Avd. Matematisk statistik

KTH Matematik

EXAM IN 5B1501 PROBABILITY THEORY AND STATISTICS
TUESDAY DECEMBER 19, 2006, 2PM–7PM.

Examiner: Dan Mattsson, tel. 790 7135.

Permitted aids: List of formulae and tables in mathematical statistics. Calculator.

Notation must be explained and defined. The calculations and reasoning must be so explicit that they can be easily followed. Numerical answers must be given with at least two valid decimals. The exam consists of 6 problems. A correct solution of a problem yields 10 points. Students with at least 24 points are guaranteed to pass. Students with 22–23 points have the opportunity to complement the exam. Students who passed quiz 1 at November 20 are credited **problem 1a**, and those who passed quiz 2 at December 6 are credited **problem 2b**.

The result is reported to Ladok Tuesday, January 16, 2007 the latest, and is thenceforth available via "Mina sidor".

Uppgift 1

The probability that the amount of aflatoxin in 10 grams of peanuts exceed t micro grams is

$$\frac{1}{(1+t)^3} \quad t \geq 0.$$

a) *Not available for those that passed quiz 1.*
Let X describe the amount of aflatoxin (in micro grams) in 10 grams of peanuts and determine $E(X)$ and $D(X)$.

b) 2 kg of peanuts is required to produce a quantity of peanut butter. Determine a level a such that the probability that the produced peanut butter contains at most a micro grams aflatoxin is $\approx 99\%$. (5 p)

Those who have not solved a) may use $E(X) = 0.6$ and $D(X) = 0.9$ micro grams.

Uppgift 2

Suppose that the number of stormy days in December is described by a Poisson random variable with expectation μ and that the number of stormy days different years are independent random variables.

a) Assume that $\mu = 2$. Determine approximately the probability that the difference in the average number of stormy days in December during two different (non overlapping) 10-year periods exceeds 2 days. (5 p)

b) *Not available for those that passed quiz 2.*

Suppose that during a 10-year period one notes the following number of stormy days in December:

$$2 \quad 0 \quad 1 \quad 2 \quad 0 \quad 2 \quad 4 \quad 3 \quad 2 \quad 1$$

Calculate an approximately 95% confidence interval for μ and test the hypothesis $\mu = 2$ against $\mu \neq 2$ on approximately 5% level of significance. (5 p)

Uppgift 3

The probability that a machine manufactures a defunct unit varies with time and is described by a random variable X with distribution:

$$p_X(0) = 0.10 \quad p_X(0.1) = 0.60 \quad p_X(0.2) = 0.25 \quad p_X(0.3) = 0.05.$$

a) At a certain point in time a manufactured unit is taken. Calculate the probability that it is defunct. (5 p)

b) At a certain point in time 3 units are selected at random. These are independent of each other but are defunct with the same (random) probability. Calculate the probability that at least two of the three selected units are defunct. (5 p)

Uppgift 4

One wants to investigate the change in travelling times when the traffic was rerouted. 12 subjects were to register their travelling times to work a certain date before the experiment, and also the corresponding times a certain date after the rerouting.
The result is:

Person no	1	2	3	4	5	6	7	8	9	10	11	12	
Before	20	25	26	22	24	48	52	27	18	12	28	31	(min)
After	18	23	26	18	26	44	51	26	22	11	29	31	(min)

Suppose that the observations are independent and that the differences in travelling time for every person are described by the same normally distributed random variables.
a) Calculate a 95% confidence intervall for the average loss in travel time, Δ.
(7 p)

b) Before the experiment one thought that the rerouting would yield an travelling time on the average 3 minutes shorter. Perform a statistical test on the 5% level of the hypothesis

$$H_0 : \Delta = 3 \text{ (min)}$$

mot

$$H_1 : \Delta \neq 3 \text{ (min)}.$$

You conclusions should clearly state whether the hypothesis H_0 is rejected or not. (3 p)

Uppgift 5

Assume that the surface area of a sphere is described by an exponentially distributed random variable X with probability density

$$f_X(t) = \frac{1}{\theta} \mathrm{e}^{-t/\theta}, \quad t \geq 0$$

where $\theta > 0$ is the expectation, $\theta = E(X)$. To estimate the area θ, n measurements $y_1, \ldots, y_n$ of the sphere's diameter are made. These are modelled as a random sample of independent and identically distributed random variables (The relationship between area and diameter is $X = \pi Y^2$.)

a) Show that the probability density function for Y is

$$f_Y(t) = \frac{2\pi t}{\theta} \mathrm{e}^{-\pi t^2/\theta}, \quad t \geq 0.$$

(4 p)

b) Based on observations $y_1, \ldots, y_n$, find an expression for the Maximum likelihood estimate of θ. Calculate the estimate numerically for the values $y_1 = 0.20$, $y_2 = 0.25$ and $y_3 = 0.18$ centimetres. (4 p)

c) Is the estimate of θ unbiased (väntevärdesriktig)? (2 p)

Uppgift 6

A group of teachers in mathematics at a larger technical university has developed a new diagnostic test in mathematics which the freshmen must take. The teachers think that 40% of the students taking the test will not pass, 40% will receive grade 3 and the rest will receive grade 4 or higher.
In a random sample of 200 students, 60 freshmen received grade 3 and 45 freshmen received grade 4 or 5, the others did not pass.
Use a χ^2-test to determine whether the observations are consistent with the assumed theoretical distribution of grades. Use approximately 5% level of significance. (10 p)

Avd. Matematisk statistik

KTH Matematik

LÖSNINGAR TILL
TENTAMEN I 5B1501 SANNOLIKHETSTEORI OCH STATISTIK I
TISDAGEN DEN 19 DECEMBER 2006 KL 14.00–19.00.

Uppgift 1

Enligt uppgiftens lydelse är fördelningsfunktionen för X:

$$F_X(t) = P(X \leq t) = 1 - P(X > t) = 1 - \frac{1}{(1+t)^3} \quad t \geq 0$$

dvs täthetsfunktionen är

$$f_X(t) = \frac{d}{dt} F_X(t) = \frac{3}{(1+t)^4} \quad t \geq 0.$$

Väntevärdet beräknas till

$$E(X) = \int x\, f_X(x)\, dx = \int_0^\infty x \frac{3}{(1+x)^4}\, dx = \int_1^\infty (t-1)\frac{3}{t^4}\, dx = \left[-\frac{3}{2t^2} + \frac{3}{3t^3}\right]_1^\infty = \frac{1}{2} = \underline{0.5}.$$

Vidare så är

$$\begin{aligned} E\left(X^2\right) &= \int x^2\, f_X(x)\, dx = \int_0^\infty x^2 \frac{3}{(1+x)^4}\, dx = \int_1^\infty (t-1)^2 \frac{3}{t^4}\, dx \\ &= 3\left[-\frac{1}{t} + \frac{2}{2t} - \frac{1}{3t^2}\right]_1^\infty = 1 \end{aligned}$$

så $V(X) = E(X^2) - (E(X))^2 = 1 - (1/2)^2 = 3/4$ och $D(X) = \underline{\sqrt{3}/2} \approx \underline{0.86}$.
b) Den totala mängden aflatoxin i jordnötssmöret $Y = X_1 + X_2 + \cdots + X_n$, $n = 200$, är enligt Centrala gränsvärdessatsen en approximativt normalfördelad stokastisk variabel med väntevärde $\mu = 200 \cdot 0.5 = 100$ och standardavvikelse $\sigma = \sqrt{200 \cdot 3/4} = \sqrt{150}$. Vi söker a så att

$$0.99 = P(Y \leq a) = P\left(\frac{Y-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) \approx \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Men $\Phi(\lambda_{0.01}) = 0.99$ så

$$\frac{a-\mu}{\sigma} = \lambda_{0.01} \quad \Rightarrow \quad a = \mu + \lambda_{0.01}\sigma = 100 + 2.3263 \cdot \sqrt{150} = \underline{128 \text{ mikrogram}}.$$

(Med valet $E(X) = 0.6$ och $D(X) = 0.9$ fås svaret 150 mikrogram.)

Uppgift 2

Låt $X_1, \ldots, X_n$ beskriva antalet stormar under $n = 10$ år. Då är $X_1, \ldots, X_n$ oberoende Po(2)-fördelade och $\sum_{i=1}^{10} X_i$ är Po(20) $\approx$ N$(20, \sqrt{20})$-fördelad ty $20 \geq 15$. Alltså är $\overline{X}$

approximativt $\mathrm{N}(2, \sqrt{0.2})$. På motsvarande sätt är medelantalet stormar under en annan tidsperiod $\overline{Y}$ approximativt $\mathrm{N}(2, \sqrt{0.2})$ oberoende av $\overline{X}$.
a) Alltså är $\overline{Y} - \overline{X}$ approximativt $\mathrm{N}(0, \sqrt{0.4})$ och

$$\begin{aligned} P\left(|\overline{Y} - \overline{X}| > 2\right) &= P\left(\overline{Y} - \overline{X} > 2\right) + P\left(\overline{Y} - \overline{X} < -2\right) \\ &\approx (1 - \Phi\left(\frac{2-0}{\sqrt{0.4}}\right)) + \Phi\left(\frac{-2-0}{\sqrt{0.4}}\right) = 1 - \Phi(3.16) + \Phi(-3.16) \\ &= 2(1 - \Phi(3.16)) = \underline{0.0015}. \end{aligned}$$

b) Här är skattas μ med $\overline{x} = \frac{1}{n}\sum_i^n x_i = 1.7$. Eftersom $\sum_{i=1}^n x_i$ är ett utfall av den $\mathrm{Po}(n\mu)$-fördelade stokastiska variabeln $\sum_{i=1}^n X_i$ där $n\mu$ skattas med 17 så approximeras fördelningen för $\overline{X}$ av $\mathrm{N}(\mu, \sqrt{\mu/n})$.
Ett konfidensintervall för μ med konfidensgrad approximativt $1 - \alpha = 95\%$ ges av

$$\mu \in \overline{x} \pm \lambda_{\alpha/2}\sqrt{\frac{\overline{x}}{n}} = 1.7 \pm 1.96\sqrt{\frac{1.7}{10}} = 1.7 \pm 0.8 = \underline{(0.9,\ 2.5)}.$$

Eftersom $\mu = 2$ inte är ett orimligt värde enligt konfidensintervallet kan hypotesen $\mu = 2$ ej förkastas.

Uppgift 3

Låt A vara händelsen att den utvalda enheten är defekt. Då är

$$\begin{aligned} P(A) &= P(A|X=0)\,P(X=0) + P(A|X=0.1)\,P(X=0.1) \\ &\quad + P(A|X=0.2)\,P(X=0.2) + P(A|X=0.3)\,P(X=0.3) \\ &= 0 \cdot 0.10 + 0.1 \cdot 0.60 + 0.2 \cdot 0.25 + 0.3 \cdot 0.05 = \underline{0.125}. \end{aligned}$$

Låt Y vara antalet defekta av de tre utvalda. Då är

$$\begin{aligned} P(Y \geq 2|X=0) &= \binom{3}{2}0^2(1-0)^1 + \binom{3}{3}0^3(1-0)^0 = 0 \\ P(Y \geq 2|X=0.1) &= \binom{3}{2}0.1^2(1-0.1)^1 + \binom{3}{3}0.1^3(1-0.1)^0 = 0.028 \\ P(Y \geq 2|X=0.2) &= \binom{3}{2}0.2^2(1-0.2)^1 + \binom{3}{3}0.2^3(1-0.2)^0 = 0.104 \\ P(Y \geq 2|X=0.3) &= \binom{3}{2}0.3^2(1-0.3)^1 + \binom{3}{3}0.3^3(1-0.3)^0 = 0.216 \end{aligned}$$

och enligt lagen om total sannolikhet

$$\begin{aligned} P(Y \geq 2) &= P(Y \geq 2|X=0)\,P(X=0) + P(Y \geq 2|X=0.1)\,P(X=0.1) \\ &\quad + P(Y \geq 2|X=0.2)\,P(X=0.2) + P(Y \geq 2|X=0.3)\,P(X=0.3) \\ &= 0 \cdot 0.10 + 0.028 \cdot 0.60 + 0.104 \cdot 0.25 + 0.216 \cdot 0.05 = \underline{0.0536}. \end{aligned}$$

Uppgift 4

a) Det handlar om jämförelse av väntevärden med *stickprov i par*.

Före	20	25	26	22	24	48	52	27	18	12	28	31
Efter	18	23	26	18	26	44	51	26	22	11	29	31
z_i = Före − Efter	2	2	0	4	-2	4	1	1	-4	1	-1	0

Vi ser z_i, $i = 1, \ldots, 12$ som utfall av oberoende $\mathrm{N}(\Delta, \sigma)$-fördelade stokastiska variabler. Δ skattas med $\overline{z} = 8/12 = 0.67$, ett utfall av en $\mathrm{N}(\Delta, \sigma/\sqrt{n})$-fördelad stokastisk variabel $\overline{Z}$. Med $s_z = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})^2} = 2.31$ fås konfidensintervallet

$$\Delta \in \overline{z} \pm t_{0.025}(n-1)\frac{s_z}{\sqrt{n}} = 0.67 \pm 2.20\frac{2.31}{\sqrt{12}} = 0.67 \pm 1.47 = \underline{(-0.80,\ 2.13)}.$$

b) Eftersom $\Delta = 3$ inte är rimligt värde enligt det framtagna konfidensintervallet förkastas hypotesen $\Delta = 3$ på signifikansnivån 5%.

Uppgift 5

Med $Y = \sqrt{X/\pi}$ så har Y fördelningsfunktion

$$F_Y(t) = P(Y \le t) = P\left(\sqrt{X/\pi} \le t\right) = P\left(X \le \pi t^2\right) = F_X(\pi t^2)$$

dvs täthetsfunktion

$$f_Y(t) = \frac{d}{dt}F_Y(t) = \frac{d}{dt}F_X(\pi t^2) = f_X(\pi t^2)2\pi t = \underline{\frac{2\pi t}{\theta}\mathrm{e}^{-\pi t^2/\theta}}$$

för $t \ge 0$.
b) ML-skattningen av θ är det värde som maximerar

$$L(\theta) = f_{Y_1}(y_1)\cdots f_{Y_n}(y_n) = \frac{2\pi y_1}{\theta}\mathrm{e}^{-\pi y_1^2/\theta}\cdots\frac{2\pi y_n}{\theta}\mathrm{e}^{-\pi y_n^2/\theta} = \frac{(2\pi)^n(y_1\cdots y_n)}{\theta^n}\mathrm{e}^{-\frac{\pi}{\theta}\sum_{i=1}^{n} y_i^2}.$$

Det är samma värde som maximerar

$$\ln(L(\theta)) = \ln((2\pi)^n(y_1\cdots y_n)) - n\ln(\theta) - \frac{\pi}{\theta}\sum_{i=1}^{n} y_i^2.$$

Lösning av

$$0 = \frac{d}{d\theta}\ln(L(\theta)) = -\frac{n}{\theta} + \frac{\pi}{\theta^2}\sum_{i=1}^{n} y_i^2 = \frac{-n}{\theta^2}\left[\theta - \frac{\pi}{n}\sum_{i=1}^{n} y_i^2\right]$$

ger ML-skattningen $\underline{\theta^*{}_{\mathrm{obs}} = \frac{\pi}{n}\sum_{i=1}^{n} y_i^2}$.
För de tre observationerna $y_1 = 0.20$, $y_2 = 0.25$ och $y_3 = 0.18$ fås skattningen

$$\theta^*{}_{\mathrm{obs}} = \frac{\pi}{3}(0.20^2 + 0.25^2 + 0.18^2) = \underline{0.14\ \mathrm{cm}^2}.$$

c) Nu är

$$\theta^* = \frac{\pi}{n}\sum_{i=1}^{n} Y_i^2 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

där $X_1, \ldots, X_n$ är exponentialfördelade med väntevärde θ. Alltså är $E(\theta^*) = \theta$ och skattningen $\underline{\theta^*{}_{\mathrm{obs}}\ \text{är väntevärdesriktig}}$.

Uppgift 6

Vi vill testa H_0 formulerad av matematiklärarna, det vill säga

$$H_0: P(\text{underkänt}) = 0.40, \quad P(\text{betyg 3}) = 0.40, \quad P(\text{betyg 4 eller 5}) = 0.20.$$

Vi förkastar H_0 för stora värden på

$$q = \frac{(95 - 200 \cdot 0.4)^2}{200 \cdot 0.4} + \frac{(60 - 200 \cdot 0.4)^2}{200 \cdot 0.4} + \frac{(45 - 200 \cdot 0.2)^2}{200 \cdot 0.2} = 8.4375$$

som om H_0 är sann är ett utfall från en approximativt $\chi^2(3-1)$-fördelad stokastisk variabel. (Notera att de förväntade värdena $np_i \geq 5$.) Ur $\chi^2(2)$-tabell fås att $\chi^2_{0.05} = 5.99$ så testet som förkastar H_0 då $q > 5.99$ har signifikansnivå ≈ 0.05.
Här är $q = 8.4375 > 5.99$, så vi <u>förkastar H_0</u>. Observationerna är inte förenliga med den tänkta betygsfördelningen.