ACCP Critical Care PRN: Guide for Descriptive Statistics

PURPOSE

To provide a brief introduction to pharmacy residents on research data set distribution, variability, and selection of central tendency descriptors.

DISCLAIMER

This document is to be used as an introductory guide and is not intended to replace a trained professional, statistician, or experienced researcher and is not intended to be the sole resource for individualized resident research projects.

ACKNOWLEDGEMENTS

This guide was created by members of the Critical Care PRN Research Committee with advanced education in statistical analysis in 2018. We would like to acknowledge the following members for their assistance in creating this resource: Melissa Thompson Bastin, PharmD, BCPS; Brittany Bissell, PharmD, BCCCP; Benjamin Hohlfelder, PharmD, BCPS; Joshua DeMott, PharmD, MSc, BCPS, BCCCP; Kendall Gross, PharmD, BCPS, BCCCP; Zachary Smith, PharmD, BCPS, BCCCP; Adrian Wong, PharmD, MPH, BCPS, BCCCP (chair).

DESCRIPTIVE STATISTICS[1]

Descriptive statistics include measures of central tendency and variability. Measures of central tendency include mean, median, and mode. These descriptions of central tendency can be appropriately applied based on the measurement scale being utilized and distribution of data (Table 1).

- Mean: Arithmetic average of data and is affected by outliers, which are extreme values of a data set. This is not true of other measures of central tendency.
- Median: Data point above or below which half of the data points lie. Alternatively, the median is the 50th percentile value of a distribution. The median is unaffected by outliers and may be more useful than the mean to describe data when outliers exist or when continuous data are not normally distributed (i.e., non-parametric).
- Mode: Most commonly obtained value or values on a data scale, or the highest point of a peak on a frequency distribution. The mode is most useful when two clusters of data exist (bimodal distribution).

Table 1. Applicability of Central Tendency Measures Based on Measurement Scale and Data Distribution

Type of Scale	Mean	Median	Mode
Interval	Yes	Yes	Yes
Ratio	Yes	Yes	Yes
Nominal	No	No	Yes
Ordinal	No	Yes	Yes
Affected by non-parametric data	Yes	No	No

MEASURES OF DATA SET VARIABILITY[1]

Data set variability is the spread or distribution of data points along the measurement scale. The most common methods used to describe data set variability include range, interquartile range, standard deviation, and standard error of the mean. These descriptions of data set variability can be appropriately applied based on the measurement scale being utilized and distribution of data (Table 2).

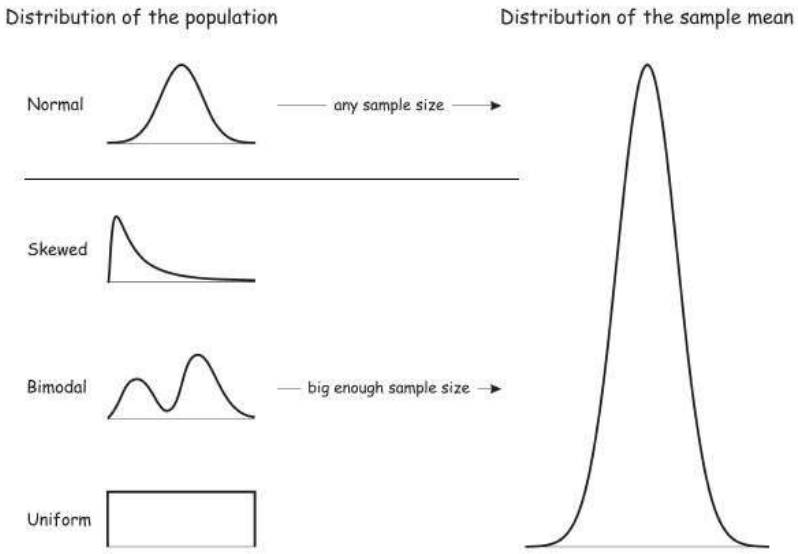
- Range: Difference between the lowest and highest values within a data group.
- Interquartile range (IQR): Measure of variability directly related to the median. The IQR describes the interval between the 25th and 75th percentile values of the data set.
- Standard deviation (SD): Provides an estimate of the degree of scatter of individual sample data points about the sample mean. The usefulness of the SD lies in its properties as related to or normal distribution.
- Standard error of the mean (SEM): Statistic derived from the standard deviation. The SEM should be used to estimate the reliability of a sample, as it relates to the population from which the sample was drawn. The SEM does not provide an estimate of the scatter of sample data and should not be used as such. The SEM is useful because it can be used to calculate a confidence intervals.

Table 2. Applicability of Measures of Variability Based on Measurement Scale

Type of Scale	Range	Interquartile Range	Standard Deviation	Standard Error of the Mean
Interval	Yes	Yes	Yes	Yes
Ratio	Yes	Yes	Yes	Yes
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No

The central limit theorem states that regardless of the true population distribution, as the study sample size increases, the study sample mean tends to be normally distributed around the true population mean, and its SD shrinks as the study population increases.[2] Given this theorem, while the study mean may be identical to the true population, it does not indicate if the study data is normally distributed (Figure 1). Data set distribution is an important characteristic to determine prior to performing statistical testing.

Figure 1. Graphical Representation of the Central Limit Theorem[2]



DATA DISTRIBUTION[1]

There are two ways of classifying the distribution of a data set, normal distribution (bell-shaped curve) and non-normal distribution. Two common methods used to determine the frequency of data set distribution are graphical and quantitative strategies. The graphical methods include the histogram and Q-

Q plot. The quantitative methods include Shapiro-Wilk and Kolmogorov-Smirnov test. There are multiple types of data distribution, including normal (Gaussian or "bell-shaped curve"; see Figure 2), binomial, and skewed distribution. There can be nominal (Gaussian) or non-normal (non-Gaussian) data that creates a bell curve distribution (Figure 3), highlighting the importance of assessing the data set distribution.

The histogram is a visual method plotting numerical data on a chart with the data set range plotted on the x-axis with the y-axis describing the numerical count of each observed data point on the x-axis. The histogram visually depicts the frequency distribution (shape) of a data set. Visual representation of data in this manner allows for assessment of underlying distribution with common distribution patterns being normal, skewed, or bimodal distribution (Figures 2, 4, 5).

Figure 2. Histogram – Normal (Gaussian) Distribution[1]

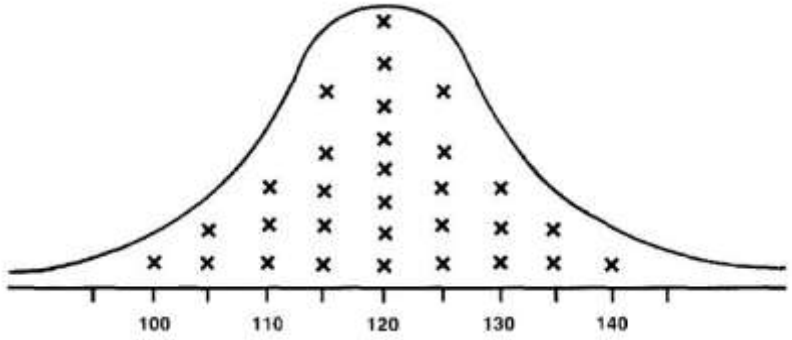


Figure 3. Histogram – Normal (A) and Non-Normal (B) Distribution[3]

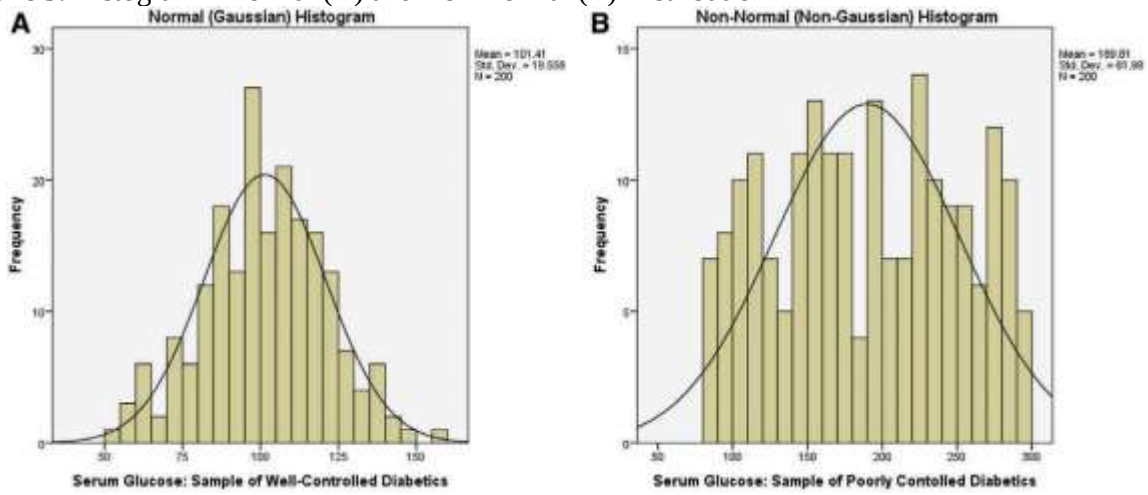


Figure 4. Histogram –Data Skewed to the Left[1]

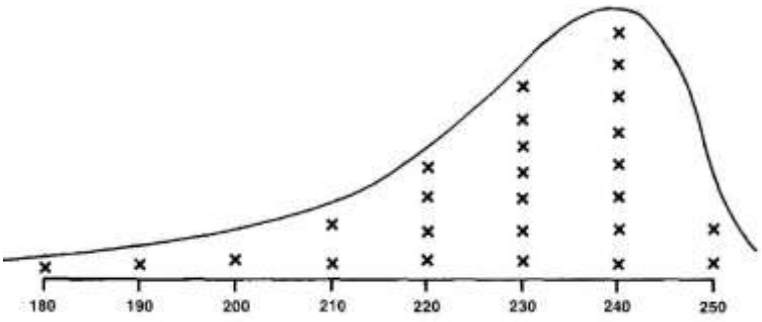
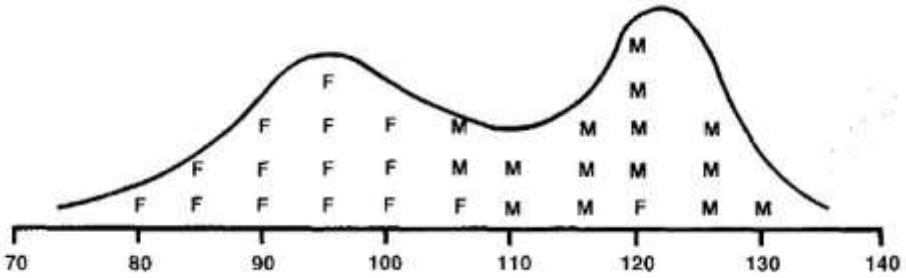
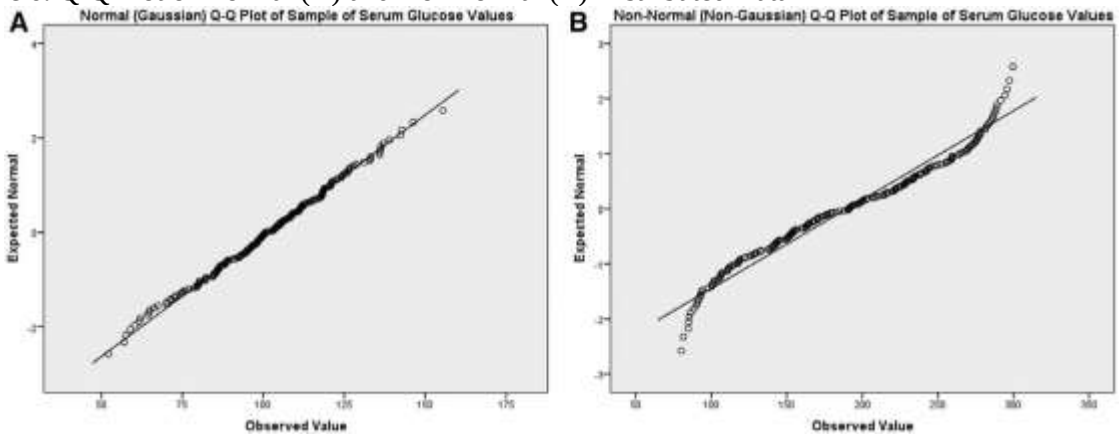


Figure 5. Histogram – Bimodal Distribution[1]



The Q-Q plot is a second graphical method to visually assess the normality of a data set. It is a scatterplot of a normal distributed data set on x-axis and the quantiles of the actual sample data set on the y-axis. If the actual data set is normally distributed it will align with the 45° reference normally distributed data set (Figure 6A). If the individual data points stray from the normally distributed data set line, the data set is not normally distributed (Figure 6B).

Figure 6. Q-Q Plot of Normal (A) and Non-normal (B) Distributed Data[3]



The Shapiro-Wilk test and Kolmogorov-Smirnov test are quantitative analysis to assess if data is normally distributed. Both tests compare the users study data set with a normally distributed data set that has the same mean and SD. The null hypothesis of these tests is that user's study data set is normally distributed. This means if the p-value is < 0.05, then the study data set is not normally distributed. The Shapiro-Wilk test is more appropriate for data sets with less than 50 samples (n), but it can be applied to larger data sets.

REFERENCES

1. Gaddis GM, Gaddis ML. Introduction to biostatistics: part 2, descriptive statistics. Ann Emerg Med. 1990;19(3):309-15.
2. Curran-Everett D. Explorations in statistics: the assumption of normality. Adv Physiol Educ. 2017;41(3):449-53.
3. Vetter TR. Fundamentals of research data and variables: the devil is in the details. Anesth Analg. 2017;125(4):1375-80.