Tentamen Statistiek voor KI/Inf/BMT (Külske) - Solutions

Monday 28 January 2008

All books, written notes, and all calculators allowed.
Cell phones and laptops not allowed.

1. Suppose that $f_Y(y) = 2y$ for $0 \leq y \leq 1$ is the density of a random variable Y.

   (a) Compute the expected value $E(Y)$ and the variance.

   (b) Compute the cdf (the cumulative distribution function).

   (c) Suppose that 8 random variables are drawn from this distribution. What is the probability that precisely 2 of these random variables have values in the interval $[0, \frac{1}{\sqrt{2}}]$?
   What is the probability that precisely 6 of these random variables have values in the interval $[0, \frac{1}{\sqrt{2}}]$?

   Solution:
   a) $E(Y) = \frac{2}{3}$, $VarY = EY^2 - (EY)^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$
   b) $F(y) = y^2$
   c) $P(0 \leq Y \leq \frac{1}{\sqrt{2}}) = \frac{1}{2}$
   and so $\binom{8}{2}\frac{1}{2^8} = \frac{7}{67} = 0.109375$ *in both cases*

2. Consider the quadratic density function on the unit interval we know from the lectures, given by $f_Y(y; 0) = 6y(1-y)$ for $y \in [0, 1]$.

   On the basis of this function we consider now a *shifted* density function, given by $f_Y(y; \theta) = 6(y-\theta)(1-y+\theta)$ for $y \in [\theta, 1+\theta]$ with the unknown mean value $\theta + \frac{1}{2}$.

   (a) Make a picture of this density, for a fixed θ.

   (b) Suppose that two realizations of this density are drawn with values 0.1 and 0.2. How would you naively estimate θ? Give a reason why $-\frac{7}{20}$ is a good estimate for θ!
   Show that it is also the maximum-likelihood estimate for θ!

   (c) Give the definition of an unbiased estimator.
   Is the estimator from part (b) unbiased?

   Solution:
   b) Choose θ *such that* $\frac{y_1+y_2}{2} = \theta + \frac{1}{2}$ *with* $y_1 = 0.1$ *and* $y_2 = 0.2$*, because of symmetry. This is also the moment estimate since the left hand side is the sample mean and the right hand side the expected value. Solving this equation gives* $\theta = -\frac{7}{20}$.

   For the maximum likelihood estimate we must have

$$0 = \frac{d}{d\theta} \log[6(0.1-\theta)(1-0.1+\theta)6(0.2-\theta)(1-0.2+\theta)]$$
$$= -\frac{1}{0.1-\theta} + \frac{1}{0.9+\theta} - \frac{1}{0.2-\theta} + \frac{1}{0.8+\theta}$$

It is simple to check by a computation that $\theta = -\frac{7}{20}$ solves this equation.

An estimator is unbiased if its expected value equals the parameter we want to estimate. But we defined our estimator $\hat{\theta}$ above in such a way that $\frac{Y_1+Y_2}{2} = \hat{\theta} + \frac{1}{2}$. If θ is the true parameter we have $\mathbf{E}_\theta(\frac{Y_1+Y_2}{2}) = \theta + \frac{1}{2}$, so that we get $\mathbf{E}_\theta\hat{\theta} = \theta$. As a conclusion, our estimator is unbiased. So, it is an example where the maximum likelihood estimator is unbiased.

3. Consider a random sample $Y_1, \ldots, Y_N$
with expected value μ and variance σ^2.

(a) Is the maximum likelihood estimator always unbiased?
Explain or give a counter-example!

(b) Suppose that μ is unknown.
Is $\frac{1}{N}\sum_{i=1}^N (Y_i - \bar{Y})^2$ where $\bar{Y} = \frac{1}{N}\sum_{i=1}^N Y_i$ an unbiased estimator for the variance σ^2?

(c) Suppose that μ is known.
Is $\frac{1}{N}\sum_{i=1}^N (Y_i - \mu)^2$ an unbiased estimator for the variance σ^2?

Solution:
a) No. One example is the estimator in part b). That the estimator from part b) is the maximum likelihood estimator for a normal sample with unknown μ and σ we saw in the lecture by a computation.

b) No. We know that $\frac{1}{N-1}\sum_{i=1}^N (Y_i - \bar{Y})^2$, which is known as the sample variance, is an unbiased estimator for the variance σ^2, see book. So the prefactor is wrong.

c) Yes. We have

$$\mathbf{E}\frac{1}{N}\sum_{i=1}^N (Y_i-\mu)^2 = \frac{1}{N}\sum_{i=1}^N \mathbf{E}(Y_i-\mu)^2 = \mathbf{E}(Y_1-\mu)^2 = var(Y_1) = \sigma^2$$

4. There were 5 open positions at a bank for which 100 candidates applied. A group A of 10 candidates went to school with the chairman's son, a group B of 90 candidates did not.

It turns out that 3 of the 5 people who finally got the job were from group A, and 2 of them from group B.

Does this sound sufficiently unfair to you, statistically speaking?
To answer use the hypergeometric distribution as a null hypothesis of fairness and compute the corresponding P-value. To do this write down the general definition of the P-value first!

Solution: See book for the general definition. The P-value is in this case the probability, under the assumption of fairness, to see 3 or more people from group A getting the job. This probability is computed to be

$$\frac{\binom{10}{5}\binom{90}{0} + \binom{10}{4}\binom{90}{1} + \binom{10}{3}\binom{90}{2}}{\binom{100}{5}} = \frac{631}{95060} = 0.00663791$$

This value is smaller than 1 percent, so the procedure looks indeed unfair, by any reasonable standards.

5. Are the dates of births randomly distributed over the month of the year?

   Consider the data below which describe the number of births in the months january - december 2006 in a big hospital.

   $$252, 255, 240, 294, 281, 266, 295, 230, 257, 227, 229, 267$$

   Can we reject the Null-hypothesis of same chances for a birth of $\frac{1}{12}$ for all months, at the level of 5 percent?

   Solution:

   We must perform a χ^2-test with null hypothesis that $p_i = \frac{1}{12}$ for all months ($\equiv$ classes), $i = 1, \ldots, 12$.

   Denote by k_i the number of births in month i and by $N = 3093$ the total number of births. The test statistic to be used becomes

   $$T = \sum_{i=1}^{12} \frac{(k_i - N/12)^2}{N/12}$$

   The value of this statistic on the data is $T \approx 24.4200$.

   Under the null-hypothesis T would have a χ^2-distribution with 11 degrees of freedom. But $T \approx 24.4200$ turns out to be bigger than the corresponding 5-percent percentile (see table in the book). So the observed value of the test statistic is more extreme and we will reject the null hypothesis. It seems that there are indeed some preferred months.