Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

Bokmål

Faglig kontakt under eksamen:
Gunnar Taraldsen 73 59 2641

EKSAMEN I EMNE ST1201/ST6201 STATISTISKE METODER

Onsdag 5. desember 2007
Tid: 09:00–13:00

Hjelpemidler: Alle trykte og håndskrevne hjelpemidler tillatt.
Alle kalkulatorer tillatt.

Sensur er ferdig: 21. desember 2007. Oppgaven består av 28 delpunkter som teller likt i sensuren. Oppgave 1 bør gjøres før Oppgave 2.

Oppgave 1 La $U_1, \ldots, U_n$ være et tilfeldig utvalg fra en sannsynlighetsfordeling som antas kjent og hvor $\mathrm{E}\, U_i = 0$, $\mathrm{Var}\, U_i = 1$. Anta at observasjonene i et eksperiment er modellert av $X_1, \ldots, X_n$ hvor $X_i = \mu + \sigma U_i$ og $n > 1$. La $\overline{X}$ og S^2 være empirisk middel og forventningsrett empirisk varians til observasjonene

a) Vis at $\mathrm{E}\, X_i = \mu$.

b) Vis at $\mathrm{Var}\, X_i = \sigma^2$.

c) Fordelingen til $W = \sqrt{n}(\overline{X} - \mu)/S$ avhenger ikke av modell parametrene (μ, σ^2). La w_α være slik at $\mathrm{P}(|W| > w_\alpha) = \alpha$. Utled et konfidensintervall for μ.

d) La $H_0 : \mu = \mu_0$. Utled en hypotesetest som har eksakt nivå α.

e) La $U \sim \mathrm{Uniform}(-\sqrt{3}, \sqrt{3})$. Vis at $\mathrm{E}\, U = 0$ og at $\mathrm{Var}\, U = 1$.

f) Vis at $X = \mu + \sigma U \sim \mathrm{Uniform}(a, b)$ med $a = \mu - \sqrt{3}\sigma$ og $b = \mu + \sqrt{3}\sigma$.

Oppgave 2 La $X_1, \ldots, X_n$ være et tilfeldig utvalg fra den uniforme fordelingen Uniform(a, b) med $n > 1$. Statistikken $T = (T_1, T_2) = (X_{(1)}, X_{(n)})$ gitt ved minste og største observerte verdi er en komplett og tilstrekkelig statistikk.

a) Bevis at T er tilstrekkelig ved faktorisering av rimelighetsfunksjonen L.

b) Vis at T er rimelighetsestimatoren for (a, b).

c) Vis at estimatorene $\hat{\mu} = (T_1 + T_2)/2$ og $\hat{\sigma} = (T_2 - T_1)/(2\sqrt{3})$ er rimelighetsestimatorene for $\mathrm{E}\, X_1$ og $\sqrt{\mathrm{Var}\, X_1}$. Hint: Oppgave 1 f).

d) Det kan vises at $\hat{\mu}$ er forventningsrett og at $\mathrm{Var}\, \hat{\mu} = (b - a)^2/[2(n + 1)(n + 2)]$. Vis at $\mathrm{Var}\, \hat{\mu} \leq \mathrm{Var}\, \overline{X}$ ved å beregne $\mathrm{Var}\, \overline{X}$.

e) Gi et generelt argument for at $\mathrm{Var}\, \hat{\mu} \leq \mathrm{Var}\, \overline{X}$ uten å beregne variansene som inngår.

f) Vis at fordelingen til $V = \sqrt{n}(\hat{\mu} - \mu)/\hat{\sigma}$ ikke avhenger av modell parametrene (a, b).

g) La v_α være slik at $\mathrm{P}(|V| > v_\alpha) = \alpha$. Finn formler for standard og utvidet feil når $\hat{\mu}$ brukes som estimator for μ.

Et eksperiment gir $(n, \overline{x}, s, x_{(1)}, x_{(n)}) = (100, 3.444, 2.498, -1.456, 7.756)$.

h) Estimer μ og beregn standard og utvidet feil ved metodene i Oppgave 1 og 2. Utvidet feil skal beregnes med utgangspunkt i $v_{5\%} = 0.53$, $w_{5\%} = 1.98$.

i) Diskuter kort fordeler og ulemper med metodene for estimering av μ gitt i Oppgave 1-2.

Oppgave 3 Storbonden Otte driver med poteter, og vil undersøke om de tre gjødselproduktene *New wave* (1), *Potato in the sky* (2), og *Natural earth* (3) er likeverdige. Ved å ta stikkprøver ved innhøstingen finnes verdien til en stokastisk variabel $X_{i,j}$ hvor $j = 1, 2, 3$ er produktnummer og i er nummeret til stikkprøven for det gitte gjødselproduktet. Variabelen selv tilsvarer netto gevinst per kvadratmeter, og har dermed enhet kroner$/m^2$.

a) Formuler en statistisk modell for å undersøke om forskjellen på gjødselproduktene er statistisk signifikant. Ta utgangspunkt i normalfordeling og randomisert blokk design.

b) Hva er null hypotesen H_0?

c) Gi en beskrivelse av en prosedyre for å utføre en test av hypotesen over.

d) Gi en begrunnelse for at randomisert blokk design er et naturlig valg for storbonden Otte.

e) Antagelsen om normalfordeling kan være vanskelig å begrunne. Formuler en alternativ ikke-parametrisk modell.

f) Gi en beskrivelse av en prosedyre for å utføre en test av hypotesen over for den alternative modellen.

Oppgave 4 Den kvinnelige befolkningen i Norge i årene 1900, 1930, 1960, 2007 var på 1143156, 1435860, 1789949, 2355346 personer (www.ssb.no). Ved bruk av 1900 som nullpunkt og avrunding til millioner finnes $\sum_i x_i y_i = 403.12$, $\sum_i x_i = 197$, $\sum_i y_i = 6.73$, $\sum_i x_i^2 = 15949$. Dette kan brukes i tallarbeidet under.

a) Bruk minste kvadrats metode til å finne en rett linje som modell for disse befolkningstallene.

b) Tegn grafen til linjen sammen med datapunktene.

c) Formuler en statistisk modell som begrunner bruk av minste kvadrats metode.

Den mannlige befolkningen i Norge i årene 1900, 1930, 1960, 2007 var på 1074815, 1363853, 1777758, 2325788 personer (www.ssb.no). Dette gir $\sum_i x_i y_i = 396.91$, $\sum_i x_i = 197$, $\sum_i y_i = 6.54$, $\sum_i x_i^2 = 15949$.

d) Bruk minste kvadrats metode på disse datapunktene også.

e) Tegn linjen og datapunktene i samme figur som for de kvinnelige befolkningstallene.

Stigningstallene til de to linjene er ikke like.

f) Formuler en null hypotese H_0 som kan brukes til å avgjøre om forskjellen er signifikant.

g) Formuler en prosedyre som kan brukes til å utføre en hypotesetest tilsvarende H_0 med et signifikansnivå $\alpha = 5\%$. Det er ikke meningen at du skal utføre denne hypotesetesten nå.

Noregs teknisk–
naturvitskaplege universitet
Institutt for matematiske fag

Nynorsk

Fagleg kontakt under eksamen:
Gunnar Taraldsen 73 59 26 41

EKSAMEN I EMNE ST1201/ST6201 STATISTISKE METODER

Onsdag 5. desember 2007
Tid: 09:00–13:00

Hjelpemidler: Alle trykte og handskrivne hjelpemiddel tillatne.
Alle kalkulatorar tillatne.

Sensur er ferdig seinast 21. desember 2007. Oppgåva består av 28 punkt som teller likt i sensuren. Gjer Oppgåve 1 føre Oppgåve 2.

Oppgåve 1 La $U_1, \ldots, U_n$ vere eit tilfeldig utval frå ein kjent sannsynsfordeling med $\mathrm{E}\, U_i = 0$, $\mathrm{Var}\, U_i = 1$. Anta at observasjonane i et eksperiment er modellert av $X_1, \ldots, X_n$ hvor $X_i = \mu + \sigma U_i$ og $n > 1$. La $\overline{X}$ og S^2 vere empirisk middel og forventningsrett empirisk varians til observasjonane

a) Vis at $\mathrm{E}\, X_i = \mu$.

b) Vis at $\mathrm{Var}\, X_i = \sigma^2$.

c) Fordelinga til $W = \sqrt{n}(\overline{X} - \mu)/S$ avhenger ikkje av modell parametrane (μ, σ^2). La w_α være slik at $\mathrm{P}(|W| > w_\alpha) = \alpha$. Utlei et konfidensintervall for μ.

d) La $H_0 : \mu = \mu_0$. Utlei ein hypotesetest som har eksakt nivå α.

e) La $U \sim \mathrm{Uniform}(-\sqrt{3}, \sqrt{3})$. Vis at $\mathrm{E}\, U = 0$ og at $\mathrm{Var}\, U = 1$.

f) Vis at $X = \mu + \sigma U \sim \mathrm{Uniform}(a, b)$ med $a = \mu - \sqrt{3}\sigma$ og $b = \mu + \sqrt{3}\sigma$.

Oppgåve 2 La $X_1, \ldots, X_n$ være et tilfeldig utval frå den uniforme fordelinga Uniform(a, b) med $n > 1$. Statistikken $T = (T_1, T_2) = (X_{(1)}, X_{(n)})$ gitt ved minste og største observerte verdi er ein komplett og tilstrekkelig statistikk.

a) Bevis at T er tilstrekkelig ved faktorisering av rimeleghetsfunksjonen L.

b) Vis at T er rimeleghetsestimatoren for (a, b).

c) Vis at estimatorene $\hat{\mu} = (T_1 + T_2)/2$ og $\hat{\sigma} = (T_2 - T_1)/(2\sqrt{3})$ er rimeleghetsestimatorene for $\mathrm{E}\, X_1$ og $\sqrt{\mathrm{Var}\, X_1}$. Hint: Oppgåve 1 f).

d) Det kan vises at $\hat{\mu}$ er forventningsrett og at $\mathrm{Var}\, \hat{\mu} = (b-a)^2/[2(n+1)(n+2)]$. Vis at $\mathrm{Var}\, \hat{\mu} \leq \mathrm{Var}\, \overline{X}$ ved å berekne $\mathrm{Var}\, \overline{X}$.

e) Gi et generelt argument for at $\mathrm{Var}\, \hat{\mu} \leq \mathrm{Var}\, \overline{X}$ utan å berekne variansane som inngår.

f) Vis at fordelinga til $V = \sqrt{n}(\hat{\mu} - \mu)/\hat{\sigma}$ ikkje avhenger av modell parametrane (a, b).

g) La v_α være slik at $\mathrm{P}(|V| > v_\alpha) = \alpha$. Finn formlar for standard og utvida feil når $\hat{\mu}$ brukast som estimator for μ.

Et eksperiment gir $(n, \overline{x}, s, x_{(1)}, x_{(n)}) = (100, 3.444, 2.498, -1.456, 7.756)$.

h) Estimer μ og berekn standard og utvida feil ved metodane i Oppgåve 1 og 2. Utvida feil skal bereknas med utgangspunkt i $v_{5\%} = 0.53$, $w_{5\%} = 1.98$.

i) Diskuter kort fordeler og ulemper med metodane for estimering av μ gitt i Oppgåve 1-2.

Oppgåve 3 Storbonden Otte driver med poteter, og vil undersøke om de tre gjødselprodukta *New wave* (1), *Potato in the sky* (2), og *Natural earth* (3) er likeverdige. Ved å ta stikkprøver ved innhøstinga finnes verdien til ein stokastisk variabel $X_{i,j}$ der $j = 1, 2, 3$ er produktnummer og i er nummeret til stikkprøven for det gitte gjødselproduktet. Variabelen sjølv tilsvarer netto gevinst per kvadratmeter, og har dermed eining kroner$/m^2$.

a) Formuler ein statistisk modell for å undersøke om forskjellen på gjødselprodukta er statistisk signifikant. Ta utgangspunkt i normalfordeling og randomisert blokk design.

b) Hva er null hypotesen H_0?

c) Gi ein skildring av ein prosedyre for å utføre ein test av hypotesen over.

d) Gi ein grunn for at randomisert blokk design er et naturlig val for storbonden Otte.

e) Trua om normalfordeling kan være vanskelig å grunngje. Formuler ein alternativ ikkje-parametrisk modell.

f) Gi ein beskriving av ein prosedyre for å utføre ein test av hypotesen over for den alternative modellen.

Oppgåve 4 Den kvinnelige befolkninga i Noreg i årene 1900, 1930, 1960, 2007 var på 1143156, 1435860, 1789949, 2355346 personar (www.ssb.no). Ved bruk av 1900 som nullpunkt og avrunding til millionar finnes $\sum_i x_i y_i = 403.12$, $\sum_i x_i = 197$, $\sum_i y_i = 6.73$, $\sum_i x_i^2 = 15949$. Dette kan brukast i tall arbeidet under.

a) Bruk minste kvadrats metode til å finne ein rett linje som modell for disse befolkningstala.

b) Teikn grafen til linja saman med data punkta.

c) Formuler ein statistisk modell som grunngje bruk av minste kvadrats metode.

Den mannlige befolkninga i Noreg i årene 1900, 1930, 1960, 2007 var på 1074815, 1363853, 1777758, 2325788 personar (www.ssb.no). Dette gir $\sum_i x_i y_i = 396.91$, $\sum_i x_i = 197$, $\sum_i y_i = 6.54$, $\sum_i x_i^2 = 15949$.

d) Bruk minste kvadrats metode på disse data punkta også.

e) Teikn linja og data punkta i same figur som for de kvinnelige befolkningstala.

Stigningstala til de to linjene er ikkje like.

f) Formuler ein null hypotese H_0 som kan brukast til å avgjøre om forskjellen er signifikant.

g) Formuler ein prosedyre som kan brukast til å utføre ein hypotesetest tilsvarande H_0 med et signifikansnivå $\alpha = 5\%$. Det er ikkje meiningen at du skal utføre denne hypotesetesten nå.

Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

English

Contact during exam:
Gunnar Taraldsen 73 59 26 41

EXAM IN ST1201/ST6201 STATISTICAL METHODS

Wednesday December 5th 2007
Time: 09:00–13:00

Aids: All kinds of typed material and hand written notes are allowed.
All calculators allowed.

Grading finished: December 21th 2007. The exam has 28 parts which have equal weight in the grading. Oppgave 1 should be done before Oppgave 2.

Oppgave 1 Let $U_1, \dots, U_n$ be a random sample from a probability distribution which is assumed known and where $\mathrm{E}\, U_i = 0$, $\mathrm{Var}\, U_i = 1$, and $n > 1$. Let $X_1, \dots, X_n$ represent the observations in an experiment, where $X_i = \mu + \sigma U_i$. As usual $\overline{X}$ and S^2 denote the empirical mean and unbiased empirical variance of the observations.

a) Show that $\mathrm{E}\, X_i = \mu$.

b) Show that $\mathrm{Var}\, X_i = \sigma^2$.

c) The distribution of $W = \sqrt{n}(\overline{X} - \mu)/S$ do not depend on μ and σ^2. Let w_α be such that $\mathrm{P}(|W| > w_\alpha) = \alpha$. Derive a formula for a confidence interval for μ.

d) Let $H_0 : \mu = \mu_0$. Derive a test with exact level α.

e) Let $U \sim \mathrm{Uniform}(-\sqrt{3}, \sqrt{3})$. Show that $\mathrm{E}\, U = 0$ and $\mathrm{Var}\, U = 1$.

f) Show that $X = \mu + \sigma U \sim \mathrm{Uniform}(a, b)$ with $a = \mu - \sqrt{3}\sigma$ and $b = \mu + \sqrt{3}\sigma$.

Oppgave 2

Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution Uniform(a, b) with $n > 1$. The statistic $T = (T_1, T_2) = (X_{(1)}, X_{(n)})$ given by the smallest and the largest observation is complete and sufficient.

a) Prove that T is sufficient by factorization of the likelihood L.

b) Show that T is the maximum likelihood estimator for (a, b).

c) Prove that the estimators $\hat{\mu} = (T_1 + T_2)/2$ and $\hat{\sigma} = (T_2 - T_1)/(2\sqrt{3})$ are the maximum likelihood estimators for $\mathrm{E}\, X_1$ and $\sqrt{\mathrm{Var}\, X_1}$. Hint: Oppgave 1 f).

d) It can be proven that $\hat{\mu}$ is unbiased and that $\mathrm{Var}\, \hat{\mu} = (b-a)^2/[2(n+1)(n+2)]$. Show that $\mathrm{Var}\, \hat{\mu} \leq \mathrm{Var}\, \overline{X}$ by calculation of $\mathrm{Var}\, \overline{X}$.

e) Give a general argument which proves $\mathrm{Var}\, \hat{\mu} \leq \mathrm{Var}\, \overline{X}$ without actual calculation.

f) Show that the distribution of $V = \sqrt{n}(\hat{\mu} - \mu)/\hat{\sigma}$ is not depending on the parameters a, b.

g) Let v_α be such that $\mathrm{P}(|V| > v_\alpha) = \alpha$. Find formulas for the standard error and the expanded error which can be used when $\hat{\mu}$ is used as an estimator for μ.

An experiment gives $(n, \overline{x}, s, x_{(1)}, x_{(n)}) = (100, 3.444, 2.498, -1.456, 7.756)$.

h) Estimate μ and calculate the standard error and the expanded error corresponding to the methods in Oppgave 1 and 2 using $v_{5\%} = 0.53$, $w_{5\%} = 1.98$.

i) Discuss briefly the advantages and disadvantages with the methods given by Oppgave 1 and 2 for the estimation of μ.

Oppgave 3 The farmer Otte has a potato field, and wishes to compare the three fertilizer products *New wave* (1), *Potato in the sky* (2), and *Natural earth* (3). Values of the random variable $X_{i,j}$ are obtained by taking random samples during the harvesting. The variable $X_{i,j}$ represents the profit per square meter (kroner/m^2), $j = 1, 2, 3$ indicates the product number, and the index i gives the sample number for the given product.

a) Formulate a statistical model which can be used to determine if there are significant differences between the products. Use a randomized block design based on the normal distribution.

b) Specify and describe the H_0 hypothesis.

c) Specify a statistical procedure which can be used to test H_0.

d) Give arguments which support the choice of using a randomized block design in this case.

e) It can be difficult to justify the assumption which involves the normal distribution. Formulate an alternative non-parametric model.

f) Specify a statistical procedure which can be used to test H_0 for the alternative model.

Oppgave 4 The female population in Norway for the years 1900, 1930, 1960, 2007 consisted of 1143156, 1435860, 1789949, 2355346 persons (www.ssb.no). The calculations below can be simplified by choosing 1900 as year 0 and by rounding to millions, which give $\sum_i x_i y_i = 403.12$, $\sum_i x_i = 197$, $\sum_i y_i = 6.73$, $\sum_i x_i^2 = 15949$.

a) Use the method of least squares to determine a line which models the above population data.

b) Plot the graph of the line together with the data points.

c) Formulate a statistical model which leads to the use of the method of least squares.

The male population in Norway for the years 1900, 1930, 1960, 2007 consisted of 1074815, 1363853, 1777758, 2325788 persons (www.ssb.no). This gives $\sum_i x_i y_i = 396.91$, $\sum_i x_i = 197$, $\sum_i y_i = 6.54$, $\sum_i x_i^2 = 15949$.

d) Use the method of least squares on this data set also.

e) Plot the graph of the line together with the data points in the figure which contains the female model.

The slopes of the two lines are not equal.

f) Formulate a hypothesis H_0 which can be used to decide whether the difference between the slopes are significant.

g) Specify a statistical procedure which can be used to test H_0 with a level $\alpha = 5\%$. You are not supposed to actually do the test.