DA503 Applied Statistics

Lecture 08

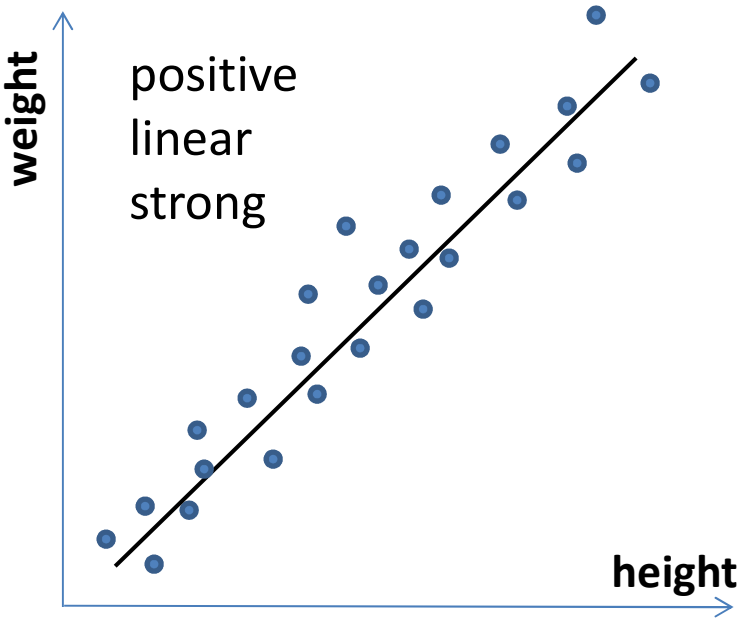
Correlation Between Variables

Introduction

- Examining relationships between two variables
- Three possibilities exist:
  - Both variables are continuous (numerical)
    - How one variable, called a dependent variable, changes in relation to changes in the other variable, called the independent variable
  - One variable is categorical, and the other is continuous
    - Is there an association between a continuous variable and different levels of a categorical variable (side-by-side boxplots provide a good visual aid)?
  - Both variables are categorical
    - Is there an association (dependency) between different levels of 2 categorical variables?

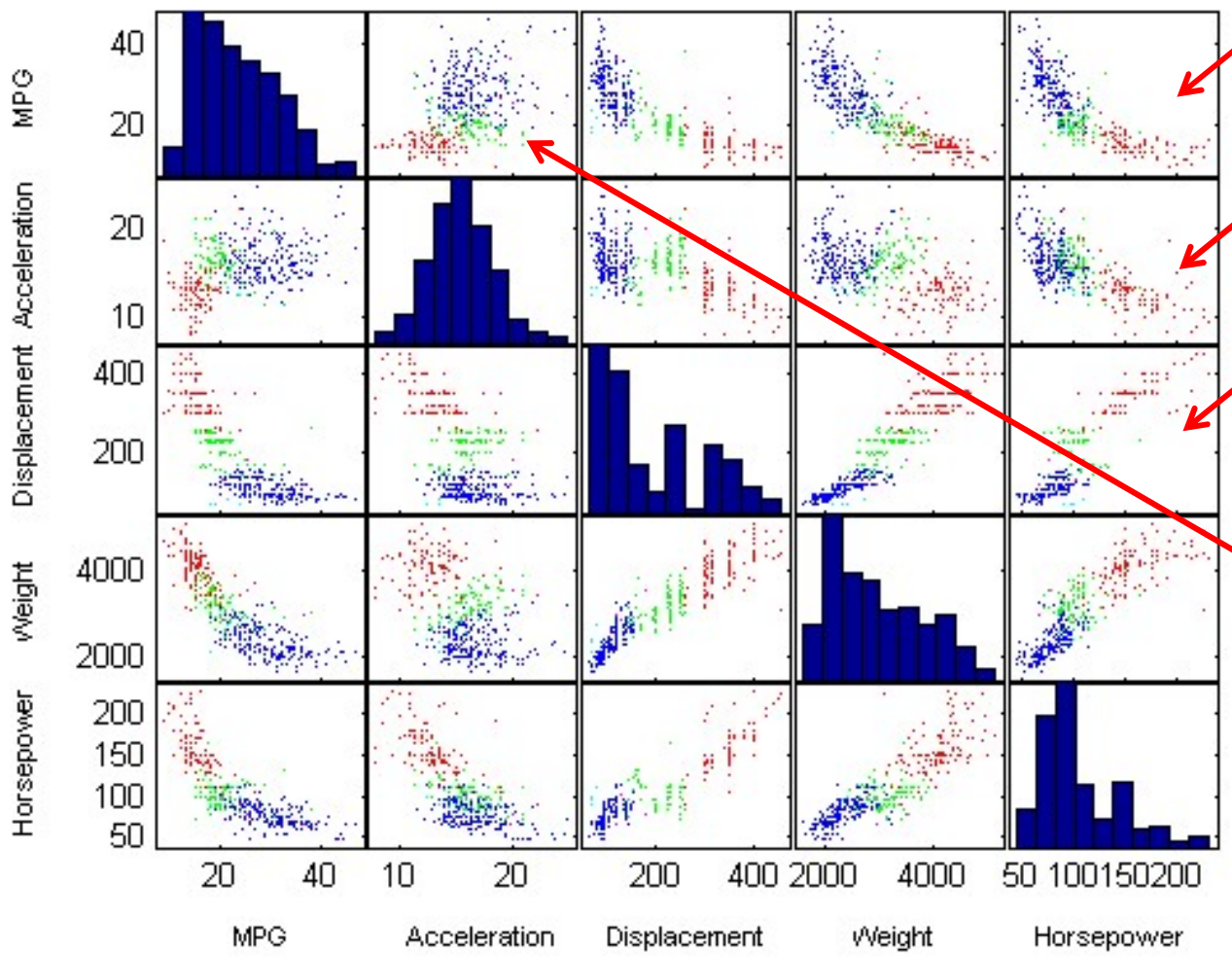
Correlation between 2 continuous variables

- We want to examine the relationship between two numerical values:
  - Between target vs predictor or predictor vs predictor
  - Is there any **correlation** between the two?
- In summarizing the relationship between two quantitative variables, we need to consider:
  - Association/Direction (i.e. positive or negative)
  - Form (i.e. linear or non-linear)
  - Strength (weak, moderate, strong)



Continuous vs Continuous

- Relationship between two variables (target-predictor and predictor-predictor)



form: nonlinear

direction: negative

direction: positive

strength: weak

But how do we measure strength?

Pearson's correlation

- **Pearson's correlation** is a statistical measure of the strength of a linear relationship between paired data (**a parametric test**)
- Assumptions
  - Two variables have to be measured on either an interval or ratio scale (don't have to be of the same type or unit)
  - Linearly related data (the relation between the two variables is linear – can be checked via a scatterplot)
- Pearson's correlation determines the degree to which a relationship is linear. So, you should not pursue a Pearson's correlation to determine the strength and direction of a relationship when you already know the relationship is not linear.

Covariance

- A measure of how changes in one variable are associated with changes in the second variable (to what degree these two variables are linearly associated).
- Similar to variance, but variance tells you how a single variable varies whereas covariance tells you how two variables vary together.
- A positive value for covariance indicates a direct or increasing linear relationship. A negative value indicates a decreasing relationship.

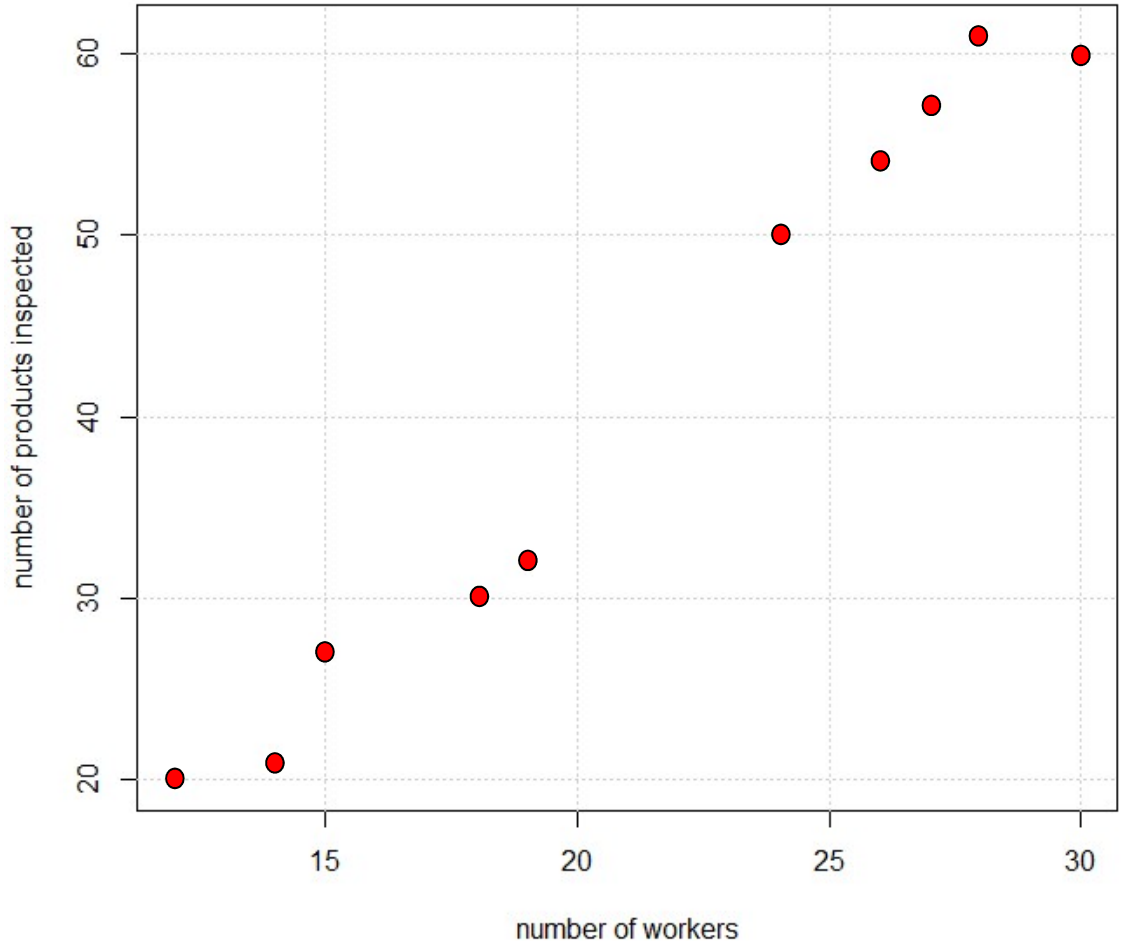
$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$ Population covariance

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = E[(X - \mu_x)(Y - \mu_y)]$$ Sample covariance

Covariance – cont'd

- **Example:** In a small factory, we have data for the number of workers (x) and the number of products (y) inspected for quality in a 30 minute time period:

#	x	y
1	12	20
2	30	60
3	15	27
4	24	50
5	14	21
6	18	30
7	28	61
8	26	54
9	19	32
10	27	57



From Statistics 101: Understanding Covariance, by Brandon Foltz

Covariance – cont'd

- Let's compute the covariance in our example:

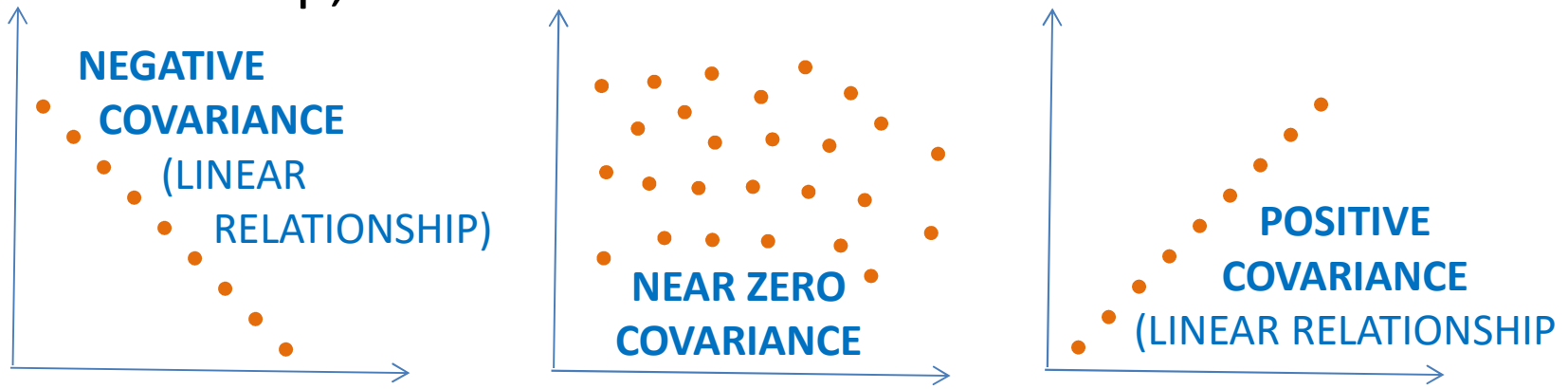
```
x = [12,30,15,24,14,18,28,26,19,27]
y = [20,60,27,50,21,30,61,54,32,57]
np.cov(x,y)
array([[ 42.01, 106.93],
       [106.93, 278.40]])
```

Var(x)

Var(y)

(sign for cov(x,y) is positive, so a positive covariance)

- A zero (or near zero) covariance tells us that these two variables are not related at all. Covariance, however, does not give us a feeling about how strong or weak the association between the variables. For the strength of that relationship, we use the correlation coefficient.



Pearson's correlation

- **(Pearson's) Correlation Coefficient**:

$$\rho = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

where:

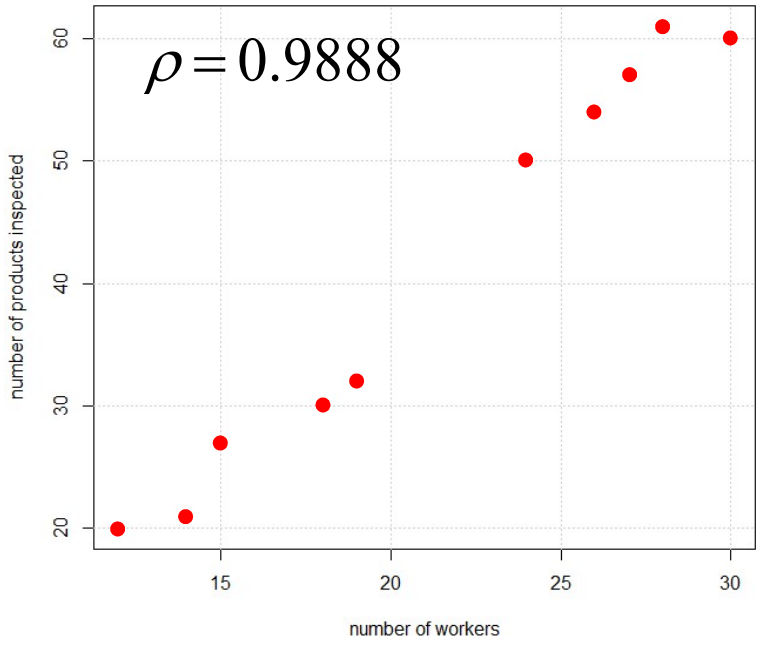
$\mathbf{S_{xy}}$: covariance of X & Y

$\mathbf{S_x}$: std deviation for X

$\mathbf{S_y}$: std deviation for Y

(all for the sample)

- Interpreting correlation:
  - Bounded between [-1 , +1]
  - Value of $\boldsymbol{\rho}$ doesn't depend on the units of X and Y
  - $\boldsymbol{\rho}$ has the same sign as $\boldsymbol{S_{xy}}$
  - Values of $\boldsymbol{\rho}$ closer to -1 or +1 indicate a strong linear relationships between the two random variables



Pearson's correlation – cont'd

```
from scipy.stats.stats import pearsonr
print('Correlation coefficient:', pearsonr(x,y))
>>> Correlation coefficient: (0.98877, 6.8586e-08)
```

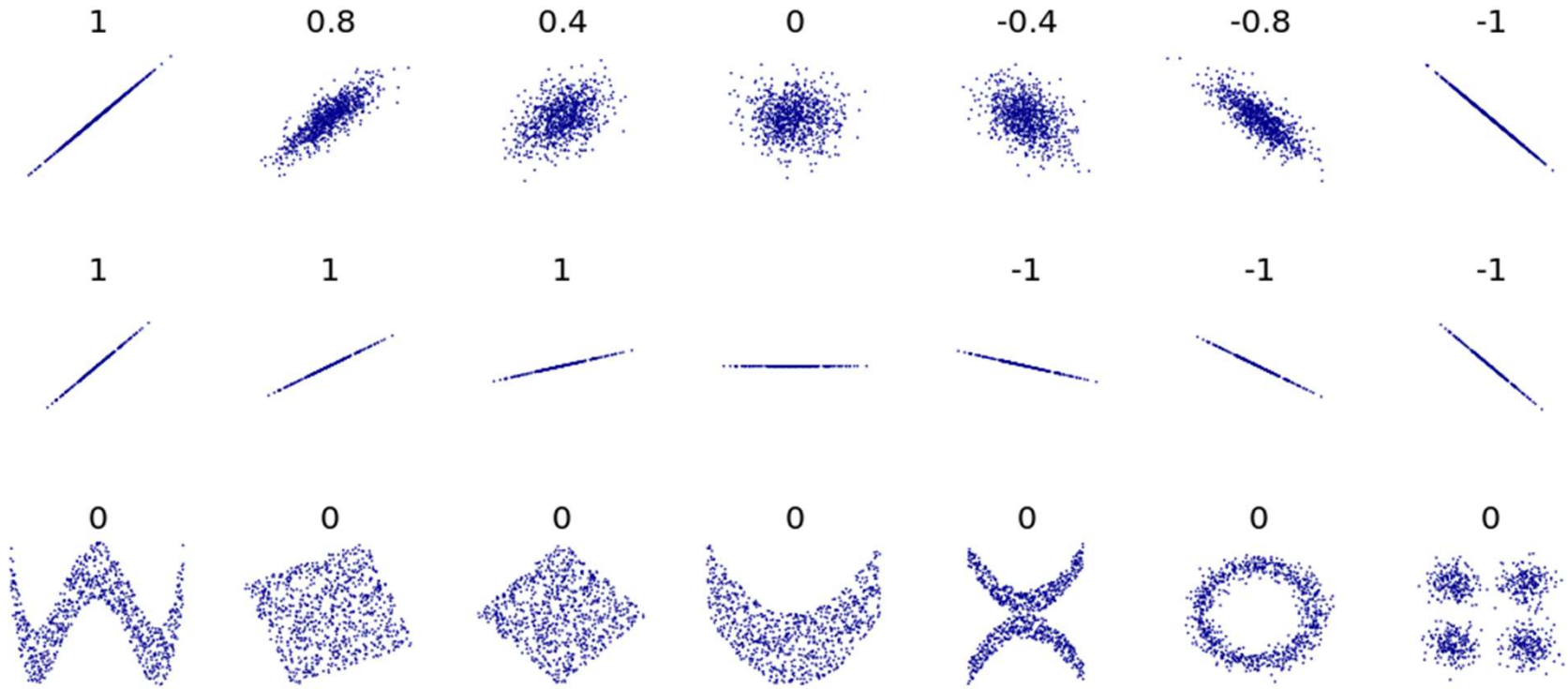
```
import numpy as np
np.corrcoef(x, y)[0,1]
>>> 0.98877
```

ρ	strength
0.80 - 1.00	Very strong
0.60 - 0.79	Strong
0.40. - 0.59	Moderate
0.20 - 0.39	Weak
0.0 - 0.19	Very weak

- Correlation tests, the Null hypothesis claims there is no correlation between the two variables. With a p value smaller than α , we can say we have evidence to reject the Null. For the above problem, a p value nearly zero tells us that, under the Null Hypothesis, probability of seeing a correlation coefficient as extreme as the found in our data set is very slim.

Variability in a correlation

- Different values of correlation coefficients and associated scatter plots

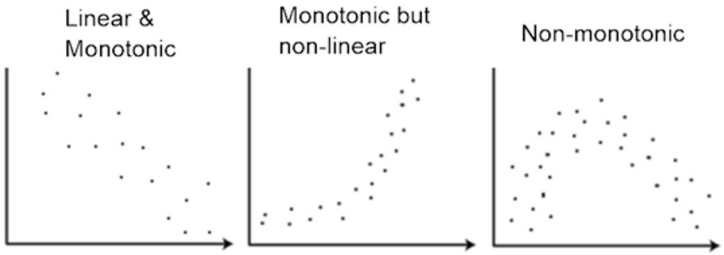


Examples of datasets with a range of correlations

Image credit: https://commons.wikimedia.org/wiki/File:Correlation_examples2.svg

Spearman's rank-order correlation

- **Spearman's** rank-order **correlation**: A statistical measure of the direction & strength of a **monotonic relationship** between paired data (**a non-parametric test**).
- Monotonic relationship:
  - Variables change together, but not in a linear way



- Monotonic relationship is not an assumption, but the measure will be low if the relationship is non-monotonic
- **Assumptions**: Variables are interval, ratio, or **ordinal**
- Spearman's correlation is used when:
  1. At least one variable is ordinal,
  2. Two variables are related, but not linearly
  3. There are significant outliers (insensitive to outliers as it's based on ranks)

Spearman's correlation – cont'd

- **Spearman's** rank-order correlation:
- There are two methods to calculate Spearman's correlation depending on whether:

1. Your data does not have tied ranks:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

d_i = difference in paired ranks, n = number of cases

2. Your data has tied ranks:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where *i = paired score*

Spearman's correlation – cont'd

- **Hypothesis**: Males with more testosterone are more aggressive

(3+4+5) / 3 = 4

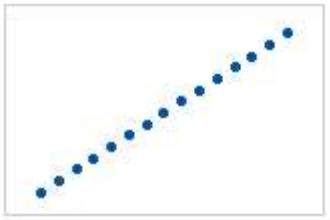
Participant	Testosterone	Rank	Aggression	Rank
1	**5**	4 (3)	**4**	2.5 (3)
2	**6**	6	**10**	6
3	**5**	4 (4)	**7**	5
4	**4**	2	**4**	2.5 (2)
5	**3**	1	**2**	1
6	**5**	4 (5)	**6**	4

```
from scipy.stats import spearmanr
x = [5, 6, 5, 4, 3, 5] ; y = [4, 10, 7, 4, 2, 6]
coef, p = spearmanr(x, y)
print('Spearman\'s coeff:', coef)
print('p-value          :', p)
Spearman's coeff: 0.8932596
p-value         : 0.0164822
```

Pearson's correlation on ranks yields the same results

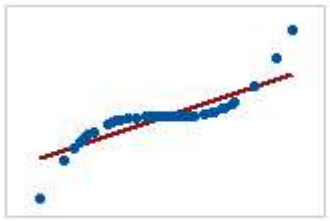
Comparison of Pearson & Spearman

- Coefficients can range in value from −1 to +1 for both



Pearson = +1, Spearman = +1

when one variable increases then the other variable increases by a consistent amount



Pearson = +0.851, Spearman = +1

one variable increases when the other increases, but the amount is not consistent

Pearson = −0.093, Spearman = −0.093

when a relationship is random or non-existent, then both correlation coefficients are nearly 0

Correlation between cont. and cat. variables

- **Point-Biserial Correlation**
  - A special form of Pearson's correlation. It measures to what degree a **continuous** variable and a **binary categorical** (dichotomous) variable move together in a linear fashion
- Example: Test score vs Gender

To calculate the correlation coef., we can just apply the Pearson's correlation to this pair:

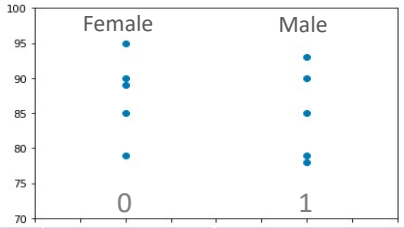
```
score = [90,85,78,93,79,85,79,89,90,95]
gender=[1,1,1,1,1,0,0,0,0,0]
np.corrcoef(score,gender)
>>> -0.2249
```

Or we can use the Scipy library:

```
score = [90,85,78,93,79,85,79,89,90,95]
gender=[1,1,1,1,1,0,0,0,0,0]
scipy.stats.pointbiserialr(score,gender)

PointbiserialrResult(correlation=-
0.22490810920, pvalue=0.5321537745)
```

p-value > 0.05, correlation is not statistically significant!



Score	Gender	Gender (num)
90	Male	1
85	Male	1
78	Male	1
93	Male	1
79	Male	1
85	Female	0
79	Female	0
89	Female	0
90	Female	0
95	Female	0

Correlation between categorical variables

- A Chi-square test tells us if a statistically significant relationship exists between categorical variables, but it doesn't tell us the strength of that relationship
- We have two approaches (among others) for computing the strength of the correlation:
  - **Phi** correlation: For dichotomous (binary) variables
  - **Cramer's V** correlation: For multi-categorical (with 2 or more categories) variables
- **Phi Coefficient**: A measure of association between two binary variables.
  - Phi coefficient (aka Matthews correlation coefficient) is a symmetrical statistics, i.e., the dependent and independent variables are interchangeable.

Correlation between categorical variables– cont'd

- **Phi Correlation coefficient** (ϕ) for a 2×2 contingency table where A, B, C, and D represent the observation frequencies:

$$\Phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Variable2 \ Variable1	True	False
True	**A**	**B**
False	**C**	**D**

Note that the Pearson correlation for two dichotomous variables is the same as the Phi coefficient, thus the interpretation is similar (range [-1 , 1])

→ 0 : no relationship

→ ±1 : perfect relationship

- **Interpreting the strength of the relationship**:

Phi correlation coefficient		Strength
-1.0 to -0.70	0.70 to 1.0	Very strong
-0.69 to -0.40	0.40 to 0.69	Strong
-0.39 to -0.30	0.30 to 0.39	Moderate
-0.29 to -0.20	0.20 to 0.39	Weak
-0.19 to -0.01	0.01 to 0.19	None / Negligible

Correlation between categorical variables– cont'd

- **Phi** Correlation coefficient (ϕ) for a 2×2 contingency table
- **Example**:

		Sickness	
		Yes	No
Smoking	Yes	**56** (A)	**27** (B)
	No	**18** (C)	**39** (D)

- Given the current level of association between the variables, ϕ coefficient yields: ϕ = 0.35 (considered moderate)

Odds ratio: (A/C) / (B/D)

- Odds ratio for this example turns out to be around 4.5. This means that those who smoke are 4.5 times more likely to develop the disease than those who don't smoke. Keep in mind that this is a very simplistic interpretation and has nothing to do with the statistical significance.

Correlation between categorical variables– cont'd

- **Cramer's V** Correlation Coefficient (ϕ_c):
- Used for categorical variables with more than 2 unique levels per category, including (2,3) and larger tables.
- ϕ_c is the strength of association ranging from 0 to 1, where
  - **0** indicates **no association**, **1** indicates **perfect association**

$$\Phi_c = \sqrt{\frac{\chi^2}{N\,k}}$$

χ^2 : Chi-square statistic
N : sample size involved in the test
k : min(# of rows-1 , # of columns-1)

- Interpretation:

ϕ_c	effect
> 0.5	High
0.3 to 0.5	Moderate
0.1 to 0.3	Low
0 to 0.1	None/Small

based on the degrees of freedom:

k	Small	Medium	Large
1	0.10	0.30	0.50
2	0.07	0.21	0.35
3	0.06	0.17	0.29
4	0.05	0.15	0.25
5	0.04	0.13	0.22

Correlation between categorical variables– cont'd

- **Cramer's V** Correlation Coefficient (ϕ_c):
- **Example**: We want to know the association between the preference of music and the study major. 200 students are questioned with the following contingency table:

	Pop	Rock	Jazz	Classical	Other
Psychology	12	12	8	8	0
Economics	4	4	20	6	6
Law	2	2	8	30	18
Other	2	2	4	16	36

With $\chi^2 \sim 113$, N = 200 and k = min(4-1 , 5-1) = 3, we find:

$$\Phi_c = \sqrt{\frac{113}{200 * 3}} = 0.43$$

$\Rightarrow$ This is a LARGE effect
(see table – previous slide)

Correlation between categorical variables– cont'd

- **Bias correction for Cramer's V** :
- Cramér's V can be a heavily biased estimator of its population counterpart and will tend to overestimate the strength of association. A bias correction, using the same notation, is given by Bergsma, Wicher (2013): "A bias correction for Cramér's V and Tschuprow's T", *Journal of the Korean Statistical Society*. **42** (3): 323–328.

$$\bar{\Phi}_c = \sqrt{\frac{\phi^2}{\min(\bar{c}-1, \bar{r}-1)}}$$

χ^2 : Chi-square statistic
N : sample size involved in the test
c : # of columns
r : # of rows

$$\phi^2 = \max(0\ ,\ \frac{\chi^2}{N} - \frac{(c-1)(r-1}{N-1})$$

$$\bar{c} = c - \frac{(c-1)^2}{N-1} \qquad \bar{r} = r - \frac{(r-1)^2}{N-1}$$

A/B Testing – revisited

- Let's go back to the A/B testing problem from Lecture-07. We found the p-value to be 0.042 (statistically significant).
- What is the effect size for this problem?
- Effect size is indeed the correlation between the message location (side-bar vs in-your-face) on the web page and the conversion (click vs non-click), which is the **Phi** (ϕ) coefficient:

$$\Phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

	Click	Non-click
Sidebar	**105 (A)**	**13895 (B)**
İn-your-face	**110 (C)**	**19390 (D)**

- **Effect size?** ϕ = 0.01148 is the correlation coefficient. We have a **very small effect** for the test statistic computed and the number of sample size given.
- **Odds ratio** = (105/110) / (13895/19390) = 1.33, i.e., sidebar visitors are 1.33 times more likely to click on the ad.

A/B Testing – revisited

- Calculation of **Phi coefficie**nt:

```
A=105 ; B=13895 ; C=110 ; D=19390
table_ = [[A,B],[C,D]]

num=A*D-B*C
denom=(A+B)*(C+D)*(A+C)*(B+D)
print('Effect size:', num/np.sqrt(denom))
>>> Effect size: 0.01148183264508
```

- Same result could've been obtained by using **Cramer's V** as well:

```
n = np.sum(table_)
r = len(table_[0]) ; c = len(table_[1])
k = min(r-1, c-1)

chi2s = stats.chi2_contingency(table_, correction=False)
print('Effect size:', np.sqrt(chi2s[0]/(n*k)))
>>> Effect size: 0.01148183264508
```

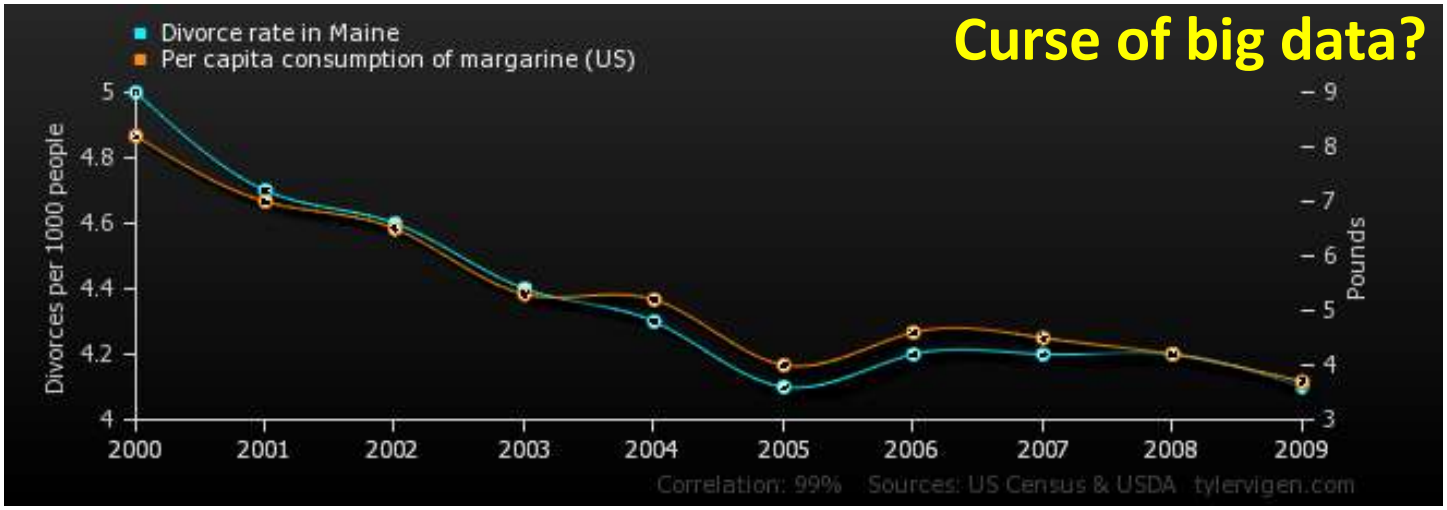
Summary: Correlation vs tool table

Variable 1	Variable 2	Method
Continuous	Continuous	• Pearson's corr (linear relation) • Spearman's corr (nonlinear relation) • Spearman's corr (at least one is ordinal
Continuous	Categorical (binary)	Point-Biserial
Categorical (binary)	Categorical (binary)	Phi coefficient
Categorical (multi)	Categorical (multi)	Cramer's V
Continuous	Categorical (multi)	• η^2 (eta-squared, effect size for ANOVA) • Logistic regression: measured by the accuracy and the degree of fit • Linear regression: square root of R^2, or the Person's correlation between the observed and fitted values in the regression analysis.

The advantage of Logistic regression as a measure of correlation between continuous and categorical variables: It doesn't make any assumptions like normality, linearity, homoscedasticity, etc. It doesn't suffer from multicollinearity as there is only 1 predictor. But it requires a balanced distribution among the levels in the categorical variable.

Correlation ≠ Causation

- Correlation doesn't imply causation… until it does...



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine (divorce per 1000 people)	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US)	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7
Correlation	**0.992558**									

Source: http://www.tylervigen.com/view_correlation?id=1703