TMA4240 Statistikk
Høst 2018

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Innlevering 6

Dette er den siste skriftlige innleveringen. Denne øvingen dekker hele pensum. Spesielt er det i denne innleveringen fokus på hypotesetesting og lineær regresjon. 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

Kraften som er nødvendig for å trekke ut en kork fra en vinflaske er en viktig egenskap ved korken. Dersom kraften er for liten gir ikke korken nok beskyttelse mot innsig av luft for vinen inni flasken. Dersom kraften er for stor, vil korken være vanskelig å fjerne. Kraften (Newton) for en bestemt korktype kan antas normalfordelt (Gaussisk fordelt) med forventning μ og standardavvik σ.

a) Anta i dette punktet at kraften har forventning 310.0 og standardavvik 36.0.

Hva er sannsynligheten for at kraften som er nødvendig for å trekke ut en kork er mellom 300.0 og 310.0?

En kork ble utsatt for en kraft på 330 uten at korken gikk ut. Hva er sannsynligheten for at en kraft større enn 360 er nødvendig for å trekke ut korken?

For et utvalg på 8 korker, hva er sannsynligheten for at utvalgsgjennomsnittet av kreftene er større enn 320?

Et utvalg av 8 tilfeldig valgte flasker med samme korktype ble plukket ut og kreftene som var nødvendig for å fjerne korkene var:

305.98 205.48 322.97 198.58 191.76 288.50 341.18 222.62

En ønsker at kraften som er nødvendig for å fjerne korken har forventning 310. Da gir korken god beskyttelse og er samtidig enkel å ta ut.

b) Anta i dette punktet at kraften har standardavvik 36.0

Test hypotesen $H_0 : \mu = 310$ mot alternativ hypotese $H_1 : \mu \neq 310$ ved signifikansnivå 1%

Dersom μ i virkeligheten er 250, hva er sannsynligheten for forkastning av H_0?

Generelt ønsker en lavest mulig standardavvik for korkene, og det er et krav at standardavviket for kraften ikke er større enn 36.0. Produsenten av korkene hevder at dette kravet er oppfylt.

c) Test $H_0 : \sigma = 36$ mot $H_1 : \sigma > 36$ ved signifikansnivå 5%.

Hva menes med p-verdi?

Finn p-verdien for testen over.

Oppgave 2

Det er kjent at matvareforretninger strategisk plasserer søtsaker i nærheten av kassene i forretningen for å hjelpe på salget. En butikkeier har til nå hatt søtsakene plassert tilfeldig rundt i butikken og erfart at kundene uavhengig av hverandre kjøper søtsaker med sannsynlighet $p_0 = 0.25$.

Butikkeieren ønsker å teste om strategien med å plassere søtsakene i nærheten av kassene i butikken kan hjelpe på salget også i hans butikk, dvs. at personer handler søtsaker med større sannsynlighet enn $p_0 = 0.25$. Han plasserer derfor nå søtsakene i nærheten av kassene og observerer i en undersøkelse at av n tilfeldig valgte kunder handlet nå X kunder søtsaker. Anta at $X \sim b(x; n, p)$, dvs. binomisk fordelt med parametre n og p og at p er ukjent.

a) Vis at sannsynlighetsmaksimeringsestimatoren (maximum likelihood estimatoren) for p er

$$\hat{p} = \frac{X}{n}.$$

Regn ut forventning og varians til estimatoren $\hat{p}$.

Anta at $n = 18$ i undersøkelsen over og at 8 av disse kundene handlet søtsaker.

b) Selger butikkeieren mer søtsaker ved å plassere søtsakene i nærheten av kassene? Formuler problemet som en hypotesetest. Hva blir beslutningen i testen med et signifikansnivå 0.05? (Merk at vi *ikke* har tilstrekkelig antall observasjoner til at sentralgrenseteoremet kan benyttes i testen).

Oppgave 3

Vi ser på konsentrasjonen av et giftstoff i havbunnen like utenfor en fabrikk. Miljøforskriftene sier at konsentrasjonen ikke skal overstige 12 [g/cm^3]. For å kontrollere dette tas prøver av havbunnen. Anta at en prøveverdi Y er normalfordelt med forventning μ og standardavvik σ.

a) De observerte måleverdiene er

$$11{,}7 \quad 12{,}4 \quad 12{,}8 \quad 12{,}9 \quad 13{,}3.$$

Kan vi på grunnlag av dette konkludere med at giftkonsentrasjonen på havbunnen like ved fabrikken er over 12? Formuler problemstillingen som en hypotesetest og utfør testen på signifikansnivå 0,05.

b) Det blir tatt 10 nye målinger, men denne gang i ulike avstander x fra fabrikken. Målingene er

x	10	20	30	40	50	60	70	80	90	100
y	9,9	11,1	9,3	10,6	9,2	9,3	10,0	9,2	10,3	8,4

I tillegg kommer de fem målingene i b). Her er $x = 0$. Det oppgis at $\sum x_i = 550$, $\sum (x_i - \bar{x})^2 = 18\,333{,}33$, $\sum y_i = 160{,}4$ og $\sum x_i y_i = 5245$.

Vi velger å utføre en lineær regresjonsanalyse med Y som avhengig variabel og x som uavhengig variabel. Modellen er

$$E(Y \mid x) = \alpha + \beta x.$$

Beregn estimatene for α og β. Forklar hva estimatet for α beskriver i dette eksemplet.

Regresjonsanalysen gir oss ikke grunnlag for å konkludere med at $\alpha > 12$. Hvorfor ikke? Sammenlign resultatet fra denne analysen med resultatet i b) og kommenter. Hvorfor kan det skje at to slike analyser gir forskjellig konklusjon? Bruk gjerne figur i forklaringen.

Oppgave 4

Et apparat for registrering av stråling er tatt inn for kalibrering og kontroll. Vi skal i denne oppgaven anta at følgende relasjon gjelder mellom apparatets registrerte måleverdi Y og strålingsintensiteten x til en strålingskilde som plasseres i henhold til et gitt forsøksoppsett:

$$Y = \alpha + \beta\, x + \varepsilon$$

Her er α og β konstanter, og ε er en stokastisk (tilfeldig) variabel som, sammen med α, representerer effekten av bakgrunnsstrålingen. Det antas at ε er normalfordelt med forventningsverdi $\mathrm{E}(\varepsilon) = 0$ og varians $\mathrm{Var}(\varepsilon) = \sigma^2$.

a) Kalibreringen innledes ved å registrere m uavhengige måleverdier $y_1, \ldots, y_m$ for Y uten noen strålingskilde, dvs. med bare bakgrunnsstråling. Disse måleverdiene kan da betraktes som et tilfeldig utvalg fra en normalfordelt populasjon.

La $\overline{Y} = \frac{1}{m} \sum_{i=1}^{m} Y_i$ være middelet (gjennomsnittet) basert på dette tilfeldig utvalget. Hvilken fordeling har $\overline{Y}$?

Forklar at en rimelig estimator for α i dette tilfellet er $\overline{Y}$.

Anta i resten av oppgaven at α og σ er *kjente* parametere.

Andre fase i kalibreringen foregår ved å foreta målinger med et utvalg strålingsintensiteter $x_1, \ldots, x_n$, som gir måleverdiene $y_1, \ldots, y_n$. Vi kan da betrakte $y_i - \alpha - \beta\, x_i$, $i = 1, \ldots, n$, som et tilfeldig utvalg fra en normalfordelt populasjon med forventning 0 og varians σ^2.

b) Bruk prinsippet for sannsynlighetsmaksimering (maximum likelihood) til å finne en estimator for koeffisienten β. Alle steg i utledningen av uttrykket for estimatoren skal vises.

Utled også minste kvadratsums-estimatoren for β.

Sammenlign de to estimatorene.

Oppgave 5

En 45-åring startet med løpetrening for 9 år siden, og har hvert år siden deltatt i samme mosjonsløp. Anvendt tid, i minutter, er gitt i tabellen nedenfor.

år i	1	2	3	4	5	6	7	8	9
alder x_i	37	38	39	40	41	42	43	44	45
tid y_i	45.54	41.38	42.50	38.80	41.26	37.20	38.19	38.05	37.45

Det oppgis at $\sum_{i=1}^{9} x_i = 369$, $\sum_{i=1}^{9} y_i = 360.37$, $\sum_{i=1}^{9}(x_i - \bar{x})^2 = 60$, $\sum_{i=1}^{9}(y_i - \bar{y})^2 = 63.28$ og $\sum_{i=1}^{9}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{9}(x_i - \bar{x})y_i = -52.57$.

Vi skal anta at observasjonene kan ses på som realisasjoner av uavhengige normalfordelte variable $Y_1, \ldots, Y_9$, hvor $\mathrm{E}(Y_i) = \alpha + \beta x_i$ og $\mathrm{Var}(Y_i) = \sigma^2$.

a) Skriv opp de vanlige forventningsrette estimatorene $\hat{\alpha}$, $\hat{\beta}$ og $\hat{\sigma}^2$ for α, β og σ^2. Regn ut estimatene for α og β for de gitte dataene. Plott datasettet og den estimerte regresjonslinjen.

Det oppgis at estimatet for σ^2 er 1.568^2.

b) Regn ut et uttrykk for variansen til estimatoren $\hat{\beta}$.

Gjennomfør en test av $H_0: \ \beta = 0$ mot $H_1: \ \beta \neq 0$, på signifikansnivå 1%.

Hva blir den praktiske fortolkningen av testen over?

Løperen ønsker å predikere anvendt tid på mosjonsløpet neste gang (alder $x_0 = 46$ år).

c) Regn ut predikert tid.

Det oppgis at $\mathrm{Var}(\hat{\alpha} + \hat{\beta}x_0) = \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$. Utled et 95% prediksjonsintervall for Y ved $x_0 = 46$ år. Hva blir intervallet med de oppgitte data?

Hvis løperen ber deg predikere anvendt tid om 15 år (alder 60 år), hva vil du svare da?

Fasit

1. **a**) 0.1103,0.28,0.2160 **b**) 0.9838 **c**) 0.005

2. **b**) 0.057

3. **a**) Forkaster H_0

5. **a**) $\widehat{\alpha} = 75.96, \widehat{\beta} = -0.876$ **b**) $\mathrm{Var}(\widehat{\beta}) = \sigma^2 / \sum_{i=1}^{n}(x_i - \overline{x})^2$, Forkast H_0 **c**) 35.66, [31.09, 40.23]