What is statistics and the need for data management!

Fares Qeadan, Ph.D

Department of Internal Medicine

Division of Epidemiology, Biostatistics, & Preventive Medicine
University of New Mexico Health Sciences Center

November 2, 2015

UNM HEALTH SCIENCES CENTER | Clinical & Translational Science Center

MOUNTAIN WEST CTR-IN

ALASKA HAWAII IDAHO MONTANA NEVADA NEW MEXICO WYOMING
CLINICAL TRANSLATIONAL RESEARCH - INFRASTRUCTURE NETWORK

- What is statistics
    - Definitions: *Population, Sample, Parameter(s) and Statistic(s)*
    - Descriptive Statistics
    - Inferential Statistics
    - Sampling Methods
    - Sample Size Calculation

- What is statistics
    - Definitions: *Population, Sample, Parameter(s) and Statistic(s)*
    - Descriptive Statistics
    - Inferential Statistics
    - Sampling Methods
    - Sample Size Calculation
- Data Management
    - Data management cycle
    - Sources of data
    - Softwares for data analysis and management
    - Guidelines for Effective Data Management
    - How to deal with Big Data

- What is statistics
  - Definitions: *Population, Sample, Parameter(s) and Statistic(s)*
  - Descriptive Statistics
  - Inferential Statistics
  - Sampling Methods
  - Sample Size Calculation
- Data Management
  - Data management cycle
  - Sources of data
  - Softwares for data analysis and management
  - Guidelines for Effective Data Management
  - How to deal with Big Data
- References

- What is statistics
    - Definitions: *Population, Sample, Parameter(s) and Statistic(s)*
    - Descriptive Statistics
    - Inferential Statistics
    - Sampling Methods
    - Sample Size Calculation
- Data Management
    - Data management cycle
    - Sources of data
    - Softwares for data analysis and management
    - Guidelines for Effective Data Management
    - How to deal with Big Data
- References
- Citation

Statistics as a Science:

It's the use of numerical or categorical data to explain a phenomenon or an experiment. Therefore, statistics involves the development and application of methods to:

- collect
- analyze and
- interpret

data. The ultimate goal of statistics is to estimate the unknown parameters of a particular population via statistics from the sample data.

Statistics as a Science:

It's the use of numerical or categorical data to explain a phenomenon or an experiment. Therefore, statistics involves the development and application of methods to:

- collect
- analyze and
- interpret

data. The ultimate goal of statistics is to estimate the unknown parameters of a particular population via statistics from the sample data.

- **Population:** All subjects of interest.
- **Sample:** A subset of the population (drawn from the sampling frame).

Statistics as a Science:

It's the use of numerical or categorical data to explain a phenomenon or an experiment. Therefore, statistics involves the development and application of methods to:

- collect
- analyze and
- interpret

data. The ultimate goal of statistics is to estimate the unknown parameters of a particular population via statistics from the sample data.

- **Population:** All subjects of interest.
- **Sample:** A subset of the population (drawn from the sampling frame).
- **Parameter:** A numerical measurement describing some characteristic of a population.

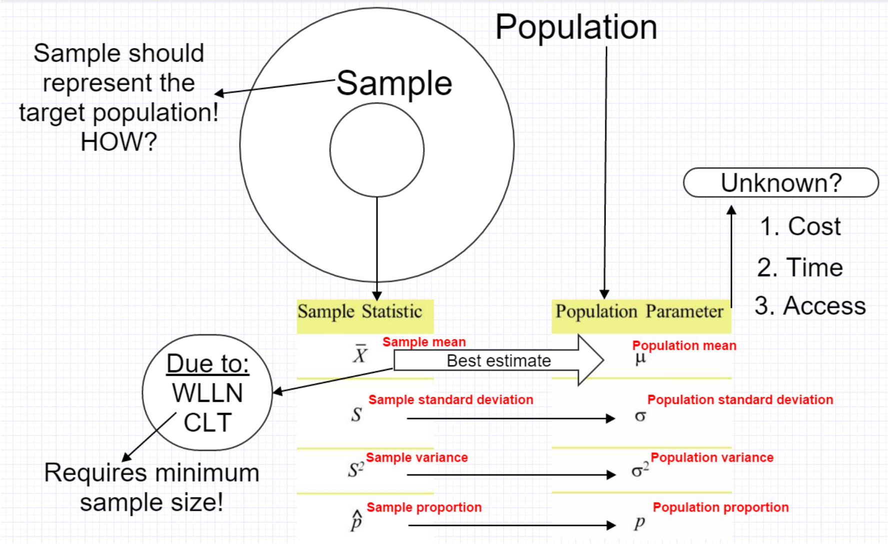
Statistics as a Science:

It's the use of numerical or categorical data to explain a phenomenon or an experiment. Therefore, statistics involves the development and application of methods to:

- collect
- analyze and
- interpret

data. The ultimate goal of statistics is to estimate the unknown parameters of a particular population via statistics from the sample data.

- **Population:** All subjects of interest.
- **Sample:** A subset of the population (drawn from the sampling frame).
- **Parameter:** A numerical measurement describing some characteristic of a population.
- **Statistic:** A numerical measurement describing some characteristic of a sample.



Population

Sample

Sample should represent the target population! HOW?

Unknown?

1. Cost
2. Time
3. Access

Sample Statistic

Population Parameter

$\bar{X}$ Sample mean → Best estimate → Population mean μ

S Sample standard deviation → Population standard deviation σ

S^2 Sample variance → Population variance σ^2

$\hat{p}$ Sample proportion → Population proportion p

Due to: WLLN CLT

Requires minimum sample size!

Descriptive Statistics:

It's a branch of statistics in which data are only used for descriptive purposes and are not employed to make inferences. Thus, descriptive statistics is concerned with numerical or graphical description of observed data (i.e. the sample data) via their values and summary statistics. The main graphical descriptive methods are pie-chart, bar-chart, box-plot, histogram an stem & leaf.

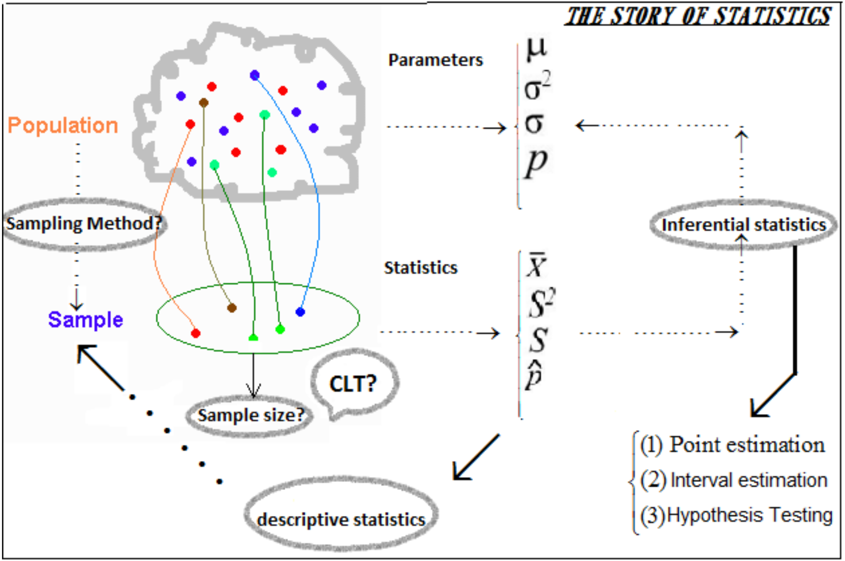
Descriptive Statistics:

It's a branch of statistics in which data are only used for descriptive purposes and are not employed to make inferences. Thus, descriptive statistics is concerned with numerical or graphical description of observed data (i.e. the sample data) via their values and summary statistics. The main graphical descriptive methods are pie-chart, bar-chart, box-plot, histogram an stem & leaf.

Inferential Statistics:

It's a branch of statistics in which conclusions or generalizations are made about the population parameters by using the sample statistics. The main components of inferential statistics are:

- Point estimation
- Interval estimation and
- Hypothesis testing


THE STORY OF STATISTICS
Population
Sampling Method?
Sample
Sample size?
CLT?
descriptive statistics
Parameters
μ
σ^2
σ
p
Statistics
$\bar{X}$
S^2
S
$\hat{p}$
Inferential statistics
(1) Point estimation
(2) Interval estimation
(3) Hypothesis Testing

Sampling Methods:

They are techniques that we use to draw a sample from a particular population (in practice from the sampling frame). Sampling Methods can be classified into one of two categories:

- Random sampling (Probability Sampling)
- Non-random Sampling (Non-probability Sampling)

Sampling Methods:

They are techniques that we use to draw a sample from a particular population (in practice from the sampling frame). Sampling Methods can be classified into one of two categories:

- Random sampling (Probability Sampling)
- Non-random Sampling (Non-probability Sampling)

- Probability (Random) Sampling
  - Simple random sampling (SRS)
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling

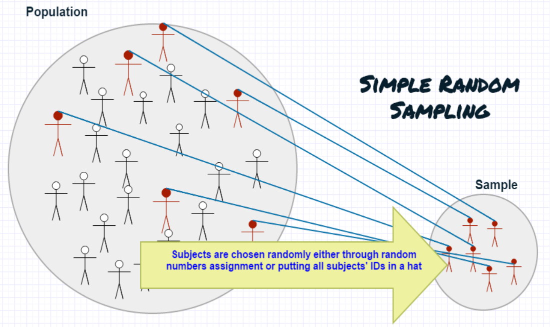
Sampling Methods:

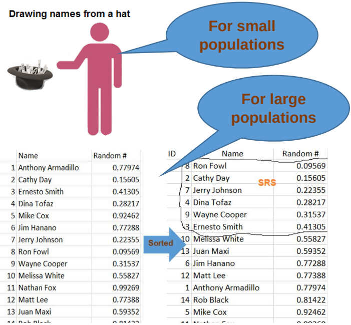
They are techniques that we use to draw a sample from a particular population (in practice from the sampling frame). Sampling Methods can be classified into one of two categories:

- Random sampling (Probability Sampling)
- Non-random Sampling (Non-probability Sampling)

- Probability (Random) Sampling
  - Simple random sampling (SRS)
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling
- Non-Probability Sampling
  - Convenience sampling
  - Volunteer sampling
  - Judgment (Purposive), Snowball, and Quota sampling

Simple Random Sampling (SRS): It's a sampling method in which each subject of the *sampling frame* has an equal chance of being selected into the sample [1]. SRS is the most popular method of random sampling. There are two types of SRS: with replacement and without replacement. SRS with replacement is less common.

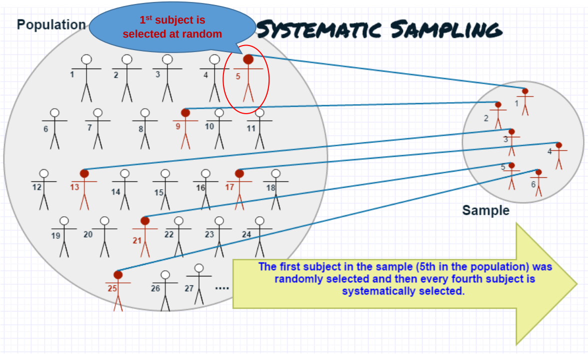

Population
SIMPLE RANDOM SAMPLING
Sample
Subjects are chosen randomly either through random numbers assignment or putting all subjects' IDs in a hat



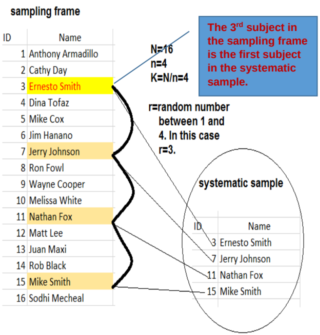
	Name	Random #
1	Anthony Armadillo	0.77974
2	Cathy Day	0.15605
3	Ernesto Smith	0.41305
4	Dina Tofaz	0.28217
5	Mike Cox	0.92462
6	Jim Hanano	0.77288
7	Jerry Johnson	0.22355
8	Ron Fowl	0.09569
9	Wayne Cooper	0.31537
10	Melissa White	0.55827
11	Nathan Fox	0.99269
12	Matt Lee	0.77388
13	Juan Maxi	0.59352
14	Rob Black	0.81422

ID	Name	Random #
8	Ron Fowl	0.09569
2	Cathy Day	0.15605
7	Jerry Johnson	0.22355
4	Dina Tofaz	0.28217
9	Wayne Cooper	0.31537
3	Ernesto Smith	0.41305
10	Melissa White	0.55827
13	Juan Maxi	0.59352
6	Jim Hanano	0.77288
12	Matt Lee	0.77388
1	Anthony Armadillo	0.77974
14	Rob Black	0.81422
5	Mike Cox	0.92462
11	Nathan Fox	0.99269

Systematic sampling: It's a sampling method in which subjects are chosen in a systematic way such that one first randomly picks the first subject from the sampling frame and then selects each *kth* subject from the list ($k = N/n$) [1]. If the sampling frame is randomly shuffled, then systematic sampling is equivalent to SRS.



The first subject in the sample (5th in the population) was randomly selected and then every fourth subject is systematically selected.



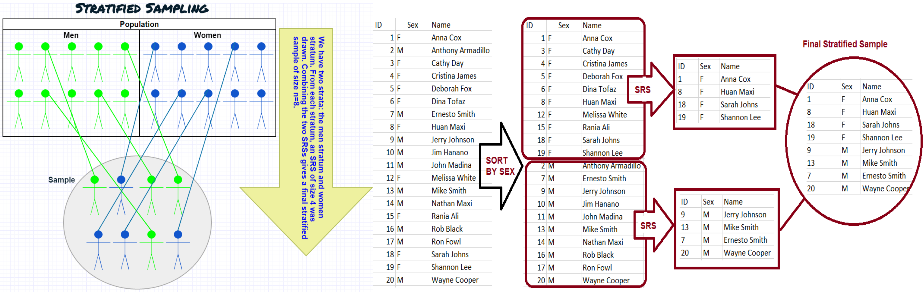
ID	Name
1	Anthony Armadillo
2	Cathy Day
3	Ernesto Smith
4	Dina Tofaz
5	Mike Cox
6	Jim Hanano
7	Jerry Johnson
8	Ron Fowl
9	Wayne Cooper
10	Melissa White
11	Nathan Fox
12	Matt Lee
13	Juan Maxi
14	Rob Black
15	Mike Smith
16	Sodhi Mecheal

The 3rd subject in the sampling frame is the first subject in the systematic sample.

r=random number between 1 and 4. In this case r=3.

ID	Name
3	Ernesto Smith
7	Jerry Johnson
11	Nathan Fox
15	Mike Smith

Stratified sampling: It's a sampling method in which a sample is obtained by firstly dividing the population into subpopulations (strata) based on some characteristics and then an SRS is taken from each stratum [1]. Combining the obtained SRSs will give the final stratified sample. Minority subgroups of interest can be ensured by stratification. There are two types of stratified sampling: proportionate and disproportionate. In the proportionate one, we draw a sample from each stratum in proportion to its share in the target population. By this method, each stratum should be internally homogeneous.



We have two strata: the men stratum and women stratum. From each stratum, an SRS of size 4 was drawn. Combining the two SRSs gives a final stratified sample of size n=8.

ID	Sex	Name
1	F	Anna Cox
2	M	Anthony Armadillo
3	F	Cathy Day
4	F	Cristina James
5	F	Deborah Fox
6	F	Dina Tofaz
7	M	Ernesto Smith
8	F	Huan Maxi
9	M	Jerry Johnson
10	M	Jim Hanano
11	M	John Madina
12	F	Melissa White
13	M	Mike Smith
14	M	Nathan Maxi
15	F	Rania Ali
16	M	Rob Black
17	M	Ron Fowl
18	F	Sarah Johns
19	F	Shannon Lee
20	M	Wayne Cooper

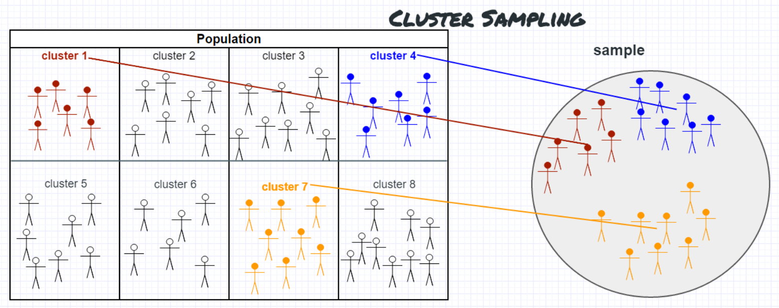
ID	Sex	Name
1	F	Anna Cox
3	F	Cathy Day
4	F	Cristina James
5	F	Deborah Fox
6	F	Dina Tofaz
8	F	Huan Maxi
12	F	Melissa White
15	F	Rania Ali
18	F	Sarah Johns
19	F	Shannon Lee
2	M	Anthony Armadillo
7	M	Ernesto Smith
9	M	Jerry Johnson
10	M	Jim Hanano
11	M	John Madina
13	M	Mike Smith
14	M	Nathan Maxi
16	M	Rob Black
17	M	Ron Fowl
20	M	Wayne Cooper

ID	Sex	Name
1	F	Anna Cox
8	F	Huan Maxi
18	F	Sarah Johns
19	F	Shannon Lee

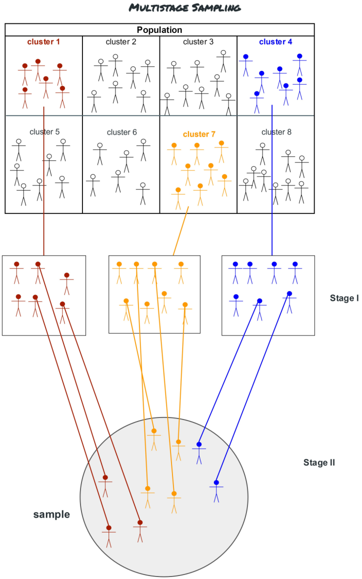
ID	Sex	Name
9	M	Jerry Johnson
13	M	Mike Smith
7	M	Ernesto Smith
20	M	Wayne Cooper

ID	Sex	Name
1	F	Anna Cox
8	F	Huan Maxi
18	F	Sarah Johns
19	F	Shannon Lee
9	M	Jerry Johnson
13	M	Mike Smith
7	M	Ernesto Smith
20	M	Wayne Coope

Cluster sampling: It's a sampling method in which the target population is first divided into naturally occurring clusters and then a random sample of clusters is obtained such that all subjects in the randomly selected clusters are included in the sample [1]. Sometimes, we include an SRS from each selected cluster instead of including all subjects which makes the sampling method to be called a two-stage sampling method. By this method, clusters should be internally as heterogeneous as the target population itself.



Multistage sampling: It's a sampling method in which we use combinations of two or more sampling methods at least one of which involves randomness [2].



Sample Size Calculation: It's an important part of the study design to ensure validity, accuracy, reliability and, scientific and ethical integrity of the study [3]. In general, the main aim of a sample size calculation is to determine the number of participants needed to detect a clinically relevant treatment effect. Formulas for sample size calculation depend on four factors:

- The significance level α
- The power of the test $1 - \beta$
- The type of the conducted test (t-test, z-test, chi-square test, etc.)
- The type of the design (case-control versus prospective)

Sample Size Calculation: It's an important part of the study design to ensure validity, accuracy, reliability and, scientific and ethical integrity of the study [3]. In general, the main aim of a sample size calculation is to determine the number of participants needed to detect a clinically relevant treatment effect. Formulas for sample size calculation depend on four factors:

- The significance level α
- The power of the test $1-\beta$
- The type of the conducted test (t-test, z-test, chi-square test, etc.)
- The type of the design (case-control versus prospective)

For sufficiently large sample size, both the law of large numbers (WLLN) as well as the central limit theorem (CLT) will work:

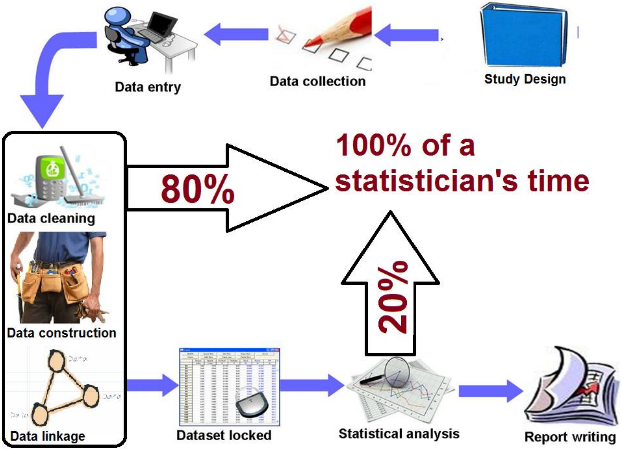
CLT:

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

WLLN:

$$\bar{x}_n \longrightarrow \mu, n \longrightarrow \infty$$

Data management cycle: It's the set of all steps taken starting from the conception of the study (study design) through the reporting of the results and archiving the data for future reusability [1].



[1] This figure is a modification of a figure taken from [4].

Sources of data:

- Censuses
- Surveys
- Experiments
- Registries
- Electronic Medical Records
- Secondary data (BRFSS, NHIS, Medicare & Medicaid, etc.)
- Social Media Data
- Publications

Softwares for data analysis and management:

- Excel
- Access (only for database management)
- SAS (has its own SQL Structured Query Language)
- SUDAAN (good for complex sample surveys)
- R
- SPSS
- Minitab
- STATA
- S-Plus
- PASS (only for sample size calculation)
- Epi Info (free by the CDC)
- REDCap (is a mature, secure web application for building and managing online surveys and databases)

REDCap™
Research Electronic Data Capture

Link to REDCap's website: http://project-redcap.org/

Link to REDCap's website from UNM:
http://hsc.unm.edu/research/ctsc/informatics/REDCap.shtml

Workshop on how to use REDCap by the CTSC:
Next Class:
Wednesday, November 18, 2015
10:00 am - 12:00 pm
HSC Library Room 226
Follow link below to register
http://hsc.unm.edu/research/ctsc/Informatics/REDCapTraining.shtml

Guidelines for Effective Data Management[5]:

- Use scripted programs for analysis and not ”GUI-driven” ones [for reusability and modification purposes].

Guidelines for Effective Data Management[5]:

- Use scripted programs for analysis and not "GUI-driven" ones [for reusability and modification purposes].
- Store data in non-proprietary software formats (comma delimited text file, .csv versus excel) [for compatibility and format purposes]

Guidelines for Effective Data Management[5]:

- Use scripted programs for analysis and not ”GUI-driven” ones [for reusability and modification purposes].
- Store data in non-proprietary software formats (comma delimited text file, .csv versus excel) [for compatibility and format purposes]
- Always store an uncorrected data file with all of its bumps and warts. Do not make any corrections to this file, make corrections within a scripted language [for accidental mistakes you could always go back to this original file]

Guidelines for Effective Data Management[5]:

- Use scripted programs for analysis and not ”GUI-driven” ones [for reusability and modification purposes].
- Store data in non-proprietary software formats (comma delimited text file, .csv versus excel) [for compatibility and format purposes]
- Always store an uncorrected data file with all of its bumps and warts. Do not make any corrections to this file, make corrections within a scripted language [for accidental mistakes you could always go back to this original file]
- Use descriptive and informative names for your data files [let the name capture both the place and time]

Guidelines for Effective Data Management[5]:

- Use scripted programs for analysis and not "GUI-driven" ones [for reusability and modification purposes].
- Store data in non-proprietary software formats (comma delimited text file, .csv versus excel) [for compatibility and format purposes]
- Always store an uncorrected data file with all of its bumps and warts. Do not make any corrections to this file, make corrections within a scripted language [for accidental mistakes you could always go back to this original file]
- Use descriptive and informative names for your data files [let the name capture both the place and time]
- All cells within each column should contain only one data type (i.e. either text, numerical or date.)

Guidelines for Effective Data Management[5]:

- Use scripted programs for analysis and not "GUI-driven" ones [for reusability and modification purposes].
- Store data in non-proprietary software formats (comma delimited text file, .csv versus excel) [for compatibility and format purposes]
- Always store an uncorrected data file with all of its bumps and warts. Do not make any corrections to this file, make corrections within a scripted language [for accidental mistakes you could always go back to this original file]
- Use descriptive and informative names for your data files [let the name capture both the place and time]
- All cells within each column should contain only one data type (i.e. either text, numerical or date.)
- Avoid having more than one record for each ID, instead create a relational database (sometimes it's no possible!)

Guidelines for Effective Data Management[5]:

- Use scripted programs for analysis and not ”GUI-driven” ones [for reusability and modification purposes].
- Store data in non-proprietary software formats (comma delimited text file, .csv versus excel) [for compatibility and format purposes]
- Always store an uncorrected data file with all of its bumps and warts. Do not make any corrections to this file, make corrections within a scripted language [for accidental mistakes you could always go back to this original file]
- Use descriptive and informative names for your data files [let the name capture both the place and time]
- All cells within each column should contain only one data type (i.e. either text, numerical or date.)
- Avoid having more than one record for each ID, instead create a relational database (sometimes it's no possible!)
- Create metadata/dictionary/codebook.

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)
  - Check for data types (numeric versus character versus date)

- Data Construction

- Data Cleaning (Maintain Data Quality)
    - Check variables names (make them informative)
    - Check for data types (numeric versus character versus date)
    - Check for missing values (Tabulate frequency)
- Data Construction
- Data Linkage

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)
  - Check for data types (numeric versus character versus date)
  - Check for missing values (Tabulate frequency)
  - Check for default values (unknown/refused: Tabulate)
- Data Construction
- Data Linkage

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)
  - Check for data types (numeric versus character versus date)
  - Check for missing values (Tabulate frequency)
  - Check for default values (unknown/refused: Tabulate)
  - Check for uniqueness (duplication: subsets by IDs)
- Data Construction
- Data Linkage

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)
  - Check for data types (numeric versus character versus date)
  - Check for missing values (Tabulate frequency)
  - Check for default values (unknown/refused: Tabulate)
  - Check for uniqueness (duplication: subsets by IDs)
  - Check for accuracy (typos in data entry: Tabulate)
- Data Construction
- Data Linkage

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)
  - Check for data types (numeric versus character versus date)
  - Check for missing values (Tabulate frequency)
  - Check for default values (unknown/refused: Tabulate)
  - Check for uniqueness (duplication: subsets by IDs)
  - Check for accuracy (typos in data entry: Tabulate)
  - Check for quality controls (min and max values: Box-plots, Quality control charts)
- Data Construction
- Data Linkage

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)
  - Check for data types (numeric versus character versus date)
  - Check for missing values (Tabulate frequency)
  - Check for default values (unknown/refused: Tabulate)
  - Check for uniqueness (duplication: subsets by IDs)
  - Check for accuracy (typos in data entry: Tabulate)
  - Check for quality controls (min and max values: Box-plots, Quality control charts)
- Data Construction
  - Create new variables (from continuous/categorical to categorical)
- Data Linkage

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)
  - Check for data types (numeric versus character versus date)
  - Check for missing values (Tabulate frequency)
  - Check for default values (unknown/refused: Tabulate)
  - Check for uniqueness (duplication: subsets by IDs)
  - Check for accuracy (typos in data entry: Tabulate)
  - Check for quality controls (min and max values: Box-plots, Quality control charts)
- Data Construction
  - Create new variables (from continuous/categorical to categorical)
  - Transform variables (from continuous to continuous)
- Data Linkage

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)
  - Check for data types (numeric versus character versus date)
  - Check for missing values (Tabulate frequency)
  - Check for default values (unknown/refused: Tabulate)
  - Check for uniqueness (duplication: subsets by IDs)
  - Check for accuracy (typos in data entry: Tabulate)
  - Check for quality controls (min and max values: Box-plots, Quality control charts)
- Data Construction
  - Create new variables (from continuous/categorical to categorical)
  - Transform variables (from continuous to continuous)
  - Create constructs
- Data Linkage

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)
  - Check for data types (numeric versus character versus date)
  - Check for missing values (Tabulate frequency)
  - Check for default values (unknown/refused: Tabulate)
  - Check for uniqueness (duplication: subsets by IDs)
  - Check for accuracy (typos in data entry: Tabulate)
  - Check for quality controls (min and max values: Box-plots, Quality control charts)
- Data Construction
  - Create new variables (from continuous/categorical to categorical)
  - Transform variables (from continuous to continuous)
  - Create constructs
- Data Linkage
  - Identify linkage type before merging two or more data sets (one-to-one, one-to-many and many-to-many linkage)

- Data Cleaning (Maintain Data Quality)
  - Check variables names (make them informative)
  - Check for data types (numeric versus character versus date)
  - Check for missing values (Tabulate frequency)
  - Check for default values (unknown/refused: Tabulate)
  - Check for uniqueness (duplication: subsets by IDs)
  - Check for accuracy (typos in data entry: Tabulate)
  - Check for quality controls (min and max values: Box-plots, Quality control charts)
- Data Construction
  - Create new variables (from continuous/categorical to categorical)
  - Transform variables (from continuous to continuous)
  - Create constructs
- Data Linkage
  - Identify linkage type before merging two or more data sets (one-to-one, one-to-many and many-to-many linkage)
  - Decide on the linkage method (Exact versus Probabilistic Linkage [SSN, name, address, etc.])

How to deal with Big Data[6]:

- Bigger hardware: Use super-computers if available.

How to deal with Big Data[6]:

- Bigger hardware: Use super-computers if available.
- Use parallel programming if applicable (There are several packages for parallel computation in R such as: Rmpi, nws, snow, sprint, foreach, multicore, and parallel)

How to deal with Big Data[6]:

- Bigger hardware: Use super-computers if available.
- Use parallel programming if applicable (There are several packages for parallel computation in R such as: Rmpi, nws, snow, sprint, foreach, multicore, and parallel)
- Store data on hard disc and analyze it chunkwise: Keep only relevant variables in the dataset you are analyzing (In this way you minimize dynamic memory allocation overhead)

How to deal with Big Data[6]:

- Bigger hardware: Use super-computers if available.
- Use parallel programming if applicable (There are several packages for parallel computation in R such as: Rmpi, nws, snow, sprint, foreach, multicore, and parallel)
- Store data on hard disc and analyze it chunkwise: Keep only relevant variables in the dataset you are analyzing (In this way you minimize dynamic memory allocation overhead)
- Sampling: Work on training sets (A training set is typically a random sample from the complete dataset). Cross-validation is commonly used in statistics and extensively used in machine learning.

How to deal with Big Data[6]:

- Bigger hardware: Use super-computers if available.
- Use parallel programming if applicable (There are several packages for parallel computation in R such as: Rmpi, nws, snow, sprint, foreach, multicore, and parallel)
- Store data on hard disc and analyze it chunkwise: Keep only relevant variables in the dataset you are analyzing (In this way you minimize dynamic memory allocation overhead)
- Sampling: Work on training sets (A training set is typically a random sample from the complete dataset). Cross-validation is commonly used in statistics and extensively used in machine learning.
- Integration of higher performing programming languages like C++ or Java

References

- [1]. Lohr, Sharon (2009). Sampling: design and analysis. Cengage Learning.
- [2]. Foreman, E. K. (1991). Survey sampling principles. CRC Press.
- [3]. Lehana Thabane (2004). Sample Size Determination in Clinical Trials. HRM-733 Class Notes. Department of Clinical Epidemiology & Biostatistics Faculty of Health Sciences, McMaster University, Hamilton ON.
- [4]. Gwenlian Stifin & Aude Espinasse (October 5th, 2011). Data Management seminar. South East Wales Trials Unit, Cardiff University.
- [5]. Elizabeth T. Borer and et al. (2009). Some Simple Guidelines for Effective Data Management. Bulletin of the Ecological Society of America 90:205-214.
- [6]. Oliver Bracht (2013). Five ways to handle Big Data in R. Obtained on November 1st, 2015 from http://www.r-bloggers.com/five-ways-to-handle-big-data-in-r/.

Thank you.

For questions, Email: FQeadan@salud.unm.edu

How to cite this work:

This work was funded by the NIH grants (1U54GM104944-01A1) through the National Institute of General Medical Sciences (NIGMS) under the Institutional Development Award (IDeA) program and the UNM Clinical & Translational Science Center (CTSC) grant (UL1TR001449). Thus, to cite this work please use:

Fares Qeadan (2015). What is statistics and the need for data management. A seminar in biostatistics for the Mountain West Clinical Translational Research Infrastructure Network (grant 1U54GM104944) and UNM Clinical & Translational Science Center (CTSC) (grant UL1TR001449). University of New Mexico Health Sciences Center. Albuquerque, New Mexico.