“Pencil and Paper” Assignment 1

1 Computing "Measures"

1.1 Calculation Method

Suppose you have a data set consisting of n values x_i. Define the sum of these values, and the sum of their squares

$$s_n = \sum_{i=1}^{n} x_i$$

$$q_n = \sum_{i=1}^{n} x_i^2$$

Show that the following is true

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \frac{q_n}{n} - \frac{s_n^2}{n^2} = \frac{1}{n}\sum_{k=1}^{n}\left(x_k - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

thus providing an alternate formula (the leftmost expression) for computing the “population variance” (the rightmost expression): “*the mean of the squares minus the square of the mean”*.

Hint: You should expand the square in the rightmost formula (do it correctly!), and distribute the common factor $\frac{1}{n}$ through to each of the three terms. Note also the result of an expression like $\sum_{i=1}^{n} c = nc$, that is, if you have a sum of n terms, and they are all equal, the result is n times that value.

Note: Even if you cannot prove the formula, keep it in mind

Remark: This identity has more than one useful consequence. First of all, note how the rightmost expression is never negative (and zero only when all the values x_k are equal), hence the same is true for the leftmost expression – a non obvious fact. Further, note how knowledge of s_n and q_n is enough to compute both sample mean and “population variance”. Of course, if S^2 is the “population” variance, the “sample” variance is found immediately: $s^2 = \frac{n}{n-1} S^2$

1.2 Calculate Measures

A statistician reports that data a set consists of 150 values, with the following summaries

- Sum of the data 535
- Sum of the squares of the data 1,048

This cannot be. Why?

2 Manipulating Probability Distributions

"Grading on a Curve"

The precise procedure for "grading on a curve" is the following. Your instructor assumes that the outcome of an assessment should be normally distributed, with a certain mean (say, μ), and a certain standard deviation, say σ. The actual result will have an average of, say, m and a sample standard deviation of, say s. Now, each grade is shifted so that the "corrected" average becomes μ, and the "corrected" standard deviation σ. Needless to say, this is quite arbitrary, since, among many other objections, it is far from obvious that a normal distribution is necessarily a good model for classroom performance.

Anyway, suppose n students take a test, resulting in grades x_k ($k = 1, 2, ..., n$). The "curving" transformation can be performed as follows:

1 "Standardize" each grade, so that the new list has average 0 and "population standard deviation" 1 (note that in no way can we think of this data set as a "sample")

2 Rescale each standardized grade by the reverse formula that changes a standard normal variable to a normal variable with assigned mean and variance.

2.1 What is expected

Suppose grades in a class are given as percentages, with an average of 0.55, and a standard deviation of 0.25, while the "theoretical" numbers are, in the mind of this instructor would be 0.7 as average and 0.1 as standard deviation. Assuming these parameters reflect the instructor's expectations, what is the probability, according to these expectations, that a student might score

1. less than 0.5
2. more than 0.9

2.2 Curving the Scores

What is the "corrected" grade for a student who scored 0.4 and for a student who scored 0.85?