Systematic Analysis of Bicistronic Reporter Assay Data

Jonathan L. Jacobs, Jonathan D. Dinman
Department of Cell Biology & Molecular Genetics
2135 Microbiology Building
University of Maryland
College Park, MD 20742.

Online Tutorial

---

1. Introduction

[TAKE ME BACK TO THE TOP]

This tutorial is intended to assist you in the analysis of bicistronic reporter data and to guide you in the creation of an Excel spreadsheet that performs the necessary calculations. We therefore assume that you have read the original publication.

The following tutorial is directed specifically towards analysis of bicistronic reporter data from the dual luciferase assay (DLA) system (references1 & 2) for the purpose of studying programmed ribosomal frameshifting. It can, however, be easily adapted to any data set that represents any bicistronic reporter assay system.

Our DLA system requires the collection of data from two different plasmids; a “zero-frame” control plasmid and a “frameshift” plasmid which are each incipiently transformed into identical yeast strains. Both plasmids encode tandem Renilla and luciferase firefly genes which are separated by a multiple cloning site (MCS). The only difference between the two plasmids is the reading frame of Fluc relative to Rluc. In the zero-frame control, Fluc and Rluc are in the same reading frame. Translation of these two ORFs produces a fusion protein that has both Rluc and Fluc activity. Our frameshift reporters differ in that Fluc is out of frame relative to Rluc and a frameshift signal (either a +1 PRF or -1 PRF signal) has been cloned into the MCS. This allows us to study both +1 and -1 frameshifting using the same system. What we are interested in is the ratio of firefly to renilla expression in an experimental reporter normalized to the same ratio in the zero frame control reporter. A schematic of this DLA system is outlined below in Figure 1. More detailed information on our DLA system can be found in Reference #2.

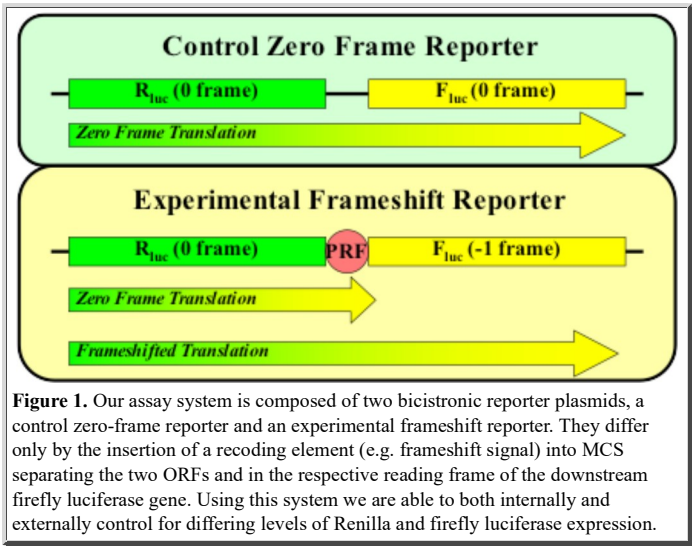


Figure 1. Our assay system is composed of two bicistronic reporter plasmids, a control zero-frame reporter and an experimental frameshift reporter. They differ only by the insertion of a recoding element (e.g. frameshift signal) into MCS separating the two ORFs and in the respective reading frame of the downstream firefly luciferase gene. Using this system we are able to both internally and externally control for differing levels of Renilla and firefly luciferase expression.

The main point here is that, once the data has been collected, you have 2 data points from each trial representing the expression levels of each of two genes. In addition, after the data collection is complete, you will also have a two sets of paired data at a minimum: one set of paired data representing the zero-frame control; and a second of paired data representing the experimental reporter (in our case, a frameshift reporter). In the next section we will briefly outline what this data might look like in an Excel spreadsheet and suggest a few good way to organize it.

For the purposes of this tutorial, we will be analyzing EIGHT data sets and, in the end, will have recreated the same data set that was presented in the NAR publication. A summary of the eight data set is presented below.

Data set	Plasmid	Description
F1	pJD376	A DLA reporter with the L-A virus -1 PRF frameshift signal.
F2		A DLA reporter with a putative -1 PRF signal cloned from the yeast gene.
F3		A DLA reporter with a putative -1 PRF signal cloned from the yeast gene.
F4+		A DLA reporter with the SARS virus -1 PRF frameshift signal in the presence of 20 µg of Anisomycin.
F4-		A DLA reporter with the SARS virus -1 PRF frameshift signal without Anisomycin.
C1	pJD375	Our standard DLA zero-frame control for yeast.
C2+		Our standard DLA zero-frame control for mammalian cell culture in the presence of 20 µg of Anisomycin.
C2-		Our standard DLA zero-frame control for mammalian cell culture without Anisomycin.

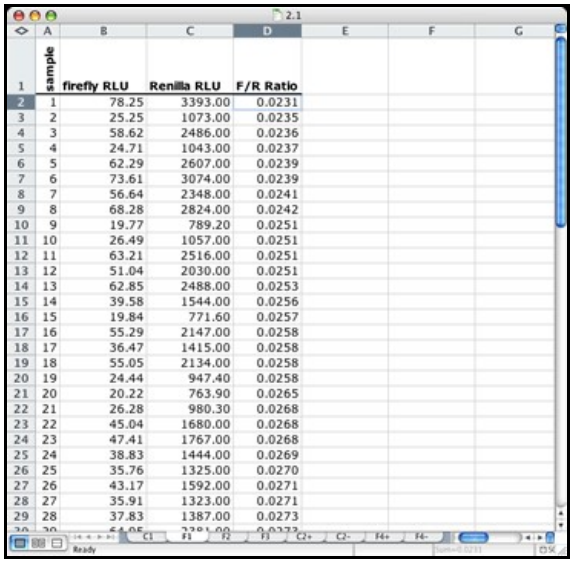
As a final note: the following sections will show screen shots taken from Microsoft Excel 2004 on a Mac running OS X. While the individual screens may look aesthetically different from similar screen shots taken using other versions of Excel on other operating systems, the functions, techniques and charting strategies will be the same.

---

2. Getting Started

[TAKE ME BACK TO THE TOP]

While you are collecting your paired data sets, it's important to organize it efficiently from the start. Below is a screen shot from MS Excel 2004 on OS X of Data set F1 collected from pJD376 (see above).

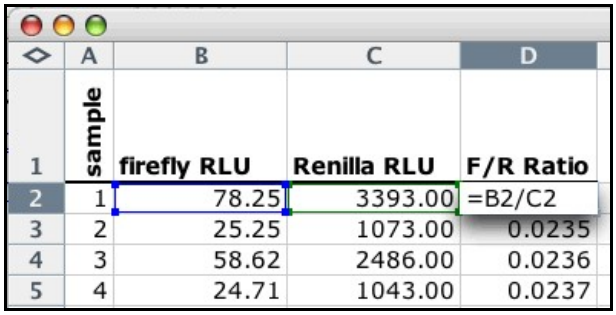


	A	B	C	D	E	F	G
1	sample	firefly RLU	Renilla RLU	F/R Ratio			
2	1	78.25	3393.00	0.0231			
3	2	25.25	1073.00	0.0235			
4	3	58.62	2486.00	0.0236			
5	4	24.71	1043.00	0.0237			
6	5	62.29	2607.00	0.0239			
7	6	73.61	3074.00	0.0239			
8	7	56.64	2348.00	0.0241			
9	8	68.28	2824.00	0.0242			
10	9	19.77	789.20	0.0251			
11	10	26.49	1057.00	0.0251			
12	11	63.21	2516.00	0.0251			
13	12	51.04	2030.00	0.0251			
14	13	62.85	2488.00	0.0253			
15	14	39.58	1544.00	0.0256			
16	15	19.84	771.60	0.0257			
17	16	55.29	2147.00	0.0258			
18	17	36.47	1415.00	0.0258			
19	18	55.05	2134.00	0.0258			
20	19	24.44	947.40	0.0258			
21	20	20.22	763.90	0.0265			
22	21	26.28	980.30	0.0268			
23	22	45.04	1680.00	0.0268			
24	23	47.41	1767.00	0.0268			
25	24	38.83	1444.00	0.0269			
26	25	35.76	1325.00	0.0270			
27	26	43.17	1592.00	0.0271			
28	27	35.91	1323.00	0.0271			
29	28	37.83	1387.00	0.0273			

In total, there are 43 data points and each represents a the expression of firefly and Renilla from an independent cell lysate. The first step is to ***calculate the ratio of Fluc to Rluc*** and then ***rank order the entire set*** from lowest to highest Fluc/Rluc ratio.

Throughout this tutorial, we will present a screen shot of the Excel formula and the matching formula from the publication any time we present a new calculation. It's obvious in this case, but here they are just for consistency.

$$x_i = \frac{F_{RLU}}{R_{RLU}} \qquad [1]$$



	A	B	C	D
1	sample	firefly RLU	Renilla RLU	F/R Ratio
2	1	78.25	3393.00	=B2/C2
3	2	25.25	1073.00	0.0235
4	3	58.62	2486.00	0.0236
5	4	24.71	1043.00	0.0237

This calculation is repeated for each of the eight data sets for every data point.

---

3. Chart Data and Remove Outliers

[TAKE ME BACK TO THE TOP]

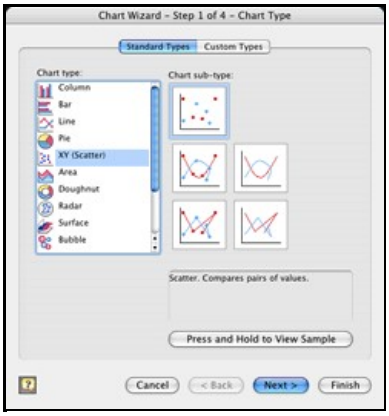
The very first step in any analysis pipeline is to first identify and remove outliers from your data set. We will again be using Data set F1 as an example, but these steps are done for each of the other seven data sets as well.

Below is a screen shot of F1 data charted as Fluc vs Rluc expression values (e.g. reporter gene 1 vs reporter gene 2). This is done simply using the Excel Chart Wizard, choosing XY (Scatter) as your chart type and using the Fluc and Rluc values as the input data. In the context of the Dual Luciferase Assay system, the relationship between Fluc and Rluc should be linear across 10 orders of magnitude. This highly regular linearity in the data can be exploited to give us a first hand, rough estimate of the quality of our data. Your mileage may vary, however, depending on the details of your assay system in use.

Charting the Expression of the Two Genes

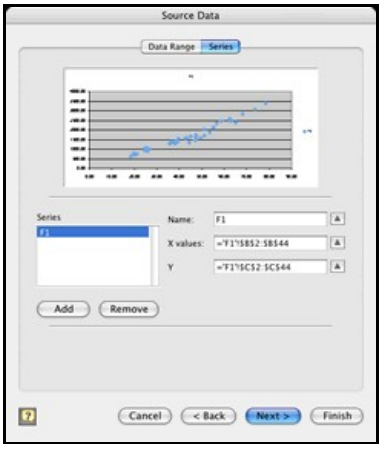
First, chart your data to get a ***qualitative*** view of all your data points. Keep in mind, we are looking for linearity.

1. Select an empty cell on the worksheet that has the data you wish to chart.
2. Open the Chart Wizard. Choose XY (Scatter) as your chart type.



3. Click "NEXT" and choose "Series" at the top (step 2 of 4)
.

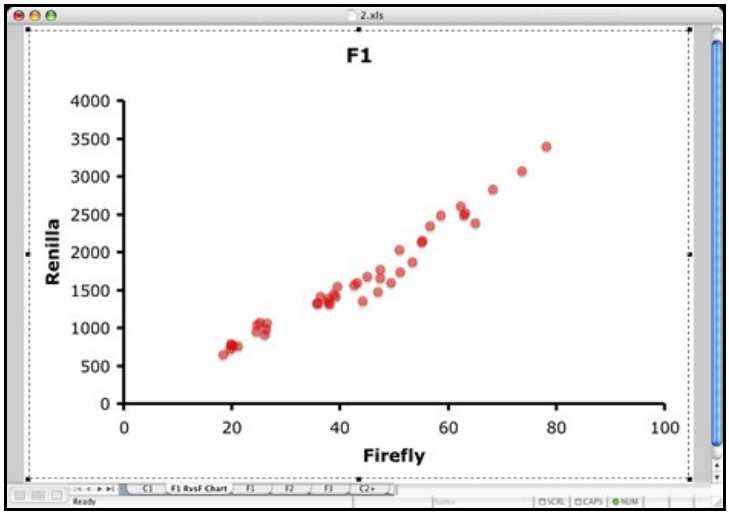
4. Choose "Add" and then set the X values to the range of cells for the Fluc data and the Y values to the Rluc data.

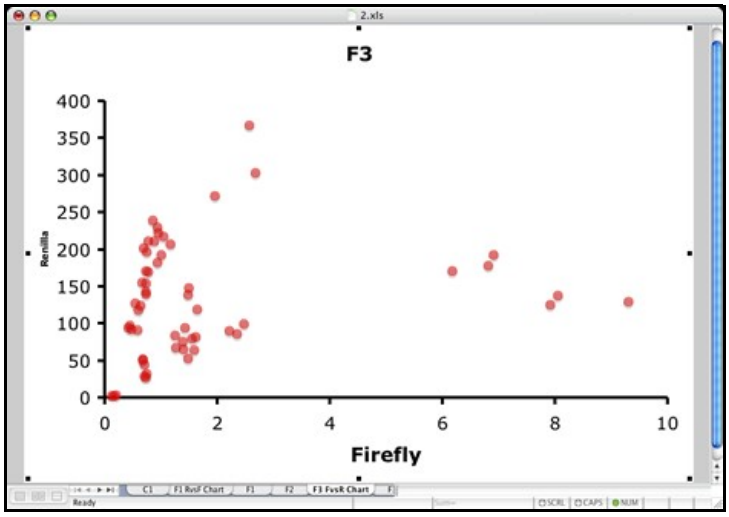


5. Click "Next". Fill out the form that follows and continue choosing "Next".
6. Eventually, you should see a window that's titled "Chart Location". For simplicity, it's usually easier just to make the new chart it's own sheet in the Excel file. Click "OK".

7. Edit the chart fonts and colors to your liking and then your done.

Your new chart should look something like the following one. The F1 data set is generally very linear, as we would expect. If, however, you were to create the same chart with the F3 Data set, you would immediately recognize that this data set has some problems.





You might be tempted to immediately throw away the data from the F3 set, but in general this is not a good practice because this technique is more of a qualitative measure of our Data set's linearity. There are other quantitative methods we will use later on in this tutorial that will *formally* exclude F3 as being a valid Data set

Finding and Excluding Outlier Data

Once your data has been charted in this way, you should next ***quantitatively*** set the boundaries for excluding outliers in your data set. Outliers can be a major source of skewness in any Data set that is assumed to be normally distributed (an assumption we will test later). Therefore, it's important that we exclude out outliers using a method that is specifically "distribution free" so that we don't introduce possible bias into our analysis early on.

The method we use is commonly referred to as the Quartile or Fourth-Spread method (Devore, 2002). Essentially, you identify the boundaries of each of the quartiles in your data set, measure the fourth-spread (*fs*, which is the distance between the lower and upper quartiles), and set the upper and lower outlier boundaries as a function of *fs*. Below are both the formulas (from the publication) and Excel screen shots for doing this. Also, we'll need to create a new sheet in the Excel file ("**Summary**") for collected statistics. And again, we'll start by working on *ratio values* (Fluc/ Rluc) from Data set F1.

1. Find the median, minimum, maximum, 25th and 75th percentile boundaries of the ratios.

maximum	=QUARTILE('F1'!$D$2:$D$44,4)	**0.0327**
75th percentile	=QUARTILE('F1'!$D$2:$D$44,3)	**0.0278**
median	=QUARTILE('F1'!$D$2:$D$44,2)	**0.0268**
25th percentile	=QUARTILE('F1'!$D$2:$D$44,1)	**0.0251**
minimum value	=QUARTILE('F1'!$D$2:$D$44,0)	**0.0231**

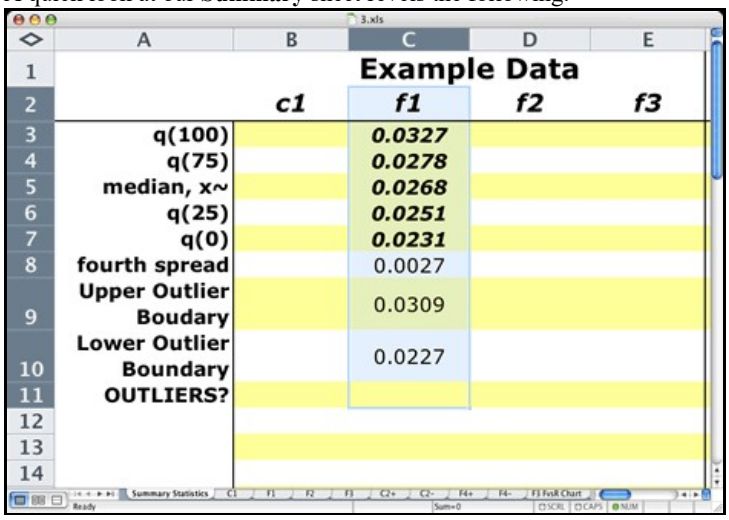
2. The difference between the 75th and 25th percentiles is the fourth spread.

fourth spread (*fs*)	=C4-C6	**0.0027**

3. The Upper and Lower Outlier Boundaries are equal to 1.5 * *fs* above and below the median.

Upper OB	=C5+1.5*C8	**0.0309**
Lower OB	=C5-1.5*C8	**0.0227**

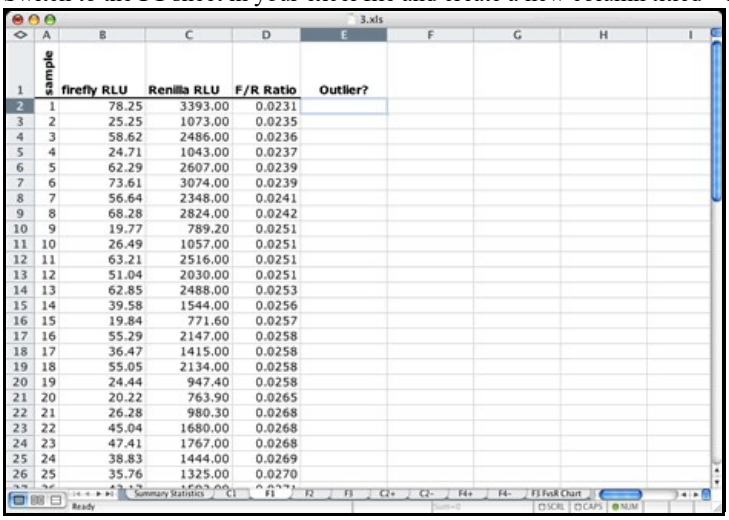
A quick look at our **Summary** sheet revels the following.



	A	B	C	D	E
1			Example Data		
2		c1	f1	f2	f3
3	q(100)		0.0327		
4	q(75)		0.0278		
5	median, x~		0.0268		
6	q(25)		0.0251		
7	q(0)		0.0231		
8	fourth spread		0.0027		
9	Upper Outlier Boudary		0.0309		
10	Lower Outlier Boundary		0.0227		
11	OUTLIERS?				
12					
13					
14					

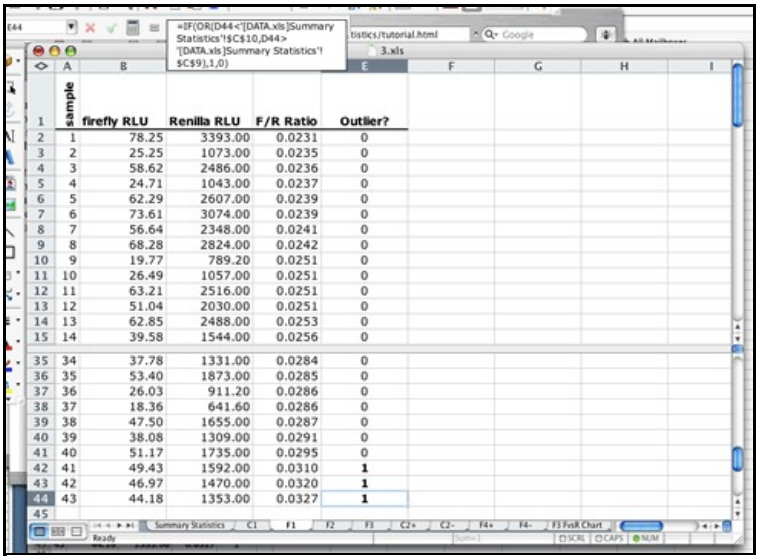
At this point you can remove any outliers from the Data set by hand, or if you choose you can have excel automatically point them out to you using a simple formula. To have your outliers spotted and counted by Excel, continue to Step 4.

4. Switch to the **F1** sheet in your excel file and create a new column titled **"Outlier?"**.



	A	B	C	D	E
1	sample	firefly RLU	Renilla RLU	F/R Ratio	Outlier?
2	1	78.25	3393.00	0.0231	
3	2	25.25	1073.00	0.0235	
4	3	58.62	2486.00	0.0236	
5	4	24.71	1043.00	0.0237	
6	5	62.29	2607.00	0.0239	
7	6	73.61	3074.00	0.0239	
8	7	56.64	2348.00	0.0241	
9	8	68.28	2824.00	0.0242	
10	9	19.77	789.20	0.0251	
11	10	26.49	1057.00	0.0251	
12	11	63.21	2516.00	0.0251	
13	12	51.04	2030.00	0.0251	
14	13	62.85	2488.00	0.0253	
15	14	39.58	1544.00	0.0256	
16	15	19.84	771.60	0.0257	
17	16	55.29	2147.00	0.0258	
18	17	36.47	1415.00	0.0258	
19	18	55.05	2134.00	0.0258	
20	19	24.44	947.40	0.0258	
21	20	20.22	763.90	0.0265	
22	21	26.28	980.30	0.0268	
23	22	45.04	1680.00	0.0268	
24	23	47.41	1767.00	0.0268	
25	24	38.83	1444.00	0.0269	
26	25	35.76	1325.00	0.0270	

5. Using Excel's built in logic formulas, you can create a simple formula to check the value of the ratio column and see if it is above or below either one of the outlier boundaries. A "1" indicates that an individual sample (a row in the Data set) should be considered an outlier. To make things more readable, I usually then bold any outliers that are found. Finally, since we previously sorted our ratios in an ascending order, if any outliers *are* identified, they will always be on the top and bottom of the column.



sample	firefly RLU	Renilla RLU	F/R Ratio	Outlier?
1	78.25	3393.00	0.0231	0
2	25.25	1073.00	0.0235	0
3	58.62	2486.00	0.0236	0
4	24.71	1043.00	0.0237	0
5	62.29	2607.00	0.0239	0
6	73.61	3074.00	0.0239	0
7	56.64	2348.00	0.0241	0
8	68.28	2824.00	0.0242	0
9	19.77	789.20	0.0251	0
10	26.49	1057.00	0.0251	0
11	63.21	2516.00	0.0251	0
12	51.04	2030.00	0.0251	0
13	62.85	2488.00	0.0253	0
14	39.58	1544.00	0.0256	0
34	37.78	1331.00	0.0284	0
35	53.40	1873.00	0.0285	0
36	26.03	911.20	0.0286	0
37	18.36	641.60	0.0286	0
38	47.50	1655.00	0.0287	0
39	38.08	1309.00	0.0291	0
40	51.17	1735.00	0.0295	0
41	49.43	1592.00	0.0310	1
42	46.97	1470.00	0.0320	1
43	44.18	1353.00	0.0327	1

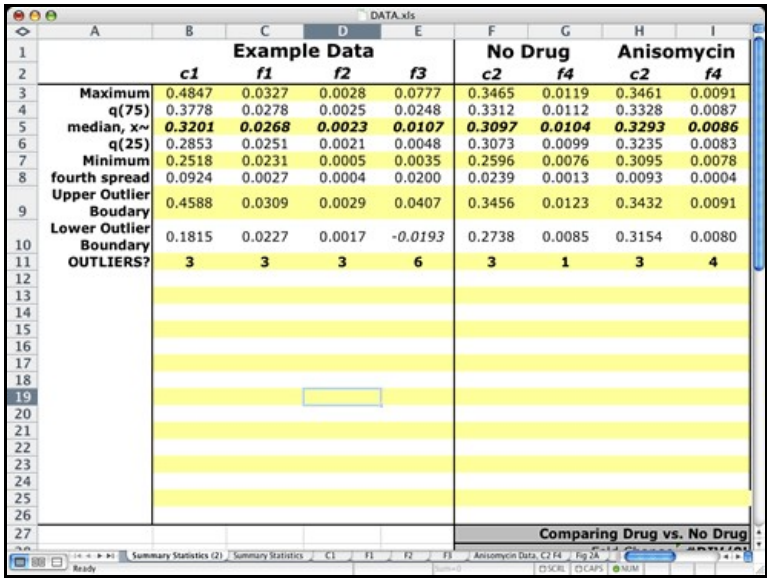
A closer look at the logic formula in cell **E44** (highlighted in blue above):

```
=IF(OR(D44<'[DATA.xls]Summary
Statistics'!$C$10,D44>
'[DATA.xls]Summary Statistics'!
$C$9),1,0)
```

The formula above basically says: "If the ratio of cell **D44** is greater than the Upper boundary or less than the Lower boundary, then this cell is equal to 1. Otherwise, zero."

6. Repeat these steps for all data sets.

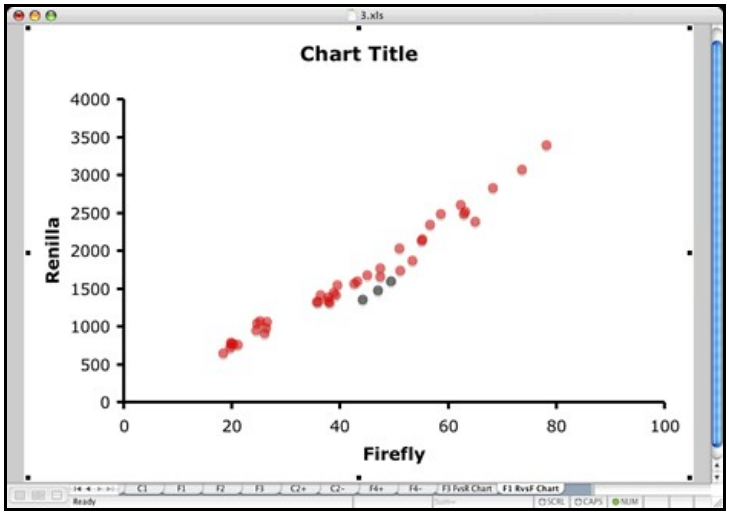
Finally, once you have done this for all your data sets, your **Summary Page** should look something like the screen shot below:

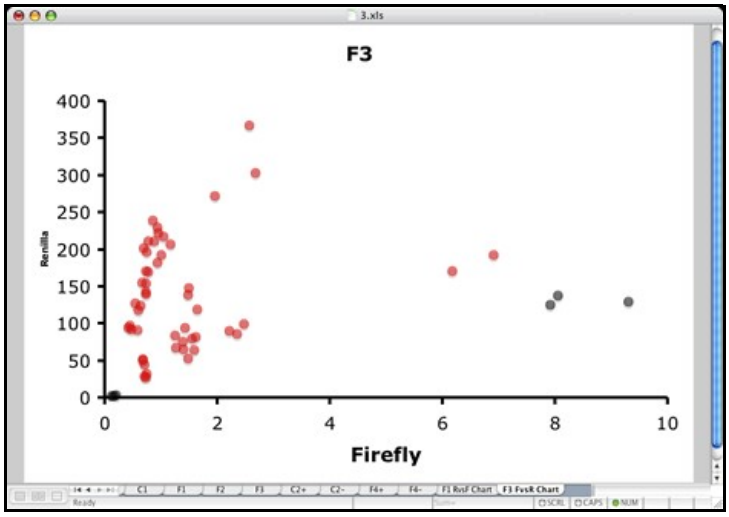


	Example Data				No Drug		Anisomycin	
	c1	f1	f2	f3	c2	f4	c2	f4
Maximum	0.4847	0.0327	0.0028	0.0777	0.3465	0.0119	0.3461	0.0091
q(75)	0.3778	0.0278	0.0025	0.0248	0.3312	0.0112	0.3328	0.0087
median, x~	0.3201	0.0268	0.0023	0.0107	0.3097	0.0104	0.3293	0.0086
q(25)	0.2853	0.0251	0.0021	0.0048	0.3073	0.0099	0.3235	0.0083
Minimum	0.2518	0.0231	0.0005	0.0035	0.2596	0.0076	0.3095	0.0078
fourth spread	0.0924	0.0027	0.0004	0.0200	0.0239	0.0013	0.0093	0.0004
Upper Outlier Boudary	0.4588	0.0309	0.0029	0.0407	0.3456	0.0123	0.3432	0.0091
Lower Outlier Boundary	0.1815	0.0227	0.0017	-0.0193	0.2738	0.0085	0.3154	0.0080
OUTLIERS?	3	3	3	6	3	1	3	4

Comparing Drug vs. No Drug

On a final note about outliers and charting you data; these two initial steps can be tied together so that on your XY scattergram you can visually see which data points along your linear spread are actually being treated as outliers. In the figures below, I have changed the color of points in the chart to black if they are outliers. Notice how this method pulls out outliers that might have otherwise gone unnoticed (in the case of F1), but does not "fix" the data in F3.





---

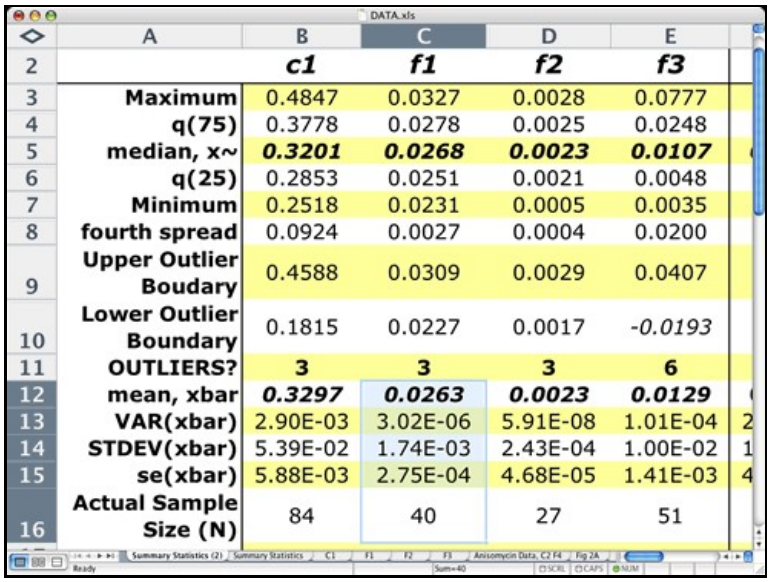
4. Generate Descriptive Statistics

[TAKE ME BACK TO THE TOP]

The next step is to simply generate the standard batch of statistical metrics (for the Fluc/Rluc *ratios*) commonly seen in research papers. The following table summarizes the metrics needed for this tutorial, the formulas used to calculate them, and how you might use excel to find these values. The sample Excel formulas are directed towards the F1 Data set, ***excluding*** the three outliers we previously identified. It's important to remember that, once you have designated a data point as an outlier it should not be included in with your statistical analysis.

Metric	Formula	Expression #	Excel Sample Formula (F1 Data set)
mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	[5]	=AVERAGE('F1'!D3:D42)
variance	$s_{N-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$	[7]	=VAR('F1'!D3:D42)
standard deviation	$s_{N-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$	[8]	=STDEV('F1'!D3:D42)
standard error	$s_e = \frac{s_{N-1}}{\sqrt{n}}$	[9]	=C14/SQRT(C18)
sample size	N	-	=COUNTIF('F1'!E3:E45,"<1")

Once these formulas are applied to each Data set, your **Summary** sheet should look something like the follow screen shot The five values for Data set F1 from the table above and their respective formulas are shown highlighted in blue.



	A	B	C	D	E
2		***c1***	***f1***	***f2***	***f3***
3	**Maximum**	0.4847	0.0327	0.0028	0.0777
4	**q(75)**	0.3778	0.0278	0.0025	0.0248
5	**median, x~**	***0.3201***	***0.0268***	***0.0023***	***0.0107***
6	**q(25)**	0.2853	0.0251	0.0021	0.0048
7	**Minimum**	0.2518	0.0231	0.0005	0.0035
8	**fourth spread**	0.0924	0.0027	0.0004	0.0200
9	**Upper Outlier Boudary**	0.4588	0.0309	0.0029	0.0407
10	**Lower Outlier Boundary**	0.1815	0.0227	0.0017	*-0.0193*
11	**OUTLIERS?**	**3**	**3**	**3**	**6**
12	**mean, xbar**	***0.3297***	***0.0263***	***0.0023***	***0.0129***
13	**VAR(xbar)**	2.90E-03	3.02E-06	5.91E-08	1.01E-04
14	**STDEV(xbar)**	5.39E-02	1.74E-03	2.43E-04	1.00E-02
15	**se(xbar)**	5.88E-03	2.75E-04	4.68E-05	1.41E-03
16	**Actual Sample Size (N)**	84	40	27	51

---

5. Is my data normally distributed?

[TAKE ME BACK TO THE TOP]

By this point you have qualitatively checked your data for linearity, identified and excluded the outliers from each Data set, and generated the standard statistical measurements most often seen in research papers. However, the assumptions of these statistical measures include: 1) that the data is normally distributed (fits a "bell curve"); and 2) that you have done enough replicates for a given degree of statistical confidence. This section deals with verifying the first assumption: that your data is indeed normally distributed.

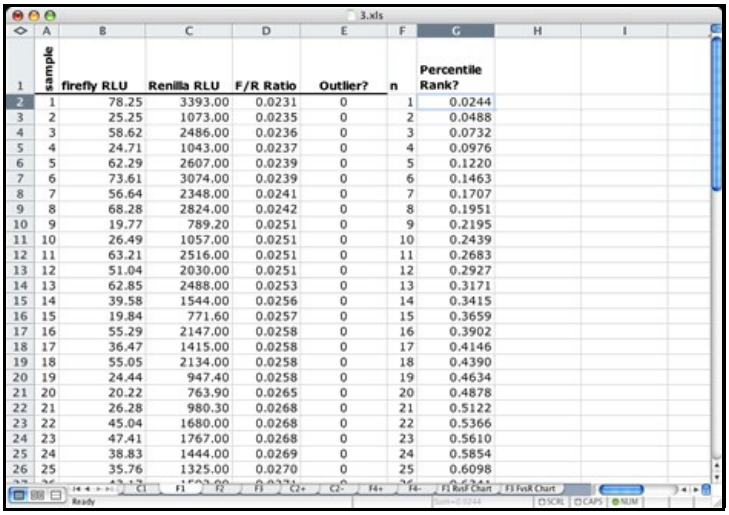
It should be noted that several methods can be found in the statistical literature that can test for whether or not a particular Data set fits a given distribution. In this tutorial (and in our publication) we suggest the creation of a normal probability plot and to find the probability plot correlation coefficient (PPCC). We have found that this method works well for the Dual Luciferase system in use in our laboratory. However, interested readers are directed to the NIST/SEMATECH e-Handbook of Statistical Methods for a broader discussion of other possible solutions. The original publication was by Filliben, 1975.

Finding the PPCC of a given data set involves rank ordering the data and generating *expected* z-scores for each data point. A scattergram (XY) plotting the expected z-score on the x-axis against the actual Fluc/Rluc ratio on the y-axis. Linear least squares regression fits the data to a linear trendline and the correlation coefficient between the two values is the PPCC.

Once you have the PPCC in hand, you can check your value against a table of critical values. If your PPCC exceeds the critical value for your sample size and desired level of confidence, then your data can be considered "*approximately normal*". If it does not, then your data is not normally distributed; i.e. you shouldn't continue with any statistical analysis that assume a normal distribution (t-tests, etc).

Fortunately, with the help of Excel, this process is fairly straight forward. We'll calculate the PPCC of Data set F1 using these steps:

1. Rank order the data by the gene1/gene2 ratios (Fluc/Rluc in our system) if you have not already done so.
2. Create two new columns next to the "Outlier?" column and title them "**n**" and "**Percentile Rank**". The "n" column is going to represent the rank of each data point that was included after we excluded outliers (hence the difference between this column and the first column title "sample"). The second column is a percentile rank of "n" assuming that you have ***N+1*** samples. This is so that the last value is does not have a percentile rank of 100% (which is a problem when you are calculating expected z-scores). A screen shot of the results is below.



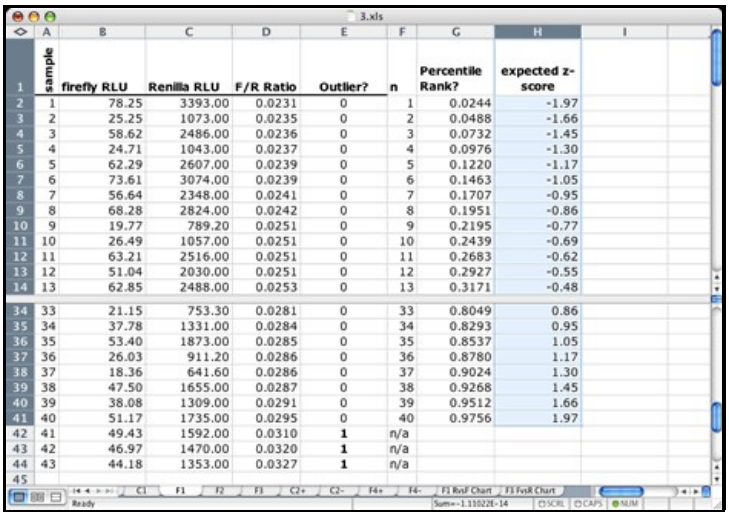
	A	B	C	D	E	F	G	H	I
1	sample	firefly RLU	Renilla RLU	F/R Ratio	Outlier?	n	Percentile Rank?		
2	1	78.25	3393.00	0.0231	0	1	0.0244		
3	2	25.25	1073.00	0.0235	0	2	0.0488		
4	3	58.62	2486.00	0.0236	0	3	0.0732		
5	4	24.71	1043.00	0.0237	0	4	0.0976		
6	5	62.29	2607.00	0.0239	0	5	0.1220		
7	6	73.61	3074.00	0.0239	0	6	0.1463		
8	7	56.64	2348.00	0.0241	0	7	0.1707		
9	8	68.28	2824.00	0.0242	0	8	0.1951		
10	9	19.77	789.20	0.0251	0	9	0.2195		
11	10	26.49	1057.00	0.0251	0	10	0.2439		
12	11	63.21	2516.00	0.0251	0	11	0.2683		
13	12	51.04	2030.00	0.0251	0	12	0.2927		
14	13	62.85	2488.00	0.0253	0	13	0.3171		
15	14	39.58	1544.00	0.0256	0	14	0.3415		
16	15	19.84	771.60	0.0257	0	15	0.3659		
17	16	55.29	2147.00	0.0258	0	16	0.3902		
18	17	36.47	1415.00	0.0258	0	17	0.4146		
19	18	55.05	2134.00	0.0258	0	18	0.4390		
20	19	24.44	947.40	0.0258	0	19	0.4634		
21	20	20.22	763.90	0.0265	0	20	0.4878		
22	21	26.28	980.30	0.0268	0	21	0.5122		
23	22	45.04	1680.00	0.0268	0	22	0.5366		
24	23	47.41	1767.00	0.0268	0	23	0.5610		
25	24	38.83	1444.00	0.0269	0	24	0.5854		
26	25	35.76	1325.00	0.0270	0	25	0.6098		

A close up of the Percentile Rank column (column **G**) shows the formula we use is

```
=F2/(COUNT($F$2:$F$41)+1)
```

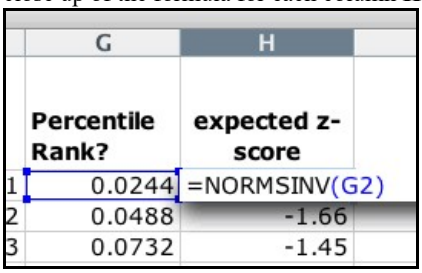
The calculated percentile rank is equivalent to an expected cumulative probability value for obtaining a ratio equal to or less than current one. Because of this, the percentile rank can be used to find the *expected* z-score for each observed F/R ratio.

3. Create another column and title it "**expected z-score**". For each cell in the column, use the Excel function NORMSINV(#) to find the z-score for a given percentile rank. Here's another screen shot:



	A	B	C	D	E	F	G	H	I
1	sample	firefly RLU	Renilla RLU	F/R Ratio	Outlier?	n	Percentile Rank?	expected z-score	
2	1	78.25	3393.00	0.0231	0	1	0.0244	-1.97	
3	2	25.25	1073.00	0.0235	0	2	0.0488	-1.66	
4	3	58.62	2486.00	0.0236	0	3	0.0732	-1.45	
5	4	24.71	1043.00	0.0237	0	4	0.0976	-1.30	
6	5	62.29	2607.00	0.0239	0	5	0.1220	-1.17	
7	6	73.61	3074.00	0.0239	0	6	0.1463	-1.05	
8	7	56.64	2348.00	0.0241	0	7	0.1707	-0.95	
9	8	68.28	2824.00	0.0242	0	8	0.1951	-0.86	
10	9	19.77	789.20	0.0251	0	9	0.2195	-0.77	
11	10	26.49	1057.00	0.0251	0	10	0.2439	-0.69	
12	11	63.21	2516.00	0.0251	0	11	0.2683	-0.62	
13	12	51.04	2030.00	0.0251	0	12	0.2927	-0.55	
14	13	62.85	2488.00	0.0253	0	13	0.3171	-0.48	
34	33	21.15	753.30	0.0281	0	33	0.8049	0.86	
35	34	37.78	1331.00	0.0284	0	34	0.8293	0.95	
36	35	53.40	1873.00	0.0285	0	35	0.8537	1.05	
37	36	26.03	911.20	0.0286	0	36	0.8780	1.17	
38	37	18.36	641.60	0.0286	0	37	0.9024	1.30	
39	38	47.50	1655.00	0.0287	0	38	0.9268	1.45	
40	39	38.08	1309.00	0.0291	0	39	0.9512	1.66	
41	40	51.17	1735.00	0.0295	0	40	0.9756	1.97	
42	41	49.43	1592.00	0.0310	1	n/a			
43	42	46.97	1470.00	0.0320	1	n/a			
44	43	44.18	1353.00	0.0327	1	n/a			

I have split the screen so that you can see how the last three values are outliers and are not being considered in this analysis. Below is a close up of the formula for each column **H** cell. It is simply :



	G	H
	Percentile Rank?	expected z-score
1	0.0244	=NORMSINV(G2)
2	0.0488	-1.66
3	0.0732	-1.45

4. Next, create another column title "**observed z-score**" in column **I**. To find the *observed* z-score we use the following calculation:

z-score $z_{Obs} = \dfrac{x_i - \bar{x}}{s_{N-1}}$ **[10]**

```
=(D2-AVERAGE($D$2:$D$41))/
STDEV($D$2:$D$41)
```

Below is a screen shot of the **F1** worksheet with the formula above in each cell in column **I**.

3.xls

sample	firefly RLU	Renilla RLU	F/R Ratio	Outlier?	n	Percentile Rank?	expected z-score	observed z-score
1	78.25	3393.00	0.0231	0	1	0.0244	-1.97	-1.86
2	25.25	1073.00	0.0235	0	2	0.0488	-1.66	-1.59
3	58.62	2486.00	0.0236	0	3	0.0732	-1.45	-1.56
4	24.71	1043.00	0.0237	0	4	0.0976	-1.30	-1.50
5	62.29	2607.00	0.0239	0	5	0.1220	-1.17	-1.38
6	73.61	3074.00	0.0239	0	6	0.1463	-1.05	-1.35
7	56.64	2348.00	0.0241	0	7	0.1707	-0.95	-1.25
8	68.28	2824.00	0.0242	0	8	0.1951	-0.86	-1.22
9	19.77	789.20	0.0251	0	9	0.2195	-0.77	-0.71
10	26.49	1057.00	0.0251	0	10	0.2439	-0.69	-0.71
11	63.21	2516.00	0.0251	0	11	0.2683	-0.62	-0.67
12	51.04	2030.00	0.0251	0	12	0.2927	-0.55	-0.66
13	62.85	2488.00	0.0253	0	13	0.3171	-0.48	-0.59
33	21.15	753.30	0.0281	0	33	0.8049	0.86	1.03
34	37.78	1331.00	0.0284	0	34	0.8293	0.95	1.20
35	53.40	1873.00	0.0285	0	35	0.8537	1.05	1.28
36	26.03	911.20	0.0286	0	36	0.8780	1.17	1.31
37	18.36	641.60	0.0286	0	37	0.9024	1.30	1.34
38	47.50	1655.00	0.0287	0	38	0.9268	1.45	1.39
39	38.08	1309.00	0.0291	0	39	0.9512	1.66	1.61
40	51.17	1735.00	0.0295	0	40	0.9756	1.97	1.84
41	49.43	1592.00	0.0310	1	n/a			
42	46.97	1470.00	0.0320	1	n/a			
43	44.18	1353.00	0.0327	1	n/a			

Now that you have both *expected* and *observed* z-scores for each data point, you can calculate the correlation coefficient between them (i.e. PPCC value) and plot the data as F/R ratio vs. expected z-score.

5. The PPCC is a degree of correlation between the two expected and observed z-scores for your data set. A value of 1 indicates perfect correlation (i.e. you data fit a normal distribution perfectly). A value of 0 indicates no correlation. Below is the formula for this calculation, and fortunately Excel has this function built in...

$$PPCC(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

[11] `=CORREL('F1'!H3:H42,'F1'!I3:I42)`

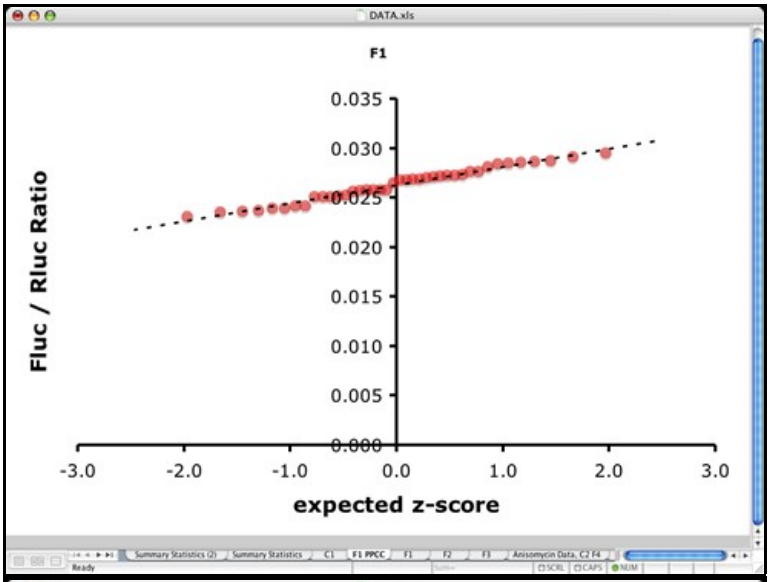
Switch back to the **Summary** worksheet and add another row for the PPCC value for each data set. Below is screen shot of this worksheet with all the values filled out for each Data set You will also notice that we have added a second row titled "Critical Value". This is the minimum value that your PPCC value must exceed in order for you to safely assume that your data is at approximately normally distributed. A more detailed discussion of PPCC values and the associated critical values can be found here (**NIST**), and the PPCC critical values can be found here as well (**PPCC Critical Values**).

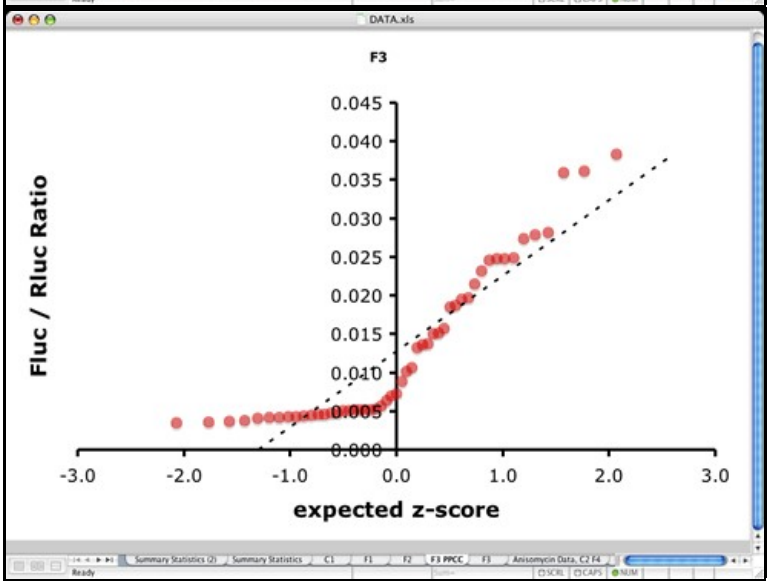
DATA.xls

		Example Data				No Drug	
		c1	f1	f2	f3	c2	f4
3	Maximum	0.4847	0.0327	0.0028	0.0777	0.3465	0.01
4	q(75)	0.3778	0.0278	0.0025	0.0248	0.3312	0.01
5	median, x~	***0.3201***	***0.0268***	***0.0023***	***0.0107***	***0.3097***	***0.01***
6	q(25)	0.2853	0.0251	0.0021	0.0048	0.3073	0.00
7	Minimum	0.2518	0.0231	0.0005	0.0035	0.2596	0.00
8	fourth spread	0.0924	0.0027	0.0004	0.0200	0.0239	0.00
9	Upper Outlier Boudary	0.4588	0.0309	0.0029	0.0407	0.3456	0.01
10	Lower Outlier Boundary	0.1815	0.0227	0.0017	*-0.0193*	0.2738	0.00
11	OUTLIERS?	3	3	3	6	3	1
12	mean, xbar	***0.3297***	***0.0263***	***0.0023***	***0.0129***	**0.3163**	**0.01**
13	VAR(xbar)	2.90E-03	3.02E-06	5.91E-08	1.01E-04	2.86E-04	6.63E
14	STDEV(xbar)	5.39E-02	1.74E-03	2.43E-04	1.00E-02	1.69E-02	8.14E
15	se(xbar)	5.88E-03	2.75E-04	4.68E-05	1.41E-03	4.37E-03	1.97E
16	Actual Sample Size (N)	84	40	27	51	15	17
17	PPCC	0.9779	0.9896	0.9865	0.9191	0.9325	0.99
18	critical value	0.9771	0.9576	0.9413	0.9654	0.9080	0.91
19	Normal?	YES	YES	YES	NO	YES	YE

The F1 data set is highlighted in blue, but if you notice the F3 data set is *not* normally distributed and it fails to pass this test. So, although we knew that F3 was a "bad" Data set after we charted Fluc vs. Rluc on a scatterplot, we now have a quantitative metric that systematically rejects the entire F3 Data set from any further analysis.

6. Although technically optional, the final step involves actually plotting the data. Simple construct an XY scatterplot using the observed ratios and the expected z-scores for those ratios. You should end up with a chart similar to the ones shown below.





As you can see from the above charts, F1 is very linear even when compared to an "idealized" or expected set of values. Furthermore, the values are more heavily distributed around the middle range; much like a bell curve. The F3 data is again shown to be poor quality and it fails it's PPCC test.

So in the end all the data sets in this tutorial pass the test for normal distribution except for F3. The next step is the see if we carried out enough replicates in order to get a good estimate of our sample mean.

---

6. Have I done enough replicates?

[TAKE ME BACK TO THE TOP]

It's typical to simply want to do three replicate experiments. Unfortunately, the number of replicates you need to do for any experiment is directly related to 1) the degree of confidence you want to have in your estimate of the sample mean and 2) the degree of variability present in your samples (i.e. the variance). The second point is commonly overlooked (read: ignored) by most molecular biologists because 3 replicates is usually enough, right?

The most common method to estimate the "minimum sample size" for any experimental data set is shown below and can be found in most statistics handbooks. The problem with the formula is that it often *underestimates* the true minimum number of replicates needed in practical situations. In other words, it is an idealized uncorrected minimum sample size estimator. Kupper & Hafner published a conversion table that will provide a researcher with a "corrected minimum sample size" for a given confidence level and error rate.

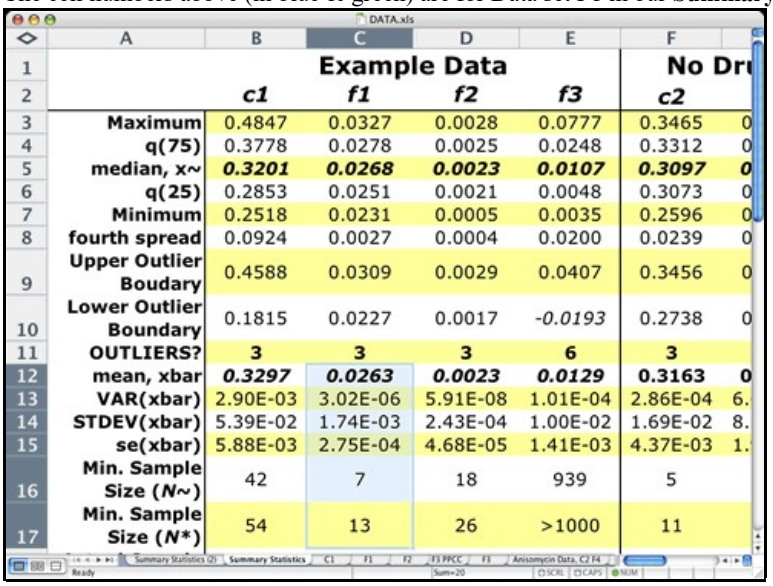
So, to get started follow the steps below. I'll be using the F1 and F3 data sets from the publication as examples of "good" and "bad" data sets. Regardless, "three's a charm" does not apply for either of them.

1. Decide *a priori* what your confidence level is for your sample mean. Usually 95% confidence level is sufficient (i.e. alpha = 0.05), with a z(alpha) = 1.96. A table of alpha values and z scores can be found HERE.
2. Decide next what the acceptable degree of error is (E). This is like saying "If my sample mean is X, then I am willing to be as much as Y% wrong from the *true* value of the mean (something that can *never* be directly measured, it's always an estimate, hence "sample mean"). I usually stick to an E of 10%, smaller values of E make the sample size requirements really big fast.
3. Next, use the following formula. Below is the formula from the publication, the expression number, and the same calculation done using Excel.

$$\tilde{N} = \left\lceil \left(2z_{\alpha/2} \times \frac{S_{N-1}}{E} \right)^2 \right\rceil \qquad [12]$$

`=CEILING((2*1.96*C14/(0.1*C12))^2,1)`

The cell numbers above (in blue & green) are for Data set F1 in our **Summary** worksheet. Below is a screen shot of how this all looks:



	A	B	C	D	E	F	
1		Example Data				No Dr	
2		c1	f1	f2	f3	c2	
3	Maximum	0.4847	0.0327	0.0028	0.0777	0.3465	0
4	q(75)	0.3778	0.0278	0.0025	0.0248	0.3312	0
5	median, x~	0.3201	0.0268	0.0023	0.0107	0.3097	0
6	q(25)	0.2853	0.0251	0.0021	0.0048	0.3073	0
7	Minimum	0.2518	0.0231	0.0005	0.0035	0.2596	0
8	fourth spread	0.0924	0.0027	0.0004	0.0200	0.0239	0
9	Upper Outlier Boudary	0.4588	0.0309	0.0029	0.0407	0.3456	0
10	Lower Outlier Boundary	0.1815	0.0227	0.0017	-0.0193	0.2738	0
11	OUTLIERS?	3	3	3	6	3	
12	mean, xbar	0.3297	0.0263	0.0023	0.0129	0.3163	0
13	VAR(xbar)	2.90E-03	3.02E-06	5.91E-08	1.01E-04	2.86E-04	6.
14	STDEV(xbar)	5.39E-02	1.74E-03	2.43E-04	1.00E-02	1.69E-02	8.
15	se(xbar)	5.88E-03	2.75E-04	4.68E-05	1.41E-03	4.37E-03	1.
16	Min. Sample Size (N~)	42	7	18	939	5	
17	Min. Sample Size (N*)	54	13	26	>1000	11	

Notice how the value of N~ (uncorrected minimum sample size) is only 7 replicates. But if you look at the F3 data set, N~ is 939! This is because the variance of F3 is so huge (relative to F1). Remember, the larger the variance, the more replicates you need to achieve the same level of confidence in your mean's estimate.

4. The next step to do is to use Kupper & Hafner's table to find the closest estimate of the *corrected* sample size (N*), this is unfortunately always a larger value than N~; but statistically more sound. In the case of the F1 Data set, our corrected minimum sample size (N*) ends up being 13.

This concludes this section. If you also complete the previous section "Is my data normally distributed?", then two fundamental assumptions about the data have been tested and we can continue on to calculating the *ratiometric* data and how to test to see if two Data set are statistically different or not.

---

7. Calculate *Ratiometric* Statistics

[TAKE ME BACK TO THE TOP]

In this section we are going to calculate the *ratiometric* statistics for each of the experimental data sets The term "ratiometric" is jargon that simply means: the ratio of the experimental data relative to some control data. Some labs simply refer this as "fold - change" statistics, % frameshifting, recoding efficiency, etc. Calculating ratiometric means and variances are not, however, as simply as you may expect. Finding the correct estimate for the variance of your ratio, for example, does not use the same formula for calculating variance of primary observational data. This is primarily because of "propagation of error" (PoE), a problem that arise whenever you are combining two sets of data together. PoE arises because the statistics from each Data set are *estimates* of their true values; e.g. the sample mean and sample variance of the F1 Data set is an estimate of the true population mean and population variance. When you combine two data sets together you end up compounding the problem because you are also combing these estimates together. In a way, making an estimate of an estimate. A better discussion of this topic can be found HERE.

1. Relative Expression of Experimental Reporter

The **relative** expression of your **experimental** reporter should be normalized to some **control** reporter. In the context of our dual luciferase system, this is termed % frameshifting. Below is the formula from the publication, the expression number, and a screen shot of the Excel cell formula.

$$\bar{x}_R = \frac{\bar{x}_E}{\bar{x}_C}$$ **[13]**

```
=C12/B12
```

DATA.xls

	A	B	C	D	E	F	G	H	I	J
1			Example Data			No Drug		Anisomycin		
2		c1	f1	f2	f3	c2	f4	c2	f4	
3	Maximum	0.4847	0.0327	0.0028	0.0777	0.3465	0.0119	0.3461	0.0091	
4	q(75)	0.3778	0.0278	0.0025	0.0248	0.3312	0.0112	0.3328	0.0087	
5	median, x~	0.3201	0.0268	0.0023	0.0107	0.3097	0.0104	0.3293	0.0086	
6	q(25)	0.2853	0.0251	0.0021	0.0048	0.3073	0.0099	0.3235	0.0083	
7	Minimum	0.2518	0.0231	0.0005	0.0035	0.2596	0.0076	0.3095	0.0078	
8	fourth spread	0.0924	0.0027	0.0004	0.0200	0.0239	0.0013	0.0093	0.0004	
9	Upper Outlier Boudary	0.4588	0.0309	0.0029	0.0407	0.3456	0.0123	0.3432	0.0091	
10	Lower Outlier Boundary	0.1815	0.0227	0.0017	-0.0193	0.2738	0.0085	0.3154	0.0080	
11	OUTLIERS?	3	3	3	6	3	1	3	4	
12	mean, xbar	0.3297	0.0263	0.0023	0.0129	0.3163	0.0106	0.3272	0.0086	
13	VAR(xbar)	2.90E-03	3.02E-06	5.91E-08	1.01E-04	2.86E-04	6.63E-07	3.72E-05	5.36E-08	
14	STDEV(xbar)	5.39E-02	1.74E-03	2.43E-04	1.00E-02	1.69E-02	8.14E-04	6.10E-03	2.31E-04	
15	se(xbar)	5.88E-03	2.75E-04	4.68E-05	1.41E-03	4.37E-03	1.97E-04	1.57E-03	6.19E-05	
16	Min. Sample Size (N~)	42	7	18	939	5	10	1	2	
17	Min. Sample Size (N*)	54	13	26	>1000	11	17	6	7	
18	Actual Sample Size (N)	84	40	27	51	15	17	15	14	
19	Enough Sampling?	YES	YES	YES	NO	YES	YES	YES	YES	
20	PPCC	0.9779	0.9896	0.9865	0.9191	0.9325	0.9920	0.9749	0.9733	
21	critical value	0.9771	0.9576	0.9413	0.9654	0.9080	0.9160	0.9080	0.9029	
22	Normal?	YES	YES	YES	NO	YES	YES	YES	YES	
23	frameshifting, xbar_r		0.0798	0.0070	N/C		0.0335		0.0264	
24	se(xbar_r)		1.65E-03	1.89E-04	N/C		7.77E-04		2.28E-04	
25	stdev(xbar_r)		1.41E-02	1.36E-03			3.14E-03		8.61E-04	
26	var(xbar_r)		1.98E-04	1.84E-06	N/C		9.83E-06		7.42E-07	
27						Comparing Drug vs. No Drug				

For F1, xbar_r is 0.0798 and is simply the ratio 0.0263 / 0.3297.

2. Finding the Variance of xbar_r

Finding the relative sample variance, var(xbar_r), requires the use of a specialized formula (see Kendall & Stuart, 1994) that properly accounts for the relative contribution of the variances from each data set (F1 and C1 in this example).

$$s_R^2 = \frac{s_E^2}{\left(\bar{x}_C\right)^2} + \frac{\left(\bar{x}_E\right)^2 s_C^2}{\left(\bar{x}_C\right)^4}$$ **[14]**

```
=C13/(B12^2)+((C12^2)*B13)/(B12^4)
```

The screen shot above of the summary page may provide clarity.

The values s_E^2 and s_C^2 are the sample variance from the F1 and C1 data set respectively (cells C13 & C12).

3. Finding the Standard Deviation of xbar_r

Similarly, the standard deviation of a combined ratiometric mean is found using the following formula (a derivation of the above formula from reference #4):

$$s_R = \bar{x}_R \times \sqrt{\left(\frac{s_E^2}{\bar{x}_E}\right)^2 + \left(\frac{s_C^2}{\bar{x}_C}\right)^2}$$ **[15]**

```
=C23*SQRT((B13/B12)^2+(C13/C12)^2)
```

The screen shot above of the summary page may provide clarity.

4. Finding the Standard Error of xbar_r

This value is commonly used to determine the size of the error bars on charts reporting xbar_r (i.e. % frameshifting). A bit more formally, it is an measure of the accuracy of your estimate of the sample mean.The **Summary Page** in the Excel spreadsheet uses the following formula (from a personal communication with Dr. Ray Koopman):

$$s_e\left(x_R\right) = \bar{x}_R \times \sqrt{\frac{s_E^2 / N_E}{\left(\bar{x}_E\right)^2} + \frac{s_C^2 / N_C}{\left(\bar{x}_C\right)^2}}$$ **[16]**

```
=C23*SQRT((C13/C18)/(C12^2)+(B13/B18)/(B12^2))
```

8. How do I compare results from two experiments?

[TAKE ME BACK TO THE TOP]

Quote from the paper:

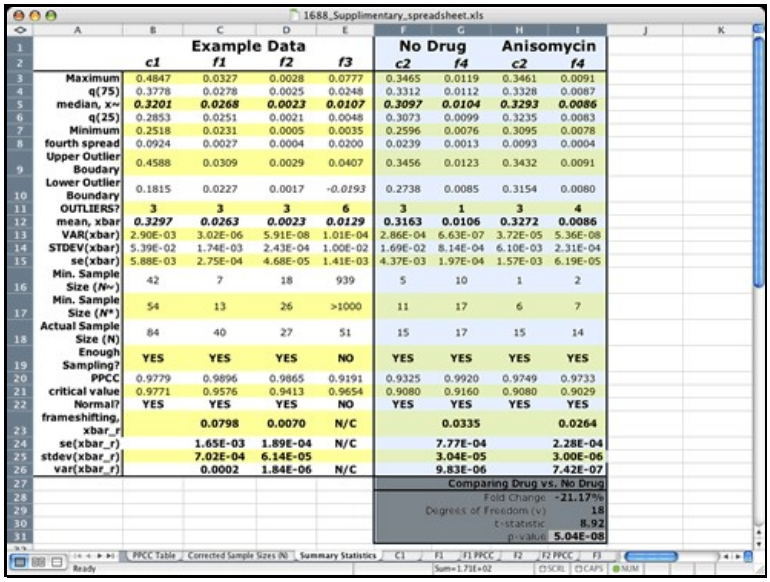
> **... Comparing Data sets:** The final stage is to determine if two experiments, each with their own respective values of xbar_r and var(xbar_r) are statistically different. The published record of studies utilizing various bicistronic reporters shows a wide variety of methods including fold-change, z-tests, or chi-square tests. For comparisons between data sets, a z-test is appropriate only for larger data sets with at least 40 samples each. Data sets for bicistronic reporter systems are usually not this large. Furthermore, a chi-square test is inappropriate as it requires both large sample sizes and that the data be separated of into discrete categorical values. We instead use the unpaired two-sample t-test (see expressions [17] and [18]) since it is more appropriate for smaller continuous data sets The requirements of this test are that the data must be normally distributed and independent, which are satisfied by the bicistronic assay data sets presented here. The hypothesis tested against states that two data sets (X & Y) come from the same population. A rejected hypothesis therefore affirms that the two data sets are indeed statistically different at some predefined confidence level (e.g. 95%, alpha = 0.05). The p-value obtained from this test is an estimation of the probability of an incorrect conclusion...

So, an unpaired two-sample t-test is carried out by first determining the degrees of freedom (no guessing!) and then calculating the ***t*** statistic. The value of ***t*** is then compared to a critical value on a lookup table(this table can be found in nearly all statistics textbooks) for a given confidence interval and Degrees of Freedom. If the t-statistic is greater than the critical value, then your two data sets are statistically different under the assumptions of this test (normally distributed, sufficient samples, etc). This test is precisely the reason why we go to such length is calculating an accurate value for the variance of each of the experiments; var(xbar_r). Without an accurate measure of the variance, the results of this t-test will be incorrect.

In the following example we will be comparing the two F4 data sets with and without the inclusion of the drug Anisomycin (see the paper for more details).

1. Find the degrees of freedom, *v*

From the sample Excel spreadsheets summary page, ***v*** is the value in cell **I29**:



1688_Supplimentary_spreadsheet.xls

		Example Data				No Drug		Anisomycin	
		c1	f1	f2	f3	c2	f4	c2	f4
3	Maximum	0.4847	0.0327	0.0028	0.0777	0.3465	0.0119	0.3461	0.0091
4	q(75)	0.3778	0.0278	0.0025	0.0248	0.3312	0.0112	0.3328	0.0087
5	median, x~	0.3201	0.0268	0.0023	0.0107	0.3097	0.0104	0.3293	0.0086
6	q(25)	0.2853	0.0251	0.0021	0.0048	0.3073	0.0099	0.3235	0.0083
7	Minimum	0.2518	0.0231	0.0005	0.0035	0.2596	0.0076	0.3095	0.0078
8	fourth spread	0.0924	0.0027	0.0004	0.0200	0.0239	0.0013	0.0093	0.0004
9	Upper Outlier Boudary	0.4588	0.0309	0.0029	0.0407	0.3456	0.0123	0.3432	0.0091
10	Lower Outlier Boundary	0.1815	0.0227	0.0017	-0.0193	0.2738	0.0085	0.3154	0.0080
11	OUTLIERS?	3	3	3	6	3	1	3	4
12	mean, xbar	0.3297	0.0263	0.0023	0.0129	0.3163	0.0106	0.3272	0.0086
13	VAR(xbar)	2.90E-03	3.02E-06	5.91E-08	1.01E-04	2.86E-04	6.63E-07	3.72E-05	5.36E-08
14	STDEV(xbar)	5.39E-02	1.74E-03	2.43E-04	1.00E-02	1.69E-02	8.14E-04	6.10E-03	2.31E-04
15	se(xbar)	5.88E-03	2.75E-04	4.68E-05	1.41E-03	4.37E-03	1.97E-04	1.57E-03	6.19E-05
16	Min. Sample Size (N~)	42	7	18	939	5	10	1	2
17	Min. Sample Size (N*)	54	13	26	>1000	11	17	6	7
18	Actual Sample Size (N)	84	40	27	51	15	17	15	14
19	Enough Sampling?	YES	YES	YES	NO	YES	YES	YES	YES
20	PPCC	0.9779	0.9896	0.9865	0.9191	0.9325	0.9920	0.9749	0.9733
21	critical value	0.9771	0.9576	0.9413	0.9654	0.9080	0.9160	0.9080	0.9029
22	Normal?	YES	YES	YES	NO	YES	YES	YES	YES
23	frameshifting, xbar_r		0.0798	0.0070	N/C		0.0335		0.0264
24	se(xbar_r)		1.65E-03	1.89E-04	N/C		7.77E-04		2.28E-04
25	stdev(xbar_r)		7.02E-04	6.14E-05			3.04E-05		3.00E-06
26	var(xbar_r)		0.0002	1.84E-06	N/C		9.83E-06		7.42E-07
27							Comparing Drug vs. No Drug		
28							Fold Change		-21.17%
29							Degress of Freedom (v)		18
30							t-statistic		8.92
31							p-value		5.04E-08

Is is found, ***v*=18**, using the following calculation:

$$v_{a,b} = \left\lfloor \frac{\left(\frac{s_{R_a}^2}{n_a} + \frac{s_{R_b}^2}{n_b}\right)^2}{\frac{\left(s_{R_a}^2/n_a\right)^2}{n_a - 1} + \frac{\left(s_{R_b}^2/n_b\right)^2}{n_b - 1}} \right\rfloor \quad [17]$$

```
=FLOOR((I26/I18+G26/G18)^2/((I26/I18)
 ^2/(I18-1)+(G26/G18)^2/(G18-1)),1)
```

2. Find the t statistic

On the Summary Page, this is **t = 8.92**

$$t_{a,b} = \left| \frac{\bar{x}_{R_a} - \bar{x}_{R_b}}{\sqrt{\frac{s^2_{R_a}}{n_a} + \frac{s^2_{R_b}}{n_b}}} \right|$$

[18] `=ABS((G23-I23)/SQRT(G26/G18+I26/I18))`

3. Then calculate your p-value using the t statistic

`=TDIST(I30,I29,2)`

The TDIST function in Excel uses numerical computation to give a very good estimation of the correct p-value for your test.

`TDIST(x, deg_freedom, tails)`

where x is the t statistic, deg_freedom is v, and tails is 2 (for a two-tailed test). A quote from the Excel help file on this function:

> **TDIST**
> Returns the Percentage Points (probability) for the Student t-distribution where a numeric value (x) is a calculated value of t for which the Percentage Points are to be computed. The t-distribution is used in the hypothesis testing of small sample data sets. Use this function in place of a table of critical values for the t-distribution.

In our sample data set, the final p-value is 5.04E-08; a very significant value that says "Yes, these two Data set are very likely different.".

Thank you.

9. References

1. Grentzmann, G., Ingram, J.A., Kelly, P.J., Gesteland, R.F. and Atkins, J.F. (1998) Rna, 4, 479-486.
2. Harger, J.W. and Dinman, J.D. (2003) Rna, 9, 1019-1024.
3. Kupper, L.L. and Hafner, K.B. (1989) American Statistician, 43, 101-105.
4. Croarkin, C. and Tobias, P. NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, (2004).
5. Kendall, M.G., Stuart, A., Ord, J.K. and O'Hagan, A. (1994) Kendall's advanced theory of statistics. 6th ed. Halsted Press, New York.
6. Devore, J.L. (2000) Probability and statistics for engineering and the sciences. 5th ed. Duxbury, Pacific Grove, CA.
7. Filliben, J.J. (1975) Technometrics, 17, 111-117.

-- Jonathan Jacobs
-- jacobsjo@umd.edu