Binomialfordelingstest

I en population delt op efter alder er hyppigheden af blodsygdommen seglcelleanæmi forskellig alt efter alder. Sygdommen som er autosomal recessiv skyldes en enkelt punktmutation i genet for hæmoglobin. De raske har derfor genotypen *SS* eller *Ss* og de syge har genotypen *ss*. Binomialfordelingen kan anvendes til at teste de forskellige aldersgrupper inden for populationen, fordi hver gruppe er inddelt i to grupper; nemlig raske og syge.

Børn 0-5 år

I det følgende laves en analyse af børn fra 0 til 5 år i en population med seglcelleanæmi. Data fremgår af 'Regn med biologi' side 72.

Det ses at der er 97 børn og 5 af dem har seglcelleanæmi. Under antagelse af Hardy-Weinberg-ligevægt og med en sandsynlighedsparameter på 0,04, kan man forvente 0,04 · 97 = 3,88 børn med seglcelleanæmi – reelt svarer det til 3-4 børn med seglcelleanæmi. Det er dog ikke hver gang at en gruppe på 97 børn har fx 4 børn med seglcelleanæmi. Man kan udregne denne sandsynlighed ved i Excel at indtaste formlen: =BINOMIAL.FORDELING(4;97;0,04;FALSK). Resultatet af dette giver 0,1991. Det betyder at hvis man udvalgte en anden gruppe børn mellem 0 og 5 år i samme population ville der være 19,9 % sandsynlighed for at denne gruppe indeholdt præcis 4 børn med seglcelleanæmi. Sandsynligheden for 3 børn er lidt større; nemlig 20,3 %. Generelt kan formlen der indtastes i Excel skrives på denne måde: =BINOMIAL.FORDELING(*X*;*Y*;*Z*;FALSK). *X* er det antal man vil finde sandsynligheden for ud af hele gruppen (*Y*). Herefter skrives den sandsynlighed mellem 0 og 1 som begivenheden foregår med. Skiftes 'FALSK' ud med 'SANDT' fås den kumulerede sandsynlighed, svarende til op til *X* personer.

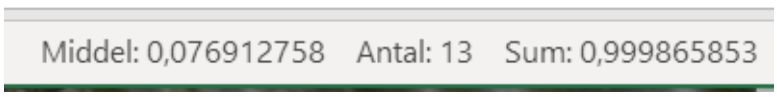
Når man skal afbillede sandsynlighederne for et bestemt antal børn der har seglcelleanæmi ud af 97 børn gøres følgende i Excel:

1) Skriv i kolonne A tallene fra 0 til 97, så 0 står i A1 og 97 står i A98. Skriv i de første 3 til 4 celler, markér disse celler og før musen ned til nederste højre hjørne i de markerede celler. Et +-tegn dukker op. Træk det med musen ned til A98.

2) Skriv i B1: '=BINOMIAL.FORDELING(A1;97;0,04;FALSK)'. Herved udregnes sandsynligheden for at netop 0 børn (det tal der står i A1) ud af 97, får seglcelleanæmi.

3) Kopier formlen i B1 således den står i alle celler ned til B98.

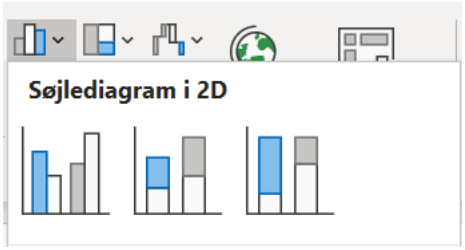
Derved har man de data man skal bruge for at lave et søjlediagram. Markér celler i B-rækken og kun i B-rækken. Begynd i B1 og hold øje med lynberegneren nede i regnearkets nederste kant, se figur 1. Når de allerfleste data er med i markeringen – og man er meget tæt på 1 i 'Sum', kan man stoppe med at markere data. Generelt er det vigtigt at få så mange tal som muligt med omkring den beregnede sandsynlighed, hvilket i dette tilfælde er 3,88.



Figur 1. En lynberegner. I dette tilfælde holdes øje med summen angivet længst til højre.

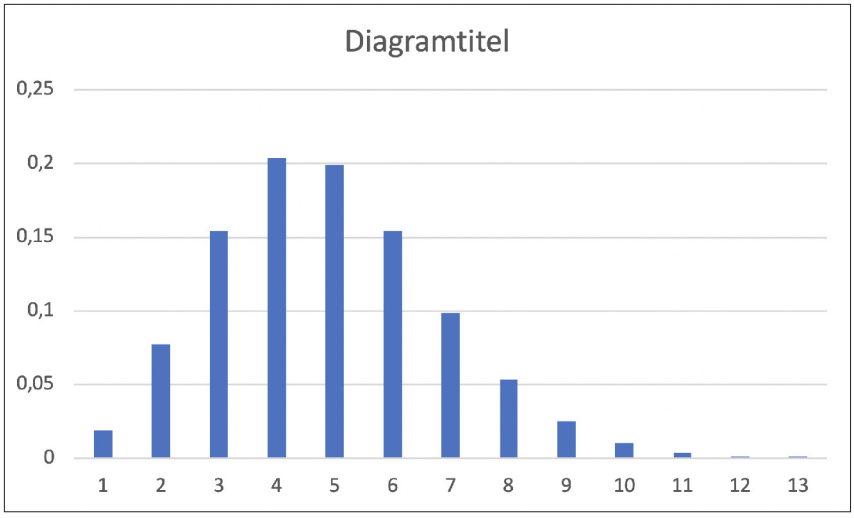
REGN MED BIOLOGI · ISBN 978-87-90363-98-7

De markerede data – i dette eksempel fra 0 til 12 børn med seglcelleanæmi – skal nu bruges til at lave et søjlediagram. Klik på fanebladet 'Indsæt' og klik i 'Diagrammer' på søjlediagram, og vælg den diagramtype længst til venstre under 'Søjlediagram i 2D', se figur 2.



Figur 2. Søjlediagram i 2D længst til venstre vælges under 'Diagrammer' i fanebladet 'Indsæt'.

Excel fremstiller automatisk et diagram som vist på figur 3. Det er *ikke* som det skal være, fx passer x-aksens værdier ikke, idet diagrammet laves med udgangspunkt i at man begynder med 1 og tæller opad. I dette tilfælde begyndes med 0, og det vil normalt være tilfældet når man laver binomialfordelinger. Desuden mangler der aksetitler. Disse fejl og mangler mangler ved Excels forslag til grafisk fremstilling kan man modificere.



Figur 3. Søjlediagram som Excel umiddelbart laver.

Modifikation af tegnet diagram

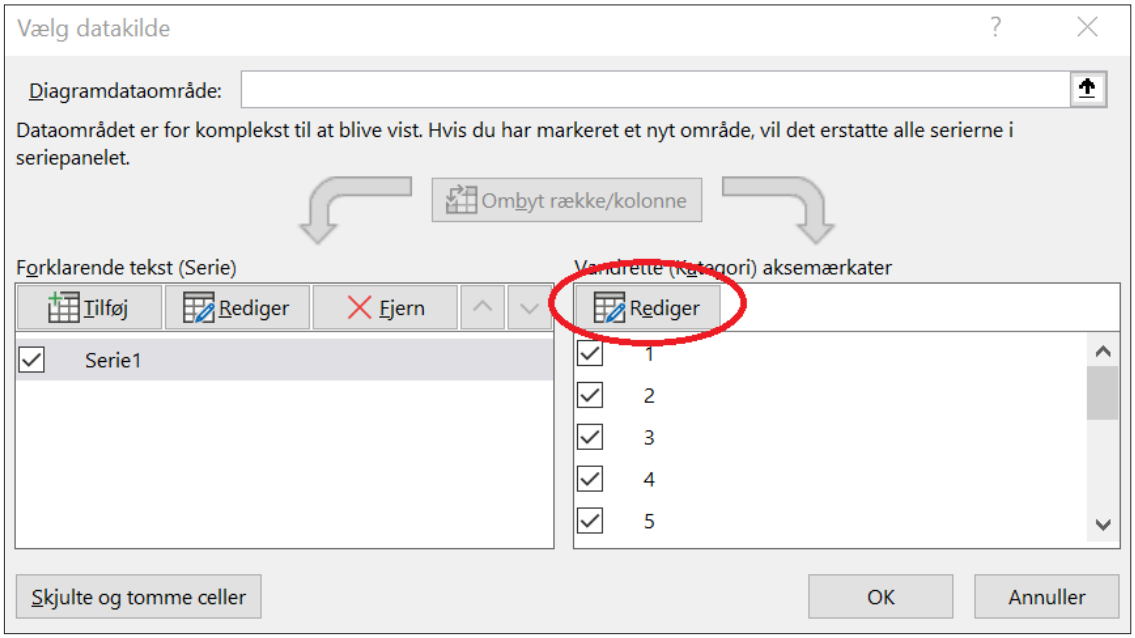
Pc (afsnit springes over hvis man anvender Mac): Højreklik på diagrammet og klik på 'Vælg data'. Klik på 'Rediger ' længst til højre, se den røde ring i figur 4. Markér cellerne A1 til A13, hvor tallene 0 til 12 står og tryk på <Enter>. Tryk derefter på 'OK'.

Mac (afsnittet springes over hvis man anvender pc): Højreklik på diagrammet og klik på 'Marker data'. Ud for 'Vandrette aksenavne (Kategori):' skal de rigtige data indsættes, se rød ring figur 5. Markér cellerne A1 til A13, hvor tallene 0 til 12 står, og tryk på <Enter>. Tryk derefter på 'OK'.

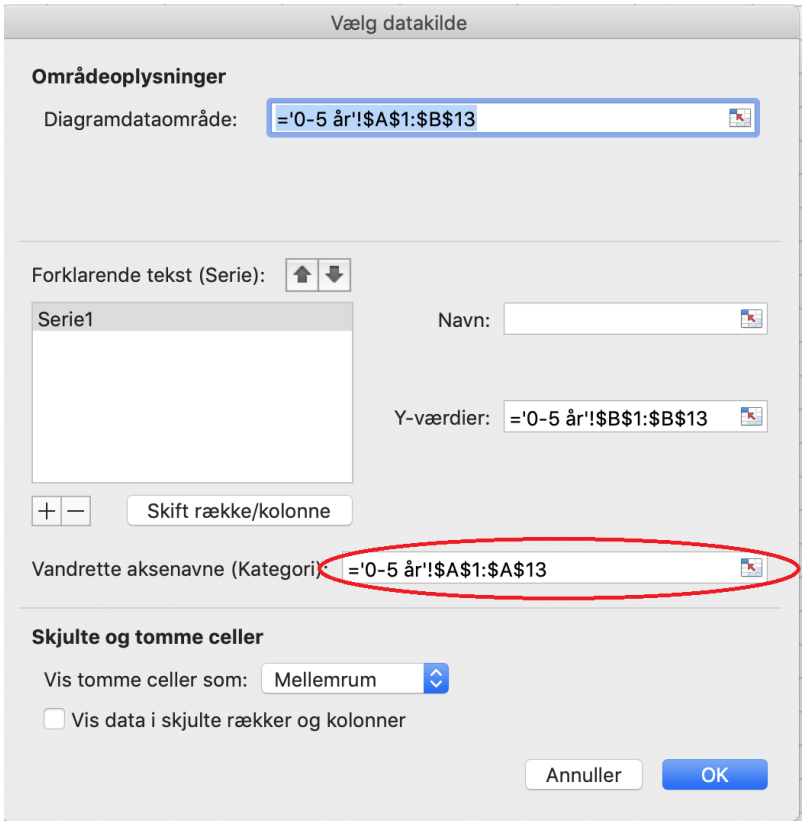
Markér felterne A1 til A13 (der hvor tallene 0 til 12 står).' Tryk herefter 'OK'. I fanebladet 'Diagramdesign' vælges 'Tilføj diagramelement'. Vælg 'Tilføj Aksetitler' → 'Primær vandret' og deref-

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

ter 'Tilføj Aksetitler' → 'Primær lodret', se figur 6. Ret nu aksetitlerne på diagrammet i regnearket, så der på x-aksen står 'Antal børn med seglcelleanæmi ud af 97 fødsler' og på y-aksetitlen står 'Sandsynligheden for seglcelleanæmi'. Dobbeltklik på 'Aksetitel' og skriv den rigtige tekst. Slet overskriften (Diagramtitel) – den skal alligevel ikke bruges. Det færdige resultat kan ses på figur 7.

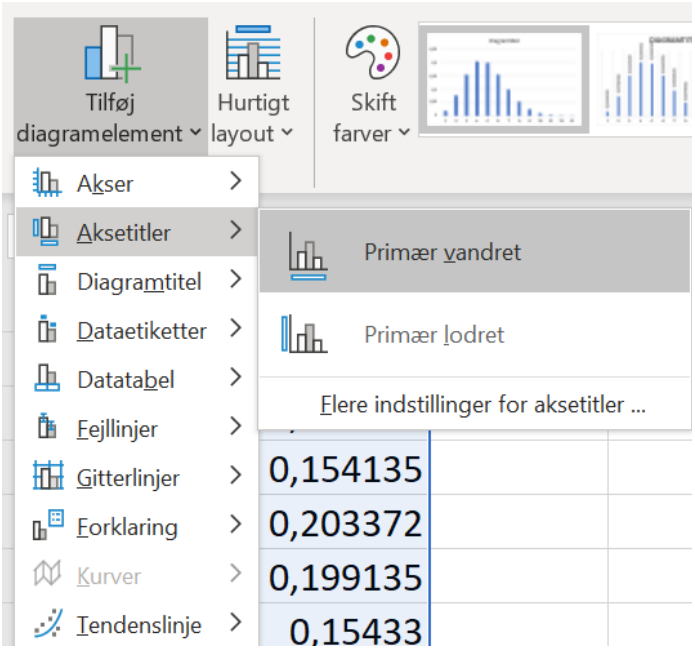


Figur 4. Ændring af dataserie på x-aksen, så den begynder med det rigtige tal – fra en pc-version af Excel.

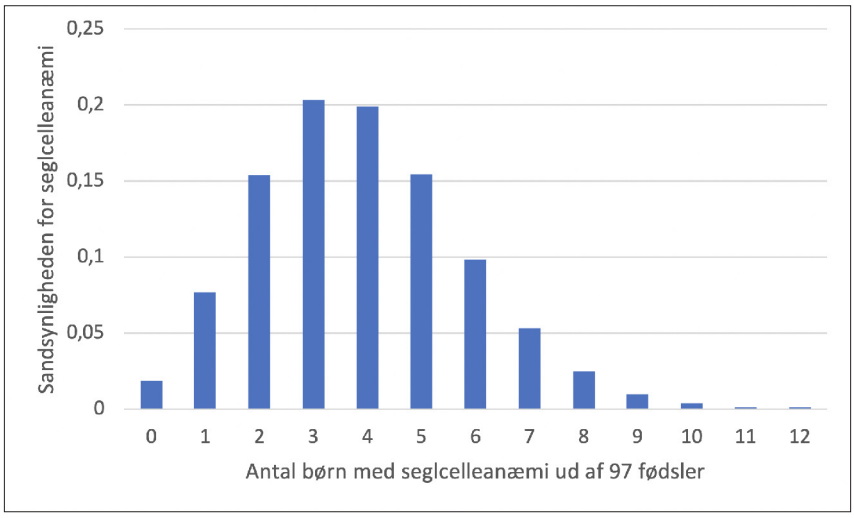


Figur 5. Ændring af dataserie på x-aksen, så den begynder med det rigtige tal – fra en Mac-version af Excel.

REGN MED BIOLOGI · ISBN 978-87-90363-98-7



Figur 6. Valg af 'Tilføj diagramelement'. Tilføj Aksetitler → Primær vandret og Primær lodret. Det skal gøres af to omgange.



Figur 7. Det færdige diagram der viser sandsynligheden for et bestemt antal børn får seglcelleancemi i en population på 97 børn.

Analyse af resultater

Tosidet test

Tidligere kunne det konstateres at 5 børn ud af 97 havde seglcelleanæmi. Man kan nu se om det er inden for normalområdet. Eller i dette tilfælde: Om der er Hardy-Weinberg-ligevægt. Først skal man beslutte sig for et konfidensinterval eller sagt med andre ord: Hvor stor en del er normalområdet og hvor stor en del er udenfor normalområdet. Det er normalt at anvende 5 % som konfidensinterval. Det skrives normalt som $p = 0{,}05$. Det vil sige at de yderste 5 % ikke er inden for normalområdet – det kaldes det kritiske område. Med en tosidet test fordeler det kritiske

område sig med 0,025 i hver sin ende af grafen. Det er sjældent at man rammer denne værdi mellem to individer præcist, og derfor er det nærmest altid sådan at ét af individerne er lidt uden for normalområdet og lidt inde i normalområdet. For at være sikker på antallet af individer der er uden for normalområdet, har man besluttet at lige så snart et individ er lidt inden for normal-området så er det i normalområdet. Forestiller man sig at der er 0 børn ud af 97 med seglcelleanæmi, kan det ses på grafen og i Excel arket at sandsynligheden for netop 0 børn ud af 97 er 0,019. Dette er under 0,025 og det er derfor ikke sandsynligt kun at have 0 børn med seglcelleanæmi. Marker herefter sandsynlighederne for 0 og 1 barn, svarende til cellerne B1 og B2 og aflæs summen nede i lynberegneren i nederste ramme (se eventuelt figur 1). Det ses at summen af B1 og B2 er 0,0961. Da det tal er større end 0,025 er det inden for normalområdet og der er 1 barn med seglcelleanæmi i en gruppe på 97 børn. Derved har man fundet det laveste antal børn som er inden for normalområdet. For at finde det største antal børn med seglcelleanæmi der normalt kan være i en gruppe med 97 børn, markerer man det nederste tal i kolonne B, svarende til B98. Ved at køre op med musen så alle tal er markerede samtidig med man holder øje med summen i lynberegneren, kan man finde det sted i B-kolonnen hvor summen holder sig under 0,025. Når man når op til B10 er summen vist i lynberegneren 0,0159. Dette tal er under 0,025. Det svarer til 9 børn, som man kan aflæse i A-kolonnen (A10). Markerer man også B9 fås en sum på 0,0408. Det er over 0,025 og derfor er det inden for normalområdet. Det tal står ud for 8 børn. Man kan derfor forvente at få mellem 1 og 8 børn med seglcelleanæmi i en gruppe på 97 børn. Det er det normale, og at have en gruppe med 5 børn med seglcelleanæmi er derfor inden for normalområdet og dermed er denne gruppe i Hardy-Weinberg-ligevægt.

Ensidet test

En ensidet test kan bruges når man fx stiller et af spørgsmålene:

1) Er der flere end forventet med et bestemt træk/en egenskab?

2) Er der færre end forventet med et bestemt træk/en egenskab?

Hvis et af de spørgsmål skal besvares, kan man anvende en ensidet test. En ensidet test placerer hele normalområdet i den ene side af fordelingskurven, og det kritiske område i den anden del af fordelingskurven. Stiller man spørgsmål 1, ligger hele det kritiske område (normalt $p = 0{,}05$) i den høje del eller den højre del af fordelingskurven. Stiller man spørgsmål 2, ligger hele det kritiske område i den lave del eller den venstre del af fordelingskurven.

I tilfældet med seglcelleanæmi er det naturligt at regne med at der er færre personer end forventet med seglcelleanæmi fordi sygdommen er dødelig. Det er derfor helt legitimt at stille spørgsmålet: Er der færre børn med seglcelleanæmi i gruppen med børn under 5 år. Det kritiske område er derfor 0,05 i det lave område. Det betyder at man ser på sandsynligheden i den lave ende i regnearket. Sandsynligheden for at få 0 børn med seglcelleanæmi er 0,019. Dette er under 0,05 og derfor er det ikke sandsynligt kun at have 0 børn med seglcelleanæmi. Herefter markeres sandsynlighederne for 0 og 1 barn, svarende til cellerne B1 og B2, og summen aflæses i lynberegneren i nederste ramme som 0,0961. Det er over 0,05, og derfor er det inden for normalområdet at have 1 barn med seglcelleanæmi. Det vil sige at 5 børn i gruppen 0 – 5 år, ikke er færre end forventet. Kun hvis der havde været 0 børn, ville der være færre børn i gruppen end forventet.

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

Hypoteseformuleringer

Når man skal teste ved at bruge binomialfordelingstests, skal man altid starte med at teste en H_0-hypotese som siger at de to grupper ikke er forskellige. Man viser at grupper er forskellige ved at forkaste H_0-hypotesen. Der er tre muligheder for binomialfordelingstests, nemlig; en tosidet test, en ensidet test, hvor man undersøger om noget er større end det H_0-hypotesen angiver (en såkaldt højrestillet binomialfordelingstest) og en ensidet test, hvor man undersøger om noget er mindre end det H_0-hypotesen angiver (en såkaldt venstrestillet binomialfordelingstest).

De har typisk formuleringer som angivet nedenfor, hvor seglcelleanæmieksemplet anvendes:

En tosidet test

H_0-hypotesen er at andelen af børn med seglcelleanæmi ikke er *forskellig fra* den seglcelleanæmifrekvens der er i hele populationen. Læg mærke til at hvis man spørger til om en gruppe *er i Hardy-Weinberg-ligevægt*, svarer det til at bruge formuleringen '*forskellig fra*', fordi man ikke tager stilling til om der er flere eller færre individer end de der svarer til at der er Hardy-Weinberg-ligevægt.

Ensidet venstrestillet test

H_0-hypotesen er at andelen af voksne med seglcelleanæmi ikke er '*mindre end*' den seglcelleanæmifrekvens der er i hele populationen.

Ensidet højrestillet test

H_0-hypotesen er at andelen af børn med seglcelleanæmi ikke er '*større end*' den seglcelleanæmifrekvens der er i hele populationen.

I det første tilfælde placerer man den angivne *p*-værdi på begge sider af fordelingsintervallet. I det andet tilfælde placerer man den angivne *p*-værdi på venstre side af fordelingsintervallet. I det tredje tilfælde placerer man den angivne *p*-værdi på højre side af fordelingsintervallet.

Børn 6-16 år og voksne

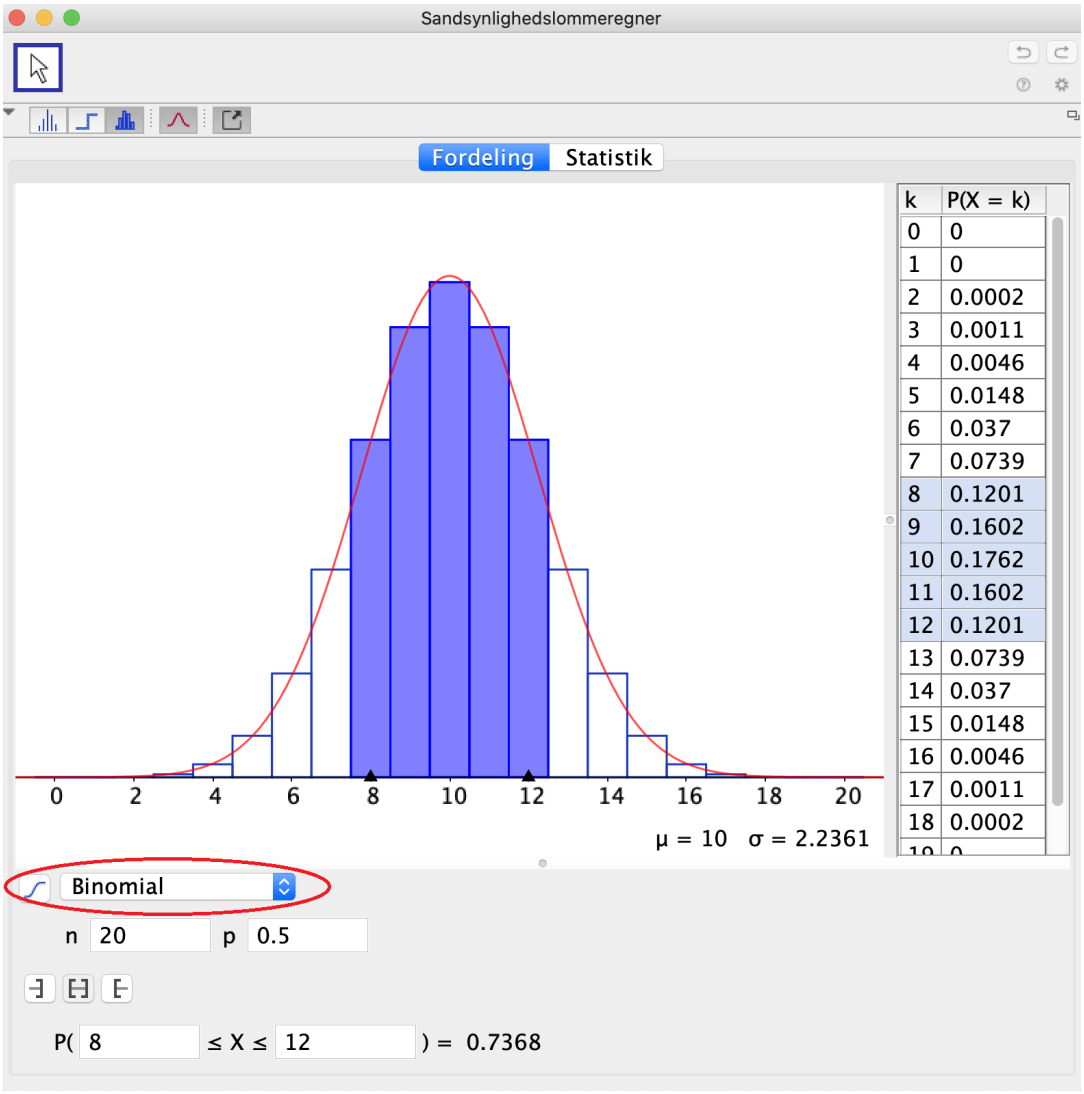
Man kan på samme måde som angivet ovenfor undersøge de to andre grupperinger for om de er i Hardy-Weinberg-ligevægt. I og med de homozygot recessive (*ss*) dør tidligt, må man forvente at der i de ældre generationer ikke er Hardy-Weinberg-ligevægt. Ligeledes kan man undersøge om hele populationen er i Hardy-Weinberg-ligevægt ved at summere alle syge og alle raske.

Brug af GeoGebras sandsynlighedslommeregner

Man kan også anvende GeoGebra som applikation for at udføre binomialfordelingstests. Det er en rigtig god og overskuelig beregner, men knap så god til afbillede data. Specielt hvis populationsstørrelsen og dermed antalsparameteren er stor, og sandsynlighedsparameteren er lille, idet det ikke umiddelbart er muligt at ændre skaleringen på x-aksen. GeoGebra anvendes på følgende måde: Start GeoGebra. Vælg 'Vis' → 'Sandsynlighedslommeregner'. Som standard begynder programmet med normalfordelingen, men ved klik på boksen, se rød ring i figur 8, ændres 'Normal' til 'Binomial'. Man har mulighed for at angive antalsparameteren (*n*) og sand-

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

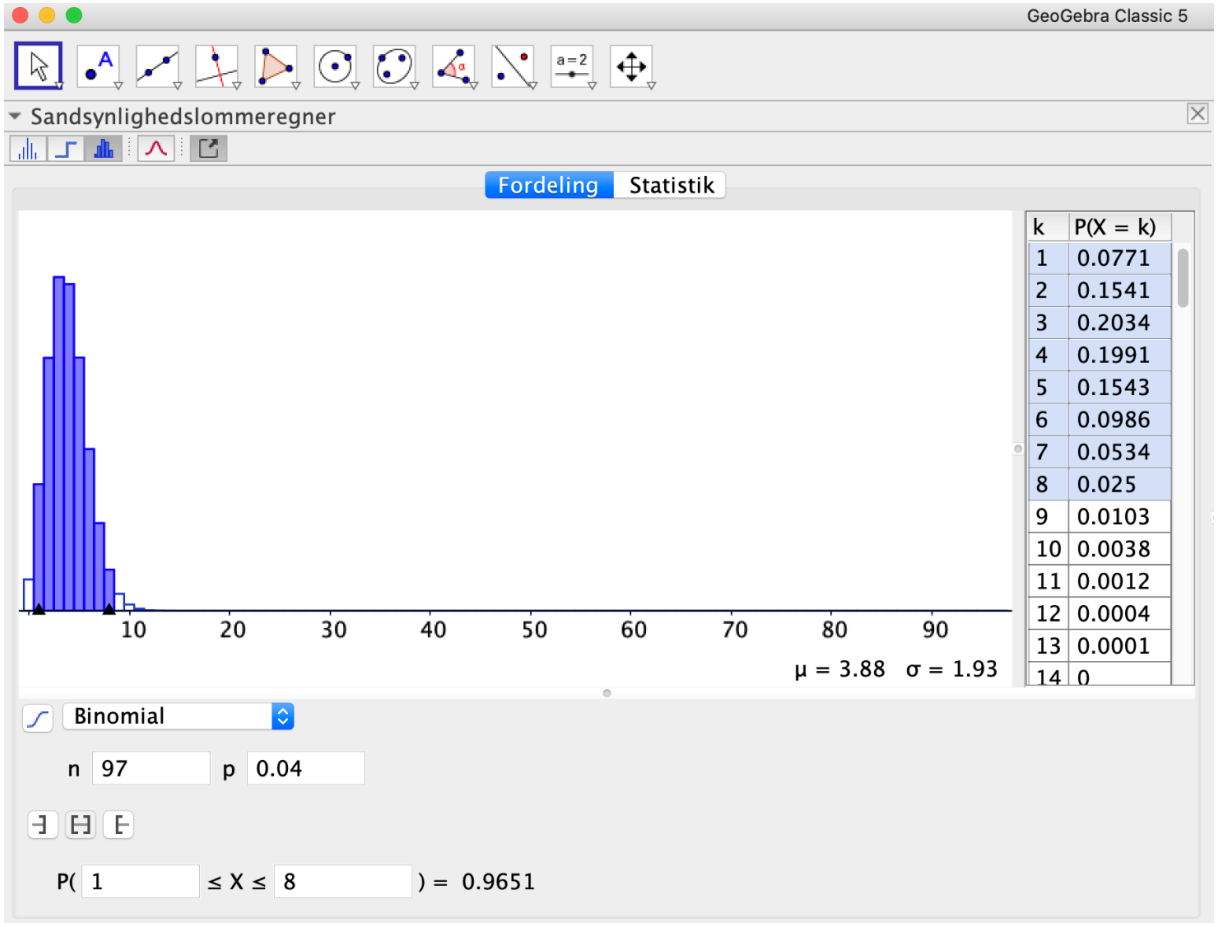
synlighedsfaktoren (*p*) og herudfra fremstilles automatisk et binomialfordelingsplot. Hvis man skulle løse samme problematik i GeoGebra som ovenfor er vist løst med Excel, skal man indtaste antalsparameteren (*n*) til 97 og sandsynlighedsfaktoren (*p*) til 0.04 (GeoGebra bruger '.' til at adskille decimaltal). Der tegnes straks et binomialfordelingsplot for alle mulige udfald (fra 0-97). Under de indtastede værdier for *n* og *p* er der tre regneikoner symboliseret ved -], [-] og [-. De tre ikoner symboliserer henholdsvis en ensidet venstrestillet test (*mindre end*), en intervalberegner og en højrestillet ensidig test (*større end*). Under disse tre regneikoner er der en sandsynlighedsformel med to blanke felter. Formlen ændrer udseende alt efter, hvilket regneikon der er valgt: P(X ≤____)=_____ for den venstrestillede test, P(X ≤____≤_____)=0 for intervalberegneren og P(____X ≤)=_____ for den højrestillede test.



Figur 8. Opstartsbilledet i GeoGebra, når man vælger sandsynlighedsberegneren for binomialfordelingen (se rød ring).

Til at skille acceptområdet fra det kritiske område giver det sig selv at man skal anvende den venstrestillede test når der spørges til sandsynligheden for at noget er *mindre end* forventet, og den højrestillede test når der spørges til sandsynligheden for at noget er *større end* forventet. Når der spørges til en tosidet test, anvendes først den venstrestillede og derefter den højrestil-

lede test, hvor man halverer den angivne *p*-værdi for hver af testene. Hvis man ønsker på p = 0,05-niveuaet at vide hvor acceptområdet i en tosidet test begynder, klikkes på ikonet for venstrestillet test og nede i den underliggende formel skrives den halve værdi af 0.05 efter lighedstegnet, svarende til 0.025, og der trykkes på <Enter>. Det indtastede ændres straks til formlen: P(X ≤ 1)=0.0961. Det resultat betyder, at den laveste værdi i acceptområdet er 1, og sandsynligheden for at få 1 eller færre med seglcelleanæmi er 0,0961, når antalsparameteren er 97 og sandsynlighedsparameteren er 0.04. Hvis 1 er det laveste tal i acceptområdet må det højeste tal i det kritiske område på venstre side af acceptområdet være 0. Ved en højrestillet test klikker man først på det regneikon der gælder for den højrestillede test (-]), og taster derefter 0.025 ind i formlen efter lighedstegnet og trykker på <Enter>. Formlen ændres til P(8 ≤ X) = 0.0409 hvilket skal tolkes som at det øverste tal i acceptområdet er 8, og sandsynligheden for at finde 8 eller flere med seglcelleanæmi er 0,0409. Hvis 8 er det højeste tal i acceptområdet er 9 det laveste tal i det kritiske område på højre side af acceptområdet. Hermed er acceptområdet i den tosidede test heltal fra og med 1 til og med 8. Klikker man herefter på regneikonet for intervalberegneren og skriver 1 ind i feltet til venstre og 8 ind i feltet til højre, så der står P(X ≤ 1 ≤ 8) fås resultatet 0,9651. Det betyder at hele acceptområdet ikke fylder 0,95, (1-0,025-0,025) som man ellers anvender som udgangspunkt, men 0,9651 i dette tilfælde. Acceptområdet vil altid være større end 1 – den anvendte *p*-værdi, fordi acceptområdet alene består af heltal. Se afbildning og beregning i figur 9.

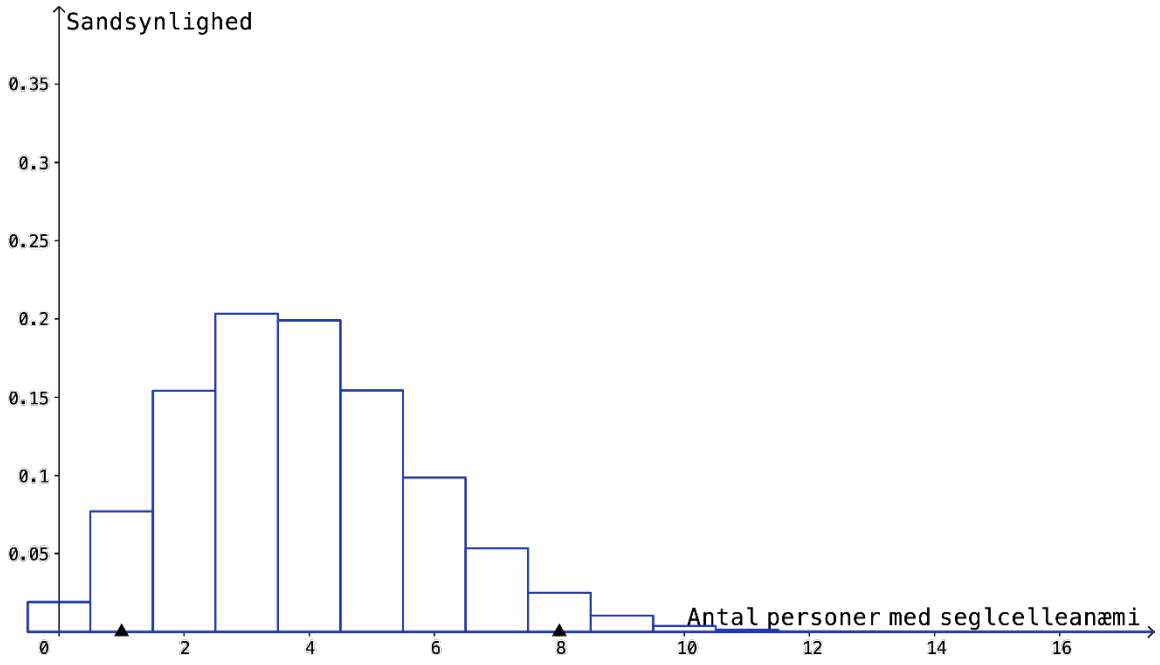


Figur 9. Skærmbillede af binomialfordelingen med en antalsparameter på 97 og en sandsynlighedsparameter på 0.04. Acceptområdet fra 1 til og med 8 er markeret og dækker 0.9651.

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

Afbildning af graf i GeoGebra

Højreklik på grafen vist i figur 9 og vælg 'kopier til tegneblok'. Klik derefter på 'vis' i menulinjen og derefter på 'tegneblok'. Højreklik på tegneblok og vælg det nederste menupunkt hvor der er et tandhjul og der står tegneblok ... Øverst i dialogboksen kan man klikke på xAkse. Ud for Navn skrives: 'Antal personer med seglcelleanæmi'. Sæt flueben i 'Vis kun positiv retning'. Klik på yAkse og ud for Navn skrives: 'sandsynlighed'. Sæt flueben i 'Vis kun positiv retning'. Under 'Basis' kan man manipulere forholdet mellem x-akse og y-akse ved i feltet 'xAkse : yAkse' at skrive fx 25 i xAkse-feltet, hvorved enheden på x-aksen bliver 25 gange længere end enheden på y-aksen. Træk i grafen og skift størrelse på tegneblokken ved at trække i vinduets sider – det tager lidt tid. Når man har et tilfredsstillende resultat, vælger man i menulinjen; Fil → Eksporter → Tegning som billede. Tryk gem og gem billedet på computeren. Sæt det derefter ind i besvarelsen. Et eksempel på et plot kan ses i figur 10. Sammenligner man figur 10 med figur 7, ses det at Excel i forhold til afbildning har væsentlige fordele i forhold til GeoGebra. Men i forhold til hurtig beregning, er GeoGebra at foretrække.



Figur 10. Afbildning af binomialfordelingen i GeoGebra.