TMA4240 Statistikk
Høst 2018

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Innlevering 5

Dette er andre av tre innleveringer i blokk 2. Denne øvingen skal oppsummere pensum forelest frem til og med uke 44. Spesielt er det i denne innleveringen fokus på konfidens- og prediksjonsintervall. For å få godkjent innleveringen kreves det at minimum 40% av svarene er riktige, og at man har gjort et ordentlig forsøk på å løse alle oppgavene. Alle deloppgaver teller like mye.

Oppgave 1

En metallurg har vært med på å utvikle en ny legering, og skal presentere ulike egenskaper ved legeringen til sine kolleger. Vi skal her se på bestemmelse av smeltepunktet til legeringen.

På grunn av mindre variasjoner i sammensetningen av legeringen, målefeil og lignende, kan gjentatte målinger av smeltepunktet til legeringen antas å være realisasjoner av uavhengige og normalfordelte variabler med forventningsverdi μ og kjent standardavvik $\sigma = 2^\circ$ C.

a) Anta først (kun i dette punktet) at forventningsverdien er $\mu = 1468^\circ$C.

Metallurgen tar en måling av smeltepunktet. Hva er sannsynligheten for at observert smeltepunkt er lavere enn 1467°C?

Anta så at metallurgen tar åtte uavhengige målinger av smeltepunktet, $X_1, X_2, ..., X_8$. Hva er sannsynligheten for at gjennomsnittet av de åtte målingene av smeltepunktet, $\bar{X} = \frac{1}{8}\sum_{i=1}^{8} X_i$, ligger mellom 1467°C og 1469°C?

Anta i resten av oppgaven at forventningsverdien til legeringens smeltepunkt, μ, er ukjent, men at standardavviket er kjent og lik $\sigma = 2^\circ$ C.

b) Utled et 90% konfidensintervall for μ basert på n uavhengige målinger av smeltepunktet, $X_1, X_2, ..., X_n$.

Metallurgen vil ikke at lengden på intervallet skal overstige 3°C, og har funnet ut at $n \geq 5$ målinger garanterer dette. Bruk de 5 første observasjonene gitt i tabell 1 til å bestemme intervallet numerisk.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1467.4	1468.0	1471.6	1468.6	1468.8	1471.9	1469.4	1466.0

Tabell 1: *Data over smeltepunkt i °C. Det oppgis at gjennomsnittet av de 8 observasjonene er 1469.0°C.*

Oppgave 2

Vi skal i denne oppgaven se på høydefordelingene til menn og kvinner. Anta at høyden til menn er normalfordelt med forventningsverdi $\mu_M = 179$ og varians $\sigma_M^2 = 6^2$, og at høyden til kvinner er normalfordelt med forventningsverdi μ_K og varians σ_K^2

a) Finn sannsynligheten for at en tilfeldig valgt mann er over 185 cm.

Finn sannsynligheten for at en mann er over 185 cm gitt at han er over 179 cm.

b) La $\mu_K = \beta \cdot \mu_M$, der β er en ukjent parameter vi ønsker å estimere. Anta at vi har høydedata fra et tilfeldig utvalg på $n = 5$ kvinner; $X_i \sim N(\mu_K, \sigma_K^2)$, $i = 1, 2, \ldots, n$.

Som estimator for β velges $\hat{\beta} = \dfrac{\sum_{i=1}^n X_i}{n\mu_M}$

Er $\hat{\beta}$ forventningsrett?

Utled et uttrykk for et 95% konfidensintervall for β, og finn tallsvar når vi fra data får $\bar{x} = \sum_{i=1}^n x_i/n = 167.5$ og $s^2 = \dfrac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 5.1^2$.

Oppgave 3

En sjokoladefabrikk ønsker å måle porøsiteten i sjokoladeplatene på samlebåndet. Eieren har anskaffet seg nytt måleutstyr. Utstyrsleverandøren oppgir at måleutstyret har normalfordelt feil med forventning $\mu = 0$ og varians $\sigma_m^2 = 0.0009$.

a) Anta at oppgitt målefeil er korrekt.

Regn ut sannsynligheten for at utstyret måler en porøsitet som er høyere enn det porøsiteten egentlig er.

Regn ut sannsynligheten for at målefeilen er større enn 0.05.

Hvis en tar to uavhengige målinger med utstyret og bruker gjennomsnittet av disse som måling, hva er da sannsynligheten for at avviket mellom målt og sann porøsitet er større enn 0.05?

Sjokoladefabrikkeieren godtar at $\mu = 0$, men tviler på at målevariansen som er oppgitt er korrekt. Han utfører derfor fem uavhengige måleforsøk på en referanse-sjokoladeplate med kjent porøsitet på eksakt 0.15:

0.153	0.132	0.128	0.174	0.163

Du kan se bort fra avrundingsfeilen i fjerde desimal i målingene.

b) Skriv opp den beste forventningsrette estimatoren for målevariansen basert på måleserien og antakelsene over. Begrunn valget ditt.

Skriv opp fordelingen til estimatoren, vis at den er forventningsrett og finn variansen til estimatoren.

Utled et 95% konfidensintervall for den korrekte målevariansen. Kommentér resultatet.

Oppgave 4

En målestasjon for luftforurensning registrerer innholdet av såkalt inhalerbart støv (dvs. partikler med midlere diameter opptil 1/100 mm) i lufta. La X (med enhet milliondels gram pr m^3) være en måling av innholdet av inhalerbart støv på en tilfeldig valgt dag. Målinger av X på ulike dager antas stokastisk uavhengige.

a) Under vanlige forhold antas at X er normalfordelt med forventning $\mu = 35$ og varians $\sigma^2 = 25$. (Dette skal antas bare i dette punktet).

Finn sannsynligheten for at en måling X er over 40, $P(X > 40)$.

Finn også sannsynligheten for at X er mellom 30 og 40, $P(30 < X < 40)$.

Finn sannsynligheten for at summen av målinger på to ulike dager er over 80. Anta at de to målingene er uavhengige.

I en kommune finnes n slike målestasjoner. Fra disse får vi i løpet av en dag målinger $X_1, X_2, \ldots, X_n$ som antas uavhengige og identisk normalfordelte med forventning μ og varians σ^2, der både μ og σ er ukjente parametre. Her vil μ angi graden av forurensning.

Anta at $n = 5$ og at det en dag gjøres følgende målinger av $X_1, \ldots, X_5$:

$$252, 311, 268, 287, 302$$

For senere bruk oppgis at $\sum_{i=1}^{5} X_i = 1420$ og $\sum_{i=1}^{5}(X_i - \bar{X})^2 = 2342$.

b) Hvilke egenskaper bør en god estimator ha?

Sett opp en forventningsrett estimator $\hat{\mu}$ for μ basert på $X_1, \ldots, X_n$. Finn variansen til $\hat{\mu}$. (Vis hvordan du regner den ut).

Sett også opp en estimator for σ^2 og bruk denne til å finne en estimator for variansen til $\hat{\mu}$.

c) Utled et 95% konfidensintervall for μ.

Regn ut tallsvar for intervallgrensene når dataene er som gitt ovenfor.

Anta at det egentlig er 6 målestasjoner i kommunen, men at det på den dagen de 5 observasjonene ovenfor ble gjort, skjedde en feil ved avlesningen fra den siste stasjonen.

d) Utled et intervall som med en sannsynlighet på 95% inneholder den ukjente målingen ved denne stasjonen.

Hva kaller vi et slikt intervall? Hvorfor er dette intervallet bredere enn konfidensintervallet fra punkt c)?

Oppgave 5

En gartner har spesialisert seg på tomatproduksjon i drivhus. Det er kjent at vekten på tomatene er avhengig av lysintensiteten i drivhuset. La Y være vekten (i gram) på en vilkårlig tomat og x være lysintensiteten. Da har en

$$Y = 100 + \beta(x - x_r) + E,$$

hvor x_r er en kjent referanse-lysintensitet som gartneren tradisjonelt bruker, β er en ukjent parameter og E er en normalfordelt (Gaussisk) tilfeldig variabel med forventning 0 g og varians

$\sigma^2 = 15^2$ g^2. Vi har altså at Y er normalfordelt med forventning $\mu = \mu(x;\beta) = 100 + \beta(x - x_r)$ og standardavvik $\sigma = 15$.

En sesong har gartneren brukt referanse-lysintensiteten $x = x_r$ slik at Y er normalfordelt med forventningsverdi 100 g og varians 15^2 g^2. Etter sesongen kontrollerer han tomatene.

a) Regn ut sannsynligheten for at en vilkårlig tomat veier mer enn 110 g.

Regn ut sannsynligheten for at en vilkårlig tomat veier mellom 90 g og 110 g.

Gartneren henter ut to vilkårlige tomater. Regn ut sannynligheten for at den ene er mer enn dobbelt så tung som den andre.

Gartneren har fem drivhus, og en sesong ønsker han å undersøke hvordan tomatenes vekt avhenger av lysintensiteten. Han setter lysintensiteten konstant men ulik i de fem drivhusene.

Etter sesongen velger han ut tre tomater vilkårlig fra hvert drivhus og veier dem.

b) Vis at sannsynlighetsmaksimeringsestimatoren for β, basert på observasjonene beskrevet over, er

$$\hat{\beta} = \frac{\sum_{j=1}^{3}\sum_{i=1}^{5}(Y_{ij} - 100)(x_i - x_r)}{3\sum_{i=1}^{5}(x_i - x_r)^2}$$

Vis at estimatoren er forventningsrett.

Utled et uttrykk for variansen til estimatoren.

Sesongen etter bestemmer han seg for å bruke samme konstante lysintensitet x_0, som er ulik x_r, i alle drivhusene.

c) Vi er interessert i å predikere vekten av en vilkårlig tomat etter denne sesongen. Vi vil altså predikere Y når $x = x_0$, og vi betegner dette med Y_0.

Forventningsverdien til Y_0 kan estimeres ved å estimere β; $\hat{\mu}_{Y_0} = \hat{\mu}(x_0;\beta) = \mu(x_0;\hat{\beta}) = 100 + \hat{\beta}(x_0 - x_r)$. Hva er estimatet når $\hat{\beta}$ er 2.0 og $x_0 - x_r = 5$?

Vis at forventningsverdien til $Y_0 - \hat{\mu}_{Y_0}$ er 0, og at variansen er $15^2 + (x_0 - x_r)^2\mathrm{Var}(\hat{\beta})$. (Hint: $\hat{\beta}$ er forventningsrett og er kun en funksjon av observasjonene fra forrige sesong.)

Ta utgangspunkt i

$$Z = \frac{Y_0 - \hat{\mu}_{Y_0}}{\sqrt{15^2 + (x_0 - x_r)^2\mathrm{Var}(\hat{\beta})}}.$$

Bruk at estimatoren for β er normalfordelt (hvilken fordeling har da Z?) til å utlede et uttrykk for et 95%-prediksjonsintervall for vekten av en vilkårlig tomat etter sesongen.

Bruk at estimatoren for β har varians 1.20, samt at $\hat{\beta}$ er 2.0 og at $x_0 - x_r = 5$ til å regne ut tallsvar.

Fasit

1. a) $0.301, 0.84$ **b)** $[1467.41, 1470.35]$

2. **a**) $0.1587, 0.3173$ **b**) $[0.90, 0.97]$

3. **a**) $0.5, 0.0475$ **b**) $[0.000122, 0.00188]$

4. **a**) $0.16, 0.68, 0.08$ **c**) $[254.0, 314.0]$

5. **a**) 0.2514,0.4972,0.0028 **c**) [78.7,141.3]