Norwegian University of Science and Technology
Department of Mathematical Sciences

Corrected 22 and 23 May 2014

TMA4245 2014V – Solutions

Problem 1 Collectable cards

a) The second card can be any of 50 equally likely cards (we consider this the sample space), of which 49 correspond to the event that the second card is different from the first card. So the probability that the two cards are different, is 49/50.

The eight cards can be considered an ordered (e.g. by the order they where bought) sample from a sample space of 50^8 equally likely samples of eight cards. Of those, $50 \cdot 49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44 \cdot 43$ correspond to the event that all eight cards are different, giving a probability of $50 \cdot 49 \cdots 43/50^8 = 0.554$ that all are different.

Alternatively, the second question can be answered by conditional probability and the product rule. Let A_i be the event that the ith card is different from the cards 1, 2, …, $i-1$, where $i = 2, 3, \ldots, 8$. Then

$$P(i \text{ first cards are different} \mid i-1 \text{ first cards are different}) = P(A_i \mid A_2 \cap A_3 \cap \cdots \cap A_{i-1}) = \frac{51-i}{50}$$

(a sample space of 50 equally likely cards for the ith card, of which $51-i$ correspond to the event) for $i = 3, 4, \ldots, 8$, so that

$$\begin{aligned} P(\text{all eight cards are different}) &= P(A_2 \cap A_3 \cap \cdots \cap A_8) \\ &= P(A_2)\, P(A_3 \mid A_2)\, P(A_4 \mid A_2 \cap A_3) \cdots P(A_8 \mid A_2 \cap \cdots \cap A_7) \\ &= \frac{49}{50}\frac{48}{50}\frac{47}{50}\frac{46}{50}\frac{45}{50}\frac{44}{50}\frac{43}{50} = 0.554. \end{aligned}$$

b) The cumulative probability distribution of each X_i is given by $P(X_i \le x) = \sum_{j=1}^{x} P(X_i = j) = \sum_{j=1}^{x} 1/\theta = x/\theta$.

$P(\max X_i \le x) = P(X_1 \le x \cap X_2 \le x \cap \cdots \cap X_n \le x) = P(X_1 \le x)\, P(X_2 \le x) \cdots P(X_n \le x) = (P(X_i \le x))^n = (x/\theta)^n$.

With $n = 20$ and $\theta = 200$ we get $P(\max X_i \le 170) = (170/200)^{20} = 0.0388$.

c) In Agnes' case the null hypothesis is rejected, since 170 is in the critical region.

The significance level of the test is the probability of rejecting the null hypothesis when the null hypothesis is true, in this case $P(\max X_i \le 172) = (172/200)^{20} = 0.0490$.

Test power is the probability of rejecting the null hypothesis as a function of θ. For $\theta = 180$ the power is $P(\max X_i \le 172) = (172/180)^{20} = 0.403$. For $\theta = 160$, $\max X_i$ is always in the critical region, and the power is 1.

d) The joint probability mass function is given by $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^{n} 1/\theta = 1/\theta^n$ for x_1, x_2, $\dots x_n$ all in the range 1, 2, $\dots$, θ and it is equal to zero otherwise. Considered a function of θ, this also gives the likelihood function. Smaller values of θ give larger values of the likelihood, however, $\theta < \max x_i$ gives a likelihood of zero. The maximum is thus attained at $\theta = \max x_i$, so that the maximum likelihood estimator is $\max X_i$, and Agnes' estimate is 170.

By definiton of expected value, $E \max X_i = \sum_{x=1}^{\theta} x\, P(\max X_i = x)$. For $\max X_i$ to be unbiased, all probability mass of $\max X_i$ would have to be concentrated in θ, that is, $P(\max X_i = \theta) = 1$ and $P(\max X_i = x) = 0$ for all other x. All other probability mass functions of $\max X_i$ on 1, 2, $\dots$, θ will give a smaller expected value than θ. Obviously, $P(\max X_i = \theta) < 1$ – indeed, $P(\max X_i = \theta) = 1 - P(\max X_i \leq \theta - 1) = 1 - (1 - 1/\theta)^n < 1$. So the estimator is not unbiased.

Problem 2 Phosphorous at cleaning plant

a) The observations from Vik seem not to come from a normal distribution as they are not symmetric around the median – the larger observations are more spread out than the smaller. The observations from Nes might come from a normal distributions as they are symmetric around the median.

The sample medians are almost the same, but that of Nes is slightly smaller. Possibly the same applies to the medians of the distributions.

Because the Vik observations above the sample median are more spread out than those below, the sample mean will be above the sample median. As observations from Nes seem to be symmetric, sample mean and median are the same. Hence, Vik has larger sample mean than Nes, suggesting that the expected value of Vik is larger than the expected value of Nes.

Vik has more spread out observations, which suggests a larger variance than Nes.

b) Let Z denote a standard normal variable. Then

$$P(Y < 0.5) = P\left(\frac{Y - 0.3}{0.1} < \frac{0.5 - 0.3}{0.1}\right) = P(Z < 2) = 0.9772,$$

$P(Y > 0.3) = 0.5$ since Y has a probability density function that is symmetric around the mean 0.3, and

$$\begin{aligned} &P(Y < 0.5 \mid Y > 0.3) \\ &\qquad = \frac{P(0.3 < Y < 0.5)}{P(Y > 0.3)} = \frac{P(Y < 0.5) - P(Y < 0.3)}{P(Y > 0.3)} = \frac{0.9772 - 0.5}{0.5} = 0.9545. \end{aligned}$$

c) The phosphorous content at flow rate q is $Y = \alpha + \beta q + \epsilon$, with ϵ normally distributed with mean 0 and variance 0.05^2. This is a linear function of the normally distributed ϵ, so Y has a normal distribution. The expected value is $EY = E(\alpha + \beta q + \epsilon) = \alpha + \beta q + E\epsilon =$

$\alpha + \beta q + 0 = \alpha + \beta q$, and the variance is $\operatorname{Var}(\alpha + \beta q + \epsilon) = \operatorname{Var}\epsilon = \sigma_\epsilon^2$. With $\alpha = 0.05$, $\beta = 0.3$, $\sigma_\epsilon^2 = 0.05^2$, the expected value evaluates to $0.05 + 0.3 \cdot 0.5 = 0.2$ for $q = 0.5$ and to $0.05 + 0.3 \cdot 1.0 = 0.35$ for $q = 1.0$. In both cases the variance is 0.05^2.

Let Y_1, Y_2 and Y_3 be three independent phosphorous measurements at flow rate 1.0. Then

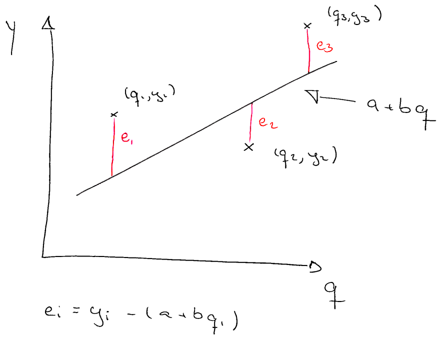
$$\begin{aligned}
&P(\text{the largest is larger than } 0.4) = 1 - P(\text{the largest is smaller than } 0.4)\\
&= 1 - P(X_1 \le 0.4, X_2 \le 0.4, X_3 \le 0.4) = 1 - (P(X_i \le 0.4))^3\\
&= 1 - \left(P\left(\frac{X_i - 0.35}{0.05} \le \frac{0.4 - 0.35}{0.05}\right)\right)^3 = 1 - (P(Z \le 1))^3 = 1 - 0.8413^3 = 0.4044,
\end{aligned}$$

where Z has the standard normal distribution.

Let Y_4 and Y_5 be independent phosphorous measurements at flow rates 0.5 and 1.0, respectively. Then $Y_4 - Y_5$ is a linear combination of independent variables having normal distributions, hence it is normally distributed. Its expected value is $E(Y_4 - Y_5) = EY_4 - EY_5 = 0.2 - 0.35 = -0.15$ and its variance $\operatorname{Var}(Y_4 - Y_5) = \operatorname{Var} Y_4 + (-1)^2 \operatorname{Var} Y_5 = 0.05^2 + 0.05^2 = 0.005$. Then

$$\begin{aligned}
P(Y_4 > Y_5) = P(Y_4 - Y_5 > 0) &= P\left(\frac{Y_4 - Y_5 - (-0.15)}{\sqrt{.005}} > \frac{-(-0.15)}{\sqrt{.005}}\right)\\
&= P(Z > 2.12) = 1 - P(Z \le 2.12) = 1 - 0.9831 = 0.0169.
\end{aligned}$$

d) In the method of least squares, the estimates a and b of α and β, respectively, are those minimizing the error sum of squares, $\text{SSE} = \sum_{i=1}^n (y_i - (a + bq_i))^2$, that is, the sum of squares of the difference between observed and fitted values.



The minimization is done by setting the partial derivatives equal to zero and solving the system of two equations for a and b,

$$0 = \frac{\partial}{\partial a}\text{SSE} = -2\sum_{i=1}^n (y_i - a - bq_i) \qquad \text{and} \qquad 0 = \frac{\partial}{\partial b}\text{SSE} = -2\sum_{i=1}^n q_i(y_i - a - bq_i).$$

e) Consider a new observation $Y_0 = \alpha + \beta q_0 + \epsilon_0$ independent of the data set. Then Y_0 and its predictor $\hat{Y}_0 = A + Bq_0$ are independent since the estimators A and B of α and β, respectively, are functions (even linear combinations) of the Y_i of the data set. Further, $Y_0 - \hat{Y}_0$ has a normal distribution, since it is a linear combination of Y_0 and the Y_i of the data set. Its expected value is $E(Y_0 - \hat{Y}_0) = EY_0 - E\hat{Y}_0 = \alpha + \beta q_0 - E(A + Bq_0) = \alpha + \beta q_0 - EA - q_0 EB = \alpha + \beta q_0 - \alpha - q_0\beta = 0$ and its variance is

$$\begin{aligned}
\sigma^2_{Y_0 - \hat{Y}_0} &= \operatorname{Var}(Y_0 - \hat{Y}_0) = \operatorname{Var} Y_0 + (-1)^2 \operatorname{Var}\hat{Y}_0\\
&= \sigma_\epsilon^2 + \sigma_\epsilon^2\left(\frac{1}{n} + \frac{(q_0 - \bar{q})^2}{\sum_{i=1}^n (q_i - \bar{q})^2}\right) = \sigma_\epsilon^2\left(1 + \frac{1}{n} + \frac{(q_0 - \bar{q})^2}{\sum_{i=1}^n (q_i - \bar{q})^2}\right).
\end{aligned}$$

Hence,

$$P\Big(- z_{\alpha/2} < \frac{Y_0 - \hat{Y}_0}{\sigma_{Y_0 - \hat{Y}_0}} < z_{\alpha/2} \Big) = 1 - \alpha,$$

and solving for Y_0 yields

$$P(\hat{Y}_0 - z_{\alpha/2}\sigma_{Y_0-\hat{Y}_0} < Y_0 < \hat{Y}_0 + z_{\alpha/2}\sigma_{Y_0-\hat{Y}_0}),$$

and by substituting estimates for estimators we get a prediction interval

$$\big[\hat{y}_0 - z_{\alpha/2}\sigma_{Y_0-\hat{Y}_0},\, \hat{y}_0 + z_{\alpha/2}\sigma_{Y_0-\hat{Y}_0}\big],$$

where $\hat{y}_0 = a + bq_0$.

If a random experiment is performed, and a $100(1-\alpha)\%$ confidence interval for a parameter is constructed on the basis of data of the experiment, the probability that the interval will contain the parameter is $1-\alpha$.

If a random experiment is performed, and a $100(1-\alpha)\%$ prediction interval for an observation, that is independent of the data of the experiment but created in the same way as the data of the experiment, is constructed on the basis of data of the experiment, the probability that the interval will contain the observation is $1-\alpha$.

In our context, a $100(1-\alpha)\%$ prediction interval for Y_0 is always wider than a $100(1-\alpha)\%$ confidence interval for $EY_0 = \alpha + \beta q$, since when constructing the prediction interval we have to consider the randomness of Y_0.

A 95% prediction interval should contain about 95% of the observations. (It should actually contain about 95% of new independent observations, but when n is large, as here, the prediction intervals based on $n-1$ of the observations will not differ much from the prediction intervals based on all of them.) The interval in Figure 3 contains a far smaller proportion of the observations and is probably a confidence interval.

f) Assumptions of the regression model: The expected value of the phosphorous content is a linear function of the water flow rate, $EY_i = \alpha + \beta q_i$. The discrepancies from the expected value are independent, normally distributed and have equal variance.

Figures 2 and 4 (left) suggest that the relationship might not be linear. There seems to be a change point around $q = 0.8$. The residuals $\hat{y}_i - y_i$ seem to depend on q; small q have positive residuals, medium q negative residuals and large q positive residuals. This non-linearity of the data points may be because the linearity assumption of the model is not satisfied, or due to dependent discrepancies. Further, the variance seems to be larger for values above 0.8.

The residuals from the QQ-plot in Figure 4 seem to come from a normal distribution as they approximately follow the straight line.