



Eksamen i: Stat100 - Statistikk.

Tid: Mandag 11. desember 2006, kl. 9.00 - 12.30 (tre og en halv time).

Alle typer kalkulator og alle andre hjelpemidler tillatt (hjelpemiddelkode C2).

Oppgaveteksten er på tre sider.

Ved bedømmelsen teller alle delpunkter likt.

Kontakt under eksamen er Torfinn Torp, telefon 91325370.

Oppgave 1 Aggressivitet

Det er en påstand at psykologer har en tendens til å tro at det er en sammenheng mellom en persons aggressivitet og om vedkommende er førstefødt eller ikke. For å undersøke dette valgte en psykolog tilfeldig et stort antall studenter og basert på en aggressivitetstest ble hver student klassifisert til en av fire grupper. Antall prosent i de fire gruppene framgår av tabellen nedenfor. I denne oppgaven antar vi at dette er de sanne prosentene i hver gruppe.

Aggressivitet	Fødselsnummer	
	Førstefødt	Ikke førstefødt
Aggressiv	15%	15%
Ikke aggressiv	25%	45%

- a)** Hva er sannsynligheten for at en tilfeldig valgt student er førstefødt?
- b)** Hva er sannsynligheten for at en tilfeldig valgt student er aggressiv, når vi vet at studenten er førstefødt?
- c)** Vis at kjennetegnene "en tilfeldig valgt student er aggressiv" og "en tilfeldig valgt student er førstefødt" ikke er uavhengige.
- d)** Dersom aggressivitet og fødselsnummer hadde vært uavhengige faktorer, hvilke fire prosenter ville da stått i tabellen (forutsatt at marginalfordelingene for begge faktorene er som ovenfor)?

Oppgave 2 Barns mentale ferdighet

En måte å vurdere små barns mentale ferdighet på er å gi dem klosser og be dem bygge et tårn så høyt de klarer. Et slikt byggeklossforsøk ble gjentatt to ganger med en måneds mellomrom. I undersøkelsen deltok 15 barn. Observerte verdier på responsvariabelen, som her er antall sekunder barnet brukte på byggingen, er gitt i tabellen nedenfor.

Barn	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Gjennomsnitt
1. forsøk, X	30	19	19	23	29	178	27	20	12	39	14	81	17	31	52	39.4
2. forsøk, Y	30	6	14	8	14	52	14	22	17	8	11	30	14	17	15	18.1

For alle 15 barna ble differansen mellom tiden på første og andre forsøk beregnet. Utvalgsstandardavviket for disse differansene er 32.82.

- a) Anta at $E(X_i) = \lambda_i + \mu$ og $E(Y_i) = \lambda_i$, der λ_i -ene og μ er ukjente parametre og i angir barnets nummer. Formuler en nullhypotese med alternativ hypotese for å undersøke en påstand om at barn bruker en annen tid på en slik bygging i andre forsøk enn i første forsøk. Angi en passende testobservator. Beregn verdien til testobservatoren, basert på det tilgjengelige datamaterialet, og utfør testen med 1% nivå. Presiser kort de forutsetningene som gjøres.

Det viste seg i ettertid at tiden 178 registrert i første forsøk for barn nummer 6 var gal og i virkeligheten skulle vært 78. Når denne feilen ble korrigert, ble utvalgsgjennomsnittet og utvalgsstandardavviket for de 15 differansene (mellom første og andre registrerte tid) henholdsvis 14.6 og 15.73.

- b) Beregn et 99% konfidensintervall for μ basert på det korrigerede datamaterialet og bruk dette intervallet til å teste hypotesen fra punkt a.
- c) La D_1, D_2, \dots, D_n betegne n differanser (mellom første og andre registrerte tid) og anta at de er uavhengige tilfeldige variabler som alle er normalfordelte med forventning μ og varians σ^2 . Betrakt følgende to estimatorene for σ^2 :
- $$\widehat{\sigma}_{(1)}^2 = \frac{1}{n} \sum_{i=1}^n (D_i - \bar{D})^2 \quad \text{og} \quad \widehat{\sigma}_{(2)}^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2.$$
- Finn forventning og varians til begge estimatorene. Hvilken av de to estimatorene vil du bruke? Begrunn svaret. (Hint: $\frac{1}{\sigma^2} \sum_{i=1}^n (D_i - \bar{D})^2$ er kjikvadratfordelt med $n-1$ frihetsgrader og har derfor forventning lik $n-1$ og varians lik $2(n-1)$).

Oppgave 3 Protein i melk

Melk er en viktig proteinkilde. Vi ønsker å studere hvordan innholdet av protein i melk varierer med kuas melkeproduksjon. For dette formålet har vi tilgjengelig et datamateriale fra 14 kyr med observerte verdier på x = melkeproduksjon (kg pr. dag) og tilhørende Y = innhold av melkeprotein (kg pr. dag). Dataene er gitt i tabellen nedenfor.

Ku	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Gjennomsnitt
x	27.2	38.2	37.6	22.7	26.6	32.2	40.2	21.8	20.2	21.3	28.0	42.7	23.0	32.2	29.56
Y	0.87	1.07	1.13	0.76	0.77	1.07	1.16	0.69	0.64	0.72	0.85	1.20	0.74	0.96	0.9021

Det ble bestemt å bruke en enkel, lineær regresjonsanalyse. Noen av resultatene fra en slik analyse, utført i Minitab, er gitt nedenfor.

<i>Predictor</i>	<i>Coef</i>	<i>SE Coef</i>	<i>T</i>	<i>P</i>
Constant	0,17558	0,04640	3,78	0,003
x	0,024576	0,001523	16,14	0,000

Analysis of Variance

<i>Source</i>	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>P</i>
Regression	1	0,46023	0,46023	260,48	0,000
Residual Error	12	0,02120	0,00177		
Total	13	0,48144			

- a) Skriv ned den modellen som ligger til grunn for analysen. Angi estimater for alle modellens ukjente parametre. Dersom melkeproduksjonen for en ku øker fra 25 til 35 kg pr. dag, angi da et estimat for den forandringen du forventer i innhold av melkeprotein.
- b) Prediker innholdet av melkeprotein dersom melkeproduksjonen for en ku er 30 kg pr. dag. Beregn også et 95% prediksjonsintervall for dette innholdet av melkeprotein.
- c) Fra den modellen som er tilpasset ovenfor, synes du det da er riktig uten videre å påstå at melk inneholder en fast forventet prosent protein (uavhengig av kuas melkeproduksjon)? Begrunn svaret.

Torfinn Torp

John Tyssedal