MAD 2022-23, Assignment 3

Helga Rykov Ibsen <mcv462> Hold 6

1. september 2023

Opgave 1

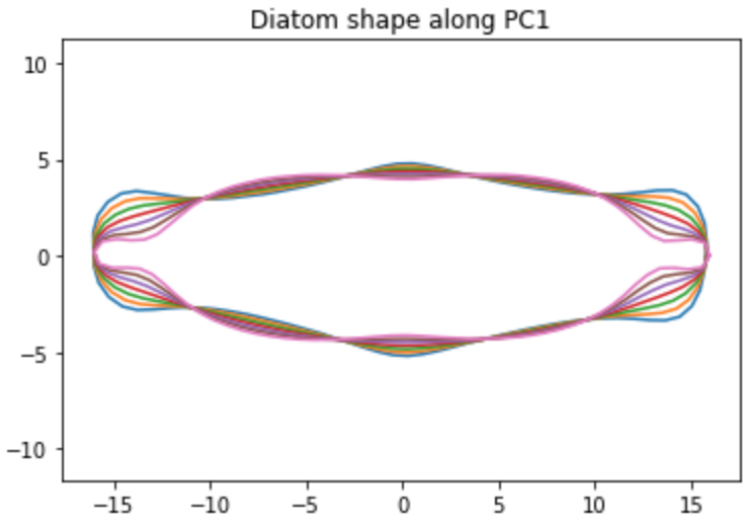
(a)

```
Proportion of variance explained by the first 1 principal components: 0.7718721493017529
Proportion of variance explained by the first 2 principal components: 0.9276996293043025
Proportion of variance explained by the first 3 principal components: 0.9521198453942007
Proportion of variance explained by the first 4 principal components: 0.9637878603999529
Proportion of variance explained by the first 5 principal components: 0.9739084497954094
Proportion of variance explained by the first 6 principal components: 0.98236065164916
Proportion of variance explained by the first 7 principal components: 0.9889975933245944
Proportion of variance explained by the first 8 principal components: 0.9910287023941854
Proportion of variance explained by the first 9 principal components: 0.9926692113360289
Proportion of variance explained by the first 10 principal components: 0.9939926229665051
```

Figur 1: Andel af varians-værdier for de første 10 hovedkomponenter

Som det tydeligt kan ses på billede 1, for at dække 90 % af variansen vil 2 første hovedkomponenter række, 95% — 3 første hovedkomponenter og 99% — 8 første hovedkomponenter.

(b)



Figur 2: Den gennemsnitlige form af en algae repræsenteret af d. 4. hovedkomponent

Figur 2 viser formen af en algae, eller rettere sagt, plottet beskriver hvordan en algae kommer til at se ud, hvis vi prøver at forudse den mest typiske form på en algae. De forskelligt farvede linjer står for de gennemsnitlige forme på en algae, hvis man ændrer på faktorstørrelsen (-3, -2...). Da linjerne ser symmetrisk ud, både langs x- og y-aksen, kan vi konkludere at algeeksemplar nr 4 er rigtig god til at forudse den mest typiske algeform. Hvis man vælger andre hovedkomponenter (fx den femte), så kan man se modeller der kan være "overfitted" og dermed ikke gode til at forudse den generelle udvikling i data'en.

Opgave 2

Vi er givet en tilfældig variabel X med middelværdi μ og varians σ^2. Vi skal vise at:

$$E[(X-\mu)^4] \geq \sigma^4 \tag{1}$$

Vi anvender Jensens' ulighed, som siger at middelværdi af X^2 er større end kvadreret middelværdi af X:

$$EX^2 \geq (EX)^2 \tag{2}$$

Hvis vi definerer en funktion $g(x) = x^2$, så kan (2) skrives som:

$$E[g(X)] \geq g(E[X]) \tag{3}$$

Vi ved at varians er den gennemsnitlige sum af kvadratiske afvigelser:

$$Var = \sigma^2 = E[(x-\mu)^2]$$

og kan derfor omskrive uligheden i (1) som:

$$E[((x-\mu)^2)^2] \geq (E[(x-\mu)^2])^2 \tag{4}$$

Vi indfører $g(x) = x^2$ og omskriver uligheden i (4) som Jensens' ulighed i (3):

$$E(g(x-\mu)^2) \geq g(E(x-\mu)^2) \tag{5}$$

Da vi ved at $E(x-\mu)^2$ af den højre side i (5) er lige med varians (i.e. σ^2), svarer den højre side af uligheden til σ^4 når $g(x) = x^2$.

Opgave 3

(a)

Vi følger proceduren som beskrevet på s. 1050 i bogen "Mathematical statistics" og bestemmer forskriften for γ-konfidenceintervallet for estimatet for middelværdi μ (NB: Ifølge opgavebeskrivelsen er gennemsnittet af observationerne $\overline{x}$ defineret som estimatet for middelværdi $\hat{\mu}$) som:

$$CONF_\gamma\{\hat{\mu} - k \leq \mu \leq \hat{\mu} + k\} \tag{6}$$

Ifølge proceduren findes k som:

$$k = c \cdot \sigma / \sqrt{n}$$

Da vi skal bruge estimatet for spredning $\hat{\sigma}$ i stedet for σ, som er givet som:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2}$$

kan vi omskrive udtrykket i (6) som:

$$CONF_\gamma\left\{\hat{\mu} - \frac{c}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2} \leq \mu \leq \hat{\mu} + \frac{c}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2}\right\}$$

(b)

```
99.0%-confidence interval:
b) Not matching in 131 (out of 10000) experiments, 1.31%
```

Figur 3: Antallet af eksperimenter, hvor den valgte middelværdi ligger udenfor 99.0-konfidensintervallet

Opgave 4

(a)

Vi opstiller nulhypotesen:
Vi forventer at forskellen på blomstringstiderne ville være normaltfordelt med middelværdien 0 (dvs. $\mu = \{(X_1 - Y_1), ..(X_5 - Y_5)\} = 0$). Med andre ord forventer vi at blomstringstiderne ville være de samme og at det ville ikke ville gøre nogen forskel at slå genet ud. (Alternativ hypotese er det modsatte af nulhypotesen)

(b)

Vi skal lave en $t - test$ til niveauet 0.05, og teste om middelværdi er 0 som vi antog i nulhypotesen. M.a.o. skal vi undersøge hvor sikker vi er på at estimatet $\hat{\mu}$ ligger inden for 0.95-konfidensintervallet.
Hvis den fundne sandsynlighed kommer til at ligge under 5%, kan vi forkaste nulhypotesen, og ellers — beholde den.

Blomstringstid	KUK	KMK	$X_i - Y_i$	$(X_i - Y_i) - \hat{\mu})^2$
1	4.1	3.1	1	0.16
2	4.8	4.3	0.5	0.01
3	4	4.5	-0.5	1.21
4	4.5	3	1.5	0.81
5	4	3.5	0.5	0.01
$\sum(X_i - Y_i)$			3	
$\hat{\mu} = \sum(X_i - Y_i)/n$			0.6	
$\sum(X_i - Y_i) - \hat{\mu})^2$				2.2
$\hat{\sigma} = \sqrt{\sum(X_i - Y_i) - \hat{\mu})^2/n - 1}$				0.74

Tabel 1: T-test af nulhypotesen $\mu = 0$ (k=1)

Note: KUK — klone uden knockout, KMK — klone med knockout.

Ved hjælp af CAS-værktøj anvendte vi "T test af middel" med følgende input: middel = 0.6; spredning = 0.74, N = 5; degree of freedom = 4.

Resultatet vi fik gav os sandsynlighed P = 0.14 eller 14%. Da det er meget mere end 5%, betyder det at vi ikke kan forkaste vores nulhypotese.

(c)

Denne gang gentager vi eksperimentet i 4(b) tre gange. Man kunne have gjort det endnu flere gange, men vi har vurderet at tre gange var nok til at forkaste nulhypotesen.

Blomstringstid	KUK	KMK	$X_i - Y_i$	$(X_i - Y_i) - \hat{\mu})^2$
1	4.1	3.1	1	0.16
2	4.8	4.3	0.5	0.01
3	4	4.5	-0.5	1.21
4	4.5	3	1.5	0.81
5	4	3.5	0.5	0.01
1	4.1	3.1	1	0.16
2	4.8	4.3	0.5	0.01
3	4	4.5	-0.5	1.21
4	4.5	3	1.5	0.81
5	4	3.5	0.5	0.01
1	4.1	3.1	1	0.16
2	4.8	4.3	0.5	0.01
3	4	4.5	-0.5	1.21
4	4.5	3	1.5	0.81
5	4	3.5	0.5	0.01
$\sum(X_i - Y_i)$			9	
$\hat{\mu} = \sum(X_i - Y_i)/n$			0.6	
$\sum(X_i - Y_i) - \hat{\mu})^2$				6.6
$\hat{\sigma} = \sqrt{\sum(X_i - Y_i) - \hat{\mu})^2/n - 1}$				0.69

Tabel 2: T-test af nulhypotesen $\mu = 0$ (k=3)

Ved hjælp af CAS-værktøj anvendte vi "T test af middel" med følgende input: middel = 0.6; spredning = 0.69, N = 15; degree of freedom = 14.

Resultatet vi fik gav os sandsynlighed P = 0.0046 eller 0.5%. Da det er

langt under 5%, betyder det at vi kan forkaste vores nulhypotese. Og det var det vi gerne ville.