Paired Data

- Bivariate Data that are coupled or matched together. They are not independent.

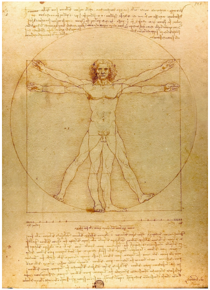
Example:

- Height and Weight measurements of individuals.
- Response reading before and after treatment of individuals.

Paired Data

Example:

- Leonardo da Vinci's *Vitruvian Man*.



- The outstretched arms and legs within circles and square.
- Ideal human proportions described by ancient Roman architect Vitrivius: height is same as length of arm span.

Paired Data

Key Tools to understand Data

- Plot to gauge relationship.
- Correlation between the variables.
- Trends

Paired Data

Consider `fat` dataset in `UsingR` package. The dataset contains body dimensions of 250 males.

```
> require(UsingR)
> names(fat)
```

```
 [1] "case"          "body.fat"      "body.fat.siri" "density"
 [5] "age"           "weight"        "height"        "BMI"
 [9] "ffweight"      "neck"          "chest"         "abdomen"
[13] "hip"           "thigh"         "knee"          "ankle"
[17] "bicep"         "forearm"       "wrist"
```

Paired Data

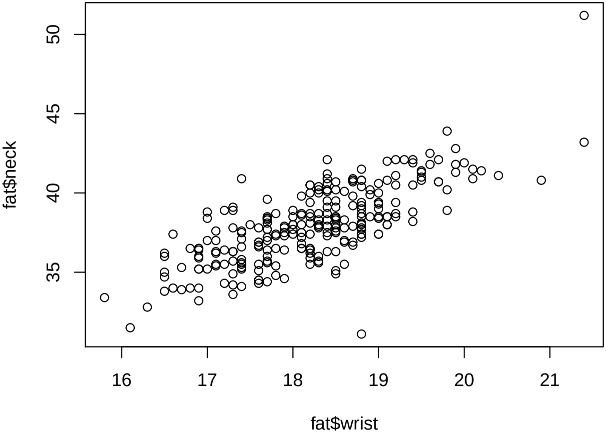
- Suppose we are interested in relation between neck and wrist.

We can first compare averages in two ways:

```
> z = mean(fat$neck)/mean(fat$wrist)
> z
[1] 2.084068
> y = mean(fat$neck/fat$wrist)
> y
[1] 2.084477
```

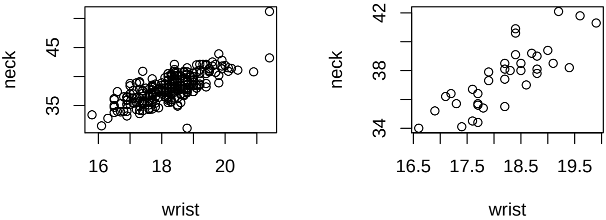
Paired Data: `fat` dataset in UsingR

```
> plot(fat$wrist, fat$neck)
```



Paired Data: `fat` dataset in UsingR

```
> par(mfrow=c(1,2))
> plot(neck~wrist, data=fat)
> plot(neck~wrist, data=fat, subset=20<=age &age <30)
```



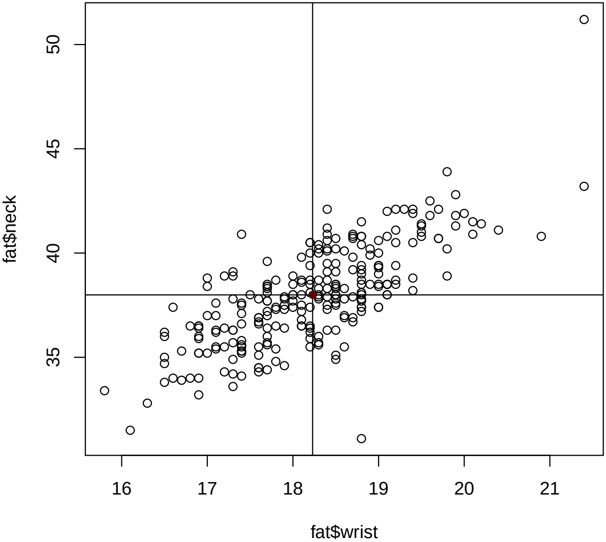
The variables seem related and also by a linear relationship

Paired Data: Correlation

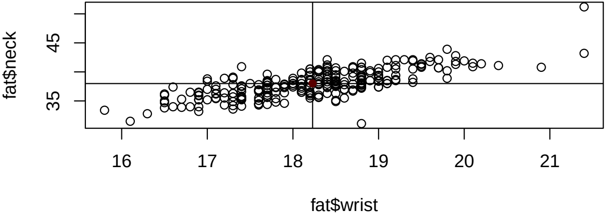
- Assume Linear Relationship between the data
- Correlation is a measure of how close the relationship is.

Before defining the term let us try to understand the plot better.

Data in four regions by means



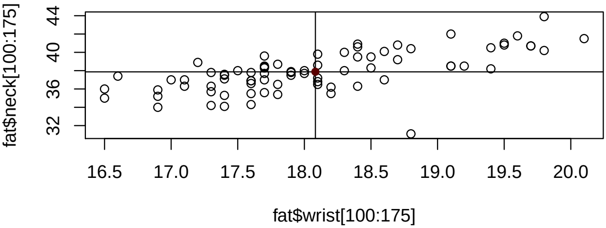
Data in four regions by means



- Understand data by those above average values and those below.
- If related then most of data should be in first and third box.

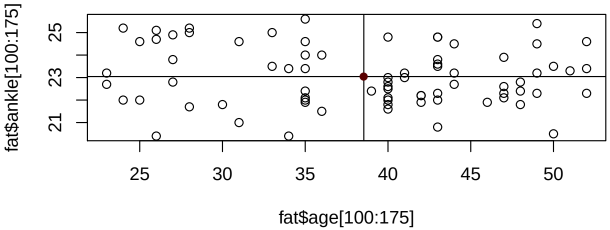
Paired Data: fat dataset in UsingR

```
> plot(fat$wrist[100:175], fat$neck[100:175])
> abline(v=mean(fat$wrist[100:175]))
> abline(h=mean(fat$neck[100:175]))
> points(mean(fat$wrist[100:175]), mean(fat$neck[100:175]),
+ pch=16, col=rgb(.35,0,0))
```



Paired Data: fat dataset in UsingR

```
> plot(fat$age[100:175], fat$ankle[100:175])
> abline(v=mean(fat$age[100:175]))
> abline(h=mean(fat$ankle[100:175]))
> points(mean(fat$age[100:175]), mean(fat$ankle[100:175]),
+ pch=16, col=rgb(.35,0,0))
```



Covariance

Covariance measures the difference between the two variables in the four regions. Suppose we have a dataset $\{(x_i, y_i) : 1 \leq i \leq n\}$ then

$$\mathrm{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Data with strong linear relationship $(x_i - \bar{x})(y_i - \bar{y})$ will have the same sign. (i.e if data lies in first and third box or in second and fourth box).
- In such cases covariance will be large in absolute value.

Pearson Correlation Coefficient

Correlation is Covariance in standardised scale. Suppose we have a dataset $\{(x_i, y_i) : 1 \leq i \leq n\}$ then

$$\mathrm{Cor}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{S_x} \right) \left(\frac{(y_i - \bar{y})}{S_y} \right)$$

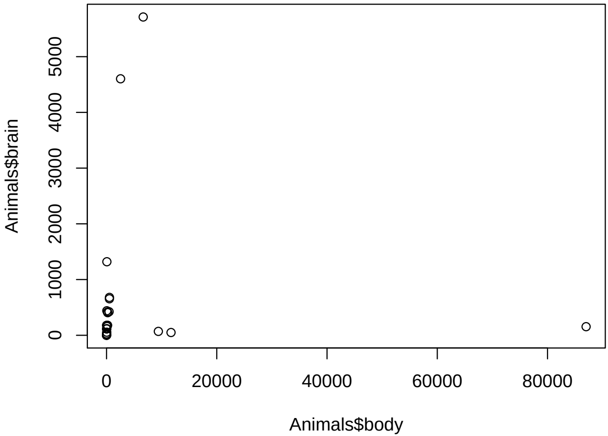
- $\mathrm{Cor}(x, y)$ is between -1 and 1.
- $\mathrm{Cor}(x, y) \in \{1, -1\}$ indicates perfect linear relationship.
- $\mathrm{Cor}(x, y) = 0$ indicates no linear relationship.

Paired Data: `fat` dataset in UsingR

```
> cor(fat$wrist, fat$neck)
[1] 0.7448264
> cor(fat$wrist, fat$height)
[1] 0.3220653
> cor(fat$age, fat$ankle)
[1] -0.1050581
```

Pearson Correlation Coefficient

```
> require(MASS)
> plot(Animals$body,Animals$brain)
```



Spearman Correlation Coefficient

```
> require(MASS)
> cor(Animals$body,Animals$brain)
[1] -0.005341163
```

- One way is to exclude the outliers.
- Another method is to transform the dataset by placing data in order and assigning a rank. Use `rank`.

```
> require(MASS)
> cor(rank(Animals$body), rank(Animals$brain))
[1] 0.7162994
```

or

```
> require(MASS)
> cor(Animals$body, Animals$brain, method="spearman")
[1] 0.7162994
```

Spearman Correlation Coefficient

Suppose we have a dataset $\{(x_i, y_i) : 1 \leq i \leq n\}$ then first rank themto get .$\{(r_{x_i}, r_{y_i}) : 1 \leq i \leq n\}$

$$\text{Spearman Correlation}(x, y) = \text{Cor}(r_x, r_y)$$

- measurement of relationship of monotonic data.
- not restricted to linear.

Chocolates and Noble Prizes

SCIENTIFIC AMERICAN™

Sign In / Register

Search ScientificAmerican.com

Subscription Center
Subscribe to Print & Tablet »
Subscribe to Print »
Give a Gift »
View the Latest Issue »

Subscribe | News & Features | Topics | Blogs | Videos & Podcasts | Education | Citizen Science | SA Magazine | SA Mind | Products

Blogs

About the SA Blog Network | Choose a blog....

The Curious Wavefunction

Musings on chemistry and the history and philosophy of science

The Curious Wavefunction Home

Chocolate consumption and Nobel Prizes: A bizarre juxtaposition if there ever was one

By Ashutosh Jogalekar | November 20, 2012 | 8

Share Email Print

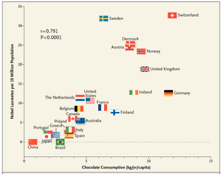


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Correlation of chocolate consumption with Nobel Laureates (Image credit: New England Journal of Medicine)

What makes a Nobel Prize winner? There's several suggested factors: Perseverance? Good luck? Good mentors and students? Here's one possible

More from Scientific American

MIND » Classics » DIGITAL »

Blog Network Highlights

The Thoughtful Animal »

The Animal Lover's Gift Guide

A Better Answer to Cl… Change Is Hidden in t…

Most Read Posts | Latest Posts

Plugged In
Air pollution stretches from Beijing to Shanghai, as seen from space

Observations
Glow Sticks Prove the Math Theorem behind the Famous Flatiron Building

factor that I would have never imagined in my wildest dreams; chocolate consumption. Chocolate consumption tracks well with the number of Nobel Laureates produced by a country.

At least that's what a paper published in the New England Journal of Medicine – one of the world's premier journals of medical research – claims. I have to say I found the study bizarre when I read it, and a few hours of strenuous, perplexed thought have done nothing to shake that feeling off. The study itself is amusing and rather brief and I think it makes for entertaining reading; what I am left contemplating is why this paper constitutes serious research and why it would have been published in a journal which over the years has presented some of the definitive medical findings of our time.

The paper starts by assuming – entirely reasonably – that winning a Nobel Prize must somehow be related to cognitive ability. It then goes on to describe a link between flavanols – organic molecules found among other foods in chocolate, green tea and red wine – and cognitive ability. Now I haven't read the literature on flavanols in detail, but I am sure that flavanols themselves couldn't possibly be responsible for improved cognitive effect, especially when they are part of a complex cocktail of dietary and environmental factors affecting brain function.

But let's say that's true; flavanols are indeed a strong indicator of cognitive function. From this idea the author basically jumps to the dubious and frankly bizarre question of whether chocolate consumption could possibly account for Nobel Prize winning ability. However, from a purely scientific standpoint the hypothesis is testable, so the author decides to simply plot the number of Nobel prize winners per 10 million people in different countries counted from 1900-2011 vs the chocolate consumption in those countries. The figures for chocolate consumption come from Caobisco and Chocosuisse and cover only four years, none before 2002. This fact itself makes any such comparison dubious to say the least; how can you compare two variables when they are sampled from such radically dissimilar sample spaces? And what about other compounds containing flavanols; why not also consider red wine or green tea?

In any case, a plot of chocolate consumption vs number of Nobel Prizes reveals a strong correlation of 0.79. Sweden is an anomaly (and the author thinks it could be a result of "patriotic bias" from the Nobel Committee); take it out and the correlation improves to 0.86. The graph in all its glory is illustrated above.

What does one make of this? Well, I have said before that if only three rules of scientific deduction were inscribed on the doors of every university and research organization in the world, one of them should be that "correlation does not mean causation". Conflating the two can lead you to believe, for instance, that storks deliver babies. Now the author recognizes this, but what I find absolutely baffling is that he makes no attempt to dissect other possible contributing factors. In fact at the end of the article he acknowledges the existence of such factors and then proceeds to dismiss them by saying that *"differences in socioeconomic status from country to country and geographic and climatic factors may play some role, but they fall short of fully explaining the close correlation observed."*

Observations
Are Genes Really Selfish? [Video]

Observations
Can We Avert the End of Elephants?

Observations
The FDA's Action on Agricultural Antibiotics Is Overdue—and Utterly Insufficient

Follow Us:

See what we're tweeting about
Scientific American Editors

Free Newsletters
Get the best from Scientific American in your inbox

Email address Go

Latest Headlines on ScientificAmerican.com
The Life of Dan Wegner: A Meeting Place for Joy and Intelligence
Energy Secretary: time to revisit US ban on crude exports [updated]
The Science of U.S. Energy: A Q&A with Secretary Ernest J. Moniz
Harnessing Disruption to Move Forward with Biotech [Video]
The Animal Lover's Gift Guide

Latest from SciLogs
Citation Rates Highlight Uphill Battle for Women in Research Careers
Let it snow, let it snow, let it snow!
Airhead Alioramus: Why Theropods Were Light-headed
Why Science Blogging Requires Story-Telling
What is Science Blogging?
More »

Chocolates and Noble Prizes

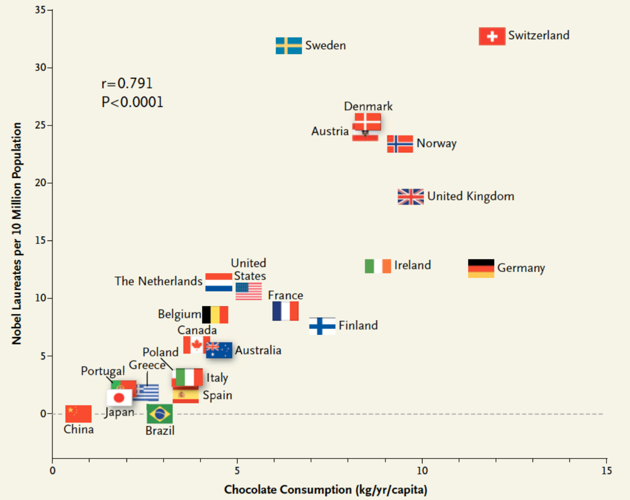


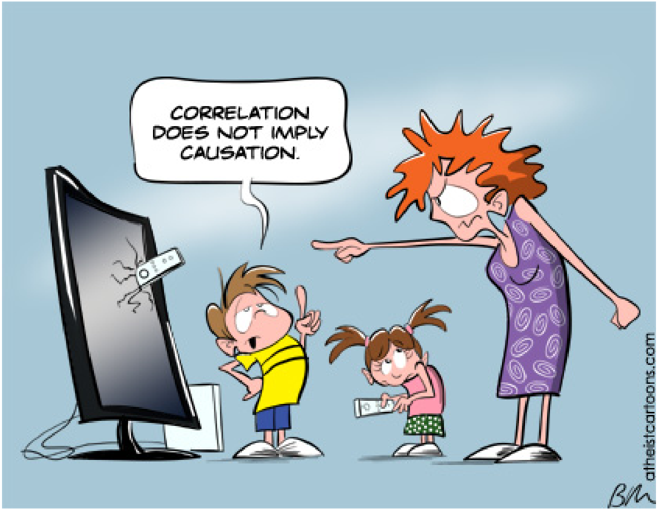
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Chocolates and Noble Prizes

Noticed: Countries with more per capita chocolate consumption have more per capita Nobel laureates.

Conclude: Chocolate consumption cause better scientific research !

Chocolates and Noble Prizes

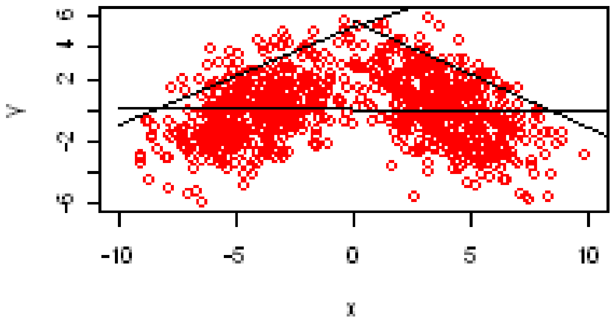


Chocolates and Noble Prizes

- Spurious: Facebook Users and Marks of users
- Causality: Smoking and lung cancer, Wine and heart risk.

Correlation

- Non-linear relationship
- 0 correlation



Correlation

- Pearson correlation coefficient is a measure of the linearity of the (possible) relationship between two variables X and Y.
- Even if correlation coefficient is high, it does not mean there is causal relationship between X and Y. Does not tell you cause and effect ?
- Care to be taken when used for predictive purposes.
- Causality: Domain Knowledge, design a good control experiment.

Simple Linear Regression: Relationship in Bivariate Data

- Key: conditional mean of response variable given the predictor cariable is a linear function.
- Model: For data points (x_i, y_i) with $1 \leq i \leq n$,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

  where ε_i assumed to be mean 0 and variance σ^2 Normal random variables.
- Observe only (x_i, y_i) for $1 \leq i \leq n$.

Simple Linear Regression: Relationship in Bivariate Data

- Find β_0, β_1 such that

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

  is minimized.
- Can be solved: Calculus and Linear Algebra

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \text{correlation}(x, y)\frac{S_x^2}{S_y^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Observations:

- Slope of line is function of Correlation in standarised scale.
- Line passes through $(\bar{x}, \bar{y})$
- Roles of y and x are not interchangeable.

Simple Linear Regression

```
[1] 0.7448264
```