Statistical Inference

ESSEC

Olga Klopp

Home work 2

This Homework is to be done by teams of two students. Write your names, homework number, course title and date. A script of your code in pdf or html should be submitted on the Moodle.

Deadline: March 30

1. **Comparison of estimators (5pts)**.

   Let X be a random variable following the Binomial distribution $\text{Bin}(n, \theta)$. In this exercise, we try to estimate θ.

   (a) Compute the maximum likelihood estimator (MLE) denoted by $\hat{\theta}_{\text{MLE}}$, as well as its expectation and its variance.

   (b) Let consider a different estimator: $\hat{\theta}_{\text{alt.}} = (X+1)/(n+2)$. Calculate its expectation and variance

   (c) The quadratic risk (the mean square error, MSE) of an estimator $\hat{\theta}$ is defined as the sum of its squared bias and variance:

   $$\text{MSE}(\hat{\theta}) = |\mathsf{E}[\hat{\theta}] - \theta|^2 + \text{Var}(\hat{\theta}).$$

   Compute the MSE of the estimators $\hat{\theta}_{\text{MLE}}$ and $\hat{\theta}_{\text{alt.}}$. Conclude on the comparison of these two estimators $\hat{\theta}_{\text{MLE}}$ et $\hat{\theta}_{\text{alt.}}$. Which one would you choose?

   (d) Illustrate the comparison by a simulation:

   - Simulate a sample from a binomial distribution for some fixed θ and n.
   - Calculate the two estimators, their bias and their variance
   - Represent (on the same graph), the MSE of $\hat{\theta}_{\text{MLE}}$ et $\hat{\theta}_{\text{alt.}}$ as a function of $\theta \in [0, 1]$ for fixed n.

2. **Robustness of the estimators(5pts)**.

   In this exercise we suppose that $X_1, \ldots, X_n$ is a n-sample of unknown distribution Q. We consider a Gaussian model, that is, $Q \in \{\mathcal{N}(\theta, 1); \theta \in \mathbb{R}\}$.

   (a) Compute the maximum likelihood estimator $\hat{\theta}^{MV}$. Deduce an estimator $\hat{Q}^{MV}$ of the distribution Q

   (b) In practice, is it possible that Q is not in the proposed model (i.e. that the real distribution is not a Gaussian)? What to think of the estimator $\hat{Q}^{MV}$ as the estimator of Q in this case?

   (c) For $\theta \in \mathbb{R}$, we denote by P_θ the distribution $\mathcal{N}(\theta, 1)$. Assume that $Q = 0.99P_0 + 0.01P_{300}$. Does Q belong to the model $\{\mathcal{N}(\theta, 1); \theta \in \mathbb{R}\}$? Give the density of Q.

   (d) Intuitively, is the estimator $\hat{Q}^{MV}$ close to Q?

   (e) Propose another method of estimating θ which does not suffer from this drawback, that is an estimator which is "robust" to the bad specification of the model. Hint: $\Phi^{-1}(1/(2 \times 0.99)) \approx 0.013$ where Φ^{-1} is the quantile function of the distribution $\mathcal{N}(0, 1)$.

3. **Hypothesis testing and doping controls (5pts)**.

   During a sport meeting, J.C. and S.R. are subject to an unannounced doping control. Doctors measure the hematocrit levels in their blood. Normally this rate is equal to $\tau_0 = 45\%$ but it can be increased by taking some drug. The measure of this rate is

assumed to be Gaussian with a standard deviation of 2. The observed value of J.C. is 48 and the one of S.R. is 50. We want to know if these values are abnormal, i.e. whether they have taken a drug, or if they are just the result of the imprecision of the measurements.

(a) The *first step* for solving a hypothesis testing is to write the *statistical model*. Specify the statistical model for the doping control of J.C. and S.R..

(b) The *second step* is to choose and write the *hypotheses*. They are formally written as equations on the parameters of the model. To choose between H_0 and H_1, we apply the same rule as in a court trial with H_0 being typically a statement reflecting the status quo. And we grant H_0 the benefit of the doubt. Write the hypotheses for the doping control.

(c) The hypotheses are written thanks to parameters of the model, but as always in statistics, these parameters are not known. We then use an estimator to approximate these parameters. The *third step* consists in choosing an *estimator* (a test statistic) for the parameter which is tested and precise its distribution under H_0 (and H_1 if you can). Propose an estimator for the parameter you want to test for J.C. and S.R..

(d) The *fourth step* is to choose the shape of the *rejection region*. Looking at the hypotheses, you can choose in which qualitative case an estimator can be considered as too extreme and tilts the balance to the side of the culpability of the defendant. Propose a shape for the rejection region.

(e) The *fifth step* is to compute the *boundaries* of the rejection region using the chosen *significance level*. In the following, we want to choose τ_c such that we reject H_0 if the measured rate is larger than this threshold τ_c.

   i. You propose to reject H_0 as soon as the measured rate is larger than 45. What is the probability of wrongly concluding that an athlete is doped?

   ii. The previous probability of accusing an innocent person is far too large: no jury would convict a defendant knowing it had such a chance of sending an innocent person to jail! So you decide to change the rule and propose to reject H_0 as soon as the measured rate is larger than 60. What is the probability of wrongly concluding that an athlete is doped?

   iii. Setting τ_c that large errs in the other direction by giving the null hypothesis too much benefit of the doubt. We then decide to choose τ_c such that the probability of wrongly concluding that an athlete is doped, i.e. the probability of the type I risk, equals 0.05. Find such a threshold τ_c. 0.05, usually denoted by α is called the level of significance of the test, it is the probability to reject H_0 while it's true.

(f) The *sixth step* is to write the test, i.e. summarizing the previous steps by specifying the *rule of rejection for* H_0 and then *conclude considering the observed values*.

   i. Summarize the previous conclusions into a decision rule.

ii. What are your conclusions for J.C. and S.R.?

iii. When you observe a measure leading to the rejection of H_0, it doesn't prove that the athlete is doped. In other words, it must be remembered, that rejecting H_0 does not prove that H_0 is false, no more than a jury's decision to convict guarantees that the defendant is guilty. It just means that if the true rate is 45, measurements of this rate as large or larger than τ_c are expected to occur only 5% of the time. Because of that small probability, a reasonable conclusion when the measurement of this rate is as large or larger than τ_c is that the true rate is larger than 45. Check that with your rule you wrongly reject H_0 with an average frequency of 0.05.

(g) As a seventh step, you could check the power of the test at the end. You could also compute the p-value (PC8) for the observed value.

i. Plot the power function associated to your hypotheses test.

ii. What is the probability of detecting an abnormal hematocrit level when it is equal to 50?
Note that the choices for steps 3 and 4 are optimal (best power) using Neyman-Pearson theory in the case of simple hypotheses.

4. **Data analysis (5pts)**.

We are interested in the weights of adult female octopuses. For this we have a sample of 240 female octopuses caught off the coast of Mauritania. We want to know an estimate of the mean of the weight and a confidence interval for this mean at the level 95%.

(a) Load dataset `poulpeF.csv` into R.

(b) Under the assumption of a Gaussian model, calculate in R the maximum likelihood estimator for the mean and the variance of the data.

(c) Plot the histogram. Is the Gaussian model reasonable ? Is it a problem ?

(d) till using the Gaussian model, calculate a 95% confidence interval for the average octopus size.