[Print: November 9, 2017]
[]

Statistics Project Directions

Along with this set of directions you will find a `.csv` file containing three columns of numbers. The name of the file is one of `CUcccxx.csv, Gcccxx.csv, Bcccxx.csv`. The first letter(s) of the file name indicates the type of distribution contained in the first column of the table (`CU` means continuous uniform, `G` means gamma, and `B` means binomial). The characters `ccc` indicate the presence of some number of other characters and the digits `xx` (or in some cases only `x`) are an identification number for the file.

Column A. When the file is opened with Excel (or any other spreadsheet that can read `.csv` files) the first column should contain these items:

```
Distribution Name
Distribution Variance
<a blank cell>
data point 1
data point 2
    ...
data point n
```

Column A Tasks.

A.1 For the type of distribution presented in the first column, plot a histogram (normalized to unit area) of the data.

A.2 Estimate the mean and the variance for the column A distribution. For comparison, the exact variance is given in the cell labelled (here) `Distribution Variance`.

A.3 The distribution found in the first column depends on two parameters as follows:

Continuous Uniform	a and b (the endpoints of the interval)
Gamma Distribution	α and β
Binomial Distribution	n and p

Estimate the two parameters associated with the distribution given in Column A and determine a 96% confidence interval around each parameter.

A.4 How large a data set is needed to get 96% confidence intervals of width 0.01 or smaller around the two parameters? (Assume $\bar{X}$ and S^2 do not change significantly with N when N is large.)

A.5 Plot a graph of the density function for the distribution in column A using the estimated parameter values determined in part A.3. Compare this graph to your normalized histogram.

Column B. The second column is similar to the first and looks like this:

```
Normal
<a blank cell>
<a blank cell>
data point 1
data point 2
    ...
data point 10000
```

This column contains 10,000 values from some Normal distribution.

Column B Tasks.

B.1 From the B column select two non-overlapping chunks[1] of consecutive data points. The first chunk should contain a large number of data points. The second chunk should contain exactly 25 data points. Estimate the mean and variance of this normal distribution using your first (large) chunk of data.

B.2 Compute 98% confidence intervals around each of the parameters μ and σ based on the large chunk of data.

B.3 Test the claim $\mu \geq 4$ using the second (small) chunk of data in a significance test. Report the P-value and the value of the statistic (the t-stat or a z-stat value as appropriate).

Column C. The third column is similar to the second column and looks like this:

```
Uniform
<a blank cell>
<a blank cell>
data point 1
data point 2
    ...
data point 1000
```

This column contains 1,000 random values that follow a continuous uniform distribution on $[0, \frac{1}{2}]$. Call this list of uniform samples U-samples.

Column C Tasks.

C.1 The situation: A 2 Hz cosine wave is being used to generate random values between -1 and $+1$. This is accomplished by selecting a time randomly and uniformly from the interval $[0, 0.5]$ and evaluating the amplitude of a particular 2 Hz cosine waveform at that time.

The U-samples of Column C represent times chosen randomly and uniformly from the interval 0 to 0.5 seconds. These times are used to generate C samples by evaluating the expression $C = \cos(4\pi U)$ at each U sample. Using these C-samples, produce a histogram that closely approximates the *density function* for the random variable C.

C.2 Estimate the mean and the variance of C and determine a 92% confidence interval about the mean.

C.3 The following reasoning is applied: the graph of the given cosine wave seems to lie above the x-axis as much as it lies below. It would seem plausible then to assume that the mean of C is 0. Perform a significance test on the claim $\mu = 0$ using all 1,000 samples. Report the P-value, and comment on whether your data supports the guess about the mean or not.

C.4 The actual density for C is given by $f_C(c) = 1/(\pi\sqrt{1-c^2})$ for $-1 < c < 1$. Plot a graph of the actual density and compare it with your Part C.1 histogram. What are the theoretical mean and variance of C?

[1] a technical term... don't try to understand this!

The Work.

Divide up the responsibility for solving these problems among the team members. In your report, explain who was responsible for which aspects of the work. In the end, it is important for everyone to understand how the problems were solved — this will make the task of preparing for the final exam easier.

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` (to name a few), in lieu of the tables in the back of the book. Make sure you know what these functions are reporting. Do not use Excel 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.

The Report.

Write up (in a few pages) who performed the analysis and how the data was analyzed. Describe any estimators, equations, theorems, etc. used in performing the confidence interval estimates. Explain what you are doing for a significance test 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. Use graphs, charts, and summary tables as needed to support your work. Above all, *make this report readable!* . . . omit needless words, be succinct!