BESKRIVANDE STATISTIK

1. GRUNDBEGREPP

Följande begrepp används ofta vid beskrivning av ett statistiskt material:

LÄGESMÅTT (medelvärde, median och **typvärde)**:

Låt D=[x_1,x_2,...,x_n] vara en tallista som beskriver ett statistiskt material.

i) **Medelvärdet** (mer precis "det aritmetiska medelvärdet") betecknas oftast med μ eller $\bar{x}$

och definieras som $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$.

ii) **Median:** För att bestämma medianen sorterar vi tallista. Om mängden har udda antal element väljer vi det mittersta element i den sorterade tabellen. Om antalet element är jämnt då är medianen medelvärdet av två mittersta element i den sorterade tabellen.

iii) **Typvärde** i en tallista är det värde som förekommer flest gånger.

SPRIDNINGSMÅTT (Varians och standardavvikelse)

iv) **Varians och standardavvikelse** är statistiska mått som visar hur mycket de olika värdena i ett statistiskt material avviker från medelvärdet. Om värdena ligger nära medelvärdet blir standardavvikelsen liten; om värdena är spridda långt över och under medelvärdet blir standardavvikelsen stor. Följande två typer av standardavvikelse används i statistiken:

Typ1. Om datatabell D=[x_1,x_2,...,x_n] beskriver hela populationen då definieras variansen och standardavvikelse enligt följande:

$$Variansen = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Standardavvikelsen} = \sigma = \sqrt{\text{Variansen}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Typ 2. Om vi inte har data för hela populationen, utan ett **stickprov,** då används så kallade **stickprovets varians och stickprovets standardavvikelse.** I detta fall använder vi följande formel ("n-1 formel")

$$Variansen = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Standardavvikelsen} = \sigma = \sqrt{\text{Variansen}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

där $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$

(Anmärkning: I några böcker använder man beteckning μ och σ för medelvärde och standardavvikelse endast om materialet täcker hela populationen, medan $\bar{x}$ och s (eller μ_{obs} och σ_{obs}) betecknar medelvärde och standardavvikelse i fall det handlar om ett stickprov.)

v) Variationsbredden är skillnaden mellan max- och min-värdet i en datatabell

Uppgift 1. Beräkna

a) medianen b) medelvärdet c) typvärde
d) stickprovets varians, (dvs (n-1)-formel)
e) stickprovets standardavvikelse, (dvs (n-1)-formel)
f) största, minsta värde och variationsbredden

för följande data: D1= [25, 22, 24, 22, 21, 23, 34, 25, 22, 33].

Lösning a) Först sorterar vi datatabell:
D1(sort)=[21, 22, 22, 22, 23, 24, 25, 25, 33, 34].
Notera att tabellen har jämnt antal (10) element
Medianen =(23+24)/2=23.5

Svar:

a) medianen=23.5 b) medelvärdet = 25.1 c) typvärde =22
d) variansen=21.43 e) standardavvikelsen =4.63
f) största värdet =34, minsta värde=21, variationsbredden =13

Uppgift 2. Beräkna

a) medianen b) medelvärdet c) typvärde
d) stickprovets varians, dvs (n-1)-formel)
e) stickprovets standardavvikelse, dvs (n-1)-formel)
f) största, minsta värde och variationsbredden

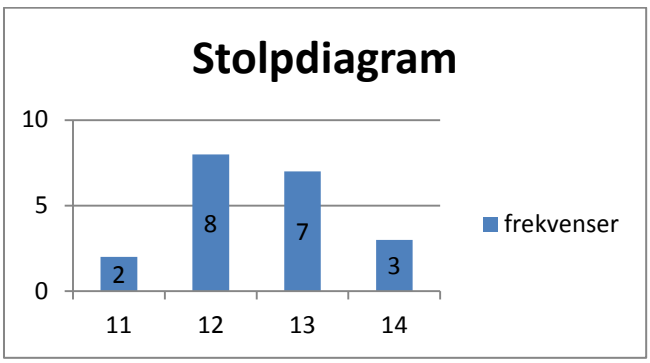
för följande data: D1= [32, 34, 32,38, 33].
a) medianen = 33 b) medelvärdet = 33.8 c) typvärde =32
d) stickprovets varians= 6.2
e) stickprovets standardavvikelse=2.49
f) största = 38, minsta värde =32 och variationsbredden =6

1. GRUPPERAT DATAMATERIAL

Anta att vi har ett stickprov med följande observationer
D2= [12, 13, 13, 12, 11, 13, 14, 12, 12, 13, 12, 14, 13, 12, 11, 13, 14, 12, 12, 13] .
I ovanstående datamängder har vi n= 20 observationer där några upprepas så att vi har k=4 olika resultat.
Vi ser att 11 förekommer 2 gånger i datamängder och säger att tillhörande frekvens är 2.
I allmänt är frekvensen f_i lika med antalet gånger elementet x_i förekommer bland observerade enheter.
Den kumulativa frekvensen till x_i visar antalet observationer som är $\leq x_i$. Den får man genom att successivt addera frekvenser.
Den relativa frekvensen är kvoten f_i /n (anges ibland i procent)
Den kumulativa relativa frekvensen är kvoten *(kumulativa frekvensen)*/n (anges ibland i procent)
Vi kan beskriva ovanstående observationer med följande tabell:

	(x_i)	frekvenser (f_i)	kumulativa frekvenser	relativa frekvenser	kumulativa relativa frekvenser
1	11	2	2	0.1	0.1
2	12	8	2+8=10	0.4	0.5 (=0.1+0.4)
3	13	7	2+8+7=17	0.35	0.85 (=0.1+0.4+0.35)
4	14	3	2+8+7+3=20	0.15	1
k=4		$n=\sum_{i=1}^{k} f_i$ =20			

Vi kan åskådligt göra frekvenser med hjälp av ett stolpdiagram



Medelvärde och (stickprovets) varians för ett frekvensindelat material kan beräknas med följande formler

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i$$

$$Variansen = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2 \ .$$

Vi har

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i = \frac{1}{20}(2\cdot 11 + 8\cdot 12 + 7\cdot 13 + 3\cdot 14) = \frac{251}{20} = 12.55$$

och

$$Variansen = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2$$

$$= \frac{1}{19}\left(2\cdot(11-12.55)^2 + 8\cdot(12-12.55)^2 + 7\cdot(13-12.55)^2 + 3\cdot(14-12.55)^2\right)$$

$$= 299/380 = 0.7868421$$

Standardavvikelse $\sigma = \sqrt{Var} = 0.887$

==========

Relativa frekvenser och sannolikheter.

Betrakta följande experiment. Vi tar på måfå ett tal från ovanstående datamängden D2.

Vad är sannolikheten att vi får 13.

Lösning:

Talet 13 förekommer 7 gånger i datamängden (dvs har frekvensen 7) bland total n=20 st observationer.

Sannolikheten att få 13, när vi tar ett tal i D2 på måfå, är

P= (antalet gynnsamma fall)/(antalet alla möjliga fall) .

Alltså

P(Vi får 13) $= \frac{g}{n} = \frac{7}{20} = 0.35$ (samma som relativa frekvensen i D2 för talet 13)

Svar P=0.35.

I allmänt gäller följande:

P(Vi får x_i) $= \frac{g}{n} = \frac{f_i}{n} =$ **den relativa frekvensen** av x_i .

Om vi tar på måfå ett element från en datamängd då är sannolikheten att få ett tal x_i lika med den **relativa frekvensen** för x_i.

På samma sätt inser vi följande samband mellan kumulerade relativa frekvenser och sannolikheter:

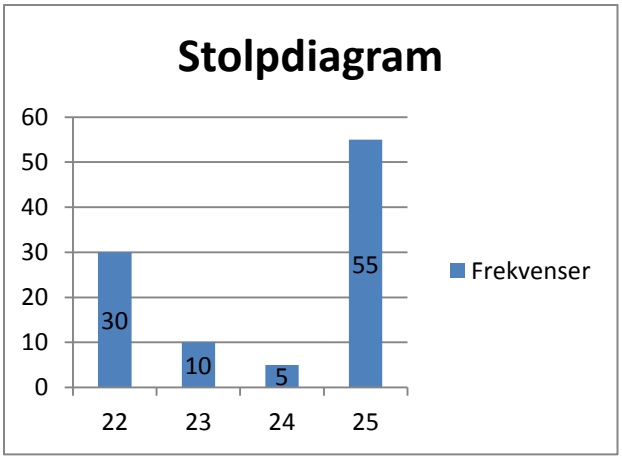
Om vi tar på måfå ett element från en datamängd då är sannolikheten att få ett tal som är $\leq x_i$ lika med den **kumulativa relativa frekvensen** för x_i.

==========

Uppgift 3. Bestäm medelvärdet, variansen och standardavvikelsen för följande klassindelat statistiskt material. Rita stolpdiagram.

k	(x_i)	frekvenser (f_i)
1	22	30
2	23	10
3	24	5
4	25	55
		$n= \sum_{i=1}^{k} f_i =100$

Svar: Medelvärdet= 23.85, variansen= 1.84596, standardavvikelsen= 1.35866



KORRELATIONSKOEFFICIENT.

JÄMFÖRELSE MELLAN TVÅ DATA-MÄNGDER

För att undersöka LINJÄRT samband mellan parade observationer

$X=(x_1,x_2,...,x_n)$ och

$Y=(y_1,y_2,...,y_n)$

används ofta korrelationskoefficient som definieras enligt följande

$$r_{XY}=\frac{\sum_{i=1}^{n}\left((x_i-\bar{x})(y_i-\bar{y})\right)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}.$$

Egenskaper: Korrelationskoefficient mellan X-data och Y-data ligger alltid mellan –1 och 1.

$$-1\le r_{XY}\le 1$$

Korrelationskoefficient visar hur starkt LINJÄRT samband är mellan X-data och Y-data. Om punkterna (x_k, y_k) ligger på linjen $y=ax+b$ då är $r_{XY}=\begin{cases}1 \text{ om } a>0\\-1 \text{ om } a<0\end{cases}$.

Om r_{XY} är nära 1 då ligger punkter (x_k, y_k) nära en rät linje med positiv lutning.

Om r_{XY} är nära –1 då ligger punkter (x_k, y_k) nära en rät linje med negativ lutning.

Anmärkning: Om det **inte finns ett linjärt samband** mellan X-data och Y-data, fortfarande kan det finnas ett annat (**ickelinjärt)** matematiskt samband mellan X och Y (till ex Y=X^2+3X+8, Y=7+5ln(X), Y=3+8eX, Y=5sinX+cosX, ….)

Uppgift 4.

a) Bestäm korrelationskoefficienten mellan X och Y-data:

X=[44, 36, 39, 48, 50, 20, 15]

Y=[100, 79, 80, 101, 112, 55, 41]

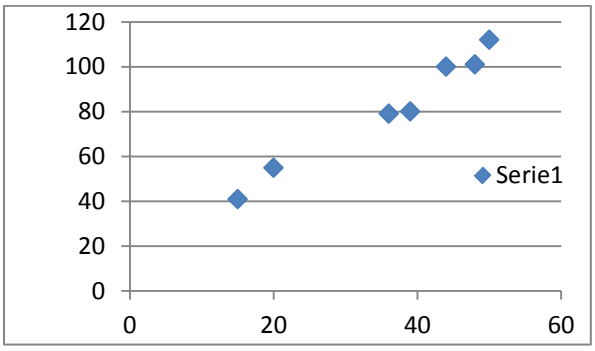
b) Finns det ett linjärt samband mellan X och Y.

c) Rita alla punkter (x_k, y_k) i ett koordinatsystem.

Svar: a) $r_{XY} = 0.9862$

b) korrelationskoefficienten är nära 1 som visar starkt linjär samband mellan X och Y.

c)



Anmärkning: Uttrycket $r_{XY} = \dfrac{\sum_{i=1}^{n}\left((x_i-\bar{x})(y_i-\bar{y})\right)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ kan skrivas på flera ekvivalenta sätt,

t ex.

$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left((x_i-\bar{x})(y_i-\bar{y})\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left((x_i-\bar{x})(y_i-\bar{y})\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{c_{xy}}{\sigma_x\sigma_y},$$

där $c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\big((x_i - \bar{x})(y_i - \bar{y})\big)$ (uttrycket kallas för kovarians mellan X och Y).

$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ och $\sigma_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$.

Anmärkning: I några böcker definieras kovarians som $c_{xy} = \frac{1}{n}\sum_{i=1}^{n}\big((x_i - \bar{x})(y_i - \bar{y})\big)$.

Uppgift 5. (gammal tenta)

Korrelationskoefficient $r = \frac{\frac{1}{n-1}\sum(x_k - \bar{x})(y_k - \bar{y})}{\sigma_x \sigma_y}$, där σ_x och σ_y standardavvikelser för X och Y, används som ett mått på hur starkt är LINJÄRT samband mellan variablerna X och Y. Bevisa följande påstående om r: Om punkterna (x_i, y_i) ligger exakt på linjen $y=ax+b$ och $a > 0$ (resp. $a < 0$) då är $r=1$ (resp. $r = -1$).

Lösning:

Om punkterna (x_i, y_i) ligger exakt på linjen $y=ax+b$ då gäller $y_k = ax_k + b$.

Först beräknar vi $\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k = \frac{1}{n}\sum_{k=1}^{n}(ax_k + b) = a\frac{\sum_{k=1}^{n} x_k}{n} + \frac{nb}{n} = a\bar{x} + b$

och

$$\sigma_y = \sqrt{\frac{\sum(y_k - \bar{y})^2}{n-1}} = \sqrt{\frac{\sum[(ax_k + b) - (a\bar{x} + b)]^2}{n-1}} = \sqrt{\frac{\sum a^2[(x_k - \bar{x}]^2}{n-1}}$$

$$= |a| \sqrt{\frac{\sum[(x_k - \bar{x}]^2}{n-1}}$$

Nu förenklar vi

$$r = \frac{\frac{1}{n-1}\sum(x_k - \bar{x})(y_k - \bar{y})}{\sigma_x \sigma_y} = \frac{\frac{1}{n-1}\sum(x_k - \bar{x})(ax_k + b - (a\bar{x} + b))}{\sqrt{\frac{\sum[x_k - \bar{x}]^2}{n-1}}\,|a|\sqrt{\frac{\sum[x_k - \bar{x}]^2}{n-1}}}$$

$$= \frac{\frac{a}{n-1}\sum (x_k - \bar{x})(x_k - \bar{x})}{\sqrt{\frac{\sum [x_k - \bar{x}]^2}{n-1}}\,|a|\sqrt{\frac{\sum [x_k - \bar{x}]^2}{n-1}}} == \frac{\frac{a}{n-1}\sum (x_k - \bar{x})^2}{|a|\frac{\sum (x_k - \bar{x})^2}{n-1}}$$

$$= \frac{a}{|a|} = \begin{cases} 1 \text{ om } a > 0 \\ -1 \text{ om } a < 0 \\ (\text{ej def om } a = 0) \end{cases}, \quad \text{vilket skulle bevisas.}$$