Stx

Chi-i-anden test i Biologi A

Jette H. Vestergaard

Dronninglund Gymnasium

Januar 2024

Indholdsfortegnelse

Forside illustration:
https://pixabay.com/da/photos/sommerfugl-insekt-vinger-blomster-1218884/

Forord

Dette hæfte er skrevet til Biologi A på stx. Hæftet er skrevet efter ønske fra Fagkonsulenten i biologi. Hæftet er et supplement til hæftet *Matematik i Biologi A*, der udkom i januar 2022.
I hæftet *Matematik i Biologi A* står der, at eleverne i biologi A skal kunne anvende binomialtest i opgaver om statistiske tests. Vejledningen til biologi A blev imidlertid ændret i juni 2023.

> ” Følgende ændringer er foretaget i vejledningen i juni 2023:
> Binomialfordeling- og test erstattes af χ^2-testen, goodness-of-fit, på A niveau. ”

Kilde:
https://www.uvm.dk/-/media/filer/uvm/udd/gym/pdf23/jun/230608-vejledning-til-biologi-a--b-og-c--stx.pdf

χ^2-testet er ikke kernestof på matematik B. Eleverne skal derfor lære at anvende χ^2-testet i biologiundervisningen. Dette hæfte er skrevet til dette formål.
I hæftet gennemgås teori og eksempler på χ^2-testet Goodness-Of-Fit. Sidst i hæftet findes en række opgaver. Til eksamen i biologi A vil testet indgå i en større opgave om en case. Opgaverne bag i dette hæfte indeholder kun det delspørgsmål, som handler om χ^2-testet.
I opgaver om χ^2-test skal man bruge antallet af frihedsgrader til at vurdere teststørrelsen. Dette antal vil altid være angivet i eksamensopgaver.

Der findes endnu et χ^2-test, nemlig χ^2-testet for uafhængighed. Dette test indgår ikke på Biologi A og vil derfor ikke bliver behandlet i dette hæfte.

Kapitel 6 og 7 er for den videbegærlige læser, der ønsker at vide mere om χ^2-fordelingen og mere om fejltyper. Det er ikke nødvendigt at læse disse afsnit for at forstå selve testet.

1. Hvornår anvendes Goodness-Of-Fit testet?

χ^2-testet Goodness-Of-Fit, også kaldet GOF-testet, kan bruges til at undersøge om en fordeling er som forventet. Udgangspunktet er hver gang, at der udtages en repræsentativ stikprøve af en større population. Man vil derefter gerne kunne konkludere noget om fordelingen i hele populationen ved at betragte stikprøven.
Stikprøven må ikke være for stor i forhold til populationen, og den må heller ikke være for lille.
Der er ingen faste regler for, hvor stor en stikprøve må være. Som tommelfingerregel kan vi anvende et statistisk test, når stikprøven maksimalt udgør 10 % af populationen.[1]
Mht. en nedre grænse for stikprøvens størrelse er det lidt mere kompliceret, så den vender vi tilbage til i afsnit 2.6.

1.1 Eksempler

1) Majskolber
For en bestemt type majs ønsker man at undersøge hypotesen om 2-gen udspaltning, hvor to uafhængige gener giver 4 fænotyper, der vil fordele sig i forholdet 9:3:3:1.
Der udtages en stikprøve af denne type majsplante. Resultatet fremgår af tabellen.

Udseende	Violet og glat	Violet og rynket	Gul og glat	Gul og rynket
Antal kerner	648	237	195	63

Kilde: MATBSTX, 28. maj 2015, opgave 13:
https://www.xn--prvebanken-1cb.dk/proevematerialer/GYMUDD/MAT/MATBSTX/materialesamling/98ea81f1-748a-44ef-9116-a5b1f25d7112

a) Undersøg ved et χ^2-test om denne population af majsplanter følger den biologiske model.

2) Orangutanger på Borneo
Proteinet α-globin indgår i hæmoglobin. Hos orangutanger findes to alleler af α-globin-genet: W_1 og W_2.
En gruppe forskere har udtaget en stikprøve blandt orangutanger på Borneo. Resultatet fremgår af tabellen.

Genotype	*W1W1*	*W1W2*	*W2W2*
Antal orangutanger	31	13	10

Kilde: BIOASTX, 2. juni 2015, opgave 2:
https://www.xn--prvebanken-1cb.dk/proevematerialer/GYMUDD/BIO/BIOASTX/materialesamling/c3e85a45-1773-48a6-a124-420b5dfd23e4

a) Beregn allelfrekvenserne p(*W1*) og q(*W2*).
b) Undersøg ved et χ^2-test om denne population af orangutanger er i Hardy-Weinberg ligevægt.

[1] Kilde: Jan Sørensen, Aalborg City Gymnasium

3) Lactoseintolerans

I mælk findes lactose. Evnen til at kunne nedbryde lactose som voksen kaldes lactosetolerans.
Tidligere undersøgelser har vist, at 5 % af befolkningen i Danmark er lactose**in**tolerant.
En gruppe forskere ønsker at undersøge, om forekomsten af lactose**in**tolerans i Grønland er den samme som i Danmark. De udtager derfor en stikprøve blandt befolkningen i Grønland. Resultatet fremgår af tabellen.

Lactose-status	tolerant	**in**tolerant
Antal grønlændere	31	119

Kilde:
https://www.apoteket.dk/sygdom/allergi-og-hoefeber/laktoseintolerans

a) Undersøg ved et χ^2-test om forekomsten af lactose**in**tolerans er den samme i Grønland som i Danmark.

Man kan altså anvende GOF-testet til at undersøge

- om en population følger en udspaltningsmodel
- om en population er i Hardy-Weinberg ligevægt
- om fordelingen i én population er den samme som fordelingen i en anden population, hvis man kender fordelingen i den anden population
- om fordelingen i en population har ændret sig i forhold til tidligere, hvis man kender fordelingen fra tidligere

Der findes også andre situationer, hvor man kan anvende GOF-testet, men i biologi vil det ofte være et af disse fire tilfælde, der er tale om.

2. Teori og beregninger – Eksempel 1

2.1 Observerede værdier

Vi betragter eksemplet med majskolberne. Fordelingen i stikprøven var følgende:

Udseende	Violet og glat	Violet og rynket	Gul og glat	Gul og rynket
Antal kerner	648	237	195	63

Disse værdier kaldes de observerede værdier.

2.2 Nulhypotese

Når man vil teste, om fordelingen i en population passer med en forventet fordeling, skal man starte med at opstille en nulhypotese. Nulhypotesen skal altid være, at fordelingen i den population, vi har taget stikprøven fra, er den samme som den forventede fordeling. I dette tilfælde bliver nulhypotesen:

H_0: Populationen af majsplanter følger udspaltningsmodellen 9:3:3:1.

Man skal også have opstillet en alternativ hypotese. Den alternative hypotese er altid "det modsatte" af nulhypotesen. I dette tilfælde bliver den:

H_A: Populationen af majsplanter følger **ikke** udspaltningsmodellen 9:3:3:1.

Hvis nulhypotesen forkastes, er det den alternative hypotese, der gælder. Man opskriver imidlertid ikke altid den alternative hypotese, da den ofte antages at være underforstået.

2.3 Forventede værdier

De forventede værdier skal beregnes ud fra forventningen i nulhypotesen. Hvis populationen af majsplanter følger den biologiske model, vil fordelingen følge udspaltningsforholdet 9:3:3:1. Vi kan derfor beregne de forventede værdier, ved at gange det samlede antal kerner med brøkdelene: $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$ og $\frac{1}{16}$.
Antal kerner i alt: $648 + 237 + 195 + 63 = 1143$

Udseende	Violet og glat	Violet og rynket	Gul og glat	Gul og rynket
Antal kerner	$1143 \cdot \frac{9}{16} = 642{,}9$	$1143 \cdot \frac{3}{16} = 214{,}3$	$1143 \cdot \frac{3}{16} = 214{,}3$	$1143 \cdot \frac{1}{16} = 71{,}4$

Det ses, at de forventede værdier ikke helt er de samme som de observerede værdier.

Spørgsmålet er nu:
Afviger de forventede værdier så meget fra de observerede værdier i stikprøven, at de faktisk tyder på en anden fordeling i hele populationen af majsplanter (en signifikant forskel), eller er der blot tale om tilfældige udsving inden for det sandsynlige område?

Dette spørgsmål kan vi svare på ved at udføre GOF-testet.

2.4 χ^2-teststørrelsen

χ^2-teststørrelsen skal bruges til at vurdere, om de observerede værdier stemmer nogenlunde overens med de forventede værdier. En mulig teststørrelse kunne være at beregne forskellene mellem de observerede og de forventede værdier og så lægge alle disse forskelle sammen. Jo større forskelle desto mindre stemmer de observerede værdier overens med de forventede.

Som illustration betragter vi følgende 2 eksempler:

Eksempel 1

Udfald	A	B	C	Sum
Observerede	9	10	11	30
Forventede	10	10	10	30
obs-forv	-1	0	1	0

Eksempel 2

Udfald	A	B	C	Sum
Observerede	5	10	15	30
Forventede	10	10	10	30
obs-forv	-5	0	5	0

I begge eksempler ses det, at summen af afvigelserne er 0, så summen af afvigelserne fortæller desværre ikke noget om, hvor meget de observerede afviger fra de forventede.
Vi prøver derfor at beregne afvigelserne og derefter sætter dem i anden i stedet for.

Eksempel 1

Udfald	A	B	C	Sum
Observerede	9	10	11	30
Forventede	10	10	10	30
obs-forv	-1	0	1	0
$(\text{obs-forv})^2$	1	0	1	2

Eksempel 2

Udfald	A	B	C	Sum
Observerede	5	10	15	30
Forventede	10	10	10	30
obs-forv	-5	0	5	0
$(\text{obs-forv})^2$	25	0	25	50

I eksempel 1 er afvigelserne mindre end i eksempel 2, og summen af $(\text{obs-forv})^2$ er tilsvarende mindre i eksempel 1 end i eksempel 2.

Et fornuftigt mål for, hvor meget de observerede afviger fra de forventede, kunne derfor være at beregne $(\text{obs-forv})^2$ for hver observation, og derefter beregne summen af disse størrelser.
Den endelige teststørrelse beregnes imidlertid (af hensyn til den senere vurdering af teststørrelsen) ved at dividere størrelserne $(\text{obs-forv})^2$ med de forventede og derefter lægge brøkerne sammen, dvs. at formlen for den endelige χ^2-teststørrelse bliver:

$$\chi^2 = \sum \frac{(obs - forv)^2}{forv}$$

En stor teststørrelse vil være mere kritisk overfor nulhypotesen end en lille teststørrelse.

Vi vender nu tilbage til eksemplet med majskolberne.

Udseende	Violet og glat	Violet og rynket	Gul og glat	Gul og rynket
Observerede	648	237	195	63
Forventede	642,94	214,31	214,31	71,44
$\frac{(obs-forv)^2}{forv}$	0,0399	2,4017	1,7403	0,9966

χ^2-teststørrelsen bliver så:

$$\chi^2 = 0{,}0399 + 2{,}4017 + 1{,}7403 + 0{,}9966 = 5{,}1785$$

Dette tal er et mål for afvigelserne. Hvis tallet er stort, ligger de observerede værdier langt fra de forventede værdier, og vi begynder at tvivle på vores nulhypotese.

2.5 Vurdering af χ^2-teststørrelsen

Spørgsmålet er nu:
Hvor stor skal χ^2-teststørrelsen så være, for at vi ikke længere tror på nulhypotesen og forkaster den?

Som hovedregel vælger vi at sige, at hvis sandsynligheden for at få en χ^2-teststørrelse der er lig med eller større end den observerede, altså sandsynligheden for i en ny stikprøve at få data, der passer lige så godt eller dårligere med nulhypotesen, er mindre end 5 %, så forkaster vi nulhypotesen.
Denne grænse (de 5 %) kaldes ***signifikansniveauet***.
Sandsynligheden for at få en χ^2-teststørrelse, der er lig med eller større end den observerede, kaldes ***p-værdien***.
Men hvorfor så lige 5 %, kunne man spørge. Der ligger forskellige statistiske argumenter bag dette valg, men dem vil vi ikke komme ind på her. Man kan godt komme ud for, at der er valgt et andet signifikansniveau ved et test, men som regel er det 5 %.

Spørgsmålet er nu:
Hvordan finder vi ud af, om sandsynligheden for at få en χ^2-teststørrelse, der er lig med eller større end den observerede, er mindre end 5 %?

Hvis man sætter en computer til at simulere forsøget rigtig mange gange, vil man få en række forskellige χ^2-teststørrelser. Det viser sig, at fordelingen af disse χ^2-værdier passer rimeligt godt med en teoretisk fordeling, som kaldes en χ^2-fordeling. Vha denne fordeling kan man finde p-værdien.
Når man anvender denne teoretiske χ^2-fordeling til at finde p-værdien, skal man angive antallet af frihedsgrader.
Antal frihedsgrader angiver hvor mange af de observerede værdier, der kan variere, når vi kender det samlede antal observationer. I eksemplet med majskolberne er der 3 frihedsgrader, da de første 3 værdier i tabellen kan variere, mens den fjerde værdi kan beregnes ud fra de første 3 værdier og det samlede antal.

Hovedregel: Hvis der er n kategorier, er antal frihedsgrader givet ved: $n-1$.

Undtagelse:
Hvis de forventede værdier beregnes ud fra en nulhypotese om Hardy-Weinberg ligevægt, er der kun 1 frihedsgrad, selv om der er tre kategorier i tabellen. Det skyldes, at når vi angiver antal frihedsgrader som antal kategorier (n) minus 1, så dækker det i virkeligheden over, at der er n parametre (sandsynligheder) i hypotesen og dermed $n-1$ frie parametre, da den sidste sandsynlighed kan bestemmes ud fra de $n-1$ første sandsynligheder, idet summen af alle sandsynlighederne er 1.
Men i en hypotese om Hardy-Weinberg ligevægt er der en sammenhæng mellem sandsynlighederne. De vil under HW-ligevægt kunne beregnes på følgende vis

$$p^2 \ : \ 2 \cdot p \cdot (1-p) \ : \ (1-p)^2$$

Dvs at der kun er 1 fri parameter, nemlig p. Og dermed kun 1 frihedsgrad.
Men i alle andre situationer, som eleverne vil komme ud for, vil der være $n-1$ frihedsgrader.
Antallet af frihedsgrader vil altid være angivet i eksamensopgaver.

I eksemplet med majskolberne bliver p-værdien 0,159. Sandsynligheden for at få en χ^2-teststørrelse, der er lig med eller større end den observerede, er derfor 15,9 %. Der er altså 15,9 % chance for i en ny stikprøve at få data, der passer lige så godt eller dårligere med nulhypotesen, hvis nulhypotesen gælder.

Konklusion:
Da p-værdien er større end 5 %, er denne stikprøve ikke usædvanlig under denne nulhypotese. Ved et signifikansniveau på 5 % giver denne stikprøve derfor *ikke* anledning til at forkaste nulhypotesen, dvs. at populationen af majsplanter følger udspaltningsmodellen 9:3:3:1.

2.6 Kriterier for anvendelse af χ^2-testet

Den generelle regel er, at hvis alle de forventede værdier er større end 5, så er χ^2-fordelingen en god tilnærmelse til fordelingen af χ^2-teststørrelsen.
Man kan derfor kun anvende χ^2-testet, hvis alle de forventede værdier er større end 5.
Ellers vil beregningen af p-værdien, som jo afhænger af tilnærmelsen til χ^2-fordelingen, være behæftet med for stor usikkerhed.
Morris H DeGroot skriver dog i *Probability and Statistics*, at tilnærmelsen er *meget god*, når de forventede er større end 5, og *tilfredsstillende*, hvis de forventede er større end 1,5.[2] Man må derfor antage, at man også med en vis rimelighed kan anvende χ^2-testet hvis blot alle de forventede er større end 1,5.

[2] DeGroot, Morris H, *Probability and Statistics*, 4. udgave, 2012.

3. Tastevejledning til Excel

Indtast følgende i et regneark i Excel:

	A	B	C	D	E	F
1	**Majskolber**					
2						
3	**Type**	violet og glat	violet og rynket	gul og glat	gul og rynket	**sum**
4	**Observerede**	648	237	195	63	
5	**procenter under H0**					
6	**Forventede**					
7	**(obs-forv)^2/forv**					

I celle F4 skal vi have beregnet summen af observationerne.
Skriv derfor i celle F4: **=sum(B4:E4)** og tast enter. Så bliver summen 1143.
Placér nu curseren i nederste højre hjørne i celle F4, så der kommer et sort kryds. Peg på krydset og hold venstre musetast nede med der trækkes ned til celle F7. Nu kopieres formlen ned til celle F7.

I række 5 indtastes nu procenterne under nulhypotesen. I dette eksempel er procenterne givet ved brøker. Excel omdanner brøker til datoer med mindre, man ændrer lidt på indstillingerne. Markér derfor række 5, højreklik og vælg **Formatér celler**. Under fanen **Tal** vælges **Tal**.
Nu kan brøkerne indtastes for at beregne procenterne under nulhypotesen.
Skriv i celle B5: **=9/16** og tast enter.
Fortsæt i cellerne B6 til B8 med brøkerne 3/16, 3/16 og 1/16.
Læg mærke til at summen bliver 1 i celle F5. Hvis den ikke bliver 1 er der en fejl i procenterne.

I række 6 skal de forventede værdier beregnes.
Skriv i celle B6: **=B5*F4** og tast enter.
Så bliver procenttallet fra B5 ganget med summen i F4. Når man sætter et $-tegn før en række- eller kolonne-henvisning, så låses henvisningen, dvs rækken og/eller kolonnen fastlåses, når formlen kopieres.
Kopiér nu denne formel over til celle E6.
Læg mærke til at summen i celle F6 bliver det samme som i celle F4. Hvis den ikke bliver det, er der en fejl i beregningerne af de forventede.
Læg desuden mærke til, at alle de forventede værdier er større end 5.

I række 7 skal brøkerne $\frac{(obs-forv)^2}{forv}$ beregnes.
Skriv i celle B7: **=(B4-B6)^2/B6** og tast enter. Kopiér nu denne formel over til celle E7.

Resultatet er nu:

3	**Type**	violet og glat	violet og rynket	gul og glat	gul og rynket	**sum**
4	**Observerede**	648	237	195	63	**1143**
5	**procenter under H0**	0,56	0,19	0,19	0,06	**1**
6	**Forventede**	642,94	214,31	214,31	71,44	**1143**
7	**(obs-forv)^2/forv**	0,0399	2,4017	1,7403	0,9966	**5,17847769**

Nu kan χ^2-teststørrelsen aflæses i celle F7. Den bliver χ^2=5,18.
Hvis man vil ændre på antallet af decimaler, klikker man på knapperne 
Den venstre giver flere decimaler. Den højre giver færre decimaler.

Skriv nu følgende i række 9-12:

9	**Chi-i-anden teststørrelsen:.**				
10	**Antal frihedsgrader:**				
11	**p-værdi:**				
12	**p-værdi i procent:**				

Skriv i celle C9: **=F7** og tast enter. Så bliver χ^2-teststørrelsen kopieret til celle C9.

Skriv i celle C10 antal frihedsgrader. Her: **3** da der er 4 observationer i tabellen.

Skriv i celle C11: **=CHIFORDELING(C9;C10)** og tast enter. Så beregnes p-værdien. Her 0,159.

Skriv i celle C12: **=C11*100** og tast enter. Så beregnes p-værdien i procent. Her 15,9 %.

Skriv til sidst en fornuftig konklusion, f.eks.:

	A	B	C	D	E	F
1	**Majskolber**					
2						
3	**Type**	violet og glat	violet og rynket	gul og glat	gul og rynket	**sum**
4	**Observerede**	648	237	195	63	**1143**
5	**procenter under H0**	0,56	0,19	0,19	0,06	**1**
6	**Forventede**	642,94	214,31	214,31	71,44	**1143**
7	**(obs-forv)^2/forv**	0,0399	2,4017	1,7403	0,9966	**5,1785**
8						
9	**Chi-i-anden teststørrelsen:**		5,1785			
10	**Antal frihedsgrader:**		3			
11	**p-værdi:**		0,1592			
12	**p-værdi i procent:**		15,92			
13						
14	**Konklusion:**					
15	Da p-værdien er større end 5 %, er denne stikprøve **ikke** usædvanlig under denne nulhypotese.					
16	Ved et signifikansniveau på 5 % giver denne stikprøve derfor **ikke** anledning til at forkaste nulhypotesen,					
17	dvs. at populationen af majsplanter følger udspaltningsmodellen 9:3:3:1.					

4. Eksempel 2

I eksempel 2 betragtes en stikprøve blandt orangutanger på Borneo.

Genotype	$W_1\,W_1$	$W_1\,W_2$	$W_2\,W_2$
Antal orangutanger	31	13	10

Nulhypotesen er her:

H0: Population af orangutanger er i Hardy-Weinberg ligevægt.

Først beregnes allelfrekvenserne.

$$p = \frac{31}{54} + 0{,}5 \cdot \frac{13}{54} = 0{,}69$$
$$q = \frac{10}{54} + 0{,}5 \cdot \frac{13}{54} = 0{,}31$$

Da H0 siger, at der er Hardy-Weinberg ligevægt, kan vi finde procenterne under H0 vha. formlerne

W1W1: $p^2 = 0{,}69^2$
W1W2: $2pq = 2 \cdot 0{,}69 \cdot 0{,}31$
W2W2: $q^2 = 0{,}31^2$

Disse formler indtastes i række 5, dvs.
Skriv i celle B5: **=0,69^2** og tast enter.
Skriv i celle C5: **=2*0,69*0,31** og tast enter.
Skriv i celle D5: **=0,31^2** og tast enter.
Resten af eksemplet indtastes på samme måde som i eksempel 1. Man kan med fordel kopiere eksempel 1 og så blot slette kolonne E, da der kun er 3 observationer i dette tilfælde. Så er formlerne allerede indtastet.

	A	B	C	D	E
1	**Orangutanger**				
2					
3	**Type**	W1W1	W1W2	W2W2	**sum**
4	**Observerede**	31	13	10	**54**
5	**procenter under H0**	0,48	0,43	0,10	**1**
6	**Forventede**	25,71	23,10	5,19	**54**
7	**(obs-forv)^2/forv**	1,0887	4,4168	4,4595	**9,9650**
8					
9	**Chi-i-anden teststørrelsen:.**		9,97		
10	**Antal frihedsgrader:**		1		
11	**p-værdi:**		0,0016		
12	**p-værdi i procent:**		0,16		
13					
14	**Konklusion:**				
15	Da p-værdien er mindre end 5 %, **er** denne stikprøve usædvanlig under denne nulhypotese.				
16	Ved et signifikansniveau på 5 % giver denne stikprøve derfor anledning til at **forkaste** nulhypotesen,				
17	dvs. at populationen af orangutanger på Borneo **ikke** er i Hardy-Weinberg ligevægt.				

5. Eksempel 3

I eksempel 3 betragtes en stikprøve blandt befolkningen i Grønland.

Lactose-status	tolerant	**in**tolerant
Antal grønlændere	31	119

Samtidig oplyses det, at 5 % af befolkningen i Danmark er lactose**in**tolerant.
Nulhypotesen er her:

H0: Forekomsten af lactose**in**tolerans er den samme i Grønland som i Danmark.

Da H0 siger, at forekomsten er den samme i Grønland som i Danmark, bliver procenterne 0,95 og 0,05.
Disse indtastes i række 5, dvs.
Skriv i celle B5: **0,95**
Skriv i celle C5: **0,05**
Resten af eksemplet indtastes på samme måde som i eksempel 1. Man kan med fordel kopiere eksempel 1 og så blot slette kolonne D og E, da der kun er 2 observationer i dette tilfælde.

	A	B	C	D	E	F
1	**Lactoseintolerans**					
2						
3	**Type**	tolerant	intolerant	**sum**		
4	**Observerede**	31	119	**150**		
5	**procenter under H0**	0,95	0,05	**1**		
6	**Forventede**	142,50	7,50	**150**		
7	**(obs-forv)^2/forv**	87,2439	1657,6333	**1744,8772**		
8						
9	**Chi-i-anden teststørrelsen:.**		1745			
10	**Antal frihedsgrader:**		1			
11	**p-værdi:**		0,0000			
12	**p-værdi i procent:**		0,00			
13						
14	**Konklusion:**					
15	Da p-værdien er mindre end 5 %, **er** denne stikprøve usædvanlig under denne nulhypotese.					
16	Ved et signifikansniveau på 5 % giver denne stikprøve derfor anledning til at **forkaste** nulhypotesen,					
17	dvs. at forekomsten af lactoseintolerans **ikke** er den samme i Grønland som i Danmark.					

6. Appendiks om χ^2-fordelingen

Definition

En χ^2-fordelt stokastisk variabel med k frihedsgrader er en sum af kvadraterne på k standardnormalfordelte stokastiske variabler.

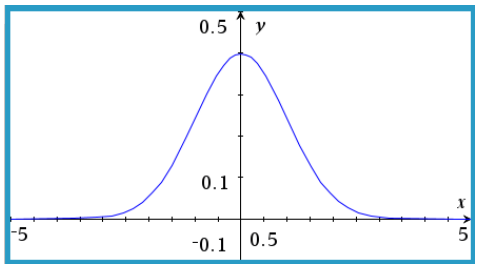
Dvs at hvis $Z_1, Z_2, Z_3, \ldots, Z_k$ er k uafhængige standardnormalfordelte stokastiske variabler, så vil summen af deres kvadrater, altså

$$Q = {Z_1}^2 + {Z_2}^2 + {Z_3}^2 + \ldots + {Z_k}^2$$

være χ^2-fordelt med k frihedsgrader.

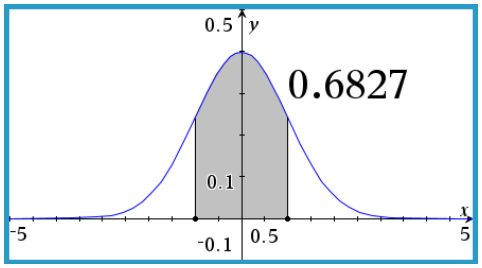
Kilde: https://da.wikipedia.org/wiki/Chi_i_anden-fordelingen

En standardnormalfordelt stokastiske variabel har middelværdi 0 og spredning 1.
Tæthedsfunktionen ses på Figur 1. Tæthedsfunktionen kan bruges til at finde ud af, hvad sandsynligheden er, for at den stokastiske variabel ligger i et bestemt interval.



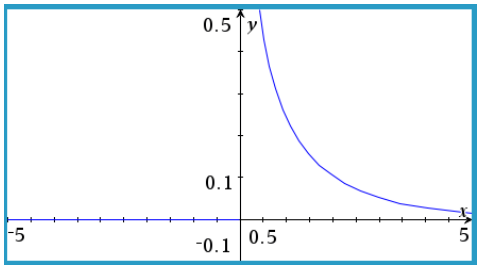
Figur 1

Hvis man vil finde sandsynligheden for at den stokastiske variabel f.eks. ligger mellem -1 og 1, bestemmes arealet under grafen i intervallet [-1;1].
Arealet ses på Figur 2. Det bliver 0.6827, dvs at sandsynligheden for at den stokastiske variabel ligger mellem -1 og 1 er 68,27 %.



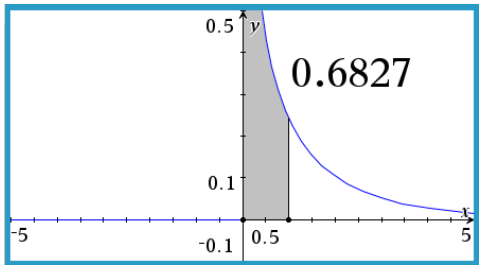
Figur 2

Hvis man kvadrerer en standardnormalfordelt stokastisk variabel, vil man få en stokastisk variabel, der kun antager positive værdier. Tæthedsfunktionen vil derfor kun ligge til højre for y-aksen.
Den nye stokastiske variabel er χ^2-fordelt med 1 frihedsgrad.
Tæthedsfunktionen for den ses på Figur 3.



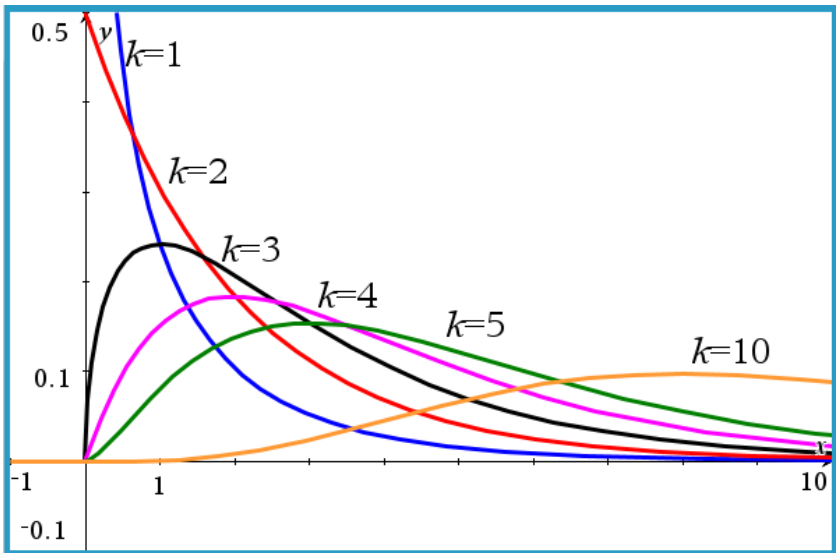
Figur 3

Arealet under denne tæthedsfunktion i intervallet [0;1] vil være det samme som arealet under tæthedsfunktionen for standard-normalfordelingen i intervallet [-1;1], hvilket ses på Figur 4.



Figur 4

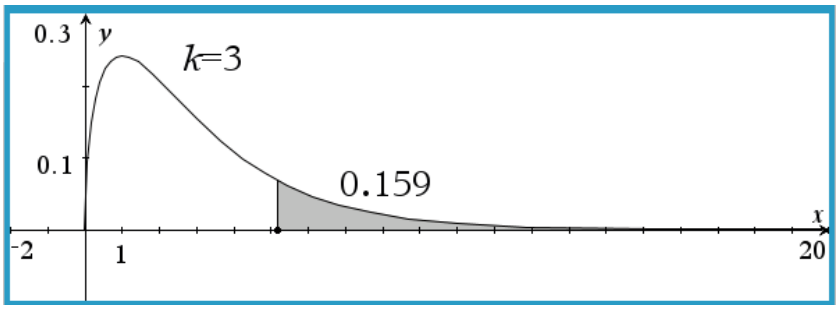
Hvis man lægger flere og flere af disse positive stokastiske variabler sammen, vil man få stokastiske variabler, der bliver større og større. Tæthedsfunktionen for χ^2-fordelingen "flytter" sig derfor mod højre, jo flere frihedsgrader der er, hvilket ses på Figur 5.



Figur 5

Når man laver et χ^2-test, beregner man teststørrelsen $\chi^2 = \sum \frac{(obs-forv)^2}{forv}$.
Hver brøk angiver afvigelsen for én observation. Jo flere observationer der er, desto større en teststørrelse forventer vi. Det giver derfor god mening at tage højde for antallet af brøker, når teststørrelsen vurderes.
Jo flere brøker der er, desto større en teststørrelse kan vi acceptere uden af forkaste nulhypotesen.
Når man vurderer teststørrelsen, er man interesseret i at vide, hvad sandsynligheden er, for at få en teststørrelse der er det samme eller større, da dette er sandsynligheden for at få det samme eller noget der passer endnu dårligere med nulhypotesen ved en ny stikprøve.
p-værdien beregnes derfor som arealet under grafen i intervallet [χ^2-teststørrelsen ; ∞ [.

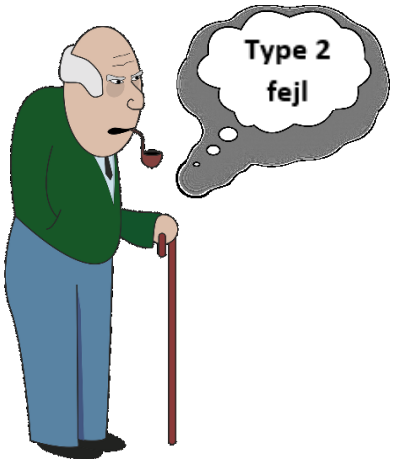
I eksemplet med majskolberne fik vi en χ^2-teststørrelse på 5,1785. Der var 4 kategorier og dermed 3 frihedsgrader. p-værdien beregnes derfor som arealet under tæthedsfunktionen med 3 frihedsgrader i intervallet [5,1785 ; ∞ [.
Arealet bliver 0.159, hvilket ses på Figur 6.
p-værdien bliver derfor 15,9 %, hvilket præcis var det, vi fandt i afsnit 2.5 side 8.



Figur 6

7. Appendiks om Type 1 og Type 2 fejl

Når man laver statistiske tests, kan man aldrig være 100 % sikker på, at man kommer frem til den rigtige konklusion.

Nulhypotese: **Personen er gravid**	Nulhypotesen er **sand**	Nulhypotesen er **falsk**
Nulhypotesen **Accepteres**		
Nulhypotesen **Forkastes**		

Figur 7

Kilde:
https://blog.bionyt.dk/biostatistik-statistik/
https://pixabay.com/da/illustrations/graviditet-mor-kommende-mor-gravid-2700659/
https://pixabay.com/da/vectors/sukkerr%C3%B8r-%C3%A6ldre-bedstefar-gnaven-1293369/
https://pixabay.com/da/vectors/t%C3%A6nker-taleboble-tegneserie-bobler-148170/

Type 1 fejl

Når man udfører et statistisk test, kan man komme ud for, at man observerer noget, der stemmer meget dårligt overens med nulhypotesen, selv om den egentlig er sand. I den situation vil man forkaste en sand nulhypotese. Fejl af denne type kaldes Type 1 fejl.

I eksemplet nederst til venstre forkastes hypotesen om graviditet, selv om kvinden er gravid. Her forkastes en sand nulhypotese.

Denne fejl vil forekomme i 5 % af de tilfælde, hvor nulhypotesen er sand, hvis testet udføres ved et signifikansniveau på 5 %.

Når man udfører et test ved et signifikansniveau på 5 %, betyder det jo netop, at man forkaster nulhypotesen, hvis man observerer noget, der er under 5 % chance for, velvidende, at det faktisk forekommer i 5 % af tilfældene.

Type 2 fejl

Omvendt kan man også komme ud for, at man observerer noget, der stemmer fint overens med nulhypotesen på trods af, at den er falsk. I dette tilfælde accepterer man en falsk nulhypotese. Fejl af denne type kaldes Type 2 fejl.

I eksemplet øverst til højre accepteres hypotesen om graviditet, selv om manden ikke er gravid. Her accepteres en falsk nulhypotese.

Der er ikke en bestemt procentdel fejl af denne type. Procentdelen afhænger af, hvor meget nulhypotesen afviger fra den sande situation.

Opgaver

Opgave 1 (Mendels 1. lov)
Dominante og recessive gener var ikke en term man kendte til før år 1865, da den tjekkiske munk Gregor Johann Mendel publicerede sine resultater fra de forsøg, han havde lavet med ærteplanter i klosterets baghave.
Kilde: https://www.biotechacademy.dk/undervisning/gymnasiale-projekter/genetik/mendels-arvelighed/

Gregor Johann Mendel studerede, hvordan egenskaber nedarves i ærteplanter. Ud fra sine eksperimenter formulerede han 2 arvelighedslove.
Mendel mente, at det første eksperiment var udformet sådan, at han skulle se en 1:3 udspaltning i gule og grønne. Resultatet af Mendels eksperiment fremgår af tabellen.

Farve	gule	grønne
Antal ærter	152	428

Kilde: https://data.math.au.dk/interactive/istar/Bog101.html

Forsøget betragtes som en stikprøve af en større population af ærteplanter.

a) Undersøg ved et χ^2-test om Mendels population af ærteplanter følger hans biologiske model om 1:3 udspaltning i gule og grønne.

Opgave 2 (Mendels 2. lov)
Georg Mendel lavede desuden ærteforsøg for at underbygge sin hypotese om 2-gen udspaltning, hvor to uafhængige gener giver 4 fænotyper, der vil fordele sig i forholdet 9:3:3:1.
Resultatet af Mendels eksperiment fremgår af tabellen.

Udseende	Gul og rund	Grøn og rund	Gul og rynket	Grøn og rynket
Antal ærter	315	108	101	32

Kilde: http://bionyt.s807.sureserver.com/biostatistik-statistik/

Forsøget betragtes som en stikprøve af en større population af ærteplanter.

a) Undersøg ved et χ^2-test om Mendels population af ærteplanter følger hans biologiske model om 9:3:3:1 udspaltning i de fire fænotyper.

Opgave 3

En gruppe forskere ønsker at undersøge, om genet *RNASE3* har betydning for, om personer smittet med malaria udvikler den dødelige hjernemalaria. Hos mennesket findes to alleler af *RNASE3*-genet: R1 og R2. Tidligere undersøgelser har vist, at en rask population vil være i Hardy-Weinberg ligevægt mht *RNASE3*-genet.

Forskerne ønsker at undersøge, om børn fra Ghana med hjernemalaria er i Hardy-Weinberg ligevægt mht *RNASE3*-genet. De udtager derfor en stikprøve blandt børn fra Ghana med hjernemalaria.
Resultatet fremgår af tabellen.

Genotype	R1R1	R1R2	R2R2
Antal børn	116	62	31

Kilde: BIOASTX, 30. maj 2017, opgave 4:
https://www.xn--prvebanken-1cb.dk/proevematerialer/GYMUDD/BIO/BIOASTX/materialesamling/74d71a50-4409-499f-9f3a-3c029ca8a3b3

a) Beregn allelfrekvenserne p(R1) og q(R2).
b) Undersøg ved et χ^2-test om denne population af børn fra Ghana med hjernemalaria er i Hardy-Weinberg ligevægt.

Opgave 4

Genet for mælkeproteinet β-casein forekommer i to alleler, A1 og A2.
En gruppe forskere ønsker at undersøge, om en population af jerseykøer er i Hardy-Weinberg ligevægt mht. β-casein-genet. De udtager derfor en stikprøve blandt jerseykøerne. Resultatet fremgår af tabellen.

Genotype	A1A1	A1A2	A2A2
Antal jerseykøer	489	241	22

Kilde: BIOASTX, 26. august 2019, opgave 2:
file:///C:/Users/jett0742/Downloads/Biologi%20A,%20stx,%20gammel%20ordning,%2026.%20august%202019.pdf

a) Beregn allelfrekvenserne p(*A1*) og q(*A2*).
b) Undersøg ved et χ^2-test om denne population af jerseykøer er i Hardy-Weinberg ligevægt.

Opgave 5

Tidligere undersøgelser har vist, at fordelingen af blodtyper i den danske befolkning er:

Blodtype	0 (nul)	A	B	AB
Procent (DK)	41	44	10	5

Kilde: https://givblod.dk/fakta-om-blod/#blodtypefordeling

En gruppe forskere ønsker at undersøge, om fordelingen af blodtyper er den samme i Japan.
De udtager derfor en stikprøve blandt befolkningen i Japan. Resultatet fremgår af tabellen.

Blodtype	0 (nul)	A	B	AB
Antal (Japan)	75	95	55	25

Kilde: https://science-gym.dk/mat/20002010/Fordeling_af_AB0.doc

a) Undersøg ved et χ^2-test om fordelingen af blodtyper er den samme i Japan som i Danmark.

Opgave 6

Tidligere undersøgelser har vist, at ca. 6,2 procent af den danske befolkning har diabetes.
(type-1-diabetes: 0,6%, type-2-diabetes: 5,5 %, andre typer: 0,1 %).

Kilde: https://videncenterfordiabetes.dk/viden-om-diabetes/generelt-om-diabetes/diabetes-i-tal

En gruppe forskere ønsker at undersøge, om forekomsten af diabetes er den samme i USA som i Danmark.
De udtager derfor en stikprøve blandt befolkningen i USA. Resultatet fremgår af tabellen.

Diabetes-status	diabetes	Ikke-diabetes
Antal amerikanere	113	887

Kilde: https://diabetesresearch.org/diabetes-statistics/

a) Undersøg ved et χ^2-test om forekomsten af diabetes er den samme i USA som i Danmark.