Beskrivende statistikk

Notat til STK1100

Ørnulf Borgan
Matematisk institutt
Universitetet i Oslo

Januar 2004

Innledning

I forbindelse med eksperimentelle studier i fysikk og kjemi, feltstudier i biologi, kliniske utprøvninger i medisin, meningsmålinger, studier av påliteligheten av tekniske komponenter, osv. samles det inn data for å belyse de relevante problemstillingene. Mengden av data kan være svært varierende fra studie til studie, fra et titalls til hundre- eller tusenvis av observasjoner.

Formålet med en statistisk analyse av et datasett er å trekke slutninger om det fenomenet som studeres på bakgrunn av de foreliggende data. For å kunne ta hensyn til "usikkerheten" (tilfeldige variasjoner) knyttet til observerte data vil en slik analyse oftest være basert på en sannsynlighetsmodell av det fenomenet som studeres. For å få oversikt over viktige trekk ved et datamateriale er det imidlertid også verdifullt (ofte i startfasen av en mer omfattende analyse) å benytte beskrivende statistiske metoder. Her skal vi kort beskrive noen slike.

Histogram

En nyttig måte å illustrere grafisk hvordan observasjonene i et datamateriale fordeler seg, er ved et histogram. En "oppskrift" for å lage et (normert) histogram er som følger:

- Finn den minste og den største observasjonen i datamaterialet.
- Velg et antall disjunkte delintervaller (som ikke alle trenger å være av samme lengde) som tilsammen dekker hele intervallet fra den minste til den største observasjonen.
- Tell opp antall observasjoner som faller i hvert delintervall.
- Bestem den relative frekvensen for hvert delintervall ved å dividere antall observasjoner i delintervallet på det samlede antall observasjoner.
- Avmerk delintervallene langs x-aksen og tegn for hvert av disse et rektangel med delintervallet som grunnlinje og areal lik delintervallets relative frekvens.

Det diagrammet som framkommer ved denne oppskriften er et (normert) histogram.

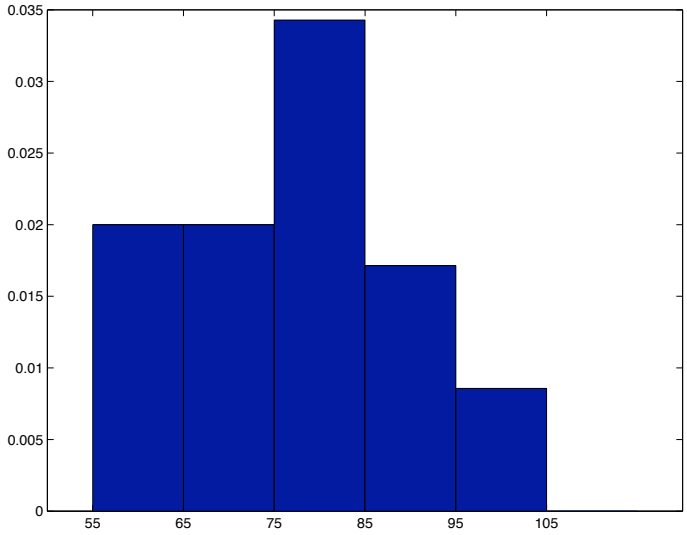
Eksempel

Nedenfor har vi gitt hvilepulsen til 35 kvinnelige studenter:

96	62	78	82	100	68	96	78	88	62	80	84
61	64	94	60	72	58	88	66	84	62	66	80
78	68	72	82	76	87	90	78	68	86	76	

I intervallene $[55\,,65)$, $[65\,,75)$, $[75\,,85)$, $[85\,,95)$ og $[95\,,105)$ er det henholdsvis 7, 7, 12, 6 og 3 pulsmålinger. De tilhørende relative frekvensene blir derfor 7/35=0.20, 7/35=0.20, 12/35=0.34, 6/35=0.17 og 3/35=0.09. I histogrammet har stolpene (rektanglene) areal lik disse relative frekvensene. Da grunnlinjen er 10 enheter lang, blir høyden av stolpene en tidel av de relative frekvensene.

Histogrammet blir som på figuren:



Figur 1: Normert histogram for pulsmålingene.

Gjennomsnitt og empirisk median

I tillegg til en grafisk framstilling som illustrerer hvordan et datamateriale fordeler seg, vil en ofte også være interessert i ett (eller flere) summarisk(e) mål for "den representative verdi" i datamaterialet, dvs. den verdi som observasjonene fordeler seg omkring. Det fins flere mål som benyttes til dette formålet. De to viktigste er gjennomsnittsverdien og den empiriske median. Den første beregnes som navnet antyder som gjennomsnittet av alle observasjonene, mens den andre er den midterste av observasjonene når disse ordnes i stigende rekkefølge fra den minste til den største. (Hvis det er et like antall observasjoner, tar en gjennomsnittet av de to midterste observasjonene.)

Eksempel (fortsatt)

Den gjennomsnitlige hvilepulsen til de 35 kvinnelige studentene er 76.7. Den empiriske medianen er 78, dvs. observasjon nummer 18 når observasjonenen ordnes i stigende rekkefølge. Da pulsdataene er nokså symmetrisk fordelt, blir det liten forskjell på gjennomsnitsverdien og den empiriske medianen.

Empirisk standardavvik og kvartildifferanse

Det er også av interesse å beregne empiriske mål som angir hvor mye observasjonene varierer rundt "den representative verdien". Det viktigste variasjonsmålet er det empiriske standardavviket. Dette er gitt ved

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

der $x_1, \ldots x_n$ er våre observasjoner og $\overline{x}$ er deres gjennomsnittsverdi.

Et annet viktig variasjonsmål er den empiriske kvartildifferansen. Denne er definert som følger: Nedre empiriske kvartil Q_1 er den verdien som er slik at 25% av observasjonene er mindre enn denne. Øvre empiriske kvartil Q_3 er tilsvarende den verdien som er slik at 25% av observasjonene er større enn denne. Den empiriske kvartildifferansen er differansen mellon øvre og nedre empiriske kvartil, dvs. $Q_3 - Q_1$.

Eksempel (fortsatt)

Det empiriske standardavviket til hvilepulsen for de 35 kvinnelige studentene er 11.6. Når observasjonene ordnes i stigende rekkefølge, blir nedre empiriske kvartil observasjon nummer ni. Altså har vi $Q_1 = 66$. Tilsvarende blir øvre empiriske kvartil observasjon nummer 27 når vi ordner de i stigende rekkefølge, dvs. $Q_3 = 86$. Den empiriske kvartildifferansen er dermed $Q_3 - Q_1 = 86 - 66 = 20$.