

Oppgave 1

To utvalg, hver av størrelse n , tas fra en normalfordeling med ukjent gjennomsnitt μ og ukjent standardavvik σ . Et 90% konfidensintervall for μ konstrueres med det første utvalget A , og et 95% konfidensintervall for μ konstrueres med det andre utvalget B .

- a) Vil 95% konfidensintervallet basert på utvalg B nødvendigvis være bredere enn 90% konfidensintervallet basert på utvalg A ? Forklar.

Oppgave 2

Vi ønsker å studere sammenhengen mellom maksimal hjerterefrekvens (Y) i slag pr minutt og alder (x) i år for kvinner og innhenter et tilfeldig utvalg av størrelse $n = 52$. For hver kvinne i måles maksimal hjerterefrekvens i slag per minutt y_i og alder x_i , $1 \leq i \leq n$. Anta at parene av målinger er uavhengige, og at Y_i er normalfordelt med forventningsverdi $\beta_0 + \beta_1 x_i$ og standardavvik σ , der β_0 , β_1 og σ er ukjente parametre.

Ved minste kvadraters metode blir β_0 estimert til $\hat{\beta}_0 = 197$ og β_1 estimert til $\hat{\beta}_1 = -0.66$. Dessuten får vi $\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 5581$ og $\sum_{i=1}^n (x_i - \bar{x})^2 = 6969$.

- a) Gi en kort forklaring av minste kvadraters metode (også kalt minste kvadratsums metode eller method of least squares) for linjetilpasning. Hvordan ville du forklare til en person som ikke kjenner til lineær regresjon, hva det betyr at $\hat{\beta}_0 = 197$ og $\hat{\beta}_1 = -0.66$.

Vi ønsker å gjennomføre en hypotesetest for å undersøke om det er en lineær sammenheng mellom x og Y .

- b) Angi passende hypoteser H_0 og H_1 for denne situasjonen. Velg en testobservator og utled en tilhørende beslutningsregel med signifikansnivå 5%. Hva blir konklusjonen på testen?

Oppgave 3

Kathrine kan velge mellom tre alternative ruter når hun hver morgen kjører el-sykkel til jobb. For å finne ut om de tre rutene har ulik forventet kjøretid, og i tilfelle hvilken rute som er raskest, har hun kjørt hver av de tre rutene på hver av de fem arbeidsdagene i en uke. For hver ukedag, for eksempel mandager, har hun trukket en tilfeldig rekkefølge for de tre ulike rutene. Kathrine har notert kjøretid i minutter for de til sammen 15 kjøreturene.

Vi lar Y_{ij} være kjøretid (minutter) for rute nummer j på dag nummer i , der $j = 1, 2, 3$ og $i = 1, 2, 3, 4, 5$. Dag 1 er mandag, dag 2 er tirsdag osv.

De observerte kjøretidene (i minutter) er gitt i tabellen nedenfor.

Ukedag	Rute 1	Rute 2	Rute 3	Gj.snitt
Mandag	22	25	26	24.33
Tirsdag	26	27	29	27.33
Onsdag	25	28	33	28.67
Torsdag	25	26	30	27.00
Fredag	31	29	33	31.00
Gj.snitt	25.80	27.00	30.20	27.67

Kathrines forsøksopplegg skal i denne oppgaven modelleres som et randomisert blokkdesign (*randomized block design*). Observasjonene representeres da ved

$$Y_{ij} = \mu_j + \beta_i + \epsilon_{ij}; \quad j = 1, 2, 3; i = 1, 2, 3, 4, 5, \quad (1)$$

der ϵ_{ij} -ene er uavhengige og $N(0, \sigma^2)$, og der $\sum_{i=1}^5 \beta_i = 0$.

a) Forklar kort hva parametrene μ_j, β_i, σ^2 i modellen (1) uttrykker.

Hva er estimatene for henholdsvis μ_1, μ_2 og μ_3 basert på de gitte dataene?

Regn også ut estimatene for β_i for $i = 1, 2, 3, 4, 5$.

Hvordan vil du nå estimere forventet kjøretid for Rute 3 på en torsdag?

Hvis vi lar Rute svare til “treatment” og Ukedag til “block” i oppsettet i pensum, blir de beregnede kvadratsummene fra dataene gitt ved:

$$\begin{aligned} \text{SSTR} &= 51.73 \\ \text{SSB} &= 71.33 \\ \text{SSE} &= 16.27 \end{aligned}$$

- b) Sett opp den komplette ANOVA-tabellen for analysen basert på de gitte kvadratsummene. (Du skal ikke ta med p-verdier i tabellen).

Tyder resultatene på at det er forskjeller i de forventede kjøretidene for de tre rutene? Velg signifikansnivå 1 % ved testingen.

Vil konklusjonen bli den samme dersom det blir brukt en modell som ikke tar hensyn til blokkeffekten? Begrunn svaret ved regning.

- c) Forklar ideen bak Tukeys metode for parvise sammenligninger av μ_j -ene.

Hvilken fortolkning kan gis av de resulterende intervallene

$$\bar{Y}_{.s} - \bar{Y}_{.t} \pm D\sqrt{MSE}$$

som beregnes i Tukeys metode?

Dersom sannsynligheten $1 - \alpha$ i Tukeys metode settes til 0.95, vil vi ha $D = 1.81$ med de gitte dataene (du skal ikke vise dette). Bruk denne verdien til å regne ut de aktuelle intervallene for μ_1, μ_2, μ_3 .

Hva kan du konkludere fra de beregnede intervallene om forskjeller mellom de tre kjørerutene?

La $m_i = \mu + \beta_i$ for $i = 1, 2, 3, 4, 5$ betegne de forventede kjøretider for hver av de fem dagene i en arbeidsuke, der $\mu = (\mu_1 + \mu_2 + \mu_3)/3$.

Kathrine hadde før forsøket en mistanke om at trafikken på fredager er annerledes enn på de andre arbeidsdagene. Hun ønsket derfor å teste nullhypotesen at forventet kjøretid på fredager er lik gjennomsnittet av forventningene for de fire andre dagene, dvs. å teste nullhypotesen

$$H_0 : m_5 = \frac{m_1 + m_2 + m_3 + m_4}{4} \quad (2)$$

- d) Forklar hvorfor dette svarer til å teste om kontrasten

$$C \equiv \sum_{i=1}^5 c_i m_i = -\frac{1}{4}m_1 - \frac{1}{4}m_2 - \frac{1}{4}m_3 - \frac{1}{4}m_4 + m_5$$

er lik 0.

Gjør rede for at $\hat{m}_i = \bar{Y}_{i.}$ er en forventningsrett estimator for m_i for $i = 1, 2, 3, 4, 5$ og bruk dette til å sette opp en estimator $\hat{C}$ for C .

Definer så

$$SS_C = \frac{\hat{C}^2}{(1/3) \sum_{i=1}^5 c_i^2}$$

Det kan vises (du skal ikke gjøre det) at

1. $SS_C/\sigma^2 \sim \chi_1^2$ under H_0 , hvor χ_1^2 betyr kjikvadrat-fordelingen med 1 frihetsgrad.
2. SS_C and SSE er uavhengige.

Bruk denne informasjonen til å teste nullhypotesen gitt ved (2). Velg signifikansnivå 5% og beskriv resultatet med ord.

Oppgave 4

En skitrener på et skianlegg ønsker å evaluere effekten av en skiundervisning. Hun måler tiden (i sekunder) det tar for en gruppe nybegynnere å fullføre en bestemt løype før og etter en skileksjon. Treneren antar at undervisningen vil redusere tiden det tar å fullføre løypen.

Treneren samler data fra 9 ungdommer. La X_i og Y_i for $i = 1, 2, \dots, 9$ være tiden (i sekunder) ungdom i trenger for å fullføre løypen før og etter skiundervisningen henholdsvis. La D_i være $X_i - Y_i$ hvor de observerte verdiene d_i for D_i ble:

12 8 -3 11 -7 3 3 2 7

- a) Hvilke forutsetninger må du gjøre for å bruke en paret t-test (paired t-test) til å undersøke om det er grunnlag for å konkludere at skiundervisningen reduserer tiden det tar for å fullføre løypen? Formuler nullhypotesen H_0 og alternativ hypotese H_1 . Sett opp testobservatoren og gjennomfør testen når signifikansnivå settes til 5%. Hva blir konklusjonen på testen?

Du kan bruke at $\sum_{i=1}^9 d_i = 36$, $\sum_{i=1}^9 d_i^2 = 458$.

- b) Det ble stilt spørsmål rundt forutsetningene for den paret t-testen, og den ble derfor satt fram et ønske for å bruke tegntesten for paret data.

Denne testen baserer seg på testobservatoren U som er antall av de $n = 9$ par (X_i, Y_i) der $X_i > Y_i$. La $p = P(X_i > Y_i)$ og spesifiser antakelser som gir at U er binomisk fordelt med parametrene $n = 9$ og sannsynlighet p .

Forklar hvordan testingen kan gjøres ved hjelp av tegntesten. Formuler nullhypotesen H_0 og alternativ hypotese H_1 , og forklar hvorfor de kan skrives på den måten. Utfør testen ved å regne ut p-verdien. Hva blir konklusjonen dersom signifikansnivået settes til 5%? (Her skal det ikke brukes tilnærming til normalfordeling).

- c) En alternativ ikke-parametrisk test er Wilcoxon fortegn rang-test (Wilcoxon signed rank test), også kalt Wilcoxons ett-utvalg test.

Forklar hva denne testen går ut på og hvilke forutsetninger den bygger på. Sammenlign med forutsetningene som ble gjort for tegntesten i forrige punkt.

Formuler nullhypotesen H_0 og alternativ hypotese H_1 .

Hvilken testobservator brukes i denne testen? Forklar hvordan man regner den ut og gjennomfør beregningen med de gitte dataene. Hva blir konklusjonen dersom signifikansnivået settes til 5%?