UNIVERSITY OF LUGANO
FACULTY OF INFORMATICS
PROF. CLAUDIO ORTELLI
SS 2011
PROF. DAVIDE LA VECCHIA

Statistics

ALaRI Exam 27 June 2011

- Duration: 2 hours and 30 minutes
- Open book exam
- Solve all exercises

1 Exercise 1 (20 pts)

1.1 (**10 pts**) **Binomial Distribution:** Let X be a random variable having a Binomial distribution, with parameters n and p.

(i) For $n = 10$ and $p = 0.2$, compute, $\mu_X := E(X)$, $\sigma_X = \sqrt{Var(X)}$ and $P(X < \mu_X - 2\sigma_X)$;

(ii) Let us keep $n = 10$ and let p be any real in $[0, 1]$. Find the value of p such that σ_X^2 is maximized.

1.2 (**9 pts**) **Poisson Distribution:** Let X be a random variable having a Poisson distribution such that $P(X = 0) = P(X = 1)$.

(i) Find $\mu_X := E(X)$ and $\sigma_X = \sqrt{Var(X)}$; (ii) Find an upper bound for $P(|X - \mu_X| \geq 2\sigma_X)$ and comment briefly the result.

1.3 (**1 pts**) **Gamma/Chi-square Distribution:** Let X be a random variable having a Gamma distribution $\Gamma(\alpha; \lambda)$, for $\alpha = 2$ and $\lambda = 2$. Compute $P(X < 11.1433)$. **Remark**: *In fact, a random variable having a* Γ*-distribution* $\Gamma(\alpha; \lambda)$*, with* $\lambda = 2$ *has a* χ^2*-distribution, with* 2α *degrees-of-freedom. Thus, apply the table of the* χ^2 *to solve the exercise.*

2 Exercise 2 (80 pts)

Let X be a positive random variable, measuring the time (in years) that a student at ALaRI needs in order to pass the exam of Statistics. Let us assume that X has an exponential distribution:

$$f(x;\lambda) = \lambda \exp^{-\lambda x}, \tag{1}$$

for $\lambda \in \mathbb{R}^+$ and $x \geq 0$.

2.1 **(5 pts)** Compute, as a function of λ:

2.1.1 (i) $P(X > 1.2)$;
(ii) $P(1.5 \leq X \leq 3)$;
(iii) $P(1.5 < X < 3)$;
(iv) $P(X \leq 0.5)$;
(v) $P(X = 3.55)$;
(vi) $P(X \geq 2.5 | X > 2)$;

2.1.2 Find x_α such that $F_X(x_\alpha) = \alpha$, for any given value of $\alpha \in [0, 1]$.

2.2 **(10 pts)** Estimation:

2.2.1 Compute the Maximum Likelihood Estimator (MLE) of λ and $\frac{1}{\lambda}$;

2.2.2 Derive analytically the expression for $Var(X)$. Then, provide the MLE for $Var(X)$. Justify your answer.

2.2.3 Provide the MLE of the tail area: $P(X > 10)$. Justify your answer.

2.3 **(35 pts)** Now let us consider the random variable

$$S_n := \sum_{i=1}^{n} X_i, \tag{2}$$

which is defined as the sum of n i.i.d. random variables, each having an exponential density as in Eq.(1), for n fix.

2.3.1 What is the exact distribution (expressed as a function of λ) of S_n? (Hint: apply the Laplace transform of X_i).

2.3.2 Provide the MLE of $P(S_n \geq s)$, for $s > 0$. Justify your answer.

2.3.3 Using the results in 2.3.1, explain carefully how one can define a test with level $\alpha = 5\%$, for:

$$H_0 : \lambda = \lambda_0 \quad vs \quad H_1 : \lambda > \lambda_0. \tag{3}$$

We assume that the value λ_0 is larger than zero. (Hint: find the distribution of S_n under H_0).

2.3.4 Assume that we are given a sample of $x_1, x_2, ..., x_n$ observations of X. For $n = 50$, we obtain $s = \sum_{i=1}^{50} x_i = 100$. According to the test defined in the previous point 2.3.3, do you accept the null hypothesis H_0 in (3), for $\lambda_0 = 0.5$?

2.4 **(30 pts)** Now let us consider the new random variable:

$$\bar{X}_n := \frac{1}{n} S_n. \tag{4}$$

2.4.1 Provide the expression for $E(\bar{X}_n)$ and $Var(\bar{X}_n)$;

2.4.2 Consider also the random variable:

$$Z_n := \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{Var(\bar{X}_n)}}. \tag{5}$$

(i) What is the distribution of Z_n, for n large?

(ii) Define $Y := Z_n^2$. What is the distribution of Y, when n is large?

(iii) Explain how one can compute: $P(Z_n > z)$ and $P(Y > z^2)$, for $z \in \mathbb{R}^+$ and n large.

2.4.3 For $n = 1000$, provide an approximation to the distribution (expressed as a function of λ) of $S_n = n\bar{X}_n$; see Eq.(4). Compare this result to the exact distribution of S_n, as derived in the question 2.3.1, and comment this finding.

3 Exercise 3 (40 pts)

Suppose we have a random sample X_1, ...,X_n from an exponential distribution with unknown parameter λ, i.e. X_i are $i.i.d. \sim Exp(\lambda)$. Suppose we want to estimate $\frac{1}{\lambda}$. Let us define $M_n = \min(X_1, ..., X_n)$.

3.1 **(8 pts)** Show that the estimator $T_1 = \bar{X}_n = (X_1 + ... + X_n)/n$ is an unbiased estimator of $1/\lambda$.

3.2 **(8 pts)** Derive and explain in details the following formula: $P(M_n \leq x) = 1 - \prod_{i=1}^{n}(1 - P(X_i \leq x))$.

3.3 **(8 pts)** Show that $M_n \sim Exp(n\lambda)$. Give detailed explanation.

3.4 **(8 pts)** Show that $T_2 = nM_n$ is an unbiased estimator of $1/\lambda$.

3.5 **(8 pts)** Which of the estimators T_1 and T_2 would you choose for estimating the mean $1/\lambda$? Substantiate your answer.

4 Exercise 4 (30 pts)

Someone is proposing two unbiased estimators U and V, with the same variance $Var(U) = Var(V)$. It therefore appears that we would not prefer one estimator over the other. However, we could go for a third estimator, namely $W = (U + V)/2$.

4.1 **(5 pts)** Show that W is unbiased.

To judge the quality of W we want to compute its variance. Lacking information on the joint probability distribution of U and V, this is impossible. However, we should prefer W in any case! To see this,

4.2 **(15 pts)** Show by means of the variance-of-the-sum rule that the relative efficiency of U with respect to W, is equal to

$$\frac{Var\left((U+V)/2\right)}{Var(U)} = \frac{1}{2} + \frac{1}{2}\rho(U,V).$$

Here $\rho(U,V) = \frac{Cov(U,V)}{\sqrt{Var(U)\,Var(V)}}$ is the correlation coefficient.

4.3 **(10 pts)** Why does this result imply that we should use W instead of U (or V)?

5 Exercise 5 (30 pts)

We model the delivery time y of a given web service as a linear regression model (without intercept) where the deterministic variable x denotes the size in megabytes of the input:

$$y_i = \theta x_i + \epsilon_i \quad \text{for } i = 1, 2, ..., n.$$

As usual, the ϵ_i here are independent random variables with $E[\epsilon_i] = 0$, and $Var(\epsilon_i) = \sigma^2$. We consider three estimators for the slope θ of the line $y = \theta x$:

1. the least squares estimator
$$T_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2};$$
2. the average slope estimator
$$T_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i};$$
3. the slope of the averages estimator
$$T_3 = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}.$$

5.1 **(8 pts)** Show that all estimators are linear estimators, i.e. they can be written as
$$\sum_{i=1}^{n} a_i y_i.$$
Determine the weights a_i for all three estimators.

5.2 **(7 pts)** Which one of the three estimators are unbiased estimators of θ? Show the calculations.

5.3 **(10 pts)** Compute the variance of each estimator.

5.4 **(5 pts)** Which estimator is to be preferred? Why?