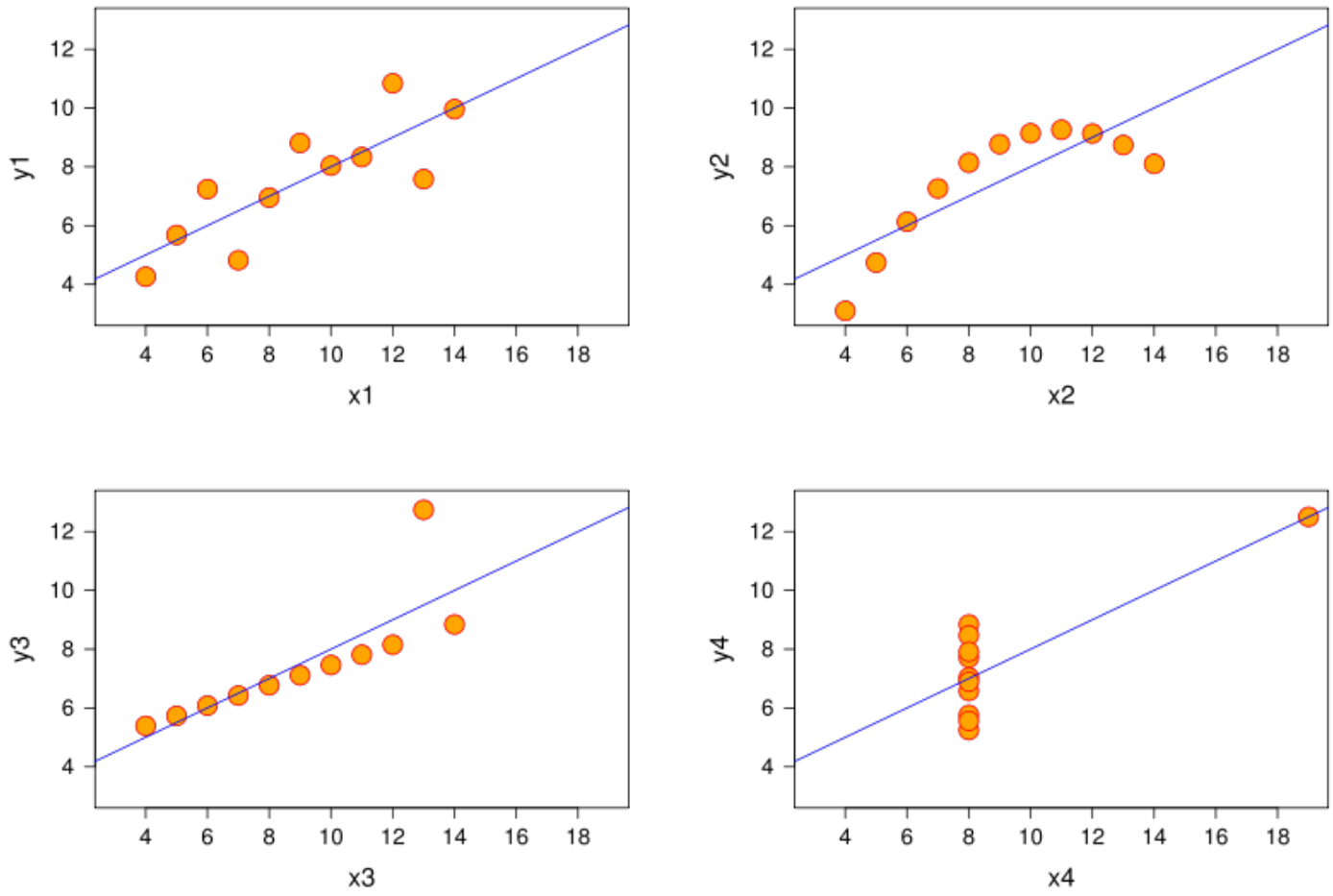
Pearson’s Correlation

Correlation – the degree to which two variables are associated (co-vary).

- Covariance may be either positive or negative.
- Its magnitude depends on the units of measurement.
- Assumes the data are from a bivariate normal population.
- Does not *necessarily* imply causation.

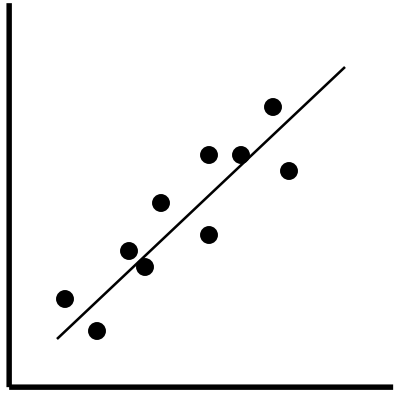


The four *y* variables have the same mean (7.5), standard deviation (4.12), correlation (0.81) and regression line ($y = 3 + 0.5x$).

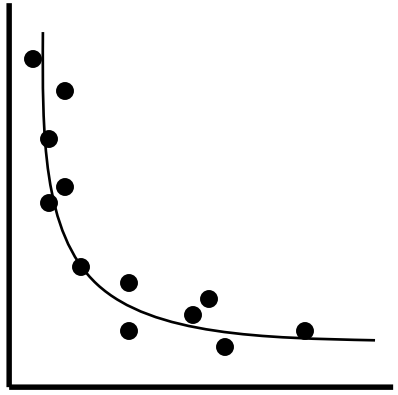
Pearson's correlation coefficient is a measure of the intensity of the *linear* association between variables.

- It is possible to have *non-linear* associations.
- Need to examine data closely to determine if any association exhibits linearity.

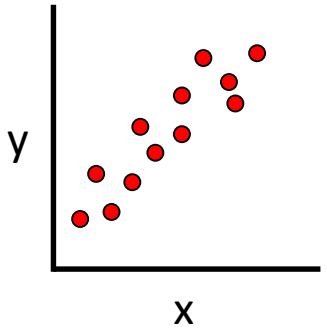
Linear



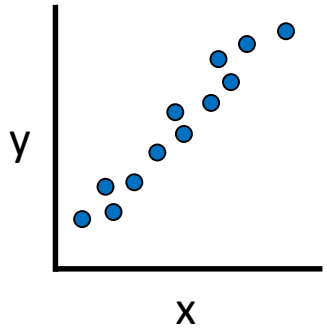
Non-linear



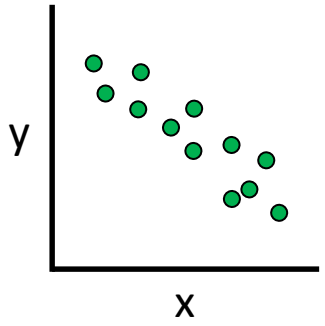
Correlation coefficient values range -1 to +1. The closer to 1 the correlation coefficient gets the ‘stronger’ the correlation.



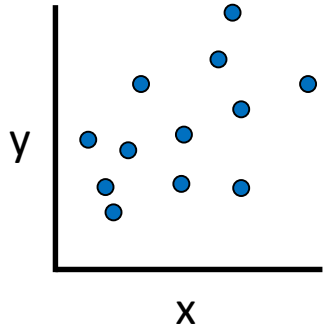
Positive correlation



Strong correlation



Negative correlation



Weak correlation

The *Pearson's Correlation Coefficient*.

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \qquad where \qquad \sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$\sum xy = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

The correlation coefficient is a measure of the *intensity* of the association between variables.

- r is a unit-less number.
- It can not be used to extrapolate a change in y based on a change in x.
- If variables are highly correlated, then we may want to investigate their association further to determine if there is a causal mechanism operating.

1 versus 2-tailed hypotheses

- 2-tailed hypotheses concerning r would state that there is a significant correlation between two variables.
    - e.g. $H_o: r = 0, H_a: r \neq 0$
- 1-tailed hypotheses concerning r would state that the association is either positive or negative.
    - e.g. $H_o: r \leq 0, H_a: r > 0$

Significance Testing for *r*

If the data are normally distributed we can calculate a t-statistic for the correlation coefficient (r) using the equation:

$$t = \frac{r}{s_r} \quad where \quad s_r = \sqrt{\frac{1-(r)^2}{n-2}}$$

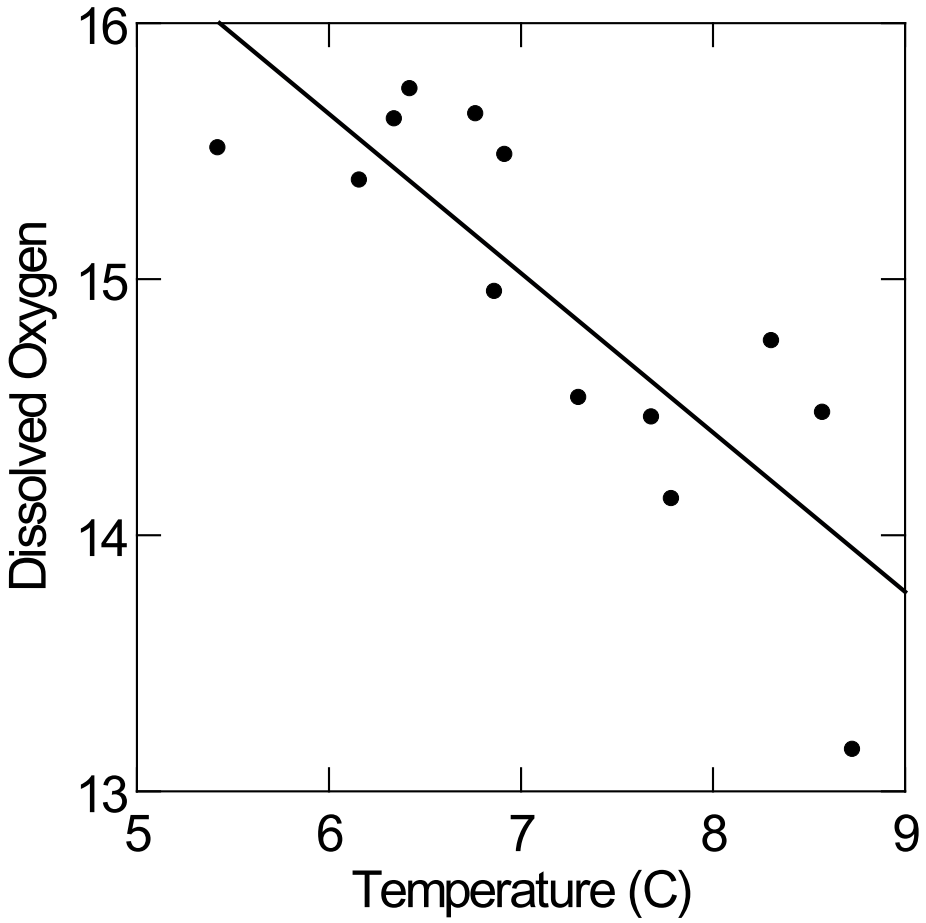
df = n-2… since there is one *df* for each column.

Here we are testing the null hypothesis that *r* = 0.

Temperature and Dissolved Oxygen

Temp (*X*)	DO (*Y*)
5.419	15.515
6.156	15.389
6.338	15.628
6.419	15.746
6.762	15.648
6.860	14.954
6.913	15.489
7.298	14.540
7.677	14.464
7.781	14.145
8.302	14.762
8.568	14.482
8.724	13.166

Temperature and Dissolved Oxygen



We will perform a 1-tailed test since our graph suggests that there may be a significant negative (or inverse) association between temperature and dissolved oxygen.

Ho: There is not a significant negative correlation between temperature and dissolved oxygen.

Ha: There is a significant negative correlation between temperature and dissolved oxygen.

Temp (X)	DO (Y)	X^2	Y^2	X*Y
5.419	15.515	29.37	240.72	84.08
6.156	15.389	37.90	236.82	94.73
6.338	15.628	40.17	244.23	99.05
6.419	15.746	41.20	247.94	101.07
6.762	15.648	45.72	244.86	105.81
6.860	14.954	47.06	223.62	102.58
6.913	15.489	47.79	239.91	107.08
7.298	14.540	53.26	211.41	106.11
7.677	14.464	58.94	209.21	111.04
7.781	14.145	60.54	200.08	110.06
8.302	14.762	68.92	217.92	122.55
8.568	14.482	73.41	209.73	124.08
8.724	13.166	76.11	173.34	114.86
$\sum$ 93.22	193.93	680.39	2899.79	1383.12

$n = 13 \qquad df = n - 2, 13 - 2 = 11$

$\sum X = 93.22 \qquad \sum Y = 193.93 \ (sum\ of\ the\ original\ data)$

$\sum X^2 = 680.39 \qquad \sum Y^2 = 2899.79 \ (sum\ of\ the\ squared\ observations)$

$\sum XY = 1383.12 \ (X \times Y\ then\ sum)$

$$\sum x^2 = 680.39 - \frac{(93.22)^2}{13} = 11.93 \qquad \sum y^2 = 2899.79 - \frac{(193.93)^2}{13} = 6.8$$

$$\sum xy = 1383.12 - \frac{(93.22)(193.93)}{13} = -7.51$$

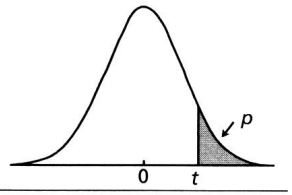
$$r = \frac{-7.51}{\sqrt{(11.93)(6.8)}} = -0.83$$

$$s_r = \sqrt{\frac{1 - (-0.83^2)}{13 - 2}} = \sqrt{\frac{0.311}{11}} = 0.168 \qquad t = \frac{-0.83}{0.168} = -4.94 \ (ignore\ the\ sign)$$

$t_{Critical}$ = 1.796 *4.94 > 1.796 reject Ho*

Table A.3 **Student's *t* distribution**

For various degrees of freedom (df), the tabled entries represent the critical values of t above which a specified proportion p of the t distribution falls. (Example: for df = 9, a t of 2.262 is surpassed by .025 or 2.5% of the total distribution.



	p (one-tailed probabilities)				
df	.10	.05	.025	.01	.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Adapted from Table III of Fisher and Yates (1974).

Since the value of *r* can be either positive or negative the critical value is a range… in this case:

-1.796 to +1.796

Our t-statistic (-4.94) falls beyond that range… so we reject H_0.

There is a significant inverse (negative) correlation between temperature and dissolved oxygen ($t_{-4.94,}$ $p < 0.005$, $r=-0.83$).

Since the range of the correlation coefficient is from -1 to +1, what does an r value of -0.83 tell us?

1. The association is *inverse*… meaning as one variable increases the other decreases.

2. The intensity of this inverse association between temperature and dissolved oxygen is fairly high.

Be aware that *r* is influenced by samples size.

Table 5.2 **Minimum values of *r* required for significance**

Sample size, *n*	Minimum absolute value of *r* needed to attain significance (using $\alpha = 0.05$)
15	.514
20	.444
30	.361
50	.279
100	.197
250	.124

For large *n*, r_{crit} is approximately $2/\sqrt{n}$.

Therefore if the sample size is large, even smaller *r* values may be important.

Do by hand on the board:

State	Black Lung Rate	Underground Miners
Kentucky	6.03	12947
West Virginia	17.37	14329
Ohio	2.07	1759
Utah	3.09	1922
Oklahoma	0.41	36
Tennessee	1.38	390
Alabama	1.92	4014
Virginia	3.87	5101
Pennsylvania	16.41	6202
Colorado	1.95	923

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \quad where$$

$$\sum x^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n}$$

$$\sum y^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

$$\sum xy = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}$$

$$t = \frac{r}{s_r} \quad where \quad s_r = \sqrt{\frac{1-(r)^2}{n-2}}$$

SPSS data set s:\GEO\pgmarr\Quantitative Methods\SPSS Data\BlackLung.sav

Correlations

		BlackLung	Underground
BlackLung	Pearson Correlation	1	.729*
	Sig. (2-tailed)		.017
	N	10	10
Underground	Pearson Correlation	.729*	1
	Sig. (2-tailed)	.017	
	N	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

Correlation vs Regression

Correlation measures of association, but no causal relationship is implied.

- There are no dependent or independent variables.

Regression measures association where a causal relationship is believed to exist.

- A dependent and 1+ independent variables are assumed.

Both correlation and regression assume that the relationship under investigation is linear, but it may be either positive (direct) or negative (inverse).

Boas Native American Tribe
Anthropometric Measurements

Correlations

		Male Arm Length	Male Leg Length
Male Arm Length	Pearson Correlation	1	.642**
	Sig. (1-tailed)		.000
	N	97	97
Male Leg Length	Pearson Correlation	.642**	1
	Sig. (1-tailed)	.000	
	N	97	97

**. Correlation is significant at the 0.01 level (1-tailed).

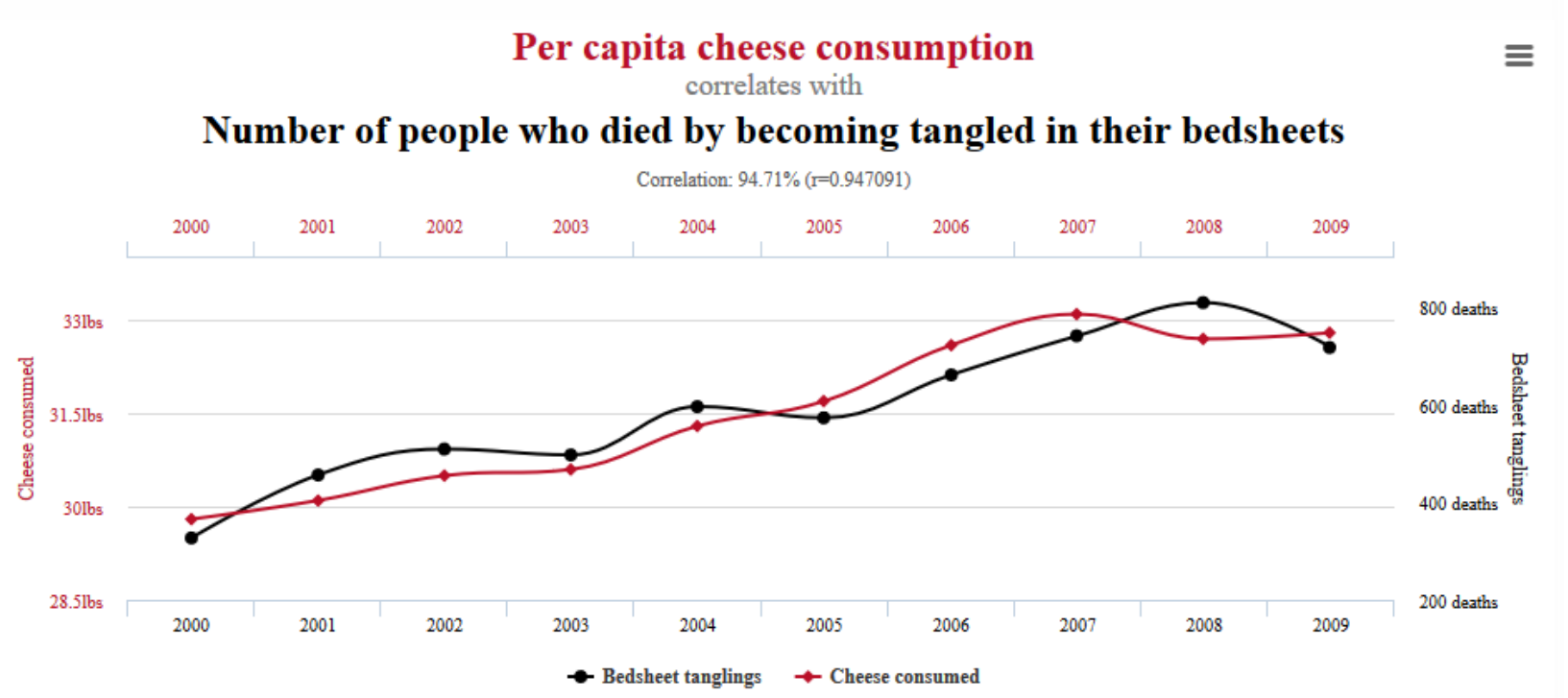
Note that arm and leg lengths are significantly correlated. However, longer arms *do not cause* longer legs. The causal mechanism is body proportions, meaning that larger individuals tend to have both longer arms *and* legs.

Remember that causation will result in correlation, but that correlation does not necessarily result in causation.

Therefore, correlation is a necessary but not sufficient condition to make causal inferences regarding our data.

Causation can really only be determined through controlled data analysis and a firm understanding of the underlying mechanisms which may result in a causal relationship.

There are situations where correlation is simply by chance, as seen below:



Source: https://www.tylervigen.com/spurious-correlations

There is no causal link between strangulation by bedsheets and cheese consumption but it has a high r value (0.947).

When performing correlation analysis:

- Test each variable for normality.
- Examine you data carefully.
- Formulate why you think the variable *should or should not* be correlated before your analysis.
- Remember that sample size influences r.
  e.g. Small r values are important in large samples.
- Remember that correlation does not equal causation.