# Clean data, and you!

Internal Validity requires intense preparation!

Cecil H. McManus, Ph.D.

11/16/22

Research Seminar Series A

PROBE Consulting Company, LLC
PO BOX 5308
Cary, NC  27512-5308
919-395-4568
information@probeconsulting.com
https://probeconsulting.com/
www.linkedin.com/in/probe-consulting-268b0139

*A day without clean data is like a day without Sunshine!*

© 2022 Cecil Harold McManus, Ph.D./PROBE Consulting Company LLC

# What is all this I am hearing about clean data?

We have a philosophy at PROBE Consulting Company, LLC. It is such a part of our history and daily operations that we placed it prominently on our website—**A Day without clean data is like a day without Sunshine!** We always f**ocus on the end at the start of a research project. The quality of any answer is directly related to the quality of how the question was asked. The design has to be clean.** Before getting into the nuts and bolts of Clean Data, let me present a metaphor that illustrates the point.

## Clean data is like painting a room!

My experience has been that an eight-hour paint job involves about two hours of painting. Most of the time on the project is dedicated to planning, preparation, and observation. A twelve-step paint process may take this course:

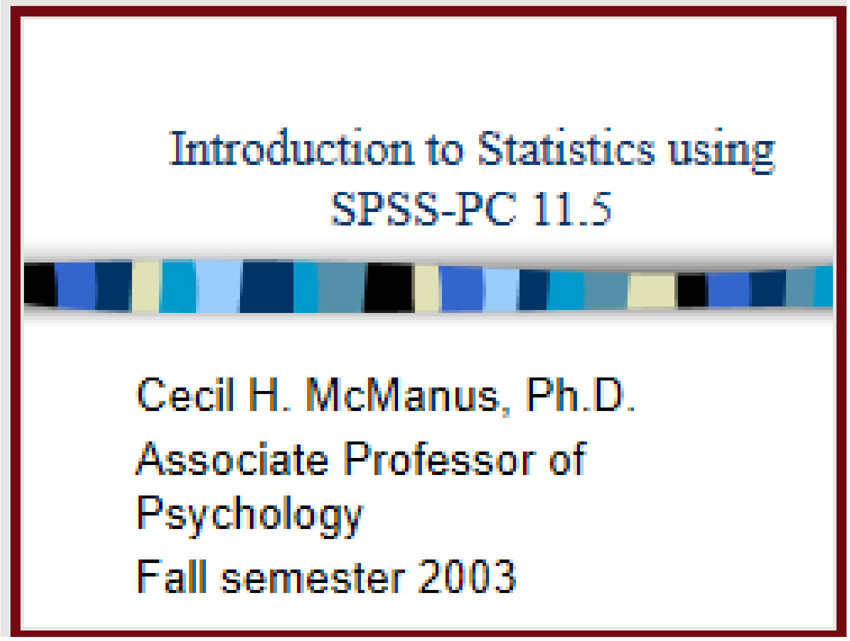
01. Selecting color or colors
02. Getting all brushes and drop clothes ready
03. Removing all face plates and window treatments
04. Applying masking tape to all borders and trim areas
05. Covering furniture and fixtures with drop cloth
06. Painting the room/area
07. Removing all masking tape and drop cloths
08. Spot-checking trouble areas
09. Packing up all drop cloths
10. Cleaning or disposing of brushes and rollers
11. Replacing all face plates and window treatments
12. Observing the drying process and touching up any problem areas

If one follows this process, the prospects for a successful project increase exponentially! If one does not, the time and expense required to repaint and replace damaged items will break the bank. In addition, the room looks very messy. It pays to prepare appropriately and measure twice, and cut once. With contemporary statistical packages, measuring thousands of times per second is possible before designing and developing an assessment system. Preparation time needs to exceed explanation time; Internal and external validity concerns demand the proper proportion of planning to prognostication.

## Clean data and Validity—no functional external exists without a clean internal!

Reliability and Validity are concepts used to evaluate the quality of research. They indicate how well a method, technique, or test measures what it is intended to measure. Reliability is about the consistency of a measure, and Validity is about the accuracy of an action (1). Let us further explore these two critical concepts. The stronger the extent to which one uses the proper research methodology and design to develop a project, the stronger the internal Validity therein. Are the participants chosen at random? Are there sufficient participants to achieve a desirable level of power? How does the distribution look---is it normal or skewed? What is the Level of Measurement inherent in the data? The Level of Measurement will determine which statistic is chosen to evaluate the hypotheses. Yes, one will need hypotheses if one aims to make an inference. Let’s walk through a few slides and lectures from the past!

# Part I:  The Basics (Skip if you don't need to review this information)



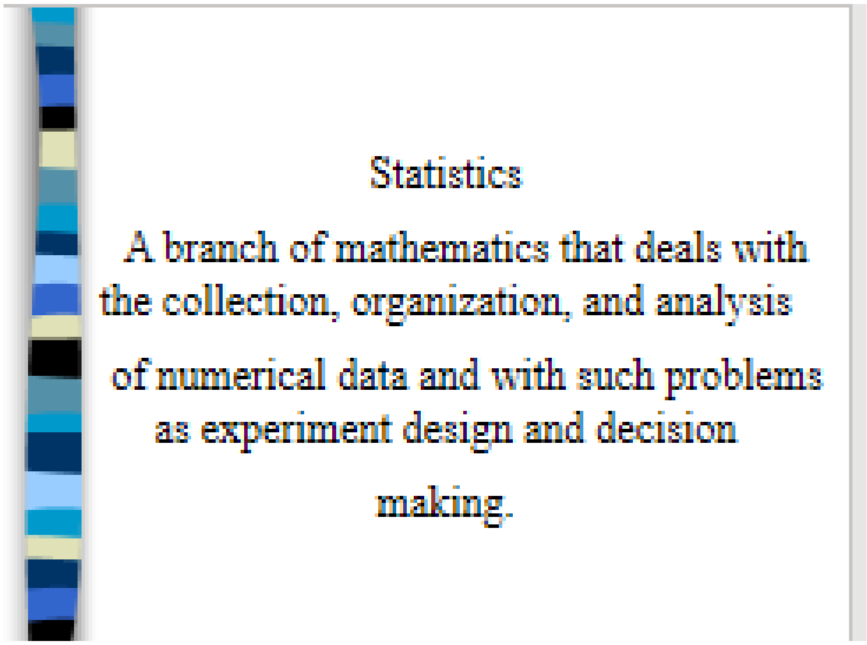
Introduction to Statistics using SPSS-PC 11.5

Cecil H. McManus, Ph.D.

Associate Professor of Psychology

Fall semester 2003

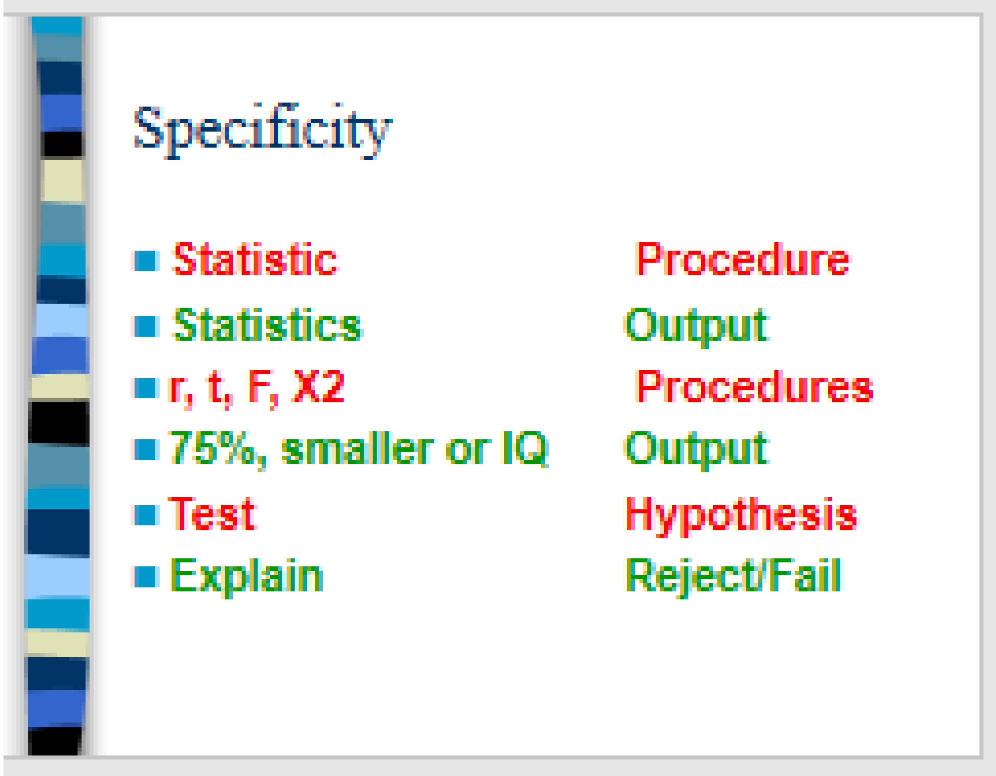
Nineteen years ago! Oh My! I was sporting a fade and wearing bell bottoms!



Statistics

A branch of mathematics that deals with the collection, organization, and analysis of numerical data and with such problems as experiment design and decision making.

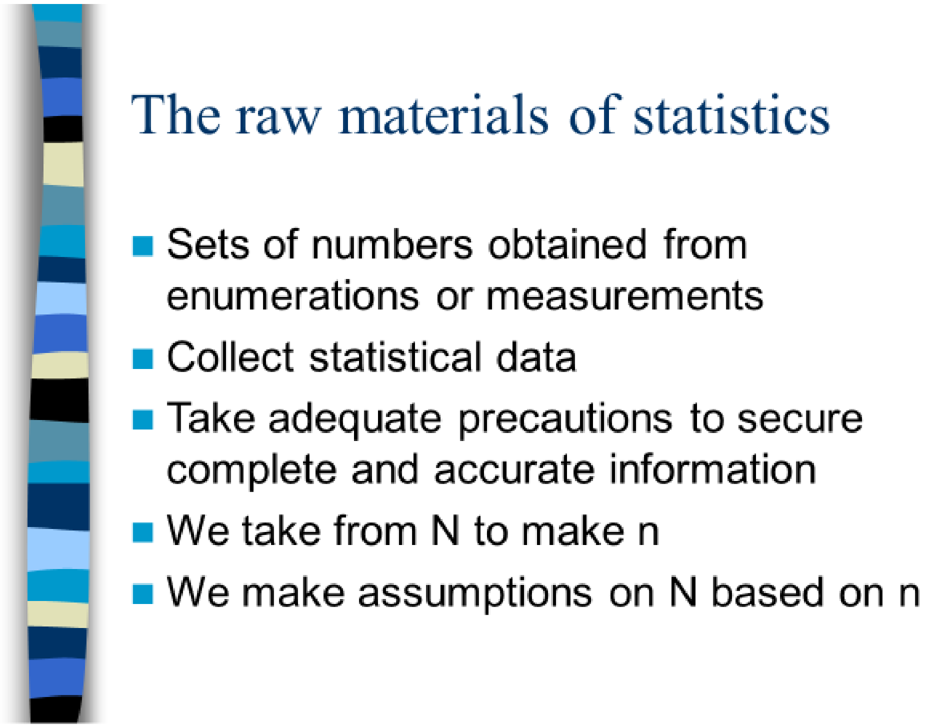
This is the foundational topic for this brief e-document



## Specificity

| | |
|---|---|
| ■ Statistic | Procedure |
| ■ Statistics | Output |
| ■ r, t, F, X2 | Procedures |
| ■ 75%, smaller or IQ | Output |
| ■ Test | Hypothesis |
| ■ Explain | Reject/Fail |

Our nomenclature refers to the procedure (t-test, ANOVA, Chi-Square, etc.) as a statistic and the summarized data/information as statistics. You may hear a sports commentator say, "let's look at the statistics" when describing how many plays were run in the first quarter or the number of sacks in the first half. They are describing data, not statistics. For data to become statistics, it must be manipulated in some manner. Go back to the commentator; if they were to say, "the percentage of running plays in the first quarter was 45%, or 90% of all sacks occurred in the 4th quarter,  they would be talking about statistics. We go a bit further by using data and statistics to ask and test questions germane to some probable outcome—we test hypotheses.



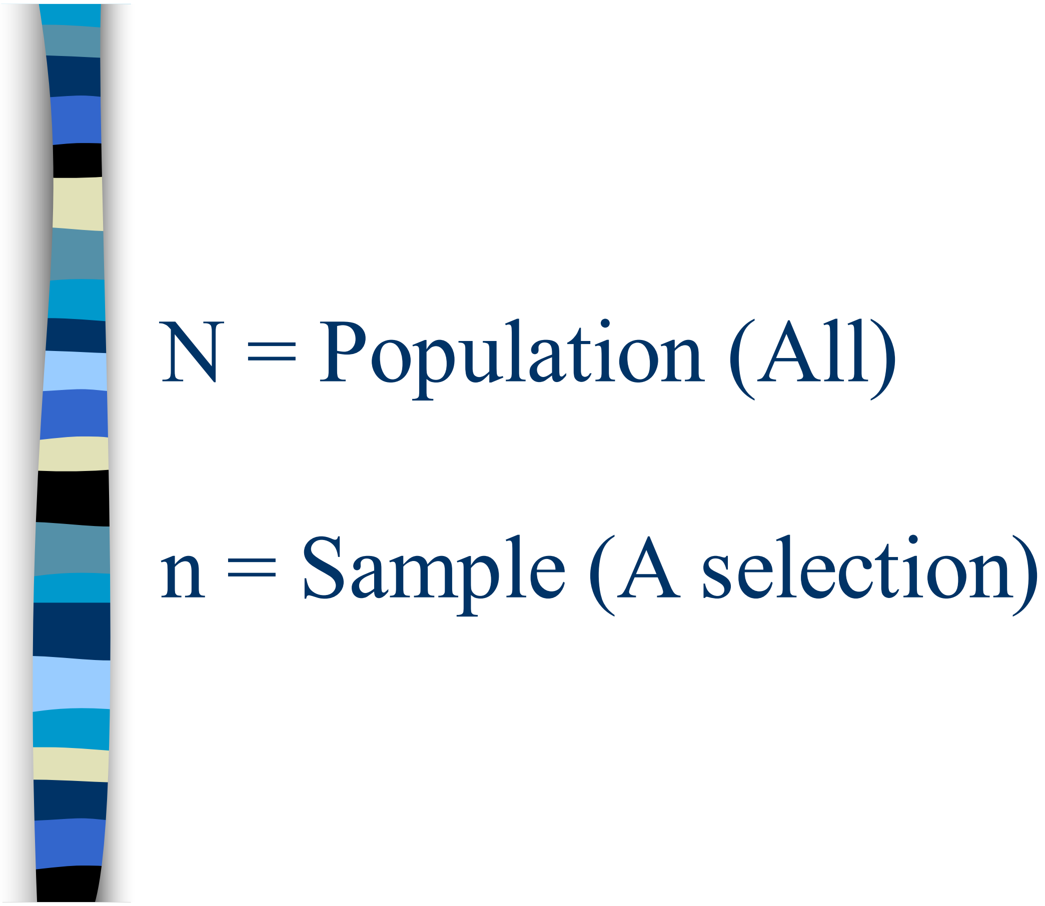
## The raw materials of statistics

- Sets of numbers obtained from enumerations or measurements
- Collect statistical data
- Take adequate precautions to secure complete and accurate information
- We take from N to make n
- We make assumptions on N based on n

I will stick with that first bullet point—we collect data. I would remove the term "statistical" from the second bullet point. I am sure I said, "we collect data in support of statistical analysis," during my engaging and dynamic lecture nineteen years ago. If not, I'm saying it now. 😊 The next bullet point starts the "CLEAN DATA" (CDAT) discussion. The "adequate precautions" terminology is the crux of our CDAT argument and centers on Internal Validity.

Internal Validity truly matters. It makes the conclusions of a causal relationship credible and trustworthy. Without high internal Validity, an experiment cannot demonstrate a causal link between two variables. To make an investigation more relevant, one must avoid confounding and extraneous variables and have built-in safeguards to ensure that the intervention consistently precedes results. The last two bullet points describe the Population (N) and the Sample (n).

This is the basic idea. One pulls a sample from the population.

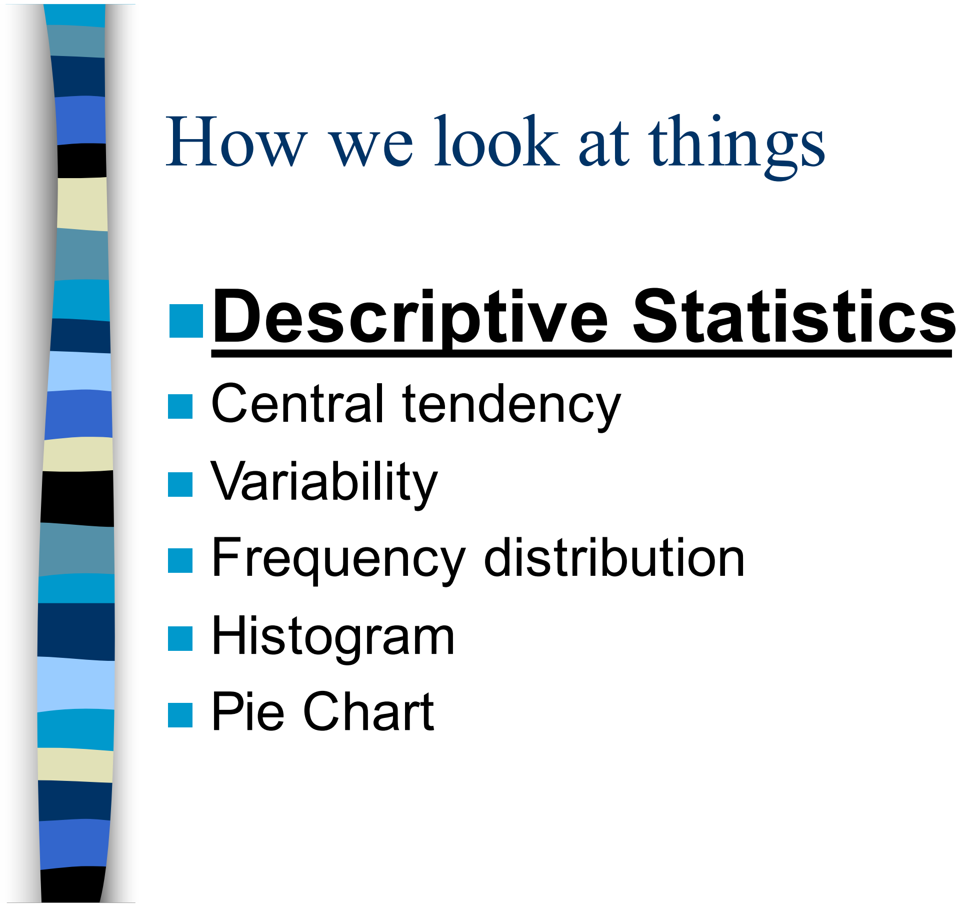


N = Population (All)

n = Sample (A selection)

Ok, you have used the proper methodology to collect data from your target population. You have chosen a sufficient number using a random process from a normal distribution (We will discuss the Central Limit Theorem in a later post); let's conclude this session with how we look at the data and then make assumptions about it.

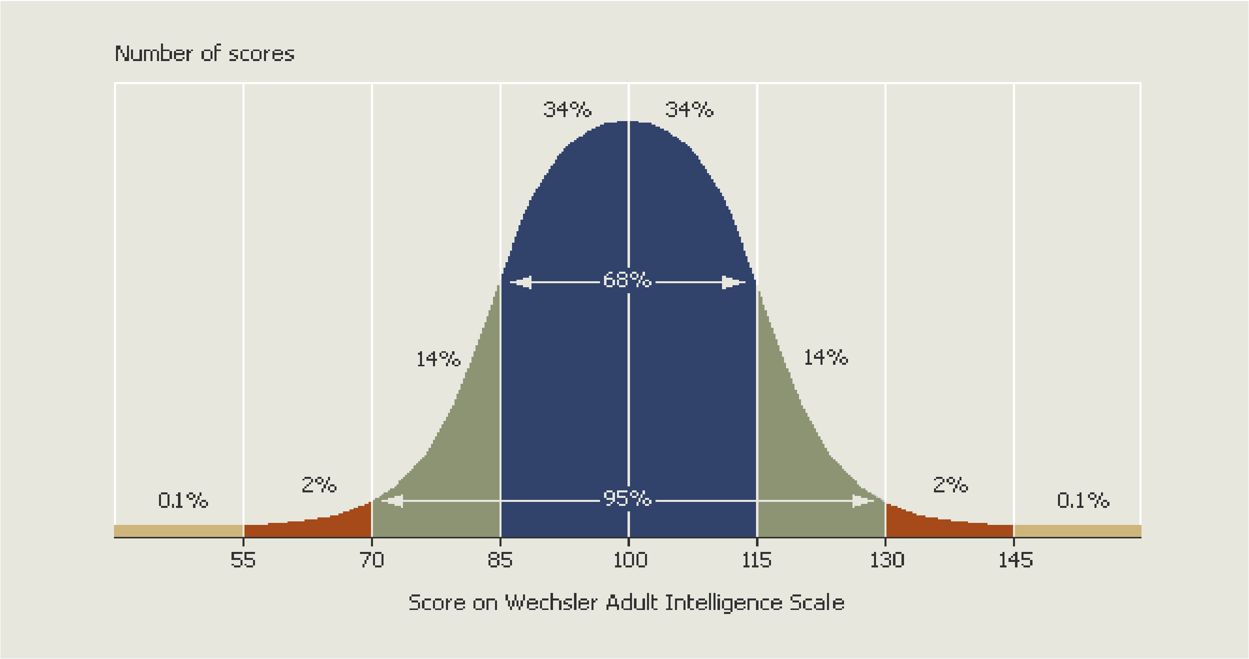
## Part II:  Looking at clean data on the normal curve!



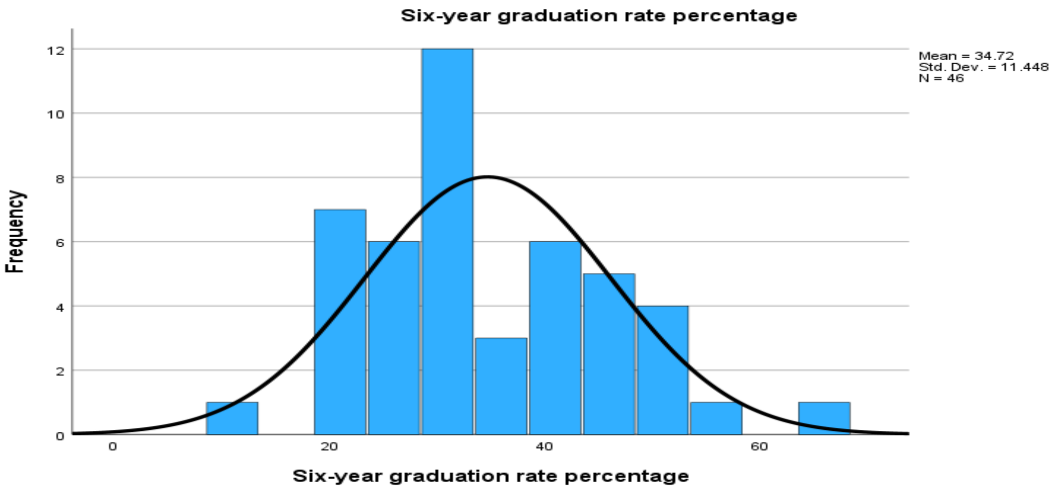
### How we look at things

- **Descriptive Statistics**
- Central tendency
- Variability
- Frequency distribution
- Histogram
- Pie Chart

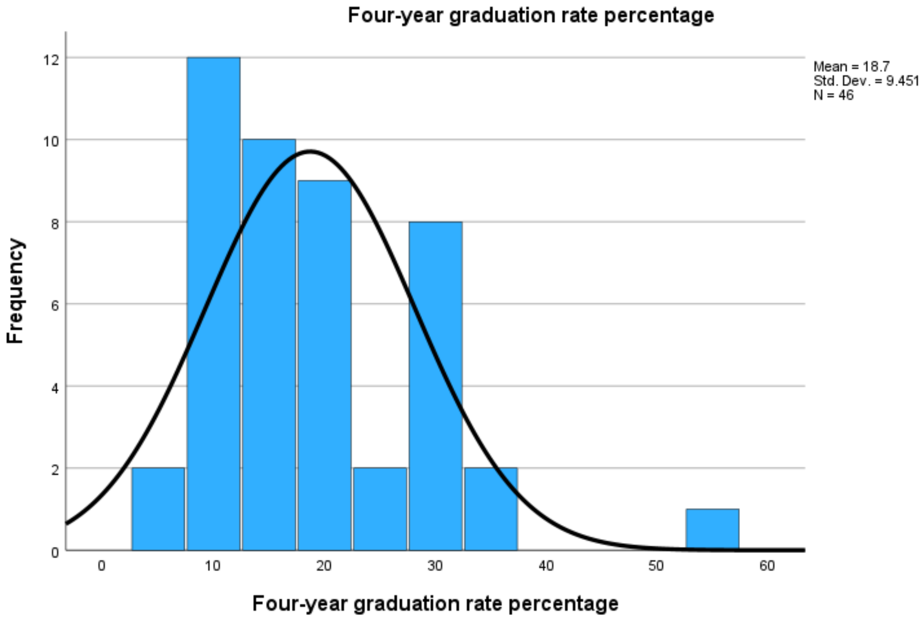
After collecting the data, the first step will be to look at it. Descriptive Statistics allows one to do just that. Central Tendency indicators (Mean, Median, and Mode) will enable one to observe how the data tends to fall toward the center of the distribution. Variability, or Measures of Dispersion, allows one to see how the data falls away from the center of the distribution.



This is a distribution curve describing IQ Scores. The Mean (Central Tendency) is 100, and the Standard Deviation is 15. THIS IS A NORMAL DISTRIBUTION.



This is a distribution curve describing Six-year graduation rates at 47 Colleges. The Mean (Central Tendency) is 34.72, and the Standard Deviation is 11.448.  THIS IS A POSITIVELY SKEWED DISTRIBUTION. WE WILL DISCUSS SKEWNESS IN A LATER DOCUMENT.



This is a distribution curve describing Four-year graduation rates at 47 Colleges. The Mean (Central Tendency) is 18.7, and the Standard Deviation is 9.451. THIS IS A POSITIVELY SKEWED DISTRIBUTION. WE WILL DISCUSS SKEWNESS IN A LATER DOCUMENT. The previous three charts used a histogram to examine the mean and standard deviation. These are essential concepts to understand. See if this makes sense to you:

IF 100 PEOPLE TAKE A TEST, AND ALL 100 SCORE 55 ON THE TEST, THE MEAN WOULD BE 55, AND THE STANDARD DEVIATION WOULD BE 0.

- The mean is the mathematical average of a set of two or more numbers.
- The Standard Deviation measures the dispersion of a dataset relative to its mean.

MEAN—SUM OF ALL SCORES DIVIDED BY THE NUMBER OF SCORES.

STANDARD DEVIATION—GET THE MEAN OF ALL SCORES. SUBTRACT EACH SCORE FROM THAT MEAN. WHAT IS THE AVERAGE OF EACH SCORE AWAY FROM THE MEAN?

**Statistics**

| | | Fall-to-fall retention rate percentage. | Six-year graduation rate percentage | Four-year graduation rate percentage | The ratio of faculty members to students |
|---|---|---|---|---|---|
| N | Valid | 47 | 46 | 46 | 47 |
| | Missing | 0 | 1 | 1 | 0 |
| Mean | | 64.43 | 34.72 | 18.70 | 14.89 |
| Median | | 65.00 | 33.00 | 17.00 | 15.00 |
| Mode | | 63[a] | 30[a] | 8[a] | 15[a] |
| **Std. Deviation** | | **12.404** | 11.448 | 9.451 | 4.007 |
| Range | | **59** | 53 | 48 | 17 |
| **Minimum** | | **32** | 11 | 5 | 8 |
| **Maximum** | | **91** | 64 | 53 | 25 |

a. Multiple modes exist. The smallest value is shown

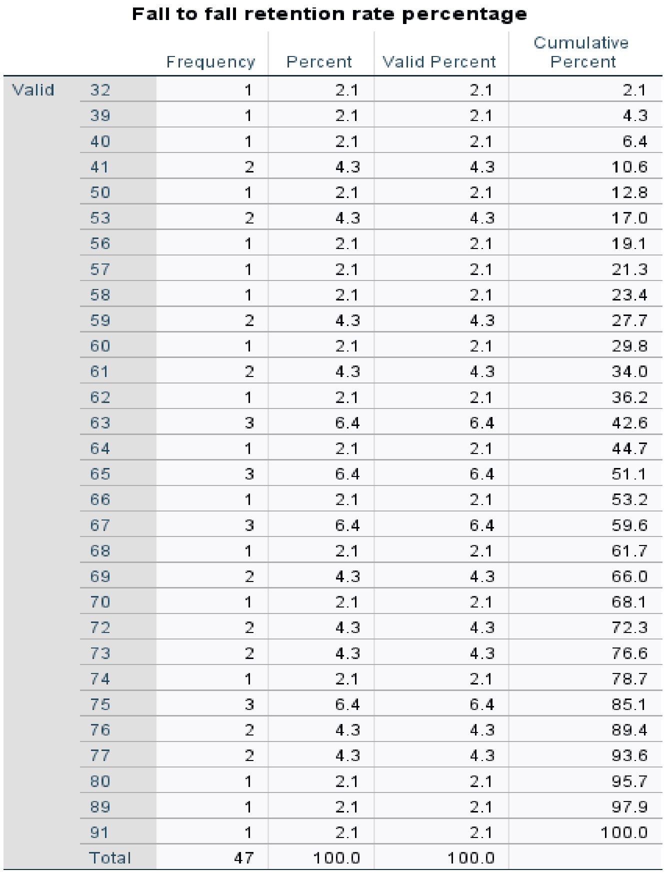
The example above is output from SPSS. The Measures of Central Tendency are displayed in the table. The Mean for the 47 Schools on Fall-to-fall retention was 63.43%, the Median was **65%,** and the Mode was **63%**. The "Typical" score for this variable was 64% for the 47 schools. The fact that the Median and Mode were very similar to the Mean suggests that there were not many extreme scores in this sample.

The Measures of Variability are also displayed in this table. The Standard Deviation for the 47 Schools on Fall-to-fall retention was 12.404%, and the Range was 59, with a low percentage of 32 and a high of 91—[91 – 32 = 59]. There are four variables here, and this table describes the central tendencies and variability germane to them.

# Part III:  Data Veracity and graphics!

■ Frequency distribution

The Frequency Distribution lets you know how many times a score occurred in your sample.



**Fall to fall retention rate percentage**

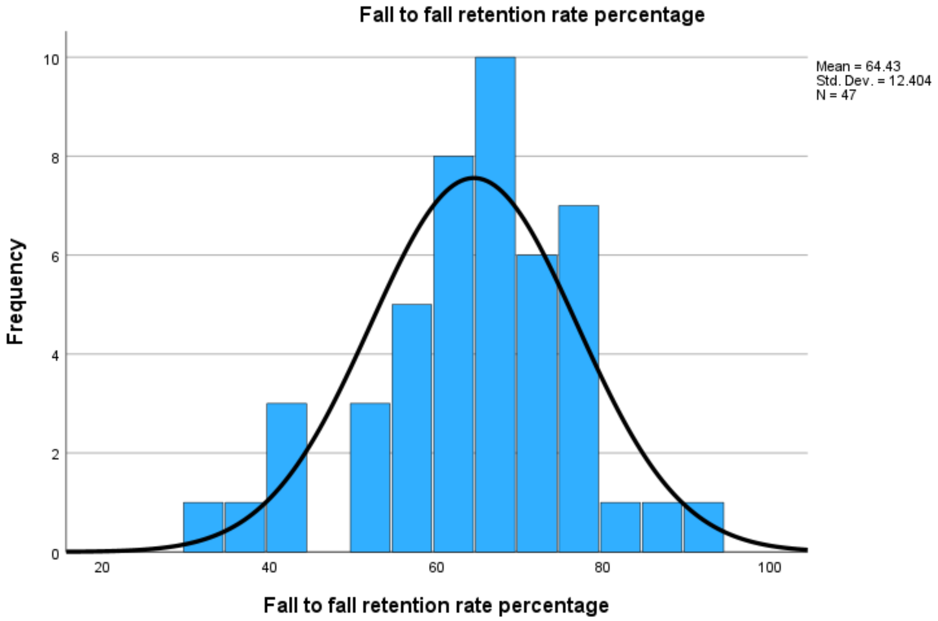
|  |  | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 32 | 1 | 2.1 | 2.1 | 2.1 |
|  | 39 | 1 | 2.1 | 2.1 | 4.3 |
|  | 40 | 1 | 2.1 | 2.1 | 6.4 |
|  | 41 | 2 | 4.3 | 4.3 | 10.6 |
|  | 50 | 1 | 2.1 | 2.1 | 12.8 |
|  | 53 | 2 | 4.3 | 4.3 | 17.0 |
|  | 56 | 1 | 2.1 | 2.1 | 19.1 |
|  | 57 | 1 | 2.1 | 2.1 | 21.3 |
|  | 58 | 1 | 2.1 | 2.1 | 23.4 |
|  | 59 | 2 | 4.3 | 4.3 | 27.7 |
|  | 60 | 1 | 2.1 | 2.1 | 29.8 |
|  | 61 | 2 | 4.3 | 4.3 | 34.0 |
|  | 62 | 1 | 2.1 | 2.1 | 36.2 |
|  | 63 | 3 | 6.4 | 6.4 | 42.6 |
|  | 64 | 1 | 2.1 | 2.1 | 44.7 |
|  | 65 | 3 | 6.4 | 6.4 | 51.1 |
|  | 66 | 1 | 2.1 | 2.1 | 53.2 |
|  | 67 | 3 | 6.4 | 6.4 | 59.6 |
|  | 68 | 1 | 2.1 | 2.1 | 61.7 |
|  | 69 | 2 | 4.3 | 4.3 | 66.0 |
|  | 70 | 1 | 2.1 | 2.1 | 68.1 |
|  | 72 | 2 | 4.3 | 4.3 | 72.3 |
|  | 73 | 2 | 4.3 | 4.3 | 76.6 |
|  | 74 | 1 | 2.1 | 2.1 | 78.7 |
|  | 75 | 3 | 6.4 | 6.4 | 85.1 |
|  | 76 | 2 | 4.3 | 4.3 | 89.4 |
|  | 77 | 2 | 4.3 | 4.3 | 93.6 |
|  | 80 | 1 | 2.1 | 2.1 | 95.7 |
|  | 89 | 1 | 2.1 | 2.1 | 97.9 |
|  | 91 | 1 | 2.1 | 2.1 | 100.0 |
|  | Total | 47 | 100.0 | 100.0 |  |

From our SPSS output, 32% occurred once for fall-to-fall retention  [first number in the table above]. The number 42% occurred two times [the fourth number in the table above]. You can also use this table to check for errors in your data. Since this variable is a percent, it should range from 0 to 100. If I saw a "125" on the table, I would know something was incorrect, and I would go back and correct that number by looking for it in my data set and then re-run my data.

# Pictures are worth a thousand words. Here are examples of images of descriptive statistics.

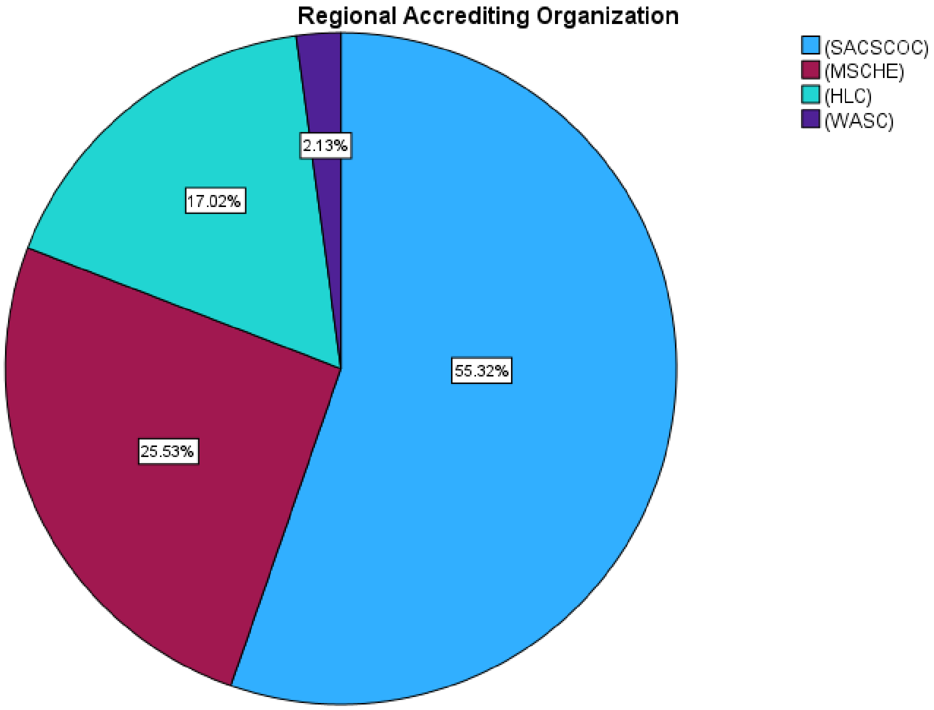
- Histogram
- Pie Chart

A histogram is one of the most commonly used graphs to show the frequency distribution. A pie chart is also ubiquitous.



From our SPSS OUTPUT, this is an example of a histogram with a curve imposed upon the data. Histograms should be used with Ratio Level data.

This is a Pie Chart example from SPSS:



Pie Charts should be used with Nominal Level data. In this case, the Names and Percentages of Regional Accrediting Agencies for the 47 schools in our samples. Over half are covered by the Southern Association of Colleges and Schools---Commission on College. What does that tell you? Most of the schools are in the Southern region of the United States.

TAKE A BASIC STAT TEST