2. Grafiske framstillinger.

2.1. Innledning.

Etter at vi har samlet inn våre data, må de bearbeides for at de kan gi nyttig informasjon. Slik bearbeiding kan bestå av at data framstilles grafisk, og / eller at visse nøkkeltall beregnes. Tidligere ble dette gjort manuelt. I dag brukes helst dataverktøy. Men de fleste teknikkene som brukes i dag, har sin bakgrunn i manuelle metoder.

Jeg skal benytte poeng fra en eksamenssensur, der 0 poeng er en helt verdiløs besvarelse mens 100 poeng er en perfekt besvarelse. Dataene ser slik ut:

44	**51**	**24**	**24**	**44**	**55**	**44**	**42**	**41**	**11**	**100**	**83**
58	**41**	**42**	**37**	**56**	**24**	**20**	**49**	**20**	**40**	**43**	**31**
41	**26**	**64**	**35**	**15**	**9**	**37**	**45**	**7**	**24**	**17**	**41**
6	**9**	**52**	**42**	**10**	**53**	**20**	**43**	**83**	**54**	**16**	**83**
35	**66**	**19**	**44**	**50**	**22**	**30**	**57**	**65**	**63**	**48**	**36**
22	**79**	**49**	**25**	**65**	**50**	**35**	**63**	**21**	**43**	**55**	**24**
44	**55**	**17**	**14**	**62**	**28**	**33**	**8**	**67**	**55**	**32**	**42**
37	**8**										

Det kan ofte lønne seg å starte med å sortere dataene i stigende rekkefølge. Da får vi:

6	**7**	**8**	**8**	**9**	**9**	**10**	**11**	**14**	**15**	**16**	**17**
17	**19**	**20**	**20**	**20**	**21**	**22**	**22**	**24**	**24**	**24**	**24**
24	**25**	**26**	**28**	**30**	**31**	**32**	**33**	**35**	**35**	**35**	**36**
37	**37**	**37**	**40**	**41**	**41**	**41**	**41**	**42**	**42**	**42**	**42**
43	**43**	**43**	**44**	**44**	**44**	**44**	**44**	**45**	**48**	**49**	**49**
50	**50**	**51**	**52**	**53**	**54**	**55**	**55**	**55**	**55**	**56**	**57**
58	**62**	**63**	**63**	**64**	**65**	**65**	**66**	**67**	**79**	**83**	**83**
83	**100**										

Når vi skal karakterisere et slikt datasett, er det nyttig å vite hva vi skal se etter. De viktigste egenskapene er:

- **Midtpunkt:** Finn et eller annet mål for midtpunktet i datasettet.
- **Spredning:** Ligger alle dataene relativt samlet, eller er det stor spredning rundt midtpunktet?
- **Fordeling:** Er dataene pent samlet nær midtpunktet, eller fins det flere ”sentra”? Er fordelingen symmetrisk, eller strekkes fordelings-kurven ut til den ene siden?
- **Ekstrem-verdier:** Er det noen verdier som avviker påfallende fra resten av dataene?

Etter å ha klargjort dataene, skal jeg gruppere dataene og lage grafiske framstillinger. I et senere notat skal jeg bruke de samme dataene til å beregne nøkkeltall for dataene.

2.2. Gruppering, frekvenstabell.

For å se hvordan dataene fordeler seg, er det vanlig å sette opp en *frekvenstabell*. Vi deler da inn dataene i et passe antall *klasser*. Det bør være minst 5 klasser og maksimalt ca. 15 klasser (noen lærebøker anbefaler litt andre tall). En røff regel er at antall klasser kan være omtrent kvadratrota av antall dataelement (eller litt lavere).

Like viktig er det at vi får ”pene” *klassegrenser* og ”pen” *klassevidde*. I vårt eksempel med poengsummer til eksamen kan det være fristende å bruke karaktergrensene som klassegrenser.

Fra et statistisk synspunkt er det mye bedre å benytte lik klassevidde for alle klassene. Denne bredden bestemmes ved at vi først beregner

$$\text{klassebredde} = \frac{\text{største dataverdi - minste dataverdi}}{\text{antall klasser}}$$

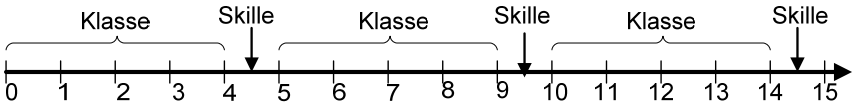
og avrunder slik at vi får "pene" tall både for klassegrenser og klassevidde.

I vårt eksempel har vi 86 dataelementer. Dette skulle tilsi ca. 9 klasser. Da blir

$$\text{klassebredde} = \frac{100-6}{9} \approx 10.4\,.$$

Vi justerer oss litt for å få "pene" tall, og prøver oss med 10 klasser og klassebredde 10.

Nå må vi se nærmere på *klassegrensene*, og forlater eksemplet vårt et øyeblikk. Hvis for eksempel dataene våre består av heltallene 0, 1, 2, … , og vi har valgt klassebredde på 5, vil klassene inneholder tallene 0 … 4, 5 … 9, 10 … 14 osv. Da er 0, 5, 10 osv. *nedre klassegrenser*, mens 4, 9, 14 osv. blir *øvre klassegrenser*. Vi kan da definere et *skille* mellom klassene på 4.5, 9.5, 14.5 osv. slik figuren nedenfor viser:



De engelske betegnelsene er *lower* og *upper class limits*, mens skillet mellom klassene kalles *class boundary*. Vi ser at klassebredden kan defineres som avstanden fra ett klasseskille til det neste, eller som avstanden mellom nedre grense i en klasse og nedre grense i neste klasse. Eller vi kan benytte avstanden mellom de øvre grensene i to naboklasser.

Hittil har vi kun sett på *heltallige* data. Men i praksis har vi ofte *kontinuerlige* data, for eksempel dersom dataene er lengde, vekt, tid eller liknende. Med slike data og klassebredde på 7, er det naturlig å definere klassene slik:

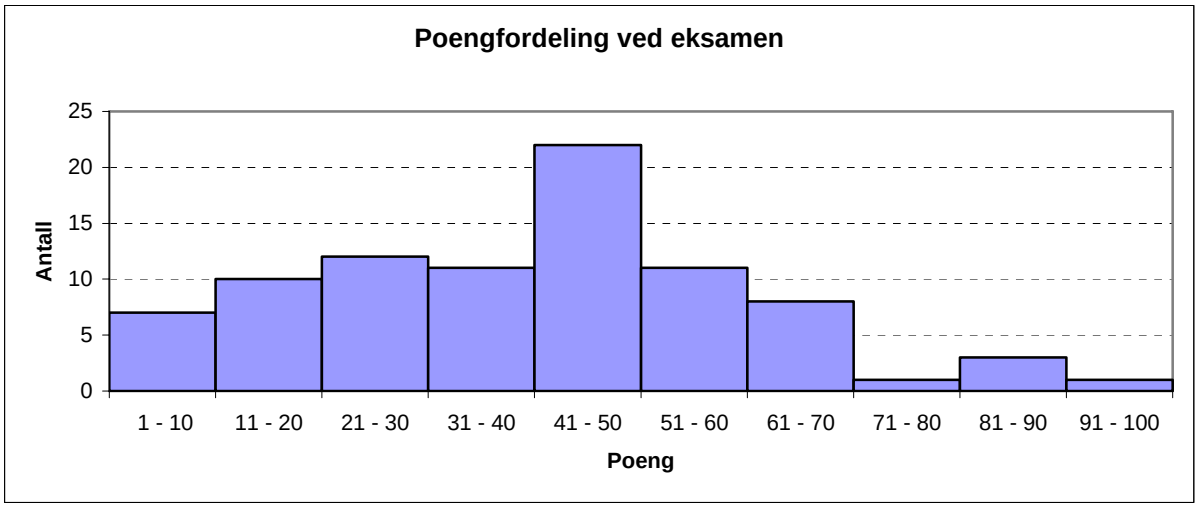
$[0,7\rangle$, $[7,14\rangle$, $[14,21\rangle$ osv.

Med slike data kan vi kun definere en av klassegrensene eksakt (i dette tilfellet nedre klassegrense), og det er ikke mulig å definere noe skille mellom klassene (*class boundary*). Klassevidden blir da avstanden mellom de nedre grensene i to naboklasser.

Nå skal vi vende tilbake til vårt eksempel med poeng til eksamen. Vi ser at ingen kandidater har 0 poeng. Det er naturlig siden kandidater som har levert blankt ikke er tatt med i denne oversikten. Vi lar da den nederste klassen inneholde verdiene 1 … 10, neste klasse inneholder 11 … 20 osv. til øverste klasse som inneholder 91 … 100. Da får vi 10 klasser med klassebredde 10 slik vi ønsket. Nå gjenstår det bare å telle opp antall dataelementer i hver klasse. Dette antallet kalles *frekvensen* for denne klassen. Med våre data får vi:

Poeng	**Antall**
1 - 10	7
11 - 20	10
21 - 30	12
31 - 40	11
41 - 50	22
51 - 60	11
61 - 70	8
71 - 80	1
81 - 90	3
91 - 100	1

Når dataene i tabellen ovenfor tegnes opp grafisk får vi et diagram som vist nedenfor. Et slikt diagram kalles et ***histogram***.

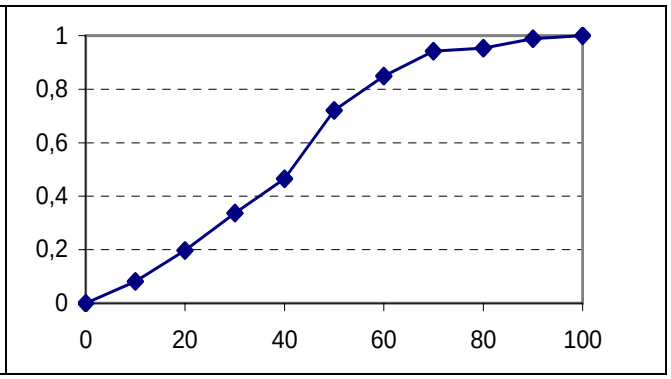


På figuren har vi angitt *antallet* (*frekvensen*) langs *y*-aksen. Vi kunne også benyttet *andelen* eller *prosentvis andel* i hver klasse langs *y*-aksen. Da ville vi fått et ***relativ frekvens-histogram***. En annen mulighet hadde vært å dele antallet eller andelen på klassebredden, og føre opp dette tallet langs *y*-aksen. Da ville vi fått frekvens eller relativ frekvens pr enhet som dataene er målt med. En slik framgangsmåte ville vært naturlig dersom vi ikke hadde hatt lik bredde på alle klassene, eller dersom vi vil sammenlikne flere histogram som ikke har samme klassebredde.

2.3. Kumulert frekvens.

Det er ofte nyttig å vite hvor stor del av dataene som har en verdi mindre enn (eller lik) en viss grense. Med utgangspunkt i vår frekvenstabell skaffer vi oss en slik oversikt ved å summere frekvensene i klasser med dataverdier som er mindre enn eller lik vår grense. Da får vi en *kumulert frekvenstabell*. Hvis vi i tillegg deler på det totale antallet, får vi en *relativ kumulert frekvens-tabell*. Med dataene i vårt eksempel får vi tabellen til venstre nedenfor:

Øvre klasse-grense	Frekvens	Kumulert frekvens	Relativ kumulert frekvens
0	0	0	0,00
10	7	7	0,08
20	10	17	0,20
30	12	29	0,34
40	11	40	0,47
50	22	62	0,72
60	11	73	0,85
70	8	81	0,94
80	1	82	0,95
90	3	85	0,99
100	1	86	1,00



Når den relative kumulerte frekvensen plottes som funksjon av klassegrensene, får vi diagrammet ovenfor til høyre. Det gir god oversikt over hvor stor andel av dataene som ligger under bestemte verdier. Av diagrammet (og tabellen) ser vi for eksempel at 72 % av kandidatene oppnådde 50 poeng eller mindre til eksamen.

2.4. Andre former for grafiske framstillinger.

Det fins mange andre former for grafiske framstillinger. De som er nevnt ovenfor, er (etter min oppfatning) de viktigste. Men du kan også komme over mange andre typer, for eksempel:

- ***Punktplot*** (*dotplot)* er en enkel form for histogram som brukes noe dersom samme verdi går igjen mange ganger. Man avsetter da alle aktuelle verdier langs en akse, og avsetter en prikk (eller et annet merke) hver gang en verdi forekommer.

- ***Stamme-og løv-plot*** minner om punktplot, men brukes dersom de aktuelle data-verdiene angis med to sifre. Da avsettes første siffer langs aksen, og andre siffer ”stables” vinkelrett på aksen. Til slutt sorteres gjerne andre-sifrene. Når våre eksamens-data framstilles i et ”stamme-og-løv”-plot, får vi:

0	6 7 8 8 9 9
1	0 1 4 5 6 7 7 9
2	0 0 0 1 2 2 4 4 4 4 4 5 6 8
3	0 1 2 3 5 5 5 6 7 7 7
4	0 1 1 1 1 2 2 2 2 3 3 3 4 4 4 4 4 5 8 9 9
5	0 0 1 2 3 4 5 5 5 5 6 7 8
6	2 3 3 4 5 5 6 7 9
7	9
8	3 3 3
9	
10	0

  Kolonnen til venstre er ”stammen” (tier-tallet i poengsummen). ”Løvet” eller små-greinene er ener-tallene i poengsummene, som listes opp horisontalt i sortert rekkefølge. Vi ser at vi får en slags liggende histogram. Merk at klasseinndelingen i ”stamme-og-løv”-plottet må bli 0 … 9, 10 … 19 osv., ikke 1 … 10, 11 … 20 osv. som i histogrammet.

- ***Pareto-grafer*** er en form for histogram der hver søyle representerer antall elementer som har en bestemt ikke-numerisk egenskap (farge, nasjonalitet, politisk parti, …). Søylene sorteres deretter i synkende rekkefølge.

- ***Kake-diagram*** består av en sirkel som er delt opp i segmenter (”kakestykker”) der arealet av hvert segment avspeiler antall dataelementer som hører hjemme i det segmentet. Det brukes ofte istedenfor Pareto-graf.

Dersom vi skal illustrere en sammenheng mellom to variabler (for eksempel høyde og vekt hos voksne menn), brukes et ***spredningsdiagram*** (*scatter plot*). Her avsettes samhørende verdier av de to variablene som punkter i et koordinatsystem. Hvis det er naturlig å anta at den ene variabelen avhenger av den andre, avsettes gjerne den avhengige variabelen langs *y*-aksen. I et spredningsdiagram for høyde og vekt er det naturlig å avsette høyde langs *x*-aksen mens vekten avsettes langs *y*-aksen for å illustrere hvordan vekten avhenger av høyden.

Dersom vi skal illustrere hvordan noe avhenger av *tiden*, bruker vi en ***tidsserie-graf***. Som navnet antyder, avsettes tid langs *x*-aksen mens den størrelsen som avhenger av tiden avsettes langs *y*-aksen.