Sobre la Correlacio I Associacio

Alex Sanchez-Pla

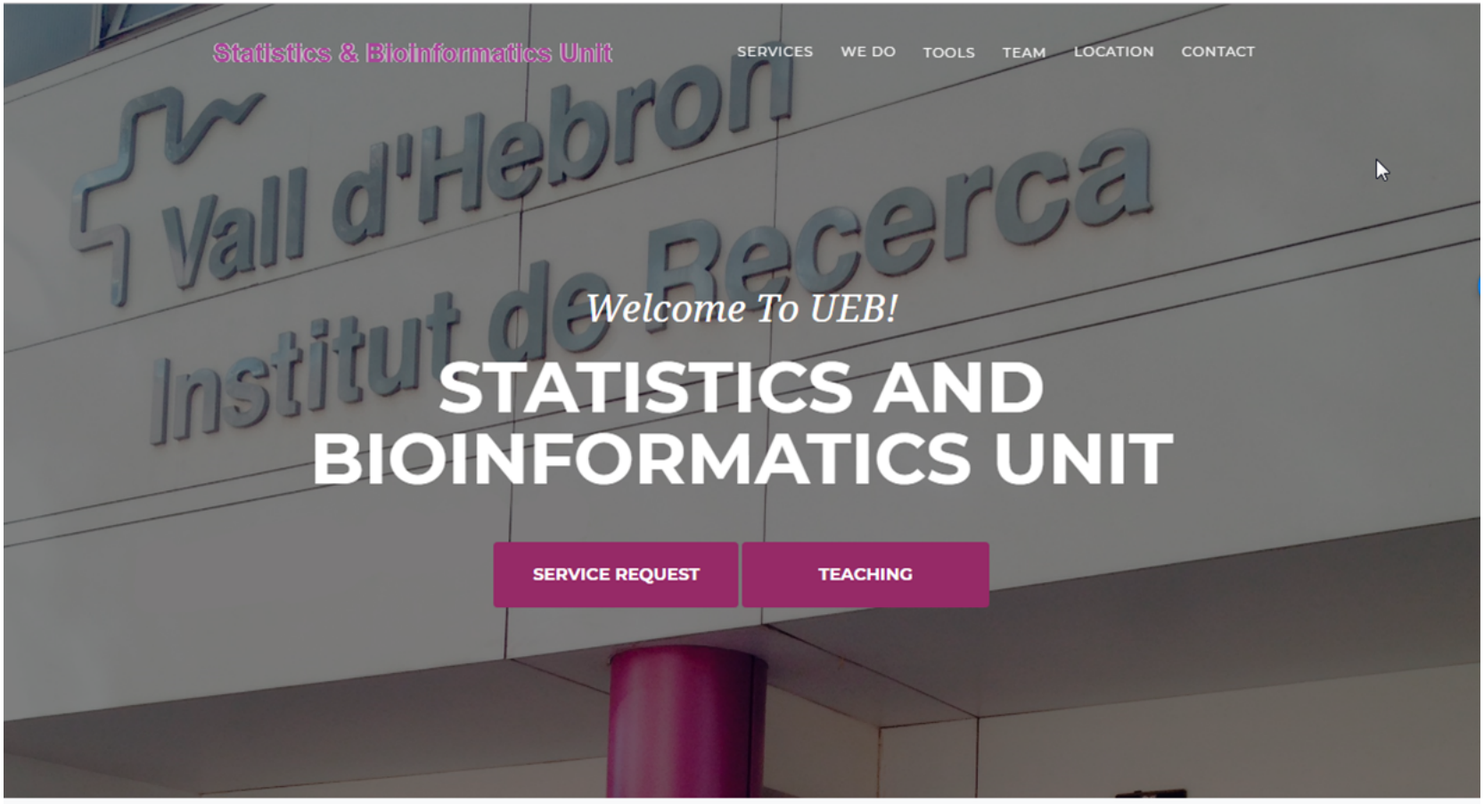
Unitat d'Estadística i Bioinformàtica (VHIR)

Dept. de Genetica, Micro I Estadistca (UB)

Dilluns 21 d'Octubre 2019 de 12:30 a 13:30
Sala d'Actes de Traumatologia i Rehabilitació

Les píndoles estadístiques son sessions divulgatives, organitzades per la Unitat d'Estadística i Bioinformàtica (UEB) del VHIR, on es presenten problemes i solucions estadístiques dirigides als professionals interessats del Campus Vall d'Hebron

Statistics and Bioinformatics Unit (UEB)



http://ueb.vhir.org

Outline of the talk

- Why this pill (some examples)
- Some basic ideas we have all heard about
- What about significance (p-values)
- Correlation false-friends
  - *Regression, Relation, Agreement, Causation* # Spurious, Ecological phallacy,
- Moving forward
  - More dimensions, More methods ...
- Wrap-up

Motivation (*Why this pill*)

- Everybody uses the term correlation, in science and normal life.

Gait Posture. 2017 Sep;57:241-245. doi: 10.1016/j.gaitpost.2017.06.014. Epub 2017 Jun 22.

Correlation of the torsion values measured by rotational profile, kinematics, and CT study in CP patients.

N Engl J Med. 2012 Oct 18;367(16):1562-4. doi: 10.1056/NEJMon1211064. Epub 2012 Oct 10.

Chocolate consumption, cognitive function, and Nobel laureates.

Messerli FH[1]. *Could there be a correlation between a country's level of chocolate consumption and its total number of Nobel laureates per capita?*

PLoS One. 2011; 6(10): e24828. PMCID: PMC3197194
Published online 2011 Oct 17. doi: 10.1371/journal.pone.0024828 PMID: 22043277

Importance of Correlation between Gene Expression Levels: Application to the Type I Interferon Signature in Rheumatoid Arthritis

Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine

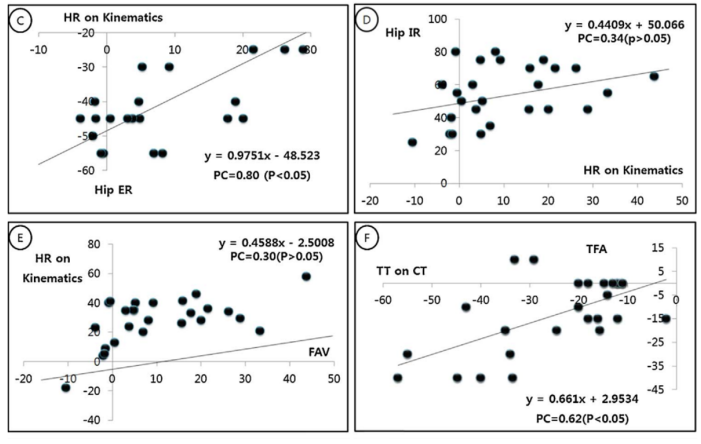
In addition, presence of miRNA molecules could be correlated to certain types of cancer,

Motivation (*Why this pill*)

- The concept, however is often misused or abused.
  - Correlation used instead of agreement
  - Regression used when there is no dependent/independent variables
  - Significance applied when assumptions do not hold
  - ...

Gait Posture. 2017 Sep;57:241-245. doi: 10.1016/j.gaitpost.2017.06.014. Epub 2017 Jun 22.

Correlation of the torsion values measured by rotational profile, kinematics, and CT study in CP patients.

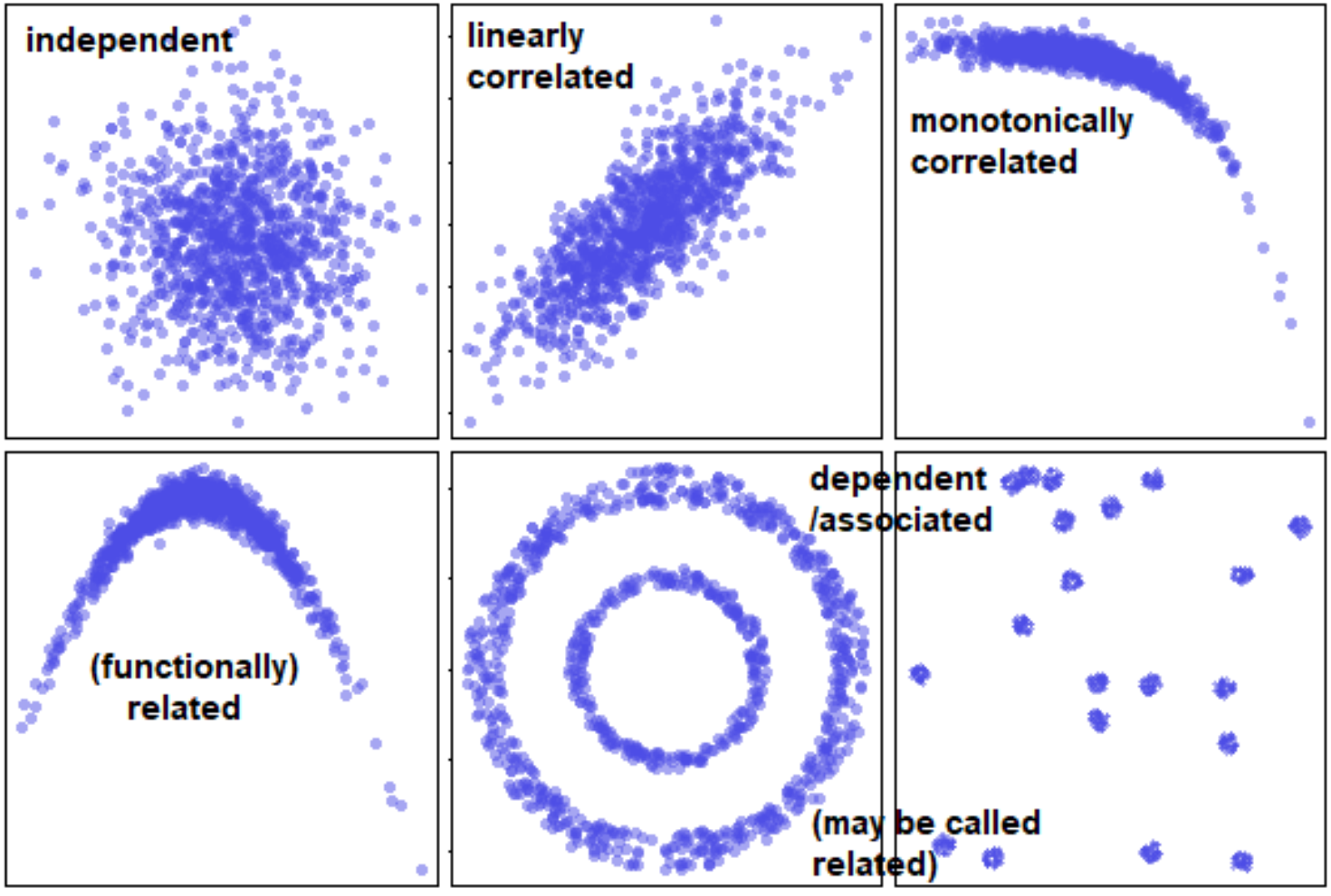


It may be worth to review a few concepts!

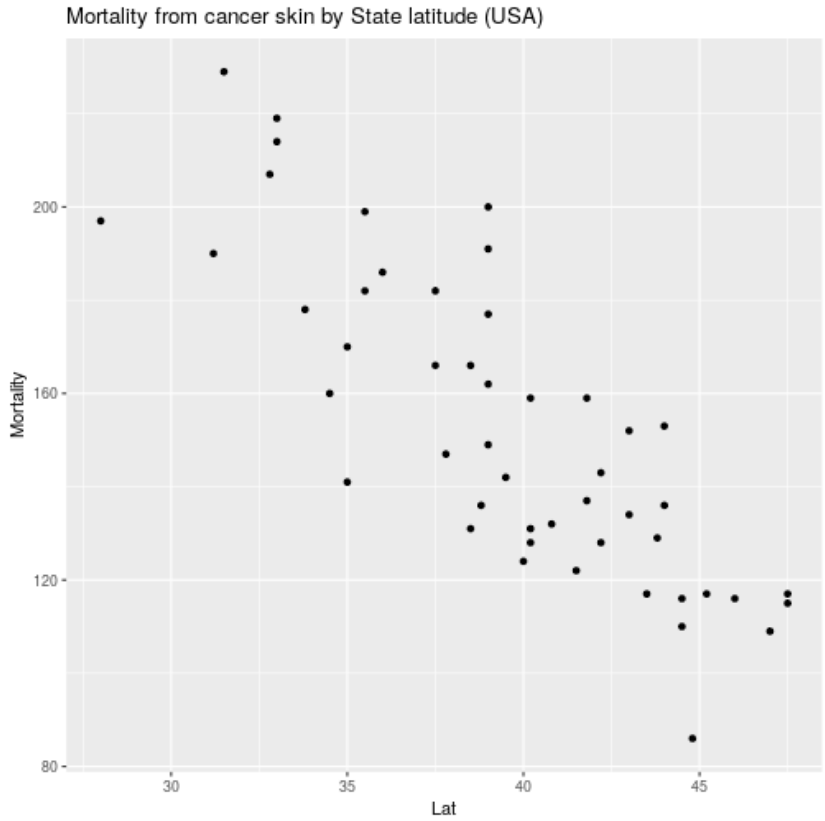
But what is correlation?

- In statistics, **dependence** or **association** is *any statistical relationship, whether causal or not, between two (random) variables*.
- In the broadest sense **correlation** is *any statistical association*, though
- it commonly refers to *the degree to which a pair of variables are linearly related* (Pearson correlation coefficient).

Association can take many forms

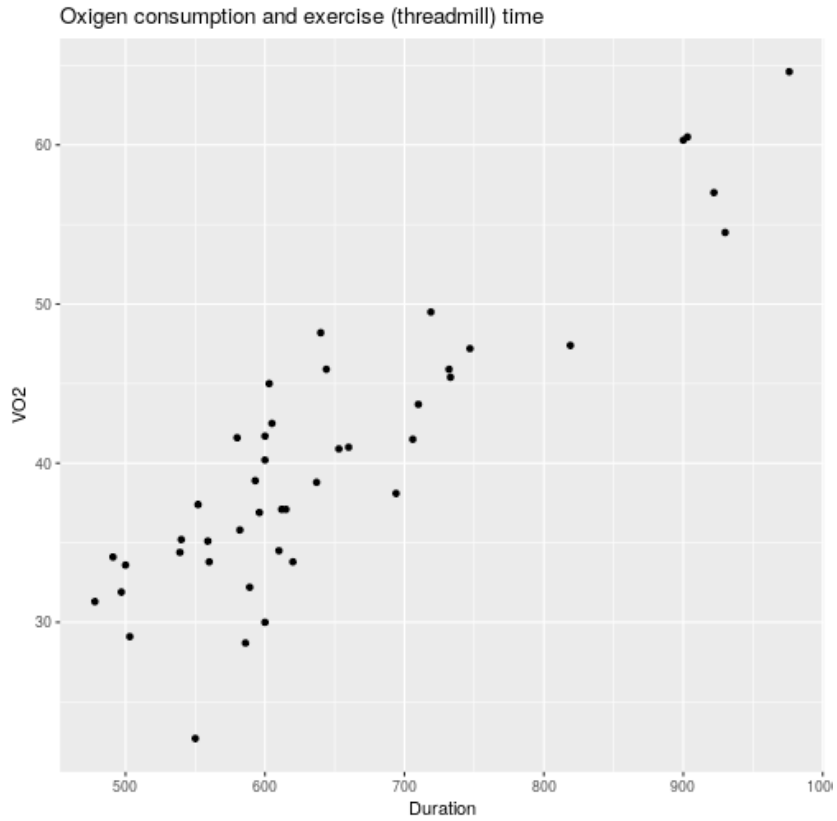


Example: Mortality & Latitude



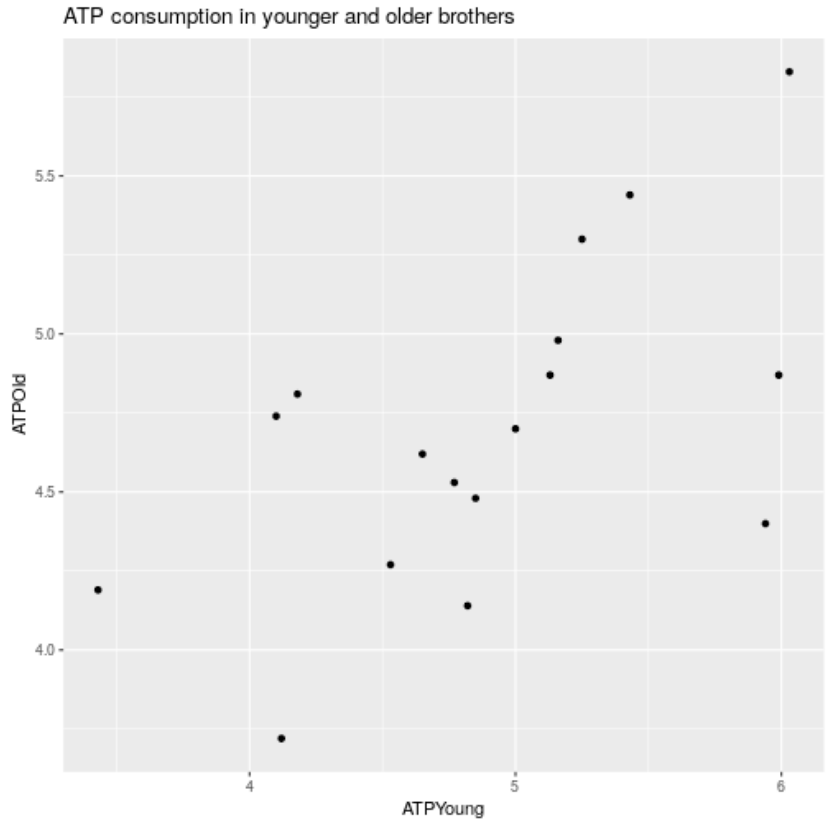
- Mortality Rate [per 10 Million (107)] of White Males Due to Malignant Melanoma of the Skin for the Period 1950–1959 by State and Some Related Variables

Example: Exercise & O_2



- Exercise Data for Healthy Active Males

Example: ATP consumption



- Erythrocyte Adenosine Triphosphate (ATP) Levelsa in Youngest and Oldest Sons in 17 Families Together with Age (Before Storage)

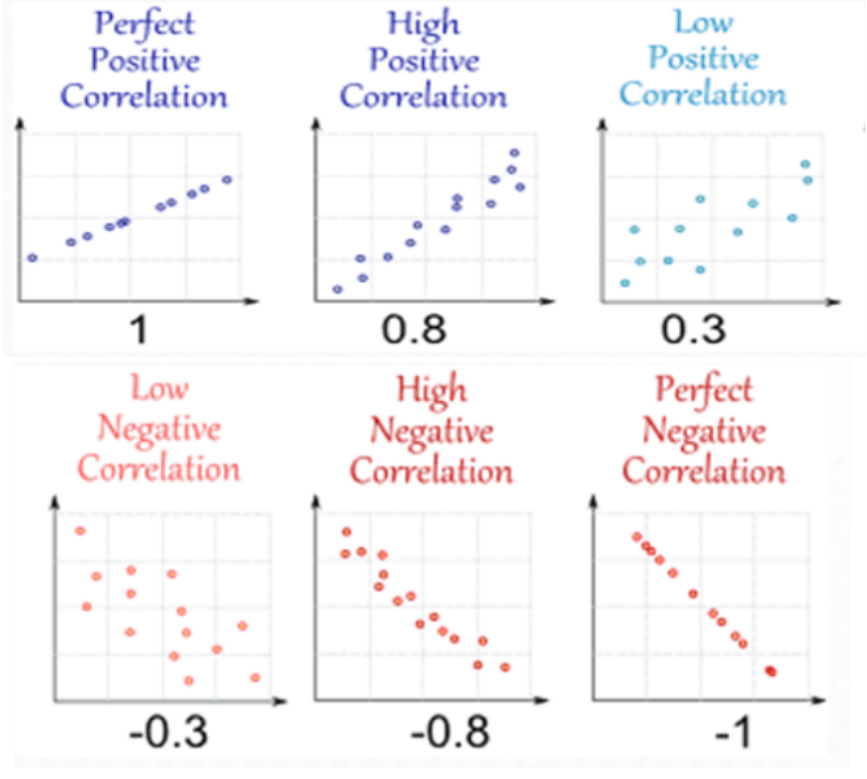
Reviewing correlation coefficients

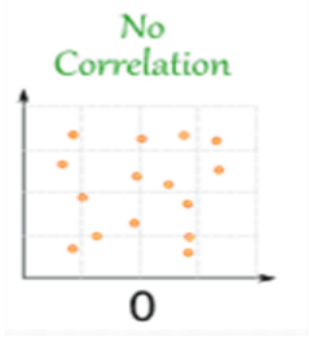
What's in my toolbox?

- Pearson (linear) correlation coefficient
- Sperman (ranks) correlation coefficient
- Kendall's tau
- Intraclass correlation coefficients
- Many other: Distance correlation, Mutual Information, etc.

Pearson correlation coefficient

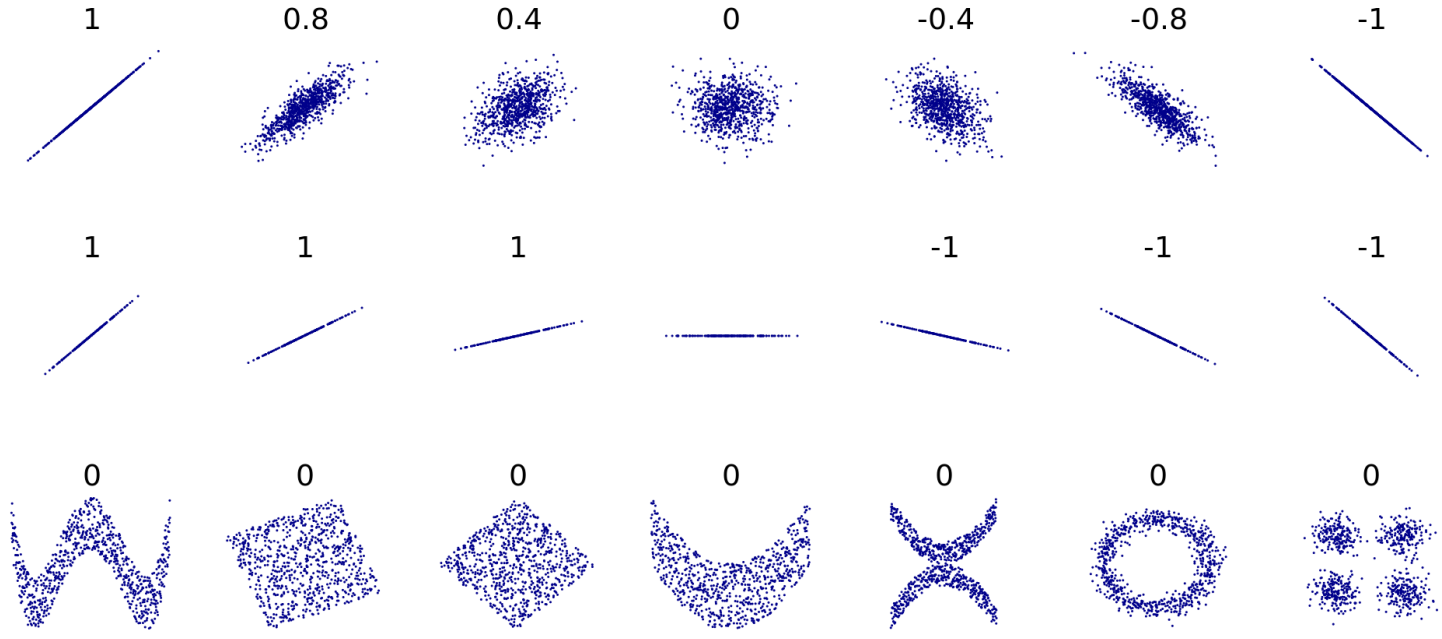
- Most commonly used correlation coefficient
- Measures the degree of **linear** relation between a pair of *quantitative* variables.
- It takes values between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.



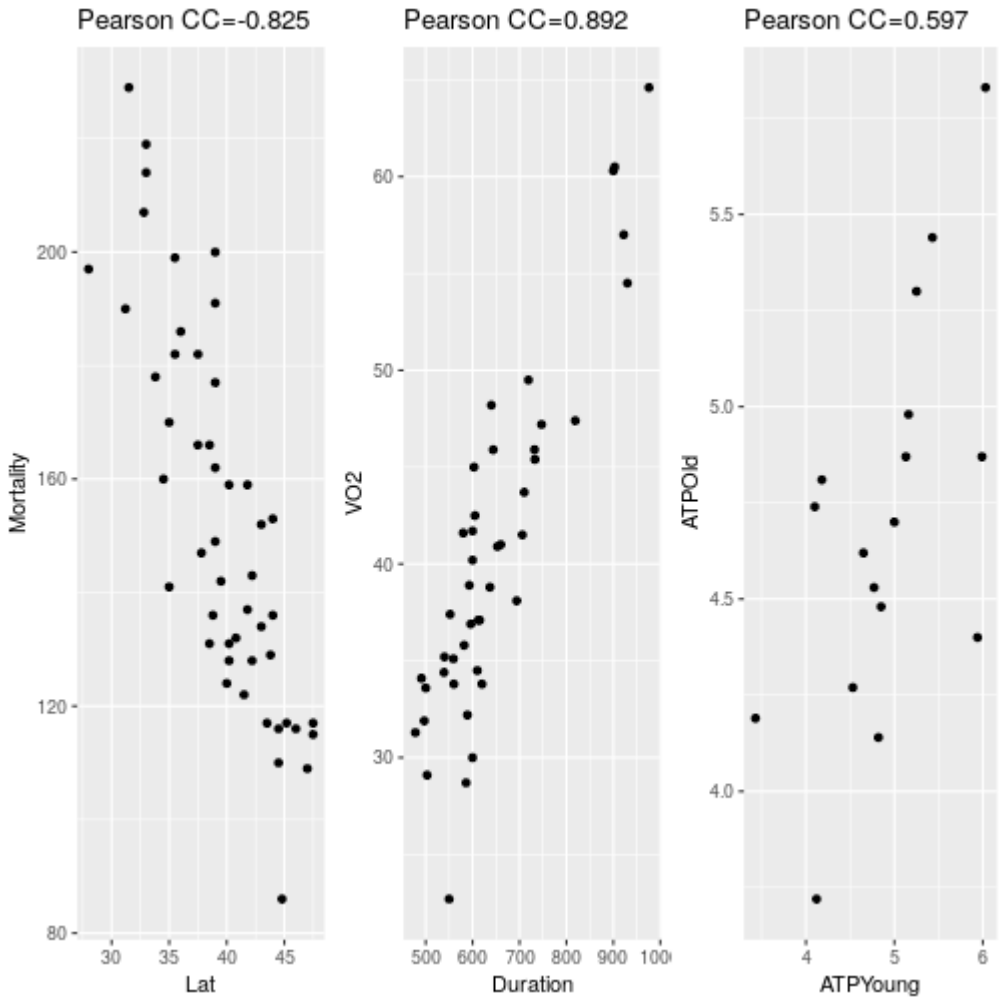


Pearson CC measures linear association

- Pearson CC changes with noise
- It is not affected by changes in slope
- If relation is not linear it becomes useless



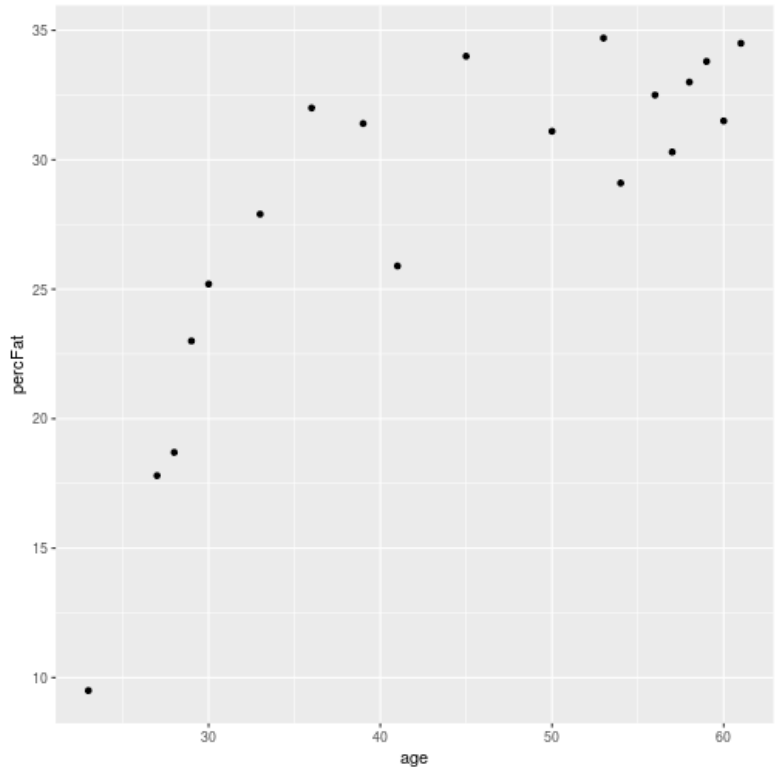
Pearson CC examples



Spearman correlation

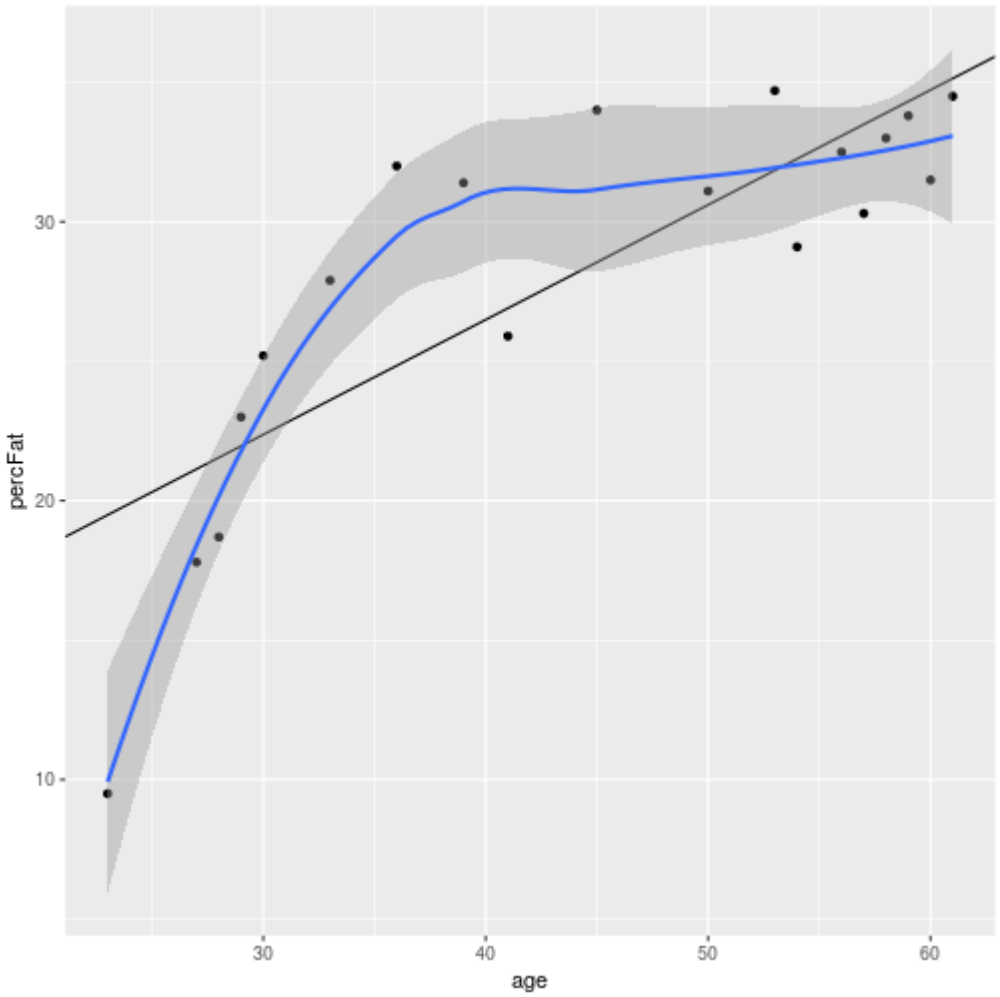
- If the relation is *monotone* or the variables are in an *ordinal scale* use *Spearman Ranks Correlation Coefficient* instead of Pearson's.

	age	percFat
1	23	9.5
3	27	17.8
4	28	18.7
5	29	23.0
9	30	25.2
2	33	27.9
12	36	32.0
6	39	31.4
7	41	25.9
8	45	34.0
10	50	31.1
11	53	34.7
13	54	29.1
14	56	32.5
15	57	30.3
16	58	33.0



Based on - Maze et al (1984).

Relation is monotonic, but not linear



Spearman CC is based on **ranks**

	age	rankAge	percFat	rankFat
1	23	1	9.5	1
3	27	2	17.8	2
4	28	3	18.7	3
5	29	4	23.0	4
9	30	5	25.2	5
2	33	6	27.9	7
12	36	7	32.0	13
6	39	8	31.4	11
7	41	9	25.9	6
8	45	10	34.0	17
10	50	11	31.1	10
11	53	12	34.7	19
13	54	13	29.1	8
14	56	14	32.5	14
15	57	15	30.3	9
16	58	16	33.0	15
17	59	17	33.8	16
18	60	18	31.5	12
19	61	19	34.5	18

Spearman CC = Pearson CC computed on ranks

```
cor(age,percFat,
    method = "spearman")
```

```
## [1] 0.777193
```

```
cor(rankAge, rankFat,
    method = "pearson")
```

```
## [1] 0.777193
```

What about Kendall's tau

- Kendall's tau is a correlation coefficient based on ranks
- That is, it can be seen as *an alternative to Spearman CC*
- Indeed the reason why/when we should choose Spearman or Kendall is not clearly documented

https://stats.stackexchange.com/questions/3943/kendall-tau-or-spearmans-rho

So what? Pearson or Spearman?

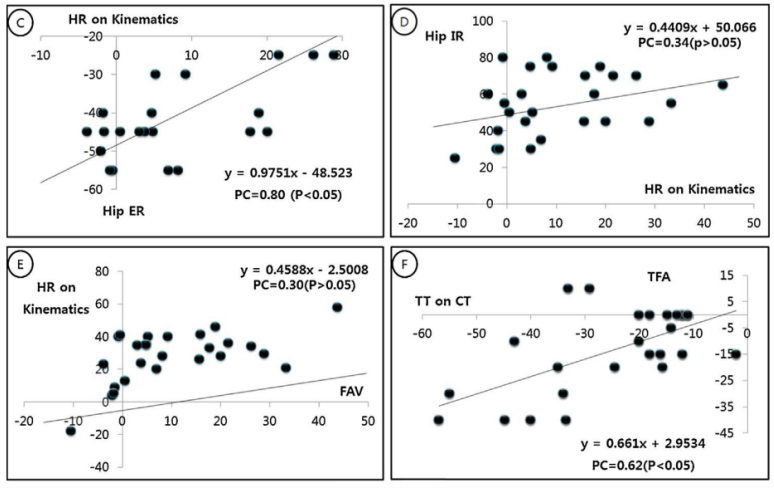
- Start with a plot
- If the relation ship is *approximately* linear you can use Pearson CC.
- If the relation is not linear but it is *monotonically increasing* or *monotonically decreasing* you better use Spearman CC.
- The *measurement scale* is also a criteria
    - Use Pearson if the data is in an "interval" or "quotient" scale
    - Use Spearman if the data are on an "ordinal" scale

Review : Measurement scales

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

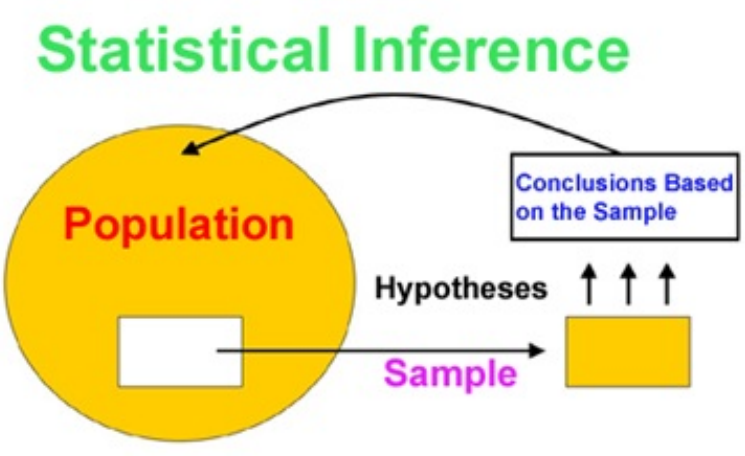
Significance of correlation

- Common correlation analysis report
  - An estimate of correlation
  - A least squares straight line
  - A significance p-value
- Is everything appropriate?



Can we make inferences on correlation?

- Statistical inference deals with inferring properties in population characteristics from representative samples.



When dealing with correlation it means answering questions such as:

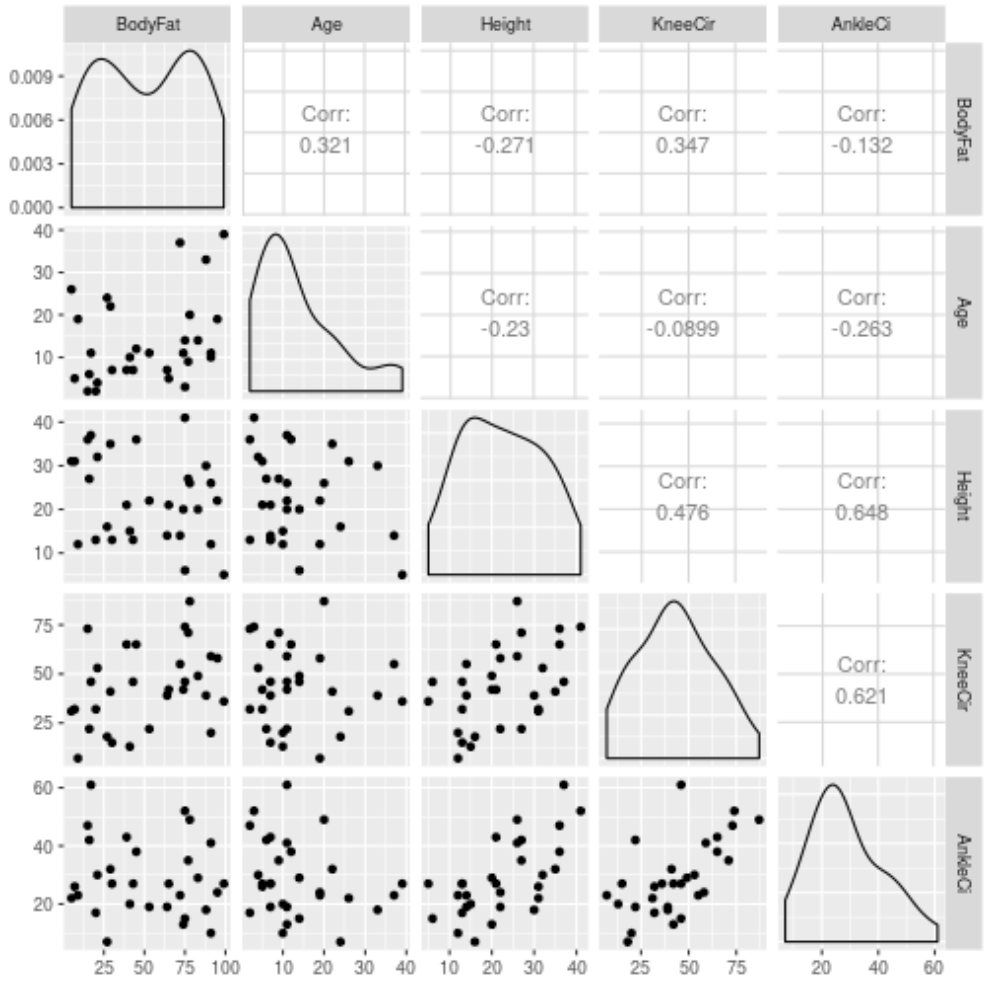
- What is the estimated correlation coefficient of the population?
  - How precise is this estimate?
- Is there any correlation in the population, or is the sample correlation just the luck of the draw?

Can we put a p-value on *r*?

- It is possible to make some inferences on the population correlation coefficient, ρ.
- It is not so straightforward as computing the sample correlation coefficient r.
- For the p-value to be "valid" some assumptions must hold:
  - We have continuous variables in an interval or quotient scale, with no outliers.
  - Data comes from a simple random sample
  - From a bivariate normal distribution
- If these assumptions hold it is possible to tests the hypotheses:

$$H_0 : \ \rho = 0, \ \text{vs} \ \rho \neq 0.$$

Example: Body fat and measures

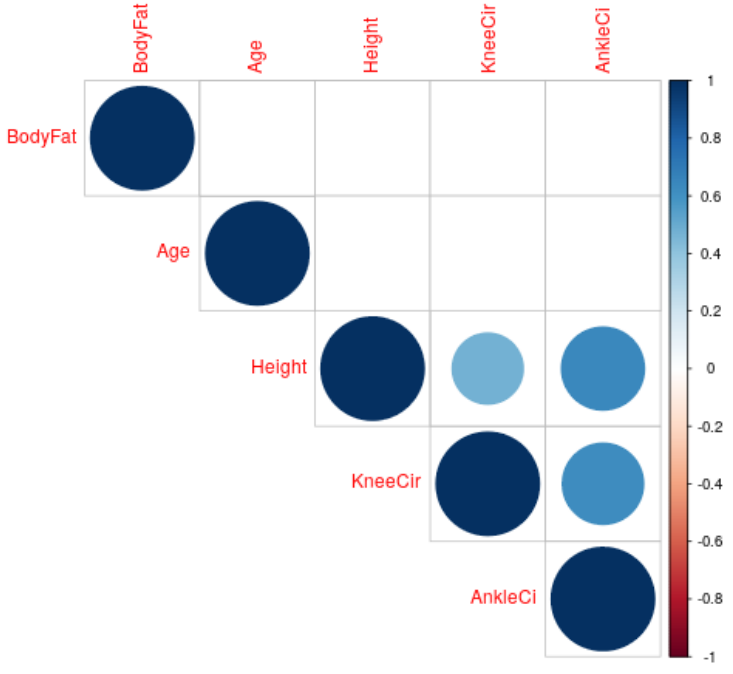


Source: "http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_BMI_Regression"

P-values for correlations

	BodyFat	Age	Height	KneeCir	AnkleCi
BodyFat	1.000	0.084	0.148	0.060	0.488
Age	0.321	1.000	0.222	0.636	0.160
Height	-0.271	-0.230	1.000	0.008	0.000
KneeCir	0.347	-0.090	0.476	1.000	0.000
AnkleCi	-0.132	-0.263	0.648	0.621	1.000

- Correlation matrix (below) with
- significance values (above)



- Correlation depicted by circle size
- Significance: Values above threshold left blank

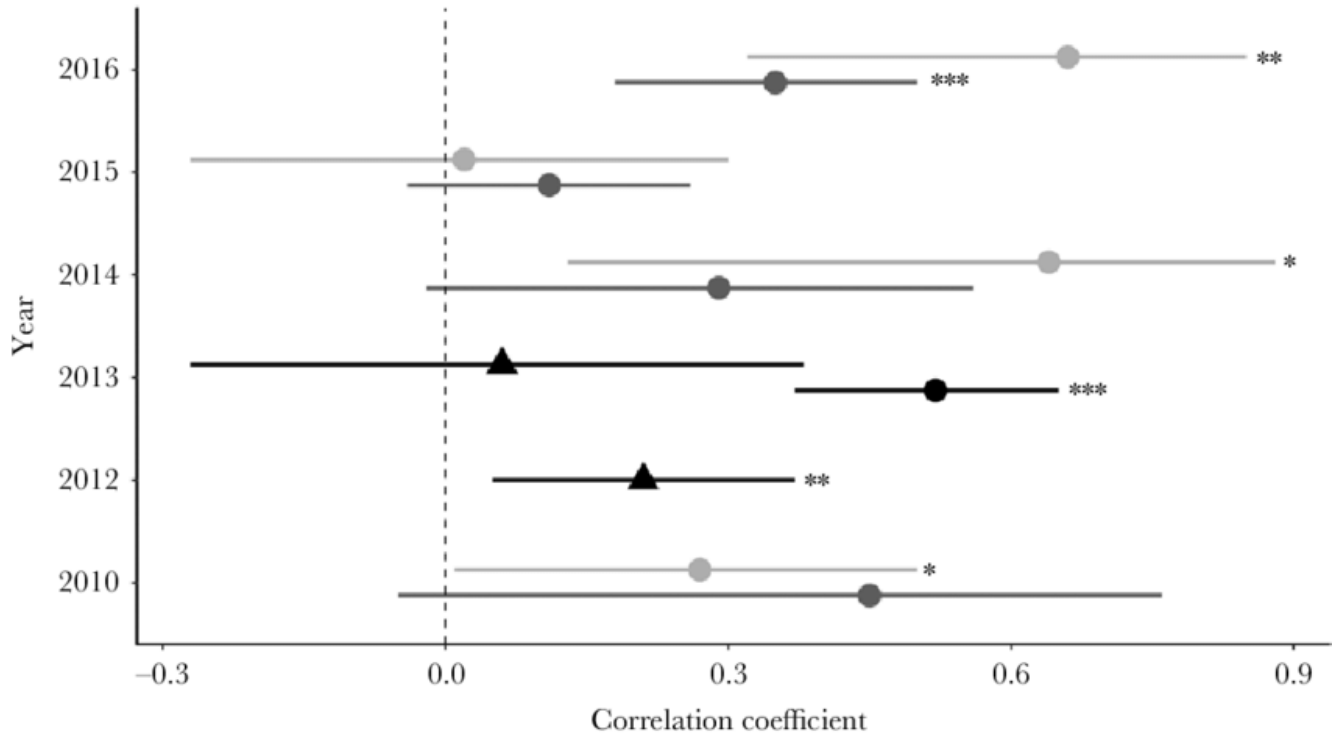
What does this test tell?

- Correlation significance test **tells nothing about how big the population correlation is**.
- It only informs about if **it can be considered to be distinct than zero**.
- Besides this it can only be considered valid if the assumptions hold.
- And, if it were not enough, results are **very sensitive to sample size**
  - Something not significant for small sample size.
  - Becomes significant as sample size increases.
- In summary: *Too much effort for a very small prize?*

Confidence intervals for ρ

- A better alternative: **Compute confidence intervals for the correlation coefficient**.
  - Again validity of assumptions is an additional difficulty
  - However it is possible to obtain good approximations based on computational intensive approximations such as the **bootstrap**.

Correlation coefficient CI example



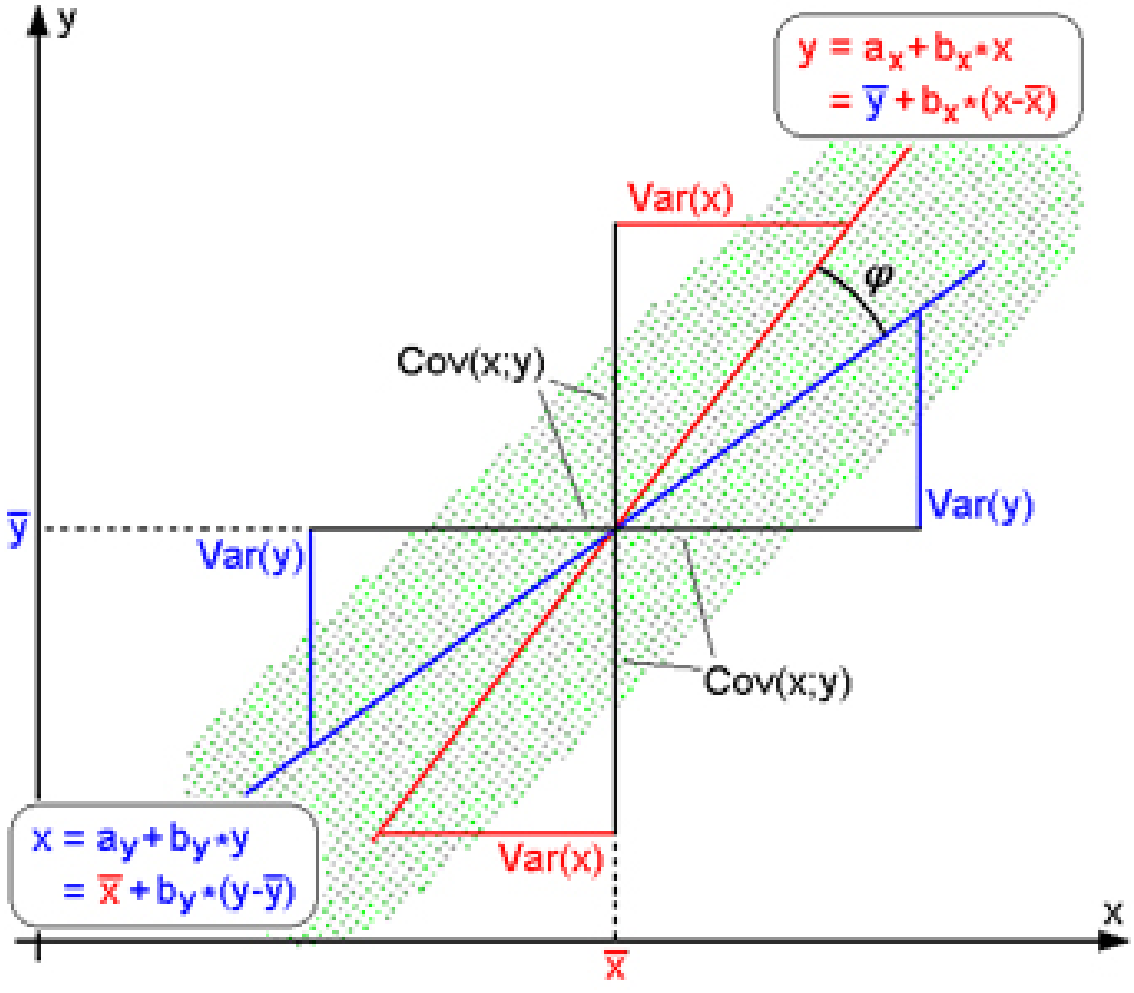
- Correlation coefficient with respective 95% confidence interval of mean body surface affected by a chronic dermathological disease.
- Correlation coefficients are given for each year and are separated by work type.
- The black dotted line represents zero correlation. Significant correlations are marked with $*$ P < 0.05, $**$ P < 0.01 and $***$ P < 0.001.

False friends (1): Regression

- Correlation and regression are often presented together.
- They are related, but they are not the same:

Basis for Comparison	Correlation	Regression
Meaning	Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describes how an independent variable is numerically related to the dependent variable.
Usage	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
Dependent and Independent variables	No difference	Both variables are different.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

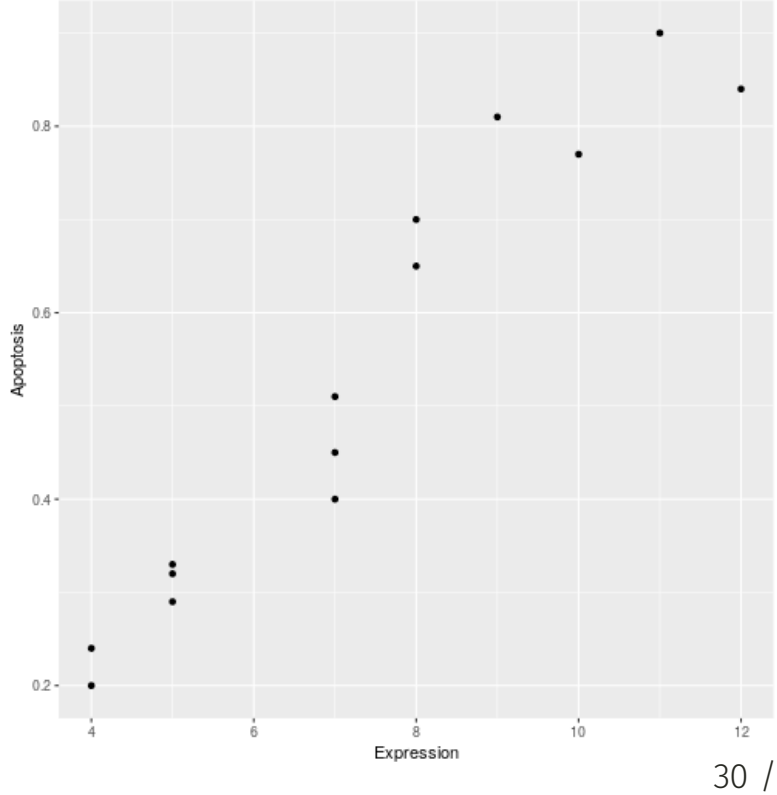
One correlation vs two regressions



False friends (2): Class comparison

- Is there correlation between disease and expression of a certain biomarker gene?
  - OK to compute Pearson CC between `Expression` and `Apoptosis` and say "There is"/"There isn't"
  - Not OK to test differences between `HIGH` and `LOW` and say "There is"/"There isn't"

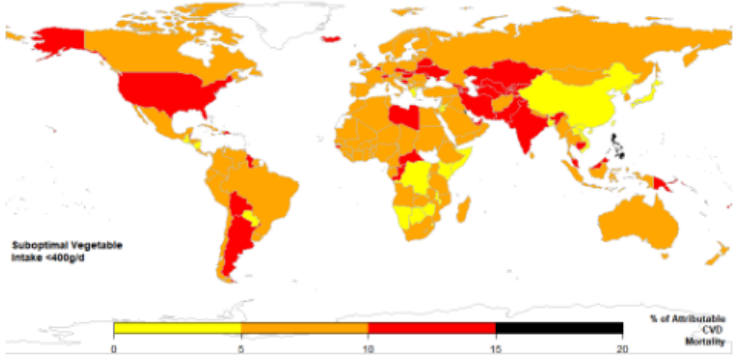
Expression	Apoptosis	Disease_cat
4	0.20	LOW
7	0.40	NA
5	0.33	LOW
8	0.70	HIGH
8	0.65	NA
9	0.81	HIGH
11	0.90	HIGH
5	0.32	LOW
7	0.45	NA
12	0.84	HIGH
10	0.77	HIGH
4	0.24	LOW
5	0.29	LOW
7	0.51	NA



FF (3): Correlation is not causation

Millones de muertes cardiovasculares por no comer suficiente fruta y verdura

Se estima que aproximadamente 1 de cada 7 muertes cardiovasculares podrían atribuirse a no comer la cantidad de fruta recomendada y 1 de cada 12 se adjudicarían al inadecuado consumo de verduras



Este es el porcentaje de muertes cardiovasculares atribuible al bajo consumo de verdura (menos de 400 gramos al día) en todo el mundo.
Escuela Friedman de Ciencias y Políticas de Nutrición en la Universidad de Tufts.

- Correlation informs about the association between two variables
- **It cannot tell anything** about if changes in the values of one variable are due to changes in the values of the other variable.

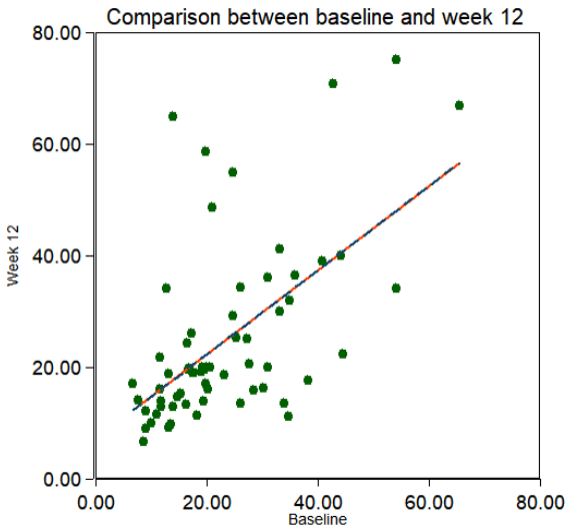
FF (4): Correlation is not agreement

- Correlation and agreement are similar and related concepts but, *although similar and related, _they represent completely different notions of association*.
  - Agreement appears when we wish to compare two methods of measuring *the same* thing.
  - Correlation assumes that we are measuring variables that may be related, but represent *different* constructs.

- Examples:
  - Lung function measured using a spirometer (expensive, accurate) or peak flowmeter (cheap, less accurate).
  - Two devices (oropharyngeal and conventional) used to mesured acidity (pH) in the esophagus as a marker of reflux.

FF (4): Correlation is not agreement

- *Aorta pulsatility* is measured in week 0 and week 12 in the same body location in stable patients.
- Both variables are measuring the same (AP) therefore the correlation does not make sense.
- Instead we compute a *concordance coefficient* (intraclass correlation) which accounts for the repeted measurement effect.



FF (4): Correlation is not agreement

Concordance correlation coefficient (Lin, 1989)					
rho_c	SE(rho_c)	Obs	[95% CI]	P value	CI type
0.563	0.083	64	[0.401; 0.725]	0.000	asymptotic
			[0.380; 0.703]	0.000	z-transform

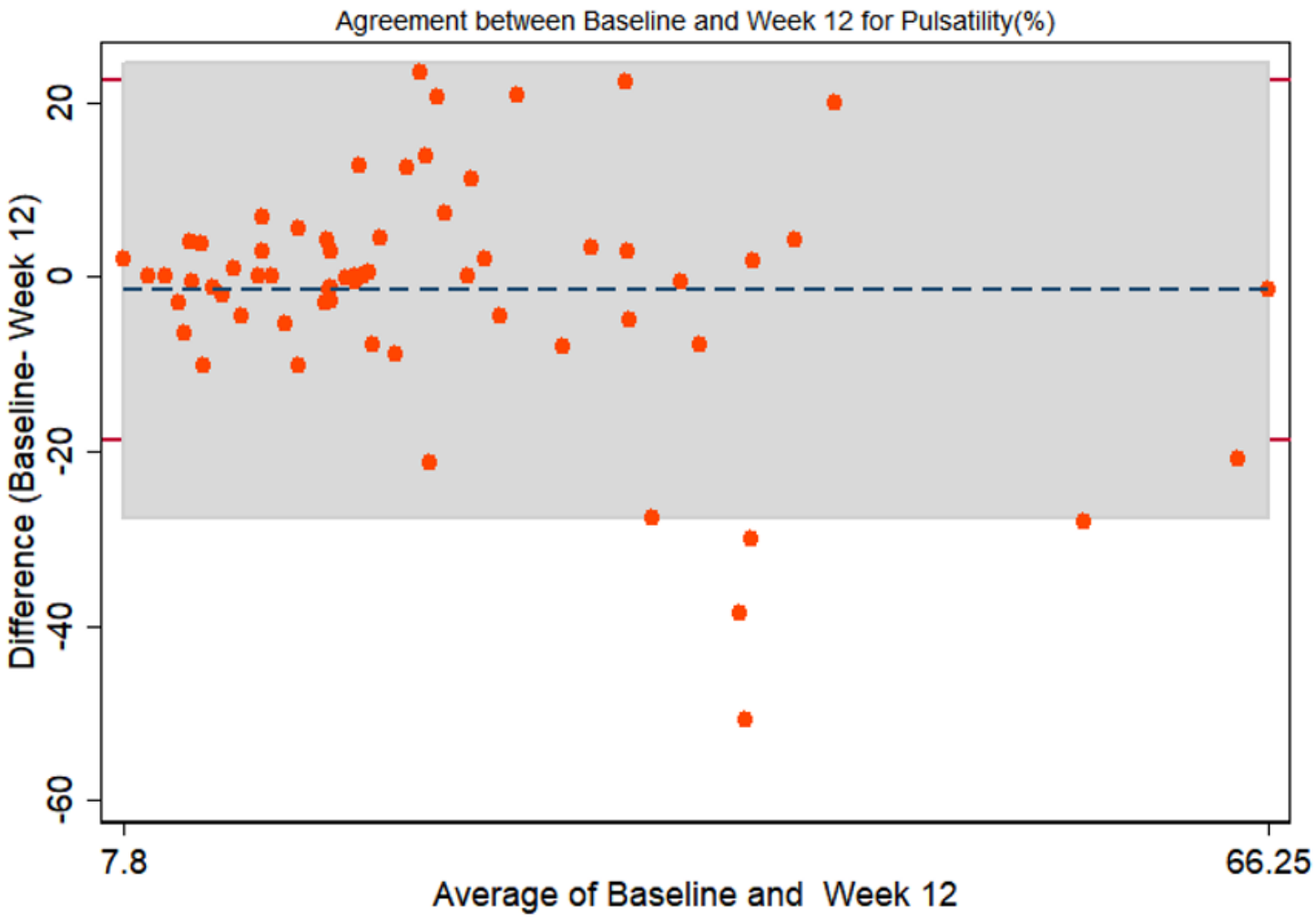
Pearson's r = 0.584 Pr(r = 0) = 0.000 C_b = rho_c/r = 0.963

Reduced major axis: Slope = 0.775 Intercept = 4.073

Difference (ivuspulsatilitat3 - ivuspulsatilitat2)		95% Limits Of Agreement
Average	Std. Dev.	(Bland & Altman, 1986)
1.535	13.291	-24.514 27.583

Bland-Altman comparison of ivuspulsatilitat2 and ivuspulsatilitat3	
Limits of agreement (Reference Range for difference)	-28.116 to 25.047
Mean difference	-1.535 (CI -4.854 to 1.785)
Range	7.800 to 66.250
Pitman's Test of difference in variance	r = -0.303, n = 64, p = 0.026

FF (4): Correlation is not agreement



Going multivariate

- What about correlation when we deal with several variables simultaneously?
- Obvious step 1: Multiple correlations: every variable vs every other variable.

```
##          BdyFt Age   Heght KneCr AnklC
## BodyFat   1.00
## Age       0.32  1.00
## Height   -0.27 -0.23  1.00
## KneeCir   0.35 -0.09  0.48  1.00
## AnkleCi -0.13 -0.26  0.65  0.62  1.00
```

- A good idea: adjust correlation between any 2 variables removing what is explained by their correlation with other variables.>

```
##          BdyFt Age   Heght KneCr AnklC
## BodyFat   1.00
## Age       0.28  1.00
## Height   -0.37  0.02  1.00
## KneeCir   0.60 -0.09  0.32  1.00
## AnkleCi -0.23 -0.11  0.35  0.51  1.00
```

Conclusions and recap

- Correlation is a useful and informative tool to quantify the relation between pairs of variables.
- Distinct situations may require distinct approaches
- If certain assumptions hold, it is possible to compute a p-value for the CC
  - Although it only informs if there is *any* correlation
  - To estimate *how big* the correlation is, better use confidence intervals.
- Correlation has many *false friends*
  - Regression, Agreement, Causation, Spureous, ...
  - Learn to know them and to know when to use them.
- Correlation can be extended in many directions
  - More coefficients
  - More dimensions

References

- Aggarwal, R., & Ranganathan, P. (2016). Common pitfalls in statistical analysis: The use of correlation techniques. Perspectives in Clinical Research, 7(4), 187. https://doi.org/10.4103/2229-3485.192046
- Asuero, A. G., Sayago, A., & González, A. G. (2006). The Correlation Coefficient: An Overview. Critical Reviews in Analytical Chemistry, 36(1), 41–59. https://doi.org/10.1080/10408340500526766
- Bland, J. M., & Altman, D. G. (n.d.). STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. https://doi.org/10.1016/j.gaitpost.2017.06.014
- Kamel, H. F. M., & Al-Amodi, H. S. A. B. (2017, August 1). Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine. Genomics, Proteomics and Bioinformatics. Beijing Genomics Institute. https://doi.org/10.1016/j.gpb.2016.11.005
- Kim, H. Y., Cha, Y. H., Chun, Y. S., & Shin, H. S. (2017). Correlation of the torsion values measured by rotational profile, kinematics, and CT study in CP patients. Gait & Posture, 57, 241–245.
- Xu, W., Hou, Y., Hung, Y. S., & Zou, Y. (2012). A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. Signal Processing, 93, 261–276. https://doi.org/10.1016/j.sigpro.2012.08.005

QUESTIONS?