Introduksjon

Eksempel på data:

Karakterer i «Stat class»

Microsoft Excel

	A	B	C	D	E	F	G	H
1	ID	Exam1	Exam2	Homework	Final	Project	TotalPoints	Grade
2	101	89	94	88	87	95	899	B
3	102	78	84	90	89	94	866	B
4	103	71	80	75	79	95	780	C
5	104	95	98	97	96	93	962	A
6	105	79	88	85	88	96	861	B

Viktige begreper for å beskrive data:

- Enheter som er objektene i datasettet
- «label» som av og til brukes for å skille enhetene
- En variabel er en karakteristikk av hver enhet
- Variablene angis med verdier

Viktig skille mellom to typer variable:

- Kategoriske variabler klassisfiserer enhetene i to eller flere grupper.

- Kvantitative variabler antar numeriske verdier som gjør aritmetiske operasjoner som addisjon og subtraksjon meningsfylt.

Fordelingen til en variabel angir verdiene variabelen antar og hvor ofte.

- Velg «labels» med omhu

- Kvantitative variable må være på samme skala, f. eks. minutter, sekunder

Eksempel

Karakterer i «Stat class»

Microsoft Excel

	A	B	C	D	E	F	G	H
1	ID	Exam1	Exam2	Homework	Final	Project	TotalPoints	Grade
2	101	89	94	88	87	95	899	B
3	102	78	84	90	89	94	866	B
4	103	71	80	75	79	95	780	C
5	104	95	98	97	96	93	962	A
6	105	79	88	85	88	96	861	B

Microsoft Excel

	A	B	C	D	E	F
1	ID	TotalPoints	Grade	Gender	PrevStat	Year
2	101	899	A	F	Yes	4
3	102	866	B	M	Yes	3
4	103	780	C	M	No	3
5	104	962	A	M	No	1
6	105	861	B	F	No	4

Den siste tabellen organiserer data på en nyttig måte
hvis man er interessert i hvordan ulike studentgrupper presterer
i kurset.

Beskrivelse av data med grafer

Første steg i analyse av data er å sette seg inn i bakgrunnen så man forstår hva som beskrives og måles.

Neste steg er en mer uformell analyse, en såkalt eksplorativ dataanalyse. Her er ulike grafiske fremstillinger og verktøy svært nyttige.

Vi ser først på hvordan kategoriske variable kan presenteres.

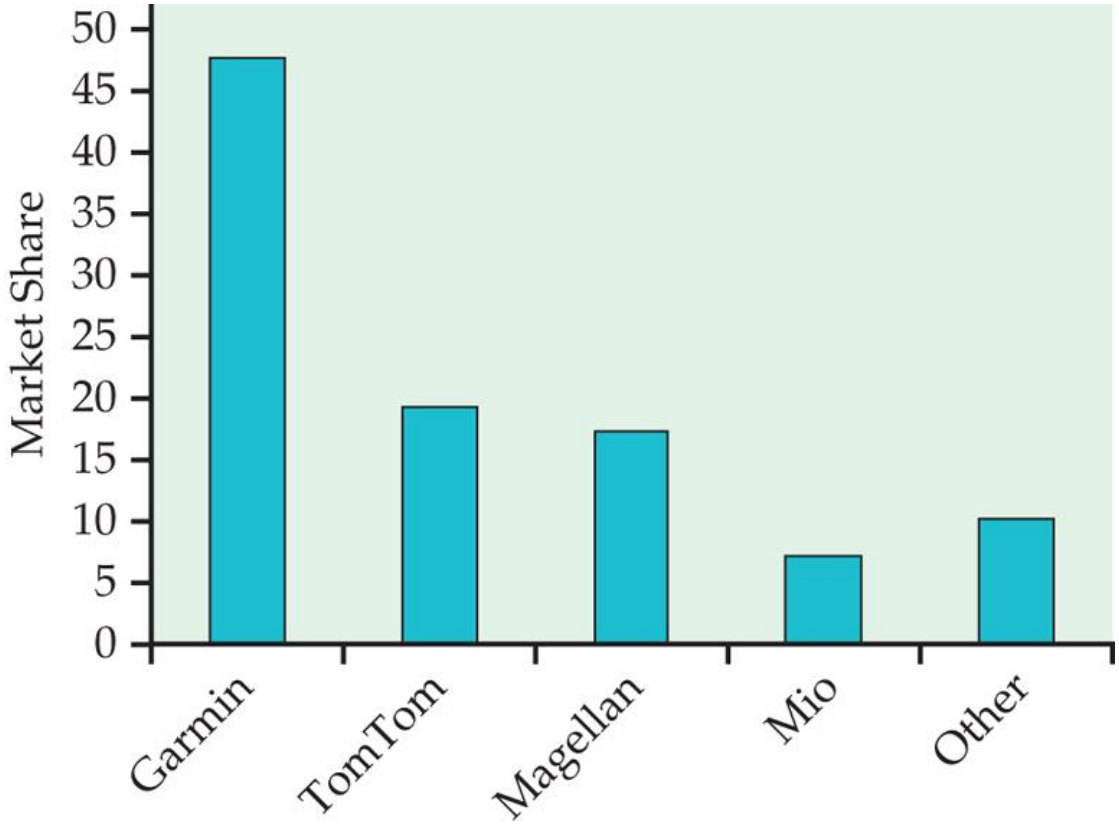
Fordelingen angis med hvor mange eller hvor

Her er det to viktige verktøy: Stolpediagram og

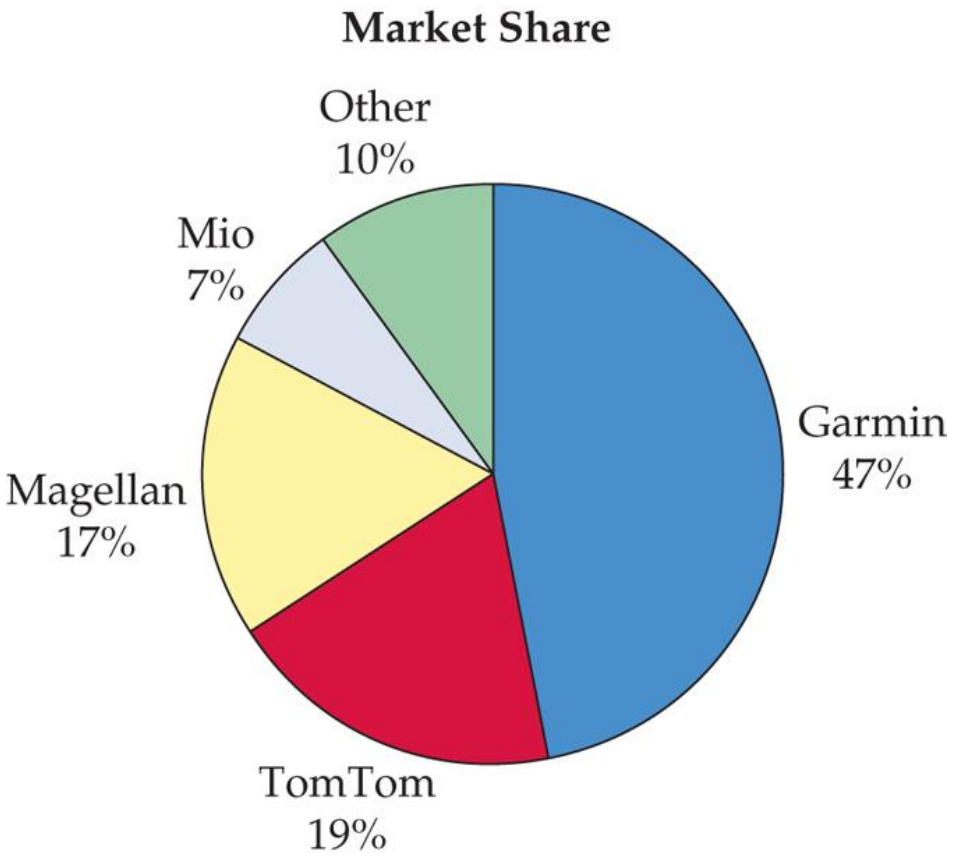
kakediagrammer.

Eksempel. Fordelingen av markedsandelen til fem GPS firmaer i USA er 47, 19, 17, 7, 10 prosent.

Her er stolpediagrammet:



og kakediagrammet



Merk

- Orden langs x-aksen i stolpediagram likegyldig. Velg det som passer best

- Stolpediagram mest fleksible. Verdiene trenger ikke summere seg til 100%

Vi ser dernest på hvordan kvantitative variabler kan presenteres.

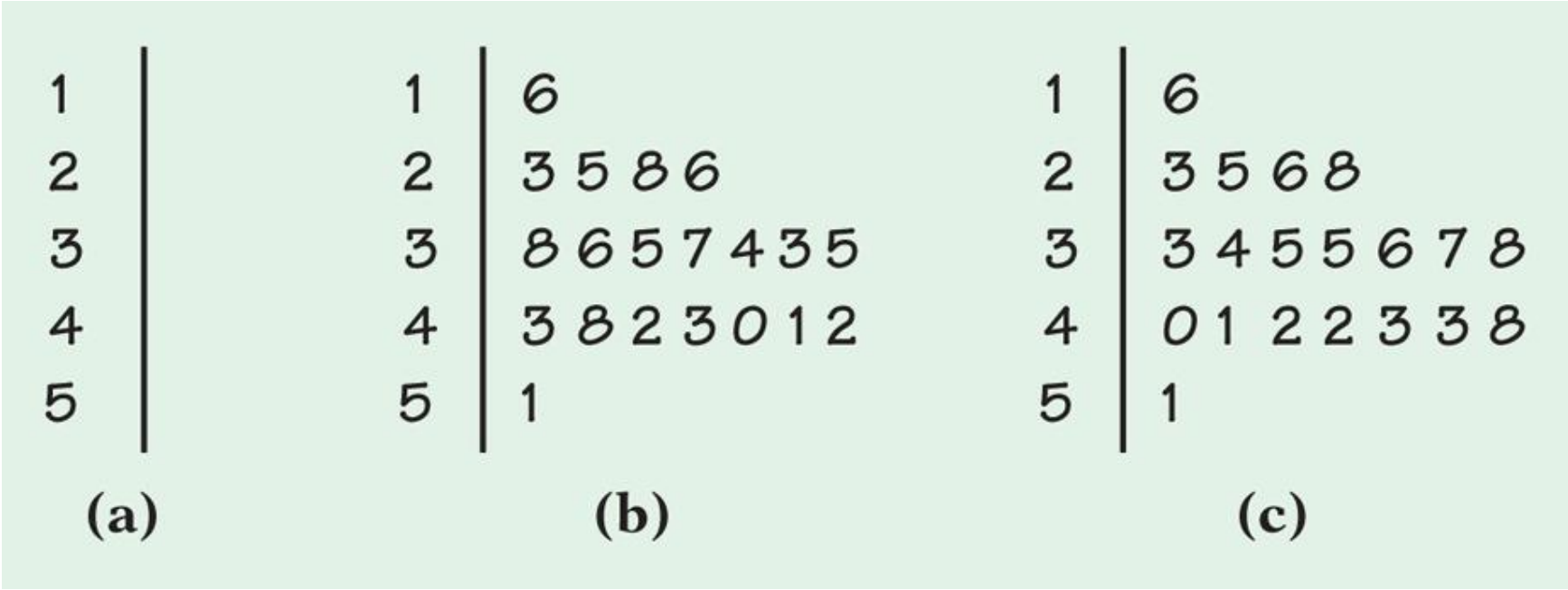
Vi ser først på såkalt stilk-og-blad plott og histogrammer.

Stilk-og-blad plott

20 målinger av 25-hydrox D-vitamin (i nanogram) blant jenter i alder 11-14 år gav følgende:

16 43 38 48 42 23 36 35 37 34 25 28 26 43 51 33 40 35 41 42

Stilk-og-blad plottet ser slik ut

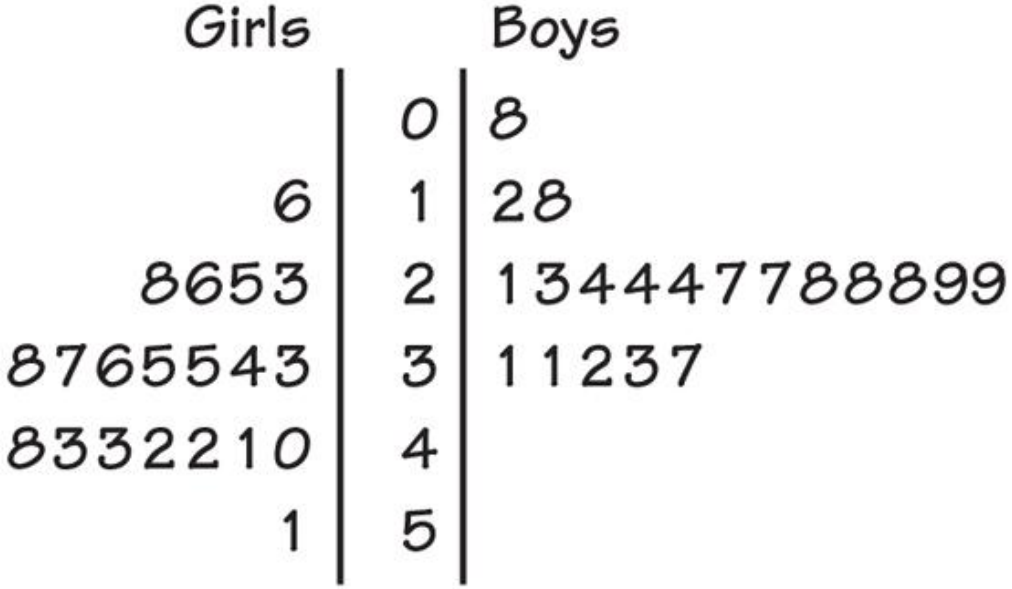


Stilk-og-blad plott er best for positive verdier.

Fremgangsmåten er slik:

1. Skill data etter de første siffrene (stilken) og siste siffer (bladene)
2. La stilken være den vertikale kolonnen
3. La bladene være en rad tilordnet stilken ordnet i stigende rekkefølge

Mer detaljerte versjoner for sammenligning av to fordelinger og finere oppdeling av stilken er også mulig



Histogrammer

Ulempe ved silk-og-blad plott:

- Problematisk ved store datasett
- Gruppinndeling fastlagt, ikke rom for skjønn

Ved histogrammer deles mulige verdier av en variabel i grupper og antall eller (prosent)andel av observasjonene som faller i hver gruppe angis.

Merk Antall klasser er valgfritt - men bruk samme vidde i de valgte gruppene

Eksempel 60 målinger av IQ for 60 femteklassinger.

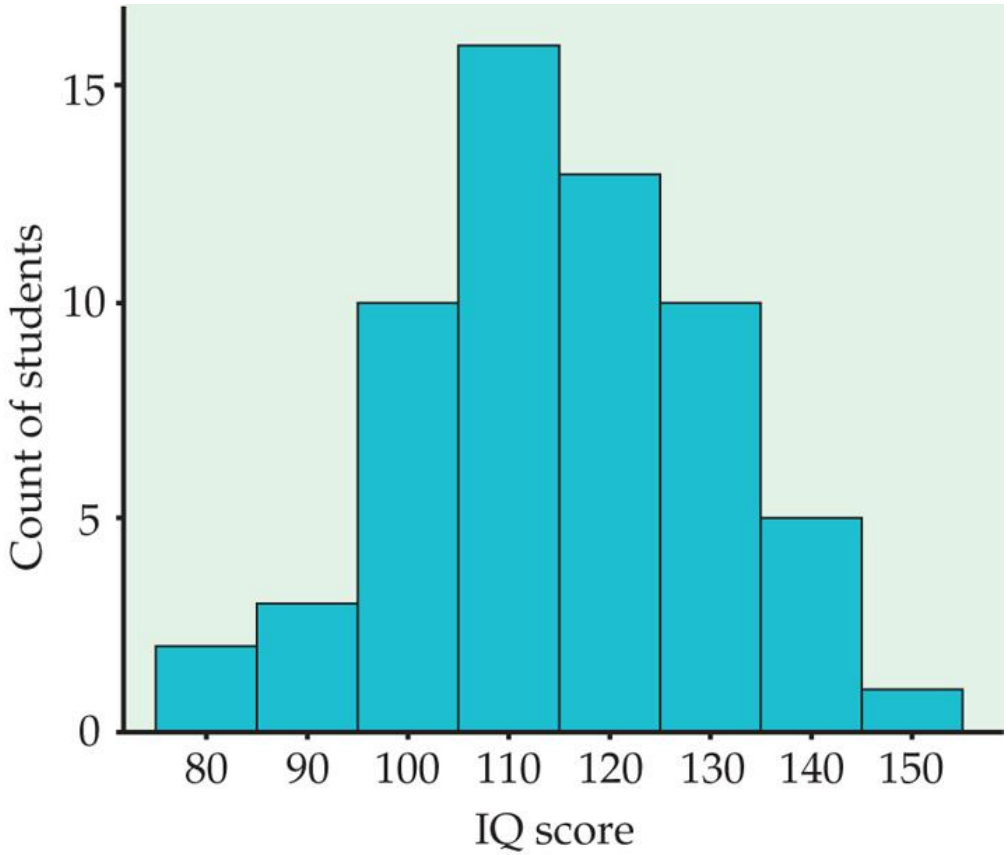
TABLE 1.1

IQ test scores for 60 randomly chosen fifth-grade students

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

Histogram:

1. Velg klasser, [75,85), ..., [145,155)
2. Tell antall i hver klasse
3. Tegn histogram


Count of students
15
10
5
0
80
90
100
110
120
130
140
150
IQ score

- Antallet i hver klasse kalles frekvens og en tabell for antall i hver klasse kalles en frekvenstabell.

- Benevningen langs y-aksen kan være prosent eller antall.

- Areal i histogram viktig visuelt. Det er derfor essensielt å bruke like vide klasser.

- Valg av antall klasser viktig. Bruk skjønn: skyskrapere og pannekaker er ikke spesielt informative. Utseende til histogrammet kan endres når antall klasser og vidden endres.

- Søylediadram og histogram er vesensforskjellige. Plasseringen langs x-aksen er vilkårlig i søylediagram. Bruk mellomrom mellom søylene i søylediagram og ingen mellomrom i histogrammer.

- Statistikkpakker har prosedyrer for å lage histogrammer. Default verdiene for valg av antall klasser er vanligvis fornuftige.

Hensikten med de grafiske framstillingene er å få et bilde av fordelingen til en variabel for eventuelt å kunnne sammenligne med andre variable. Dette gjelder både mer almenne trekk, men også identifisering av enheter der verdiene til variabelen skiller seg ut.

N= 31492 målinger er tatt for lengden av samtaler ved et «call center». Tabellen viser 80.

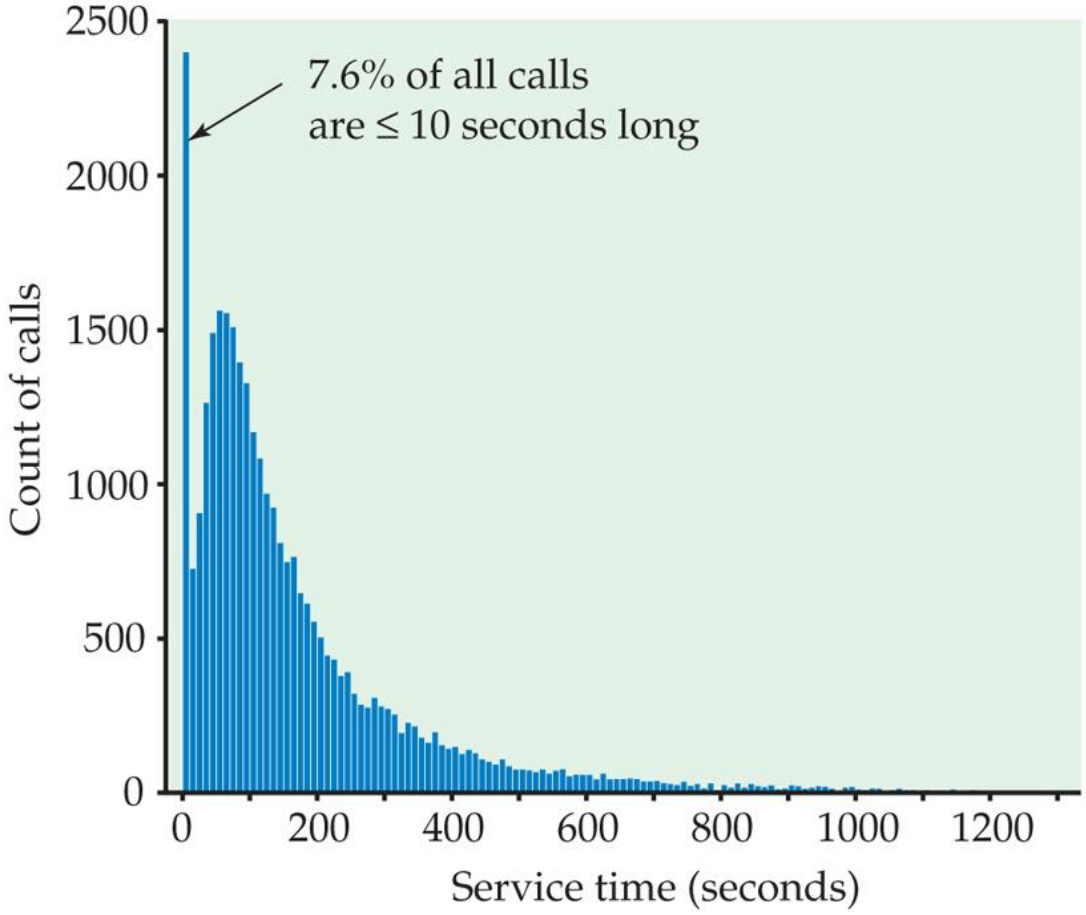
Følgende 80 målinger er en del av de n=31492 registreringer av lengden på samtaler ved et «call center»

TABLE 1.2

Service times (seconds) for calls to a customer service center

77	289	128	59	19	148	157	203
126	118	104	141	290	48	3	2
372	140	438	56	44	274	479	211
179	1	68	386	2631	90	30	57
89	116	225	700	40	73	75	51
148	9	115	19	76	138	178	76
67	102	35	80	143	951	106	55
4	54	137	367	277	201	52	9
700	182	73	199	325	75	103	64
121	11	9	88	1148	2	465	25

som gir histogrammet for hele datasettet.



Her er de almenne trekkene gitt ved

Fordelingen har ett /(to?) toppunkt, dvs den er unimodal (bimodal?) .

Fordelingen ikke symmetrisk, dvs klokkeformet, men skjev mot høyre.

Andre viktige slike almenne kjennetegn er midtpunkt i fordelingen og spredningen.

Små og store verdier av variabelen er i halen til fordelingen.

Avviket er opphopingen av korte samtaler er det tydeligste avviket i dette tilfellet. En individuell verdi som faller utenfor det almenne mønsteret kalles en outlier.

Om en enhet er en outlier eller ikke må vurderes. En mulighet er simpelthen målefeil, feil med måleutstyr eller punchefeil.

Hvis det er mulig, er plott av observasjonene som en funksjon av tidspunktet de ble hentet inn nyttig. Det kan avsløre outliere.

Bare å klassifisere ekstreme observasjoner som outliere er for enkelt som følgende eksempel viser.

Beskrive fordelinger med tall

Å beskrive fordelinger med grafiske fremstillinger og numeriske oppsummeringer utfyller hverandre og kaster lys over ulike aspekter. Vi skal spesielt se nærmere på mål for senter i fordelinger, som gjennomsnitt og median, og mål for spredning, som standard avvik og interkvartil rang.

Gjennomsnitt: La x_1,...,x_n betegne observasjonene. Da er gjennomsnittet $\bar{x}$=(x_1 +...+x_n)/n.

Medianen er grovt sett den «midterste» av observasjonene. Den beregnes på følgende måte.

1. Sorterter eller ordn observasjonene i stigende rekkefølge.

2. Hvis antallet observasjoner, n, er odde dvs. n=2k+1 er medianen observasjon nummer k+1=(n+1)/2. Posisjonen til medianen er også k+1.

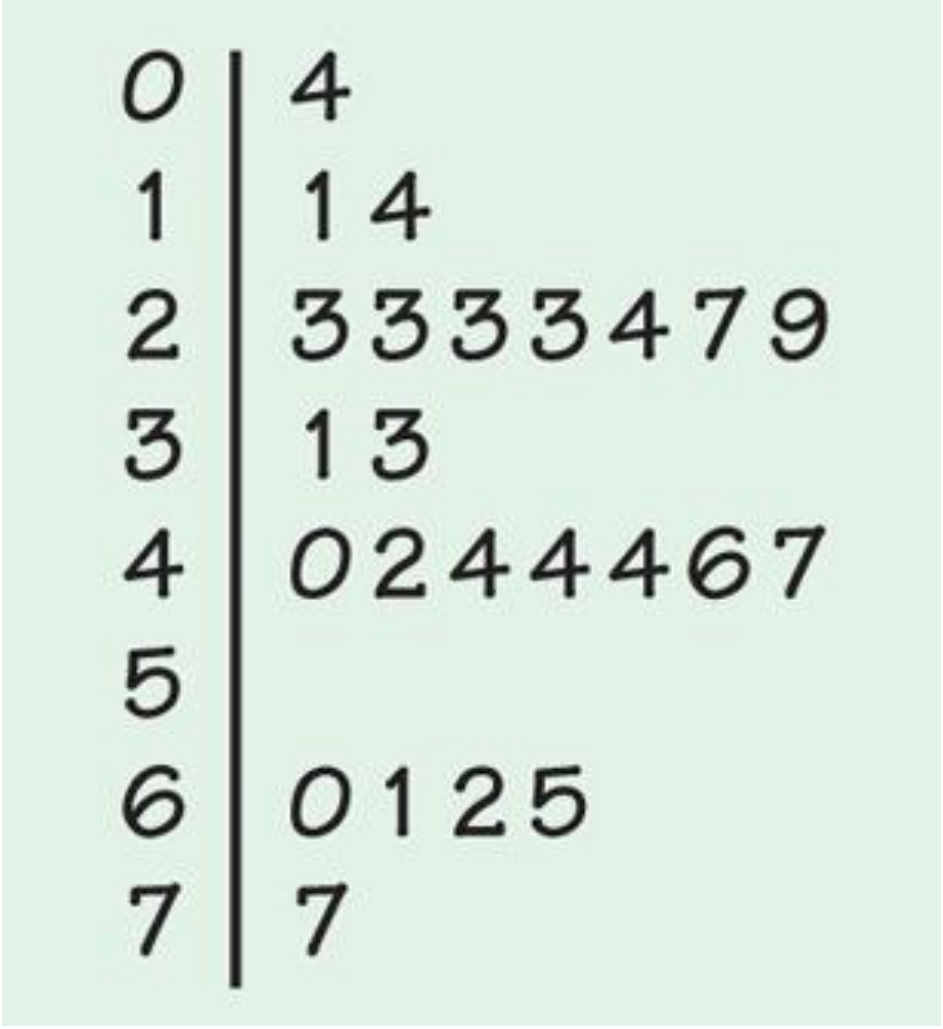
3. Hvis antallet observasjoner, n, er like dvs. n=2k er medianen gjennomsnitt av observasjon nummer k=n/2 og observasjon k+1=n/2+1. Posisjonen til medianen er (n+1)/2.

Mer formelt : La $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ være observasjonene ordnet eller sortert i stigende rekkefølge.

Anta at n er odde, dvs. $n = 2k + 1$. Da er den miderste verdien $x_{(k)}$, altså er medianen $M = x_{(k)}$.

Anta at n er like, dvs. $n = 2k$. Da er de to miderste verdiene $x_{(k)}$ *og* $x_{(k+1)}$, altså er medianen $M = (x_{(k)} + x_{(k+1)}) / 2$.

Eksempel Stilk-og-blad plottet nedenfor viser tiden det tar i gjennomsnitt å etablere et foretak i n=24 land.



De ordnede observasjonen er

4, 11,14,23,23,23,23,24,27,29,31,33,
40,42,44,44,44,46,47,60,61,62,65,77 .

Da er gjennomsnittet (4+11+23+...+65+77)/24=897/24=37.375.

Siden det er et like antall observasjoner er posisjonen til medianen (n+1)/2=25/2=12.5.

Medianen er derfor M= (33+40)/2 = 36.5.

Medianen er mer robust/resistent enn gjennomsnitt i den forstand at den er mindre påvirket av ekstreme observasjoner.

For tilnærmede eller eksakt symmetriske fordelinger er ikke forskjellen så stor.

For ikke symmetriske fordelinger dras gjennomsnittet mot tyngre eller lengre haler.

For eksempel kan inntekts og formuesfordelinger være veldig skjeve.

Hva skjer med formues fordelingen hvis Fredriksen flytter hjem?

Outliers igjen: Se på data, identifiser outliere og hva årsaken kan være.

Hvis det er målefeil eller feilregistrering, bør de rettes hvis det er mulig. Ellers kan outliere droppes i endel tilfeller.

Å undersøke dem spesielt kan være nødvendig. I de tilfellene der outlierne ikke kan droppes anbefales å bruke robuste/resistente metoder.

En annen viktig egenskap for å beskrive fordelinger er som vi har sett, spredingen. Vi skal se på to typer standardavvik og interkvartil avstand.

Som navnet antyder bygger interkvartil avstand på kvartiler.
Grovt sagt deler kvartilene de ordnede observasjonene i fire like store deler. Kvartilene er spesialtilfeller av persentiler.
Første kvartil, Q_1, er 25. persentil, medianen er 50. persentil og tredje kvartil, Q_3, er 75. persentil.

Persentiler kan defineres som den verdien der 100p% av observasjonene er mindre eller lik verdien.

Formell definisjon av Q_1 og Q_3:

1. Sorter eller ordn observasjonene i stigende rekkefølge og finn medianen M.

2. Første kvartil, Q_1, er medianen av observasjonene til venstre for M.

3. Tredje kvartil, Q_3, er medianen av observasjonene til høyre for M.

Merk

- Kvartilene er resistente/robuste.

- Pass på observasjoner med samme numeriske verdier. Anvend samme regler som om verdiene er distinkte

En nyttig sammenfatning av mål for senter og spredning i en fordeling er en <u>såkalt fem-talls oppsummering</u> og det tilsvarende grafiske <u>boks plott</u>.

Fem-talls oppsummering:

$$x_{(1)} = \min_{1 \le i \le n} x_i,\ Q_1,\ \mathrm{M},\ Q_3,\ x_{(n)} = \max_{1 \le i \le n} x_i.$$

Boks plott:

Boks begrenset av kvartilene Q_1 og Q_3.

Horisontal linje som markerer medianen M.

Linje fra boksen til minste og største observasjon.

Eksempel Her er de 80 samtalelengdene vi har sett på tidligere.

TABLE 1.2

Service times (seconds) for calls to a customer service center

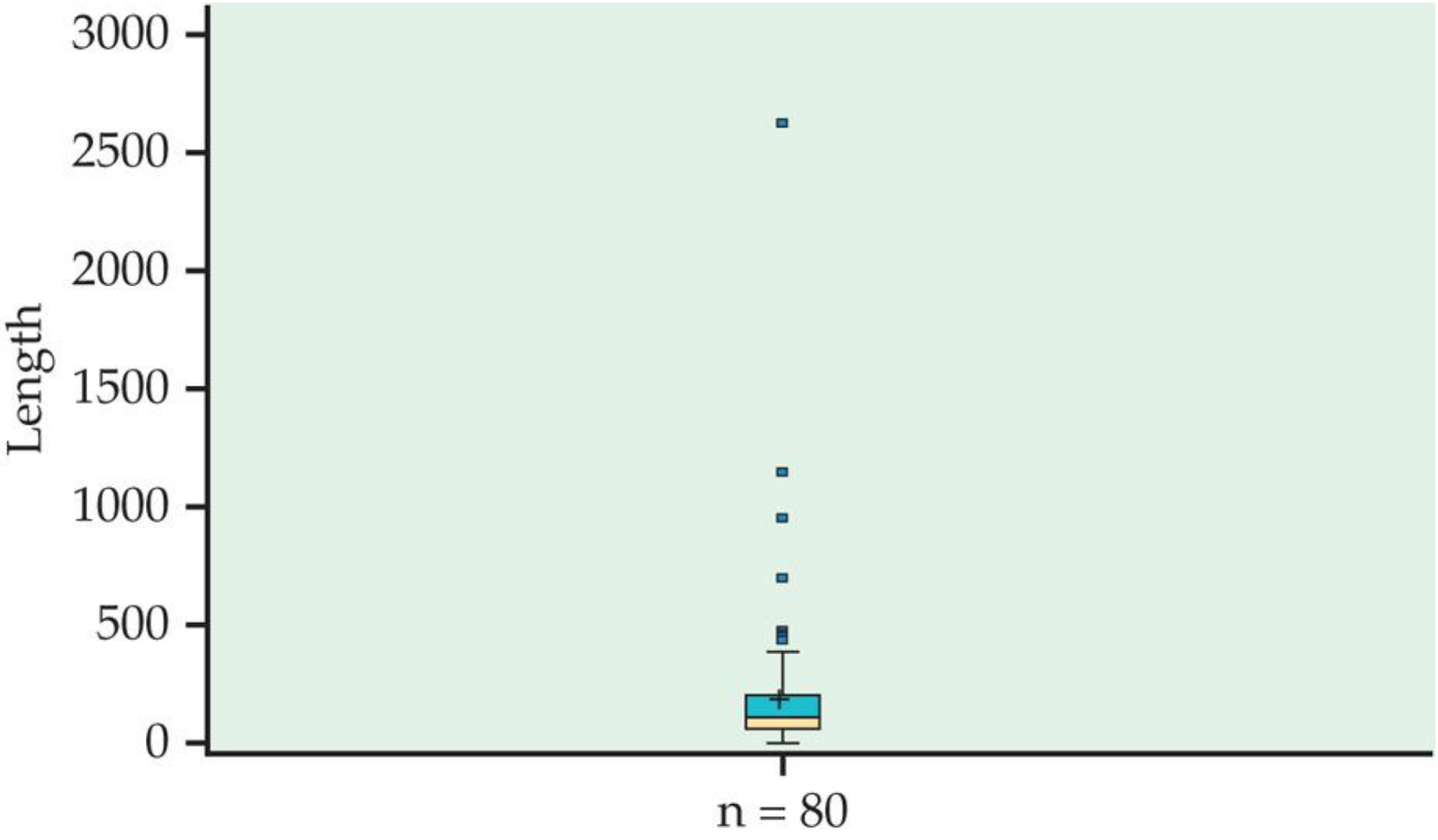
77	289	128	59	19	148	157	203
126	118	104	141	290	48	3	2
372	140	438	56	44	274	479	211
179	1	68	386	2631	90	30	57
89	116	225	700	40	73	75	51
148	9	115	19	76	138	178	76
67	102	35	80	143	951	106	55
4	54	137	367	277	201	52	9
700	182	73	199	325	75	103	64
121	11	9	88	1148	2	465	25

Femtallsoppsummeringen er:

1.0, 54.5, 103.5, 200, 2631

som viser at fordelingen er skjev.

Dette vises også fra det (modifiserte) boksplottet). Her er gjennomsnittet markert med «+» ikke langt fra tredje kvartil. Dessuten angis ”eksteme” verdier.



Spredningsmålet definert fra kvartilene kalles interkvartil avstand,IQR, og er avstanden mellom tredje og første kvartil.

Interkvartil avstand: IQR= Q_3-Q_1.

Interkvartil avstand brukes til å identifisere outliere, som defineres som observasjoner som er mindre enn Q_1-1.5×IQR eller større enn Q_3 +1.5×IQR.

Et boksplott som spesifiserer mulige outliere kalles et modifisert boksplott og ble vist for dataene fra call senteret. Der var 1.5×IQR= 1.5×145.5=218.25, slik at verdier under 54.5 – 218.25 eller over 200+218.25= 418.25 markeres. Det svarer til samtalene av lengde 438 465 479 700 700 951 1148 2631.

Det vanligste målet for spredning er **ikke** interkvartil avstand, men standardavvik :

La $s^2 = [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] / (n-1).$

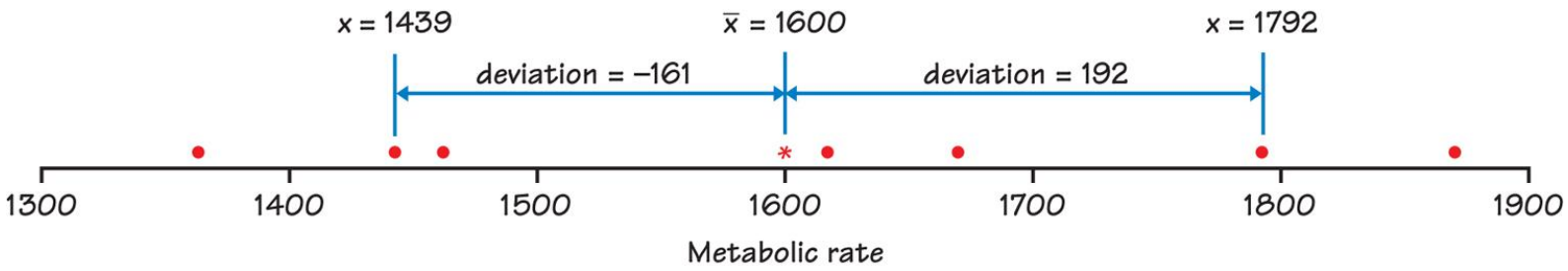
Da er standardavviket definert som $s = \sqrt{s^2}$.

Størrelsen s^2 kalles variansen til x_i - ene.

Eksempel Stoffskifte rate (kalorier pr 24 timer)

Målinger 1792 1666 1362 1614 1460 1867 1439

Gjennomsnitt: 1600 kalorier, std. avvik. 189.24



Merk

- $\bar{x}$ minimerer $(x_1 - a)^2 + \cdots + (x_n - a)^2$ med hensyn på *a,* så kvadrering av avik fra $\bar{x}$ er rimelig.

- Standardavviket s er en vikig størrelse i normalfordeling, og s er her et mer naturlig spredningsmål enn s^2.

- Standardavviket angis på samme skala som observasjonene.

- Deler på n-1 fordi $(x_1 - \bar{x})$ + ... + $(x_n - \bar{x})$ =0 medfører en restriksjon, så n-1 er antallet urrelaterte tall.

- n-1 kalles antallet frihetsgrader til s eller s^2.

Standardavviket har følgende egenskaper

- Standardavviket måler spredning rundt gjennomsnitt og bør derfor ikke brukes sammen med andre mål for senteret i fordelingen enn gjennomsnitt.

- s=0 svarer til ingen spredning, dvs $x_1 = \cdots = x_n = \bar{x}$.

- s er ikke resistent/robust. Det er enda mer enn gjennomsnitt påvirket av noen få ekstreme observasjoner.

- For symmetriske eller tilnærmet symmetriske fordelinger beskriver gjennomsnitt og standardavvik fodelingen godt. For skjeve fordelinger gir fem-talls oppsummering mer informasjon.

Variabler kan angis i ulike enheter. Eksempler er lengder i centimeter eller tommer og temperaturer på en Celcius eller Fahrenheit skala. Sammenhengen beskrives ved en lineær transformasjon.

Lineær transformasjon: x_{new} = a + bx_{old} .

Merk

- Ved multiplikasjon av alle verdier med en konstant endres måleenheten.

- Ved å addere en konstant til alle verdier endres nullpunktet.

Eksempel

Overgang fra kilometer til «miles»: x_{new} = 0.62 x_{old}

Overgang fra Fahrenheit til Celcius skala: x_{new} = 5(x_{old}-32)/9.

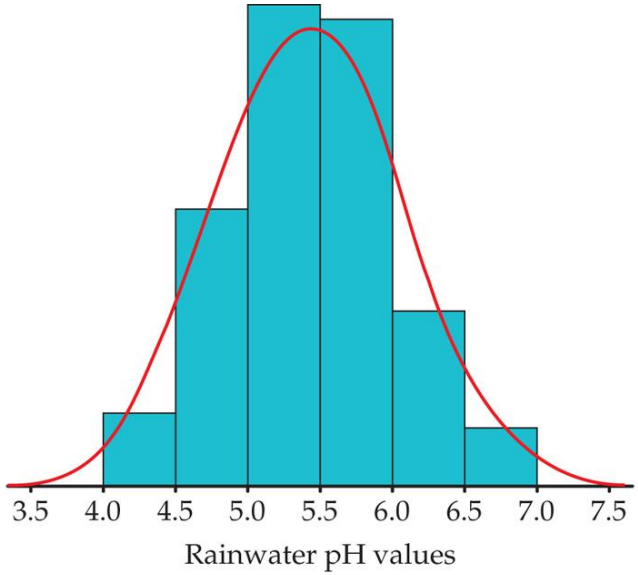
Merk

- Lineære transformasjoner endrer ikke formen på en fordeling.
- Hvis x_{new} = b x_{old}, endres M til bM, Q_1 til b Q_1 , Q_3 til b Q_3 og $\bar{x}$ til b $\bar{x}$.
- Hvis x_{new} = x_{old} +a, endres $\bar{x}$ til $\bar{x}$ +a , M til M+a.

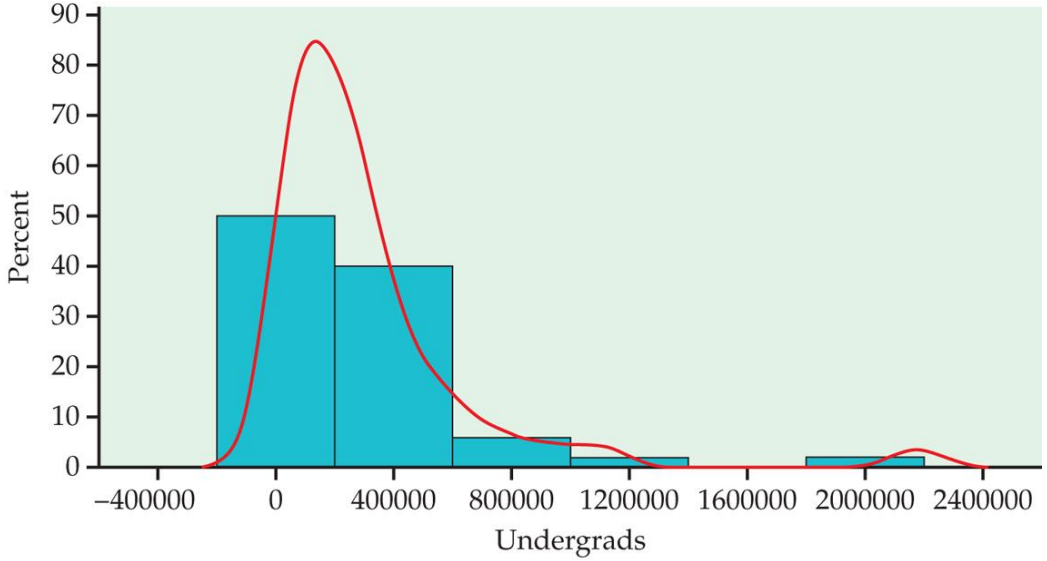
Tetthetskurver og normalfordeling

Fordelingen til dataene, som også kalles den empiriske fordelingen, kan tilnærmes med en del standard (teoretiske) fordelinger.

En vanlig måte å gjøre det på er å tilnærme histogrammer med tetthetskurver.


3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
Rainwater pH values

(a)


Percent
90
80
70
60
50
40
30
20
10
0
–400000
0
400000
800000
1200000
1600000
2000000
2400000
Undergrads

(b)

Den vanligste tetthetskurven er den klokkeformede symmetriske normalfordelingskurven.

Tetthetskurver er kurver som

- Alltid er over den horisontale x-aksen

- Har areal lik 1 under kurven

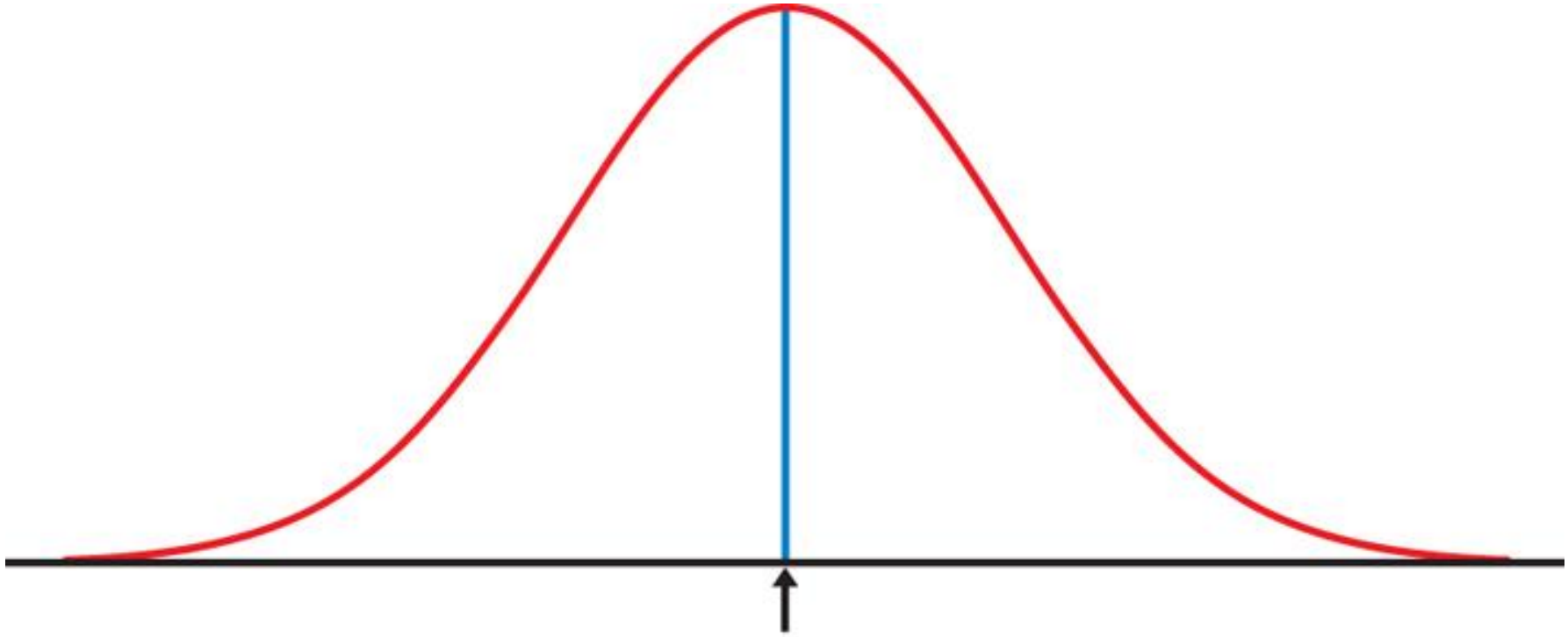
Arealt under en tetthetskurve representerer andeler på samme måte som for et histogram (hvis andeler eller prosenter brukes på den vertikale aksen).

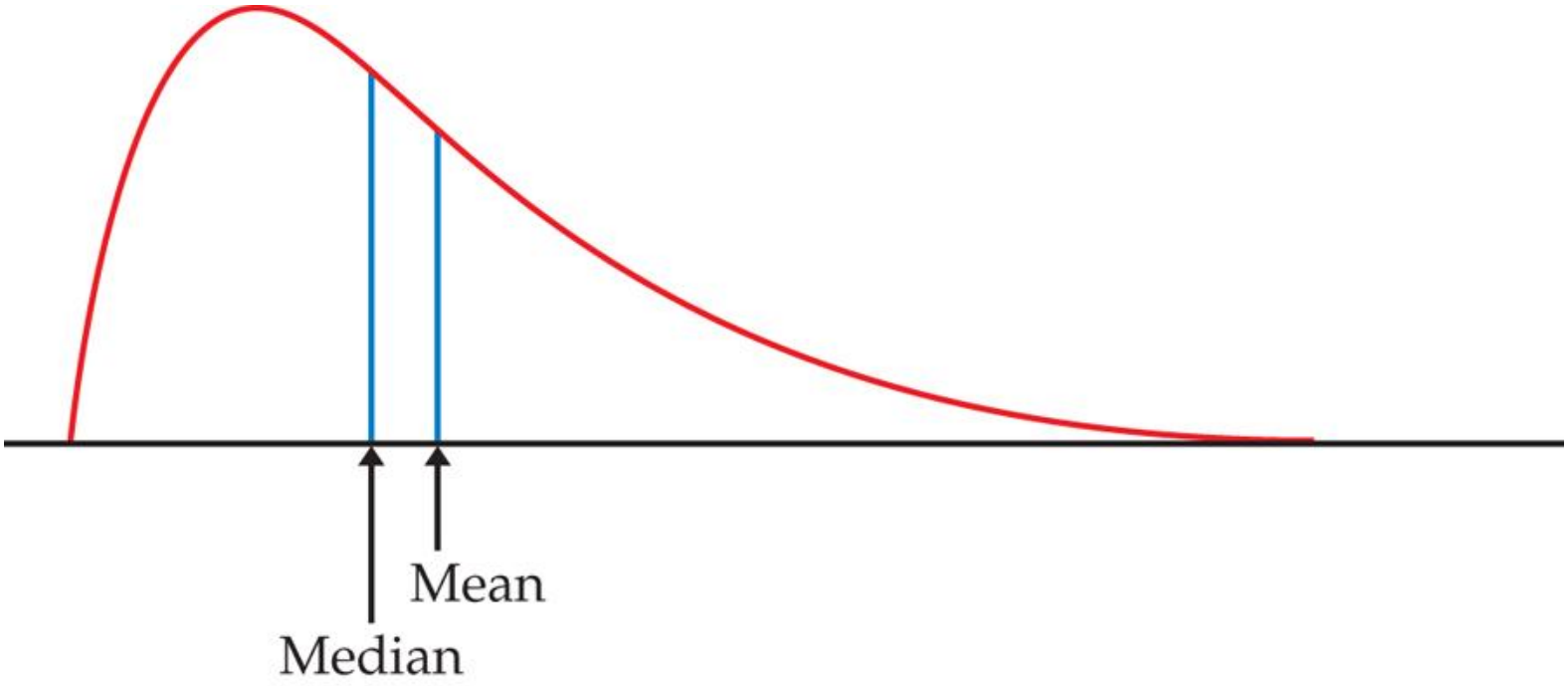
Mål for senter og spredning kan brukes for tetthetskurver på samme måte som for fordelingen til observasjonene, dvs. den empiriske fordelingen.

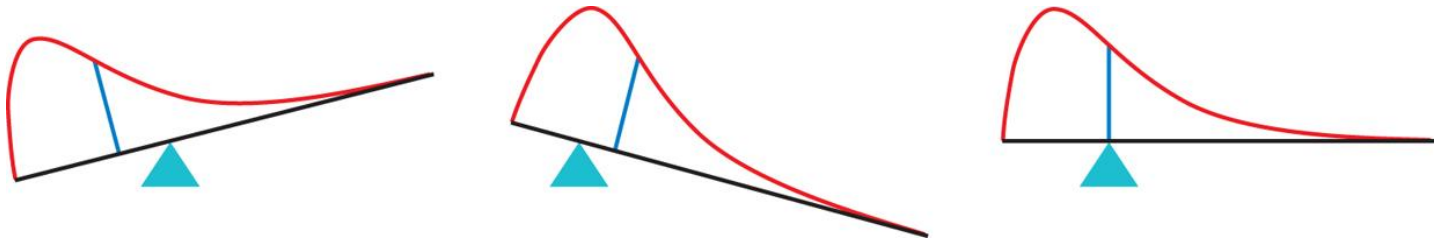
Medianen svarer til det punktet som deler arealet under tetthetskurven i to like store deler.

For de teoretiske tetthetskurven svarer **forventingen** til gjennomsnitt i empiriske fordelinger. På engelsk brukes «mean» om begge deler.

Forventningen kan oppfattes som et tyngdepunkt.


Median and mean


Mean
Median



Normalfordelingskurven er en særdeles viktig tetthetskurve.

- Den er helt bestemt ved å spesifisere forventnigen μ og standardavviket σ.
- Den er symmetrisk rundt forventningen μ som derfor er senteret i fordelingen.
- Standardavviket σ bestemmer spredningen og kan beskrives som det punktet der tetthetskurven endrer krumning.
- Fordelingen betegnes med $N(\mu,\sigma)$.

Normalfordelingen er viktig fordi.

- Den er ofte en god beskrivelse av aktuelle data, for eksempel testdata og mange biologiske data.

- Den gir en god beskrivelse av resulatet av tilfeldige forsøk som gjentas mange ganger, for eksempel myntkast.

- Statistiske prosedyrer utarbeidet under forutsetning at observasjonene beskrives med en normalfordeling, virker godt for data fra tilnærmet symmetriske fordelinger.

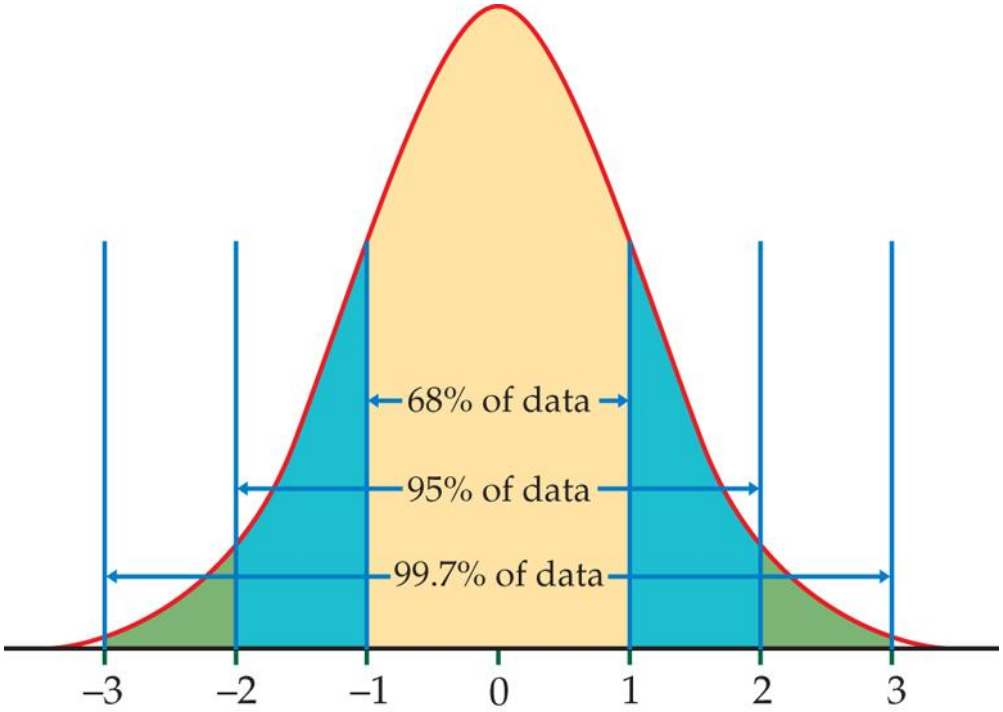
« Å anta at tetthetskurvene er symmetrisk er 90% på vei til å anta at den er normal»

68-95-99.7 regel for en N(μ,σ) fordeling:

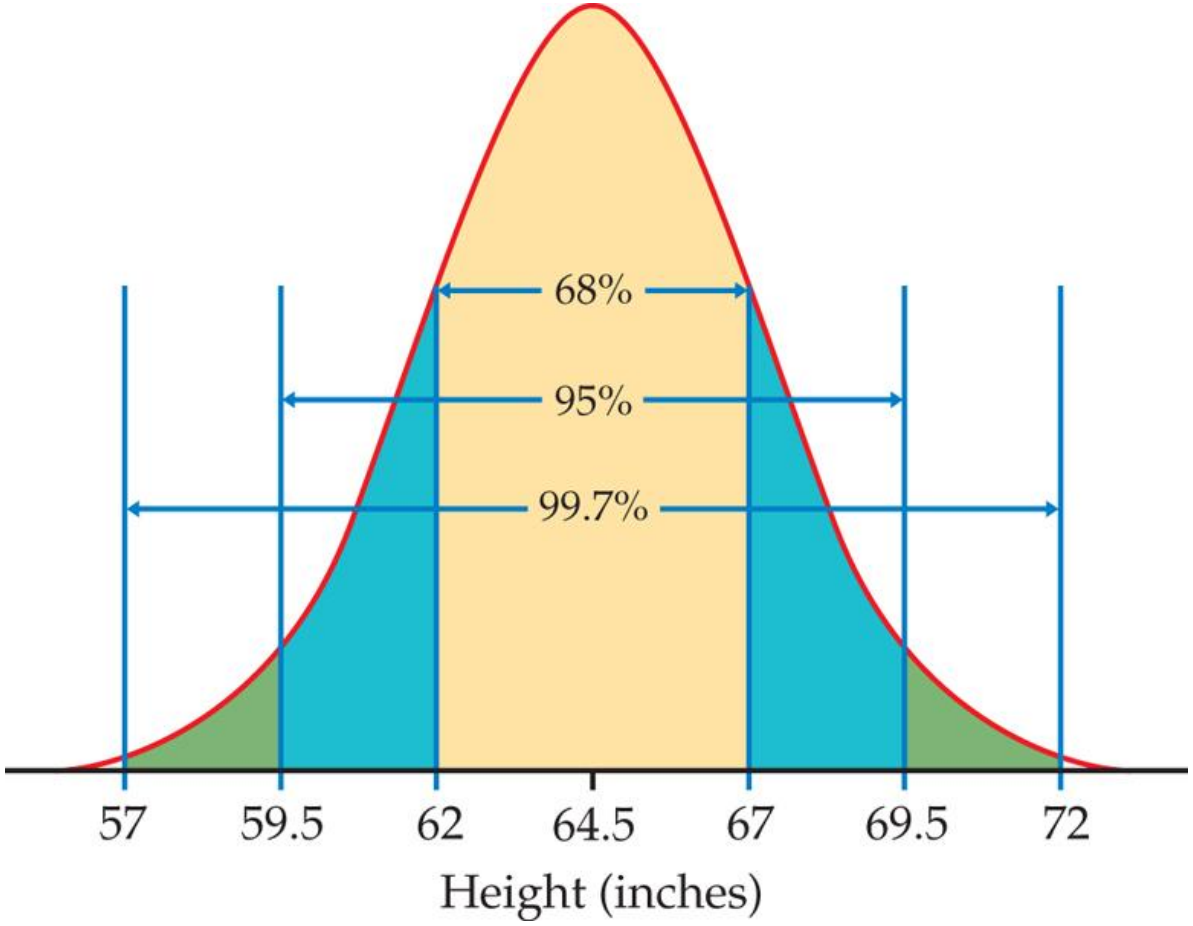
68% av observasjonene ligge i intervallet (μ-σ, μ+σ).

95% av observasjonene ligge i intervallet (μ-2σ, μ+2σ).

99.7% av observasjonene ligge i intervallet (μ-3σ, μ+3σ).



Eksempel Fordelingen av høyden til (amerikanske) kvinner mellom 18 og 24 år er tilnærmet normalfordelt med forventning 64.5 tommer og standardavvik 2.5 tommer.



La x være en observasjon fra en fordeling med forventning μ og standardavvik σ.

Standarisert verdi elle z-skåre: $z = \frac{x-\mu}{\sigma}$

z-skåren angir hvor mange standardavvik observasjonen ligger fra forventningen μ.

Eksempel For unge amerikanske kvinner er forventningen 64.5 Og standardavviket 2.5 tommer. Det betyr at en kvinne som er 68 tommer høy har en z-skåre på z= (68-64.5)/2.5=1.4, eller 1.4 standardavvik over forventningen.

Notasjon: En variabel som er gitt ved en teoretisk fordeling som for eksempel den normale $N(\mu,\sigma)$, betegner vi med store bokstaver som X.

Verdien den antar betegnes med små bokstaver x.

I eksemplet ovenfor vil X være høyden til unge amerikanske kvinner og «X <68» betyr derfor «høyden til en ung amerikansk» kvinne er mindre enn 68 tommer».

Merk:

Standarisering er eksempel på en lineær transformasjon.

Standariserte variable har forventning 0 og standardavvik 1.

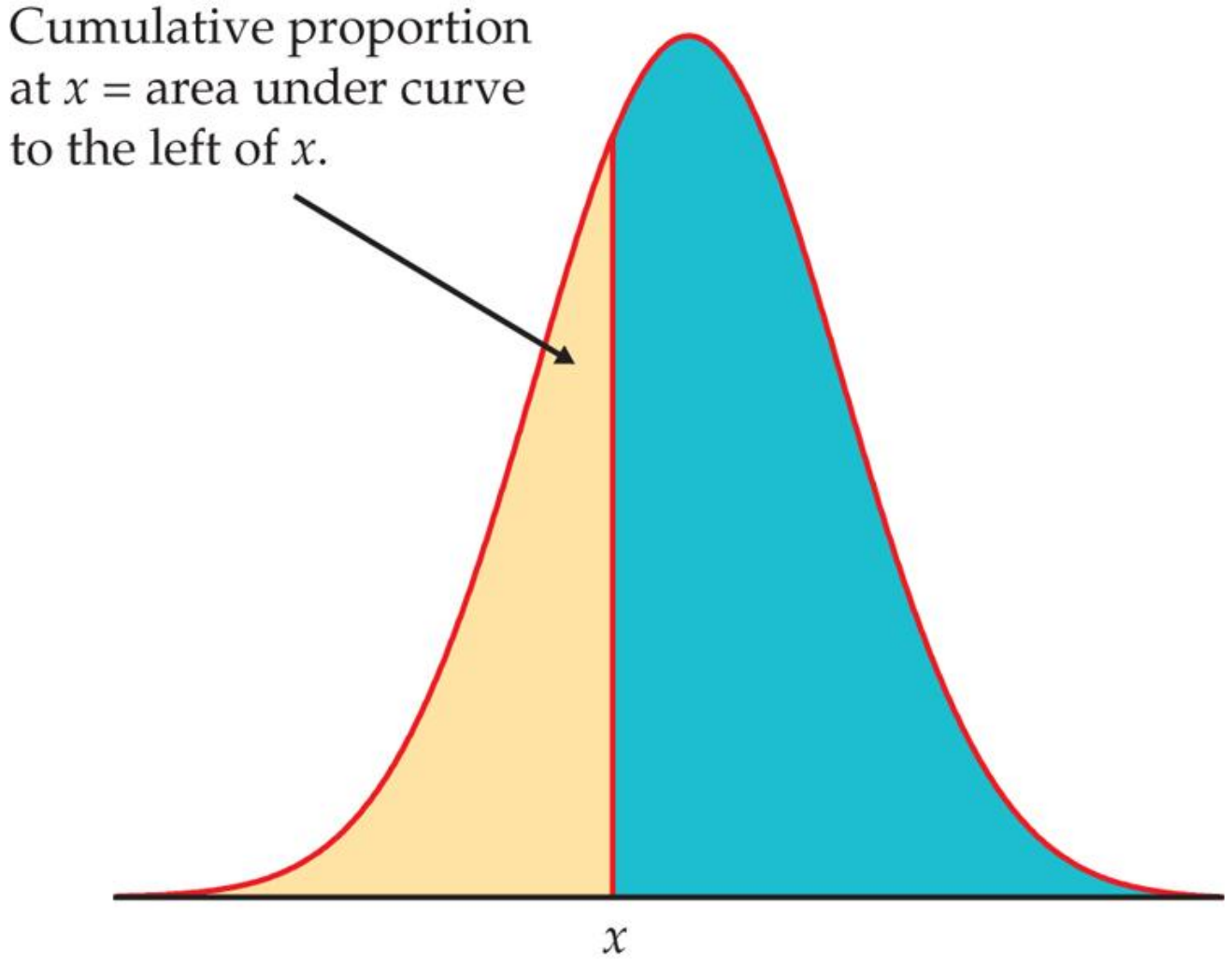
Standariserte normalfordelte variable er normale. Hvis
X er $N(\mu,\sigma)$ fordelt, er $Z=(X-\mu)/\sigma$ $N(0,1)$ fordelt.

Fordelingen N(0,1) kalles standard normalfordelingen.

Arealet under en tetthetskurve angir andeler.

Alle statistikk pakker kan beregne slike andeler for de vanligste teoretiske fordelingene spesielt N(μ,σ) fordelingen.
Da må arealet spesifiseres og også μ og σ.

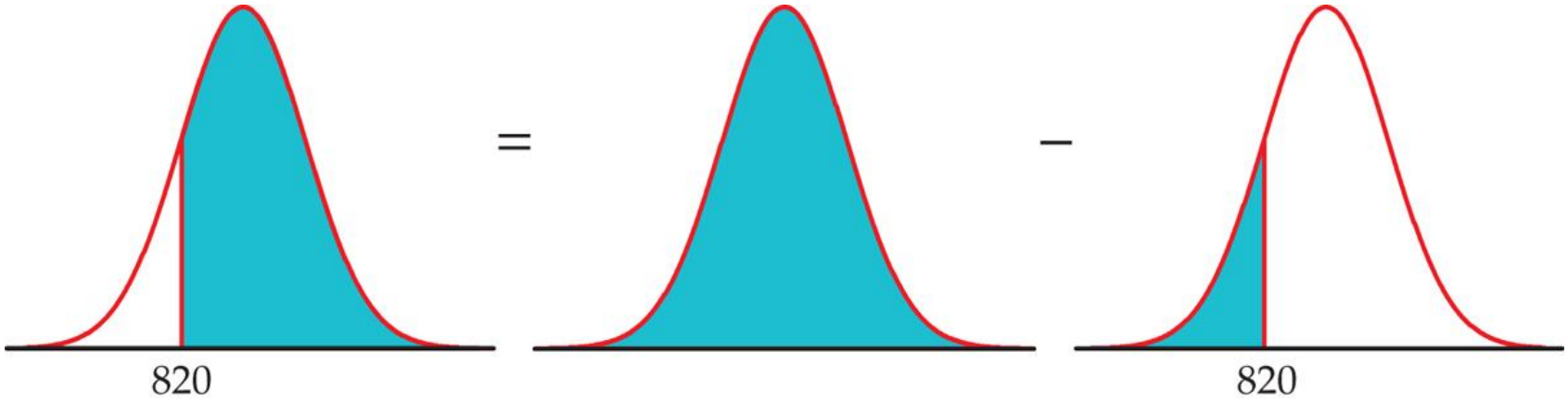
Arealet til venstre for en verdi x kalles kumulative andeler, og arealene svarer til intervallene $(-\infty, x)$.

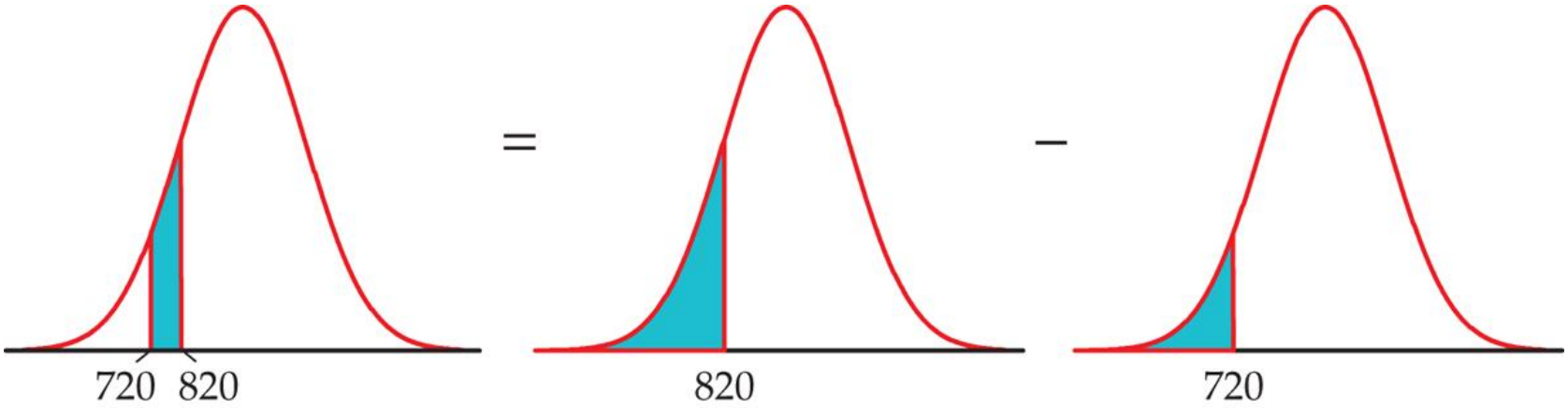

Cumulative proportion
at x = area under curve
to the left of x.
x

Siden det totale areal under tetthetskurver er 1, kan de kumulative andelene brukes til å finne andeler over andre arealer enn $(-\infty, x)$.

Eksempel Resultatene fra de vanlige amerikanske SAT testene for verbale og matematiske ferdigheter kan tilnærmes med en N(1026,209) fordeling.

Kravet for medlemmer av idrettslagene er 820. Andelen som får bedre resultat enn 820, og andelen som ligger mellom 720 og 820, kan derfor illustreres som følger.


=
−
820
820


720
820
=
820
–
720

For standard normalfordelingen er arealet til venstre for z tabullert for z=-3.49, -3.48,...,3.48,3.49 i tabell A bak i læreboka, dvs de kumulative andelene som svarer til intervallene $(-\infty, z)$.

Disse tabellene kan brukes til å beregne $X > 820$ i eksemplet med SAT tester.

«$X > 820$» svarer til

«$Z=(X -1026)/209 > (820-1026)/209=-0.99$».

Fra tabell A i boka er andelen observasjoner mindre enn -0.99 lik 0.1611 slik at andelen som er større er 1-0.1611=0.8379.

Man kan også stille det omvendt spørsmål: Hva er resultatet ,x, som gjør at 10% av studentene har bedre resultat enn x?

Statistikkpakker gir svar, men tabellene kan også brukes.

Fra tabell A i boka ser en at andel 0.9 svarer omtrent til Z=1.28. Da er $1.28=z=(x-1026)/209$ som gir $x=1026 + 209 \times 1.28= 1293.52$.

Histogrammer kan brukes til å vurdere om den empiriske fordelingen i et datasett kan tilnærmes med en normalfordeling.

Normalfordelingsplott er bedre.

Normalfordelingsplott:

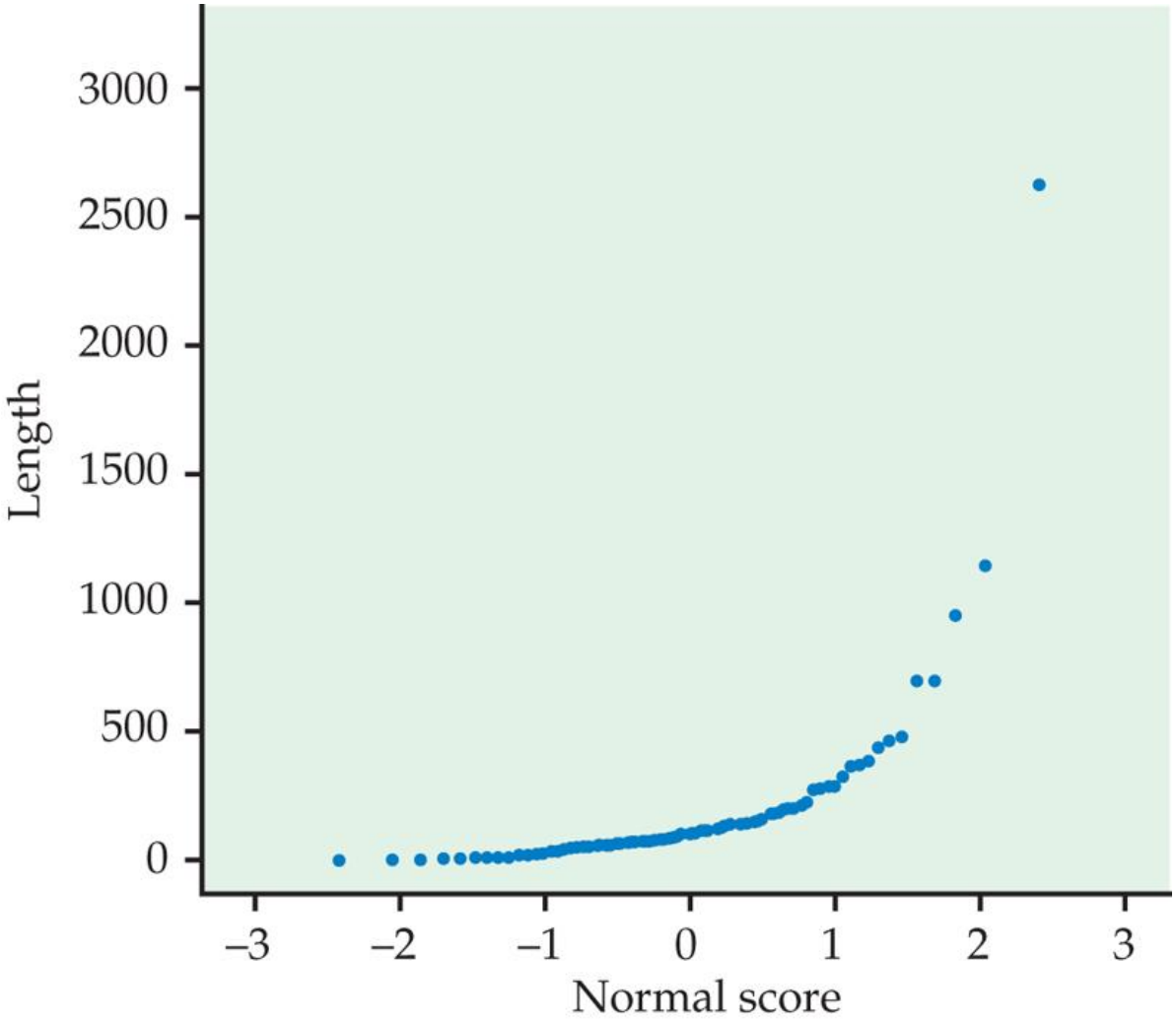
1. Sorter eller ordn dataene i stigende rekkefølge og bestem persentilene som verdiene svarer til.

2. Finn z-verdiene som svarer til persentilene fra punkt 1.

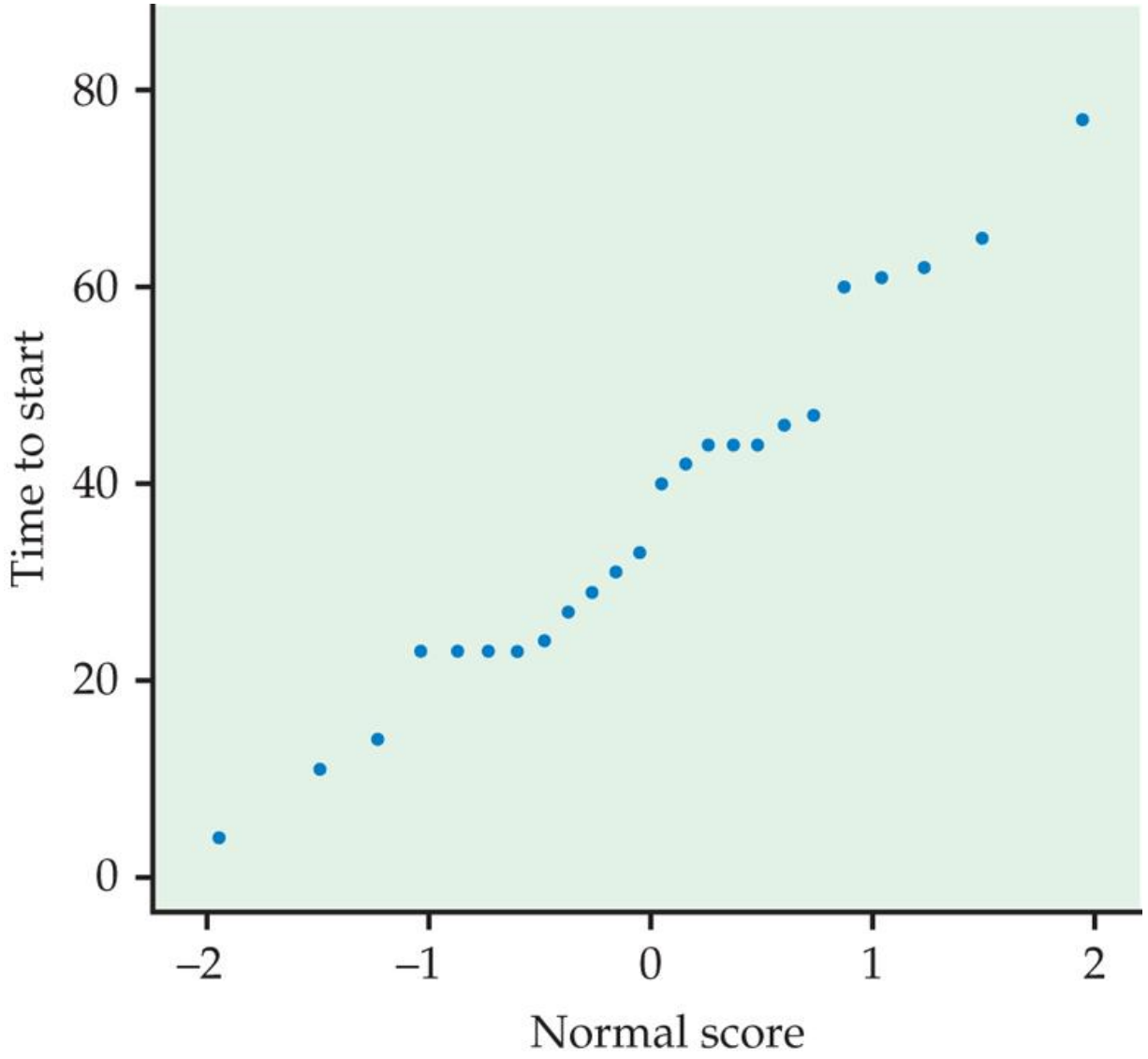
3. Plott verdiene til datapunktene mot z-verdiene fra punkt 2.

Merk Tilnærmede rette linjer indikerer at normalfordelingen er en god tilnærmelse. Store avvik tyder på ikke-normalitet.

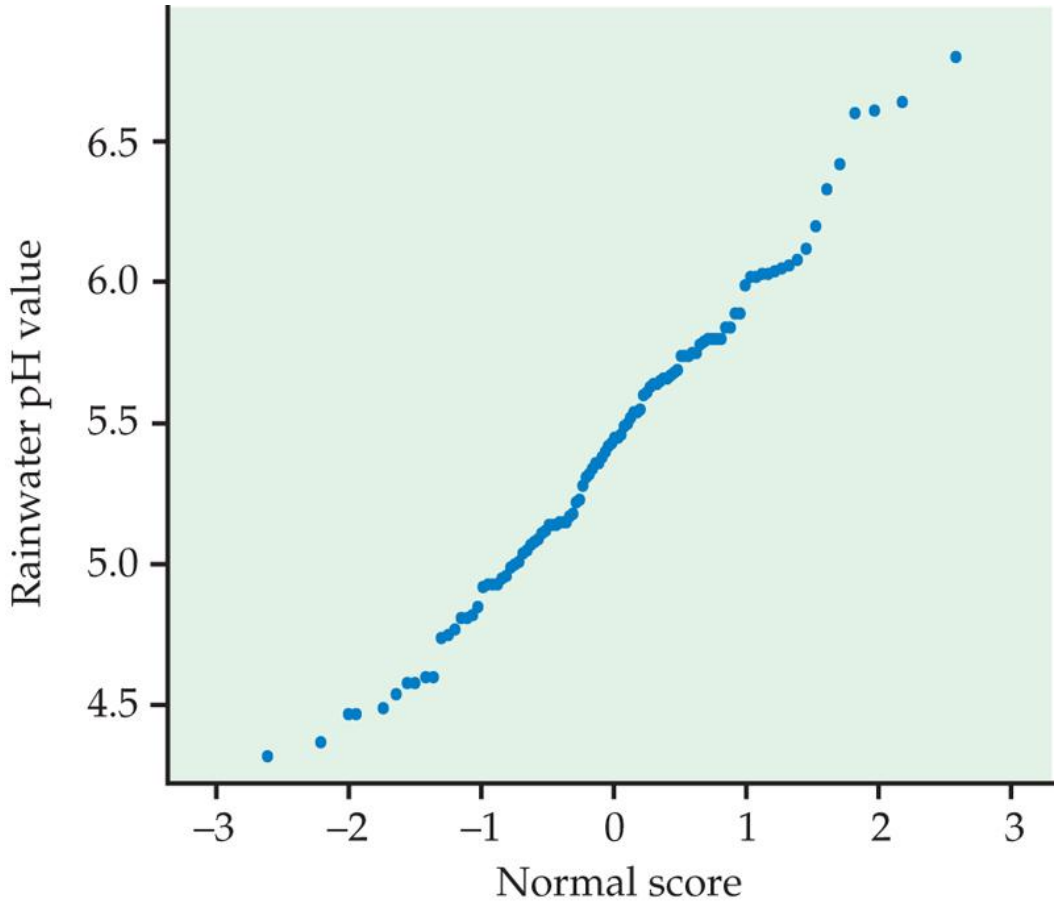
Outliere viser seg som avvikende punkter.



Lenge av 80 samtaler i call senter.



Antall dager til å starte foretak.

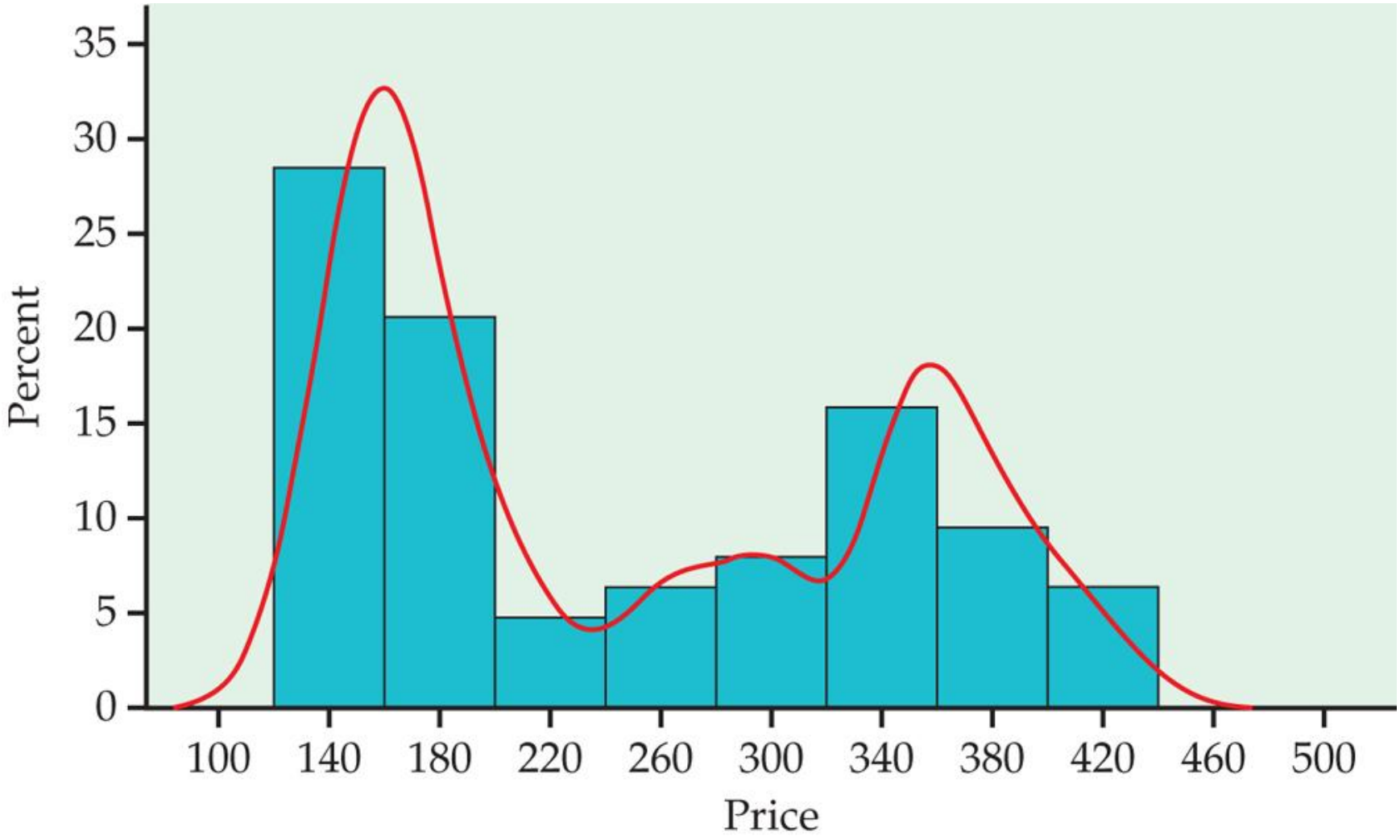


Data for surhet i 105 nedbørsprøver.

Moderne statistikkpakker har «automatiske» tetthetsestimatorer, som i praksis betyr at histogrammene tilnærmes med glatte kurver.

Nedenfor er et eksempel for pris på billetter til et idrettarrangement fra en wedside som auksjonerer slike.

Fordelingen ser bimodal ut.


Percent
35
30
25
20
15
10
5
0
100
140
180
220
260
300
340
380
420
460
500
Price