# *UNIVERSITY OF OSLO*
# *DEPARTMENT OF ECONOMICS*

Exam: **ECON4130 – Statistics 2**

Date of exam: Monday, December 10, 2018 **Grades are given:** January 4, 2019

Time for exam: 14.30 – 17.30

The problem set covers 6 pages (incl. cover sheet)

Resources allowed:

- All written and printed resources – as well as two alternative calculators: Aurora HC106 or Casio FX-85EX - are allowed

The grades given: A-F, with A as the best and E as the weakest passing grade. F is fail.

# Econ 4130 Regular Exam 2018H

## Problem 1

**A.** Let $X$ be a continuous random variable (rv) with probability density function (pdf)

$$f_X(x) = \begin{cases} 2 \cdot (1-x) & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

**i)** Show that the cumulative distribution function (cdf) of $X$, $F_X(x) = P(X \le x)$ is equal to $2x - x^2$ when $0 < x < 1$.

**ii)** Sketch a graph of $F_X(x)$ when $-1 \le x \le 2$.

**iii)** Find $P(X \ge 0.5)$ and $P(0.5 \le X \le 2)$.

**B.**

**i)** Show that $E(X) = \frac{1}{3}$ and $\text{Var}(X) = \frac{1}{18}$.

**ii)** Show that the quantile function, $q_p$, for $X$ (i.e., where $q_p$ by definition solves the equation $P(X \le q_p) = p$ for $0 < p < 1$), is $q_p = 1 - \sqrt{1-p}$ for $0 < p < 1$.

[**Hint.** You are reminded that the solution of a second degree equation, $ax^2 + bx + c = 0$, is $x = \frac{1}{2a}\left(-b \pm \sqrt{b^2 - 4ac}\right)$ if $b^2 \ge 4ac$. ]

**iii)** Suppose that observations of $X$ are small relative frequencies. Often relative frequencies are presented as percentages in reports. Find the pdf of the percentage, $Z = 100X$.

**C.** Let $Y$ be another rv such that the conditional distribution of $Y$ given that $X = x$ is fixed, is normal with expectation $2x$ and variance $x$ (written in short $Y \mid x \sim N(2x, x)$ ). In addition, let the rv $X$ be distributed as in section **A**.

**i)** Write up a formula for the joint pdf of the random pair, $(X, Y)$.

**ii)** Explain why $(X, Y)$ is *not* bivariate normally (i.e., two-dimensional normally) distributed.

**iii)** Let $R = \frac{Y}{\sqrt{X}}$. Find the regression function, $E(R \mid x)$, of $R$ with respect to $X$. Is the regression homoscedastic or heteroskedastic? Give a reason for your answer.

[**Hint.** Note that $R = \frac{Y}{\sqrt{x}}$ when $X$ is fixed to the value $x$.]

**D.** **i)** Find $E(Y)$ and $\mathrm{Var}(Y)$.

**ii)** Find the correlation coefficient, $\rho = \rho(X, Y)$, between $X$ and $Y$.

**E.** We want to simulate 3 independent observation pairs of $(X, Y)$, i.e., $(x_1, y_1),\ (x_2, y_2),\ (x_3, y_3)$.

**i)** Let 0.88, 0.11, 0.39 be 3 independent draws from the standard uniform distribution (i.e., the uniform distribution over the interval [0,1]). Transform these numbers into 3 independent observations of $X$, called $x_1, x_2, x_3$.

**ii)** In addition we let Stata produce (simulate) 3 independent draws from the standard normal distribution, $N(0,\ 1)$, with the result, 0.20, $-1.90$, 0.46. Transform these numbers into 3 independent observations, $y_1, y_2, y_3$ of $Y$ in such a way that $(x_i, y_i),\ i = 1, 2, 3$ can be considered as 3 independent observation pairs of $(X, Y)$.

**[Hint.** It may be helpful to utilize the fact that if $W \sim N(a, b)$, then $V = \frac{W - a}{\sqrt{b}} \sim N(0,1)$. Then express $W$ by $V$. ]

**iii)** What statistical model would you suggest for the data set represented by the 3 pairs of numbers, $(x_i, y_i),\ i = 1, 2, 3$?

## Problem 2

**Introduction.** Table 1 shows prevalence rates[1] of a certain disease in $n = 50$ cities in a country (a certain point in time). The rates are percentages so that, for example, in city no. 2, 12.8% of the risk population has the disease. The relative frequency in city no. 2 of people with the disease is obtained by dividing by 100 (e.g., the relative frequency in city no. 2, is 0.128). The model used will be dealing with relative frequencies rather than with percentages.

[1] Simulated data.

**Table 1** Prevalence rates in 50 cities.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 12.8 | 13.9 | 1.3 | 9 | 8.2 | 0.8 | 4.3 | 0.9 | 8.4 |
| 10 | 3 | 1.3 | 6 | 2.3 | 1 | 4.7 | 4.9 | 0.9 | 4 |
| 21.8 | 7.2 | 8.9 | 9.6 | 9.6 | 0.1 | 7 | 9.7 | 0.6 | 10.2 |
| 4.7 | 2.5 | 8.1 | 3.3 | 4.7 | 3.8 | 0.1 | 0.3 | 6.8 | 16.1 |
| 1.9 | 1.3 | 3.3 | 3.8 | 7.8 | 1.7 | 16.1 | 1.3 | 0.7 | 2.9 |

**Some intermediate results:** If, for city no. *i*, $z_i$ is the percentage given in **table 1** and $x_i = z_i/100$ the relative frequency, $i = 1, 2, \ldots, n = 50$, we have

$$\sum_1^{50} x_i = 2.737 \quad \text{and} \quad -\sum_1^{50} \ln(1 - x_i) = 2.8829$$

**(1) Model.** Let the rv $X_i$ represent the relative frequency of the disease in city no. *i*. Assume that $X_1, X_2, \ldots, X_n$ $(n = 50)$, are independent and identically distributed (*iid*) with the common pdf

$$f(x;\alpha) = \begin{cases} (\alpha+1)(1-x)^{\alpha} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha > -1$ is a parameter.

**Questions.**

**A.**

**i)** Show that the maximum likelihood estimator (mle) of $\alpha$ is

$$\hat{\alpha} = -1 - \frac{n}{\sum_{i=1}^{n} \ln(1 - X_i)}$$

Calculate the estimate of $\alpha$ based on the data in **table 1**.

**ii)** According to general mle theory, $\sqrt{nI(\alpha)}(\hat{\alpha} - \alpha)$ converges in distribution to a normal distribution as *n* increases, i.e., $\sqrt{nI(\alpha)}(\hat{\alpha} - \alpha) \underset{n\to\infty}{\overset{D}{\to}} N(0,1)$ where $I(\alpha)$ is the Fisher information for one observation. Or, in other words, $\sqrt{n}(\hat{\alpha} - \alpha) \underset{n\to\infty}{\overset{D}{\to}} N(0, b(\alpha))$, where $b(\alpha)$ is a function of $\alpha$. Find $b(\alpha)$ as a function of $\alpha$.

**iii)** It follows that, for fixed large *n*, $\hat{\alpha}$ is approximately normally distributed. Write up the parameters (i.e., the expectation and variance) of this normal distribution.

**B.** **i)** Construct an approximately 95% confidence interval (CI) for $\alpha$ based on the normal approximation in section **A**. If you need Slutsky's lemma in this construction, explain how it is used. Calculate the observed CI based on the data in **table 1**.

**ii)** Test $H_0 : \alpha = 20$ versus $H_1 : \alpha \neq 20$ at the level of significance approximately 5%, using the data in **table 1** and formulate a conclusion.

**C.** **i)** Show that the cdf of $X_i$ under model (1) is

$$F_{X_i}(x) = P(X_i \leq x) = \begin{cases} 1-(1-x)^{\alpha+1} & \text{for } 0 < x < 1 \\ 0 & \text{for } x \leq 0 \\ 1 & \text{for } x \geq 1 \end{cases}$$

**ii)** We will investigate if there is evidence in the data against the model (1) and will use the Pearson Chi-square test for this. In **table 2** we have divided the interval $(0,1)$ into 6 suitable intervals of various length as indicated in the table. Let $Z_j$ be the number out of 50 relative frequencies, observed as in **table 1,** that fall in the interval *j*, *j*=1,2,…,6. We assume that $Z_1, Z_2, \ldots, Z_6$ are multinomially distributed with parameters $n = 50$, $p_1, p_2, \ldots, p_6$, and where $p_j$ denotes the probability that an observed relative frequency falls in interval *j*.

The null-hypothesis is that $X_1, X_2, \ldots, X_n$ satisfy model (1) with $f(x;\alpha)$ as the common pdf.

**Table 2**

| *j* | Interval | Observed frequencies $O_j = Z_j$ | Expected frequencies under $H_0$ $E_j$ | $\frac{(O_j - E_j)^2}{E_j}$ |
|---|---|---|---|---|
| 1 | $\leq 0.01$ | 10 | 7.6 | 0.758 |
| 2 | 0.01 - 0.02 | 6 | 6.5 | 0.038 |
| 3 | 0.02 - 0.04 | 9 | 10.3 | 0.164 |

| | | | | |
|---|---|---|---|---|
| 4 | 0.04 - 0.06 | 6 | 7.5 | 0.300 |
| 5 | 0.06 - 0.09 | 9 | 7.5 | 0.300 |
| 6 | > 0.09 | 10 | 10.7 | 0.046 |

(The upper limits in each interval are inclusive, and the lower limits exclusive)

Explain how the expected frequencies under $H_0$ are determined and describe the structure implied by $H_0$ on the interval probabilities, $p_1, p_2, \ldots, p_6$.

**iii)** Perform the chi-square test, based on **table 2,** choosing the level of significance approximately 5%. Comment on the result.

**D.** **i)** Assume the model (1) to be true. Show that the expected value, $\mu$, of $X_i$ is

$$\mu = E(X_i) = \frac{1}{\alpha + 2}$$

**[Hint.** Find first $E(1 - X_i)$. **]**

**ii)** Derive the moment method estimator (mme), $\tilde{\alpha}$, of $\alpha$ based on $X_1, X_2, \ldots, X_n$ and calculate the estimate, $\tilde{\alpha}_{obs}$. Why is $\tilde{\alpha}$ consistent?

**iii)** Derive and calculate an approximately 95% CI for $\alpha$ based on the central limit theorem (CLT) applied to $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. It is useful to know that the variance, $\sigma^2$, of $X_i$, in this case is a function of $\alpha$ (that you do not need to prove here) given by,

$$\mathrm{Var}(X_i) = \sigma^2 = \sigma^2(\alpha) = \frac{\alpha + 1}{\alpha + 3} - \left[\frac{\alpha + 1}{\alpha + 2}\right]^2$$

[**Hint.** The construction of the CI for $\alpha$ may be done in two steps. Find first a CI for $\mu = E(X_i)$, using the CLT combined with a consistent estimate of the standard error. Then transfer this CI for $\mu$ to a CI for $\alpha$ with the same degree of confidence, utilizing that $\mu$ is a decreasing function of $\alpha$.]