# Econ 4130 Regular Exam 2020H

## ANSWERS

Settet er inndelt i 9 deloppgaver, A,B,C,… som alle anbefales å telle likt for å gjøre det litt lettere å stå. Svar er gitt i <<< ….. >>>.

## 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 $P\left(\frac{m}{4} < X < \frac{3m}{4}\right)$ .

---

<<< **Answer:** (i) Since $X$ can take values in the interval (0, $m$) only, $U$ takes values in (0, 1). The cdf of $X$ is

$$F_X(x) = \begin{cases} 0 & \text{for } x \le 0 \\ \frac{x}{m} & \text{for } 0 < x < m \\ 1 & \text{for } x \ge m \end{cases}$$

For $u \in (0,1)$ the cdf of $U$ becomes $F_U(u) = P(U \le u) = P(X \le mu) = F_X(mu) = u$ since $0 < mu < m$.

For $u \le 0$, $mu \le 0 \Rightarrow F_U(u) = F_X(mu) = 0$.

For $u \ge 1$, $mu \ge m$ implying $F_U(u) = F_X(mu) = 1$.

The cdf $F_U(u)$ determines the standard uniform distribution and, therefore, proves the result.

(ii) $X = mU \Rightarrow E(X) = mE(U) = \frac{m}{2}$ and $\text{var}(X) = m^2\text{var}(U) = \frac{m^2}{2}$

$$E(U^2) = \text{var}(U) + \left[E(U)\right]^2 = \frac{1}{12} + \frac{1}{4} = \frac{1}{3} \Rightarrow \text{E}(X^2) = m^2 E(U^2) = \frac{m^2}{3}$$

(iii) Since $P\left(X=\frac{3m}{4}\right)=0,\quad P\left(\frac{m}{4}<X<\frac{3m}{4}\right)=F_X\left(\frac{3m}{4}\right)-F_X\left(\frac{m}{4}\right)=\frac{3}{4}-\frac{1}{4}=\frac{1}{2}$ .

>>>

---

**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 $\text{var}(Y)=\frac{5}{12}$.

---

<<< **Answer:**

(i) If the rv *W* is exponentially distributed with parameter $\lambda$, we know that $E(W)=\frac{1}{\lambda}$ and $\text{var}(W)=\frac{1}{\lambda^2}$ . Hence, since $Y\mid x\ \sim\ \exp\left(\lambda=\frac{m}{x}\right)$,

$$E(Y\mid x)=\frac{x}{m}\quad\text{and}\quad\text{var}(Y\mid x)=\frac{x^2}{m^2}$$

The regression function, $\mu(x)=E(Y\mid x)$ is, therefore, linear in *x* and heteroscedastic. It is, as is the case for the conditional pdf $f(y\mid x)$, well defined only for *x*'s that may occur as observations of *X*, that is, for $0<x<m$. (Some may answer $0<x\le m$, which is acceptable, of course.)

(ii) We get from **A** that $E(Y)=E\left[E(Y\mid X)\right]=E\left(\frac{X}{m}\right)=\frac{1}{2}$ and

$$\text{var}(Y)=E\left[\text{var}(Y\mid X)\right]+\text{var}\left[E(Y\mid X)\right]=\frac{1}{m^2}E\left(X^2\right)+\text{var}\left(\frac{X}{m}\right)=$$

$$=\frac{1}{m^2}\left[E\left(X^2\right)+\text{var}\left(X\right)\right]\ \overset{\text{A.}}{=}\ \frac{1}{m^2}\left[\frac{m^2}{3}+\frac{m^2}{12}\right]=\frac{5}{12}$$

>>>

---

**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 $E(XY) = E\big[X \cdot E(Y \mid X)\big]$, 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. ]

---

<<< **Answer:**

(i) We have $\text{cov}(X,Y) = E(XY) - E(X)E(Y)$, where

$$E(XY) = E\big[X \cdot E(Y \mid X)\big] = E\left(\frac{X^2}{m}\right) = \frac{1}{m} \cdot \frac{m^2}{3} = \frac{m}{3}$$

Hence, $\text{cov}(X,Y) = \frac{m}{3} - \frac{m}{2} \cdot \frac{1}{2} = \frac{m}{12}$ and the correlation

$$\rho(X,Y) = \frac{\frac{m}{12}}{\sqrt{\frac{m^2}{12} \cdot \frac{5}{12}}} = \frac{1}{\sqrt{5}}$$

(ii) Let $V = \frac{Y}{X}$. Following the hint, we get the conditional cdf of $V$ given that $X = x$ is fixed:

For any $v > 0$, $P(V \le v \mid x) = P\left(\frac{Y}{X} \le v \;\middle|\; x\right) = P\left(\frac{Y}{x} \le v \;\middle|\; x\right) = P\big(Y \le xv \;\big|\; x\big) =$

$$= 1 - e^{-\frac{m}{x}vx} = 1 - e^{-mv} \quad \text{(and } P(V \le v \mid x) = 0 \text{ for } v \le 0\text{)}$$

Hence, $(V \mid x)$ is exponentially distributed $(m)$ - whatever value of $x$ – with the pdf $f_V(v \mid x) = me^{-mv}$ for $v > 0$, that does not depend on $x$. We have shown in the lectures that this implies that $V$ and $X$ are independent.

>>>

---

**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 $P\big(W \le q(p)\big) = p$. ]

---

<<< **Answer:**
If $U \sim \text{unif}(0,1)$, then $X = 10U \sim \text{unif}(0,10)$. Hence $x = 10 * 0.21 = 2.1$ is an observation of $X$.

We have $Y \mid x \sim \exp(10/x)$ distributed with cdf $F(y \mid x) = 1 - e^{-\frac{10}{x}y}$ for $y > 0$, and quantile function, $q(p) = -\frac{x}{10}\ln(1-p)$. Then $q(U) = -\frac{x}{10}\ln(1-U) \sim \exp(10/x)$.
From this we obtain $y = q(0.56) = -\frac{2.1}{10}\ln(1-0.56) = 0.172...$,
and $(2.1,\ 0.172)$ is a draw from the distribution of $(X, Y)$.

>>>

---

## Problem 2

**A.** Let the random variable (rv) $U$ be gamma distributed with shape parameter $\alpha$ and scale parameter $\lambda$ (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

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

---

<<< **Answer:**

$$E\left(U^{-1}\right) = \frac{\Gamma(\alpha-1)}{\lambda^{-1}\Gamma(\alpha)} = \frac{\Gamma(\alpha-1)}{\lambda^{-1}(\alpha-1)\Gamma(\alpha-1)} = \frac{\lambda}{\alpha-1}$$

$$E\left(U^{-2}\right)=\frac{\Gamma(\alpha-2)}{\lambda^{-2}\Gamma(\alpha)}=\frac{\Gamma(\alpha-2)}{\lambda^{-2}(\alpha-1)(\alpha-2)\Gamma(\alpha-2)}=\frac{\lambda^2}{(\alpha-1)(\alpha-2)}$$

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

$$=\frac{\lambda^2}{(\alpha-1)^2(\alpha-2)}$$

>>>

---

**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 $(b,\theta)$ with cdf

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

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 $\theta$ based on $X_1,X_2,\ldots,X_n$, is

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

Explain briefly why $Y$ is $\Gamma(n,\theta)$ distributed. Show also that the mle of $\theta$ 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 $\theta$, obtained as a modification of $\hat{\theta}$ (i.e., by multiplying $\hat{\theta}$ by a suitable constant).

**(iii)** Show that $MSE(\hat{\theta})=\dfrac{n+2}{(n-1)(n-2)}\theta^2$, 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 $\text{var}\left(\hat{\theta}\right) = \dfrac{n^2\theta^2}{(n-1)^2(n-2)}$ 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 $n^2+n-2 = (n-1)(n+2)$. ]

---

<<< **Answer:** (i): The exp($\theta$) distribution is the same as $\Gamma(1,\theta)$. We know that a sum of independent gamma variables with the same scale parameter is gamma distributed with shape equal to the sum of the individual shapes. Since all $Y_i \sim \Gamma(1,\theta)$ have the same scale $\theta$, $Y = \sum_{i=1}^{n} Y_i$ must be $\Gamma(n,\theta)$ distributed.

The joint distribution of $Y_1, Y_2, \ldots, Y_n$ is

$f(y_1, y_2, \ldots, y_n \mid \theta) = f(y_1 \mid \theta) f(y_2 \mid \theta) \cdots f(y_n \mid \theta) = \theta^n e^{-\theta \sum_{i=1}^{n} y_i}$. Substituting the observed values for $y_1, y_2, \ldots, y_n$ leads to the log likelihood $l(\theta) = n \ln\theta - \theta \sum_{i=1}^{n} y_i$ with derivative $l'(\theta) = \dfrac{n}{\theta} - \sum_{i=1}^{n} y_i = 0$ for $\theta = \theta^* = \dfrac{n}{y}$ (the ml estimate) where $y = \sum_{i=1}^{n} y_i$, giving the ml estimator $\theta^* = \dfrac{n}{Y}$, which is the same as $\hat{\theta}$.

(ii) Since $Y \sim \Gamma(\alpha = n,\ \lambda = \theta)$, we get from section **A,**

$E(\hat{\theta}) = nE\left(\dfrac{1}{Y}\right) = n\dfrac{\theta}{n-1} = c_n\theta$ where $c_n = \dfrac{n}{n-1}$**.** Hence, $\tilde{\theta} = \dfrac{(n-1)}{n}\hat{\theta} = \dfrac{n-1}{Y}$ is unbiased.

(iii) For any estimator $\hat{\hat{\theta}}$, MSE is $E\left[\left(\hat{\hat{\theta}} - \theta\right)^2\right] = \text{var}\left(\hat{\hat{\theta}}\right) + \left[E\left(\hat{\hat{\theta}}\right) - \theta\right]^2$. Hence

$$MSE(\hat{\theta}) = \text{var}(\hat{\theta}) + \left(E(\hat{\theta}) - \theta\right)^2 = \frac{n^2\theta^2}{(n-1)^2(n-2)} + \left(\frac{n\theta}{n-1} - \theta\right)^2 =$$

$$= \theta^2\left[\frac{n^2}{(n-1)^2)(n-2)} + \left(\frac{n}{n-1} - 1\right)^2\right] = \theta^2\left[\frac{n^2}{(n-1)^2)(n-2)} + \frac{1}{(n-1)^2}\right] =$$

$$= \frac{\theta^2}{(n-1)^2} \cdot \frac{n^2+n-2}{(n-2)} = \frac{\theta^2(n-1)(n+2)}{(n-1)^2(n-2)} = \frac{n+2}{(n-1)(n-2)}\theta^2$$

while

$$MSE(\tilde{\theta}) = \text{var}(\tilde{\theta}) = (n-1)^2 \cdot \text{var}\left(\frac{1}{Y}\right) \overset{\text{section A}}{=} (n-1)^2 \frac{\theta^2}{(n-1)^2(n-2)} = \frac{\theta^2}{n-2}$$

The unbiased $\tilde{\theta}$ has smallest MSE and is, therefore, preferable by the MSE-criterion.
Estimates: $\hat{\theta}_{obs} = 3.813$, $\tilde{\theta}_{obs} = 3.811$ (i.e., not much difference…).

>>>

---

**C.** Suppose that $X$ is Pareto distributed $(b,\theta)$ 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 $\theta$ as in the distribution of $X$.

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

---

<<< **Answer:**

We have for $v > c$,

$$P(V \le v) = P(X \le v \mid X > c) = \frac{P(c < X \le v)}{P(X > c)} = \frac{1-\left(\frac{b}{v}\right)^{\theta} - 1 + \left(\frac{b}{c}\right)^{\theta}}{\left(\frac{b}{c}\right)^{\theta}} =$$

$$= 1-\left(\frac{c}{b}\frac{b}{v}\right)^{\theta} = 1-\left(\frac{c}{v}\right)^{\theta} \quad \Rightarrow \quad V \sim \text{Pareto}(c,\theta)$$

>>>

---

**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 $\theta$ 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$(b, \theta)$, 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.

---

<<< **Answer:**

**(i):** The expected frequency for category *j* under $H_0$ is $E_j = n\hat{p}_j(\theta) = np_j(\hat{\theta}_{II,obs})$, where $p_j(\theta)$ is the category probability under $H_0$. Since $\hat{\theta}_{II,obs} = 2.869$, we get

$$p_{13}(\hat{\theta}_{II,obs}) = \hat{F}_{II}(1\,000\,000) - \hat{F}_{II}(800\,000) = \left(\frac{350\,000}{800\,000}\right)^{2.869} - \left(\frac{350\,000}{1\,000\,000}\right)^{2.869} =$$

$$0,0933 - 0.0492 = 0.0441 \quad \Rightarrow \quad E_{13} = (0.0441)(556) = 24.53$$

and $\quad Q_{13} = \dfrac{(O_{13} - E_{13})^2}{E_{13}} = \dfrac{(17 - E_{13})^2}{E_{13}} = 2.31$

$$p_{14}(\hat{\theta}_{II,obs}) = 1 - \hat{F}_{II}(1\,000\,000) = \left(\frac{350\,000}{1\,000\,000}\right)^{2.869} = 0.0492 \quad \Rightarrow \quad E_{14} = (0.0492)(556) = 27.35$$

and $\quad Q_{14} = \dfrac{(O_{14} - E_{14})^2}{E_{14}} = \dfrac{(30 - E_{14})^2}{E_{14}} = 0.26$

**(ii)** We want to test $H_0 : X \sim \text{Pareto}(b, \theta)$ vs $H_1 : "X \text{ has some other distribution}"$ for both groups with $b = 250\,000$ for group I and $b = 350\,000$ for group II.

Under $H_0$, $\quad Q = \sum_j Q_j \overset{\text{approximately}}{\sim} \chi^2_{k-1-s} = \begin{cases} \chi^2_{12} & \text{for group I} \\ \chi^2_8 & \text{for group II} \end{cases}$,

where *k* is the number of categories and *s* is the number of parameters estimated under $H_0$. The 1% critical value is then 26.22 for group I and 20.09 for group II according to the Rice $\chi^2$ table.

Hence, we reject $H_0$ for group I but not for group II.

**(iii)** For group II the critical value at 10% is 13.36. Since we do not reject $H_0$ at this level of significance, the p-value must be larger than 10%, and not smaller.

>>>

---

**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 $\theta$ 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 $\chi^2_d$-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.

---

<<< **Answer:**

**(i)**: The mgf of $Y \sim \Gamma(n,\theta)$ is $M_Y(t) = \left(\frac{\theta}{\theta - t}\right)^n$. $U = 2\theta Y$ is a linear transformation of *Y* with mgf, $M_U(t) = M_Y(2\theta t) = \left(\frac{\theta}{\theta - 2\theta t}\right)^n = \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^n$, i.e., the mgf of $\Gamma\left(n, \frac{1}{2}\right) = \Gamma\left(\frac{2n}{2}, \frac{1}{2}\right) = \chi^2_{2n}$. The uniqueness of the mgf determines the $\chi^2_{2n}$ distribution for $2\theta Y$.

**(ii):** Let $\chi^2_{d,p}$ be the p-quantile (percentile) in the $\chi^2_d$ distribution. Then $0.95 = P\left(\chi^2_{2n,0.025} \leq 2\theta Y \leq \chi^2_{2n,0.975}\right) = P\left(\frac{\chi^2_{2n,0.025}}{2Y} \leq \theta \leq \frac{\chi^2_{2n,0.975}}{2Y}\right)$, giving the 95% CI:

$\left[\frac{\chi^2_{2n,0.025}}{2Y}, \frac{\chi^2_{2n,0.975}}{2Y}\right]$. Since the $df = 2*556 = 1112$, we get from the approximation formula,

$$\chi^2_{1112,0.025} = 0.5*\left(-1.96 + \sqrt{2*1112-1}\right)^2 = 1021.0093$$

$$\chi^2_{1112,0.975} = 0.5*\left(1.96 + \sqrt{2*1112-1}\right)^2 = 1205.8323$$

and the observed 95% CI for $\theta$, $\left[\frac{1021.0093}{2Y_{obs}}, \frac{1205.8323}{2Y_{obs}}\right] = [2.634,\ 3.111]$

**(iii):** $H_0 : \theta \geq 3$ against $H_1 : \theta < 3$ and $2\theta Y \sim \chi^2_{2n}$, suggest the test statistic $Z = 6Y = \frac{3}{\theta} 2\theta Y$ which tends to get larger under the alternative hypothesis.

An $\alpha$ - level test is therefore: "Reject $H_0$ if

$Z > \chi^2_{2n,1-\alpha} = (0.5)\left(1.65 + \sqrt{2 \cdot 1112 - 1}\right)^2 = 1190.66$ for $\alpha = 0.05$

Since $Z_{obs} = 6 \cdot (193.7957) = 1162.77...$ , $H_0$ is not rejected at the 5% level of significance. So, there is not sufficient evidence (at 5%) to claim that $\theta$ is lower than 3, but maybe lower than 3.1 according to the CI. In any case, the results of this exercise indicate that the cut-off *b* = 250 000 used in the lectures for these data, and chosen rather ad hoc, was set too low for female incomes in 1998.

>>>

---