Mats Rudemo, tel 0708 626472, email rudemo@chalmers.se

**SHORT SOLUTIONS to Home Exam June 3, 2020, 14.00-18.00**

Course code TMS016/MSA301

Literature and notes may be used in this written Home examination. All types of pocket calculators and computers are allowed. You are not allowed to communicate with any individual in any way. In the written examination there are two pages and two problems. You are supposed to answer both problems, and in the judgement they have the same weight. Answers may be given in English or Swedish.

## Problem 1.

The left part of Figure 1 shows a histogram of non-zero wind power observations measured at 336 stations in Denmark. The locations of the wind power stations are shown in the right part of Figure 1 together with estimated (predicted) mean values of wind power computed from the wind power station measurements. From the colour bar we note that some estimated mean values are negative, but would then be interpreted as zero.

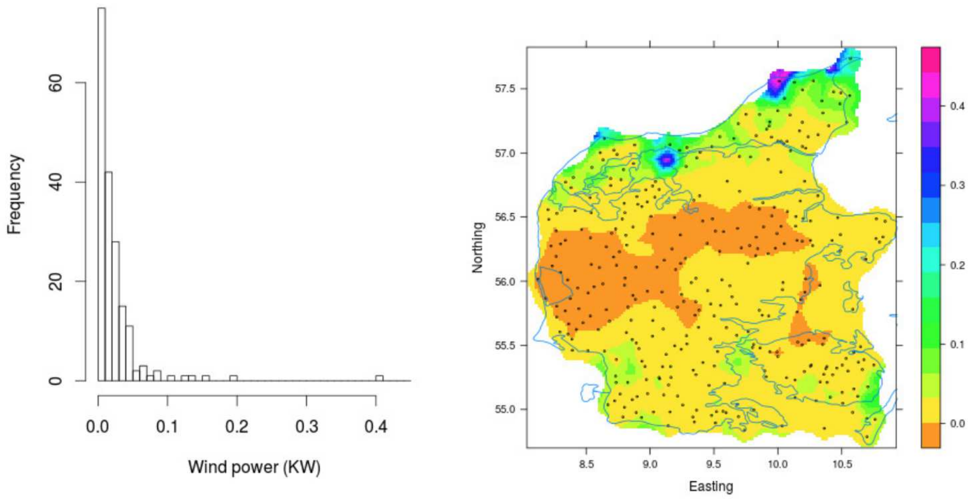


Figure 1: Left: histogram of the non-zero wind power observations measured at a specific day and time interval 2009 at 336 stations. Right: estimated mean wind power in Denmark (for the specific daytime) together with the wind power stations shown as black dots.

**a)** Suggest a model and a method that could be used to produce an estimated mean map such as shown in right part of Figure 1. Give details with suitable formulas. NOTE: an accurate model of the data may be out of scope for the present course, but a bold approximation could be quite useful.

**b)** Show with suitable details how one could produce a map similar to the right part of Figure 1 but with estimated standard deviations instead of estimated means.

## SOLUTION to Problem 1.

**a)** From Figure 1 left we see that the wind power distribution has a long tail to the right. Typically we should then use a transformation such as a log transformation. Let $Z(s)$ denote the wind power at location $s$, put $Y(s) = \log(Z(s) + a)$ with a parameter $a$ and assume that

$$Y(s) = \sum_{k=1}^{K} B_k(s)\beta_k + \epsilon(s) \tag{1}$$

where $B_1(s), \ldots, B_K(s)$ are suitable covariates at site $s$ such as height, distance to the sea et cetera, and $\epsilon(s)$ are $N(0, \sigma^2)$ noise variables, independent for different locations $s$.

Let $s_1, \ldots, s_N$, $N = 336$, be the locations of the wind power stations. The log-likelihood for our observations is

$$\ell(a, \beta_1, \ldots, \beta_K, \sigma) = \sum_{i=1}^{N} \log\left\{\frac{1}{\sigma}\phi\left(\frac{Y(s_i) - \sum_{k=1}^{K} B_k(s_i)\beta_k}{\sigma}\right)\right\}, \tag{2}$$

where $\phi$ is the density of a standard normal variable. Maximization of the log-likelihood (by a computer method) gives maximum likelihood estimates $\hat{a}, \hat{\beta}_1, \ldots, \hat{\beta}_K, \hat{\sigma}$.

We note that

$$Z(s) = -a + e^{\epsilon(s)} \exp\{\sum_{k=1}^{K} B_k(s)\beta_k\}. \tag{3}$$

A simple computations shows that $\boldsymbol{E}e^{\epsilon(s)} = e^{\sigma^2/2}$. Thus we find

$$\mu(s) = \boldsymbol{E}\{Z(s)\} = -a + e^{\sigma^2/2} \exp\{\sum_{k=1}^{K} B_k(s)\beta_k\}, \tag{4}$$

which we estimate by

$$\hat{\mu}(s) = -\hat{a} + e^{\hat{\sigma}^2/2} \exp\{\sum_{k=1}^{K} B_k(s)\hat{\beta}_k\}. \tag{5}$$

Plotting $\hat{\mu}(s)$ as a function of $s$ should give a plot similar to Figure 1, right part.

Further we note that the model (1) is an OLS model. We could go on to consider GLS or ML models, as in Lecture Notes Sections 5.4.2 and 5.4.3, but that would be more complicated.

**b)** Let us now consider estimation of the standard deviation (or equivalently variance) instead of the mean. We note that

$$\mathrm{Var}(Z(s)) = \boldsymbol{E}(Z(s))^2 - (\mu(s))^2 \tag{6}$$

and $(\mu(s))^2$ we can estimate by $(\hat{\mu}(s))^2$. It remains to estimate $\boldsymbol{E}(Z(s))^2$. We note that $\boldsymbol{E}e^{2\epsilon(s)} = e^{2\sigma^2}$ and find

$$\begin{aligned}
\boldsymbol{E}(Z(s))^2 &= \boldsymbol{E}\left(-a + e^{\epsilon(s)} \exp\{\sum_{k=1}^{K} B_k(s)\beta_k\}\right)^2 \\
&= a^2 - 2a\boldsymbol{E}(e^{\epsilon(s)}) \exp\{\sum_{k=1}^{K} B_k(s)\beta_k\} + \boldsymbol{E}(e^{2\epsilon(s)}) \exp\{2\sum_{k=1}^{K} B_k(s)\beta_k\} \\
&= a^2 - 2ae^{\sigma^2/2} \exp\{\sum_{k=1}^{K} B_k(s)\beta_k\} + \exp\{2\sigma^2 + 2\sum_{k=1}^{K} B_k(s)\beta_k\}.
\end{aligned} \tag{7}$$

This second order moment can be estimated by replacing parameters with their estimates, and then we proceed as in the solution of **a** to produce the wanted map.

## Problem 2.

Figure 2 shows results from an experiment with $16\times24 = 384$ colonies of yeast mutants grown under normal conditions (left) and in a nutrition solution with arsenic added (right). It is the same mutant grown in corresponding positions on both plates, for instance in the top left spot in both images. The object is to analyze the effect of arsenic on the different mutants.

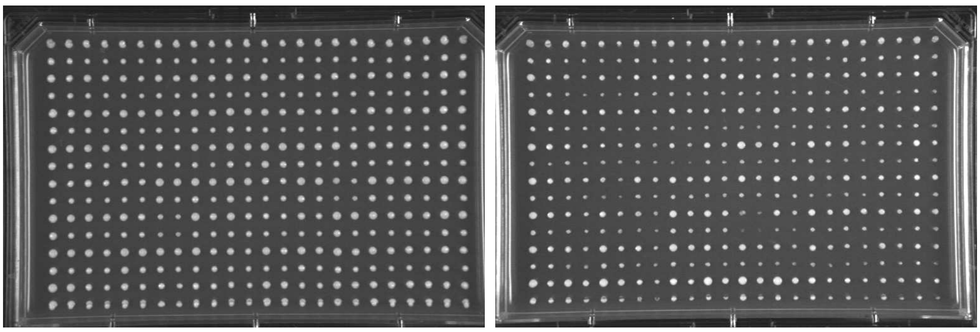

Figure 2: Images of two plates showing size of yeast colonies grown under normal conditions (left) and with arsenic added (right).

**a)** Suggest a method for computing the spot area of the 384 yeast colonies in each plate.

**b)** There are actually 96 different mutants studied in this experiment and each mutant is grown in a group of four positions in the following way: It is the same mutant in
row 1, column 1; row 1, column 2; row 2, column 1; row 2, column 2
and similarly in
row 1, column 3; row 1, column 4; row 2, column 3; row 2, column 4
and so on.

Further, in each group of four colonies for the same mutant the concentration decreases in the order shown above. (Check for yourself by looking at the images that this seems reasonable.) How it decreases is not precisely known, but it can be assumed that it is the same start amount (before growth) for the colonies in corresponding positions in the two plates.

Suggest a suitable statistical model for estimating the effect of arsenic on the growth of each of the 96 mutants. Assume that the growth of each colony is described by the corresponding spot area.

How can you for each of the 96 mutants test the hypothesis that arsenic has no effect on the growth of colonies?

**c)** How can you test the hypothesis that arsenic generally has no effect on the growth of yeast colonies? Discuss how valid the test is.

## SOLUTION to Problem 2.

**a)** Start by finding two thresholds $t_L$ and $t_R$ for the left and right plates, respectively. Compute and inspect for each plate the histogram of grey values. Find a suitable method to compute threshold. Perhaps it will work with taking the mean between two peaks, one peak for white and one for black pixels.

Associate with each of the spots disjoint quadratic areas safely containing the white pixels of the corresponding spot and denote by $S_{mct}$ the number of white pixels (above the threshold) in the quadratic area for mutant $m, m = 1, \ldots, M$, with $M = 96$, concentration $c, c = 1, \ldots, 4$, and treatment $t, t = 1, 2$, where $t = 1$ corresponds to the left plate and $t = 2$ corresponds to the right plate.

**b)** Put

$$Y_{mc} = \log(S_{mc1}/S_{mc2}) \tag{8}$$

and assume that $Y_{mc}, c = 1, \ldots, 4, m = 1, \ldots, M$, are independent and $N(\mu_m, \sigma_m^2)$. Thus $\mu_m$ is a measure of the effect of arsenic on mutant $m$. To test that arsenic has no effect on mutant $m$ we will test the hypothesis

$$\mathrm{H}_{0m} : \mu_m = 0. \tag{9}$$

A suitable test variable for this hypothesis is

$$t_m = \frac{\overline{Y}_m}{s_m/\sqrt{4}}, \tag{10}$$

where $\overline{Y}_m = (1/4)\sum_{c=1}^{4} Y_{mc}$ and $s_m^2 = (1/3)\sum_{c=1}^{4}(Y_{mc} - \overline{Y}_m)^2$. We reject the hypothesis $\mathrm{H}_{0m}$ on the 5% level if

$$|t_m| > t_{.975,3}, \tag{11}$$

where $t_{.975,3}$ is the 0.975 quantile of a $t$-distribution with 3 degrees of freedom. (A one-sided test with rejection if $t_m > t_{.95,3}$ could also be motivated.)

**c)**

To test that arsenic has no effect on any of the mutants we will test the hypothesis

$$\mathrm{H}_0 : \mu_m = 0, m = 1, \ldots, M \tag{12}$$

Assume for simplicity that $\sigma_m^2 = \sigma^2$ for all $m$. A suitable test variable is now

$$t = \frac{\overline{Y}}{s/\sqrt{4M}}, \tag{13}$$

where $\overline{Y} = (1/4M)\sum_{c=1}^{4}\sum_{m=1}^{M} Y_{mc}$ and $s^2 = (1/(3M))\sum_{m=1}^{M}\sum_{c=1}^{4}(Y_{mc} - \overline{Y}_m)^2$. We reject the hypothesis $\mathrm{H}_0$ on the 5% level if

$$|t| > t_{.975,3M}. \tag{14}$$

One could check the validity of assumptions by plotting the histogram of all $4M$ residuals $Y_{mc} - \overline{Y}_m$ and see if it looks normal.