Associations, Correlations, and Causation

Visualization ReCap

- http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/
- https://www.gapminder.org/tools/#_chart-type=bubbles
- http://www.informationisbeautiful.net/
- www.bewitched.com

Association: Peeling the Onion

I See Dead People

What's wrong with this statement?

Associations are not Causations

Everyone who ate pickles in the year 1743 is now dead.

Therefore, pickles are fatal?

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

True statement:

Palm size correlates with your life expectancy

The larger your palm, the less you will live, on average.

Try it out - look at your neighbors and you'll see who is expected to live longer.

Associations are ??? Causations

Why?

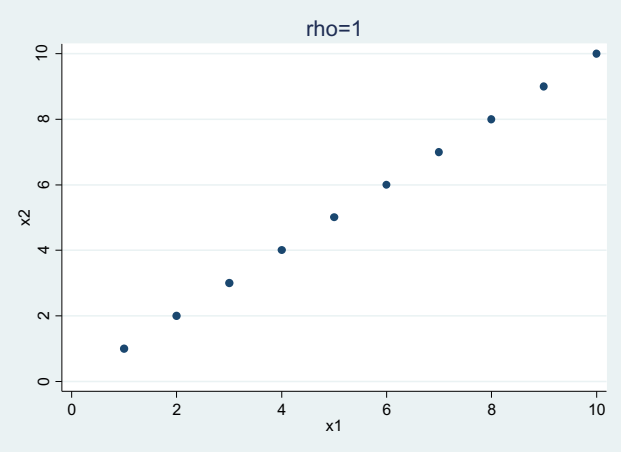
Women have smaller palms and live 6 years longer on average

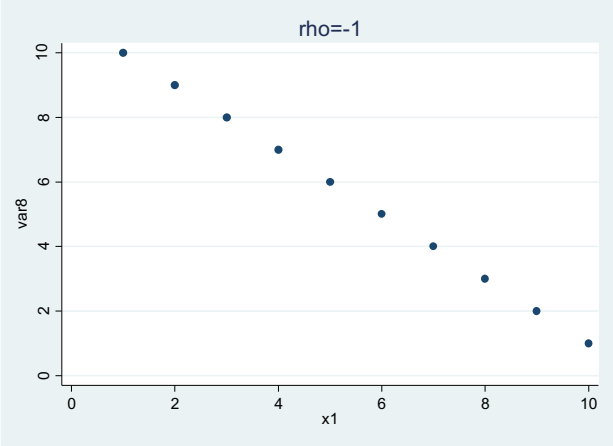
Don’t stop at the first association.
Ask yourself (and the data) WHY?

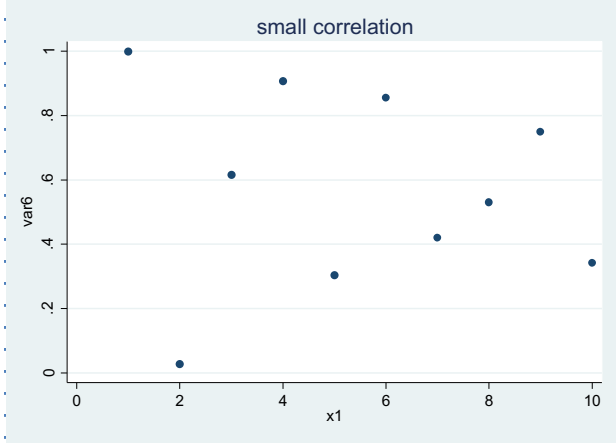
Correlation

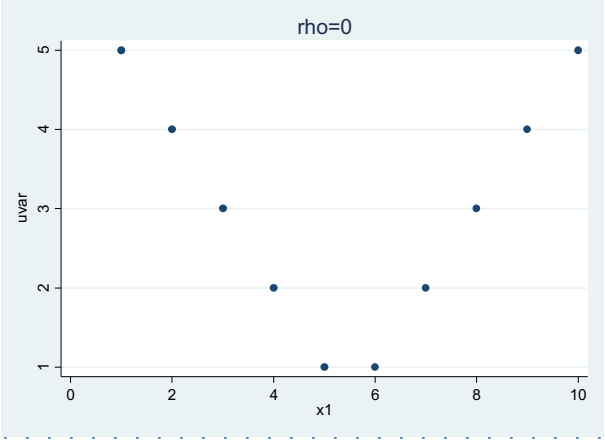
- Correlation is a examines the relationship between 2 (usually) continuous variables
  - Does one increase with the other?
    - e.g. Does CD4 count increase with increasing age?
- Both variables are measured on the same people
- Correlation assumes a linear relationship between the two variables
- Correlation is symmetric
  - The correlation of A with B is the same as the correlation of B with A
  - Designated by ρ (rho)

Correlation









Correlation

- ρ lies between -1 and 1
- -1 and 1 are perfect correlations, 0 is no correlation
- An estimator of the population correlation ρ is Pearson's correlation coefficient denoted r
- It is estimated by:

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

Correlation: hypothesis testing

- To test whether there is a correlation between two variables, our hypotheses are

  $H_0 : \rho=0$ and $H_A : \rho\neq 0$

- We need to calculate a test statistic for r
- The test statistic is

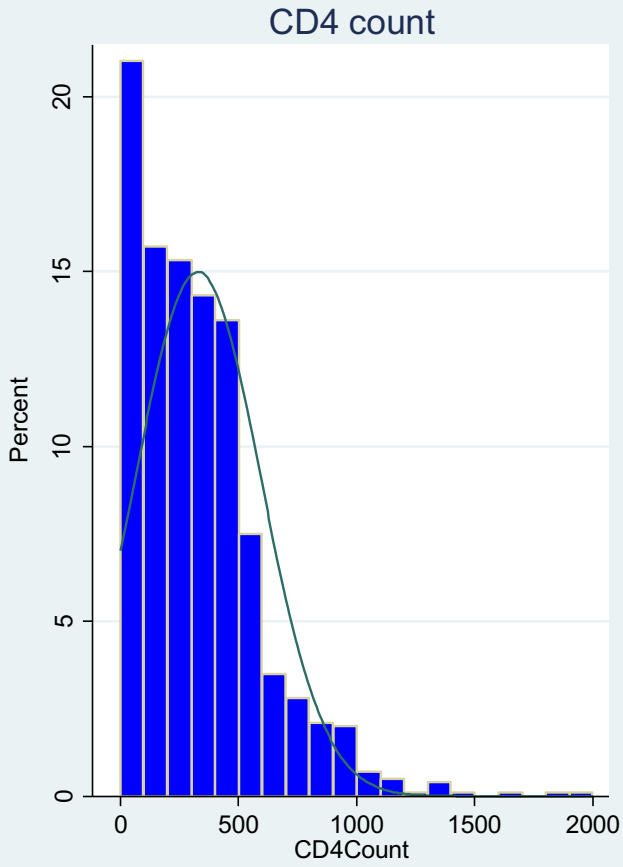
$$t = \frac{r-0}{se(r)}$$

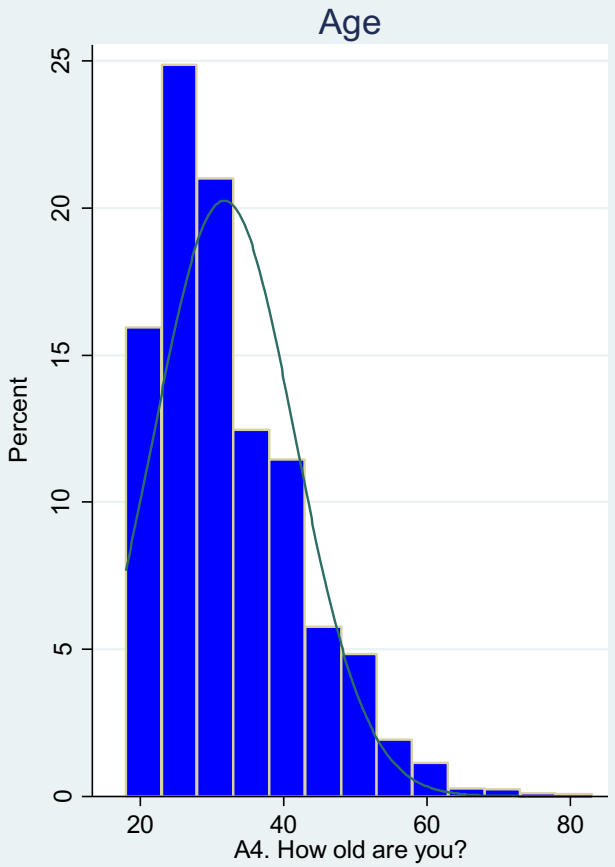
$$\text{where } se(r) = \sqrt{\frac{1-r^2}{n-2}}$$

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

Correlation: hypothesis testing

- The test statistic follows a t distribution with n-2 degrees of freedom under the null
- The assumptions
  - The pairs of observations (x_i , y_i) were obtained from a random sample
  - X and Y are normally distributed


CD4 count
Percent
0
5
10
15
20
0
500
1000
1500
2000
CD4Count


Age
Percent
0
5
10
15
20
25
20
40
60
80
A4. How old are you?

Correlation example

```
corr cd4count age
(obs=999)

             | cd4count    age_b
-------------+------------------
   cd4count  |   1.0000
       age_b |  -0.1133    1.0000
```

Note that the hypothesis test is only of $\rho=0$, no other null
Also note that the correlation is the linear relationship only

Correlation example

pwcorr var1 var2, sig obs

```
. pwcorr cd4count age, sig obs

             | cd4count    age_b
-------------+------------------
   cd4count |   1.0000
             |
             |      999
             |
       age_b |  -0.1133   1.0000
             |   0.0003
             |      999      3387
             |
```