UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — Statistiske metoder og dataanalyse 1.
Eksamensdag:	Tirsdag 11. desember 2012.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 5 sider.	
Vedlegg:	Tabell over Poisson-, normal-, t-, og F-fordeling.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110.

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1.
Lånekassen definerer en *fulltidsstudent* som en som samler minst 30 studiepoeng per semester. Vi skal ved hjelp av en spørreundersøkelse finne ut om andelen fulltidsstudenter er forskjellig ved NTNU i Trondheim og ved UiO i Oslo. Kall andelen fulltidsstudenter ved UiO for p_1 og andelen fullltidsstudenter ved NTNU for p_2. Vi spør et tilfeldig utvalg på størrelse n fra UiO og et tilfeldig utvalg på størrelse m fra NTNU om de oppfyller dette kravet inneværende semester.

a) Utled en test på nivå $\alpha = 0.05$ for å teste

$$\mathrm{H}_0 : p_1 - p_2 = 0$$

mot en tosidig alternativhypotese. Anta at n og m er store nok til at du kan bruke tilnærming til normalfordeling.

b) Konstruer et 95% konfidensintervall for forskjellen $p_1 - p_2$. Forklar hvorfor uttrykket for variansen blir forskjellig fra det du brukte i a).

c) La $n = 200$ og $m = 400$. 102 av de spurte studentene ved UiO og 248 av de spurte ved NTNU kvalifiserte som fulltidsstudenter. Sett inn både i a) og b) og kommenter.

Oppgave 2.
Et advokatfirma i USA blir kontaktet av en gruppe kvinnelige ansatte i et elektronikkfirma, som mener at firmaet de jobber i gir kvinnene dårligere lønnsvilkår enn mennene. Advokatfirmaet plukker ut tilfeldige ansatte i elektronikkfirmaet, 10 kvinner og 9 menn, og gjennomgår deres lønnsopplysninger. Spesielt samler de tall for utbytte ved siste

(Fortsettes på side 2.)

personlige lønnsforhandling (i \$) og en vurdering (0-100 poeng) for hvor tilfreds nærmeste overordnet er med medarbeiderens arbeid det siste året. Det sies fra firmaet at utbyttet ved lønnsforhandling blant annet baseres på denne vurderingen (variabelen 'score'). Data er lagt ved til slutt i oppgavesettet, sammen med en utskrift fra R og et plott. Lønnsøkning fra lønnsforhandlingene heter der 'salary_incr'. Variabelen 'sex' er medarbeiderens kjønn, kodet som 1 for kvinner og 0 for menn.

Det er opplagt at kvinnene i utvalget i gjennomsnitt har fått langt lavere lønnstillegg enn mennene. Firmaet svarer advokatene at det må være fordi kvinnene generelt har fått lavere poengsum for arbeidsinnsatsen. Du skal først undersøke disse poengene for arbeidsinnsats (score). Gjennomsnittlig score i utvalget for kvinner var 43.5, og for menn var den 55.56. Vi oppgir også at estimerte varianser var 922.50 for score for kvinner og 552.78 for score for menn.

a) Vi antar at variabelen score er normalfordelt både for kvinner og for menn, med forventning henholdsvis μ_1 og μ_2 og samme varians σ^2. Gjennomfør en t-test for å avgjøre om det er grunnlag i dataene for å påstå at kvinnene gjør en dårligere jobb (målt ved score fra overordnet) enn mennene. Bruk nivå $\alpha = 0.05$. Du må skrive opp hypotesene og utrykket for testobservator, og spesifisere hvilken fordeling du har brukt. Skriv en konklusjon. Du kan støtte deg på utskriften fra R.

b) I a) har vi antatt lik varians i fordelingene til score for kvinner og for menn. Er dette en rimelig antakelse? Svar ved hjelp av en hypotesetest. Sett opp hypoteser, testobservator og fordeling, og gjennomfør testen på nivå $\alpha = 0.1$. Hva blir din konklusjon? Du kan bruke informasjon fra R-utskriften.

Heretter skal vi betrakte score som en fast forklaringsvariabel (ikke stokastisk). For at advokatene skal kunne bruke tallmaterialet i saken mot elektronikkfirmaet, bestemmer firmaets statistikk-konsulent at man skal utføre en multippel regresjon, der utbytte i lønnsforhandling er responsen som forsøkes forklart av score og kjønn. La responsen være y, score x_1 og kjønn x_2. Modellen vi bruker har med et interaksjonsledd, og ser slik ut:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

der ϵ_i er uavhengige og $\mathrm{N}(0, \sigma^2)$, $i = 1, \cdots, n$.

c) Spesifiser hvordan modellen over blir for hhv. kvinner og menn, og forklar hvordan parameterne kan brukes til å vurdere to typer lønnsdiskriminering: mulig forskjell i grunnleggende lønnstillegg, og mulig forskjell i uttelling for hvert poeng i arbeidsvurderingen (score).

d) Du finner den tilpassede regresjonsmodellen, basert på de $n = 19$ observasjonene, i utskriften fra R. Forklar på bakgrunn av de estimerte parameterverdiene og tilhørende P-verdier hvorfor kvinnene og advokatene deres har en god sak.

(Fortsettes på side 3.)

e) Konstruer et 99% konfidensintervall for interaksjonsparameteren β_3. Du kan hente noen av tallene du trenger fra utskriften. Forklar hvordan et slikt konfidensintervall skal tolkes.

f) Hvordan er 'Multiple R-squared' beregnet, og hvordan skal den tolkes? Hvordan vurderer du den tilpassede modellen i forhold til vedlagte plott av dataene? Hvilke andre plott ville du ha konstruert for å vurdere modellens egnethet?

Oppgave 3.

Du skal i denne oppgaven studere en enkel lineær regresjonsmodell som vi tvinger til å gå gjennom origo, dvs. at skjæringspunktet med y-aksen er satt lik null. La responsen være y og forklaringsvariabelen x. Vi har n observasjonspar (x_i, y_i). Modellen er da altså

$$y_i = \beta x_i + \epsilon_i, \qquad\qquad i = 1, \cdots, n,$$

der ϵ_i-ene er uavhengige og normalfordelte med forventning 0 og varians σ^2.

a) Finn minste kvadraters estimator (least squares estimator) $\hat{\beta}_{MK}$ for β. Vis at denne er forventningsrett og finn et uttrykk for estimatorens varians.

b) En alternativ estimator for β er $\hat{\beta}_A = \sum_i^n y_i / \sum_i^n x_i$. Vis at også denne er forventningsrett. Finn variansen til $\hat{\beta}_A$ og sammenlign med variansen til $\hat{\beta}_{MK}$. Hvilken estimator vil du bruke? Begrunn svaret.

Oppgave 4.

Antall tilfeller X av en sjelden, medfødt sykdom i Norge per år kan antas å følge en Poisson-fordeling med parameter λ, dvs. $X \sim \text{Poisson}(\lambda)$, slik at $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ for $k = 0, 1, 2, ...$

a) Forventet antall barn født med denne sykdommen har vært ett per år. Hvor mange tilfeller må man minst observere et gitt år for å kunne forkaste $\text{H}_0 : \lambda = 1$ til fordel for $\text{H}_a : \lambda > 1$ på nivå $\alpha = 0.05$?

b) Hvor sannsynlig er det at man med testen ovenfor oppdager at λ faktisk er økt, dersom den i virkeligheten er doblet i forhold til tidligere (dvs. $\lambda = 2$)? Hva er sannsynligheten for type-II-feil i denne situasjonen?

c) Forskere vil studere forekomsten av sykdommen nærmere. For å estimere sannsynligheten p for å få et barn med sykdommen, observerer man antall fødte barn n_i og antall tilfeller X_i for hvert år i, der $i = 1, \cdots, m$. Vi antar at sannsynligheten p er konstant i disse m årene. Dessuten er n_i-ene store og p liten, slik at vi kan anta at $X_i \sim \text{Poisson}(n_i p)$, $i = 1, \cdots, m$.

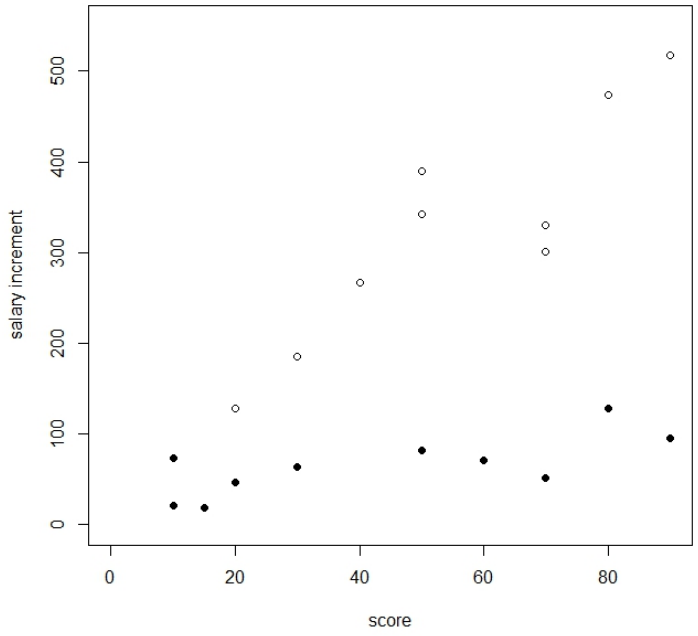
Finn sannsynlighetsmaksimerings-estimatoren (Maximum Likelihood Estimator) for p i denne situasjonen. Finn estimatorens forventning og varians.

(Fortsettes på side 4.)

Datasett oppgave 1:

	sex	score	salary_incr
1	1	10	21
2	1	90	96
3	1	20	47
4	1	80	128
5	1	30	64
6	1	70	52
7	1	10	73
8	1	15	19
9	1	50	82
10	1	60	71
11	0	20	128
12	0	80	474
13	0	50	342
14	0	70	330
15	0	30	185
16	0	70	301
17	0	40	267
18	0	90	517
19	0	50	390

Plott av data, sort punkt kvinner, hvitt punkt menn:



(Fortsettes på side 5.)

Utskrift fra R, oppgave 1:

```
> t.test(score[1:10],score[11:19],alternative="less", var.equal=T)

        Two Sample t-test

data:  score[1:10] and score[11:19]
t = -0.959, df = 17, p-value = 0.1755
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 9.812306
sample estimates:
mean of x mean of y
 43.50000  55.55556

> var.test(score[1:10],score[11:19])

        F test to compare two variances

data:  score[1:10] and score[11:19]
F = 1.6688, num df = 9, denom df = 8, p-value = 0.4822
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3830055 6.8455251
sample estimates:
ratio of variances
          1.66884

> fit=lm(salary_incr~score+sex+score*sex)
> summary(fit)

Call:
lm(formula = salary_incr ~ score + sex + score * sex)

Residuals:
    Min      1Q  Median      3Q     Max
-93.626 -21.684   0.266  29.609  90.394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.0553    40.7630   1.522 0.148723
score         4.7510     0.6815   6.972 4.49e-06 ***
sex         -31.1230    48.3228  -0.644 0.529259
score:sex    -3.9609     0.8437  -4.695 0.000288 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 45.32 on 15 degrees of freedom
Multiple R-squared: 0.9327,     Adjusted R-squared: 0.9192
F-statistic: 69.29 on 3 and 15 DF,  p-value: 5.101e-09
```

SLUTT