Stx

Deskriptiv statistik i Biologi A

2. Udgave

Jette H Vestergaard

Dronninglund Gymnasium

Januar 2022

Indholdsfortegnelse

2. udgave

Illustrationer
Forside:
https://pixabay.com/da/photos/vanilje-blomst-vanilje-hvid-gul-542019/

1. Boksplot

1.1 Hvad er et boksplot

Et boksplot er et diagram, som illustrerer de 5 værdier, der indgår i *det udvidede kvartilsæt* for en række observationer. De 5 værdier er: minimum, nedre kvartil, median, øvre kvartil og maksimum.
De tre midterste værdier (nedre kvartil , median , øvre kvartil) kaldes kvartilsættet og opdeler observationsrækken ved 25 %, 50 % og 75 %.

Eksempel
I en klasse måles højden af alle drengene.
Resultatet er følgende: 191 , 180 , 188 , 198 , 185 , 179 , 188 , 172 , 165 , 190.
Observationerne stilles op i rækkefølge: 165 , 172 , 179 , 180 , 185 , 188 , 188 , 190 , 191 , 198.

Median, Med
Medianen er den midterste observation, hvis der er et ulige antal observationer, og gennemsnittet af de to midterste observationer, hvis der er et lige antal observationer.
Her er der 10 observationer. Medianen er derfor gennemsnittet af de to midterste observationer (185 og 188), dvs medianen er 186,5.

Nedre kvartil, Q1
Den nedre kvartil findes på samme måde som medianen, men blandt observationerne til venstre for medianen (165 , 172 , 179 , 180, 185). Her er der 5 observationer. Nedre kvartil, Q1, er derfor lig med 179.

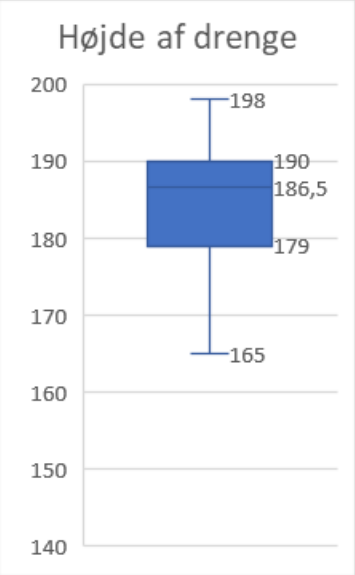
Øvre kvartil, Q3
Den øvre kvartil findes på samme måde som medianen, men blandt observationerne til højre for medianen (188 , 188 , 190 , 191 , 198). Her er der 5 observationer. Øvre kvartil, Q3, er derfor lig med 190.

Det udvidede kvartilsæt
Det udvidede kvartilsæt er (165 , 179 , 186,5 , 190 , 198).

Boksplot
Når man skal tegne et boksplot, tegnes en lodret akse, som mindst dækker værdierne i observationsrækken. Derefter markeres de 5 værdier i det udvidede kvartilsæt med en vandret streg. Der tegnes en lodret streg gennem de 5 vandrette streger fra minimum til maksimum. Til sidst tegnes en "kasse" mellem Q1 og Q3. Bredden af kassen er underordnet.
Figur 1 viser boksplottet for drengene i denne klasse.

I Excel kaldes et boksplot for en **kasse med hale**.

I nogle programmer ligger aksen og boksplottet vandret, men betydningen er den samme.



Figur 1

1.2 Hvad kan et boksplot bruges til

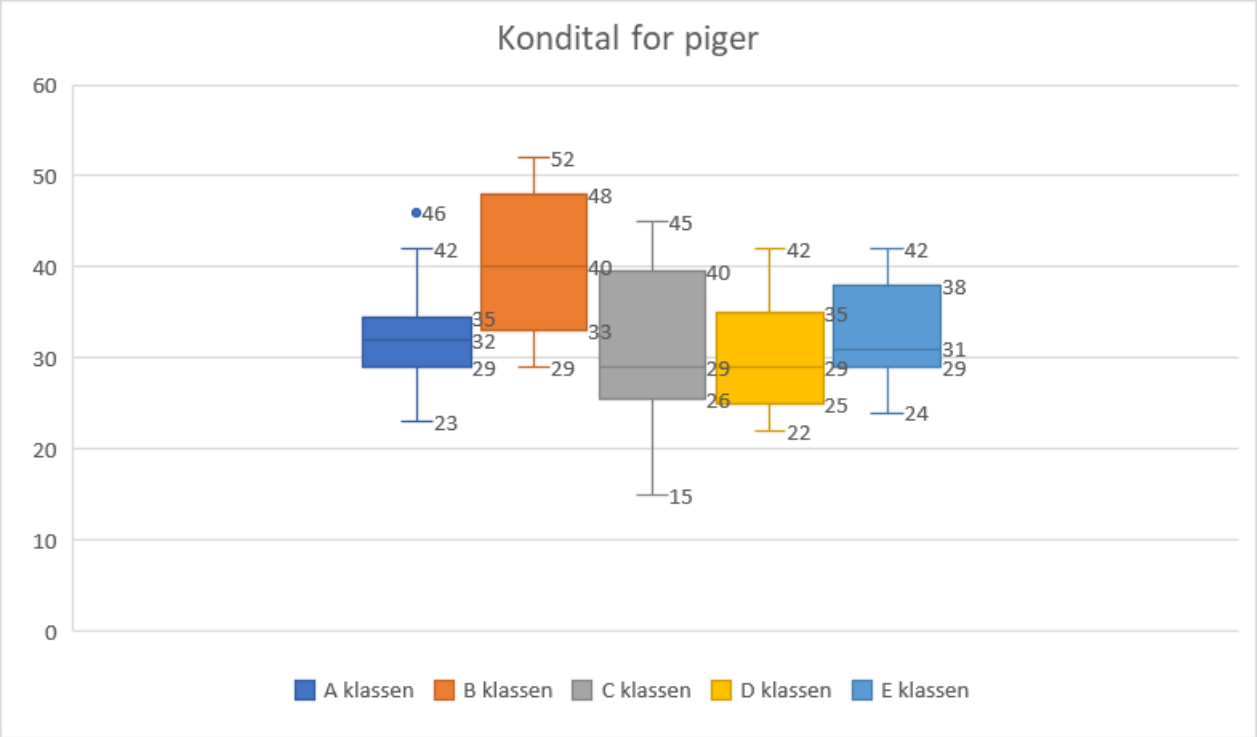
Et boksplot illustrerer ikke andet end de 5 værdier i det udvidede kvartilsæt, så hvis man kun har én enkelt række observationer, kan man lige så godt betragte det udvidede kvartilsæt.
Hvis man til gengæld har mange observationsrækker, kan et diagram med flere boksplot ved siden af hinanden give et bedre overblik over data, end et skema med det udvidede kvartilsæt for hver af observationsrækkerne ville give.

Eksempel
På et gymnasium testes de nye 1.g elever, og deres kondital noteres. Da drenges og pigers kondital vurderes forskelligt, opdeles data efter køn. Tabel 2 viser det udvidede kvartilsæt for pigerne i 5 klasser.

	A klassen	B klassen	C klassen	D klassen	E klassen
Minimum	23	29	15	22	24
Nedre kvartil, Q1	29	33	25,5	25	29
Median, Med	32	40	29	29	31
Øvre kvartil, Q3	34,5	48	39,5	35	38
Maksimum	46	52	45	42	42

Tabel 2

Det ses, at det er svært at danne sig et overblik over data, når de præsenteres på denne måde.
Figur 3 viser boksplots for pigerne i de 5 klasser.



Figur 3

Det ses tydeligt, at det giver et meget bedre overblik over data, når de præsenteres på denne måde.
Excel er for overskuelighedens skyld indstillet til 0 decimaler, så kvartilerne er rundet af til hele tal.

1.3 Sammenligning af flere observationsrækker vha boksplot

Der findes ingen regler for, hvad man præcist skal komme ind på i en sammenligning af flere boksplots. Men det kan være en god ide at komme ind på følgende 3 ting:

- Et par sammenligninger af interessante kvartiler
- Sammenligning af variationen i klasserne
- Opsamlende konklusion

Eksempel
I eksemplet fra afsnit 1.2 med pigerne i de 5 klasser kunne en sammenligning være:

Interessante kvartiler
B klassens Q3 ligger over maksimum i alle de andre klasser. Altså har de 25 % bedste i B klassen et højere kondital end alle pigerne i de andre klasser.
Samtidig ses det, at B klassens minimum ligger over Q1 i alle de andre klasser. Altså er den svageste elev i B klassen bedre end de 25 % dårligste i hver af de andre klasser.

Variation i klasserne
Det ses, at variationen (forskellen mellem maksimum og minimum) i C klassen er størst, og at variationen i E klassen er mindst. Dvs at eleverne i C klassen er mest forskellige mht kondital og eleverne i E klassen er mest ens mht kondital.
Samtidig ses det, at "kassen" i A klassen er mindst og at "kassen" i B klassen er størst. Dvs at de 50 % midterste elever er mest ens i A klassen, mens de er mest forskellige i B klassen.

Konklusion
Det ses, at B klassen generelt har et bedre kondital end de andre klasser. De har en "top" der ligger over de andre klasser, mens deres svageste elev er bedre end "bunden" i de andre klasser.
Des ses desuden, at eleverne i E klassen er mest ens mht kondital, mens eleverne i C klassen er mest forskellige mht kondital.

1.4 Hvad kan et boksplot ikke bruges til

Et boksplot hører til i den del af statistikken, der hedder "deskriptiv statistik". Det betyder, at det er beskrivende statistik. Altså "hvordan ser data ud lige netop i disse undergrupper".
Et boksplot er *ikke* et matematisk test. Man kan derfor *ikke* konkludere noget signifikant om en større population ud fra nogle boksplots for nogle undergrupper.
Et boksplot er altså rent beskrivende.

1.5 Tastevejledning til Excel (PC)

Data indtastes i et regneark. *Tabel 4.*
Husk at navngive observationsrækkerne i række 1. (de røde tal er ikke observationer)
Hvis der er et lige antal observationer i observationsrækken, beregner Excel desværre ikke Q1 og Q3 på den rigtige måde. Hvis man vil have beregnet dem på den rigtige måde, kan man på snedig vis godt få Excel til det. Man skal blot tilføje en ekstra observation, medianen (de røde tal), i rækkerne med et lige antal observationer.
I celle A18 skriver man **=kvartil(A2:A17;2)** og taster enter.
A2 til A17 henviser til de 16 observationer. 2-tallet til sidst henviser til den 2. kvartil, altså medianen.
På tilsvarende vis skrives **=kvartil(B2:B15;2)** i celle B16 og **=kvartil(E2:E19;2)** i celle E20.
Markér derefter alle udfyldte celler (også række 1 med navnene).
Vælg **Indsæt – Anbefalede diagrammer**.
Vælg **Alle diagrammer – Kasse med hale**.
Tast **OK**.
Så tegnes de 5 boksplot.

	A	B	C	D	E
1	A klassen	B klassen	C klassen	D klassen	E klassen
2	23	29	45	42	27
3	25	30	45	32	41
4	27	30	44	22	29
5	29	33	44	22	32
6	29	33	35	27	30
7	31	37	34	31	39
8	31	39	33	27	42
9	32	41	33	29	24
10	32	41	29	40	34
11	32	46	28	27	25
12	33	48	28	30	30
13	34	48	27	38	29
14	35	50	27	29	31
15	42	52	24	23	38
16	42	40	24	23	31
17	46		21	29	24
18	32		15	38	38
19					33
20					31

Tabel 4

Man kan nu finpudse sit diagram med aksetitler, dataetiketter, osv. på følgende måde.
Aktivér diagrammet ved at klikke i det så der kommer små cirkler hele vejen rundt.
Så kommer der en firkant med et kryds i til højre for diagrammet. Klik på krydset.
Så får man valgmulighederne vist i *Figur 5.*
Aksetitler giver mulighed for at skrive noget på akserne.

Figur 5

Figur 6

Dataetiketter viser værdierne for det udvidede kvartilsæt. Når man peger på *Dataetiketter* kommer der en pil til højre som i *Figur 6.* Herefter kan man vælge, hvor værdierne skal stå og også antallet af decimaler (ved at vælge *Flere indstillinger – Tal*).
Forklaring viser navnene på de 5 observationsrækker. Pilen til højre i dén linje giver mulighed for at vælge placeringen af navnene.

Hvis man vil have fjernet 1-tallet i bunden klikker man på det og vælger Klip.

Som udgangspunkt er der et kryds i midten af boksplottene. Det viser middelværdien i den pågældende observationsrække. Hvis man gerne vil have fjernet dette kryds, aktiverer man det enkelte boksplot ved at klikke i det, og fjerner derefter "fluebenet" ved **vis mærker for middelværdi** i rullegardinet ude til højre.

Diagrammet skulle nu helst se ud som diagrammet i afsnit 1.2.

2. Middelværdi og spredning

2.1 Hvordan beregnes middelværdi og spredning

Betragt en observationsrække med n observationer

$$x_1, x_2, x_3, \ldots, x_n$$

Middelværdien for observationsrækken er gennemsnittet af observationerne og beregnes ved formlen

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Spredningen for observationsrækken er et mål for, hvor meget de observerede værdier afviger fra middelværdien, og beregnes ved formlen

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

I formlen beregner man forskellen mellem de observerede værdier og middelværdien. Hvis disse forskelle er store bliver spredningen stor. Hvis disse forskelle er små bliver spredningen lille. Spredningen fortæller derfor noget om variationen i observationsrækken.

Det ses, at formlen for spredning ret regnetung, så i praksis vil vi ikke anvende denne formel manuelt men nøjes med at lade Excel beregne spredningen for os.

I nogle C-niveau matematikbøger divideres der med n i stedet for $n - 1$, så eleverne kan have stødt på en anden formel. Men i alle indbyggede statistikpakker i regneark eller CAS-programmer anvendes formlen med $n - 1$. Forskellen er imidlertid heller ikke ret stor, men mindre der er tale om en meget lille stikprøve.

2.2 Hvordan fortolkes middelværdi og spredning

Middelværdien er gennemsnittet af en observationsrække og er simpel at fortolke for én observationsrække. Spredningen er et mål for variationen i en observationsrække, men det giver til gengæld ingen mening at fortolke dén for én observationsrække. Den kan kun bruges i forbindelse med en sammenligning af flere observationsrækker.

Eksempel

Vi betragter eksemplet fra afsnit 1 med pigernes kondital.

Middelværdien og spredningen for de 5 observationsrækker beregnes til

	A klassen	B klassen	C klassen	D klassen	E klassen
Middelværdi	33	40	32	30	32
Spredning	6,2	8,0	8,9	6,3	5,6

Tabel 7

Det ses, at B-klassen klart har den højeste middelværdi, altså har de som gennemsnit det højeste kondital.

Det ses desuden, at spredningen er størst i C-klassen mens den er mindst i E-klassen. Det betyder, at variationen er størst i C-klassen mens variationen er mindst i E-klassen. Altså er eleverne i C-klassen mest forskellige mht kondital, mens eleverne i E-klassen er mest ens mht kondital.

Konklusionen her minder meget om konklusionen på de 5 boksplot.

2.3 Hvad kan vi ikke konkludere ud fra middelværdi og spredning

Middelværdi og spredning hører ligesom boksplot til i den del af statistikken, der hedder "deskriptiv statistik". Det betyder, at det er beskrivende statistik. Altså "hvordan ser data ud lige netop i disse undergrupper". Middelværdi og spredning er derfor *ikke* et matematisk test. Man kan altså *ikke* konkludere noget signifikant om en større population ud fra middelværdi og spredning for nogle undergrupper.
Middelværdi og spredning er rent beskrivende.

2.4 Tastevejledning til Excel (PC)

I regnearket med konditallene fra afsnit 1 forskydes data én kolonne til højre ved at højreklikke på A'et for oven og vælge **Indsæt**, så der kommer en "tom" kolonne A.
Derefter skrives **middelværdi** i celle A21 og **spredning** i celle A22.
I celle B21 skrives formlen **=MIDDEL(B2:B19)** og der tastes enter, hvorved middeltallet beregnes.
Derefter kopieres formlen over i cellerne C21 til F21 på følgende måde:

- klik i celle B21
- flyt curseren ned i højre hjørne så der kommer et sort kryds
- hold venstre musetast nede og træk derefter musen over til celle F21

I celle B22 skrives formlen **=STDAFV.S(B2:B19)** og der tastes enter, hvorved spredningen beregnes.
Derefter kopieres formlen over i cellerne C22 til F22 på samme måde som før.
Antallet af decimaler kan ændres ved at klikke på knapperne → 

	A	B	C	D	E	F
1		A klassen	B klassen	C klassen	D klassen	E klassen
2		32	46	29	42	27
3		23	30	24	32	41
4		27	39	27	22	29
5		42	33	44	22	32
6		35	48	45	27	30
7		31	41	34	31	39
8		29	33	33	27	42
9		34	29	44	29	24
10		29	30	15	40	34
11		42	52	27	27	25
12		25	50	33	30	30
13		32	48	24	38	29
14		33	37	28	29	31
15		32	41	35	23	38
16		46		28	23	31
17		31		21	29	24
18				45	38	38
19						33
20						
21	**middelværdi**	**33**	**40**	**32**	**30**	**32**
22	**spredning**	**6,2**	**8,0**	**8,9**	**6,3**	**5,6**

Tabel 8

3. Relevant databehandling

3.1 "Relevant databehandling" – hvad er det

I mere åbne opgaver kan man komme ud for formuleringen:

"Foretag relevant databehandling, der kan vise…"

Der er ikke noget entydigt svar på et sådant spørgsmål. Men selvfølgelig er der noget databehandling, der er mere relevant end andet.
Hvis data består af flere observationsrækker, hvor der hver gang er målt på det samme men i flere undergrupper, vil det ofte være relevant at lave noget sammenlignende deskriptiv statistik.
Hvis data består af to observationsrækker med sammenhørende værdier (altså nogle målepunkter med to variabler) vil det ofte være relevant at plotte data og lave forskellige regressioner for at undersøge data.

Eksempel
I det Vejledende Sæt 1 forekommer denne formulering i opgave 1, som vi vil se på her. Opgaveformuleringen ses s. 11. Data består af 3 observationsrækker, som man gerne vil have sammenlignet. Det vil derfor være oplagt at lave noget deskriptiv statistik.
Til opgaven hører et bilag med data fra de tre stikprøver.

Man skal i spørgsmål 3 vise: "Hvilken af de tre stikprøvestørrelser, der kommer nærmest på den faktiske tæthed af mariehøns på papiret."
Spørgsmålet er en anelse uheldigt formuleret, idet alle tre stikprøvestørrelser er på 50 observationer. Der menes nok, at man skal vise, hvilken af de tre *stikprøver*, der giver det bedste bud på populationstætheden.

I opgaven står der desuden: "Den faktiske tæthed af mariehøns på papiret er 0,064 pr. cm^2, idet det totale antal mariehøns på papiret er 120, og arealet af papiret er 1871 cm^2."

Opgaven går altså ud på at vurdere, hvilken stikprøve, der giver det bedste bud på antal mariehøns pr cm^2.
Det oplagte vil her være at beregne det gennemsnitlige antal mariehøns pr cm^2 for hver stikprøve.
Det gennemsnitlige antal mariehøns i hver stikprøve beregnes ved at finde middelværdien i hver række som beskrevet i afsnit 2.
For hver stikprøve divideres det gennemsnitlige antal mariehøns med arealet af plastikskiven for at find det gennemsnitlige antal mariehøns pr cm^2.

	Diameter 3 cm	Diameter 6 cm	Diameter 9 cm	Populationstæthed (antal/cm^2)
Middelværdi af antal mariehøns	0,92	2,76	5,34	
areal af skive (cm^2)	7,07	28,27	63,62	
gennemsnitligt antal mariehøns pr cm^2	0,130	0,098	0,084	0,064

Tabel 9

Det ses, at den stikprøve, der kommer tættest på den faktiske tæthed af mariehøns på papiret, er stikprøven, hvor diameteren er 9 cm.
Det ses også, at populationstætheden bliver beregnet for stor ved alle tre stikprøver, men at den bliver mindre og mindre og nærmer sig den sande værdi jo større diameteren er. Dette kunne tyde på, at man ville få et endnu mere præcist resultat, hvis man valgte en endnu større plastikskive. Men det er klart, at der er en øvre grænse for skivens areal, idet skiven skal lande indenfor papiret, som i alt er 1871 cm^2.
At tætheden bliver estimeret for stor ved alle tre skiver kunne tyde på en fejl ved designet af forsøget. Man kunne måske forestille sig, at eleverne ikke kaster skiven tilfældigt, men ubevidst forsøger at kaste skiven derhen, hvor der er nogle mariehøns. Man kunne måske også forestille sig, at de tæller nogle mariehøns med, som delvist ligger udenfor skiven.

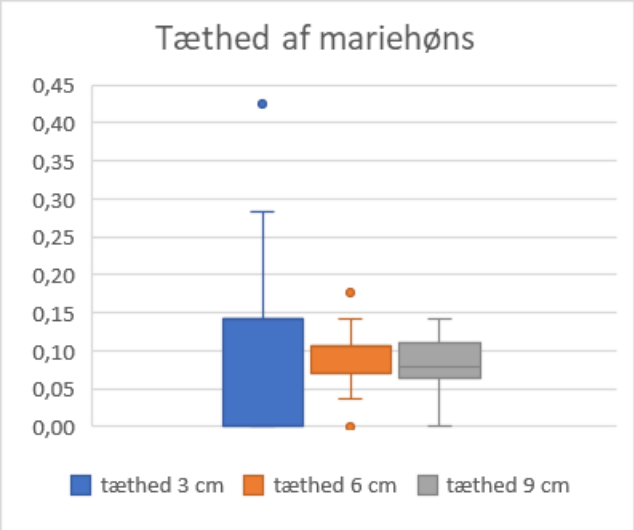
Ud over at beregne antal mariehøns pr cm^2 kunne man også lave boksplots for at undersøge de tre observationsrækker nærmere.
De tre rækker er imidlertid ikke direkte sammenlignelige, da de tre skiver er forskellige, så man skal først beregne den estimerede tæthed ved hver enkelt observation, ved at dividere antallet af mariehøns med arealet af skiven. De estimerede tætheder vil være sammenlignelige, da de beskriver det samme.
Man laver derfor 3 nye observationsrækker ved siden af de oprindelige observationsrækker.

Navnene på de nye observationsrækker skrives i cellerne F1, G1 og H1.
I celle F2 skrives **=B2/7,07** og der tastes enter.
(da de oprindelige observationer står i kolonne B)
Denne formel kopieres ned til celle F51.
I celle G2 skrives **=C2/28,27** og der tastes enter.
Denne formel kopieres ned til celle G51.
I celle H2 skrives **=D2/63,62** og der tastes enter.
Denne formel kopieres ned til celle H51.
De første 8 rækker ses i *Tabel 10.*

F	G	H
tæthed 3 cm	tæthed 6 cm	tæthed 9 cm
0,00	0,07	0,00
0,28	0,14	0,13
0,14	0,04	0,05
0,00	0,11	0,09
0,14	0,11	0,08
0,14	0,11	0,11
0,28	0,11	0,06

Tabel 10

Derefter tilføjes medianerne (da der er et lige antal observationer), ved at skrive **=KVARTIL(F2:F51;2)** i celle F52.
Denne formel kopieres over i celle G52 og H52.
Boksplottene ses i *Figur 11.*



Figur 11

Det ses, at der er størst variation i den estimerede tæthed ved brug af den mindste skive.

Prikkerne yderst i det blå og i det orange boksplot er det man kalder *outliers*. Det er observationer, der ligger "langt væk" fra de andre observationer.
"Langt væk" defineres som mere end halvanden kvartilbredde under Q1 eller over Q3.
Kvartilbredden er afstanden mellem Q1 og Q3.
Kvartilbredden for det blå boksplot er 0,14. Dette tal ganges med 1,5 og giver så 0,21.
Q3+0,21 giver 0,35. Dvs at observationer over 0,35 er outliers i det blå boksplot.

4. Opgaver

Opgave 1

Kilde: Vejledende Sæt 2 opgave 4 uddrag

En gruppe gymnasieelever opstillede den hypotese, at hvilepulsen er lavere hos trænede personer end hos utrænede personer. For at teste hypotesen hvilede eleverne i 5 minutter, hvorefter antal pulsslag på håndleddet blev talt i 30 sekunder og derefter omregnet til slag pr minut. Eleverne skulle selv angive, om de var trænede eller utrænede.

Resultaterne af målingerne findes i vedlagte Excel-dokument.

Puls (slag/min) trænede	62	58	60	74	55	55	58	65	59	48	62	68	66	56	54	64	60
Puls (slag/min) utrænede	60	72	80	67	68	60	71	68	68	72	82	79	66	86	78	71	80

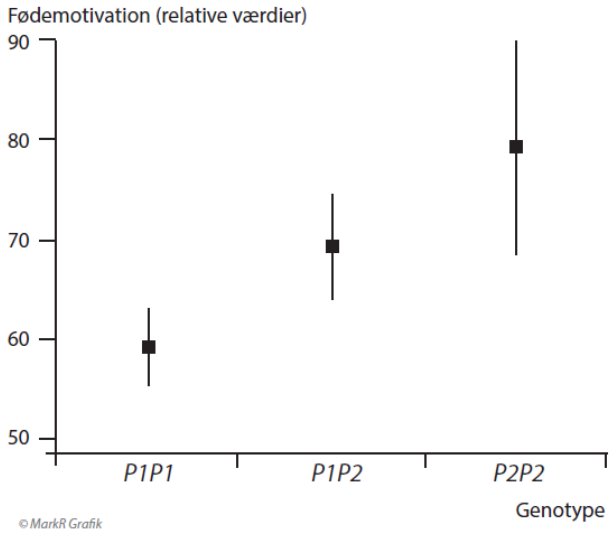
3. Afbild resultaterne af gymnasielevernes målinger som boksplot.

4. Vurder, om resultaterne bekræfter hypotesen om, at hvilepulsen er lavere hos trænede personer end hos utrænede personer.

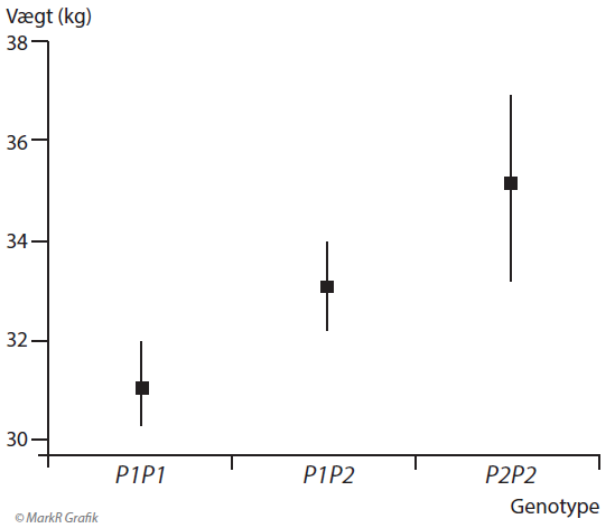
Opgave 2

Kilde: Vejledende Sæt 2 opgave 2 uddrag

Genotypen med hensyn til normalallelen, *P1*, og mutantallelen, *P2*, blev bestemt hos en gruppe labrador retrievere. Sammenhæng mellem genotype og henholdsvis vægt og fødemotivation i forbindelse med træning hos hundene er vist i *figur 3* og *figur 4*.



Figur 3.
Sammenhæng mellem genotype og fødemotivation. For hver gentype er angivet middelværdi og spredning.



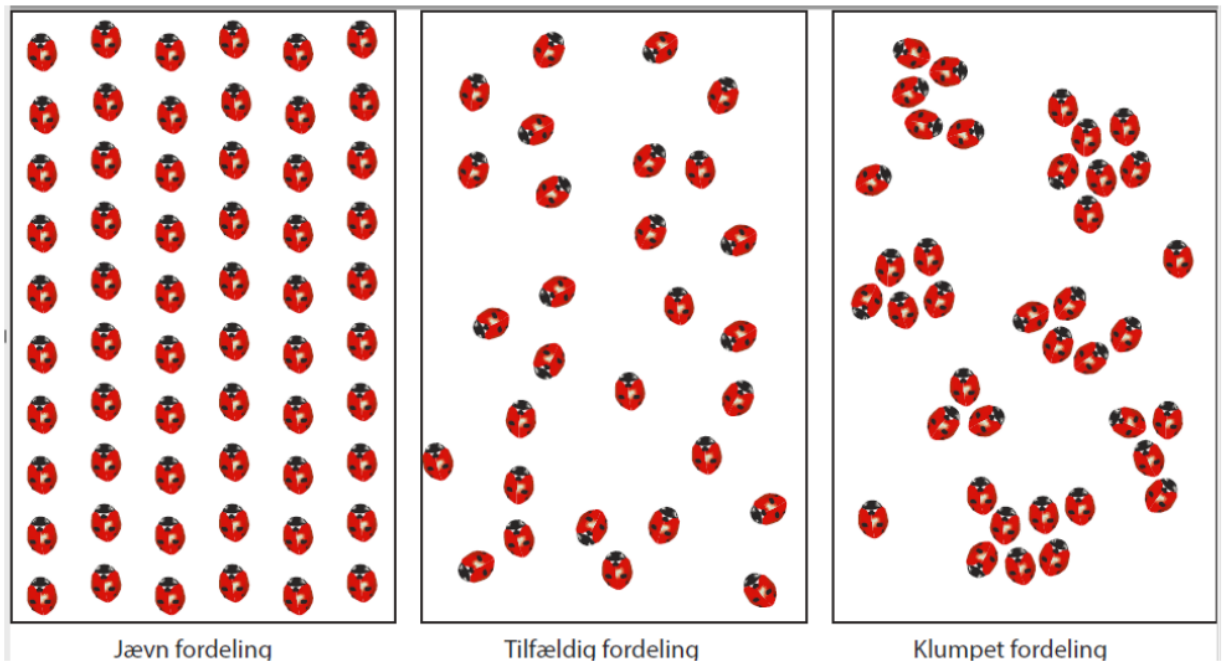
Figur 4.
Sammenhæng mellem genotype og vægt. For hver gentype er angivet middelværdi og spredning.

3. Analysér resultaterne, vist i *figur 3* og *figur 4*.

Opgave 3

Kilde: Vejledende Sæt 1 opgave 1 uddrag

Populationer af planter og dyr kan være fordelt på forskellig måde i en biotop, se *figur 1*.
Individer af en art er oftest klumpet fordelt i naturen.



Figur 1
Tre eksempler på fordeling af individer i en biotop.

En gruppe gymnasieelever arbejder med at simulere bestemmelse af populationstæthed af mariehøns ved hjælp af stikprøver. Eleverne har den hypotese, at nøjagtigheden af bestemmelsen af populationstætheden vil øges i takt med at arealet af stikprøven forøges.
Eleverne udfører tre serier á 50 stikprøver. I første serie anvendes en plastikskive med et areal på 7,07 cm^2, i anden serie en skive med et areal på 28,27 cm^2 og i tredje serie en skive med et areal på 63,62 cm^2.
Den faktiske tæthed af mariehøns på papiret er 0,064 pr. cm^2, idet det totale antal mariehøns på papiret er 120, og arealet af papiret er 1871 cm^2.
De ubehandlede data fra de tre serier findes i vedlagte Excel-dokument.

3. Foretag relevant databehandling, der kan vise, hvilken af de tre stikprøvestørrelser, der kommer nærmest på den faktiske tæthed af mariehøns på papiret. Begrund dit valg af databehandling.

4. Skriv en konklusion på gymnasieelevernes populationsbestemmelse ved stikprøve på grundlag af fremgangsmåden, vist i den vedlagte film og din databehandling.