



[www.epilepsy.va.gov/Statistics](http://www.epilepsy.va.gov/Statistics)

# Statistics in Evidence Based Medicine

## Lecture 2: Descriptive Summary Statistics

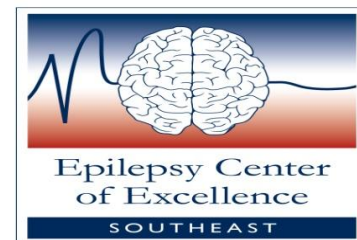
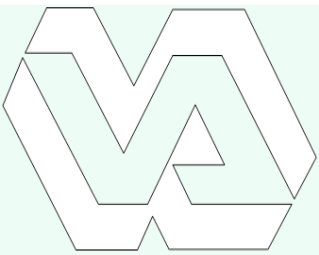
**Rizwana Rehman, PhD**

Regional Statistician

Southeast Epilepsy Center of Excellence  
Durham VA Medical Center, Durham NC

[Rizwana.Rehman@va.gov](mailto:Rizwana.Rehman@va.gov)

(919)286-0411 ext: 5024





# Overview

---

- Types of data and significant figures
- Frequency tables and distributions
- Descriptive summary statistics
  - Measures of central tendency
    - Mean, Median, Mode
  - Measures of dispersion
    - Range, Interquartile Range, Standard Deviation
- Reporting summary statistics
- Using Excel and Openstat

---

Audio Information: Dial 1-888-767-1050  
Conference ID 59058061



# Types of Data

## Quantitative

**Continuous**  
height, age

**Discrete**  
number of  
seizures per  
month

## Qualitative

### Categorical

**Ordinal**  
grade of  
breast cancer

**Nominal**  
sex, blood  
group



# Significant Figures

---

- After performing a computation only the first few non-zero digits of a number are important and we call these significant figures.
- The leading zeros in a number are not significant.

0.001096 has four significant figures. To three significant figures the number is 0.00110.

<http://www.usca.edu/chemistry/genchem/sigfig.htm>

---

Audio Information: Dial 1-888-767-1050  
Conference ID 59058061



# Frequency Table/Distribution

---

**A table showing the frequency of values in a data set.**

**Example: Test scores in a class ( $n = 23$ )**

65, 65, 70, 70, 70, 75, 75, 75, 75, 80, 80,  
80, 80, 80, 85, 85, 85, 85, 90, 90, 90,  
95, 95

---

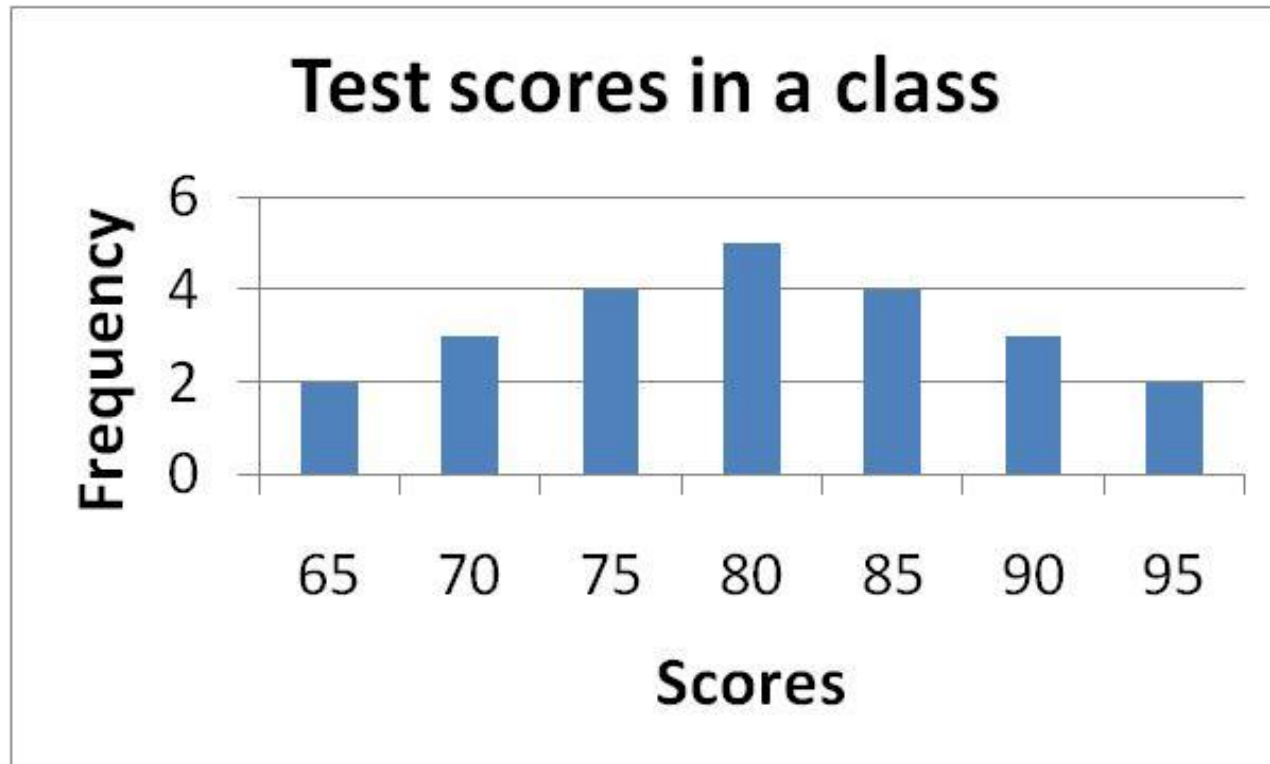
**Audio Information: Dial 1-888-767-1050  
Conference ID 59058061**



# Frequency Distribution of Test Scores

Test Scores	Frequency
65	2
70	3
75	4
80	5
85	4
90	3
95	2

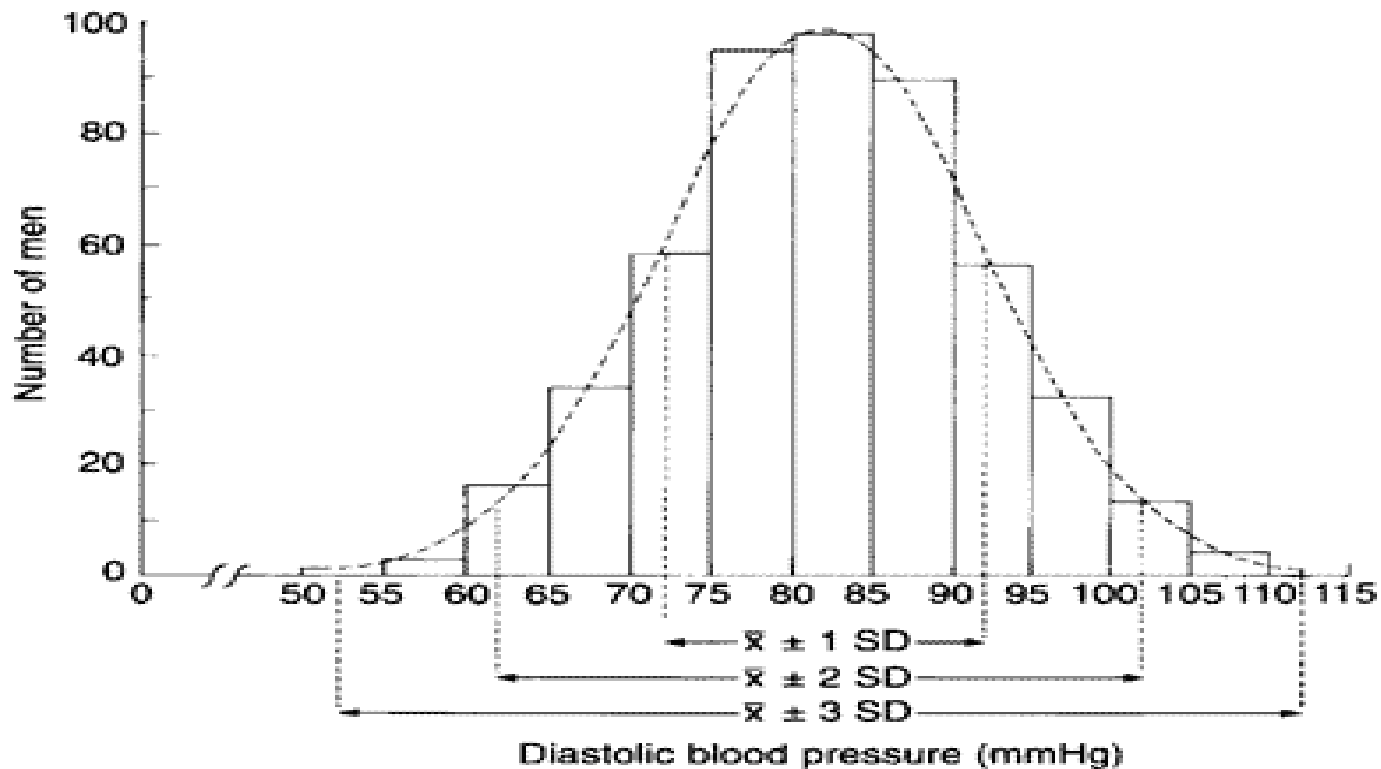
# Excel Column Graph of Test Scores



---

**Audio Information: Dial 1-888-767-1050  
Conference ID 59058061**

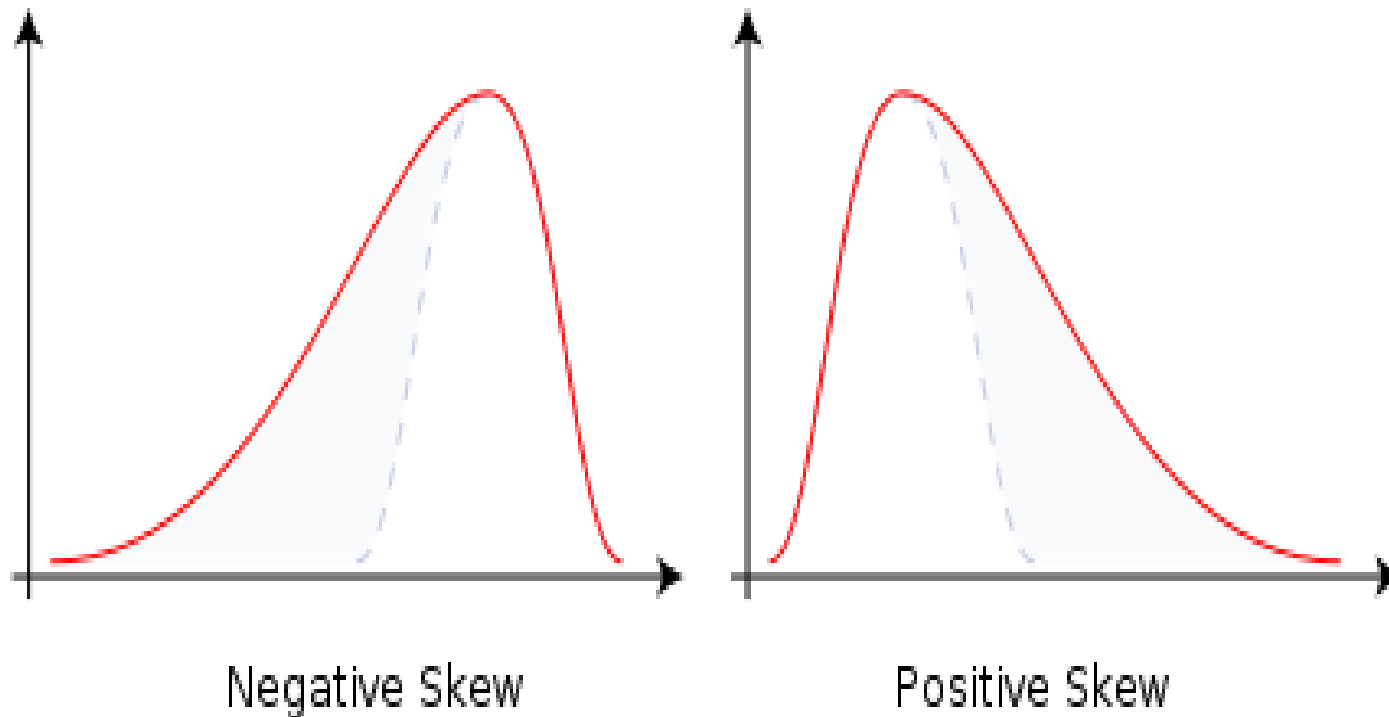
# Normal Frequency Distribution



Data distribution for 500 men  
Statistics at Square One: Figure 2.1



# Asymmetrical or Skewed Distributions



Audio Information: Dial 1-888-767-1050  
Conference ID 59058061



# Summary Statistics for Distributions

- Central Tendency
  - Where is the center?
  - What value is most common?
  - What should we use?
- Dispersion
  - How disperse is data?
  - What shape is it?
  - What should we use?

---

Audio Information: Dial 1-888-767-1050  
Conference ID 59058061



# Median: Where is Center?

---

- Median (midpoint) divides a dataset into two groups. Half of the data values lie below the median and half lie above the median.
- No conventional symbol. Sometimes  $Md$
- Median is known as **measure of location**.
- Median **may not** belong to the data set under investigation.



# Median for Odd Number of Values

---

Find the median of nine values:

0.6, 2.6, 1.1, 0.1, 0.4, 1.3, 1.2, 2.2, 1.9

**Step one:** Arrange observations in ascending order

0.1, 0.4, 0.6, 1.1, **1.2**, 1.3, 1.9, 2.2, 2.6

**Step two:** Look for the middle  $(n+1)/2$ th value

**Answer:** Median is **1.2** (in the data set)



# Median for Even Number of Values

---

Find the median of ten values:

0.6, 2.6, 1.1, 0.1, 0.4, 1.3, 1.2, 2.2, 1.9, 1.9

**Step one:** Arrange observations in ascending order

0.1, 0.4, 0.6, 1.1, 1.2, 1.3, 1.9, 1.9, 2.2, 2.6

**Step two:** Take the average of middle two observations.  $(1.2+1.3)/2$

**Answer:** The median is **1.25** (not in the data set)



# Properties of Median

---

- Median is **robust** to extreme values (outliers).

Median for 1.2, 10, 1.3, 1.3, 2.3 is **1.3**.

Median for 1.2, 1.3, 1.3, 2.3 is **1.3**.

- It is the closest point to all the observations.

$$|1.2-1.3|+|1.3-1.3|+|1.3-1.3|+|2.3-1.3|=1.1$$

**No other point will give us smaller difference than 1.1**

- **Median is less efficient; consider mean**



# Arithmetic Mean or Average

- Add all the observations and divide sum by number of observations to get the mean.

$$\bar{x} = \frac{\sum x}{n}$$

The mean of numbers 1.2, 10, 1.3, 1.3, 2.3 is 3.22

The mean of numbers 1.2, 1.3, 1.3, 2.3 is 1.525

- Mean is very sensitive to outliers.
- Mean of a population is denoted by  $\mu$

---

Audio Information: Dial 1-888-767-1050  
Conference ID 59058061



# Properties of Mean

---

- Sum of the deviations of a set of values from their arithmetic mean is **0**.

The mean of numbers 1.2, 1.3, 1.3, 2.3 is **1.525**

$$(1.2-1.525)+(1.3-1.525)+(1.3-1.525)+(2.3-1.525)=0$$

- It minimizes sum of squares of deviations from a point.

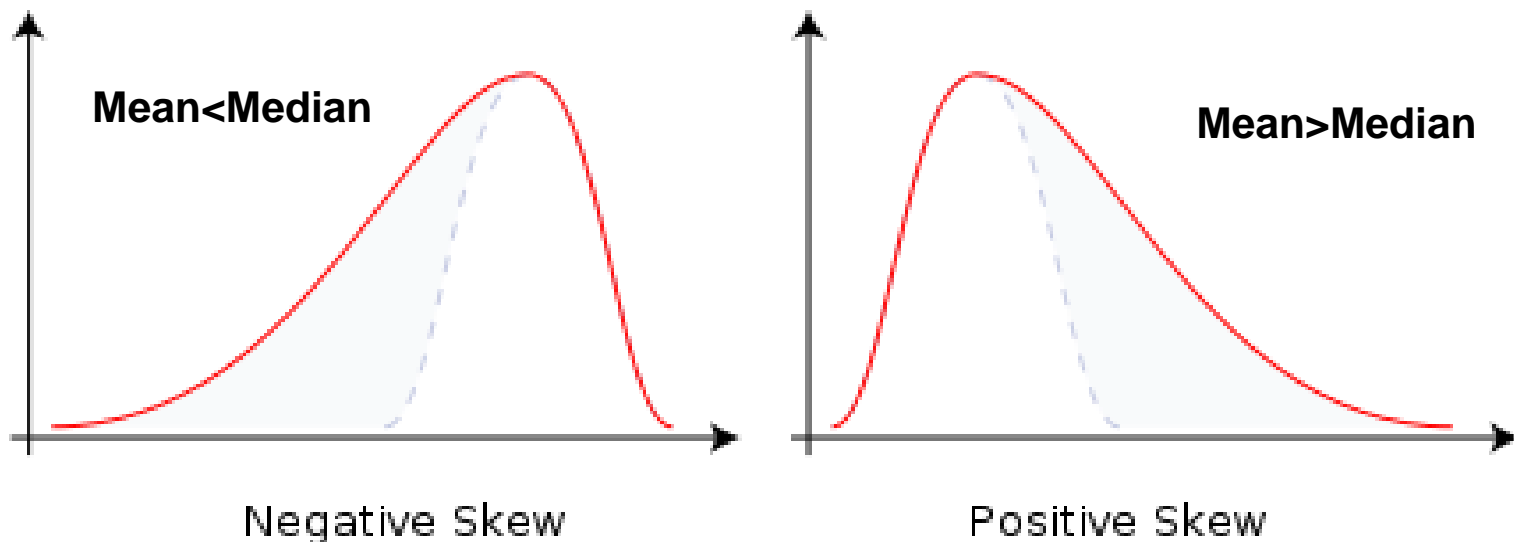
$$|1.2-1.525|^2+|1.3-1.525|^2+|1.3-1.525|^2+|2.3-1.525|^2= 0.8075$$

**No other point will give us smaller squared difference than 0.8075.**



# Using Median or Mean?

- For a symmetrical distribution (normal) mean and median **both are equal**
- For a skewed distribution use **median**





# Mean & Median both Useful!

---

Consider a variable that takes value one for males and zero for females

1, 0, 0, 1, 1, 1, 0

Mean is 0.57 & Median is 1

- Mean tells us the proportion of males
- Median tells us which group contained more than 50% of the people



## A Misconception about Mean

---

~~Mean shouldn't be used for ordinal data~~

**Mean from ordered categorical variables can be more useful some times.**

Consider rating of a lecture on a scale of 1 (poor) to 5 (excellent). Mean is a better summary statistic than median.



# Mode

---

**Mode is the value that occurs most frequently**

- The mode of data set 1, 3, 4, 5, 6, 6, 6, 2 is 6
- The modes of data set 1, 3, 4, 4, 4, 6, 6, 6 are 4 and 6

**Mode is used for bimodal distributions**



# Measure of Spread: The Range

---

**Range is the difference between smallest and largest observation.**

For data set 1, 2, 3, 10, 12 the range is 11.

---

Audio Information: Dial 1-888-767-1050  
Conference ID 59058061



# Interquartile Range

---

**Difference between the first and third quartiles is known as interquartile range (IQR). It contains the central 50% of observations.**

- Quartiles are points that divide the data into four quarters.
- 25% of observations lie below first quartile.
- 50% of observations lie below the second quartile (**median**) and 50% lie above it.
- 25% of observations lie above third quartile.



# Computing Quartiles

---

- Arrange the observations in ascending order.
- For  $n$  observations  $(n+1)/4^{\text{th}}$  observation is the first quartile.
- $(n+1)/2^{\text{th}}$  observation is the second quartile.
- $3(n+1)/4^{\text{th}}$  observation is the third quartile.

---

Audio Information: Dial 1-888-767-1050  
Conference ID 59058061



# Interquartile Range for Odd n

---

For data set 1, 1.5, 2, 2.75, 4, 5.5, 7.5

$n=7$

$(n+1)/4^{\text{th}}$  observation is  $2^{\text{nd}}$  observation  $\rightarrow Q_1=1.5$

$3(n+1)/4^{\text{th}}$  observation is  $6^{\text{th}}$  observation  $\rightarrow Q_3=5.5$

$$\text{IQR}=5.5-1.5=4.0$$





# Interquartile Range for Even n

For data set 1, 1.5, 2, 2.75, 4, 5.5, 7.5, 8

$n=8$

$(n+1)/4^{\text{th}}$  observation is  $2.25^{\text{th}}$

observation  $\rightarrow Q_1 = 1.5 + 0.25(2 - 1.5) = 1.625$

$3(n+1)/4^{\text{th}}$  observation is  $6.75^{\text{th}}$  observation

$\rightarrow Q_3 = 5.5 + 0.75(7.5 - 5.5) = 7.0$

$IQR = 7.0 - 1.625 = 5.375$



# Standard Deviation *SD*

---

- Standard Deviation is a measure of spread of data about their mean.

$$SD = s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

- The name of statistic before taking the square root is **variance**.
- The standard deviation of a population is denoted by  $\sigma$

# Computing Standard Deviation

$x$	$x - \bar{x}$	$(x - \bar{x})^2$
1	$1 - 3 = -2$	4
2	$2 - 3 = -1$	1
3	$3 - 3 = 0$	0
4	$4 - 3 = 1$	1
5	$5 - 3 = 2$	4
Total = 15	Total = 0	Total = 10

$$SD = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{10}{4}} = \sqrt{2.5} \cong 1.58$$



# Importance of Standard Deviation for Bell Shaped Distribution

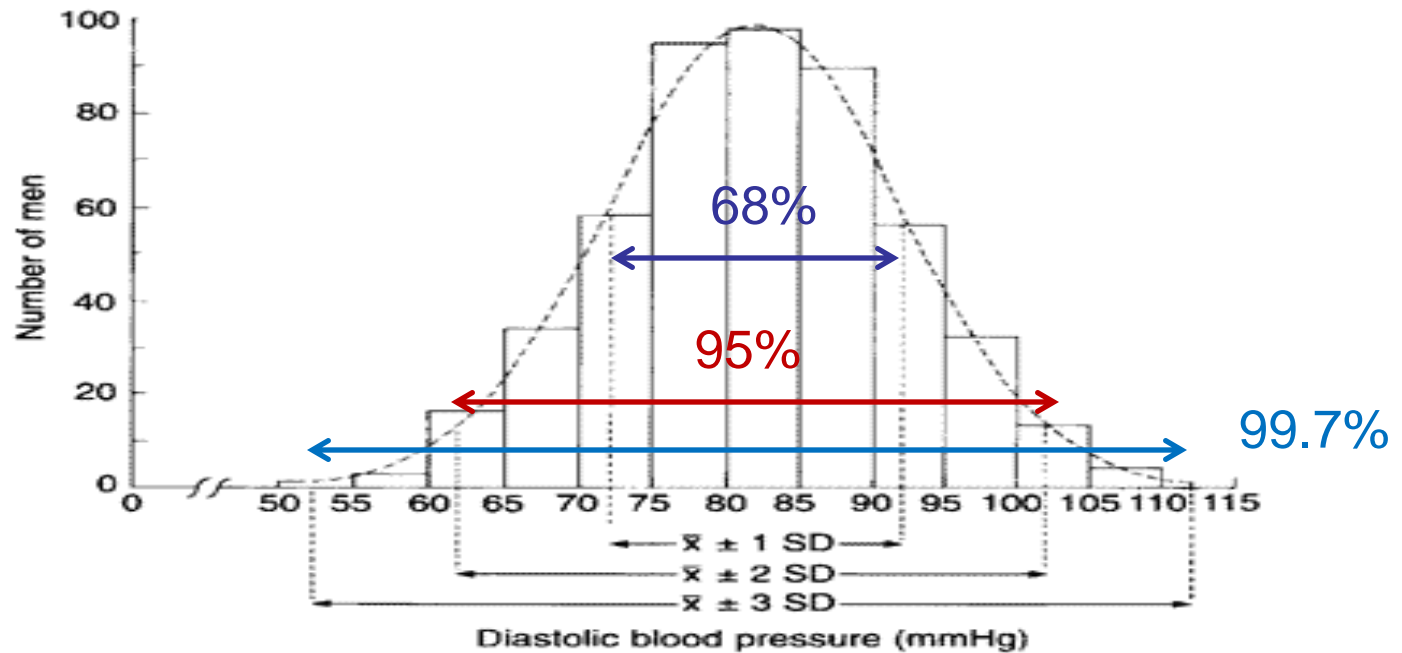
---

- 68% of observations lie between the mean  $\pm 1$  SD
- 95% of observations lie between the mean  $\pm 2$  SD
- 99.7% of observations lie between the mean  $\pm 3$  SD

---

Audio Information: Dial 1-888-767-1050  
Conference ID 59058061

# Dispersion of Data in Normal Distribution



**Mean=82mmHG, SD=10mmHG, n= 500**

**Mean 1 SD=(72 – 92) contains approximately 340 observations**

**Mean 2 SD=(62 – 102) contains approximately 475 observations**

**Mean 3 SD=(52 – 112) contains approximately 499 observations**



# The Coefficient of Variation *CV*

---

The coefficient of variation is obtained from dividing *SD* by the mean and multiplying by 100. It is a measure of relative spread in data.

$$CV = \frac{SD}{\bar{x}} \times 100$$

Use CV when comparing data sets with different units or widely different means



## Example: The Coefficient of Variation $CV$

---

shock index data: mean = 0.69, SD = 0.20

systolic blood pressure data : mean = 138, SD = 0.26

$CV$  for shock index data = 29.0%

$CV$  systolic blood pressure data = 18.8%

Basics and Clinical Biostatistics: Chapter 3

Kline et al. (2002)



# Using Different Measures of Dispersion

---

- Use standard deviation with mean for a symmetric distribution only.
- With skewed data use median and interquartile range. Also report mean.
- Use range to show extreme values.
- For comparison of distributions measured on different scales use coefficient of variation.





# Displaying Summary Statistics

---

- Display mean to one more significant digit than data.

The mean of numbers 1.2, 1.3, 1.3, 2.3 is 1.525; report mean=**1.53**

- Display standard deviation to two more significant figures than data.

The SD Of number 1, 2, 3, 4, 5 is 1.58113883008419----. report standard deviation=**1.58**



# Displaying Summary Statistics

---

- When quoting a range or interquartile range, give the two numbers that define it.

For data set 1, 2, 3, 10, 12 report range= (1–12) or range=1 to 12

For data set 1, 1.5, 2, 2.75, 4, 5.5, 7.5 report  
interquartile range=(1.5–5.5)

- Median and IQR should be given to the same accuracy as the data or one extra significant digit if average of two numbers is needed.

For data set 1, 1.5, 2, 2.75, 4, 5.5, 7.5, 8 report median=3.375  
interquartile range =(3.125–7.0)



# Checking Skewness using Summary Statistics

---

If mean or median is near the lower limit of range or interquartile range, then distribution is positively skewed and vice versa.

median=0.46, mean=0.51, SD=0.22, range=0.51 to 1.66,  
interquartile range=0.35 to 0.60

positive skewness

[http://www-users.york.ac.uk/~mb55/msc/applbio/week3/sd\\_text.pdf](http://www-users.york.ac.uk/~mb55/msc/applbio/week3/sd_text.pdf)

median=9, mean=8.2, SD=4.367, range=1 to 15,  
interquartile range=6.25 to 10.75

data set: 1,2,6,7,8,10,10,11,12,15

negative skewness



# Using Excel for Today's Lecture

---

- Adding in Data Analysis tool box

<http://cameron.econ.ucdavis.edu/excel/ex01access.html>

- For descriptive summary statistics look at

<http://cameron.econ.ucdavis.edu/excel/ex21descriptivestatistics.html>

---

Audio Information: Dial 1-888-767-1050  
Conference ID 59058061



# Using Excel 2007 for Summary Statistics

---

## Test scores in a class

65, 65, 70, 70, 70, 75, 75, 75, 75, 80, 80,  
80, 80, 80, 85, 85, 85, 85, 90, 90, 90,  
95, 95



# Excel 2007 for Summary Statistics

<i>Column1</i>	
Mean	80
Standard Error	1.832944
Median	80
Mode	80
Standard Deviation	8.790491
Sample Variance	77.27273
Kurtosis	-0.78082
Skewness	0
Range	30
Minimum	65
Maximum	95
Sum	1840
Count	23

# Using Openstat

<http://www.statprograms4u.com/>

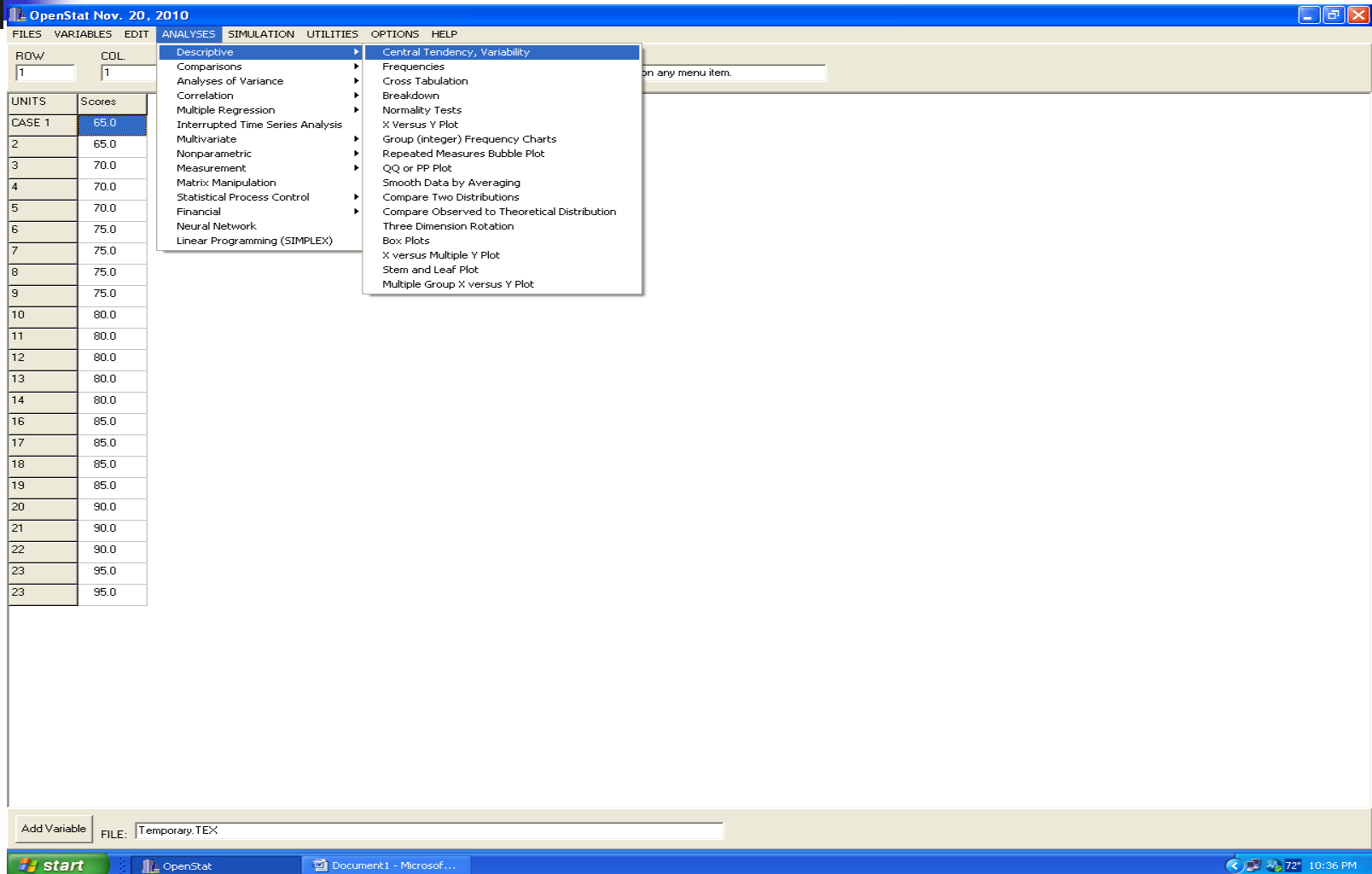
## Entering data

The screenshot displays the OpenStat software interface. The title bar reads "OpenStat Nov. 20, 2010". The menu bar includes "FILES", "VARIABLES", "EDIT", "ANALYSES", "SIMULATION", "UTILITIES", "OPTIONS", and "HELP". Below the menu bar, there are input fields for "ROW" (1), "COL" (1), "Cell Edit (Return to finish)" (65.0), "N CASES" (23), "No. VAR.S" (1), "ASCII" (13), and "STATUS:" (Press F1 for help when on any menu item.).

The main data entry area shows a table with two columns: "UNITS" and "Scores". The "UNITS" column lists "CASE 1" followed by rows 2 through 23. The "Scores" column contains the following values: 65.0, 65.0, 70.0, 70.0, 70.0, 75.0, 75.0, 75.0, 75.0, 80.0, 80.0, 80.0, 80.0, 80.0, 85.0, 85.0, 85.0, 85.0, 90.0, 90.0, 90.0, 95.0, and 95.0.

At the bottom of the window, there is a status bar with "Add Variable" and "FILE: Temporary.TEX". The Windows taskbar at the very bottom shows the "start" button, "OpenStat", "Document1 - Microsof...", and the system clock "10:35 PM".

# Choose Main Menu



OpenStat Nov. 20, 2010

FILES VARIABLES EDIT ANALYSES SIMULATION UTILITIES OPTIONS HELP

ROW COL  
1 1

UNITS Scores

CASE 1 65.0

2 65.0

3 70.0

4 70.0

5 70.0

6 75.0

7 75.0

8 75.0

9 75.0

10 80.0

11 80.0

12 80.0

13 80.0

14 90.0

16 85.0

17 85.0

18 85.0

19 85.0

20 90.0

21 90.0

22 90.0

23 95.0

23 95.0

Descriptive  
Comparisons  
Analyses of Variance  
Correlation  
Multiple Regression  
Interrupted Time Series Analysis  
Multivariate  
Nonparametric  
Measurement  
Matrix Manipulation  
Statistical Process Control  
Financial  
Neural Network  
Linear Programming (SIMPLEX)

Central Tendency, Variability  
Frequencies  
Cross Tabulation  
Breakdown  
Normality Tests  
X Versus Y Plot  
Group (integer) Frequency Charts  
Repeated Measures Bubble Plot  
QQ or PP Plot  
Smooth Data by Averaging  
Compare Two Distributions  
Compare Observed to Theoretical Distribution  
Three Dimension Rotation  
Box Plots  
X versus Multiple Y Plot  
Stem and Leaf Plot  
Multiple Group X versus Y Plot

Add Variable FILE: Temporary.TEX

start OpenStat Document1 - Microsof...

72° 10:36 PM



# Choose Options

OpenStat Nov. 20, 2010

FILES VARIABLES EDIT ANALYSES SIMULATION UTILITIES OPTIONS HELP

ROW  
1

UNITS

CASE 1

2

3

4

5

6

7

8

9

10

11

12

13

14

16 85.0

17 85.0

18 85.0

19 85.0

20 90.0

21 90.0

22 90.0

23 95.0

23 95.0

**Descriptive Statistics**

This procedure provides means, variances, standard deviations, skewness, kurtosis and range values for each variable selected. Select the variables in the left list and enter them for analysis by clicking the right arrow button. If you select the z score option, a new variable will be added to your grid for each variable you select. The new variable will contain the transformation of the original variable into a z score. If you elect the case-wise deletion option, the calculations will be done for all valid values of each variable otherwise a list-wise deletion of records will occur in any one of the variables contains

Available Variables

Variables to Analyze

Scores

Options:

- ☒ Sample size, Sum
- ☒ Mean, Var., Std.Dev.
- ☒ Std. Error of Mean
- ☐ Confidence Interval
- ☐ Geometric Mean (positive values)
- ☐ Harmonic Mean (no zeroes)
- ☒ Range
- ☒ Skewness
- ☒ Kurtosis
- ☒ Quartiles
- ☐ CaseWise Deletion
- ☐ z Scores to Grid
- ☐ Print Multiple Method Quartiles
- ☐ Print Percentile Ranks

Two-Tailed Confidence Interval: 0.95

Reset

Cancel

OK

Add Variable

FILE: Temporary.TEX

start OpenStat Document1 - Microsof...

72° 10:36 PM

# Results

OpenStat Nov. 20, 2010

FILES VARIABLES EDIT ANALYSES SIMULATION UTILITIES OPTIONS HELP

ROW 1

UNITS

CASE 1

2

3

4

5

6

7

8

9

10

11

12

13

14

16

17

18

19

20

21

22

23

23

Results Window

DISTRIBUTION PARAMETER ESTIMATES

=====

Scores (N = 23) Sum = 1840.000

Scores (N = 23) Mean = 80.000 Variance = 77.273 Std.Dev. = 8.790

Std.Error of Mean = 1.833

Range = 30.000 Minimum = 65.000 Maximum = 95.000

Skewness = 0.000 Std. Error of Skew = 0.481

Kurtosis = -0.781 Std. Error Kurtosis = 0.935

Median = 80.000

Q1 = 75.000

Q3 = 85.000

Interquartile range = 10.000

95.0

95.0

Add Variable FILE: Temporary.TEX

start OpenStat Document1 - Microsof...

71° 10:38 PM



# Thank you!

---

**Questions/Comments**

[Rizwana.Rehman@va.gov](mailto:Rizwana.Rehman@va.gov)

**(919) 286-0411 ext: 5024**

For more information, program materials,  
and to complete evaluation for CME  
credit visit

[www.epilepsy.va.gov/Statistics](http://www.epilepsy.va.gov/Statistics)