Elements of Statistics and Econometrics - 2021

Assignment 2

Problem 4: Inferential statistics

1. We start with the verification of the law of large numbers and of the central limit theorem. Thus we check if the mean $\bar{X}$ converges (in probability) to the expectation μ if the sample size increases and if its distribution converges to the normal distribution.

   (a) Simulate samples of size $n = 10, \ldots, 100000$ (with step say 5000) from a normal distribution with mean 1 and variance 1, i.e. $N(1, 1)$. For each sample compute the sample mean and the sample variance. Plot the path of the sample means as a function of n. What conclusion can we draw from the figure if we keep in mind the law of large numbers?

   (b) Heuristic perspective: just by observing the figure, how many observations do we need in order to obtain an estimator that is close enough to the true value $(\mu \pm 0.01)$?

   (c) For each sample compute manually the 95% confidence interval for the mean and add it to the plot. Do it ones with known σ and ones with an estimated. Discuss what you observe and provide explanation for your findings.

   (d) Simulate $b = 1000$ samples of size $n = 5$ from some weird distribution, for example uniform or χ^2 distribution with $df = 1$. For each sample estimate the mean, the variance and store them. Plot the histogram for the sample of means and the histogram of the sample of variances. Add the densities of the normal distribution for comparison purposes. Compare these densities with the histograms. What do you expect and why (statistical reasoning!)?

   (e) 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?

   (f) In the lecture we discussed the CLT for the sample mean. Here it seems to apply to the sample variance too. Why?

2. (This is an optional problem. You get additional points for solving it.) The objective of this part is to get a better feeling for the ML estimation procedure. Consider the daily UAH/Euro exchange rate for the last year. The data can be downloaded, for example, from www.investing.com. Financial data typically has heavy tails and better fits the t-distribution, than the normal distribution.

(a) Compute the log-returns of the exchange rate: $r_t = ln(ex_t/ex_{t-1})$, where ex_t is the exchange rate on day t. Using your software run a suitable goodness-of-fit test (for example, Kolmogorov-Smirnov) to verify if the data is normally distributed. Write down the hypotheses, the test statistics and the p-value. What do we conclude?

(b) Now assume that the returns follow a (generalized) t-distribution with some unknown degree of freedom. The density function of the t_{df}-distribution is given by

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

where $B(\cdot,\cdot)$ is the beta function (`beta(a,b)` in R.) Write a code for the log-likelihood function evaluated at the sample of returns, i.e.

$$\ln L(\mu, \sigma, df) = \sum_{t=1}^{T} \ln \left(\frac{\left(1 + \frac{1}{df}\left(\frac{r_t-\mu}{\sigma}\right)^2\right)^{-\frac{df+1}{2}}}{B(df/2, 1/2)\sqrt{df}\sigma} \right)$$

and maximize it w.r.t. (μ, σ, df). Use a build-in optimization function. Note, that $\sigma > 0$ and $df > 0$.

(c) Using the estimated parameters compute the probability that

- the exchange rate return is between -0.01 and 0.01;
- the exchange rate return is larger than the mean;
- the exchange rate return is negative.

Recall for this task that $P(a < X \leq b) = F(b) - F(a)$ for any distribution function F. Use again the build-in function for the cdf of the generalized t distribution.

3. Here we run a few tests for the data used in the first assignment. For every test write down: the two hypotheses; the formula for the test statistic with inserted values; rejection area computed manually; p-value from your software; your decision and the verbal conclusion (not just sth like "H_0 is rejected"). Use for all tests $\alpha = 5\%$.

(a) The real estate company argues that the mean sale price is larger than 180000. Can this be confirmed with an appropriate test?

(b) We would like to check if the sale price of two-story houses is significantly different than the sale price of one-story houses (`HouseType`). Assume that the variances are unknown and not necessarily equal.

(c) We suspect that the sale price is not related to the year of construction (`YearBuilt`). Check this using the Pearson correlation coefficient.

(d) Test if the fraction of one-story houses (check the feature `House-Style`) is significantly higher than 50%.

(e) All the above tests rely on the normality assumption of `SalePrice`. Check this assumption using the Kolmogorov-Smirnov test. Use only built-it functions, no manual computations here.

4. The next 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 be rejected.

   (b) 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 $p_i = 1$ if 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.

5. **Power of a test**: The next aim is to assess the probability of type 2 error (power of a test) of a 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 $df = 2, ..., 50$ 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.