[Print: May 18, 2020]
[]

Statistics Project Directions

Along with this set of directions you will find a `.csv` file containing one column of 2,500 numbers. These numbers represent sample values drawn from one of the following distributions [the distribution's parameters appear below the name in this table]:

Normal	Gamma	Beta
$\mu \quad \sigma^2$	$\alpha \quad \beta$	a b
Chi-square	Continuous Uniform	Exponential
ν (d.f.)	$a \quad b$	β
	Weibull	
	$\alpha \quad \beta$	

In creating these data sets, the notation and definitions found in the course textbook were followed. Many implementations of the Weibull distribution were found that have different meanings for the two parameters. [See the Statistical Goodies write-up for more information].

Tasks.

1 Create a histogram from your data using all 2500 data points. The graph you see is approximately proportional to the density function of one of these distributions. Report the span of the histogram (that is, the values: left edge of the leftmost bin and right edge of the rightmost bin), number of bins and the bin width.

2 Look for graphs of the density function for each of the seven distributions listed above. The parameters of the distribution are going to affect the shape of the density curve so you may need to look at several varieties of curves for each distribution. [Wikipedia is a good source for this information.] Based on this search pick out no more than two of those distributions that might be candidates for the distribution found in the `.csv` file. [For example, based on the shape of the density function, you might conclude that the mystery distribution is likely either exponential or Weibull.]

3 For each of the candidates you found in step (2), note the distribution's parameters that are going to be required to identify this distribution exactly – they are noted in the table above. Estimate the mean and variance for each distribution you are considering.

4 Use the method of moments (or other techniques you might know) to estimate the parameters of each distribution under consideration.

5 Using the estimated parameters from item 4, perform a goodness of fit significance test to try to resolve the question concerning the type of distribution exhibited by the data. You will need to perform one goodness-of-fit *significance* test for each distribution under consideration.

6a If all of the P-values from step 5 are below 90%, locate the data set that had the largest P-value and perform a goodness of fit *hypothesis* test on that data to resolve the issue of its distribution. Report the P-value and the stat value. Even if you cannot get a positive resolution, report your findings and go on the step (7)– the distribution you found with the best goodness of fit P-value will stand in for your candidate distribution.

6b If some P-values from step (5) are above 90%, assume that the data set with the largest of these P-values represents the distribution of the data. This is your candidate distribution. Report the P-value you found and the stat value.

7 Report the name of your candidate distribution and its parameters. Scale the histogram of the candidate distribution from part 1 so that it best represents a density function (that is, convert it from a raw histogram to one normalized to unit area). Plot the normalized histogram next to the theoretical density function for your data set based on your analysis and the estimates of the distribution's parameters. [Ideally both of these graphs will appear on the same set of axes.]

8 Assuming that your estimated variance is really incredibly close to the true variance – that is $\sigma^2 = S^2$, something that almost never happens. Give a 96% confidence interval around the mean of your candidate distribution. With the variance assumption in place, determine 96% confidence intervals around the parameters of the candidate distribution. [You'll need to use the method of moments, the CI around the mean, and some algebraic fiddling to make this work.]

The Work.

Many spreadsheets contain packages that can do a lot of these calculations for you. Avoid using these since a computer of any kind is unavailable for the final exam – it is best to learn how to perform these computations from the basic ideas. You may use spreadsheet functions, such as `NORMINV`, `NORMDIST`, `CHIINV`, `TINV`, `AVERAGE`, `STDEV`, `VAR` (to name a few) and most spreadsheets have implementations of `GAMMA`, `BETA` and `WEIBULL` but note that there are inconsistencies in the definitions of the parameters with respect to the `WEIBULL` distribution).

Do not use spreadsheet data analysis functions to determine histograms, confidence intervals, hypothesis/significance tests, and so on. For example, histograms may be constructed by judicious use of the `INT` and `COUNTIF` spreadsheet functions, not the `HISTOGRAM` function. Most of these procedures needed for this project are written up in the Statistical Goodies notes.

The Report.

In your report, summarize your findings. In addressing the previous 8 items, give all of the information required so that someone could duplicate your calculations. Use formulas, graph, charts, and tables where appropriate. Describe any estimators, equations, theorems, etc. used in performing the confidence interval estimates. Explain what you are doing for the significance/ hypothesis tests and interpret the results. Include any remarks about the data (or the results) that you feel are pertinent.

It isn't necessary (or desired) to turn in pages and pages of spreadsheet computations. The results and your interpretations of them must suffice. Above all, *make this report readable!* . . . omit needless words, be succinct!