

Elementær statistik

Lektion 1

Peter Tibert Stoltze
stat@peterstoltze.dk

1. februar 2010

Dagens program

- ▶ Praktiske forhold
- ▶ Kapitel 1: Statistiske grundbegreber
- ▶ Kapitel 2: Frekvensfordelinger
- ▶ Kapitel 3: Central tendens og spredning

Del I

Praktiske forhold

Oversigt

Lærebogen

Undervisningsplan

Undervisningsform og forberedelse

Om opgaveregning

Kommunikation

Lærebogen

Kapitel	Sider	Opgaver	Titel
1	5	0	Statistiske grundbegreber
2	6	1	Frekvensfordelinger
3	7	1	Centraltendens og spredning
4	8	3	Sandsynlighed og statistiske modeller
5	7	2	Estimation
6	4	0	Signifikanstestning
7	20	5	Forskelle mellem centraltendenser
8	7	4	Chi-i-anden prøven
9	6	1	Test for normalitet
10	7	1	Simpel korrelation
11	11	3	Lineær regression
12	12	2	Variansanalyse
13	8	2	Reliabilitet og enighed
Σ	108	25	

Undervisningsplan

Lektion	Uge	Dato	Tid	Kapitel	Opgaver	Kommentar
1	5	01.02.10	8-11	1,2,3	1,2	
2	6	08.02.10	8-11	3,4	3,4,5	
.	7	15.02.10	.	.	.	Vinterferie
3	8	22.02.10	8-11	4,5	6,7	
4	9	01.03.10	8-11	6,7	8,9,10	
5	10	08.03.10	8-11	7	11,12	
6	11	15.03.10	8-11	8,9	13,14,15	
7	12	22.03.10	8-11	8,9	16,17	
.	13	29.03.10	.	.	.	Påskeferie
.	14	05.04.10	.	.	.	2. påskedag
8	15	12.04.10	8-11	10,11	18,19	
9	16	19.04.10	8-11	10,11	20	
10	17	26.04.10	8-11	11,12	21	
11	18	03.05.10	8-11	12	22	
12	19	10.05.10	8-11	12	23	
13	20	17.05.10	8-11	13	24,25	

Undervisningsform og forberedelse

- ▶ Undervisning
 - ▶ gennemgang af teoristof (hele lærebogen)
 - ▶ regning af opgaver
 - ▶ studerende der fremlægger opgave
- ▶ Forberedelse
 - ▶ læse relevante kapitler
 - ▶ ideelt set forsøge at løse opgaver

Om opgaveregning

- ▶ Forståelsen af stoffet bliver afprøvet
- ▶ Det abstrakte bliver konkret (for det meste)
- ▶ Bedst med blyant, papir og lommeregner
- ▶ Dernæst kan vi se på fx. Excel eller SAS JMP

Kommunikation

- ▶ Jeg kan kontaktes via mail: stat@peterstoltze.dk
- ▶ Materiale til undervisningen samles (indtil videre) på <http://peterstoltze.dk/stat>

Del II

Kapitel 1: Statistiske grundbegreber

Overzicht

Indledning

Population versus stikprøve

Variabeltyper og måleskalaer

Parametrisk versus ikke-parametrisk statistik

Indledning

- ▶ Statistik bearbejder data systematisk
- ▶ Statistik giver reproducérbare resultater
- ▶ Statistik giver mulighed for at generalisere
- ▶ Statistik benyttes i videnskabelige artikler

Eksempel på indsamlede data

Tabel 1.1

Opdigtede læsescores for 30 piger og 30 drenge i tredje klasse

Køn	Score
drenge	55 46 32 54 60 77 49 43 64 32
	24 44 59 79 89 19 21 48 57 52
	35 68 55 88 69 38 49 79 24 48
piger	22 63 40 78 64 55 45 88 33 53
	54 69 74 62 51 58 71 66 68 41
	67 51 38 42 63 71 84 95 55 71

Læsescores (fortsat)

- ▶ Svært at få overblik og drage nogle (fornuftige) konklusioner
- ▶ Kan gruppere data i en tabel
→ Frekvensfordeling (kapitel 2)
- ▶ Kan beskrive typisk værdi og variation
→ Central tendens og spredning (kapitel 3)
- ▶ Kan sammenligne stikprøvegennemsnit
→ t -test (kapitel 7)

Population versus stikprøve

- ▶ I de fleste undersøgelser vil vi generalisere til en *population* på baggrund af undersøgelse af en *stikprøve*
- ▶ Kræver typisk at stikprøvens elementer er udtaget *simpelt tilfældigt*
- ▶ En population kan være endelig (*finit*) eller uendelig (*infini*)

Variabeltyper

- ▶ Afhængige og uafhængige variable
- ▶ Diskrete og kontinuerte variable

Måleskalaer

- Nominalskala** Måler tilhørsforhold til ikke-ordnede kategorier (fx. køn og ordklasse)
- Ordinalskala** Ordnet skala (fx. karakterer) hvor forskellen mellem 6 og 7 ikke kan sammelignes med forskellen mellem 7 og 8, eller hvor 10 ikke kan siges at repræsentere det dobbelte af 5
- Intervalskala** Ordnet og med sammenlignelige forskelle (fx. tidsskala eller temperatur målt i celcius), så spring af samme længde repræsenterer samme ændring
- Ratioskala** Har giver også forholdstal mening (længdemål eller temperatur målt i ° kelvin), så 10 er det dobbelte af 5

Parametrisk versus ikke-parametrisk statistik

- Parametrisk** Parametriske metoder er de klassiske metoder, der baseres på test af hypoteser om parametre i statistiske modeller — bygger ofte på antagelser om fordelinger
- Non-parametrisk** Er antagelserne ikke rimelige kan man ofte finde et non-parametrisk alternativ — disse er dog ofte mindre følsomme

Del III

Kapitel 2: Frekvensfordelinger

Oversigt

Indledning

Grafik af frekvensfordelinger

Frekvensfordeling med Excel

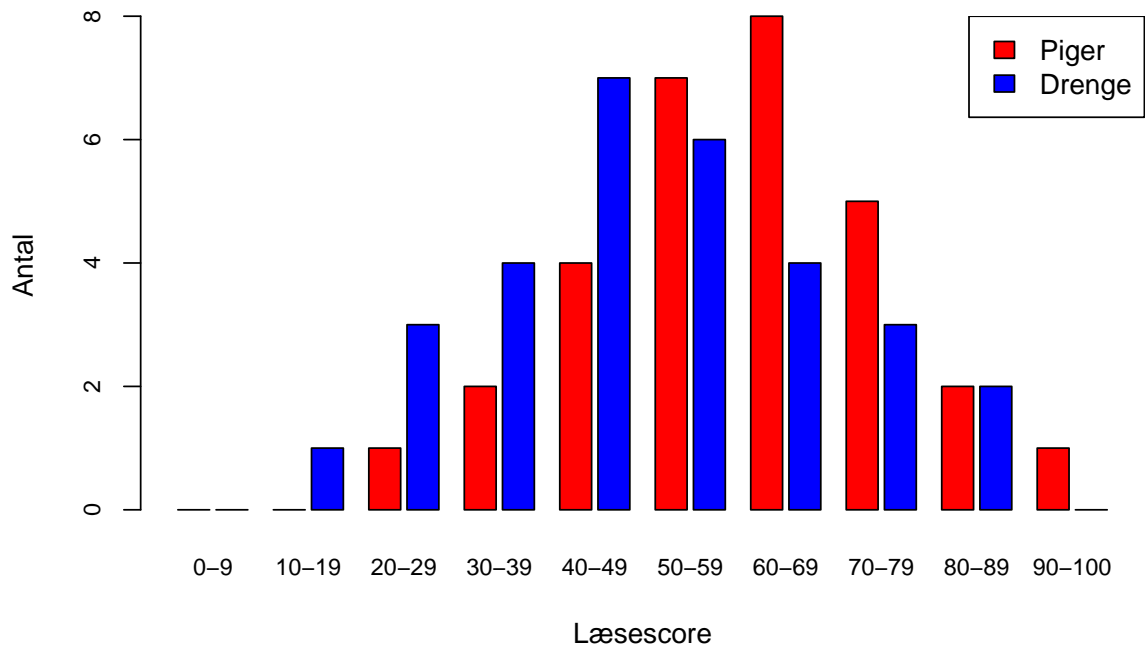
Frekvensfordeling

- ▶ Optælling af antal observationer i en række passende intervaller
- ▶ Ikke for snævre og ikke for brede. . .
- ▶ Mest almindeligt med samme bredde for alle intervaller (evt fraset de to yderste)

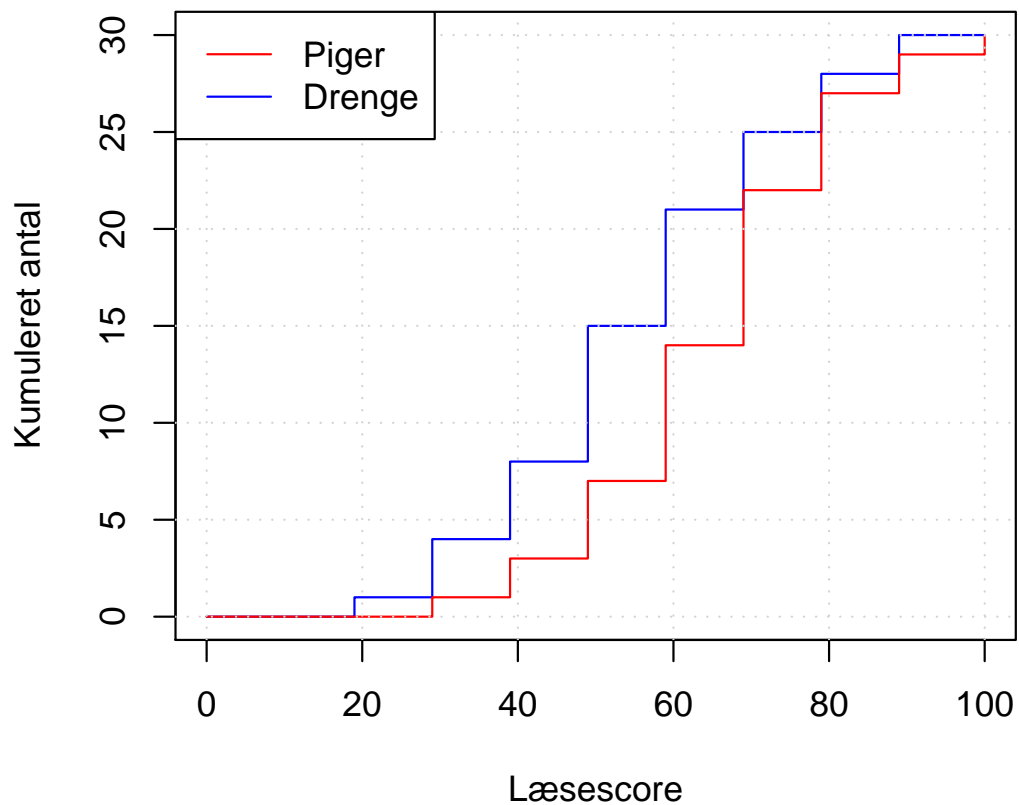
Læsescores som tabel

Score	Antal		Σ Antal	
	Piger	Drenge	Piger	Drenge
0-9	0	0	0	0
10-19	0	1	0	1
20-29	1	3	1	4
30-39	2	4	3	8
40-49	4	7	7	15
50-59	7	6	14	21
60-69	8	4	22	25
70-79	5	3	27	28
80-89	2	2	29	30
90-100	1	0	30	30
Σ	30	30	-	-

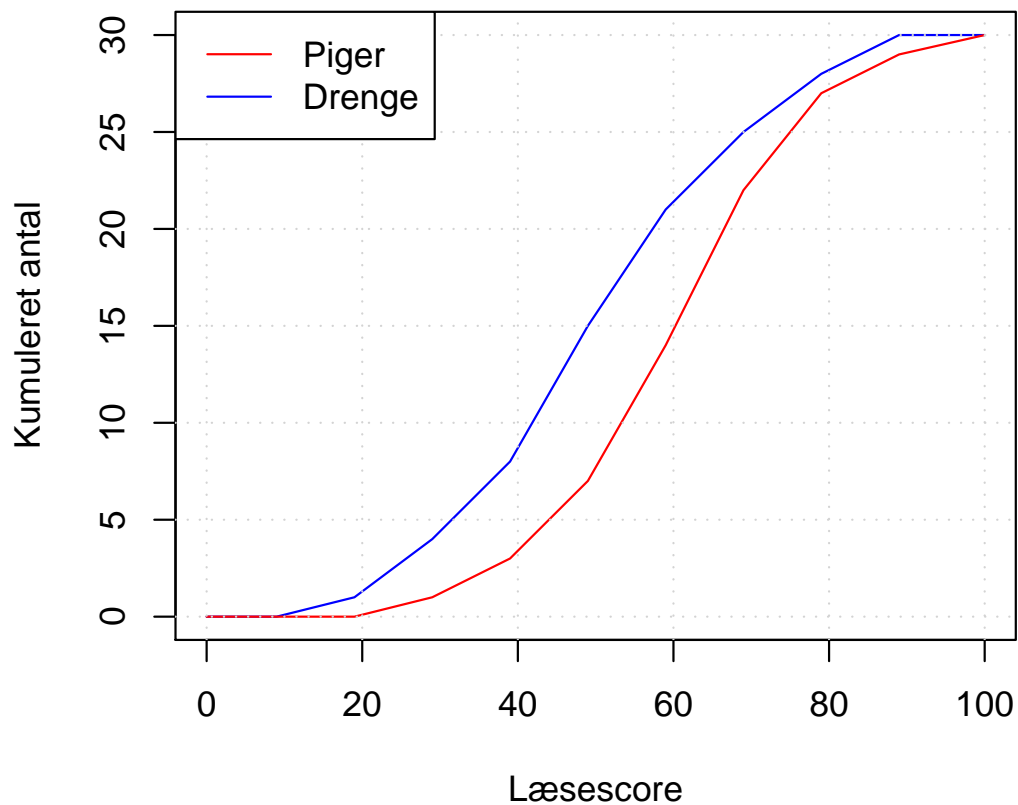
Læsescores som histogram



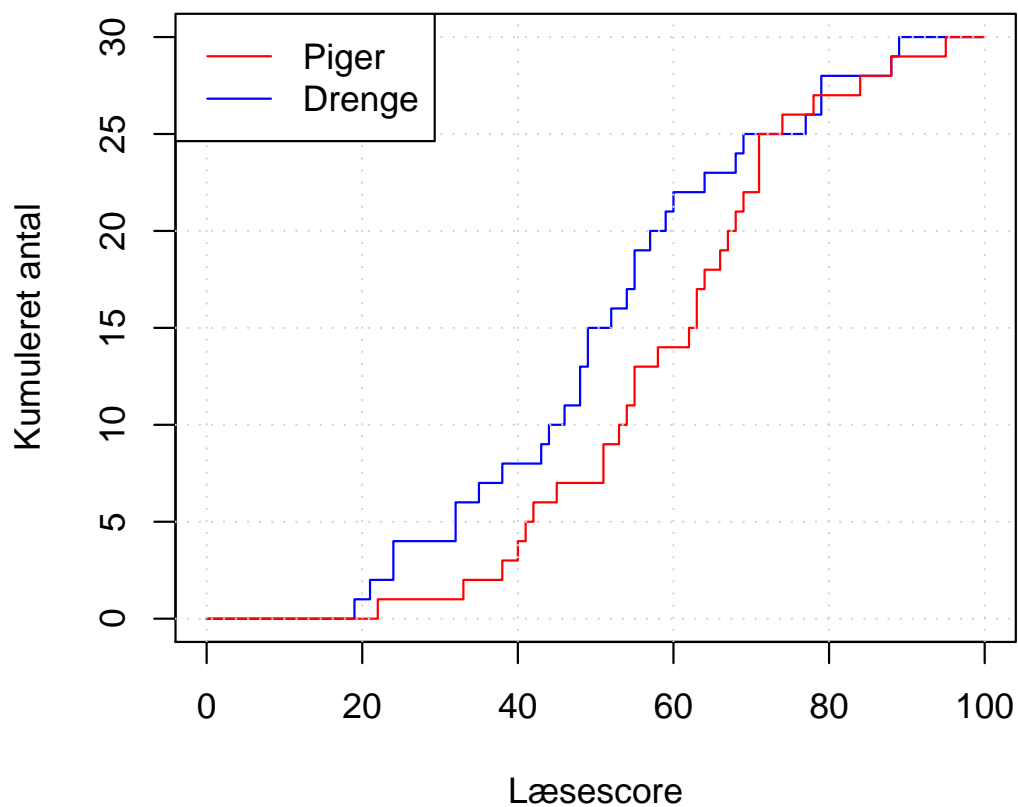
Kumulativ fordeling af læsescores (1/3)



Kumulativ fordeling af læsescores (2/3)

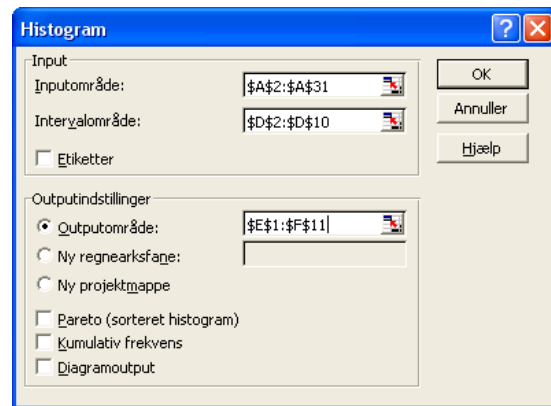
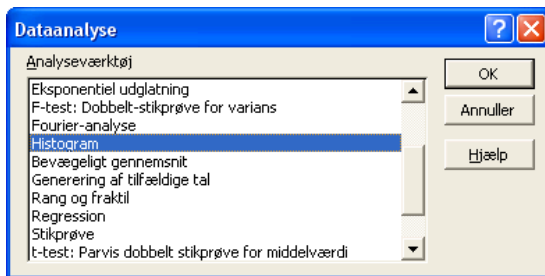


Kumulativ fordeling af læsescores (3/3)



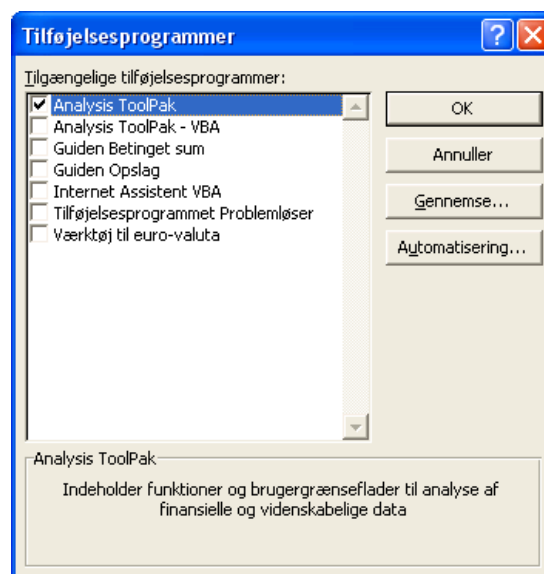
Frekvensfordeling med Excel

- ▶ Kan beregnes vha DataAnalyse | Histogram
- ▶ Skal specificere passende intervalafgrænsning
- ▶ Laves for et materiale ad gangen (drengene og piger hver for sig)



Analysis ToolPak

Før man kan lave optælling af fordelinger skal tilføjesprogrammet Analysis ToolPak (sic!) installeres:



Del IV

Kapitel 3: Central tendens og spredning

Oversigt

Indledning

Central tendens

Spredning

Fraktiler

Opgaveregning

Indledning

- ▶ I kapitel 2 omsatte vi de rå data til en tabel, der bedre viste materialets fordeling
- ▶ Fordelingen illustrerede vi med forskellige former for grafik
- ▶ Nu vil vi gerne karakterisere fordelingerne kvantitativt gennem deres **beliggenhed** og **variation**

Centraltendens

- ▶ Typetal eller modus (eng: mode)
- ▶ Aritmetisk middelværdi eller stikprøvegennemsnit (eng: mean or sample mean)
- ▶ Median eller 50%-fraktil

Modus

- ▶ **Modus** eller typetallet er den hyppigst forekommende værdi
- ▶ Eneste anvendelige mål for data målt på nominalskala (m/k, ja/nej) men vel nok mest anvendt ifm. data målt på ordinalskala (karakterer, scores)
- ▶ Læsescores: 71 for pigerne mod 55 for drengene
- ▶ Beregnes i Excel med funktionen `hyppigst`

Aritmetisk middelværdi

- ▶ Den **aritmetiske middelværdi** er summen af observationerne i forhold til antallet af observationer

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

- ▶ Læsescores: 59,7 for pigerne mod 51,9 for drengene
- ▶ Kan beregnes i Excel med funktionen `midde1`

Median

- ▶ **Medianen** er den midterste observation (gennemsnittet af de to midterste hvis n er lige)
- ▶ Specialtilfælde af generelt fraktilbegreb (50%-fraktil)
- ▶ Læsescores: 62,5 for pigerne mod 50,5 for drengene
- ▶ Beregnes i Excel med funktionen `median`

Centraltendenser for læsescores

Score	Dreng	Piger
n	30	30
Modus	55	71
\bar{x} (gennemsnit)	51,9	59,7
Median	50,5	62,5

Spredning og varians

- ▶ **Spredningen** s er kvadratroden af **variansen** s^2
- ▶ Variansen s^2 er kvadratet på spredningen s
- ▶ Variansen beregnes efter følgende formel:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\text{SAK}_x}{n-1}$$

hvor

$$\text{SAK}_x = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$$

- ▶ Hvis der *ikke* er tale om en stikprøve *kan* man benytte n i stedet for $n-1$ i nævneren

Spredning og varians

- ▶ Spredning kaldes også for standardafvigelse (eng: standard deviation)
- ▶ I Excel beregnes spredning med `stdafv` og varians med `varians`
- ▶ Spredning på læsescores: 16,6 for pigerne mod 19,2 for drengene

Fraktiler

- ▶ Om $P\%$ -fraktilen gælder, at P procent af observationerne er mindre end eller lig denne værdi
- ▶ Medianen er 50%-fraktil
- ▶ Andre navne for bestemte fraktiler er
 - ▶ Kvartiler (25, 50, 75)
 - ▶ Deciler (10, 20, ..., 90)
 - ▶ Percentiler (1, 2, ..., 99)
- ▶ Specielt er 25% fraktilen den **nedre kvartil** og 75% fraktilen den **øvre kvartil**
- ▶ Forskellen mellem øvre og nedre kvartil kaldes for **interkvartilafstanden** (IQR)

Beregning af fraktiler for grupperede data

$$P\% = L + \frac{k(\frac{Pn}{100} - F)}{f}$$

hvor

- ▶ P er den ønskede fraktil
- ▶ L nedre grænse i klassen, hvor den ønskede fraktil befinder sig
- ▶ k er klassebredden
- ▶ n er antal observationer
- ▶ F er antal observationer op til nedre grænse i den klasse, hvor fraktilen befinder sig
- ▶ f er antal observationer i den klasse, hvor fraktilen befinder sig

Eksempel: Beregning af 25% fraktil

- ▶ Følgende tabel er uddrag af kumulativ fordeling for 30 observationer:

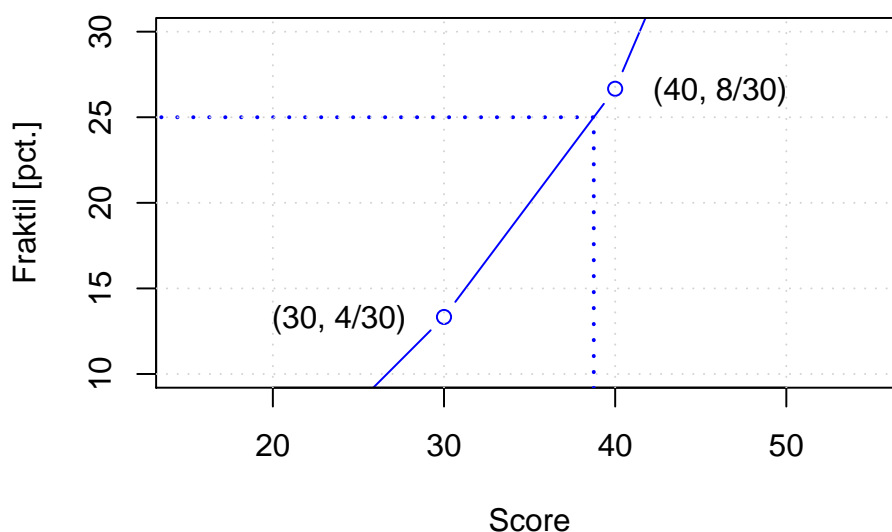
Nedre grænse	Øvre grænse	Obs	\sum Obs
0	10	0	0
10	20	0	0
20	30	4	4
30	40	4	8
.	.	.	.

- ▶ 30 er 13.3% fraktil og 40 er 26.6% fraktil, men vi vil gerne kende 25% fraktilen:

$$L + \frac{k\left(\frac{Pn}{100} - F\right)}{f} = 30 + \frac{10 \cdot \left(\frac{25 \cdot 30}{100} - 4\right)}{4} = 38.75$$

Alternativ beregning

Kan gøres efter et geometrisk princip:



$$\frac{40 - 30}{8/30 - 4/30} = \frac{x - 30}{7,5/30 - 4/30} \Rightarrow \frac{10}{4} = \frac{x - 30}{3,5} \Rightarrow x = \frac{35}{4} + 30 = 38,75$$

Beregning med Excel

- ▶ Beregnes i Excel med funktionen `fraktil`
- ▶ Der benyttes her en lidt anden definition end den her anvendte, men resultaterne minder en del om hinanden (specielt for store n)
- ▶ Beregning med Excel af de tre kvartiler samt interkvartilafstand (IQR) og spredning (s) for læsescores:

Fraktil	Drenge	Piger
25% fraktil	39,3	55,0
50% fraktil	49,0	62,7
75% fraktil	65,0	67,6
IQR	25,8	12,6
s	19,2	16,6

Opgaveregning

- ▶ Opgave 1 og 2
- ▶ Meget gerne frivillige til at gennemgå én opgave hver næste gang (cirka 5-10 minutter)