

Elementær statistik

Lektion 2

Peter Tibert Stoltze
stat@peterstoltze.dk

8. februar 2010

Dagens program

- ▶ Kapitel 2: Frekvensfordelinger
 - ▶ Kort opsamling
 - ▶ Opgave 1
- ▶ Kapitel 3: Central tendens og spredning
 - ▶ Gennemgang (fortsat)
 - ▶ Opgave 2
- ▶ Kapitel 4: Sandsynlighed og statistiske modeller
 - ▶ Gennemgang
 - ▶ Opgave 3, 4 og 5

Del I

Kapitel 2: Frekvensfordelinger

Opsamling

- ▶ Opsummering af større datamængder i frekvenstabeller ved optælling af antal observationer i en række skarpt definerede klasser, der dækker hele variationsområdet
- ▶ Valg af passende optællingsklasser kan kræve nogle eksperimenter. . .
- ▶ Den kumulative frekvens er antal observationer lavere end eller lig klassens øvre grænse
- ▶ Den kumulative fordeling kan udtrykkes i procent hvis fordelinger med forskelligt antal observationer skal sammenlignes
- ▶ Et **histogram** er frekvensfordelingen repræsenteret grafisk som søjler
- ▶ Den kumulative frekvensfordeling repræsenteres oftest som en kurve

Opgave 1

Tavlegennemgang...

Del II

Kapitel 3: Central tendens og spredning

Oversigt

Centraltendens

Spredning

Praktisk beregning

Fraktiler

Opgave 2

Centraltendens — opsamling

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

Eksempel

Centraltendenser for opdigtede læsescores for 30 piger og 30 drenge i tredje klasse

Score	Drenge	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, men det er absolut undtagelsen!

$n - 1$ giver centrale estimater

						$\frac{\text{SAK}}{n}$	$\frac{\text{SAK}}{n-1}$
Population	7	9	13	14	27	48,80	61,00
Stikprøve 1	7	9	13	14		8,19	10,92
Stikprøve 2	7	9	13		27	61,00	81,33
Stikprøve 3	7	9		14	27	60,69	80,92
Stikprøve 4	7		13	14	27	53,19	70,92
Stikprøve 5		9	13	14	27	45,69	60,92
Gennemsnit						45,75	61,00

Spredning og varians

- ▶ Spredning kaldes også for **standardafvigelse** (eng: standard deviation)
- ▶ Må ikke forveksles med **standardfejl** (eng: standard error), der er spredningen på middelværdien:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- ▶ I Excel beregnes spredning med `stdafv` og varians med `varians`
- ▶ Spredning på læsescores: 16,6 for pigerne mod 19,2 for drengene

Først lidt nomenklatur

- ▶ Vi regner på en simpelt tilfældigt udtaget **stikprøve** fra en **population**.
- ▶ Vi antager at observationerne er målt på en intervalskala eller en ratioskala.
- ▶ Gennemsnittet i stikprøven \bar{x} er et **estimat** for middelværdien μ i populationen:

$$\hat{\mu} = \bar{x}$$

- ▶ Spredningen i stikprøven s er estimat for spredningen σ i populationen:

$$\hat{\sigma} = s$$

- ▶ Spredningen er målt på samme skala som observationerne — det er variansen ikke.

Praktisk beregning af middelværdi og spredning

- ▶ Data (stikprøven) præsenteres på en liste på følgende måde

$$x = \{9; 2; 8; 11; 16; 16; 6; 5; 4; 6\}$$

- ▶ Start med at bestemme antallet af observationer

$$n = 10$$

- ▶ og dernæst summen af observationerne (indeks på sumtegnene er udeladt, da der summeres over alle observationer)

$$\sum x = 9 + 2 + 8 + 11 + 16 + 16 + 6 + 5 + 4 + 6 = 83$$

- ▶ Nu kan du bestemme gennemsnittet i stikprøven

$$\bar{x} = \frac{\sum x}{n} = \frac{83}{10} = 8,3$$

Praktisk beregning af middelværdi og spredning

- ▶ Så beregner du summen af de kvadrerede observationer (kvadratsummen)

$$\sum x^2 = 9^2 + 2^2 + 8^2 + 11^2 + 16^2 + 16^2 + 6^2 + 5^2 + 4^2 + 6^2 = 895$$

- ▶ så du kan beregne summen af de kvadrerede afvigelser fra gennemsnittet (summen af afvigelsesernes kvadrater)

$$SAK_x = \sum (x_i - \bar{x})^2 = \sum x^2 - \frac{1}{n}(\sum x)^2 = 895 - \frac{83^2}{10} = 206,1$$

- ▶ Så kan du beregne variansen for x 'erne i stikprøven

$$\hat{\text{var}}(x) = \frac{SAK_x}{n-1} = \frac{206,1}{10-1} = 22,9$$

- ▶ og endelig spredningen (stikprøvestandardafvigelsen)

$$s = \sqrt{\hat{\text{var}}(x)} = \sqrt{22,9} = 4,8$$

Opgave

- ▶ Ekstraopgave med praktisk beregning af middelværdi og spredning

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)

Direkte beregning af fraktiler

- ▶ Lad en stikprøve med n elementer være opstillet i rækkefølge, således at x_1 er den mindste observation og x_n er den største
- ▶ Da er den i 'te observation P -fraktilen i stikprøven, hvor

$$P = \frac{i - 0,5}{n}$$

- ▶ For store n er således $P \approx 0$ for $i = 1$ og $P \approx 1$ for $i = n$
- ▶ Ønsker man at kende en bestemt fraktil, da kan man regne baglæns i ovenstående udtryk, hvor resultatet dog kun sjældent vil være heltalligt i og dermed en bestemt observation. Dette kan løses ved **lineær interpolation**, som gennemgås senere. . .

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
<i>IQR</i>	25,8	12,6
s	19,2	16,6

Beregning af fraktiler for grupperet 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

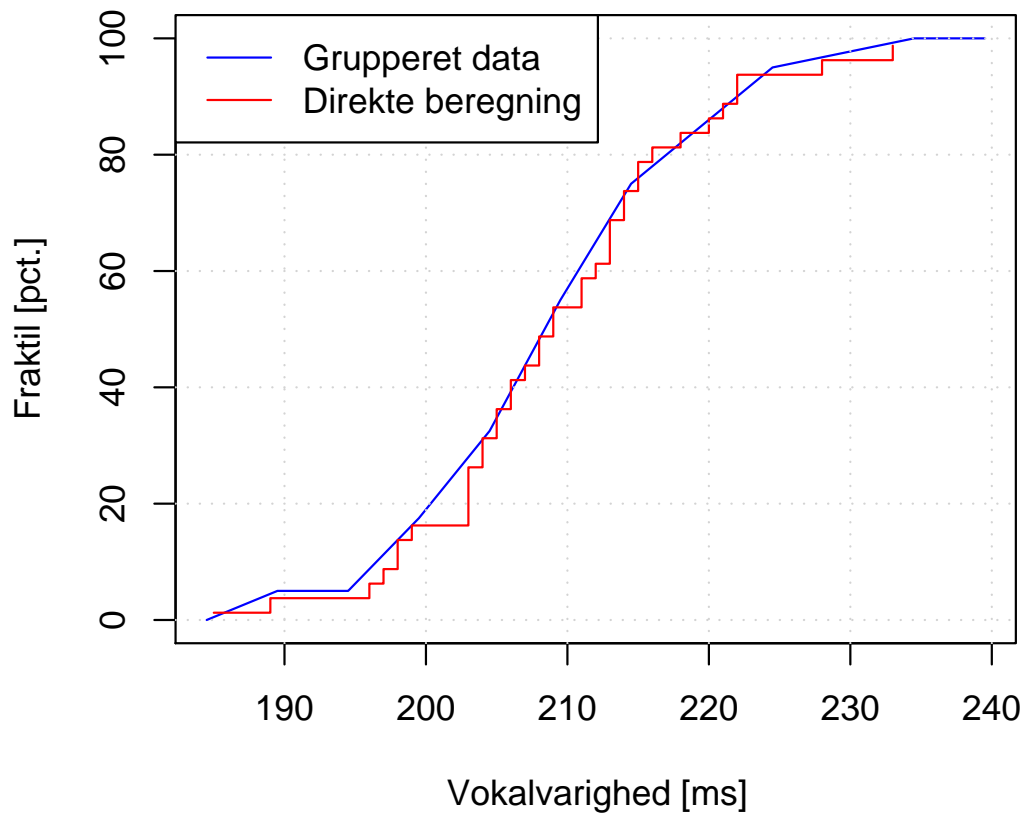
Eksempel: Vokalvarighed

Tabel 3.1

Frekvensfordeling for vokalvarighed i ms. Klassebredde er 5ms.

Nedre	Øvre	Frekvens	Kumulativ frekvens	Relativt
	184,5	0	0	0,0
184,5	189,5	2	2	5,0
189,5	194,5	0	2	5,0
194,5	199,5	5	7	17,5
199,5	204,5	6	13	32,5
204,5	209,5	9	22	55,0
209,5	214,5	8	30	75,0
214,5	219,5	4	34	85,0
219,5	224,5	4	38	95,0
224,5	229,5	1	39	97,5
229,5	234,5	1	40	100,0
234,5		0	40	100,0
		40		

Kumulativ fordeling af vokalvarighed



Eksempel på beregning

- ▶ Vi vil beregne 50% fraktilen (medianen) for datasættet med vokalvarighed:

$$Median = L + \frac{k(\frac{Pn}{100} - F)}{f} = 204,5 + \frac{5(\frac{50 \cdot 40}{100} - 13)}{9} = 208,39$$

- ▶ Vi kunne også lave interpolation i tabellen:

204,5	32,5
x^*	50,0
209,5	55,0

dvs. bestemme x^*

Lineær interpolation — generelt

- ▶ Vi kender punkterne (x_1, y_1) og (x_2, y_2) og ønsker at bestemme punktet (x^*, y^*) idet vi kender den ene af koordinaterne.
- ▶ Vi antager at der gælder

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y^* - y_1}{x^* - x_1}$$

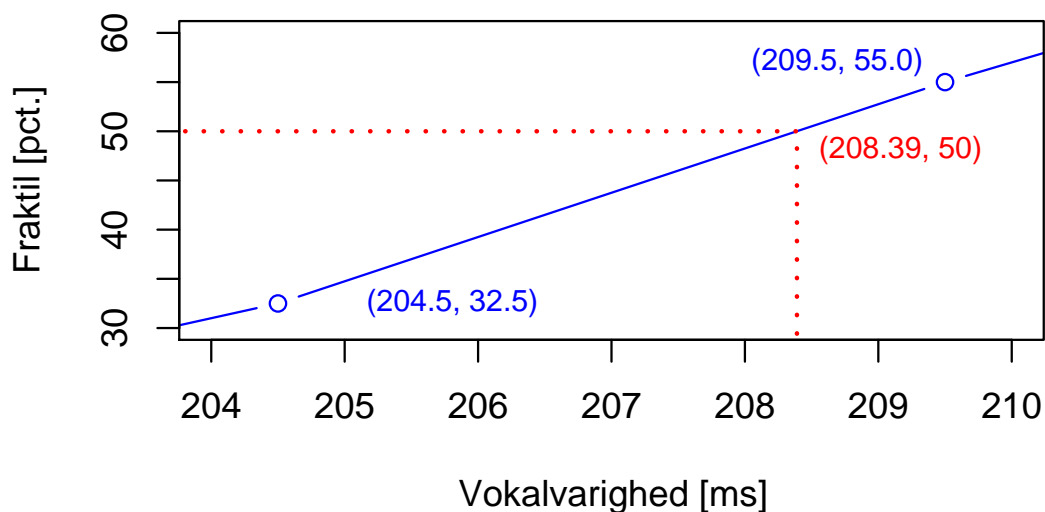
- ▶ Kender vi x^* kan vi bestemme y^* som

$$y^* = y_1 + \frac{y_2 - y_1}{x_2 - x_1} \cdot (x^* - x_1)$$

- ▶ Kender vi omvendt y^* kan vi bestemme x^* som

$$x^* = x_1 + \frac{x_2 - x_1}{y_2 - y_1} \cdot (y^* - y_1)$$

Lineær interpolation — grafisk



$$\begin{aligned} \frac{55,0 - 32,5}{209,5 - 204,5} &= \frac{50,0 - 32,5}{x^* - 204,5} \\ \Rightarrow x^* &= 204,5 + \frac{209,5 - 204,5}{55,0 - 32,5} \cdot (50,0 - 32,5) = 208,39 \end{aligned}$$

Opgave 2

Tavlegennemgang...

Del III

Kapitel 4: Sandsynlighed og statistiske modeller

Oversigt

Indledning

Sandsynlighed i binomialfordelingen

Normalfordelingen

Modelkontrol med normalfordelingen

Opgaver

Generalisering fra stikprøve til population

- ▶ Idé: Opstil en model for populationen og estimér modellens parametre på baggrund af stikprøven
- ▶ Kontrollér at stikprøven ikke er i modstrid med modellen
- ▶ Eksempel: 95% konfidensinterval for middelværdien i en normalfordeling

Binomialfordelingen - uformelt

- ▶ Lyttetest: En person har i tre ud af tre sætninger korrekt hørt forskel på **bas** og **pas** - kan det være tilfældigt?
- ▶ Vi gentager et eksperiment tre gange, hvor der hver gang er 50% sandsynlighed for at få succes ved en tilfældighed (fx. få krone)
- ▶ Hvad er sandsynligheden for at få krone tre gange i træk?
- ▶ Og hvorfor er det interessant?

Uformelt... *fortsat*

- ▶ Der er otte mulige udfald ved tre kast: KKK, KKP, KPK, PKK, KPP, PKP, PPK, PPP
- ▶ Alle otte udfald er lige sandsynlige og netop ét udfald svarer til tre gange krone
- ▶ Laplace's lov: Sandsynlighed er antal gunstige divideret med antal mulige
- ▶ Sandsynligheden for netop tre gange krone er således $1/8 = 0,125 = 12,5\%$

Uformelt. . . fortsat

- ▶ Tilbage til lyttetest: Der er altså en sandsynlighed (risiko) på 12,5% for, at personen ikke kan høre forskel på bas og pas selvom der blev svaret rigtigt i 3 ud af 3 tilfælde.
- ▶ Er dette acceptabelt og hvis ikke: Hvordan kan man så lave eksperimentet bedre?

Binomialfordelingen - formelt

- ▶ n Bernoulli-forsøg med sandsynligheden p for sandt (og følgelig sandsynligheden $1 - p$ for falsk)
- ▶ Punktsandsynligheder er givet ved

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x}, x = 0, 1, \dots, n$$

hvor $K(n,x)$ er binomialkoefficienten

$$\binom{n}{x} = \frac{n \cdot (n-1) \cdot \dots \cdot (n-x+1)}{x \cdot (x-1) \cdot \dots \cdot 1} = \frac{n!}{x!(n-x)!}$$

Opgave 3

Opgave 3

184
07.02.09

Sandsynligheden for at summe korrekt
 n gange i træk er

$$p = 0,5^n$$

Vi skal løse uligheden

$$p = 0,5^n < 0,05$$

Vi bruger regnereglen $\log(y^x) = x \cdot \log(y)$:

$$0,5^n < 0,05$$

$$\Downarrow n \cdot \log(0,5) < \log(0,05)$$

$$\Downarrow n > \frac{\log(0,05)}{\log(0,5)} = 4,32$$

Bemærk at ulighedstegnet bliver vendt
da $\log(0,5) < 0$

Vi skal altså mindst spørge fem
for at sandsynligheden bliver
mindre end 5 pct.

Er kravet 0,1 pct. finder vi

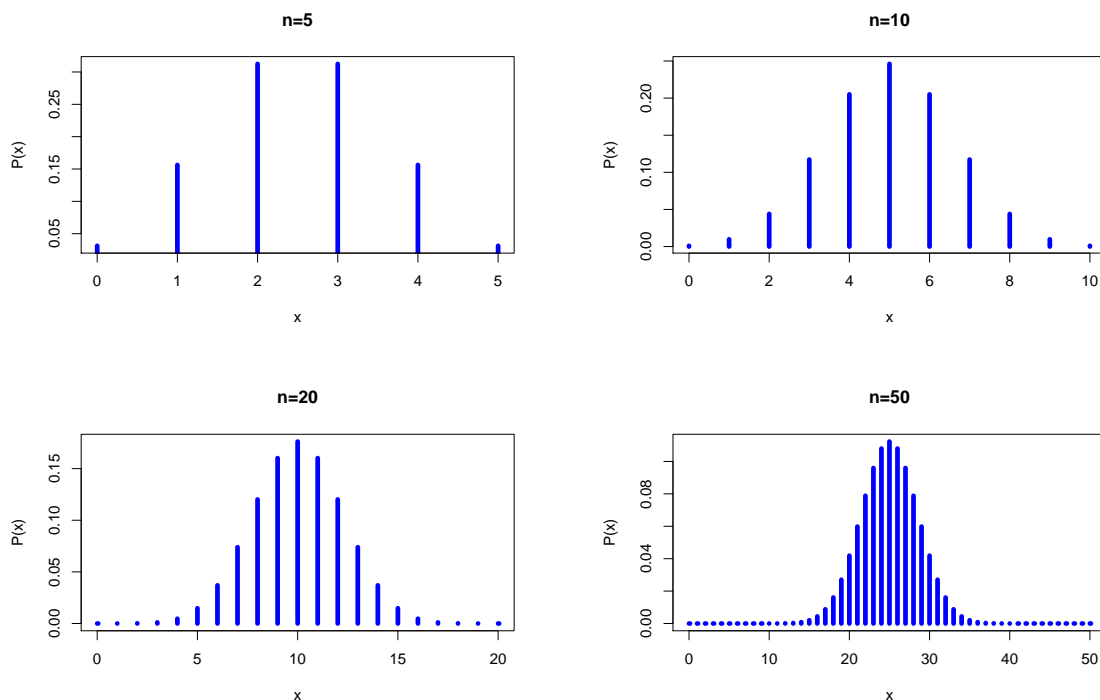
$$n > \frac{\log(0,001)}{\log(0,5)} = 9,97$$

Dvs. vi skal mindst spørge ti

Normalfordelingen

- ▶ Normalfordelingen er en **kontinuert** fordeling mens binomialfordelingen er en **diskret** fordeling
- ▶ **Histogrammet** for en binomialfordeling med $p = 0,5$ og meget højt n ligner **tæthedsfunktionen** for en normalfordeling

Normalfordelingen som grænsefordeling for binomialfordelingen med $p = 0,5$



Normalfordelingen

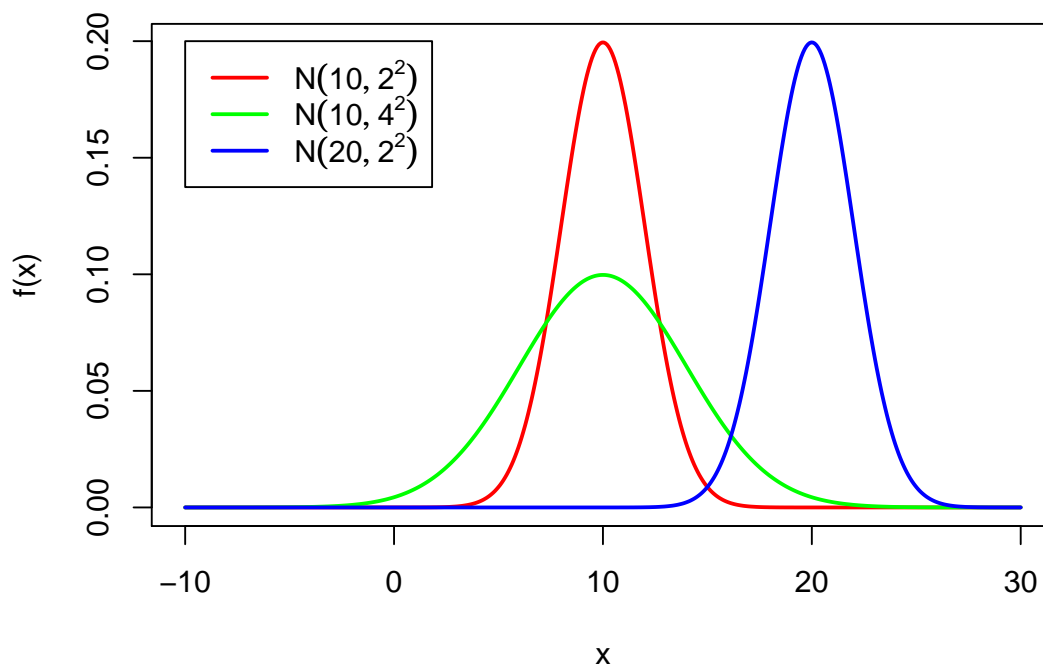
- ▶ Der findes uendeligt mange normalfordelinger, der hver især er karakteriseret ved deres **middelværdi** μ og deres **spredning** σ
- ▶ Middelværdi μ og spredning σ er parametre i normalfordelingen, og vi skriver $N(\mu, \sigma^2)$
- ▶ **Tæthedsfunktionen** er en klokkeformet kurve:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

(vi bruger heldigvis næsten altid tabeller)

- ▶ Kurven har toppunkt for $x = \mu$
- ▶ Større spredning giver fladere tæthedsfunktion

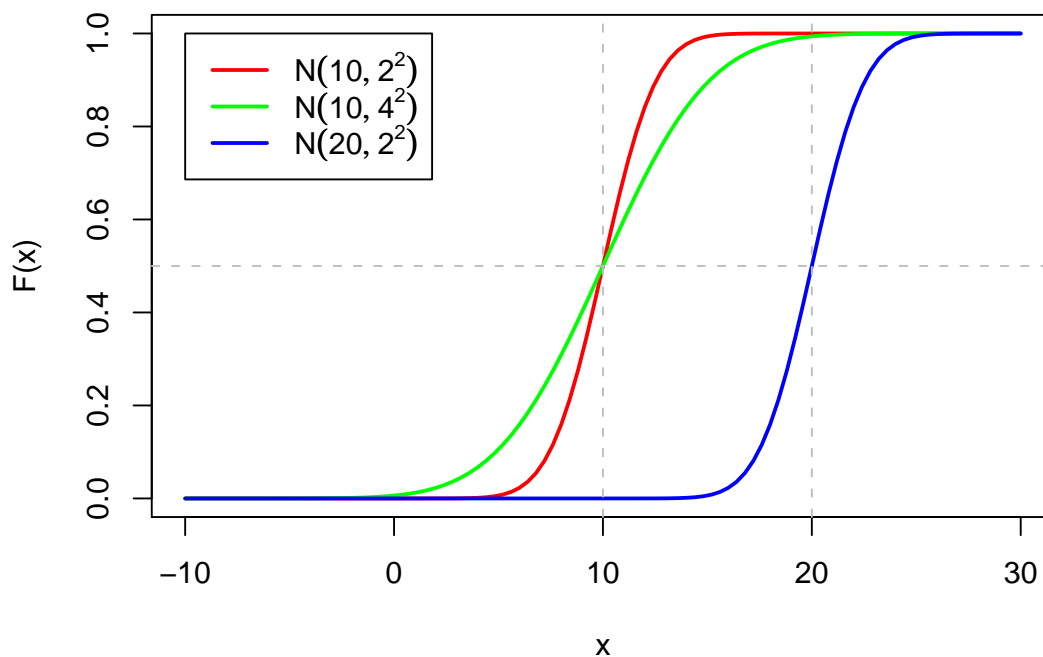
Tætheddsfunktion for 3 forskellige normalfordelinger



Fordelingsfunktion og standardnormalfordeling

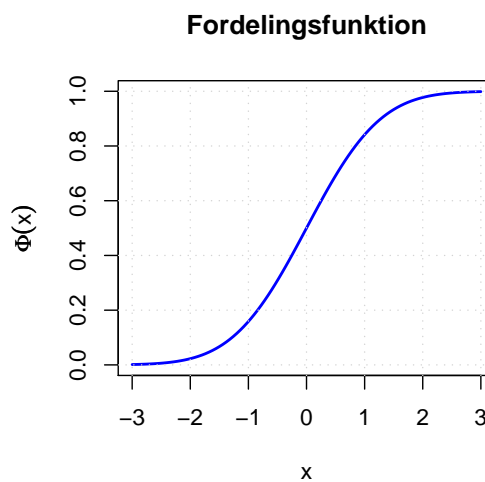
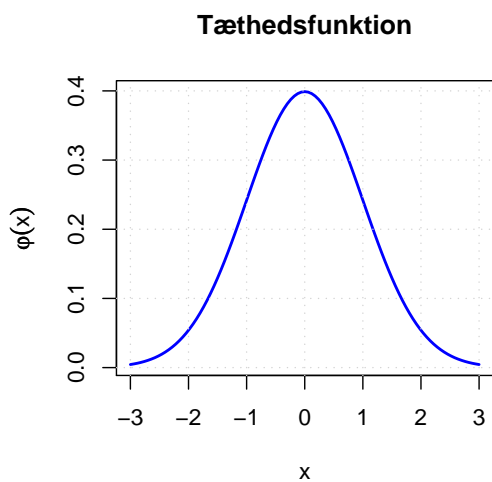
- ▶ Ved bestemt integration af tæthedsfunktionen kommer vi frem til fordelingsfunktionen, der er en slags kummuleret frekvensfordeling
- ▶ Fraktiler i en normalfordeling er nyttige ifm udsagn af typen:
 - ▶ 50% af eleverne kan forventes at score mellem 22 og 87 i den forelagte prøve
 - ▶ 5% af eleverne forventes at score mindre end 12
- ▶ Fordelingsfunktionen går gennem $(\mu, 0.5)$
- ▶ Lavere spredning giver stejlere fordelingsfunktion

Fordelingsfunktion for 3 forskellige normalfordelinger



Standardnormalfordelingen

- ▶ Der findes uendeligt mange normalfordelinger, men vi kan i praksis klare os med én, nemlig **standardnormalfordelingen** $N(0, 1)$
- ▶ Fordelingsfunktionen $\Phi(x)$ fremkommer ved integration af tæthedsfunktionen $\varphi(x)$



Eksempel på brug af Φ

- ▶ Antag at vi har lavet en undersøgelse, hvor gennemsnittet af scorene er 17 og standardafvigelsen er 3. Vi antager desuden, at scorene følger en normalfordeling.
- ▶ Vi vil nu gerne kende sandsynligheden for, at en tilfældig score er mindre end 14.
- ▶ Vi normaliserer ved at beregne den såkaldte z-værdi:

$$z = \frac{x - \bar{X}}{s} = \frac{14 - 17}{3} = -1$$

- ▶ Ved opslag i Tabel A kan vi nu se at

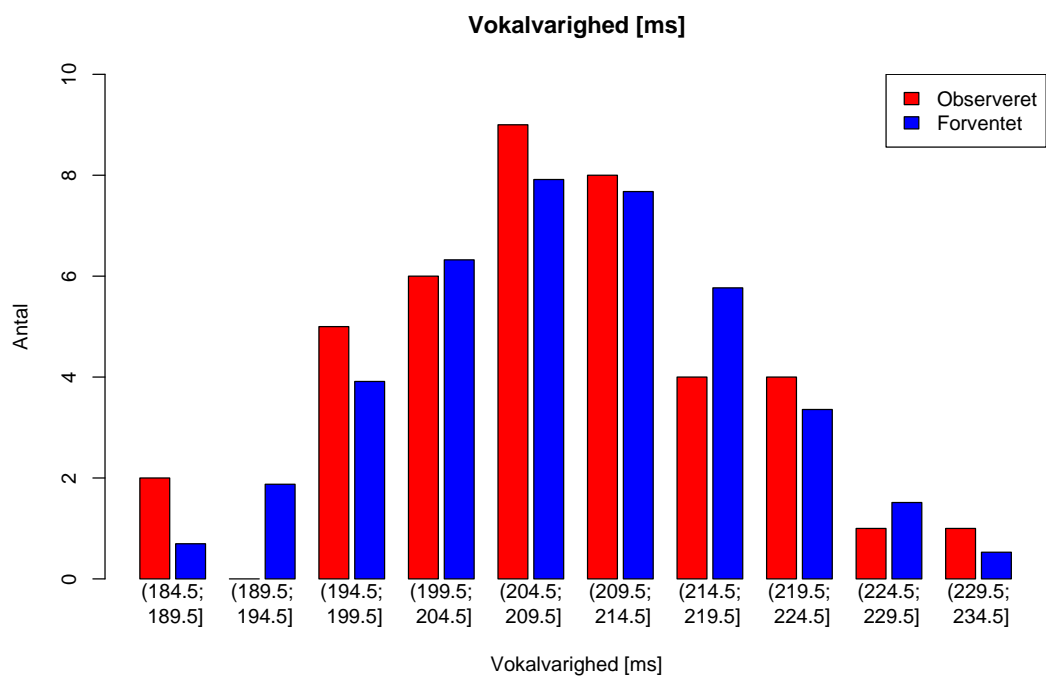
$$p = P(x \leq 14) = \Phi(-1) = 0,159$$

- ▶ Sandsynligheden for at en tilfældig score er mindre en 14 er altså cirka 16%

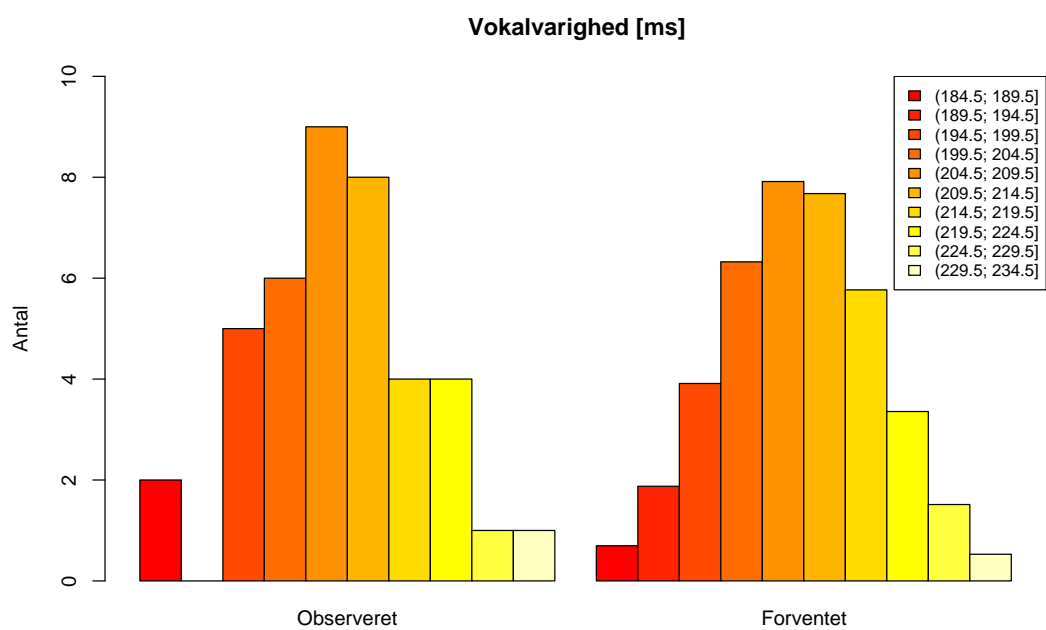
Modelkontrol

- ▶ Vi er ofte interesserede i at se, hvor godt vores stikprøve egentlig stemmer overens med normalfordelingsantagelsen
- ▶ For Tabel 2.5 (vokalvarighed i ms) beregner vi det forventede antal observationer i et bestemt interval under antagelsen om normalitet og sammenligner med det observerede
- ▶ Vi beregner $\bar{x} = 208,9$ og $s = 9,79$
- ▶ For klassen afgrænset ved $(204,5; 209,5]$ beregnes to z-værdier til -0,45 og 0,06
- ▶ Via Tabel A findes tilhørende sandsynligheder p som 0,326 og 0,524
- ▶ Sandsynligheden for at være i intervallet er derfor $0,524 - 0,326 = 0,198$
- ▶ Da stikprøven omfatter 40 enheder forventer vi at finde $40 \cdot 0,198 = 7,92$ enheder i intervallet
- ▶ Der var faktisk 9. . .

Grafisk modelkontrol



Grafisk modelkontrol



Opgaver

- ▶ Opgave 4
 - ▶ Regnes til næste gang
- ▶ Opgave 5
 - ▶ Gennemgås på tavlen nu