HG
Nov. 14

ECON 4130 H14

Extra exercises for no-seminar week 46

(Solutions will be put on the net Thursday 13 nov.)

Exercise A

(*Based on* ***Exam 2010H -postponed.***)

Problem 1

Introduction. When counting the number of members in groups such as the number of bacteria per colony, the number of people per household, or the number of animals per litter, every single count must necessarily be larger or equal to 1. This excludes, e.g., the poisson distribution as a model for such counts. On the other hand the so-called *logarithmic series* distribution often proves useful:

A discrete random variable (rv), Y, taking values in $\{1,2,3,\ldots\}$, is said to be *logarithmic series* distributed if the probability mass function (pmf) is

$$(1) \qquad P(Y=y) = f(y;\theta) = -\frac{1}{\ln(1-\theta)}\cdot\frac{\theta^y}{y}, \qquad y = 1,2,3,\ldots$$

where θ is a parameter such that $0<\theta<1$.

A. Calculate $P(Y=1)$ and $P(Y\geq 2)$ when $\theta = 0.5$.

B. Show that the expected value of Y is

$$E(Y) = c\frac{\theta}{1-\theta} \quad \text{where} \quad c = -\frac{1}{\ln(1-\theta)}$$

[**Hint:** You are reminded of the sum of a geometric series $\sum_{i=0}^{\infty} a^i = 1 + a + a^2 + a^3 + \cdots = \frac{1}{1-a}$ which is valid if $|a|<1$]

C. **(i)** Show that the moment generating function (mgf) for Y is given by

$M(t) = -c\ln(1-\theta e^t)$, where c is as given in section **B**. Explain why $M(t)$ is well defined in an open interval around 0.

[Hint: You may need the following result (which you do not need to prove) from the theory of series:

$$\sum_{i=1}^{\infty}\frac{a^i}{i} = a + \frac{a^2}{2} + \frac{a^3}{3} + \cdots = -\ln(1-a) \text{ whenever } |a|<1 \text{]}$$

(ii) Find $\mathrm{Var}(Y)$ as a function of θ.

D. The data in table 1[1] are the result of investigating the number of bacteria in each of 675 colonies of a certain type of soil bacteria. For example, the table shows that 146 of the colonies consisted of 2 bacteria, and 359 colonies had one bacterium only.

Table 1 Frequency table of the size of colonies

Bacteria per colony (j)	1	2	3	4	5	6	7	Sum
Number of colonies observed f_j	359	146	57	41	26	17	29	675

To ease the calculations we give: $\sum_{j=1}^{7} j \cdot f_j = 1421$

Let Y_i denote the number of bacteria in colony *i*. Assume that $Y_1, Y_2, \ldots, Y_n$ (where $n = 675$) are independent and identically distributed (iid) where Y_i is logarithmic series distributed with unknown parameter θ.

(i) Show that the maximum likelihood estimator (mle) of θ, $\hat{\theta}$, satisfies the equation

[1] *Source:* C.A.Bliss and R.A.Fisher, "Fitting the Negative Binomial Distribution to Biological Data", *Biometrics* 9 (1953): 176-200.

(2) $$\overline{Y} = \frac{\hat{\theta}}{-(1-\hat{\theta})\ln(1-\hat{\theta})}, \quad \text{where} \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

[**Hint:** You can skip proving that the solution of (2) actually maximizes the log likelihood since the second derivative of the log likelihood is slightly complicated here.]

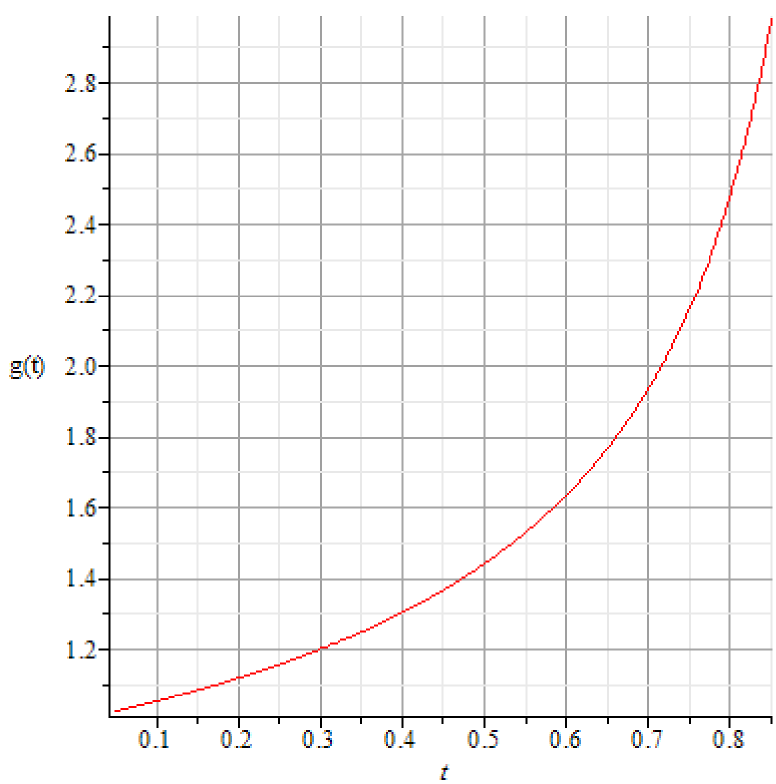
(ii) Explain why the moment method estimator (mme) and the mle are the same in this case.

E. **(i)** Introduce the function $g(t)$ defined by

$$g(t) = \frac{t}{-(1-t)\ln(1-t)} \quad \text{for} \quad 0 < t < 1$$

In **figure 1** we have plotted the function, $g(t)$, for $0.05 < t < 0.9$

Figure 1



Use the plot to obtain an approximate (rough) value for the mle of θ based on the data at hand.

(ii) We are also interested in the mean colony size in the population, i.e., $E(Y)$. Explain why the maximum likelihood estimator for $E(Y)$, based on the proposed model, is simply $\hat{E}(Y) = \bar{Y}$. Present your estimate, based on the data in table 1, of the mean number of bacteria per colony for this population.

Problem 2

There is a complication with the data in table 1 in Problem 1. At the time when the data were collected (before 1953) it turned out difficult to count colonies with more than 6 bacteria. Hence all the 29 colonies registered as having 7 bacteria in table 1 should, more correctly, be registered as having 7 *or more* bacteria. The correct heading in the table should have been ≥ 7 instead of only 7. In this problem we will try to accommodate this complication.

A. The complication described above implies that the observed values of $\bar{Y}$ and the mle, $\hat{\theta}$, found in Problem 1 are not entirely correct. Are the true values of $\bar{Y}$ and $\hat{\theta}$ larger or smaller than the values found in Problem 1 when replacing the count 7 by ≥ 7? Give a reason for your answer.

[**Hint:** You can take for granted that the function $g(t)$ is strictly increasing for $0 < t < 1$. Confer also figure 1 in Problem 1.]

B. **Introduction.** As a result of replacing the count 7 by ≥ 7 the true value of $\bar{Y}$ cannot be calculated. But we can still estimate $E(Y)$ credibly if our chosen model for the distribution of *Y* is realistic. A possible test of the realism of our model is the well known Pearson χ^2 - test. The H_0 hypothesis states that the model in Problem 1 is true (i.e., $Y_1, Y_2, \ldots, Y_n$ (where $n = 675$) are *iid* where Y_i is logarithmic series distributed with unknown parameter θ). Under H_0 the frequencies in table 1 are then multinomially distributed with 7 categories, $\{1\}, \{2\}, \ldots, \{6\}, \{\geq 7\}$ and respective probabilities, $p_j(\theta)$ for $j = 1, 2, \ldots, 7$, where

$$p_j(\theta) = \begin{cases} P(Y = j) = -\dfrac{1}{\ln(1-\theta)} \cdot \dfrac{\theta^j}{j} & \text{for } j = 1, 2, \ldots, 6 \\ P(Y \geq 7) = 1 - p_1(\theta) - p_2(\theta) - \cdots - p_6(\theta) & \text{for } j = 7 \end{cases}$$

Based on this specification we can determine a corrected mle, the observed value of which turns out to be $\hat{\theta}_{obs} = 0.7569$ (you do not need to prove this).

(Note that the index *obs* stands for the observed value of the random variable in focus.)

Question. Perform the χ^2-test and formulate a conclusion based on 10% level of significance. Some of the calculations necessary have been done in table 2. Complete the table by filling in the cells with question mark.

Table 2 Partial table of quantities underlying the Pearson χ^2-test based on the corrected mle $\hat{\theta}_{obs} = 0.7569$[2]

Category j	Observed frequency (O_j)	Mle under H_0 ($p_j(\hat{\theta})$)	Estimated frequency under H_0 (E_j)	$\frac{(O_j - E_j)^2}{E_j}$
1	359	0.535	361.23	0.01
2	146	0.203	136.71	0.63
3	57	0.102	68.99	2.08
4	41	0.058	39.17	0.09
5	26	0.035	23.72	0.22
6	17	?	?	?
7	29	?	?	?
Sum	675	1	675	?

[2] The index *obs* stands for the observed value of the random variable in focus.

C. **(i)** Explain why the likelihood function for θ used in Problem 1 cannot be used for the corrected data where category 7 is replaced by ≥ 7.

(ii) Assuming H_0 in section **B** to be true, set up an expression for the log likelihood function for θ based on derived multinomial model in section **B**.

D. **(i)** Maximizing the log likelihood based on the given data, which requires iterations, gives the maximum likelihood estimate, $\hat{\theta}_{obs} = 0.7569$ (which you don't have to show here). Use this to compute a corrected estimate of the mean population colony size, $E(Y)$ from Problem 1. Compare with the corresponding estimate in Problem 1.

(ii) It turns out (you don't need to show this) that the Fisher information for one observation (trial) (out of $n = 675$ trials) in this (multinomial) model is

$I(\theta) = \sum_{j=1}^{7} \frac{\left(p_j'(\theta)\right)^2}{p_j(\theta)}$, where $p_j'(\theta)$ is the derivative with respect to θ of $p_j(\theta)$.

It turns out that the estimated value of $I(\theta)$ is

$$I(\hat{\theta}_{obs}) = I(0.7569) = 6.6076$$

Use this information to calculate an approximate 95% confidence interval (CI) for θ and justify the interval from general maximum likelihood theory and Slutsky's lemma.

(iii) Determine approximate 95% confidence limits for the population mean $E(Y)$ based on figure 1 in Problem 1.

Exercise B

Problem 2 from postponed exam 2005H

a. Let Z be chi-square distributed with *r* degrees of freedom (written in short $Z \sim \chi^2(r)$). According to the textbook, this is the same as the gamma distribution with parameters $\alpha = r/2$ and $\lambda = 1/2$ (in short $Z \sim \Gamma\left(\frac{r}{2}, \frac{1}{2}\right)$). Use this and known properties of the gamma distribution to show that

(i) $\mathrm{E}(Z) = r, \qquad \mathrm{Var}(Z) = 2r$

(ii) For *r* = 2, find the median of *Z*. Is the median smaller or larger than E(*Z*)?

b. According to the textbook, if X is standard normally distributed ($X \sim N(0,1)$), then $X^2 \sim \chi^2(1)$.

(i) Use this and known properties of the gamma distribution to show that, if $X_1, X_2, \ldots, X_n$ are *iid* with $X_i \sim N(0,1)$, then $Z = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$. (*Note that iid means "independent and identically distributed"*.)

(ii) Use for example the central limit theorem to justify that $Z = \sum_{i=1}^{n} X_i^2$ is approximately $N(n, 2n)$ distributed for large *n*.

c. Let q_n denote the 95% quantile (i.e., such that $P(Z \le q_n) = 0.95$) in the $\chi^2(n)$ distribution. Exact values for q_n can be found in the χ^2 - table in Rice. Let q_n^* denote the approximate 95% quantile determined by the approximate distribution for *Z* derived in section **b.** (i.e., $N(n, 2n)$). Show that

$$q_n^* = n + \sqrt{2n} \cdot 1{,}64$$

Calculate the approximation error $q_n - q_n^*$ for $n = 30, 60, 120$, and comment on the result.

d. The approximate quantile in section **c.** can be improved somewhat. Let $X_1, X_2, \ldots, X_n$ ~ $N(0, 1)$ as in section **b.**, and put

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Then, by the central limit theorem combined with some other theory, it can be proven that $\sqrt{2n}\left(\sqrt{\bar{Z}} - 1\right) \xrightarrow{D} N(0,1)$ as $n \to \infty$ (you do not need to justify this here). Show that, based on the approximate normal distribution for $\sqrt{\bar{Z}}$, we can derive an alternative approximation, q_n^{**}, to q_n (i.e., the 95% quantile of $Z = n\bar{Z}$), given by

$$q_n^{**} = \frac{1}{2}\left(1.64 + \sqrt{2n}\right)^2.$$

Compare q_n^{**} with q_n^{*} .for n = 30, 60, 120.

[*Note that this approximation is similar but slightly different from the suggestion given in the* χ^2*-table in Rice, which represents a third approximation*.]

Exercise C

(A slightly altered version of Problem 1 from Exam 2005H –postponed.)

a. Let X and Y be two continuous random variables, both varying in the interval [0, 1]. The joint cumulative probability function (cdf) of X and Y is given by

$$F(x, y) = xy[1 + (0.8)(1-x)(1-y)] \qquad \text{for} \quad 0 \le x \le 1 \quad \text{and} \quad 0 \le y \le 1$$

Show that the joint density function for X and Y is given by

$$f(x, y) = 1 + (0.8)(1-2x)(1-2y) \qquad \text{for} \quad 0 \le x \le 1 \quad \text{og} \quad 0 \le y \le 1.$$

b. Show that both X and Y marginally are uniformly distributed over [0, 1]. Find the expectation and variance for X and Y. Are X and Y stochastically independent? Calculate $P(Y \le 0.5)$ and $P(Y \le 0.5 \mid X \le 0.5)$.

c. Sketch the conditional density for Y, $f(y \mid x)$, in a graph when x = 1/2, and

$x = 1$ respectively (i.e. altogether two graphs).

Calculate the regression function, $\mathrm{E}(Y \mid x)$, and sketch a graph of this.

d. Find $\mathrm{E}(XY)$ by the "double expectation" theorem, i.e.,

$$\mathrm{E}(XY) = \mathrm{E}[\mathrm{E}(XY \mid X)] = \mathrm{E}[X\ \mathrm{E}(Y \mid X)] \text{ etc.}$$

Calculate the correlation coefficient between X and Y.