UBINET, Master 1 IFI Introduction to Probability and Statistics

Final Exam, May 2014

2 hours

Course and manuscript notes allowed. Not computers, cellphones, calculators, books.
When necessary, computation results are given at the end of the exercise.

Instruction and comments: the points awarded for your answer will be based on the correctness of your answer as well as the clarity of the main steps in your reasoning. All proposed solutions must be proved. **All the exercises are independent.** *The points are indicated so you may adapt your effort.*

Exercise 1 (Moments and Quantiles, 8 points) The exponential distribution with parameter λ is defined for positive values as : $f(x) = \lambda e^{-\lambda x}$ for all $x \geq 0$. (Note: question 4 is independent from the other ones)

1. Compute the mean μ and variance σ^2 of an exponential distribution with parameter λ.
2. Compute the coefficient of variation C defined as $C = \frac{\sigma}{\mu}$ of an exponential distribution with parameter λ.
3. Give the general interpretation of C and explain why it is an interesting metric when comparing two random samples.
4. The n-th quantile of a distribution that takes positive values is the value x_n that can be computed by the following formula:
$$\frac{n}{100} = \int_0^{x_n} f(u)du$$
   (a) Give a close formula for the n-th quantile of an exponential distribution with parameter λ.
   (b) Compare the mean and the median of an exponential distribution with parameter λ.
5. The skewness was defined in the course with the slide in Figure 1.
   (a) Give the definition of μ_2 and μ_3.
   (b) Compute the skewness for the case of an exponential distribution with parameter λ.
   (c) Does the sign of the skewness comply with the interpretation at the end of the slide in Figure 1?

Exercise 2 (Descriptive,6 points) One collects the following dataset during an experiment:
$X = \{84.0717255983663; 25.4282178971531; 81.4284826068816$
$24.3524968724989; 92.9263623187228; 34.9983765984809$
$19.6595250431208; 25.1083857976031; 61.6044676146639$
$47.3288848902729; 35.1659507062997; 83.0828627896291$
$58.5264091152724; 54.9723608291140; 91.7193663829810$
$28.5839018820374; 75.7200229110721; 75.3729094278495$
$38.0445846975357; 56.7821640725221; 18000\}$

1. Figure 2 presents both a boxplot and a cumulative distribution function of X.
   (a) Explain why one ends up with such figures that are difficult to read?
   (b) Propose two solutions to improve the readability of the figures: (i) if you are allowed to set aside some of the values in X (which one(s)?) and (ii) if you must keep all values.
2. While collecting X, a second dataset Y was simultaneously collected, whose values are: $Y = \{4.43167031066947; 3.23585949821918; 4.39972512098001$
$3.19263438523668; 4.53180737648718; 3.55530167751315$
$2.97856195588807; 3.22320188586229; 4.12073439412165$
$3.85712078332523; 3.56007830497819; 4.41983845664280$

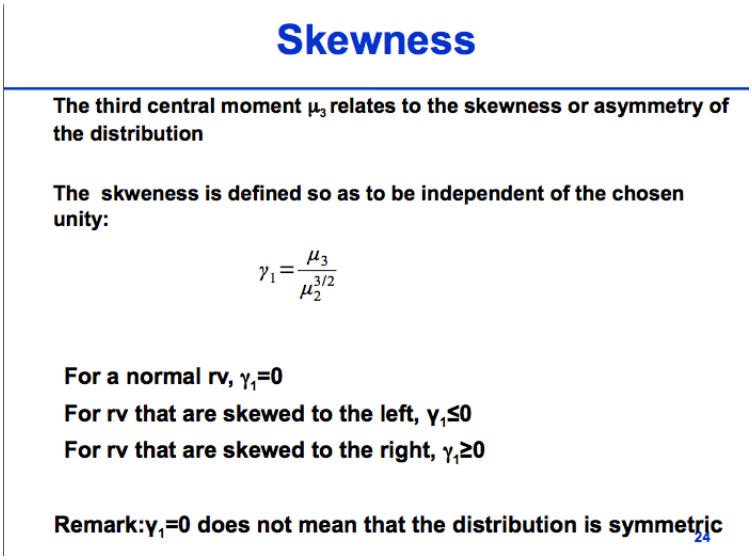

Figure 1: Skewness

4.06947809023830; 4.00683052854201; 4.51873354979948
3.35284368809688; 4.32704262891731; 4.32244791904546
3.63875875348363; 4.03922226360705; 9.79812703687830}

Figure 3 presents the scatterplot and the qqplot of (X, Y).

(a) What does the scatterplot tell you?

(b) The solid bold line in the qqplot is the bissector line (the qqplot itself is the set of crosses). How do you interpret the qqplot?

(c) How will the scatterplot evolve if one shuffles the values of X (no modification of Y)?

(d) Same question for the qqplot: How will it evolve if one shuffles the values of X (no modification of Y)?

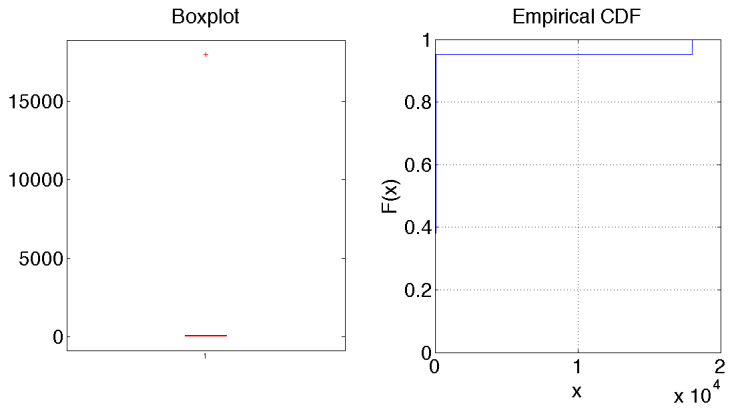

Figure 2: Boxplot/CDF

Exercise 3 (A little bit of sport, 5 points) *A runner is training on a stadium. He is monitoring its lap times. We suppose that its lap times follow a normal distribution* $\mathcal{N}(60, 10)$ *of expectation* $E = 60$ *and of standard deviation* $\sigma = 10$. *We consider a random sample of* n *lap times.*

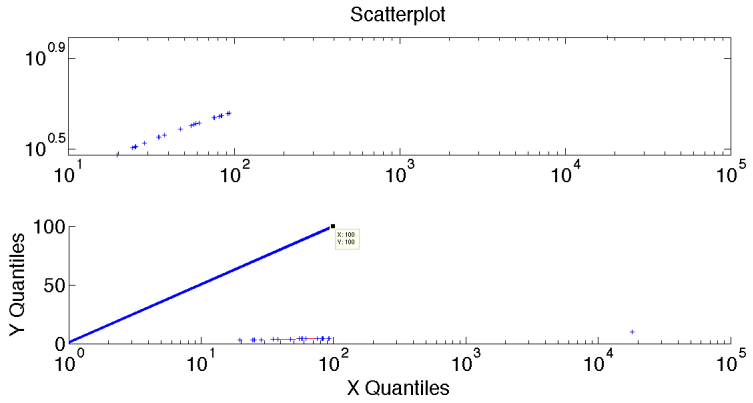

Figure 3: Scatterplot/Qqplot

For all the following questions, write the formulas as a function of n and do numerical computations for $n = 1$, $n = 9$, $n = 100$.

1. *What is the distribution of the mean m_n of the sample marks?*
2. *What is the probability that $m_n \geq 62$?*
3. *What is the probability that $58 \leq m_n \leq 62$?*
4. *Determine an interval $[E - a_n, E + a_n]$ centered in E such that the probability that m_n is in this interval is 90%.*

To obtain numerical results, you may use the following computation results: $3.33 * 0.65 = 2.16$, $3.33 * 0.67 = 2.23$, $3/5 = 0.6$

Exercise 4 (Can't wait for the ski!, 5 points) The company *Omlet* has built a cable car which collapses as soon as the total weights of the passengers is over 3000 kg. The security engineer knows that the passenger average weight is 70kg. He decides to limit the capacity of the cable car to 40 persons.

A random group of 40 people takes the cable car.

1. We suppose first that the standard deviation of the distribution of the passenger weight is $\sigma = 10$. What is the risk that the cable car collapse? Write first the formulas as a function of σ and then do the numerical computations.
2. We then suppose that σ is unknown. Explain which method you would follow to answer the question and give formulas.

To obtain numerical results, you may use the following computation results: $\sqrt{40} = 6.32$, $\sqrt{10} = 3.16$.

Exercise 5 (Shake, Serve, Drink, 5 points) The company ORANGINOR is specialized in the production of softdrink bottles. The trade manager has just changed the advertising agency. This agency has decided to axe its advertising campaign on a new slogan: "Shake, serve, drink". An advertising campaign is launched on television. To estimate the impact of this campaign, the trade manager proceeds to a poll with a random sample of 2000 people (taken in a population that is supposed infinite). Each person was aked: "Do you know ORANGINOR's new slogan?" 300 people answered yes.

1. Give a 95% confidence interval of the proportion of the global population knowing the new slogan.

   Would you reject with a risk lower than 5% the hypothesis H_0: 20% of the population knows the slogan?

To obtain numerical results, you may use the following computation results: $\sqrt{0.1275} = 0.36$, $\frac{300}{2000} = 0.15$, $1.96 * 0.008 = 0.01565$, $0.15 * 0.85 = 0.1275$, $\sqrt{2000} = 44.7$, $0.36/44.7 = 0.008$.