Skriftlige prøveopgaver i dataanalyse – sæt 1

Fysik 2 & Nano 4
Kursusholder: Ege Rubak

Eksamensdato: Tirsdag den 11. Maj 2010

Bemærk: Til eksamen vil der typisk indgå 2-3 af sådanne dataanalyseopgaver sammen med 2-3 differentialligningsopgaver.

Ved bedømmelsen vil der blive lagt vægt på såvel korrekt metode og korrekt svar derfor skal metoden klart fremgå af besvarelsen – held og lykke.

Opgave 1. En sælger ved en brugtvognsforhandler er på provision, og når han sælger en bil, får han 4200 kr for en personbil og 4800 kr for en varebil. Han forventer at kunne sælge et antal person- og varebiler pr. dag med flg. sandsynligheder:

Antal personbiler	0	1	2	3
Sandsynlighed	0.3	0.4	0.2	0.1

Antal varebiler	0	1	2
Sandsynlighed	0.4	0.5	0.1

1. Beregn det forventede antal personbiler og det forventede antal varebiler, sælgeren kan sælge på en dag.
   $\mathrm{E}(X_p) = 1.1$ $\mathrm{E}(X_v) = 0.7$
2. Beregn standard afvigelsen på antallet af personbiler og standardafvigelsen på antallet af varebiler, sælgeren kan sælge på en dag.
   $\sigma_p = 0.94$ $\sigma_v = 0.64$
3. Beregn den forventede samlede provision for både personbiler og varebiler, som sælgeren har på en dag.
   Lad $Y = 4200X_p + 4800X_v$, så er $\mathrm{E}(Y) = 7980$
4. Beregn standardafvigelsen på sælgerens samlede provision på en dag, idet vi antager, at antal solgte personbiler og antal solgte varebiler er uafhængige.
   $\sqrt{\mathrm{Var}(Y)} = 7020$

Opgave 2. Antallet af minutter, det tager at reparere en automatisk påfyldningsmaskine til et bestemt fødevareprodukt, er normalfordelt med middelværdi 120 minutter og varians 16 minutter^2. Hvis påfyldningsmaskinen er nede mere end 125 minutter, skal maskinen rengøres og allerede producerede fødevarerprodukter skal kasseres. Dette er bekosteligt og ikke ønskværdigt.

1. Hvad er sandsynligheden for, at påfyldningsmaskinen er nede mere end 125 minutter?
   I begge disse opgaver handler det om, at bruge formlen for at standardisere en stokastisk variabel. Dvs. hvis $X \sim N(\mu, \sigma^2)$ så er gælder for den standardiserede stokastiske variabel $Z = (X - \mu)/\sigma$, at $Z \sim N(0, 1)$. I denne opgave får vi: $P(X > 125) = P(\frac{X-120}{4} > \frac{125-120}{4}) = P(Z > 1.25) = 1 - P(Z < 1.25) = 1 - 0.89 = 0.11$ (de 0.89 er fundet i Matlab, hvis man slår op i tabellen kan man kun se, at $P(Z \leq 1.25) \approx 0.90$, så et svar på 0.10 er helt fint).
2. En medarbejder ønsker at kende et tidsinterval, som han med 95% sandsynlighed kan forvente, at maskinen er nede i. Find det 95% sandsynlighedsinterval, der netop er symmetrisk omkring middelværdien.
   Vi skal bestemme k så $P(\mu - k < X < \mu + k) = 0.95$.
   Pga. symmetri er det det samme som: $P(X < \mu + k) = 0.975$
   Dette er også: $P(\frac{X-\mu}{\sigma} < \frac{\mu+k-\mu}{\sigma}) = P(Z < \frac{k}{\sigma}) = 0.975$
   Dette betyder, at $\frac{k}{\sigma} = 1.96$ (fra tabellen). Intervallet er så $\mu \pm 1.96\sigma$, hvilket med tallene her er $[112.16; 127.84]$.

Opgave 3. Nedenstående data angiver vægten af 10 studerende hhv. før og efter en måneds fælles slankekur.

Før:	76	86	71	76	78	78	79	83	76	76
Efter:	74	83	72	77	74	76	78	81	74	75

1. Har slankekuren haft en signifikant effekt? (brug et signifikansniveau på 5%)
   Ja. Kritisk værdi, $t_{1-\alpha/2}(9) = 2.26$. Teststørrelse fra data, $t_{\text{obs}} = 3$. P-værdi, $p = P(|T| > |t_{\text{obs}}|)$: $0.01 < p < 0.02$.
2. Bestem et 99% konfidensinterval for vægtforskellen. $[-.125; 3.125]$

Opgave 4. Væggen i en 2 liters plastikflaske skal have en vis tykkelse, således at flasken ikke revner ved slag eller lignende. Ved en kvalitetskontrol udtages derfor en stikprøve på 25 flasker og der måles for vægtykkelsen et stikprøvegennemsnit på $\bar{x} = 4.05mm$ samt en stikprøvespredning på $s = 0.08mm$. Det antages, at observationerne er uafhængige normalfordelte.

1. Bestem et 95% konfidensinterval for middelværdien af vægtykkelsen.
   $[4.017; 4.083]$
2. Test på 5% signifikansniveau om vægtykkelsen er signifikant mindre end $4mm$.
   $H_0 : \mu \geq 4$, $H_1 : \mu < 4$. Kritisk værdi, $t_\alpha(24) = -1.71$ (da testet kun er en-sidet er det kun negative værdier af teststørrelsen der er kritisk). Teststørrelse fra data, $t_{\text{obs}} = 3.125$. Tydeligvis er 3.125 ikke mindre end -1.71 og vi kan altså ikke afvise nulhypotesen om at vægtykkelsen er mindst $4mm$. Dvs. den er ikke signifikant mindre end $4mm$.

Opgave 5. I løbet af juni måler en hospitalslæge hver dag højeste dagstemperatur, t, samt antallet af patienter, n, der ankommer med symptomer på dehydrering. Fra de målte data beregner lægen følgende: Stikprøvegennemsnittet for t, $\bar{t} = 22.4$, stikprøvegennemsnittet for n, $\bar{n} = 111.4$, stikprøvevariansen for t, $s_t^2 = 10.5$, stikprøvevariansen for n, $s_n^2 = 4.0$, samt stikprøvekovariansen mellem t og n, $s_{tn} = 5.7$.
Bemærk at notationen kan være lidt forvirrende her. Der er observeret i hele juni, dvs. 30 dage. Vores målinger er altså antallene $n_1, \ldots, n_{30}$ givet temperaturerne $t_1, \ldots, t_{30}$. Normalt kalder vi disse hhv. $\mathbf{y}$ og $\mathbf{x}$ og betegner antallet af målinger $n = 30$.

1. Estimer en regressionsmodel for dehydration givet temperaturen.
   $\hat{\beta}_1 = 5.7/10.5 = 0.54$, $\hat{\beta}_0 = 111.4 - 0.54 \cdot 22.4 = 99.24$.
2. Er der en signifikant effekt af temperaturen?
   Ja. Vi laver et hypotesetest: $H_0 : \beta_1 = 0$, $H_1 : \beta \neq 0$. Teststørrelsen er $t_{\text{obs}} = 9.8$ og den kritiske værdi er 2.05, så vi kan afvise, at der ikke er en effekt af temperaturen. Med andre ord, så er der en effekt af temperaturen.
3. Angiv en approximativ p-værdi for effekten af temperaturen.
   $p = P(|T| > |t_{\text{obs}}|) = P(|T| > 9.8) = 2P(T > 9.8) < 2P(T > 2.76) = 2(1 - 0.995) = 0.01$. Dvs. p-værdien er mindre end 0.01. Dette er ikke en specielt god approksimation, men den bedste vi kan lave ud fra tabellen.
4. Bestem et 95% konfidensinterval for begge regressionsparametre.
   Konf. int. for $\beta_0 : 99.24 \pm 2.05 \cdot 1.26 = 99.24 \pm 2.58$.
   Konf. int. for $\beta_1 : 0.54 \pm 2.05 \cdot 0.056 = 0.54 \pm 0.11$.
5. Beregn determinationskoefficienten.
   $R^2 = 0.77$

Opgave 6. Et skrabelodslotteri angives til at have følgende gevinstsandsynligheder:

Gevinst i kr:	0	50	500	1000	5000	10000
Chance i pct:	80	10	5	4	0.9	0.1

Der udtages tilfældigt 1000 lodder og antallet af de forskellige gevinster er opsummeret her:

Gevinst i kr:	0	50	500	1000	5000	10000
Antal lodder:	820	90	45	39	6	0

1. Test på 5% signifikansniveau om den angivne fordeling er korrekt.
   Vi kan ikke afvise at den angivne fordeling er korrekt. Den observerede teststørrelse er: $X^2_{\text{obs}} = 4.025$. Kritisk værdi: 11.07.
2. Angiv en approximativ p-værdi for det udførte test.
   $p = P(X^2 > X^2_{\text{obs}})$: $0.50 < p < 0.75$.

Husk at angive studienummer på alle afleverede ark samt hvor mange sider din besvarelse består af.