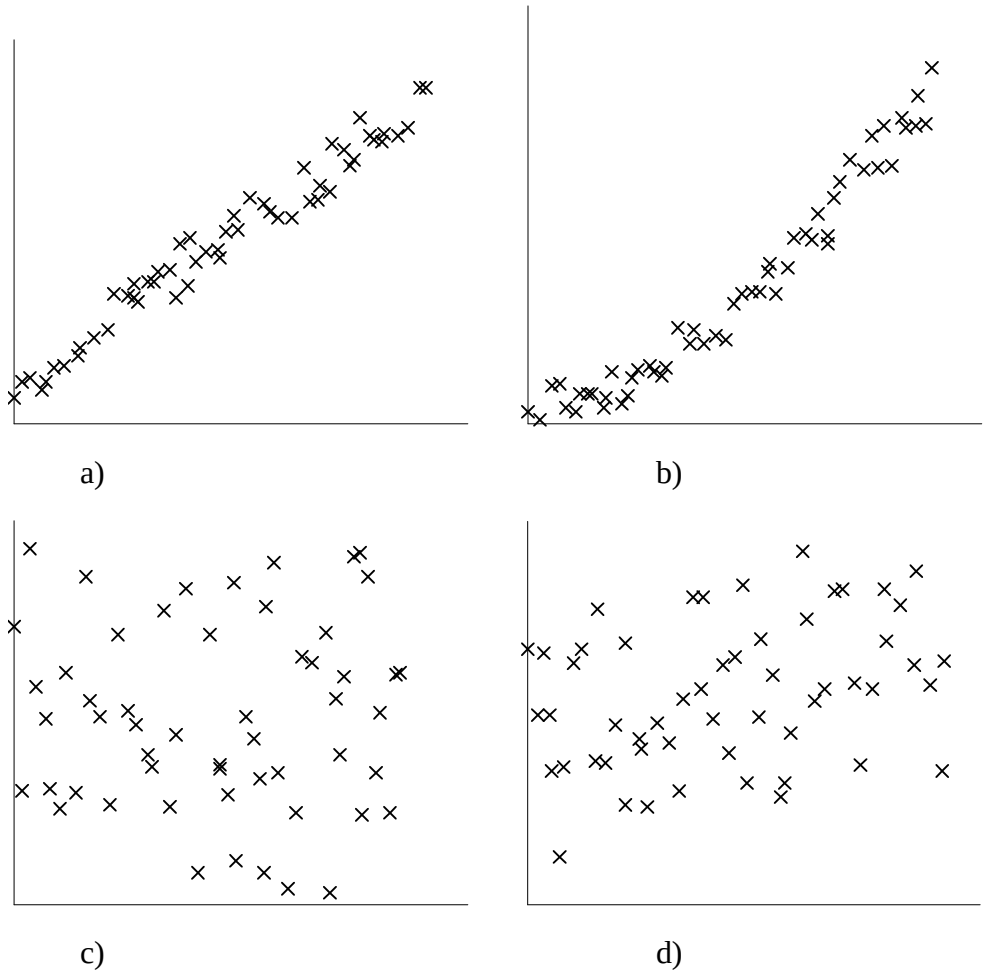
CORRELATION OF EXPERIMENTAL DATA

In this unit of the course we use statistical methods to look for trends in data. Most typically (for your engineering course) we conduct experiments where we have an independent variable, *x*, which we vary systematically. We then look for the resulting changes in the dependant variable, *y*. If we generate a scatter plot of *y* vs. *x* we could get a graph with many different looks. Most often we wish to identify whether there is a relationship between *x* and *y*, and if so, what that relationship might be.

Look at the following four graphs and try to assess whether there is a relationship between *x* and *y*. What, if anything, can you say about that relationship?



Clearly, if we got a graph like a) we would immediately conclude that there is a strong relationship between *x* and *y*, and that it is probably linear. For graph b) we would probably conclude that there is a nonlinear relationship between *x* and *y*.

For graph c) we might conclude that there does not seem to be any relationship between *x* and *y*.

Graph d) causes a problem. It is not clear whether there might be a weak relationship, or maybe there is none at all.

To help us decide, we use a *correlation coefficient*. You have probably seen this in your previous Excel coursework. The correlation coefficient gives a numerical measure of the functional relationship between two variables. It is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}}$$

where $\bar{x}$ and $\bar{y}$ are the average x and y values respectively. The resulting correlation coefficient is bounded in the range:

$$-1 \le r_{xy} \le +1$$

As a note, if you use Excel to do linear regression, the output is usually $(r_{xy})^2$, not r_{xy}. $(r_{xy})^2$ is called the *coefficient of determination.*

> $r_{xy} = +1$ indicates a perfect linear correlation between *x* and *y*, with a positive slope.
>
> $r_{xy} = -1$ indicates a perfect linear correlation between *x* and *y*, with a negative slope.
>
> $r_{xy} = 0$ indicates that there is no linear correlation between *x* and *y*.

Most typically r_{xy} will not be exactly +1, -1 or zero. Indeed, if you get any of these numbers it is highly suggestive that there has been some data manipulation (cheating!). So how do we interpret the numbers we get for the correlation coefficient?

The answer is that we compare our computed correlation coefficient with tabulated values. The tabulated values give the <u>maximum correlation coefficient values that might be expected by pure chance if there is no correlation between the variables.</u> The values depend on the number of samples and a chosen level of significance, α.

Table: Maximum correlation coefficient that might be expected by pure chance

	Level of Significance, α				
n	20%	10%	5%	2%	1%
3	0.951	0.988	0.997	1.000	1.000
4	0.800	0.900	0.950	0.980	0.990
5	0.687	0.805	0.878	0.934	0.959
6	0.608	0.729	0.811	0.882	0.917
7	0.551	0.669	0.754	0.833	0.875
8	0.507	0.621	0.707	0.789	0.834
9	0.472	0.582	0.666	0.750	0.798
10	0.443	0.549	0.632	0.715	0.765
11	0.419	0.521	0.602	0.685	0.735
12	0.398	0.497	0.576	0.658	0.708
13	0.380	0.476	0.553	0.634	0.684
14	0.365	0.458	0.532	0.612	0.661
15	0.351	0.441	0.514	0.592	0.641
16	0.338	0.426	0.497	0.574	0.623
17	0.327	0.412	0.482	0.558	0.606
18	0.317	0.400	0.468	0.543	0.590
19	0.308	0.389	0.456	0.529	0.575
20	0.299	0.378	0.444	0.516	0.561
25	0.265	0.337	0.396	0.462	0.505
30	0.241	0.306	0.361	0.423	0.463
35	0.222	0.283	0.334	0.392	0.430
40	0.207	0.264	0.312	0.367	0.403
45	0.195	0.248	0.294	0.346	0.380
50	0.184	0.235	0.279	0.328	0.361
100	0.129	0.166	0.197	0.233	0.257
200	0.091	0.116	0.138	0.163	0.180

EXAMPLE: We conduct an experiment and take 15 sample (x, y) pairs. The correlation coefficient between x and y was 0.500. At the 5% level of significance (95% degree of confidence) is there a linear relationship between x and y?

ANSWER: The tabulated value of correlation for 15 samples and 5% level of significance is 0.514. The experimental value is smaller than the tabulated value. We conclude that the experimental correlation could have resulted purely by chance and therefore there is no evidence of a linear relationship between x and y.

EXAMPLE: We conduct an experiment and take 15 sample (x, y) pairs. The correlation coefficient between x and y was 0.600. At the 5% level of significance (95% degree of confidence) is there a linear relationship between x and y?

ANSWER: The tabulated value of correlation for 15 samples and 5% level of significance is 0.514. The experimental value is greater than the tabulated value. We conclude that the experimental correlation is greater than that possible purely by chance. Therefore there is <u>no evidence that there is not</u> a linear relationship between *x* and *y*.

Note the double-negative in our conclusion
What does this mean?

This is an example of *Hypothesis Testing*[*]. Because of the nature of the correlation coefficient table, our hypothesis was that "the experimental correlation coefficient was the result of pure chance." In this example we disproved that hypothesis. That is not quite the same as saying we proved that a linear relationship exists! This very nature of hypothesis testing can lead to some interesting arguments:

"Is there a relationship?"

"I can't say that there isn't a relationship."

"So, is there a relationship?"

"I don't know. All I can say is that there is no evidence that there isn't a relationship!"

What about the common presumption of
Innocent until proven guilty?

In court the hypothesis is *guilty*, and the jury either agrees or disagrees with the hypothesis (*guilty* or *not guilty*). The court does not find anyone *innocent.* So hypothesis testing shows that O.J. was not guilty (court finding), but that does not mean he was innocent!

EXAMPLE: We conduct an experiment and take 20 sample (*x*, *y*) pairs. The correlation coefficient between *x* and *y* was 0.400. Is there a linear relationship between *x* and *y*?

ANSWER: The tabulated values of correlation for 20 sample pairs are:

0.299 (at 20%)
0.378 (at 10%)
0.444 (at 5%)
0.516 (at 2%)
0.561 (at 1%)

We therefore conclude that at 5% significance or less (i.e., 95% degree of confidence or more) the correlation coefficient we obtained for our data is less than that possible by pure chance. Therefore there is no evidence of a linear relationship between *x* and *y*.

[*] On this course we do not study hypothesis testing and the types of errors that can occur.

At the 10% level of significance (90% degree of confidence), the experimental correlation coefficient is greater than that possible by pure chance. Therefore there is no evidence that there is not a linear relationship between *x* and *y*.

CAUTIONS:

When you try to decide if a straight line is a good model for the data, you should not just use the numerical values of correlation coefficient, r_{xy} . You must also look at the graph. This is because the correlation function analysis fails under some circumstances. For example:

> Parabolic and other polynomial function data can generate large correlation coefficients even though the functions are not linear.
>
> Functions such as sine and cosine can produce very low values of correlation coefficient, even though there is a strong relationship (albeit not a linear one) between *x* and *y*.
>
> Just because two variables are correlated, and a relationship exists between them, it does not mean the relationship is causal. Wrongly assuming causality can lead to serious prediction errors.
>
>> If you increase the load on a beam, it deflects more. This is a causal relationship because the increased load caused the beam to deflect more. The opposite is not causal – an increased deflection does not cause an increased load.
>>
>> Plotting vehicular wind noise against engine vibration will show an apparent trend. But this is not a causal relationship. The engine vibration is not causing the wind noise, and the wind noise is not causing the engine vibration. In this case they are both being caused by (unmeasured) external effects – vehicle and engine speeds.
>>
>> If you wrongly assumed causality, you might consider changing the aerodynamic properties of the vehicle in an attempt to reduce engine vibration!

PROBLEMS

Correlation problems are assigned in the Regression handout. There are no correlation-only problems.