**Part A** Insert your answers below and save as a Word document. Show your work. Save the STATA output as an rtf file to be able to include in this document.

1. (1) Calculate the mean and standard deviation for the data below.

(2) Construct a stem and leaf plot and comment on the shape of the plot.

10 50 35 20 11 10 10 20 40 15 20 35 10 22 10 40 10 35 25 30 29 15

2. The Excel file RTOG2trials contains patient characteristics of two RTOG randomized trials (study numbers 9003 and 9501) on testing new therapies for advanced head and neck cancer patients. Bring the data into STATA and perform the following analyses:

1. For the variables listed below specify if categorical, discrete, or continuous
2. Generate summary statistics (mean, std, min, max) on the variable *age for all patients*
3. Generate summary statistics (mean, std, min, max) on the variable *age by study*
4. Find the 25th 50th and 75th percentiles for age (i.e. find Q1, Q2, Q3) and calculate IQR
5. Which study has the highest percentage of Native Americans? How about Hispanics? (Hint: create a frequency table for race by study number)
6. Generate a Boxplot of age by study (represented in one figure)
7. Generate a histogram of age for all patients, and then histograms of age by study
8. Describe the age distributions in these two studies based on the summary statistics and graphics

***Variable in database Description Codes***

|  |  |  |
| --- | --- | --- |
| Sn | RTOG study number (=9003 or 9501) |  |
| Cn | Case number |  |
| Age | Age, in years |  |
| gender | Gender | 1= male |
|  |  | 2= female |
| Race | Race | 1= white |
|  |  | 2= hispanic |
|  |  | 3= black |
|  |  | 4= oriental |
|  |  | 6= native American |
|  |  | 7= other |

**Reading Assignment**

• Chapters 6.1, 6.2, 7.1, 7.2 and 7.4 in the textbook

**Part B:**

Complete the following problems using the methods we learned in class.

1. A tetrapeptide (a compound consisting of 4 amino acids linked in a chain) has the following amino acid composition: alanine (A), glutamic acid (G), lysine (L), and histidine (H). For example, ALGH and LGHA are two possible chains that can be constructed from these amino acids.

1. Draw a tree diagram to represent the 24 ways in which these amino acids can form a tetrapeptide.
2. If each chain is equally likely, what is the probability of finding A at either end of the chain?
3. What is the probability that lysine is not found at either end of the chain?

2. In guinea pigs, long hair (L) is dominant to short hair (l), and black fur (B) is dominant to albino fur (b). Answer the following questions.

(a) What are the possible genotypes of the offspring of a black haired parent and an albino, and the probability of each genotype?

(b) If two longhaired parents are mated, what is the probability of a short-haired offspring?

3. Toxicity monitoring is very important for patients receiving cancer treatment and mandatory in clinical trials. For example, breast cancer patients receiving HER2 targeted therapy may have a 10% probability to have cardiac toxicity. Such toxicity is closely monitored. A study will enroll 30 patients with such conditions.

1. Why is this a suitable problem to apply the binomial distribution theory?
2. What is the probability of observing one patient with a cardiac event among the 30 patients?
3. What is the probability that at least one patients having cardiac events among the 30?
4. What is the probability that at least two patients having cardiac events among the 30?

4. Pagano, page 193 #12:

According to the Behavioral Risk Factor Surveillance System, 58% of all Americans adhere to a sedentary lifestyle.

1. If you selected repeated samples of size twelve from the U.S. population, what would be the mean number of individuals per sample who do not exercise regularly? What would be the standard deviation?
2. Suppose that you select a sample of twelve individuals and find that ten of them do not exercise regularly. Assuming that the Surveillance System is correct, what is the probability that you would have obtained results as bad as or worse than those you observed?
3. Also explain why the binomial distribution is suitable here.

5. Pagano, page 194 #18:

Among females in the United States between 18 and 74 years of age, diastolic blood pressure is normally distributed with mean µ = 77 mm Hg and standard deviation σ = 11.6 mm Hg.

1. What is the probability that a randomly selected woman has a diastolic blood pressure less than 60 mm Hg?
2. What is the probability that she has a diastolic blood pressure greater than 90 mm Hg?
3. What is the probability that she has a diastolic blood pressure between 60 and 90 mm Hg?
4. Find the upper and lower bounds of diastolic blood pressure between which 90% of the women find their diastolic blood pressure values.