Objective Analysis of Student Data

R. Jones
Physical Sciences Department
Emporia State University
Emporia, Kansas
**Introduction**

“How good is my data?” How much data do I need to take?” These are common student questions and are often difficult to answer. In the present paper we outline methods for objectively deciding such matters.

**Linear Relationships**

In our typical first year physics course and advanced laboratory course many of the relationships we have students investigate are linear:

\[ F = ma \quad \text{plot} \quad F \ vs \ a \]

\[ F = kx \quad \text{plot} \quad F \ vs \ x \]

\[ PV = NkT \quad \text{plot} \quad P \ vs \ \frac{1}{V} \]

\[ Q = mL \quad \text{plot} \quad Q \ vs \ m \]

\[ f = \mu N \quad \text{plot} \quad f \ vs \ N \]

\[ V = IR \quad \text{plot} \quad V \ vs \ I \]

\[ R = \frac{\rho L}{A} \quad \text{plot} \quad R \ vs \ L \]

\[ I^2 R \Delta t = mc \Delta T \quad \text{plot} \quad \Delta T \ vs \ \Delta t \]

\[ n\lambda = \frac{dy}{L} \quad \text{plot} \quad y \ vs \ n \]

\[ \frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i} \quad \text{plot} \quad \frac{1}{d_o} \ vs \ \frac{1}{d_i} \]

\[ B = \frac{\mu_0 N}{a} I \left( \frac{4}{5} \right)^{\frac{3}{2}} \quad \text{plot} \quad B \ vs \ I \]

\[ qV = \frac{hc}{\lambda} \quad \text{plot} \quad V \ vs \ \frac{1}{\lambda} \]
In other cases where the equations are nonlinear we recast the relationship in linear form by suitable mathematical transformations:

\[ y = \frac{1}{2} gt^2 \]  
plot \hspace{1em} y \text{ vs } t^2

\[ F = mrw^2 \]  
plot \hspace{1em} F \text{ vs } w^2

\[ T = 4 \frac{L^2 f^2 \mu}{n^2} \]  
plot \hspace{1em} T \text{ vs } \frac{1}{n^2}

\[ \tau = 2\pi \sqrt{\frac{m}{k}} \]  
plot \hspace{1em} \tau \text{ vs } \sqrt{m}

\[ \tau^2 = \frac{4\pi^2}{g} L \]  
plot \hspace{1em} \tau^2 \text{ vs } L

\[ n_1 \sin \theta_1 = n_2 \sin \theta_2 \]  
plot \hspace{1em} \sin \theta_1 \text{ vs } \sin \theta_2

\[ \ln \frac{N}{N_o} = -\frac{x}{\lambda} \]  
plot \hspace{1em} \ln N \text{ vs } x

**The Pearson Correlation Coefficient**

Having cast everything in simple linear form we can investigate the quality of our data using the common Pearson “R” correlation coefficient:\(^1\)

\[
R = \frac{\sum_{i=1}^{N} X_i Y_i - \left( \frac{\sum_{i=1}^{N} X_i}{N} \right) \left( \frac{\sum_{i=1}^{N} Y_i}{N} \right)}{\sqrt{\sum_{i=1}^{N} X_i^2 - \left( \frac{\sum_{i=1}^{N} X_i}{N} \right)^2} \sqrt{\sum_{i=1}^{N} Y_i^2 - \left( \frac{\sum_{i=1}^{N} Y_i}{N} \right)^2}}
\]

where \( X \) is the independent variable, \( Y \) is the dependent variable, and \( N \) \( X_i, Y_i \) data points have been measured.
Many computer plotting software utilities provide the Pearson $R$ value (or, alternatively, $R^2$) along with a graph $X_i$ vs $Y_i$. MS Excel is one common example which we use. Pocket calculators (like the TI-83) can also compute $R$ as one of their statistical analysis functions.

Guilford\(^2\) says that $R \geq .9$ indicates “very high correlation,” a “very dependable relationship” and we usually adopt $R \geq .9$ as our definition of “good data.” Work in the social sciences might require a relaxation of this standard.

If one wishes to minimize the amount of math involved, it is possible to supply the students with a graph like figure 1. Data points have been chosen to give a correlation coefficient of $R = .9$ for this graph. Students can be instructed to hold their data plots alongside this model. Qualitatively, if their data looks as good or better than figure 1 then $R \geq .9$ and they have “good data.”

**How much data should I take?**

It is possible to use a similar method to decide when enough data has been taken.

Suppose students have collected the following data:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.1</td>
</tr>
<tr>
<td>2.1</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>5.1</td>
</tr>
<tr>
<td>3.1</td>
<td>7</td>
</tr>
</tbody>
</table>

A calculation of the correlation coefficient gives $R = .79$. If we are striving for $R \geq .9$, we are led to take more data, perhaps arriving at:
<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.1</td>
</tr>
<tr>
<td>2.1</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>5.1</td>
</tr>
<tr>
<td>3.1</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>1.1</td>
<td>2.5</td>
</tr>
</tbody>
</table>

The Pearson $R$ value is now $R = .92$ and “enough” data has now been taken. It is worth pointing out to students that you only know how much data to take after you have begun to take measurements, not before. Data taking and analysis should be interspersed. One should not try to separate data taking from data analysis. Of course in real life one might also have to stop when you run out of time or out of resources.
