Matematikopgaver i systembiologi: søvndata

Af Jan Brønnum Sørensen, Aalborg City Gymnasium

Opgaverne er tænkt i samspil med et forløb i biologi eller bioteknologi, hvor data fra et forsøg om bestemmelse af gener på gymnasielever og fra et tilhørende spørgeskema på flere skoler indgår. Opgaverne forudsætter kendskab til deskriptiv statistik, binomialtest og konfidensintervaller for andele.
I forbindelse med nogle af opgaverne anvendes desuden konfidensinterval for sammenligning af andele i to population og chi^2-test for fordeling og for uafhængighed, hvilket er udover normalt pensum i matematik, men af hensyn til den praktiske anvendelse både her og andre steder i biologien giver god mening at have med alligevel. Der er links til teori om disse emner.
Opgaverne handler således om statistik og databehandling af et større, autentisk datasæt, som eleverne optimalt set selv har leveret egne data til fra forsøget i biologi.
Der er løsningsforslag til opgaverne sidst i materialet.

Opgave 1: Deskriptiv statistik

Først ses på hver enkelt af de 3 måder til at inddele i typer på – egen vurdering i spørgsmål 19, pointsummen i spørgeskemaet og genbestemmelse i forsøget.
1.1: Lav et pindediagram over svarene på spørgsmål 19 (egen vurdering af A- eller B-menneske)
1.2: Lav et pindediagram over pointsummen for hver elev.
1.3: Grupper data for pointsum i forhold til skemaet for inddeling i "type" efter point, og tegn et histogram.
1.4: Lav et pindediagram for fordelingen efter gener.
1.5: Overvej og diskuter, hvad disse figurer viser, og find evt. selv på flere figurer eller beregninger til at illustrere hver af disse 3 metoder til inddeling.

Se nu på de 3 metoder parvis ved at lave en antalstabel / krydstabel.
1.6: Lav en antalstabel for "egen vurdering" mod "gentype". Dvs., at du skal udfylde en tabel som denne med antal personer i hver kombination.

	Absolut aften	Mest aften	Mest morgen	Absolut morgen
4+4 gener				
4+5 gener				
5+5 gener				

1.6: Overvej og diskuter, hvad tallene i tabellen siger om sammenhængen eller mangel på samme mellem egen vurdering og egne gener.
1.7: Lav samme type tabel for "egen vurdering" og "pointsum".
1.8: Lav samme type tabel for "gentype" og "pointsum".

novo nordisk fonden

Opgave 2: Binomialtest og konfidensinterval for andel

En norsk undersøgelse typer på, at fordelingen af gener er sådan, at ca. 44% har 4+4, ca. 44% har 4+5 og blot ca. 11% har 5+5 Det vil vi naturligvis undersøge yderligere, end pindediagrammet i opgave 1.4, men start med at kigge på dette pindediagram.
Vi vil se vores data som en stikprøve fra populationen bestående af alle danske gymnasieelever.
2.1: Vurder og diskuterer, om stikprøven kan anses for repræsentativ for populationen i forhold til søvntype.
2.2: Test hypotesen, at 1/9 af hele populationen har 5+5 gener på 5% signifikansniveau. Herunder skal du finde teststørrelsen, acceptområdet, det kritiske område og p-værdien.
2.3: Lav et 95% konfidensinterval for andelen af personer i populationen, som har 5+5 gener.
2.4: Hvor bredt ville et 95% konfidensinterval for andelen blive, hvis der var 1500 personer i stikprøven? Hvad hvis der var 5000 elever?
2.5: Hvor stor skal stikprøven være, hvis et 95% konfidensinterval for andelen skal være 1% bredt?
2.6: Lav et 99% konfidensinterval for andelen af personer i populationen, som har 5+5 gener.
2.7: Hvor bredt ville et 99% konfidensinterval for andelen blive, hvis der var 1500 personer i stikprøven? Hvad hvis der var 5000 elever?
2.8: Hvor stor skal stikprøven være, hvis et 99% konfidensinterval for andelen skal være 1% bredt?

Opgave 3: Sammenligning af to populationer vha. konfidensintervaller

Lad os nu undersøge, om der er forskel på de to køn mht. gentyper.
3.1: Hvordan fordeler hver af de to køn sig på de tre forskellige gentyper?
3.2 Lav et 95% konfidensinterval for andelen af piger med gentype 5+5 og tilsvarende et 95% konfidensinterval for andelen af drenge med gentype 5+5.
3.3 Overvej og diskuter, om man vha. disse to konfidensintervaller kan afgøre, om der er forskel på andelen af piger og andelen af drenge med gentype 5+5.

Vi får brug for lidt teori om konfidensintervaller, der ligger lidt ud over almindeligt pensum, så vi kan sammenligne andele i to populationer. Så hvis du ikke allerede har hørt om, hvordan man gør det, kan du f.eks. læse om det her eller se denne video på Restudy (Restudy kræver abonnment).
3.4 Lav et 95% konfidensinterval for forskellen mellem de to andele.

novo nordisk fonden

Opgave 4: Undersøgelse af fordeling af gentyper (4+4, 4+5 og 5+5)

I opgave 2 så vi på, om andelen er gentype 5+5 kan være 1/9, og dermed om 4+4 og 4+5 tilsammen kan være 8/9. Men, hvad hvis vi ønsker at se på opdelingen i de 3 typer samtidigt? Da binomial jo, som navnet antyder, kun handler om "2 muligheder", kan binomialtest og konfidensintervaller ikke benyttes, når der er 3 eller flere muligheder. Vi skal derfor yderlige have fat i noget teori, som ligger lidt ud over almindeligt pensum, nemlig chi^2-test for fordeling.
Hvis du ikke har haft om det før, kan du f.eks. læse om det her Mathhx eller se denne video på Restudy.

4.1: Lav et chi^2-Goodness of Fit test for hypotesen, at gentyperne fordeles sig med 4/9, 4/9 og 1/9 på 5% signifikansniveau.

Opgave 5: Undersøgelse af forskel på køn mht. gentyper

I opgave 3 så vi på, om andelen er gentype 5+5 kan være den samme for de to køn.
Hvis vi i stedet for blot at se på de to muligheder (5+5 eller andet) vil se på alle 3 muligheder på én gang, skal vi igen have lidt ny teori. Det er også et chi^2-test, men denne gang for uafhængighed. Her, om der er uafhængighed mellem køn og gentype.
Hvis du ikke har haft om det før, kan du f.eks. læse om der her Mathhx eller se denne video på Restudy.

5.1: Lav et chi^{2-} uafhængighedstest for hypotesen, at køn og gentype er uafhængige på 5% signifikansniveau.

novo nordisk fonden

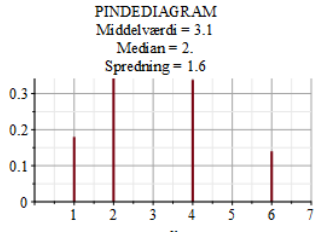
Løsning til opgave 1.

Disse løsninger er lavet i Maple på et testdatasæt, da det rigtige datasæt ikke var tilgængeligt endnu. Så graferne svarer ikke til dine resultater, som du derfor ud over at kigge her for metoden også bør vise til din lærer.

1.1

19. Du har måske hørt om A-mennesker (”morgentyper”) og B-mennesker (”aftentyper”)
[6] Absolut en morgentype
[4] Snarere en morgentype end en aftentype
[2] Snarere en aftentype end en morgentype
[1] Absolut en aftentype.

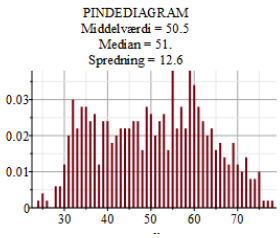
plotPindediagram(spg19)



1.2

Score	16-30	31-41	42-58	59-69	70-86
Type	Absolut ”aftentype”	Moderat ”aftentype”	Mellemtype	Moderat ”morgentype”	Absolut ”morgentype”

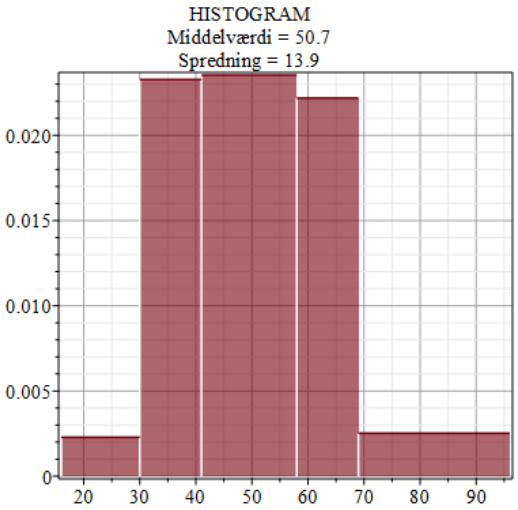
plotPindediagram(spgsum)



1.3

$$grppoint := grupperData(spgsum, \langle 16, 30, 41, 58, 69, 96 \rangle) = \begin{bmatrix} 16..30 & 16 \\ 30..41 & 128 \\ 41..58 & 200 \\ 58..69 & 122 \\ 69..96 & 34 \end{bmatrix}$$

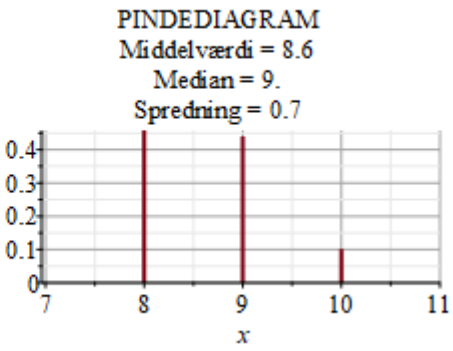
plotHistogram(grppoint)



novo nordisk fonden

1.4

$plotPindediagram(spggener)$



1.6

Direkte svar mod gener.

$M1 := simsøvndata[\ ..,\ [20,\ 23]]:$

$A1 := antalstabel(M1) =$

"Observeret"	8.0	9.0	10.0	"i alt"
1.0	74	16	0	90
2.0	99	65	7	171
4.0	46	101	22	169
6.0	10	38	22	70
"i alt"	229	220	51	500

1.7

Direkte svar mod pointscore.

$M2 := simsøvndata[\ ..,\ [20,\ 22]]:$

$A2 := antalstabel(M2) =$

"Observeret"	0.	1.0	2.0	3.0	4.0	"i alt"
1.0	11	51	25	3	0	90
2.0	5	61	71	31	3	171
4.0	0	14	84	57	14	169
6.0	0	2	20	31	17	70
"i alt"	16	128	200	122	34	500

1.8

Pointscore mod gener.

$M3 := simsøvndata[\ ..,\ [22,\ 23]]:$

$A3 := antalstabel(M3) =$

"Observeret"	8.0	9.0	10.0	"i alt"
0.	16	0	0	16
1.0	127	1	0	128
2.0	86	113	1	200
3.0	0	99	23	122
4.0	0	7	27	34
"i alt"	229	220	51	500

novo nordisk fonden

Løsning til opgave 2.

2.1 Med elever af begge kønne fra flere skoler spredt ud over landet, er det ikke umiddelbart noget, som indikerer, at stikprøven ikke skulle være repræsentativ.

2.2

I datasætter har 51 ud af 500 personer 5+5 gener, så teststørrelse af 51, svarende til ca. 10%.

$$invbin\left(500, \frac{1}{9}, 0.025\right) = 42.$$

$$invbin\left(500, \frac{1}{9}, 0.975\right) = 70.$$

Så acceptområdet er [42, 70] og det kritiske område er [0, 41] og [71, 500].
Da stikprøvestørrelsen ligger i acceptområde, forkastes hypotesen ikke.
Det kan derfor ikke afvises, at 1/9 af alle i populationen har 5+5 gener.

$$pværdi := 2 \cdot bincdf\left(500, \frac{1}{9}, 51\right) = 0.5723001610$$

Så pværdien er ca. 57%, hvilket ligeledes viser, at hypotesen ikke kan forkastes.

2.3

$$konfidensInterval\left(\frac{51}{500}, 500, 0.95\right) = [0.075472, 0.128528]$$

2.4 Her kan man diskutere, hvad man skal sætte den forventede andel til, da den jo ikke kendes. Jeg har blot sat andelen til samme andel, som vi har set i vores stikprøve.

$$konfidensInterval\left(\frac{51}{500}, 1500, 0.95\right) = [0.086684, 0.117316]$$

$$konfidensInterval\left(\frac{51}{500}, 5000, 0.95\right) = [0.093611, 0.110389]$$

2.5 Man kan diskutere, om man skal anvende faktoren 1,96 (svarer til 2,5% fraktilen i normalfordelingen) eller faktoren 2 (formelsamlings værdi). Her er 1,96 anvendt. Hvis man bruger 2, bliver facit lidt anderledes.

$$solve\left(1.96 \cdot \sqrt{\frac{phat \cdot (1 - phat)}{n}} = 0.005\right) = 14075.00774$$

$$konfidensInterval(phat, 14075, 0.95) = [0.097000, 0.107000]$$

2.6-2.8

$$konfidensInterval\left(\frac{51}{500}, 500, 0.99\right) = [0.067137, 0.136863]$$

$$konfidensInterval\left(\frac{51}{500}, 1500, 0.99\right) = [0.081872, 0.122128]$$

$$konfidensInterval\left(\frac{51}{500}, 5000, 0.99\right) = [0.090975, 0.113025]$$

$$phat := \frac{51}{500} : z := invnorm(0.995) = 2.57582930355009$$

$$solve\left(z \cdot \sqrt{\frac{phat \cdot (1 - phat)}{n}} = 0.005\right) = 24309.19956$$

$$konfidensInterval(phat, 24309, 0.99) = [0.097000, 0.107000]$$

novo nordisk fonden

Løsning til opgave 3.

3.1

Køn og genfordeling

$$M := antalstabel(simsøvndata[\ ..,\ [1,\ 23]\ ]) = \begin{bmatrix} \text{"Observeret"} & 8.0 & 9.0 & 10.0 & \text{"i alt"} \\ \text{"K"} & 134 & 121 & 25 & 280 \\ \text{"M"} & 95 & 99 & 26 & 220 \\ \text{"i alt"} & 229 & 220 & 51 & 500 \end{bmatrix}$$

3.2

$$ppiger := \frac{25.}{280} : npiger := 280 : pdrenge := \frac{26.}{220} : ndrenge := 220 :$$

$$konfidensInterval(ppiger,\ npiger,\ 0.95) = [0.055885,\ 0.122686]$$

$$konfidensInterval(pdrenge,\ ndrenge,\ 0.95) = [0.075524,\ 0.160840]$$

3.3 Man kan helt generelt ikke blot se, om de to individuelle konfidensintervaller overlapper, da 5% signifikans i den ene og 5% signifikans i den anden IKKE tilsammen giver 5% signifikans. Man skal derfor benyttes et fælles konfidensinterval.

3.4 Om man benytter 1,96 eller 2 som faktor er en smagssag.

$$pforskel := pdrenge - ppiger = 0.02889610391$$

$$CI95 := \left[pforskel - 1.96 \cdot \sqrt{\frac{pdrenge \cdot (1 - pdrenge)}{ndrenge} + \frac{ppiger \cdot (1 - ppiger)}{npiger}} \right.$$

$$\left. ,\ pforskel + 1.96 \cdot \sqrt{\frac{pdrenge \cdot (1 - pdrenge)}{ndrenge} + \frac{ppiger \cdot (1 - ppiger)}{npiger}} \right] = [-0.02528328531,\ 0.08307549313]$$

Dvs, at der er 95% sikkerhed for, at forskellen i andele ligger mellem -2,5% og 8,3%.
Da 0 ligger i dette interval, kan vi ikke forkaste en hypotese om, at andelene er ens.

Der er således ikke statistisk signifikans for, at kønnenes er forskellige mht. andel med gentype 5+5.

novo nordisk fonden

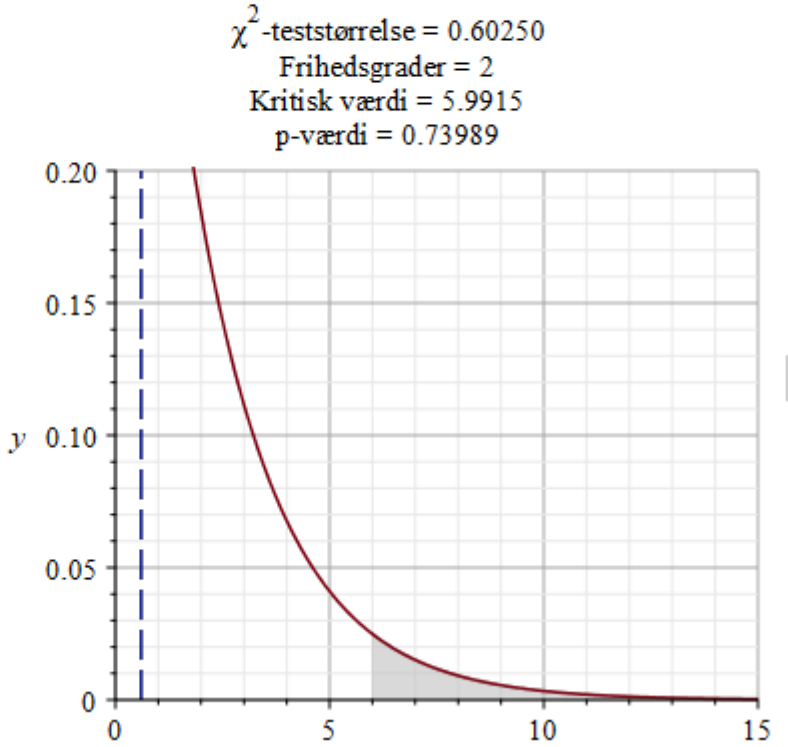
Løsning til opgave 4.

4.1

$$obs := [229, 220, 51]:$$

$$forv := \left[\frac{4}{9}, \frac{4}{9}, \frac{1}{9}\right] \cdot 500. = [222.2222222, 222.2222222, 55.55555556]$$

$$chi2GOFtest(obs, forv)$$



Da p-værdien på 74% er over signifikansniveauet på 5%, forkastes hypotesen ikke.
Det kan derfor ikke afvises, at fordelingen er 4/9, 4/9, 1/9. Husk dog, at denne facitliste er lavet ud fra fiktive data, så det er interessant, om resultaterne er de samme på faktiske data.

novo nordisk fonden

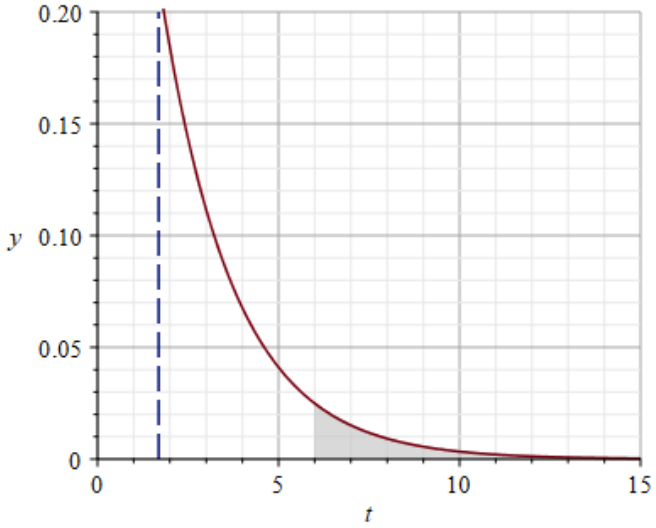
Løsning til opgave 5.

5.1

$$M = \begin{bmatrix} \text{"Observeret"} & 8.0 & 9.0 & 10.0 & \text{"i alt"} \\ \text{"K"} & 134 & 121 & 25 & 280 \\ \text{"M"} & 95 & 99 & 26 & 220 \\ \text{"i alt"} & 229 & 220 & 51 & 500 \end{bmatrix}$$

$chi2Utest(M)$

χ^2-teststørrelse = 1.6858
Frihedsgrader = 2
Kritisk værdi = 5.9915
p-værdi = 0.43046



Da p-værdien på 43% er over signifikansniveauet på 5%, forkastes hypotesen ikke.
Det kan derfor ikke afvises, at der er uafhængighed mellem køn og gentyper. Med andre ord, at piger og drenge har samme fordeling blandt de 3 gentyper.

novo nordisk fonden