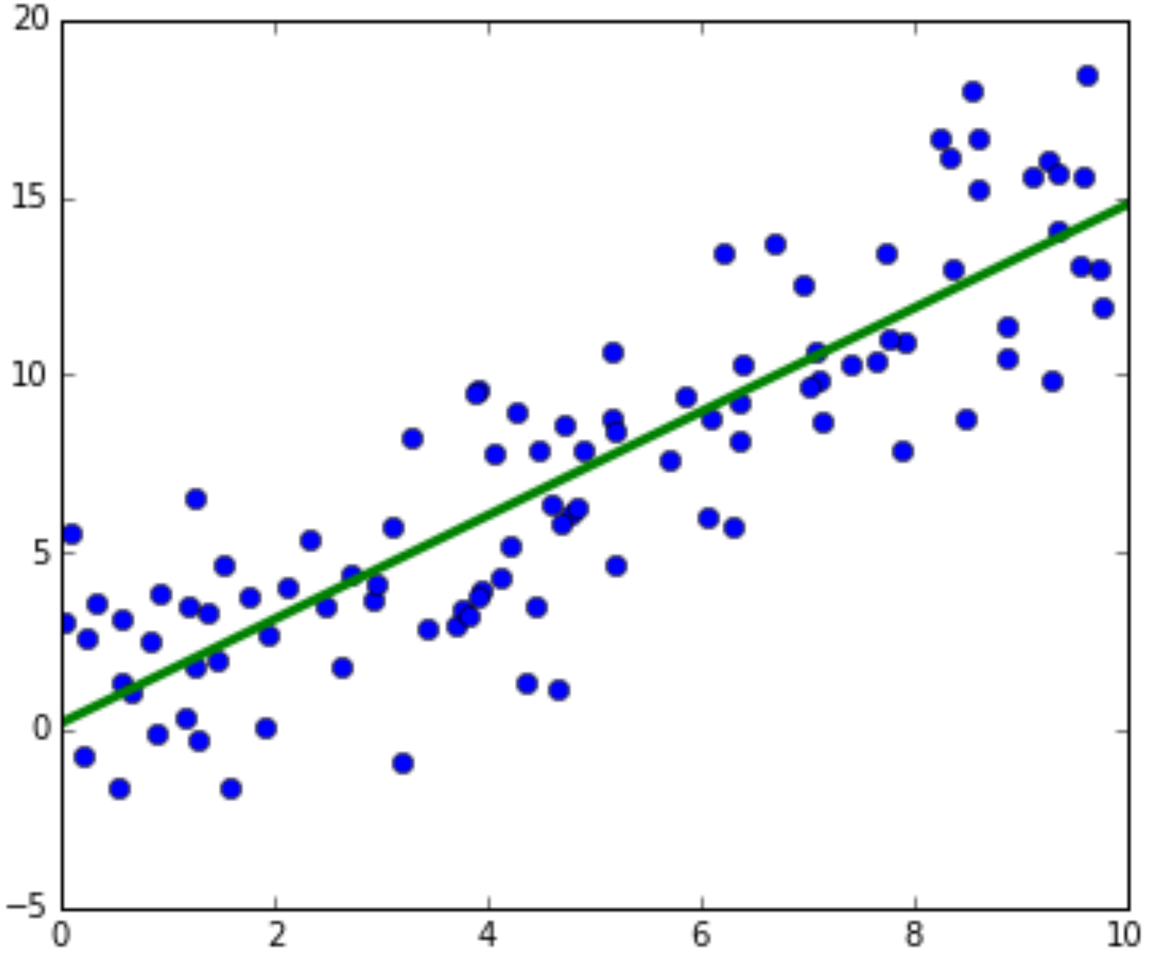
Info 2950, Lecture 22

25 Apr 2017

Prob Set 6: due Mon night 24 Apr

Prob Set 7: to be issued tonight, due mid next week

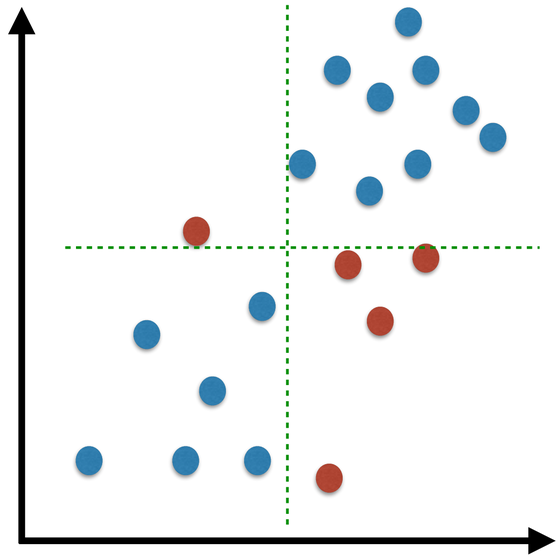
Prob Set 8: due 11 May (end of classes)


20
15
10
5
0
−5
0
2
4
6
8
10

$$Cov(X,Y) = E[(X-E[X])(Y-E[Y])]$$

$$Cov(X,X) = Var(X)$$

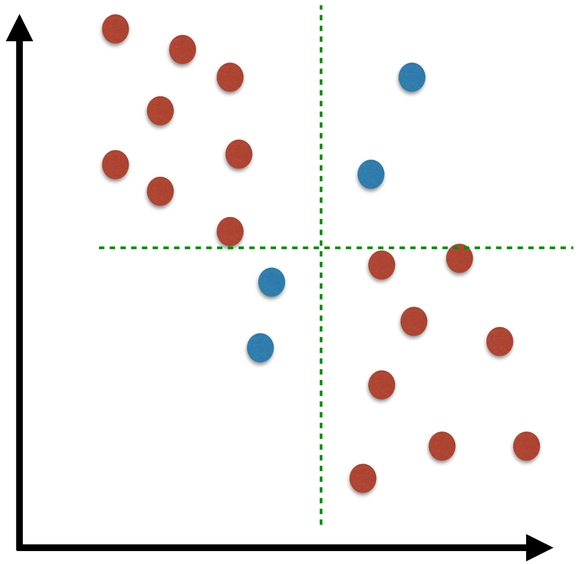
X>E[X], Y>E[Y]



X<E[X], Y<E[Y]

$Cov(X,Y) > 0$

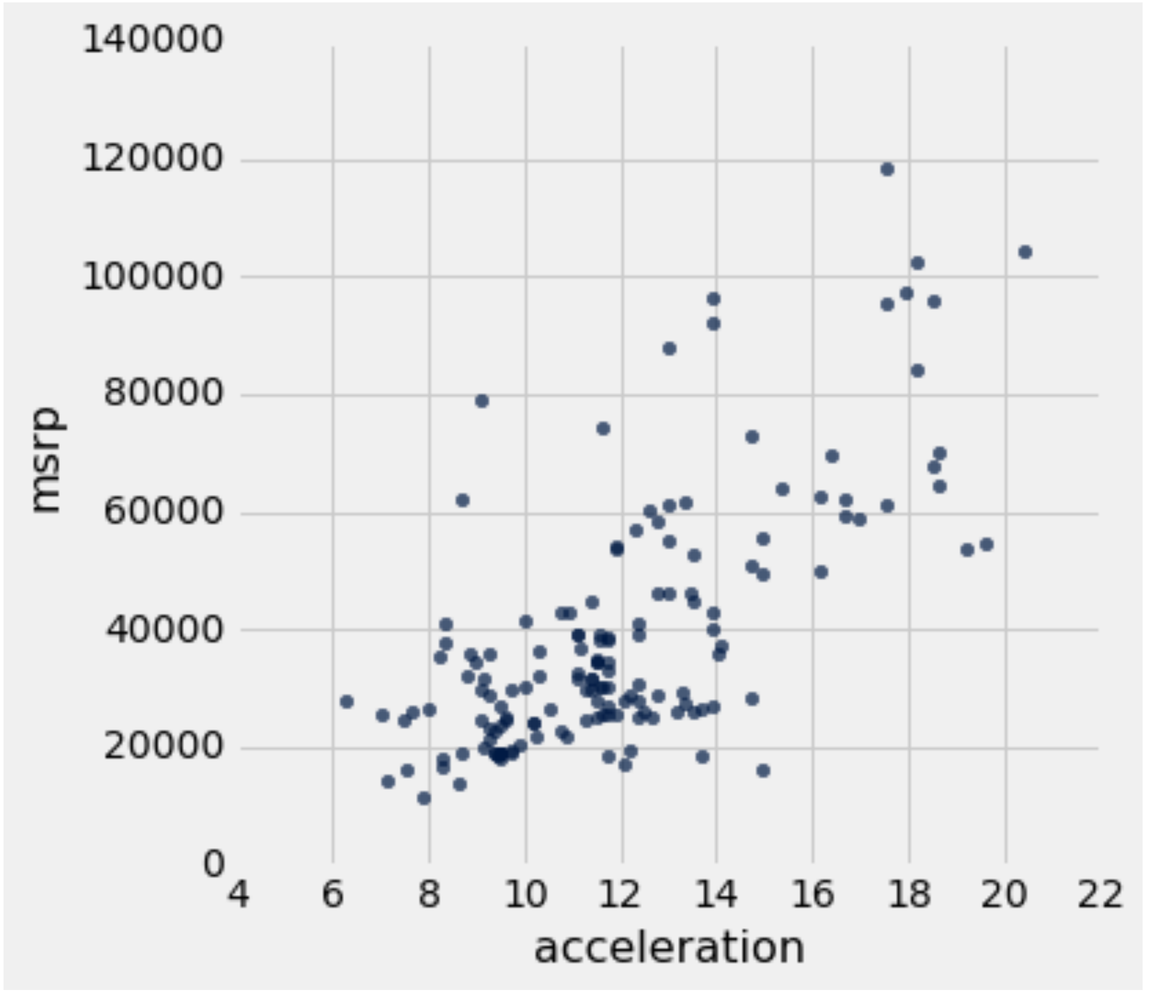
X<E[X], Y>E[Y]

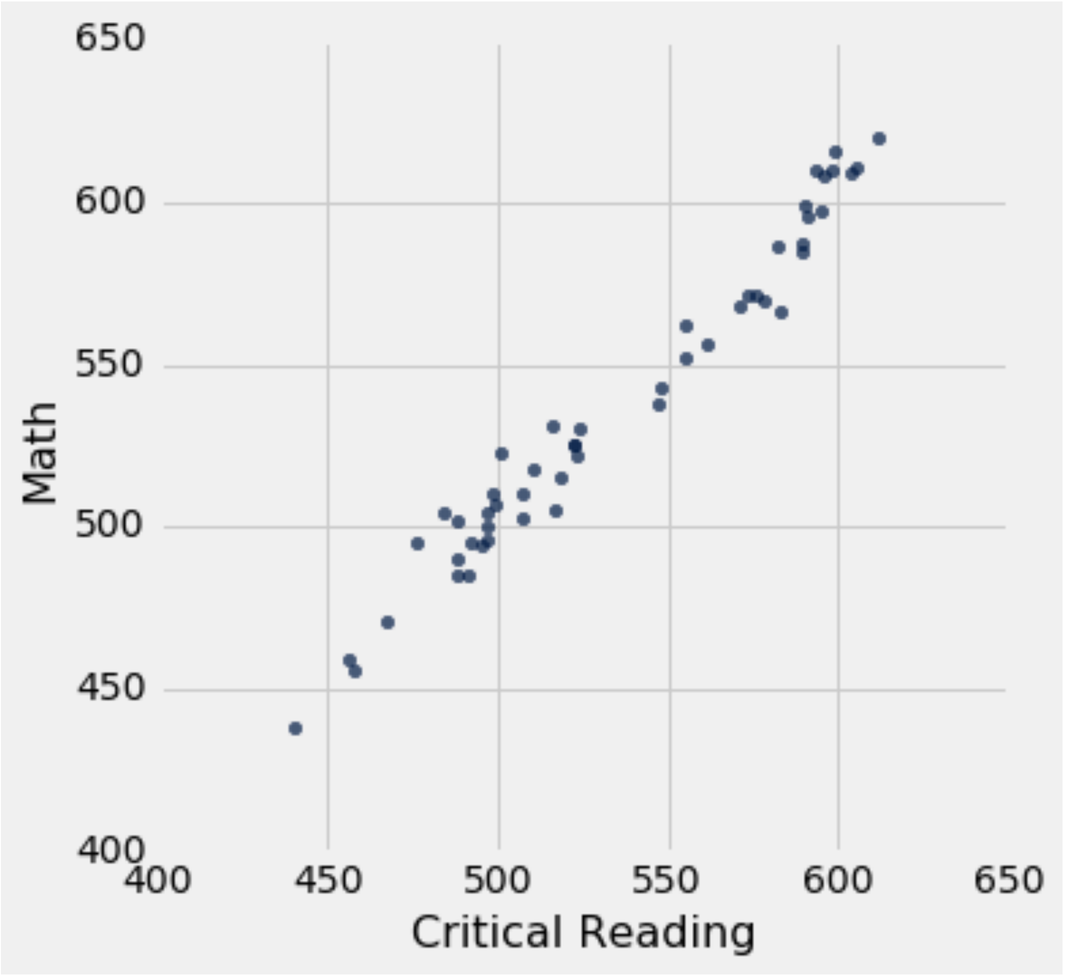


X>E[X], Y<E[Y]

$Cov(X,Y) < 0$

https://www.inferentialthinking.com/chapters/13/1/correlation.html





https://www.inferentialthinking.com/chapters/13/1/correlation.html

Correlation coefficient only measures association.
Correlation does not imply causation.

Though the correlation between the weight and the math ability of children in a school district may be positive, that does not mean that doing math makes children heavier or that putting on weight improves the children's math skills.

Age is a confounding variable: older children are both heavier and better at math than younger children, on average

https://www.inferentialthinking.com/chapters/13/1/correlation.html

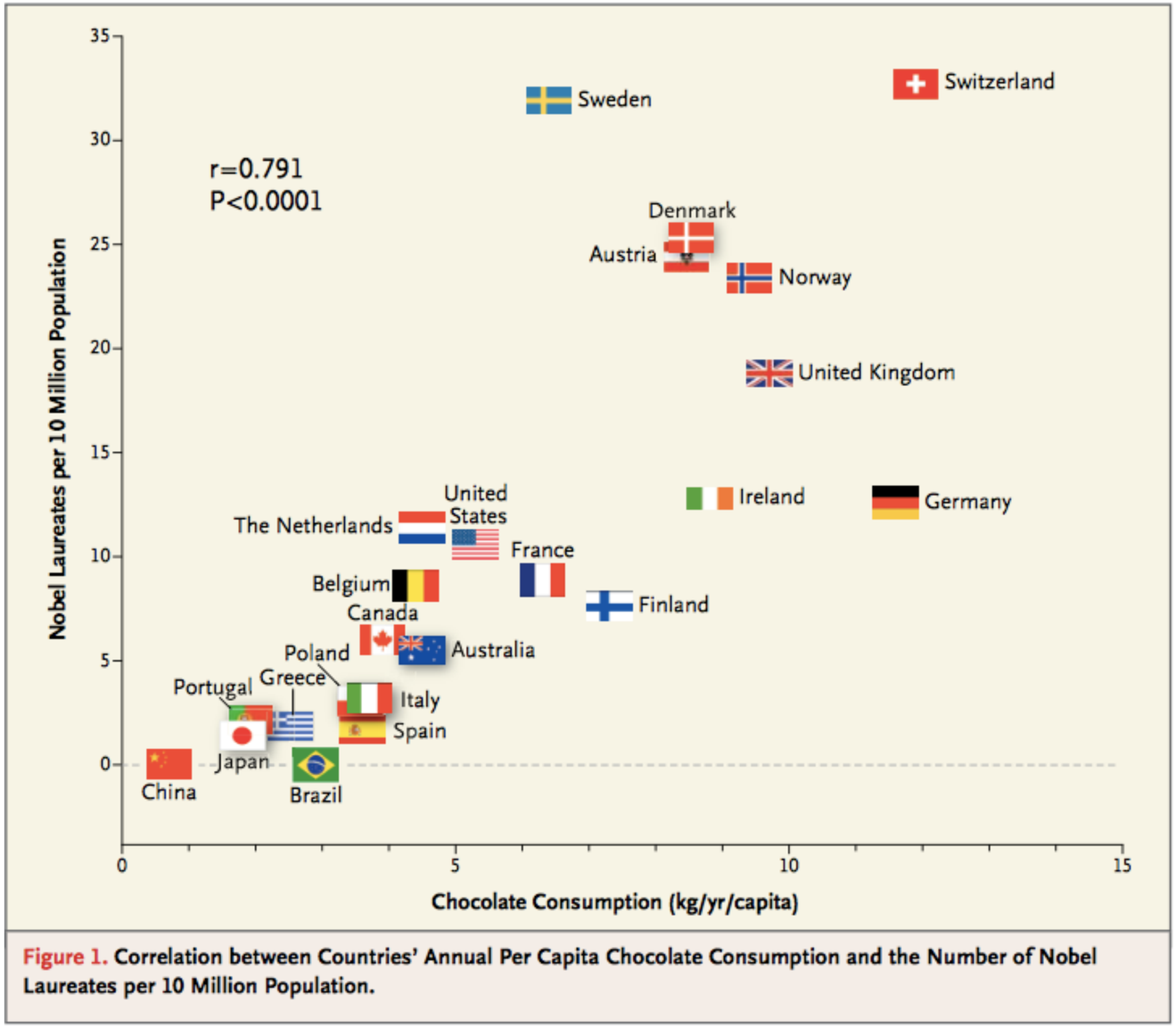


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

see also http://www.tylervigen.com/spurious-correlations

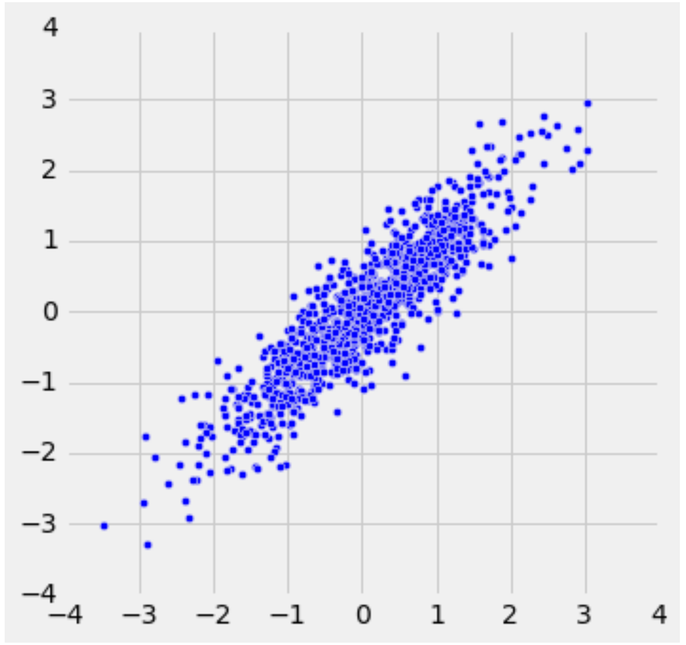
Recall that if we rescale the data, x -> x/c, then that divides the standard deviation σ[x] by the same c.

In particular, we can divide by c= σ[x], which gives a new distribution with standard deviation normalized to 1 (as is done to calculate z values).

Note that Pearsonr(x,y) = Cov[x,y] / σ[x] σ[y] is unchanged by rescaling both x and y.

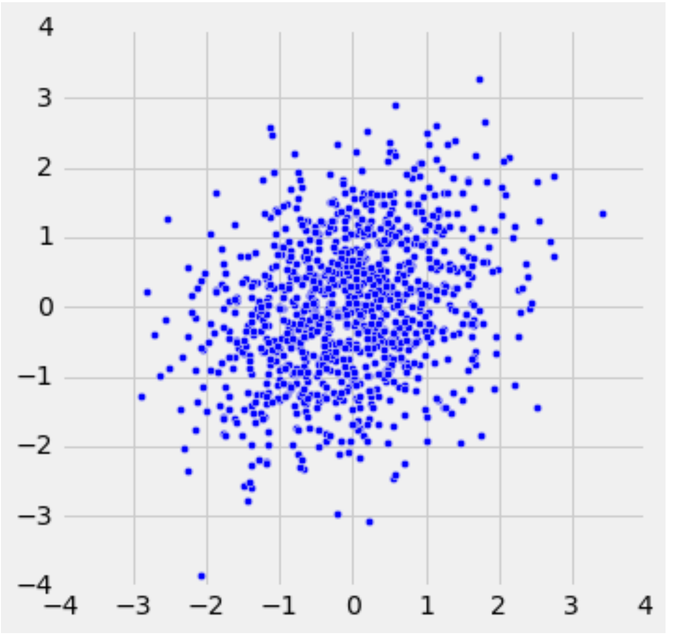
On the other hand, if we rescale x-> x/σ[x] and y-> y/σ[y] , then the linear regression slope Cov[x,y]/Var[x] -> Cov[x,y].

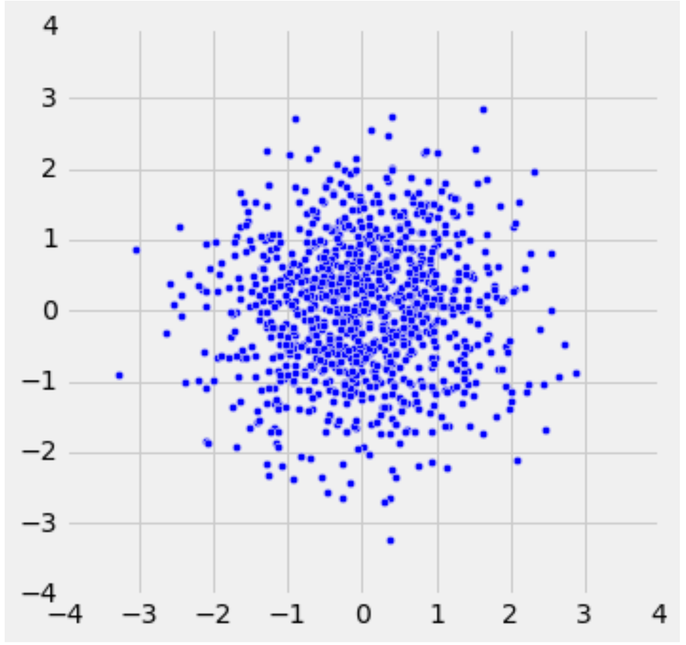
Therefore the Pearsonr(x,y) is just the linear regression slope when the variables are rescaled to have standard deviation equal to 1.


4
3
2
1
0
−1
−2
−3
−4
−4 −3 −2 −1 0 1 2 3 4

.9

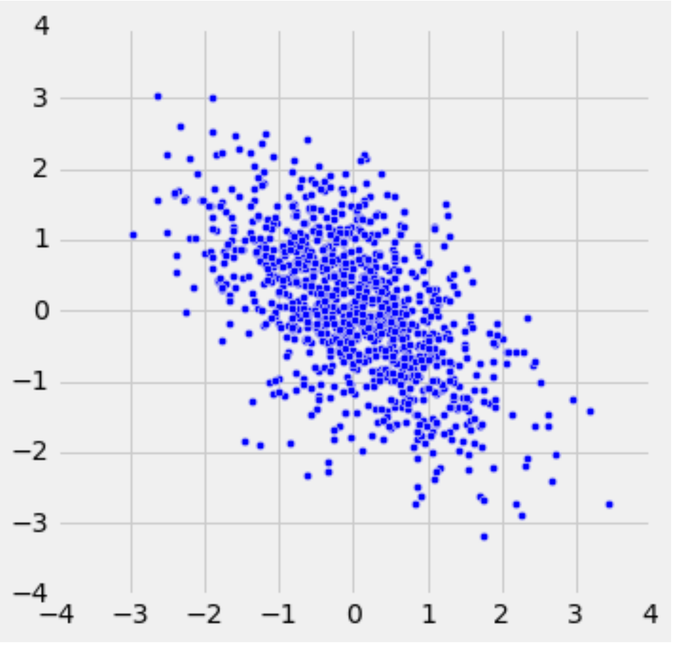
.25

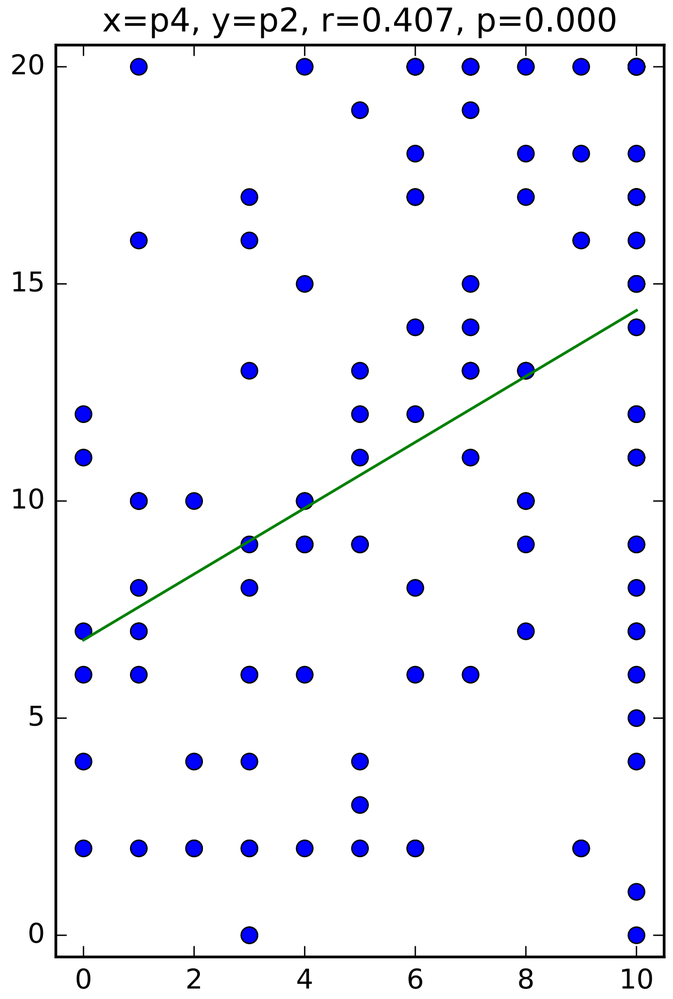

4
3
2
1
0
−1
−2
−3
−4
−4 −3 −2 −1 0 1 2 3 4

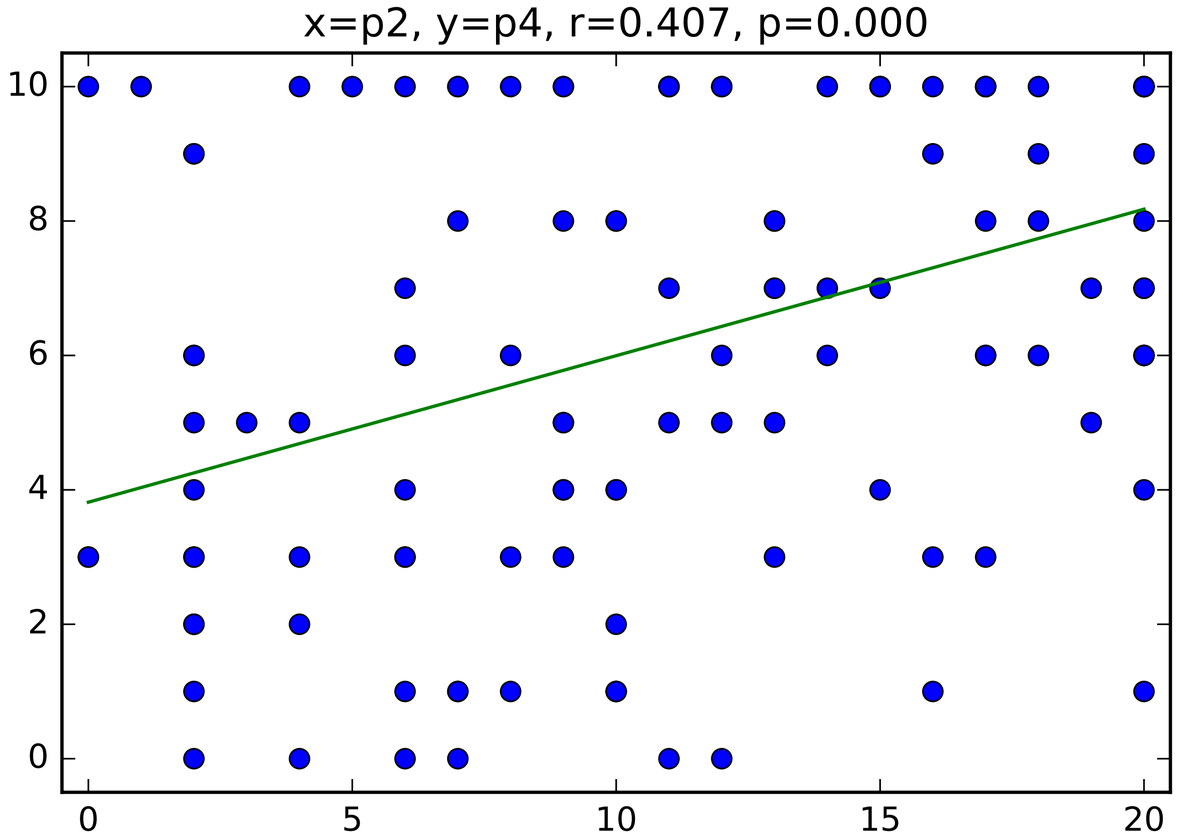

4
3
2
1
0
−1
−2
−3
−4
−4 −3 −2 −1 0 1 2 3 4

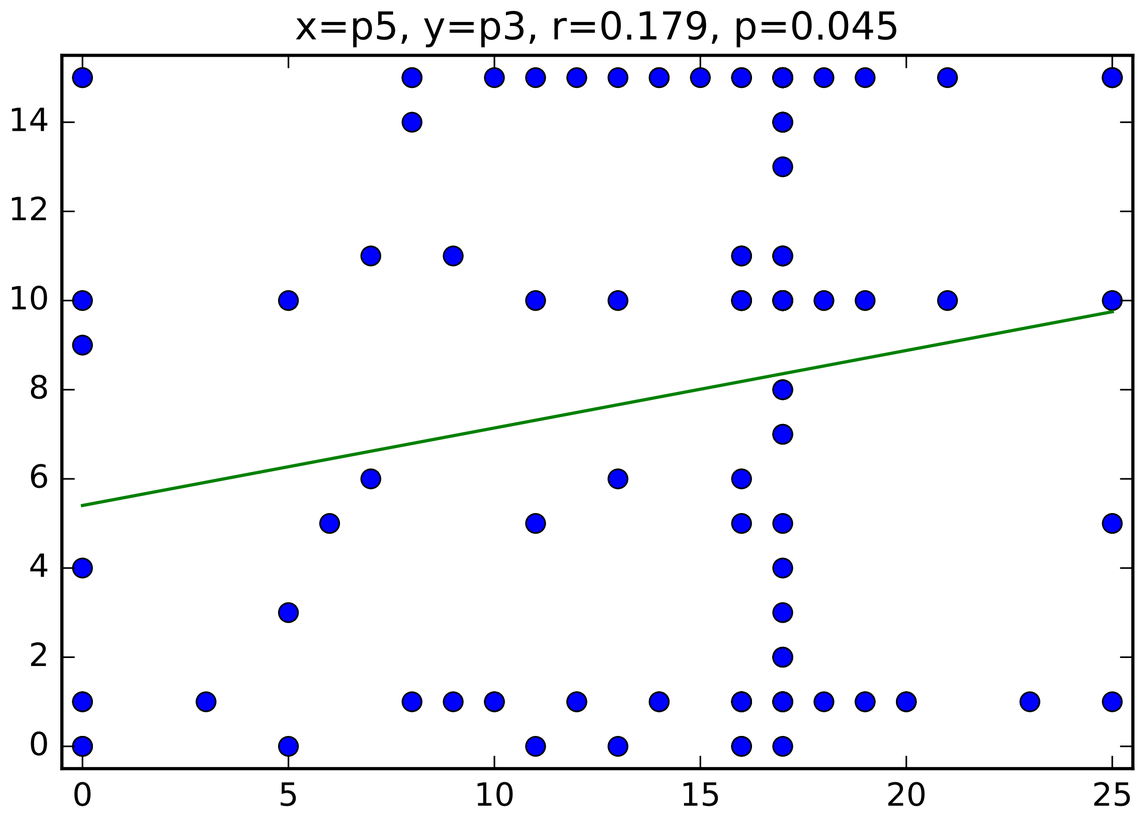
0

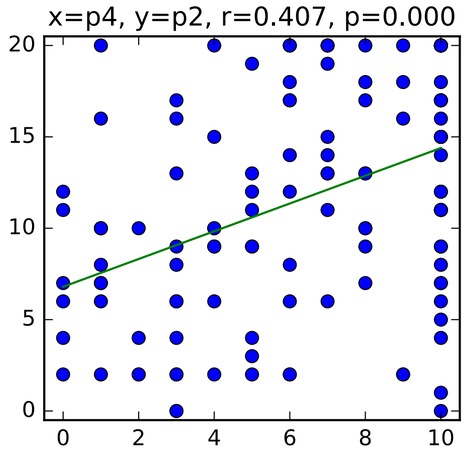
-.55

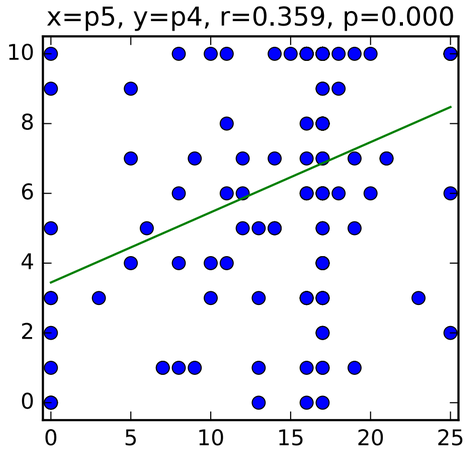

4
3
2
1
0
−1
−2
−3
−4
−4 −3 −2 −1 0 1 2 3 4

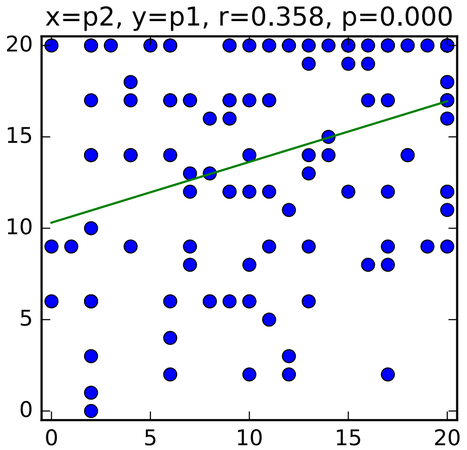

x=p4, y=p2, r=0.407, p=0.000
20
15
10
5
0
0
2
4
6
8
10

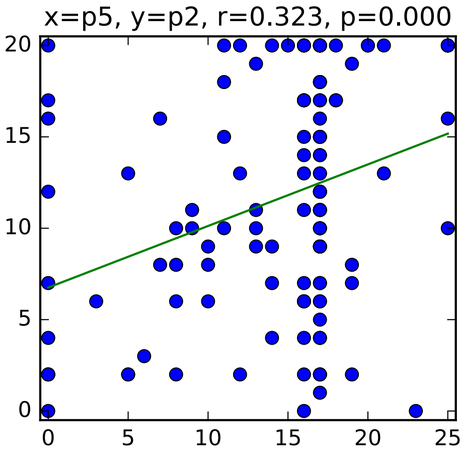

x=p2, y=p4, r=0.407, p=0.000
10
8
6
4
2
0
0
5
10
15
20

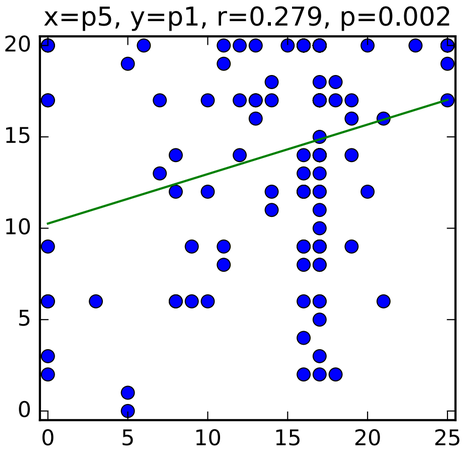

x=p5, y=p3, r=0.179, p=0.045
0
2
4
6
8
10
12
14
0
5
10
15
20
25

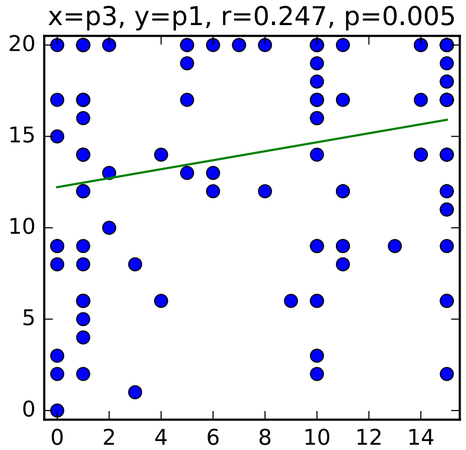


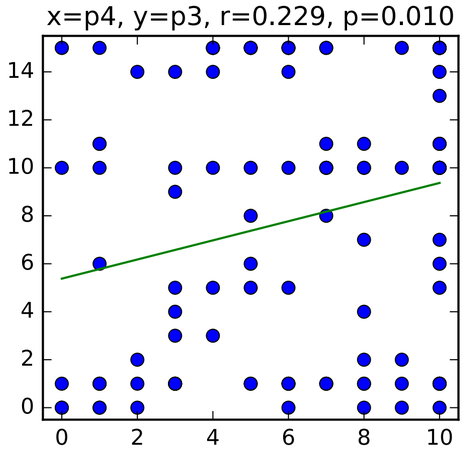


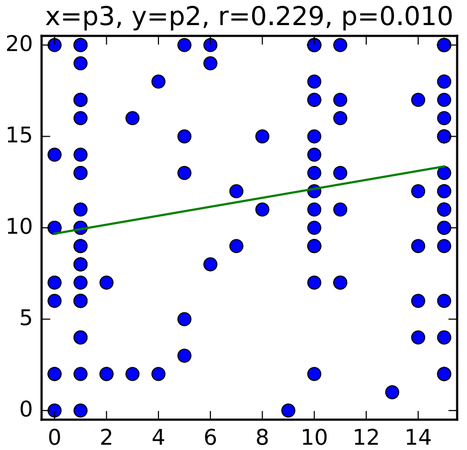


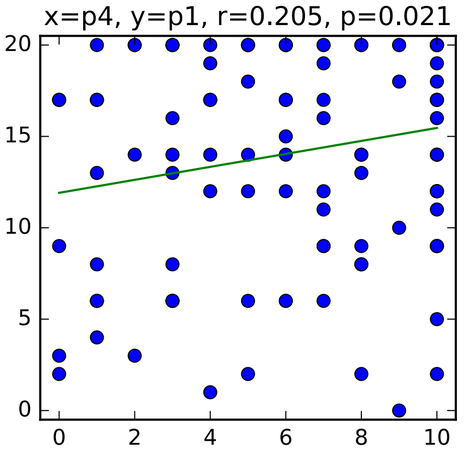


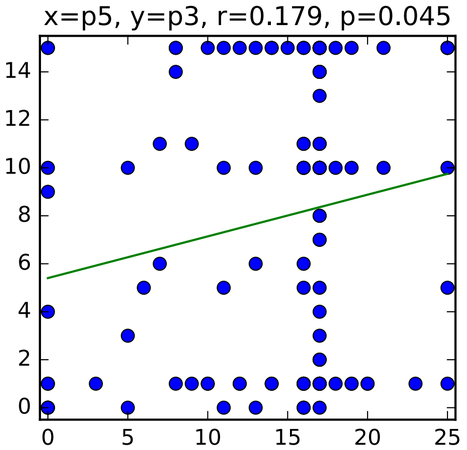






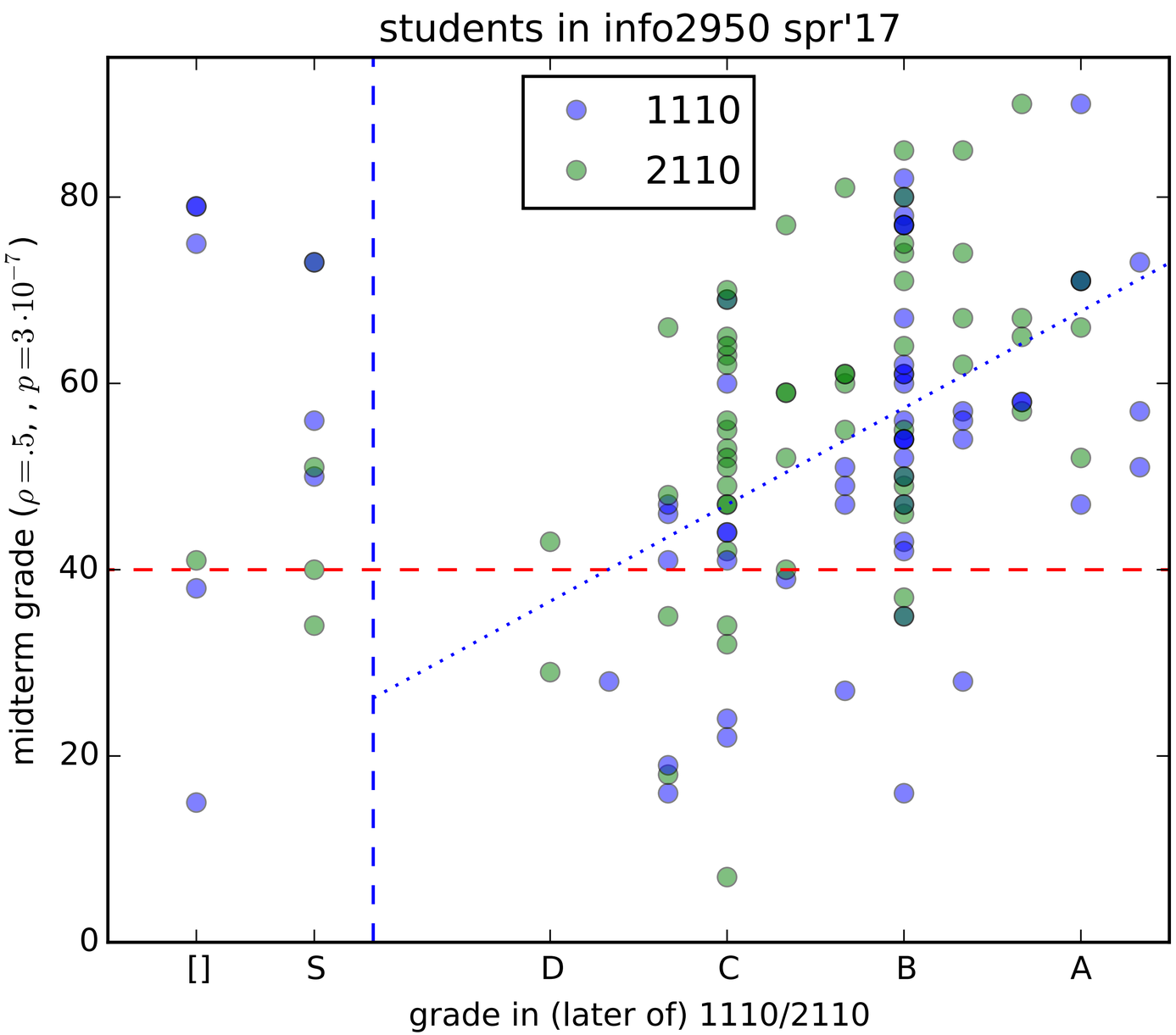




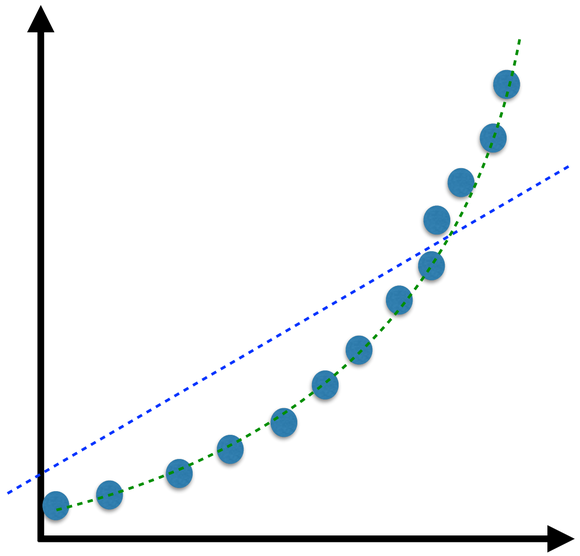


[20,20,15,10,25]

1) naive bayes, 2) simple probab, 3) wikipedia, 4) python, 5) research lit


students in info2950 spr'17
1110
2110
midterm grade ($\rho = .5$, , $p = 3 \cdot 10^{-7}$)
0
20
40
60
80
[]
S
D
C
B
A
grade in (later of) 1110/2110

Dependences are not always linear



(Except in …)

http://www.nccoast.org/uploads/documents/CRO/2012-5/SLR-bill.pdf

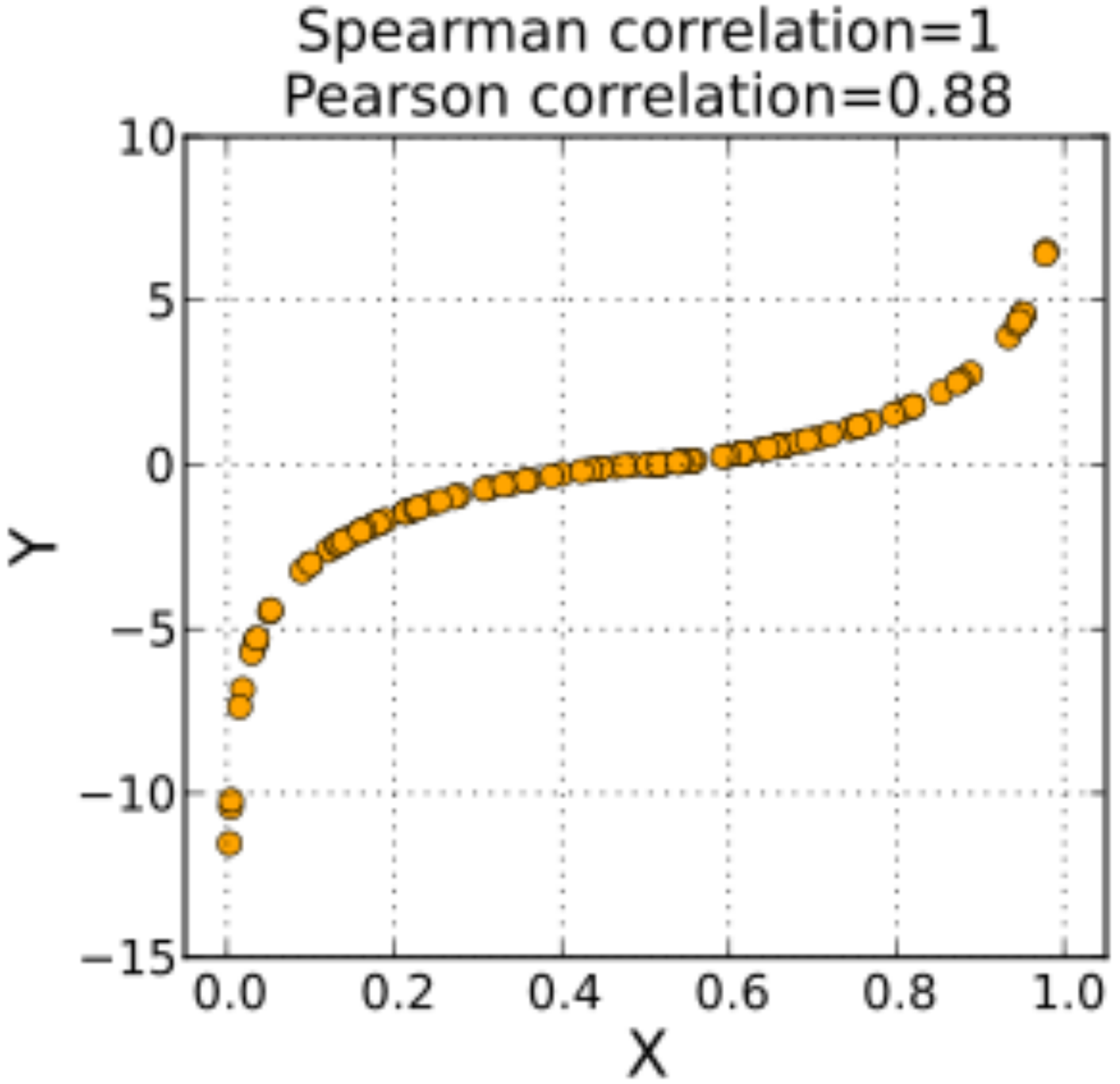
GENERAL ASSEMBLY OF NORTH CAROLINA
SESSION 2011

H **D**

HOUSE BILL 819
Committee Substitute Favorable 4/21/11
PROPOSED SENATE COMMITTEE SUBSTITUTE H819-CSLH-38 [v.18]

4/25/2012 11:05:28 AM

(e) The Division of Coastal Management shall be the only State agency authorized to develop rates of sea-level rise and shall do so only at the request of the Commission. These rates shall only be determined using historical data, and these data shall be limited to the time period following the year 1900. Rates of sea-level rise may be extrapolated linearly to estimate future rates of rise but shall not include scenarios of accelerated rates of sea-level rise. Rates of sea-level rise shall not be one rate for the entire coast but, rather, the Division shall consider separately oceanfront and estuarine shorelines. For oceanfront shorelines, the Division shall use no fewer than the four regions defined in the April 2011 report entitled "North Carolina Beach



https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

The Pearson correlation coefficient misses non-linear relationships and is also sensitive to outliers — the Spearman correlation can sometimes find correlations that Pearson misses.

It is defined as the Pearson correlation of the rank order of the data.
That means it also varies from -1 (perfectly anti-correlated) to $+1$ (perfectly correlated), with 0 meaning uncorrelated.

If the data has $x = [.6, .4, .2, .1, .5]$ then the ranks are $r = [5, 3, 2, 1, 4]$.

For data $y = [403, 54, 7, 2, 148]$, the ranks $s = [5, 3, 2, 1, 4]$ are the same.*

so the Spearman correlation is 1, whereas the Pearson is less than one.
Both functions are available in scipy.stats (as pearsonr() and spearmanr()).

[*Actually the second was generated from the first by taking the integer part of $\exp(10x)$]

Defined as the Pearson correlation for the ranks, the Spearman correlation is written

$$\rho = \frac{\mathrm{Cov}[r, s]}{\sigma[r]\sigma[s]} \ , \tag{1}$$

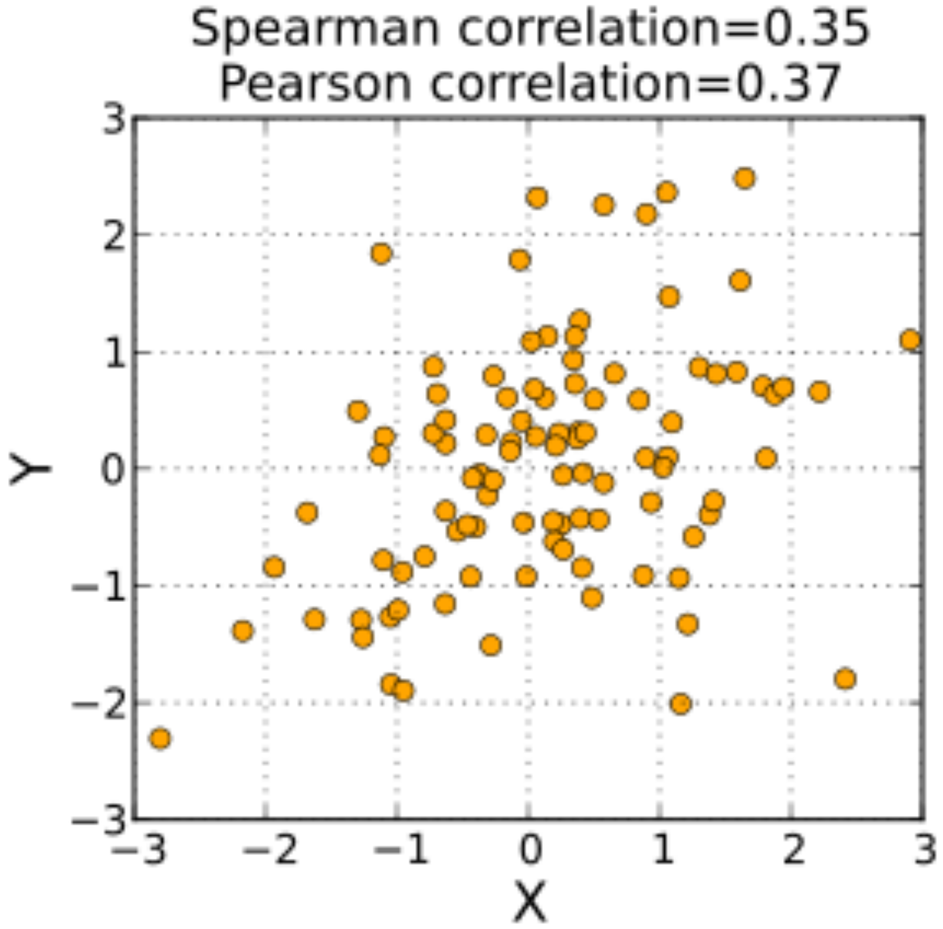
where $\mathrm{Cov}[r, s] = E[(r - E[r])(s - E[s])]$

(generalizing the $\mathrm{Var}[x] = E[(x - E[x])^2]$, with $\mathrm{Cov}[x, x] = \mathrm{Var}[x]$).

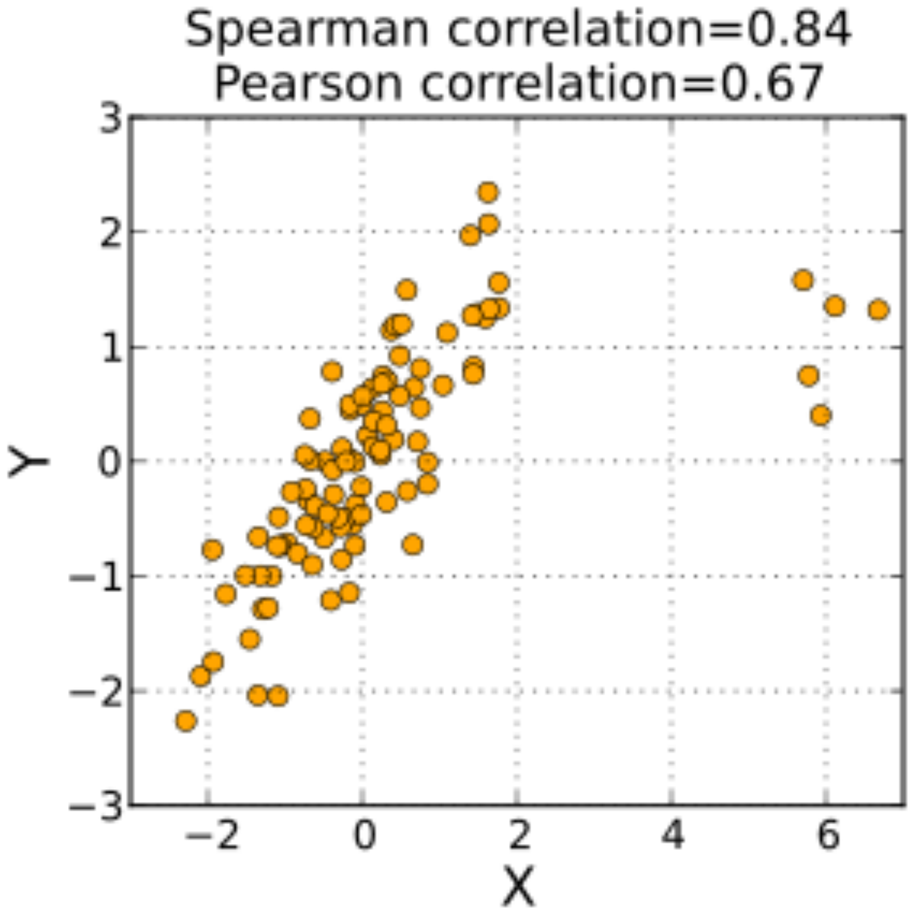
The formula for the Spearman correlation coefficient is given at
http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient
in terms of the difference $d_i = r_i - s_i$ between ranks, in this easily calculable form:

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

It is straightforward to verify that (1) reduces to (2) (see linked notes)



roughly elliptically distributed and there are no prominent outliers: same



less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples (limited to value of rank)

https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Slate

The Democrats Show Some Spunk

Two days of obstruction on Capitol Hill.

By *Jim Newell*

dynamically generated, so scrape with selenium

We are on Day 12 of the Trump administration, and Senate Democrats have begun slowing the pace of Capitol Hill down to a crawl.

JIM NEWELL

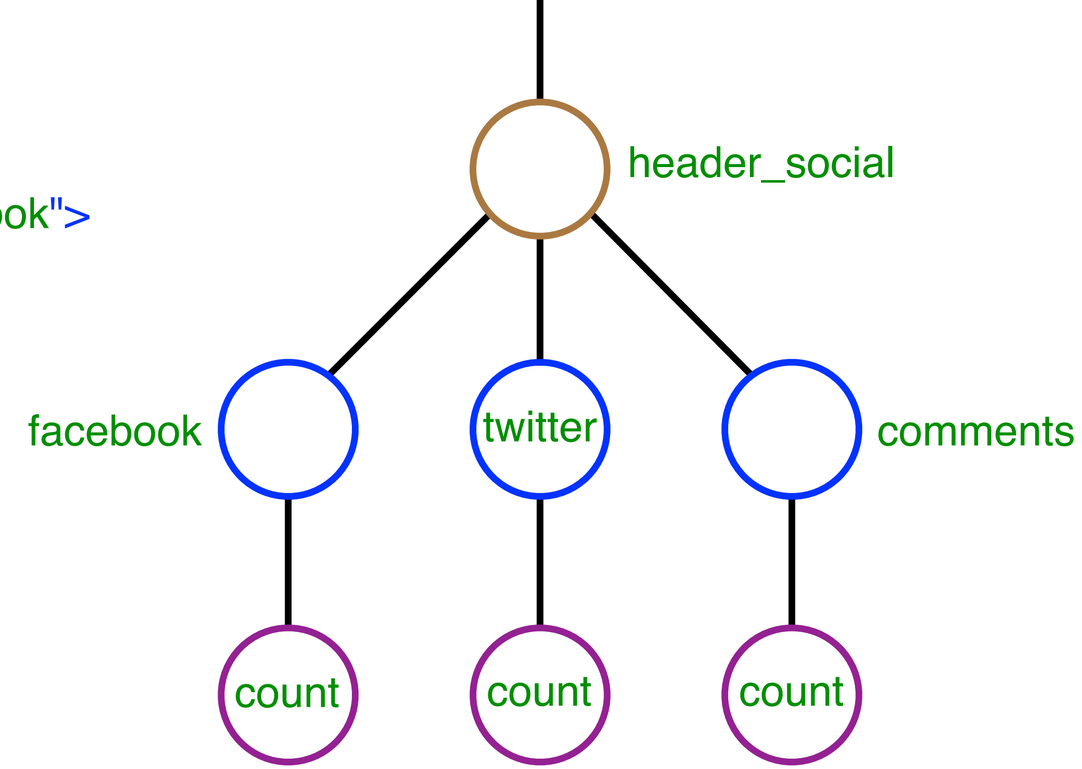
It started Monday night with a surprise procedural move, just as Congress was returning to session. The Senate Finance Committee

```
<div class="social social-with-popup sharing-buttons header_social">

  <a href="#" data-share="http://www.facebook.com/..." class="facebook">
   <span class="icon"></span>
   <span class="count">8.9k</span>
  </a>

  <div class="twitter">
   <a href="#" data-share="http://twitter.com/..." class="twitter">
     <span class="icon"></span>
   </a>
   <a class="tw-count" href="http://twitter.com/search?q=...">
       <span class="count">2.1k</span>
   </a>
  </div>

   <a href="#comments" class="comments">
    <span class="icon"></span>
    <span class="count">1.4k</span>
   </a>
</div>
```




```
allsoc=[]
for url in urls:
  driver.get(url)
  social = driver.find_element_by_class_name('header_social')
  soc=[] #temporary list for this url
  for cl in ('facebook','twitter','comments'):
    elt = social.find_element_by_class_name(cl)
    subelt = elt.find_element_by_class_name('count')
    soc.append(subelt.get_attribute('innerHTML'))
allsoc.append(soc)
```