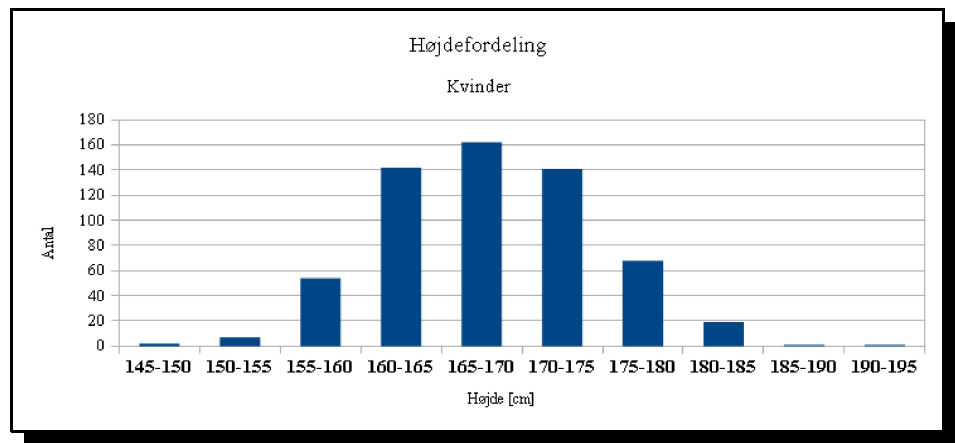


B I O L O G I



OM VURDERING AF TALMATERIALE

Indhold

	Indledning	<i>side 1</i>
1	Resultatbehandling	
	<i>I Skemaopstilling</i>	<i>side 3</i>
	<i>II Variation og spredning</i>	<i>side 4</i>
	<i>III Grafisk vurdering</i>	<i>side 6</i>
2	Statistisk analyse	
	<i>I Test af normalfordeling</i>	<i>side 7</i>
	<i>IIa F-test og t-test</i>	<i>side 7</i>
	<i>IIIb Variansanalyse</i>	<i>side 10</i>
	<i>III Tendenslinier</i>	<i>side 12</i>
	<i>IV χ^2 test</i>	<i>side 15</i>
3	Tillæg	
	<i>A Beregning af gns. og spredning fra grupperede data</i>	<i>side 16</i>
	<i>B Tabel over udvalgte χ^2 værdier</i>	<i>side 16</i>
4	Litteratur	<i>side 17</i>
5	Register	<i>side 18</i>

□ Thorkild Steenberg □

4. udg. 2020

Indledning

Resultater af biologiske forsøg eller undersøgelser vil ofte have form af en række måledata eller beregninger på basis af målinger.

Hvordan overskues og behandles sådanne forsøgsresultater?

Ofte ønsker man tillige svar på spørgsmålet: "Er der forskel på resultaterne af denne og hin måling?"

Hæftet beskriver de mest grundlæggende emner i talbehandling: opstilling af resultater i tabeller, import af data, beregning af gennemsnit og spredning, tegning af fordelingsdiagram, statistisk test af om to gennemsnit er ens, statistisk test af om flere gennemsnit er ens, statistisk test af om en faktisk fordeling afviger fra det forventede og tilpasning af data til en ret linie eller andre kurveformer.

□ □ □

Resultatbehandling

I Oversigtsskema og gennemsnitsberegning

To stikprøver på hver 50 stk grønne bønner og gule bønner vejes for at kunne besvare følgende spørgsmål:

- 1 *Hvad er gennemsnitsvægten af grønne bønner og gule bønner?*
- 2 *Er der forskel på bønnerne med hensyn til vægt?*

Måledata indtastes i et regnearkskema og gennemsnitsværdier beregnes (figur 1).

Er der i stedet tale om data i fil-format - typisk udtræk fra databaser eller lignende - indlæses disse data direkte i regnearket (se tillæg side 16 for beregning ud fra grupperede data).

	Grønne bønner		Gule bønner	
	vægt i mg			
	161	127	230	160
	248	92	180	170
	169	143	240	240
	108	157	200	220
	221	124	200	220
	102	168	230	110
	251	234	190	180
	182	159	230	140
	148	109	220	150
	180	182	220	180
	128	182	310	210
	161	152	190	180
	205	159	270	250
	178	129	210	150
	178	153	230	240
	164	153	250	170
	136	138	310	210
	153	136	150	250
	146	155	150	270
	215	164	190	240
	116	307	190	280
	161	183	150	190
	119	141	190	220
	108	180	200	180
	185	133	180	250
Gns	162		207	

Figur 1

Vægte for grønne og gule bønner; nederst i tabellen er gennemsnitsvægtene beregnet.

Svaret på spørgsmål 1 er gennemsnitsvægtene i figur 1:

162 mg for grønne bønner og 207 mg for gule bønner.

II Resultaternes variation

Spørgsmål 2 kan derimod ikke umiddelbart besvares. Om der er reel forskel på gennemsnitsværdierne afhænger af hvor "godt" de er bestemt, dvs. hvor stor variation der er imellem de enkelte måleresultater.

Det kan give et overblik at tegne søjlediagrammer (histogrammer) over resultaternes fordeling. Vælg et passende antal, lige store intervaller og optæl hvor mange værdier, der falder i hvert interval. Værdier, der er lig med et intervals øvre grænse, medtages i dette interval, medens værdier, der er lig med nedre intervalgrænse, medtages i forrige interval. Figur 2 og 3 viser en fordelingstabel og det tilsvarende søjlediagram.

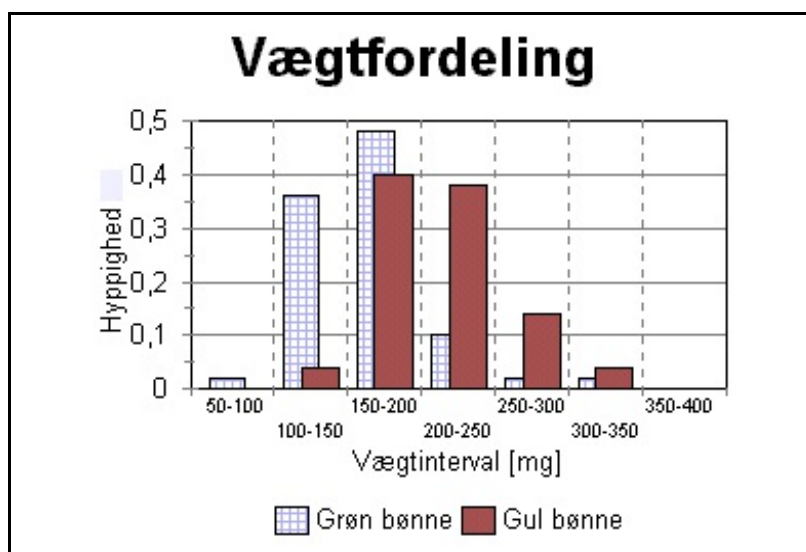
Figur 2.

Eksempel på fordelings-tabel. Intervalhyppighederne er angivet som antal og som decimaltal.

Interval	Grøn bønne		Gul bønne	
	Antal	Hyppighed	Antal	Hyppighed
50-100	1	0,02	0	0
100-150	18	0,36	2	0,04
150-200	24	0,48	20	0,4
200-250	5	0,1	19	0,38
250-300	1	0,02	7	0,14
300-350	1	0,02	2	0,04
350-400	0	0	0	0
Sum	50	1	50	1

Figur 3

Diagram med vægtintervaller afsat på x-aksen og intervalhyppighed afsat på y-aksen.



Resultaterne fordeler sig mere eller mindre jævnt på begge sider af en gennemsnitsværdi. Hvis man forestiller sig et ideelt, meget stort forsøgsmateriale ville resultaterne fordele sig symmetrisk omkring gennemsnitsværdien og udgøre det, man kalder en normalfordeling.

Det faktiske måleresultat kan så betragtes som en stikprøve af dette ideelle, normalfordelte forsøgsmateriale, og det man behøver er et mål for stikprøvens pålidelighed som udtryk for denne fordeling.

Beregner man den gennemsnitlige afvigelse mellem måleværdierne og gennemsnittet fås resultaternes **spredning** (σ):

$$\sigma = \sqrt{\frac{\sum (\bar{x} - x_i)^2}{n-1}} ; \left[\begin{array}{l} \Sigma = \textit{summationstegn} \\ \bar{x} = \textit{gennemsnit} \\ x_i = \textit{måleværdi} \\ n = \textit{antal måleværdier} \end{array} \right]$$

Spredningen udtrykker resultaternes "samling" om gennemsnittet.

Jo mindre σ er, des bedre er gennemsnitsværdien bestemt, og man kan så tillade sig at betragte to gennemsnit som forskellige, hvis deres respektive spredningsintervaller (gns $-\sigma$ til gns $+\sigma$) ikke overlapper eller kun overlapper lidt.

I figur 4 er resultaterne vist med udregnet spredning (se side 16 hvis man arbejder med grupperede data).

	Grønne bønner	Gule bønner
Gns	162	207
Spr	41	43

Figur 4. Beregnet gennemsnit og beregnet spredning

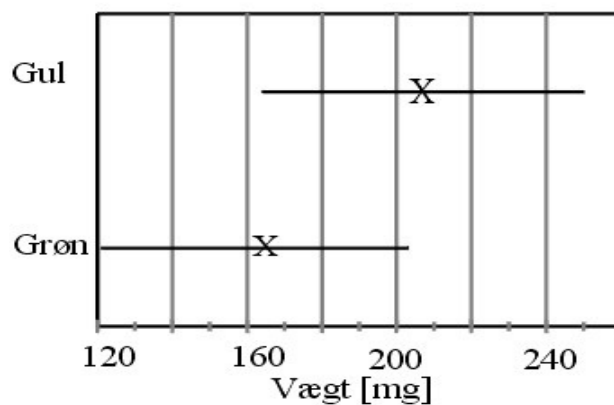
Man har valgt $\bar{x} \pm \sigma$ som en standard resultatangivelse.

Svaret på spørgsmålet bør derfor korrekt skrives:

Grønne bønner:	162 \pm 41 mg	Gule bønner:	207 \pm 43 mg
----------------	-----------------	--------------	-----------------

III Grafisk vurdering af resultater

Svaret på spørgsmål 2 kan man nu få ved at sammenligne spredningsintervallerne (gennemsnit - σ til gennemsnit + σ ; figur 6) for forsøgsresultaterne.



Figur 5.

Diagram der viser spredningsintervaller for måleresultater. (X markerer gennemsnit)

Spredningsintervallerne overlapper noget; men ikke mere end at det må kunne påståes, at gennemsnitsvægtene er forskellige.

□ □ □

Statistisk analyse

Længere end ovenstående kan man ikke komme uden en egentlig statistisk analyse.

I Test af normalfordeling

Den grundlæggende forudsætning bør testes: nemlig at der er tale om normalfordelte værdier i en stikprøve.

Ud fra fordelingstabellerne (side 4) beregnes den kumulerede hyppighed ved at fordelingshyppighederne (%) lægges sammen løbende (1. felt alene, 2. felt lægges sammen med 1. felt, 3. felt lægges sammen med 1. og 2. felt, etc). Disse kumulerede hyppigheder afsættes derefter i et digram på statistikpapir (normalfordelingspapir). Ligger de afsatte punkter jævnt om - eller på - en ret linie er der tale om en normalfordeling.

II Afgørelse af om der er forskel på gennemsnitsværdier

Her må man skelne mellem

A) tilfælde hvor kun to gennemsnitsværdier skal sammenlignes og

B) tilfælde hvor flere end to gennemsnitsværdier skal sammenlignes.

II A Sammenligning af to middelværdier (*t*-test)

Den statistiske test der anvendes til sammenligning af to middelværdier (*t*-test) har som forudsætning, at varianserne ($\text{var} = \sigma^2$) for de to stikprøver, der skal sammenlignes, er ens. Derfor må det først afgøres, om de foreliggende varianser er tilstrækkeligt ens.

II A-1 Denne test kaldes en *F*-test

$$F = \frac{\text{største varians}}{\text{mindste varians}} ; \left[\begin{array}{l} m = \text{antal målinger i tælleren} - 1 \\ n = \text{antal målinger i nævneren} - 1 \end{array} \right]$$

Hvis $\sigma_1^2 = \sigma_2^2$ er $F = 1$.

Det vil sige, hvis F værdien er tilstrækkeligt tæt på 1, er der ikke noget, der strider mod at acceptere, at de to stikprøvevarianser er ens.

F værdien kan slås op i en tabel (m og n kaldes frihedsgrader), og sandsynligheden for at varianserne er ens kan aflæses. Alternativt kan man med de indbyggede statistiske funktioner i regnearket beregne F -værdien og sandsynligheden på én gang (eksempel figur 6).

Hvis F er mindre end den kritiske værdi - det samme som at sandsynligheden $P(F \leq f) \geq 0.05$ - accepteres det, at varianserne er ens.

I figur 6 sammenlignes varianserne for de to bønner ved en F -test:

F-test: Varianser for to stikprøver		
	Grøn	Gul
Middelværdi	161,66	207,4
Varians	1705,3718	1848,20408
Observationer	50	50
df	49	49
F	1,0837543	
P(F<=f) enhalet	0,4227166	
F-kritisk enhalet	1,4463807	

Figur 6 Eksempel på F-test med regneark

F er mindre end den kritiske værdi, dvs det accepteres at varianserne er ens.

II A-2 Derefter fortsættes med den egentlige sammenligning af middelværdierne: en t -test

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right) \frac{\sum x_i^2 - n_x \bar{x}^2 + \sum y_j^2 - n_y \bar{y}^2}{n_x + n_y - 2}}}; \quad \left[\begin{array}{l} \bar{x} \text{ og } \bar{y} = \text{gns. 1 og 2} \\ n_x \text{ og } n_y = \text{antal data} \\ x_i \text{ og } y_j = \text{måleværdier} \end{array} \right]$$

Hvis $gns_1 = gns_2$ er $t = 0$.

Det vil sige, hvis $|t|$ -værdien er tilstrækkeligt tæt på 0, er der ikke noget, der strider mod at acceptere, at de to gennemsnit er ens.

t -værdien kan slås op i en tabel ($n_x + n_y - 2 =$ antal frihedsgrader) og sandsynligheden for at gennemsnittene er ens kan aflæses; men her er det absolut en fordel at bruge regnearkets statistiske funktioner.

t-Test: to stikprøver med ens varianser		
	Grøn	Gul
Middelværdi	161,66	207,4
Varians	1705,3718	1848,20408
Observationer	50	50
Samlet varians	1739,205	
Hypotetisk middelforskel	0	
df	48	
t	-4,025227	
P(T<=t) enhalet	0,0001006	
t-kritisk enhalet	1,6772242	
P(T<=t) tohalet	0,0002013	
t-kritisk tohalet	2,0106348	

Figur 7 *Eksempel på t-test med regneark*

Hvis $|t|$ er mindre end den kritiske værdi - det vil sige det samme som at sandsynligheden $P(T \leq t) \geq 0,05$ - accepteres, at gennemsnittene er ens.

I figur 7 sammenlignes middelværdierne for grønne bønner og gule bønner ved en t -test:

$|t|$ er større end den kritiske værdi, dvs de to middelværdier er ikke ens.

□ □ □

II B Sammenligning af flere middelværdier (variansanalyse)

Man kan bruge en *variansanalyse* til at teste hypotesen at et antal stikprøver har det samme gennemsnit. Variansen mellem grupper sammenlignes med variansen inden for grupperne; der anvendes også her en *F*-test:

$$F = \frac{\sum_1^G (\bar{x}_g - \bar{x})^2}{(G - 1)} \div \frac{\sum_1^G \sum_1^{n_g} (x - \bar{x}_g)^2}{(N - G)} ; \left[\begin{array}{l} x = \text{enkelte data} \\ \bar{x}_g = \text{gruppegns.} \\ G = \text{antal grupper} \\ N = \text{total antal} \\ \bar{x} = \text{total gns} \\ n_g = \text{antal data i gruppen} \end{array} \right]$$

Tælleren udtrykker gruppegennemsnitenes afvigelse fra totalgennemsnit og nævneren udtrykker de enkelte datas afvigelse fra deres gruppegennemsnit. Hvis de to variansudtryk er ens er *F* = 1.

Det vil sige, at hvis *F*-værdien er tilstrækkeligt tæt på 1, er der ikke noget der strider mod at acceptere, at stikprøvegennemsnittene er ens.

Hvis *F* er mindre end den kritiske værdi - det samme som at sandsynligheden $P(F \leq f) \geq 0,05$ - accepteres det, at gennemsnittene er ens.

Afstand mellem træer i seks prøveflader: 1-4: 25 kvm; 4-6 100 kvm						
	1	2	3	4	5	6
	2,300	0,450	2,630	1,200	2,300	1,750
	2,800	2,450	2,630	1,700	2,300	1,000
		1,000	1,220	0,710	1,900	1,000
		0,450	1,220	3,740	2,270	2,300
		0,450		0,870	2,000	2,500
		0,550		2,500	1,900	1,800
				1,900	3,470	1,800
				1,400	1,550	2,280
					0,800	2,870
					0,800	1,200
						0,830
						0,830
gns	2,550	0,892	1,925	1,753	1,929	1,680
spr	0,354	0,793	0,814	0,988	0,779	0,705

Figur 8. Afstandsmålinger fra seks prøveflade skal sammenlignes

I figur 8 er vist målinger af afstanden mellem træer i seks prøveflader. Er de seks prøveflader sammenlignelige, således at de kan betragtes som en stor prøveflade? Her skal bruges en variansanalyse. Resultatet af analysen ses i figur 9.

Variansanalyse - enkelt faktor						
RESUME						
<i>Grupper</i>	<i>Antal</i>	<i>Sum</i>	<i>Gennemsnit</i>	<i>Varians</i>		
Kolonne 1	2	5,1	2,55	0,125		
Kolonne 2	6	5,35	0,8916667	0,628417		
Kolonne 3	4	7,7	1,925	0,6627		
Kolonne 4	8	14,02	1,7525	0,975507		
Kolonne 5	10	19,29	1,929	0,607321		
Kolonne 6	12	20,16	1,68	0,496982		
Beregning						
<i>Variationskilde</i>	<i>SK</i>	<i>fg</i>	<i>MK</i>	<i>F</i>	<i>P-værdi</i>	<i>F krit</i>
Mellem grupper	6,118024	5	1,2236049	1,913841	0,116175	2,477165
Inden for grupper	23,01642	36	0,6393451			
I alt	29,13445	41				

Figur 9. Variansanalyse udført på resultaterne i figur 8

F er mindre end den kritiske værdi; det vil sige at gennemsnitsværdierne i eksemplet er sammenlignelige.

□ □ □

III Tendenslinier (regressionsanalyse).

Ofte har man i biologiske forsøg samtidige målinger af to eller flere variabler, som man vil undersøge den indbyrdes sammenhæng imellem.

Målingerne afsættes i et x-y diagram, som punkter uden forbindelseslinier.

Regnearket indeholder standard tendenslinier - fx lineær, logaritmisk eller potens funktion, som kan anvendes til at analysere den eventuelle sammenhæng.

Regnearket beregner konstanter i den valgte funktion og kan vise funktionen i diagrammet.

Desuden får man en beregning af korrelationskoefficienten r , som angiver graden af indbyrdes afhængighed mellem de to variable:

$r = 0$ ingen afhængighed til $r = 1$ total afhængighed.

Det er vigtigt at understrege, at der ikke nødvendigvis er tale om en årsagssammenhæng mellem de to variable selv om korrelationskoefficienten er høj.

Sandsynligheden for at en r -værdi er forskellig fra 0 kan findes med en t -test:

$$t = \sqrt{v r^2 / (1 - r^2)} ; v = n - 2$$

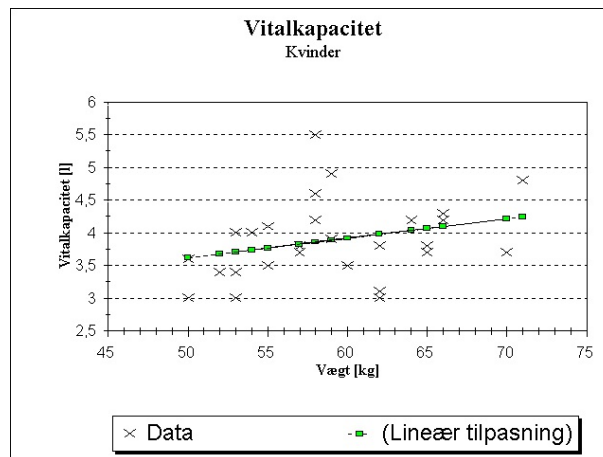
Tabel over udvalgte værdier af t_k og n	
t_k	n
1,960	∞
1,980	120
2,048	30
2,101	20
2,306	10

Hvis $t > t_k$ (alternativt $0,025 \leq P(T \leq t) \leq 0,0975$) kan det ikke accepteres at r er lig med 0 (dvs. r signifikant forskellig fra 0).

Hosstående tabel viser udvalgte værdier af t_k og n .

Figur 10 viser som eksempel målinger af vægt og vitalkapacitet (dvs forskel mellem maksimal indånding og maksimal udånding) hos kvinder

Målingerne er afsat i et x-y diagram. Den mest simple sammenhæng vil være en lineær funktion, og opgaven består således i at bestemme et udtryk for den rette linie, der bedst muligt tilpasses punktmængden.



Figur 10 Vitalkapacitet som funktion af vægt hos kvinder

Man vælger linien således, at summen af kvadratet på afstanden mellem de beregnede y-koordinater og de faktiske y-koordinater bliver mindst mulig:

$$\sum (y_i - y_{bi})^2 ; \left[\begin{array}{l} y_i = \text{målt } y\text{-koordinat} \\ y_{bi} = \text{beregnet } y\text{-koordinat} \end{array} \right]$$

Metoden kaldes *lineær tilpasning* efter mindste kvadraters metode.

Ligningen for linien er $y = \alpha x + b$ og hældningskoefficienten α og skæringspunktet med y-aksen b kan beregnes efter følgende udtryk:

$$\alpha = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} , \quad b = \bar{y} - \alpha \bar{x} ; \quad \left[\begin{array}{l} \bar{x} = \text{gns. af } x\text{-data} \\ \bar{y} = \text{gns. af } y\text{-data} \\ x_i, y_i = \text{datasæt} \end{array} \right]$$

Regnearket kan beregne hældningskoefficient og skæringspunkt med y-aksen (se fig 11).

I figur 11 er også vist korrelationskoefficienten r (ofte vist som r^2)

Ligningen for eksemplet i figur 10 bliver ved indsætning af værdierne i figur 11:

$$\text{vitalkap}_{\text{kvinder}} = 0,0298 * \text{vægt}_{\text{kvinder}} + 2,1299; \quad r^2 = 0,085179;$$

Udføres en *t*-test på korrelationskoefficienten som beskrevet ovenfor fås at r ikke adskiller sig fra 0; det vil sige at datamaterialet ikke viser sammenhæng eller at datamaterialet ikke er tilstrækkeligt til at vise en sammenhæng mellem vitalkapacitet og vægt.

Regressionsdata for eksempel i figur 10			
Konstant			2,129942
Standardafv. for Y-estimat			0,577577
Kvadratet på R			0,085179
Antal observationer			28
Frihedsgrader			26
<hr/>			
X-koefficient(er)		0,029788	
Standardafv. for koefficient		0,019145	

Figur 11 Eksempel på lineær tilpasning vha regneark

□ □ □

IV Vurdering af om observerede værdier er tilstrækkeligt tæt på forventede værdier (hypotese): χ^2 test

Arbejde med fænotypiske fordelinger rejser ofte spørgsmålet: “Er den observerede fordeling i oversensstemmelse med den forventede, teoretiske fordeling?”

Svaret på sådanne spørgsmål kan findes med en χ^2 test:

$$\chi^2 = \sum \frac{(\text{obs.} - \text{forv.})^2}{\text{forv.}}; \text{obs.} = \text{observerede og forv.} = \text{forventede værdier}$$

Testen beregner et udtryk for summen af kvadrerede afvigelser mellem forventede og observerede værdier.

I nedenstående tabel er vist fordelingen af blodtyper i fire studenterklasser 2014-2017.

	A	O	B	AB	Total
Observeret	25	22	6	5	58
Forventet	25,25	24,36	5,8	2,32	58
Afvigelse	-0,52	-2,36	0,2	2,68	

Den forventede fordeling er beregnet ud fra fordelingen i Danmark:

A	O	B	AB
44 %	42 %	10 %	4 %

Udregnes udtrykket ovenfor med afvigelserne fra tabellen indsat fås: $\chi^2 = 3,34$
 χ^2 værdien kan slåes op i en tabel (*frihedsgrader fås ved at trække 1 fra antal kolonner (klasser)*) og **sandsynligheden for at en χ^2 værdi mindst så stor som den fundne kunne ventes at fremkomme ved tilfældighed alene** kan aflæses. Alternativt kan sandsynligheden beregnes i regnearkets statistiske funktioner.

I tillægget er vist en tabel over udvalgte χ^2 værdier.

Aflæs den fundne χ^2 værdi i denne tabel fås sandsynligheden ca 20 %; dvs. der er ikke noget, der strider mod at acceptere at blodtypefordelingen følger fordelingen i den danske befolkning.

□ □ □

Tillæg

A Formler til beregning af gennemsnit og spredning af grupperede data

Beregning af gennemsnit ud fra fordelingstabeller (grupperede data):

$$\bar{x} \approx \frac{\sum (n_x \cdot x)}{\sum n_x} ; \left[\begin{array}{l} \bar{x} = \text{gennemsnit} \\ \sum = \text{summationstegn} \\ n_x = \text{antal i klasse} \\ x = \text{klassemidte} \end{array} \right]$$

Lav et antal kolonner i regnearket til udregning af formelen.

Beregning af spredning ud fra fordelingstabeller (grupperede data):

$$\sigma \approx \sqrt{\frac{\sum (n_x \cdot x^2) - \frac{(\sum (n_x \cdot x))^2}{\sum n_x}}{(\sum n_x) - 1}} ; \left[\begin{array}{l} \sigma = \text{spredning} \\ \sum = \text{summationstegn} \\ n_x = \text{antal i klasse} \\ x = \text{klassemidte} \end{array} \right]$$

Lav et antal kolonner i regnearket til udregning af formelen

B Tabel over udvalgte χ^2 værdier

Tabellen nedenfor angiver sandsynligheden for at en χ^2 værdi mindst så stor som den fundne kunne ventes at fremkomme ved tilfældighed alene; frihedsgrader findes ved at trække 1 fra antal af klasser (kolonner) i tabelopstillingen.

frihedsgrader	χ^2 værdier							
1	0,0002	0,004	0,455	1,074	1,642	2,706	3,841	6,635
2	0,020	0,103	1,386	2,408	3,219	4,605	5,991	9,210
3	0,115	0,352	2,366	3,665	4,642	6,251	7,817	11,35
$p(\chi^2 > \text{fundet})$	99	95	50	30	30	10	5	1

Litteratur

- 1 Bernhard Andersen & Erling B. Andersen:
Grundlæggende statistik
Gyldendal. 1978.
- 2 Bernhard Andersen & Erling B. Andersen:
Matematisk statistik i grundtræk
Gyldendal. 1973.
- 3 Arne Nielsen, Jørgen Hilden & Kirsten Fenger:
Statistik og sandsynlighed anvendt i medicin
FADL 2.udg. 1976.
- 4 Anon.
Applied Statistics with TI Programmable 58/59
Texas Instruments Inc. 1977.
- 5 Bent Christensen, Hans Greve & Svend Sporning (red.)
Biologisk Forskning gml. udgave 6. oplag
Gyldendal 1976

□ □ □

Register

σ , spredning		Normalfordeling	4
beregning	5	test af	7
definition	5	Resultatangivelse	5
resultatangivelse	5	Spredning, σ	
F-test		definition	5
definition	7	Spredningsintervaller	5
Fordelingstabel	4	grafisk sammenligning af	6
beregning af gennemsnit	16	Statistisk analyse	7
beregning af spredning	16	Stikprøve	5
Gennemsnit		Søjlediagram	4
beregning (grupperede data)	16	t-test	
resultatangivelse	5	beregning	7
t-test; sammenligning af to		definition	8
gennemsnit	7	Varians	7
variansanalyse; sammenligning af		Variansanalyse	10
flere gns.	10	definition	10
Grafisk vurdering	6	Variation	4
Hyppigheder		x-y diagrammer	
kumulerede	7	lineær tilpasning	12
Intervalhyppighed	4		
Kommafiler			
data i fil-format	3		
Lineær tilpasning			
definition	12, 13		
Måledata	1, 12		

