EXAM IN 5B1501 SANNOLIKHETSTEORI OCH STATISTIK I
TUESDAY DECEMBER 20, 2005 08.00AM–01.00PM.

Examiner: Dan Mattsson, tel. 790 7135.

Permissible aids: Formel- och tabellsamling i Matematisk statistik. Calculator.

Notation must be explained and defined. Arguments and calculations must be so detailed that they are easy to follow. Numerical answers must be given with 2 correct decimals. The exam consists of 6 problems. A correct solution yields 10 points. A student with at least 24 points is guaranteed to pass. Students with 22–23 points have the opportunity to complement the exam.

The results will be posted (at the latest) on Friday the 13th of January, 2006, on the Mathematical Statistics bulletin board Lindstedtsvägen 25, straight ahead from the entrance.

The corrected exam will be available at ”elevexpeditionen” up to seven weeks after the date of the exam.

Uppgift 1

A resistor has a nominal resistance of 100Ω and tolerance 10%. Hence, assume that the resistance R is a random variable, uniformly distributed on the interval $[90,\ 110]$ Ohm. A voltage of 5V is applies over the resistor and the current through the resistor is, by Ohm's law, $I = 5/R$.

a) Calculate $E(I)$ and $D(I)$. (4 p)

b) A circuit consists of $n = 100$ independent resistors as above, connected in parallel, with a voltage of 5V over the circuit. Calculate (approximately) the probability that the total current through the circuit exceeds 5.1 Ampere. (The total current is the sum of the currents in every branch of the parallel connection.) (6 p)

Those who haven't solved a) may use the (incorrect) values $E(I) = 0.05$ and $D(I) = 0.005$ Ampere.

Uppgift 2

Bits (zeros and ones) may be transmitted erroneously over a noisy channel. Suppose that a bit (a zero or one) is erroneously transmitted with probability p, independently of other bits. For channel A is the probability of error $p = 0.005$ and for channel B is $p = 0.015$.
A bit stream of $n = 1000$ bits were received of which $x = 10$ bits were erroneously transmitted.

a) Estimate, using the method of maximum likelihood, which channel (A or B) was used in the transmission. (5 p)

b) Suppose that 75% of all bit streams is carried by channel A, the rest by channel B. Using the notation

$$A = \{\text{Channel A used}\} \quad \text{and} \quad B = \{\text{Channel B används}\} = A^*$$

then $P(A) = 0.75$ and $P(B) = 0.25$.
Calculate $P(A|10 \text{ errors among } 1000 \text{ bits})$ and $P(B|10 \text{ errors among } 1000 \text{ bits})$ and estimate the channel used for transmission, using these conditional probabilities. (5 p)

Hint: The estimate is a so-called *Bayes*-estimate.

Uppgift 3

Micro circuits are delivered in lots of size $n = 1000$ units. Assume that every circuit is defect with probability $p = 0.015$, independently of other circuits. A whole circuit generates a net profit of 2 SEK when traded, a defective circuit generates a net loss of 20 SEK.

a) Let Y describe the total profit when the lot is traded. Determine $E(Y)$ and $D(Y)$.

b) Determine (approximately) the probability that the total profit is less than 1500 SEK, i.e., $P(Y < 1500)$.

Uppgift 4

In order to predict the sheer strength of steel girders two different methods are used, here denoted method 1 and method 2. In a comparative study the quotient between the predicted value and the real value was calculated for 5 different girders with the two methods. The values obtained were as follows:

Girder no	1	2	3	4	5
Method 1	1.19	1.15	1.32	1.34	1.20
Method 2	1.06	0.99	1.06	1.06	1.07

Result k, $k = 1, \ldots, n$, for method 1 is seen as an observation from $\mathrm{N}(\mu_k + \Delta, \sigma_1)$ and result k from method 2 as an observation from $\mathrm{N}(\mu_k, \sigma_2)$. All random variables are assumed to be independent.

a) Calculate a 95% confidence interval for the systematic difference Δ. (7 p)

b) Test the hypothesis $H_0 : \Delta = 0.25$ against the hypothesis $H_1 : \Delta \neq 0.25$ on the level 5%. It should be clearly stated if H_0 is rejected or not. (3 p)

Uppgift 5

In an opinion poll, one wanted to investigate the preferences for competing brands A and B for different consumer populations.
100 consumers in the age-span 18–30 years tried both products A and B and said whether they preferred A or B or if they did not see any difference between them. The results are summarised in the table

	Prefer A	Prefer B	No difference	
Age-span 18–30 years	23	47	30	Total: 100

The same investigation was performed on 50 consumers in the age span 30–65 years. The results are summarised by

	Prefer A	Prefer B	No difference	
Age-span 30–65 years	16	13	21	Total: 50

Test, using an appropriate method, if the is any difference in preferences between the two groups of consumers. Use significance level $\approx$10%.
State your hypothesis clearly, and indicate whether the null hypothesis should be rejected or not. (10 p)

Uppgift 6

In a paper was the connection between pressure sustainability x and permeability y for different mixtures of concrete presented.
The measurements for $n = 14$ pairs (x_i, y_i) of data are summarised by:

$$\sum_{i=1}^{14} y_i = 572 \qquad \sum_{i=1}^{14} y_i^2 = 23530$$

$$\sum_{i=1}^{14} x_i = 43 \qquad \sum_{i=1}^{14} x_i^2 = 157.42$$

$$\sum_{i=1}^{14} x_i y_i = 1697.80.$$

Suppose that (x_1, y_1),...,(x_{14}, y_{14}) are given by a linear regression model, i.e., are samples from (x_1, Y_1),...,(x_{14}, Y_{14}), where Y_i follows a normal distribution $N(\alpha + \beta x_i, \sigma)$.

a) Calculate the estimates α^*_{obs}, β^*_{obs} of the coefficients in the regression line. (3 p)

b) Use the regression line to estimate the expected permeability when the pressure sustainability is $x = 4.3$. (2 p)

d) Construct a confidence intervals for α and β with confidence level 95%. (5 p)

Avd. Matematisk statistik

KTH Matematik

LÖSNINGAR TILL
TENTAMEN I 5B1501 SANNOLIKHETSTEORI OCH STATISTIK I
TISDAGEN DEN 20 DECEMBER 2005 KL 08.00–13.00.

Uppgift 1

Om U är $\mathrm{U}(90, 110)$ så är $f_U(x) = \frac{1}{20}$ för $90 \leq x \leq 110$. Med $I = 5/U$ så är

$$\begin{aligned} E(I) &= E\left(\frac{5}{U}\right) = \int_{-\infty}^{\infty} \frac{5}{x} f_U(x)\,dx = \int_{90}^{110} \frac{5}{x} \cdot \frac{1}{20}\,dx = \frac{5}{20}\left[\ln(x)\right]_{90}^{110} \\ &= \underline{\frac{1}{4}\ln(11/9) \approx 0.050168} \end{aligned}$$

och

$$\begin{aligned} E\left(I^2\right) &= E\left((5/U)^2\right) = \int_{-\infty}^{\infty} \frac{25}{x^2} f_U(x)\,dx = \int_{90}^{110} \frac{25}{x^2} \cdot \frac{1}{20}\,dx = \frac{25}{20}\left[-\frac{1}{x}\right]_{90}^{110} \\ &= \frac{5}{4}\left(\frac{1}{90} - \frac{1}{110}\right) = \frac{1}{396} \approx 0.0025253. \end{aligned}$$

Alltså är

$$V(I) = E\left(I^2\right) - (E(I))^2 = \frac{1}{396} - \left(\frac{1}{4}\ln(11/9)\right)^2 \approx 8.457 \cdot 10^{-6}$$

och $D(I) = \sqrt{V(I)} \approx \underline{0.0029081}$.
b) Den totala strömmen $I = I_1 + I_2 + \cdots + I_n$, $n = 100$, där $I_1, \ldots, I_n$ är oberoende och likafördelade. Enligt CGS är I approximativt $\mathrm{N}(\mu, \sigma) = \mathrm{N}(100 \cdot 0.0502, \sqrt{100} \cdot 0.0025) = \mathrm{N}(5.02, 0.0291)$. Alltså är

$$P(I > 5.1) = P\left(\frac{I - \mu}{\sigma} > \frac{5.1 - \mu}{\sigma}\right) \approx 1 - \Phi(2.862) \approx 1 - 0.9979 = \underline{0.0021}.$$

(Med $E(I) = 0.05$ och $D(I) = 0.005$ fås svaret 0.0228.)

Uppgift 2

Antalet felaktigt överförda bitar, $x = 10$, är ett utfall av en binomialfördelad stokastisk variabel X, X är $\mathrm{Bin}(n, p) = \mathrm{Bin}(1000, p)$.
Likelihoodfunktionen blir

$$\begin{aligned} L(p) &= P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \\ &= \begin{cases} \binom{1000}{10} 0.005^{10}(1 - 0.005)^{990} = 0.018 & \text{om } p = 0.005 \\ \binom{1000}{10} 0.015^{10}(1 - 0.015)^{990} = 0.048 & \text{om } p = 0.015 \end{cases} \end{aligned}$$

Då $L(0.015) > L(0.005)$ är ML-skattningen att <u>kanal B användes</u>. (Man kan använda Poissonapproximation och tabell 5 för att bestämma sannolikheterna. Då får man att $L(0.015) = 0.049$ och $L(0.005) = 0.018$.)
b) Med Bayes formel fås

$$\begin{aligned} P(A|X=10) &= \frac{P(A \cap \{X=10\})}{P(X=10)} = \frac{P(X=10|A)\,P(A)}{P(X=10|A)\,P(A) + P(X=10|B)\,P(B)} \\ &= \frac{L(0.005)P(A)}{L(0.005)P(A) + L(0.015)P(B)} = 0.528 \end{aligned}$$

och $P(B|X=10) = 1 - P(A|X=10) = 0.472$. Här är $P(A|X=10) > P(B|X=10)$ så skattningen blir att <u>kanal A användes</u>.

Uppgift 3

Låt X beskriva antalet defekta kretsar i ett parti om $n = 1000$ enheter. Med modellen att X är $\text{Bin}(n, p) = \text{Bin}(1000, 0.015)$ så beskrivs nettovinsten av

$$Y = \text{vinst} - \text{förlust} = 2 \cdot (n - X) - 20 \cdot X = 2n - 22X$$

Då är

$$E(Y) = E(2n - 22X) = 2n - 22E(X) = 2n - 22np = \underline{1670 \text{ kronor}}$$

och

$$V(Y) = V(2n - 22X) = (-22)^2 V(X) = 484 \cdot np(1-p) = 7151.1$$

så $D(Y) = \sqrt{V(Y)} = \underline{84.564 \text{ kronor}}$.
b) Eftersom $V(X) = np(1-p) = 14.775 > 10$ är X approximativt normalfördelad och så även Y, dvs Y är approximativt $\text{N}(\mu, \sigma) = \text{N}(1670,\ 84.564)$. Nu är

$$\begin{aligned} P(Y < 1500) &= P\left(\frac{Y-\mu}{\sigma} < \frac{1500-\mu}{\sigma}\right) \approx \Phi(-2.01) = 1 - \Phi(2.01) = 1 - 0.9778 \\ &= \underline{0.0222}. \end{aligned}$$

Uppgift 4

De parvisa skillnaderna

Stålbalk nr	1	2	3	4	5
y_i : (Metod 1 – Metod 2)	0.13	0.16	0.26	0.28	0.13

är utfall av $\text{N}(\Delta, \sigma)$-fördelade stokastiska variabler, där Δ skattas med $\overline{y} = 0.192$ och σ med $s = 0.072595$. Då $\overline{y}$ är ett utfall av $\overline{Y}$, $\overline{Y}$ är $\text{N}(\Delta, \sigma/\sqrt{n})$, så är

$$\frac{\overline{Y} - \Delta}{S/\sqrt{n}} \quad t(n-1)\text{-fördelad.}$$

Ur $t(n-1) = t(4)$-tabell fås $t_{0.025} = 2.78$ så ett 95% konfidensintervall för Δ blir

$$\Delta \in \overline{y} \pm t_{0.025} \frac{s}{\sqrt{n}} = 0.192 \pm 2.78 \cdot 0.0325 = \underline{0.192 \pm 0.090} = \underline{(0.10,\ 0.28)}.$$

b) Eftersom $\Delta = 0.25$ inte är ett orimligt värde på Δ enligt konfidensintervallet kan H_0 inte förkastas på nivå 5%.

Uppgift 5

Homogenitetstest. Nollhypotesen är att det inte är någon skillnad i preferenser mellan de två åldersgrupperna. Observationerna sammanfattas i tabellen

x_{ij}	Föredrar A	Föredrar B	Ingen skillnad	
Åldersgrupp 18–30 år	23	47	30	$n_1 = 100$
Åldersgrupp 30–65 år	16	13	21	$n_2 = 50$
Totalt:	$m_1 = 39$	$m_2 = 60$	$m_3 = 51$	$N = 150$

Vi förkastar nollhypotesen för stora värden på

$$q = \sum_{i,j} \frac{(x_{ij} - \frac{m_i n_j}{N})^2}{\frac{m_i n_j}{N}} = 6.1252$$

som om nollhypotesen är sann är ett utfall från en (approximativt) $\chi^2((3-1)(2-1)) = \chi^2(2)$-fördelad stokastisk variabel. (Approximationen ok ty $n_i m_j / N > 5$ för alla i, j.) Ur χ^2-tabell fås att $\chi^2_{0.10} = 4.61 < q$ så hypotesen om en gemensam preferens förkastas på nivå 10%. (Hypotesen förkastas även på nivå 5%).

Uppgift 6

a) Vi får $\bar{x} = 43/14$ samt $\bar{y} = 572/14$. Vidare är (med beteckningar från formelsamlingen)

$$S_{xx} = \sum_1^{14} x_i^2 - 14(\bar{x})^2 = 157.42 - 14 \cdot (43/14)^2 = 25.3486$$

$$S_{xy} = \sum_1^{14} x_i y_i - 14\bar{x}\bar{y} = 1697.80 - 14 \cdot (43/14) \cdot (572/14) = -59.0571$$

$$S_{yy} = \sum_1^{14} y_i^2 - 14(\bar{y})^2 = 23530 - 14 \cdot (572/14)^2 = 159.7143$$

Detta ger

$$\beta^*_{\text{obs}} = S_{xy}/S_{xx} = -2.3298,\ \alpha^*_{\text{obs}} = \bar{y} - \beta^*_{\text{obs}}\bar{x} = 572/14 - (-2.3298) \cdot (43/14) = 48.0130.$$

Skattad linje är alltså $y = 48.0130 - 2.3298x$.
b) Vi skattar med $\alpha^*_{\text{obs}} - \beta^*_{\text{obs}} 4.3 = 48.0130 - 2.3298 \cdot 4.3 = 37.9949$.
c) Vi skattar σ med s där

$$(14-2)s^2 = S_{yy} - (\beta^*_{\text{obs}})^2 S_{xx} = 22.1229$$

som ger $s = \sqrt{22.1229/12} = 1.3578$. Vi får konfidensintervallen

$$\begin{aligned} I_\alpha &= \alpha^*_{\text{obs}} \pm t_{0.025}(12) s \sqrt{\frac{1}{14} + \frac{(\bar{x})^2}{S_{xx}}} = 48.0130 \pm 2.18 \cdot 1.3578 \sqrt{\frac{1}{14} + \frac{(43/14)^2}{25.3486}} \\ &= 48.0130 \pm 1.9714, \\ I_\beta &= \beta^*_{\text{obs}} \pm t_{0.025}(12) \frac{s}{\sqrt{S_{xx}}} = -2.3298 \pm 0.5879. \end{aligned}$$