Econ 4130 Regular Exam 2020H

Problem 1

A. Let the random variable (rv), X, be continuous and uniformly distributed over the interval $(0, m)$, excluding the end points, where m is a positive constant ($m > 0$).

(i) Show that $U = \frac{X}{m}$ is uniformly distributed over the unit interval (0, 1) (i.e., standard uniformly distributed).

(ii) Use the well-known fact that $E(U) = \frac{1}{2}$ and $\text{var}(U) = \frac{1}{12}$ to show that

$$E(X) = \frac{m}{2}, \quad E(X^2) = \frac{m^2}{3}, \text{ and } \text{var}(X) = \frac{m^2}{12}.$$

(iii) Find the probability $P\left(\frac{m}{4} < X < \frac{3m}{4}\right)$.

B. In addition to X from section **A**, we have another rv, Y, that is such that, given that X is fixed to the positive value $x < m$, then the distribution of $(Y \mid X = x)$ is exponential with parameter $\lambda = \frac{m}{x}$, and cumulative distribution function (cdf)

$$F(y \mid x) = P(Y \le y \mid X = x) = \begin{cases} 1 - e^{-\frac{m}{x}y} & \text{for } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

(i) Set up the regression function ($E(Y \mid x)$) for Y with respect to X and also the conditional variance of Y given that $X = x$. Is the regression linear or non-linear? Is it homoscedastic or heteroscedastic? For which x's is the regression function well defined?

[**Hint:** You may use formulas for expectation and variance in an exponential distribution without having to justify them.]

(ii) Show that $E(Y) = \frac{1}{2}$ and $\text{var}(Y) = \frac{5}{12}$.

C. (i) Find the covariance, $\text{cov}(X,Y)$, and show that the correlation coefficient between X and Y is $\frac{1}{\sqrt{5}}$.

[**Hint:** Remember that $E(XY) = E\left[X \cdot E(Y \mid X)\right]$, a fact you do not need to justify here.]

(ii) Show that $V = \frac{Y}{X}$ and X are stochastically independent.

[**Hint:** Determine the conditional distribution of V given that $X = x$ is fixed. Remember that if $X = x$ is fixed, the distributions of $\frac{Y}{X}$ and $\frac{Y}{x}$ must be the same.]

D. Let 0.21 and 0.56 be two independent draws from a standard uniformly distributed rv, $U \sim \text{uniform}(0,1)$. Transform these two observations to an observation pair, (x, y), of (X,Y), where (X,Y) has joint distribution as in section **B** with $m = 10$.

[**Hint.** Remember that, if $W \sim \exp(\lambda)$, the quantile function is $q(p) = -\frac{1}{\lambda}\ln(1-p)$, a fact you do not have to prove here. The quantile function satisfies $P\left(W \le q(p)\right) = p$.]

Problem 2

A. Let the random variable (rv) U be gamma distributed with shape parameter α and scale parameter λ (in short: $U \sim \Gamma(\alpha,\lambda)$).

Assume that $\alpha > 2$. Let $\Gamma(t)$ denote the gamma function.

Use the formula (derived in the lectures), $E\left(U^r\right) = \frac{\Gamma(\alpha+r)}{\lambda^r\Gamma(\alpha)}$, which holds for any real $r > -\alpha$, to show that

$$E\left(\frac{1}{U}\right) = \frac{\lambda}{\alpha-1} \quad \text{and} \quad \text{var}\left(\frac{1}{U}\right) = \frac{\lambda^2}{(\alpha-1)^2(\alpha-2)}$$

B. Let X be the income of a randomly chosen person from a population of people with higher income - defined as incomes higher than a cut-off value *b*.
Assume that X is Pareto distributed (b,θ) with cdf

$$F_X(x) = P(X \le x) = \begin{cases} 1-\left(\frac{b}{x}\right)^{\theta} & \text{for } x > b \\ 0 & \text{for } x \le b \end{cases}$$

where $\theta > 0$ is a parameter, and *b* a known cut-off value.

(i) We have observations of $X_1, X_2, \ldots, X_n$ that are independent and identically distributed (*iid*), all with the same distribution as X. From the lectures we know that the maximum likelihood estimator (mle) of θ based on $X_1, X_2, \ldots, X_n$, is

$\hat{\theta} = \frac{n}{Y}$, where $Y = \sum_{i=1}^{n} \ln \frac{X_i}{b}$ is $\Gamma(n,\theta)$ distributed, and where $Y_i = \ln \frac{X_i}{b}$ is exponentially distributed with parameter θ.

Explain briefly why Y is $\Gamma(n,\theta)$ distributed. Show also that the mle of θ based on $Y_1, Y_2, \ldots, Y_n$ (instead of on $X_1, X_2, \ldots, X_n$) is equal to the given $\hat{\theta}$.

(ii) Show that $E(\hat{\theta}) = c_n\theta$, where c_n is a constant depending on *n*. Find c_n and use this to set up an unbiased estimator, $\tilde{\theta}$, for θ, obtained as a modification of $\hat{\theta}$ (i.e., by multiplying $\hat{\theta}$ by a suitable constant).

(iii) Show that $MSE(\hat{\theta}) = \frac{n+2}{(n-1)(n-2)}\theta^2$, where the mean squared error (*MSE*) is defined as $MSE(\hat{\theta}) = E\left[(\hat{\theta}-\theta)^2\right]$. Compare $\hat{\theta}$ and $\tilde{\theta}$ by their mean squared errors, and state which one of the two is the most preferable estimator based on the *MSE*s.
The data used in the lectures were $n = 2361$ incomes above $b = 250\,000$ for females in Norway in 1998. The mle estimate was $\hat{\theta}_{obs} = 3.813$. Calculate the estimate from the unbiased estimator and compare.

[**Hint.** You may use, without having to show it, that
$\text{var}\left(\hat{\theta}\right) = \frac{n^2\theta^2}{(n-1)^2(n-2)}$ and the fact that
$E\left[(W-c)^2\right] = \text{var}(W) + \left(E(W)-c\right)^2$, where *W* is a rv and *c* a constant. In addition, you may need that
$n^2 + n - 2 = (n-1)(n+2)$.]

C. Suppose that X is Pareto distributed (b,θ) as in section **B.** Let V be the income of a randomly chosen individual from the subpopulation with incomes larger than c, where $c > b$. We then may assume that V is distributed as X when $X > c$.

Show that $V \sim \text{Pareto}(c,\theta)$ with the same θ as in the distribution of X.

[**Hint.** Find $P(V \le v)$ for $v > c$, noting that $P(V \le v) = P(X \le v \mid X > c)$.]

D. **Introduction.** In the lectures a sample of $n = 2361$ women from Norway in 1998, all with income larger than $b = 250\,000$ NOK, was analyzed based on the Pareto model.

Let *group I* denote women with income larger than 250 000 NOK ($n = 2361$ in the sample), and *group II* women with income larger than 350 000 NOK ($n = 556$ in the sample).

Using the Pareto model in section **B**, the mle estimate of θ was calculated as $\hat{\theta}_{I,obs} = 3.813$ for group I and $\hat{\theta}_{II,obs} = 2.869$ for group II.

We want to test the Pareto assumption using the Pearson chi-square test for the two groups. We then divide the income scale from $b = 250\,000$ and upwards into 14 intervals (categories) and count the frequencies in each interval as indicated in **table 1**.

Table 1 shows only the upper limit in each interval. Thus, the first interval for group I contains incomes from 250 000 to 275 000. (In all intervals the upper limit is inclusive and the lower limit exclusive.) The number of incomes (i.e., the frequency) in the first interval was 759 in the data, according to the table.
The second interval goes from 275 000 to 300 000 with frequency 532. And so on. The last interval comprises incomes larger than 1 000 000 (frequency 30).

For group II the first interval goes from 350 000 to 400 000 (with frequency 198), and so on.

We assume $X_1, X_2, \ldots, X_n$ are iid for both groups where $n = 2361$ for group I and $n = 556$ for group II. The hypothesis, H_0, to be tested is that $X_i \sim$ Pareto(b,θ), where $b = 250\,000$ for group I and $b = 350\,000$ for group II.
Let O_j be the observed frequency and E_j the expected frequency under H_0 for category *j*. In addition, we write $Q_j = \dfrac{(O_j - E_j)^2}{E_j}$ in the table.

(End of introduction.)

Table 1

	Group I	Cut-off $b = 250\,000$				Group II	Cut-off $b = 350\,000$	
Category	Upper interval limit	Frequency O_j	E_j	Q_j		Frequency O_j	E_j	Q_j
1	275 000	759	719.41	2.18				
2	300 000	532	463.51	10.12				
3	325 000	298	309.87	0.45				
4	350 000	216	213.72	0.02				
5	400 000	198	261.14	15.27		198	176.95	2.50
6	450 000	107	142.32	8.77		107	108.69	0.03
7	500 000	72	83.06	1.47		72	70.53	0.03
8	550 000	37	51.19	3.93		37	47.81	2.44
9	600 000	34	32.98	0.03		34	33.58	0.01
10	650 000	17	22.05	1.16		17	24.3	2.19
11	700 000	19	15.21	0.94		19	18.03	0.05
12	800 000	25	18.58	2.22		25	24.22	0.03
13	1 000 000	17	16.04	0.06		17	?	?
14	Larger than 1 000 000	30	11.95	27.26		30	?	?
	Sum	2361		73.88		556		9.84

Questions.

(i) Calculate numbers to fill in the 4 cells with question marks in the table. Explain your calculations.

(ii) Perform the Pearson chi-square test both for group I and for group II. Use the level of significance 1%. (Note that the total sum of Q_j for group II is a sum over all 10 categories.).

(iii) Is the p-value of the test for group II smaller than 10%? Give a reason for your answer.

E. Let $X_1, X_2, \ldots, X_n$ be iid such that $X_i \sim \text{Pareto}(b, \theta)$ distributed as in section **B**, and let $Y = \sum_{i=1}^{n} \ln \frac{X_i}{b} = \sum_{i=1}^{n} Y_i$, where $Y_i = \ln \frac{X_i}{b}$.

(i) Use the moment generating function (mgf) of the gamma distribution to show that $2\theta Y$ is chi square distributed with $2n$ degrees of freedom (in short, $2\theta Y \sim \chi^2_{2n}$).

[Hint. Remember that the chi square distribution with *r* degrees of freedom is the same as the $\Gamma\left(\frac{r}{2}, \frac{1}{2}\right)$ distribution.]

(ii) Use **(i)** to derive a formula for a 95% confidence interval (CI) for θ based on Y from group II. Calculate the CI in group II when the observed value in group II of Y is $Y_{obs} = 193.7957$ (with $n = 556$).

[**Hint**. For a large degrees of freedom you may need the approximation formula given in table 3 (of appendix B in Rice), for the p-quantile (i.e., p-percentile), $\chi^2_{d,p}$, of the χ^2_d-distribution with d degrees of freedom,

$\chi^2_{d,p} = \frac{1}{2}\left(z_p + \sqrt{2d-1}\right)^2$, where z_p is the p-quantile in the standard normal distribution, $N(0,1)$.]

(iii) We want to test $H_0 : \theta \geq 3$ against $H_1 : \theta < 3$. Set up a test for this problem based on Y and the fact that $2\theta Y \sim \chi^2_{2n}$. The observed value of Y is given in (**ii**). Perform the test at the 5% level of significance and comment on the result.