ISyE 6644 – Spring 2023

Homework #4 — Due Monday, March 27

Follow the instructions on the syllabus to submit a single archive containing all codes and a single PDF document containing your discussion and findings. Name the archive as `LastName-FirstName.zip`.

1. For each of the following distributions, derive formulas for the MLEs of the indicated parameters, assuming that we have IID data $\{X_1, \ldots, X_n\}$ from the distribution in question. The distributions are tabulated in Tables 6.3–6.4 of Law (2015) and in the User Guide of ExpertFit. In each case compare the MLE(s) with the estimator(s) obtained from the method of moments.
   (a) Lognormal, joint MLES for μ and σ^2. Hint: The log-data are IID normal.
   (b) binomial(t, p), MLE for p assuming that the number of trials t is known.

2. Consider the two-parameter exponential distribution with density function

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-(x-\gamma)/\beta}, & \text{if } x \geq \gamma \\ 0, & \text{otherwise,} \end{cases}$$

   where γ is the location parameter and $\beta > 0$ is the scale parameter. Given an IID sample $\{X_1, \ldots, X_n\}$ from this distribution, find the MLEs of γ and β. Hint: Remember that γ cannot exceed any data point X_i. Obtain the log-likelihood function, and for β fixed, find the value of γ that maximizes the log-likelihood function.

3. The attached text file `6644hw4s23-A.txt` contains loading times for vessels at a seaport, in days. Import the dataset into ExpertFit, choose the *Advanced* mode, and attempt to fit a distribution to this data set. Feel free to examine both unbounded and bounded distributions. Create a concise report that includes
   - the table with the summary statistics;
   - a scatter plot and a lag-correlation plot to asses the independence of the data;
   - the formula of the fitted density function and the estimates of the model parameters (look in the ExpertFit User's Guide);
   - the P-P plot;
   - tables from the Anderson-Darling and Kolmogorov-Smirnov tests;
   - the Simio expression for generating observations from this distribution; and
   - a brief paragraph assessing the appropriateness of the proposed model.

   Then redo this problem assuming that the loading times have a lower bound of 0.5 days. Do you get a better fit?

4. The attached file `6644hw4s23-B.txt` contains demands for a product during consecutive weeks in a calendar year.

Import the dataset into an ExpertFit and specify that the data are integer-valued. Attempt to fit a discrete distribution to this data set. If the fit is good, create a concise report that includes

- the table with the summary statistics;
- a scatter plot and a lag-correlation plot to asses the independence of the data;
- the formula for the fitted probability mass function and the estimates of the model parameters (look in the ExpertFit User's Guide);
- the P-P plot;
- the table from the equiprobable chi-square goodness-of-fit test (feel free to alter the number of cells provided that the expected cell count is ≥ 5);
- the Simio expression for generating observations from this distribution; and
- a brief paragraph assessing the suitability of the proposed model.

Otherwise, justify your objection to the best model selected by ExpertFit using the P-P plot and the table from the equal-width chi-square goodness-of-fit test.

5. A potentially better model for the data in Problem 4 is the negative binomial distribution with a non-integer parameter r; see http://en.wikipedia.org/wiki/Negative_binomial_distribution. We covered this model during the review of probability distributions. This model has probability mass function

$$f(k; p) = \Pr(X = k) = \frac{\Gamma(r + k)}{\Gamma(r)k!} p^r (1 - p)^k, \quad k = 0, 1, 2, \dots,$$

mean $\mu = r(1-p)/p$ and variance $r(1-p)/p^2$. Further, it allows matching of the first two moments and can accommodate coefficients of variation (standard deviation over mean) that are significantly larger that 1.

The web site http://www.wessa.net/rwasp_fitdistrnegbin.wasp fits negative binomial models and computes estimates of the "size" parameter r and the mean μ. Upload the data and use the estimates of r and μ to estimate p and the variance. Your report should contain a screen capture with the fitted parameters. Are the latter estimates close to the sample moments?

6. The attached file `6644hw4s23-C.txt` contains repair times for a piece of equipment (in minutes).

   (a) Using ExpertFit, attempt to find a distribution that fits the data well. A good model should pass the Anderson-Darling and Kolmogorov-Smirnov tests at various levels of significance α. If ExpertFit fails to find a good model, give a reasonable explanation.

   (b) Compute a kernel density estimate (KDE) using Python (look for a function), R (look at the documentation in https://www.rdocumentation.org/packages/ks/versions/1.10.7 for the `ks` package and the `kde` function within), or the online tool at www.wessa.net/rwasp_density.wasp. You can use the Normal$(0, 1)$ kernel and the bandwidth h supplied by the package (feel free to experiment). Some packages (such as the `kde` function of R allow for restricted domains).

(c) Use the KDEVC algorithm ("VC" stands for variance correction) of Hormann and Leydold (2000) (see http://informs-sim.org/wsc00papers/089.PDF) to generate 1000 observations from the KDE in part (b). The quantity σ_K^2 is the variance of the symmetric kernel; hence it is equal to 1 for the Normal$(0, 1)$ kernel. To avoid generating negative observations, use the "mirroring" trick in the Remark following the KDEVC algorithm. Then create a histogram for the generated data and compare it with the KDE in part (b).

7. Use the method discussed in class to fit a lognormal distribution with location parameter $\gamma = 9$, a most-likely value (mode) $m = 22$, and a 95th percentile $x_{0.95} = 30$. List the fitted parameters $\tilde{\mu}$ and $\tilde{\sigma}$ and, if possible, plot the fitted density and CDF. Finally, compute estimates of the mean and standard deviation of the fitted model.

8. **[Bonus problem, 3 points]** Suppose that $\{X_1, X_2, \ldots, X_n\}$ is an independent sample from the beta distribution with density function $f(x; p) = p(1-x)^{p-1}$, $0 < x < 1$, and an unknown shape parameter $p > 0$.

   (a) Find the mean of this distribution and use the method of moments to obtain an estimator of the power p.
   (b) Find the maximum likelihood estimator (MLE) of p. Verify that the estimator maximizes the log-likelihood function.
   (c) Consider the following data set:
   0.08, 0.10, 0.26, 0.35, 0.39, 0.24, 0.47, 0.11, 0.37, 0.09
   Use the Kolmogorov-Smirnov test with type I error $\alpha = 0.10$ to assess the fit of the beta distribution with $p = 3$. Build a detailed table similar to the one on the slides (for the exponential distribution).