Lab FDSAI

The χ^2 goodness-of-fit test

1. A user wishes to study the effectiveness of the anti-spam program used to protect their email account. To this end, over 50 consecutive days, he records the number of spam emails (file spam.txt).

   Conduct a goodness-of-fit test at a significance level of 1% to verify whether the number of spam emails follows an exponential distribution.

The Kolmogorov-Smirnov test

1. In the file `pretcarburant.csv`, you will find the daily price trends for gasoline and diesel over the years 2020-2021 in Romania (Oct. 22, 2020 - Oct. 20, 2021, source: `http://www.peco-online.ro`). The first column represents the daily gasoline price, and the second column represents the diesel price.

   (a) Use the Kolmogorov-Smirnov test to assess whether the prices of the two fuel types follow any known distribution (Normal, Log-Normal, Gamma, Weibull, Raleigh).

   (b) Using the Kolmogorov-Smirnov test, check whether the two samples come from the same distribution.

2. A random variable X is distributed *log-normally* with parameters μ and σ if $\ln X \sim N(\mu, \sigma)$. Use a Kolmogorov-Smirnov test at a significance level of 0.05 ($k_{0.95} = 1.36$) to determine whether the following lifespans (in days) of laboratory mice, obtained from a study of a cancer treatment, originates from a log-normal distribution with parameters $\mu = 3$ and $\sigma = 4$:

$$24, 12, 36, 40, 16, 10, 12, 30, 38, 14, 22, 18$$

The bootstrap method

1. Over the 5 working days of the week, a student spends 2, 2, 3, 3, and 5 hours, respectively, on homework.

   (a) How many bootstrap resampling sets exist?

   (b) Using suitable software (for example, Octave, Matlab, Python, or another), generate $r = 1000$ samples and determine the bootstrap distribution of the sample median.

   (c) Using the distribution determined in the previous step, estimate the variance of the sample median and the bias (the shift from the actual median time spent on homework).

2. The following data represent the GRE scores obtained by a sample of 16 candidates from Bucharest:

$$5.22, 4.74, 6.44, 7.08, 4.66, 5.34, 4.22, 4.80, 5.02, 6.55, 4.18, 4.64, 6.00, 4.12, 5.30, 5.64$$

   Using suitable software (e.g. Matlab, Python, or another), generate $r = 10000$ samples and determine the bootstrap distribution of the sample mean. Then estimate the average score of Bucharest candidates.

3. (Hastie&Tibshirani, *Statistical Learning*) This exercise uses the `Boston` dataset from the Python `sklearn` library, which contains information on housing in Boston.

   (a) From this dataset, extract the feature `medv` (the median value of owner-occupied homes, in thousands of dollars). Let $x_1, \ldots, x_n$ be the sample obtained this way. Estimate the mean value of the `medv` feature, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

   (b) Estimate the standard error of the estimator $\overline{X}$ using the formula $\widehat{MSE}(\overline{X}) = \frac{s^2}{n}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is the sample variance. Interpret the result.

   (c) Now estimate the standard error of $\overline{x}$ using the bootstrap method by generating $B = 10000$ bootstrap samples, and compare this result with the one obtained in the previous point.

(d) Based on the bootstrap estimation from point 3c, construct a 95% confidence interval for the mean of the `medv` feature. Compare this result with the 95% confidence interval built solely from the initial sample (since the variance is unknown, use the quantiles of the Student distribution $t_{\alpha;n}$).

(e) Estimate the median value of the `medv` feature based on the initial dataset. Let $\widehat{M}$ be this estimator.

(f) As we did above for the mean, we now seek an estimate of the mean squared error of the median estimator. Unlike the previous case, there is no formula for calculating $\widehat{MSE}(\widehat{M})$. Instead, use the information from the bootstrap samples to estimate $MSE(\widehat{M})$ using the formula provided by the quantiles of the bootstrap distribution determined in point 3c.