på kategoriske variable.

BEMÆRK: En kategorisk variabels (bogstav)værdier kan altid omdefineres til passende valgte talværdier. Det ændrer imidlertid ***ikke*** variablen fra at være kategorisk til kvantitativ.

Eksempel: Boligpriser

Variablen *Kommune* er fortsat kategorisk, selv hvis vi erstatter de enkelte kommunernes navne med talværdier, f.eks. 1 = Kolding, 2 = Fredericia osv. Det skyldes, at de valgte talværdier ikke har nogen naturlig talmæssig fortolkning (det er f.eks. meningsløst at beregne og fortolke et gennemsnit af de valgte talværdier).

KVANTITATIV/KATEGORISK

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Der er behov for forskellige metoder til analyse af henholdsvis kvantitative og kategoriske variable. Derfor er det vigtigt at være opmærksom på, hvilken af de to typer en given variabel er.

For ***kvantitative*** variable giver det mening at lave beregninger med variablens værdier (eksempelvis lægge værdier sammen eller trække værdier fra hinanden), mens den slags ikke giver mening for ***kategoriske*** variable.

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

Det er muligt at beregne et gennemsnit af værdierne af den kvantitative variabel *BoligM2* for at finde det gennemsnitlige boligareal blandt de betragtede boliger.

Det er ikke muligt at beregne et gennemsnit af værdierne af den kategoriske variabel *Kommune*.

KVANTITATIV/KATEGORISK

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

I visse tilfælde afhænger det af den konkrete sammenhæng, om en variabel opfattes som kvantitativ eller kategorisk.

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

Variablen *Opførelsesår* kan både opfattes som kvantitativ og kategorisk. Det giver umiddelbart fin mening at lave beregninger med variablens værdier (f.eks. trække to værdier fra hinanden for at beregne forskellen i alder på to boliger).

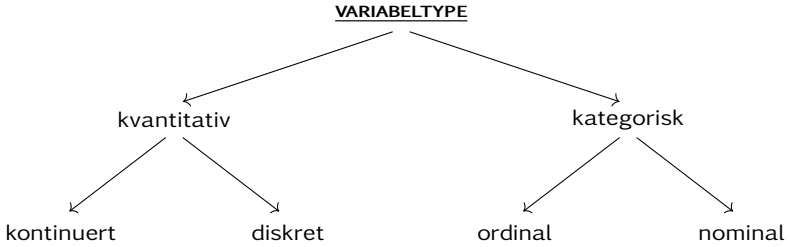
Hvis man vil lave beregninger med variablens værdier, svarer det til at opfatte variablen som kvantitativ. Hvis man i stedet blot vil bruge variablen til at gruppere boligerne efter, hvornår de er opført, svarer det til at opfatte variablen som kategorisk.

Variablen *Salgsdato* er et andet eksempel på en variabel, der både kan opfattes som kvantitativ og kategorisk.

KVANTITATIV/KATEGORISK

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Kvantitative og kategoriske variable opdeles hver især i to undergrupper af variable:



Databehandlingsprogrammet JMP angiver variabeltyperne ved følgende symboler:

- : kvantitativ (kontinuert eller diskret)
- : ordinal
- : nominal

BEMÆRK: JMP skelner ikke mellem kontinuerte og diskrete variable, fordi man i praksis ofte kan lave de samme analyser for de to variabeltyper.

KONTINUERT/DISKRET

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Når vi skal beskrive indholdet af en ***kvantitativ*** variabel i et datamateriale, skelner vi mellem to forskellige typer af kvantitative variable.

En kvantitativ variabel kaldes...

- **kontinuert** *("continuous")* hvis dens mulige værdier kan ligge vilkårligt tæt på hinanden (variable som f.eks. *højde, vægt, indkomst*)
- **diskret** *("discrete")* hvis dens mulige værdier ***ikke*** kan ligge vilkårligt tæt på hinanden
  (variable som f.eks. *antal biler pr. husstand, antal fejlbehæftede varer pr. leverance, antal telefonopkald pr. dag*)

For en ***diskret*** variabel har de mulige værdier ***afstand*** mellem sig
(f.eks. har variablen *antal biler pr. husstand* de mulige værdier 0, 1, 2, 3 osv., og der er således en afstand på 1 mellem de mulige værdier)

En diskret variabel er ofte en variabel, der måler "et antal" af noget
(biler, fejlbehæftede varer, telefonopkald osv.)

For en ***kontinuert*** variabel har de mulige værdier ***ikke afstand*** mellem sig
(f.eks. har variablen *indkomst* alle positive tal som sine mulige værdier, og de mulige værdier ligger således uendelig tæt på hinanden)

KONTINUERT/DISKRET

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

I visse tilfælde vil vi tænke på en ***diskret*** variabel, *som om* den er kontinuert, fordi det kan gøre det lettere at analysere variablen.

Hvis en diskret variabel har...

- ***få*** mulige værdier, vil vi som regel tænke på den som ***diskret*** (fordi det typisk vil være vigtigt med en tydelig skelnen mellem på hinanden følgende værdier af variablen)
- ***mange*** mulige værdier, vi vi som regel tænke på den som ***kontinuert*** (fordi det typisk *ikke* vil være vigtigt med en tydelig skelnen mellem på hinanden følgende værdier af variablen)

KONTINUERT/DISKRET

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

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

Variablen *AntalRum* angiver antal rum i boligen. Variablens mulige værdier er 0, 1, 2 osv. Der er således en afstand (på 1) mellem variablens mulige værdier, og variablen klassificeres derfor som diskret.

Langt de fleste boliger har et begrænset antal rum (og variablen *AntalRum* har derfor et begrænset antal mulige værdier). Det gør det mest naturligt at klassificere variablen som diskret.

Variablen *Salgspris* angiver boligens salgspris. Eftersom de mulige værdier principielt kan være alle positive tal, så ligger de mulige værdier vilkårligt tæt på hinanden, og variablen klassificeres derfor som kontinuert.

Selv hvis variablen *Salgspris* opgøres i hele kr., og der således er en afstand på 1 kr. mellem de mulige værdier, er der fortsat mange mulige værdier, og vi vil derfor alligevel tænke på det, som om de mulige værdier ligger vilkårligt tæt på hinanden.

Det er ikke vigtigt at skelne mellem, om salgsprisen på en bolig eksempelvis er 2.000.000 kr. eller 2.000.001 kr. (= to på hinanden følgende værdier), og derfor er det mest naturligt at klassificere variablen *Salgspris* som kontinuert.

ORDINAL/NOMINAL

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Når vi skal beskrive indholdet af en ***kategorisk*** variabel i et datamateriale, skelner vi mellem to forskellige typer af kategoriske variable.

En kategorisk variabel kaldes...

- **ordinal** *("ordinal")* hvis dens værdier på naturlig vis kan ordnes i rækkefølge, f.eks. "god" -> "bedre" -> "bedst" eller "lille" -> "mellem" -> "stor"

  (variable som f.eks. *kundetilfredshed, arbejdsindsats, forventninger til fremtidig vækst*)
- **nominal** *("nominal")* hvis dens værdier ***ikke*** på naturlig vis kan ordnes i rækkefølge

  (variable som f.eks. *køn, øjenfarve, uddannelse*)

Kort sagt:
En kategorisk variabel er ***ordinal***, hvis *rækkefølgen* af dens værdier har en betydning, og ellers er den ***nominal***.

ORDINAL/NOMINAL

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

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

Variablen *Kommune* angiver hvilken kommune, boligen er beliggende i. Variablens værdier (kommunenavne) kan ikke umiddelbart på nogen naturlig måde ordnes i rækkefølge, og variablen *Kommune* klassificeres derfor som nominal.

Variablen *Boligstørrelse* med de mulige værdier "Lille", "Mellem" og "Stor" (defineret som boliger med boligareal på henholdsvis under 100 m^2, mellem 100 og 200 m^2 og over 200 m^2) kan på naturlig vis ordnes i rækkefølge enten som "Lille" -> "Mellem" -> "Stor" eller som "Stor" -> "Mellem" -> "'Lille". Variablen *Boligstørrelse* klassificeres derfor som ordinal.

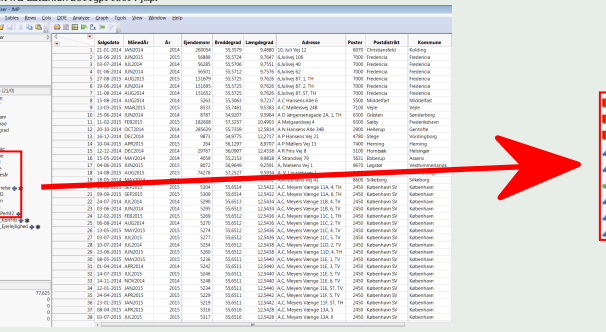
ORDINAL/NOMINAL

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Eksempel: Boligpriser

I JMP kan variabeltypen for hver enkelt variabel i datasættet aflæses direkte i programmet:

Udsnit fra datafilen `Boligpriser.jmp`:



Eksempelvis kan man se, at JMP opfatter variablen *Kommune* som nominal, variablen *Boligstørrelse* som ordinal, variablen *Salgspris* som kvantitativ osv.

BESKRIVELSE AF DATA

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Når vi skal beskrive en variabel i et datamateriale, er det en god idé at kombinere grafiske og numeriske (dvs. talmæssige) værktøjer.

Vi vil som regel starte med at lave én eller figurer af variablens observationer i datamaterialet og herefter udregne én eller flere numeriske størrelser, der kan give en mere specifik opsummering af variablens observationer.

Formålet med både figurer og numeriske størrelser er at få et overblik over variablens **fordeling** *("distribution")*, dvs. få et overblik over hvordan variablen "fordeler sig" på sine mulige værdier.

Variablens ***fordeling*** er defineret som information om...

- ***hvilke værdier*** variablen kan antage
- ***hvor ofte*** vi observerer de enkelte værdier

Når man ser på en variabels fordeling baseret på observationer af variablen fra et konkret datamateriale, vil vi ofte kalde det for variablens ***empiriske fordeling*** (til forskel fra en variabels teoretiske fordeling, som vi siden hen skal se på).

BESKRIVELSE AF DATA

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Vi kan analysere en variabels empiriske fordeling ved at se på **hyppigheden** af hver af variablens mulige værdier i form af...

- den ***absolutte*** hyppighed *("frequency")* (= antal gange hver værdi optræder i datasættet)
- den ***relative*** hyppighed *("relative frequency")* (= antal gange hver værdi optræder i datasættet divideret med det samlede antal observationer i datasættet)

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

Variablen *Kommune* har bl.a. de mulige værdier Kolding og Fredericia.

Den empiriske fordeling af variablen *Kommune* angiver for hver af de mulige værdier, hvor mange gange værdien optræder i datamaterialet. Eksempelvis viser værdien Kolding sig at optræde med en hyppighed på i alt 989 gange i datamaterialet svarende til en relativ hyppighed på 0,01274 ($= \frac{989}{77.625}$, idet datamaterialet består af i alt 77.625 observationer).

BESKRIVELSE AF DATA

Variabeltyper
Kvant/kategorisk
Kontinuert/diskret
Ordinal/nominal
Beskrivelse
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

For en ***kategorisk*** variabel kan variablens empiriske fordeling beskrives grafisk v.hj.a....

- et **søjlediagram** *("bar graph")*
- et **cirkeldiagram** *("pie chart")*

og numerisk ved hjælp af en **hyppighedstabel** *("frequency table")*.

For en ***kvantitativ*** variabel kan variablens empiriske fordeling beskrives grafisk v.hj.a....

- et **histogram** *("histogram")*
- et **CDF plot** *("cumulative density function")*
- et **box plot** *("box plot")*

og numerisk ved hjælp af en række forskellige nøgletal.

I det følgende vil vi først se på beskrivelse af den empiriske fordeling for en kategorisk variabel, og dernæst se på beskrivelse af den empiriske fordeling for en kvantitativ variabel.

1 Variabeltyper

2 Kategorisk data – grafisk og numerisk beskrivelse

3 Kvantitativt data – grafisk beskrivelse

4 Kvantitativt data – numerisk beskrivelse

5 OPSUMMERING

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

For en ***kategorisk*** variabel giver en ***hyppighedstabel*** en oversigt over, hvor mange gange (dvs. hvor hyppigt) vi har observeret variablens forskellige mulige værdier.

Det er ligegyldigt, om man opstiller hyppighedstabellen på baggrund af den absolutte eller den relative hyppighed (fordi den ene blot er en skalering af den anden).

For at skabe sig et overblik over en kategorisk variabel, kan man grafisk illustere hyppighedstabellen ved hjælp af enten et ***søjlediagram*** eller et ***cirkeldiagram***.

Ofte er søjlediagrammet lettere at aflæse end cirkeldiagrammet, men principielt er de to diagramtyper akkurat lige anvendelige.

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Eksempel: Boligpriser

Variablen *Landsdel* er en kategorisk variabel, der angiver i hvilken landsdel den solgte bolig er beliggende. Variablens hyppighedstabel angiver hvordan datamaterialets 77.625 solgte boliger er fordelt på de forskellige landsdele.

Frequencies

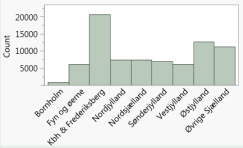
Level	Count	Prob
Bornholm	571	0,00736
Fyn og øerne	5836	0,07518
Kbh & Frederiksberg	20457	0,26354
Nordjylland	7195	0,09269
Nordsjælland	7086	0,09129
Sønderjylland	6838	0,08809
Vestjylland	5984	0,07709
Østjylland	12551	0,16169
Øvrige Sjælland	11107	0,14309
Total	77625	1,00000

N Missing 0
9 Levels

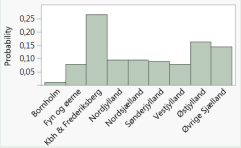
Tabellen her viser både den absolutte hyppighed og den relative hyppighed. Eksempelvis viser tabellen, at 571 (= absolut hyppighed) af de solgte boliger er beliggende på Bornholm, svarende til andel på $\frac{571}{77.625} = 0{,}00736$ (= relativ hyppighed), dvs. 0,736%, af samtlige solgte boliger.

Søjlediagrammer af henholdsvis de absolutte og de relative hyppigheder viser, at de giver anledning til præcis samme figur. Figurerne adskiller sig udelukkende ved hvilke værdier, der står op ad 2. aksen.

Absolut hyppighed:



Relativ hyppighed:



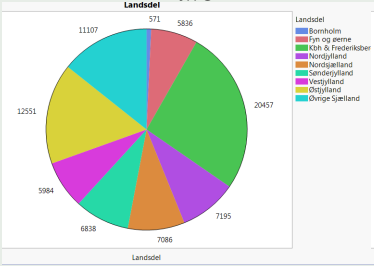
JMP-video [Analyze -> Distribution ; Graph -> Graph Builder]

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

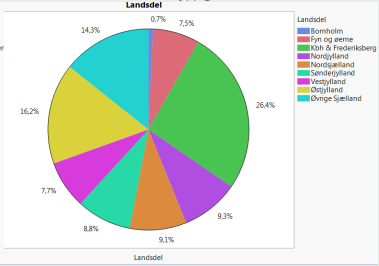
Eksempel: Boligpriser (fortsat)

Som alternativ til søjlediagrammerne kan man i stedet illustrere hyppighederne i cirkeldiagrammer.

Absolut hyppighed:



Relativ hyppighed:



Det er nøjagtig den samme information som i søjlediagrammerne blot repræsenteret på en anden måde.

1 Variabeltyper

2 Kategorisk data – grafisk og numerisk beskrivelse

3 Kvantitativt data – grafisk beskrivelse

Histogram • *Fraktil* • *CDF plot* • *Box plot*

4 Kvantitativt data – numerisk beskrivelse

5 OPSUMMERING

Histogram

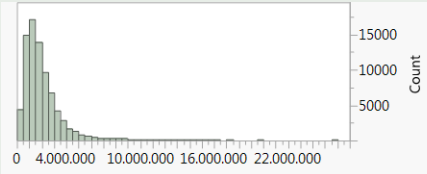
Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Et ***histogram*** giver et overblik over, hvor ofte vi observerer en ***kvantitativ*** variabels forskellige mulige værdier, dvs. et overblik over variablens fordeling.

Eksempel: Boligpriser

Variablen *Salgspris* er en kvantitativ variabel, og vi kan derfor tegne et histogram over dens værdier. Histogrammet er en figur, der opsummerer samtlige datamaterialets 77.625 observationer af variablen *Salgspris*.

Histogrammet inddeler de 77.625 observerede salgspriser i intervaller efter størrelse (i eksemplet her er længden af intervallerne 500.000 kr.).



Højden af søjlerne angiver hvor mange observationer, der findes i hvert af intervallerne.

Af histogrammet kan man blandt andet se, at der er solgt flest boliger til priser mellem 1.000.000 kr. og 1.500.000 kr.

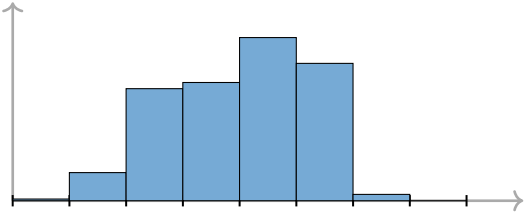
JMP-video [Analyze -> Distribution]

Histogram

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Et histogram konstrueres på følgende måde:

1. Inddel 1. aksen i et passende antal intervaller
2. Tæl for hvert interval antallet af observationer i det pågældende interval
3. Tegn for hvert interval en søjle op ad 2. aksen, der angiver den absolutte (eller relative) hyppighed



Bemærk: Det er ligegyldigt, om man tegner den absolutte eller den relative hyppighed op ad 2. aksen. Det giver anledning til præcis samme figur (eneste forskel er hvilke talværdier, der står på 2. aksen).

Bemærk: JMP vælger automatisk en passende intervalinddeling af 1. aksen, når programmet tegner histogrammer.

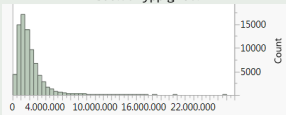
Histogram

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

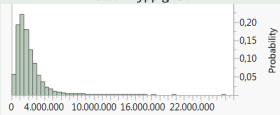
Eksempel: Boligpriser

Histogrammet for variablen *Salgspris* kan tegnes på to forskellige måder, afhængig af hvad der tegnes op ad 2. aksen (absolut eller relativ hyppighed):

Absolut hyppighed:



Relativ hyppighed:



De to figurer er ens og viser dermed præcis det samme om fordelingen af variablen *Salgspris*. Figurerne adskiller sig udelukkende ved hvilke værdier, der står op ad 2. aksen.

Eksempelvis viser begge histogrammer, at der er solgt flest boliger til en pris mellem 1.000.000 kr. og 1.500.000 kr. Helt præcist er der solgt i alt 17.003 boliger (= absolut hyppighed) til en pris mellem 1.000.000 kr. og 1.500.000 kr., svarende til en andel på $\frac{17.003}{77.625} = 21{,}90\%$ (= relativ hyppighed) af samtlige solgte boliger.

Ligeledes viser begge histogrammer eksempelvis (hvis man ser godt efter), at langt de fleste boliger er solgt til en pris under 5.000.000 kr., men at der også er solgt enkelte boliger til en pris over 15.000.000 kr.

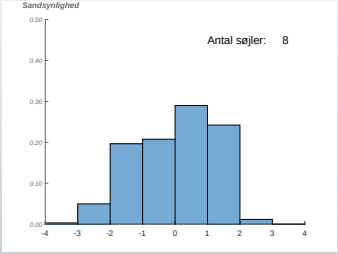
Histogram

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

For datasæt med mange observationer vil en variabels histogram minde mere og mere om en pæn glat kurve i takt med, at man øger antallet af søjler i histogrammet.

Nedenstående konstruerede eksempel viser, hvorledes histogrammet ændrer sig for en variabel med 1.000.000 observationer i intervallet fra -4 til 4 i takt med, at antallet af søjler i histogrammet øges.

ANIMATION: Fra histogram til glat kurve



Start animation

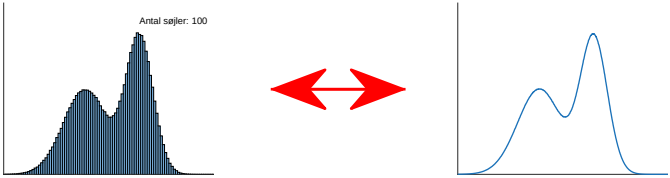
Histogram

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Fordelingen af en variabel er således i praksis beskrevet ved en pæn glat kurve, hvis blot datamaterialet er stort nok, og såfremt inddelingen af 1. aksen i intervaller er tilstrækkeligt fin, når man tegner sit histogram.

Når man derfor...

- ser på en variabel i et konkret *datamateriale*, så tegner man variablens empiriske fordeling ved hjælp af et *histogram*.
- diskuterer variablens generelle *teoretiske egenskaber*, tegner man i stedet variablens fordeling ved hjælp af en pæn *glat kurve*.



Man kan tænke på den glatte kurve som en teoretisk abstraktion af histogrammet.

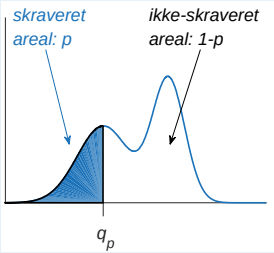
FRAKTIL

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

For en ***kvantitativ*** variabel er det i mange tilfælde nyttigt at beskrive variablens fordeling ved hjælp af begrebet "fraktil".

Definition (fraktil)

*En variabels p-**fraktil** ("percentile"/"quantile") q_p er den talværdi, som en variabel med sandsynlighed p er mindre end.*



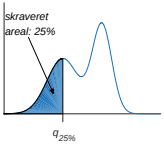
I et datamateriale er en variabels p-fraktil q_p den talværdi, som lige netop en andel p af variablens observationer er mindre end.

Fraktil

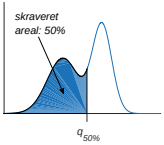
Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

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

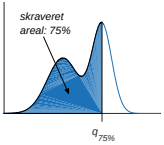
- 25%-fraktilen ($q_{25\%}$) kaldes også **1. kvartil**:



- 50%-fraktilen ($q_{50\%}$) kaldes også 2. kvartil eller **medianen**:



- 75%-fraktilen ($q_{75\%}$) kaldes også **3. kvartil**:



Bemærk: Af og til anvendes også notationen $Q1$ for 1. kvartil, $Q2$ for medianen (2. kvartil) og $Q3$ for 3. kvartil.

Fraktil

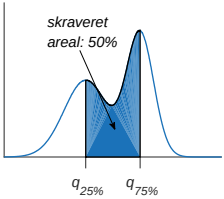
Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Hvis man ordner alle observationerne i et datamateriale efter størrelse fra mindst til størst, så er...

- 1. kvartil ($q_{25\%}$) den værdi, som netop 25% af alle observationerne er mindre end
- medianen ($q_{50\%}$) den værdi, som netop 50% af alle observationerne er mindre end. Med andre ord, medianen er den midterste observation, når data ordnes efter størrelse
- 3. kvartil ($q_{75\%}$) den værdi, som netop 75% af alle observationerne er mindre end

Desuden ser man ofte også på forskellen mellem 1. og 3. kvartil, der kaldes **inter-kvartilbredden** *("interquartile range")*

$$\text{IQR} = q_{75\%} - q_{25\%}$$



Fraktil

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Eksempel: Boligpriser

Softwareprogrammet JMP beregner automatisk en række forskellige fraktiler. For variablen *SalgsprisPerM2* er resultatet følgende:

Quantiles

100.0%	maximum	59898,477157
99.5%		49264,705882
97.5%		40425,531915
90.0%		31111,111111
75.0%	quartile	22313,432836
50.0%	median	14662,921348
25.0%	quartile	9349,4542822
10.0%		6000
2.5%		3444,4444444
0.5%		2338,1294964
0.0%	minimum	2000

Eksempelvis fremgår det, at 25% af samtlige solgte boliger er solgt til en kvadratmeterpris på 9.349 kr. eller derunder, og at 50% (dvs. halvdelen) af boligerne er solgt til en kvadratmeterpris på 14.663 kr. eller derunder.

Ligeledes kan vi se, at 99,5% af boligerne er solgt til en kvadratmeterpris på 49.265 kr. eller derunder, eller med andre ord: 0,5% er solgt til en kvadratmeterpris på 49.265 kr. eller *derover*.

Udover ovenstående fraktiler, som er dem JMP beregner som standard, kan man beregne yderligere fraktiler efter behov. Som eksempel er her beregnet 80%-fraktilen:

Custom Quantiles

Quantiles

Quantile	Estimate	Lower 95%	Upper 95%	Actual Coverage
80%	24656	24528,3	24811,3	95,01

Det fremgår således, at 80% af samtlige solgte boliger er solgt til en kvadratmeterpris på 24.656 kr. eller derunder (det er tallet under "Estimate", der skal bruges).

JMP-video [Analyze -> Distribution]

Fraktil

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Fordelingen af en ***kvantitativ*** variabel i et datamateriale kan illustreres ved hjælp af et histogram og kan ligeledes angives numerisk ved at beregne variablens fraktiler.

Der er således en tæt sammenhæng mellem en variabels histogram og en variabels fraktiler.

Mere præcist gælder, at hvis vi...

- opfatter histogrammets samlede ***areal*** som lig med 100%
- og læser histogrammet "fra venstre mod højre"

så kan vi (i princippet) aflæse variablens fraktiler blot ved at se på 1. aksen i histogrammet.

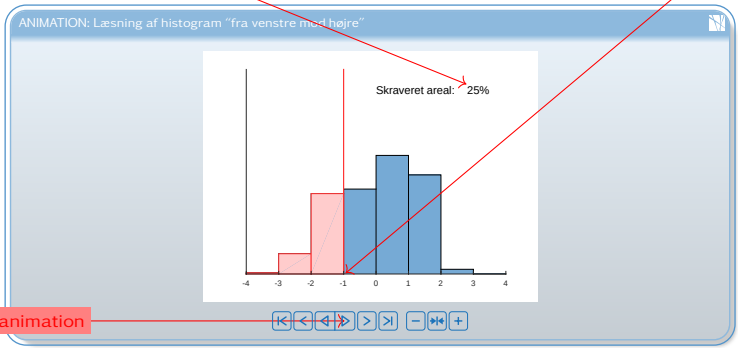
Fraktil

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Nedenstående konstruerede eksempel viser, hvorledes man af histogrammet kan aflæse, hvor stor en andel af en variabels observationer (= det rødt skraverede område), der er mindre end en given værdi (= den lodrette røde streg).

Med andre ord så viser eksemplet, hvordan man (i princippet) kan aflæse en variabels fraktiler ud fra 1. aksen i variablens histogram.

Eksempelvis er 25% af variablens observationer mindre end værdien -1 i nedenstående eksempel (dvs. $q_{25\%} = -1$).



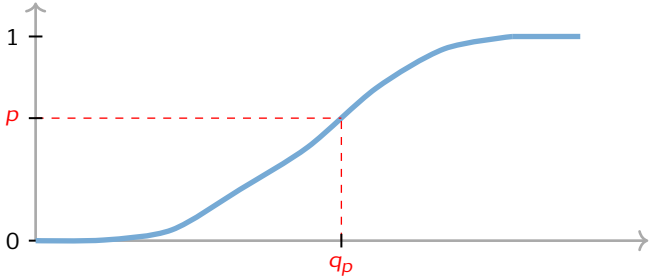
CDF plot

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Et ***CDF plot*** giver – ligesom et histogram – et grafisk overblik over en ***kvantitativ*** variabels fordeling.

CDF plottet repræsenterer i princippet præcis samme information som histogrammet, men illustrerer det blot på en anden måde.

CDF plottet fremkommer ved at tegne alle værdier af p fra 0 til 1 op ad 2. aksen og tegne de tilhørende q_p-fraktiler ud ad 1. aksen.



Til forskel fra et histogram, hvor det i praksis er svært at aflæse en variabels fraktiler, er det lige til at aflæse fraktiler i et CDF plot.

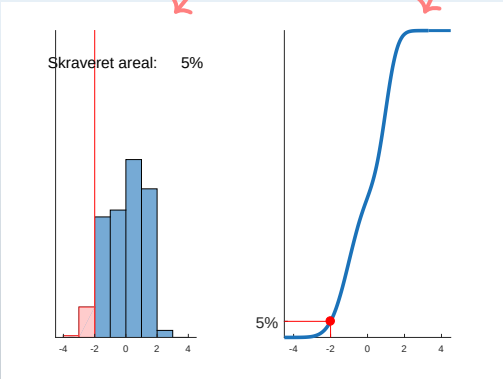
CDF PLOT

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Nedenstående konstruerede eksempel for en variabel med værdier mellem -4 og 4 viser sammenhængen mellem en variabels histogram og dens CDF plot.

Eksempelvis er 5% af variablens observationer mindre end værdien -2 i nedenstående eksempel (dvs. $q_{5\%} = -2$).

ANIMATION: Sammenhæng mellem histogram og CDF plot



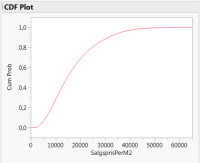
Start animation

CDF plot

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

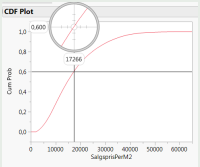
Eksempel: Boligpriser

Variablen *SalgsprisPerM2* har nedenstående CDF plot:



Figuren viser, at størstedelen af de solgte boliger har en kvadratmeterpris på mellem 2.500 kr. og 50.000 kr., idet der stort set er 0% sandsynlighed til venstre for 2.500 kr., og der stort er 100% sandsynlighed, når vi når til 50.000 kr.

Hvis man aflæser specifikke værdier i CDF plottet, kan man eksempelvis se, at 60%-fraktilen ($q_{60\%}$) er cirka 17.266 kr.



Med andre ord, så har 60% af de solgte boliger en kvadratmeterpris på 17.266 kr. eller derunder.

▶ JMP-video [Analyze -> Distribution]

Box plot

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Et ***box plot*** giver – ligesom et histogram og et CDF plot – et grafisk overblik over en ***kvantitativ*** variabels fordeling.

Et box plot opsummerer nogle få, udvalgte karakteristika omkring variablens fordeling, og giver derfor et mindre detaljeret overblik end et histogram/CDF plot.

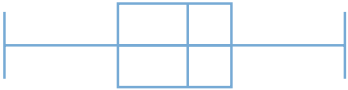
Den mindre detaljerede beskrivelse betyder omvendt, at det ofte er lettere at sammenligne fordelinger for flere forskellige variable på én gang ved at sammenligne deres box plots frem for ved at sammenligne deres histogrammer/CDF plots.

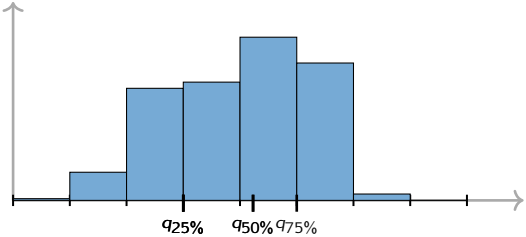
Box plot

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

På baggrund af en variabels ***histogram*** konstrueres samme variabels ***box plot*** ved at lave en figur med følgende værdier:

- den værdi, som 25% af observationerne er mindre end (25%-fraktilen)
- den værdi, som 50% af observationerne er mindre end (50%-fraktilen)
- den værdi, som 75% af observationerne er mindre end (75%-fraktilen)





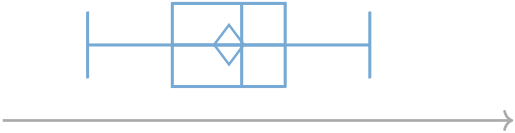
De tre værdier (lodrette streger) forbindes herefter med vandrette linjer, så der hermed er tegnet en figur (*"plot"*) indeholdende en kasse (*"box"*).

Box plot

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

De tre primære værdier (lodrette streger) i et box plot er 25%-, 50%- og 75%-fraktilerne, som beskriver:

- fordelingens ***centrum*** (50%-fraktilen)
- fordelingens ***variation*** (interkvartilbredden = afstanden mellem 75%- og 25%-fraktilerne)



Desuden vises datamaterialets gennemsnit som midtpunktet i en diamant.

Udover disse værdier indeholder et box plot yderligere en eller flere værdier, der beskriver fordelingens største og mindste observationer, fordelingens såkaldte **haler**, samt særlige ekstraordinært store eller ekstraordinært små værdier, såkaldte **outliers**.

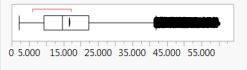
Hvilke værdier der specifikt er tale om varierer lidt, afhængig af præcis hvordan man tegner sit box plot. Softwareprogrammet JMP tegner eksempelvis to forskellige slags box plots: "outlier box plot" og "quantile box plot".

Box plot

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Eksempel: Boligpriser

Et ***outlier box plot*** for variablen *SalgsprisPerM2* har følgende udseende (her med en intervalinddeling på 2.500 kr.):



Outlier box plottet viser 25%- (≈ 9.000 kr.), 50%- (≈ 15.000 kr.) og 75%-fraktilen (≈ 22.500 kr.) i fordelingen af variablen *SalgsprisPerM2*. Derudover indeholder outlier box plottet bl.a. den mindste (≈ 2.500 kr.) og den største (≈ 42.500 kr.) af alle de observationer i datamaterialet, som er indeholdt i intervallet

$$\left[q_{25\%} - 1,5 \cdot IQR;\ q_{75\%} + 1,5 \cdot IQR\right]$$

Ideen er at opfatte alle observationer der ligger i dette interval, som de værdier af variablen *SalgsprisPerM2* vi umiddelbart vil forvente at observere. Enhver observation, der *ikke* befinder sig i dette interval, er eksplicit optegnet med en sort prik i outlier box plottet og kan opfattes som en potentiel outlier. På den måde giver outlier box plottet en idé om, hvor mange ekstraordinært store hhv. små observationer (dvs. outliers), der findes i datamaterialet.

Konklusionen på baggrund af outlier box plottet er, at...

- det generelle niveau for kvadratmeterprisen på solgte boliger er ca. 15.000 kr. pr kvm.
- kvadratmeterprisen for en gennemsnitlig bolig kan umiddelbart forventes at variere mellem 2.500 kr. og 42.500 kr.
- der findes en del boliger, som er solgt til væsentligt højere kvadratmeterpriser end normalen.

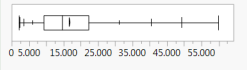
JMP-video [Analyze -> Distribution]

Box plot

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Histogram
Fraktil
CDF plot
Box plot
Kvantitativ (numerisk)
OPSUMMERING

Eksempel: Boligpriser

Et ***quantile box plot*** for variablen *SalgsprisPerM2* har følgende udseende (her med en intervalinddeling på 2.500 kr.):



Quantile box plottet viser - ligesom outlier box plottet - 25%-, 50%- og 75%-fraktilen i fordelingen af variablen *SalgsprisPerM2*.

Derudover viser plottet en række yderligere fraktiler i fordelingen. Præcis hvilke fraktiler fremgår af den tilhørende fraktiltabel:

Quantiles

100.0%	maximum	59898,477157
99.5%		49264,705882
97.5%		40425,531915
90.0%		31111,111111
75.0%	quartile	22313,432836
50.0%	median	14662,921348
25.0%	quartile	9349,4542822
10.0%		6000
2.5%		3444,4444444
0.5%		2338,1294964
0.0%	minimum	2000

Med quantile box plottet får man - i forhold til outlier box plottet - en lidt anderledes beskrivelse af fordelingens haler (altså i hvilket omfang der er ekstraordinært store eller små observationer). Men som regel vil konklusionerne på baggrund af de to box plots i al væsentlighed være de samme.

1 Variabeltyper

2 Kategorisk data – grafisk og numerisk beskrivelse

3 Kvantitativt data – grafisk beskrivelse

4 Kvantitativt data – numerisk beskrivelse

Centrum • *Variation* • *Symmetri* • *Den empiriske regel*

5 OPSUMMERING

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

En beskrivelse af en ***kvantitativ*** variabels fordeling bør involvere både en grafisk og en numerisk beskrivelse.

Den grafiske beskrivelse kan laves ved hjælp af histogram, CDF plot og box plot af variablens fordeling.

Den numeriske beskrivelse består i beregning af forskellige nøgletal, der opsummerer en række karakteristika ved fordelingen:

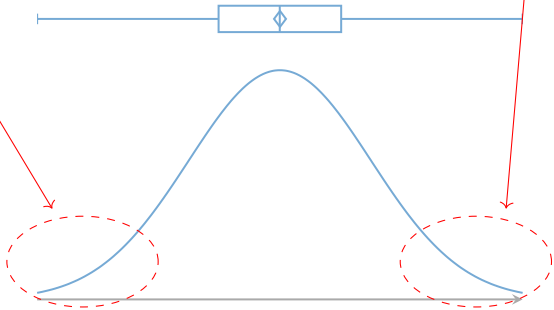
- ***centrum*** (hvad er de typiske værdier variablen antager?)
- ***variation*** (hvor meget varierer variablens værdier fra observation til observation?)
- ***(a)symmetri*** (er der flere/færre store end små observationer?)
- ***haler*** (hvad er variablens største hhv. mindste værdier?)
- ***outliers*** (er der enkelte observationer med ekstremt store eller ekstremt små værdier?)

Centrum og variation er fordelingens væsentlige karakteristika. Men inden vi ser på dem, er det nødvendigt først at se nærmere på begreberne haler og outliers.

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

For en kvantitativ variabel angiver fordelingens højre og venstre ***haler*** variablens største henholdsvis mindste værdier.

Hvis vi tegner fordelingens histogram og box plot, så illustrerer fordelingens ***venstre*** hale variablens ***mindste*** værdier, mens fordelingens ***højre*** hale illustrerer variablens ***største*** værdier.



Når vi taler om fordelingens venstre (højre) hale, taler vi således blot om fordelingens mindste (største) værdier.

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Hvis der i fordelingens haler findes enkelte ekstraordinært store eller ekstraordinært små observationer, kalder vi disse observationer for ***outliers***. Outliers afviger per definition markant fra alle øvrige observationer i datamaterialet, og de kan have en uhensigtsmæssig stor betydning, når man forsøger at beskrive en kvantitativ variabel ved hjælp af nøgletal.

Outliers kan eksempelvis opdages ved at optegne...

- fordelingens histogram og se efter, om der er enkelte observationer med ekstraordinært store og/eller små værdier i forhold til det øvrige datamateriale
- fordelingens ***quantile box plot***. I dette plot opfattes alle observationer udenfor intervallet

$$\left[q_{25\%} - 1,5 \cdot IQR;\ q_{75\%} + 1,5 \cdot IQR\right]$$

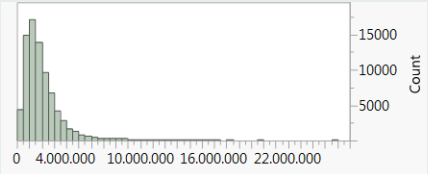
  som *potentielle* outliers og er derfor markeret enkeltvis

Om en given observation er en outlier eller ej vil altid bero på en individuel vurdering af, om observationen skiller sig så meget ud, at den ikke på rimelig vis kan sammenlignes med de øvrige observationer.

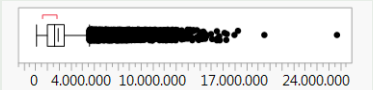
Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Eksempel: Boligpriser

Histogrammet af variablen *Salgspris* viser, at der er enkelte boliger, som er solgt til en markant højere pris end alle øvrige boliger.



Samme billede får vi også, hvis vi i stedet ser på fordelingens quantile box plot.



Hvis vi derfor eksempelvis beregner gennemsnittet af alle boligers salgspriser, så vil gennemsnittet blive uforholdsmæssigt højt på grund af nogle få boliger med ekstraordinært høje salgspriser.

Gennemsnittet bliver dermed ikke umiddelbart retvisende for salgsprisen på en gennemsnitlig bolig i datamaterialet.

De få boliger med ekstraordinært høje salgspriser kan opfattes som ***outliers***, fordi de skiller sig markant ud i forhold til alle øvrige observationer i datamaterialet.

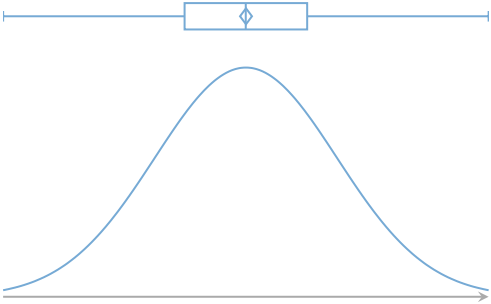
Centrum

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

For en kvantitativ variabel angiver fordelingens **centrum** det punkt, som variablens værdier er "centreret omkring".

Fordelingens centrum er med andre ord en indikator for hvilken værdi, vi umiddelbart vil forvente at variablen vil antage.

Hvis vi tegner fordelingens histogram og box plot, vil centrum ofte være nogenlunde i midten i hver af de to figurer.



Centrum

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Definition (gennemsnit)

For en kvantitativ variabel med observerede værdier $x_1, \ldots, x_n$ beregnes variablens ***gennemsnit*** *("average") $\overline{x}$ som*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Gennemsnittet ($\overline{x}$) er et mål for fordelingens centrum.

Forklaring af notation:

Symbolet Σ kaldes for et "sumtegn" (det græske bogstav Sigma). Notationen betyder, at man skal summere alle værdier af størrelsen x_i *fra* $i = 1$ op til og med $i = n$, dvs.

$$\sum_{i=1}^{n} x_i = x_1 + \ldots + x_n$$

Definitionen af $\overline{x}$ udtrykker således blot beregningen af det helt almindelige gennemsnit af observationerne $x_1, \ldots, x_n$ som summen af alle observationerne divideret med antallet af observationer n (= ganget med $\frac{1}{n}$).

Centrum

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Fordelingens centrum kan måles ved følgende to nøgletal:

- gennemsnit ($\overline{x}$)
- median ($q_{50\%}$)

Det almindelige gennemsnit $\overline{x}$ er normalt den mest naturlige måde at beregne fordelingens centrum/forventede værdi på. Gennemsnittet er imidlertid følsomt overfor eventuelle outliers.

Som alternativ til gennemsnittet ser man derfor også på fordelingens median, der er robust overfor (dvs. ikke er påvirket af) outliers.

Centrum

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Eksempel: Boligpriser

Beregning af henholdsvis gennemsnit og median for variablen *Salgspris* giver to forskellige indikationer på variablens centrum, dvs. indikationer på størrelsen af den typiske salgspris for boliger i datamaterialet.

Den gennemsnitlige salgspris blandt alle boliger i datamaterialet er $\overline{x} = 1.911.910$ kr., mens medianen er $q_{50\%} = 1.595.000$ kr. og dermed væsentligt mindre.

Det skyldes, som vi har set, at enkelte boliger har en markant højere salgspris end alle øvrige boliger. Det påvirker gennemsnittet $\overline{x}$ i opadgående retning, mens det ikke påvirker medianen $q_{50\%}$.

Umiddelbart virker medianen på 1.595.000 kr. derfor som det mest retvisende mål for fordelingens centrum, dvs. som den mest retvisende indikation af hvilken pris boligerne i datamaterialet typisk er blevet solgt til.

Quantiles

100.0%	maximum	25750000
99.5%		8350000
97.5%		5350000
90.0%		3500000
75.0%	quartile	2450000
50.0%	median	1595000
25.0%	quartile	1000000
10.0%		640000
2.5%		372000
0.5%		240414,96
0.0%	minimum	70000

Summary Statistics

Mean	1911910,2
Std Dev	1364389,6
Std Err Mean	4897,0846
Upper 95% Mean	1921508,5
Lower 95% Mean	1902312
N	77625

Centrum

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

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 variablen *Lemvig*, der indeholder den gennemsnitlige årlige indkomst (i 1.000 kr.) for personer bosiddende i Lemvig kommune, er gennemsnittet 276,2 (dvs. 276.200 kr.) og medianen 279 (dvs. 279.000 kr.).

På grund af datasættets begrænsede størrelse (5 observationer) er det nemt at illustrere, hvorledes de to nøgletal beregnes.

Gennemsnittet beregnes som

$$\overline{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}\left(x_1 + x_2 + x_3 + x_4 + x_5\right)$$

$$= \frac{1}{5}\left(265 + 266 + 280 + 279 + 291\right) = 276,2$$

mens medianen $q_{50\%} = 279$ fremkommer som den midterste observation, når alle observationerne ordnes efter størrelse:

$$265, 266, \mathbf{279}, 280, 291$$

Quantiles

100.0%	maximum	291
99.5%		291
97.5%		291
90.0%		291
75.0%	quartile	285,5
50.0%	median	279
25.0%	quartile	265,5
10.0%		265
2.5%		265
0.5%		265
0.0%	minimum	265

Summary Statistics

Mean	276,2
Std Dev	10,848963
Std Err Mean	4,8518038
Upper 95% Mean	289,67077
Lower 95% Mean	262,72923
N	5

Variation

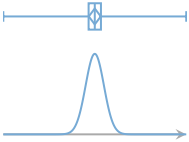
Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

For en kvantitativ variabel angiver fordelingens **variation**, hvor meget variablens værdier typisk "varierer" (ændrer sig) fra observation til observation.

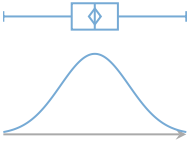
Fordelingens variation er med andre ord en indikator for, hvor stor forskel der er på de enkelte observationer i datamaterialet.

Hvis vi tegner fordelingens histogram og box plot vil "'bredden" af hver af de to figurer illustrere størrelsen af fordelingens variation.

En fordeling med ***lille*** variation vil have et ***smalt*** box plot og histogram, mens en fordeling med ***stor*** variation vil have et ***bredt*** box plot og histogram.



Lille variation



Stor variation

VARIATION

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Definition (varians, standardafvigelse)

For en kvantitativ variabel med observerede værdier $x_1, \ldots, x_n$ beregnes variablens **varians** ("variance") s^2 *som*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

og variablens **standardafvigelse** ("standard deviation") s *som*

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2} = \sqrt{s^2}$$

Variansen (s^2) og standardafvigelsen (s) er begge to mål for fordelingens variation.

Varians og standardafvigelse er essentielt to forskellige måder at udtrykke præcis det samme på, nemlig i hvilket omfang observationerne i datamaterialet afviger fra datamaterialets gennemsnitlige værdi ($\overline{x}$).

Variation

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Variansen (s^2) er umiddelbart den mest intuitive størrelse af de to, idet den måler observationernes gennemsnitlige kvadrerede afvigelse $(x_i - \overline{x})^2$ fra gennemsnittet $\overline{x}$.

Standardafvigelsen (s) er omvendt den lettest fortolkelige af de to, fordi den måler fordelingens variation i samme enhed (f.eks. kr., kg, dage) som variablen selv, mens variansen måler variation i "enheden kvadreret" (f.eks. kr.^{2}, kg^2, dage^2), der ikke har nogen naturlig fortolkning.

Variansen og standardafvigelsen skelner ikke (fordi der kvadreres) mellem positive og negative afvigelser fra gennemsnittet. Begge mål ser alene på i hvilket omfang hver enkelt observation (x_i) ligger tæt på eller langt fra gennemsnittet ($\overline{x}$).

Både variansen og standardafvigelsen er altid større end eller lig med 0. Desuden er såvel varians som standardafvigelse kun lig 0 i det specielle tilfælde, hvor alle datamaterialets observationer er identiske (dvs. $x_1 = x_2 = \ldots = x_n$).

Variation

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Fordelingens variation kan måles ved følgende to nøgletal:

- standardafvigelse (s) (eller alternativt variansen s^2)
- interkvartilbredden (IQR)

Standardafvigelsen s er normalt den mest naturlige måde at beregne fordelingens variation på. Standardafvigelsen er imidlertid følsom overfor eventuelle outliers.

Som alternativ til standardafvigelsen ser man derfor også på fordelingens interkvartilbredde (IQR), der er robust overfor outliers.

Desto ***større*** variation en fordeling har, desto ***større*** bliver...

- standardafvigelsen
- variansen
- interkvartilbredden

Variation

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Eksempel: Boligpriser

Beregning af henholdsvis standardafvigelse og interkvartilbredde for variablen *Salgspris* giver to indikationer på variablens variation, dvs. på hvor meget salgsprisen varierer fra bolig til bolig i datamaterialet.

Standardafvigelsen af salgspris blandt alle boliger i datamaterialet er $s = 1.364.390$ kr., mens interkvartilbredden er $IQR = q_{75\%} - q_{25\%} = 2.450.000 - 1.000.000 = 1.450.000$ kr.

De to mål er således nogenlunde enige om størrelsen af variationen i salgspris blandt boligerne i datamaterialet.

Hvis man ønsker det, kan man ud fra standardafvigelsen også beregne fordelingens varians som $s^2 = 1.364.390^2 = 1.861.559.056.415$. Dette tal har imidlertid ikke nogen naturlig fortolkning (fordi s^2 måler i enheden kr.^{2}, dvs. "kroner-i-anden-potens").

Hvis man ønsker det, kan både varians og interkvartilbredde også udregnes direkte i JMP.

Quantiles

100.0%	maximum	25750000
99.5%		8350000
97.5%		5350000
90.0%		3500000
75.0%	quartile	2450000
50.0%	median	1595000
25.0%	quartile	1000000
10.0%		640000
2.5%		372000
0.5%		240414,96
0.0%	minimum	70000

Summary Statistics

Mean	1911910,2
Std Dev	1364389,6
Std Err Mean	4897,0846
Upper 95% Mean	1921508,5
Lower 95% Mean	1902312
N	77625

Summary Statistics

Mean	1911910,22363929
Std Dev	1364389,62778789
Std Err Mean	4897,08455304585
Upper 95% Mean	1921508,48265487
Lower 95% Mean	1902311,96462371
N	77625
Variance	1861559056415,18
Interquartile Range	1450000

JMP-video [Analyze -> Distribution]

Variation

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

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 variablen *Lemvig*, der indeholder den gennemsnitlige årlige indkomst (i 1.000 kr.) for personer bosiddende i Lemvig kommune, er standardafvigelsen 10,849 (dvs. 10.849 kr.) og variansen 117,7 (der ikke har nogen naturlig måleenhed).

Variansen beregnes som (husk at vi tidligere har fundet gennemsnittet til $\overline{x} = 276,2$)

$$s^2 = \frac{1}{5-1}\sum_{i=1}^{5}(x_i - \overline{x})^2 = \frac{1}{4}\left((x_1-\overline{x})^2 + (x_2-\overline{x})^2 + (x_3-\overline{x})^2 + (x_4-\overline{x})^2 + (x_5-\overline{x})^2\right)$$

$$= \frac{1}{4}\Big((265-276,2)^2 + (266-276,2)^2 + (280-276,2)^2 + (279-276,2)^2 + (291-276,2)^2\Big) = 117,7$$

mens standardafvigelsen herefter kan findes som

$$s = \sqrt{s^2} = \sqrt{117,7} = 10,849$$

Summary Statistics

Mean	276,2
Std Dev	10,848963
Std Err Mean	4,8518038
Upper 95% Mean	289,67077
Lower 95% Mean	262,72923
N	5
Variance	117,7

Symmetri

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

En fordeling kaldes ***symmetrisk***, ***højreskæv*** eller ***venstreskæv*** alt afhængig af om den er symmetrisk omkring sit centrum eller ej.

Hvis fordelingen er højreskæv (venstreskæv) betyder det, at der er visse meget store (små) observationer i datamaterialet. Disse observationer vil føre til en større (mindre) gennemsnitlig værdi end i tilfældet med en symmetrisk fordeling.

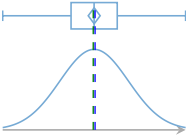
Fordelingens (a)symmetri kan derfor både aflæses af fordelingens histogram, af dens box plot og ved at sammenligne dens **gennemsnit** og **median**.

Symmetri

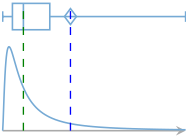
Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Hvis en fordeling er...

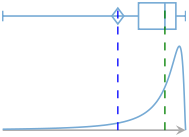
- **symmetrisk** *("symmetric")* er **gennemsnit** og **median** identiske. Desuden er histogram og box plot symmetriske omkring fordelingens centrum.



- **højreskæv** *("right-skew")* er højre hale i histogrammet længere end venstre hale. Desuden er **gennemsnittet** ***større*** end **medianen**.



- **venstreskæv** *("left-skew")* er venstre hale i histogrammet længere end højre hale. Desuden er **gennemsnittet** ***mindre*** end **medianen**.



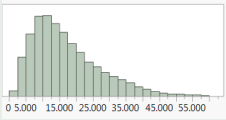
Bemærk: Skævhed kan ikke altid aflæses af fordelingens box plot, der kan indikere symmetri, selv om fordelingen er en smule skæv. Hvis fordelingen imidlertid er tilstrækkeligt højreskæv (venstreskæv) vil box plottet være asymmetrisk, idet 3. kvartil (1. kvartil) så vil ligge længere væk fra medianen end 1. kvartil (3. kvartil).

Symmetri

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

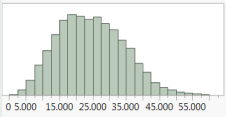
Eksempel: Boligpriser

Histogrammet af variablen *SalgsprisPerM2* viser, at variablens fordeling er tydeligt højreskæv, fordi der er enkelte boliger, som er solgt til en markant højere kvadratmeterpris end alle øvrige boliger.

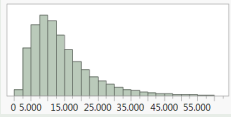


Hvis vi inddeler observationerne efter den solgte boligs type, ser vi, at der er væsentlige forskellige på fordelingen af *SalgsprisPerM2* indenfor de enkelte boligtyper.

Ejerlejlighed:



Villa:



Fordelingen af kvadratmeterprisen er stort set symmetrisk, hvis vi alene ser på salgspriserne blandt ejerlejligheder, mens der er en tydelig højreskævhed i salgspriserne blandt villaer.

Der er således – alt andet lige – en noget mere jævn fordeling af kvadratmeterprisen blandt de solgte ejerlejligheder, mens der i højere grad blandt de solgte villaer er enkelte boliger, som er solgt til en væsentligt højere kvadratmeterpris end de øvrige.

JMP-video [Analyze -> Distribution]

Den empiriske regel

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Formålet med dataanalyse er at bruge eksisterende historisk talmateriale til at forsøge at udtale sig om, hvad vi rimeligvis kan forvente os om fremtiden.

Ved for en given variabel at kombinere de beregnede værdier af gennemsnit ($\overline{x}$) og standardafvigelse (s) får vi et første og ganske stærkt værktøj til rådighed, der kan bruges til netop dette formål.

Det følgende resultat fortæller os, hvad vi umiddelbart kan forvente os om fremtidige værdier af en given variabel.

Resultat ["Den empiriske regel"]

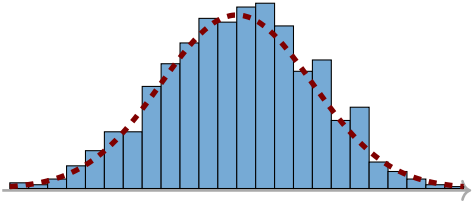
Hvis en kvantitativ variabel har en klokkeformet ("bell-shaped") fordeling, så vil...

- *ca. 68% af variablens observationer ligge i intervallet* $\left[\overline{x}-s;\ \overline{x}-s\right]$
- *ca. 95% af variablens observationer ligge i intervallet* $\left[\overline{x}-2s;\ \overline{x}+2s\right]$
- *næsten alle variablens observationer ligge i intervallet* $\left[\overline{x}-3s;\ \overline{x}+3s\right]$

Den empiriske regel

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Med "klokkeformet fordeling" menes, at fordelingens histogram har form som en pæn, symmetrisk (kirke)klokke.



Som vi siden hen skal se, vil ganske mange variable i praksis have en sådan form, og for alle sådanne variable kan vi bruge "den empiriske regel".

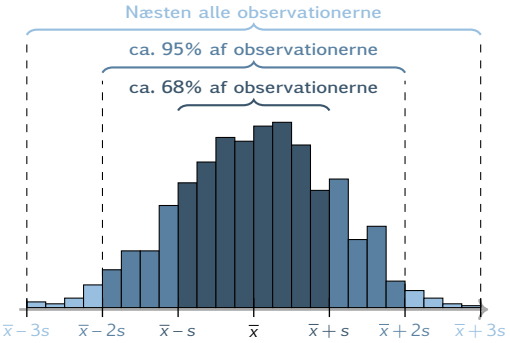
Den empiriske regel fortæller os, at når en variabel har en (nogenlunde) klokkeformet fordeling, så vil variablens værdier ligge i nærheden af sit gennemsnit ($\overline{x}$). Desuden fortæller den, at hvor meget vi kan forvente værdierne vil afvige fra gennemsnittet er bestemt af standardafvigelsen (s).

Den empiriske regel

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Mere præcist, så fortæller den empiriske regel, at...

- ca. 68% af variablens observationer vil være mindre end 1 standardafvigelse væk fra gennemsnittet.
- ca. 95% af variablens observationer vil være mindre end 2 standardafvigelser væk fra gennemsnittet.
- næsten alle variablens observationer vil være mindre end 3 standardafvigelser væk fra gennemsnittet.



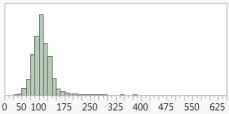
Den empiriske regel

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Eksempel: Boligpriser

Variablen *BoligM2* angiver størrelsen af boligarealet (i m^2) for hver af de solgte boliger. Histogrammerne af *BoligM2* for henholdsvis rækkehuse og villa viser, at variablens fordeling for begge boligtyper er nogenlunde klokkeformet.

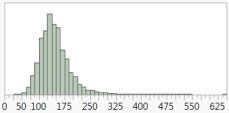
Rækkehus:



Summary Statistics

Mean	106,99756
Std Dev	24,300127
Std Err Mean	0,3100656
Upper 95% Mean	107,6054
Lower 95% Mean	106,38972
N	6142

Villa:



Summary Statistics

Mean	144,99603
Std Dev	42,12135
Std Err Mean	0,1972946
Upper 95% Mean	145,38273
Lower 95% Mean	144,60933
N	45580

Selv om der er højreskævhed i begge fordelinger, er det ikke et alvorligt problem, fordi den højre hale for begge fordelingers vedkommende er meget lav/flad (og dermed kun indeholder relativt få observationer).

Af histogrammerne er det tydeligt at se, at boligarealet (ikke overraskende) generelt set er større i villaer end i rækkehuse (centrum ligger længere til højre i histogrammet for villaer). Ligeledes viser histogrammerne, at der er større spredning i boligarealet blandt villaer end blandt rækkehuse (histogrammet for villaer er bredere).

Den empiriske regel

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
Centrum
Variation
Symmetri
Empirisk regel
OPSUMMERING

Eksempel: Boligpriser (fortsat)

Ved anvendelse af den empiriske regel får vi, at boligarealet for et tilfældigt udvalgt rækkehus *i datamaterialet*...

- med ca. 68% sandsynlighed vil ligge i intervallet

$$\left[\overline{x} - s\ \overline{x} + s\right] = \left[107 - 24,30;\ 107 + 24,30\right] = \left[82,70;\ 131,30\right]$$

- med ca. 95% sandsynlighed vil ligge i intervallet

$$\left[\overline{x} - 2s\ \overline{x} + 2s\right] = \left[107 - 24,30;\ 107 + 24,30\right] = \left[58,40;\ 155,60\right]$$

- med næsten 100% sandsynlighed vil ligge i intervallet

$$\left[\overline{x} - 3s\ \overline{x} + 3s\right] = \left[107 - 24,30;\ 107 + 24,30\right] = \left[34,10; 179,90\right]$$

Ovenstående udsagn vedrører alene de boliger, der findes i datamaterialet, dvs. de vedrører boliger som er blevet solgt i løbet af den tidsperiode, datamaterialet dækker.

Såfremt datamaterialet imidlertid også er nogenlunde retvisende for boligsalget i dag og i en periode fremover, (dvs. såfremt de faktorer der er afgørende for boligsalget er nogenlunde uændrede fra dataperioden og fremefter), vil ovenstående udsagn også gælde for de rækkehuse, der sælges fremadrettet.

Eksempelvis kan vi så slutte, at vi vil forvente, at boligarealet på *fremtidige* solgte rækkehuse, med ca. 68% sandsynlighed vil være mellem 82,7 m^2 og 131,3 m^2.

1 Variabeltyper

2 Kategorisk data – grafisk og numerisk beskrivelse

3 Kvantitativt data – grafisk beskrivelse

4 Kvantitativt data – numerisk beskrivelse

5 OPSUMMERING

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Kort opsummering af dette notesæt:

En variabel i et datamateriale er én af følgende typer:

- kvantitativ (kontinuert eller diskret)
- kategorisk (ordinal eller nominal)

En variabel kan illustreres v.hj.a. følgende figurer:

- kategorisk variabel : søjlediagram, cirkeldiagram
- kvantitativ variabel : histogram, box plot

En kvantitativ variabel har følgende karakteristika:

- centrum (måles ved gennemsnit og median)
- variation (måles ved standardafvigelse og interkvartilbredde)
- (a)symmetri (vurderes grafisk og måles ved fraktiler)
- haler (vurderes grafisk og måles ved fraktiler)
- outliers (vurderes grafisk og måles ved fraktiler)

Hvis et datamateriale indeholder outliers, vil de...

- påvirke beregningen af gennemsnit ($\bar{x}$), varians (s^2) og standardafvigelse (s)
- *ikke* påvirke beregningen af median ($q_{0,50}$) og interkvartilbredde (IQR)

OPSUMMERING

For en variabel med en klokkeformet fordeling vil...

- ca. 68% af observationerne ligge i intervallet $[\overline{x} - s;\ \overline{x} - s]$
- ca. 95% af observationerne ligge i intervallet $[\overline{x} - 2s;\ \overline{x} + 2s]$
- næsten alle observationerne ligge i intervallet $[\overline{x} - 3s;\ \overline{x} + 3s]$

INDEKS

Box plot *s. 37*
Centrum *s. 47*
CDF plot *s. 34*
Cirkeldiagram *s. 19*
Diskret variabel *s. 9*
Empirisk regel *s. 61*
Fordeling *s. 15*
Fraktil *s. 28*
Gennemsnit *s. 48*
Hale *s. 44*
Histogram *s. 23*
Hyppighed *s. 16*
Højreskæv *s. 59*
Interkvartilbredde *s. 30*
Kategorisk variabel *s. 4*
Kontinuert variabel *s. 9*
Kvantitativ variabel *s. 4*
Kvartil *s. 29*
Median *s. 29*
Nominal variabel *s. 12*
Ordinal variabel *s. 12*
Outliers *s. 45*
Standardafvigelse *s. 53*
Symmetrisk *s. 59*
Søjlediagram *s. 19*
Varians *s. 53*
Variation *s. 52*
Venstreskæv *s. 59*

Variabeltyper
Kategorisk
Kvantitativ (grafisk)
Kvantitativ (numerisk)
OPSUMMERING

Nye funktionaliteter i dette notesæt:

- *Analyze -> Distribution:*
  - Beregn hyppighedstabel (for kategorisk variabel)
  - Tegn søjlediagram (for kategorisk variabel)
  - Beregn fraktiler, gennemsnit, varians, standardafvigelse (for kvantitativ variabel)
  - Tegn histogram, CDF plot og box plot (for kvantitativ variabel)
- *Graph -> Graph Builder:*
  - Cirkeldiagram (for kategorisk variabel)

JMP-videoer:

s. 20: ▶ [Analyze -> Distribution ; Graph -> Graph Builder]
s. 23: ▶ [Analyze -> Distribution] (histogram)
s. 31: ▶ [Analyze -> Distribution] (fraktiler)
s. 36: ▶ [Analyze -> Distribution] (CDF plot)
s. 40: ▶ [Analyze -> Distribution] (box plots)
s. 56: ▶ [Analyze -> Distribution] (ekstra nøgletal)
s. 60: ▶ [Analyze -> Distribution] (opdeling på variabel)