Z.0 PRECISION AND BIAS

Z.1 PAIRED SAMPLING TRAIN PRECISION

Z.1.1 What Is Precision and Why Is It Important?

Precision is a measure of the agreement between two or more independent observations of the same physical quantity obtained under the same conditions. When paired sampling trains are used, assessing the precision of the two data points collected for a given test run is important in identifying possible problems with the sampling train or sample collection method. Although the PM concentrations measured by the two sampling trains should be the same for a test run, there invariably are differences in the two measurements. Typically, the differences in emissions between the two sampling trains are insignificant. In some cases, it is unclear whether the two data points are in relative agreement or if there are significant differences, which could indicate a problem with the sampling instruments, collection methods, or analytical procedures.

Z.1.2 Does PS-11 Require Me To Evaluate the Precision of Paired Sampling Train Data?

Evaluating the precision of paired sampling train data is not required under PS-11. However, in Section 8.6(1)(i) of PS-11, it is recommended that you use paired sampling trains and screen the data for precision when conducting a correlation test or an RCA Likewise, the procedures described here for evaluating paired sampling train precision are recommendations and not requirements. Section 8.6(3) of PS-11 specifies that when conducting a correlation test, you may reject the results of up to five test runs without explanation; additional test runs can be rejected under certain conditions. The same provisions apply to RCAs. The procedures described in the following paragraphs can help in deciding which test runs, if any, you should consider rejecting.

Z.1.3 What Methods Can I Use To Determine the Precision of Paired Sampling Train Data?

This section discusses two methods for evaluating the precision of paired sampling train data. Section Z.1.4 describes a method that is based on the determination of the relative standard deviation (RSD). Section Z.1.5 presents an alternative method, which involves calculating the regression residuals for the data set and comparing them with the standardized residuals for the test run. Section Z.1.6 compares the two methods and presents recommendations on which procedure to use and how to use the results of evaluations of data precision for initial correlation tests and RCAs.

Z.1.4 How Do I Evaluate Precision Using the Relative Standard Deviation?

Determining the RSD is a relatively simple and straightforward procedure for evaluating the precision of paired sampling data. This procedure, which was included in earlier proposals of PS-11, currently appears in Section 12.2 of Method 5i.

The RSD presented in this discussion and in Method 5i is not the true RSD but is actually a measure of the standardized half range. For two data points, the standardized half range differs from the true RSD by a factor of $(1/\sqrt{2})$ or 0.707; that is, the RSD (as presented here) equals the true RSD × 0.707.

Evaluating the precision of paired sampling train data using the RSD entails the following steps:

1. Calculate the RSD for each test run;

2. Determine the average PM concentration for each run;

3. Determine the recommended maximum RSD for each run, based on the average PM concentration for the run; and

4. Compare the calculated RSD with the recommended maximum RSD for each run and flag those test runs for which the calculated RSD exceeds the recommended maximum RSD value.

The following paragraphs describe each of these steps in more detail. Example calculations are also presented.

Step 1: Calculate RSD

The RSD is calculated for each test run using Equation Z.1-1:

$$RSD = 100\% \times \frac{\left|\left(C_a - C_b\right)\right|}{\left(C_a + C_b\right)} \qquad \text{(Equation Z.1-1)}$$

where

RSD	=	relative standard deviation, percentage
C_a	=	concentration measured using Train A, units of concentration
C_b	=	concentration measured using Train B, same units as C_a
$\lvert C_a\text{-}C_b \rvert$	=	absolute value of the difference between C_a and C_b, same units of concentration.

The concentration units for the variables C_a and C_b do not matter, provided the same units are used for both concentrations (e.g., both concentrations could be in units of mg/dscm or both could be in units of mg/acm).

In the following example calculation, assume the concentration for Train A was 9.2 mg/acm, and the concentration for Train B during the same test run was 7.6 mg/acm.

$$RSD = 100\% \times \frac{|(9.2 - 7.6)|}{(9.2 + 7.6)} = \frac{1.6}{16.8} \times 100 = 9.5\%$$

Step 2: Determine Average PM Concentration

The recommended maximum RSD (RSD_r) is based on the average PM concentration (C_{ave}) for the test run. For the previous example, the average PM concentration for the run is

$$C_{ave} = \frac{C_a + C_b}{2} = \frac{9.2 + 7.6}{2} = 8.4$$

Step 3: Determine Recommended Maximum RSD

The recommended maximum RSD is a function of the average PM concentration for the two data points, as follows:

- For average PM concentrations that are at least 10 mg/dscm (or at least 10 mg/acm), the recommended maximum RSD is 10 percent;
- For average PM concentrations that are less than or equal to 1.0 mg/dscm (or less than or equal to 1.0 mg/acm), the recommended maximum RSD is 25 percent; and
- For average PM concentrations between 1.0 and 10 mg/dscm (or between 1.0 and 10 mg/acm), the recommended maximum RSD is determined using Equation Z.1-2, which is simply a linear interpolation between 10 and 25 percent:

$$RSD_r = (26.67 - 1.67C_{ave}) \qquad \text{(Equation Z.1-2)}$$

where

RSD_r = recommended maximum RSD for average PM concentrations between 1.0 and 10, percent
C_{ave} = average PM concentration for the two sampling trains, units of concentration.

For the above example, the average PM concentration is greater than 1.0 mg/dscm and less than 10 mg/dscm; therefore, Equation Z.1-2 is used to determine the recommended maximum RSD as follows:

$$RSD_r = 26.67 - 1.67 \times C_{ave} = 26.67 - 1.67 \times 8.4 = 12.6\%$$

Step 4: Compare Calculated RSD to Recommended Maximum RSD

The calculated RSD is 9.5 percent, which is less than the recommended maximum RSD of 12.6 percent. Therefore, the test run satisfies the criterion for RSD.

Example Problem: How Do I Evaluate the Precision of Paired Sampling Train Data Using the RSD?

For the initial correlation test of a PM CEMS, 20 test runs are conducted using paired sampling trains. The results of the test are summarized in Table Z.1-1. The table also shows for each test run the average PM concentration, the RSD calculated using Equation Z.1-1, and the recommended maximum RSD. Finally, the table indicates whether each test run met the RSD criterion.

Table Z.1-1. Example RSD Calculations – Results of Initial Correlation Test

Run	PM Concentration, mg/dscm			Calculated RSD, %	Recommended maximum RSD, %	Does the test run satisfy the RSD criterion?
	Train A	Train B	Average			
1	2.6	4.2	3.4	23.5	21	No
2	3.9	4.4	4.2	6	19.7	Yes
3	4.2	2.8	3.5	20	20.8	Yes
4	8.6	10.3	9.5	9	10.9	Yes
5	15.1	14.5	14.8	2	10	Yes
6	18.3	19.7	19.0	3.7	10	Yes
7	22.4	21.3	21.9	2.5	10	Yes
8	24.5	20.2	22.4	9.6	10	Yes
9	24.3	36.9	30.6	20.6	10	No
10	32.9	32.2	32.6	1.1	10	Yes
11	6.4	7.4	6.9	7.2	15.1	Yes
12	9.5	14.2	11.9	19.8	10	No
13	12.4	15.1	13.8	9.8	10	Yes
14	14.2	11.2	12.7	11.8	10	No
15	7.8	7.1	7.5	4.7	14.2	Yes
16	21.3	18.3	19.8	7.6	10	Yes
17	4.6	13.5	9.1	49.2	11.6	No
18	28.7	24.2	26.5	8.5	10	Yes
19	12.6	14.8	13.7	8	10	Yes
20	10.5	11.3	10.9	3.7	10	Yes

As shown in the table, Runs 1, 9, 12, 14, and 17 did not meet the precision criterion for RSD. For Run 1, the average PM concentration was 3.4 percent, and the RSD was 23.5 percent. Since the average PM concentration is between 1.0 and 10 mg/dscm, Equation Z.1-2 is used to determine the recommended maximum RSD for this run. Using that equation, the recommended maximum RSD for Run 1 is calculated to be 21 percent (26.67 - 1.67 × 3.4 = 21.0). For Run 9, the RSD (20.6 percent) exceeded the recommended maximum RSD of 10 percent. For Runs 12 and 14, the RSDs were 19.8 percent and 11.8 percent, respectively, both of which exceed the recommended maximum of 10 percent. Finally, the RSD for Run 17 was 49.2 percent, which exceeds the recommended maximum RSD of 11.6 percent for that run, as determined by Equation Z.1-2.

Section 8.5(3) of PS-11 requires a minimum of 15 test runs for the initial correlation test, and paragraph 8.5(3)(iii) allows up to 5 test runs to be discarded without explanation. In the above example, the facility might consider discarding the 5 test runs for which the RSD exceeded the recommended maximum RSD; if it did so, there still would be a sufficient number of test runs to develop the initial correlation. However, the facility would not be required to reject any of the data; it could decide to keep the results of any or all of the test runs that did not meet the RSD criterion.

Z.1.5 How Do I Evaluate the Precision of Paired Sampling Train Data Using Residuals?

A residual is defined as the difference between an observed value and the fitted value that is determined using a regression equation. Using residuals to evaluate the precision of paired sampling train data is an iterative process, which consists of the following steps:

1. Determine the linear regression equation for the data;

2. Calculate the residual and standardized residual for each test run;

3. Eliminate any run for which the absolute value of the standardized residual exceeds 3.0;

4. Recalculate the residuals and standardized residuals for each test run; and

5. Repeat Steps 3 and 4 until the absolute value of the standardized residual for each run is less than or equal to 3.0.

The following paragraphs describe these steps in more detail. An example problem to illustrate all of the calculations for evaluating paired sampling data precision using residuals also is presented.

Step 1: Determine Regression Equation

Before the residuals can be calculated, the regression equation must be determined. For paired sampling train data, a linear regression model is used. The procedure for determining the linear regression equation is identical to that specified in Section 12.3(1) of PS-11 for developing a linear correlation model. The data from either sampling train can be used for the dependent

variable (y) or independent variable (x). For this discussion, the data from Train B represent the dependent variable, and the data from Train A represent the independent variable. The regression equation is of the form presented in Equation Z.1-3:

$$\hat{y} = b_0 + b_1 x \qquad \text{(Equation Z.1-3)}$$

where

$\hat{y}$ = predicted (fitted) PM concentration for Train B
b_0 = intercept of the regression line, as calculated using Equation Z.1-4
b_1 = slope of the regression line, as calculated using Equation Z.1-6
x = PM concentration measured by Train A.

The y intercept (b_0) of the regression line is determined using Equation Z.1-4:

$$b_0 = \bar{y} - b_1 \cdot \bar{x} \qquad \text{(Equation Z.1-4)}$$

where

$\bar{x}$ = mean of the Train A data, as calculated using Equation Z.1-5
$\bar{y}$ = mean of the Train B data, as calculated using Equation Z.1-5.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{(Equation Z.1-5)}$$

where

x_i = PM concentration measured using Train A for run i
y_i = PM concentration measured using Train B for run i
n = number of data pairs.

The slope (b_1) of the regression line is determined using Equation Z.1-6:

$$b_1 = \frac{S_{xy}}{S_{xx}} \qquad \text{(Equation Z.1-6)}$$

where

S_{xx}, S_{xy} = as calculated using Equation Z.1-7:

$$S_{xx} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2, \quad S_{xy} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) \qquad \text{(Equation Z.1-7)}$$

where

x_i, $\bar{x}$, y_i, and $\bar{y}$ = as defined above for Equations Z.1-4 and Z.1-5.

For the data listed in Table Z.1-1, the calculated values are as follows:

$$\bar{x} = \frac{1}{20}\sum_{i=1}^{20}(2.6 + 3.9 + \ldots + 10.5) = 14.24$$

$$\bar{y} = \frac{1}{20}\sum_{i=1}^{20}(4.2 + 4.4 + \ldots + 11.3) = 15.18$$

$$S_{xx} = \sum_{i=1}^{20}\left((2.6 - 14.24)^2 + (3.9 - 14.24)^2 + \ldots + (10.5 - 14.24)^2\right) = 1{,}511$$

$$S_{xy} = \sum_{i=1}^{20}\left((2.6\text{-}14.24)(4.2\text{-}15.18) + \ldots + (10.5\text{-}14.24)(11.3\text{-}15.18)\right) = 1{,}348$$

$$b_1 = \frac{1{,}348}{1{,}511} = 0.892$$

$$b_0 = 15.18 - (0.892 \times 14.24) = 2.47$$

The regression equation is then expressed as

$$\hat{y} = 2.47 + 0.892x$$

Step 2: Calculate Residuals and Standardized Residuals

For a paired sampling train test run, the residual is simply the difference between the concentration (y) measured by one of the trains (e.g., Train B) and the concentration determined using the regression equation ($\hat{y}$), as shown in Equation Z.1-8:

$$e = y - \hat{y} \qquad \text{(Equation Z.1-8)}$$

where

- e = residual
- y = concentration measured by Sampling Train B
- $\hat{y}$ = predicted (fitted) concentration for Sampling Train B calculated using the regression equation.

The standardized residual for a test run is a function of the residual for the run and the residual variance, as calculated using Equation Z.1-9:

$$e_S = \frac{e_i}{\sqrt{Var(e)}} \qquad \text{(Equation Z.1-9)}$$

where

e_s = standardized residual
e_i = residual for Run i
Var(e) = residual variance, as determined using Equation Z.1-10.

$$Var(e) = \frac{\sum_{i=1}^{n} e_i^{\,2}}{(n-1)} \qquad \text{(Equation Z.1-10)}$$

where

Var(e), e_i, and n = as defined previously.

Based on the data presented in Table Z.1-1, the fitted y values for Runs 1 and 2 are

$$Run\,1\text{:}\quad \hat{y} = 2.47 + (0.892 \times 2.6) = 4.79$$
$$Run\,2\text{:}\quad \hat{y} = 2.47 + (0.892 \times 3.9) = 5.95$$

From Equation Z.1-8, the residuals for Runs 1 and 2 are

$$Run\,1\text{:}\quad e_1 = 4.2 - 4.79 = -0.59$$
$$Run\,2\text{:}\quad e_2 = 4.4 - 5.95 = -1.55$$

From Equation Z.1-10, the residual variance for the data in Table Z.1-1 is

$$Var(e) = \frac{\sum_{1=1}^{20} \left(-0.59^2 + (-1.55^2) + \ldots + (-0.54^2\right)}{(20-1)} = 16.19$$

From Equation Z.1-9, the standardized residual for Run 1 is

$$e_s = \frac{-0.59}{\sqrt{16.19}} = -0.147$$

For this run, the absolute value of the standardized residual is 0.147, which is less than 3.0. Therefore, Run 1 meets the criterion for precision based on the residuals.

Example Problem: How Do I Evaluate the Precision of Paired Sampling Train Data Using Residuals?

This example uses the same initial correlation test data presented in Table Z.1-1. A total of 20 test runs were conducted using paired sampling trains. The first step is to perform a linear regression using the data from sampling Train A as the independent variable (x values) and the data for Train B as the dependent variable (y values). Table Z.1-2 summarizes the results of the regression analysis. The resulting regression equation is $\hat{y} = 2.473 + 0.892\ y$.

Table Z.1-2. Summary of Regression Calculations

Variable	Calculated value	Equation no.
$\bar{x}$	14.24	Z.1-5
$\bar{y}$	15.18	Z.1-5
n	20	Z.1-5
S_{xx}	1.511	Z.1-7
S_{xy}	1348	Z.1-7
b_0	2.473	Z.1-4
b_1	0.892	Z.1-6

The residuals are calculated using Equation Z.1-8, and the standardized residuals are calculated using Equations Z.1-9 and Z.1-10. Table Z.1-3 shows the predicted (fitted) y value, the residual, and the standardized residual for each test run. As shown in the table, all runs meet the precision criterion except Run 9. For that run, the absolute value of the standardized residual is 3.167, which exceeds 3.0.

Table Z.1-3. Summary of Residuals for Example Problem – First Iteration

Run	PM concentration, mg/dscm		Predicted PM concentration($\hat{y}$)	Residual (e_i)	Standardized residual (e_s)
	Train A	Train B			
1	2.6	4.2	4.79	-0.593	-0.147
2	3.9	4.4	5.95	-1.553	-0.386
3	4.2	2.8	6.22	-3.421	-0.850
4	8.6	10.3	10.15	0.153	0.038
5	15.1	14.5	15.95	-1.447	-0.360
6	18.3	19.7	18.80	0.897	0.223
7	22.4	21.3	22.46	-1.161	-0.289
8	24.5	20.2	24.34	-4.135	-1.028

(continued)

Table Z.1-3. (continued)

Run	PM concentration, mg/dscm		Predicted PM concentration($\hat{y}$)	Residual (e_i)	Standardized residual (e_s)
	Train A	Train B			
9	24.3	36.9	24.16	12.743	3.167
10	32.9	32.2	31.83	0.369	0.092
11	6.4	7.4	8.18	-0.784	-0.195
12	9.5	14.2	10.95	3.250	0.808
13	12.4	15.1	13.54	1.562	0.388
14	14.2	11.2	15.14	-3.944	-0.980
15	7.8	7.1	9.43	-2.333	-0.580
16	21.3	18.3	21.48	-3.180	-0.790
17	4.6	13.5	6.58	6.922	1.720
18	28.7	24.2	28.08	-3.883	-0.965
19	12.6	14.8	13.72	1.083	0.269
20	10.5	11.3	11.84	-0.543	-0.135

The next step is to remove Run 9 from the data set and recalculate the regression equation and standardized residuals to determine if any of the data points exceed 3.0 after the outlier (Run 9) has been removed. Table Z.1-4 summarizes the regression parameters.

Table Z.1-4. Summary of Regression Calculations for Example Problem with Run 9 Removed

Variable	Calculated value
$\bar{x}$	13.71
$\bar{y}$	14.04
n	19
S_{xx}	1404
S_{xy}	1118
b_0	3.12
b_1	0.796

The residuals and standard residuals are recalculated. The revised calculations, summarized in Table Z.1-5, show that the absolute value of the standardized residual is less than 3.0 for all runs. Therefore, the evaluation is complete with only one outlier (Run 9) identified. The facility can decide to reject that run, but it is not required to do so.

Table Z.1-5. Revised Calculation of Residuals after Removal of Run 9

Run	PM concentration, mg/dscm		Predicted PM concentration($\hat{y}$)	Residual (e_i)	Standardized residual (e_s)
	Train A	Train B			
1	2.6	4.2	5.19	-0.99	-0.378
2	3.9	4.4	6.23	-1.825	-0.696
3	4.2	2.8	6.46	-3.664	-1.398
4	8.6	10.3	9.97	0.332	0.127
5	15.1	14.5	15.14	-0.643	-0.245
6	18.3	19.7	17.69	2.009	0.766
7	22.4	21.3	20.96	0.344	0.131
8	24.5	20.2	22.63	-2.428	-0.926
10	32.9	32.2	29.32	2.884	1.1
11	6.4	7.4	8.22	-0.816	-0.311
12	9.5	14.2	10.68	3.516	1.341
13	12.4	15.1	12.99	2.107	0.804
14	14.2	11.2	14.43	-3.227	-1.231
15	7.8	7.1	9.33	-2.231	-0.851
16	21.3	18.3	20.08	-1.780	-0.679
17	4.6	13.5	6.78	6.717	2.563
18	28.7	24.2	25.97	-1.772	-0.676
19	12.6	14.8	13.15	1.647	0.629
20	10.5	11.3	11.48	-0.180	-0.069

Z.1.6 Which Procedure Should I Use for Evaluating Paired Sampling Train Data Precision?

The methods presented in Sections Z.1.3 and Z.1.4 for evaluating precision differ in terms of simplicity and results. The method based on residuals analysis is a relatively complex, iterative procedure. The RSD method is much simpler and requires a single round of calculations. However, as indicated by the results of the example problem, the RSD method is more stringent in terms of identifying outliers. In the example problem, the results of five test runs were identified as possible outliers based on RSD, but the residuals analysis of the same data set identified only one test run as a potential outlier. Although either method can be used to evaluate paired sampling train precision, a recommended procedure is to use the RSD method initially to evaluate data precision. If the RSD analysis shows that the data for all test runs are within the recommended maximum RSD, no further assessment of data precision should be necessary because the RSD method is the more stringent of the two methods presented here. If the RSD values for one or more test runs indicates that the data set contains outliers, an analysis of the residuals is recommended.

Regardless of which method is used, any potential outliers should be investigated to determine possible causes for the lack of precision. If problems are identified in how the samples were collected or analyzed, the suspect data pairs should be discarded. If no problems can be identified, it may be that the apparent lack of precision is simply evidence of the natural variations in the data. In any case, the facility has the option of discarding the results of up to five test runs without explanation, provided there remains a minimum number of test runs to satisfy PS-11 (i.e., 15 runs for developing a correlation or 12 runs for an RCA). The results of additional test runs can be rejected if the basis for rejecting the runs is specified in PS-11, the test method, or the quality assurance plan for the facility. As stated above, a minimum of 15 test runs is needed to develop a correlation, and a minimum of 12 test runs is required for an RCA.

Z.2 PAIRED SAMPLING TRAIN BIAS

Z.2.1 What Is Bias and Why Is It Important?

Bias is a systematic error in data. Bias can arise from differences in parameter measurement methods, data collection methods, or data reduction methods. In paired sampling trains, bias can occur when the instruments used to measure emissions differ from one train to the other, or when the test crew operates the two sampling trains differently. For example, if two sampling trains include heated filters, and the filter temperatures are significantly different, more PM can condense on the filter with the lower temperature, resulting in biased data. Because it may not be apparent that there are differences between the sampling trains or data collection methods, it is important to evaluate the data for bias. When developing the correlation equation for PS-11, bias can introduce more scatter in the data, skew the correlation equation, or result in failure to meet the correlation acceptance criteria specified in Section 13.2 of PS-11.

Z.2.2 How Do I Determine Bias in Paired Sampling Train Data?

There are several methods that can be used to identify bias in paired sampling train data. The method described in the following paragraphs consists of three steps:

1. Determine the linear regression equation for the data;
2. Calculate the confidence interval for the slope of the regression line; and
3. Calculate the confidence interval for the intercept of the regression line.

The following paragraphs describe these steps in more detail.

Step 1: Determine Regression Equation

The procedure for determining the linear regression equation is the same as is described in Section Z.1.2.1 for evaluating residuals for precision. The data from either of the paired sampling trains can be used for the dependent variable or the independent variable, and the resulting linear regression equation has the form indicated by Equation Z.1-3.

$$\hat{y} = b_0 + b_1 x$$

Step 2: Determine Confidence Interval for Slope

The next step is to determine the 95 percent confidence interval for the slope of the regression line (b_1). The criterion for the slope of the regression line is that the value 1.0 falls within the confidence interval. The 95 percent confidence interval for the slope is calculated using Equation Z.2-1.

$$CI_{b1} = b_1 \pm t_{95\%} \times SE(b_1) \qquad \text{(Equation Z.2-1)}$$

where

CI_{b1} = 95 percent confidence interval for the slope
b_1 = slope of the regression line
$t_{95\%}$ = t(n-2,0.975) is the 100(1-0.005/2) percentage point of a Student's *t*-distribution with n-2 degrees of freedom (see Appendix A)
$SE(b_1)$ = standard error of the slope.

The standard error of the slope is calculated using Equation Z.2-2:

$$SE(b_1) = \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \text{(Equation Z.2-2)}$$

where

x_i and $\bar{x}$ = as defined above for Equations Z.1-4 and Z.1-5
S = root mean square error.

The root mean square error (S) is calculated using Equation Z.2-3:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)}} \qquad \text{(Equation Z.2-3)}$$

where

y_i, $\hat{y}_i$, and n = as defined above for Equations Z.1-3 through Z.1-5.

For the data presented in Table Z.1-1, the regression equation was previously determined to be $\hat{y} = 2.47 + 0.892x$. Using that equation and the Train B data from Table Z.1-1, the root mean square error is calculated using Equation Z.2-3 as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{20} (4.2 - 4.79)^2 + (4.4 - 5.95)^2 + \ldots + (11.3 - 11.84)^2}{(20 - 2)}} = 4.13$$

The mean x value was previously determined to be 14.24. When the value for S and the data for Train A from Table Z.1-1 are inserted in Equation Z.2-2, the standard error of the slope is calculated as follows:

$$SE(b_1) = \frac{4.13}{\sqrt{\sum_{i=1}^{20} \left((2.6 - 14.24)^2 + (3.9 - 14.24)^2 + \ldots + (10.5 - 14.24)^2\right.}} = 0.106$$

For 18 degrees of freedom (n - 2 = 20 - 2 = 18), the value for the *t* statistic (refer to Appendix A) is 2.101. Inserting that value and the results of the above calculations for root mean square error and the standard error of the slope into Equation Z.2-1, the confidence interval for the slope is as follows:

$$CI_{b1} = 0.892 \pm (2.101 \times 0.106) = 0.892 \pm 0.223$$

This can also be written as

$$0.669 \leq b_1 \leq 1.116 \text{ with 95\% confidence}$$

Since this interval includes the value 1.0, the data satisfy the bias criterion for slope.

Step 3: Determine Confidence Interval for Intercept

The final step is to determine the 95 percent confidence interval for the intercept of the regression line (b_0). If the value 0.0 falls within the confidence interval for the intercept, the data satisfy the bias criterion and are acceptable. The 95 percent confidence interval for the intercept is calculated using Equation Z.2-4:

$$CI_{b0} = b_0 \pm t_{95\%} \times SE(b_0) \qquad \text{(Equation Z.2-4)}$$

where

CI_{b0} = 95 percent confidence interval for the intercept
b_0 = intercept of the regression line
$t_{95\%}$ = as defined above for Equation Z.2-1
$SE(b_0)$ = standard error of the intercept.

The standard error of the intercept is calculated using Equation Z.2-5:

$$SE(b_0) = S\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})}} \qquad \text{(Equation Z.2-5)}$$

where all variables are as defined above in Equations Z.1-3 to Z.1-5 and Z.2-4.

For the data presented in Table Z.1-1, the parameter $SE(b_0)$, which is determined from Equation Z.2-5, is as follows:

$$SE(b_0) = 4.13\sqrt{\frac{\sum\left(2.6^2 + 3.9^2 + \ldots + 10.5^2\right)}{20\sum\left((2.6 - 14.24) + (3.9 - 14.24) + \ldots + (10.5 - 14.24)\right)}} = 1.774$$

Substituting into Equation Z.2-4, the confidence interval for the intercept is as follows:

$$CI_{b0} = 2.47 \pm (2.10 \times 1.77) = 2.47 \pm 3.73$$

This can also be written as

$$-1.25 \leq b_0 \leq 6.20 \text{ with 95 percent confidence}$$

Since this interval includes the value 0.0, the data satisfy the bias criterion for intercept.

Example Problem A: How Do I Evaluate Paired Sampling Train Data for Bias?

In this example, 15 test runs were performed using paired sampling trains. Table Z.2-1 summarizes the data for the two sampling trains.

Table Z.2-1. Example A – Summary of Test Results

Run	PM Concentration	
	Train A (x)	Train B (y)
1	2.3	2.5
2	3.9	4.4
3	4.2	4.1
4	8.6	10.3
5	14.2	13.1
6	15.1	14.5
7	18.3	19
8	22.4	21.3
9	24.5	22.8
10	28.1	28.2
11	32.9	32.2
12	6.4	6.9
13	7.2	6.9
14	12.6	12.4
15	10.5	11.3

Table Z.2-2 summarizes the results of the regression calculations.

Table Z.2-2. Summary of Regression Calculations

Variable	Calculated value	Equation No.
$\bar{x}$	14.08	Z.1-5
$\bar{y}$	13.99	Z.1-5
n	15	Z.1-5
S_{xx}	1239	Z.1-7
S_{xy}	1182	Z.1-7
b_0	0.556	Z.1-4
b_1	0.954	Z.1-6
S	0.793	Z.2-3
$t_{95\%}$	2.16	Appendix A
$SE(b_1)$	0.0225	Z.2-2
$SE(b_0)$	0.378	Z.2-5
CI_{b1}	$0.906 < b_1 < 1.003$	Z.2-1
CI_{b0}	$-0.260 < b_0 < 1.372$	Z.2-4

The resulting regression equation is

$$\hat{y} = 0.556 + 0.954x$$

The confidence interval for the slope is determined from Equation Z.2-1 to be $0.906 < b_1 < 1.003$. Because the confidence interval includes the value 1.0, the data are considered to be unbiased with respect to regression line slope. The confidence interval for the intercept is determined from Equation Z.2-4 to be $-0.260 < b_0 < 1.372$. Because the confidence interval includes the value 0.0, the data are considered to be unbiased with respect to regression line intercept. Therefore, the data are considered to be unbiased.

Example Problem B: How Do I Evaluate Paired Sampling Train Data for Bias?

In this example, 18 test runs were performed using paired sampling trains. Table Z.2-3 summarizes the data for the two sampling trains.

Table Z.2-3. Example B – Summary of Test Results

Run	PM Concentration	
	Train A (x)	Train B (y)
1	2.3	2.5
2	3.9	4.4
3	4.2	4.8
4	8.6	10.3
5	14.2	14.1
6	15.1	16.2
7	18.3	19.0
8	22.4	22.5
9	24.5	26.8
10	28.1	28.2
11	5.2	7.7
12	6.4	6.9
13	7.2	8.5
14	12.6	13.9
15	10.5	11.3
16	22.4	24.3
17	29.3	29.1
18	26.9	26.1

The resulting regression equation is

$$\hat{y} = 1.18 + 0.974x$$

Table Z.2-4 summarizes the results of the regression calculations. The confidence interval for the slope is determined as $0.924 < b_1 < 1.024$. Because the confidence interval includes the value 1.0, the data are considered to be unbiased with respect to regression line slope. The confidence interval for the intercept is calculated to be $0.325 < b_0 < 2.036$. In this case, the confidence interval does not include the value 0.0. Consequently, the data are considered to be biased. A closer look at the data shows that the PM concentration measured by Train B was higher than the PM concentration for the same run measured by Train A in 15 of the 18 test runs. The test does not reveal which sampling train provided the more accurate measurements, so both trains should be checked for possible sources of bias. If the source of the bias can be determined, only the data from the train that provided the unbiased results should be used to develop the correlation equation. If the source of the bias cannot be determined, the owner or operator can decide to use the data or repeat the correlation test. However, if paired sampling data are required by regulation, the emission test should be repeated.

Table Z.2-4. Summary of Regression Calculations

Variable	Calculated value	Equation No.
$\bar{x}$	14.56	Z.2-3
$\bar{y}$	15.37	Z.2-3
S_{xx}	1425	Z.2-5
S_{xy}	1388	Z.2-5
b_0	1.18	Z.2-2
b_1	0.974	Z.2-4
S	0.893	Z.2-8
$t_{95\%}$	2.12	Appendix A
$SE(b_1)$	0.0236	Z.2-7
$SE(b_0)$	0.404	Z.2-10
CI_{b1}	$0.924 < b_1 < 1.024$	Z.2-6
CI_{b0}	$0.325 < b_0 < 2.036$	Z.2-9

Z.2.5 Using the Bias Check Spreadsheet

This section provides instructions for using the Bias Check Spreadsheet. The spreadsheet can be used to check paired sampling train data for bias using the procedure described in the previous sections.

The spreadsheet includes three worksheets. The worksheets appear as tabs at the bottom of the screen. You can move from one worksheet to another by clicking on the appropriate worksheet

tab. The initial worksheet (Test Data) is used for data entry. The Calculations worksheet shows the intermediate calculations that are needed for checking bias. The Summary of Results worksheet presents the results of the bias check

The Calculations and Summary worksheets are completely locked, which means that the user can open the worksheet and view the results of the calculations but cannot modify any of the cells in the worksheet. The Test Data worksheet is partially locked. The user enters the test data for one sampling train (Train A) in cells A6 through A65 and the data for the other sampling train (Train B) in cells B6 to B65. It makes no difference in the analysis which train is considered Train A and which is considered Train B. Cells F3 to F6 are used to enter the facility name, location, emission unit, and test dates. All other cells in the Test Data worksheet are locked. Figure Z.2-1 shows the Test Data worksheet for Example Problem A described previously.

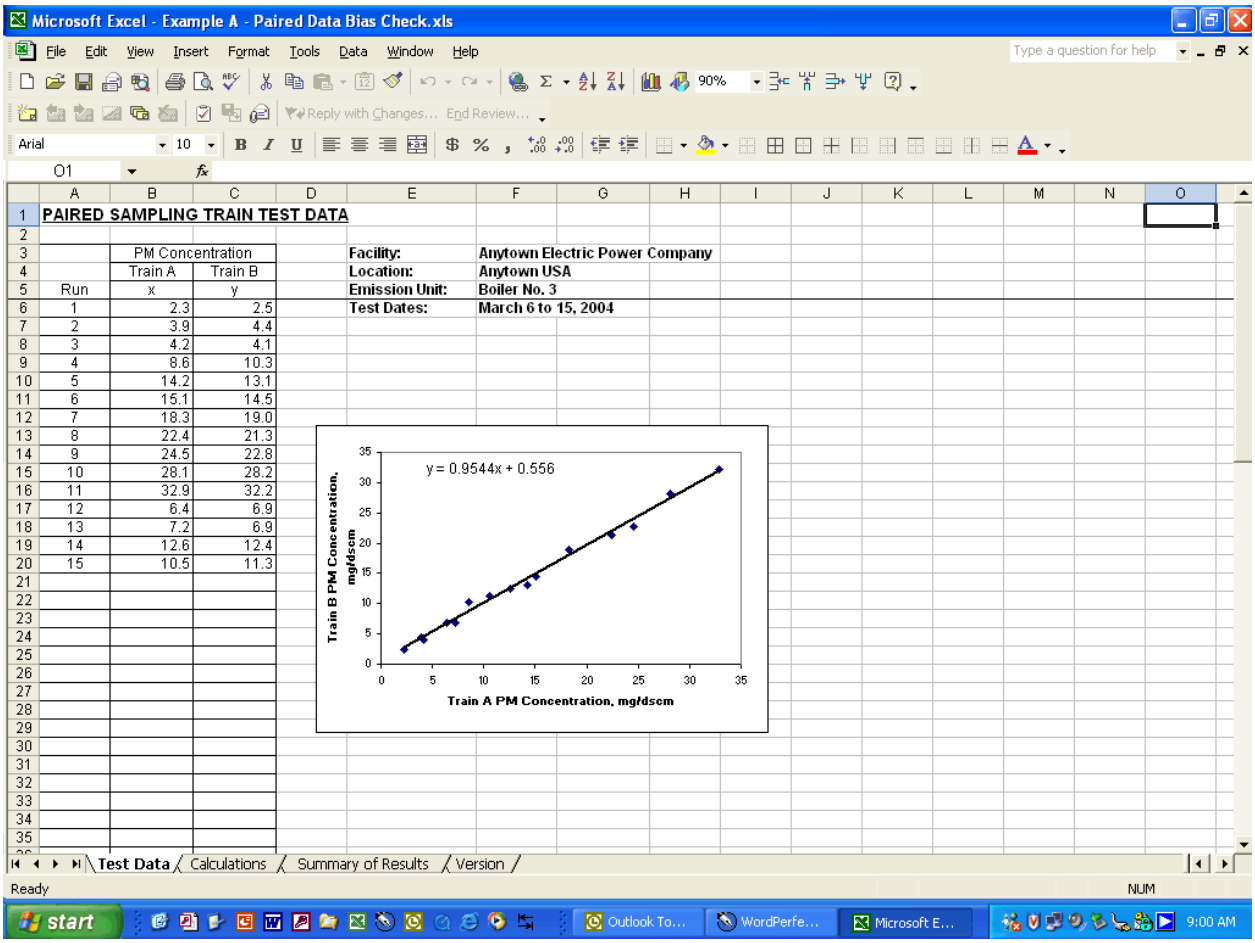
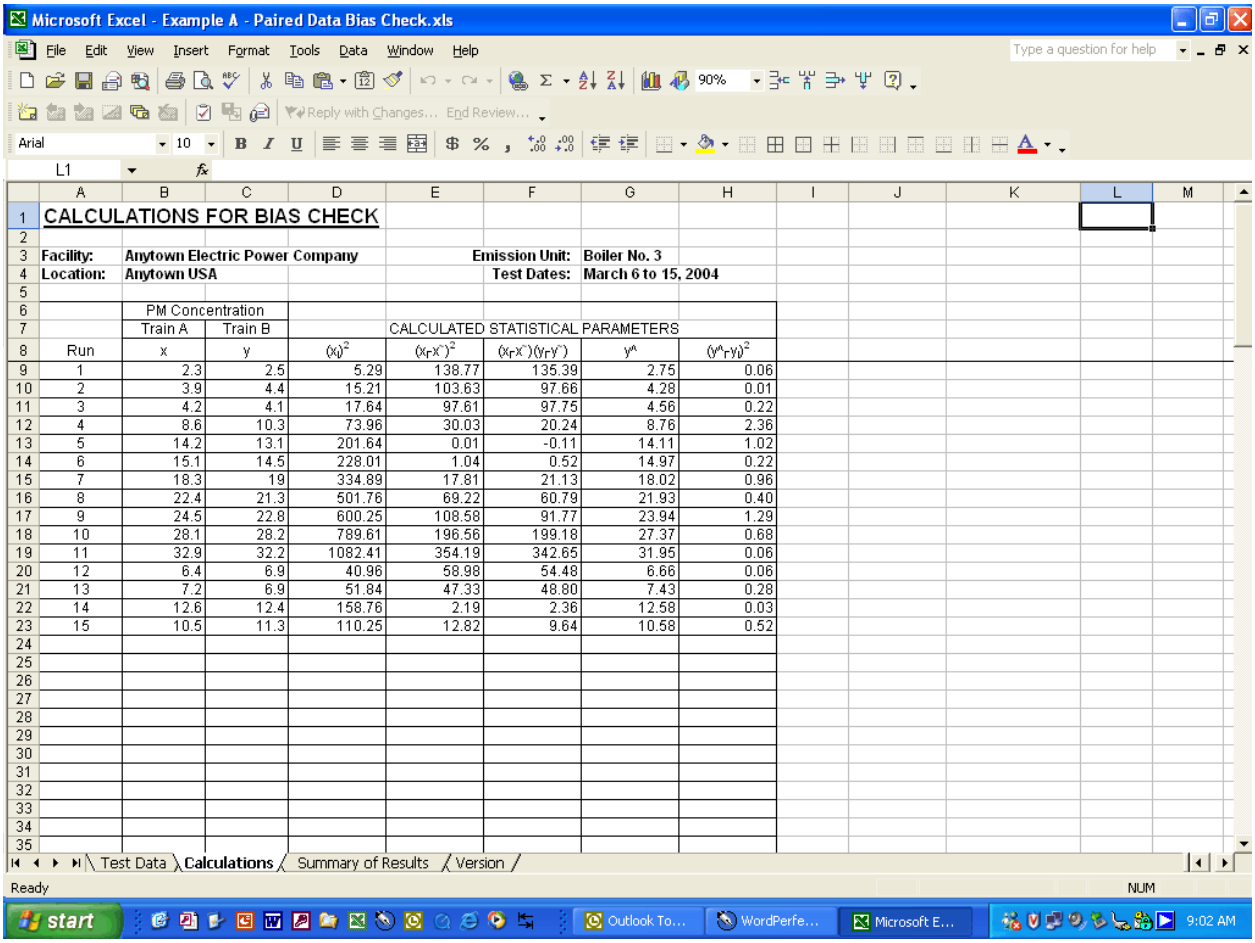


Figure Z.2-1. Example of Test Data worksheet for Bias Check Spreadsheet (Example Problem A).

The Calculations worksheet consists of several columns that display the results of the various calculations needed for the linear regression. All cells are locked in this worksheet. Figure Z.2-2 shows the Calculations worksheet for Example Problem A.



Microsoft Excel - Example A - Paired Data Bias Check.xls

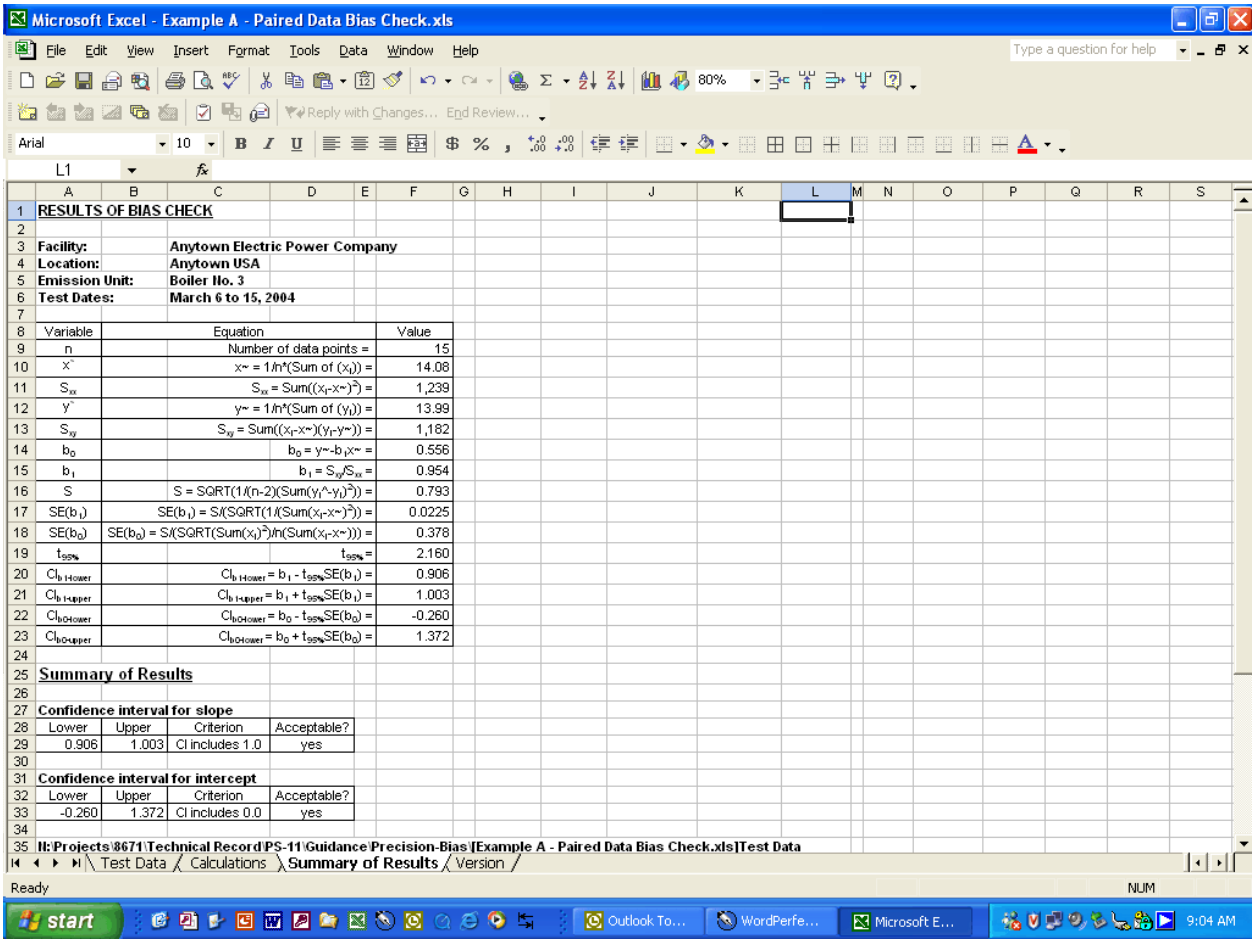
	A	B	C	D	E	F	G	H
1	CALCULATIONS FOR BIAS CHECK							
2								
3	Facility:	Anytown Electric Power Company				Emission Unit:	Boiler No. 3	
4	Location:	Anytown USA				Test Dates:	March 6 to 15, 2004	
5								
6		PM Concentration						
7		Train A	Train B		CALCULATED STATISTICAL PARAMETERS			
8	Run	x	y	$(x_i)^2$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})(y_i-\bar{y})$	$\hat{y}$	$(\hat{y}_i-y_i)^2$
9	1	2.3	2.5	5.29	138.77	135.39	2.75	0.06
10	2	3.9	4.4	15.21	103.63	97.66	4.28	0.01
11	3	4.2	4.1	17.64	97.61	97.75	4.56	0.22
12	4	8.6	10.3	73.96	30.03	20.24	8.76	2.36
13	5	14.2	13.1	201.64	0.01	-0.11	14.11	1.02
14	6	15.1	14.5	228.01	1.04	0.52	14.97	0.22
15	7	18.3	19	334.89	17.81	21.13	18.02	0.96
16	8	22.4	21.3	501.76	69.22	60.79	21.93	0.40
17	9	24.5	22.8	600.25	108.58	91.77	23.94	1.29
18	10	28.1	28.2	789.61	196.56	199.18	27.37	0.68
19	11	32.9	32.2	1082.41	354.19	342.65	31.95	0.06
20	12	6.4	6.9	40.96	58.98	54.48	6.66	0.06
21	13	7.2	6.9	51.84	47.33	48.80	7.43	0.28
22	14	12.6	12.4	158.76	2.19	2.36	12.58	0.03
23	15	10.5	11.3	110.25	12.82	9.64	10.58	0.52

Test Data | Calculations | Summary of Results | Version

Ready NUM

Figure Z.2-2. Example of Calculations worksheet for Bias Check Spreadsheet (Example Problem A).

The Summary of Results worksheet includes the information about the facility, emission unit, and the emission test date that were entered on the Data Entry worksheet, followed by a table displaying the values for the statistical parameters that must be calculated to evaluate the data for bias. Finally, the summary of the results is displayed. This summary includes the upper and lower bounds for the confidence intervals for the slope and intercept of the regression line. The spreadsheet also indicates whether the data satisfy the bias criteria for slope and intercept. Figure Z.2-3 shows the Summary of Results worksheet for Example Problem A.



	A	B	C	D	E	F
1	**RESULTS OF BIAS CHECK**					
2						
3	**Facility:**		**Anytown Electric Power Company**			
4	**Location:**		**Anytown USA**			
5	**Emission Unit:**		**Boiler No. 3**			
6	**Test Dates:**		**March 6 to 15, 2004**			
7						
8	Variable		Equation			Value
9	n		Number of data points =			15
10	$\bar{x}$		$\bar{x} = 1/n*(\text{Sum of } (x_i)) =$			14.08
11	S_{xx}		$S_{xx} = \text{Sum}((x_i-\bar{x})^2) =$			1,239
12	$\bar{y}$		$\bar{y} = 1/n*(\text{Sum of } (y_i)) =$			13.99
13	S_{xy}		$S_{xy} = \text{Sum}((x_i-\bar{x})(y_i-\bar{y})) =$			1,182
14	b_0		$b_0 = \bar{y}-b_1\bar{x} =$			0.556
15	b_1		$b_1 = S_{xy}/S_{xx} =$			0.954
16	S		$S = \text{SQRT}(1/(n-2)(\text{Sum}(y_i\hat{}-y_i)^2)) =$			0.793
17	$SE(b_1)$		$SE(b_1) = S/(\text{SQRT}(1/(\text{Sum}(x_i-\bar{x})^2)) =$			0.0225
18	$SE(b_0)$		$SE(b_0) = S/(\text{SQRT}(\text{Sum}(x_i)^2)/n(\text{Sum}(x_i-\bar{x}))) =$			0.378
19	$t_{95\%}$		$t_{95\%} =$			2.160
20	$CI_{b\,1\text{-lower}}$		$CI_{b\,1\text{-lower}} = b_1 - t_{95\%}SE(b_1) =$			0.906
21	$CI_{b\,1\text{-upper}}$		$CI_{b\,1\text{-upper}} = b_1 + t_{95\%}SE(b_1) =$			1.003
22	$CI_{b0\text{-lower}}$		$CI_{b0\text{-lower}} = b_0 - t_{95\%}SE(b_0) =$			-0.260
23	$CI_{b0\text{-upper}}$		$CI_{b0\text{-lower}} = b_0 + t_{95\%}SE(b_0) =$			1.372
24						
25	**Summary of Results**					
26						
27	**Confidence interval for slope**					
28	Lower	Upper	Criterion	Acceptable?		
29	0.906	1.003	CI includes 1.0	yes		
30						
31	**Confidence interval for intercept**					
32	Lower	Upper	Criterion	Acceptable?		
33	-0.260	1.372	CI includes 0.0	yes		
34						
35	I:\Projects\8671\Technical Record\PS-11\Guidance\Precision-Bias\[Example A - Paired Data Bias Check.xls]Test Data					

Figure Z.2-3. Example of Summary of Results worksheet for Bias Check Spreadsheet (Example Problem A).

Z.2.6 Using Excel Data Analysis Tools to Evaluate Bias

The data analysis tools included in Microsoft Excel also can be used to perform the bias check.

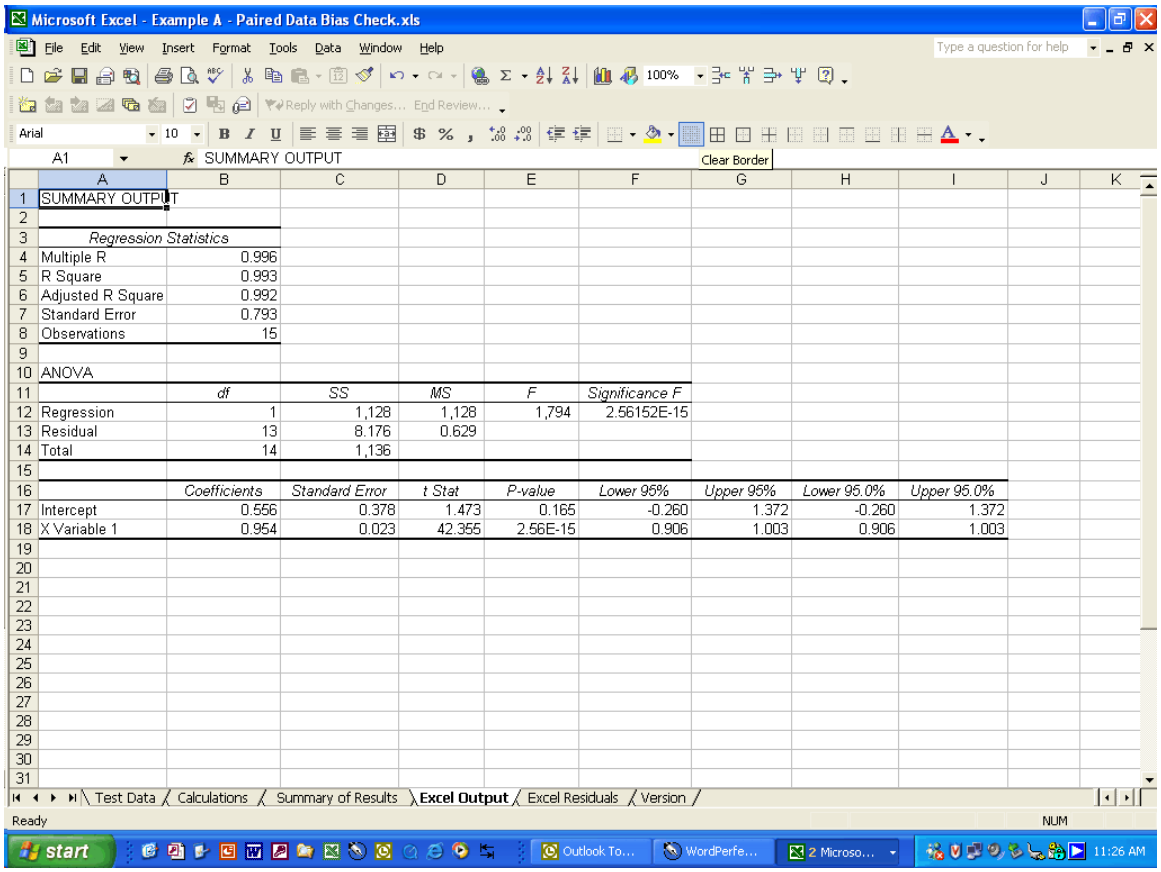
- Key in the data for the two sampling trains in separate columns.
- Under the Tools menu, select Data Analysis.
- Scroll down the list of options to Regression, then click OK.
- Click on the icon in the Input Y Range cell, then block off the data for sampling Train B and press enter.
- Click on the icon in the Input X Range cell, then block off the data for sampling Train A and press enter.

- Check that the Confidence Level is indicated as "95%." If it is not, click on Confidence Level, then key in 95 in the adjacent cell.

- Choose where you want the results to be displayed. If you click on the icon in the Output Range cell, you must select a cell in the spreadsheet where you want the first cell of the results to appear and press enter. If you click on New Worksheet Ply, a new worksheet will be inserted with the results displayed. If you click on New Workbook, a new workbook will be created with the results displayed. Selecting the other options, such as Labels and Residuals, provides additional results, but is not necessary for checking the data for bias.

- After deciding where you want the results displayed, click OK.

The results will then be displayed in the selected location (starting in the cell you indicated for Output Range, new worksheet, or new workbook). Figure Z.2-4 shows the output for the Example Problem A. The last two lines of the results pertain to the regression line intercept and the slope (labeled as "X Variable 1"). The lower and upper limits of the 95 percent confidence interval are displayed several columns to the right of the intercept and slope. If the confidence interval for the intercept includes the value 0, and the confidence interval for the slope includes the value 1, the data satisfy the bias criteria.



	A	B	C	D	E	F	G	H	I	J	K
1	SUMMARY OUTPUT										
2											
3	*Regression Statistics*										
4	Multiple R	0.996									
5	R Square	0.993									
6	Adjusted R Square	0.992									
7	Standard Error	0.793									
8	Observations	15									
9											
10	ANOVA										
11		*df*	*SS*	*MS*	*F*	*Significance F*					
12	Regression	1	1,128	1,128	1,794	2.56152E-15					
13	Residual	13	8.176	0.629							
14	Total	14	1,136								
15											
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*		
17	Intercept	0.556	0.378	1.473	0.165	-0.260	1.372	-0.260	1.372		
18	X Variable 1	0.954	0.023	42.355	2.56E-15	0.906	1.003	0.906	1.003		

Figure Z.2-4. Output of Excel Regression Analysis for Example Problem A.

Appendix A

Student's T-Distribution

Student's T-Distribution: One-Sided at 95% Confidence Interval

Number of pairs (n)	Degrees of freedom	Student's $t_{95\%}$ value
8	6	2.447
9	7	2.365
10	8	2.306
11	9	2.262
12	10	2.228
13	11	2.201
14	12	2.179
15	13	2.160
16	14	2.145
17	15	2.131
18	16	2.120
19	17	2.110
20	18	2.101
21	19	2.093
22	20	2.086
23	21	2.080
24	22	2.074
25	23	2.069
26	24	2.064
27	25	2.060
28	26	2.056
29	27	2.052
30	28	2.048
31	29	2.045
32	30	2.042
33	31	2.040
34	32	2.037
35	33	2.035
36	34	2.032
37	35	2.030
38	36	2.028
39	37	2.026
40	38	2.024
41	39	2.023
42	40	2.021
43	41	2.020
44	42	2.018

(continued)

Student's T-Distribution (continued)

Number of pairs (n)	Degrees of freedom	Student's $t_{95\%}$ value
45	43	2.017
46	44	2.015
47	45	2.014
48	46	2.013
49	47	2.012
50	48	2.011
51	49	2.010
52	50	2.009
53	51	2.008
54	52	2.007
55	53	2.006
56	54	2.005
57	55	2.004
58	56	2.003
59	57	2.002
60	58	2.002
61	59	2.001
62	60	2.000
63	61	2.000
64	62	1.999
65	63	1.998
66	64	1.998
67	65	1.997
68	66	1.997
69	67	1.996
70	68	1.995
71	69	1.995
72	70	1.994
73	71	1.994
74	72	1.993
75	73	1.993
76	74	1.993
77	75	1.992
78	76	1.992
79	77	1.991
80	78	1.991