

**Statistic of the Week**

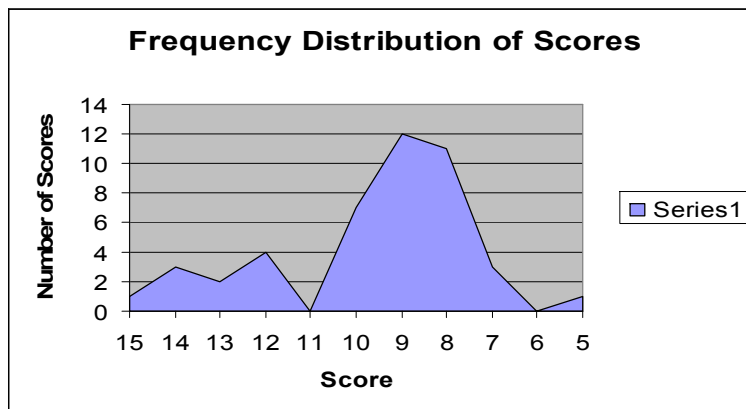
**Normal Distributions, Measures of Central Tendency and Measures of Dispersion**

**Properties of Frequency Distributions**

Suppose you were the teaching assistant in a small class of undergraduates. After the first quiz, one of the students asked "How'd I do in relation to the rest of the class?" The student's raw score isn't much help in today's competitive society, so you illustrated the student's test score among the scores of the rest of the class by making a frequency distribution table (like from last week):

Scores	Frequency of Scores	You Say "this many of these"
15	1-----	"one fifteen
14	3	three fourteens
13	2	two thirteens
12	4	four twelves
11	0	no elevens
10	7	seven tens ->your student got one of the tens
09	12	twelve nines
08	11	eleven eights
07	3	three sevens
06	0	no sixes
05	1	one five

Graphically you could draw the distribution of scores like this:



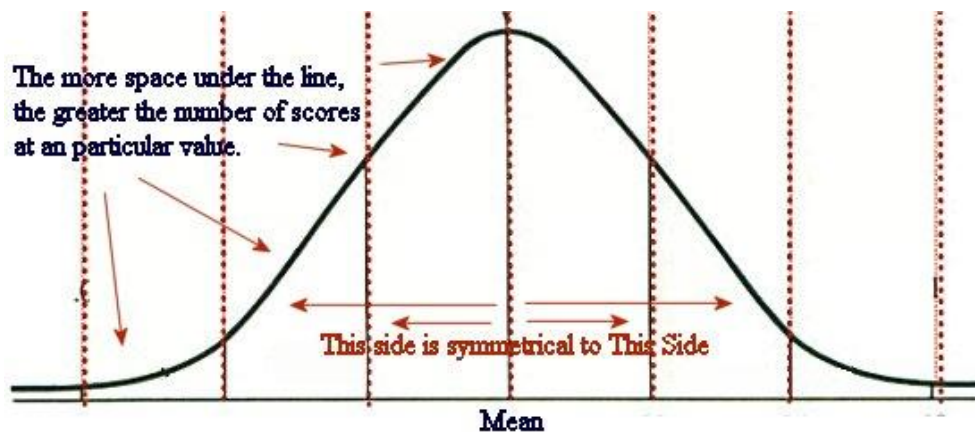
The value of frequency distributions lies in their graphic representation properties. Anyone can look at a frequency distribution and see with their eyes how scores on any measure are stacking up. Even more important is the usefulness of frequency distributions to the centerpiece of all statistics.

## That Centerpiece would be ... **The Central Limit Theorem**

The Central Limit Theorem is a very big deal in statistical world, and it goes a little something like this: Statisticians believe (i.e., pretend or assume) that most measures, when taken from large enough samples (say an entire population), will deliver a frequency distribution known as a NORMAL DISTRIBUTION. When a measure is NORMALLY DISTRIBUTED it follows the **normal probability distribution** and it exhibits certain immutable properties.

The Central Limit Theorem (CLT) says that as sample size is increased, average scores **cluster around the true mean of the population** (keep in mind that the sample is a representation of the population).

Further the distribution of the sample averages also increasingly resembles the normal probability distribution (a.k.a. the normal curve, the bell shaped curve). Here's one now:



When we randomly sample from any population, we expect the responses we get to be normally distributed. Had we given this same quiz to 100 classes of approximately the same size as the one above, or given the quiz to a class 100 times as big as this one, we would probably get a distribution that looks more normal and bell shaped, like the one I gave you in the lesson on **elementary things**.

You will recall that there are three slightly different ways to think about measures of central tendency:

- Means are average scores of a given sample of measures on a variable.
- Modes are the scores that occur most often within a given sample of measures on a variable
- Medians are the midpoint, or middlemost score, along a range of scores of a given sample of measures on a variable.

You can see how a mean can be different from a mode, and a mode different from a median, and a median different from a mean, and so on, if a variable is not really, truly normally distributed.

The key phrase in the definition of Central Limit is **average scores cluster around the true mean of the population**. When data are normally distributed, the MEAN, MODE AND MEDIAN ARE ALL THE SAME SCORE and they all fall at the exact midpoint of the distribution.

Of course, measures taken in FCS research are hardly ever really normally distributed, but we have to pretend they are so we can calculate inferential statistics. This is a temporary article of faith.

**Calculating the Mean in two easy steps:** When presented with a set of scores (sometimes called a dataset), simply 1) add all the scores up and 2) divide by the number of scores. This works well when the number of responses (NSize) is small (anyone can add up a column of 10 scores). But what if you have 2000 scores, and many of the responses are repeated (i.e., there is more than one response with the same value/score)?

In that case, you'd need to do what statisticians do in four more tedious steps:

1. Present the data in a distribution table with columns labeled X, f, and f(X), like the table below.
2. Multiply each score by it's frequency - f(X)
3. Sum the f(X)s ----- $\sum f(X)$
4. Divide by the NSIZE ----- $\sum f(X)/N$

Making a habit of thinking about data in terms of distribution tables is a useful technique. The alternative is to be **lost in a raging sea of inexactitude**. There's enough of that going around already.

Along the left column there are the scores (these listed in reverse order of magnitude, but they don't have to be). Individual scores will always be called X or Y in capital letters.

The next column is the Frequency and this will always be called f in lower case letters.

**Watch out!!!!!!!!!!** The third column is f(X) is simply the score times the frequency or X times f (The fourth column is blank for now. There will be many more columns before we're done with the class). Just look at the table for a while – See it, Know it, Be it!

<b>X</b>	<b>f</b>	<b>f(X)</b>	
<b>Scores</b>	<b>Freq</b>	<b>Score x Freq</b>	
15	01	15*01=15	
14	03	14*03=42	
13	02	13*02=26	
12	04	12*04=24	
11	00	11*00=00	
10	07	10*07=70	
09	12	09*12=108	
08	11	08*11=88	
07	03	07*03=21	
06	00	06*00=00	
05	01	05*01=05	
	NSIZE=44	$\sum f(X)=423$	$\sum (f(X)/N = 423/44 = 9.614$

To calculate the average, just add up the column of f(X)'s to get 423 then divide by the number of X's (divide by 44) and you get 9.614 The Mean Score is 9.614 (which is a score no one in the sample really got – this is how the average number of people in a family can be 3.9 when there's no such thing as a .9 person – except maybe for politicians).

We call averages MEANS in statistics and they are symbolized by the term  $\bar{x}$  or  $\bar{x}$ . The reason why Means are measures of central tendency should be crystal clear to you by now. They are in the middle of the frequency distribution.

There are two other measures of Central Tendency

- - the midpoint of the scores or MEDIAN. To calculate the Median - just count to the middle of the range of scores 5 6 7 8 9 **10** 11 12 13 14 15 Counting from the left, The sixth out of eleven scores (a 10) is the Median.
- - the score that occurs most often among all the scores or MODE. To calculate the Mode (the score that occurs most often) just look in the frequency distribution table for the score that has the biggest frequency - in this case it was a 9.

The second half of the discussion of measures of central tendency have to do with **Measures of Dispersion which describe the way:**

- scores are distributed across the values of a variable and
- scores spread away from the middle of the curve.

The two important Measures of Dispersion are: **Variance** and **Standard Deviation**. Calculating Variance and Standard Deviation is where we'll use those extra columns in the frequency distribution table.

If you were to calculate a mean, then take each score and subtract the mean from it, you'd have a column of deviations from the mean. **Variance is the average squared deviations from the mean.** Deviations from the mean are squared because some of them (half of them) will be negative and half positive values. If we add the deviations without squaring them, we'd get zero (0) for a sum, which is a good way to check for normal distribution.

- $\sigma^2$  - the lower case greek letter sigma squared is the symbol used for Variance.
- $\sigma$  is the square root of the variance.

To arrive at the variance and the standard deviation, we have to calculate the square root of the equation  $\sum (X - \bar{x})^2 / N$  which is computed using the following steps from the Distribution Table:

1. Construct a Frequency Distribution Table (been there)
  2. Calculate the Mean (done that)
  3. Calculate score deviations from the mean (that 4<sup>th</sup> column)
  4. Square the deviations (a fifth column)
  5. Sum the squared deviations (at the bottom of the 5<sup>th</sup> column)
  6. Divide by the NSIZE (number of cases or observations at the bottom of 5<sup>th</sup> column).
- To get the Standard Deviation, simply take the square root of the Variance and there you have the standard deviation.

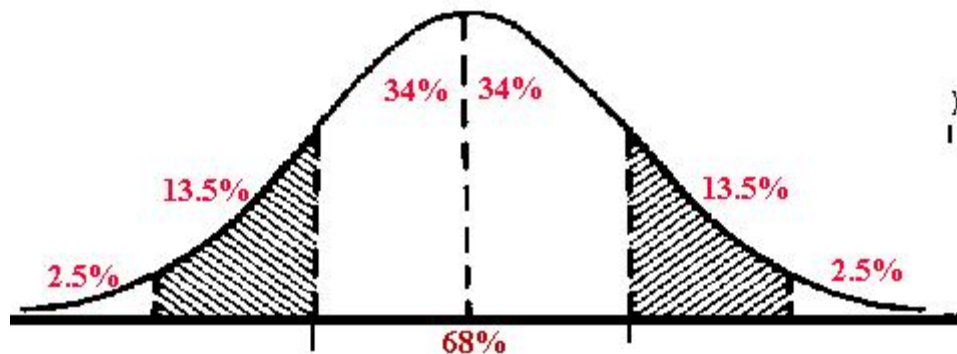
Again, using the test score data:

X Scores	f Freq.	f(X)	Deviation $X-\bar{x}$	$(X-\bar{x})^2$
15	1	15	15-9.614=5.386	29.009*1 = 29.009
14	3	42	14-9.614=4.386	19.237*3 = 88.760
13	2	23	13-9.614=3.386	11.465*2 = 22.930
12	4	48	12-9.614=2.386	5.693*4= 22.772
11	0	0	11-9.614=1.386	1.912*0= 0
10	7	70	10-9.614= 0.386	0.149*7= 1.043
09	12	108	09-9.614= -0.614	0.377*12= 4.524
08	11	88	08-9.614= -1.614	2.605*11= 28.655
07	3	21	07-9.614= -2.614	6.833*3= 20.499
06	0	0	06-9.614= -3.614	13.061*0= 0
05	1	5	05-9.614= -4.614	21.289*1= 21.289
	NSIZE=44	$\Sigma(X)=423$	$\Sigma(X-\bar{x}) = 0$	$\Sigma(X-\bar{x})^2= 240.470$
		$\Sigma(X)/N=423/44$		$\sigma^2=\Sigma(X-\bar{x})^2/44=240.470/44$
		Mean= 9.614		$\sigma^2 =5.465$
				The square root of $\sigma^2$ is $\sigma$
				$\sigma= 2.34$

- If Variance is small, it means that each of the deviations are small and the scores are narrowly dispersed across the range of scores.
- If Variance is large, it means that the distribution of scores is widely dispersed across the range of scores.

**According to the Central Limit Theorem, for any measure, for any sample of subjects, the standard deviation and the mean will distribute the scores normally so that the greatest frequency of scores will fall on or near the Mean Score.**

Standard Deviation units away from the mean will account for precise amounts of score under the normal curve - these are those certain temporary articles of faith mentioned at the top of the discussion :



Here those immutable properties / temporary articles of faith on a normally distributed variable:

- **+1 and -1 Std. Deviation around the mean accounts for 68% of the scores in the measure**
- **+2 and -2 Std. Deviations around the mean accounts for 95% of the scores**
- **What's left is 5% or .05 of the scores, this will become important as a point for significance.**

You all have heard the phrase “This is significant at the .05 level!” Well, the .05 referred to is what is left after two standard deviations on either side of the mean.

These standard intervals will allow statisticians to make inferences and predict the outcome of dependent variables based on what they’ve observed about independent variables. They rely on these theoretical assumptions to help in prediction again and again and again.

### **Two more small things before you go to do your homework.**

Frequency distributions of measures taken from sample data will vary depending on the size of the sample and the natural skewness of the variable (measure).

**Skewness** - The shape of this distribution is skewed slightly toward the lower end of score values. Distributions may be skewed right or left.

**Kurtosis** - They may also suffer from the dreaded numerical disease - Kurtosis. This is when too many scores (or not enough scores) fall into the middle or the ends of the curve.

See homework next page:

**Homework**

Name \_\_\_\_\_

**Statistics - Central Tendency**

Example #1 - Suppose you are interested in the *Desirability Ratings* for putting a nuclear waste disposal facility in a nearby town. The question you asked respondents was:

“Taken altogether - the jobs being created, the possibility of contamination, the increased chance of cancer among you and your neighbors, and the increase in electrical power you will receive - How do you feel about having the Nuclear Waste Facility in BoBoTown?”

“Please select a response that most closely reflects your feelings:

1=Strongly Disagree 2=Disagree 3=Disagree a little 4=No Opinion 5=Agree a little 6=Agree 7=Strongly Disagree”

Here's the Data:

X	f	F(X)	(X- $\bar{x}$ )	(X- $\bar{x}$ ) <sup>2</sup>			
1	3	3*1= 3	1-3.9= -2.9	8.41*3= 25.23			
2	7	7*2= 14	2-3.9= -1.9	3.61*7= 25.27			
3	11	11*3= 33	3-3.9= -0.9	0.81*11= 8.91			
4	10	10*4= 40	4-3.9= 0.1	0.01*10= 0.10			
5	10	10*5= 50	5-3.9= 1.1	1.21*10= 12.10			
6	8	8*6= 48	6-3.9= 2.1	4.41*8= 35.28			
7	1	1*7= 7	7-3.9=3.1	9.60*1= 9.60			

What is the sample size NSIZE? \_\_\_\_\_

Calculate the Mean \_\_\_\_\_

Calculate the Mode \_\_\_\_\_

Calculate the Median \_\_\_\_\_

Calculate the Variance \_\_\_\_\_

Calculate the Standard Deviation (Std.) \_\_\_\_\_

Problem #2 next page

#2 - You are given the following data concerning the number of letters a random sample of 200 freshpersons wrote to their parents during the first term at college.

#Letters	#Freshpersons	f(X)-----	Deviation-----	Deviation <sup>2</sup> -----
0	40			
1	60			
2	50			
3	40			
4	10			
5	5			

What is the sample size NSIZE? \_\_\_\_\_ Calculate the Mean \_\_\_\_\_

Calculate the Mode \_\_\_\_\_ Calculate the Median \_\_\_\_\_

Calculate the Variance \_\_\_\_\_ Calculate the Standard Deviation (Std.) \_\_\_\_\_

#3 - You are the Quality Control Engineer for the People s Cannery in the Republic of Lower Gizmania. You have collected data on all 12 of your Dedicated Workers in the Cannery Workers Union #232. You have been instructed to increase production of Gumbo Units by offering a prize for the worker with the highest Gumbo Production Rate (GPR). Right now the average production rate is 16 Gumbos per hour.

(By the way - the prize is a night on the town with the 1946 Tractor Calendar Boy!)

Here s the Data: (there's a trick here!The Worker # is the frequency – there are 7 Hymie’s in this factory)

GuPH stands for Gumbo Units Per Hour

Worker#	GuPH	f(X)-----	Deviation-----	Deviation <sup>2</sup> -----
1 Vlad	19			
2 Emir	17			
3 Olaf	15			
4 Peg	18			
5 Gurn	24			
6 Bud	11			
7 Hymie	15			
8 Oscar	13			
9 Loeb	16			
10 Yair	22			
11 Fleen	21			
12 Dufus	20			

NSIZE? \_\_\_\_\_

Calculate the Mean \_\_\_\_\_

Calculate the Mode \_\_\_\_\_

Calculate the Median \_\_\_\_\_

Calculate the Variance \_\_\_\_\_ Calculate the Standard Deviation (Std.) \_\_\_\_\_

#4 Your social life is boring and you are interested in *ascertaining whether blondes really do have more fun than people with non-blonde hair*. To address this personally important question, you gather data from a sample of 40 blonde and 30 non-blonde undergraduates who are currently on the dating circuit.

You ask them this question: *Think about the fun you have on dates. Rate your personal dating experience on the scale below. Be careful to include even the so-so dates, & circle a number that most closely reflects the fun you have.*

*My Dates are generally:*

- 9= Far Far Above Average
- 8= Far Above Average
- 7= Above Average
- 6=Average
- 5= BelowAverage
- 4=Far Below Average
- 3=Far Far Below Average
- 2= Far Far Far Below Average
- 1= Really, Incredibly Below Average
- 0=Worse Than You Could Imagine

Dating	Non-Blonde	Blonde
Rating	Students	Students
0	1	1
1	3	4
2	5	4
3	6	5
4	8	10
5	3	7
6	2	6
7	1	2
8	1	1
9	0	0

(Careful, you are going to have to calculate two sets of everything!)

What is the sample size NSIZES? \_\_\_\_\_

Calculate the Means \_\_\_\_\_

Calculate the Modes \_\_\_\_\_

Calculate the Medians \_\_\_\_\_

Calculate the Variances \_\_\_\_\_

Calculate the Standard Deviations (Std.) \_\_\_\_\_

- a. Do blondes have more fun on dates? How d you arrive at your conclusion?
- b. What conclusions can you draw from the above data concerning fun and dates and being a blonde?
- c. Which of the students are most likely to flunk out of school by the end of the term & why?