TECHNISCHE UNIVERSITEIT EINDHOVEN
Faculteit Wiskunde en Informatica

Exam Pre-master Probability and Statistics (2DI50) on 23 March 2007, 9.00-12.00

- At the exam it is allowed to use the 'Statistisch Compendium' and a pocket calculator.
- The exam has 8 exercises with 20 items.
- The answers have to be in ENGLISH. Give not just answers, but also arguments.
- For each item one get get two points. In total one can get 40 points.

Problem 1:

In a population of 100 individuals, 5 of them have disease D_1, 10 of them have disease D_2, and 85 of them have neither disease D_1 nor D_2. None of the individuals has both diseases.
The same symptom appears among 4 people having disease D_1, among 6 people having disease D_2 and among 5 people that have neither disease D_1 nor D_2.

(a) What is the probability that a randomly chosen (from the population of 100 individuals) person showing the symptom has either disease D_1 or disease D_2 ?
(b) What is the probability that a randomly chosen person showing the symptom has neither disease D_1 nor D_2.

Solution: Let
A_1 denote the event that a person has disease D_1,
A_2 denote the event that a person has disease D_2,
A_3 denote the event that a person has none of the both diseases,
S denote the event that a person shows the symptom.

(a) We are looking for $\mathbb{P}(A_1 \cup A_2|S)$. Now

$$\mathbb{P}(A_1 \cup A_2|S) = \mathbb{P}(A_1|S) + \mathbb{P}(A_2|S) - \mathbb{P}(A_1 \cap A_2|S)$$

We have $\mathbb{P}(A_1 \cap A_2|S) = 0$ because $A_1 \cap A_2 = \emptyset$ (none of the individuals has both diseases). On the other hand

$$\mathbb{P}(A_1|S) = \frac{\mathbb{P}(S|A_1)\mathbb{P}(A_1)}{\mathbb{P}(S)} = \frac{\frac{4}{5}\frac{5}{100}}{\frac{15}{100}} = \frac{4}{15}$$

$$\mathbb{P}(A_2|S) = \frac{\mathbb{P}(S|A_2)\mathbb{P}(A_2)}{\mathbb{P}(S)} = \frac{\frac{6}{10}\frac{10}{100}}{\frac{15}{100}} = \frac{6}{15}$$

which implies $\mathbb{P}(A_1 \cup A_2|S) = \frac{10}{15} = \frac{2}{3}$
(b) We are looking for $\mathbb{P}(A_3|S)$. As before, one has

$$\mathbb{P}(A_3|S) = \frac{\mathbb{P}(S|A_3)\mathbb{P}(A_3)}{\mathbb{P}(S)} = \frac{\frac{5}{85}\frac{85}{100}}{\frac{15}{100}} = \frac{5}{15} = \frac{1}{3}$$

Alternatively one might observe that $\mathbb{P}(A_3|S) = 1 - \mathbb{P}(\bar{A}_3|S) = 1 - \mathbb{P}(A_1 \cup A_2|S) = 1 - \frac{2}{3} = \frac{1}{3}$.

Problem 2:

The amount of bread (measured in hundreds of Kg) sold by a bakery in a day is a random variable X having probability density

$$f_X(x) = \begin{cases} c\,x & 0 \leq x < 3 \\ c(6-x) & 3 \leq x < 6 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the constant c.
(b) What is the probability that in one day more than 300 Kg (event A) are sold? What is the probability that in one day between 150 and 450 Kg (event B) are sold?
(c) Are the events A and B independent?

Solution:
(a) We have to impose the normalization condition: $\int_{-\infty}^{+\infty} f_X(x)dx = 1$. That is

$$\int_0^3 c\,x dx + \int_3^6 c(6-x)dx = 1$$

This implies $c = \frac{1}{9}$.
(b) We have

$$\mathbb{P}(A) = P(X > 3) = \int_3^{+\infty} f_X(x)dx = \frac{1}{2}$$

$$\mathbb{P}(B) = P(1.5 < X < 4.5) = \int_{1.5}^{4.5} f_X(x)dx = \frac{3}{4}$$

(c) Yes, A and B are independent. Indeed we have

$$\mathbb{P}(A \cap B) = P(3 < X < 4.5) = \int_3^{4.5} f_X(x)dx = \frac{3}{8} = \mathbb{P}(A)\mathbb{P}(B)$$

Problem 3:

A man with n keys wants to open his door and tries the keys ar random. Exactly one key will open the door.
(a) Suppose that unsuccessful keys are NOT eliminated from the n keys. Determine the probability distribution of X, the number of trials until he opens the door. Find the range of X and its expected number of trials.
(b) Suppose that unsuccessful keys are eliminated from the n keys. Determine the probability distribution of the number of trials until he opens the door. Find the expected number of trials.

Solution:
(a) Denoting by X the number of trials until he opens the door, we have that X has a geometric probability distribution with parameter $p = \frac{1}{n}$, since the probability of successfully opening the door at a generic trial is constant. The range of X will then be $\mathbb{N}$ and its expectation is $\mathbb{E}(X) = \frac{1}{p} = n$.
(b) In this case the man will open the door by at most n trials. We have

$$\mathbb{P}(X = 1) = \mathbb{P}(\text{open door} 1^{\text{st}} \text{trial}) = \frac{1}{n}$$

and hence

$$\mathbb{P}(X > 1) = 1 - \mathbb{P}(\text{open door} 1^{\text{st}} \text{trial}) = \frac{n-1}{n}$$

After k unsuccessful attempts, there are $n-k$ keys left from which we choose randomly the next one. The probability that this key opens the door is thus

$$\mathbb{P}(X = k+1 \,|\, X > k) = \mathbb{P}(\text{open door}(k+1)^{\text{th}} \text{trial} \,|\, k^{\text{th}} \text{attempt unsuccessful}) = \frac{1}{n-k}$$

which also implies

$$\begin{aligned} \mathbb{P}(X > k+1 \,|\, X > k) &= \mathbb{P}((k+1)^{\text{th}} \text{attempt unsuccessful} \,|\, k^{\text{th}} \text{attempt unsuccessful}) \\ &= 1 - \frac{1}{n-k} = \frac{n-k-1}{n-k} \end{aligned}$$

Hence by the rule of total probability

$$\begin{aligned} \mathbb{P}(X = k+1) &= \mathbb{P}(X = k+1 \,|\, X > k)\, \mathbb{P}(X > k \,|\, X > k-1) ... \mathbb{P}(X > 2 \,|\, X > 1) \mathbb{P}(X > 1) \\ &= \frac{1}{n-k} \frac{n-k}{n-k+1} ... \frac{n-2}{n-1} \frac{n-1}{n} = \frac{1}{n}. \end{aligned}$$

So X is uniformly distributed in $\{1, \dots, n\}$. The expectation is $\mathbb{E}(X) = \frac{n+1}{2}$.

Problem 4:

The length of the time required to complete a college test is found to be normally distributed with mean 50 minutes and standard deviation 12 minutes.

(a) When should the test be terminated if we wish to allow sufficient time for 90% of the students to complete the test?
(b) What proportion of students will finish the test between 30 and 60 minutes?

Solution:
(a) Let X be the length of time to complete the test. Then $Z = \frac{X-50}{12} \sim N(0,1)$. We need to find c such that $\mathbb{P}(X < c) = 0.9$. From the Statistisch Compendium you see that $\mathbb{P}(Z < 1.28) = 0.9$, which implies

$$\frac{c-50}{12} = 1.28$$

The solution is $c = 65.36$
(b) One has to compute

$$\mathbb{P}(30 < X < 60) = \mathbb{P}(-1.67 < Z < 0.83) = \mathbb{P}(Z < 0.83) - \mathbb{P}(Z < -1.67) = 0.7967 - 0.0475 = 0.7492$$

The proportion of students that will finish the test between 30 and 60 minutes is 75%.

Problem 5:

Wires are manufactured with a nominal length of 20 cm. It is assumed that the length of a wire is a normal gaussian variable with mean μ and variance σ^2. A random sample of 25 measurements is taken and it is found a sample mean $\bar{x} = 16$ cm and a sample variance $S^2 = 100$ cm^2.
(a) Give a 99% confidence interval for the mean μ.
(b) Discuss whether the nominal length is a plausible value for the mean μ given the result of part (a).
(c) Give a 99% confidence interval for the variance σ^2.

Solution:
(a) The confidence interval is given by the formula:

$$(\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}})$$

For $n = 25$ and $\alpha = 0.01$ we find (use Statistisch compendium) $t_{24,0.005} = 2.80$. The given values $\bar{x} = 16$ cm and $s = 10$ cm imply a CI $= (10.4, 21.6)$ cm.
(b) The interval includes the value 20 cm, so on the basis of the confidence interval the hypothesis $\mu = 20$ would be accepted with a 99% confidence.
(c) From the Statistisch Compendium, we see that the confidence interval is given by

$$\frac{(n-1)\,S^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)\,S^2}{\chi^2_{n-1,1-\alpha/2}}.$$

Here, we have $n = 25$, $S^2 = 100$ and $\alpha = 0.01$. From the Statistisch Compendium, we see that $\chi^2_{24,0.005} = 45.6$, $\chi^2_{24,0.995} = 9.89$. Hence,

$$52.6 = \frac{24 \times 100}{45.6} < \sigma^2 < \frac{24 \times 100}{9.89} = 242.7$$

and the confidence interval is given by $CI = (52.6, 242.7)$.

Problem 6:

Two factories produce light bulbs whose lifetimes (measured in thousands of hours) are exponential random variables with parameter $\lambda_1 = 1$ for the first factory and parameter $\lambda_2 = 2$ for the second factory.
(a) What is the probability that the lifetime of a light bulb produced by the first factory is less then 1000 hours? What is the probability that the lifetime of a light bulb produced by the second factory is less then 1000 hours?

In a batch of 300 light bulbs, 200 have been produced by first factory and 100 have been produced by the second factory. A randomly chosen light bulb is tested and it is found that its lifetime is less than 1000 hours.
(b) What is the probability that this light bulb was produced by the first factory? (Hint: use Bayes formula).

Solution:
(a) Denoting by X the lifetime of the light bulb, we are asked to compute the $\mathbb{P}(X < 1)$. This is nothing else than the cumulative distribution function evaluated at 1. Being the lifetime exponential we have $F_X(x) = 1 - e^{-\lambda x}$. If the light bulb is produced by the first factory then we have $\mathbb{P}(X < 1) = 1 - e^{-1}$; if the light bulb is produced by the second factory then we have $\mathbb{P}(X < 1) = 1 - e^{-2}$.
(b) Let I' and I'' denote the events that the chosen light bulb comes from the first factory, respectively second factory. We have $\mathbb{P}(I') = \frac{2}{3}$ and $\mathbb{P}(I'') = \frac{1}{3}$. We are asked to compute $\mathbb{P}(I'|\{X < 1\})$. By the Bayes' theorem we have

$$\mathbb{P}(I'|\{X < 1\}) = \frac{\mathbb{P}(\{X < 1\}|I')\mathbb{P}(I')}{\mathbb{P}(\{X < 1\}|I')\mathbb{P}(I') + \mathbb{P}(\{X < 1\}|I'')\mathbb{P}(I'')}$$

Using the results of the previous item we find:

$$\mathbb{P}(I'|\{X < 1\}) = \frac{(1 - e^{-1})\frac{2}{3}}{(1 - e^{-1})\frac{2}{3} + (1 - e^{-2})\frac{1}{3}}$$

Problem 7:

Let U be a random variable which takes values 0, 1 and 4 with probabilities $1 - 2\theta + \frac{3}{2}\theta^2$, $2\theta(1 - \theta)$ and $\frac{1}{2}\theta^2$, respectively.

(a) Show that $\hat{\Theta} = U/2$ is an unbiased estimator for θ.
(b) Compute the mean square error of $\hat{\Theta}$

Solution
(a)We have

$$\mathbb{E}(U) = 0 \times \{1 - 2\theta + \frac{3}{2}\theta^2\} + 1 \times \{2\theta(1-\theta)\} + 4 \times \{\frac{1}{2}\theta^2\} = 2\theta$$

which implies

$$\mathbb{E}(\hat{\Theta}) = \theta$$

(b) Being the estimator unbiased it suffices to compute the variance. You find:

$$MSE(\hat{\Theta}) = V(\hat{\Theta}) = \frac{1}{2}\theta(1+\theta)$$

Problem 8:

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter p. The sample mean $\bar{X}$ and sample variance S^2 are defined as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad\qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

(a) Show that S^2 is an unbiased estimator for the population variance $\sigma^2 = p(1-p)$.
(b) Show that $\bar{X} - S^2$ is an unbiased estimator for p^2.
(c) A measure is taken for a random sample of size $n = 80$. This yields a sample mean of $\bar{x} = 0.275$. Compute an approximate 90% confidence interval for p, stating clearly any assumption which are necessary for the approximation to be valid.
(d) Suppose now that you would like to estimate p with a precision of ± 0.02, in the sense that a 90% confidence interval should have half-width no greater than 0.02. How large a sample size would be necessary to achieve this precision? Use the sample mean from the previous item involving 80 samples.

Solution:
(a) We have

$$\mathbb{E}(S^2) = \frac{1}{n-1}\left(\mathbb{E}(\sum_{i=1}^{n} X_i^2) - n\mathbb{E}(\bar{X}^2)\right)$$

Since $\mathbb{E}(X_i^2) = p$ and $\mathbb{E}(\bar{X}^2) = \frac{1}{n}p + \frac{n-1}{n}p^2$, we find

$$\mathbb{E}(S^2) = \frac{1}{n-1}\left(np - p - (n-1)p^2\right) = p(1-p)$$

(b) Combining together the previous answer and the fact that the sample mean is an unbiased estimator of the population mean (i.e. $\mathbb{E}(\bar{X}) = p$), we obtain

$$\mathbb{E}(\bar{X} - S^2) = p - p(1-p) = p^2$$

(c) We use the normal approximation to the binomial distribution, namely

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximatively standard normal. For the approximation to be adequate it is necessary that the value of n is large enough (which is certainly the case for $n = 80$) and that the value of p is not close to zero or one (which is the case here since the estimated proportion is 0.275). The $100(1-\alpha)\%$ CI is derived from

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{x} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Replacing p by $\bar{x}$ in the denominators of the previous expression one gets

$$\mathbb{P}(\bar{x} - z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \leq p \leq \bar{x} + z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}})$$

so that

$$\bar{x} - z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \leq p \leq \bar{x} + z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}$$

Imposing $\alpha = 0.1$ implies (use Statistisch compendium) $z_{0.05} = 1.645$. Inserting $n = 80$, $\bar{x} = 0.275$ in the previous formula we deduce $\text{CI} = (0.193, 0.357)$.

(d) We impose that the half-width of the confidence interval is smaller than 0.02

$$z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \leq 0.02$$

Solving for n we find $n \geq 1349$, so a further 1269 trials need to be examined in addition to the 80 in the initial sample.