Frequency Response Standard Field Trial Analysis

Howard F. Illian, September 7, 2012

Executive Summary:

Frequency Response Standard Field Trial data analysis leads to three important conclusions.

Analysis indicates that a single event based compliance measure is unsuitable for compliance evaluation when based on data that has the large degree of variability demonstrated by the field trial. Only three out of 19 BAs would be compliant for all events with a standard based on a single event measure on the Western Interconnection. Only one out of 31 BAs would be compliant for all events with a standard based on a single event measure for the Eastern Interconnection. The general consensus of the industry is there is not a reliability issue with insufficient Frequency Response on any of the North American Interconnections at this time. Therefore, it is unreasonable to even consider a standard that would indicate over 90% of the BAs in North American to be non-compliant with respect to maintaining sufficient Frequency Response to support adequate reliability.

Analysis confirms that the sample size selected is sufficient to stabilize the result and alleviate the perceived problem associated with outliers. BAs with large measurement variation had enough samples to mitigate the risk associated with outliers. This demonstrates the sample size chosen (20 to 25 events) is sufficient to stabilize all three methods of measuring FRM. Therefore, it can be concluded that none of the methods are unduly influenced by outliers and the selection of the measurement method should be based on other factors.

Linear regression is the preferred method to use as the basis for the Frequency Response Measure. During evaluation of the results, the graphs showed that regression provides a higher estimate of FRM than the median. A comparison was made between the FRM as measured by the median and the FRM as measured by the regression. The results of that analysis reveal the regression shows a performance for all samples that is 8.7% of their FRO higher than is the median's performance on the Eastern Interconnection and 11.7% of their FRO higher than the median's performance on the Western Interconnection. In an unbiased analysis, one would expect that the median and regression to yield the same result. Therefore, this indicates that there is a statistical bias affecting the results of the analysis.

The statistical bias causing the difference between the median and regression results is explained by an attribute of Frequency Response. As the frequency deviation increases for larger frequency Disturbance events, the Frequency Response also increases. In simple terms, the regression includes the effect of this non-linear attribute and the median does not. The median underestimates the FRM. It cannot evaluate this non-linear attribute correctly. Regression is the only measurement method that captures the non-linear Frequency Response.

Introduction:

This paper presents the first evaluation of extensive data developed from the standardized methods developed by the Resources Subcommittee (RS), the Frequency Working Group (FWG), the Frequency Response Standard Drafting Team (FRSDF) and NERC Staff.

This paper provides the first statistical analysis and evaluation on field trial data with similar sample sizes to those specified in the draft Standard BAL-003-1 Frequency Response and Frequency Bias Setting and answers three critical questions for the FRSDT.

1. Should compliance be based upon a single event measure?
2. Is a sample size of at least 20 events sufficient to provide stable results?
3. Is Median, Mean or Regression the best method for determination of a Frequency Response Measure (FRM) for use in compliance evaluation?

Data Preparation:

This report required extensive data preparation to perform the analysis upon which to base substantive conclusions. Three areas of data collection and preparation were required.

BA Data:

Sixty of the BAs on the Eastern and Western Interconnections provided data on the FRS Form1 for 2011. The analysis was not performed for either of the single BA interconnections, ERCOT or Quebec. Of the 60 BAs that provided data, only 50 provided data of sufficient quality to be used in the analysis. BAs that were excluded provided frequency data that was either obviously incorrect (i.e. frequency data in Hertz instead of change in Hertz) or frequency data that was uncorrelated to the interconnection measured frequency.

Normalization:

Since the data provided by the BAs is confidential, the BA data was normalized to hide the identity of individual BAs. This normalization was performed by dividing the change in actual net interchange by the Frequency Response Obligation (FRO) for each BA. This normalization converts all of the data from the actual Frequency Response of the BA to a per unit Frequency Response value where 1.0 indicates that the Frequency Response is equal to the BA's FRO.

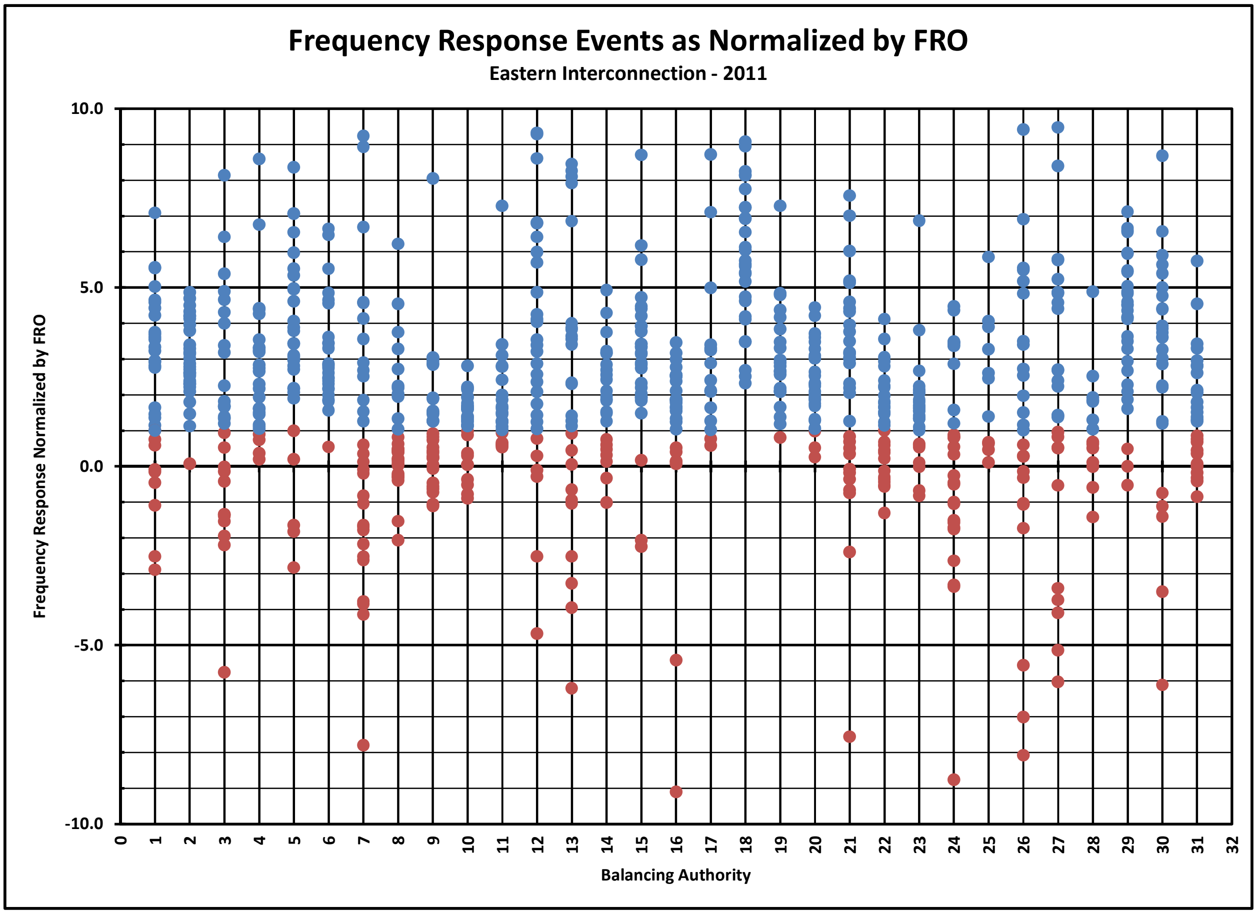
This normalization process required the development of the some of the data that would appear on the equivalent of the CPS2 Bounds Report as it would appear under this revised standard. The required data was extracted from the FERC Form No. 714 Reports for the year 2009. The data was estimated for those BAs that did not submit 714 Reports. The equivalent data was estimated based on other sources. The validity of this statistical analysis is not dependent upon the accuracy of the FRO estimates. It is only necessary for these estimates to be close to the actual values for firm conclusions to be drawn and to put the results in the proper context.

Once the FROs were estimated for all of the BAs on the Eastern and Western Interconnections, they were transcribed onto the FRS Form1s for each BA included in the analysis. The final step was to write VBA programs to automate the evaluation of the field trial data. This completed the data preparation required.

Single Event Compliance:

The variability of the measurement of Frequency Response for an individual BA for an individual Disturbance event was evaluated to determine its suitability for use as a compliance measure. The individual Disturbance events were normalized and plotted for each BA on the Eastern and Western Interconnections. This data was plotted with a dot representing each event. Events with a measured Frequency Response above the FRO were shown as blue dots and events with a measured Frequency Response below the FRO were shown as red dots. In order to show the full variability of the results the plots have been provide with two scales, a large scale to show all of the events and small scale to show the events closer to the FRO or a value of 1.0. Appendix 1 shows these Frequency Response Events as Normalized by FRO. One of these graphs for the Eastern Interconnection is shown below.

Analysis of this data indicates a single event based compliance measure is unsuitable for compliance evaluation when the data has the large degree of variability shown in the charts in Appendix 1. Based on the field trial data provided, only three out of 19 BAs on the Western Interconnection would be compliant for all events with a standard based on a single event measure. Only one out of 31 BAs on the Eastern Interconnection would be compliant for all events with a standard based on a single event measure. The general consensus of the industry is that there is not a reliability issue with insufficient Frequency Response on any of the North American Interconnections at this time. Therefore, it is unreasonable to even consider a standard that would indicate over 90% of the BAs in North American to be non-compliant with respect to maintaining sufficient Frequency Response to maintain adequate reliability.



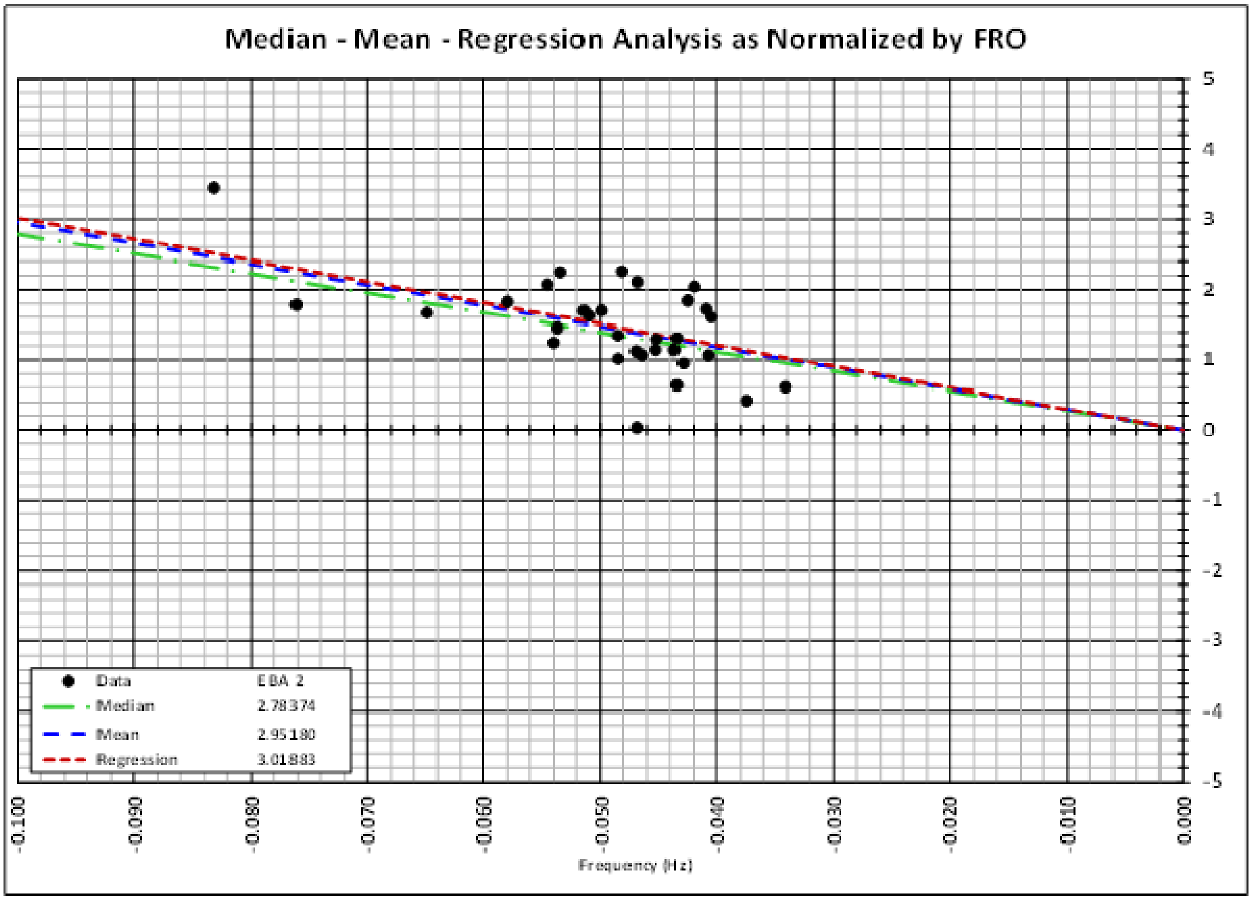
Event Sample Size:

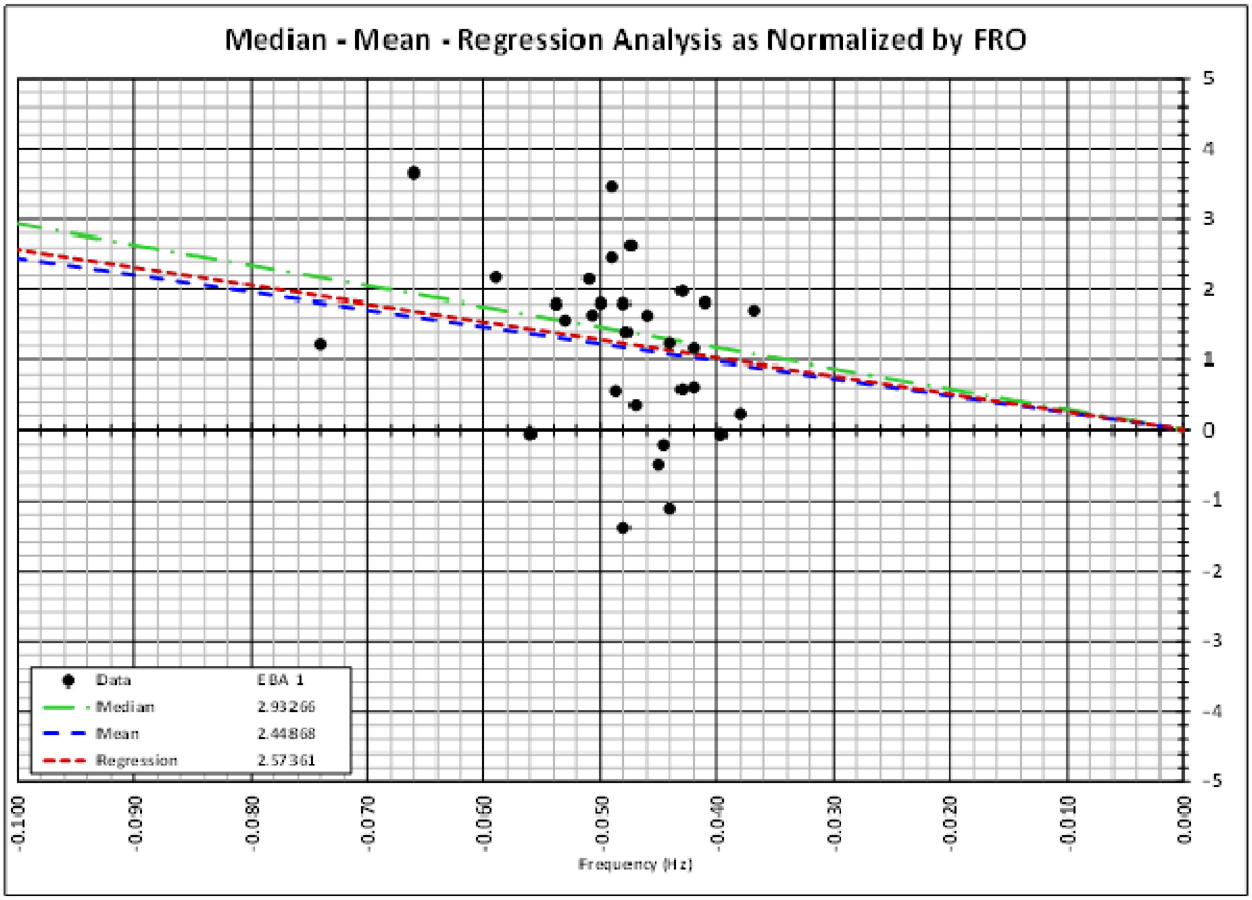
Previous studies recommended a sample size sufficient to provide a stable measure of Frequency Response of 20 to 25 events. These previous studies were performed on limited data and a limited number of BAs. The field trial data set allows conclusions to be drawn with respect to the sample size specified for FRM calculation in the draft standard. Field trial data analysis indicates whether or not the sample sizes specified would provide a stable result.

Median, Mean or Regression Results:

Field trial analysis also answers the question of which of the candidate measures of FRM would perform best when applied in a compliance environment. Since the questions related to sample size and method of measurement are similar, both can be answered with a single study.

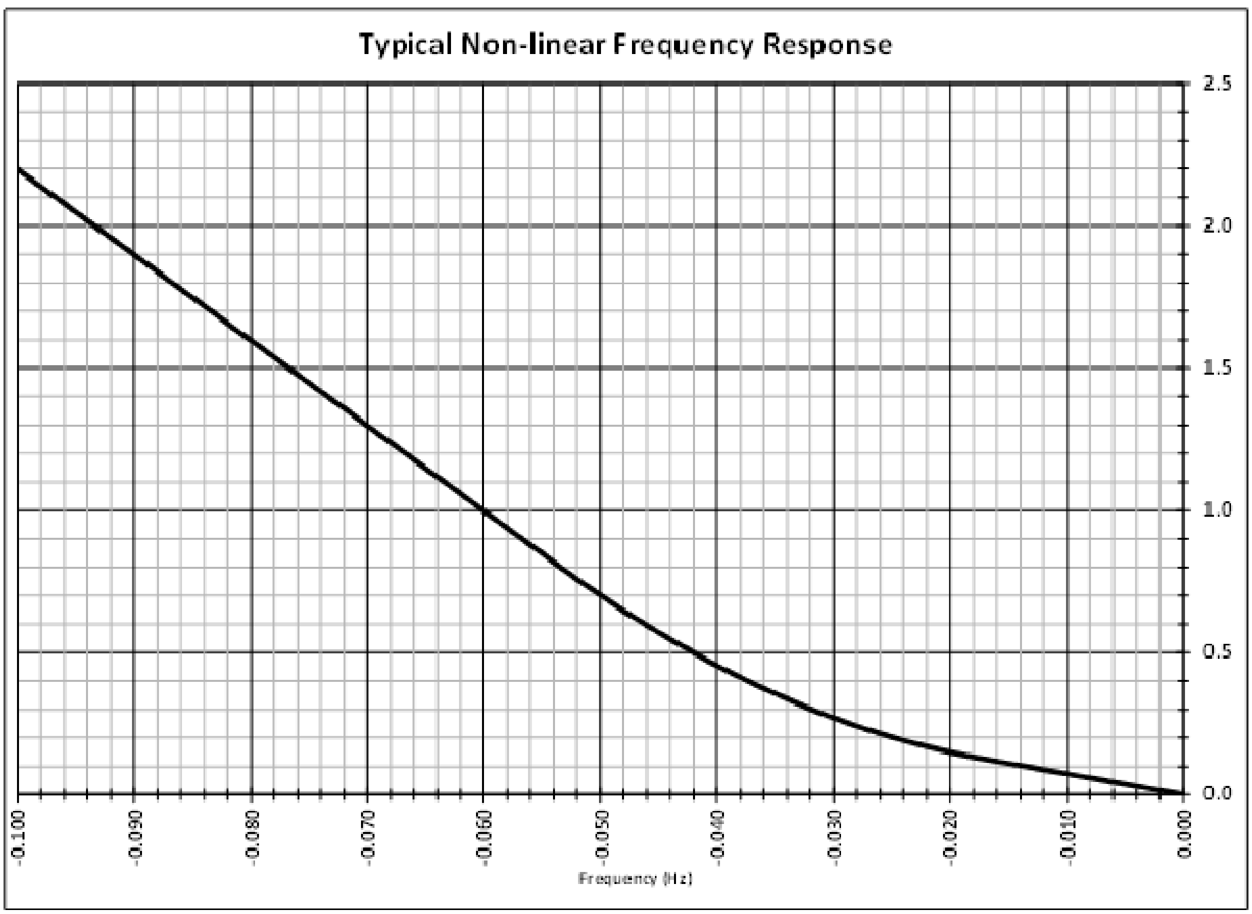
All of the normalized data were evaluated using all three candidate methods for measuring FRM. Appendix 2 presents the series of graphs indicating results for each BA. Each graph shows all of the individual data points use to determine the median, mean and regression lines. The median line is green, the mean line is blue and the regression line is red. The value of the Normalized Frequency Response (vertical axis) where the line intercepts the value of frequency (horizontal axis) at a value of 0.1 Hz indicates compliance. Values above 1.0 indicate a FRM above the FRO and values below 1.0 indicate a FRM below the FRO. Two graphs from Appendix 2 are shown here. The first is a graph with a small degree of variability in the measured Frequency Response for each individual event. The second is a graph with a large degree of variability in the measured Frequency Response for each individual event.


Median - Mean - Regression Analysis as Normalized by FRO
Data EBA 2
Median 2.78374
Mean 2.95180
Regression 3.01883
Frequency (Hz)


Median - Mean - Regression Analysis as Normalized by FRO
Data EBA 1
Median 2.93266
Mean 2.44868
Regression 2.57361
Frequency (Hz)

Review of these graphs indicates that the outlier problem, as previously described in the draft background document, did not present itself. There were no BAs that had a small degree of variability in the measured single event Frequency Response for most of the events with a few outliers. The variability appeared similar for all events for each BA indicating that the sample size of 20 to 25 events is sufficient to stabilize the result and eliminate any undue influence from potential outliers. In those BAs with large variations in measured single event response, the sample size was sufficient to collect enough samples that no single outliers unduly influenced the result as was feared. BAs with large measurement variation still had enough samples to mitigate the risk associated with outliers. This demonstrates that the sample size chosen is sufficient to stabilize all three methods of measuring FRM. Therefore, it can be concluded that none of the methods are unduly influenced by outliers and the selection of the measurement method should be based on other factors.

During evaluation, the graphs appeared to show that the regression provided a higher estimate of FRM than the median. Consequently, a comparison was made between the FRM as measured by the median and the FRM as measured by the regression. The results of that analysis reveal the regression shows a performance for all samples that is 0.087% of their FRO higher than is the median's performance on the Eastern Interconnection and 0.117% of their FRO higher than the median's performance on the Western Interconnection. In an unbiased analysis, one would expect the median and regression would yield the same result. Therefore, this would indicate there is some unknown statistical bias affecting the results of the analysis.



The bias causing the difference between the median and regression results can be explained by an attribute of Frequency Response. As the frequency deviation increases for larger Disturbance events, the Frequency Response increases, but it does so disproportionately. This is shown in the Typical Non-linear Frequency Response graph above. This attribute of

Frequency Response has been demonstrated in technical papers.[1,2] It has also been implemented in the variable Frequency Bias Settings used by ERCOT, BPA and BC Hydro. In simple terms, the regression includes the effect of this non-linear attribute and the median does not. The regression readily accommodates the disproportionality on the slope of the regression line. In this case the effect tends to be upward—ever bigger MWs per increment in size of larger frequency error. The median is biased against any disproportionate increase in response per increase in size of frequency error as part of the median's blindness to outliers. The median will give no credit for the ever growing amount of MWs deployed per added increment in size of frequency error. All the median does is count the number of your MW responses regardless of size and, to represent all the MW responses, it chooses the one that occurred half-way in the sequence of decreasingly negative and increasingly positive frequency errors. As a consequence, the median underestimates the FRM because it cannot evaluate the non-linear attribute correctly. It doesn't see or notice that attribute at all through its blinders exclusively of numerical order or placement in a sequence. Regression is the only measurement method that captures the non-linear Frequency Response correctly.

Median, Mean, Regression Descriptions:

Median is the numerical value separating the higher half of a one-dimensional sample, a one-dimensional population, or a one-dimensional probability distribution, from the lower half. The Median of a finite list of numbers is found by arranging all the observations from lowest value to highest value and picking the middle one. When the number of observations is even, there is no single middle value; the Median is arbitrarily defined as the mean of the two middle values.

In a sample of data, or a finite population, there may be no member of the sample whose value is identical to the Median (in the case of an even sample size), and, if there is such a member, there may be more than one so that the Median may not uniquely identify a sample member. Nonetheless, the value of the Median is uniquely determined with the usual definition. A Median is also a central point that minimizes the arithmetic mean of the absolute deviations. However, a Median need not be uniquely defined. Where exactly one Median exists, statisticians speak of "the Median" correctly; even when no unique Median exists, some statisticians speak of "the Median" informally.

The Median can be used as a measure of location when a distribution is skewed, when end-values are not known, or when one requires reduced importance to be attached to outliers, e.g., because they may be measurement errors. A Median-unbiased estimator minimizes the risk with respect to the absolute-deviation loss function, as observed by Laplace.[3] For continuous probability distributions, the difference between the Median and the Mean is never more than one standard deviation. Calculation of Medians is a popular technique in summary statistics and summarizing statistical data, since it is simple to understand and easy to calculate, while also giving a measure that is more robust in the presence of outlier values than is the Mean.

Mean is the numerical average of a one-dimensional sample, a one-dimensional population, or a one-dimensional probability distribution. A Mean-unbiased estimator minimizes the risk (expected loss or estimate error) with respect to the squared-error loss function, as observed by Gauss.[4] The Mean is more sensitive to outliers for the very reason that it is a better estimator; it minimizes the squared-error loss function.

---

[1] Hoffman, Stephen P., Frequency Response Characteristic Study for ComEd and the Eastern Interconnection, Proceedings of the American Power Conference, 1997.

[2] Kennedy, T., Hoyt, S. M., Abell, C. F., Variable, Non-linear Tie-line Frequency Bias for Interconnected Systems Control, IEEE Transactions on Power Systems, Vol. 3, No. 3, August 1988.

[3] An absolute-deviation loss function is used to minimize the risk of estimate error when dealing with uniform distributions. Appendix 3 provides a description of Uniform Distributions and a derivation of the Median.

[4] A squared-error loss function is used to minimize the risk when dealing with normal (Gaussian) distributions. Appendix 4 provides a description of normal (Gaussian) distributions and a derivation of the Mean.

Linear Regression is the linear average of a multi-dimensional sample, or a multi-dimensional population. A Linear Regression-unbiased estimator minimizes the risk (expected loss or estimate error) with respect to the squared-error loss function in multiple dimensions, as observed by Gauss.[5] The Linear Regression is also sensitive to outliers for the very reason that it is a better estimator; it minimizes the squared-error loss function.

Important Considerations:

The following issues have been raised as important to consider with respect to the selection of the best method for measuring Frequency Response.

Two dimensional measurement of Frequency Response provides the best representation of the change in MWs divided by the change in frequency and is used to estimate the Frequency Bias Setting which indicates the Frequency Response in MWs provided at actual frequency as compared to scheduled frequency.

The Non-linear attribute of Frequency Response has been demonstrated on all of the North American interconnections and is an important consideration in the representation of Frequency Response.

A single best estimator of Frequency Response is a necessary result for use in compliance evaluation.

A linear system[6] is assumed in the development of the individual Frequency Response Obligation for each BA on a multiple BA interconnection and is used to distribute the Interconnection Frequency Response Obligation among the BAs on that interconnection. If the system within which it has been developed and measured is a Non-linear System,[7] then the conclusion that, "If all BAs provide their Frequency Response Obligation, the interconnection will achieve its total required frequency Response cannot be logically concluded.

Bi-modal distributions occur whenever a reconfiguration of BAs occurs within a compliance year. Unless the method chosen can correctly represent bi-modal distributions, reconfigured BAs cannot be effectively measured for compliance.

Quality statistics should be available for use in compliance evaluation. The measure of Frequency Response is used to determine compliance with minimum provision of the BAs obligation for providing its share of Frequency Response for the interconnection. Using a measure for compliance includes with it the responsibility of assuring that the measure also provides a reasonable level of confidence that it is a fair representation of the BA's performance. There is still a presumption that an indication of non-compliance should not occur due to pure chance.

Reducing influence of noise in the data is considered an important attribute in the measurement method. All measurements of Frequency Response will be affected by noise in the measurement process.

Reducing influence of outliers in the data is considered the most important attribute in the measurement method. All measurements of Frequency Response will be affected by true outliers. The risk associated with the reduction in the influence of outliers is that valid information about the measure is also lost when an outlier reduction method is used.

Ease of calculation is an important consideration.

A familiar indicator is important for communication with the industry.

[5] Appendix 5 provides a derivation of the Linear Regression.
[6] A Linear System is a system in which the sum of the parts is equal to the whole.
[7] A Non-linear System is a system in which the sum of the parts is not equal to the whole.

The advantages of each method of measurement are presented in Table 1 – Median, Mean & Regression Comparison shown below.

Table 1 – Median, Mean & Regression Comparison

FRM Measurement Method	Median	Mean	Regression
Provides two dimensional measurement	[8]	[8]	Y
Represents non-linear attributes	[9]	[9]	Y
Provides a single best estimator (single value)	[10]	Y	Y
Is part of a linear system		Y	Y
Represents bi-modal distributions	[11]	Y	Y
Quality statistics available	[12]	Y	Y
Reducing influence of noise	Y[13]		P[14]
Reducing influence of outliers	Y		P[15]
Easy to calculate	Y	Y	[16]
Familiar indicator	Y	Y[17]	
Currently used as the measure in BAL-003		Y	

Y = Yes **P = Partial**

[8] Neither Median nor Mean can evaluate the two dimensional nature of Frequency Response.

[9] Neither Median nor Mean can capture the Non-linear attribute of Frequency Response and both underestimate the typical non-linear Frequency Response.

[10] Median is arbitrarily defined as the average of the two central values when there is an even number of values in the data set. The decision to further constrain this central range of values to a single value that is the average of the ends of that range is unsupported by any mathematical construct. It is only the desire of those looking for simplicity in the result that supports this singular definition of Median.

[11] The Median fails to provide a valid estimate of Frequency Response when the distribution of frequency event responses is bi-modal due to Balancing Authority reconfiguration or changes in responsibility for control such as partial period Overlap of Supplemental Control.

[12] The Median fails to provide any methods to determine the quality, significance or confidence associated with the measure.

[13] The Median reduces the influence of noise in the data, but that noise reduction comes with the cost of eliminating the availability of any quality statistics.

[14] Linear Regression provides a result that weights the data according to the change in frequency. Since the noise in the data is independent of change in frequency, Linear Regression provides a superior method for reducing the influence of noise in the resulting estimate of Frequency Response.

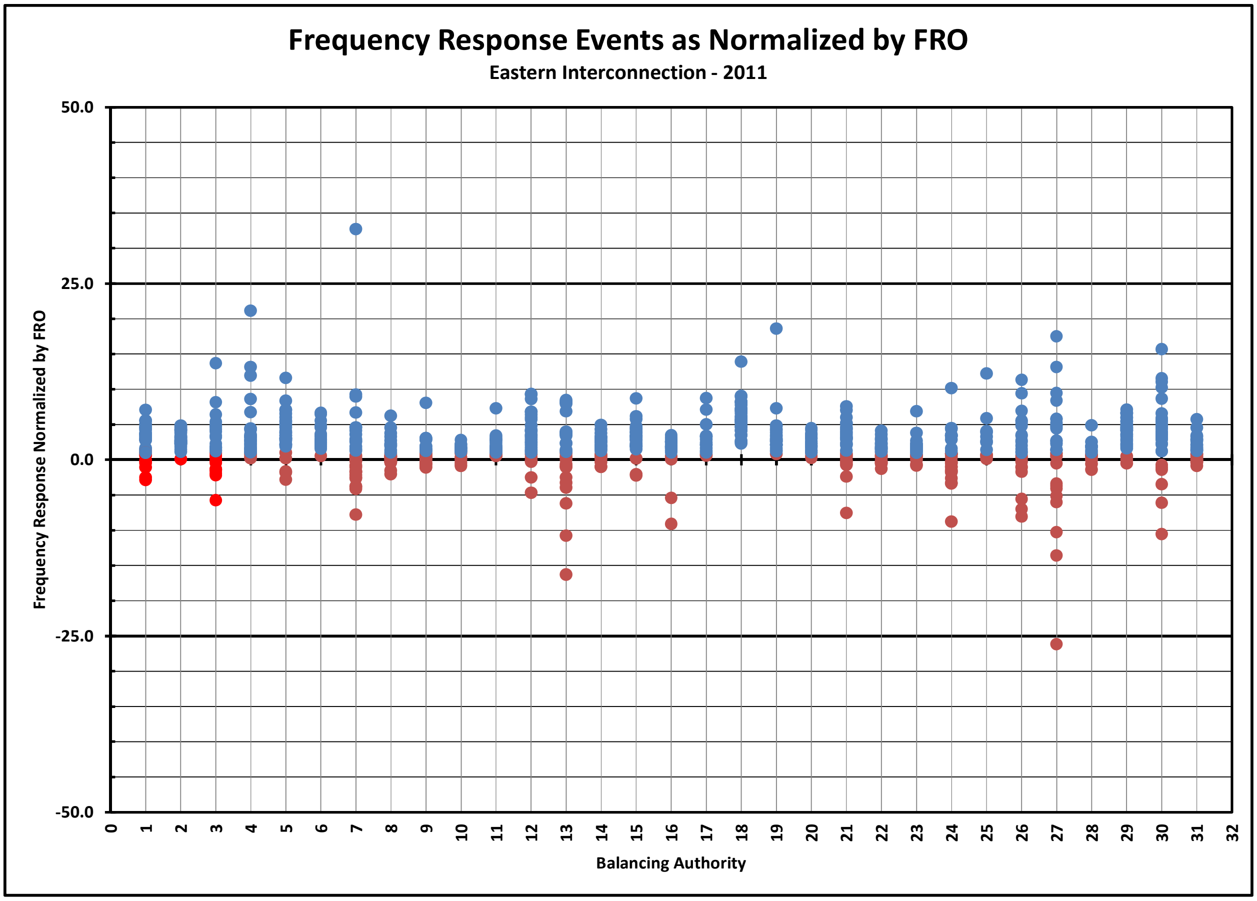
[15] Linear Regression is less sensitive to outliers and large data errors than the Mean.

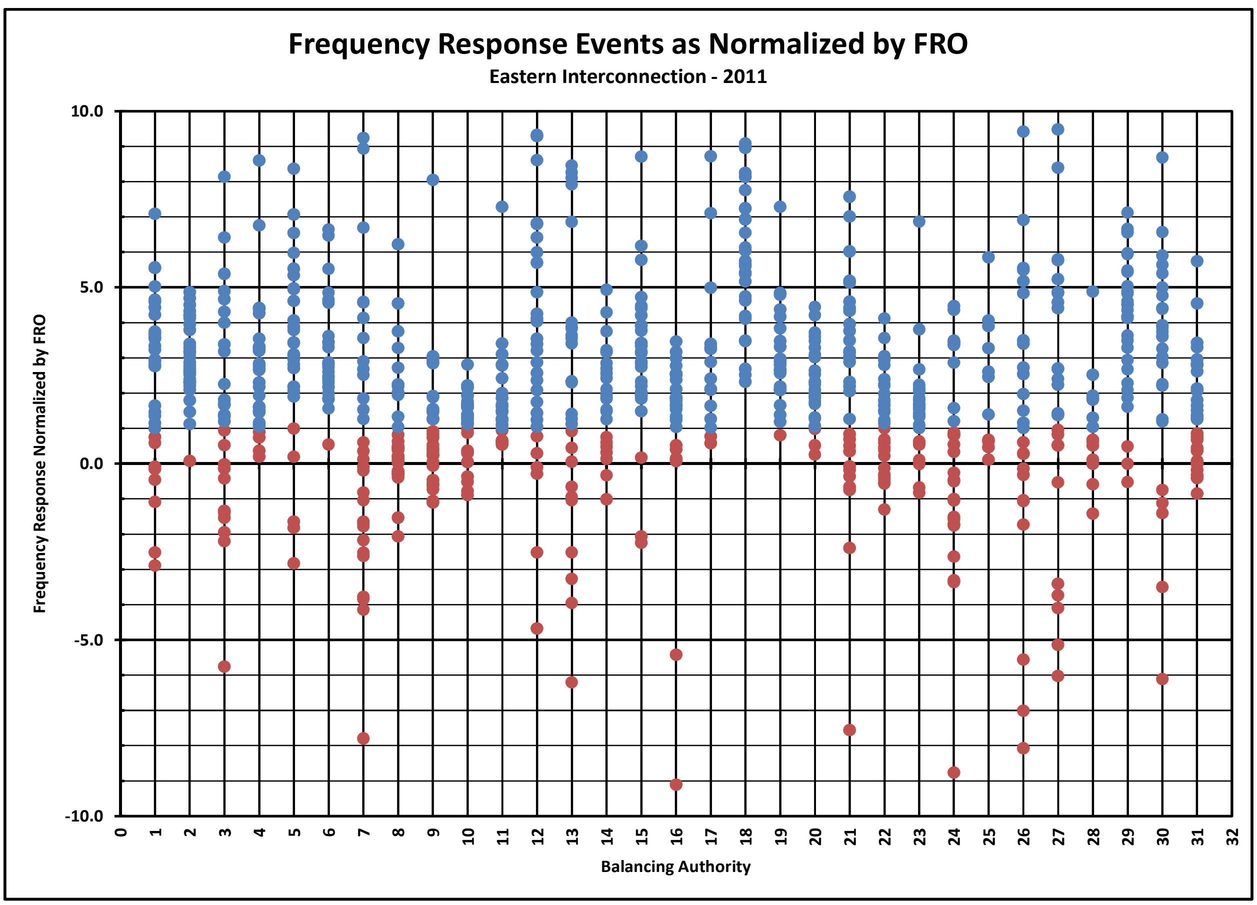
[16] Linear Regression is more complex and requires more effort to calculate, but that additional effort is small when the evaluation process has been automated.

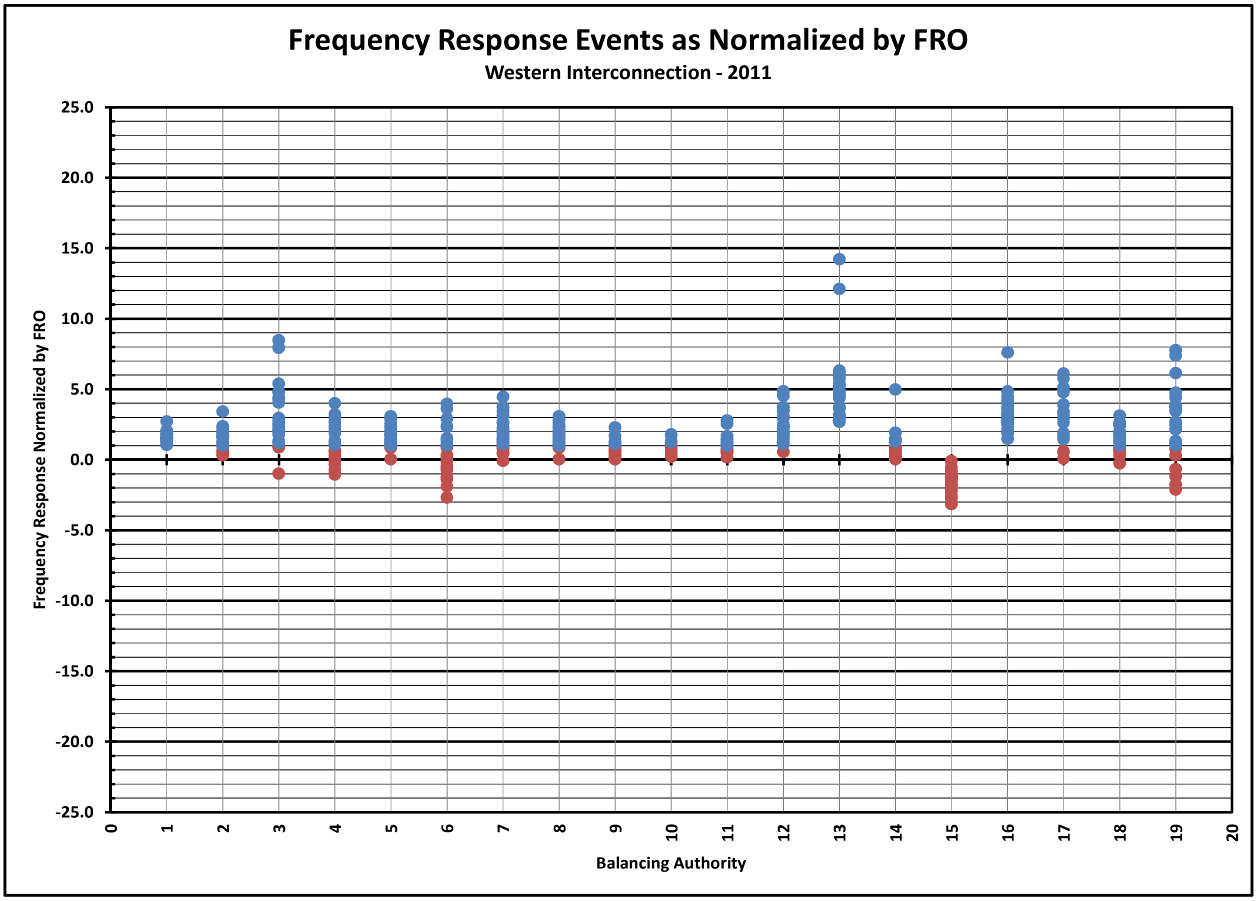
[17] Mean is currently used as the measure in BAL-003.

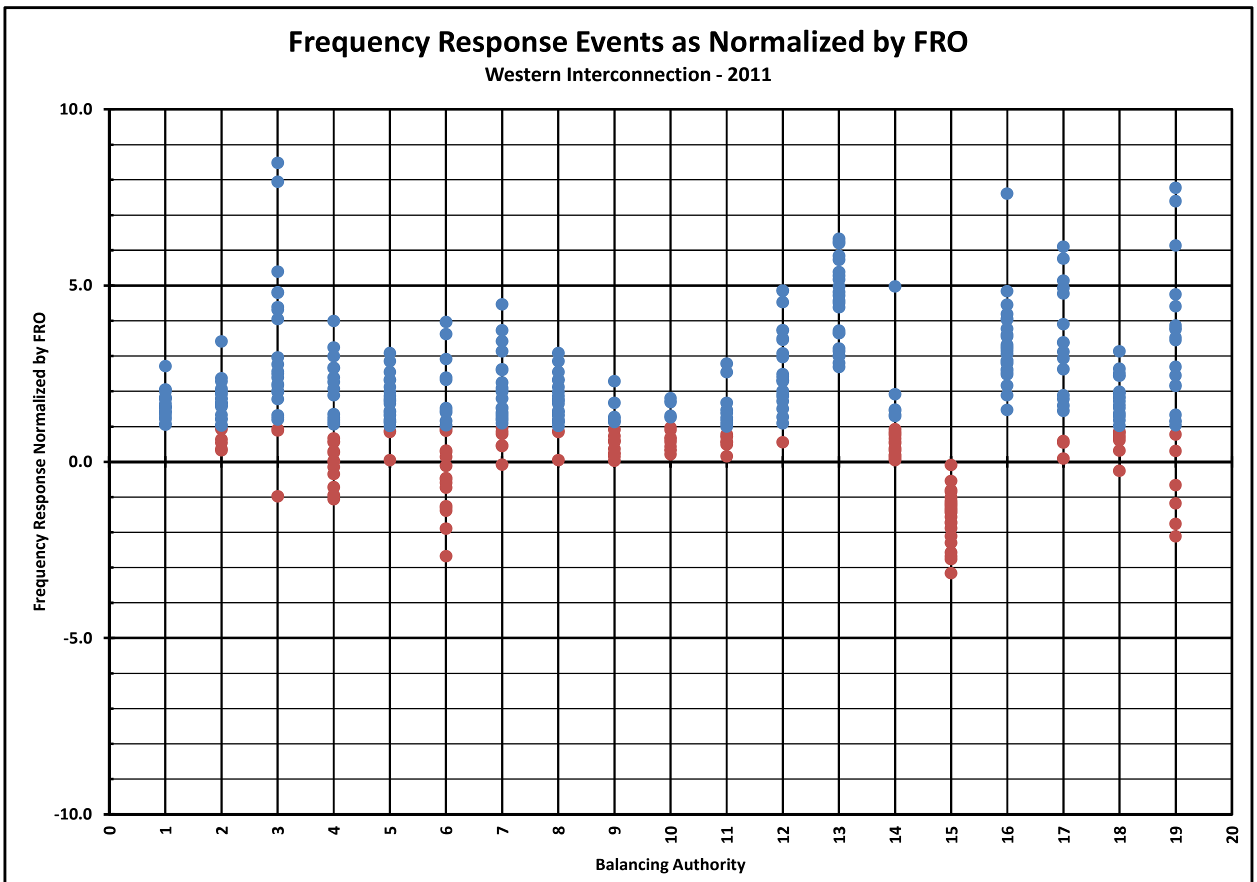
Recommendation:

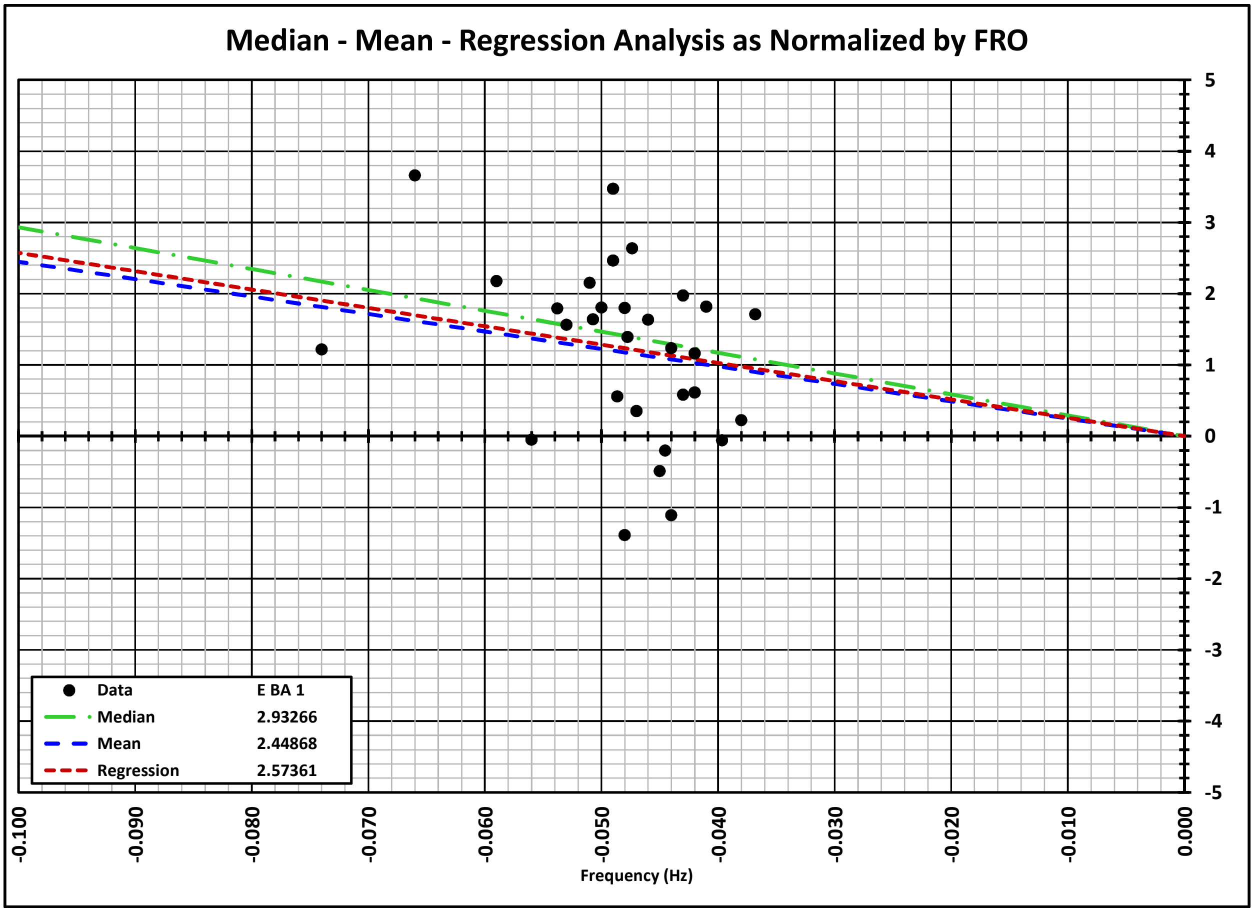
Based on the results of the above analysis, when one considers the mitigating effect of the sample size with respect to outliers, one concludes that linear regression is the preferred method to use as the basis for the Frequency Response Measure.


Frequency Response Events as Normalized by FRO
Eastern Interconnection - 2011
50.0
25.0
0.0
-25.0
-50.0
Frequency Response Normalized by FRO
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Balancing Authority

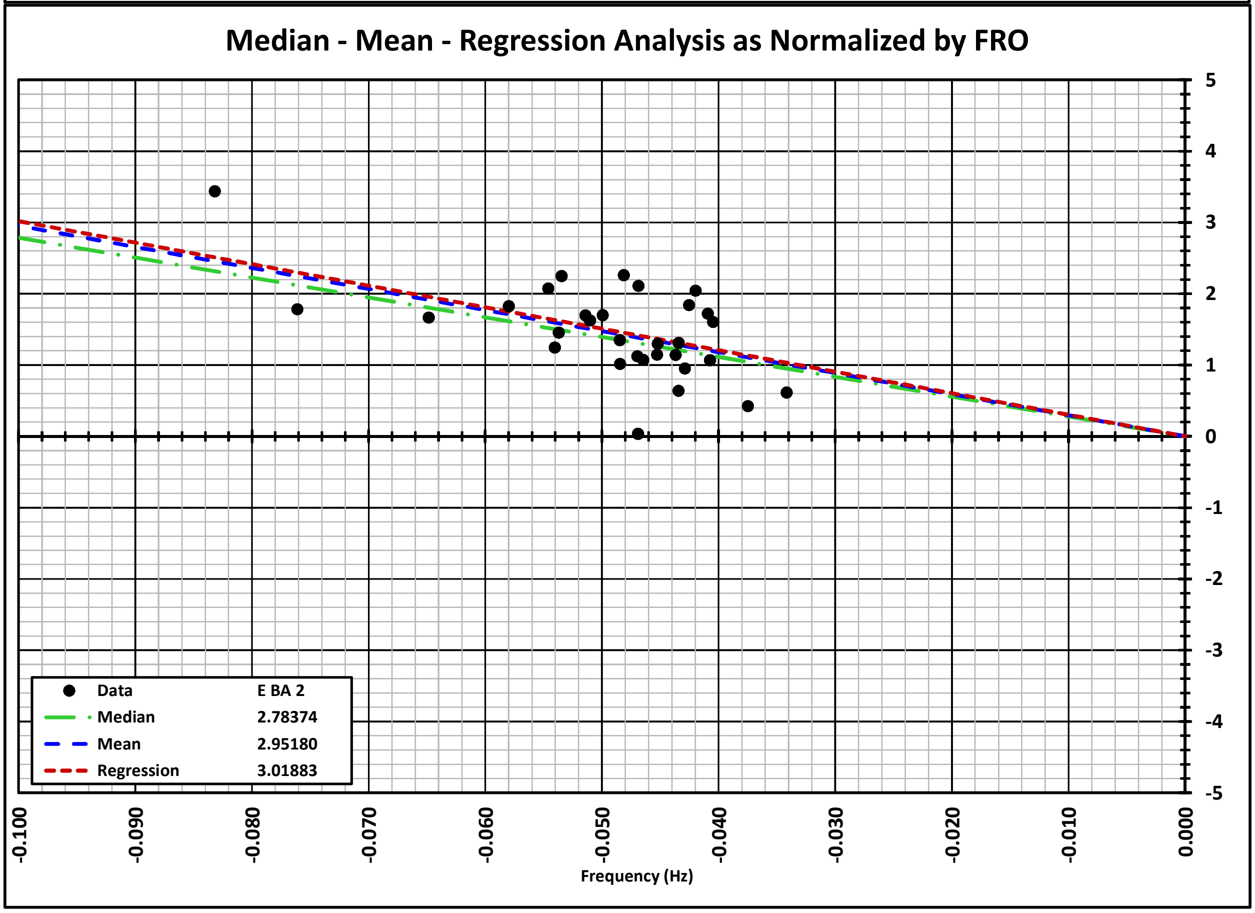

Frequency Response Events as Normalized by FRO
Eastern Interconnection - 2011
10.0
5.0
0.0
-5.0
-10.0
Frequency Response Normalized by FRO
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Balancing Authority


Frequency Response Events as Normalized by FRO
Western Interconnection - 2011
Frequency Response Normalized by FRO
25.0
20.0
15.0
10.0
5.0
0.0
-5.0
-10.0
-15.0
-20.0
-25.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Balancing Authority

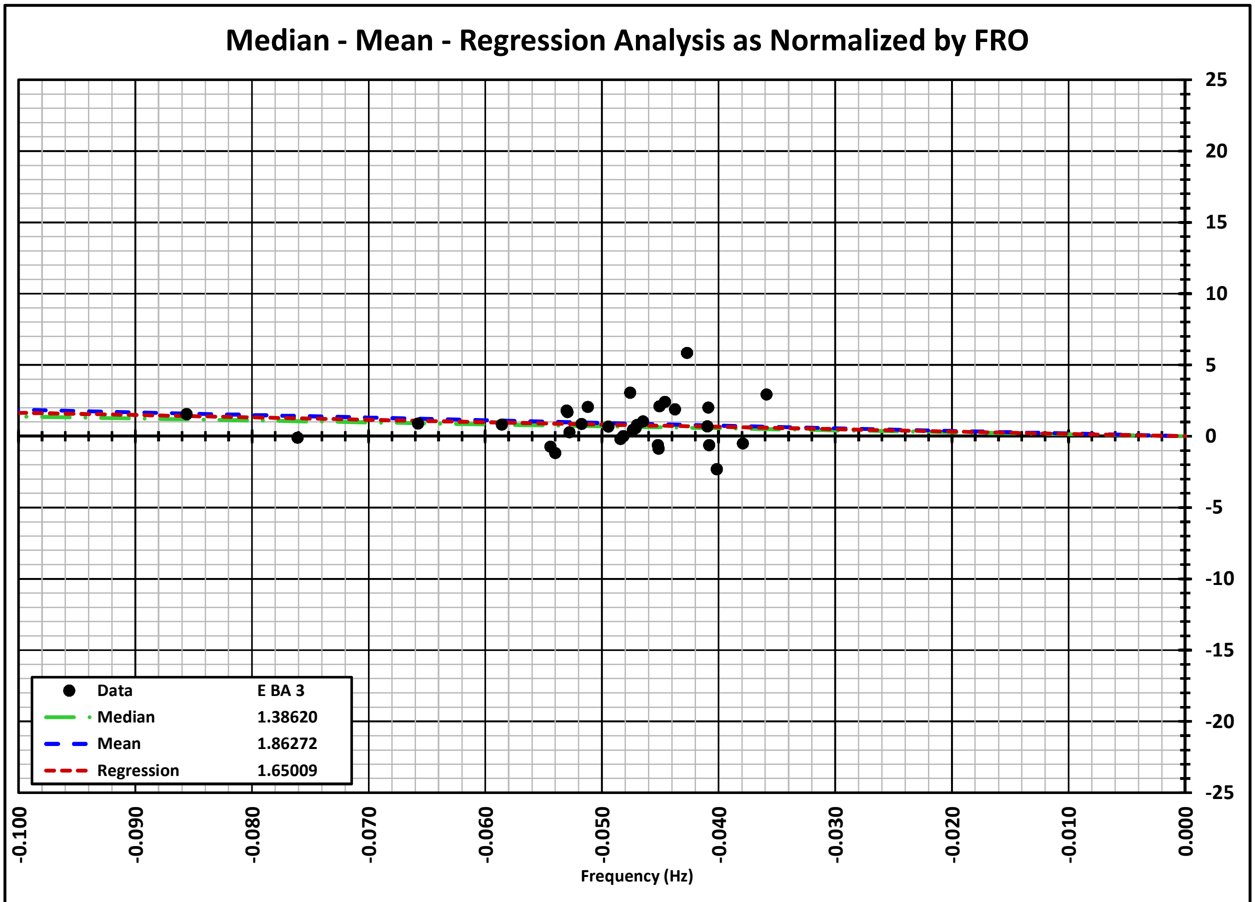

Frequency Response Events as Normalized by FRO
Western Interconnection - 2011
Frequency Response Normalized by FRO
10.0
5.0
0.0
-5.0
-10.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Balancing Authority

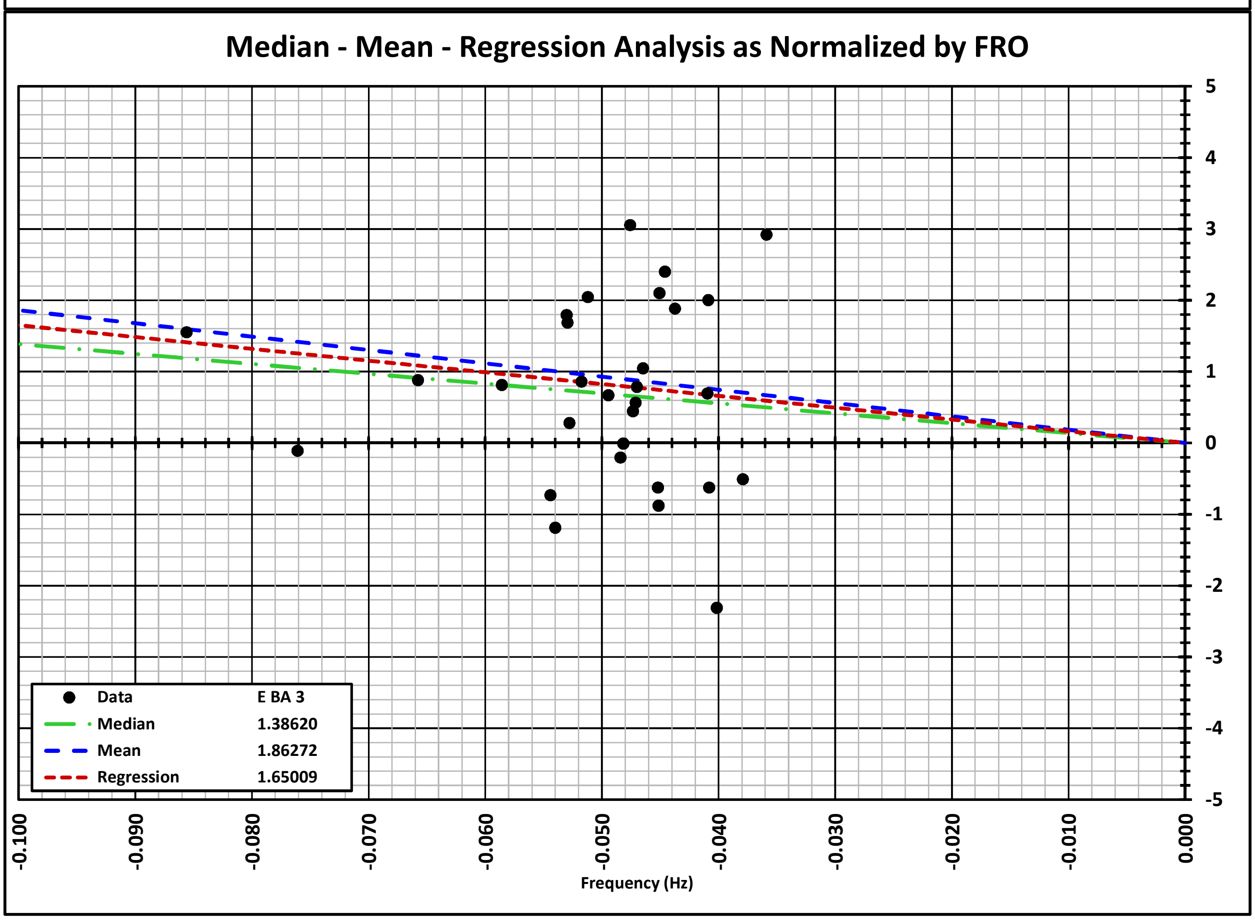


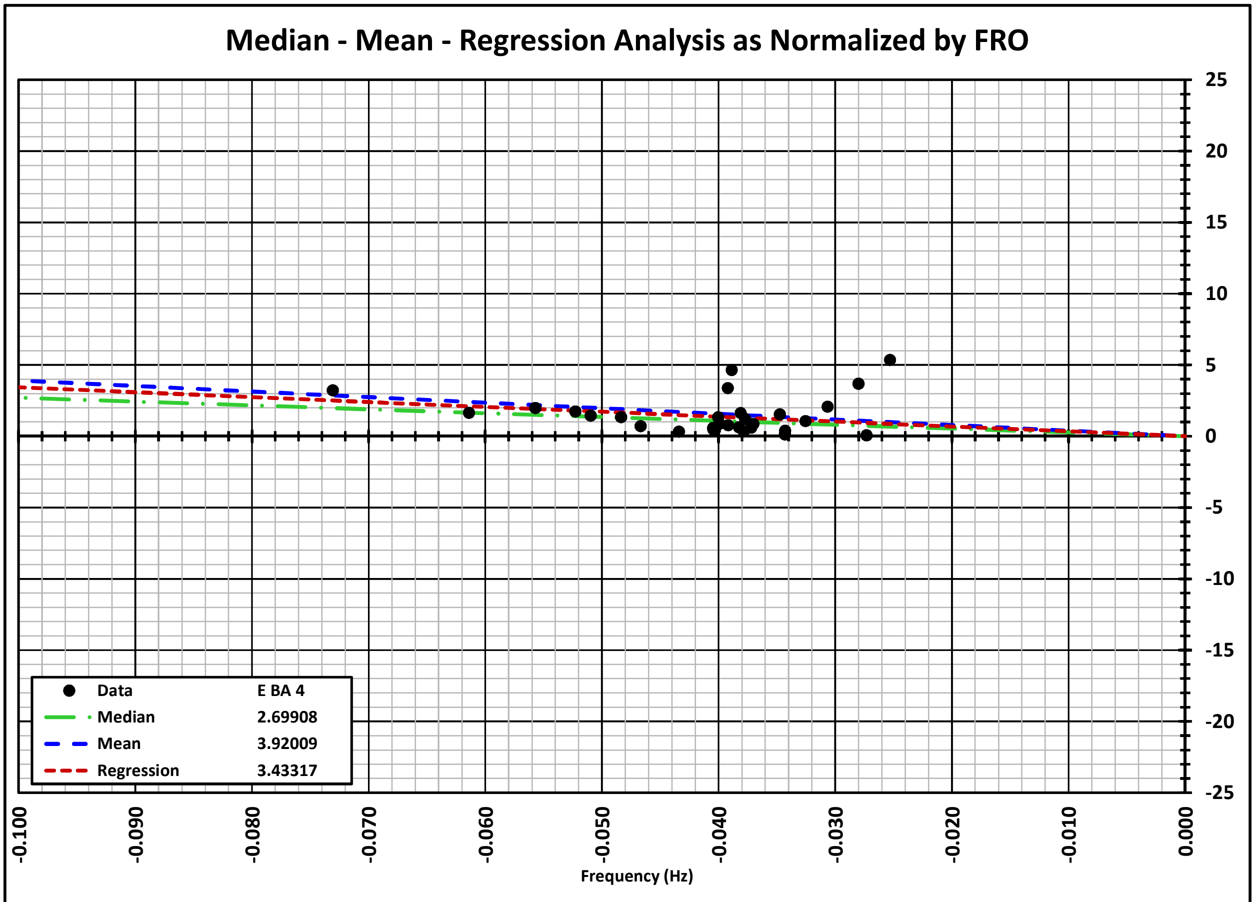
Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 1
Median 2.93266
Mean 2.44868
Regression 2.57361
Frequency (Hz)

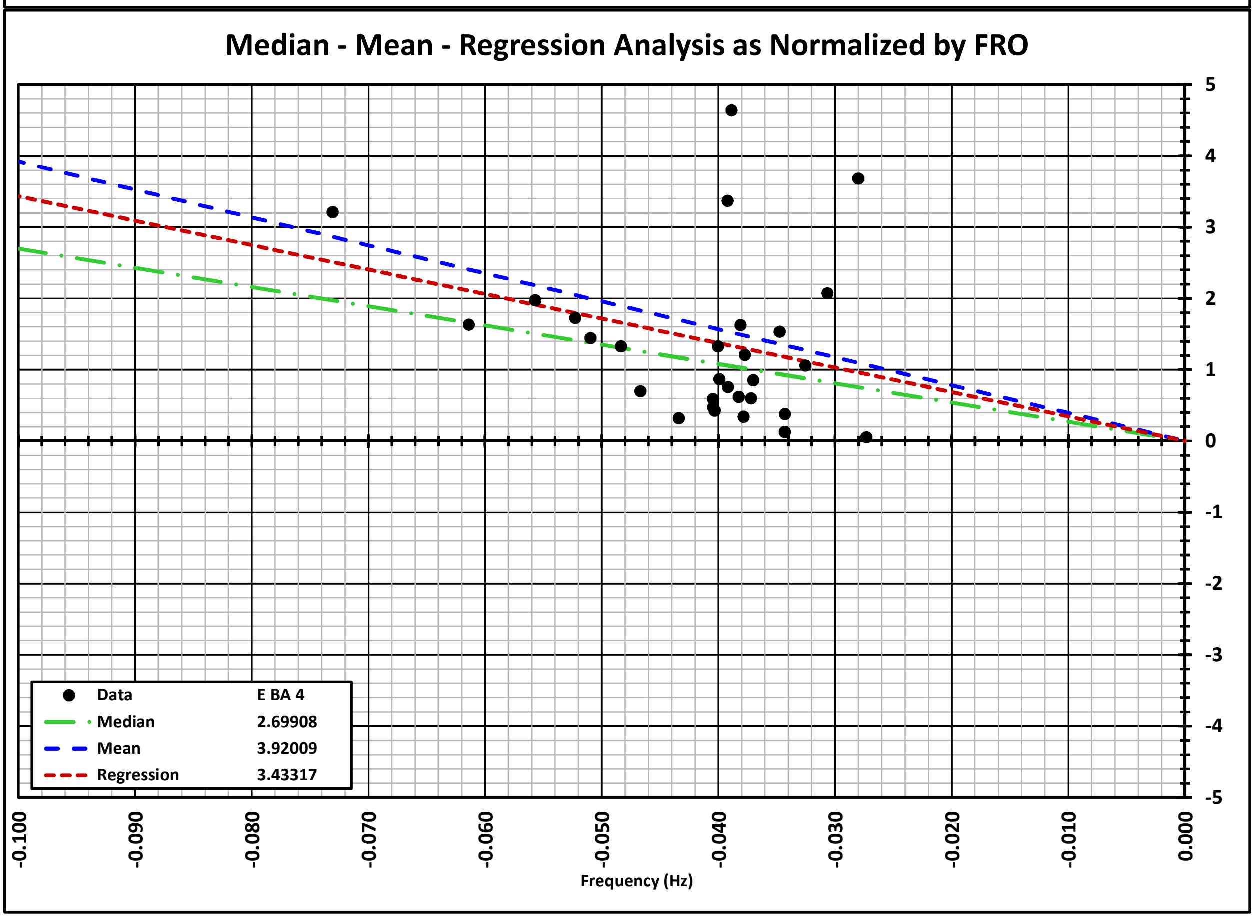


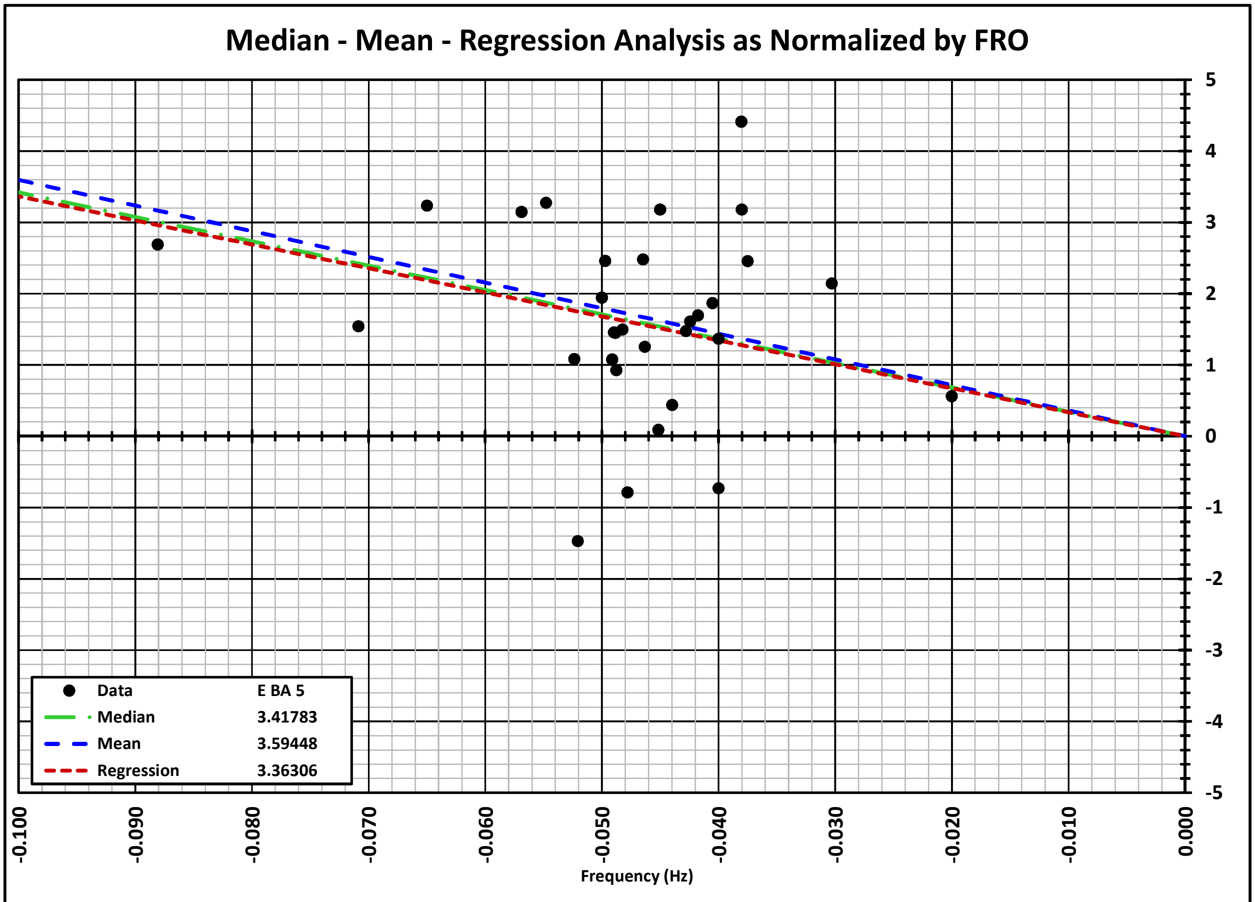
Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 2
Median 2.78374
Mean 2.95180
Regression 3.01883
Frequency (Hz)

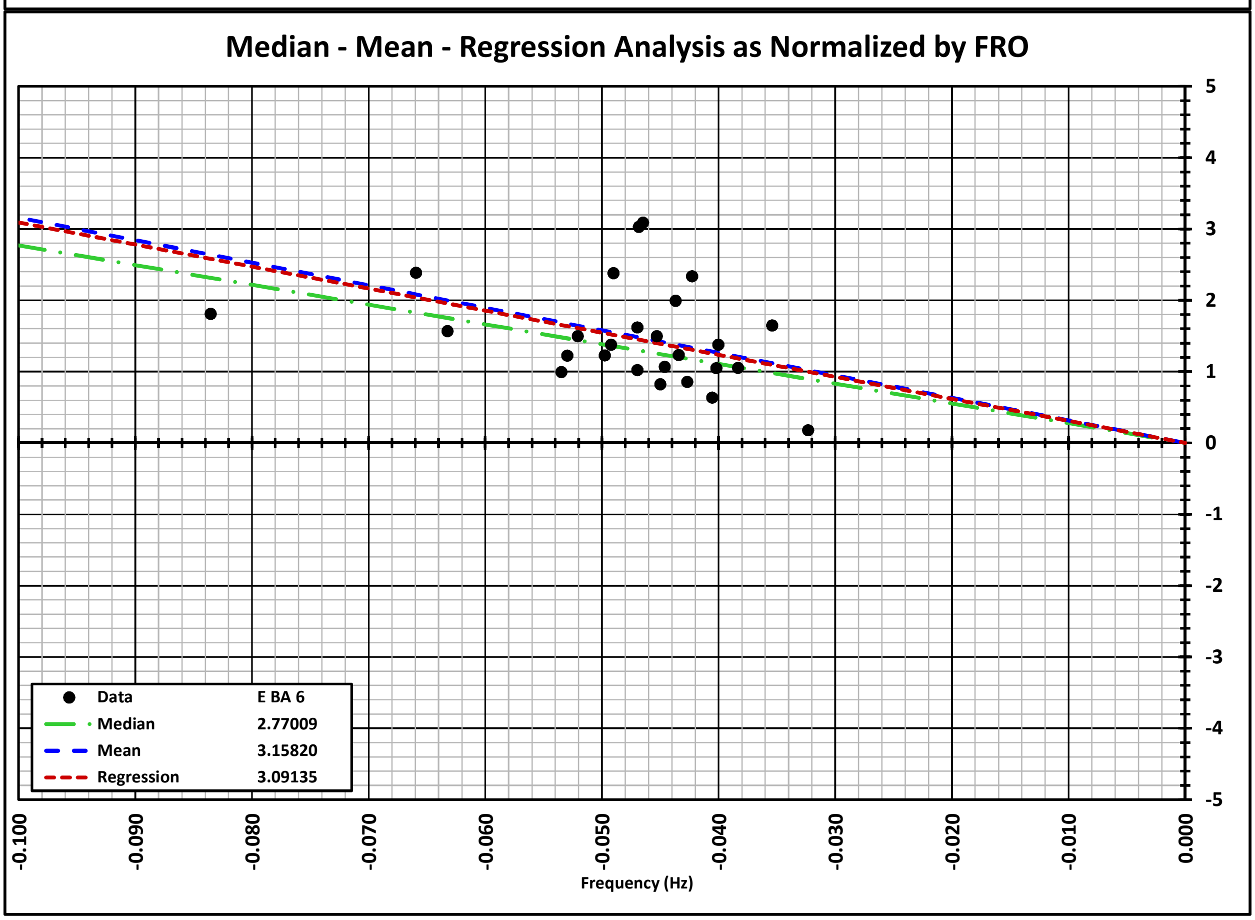

Median - Mean - Regression Analysis as Normalized by FRO
25
20
15
10
5
0
-5
-10
-15
-20
-25
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)
Data E BA 3
Median 1.38620
Mean 1.86272
Regression 1.65009

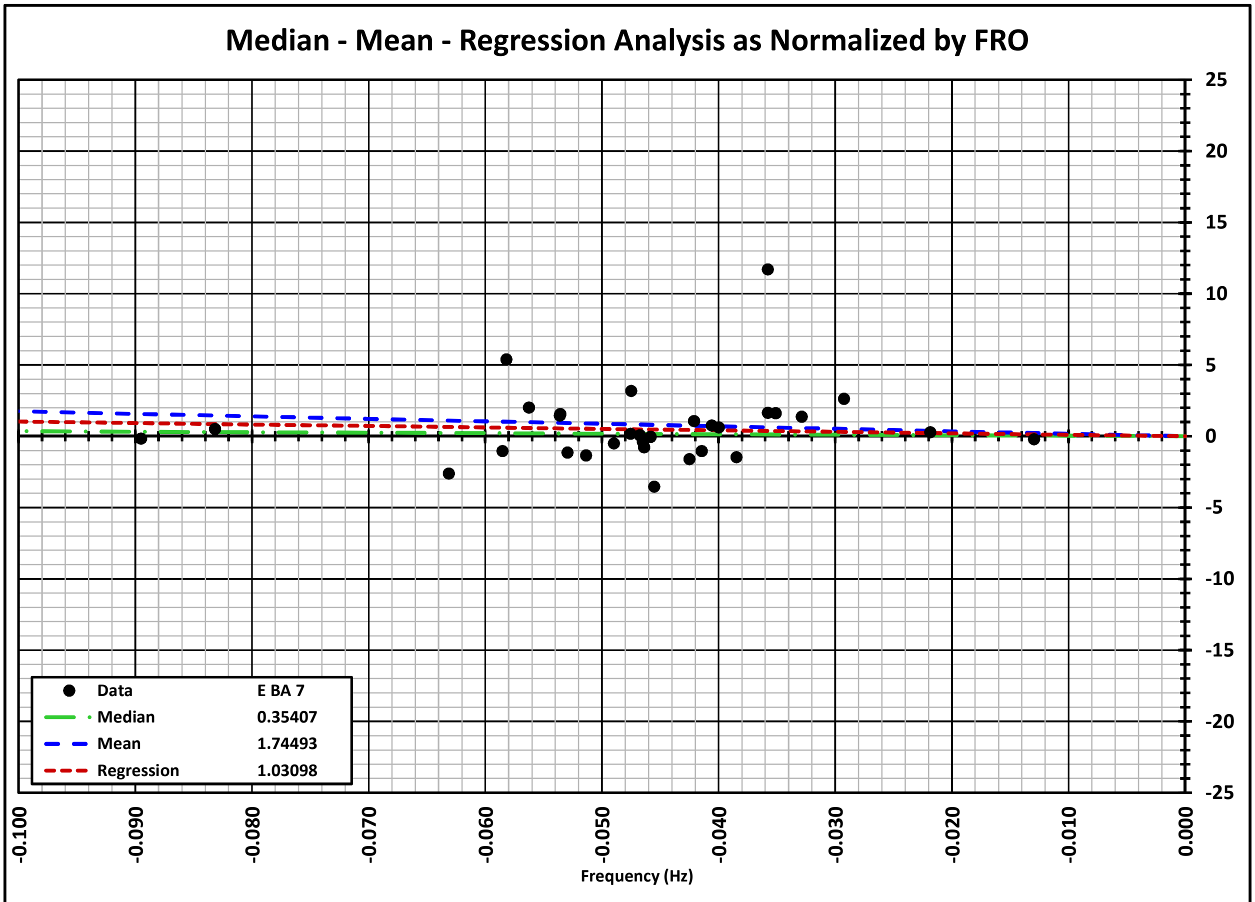

Median - Mean - Regression Analysis as Normalized by FRO
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)
Data E BA 3
Median 1.38620
Mean 1.86272
Regression 1.65009

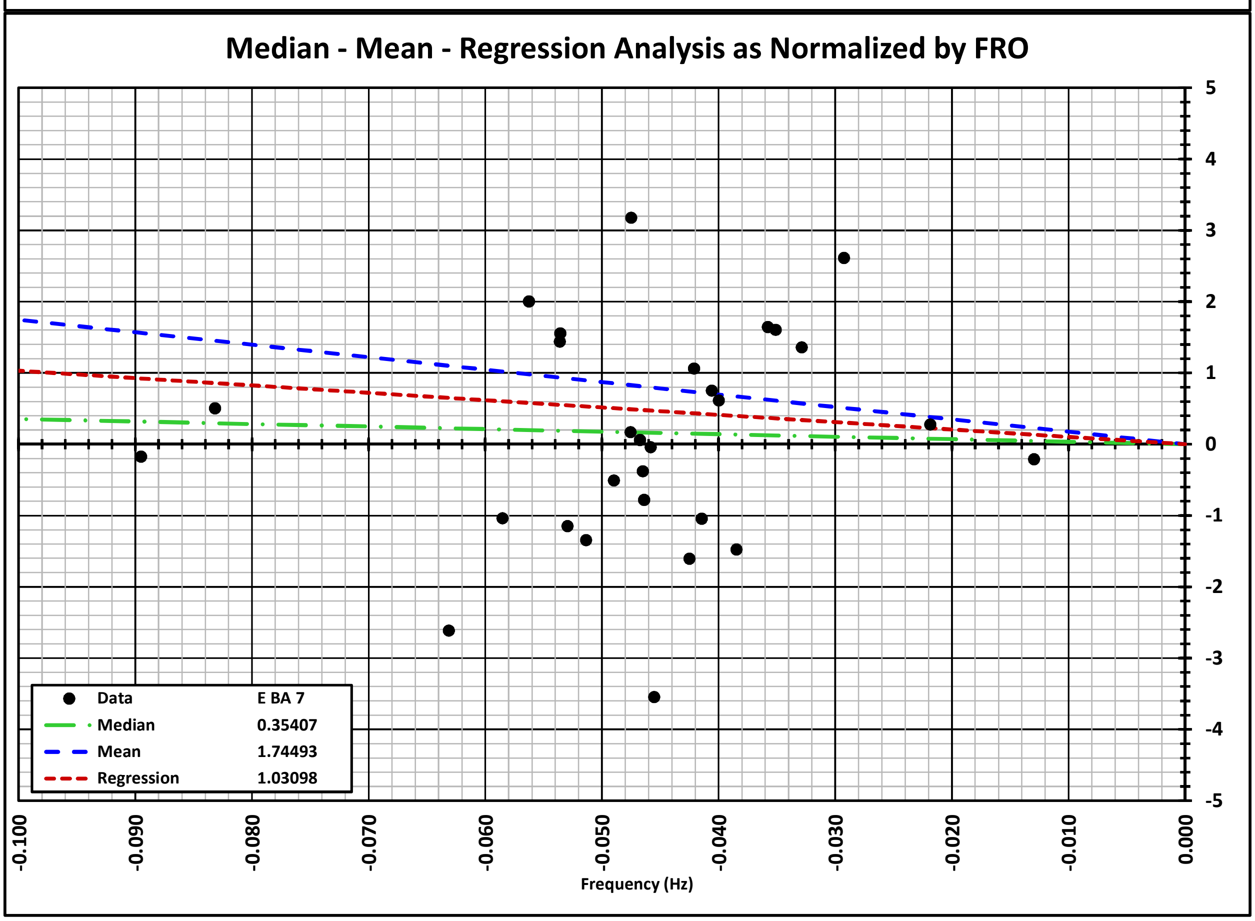

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 4
Median 2.69908
Mean 3.92009
Regression 3.43317
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
25
20
15
10
5
0
-5
-10
-15
-20
-25

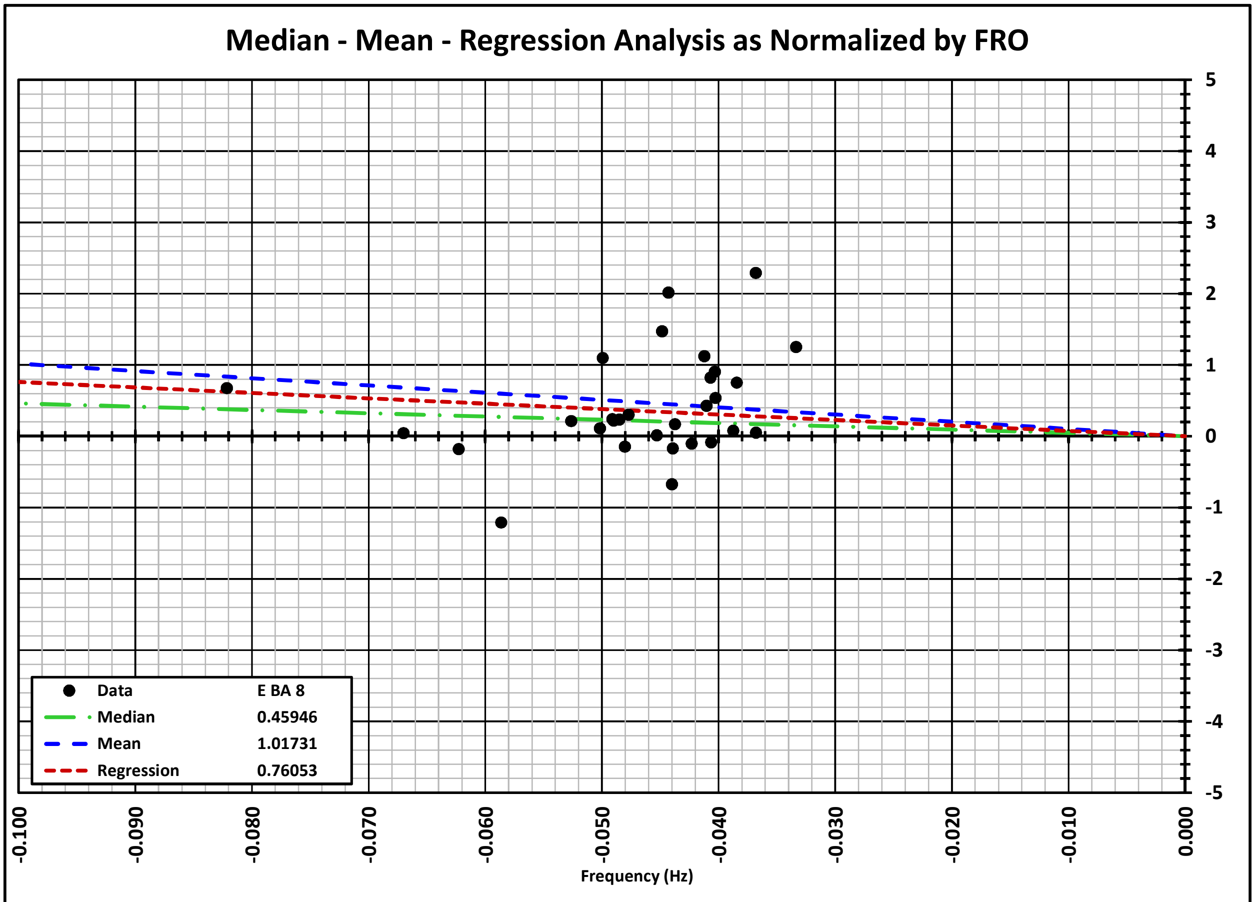

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 4
Median 2.69908
Mean 3.92009
Regression 3.43317
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
5
4
3
2
1
0
-1
-2
-3
-4
-5

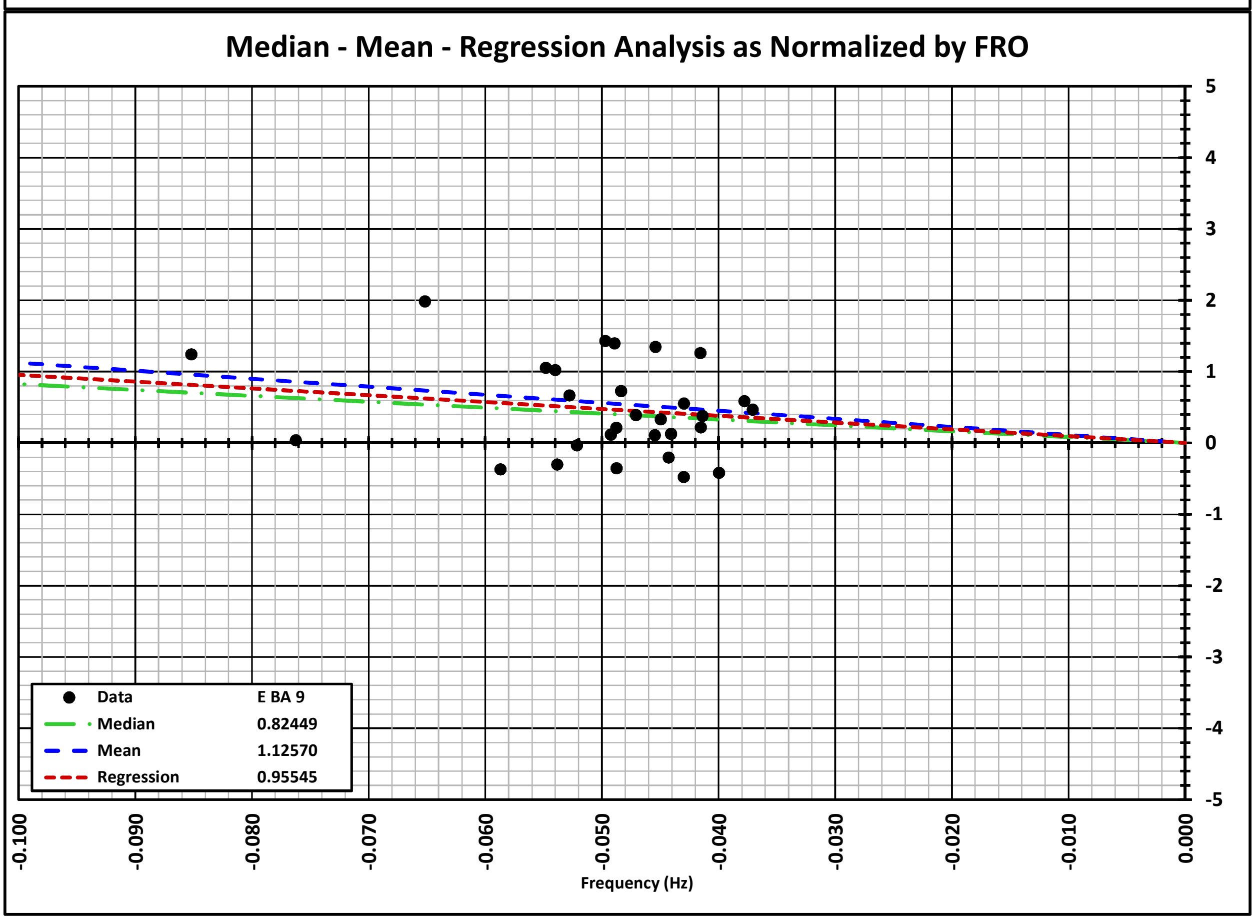

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 5
Median 3.41783
Mean 3.59448
Regression 3.36306
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
5
4
3
2
1
0
-1
-2
-3
-4
-5

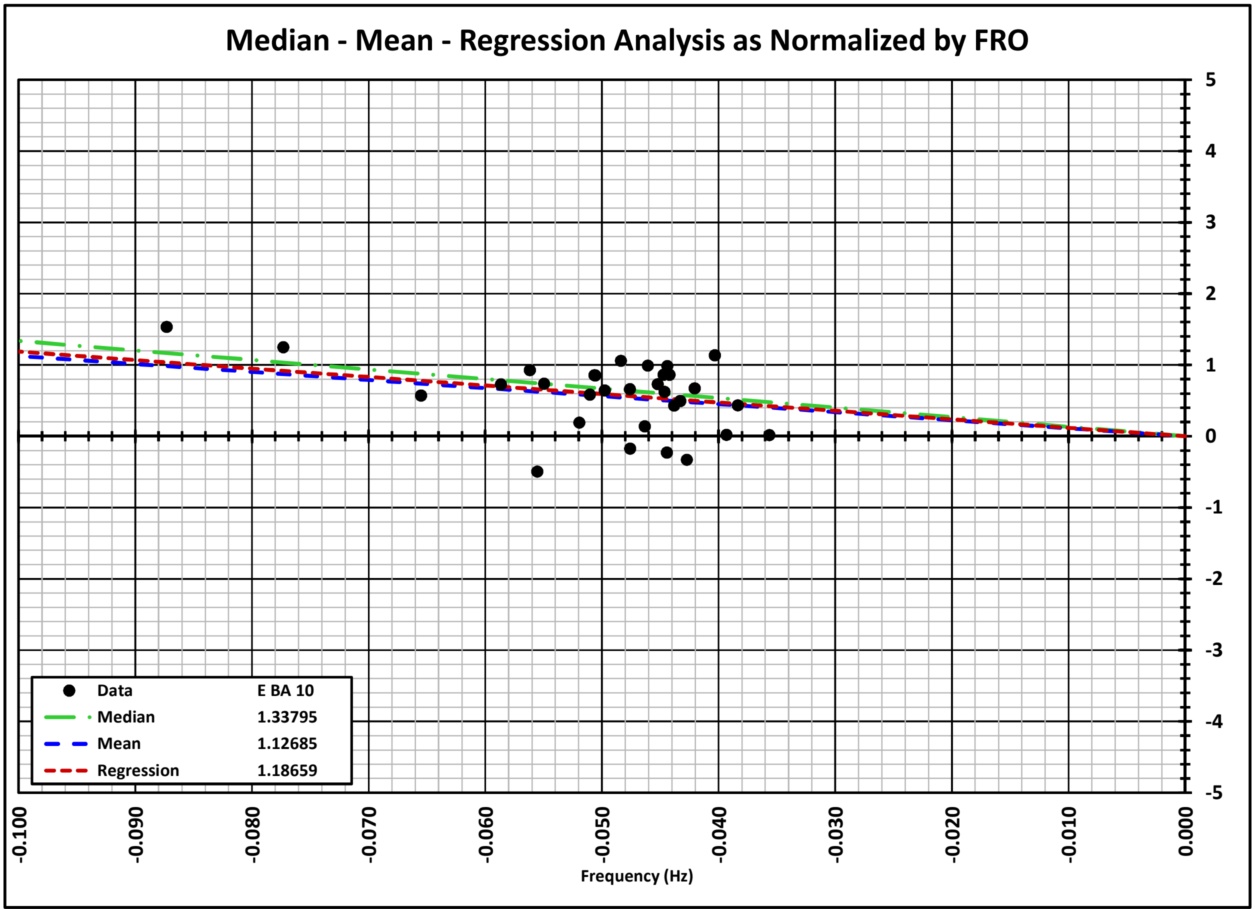

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 6
Median 2.77009
Mean 3.15820
Regression 3.09135
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
5
4
3
2
1
0
-1
-2
-3
-4
-5

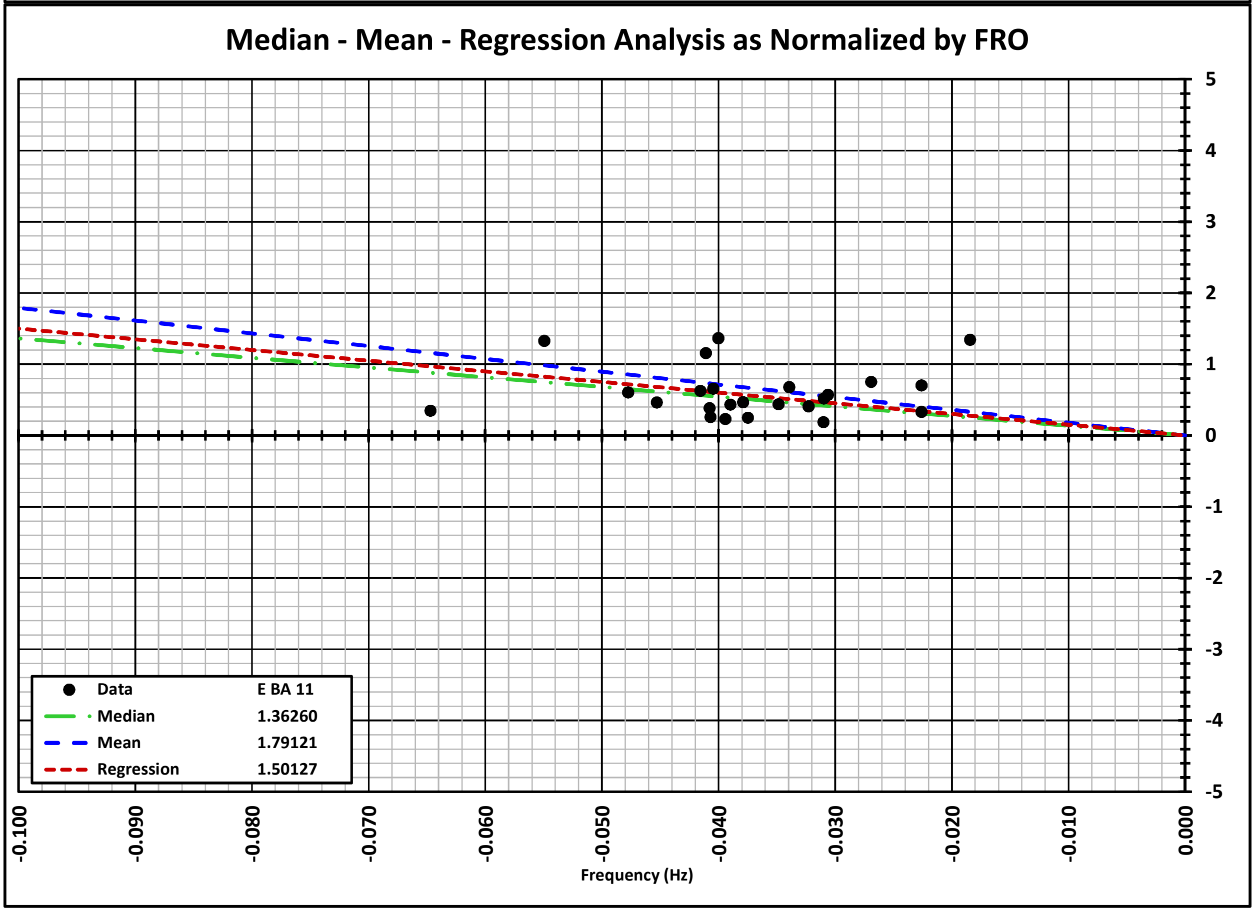

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 7
Median 0.35407
Mean 1.74493
Regression 1.03098
Frequency (Hz)

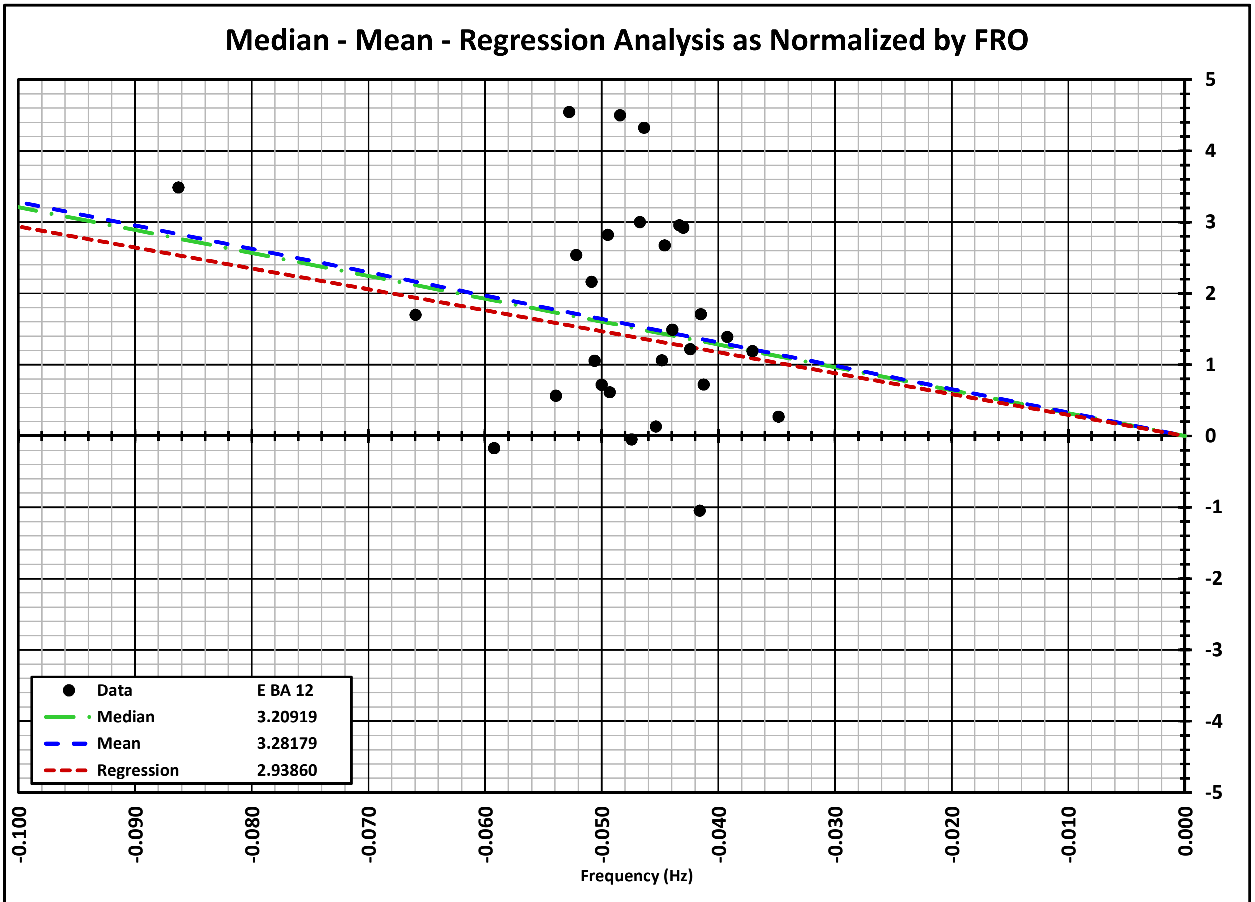

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 7
Median 0.35407
Mean 1.74493
Regression 1.03098
Frequency (Hz)

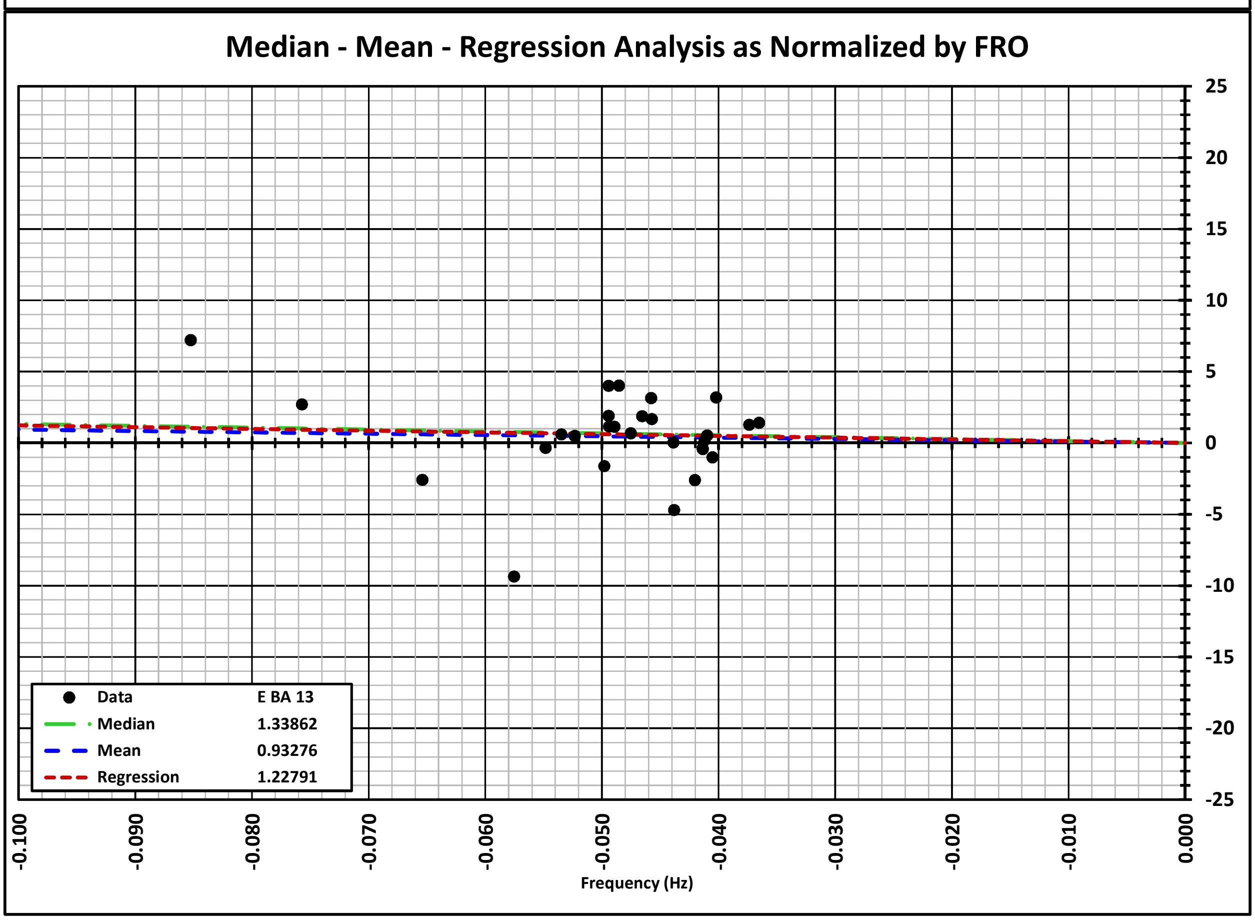

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 8
Median 0.45946
Mean 1.01731
Regression 0.76053
Frequency (Hz)
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000

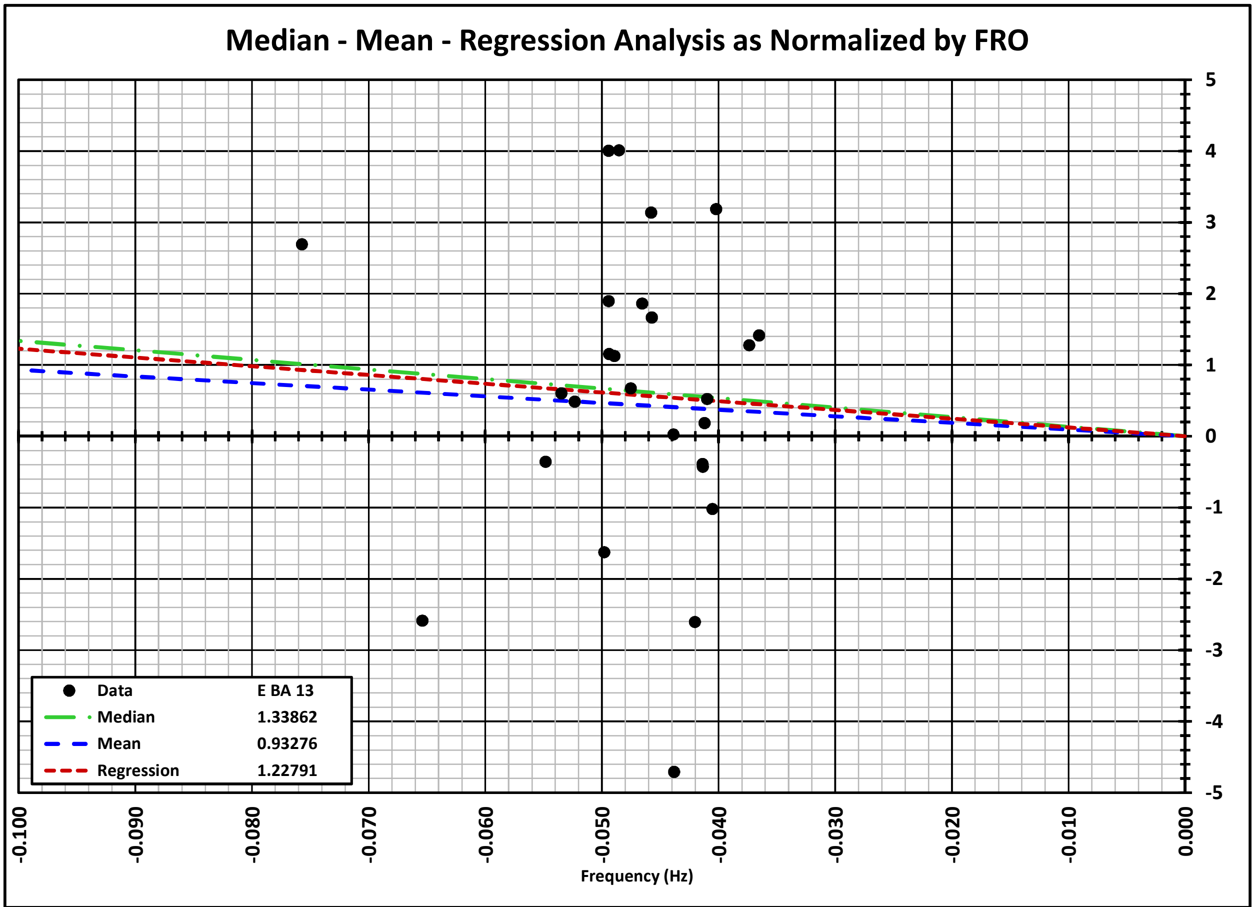

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 9
Median 0.82449
Mean 1.12570
Regression 0.95545
Frequency (Hz)
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000

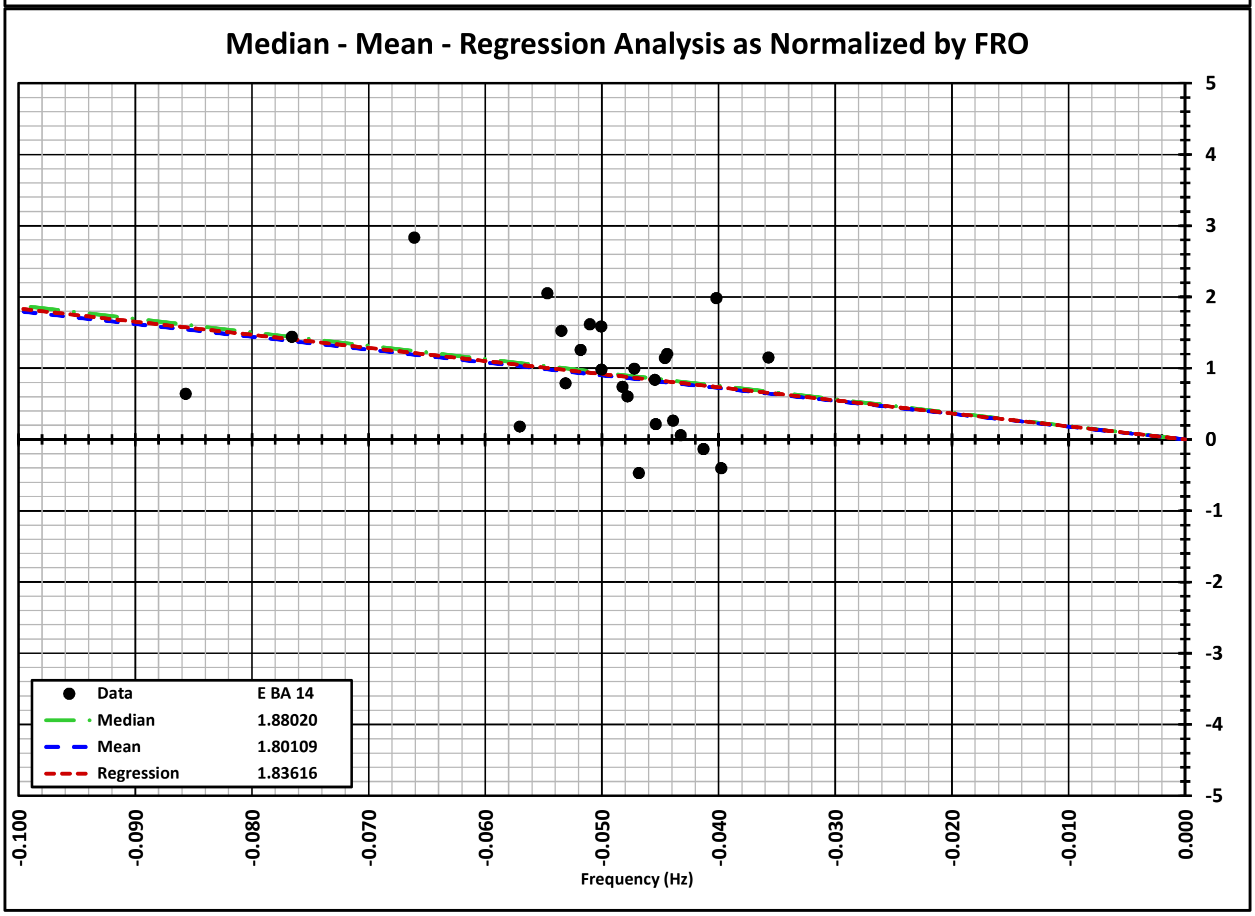

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 10
Median 1.33795
Mean 1.12685
Regression 1.18659
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

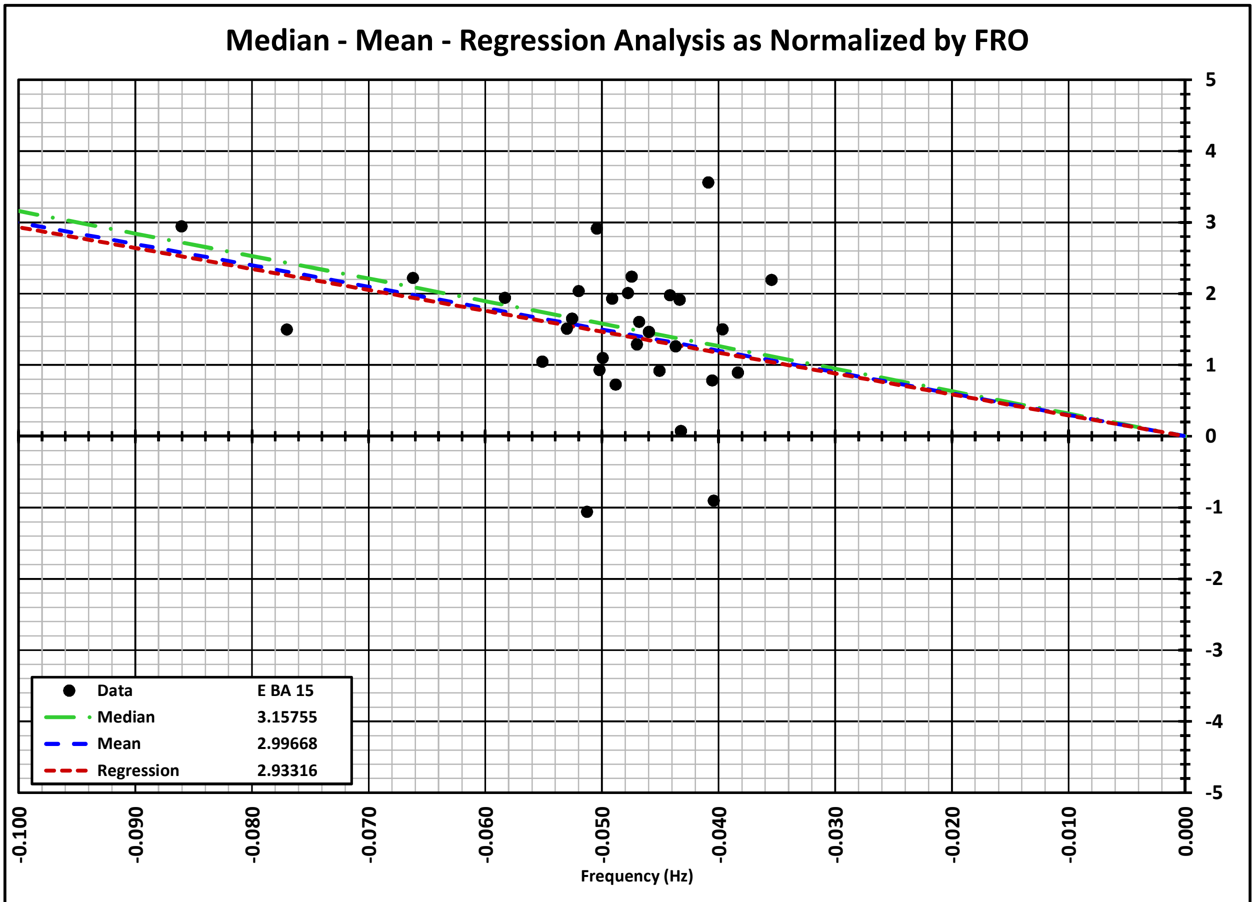

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 11
Median 1.36260
Mean 1.79121
Regression 1.50127
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)


Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 12
Median 3.20919
Mean 3.28179
Regression 2.93860
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

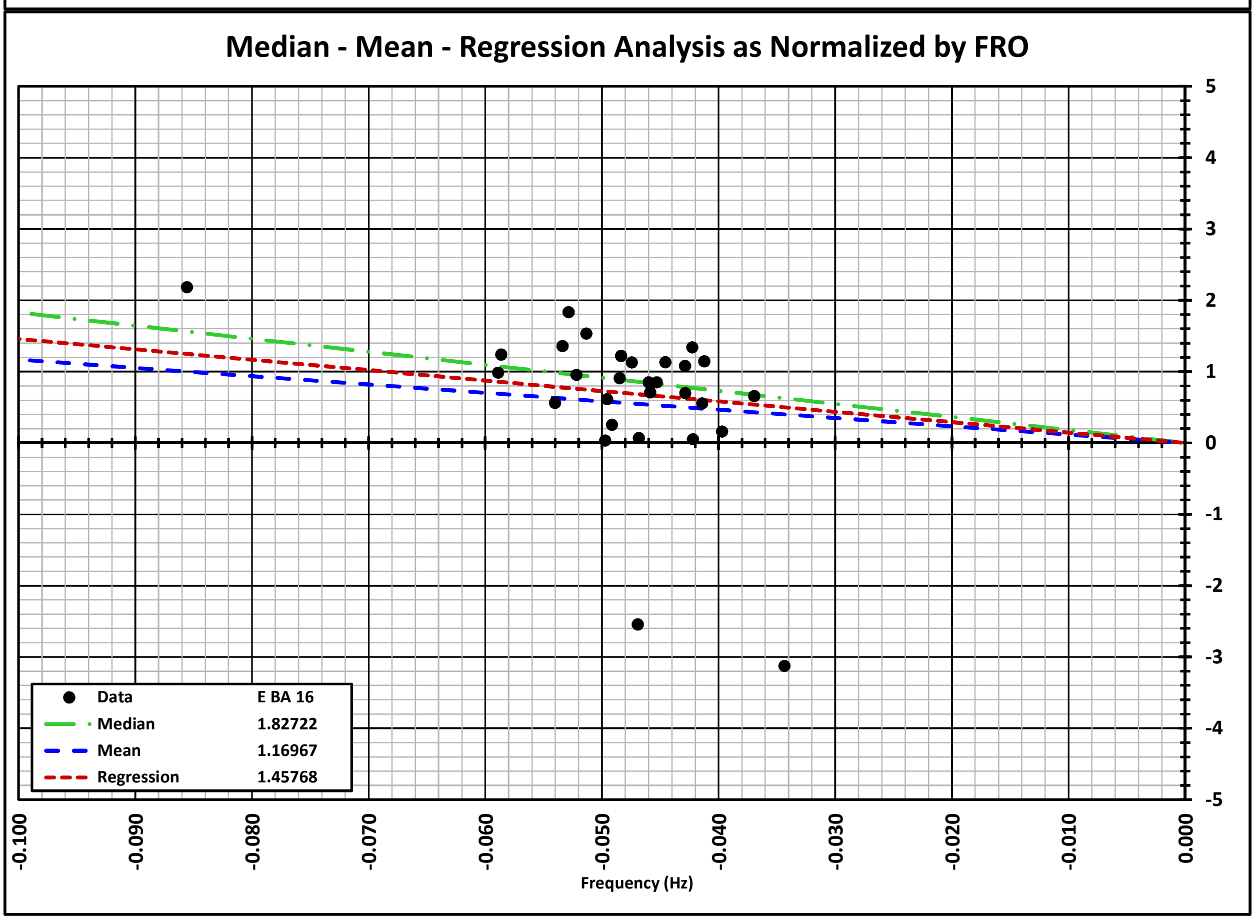

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 13
Median 1.33862
Mean 0.93276
Regression 1.22791
25
20
15
10
5
0
-5
-10
-15
-20
-25
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)


Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 13
Median 1.33862
Mean 0.93276
Regression 1.22791
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

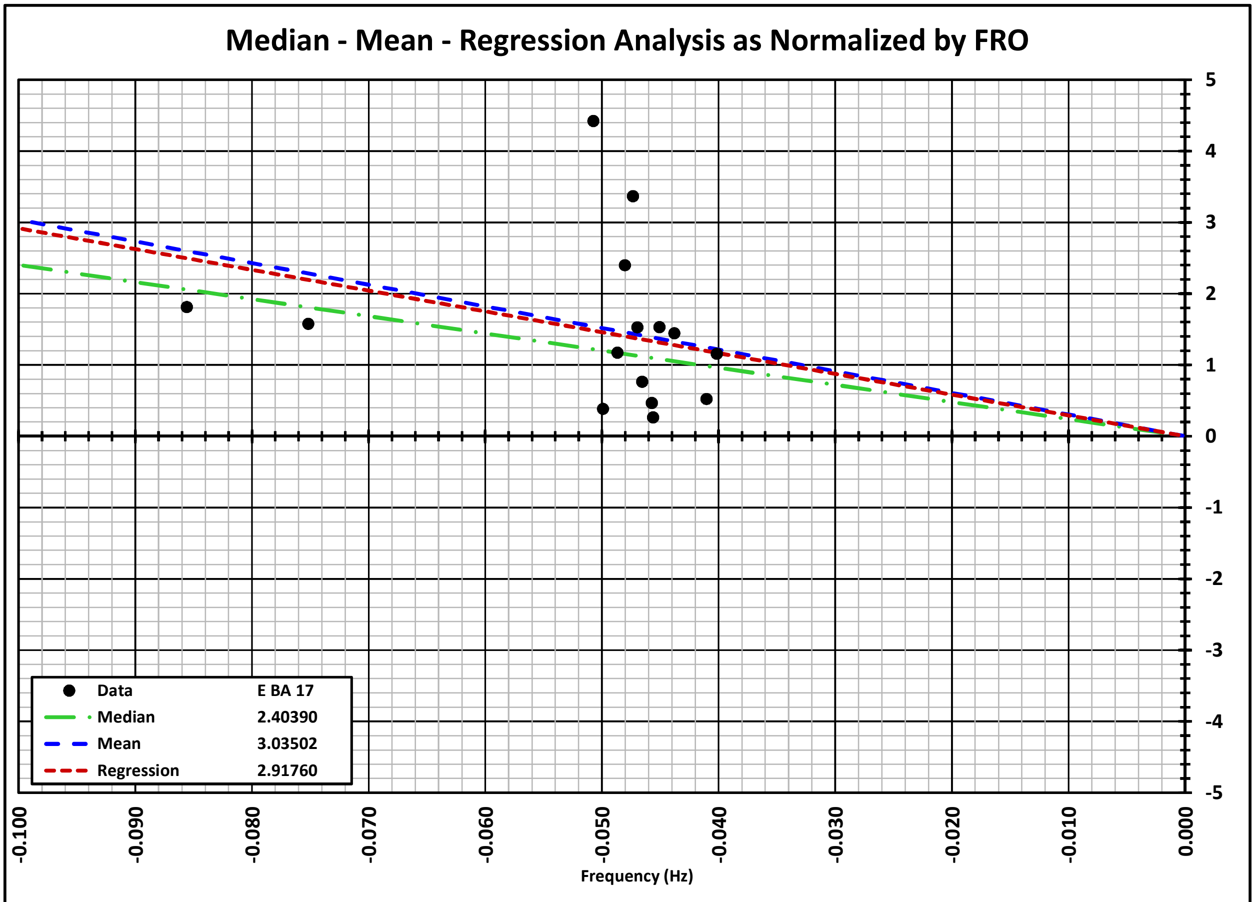

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 14
Median 1.88020
Mean 1.80109
Regression 1.83616
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

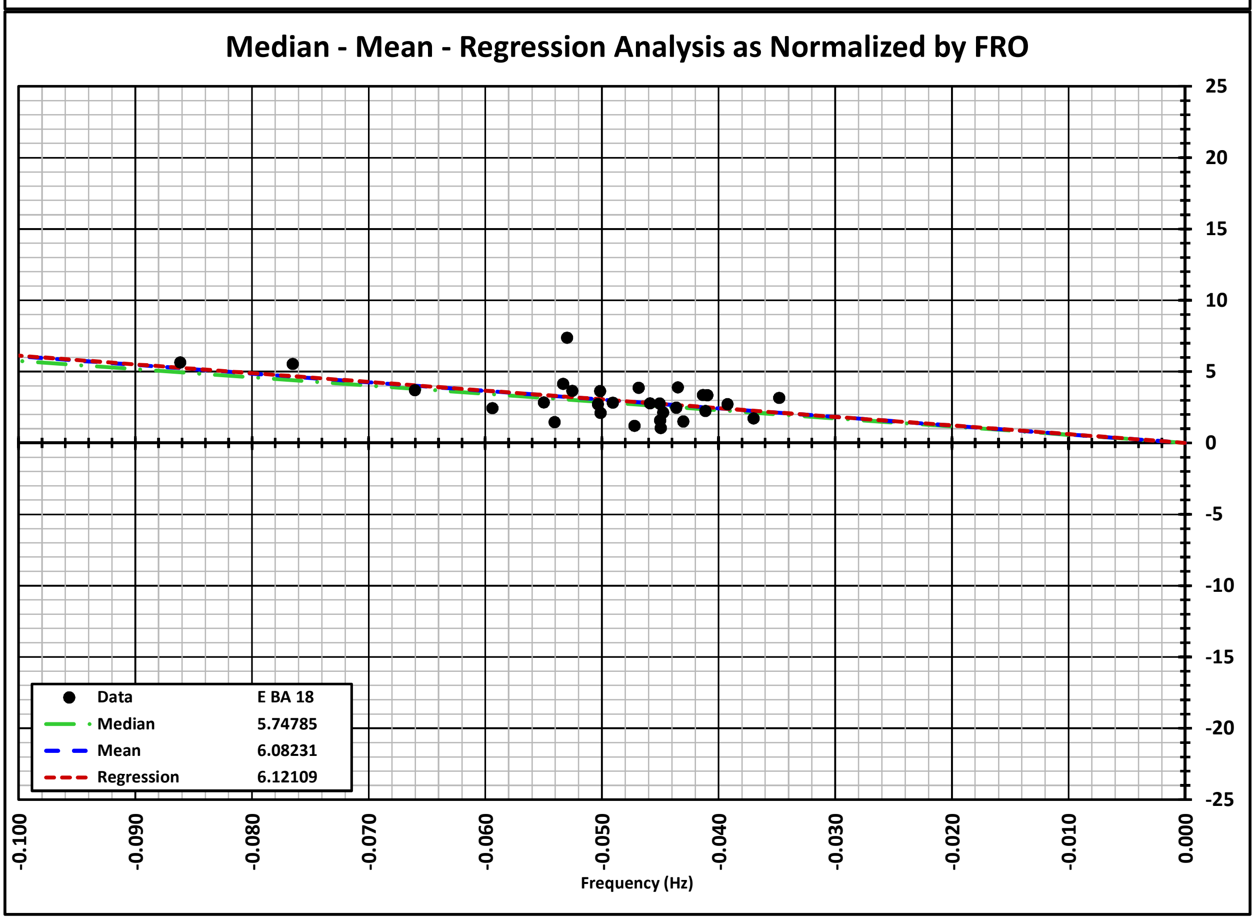


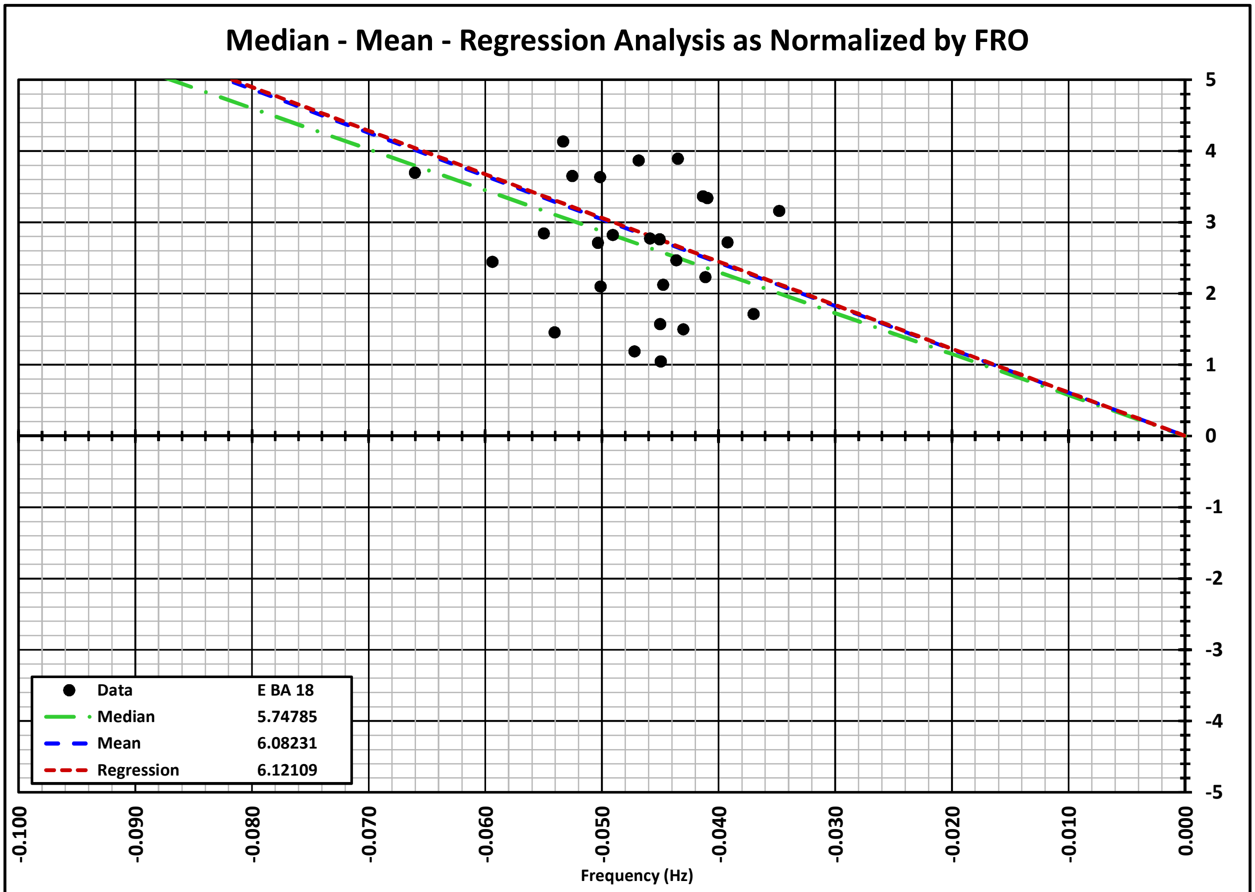
Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 15
Median 3.15755
Mean 2.99668
Regression 2.93316
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

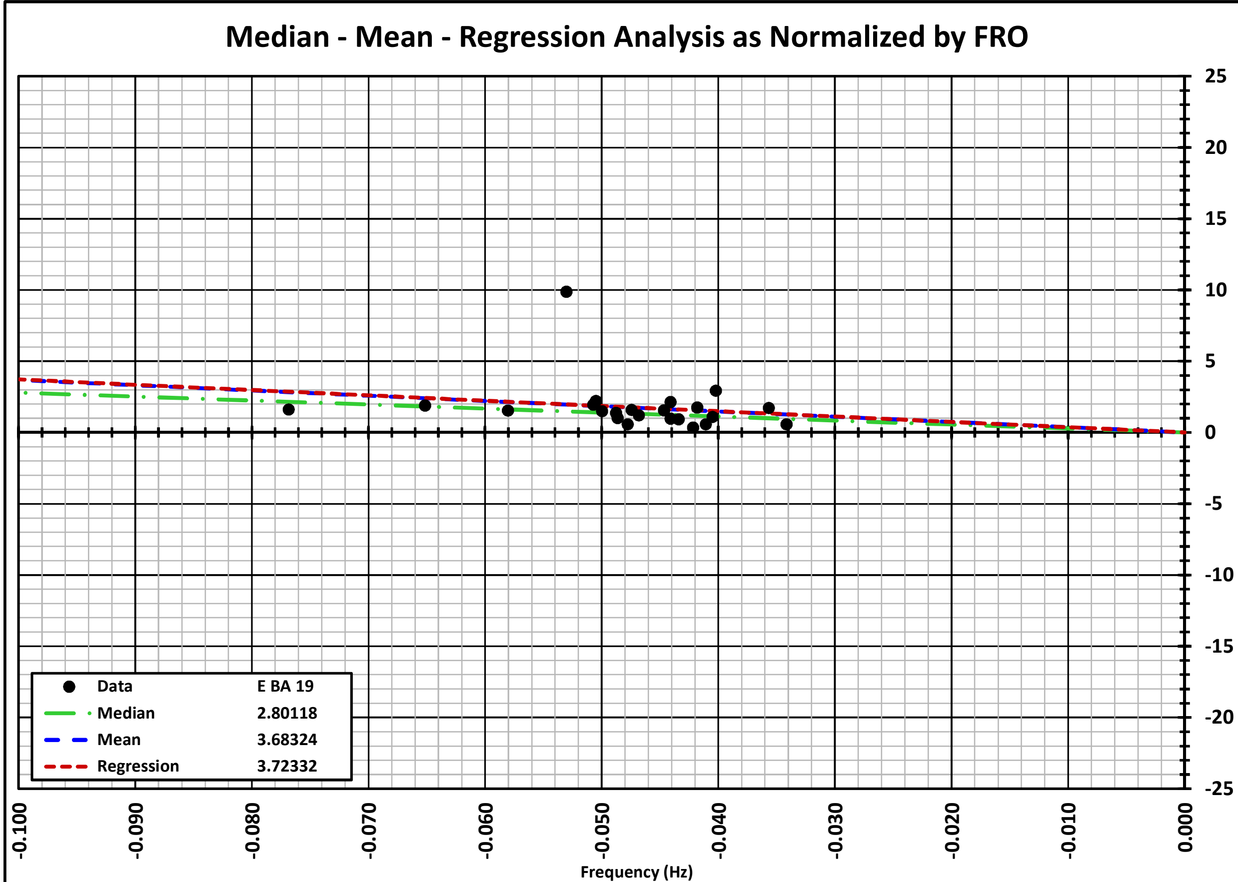


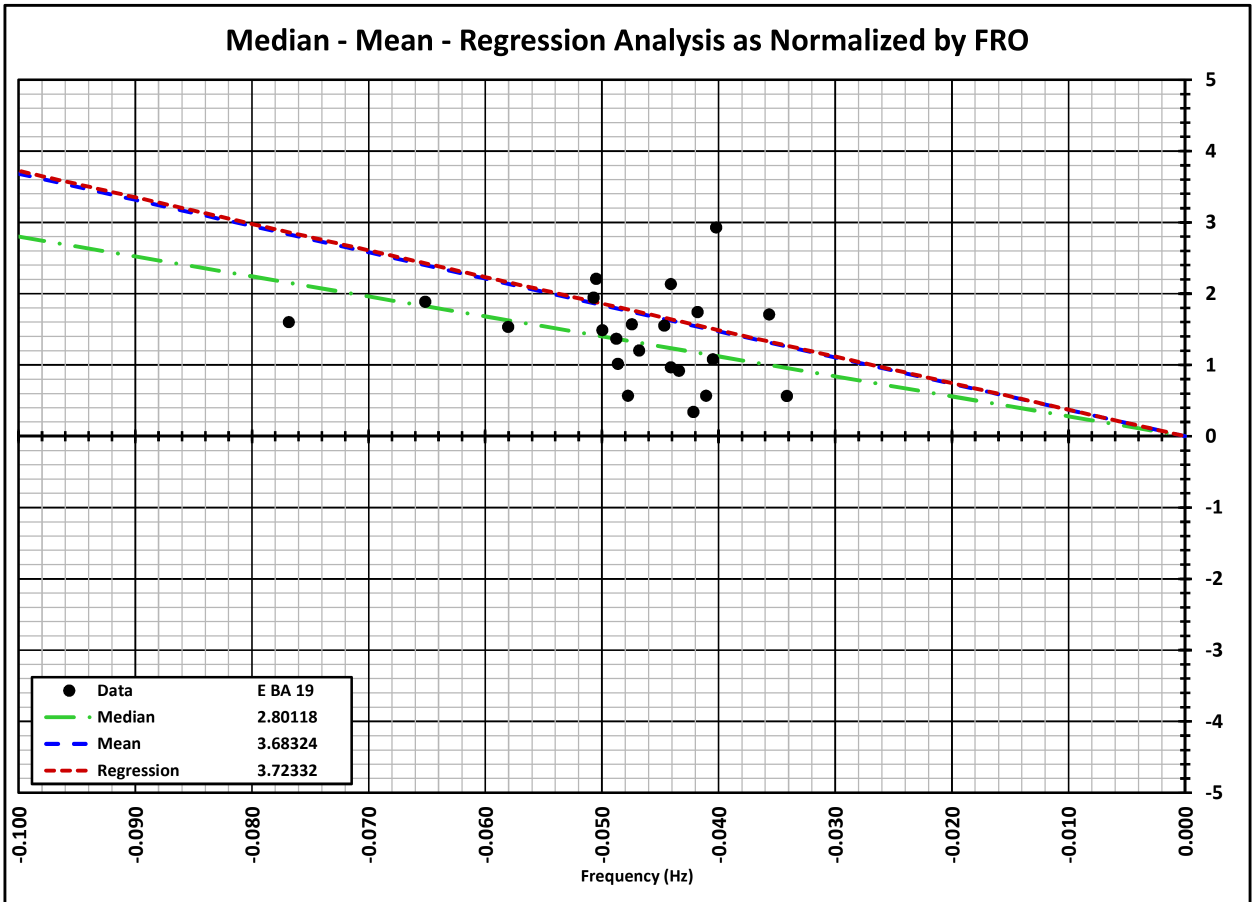
Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 16
Median 1.82722
Mean 1.16967
Regression 1.45768
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

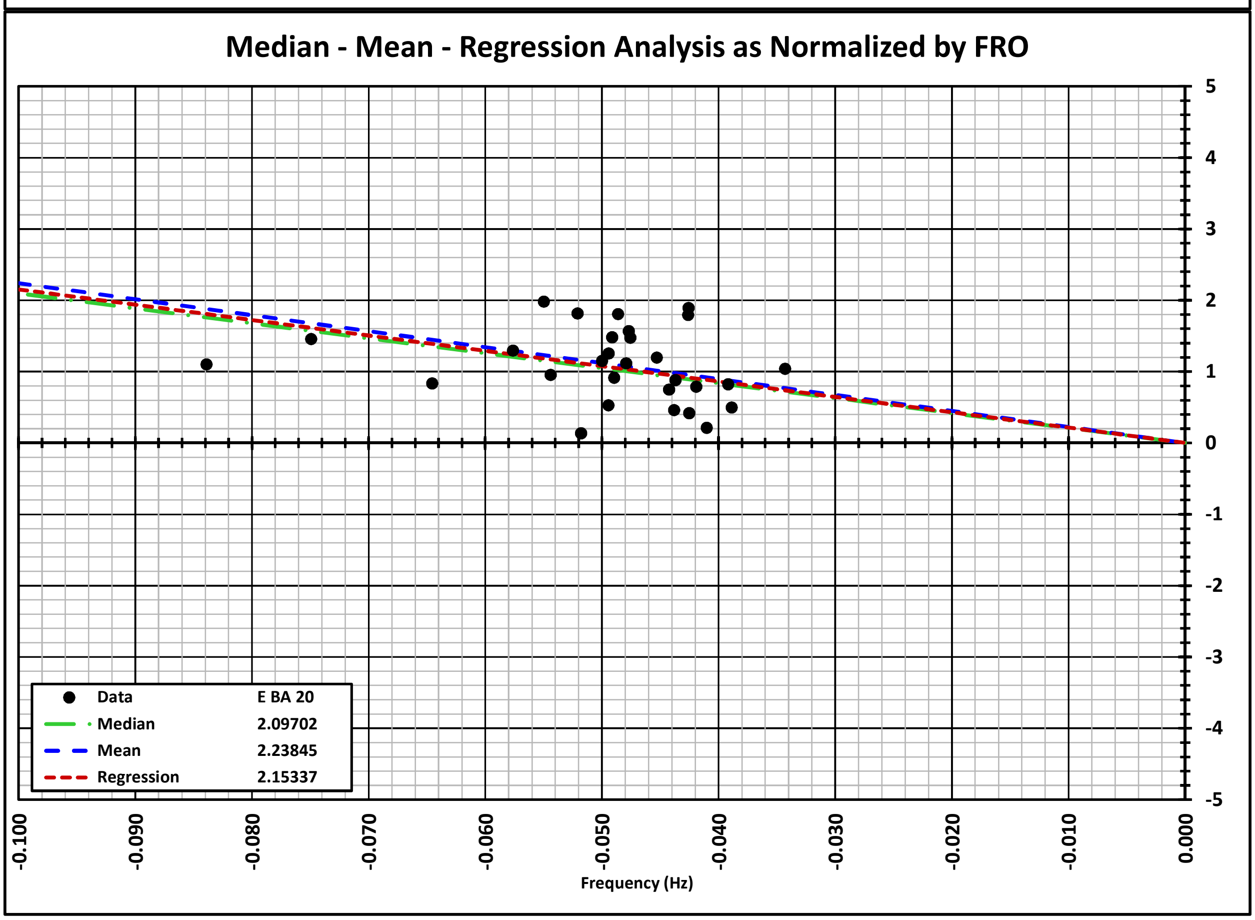

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 17
Median 2.40390
Mean 3.03502
Regression 2.91760
Frequency (Hz)
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000

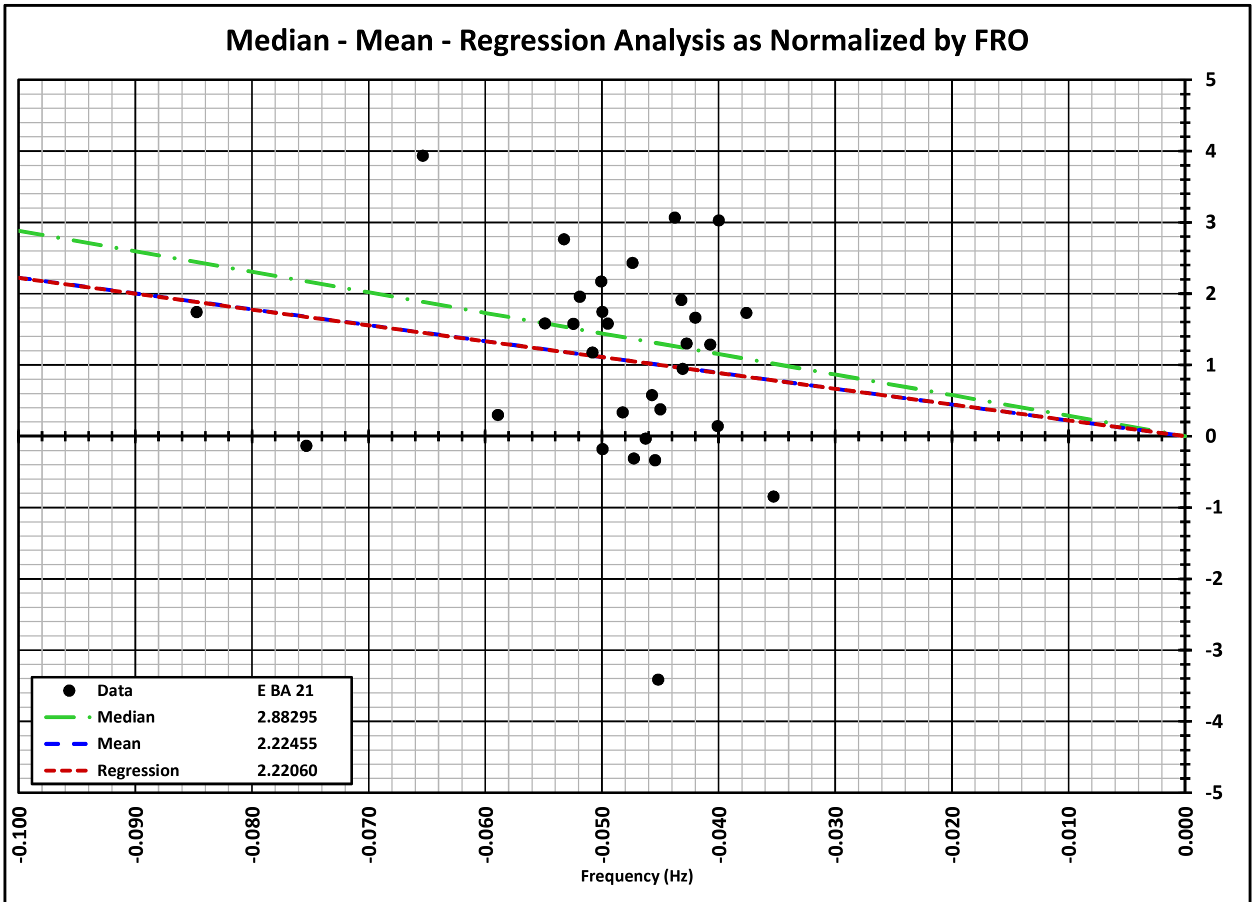

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 18
Median 5.74785
Mean 6.08231
Regression 6.12109
Frequency (Hz)
25
20
15
10
5
0
-5
-10
-15
-20
-25
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000


Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 18
Median 5.74785
Mean 6.08231
Regression 6.12109
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
5
4
3
2
1
0
-1
-2
-3
-4
-5

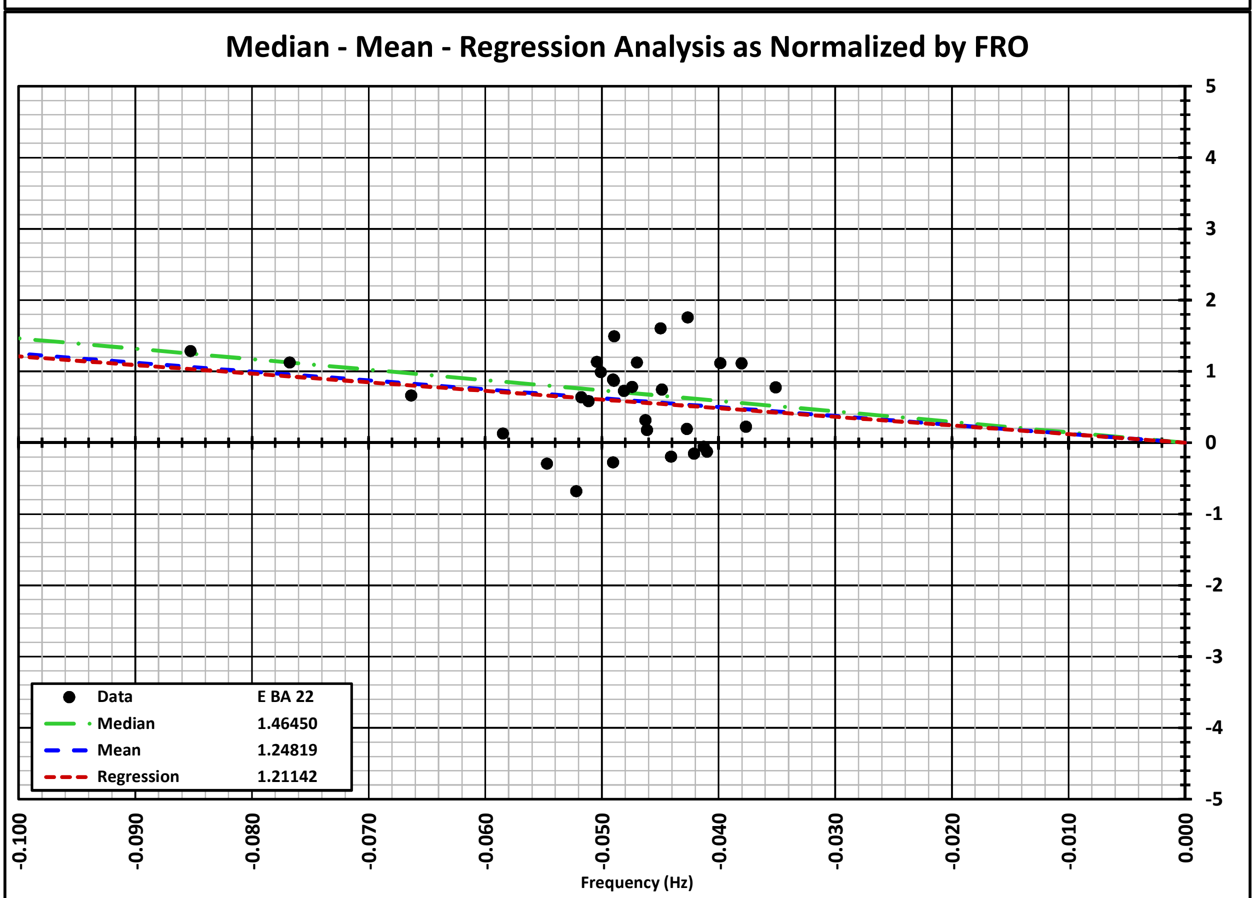

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 19
Median 2.80118
Mean 3.68324
Regression 3.72332
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
25
20
15
10
5
0
-5
-10
-15
-20
-25


Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 19
Median 2.80118
Mean 3.68324
Regression 3.72332
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

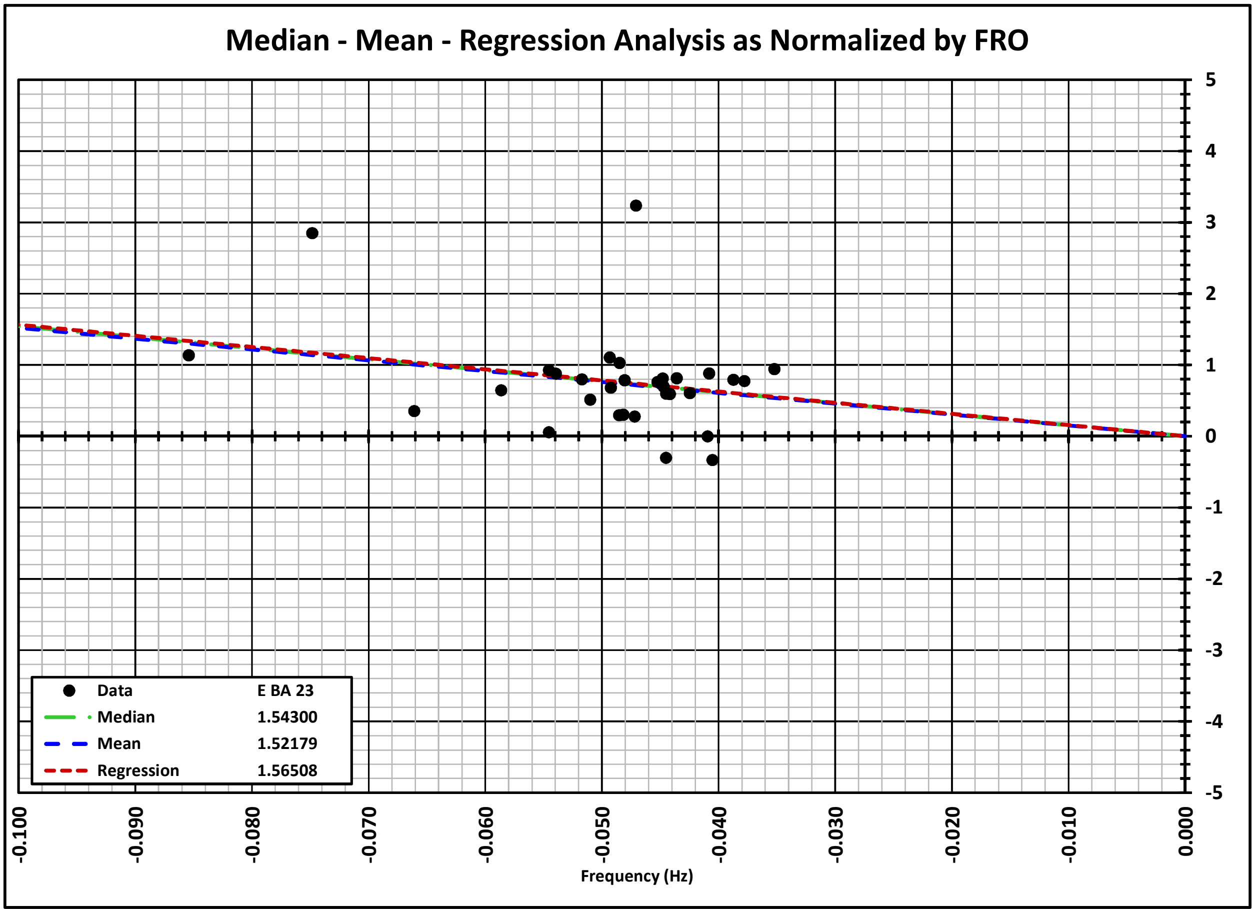

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 20
Median 2.09702
Mean 2.23845
Regression 2.15337
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

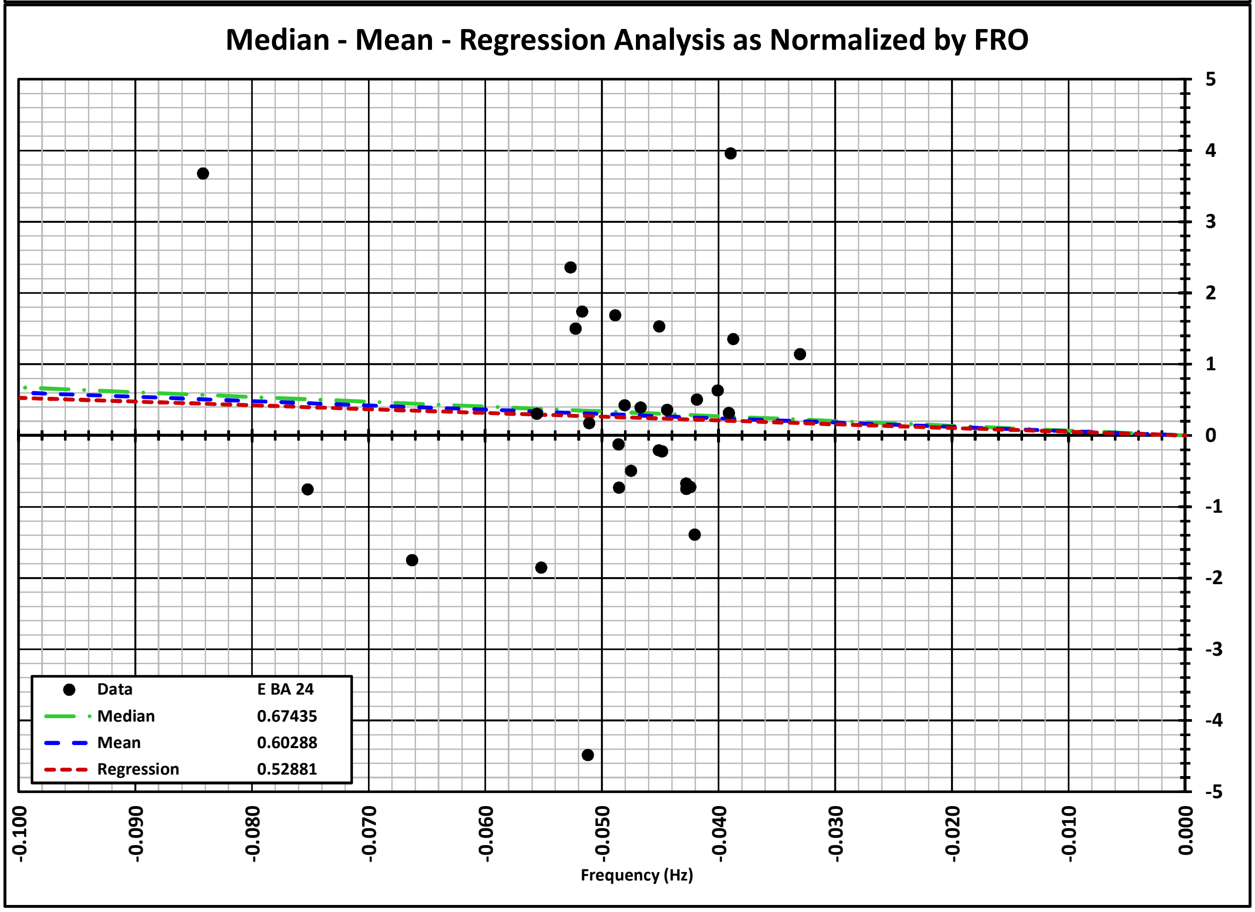


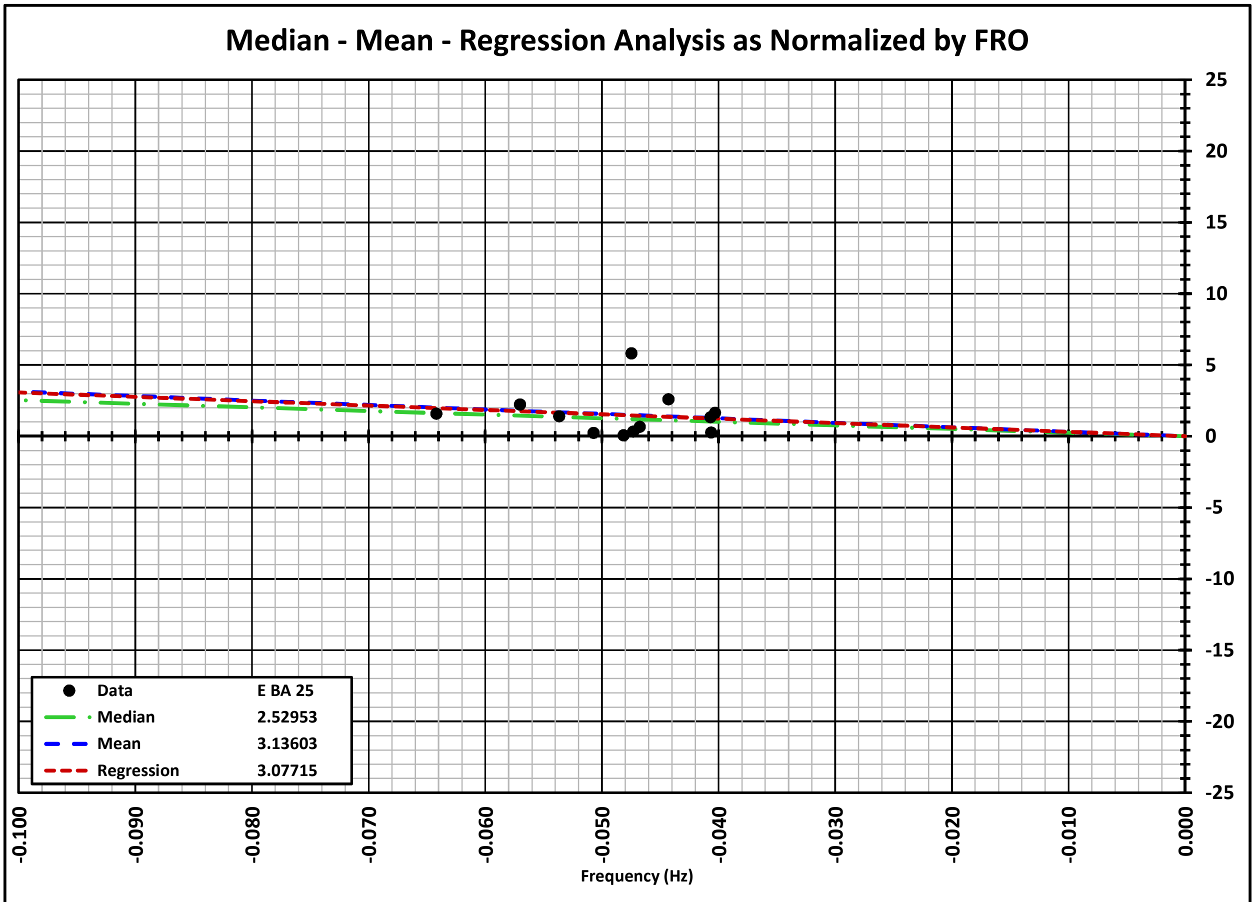
Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 21
Median 2.88295
Mean 2.22455
Regression 2.22060
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
5
4
3
2
1
0
-1
-2
-3
-4
-5

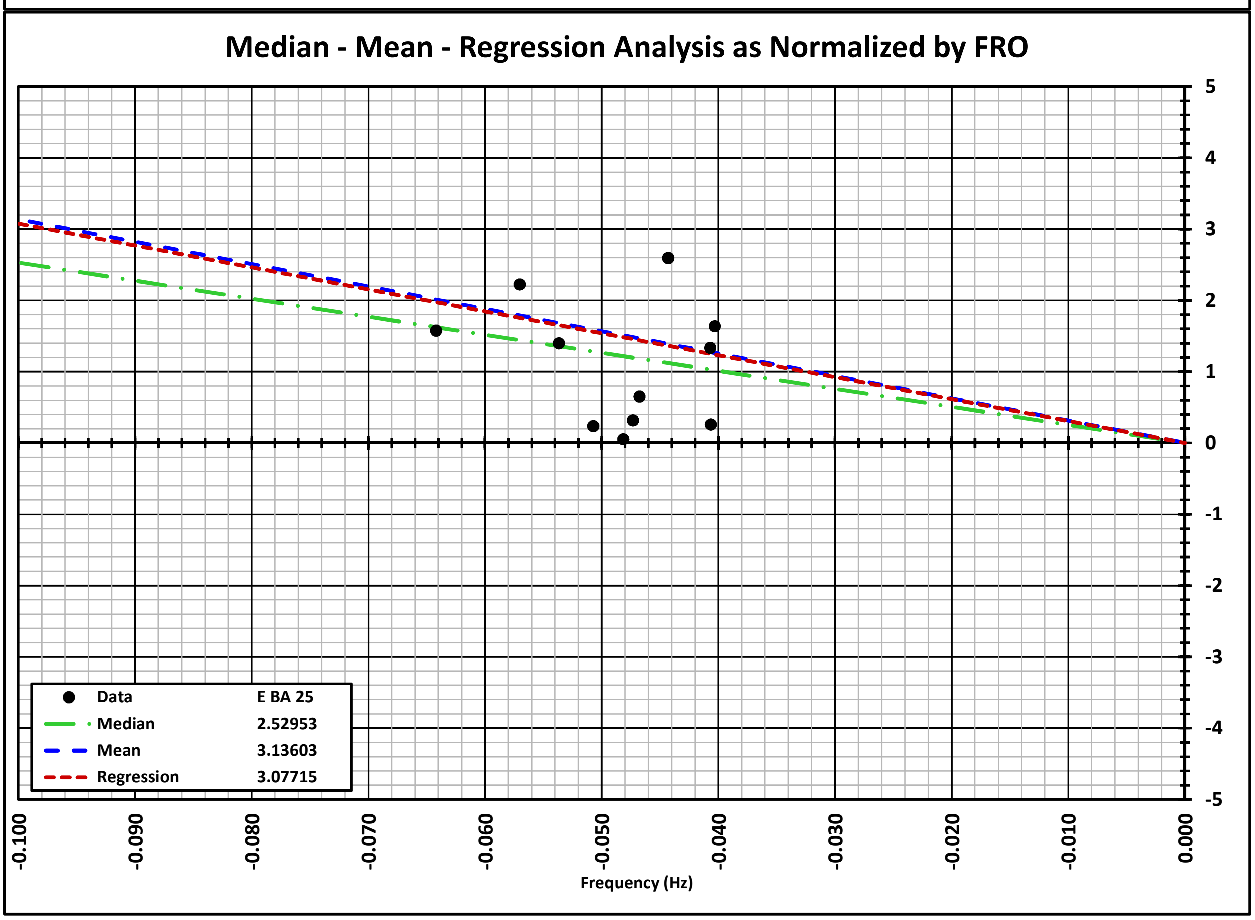


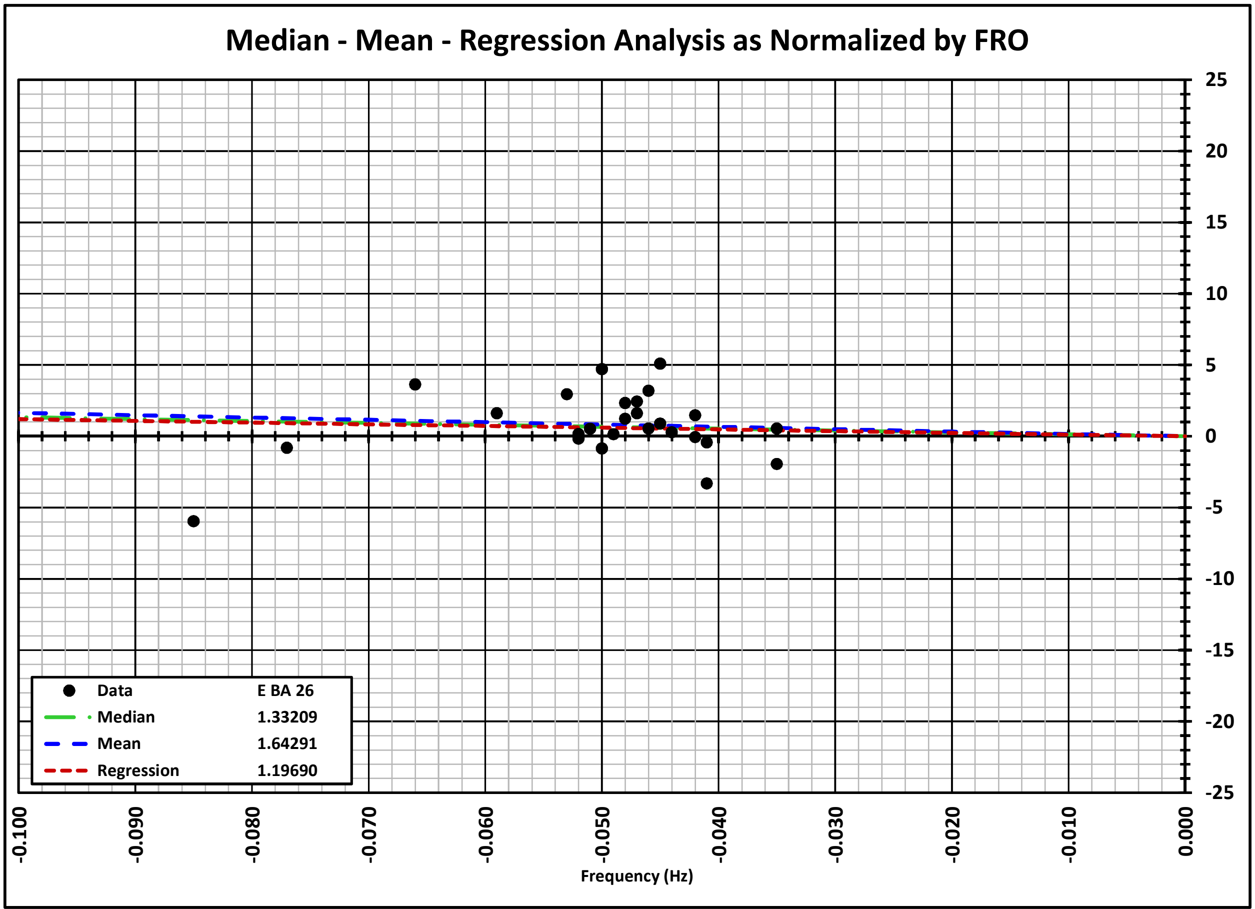
Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 22
Median 1.46450
Mean 1.24819
Regression 1.21142
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
5
4
3
2
1
0
-1
-2
-3
-4
-5

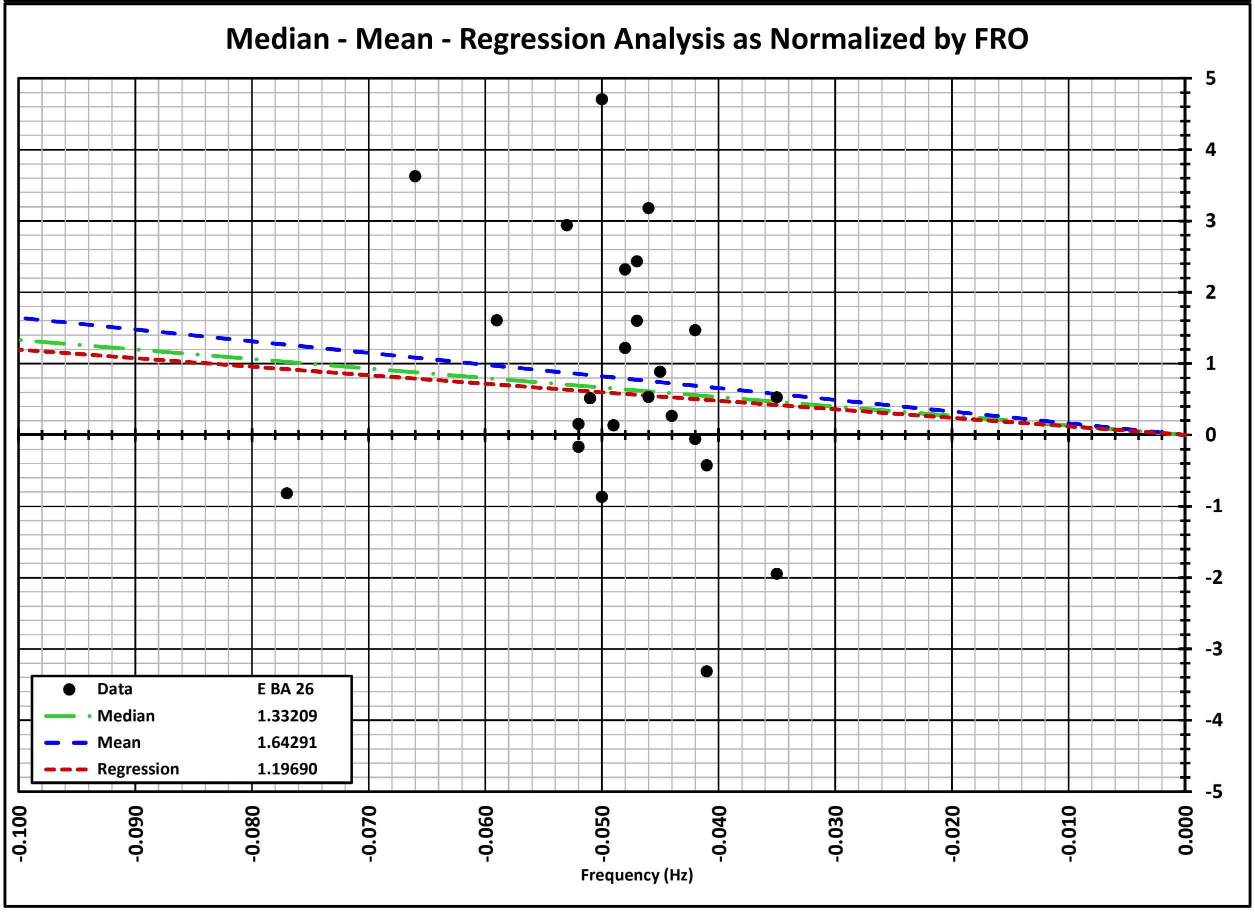

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 23
Median 1.54300
Mean 1.52179
Regression 1.56508
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
5
4
3
2
1
0
-1
-2
-3
-4
-5

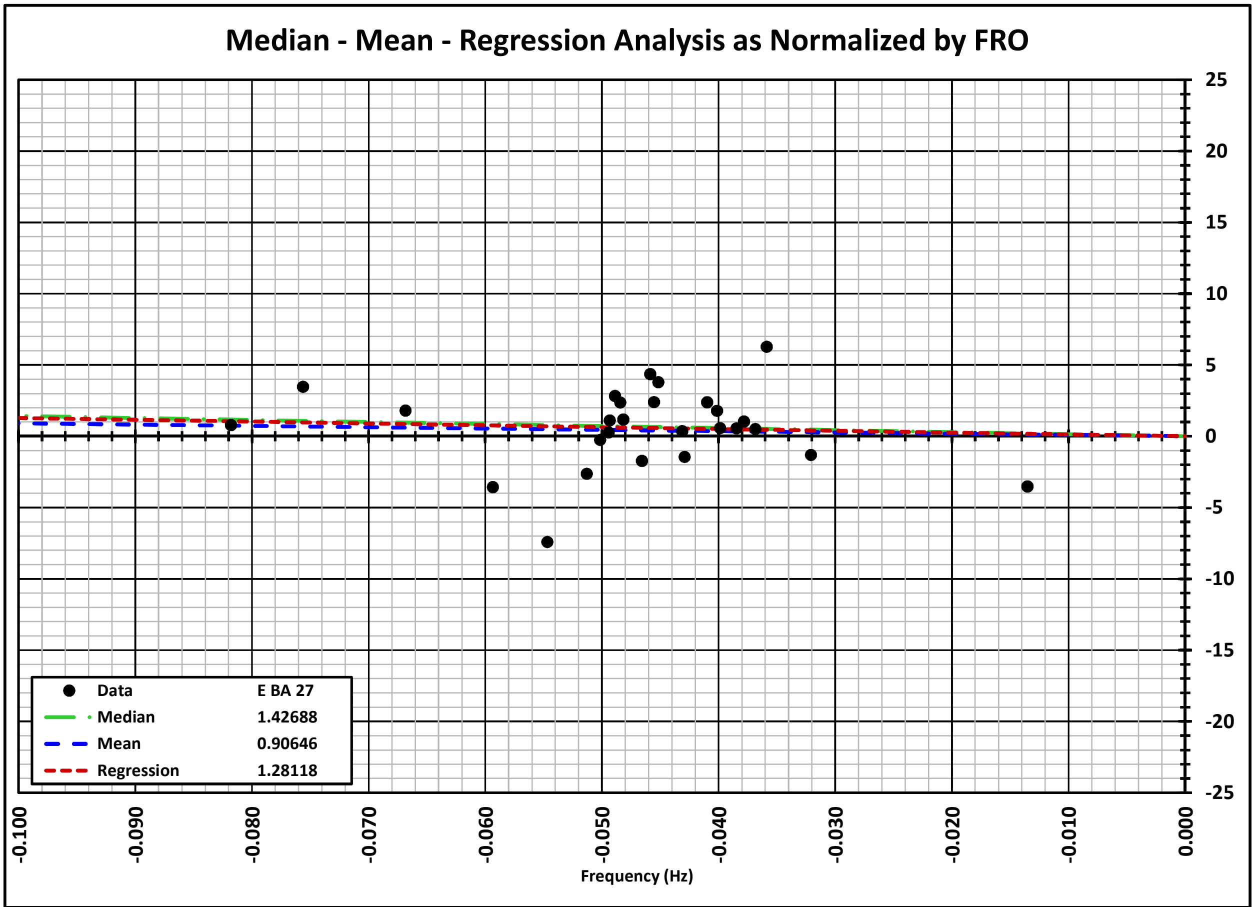

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 24
Median 0.67435
Mean 0.60288
Regression 0.52881
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
5
4
3
2
1
0
-1
-2
-3
-4
-5

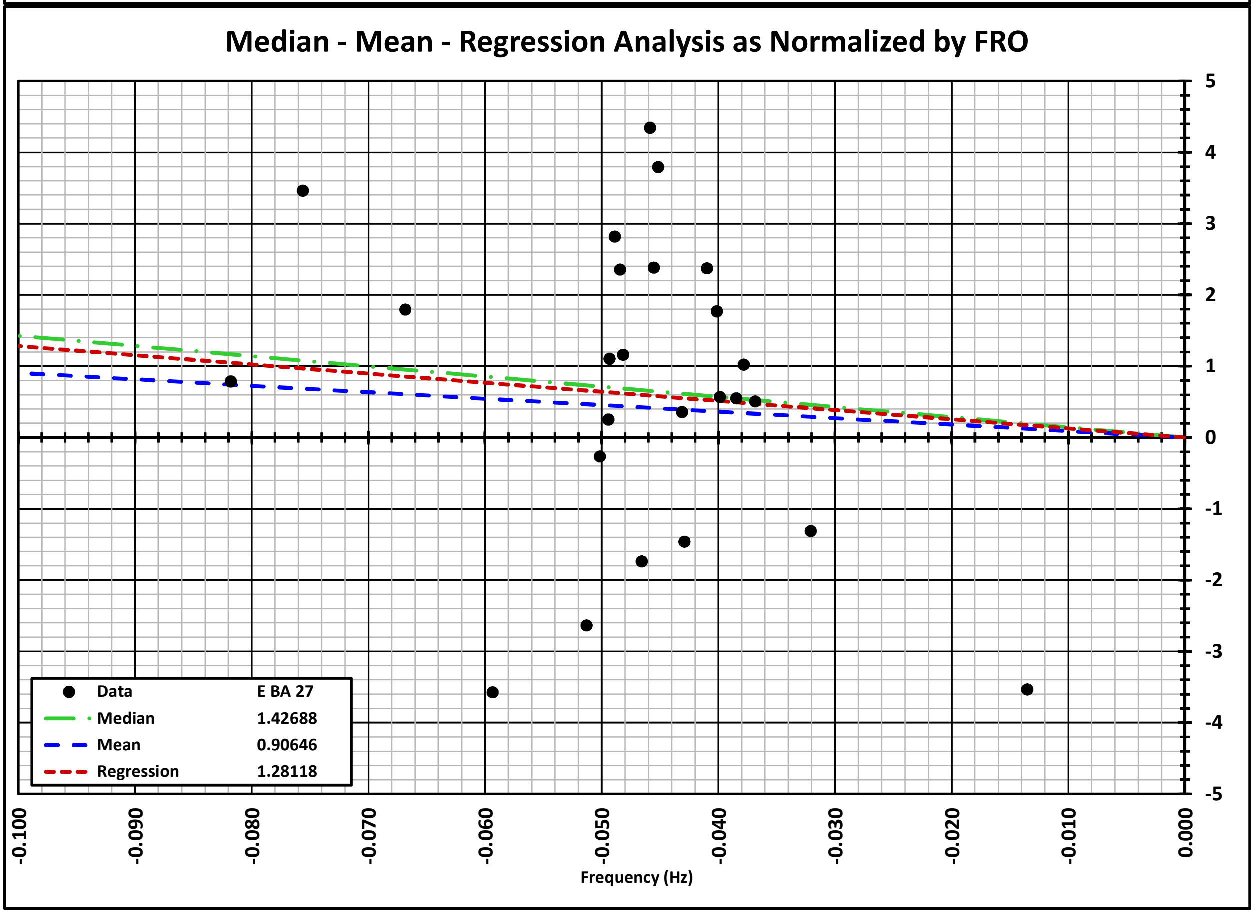

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 25
Median 2.52953
Mean 3.13603
Regression 3.07715
Frequency (Hz)
25
20
15
10
5
0
-5
-10
-15
-20
-25
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000

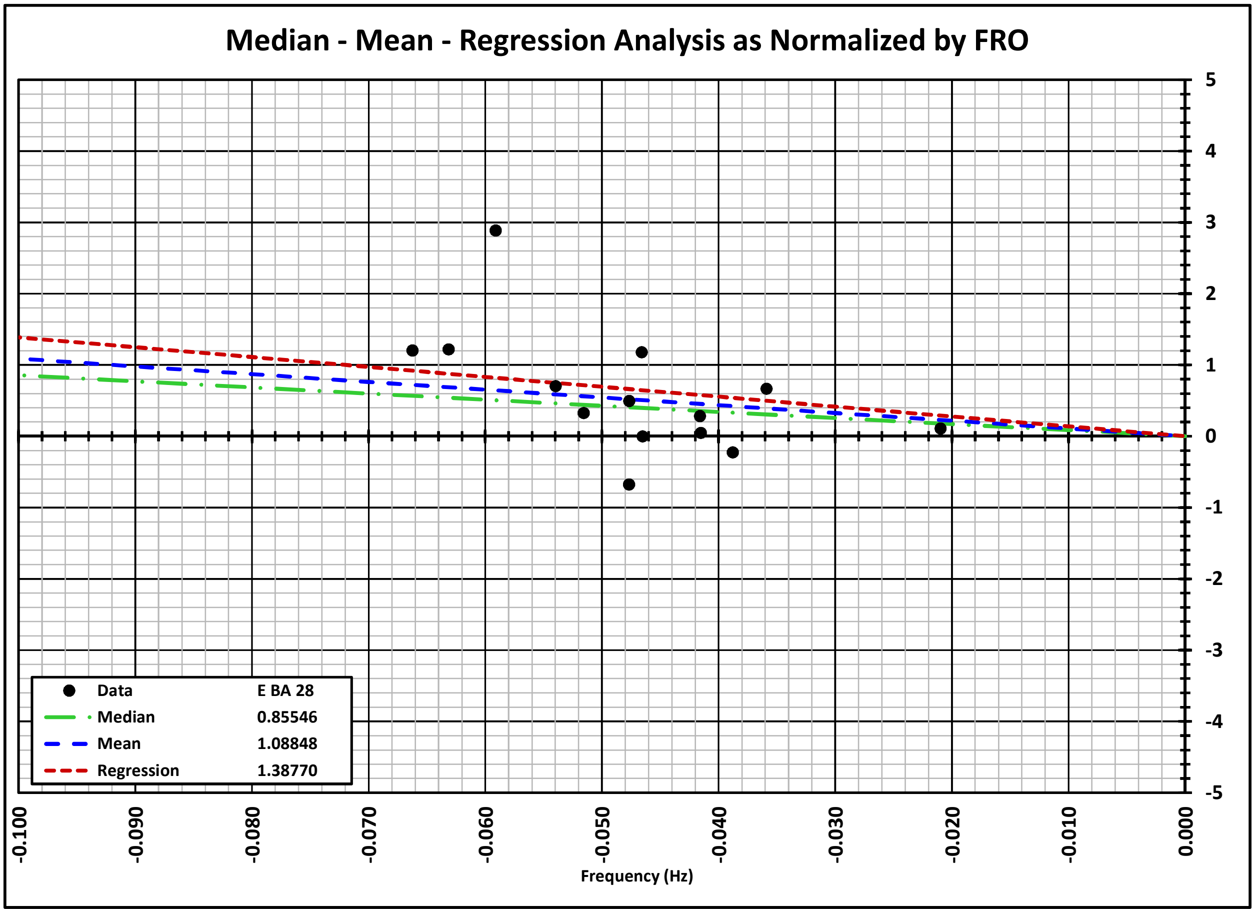

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 25
Median 2.52953
Mean 3.13603
Regression 3.07715
Frequency (Hz)
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000

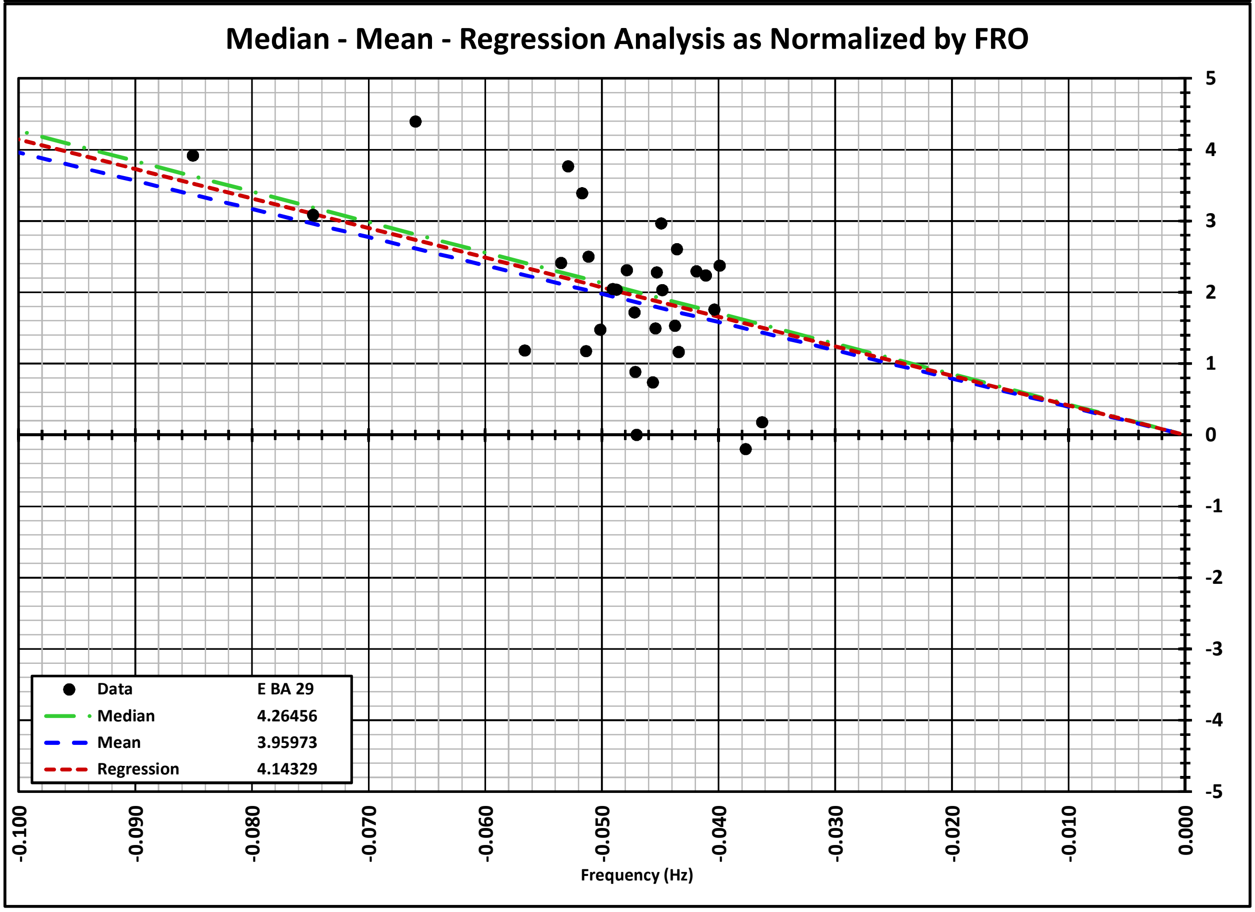

Median - Mean - Regression Analysis as Normalized by FRO
25
20
15
10
5
0
-5
-10
-15
-20
-25
Data E BA 26
Median 1.33209
Mean 1.64291
Regression 1.19690
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

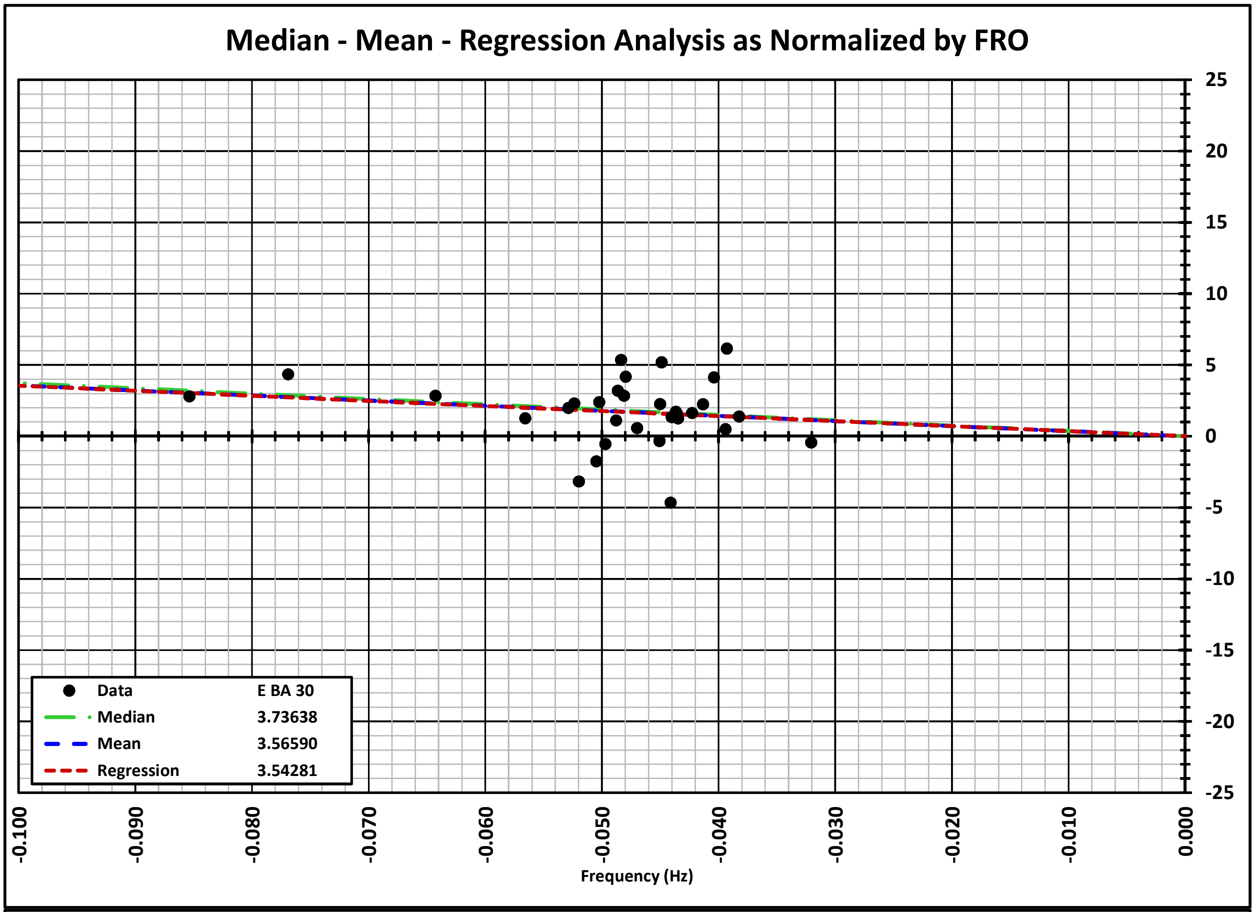

Median - Mean - Regression Analysis as Normalized by FRO
5
4
3
2
1
0
-1
-2
-3
-4
-5
Data E BA 26
Median 1.33209
Mean 1.64291
Regression 1.19690
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

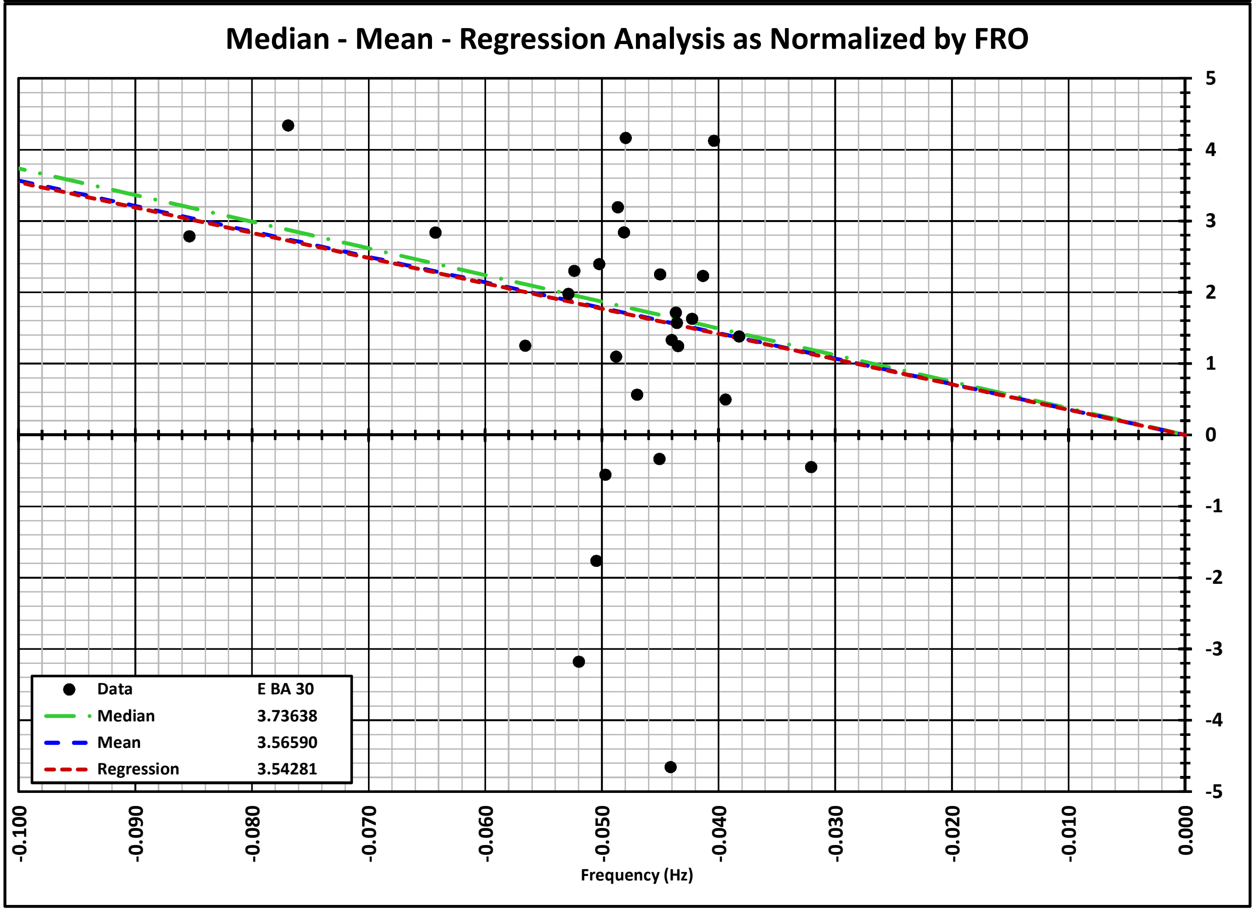

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 27
Median 1.42688
Mean 0.90646
Regression 1.28118
Frequency (Hz)
25
20
15
10
5
0
-5
-10
-15
-20
-25
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000

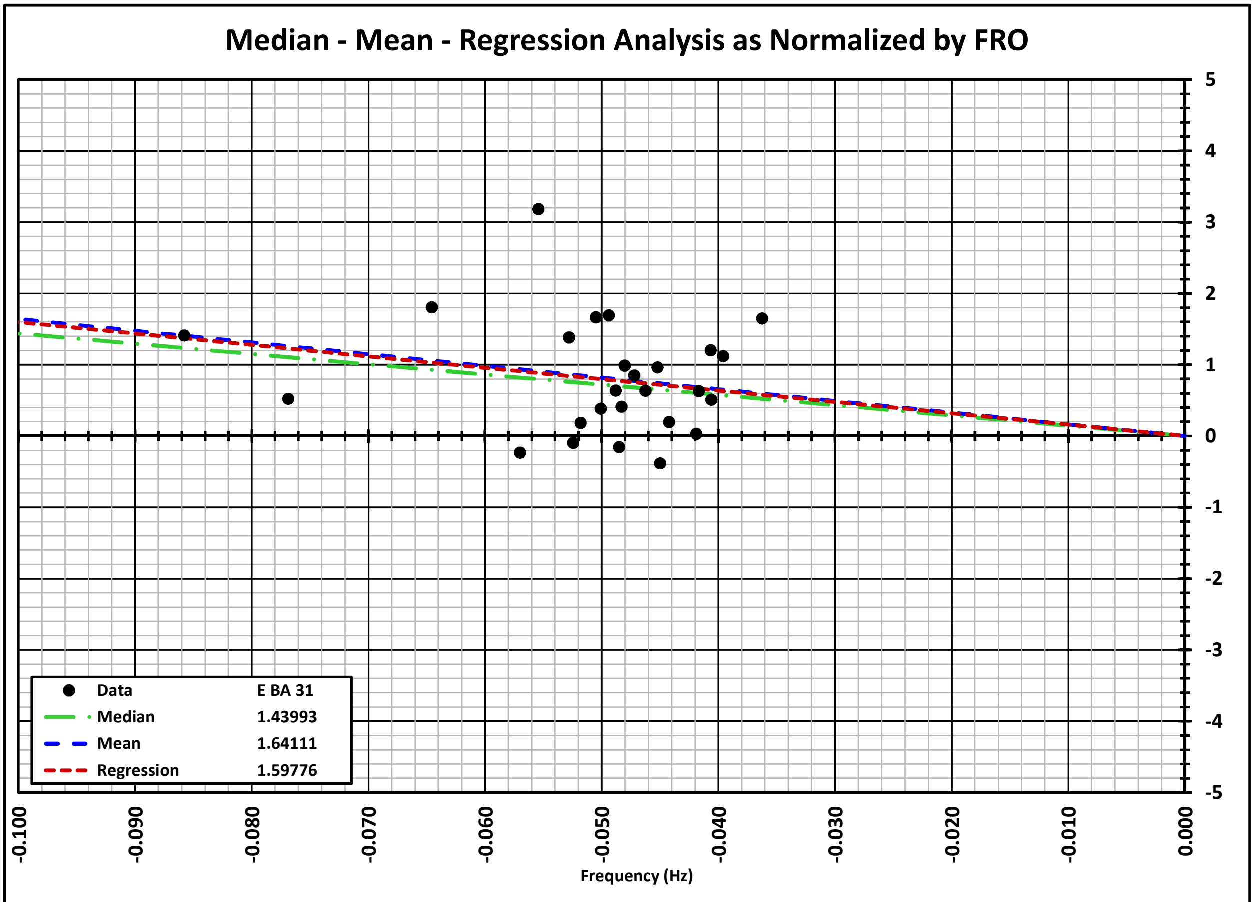

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 27
Median 1.42688
Mean 0.90646
Regression 1.28118
Frequency (Hz)
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000

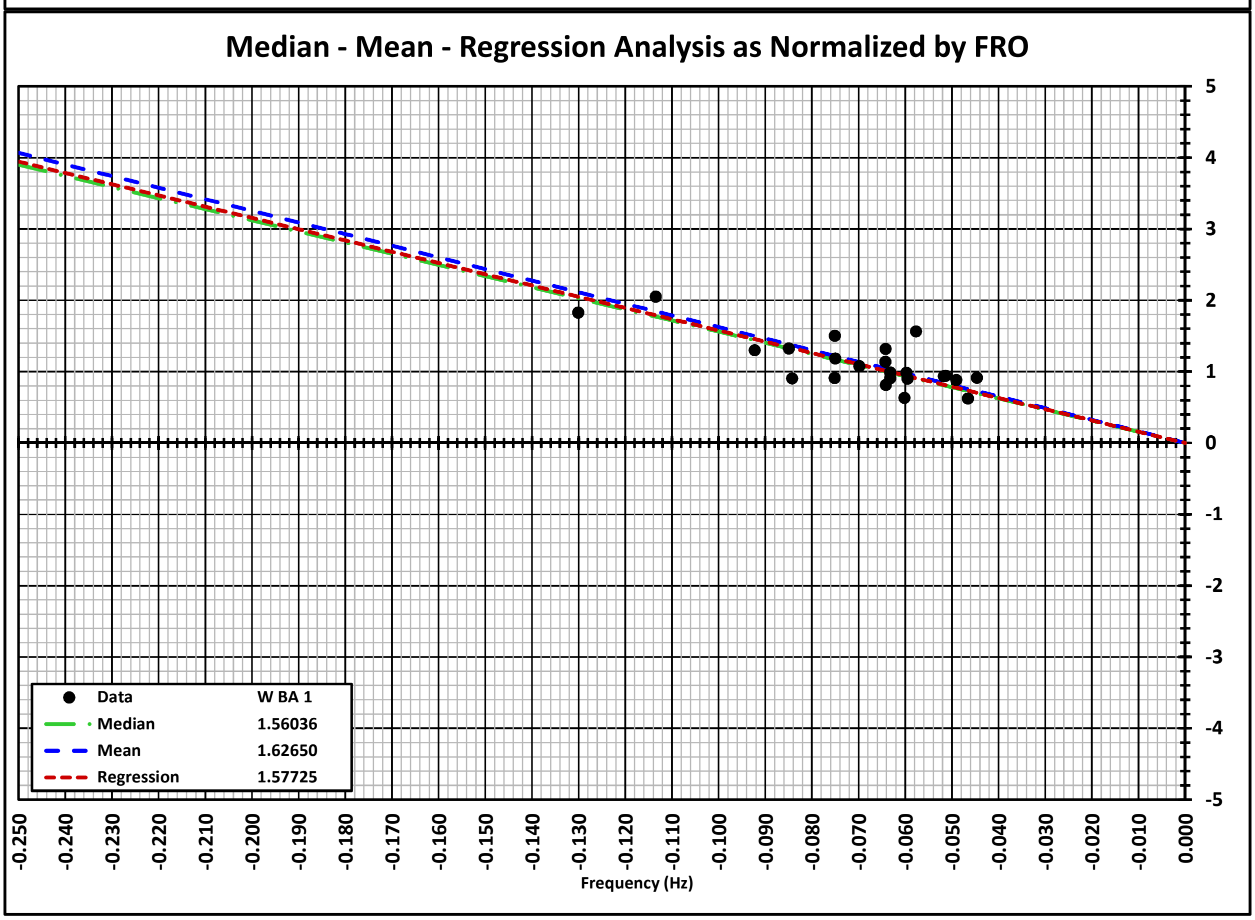

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 28
Median 0.85546
Mean 1.08848
Regression 1.38770
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

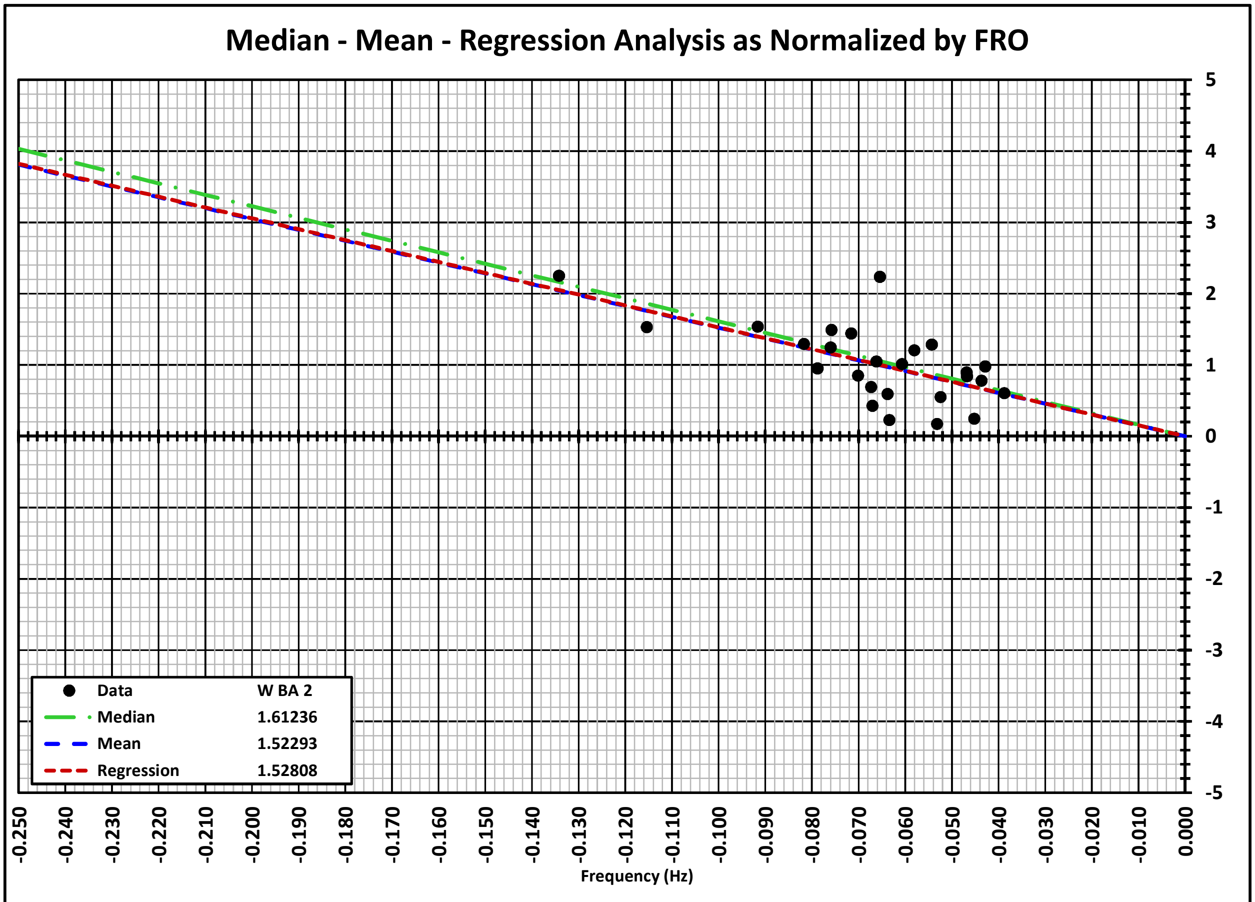

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 29
Median 4.26456
Mean 3.95973
Regression 4.14329
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

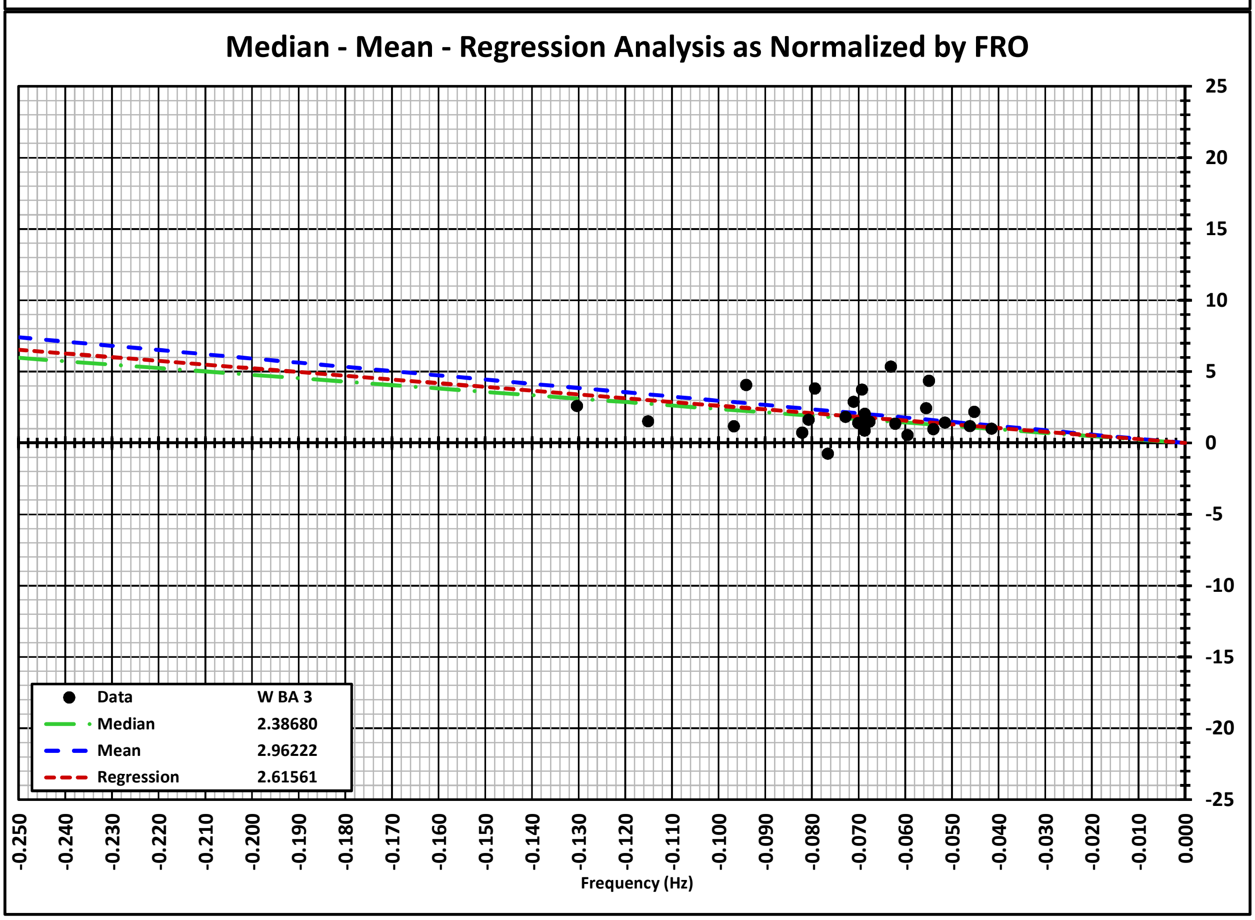

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 30
Median 3.73638
Mean 3.56590
Regression 3.54281
Frequency (Hz)

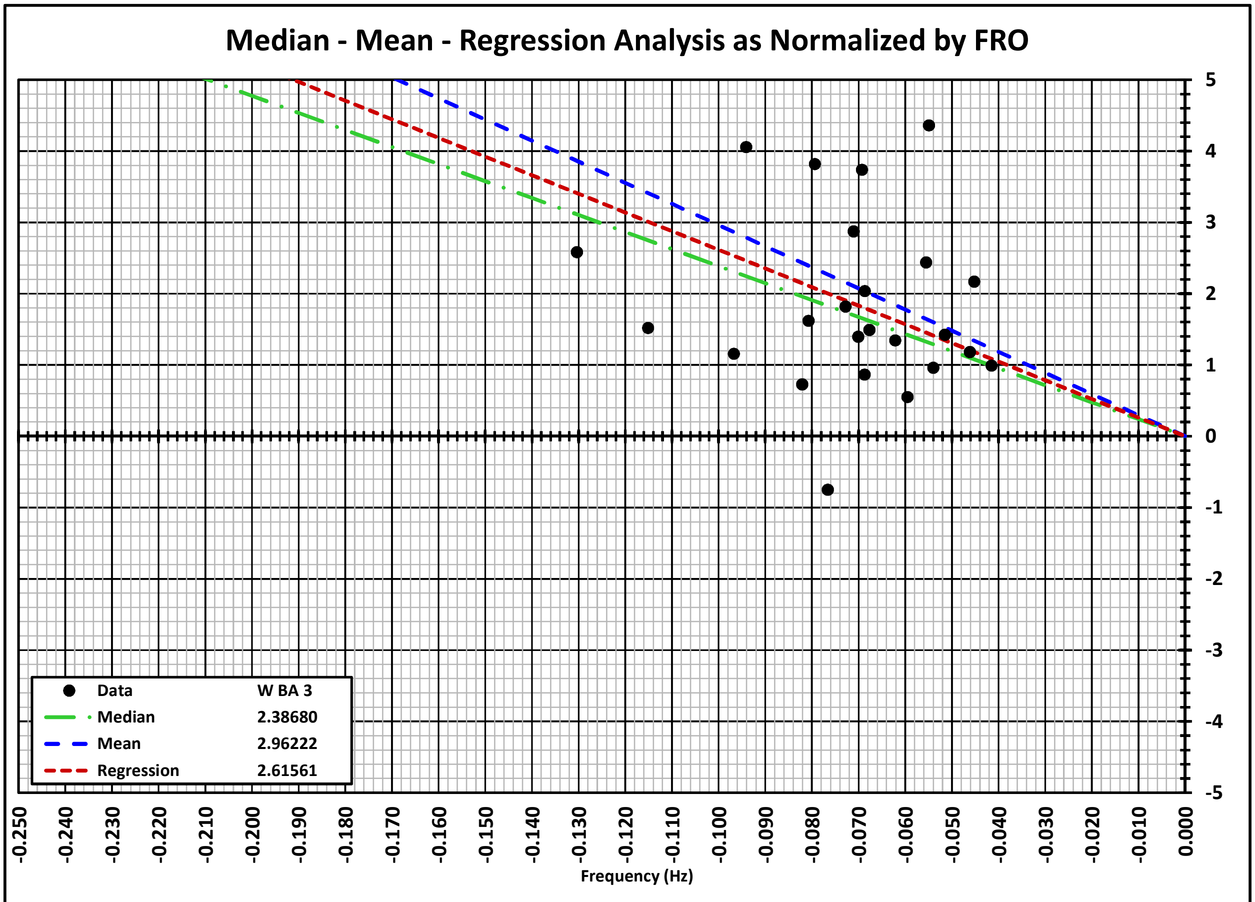

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 30
Median 3.73638
Mean 3.56590
Regression 3.54281
Frequency (Hz)

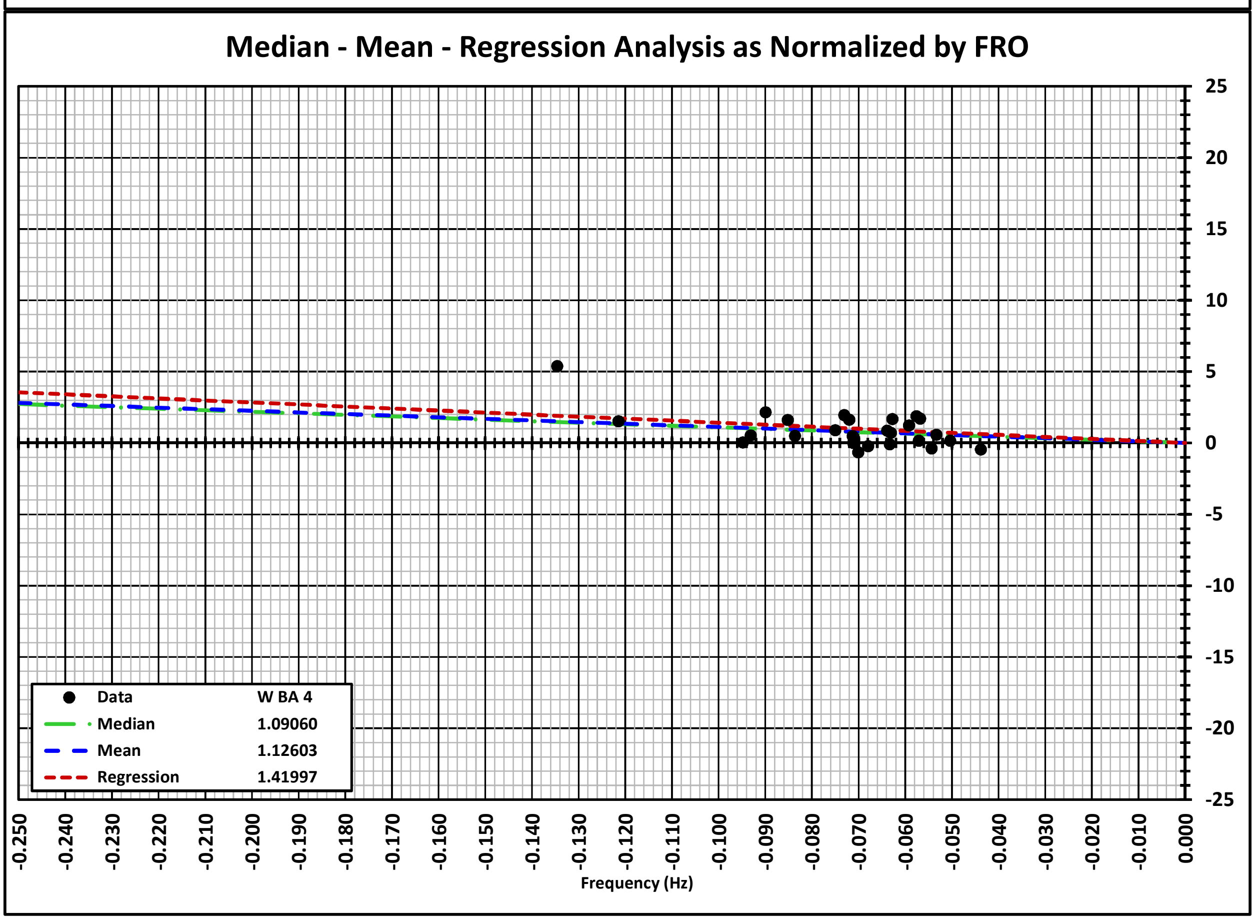

Median - Mean - Regression Analysis as Normalized by FRO
Data E BA 31
Median 1.43993
Mean 1.64111
Regression 1.59776
Frequency (Hz)
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
5
4
3
2
1
0
-1
-2
-3
-4
-5

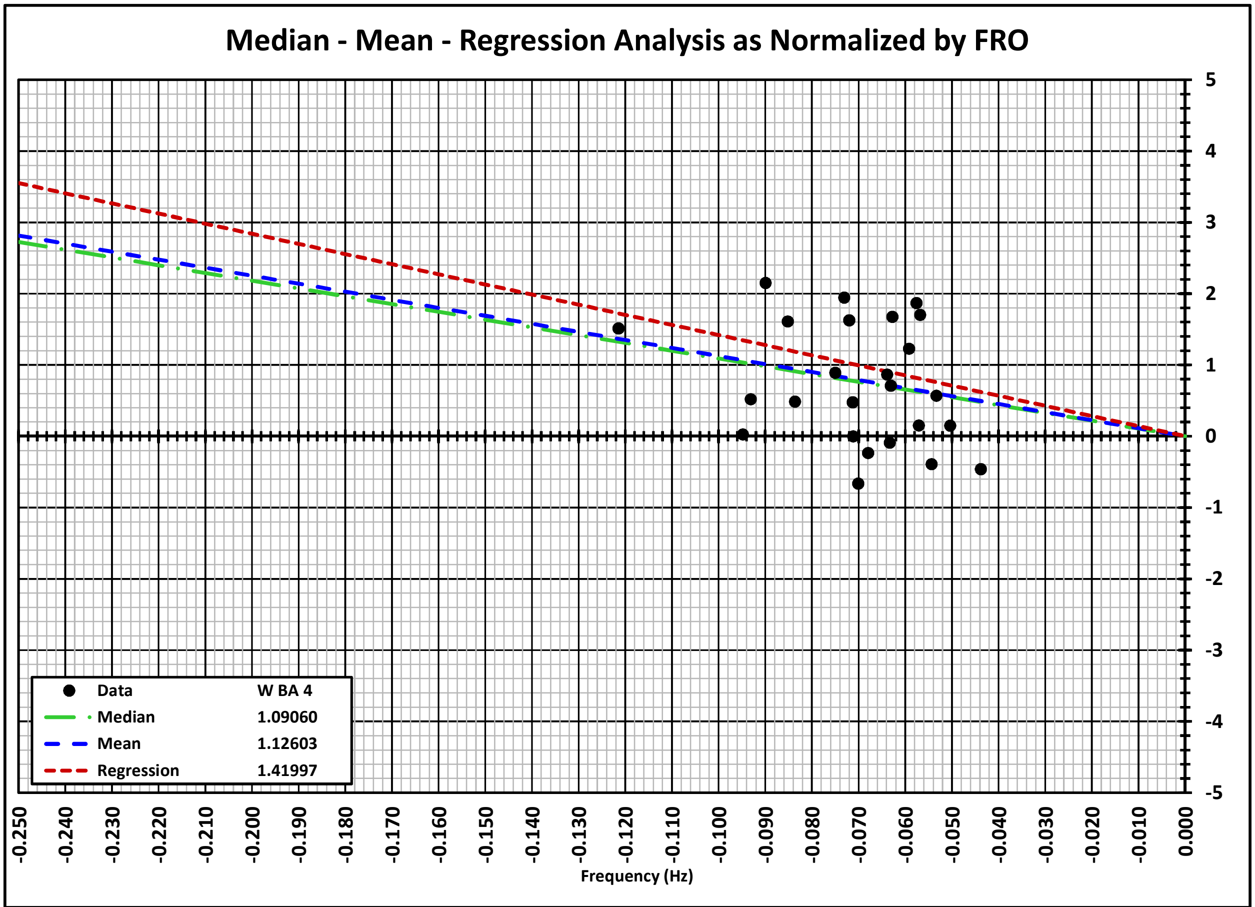

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 1
Median 1.56036
Mean 1.62650
Regression 1.57725
Frequency (Hz)
-0.250
-0.240
-0.230
-0.220
-0.210
-0.200
-0.190
-0.180
-0.170
-0.160
-0.150
-0.140
-0.130
-0.120
-0.110
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
5
4
3
2
1
0
-1
-2
-3
-4
-5

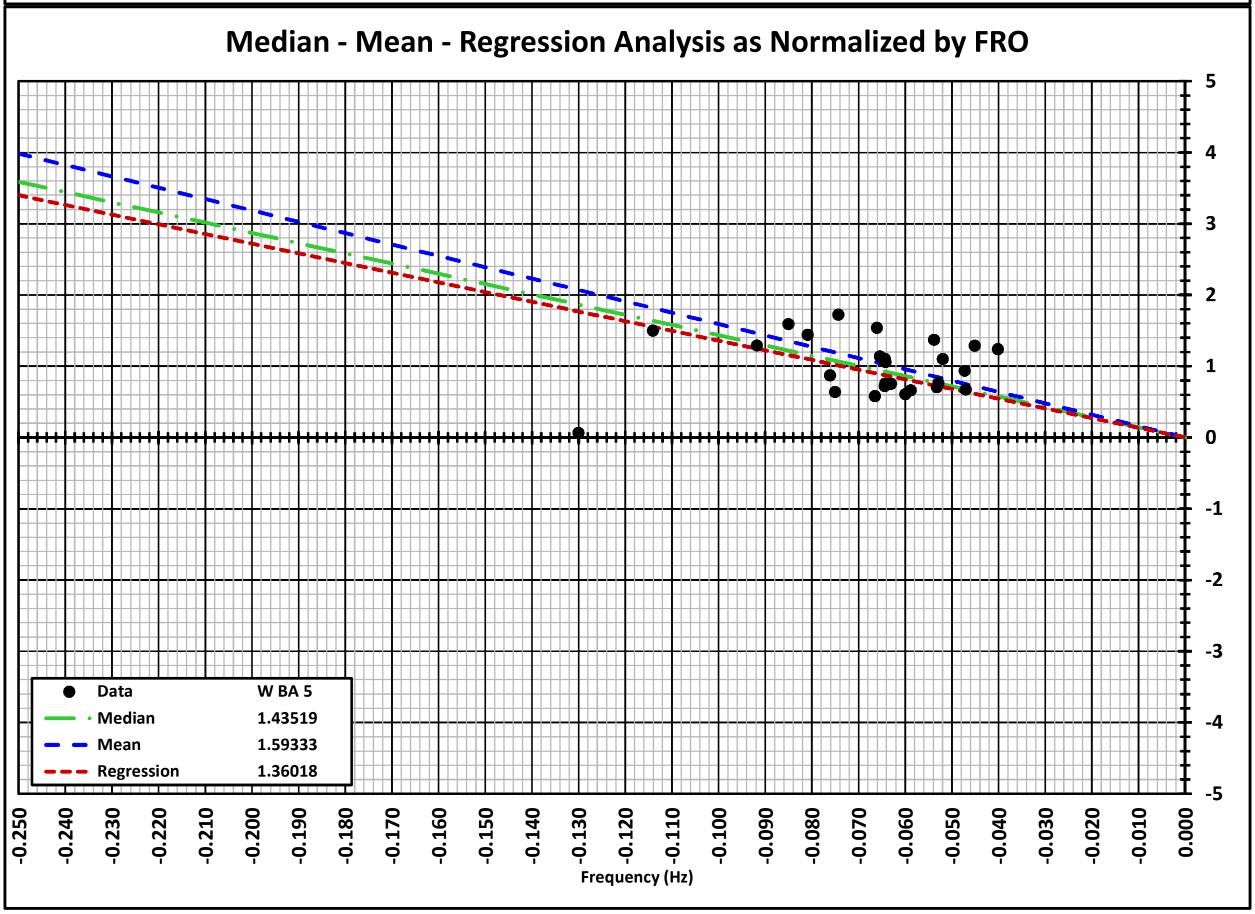

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 2
Median 1.61236
Mean 1.52293
Regression 1.52808
Frequency (Hz)
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.250
-0.240
-0.230
-0.220
-0.210
-0.200
-0.190
-0.180
-0.170
-0.160
-0.150
-0.140
-0.130
-0.120
-0.110
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000

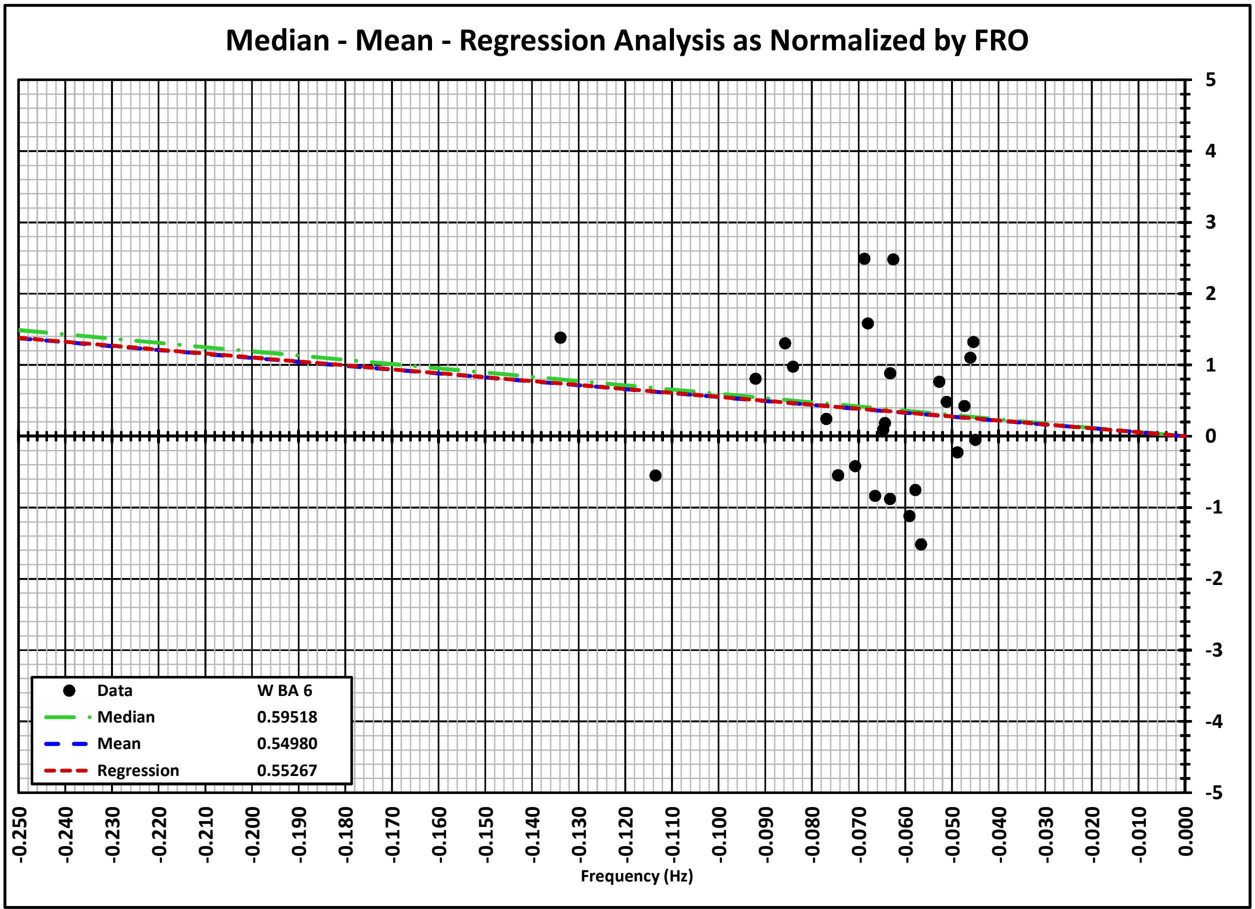

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 3
Median 2.38680
Mean 2.96222
Regression 2.61561
Frequency (Hz)
25
20
15
10
5
0
-5
-10
-15
-20
-25
-0.250
-0.240
-0.230
-0.220
-0.210
-0.200
-0.190
-0.180
-0.170
-0.160
-0.150
-0.140
-0.130
-0.120
-0.110
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000

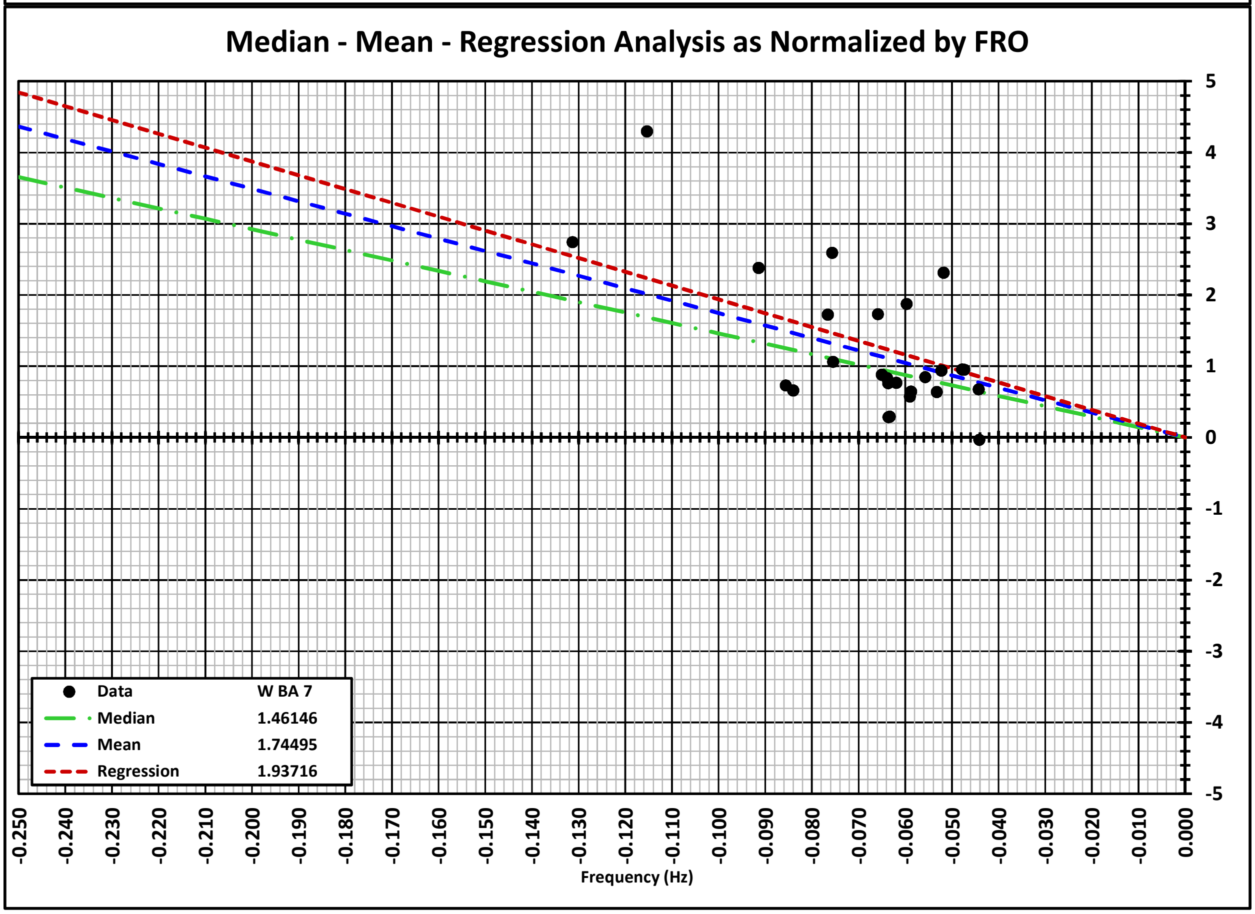

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 3
Median 2.38680
Mean 2.96222
Regression 2.61561
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.250
-0.240
-0.230
-0.220
-0.210
-0.200
-0.190
-0.180
-0.170
-0.160
-0.150
-0.140
-0.130
-0.120
-0.110
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

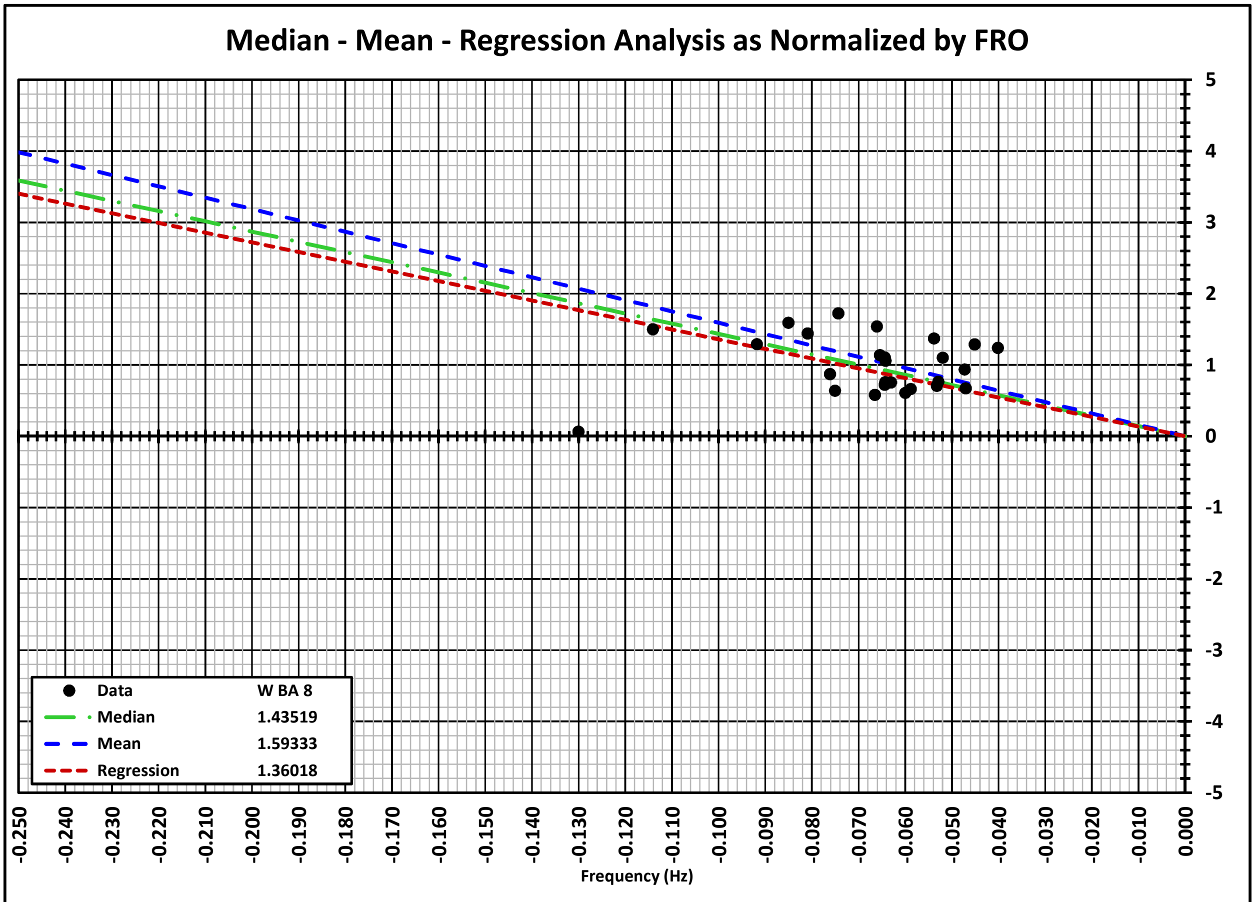

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 4
Median 1.09060
Mean 1.12603
Regression 1.41997
25
20
15
10
5
0
-5
-10
-15
-20
-25
-0.250
-0.240
-0.230
-0.220
-0.210
-0.200
-0.190
-0.180
-0.170
-0.160
-0.150
-0.140
-0.130
-0.120
-0.110
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

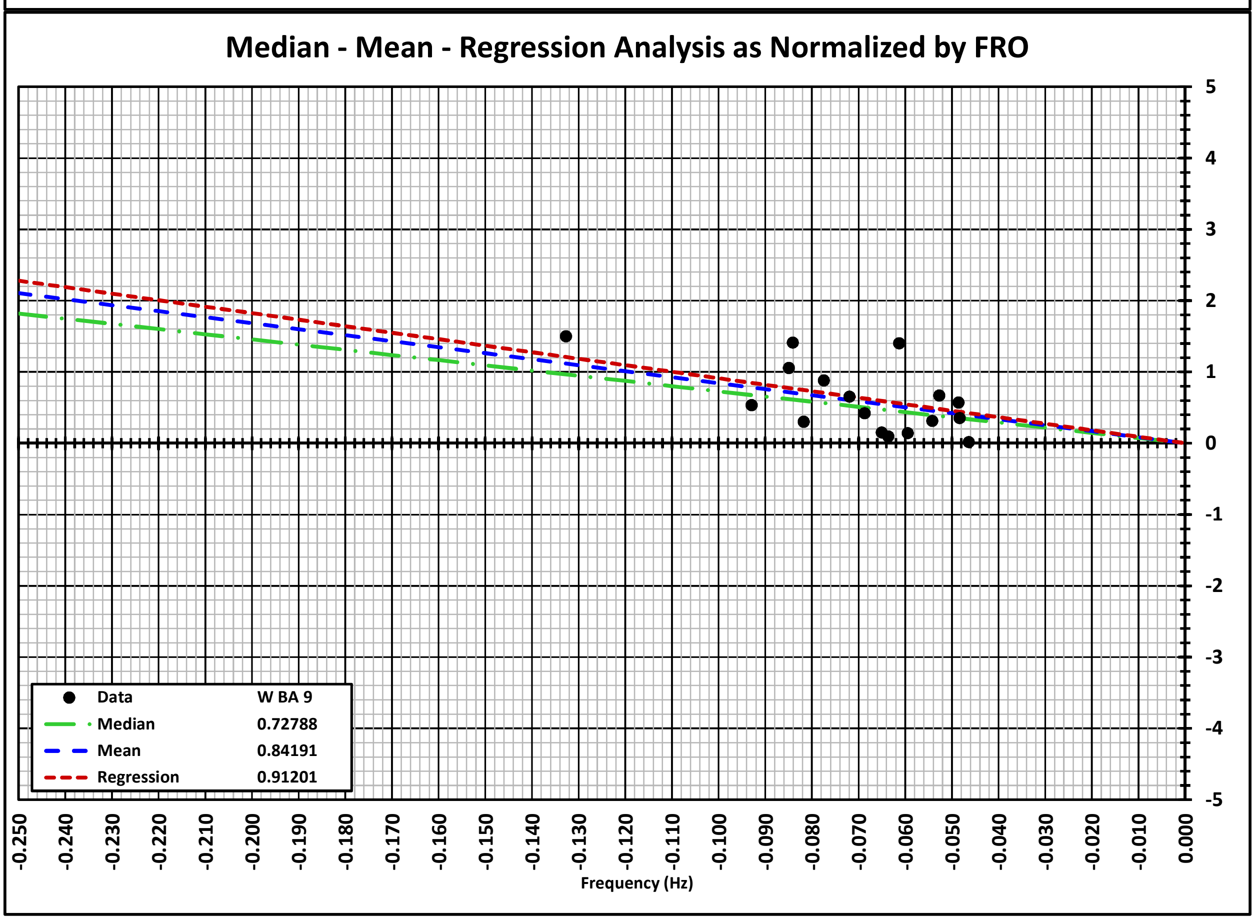

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 4
Median 1.09060
Mean 1.12603
Regression 1.41997
Frequency (Hz)

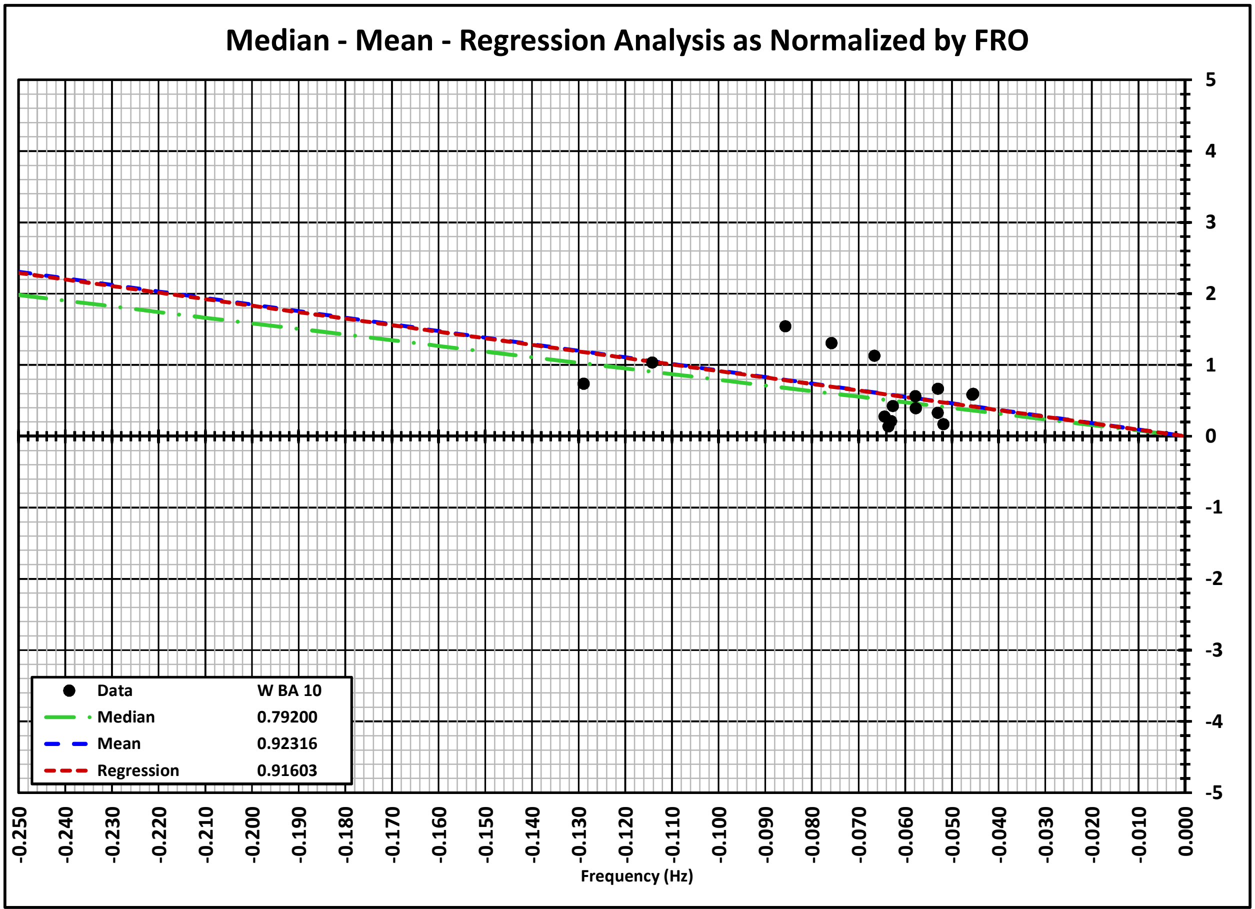

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 5
Median 1.43519
Mean 1.59333
Regression 1.36018
Frequency (Hz)

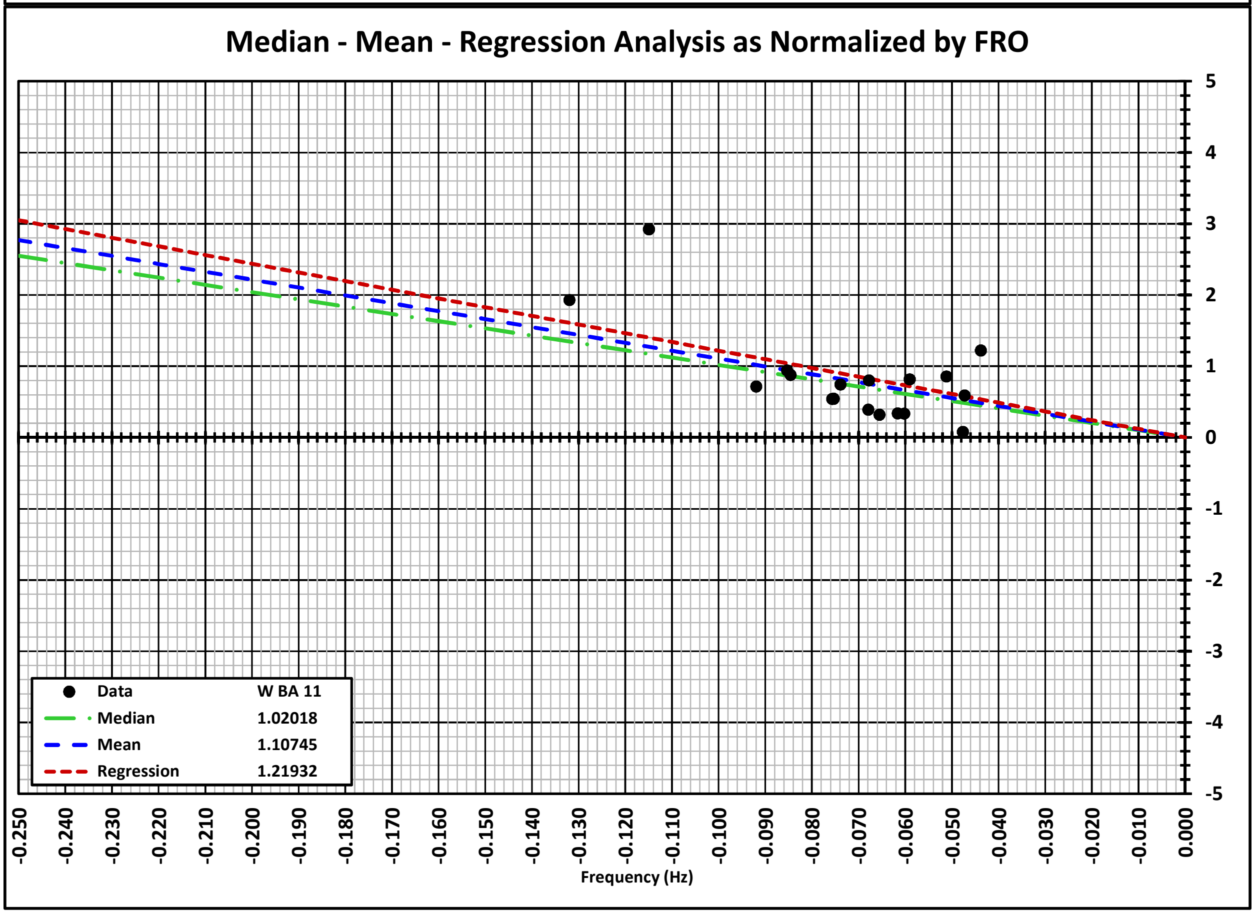

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 6
Median 0.59518
Mean 0.54980
Regression 0.55267
Frequency (Hz)
-0.250 -0.240 -0.230 -0.220 -0.210 -0.200 -0.190 -0.180 -0.170 -0.160 -0.150 -0.140 -0.130 -0.120 -0.110 -0.100 -0.090 -0.080 -0.070 -0.060 -0.050 -0.040 -0.030 -0.020 -0.010 0.000
5 4 3 2 1 0 -1 -2 -3 -4 -5

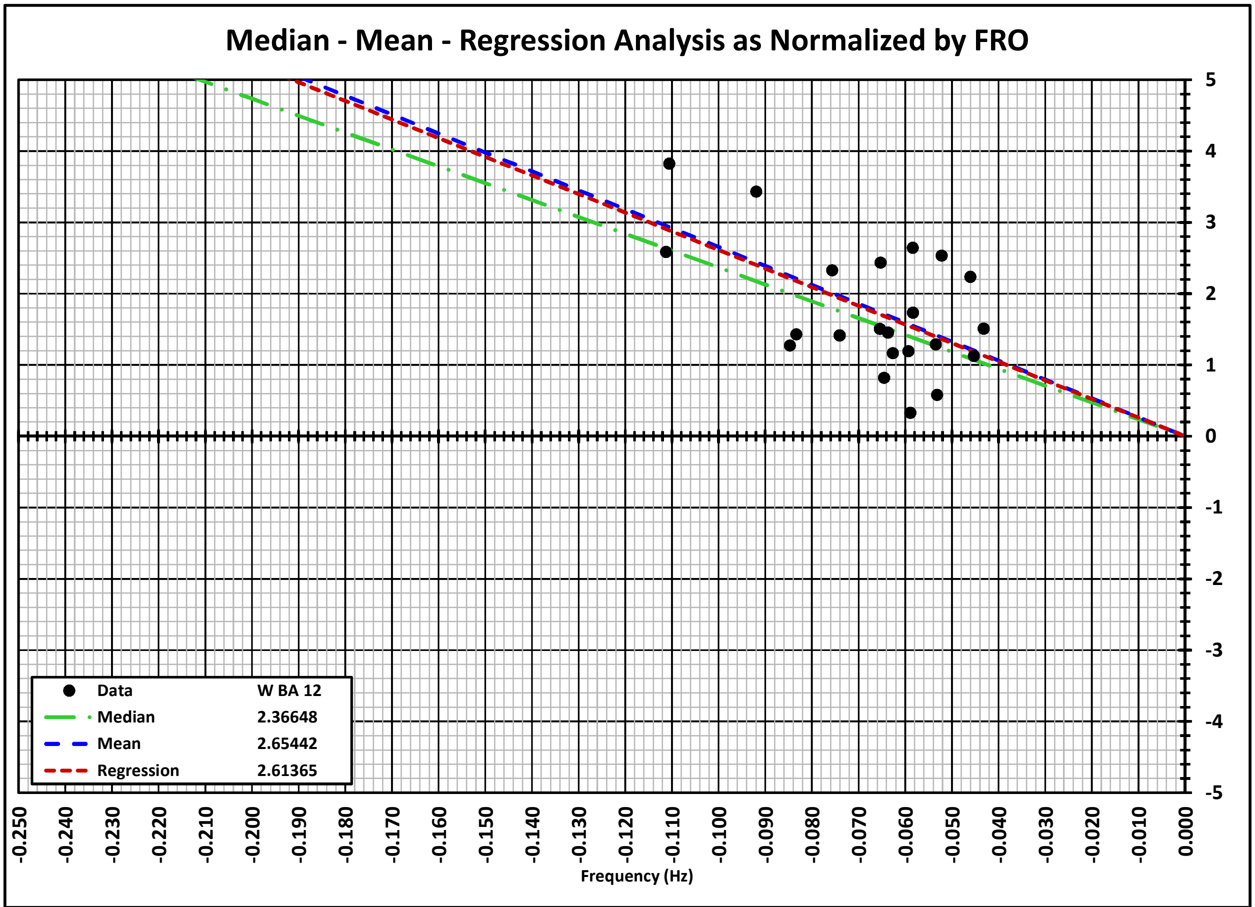

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 7
Median 1.46146
Mean 1.74495
Regression 1.93716
Frequency (Hz)
-0.250 -0.240 -0.230 -0.220 -0.210 -0.200 -0.190 -0.180 -0.170 -0.160 -0.150 -0.140 -0.130 -0.120 -0.110 -0.100 -0.090 -0.080 -0.070 -0.060 -0.050 -0.040 -0.030 -0.020 -0.010 0.000
5 4 3 2 1 0 -1 -2 -3 -4 -5

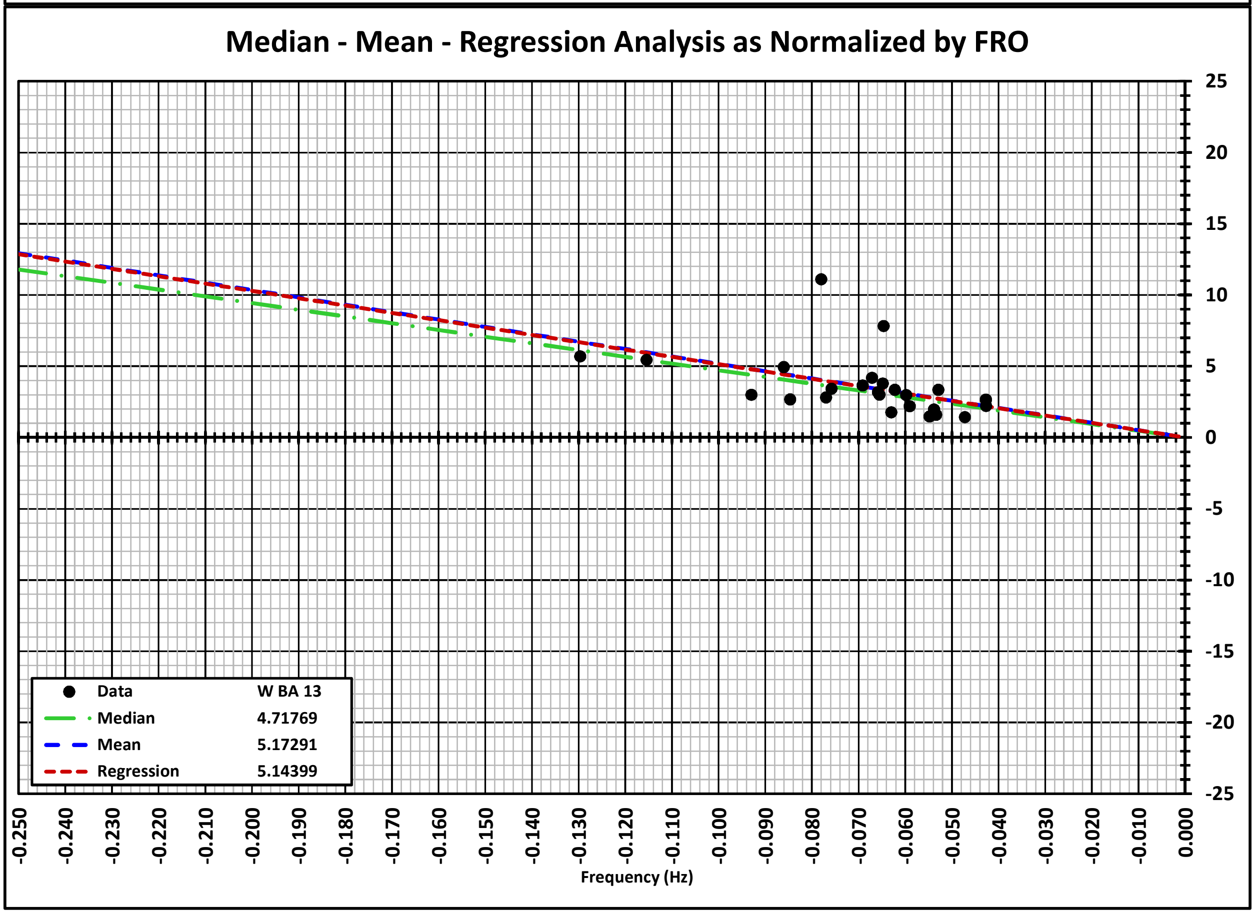

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 8
Median 1.43519
Mean 1.59333
Regression 1.36018
Frequency (Hz)
-0.250 -0.240 -0.230 -0.220 -0.210 -0.200 -0.190 -0.180 -0.170 -0.160 -0.150 -0.140 -0.130 -0.120 -0.110 -0.100 -0.090 -0.080 -0.070 -0.060 -0.050 -0.040 -0.030 -0.020 -0.010 0.000
5 4 3 2 1 0 -1 -2 -3 -4 -5

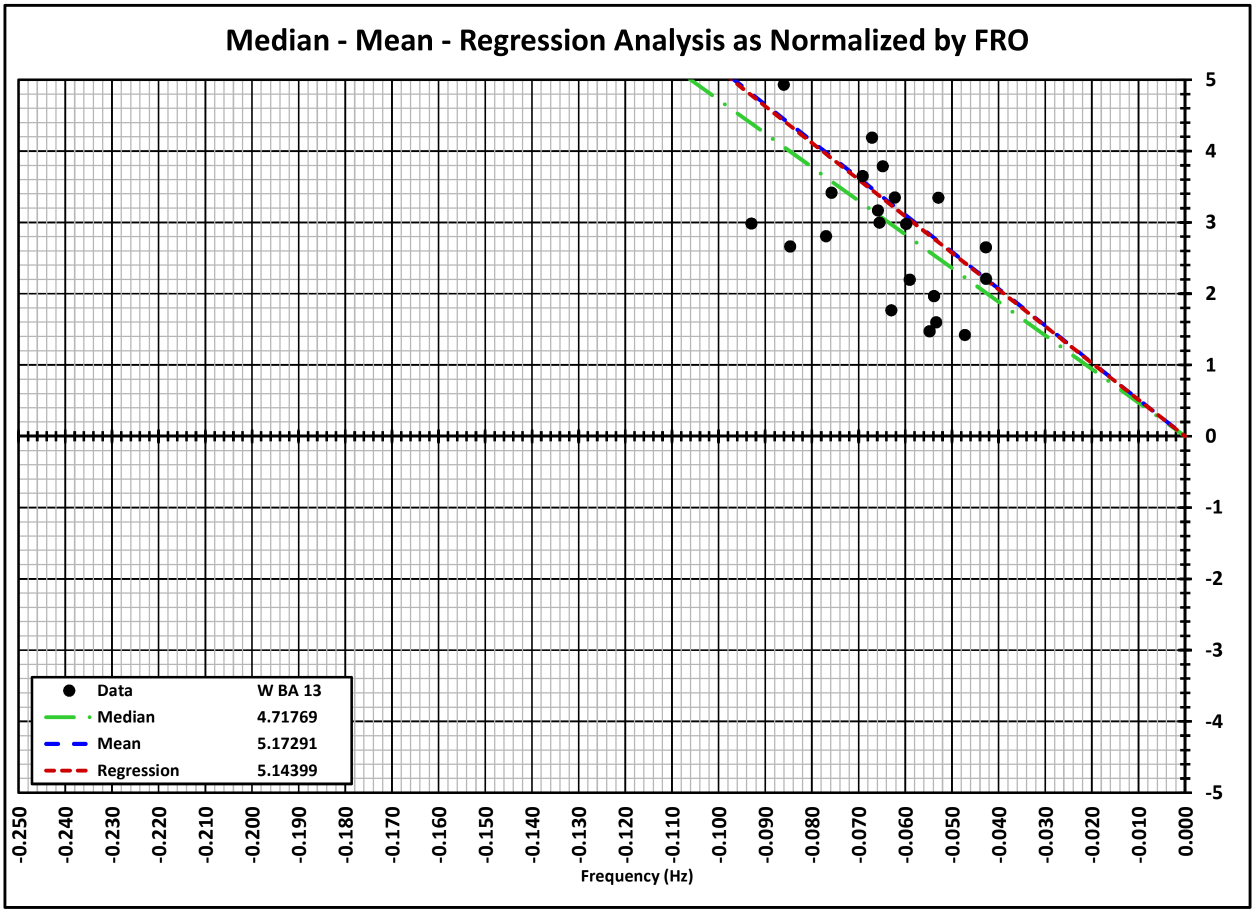

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 9
Median 0.72788
Mean 0.84191
Regression 0.91201
Frequency (Hz)
-0.250 -0.240 -0.230 -0.220 -0.210 -0.200 -0.190 -0.180 -0.170 -0.160 -0.150 -0.140 -0.130 -0.120 -0.110 -0.100 -0.090 -0.080 -0.070 -0.060 -0.050 -0.040 -0.030 -0.020 -0.010 0.000
5 4 3 2 1 0 -1 -2 -3 -4 -5

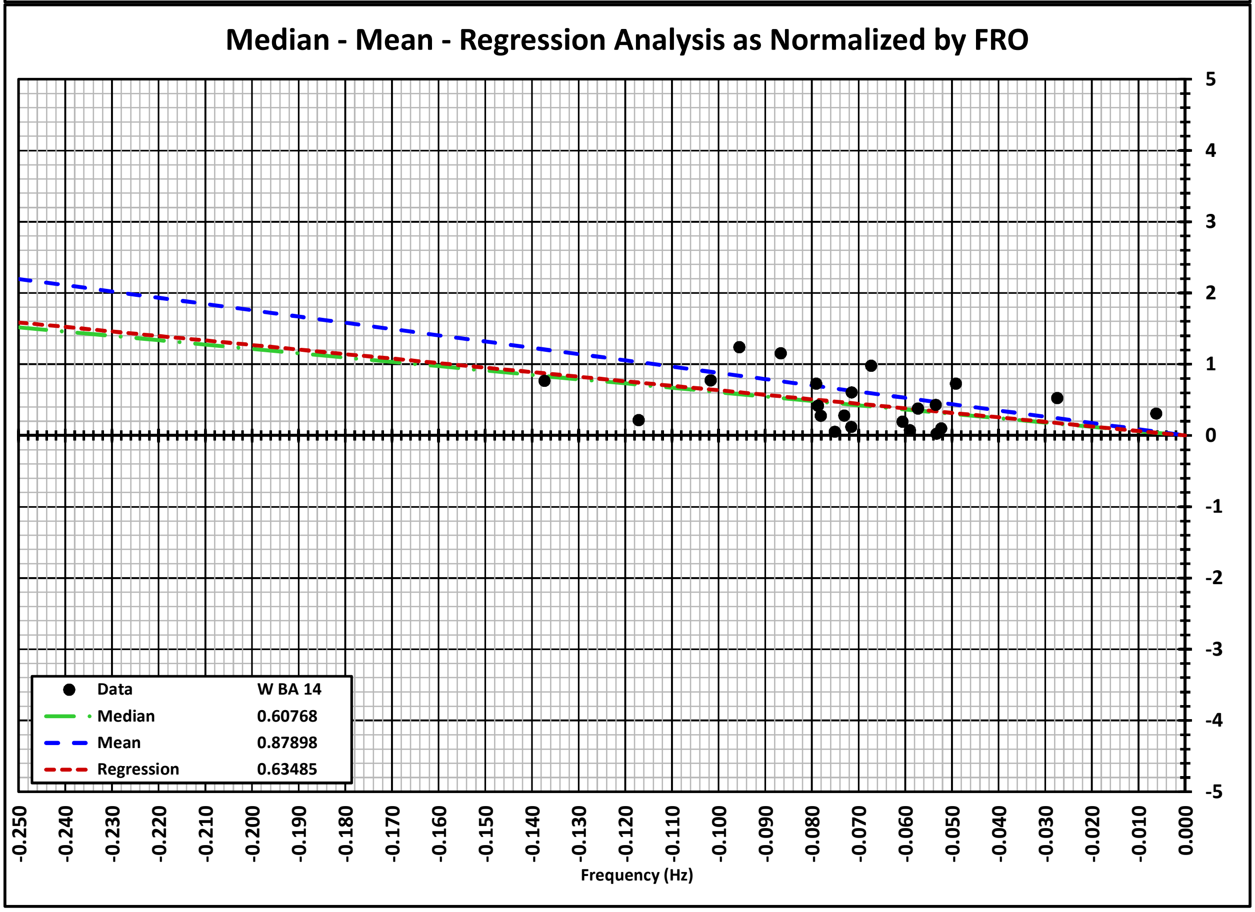

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 10
Median 0.79200
Mean 0.92316
Regression 0.91603
Frequency (Hz)

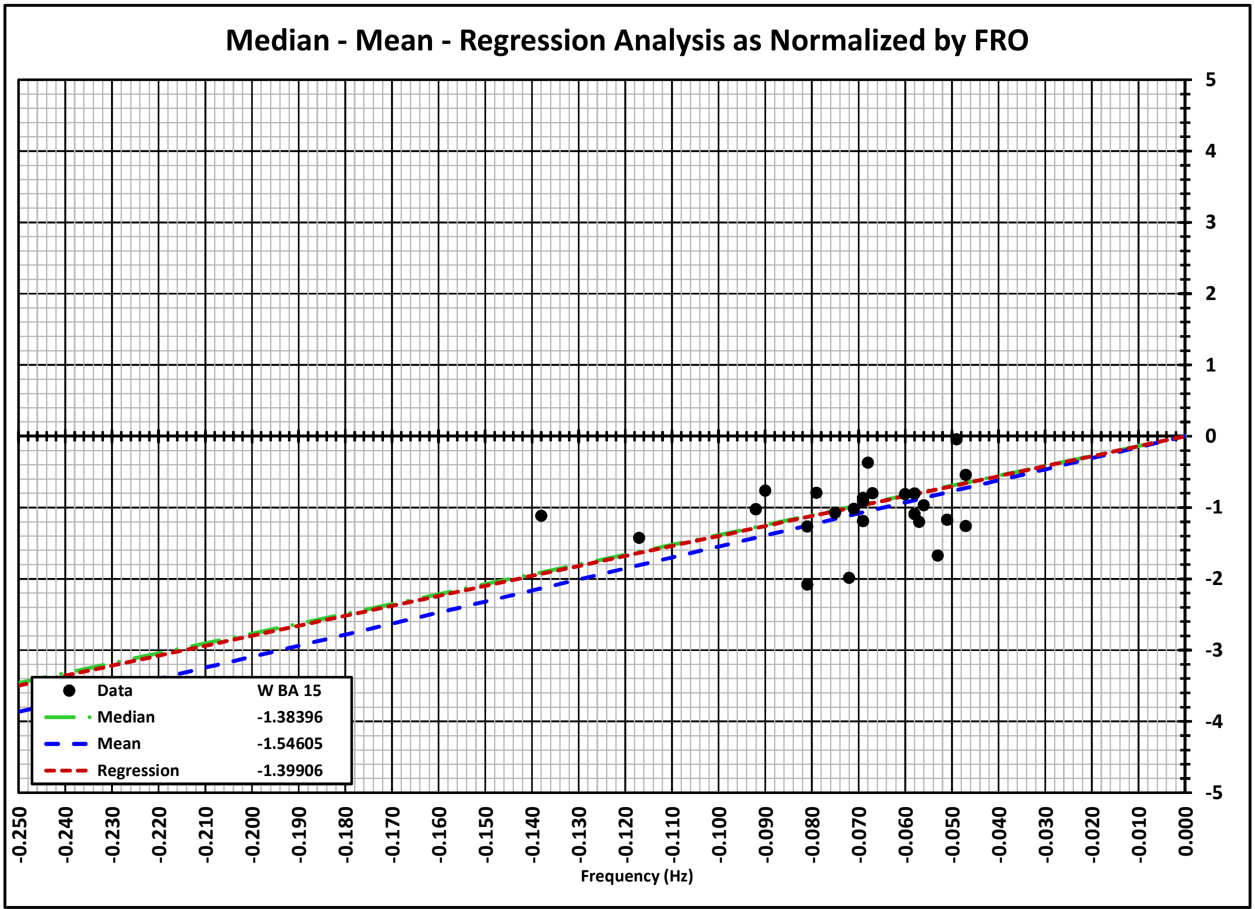

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 11
Median 1.02018
Mean 1.10745
Regression 1.21932
Frequency (Hz)

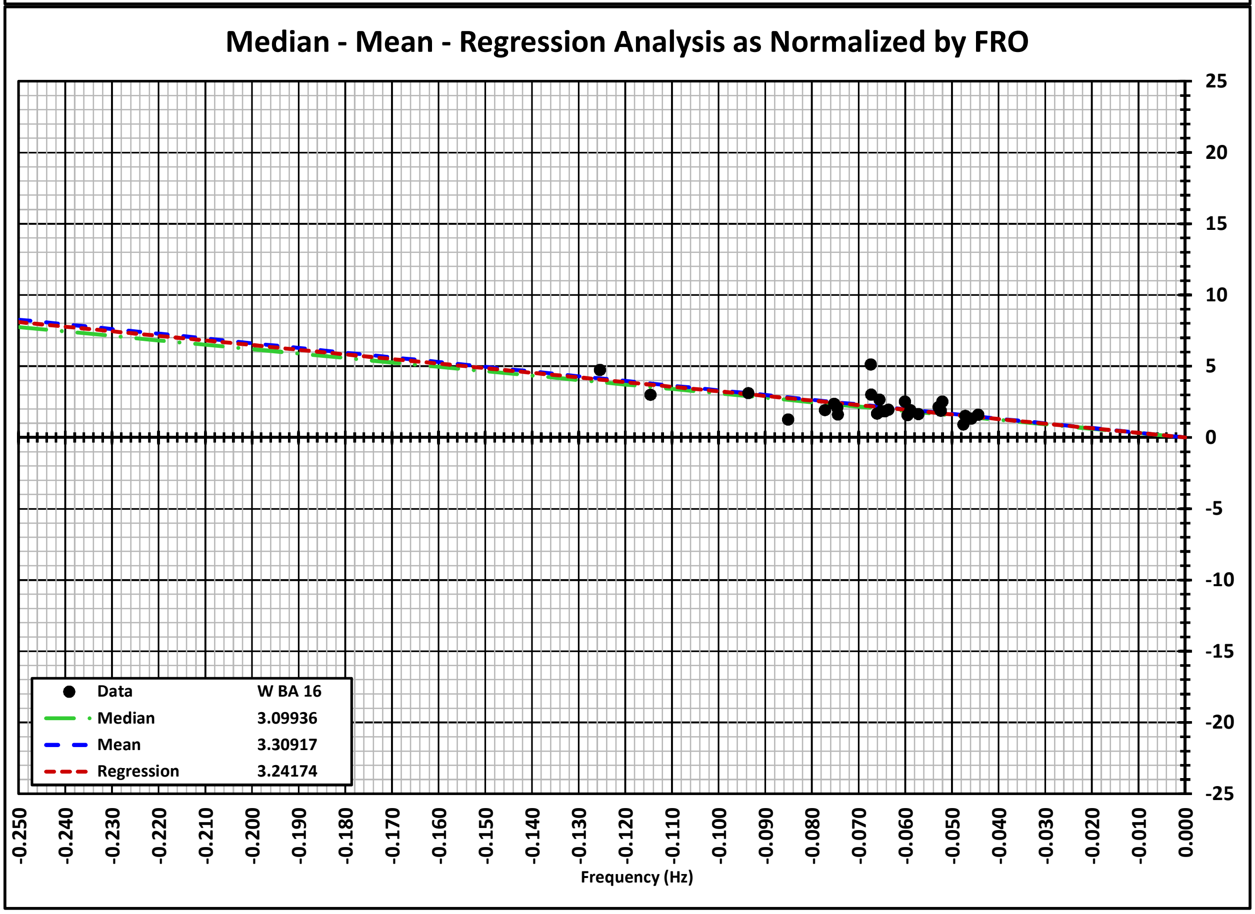

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 12
Median 2.36648
Mean 2.65442
Regression 2.61365
Frequency (Hz)

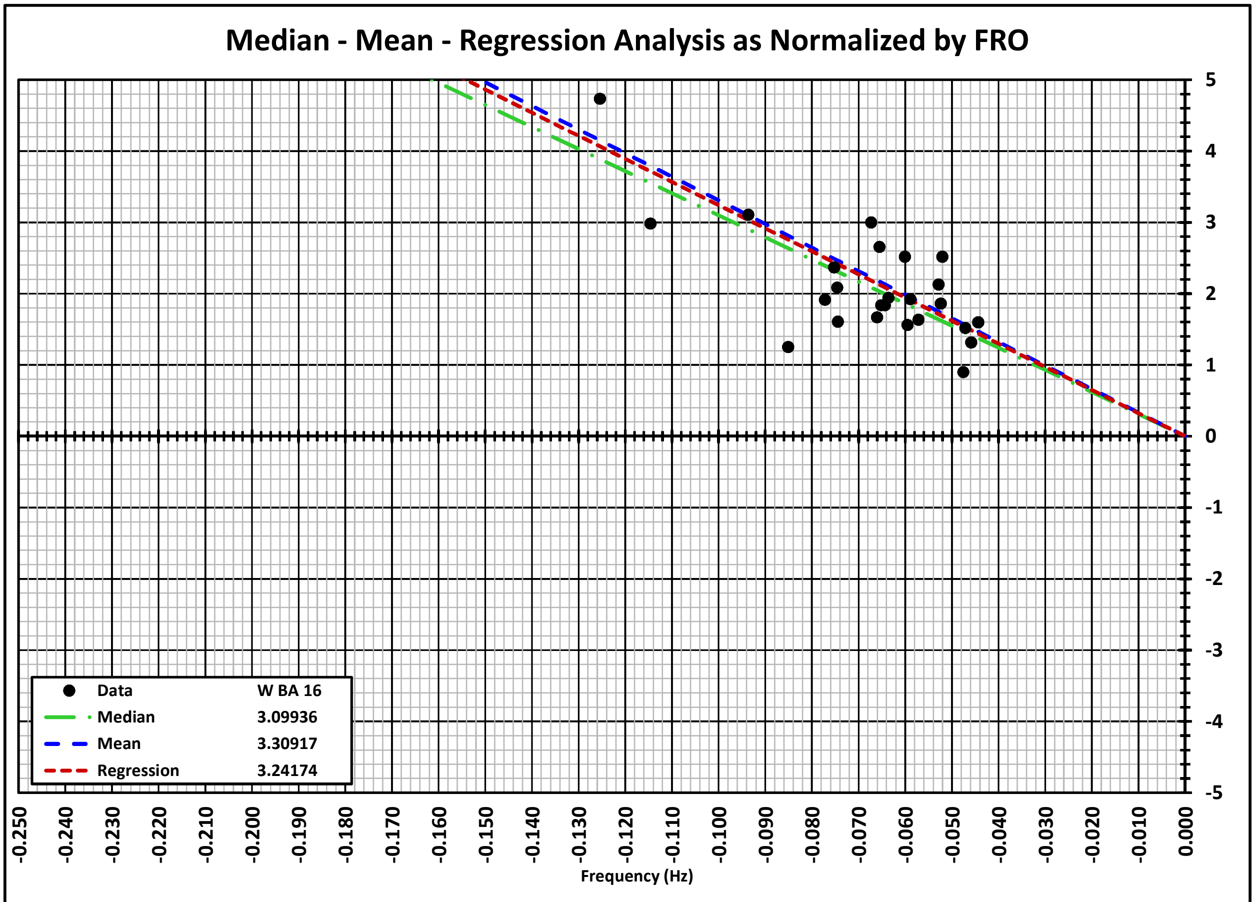

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 13
Median 4.71769
Mean 5.17291
Regression 5.14399
Frequency (Hz)

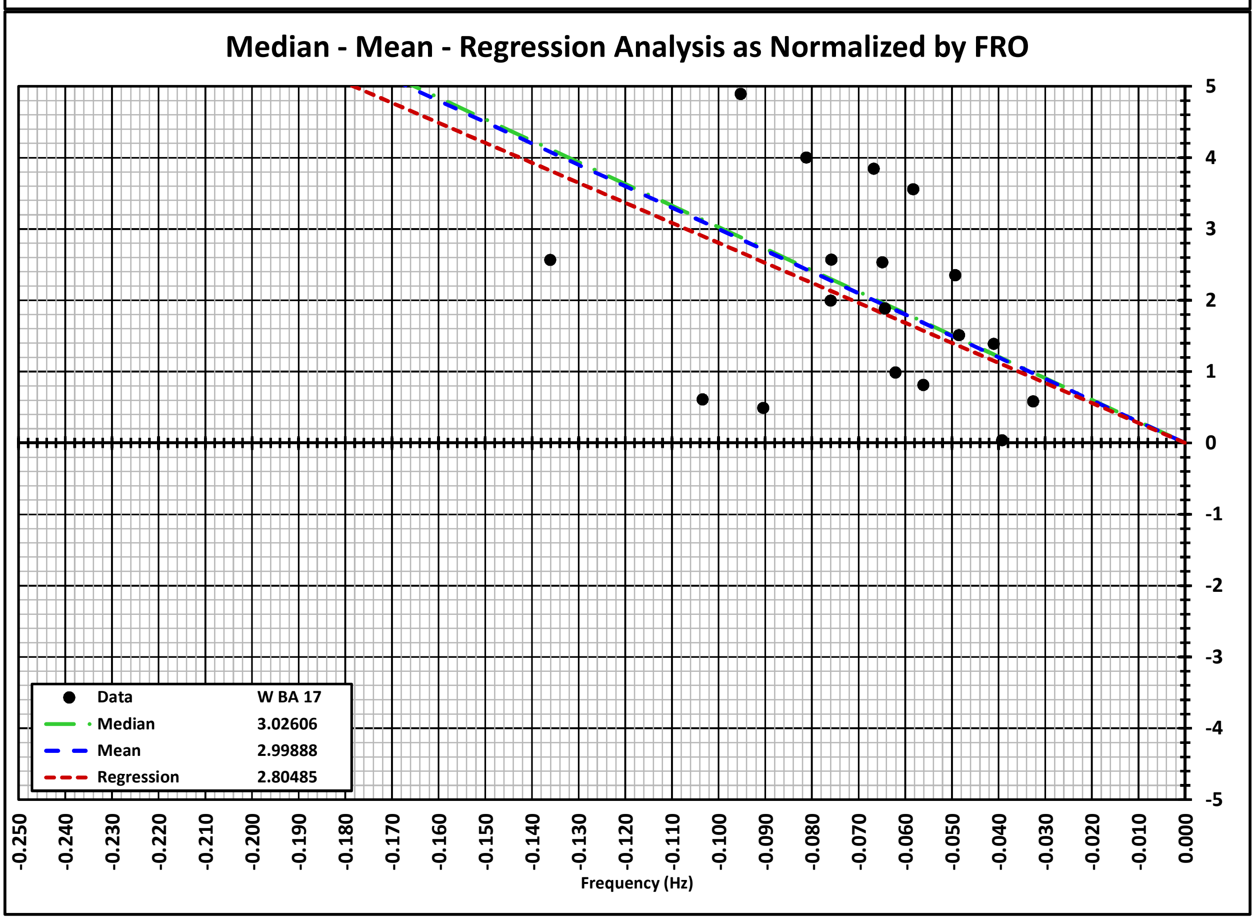

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 13
Median 4.71769
Mean 5.17291
Regression 5.14399
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.250
-0.240
-0.230
-0.220
-0.210
-0.200
-0.190
-0.180
-0.170
-0.160
-0.150
-0.140
-0.130
-0.120
-0.110
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

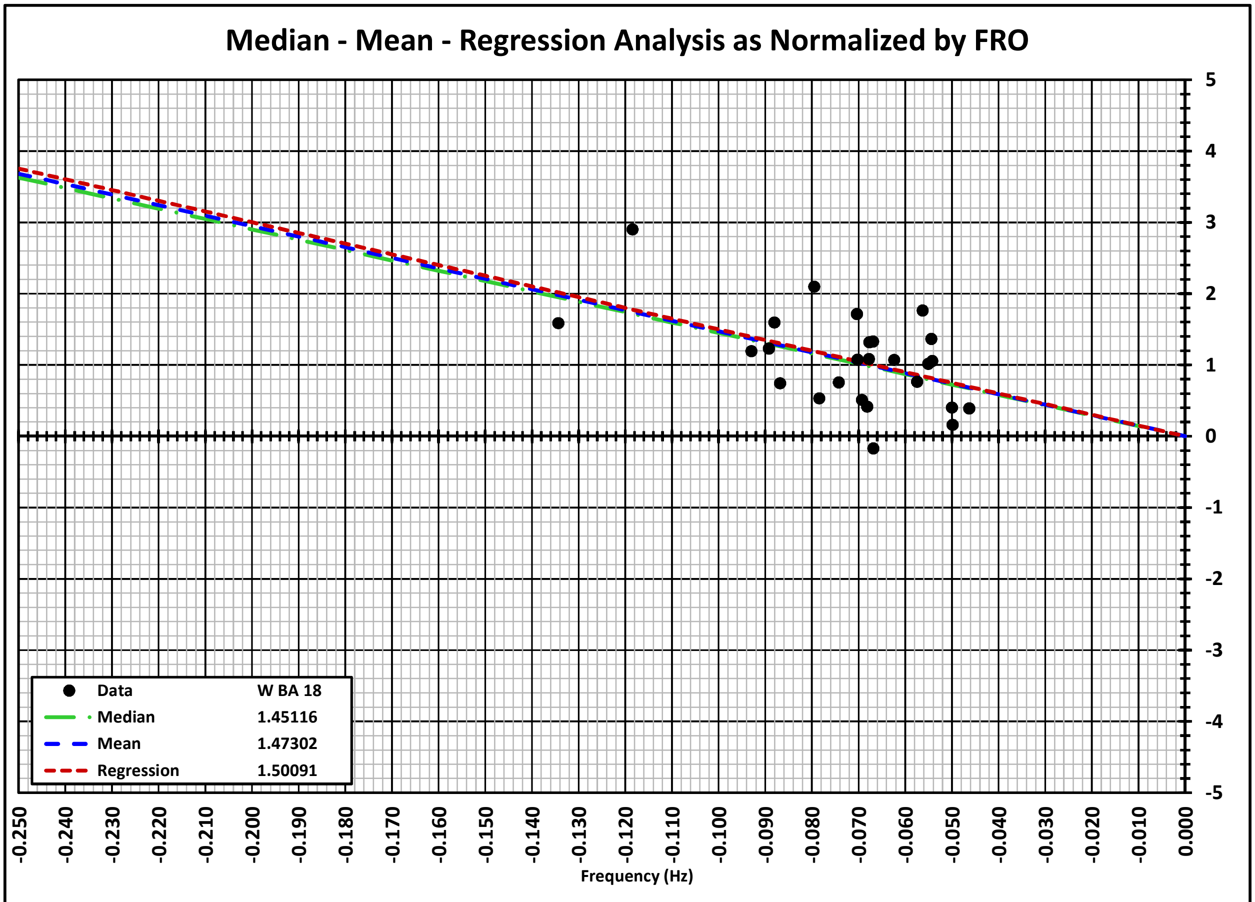

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 14
Median 0.60768
Mean 0.87898
Regression 0.63485
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.250
-0.240
-0.230
-0.220
-0.210
-0.200
-0.190
-0.180
-0.170
-0.160
-0.150
-0.140
-0.130
-0.120
-0.110
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

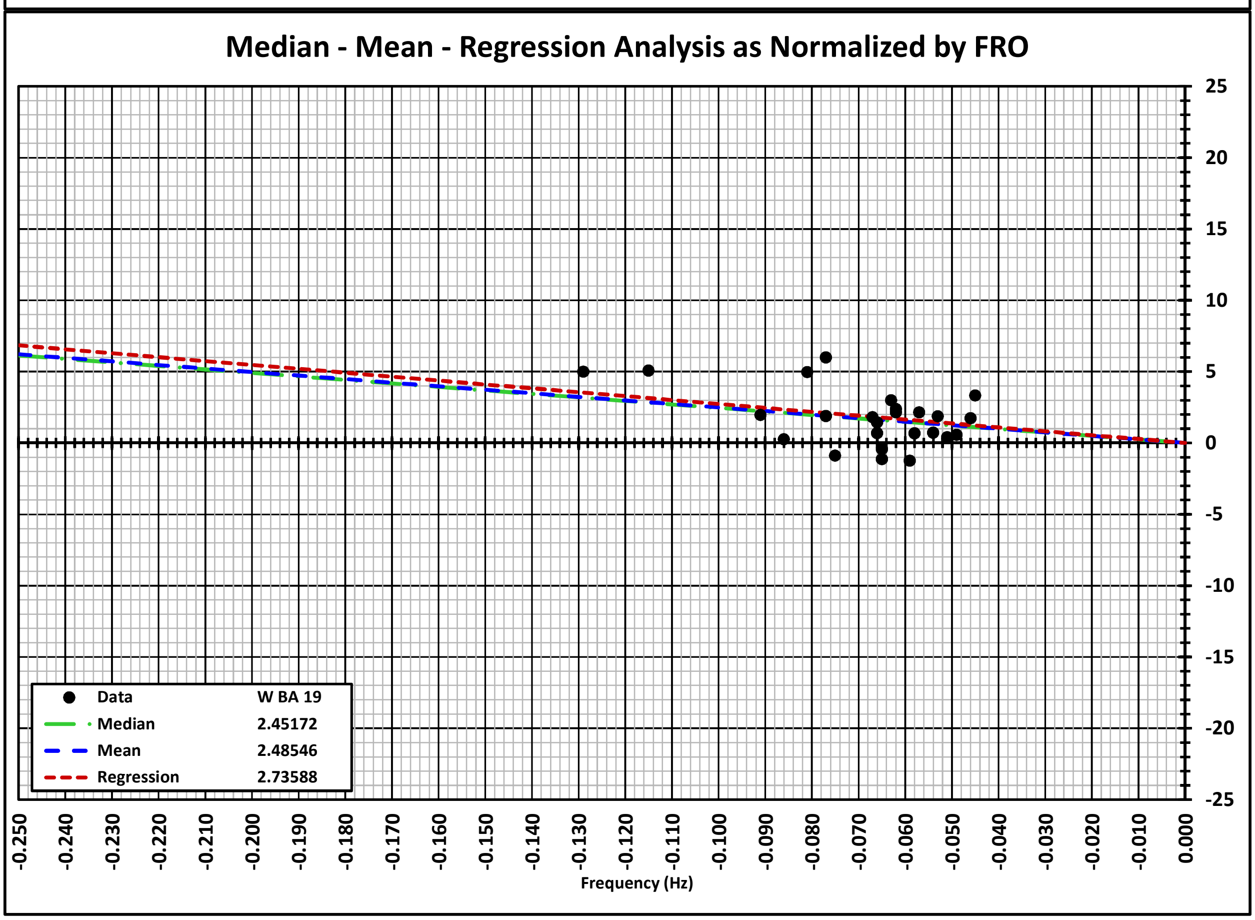

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 15
Median -1.38396
Mean -1.54605
Regression -1.39906
Frequency (Hz)
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.250
-0.240
-0.230
-0.220
-0.210
-0.200
-0.190
-0.180
-0.170
-0.160
-0.150
-0.140
-0.130
-0.120
-0.110
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000

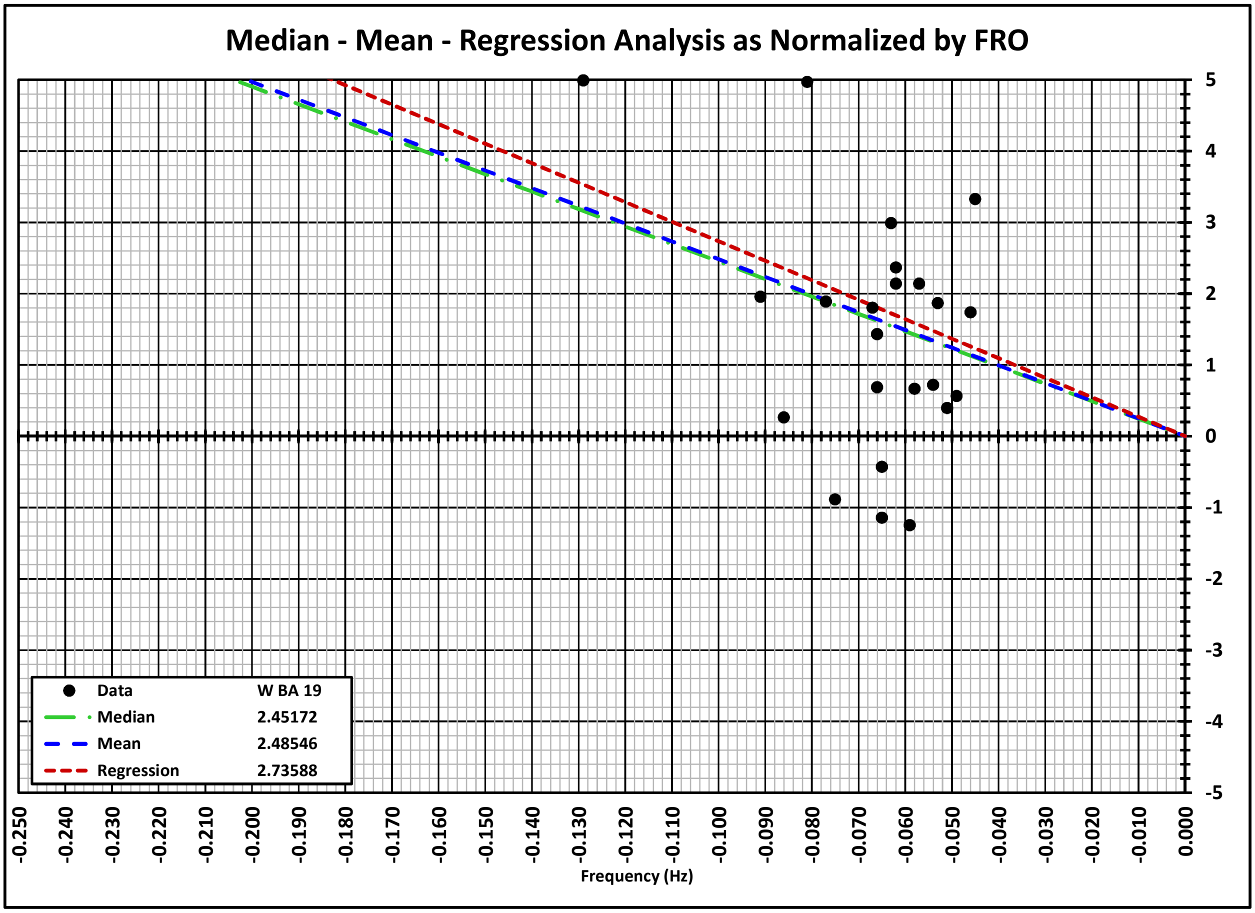

Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 16
Median 3.09936
Mean 3.30917
Regression 3.24174
Frequency (Hz)
25
20
15
10
5
0
-5
-10
-15
-20
-25
-0.250
-0.240
-0.230
-0.220
-0.210
-0.200
-0.190
-0.180
-0.170
-0.160
-0.150
-0.140
-0.130
-0.120
-0.110
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000


Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 16
Median 3.09936
Mean 3.30917
Regression 3.24174
Frequency (Hz)
-0.250 -0.240 -0.230 -0.220 -0.210 -0.200 -0.190 -0.180 -0.170 -0.160 -0.150 -0.140 -0.130 -0.120 -0.110 -0.100 -0.090 -0.080 -0.070 -0.060 -0.050 -0.040 -0.030 -0.020 -0.010 0.000
5 4 3 2 1 0 -1 -2 -3 -4 -5


Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 17
Median 3.02606
Mean 2.99888
Regression 2.80485
Frequency (Hz)
-0.250 -0.240 -0.230 -0.220 -0.210 -0.200 -0.190 -0.180 -0.170 -0.160 -0.150 -0.140 -0.130 -0.120 -0.110 -0.100 -0.090 -0.080 -0.070 -0.060 -0.050 -0.040 -0.030 -0.020 -0.010 0.000
5 4 3 2 1 0 -1 -2 -3 -4 -5


Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 18
Median 1.45116
Mean 1.47302
Regression 1.50091
Frequency (Hz)


Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 19
Median 2.45172
Mean 2.48546
Regression 2.73588
Frequency (Hz)


Median - Mean - Regression Analysis as Normalized by FRO
Data W BA 19
Median 2.45172
Mean 2.48546
Regression 2.73588
5
4
3
2
1
0
-1
-2
-3
-4
-5
-0.250
-0.240
-0.230
-0.220
-0.210
-0.200
-0.190
-0.180
-0.170
-0.160
-0.150
-0.140
-0.130
-0.120
-0.110
-0.100
-0.090
-0.080
-0.070
-0.060
-0.050
-0.040
-0.030
-0.020
-0.010
0.000
Frequency (Hz)

The Median best represents a "uniform" one-dimensional data set !

Uniform Distribution: In probability theory and statistics, the continuous uniform distribution or rectangular distribution is a family of probability distributions such that for each member of the family, all intervals of the same length on the distribution's support are equally probable. The support is defined by the two parameters, a and b, which are its minimum and maximum values.

Median: We have been taught in statistics that minimizing the sum of the differences error term provides the best estimate for the value for a uniform data set. Define a data set as one dimensional with values $\{x_1, x_2, \ldots, x_n\}$. The objective is to select a single value that best represents this data set by minimizing the sum of the residuals.

$$SDE = \sum_{i=1}^{n} \left(x_i - x_m \right)$$

Where: x_m = Best single value to represent the data set.

The result is undefined using calculus. Therefore, other logic must be used.

Organize the data in order from smallest to largest. Then investigate the change in total difference as the candidate Median value is raised from the smallest to the largest value in the data set.

When the candidate Median value is raised above the smallest data value the difference between the candidate Median value and the smallest value increases, but the difference between the candidate Median value and all other data values decreases by an amount equal to the increase in the difference for the smallest value times the number of data values above the candidate Median value. As the candidate Median value increases, the total difference from all values will decrease until exactly one half of the data values are above the candidate Median value and exactly one half of the data values are below the candidate Median value. If there are an even number of data values in the set, any change in the candidate Median value between the data value immediately below the half and the data point immediately above the half will not change the total difference because the difference change in the increasing direction and the difference change in the decreasing direction offset each other. However, if there an odd number of data values in the data set, the candidate Median value equal to the center data value will result in a minimum of the differences.

This demonstrates that the Median is the best estimate for a set of uniform data because it minimizes the sum of the error terms for the data set.

The real question is not whether the Median is an appropriate estimator, but is the Median an appropriate estimator for the data being analyzed?

The Mean best represents a "normal" one dimensional data set !

Normal (Gaussian) Distribution: In probability theory, the normal (or Gaussian) distribution is a continuous probability distribution that has a bell-shaped probability density function, known as the Gaussian function or informally the bell curve, where parameter μ is the mean or expectation (location of the peak) and σ□^{2} is the variance, the mean of the squared deviation, (a "measure" of the width of the distribution). σ is the standard deviation. The distribution with μ = 0 and σ□^{2} = 1 is called the standard normal. A normal distribution is often used as a first approximation to describe real-valued random variables that cluster around a single mean value.

The normal distribution is considered the most prominent probability distribution in statistics. There are several reasons for this:

- First, the normal distribution is very tractable analytically, that is, a large number of results involving this distribution can be derived in explicit form.
- Second, the normal distribution arises as the outcome of the central limit theorem, which states that under mild conditions the sum of a large number of random variables is distributed approximately normally.
- Finally, the "bell" shape of the normal distribution makes it a convenient choice for modeling a large variety of random variables encountered in practice.

For this reason, the normal distribution is commonly encountered in practice, and is used throughout statistics, natural sciences, and social sciences as a simple model for complex phenomena. For example, the observational error in an experiment is usually assumed to follow a normal distribution, and the propagation of uncertainty is computed using this assumption. Note that a normally-distributed variable has a symmetric distribution about its mean. Quantities that grow exponentially, such as prices, incomes or populations, are often skewed to the right, and hence may be better described by other distributions, such as the log-normal distribution or Pareto distribution. In addition, the probability of seeing a normally-distributed value that is far (i.e. more than a few standard deviations) from the mean drops off extremely rapidly. As a result, statistical inference using a normal distribution is not robust to the presence of outliers (data that is unexpectedly far from the mean, due to exceptional circumstances, observational error, etc.). When outliers are expected, data may be better described using a heavy-tailed distribution such as the Student's t-distribution.

Mean: We have been taught in statistics that minimizing the sum of the squares of the error term provides the best estimate for the value for a normal data set. Let's define a data set as one dimensional with values $\{x_1, x_2, \ldots, x_n\}$. The objective is to select a single value that best represents this data set by minimizing the sum of the squares of the residuals.

$$SSE = \sum_{i=1}^{n} \left(x_i - x_m\right)^2$$

Where: x_m = Best single value to represent the data set.

$$SSE = \sum_{i=1}^{n} \left(x_i^2 - 2x_i x_m + x_m^2\right)$$

$$SSE = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i x_m + \sum_{i=1}^{n} x_m^2$$

$$SSE = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i x_m + nx_m^2$$

Take the derivative of SSE with respect to x_m, and set that derivative equal to zero.

$$\frac{\partial}{\partial x_m} SSE = \frac{\partial}{\partial x_m}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i x_m + nx_m^2\right)$$

$$\frac{\partial}{\partial x_m} SSE = \frac{\partial}{\partial x_m}\left(\sum_{i=1}^{n} x_i^2\right) - \frac{\partial}{\partial x_m}\left(\sum_{i=1}^{n} 2x_i x_m\right) + \frac{\partial}{\partial x_m}\left(nx_m^2\right)$$

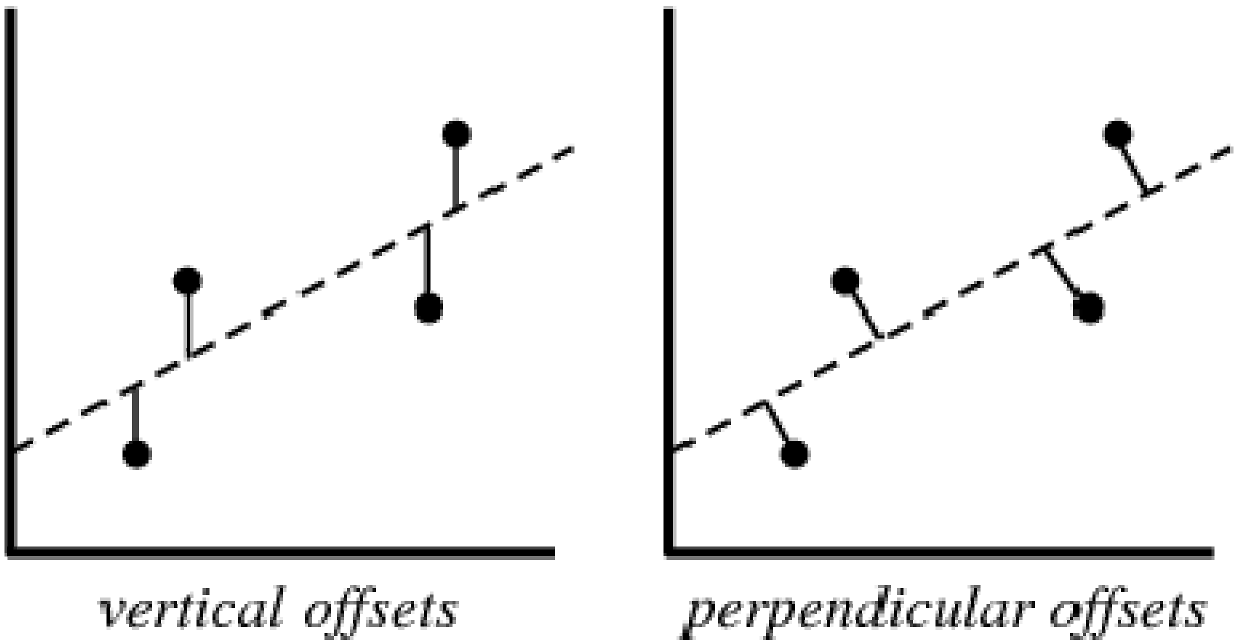
$$\frac{\partial}{\partial x_m} SSE = -2\sum_{i=1}^{n} x_i + 2nx_m = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i = x_m = \bar{x}$$

This demonstrates that the Mean is the best estimate for a set of normal data because it minimizes the sum of the squares of the error terms for the data set.

A Linear Regression best represents a "normal" two dimensional data set !

Linear Regression: As with the one dimensional data set, the objective is to minimize the sum of the squares of the error terms. However, there may be differences that depend upon how we define the error terms.



There are three alternatives available for defining the error term. It can be defined with respect to the dependent variable alone as shown in the vertical offsets plot above. The second is to define the error in terms of the horizontal offsets (not shown). That alternative is the same as the first alternative when the independent variable is exchanged with the dependent variable. The third alternative is to define the error as the perpendicular distance from the best fit line. This is shown in the perpendicular offsets plot above. When the regression is solved using the perpendicular offsets, both variables are considered equal with respect to contribution to error, and the ranking of variables is not necessary.

Solution assuming an independent / dependent variable relationship !

In the first example the error term is defined as one dimensional on the dependent variable axis. This is based on the vertical offsets shown above. The result is derived as follows:

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

Where: $\hat{y}_i$ = Best y value to represent the data set at a given x value.

Substitute a linear equation, $\hat{y}_i = ax_i + b$, for the estimated y value.

$$SSE = \sum_{i=1}^{n} \left(y_i - ax_i - b \right)^2$$

Since we now have two variables, a and b, the derivative must be taken with respect to each variable. Setting each derivative equal to zero will provide two equations that can be solved for the two unknowns, a and b.

$$\frac{\partial}{\partial b}SSE=\frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i-ax_i-b)^2=-2\sum_{i=1}^{n}(y_i-ax_i-b)=0$$

$$\frac{\partial}{\partial a}SSE=\frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i-ax_i-b)^2=-2\sum_{i=1}^{n}(x_iy_i-ax_i^2-bx_i)=0$$

Rearrange terms and solve the two equations. Solve for b first.

$$-\sum_{i=1}^{n}y_i+a\sum_{i=1}^{n}x_i+nb=0 \quad\Rightarrow\quad b=\tfrac{1}{n}\sum_{i=1}^{n}y_i-a\tfrac{1}{n}\sum_{i=1}^{n}x_i \quad\Rightarrow\quad b=\overline{y}-a\overline{x}$$

Substitute the result for b into the second equation and solve for a.

$$-\sum_{i=1}^{n}x_iy_i+a\sum_{i=1}^{n}x_i^2+(\overline{y}-a\overline{x})\sum_{i=1}^{n}x_i=0 \qquad\Rightarrow\qquad a=\frac{\sum_{i=1}^{n}x_iy_i-n\overline{yx}}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}$$

Calculate the value of a and substitute into the first equation to get the value of b. These are the most common equations used for linear regression. However, they assume that the dependent and independent variables can be identified and that the error in the dependent variable is more important than the error in the independent variable.

Solution without the independent / dependent variable relationship assumption !

In this section, the problem is solved using the perpendicular offsets to determine the error terms. This provides a solution that is not dependent upon any assumption concerning the relationship between the variables.

The first step in this solution is to determine the square of the perpendicular offset from the regression line that represents the error term.

$$SSE=\sum_{i=1}^{n}\left(\frac{[y_i-(ax_i+b)]^2}{1+a^2}\right)$$

Since we again have two variables, a and b, the derivative must be taken with respect to each variable. Setting each derivative equal to zero will provide two equations that can be solved for the two unknowns, a and b.

$$\frac{\partial}{\partial b}SSE=\frac{\partial}{\partial b}\sum_{i=1}^{n}\left(\frac{[y_i-(ax_i+b)]^2}{1+a^2}\right)=\frac{-2}{1+a^2}\sum_{i=1}^{n}(y_i-ax_i-b)=0$$

$$\frac{\partial}{\partial a}SSE=\frac{\partial}{\partial a}\sum_{i=1}^{n}\left(\frac{[y_i-(ax_i+b)]^2}{1+a^2}\right)$$

$$\frac{\partial}{\partial a}SSE = \frac{-2}{1+a^2}\sum_{i=1}^{n}(y_i - ax_i - b)x_i - \sum_{i=1}^{n}\frac{(y_i - ax_i - b)^2(2a)}{(1+a^2)^2} = 0$$

Rearrange terms and solve the two equations. Solve for b first.

$$-\sum_{i=1}^{n} y_i + a\sum_{i=1}^{n} x_i + nb = 0 \quad \Rightarrow \quad b = \tfrac{1}{n}\sum_{i=1}^{n} y_i - a\tfrac{1}{n}\sum_{i=1}^{n} x_i \quad \Rightarrow \quad b = \overline{y} - a\overline{x}$$

This is the same result as before. Substitute the result for b into the second equation and solve for a. The detailed intermediate equations for this solution can be found at http://mathworld.wolfram.com/LeastSquaresFittingPerpendicularOffsets.html. After much manipulation the following equations result.

$$A = \frac{1}{2}\frac{\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) - \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)}{n\overline{y}\overline{x} - \sum_{i=1}^{n} x_i y_i} \qquad \Rightarrow \qquad a = -A \pm \sqrt{A^2 + 1}$$

This solution is somewhat more complex than the vertical offset solution. That is the reason that the vertical offset solution is commonly used. In most cases, the vertical offset solution provides an adequate answer to the problem without the added complexity of the perpendicular offset solution. However, when the vertical offset solution is used, it makes a difference which variable is considered the independent variable and the dependent variable. This can significantly affect the results when the slope is large.

Additional information requires a special case linear regression !

The calculation of Frequency Response requires the use of a special case linear regression. Frequency Response is defined as to be equal to zero when the frequency error is equal to zero. This information requires the modification of the linear regression used to provide the best representation of the data. The appropriate linear regression for representing Frequency Response is a regression where the regression line crosses the origin of the axis representing the two variables, frequency and Frequency Response (MW). Therefore, the previously developed general solution to the problem requires modification. This is done by setting the variable that represents the ***y-intercept*** to zero. In the above examples, the b term must be set to zero.

Special case solution assuming an independent/dependent variable relationship !

In the first example the error term is defined as one dimensional on the dependent variable axis. This is based on the vertical offsets but in this case the variable representing the intercept is eliminated. The result is derived as follows:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where: $\hat{y}_i$ = Best y value to represent the data set at a given x value.

Substitute a linear equation, $\hat{y}_i = ax_i$, for the estimated y value.

$$SSE = \sum_{i=1}^{n}(y_i - ax_i)^2$$

Since we now have a single variables, a, the derivative must be taken with respect to that variable. Setting the derivative equal to zero will provide an equation that can be solved for the unknown, a.

$$\frac{\partial}{\partial a}SSE = \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - ax_i)^2 = -2\sum_{i=1}^{n}(x_i y_i - ax_i^2) = 0$$

Rearrange terms and solve the equation.

$$-\sum_{i=1}^{n} x_i y_i + a\sum_{i=1}^{n} x_i^2 = 0 \qquad \Rightarrow \qquad a = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

This equation is somewhat simpler than the equation using a non-zero intercept. In the specific case that we are considering, the estimate of Frequency Response, the slope of the regression line is not expected to be large, near vertical. Therefore, the assumption of dependent and independent variables is not important to the solution. In this case, the additional complexity added by considering the horizontal offsets is not significant to the solution and has been eliminated from consideration.