STATISTIK

x-klasserne

Gammel Hellerup Gymnasium

November 2023 ; Michael Szymanski ; mz@ghg.dk

Indholdsfortegnelse

INDLEDNING

Det er umuligt at komme med en fyldestgørende beskrivelse af, hvordan man inden for naturvidenskaberne er kommet frem til den enorme mængde viden, der til stadighed udbygges, justeres, glemmes og forkastes. Viden kan være opstået gennem gode ideer, tilfældigheder, opmærksomme iagttagelser af uventede hændelser, forkerte udregninger og fejl, der udligner hinanden, gruppers systematiske arbejde og enkeltpersoners vedholdende fokusering på problemer. Inden for videnskabsteorien forsøger man at give en slags ideal for, hvordan videnskab burde bedrives, og hvordan man undgår "forkert" viden.

Inden for naturvidenskaben anvendes den såkaldt *hypotetisk-deduktive* metode koblet sammen med falsifikationsprincippet. Det går kort fortalt ud på, at man har en hypotese (også kaldet en teori), som man udleder nogle konsekvenser af. Disse konsekvenser skal ifølge falsifikationsprincippet testes med henblik på at få forkastet hypotesen. Hvis dette mislykkes, er hypotesen ikke blevet forkastet, men derimod styrket.

Det er i testningen af konsekvenserne, at matematikken - eller mere præcist statistikken - kommer ind i billedet. Konsekvenserne kan f.eks. være en formel (ofte i fysik) eller en hypotese (ofte i biologi). Man taler om hypotesetest. Det er dog vigtigt at bemærke, at denne anvendelse af ordet 'hypotese' ikke svarer til anvendelsen af ordet i den hypotetisk-deduktive metode, men derimod til de konsekvenser, der udledes ved metoden. Hvis man bruger biologi-hypoteser og hypoteserne i hypotesetest som udgangspunktet i hypotetisk-deduktiv metode, vil man lede forgæves efter deduktionen.

Statistik går ud på at:

1) Indsamle datamateriale.
2) Organisere det indsamlede materiale, så det kan behandles.
3) Analysere og/eller teste det indsamlede materiale i forhold til en hypotese.
4) Vurdere analysen eller testresultatet.
5) Præsentere data og konklusioner på en overskuelig måde.

I dette hæfte behandles først punkterne 2) og 5), der er den såkaldte deskriptive statistik. Så gennemgås punkt 1). Og endelig punkterne 3) og 4).

Statistik: Metodisk indsamling, systematiseret opstilling og matematisk testning af hypotese på baggrund af talmateriale.

DESKRIPTIV STATISTIK

Deskriptiv statistik kaldes også for *beskrivende statistik*.

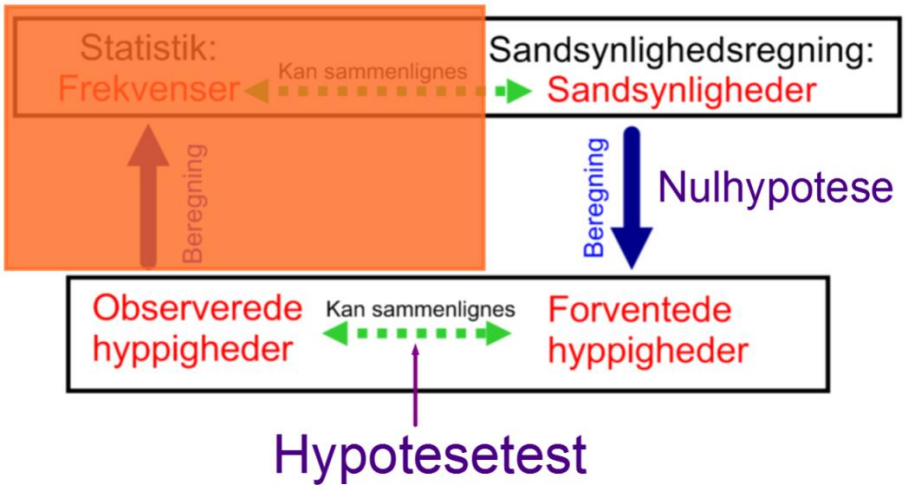
Bemærk, at det er **beskrivende** statistik. Deskriptiv statistik indeholder ingen analyse eller vurdering. Analyserne og vurderingerne kommer, når vi er færdige med at opstille vores data.

Vi har allerede nærmet os deskriptiv statistik i forbindelse med funktionsbegrebet, når vi har set på modeller, regression og residualer. I den del af den deskriptive statistik, som vi nu skal beskæftige os med, og som normalt er det, man forbinder med deskriptiv statistik, handler det om **at udregne nogle karakteristiske størrelser (deskriptorer), der fortæller noget centralt om et datamateriale, samt at få opstillet datamaterialet på en overskuelig måde, så man efterfølgende kan vurdere det**.

Deskriptorerne, vi skal se på, er: **Observationssættets størrelse, typetal, middelværdi, median, skævhed, mindste og største observation, fraktiler, kvartiler, varians og spredning**.

Når vi arbejder inden for deskriptiv statistik, anvender vi *frekvenser* i stedet for *sandsynligheder*, fordi vi indsamler noget datamateriale og foretager beregninger på dette uden inddragelse af nogen teori, hvorudfra vi kunne have ræsonneret os frem til nogle sandsynligheder. Frekvenserne beregnes ud fra de observerede hyppigheder.

Når vi senere skal se på hypotesetest, arbejder vi både med hyppigheder og sandsynligheder, da vi tager udgangspunkt i en såkaldt nulhypotese, der giver os nogle sandsynligheder, som vi kan anvende til at beregne forventede hyppigheder, der skal sammenlignes med observerede hyppigheder. Vores skema fra 'Sandsynlighedsregning og kombinatorik' vil derfor blive omformet til følgende:



DESKRIPTIV STATISTIK

Grupperede og ikke-grupperede observationssæt

Vi tager nu udgangspunkt i, at man har indsamlet noget datamateriale. Det kunne være:

- Man har målt højden af alle de unge mænd, der var til session i 2014.
- Man har registreret antallet af sygedage i 2018 for alle landets gymnasielærere.
- Man har registreret karaktererne til den skriftlige sommereksamen for matematik A-elever på stx i 2017.
- Man har målt den gennemsnitlige årlige nedbørsmængde i Danmark i perioden 1890-2014.
- Man har registreret antallet af biler for hver husstand i Danmark.

Hele pointen med deskriptiv statistik er som nævnt, at man skal have opstillet sine data på en så overskuelig måde som muligt, samt at man skal have beregnet nogle størrelser, der fortæller noget om datamaterialet.

I forbindelse med opstillingen af data, skal man som det første afgøre, om man skal gruppere sit datamateriale eller ej. Hvis man, som det f.eks. er tilfældet med højden af unge mænd på session, har foretaget en tilsvarende indsamling tidligere, eller hvis det, som f.eks. i tilfældet med karaktererne, giver sig selv fra start, kan afgørelsen være foretaget inden indsamlingen.

Men ellers skal man kigge på sit materiale og se, hvad der vil være mest hensigtsmæssigt med henblik på den mest overskuelige opstilling.

Bemærk, at dette er en væsentlig forskel fra hypotesetest, hvor det er en dødssynd først at vælge sit statistiske test efter at have set på datamaterialet.

Ikke-grupperede observationssæt:

Dette vælges, hvis den observerede størrelse er af en sådan art, at den kun kan antage et ikke for stort antal veldefinerede værdier. F.eks. hvis man observerer elevers karakterer. Her er mulighederne -3, 0, 2, 4, 7, 10 og 12.

Eller hvis man observerer antal hundehvalpe ved en fødsel med mulighederne 1, 2, 3, 4, … , 15 (hvor det "veldefinerede" maksimum kan sættes ud fra det højest observerede).

I vores eksempler fra før vil det være oplagt ikke at gruppere observationssættet i tilfældene med karakterer og antal biler.

Hvis man har et meget stort datasæt med mange forskellige værdier, kan man også undlade at gruppere, hvis man kun er interesseret i at udregne deskriptorer og opstille diagrammer, der alene er baseret på deskriptorer (f.eks. et såkaldt *boksplot*, som vi senere skal se på).

Grupperet observationssæt:

Her grupperes observationerne i passende intervaller. Antallet af intervaller må ikke være for småt, da man så mister for meget information om det indsamlede materiale. Det må heller ikke være for stort, hvis det går ud over overskueligheden. Selve intervalstørrelsen skal man også selv vælge. De fleste intervaller bør for overskuelighedens skyld være lige store, men sommetider kan man med fordel gøre intervallerne i enderne større, da man så undgår flere intervaller med få eller ingen observationer. Man kan evt. også gøre de centrale intervaller mindre, hvis man i dette område ønsker ikke at miste for meget information.

Det er vigtigt at bemærke, at man grupperer for overskuelighedens skyld, men at det sker på bekostning af noget tabt information.

Man har vedtaget, at intervallerne er lukkede mod højre og åbne mod venstre, dvs. f.eks.]5,10].

Skemaer

Observationssættets størrelse (n): ***

Observation (f.eks. **Karakter/Højde i cm**)	**-3 /]140;150]**	**0 /]150;155]**	**…**
Hyppighed / Intervalhyppighed (*h*)	**2**	**8**	**…**
Frekvens / Intervalfrekvens (%) (*f*)	**3%**	**12%**	**…**
Kumuleret frekvens / intervalfrekvens (%)	**3%**	**15%**	**100%**

Frekvenser udfyldes ved at regne på hyppighederne:

***Hyppighed*:** Antallet af den pågældende observation. Dvs. at ovenfor har 2 elever fået målt en højde mellem 141 cm og 150 cm (begge inklusive).

***Frekvens*:** Beregnes som: $\text{Frekvens} = \frac{\text{Hyppighed}}{\text{Observationssættets størrelse}}$ dvs. $f = \frac{h}{n}$

Man får så et decimaltal mellem 0 og 1, der **kan** angives i procent.

Kumulerede frekvenser udfyldes ved hjælp af frekvenserne:

***Kumuleret frekvens*:** Frekvenserne til og med den pågældende observation lægges sammen. Dvs. at ovenfor viser den kumulerede frekvens, at 15% af eleverne har højder på 155 cm eller mindre.

Kontrol: Den kumulerede frekvens skal ved sidste observation give 100% (evt. kan der være en decimal til forskel pga. afrundinger undervejs).

Eksempel 1a (Ikke-grupperet): Matematik A-niveau skriftlig eksamen 2018.

$n = 10657$

Karakter	-3	00	02	4	7	10	12
Hyppighed	83	1085	723	1927	3199	2566	1074
Frekvens	0,8%	10,2%	6,8%	18,1%	30,0%	24,1%	10,1%
Kumuleret frekvens	0,8%	11,0%	17,8%	35,9%	65,9%	90,0%	100,1%

Regneeksempler:

Frekvens for karakteren 4: $f = \frac{h}{n} = \frac{1927}{10657} = 0{,}1808201182 = 18{,}1\%$

Kumuleret frekvens for karakteren 4: $0{,}8\% + 10{,}2\% + 6{,}8\% + 18{,}1\% = 35{,}9\%$

Dvs. at 35,9% af eleverne fik karakteren 4 eller derunder.

Det bemærkes, at den kumulerede frekvens pga. afrundninger giver 100,1% i stedet for 100% for karakteren 12.

Eksempel 2a (Ikke-grupperet): Matematik B-niveau skriftlig eksamen 2018.

$n = 11135$

Karakter	-3	00	02	4	7	10	12
Hyppighed	405	2287	1145	2225	2786	1844	443

I Maple kan værdierne indtastes i en matrix. I dette tilfælde 7 rækker og 2 søjler:

$$EksamenB := \begin{bmatrix} -3 & 405 \\ 0 & 2287 \\ 2 & 1145 \\ 4 & 2225 \\ 7 & 2786 \\ 10 & 1844 \\ 12 & 443 \end{bmatrix}$$

Med Gym-pakkens *frekvensTabel* kan man derefter få udregnet skemaet:

$frekvensTabel(EksamenB)$

observation	hyppighed	frekvens(%)	kumuleret(%)
-3	405	3.637	3.64
0	2287	20.54	24.2
2	1145	10.28	34.5
4	2225	19.98	54.4
7	2786	25.02	79.5
10	1844	16.56	96
12	443	3.978	100

Eksempel 3a (Grupperet): Bestemmelse af højden af Det Skæve Tårn i Pisa.

24 elever er med samme metode kommet frem til følgende værdier angivet i meter ($n = 24$):

63,9	48,4	49,5	38,9	33,5	52,5	51,8	56,7	52,5
10,9	49,2	42,4	72,5	39,8	271,5	73,1	169,1	48,4
48,8	49,2	48,4	31,9	59,3	45,9			

Her skal man gruppere observationerne, hvis man ønsker at lave et skema. Hvis man ikke grupperede observationssættet, ville man få 20 forskellige observationer med hyppigheden 1 for alle observationer bortset fra 48,4 (3), 49,2 (2) og 52,5 (2). Et sådant skema ville ikke kunne bruges.

Man skal nu vælge nogle passende intervaller, og her skal man bl.a. bemærke de med gult fremhævede målinger. 10,9 er mindste observation, 271,5 er største observation og 169,1 er en måling, der afviger markant fra hovedparten af målingerne. Intervallerne skal (selvfølgelig) indeholde disse målinger, men der må heller ikke være for mange intervaller (så opstår samme problem, som hvis man ikke grupperede observationerne). Og hvis intervallerne bliver for store, fordi man skal have 271,5 med, mister man meget information i området 30-80, hvor 21 ud af 24 målinger ligger.

I dette tilfælde er man derfor nødt til at anvende intervaller med forskellig intervalbredde. Man kunne f.eks. vælge:

Højde	]10,20]	]20,30]	]30,40]	]40,50]	]50,60]	]60,70]	]70,80]	]80,100]	]100,200]	]200,300]
Intervalhyppighed	1	0	4	9	5	1	2	0	1	1
Intervalfrekvens	4,2%	0,0%	16,7%	37,5%	20,8%	4,2%	8,3%	0,0%	4,2%	4,2%
Kum. intervalfre.	4,2%	4,2%	20,9%	58,4%	79,2%	83,4%	91,7%	91,7%	95,9%	100,1%

Intervalfrekvenserne og de kumulerede intervalfrekvenser er beregnet på samme måde som frekvenser og kumulerede frekvenser (se Eksempel 1a).

Eksempel 4a (Grupperet): Bestemmelse af længden af bymuren i Lucca (angivet i meter). $n = 54$

4908	3826	3746	4143	4070	4095	3780	5037	3875
4091	4663	4321	4375	4443	3737	4235	4039	3945
4387	4175	4101	4326	4373	4254	4175	3090	3948
4877	4225	4163	4170	4396	4344	4171	5280	4164
4172	4977	4427	4614	4515	3981	4568	4225	4139
4397	4329	4301	4576	4269	4141	4076	3740	4220

Den mindste værdi er 3090, og den største er 5280. *Variationsbredden* (forskellen mellem max og min) er altså 2190. Man kan godt anvende ens intervalbredde i dette tilfælde, da målingerne er fordelt tilpas jævnt mellem max og min. Man kunne f.eks. vælge 12 intervaller med intervalbredden 200 eller 8 intervaller med intervalbredden 300. Her er valgt 8 intervaller:

Længde	]3000,3300]	]3300,3600]	]3600,3900]	]3900,4200]	]4200,4500]	]4500,4800]	]4800,5100]	]5100,5400]
Int. Hyp.	1	0	6	19	18	5	4	1

I Maple indtastes intervaller på den sædvanlige måde med "..":

$$Bymurlængde := \begin{bmatrix} 3000..3300 & 1 \\ 3300..3600 & 0 \\ 3600..3900 & 6 \\ 3900..4200 & 19 \\ 4200..4500 & 18 \\ 4500..4800 & 5 \\ 4800..5100 & 4 \\ 5100..5400 & 1 \end{bmatrix}$$

frekvensTabel(*Bymurlængde*)

observation	hyppighed	frekvens(%)	kumuleret(%)
3000 .. 3300	1	1.852	1.85
3300 .. 3600	0	0	1.85
3600 .. 3900	6	11.11	13
3900 .. 4200	19	35.19	48.1
4200 .. 4500	18	33.33	81.5
4500 .. 4800	5	9.259	90.7
4800 .. 5100	4	7.407	98.1
5100 .. 5400	1	1.852	100

Diagrammer

Skemaerne kan bruges til at tegne kurver eller diagrammer, der skal gøre det observerede talmateriale overskueligt for læseren. I alle tilfælde angives observationerne ud af 1. aksen.

Husk, at skalaen på 1. aksen skal være jævn (med mindre der er en god grund til f.eks. at gøre den logaritmisk eller andet). Så hvis f.eks. karakterne -3, 0, 2, 4, 7, 10 og 12 skal angives, skal afstanden mellem 4 og 7 være 1½ gang så stor som afstanden mellem 2 og 4.

Med grupperede observationssæt afsættes intervalendepunkterne. **Og husk endnu engang, at skalaen skal være jævn, dvs. store intervaller kommer til at fylde mere på aksen.**

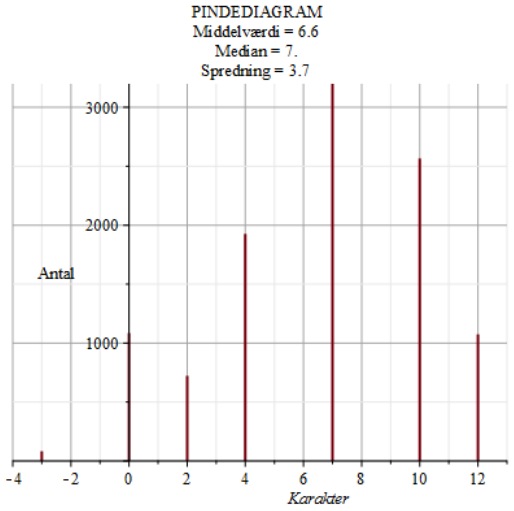
Man kan for både grupperede og ikke-grupperede observationssæt afbilde frekvenser og kumulerede frekvenser, og der er derfor 4 forskellige typer af diagrammer, der kan tegnes.

<table>
<tr><th></th><th>Ikke-grupperede observationssæt</th><th>Grupperede observationssæt</th></tr>
<tr><td>Frekvens</td><td>Pinde-, søjle- eller stolpediagram
Hyppigheden eller frekvensen angives op ad 2. aksen (ofte angivet i %).</td><td>Histogram
Ofte vælger man ikke at have en 2. akse. I stedet anvendes et rektangel (kvadrat eller aflang) til at angive, hvordan der omregnes fra areal til %. F.eks. 5%
Frekvensen for hvert interval omregnes til et areal, der afsættes som en søjle, hvor grundfladen i søjlen bestemmes af intervalbredden, og hvor søjlens højde skal afsættes, så arealet af søjlen kommer til at passe. Hvis to intervaller, hvor det ene er dobbelt så bredt som det andet, indeholder lige mange observationer, vil søjlehøjden i det bredeste altså blive halvt så stor som i det andet.
Undtagelse: Hvis alle intervaller er lige store, kan der godt laves en 2. akse med enheden %, hvor man så ikke længere skal angive et areal, men hvor søjlehøjden direkte angiver frekvensen. Denne enhed er dog strengt taget forkert.
Generel 2. akse: Man kan i alle tilfælde anvende en 2. akse, hvis man benytter sig af den rigtige enhed.
Hvis enheden på 1. aksen er kg, skal enheden på 2. aksen være $\frac{\%}{kg}$, og hvis enheden på 1. aksen er m^2, skal enheden på 2. aksen være $\frac{\%}{m^2}$. Så vil frekvensen kunne aflæses som arealet af søjlen.
Denne generelle metode gør det nemmere at sammenligne histogrammer med de klokkeformede normalfordelinger.</td></tr>
<tr><td>Kumuleret frekvens</td><td colspan="2">Trappediagram /Trappekurve
2. Aksen går fra 0% til 100%.
Den kumulerede frekvens for hver observation afsættes som et punkt, og fra hvert punkt tegnes linjestykker lodret ned og vandret mod højre, så der dannes en trappe.</td><td>Sumkurve
2. aksen går fra 0% til 100%.
Der sættes først et punkt på 1. aksen ved første intervals venstre endepunkt. Derefter afsættes den kumulerede frekvens for alle intervallerne som et punkt ved højre intervalendepunkt.
Punkterne forbindes med rette linjestykker.</td></tr>
</table>

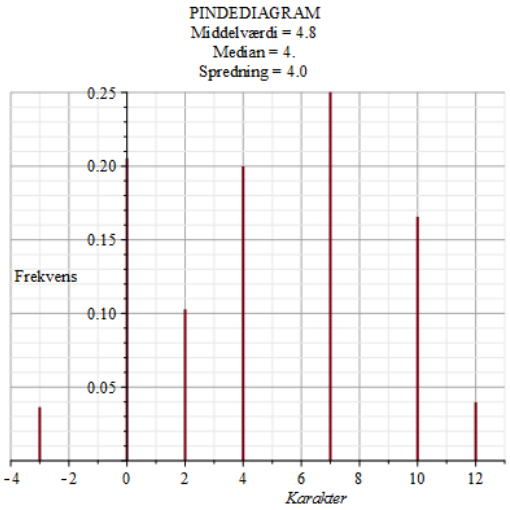
Eksempel 1b og 2b: Pindediagrammer og trappekurver

Når man har indtastet sine målinger i en matrix (se Eksempel 1a), kan Gym-pakkens *plotPindediagram* tegne pindediagrammer med enten antallet (tilføj *y_akse=antal*) eller frekvenser:

plotPindediagram(EksamenA, y_akse = antal)

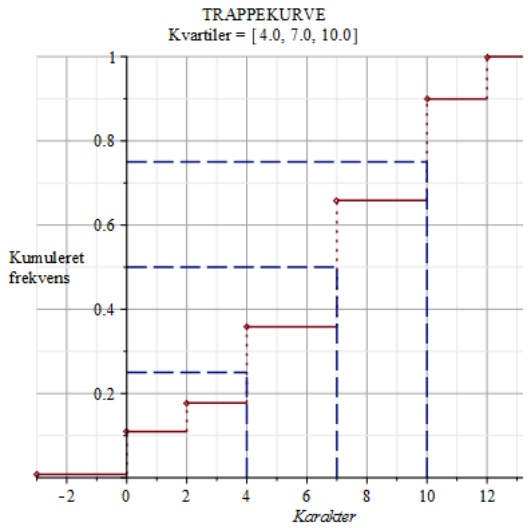


plotPindediagram(EksamenB)

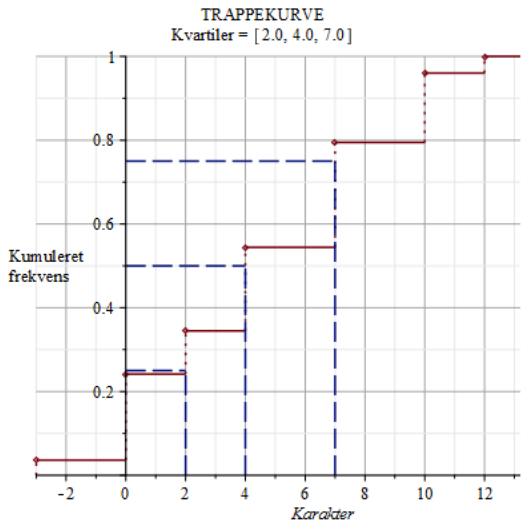


Med kommandoen *plotTrappekurve*, kan man få tegnet trappekurver:

plotTrappekurve(EksamenA)



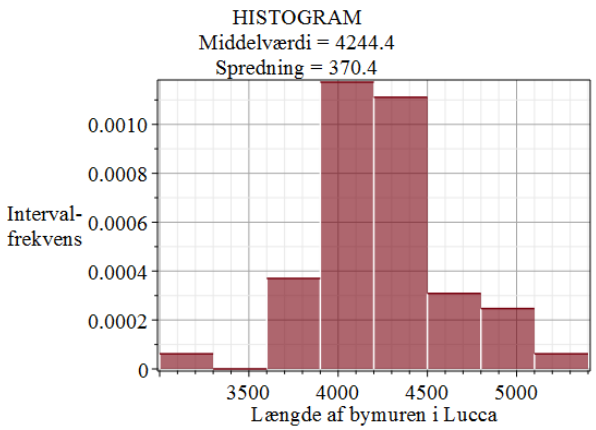
plotTrappekurve(EksamenB)



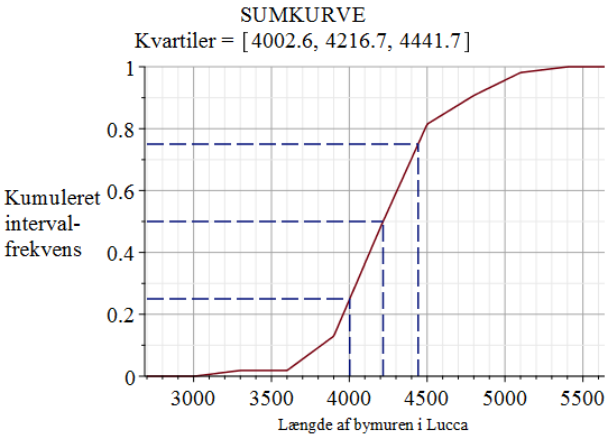
Eksempel 4b: Længden af bymuren i Lucca:

Med kommandoerne *plotHistogram* og *plotSumkurve* får man:

plotHistogram(Bymurlængde)



plotSumkurve(Bymurlængde)

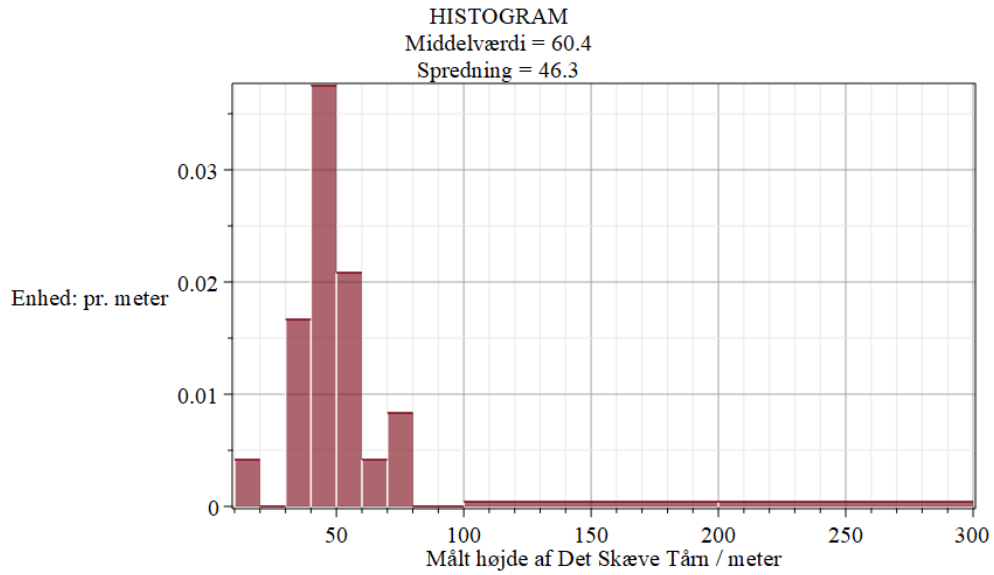


Udover at tegne diagrammerne udregner Gym-pakkens kommandoer også nogle deskriptorer (middelværdi, median, spredning, kvartilsæt), som vi skal se nærmere på i næste afsnit.

Bemærk, at man i Eksempel 4b kan afsætte intervalfrekvensen ud ad 2. aksen, fordi intervalbredderne er lige store. I nedenstående eksempel er intervalbredderne forskellige, og derfor ændres enheden på 2. aksen:

Eksempel 3b: Højden af Det Skæve Tårn i Pisa

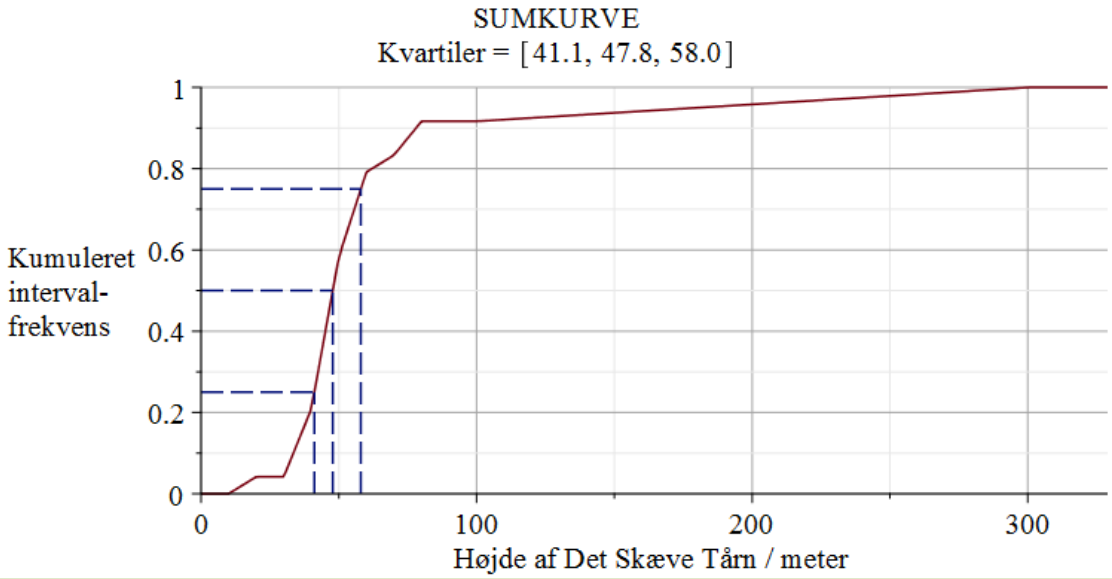
plotHistogram(*HøjdeSkæveTårn*)



Bemærk, at søjlerne for intervallerne]10,20] og]60,70] er meget højere end søjlerne for intervallerne]100,200] og]200,300], selvom de alle hver især indeholder 1 måling.

Dette skyldes som nævnt intervalbredderne, da målingen "fordeles" jævnt ud på hele intervallet.

plotSumkurve(*HøjdeSkæveTårn*)



Statistiske deskriptorer

Ud fra skemaet eller diagrammerne kan man aflæse eller udregne de statistiske deskriptorer (se næste side). Bemærk, at nogle af deskriptorerne allerede kendes fra sandsynlighedsregning:

Ikke-grupperede observationssæt	*Grupperede observationssæt*
Observationssættets størrelse (betegnes i nedenstående med *n*) Antallet af observationer. Hvis man f.eks. har målt længden af præriehundes fortænder, er antallet af observationer det antal præriehunde, man har målt på.	
Typetal Typetallet er observationen med den største hyppighed. Der kan godt være mere end ét typetal.	**Typeinterval** Observationsintervallet med den største hyppigheds**tæthed**. Det ses nemmest på histogrammet, hvor det er intervallet med den højeste søjle. Der kan godt være mere end ét typeinterval.
Middelværdi/Gennemsnit Angives som μ eller $\bar{x}$. Når x_i betegner den i'te observation og h_i hyppigheden og f_i frekvensen af denne, beregnes middelværdien ved: $\mu = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot h_i$ eller $\mu = \sum_{i=1}^{k} x_i \cdot f_i$	**Middelværdi/Gennemsnit (**Angives som μ eller $\bar{x}$). Når m_i betegner midtpunktet af det i'te observationsinterval og h_i hyppigheden og f_i frekvensen for intervallet, beregnes μ ved: $\mu = \frac{1}{n} \cdot \sum_{i=1}^{k} m_i \cdot h_i$ eller $\mu = \sum_{i=1}^{k} m_i \cdot f_i$ **Bemærk:** Denne udregnede middelværdi vil som udgangspunkt afvige en smule fra den middelværdi man ville finde, hvis man fandt den for alle observationerne, når de IKKE var grupperede.
Median Den midterste observation. Ekstreme værdier påvirker middelværdien, men ikke medianen. Derfor er medianen et bedre udtryk for hovedtendensen i et observationssæt.	**Median** Den midterste observation. Aflæses på sumkurven ved at gå vandret ud fra 2. aksen ved 50% indtil kurven rammes, hvorefter man går lodret ned til aflæses af medianen. Bemærk, at medianen IKKE er observationsintervallet, men et tal i intervallet (evt. med decimaler).
Nedre og øvre kvartil Den midterste observation i den nedre halvdel af observationssættet og den midterste observation i den øvre halvdel.	**Nedre og øvre kvartil** Den midterste observation i den nedre halvdel af observationssættet og den midterste observation i den øvre halvdel. Aflæses ligesom medianen – blot skal man gå ud fra henholdsvis 25% og 75%.
(Fraktiler) Hvis man vil aflæse fraktiler, skal man anvende samme metode som med grupperede observationssæt – bare anvendt på trappediagrammet i stedet for på sumkurven.	**Fraktiler** (angives altid med en % foran) Angiver afgrænsningen af en vis andel observationer. F.eks. er 10%-fraktilen den observation, hvor 10% af observationerne ligger under. Fraktilerne aflæses ligesom medianen, der bare er et andet ord for 50%-fraktilen, ligesom nedre kvartil er 25%-fraktilen og øvre kvartil 75%-fraktilen.
Mindste observation og største observation **Variationsbredde = max - min**	**Mindste observation og største observation** Bemærk, at hvis man ikke kender de oprindelige data (og dermed min og max, skal man bruge venstre intervalendepunkt for det første interval og højre endepunkt for sidste interval.
Varians Med de samme betegnelser som for middelværdien beregnes variansen: $\text{var}(x) = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \mu)^2 \cdot h_i$	**Varians** Med de samme betegnelser som for middelværdien, beregnes variansen: $\text{var}(x) = \frac{1}{n} \cdot \sum_{i=1}^{k} (m_i - \mu)^2 \cdot h_i$ eller $\text{var}(x) = \sum_{i=1}^{k} (m_i - \mu)^2 \cdot f_i$
Standardafvigelse/Spredning Beregnes som $\sigma(x) = \sqrt{\text{var}(x)}$	**Standardafvigelse/Spredning** $\sigma(x) = \sqrt{\text{var}(x)}$

Typetal og typeinterval

Typetal optræder ved ikke-grupperede observationssæt, og *typeintervaller* optræder ved grupperede observationssæt.

Vi husker definitionen fra sandsynlighedsregning:

Definition: I følgende definition anvendes entalsbetegnelser. Hvis der er flere størrelser, der opfylder betingelserne, er der flere typetal eller typeintervaller.

a) Inden for statistik er *typetallet* observationen med den største hyppighed.

b) Inden for statistik er *typeintervallet* observationsintervallet med den største **tæthed**.

c) Inden for sandsynlighedsregning er *typetallet* den værdi af den stokastiske variabel, der har størst sandsynlighed.

Typetal er et meget simpelt begreb, der ikke plejer at volde problemer. Det er simpelthen det (eller de) tal i observationssættet, der optræder flest gange, dvs. observationen (eller observationerne) med den største hyppighed/frekvens. Som det fremgår af ovenstående formulering, kan der godt være flere typetal, nemlig hvis der er mere end én observation med den egenskab, at ingen anden observation har en større hyppighed.

Et *typeinterval* er et lidt mere kompliceret begreb. I hvert fald skal man passe lidt på, da man i sjældne tilfælde kan gå i en "fælde". Der kan ligesom med typetal godt være flere typeintervaller, men forskellen mellem de to begreber er, at typeintervallet **ikke** nødvendigvis er intervallet (eller intervallerne) med den største intervalhyppighed.

Typeintervallet er det (eller de) interval(ler), der har den største *intervalhyppighedstæthed*, dvs. der hvor intervalhyppigheden delt med intervalbredden giver det største tal. Dette ses nemmest i et histogram, hvor typeintervallet er intervallet med ***den højeste søjle***.

Hvis man arbejder med intervaller med konstant intervalbredde, kan man godt nøjes med at se på intervalhyppigheden, men hvis man arbejder med forskellige intervalbredder, er det vigtigt at anvende den rigtige definition. Man kan godt anvende Gym-pakkens kommando *typetal*, men der er fejl i kommandoen *typeinterval*, da den ikke tager hensyn til forskellige intervalbredder.

Eksempel 1c, 2c, 3c og 4c: Typetal og typeintervaller.
Matematik A-niveau: Typetallet (typekarakteren) er 7 (3199 er den største hyppighed)
Matematik B-niveau: Typetallet er 7 (2786 er den største hyppighed)
Højde af Det Skæve Tårn: Typeintervallet er]40,50]
Længde af bymuren i Lucca: Typeintervallet er]3900,4200]

Middelværdi / Gennemsnit

Middelværdien udregnes enten med formlen kendt fra sandsynlighedsregning, hvor man blot erstatter sandsynligheder med frekvenser, eller ved hjælp af hyppigheder og n.

Eksempel 1d: Middelværdi matematik A:

Formlen $\mu = \sum_{i=1}^{k} x_i \cdot f_i$ benyttes. Der er anvendt flere decimaler i udregningen, end man kan se i skemaet i Eksempel 1a:

$\mu = -3 \cdot 0.00779 + 0 \cdot 0.10181 + 2 \cdot 0.06784 + 4 \cdot 0.18082 + 7 \cdot 0.30018 + 10 \cdot 0.24078 + 12 \cdot 0.10078 = 6.55401$

Dvs. middelværdien er **6,6**

Eksempel 2d: Middelværdi for matematik B:

Formlen $\mu = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot h_i$ benyttes:

$$\mu = \frac{1}{11135.} \cdot (-3 \cdot 405 + 0 \cdot 2287 + 2 \cdot 1145 + 4 \cdot 2225 + 7 \cdot 2786 + 10 \cdot 1844 + 12 \cdot 443) = 4.780691513$$

Middelværdien kan også bestemmes med Gym-pakkens kommandoer *middel* og *gennemsnit*:

$$middel(EksamenB) = 4.78069151324652$$
$$gennemsnit(EksamenB) = 4.78069151324652$$

Ved grupperede observationer skal man anvende intervalmidtpunkterne som observationer. Da intervalopdelingen ikke er entydig, er der heller ikke noget entydigt svar på den udregnede middelværdi. Hvis man kender det oprindelige datasæt (før grupperingen), kan man selvfølgelig udregne middelværdien på dette sæt.

Eksempel 3d: Middelværdi for højden af Det Skæve Tårn:

Formlen $\mu = \frac{1}{n} \cdot \sum_{i=1}^{k} m_i \cdot h_i$ benyttes. Bemærk altså, at det er intervalmidtpunkterne, der indgår i udregningen. Desuden kan man igen benytte Gym-pakkens *middel* og *gennemsnit*:

$$\mu = \frac{1}{24.} \cdot (15 \cdot 1 + 25 \cdot 0 + 35 \cdot 4 + 45 \cdot 9 + 55 \cdot 5 + 65 \cdot 1 + 75 \cdot 2 + 90 \cdot 0 + 150 \cdot 1 + 250 \cdot 1) = 60.41666667$$
$$middel(HøjdeSkæveTårn) = 60.4166666666667$$

Hvis man beregner middelværdien ud fra de oprindelige data (inden gruppering), får man 62,8.

Eksempel 4d: Middelværdi for længden af bymuren i Lucca:
Her bruges Gym-pakkens kommando:

$$gennemsnit(Bymurlængde) = 4244.44444444444$$

Beregning på de oprindelige data giver **4252 meter.**

Der er altså en forskel, men afvigelse vil i de fleste tilfælde være ubetydelig.

Kvartilsæt og udvidet kvartilsæt (median, nedre kvartil, …)

Kvartilsættet består af medianen (*m* eller Q_2), nedre kvartil (Q_1) og øvre kvartil (Q_3).

Et udvidet kvartilsæt består udover ovennævnte også af mindste observation (min) og største observation (max).

Kvartilsættet angives (Q_1, m, Q_3)

Det udvidede kvartilsæt angives $(\min, Q_1, m, Q_3, \max)$

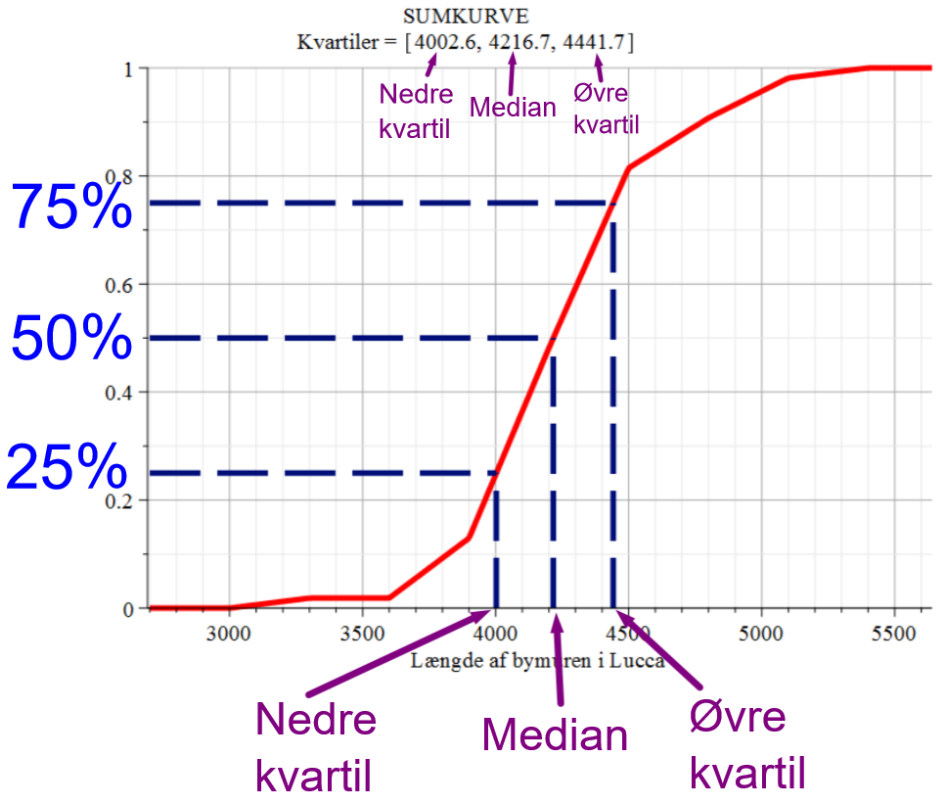
Kvartilbredde, IQR (Interquartile range) er afstanden mellem nedre og øvre kvartil: $IQR = Q_3 - Q_1$

For **grupperede observationer** bestemmes kvartilsættet ved hjælp af sumkurven.

For **ikke-grupperede observationer** findes der (mindst) to forskellige metoder til at bestemme kvartilsættet, og faktisk kan de to metoder i nogle tilfælde give (oftest kun lidt) forskellige resultater. Den slags er naturligvis total uhørt i "rigtig" matematik, men det er ikke noget problem i lige netop denne sammenhæng.

Sumkurvemetoden (grupperede observationer):

Eksempel 4e: Sumkurvemetoden er illustreret i Eksempel 4b, da Gym-pakkens *plotSumkurve* automatisk aflæser kvartilsættet:



Man går vandret ind til grafen fra 25%, 50% og 75%, og derefter lodret ned og aflæser henholdsvis nedre kvartil, median og øvre kvartil.

Disse tal fortæller, at:

25% af eleverne har målt længder mindre end 4003 m.

50% af eleverne har målt længder mindre end 4217 m.

75% af eleverne har målt længder mindre end 4442 m.

Da man med en sumkurve har fået smurt målingerne jævnt ud over hvert interval, er der ikke længere nogen målinger for de enkelte værdier. Man taler nu kun om intervaller. F.eks. kan man sige, at 50% af målingerne ligger mellem 4003 m og 4442 m.

Derfor kunne man også ovenfor have sagt, at 25% af eleverne har målt længder på højst 4003 m.

Kvartilsættet er (4003 m, 4217 m, 4442 m)

Det udvidede kvartilsæt er (3090 m, 4003 m, 4217 m, 4442 m, 5280 m)

$IQR = 4441,7\,\text{m} - 4002,6\,\text{m} = 439,1\,\text{m}$

Variationsbredden = 5280 m – 3090 m = 2190 m

(Mindste og største observation er fundet i det oprindelige datasæt i Eksempel 4a)

Ordnet-følge-metoden (ikke-grupperede observationer):

Det er denne metode, som TI n'spire, Excel og Gym-pakkens *kvartiler* anvender (og som vist nok er den mest anvendte internationalt set).

Observationerne stilles op i en ordnet følge (dvs. efter størrelse med den mindste først):

F.eks. 1, 2, 2, 3, 5, 5, 5, 6, 7, 7, 7, 7, 9, 11, 14, 14, 14, 15, 18

Hvis der er et ulige antal observationer, er medianen det midterste tal i følgen.
Hvis der er et lige antal observationer, er medianen gennemsnittet af de to midterste tal.

Medianen deler observationssættet i to lige store dele (hvis der er et ulige antal observationer, og medianen derfor rammer et tal i følgen, fjernes dette tal og indgår altså ikke i nogen af delene).

Den nedre kvartil bestemmes efterfølgende som medianen af den nedre halvdel, mens den øvre kvartil er medianen af den øvre halvdel.

Eksempel 5*:* Vi ser igen på 1, 2, 2, 3, 5, 5, 5, 6, 7, 7, 7, 7, 9, 11, 14, 14, 14, 15, 18

I ovenstående følge er der 19 observationer. Det tiende tal, der er 7, er derfor medianen.

Dette tal fjernes og deler nu observationssættet i:

1, 2, 2, 3, 5, 5, 5, 6, 7 nedre halvdel ; 7, 7, 9, 11, 14, 14, 14, 15, 18 øvre halvdel.

Der er et ulige antal observationer i disse halvdele – nemlig 9 – så den nedre kvartil er det femte tal i den nedre halvdel (dvs. 5) og den øvre kvartil er det femte tal i den øvre halvdel (dvs. 14).

Desuden er den mindste observation 1 og den største observation 18.

Hermed er kvartilsættet (5,7,14)

Det udvidede kvartilsæt er (1,5,7,14,18)

$IQR = 14 - 5 = 9$ Variationsbredden $= 18 - 1 = 17$

Eksempel 6a*: Et nyt observationssæt er 0, 0, 2, 2, 4, 7, 7, 10, 12, 12*

Der er et lige antal observationer – nemlig 10 – og derfor er medianen gennemsnittet af den femte og den sjette observation (der er 4 og 7). Medianen er altså 5,5, selvom der ikke er nogen observation, der har denne værdi.

Mediansnittet ligger mellem 4 og 7, så ingen observationer fjernes, når observationssættet deles i to lige store dele:

0, 0, 2, 2, 4 nedre halvdel ; 7, 7, 10, 12, 12 øvre halvdel

Der er et ulige antal observationer i disse halvdele – nemlig 5 – så den nedre kvartil er det tredje tal i den nedre halvdel (dvs. 2) og den øvre kvartil er det tredje tal i den øvre halvdel (dvs. 10).

Desuden er den mindste observation 0 og den største observation 12.

Hermed er kvartilsættet (2,5.5,10). En anden skrivemåde er (2;5,5;10)

Det udvidede kvartilsæt er (0,2,5.5,10,12)

$IQR = 10 - 2 = 8$ Variationsbredden $= 12 - 0 = 12$

Gym-pakken har kommandoerne *median* og *kvartiler*:

Eksempel 1e og 2e:

$median(EksamenA) = 7.$ $median(EksamenB) = 4.$

$kvartiler(EksamenA) = [4., 7., 10.]$ $kvartiler(EksamenB) = [2., 4., 7.]$

Man kan sammenligne med trappediagrammerne i Eksempel 1b og 2b.

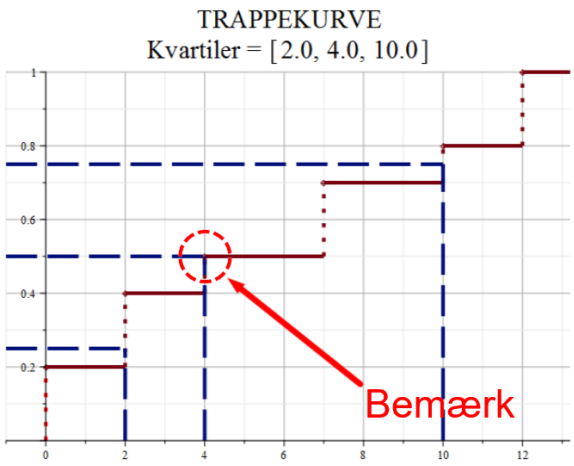
Trappediagramsmetoden:

Denne metode svarer til sumkurvemetoden for grupperede observationer.
Kvartilerne (nedre kvartil, median og øvre kvartil) bestemmes ved at gå vandret ud fra 2. aksen ved frekvenserne 25%, 50% og 75%, indtil man rammer trappen. Derfra går man lodret ned på 1. aksen og aflæser nedre kvartil, median og øvre kvartil. Sådan gør Gym-pakkens *plotTrappekurve* (se Eksempel 1b og 2b).

Denne metode **kan** give kvartilsæt, der afviger fra ordnet-følge-metoden, da den altid kun vil give værdier fra observationssættet, mens ordnet-følge-metoden sommetider giver gennemsnitsværdier, der ikke optræder i observationssættet.

Eksempel 6b: Med tallene fra Eksempel 6a giver trappediagramsmetoden:

$Eksempel6 := [0, 0, 2, 2, 4, 7, 7, 10, 12, 12]$:
$plotTrappekurve(Eksempel6)$



Fra Eksempel 6a ved vi, at ordnet-følge-metoden giver kvartilsættet (2,5.5,10).

Forskellen opstår, når den vandrette linje, der udgår fra 2. aksen, ikke rammer en lodret del af et trappetrin, men derimod flugter med en vandret del af et trappetrin (se figuren ovenfor).

Når man skal sætte ord på kvartilsættet, er det trappediagramsmetoden, der giver bedst mening:

De 25% laveste målinger er på 2 eller derunder.
De 50% laveste målinger er på 4 eller derunder (4 kan evt. erstattes af 5,5).
De 75% laveste målinger er på 10 eller derunder.

Bemærk, at man med trappekurver **IKKE** kan anvende formuleringen "har fået mindre end 2", da der her er målinger med værdien 2.

Om middelværdier og medianer

Middelværdien og medianen for et datasæt vil typisk ligge tæt på hinanden. Hvis tæthedsfunktionen (evt. et pindediagram eller et histogram) er symmetrisk omkring den lodrette linje gennem middelværdien, vil median og middelværdi være ens.

Men der kan også være store eller væsentlige forskelle på de to. Selve værdierne af de enkelte observationer indgår i beregningen af middelværdien, mens medianen kun er baseret på ordningen af de enkelte observationer. F.eks. vil observationssættene 1,2,3,4,5 og 1,2,3,4,10000000000 have ens medianer (3), mens middelværdierne vil være vidt forskellige (3 og 2000000002).

Medianen er altså ikke påvirket af ekstreme værdier, og derfor vil den i en del situationer være et bedre mål for en søgt værdi. Lad os se på en oversigt over vores 4 observationssæt:

Observationssæt		Middelværdi	Median	Rigtig værdi
Matematik A		6,6	7	--
Matematik B		4,8	4	--
Højde af Det Skæve Tårn	Rådata	62,8 m	49,2 m	55,9 m
	Grupperet	60,4 m	47,8 m	
Længde af bymuren i Lucca	Rådata	4252 m	4223 m	4223 m
	Grupperet	4244 m	4217 m	

- I matematik A og B ligger medianen (stort set) så tæt på middelværdien, som det er muligt. I disse tilfælde giver det ikke mening at tale om en "rigtig" værdi, for her **er** 6,6 den rigtige middelværdi, mens 7 **er** den rigtige median.
- Vi ser i bestemmelsen af højden af Det Skæve Tårn, at grupperingen (som forventet) har ændret lidt på værdierne for middelværdi og median. Det bemærkes også, at middelværdien er væsentlig højere end medianen, hvilket skyldes de to ekstreme målinger 271,5 m og 169,1 m (der slet ikke opvejes af den ekstremt lille værdi 10,9 m). I dette tilfælde er medianen dog kun en anelse tættere på den rigtige værdi end middelværdien.
- Ved længden af bymuren i Lucca bemærkes det, at der kun er en relativ lille forskel på middelværdi og median. Det er tilfældigt, at medianen lige præcis rammer den rigtige værdi (Og det kan være svært at tale om en præcis rigtig værdi, da det ikke er klart, hvordan man skal måle længden af bymuren. En angivelse på 4,2 km er nok mere rimelig).

Vi skal snart se på begrebet *skævhed*, der forsøger at sætte tal på den asymmetri, der også kan give forskel på middelværdi og median.

Fraktiler

Kvartilerne er nogle særlige *fraktiler*, dvs. *fraktil* er et mere overordnet begreb end *kvartil*.

p%-fraktilen er den værdi på førsteaksen, hvorom man kan sige, at de p% laveste målinger har denne værdi eller derunder.

Den nedre kvartil er 25%-fraktilen, medianen er 50%-fraktilen og øvre kvartil er 75%-fraktilen.

Fraktilerne aflæses på trappekurver eller sumkurver på samme måde som kvartilerne.

Eksempel 1f, 2f, 3f og 4f: I Gym-pakken findes kommandoen *fraktil*:

$$fraktil(EksamenA,\ 0.17) = 2$$
$$fraktil(EksamenA,\ 0.83) = 10$$

$$fraktil(EksamenB,\ 0.17) = 0$$

$$fraktil(HøjdeSkæveTårn,\ 0.40) = 45.11111111$$

$$fraktil(Bymurlængde,\ 0.90) = 4776.000000$$

Til matematik A-eksamen har de 17% af eleverne med laveste karakterer fået karakteren 02 eller derunder. De 83% af eleverne med laveste karakterer har fået karakteren 10 eller derunder. Hermed kan man også sige, at de 17% med de højeste karakterer har fået karakteren 10 eller derover.

De 17% af matematik B-eleverne med laveste karakterer har fået karakteren 00 eller derunder.

40% af eleverne har målt Det Skæve Tårn til at være mindre end 45,1 m.
60% af eleverne har målt Det Skæve Tårn til at være større end 45,1 m.
90% af eleverne har målt bymuren i Lucca til at være kortere end 4776 m.

Varians og spredning (standardafvigelse)

Begreberne *varians* og *spredning* kender vi allerede fra sandsynlighedsregning, og formlerne er de samme, bortset fra at sandsynligheder er erstattet af frekvenser.

Vi skal senere se på formlerne i forbindelse med stikprøver fra en population, hvor de ser lidt anderledes ud. Gym-pakken har kommandoerne *varians* og *spredning*, som man kan bruge:

Eksempel 1g, 2g, 3g og 4g: Med vores velkendte datasæt får man:

$$\text{var}(x) = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \mu)^2 \cdot h_i \qquad \sigma(x) = \sqrt{\text{var}(x)}$$

$varians(EksamenA) = 13.5786029663090$

$\sqrt{13.5786029663090} = 3.684915599$

$spredning(EksamenA) = 3.68491559826124$

$spredning(EksamenB) = 3.95348400828980$

$spredning(HøjdeSkæveTårn) = 46.2537536014346$

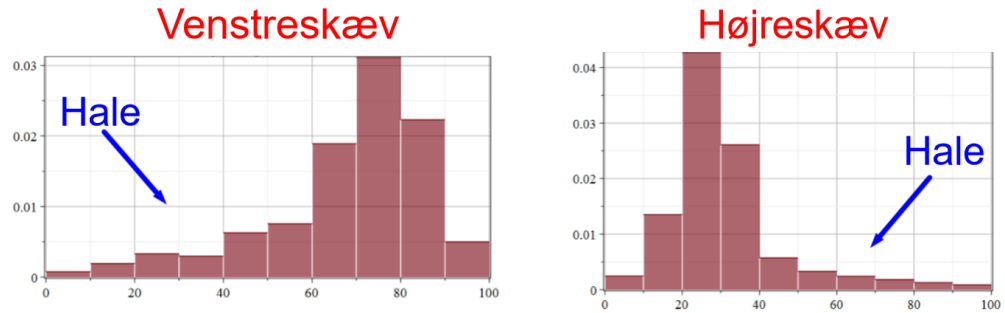
$spredning(Bymurlængde) = 370.393517795184$

Spredningen ved B-niveau-eksamen var altså større end ved A-niveau-eksamen.
Spredningen ved målingen af højden på Det Skæve Tårn er 46,3 m
Spredningen ved målingen af længden af bymuren er 370 m

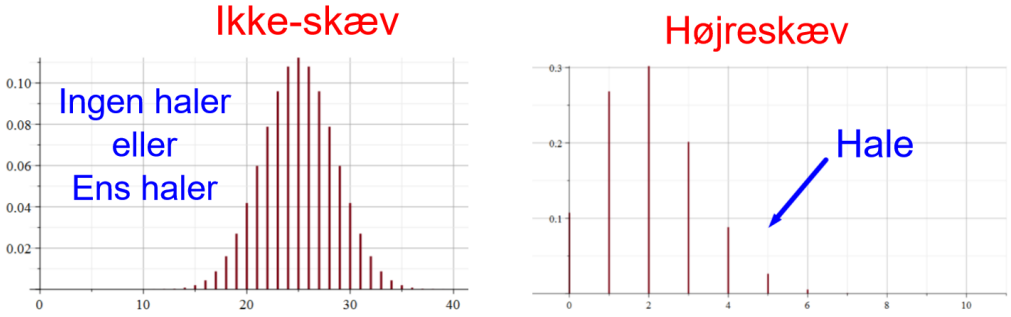
Skævhed

Med begrebet *skævhed* forsøger man at indfange asymmetrier i tæthedsfunktionerne. I den forbindelse benytter man begrebet *hale* (se nedenfor). Hvis halen ligger til højre, er fordelingen højreskæv, og hvis halen ligger til venstre, er fordelingen venstreskæv. Begrebet kan bruges på alle former for tæthedsfunktioner.

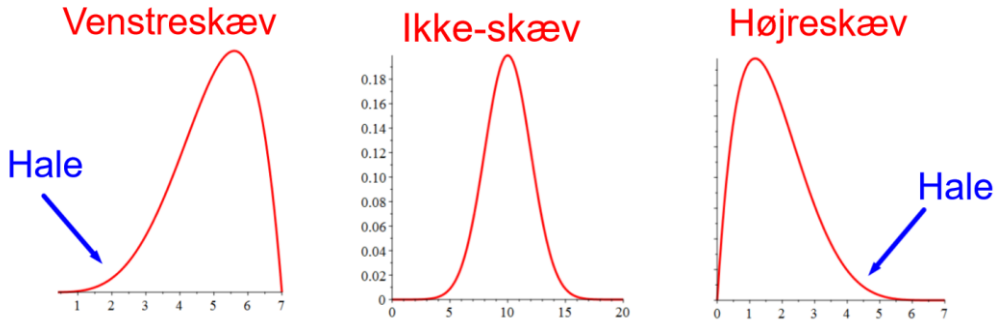
Grupperede observationssæt



Ikke-grupperede observationssæt



Kontinuerte tæthedsfunktioner



Man kan godt beregne en værdi for skævheden (betegnet med et lille gamma, γ). Vores formelsamling arbejder ikke med et mål for skævheden, men kun kvalitative betegnelser, der bestemmes ud fra middelværdien μ og medianen *m*:

Skævhed

"Korrekt":		**Formelsamling:**
$\gamma < 0$: Venstreskæv		$\mu < m$: Venstreskæv
$\gamma = 0$: Ikke-skæv	$\gamma = \frac{1}{n} \cdot \sum_{i=1}^{k} \frac{(x_i - \mu)^3}{\sigma^3} \cdot h_i$	$\mu = m$: Ikke-skæv
$\gamma > 0$: Højreskæv		$\mu > m$: Højreskæv

Problemet med begrebet *skævhed* er, at der vist ikke er nogen enighed om definitionen. På wikipedia fandt jeg hurtigt 7 forskellige formler. Formelsamlingens beskrivelse repræsenterer en formel, der i bedste fald er gammeldags. I værste fald forkert.

Formelsamlingens tommelfingerregel ser ud til at være baseret på en af formlerne $\gamma = \frac{\mu - m}{\sigma}$ og $\gamma = 3 \cdot \frac{\mu - m}{\sigma}$. Disse formler udnytter, at en hale løst sagt påvirker middelværdien mere end medianen.

Formlen $\gamma = \frac{1}{n} \cdot \sum_{i=1}^{k} \frac{(x_i - \mu)^3}{\sigma^3} \cdot h_i$ er med kuberne tydeligvis baseret på den idé, at afvigelserne fra middelværdien skal vægte højere, jo større de er. Og så vil jeg tro, at eksponenten 3 er lidt vilkårlig, men at den blot skal sikre, at fortegnet på afvigelserne fra middelværdien beholdes (hvor man jo med variansens eksponent 2 sikrer sig, at alle afvigelser regnes positive).

Man kan med den "rigtige" formel komme ud for, at middelværdien ligger til højre for medianen i en venstreskæv fordeling (dvs. tommelfingerreglen fra formelsamlingen holder i så fald ikke).

Eksempel 1h, 3h og 4h: Udregningerne i det følgende er foretaget med formlerne:

Formel 1: $\gamma = 3 \cdot \frac{\mu - m}{\sigma}$ Formel 2 (rigtige): $\gamma = \frac{1}{n} \cdot \sum_{i=1}^{k} \frac{(x_i - \mu)^3}{\sigma^3} \cdot h_i$

Eksamen A-niveau:	Formel 1: $\gamma = -0,36$	Formel 2: $\gamma = -0,0049$ (venstreskæv)
Højde af Det Skæve Tårn:	Formel 1: $\gamma = 0,80$	Formel 2: $\gamma = 3,10$ (højreskæv)
Længden af bymuren:	Formel 1: $\gamma = 0,24$	Formel 2: $\gamma = 0,15$ (højreskæv)

I vores tre tilfælde ender man med den samme kvalitative konklusion med de to formler (og dermed også med tommelfingerreglen). Men som sagt gælder det ikke altid.

Det bemærkes, at den rigtige formel giver en højere værdi for skævheden end Formel 1 i tilfældet med højden af Det Skæve Tårn, hvor der er to ekstremt høje målinger, mens den rigtige formel giver en mindre værdi end Formel 1 i tilfældet med bymuren, hvor der ikke er nogen ekstreme værdier.

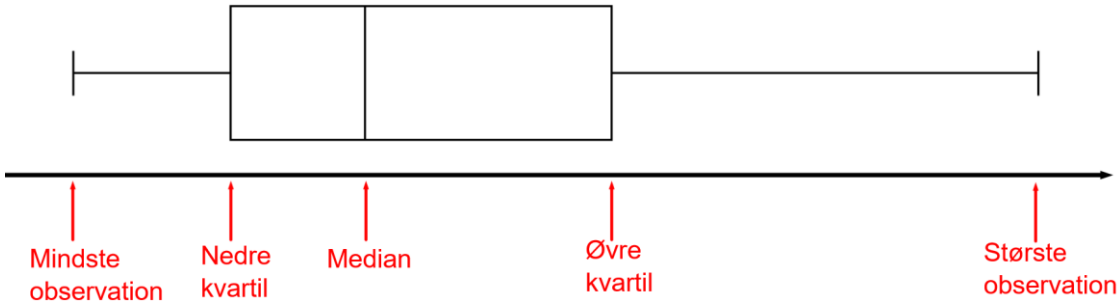
Boksplot / Boxplot

Median, nedre kvartil, øvre kvartil, mindste observation og **største observation** er fem deskriptorer, der fortæller en del om selve observationssættet, og de kan opstilles overskueligt i et såkaldt *boksplot* (opfundet i 1969).

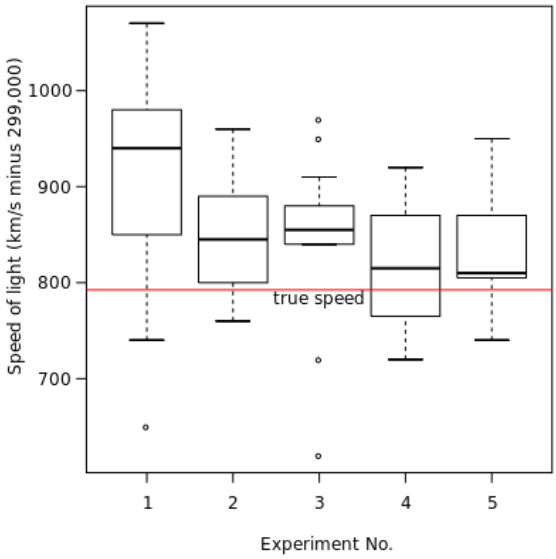
Først tegnes en 1. akse, præcis som hvis det var et pindediagram eller histogram (dvs. husk, at skalaen som udgangspunkt skal være jævn).

Der er ingen 2. akse.

I en vilkårlig højde over 1. aksen tegnes 5 lodrette linjer, der angiver henholdsvis mindste observation, nedre kvartil, median, øvre kvartil og største observation. De to yderste linjer tegnes lidt mindre end de andre. Derefter tegnes 4 vandrette streger, så man får en "boks" med udseendet:



Et boksplot kan også tegnes lodret (hvor 1.aksen altså også angives lodret), og man vil ofte afbilde flere bokse i samme diagram, så boksene kan sammenlignes:



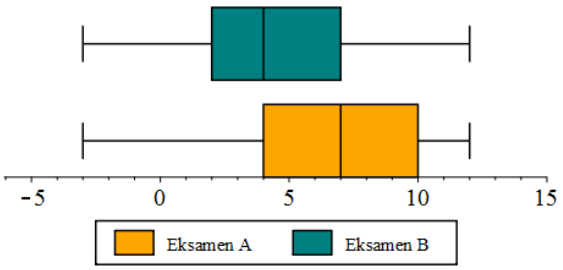
På figuren ovenfor er der noget, der ser mystisk ud. Der er tilsyneladende målinger, der ligger uden for mindste og største observation (se de små cirkler). Det skyldes dog blot, at angivelsen af de yderste linjer kan være baseret på forskellige regler.

Som sagt bruger vi mindste og største observation, men man kan godt vælge at frasortere ekstreme målinger, eller man kan benytte andre deskriptorer til fastsættelsen.

Med Gym-pakken kan man tegne boksplot. Nogle gange giver det mening at tegne to bokse i samme diagram, så man kan sammenligne dem (f.eks. matematik A og matematik B), mens det andre gange ville være meningsløst (f.eks. højde af Det Skæve Tårn og længde af bymuren i Lucca).

Eksempel 1i, 2i, 3i og 4i: Gym-pakken har kommandoen *boksplot*:

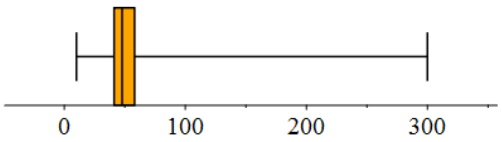
boksplot(*EksamenA*, *EksamenB*)



De 50% bedste på matematik A ligger i samme karakterområde som de 25% bedste på matematik B. Og de 50% svageste på matematik B ligger i samme karakterområde som de 25% svageste på matematik A.

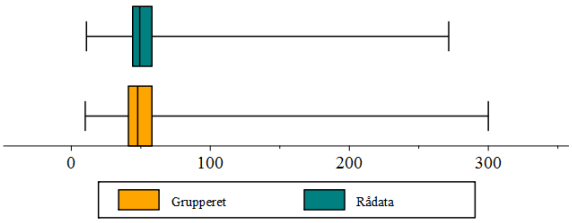
boksplot(*HøjdeSkæveTårn*)

Kvartiler = [41.11, 47.78, 58.00]



Dette kan sammenlignes med boksplottet baseret på rådata:

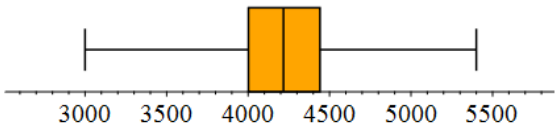
boksplot(*HøjdeSkæveTårn*, *SkæveTårnRådata*)



Der er forskel på de to, men forskellen er ikke væsentlig, når man blot skal overskue data.

boksplot(*Bymurlængde*)

Kvartiler = [4003., 4217., 4442.]

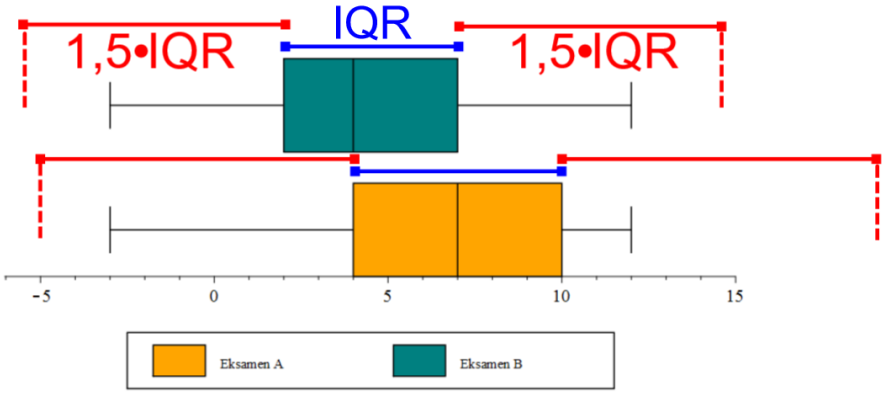


Højden af Det Skæve Tårn og længden af bymuren giver begge boksplot med relativt smalle bokse, dvs. kvartilbredden er væsentlig mindre end variationsbredden. Det er en vigtig observation, så vi skal se på nu.

Kvartilbredden kan benyttes til at definere ekstreme målinger, som man kalder *outliers.*

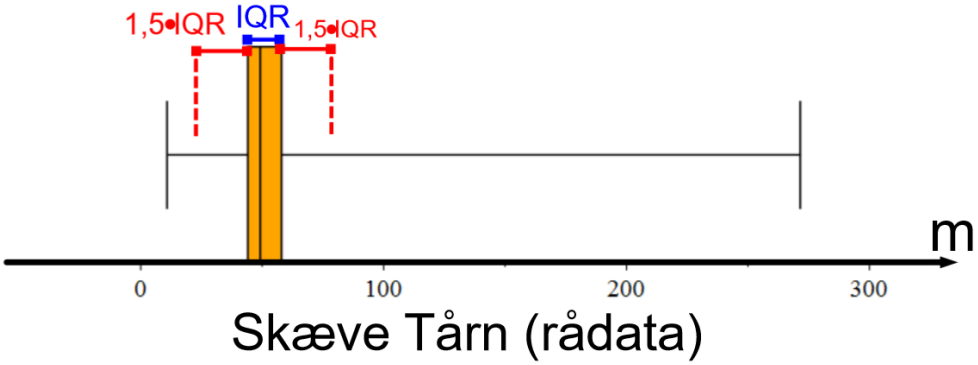
Outlier: Hvis en observation ligger mere end $1{,}5 \cdot IQR$ under nedre kvartil eller mere end $1{,}5 \cdot IQR$ over øvre kvartil, kaldes den en *outlier*.

Eksempel 1j, 2j, 3j og 4j: Vi ser på, om der findes *outliers* i vores 4 datasæt:



Der er ingen observationer uden for de stiplede røde linjer, så der er ingen *outliers* i karaktererne for matematik A og B.

For observationssættet med højden af Det Skæve Tårn ser det anderledes ud:



Det er *outliers* til begge sider. Det kan være svært at vurdere ud fra figuren, hvor mange af de høje værdier, der er *outliers*, men hvis det er vigtigt, kan man regne på det.

Kvartilsættet er (44.15 , 49.2 , 58)

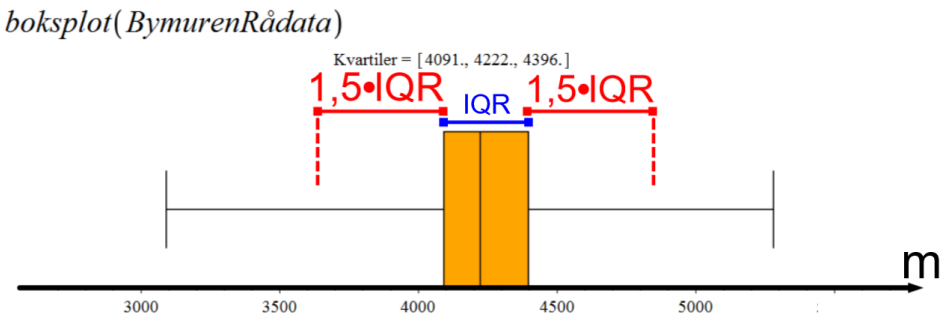
Dermed er:

$IQR = 58 - 44{,}15 = 13{,}85$

$44{,}15 - 1{,}5 \cdot 13{,}85 = 23{,}375$

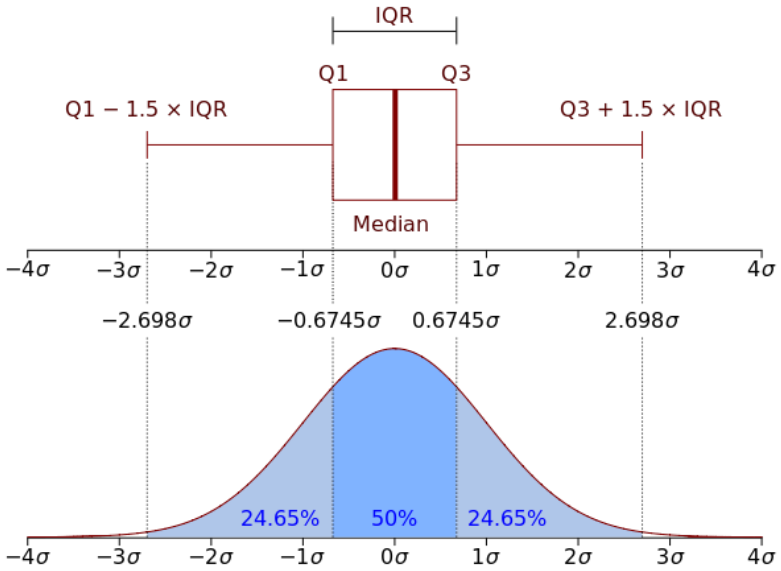
$58 + 1{,}5 \cdot 13{,}85 = 78{,}775$

Dvs. alt under 23,375 m og alt over 78,775 m er *outliers* (altså 10,9 m, 169,1 m og 271,5 m)



Her er der også *outliers*. Find selv ud af hvilke.

Man kan omsætte normalfordelinger til følgende boksplot (mindste og største observation er erstattet af grænserne for *outliers*, da normalfordelinger ikke har mindste og største observationer):



Hvis en størrelse følger en normalfordeling, vil man altså have 0,7% *outliers*, **hvis** man anvender denne regel fra boksplottet.

Tallet 1,5 er ikke noget eksakt udregnet tal. John Tukey, der opfandt boksplottet, skulle have udtalt, at tallet 1,5 kommer fra, at 1 er for lidt og 2 for meget.

Generelt er der ikke nogen fast regel for, hvad er outlier er. I det danske gymnasium skal du bruge:

Outlier: En observation, der er mere end $1{,}5 \cdot IQR$ fra nærmeste kvartil.

Exceptionelt udfald: Et udfald, der er mere end 3 spredninger fra middelværdien.

Median og *IQR* er *robuste* begreber, da de ikke påvirkes af ekstreme udfald. Middelværdi og spredning er ikke robuste begreber. Begrebet *outlier* er altså knyttet til de robuste begreber.

Vigtigt: *Outliers* (og *exceptionelle udfald*) er mere end bare ord. Begreberne kan fungere som en regel for, hvornår en måling skal smides væk. Dvs. efter at have udført en hel måleserie (hvilket er nødvendigt, da man skal kende kvartilsættet for at kunne bestemme *IQR* og dermed afgøre, hvilke målinger der er outliers), kan man vælge at smide alle outliers væk, **inden** man udregner de statistiske deskriptorer (middelværdi, spredning, ...), da de vil give et misvisende billede.

Til grund for denne handlemåde ligger den tanke, at outliers er en slags fejl (f.eks. målefejl) eller udtryk for en effekt, der vil være misvisende, hvis den inddrages (Skal dværges højde inddrages, hvis man vil bestemme menneskers gennemsnitshøjde? Skal åndssvage inddrages, hvis man skal bestemme den gennemsnitlige IQ? ...)

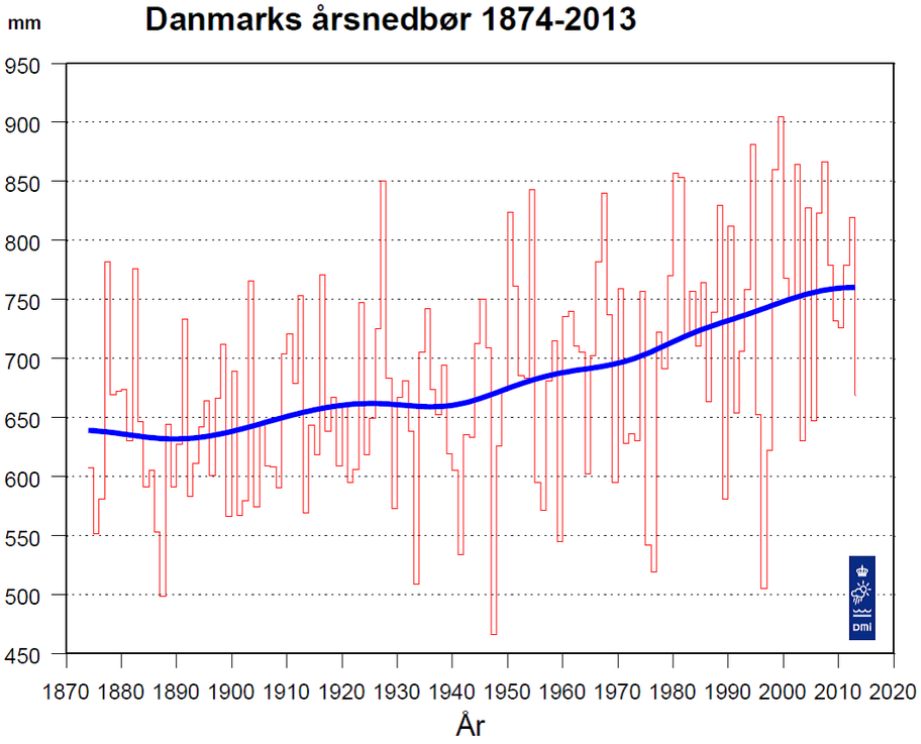
MEN man skal være meget varsom med bare at smide outliers væk, for de kunne jo indeholde noget "virkelig" information. F.eks. opdagede man hullet i ozonlaget nogle år senere, end man kunne have gjort, fordi softwaren i de satellitter, der målte på ozonlaget, smed disse ekstreme, men rigtige, målinger væk. Det var først, da nogle forskere målte fra jorden, at man blev opmærksom på denne fejl.

John Tukey opfandt boksplottet som en måde at opstille resultater visuelt på, og for at man kunne sammenligne forskellige observationssæt (f.eks. kvinders løn vs. mænds løn). Dette var tænkt som et alternativ til de test, vi senere skal beskæftige os en hel del med. Tukey mente, at disse test blot var en jagt på tal, der fører til konklusioner, og at det oftest ikke er hensigtsmæssig, da sådanne test tit forudsætter, at datamaterialet følger en bestemt fordeling (oftest en normalfordeling).

Vi skal senere se på disse test, der giver os nogle tal, som vi skal lære at forholde os til.

Som opsamling på den deskriptive statistik ses her et eksempel, der bl.a. viser, hvad man kan gøre, hvis matricerne får mere end 10 rækker, og som har en meget vigtig pointe til sidst.

Eksempel 7 (grupperet observationssæt): Årlig nedbørsmængde i Danmark.
Vi ser på følgende indsamlede data:



Egentlig har man allerede anvendt deskriptiv statistik på datamaterialet, da man har indsat nedbørsmængden som funktion af tiden, hvilket viser en klar tendens til øget nedbørsmængde. Vi vil nu beskrive datamaterialet på en anden måde, nemlig ved at gruppere materialet og lave histogram og sumkurve.

Vi skal først have vurderet nogle passende intervaller for nedbørsmængden. Det virker oplagt med en intervalbredde på 50 mm, men vi kan se, at der ligger mange målinger i området 600-750 mm, og derfor inddeler vi i intervaller på 25 mm i dette område. Desuden gøres de to yderste intervaller 100 mm brede (observationssættets størrelse blev 136, så jeg må have overset et par år):

Nedbørsmængde	]450,550]	]550,600]	]600,625]	]625,650]	]650,675]	]675,700]	]700,725]	]725,750]	]750,800]	]800,850]	]850,950]
Intervalhyppighed	8	18	14	17	14	9	14	10	16	9	7
Her skulle have været intervalfrekvens og kumuleret intervalfrekvens, men jeg anvender Gym-pakken til udregningerne.											

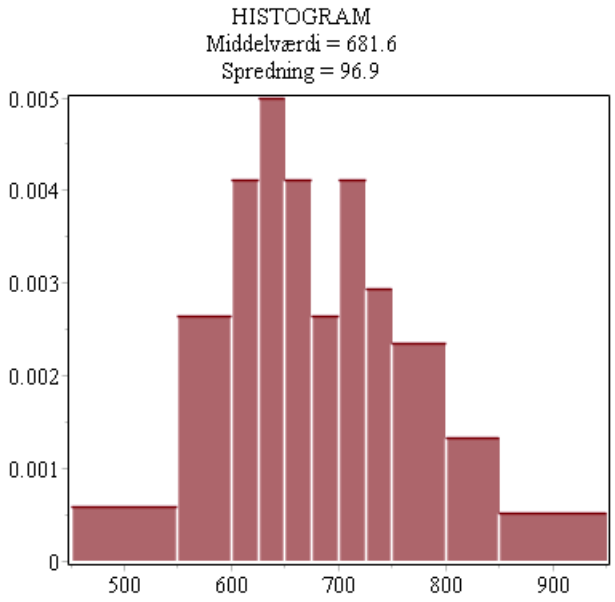
```
N := (450..550, 550..600, 600..625, 625..650, 650..675, 675..700, 700..725, 725..750, 750..800, 800..850, 850..950|8, 18, 14, 17, 14, 9, 14, 10, 16, 9, 7) :
frekvensTabel(N)
```

observation	hyppighed	frekvens	kumuleret
450 .. 550	8	0.0588	0.0588
550 .. 600	18	0.132	0.191
600 .. 625	14	0.103	0.294
625 .. 650	17	0.125	0.419
650 .. 675	14	0.103	0.522
675 .. 700	9	0.0662	0.588
700 .. 725	14	0.103	0.691
725 .. 750	10	0.0735	0.765
750 .. 800	16	0.118	0.882
800 .. 850	9	0.0662	0.949
850 .. 950	7	0.0515	1

Bemærk, hvordan man indtaster grupperede observationssæt i Maple. Det bliver en 11x2-matrix, da der står en lodret streg efter angivelsen af de 11 intervaller.

Vi kan nu anvende Gym-pakken til at tegne et histogram:

$plotHistogram(N) =$



Vi kan her aflæse, at **typeintervallet** er]625,650], da det er den højeste søjle. Bemærk, at det IKKE er intervallet med den største intervalhyppighed, der er]550,600].

Middelværdien er beregnet ved at anvende intervalmidtpunkterne:

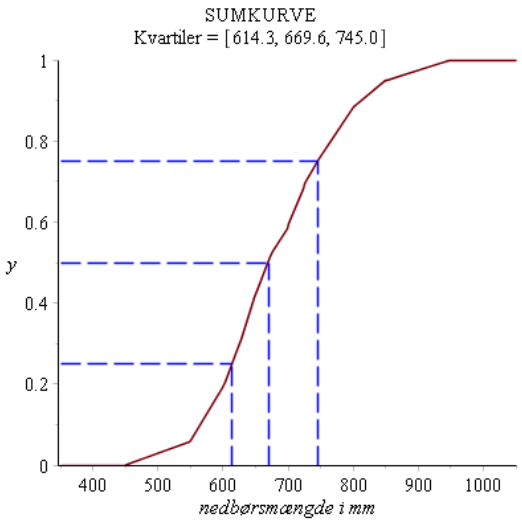
$$\mu = \frac{1}{n} \cdot \sum_{i=1}^{11} m_i \cdot h_i = \frac{1}{136} \cdot (500 \cdot 8 + 575 \cdot 18 + 612{,}5 \cdot 14 + 637{,}5 \cdot 17 + 662{,}5 \cdot 14 + ... + 900 \cdot 7) = 682$$

Der er altså gennemsnitligt faldet 682 mm nedbør om året i Danmark i perioden 1874-2013.

Bemærk, at dette tal højst sandsynligt vil afvige lidt fra en værdi, der var beregnet som et gennemsnit af hvert enkelt år.

På samme måde kan variansen udregnes ud fra intervalmidtpunkterne.
Endelig kan man med Gym-pakken få tegnet en sumkurve:

$plotSumkurve(N) =$



Vi har her fået oplyst kvartilsættet, der bl.a. fortæller os, at i de 25% mest nedbørsrige år faldt der mindst 745 mm nedbør, og i halvdelen af årene i perioden er der faldet højst 670 mm nedbør.

Lige som med de ikke-grupperede observationssæt, kan man desuden bestemme fraktiler ved:

$$fraktil(N,\ 0.1) = 565.5555556$$

Dvs. at i de 10% mindst nedbørsrige år faldt der højst 566 mm nedbør.

Man kan også være interesseret i at svare på den "modsatte" type spørgsmål, f.eks. "*I hvor stor en del af årene faldt der over 700 mm nedbør?*" eller "*I hvor stor en del af årene faldt der mellem 600 og 700 mm nedbør?*".

For at kunne besvare denne type spørgsmål skal man arbejde med sumkurven som et funktionsudtryk. Dette ordnes med Gym-pakken ved:

$$M(t) := sumkurve(N,\ t)\ :$$

Hvis man vil se, hvordan en sumkurve ser ud som funktion (ved en gaffelforskrift), skriver man :

$$M(t) = \begin{cases} 0 & t < 450 \\ 0.000588235294117647\ t - 0.264705882352941 & 450 \le t \text{ and } t < 550 \\ 0.00264705882352941\ t - 1.39705882352941 & 550 \le t \text{ and } t < 600 \\ 0.00411764705882353\ t - 2.27941176470588 & 600 \le t \text{ and } t < 625 \\ 0.00500000000000000\ t - 2.83088235294118 & 625 \le t \text{ and } t < 650 \\ 0.00411764705882353\ t - 2.25735294117647 & 650 \le t \text{ and } t < 675 \\ 0.00264705882352941\ t - 1.26470588235294 & 675 \le t \text{ and } t < 700 \\ 0.00411764705882353\ t - 2.29411764705882 & 700 \le t \text{ and } t < 725 \\ 0.00294117647058823\ t - 1.44117647058823 & 725 \le t \text{ and } t < 750 \\ 0.00235294117647059\ t - 1.00000000000000 & 750 \le t \text{ and } t < 800 \\ 0.00132352941176471\ t - 0.176470588235294 & 800 \le t \text{ and } t < 850 \\ 0.000514705882352942\ t + 0.511029411764706 & 850 \le t \text{ and } t < 950 \\ 1 & 950 \le t \end{cases}$$

Vi ser først, hvordan vi med funktionsudtrykket finder 10%-fraktilen:

$$solve(M(t) = 0.1) = 565.5555556$$

Vi ser, at det stemmer med det tidligere udregnede resultat.

Spørgsmålet: *I hvor stor en del af årene faldt der over 700 mm nedbør?*
Vi skal huske, at funktionsværdien angiver hvor stor en procentdel af observationerne, der ligger på eller **under** den indsatte værdi, så vi skal have:

$$1 - M(700) = 0.411764705882353$$

Dvs. at i 41% af årene faldt der mindst 700 mm nedbør.

Spørgsmålet: *I hvor stor en del af årene faldt der mellem 600 og 700 mm nedbør?*
Vi skal her have procentdelen mellem de to værdier, dvs:

$$M(700) - M(600) = 0.397058823529412$$

Dvs. at i 40% af årene faldt der mellem 600 og 700 mm regn.

Man kan gøre præcis det samme med ikke-grupperede observationssæt ved at erstatte *sumkurve* med *trappekurve*.

Vores behandling af datamaterialet i ovenstående eksempel leder hen til det vigtige spørgsmål, som man altid bør stille sig, når man arbejder med deskriptiv statistik: *Hvad er det egentlig, jeg vil illustrere, og har jeg valgt de rette redskaber til dette?*

I vores eksempel ser det temmelig tåbeligt ud, hvad jeg har foretaget mig. Bemærk i eksemplet, at vi begynder med en grafisk fremstilling, der tydeligt viser en tendens til øget nedbørsmængde. Denne information går fuldstændig tabt, når vi går over til at tegne et histogram og en sumkurve. Og endnu værre: Vores diagrammer og deskriptorer er misvisende, for vi kan f.eks. ikke længere forvente, at vi kun hvert fjerde år får en nedbørsmængde over 745 mm (øvre kvartil). Langt de fleste år efter år 2000 har haft nedbørsmængder over 745 mm, fordi nedbørsmængden er steget.

Hvis vi ville illustrere denne tendens til øget nedbørsmængde ved histogrammer, sumkurver eller boksplot, kunne vi have inddelt vores interval i to (f.eks. før og efter 1940) og så f.eks. tegnet boksplot for begge disse intervaller.

Opgaverne 400*

NORMALFORDELINGER

Deskriptiv statistik har meget lidt med sandsynlighedsregning at gøre, men hvis man beskæftiger sig tilpas længe med deskriptiv statistik og får tegnet en masse pindediagrammer og histogrammer over mange forskellige ting, vil man bemærke, at man temmelig ofte får afbildninger, der kan tilnærmes med en klokkeform, der stammer fra sandsynlighedsregning.

Denne klokkeform er en *normalfordelingskurve* – også kaldet en *gausskurve* – og den er grafen for en tæthedsfunktion med funktionsforskriften:

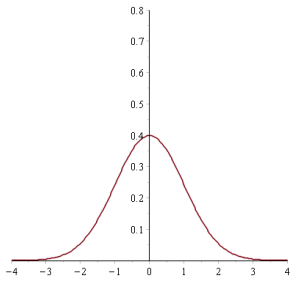
$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\cdot\sigma^2}}$$

Her er μ middelværdien, og σ er spredningen.

Eksempel 8: Normalfordelingskurver med middelværdien 0 og forskellige spredninger:

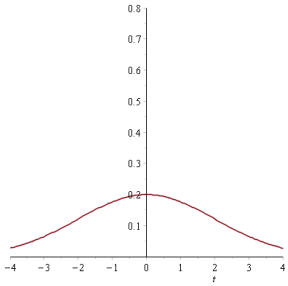
Spredningen 1

$plot(normalpdf(0, 1, t), t = -4..4)$

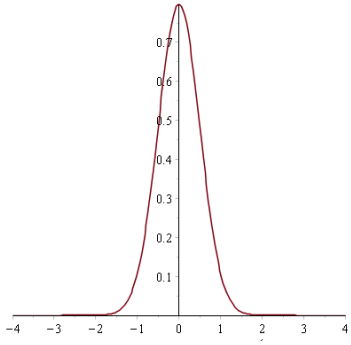


Spredningen 2

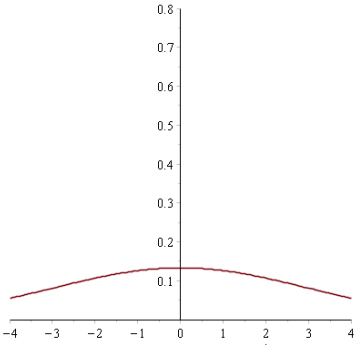
$plot(normalpdf(0, 2, t), t = -4..4)$



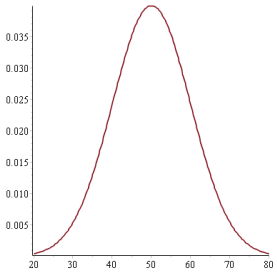
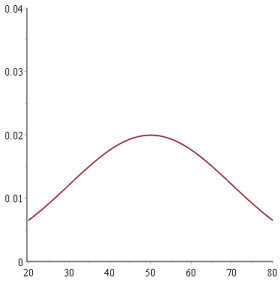
Spredningen 0,5



Spredningen 3



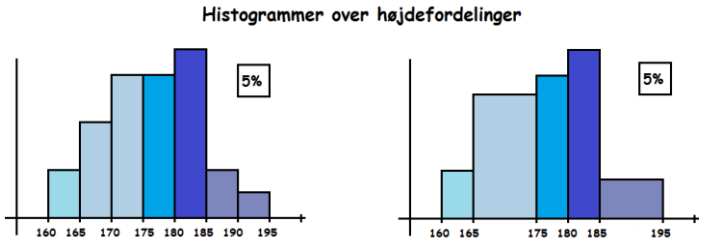
Eksempel 9: Et par normalfordelinger med middelværdien 50:

Middelværdi 50 og spredning 10	Middelværdi 50 og spredning 20
$plot(normalpdf(50,\ 10,\ t),\ t = 20\ ..80)$	$plot(normalpdf(50,\ 20,\ t),\ t = 20\ ..80,\ y = 0\ ..0.04)$
	

Om areal og enheder i histogrammer og tæthedsfunktioner

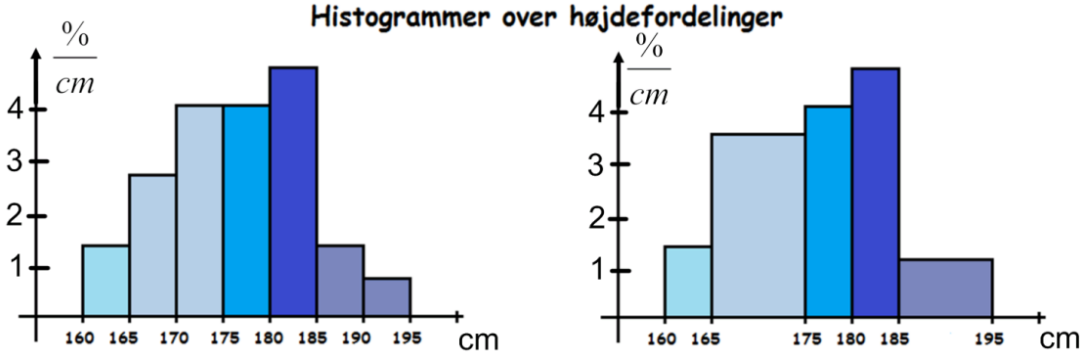
Det samlede areal under en normalfordelingskurve er 1 (100%). På samme måde har man i et histogram, at summen af alle søjlernes areal er 1 (100%). Med udgangspunkt i dette kan man ræsonnere sig frem til enheden på 2. aksen - hvis der er en 2. akse.

Det anbefales ofte ikke at angive en 2. akse, når man laver et histogram, men i stedet tegne et rektangel et sted på figuren og angive, hvor mange procent det pågældende areal svarer til:



Oftest begrundes dette med, at man vil få forkerte resultater, hvis man arbejder med forskellige intervalbredder, som det ses i figuren ovenfor til højre. Der er flere personer med højder mellem 165 og 175, end der er mellem 175 og 180, selvom søjlen for 165-175 er den laveste af de to. Det afgørende er arealerne af de to søjler. Hvis man havde haft frekvensen op ad 2. aksen, ville man altså have fået et forkert resultat.

MEN man kan slippe helt uden om dette problem ved at anvende den rigtige enhed på 2. aksen. Enheden skal være $\frac{\%}{enheden\ på 1.aksen}$, dvs. i vores tilfælde:



Hvis man anvender denne enhed, får det ingen betydning, når man slår intervaller sammen eller deler intervaller.

Aflæsninger fungerer ved, at hvis man vil vide hvor stor en procentdel af personerne, der har en højde mellem 175 cm og 180 cm, aflæser man for denne søjle værdien $4{,}1\frac{\%}{cm}$ på 2. aksen, og procentdelen udregnes så ved:

$$4{,}1\frac{\%}{\text{cm}}\cdot(180\,\text{cm}-175\,\text{cm})=4{,}1\frac{\%}{\text{cm}}\cdot 5\,\text{cm}=20{,}5\%\,.$$

Procentdelen af personer med en højde mellem 165 cm og 175 cm udregnes ud fra histogrammet til højre:

$$3{,}7\frac{\%}{\text{cm}}\cdot(175\,\text{cm}-165\,\text{cm})=3{,}7\frac{\%}{\text{cm}}\cdot 10\,\text{cm}=37\%$$

Man kan også udregne procentdelen for dele af intervaller, f.eks. 161-163 cm. Her aflæses værdien 1,5 på 2. aksen, så den samlede procentdel bliver:

$$1{,}5\frac{\%}{\text{cm}}\cdot(163\,\text{cm}-161\,\text{cm})=1{,}5\frac{\%}{\text{cm}}\cdot 2\,\text{cm}=3\%\,.$$

For normalfordelingskurver gælder præcis det samme som for histogrammer, når man skal have en enhed på 2. aksen.

Arealet under grafen er 100%, og det opnår man ved på 2. aksen at anvende enheden $\frac{\%}{enhed\ på\ 1.aksen}$, dvs. hvis man f.eks. har målt på kræfter og derfor har newton (N) ud af 1. aksen, skal enheden på 2. aksen være "pr. N", "N^{-1}"eller $\frac{\%}{N}$.

Bemærk forskellen mellem histogrammer og normalfordelingskurver.

Histogrammer er baseret på et endeligt antal intervaller, der giver anledning til et endeligt antal rektangler, der hver har et areal.

Normalfordelingskurver er baseret på differentiable tæthedsfunktioner, hvor der til ethvert argument er knyttet en funktionsværdi, men hvor man ikke har et eneste rektangel med et areal (eller også kan man løst sige, at man har uendelig mange rektangler hvert med arealet 0, men her er det centrale ord ”løst”).

Ovenfor så vi, hvordan man med histogrammer bestemmer hvilken procentdel af observationerne, der ligger i et interval.

Man kan gøre det samme med normalfordelinger, men her foregår det hele med funktionsværdier.

Vi **antager** nu, at vores nedbørsmængder fra Eksempel 7 med grupperede observationssæt var normalfordelt med middelværdien 681,6 mm og spredningen 96,9 mm (dvs. de beregnede værdier fra eksemplet). Vi stiller nu de samme spørgsmål, som vi stillede i Eksempel 7:

Spørgsmålet: *I hvor stor en del af årene faldt der over 700* mm *nedbør?*
Vi definerer først funktionen i Maple og finder derefter den del af arealet under grafen, der ligger længere ude end 700 mm:

$$f(t) := normalpdf(681.6,\ 96.9,\ t)\ :$$
$$1-\int_{-\infty}^{700} f(t)\ \mathrm{d}t = 0.4246990436$$

Ifølge normalfordelingsmodellen skulle det altså være i 42% af årene, at nedbørsmængden var over 700 mm (det rigtige tal var 41%).

Spørgsmålet: *I hvor stor en del af årene faldt der mellem 600 og 700* mm *nedbør?*

$$\int_{600}^{700} f(t)\ \mathrm{d}t = 0.3754364388$$

Dvs. normalfordelingsmodellen giver, at i 38% af årene var nedbørsmængden mellem 600 mm og 700 mm (det rigtige tal var 40%).

Nogle vigtige værdier for normalfordelinger

Man kan med Gym-pakkens *normalpdf* beregne nogle vigtige, generelle tal:

$$f(t) := normalpdf(m, s, t) :$$

$$\int_{-\infty}^{\infty} f(t)\ \mathrm{d}t \xrightarrow{\text{assuming positive}} 0.9999999995$$

$$\int_{m-s}^{m+s} f(t)\ \mathrm{d}t = 0.6826894918$$

$$fsolve\left(\int_{m-a\cdot s}^{m+a\cdot s} f(t)\ \mathrm{d}t = 0.95\right) = 1.959963989$$

$$\int_{m-2\,s}^{m+2\,s} f(t)\ \mathrm{d}t = 0.9544997356$$

$$fsolve\left(\int_{m-a\cdot s}^{m+a\cdot s} f(t)\ \mathrm{d}t = 0.99\right) = 2.575829321$$

$$\int_{m-3\,s}^{m+3\,s} f(t)\ \mathrm{d}t = 0.9973002034$$

Dvs. vi har fundet tallene i nedenstående oversigt. Tjek, at du kan se sammenhængen mellem det indtastede i Maple (ovenfor) og nedenstående tal, og prøv selv at foretage nogle af indtastningerne.

Oversigt over vigtige værdier i forbindelse med normalfordelinger:

Det samlede areal under gausskurven er 1.

Arealet under kurven i intervallet $[\mu-\sigma,\ \mu+\sigma]$ er 0,683.

Arealet under kurven i intervallet $[\mu-1{,}95996\cdot\sigma\ ;\ \mu+1{,}95996\cdot\sigma]$ er 0,95.

Arealet under kurven i intervallet $[\mu-2\sigma\ ,\ \mu+2\sigma]$ er 0,954.

Arealet under kurven i intervallet $[\mu-2{,}57583\cdot\sigma\ ;\ \mu+2{,}57583\cdot\sigma]$ er 0,99.

Arealet under kurven i intervallet $[\mu-3\sigma\ ;\ \mu+3\sigma]$ er 0,9973.

Dvs. at 68,3% af observationerne i et normalfordelt observationssæt ligger inden for én standardafvigelse af middelværdien.

Tilsvarende ligger 95,4% af observationerne inden for to standardafvigelser (og disse betegnes som **normale udfald**), mens 99,73% ligger inden for tre standardafvigelser.

De 0,27% af udfaldene, der ligger mere end tre standardafvigelser fra middelværdien, kaldes **exceptionelle udfald.**

5% ligger mere end 1,96 standardafvigelser fra middelværdien.

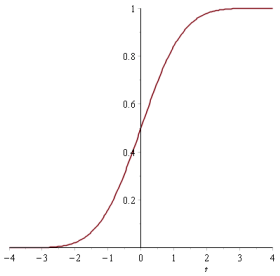
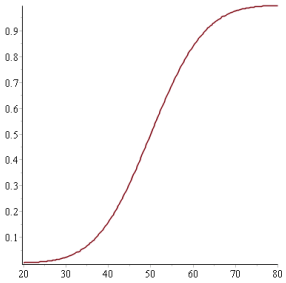
1% ligger mere end 2,58 standardafvigelser fra middelværdien.

Fordelingsfunktioner:

Som vi ved fra sandsynlighedsregning, kaldes arealfunktionen af en tæthedsfunktion for en *fordelingsfunktion*. Dvs. en fordelingsfunktion kumulerer sandsynlighederne:

$$\Phi(x) = \int_{-\infty}^{x} f(t)\,dt$$

Et par eksempler på fordelingsfunktioner for normalfordelinger kan tegnes med Gym-kommandoen *normalcdf*:

Middelværdi 0 og spredning 1	Middelværdi 50 og spredning 10
$plot(normalcdf(0, 1, t), t = -4..4)$	$plot(normalcdf(50, 10, t), t = 20..80)$
	

Eftersom fordelingsfunktionen angiver arealet under grafen for tæthedsfunktionen i intervallet $]-\infty, x]$, kan man besvare vores nu velkendte spørgsmål fra Eksempel 7 på følgende måde:

Spørgsmålet: *I hvor stor en del af årene faldt der over 700* mm *nedbør?*

$$f(t) := normalcdf(681.6, 96.9, t):$$
$$1 - f(700) = 0.4246990432$$

Spørgsmålet: *I hvor stor en del af årene faldt der mellem 600 og 700* mm *nedbør?*

$$f(700) - f(600) = 0.3754364390$$

Øvelse 1: Benyt Maple til at bestemme nogle af nedenstående udvalgte værdier.

$\Phi(\mu - 3\sigma) = 0,00135$	$\Phi(\mu + \sigma) = 0,841$
$\Phi(\mu - 2\sigma) = 0,023$	$\Phi(\mu + 2\sigma) = 0,977$
$\Phi(\mu - \sigma) = 0,159$	$\Phi(\mu + 3\sigma) = 0,99865$
$\Phi(\mu) = 0,5$	$\Phi(x) \to 1$ for $x \to \infty$

Opgaverne 401*

Den Centrale Grænseværdisætning

Normalfordelinger er helt centrale inden for statistiske test. Alle de test, vi snart skal beskæftige os med, er baseret på en antagelse om, at de undersøgte størrelser (tilnærmelsesvis) er normalfordelte. Det skyldes *Den Centrale Grænseværdisætning*, der i en meget kort, upræcis og **ikke korrekt** version lyder:

Alt er tilnærmelsesvis normalfordelt.

Ordet 'central' henviser til sætningens vigtighed inden for sandsynlighedsregning og statistik.

Som sagt er ovennævnte ordlyd ikke korrekt, men den giver et meget godt billede af normalfordelingens betydning. Den Centrale Grænseværdisætning findes i mange forskellige – mere eller mindre stærke – versioner.

Den Centrale Grænseværdisætning: Lad $S_n = \frac{X_1 + X_2 + ... + X_n}{n}$ være en stokastisk variabel, der angiver det aritmetiske gennemsnit for *n* uafhængige og identisk fordelte stokastiske variable X_i ($i = 1, 2, 3, ..., n$) med endelig middelværdi μ og spredning σ.

Så vil S_n tilnærmelsesvis være normalfordelt med middelværdien μ og spredning $\frac{\sigma}{\sqrt{n}}$, når *n* er tilpas stor, og tilnærmelsen bliver bedre, jo større *n* er.

Betingelsen 'uafhængige og identisk fordelte' anvendes så ofte inden for statistik, at man har forkortet den til *i.i.d.* (eller *iid* eller *IID*) - *independent and identically distributed*.

Eksempel 10: Vi ved fra tidligere, at hvis man måler baggrundsstrålingen inden for 10 sekunder, vil den stokastiske variabel *X*, der angiver tælletallet, være poissonfordelt. Poissonfordelingen med middelværdien λ har spredningen $\sqrt{\lambda}$, og vores stokastiske variabel *X* opfylder altså betingelsen om endelig middelværdi og spredning.

Se nu på den stokastiske variabel $S_{100} = \frac{X_1 + X_2 + ... + X_{100}}{100}$, der svarer til, at man 100 gange har målt baggrundsstrålingen i 10 sekunder og taget gennemsnittet af tælletallene.

Denne stokastiske variabel vil tilnærmelsesvis være normalfordelt med middelværdien λ og spredningen $\frac{\sqrt{\lambda}}{\sqrt{100}} = \frac{\sqrt{\lambda}}{10}$. Dvs. at hvis du masser af gange 100 gange måler baggrundsstrålingen i 10 sekunder (det bliver til rigtig mange målinger) og laver et histogram baseret på de målte værdier for S_{100}, så vil dette histogram danne normalfordelingernes klokkeform.

Der er flere væsentlige ting at bemærke her:

- **Selve tælletallet** er poissonfordelt, men hvis du måler tilpas mange gange og tager gennemsnittet af målingerne, så vil **dette gennemsnit** tilnærmelsesvis være normalfordelt. Og dette gælder ikke bare for poissonfordelinger. Det gælder for alle fordelinger med endelig middelværdi og spredning.
- Spredningen for den pågældende normalfordeling bliver mindre, jo flere målinger *n*, der foretages, da spredningen er $\frac{\sigma}{\sqrt{n}}$. Dvs. at sandsynligheden for at ramme inden for et fastsat interval omkring middelværdien bliver større, jo flere målinger, der foretages.
- Det er meget vigtigt at skelne mellem observationssættets spredning (den ændrer sig ikke - bortset fra tilfældige udsving - uanset hvor mange observationer, der er i sættet) og spredningen på middelværdien (den bliver mindre, jo flere observationer, der er i sættet).
- Vi har ovenfor taget Den Centrale Grænseværdisætning i sin "skrappeste" version. Vi har f.eks. forlangt, at de stokastiske variable X_i skulle være *i.i.d.*. Faktisk er det ikke altid nødvendigt. Sætningen gælder som udgangspunkt også, hvis de blot er uafhængige.
- Og faktisk gælder den også for visse tilfælde med uendelige spredninger. Sandsynlighederne skal bare falde "hurtigt nok".

Opgaverne 402*

Binomialformlen vs. normalfordelingskurven

Binomialfordelingen er som set under forløbet om sandsynlighedsregning og kombinatorik en sandsynlighedsfordeling, hvor sandsynligheden for at få *r* succeser er:

$$p\left(X=r\right)=K\left(n,r\right)\cdot p^{r}\cdot\left(1-p\right)^{n-r}$$

, hvor *n* er antalsparameteren og *p* successandsynligheden for én udførelse af det pågældende eksperiment.

Middelværdien og spredningen er:

$$\mu=E(X)=n\cdot p$$

$$\sigma=\sigma(X)=\sqrt{n\cdot p\cdot\left(1-p\right)}$$

Dvs. binomialfordelingen har veldefineret middelværdi og spredning. Derfor ved vi fra Den Centrale Grænseværdisætning, at hvis man udfører tilpas mange forsøg og tager gennemsnittet af disse, vil dette gennemsnit tilnærmelsesvis følge en normalfordeling. Naturligvis kan det aldrig helt blive en normalfordeling, for normalfordelingen er en kontinuert fordeling, og vores gennemsnit kan – uanset hvor mange gange *n* forsøget udføres – kun give rationale værdier. Men ligesom et irrationalt tal kan være grænseværdien for en følge af rationale tal, så kan en kontinuert fordeling også være grænsen for en følge af diskrete fordelinger. Tænk på følgende:

Når man som beskrevet i Den Centrale Grænseværdisætning tager en fordeling og danner sit gennemsnit ved at udføre sit eksperiment *n* gange, kan man som nævnt ovenfor få brøker som resultat, selvom den bagvedliggende fordeling kun kan give hele tal (hvilket f.eks. er tilfældet for bl.a. binomialfordelingen, poissonfordelingen, den negative binomialfordeling, den hypergeometriske fordeling og den negative hypergeometriske fordeling). Og de mulige brøker bliver flere og flere, jo større *n* er, så man så at sige "udfylder" områderne mellem de hele tal med flere og flere mulige værdier. På den måde kan man godt se, at det giver mening, at en diskret fordeling kan nærme sig den kontinuerte normalfordeling.

Men binomialfordelingen er i sig selv sådan set "bare" et opskaleret gennemsnit af en række binomialeksperimenter. Man lægger antal succeser fra hvert enkelt binomialeksperiment sammen, men undlader at dividere med antal binomialeksperimenter (*n*) og får dermed $n\cdot S_n$ (jf. Den Centrale Grænseværdisætning).

Derfor vil binomialfordelingen i sig selv også nærme sig normalfordelingen, når bare *n* bliver tilpas stor (og for mindre *n* værdier: Jo tættere *p* er på 0,5).

Umiddelbart kan det lyde helt forkert at prøve at sammenligne en diskret fordeling, der kun kan give heltallige værdier, med den kontinuerte normalfordeling. Og det bliver endnu værre, når man bemærker, at udfaldene i binomialfordelingen er begrænset i begge ender ($0\leq r\leq n$), mens normalfordelingens tæthedsfunktion har hele $\mathbb{R}$ som definitionsmængde.

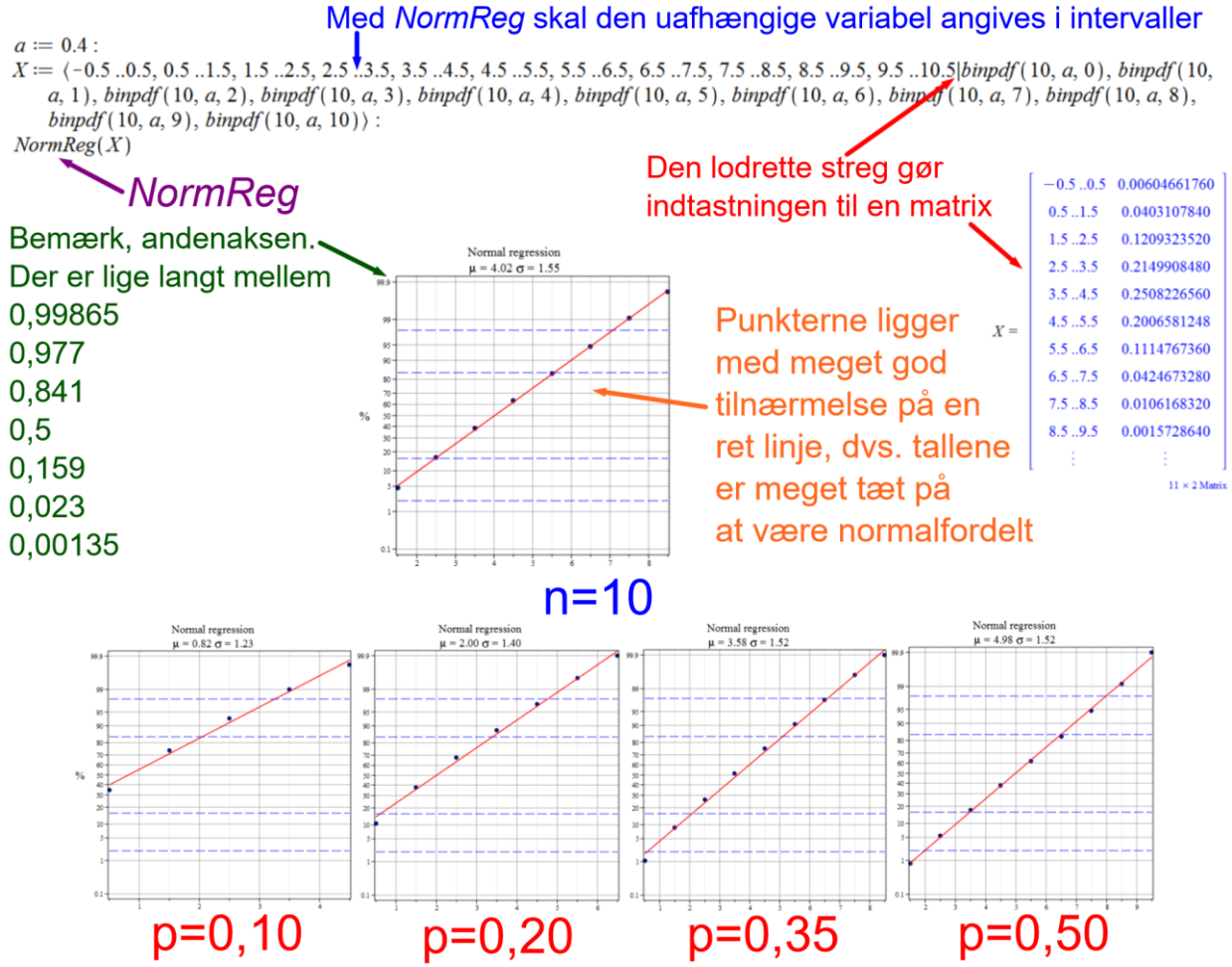
Hvis man imidlertid vælger at se bort fra disse grundlæggende forskelle og sammenligner binomialfordelingen med den normalfordeling, der er fremkommet ved at anvende middelværdi og spredning fra binomialfordelingen, så gælder det, at binomialfordelingen tilnærmelsesvis følger normalfordelingen forstået på den måde, at pindediagrammet over binomialfordelingen tilnærmelsesvis får normalfordelingens karakteristiske klokkeform.

Tilnærmelsen bliver som sagt bedre, jo større *n* er, og jo tættere successandsynligheden *p* er på 0,5.

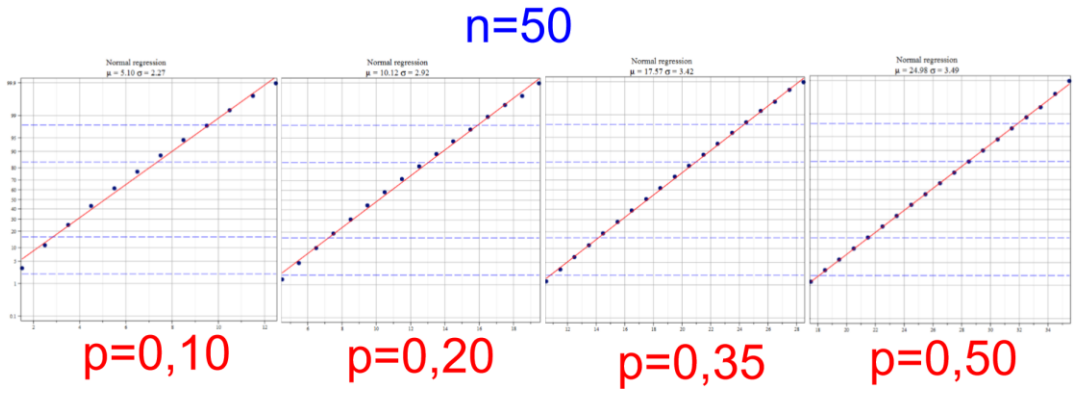
Se bilag A for eksempler.

Man kan undersøge, om et datamateriale kan beskrives ved en lineær model, ved at afbilde det i et almindeligt koordinatsystem og se, om punkterne tilnærmelsesvis danner en ret linje. Og ved at afbilde data i et enkeltlogaritmisk koordinatsystem og se, om punkterne danner en ret linje, kan man vurdere, om man kan anvende en eksponentiel udvikling som model. Potensfunktioner genkendes som rette linjer i et dobbeltlogaritmiske koordinatsystem. Man har også lavet normalfordelingspapir med en specielt konstrueret ordinatakse, der anvendes til at tjekke, om en fordeling kan tilnærmes med en normalfordeling. Man plotter sumkurven (grupperet) og ser, om punkterne tilnærmelsesvis ligger på en ret linje. Binomialfordelingen er jo ikke grupperet, men man anvender så intervaller med bredden 1 symmetrisk omkring stedet, dvs. i stedet for f.eks. 7 anvendes [6.5,7.5].

Dette normalfordelingspapir fremkommer i Maple, når man anvender Gym-pakkens *NormReg*.



Bemærk , at punkterne med bedre og bedre tilnærmelse ligger på en ret linje, når succes-sandsynligheden kommer tættere på 0,5. De største afvigelser ses i enderne. Samme mønster ses nedenfor, hvor tilnærmelserne bare er bedre, da *n* er større.



Er mit konkrete eksperimentelle datasæt normalfordelt?

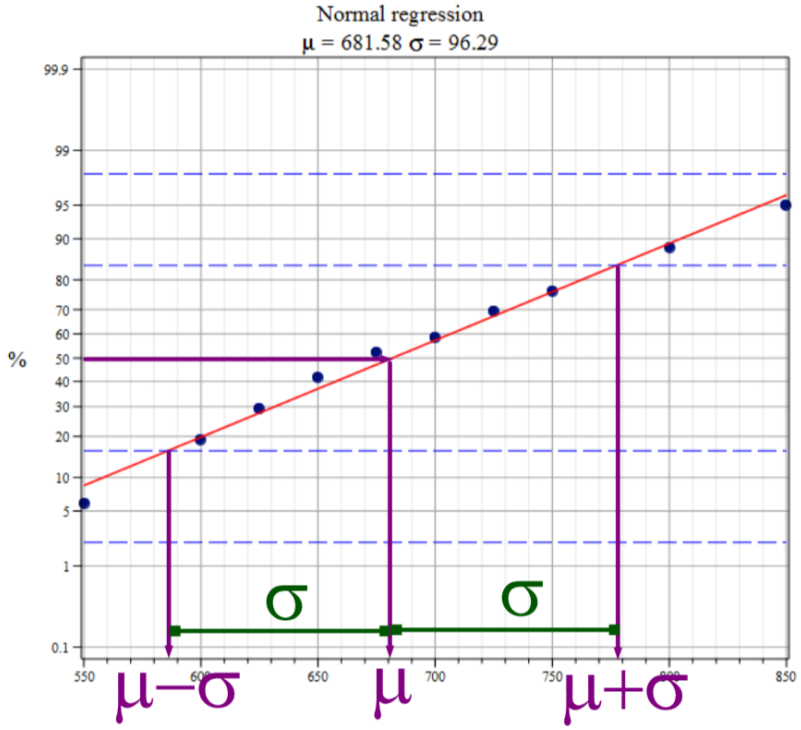
Vi har lige set, hvordan man med *NormReg* kan se, at binomialfordelingen med god tilnærmelse følger en normalfordeling, og at tilnærmelsen er bedre, jo større *n* er, og jo tættere *p* er på 0,5. Det var lidt kunstigt, da vi var nødt til at danne intervaller omkring de enkelte steder, men det slipper vi jo for, når vi ser på eksperimentelle datasæt, der helt naturligt skal grupperes.

Vi ser på datasættet fra Eksempel 7 (Årlig nedbørsmængde i Danmark):

$$N = \begin{bmatrix} 450..550 & 8 \\ 550..600 & 18 \\ 600..625 & 14 \\ 625..650 & 17 \\ 650..675 & 14 \\ 675..700 & 9 \\ 700..725 & 14 \\ 725..750 & 10 \\ 750..800 & 16 \\ 800..850 & 9 \\ \vdots & \vdots \end{bmatrix}$$

11 × 2 Matrix

$NormReg(N)$



Punkterne ligger nogenlunde på en ret linje. Dvs. nedbørsmængden kan med rimelig tilnærmelse beskrives som normalfordelt (jf. histogrammet i Eksempel 7, der nogenlunde følger klokkeformen). Egentlig burde man forvente en bedre tilnærmelse, for nedbørsmængden er en størrelse, der burde være normalfordelt, og datasættet er stort, men som tidligere omtalt ses en systematisk forøgelse af nedbørsmængden over tid, hvilket slører billedet.

De stiplede blå linjer markerer procenterne for en og to spredninger omkring middelværdien i normalfordelingen, og de kan sammen med en vandret linje fra 50% som vist ovenfor anvendes til at aflæse middelværdi og spredning (hvilket selvfølgelig er overflødigt, når de allerede er udregnet, men hvis man kun har en graf, er det fremgangsmåden).

QQ-plot

Man kan dog også undersøge, om et datamateriale er normalfordelt, uden først at gruppere det. Det gøres med et såkaldt QQ-plot (fraktilplot). Gym-kommandoen er *QQplot*.

Igen er ideen, at man skal se på, om punkterne tilnærmelsesvis ligger på en ret linje. Hvis de gør det, er datamaterialet normalfordelt.

Et QQ-plot fremkommer ved, at punkternes førstekoordinat er selve måletallet fra datasættet, mens andenkoordinaten er den tilsvarende fraktil i standardnormalfordelingen (alias *u*-fordelingen alias normalfordelingen med middelværdi 0 og spredning 1).

Fremgangsmåden er (næsten) følgende:

Antag, at man har en måleserie på 200 målinger, der er stillet op i ordnet rækkefølge, og at tal nummer 60 i denne række er 1729.

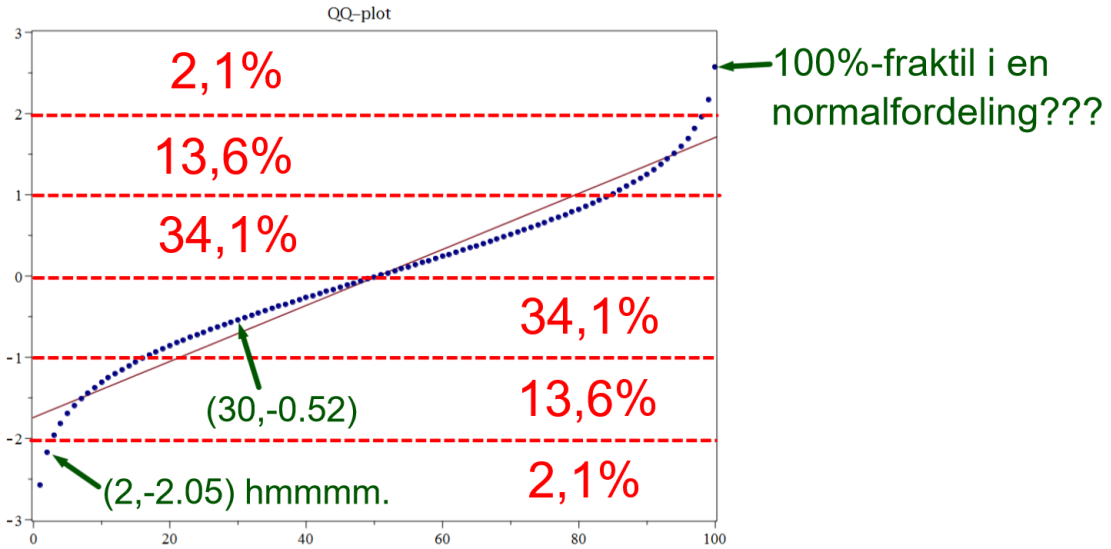
- Tallet 1729 er så 30%-fraktilen i datasættet ($\frac{60}{200} = 0,30 = 30\%$). 1729 er førstekoordinaten i det punkt, vi skal have placeret i koordinatsystemet.
- 30%-fraktilen i *u*-fordelingen er -0,52 (se udregningen med Maple nedenfor)

$$fsolve(normalcdf(0,\ 1,\ b) = 0.30,\ b) = -0.5244005128$$

- Dvs. -0,52 er andenkoordinaten for punktet, der altså er (1729,-0.52).

I nedenstående eksempel ses på tallene 1, 2, 3, 4, 5, 6, …, 100. Dette datasæt er oplagt ikke normalfordelt, da tallene er fuldstændig jævnt fordelt. Pindediagrammet ville give 100 lige høje pinde (højden 1), dvs. ikke nogen klokkeform. Og i koordinatsystemet nedenfor ses det også, at punkterne afviger systematisk fra den rette linje:

$QQplot(A,\ view = [0\ ..100, -3\ ..3])$



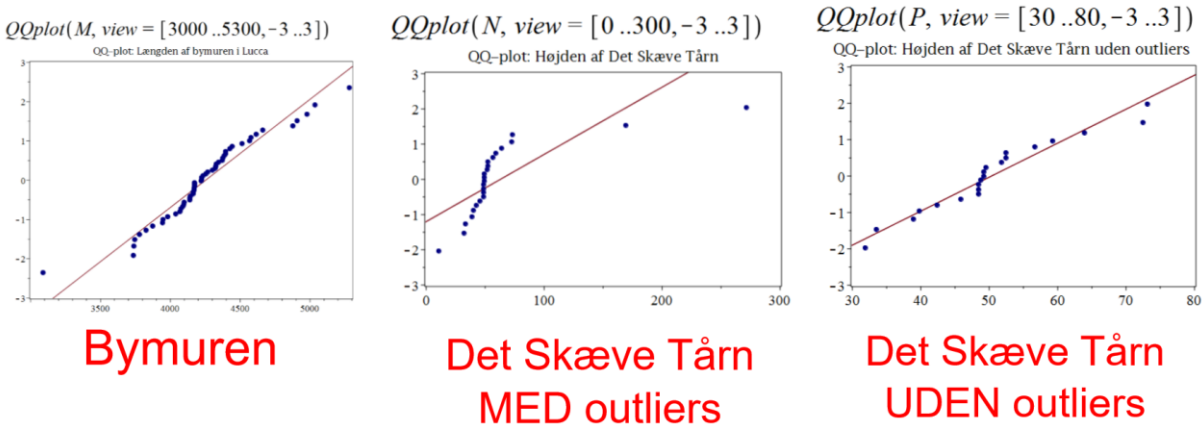
De stiplede røde linjer opdeler koordinatsystemet med udgangspunkt i *u*-fordelingen. Så vi kan regne ud, hvor mange punkter der vil være i hver del:

$$normalcdf(0,\ 1, -2) - normalcdf(0,\ 1, -3) = 0.0214002340$$
$$normalcdf(0,\ 1, -1) - normalcdf(0,\ 1, -2) = 0.1359051220$$
$$normalcdf(0,\ 1,\ 0) - normalcdf(0,\ 1, -1) = 0.3413447460$$

Der vil altså være 2 punkter i den nederste del, 14 i den næste (tæl selv efter) og 34 i området lige under midten. Bemærk, at disse procentdele altid gælder for områderne opdelt på denne måde. Man kan kun lave en tilsvarende opdeling med lodrette, ækvidistante linjer, HVIS datasættet er normalfordelt. Man skal så bruge middelværdi og spredning til opdelingen.

Vi kan også se på de enkelte punkter. Tallet 30 er 30%-fraktilen, og vi ved fra før, at i *u*-fordelingen er 30%-fraktilen -0,52. Derfor har man punktet (30,-0.52). Og dog… kun næsten. For der er lige det problem, at man jo ikke har en 100%-fraktil i normalfordelinger, da de er ubegrænsede. Man er derfor nødt til at lave nogle tricks for at få alle tal med, og det rykker alting en lille smule. Se f.eks. det punkt, der skulle være (2,-2.05). Jeg ved ikke, hvordan det konkret er gjort i Gym-pakken. Der er forskellige metoder, og ingen af dem er "den rigtige", for det er et problem, der ikke **kan** løses rigtigt, da normalfordelingen **er** ubegrænset. Lige som man ikke kan stemme et klaver "rigtig", men f.eks. "veltempereret", "ligesvævende" eller "pythagoræisk".

Eksempel 11: Vi ser på vores velkendte bestemmelser af højden af Det Skæve Tårn i Pisa og længden af bymuren i Lucca:



Bymuren | Det Skæve Tårn MED outliers | Det Skæve Tårn UDEN outliers

Bymuren: Der er nogle systematiske afvigelser, så der er kun antydninger af en normalfordeling.

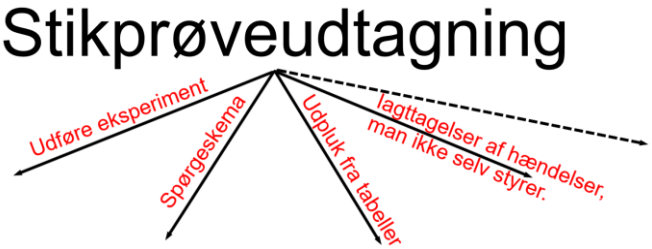
Det Skæve Tårn: Her ses det, at man for det fulde datasæt får noget, der absolut ikke er normalfordelt. Outlierne ødelægger fuldstændig billedet.

Når de tre outliers er fjernet, ser man, at punkterne med god tilnærmelse ligger på en ret linje (afvigelserne er ikke systematiske), så disse målinger er normalfordelte. Man kan her se betydningen af at fjerne outliers, HVIS der er belæg for at betragte dem som fejlmålinger.

Opgaverne 403*

STIKPRØVEUDTAGNING OG EKSPERIMENTELT ARBEJDE

Inden for naturvidenskaberne indsamler man typisk data ved at måle eller iagttage nogle størrelser i opstillede forsøg eller – f.eks. inden for astronomien - i forbindelse med hændelser, man ikke selv kontrollerer. Andre eksempler på indsamlinger kunne være spørgeskemaundersøgelser, interviews eller opslag i et tabelværk, hvis data allerede er indsamlet.



Hovedtanken bag stikprøveudtagning og målinger og beregninger på denne er, at man har et begreb eller en mængde af størrelser, som man gerne med en vis nøjagtighed vil kunne tilskrive en værdi eller en række egenskaber, og det vil man gøre ved at slutte induktivt fra stikprøven (*induktivt* forstået i den version af ordet, at man slutter fra det specielle til det generelle).

I forbindelse med bestemmelsen af en fysisk størrelse er et af problemerne måleusikkerheder, der gør det umuligt at bestemme en værdi præcist. Og generelt gælder om fysiske love, at de skal gælde til alle tider og alle steder og i alle forbindelser, hvilket man selvsagt ikke kan måle. Man foretager derfor et begrænset antal målinger (svarende til at udtage en stikprøve), og ud fra disse målinger slutter man sig til noget angående den fysiske størrelse eller den fysiske lov. Bemærk, at dette er en induktiv slutning. Både forstået som *en slutning fra det specielle til det generelle* og som at *forsøgene yder støtte, men ikke sikkerhed, for konklusionen*. Der er altså ikke noget logisk gyldigt i selve slutningen, og man ender også "kun" med at kunne angive resultatet med en vis usikkerhed angivet med anvendelsen af sandsynligheder.

Der er altså to grundlæggende vilkår, vi ikke kan komme uden om:

1) Vores slutninger fra stikprøve til population er induktiv og derfor ikke logisk gyldig.
2) Vi må anvende sandsynligheder i angivelsen af vores konklusion.

Oversigt over og kort forklaring på centrale begreber

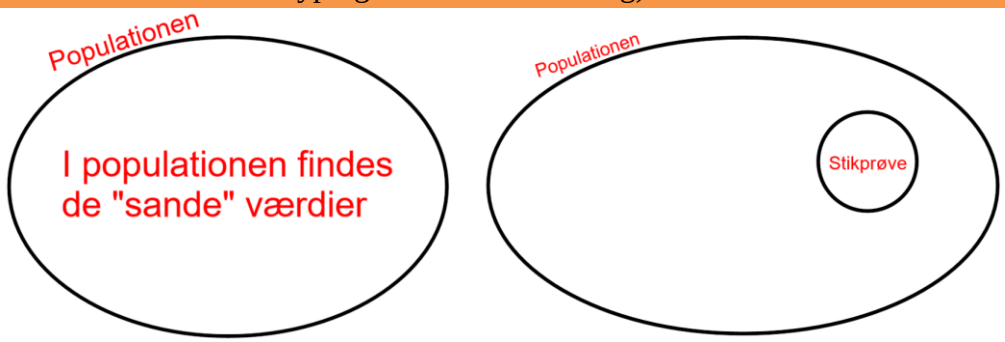
Population: Den mængde af størrelser, om hvilken man ønsker at kunne drage nogle konklusioner.

Eksempler på mulige populationer:

a) Danske gymnasieelever.
b) Verdens befolkning.
c) Beviser for matematiske sætninger i danske undervisningsbøger gennem tiderne.
d) 3 mm skruer fra firmaet Jernmand.
e) Lysets hastighed som den optræder i alle sammenhænge i naturen.
f) Newtons 2. lov til beskrivelse af enhver kraftpåvirkning af ethvert legeme.

Stikprøve: En stikprøve er en delmængde af en population.

Det er ud fra stikprøven, at der skal kunne drages konklusioner vedrørende populationen. Styrken af konklusionerne vil vokse, når stikprøvestørrelsen øges, men den vil vokse langsommere og langsommere, så der efterhånden skal en stor forøgelse af stikprøvestørrelsen til at give en lille forøgelse af styrken. Og hvis stikprøven udtages med henblik på hypotesetest, skal man f.eks. ved en binomialtest sikre sig, at stikprøvens størrelse er så meget mindre end populationen, at man kan se bort fra, at man ikke arbejder **med tilbagelægning**, der er karakteristisk for binomialfordeling (og ellers må man teste med den hypergeometriske fordeling).



Eksempler på stikprøver (i tilknytning til ovenstående populationer):

a) 100 elever fra københavnske gymnasier, 100 elever fra århusianske gymnasier og 100 elever fra fynske gymnasier (vi skal senere se, at dette er en dårlig stikprøve, bl.a. fordi der ikke går lige mange elever de tre steder, hvorfor antallene i stikprøven heller ikke skal være de samme).
b) 5 tilfældigt valgte kvinder fra hvert land i verden.
c) Samtlige matematiske beviser i 4 tilfældigt udvalgte matematikbøger på et bibliotek.
d) 100 tilfældigt udvalgte skruer fra firmaet Jernmand.
e) Bestemmelsen af lysets hastighed ved gentagelse af det samme forsøg 20 gange.
f) Måling af accelerationen af 10 forskellige genstande udsat for hver 5 forskellige kraftpåvirkninger.

Opgaverne 410*

Bias: En stikprøve siges at være *biased*, hvis nogle af elementerne i populationen har haft mindre sandsynlighed for at komme med i stikprøven end andre (jf. eksemplerne a og b ovenfor).

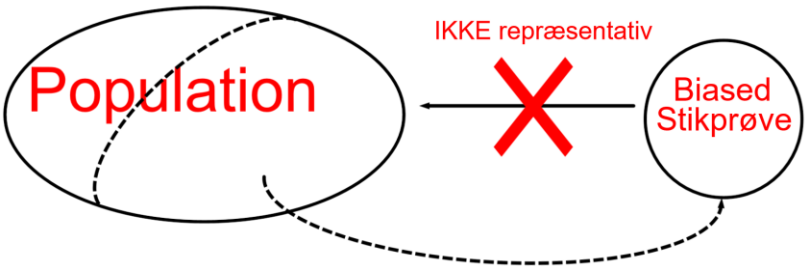
Bias betyder 'skævhed'. En biased stikprøve er **ikke** *repræsentativ* for populationen, dvs. den kan ikke bruges til at estimere de statistiske deskriptorer for populationen.

1) En forsker ønsker at undersøge vægten på skovmus og fanger dem i fælder ved at lokke dem med mad. Stikprøven bliver biased, fordi skovmusene skal overvinde frygten for fælden for at gå i den, hvilket er mere sandsynligt for en sulten mus, der altså som udgangspunkt er tyndere end en 'gennemsnitsmus'.

2) Den danske befolknings tv-forbrug undersøges ved en telefonundersøgelse med fastnetnumre (det er et forældet eksempel). Stikprøven bliver biased, fordi mennesker, der oftere er hjemme og kan tage telefonen, nok ser mere fjernsyn end dem, der sjældnere er hjemme (og derfor "har haft mindre sandsynlighed for at komme med i stikprøven end andre").

3) Alle undersøgelser, hvor personer selv kan vælge at deltage, bliver biased, da motivationen for at deltage på en eller anden måde kan hænge sammen med svarene.

4) I 1936 indsamlede *The Literary Digest* mere end 2 millioner tilkendegivelser fra bladets læsere samt bil- og/eller telefon-ejere om, hvem de ville stemme på, og kom frem til en kæmpe sejr til republikaneren Alf Landon over demokraten Franklin Roosevelt. De havde spurgt 10 millioner. Stikprøven var sandsynligvis biased, fordi alle tre grupper af deltagere var mere velhavende end formuemedianen, eller pga. en eller anden motivationsfaktor (det er vist stadig ikke helt afklaret).



Korrelation: For at forstå beskrivelsen af næste begreb skal man forstå udtrykket *at korrelere*.

Kort sagt siges to størrelser (variable) at korrelere, hvis de (gensidigt) afhænger af hinanden.

Lidt længere sagt korrelerer to variable, der observeres i par, hvis man, når man kigger på de par, hvor den første variabel er større end sin gennemsnitsværdi oftest også har, at den anden variabel er større end sin gennemsnitsværdi (positiv korrelation) eller mindre end sin gennemsnitsværdi (negativ korrelation).

En noget længere (og helt præcis) formulering kræver formler.

I tilfælde 1) med skovmusene er variablen *vægt* negativt korreleret med variablen *sult*, hvis det oftest er sådan, at når man har en mus, der vejer mindre end gennemsnittet, så er den mere sulten end gennemsnittet.
I samme eksempel er de to variable *sult* og *vovemod* positivt korrelerede, hvis det er sådan, at en mus, der er mere sulten end gennemsnittet også vover mere end gennemsnittet.

Skjulte variable: En skjult variabel er en uvedkommende variabel, der korrelerer – enten positivt eller negativt – med både den uafhængige og den afhængige variabel.
Begreberne uafhængig og afhængig variabel skal i denne sammenhæng forstås som følgende eksempler viser:

1) Skovmusene: Her er den uafhængige variabel sandsynligheden for at fange den enkelte mus, mens vægten af den enkelte mus er den afhængige variabel. Den skjulte variabel er 'sult', da den korrelerer positivt med sandsynligheden for at musen fanges (jo mere sult, jo større sandsynlighed) og korrelerer negativt med musens vægt (jo mere sult, des mindre vægt).

2) Undersøgelse af tv-forbrug: Her er den uafhængige variabel sandsynligheden for at få fat på en person (dvs. at få foretaget telefoninterviewet), mens den afhængige variabel er tv-forbruget. Den skjulte variabel er så 'ophold i hjemmet', fordi den korrelerer positivt med både sandsynligheden for at få fat på folk (jo mere tid der bruges i hjemmet, jo større er chancen for at være hjemme når fastnettelefonen ringer) og med tv-forbruget (jo mere man er i hjemmet, jo mere fjernsyn vil man som oftest se).

4) Meningsmålingen: Her er den uafhængige variabel sandsynligheden for at få en tilkendegivelse fra en stemmeberettiget amerikaner, mens den afhængige variabel kunne være sandsynligheden for at stemme på Franklin Roosevelt. Her er den skjulte variabel 'højere formue eller indkomst end medianen', da den korrelerede positivt med sandsynligheden for at tilkendegive sin stemme (fordi flere penge øger sandsynligheden for at holde et blad eller eje telefon eller bil) og korrelerede negativt med det at stemme på Franklin Roosevelt (som var mere venstreorienteret end Alf Landon).
(Hvis "Franklin Roosevelt" var erstattet med "Alf Landon", havde korrelationen været positiv.)

Systematiske fejl: Afvigelserne mellem de statistiske deskriptorer i en model opstillet ud fra en biased stikprøve og de statistiske deskriptorer for selve populationen.

Da der grundet usikkerheder som udgangspunkt altid vil være afvigelser mellem en model og virkeligheden, kræver overstående korte formulering en uddybning.

Hvis man forestiller sig, at man kan tage stikprøver, der ikke er biased, fra en population, vil værdierne for deskriptorerne for de enkelte stikprøver godt nok afvige fra populationens sande værdier, men ved at tage gennemsnittet af stikprøverne og øge disses antal, vil man komme tættere og tættere på de sande deskriptorer.

Dette er ikke tilfældet, hvis de enkelte stikprøver er biased. Så vil gennemsnitsværdierne for de enkelte deskriptorer ikke nærme sig de sande værdier, når antallet af stikprøver øges, men derimod nogle bestemte 'falske' værdier for deskriptorerne.

De systematiske fejl er altså i princippet afvigelserne mellem de sande værdier og ovennævnte falske værdier, der fås ved at tage gennemsnittet af 'uendelig' mange biased stikprøver (hvor det er underforstået, at stikprøverne udtages på samme måde hver gang, så der altså er tale om den samme form for bias).

En kvalitativ beskrivelse af de systematiske fejl i vores eksempler er:

1) Her bliver den systematiske fejl, at undersøgelsen viser en for lille gennemsnitsvægt af musene, fordi den skjulte variabel 'sult' korrelerer positivt med sandsynligheden for at musen fanges og negativt med musens vægt.
2) Her bliver den systematiske fejl, at undersøgelsen viser et for stort tv-forbrug, fordi den skjulte variabel "ophold i hjemmet" korrelerer positivt med både sandsynligheden for at få fat på folk og med tv-forbruget.
4) Her er den systematiske fejl, at undersøgelsen viser en for lille vælgertilslutning til Roosevelt, fordi den skjulte variabel "større formue eller indkomst" korrelerede positivt med sandsynligheden for at tilkendegive sin stemme og negativt med det at stemme på Roosevelt.

Stratifikation: Inddeling af populationen i disjunkte delmængder (dvs. at alle elementer i populationen placeres i en og kun en delmængde).

Stratifikation betyder laginddeling (strata ~ lag). For at sikre sig mod, at en stikprøve bliver biased, kan man inddele population i nogle dele baseret på en forhåndsvurdering eller efter at have noteret sig en skævhed i stikprøven (poststratifikation). Ved en vælgerundersøgelse kan man f.eks. vurdere, at *mænd* og *kvinder* stemmer forskelligt, *unge*, *midaldrende* og *ældre* stemmer forskelligt og *underklassen*, *middelklassen* og *overklassen* stemmer forskelligt. Man skal så opdele i $2 \cdot 3 \cdot 3 = 18$ forskellige disjunkte delmængder.

Stratifikationsregel: Hver delmængde skal være repræsenteret med samme procentdel i stikprøven som i populationen.

En opsummerende sætning: Hvis stikprøveudtagningen indeholder en skjult variabel, bliver stikprøven biased, og dermed indeholder undersøgelsen en systematisk fejl.

Begreberne anvendt inden for naturvidenskaberne

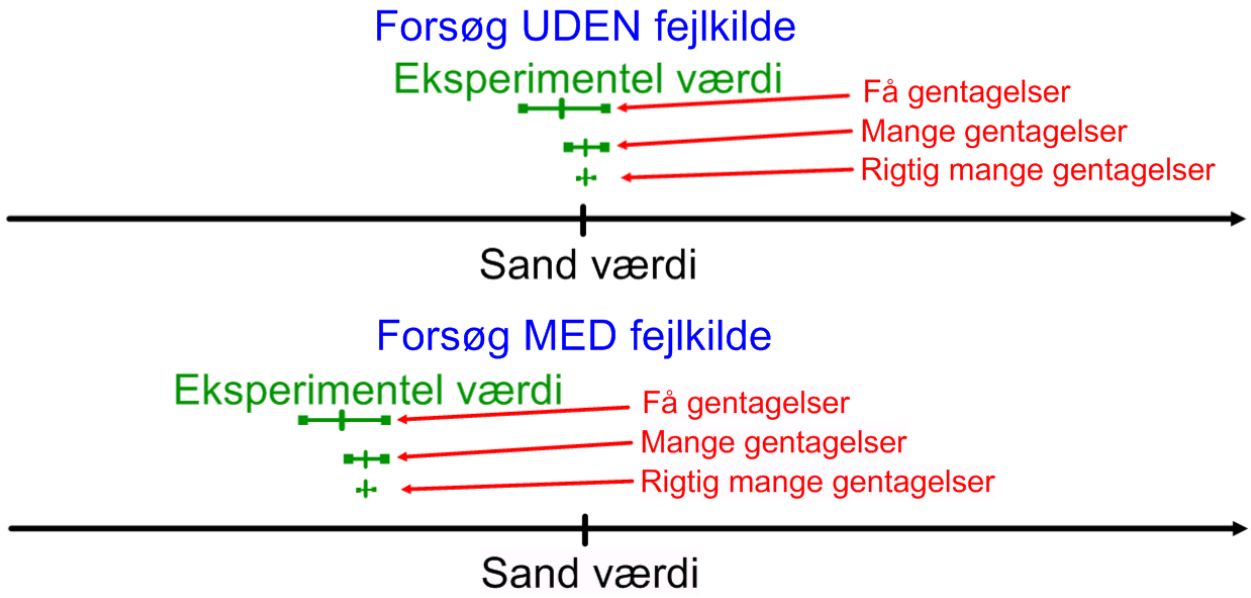
Skjulte variable - fejlkilder: Begrebet *skjult variabel* kaldes inden for naturvidenskaberne for *en fejlkilde*. Fejlkilder vil påvirke forsøgsresultaterne, så der ikke kommer overensstemmelse mellem teorien og eksperimentet, uanset hvor mange gange man udfører forsøget.

Dette kan illustreres med følgende figur:

Vi skal måle på en fysisk størrelse og antager, at der er en "sand" værdi af denne (angivet med den lille lodrette sorte streg), som vi skal bestemme. Vi udfører forsøget 8 gange, og som de røde cirkler angiver, får vi forskellige resultater hver gang, men de fordeler sig nogenlunde ligeligt på hver side af den sande værdi. Vores eksperimentelle værdi er angivet med usikkerheder, dvs. vi angiver et interval, hvor midten er vores eksperimentelle værdi, og som vi hævder, at den sande værdi med en vis sandsynlighed skal ligge inden for (hvilket den gør på vores illustration).

Men hvis der er en fejlkilde i forsøget, vil vores eksperimentelle værdier ikke fordele sig nogenlunde ligeligt omkring den sande værdi. Alle eller en klar overvægt af værdier vil placere sig på den ene side, og den sande værdi vil derfor ikke ligge inden for det interval, der er bestemt eksperimentelt.

Når et forsøg udføres mange gange, svarer det til at øge stikprøvens størrelse, og det vil mindske de relative usikkerheder, men det vil ikke hjælpe os til at komme nærmere den sande værdi, hvis der er fejlkilder i eksperimentet (se illustrationen):



Fejlkilder er størrelser eller fænomener, der er en (væsentlig) del af det fysiske system, man forsøger at beskrive, men som ikke indgår i den formel eller teori, man anvender til at beskrive systemet.

Eksempel: Man vil undersøge faldloven $s(t)=\frac{1}{2}\cdot g\cdot t^2$ ved at lade en genstand falde fra forskellige højder over jordoverfladen og måle den tid, faldet tager.

Luftmodstanden er en skjult variabel (fejlkilde), da den kan siges at korrelere med både strækningen og tiden. Jo større luftmodstand, jo kortere strækning (på en fast tid), og jo større luftmodstand, jo længere tid (med en fast strækning). Hvis det er strækningen, man vil behandle som uafhængig variabel, kan man sige, at man vil måle for store tider.

Man kan sige, at stikprøven (forsøgene) er biased, da sandsynligheden for at måle for korte tider er (meget) mindre end sandsynligheden for at måle for store tider.

Og dermed får man en systematisk fejl (man måler for store tidsrum, og formlen ser altså ud til at skulle forkastes).

Estimater af deskriptorer ud fra stikprøver

Formålet med stikprøver er som nævnt at anvende dem til at kunne sige noget om hele populationen. Men slutningen fra stikprøve til population er en induktiv slutning, og vi kan altså ikke være sikre på, at vores værdier fra stikprøven er rigtige. Man taler derfor om *estimater* af værdierne i stedet for om bestemmelse af værdierne.

F.eks. er der et gennemsnitligt antal stykker slik pr. pose i en konkret produktion, og dette gennemsnit kunne man bestemme ved at tælle antal stykker slik i hver pose i produktionen. Dette gennemsnit er den sande værdi for gennemsnittet.

Men hvis man udtager en stikprøve på f.eks. 50 poser og finder et gennemsnit ved at regne på disse 50 poser, så finder man et estimat for gennemsnittet, og dette estimat kan godt afvige (lidt) fra den sande værdi.

Formler for estimaterne af gennemsnit og spredning er:

Stikprøve-estimater: De ud fra stikprøven $\{x_1, x_2, x_3, ..., x_n\}$ bestemte estimater $\bar{x}$ og s for henholdsvis middelværdien μ og spredningen σ for populationen er:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

Med Gym-pakken bestemmes estimatet for spredningen med en af kommandoerne *standardafvigelse* og *stikprøvespredning*

Bemærk nævneren i brøken under kvadratrodstegnet. Det er den eneste forskel fra vores tidligere formler. Vi skal altså ikke dele med stikprøvens størrelse, men med én mindre end stikprøvens størrelse. Estimatet for spredningen (som er vores bedste bud på populationens spredning) bliver altså (lidt) større end den værdi, vi ville have fået, hvis vi anvendte vores tidligere formel for spredningen.

Nævneren $(n-1)$ skyldes, at vores sum under kvadratrodstegnet vil give en lidt for lille værdi, fordi vi udregner afvigelser for værdier i forhold til en størrelse, der er beregnet ud fra disse værdier. HVIS vi af en eller anden årsag allerede kender middelværdien for populationen og kun er interesseret i spredningen, skal vi anvende den kendte middelværdi og vores tidligere formel for spredning.

Nogle tal inden for statistik er lidt vilkårlige (f.eks. 1,5 i forbindelse med bestemmelse af *outliers*), men $(n-1)$ er IKKE en tilfældig nævner. Det kan bevises (hvilket vi ikke kaster os ud i), at det giver det bedste bud på populationens spredning.

Egentlig dækker begreberne *spredning* og *standardafvigelse* over præcis det samme (de er synonymer). Når Gym-pakken skelner mellem de to, er det altså blot for at kunne anvende to forskellige formler:

$$A := [4, 10, 15, 23, 5, 7]:$$
$$spredning(A) = 6.59966329173441$$
$$standardafvigelse(A) = 7.22956891357528$$
$$stikprøvespredning(A) = 7.22956891357528$$

Opgaverne 412*

Konfidensintervaller

Man kan bruge stikprøver til at estimere parametre for en population og angive disse ved punktangivelser, f.eks. "Vi har målt tyngdeaccelerationen til $9{,}81\frac{\text{m}}{\text{s}^2}$" eller "Ifølge meningsmålingen vil 13,4% af vælgerne stemme på partiet X".

Men en sådan punktangivelse fortæller intet om, hvor sikre vi er på resultatet. Vi ved, at det er et estimat, men der er forskel på, om tallet 9,81 er fremkommet ved at foretage to målinger på henholdsvis 7,81 og 11,81 eller 100 målinger, der alle lå mellem 9,7 og 9,9. Det er klart, at hvis vores stikprøve er repræsentativ, er 9,81 et bedre estimat i sidstnævnte tilfælde.

For at kunne inddrage dette aspekt benytter man spredninger til at konstruere såkaldte *konfidensintervaller*, dvs. vi angiver vores resultat ved et interval. For at kunne gøre det, skal vi først have indført nogle begreber:

α : Signifikansniveau

Et tal mellem 0 og 1 (mellem 0% og 100%). Ofte anvendes $\alpha = 5\%$. Det er egentlig et begreb, vi først skal anvende under 'test', men det hænger direkte sammen med …

$1-\alpha$: Konfidensniveau

Et tal mellem 0 og 1 (mellem 0% og 100%). Ofte anvendes $1-\alpha = 95\%$

κ : Kritisk værdi

Et tal direkte knyttet til konfidensniveauet. Hver fordeling har sin egen omregning fra konfidensniveau til kritisk værdi. Vi vil kun regne på situationer, hvor vi anvender normalfordelinger, og her har vi set omregningsmetoden i afsnittet *Nogle vigtige værdier for normalfordelingen*, så den gule boks i afsnittet har allerede givet os nogle af følgende værdier:

For normalfordelinger	
$1-\alpha$	κ
68,3%	1
90%	1,645
95%	1,960
95,4%	2
98%	2,326
99%	2,576
99,73%	3

Den kritiske værdi fortæller os altså, hvor mange gange spredningen vi skal gå ud til begge sider fra middelværdien for at finde en procentdel svarende til konfidensniveauet i intervallet.

$\left[\hat{v}-\kappa\cdot\sigma,\hat{v}+\kappa\cdot\sigma\right]$: $(1-\alpha)$ – Konfidensinterval

$\hat{v}$: Vores estimat af den parameter for populationen, vi vil bestemme.

σ : Spredningen på estimatet. Denne spredning kan være kendt eller estimeret med $(n-1)$-formlen, men en væsentlig pointe er, at det er spredningen på $\hat{v}$, og hvis $\hat{v}$ er en middelværdi, skal vi også inddrage Den Centrale Grænseværdisætning efter $(n-1)$-formlen.

Vigtigt

- Alt, hvad vi siger om konfidensintervaller, forudsætter, at vores stikprøve er repræsentativ. Hvis vores stikprøve er biased, kan vi slet ikke sætte procenter på.
- Konfidensniveauet skal vælges, **inden** man indsamler sine data. Dvs. **inden** man har nogen oplysninger om de konkrete data, skal man sige f.eks. "Jeg vælger at arbejde med et 95%-konfidensinterval."

Eksempel 12 – 0,01: En fabrik producerer et parti skruer bestående af 300.000 "ens" skruer.

Scenarie 1: Samtlige skruer passerer igennem et måleapparat, der uhyre nøjagtigt måler længden af den enkelte skrue. Da samtlige skruer måles, kan fabrikken bestemme gennemsnitslængden samt spredningen på produktionen med formlerne:

$$\mu = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} \qquad \sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu)^2} \qquad (n = 300.000)$$

Dette er de sande værdier for gennemsnitslængden og spredningen. Hvis skruerne skal anvendes til noget, hvor der kun må være meget små forskelle på længderne, skal spredningen σ være meget lille. Dette kan kun sikres gennem det udstyr, der bruges til fremstillingen. I teorien er σ uafhængigt af partiets størrelse, dvs. det ville ikke ændre spredningen, hvis man fremstillede 300 milliarder skruer (under antagelse af at udstyret ikke bliver slidt).

Scenarie 2: Det antages nu, at man kun måler skruelængderne i en stikprøve på 100 af skruerne. Med nedenstående formler kan man estimere gennemsnitslængden og spredningen i selve partiet på 300.000 skruer:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} \qquad s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \mu)^2} \qquad (n = 100)$$

Disse værdier er vores bedste estimat af de sande værdier μ og σ fra scenarie 1. Men pga. tilfældigheder vil der som udgangspunkt være forskel mellem de sande værdier og vores estimater. Vi kan ikke sige noget sikkert om den konkrete forskel, men vi kan sige noget om vores forventninger til forskellen mellem μ og $\bar{x}$ (se næste scenarie):

Scenarie 3: Vi ser nu på betydningen af stikprøvens størrelse, dvs. vi sammenligner stikprøver med størrelserne 100 (scenarie 2) og 500.
Begge stikprøver kan med formlerne fra scenarie 2 bruges til at bestemme vores bedste estimat for de sande værdier $(n = 100 \quad \text{eller} \quad n = 500)$.

Men hvorfor er der så forskel på en stikprøve på 100 og en på 500 skruer?
Fra Den Centrale Grænseværdisætning ved vi, at vores udregnede gennemsnit med god tilnærmelse vil kunne beskrives med en normalfordeling med spredningen

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad (n = 100 \quad \text{eller} \quad n = 500)$$

Da *s* i princippet ikke afhænger af stikprøvens størrelse, vil vores stikprøve på 500 give en mindre spredning på det estimerede gennemsnit end stikprøven på 100.

Men bemærk: Dette er spredningen på det estimerede gennemsnit af længden. Denne spredning kan vi gøre mindre ved at gøre stikprøven større. Men det er IKKE spredningen i produktionen. Dvs. den fortæller ikke noget om σ fra scenarie 1. Dvs. man kan ikke nødvendigvis bruge skruerne til noget, der kræver meget små forskelle i længderne, blot fordi man har fået bestemt gennemsnitslængden i produktionen meget præcist.

Eksempel 12: Man har valgt at arbejde med et 95%-konfidensinterval og dermed den kritiske værdi 1,960 (fordi vi ønsker at finde et gennemsnit, og ifølge Den Centrale Grænseværdisætning ved vi, at dette gennemsnit tilnærmelsesvist vil være normalfordelt). Der foretages derefter 16 målinger af tyngdeaccelerationen (målt i enheden $\frac{m}{s^2}$) svarende til at udtage en stikprøve med størrelsen 16.

Gym-pakken benyttes til at bestemme stikprøve-estimaterne (bemærk, at det er kommandoen *standardafvigelse*, der benyttes – man kunne også have brugt *stikprøvespredning*):

$Tyngdeaccelerationen := [9.93, 9.82, 9.71, 9.62, 9.89, 10.12, 9.55, 9.78, 9.96, 9.91, 9.80, 9.74, 9.81, 10.01, 9.84, 9.67]$:
$middel(Tyngdeaccelerationen) = 9.822500000$
$standardafvigelse(Tyngdeaccelerationen) = 0.147670805050731$

Vi benyttede stikprøven til at estimere middelværdi og spredning for "populationen". Da vi jo går ud fra, at tyngdeaccelerationen det pågældende sted har én sand værdi, kan det lyde underligt, at der skulle være en spredning, men her skal 'populationen' forstås som en uendelig mængde målinger af tyngdeaccelerationen, hvor man pga. måleusikkerhed vil få forskellige værdier, men hvor vi går ud fra, at gennemsnittet af denne uendelige mængde målinger er den sande værdi for tyngdeaccelerationen.

Vi er dog ikke interesserede i spredningen på målingerne, men på **spredningen på middelværdien**, og derfor skal vi tage Den Centrale Grænseværdisætning i brug:

$$\text{Spredningen på middelværdien: } \sigma = \frac{0.147670805050731}{\sqrt{16}} = 0.03691770128$$

$$\text{Venstre intervalendepunkt: } 9.8225 - 1.96 \cdot 0.03691770128 = 9.750141305$$
$$\text{Højre intervalendepunkt: } 9.8225 + 1.96 \cdot 0.03691770128 = 9.894858695$$

Med middelværdi, kritisk værdi og spredning på middelværdien, har vi så fundet:

$$\text{95\%-konfidensintervallet (med enheden } \frac{m}{s^2}\text{): } [9.75, 9.89]$$

Opgaverne 414*

Da Den Centrale Grænseværdisætning fortæller os, at spredningen på det estimerede gennemsnit findes ved $\frac{\sigma}{\sqrt{n}}$, hvor n er stikprøvens størrelse, vil en større stikprøve give et smallere konfidensinterval. Og da et højere konfidensniveau giver en højere kritisk værdi, vil et højere konfidensniveau give bredere konfidensintervaller. Vi har altså:

Alt andet lige…:	En større stikprøve giver et smallere konfidensinterval.
	Et højere konfidensniveau giver et bredere konfidensinterval.

Men hvad fortæller vores konfidensintervaller os?

Lad os begynde med fælden og se på, hvad de IKKE fortæller os. Når man har udregnet et konfidensinterval, vil den sande værdi enten ligge inden for intervallet eller uden for. Dvs. enten det ene eller det andet. Derfor kan man IKKE sige, at når man har fundet et 95%-konfidensinterval, så er der 95% chance for, at den sande værdi ligger inden for intervallet. For når man **har** konfidensintervallet, er der ikke længere tale om sandsynligheder.

Men det, som man KAN sige, er (udtrykt med konfidensniveauet 95%):

Ækvivalente betydninger af konfidensintervaller:

- Inden jeg udtager min (repræsentative) stikprøve, er der 95% chance for, at det 95%-konfidensinterval, som jeg vil beregne ud fra stikprøven, vil indeholde den sande værdi.
- 95% af de 95%-konfidensintervaller, der beregnes ud fra stikprøver, indeholder den sande værdi.

Konfidensinterval for hældning

I Eksempel 12 målte vi samme størrelse mange gange. Med et *QQplot* afgjorde vi, om målingerne var normalfordelte. Men ofte måler man på størrelser, der afhænger af hinanden (f.eks. i 'opvarmning af vand' den afhængige variabel ΔT og den uafhængige variabel Q). Hvis der er tale om lineære sammenhænge, vil hældningen direkte eller indirekte kunne fortælle os noget om en central størrelse (f.eks. den specifikke varmekapacitet for vand). Man vil så gerne kunne sige noget om nøjagtigheden af den fundne værdi. Vi skal derfor nu se på konfidensintervaller for hældninger.

Sætning 1: Konfidensinterval for hældning baseret på *n* målinger (Gym-pakke: *testLin*)

Forudsætninger:

1) Det er en lineær sammenhæng (punkterne danner en ret linje i et almindeligt koordinatsystem).
2) **Enten** er antallet af målepunkter tilpas stort **eller** residualerne er normalfordelt (kan testes med et QQplot - Gym-pakken har kommandoen *residualQQplot*).

Bestemmelse af konfidensintervallet $\left[\hat{v}-\kappa\cdot\sigma,\hat{v}+\kappa\cdot\sigma\right]$ **(brug *testLin*)**

$\hat{v}$**:** Estimatet er vores hældning bestemt ved lineær regression (mindste kvadraters metode).

κ: Den kritiske værdi skal bestemmes ud fra en såkaldt *t*-fordeling med $n-2$ frihedsgrader (se afsnittet om test). Værdien afviger lidt fra værdien bestemt ud fra normalfordelingen.

σ: $\sigma=\dfrac{s_{res}}{\sqrt{n}\cdot s_x}$, hvor s_x er spredningen på den uafhængige variabel, og s_{res} er residualspredningen.

Residualspredningen: $s_{res}=\sqrt{\dfrac{\sum_{i=1}^{n} r_i^2}{n-2}}$, r_i'erne er residualerne. (Gym-pakken: *residualspredning*)

En **løs** forklaring på formlerne (dvs. ikke et bevis): Udtrykket $(n-2)$ i nævneren på residual-spredningen skyldes, at 2 frihedsgrader er gået til bestemmelse af hældning og skæring. Formlen for spredningen på hældningen giver mening, når man ser på, at tælleren er spredningen på residualerne, dvs. spredningen på *y*-værdier, mens den ene faktor i nævneren er spredningen på *x*-værdierne ($a=\dfrac{\Delta y}{\Delta x}$).

Residualspredningen er ligesom forklaringsgraden (*r*-kvadratet) et udtryk for punkternes afstand til regressionslinjen. Jo tættere værdien er på 0, jo tættere ligger punkterne på linjen. Men residualspredningen er ikke dimensionsløs, så dens størrelse skal sammenlignes med de konkrete *y*-værdier, hvis det skal give mening.

Eksempel 13a: Et datasæt med 61 sammenhørende værdier af vægt (i kg) og tempo (min. pr. km) for en løber er hentet ind fra Excel, hvor kommaer med 'søg'-'erstat' er ændret til punktummer, og gemt i Maple i de lodrette lister *Vægt* og *Løbetid*, som pga. pladsen ikke angives her.

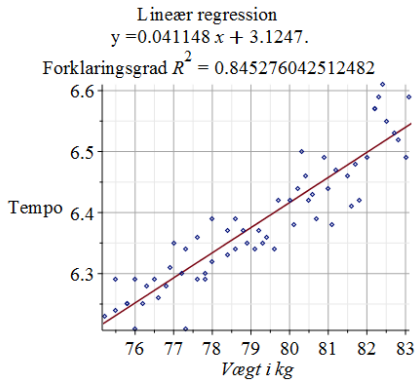
Vi kan så bestemme hældningen på sædvanlig vis ved hjælp af lineær regression:

$$l(v) := LinReg(Vægt, Løbetid, v):$$
$$l(v) = 0.0411476127767431\,v + 3.12467719457650$$

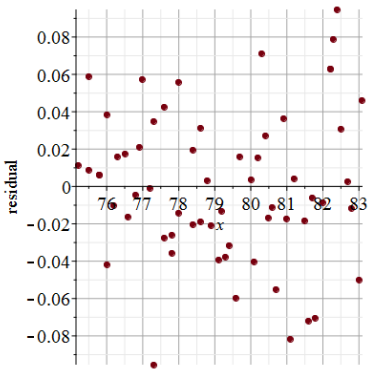
Dvs. hældningen er 0,041, hvilket fortæller os, at løberen kan lægge 0,041 minutter pr. km. til sin løbetid pr. kg. kropsvægten øges. Dette har vi set før. Men vi skal nu se på, hvor præcist vi mener, at disse 0,041 minutter er bestemt …

Eksempel 13b: Vi vil bestemme et 95%-konfidensinterval for hældningen. Vi skal derfor **først** sikre os, at der rent faktisk **er** tale om en lineær sammenhæng. Dvs. enten **ses** på, om punkterne danner en ret linje i et alm. koordinatsystem, eller om residualerne ligger usystematisk omkring 0:

LinReg(*Vægt*, *Løbetid*)



plotResidualer(*Vægt*, *Løbetid*, *1*)



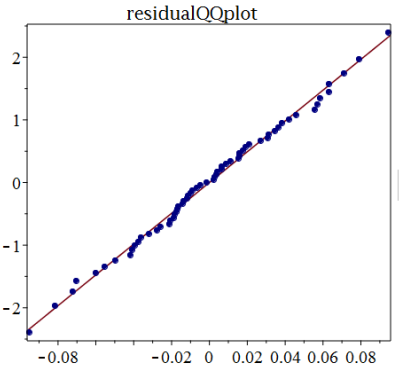
Venstre: Punkterne danner en ret linje. Afvigelserne virker usystematiske. Der er ingen buet tendens. Så en lineær model er en passende model.

Højre: Residualerne ligger usystematisk spredt omkring 0, så den lineære model er en god model.

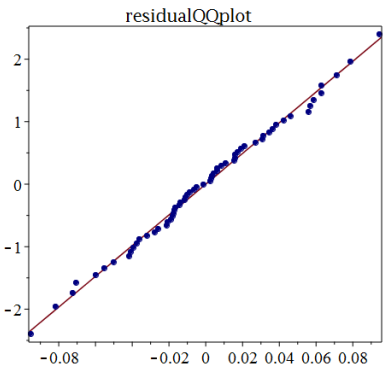
Kommentar: Faktisk fungerer eksponentielle udviklinger og potensfunktioner lige så godt. Det skyldes, at datasættet dækker så lille et vindue (8 kg ud af 83 kg og 0,4 minutter ud af godt 6,6 minutter), at alt groft sagt vil virke lineært.

Vi vil nu undersøge, om residualerne er normalfordelt, og bruger derfor *residualQQplot*. Da vi allerede har foretaget den lineære regression, kan vi bruge indtastningen til venstre nedenfor, og ellers kan man altid klare sig med indtastningen til højre (som det ses, er de identiske):

residualQQplot(*Vægt*, *Løbetid*, *1*)



residualQQplot(*Vægt*, *Løbetid*, *LinReg*)



Punkterne ligger meget tæt på den rette linje, så residualerne er helt klart normalfordelte.

Forudsætningerne er altså opfyldt for, at man må angive et konfidensinterval.

Vi benytter Gym-pakkens *testLin*:

testLin(*Vægt*, *Løbetid*, *konfidens* = 0.95)

	a	b
Koefficient	0.041148	3.124677
Standardfejl	0.002292	0.181607
t-stat	17.953392	17.205731
p-værdi	0.000000	0.000000
Nedre 95.00%	0.036562	2.761283
Øvre 95.00%	0.045734	3.488072
Frihedsgrader	59	

Vores konfidensinterval for hældningen er derfor: **[0.037,0.046].**

Hvis konfidensintervallet indeholder både negative og positive værdier – og dermed også 0 – må man forkaste ideen om en lineær sammenhæng, for man kan ikke tale om en sammenhæng, hvis der kan være både negativ og positiv korrelation. Eller i tilfældet $a = 0$, slet ingen korrelation.

Opgaverne 416*

Konfidensinterval for sandsynligheder

Konfidensintervaller for sandsynligheder møder vi ofte i forbindelse med meningsmålinger.

Situationen er: Man går ud og spørger *n* personer, hvad de vil stemme på. Hvis man så tager ét parti ad gangen, her partiet X, kan svaret omfortolkes til, at man har svaret på et spørgsmål med to svarmuligheder:

Personen stemmer på parti X (succes) eller personen stemmer ikke på parti X (fiasko).

Hvis *r* personer svarer, at de stemmer på parti X, har vi estimeret successandsynligheden $\hat{p}$ for, at en tilfældig person vil stemme på parti X, til:

$$\hat{p} = \frac{r}{n}$$

Hvis vi nu vender situationen om og tager udgangspunkt i vores estimerede successandsynlighed, så giver formlen os samtidig, at hvis vi går ud og spørger *n* personer, om de vil stemme på partiet X, vil vi i gennemsnit kunne forvente, at $r = \hat{p} \cdot n$ vil svare 'ja' (middelværdi for binomialfordeling).

Spredningen på middelværdien for binomialfordeling er $\sigma = \sqrt{n \cdot \hat{p} \cdot \left(1 - \hat{p}\right)}$.

Den relative spredning (%-vis spredning) er så $\frac{\sigma}{n} = \frac{\sqrt{n \cdot \hat{p} \cdot \left(1 - \hat{p}\right)}}{n} = \sqrt{\frac{n \cdot \hat{p} \cdot \left(1 - \hat{p}\right)}{n^2}} = \sqrt{\frac{\hat{p} \cdot \left(1 - \hat{p}\right)}{n}}$

Sætning 2: Konfidensinterval for sandsynligheder baseret på stikprøver

$$\left[\hat{p} - \kappa \cdot s_p, \hat{p} + \kappa \cdot s_p\right]$$

Gym-kommandoen: *konfidensInterval*

$\hat{p}$: Den estimerede successandsynlighed. $\hat{p} = \frac{r}{n}$

κ : Den kritiske værdi. Anvend værdierne for normalfordeling.
Nogle gange anvendes 2 i stedet for 1,96 for konfidensniveauet 95% (bl.a. i formelsamlingen)

s_p : Spredningen på den estimerede successandsynlighed. $s_p = \sqrt{\frac{\hat{p} \cdot \left(1 - \hat{p}\right)}{n}}$

$1{,}96 \cdot s_p$ eller $2 \cdot s_p$ kaldes for ***den statistiske usikkerhed***.

Eksempel 14: Vi vælger at arbejde med konfidensniveauet 95% (dvs. $\kappa = 1{,}96$)
I en stikprøve på 2734 personer svarer 845, at de vil stemme på partiet X.

Den estimerede successandsynlighed er: $\hat{p} = \frac{845}{2734} = 0{,}3090709583 = 30{,}9\%$

$$s_p = \sqrt{\frac{0{,}30907 \cdot \left(1 - 0{,}30907\right)}{2734}} = 0{,}00883785$$

Konfidensintervallet: $\left[0{,}30907 - 1{,}96 \cdot 0{,}00884 \; ; \; 0{,}30907 + 1{,}96 \cdot 0{,}00884\right] = \underline{\underline{\left[29{,}2\% \; ; 32{,}6\%\right]}}$

Statistisk usikkerhed: $1{,}96 \cdot 0{,}00883785 = 0{,}017322 = \underline{\underline{1{,}7\%}}$

Gym-pakken: $konfidensInterval(845,\ 2734,\ 0.95) = [0.291749,\ 0.326393]$

Opgaverne 418*

TEST

Inden for statistik har man valgt at sige, at man udfører *et* test. Denne (korrekte) sprogbrug er dog ikke slået helt igennem, så man kan stadig masser af steder læse, at der er foretaget *en* statistisk test.

Selve tankegangen og fremgangsmåden er:

Et test består altid i, at man vælger to hypoteser og et test som forklaringsmodel i den pågældende situation. Hypoteserne skal holdes op imod hinanden, når man har indsamlet data.

Den ene hypotese kaldes *nulhypotesen* og betegnes H_0.
Den anden hypotese kaldes den alternative hypotese og betegnes H_1.

Desuden vælger man et såkaldt *signifikansniveau* α (ofte 5%), der er et mål for, hvornår vi – under antagelse af, at nulhypotesen er rigtig – får usandsynlige resultater. Hvis $\alpha = 5\%$, betyder det, at HVIS nulhypotesen er rigtig, er der 5% risiko for, at vi kommer til at forkaste den.

Derefter udtager man sin stikprøve, så man har en række måleresultater.

På baggrund af nulhypotesen og det korrekt valgte test (vi skal lære om binomialtest, to slags χ^2-test, tre slags t-test og lidt om z-test), udregner man en *teststørrelse*, og denne kan omregnes til en sandsynlighed p for ***under forudsætning af, at nulhypotesen er sand,*** *at få det pågældende måleresultat* ***eller et måleresultat, der (endnu) mere end det pågældende måleresultat støtter den alternative hypotese frem for nulhypotesen.***

Jo mindre p er, des mindre sandsynligt er det pågældende resultat **under forudsætning af nulhypotesen**, dvs. at hvis p bliver tilstrækkelig lille, vil man forkaste nulhypotesen. Man sammenligner så p og α.

Hvis $p < \alpha$ forkastes nulhypotesen.

Hele testmetoden kan føre til to typer af fejl:

Fejl af type 1: En sand nulhypotese forkastes.
Fejl af type 2: En falsk nulhypotese forkastes ikke.

Hypoteserne

Det er nulhypotesen, der undersøges, dvs. det er nulhypotesen, man kan ende med at forkaste eller ikke forkaste.

Man kan aldrig bevise en teori, og det er derfor vigtigt at bemærke, at din konklusion altid skal omhandle forkastelse eller ikke forkastelse af nulhypotesen.

Nulhypotesen er altid den hypotese, der beskriver situationen, som den forventes at være ifølge en teori eller en tabel, eller som siger, at der ikke er nogen sammenhæng mellem forskellige størrelser.

Den alternative hypotese siger, at teorien ikke holder, at tabelværdien ikke er den rigtige (evt. at den er større/mindre) eller at der rent faktisk er en sammenhæng mellem de forskellige størrelser.

Det er meget vigtigt ikke at blande dit/forskerens ønske ind i valget af hypoteser. Ifølge vores gennemgang af videnskabelig praksis, bør man forsøge at opstille forsøg, der kan forkaste en teori, og oftest vil man inden for f.eks. samfundsvidenskaberne forsøge at finde sammenhænge mellem forskellige størrelser. Da nulhypotesen er vores udgangspunkt, kan det føre til den grundlæggende fejl, at man får formuleret en forkert nulhypotese, nemlig den hypotese, der er i overensstemmelse med ens ønske.

Eksempel 15: Ved hjælp af et svingende pendul vil man bestemme tyngdeacceleration ved jordoverfladen. Vi har en tabelværdi på $g = 9{,}82\frac{m}{s^2}$, og derfor bliver vores hypoteser:

$$H_0: \ g = 9{,}82\frac{m}{s^2} \qquad H_1: \ g \neq 9{,}82\frac{m}{s^2}$$

Eksempel 16: På et optisk gitter står der, at antallet af spalter pr. mm i et gitter (dvs. $\frac{1}{d}$) er 1200. Vi har en mistanke om, at dette tal er forkert. Vores hypoteser bliver:

$$H_0: \ \frac{1}{d} = 1200 \ mm^{-1} \qquad H_1: \ \frac{1}{d} \neq 1200 \ mm^{-1}$$

Eksempel 17: Vi ønsker at vise, at der er en sammenhæng mellem lektielæsning og opnået fagligt niveau. Vores hypoteser bliver derfor:

H_0: Der er ingen sammenhæng mellem lektielæsning og opnået fagligt niveau.

H_1: Der er en sammenhæng mellem lektielæsning og opnået fagligt niveau.

Eksempel 18: Man ønsker at undersøge, om vælgertilslutningen til partierne har ændret sig siden seneste valg. Hypoteserne bliver derfor:

H_0: Vælgertilslutningen har ikke ændret sig siden seneste valg.

H_1: Vælgertilslutningen har ændret sig siden seneste valg.

Eksempel 19: Man ønsker at undersøge, om et bestemt stof virker mod hovedpine:

H_0: Stoffet virker ikke mod hovedpine.

H_1: Stoffet virker mod hovedpine.

Eksempel 20: Vi vil teste faldloven – som vi ikke tror på - med en bold i frit fald.

H_0: Faldloven gælder.

H_1: Faldloven gælder ikke.

Eks. 21: Vi ønsker i et forsøg med opvarmning at bestemme vands specifikke varmekapacitet.

H_0: Vands specifikke varmekapacitet har den værdi, der kan slås op i databogen.

H_1: Vands specifikke varmekapacitet har en anden værdi end den, der kan slås op i databogen.

Oversigt over begreber i forbindelse med statistiske test

Et statistisk test er en procedure til at vurdere, om et datamateriale er i overensstemmelse med en fremsat hypotese.

Bemærk ordet "vurdere". Man kan **ikke** afgøre, om der er overensstemmelse, men 'kun' give en (velbegrundet) vurdering.

Nulhypotesen H_0: Den hypotese, der afprøves i et statistisk test. Den kan i mange situationer angives i form af den antagne værdi for den parameter, der testes på, f.eks. en middelværdi:

$$H_0: \ \mu = \mu_0 .$$

Den alternative hypotese H_1: Den hypotese, som nulhypotesen holdes op imod. Med ovenstående nulhypotese kan den alternative hypotese være:

a) $H_1: \ \mu \neq \mu_0$ Tosidet test

b) $H_1: \ \mu < \mu_0$ Venstresidet test

c) $H_1: \ \mu > \mu_0$ Højresidet test

Signifikans: Et resultat siges at være *statistisk signifikant*, hvis det er usandsynligt, at det er indtruffet ved et tilfælde. Det kan også udtrykkes ved, at der foreligger *signifikans*.
Signifikansniveauet α er den sandsynlighed, der fastsætter, hvad der skal regnes som 'usandsynligt'. Dette niveau er ikke fast. Det oftest benyttede er $\alpha = 0{,}05$.
Hvis resultatet er usandsynligt (dvs. hvis *p*-værdien er mindre end signifikansniveauet), forkastes nulhypotesen.

Acceptområde A: Det område (den mængde), inden for hvilket den målte parameter skal ligge, hvis nulhypotesen ikke skal forkastes.

Det kritiske område K: Det område (den mængde), inden for hvilket den målte parameter skal ligge, hvis nulhypotesen skal forkastes.

Ved et *tosidet test* ligger det kritiske område på hver sin side af acceptområdet. Hvis man f.eks. arbejder ud fra en antagelse om, at den målte parameter er normalfordelt med middelværdi μ og spredning σ og har fastsat et signifikansniveau på 5%, vil

det kritiske område være $K =]-\infty; \mu - 1{,}96 \cdot \sigma[\cup]\mu + 1{,}96 \cdot \sigma; \infty[$,

mens acceptområdet er $A = [\mu - 1{,}96\sigma; \mu + 1{,}96\sigma]$.

Ved et *venstresidet test* ligger det kritiske område til venstre for acceptområdet. Med samme antagelse som ovenfor fås det kritiske område $K =]-\infty; \mu - 1{,}645\sigma[$ og acceptområdet $A = [\mu - 1{,}645\sigma; \infty[$. Tallet 1,645 er fundet ud fra, at der skal være 5% chance for at havne i det kritiske område.

Detalje: Ved normalfordelinger og andre kontinuerte fordelinger kan 'snittet' lægges præcist ved de 5%. Dette er ikke tilfældet ved diskrete fordelinger (f.eks. binomialfordelingen). Så her er der brug for en mere præcis formulering, der siger, at signifikansniveauet er den **maksimale** sandsynlighed, der fastsætter, hvad der skal regnes som 'usandsynligt'. Man 'favoriserer' altså nulhypotesen.

Dvs. at med signifikansniveauet 5% kan man f.eks. være nødt til at vælge en kritisk mængde, som der måske kun er 1,3% sandsynlighed for at ramme inden for (bemærk, at de 1,3 bare er et eksempel på et tal mindre end 5).

Fejltyper:

	Sand nulhypotese	Falsk nulhypotese
Nulhypotesen forkastes ikke	Godt	Fejl af 2. art Type-II fejl β–fejl
Nulhypotesen forkastes	Fejl af 1. art Type-I fejl α–fejl	Godt

Meget væsentlig pointe i forbindelse med valg af alternativ hypotese:

Som angivet tidligere kan man vælge 'tosidet test', 'venstresidet test' og 'højresidet test'. Dvs. at hvis man f.eks. vil undersøge, om drenge og piger laver lige mange lektier i gymnasiet, bliver nulhypotesen, at drenge og piger laver lige mange lektier i gymnasiet, og man skal så afgøre, hvilken af følgende tre alternative hypoteser, den skal holdes op imod:

a) Drenge og piger laver ikke lige mange lektier i gymnasiet (tosidet test).

b) Drenge laver flere lektier end piger i gymnasiet (højresidet).

c) Drenge laver færre lektier end piger i gymnasiet (venstresidet).

Den væsentlige pointe er, at **man skal træffe sit valg uafhængigt af sine data, dvs. som udgangspunkt allerede inden indsamling af data.**

Årsagen til dette er, at man ellers ender med at begå dobbelt så mange type-I-fejl, dvs. man får forkastet dobbelt så mange sande nulhypoteser, som hvis man gjorde det på den rigtige måde.

Antag nemlig, at nulhypotesen om, at drenge og piger laver lige mange lektier i gymnasiet, er sand. Hvis man undersøger dette i en hel række undersøgelser, vil man som udgangspunkt (da man arbejder med stikprøver) aldrig opnå præcis samme resultat for drenge og piger, men man kan regne med, at omkring 50% af undersøgelserne viser, at drenge laver flest lektier, mens 50% viser, at piger laver flest lektier.

Hvis man nu i en konkret undersøgelse kiggede på tallene og så, at drengene i stikprøven lavede flere lektier end pigerne, og derfor valgte et højresidet test, ville man placere hele det kritiske område på f.eks. 5% ude til højre i stedet for at fordele de 5% på 2,5% yderst til højre og 2,5% yderst til venstre. Hvis stikprøven havde vist, at drengene lavede færre lektier end pigerne, ville man med denne **forkerte metode** vælge et venstresidet test og placere de 5% yderst til venstre.

Hvis man altså først kigger på data, kommer man i dette tilfælde til reelt at placere 10% (5% i hver side), når det kritiske område skal angives, og dermed har man (ubevidst) fordoblet signifikansniveauet.

Binomialtest

I Maples Gym-pakke: *binomialTest*

Vi indleder med binomialtest, selvom dette som det eneste af vores test ikke indeholder en beregnet teststørrelse, der skal omregnes til en *p*-værdi. Til gengæld kan vi i modsætning til de andre test være med hele vejen matematisk.

Der ligger altid en fordeling til grund for et test. Til grund for binomialtest ligger binomialfordelingen, der som bekendt beskæftiger sig med *n* gentagelser af et forsøg med to mulige udfald (succes og fiasko) med konstant successandsynlighed *p*.

Binomialfordelingen er så sandsynlighedsfordelingen for den stokastiske variabel, der angiver antallet *r* af succeser.

Fordelingen kan beskrives ved både en diskret tæthedsfunktion og en fordelingsfunktion.

Inden vi ser på eksempler på binomialtest, skal vi derfor lige gennemgå nogle beregninger på binomialfordelingen.

Den diskrete tæthedsfunktion $f(r)$ angiver sandsynligheden for *r* succeser:

$$f(r) = p(X = r) = K(n,r) \cdot p^r \cdot (1-p)^{n-r}$$

I Maples Gym-pakke hedder kommandoen *binpdf* (*pdf ~ probability density function*).

Eksempel 22: Udregning med $n = 50$, $p = 0{,}17$ *og* $r = 9$:

```
restart
with(Gym) :
```

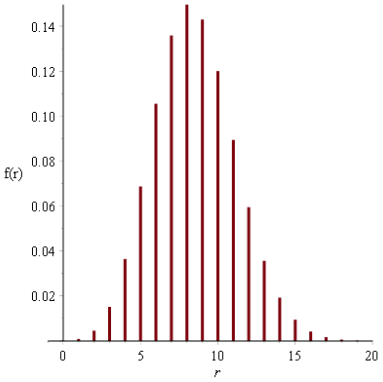
$$f(9) = P(X = 9) = K(50, 9) \cdot 0,17^{9} \cdot (1 - 0,17)^{50-9} = \frac{50!}{41! \cdot 9!} \cdot 0.17^{9} \cdot 0.83^{41} = 0.1429305268$$

$$binpdf(50, 0.17, 9) = 0.1429305268$$

Dvs. ved 50 gentagelser af et forsøg med successandsynligheden 0,17 er sandsynligheden 14,3% for at få netop 9 succeser.

Maple kan tegne et pindediagram som graf for tæthedsfunktionen:

$pindediagramBIN(50, 0.17)$



Det er sandsynligheden, der er ud ad 2. aksen, så vi kan se, at det passer fint med de 14,3% for 9 succeser.

Vi kan desuden se, at sandsynlighederne for at få mere end 20 succeser er forsvindende lille.

Fordelingsfunktionen angiver sandsynligheden for **højst** *r* succeser:

$$F(r) = p(X \le r) = \sum_{i=0}^{r} f(i) = \sum_{i=0}^{r} K(n,i) \cdot p^{i} \cdot (1-p)^{n-i}$$

Bemærk, at fordelingsfunktionen her anvender sumtegn, da binomialfordelingen er diskret, mens man for den kontinuerte normalfordeling anvender integraltegn i fordelingsfunktionen.

I Maples Gym-pakke hedder kommandoen *bincdf* (*cdf ~ cumulative distribution function*).

Udregning med $n = 50$, $p = 0{,}17$ *og* $r = 9$:

$$bincdf(50, 0.17, 9) = 0.6597105658$$

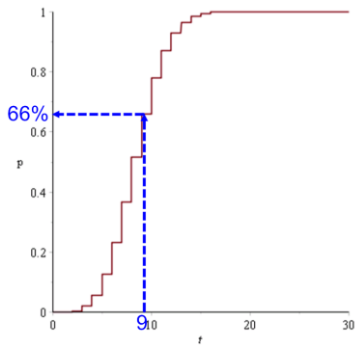
Dvs. at sandsynligheden for at få højst 9 succeser er 66,0%.

Dette kunne også være udregnet ved:

```
binpdf(50, 0.17, 0) + binpdf(50, 0.17, 1) + binpdf(50, 0.17, 2) + binpdf(50, 0.17, 3) + binpdf(50, 0.17, 4) + binpdf(50, 0.17, 5) + binpdf(50, 0.17, 6) + binpdf(50, 0.17, 7)
    + binpdf(50, 0.17, 8) + binpdf(50, 0.17, 9)
= 0.6597105657
```

Vi kan også bede Maple om at afbilde fordelingsfunktionen:

$$plot(bincdf(50, 0.17, t), t = 0..30)$$



Den stiplede blå linje er sat for at vise, at sandsynligheden for at få højst 9 succeser er 66%.

Man kan også anvende fordelingsfunktionen, hvis man skal finde sandsynligheden for at få ***mindst*** et bestemt antal succeser. Man skal i så fald udnytte:

$$p(X \geq r) = 1 - p(X \leq r-1)$$

Denne formel udnytter, at man er (100%) sikker på at få enten mindst r succeser eller højst $r-1$ succeser.

I eksemplet med $n = 50$ og $p = 0{,}17$ får man:

$$P(X \geq 10) = 1 - P(X \leq 9) = 1 - bincdf(50, 0.17, 9) = 0.3402894342$$

Dvs. der er 34% chance for at få mindst 10 succeser.

Endelig kan man også bruge fordelingsfunktionen til at bestemme sandsynligheden for at ramme inden for et vist interval ved at udnytte:

$$p(r \leq X \leq t) = p(X \leq t) - p(X < r) = p(X \leq t) - p(X \leq r-1)$$

Eksempel 23:

Sandsynligheden for at få mellem 7 og 12 succeser (begge tal inklusive) er:

$$P(7 \leq X \leq 12) = P(X \leq 12) - P(X \leq 6) = bincdf(50, 0.17, 12) - bincdf(50, 0.17, 6) = 0.6972285195$$

Dvs. sandsynligheden er 69,7%, hvilket også lidt mere besværligt kunne være fundet ved:

$$P(X=7) + P(X=8) + P(X=9) + P(X=10) + P(X=11) + P(X=12) =$$
$$binpdf(50, 0.17, 7) + binpdf(50, 0.17, 8) + binpdf(50, 0.17, 9) + binpdf(50, 0.17, 10) + binpdf(50, 0.17, 11) + binpdf(50, 0.17, 12) = 0.6972285193$$

Det er nu tid til at se på eksempler på binomialtest.

Eksempel 24:

Vi har købt en terning og vil undersøge, om det er en snydeterning, så man ikke har den rigtige sandsynlighed for at få en sekser. Oftest vil snydeterninger nok give for mange seksere, men hvis vi også holder muligheden åben for, at det kan være en snydeterning beregnet på modstanderen, laver vi et ligesidet test. Dvs. vores hypoteser bliver:

H_0: Det er ikke en snydeterning, dvs. $p_{succes} = \frac{1}{6}$ H_1: Det er en snydeterning, dvs. $p_{succes} \neq \frac{1}{6}$

Vi vælger at arbejde med signifikansniveauet 5% (dette skal også vælges inden forsøget udføres). Vi udfører et forsøg med 100 kast med terningen og får 23 seksere. Vi kan hurtigt se, at frekvensen af seksere er 23%, dvs. højere end sandsynlighedssuccesen på 16,7%, men spørgsmålet er, om forskellen er signifikant, dvs. om det er for usandsynligt med en ikke-snydeterning at få et sådant resultat eller noget, der er "værre".

Vi har valgt et ligesidet test med signifikansniveau 5%, så vi skal have placeret 2,5% i hver side. Vi kan nu gribe det an på forskellige måder:

Metode 1:

Vi finder sandsynligheden for ***under forudsætning af at nulhypotesen holder*** at få 23 eller flere seksere (dvs. mindst 23):

$$P(X \geq 23) = 1 - P(X \leq 22) = 1 - bincdf\left(100, \frac{1}{6}, 22\right) = 0.0630503270$$

Da sandsynligheden på 6,3% er større end 2,5%, er afvigelsen IKKE signifikant, dvs. vi kan IKKE forkaste nulhypotesen. Vi kan altså ikke hævde, at det er en snydeterning.

Metode 2: Vi vil først bestemme acceptområde og det kritiske område. Endnu engang husker vi på, at det er et ligesidet (tosidet) test, så vi skal have placeret (højst) 2,5% i hver side. Vi skal desuden lægge mærke til, at udregningerne i venstre side og højre side skal foretages forskelligt. Vi ved, at nulhypotesen forudsiger 16,7 seksere, så venstresiden består af hændelserne 0 - 16 seksere, mens højresiden er 17-100 seksere.

Først venstresiden:

Vi skal her se på, hvor springet forbi 2,5% sker, og vi ser på fordelingsfunktionens værdier:

$$seq\left(\left[x, bincdf\left(100, \frac{1}{6}, x\right)\right], x = 0..16\right)$$

[0, $1.207467347\ 10^{-8}$], [1, $2.535681429\ 10^{-7}$], [2, 0.000002644353490], [3, 0.00001826415109], [4, 0.00009402016947], [5, 0.0003849232800], [6, 0.001306116464], [7, 0.003780178156], [8, 0.009532371592], [9, 0.02129241151], [10, 0.04269568415], [11, 0.07771922120], [12, 0.1296708012], [13, 0.2000052479], [14, 0.2874209174], [15, 0.3876575517], [16, 0.4941589757]

Dette viser, at sandsynligheden for at få f.eks. højst 7 seksere er 0,3780178156%.
Her ses det, at springet forbi 2,5% sker fra 9 til 10 seksere.

Højresiden:

Her skal vi se på sandsynligheden for at få det på gældende antal seksere ***eller*** flere. Da dette udregnes som $P(X \geq r) = 1 - P(X \leq r-1)$, skal vores sekvens angives anderledes:

$$seq\left(\left[x, 1 - bincdf\left(100, \frac{1}{6}, x - 1\right)\right], x = 17..50\right)$$

[17, 0.5058410243], [18, 0.4005925583], [19, 0.3035300840], [20, 0.2197498431], [21, 0.1518878479], [22, 0.1001834706], [23, 0.0630503270], [24, 0.0378643687], [25, 0.0217033787], [26, 0.0118774969], [27, 0.0062087189], [28, 0.0031013887], [29, 0.0014811380], [30, 0.0006765997], [31, 0.0002957849], [32, 0.0001238040], [33, 0.0000496373], [34, 0.0000190716], [35, 0.0000070251], [36, 0.0000024818], [37, $8.412\ 10^{-7}$], [38, $2.737\ 10^{-7}$], [39, $8.55\ 10^{-8}$], [40, $2.56\ 10^{-8}$], [41, $7.4\ 10^{-9}$], [42, $2.0\ 10^{-9}$], [43, $5.\ 10^{-10}$], [44, $1.\ 10^{-10}$], [45, 0.], [46, 0.], [47, 0.], [48, 0.], [49, 0.], [50, 0.]

(Der er kun fortsat op til 50 seksere, da sandsynlighederne for flere seksere er ekstremt små).
Det væsentlige er, hvor springet forbi 2,5% sker.
Dette ses at ske fra 24 til 25 seksere. Vi får dermed følgende:

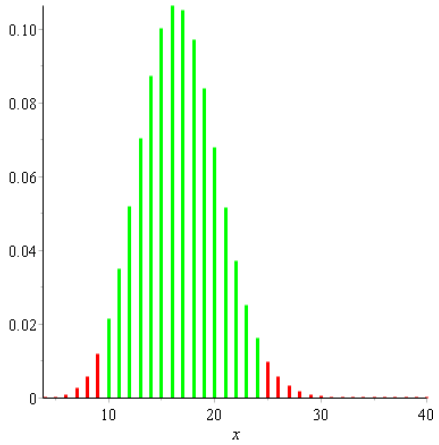
Acceptmængden er $A = \{10,11,12,...,24\}$

Den kritiske mængde er $K = \{0,1,2,...,9\} \cup \{25,26,27,...,100\}$

Da udfaldet på 23 seksere ligger i acceptmængden, forkastes nulhypotesen IKKE.

Metode 3: Med Gym-pakken kan man desuden få:

$$binomialTest\left(100, \frac{1}{6}, 0.05, tosidet\right)$$



Når man anvender 'binomialTest', skal man udover *n* og *p* angive signifikansniveauet (her 0,05) samt om testet skal være 'venstre', 'højre' eller 'tosidet' (et andet ord for *ligesidet* eller *dobbeltsidet*).

Bemærk, at det er **tæthedsfunktionen**, der er angivet, dvs. sandsynligheden for de enkelte hændelser.

Acceptmængden er angivet med grønt.

Eksempel 25: Vi har spillet med terning og fået en mistanke om, at den er skæv og giver for mange 5'ere. Dette vil vi gerne undersøge, og vi opstiller derfor hypoteserne:

H_0: Terningen er ikke skæv, dvs. $p_{succes} = \frac{1}{6}$

H_1: Terningen giver for mange 5'ere, dvs. $p_{succes} > \frac{1}{6}$

Vi vælger igen signifikansniveauet 5% og udfører et forsøg med 300 kast, hvor vi får 67 5'ere. Det er jo flere end de forventede 50 5'ere, men spørgsmålet er, om forskellen er signifikant.

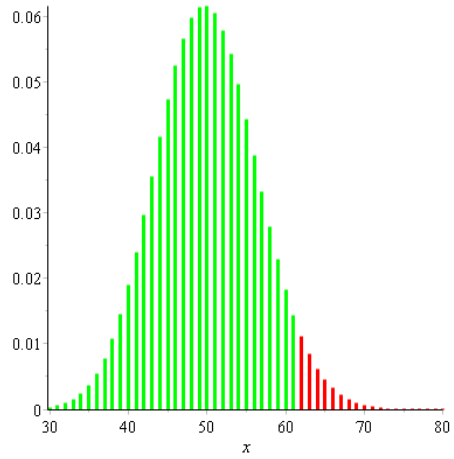
Da vi her har valgt at lave et højresidet test, skal alle 5% (den kritiske mængde) placeres til højre. Vi får nu:

$$P(X \geq 67) = 1 - P(X \leq 66) = 1 - bincdf\left(300, \frac{1}{6}, 66\right) = 0.0066785993 < 0,05$$

Da sandsynligheden for at få mindst 67 5'ere med en 'ærlig' terning er under 5%, har vi altså fået signifikant flere 5'ere end forventet, og vi må **forkaste nulhypotesen** til fordel for den alternative hypotese. Vi konkluderer altså, at terningen er skæv og giver for mange 5'ere.

Grafisk ses acceptmængden og den kritiske mængde ved:

$$binomialTest\left(300, \frac{1}{6}, 0.05, højre\right)$$



Som det ses er:

Acceptmængden: $A = \{0,1,2,...,61\}$

Kritisk mængde: $K = \{62,63,64,...,300\}$

De 67 5'ere ligger altså i den kritiske mængde.

Opgaverne 436*

Meningsmålinger af tilslutningen til politiske partier er et af de steder, hvor vi oftest møder statistiske tests præsenteret, og det er også en situation, hvor man skal passe på ikke at gå i en "fælde" (jf. fremlæggelse 10 *Multiple comparisons problem*, side 78). For hvis der f.eks. er 10 partier, og man kigger på tallene og får øje på et parti, hvor man ser en stor forandring af tilslutningen og derfor beslutter at teste, om ændringen er statistisk signifikant, skal man korrigere for, at man egentlig foretager 10 test (da man udvælger blandt 10 resultater). Hvis vi f.eks. arbejder med et signifikansniveau på 5%, skal vi teste, som om signifikansniveauet var $\frac{5\%}{10} = 0,5\%$ (*Bonferroni-korrektion*). På den måde sikres, at sandsynligheden for at begå en type-I-fejl i det samlede test (bestående af 10 enkelte test) er den samme, som hvis man kun foretog ét test.

Bonferroni-korrektion er kun én blandt flere forskellige metoder til at forsøge at undgå type-I-fejl, når man foretager mange test. Problemet er, at man med denne metode øger risikoen for type-II-fejl, dvs. at falske nulhypoteser IKKE forkastes.

I det næste eksempel antages det altså, at vi – af en eller anden årsag - **fra start** er interesseret i tilslutningen til Det Radikale Venstre, dvs. vi har **IKKE** kigget på meningsmålingen, inden vi beslutter os for at se på netop Det Radikale Venstre.

Eksempel 26: Ved folketingsvalget i 2011 fik Det Radikale Venstre 9,5% af stemmerne. En meningsmåling i januar 2015 fortæller, at de nu står til 7,5% af stemmerne. Spørgsmålet er så, om dette er en signifikant forskel.

Man kan ikke svare på dette, hvis man ikke ved, hvor stor stikprøven er. Så nu antager vi, at stikprøven består af 1300 personer.

Vi skal nu have valgt vores to hypoteser:

H₀: Det Radikale Venstre har samme tilslutning som ved folketingsvalget, dvs. $p_{succes} = 0,095$.

Vi skal nu have bestemt os for vores alternative hypotese. Vi kan se, at tilslutningen ser ud til at være gået ned og kunne derfor være fristede til at lave et venstresidet test.

MEN her er det ekstremt vigtigt at huske på, at man aldrig må vælge alternativ hypotese efter at have set på tallene, da man ellers får dobbelt så mange signifikante resultater i forhold til det rigtige antal.

Nu er det jo lidt for sent at vælge alternativ hypotese inden at have set tallene, men vi lader derfor som om, vi ikke har set resultatet af meningsmålingen og vælger derfor:

H₁: Det Radikale Venstre har ikke samme vælgertilslutning som ved valget, dvs. $p_{succes} \neq 0,095$

Vi placerer altså 2,5% i begge sider (ligesidet test).

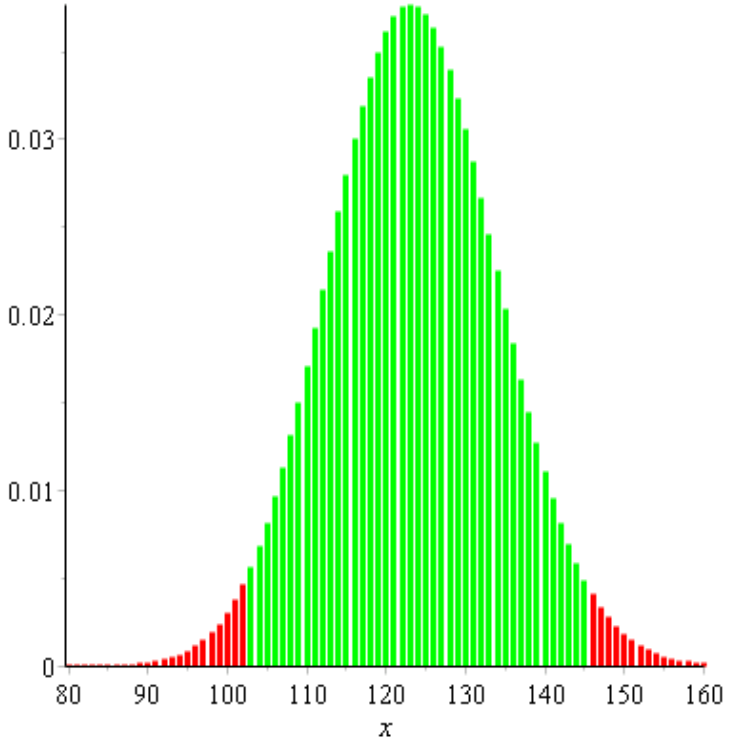
7,5% af 1300 personer svarer til $0,075 \cdot 1300 = 97,5$, dvs. 98 personer (det kunne reelt godt have været 97 personer, da det også ville give 7,5%).

Da vi befinder os på venstre side, skal vi finde sandsynligheden for højst 98 personer:

$$P(X \leq 98) = bincdf(1300, 0.095, 98) = 0.007553422444 < 0,025$$

Da sandsynligheden er under 2,5%, er vælgertilslutningen i stikprøven signifikant mindre end ved valget (nulhypotesen forkastes).

$$binomialTest(1300, 0.095, 0.05, tosidet)$$



Det ses, at acceptmængde og kritisk mængde er:

$A = \{103,104,105,...,145\}$

$K = \{0,1,...,102\} \cup \{146,...,1300\}$

Grænserne 103 og 145 i acceptmængden kan omregnes til procenterne 7,9% og 11,2%.

Dvs. at inden for $[7,9\%;11,2\%]$ ville man ikke have kunnet sige, at vælgertilslutningen var ændret.

Opgaverne 437* og 438* og 439*

χ^2-test

Til grund for alle χ^2-test ligger meget passende de såkaldte χ^2-fordelinger, som vi ser på om lidt. χ^2-fordelingerne er baseret på normalfordelingen (og det er her, at Den Centrale Grænseværdisætning for alvor bliver central), når man anvender χ^2-test.

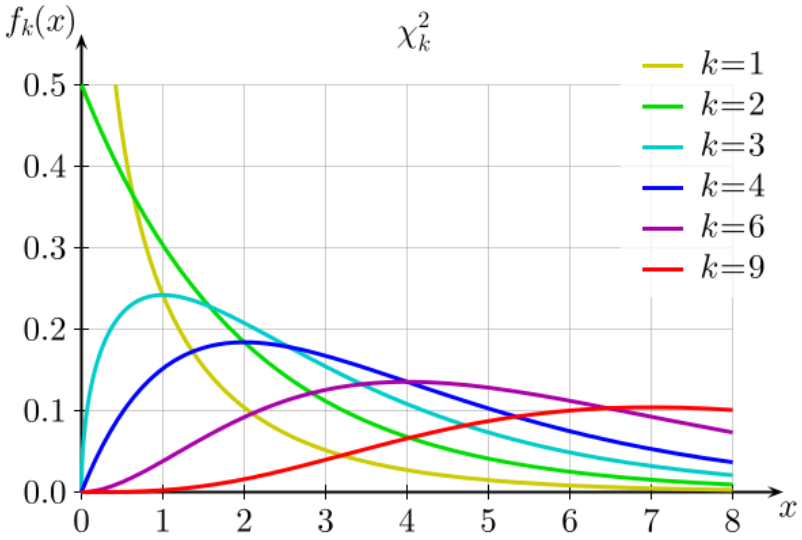
Ved et χ^2-test udregner man en såkaldt teststørrelse Q (sommetider anvendes χ^2 lidt misvisende i stedet for Q):

$$Q = \sum_{i=1}^{n} \frac{(O_i - F_i)^2}{F_i}$$

F_i : Den forventede værdi (beregnet på baggrund af nulhypotesen)
O_i : Den observerede værdi
n: Antallet af observerede kategorier (celler)

Pointen er, at denne teststørrelse Q - **HVIS nulhypotesen er sand** - med god tilnærmelse følger en χ^2-fordeling.

Grafisk ser tæthedsfunktionerne for en del af χ^2-fordelinger ud som vist nedenfor.



k-værdierne angiver antallet af *frihedsgrader*.
Q-værdien skal sammenlignes med et tal aflæst på førsteaksen.

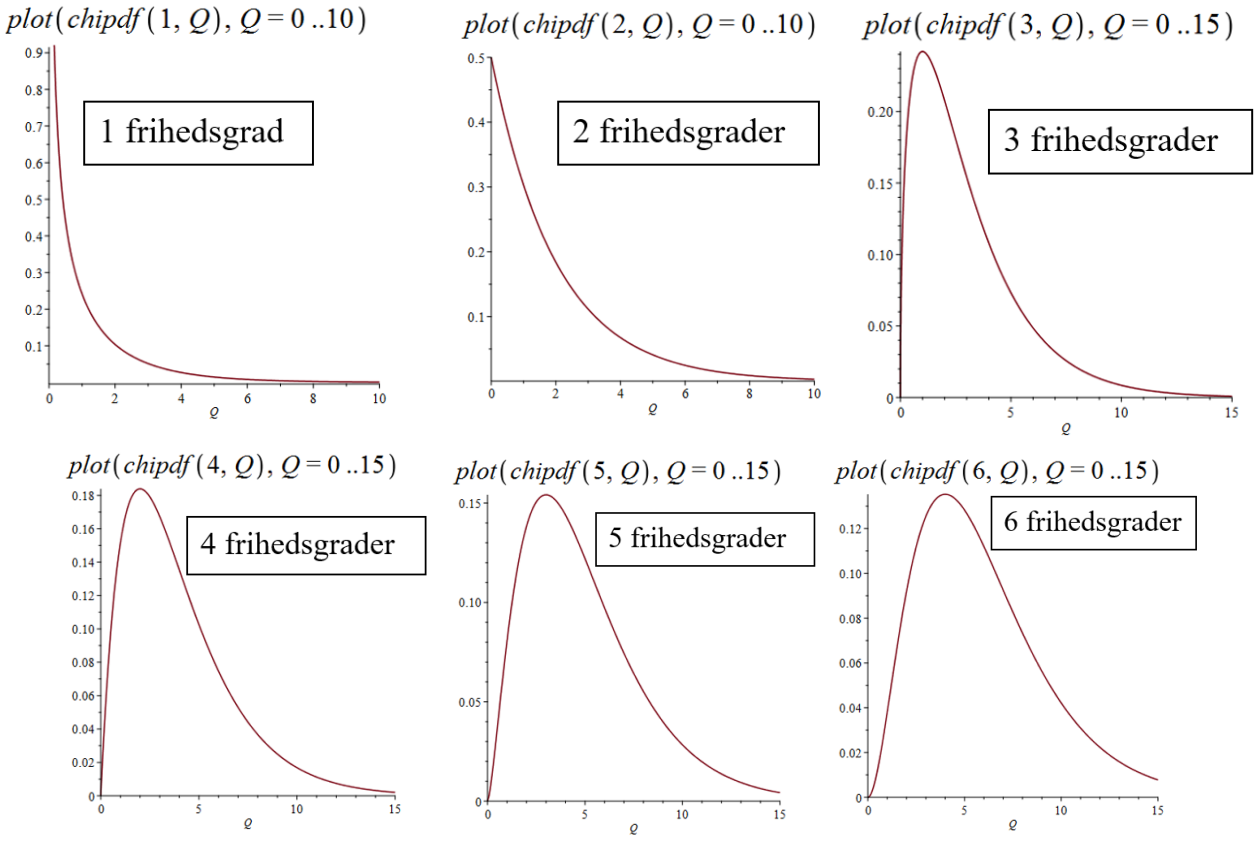
Antallet af frihedsgrader er det antal kategorier (observerede værdier), der frit kan varieres, eller sagt med andre ord, *antallet af celler man skal kende observationerne i, før man kan udregne resten af tabellen*.

Denne beskrivelse bliver nemmere at forstå, når vi ser på eksemplerne.

Bemærk, at arealet under hver af χ^2-fordelingerne (selvfølgelig) er 1 (100%).
Jo mindre Q-værdien er, jo mere støtter det nulhypotesen (se udregningen af Q-værdien, hvor Q bliver mindre, jo tættere de observerede værdier ligger på de forventede værdier, der er beregnet med udgangspunkt i nulhypotesen).

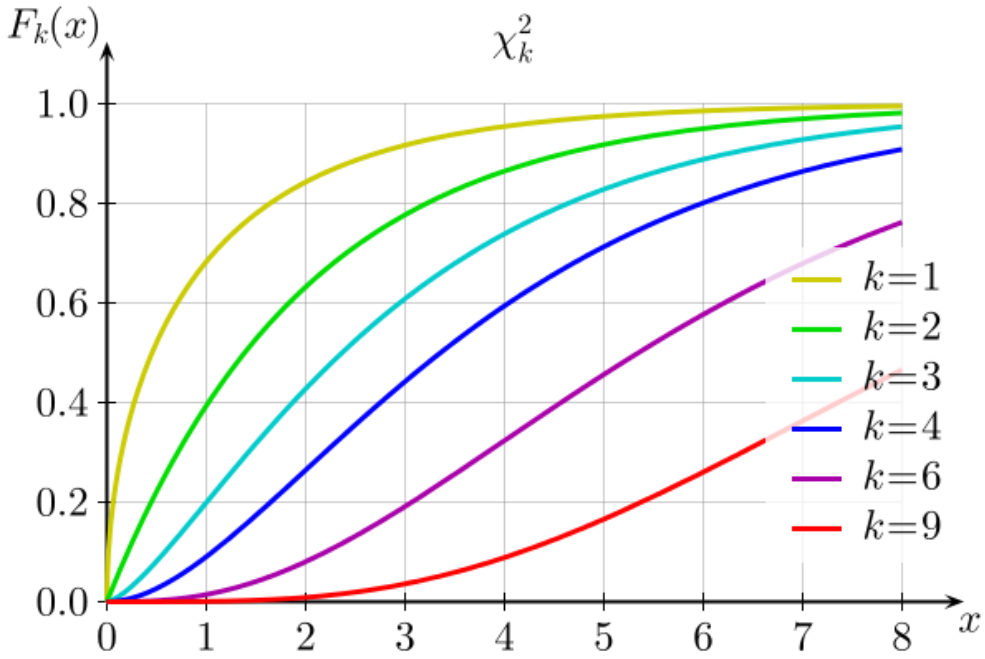
Opgaverne 440*

Med Maples Gym-pakke kan man grafisk afbilde tæthedsfunktionerne for χ^2-fordelingerne:



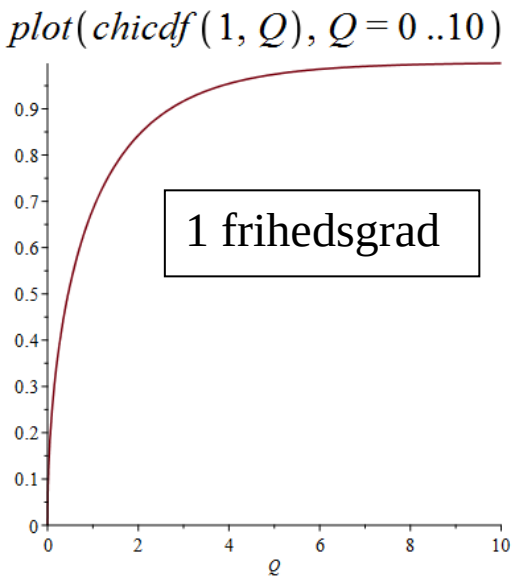
Signifikansniveauet α kan omregnes til en værdi på 1. aksen ved at finde den værdi på 1.aksen, hvor arealet under grafen til højre for denne værdi svarer til α.

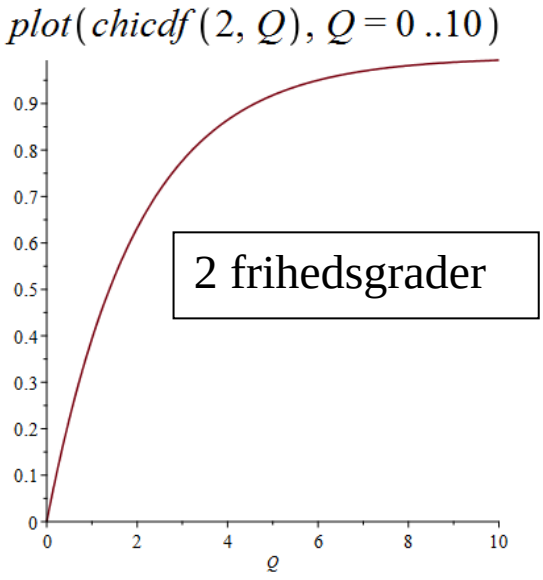
Dette foregår nemmest ved at anvende χ^2-**fordelings**funktionerne:



Husk, at det nu er sandsynlighederne for højst den pågældende værdi, der er angivet på 2. aksen, dvs. med f.eks. 6 frihedsgrader, er sandsynligheden for, at *Q*-værdien er 6 eller under, ca. 58%.

I Gym-pakken tegnes det ved:





Det er vigtigt at huske på, at *tæthedsfunktionerne* IKKE har sandsynligheder ud ad 2. aksen, men at sandsynlighederne fremkommer som **arealer** under grafen. Dvs. man kan på grafen med det passende antal frihedsgrader finde sandsynligheden for at få højst en bestemt Q-værdi ved at beregne arealet under grafen i intervallet $[0, Q]$

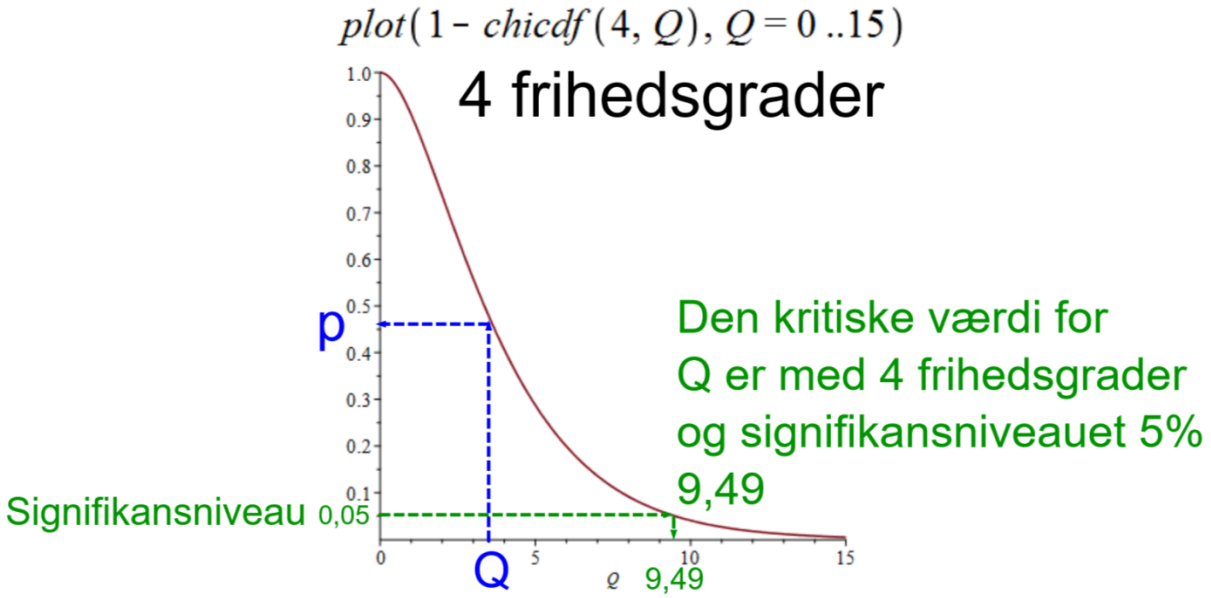
Fordelingsfunktionerne angiver derimod lige netop denne værdi. Dvs. vi kan direkte på 2. aksen aflæse sandsynligheden for højst at få den pågældende Q-værdi.

Omregning mellem p og Q

Vi har indtil videre kigget på sandsynligheder for at få *højst* en bestemt værdi. Men definitionen på signifikansniveau gør, at vi er interesseret i sandsynligheden for at få *mindst* den pågældende værdi.

Vi skal derfor ikke kigge på fordelingsfunktionerne $F(Q)$, men på $1-F(Q)$.

Vi ser nu på et eksempel med 4 frihedsgrader og signifikansniveauet 5%:



Bemærk, at jo større Q-værdien bliver, jo mindre bliver sandsynligheden p, da denne graf netop angiver sandsynligheden for - under forudsætning af nulhypotesens gyldighed - at få *mindst* den pågældende Q-værdi.

Du skal ud fra denne figur kunne forstå følgende:

- p-værdi: Hvis $p < \alpha$ (signifikansniveauet), forkastes nulhypotesen.
- Q-værdi: Hvis $Q > Q_{kritisk}$ (den kritiske værdi), forkastes nulhypotesen.

Man kan regne frem og tilbage mellem p-værdien og Q-værdien ved:
$p = 1 - chicdf(antal\ frihedsgrader, Q)$

Eksempel 27: $p = 1 - chicdf(6, 5.43) = 0.489957398076201$
Dvs. at hvis man har 6 frihedsgrader, er sandsynligheden 49% for - under forudsætning af at nulhypotesen er sand - at få en Q-værdi på mindst 5,43.

Eksempel 28: $fsolve(0.05 = 1 - chicdf(5, Q)) = 11.07049769$
Dvs. hvis man har 5 frihedsgrader og et signifikansniveau på 5%, er den kritiske værdi for Q-værdien 11,07.

Opgaverne 441*

En vigtig tabel

Det sidste eksempel viser fremgangsmåden til at udregne nedenstående vigtige tabel:

Degrees of freedom (df)	Q-value										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

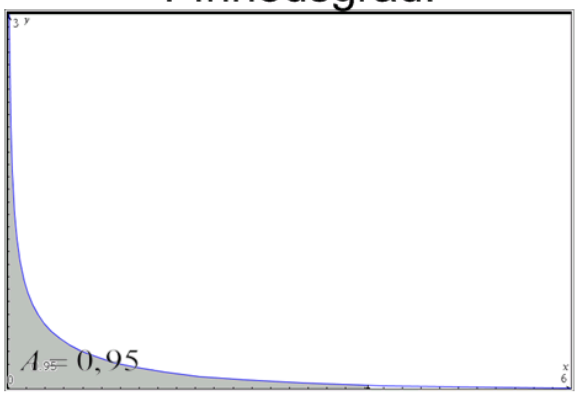
Den kritiske værdi for Q-værdien

Signifikansniveau

Tabellen viser, at hvis vi vælger signifikansniveauet 5% $(\alpha = 0,05)$ og har 3 frihedsgrader, vil værdien 7,82 på 1. aksen fungere som Q-værdiens grænse for, hvornår nulhypotesen forkastes.
Hvis Q-værdien er større end 7,82, forkastes nulhypotesen.
Hvis Q-værdien er mindre end 7,82, forkastes nulhypotesen ikke.

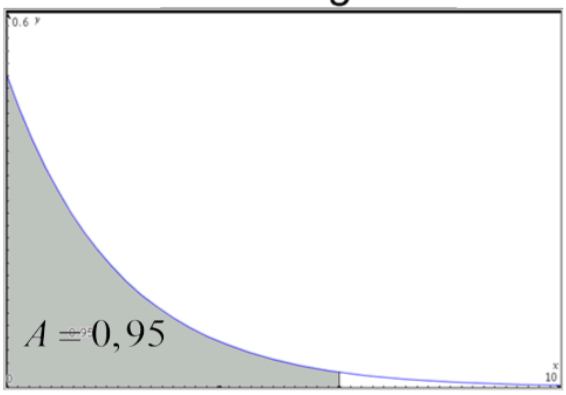
Med <u>signifikansniveauet 5%</u> har man ifølge tabellen følgende grænser:

1 frihedsgrad.



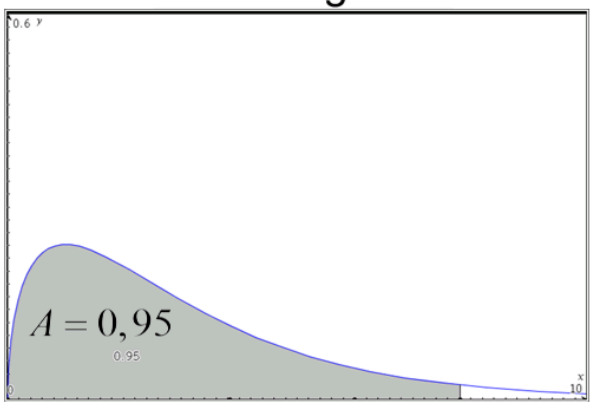
3,84

2 frihedsgrader



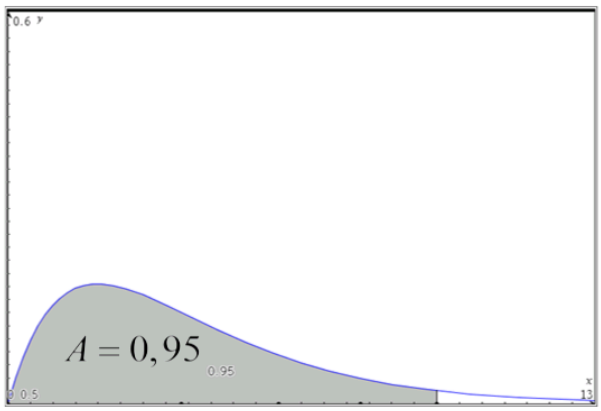
5,991

3 frihedsgrader



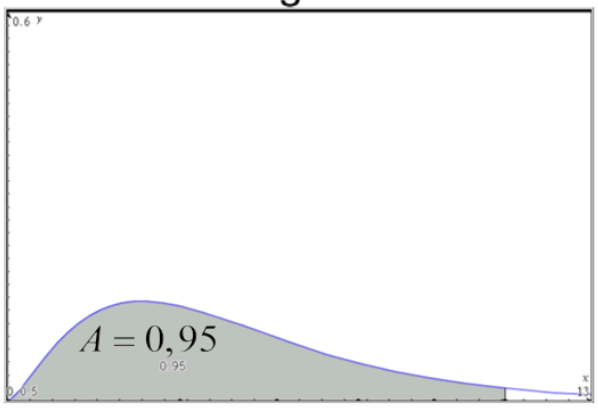
7,815

4 frihedsgrader



9,49

5 frihedsgrader



11,07

Tallene på førsteaksen (3.84, 5.991, 7.815, 9.49 og 11.07) angiver grænsen for vores *Q*-værdi, når vi arbejder med et signifikansniveau på $\alpha = 0,05 = 5\%$.

Hvis signifikansniveauet α gøres større, bliver grænsen mindre - og omvendt.

Man har altså to forskellige måder at komme frem til en konklusion på:

- *p*-værdi: Hvis $p < \alpha$, forkastes nulhypotesen.
- *Q*-værdi: Hvis $Q > Q_{kritisk}$, forkastes nulhypotesen.

χ^2-test (chi-i-anden-test) GOF

I Maples Gym-pakke: *ChiKvadratGOFtest*

GOF står for *Goodness Of Fit*. Testet kaldes også sommetider Pearsons χ^2-GOF-test (der findes også en masse andre χ^2-test).

I et GOF-test undersøger man, om et observationssæt er i overensstemmelse med en teoretisk eller forventet fordeling.
Man arbejder altid ud fra hypoteserne:
H_0: Observationssættet er i overensstemmelse med den forventede fordeling.
H_1: Observationssættet er ikke i overensstemmelse med den forventede fordeling.

Der er ikke noget med højresidet eller venstresidet test, da χ^2-fordelingerne er en slags "kvadreret u-fordeling", hvorfor man ikke kan skelne mellem positive og negative afvigelser fra middelværdien.

Fremgangsmåden er så:
a) Vælg signifikansniveau og bestem antal frihedsgrader.
b) Opstil en tabel med observerede og forventede værdier (sidstnævnte på baggrund af nulhypotesen).
c) Udregn *Q*-værdien.
d) Sammenlign *Q*-værdien med tabellen og se, om nulhypotesen skal forkastes (hvis *Q*-værdien er **større** end den kritiske værdi, skal nulhypotesen forkastes). Evt. kan *p*-værdien anvendes i stedet for *Q*-værdien, hvor en *p*-værdi **mindre** end signifikansniveauet fører til forkastelse af nulhypotesen.

Eksempel 29:

Ved valget i 2011 fordelte stemmerne sig på følgende måde:

Parti	A	B	C	F	I	K	O	V	Ø
%-del	24,8	9,5	4,9	9,2	5,0	0,8	12,3	26,7	6,7

En meningsmåling 29. januar 2015 med 1682 repræsentativt udvalgte danskere viser nu:

Parti	A	B	C	F	I	K	O	V	Ø
%-del	22,9	7,1	4,4	6,7	5,0	0,6	21,2	23,6	8,5

Vi vil gerne undersøge, om vælgertilslutningen har ændret sig.

a) Vi sætter signifikansniveauet til 5%, og da der er 9 partier, er antallet af frihedsgrader 8 ($9-1=8$), da vi frit kan vælge 8 %-dele, hvorefter den sidste procentdel er låst fast af betingelsen om, at procentdelene summeres op til 100%.

b) Det er vigtigt at bemærke, at man ikke kan arbejde med %-satser i χ^2-test, så disse skal omregnes til forventede og observerede værdier. De forventede værdier udregnes med udgangspunkt i nulhypotesen (dvs. valgresultatet), og de observerede værdier beregnes ud fra meningsmålingen.
I begge tilfælde omregnes fra % til antal ved hjælp af de 1682 adspurgte:
Eksempel: Forventet F: $1682 \cdot 9,2\% = 1682 \cdot 0,092 = 154,744 \approx 155$
Eksempel: Observeret V: $1682 \cdot 23,6\% = 1682 \cdot 0,236 = 396,952 \approx 397$

Parti	A	B	C	F	I	K	O	V	Ø
Forventet	417	160	82	155	84	13	207	449	113
Observeret	385	119	74	113	84	10	357	397	143

c) $$Q=\sum_{i=1}^{9}\frac{(O_i-F_i)^2}{F_i}=\frac{(385-417)^2}{417}+\frac{(119-160)^2}{160}+\frac{(74-82)^2}{82}+\frac{(113-155)^2}{155}+$$
$$\frac{(84-84)^2}{84}+\frac{(10-13)^2}{13}+\frac{(207-357)^2}{357}+\frac{(449-397)^2}{397}+\frac{(113-143)^2}{143}=148,4979$$

d) Vores vigtige tabel fortæller os, at med 8 frihedsgrader og signifikansniveauet 0,05, er den kritiske værdi for Q 15,51.

Da $148,4979>15,51$, forkastes nulhypotesen. Eller med andre ord: Der er signifikant forskel på valgresultatet og meningsmålingen.

Vi kan også beregne en p-værdi (hvor det udnyttes, at der er 8 frihedsgrader):

$$p=1-chicdf(8,\ 148.4979)=0.$$

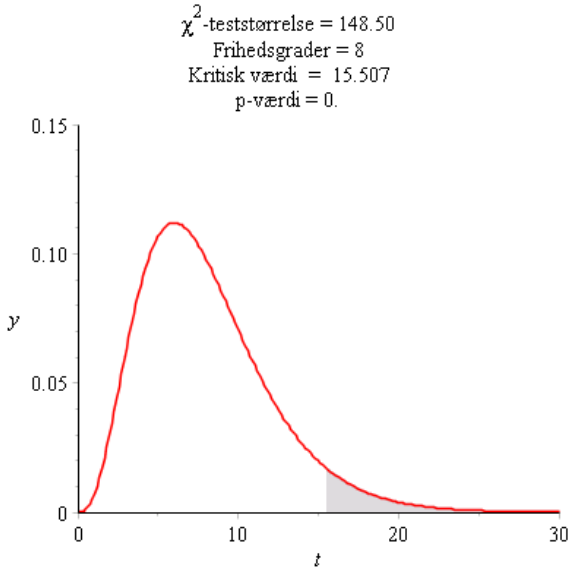
Vi får altså et tal, der er så tæt på 0, at Maple ikke kan angive det.

Da $p<5\%$, forkastes nulhypotesen.

Bemærk: Man skal selvfølgelig ikke normalt anvende både Q og p til at afgøre, om nulhypotesen forkastes. Man anvender én af dem (efter eget valg).

I Maples Gym-pakke kan testet udføres, når man har udregnet tabellerne i b):

```
with(Gym) :
obs := [385, 119, 74, 113, 84, 10, 357, 397, 143] :
forv := [417, 160, 82, 155, 84, 13, 207, 449, 113] :
ChiKvadratGOFtest(obs, forv, level = 0.05)
```



Her anvendes betegnelsen χ^2 for teststørrelsen Q, men ellers kan man se, at værdierne er de samme.

Endnu en detalje: Teststørrelsen Q er som nævnt som udgangspunkt med god tilnærmelse χ^2-fordelt. Denne tilnærmelse er dog ikke så god, hvis nogle af de **forventede** værdier er meget små. Man har for de forventede værdier fastsat værdien 5 som nedre grænse for, hvornår det er rimeligt at antage, at Q følger χ^2-fordelingen.

I vores eksempel 1 er den mindste forventede værdi 13 (partiet K), så her er der ikke problemer. Men hvis man i en eksamensopgave bliver "tvunget" til at lave et χ^2-test i en situation med en eller flere værdier under 5, bør man kommentere dette problem.

Eksempel 30:

Vi vil undersøge, om vores nyindkøbte terning er skæv. Vores nulhypotese er altså, at terningen **ikke** er skæv, dvs. at sandsynligheden for hvert udfald er $p = \frac{1}{6}$, mens den alternative hypotese er, at terningen er skæv. Vi vælger signifikansniveauet 1%, kaster terningen 600 gange og får:

Øjental	1	2	3	4	5	6
Antal	96	98	105	101	95	105

a) Antallet af frihedsgrader er 5, da vi kan udregne antallet af 6'ere, når vi kender antallet af de 5 andre øjental og ved, at der var i alt 600 kast.

b) Vi har allerede den observerede tabel ovenfor, og den forventede tabel er:

Øjental	1	2	3	4	5	6
Antal	100	100	100	100	100	100

c) $Q = \sum_{i=1}^{6} \frac{(O_i - F_i)^2}{F_i} =$

$$\frac{(96-100)^2}{100} + \frac{(98-100)^2}{100} + \frac{(105-100)^2}{100} + \frac{(101-100)^2}{100} + \frac{(95-100)^2}{100} + \frac{(105-100)^2}{100} = 0{,}96$$

d) Med 5 frihedsgrader og et signifikansniveau på 1% fortæller tabellen os, at grænsen for Q-værdien er 15,09.

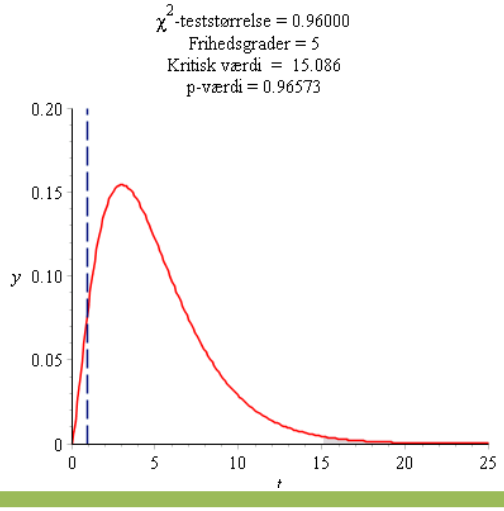
Da $0{,}96 < 15{,}09$, forkastes nulhypotesen IKKE. Dvs. vi har ikke belæg for at hævde, at terningen er skæv, da vores observationer ikke afviger signifikant fra det forventede.

Vi kunne også have udregnet *p*-værdien:

$$p = 1 - chicdf(5, 0.96) = 0.965727098837717$$

Man kan kort konkludere: "Da $p > 1\%$, forkastes nulhypotesen IKKE." Med lidt flere ord kan man sige, at hvis vores nulhypotese er sand (dvs. terningen ikke er skæv), er der 96,6% chance for at få er resultat som vores eller et, der er værre (dvs. som peger mod en skæv terning). Vores resultat er altså på ingen måde usædvanligt, og derfor forkastes nulhypotesen ikke.

```
with(Gym) :
obs := [96, 98, 105, 101, 95, 105] :
forv := [100, 100, 100, 100, 100, 100] :
ChiKvadratGOFtest(obs, forv, level = 0.01)
```



Opgaverne 445*

χ^2-test (chi-i-anden-test) Uafhængighedstest

I Maples Gym-pakke: *ChiKvadratUtest*

Dette er også et af Pearsons χ^2-test.

Det anvendes til at teste, om to forskellige størrelser er uafhængige.
Hypoteserne er derfor altid:
H_0: De to størrelser er uafhængige af hinanden.
H_1: De to størrelser er afhængige af hinanden.

Eksempel 31: Man ønsker at undersøge, om der er en sammenhæng mellem elevers præstationer i matematik og fysik. Man har en formodning om, at der er en klar sammenhæng, og man ønsker at vise dette, hvorfor man 'satser' og vælger et signifikansniveau på 0,1%. Man undersøger derefter 528 studenters præstationer og opdeler dem inden for hvert fag i 'høj karakter', 'mellem karakter' og 'lav karakter'.

Den observerede tabel bliver:

		Matematik			
		Høj	Mellem	Lav	I alt
Fysik	Høj	56	71	12	139
	Mellem	47	163	38	248
	Lav	14	42	85	141
	I alt	117	276	135	528

Der er tilføjet en række og en søjle med 'I alt', hvor tallene er fundet ved at tage summen af tallene i den tilsvarende række/søjle. Nederst i højre hjørne fås det samlede antal studenter i undersøgelsen både ved at lægge de tre røde tal ovenover sammen og ved at lægge de tre røde tal til venstre sammen. På denne måde kan man tjekke, om man har regnet forkert.

Vi ønsker nu at opstille en forventet tabel og i samme forbindelse se på antallet af frihedsgrader. Den forventede tabel er som bekendt baseret på nulhypotesen, dvs. vi går ud fra, at karaktererne i matematik og fysik er uafhængige af hinanden.
Som eksempel ser vi på det forventede antal elever, der skulle få en Mellem-karakter i matematik og en Høj-karakter i fysik.

Vi kan gribe det an fra to forskellige synsvinkler:

1. synsvinkel: Der er 276 ud af de 528 elever, der har fået Mellem i matematik, dvs. $\frac{276}{528} = 52{,}3\%$.

Der er 139 elever, der har fået Høj i fysik, og hvis der ikke er nogen sammenhæng mellem karaktererne i fysik og matematik, skulle 52,3% af disse have fået karakteren Mellem i matematik, dvs. antallet af elever med karakteren Mellem i matematik og Høj i fysik måtte forventes at være:
$52{,}3\% \cdot 139 = 0{,}523 \cdot 139 = 72{,}7 \approx 73$

2. synsvinkel: Der er 139 ud af de 528 elever, der har fået Høj i fysik, dvs. $\frac{139}{528} = 26{,}3\%$.

Der er 276 elever, der har fået Mellem i matematik, og hvis 26,3% af disse også har fået Høj i fysik, bliver det forventede antal med Mellem i matematik og Høj i fysik:
$26{,}3\% \cdot 276 = 0{,}263 \cdot 276 \approx 73$

Hvis man kigger lidt på udregningerne, kan man se, hvorfor resultatet bliver det samme.

Vi begynder nu at udfylde den forventede tabel:

		Matematik			
		Høj	Mellem	Lav	I alt
Fysik	Høj	$\frac{139\cdot 117}{528}=31$	$\frac{139\cdot 276}{528}=73$		139
	Mellem	$\frac{248\cdot 117}{528}=55$	$\frac{248\cdot 276}{528}=130$		248
	Lav				141
	I alt	117	276	135	528

Inden vi fortsætter, kan vi nu se på antal frihedsgrader, når man arbejder med tabeller. For som det fremgår af ovenstående, er vi - da vi kender alle de røde tal - nu i stand til at beregne de manglende 5 værdier:

Eksempler:

Høj-mat og Lav-fys: $Antal = 117-31-55=31$

Lav-mat og Mellem-fys: $Antal = 248-55-130=63$

Vi har altså 4 frihedsgrader i dette eksempel, da vi ud fra disse fire værdier er i stand til at beregne resten.

Generelt gælder det for en tabel med *r* rækker og *s* søjler, at antallet *f* af frihedsgrader er:

$$f=(r-1)\cdot(s-1)$$

Den forventede tabel udfyldes helt:

		Matematik			
		Høj	Mellem	Lav	I alt
Fysik	Høj	31	73	35	139
	Mellem	55	130	63	248
	Lav	31	73	37	141
	I alt	117	276	135	528

Og den observerede var som angivet tidligere:

		Matematik			
		Høj	Mellem	Lav	I alt
Fysik	Høj	56	71	12	139
	Mellem	47	163	38	248
	Lav	14	42	85	141
	I alt	117	276	135	528

Vi beregner så vores *Q*-værdi:

$$Q=\frac{(56-31)^2}{31}+\frac{(71-73)^2}{73}+\frac{(12-35)^2}{35}+\frac{(47-55)^2}{55}+\frac{(163-130)^2}{130}+$$
$$\frac{(38-63)^2}{63}+\frac{(14-31)^2}{31}+\frac{(42-73)^2}{73}+\frac{(85-37)^2}{37}=139,55$$

Vi har 4 frihedsgrader og arbejder som nævnt fra start med signifikansniveauet 0,1%. Vores tabel fortæller os så, at den kritiske værdi for vores *Q*-værdi er 18,47.

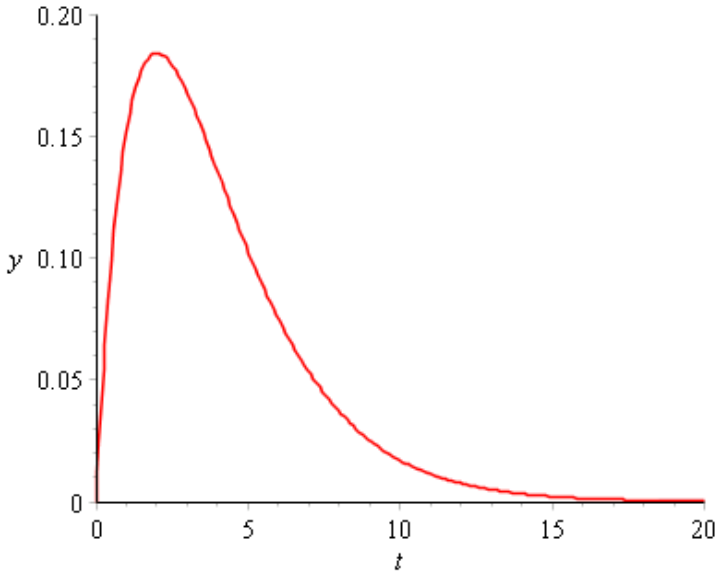
Da $139,55>18,47$, forkastes nulhypotesen, dvs. der er en signifikant sammenhæng mellem karaktererne i matematik og fysik.

I Maple skal man opskrive sin tabel i en matrix. Benyt "tab" til at bevæge dig frem mellem cellerne:

$$M := \begin{bmatrix} 56 & 71 & 12 \\ 47 & 163 & 38 \\ 14 & 42 & 85 \end{bmatrix}:$$

$$ChiKvadratUtest(M,\ level = 0.001)$$

χ^2-teststørrelse = 145.78
Frihedsgrader = 4
Kritisk værdi = 18.467
p-værdi = 0.



Teststørrelsen afviger (lidt) fra vores beregnede værdi, men det skyldes de afrundinger, vi foretog undervejs.

Med Maples Gym-pakke kan man finde den forventede tabel med *forventet(...)*:

$$forventet(M) = \begin{bmatrix} 30.801 & 72.659 & 35.540 \\ 54.955 & 129.64 & 63.409 \\ 31.244 & 73.705 & 36.051 \end{bmatrix}$$

Eksemplet viste os også, hvordan vi generelt udregner den forventede tabel ud fra vores beregnede "i alt"-celler:

	A1	A2	A3	A4	I alt
B1	$\frac{B1_{ialt} \cdot A1_{ialt}}{n}$	$\frac{B1_{ialt} \cdot A2_{ialt}}{n}$	$\frac{B1_{ialt} \cdot A3_{ialt}}{n}$	$\frac{B1_{ialt} \cdot A4_{ialt}}{n}$	$B1_{ialt}$
B2	$\frac{B2_{ialt} \cdot A1_{ialt}}{n}$	$\frac{B2_{ialt} \cdot A2_{ialt}}{n}$	$\frac{B2_{ialt} \cdot A3_{ialt}}{n}$	$\frac{B2_{ialt} \cdot A4_{ialt}}{n}$	$B2_{ialt}$
B3	$\frac{B3_{ialt} \cdot A1_{ialt}}{n}$	$\frac{B3_{ialt} \cdot A2_{ialt}}{n}$	$\frac{B3_{ialt} \cdot A3_{ialt}}{n}$	$\frac{B3_{ialt} \cdot A4_{ialt}}{n}$	$B3_{ialt}$
I alt	$A1_{ialt}$	$A2_{ialt}$	$A3_{ialt}$	$A4_{ialt}$	n

Eksempel 32: Vi vil undersøge virkningen af en slags medicin og ser derfor på 1000 personer med den sygdom, som medicinen skal hjælpe imod. Heraf får 500 medicinen, og 500 gør ikke. Vi arbejder med signifikansniveauet 5%.

Vores hypoteser bliver så:
H_0: Sygdomstilstanden er uafhængig af, om personen har fået medicin eller ej.
H_1: Sygdomstilstanden afhænger af, om personen har modtaget medicin eller ej.

Det er værd at bemærke, at nulhypotesen forkastes, hvis personerne får det signifikant værre pga. medicinen, så hvis der fremkommer signifikans, skal man ved at kigge på tallene i tabellen se, om medicinen har en positiv effekt.

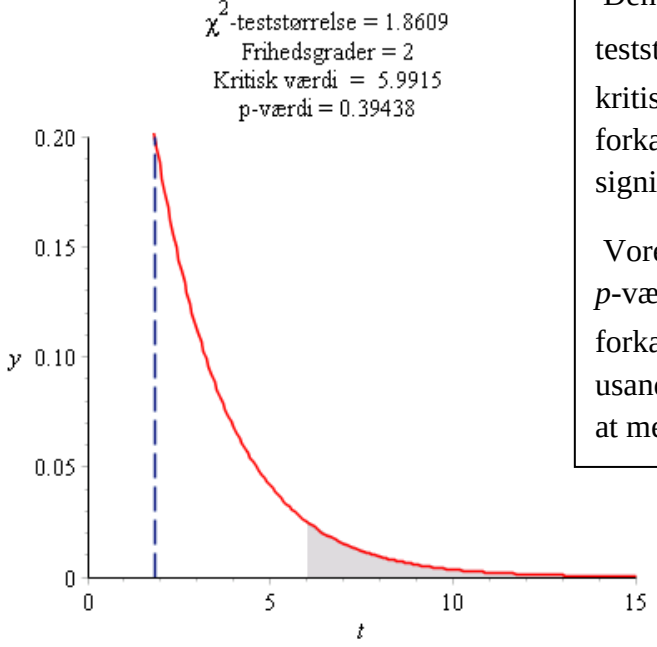
Vi har fået følgende tabel (hvor vi selv har udregnet "i alt"):

	Ingen symptomer	Svage symptomer	Syge	I alt
Fået medicin	300	80	120	500
Ingen medicin	279	86	135	500
I alt	579	166	255	1000

Vi lader Maples Gym-pakke foretage udregningerne:

$$M := \begin{bmatrix} 300 & 80 & 120 \\ 279 & 86 & 135 \end{bmatrix}:$$

$$ChiKvadratUtest(M,\ level = 0.05)$$



Maple giver os alle de værdier, vi skal bruge til noget (dvs. vi behøver ikke at anvende vores tabel).

Den kritiske værdi er 5,9915, og da vores teststørrelse Q (eller χ^2) er mindre end den kritiske værdi, skal nulhypotesen IKKE forkastes. Dvs. medicinen har ikke nogen signifikant virkning.

Vores signifikansniveau er 5%, og da vi har en p-værdi på 39%, er $p > 5\%$, så nulhypotesen forkastes IKKE. Resultatet er ikke så usandsynligt, at vi kan forkaste hypotesen om, at medicinen er virkningsløs.

Opgaverne 450*

t-test (Student's t-test)

Navnet skyldes, at William S. Gosset beskrev testene i artikler skrevet under pseudonymet Student.

t-test anvendes, når man har en forholdsvis lille stikprøve af en eller to størrelser, der formodes at være normalfordelt, men hvor man hverken kender middelværdien μ eller spredningen σ.

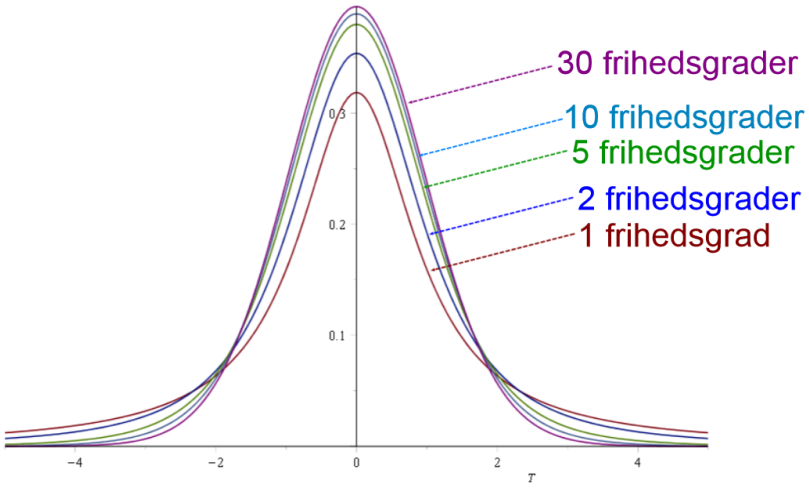
Hvis stikprøven er stor, og/eller man kender spredningen, skal man anvende *z*-test (baseret på normalfordelingen).

Til grund for *t*-testet ligger de såkaldte t-fordelinger, og vi udregner lige som tidligere teststørrelser - kaldet *T* - der, hvis nulhypotesen er sand, følger *t*-fordelingen med det passende antal frihedsgrader.

Bortset fra nye fordelinger og teststørrelser er fremgangsmåden nogenlunde den samme som ved χ^2-test.

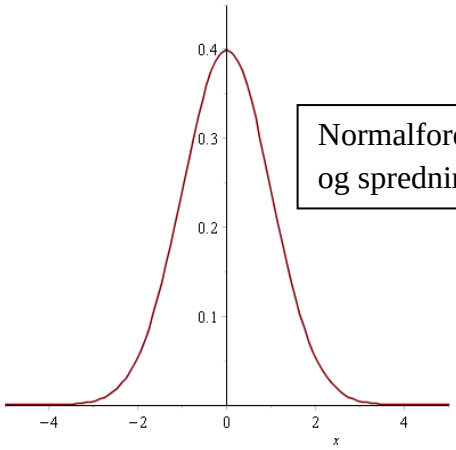
Nogle *t*-fordelingers *tæthedsfunktioner* ses her (fordelingsfunktionerne kan tegnes med *tcdf*).

$$plot([tpdf(1,T), tpdf(2,T), tpdf(5,T), tpdf(10,T), tpdf(30,T)], T=-5..5)$$



Bemærk, at tæthedsfunktionerne er symmetriske omkring 0, og at vores teststørrelser altså i dette tilfælde også kan blive negative.

En anden ting, der kan bemærkes, er, at jo flere frihedsgrader, jo tættere kommer t-fordelinger på normalfordelingen med middelværdi 0 og spredning 1, dvs. den såkaldte *u-fordeling*:



Normalfordeling med middelværdi 0 og spredning 1 (*u*-fordelingen).

Vi er her kommet tilbage til en situation, hvor vi skal vælge alternative hypoteser og kan vælge mellem *venstresidet*, *højresidet* og *ligesidet/tosidet/dobbeltsidet* test.

t-test: One-Sample-t-test

Dette test anvendes, når man vil undersøge, om en række målinger af en bestemt størrelse, der kan formodes at være normalfordelt, men hvor spredningen ikke kendes, har en forventet middelværdi (der f.eks. kunne være en tabelværdi).

Det giver altså følgende nulhypotese:

H_0: $\mu_{obs} = \mu_0$

Der er følgende valgmuligheder for alternativ hypotese (husk, den skal vælges inden målingerne udføres):

1) $H_1 : \mu_{obs} < \mu_0$ (venstresidet test, dvs. hele den kritiske mængde placeres under μ_0)

2) $H_1 : \mu_{obs} > \mu_0$ (højresidet test, dvs. hele den kritiske mængde placeres over μ_0)

3) $H_1 : \mu_{obs} \neq \mu_0$ (ligesidet test, dvs. den kritiske mængde fordeles på begge sider af μ_0)

Teststørrelsen T er i dette tilfælde:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Her er:

n: Antallet af målinger.

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ er det aritmetiske gennemsnit af de observerede værdier.

$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$ er den estimerede spredning.

$f = n - 1$ er antallet af frihedsgrader.

Antallet af frihedsgrader samt den lidt overraskende brøk $\frac{1}{n-1}$ i udtrykket for den estimerede spredning kommer af, at man "mister" en frihedsgrad, når man udregner det aritmetiske gennemsnit. For givet det aritmetiske gennemsnit, behøver man kun at kende $n-1$ værdier, da den sidste værdi kan beregnes ud fra disse og gennemsnittet.

Det er teststørrelsen *T*, der - **hvis nulhypotesen er sand** - vil følge *t*-fordelingen med de passende antal frihedsgrader.

Bemærk, at udtrykket for teststørrelsen *T* svarer til at normere en fordeling med middelværdien μ_0 og spredningen $\frac{s}{\sqrt{n}}$ (se Definition 10 og Sætning 10 i *Sandsynlighedsregning og kombinatorik*).

Spredningen $\frac{s}{\sqrt{n}}$ kommer fra Den Centrale Grænseværdisætning.

Eksempel 33: Ved et forsøg med faldende tennisbolde har vi forsøgt at undersøge, om tyngdeaccelerationen er $9{,}82\,\frac{m}{s^2}$. Vi vælger et ligesidet test og arbejder med signifikansniveauet 5%. Vi udfører forsøget 9 gange (dvs. antallet af frihedsgrader er 8):

Målt værdi i enheden $\frac{m}{s^2}$	9,69	9,81	9,57	10,02	9,48	9,94	9,21	9,54	9,37

Vi udregner teststørrelsen ved hjælp af Maple:

```
restart
with(Gym) :
g := [9.69, 9.81, 9.57, 10.02, 9.48, 9.94, 9.21, 9.54, 9.37] :
```

$$g_{gen} := \frac{\sum_{i=1}^{9} g_i}{9} = 9.625555556$$

$$s := \sqrt{\frac{1}{8} \cdot \sum_{i=1}^{9} \left(g_i - g_{gen}\right)^2} = 0.2650995620$$

$$T := \frac{g_{gen} - 9.82}{\frac{s}{\sqrt{9}}} = -2.200431142$$

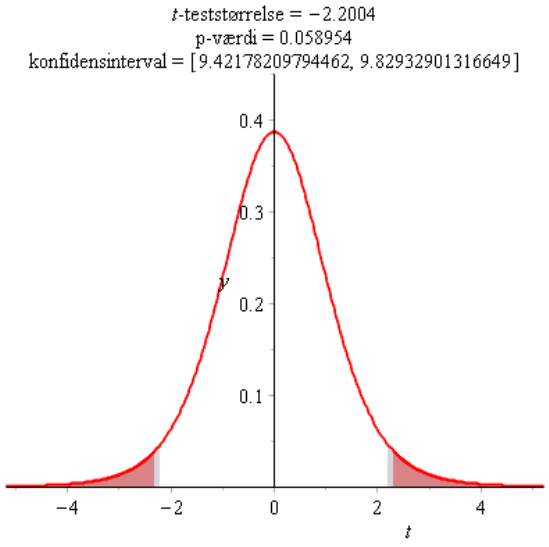
Vi har valgt et ligesidet test med signifikansniveauet 5%, så vi har 2,5% liggende yderst til venstre for 0. Vi kan udregne sandsynligheden for at få højst denne T-værdi (husk, der er 8 frihedsgrader):

$$p_{højst} = tcdf(8, -2.200431142) = 0.02947712904101 14$$

Da denne sandsynlighed er over 2,5%, forkaster vi IKKE nulhypotesen.

Vi kan også lade Maple udregne det hele, når vi allerede har defineret vores tabel.

```
tTest(g, 9.82)
```



Vi ser, at teststørrelsen passer med vores udregnede værdi.

p-værdien passer ikke med vores, men det skyldes, at det er forskellige værdier, der er udregnet. Maple udregner sandsynligheden for at havne mindst 2,2004 fra 0, hvilket giver en dobbelt så stor værdi som vores udregnede sandsynlighed, hvor vi kun ser på venstre side af grafen.

Men Maples *p*-værdi skal sammenlignes med 5%, så konklusionen vil altid blive den samme.

Eksempel 34: Vi måler på niveauet for en kræftcelles genekspression for genet c-Myc.
Vi ønsker at vide, om niveauet er over standardværdien på 100, og arbejder med et signifikansniveau på 3%.
Vi foretager seks målinger (og vi arbejder altså med 5 frihedsgrader):

Målt værdi:	114,6	112,9	98,5	106,7	109,8	103,6

Vores hypoteser bliver altså:
H_0: Gennemsnittet af vores målte værdier er 100
H_1: Gennemsnittet af vores målte værdier er over 100 (højresidet test)
Vi udregner teststørrelsen i Maple:

$$c := [114.6, 112.9, 98.5, 106.7, 109.8, 103.6]:$$

$$c_{gen} := \frac{\sum_{i=1}^{6} c_i}{6} = 107.6833333$$

$$s := \sqrt{\frac{1}{5} \cdot \sum_{i=1}^{6} (c_i - c_{gen})^2} = 6.025086444$$

$$T := \frac{c_{gen} - 100}{\frac{s}{\sqrt{6.}}} = 3.123647483$$

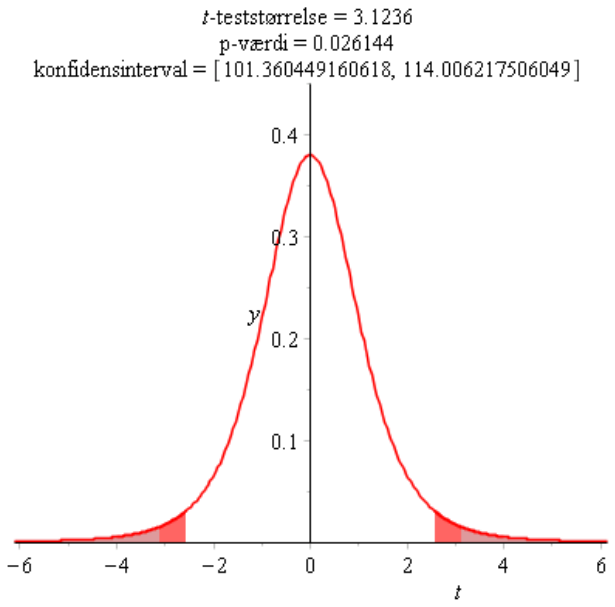
Vi har signifikansniveauet 3% i et højresidet test og har derfor placeret hele den kritiske mængde som de 3% af de normalfordelte udfald af T, der ligger mest over 100.
Vi ønsker derfor at finde ud af sandsynligheden for at få *mindst* vores T-værdi:

$$p_{mindst} = 1 - tcdf(5, 3.123647483) = 0.0130721360293218$$

Da $p_{mindst} < 3\%$, forkastes nulhypotesen. Dvs. niveauet er signifikant højere end standardværdien på 100.

Maples resultat er:

$$tTest(c, 100)$$



Igen ser vi, at teststørrelsen (naturligvis) giver det samme som vores beregning, mens *p*-værdien er dobbelt så stor som vores, da det igen er forskellige sandsynligheder, vi udregner, men også forskellige værdier, vi sammenligner med (de 2,6144% skulle være sammenlignet med 6%).

Opgaverne 465*

t-test: Two-Sample-paired-difference-t-test (parvise observationer)

Dette test anvendes, hvis man har to forskellige observationssæt, hvor man har målt på den samme størrelse, og hvor observationerne i de to sæt hører sammen parvis.

Eksempel 35:

- Hvis man har én gruppe af mennesker og afprøver to forskellige slags medicin, træning, kost eller lignende på alle i gruppen og for hver person måler virkningen af begge slags. Der er så to målinger på hver person, og disse to målinger sættes sammen parvis.
- Hvis man har to vingeprofiler for en vindmølle og tester dem ved en masse forskellige vindhastigheder. Der vil så for hver vindhastighed være to værdier (sikkert effekter), der knyttes sammen parvis.

Hele ideen med dette test er, at man går ind og kigger på differensen *d* (regnet med fortegn) mellem de parvise observationer og derefter regner dette som et one-sample-t-test med nulhypotesen:

H_0: $\mu_{obs} = 0$ (dvs. det aritmetiske gennemsnit af differenserne er 0)

Og igen er der altså tre valgmuligheder for alternativ hypotese:

1) $H_1 : \mu_{obs} < 0$ (venstresidet test, dvs. hele den kritiske mængde placeres under 0)

2) $H_1 : \mu_{obs} > 0$ (højresidet test, dvs. hele den kritiske mængde placeres over 0)

3) $H_1 : \mu_{obs} \neq 0$ (ligesidet test, dvs. den kritiske mængde fordeles på begge sider af 0)

Med udgangspunkt i observationssættene $x_1, x_2, x_3, ..., x_n$ og $y_1, y_2, y_3, ..., y_n$ har man altså:

n: Antallet af parvise observationer.

$d_i = x_i - y_i$: Er differensen mellem de parvise observationer

$T = \dfrac{\overline{d}}{\dfrac{s_d}{\sqrt{n}}} = \dfrac{\overline{d} \cdot \sqrt{n}}{s_d}$ er teststørrelsen, hvor man har:

$\overline{d} = \dfrac{\sum_{i=1}^{n} (x_i - y_i)}{n}$: Det aritmetiske gennemsnit af differenserne.

$s_d = \sqrt{\dfrac{1}{n-1} \cdot \sum_{i=1}^{n} (d_i - \overline{d})^2}$. Den estimerede spredning af differenserne.

$f = n - 1$: Antallet af frihedsgrader

Eksempel 36: Sovemidlerne A og B skal testes på 10 forsøgspersoner, og man måler søvnperioden i timer. Man vil se på, om der er forskel på A og B, og vælger signifikansniveauet 5%:

Person nr.	1	2	3	4	5	6	7	8	9	10
Sovemiddel A	7,7	5,4	6,8	5,8	6,9	10,4	10,7	7,8	7,0	9,0
Sovemiddel B	8,9	7,8	8,1	7,1	6,9	11,4	12,5	8,6	11,6	10,4

Man udregner teststørrelsen i Maple:

$$restart$$
$$with(Gym):$$
$$A := [7.7, 5.4, 6.8, 5.8, 6.9, 10.4, 10.7, 7.8, 7.0, 9.0]:$$
$$B := [8.9, 7.8, 8.1, 7.1, 6.9, 11.4, 12.5, 8.6, 11.6, 10.4]:$$
$$d := B - A = [1.2, 2.4, 1.3, 1.3, 0., 1.0, 1.8, 0.8, 4.6, 1.4]$$
$$d_{gen} := \frac{\sum_{i=1}^{10} d_i}{10} = 1.580000000$$
$$s_d := \sqrt{\frac{1}{10-1} \cdot \sum_{i=1}^{10} \left(d_i - d_{gen}\right)^2} = 1.229995483$$
$$T := \frac{d_{gen} \cdot \sqrt{10.}}{s_d} = 4.062127684$$

Man ser nu på sandsynligheden for at få *mindst* denne T-værdi:

$$p_{mindst} = 1 - tcdf(9, 4.062127684) = 0.0014164450973790 4$$

Da det skulle undersøges, om der er forskel, er det et ligesidet test, der skal foretages, så de 5% placeres i begge ender, dvs. vi skal sammenligne *p*-værdien med 2,5%, og da $0{,}14\% < 2{,}5\%$, må nulhypotesen forkastes. Der er altså signifikant forskel på de to sovemidler, og det ses, at det er B, der virker bedst (T er positiv).

Vi kunne også have set på den kritiske værdi for T:

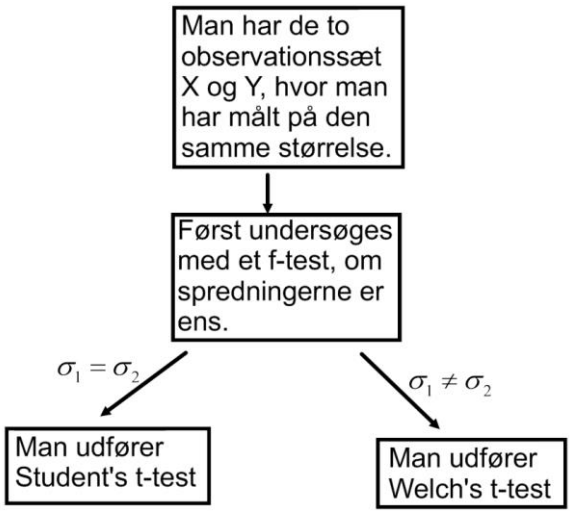
$$0.975 = tcdf(9, T) \xrightarrow{\text{solve for T}} [[T = 2.262157164]]$$

Da $4{,}06 > 2{,}26$, forkastes nulhypotesen.

Opgaverne 470*

t-test: Two-Sample-t-test (ikke-parvise observationer)

Hvis man har to observationssæt, hvor observationerne ikke kan parres, kan man stadig udføre t-test på observationssættene, men det kræver lidt mere forarbejde:



Hver af disse test har egne teststørrelser, og f-testet har (naturligvis) sin egen f-fordeling, der ligger til grund for testet.

Nogle eksempler, hvor disse test kunne anvendes, er:

- Man vil vise, at nogle fugle vejer mere om efteråret end om foråret pga. vinterforberedelse, og fanger og vejer derfor et antal fugle af en bestemt art om foråret og et antal af samme art om efteråret.
- Man vil undersøge virkningen af et sovemiddel og tester derfor på to forskellige grupper af mennesker henholdsvis sovemidlet og et placebo.
- Man vil undersøge virkningen af to forskellige sovemidler, der testes på to forskellige grupper af mennesker.

z-test

Til grund for z-test ligger u-fordelingen (dvs. normalfordelingen med middelværdien 0 og spredningen 1). Dvs. vi kan bruge vores viden om denne fordeling til at vurdere teststørrelsen (f.eks. at 95% ligger inden for afstanden 1,96 på hver side af 0). Som vi så under t-test, kommer t-fordelingerne tættere på ovennævnte normalfordeling, jo større antallet af frihedsgrader bliver, og man kan da også anvende z-test, hvis man har tilpas mange observationer. Der er ingen fast grænse, men omkring 30 observationer kan bruges som en meget løs tommelfingerregel.

Teststørrelsen bliver også den samme som ved T-test, men med en lille, ekstra detalje:

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

n: Antallet af målinger. $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ er det aritmetiske gennemsnit.

$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$ er den estimerede spredning, HVIS man ikke kender denne i forvejen.

HVIS man kender *s* i forvejen, anvender man denne værdi.

Man ville f.eks. anvende et z-test, hvis man kiggede på 2015's målte højder af mænd til session og ville se, om mændenes gennemsnitshøjde havde ændret sig fra en kendt værdi (ca. 180,6 cm).

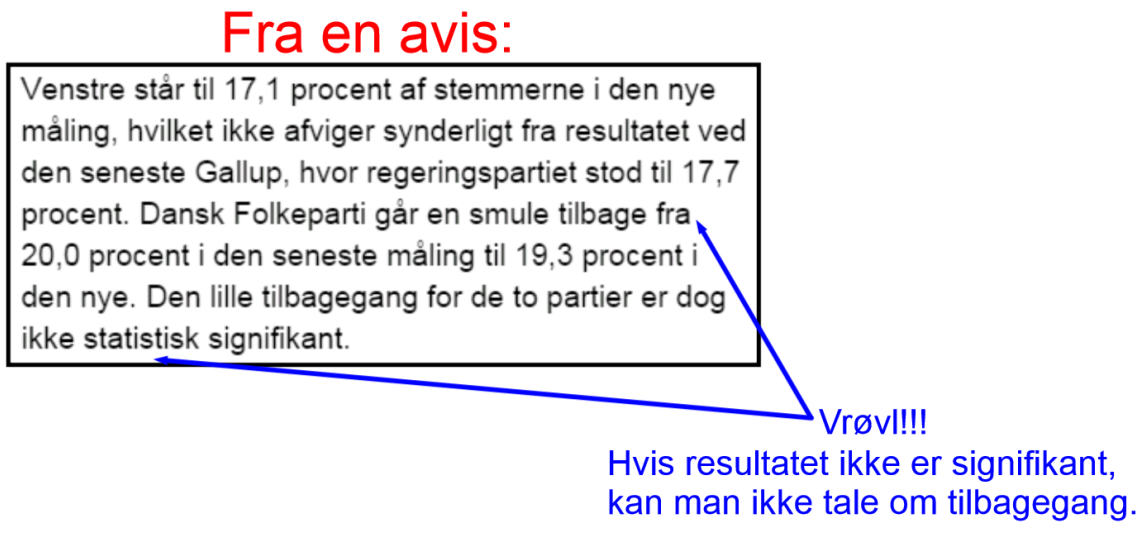
Nogle pointer

- Inden for sandsynlighedsregning anvendes sandsynligheder, og vi kan ved hjælp af sandsynlighedsregning deducere os frem til nogle værdier i konkrete situationer.
- Inden for statistik, når vi anvender stikprøver, arbejder vi med frekvenser, og vi kan slutte induktivt fra egenskaber i stikprøven til egenskaber i populationen.
- **De store tals lov**: Vores frekvenser i stikprøven kan med en vis sandsynlighed komme vilkårligt tæt på sandsynlighederne eller frekvensen i populationen, hvis vi gør stikprøven stor nok.
- Når man laver statistiske test, finder man kun statistiske sammenhænge - ikke årsagssammenhænge.
- **Den Centrale Grænseværdisætning:** Uanset hvilken fordeling man udtager stikprøver fra, vil gennemsnittene af stikprøverne være normalfordelt.

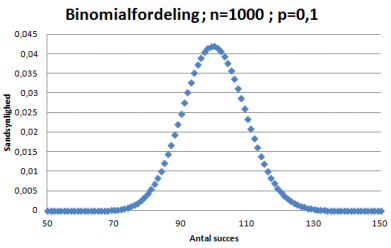
Mere om statistik (gruppeopgaver og fremlæggelser)

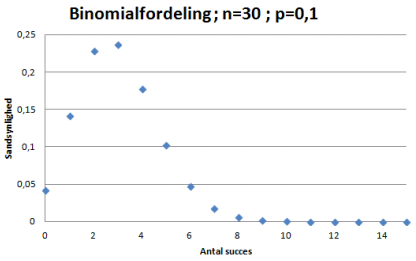
Hver gruppe skal skrive en matematikrapport (opgave), der skal afleveres, og gruppen skal holde et foredrag for klassen. Opgaverne bliver knyttet til et mundtligt eksamensspørgsmål.

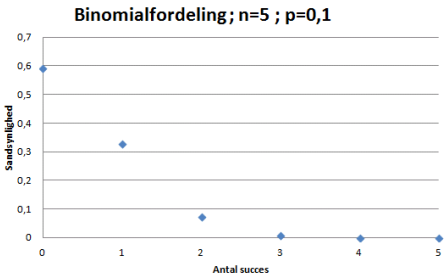
1. Hawthorne-effekten og Placebo-effekten.
2. Pygmalion-effekten (Rosenthal-effekten).
3. Snyd/manipulation med statistik.
4. De små tals lov og Post hoc ergo propter hoc.
5. Prosecutor's fallacy
6. Simpson's paradox (Yule-Simpson-effekt)
7. Artiklen: The Performance Of Mutual Funds In The Period 1945-1964 (Michael Jensen)
8. Benfords lov (loven om det første ciffer)
9. Regression mod middelværdien
10. Multiple comparisons problem (Look-elsewhere effect)

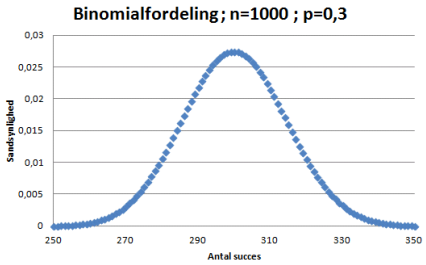


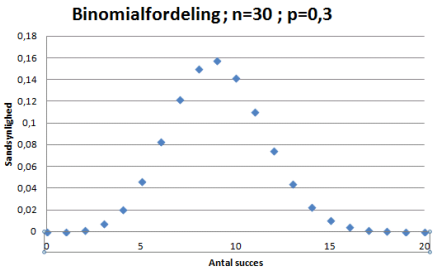
BILAG A: Binomialfordeling

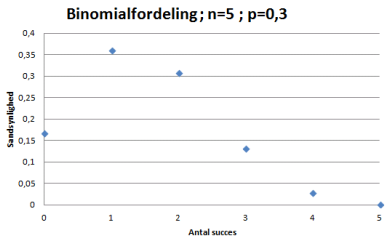


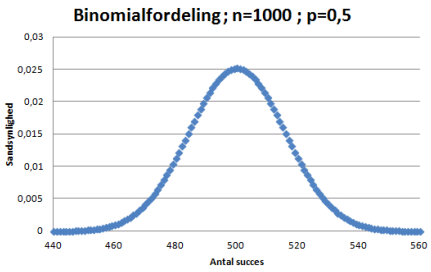


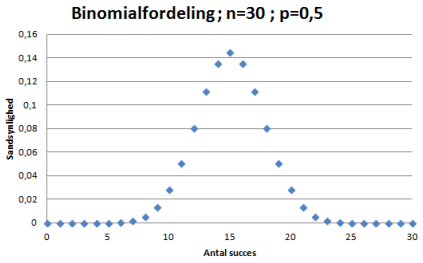


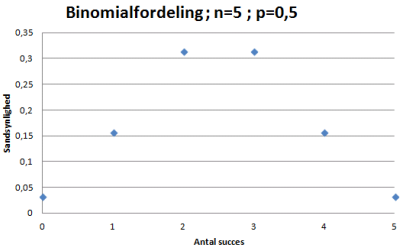












BILAG B: Signifikansniveauer

Degrees of freedom (df)	Q-value										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001