STATISTICS PRACTICAL 1

(1) 100 students each perform an experiment to estimate a parameter μ, and each one independently computes a 99% confidence interval for μ. What is the probability that there will be at least 3 students whose confidence intervals do not include μ? (Hint: Use the binomial distribution or the Poisson distribution.)

(2) Suppose we have a new IQ test, which we wish to compare to a standard IQ test, to see whether the scores on the new test have the same average as the scores on the old test. We take 200 subjects, and pick 120 at random to take the new test; the other 80 take the old test. The results are as follows:

Test	# subjects	Mean	SD
Old	80	101	11
New	120	98	8

We wish to do a Z test at the 0.01 level to determine whether the means on the tests could be the same.

(a) State the test hypotheses.
(b) State any assumptions you have to make.
(c) Calculate the test statistic, compute the p-value.
(d) Compare to the critical value and conclude.

(3) Suppose you have 4 independent observations from a normal distribution with unknown mean and unknown variance. You estimate the mean and SD from the data, and use them to compute the Student t statistic $T = (\bar{X} - 1)/(\hat{\sigma}/\sqrt{n})$. You report it to someone, who mistakenly believes that the variance is known, and thus performs a Z test for the hypothesis that the mean of the distribution is 1. What will the impact be on the statistical conclusions? What if the error goes in the other direction (that is, you compute from a known variance, but someone believes you have delivered a T statistic)? Carry out some simulations to demonstrate the error that will be made.

(a) If the person is performing two-tailed hypothesis tests at level 0.05, calculate the size of the error in each case.
(b) Carry out some simulations to demonstrate this.

(4) ***E. coli* descriptive statistics in `R`** (Optional, for morning or afternoon.) In the following sequence of questions will be using data from the completely sequenced genome of *E. coli*, more precisely strain K-12, substrain MG 1655, version M52 short. *Adapted from a lab written by Terry Speed and Bin Yu.*

(a) **Background and raw data.**
The sequence data have come from the NCBI website: `www.ncbi.nlm.nih.gov.` You may download the data by typing
`load(url('http://steinsaltz.me.uk/DTC/ecoli.rda'))`.
What was originally a single string of length 4641652 has been processed by the command `strsplit(x,'')` to create a vector `ecp` of length 4641652, each entry consisting of one letter of A, C, G, T.

(b) Count the number of Adenines, Guanines, Cytosines and Thymines (As, Gs, Cs and Ts). Then there is the dinucleotide composition, that is, the number of AAs, AGs etc. (along one strand). There further is the trinucleotide composition, that is, the number of AAAs, AAGs etc.

(i) If you had to assign a probability to observing an A at a stated position on the *E. coli* genome, what figure would you use and why? Under what assumptions does this seem appropriate?
(ii) Type `L=length(ecp)` followed by `table(ecp[-L],ecp[-1])`. What is this telling you?
(iii) Your are told there is an A at a position along the *E. coli* genome. What probability would you assign to it being followed by a G, and why?
(iv) Does the *E. coli* composition data suggest that the event we observe a G at one site is independent (in some suitable sense) of the previous two bases? Explain fully, illustrating with appropriate data.

(c) **Purine counts.**
Divide up the data into about 46,400 blocks of 100 base pairs (bp). Count the number of purines (i.e. A or G). Do the same for the 4,640 blocks of 1,000 bp, and 464 blocks of 10,000 bp.
(i) For each set of counts, calculate the mean and standard deviation of the number of purines per block, and draw histograms of these numbers.
(ii) Compare the results of (a) across the different block sizes and comment.
(iii) For each block size, calculate the fraction of the counts within 1, 2 and 3 standard deviations of the mean.
(iv) Repeat (a), (b) and (c) for *proportions* (rather than counts) of purines in each block.

(d) Compute counts of TATAAT in blocks of 5,000 bp. Assuming that these counts follow a Poisson distribution, estimate the parameter of this distribution and obtain an estimate of the standard error of your parameter estimate. This can be done by either a formula or by simulation.