TMA4245 Statistikk
Vår 2017

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Innlevering 4

Dette er den andre av to innleveringer i blokk 2, og den siste innleveringen i øvingsopplegget. Tema for oppgavene er blant annet hypotesetesting og sammenligning av utvalg (kap. 10) og lineær regresjon (kap. 10). Sannsynlighetsmaksimering (kap. 9) er også aktuelt. Alle deloppgaver teller like mye.

Oppgave 1

Vi ser på kostnaden av et veiprosjekt. La X være en kontinuerlig stokastisk (tilfeldig) variabel som angir den faktiske kostnaden pr. meter for veien som skal bygges. Kostnaden pr. meter er avhengig av grunnforhold og pris på materialer. Vi antar at X er normalfordelt med ukjent forventningsverdi $\mathrm{E}(X) = \mu$ og ukjent standardavvik $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sigma$.

En ekspertgruppe mener at forventet kostnad pr. meter for veien som skal bygges blir 10000 kr/meter. I tillegg er det samlet inn data fra $n = 9$ veiprosjekter med tilsvarende grunnforhold og materialkostnader, og dataene finnes i tabell 1. Her er x_i kostnaden i kr/meter for veiprosjekt nummer i, og det oppgis at $\sum_{i=1}^{9} x_i = 106480$ og $\sum_{i=1}^{9}(x_i - \bar{x})^2 = 49295335$.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
10099	10925	15397	11676	11823	15788	12652	8337	9783

Tabell 1: Data over kostnad i kr/meter for $n = 9$ veiprosjekter.

Prosjektlederen er skeptisk til kostnadsanslaget på $\mu = 10000$ kr/meter, og mener at grunnforholdene tilsier at forventet kostnad i kr/meter er høyere.

a) Formulér dette som en hypotesetest ved å definere nullhypotese og alternativ hypotese.

Sett opp en testobservator og finn forkastingsområdet. Hva blir konklusjonen på testen, med data gitt i tabell 1, når signifikansnivået er $\alpha = 0.01$?

Regn ut p-verdien ved å bruke tabell 2.

Oppgave 2

Ved behandling av visse kreftformer får pasientene kurer der en bestemt type medisin blir injisert i blodet i løpet av 24 timer. Alle pasienter får tilført samme dose medisin. Ved avslutningen av kuren blir konsentrasjonen av medisin i blodet målt. Medisinkonsentrasjonen måles i milligram medisin per liter blod. For at behandlingen skal ha ønsket effekt bør medisinkon-

t	1.8	1.9	2.0	2.1	2.2	2.3	2.4
$\nu = 7$	0.943	0.950	0.957	0.963	0.968	0.973	0.976
$\nu = 8$	0.945	0.953	0.960	0.966	0.971	0.975	0.978
$\nu = 9$	0.947	0.955	0.962	0.967	0.972	0.977	0.980

Tabell 2: Kumulativ sannsynlighet i t-fordelingen. For en stokastisk variabel T som er t-fordelt med ν frihetsgrader, så viser tabellen $P(T \leq t)$ for ulike verdier av t.

sentrasjonen ved avslutningen av kuren helst overstige 5 mg/l. På grunn av bivirkninger blir det ansett som uheldig om medisinkonsentrasjonen overstiger 12 mg/l. La Y betegne målt medisinkonsentrasjon ved avslutningen av en kur, og anta at Y er normalfordelt med forventningsverdi μ og varians σ^2. Målt medisinkonsentrasjon ved avslutningen av ulike kurer antas uavhengige.

Anta at μ er ukjent, mens $\sigma^2 = 2^2$ antas kjent. Fra åtte ulike kurer har man registrert dataene:

kur i	1	2	3	4	5	6	7	8
y_i	7.1	9.2	10.8	12.0	6.1	8.2	8.7	7.7

a) Skriv opp en rimelig estimator for μ, og regn ut estimatet.

Utled et 95% konfidensintervall for μ. Hva blir intervallet med de oppgitte dataene?

Legene har etterhvert funnet ut at i stedet for å gi alle pasienter samme dose medisin, vil det være gunstigere å justere dosene etter hvor syk pasienten er og hvor godt han/hun tåler bivirkningene. La x være dosen. Vi antar at x kan kontrolleres, dvs x er ikke stokastisk.

Man antar at en god lineær regresjonsmodell for sammenhengen mellom x og Y vil være

$$Y = \beta x + E,$$

der β er en ukjent konstant og E er en normalfordelt stokastisk variabel med forventningsverdi 0 og kjent varians $\sigma_E^2 = 2^2$.

b) Hvorfor er det i dette tilfellet rimelig å ikke ha med noe konstantledd i den lineære regresjonsmodellen?

Vis at sannsynlighetsmaksimeringsestimatoren (SME) for β basert på n uavhengige observasjoner blir

$$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

der x_i og Y_i er henholdsvis dose og målt medisinkonsentrasjon for observasjon nummer i.

Regn ut forventningsverdien og variansen til $\hat{\beta}$.

Det har i løpet av ti kurer på ulike pasienter blitt observert følgende sammenhørende verdier for x og Y:

kur i	1	2	3	4	5	6	7	8	9	10
x_i	4.5	4.0	5.5	7.0	8.0	8.5	9.0	6.5	6.0	5.0
y_i	6.2	5.2	7.3	8.7	9.0	10.5	10.3	8.2	7.4	7.0

Det oppgis at $\sum_{i=1}^{10} y_i x_i = 536.4$ og $\sum_{i=1}^{10} x_i^2 = 436$.

Før legene gir en pasient en viss dose x_0 ønsker de å vite noe om hvilken målt medisinkonsentrasjon Y_0 man kan regne med at dette vil gi. Du skal hjelpe legene ved å lage et 95% prediksjonsintervall.

c) Hva er tolkningen av et 95% prediksjonsintervall?

Utled et 95% prediksjonsintervall for Y_0 når $x_0 = 8$ ved å bruke de oppgitte dataene.

Oppgave 3

La $X_1, X_2, \ldots, X_n$ vere eit tilfeldig utval frå ein normalfordelt populasjon med forventningsverdi μ_X og standardavvik σ_X. Vidare er $Y_1, Y_2, \ldots, Y_m$ eit tilfeldig utval frå ein normalfordelt populasjon med forventningsverdi μ_Y og standardavvik σ_Y. Anta at dei to utvala er uavhengige av kvarandre.

Definer $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j$, $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ og $S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y})^2$.

a) Vi ønskjer å undersøke om det er grunn til å tru at μ_Y er større enn μ_X.

Sett opp nullhypotese og alternativ hypotese.

Vel testobservator, og oppgje kva sannsynsfordeling denne har når nullhypotesen er sann.

Finn ein forkastningsregel når vi vel signifikansnivå 0.05.

b) I denne oppgåva skal me undersøke målingar av absorbert fukt i 12 ulike prøvar av 2 forskjellige betongblandingar etter at prøvane har vore utsatt for fukt i 48 timar (Tabell 3). Tala er henta frå tabellen på s. 508 i læreboka (Walpole, Myers, Myers og Ye). Nytt Matlab-skriptet under til å utføre hypotesetesten du har formulert i **a)**. Du kan nytte at talet på fridomsgrader v er det største heiltalet mindre enn

$$\frac{(S_X^2/n + S_Y^2/m)^2}{\left(S_X^2/n\right)^2/(n-1) + \left(S_Y^2/m\right)^2/(m-1)}.$$

```
% Observasjoner
x = [551, 457, 450, 731, 499, 632];
y = [595, 580, 508, 583, 633, 517];

% Utvalsstorleik
n = length(x);
m = length(y);

% Estimer gjennomsnitt og varians
xbar = mean(x);
ybar = mean(y);
sx2 = var(x);
sy2 = var(y);

% Observert t-verdi
```

x	551	457	450	731	499	632
y	595	580	508	583	633	517

Tabell 3: Fuktabsorpsjon i to ulike betongblandingar.

```
tobs = (ybar - xbar)/sqrt(sx2/n + sy2/m);

% Fridomsgrader
v = (sx2/n + sy2/m)^2/((sx2/n)^2/(n-1) + (sy2/m)^2/(m-1));

% Finn kritisk t-verdi ( P(T>tc) = 0.05 )
tc = icdf('t', 0.95, floor(v));

% Testkonklusjon
fprintf('tobs = %.4f, tc = %.4f\n\n', tobs, tc)
if tobs > tc
    fprintf('==> Forkast H0')
else
    fprintf('==> Ikkje forkast H0')
end
fprintf('\n\n')
```

Korleis vil du omsetje svaret (forkast eller ikkje forkast nullhypotesen) i denne situasjonen?

c) Utlei eit uttrykk for eit 90%-konfidensintervall for forskjellen $\mu_Y - \mu_X$. Nytt Matlab-skriptet under til å finne konfidensintervallet.

```
% Rekn ut endepunkta til intervallet
L = ybar - xbar - tc*sqrt(sx2/n + sy2/m);
U = ybar - xbar + tc*sqrt(sx2/n + sy2/m);

% Skriv intervallet til skjermen
fprintf('90%% konfidensintervall:\n')
fprintf('[%.4f, %.4f]\n\n',L,U)
```

Testen over kunne vore gjort med Matlab sin innebygde t-test (merk at talverdiane for konfidensintervallet ikkje er eksakt identiske sidan Matlab nyttar det eksakte talet på fridomsgrader):

```
[h,pverdi,konfidensintervall,stats] = ttest2(y,x, 'Vartype', 'unequal','Alpha', 0.1)
```

Oppgave 4 Matlaboppgave om lineær regresjon

Fila `taxi.txt`, som er tilgjengelig på
`https://www.math.ntnu.no/emner/TMA4245/2017v/inn4/taxi.txt`,
inneholder data fra $n = 3000$ taxiturer i New York City i mars 2016. Første kolonne er kjørt distanse, i engelske mil. Andre kolonne er betalt beløp, inkludert tips, i amerikanske dollar.

Vi vil bruke en lineær regresjonsmodell til å modellere prisen på en taxitur som en funksjon av turens lengde. La derfor x_i være distansen, og y_i være betalingen for tur nr. i, for $i = 1, \ldots, n$.

a) For tur nr. i, skriv opp uttrykket for y_i som en funksjon av forklaringsvariabelen x_i, parametrene β_0 og β_1 og feilleddet ϵ_i.

Hvilke antakelser gjør vi når vi bruker denne modellen?

b) Bruk Matlab til å tilpasse modellen til dataene. Lag deretter følgende figurer:

- En figur som viser regresjonslinja sammen med observasjonene (x_i, y_i), $i = 1, \ldots n$.
- En figur hvor du plotter residualene $y_i - \hat{y}_i$ mot avstandene x_i.
- Et normal-kvantil-kvatil-plott (normal QQ-plot), hvor de empiriske kvantilene til residualene sammenlignes med de teoretiske kvantilene i standard normalfordelingen.

Hvor godt passer modellen til dataene, og i hvilken grad er antakelsene du skrev opp i forrige punkt oppfylt?

Forslag til Matlab-kode:

```
taxidata = dlmread('taxi.txt');

x = taxidata(:,1);
y = taxidata(:,2);

n = length(x);

Sxy = sum((x - mean(x)) .* y);
Sxx = sum((x - mean(x)).^2);

betahat1 = Sxy/Sxx;
betahat0 = mean(y) - betahat1*mean(x);

xx = linspace(0, 45);
y_linje = betahat0 + xx*betahat1;

y_tilpasset = betahat0 + betahat1*x;
residualer = y - y_tilpasset;

figure()
subplot(2,2,[1,2])
hold on
plot(x, y, 'o')
plot(xx, y_linje, 'r-')
hold off
box on
xlabel('Strekning (mi)')
ylabel('Pris (USD)')
```

```
title('Tilpasset regresjonsmodell')
set(gca, 'FontSize', 14)

subplot(2,2,3)
hold on
plot(x, residualer, 'o')
plot(xx,zeros(size(xx)), '--')
hold off
box on
xlabel('Strekning (mi)')
ylabel('Residualer (USD)')
title('Residualplott')
set(gca, 'FontSize', 14)

subplot(2,2,4)
qqplot(residualer)
title('Kvantil-kvantil-plott av residualer')
set(gca, 'FontSize', 14)
box on
```

c) La den tilfeldige variabelen Y_0 være betalingen for en fremtidig tur med lengde $x_0 = 5$ miles.

Finn et punktestimat $\hat{y}_0$ for Y_0.

Utled et uttrykk for et 95% konfidensintervall for Y_0. Bestem endepunktene til intervallet numerisk ved hjelp av Matlab.

Hint: Hvis den fremtidige observasjonen er $Y_0 = \beta_0 + \beta_1 x_0 + \epsilon$ og uttrykket for punktestimatet er $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, hvor $\hat{\beta}_0$ og $\hat{\beta}_1$ er tilfeldige variabler, så kan differansen mellom Y_0 og $\hat{Y}_0$ skrives

$$\begin{aligned} Y_0 - \hat{Y}_0 &= \beta_0 - \hat{\beta}_0 + (\beta_1 - \hat{\beta}_1)x_0 + \epsilon \\ &= \beta_0 - (\bar{Y} - \hat{\beta}_1 \bar{x}) + \beta_1 x_0 - \hat{\beta}_1 x_0 + \epsilon \\ &= \beta_0 + \beta_1 x_0 - \bar{Y} - (x_0 - \bar{x})\hat{\beta}_1 + \epsilon. \end{aligned}$$

Her er de to første leddene konstante.

Forslag til Matlab-kode:

```
x0 = 5;
yhat0 = betahat0 + betahat1*x0

y_tilpasset = betahat0 + betahat1*x;
residualer = y - y_tilpasset;
s2 = var(residualer);

varYhat0 = s2*(1 + 1/n + (x0 - mean(x)).^2/Sxx);
```

```
stdYhat0 = sqrt(varYhat0);

tq = icdf('t', 0.975, n-2);

y_nedre = yhat0 - tq*stdYhat0
y_oevre = yhat0 + tq*stdYhat0
```

Fasit

1. **a**) 0.029

2. **a**) $\widehat{\mu} = \overline{Y}, 8.725, [7.34, 10.11]$ **b**) $\mathrm{E}(\widehat{\beta}) = \beta$, $\mathrm{Var}(\widehat{\beta}) = \sigma_E^2 / \sum_{i=1}^n x_i^2$ **c**) $[5.64, 14.04]$

3. **a**) Testobservator: $T = (\overline{Y} - \overline{X})/(S_X^2/n + S_Y^2/m)^{1/2}$, forkast H_0 viss $t > t_{v,0.05}$ **b**) Forkastar ikkje H_0 **c**) $[-79.32, 111.32]$

4. **c**) 20.86, $(11.20, 30.52)$