Introduction to Statistics

December 5, 2017

GSA

D2D
DATA TO DECISIONS

Course Outline

- What is statistics
- Major definitions
- Exploratory analysis
- Descriptive statistics
- Inferential statistics
- Probability and statistics
- Additional training

GSA

D2D
DATA TO DECISIONS

What is Statistics

- **Webster definition:**
  - "A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."
- **The purpose of statistics is converting the data into useful information**
- **"Statistics is a science for decisions"**
  Unknown statistician
- **Etymology**
  - From Italian *statista* – politician
  - Originally *"analysis of data about the state"*
- **Modern day applications**
  - Census
  - Economics
  - Market analysis
  - Manufacturing (quality control)
  - Scientific research

GSA

D2D
DATA TO DECISIONS

Key Definitions

(Statistics and visualization tools often have different names for same things)

- **Data**
  - ➢ Pieces of information about individuals/samples/records organized into variables
- **Variable**
  - ➢ An element, feature, or factor that vary
  - ➢ Categorical (qualitative) variables
    - Place an individual/object into one of several groups
    - Can be organized into hierarchies, e.g. vertebrae-mammals-bears-grizlies
    - Also called *dimensions* or *attributes*
  - ➢ Quantitative variables
    - Represent some kind of measurement
    - Take numerical values
    - Also called *measures*
- **Dataset**
  - ➢ a set of data identified with particular circumstances

GSA

D2D
DATA TO DECISIONS

Exercise

Please name:

- ***categorical variables***
- ***quantitative variables***

Variables

Individuals	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (1=No, 2=Yes)	Race
Patient #1	M	59	175	69	1	White
Patient #2	F	67	140	62	2	Black
Patient #3	F	73	155	59	1	Asian
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Patient #75	M	48	90	72	1	White

GSA

D2D
DATA TO DECISIONS

Data Levels of Measurements

A variable can have one of four levels of measurement:

- ***Nominal: categories or names only, can be used to classify the data***
- ***Ordinal: can be ordered, e.g. small-medium-big, never-sometimes-always, etc.***
- ***Interval: meaningful differences within the measurement, e.g. temperature, dress sizes***
- ***Ratio: can do math across different ratio measurements, e.g. weight, length, currency***

OK to compute....	Nominal	Ordinal	Interval	Ratio
frequency distribution.	Yes	Yes	Yes	Yes
median and percentiles.	No	Yes	Yes	Yes
add or subtract.	No	No	Yes	Yes
mean, standard deviation, standard error of the mean.	No	No	Yes	Yes
ratio, or coefficient of variation.	No	No	No	Yes

GSA

D2D
DATA TO DECISIONS

Exploratory Data Analysis

- ***Examine distribution of variables***
  - ***What values variables can take***
  - ***How often they take those values***
- ***Establish hierarchies for categorical variables (dimensions)***
- ***Compute descriptive statistics (for numeric variables)***
- ***Visualize with Plots***
  - ***Scatter – general picture***
  - ***Histogram – frequency***

GSA

D2D
DATA TO DECISIONS

Descriptive Statistics

Descriptive statistics quantitatively describe or summarize features of data

- ***Simple summaries of a variable***

Main descriptive statistics:

- ***Distribution***
  - ***Frequency (of occurrences)***
- ***Central Tendency***
  - ***Mean (Average)***
  - ***Median***
  - ***Mode***
- ***Dispersion***
  - ***Range***
  - ***Variance***
  - ***Standard Deviation***

GSA

D2D
DATA TO DECISIONS

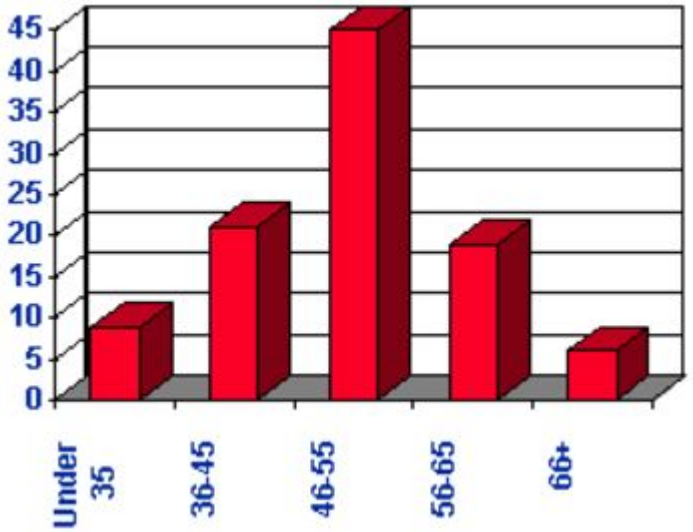
Descriptive Statistics: Distribution

A frequency of individual values or ranges of values for a variable

Can be a table or a graph (histogram)

Example: Percent of income of five age groups

Category	Percent
Under 35	9%
36-45	21
46-55	45
56-65	19
66+	6

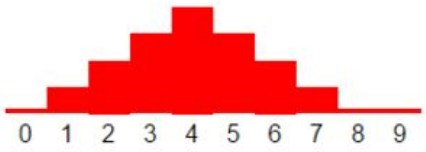


GSA

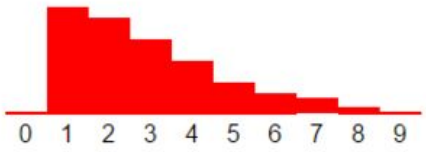
D2D
DATA TO DECISIONS

Table Vs. Histogram

- ***Table***
  - ***Compact***
  - ***Easy to see the values***
  - ***Machine readable***
- ***Histogram***
  - ***Shows patterns***
  - ***Better for human perception***



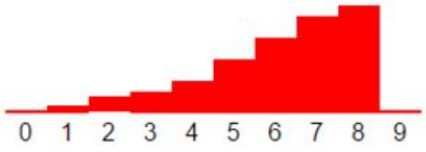
Symmetric, unimodal, bell-shaped



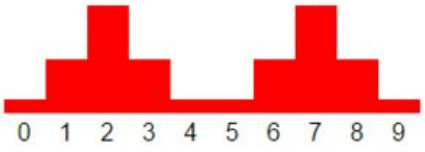
Skewed right

Non-symmetric, bimodal

Uniform



Skewed left



Symmetric, bimodal

GSA

D2D
DATA TO DECISIONS

Descriptive Statistics: Central Tendency

Central Tendency - one number to describe the variable

Three ways to estimate

- ***Mean (bias, average)***

  A = [15, 20, 21, 20, 36, 15, 25, 15]

  Mean(A) = sum(A)/count(A) = 167/8 = 20.875

- ***Median***

  A = [15,15,15,20,20,21,25,36]

  Median(A) = 20 (the mid point of A)

- ***Mode***

  The most frequently occurring number

  Mode(A) = 15

  Can be applied to both quantitative and categorical data

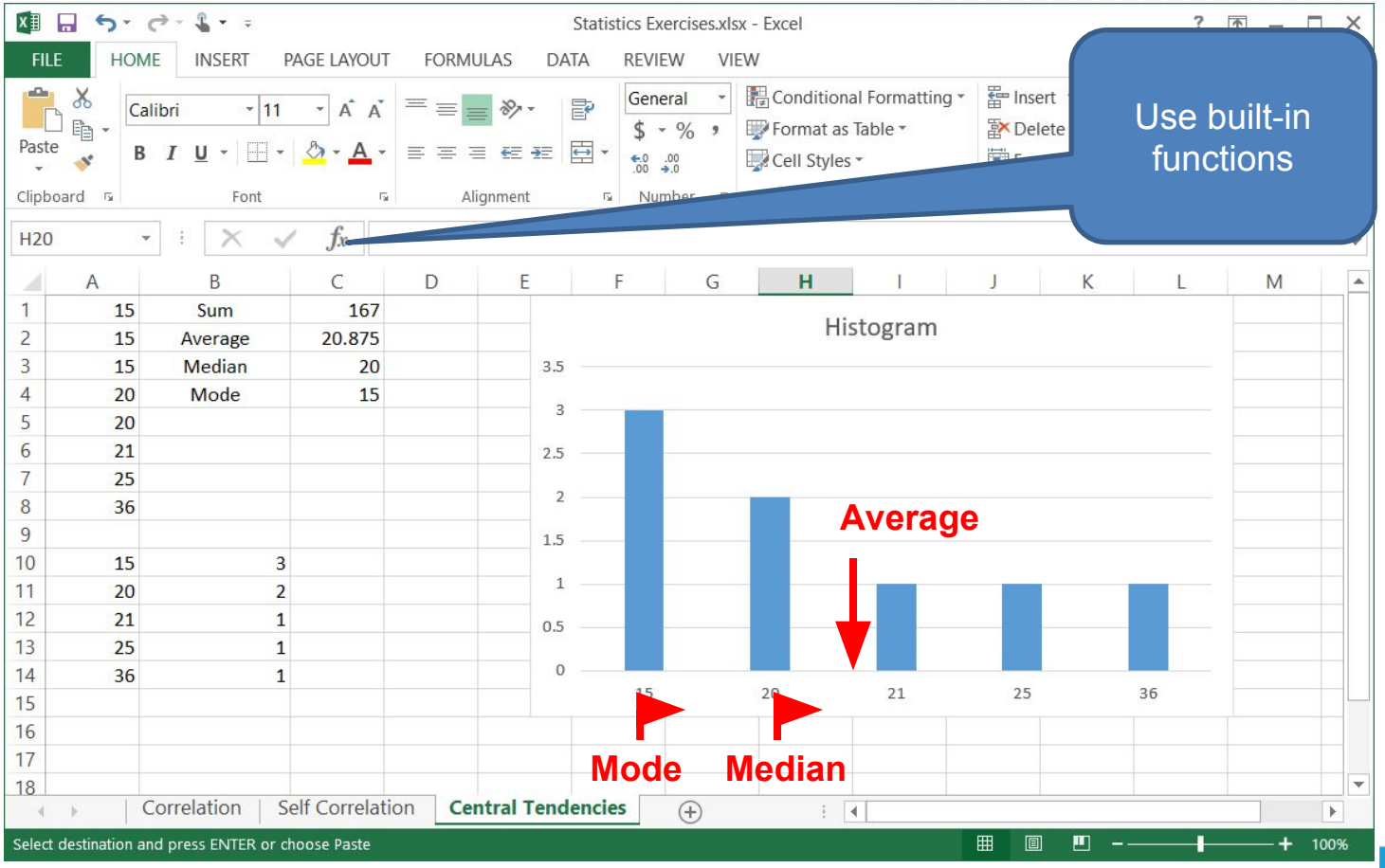
- ***Note: for normal distribution Mean = Median = Mode***

GSA

D2D
DATA TO DECISIONS

Exercise

Compute Mean, Median, and Mode in Excel



How to create histogram in Excel
https://www.ablebits.com/office-addins-blog/2016/05/11/make-histogram-excel/

Descriptive Statistics: Dispersion

Three common ways to estimate

- ***Range - simply max - min***

  A = [15, 20, 21, 20, 36, 15, 25, 15]

  Range(A) = 36 – 15 = 21

Note: a big outlier can greatly influence a result!

- ***Variance***
  - ➢ ***Show the relation of each value to the mean (variance around average)***

$$\frac{\Sigma(X - \bar{X})^2}{(n - 1)}$$

- ***Standard Deviation***
  - ➢ ***A square root of variance***

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$$

X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

GSA

D2D
DATA TO DECISIONS

Why Do We Square Differences in Variance and Standard Deviation?

Imaging a data set

A = [4, 4, 0, -4, -4] Average(A) = 0

If we would not square differences, the negatives will compensate the positives and we would not get a true perception of the dispersion, i.e.:

(4 – 0) + (4 – 0) + 0 + (-4 – 0) + (-4 – 0) = 0

Can we take absolutes? Yes, this will be called Mean Deviation

$$\text{Mean Deviation} = \frac{\Sigma|x - \mu|}{N}$$

However, Standard Deviation has proven to work better on most of datasets and is used way more often

Exercise: compute Mean and Standard Deviations for

A = [7, 1, -6, -2]

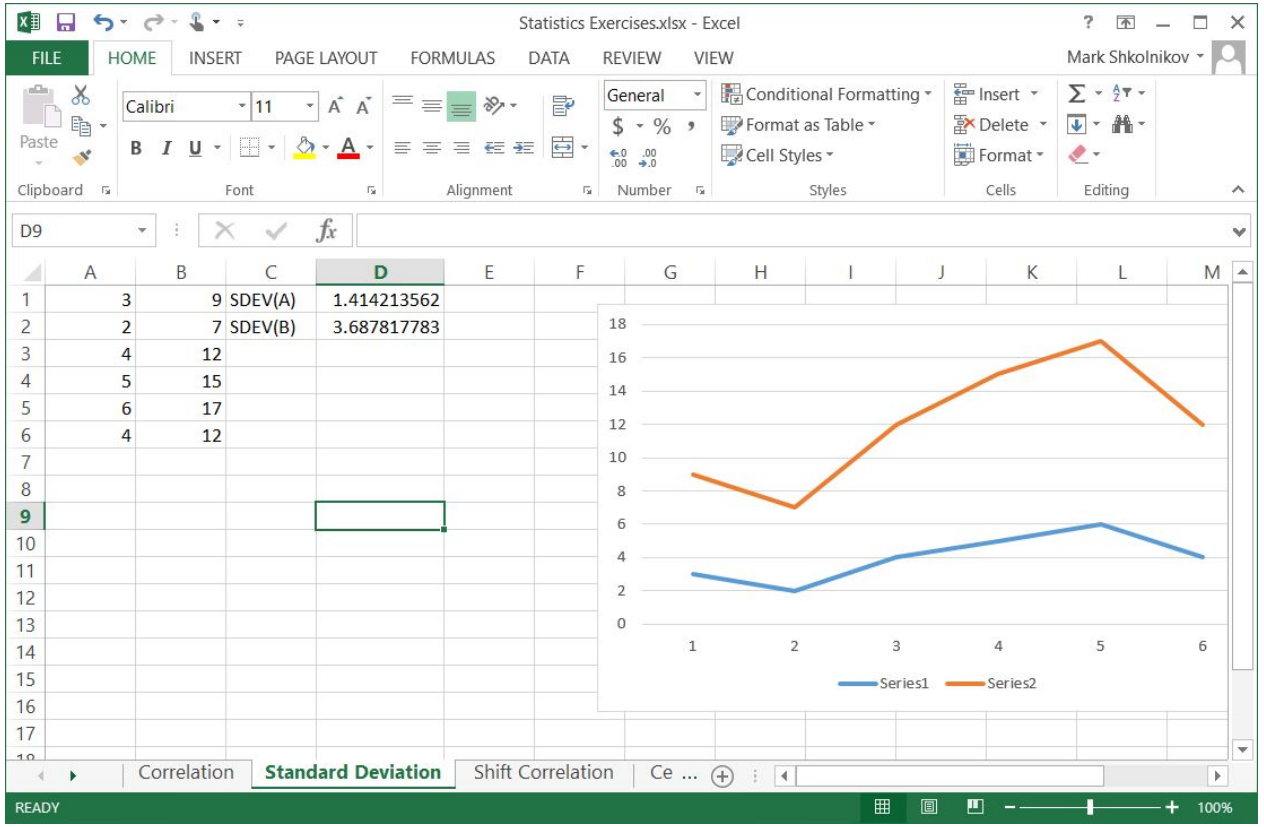
GSA

D2D DATA TO DECISIONS

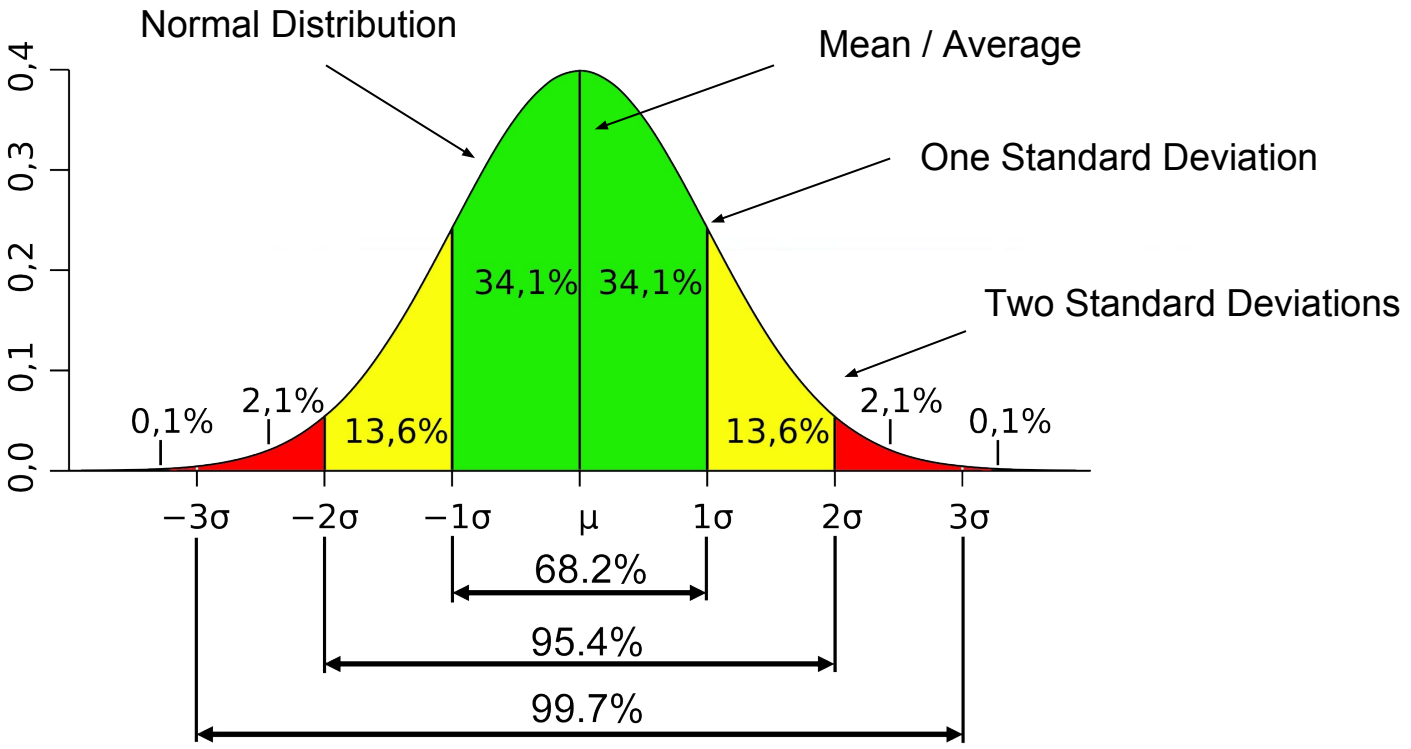
Exercise

Compute Standard Deviation

- ***For A***
- ***For B***



Normal Distribution and Standard Deviation

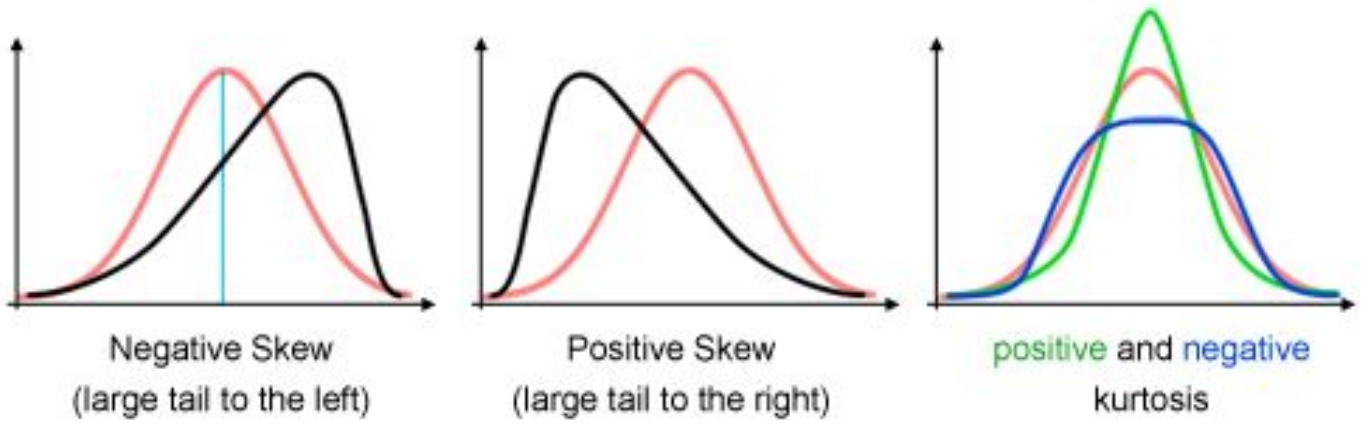


Skew(ness) and Kurtosis

***Skew** is a measure of "lack of symmetry" – the bigger is the skew the more asymmetric is the distribution*

***Kurtosis** is a measure of*

- *"Peakedness" (Wolfram MathWorld)*
- *Combined weight of the "tails" relative to the rest of the distribution (Dr. Donald Wheeler)*



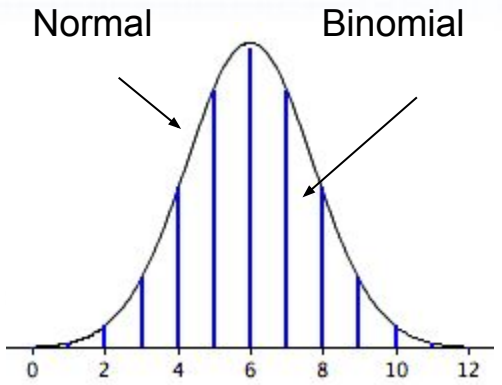
Normal distribution ***Skew = 0, Kurtosis = 3***

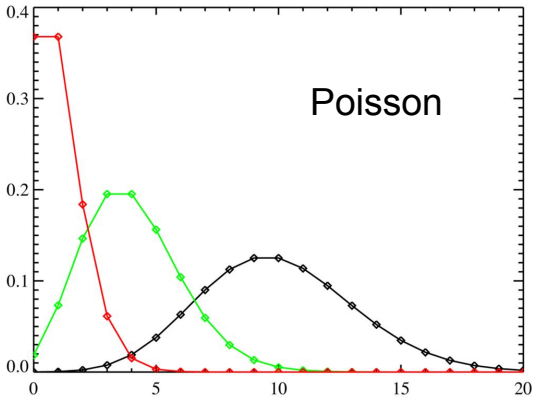
GSA

D2D
DATA TO DECISIONS

Normal, Binomial and Poisson Distributions

- ***Normal distribution*** ***(Gaussian) is a probability distribution of a (truly) random continuous variable***
- ***Binomial distribution*** ***is a probability distribution of a discrete random variable (have mutually exclusive outcomes), e.g. coin flipping***
  - ***For a large number of evens approaches normal distribution curve***
- ***Poisson distribution*** ***describes probability of a certain number of events within a certain period of time, e.g. cars passing traffic light, radioactive decay***
  - ***For large number of events approaches normal***





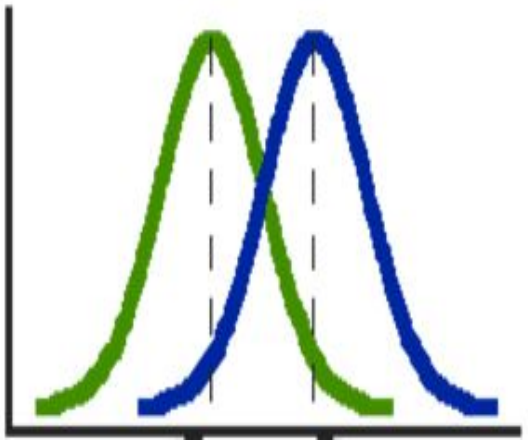
GSA

D2D
DATA TO DECISIONS

Inferential Statistics

Reaching conclusions that extend beyond the immediate data

- ***Identify trends***
- ***Establish relations between independent and dependent variables***
- ***Often involves comparison of***
  - ***Variables***
  - ***A variable to a model or pattern***
  - ***Can be qualitatively accessed by plotting***
  - ***Can be quantitatively accessed by***
    - ***Covariance and Correlation (compare patterns)***
    - ***T-test (compare means)***
    - ***ANOVA (compare variances)***



GSA

D2D
DATA TO DECISIONS

Covariance and Correlation

Both describe how two variables deviate from expected values (means/averages)

- ***Covariance is average of the products of deviations of each variable from its mean***
- ***Correlation (coefficient) is "normalized" dimensionless covariance, i.e.***

Correlation(A,B) = Covariance(A,B) / ((Standard Deviation(A))(Standard Deviation(B)))

Example

A = [3, 2, 4, 5, 6] Average(A) = 4

B = [9, 7, 12, 15, 17] Average(B) = 12

Covariance(A,B) =

= ((3 – 4)(9 – 12) + (2 – 4)(7 – 12) + (4 – 4)(12 – 12) + (5 – 4)(15 – 12) + (6 – 4)(17 – 12))/5 =

= 5.2

Exercise: use Excel to compute covariance and correlation

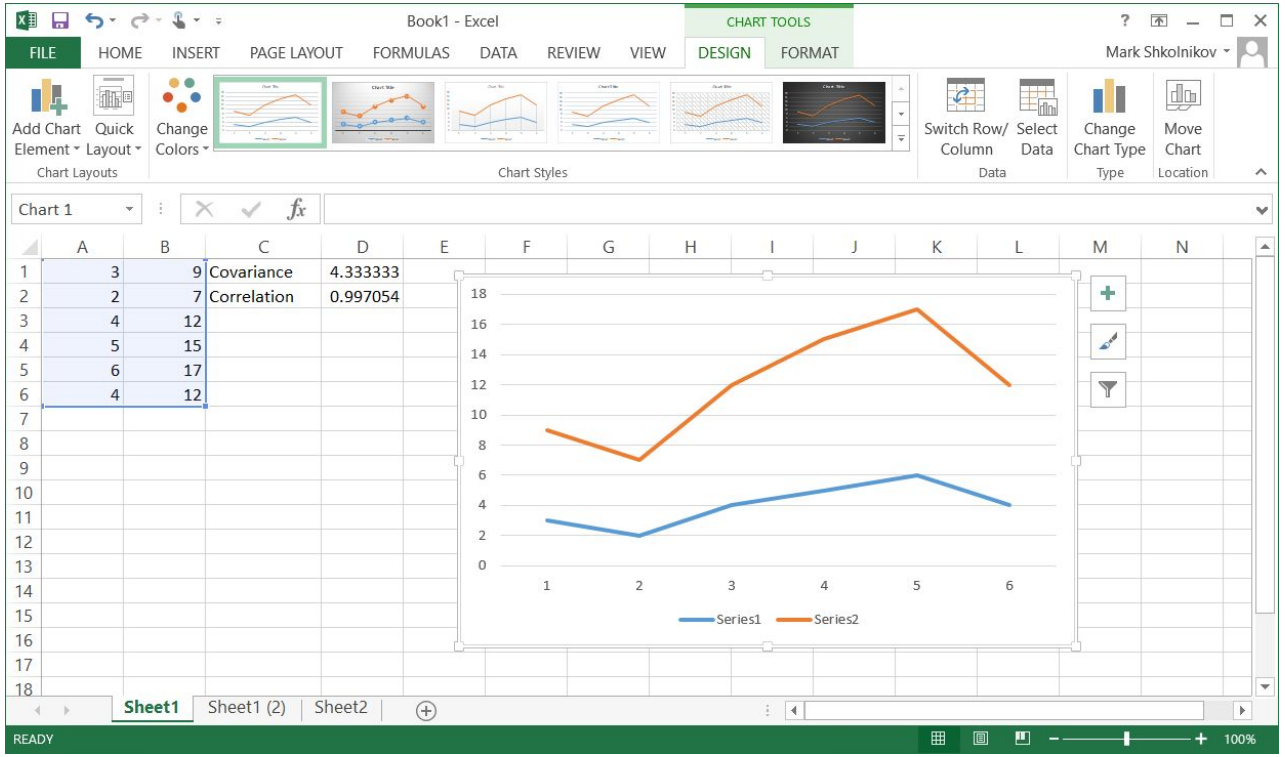
GSA

D2D
DATA TO DECISIONS

Exercise

Compute Covariance and Correlation Coefficient

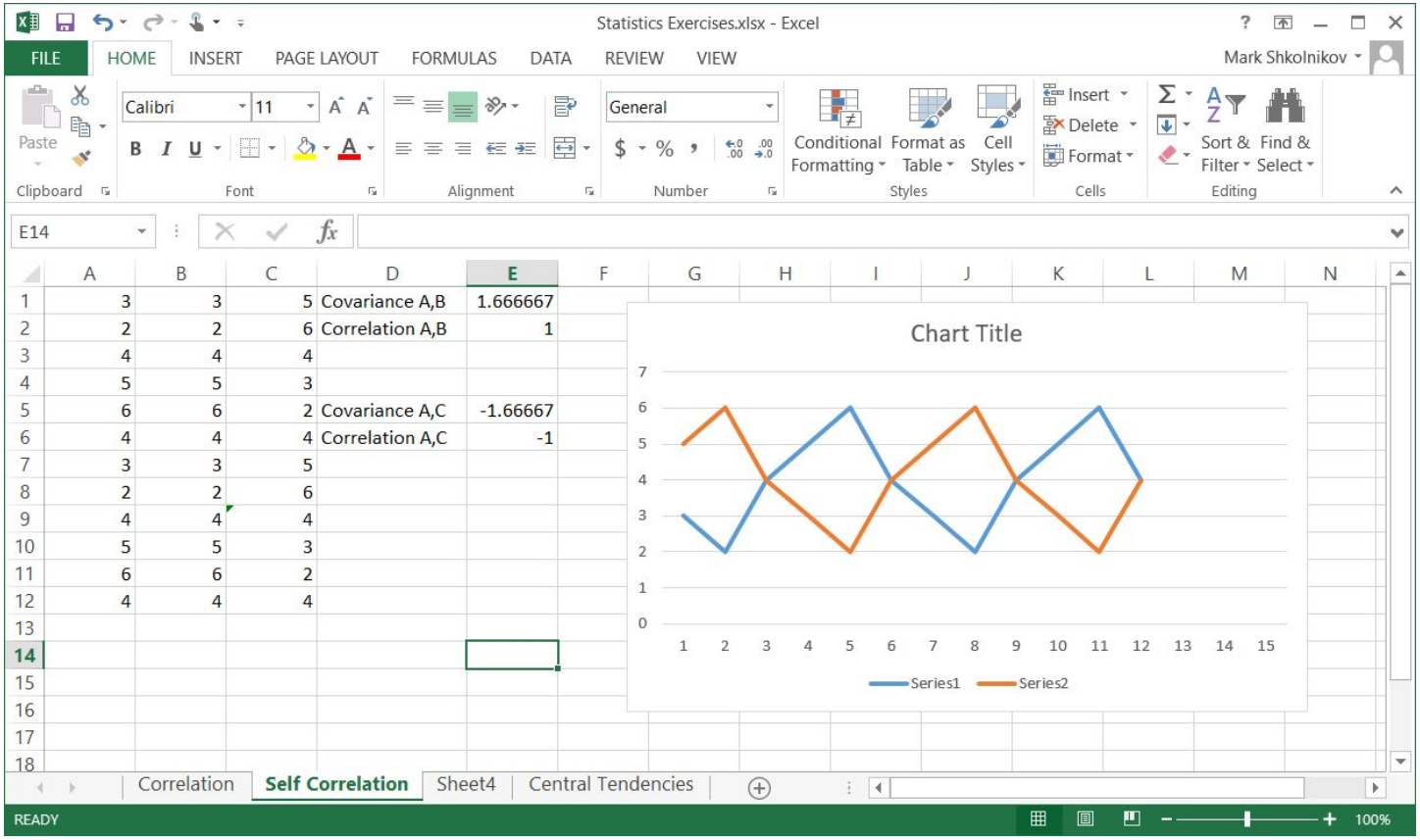
- ***For A and B***
- ***For A and A***



Exercise (cont.)

Compute Covariance and Correlation Coefficient

- ***For A and C, where C is shifted A***



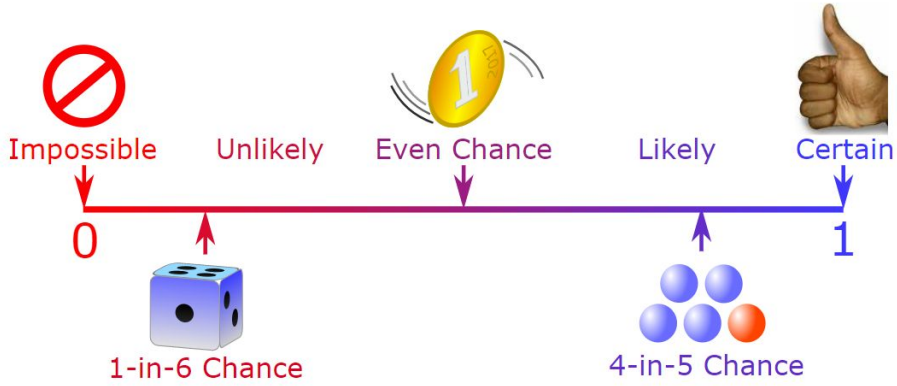
	A	B	C	D	E
1	3	3	5	Covariance A,B	1.666667
2	2	2	6	Correlation A,B	1
3	4	4	4		
4	5	5	3		
5	6	6	2	Covariance A,C	-1.66667
6	4	4	4	Correlation A,C	-1
7	3	3	5		
8	2	2	6		
9	4	4	4		
10	5	5	3		
11	6	6	2		
12	4	4	4		
13					
14					
15					
16					
17					
18					

GSA

D2D
DATA TO DECISIONS

What is Probability?

The extent to which an event is likely to occur, measured by the ratio of the favorable cases to the whole number of cases possible



P(A or B) = P(A) + P(B) ***P(A & B) = P(A) * P(B)***

Example:

Fair coin: two equal possibilities P(Head) = P(Tail) = 1/2

Probability of Head or Tail = 1/2 + 1/2 = 1

Probability of Head after Tail = 1/2 * 1/2 = 1/4

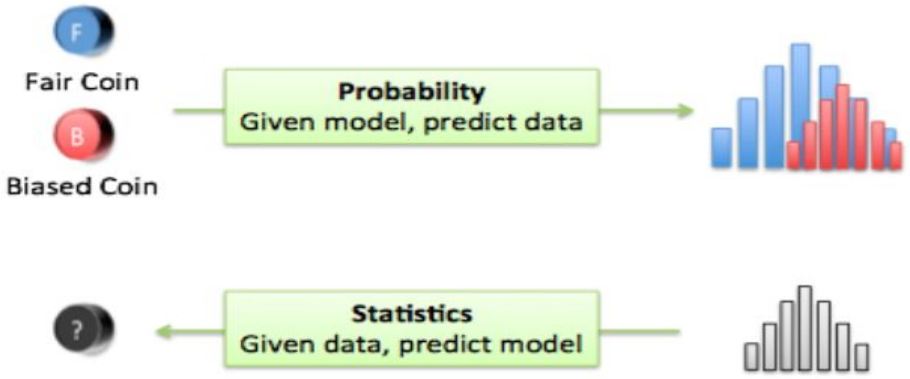
GSA

D2D
DATA TO DECISIONS

Probability and Statistics
Where is the connection?

A (very) facilitated explanation:

- ***Probability: from model to data distribution (create/analyze a model and predict an outcome)***
- ***Statistics: from data distribution to model (what effects the outcome)***
- ***Connection – math behind models and data distribution***



GSA

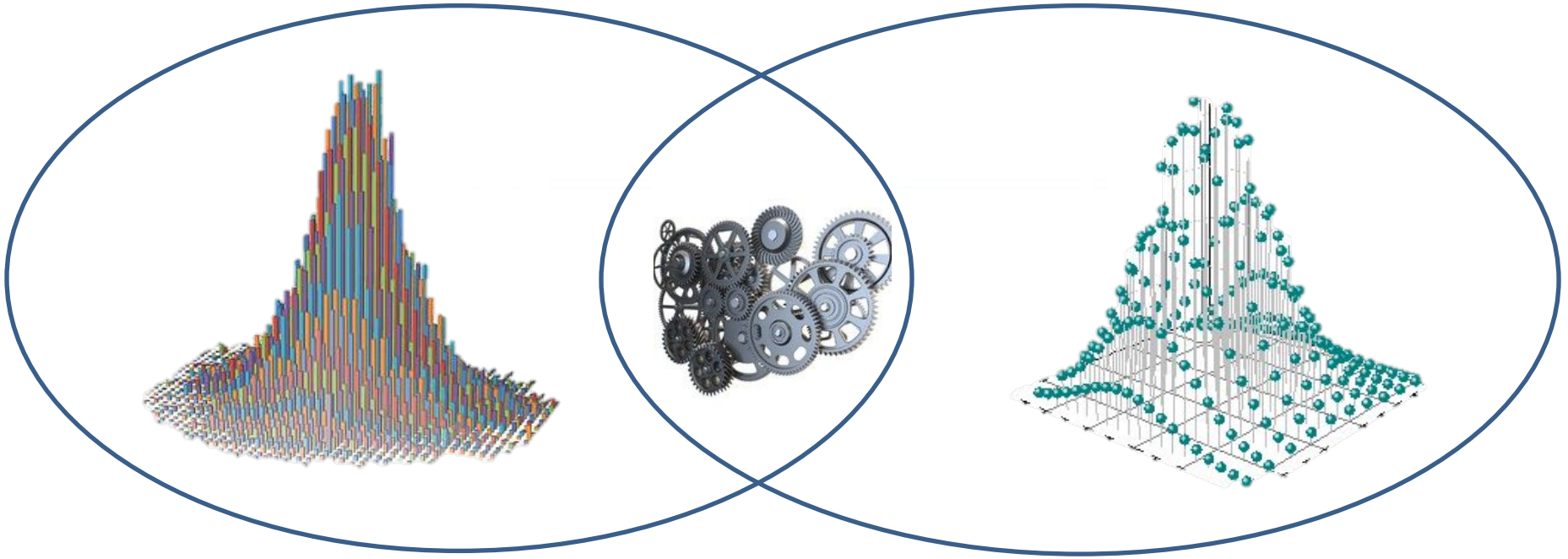
D2D
DATA TO DECISIONS

Probability and Statistics Together

Predictive Analytics

Statistics

Probability



Analyze Data

Create Model

Predict behavior under various conditions

GSA

D2D
DATA TO DECISIONS

Regression Analysis

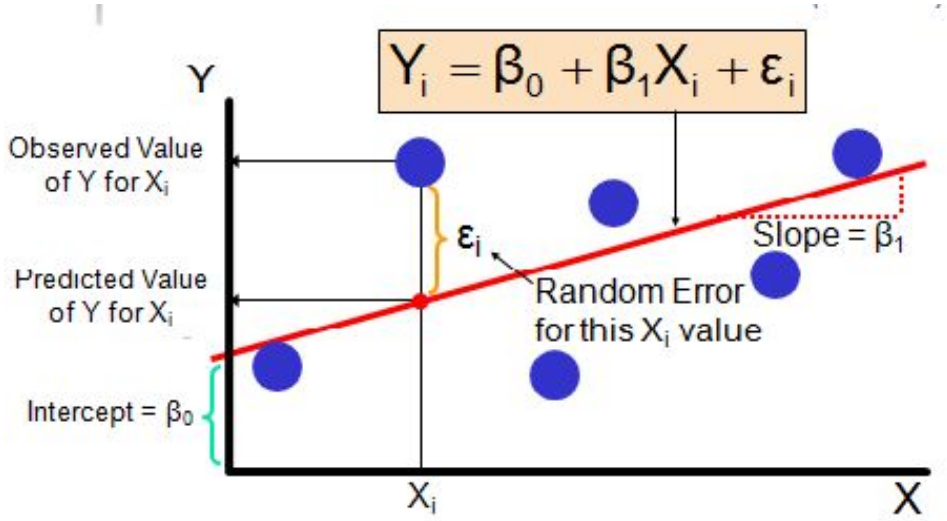
Regression is a method for fitting a curve through a set of points using some goodness-of-fit criterion

Linear regression is the most common type of regression

Least Squares Fitting is the simplest mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets ("the residuals") of the points from the curve

Regression analysis is a set of statistical processes for estimating the relationships among variables

Regression analysis is used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships



T-test

- **T-test (A.K.A. Student's t test)** is a **statistical test** which is widely used to compare the mean of two groups of samples, i.e.: to evaluate whether the means of the two sets of data are statistically **significantly** different from each other.
- T-tests are handy hypothesis tests in statistics when you want to compare means. You can compare a sample mean to a hypothesized or target value using a **one-sample t-test**. You can compare the means of two groups with a **two-sample t-test**. If you have two groups with paired observations (e.g., before and after measurements), use the paired t-test.
- The **one-sample t-test**, used to compare the mean of a population with a theoretical value.

Let x represent a set of values with size n, with mean m and with standard deviation s. The comparison of the observed mean m of the population to a theoretical value μ is performed with the formula below:

$$t = (m-\mu)/(s/\sqrt{(n)})$$

GSA

D2D
DATA TO DECISIONS

ANOVA - Analysis of Variance

A.K.A Fisher Analysis of Variance

- **ANOVA is a statistical hypothesis test: how a dataset is different from another one (or a model)**
  - ➢ **One / Two way ANOVA (for single/two factor analysis)**
  - ➢ **MANOVA / N-way ANOVA – Multivariate ANOVA (for multifactor analysis)**
- **Rather involved computation**
- **Results are typically displayed as an ANOVA Table**

Source	SS	DF	MS	F
Treatments	SST	k−1	SST/(k−1)	MST/MSE
Error	SSE	N−k	SSE/(N−k)	

SS – sum of squares
DF – degrees of freedom, DF = k – 1, where k is number of groups of samples (“treatments”)
MS – mean squares, MS = SS/DF
T – treatment, A.K.A “between”
E – error, A.K.A “within”
N – number of samples in the groups

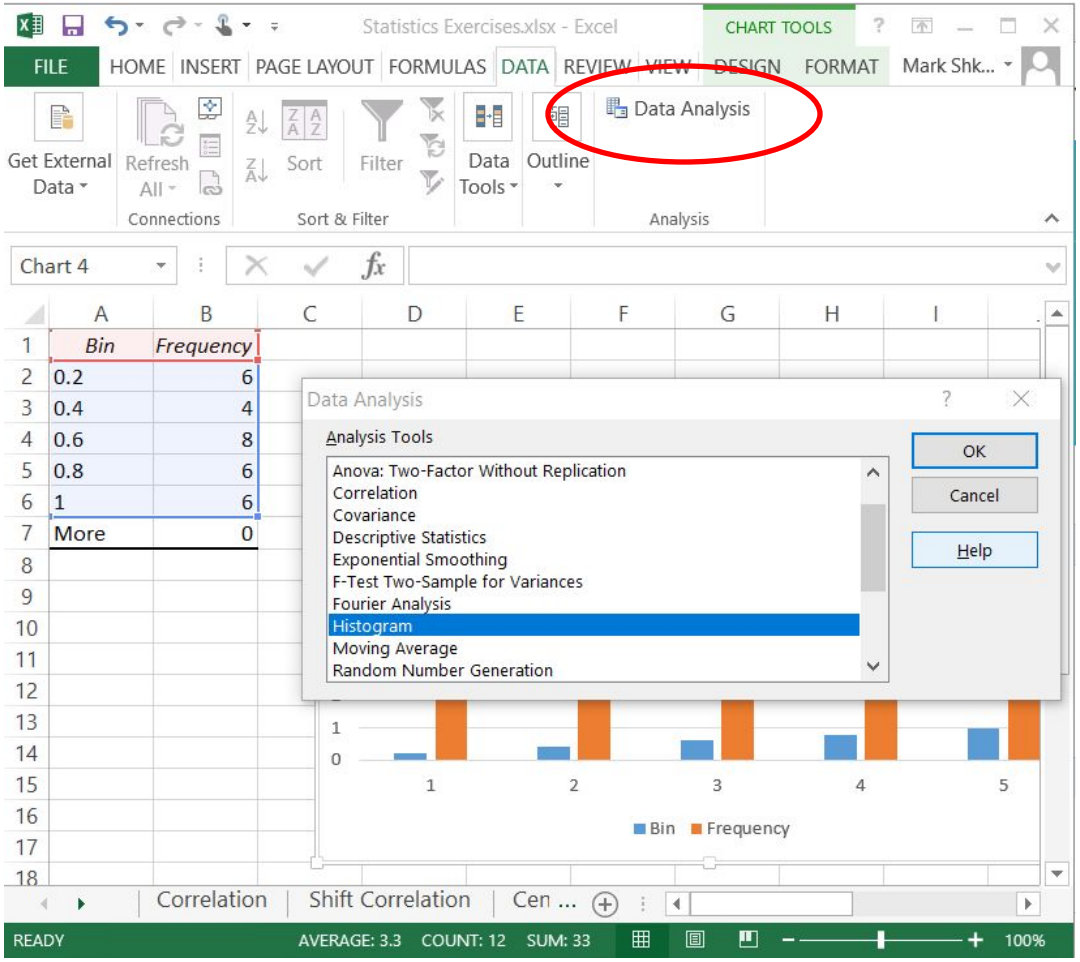
GSA

D2D
DATA TO DECISIONS

Data Analysis in Excel

Enable Data Analysis ToolPak

- ***File => Options => Add-Ins => Analysis ToolPak => Go***
- ***Check box "Analysis ToolPak"***
- ***OK***



Useful Links

- http://www.statisticshowto.com/statistics-basics/
- https://www.socialresearchmethods.net/kb/index.php
- A semester-long course
- http://online.stanford.edu/course/probability-and-statistics-self-paced
- R tutorials
- http://www.statmethods.net/index.html
- http://www.cengage.com/resource_uploads/downloads/1305115341_450336.pdf

GSA

D2D
DATA TO DECISIONS

Q & A