

5. Gene A, gene B and gene C are known to code proteins that function as a group. Studies have shown that the expression levels of these genes (denoted by A, B, C respectively) fluctuate from person to person, with mean 1.2 (for A), 1.0 (for B), 1.0 (for C), and standard deviation 0.5 (for A), 0.4 (for B), and 0.3 (for C). The correlation coefficients are 0.6 (between A and B), 0.7 (between A and C), and 0.8 (between B and C). A biologist was interested in comparing their expression levels. In particular, he first took the average of B and C because gene B and gene C are homologous. Then he used the difference between A and $(B+C)/2$ to measure the lack of coordination in expression. In fact, he found that if the difference is greater than 0.5, then it is a sign of developing one type of metabolic pathway dysfunction called MPD1. Assume that the distributions are normal. (a) What percentage of people are likely to develop MPD1?

(b) What would your answer be if A, B, C were assumed to be independent?

6. The survival lifetimes of patients after operation or after chemotherapy are recorded for one type of cancer. Assuming that decision for operation or chemotherapy is purely random. For the 100 patients in the operation group, mean survival time is 4.5 years and SD is 0.5 years. For the 400 patients in the chemotherapy group, the mean survival time is 4.8 years and the SD is 1.0 years.

(a). Find a 95% confidence interval for difference in mean survival time between chemotherapy and operation.

(b). Two children in a family have developed this type of cancer. The older brother chose chemotherapy and the younger brother chose operation. What is the probability that the younger brother survive longer than the older brother? What assumptions are made in your calculation? Are they reasonable?

7. A survey organization takes a random sample of 625 households from a city of over a million households. On the average, there are 2.30 persons per sample household and the standard deviation is 1.75. Say whether each of the following statement is true or false, and explain.

(a). The SE of the mean is 0.07.

(b) A 95% confidence interval for the average household size in the city is 2.16 to 2.44.

(c) A 95% confidence interval for the average household size in the sample is 2.16 to 2.44.

(d) 95% of the households in the city contain between 2.16 and 2.44 persons.

(e) 95% of the households in the sample contain between 2.16 and 2.44 persons.

(f) Is it likely that the distribution of the household sizes in the sample follow a normal distribution?

(g) (not covered under midterm) Is it likely that the sample mean will follow a normal distribution? Why?