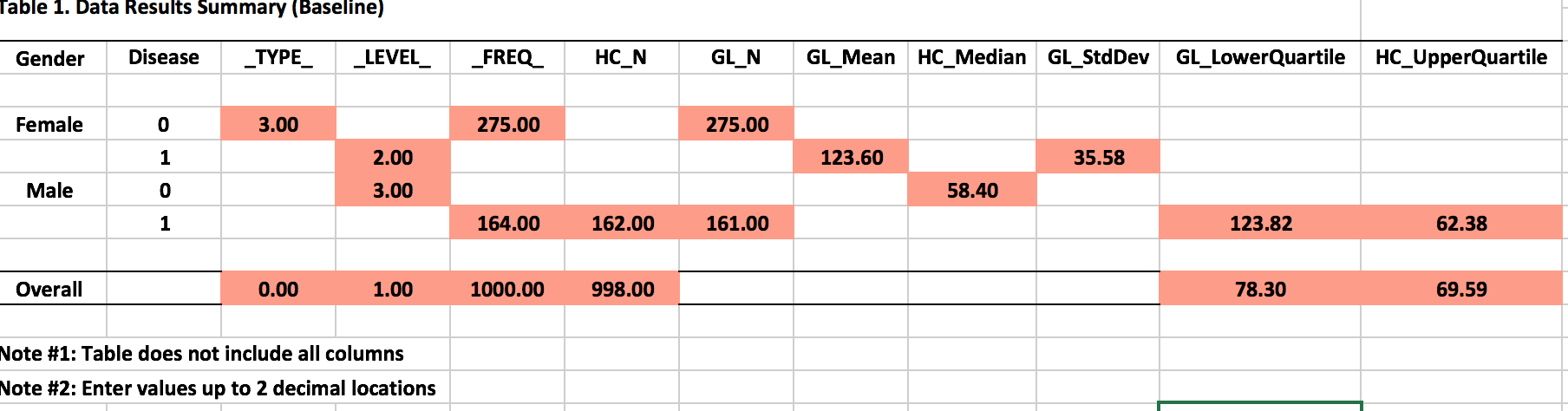
**Assignment 04**

Please answer the following questions to the best of your ability. Please show all your work and read the questions carefully.

You are the lead analyst for a clinical study that utilizes data file “data.csv”. Some of your reports are incomplete and you have yet to analyze some data. Each patient had a baseline row and a follow-up row performed six months later. The dataset contains multiple columns where

* PatientId: Unique identifier for the patient.
* TimePoint: 0 for the baseline, 1 for the follow-up
* Disease: 1 for yes and 0 for no

1. Using the provided patient data, fill in the missing cells (marked red) in Table 1 of “report.xlsx” file. Table 1 consist of generated SAS summary statistics (mean, median, standard deviation, first quartile, third quartile) of the baseline demographics categorized by gender and disease.



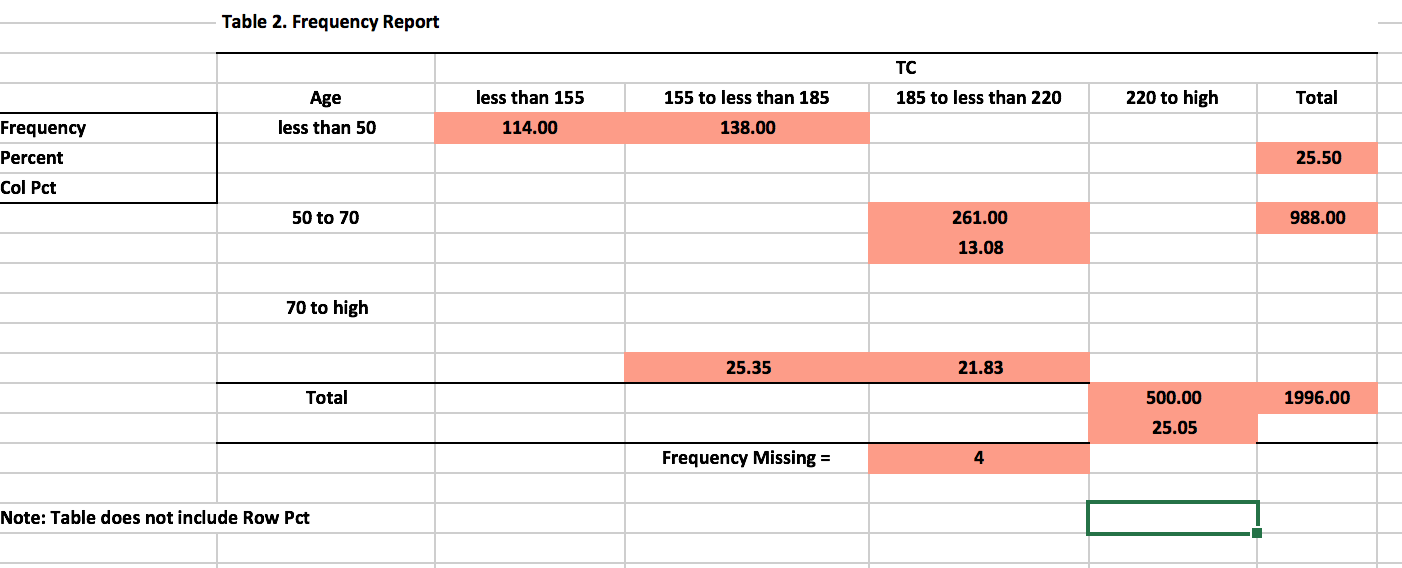
1. **a.** Identify the patient id and TR value of the patient with the 5th highest TR value. Limit patients to those with disease, follow-up time-point, and within the age range of 80 to 90 (inclusive).



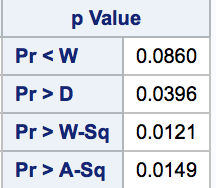
**b.** Identify the patient id and CP value of the patient with the 4th lowest CP value. Limit patients to those with disease, follow-up time-point, and within the age range of less than 50.

****

1. Fill in the missing cells (marked red) in Table 2. Frequency Report of “report.xlsx” file. Table 2 is a cross tabular frequency report of Age by TC values. Continuous data Age is categorized into three groups while continuous data TC is categorized into four categories.

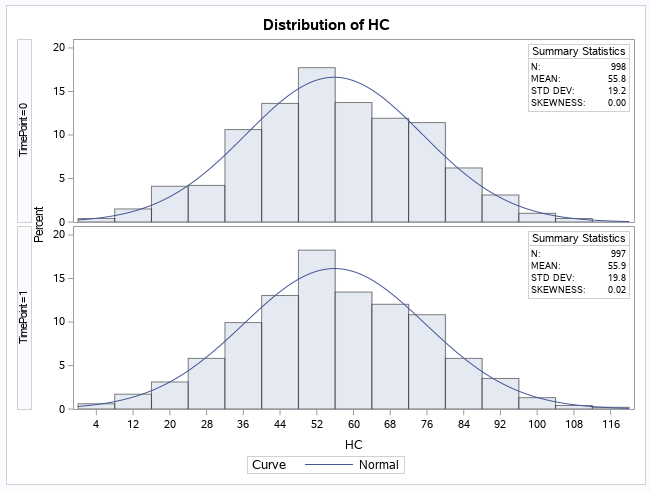


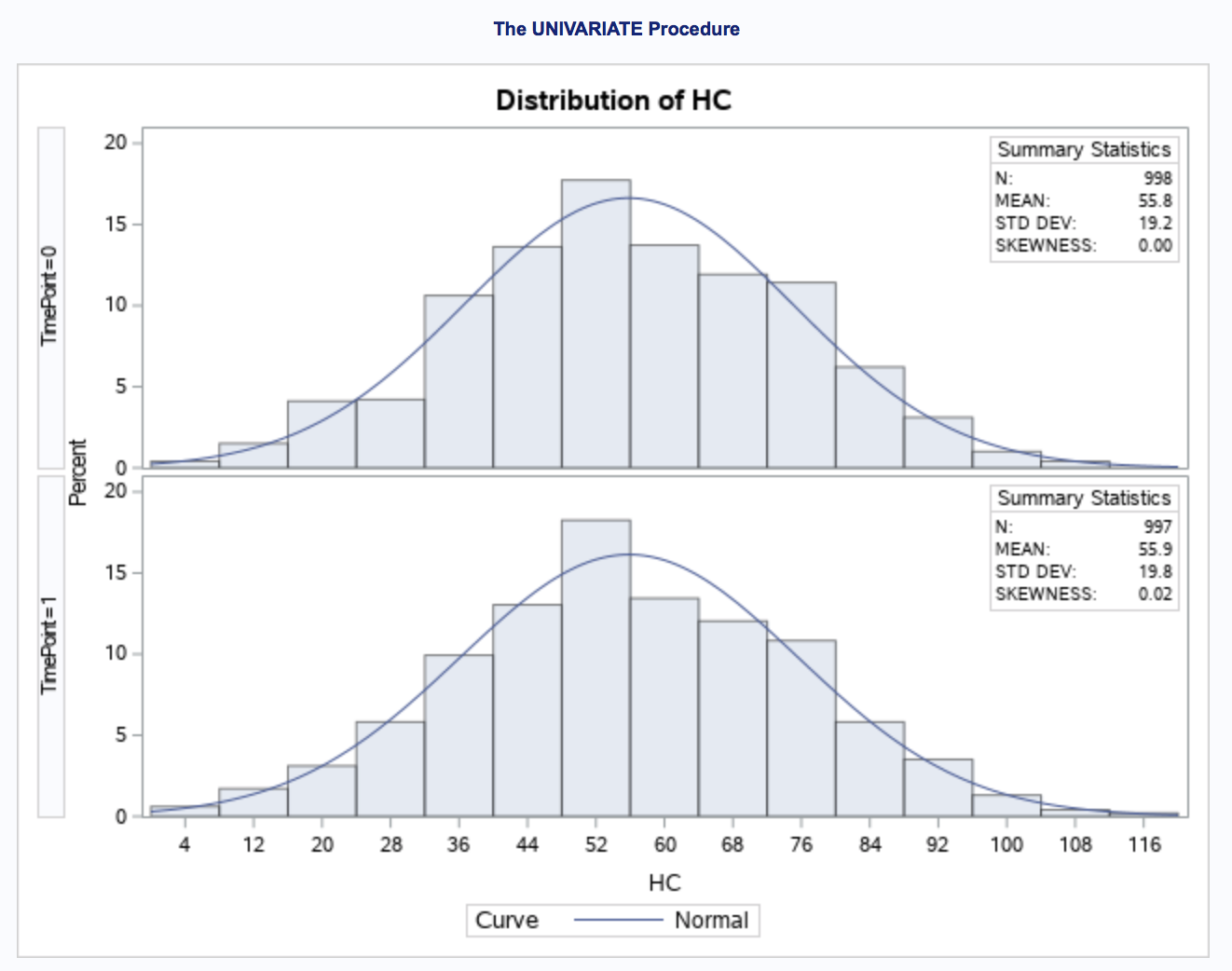
1. **a.** Analyze the provided patient data and state whether the continuous data (TC, LC, HC, TR, GL, CP) are normally distributed at (.



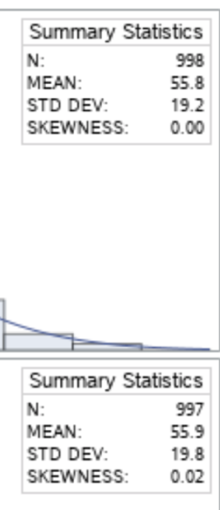
**Normally distributed because the p value > 0.05**

**b.** Replicate the histograms below of the continuous data variable HC for each timepoint. (Hint. Include inset in each histogram titled ‘Summary Statistics’ that contains n as ‘N:’, mean as ‘Mean:’, standard deviation as ‘Std Dev:’, and skewness as ‘Skewness:’ displayed in the northeast corner of the plot. Make sure to display up to 2 decimal locations for mean, standard deviation, and skewness (i.e. 1.23)).





**c.** According to the skewness for the two histograms above, are the data symmetric or skewed? If so, is it left or right skewed?



**Distribution of HC is symmetric where TimePoint = 0 and is right skewed where TimePoint = 1**

**d.** Which has higher variation: HC baseline or TR baseline?

