

Elements of Statistics and Econometrics

Assignment 2

Problem 3: Inferential statistics

1. We start with the verification of the law of large numbers. Thus we check if an estimator converges (in probability) to its true value if the sample size increases.
 - (a) Simulate samples of size $n = 100, \dots, 100000$ (with step say 1000) from a normal distribution with mean 1 and variance 1, i.e. $N(1, 1)$. For each sample estimate the mean, the variance. Plot the path of sample means as function of n . What conclusion can we draw from the figure if we keep in mind the law of large numbers?
 - (b) Frequently it is difficult to obtain more data. How many observations do we need in order to obtain an estimator which is close enough (± 0.01) to the true value?
 - (c) Add to the plot the 95% confidence intervals. Do it ones with known σ and ones with an estimated. The confidence intervals have to be constructed manually. Provide their interpretation.
 - (d) Next plot the sample variance as a function of n . If we think about the consistency of the sample variance as an estimator for σ^2 , does the figure support this property? Give the interpretation of consistency in your own words.
2. The objective of this part is to get a better feeling for the ML estimation procedures. The estimation for non-standard distributions/models usually follows the maximum-likelihood principle. The t -distribution is a popular alternative if the sample distribution is symmetric, but exhibits heavier tails compared to the normal distribution.
 - (a) Let x_1, \dots, x_n be a given sample. We assume that it stems from a t -distribution with an unknown number of degrees of freedom. Write down the corresponding log-likelihood function. The density function of the t_{df} -distribution is given by

$$f(x) = \frac{\left(1 + \frac{x^2}{df}\right)^{-\frac{df+1}{2}}}{B(df/2, 1/2)\sqrt{df}},$$

where $B(\cdot, \cdot)$ is the beta function (`beta(a,b)` in R.)

- (b) Simulate a sample of size $n = 100$ from t_5 . Maximize the log-likelihood function (numerically) for the given sample and obtain the ML estimator of the number of degrees of freedom. Compare the estimator with the true value.
 - (c) Increase the sample to $n = 5000$ and compare the new estimator with the true value. Which property of the estimator we expect to observe?
3. Next we assess the asymptotic distribution of estimators (in the sense of the central limit theorem).
- (a) Simulate $b = 1000$ samples of size $n = 100$ from a χ_2^2 distribution with mean 1 and variance 1, i.e. $N(1, 1)$. For each sample estimate the mean, the variance and keep them. Plot the histogram for one of the sample, so that you get a better feeling how the original distribution looks like.
 - (b) Plot the KDE or a histogram for the sample of means and the sample of variances. Add the density of the normal distribution for comparison purposes. Note, that the density Compare the density estimator with the normal density. What do you expect and why (statistical reasoning!)?
 - (c) In the lecture we discussed the CLT for the sample mean. Here it seems to apply to the sample variance too. Why?
 - (d) Let n take values 10^3 , 10^4 , 10^5 and 10^6 . Check the impact of n on the results. Can the statement of the CLT be confirmed?
4. The next aim objective is check if the probability of type 1 error (size of a test) is correctly attained by a simple two-sided test for the mean.
- (a) Simulate a sample of length $n = 100$ from a normal distribution with mean $\mu_0 = 500$ and variance $\sigma^2 = 50$. (Note: you may use the transformation $X = \mu + \sigma Z$, where $Z \sim N(0, 1)$.) The objective is to test the null hypothesis $H_0 : \mu = 500$. Assume that σ^2 has to be estimated. Compute the test statistics using the formulas in the lecture; determine the rejection area for $\alpha = 0.04$ and decide if H_0 can to be rejected.
 - (b) Determine the p -values using the formulas from the lecture and compare/check the results using a build-in function for this test in R or Python. Give a verbal interpretation of the obtained p -value.
 - (c) Simulate $M = 1000$ samples of size $n = 100$ and with $\mu_0 = 500$ and variance $\sigma^2 = 50$. For each sample i run the test (using a standard function) and set $p_i = 0$ if H_0 is not rejected and $p_i = 1$ is rejected. Compute $\hat{\alpha} = \frac{1}{M} \sum_{i=1}^M p_i$. $\hat{\alpha}$ is the empirical confidence level (empirical size) of the test. Compare $\hat{\alpha}$ with α . Do you expect the difference to be large or small and why? Relate it to the assumptions of the test.
 - (d) Assume now that one of the assumptions is not satisfied. For example, the data is in fact not normal. Repeat the above analysis, but simulate a sample z_1, \dots, z_n from t -distribution with 3 degrees of freedom. Compute $x_i = 500 + z_i/\sqrt{z_i}$. (Note: This will guarantee the same expectation and the same variance as the above normal distribution.) Recompute a new $hata\alpha$. What do you expect? Relate your answer to the type one error and the underlying assumptions.

(e) **Power of a test:** The first objective is to assess the probability of type 2 error (power of a test) of goodness-of-fit test. Goodness-of-fit tests for the normal distribution are of key importance in statistics, since they allow to verify the distributional assumptions required in many models. Here we check the power of the Kolmogorov-Smirnov test, i.e. is the test capable to detect deviations from normality?

- Simulate $M = 1000$ samples of size 100 from a t -distribution with $df = 2, \dots, 50$ degrees of freedom. For each sample run the Kolmogorov-Smirnov test and count the cases when the H_0 of normality is correctly rejected (for each df). How would you use this quantity to estimate the power of the test? Make an appropriate plot with the df on the X-axis. (Note: the t -distribution converges to the normal distribution as df tends to infinity. For $df > 50$ the distributions are almost identical.) Discuss the plot and draw conclusions about the reliability of the test.