UNIVERSITETET I OSLO
ØKONOMISK INSTITUTT

Eksamen i: ECON2130 – Statistikk 1
Exam: ECON2130 – Statistics 1

Eksamensdag: Torsdag 1. juni 2006
Date of exam: Thursday, June 1, 2006

Sensur kunngjøres: Fredag 16. juni
Grades will be given: Friday, June 16

Tid for eksamen: kl. 14:30 – 17:30
Time for exam: 02:30 p.m. – 05:30 p.m.

Oppgavesettet er på 5 sider
The problem set covers 5 pages

English version on page 3

Tillatte hjelpemidler:
- Alle trykte og skrevne hjelpemidler, samt lommekalkulator er tillatt

Resources allowed:
- *All printed and written resources, as well as calculator is allowed*

Eksamen blir vurdert etter ECTS-skalaen. A-F, der A er beste karakter og E er dårligste ståkarakter. F er ikke bestått.

The grades given: A-F, with A as the best and E as the weakest passing grade. F is fail.

Oppgave 1

I en by med 10 000 stemmeberettigete innbyggere pågår en heftig debatt om det skal etableres en bomring (dvs bommer på alle innfartsveiene der hver passering mot byen med bil koster en avgift) for å finansiere diverse trafikkprosjekter. På slutten av et debattprogram på lokal-TV ble seerne anmodet om å ringe inn (i løpet av programmet) og stemme for eller imot bomring. Av 5 000 innringere var 4 500 (altså 90%) imot bomring. For enkelhets skyld antar vi at tallet 10 000 bare omfatter byens stemmeberettigete og at bare stemmeberettigete ringte inn til programmet. Vi antar også at alle de stemmeberettigete har en mening for eller imot.

La M betegne motstander og T tilhenger av bomring. La S betegne det å delta i telefonavstemningen i TV og $\overline{S}$ det å ikke delta. Anta at de 10 000 stemmeberettigete er fordelt i de fire kombinasjonene av mening og stemmedeltagelse som angitt i tabell 1.

Tabell 1

	T (for)	M (mot)
S (avgir stemme)	500	4500
$\overline{S}$ (avgir ikke stemme)	2000	3000

a. For en tilfeldig valgt stemmeberettiget, la begivenhetene T, M og S stå for at vedkommende er tilhenger, motstander eller deltok i TV-avstemningen henholdsvis.

(i) Finn sannsynlighetene $P(M)$, $P(M \mid S)$, $P(M \mid \overline{S})$ og $P(M \cup S)$

(ii) Er M og S uavhengige? Begrunn svaret ditt.

b. Anta vi trekker to personer rent tilfeldig fra populasjonen av stemmeberettigete. La X være antall blant de to som er M (dvs. motstander av bomring). Anta X er binomisk fordelt.

(i) Begrunn antakelsen at X er binomisk fordelt – i hvert fall tilnærmet. Hva blir forventningen til X?

(ii) Beregn sannsynligheten for at de er enige (dvs. begge for eller begge imot bomring).

c. (*Mer krevende*). Tallene i første rad i **tabell 1** (500 og 4500) er kjente fra TV-avstemningen, som impliserer at $P(M \mid S) = 0{,}9$. Tallene i annen rad (2000 og 3000) er imidlertid bare gjetninger for regneeksemplets skyld. De er i virkeligheten ukjente. Sett $p = P(M)$ som er ukjent og av spesiell interesse. Hvilke tall måtte det ha stått i annen rad i tabell 1 (istedenfor 2000 og 3000) hvis TV-avstemningen hadde vært representativ i betydningen $p = P(M \mid S) = 0{,}9$?

Oppgave 2

For å måle graden av over- eller undervekt hos mennesker brukes ofte *kroppsmasseindeksen* (KMI), som definers som $(vekt)/(høyde)^2$ der vekt er målt i kilo (kg) og høyde i meter (m). Normalområdet for KMI regnes fra 20 til 25 kg/m^2. En KMI under 20 regnes som "undervekt" og en KMI over 25 som "overvekt".

I denne oppgaven betyr "ung kvinne" en kvinne i aldersgruppen 20 – 29 år.

a. La X være KMI for en tilfeldig valgt kvinne fra populasjonen av unge kvinner i Norge. Vi antar at X er normalfordelt med forventning $\mu = 22{,}4$ og standardavvik $\sigma = 3{,}4$ (dvs. $X \sim N(22{,}4\,,\ 3{,}4)$), der parameterverdiene er hentet fra en norsk kostholdsundersøkelse fra 1997.

(i) Vis at sannsynligheten for at en tilfeldig valgt ung kvinne er undervektig (KMI < 20), er 0,239.

(ii) Finn nedre desil (d_1) definert ved $P(X \le d_1) = 0{,}10$.

b. Betrakt et tilfeldig utvalg av 4 unge norske kvinner, og la X_i være KMI for kvinne nr. i i utvalget. Anta at X_1, X_2, X_3, X_4 er uavhengige og identisk fordelte med samme fordeling som X i punkt **a**..

(i) Finn $P(\overline{X} < 20)$ der $\overline{X} = \frac{1}{n}\sum_{i=1}^{4} X_i$.

(ii) Hva er sannsynligheten for at minst tre av de fire kvinnene i utvalget har KMI < 20?

c. En gruppe på 15 kvinnelige studenter ved en idrettshøyskole diskuterer om det er et urimelig slankepress på kvinnelige idrettsutøvere. Til å begynne med registrerte de sin egen KMI og fant gjennomsnittet for gruppen, $\bar{x} = 21,5$, og empirisk standardavvik, $s = 3,9$. Gruppen tar utgangspunkt i følgende modell:

$X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte med $X_i \sim N(\mu, \sigma)$, der n = 15, og der μ, σ begge antas ukjente. Her representerer X_i KMI for person nr. i i gruppen.

Beregn et 95% konfidensintervall for μ under disse betingelsene.
Drøft kort rimeligheten av modellen og eventuelt tilleggsforutsetninger som trengs for å kunne tolke μ som gjennomsnittlig KMI i populasjonen av yngre kvinnelige idrettsutøvere i Norge.

d. Gruppen ønsker å teste $H_0: \ \mu \geq \mu_0$ mot $H_1: \ \mu < \mu_0$, der $\mu_0 = 22,4$ som er brukt i punkt **a**. I håp om å øke utsagnskraften finner gruppen å kunne forutsette at σ er kjent lik 3,4. Forøvrig er forutsetningene som i punkt **c**. Sett opp en test med signifikansnivå 5% som utnytter at σ er kjent. Gjennomfør testen med dataene gitt i punkt **c.**, og formuler en konklusjon. Beregn også testens p-verdi.

e. Gitt testen i punkt **d**., der antall observasjoner er 15. Hva blir sannsynligheten for å begå feil av type I og sannsynligheten for å begå feil av type II hvis $\mu = \mu_0 = 22,4$? Hva blir de to feilsannsynlighetene hvis $\mu = 21,5$?

f. La nå antall observasjoner, *n*, være vilkårlig. For øvrig la modell (med kjent σ) og hypoteser være som i punkt **d**. Sett opp testobservator og forkastningområde når *n* er vilkårlig, der signifikansnivået fortsatt er 5%. Sett også opp et uttrykk for styrkefunksjonen, $\gamma(\mu)$. Hvor stor må *n* være for at styrkefunksjonen for testen skal ha en verdi på minst 0,95 hvis den sanne verdien av μ er 21,5?

ENGLISH VERSION

Problem 1

In a town containing 10 000 inhabitants with a right to vote there is an intensive debate going on concerning the possibility of establishing a “toll ring” (i.e., toll stations on all roads leading to town where each passing by car towards town requires a fee to be paid), in order to finance various traffic projects. In a debate program on local TV the viewers were asked to phone in (during the program) their opinion for or against the establishment of the toll ring. This resulted in 5 000 calls of which 4 500 (i.e., 90%) were against toll ring. For simplicity we assume that the number of 10 000 inhabitants in the town only include those that have a right to vote, and that all calls to the TV station were from people with a right to vote. In addition we assume that all inhabitants with a right to vote have an opinion for or against toll ring.

Let *M* denote opposition against and *T* support for the toll ring. Let *S* denote participation in the phone voting in the TV program and $\bar{S}$ non-participation. Assume that the 10 000 inhabitants with a right to vote fall into the four combinations of opinion and TV-vote behavior as given in table 1:

Table 1

	T (for)	M (against)
S (voter)	500	4500
$\bar{S}$ (non-voter)	2000	3000

a. For a randomly chosen person with a right to vote from the town, let the events *T*, *M* and *S* denote that he or she is a supporter, an opponent or a participant of the TV vote respectively.

(iii) Find the probabilities $P(M)$, $P(M \mid S)$, $P(M \mid \bar{S})$ and $P(M \cup S)$

(iv) Are *M* and *S* independent? Give a reason for your answer.

b. Suppose that two people are drawn at random from the population of inhabitants with a right to vote. Let *X* be the number among the two who are *M* (i.e., against toll ring). Suppose that *X* is binomially distributed.

(i) Discuss the assumption that X is binomially distributed – at least approximately. What is the expectation of *X*?
(ii) Calculate the probability that they agree (i.e., that both of them are for or both are against the toll ring).

c. (*More demanding*). The numbers in the first row of **table 1** (500 and 4500) are known from the TV voting, which implies that $P(M \mid S) = 0{,}9$. The numbers in the second row (2000 and 3000), however, are only guesses for the sake of illustration. They are in fact unknown. Put $p = P(M)$ which is unknown and of special interest. Which numbers would have to be in the second row in table 1 (instead of 2000 and 3000) if the TV-vote were representative in the sense that $p = P(M \mid S) = 0{,}9$?

Problem 2

The *body mass index* (BMI) is often used to measure the degree of over- or underweight of human beings. It is defined as $(weight)/(height)^2$ where weight is measured in kilo (kg) and height in meter (m). The normal interval for BMI is usually taken to extend from 20 to 25 kg/m^2. A BMI below 20 is counted as "underweight" and a BMI over 25 as "overweight".

In this problem the term "young woman" means a woman belonging to the age group 20 – 29 years.

a. Let *X* be the BMI for a randomly chosen woman from the population of young women in Norway. We assume that X is normally distributed with expectation $\mu = 22{,}4$ and standard deviation $\sigma = 3{,}4$ (i.e.. $X \sim N(22{,}4\,,\ 3{,}4)$). The parameter values are taken from a Norwegian study from 1997.

(**i**) Show that the probability that a randomly chosen young woman is underweight (BMI < 20), is 0,239.
(**ii**) Find the lower decile (d_1) defined by $P(X \le d_1) = 0{,}10$.

b. Let X_i be the BMI of woman no. i in a random sample of 4 young Norwegian women. Suppose that X_1, X_2, X_3, X_4 are independent and identically distributed with the same distribution as X in section **a**.

(**i**) Find $P(\overline{X} < 20)$ where $\overline{X} = \frac{1}{n}\sum_{i=1}^{4} X_i$.

(**ii**) What is the probability that at least three of the four women in the sample have BMI < 20?

c. A group consisting of 15 female students from a high school of athletics are discussing whether there is an unreasonable pressure on female athletes for slimming. Initially they all measured their own BMI and found that the mean for the group was $\overline{x} = 21{,}5$, and sample standard deviation, $s = 3{,}9$. The group decides to base their statistical analysis on the following model:

$X_1, X_2, \ldots, X_n$ are independent and identically distributed with $X_i \sim N(\mu, \sigma)$, where n = 15, and where μ, σ both are assumed to be unknown. Here X_i represents the BMI for person no. i in the group.

Calculate a 95% confidence interval for μ under these conditions.

Discuss briefly the reasonableness of the model and possible extra assumptions that are needed in order to be able to interpret μ as the mean BMI in the population of young female athletes in Norway.

d. The group wishes to test $H_0: \ \mu \ge \mu_0$ against $H_1: \ \mu < \mu_0$ where $\mu_0 = 22{,}4$ (referred to in section **a)**. Hoping to increase the strength of the information in the data, the group finds reason to assume that σ is known and equal to 3,4. Otherwise the assumptions are as in section **c**. Set up a test for the given hypotheses that utilizes σ being known, and using the level of significance 5%. Carry out the test using the data given in section **c**. State a conclusion. Finally, calculate the p-value of the test.

e. Consider the test in section **d.**, where the number of observations is 15. What is the probability of making an error of type I and what is the probability of making an error of type II if $\mu = \mu_0 = 22{,}4$? What are the two error probabilities if $\mu = 21{,}5$?

f. Now, let the number of observations, n, be arbitrary. Otherwise let the model (with known σ) and hypotheses be as in section **d**. Set up the test statistic and the rejection region when n is arbitrary, where the level of significance is 5%. Also set up an expression for the power function, $\gamma(\mu)$. How large must n be in order that the power function of the test attains a value of at least 0,95 if the true value of μ is 21,5?