Statistics –Your Friend, Not Your Foe

Presented by:

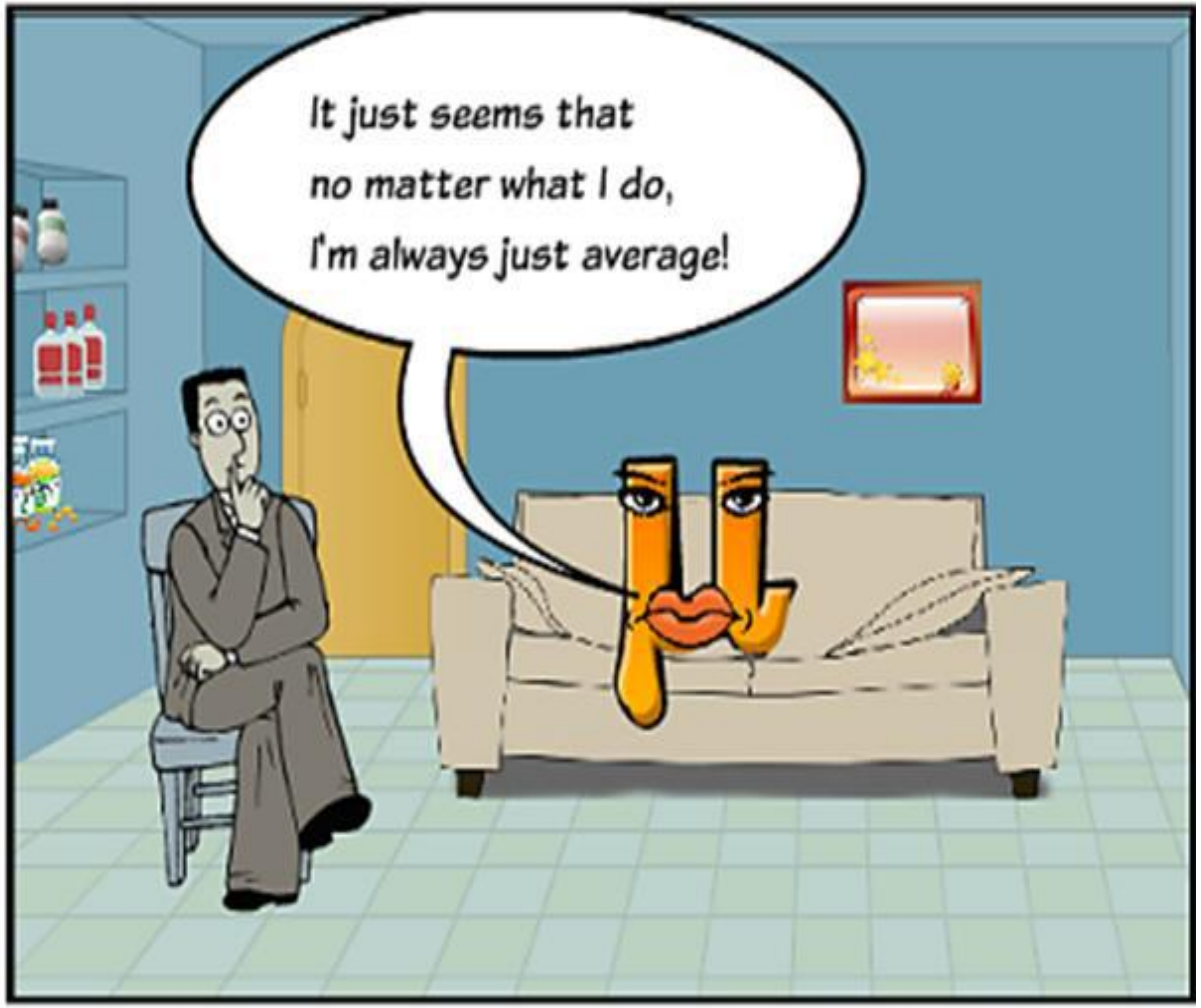
Paulo Macedo, PhD, Senior Statistician

Sewit Araia, MPH, Sr. Program Manager and Statistician

March 01, 2017

Outline

- **Introduction**
- **Statistics – Friend or Foe?**
- **The Role of Data Analytics Today**
- **Major Types of Analytics**
- **Understanding Your Data**
- **Outlier Detection Techniques/Statistical Tools**
    - Descriptive Statistics
    - Ranking and Percentile
    - Z - score
    - Box-Plot
    - Cluster Analysis
    - Predictive Modeling
- **Sampling and Extrapolation**


It just seems that no matter what I do, I'm always just average!

Statistics – Friend or Foe?

“In God we trust, all others bring data.”
- Attributed to William Edwards Deming, Statistician (1900-1993)

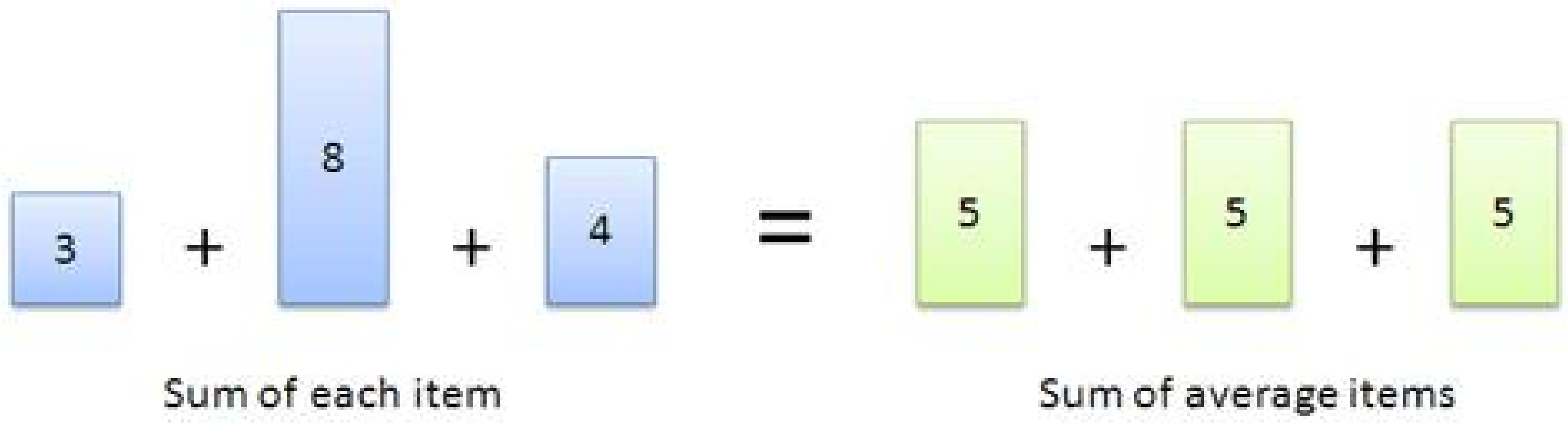
“We are drowning in information and starving for knowledge.”
- Rutherford D. Rogers, Librarian

Statistics – Friend or Foe?

- **As a first impression it looks like Statistics is a Foe:**
  - Complex subject made worse by obscure terminology.
  - Statistics is associated with steep learning curves.

- **But it actually can be a Friend:**
  - Many statistical concepts have intuitive meanings, for example:
    - The average (mean) is a number that summarizes the data in a single value.
    - Other statistical summary numbers can be used to interpret large amounts of data helping to focus decision-making processes

Mean



The Role of Data Analytics Today

The Role of Data Analytics Today

- **Data Analytics methods are commonly associated with:**
  - Statistics
  - Machine Learning
  - Data Mining

- **The methods used in the three areas are very similar — fundamentally they are the same**
  - They use the same material and almost exactly the same techniques

- **However, they have slightly different perspective due to their distinct historical development**
  - **Statistics**
    - The emphasis is on formal statistical inference (confidence intervals, hypothesis tests, optimal estimators)
    - The emphasis is also on testing models and assumptions.
  - **Machine Learning**
    - The emphasis is on making accurate predictions
    - In particular, on building software systems that make predictions
  - **Data Mining**
    - The emphasis is on valuable insights (patterns) in large databases

The Role of Data Analytics Today

- **"Drowning in information"**
  - There is an increase in data collection in both private and government sectors

- **In the Healthcare Industry this is characterized by a movement towards collecting large amount of data:**
  - Electronic health records
  - Payer claims
  - Pharmacy data
  - Laboratory test results
  - Patient registries
  - Quality Measures Data

- These developments require the use of effective analytical tools to provide oversight of health insurance transactions for compliance checking and fraud detection making smart use of limited audit resources

The Role of Data Analytics Today

- Data is collected and validated — hopefully not garbage …. Then what is next?
- Turn piles of data into actionable insights using the proper analytical tools
  - Non compliant providers can be detected → Cost saving to the program
  - Intervention programs can be developed → Mitigating program issues
  - Edits can be implemented → Continuous monitoring
  - Policies can be updated → More effective regulations

"*Big Data* is not about the data. Data is easily obtainable and cheap, and more so every day. The analytics that turn piles of numbers into actionable insights is difficult, and more sophisticated every day."

— Gary King

Major Types of Analytics

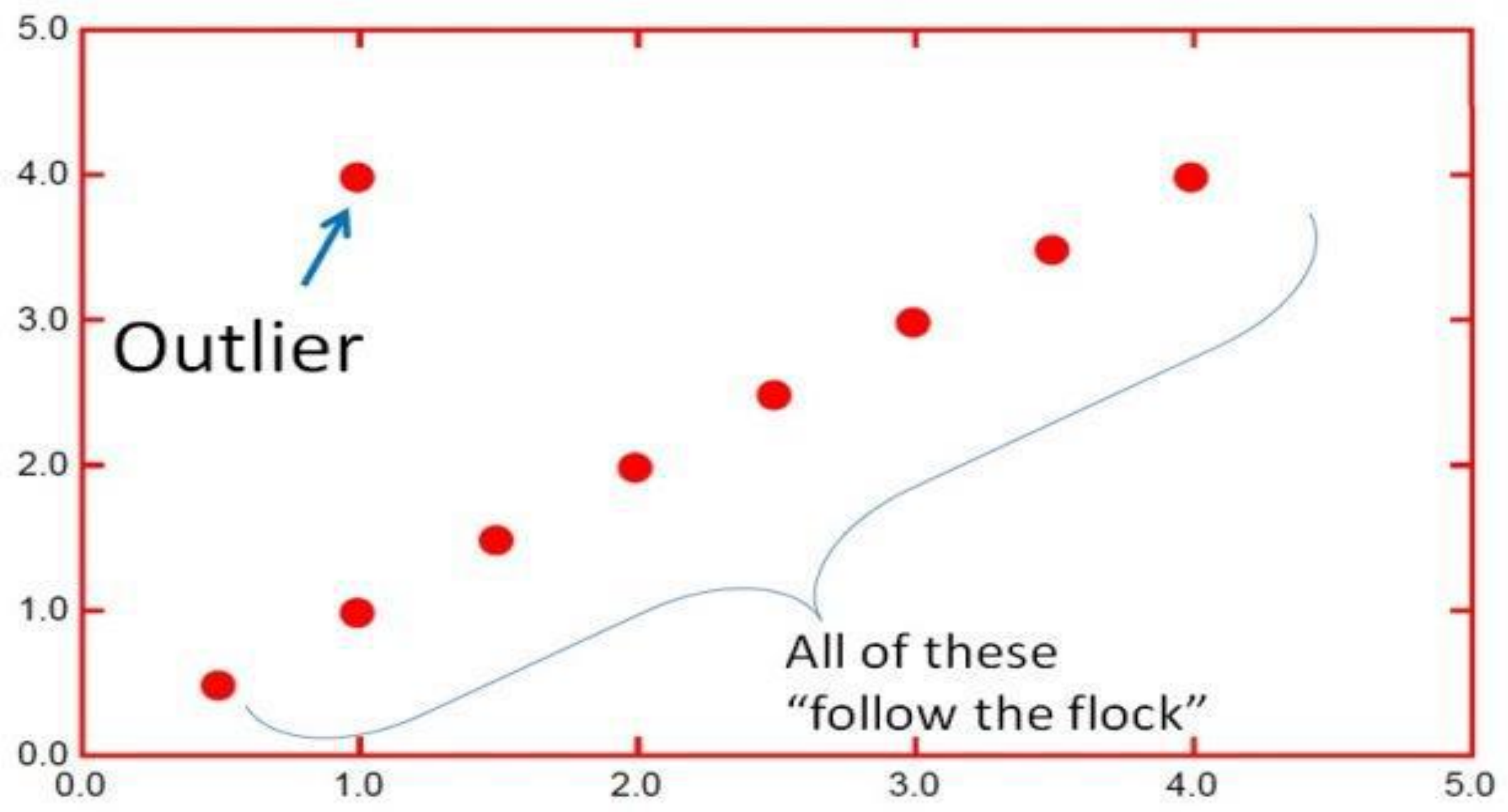
Major Type of Analytics

Machine Learning		Statistics
Network, graphs	⟷	Model
Weights	⟷	Parameters
Learning	⟷	Fitting
Supervised learning	⟷	Regression/classification
Unsupervised learning	⟷	Clustering, density estimation
Generalization	⟷	Test set performance

Major Type of Analytics

Unsupervised Learning Methods (Non-Structured Analysis)

- No prior information required
- Outlier detection
  - Classifying data in two subsets,
  - Outlier and within-the-norm providers



Some methods used in fraud detection

- **Time Series Analysis**
  - Trend analysis
  - Spike analysis
- **Cluster Analysis**
  - Based on key similarities within the groups
  - Used to identify sub-specialties among providers according to their billing pattern
- **Link Analysis**
  - Identifying connections between providers

Major Type of Analytics

- **Supervised Learning Methods (Structured Analysis)**
  - Require prior information – at least on a number of outcomes
    - A frequent outcome is “Yes” or “No”, for example, providers could be “Excluded” or “Non-excluded (Active)”
    - The goal is to find the probability that Non-excluded providers will be excluded from the healthcare network based on their billing pattern similarity with the excluded providers
  - **Some methods used in fraud detection:**
    - Logistic regression
    - Decision trees
    - Neural network

Major Type of Analytics

Banking

- Supervised Learning
- Predict credit worthiness of credit card holders:
  - Build a machine learning model to look for delinquency attributes by providing it with data on delinquent and non-delinquent customers
- Unsupervised Learning
- Segments customers by behavioral characteristics:
  - Survey prospects and customers to develop multiple segments using clustering

Healthcare

- Supervised Learning
- Predict patient readmission rates:
  - Build a regression model by providing data on the patients' treatment regime and readmissions to show variables that best correlate with readmissions
- Unsupervised Learning
- Categorize MRI data by normal or abnormal results:
  - Use cluster analysis to group the results into two – within the norm and out of the norm

Understanding Your Data

Understanding Data

More time is spent on understanding the data than conducting the statistical analysis

- Conduct descriptive analysis
- Conduct research on external sources, regulatory analysis/policy analysis related to issues
- Understand the potential outcome — but remember the data may surprise you

Data understanding is our Friend

- This is intuitive, we do this everyday - understanding the data and how the data is generated

Your analysis is only as good as your data.

Understanding Data

Public Data

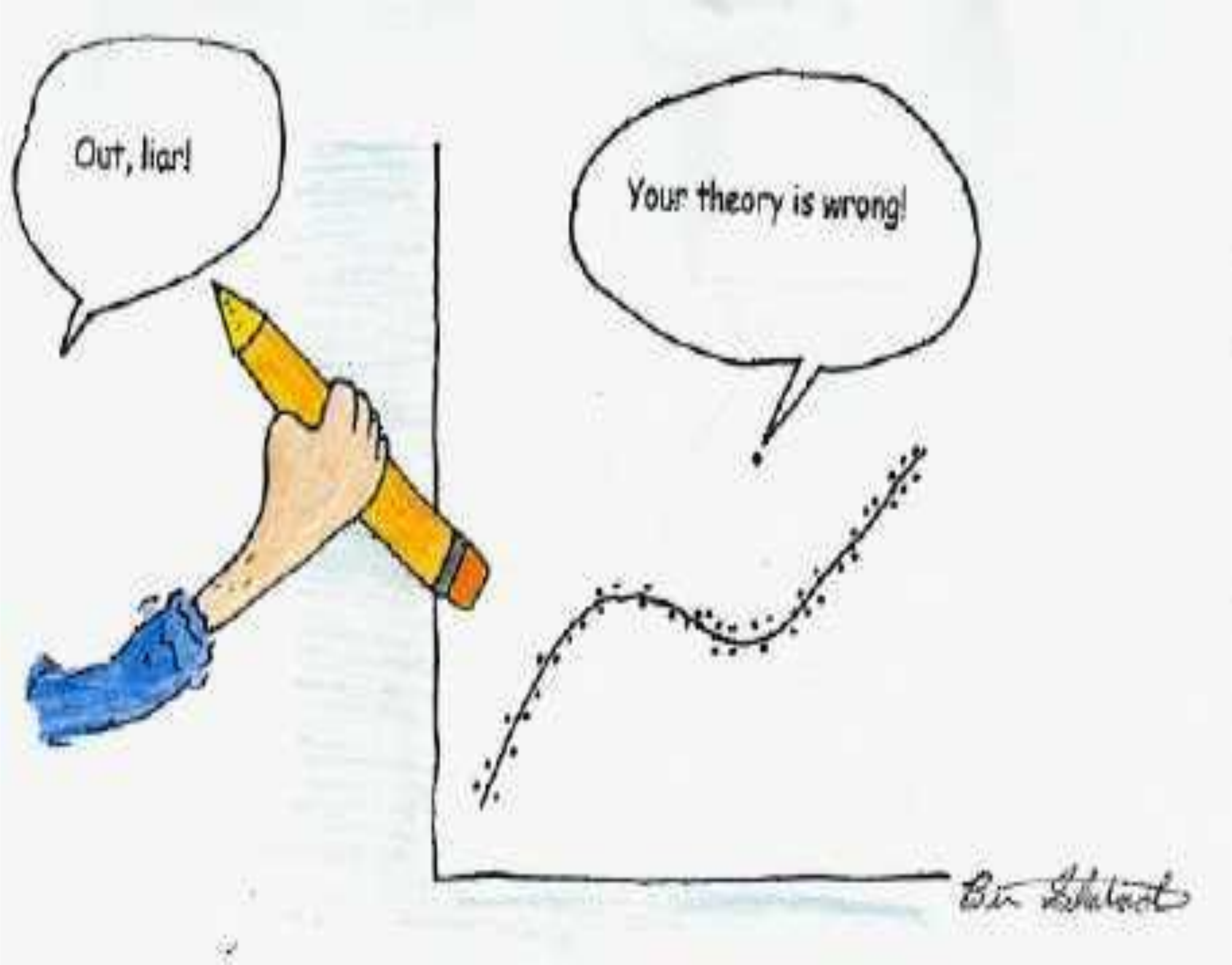
Medicare Provider Utilization and Payment Data: Physician and Other Supplier Public Use File (Physician and Other Supplier PUF)

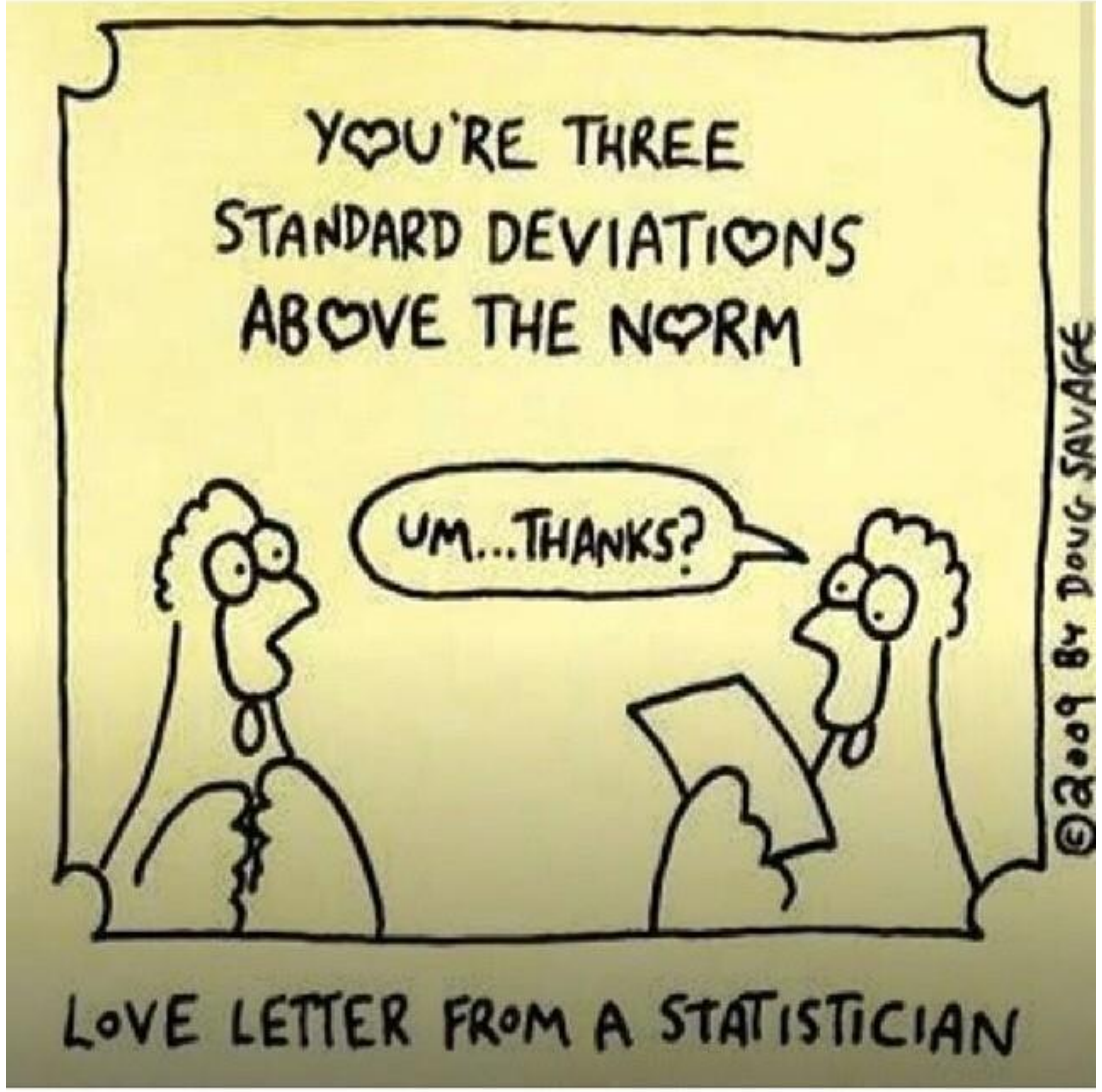
- Published by Centers for Medicare & Medicaid Services
- Data is available from 2013 — 2014
- Data includes – Procedure codes, Provider identifier , Provider demographic information, reimbursement amount per procedure code
- The data is being used to illustrate the various methods in this workshop

List of Excluded Individuals and Entities (LEIE)

- Published by the Office of Inspector General of the U.S. Department of Health & Human Services (HHS)
- Includes a List of Individuals and Entities excluded from Federal funded health care programs
- The data is being used to illustrate predictive modeling method

Outlier Detection Techniques/Statistical Tools


Out, liar!
Your theory is wrong!


YOU'RE THREE STANDARD DEVIATIONS ABOVE THE NORM
UM...THANKS?
©2009 BY DOUG SAVAGE
LOVE LETTER FROM A STATISTICIAN

Outlier Detection Techniques/Statistical Tools

Raw Data

This data will be use to illustrate the various methods:

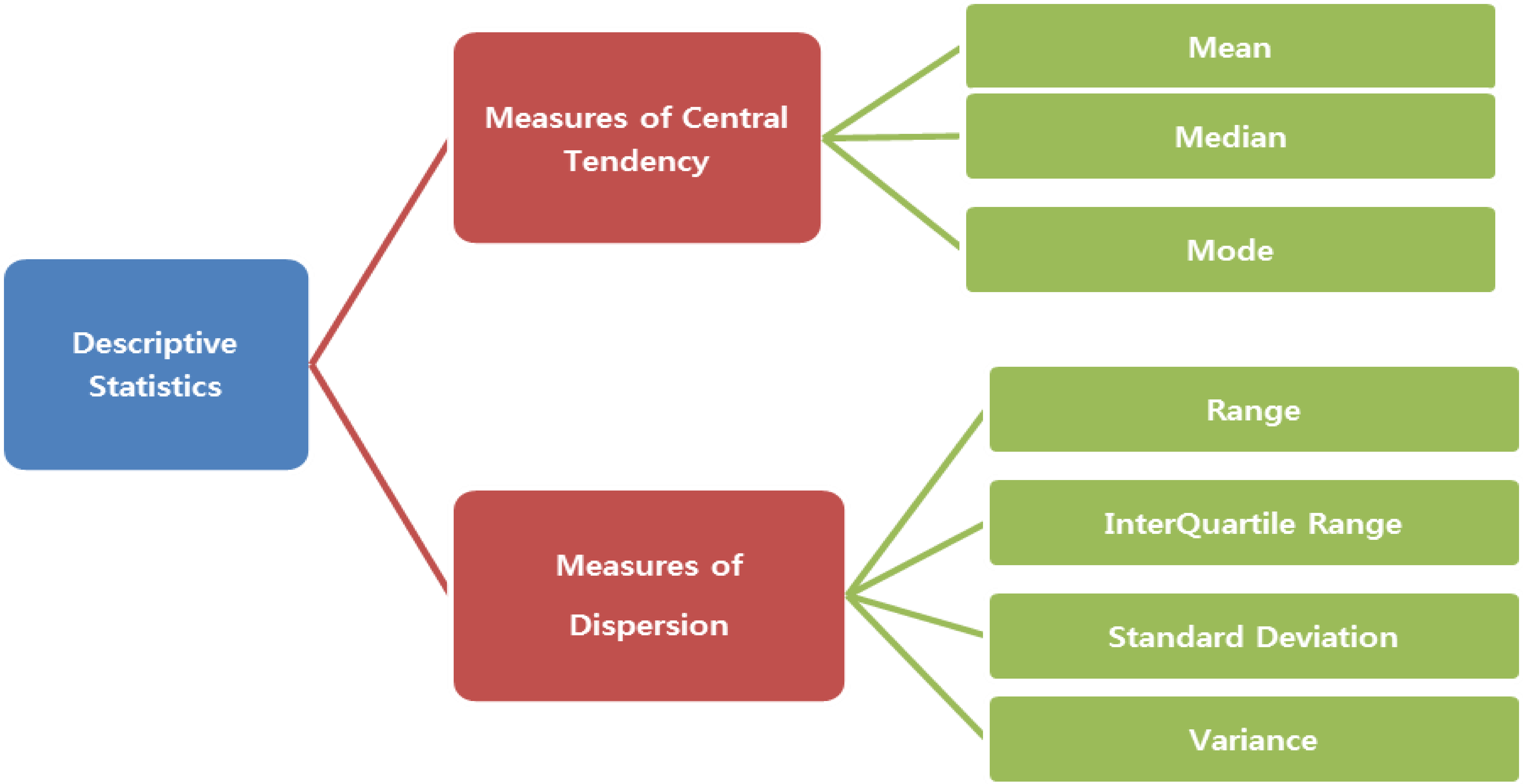
- **Five variables**
  - Payment Per Beneficiary
  - Services Per Beneficiary
  - Average Birth Year of Beneficiaries
  - Percentage of Benes with Diabetes
  - Average Health Risk Score of Beneficiaries
- **100 de-identified Providers**
  - Specialty 01, General Practice
  - Data source: CMS Public use file (PUF)

A	B	C	D	E	F
Provider ID	Payment Per Bene	Services per Bene	Inv. Age of Benes	% of Benes with Diabetes	Inv Health Risk Score of Benes
ID_646096	$190.27	2	1934	41.00%	0.514986095
ID_117056	$182.96	4	1939	29.00%	0.907688118
ID_282206	$198.67	5	1942	21.00%	1.042318115
ID_101716	$528.02	13	1939	25.00%	1.153668666
ID_803756	$176.48	2	1947	54.00%	0.343760743
ID_683709	$306.88	5	1936	39.00%	0.635525898
ID_435156	$210.55	3	1941	48.00%	0.531547334
ID_895138	$224.27	5	1942	36.00%	0.908265213
ID_761090	$206.43	3	1944	35.00%	0.255924656
ID_669911	$70.75	2	1942	38.00%	0.954653938
ID_365846	$232.02	3	1937	45.00%	0.453782275
ID_273953	$238.43	5	1940	20.00%	1.153801777
ID_916080	$128.84	5	1942	23.00%	1.124606388
ID_849495	$127.72	1	1944	52.00%	0.6222001
ID_306931	$286.01	8	1941	75.00%	0.757862827
ID_319060	$72.78	2	1940	36.00%	0.668047298
ID_656054	$100.70	2	1941	47.00%	0.64977258
ID_674144	$174.06	2	1942	46.00%	0.441559588
ID_186871	$323.51	4	1942	41.00%	0.579273591
ID_900197	$372.45	5	1937	42.00%	0.376690398
ID_243834	$171.60	2	1940	48.00%	0.46628742
ID_129149	$282.90	7	1942	42.00%	0.671276096

Outlier Detection Techniques/Statistical Tools

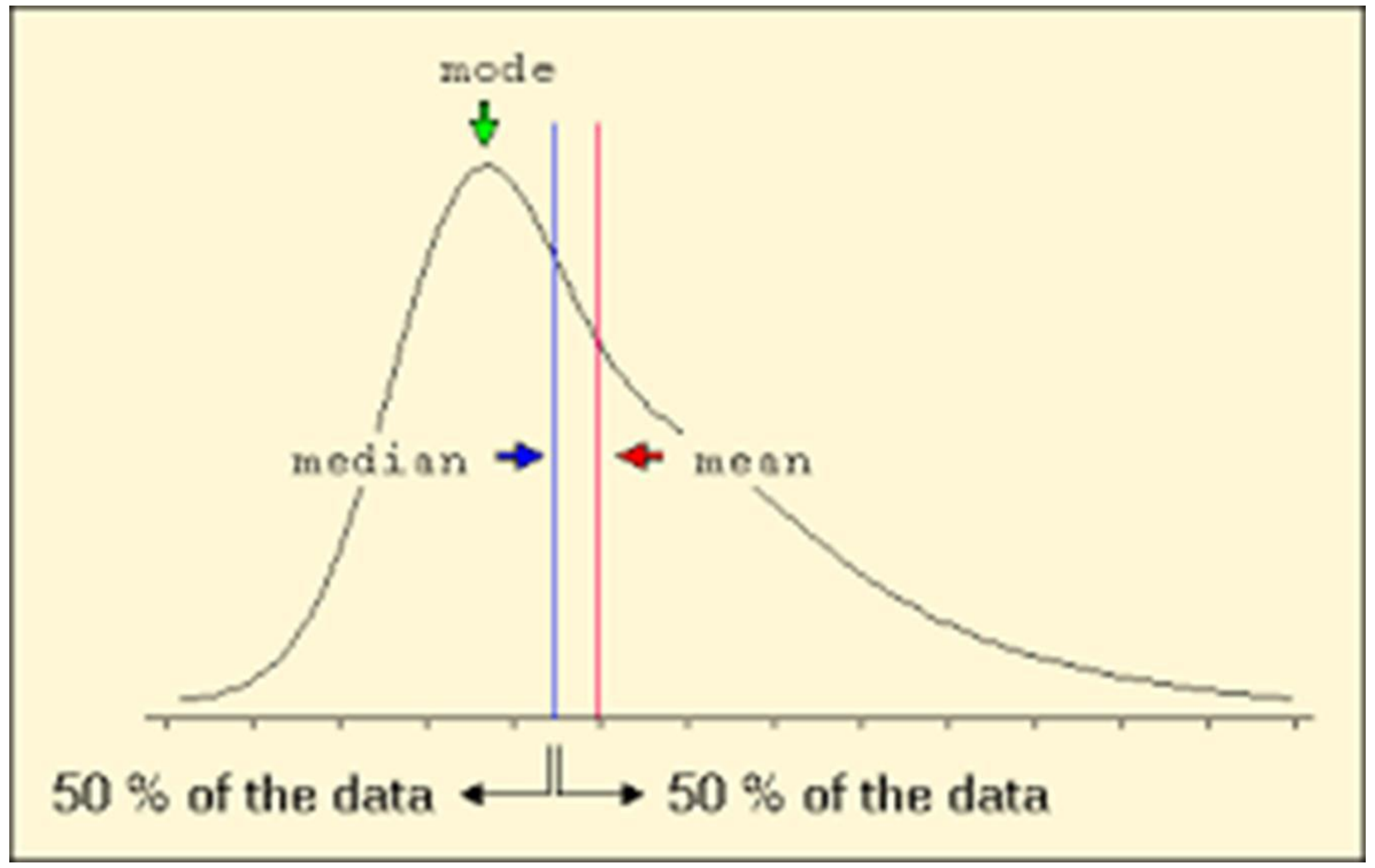
Descriptive Statistics

Descriptive statistics summarizes the data and it is essential to better understand the data



Outlier Detection Techniques/Statistical Tools

Descriptive Statistics



Mean

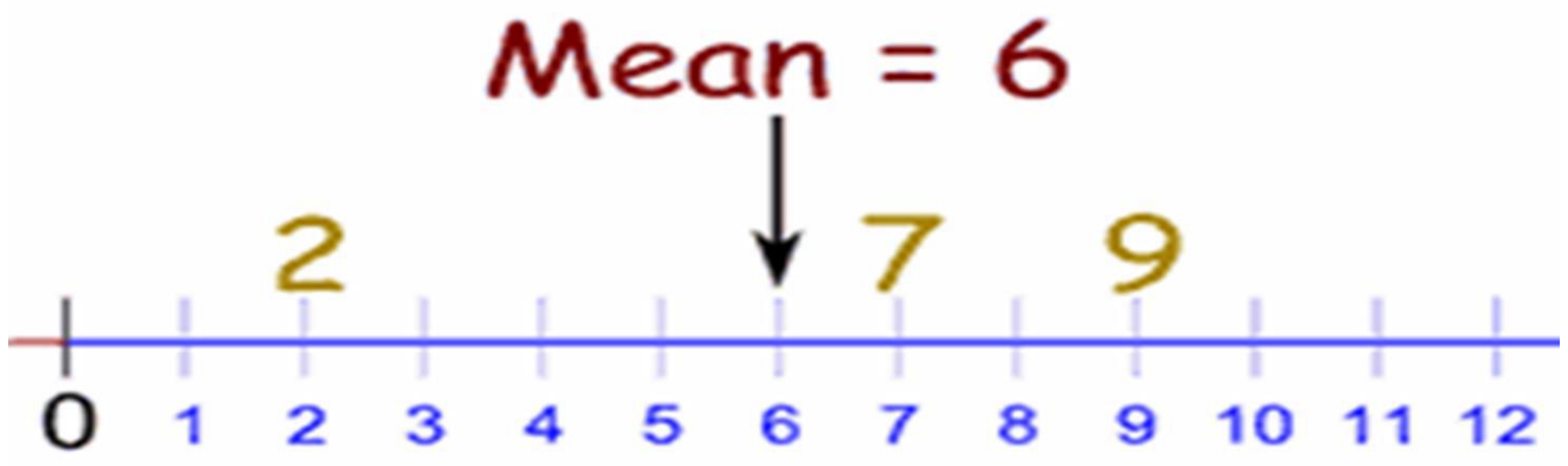
- The average of the values on a given measurement/indicator
- The mean is subject to the pull of influential points/outliers
- **3,5,5,8, Mean= 7**
- **3,5,5,43, Mean=14**

Median

- If odd set of numbers then the median is the one middle number
- If even set of numbers then the median is the mean of the two middle numbers
- The median is resilient to influential points/outlier – as long as the middle values remain the same
- **3,5,5,8, Median= (5+5)/2 = 5** (Even numbers)
- **3,5,5,43, Median= (5+5)/2 = 5** (Even numbers)

Mode

- The value that appears most often in a set of data
- Hint: Mode =“Most”
- **3,5,5,8, Mode = 5 ; 3,5,5,43, Mode = 5**
- **1,2,3,4, No Mode**



Range

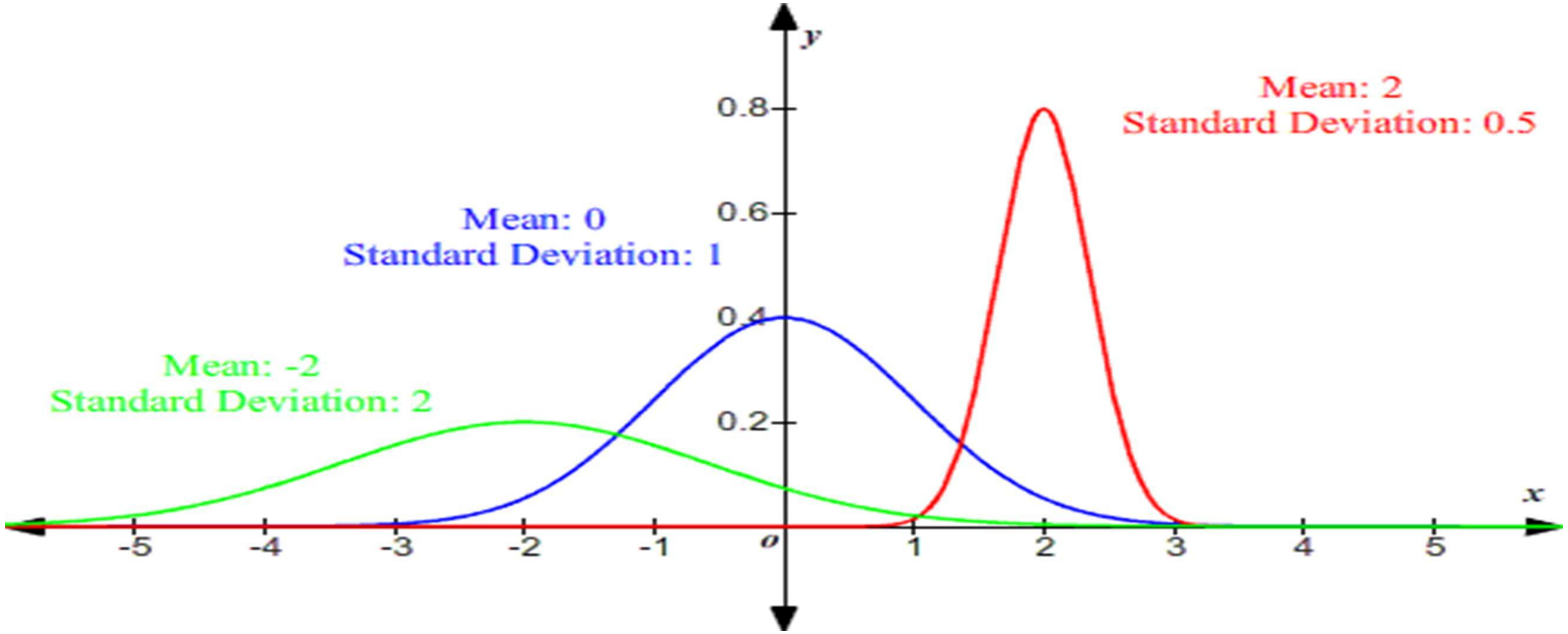
- The Range is the difference between the lowest and highest values
- **3,5,5,8, Range = 8 - 3=5 ; 3,5,5,43, Range = 43 - 3= 40**
- Illustrates the spread of the data

Outlier Detection Techniques/Statistical Tools

Descriptive Statistics

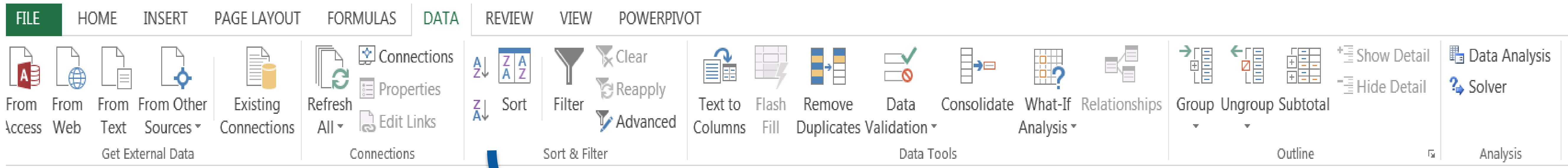
Standard Deviation

- A measure of the dispersion or variation in a distribution, lack of dispersion can result in a lack of outlier.
- If the data is close together, the standard deviation is small. If the data is spread out, the standard deviation is large.

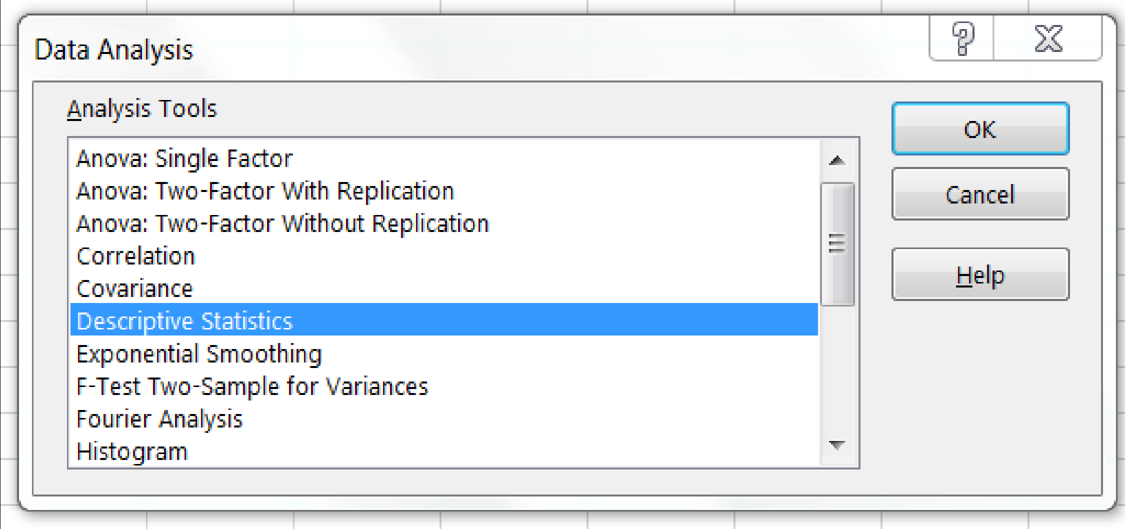


Outlier Detection Techniques/Statistical Tools

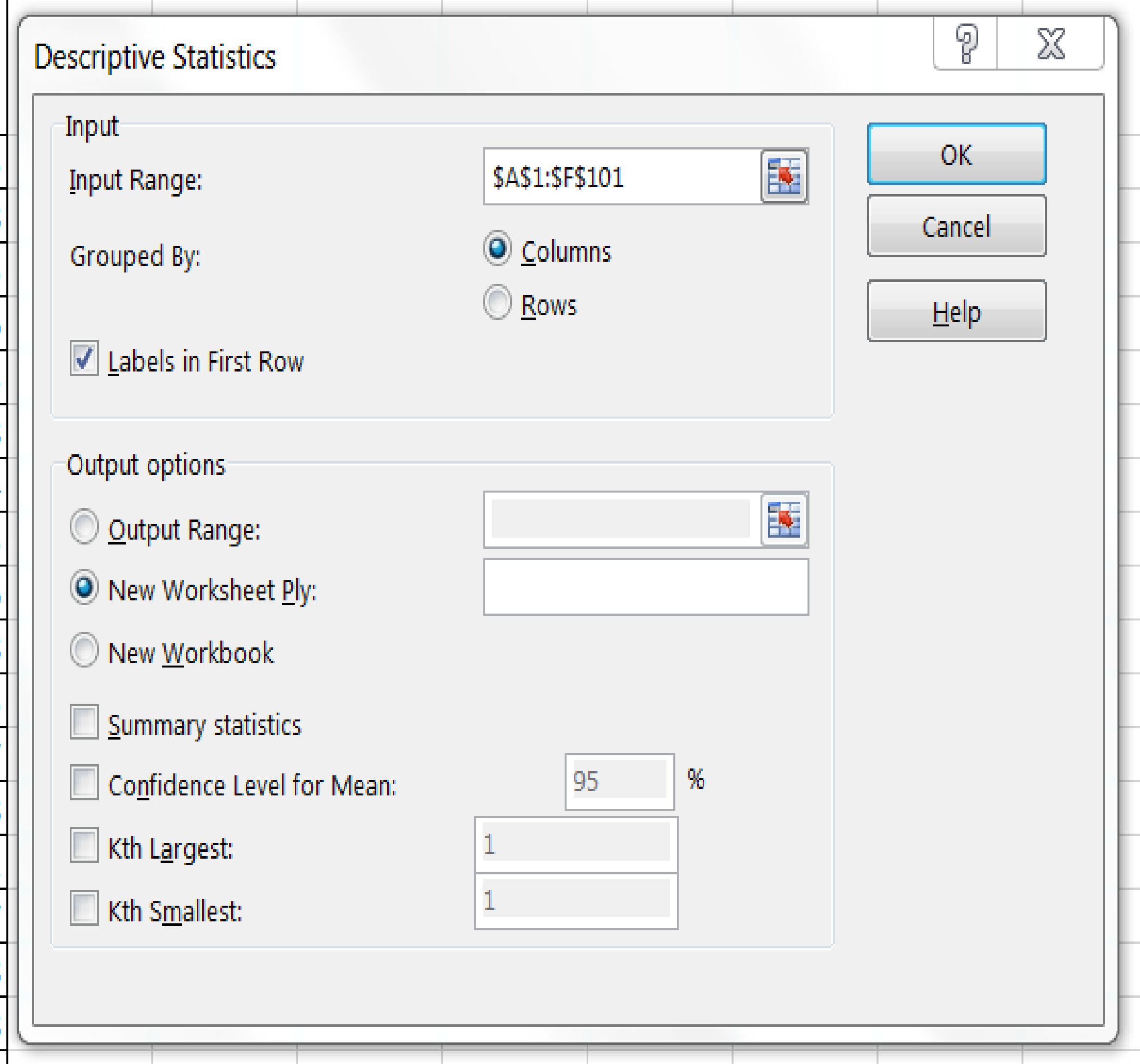
Descriptive Statistics – Excel Tool



Provider ID	Payment Per Bene	Services per Bene	Inv. Age of Benes	% of Benes with Diabetes	Inv Health Risk Score of Benes
ID_646096	$190.27	2	1934	41.00%	0.514986095
ID_117056	$182.96	4	1939	29.00%	0.907688118
ID_282206	$198.67	5	1942	21.00%	1.042318115
ID_101716	$528.02	13	1939	25.00%	1.153668666
ID_803756	$176.48	2	1947	54.00%	0.343760743
ID_683709	$306.88	5	1936	39.00%	0.635525898
ID_435156	$210.55	3	1941	48.00%	0.531547334
ID_895138	$224.27	5	1942	36.00%	0.908265213
ID_761090	$206.43	3	1944	35.00%	0.255924656
ID_669911	$70.75	2	1942	38.00%	0.954653938
ID_365846	$232.02	3	1937	45.00%	0.453782275
ID_273953	$238.43	5	1940	20.00%	1.153801777
ID_916080	$128.84	5	1942	23.00%	1.124606388
ID_849495	$127.72	1	1944	52.00%	0.6222001
ID_306931	$286.01	8	1941	75.00%	0.757862827
ID_319060	$72.78	2	1940	36.00%	0.668047298
ID_656054	$100.70	2	1941	47.00%	0.64977258
ID_674144	$174.06	2	1942	46.00%	0.441559588
ID_186871	$323.51	4	1942	41.00%	0.579273591
ID_900197	$372.45	5	1937	42.00%	0.376690398
ID_243834	$171.60	2	1940	48.00%	0.46628742
ID_129149	$282.90	7	1942	42.00%	0.671276096



Provider ID	Payment Per Bene	Services per Bene	Inv. Age of Benes	% of Benes with Diabetes	Inv Health Risk Score of Benes
ID_646096	$190.27	2	1934	41.00%	0.514986095
ID_117056	$182.96	4	1939	29.00%	0.907688118
ID_282206	$198.67	5	1942	21.00%	1.042318115
ID_101716	$528.02	13	1939	25.00%	1.153668666
ID_803756	$176.48	2	1947	54.00%	0.343760743
ID_683709	$306.88	5	1936	39.00%	0.635525898
ID_435156	$210.55	3	1941	48.00%	0.531547334
ID_895138	$224.27	5	1942	36.00%	0.908265213
ID_761090	$206.43	3	1944	35.00%	0.255924656
ID_669911	$70.75	2	1942	38.00%	0.954653938
ID_365846	$232.02	3	1937	45.00%	0.453782275
ID_273953	$238.43	5	1940	20.00%	1.153801777
ID_916080	$128.84	5	1942	23.00%	1.124606388
ID_849495	$127.72	1	1944	52.00%	0.6222001
ID_306931	$286.01	8	1941	75.00%	0.757862827
ID_319060	$72.78	2	1940	36.00%	0.668047298
ID_656054	$100.70	2	1941	47.00%	0.64977258
ID_674144	$174.06	2	1942	46.00%	0.441559588
ID_186871	$323.51	4	1942	41.00%	0.579273591
ID_900197	$372.45	5	1937	42.00%	0.376690398
ID_243834	$171.60	2	1940	48.00%	0.46628742



Integrity Management Services

Outlier Detection Techniques/Statistical Tools

Descriptive Statistics – Excel Output

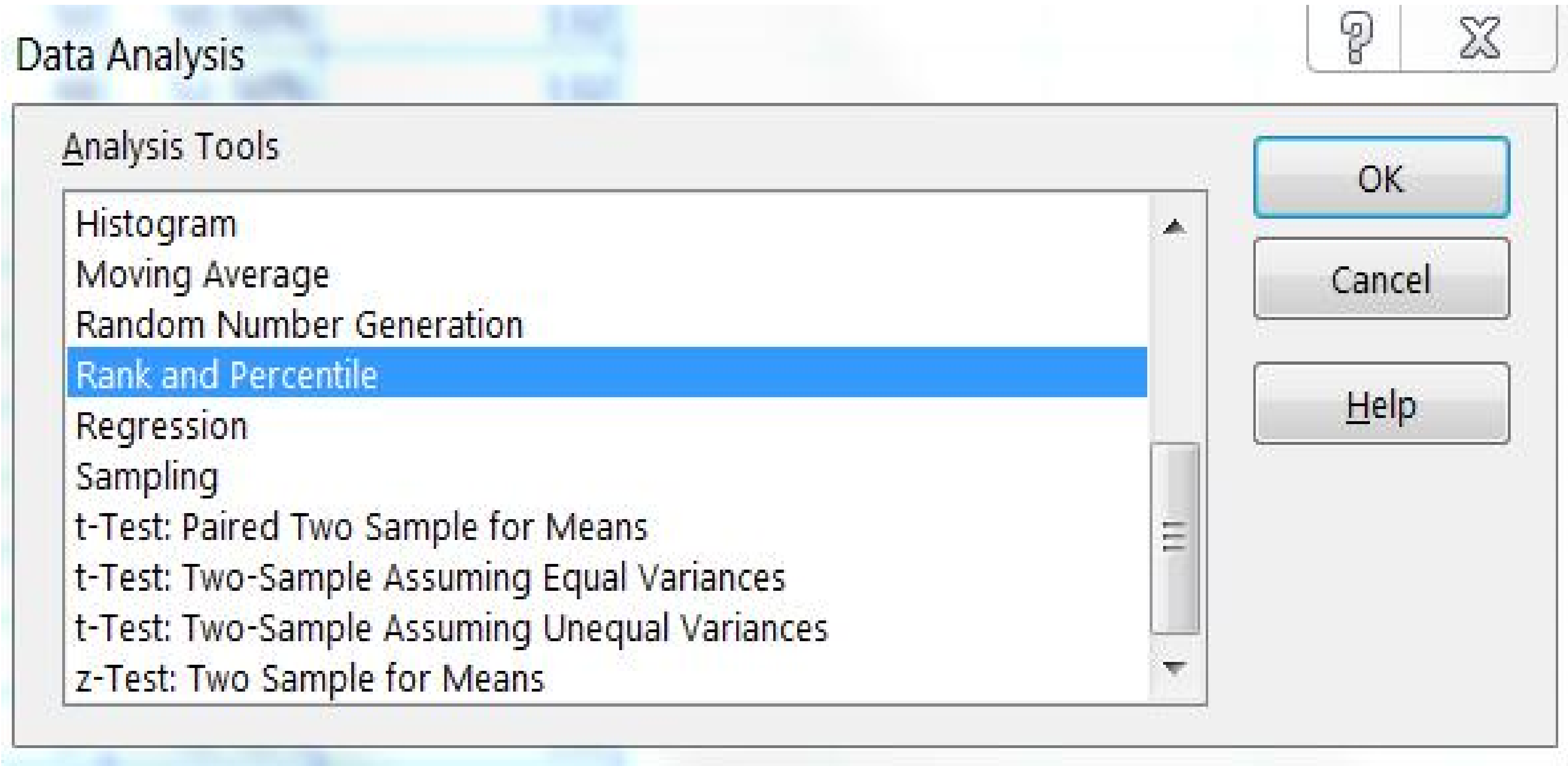
Payment Per Bene		*Services per Bene*		*Inv. Age of Benes*		*% of Benes with Diabetes*		*Inv Health Risk Score of Benes*	
Mean	$228.24	Mean	4.62	Mean	1942	Mean	37%	Mean	0.76
Standard Error	15.37	Standard Error	0.33	Standard Error	0.40	Standard Error	0.01	Standard Error	0.03
Median	$196.69	Median	3.83	Median	1941	Median	36%	Median	0.76
Mode	#N/A	Mode	#N/A	Mode	1941	Mode	41%	Mode	#N/A
Standard Deviation	153.65	Standard Deviation	3.29	Standard Deviation	4.03	Standard Deviation	0.12	Standard Deviation	0.27
Sample Variance	23,608.47	Sample Variance	10.83	Sample Variance	16.26	Sample Variance	0.01	Sample Variance	0.07
Kurtosis	20.07	Kurtosis	6.87	Kurtosis	0.86	Kurtosis	0.53	Kurtosis	-1.00
Skewness	3.50	Skewness	2.28	Skewness	0.28	Skewness	0.65	Skewness	0.22
Range	$1,245.86	Range	18.22	Range	23	Range	60%	Range	1.13
Minimum	$10.49	Minimum	1.05	Minimum	1931	Minimum	15%	Minimum	0.26
Maximum	$1,256.36	Maximum	19.27	Maximum	1954	Maximum	75%	Maximum	1.39
Sum	$22,823.87	Sum	462.08	Sum	194186	Sum	36.93	Sum	75.57
Count	100.00	Count	100.00	Count	100	Count	100.00	Count	100.00

Outlier Detection Techniques/Statistical Tools

Ranking and Percentile – Excel Tool and Output

Excel has a tool to Rank providers based on their sorted position in each variable

- Provider's ranking is generated within each indicator
- Total ranking is calculated by addition the total ranking – the lower the total ranking the more an outlier a provider is



A	N	O	P	Q	R	S	T	U	V
Provider ID	Point	% of Benes with Diabetes	Rank	Percent	Point	Inv Health Risk Score of Benes	Rank	Percent	Total Ranking
ID_573433	69	33%	59	39.30%	69	0.86	40	60.60%	110
ID_513905	83	44%	28	70.70%	83	0.79	46	54.50%	127
ID_543543	35	46%	20	76.70%	35	0.64	61	39.30%	129
ID_871960	97	57%	5	94.90%	97	0.40	93	7.00%	130
ID_306931	15	75%	1	100.00%	15	0.76	50	50.50%	132
ID_791822	92	35%	52	44.40%	92	0.76	48	52.50%	132
ID_124885	63	52%	9	89.80%	63	0.88	37	63.60%	136
ID_473715	23	71%	2	98.90%	23	0.76	49	51.50%	141
ID_365774	41	45%	25	73.70%	41	0.69	54	46.40%	146
ID_550162	30	51%	12	88.80%	30	0.43	89	11.10%	151
ID_737783	54	37%	45	53.50%	54	1.34	2	98.90%	151
ID_704678	76	57%	5	94.90%	76	0.81	45	55.50%	152
ID_158539	53	25%	82	14.10%	53	1.14	8	92.90%	158
ID_479429	60	29%	70	27.20%	60	1.11	11	89.80%	161
ID_129149	22	42%	32	67.60%	22	0.67	57	43.40%	168
ID_545940	66	31%	64	35.30%	66	0.90	36	64.60%	170
ID_101716	4	25%	82	14.10%	4	1.15	7	93.90%	175

Outlier Detection Techniques/Statistical Tools

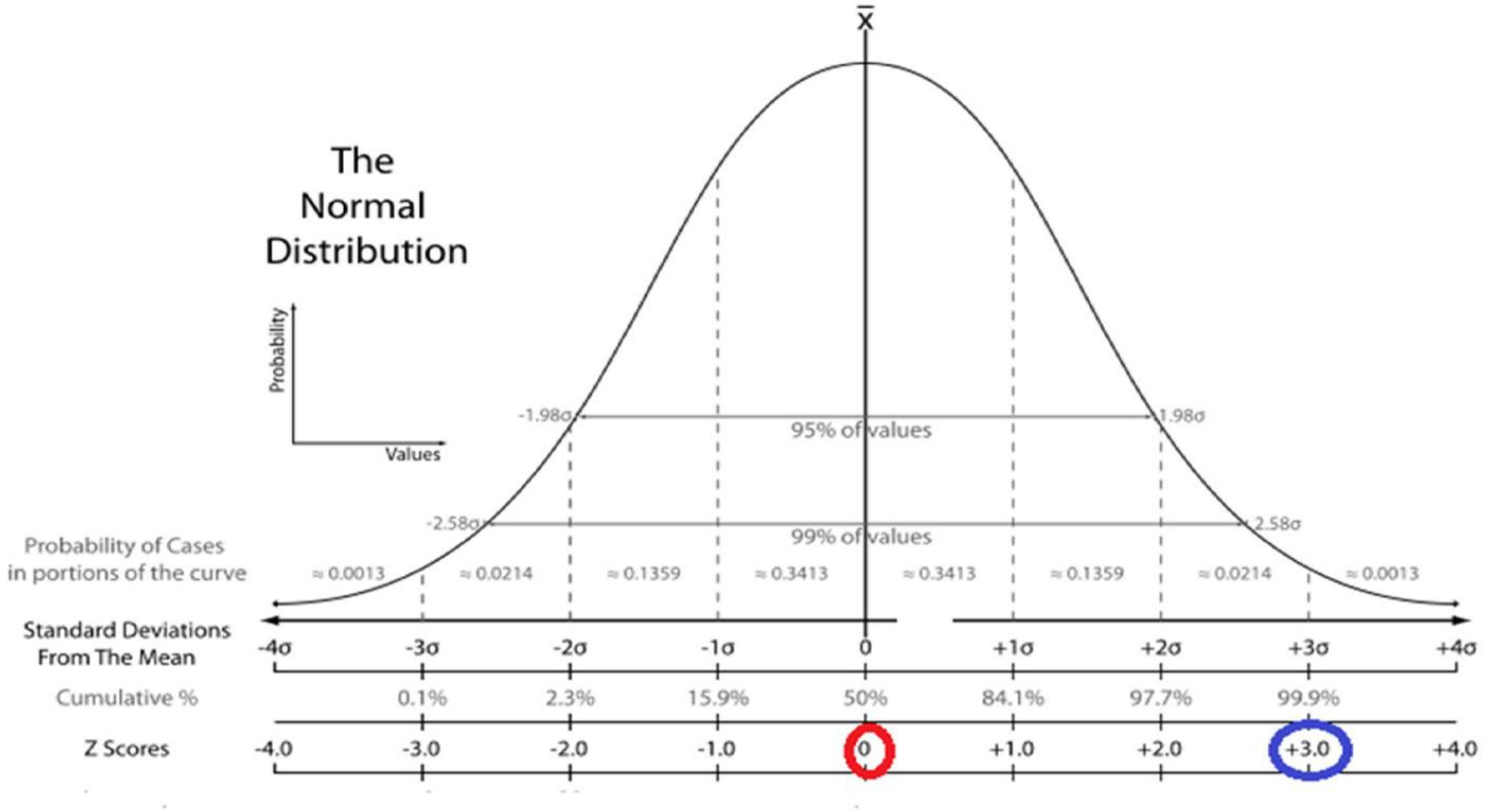
Z - Score

Z – score

- A measure of how far a value is from the mean in terms of the number of standard deviations
  - for example, provider payment per beneficiary
- The raw data is re-scaled to have mean 0 and standard deviation 1
- A raw data value that is exactly equal to the mean corresponds to a Z-score value of 0
- The z-score re-scaling of data is commonly used to identify outliers
- Z-score may be used as a ranking method using multiple indicators – Composite Ranking
- Re-scaled data loses its original interpretation (change of units))

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation



Outlier Detection Techniques/Statistical Tools

Z – Score – Excel Formulas and Output

- Computing Z- score - one variable/Indicator
  - Payment per Bene
- Mean

B	C	D	E
Payment Per Bene	Mean	St Dev	Z-s
$1,256.36	=AVERAGE($B$2:$B$91)		

- St. Dev (Standard Deviation)

A	B	C	D	E	
Provider ID	Payment Per Bene	Mean	St Dev	Z-score	RN
ID_365774	$1,256.36	$246.78	=STDEVP($B$2:$B$91)		

- Z – score

A	B	C	D	E	
Provider ID	Payment Per Bene	Mean	St Dev	Z-score	RM:
ID_365774	$1,256.36	$246.78	$149.90	=((B2-C2)/D2)	

A	B	C	D	E
Provider ID	Payment Per Bene	Mean	St Dev	Z-score
ID_365774	$1,256.36	$246.78	$149.90	6.735146
ID_998581	$678.83	$246.78	$149.90	2.882315
ID_871960	$550.05	$246.78	$149.90	2.023186
ID_573433	$530.77	$246.78	$149.90	1.894607
ID_101716	$528.02	$246.78	$149.90	1.876245
ID_871129	$422.63	$246.78	$149.90	1.173191
ID_744767	$421.01	$246.78	$149.90	1.162322
ID_550162	$414.21	$246.78	$149.90	1.117015
ID_472611	$406.85	$246.78	$149.90	1.067880
ID_900197	$372.45	$246.78	$149.90	0.838382
ID_158539	$348.24	$246.78	$149.90	0.676898
ID_121221	$333.20	$246.78	$149.90	0.576526
ID_704678	$326.39	$246.78	$149.90	0.531121

Outlier Detection Techniques/Statistical Tools

Z - Score

Composite Ranking

- Includes multiple indicators in the ranking method
- Z- score is calculated for each indicator
- Providers who are above 2 or 3 Standard deviation above the mean are consider outliers
- This is a simple way of ranking providers on multiple indicators

Disadvantages

- Loses the original interpretation of the raw data
- Gives equal weights to all indicators in composite ranking

Outlier Detection Techniques/Statistical Tools

Z – Score (Composite Ranking) – Excel Output

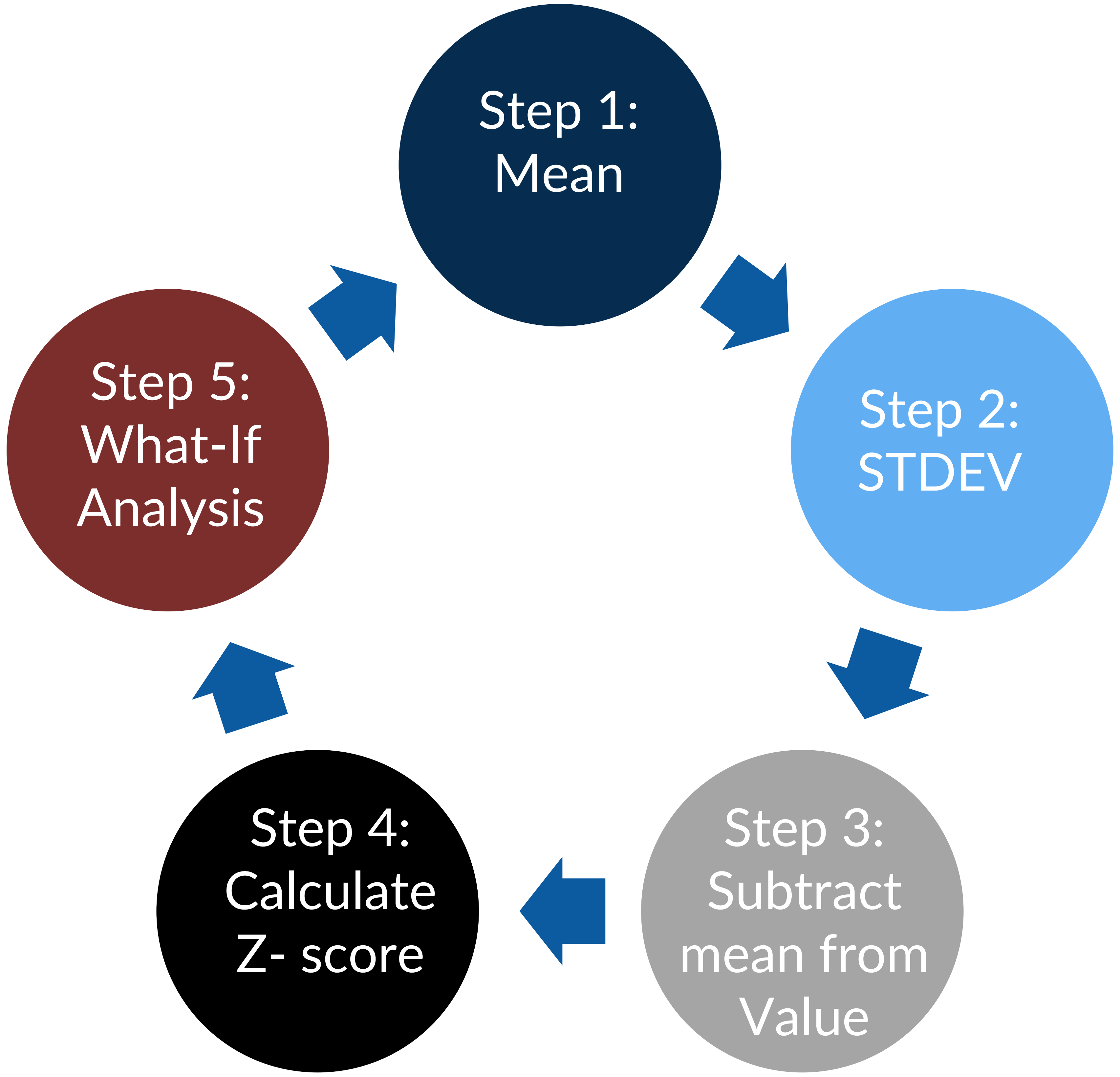
	A	AA	AB	AC	AD	AE	AF
1	**Provider ID**	**Step 5 What-If Analysis: Payment Per Bene**	**Step 5 What-If Analysis: Services per Bene**	**Step 5 What-If Analysis: Inv.Avg. Age of Benes**	**Step 5 What-If Analysis: % of Benes with Diabetes**	**Step 5 What-If Analysis: Inv. Avg. Health Risk Score of Benes**	**Total Z-Score**
2	ID_365774	1	1	0	0	0	2
3	ID_573433	0	0	1	0	0	1
4	ID_998581	0	1	0	0	0	1
5	ID_744767	0	1	0	0	0	1
6	ID_306931	0	0	0	1	0	1
7	ID_871960	0	0	0	0	0	0
8	ID_101716	0	0	0	0	0	0
9	ID_871129	0	0	0	0	0	0
10	ID_473715	0	0	0	0	0	0
11	ID_513905	0	0	0	0	0	0
12	ID_550162	0	0	0	0	0	0

Outlier Detection Techniques/Statistical Tools

Z – Score (Composite Ranking) – Excel Steps

Step 5: What-If Analysis

=IF(X>=3,1,0)



Outlier Detection Techniques/Statistical Tools

Z – Score (Composite Ranking) – Excel Output

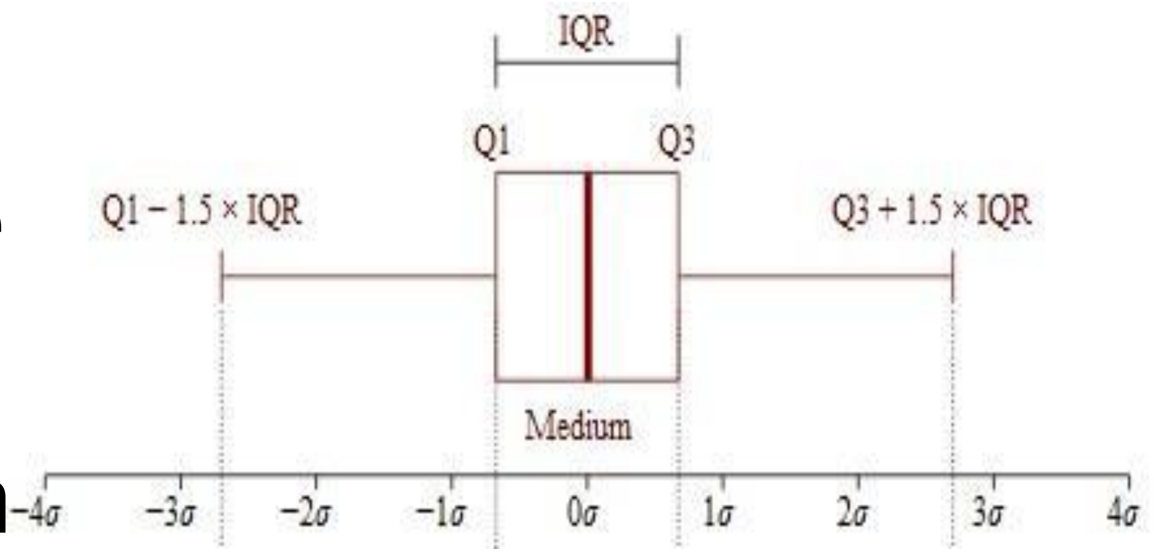
1	Provider ID	Payment Per Bene	Step 1	Step 2	Step 3	Step 4	Step 5	Total Z-Score
2	ID_365774	$1,256.36	$228.24	152.88	$1,028.12	6.72	1	2
3	ID_573433	$530.77	$228.24	152.88	$302.53	1.98	0	1
4	ID_998581	$678.83	$228.24	152.88	$450.59	2.95	0	1
5	ID_744767	$421.01	$228.24	152.88	$192.77	1.26	0	1
6	ID_306931	$286.01	$228.24	152.88	$57.78	0.38	0	1
7	ID_871960	$550.05	$228.24	152.88	$321.81	2.10	0	0
8	ID_101716	$528.02	$228.24	152.88	$299.78	1.96	0	0
9	ID_871129	$422.63	$228.24	152.88	$194.40	1.27	0	0
10	ID_473715	$279.76	$228.24	152.88	$51.53	0.34	0	0

Outlier Detection Techniques/Statistical Tools

Box plot

Box plot

- A good way to summarize large amounts of data
- A measure of spread, based on dividing a data set into quartiles
  - Q1 is the "middle" value in the *first* half of the rank-ordered data se
  - Q2 is the median value in the set.
  - Q3 is the "middle" value in the *second* half of the rank-ordered data



- Right tail Outlier= 75th Percentile + 1.5*IQR (Inter-Quartile Range),
  - where IQR = (Q3 – Q1)

Provider ID	Payment Per Bene	Services per Bene	Inv. Age of Benes	% of Benes with Diabetes	Inv Health	Box Plot Pmt Outlier? Yes = 1; No = 0	Box Plot Servs Outlier? Yes = 1; No = 0	Box Plot Age Outlier? Yes = 1; No = 0
ID 573433	$530.77	11	1954	33.00%	0.862515	=IF(B2 >= PERCENTILE.INC(B$2:B$101,0.75) + 1.5*(PERCENTILE.INC(B$2:B$101,0.75) - PERCENTILE.INC(B$2:B$101,0.25)), 1,0)		

Outlier Detection Techniques/Statistical Tools

Boxplot - Excel

A	B	C	D	E	F	G	H	I	J	K	L
Provider ID	Payment Per Bene	Services per Bene	Inv. Age of Benes	Benes with Diabetes	Inv Health Risk Score of Benes	Box Plot PMT Outlier?	Box Plot Servs Outlier?	Box Plot Age Outlier?	Box Plot Diab Outlier?	Box Plot Inv Health Outlier?	Number of Outlier (Max = 5)
ID_573433	$530.77	11	1954	33%	0.86	1	1	1	0	0	3
ID_365774	$1,256.36	19	1940	45%	0.69	1	1	0	0	0	2
ID_998581	$678.83	19	1936	24%	1.05	1	1	0	0	0	2
ID_871960	$550.05	5	1953	57%	0.40	1	0	1	0	0	2
ID_101716	$528.02	13	1939	25%	1.15	1	1	0	0	0	2
ID_744767	$421.01	15	1939	23%	1.10	0	1	0	0	0	1
ID_306931	$286.01	8	1941	75%	0.76	0	0	0	1	0	1
ID_482392	$23.85	11	1942	24%	1.15	0	1	0	0	0	1
ID_871129	$422.63	7	1933	69%	0.44	0	0	0	0	0	0
ID_550162	$414.21	6	1945	51%	0.43	0	0	0	0	0	0

Outlier Detection Techniques/Statistical Tools
Cluster Analysis

Outlier Detection Techniques/Statistical Tools

Cluster Analysis

Cluster analysis

- Systematically way of grouping providers using measure of similarity
- Summarize the descriptive statistics of the clusters, including the mean value (centroid) of each cluster

- Diverse types of variables can be used to cluster the data

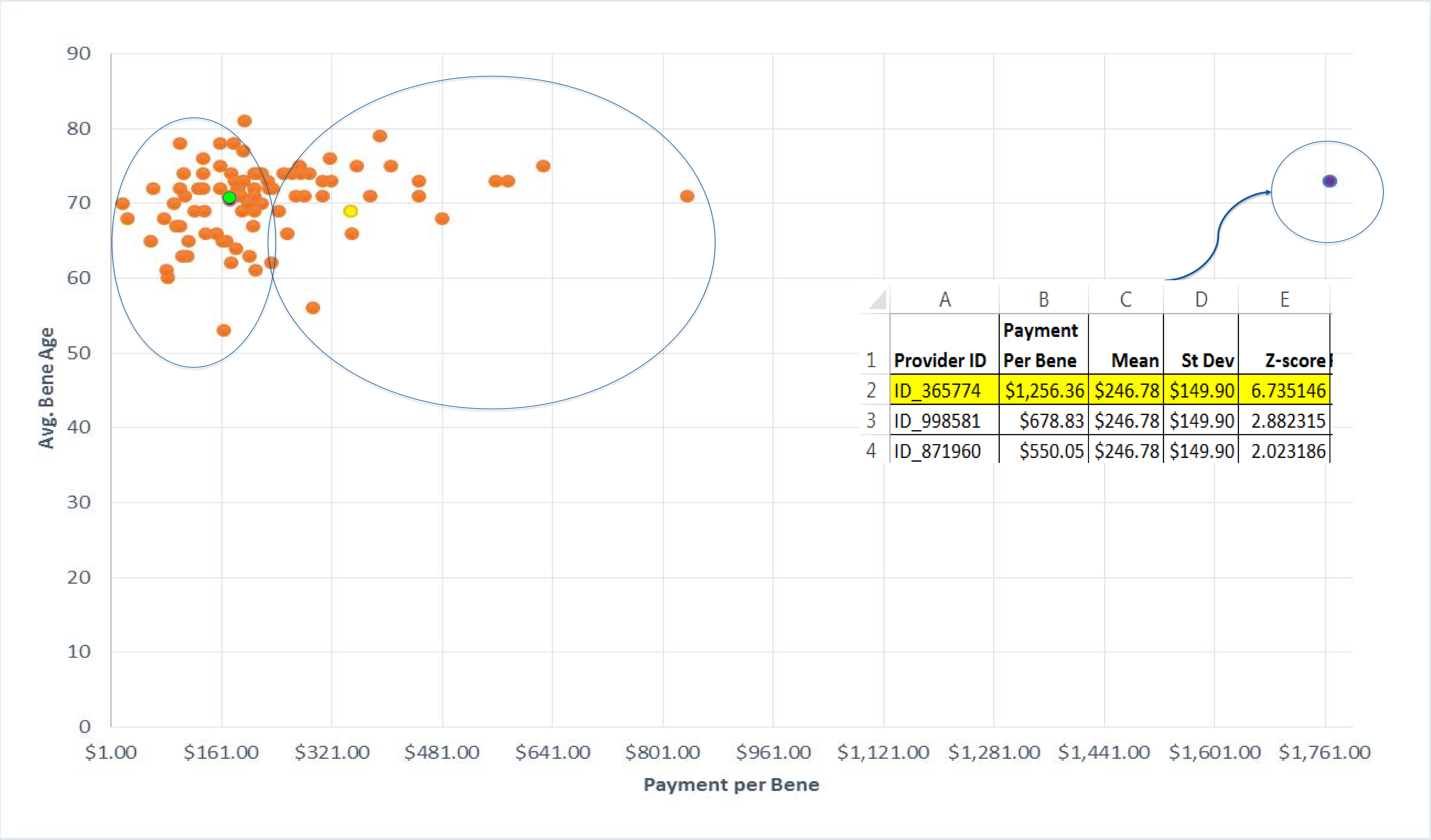
- For example, in healthcare claims data some of the indicators that can be included are:
  - Procedure codes
  - Place of service
  - Payment amount

- Understanding the data is essential because not all potential indicators will be informative in clustering the data properly, and using the most fitting indicators will reduce misclassification

Outlier Detection Techniques/Statistical Tools

Cluster Analysis – Excel Output

	A	B	C	D	E	F	G	H	I	J	K
1	Provider ID	Payment Per Bene	Avg. Age of Benes	Index	Distance to CENTROID 1	Distance to CENTROID 2	Distance to CENTROID 3	Payment Per Bene	Avg. Age of Benes	CLASS	Minimum Distance
2	ID_365774	$1,256.36	74	41	1,087.0448	955.6215	0.0029	$1,256.36	74	Cluster3	0.002875384
3	ID_998581	$678.83	78	88	509.5522	378.1216	577.5395	$678.83	78	Cluster2	378.1216224
4	ID_871960	$550.05	61	97	380.8877	249.6198	706.4262	$550.05	61	Cluster2	249.6198434
5	ID_573433	$530.77	60	69	361.6537	230.4279	725.7152	$530.77	60	Cluster2	230.4278977
6	ID_101716	$528.02	75	4	358.7220	227.2924	728.3332	$528.02	75	Cluster2	227.2923592
7	ID_871129	$422.63	81	47	253.4883	122.1386	833.7477	$422.63	81	Cluster2	122.1385945
8	ID_744767	$421.01	75	28	251.7130	120.2825	835.3481	$421.01	75	Cluster2	120.2825117
9	ID_550162	$414.21	69	30	244.9179	113.5649	842.1536	$414.21	69	Cluster2	113.5648968
10	ID_472611	$406.85	83	51	237.7992	106.5500	849.5517	$406.85	83	Cluster2	106.5500194
11	ID_900197	$372.45	77	20	203.2012	71.8051	883.9102	$372.45	77	Cluster2	71.80512016
12	ID_158539	$348.24	73	53	178.9332	47.5096	908.1116	$348.24	73	Cluster2	47.50956181



	A	B	C	D	E
1	Provider ID	Payment Per Bene	Mean	St Dev	Z-score
2	ID_365774	$1,256.36	$246.78	$149.90	6.735146
3	ID_998581	$678.83	$246.78	$149.90	2.882315
4	ID_871960	$550.05	$246.78	$149.90	2.023186

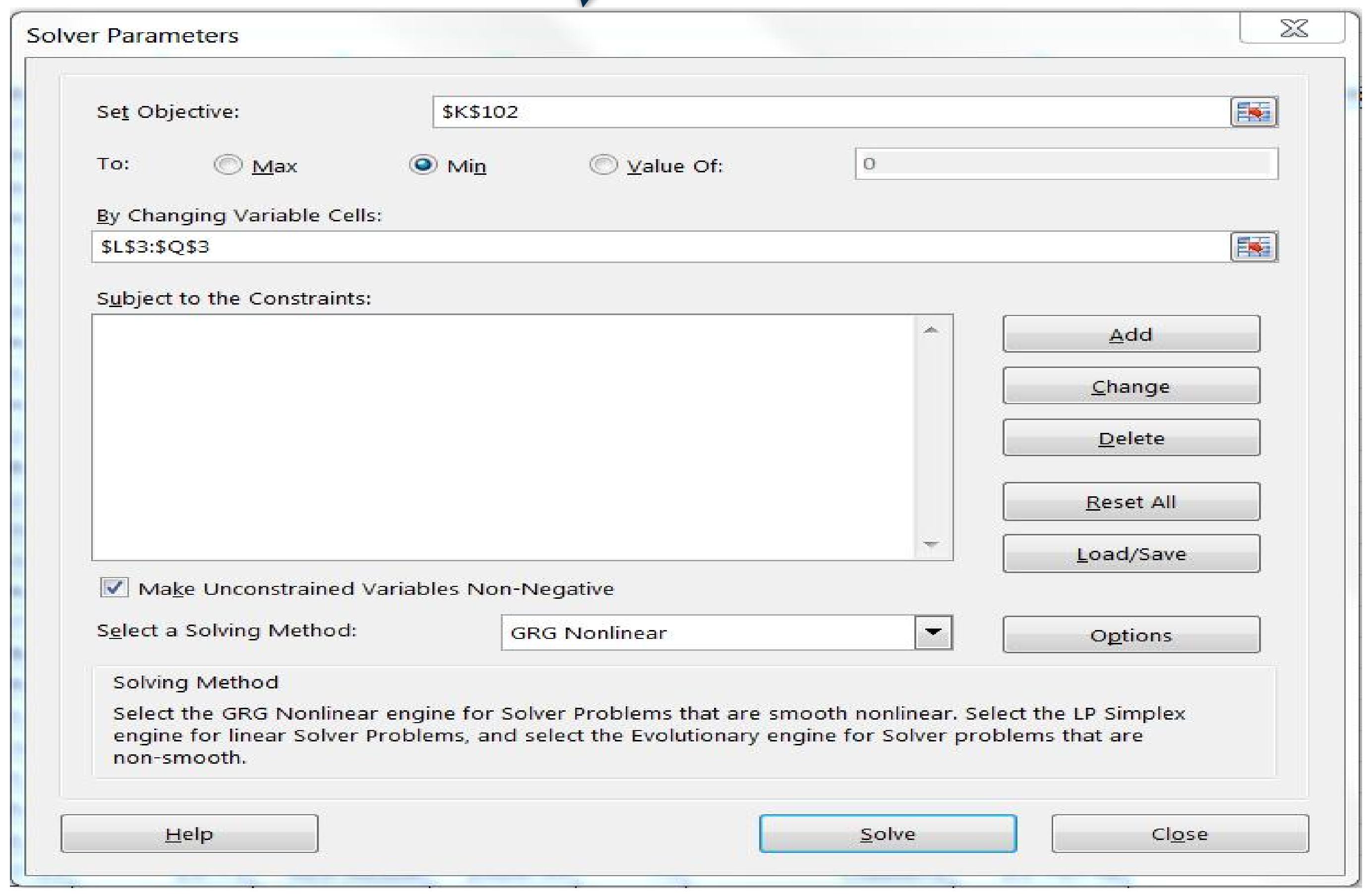
Outlier Detection Techniques/Statistical Tools

Cluster Analysis – Excel Steps

L	M	N	O	P	Q
Cluster Center 1		Cluster Center 2		Cluster Center 3	
Payment Per Bene	Avg. Age of Benes	Payment Per Bene	Avg. Age of Benes	Payment Per Bene	Avg. Age of Benes
169.3123609	71.82497411	300.7337187	73.38276732	1256.35282	73.99812514

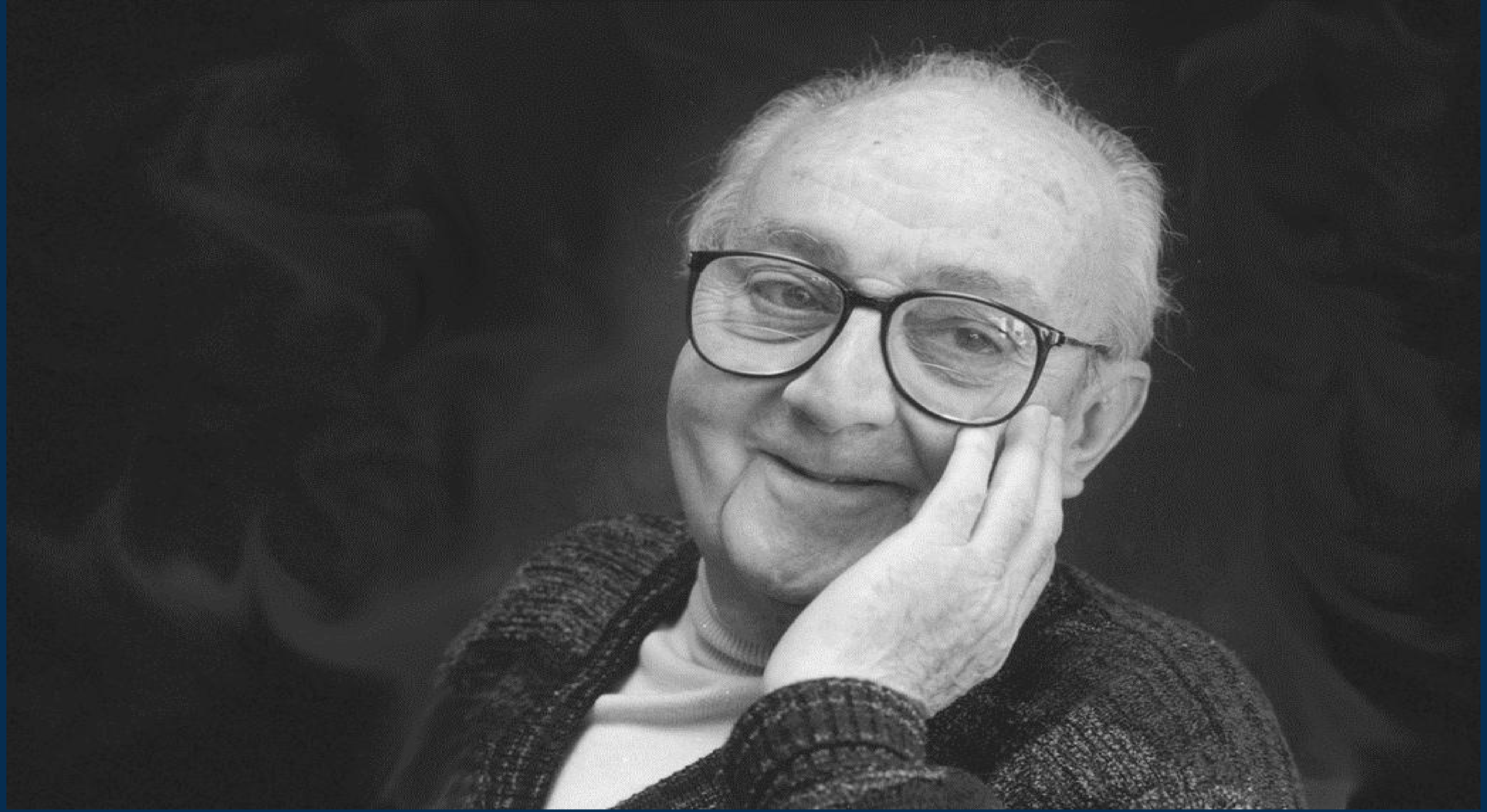
- Distance to CENTROID 1 =SQRT((B2-$L$3)^2+(C2-$M$3)^2)
- CLASS = IF(MIN(E2:G2)=E2,"Cluster1",IF(MIN(E2:G2)=F2,"Cluster2","Cluster3"))
- Minimum Distance = IF(J2="Cluster2",F2,IF(J2="Cluster3",G2,IF(J2="Cluster1",E2)))

Sum of Minimum Distance



	A	B	C	D	E	F	G	H	I	J	K
1	Provider ID	Payment Per Bene	Avg. Age of Benes	Index	Distance to CENTROID 1	Distance to CENTROID 2	Distance to CENTROID 3	Payment Per Bene	Avg. Age of Benes	CLASS	Minimum Distance
2	ID_365774	$1,256.36	74	41	1,087.0448	955.6215	0.0029	$1,256.36	74	Cluster3	0.002875384
3	ID_998581	$678.83	78	88	509.5522	378.1216	577.5395	$678.83	78	Cluster2	378.1216224
4	ID_871960	$550.05	61	97	380.8877	249.6198	706.4262	$550.05	61	Cluster2	249.6198434
5	ID_573433	$530.77	60	69	361.6537	230.4279	725.7152	$530.77	60	Cluster2	230.4278977
6	ID_101716	$528.02	75	4	358.7220	227.2924	728.3332	$528.02	75	Cluster2	227.2923592
7	ID_871129	$422.63	81	47	253.4883	122.1386	833.7477	$422.63	81	Cluster2	122.1385945
8	ID_744767	$421.01	75	28	251.7130	120.2825	835.3481	$421.01	75	Cluster2	120.2825117
9	ID_550162	$414.21	69	30	244.9179	113.5649	842.1536	$414.21	69	Cluster2	113.5648968
10	ID_472611	$406.85	83	51	237.7992	106.5500	849.5517	$406.85	83	Cluster2	106.5500194
11	ID_900197	$372.45	77	20	203.2012	71.8051	883.9102	$372.45	77	Cluster2	71.80512016
12	ID_158539	$348.24	73	53	178.9332	47.5096	908.1116	$348.24	73	Cluster2	47.50956181

Outlier Detection Techniques/Statistical Tools
Predictive Modeling



"The most that can be expected from any model is that it can supply a useful approximation to reality: All models are wrong; some models are useful." – George E.P. Box, British Statistician, 2005

Outlier Detection Techniques/Statistical Tools

Predictive Modeling – Overview

What is Predictive Modeling?

- Predictive modeling is a process through which a future outcome or behavior is predicted based on the historical data at hand
  - The probability of a provider joining the exclusion list - historical data is the exclusion list
  - The probability of a provider joining a list of providers to be investigated – the historical data is the list investigations
  - The probability of a provider (beneficiary) staying with medical group or health plan – the historical data is the list of providers who left

Why Predictive Modeling ?

- Preventing future fraud– cost saving to the programs
- Investigating providers before they can do more damage or commit more fraud
- Proactively initiating programs to retain providers or beneficiaries within the medical group or health plan

Outcome of Predictive Modeling?

- The goal is to determine the likelihood of the outcome – the higher the probability the more likely the outcome will occur
  - To be excluded from the programs
  - To be included in the investigation list
  - To stay with the medical group or health plan

Outlier Detection Techniques/Statistical Tools

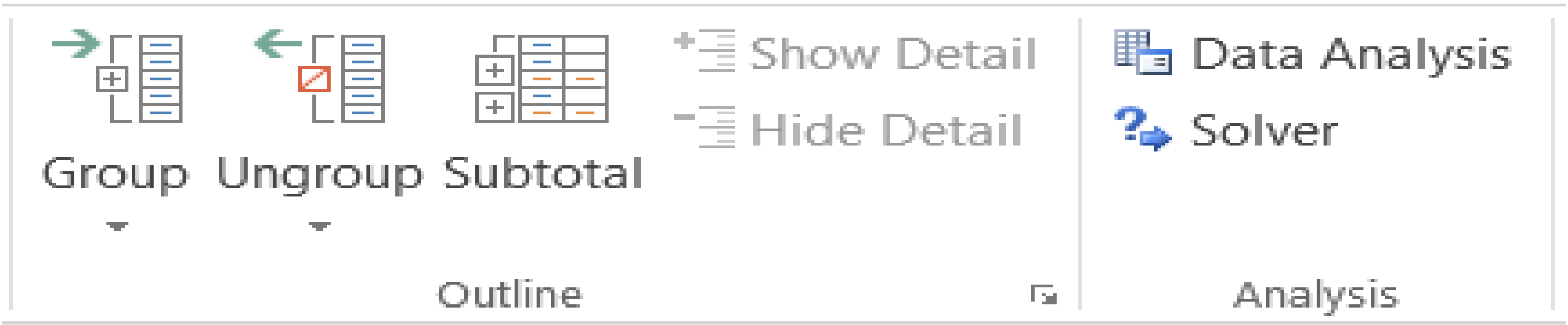
Predictive Modeling – Building Blocks

Terms commonly used in Predictive Modeling

- **Logistic Regression** – a predictive model used when the outcome is "Yes" or "No"
- **Training Dataset** - dataset that includes both historical and current data with clear distinction of the outcomes – coded 1 for "Yes" and 0 for "No"
- **Weights (Coefficients)** – numbers that express the importance of variables
- **P-values** – numbers that express the strength of association between the outcome and variables
- **Odds Ratio (OR)** – another way of expressing probability; 75% probability is equal to OR of 3
- **Log-Likelihood Algorithm** – algorithm that maximizes the likelihood of obtaining the observed data
- **Scoring Dataset** – new data on individuals/entities whose probabilities of outcomes will be computed

Excel Tools

- **Data Analysis Regression** – add-in tool available in Excel
- **Excel Solver** – add-in tool available in Excel



Outlier Detection Techniques/Statistical Tools

Predictive Modeling – Workflow

Step 1

Download PUF and LEIE data, run descriptive statistics, Identify variables and create the analytical file/training dataset

Step 2

Assign random probability numbers to the providers according to their original classification (“1” or “0”) and compute odds ratio

Step 3

Compute Odds Ratio and run Regression of Odds Ratio on the 5 variables of the analytical file

Step 4

Insert the estimated regression weights (coefficients) into the “Coefficients Table”

Step 5

Compute the Log-likelihood using the values in the “Coefficients Table”

Outlier Detection Techniques/Statistical Tools

Predictive Modeling – Workflow

Step 6

Run Excel Solver using as objective function the sum of the Log-likelihood. The Solver will find the weights (coefficients) that maximize the Log-likelihood – *the likelihood of obtaining the observed data*

Step 7

Create the model covariance matrix to test the significance of the logistic regression estimated weights

Step 8

Create a dataset of active non-excluded (PUF) providers including the relevant indicators to score their probability of being excluded using the weights generated by the logistic regression

Step 9

Score probabilities of non-excluded providers

Step 10

Review results and conduct additional analysis

Outlier Detection Techniques/Statistical Tools

Predictive Modeling – Excel Output

Provider ID	Payment Per Bene	Services Per Bene	Avg. Age of Benes	% of Benes with Diabetes	Avg. Health Risk Score of Benes	L	e to the power of L	Probability (Px)
ID_365774	$1,256.36	19	74	45.00%	1.4592	19.80	396,550,446.39	100.00%
ID_573433	$530.77	11	60	33.00%	1.1594	6.34	567.97	99.82%
ID_473715	$279.76	4	68	71.00%	1.3161	4.76	116.35	99.15%
ID_737783	$278.92	5	73	37.00%	0.7488	3.44	31.22	96.90%
ID_101716	$528.02	13	75	25.00%	0.8668	2.70	14.83	93.68%
ID_871960	$550.05	5	61	57.00%	2.5261	2.40	11.03	91.69%
ID_998581	$678.83	19	78	24.00%	0.948	0.89	2.44	70.96%
ID_158539	$348.24	8	73	25.00%	0.8751	0.51	1.67	62.56%
ID_306931	$286.01	8	73	75.00%	1.3195	0.48	1.62	61.85%
ID_124885	$217.29	5	69	52.00%	1.1407	0.40	1.50	59.93%
ID_638763	$296.67	5	75	26.00%	1.0391	0.39	1.47	59.52%

100 random providers were selected to estimate their probabilities of exclusion
The top 10 providers have at least 59% of exclusion probability

Outlier Detection Techniques/Statistical Tools

Predictive Modeling – Executing the Steps

Step 1: Training dataset

- The training dataset included
  - 18 GP physicians excluded from the Medicare program and
  - A sample of 72 physicians active in the program (non-excluded)
  - 5 variables
    - Payment Per Beneficiary
    - Number of Services Per Beneficiary
    - Average Age of Beneficiaries
    - Percentage of Beneficiaries with Diabetes
    - The average CMS-computed beneficiary health risk score

- The objective of using a training dataset is to assess the importance (weights) of the 5 variables in predicting the outcome of being excluded

Outlier Detection Techniques/Statistical Tools

Predictive Modeling – Step 1: Training Dataset

Step 1						
Provider ID	**EXCLUDED**	**Payment Per Bene**	**Services Per Bene**	**Avg. Age of Benes**	**% of Bene with Diabetes**	**Avg. Health Risk Score of Benes**
ID_331177	1	$627.41	14.9	75	40.00%	0.9924
ID_607721	1	$576.74	13.43	73	35.00%	1.0212
ID_728367	1	$836.23	10.3	71	37.00%	1.5547
ID_376426	1	$287.10	8.52	74	75.00%	1.1845
ID_136308	1	$558.54	8.5	73	47.00%	1.1187
ID_229648	0	$1,767.16	67.83	73	20.00%	1.2638
ID_277298	0	$447.73	15.93	71	38.00%	1.0345
ID_192713	0	$262.48	10.48	74	19.00%	1.146
ID_572197	0	$219.62	9.48	70	34.00%	1.015
ID_374411	0	$446.12	7.93	73	50.00%	1.6322

Outlier Detection Techniques/Statistical Tools

Predictive Modeling - Executing the Steps

Steps 2-6: Intermediary steps

- **Step 2:** Probabilities between 50-99% were assigned to excluded providers; and probabilities between 1-49% were assigned to non-excluded providers
- **Step 3:** The Odds Ratios were computed for each of the providers
- **Step 4:** Regression weights (coefficients) were estimated
- **Step 5:** Using the initial weights computed in Step 4, the Log-Likelihood was generated for each of the providers
- **Step 6:** The Excel Solver was run using the sum of the Log-likelihood and the initial weights to generate the final weights for each of the variables

Step 7: P-Value

- The P-value of each variable was computed to assess the strength of the association between the variable and the outcome
- The lower the p-value the stronger the association between variables
- Standard rule is to consider p-values <= 5% as indicative of statistically significant association

Outlier Detection Techniques/Statistical Tools

Predictive Modeling - Step 6: Final Weights

Solver Options

Max Time Unlimited, Iterations Unlimited, Precision 0.000001
Convergence 0.0001, Population Size 100, Random Seed 0, Derivatives Central
Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 1%

Objective Cell (Max)

Cell	Name	Original Value	Final Value
U	Log(Maximum Likelihood)	-37.00302969	-17.56488668

Variable Cells

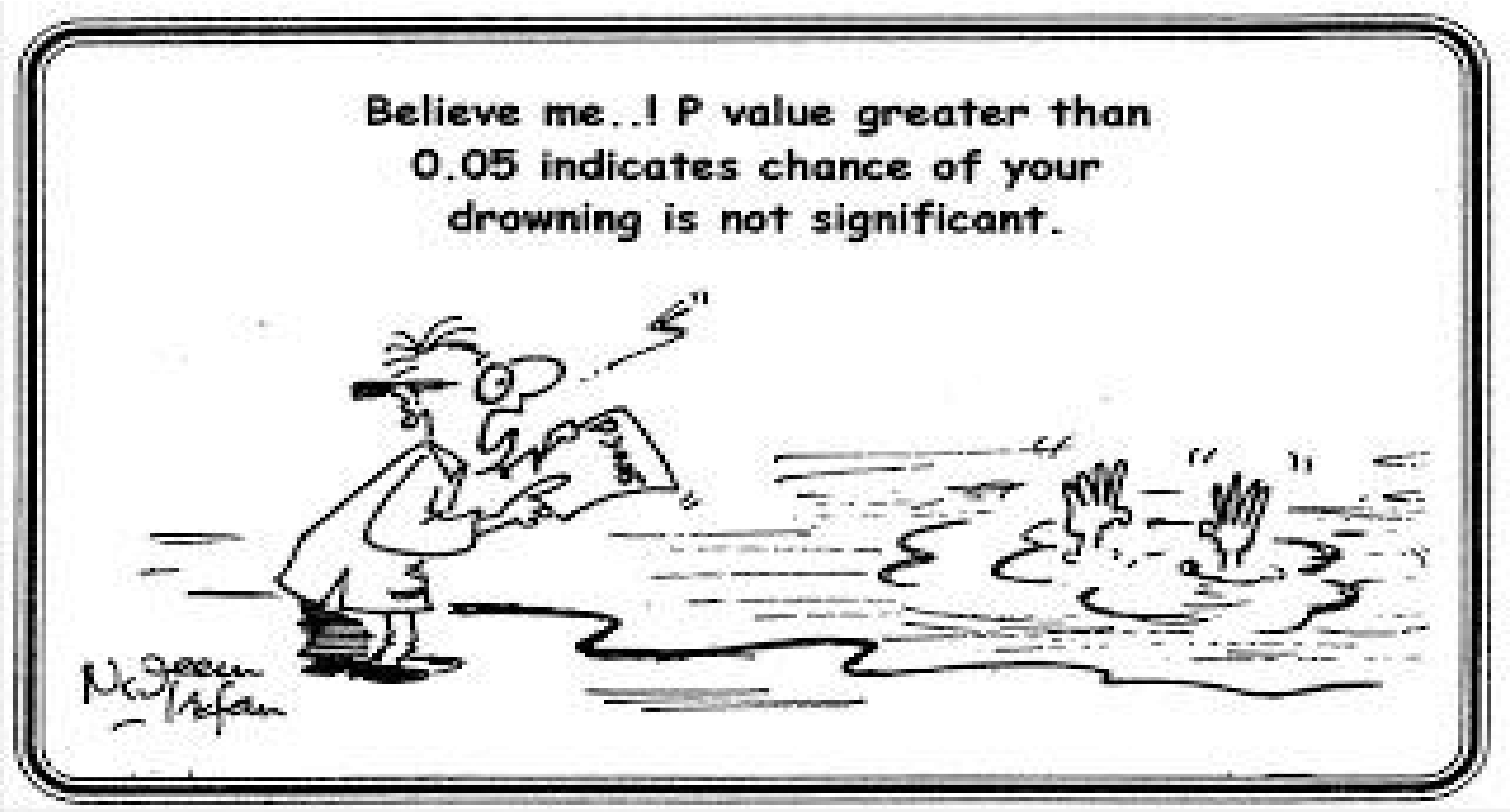
Cell	Name	Original Value	Final Value	Integer
U	Intercept	2.5150	17.7796	Contin
U	Payment Per Bene	0.0048	0.0359	Contin
U	Services Per Bene	-0.1142	-0.9372	Contin
U	Avg. Age of Benes	-0.0498	-0.2324	Contin
U	% of Bene with Diabetes	2.6324	11.1129	Contin
U	Avg. Health Risk Score of Benes	-0.8780	-8.8086	Contin

Outlier Detection Techniques/Statistical Tools

Predictive Modeling – Step 7: P- values

- P-value is calculated by:
  - The value of Wald = $(Coef/STD\ DEV)^2$
  - The value of P-value =CHISQ.DIST.RT(Wald,1), where CHISQ.DIST.RT is an Excel statistical function
  - The p-value of all the variable are less than 0.05

122		Coef	VAR	STD DEV	Wald	p-value
123	**Intercept**	17.7796	59.3103	=SQRT(C123)		0.0210
124	**Payment Per Bene**	0.0359	0.0002	0.0126	8.1835	0.0042
125	**Services Per Bene**	-0.9372	0.1892	0.4350	4.6418	0.0312
126	**Avg. Age of Benes**	-0.2324	0.0110	0.1050	4.9017	0.0268
127	**% of Bene with Diabetes**	11.1129	17.6422	4.2003	7.0001	0.0082
128	**Avg. Health Risk Score of Benes**	-8.8086	8.1025	2.8465	9.5763	0.0020


Believe me..! P value greater than 0.05 indicates chance of your drowning is not significant.

Outlier Detection Techniques/Statistical Tools

Predictive Modeling - Step 8: Scoring

New data of 100 non-excluded providers are scored using the weights

Provider ID	Payment Per Bene	Services Per Bene	Inv. Age of Benes	% of Benes with Diabetes	Inv Health Risk Score of Benes	L	e to the power of L	Probability (Px)
ID_365774	$1,256.36	19	74	45.00%	1.4592	19.80	396,550,446.39	100.00%
ID_573433	$530.77	11	60	33.00%	1.1594	6.34	567.97	99.82%
ID_473715	$279.76	4	68	71.00%	1.3161	4.76	116.35	99.15%
ID_737783	$278.92	5	73	37.00%	0.7488	3.44	31.22	96.90%
ID_101716	$528.02	13	75	25.00%	0.8668	2.70	14.83	93.68%
ID_871960	$550.05	5	61	57.00%	2.5261	2.40	11.03	91.69%
ID_998581	$678.83	19	78	24.00%	0.948	0.89	2.44	70.96%
ID_158539	$348.24	8	73	25.00%	0.8751	0.51	1.67	62.56%
ID_306931	$286.01	8	73	75.00%	1.3195	0.48	1.62	61.85%
ID_124885	$217.29	5	69	52.00%	1.1407	0.40	1.50	59.93%
ID_638763	$296.67	5	75	26.00%	1.0391	0.39	1.47	59.52%

Outlier Detection Techniques/Statistical Tools

Results From The Outlier Detection Techniques

Eight out of the 10 providers were also in the top 10 in more than one statistical tool

- One provider was in the top 10 in all the statistical tools
- Four providers were in the top 10 in 5 of the statistical tools
- Two providers were in the top 10 in 4 of the statistical tools
- One provider was an outlier in 2 statistical tool

Two providers were in the top 10 in the predictive modeling only

Provider ID	Excel Ranking	Z – Score	Composite Ranking (Z- score)	Box-Plot	Cluster Analysis	Predictive Modeling	Number of Hits by Detection Tools
ID_365774	9	1	1	2	C3	100.00%	6
ID_573433	1	4	2	1		99.82%	5
ID_473715	8		9		C2	99.15%	4
ID_737783						96.90%	1
ID_101716		5	7	5	C2	93.68%	5
ID_871960	4	3	6	4		91.69%	5
ID_998581		2	3	3		70.96%	4
ID_158539						62.56%	1
ID_306931	5		5	7	C2	61.85%	5
ID_124885	7					59.52%	2

Sampling and Extrapolation

Statistical Sampling Methods

Why sampling?

- Limited amount of available audit resources makes unfeasible reviewing 100 percent of the items of a population
- Statistically valid random samples allow for the extrapolation of the sample audit results to the whole population
- Typical goals of sampling in health care programs include:
  - Checking if enrollment application procedures or eligibility status processes are complying with regulatory requirements (auditors record results as yes or no)
  - determination of the possible existence of claim overpayments (auditors record results in dollars amount)

Sampling requirements

- The goal of the statistical requirements is to make sure that the sample is representative of the larger group
- The methodology does not need to be optimal as long as it is statistically valid - having a scientific basis with reference to regulatory guidance.
- The existence of multiple valid sample plans allows the auditor to choose the designs that less demanding in audit resources
- Proper documentation of the whole process is essential to make it fully replicable

Main types of audit-oriented sampling

- Attribute sampling – the goal is to find the *proportion of items* in the sample that meet a specified set of criteria and then estimate the number of population items in error
- Variable sampling – the goal is to determine the *dollar amount of billing errors* in the sample and then estimate the total dollar value of the errors made

Statistical Sampling Methods

Types of sample design frequently used in auditing

- Simple random sampling – involves the random selection of data from the entire population so that each possible sample is equally likely to occur
- Stratified random sampling – divides the population into smaller groups (strata) of similar characteristics and selects random sample from within each group
- If the technical statistical parameters are the same then the stratified random sampling approach saves audit resources as it requires smaller sample sizes than the simple random sampling method

Extrapolation

- Type of extrapolation
  - Error rate
  - Overpayment amounts
- The results of the sample review (audit) may or may not justify extrapolation
- The presence of "sustained or high level of payment error" in billing transactions justifies performing sampling to estimate the total dollar amount of billing errors (language taken from the Program Integrity Manual of CMS, Section 8.4.1.2)
  - Extrapolation is not justified if the error rate is low – in this case the recoupment is limited to the overpayment found in the sample

Statistical Sampling Methods

Software requirements to perform sampling

- Random number generator with the option of retaining seed numbers to make the process replicable
- Availability of key statistical distributions

Software packages/platforms to implement sampling

- **Microsoft EXCEL**
  - Includes available functions to retrieve information from embedded statistical distribution tables to be included in formulas that compute sample size and allocation of overall sample size across strata
- **RAT-STATS (OIG)**
  - Includes diverse sampling and extrapolation options by means of dropdown windows
  - Requires stratum boundary information to be fed into the system
- **SAS, SPSS, R** and other statistical packages
  - Allow for programming of procedures necessary to implement sampling and extrapolation procedures
  - Can incorporate program code to identify stratum boundaries
- **GLȲD(Σ)™**
  - Allows for the implementation of sampling or extrapolation processes without the need of coding
  - Includes the identification of stratum boundaries without the need of coding

Reference

- Dreiseitl, S. and L. Ohno-Machado, “Logistic Regression and Artificial Neural Network Classification: A Methodological Review”, Journal of Biomedical Informatics 35, 2002, 352:359 : http://www.sciencedirect.com/science/article/pii/S1532046403000340
- Hastie, T., R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009
- King, G. and L. Zeng, “Logistic Regression in Rare Events Data.” Political Analysis, 9, 2001, 137–163, Spring. Copy at http://j.mp/lBZoli
- King, G., “Big Data Is Not About The Data!” In Computational Social Science: Discovery and Prediction, ed. By M. Alvarez, Cambridge University Press, 2016: http://gking.harvard.edu/files/gking/files/prefaceorbigdataisnotaboutthedata_1.pdf
- Larsen, R., “Notes on Matrix Operations in Excel” (excerpt from Engineering with Excel by Ronald Larsen): http://www.eng.auburn.edu/~clemept/CEANALYSIS_SPRING2011/matrixoperations_notes.pdf
- Macedo, P. and C. Dorfschmid Statistics: “Friend or Foe? The Compliance Officer’s Perspective”, Journal of Health Care Compliance, Volume 14, Number 1, January-February 2012
- Zaointz, C., “Significance Testing of the Logistic Regression Coefficients” : http://www.real-statistics.com/logistic-regression/significance-testing-logistic-regression-coefficients/
- H. Aravind, C. Rajgopal and K.P. Soman, “A Simple Approach to Clustering in Excel”, International Journal of Computer Applications (0975 – 8887) Volume 11– No.7, December 2010

Thank you!

For inquiries, please contact us at:

info@integritym.com

703-683.9600

Integrity Management Services