Week 3

From the Expert: Statistics

The American Statistical Association (ASA) states, "Statistics is the science of learning from data, and of measuring, controlling, and communication uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances." In general, there are two methodologies to analyze the data: descriptive and inferential statistics.

Descriptive statistics focuses on simplifying, summarizing and organizing data. Methods include central tendency measures (mean, median, and mode), variability measures (range, standard deviation), position measures (percentiles, quartiles), tables and graphs (frequency table, bar graph, histogram, boxplot, pie chart, scatter plots, etc.). Descriptive statistics allow you to make conclusions only on the presented data set but not beyond that.

Examples of descriptive statistics are numbers of different car production in the year 2013, average accumulated snow in Colorado for the past 5 years, and top ten states with the highest rate of students graduated from high school in the past year.

Central tendency measures relate to the computed center around the measurement data. Typically, they are among the first statistics to be computed. In a symmetrical normal distribution, the value of mean, median and mode is identical. In an asymmetrical or skewed distribution, these three values can be different. Skewed right distribution has a longer tail stretching toward the right. Whereas, skewed left distribution has a longer tail stretching toward the left. Data with lower bounds are often skewed right and those with upper bounds are often skewed left. The mean is greater than the median for right skewed distribution but the mean is less than the median for left skewed distribution. To summarize the data set in a skewed distribution, it is recommended to report at least 2 metrics: mean and median, but all three metrics: mean, median and mode are preferred. After that, best-fit distributions such as Weibull, gamma, chi-square, lognormal and power lognormal will be determined for skewed right data (NIST).

Variability measures or dispersion measures describe the spread from the center. Two populations with the same mean may have different range dispersion. Therefore, understanding the data dispersion is as important as understanding central tendency.

Inferential statistics attempt to use sample data to make conclusions about the interest population by testing the sample and generalizing the findings to the whole population, in which we cannot access directly. There are two main methods in inferential statistics: estimating the parameter and testing the statistical hypothesis. Practically, statistical inferences heavily rely on making assumptions about the data distribution. If the given data set distribution is relatively close to that of the theoretical probability distribution, the assumptions from the theoretical distribution can be applied to the actual data set

calculations. Therefore, it is very crucial to know the shape of the distribution since it will guide you on what analysis should be performed.

Examples of inferential statistics are the maximum fee that clients are willing to pay for a membership, or the graphical interface design that promotes the click from the customers.

Note that, the term 'parameter' is the term used to describe numbers from a population where as 'statistic' is the term used to describe numbers from a sample.

The usages of inferential statistics are in the form of hypothesis testing, estimation (point, interval estimation), correlations, and regression analysis.

Hypothesis testing is to hypothesize what should be the value of the population parameter. The testing is performed by collecting sample data and compare with the population parameter.

Estimation is the value that best approximate the concerned population parameter. There are two types of estimation: point and interval estimation. Point estimation is a single number of sample statistics while interval estimation is a range of numbers.

Introduction to R

R is an open source software package. It is available for free download from (http://CRAN.R-project.org/) and its mirrors. R is good at graphic plotting, data analysis, and statistical modeling, etc. R is an interpretative and interactive language. Because R is open source software the source code is available for anyone to modify and distribute as far as it conforms to the source license. R has been ported to many platforms such as Mac, Windows, and Linux.

Try these commands and observe the results.
Basic calculations:

```
> 6+7
> 12*5-6
> 2^10
```

The results are not stored. To store the results for later usage, assign them to the objects.

Objects: store the result in the object or variable. Type the object name to see the results.

```
> out1 <- (9+8)^2  # to see the result type out1 on the command line
> out2 <- 7/5
> out3=out1 + out2
```

Functions: code that is made ready to use

```
> sqrt(1225)    # square root
> log(5)        # natural log
> abs(-7)       # absolute value
> exp(5)        # e raised to the power 5
> round(7.527, digit=2)  # the number to round with its decimal places
```

Vectors: sequence of numbers. Numbers can be integer, real, double, or complex numbers. Use mode() to find out if the vector is numeric and use class() to find out what type of numeric.

```
> 1:10
> M <- seq(1,10)
> N=M+5
> K= M*N
> M2 <- seq(1,50, by=5)
> N2=M2*2
> K2 = M2+N2
> Z = c(1,2,3,4,5,6,7,8,9,10) #c is the concatenation function
```

Matrices: two dimensional tables of numbers.

```
>mat1 <- matrix(data=seq(1,20),nrow=4,ncol=5) # type the name of outputs to see the results
>mat1/2
>sqrt(mat1)
```

Data frames (tables with numbers and other kinds of data)

```
> id <- c(1,2,3,4)  #c() is function for concatenation
> sex <- c('F','M','M','F')
> year <- c("freshman", "senior", "junior", "sophermore")
> GPA <- c(3.45,2.9,3.1,3.2)
> student <- data.frame(id,sex,year,GPA)
> student
```

Choosing data: subscript and subset – exclude some data from the analysis

```
> X <- rnorm(10,1,0.1)   #draw 10 random numbers from normal dist. with mean 1, std 0.1
> X[1]              # the first element
> pos <- c(1,3,5)
> X[pos]            # the first, third and fifth element
> femalestudent <- subset(student,sex=='F') # extract only female student
> femalestudent2 <- subset(student,sex=='F',select=c(year,GPA)) #extract year and GPA for female student
```

Basic Statistical Concepts

```
> data <- c(6.6,5.2,4.6,7.8,6.1,9.8,8.2,9.7,7.5,8.7,7.1,10.6,8.2)
> min(data)
> max(data)
> hist(data,col="grey",main="my data") # histogram in grey color using "data"
> boxplot(data,col="grey",xlab="data",ylab="value") #box plot using "data" set
> par(mfrow=c(1,2)) #draw 2 plots (side by side) with the following 2 graphics
> hist(data,col="brown",main="my data")
```

```
> boxplot(data,col="brown",xlab="data",ylab="value")

Measures of central tendency
> mean(data)     #average of the "data" set
> median(data)  # the middle number
> mode(data)     # return the storage mode of object, it is "numeric" in this case
> mde <- density(data)   #use these two commands to find the most frequent number
> mde$x[which(mde$y == max(mde$y))]

Measures of dispersion
> var(data)   # find the variance
> sd(data)    # standard deviation
> IQR(data)  # Interquatile range defined by the first and third quartiles
> summary(data)  # many useful statistics

Combining data sets
> x <- c("AL","AK","AZ","AR")
> y <- c("Alabama","Alaska","Arizona","Arkansas")
> paste(x,y)  #paste is a function to concatenate multiple vectors to a single vector
> paste(x,y, sep="-")  # specify "-" as the separator
> cbind(student,x,y) # combine objects by adding column (student is the object in data
frame exercise)
> student2 <- student
> rbind(student2,student) #combine objects by adding column
```

Visualization

Exploratory data analysis (EDA) is a methodology for analyzing data sets using visual techniques. The goal is to understand the data. Besides summary statistics, plots and graphs are the basic tools for EDA. The systematic methods for going though the data include plotting the distribution of different variables, plotting time series, transforming variables, and looking at the pairwise relationship by scatter plots metrics (O'Neil & Schutt, 2013). EDA was promoted by John Tukey. EDA is the inspiration of the statistical interactive computing S programming language at Bell Labs, which is the predecessor of the two later modern languages: S PLUS and R.

Summary statistics are useful for describing the data set using only few key numbers. However, they can be misleading especially when the overall distribution is ignored. Let's look at the classic and elegant example called Anscombe's Quartet created by a statistician Francis Anscombe (1973). The data consist of 4 datasets. Each dataset consists of eleven pairs of (x,y) points. Each data set has the similar summary statistical values as this following table (http://en.wikipedia.org/wiki/Anscombe's_quartet):

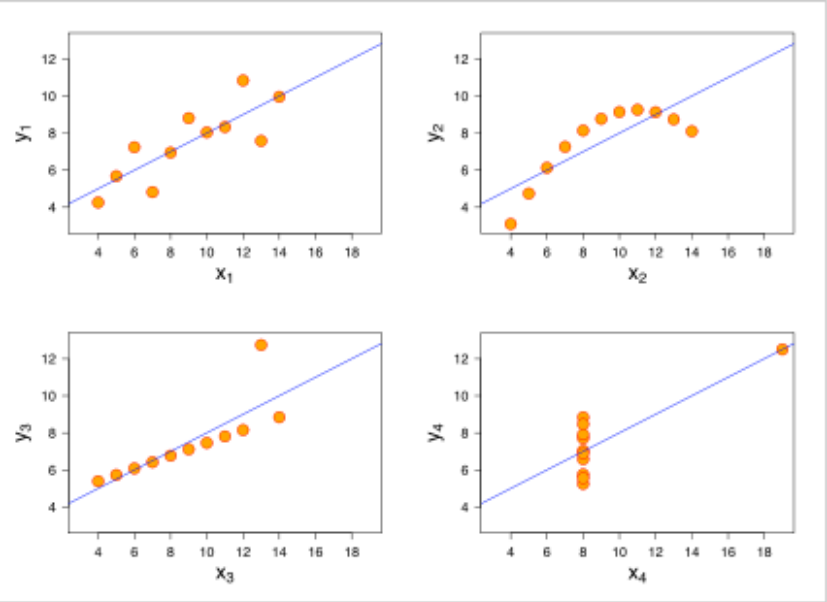
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Source: http://en.wikipedia.org/wiki/Anscombe's_quartet

All 4 datasets contain similar statistical values:
N = 11
Mean of X's = 9.0
Mean of Y's = 7.5
Sample variance of X = 11.0
Equation of regression line: Y = 3+0.5X
Correlation between X and Y = 0.82
Standard error of estimate of slope = 0.118
t = 4.24

However, the graphical displays of the data are clearly different: see http://en.wikipedia.org/wiki/Anscombe's_quartet.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Source: http://en.wikipedia.org/wiki/Anscombe's_quartet

The purpose of the Anscombe's quartet is to illustrate that graphic display is very critical before the data analysis process begins. Graphic displays can reveal different perspectives that may be hidden in the data using only basic statistics summary alone.

Data visualization is considered as a branch of descriptive statistics. Friedman (2008) states that "the main goal of data visualization is to communicate clearly and effectively through graphical means." Unfortunately, the designers often emphasize too much on creating stunning data visualization but fail to convey the ideas. Data visualization is related to statistical graphics, in the field of statistics for visualizing the quantitative data.

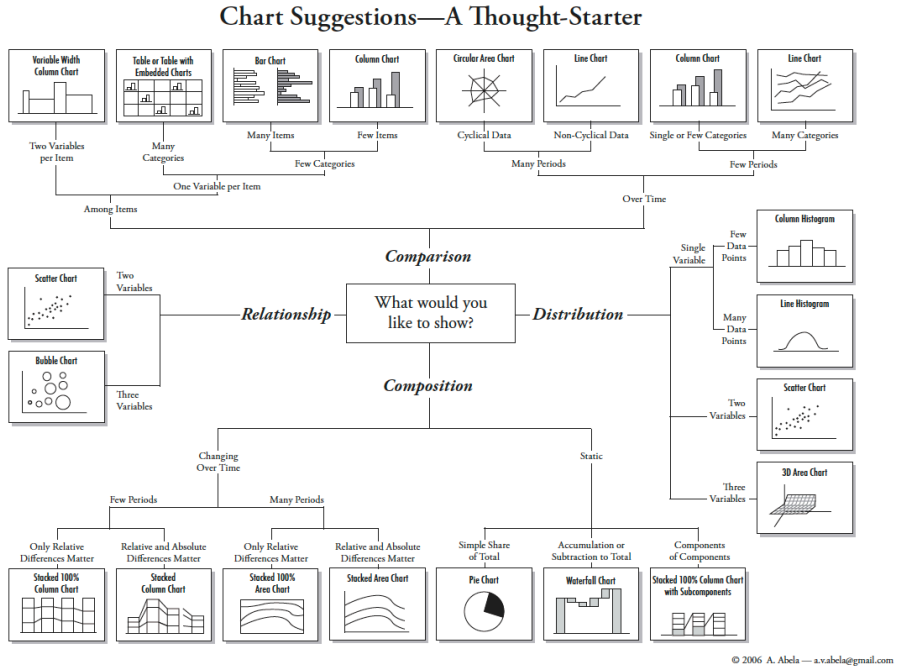
The expert in the field of visual display is Edward Tufte, who wrote the famous book "The Visual Display of Quantitative Information". In this book, numerous examples of best and worst practices in the history of visualization are illustrated. Tufte design principles consist of: http://thedoublethink.com/2009/08/tufte's-principles-for-visualizing-quantitative-information/

- Clear, detailed, label thoroughly, and scale appropriately

- The graphic effect size should be proportional to the numerical quantities. This is known as “Lie Factor”.
    - Lie factor = size of effect shown in the graphic / size of effect in data
    - A lie factor that is greater than 1 means the graph overstates the effect in the data and understates the effect in data, otherwise.
- Show data variation, not design variation.
- Maximize “Data-Ink” ratio. Data ink is the ink on a graph that represents data. For Tufte, good graphical representations should maximize data ink and erase non-data ink as much as possible. Data-ink is calculated from 1 minus the proportion of graph that can be erased without loss of data information.
- avoid “Chartjunk” Tufte coined the term “Chartjunk” to describe the graphical presentation that does not convey information, for example, graphs that demonstrate the designer graphic’s ability rather than display the data. Each field may have different standard for Chartjunk.

The general guideline is using the simplest type of chart to convey the information with the awareness of expectation and standard in the field.

Abela created this fantastic graphic as a guideline for selecting the appropriate chart.



Source: Abela, A. (September 06, 2006). *Choosing a good chart*. The Extreme Presentation ™ Method.

References

Abela, A. (September 06, 2006). *Choosing a good chart*. The Extreme Presentation ™ Method. Retrieved from http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html.

Adler, J. (2010). R in a nutshell. A desktop quick reference. O'Reilly Media. Retrieved from http://shop.oreilly.com/product/9780596801717.do.

American Statistical Association (n.d.). Retrieved from http://www.amstat.org/careers/whatisstatistics.cfm.

Anscombe, F. J. (1973). Graphs in Statistical Analysis. The American Statistician **27** (1): 17–21. JSTOR 2682899. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/00031305.1973.10478966.

Engineering Statistics Handbook. *Histogram Interpretation: Skewed (Non-normal) Right* (n.d.). Retrieved from http://www.itl.nist.gov/div898/handbook/eda/section3/histogr6.htm

Friedman, V. (January 14, 2008). *Data Visualization and infographics*. Smashing Magazine. Retrieved from http://www.smashingmagazine.com/2008/01/14/monday-inspiration-data-visualization-and-infographics/.

Friendly, M., & Denis, D. J. (2001). *Milestones in history of thematic cartography, statistical graphics, and data visualization*. Retrieved from http://www.datavis.ca/milestones/.

Knell, R. J. (n.d.). *Introductory R: A beginner's guide to data visualization, statistical analysis and programming in R*. (Sample chapters) Retrieved from http://www.introductoryr.co.uk/Introductory%20R%20example%20chapters.pdf.

O'Neil, C., & Schutt, R. (2013). *Doing data science*. O'Reilly Media. Retrieved from http://shop.oreilly.com/product/0636920028529.do.

Statistics. Retrieved: July 21,2014 from American Statistical Association: http://en.wikipedia.org/wiki/Statistics

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, CT: Graphics Press. ISBN 0-9613921-4-2. Retrieved from http://www.edwardtufte.com/tufte/books_vdqi.

Tufte, E. R. (2009.08.05). *Tufte's principles*. Retrieved from http://thedoublethink.com/2009/08/tufte%E2%80%99s-principles-for-visualizing-quantitative-information/.

Bogard, Yvonne 9/18/14 11:25 AM
Comment [1]: Emailed paul betty 091814.

Tukey, J. W. (1962). *The Future of data analysis*. The Annals of Mathematics Statistics, Vol. 33, Number 1, pgs. 1-67. Retrieved from http://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711.

Venables, W. N., & Smith, D. M. (2014-07-10). *An Introduction to R: Notes on R- A programming environment for data analysis and graphics.* Version 3.1.1. Retrieved from http://cran.r-project.org/doc/manuals/R-intro.pdf.