



Marcello Pagano

# [JOTTER 9 CORRELATIONS AND NON-PARAMETRIC TESTS]

Pearson's Correlation Coefficient, Spearman's correlation coefficient, sign test, Wilcoxon signed rank test, Wilcoxon rank sum test.

## Relationships between variates



The **odds ratio** is a means of quantifying a relationship between two dichotomous variates:

e.g. exposure (yes,no)  
and disease (yes,no)

We now continue our exploration of the relationship between two variables. Today we look at the correlation coefficient to attempt to quantify the relationship between two continuous variables, but before doing that let us briefly review what we did to measure the relationship between two dichotomous variables. To quantify that relationship we looked at the odds ratio.

<div> <div>Key</div> <div> <i>frequency</i>  <i>row percentage</i>  <i>column percentage</i> </div> </div>				OR=2.11	
What is your sex?	When you wash your hair in the shower, do you...				
	Face away	Face the	Total		
Female	2,378 60.91 55.87	1,526 39.09 37.46	3,904 100.00 46.87		
Male	1,878 42.43 44.13	2,548 57.57 62.54	4,426 100.00 53.13		
Total	4,256 51.09 100.00	4,074 48.91 100.00	8,330 100.00 100.00		
Pearson chi2(1) = 283.5208 Pr = 0.000					



Let us digress a little and look at an example of a study of the relationship between dichotomous variables. These data come from the survey you were asked to fill in if you were one of the early enrollers in this course. You guys were asked about rinsing your hair in the shower in the morning and whether you faced the showerhead or do you turn around and rinse your hair from the back of the head?

Here is how 8,330 of you responded: Of the female responders, 60% answered that they faced away, and 40% said that they faced the shower head. On the other hand, the males, responded

almost the opposite: 42% faced away, and 58% faced the shower head. I do not know why, but every time I ask this question in a class, it comes out this way. It is a 60/40 split, one way or the other.

A number of people have suggested that this behavior is perfectly predictable because it has only to do with hair length; with longer hair it is easier or preferable to rinse facing away from the shower head, and women tend to have longer hair than men. Puzzle solved.

To empirically check this theory, we also asked you to tell us whether you considered yourself to have long hair or not.

Key		OR=0.67	
frequency		1/OR = 1.48	
row percentage			
column percentage			
Do you consider your hair to be long?	When you wash your hair in the shower, do you...		Total
	Face away	Face the	
No	2,649	2,894	5,543
	47.79	52.21	100.00
	62.02	70.78	66.30
Yes	1,622	1,195	2,817
	57.58	42.42	100.00
	37.98	29.22	33.70
Total	4,271	4,089	8,360
	51.09	48.91	100.00
	100.00	100.00	100.00
Pearson chi2(1) = 71.6252 Pr = 0.000			

And you responded as shown in the above table. Of those who did not consider themselves to have long hair, 52% faced the shower and 48% did not. Of those who considered themselves to have long hair, 49% faced the shower, 51% did not. So not quite as sharp a distinction as with sex—indeed, the odds ratio is 1.48 this time, instead of the 2.11 odds ratio above, when the dichotomy was based on sex—but still significantly different from one, with a reported p-value of 0.000.

So now we are confused, is it sex-related or hair-length-related?

-> Doyouconsideryourhairtobe = Yes

OR=2.4

Key
<i>frequency</i>
<i>row percentage</i>
<i>column percentage</i>

What is your sex?	When you wash your hair in the shower, do you...		Total
	Face away	Face the	
Female	1,418 61.41 87.86	891 38.59 75.13	2,309 100.00 82.46
Male	196 39.92 12.14	295 60.08 24.87	491 100.00 17.54
Total	1,614 57.64 100.00	1,186 42.36 100.00	2,800 100.00

Pearson chi2(1) = 76.6095 Pr = 0.000

-> Doyouconsideryourhairtobe = No

OR=2.0

Key
<i>frequency</i>
<i>row percentage</i>
<i>column percentage</i>

What is your sex?	When you wash your hair in the shower, do you...		Total
	Face away	Face the	
Female	958 60.14 36.32	635 39.86 22.03	1,593 100.00 28.85
Male	1,680 42.77 63.68	2,248 57.23 77.97	3,928 100.00 71.15
Total	2,638 47.78 100.00	2,883 52.22 100.00	5,521 100.00

Pearson chi2(1) = 137.0243 Pr = 0.000

We are, of course, concerned with the Yule effect (or Simpson's paradox), so let us investigate this a little deeper. Let us first look at the people who considered their hair to be long.

There were 2,800 of you who did, and when we looked at the sex-related odds ratio we find it is 2.4. (61% of females faced away and 40% males faced away.)

Amongst the 5,521 of you who considered that you did not have long hair, the sex-related odds ratio was 2.0. (60% of females faced away and 43% males faced away.)

Neither of these is too far from the 2.11 odds ratio when length of hair was ignored, above, and both odds ratios associated with a reported p-value of 0.000 for the null hypothesis that the odds ratio is one. So the sex difference maintains in both groups.

And so the same sort of sex-related relationship maintains amongst those who consider their hair be long, as in the group who did not consider their hair to be long. So hair length does not seem to matter for this relationship.

-> Whatisyoursex = Female

Key

frequency

row percentage

column percentage

Do you consider your hair to be long?	When you wash your hair in the shower, do you...		Total
	Face away	Face the	
No	958 60.14 40.32	635 39.86 41.61	1,593 100.00 40.83
Yes	1,418 61.41 59.68	891 38.59 58.39	2,309 100.00 59.17
Total	2,376 60.89 100.00	1,526 39.11 100.00	3,902 100.00 100.00

Pearson chi2(1) = 0.6422 Pr = 0.423

OR=0.95

-> Whatisyoursex = Male

Key

frequency

row percentage

column percentage

Do you consider your hair to be long?	When you wash your hair in the shower, do you...		Total
	Face away	Face the	
No	1,680 42.77 89.55	2,248 57.23 88.40	3,928 100.00 88.89
Yes	196 39.92 10.45	295 60.08 11.60	491 100.00 11.11
Total	1,876 42.45 100.00	2,543 57.55 100.00	4,419 100.00 100.00

Pearson chi2(1) = 1.4524 Pr = 0.228

OR=1.1

One last analysis in order to convince ourselves: let us look within each sex. First with females we see that the hair-length related odds ratio is 0.95, and with males the hair-length related odds ratio is 1.1. In neither case would we reject the null hypothesis that the odds ratio is one. This should convince us that it is not hair-length related, once we know the sex of the responder. So we are still confounded by the reason for this sex-related phenomenon of whether we face the shower head or not when rinsing our hair.

One last aside from this last slide: you may hear skeptics complain that if your sample size is large enough you can always find significance. Here we see two very large samples, 3,902 in the one sample and 4,419 in the other and in neither case did we reject the null hypothesis that the odds ratio is one.

## Relationships between variates



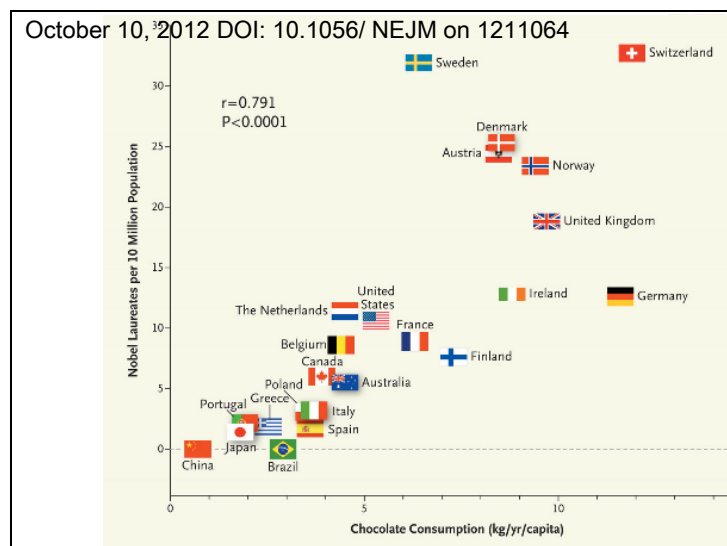
The **odds ratio** is a means of quantifying a relationship between two dichotomous variates:

e.g. exposure (yes,no)  
and disease (yes,no)

What if the two variates are  
not dichotomous? – Generalize.  
Quantify

Let us now turn our attention to quantifying the relationship between two variables when they are not both dichotomous. Let us first look at when both variables are continuous.

Pearson's Correlation Coefficient

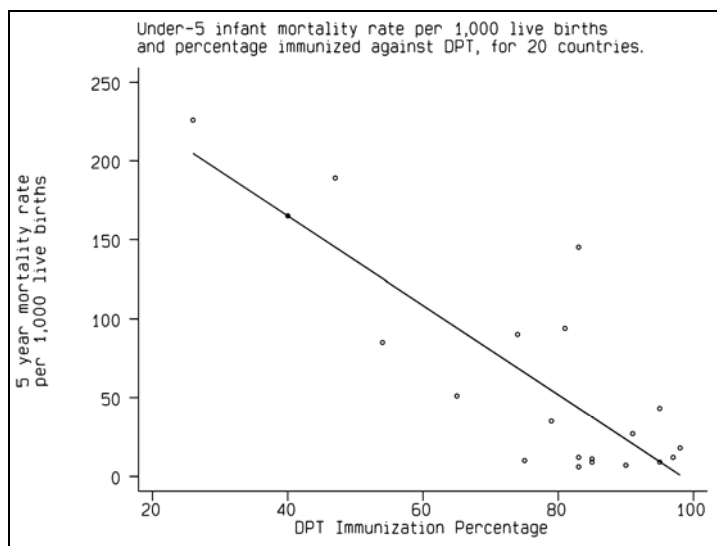


This graph just appeared in the New England Journal. On the bottom axis we have the per capita chocolate consumption—kilograms per year. On the vertical scale we have the number of Nobel laureates per 10 million inhabitants in a country. And what we see is a quasi-linear relationship between chocolate consumption and the number of Nobel laureates per country.

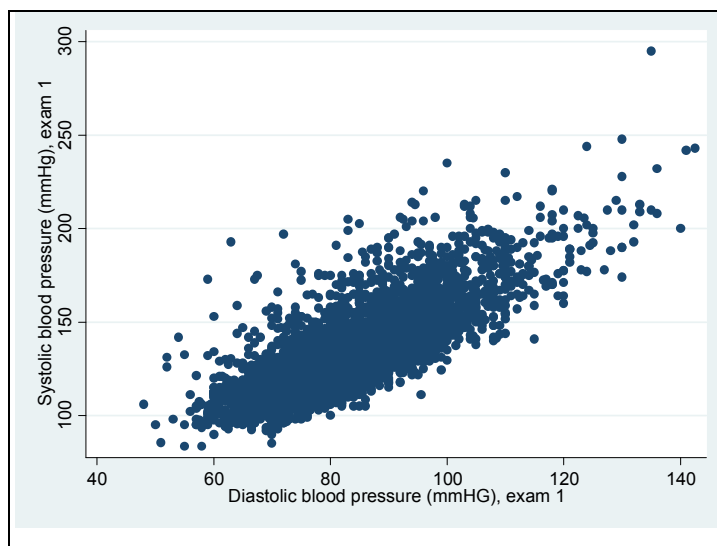
This is your classical correlation analysis. The author found a very high correlation, and deduced, tongue in check, no doubt, that chocolate has some impact on your brain cells, to explain the correlation.

Now, there is a problem with this study. Not to be a wet blanket, but if you want to be serious about this, this study suffers from what is called the ecological fallacy. We return to this issue before this week is out.

<sup>1</sup> Occasional Notes: **Chocolate Consumption, Cognitive Function, and Nobel Laureates**, Franz H. Messerli, M.D. October 10, 2012 DOI: 10.1056/NEJMon1211064



Here is another example. Once again we are dealing with countries. And here we have the five year mortality rate per 1,000 live births on the vertical axis. That is an indicator often used to monitor the health level of a country. On the horizontal we have the DPT (diphtheria, pertussis and tetanus) immunization percentage in a country. This too is a country level graph, and whatever conclusions we can draw from this should be at the country level.

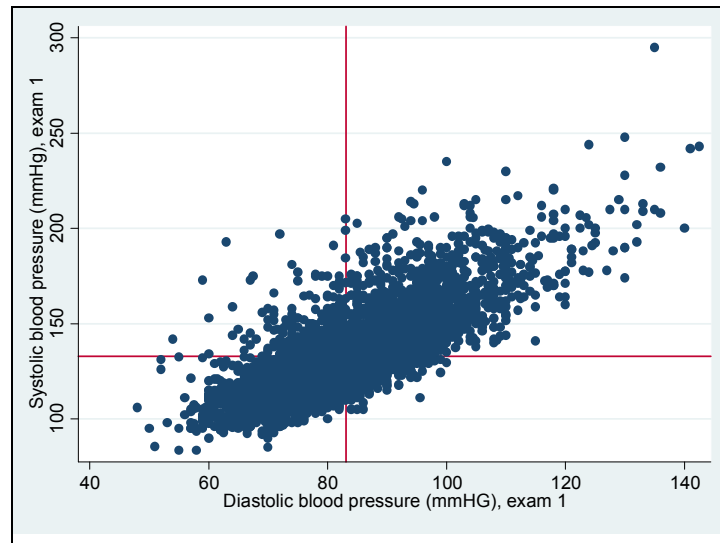


At the personal level, here is a scatter plot of the systolic and diastolic blood pressures from the first visit in the Framingham heart study. We have seen this graph and the elliptical relationship

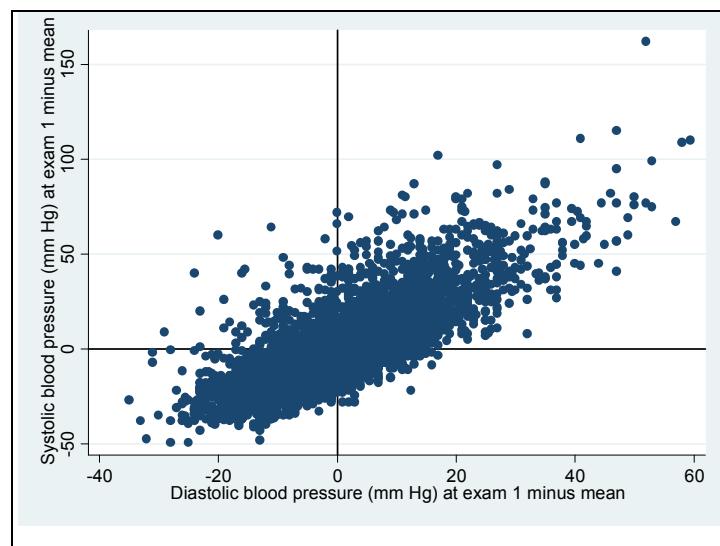


before. We see that as the diastolic blood pressure goes up, the systolic blood pressure goes up, and vice versa. They vary together. They co-vary. They both vary. But they co-vary. They both go off in the same direction, hand in hand.

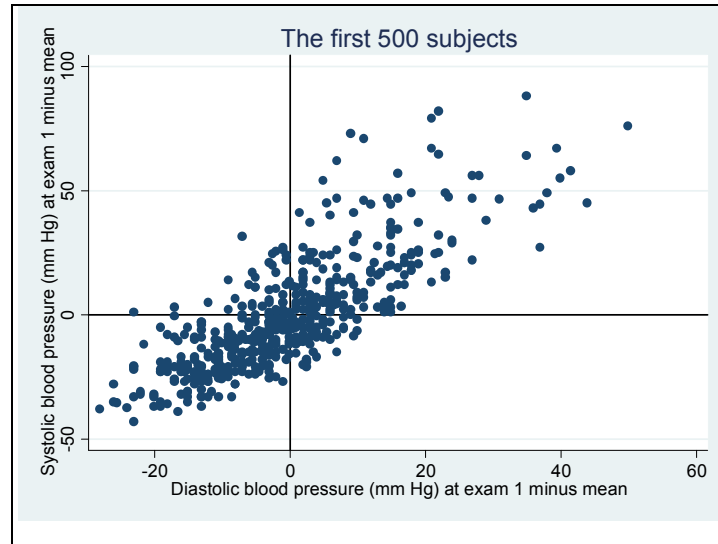
Can we quantify this behavior?



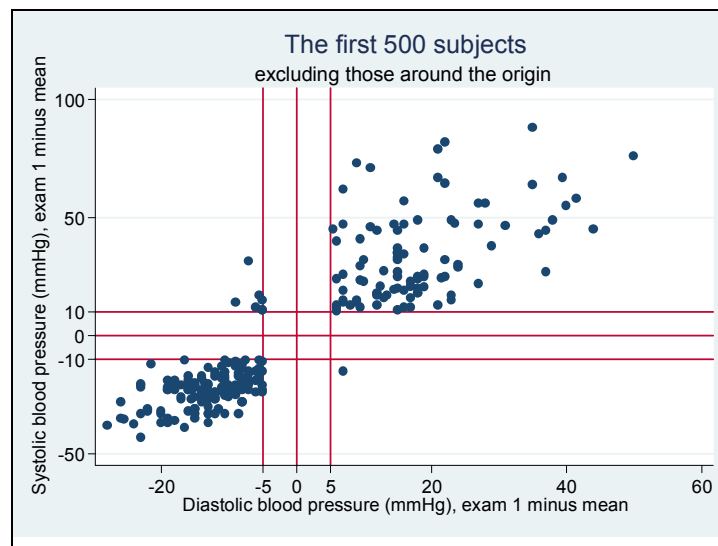
To this end, consider centering the graph by introducing axes at the means of the two. So let me subtract the mean of the systolic pressures from the systolic readings and the mean of the diastolic pressures from the diastolic readings.



After relabeling the axes we get this picture. Since there are too many dots on this graph let us just consider the first 500 subjects.



This allows us to get a better picture of what is going on.



Let us further attempt to understand what is happening by excluding all the points with small, in absolute value, components to get the picture above. It shows four quadrants only, and those are points both of whose components are sizable. Concentrating on just these values we see that most of them fall into two quadrants where the signs of both the components are the same: large positive diastolic readings go with large positive systolic readings, and large negative diastolic readings go with large negative systolic readings. You should not erase points, of

course, but this is just temporarily and for expository purposes. The challenge remains of capturing this behavior in a single number.

```
list diabp1 mdiabp1 sysbp1 msysbp1 if _n<15
```

	diabp1	mdiabp1	sysbp1	msysbp1
1.	70	-13.08356	106	-26.9078
2.	81	-2.08356	121	-11.9078
3.	80	-3.08356	127.5	-5.4078
4.	95	11.91644	150	17.0922
5.	84	.91644	130	-2.9078
6.	110	26.91644	180	47.0922
7.	71	-12.08356	138	5.0922
8.	71	-12.08356	100	-32.9078
9.	89	5.91644	141.5	8.5922
10.	107	23.91644	162	29.0922
11.	76	-7.08356	133	.0922
12.	88	4.91644	131	-1.9078
13.	94	10.91644	142	9.0922
14.	88	4.91644	124	-8.9078

Mean

diabp1=83

sysbp1=133

Let us return to looking at these numbers in a table rather than on the graph. Here are the first 14 values. The columns labeled diabp1 and sysbp1 contain the original data. These are transformed by subtracting the respective means to get the columns mdiabp1 and msysbp1. Concentrating on these two columns, we see that large negative values are paired as are large positive values, by and large.

To capture this behavior we can think back to the variance.

	FEV <sub>1</sub>	$(x_j - \bar{x})$	$(x_j - \bar{x})^2$
$\bar{x} = 2.95$	2.30	-0.65	0.423
	2.15	-0.80	0.640
	3.50	0.55	0.303
	2.60	-0.35	0.123
	2.75	-0.20	0.040
	2.82	-0.13	0.169
	4.05	1.10	1.210
	2.25	-0.70	0.490
	2.68	-0.27	0.073
	3.00	0.05	0.003
	4.02	1.07	1.145
	2.85	-0.10	0.010
	3.38	0.43	0.185
	Total	0.00	4.66

## Variance



$$\text{Variance} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\geq 0$$

e.g.

$$= \frac{4.66}{12} = 0.39 \text{ liters}^2$$

With the variance we multiplied each deviation by itself and then found the average squared-deviation.

## Covariance



$$\text{Variance} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{Covariance} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Now we have two variables, so instead of squaring the one deviation and getting the variance, we can multiply the two deviations, average them out and get what is called the *covariance*. This tells us how much the x and the y, co-vary, or vary together.

### Pearson's Correlation Coefficient. Product moment correlation.



$n$  pairs:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

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

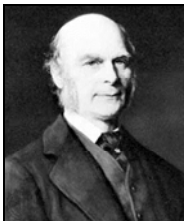
Note that  $-1 \leq r \leq 1$

0.7842 diastolic and systolic blood pressure for FHS at first visit


If we also standardize our variables by dividing the deviations by their respective standard deviations, then we get  $r$ , or what is called the sample *correlation coefficient*, or the *product moment correlation coefficient*, or *Pearson's Correlation Coefficient*.

One immediate result of this standardization is that, from the Cauchy inequality,  $-1 \leq r \leq 1$ .

In the example with diastolic and systolic blood pressure at visit one, we get that the correlation coefficient is 0.7842—very high and close to 1, but not quite 1.



## Correlation



Francis Galton  
1822-1911

Karl Pearson  
1857-1936

Measure of **linear** relationship between two continuous random variables.

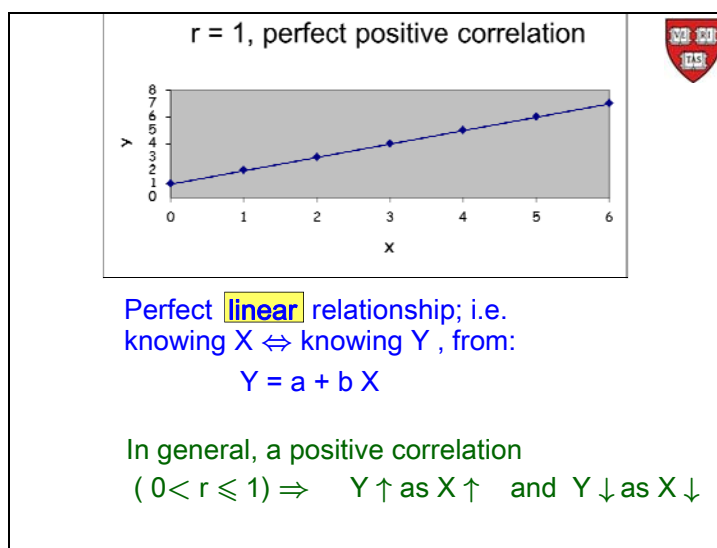
Correlation Coefficient

$$\rho = \text{average} \left[ \left( \frac{X - \mu_X}{\sigma_X} \right) \left( \frac{Y - \mu_Y}{\sigma_Y} \right) \right]$$

Pearson did the early mathematical work on correlation, but its introduction is due Francis Galton. Galton was Charles Darwin's cousin, and the story goes that he was a little bit jealous of his cousin's fame. Galton also invented the word eugenics, and left us with that perversion of his cousin's research.

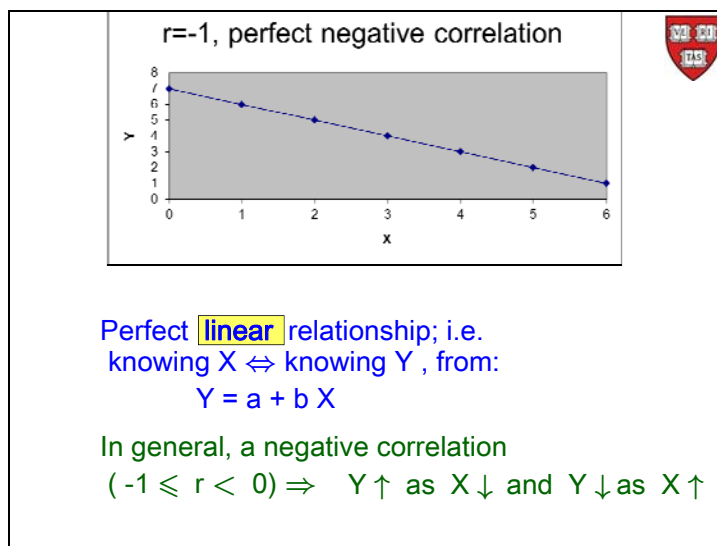
Take care that this coefficient is a measure of *linear* relationship. This word linear people sometimes ignore, but you do that at your own peril. As we discuss this coefficient, you should appreciate what that qualifier means.

Within the population, we follow our tradition of using Greek letters to label parameters. Here we use the letter  $\rho$  to denote the population parameter that we are attempting to estimate by using the  $r$  from a sample.



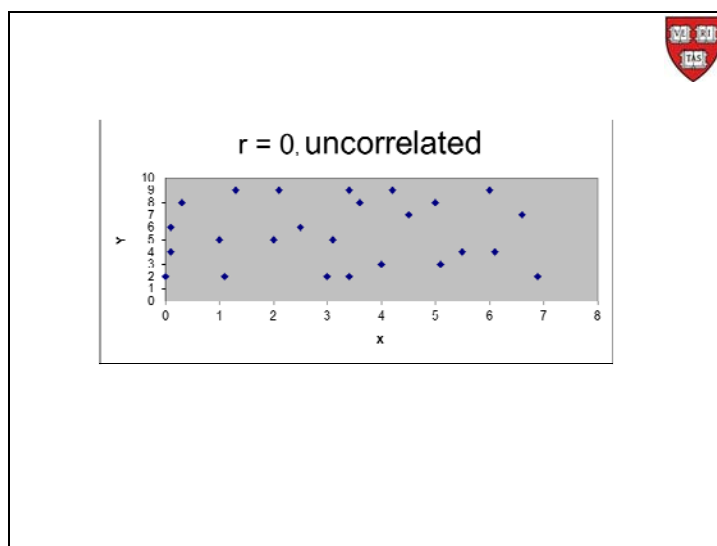
There are some special, extreme values of  $r$ , and  $\rho$ , that are worth noting. One is at the right end of the scale, at the value 1. Then we have perfect positive correlation, and that is the largest correlation one can have. What it actually means is that there is a straight line relating the two variables. They are thus basically a single random variable. The slope  $b$  is positive, so the variables increase, and decrease, together.

In general, a positive correlation means that the two variables tend to increase, and decrease, together. The relationship is only perfect at the extreme of one.

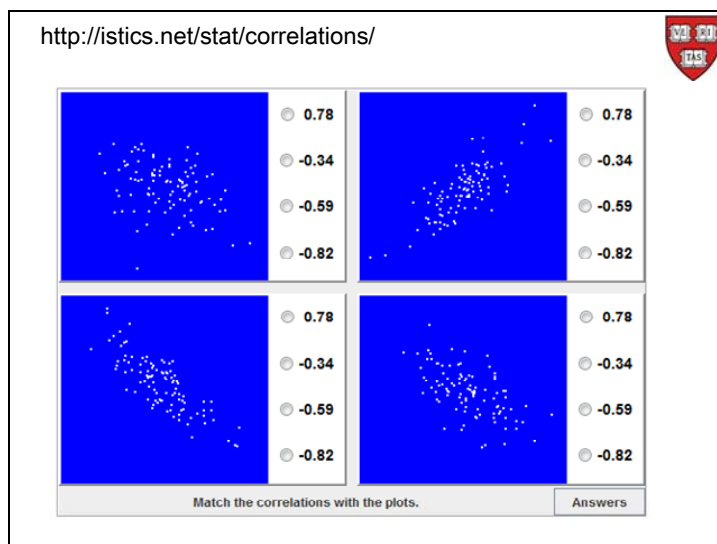


The other extreme is when  $r$ , and  $\rho$ , equal minus one. This happens when we have a perfect relationship, as with the situation above when the correlation was one, except that this time the variables go in opposite directions; so the slope  $b$  is negative, and thus one variable goes up as the other goes down

In general, when you do not have this perfect negative correlation, but the correlation coefficient is still negative, then on average one variable goes up as the other one comes down.



The other special value the correlation coefficient can take is right in the center, and that is when it is equal to zero. When that happens we say that the variables are *uncorrelated*. This might remind you of independence, but do not be confused. What this means is that, and here comes that important word again, there is no *linear* relationship between the variables.



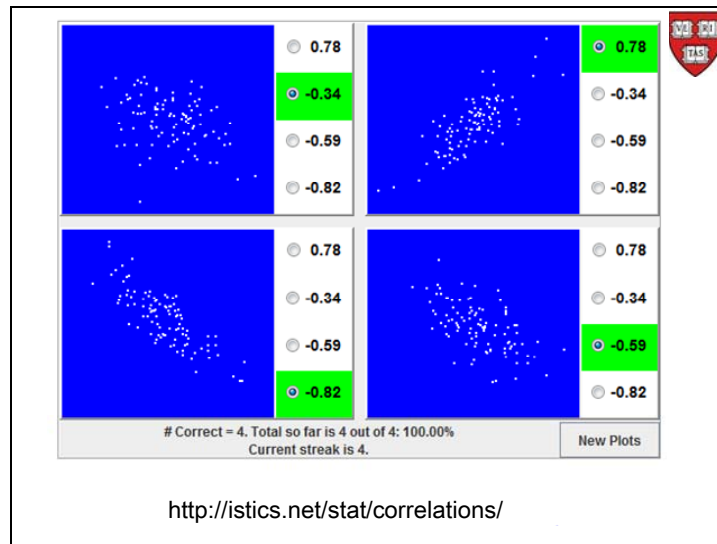
What I would like you to do is go to this website<sup>2</sup>. There you will find a game I find mesmerizing. They throw up four scatter plots of clouds of points. On the right hand columns of each plot you find four correlation coefficients, the same four at each plot. The game is to link a graph with a correlation coefficient, and you get scored on how many you get correct. That way you can get a feel for what the correlation coefficient is measuring.

For example, with this panel, we see that the top right-hand corner is sloping positively, so that one should get the 0.78 choice. The others get gradually more negative from -0.34 to -0.59 to -0.82. The trick in making the identification is to think back to the three graphs above. For the extremes at plus or minus one we had no variability around the line, whereas at a correlation of zero we had maximal variability. So grade these three according to the amount of variability in the scatters: the top left-hand corner probably has the maximum variability, so it should be identified with the -0.34. Of the bottom two, the one on the left looks tighter than the one on the right.

You make your choices, and then Answers.

<sup>2</sup> <http://istics.net/stat/correlations/> or <http://www.istics.net/Correlations/> if you prefer Java.





In this case I was correct on all four. So after all these years I was able to be right! You go head and have some fun and get some feel for how sample correlation coefficients vary.

### Inference on $\rho$

$$\rho = \text{average} \left( \left( \frac{X - \mu_x}{\sigma_x} \right) \left( \frac{Y - \mu_y}{\sigma_y} \right) \right)$$

estimated by


$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

Sampling distribution:  
If  $X$  &  $Y$  are normally distributed and  $\rho = 0$ , then

$$t_{n-2} = r / \sqrt{\frac{1-r^2}{n-2}}$$

We have a population parameter  $\rho$ , and we would like to make inference about this parameter on the basis of a sample from this population. Suppose we wish to test a hypothesis about  $\rho$ . The only hypothesis we cover in this course is the one that says that  $\rho = 0$ . So we test the hypothesis that two variables,  $X$  and  $Y$ , are uncorrelated.

So what we need is the sampling distribution of  $r$  when  $\rho$  is equal to 0. The slide gives us the sampling distribution if the two variables,  $X$  and  $Y$ , are normally distributed.



e.g.  $r = -0.829$  for DPT example

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

$$= -0.829 \sqrt{\frac{20-2}{1-(-0.829)^2}} = -6.29$$

versus  $t$  with 18 degrees of freedom, so  $p < 0.001$ .

So reject  $H_0 : \rho = 0$ .

In the DPT example, above, we have that  $r$  was  $-0.829$ , so our  $t = -6.29$  and the  $p$ -value is less than  $0.001$ , and so we would reject, at the 5% level, the null hypothesis.

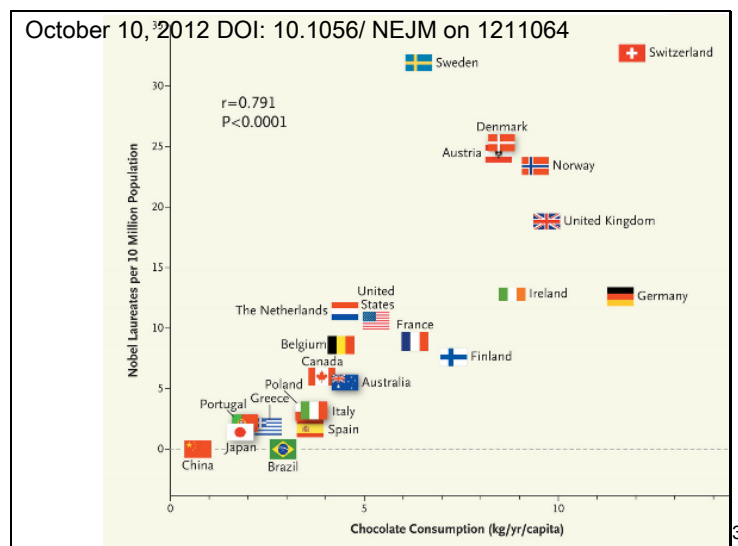
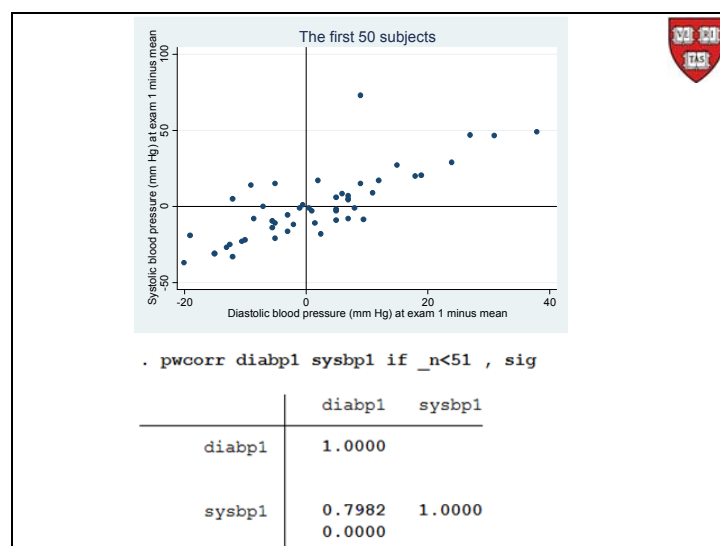



Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

In the top left-hand side of the graph we see that  $r=0.791$  and the p-value, associated with the null that  $p=0$ , is less than 0.0001. So we would reject that hypothesis.



<sup>3</sup> Franz H. Messerli, M.D. Occasional Notes, **Chocolate Consumption, Cognitive Function, and Nobel Laureates**, October 10, 2012 DOI: 10.1056/NEJMon1211064

Choosing the first 50 subjects in our Framingham heart study, just as an exercise, we test the hypothesis that  $\rho = 0$ , using the command `pwcorr`, and we get that the correlation is 0.7982 and the p-value is underneath it. It is 0.0000. So the p-value is less than 0.00005. So on the basis of these 50 observations (and this is not a proper study since this is not a random sample, just me exercising pedagogical license) we would reject the null hypothesis that diastolic and systolic blood pressure at visit one were uncorrelated.



Misconceptions:

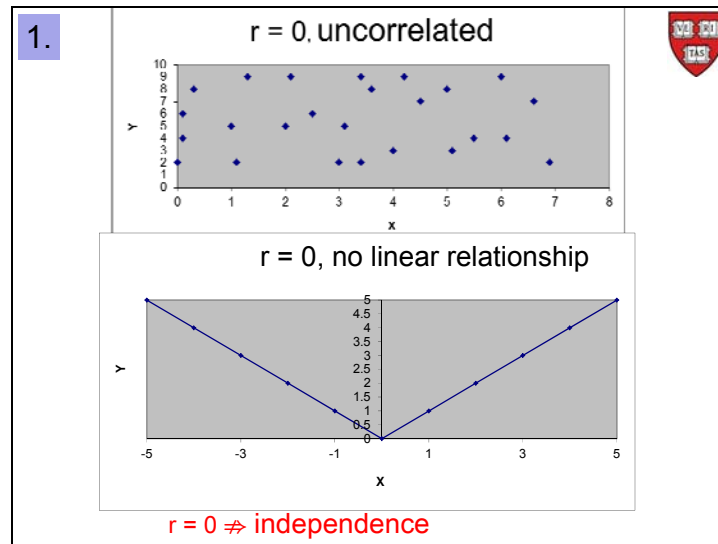
1. Correlation = 0 implies independence
2. Correlation implies causality
3. Ecological Fallacy

Now that you have seen the correlation coefficient, we look at three important misconceptions some people have about correlations.

The first misconception is that when the correlation equals zero, this implies independence. Correlation equals zero does not imply independence, the two variables are merely uncorrelated.

The second misconception is a very touchy one, and that is that correlation implies causality. It does not. It simply implies correlation. The two are not synonymous.

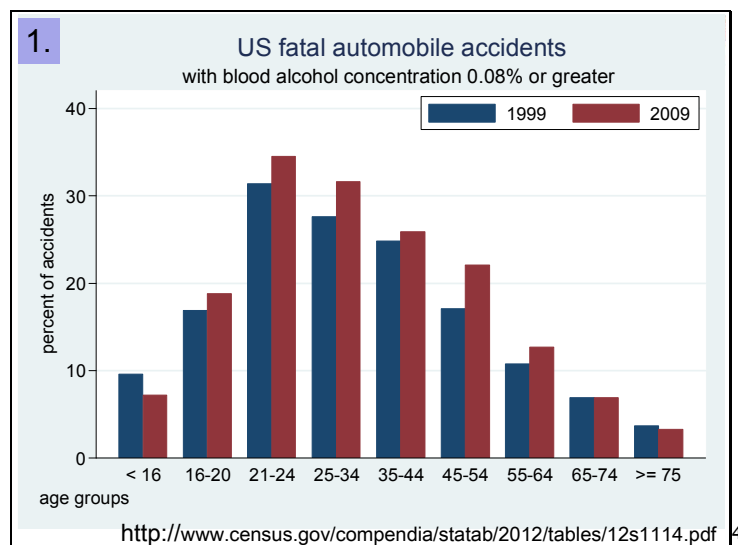
And the third misconception is that correlation at the ecological level implies correlation at the personal level—the ecological fallacy we have just seen. Let us take these misconceptions one by one.



You get an  $r=0$  from the top graph when the points in the scatter have no discernible pattern, but you also get zero correlation in the bottom graph. Although this one is reminiscent of the extreme correlations of plus and minus one, indeed the pattern it is a combination of those two patterns, you get an  $r$  that equals zero.

The perfect linear relationships on the positive side and on the negative side just sort of cancel each other out. You get this relationship when  $Y = |X|$ . It does not mean that  $X$  and  $Y$  are independent, of course, just uncorrelated.

This non-monotonic relationship, first down then up, can happen in nature, it is not just a mathematical artifact.



<sup>4</sup> <http://www.census.gov/compendia/statab/2012/tables/12s1114.pdf>

Here is an example of a U-shaped relationship (except that it is an upside-down U, but that does not affect the argument). The height of the bars represent the percent of all fatal automobile accidents in the US that involve blood level of alcohol of 0.08%, or above for the drivers. The blue bars refer to 1999 and the red bars to ten years later, namely 2009, and each pair of bars refer to a different age group.


In those 10 years we see a small improvement in the less than 16s, even though most of those do not have their driver's license so that is problematic. There has been a decrease in those over 75 and a flat spot in the 65 to 74. But with everybody else between the ages of 16 and 64 there was an increase in the percentage of fatal accidents that involved alcohol. So something is going wrong here.

That is not why I am showing you this. Of course, it is good if I give you the don't drink and drive, message, but the reason I am showing you this is that this looks very much like that U-shaped relationship we just saw, and here too the correlation coefficient is not very high. But the fact that there is a relationship between the age of the drivers and this outcome we are measuring, seems indisputable.

So what we are seeing is an example of the fact that a non-linear relationship is not measured well by the correlation coefficient. There are a number of situations like this where we just have to be very careful. So correlation zero does not imply independence.

2.

Short of  $\rho = \pm 1$ , a high correlation does not imply a cause & effect relationship.



The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning.

Stephen Jay Gould  
*The Mismeasure of Man*,  
 W.W.Norton & Co. 1981, p. 242

So moving on to the second misconception, here is what Stephen Jay Gould had to say about that.

---

2.



All causation as we have defined it is correlation, but the converse is not necessarily true, i.e. where we find correlation we cannot *always*\* predict causation. In a mixed African population of [black Africans] and Europeans, the former may be more subject to smallpox, yet it would be useless to assert darkness of skin (and not absence of vaccination) as a cause.

\* [stress added m.p.]

The Grammar of Science  
Karl Pearson  
London, Adam and Charles Black, 1900

Indeed, Pearson recognized this problem back in 1900 shortly after he introduced the correlation coefficient. So this direction has been well established.

2.

Oct 31 12:11 PM Harvard Crimson



### Correlation Still Doesn't Equal Causation in Soda Studies

The report links aspartame to increased risks of leukemia, lymphoma, and non-Hodgkin's lymphoma,.....Boston public high school students has shown that students who identified as heavy soda drinkers were more likely to engage in violent behavior.....

Neither study decisively proves the harmful effects of soda, so until more intensive studies are preformed, it looks like you're safe to enjoy a glass of your favorite soda without worrying too much about the possibility of either cancer risks or increased violence.

<http://www.thecrimson.com/article/2012/10/31/soda-studies-harvard/>

5

But just because correlation does not imply causality it does not mean that because there is correlation then there cannot be a causal. This might sound silly, but the tobacco companies

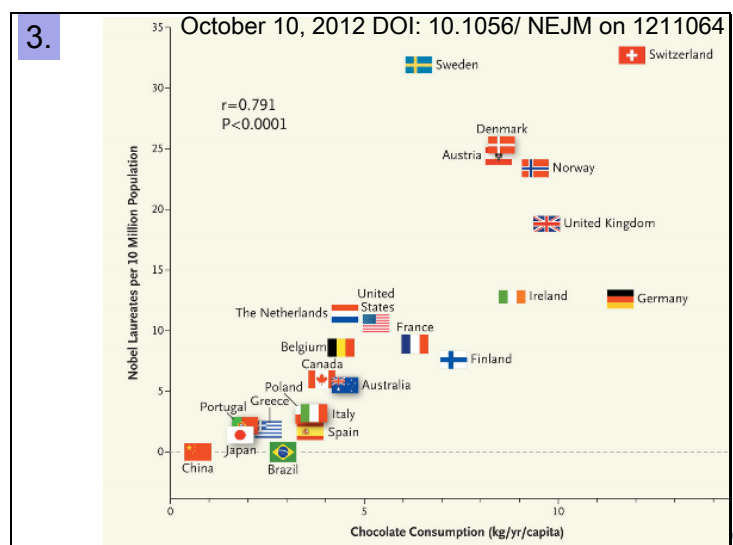
<sup>5</sup> <http://www.thecrimson.com/article/2012/10/31/soda-studies-harvard/>



kept singing this tune for some 75 years: you have only shown a correlation and correlation does not imply causality.

Here is an article that appeared in the Harvard Crimson. It was commenting on two studies that had come out of the Harvard School of Public Health dealing with the consumption of soft drinks. One of the studies showed a correlation between aspartame—an artificial sweetener used in the soft drinks—and increased risks of leukemia, lymphoma, and non-Hodgkin's lymphoma. The other study correlated high school students who were identified as heavy soda drinkers were also more likely to engage in violent behavior.

Now this undergraduate reporter had been rightly taught that correlation still does not equal causation, but she incorrectly went one step too far when she said, "It looks like you're safe to enjoy a glass of your favorite soda."




Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

The last misconception is the ecological fallacy. It is a mathematical property of the correlation coefficient that it increases when considering two groups as opposed to the correlation between the individuals in the group.

Case in point, above, when considering the correlation coefficient between countries' chocolate consumption and their Nobel prizes received. At that level of aggregation, the correlation turns

<sup>6</sup> Franz H. Messerli, M.D. Occasional Notes, **Chocolate Consumption, Cognitive Function, and Nobel Laureates** October 10, 2012 DOI: 10.1056/NEJMon1211064

out to be 0.791. Before you run out and drive up the price of chocolate, this calculation tells us nothing about the correlation coefficient when calculated at the individual level. That is the ecological fallacy; namely believing that it does.

3.


Ecological fallacy: Assuming that correlations measured at an aggregated level imply the same at an individual level.

Are people who drink “hard” water containing higher levels of calcium and/or magnesium less likely to suffer cardiovascular disease?

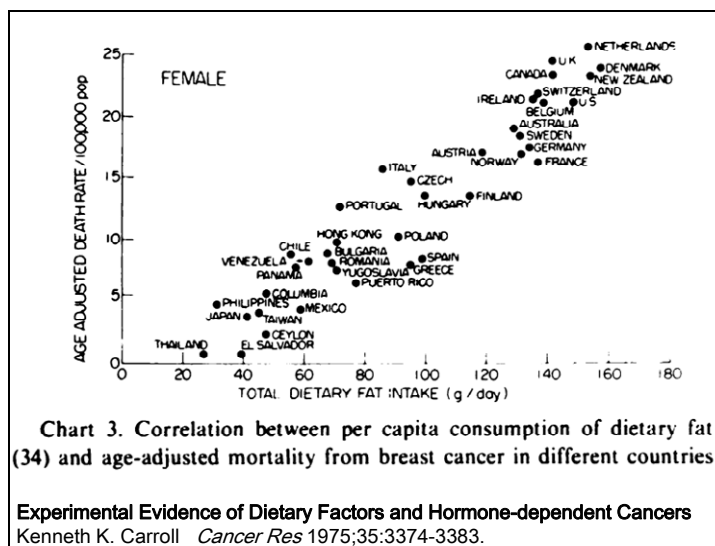
[http://www.who.int/water\\_sanitation\\_health/gdwqr/evasion/cardiofullreport.pdf](http://www.who.int/water_sanitation_health/gdwqr/evasion/cardiofullreport.pdf)

There is an ongoing debate that has lasted a number of years that is trying to establish whether people who drink hard water containing higher levels of calcium and or magnesium are less likely to suffer cardiovascular diseases.

A large number of studies have investigated the potential health effects of drinking-water *hardness*. Most of these have been *ecologic* and have found an inverse relationship between water hardness and cardiovascular mortality. Inherent weaknesses in the ecologic study design limit the conclusions that can be drawn from these studies.

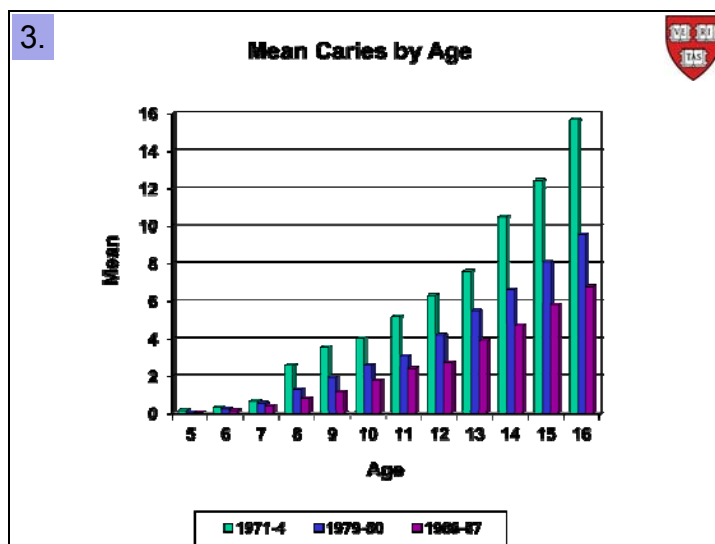
Based on identified *case-control* and *cohort studies*, there is no evidence of an association between water hardness or calcium and acute myocardial infarction or deaths from cardiovascular disease (acute myocardial infarction, stroke and hypertension). There does not appear to be an association between drinking-water magnesium and acute myocardial infarction. However, the studies do show a negative association (i.e. protective effect) between cardiovascular mortality and drinking-water magnesium. Although this association does not necessarily demonstrate causality, it is consistent with the well known effects of magnesium on cardiovascular function.

Above is shown the address of the WHO website, and you may wish to go there to follow the debate. They recognize the shortcomings of an ecological study and what they are looking for is something better, possibly case control studies and maybe cohort studies to settle the debate.



Here is another example of an ecological study that shows the per capita consumption of dietary fat and age-adjusted mortality from breast cancer in different countries. We see a wonderful, wonderful lozenge shaped relationship, but it is not telling us anything about the individuals involved. It is telling us something about the aggregate level, and the aggregate correlation is going to be higher than the individual correlation. We need further study to go into this.

But let me repeat, correlation at the aggregate level does not mean that there might not be causality at the individual level too.




Consider this case in point. We looked at how the mean caries varied by age across three surveys; the first in 1971-74, the second about eight years later, and the third about seven years after that. What we saw was that in each age group the number of caries goes down over those three studies. The explanation was that the more and more locales fluoridated their water over


the period of the studies. But this linkage between fluoride and cavities was first suspected because different regions in the country had naturally different fluoride content in the water and that was correlated with stronger teeth. As a result fewer dentists need fill in cavities.

There are any number of relationships, such as asbestos and mesothelioma, that start off as observation of correlations and that subsequently are proven to be causal. I repeat, just because it is correlation does not necessarily mean that it is not causal also. Indeed, it might be the first tip off, as it was with fluoride and asbestos, to lead to the right direction to discover the causal path.

### Spearman's Correlation Coefficient

Robustness






Note also that  $r$  is sensitive to outliers & it measures linear relationship.

Alternative: Spearman's rank correlation – same as Pearson's but replace observations by their ranks.

Charles E Spearman  
1863—1945

Pearson's correlation coefficient is not robust and is sensitive to all the observations. (You can wait till next week to understand this more fully, or research the issue for yourself.) Further, this coefficient is quantifying a linear relationship. So what Spearman—a person whose name is related to intelligence testing and all that entails—came up with a very clever idea. He argued that since we may or may not have normality, and its associated linearity, replace the observations with their ranks. Then calculate the correlation coefficient between the ranks. That is what we today call the Spearman correlation coefficient.



**Tied Ranks:**

X :	1.7	2.3	2.3	3.4
Ranks:	1	← 2 →		4
		0.5 (2+3)		
	1	2.5	2.5	4

---

X :	1.7	2.3	2.3	2.3	4
	1	3	3	3	4

$$\frac{2+3+4}{3} = 3$$

First, a quick reminder about ranks: if we have these four observations, 1.7, 2.3, 2.3, and 3.4, we might give them ranks 1, 2, 3, 4, except that the two middle ones are equal—we call those tied ranks.

There are a number of ways to handle tied ranks—for example, argue that each of the 2.3 are second smallest so each should get rank 2. The most common compromise is to average out the ranks that would have been given if we had slightly jiggered the data to break the ties, but not enough to reorder them. For example change the 2.3s to 2.31 and 2.32. Then the original 2.3s would get ranks 2 and 3. Average those out and associate the rank of 2.5 with each of the original 2.3. This is what sometimes happens in sporting events when the prize money is averaged out between contestants who tie.

It does lead to fractional ranks, although the worst it can get is, as here, introduction of a 0.5 in the ranks, and that happens if even numbers of people reach a tie. If odd numbers of people reach a tie, then the ranks remain whole. I leave that for you to prove for yourself.

e.g. fake numbers:

i	Raw Data		Ranks	
	x	y	$x_r$	$y_r$
1	1.3	14.3	2	2
2	1.7	14.7	4	3
3	0.8	18.0	1	4
4	1.4	12.1	3	1

$$r_s = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_{ri} - \bar{x}_r}{s_{x_r}} \right) \left( \frac{y_{ri} - \bar{y}_r}{s_{y_r}} \right)$$

So returning to Spearman, operationally this is what you do: first rank the x amongst themselves, then rank the y amongst themselves, all the while retaining the order of the data to maintain the proper linkage. Then ignore the original data and act as if the ranks are the original data. Now calculate the Pearson correlation coefficient with these ranks. That is the Spearman correlation coefficient,  $r_s$ .

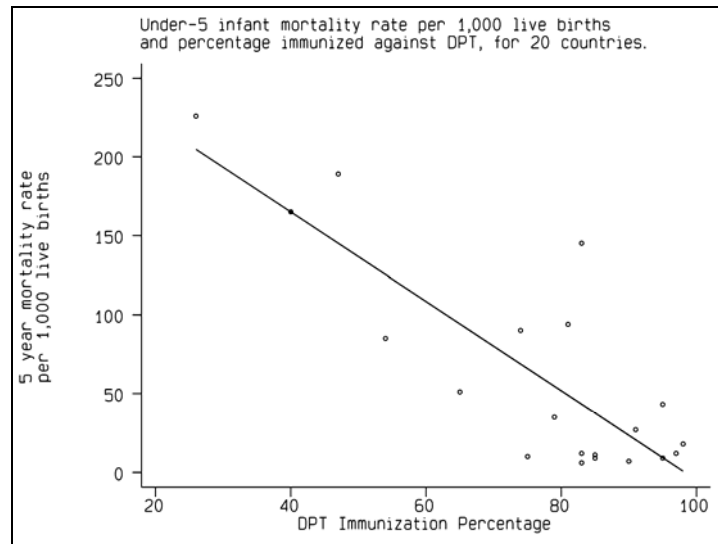
e.g. fake numbers:

i	Raw Data		Ranks		d
	x	y	$x_r$	$y_r$	
1	1.3	14.3	2	2	0
2	1.7	14.7	4	3	1
3	0.8	18.0	1	4	-3
4	1.4	12.1	3	1	2

$$r_s = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_{ri} - \bar{x}_r}{s_{x_r}} \right) \left( \frac{y_{ri} - \bar{y}_r}{s_{y_r}} \right)$$

$$= 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

One nice mathematical observation is this formula; useful if you do not have a computer to do your calculations for you!



Returning to our DPT example, chances are that neither the immunization percentages nor the mortality rates are normally distributed, so we might be tempted to calculate the Spearman's correlation coefficient in this case.

Nation	%Immun.	Rank	Death /1000	Rank	d	d <sup>2</sup>
Ethiopia	26	1	226	20	-19	361
Bolivia	40	2	165	18	-16	256
Senegal	47	3	189	19	-16	256
Brazil	54	4	85	14	-10	100
Mexico	65	5	51	13	-8	64
Turkey	74	6	90	15	-9	81
U.K.	75	7	10	5	2	4
USSR	79	8	35	11	-3	9
Egypt	81	9	94	16	-7	49
Japan	83	10	6	1	9	100
Greece	83	11	12	7.5	3.5	12
India	83	12	145	17	-6	36
Italy	85	13	11	6	7	56
Canada	85	14	9	3.5	10	100
Finland	90	15	7	2	13	169
Yugoslavia	91	16	27	10	6	36
France	95	17	9	3.5	14	196
China	95	18	43	12	5.5	30
USA	97	19	12	7.5	11.5	132
Poland	98	20	18	9	11	121
Total						2169

Here is the tabular data and all the steps necessary to calculate Spearman's correlation coefficient.

In the DPT example:

$$\begin{aligned} r_s &= 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \\ &= 1 - \frac{6(2,169)}{20(399)} \\ &= -0.631 \end{aligned}$$

Versus Pearson's:

$$r = -0.829$$

And the Spearman's correlation coefficient is -0.631. Remember that how we rank, from smallest to largest, or vice versa, is arbitrary, so we need to look how we ranked here in order to understand what the coefficient is saying. In this case we ranked the immunization with the lowest coverage getting a rank of 1, and then going up, and the lowest mortality getting a rank of 1, on then up. So a negative correlation means that as the coverage goes up the mortality goes down, as we might expect.

It turns out that with these data, the Pearson's was not that far different from the Spearman's, but it does buy us a little safety to know both, anyway.

To test **correlation** of two characteristics  
(only has power against  $\rho \neq 0$ )

$$\begin{aligned} t_s &= r_s \sqrt{\frac{n-2}{1-r_s^2}} \\ &= -0.631 \sqrt{\frac{18}{1-(0.631)^2}} \\ &= -3.45 \end{aligned}$$

versus t with 18 degrees of freedom,  
so  $0.001 < p < 0.01$

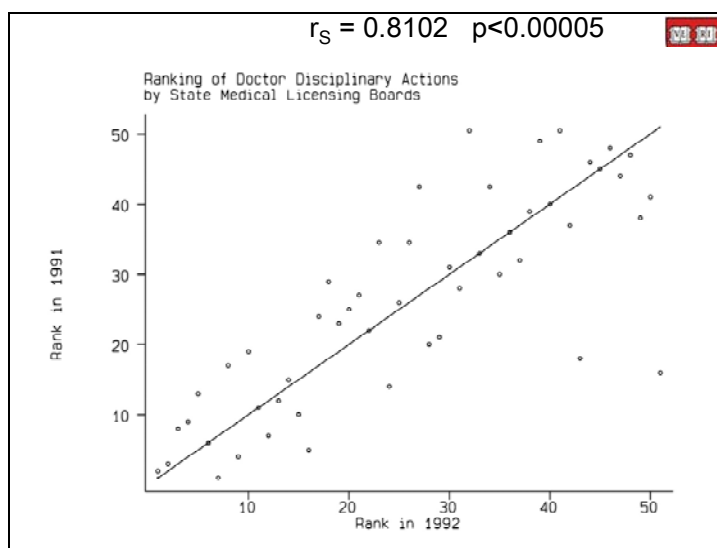
<sup>7</sup> I have edited this slide. The word correlation appears where the word independent used to be. It was a mistake.



Note that in the Spearman's correlation coefficient the word linear does not appear. That is because it is not necessarily measuring just a linear relationship between the two variables. Mathematically, we say that it is measuring the strength of a monotonic relationship. (Monotonic means that the variables travel together. Either they both go up (down) together, or one goes up as the other goes down. A special case of this is a straight line. With a positive slope they both go up (down) together, and with a negative slope one goes up as the other goes down. But the straight line is not the only relationship that is monotonic. Think of non-straight line (or non-linear) relationships such as weight gain with age; height with age; etcetera.)

So Spearman is a generalization of Pearson, but it does not solve the U-shaped relationship problem. (Test it for yourself with the same example we used above for Pearson, namely  $Y=|X|$ . That relationship yields both a Pearson and a Spearman of zero.)

Testing hypotheses with Spearman's correlation is almost identical to the Pearson case, even sharing the shortcoming that it only has power against the null hypothesis that  $\rho$  is zero.



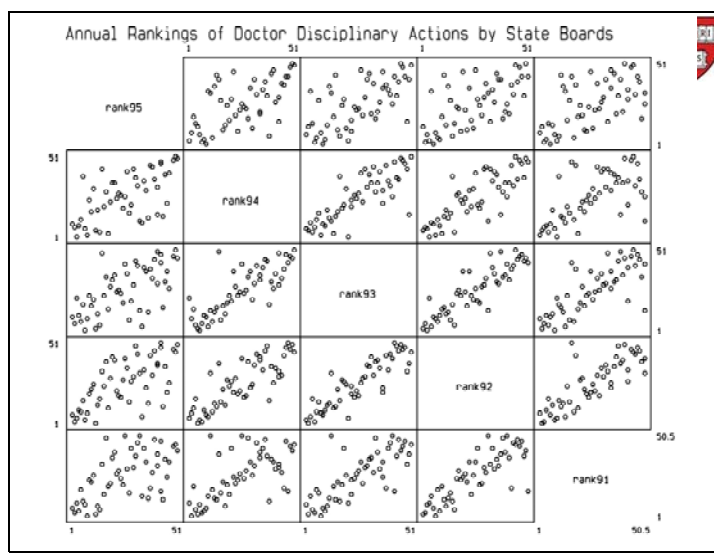
Here is another situation where you would certainly want to calculate the Spearman correlation coefficient. This shows the rankings of States (plus the District of Columbia) by a watchdog group of the medical profession in the US<sup>8</sup>. They keep an eye on the disciplinary action taken by the state medical boards.

This is a plot of the ranks in 1991 against the ranks in 1992. We see that apart from a few outliers, there seems to be a very strong correlation between these two years. In support of this observation we have that Spearman's rank correlation is 0.8102. So, here, we do not have to

<sup>8</sup> <http://www.citizen.org/Page.aspx?pid=183>

assume anything about the distribution of what it is that these people are measuring. We are just looking at their ranks.


I found it rather surprising, when I first saw this, that from one year to the next the medical boards' actions are so highly correlated. Why this should be I do not know. Do all bad doctors go to the same State?



Here is a matrix plot to capture this behavior over the five years, 1991 to 1995. To read a matrix plot we look at each diagonal entry. That variable defines all the horizontal axes for the plots in that column, and all the vertical axes for all the plots in that row. Looking at the plot as a whole, we see that the set of the lower triangle of plots is almost the same as the upper triangle set, so I could have suppressed one set of them—Stata gives me that option—but the role of column and row axes are interchanged in these two sets, and I wanted to retain that.

When looking at this particular set of variables it is interesting to look along diagonal rows. We have spoken of the main diagonal, it holds the names of the variables. If we move one up, so to speak, we get the plots when the two years (vertical and horizontal axes) are one-year apart. We get the same comparison by going down one diagonal from the main diagonal. If we go up (down) two diagonals, then we get the comparisons of years that are two-years apart. And so on.

The pattern that we are seeing is that they seem pretty tight when they are one-year apart, but that that tightness decreases as we go further away from the main diagonal; i.e. when the comparison is being made of two years further apart in time. (Remember the correlation game you played!)




```

. spearman rank95 rank94
Number of obs =   51
Spearman's rho =   0.6035
Test of Ho: rank95 and rank94 independent
Pr > |t| =   0.0000

. pwcorr rank95-rank91
      | rank95 rank94 rank93 rank92 rank91
-----+-----
rank95 | 1.0000
rank94 | 0.6057 1.0000
rank93 | 0.6321 0.8071 1.0000
rank92 | 0.6441 0.8168 0.8808 1.0000
rank91 | 0.5833 0.6292 0.7643 0.8102 1.0000

```

Here are the Spearman coefficients to accompany these plots. We see the same striation pattern as we see in the plots.



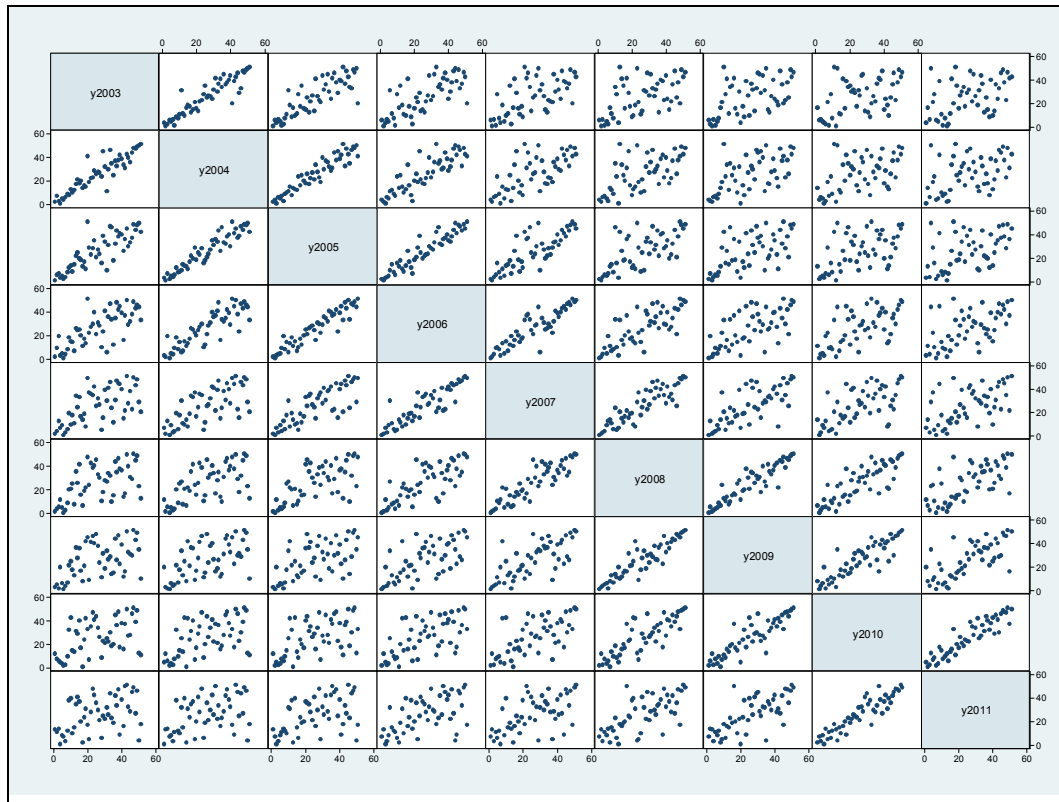
```

. pwcorr rank95-rank91 , bonf sig
      | rank95 rank94 rank93 rank92 rank91
-----+-----
rank95 | 1.0000
      |
rank94 | 0.6057 1.0000
      | 0.0000
rank93 | 0.6321 0.8071 1.0000
      | 0.0000 0.0000
rank92 | 0.6441 0.8168 0.8808 1.0000
      | 0.0000 0.0000 0.0000
rank91 | 0.5833 0.6292 0.7643 0.8102 1.0000
      | 0.0001 0.0000 0.0000 0.0000


```

You can also ask for the significance value with the command *pwcorr*. Here it is for the Bonferroni, and we see that they are all significant at the 5% level.

I was confused about why this correlation existed, but it was 15 years, or so ago. So recently I went searching for more current data and here it is for the years 2003 to 2011.



And here is that same matrix. This time the years range from 2003 to 2011. If anything, the striation pattern looks even more pronounced now than it did fifteen years ago.



```


. spearman y2003-y2011
(obs=51)

```

	y2003	y2004	y2005	y2006	y2007	y2008	y2009	y2010
y2003	1.0000							
y2004	0.9142	1.0000						
y2005	0.8287	0.9472	1.0000					
y2006	0.7353	0.8584	0.9395	1.0000				
y2007	0.6213	0.6976	0.8093	0.8927	1.0000			
y2008	0.5459	0.6207	0.6580	0.7658	0.8824	1.0000		
y2009	0.4996	0.5721	0.5879	0.6494	0.7319	0.9138	1.0000	
y2010	0.4703	0.5075	0.5360	0.6319	0.6437	0.8260	0.8757	1.0000
y2011	0.4280	0.4401	0.4771	0.5670	0.5733	0.6791	0.7206	0.9182

Public Citizen's Health Research Group Ranking of the Rate of State Medical Boards' Serious Disciplinary Actions, 2009-2011  
SM Wolfe, C Williams, and A Zaslow , May 17, 2012  
<http://www.citizen.org/documents/2034.pdf>

Here is the matrix of the Spearman correlation coefficient. The striation pattern is quite evident.



```

. spearman y2003-y2011 , p(0.05) bonferroni
(obs=51)

```


	y2003	y2004	y2005	y2006	y2007	y2008	y2009	y2010
y2003	1.0000							
y2004	0.9142	1.0000						
y2005	0.8287	0.9472	1.0000					
y2006	0.7353	0.8584	0.9395	1.0000				
y2007	0.6213	0.6976	0.8093	0.8927	1.0000			
y2008	0.5459	0.6207	0.6580	0.7658	0.8824	1.0000		
y2009	0.4996	0.5721	0.5879	0.6494	0.7319	0.9138	1.0000	
y2010	0.4703	0.5075	0.5360	0.6319	0.6437	0.8260	0.8757	1.0000
y2011		0.4401	0.4771	0.5670	0.5733	0.6791	0.7206	0.9182

And the Bonferroni results are here, and they are all significant at the 5% level, except for the bottom left corner. (Stata has changed how it reports these results from 15 years ago when it showed the p-value. Now it shows which coefficients are significant, at the chosen level of significance.)

What does this show? Is this just an example of something we know very well, which is that these medical boards tend to have a number of physicians sitting on them, so this is just a

manifestation of the observation that groups of people cannot police themselves? I leave to you to interpret these graphs.

## Non-parametrics

[http://en.wikipedia.org/wiki/Sex\\_ratio](http://en.wikipedia.org/wiki/Sex_ratio)
11/10/2012


The CIA estimates that the current world wide sex ratio *at birth* is 107 boys to 100 girls.<sup>[3]</sup> In 2010, the global sex ratio was 986 females per 1,000 males and trended to reduce to 984 in 2011.<sup>[4]</sup>

107 m : 100 f		$\Pr(\text{male}) = \frac{107}{207} = 0.5169$
(935 f : 1000 m)		
984 f : 1000 m	$\Rightarrow$	$\Pr(\text{male}) = \frac{1000}{1984} = 0.5040$
986 f : 1000 m		$\Pr(\text{male}) = \frac{1000}{1986} = 0.5035$

Spearman's rank correlation coefficient is your first example of something that is called a non-parametric statistic because it does not make any assumptions, which can be characterized up to some parameters, about the distribution of the population.

The first example I want to show you has to do with the sex ratio. So I went to Wikipedia and this is what I found: the CIA estimates that the current worldwide sex ratio at birth is 107 boys to 100 girls<sup>9</sup>. Also, The Times of India gave the ratio for 2010 and for 2011, and there they were<sup>10</sup>.

I think there is something wrong with these numbers, because if you put them all on the same scale— 1,000 males, say—then I do not think that in three years one sees this much disparity.



But accepting that, there are two points I want to make with this. The first is that we are still talking about the sex ratio at birth. We seem to be stuck on this theme for many years now. It has been used as an indicator for any number of things. Currently there is concern about the worrisome practice of using technology to influence this ratio by producing more males.

<sup>9</sup> [https://www.cia.gov/library/publications/the-world-factbook/fields/print\\_2018.html](https://www.cia.gov/library/publications/the-world-factbook/fields/print_2018.html)

<sup>10</sup> [http://articles.timesofindia.indiatimes.com/2011-08-17/india/29895810\\_1\\_ratio-abortion-and-craze-craze-for-male-child](http://articles.timesofindia.indiatimes.com/2011-08-17/india/29895810_1_ratio-abortion-and-craze-craze-for-male-child)

The other point struck me when reading the Times of India. After listening to the complaints of some students in the past who claim not to be able to understand odds, I see before me a newspaper that presumably appeals to a large, general readership talking about odds! Surely they want their readership to understand what they are writing, and yet they feel free to use odds.

For those of you who do not appreciate odds, we can calculate the probabilities, and there they are.

 <p>John Arbuthnot 1667-1735</p>	 <p>Claim: Divine providence, not chance, governs the sex ratio at birth.</p> <p>John Arbuthnot (1710) "An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes." <i>Philosophical Transactions of the Royal Society</i></p>
---	--

Why I brought up the sex ratio at birth is John Arbuthnot. He not only was a physician, to the queen yet, the inventor of John Bull, but he also was a man way ahead of his time both in measuring characteristics of people, and in the formulation of a statistical hypothesis test. This he did in the context of the sex ratio, which he used in order to utilize statistics to prove there is a god. This is one of the first times hypothesis testing was used in the literature.

This claim he made in his article, "An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes," in the *Philosophical Transactions of the Royal Society*, in 1710.

Data:

Christenings, London 1629--1710 where for 82 years he observed more boys christened than girls.

If  $\Pr(\text{boy}) = 0.5$ ,

$\text{Prob}(82 \text{ years, \# boys} > \# \text{ girls}) = (0.5)^{82}$

$$= \frac{1}{4\,8360\,0000\,0000\,0000\,0000\,0000}$$

“From whence it follows, that it is Art, not Chance, that governs.”

The way he went about his study was to look at the christening records in London for the period 1629 to 1710. For those 82 years he counted that there were more boys than girls christened. Chance to him meant that the sex ratio had to be 0.5. So to see 82 years, out of 82 years observed, where there were more boys than girls had that minute p-value above. Getting this p-value was much too small to believe Chance was governing this, so Art must have been at work!

Now, why chance had to be exactly 0.5 and not any other number he did not defend. This was one of the first examples, the other being Neumann and the data from Breslau used to disprove astrology, of the use of hypothesis testing.

Resting energy expenditure (kcal/day): cystic fibrosis & healthy, matched on age, sex, height and weight.

Pair	CF	Healthy	Diff	Sign
1	1153	996	157	+
2	1132	1080	52	+
3	1165	1182	-17	-
4	1460	1452	8	+
5	1634	1162	472	+
6	1493	1619	-126	-
7	1358	1140	218	+
8	1453	1123	330	+
9	1185	1113	72	+
10	1824	1463	361	+
11	1793	1632	161	+
12	1930	1614	316	+
13	2075	1836	239	+



What Arbuthnot did is what we now call the Sign Test. For each year he counted the number of boys christened and subtracted the number of girls born, and he came up with 82 plus signs. If  $p$ , the probability a boy is christened is 0.5, then he would have expected about 41 years where there would be more boys and 41 years where more girls were christened.

Here is a more modern example. We have 13 pairs of individuals, 13 with cystic fibrosis and another 13 controls without, who were matched with the cases according to age, sex, height, and weight. The outcome measured in each patients was the resting energy expenditure.

Taking the differences in the outcomes in each pair we see 11 positive and 2 negative differences.

### The Sign Test

Let  $D$  be the number of positive differences:


$$Z_+ = \frac{D - (n/2)}{\sqrt{n/4}}$$

$$= \frac{11 - 6.5}{1.80} = 2.50$$

$p = 0.006 + 0.006 = 0.012 < 0.05$

For small  $n$  use the binomial.

We can use a normal approximation, as above, or call on Stata.



```
. signtest CF = Healthy
Sign test
```

sign	observed	expected
positive	11	6.5
negative	2	6.5
zero	0	0
all	13	13

```
One-sided tests:
Ho: median of CF - Healthy = 0 vs.
Ha: median of CF - Healthy > 0
Pr(#positive >= 11) =
  Binomial(n = 13, x >= 11, p = 0.5) = 0.0112

Ho: median of CF - Healthy = 0 vs.
Ha: median of CF - Healthy < 0
Pr(#negative >= 2) =
  Binomial(n = 13, x >= 2, p = 0.5) = 0.9983


Two-sided test:
Ho: median of CF - Healthy = 0 vs.
Ha: median of CF - Healthy != 0
Pr(#positive >= 11 or #negative >= 2) =
  min(1, 2*Binomial(n = 13, x >= 11, p = 0.5)) = 0.0225
```

We see from the Stata output that the null hypothesis that the sign is just as likely to be positive as negative has a p-value of 0.0225, and thus we reject that hypothesis at the 5% level.

The meaning of a probability of 0.5 of a positive or negative sign of the difference is that the median of the distribution of that difference is zero. In other words, that the distributions of the controls and the cases both have the same median.

## Wilcoxon Signed Rank Test


Resting energy expenditure (kcal/day): cystic fibrosis & healthy, matched on age, sex, height and weight.



Pair	CF	Healthy	Diff	Sign
1	1153	996	157	+
2	1132	1080	52	+
3	1165	1182	-17	-
4	1460	1452	8	+
5	1634	1162	472	+
6	1493	1619	-126	-
7	1358	1140	218	+
8	1453	1123	330	+
9	1185	1113	72	+
10	1824	1463	361	+
11	1793	1632	161	+
12	1930	1614	316	+
13	2075	1836	239	+

The Sign test only looks at the signs of the differences and not much else. So small differences and large differences are all counted the same. Frank Wilcoxon argued that there is some information to be gained by looking at the magnitude of these differences.


Ranks usually 1 top  
2 next .....



2010 U.S. Suicides, by methods & sex				
Method	Male	Rank	Female	Rank
Firearms	16,962	1	2,430	2
Poisoning	3,573	3	3,026	1
Suffocation	7,592	2	1,901	3
Other	2,150	4	730	4
Total	30,277		8,087	

Consider this report of the methods of suicide in the US in 2010<sup>11</sup>. If we rank the data for males and females we find that for males the most popular method involves the use of firearms. The next most popular is what they used to call strangulation and suffocation, presumably one hangs oneself. For females, poison is the most popular way of committing suicide. Firearms is the second most popular and suffocation is third.

From looking at the rankings of the methods used, we can see that there is a qualitative difference between the sexes. If we want to act as if these were samples and we want to do some sort of t-test, then we would have to claim some normality, but if these numbers are not normally distributed, can we do something with the ranks?




One sample, or paired, Wilcoxon

e.g. Reduction in forced vital capacity (FVC) for a sample of patients with cystic fibrosis:

Frank Wilcoxon  
1892 - 1965

Wilcoxon came up with tests analogous to the t-tests but instead of using the raw data, he first ranks the data and then uses the ranks. Indeed, Wilcoxon is to Student as Spearman is to Pearson. First we look at the one-sample version of the Wilcoxon by applying it to a study of a drug to ameliorate the effects of cystic fibrosis. Forced vital capacity (FVC) is the volume of air that one can expel from one's lungs in 6 seconds. In this study they compared the reduction in FVC for patients with cystic fibrosis over two similar periods of time, once when taking the drug and once when on placebo.


<sup>11</sup> [http://webappa.cdc.gov/sasweb/ncipc/mortrate10\\_us.html](http://webappa.cdc.gov/sasweb/ncipc/mortrate10_us.html)



Pat	Placebo	Drug	Diff	Rank	"Signed Rank"
1	224	213	11	1	1
2	80	95	-15	2	2
3	75	33	42	3	3
4	541	440	101	4	4
5	74	-32	106	5	5
6	85	-28	113	6	6
7	293	445	-152	7	7
8	-23	-178	155	8	8
9	525	367	158	9	9
10	-38	140	-178	10	10
11	508	323	185	11	11
12	255	10	245	12	12
13	525	65	460	13	13
14	1023	343	680	14	14
Totals (T)					86 19

Here are fourteen pairs of readings on the fourteen patients. We obtain the differences, as in the Sign test (or the dependent t-test), and then rank the absolute values of the differences.

We then look at the total ranks of those associated with positive differences and the total ranks of those with negative differences. If the distribution of the differences is symmetric around zero, then one would intuitively expect that these two sums should be approximately the same. This is the Wilcoxon signed rank test.



$H_0: \text{Median}_x = 0$

Under  $H_0$


$$\text{mean}_T = \mu_T = \frac{1}{4}n(n+1)$$

$$\sigma_T = \sqrt{n(n+1)(2n+1)/24}$$

For large  $n$ :

$$Z = \frac{T - \mu_T}{\sigma_T} \text{ approx. stand. normal}$$

In contrast to the Sign test, we are not only taking the sign of the difference into account, but also the size of these differences by looking at the ranked differences. The Z-statistic to use, which is displayed above, has an approximately normal sampling distribution.




For the above example,  $T = 19$   
and  $n = 14$

$$Z = \frac{19 - 14(15) / 4}{\sqrt{14 \times 15 \times 29 / 24}}$$

$$= -2.10$$

$p = 0.036$

In this instance, p value is 0.036, and so we would reject the null hypothesis of equality. If we look at the data we see that the sum of the positive ranks is too high. That means that the placebo seems to cause a bigger loss than the drug, so the drug looks effective, on the basis of these data.



```
. signrank placebo = drug
Wilcoxon signed-rank test
```

sign	obs	sum ranks	expected
positive	11	86	52.5
negative	3	19	52.5
zero	0	0	0
all	14	105	105

unadjusted variance	253.75
adjustment for ties	0.00
adjustment for zeros	0.00
adjusted variance	253.75

Ho: placebo = drug  
 $z = 2.103$   
 Prob > |z| = 0.0355

## Wilcoxon Rank Sum Test<sup>12</sup>

Low Exposure				High Exposure			
nMA	Rank	nMA	Rank	nMA	Rank	nMA	Rank
34.5	2	51.0	23	28.0	1	51.0	23
37.5	6	52.0	25.5	35.0	3	52.0	25.5
39.5	7	53.0	28	37.0	4.5	53.0	28
40.0	8	54.0	31.5	37.0	4.5	53.0	28
45.5	11.5	54.0	31.5	43.5	9	54.0	31.5
47.0	14.5	55.0	34.5	44.0	10	54.0	31.5
47.0	14.5	56.5	36	45.5	11.5	55.0	34.5
47.5	16	57.0	37	46.0	13		
48.7	19.5	58.5	38.5	48.0	17		
49.0	21	58.5	38.5	48.3	18		
51.0	23			48.7	19.5		
Total ranks		467		Total ranks		313	

The Wilcoxon two sample test, called the Wilcoxon Rank Sum Test, is what we would use when the two samples are independent. Here are data to study children suffering from phenylketonuria (PKU), a disorder associated with the inability to metabolize the protein phenylalanine. The children have been divided into two groups, the first with average serum levels of phenylalanine less than 10.0 mg/dl and the second group with average phenylalanine levels above 10.0 mg/dl. The study wishes to compare the normalized mental scores for these two groups, without assuming normality of the distributions of these scores.

The test first pools the two samples in order to rank all of them together. Then take the sum of the ranks in each of the groups. Intuitively if the two population distributions are equal to each other and the two sample sizes are equal, then these two sums of ranks should be approximately equal. If the two sample sizes are different we can average things out to make the two comparable.

<sup>12</sup> Sometimes called the Mann-Whitney, or Wilcoxon-Mann-Whitney, or any permutation of these names.

Wilcoxon-Mann-Whitney Test  
for two independent samples.



I	Ranks	II	Ranks
$x_1$	$R_1$	$y_1$	$R_{n_1+1}$
$x_2$	$R_2$	$y_2$	$R_{n_1+2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{n_1}$	$R_{n_1}$	$y_{n_2}$	$R_{n_1+n_2}$

$$W = \sum_{j=1}^{n_1} R_j$$

$H_0$  : Sample I and II same population

Under  $H_0$

$$\text{Average}(W) = \mu_w = n_s(n_s + n_L + 1) / 2$$

$$\sigma_w = \sqrt{\frac{n_s n_L (n_s + n_L + 1)}{12}}$$

For large  $n_s, n_L$  ( $>10$ )

$$Z = \frac{W - \mu_w}{\sigma_w}$$

is approximately standard normal



So you'd expect the two ranks-- the two sums of the ranks be the same, if the two sample sizes are the same. If not, you'll do the appropriate weighting.



Example:  $W = 313$

$$\mu_W = n_S(n_S + n_L + 1) / 2$$

$$= 18(18 + 21 + 1) / 2 = 360$$

$$\sigma_W = \sqrt{n_S n_L (n_S + n_L + 1) / 12}$$

$$= \sqrt{18(21)(18 + 21 + 1) / 12} = 35.5$$

$$Z = \frac{W - \mu_W}{\sigma_W}$$

$$= \frac{313 - 360}{35.5} = -1.32$$

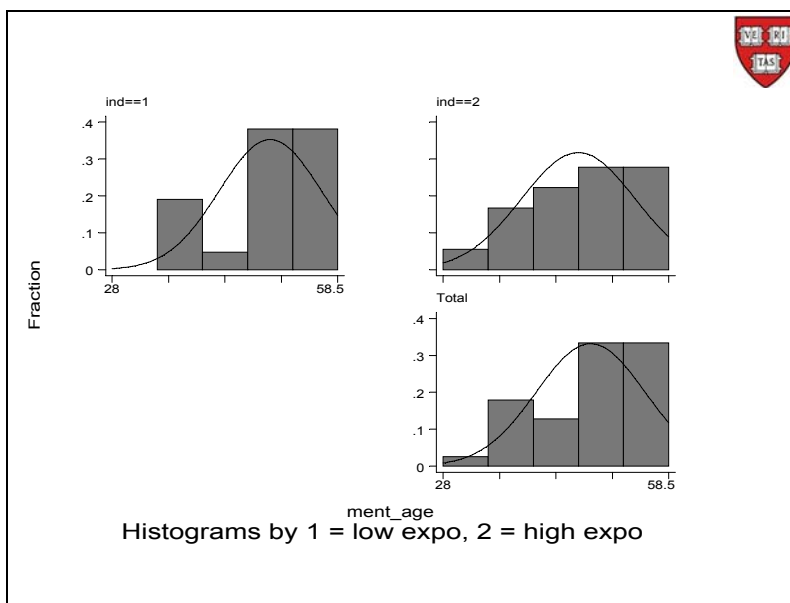
$$p = 0.186$$

In this example, the standard Z is -1.32, so the p-value is 0.186, and we do not reject the null hypothesis.

```
. tab ind, summ( ment_age )
```

1 = low expo, 2 = high expo	Summary of ment_age		
	Mean	Std. Dev.	Freq.
1	49.366667	6.9038636	21
2	46.277778	7.6722997	18
Total	47.941026	7.3384968	39

Here are the summary statistics for the raw data before we did the ranking. Pre-today, or pre you learning Wilcoxon, you might have done a t-test. These two means look approximately the same.



If we want to use the t-test we could look at the histograms. Unfortunately, none of the three (the two individual groups and the combined groups) histograms look particularly normal. So maybe we shouldn't be doing the t-test, but it does not matter, we are halfway there so let us do look at both tests.

```
. ranksum ment_age , by(ind)
```

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

ind	obs	rank sum	expected
1	21	467	420
2	18	313	360


combined	39	780	780
unadjusted variance		1260.00	
adjustment for ties		-3.19	
adjusted variance		1256.81	

Ho: ment\_age(ind==1) = ment\_age(ind==2)

z = 1.326

Prob > |z| = 0.1849

And now here's the rank-sum. If we did the Wilcoxon rank sum test we see that the p value is 0.1849,



ttest ment\_age , by(ind) unequal  
Two-sample t test with unequal variances


Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
1	21	49.36667	1.506547	6.903864	46.22407	52.50927
2	18	46.27778	1.808378	7.6723	42.46243	50.09312
combined	39	47.94103	1.1751	7.338497	45.56216	50.31989
diff		3.088889	2.353702		-1.691287	7.869065

Satterthwaite's degrees of freedom: 34.6139  
Ho: mean(1) - mean(2) = diff = 0

Ha: diff < 0	Ha: diff ~ 0	Ha: diff > 0
t = 1.3124	t = 1.3124	t = 1.3124
P < t = 0.9010	P >  t  = 0.1980	P > t = 0.0990

And if we do the t-test, the p value is 0.198. So in this set of data where the normality assumption is not too strictly adhered to, the t-test and the Wilcoxon rank sum test provide very similar results. It seems like the t-test is a little more robust than expected.

Once you have your data on the computer and have access to a decent set of computer routines and there are three different ways to do something, then explore. You will learn something about your data whether you get the same answer three times, or when you get different answers that will require some explaining.



Wilcoxon versus Student's t :

<b>Advantage</b>
We do not need to assume that the population is Normally distributed for Wilcoxon to be applicable – robust.
<b>Disadvantage</b>
When in fact the parent population is Normal, the Wilcoxon is less efficient (approximately 95% efficient).

In summary, when should we use the Wilcoxon, and when should we use Student's  $t$ ? The advantage of the Wilcoxon is that it does not require us to assume anything about the parent or population distribution of the variable in question—in particular we have no need to assume normality, whereas that is an assumption we need to make for Student's  $t$ . The Wilcoxon is more robust in that sense, and also in that it deals with ranks that are more robust to outliers.

But you are not going to get something for nothing, so how do we pay for this? Well, here's the disadvantage. If in fact you were justified in making your normality assumption—so if in fact it would be legitimate for you to use the  $t$ , how much do you lose by using the Wilcoxon? And the answer is not that much. When in fact you have normal data, the Wilcoxon is about 95% efficient, versus Student's  $t$ . So if you have 20 observations in your sample, the Wilcoxon is about as efficient as if you had 19 observations and use the  $t$  appropriately. I like the Wilcoxon, and as I said, it's not that expensive. And it's a good back up to see how different, if at all, it is to using the  $t$ .