Stx

Binomialtest i Biologi A

2. Udgave

Jette H Vestergaard

Dronninglund Gymnasium

Januar 2022

Indholdsfortegnelse

Forord

I dette hæfte skal vi beskæftige os med binomialforsøg og binomialtest. Begge dele er kernestof på stx matematik B. På stx biologi A er binomialtest ikke kernestof, men da man i matematik har valgt, at binomialtest er det eneste test indenfor kernestoffet efter reform 17, har man i biologi også valgt at inddrage binomialtest i biologiopgaverne, idet fagene dermed supplerer hinanden.
Dette hæfte er skrevet til elever på biologi A. Der er derfor visse matematiske detaljer, som er udeladt, da de behandles på matematik B men ikke forventes behandlet på biologi A. Imidlertid vil eleverne sagtens kunne udføre beregningerne, da de har lært dem på matematik B. Vi vil blot ikke gennemgå dem her.

1. Binomialfordeling

1.1 Binomialforsøg

Et binomialforsøg består af en række uafhængige gentagelser af et bestemt eksperiment, **basiseksperimentet**. At gentagelserne er uafhængige betyder, at udfaldene af gentagelserne af basiseksperimentet ikke afhænger af hinanden.
F.eks. er 10 kast med en mønt eller 10 kast med en terning eksempler på binomialforsøg. Hvert af de 10 udfald er uafhængig af de andre udfald.
Ved hver gentagelse af basiseksperimentet interesserer vi os for et bestemt udfald, som vi kalder **succes**. Dét eller de andre udfald kaldes **fiasko**.
Ved kastet med mønten kunne succes være udfaldet **plat**. Og ved kastet med terningen kunne succes være udfaldet **sekser**.
Sandsynligheden for succes kaldes p og betegnes **sandsynlighedsparameteren**.
Ved kastet med mønten er $p = 1/2$, og ved kastet med terningen er $p = 1/6$. I hvert fald hvis der er tale om en ærlig mønt og en ærlig terning.
Antallet af gentagelser kaldes n og betegnes **antalsparameteren**.
I både forsøget med mønten og terningen er $n = 10$.
Ved et binomialforsøg tæller man antal succes'er. Lad X være lig med antallet af succes'er. Man siger, at X er binomialfordelt med antalsparameter n og sandsynlighedsparameter p. Den korte skrivemåde for dét er: $X \sim b(n, p)$.
De to binomialforsøg kan kort beskrives på følgende måde:

<u>Forsøg: 10 kast med ærlig mønt</u>
Basiseksperiment: Ét kast med en mønt.
Succes: Plat.
Sandsynlighedsparameter: $p = 1/2$.
Antalsparameter: $n = 10$.
X = Antal succes'er (plat).
$X \sim b(10, 1/2)$.

<u>Forsøg: 10 kast med ærlig terning</u>
Basiseksperiment: Ét kast med en terning.
Succes: Sekser.
Sandsynlighedsparameter: $p = 1/6$.
Antalsparameter: $n = 10$.
X = Antal succes'er (sekser).
$X \sim b(10, 1/6)$.

Nogle forsøg handler om at trække en stikprøve fra en population. I sådanne tilfælde er der kun tale om et binomialforsøg, hvis der er tale om en **stikprøve med tilbagelægning**, da der så er samme sandsynlighed for succes hver gang.

Hvis der er tale om en **stikprøve uden tilbagelægning** vil sandsynligheden for succes ændre sig ved hvert eneste træk og afhænge af, om der blev trukket en succes i forrige træk eller ej.
Imidlertid vil vi alligevel ofte anvende binomialfordeling i forbindelse med stikprøver uden tilbagelægning. Det gør vi i de tilfælde, hvor stikprøven er meget lille i forhold til populationen. I sådanne tilfælde ændrer sandsynligheden for succes sig kun ganske lidt, uanset om man trækker en succes eller ej, så fejlen, vi begår, når vi siger, at der er samme sandsynlighed for succes hver gang, er ubetydeligt lille.
Som tommelfingerregel siger vi, at binomialtestet kan anvendes i forbindelse med stikprøver uden tilbagelægning, når stikprøven maksimalt udgør 10 % af populationen.[1]

1.2 Binomialfordeling

Vi er nu interesserede i at finde sandsynlighederne for de forskellige antal succes'er. Hvis der er 10 gentagelser af basiseksperimentet vil der være mulighed for mellem 0 og 10 succes'er.
Vi ønsker nu at finde sandsynligheden for at X er lig med k (altså sandsynligheden for k succes'er), for alle værdier af k mellem 0 og 10.
Den korte måde at skrive "sandsynligheden for at X er lig med k" er $P(X = k)$. Her står P'et for probability, som betyder sandsynlighed på engelsk. Denne sandsynlighed kaldes en **punktsandsynlighed**.

I forsøget med 10 kast med mønt vil vi forvente, at der cirka vil komme 5 plat'er, når vi kaster 10 gange, da sandsynligheden for plat er 1/2. Men vi ved godt, at man også kan få 4 plat eller 3 plat eller… Vi forventer bare, at sandsynligheden for 5 plat er størst og sandsynligheden for 0 eller 10 plat er mindst.
Den præcise metode til at beregne de forskellige sandsynligheder overlader vi til matematikerne.
Vi nøjes med at beregne dem i Excel.

Åben et Excel regneark og skriv følgende ind i de første 4 linjer:

	A	B	C	D
1	Basiseksperiment: Ét kast med ærlig mønt.			
2	Succes: Plat.			
3	Sandsynlighedsparameter (sandsynlighed for for succes): p=0,5			
4	Antalsparameter (antal gentagelser): n=10			

I celle A6 skrives nu **Antal succes'er**
Derefter indtastes tallene 0, 1, 2, … , 10 i cellerne A7 til A17.

I celle B6 skrives **Sandsynlighed**
Vi vil nu gerne have sandsynlighederne for de forskellige antal succes'er i cellerne B7 til B17.
Disse sandsynligheder findes vha et forprogrammeret værktøj i Excel.
I celle B7 skrives nu **= BINOMIAL.FORDELING (A7 ; 10 ; 0,5 ; FALSK)**
A7 fordi succes-værdien står i celle A7.
10 fordi der er 10 gentagelser af forsøget.
0,5 fordi sandsynlighedsparameteren er 0,5.

[1] Kilde: Jan Sørensen, Aalborg City Gymnasium

FALSK giver umiddelbart ikke nogen mening. Men det skyldes, at denne kommando også bruges til at beregne en anden slags sandsynligheder, nemlig de kumulerede frekvenser. Og man har så valgt, at der ved den ene slags skal skrives FALSK, og ved den anden skrives SAND.
Når vi har tastet Enter giver det 0,000977.

Vi vil gerne have sandsynlighederne angivet med 3 decimaler. Klik derfor på knappen  et antal gange indtil der er 3 decimaler. Hvis sandsynligheden angives med eksponentiel notation, f.eks. 4,5E-12, klikker man først på knappen med de 3 nuller for at få tallet skrevet på normal vis, inden man vælger antal decimaler.

Nu skal vi have kopieret formlen over i cellerne B8 til B17. Klik derfor i celle B7 og flyt musen ned i højre hjørne, så der kommer et sort kryds. Hold venstre musetast nede og træk ned til celle B17. Nu bliver formlen kopieret over i de andre celler og nu skulle det helst se således ud:

6	Antal succes'er	Sandsynlighed
7	0	0,001
8	1	0,010
9	2	0,044
10	3	0,117
11	4	0,205
12	5	0,246
13	6	0,205
14	7	0,117
15	8	0,044
16	9	0,010
17	10	0,001

Det ses, at sandsynligheden for 5 succes'er er 24,6 %, og det er, som vi forventede, det antal succes'er, der er det mest sandsynlige.
0 eller 10 succes'er er det mindst sandsynlige.

Disse sandsynligheder kaldes en **binomialfordeling**, fordi de beskriver fordelingen af sandsynlighederne i et binomialforsøg.

Med den korte notation skrives punktsandsynlighederne
$P(X = 0) = 0{,}001$
$P(X = 1) = 0{,}010$
$P(X = 2) = 0{,}044$
…
$P(X = 10) = 0{,}001$

Vi vil nu lave et diagram over punktsandsynlighederne.

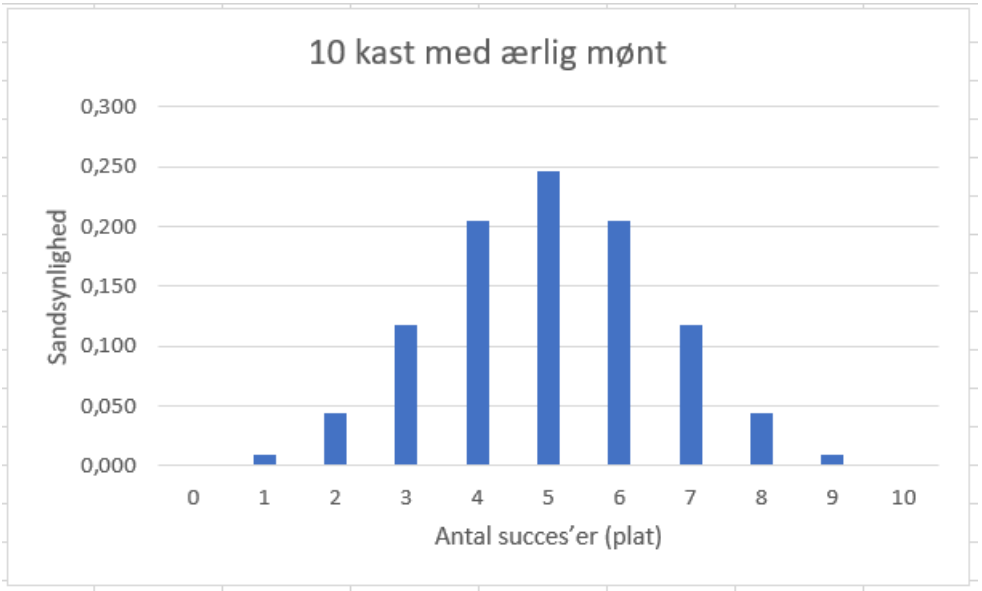
OBS: På nogle mac computere er proceduren en lidt anden. Hvis nedenstående procedure ikke giver det rigtige resultat på din mac, så følg vejledningen på bilaget til dette hæfte.

Klik i celle A7, så der kommer et fedt hvidt kryds. Hold venstre musetast nede og træk ned til celle B17. Nu bliver cellerne A7 til A17 og B7 til B17 markeret. Vi ønsker nu at lave et søjlediagram over sandsynlighederne.

Klik på fanen **Indsæt** og vælg **Anbefalede diagrammer**.
Her vælges **søjlediagrammet**.

Vi vil gerne have skrevet tekst på akserne og også have skrevet en titel på diagrammet.
Klik et tilfældigt sted i diagrammet, så der kommer en ramme med små cirkler rundt om diagrammet. Samtidig kommer der tre små knapper til højre for diagrammet. Klik på den øverste knap med krydset.
Sæt et flueben ved **Aksetitler**. Der er allerede sat et flueben ved **Diagramtitel**. Nu kan man både skrive en diagramtitel og skrive tekst ved akserne.

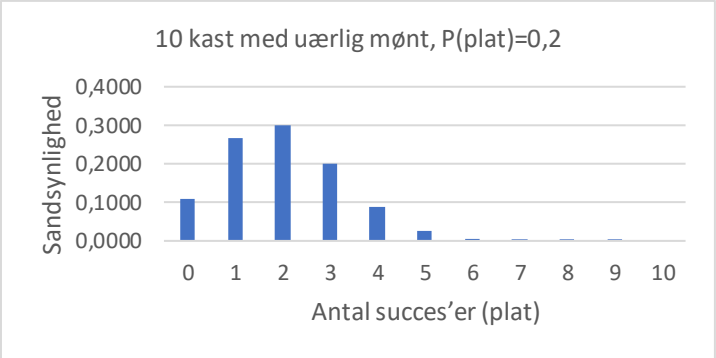
Skriv ved x-aksen: Antal succes'er (plat).
Skriv ved y-aksen: Sandsynlighed.
Skriv som diagramtitel: 10 kast med ærlig mønt.
Diagrammet ser nu således ud:



Diagrammer af denne type er centrale i det **binomialtest**, som beskrives i næste afsnit.

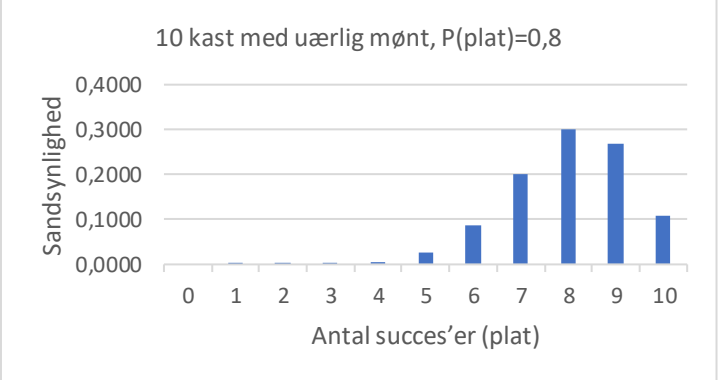
Hvis sandsynligheden for plat er mindre end 0,5, f.eks. $p = 0{,}2$, vil vi forvente at det mest sandsynlige antal succes'er er mindre end 5.

I det tilfælde vil diagrammet se således ud:



Og hvis sandsynligheden for plat er større end 0,5, f.eks. $p = 0{,}8$, vil vi forvente at det mest sandsynlige antal succes'er er større end 5.

I det tilfælde vil diagrammet se således ud:



2. Binomialtest

Hvis vi ønsker at undersøge om en hypotese er rigtig eller forkert, udfører vi ofte et forsøg for at undersøge hypotesen. Hvis forsøget er udformet som et binomialforsøg, kan binomialfordelingen anvendes til at afgøre, om hypotesen er rigtig eller forkert.
Der findes to slags binomialtest. Tosidet test og etsidet test. I biologi anvendes kun det tosidede test.

2.1 Tosidet binomialtest[2]

Vi betragter et eksempel.
Tidligere undersøgelser har vist, at 10 % af befolkningen i Danmark har blodtype B.
I en stikprøve på 90 personer blev 15 testet til at have blodtype B. Giver stikprøven grund til mistanke om en ændring i blodtype-B-andelen ?

Umiddelbart, så udgør de 15 personer $\frac{15}{90} = 0{,}167 = 16{,}7\ \%$ af stikprøve, hvilket er mere end de 10 %, som plejer at have blodtype B. Spørgsmålet er nu, om denne afvigelser er så stor, at vi kan konkludere, at der er sket en ændring. Det er dét, vi kan bruge binomialtestet til at afgøre.

Når man skal lave et binomialtest (eller andre test), skal man starte med at opstille en **nulhypotese**. Vores hypotese her er, at der er sket en ændring i blodtype-B-andelen. Imidlertid er det ikke dét, der skal stå i vores nulhypotese. Nulhypotesen skal være, at der ikke er sket nogen ændring. Vores formodning er så, at nulhypotesen forkastes. Nulhypotesen skal altid indeholde nogle kendte sandsynligheder, som vi kan regne videre med. I dette tilfælde kommer nulhypotesen til at se således ud:

H_0 : Blodtype-B-andelen er uændret, dvs der er stadig 10 %, der har blodtype B.

Vi skal også have formuleret en **alternativ hypotese**, som er det modsatte af nulhypotesen, og som er dét, der gælder, hvis nulhypotesen forkastes. Den alternative hypotese bliver her:

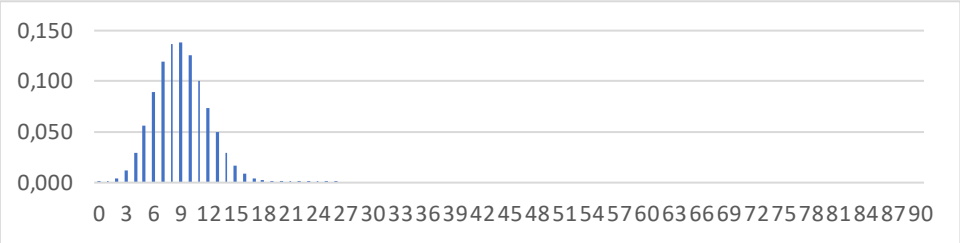
H_A : Blodtype-B-andelen har ændret sig.

Bemærk! (og nu bliver det lidt tricky!) Man kunne her godt fristes til at opstille den alternative hypotese, at andelen af blodtype B er steget, fordi vi jo kan se, at det er den i stikprøven. Det må vi imidlertid ikke! Den alternative hypotese skal altid opstilles inden stikprøven tages. Hvis man så alligevel opstiller den efter, at stikprøven er taget, skal man forestille sig, at man ikke har den viden, som stikprøven giver én, når man opstiller den alternative hypotesen.

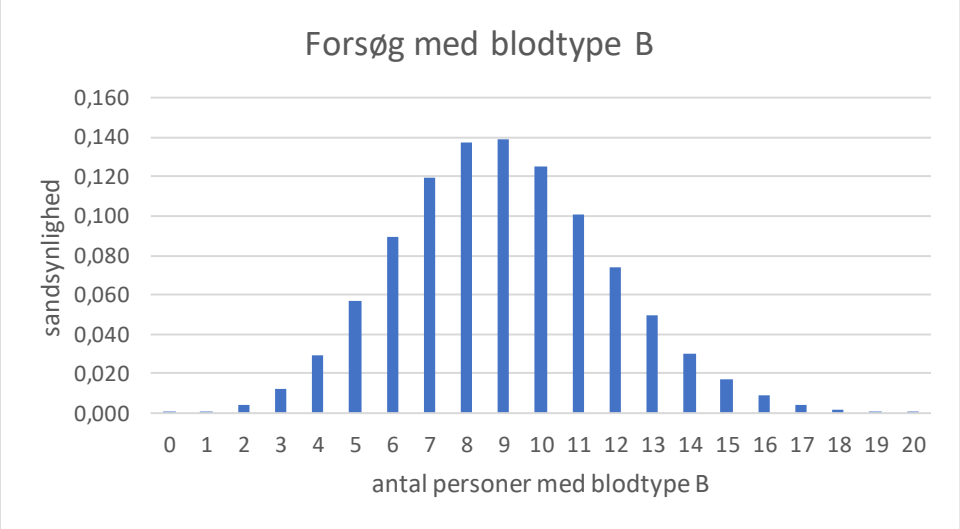
Vi skal nu i gang med at teste denne nulhypotese.
Der er her tale om en stikprøve uden tilbagelægning, men da stikprøven (90 personer) er meget lille i forhold til hele befolkningen (ca. 5,8 mio), kan vi betragte forsøget som værende med tilbagelægning.
Dermed kan vi tillade os at betragte forsøget som et binomialforsøg, som vi kort kan beskrive på følgende måde:

[2] Inspireret af Carstensen, MAT B2, s. 300-301.

Basiseksperiment: Udvælge én person.
Succes: Blodtype B.
Sandsynlighedsparameter: $p = 0{,}10$.
Antalsparameter: $n = 90$.
X = Antal succes'er.
$X \sim b(\,90\,;\,0{,}10\,)$.
Den observerede værdi af X (teststørrelsen) er $x_0 = 15$.
Vi kan nu beregne punktsandsynlighederne for X-værdier mellem 0 og 90. Det gør vi som i forrige afsnit.
Diagrammet over punktsandsynlighederne kommer til at se således ud:



Det ses, at punktsandsynlighederne for X-værdier over 20 er ubetydeligt små. Vi tegner derfor lige diagrammet igen, for X-værdier mellem 0 og 20.



Det ses, at 9 personer er det mest sandsynlige. Det var også hvad vi forventede (10 % af 90 personer).
Spørgsmålet er bare, om 15 personer er så lidt sandsynligt, at vi kan konkludere, at der er sket en ændring i blodtype B-andelen i hele befolkningen.

Vi vælger som regel at sige, at hvis vi har observeret noget, der er under 5 % sandsynligt, så må vores startantagelse (altså vores nulhypotese) være forkert.
I praksis kan det selvfølgelig godt ske, at man observerer noget meget usandsynligt. Man kan i princippet godt få 80 seksere, hvis man kaster 100 gange med en terning. Men hvis det sker, så tror vi alligevel mere på, at det er en snyde-terning (altså at nulhypotesen: at det er en ærlig terning, er forkert), end vi tror på, at vi har observeret noget ganske usandsynligt.
De 5 % kaldes testets **signifikansniveau**.

Vi skal altså have fundet de 5 % mindst sandsynlige observationer i vore blodtypeforsøg. Observationer langt fra de forventede 9 personer er de mest kritiske for nulhypotesen. Vi skal derfor have fundet de yderste observationer i begge sider (i halerne af binomialfordelingen), som i hver side har en samlet sandsynlighed på maksimalt 2,5 %.
Da det er svært at aflæse præcist på diagrammet, betragter vi derfor punktsandsynlighederne, som ses i tabellen:
Det ses, at

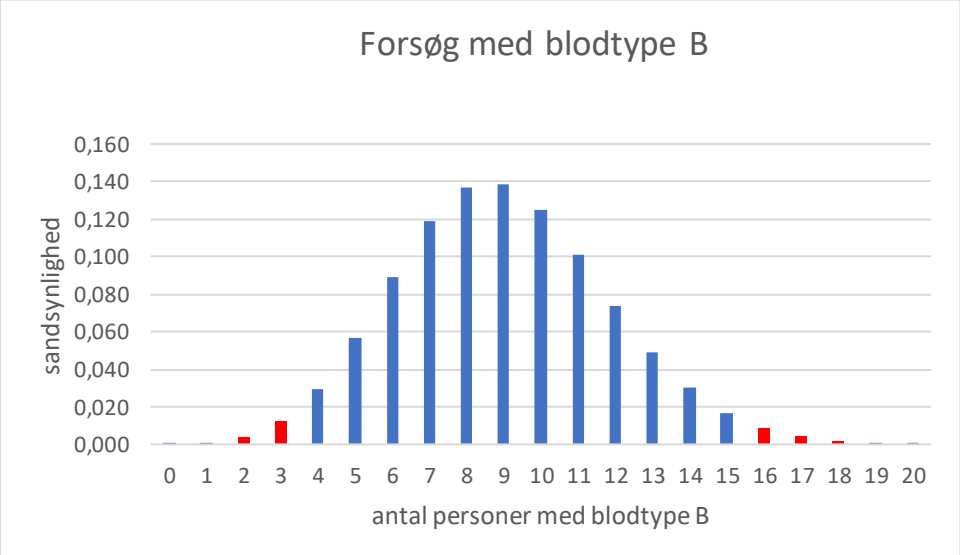
$$\begin{aligned} &P(X=0)+P(X=1)+P(X=2)+P(X=3)\\ &=0{,}000+0{,}001+0{,}004+0{,}012\\ &=0{,}017\\ &=1{,}7\,\% \end{aligned}$$

Hvis vi prøver at lægge $P(X=4)$ til, giver det 4,7 %, hvilket er langt over de 2,5 %. Altså er $X=4$ ikke en kritisk værdi.
Dvs at $X=0$, $X=1$, $X=2$ og $X=3$ er de kritiske værdier til venstre.
Til højre ses det, at

$$\begin{aligned} &P(X=19)+P(X=18)+P(X=17)+P(X=16)\\ &=0{,}001+0{,}002+0{,}004+0{,}009\\ &=0{,}016\\ &=1{,}6\,\% \end{aligned}$$

Antal succeser	sandsynlighed
0	0,000
1	0,001
2	0,004
3	0,012
4	0,030
5	0,057
6	0,089
7	0,119
8	0,137
9	0,139
10	0,125
11	0,101
12	0,074
13	0,049
14	0,030
15	0,017
16	0,009
17	0,004
18	0,002
19	0,001
20	0,000

Hvis vi lægger $P(X=15)$ til kommer vi igen over de 2,5 %. Altså er værdier af X kritiske fra 16 og opefter.
Samlet set er den kritiske mængde derfor: $K=\{0,\ldots,3,16,\ldots,90\}$.
Dette kunne f.eks. illustreres i diagrammet på følgende måde (hvor de røde søjler er de kritiske værdier):



Konklusion:
Da den observerede værdi $x_0 = 15$ ikke ligger i den kritiske mængde ved et signifikansniveau på 5 %, kan vi *ikke* forkaste nulhypotesen.
Altså må vi ved et signifikansniveau på 5 % konkludere, at der ud fra denne stikprøve ikke er belæg for at konkludere, at blodtype-B-andelen er ændret i hele befolkningen.

Hvis vi havde observeret 16 personer med blodtype B, så havde vi kunnet konkludere, at blodtypeandelen var ændret ved et signifikansniveau på 5 %.
Når man skriver "ved et signifikansniveau på 5 %", så signalerer man, at påstanden hverken er bevist eller modbevist. Man har blot gode argumenter for at antage det ene eller det andet.

2.2 Eksempel på opgavebesvarelse, Orangutanger på Borneo

Vi betragter opgave 4 om orangutanger i det vejledende sæt 1. Besvarelsen kunne f.eks. se således ud:

Spørgsmål 3
På Borneo er frekvensen for allelen W1 lig med p=0,69 og frekvensen for allelen W2 er q=0,31.
Da det forudsættes, at der er Hardy-Weinberg ligevægt, vil frekvenserne for de tre genotyper W1W1, W1W2 og W2W2 være p^2 , $2pq$ og q^2 .
Altså er frekvensen for W2W2 lig med $q^2 = 0{,}31^2 = 0{,}0961$.

Spørgsmål 4
Forskere har den hypotese, at orangutanger med genotype W2W2 har øget modstandsdygtighed over for malaria. De forventer derfor ikke, at populationen af orangutanger er i Hardy-Weinberg ligevægt. For at teste dette opstiller de nulhypotesen

H_0: Der er Hardy-Weinberg ligevægt i populationen, dvs frekvensen af W2W2 er 0,096.

For at teste hypotesen observerer de blandt 54 orangutanger, at 11 af dem har genotypen W2W2.

Der er tale om en stikprøve uden tilbagelægning. Da hypotesen ønskes testet ved et binomialtest, antager vi, at stikprøven er lille i forhold til populationen, altså at den højest udgør 10 % af populationen. Dvs at vi antager, at der mindst er 540 orangutanger på Borneo.
Dermed kan vi tillade os at betragte forsøget som et binomialforsøg, som vi kort kan beskrive på følgende måde:

Basiseksperiment: Udvælge én orangutang.
Succes: Genotype W2W2.
Sandsynlighedsparameter: $p = 0{,}096$.
(I dette spørgsmål rundes p op til 2 betydende cifre, i forhold til svaret i spørgsmål 3.)
Antalsparameter: $n = 54$.
X = Antal succes'er.
$X \sim b(54\ ;\ 0{,}096)$.
Den observerede værdi af X (teststørrelsen) er $x_0 = 11$.

Før vi går i gang med at teste vores nulhypotese skal vi have den alternative hypotese på plads.
Nulhypotesen er:

H_0 : Der er Hardy-Weinberg ligevægt i populationen, dvs frekvensen af W2W2 er 0,096.

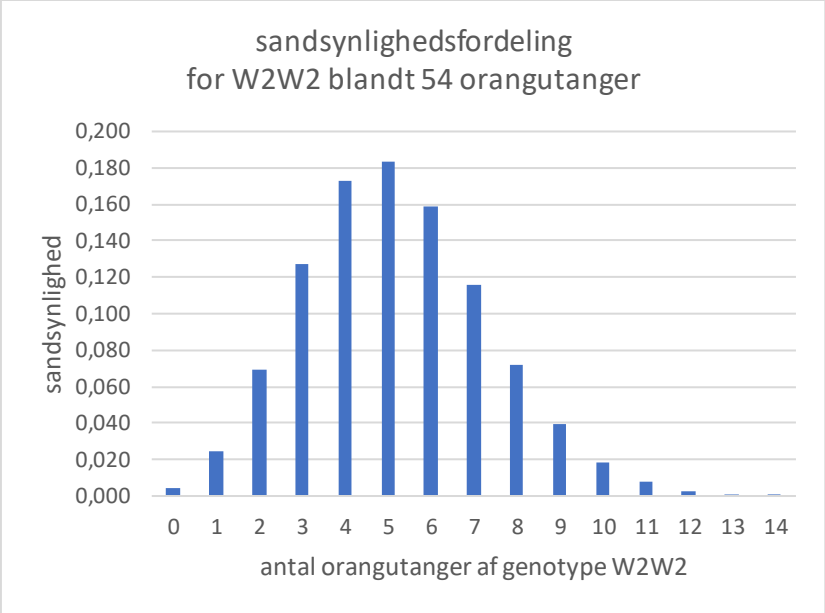
Den alternative hypotese bliver så:

H_A : Der er *ikke* Hardy-Weinberg ligevægt i populationen.

Punktsandsynlighederne beregnes i Excel vha værktøjet BINOMIAL.FORDELING. Samtidig tegnes et diagram over dem.

I tabellen og i diagrammet vises kun punktsandsynligheder over 0,001. De resterende punktsandsynligheder er alle mindre end 0,001.

Antal succes'er	sandsynlighed
0	0,004
1	0,025
2	0,069
3	0,128
4	0,173
5	0,184
6	0,159
7	0,116
8	0,072
9	0,039
10	0,019
11	0,008
12	0,003
13	0,001
14	0,000



Diagrammet svarer til diagrammet på opgavearket.

Det ses, at det mest sandsynlige antal orangutanger med genotype W2W2 er 5 ved en stikprøve på 54 orangutanger.

Da vi skal teste nulhypotesen ved et signifikansniveau på 5 %, skal vi have fundet de 2,5 % mindst sandsynlige observationer i hver side (hale) af binomialfordelingen.

Det ses, at $X = 0$ er kritisk værdi i venstre side. $X = 1$ er ikke en kritisk værdi, da 0 og 1 til sammen har en sandsynlighed over 2,5 %.

Det ses, at $X = 11$ og X-værdier over 11 er kritiske værdier i højre side. $X = 10$ er ikke en kritisk værdi, da den samlede sandsynlighed i højre side så ville blive over 2,5 %.

Den kritiske mængde bliver derfor : K = { 0 , 11, 12 , … , 54 }.

Da $x_0 = 11$ ligger i den kritiske mængde, kan vi nu *forkaste* nulhypotesen ved et signifikansniveau på 5 %.

Dette ses også i diagrammet på opgavearket, hvor søjlen ved $X = 11$ er rød (de røde er de kritiske værdier).

Spørgsmål 5

Da vi i spørgsmål 4 forkastede nulhypotesen, må vi ved et signifikansniveau på 5 % konkludere, at der *ikke* er Hardy-Weinberg ligevægt i populationen.

Dette kan måske skyldes, at genotypen W2W2 er mere modstandsdygtig overfor malaria end de to andre genotyper er. Men det har vi ikke bevist her. Her har vi blot vist, at det er overvejende sandsynligt, at der ikke er Hardy-Weinberg ligevægt i populationen.

3. Opgaver

Opgave 1

Der fødes erfaringsmæssigt lidt flere kvinder end mænd. I Danmark er fødselsprocenten for piger 51 %.
På en bestemt fødegang blev der i en bestemt periode registreret 186 nyfødte, hvoraf de 107 var piger.
Dvs at der var $\frac{107}{186} = 57{,}5\ \%$ piger i stikprøven.
Undersøg om nulhypotesen

H_0 : Kønsfordelingen er den samme på denne fødegang som i resten af Danmark.

kan forkastes ved et signifikansniveau på 5 %.

Opgave 2

Ca. 1,5 % af Danmarks befolkning er rødhårede. Den røde hårfarve skyldes mutation af genet MC1R, som begge forældre skal være bærere af for at kunne få et rødhåret barn.
I Skotland undersøgte man en stikprøve på 120 personer og fandt, at 9 af dem havde rødt hår. Dvs at der var $\frac{9}{120} = 7{,}5\ \%$ rødhårede i stikprøven.
Undersøg om nulhypotesen

H_0 : Procentdelen af rødhårede er den samme i Skotland som i Danmark.

kan forkastes ved et signifikansniveau på 5 %.

Kilde: https://www.alt.dk/artikler/rodharet-alt-du-skal-vide

Litteratur

Ballund, Henrik, m.fl., Naturfaget – en grundbog, 1. udgave, 1. oplag, 1998, Gads Forlag.

Carstensen, Jens m.fl., MAT AB2 stx opgaver, 3. udgave, 1. oplag, 2018, Systime.

Carstensen, Jens m.fl., MAT B2, 4. udgave, 1. oplag, 2018, Systime.

Clausen, Flemming m.fl., Grundbog B2, 1. udgave, 1. oplag, 2018, Gyldendal.

Hebsgaard, Thomas og Hans Sloth, Højniveaumatematik 1, 1. udgave, 2. oplag, 1999, Forlaget TRIP.

"28 fascinerende ting, du skal vide om rødhårede", 17. august 2017**",** https://www.alt.dk/artikler/rodharet-alt-du-skal-vide