Problem 1

Let X be a continuous random variable (*rv*) that varies in the unit interval (0, 1) and has *pdf* (probability density function)

$$f(x) = \begin{cases} \dfrac{c}{1+x} & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

where c is a constant such that f is a proper pdf.

A. **(i)** Show that $c = \dfrac{1}{\ln 2}$.

(ii) Sketch a graph of the pdf f.

(iii) Show that the cumulative distribution function (*cdf*) of X is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ c\ln(1+x) & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x > 1 \end{cases}$$

B. **(i)** Find $E(X)$. **(ii)** Find $\text{var}(X)$.

(iii) Show that, for any real $r > 0$, $E\left[(1+X)^r\right] = \dfrac{2^r - 1}{r \ln 2}$.

[Hint: Note that

$$\frac{x}{1+x} = 1 - \frac{1}{1+x}$$

$$\frac{x^2}{1+x} = x - 1 + \frac{1}{1+x}$$

]

C. **(i)** Let x_p be the p-quantile (percentile) of X, defined by $P(X \le x_p) = p$ for $0 < p < 1$. Show that $x_p = 2^p - 1$ and calculate the median of X .

(ii) Let $Y = c \cdot \ln(1+X)$ where $c = \dfrac{1}{\ln 2}$. Show that the rv Y is standard uniformly distributed (i.e., uniformly distributed over the unit interval $(0,1)$).

Problem 2

A number of married women from a district in Guatemala, all born during the period 1935-44, were interviewed in 1969 and asked the age at which they were married. The data[1] are given in grouped form in **table 1**.

The ages at marriage (in completed years) are grouped in 2-years intervals, as indicated in the table, where the lower limit in each interval is inclusive and the upper limit exclusive. For example, the first interval, 9 – 11, comprises 5 women who had completed 9 or 10 years at marriage. The midpoint 10 of the first interval will be considered as an estimate of the true age at marriage for this age group.

Table 1 Age at marriage (in completed years). Sample size, $n = 90$ women.

Age interval	9 - 11	11 - 13	13- 15	15 - 17	17 - 19	19 - 21	21 - 23
Midpoint, X	10	12	14	16	18	20	22
Frequency	5	11	18	28	8	7	4
Age interval	23 - 25	25 - 27	27 - 29	29 - 31	31 - 33	33 - 35	
Midpoint, X	24	26	28	30	32	34	
Frequency	5	2	0	1	0	1	

A. Assume that the frequencies in the table are observations of multinomially distributed rv's,

$$V_1, V_2, \ldots, V_{13} \sim \text{multinom}(n = 90,\ p_1, p_2, \ldots, p_{13}),$$

where $V_1 + V_2 + \cdots + V_{13} = 90$, $\quad p_1 + p_2 + \cdots + p_{13} = 1$, and where the categories are the 13 age intervals in table 1. (Note that the first category, age group 9 – 11, is slightly misleading. It is valid for the data, but should comprise all ages less than 11 in the general multinomial setup.)

(i) Explain briefly why $V = V_1 + V_2 + V_3$ is binomially distributed, $V \sim \text{bin}(90, p)$, where $p = p_1 + p_2 + p_3$ is the probability that a randomly chosen married woman from the population in question was married at age less than 15 years.

[1] Data are taken from Hand, Daly, Lunn, McConway, and Ostrowski, «Handbook of Small Data Sets», Chapman and Hall, 1994.

(ii) Explain why $\hat{p} = \frac{V}{n}$ and $\sqrt{\hat{p}(1-\hat{p})}$ are consistent estimators for p and $\sqrt{p(1-p)}$ respectively as $n \to \infty$.

(iii) We want to test $H_0 : p \le 0.25$ against $H_1 : p > 0.25$. Set up a suitable test statistic for this problem and calculate (approximately) the p-value of the test based on the data in table 1. If you need Slutsky's lemma in your argument, explain how you use it. Comment briefly on the result.

[Hint: You may use as a fact, without having to prove it, that $\sqrt{n}\frac{\hat{p}-p}{\sqrt{p(1-p)}} \underset{n\to\infty}{\overset{D}{\to}} Z \sim N(0,1)$, where $\overset{D}{\to}$ means convergence in distribution, and *Z* is standard normally distributed. **]**

B. Let X_i be the midpoint in the age interval that is true for woman number *i* in the sample, $i = 1, 2, \ldots, 90$. For simplicity we will now ignore the fact that X_i is only an approximation of the true age, and consider $X_1, X_2, \ldots, X_{90}$ as independent, identically distributed (*iid*), and continuous rv's. In addition, we assume that X_i is gamma distributed with shape, $\alpha > 0$ and known scale, $\lambda = 1$ (in short: $X_i \sim \Gamma(\alpha, 1)$).

(i) Explain briefly why $E(X_i) = \text{var}(X_i) = \alpha$.

(ii) Show that the ordinary least squares estimator (OLS) of α is the same as the moment method estimator (mme), $\tilde{\alpha} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, (where $n = 90$), in this case.

[Hint: Remember that the OLS estimator is obtained by minimizing $\sum_{i=1}^{n}\left(X_i - E(X_i)\right)^2 = \sum_{i=1}^{n}\left(X_i - \alpha\right)^2$ with respect to α .**]**

(iii) The observed value is $\tilde{\alpha}_{obs} = \bar{X}_{obs} = 16.556$. Use this and the central limit theorem (CLT) to derive and calculate an approximate 95% confidence interval (CI) for α.

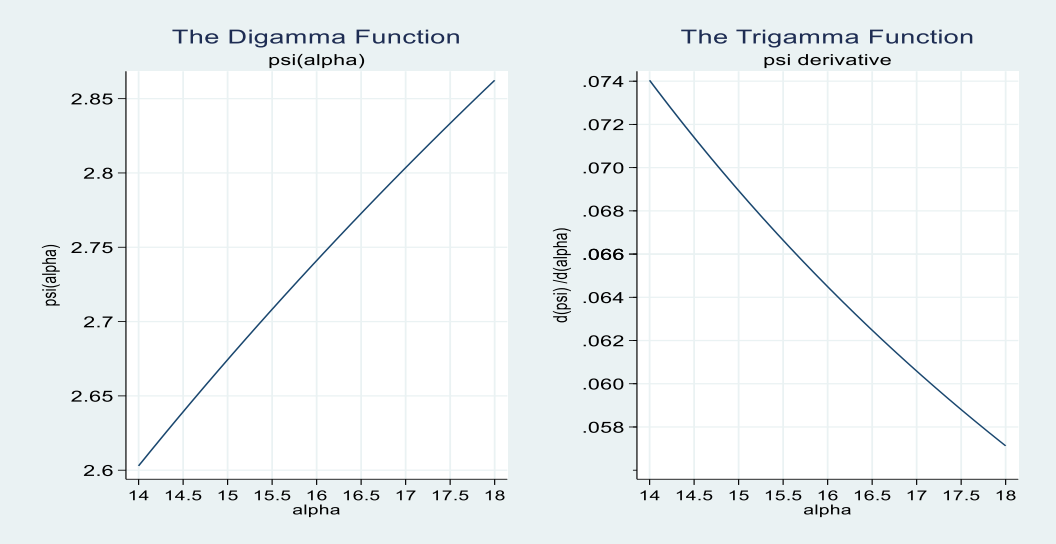
C. We continue to assume that $X_1, X_2, \ldots, X_{90}$ are iid and $X_i \sim \Gamma(\alpha, 1)$, as in **B.**

(i) Write up the log-likelihood function you would use to find the maximum likelihood estimator (mle) for α. The maximizing value of the log likelihood function cannot be found analytically since it depends on the *digamma* function, ψ, defined by $\psi(\alpha) = \frac{\partial \ln \Gamma(\alpha)}{\partial \alpha} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$, but must be determined numerically. It can be proven that $\psi(\alpha)$ is strictly increasing for all $\alpha > 0$.

Give a rough estimate of the mle, $\hat{\alpha}$, using **figure 1** and the observed value of the mean of $\ln(X_i)$, for $i = 1, 2, \ldots, 90$, which is 2.79.

[Hint: Find first the derivative of the log likelihood with respect to α **.]**

Figure 1 The di- and trigamma functions.



(ii) Show that the Fisher information for one observation is given by $I(\alpha) = \psi'(\alpha)$ (i.e., the trigamma function, defined as the derivative of the digamma function and can be proven to be continuous).

(iii) Derive an approximate 95% CI for α based on the mle, $\hat{\alpha}$. Calculate the interval using the (rough) estimate for $\hat{\alpha}_{obs}$ you found in **(i)**. If you need $\psi'(\hat{\alpha}_{obs})$, use **figure 1** to find a rough estimate.

D. **(i)** In our gamma model we have assumed that the scale parameter, λ, is 1. To see if there is any evidence against this assumption in the data, we now estimate the full gamma model, assuming $X_1, X_2, \ldots, X_{90}$ are iid with $X_i \sim \Gamma(\alpha, \lambda)$, where $\alpha > 0, \lambda > 0$.

Estimating α, λ by the maximum likelihood method and using Stata, gave the following output:

Table 2. Stata output from ml estimation of full gamma model for the grouped data.

```
Maximum likelihood estimation

Log likelihood = -253.48007                     Number of obs     =         90

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /alpha |   16.05893   2.369476     6.78   0.000     11.41484    20.70302
     /lambda |   .9700024   .1453792     6.67   0.000     .6850643     1.25494
------------------------------------------------------------------------------
```

Test $H_0 : \lambda = 1$ against $H_1 : \lambda \neq 1$ based on the output information in **table 2**. Calculate the p-value approximately and comment on the result.

(ii) Another way to test the model formulated in section **B**, is to let the model be our H_0, and test it by a Pearson chi-square test. The background model is the multinomial model used in section **A**, however after combining some age groups due to the presence of age groups with too few observations. The category probabilities under H_0 are structured according to the gamma model in section **B**. The results of doing all this are given in **table 3** below.

Table 3 The underlying data for the Pearson chi-square statistic, where O_j represents the observed frequency, and E_j the expected frequency under H_0.

j	Category	O_j	Probability under H_0	E_j	$\frac{(O_j - E_j)^2}{E_j}$
1	< 11	5	0.071	6.349	0.29
2	11 - 13	11	0.124	11.167	0.00
3	13- 15	18	0.183	16.486	0.14
4	15 - 17	28	0.198	17.819	5.82
5	17 - 19	8	0.168	15.135	3.36
6	19 - 21	7	0.118	10.614	1.23
7	21 -25	9	0.108	9.729	0.05
8	$\geq$ 25	4	0.030	2.702	0.62
	Sum	90	1	90	11.52

Perform the Pearson chi-square test for the model in section **B,** using the level of significance 5%, and comment briefly on the result.

E. **(i)** Write $\overline{\ln X} = \frac{1}{n}\sum_{i=1}^{n} \ln X_i$, where $n = 90$. Explain why the mle, $\hat{\alpha}$, is completely determined by $\overline{\ln X}$ from the data.

(ii) **Introduction.** Another weakness of the model in **B** is the grouping. X_i is not the true age at marriage for woman *i*. Let Y_i be the true age at marriage for woman

$i,\ i = 1, 2, \ldots, 90$, and write $\overline{\ln Y} = \frac{1}{n}\sum_{i=1}^{n} \ln Y_i$. Because of **(i),** we may assess the effect of grouping by comparing $\overline{\ln X}$ and $\overline{\ln Y}$. The age intervals in **table 1** are not equal on the logarithmic scale, but the largest interval is the first one, $(\ln 9,\ \ln 11)$, that has length about 0.2. A simplified model for the relation between Y_i and X_i is by postulating $\ln Y_i = \ln X_i + U_i$ where $U_1, U_2, \ldots, U_n$ are iid and uniformly distributed over the interval $(-0.1,\ 0.1)$ (i.e., $U_i \sim \text{unif}(-0.1,\ 0.1)$).

Note that this simplified model will tend to overestimate the difference between $\overline{\ln X}$ and $\overline{\ln Y}$ somewhat since all of the age intervals (except the first one) are shorter than 0.2 on the logarithmic scale.

Question: Show that $\text{var}(U_i) = \frac{1}{300}$, and that $\left|\overline{\ln Y} - \overline{\ln X}\right| = \left|\bar{U}\right|$, where $\bar{U} = \frac{1}{n}\sum_{i=1}^{n} U_i$.

(iii) Based on the approximate normal distribution of $\bar{U}$, determine a constant c such that $P\left(\left|\overline{\ln Y} - \overline{\ln X}\right| \le c\right) = P\left(|\bar{U}| \le c\right) \approx 0.95$.

F. **(i)** Let, as in **B**, $X \sim \Gamma(\alpha, 1)$. Show that the moment generating function (mgf) of $\ln X$ is $M_{\ln X}(t) = \frac{\Gamma(\alpha + t)}{\Gamma(\alpha)}$ for $t > -\alpha$.

[Hint. Note that $e^{t \ln X} = X^t$.**]**

(ii) Show that $E\left(\ln X\right) = \psi(\alpha)$ and $\text{var}\left(\ln X\right) = \psi'(\alpha)$, where ψ and ψ' are the digamma and trigamma functions respectively.