Norwegian University of Science and Technology
Department of Mathematical Sciences

Tentative solutions to TMA4240 Statistics, December 18, 2010

Problem 1 Hay fever and eczema

a) If E and H are independent events, then the probability that a child has eczema is equal to the probability that the child has eczema given that we already know that the child also has hay fever. Also, the probability that the child has both ezcema and hay fever is the product of the marginal probability that the child has ezcema and the probability that the child has hay fever.

E and H are not independent since; $P(H) \cdot P(E) = 0.04 \cdot 0.07 = 0.0028$ which is different from $P(E \cap H) = 0.009$.

Let E^* be the complemenary event of E and H^* the complementary event of H.

$$\begin{aligned} P(H^*|E^*) &= \frac{P(E^* \cap H^*)}{P(E^*)} \\ &= \frac{1 - P(E \cup H)}{1 - P(E)} \\ &= \frac{1 - (P(H) + P(E) - P(H \cap E))}{1 - P(E)} \\ &= \frac{1 - 0.04 - 0.07 + 0.009}{1 - 0.04} = \frac{0.899}{0.96} = 0.936 \end{aligned}$$

Problem 2 Sale of newspapers

X is Poisson distributed with expected value $\mathrm{E}(X) = \mu$. This means that

$$P(X = x) = \frac{\mu^x}{x!} \exp(-\mu)$$

a) Known: $\mu = 10$.

What is the probability that exactly 10 newspapers are sold:

$$P(X = 10) = \frac{10^{10} \cdot \exp(-10)}{10!} = 0.125$$

What is the probability that 13 or more newspapers are sold:

$$P(X \geq 13) = 1 - P(X \leq 12) = 1 - 0.7916 = 0.21$$

found in the table on page 20 of "Formelsamlingen".

b) $X_1, X_2, \ldots, X_n$ iid Poisson μ, where μ is unknown. Estimator for μ is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$\begin{aligned} \mathrm{E}(\bar{X}) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \\ \mathrm{Var}(\bar{X}) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mu = \frac{\mu}{n} \end{aligned}$$

Using the central limit theorem:

$$Z = \frac{\bar{X} - \mathrm{E}(\bar{X})}{\sqrt{\mathrm{Var}(\bar{X})}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\mu}{n}}} \approx \text{standard normal}$$

since $\bar{X}$ is the average of independent identically distributed random variables.

To construct a $(1-\alpha)100\%$ confidence interval we start with the $\alpha/2$ quantile of the standard normal distribution, $z_{\alpha/2}$:

$$\begin{aligned} P(-z_{\alpha/2} < Z < z_{\alpha/2}) &\approx 1 - \alpha \\ P(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sqrt{\frac{\mu}{n}}} < z_{\alpha/2}) &\approx 1 - \alpha \end{aligned}$$

This inequality is difficult to solve wrt μ so we choose to approximate the μ in the variance in the denominator with the estimator $\bar{X}$, and then solve wrt μ.

$$\begin{aligned} P(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sqrt{\frac{\bar{X}}{n}}} < z_{\alpha/2}) &\approx 1 - \alpha \\ P(\bar{X} - z_{\alpha/2}\sqrt{\frac{\bar{X}}{n}} < \mu < \bar{X} + z_{\alpha/2}\sqrt{\frac{\bar{X}}{n}}) &\approx 1 - \alpha \end{aligned}$$

This interval covers the unknown μ with approximate probability $1-\alpha$ when n is sufficiently large. Here $n = 30$ and in addition the Poisson distribution is rather symmetric for large μ, e.g. for $\mu = 10$.

Numerically:

- $z_{0.025} = 1.96$
- $n = 30$
- $\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = 10.75$

$$\mu \in [10.75 \pm 1.96 \cdot \sqrt{\frac{10.75}{30}}] = [9.58, 11.92]$$

c) We have $Y = \min(X, a)$, where X is Poisson with μ. If $y \leq (a-1)$ the newspaper is not sold out, and $P(Y = y) = P(X = y)$, but for $y = a$ we could have had a sale of at least a newspapers, that is $P(Y = a) = \sum_{x=a}^{\infty} P(X = x) = 1 - \sum_{x=0}^{a-1} P(X = x)$. Thus the distribution function for Y is

$$P(Y = y) = \begin{cases} P(X = y) & \text{when } y < a \\ 1 - \sum_{x=0}^{a-1} P(X = x) & \text{when } y = a \end{cases}$$

$$\begin{aligned} E(Y) &= \sum_{y=0}^{a} y \cdot P(Y = y) \\ &= \sum_{y=0}^{a-1} yP(X = y) + a \cdot P(Y = a) \\ &= \sum_{y=0}^{a-1} yP(X = y) + a \cdot (1 - \sum_{x=0}^{a-1} P(X = x)) \\ &= a + \sum_{y=0}^{a-1} yP(X = y) - a\sum_{x=0}^{a-1} P(X = x)) \\ &= a - \sum_{y=0}^{a-1} (a - y)P(X = y) \end{aligned}$$

Cost of buying the newspaper for the kiosk: $5a$.
Income from sale and return: $12Y + 3(a - Y)$.
Total earnings: income-cost$=12Y + 3(a - Y) - 5a = 9Y - 2a$

Now to the expected total earnings, $h(a)$.

$$\begin{aligned} h(a) &= \mathrm{E}(9Y - 2a) \\ &= 9[a - \sum_{y=0}^{a-1}(a-y)P(X=y)] - 2a \\ &= 7a - 9\sum_{y=0}^{a-1}(a-y)P(X=y) \end{aligned}$$

This is a function of a, where a is discrete and can take values $0, 1, ldots$. We would like to find the value of a maximizing $h(a)$. Since a is not continuous we can not differensiate $h(a)$ wrt. a, but instead we look at differences between $h(a)$ and $h(a-1)$. If this difference is positive, then $h(a)$ is larger than $h(a-1)$, while when the difference is negative then $h(a-1)$ is larger than $h(a)$.

$$\begin{aligned} h(a) - h(a-1) &= 7a - 9\sum_{y=0}^{a-1}(a-y)P(X=y) - 7(a-1) - 9\sum_{y=0}^{a-2}(a-1-y)P(X=y) \\ &= 7a - 7a + 7 - 9\sum_{y=0}^{a-1}(a-y)P(X=y) + 9\sum_{y=0}^{a-2}(a-1-y)P(X=y) \\ &= 7 - 9\,[aP(X=0) + (a-1)P(X=1) + \cdots \\ &\quad +(a-a+2)P(X=a-2) + (a-a+1)P(X=a-1)] \\ &\quad +9\,[(a-1)P(X=0) + (a-2)P(X=1) + \cdots \\ &\quad +(a-1-a+2)P(X=a-2)] \\ &= 7 - 9\,[(a-a+1)P(X=0) + (a-1-a+2)P(X=1) + \cdots \\ &\quad +(a-a+2-a+1+a-2)P(X=a-2) + (a-a+1)P(X=a-1)] \\ &= 7 - 9\sum_{y=0}^{a-1}P(X=y) = 7 - 9P(X \le (a-1)) \end{aligned}$$

We see that this difference is a function of the cumulative Poisson distribution, which is increasing and monotone. When is $h(a) - h(a-1) \ge 0$?

$$\begin{aligned} h(a) - h(a-1) &\ge 0 \\ 7 - 9P(X \le (a-1)) &\ge 0 \\ P(X \le (a-1)) &\le \frac{7}{9} = 0.778 \end{aligned}$$

Consulting the Poisson cumulative tables on page 20 of "Formelsamlingen", we see that with $\mu = 10$ then $P(X \le 11) = 0.6968$ and $P(X \le 12) = 0.7916$. This means that the

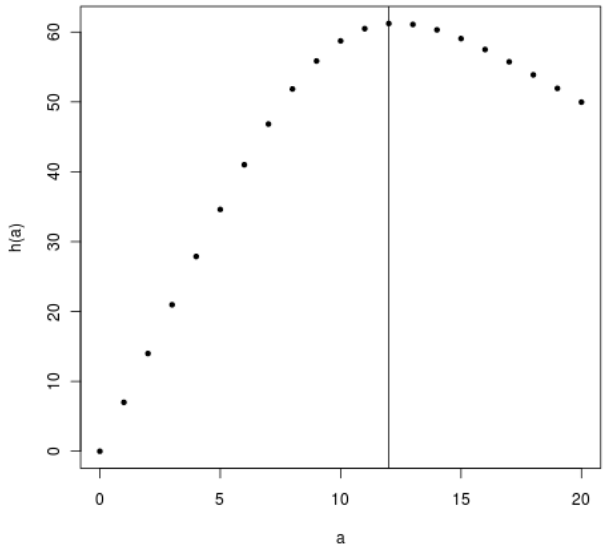
largest a for which $P(X \leq (a-1)) \leq 0.778$ is $a - 1 = 11$, giving $a = 12$ as the value with maximum expected total earnings.

Additional checks (numerical values):

$$\begin{aligned} h(13) - h(12) &= -0.12 \\ h(12) - h(11) &= 0.73 \\ h(11) - h(10) &= 1.75 \end{aligned}$$

so, yes, the maximum must be at $a = 12$.

Additional figure of $h(a)$ that would not be possible to draw when sitting for the exam - but might provide additional insight for the reader.



Problem 3 Covariance

Random variable X has $\mathrm{E}(X) = 10$ and $\mathrm{Var}(X) = 4$, while random variable Y has $\mathrm{E}(Y) = 8$ and $\mathrm{Var}(Y) = 9$. In addition X and Y are dependent variables with covariance $\mathrm{Cov}(X, Y) = 5$.

a)

$$\begin{aligned} \mathrm{E}(2X - Y) &= 2 \cdot \mathrm{E}(X) - \mathrm{E}(Y) = 2 \cdot 10 - 8 = 12 \\ \mathrm{Var}(2X - Y) &= 2^2 \cdot \mathrm{Var}(X) + (-1)^2 \mathrm{Var}(Y) + 2 \cdot 2 \cdot (-1) \mathrm{Cov}(X, Y) \\ &= 4 \cdot \mathrm{Var}(X) + \mathrm{Var}(Y) - 4 \cdot \mathrm{Cov}(X, Y) = 4 \cdot 4 + 9 - 4 \cdot 5 = 5 \end{aligned}$$

Where we have used the formula

$$\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X)+b^2\mathrm{Var}(Y)+2ab\mathrm{Cov}(X,Y)$$

We observe that $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\cdot\mathrm{E}(Y)$, so that $\mathrm{E}(XY) = \mathrm{Cov}(X,Y) + \mathrm{E}(X)\cdot\mathrm{E}(Y) = 5 + 10\cdot 8 = 85$, which we need now:

$$\begin{aligned}\mathrm{E}\left((X-3)(Y-5)\right) &= \mathrm{E}\left(XY-3Y-5X+15\right) = \mathrm{E}(XY)-3\mathrm{E}(Y)-5\mathrm{E}(X)+15\\ &= 85-3\cdot 8-5\cdot 10+15 = 85-24-50+15 = 26\end{aligned}$$

Problem 4 The experimental farm

a) Biomass, Y, is normally distributed (Gaussian) with $\mathrm{E}(Y)=5$ and $\mathrm{Var}(Y)=4$. Calculate probabilites:

$$\begin{aligned}P(Y>6) &= 1-P(Y\le 6) = 1-P(\frac{Y-5}{2}\le\frac{6-5}{2}) = 1-\Phi(0.5)\\ &= 1-0.6915 = 0.31\\ P(4<Y\le 6) &= P(Y\le 6)-P(Y\le 4) = P(\frac{Y-5}{2}\le\frac{6-5}{2})-P(\frac{Y-5}{2}\le\frac{4-5}{2})\\ &= \Phi(0.5)-\Phi(-0.5) = 0.6915-0.3085 = 0.38\\ P(Y>6\mid Y>4) &= \frac{P(Y>6\cap Y>4)}{P(Y>4)} = \frac{P(Y>6)}{P(Y>4)}\\ &= \frac{1-\Phi(0.5)}{1-\Phi(-0.5)} = \frac{1-0.6915}{1-0.3085} = \frac{0.3085}{0.6915} = 0.45\end{aligned}$$

b) Linear regression of biomass Y as a linear function of cultivation period (without intercept), but with variance dependent on cultivation period, x.

$$Y = \beta x + \varepsilon(x) \text{ for } x>0,$$

where $\varepsilon(x)$ is normally distributed (Gaussian) $\mathrm{E}(\varepsilon)=0$ and $\mathrm{Var}(\varepsilon)=\tau^2x^2$.

Measurements for $n=5$ plants at cultivation periods $x_1, x_2, \ldots, x_5$ with biomass $Y_1, Y_2, \ldots, Y_5$.

Distribution of Y_i given x_i:
Since ε_i is normally distributed, given x_i and β is a constant parameter, then Y_i is a linear function of ε_i, and Y_i is thus also normally distributed.

$$\begin{aligned}\mathrm{E}(Y_i) &= \mathrm{E}(\beta x_i+\varepsilon_i) = \beta x_i + \mathrm{E}(\varepsilon_i) = \beta x_i\\ \mathrm{Var}(Y_i) &= \mathrm{Var}(\beta x_i+\varepsilon_i) = 0+\mathrm{Var}(\varepsilon_i) = \tau^2 x_i^2\end{aligned}$$

We choose to derive the maximum likelihood estimators for β og τ^2. Alternatively, the least squares estimator for β might be derived, but since the variances differ between the εs then a weighting factor $\frac{1}{\tau^2 x_i^2}$ should be used when minimizing the sum of squared errors, $SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n \frac{1}{\tau^2 x_i^2}(y_i - \hat{y}_i)^2$. This is called weighted least squares and is not on the reading list of this course. Using the least squares without this weighting, will give partial credit.

$$\begin{aligned}
L(\beta, \tau^2, y_1, y_2, ..., y_n) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \frac{1}{\tau x_i} \exp(-\frac{1}{2\tau^2 x_i^2}(y_i - \beta x_i)^2) \\
\ln L(\beta, \tau^2, y_1, y_2, ..., y_n) &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\tau^2 - \sum_{i=1}^n \ln x_i - \frac{1}{2\tau^2}\sum_{i=1}^2 \left(\frac{y_i - \beta x_i}{x_i}\right)^2 \\
\frac{\partial \ln L(\beta, \tau^2, y_1, y_2, ..., y_n)}{\partial \beta} &= -\frac{1}{2\tau^2}\sum_{i=1}^n 2\left(\frac{y_i - \beta x_i}{x_i}\right)\cdot(-1) \\
&= \frac{1}{\tau^2}\left[\sum_{i=1}^n \frac{y_i}{x_i} - n\beta\right] = 0 \\
\beta &= \frac{1}{n}\sum_{i=1}^n \frac{y_i}{x_i}
\end{aligned}$$

This gives the following estimator, B, for β.

$$B = \frac{1}{n}\sum_{i=1}^n \frac{Y_i}{x_i}$$

We then continue with the estimator for τ^2.

$$\begin{aligned}
\frac{\partial \ln L(\beta, \tau^2, y_1, y_2, ..., y_n)}{\partial \tau^2} &= -\frac{n}{2\tau^2} - \frac{1}{2}\sum_{i=1}^n \left(\frac{y_i - \beta x_i}{x_i}\right)^2 \cdot (-\frac{1}{\tau^4}) \\
&= -\frac{n}{2\tau^4}\left[\tau^2 - \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \beta x_i}{x_i}\right)^2\right] = 0 \\
\tau^2 &= \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \beta x_i}{x_i}\right)^2
\end{aligned}$$

Inserting the estimator, B, for β, this yields the estimator $\hat{\tau}^2$ for τ^2.

$$\hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i - B x_i}{x_i}\right)^2$$

Numerical results using $n = 5$, $\sum_{i=1}^{5} \frac{y_i}{x_i} = 2.61$ and $\sum_{i=1}^{5} \frac{y_i^2}{x_i^2} = 1.59$.

$$\begin{aligned} b &= \frac{2.61}{5} = 0.52 \\ \hat{\tau}^2 &= \frac{1}{5}(\sum_{i=1}^{5} \frac{y_i^2}{x_i^2} - 2 \cdot b \sum_{i=1}^{5} \frac{y_i}{x_i} + 5b^2) = (1.58 - 2 \cdot 0.52 \cdot 2.61 + 5 \cdot 0.52^2)/5 = 0.046 \end{aligned}$$

c) Here we use the estimator

$$B = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{x_i}$$

and assume $\tau^2 = 0.04$ is known.

For the hypotheses

$$H_0 \colon \beta = 0.50 \text{ vs. } H_1 \colon \beta > 0.50$$

we would reject H_0 when $B > k$, where k can be found from

$$P(\text{Reject } H_0 \mid H_0 \text{ true}) \leq 0.05$$

To do this we need the distribution of B. Since B is a linear combination of Y_is (the x_is are constants), and the independent Y_is are normally distributed, then B is also normally distributed with

$$\begin{aligned} \mathrm{E}(B) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \mathrm{E}(Y_i) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} (\beta x_i) = \beta \\ \mathrm{Var}(B) &= \frac{1}{n^2} \sum_{i=1}^{n} (\frac{1}{x_i})^2 \mathrm{Var}(Y_i) = \frac{1}{n^2} \sum_{i=1}^{n} (\frac{1}{x_i})^2 (\tau^2 x_i^2) = \frac{\tau^2}{n} \end{aligned}$$

Back to rejection rule:

$$\begin{aligned} P(B > k \mid \beta_0 = 0.50) &\leq 0.05 \\ P(\frac{B - \beta_0}{\sqrt{\frac{\tau^2}{n}}} > \frac{k - \beta_0}{\sqrt{\frac{\tau^2}{n}}} \mid \beta_0 = 0.50) &\leq 0.05 \end{aligned}$$

In the standard normal distribution the value z_α has area α to the right (greater than), and thus $\frac{k-\beta_0}{\sqrt{\frac{\tau^2}{n}}} = z_\alpha$. Solving for k gives $k = \beta_0 + z_\alpha \sqrt{\frac{\tau^2}{n}}$ (where $\beta_0 = 0.50$). This means that the rejection rule is:
Reject H_0 when $B > \beta_0 + z_\alpha \sqrt{\frac{\tau^2}{n}}$.

Using the numerical values in the example we have

- $\beta_0 = 0.5$
- $z_{0.05} = 1.645$
- $n = 5$
- $\frac{\tau}{\sqrt{n}} = 0.089$
- $\beta_0 + z_\alpha\sqrt{\frac{\tau^2}{n}} = 0.5 + 1.645 \cdot 0.089 = 0.647$
- $B = 0.52$

Since $B = 0.52 < 0.647$ we do not reject H_0. What we have observered is not in contradiction to H_0 being true.

Additional: p-value (not asked for):

$$\begin{aligned} P(B \geq 0.52 \mid \beta_0 = 0.5) &= P(\frac{B - \beta_0}{\sqrt{\frac{\tau^2}{n}}} \geq \frac{0.52 - 0.5}{\sqrt{\frac{0.04}{5}}} \mid \beta_0 = 0.5) \\ &= P(Z \geq 0.22) = 1 - \Phi(0.22) = 1 - 0.5871 = 0.41 \end{aligned}$$

The power of the test is the probability of rejecting H_0 at a given value under H_1, here at $\beta = 0.7$.

$$\begin{aligned} \text{Power} &= P(\text{Reject } H_0 \mid \text{ true value is } \beta = 0.7) \\ &= P(B > k \mid \beta = 0.7) = P(B > \beta_0 + z_\alpha\sqrt{\frac{\tau^2}{n}} \mid \beta = 0.7) \\ &= P(\frac{B - \beta}{\sqrt{\frac{\tau^2}{n}}} > \frac{\beta_0 - \beta}{\sqrt{\frac{\tau^2}{n}}} + z_\alpha \mid \beta = 0.7) \\ &= P(Z > \frac{\beta_0 - \beta}{\sqrt{\frac{\tau^2}{n}}} + z_\alpha \mid \beta = 0.7) \\ &= 1 - \Phi(\frac{\beta_0 - \beta}{\sqrt{\frac{\tau^2}{n}}} + z_\alpha) \end{aligned}$$

With $\beta = 0.7$ and $n = 5$:

$$\text{Power} = 1 - \Phi(-2.236 + 1.645) = 1 - \Phi(-0.59) = 1 - 0.2776 = 0.72$$

For $n = 5$ the power at $\beta = 0.7$ is 0.72. This means that with our test with $n = 5$ observations the probability of rejecting H_0 when the true $\beta = 0.7$ is 0.72.