Norwegian University of Science and Technology
Department of Mathematical Sciences

Corrected 30 May and 4 June 2012

Solutions

TMA4245 Statistics

Saturday 26 May 2010 9:00–13:00

Problem 1 Snow density

a) The probability is $\int_{0.5}^{0.9} 6x(1-x)\,dx = \int_{0.5}^{0.9}(6x-6x^2)\,dx = [3x^2-2x^3]_{0.5}^{0.9} = 0.472$.

b) The likelihood function is given by

$$L(\beta) = \prod_{i=1}^{n} \beta(\beta+1)x_i(1-x_i)^{\beta-1} = \beta^n(\beta+1)^n\left(\prod_{i=1}^{n} x_i\right)\prod_{i=1}^{n}(1-x_i)^{\beta-1},$$

and the log likelihood

$$\ln L(\beta) = n\ln\beta + n\ln(\beta+1) + \sum_{i=1}^{n}\ln x_i + (\beta-1)\sum_{i=1}^{n}\ln(1-x_i),$$

which has derivative

$$(\ln L)'(\beta) = \frac{n}{\beta} + \frac{n}{\beta+1} + \sum_{i=1}^{n}\ln(1-x_i).$$

$(\ln L)'$ is decreasing on $(0,\infty)$ and the sum of two first terms tends to ∞ when $\beta \to 0^+$ and to 0 when $\beta \to \infty$, so that $(\ln L)'$ will have a single zero (the third term is negative) for $\beta > 0$ and be positive left of the zero and negative right of the zero. This means that L has its maximum at this zero. Solving for the zero,

$$\beta^2\sum_{i=1}^{n}\ln(1-x_i) + \left(2n + \sum_{i=1}^{n}\ln(1-x_i)\right)\beta + n = 0,$$

we get

$$\beta = \frac{-2n - \sum_{i=1}^n \ln(1-x_i) \pm \sqrt{4n^2 + (\sum_{i=1}^n \ln(1-x_i))^2}}{2\sum_{i=1}^n \ln(1-x_i)}$$
$$= -\frac{n}{\sum_{i=1}^n \ln(1-x_i)} - \frac{1}{2} \pm \sqrt{\Big(\frac{n}{\sum_{i=1}^n \ln(1-x_i)}\Big)^2 + \frac{1}{4}}.$$

We choose the larger zero since $(\ln L)'$ has only one zero for positive arguments (the other we found must be negative), and get the maximum likelihood estimator

$$\sqrt{\Big(\frac{n}{\sum_{i=1}^n \ln(1-X_i)}\Big)^2 + \frac{1}{4}} - \frac{n}{\sum_{i=1}^n \ln(1-X_i)} - \frac{1}{2} = \sqrt{\frac{1}{\big(\overline{\ln(1-X)}\big)^2} + \frac{1}{4}} - \frac{1}{\overline{\ln(1-X)}} - \frac{1}{2}.$$

For $n = 100$ and $\sum_{i=1}^n \ln(1-x_i) = -104.0$ the estimate is $\sqrt{1/1.04^2 + 1/4} + 1/1.04 - 1/2 = 1.545$.

(The discussion of actual attainment of maximum at the zero and of which zero to be chosen, is not required.)

Problem 2 Temperature in March and April

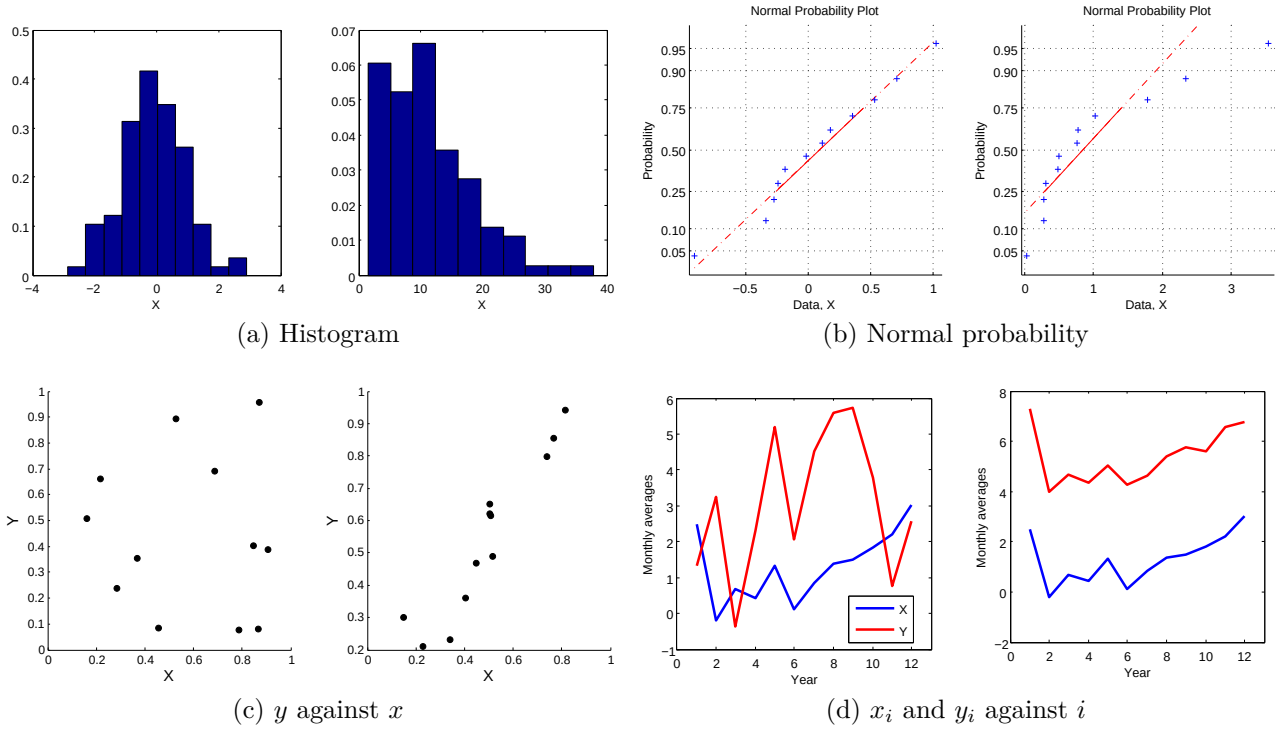
a) We can get an impression of normality of X or Y by means of a histogram (Figur 1a), and a more accurate assessment by a normal quantile–quantile plot, which should show linear relationships between ordered observations against normal quantiles, or a normal probability plot (Figure 1b) could be used, which should show linear relationships.

To assess whether X and Y are independent, we can plot the values of Y against X (Figure 1c). No special pattern should emerge, for example the points being close to a non-horizontal line (which indicates correlation). Also a plot of the values of the X_i and the Y_i against i in the same graph could reveal dependence (Figure 1d).

b) $(\bar{X} - \mu)/(S/\sqrt{n})$ has the t-distribution with $n-1$ degrees of freedom, so

$$P\Big(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\Big) = \alpha,$$

with $n-1$ degrees of freedom for $t_{\alpha/2}$. Solving the double inequality for μ, we get 99% confidence bounds $\bar{x} \pm t_{\alpha/2} s/\sqrt{n}$ for μ. Here, $n = 12$, $\bar{x} = \sum x_i/n = 9.10/12 = 0.758$, $\alpha = 0.01$, $t_{0.005} = 3.106$, $s^2 = \frac{1}{n-1}\Big(\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\Big) = \frac{1}{11}(77.07 - \frac{1}{12}(9.10)^2) = 6.379$, giving bounds $0.758 \pm 3.106\sqrt{6.379}/\sqrt{12} = 0.758 \pm 2.265$, and a confidence interval $(-1.5, 3.0)$.



(a) Histogram (b) Normal probability

(c) y against x (d) x_i and y_i against i

Figure 1: (a) Histograms and (b) normal probability plots. Data come from a normal distribution at the left but not at the right. (c) Plots of y against x and (d) of the x_i and y_i against i. X and Y are independent at the left, but not at the right.

c) We consider the null hypothesis $H_0\colon \mu_\mathrm{a}-\mu_\mathrm{m}=5$ (or $H_0\colon \mu_\mathrm{a}-\mu_\mathrm{m}\geq 5$) and the alternative hypothesis $H_1\colon \mu_\mathrm{a}-\mu_\mathrm{m}<5$.

The observations come naturally in pairs, and there is reason to believe that the March and the April temperatures of a year are dependent. Therefore we choose a paired test, which is performed as a single sample test using the differences $d_i = y_i - x_i$. Under the null hypothesis, the test statistic $T = (\bar{D}-5)/(S_D/\sqrt{12})$ has the t distibution with 11 degrees of freedom. A small value of T is indicative of H_1, and the critical value is $-t_{0.05} = -1.796$.

With our data, $\bar{d} = \bar{y}-\bar{x} = (67.40-9.10)/12 = 4.86$ and $s_D^2 = \frac{1}{n-1}\Big(\sum d_i^2 - \frac{1}{n}(\sum d_i)^2\Big) = \frac{1}{11}(364.53 - \frac{1}{12}(67.40-9.10)^2) = 7.39$. So $t = (4.86-5)/(\sqrt{7.39}/\sqrt{12}) = -0.18$, which is not in the critical region, and we do not reject H_0. At the 0.05 significance level there is not evidence to state that $\mu_\mathrm{a}-\mu_\mathrm{m}<5$.

Problem 3 Chemical factory

a) The probability that the lamp lights up when the procedure is performed once, is $P(X > 25) = P((X-20)/4 > (25-20)/4) = P(Z > 1.25) = P(Z < -1.25) = 0.1056$, where Z has the standard normal distribution.

If the procedure is performed three times, let the amounts of by-product be X_1, X_2, X_3. Then the probability that the lamp lights up at least once is $P(X_1 > 25 \cup X_2 > 25 \cup X_3 > 25) = 1 - P(X_1 \le 25,\ X_2 \le 25,\ X_3 \le 25) = 1 - (P(X_i \le 25))^3 = 1 - (1 - P(X_i > 25))^3 = 1 - (1 - 0.1056)^3 = 0.285$.

Let Y be the number of times the lamp lights up when the procedure is performed 100 times. Then Y has the binomial distribution with parameters $n = 100$ and $p = 0.1056$, and

$$P(Y \ge 15) = P(Y \ge 14.5) = P\left(\frac{Y - np}{\sqrt{np(1-p)}} \ge \frac{14.5 - 100 \cdot 0.1056}{\sqrt{100 \cdot 0.1056 \cdot 0.8944}}\right)$$
$$\approx P(Z > 1.28) = P(Z \le -1.28) = 0.100,$$

using the normal approximation with continuity correction. (The exact binomial probability is 0.104.)

(You are not penalized if you have not applied the continuity correction. If you use 15 instead of 14.5 in the normal approximation, you get 0.074, and if you use 14 (arising from $P(Y \ge 15) = 1 - P(Y \le 14)$), you get 0.131 – not very good approximations.)

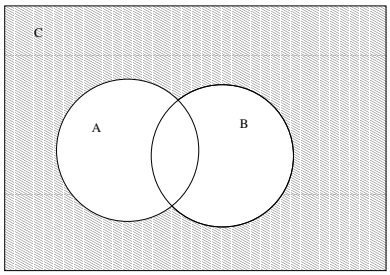
b) Let n be the number of times the procedure is performed, and let $X_1, X_2, \ldots, X_n$ be the amounts of by-product. The total amount of by-product, $Y = \sum_{i=1}^n X_i$, has the normal distribution with mean $20n$ and variance $4^2 n$. We want to find n such that

$$0.01 \le P(Y \ge 500) = P\Big(\frac{Y - 20n}{4\sqrt{n}} \ge \frac{500 - 20n}{4\sqrt{n}}\Big) = P\Big(Z \ge \frac{500 - 20n}{4\sqrt{n}}\Big),$$

that is, $(500 - 20n)/(4\sqrt{n}) \ge z_{0.01} = 2.326$, or $125 - 5n \ge z_{0.01}\sqrt{n}$, and we have a quadratic inequality $5n + z_{0.01}\sqrt{n} - 125 \le 0$ in $\sqrt{n}$. The left hand side is a downward-pointing parabola (as a function of $\sqrt{n}$) having zeros $\big(-z_{0.01} \pm \sqrt{z_{0.01}^2 + 4 \cdot 5 \cdot 125}\big)/(2 \cdot 5)$, that is, -5.24 and 4.77, meaning that $0 \le \sqrt{n} \le 4.77$ and $n \le 22.8$, that is, $n \le 22$ since n is an integer.

Problem 4 Barbecue tonight?

a) A Venn diagram is shown to the right.



$P(A \cap B) > 0$ implies that $A \cap B \neq \varnothing$, and A and B are not disjoint. $P(A)P(B) = 0.4 \cdot 0.4 = 0.16 \neq 0.2 = P(A \cap B)$, so A and B are not independent.

Note that $C = (A \cup B)'$. $P(C) = P((A \cup B)') = 1 - P(A \cup B) = 1 - (P(A) + P(B) - P(A \cap B)) = 1 - (0.4 + 0.4 - 0.2) = 0.4$.

$A \cap C = A \cap (A \cup B)' = A \cap A' \cap B' \subseteq A \cap A' = \varnothing$, so A and C are disjoint (this is is also obvious from the description of the events). The conditional probability of barbeque given no rain is $P(C \mid A') = P(C \cap A')/P(A') = P(C)/(1-P(A)) = 0.4/(1-0.4) = 0.67$.

b) $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$. Since $\bar{Y}$ is a linear combination of the Y_i, $1 \leq i \leq n$, and also $\hat{\beta}$ is a linear combination of the Y_i, also $\hat{\alpha}$ can be written as a linear combination of the Y_i, which are mutually independent and have normal distributions, thus $\hat{\alpha}$ has the normal distribution. $E\hat{\alpha} = E(\bar{Y} - \hat{\beta}\bar{x}) = \frac{1}{n}\sum EY_i - \beta\bar{x} = \frac{1}{n}\sum(\alpha + \beta x_i) - \beta\bar{x} = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha$, and $\operatorname{Var}\hat{\alpha} = \operatorname{Var}(\bar{Y} - \hat{\beta}\bar{x}) = \operatorname{Var}(\bar{Y}) + \bar{x}^2 \operatorname{Var}\hat{\beta} = \sigma^2/n + \bar{x}^2\sigma^2/\sum(x_i - \bar{x})^2 = (1/n + \bar{x}^2/\sum(x_i - \bar{x})^2)\sigma^2$ (which can be shown to be equal to $\sigma^2 \sum x_i^2/(n\sum(x_i - \bar{x})^2)$).

The assumptions are that the Y_i are independent variables having the normal distribution with mean $\alpha + \beta x_i$ and variance σ^2. The Figure indicates that relationship between EY and x might not be linear, as most points having a small or large x_i lie above the estimated regression line and most other points lie below. This could also be due to dependence between the Y_i. The assumption of constant variance seems to be OK.

c) An estimate of the temperature at 20:00 if the temperature at 13:00 is 15 °C, is $\hat{\alpha} + \hat{\beta} \cdot 15 = -6.57 + 1.43 \cdot 15 = 14.9$.

For making a 95% prediction interval for a new observation Y_0 corresponding to x_0, we consider the variable $\hat{Y}_0 - Y_0 = \hat{\alpha} + \hat{\beta}x_0 - \alpha - \beta x_0 - \epsilon$. Since $\hat{\alpha}$ and $\hat{\beta}$ are linear combinations of $Y_1, \ldots, Y_n$, $\hat{Y}_0 - Y_0$ is a linear combination of $Y_1, \ldots, Y_n$ and ϵ, which are all independent, so $\hat{Y}_0 - Y_0$ has a normal distribution. Its expected value is $E(\hat{Y}_0 - Y_0) = E(\hat{\alpha} + \hat{\beta}x_0 - \alpha - \beta x_0 - \epsilon) = 0$ and its variance $\operatorname{Var}(\hat{Y}_0 - Y_0) = \operatorname{Var}(\bar{Y} + \hat{\beta}(x_0 - \bar{x}) - \epsilon) = \sigma^2(1 + 1/n + (x_0 - \bar{x})^2/\sum_{i=1}^n (x_i - \bar{x})^2)$ ($\hat{\beta}$ and $\bar{Y}$ are independent), leading to a statistic $(\hat{Y}_0 - Y_0)/\left(\hat{\sigma}\sqrt{1 + 1/n + (x_0 - \bar{x})^2/\sum_{i=1}^n (x_i - \bar{x})^2}\right)$, which has the t-distribution with $n-2$ degrees of freedom. So

$$P\left(-t_{0.025} < \frac{\hat{Y}_0 - Y_0}{\hat{\sigma}\sqrt{1 + 1/n + (x_0 - \bar{x})^2/\sum_{i=1}^n (x_i - \bar{x})^2}} < t_{0.025} \right) = 0.95,$$

with $n - 2$ degrees of freedom for $t_{0.025}$. Solving the double inequality for Y_0, we get 95% prediction bounds $\hat{y}_0 \pm t_{0.025}\hat{\sigma}\sqrt{1 + 1/n + (x_0 - \bar{x})^2/\sum_{i=1}^n (x_i - \bar{x})^2}$ for y_0. Here, $n = 18$, $\bar{x} = 15.5$, $\hat{y}_0 = 14.9$, $t_{0.025} = 2.101$ ($n - 2 = 18$ degrees of freedom), giving bounds $14.9 \pm 2.101 \cdot 2.04\sqrt{1 + 1/20 + (15 - 15.5)^2/510.7} = 14.9 \pm 4.39$, and a prediction interval $(10.5, 19.3)$.