HG
Sept 2017

ECON 4130

Supplementary Exercises 1- 4

(for seminar week 39)

Exercise 1

Quantiles (percentiles).
Let *X* be a continuous random variable (rv.) with pdf $f(x)$ and cdf $F(x)$. For $0 < p < 1$ we define *p*-th quantile (or 100*p*-th percentile), x_p, as the solution of $P(X \le x_p) = p$ or, with other words, the solution of $F(x_p) = p$. Hence $x_p = F^{-1}(p)$. Then, e.g., $x_{0.5} = F^{-1}(0.5)$ is the median and $x_{0.25}$ the lower quartile. (If the equation $F(x_p) = p$ has several solutions, then x_p is by convention chosen as the smallest one. This may occur when $F(x)$ is flat in some regions as will be the case when *X* is discrete.)

a. Suppose *X* is exponentially distributed with parameter λ (pdf $f(x) = \lambda e^{-\lambda x}$, $x > 0, \quad \lambda > 0$). Find x_p as a function of λ . This distribution is right skewed (i.e. has a heavy tail to the right). For such distributions usually the expected value is larger than the median. Illustrate this by calculating the median and the expectation for *X* when $\lambda = 3$.

b. Based on a sample of (total) incomes from $n_1 = 14867$ men and $n_2 = 14286$ women drawn from the Norwegian population in 1998, a number of quantiles of the income distributions for men and women have been estimated. The results are given in the table.

Table 1. *Estimated quantiles of income distributions for men and women. Norway 1998*

p	x_p Men (NOK 1000)	x_p Women (NOK 1000)
0.2	125	68
0.4	214	118
0.6	279	177
0.8	370	237
0.95	650	331
0.99	1260	550

Based on the quantiles, draw a histogram for the incomes in the sample both for men and for women. [**Hint:** The class limits are given by the x_p's. Therefore the length of the intervals will vary. The relative frequency of the intervals must then be equal to the area of the corresponding boxes of the histogram. Note that the relative frequency of, e.g., the interval $(x_{0.2}, x_{0.4})$ is 0.2 etc.]

Exercise 2 (Pareto Distribution)

(Note: Section a and d are based on Problem 1 in the exam paper 2003 H)

Vilfredo Pareto (1848-1923) claimed that incomes often are distributed according to a particular form, at least incomes above a certain limit ("the upper tail"). This type of distribution is called the Pareto distribution. Let *b* denote the lower limit of incomes considered. The density of the Pareto distribution is defined by $f(x) = \theta b^{-1}(x/b)^{-\theta-1}$ for $x > b$, and $f(x) = 0$ for $x \le b$. Here b is a chosen cut-off point (usually known) and $\theta > 0$ is a parameter that determines the form of the distribution (i.e. the length of the upper tail before it gets negligible).

a. Let *X* be Pareto distributed as described. Show that that the cdf is $F(x) = 1-(b/x)^{\theta}$ for $x > b$. What is the value of $F(x)$ for $x \le b$? Find E(*X*). Note that E(*X*) does not exist if $\theta \le 1$. Why? Illustrate by considering the case $\theta = 1$.

b. Show that the *p*-th quantile is given by $x_p = b(1-p)^{-1/\theta}$ (or $\ln(x_p) = \ln(b) - \frac{1}{\theta}\ln(1-p)$).

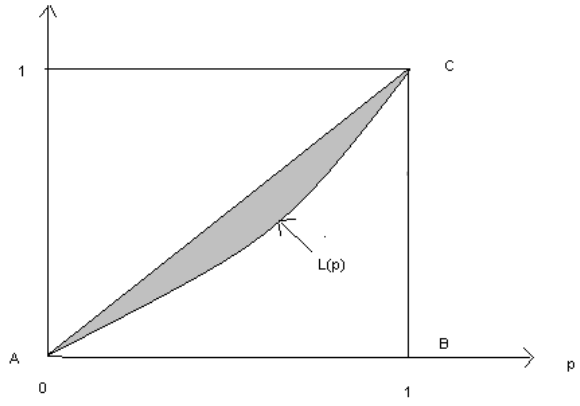
c. Let 0.21, 0.78, 0.54, 0.98 be 4 observations drawn (i.e. simulated by a computer) from the uniform distribution over [0, 1]. Transform these observations to 4 observations drawn (i.e. simulated) from the Pareto distribution with $b = \text{NOK } 250\,000$ and $\theta = 2.8$.

d. Show that $Y = \ln(X/b)$ is exponentially distributed with pdf $f_Y(y) = \theta e^{-\theta y}$ for $y > 0$.

Exercise 3 (Lorenz Curve)

Income distributions are sometimes studied by the so called Lorenz curve, $L(p) \quad 0 \le p \le 1$, that says something about the income inequality in the population. The (empirical) Lorenz curve based on a sample of incomes drawn from the population tells us, for given *p*, the percentage that the total sum of incomes from *the 100p% lowest incomes* in the sample, constitutes of the total sum of incomes from *the complete sample*. See figure 1 and the rationale below (1) for details.

Fig. 1 Lorentz Curve, L(p)



For a theoretical definition consider an income distribution given by the pdf, $f(x)$, concentrated on the positive axis, and let the rv X be distributed according to *f*. Then X represents the income of a randomly drawn member from the population. Let x_p be the *p*-th quantile of X, and put $\mu = \mathrm{E}(X) = \int_0^\infty xf(x)dx$. The Lorenz curve is defined by

(1) $$L(p) = \frac{1}{\mu}\int_0^{x_p} xf(x)dx$$

[**Rationale**: Consider a finite population with *N* members and an income distribution close to *f*. Then the mean income in the population is (close to) $E(X) = \mu$ and the total sum of incomes is (close to) $N\mu$. We now need an expression for the sum of incomes that are lower than a cut-off point, *c*. Let *Y* be *X* truncated at *c*, in the sense that $Y = X$ when $X \le c$ and $Y = 0$. when $X > c$. Then the total sum of incomes in the population that are less or equal to *c* is $N \cdot \mathrm{E}(Y)$. Now *Y* is a mixed rv, i.e. partly discrete and partly continuous (note that $P(Y = 0) = P(X > c) > 0$). The theory for finding the expectation of such a rv is not covered in this course, but it can be shown that $\mathrm{E}(Y) = 0 \cdot P(Y = 0) + \int_0^c xf(x)dx = \int_0^c xf(x)dx$. Hence, the sum of incomes that are at most *c*, divided by the total sum of incomes in the population is

$$\frac{N\mathrm{E}(Y)}{N\mu} = \frac{\mathrm{E}(Y)}{\mu} = \frac{1}{\mu}\int_0^c xf(x)dx \quad]$$

a. Suppose that *X* is Pareto distributed as defined in exercise 2 where we assume that $\theta > 1$ so that E(*X*) exists. Show that the (theoretical) Lorenz curve in this case is given by

$$L(p) = 1 - (1-p)^{1-1/\theta}$$

b. One possible measure of the degree of income inequality in the population is the so called Gini coefficient, *G*, which is defined as the shaded area in figure 1 divided by the area of the triangle ABC, i.e. *G* is equal to 2 times the shaded area. Hence $0 \le G \le 1$. Under what circumstances is $G = 0$? Show that for the Pareto distribution

$$G = \frac{1}{2\theta - 1}$$

c. For the Norwegian income data from 1998, referred to in exercise 1, it turns out that the Pareto distribution fits satisfactorily well for incomes above NOK 250 000, both for women and for men. Assume $b = 250\,000$. Then (maximum likelihood) estimates for θ based on these data are:

Women: $\hat{\theta} = 3.813$ Men: $\hat{\theta} = 2.283$

Calculate corresponding estimates for the Gini coefficient and comment shortly on the results.

(**Note.** The maximum likelihood estimation method (that comes later in the course) has the property that, if $h(\theta)$ is a transformed parameter, then $h(\hat{\theta})$ is the maximum likelihood estimator for $h(\theta)$ whenever $\hat{\theta}$ is the maximum likelihood estimator for θ.)

Exercise 4 (Chi-square distribution)

One class of distributions that occurs very often in applied statistics is the chi-square distribution with *d* degrees of freedom (*d* = 1, 2, 3, …..), (written in short χ^2_d). The pdf is given by:

$$f(z) = \frac{1}{2^{d/2}\Gamma(\frac{d}{2})} z^{d/2-1} e^{-z/2} \quad \text{for} \quad z > 0$$

$$f(z) = 0 \qquad \text{for } z \le 0$$

a. Suppose the rv *Z* is distributed according to this distribution (written $Z \sim \chi^2_d$). Show that

i) $\mathrm{E}(Z) = d$

ii) $\mathrm{var}(Z) = 2d$

iii) $\mathrm{E}(Z^r) = 2^r \dfrac{\Gamma(\frac{d}{2}+r)}{\Gamma(\frac{d}{2})}$ for any real $r > -d/2$ (even non-integers)

(**Hint:** Identify *Z* as a gamma distributed rv (i.e. $Z \sim \Gamma(\frac{d}{2}, \frac{1}{2})$), and use the formula, derived in the lectures,

$$\mathrm{E}(X^r) = \frac{\Gamma(\alpha+r)}{\lambda^r \Gamma(\alpha)}$$ for any real $r > -\alpha$, when $X \sim \Gamma(\alpha, \lambda)$.)

b. Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed (*X* ~ iid) with expectation, $\mathrm{E}(X_i) = \mu$ and variance, $\mathrm{var}(X_i) = \sigma^2$. From the introductory statistics course we know that an unbiased estimator of σ^2 is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$ (i.e. $\mathrm{E}(S^2) = \sigma^2$) where $\bar{X}$ is the sample mean.

Show that the sample standard deviation, $S = \sqrt{S^2}$, is biased in the sense that it under-estimates σ, i.e. $\mathrm{E}(S) < \sigma$.
(**Hint:** Use the result that for any rv, *Y*, that is not a constant, we have $0 < \mathrm{var}(Y) = \mathrm{E}(Y^2) - \left[\mathrm{E}(Y)\right]^2$. Use this with $Y = S$).

c. Hence *S* is a biased estimator for σ (the bias can be shown to be negligible for large *n* in most cases). We will calculate the bias in a special case. Write $\mathrm{E}(S) = c\sigma$ where *c* is a constant with $0 < c < 1$. Calculate *c* for the special case that *n* = 10 and the X_i's are normally distributed (i.e. $X_i \sim N(\mu, \sigma^2)$).

(**Hint:** Use the famous theorem (see e.g. rule 158 in Løvås, "Statistikk for universiteter og høgskoler", or see Theorem B in Rice sec. 6.3) that says that, under these conditions, $V = \dfrac{(n-1)S^2}{\sigma^2}$ is chi-square distributed with n-1 (= 9 in this case) degrees of freedom.

Note that $S = \dfrac{\sigma}{\sqrt{n-1}}\sqrt{V}$.)