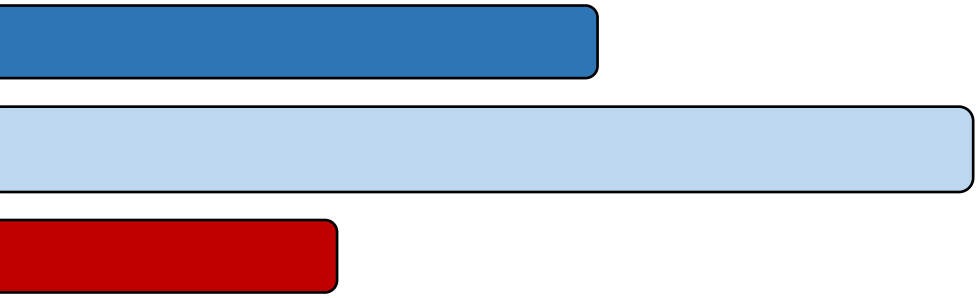
Statistik - kvantitativ metode

- Noter

$$\sum\sigma$$



Huspriser (1.000§)	Frekvens
330 til 411	4
411 til 492	11
492 til 573	12
573 til 654	3
654 til 735	6
	Total = 36

Nu får vi en mere valid distribution af frekvenserne for de enkelte variable.

Kumuleret frekvens

Det kan være, at man har behovet for at se, hvor mange observationer der falder under toppen af en given kategori, da vil kumuleret frekvens være mere fordelagtig at benytte:

Huspriser (1.000$)	Frekvens	Kumuleret frekvens
300 til 400	4	4
400 til 500	11	4+11 = 15
500 til 600	14	4+11+14 = 29
600 til 700	5	4+11+14+5 = 34
700 til 800	2	4+11+14+5+2 = 36
	Total = 36	

Kigger vi på den kumuleret frekvens i 2. række, kan vi se, at 15 af husene er solgt for mindre end 500.000$.

Relativ kumuleret frekvens

Ligesom ved eksemplet omkring sammenligningen med februar og marts, bruger man også den relative frekvens og herunder også den relative kumulative frekvens, hvis man vil sammenligne ovenstående huspriser i et andet distrikt. Det kan være, der er flere huse i det distrikt eller andet, hvorfor man skal sammenligne den relative frekvens for de to områder, hvis en valid sammenligning skal finde sted.

Eks. Taget fra kapitel 2 - vi beregner nu gennemsnit, varians og standardafvigelse

Huspriser (1.000$)	Frekvens
300 til 400	4
400 til 500	11
500 til 600	14
600 til 700	5
700 til 800	2
	Total = 36

Huspriser (1.000$)	f_i	m_i	$m_i f_i$	$(m_i - \overline{x})^2 f_i$
300 til 400	4	350	1.400	$(350 - 522)^2 \cdot 4 = 118.336$
400 til 500	11	450	4.950	$(450 - 522)^2 \cdot 11 = 57.024$
500 til 600	14	550	7.700	$(550 - 522)^2 \cdot 14 = 10.976$
600 til 700	5	650	3.250	$(650 - 522)^2 \cdot 5 = 81.920$
700 til 800	2	750	1.500	$(750 - 522)^2 \cdot 2 = 103.968$
Total	36		18.800	$SUM = 372.224$

$$\overline{x} = \frac{18.800}{36} = 522{,}22 = 522\ (afrundet)$$

$$s^2 = \frac{372.224}{36 - 1} = 10.635$$

$$s = \sqrt{10.635} = 103{,}13$$

Sandsynlighedsfunktion

Sandsynlighedsfunktionen er defineret som: $P(X = x) = P(x)$

Lille x er værdi for selvvalgt værdi, eksempelvis 2. Store X derimod er den stokastiske variabel.

En sådan funktion skal endvidere opfylde følgende:

1. $0 \leq P(x) \leq 1$
2. $\sum p(x) = 1$ (summen af alle udfald skal give 1)

Eks. Slag med en 6-sidet terning. I alt giver det 1.

1	2	3	4	5	6
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Forventet værdi

Dette er en af de mest væsentlige sandsynlighedskoncepter i statistik. Det bliver også kaldet for populationsgennemsnittet.

Det er den forventede værdi af den diskrete stokastiske variabel x, og denoteres $E(X)$ eller bare μ (Mu).

Det er endvidere essentielt at klarlægge, at den værdi med den største sandsynlighed for at finde sted, ikke skal forveksles med den forventede værdi for den stokastiske variabel.

Når vi har at gøre med en diskret stokastisk variabel, X, med værdierne $X_1, X_2, X_{3,}, X_4$ ...., hvilket sker når $P(X = x_i)$, da vil den forventede værdi skulle beregnes som:

$$E(X) = \mu = \sum x_i \, P(X = x)$$

Det er sådan set bare ligesom vægtet gennemsnit som jeg har gennemgået tidligere, men lad os prøve med et eksempel, hvor det handler om sandsynlighed:

Nedenfor har jeg lavet en tabel, der viser sandsynligheden for, hvor mange gange jeg i løbet af en uge drikker sodavand.

Opdateret med åbenbaring:

Alt giver god mening, da vi jo tager standardafvigelsen for populationen, men det er kvadratroden af n vi bruger. Altså observationer for stikprøven. Så vi bruger både noget fra populationen, men også noget for stikprøven, hvorfor det giver god mening, at vi opnår en difference mellem populationen og stikprøven (gennemsnittet) som vi kalder standard error.

Stikprøve fra en normal population

For en vilkårlig stikprøvestørrelse n, vil stikprøvefordelingen af $\overline{X}$ være normalt, hvis populationen X, hvor stikprøven er taget fra, er normalfordelt.

Hvis $\overline{X}$ er normalfordelt, da kan den tranformeres til en standard normal stokastisk variabel som:

$$Z = \frac{\overline{X} - E(\overline{X})}{se(\overline{X})} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Lad os bruge formlen:

Eks. gennemsnitsdiameteren på en pizza er 16 cm. og der er en standardafvigelse på 0,8 cm. Hvor stor er sandsynligheden så for, at 2 tilfældige udvalgte pizzaer er mindre end 15,5 cm.?

Løsning:

$$P(\overline{x} < 15{,}5) = P(Z < \frac{15{,}5 - 16}{\frac{0{,}8}{\sqrt{2}}} = P(Z < -0{,}88)$$

Nu skal vi finde den tilhørende z-værdi til -0,88. Den er 0,1894, omregnet til procent 18,94%.

Central limit theorem

Dette handler om, at ikke alle fordelinger er normalfordelt, og hvad gør vi så. Jo, så længe stikprøvestørrelsen n, er stor nok, ja så vil vi nok se en normalfordeling. En tommelfingerregel er, at man ikke kan retfærdiggøre noget er normalfordelt før $n \geq 30$.

Og ligesom før, hvis $\overline{X}$ er nogenlunde normalfordelt, da kan enhver værdi $\overline{x}$ blive omdannet til sin tilhørende værdi z, som er lig: $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$