USTH MASTER SPACE

Earth Observation - Astrophysics - Satellite Technologies

USTH

VIETNAM FRANCE UNIVERSITY

Lecture 1: Fundamental statistics quantities

Hoàng Đức Thường

Department of Space and Applications (DSA), USTH

Hanoi, 2021

Statistics

In this class I will review statistic quantities as mean, median, variance, and associated errors. The correlation between two random variables.

Statistics

At the end of this class, you should be able to do:

- Determine statistic quantities for a data set and their errors
- Optimally combine data.
- Correlation coefficient.

Statistics

- Statistics are based on data only: The collection, organization, analysis, interpretation, and presentation data.

- Statistics: A number or set of numbers that describe the data. (mean, median, mode, variance, ...)

The process of science

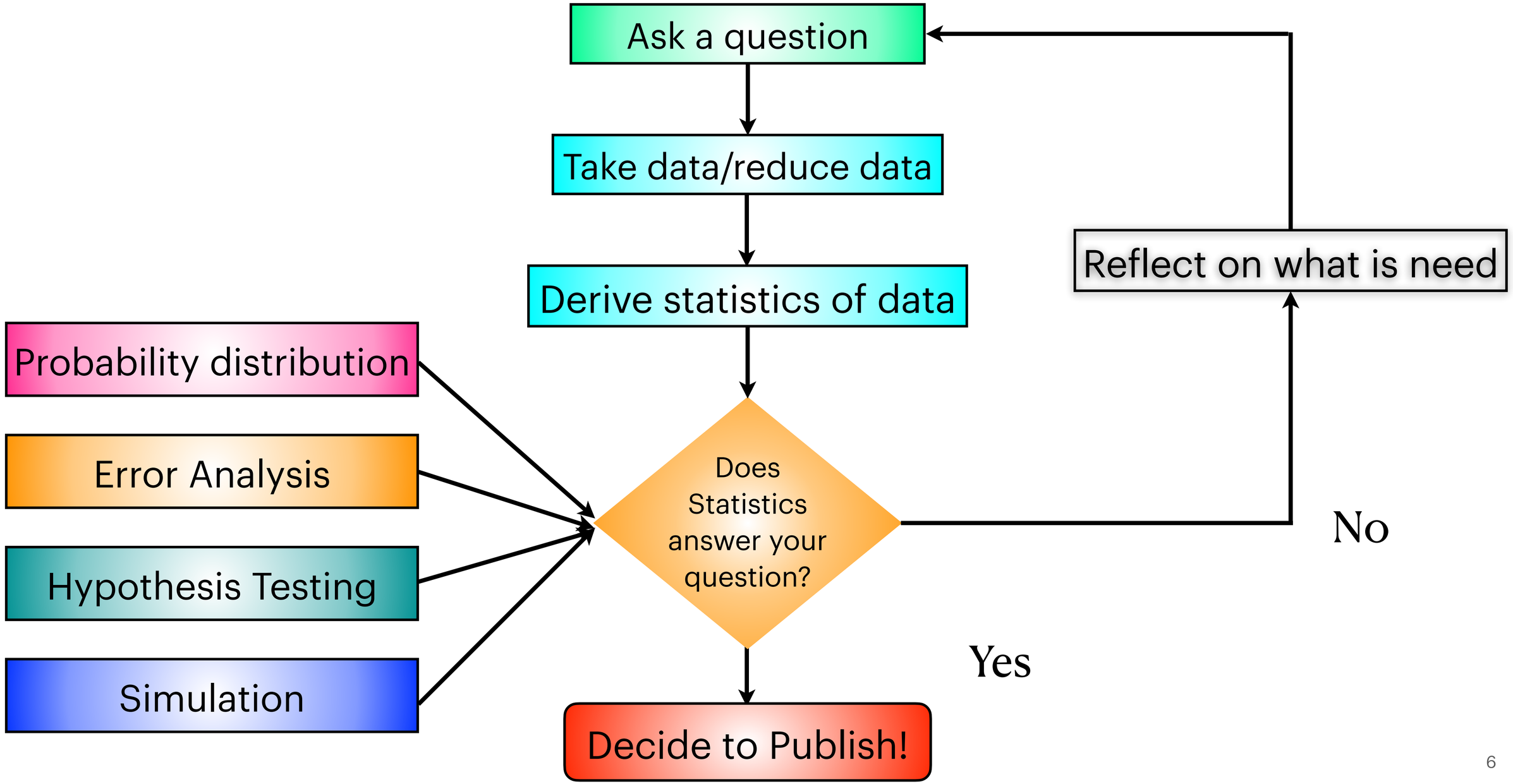
Observation/
Measurements

Analyze data

Conclusion
(Test hypothesis,
probabilities)

Scientific questions

The process of decision making



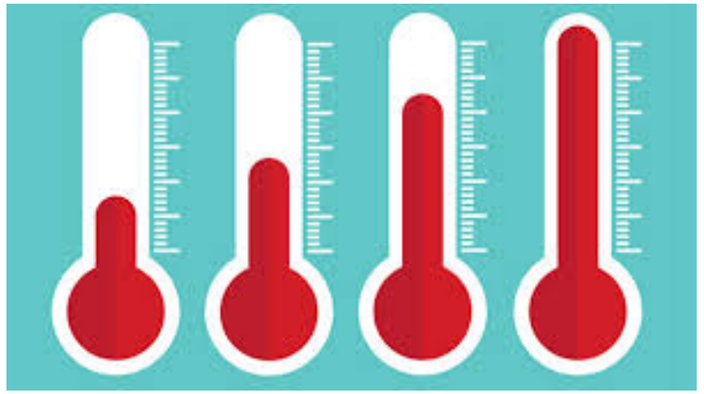
Why do we need statistics?

"If your experiment needs statistics, you ought to have done a better experiment" (E.Rutherford)

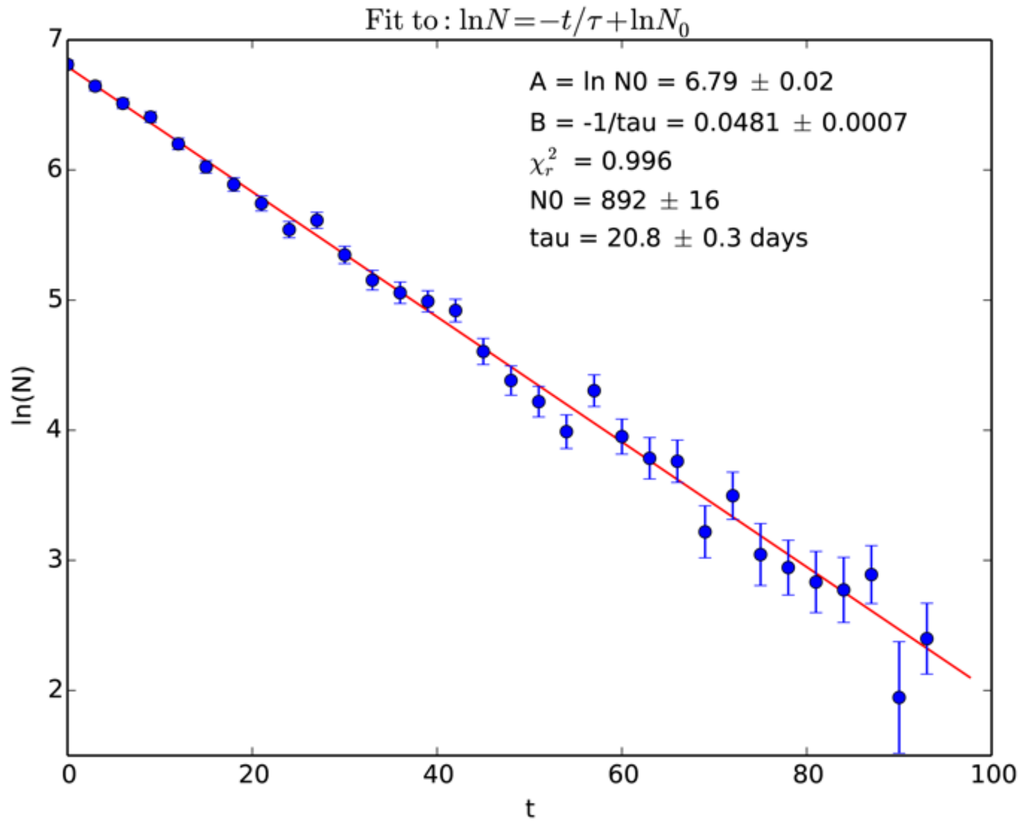
- Statistics are about making **precise decisions.**
- Statistics allow us to examine **relationships,** explore issues or **uncertainties.**
- Statistics allow us to construct concepts and **develop theories**.

When we use statistics?

- **Measuring a quantity (parameter estimation):** We can describe a data set using mean, mode, median, variance parameter, what is the uncertainty of the estimation?



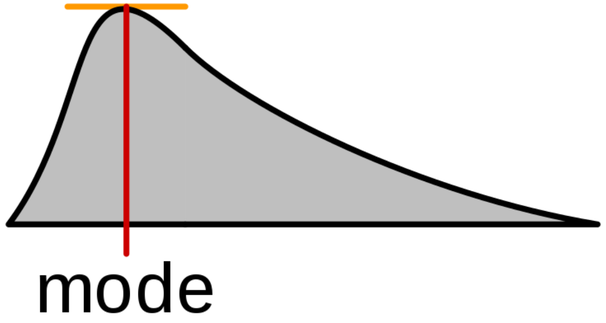
- **Correlations** between variables: There are two variables have correlated with each other, are they implying a probably physical connection?



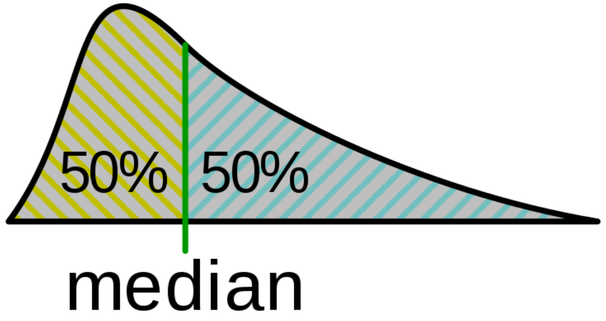
- **Testing a model**: We can fit a set of data with a theoretical model for hypothesis testing. Which model is the best?

1. Statistics quantities: Mode, Median, Mean

- A **statistic** is a quantity that summarizes/describes the data.
- Let's consider a set of data of N independent sample x_i, how do we summary this data?

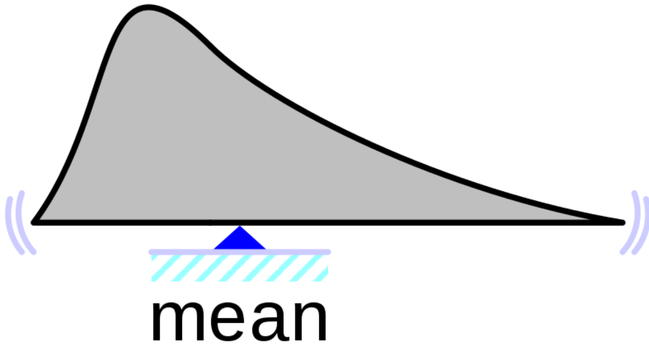


- The **mode** (The most often appear value/the peak).
- The **median** (middle of the range of values / middle value when ranked).



- The **mean** (average value):

$$E(x) = \mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N}$$



1. Statistics quantities: Mode, Median, Mean

- A **statistic** is a quantity that summarizes/describes the data.
- Let's consider a set of data of N independent sample x_i, how do we summary this data?
- The **standard deviation/Mean square deviation** σ (spread) or **variance**:

$$E(x_i - \bar{x}) = Var(x) = \sigma^2(x) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

- The **Root mean square deviation** (rms) = σ
- The **mean deviation**

$$\overline{\Delta x} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$$

1. Statistics quantities: Mode, Median, Mean

- We can describe **error** in each statistics quantity:

- The error in mean is **standard deviation divided by** $\sqrt{N}$ (It means that if we increase the number sample/measurement, the error in the mean will improve.)

- The **error in mean** $= \frac{\sigma}{\sqrt{N}}$

- The **error in median** $= 1.25\frac{\sigma}{\sqrt{N}}$

- The **error in variance** $= \sigma^2\sqrt{\frac{2}{N-1}}$

1. Statistics quantities: Mode, Median, Mean

Exercise 1:

We have N=12 measurements of a variable x_i=(9.6, 6.2, 8.3, 6.1, 7.0, 7.9, 7.3, 6.4, 7.1, 7.3, 6.8, 10.1). Estimate the mean, variance and median of this dataset. What are the errors in your estimates?

The meaning of an error bar

Example: Hubble constant measurement the expansion rate of the Universe:

$$H_0 = 70 \pm 5 \ kms^{-1}Mpc^{-1} .$$

- This almost never means H_0 is between 65 and 75
- It almost always means there is 68% (one sigma - 1σ) confidential probability that H_0 is in the region $65 < H_0 < 75$.
- It often means the distribution for H_0 is a Normal/Gaussian distribution with a mean value 70 and a standard deviation 5.
- We can describe H_0 using a probability distribution.

Estimators and bias

- Statistics are based on data, however they often use as **estimators** of probability distribution. This is a frequentist approach that helps to decide a model is good or bad.

- Example: The variance estimator $\hat{V} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$ measure the variance V.

- If an estimator is **unbiased (no error),** it means that the true value can be recovered on **average** over many realizations, $\left\langle \hat{V} \right\rangle = V$.

Notice: "^" notation means "estimator of"

$\langle \ldots \rangle$ notation means average over many experiments/measurements

Optimal combination of data

- If we have N independent estimates x_i of quantity y. Each has varying errors σ_i. What is our best combined estimate of y?

- A simple average $\hat{y} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is not the optimal combination, because we want to give **weight to the more precise estimates**. Let's weight each estimate by w_i, the best combined estimate is the weight mean:

$$\hat{y} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}.$$

Optimal combination of data

- The weights which minimize the combined error are **inverse-variance weights**

$$w_i = \frac{1}{\sigma_i^2}$$

$$\hat{y} = \frac{\sum_{i=1}^{N} x_i/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2}.$$

- The variance in the combined estimate is: $$\sigma_w^2 = \frac{1}{\sum_i^n 1/\sigma_i^2}$$

- This optimal combination of data is helpful if the errors in the data are dominated by statistics, not systematics errors.

Exercises

- **Exercise 2:**

We have N=10 measurements of a quantity y (6.8±2.0, 6.5±1.1, 4.3±1.7, 5.5±0.5, 6.0±2.5, 7.1±1.3, 4.7±1.2, 5.8±1.1, 6.5±0.5, 5.4±2.6). What is the optimal estimate of this quantity and the error in that estimate?

(measurements of phone screen size!)

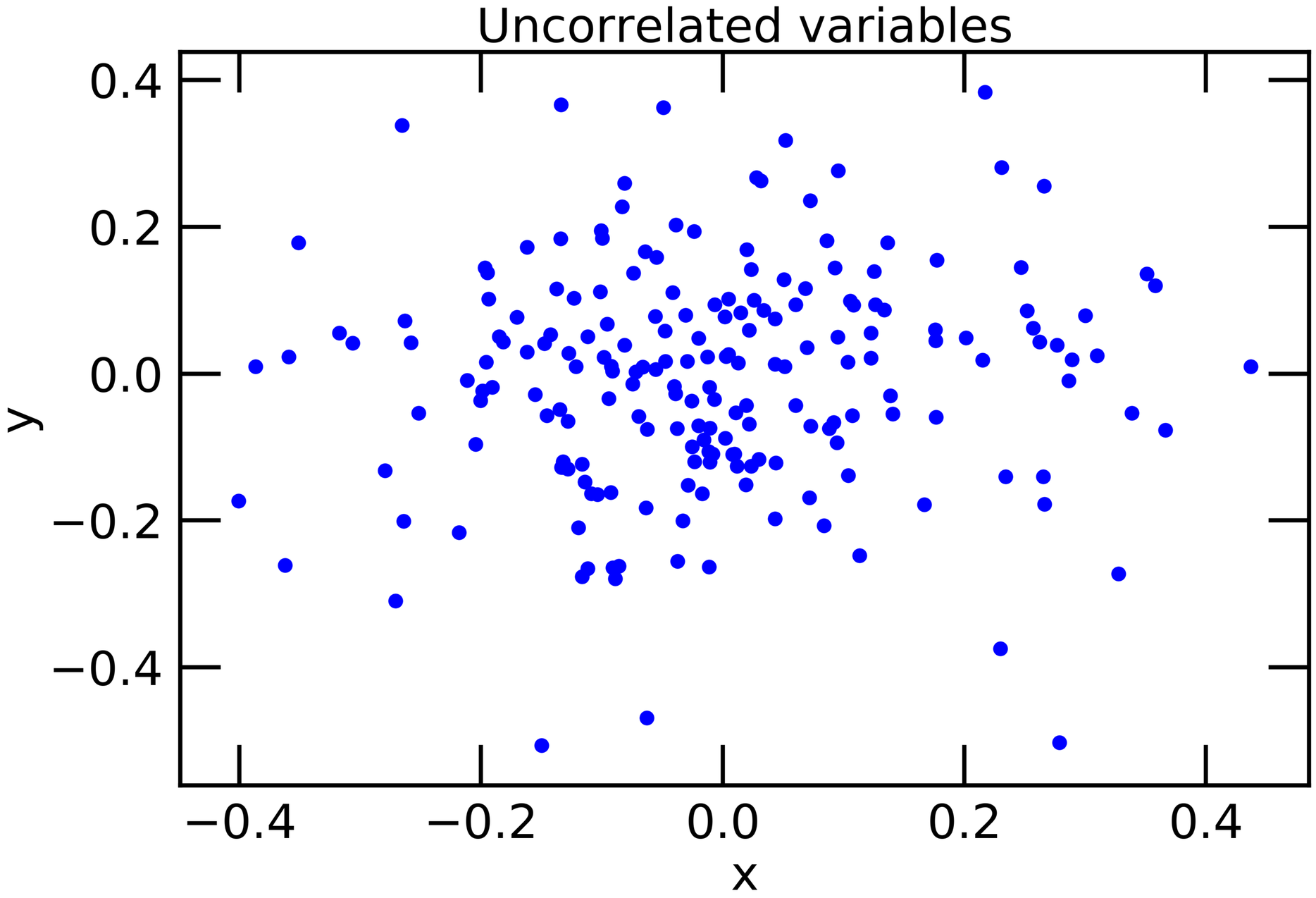
- A further measurement 5.0±0.2 is added. How should our estimate change?

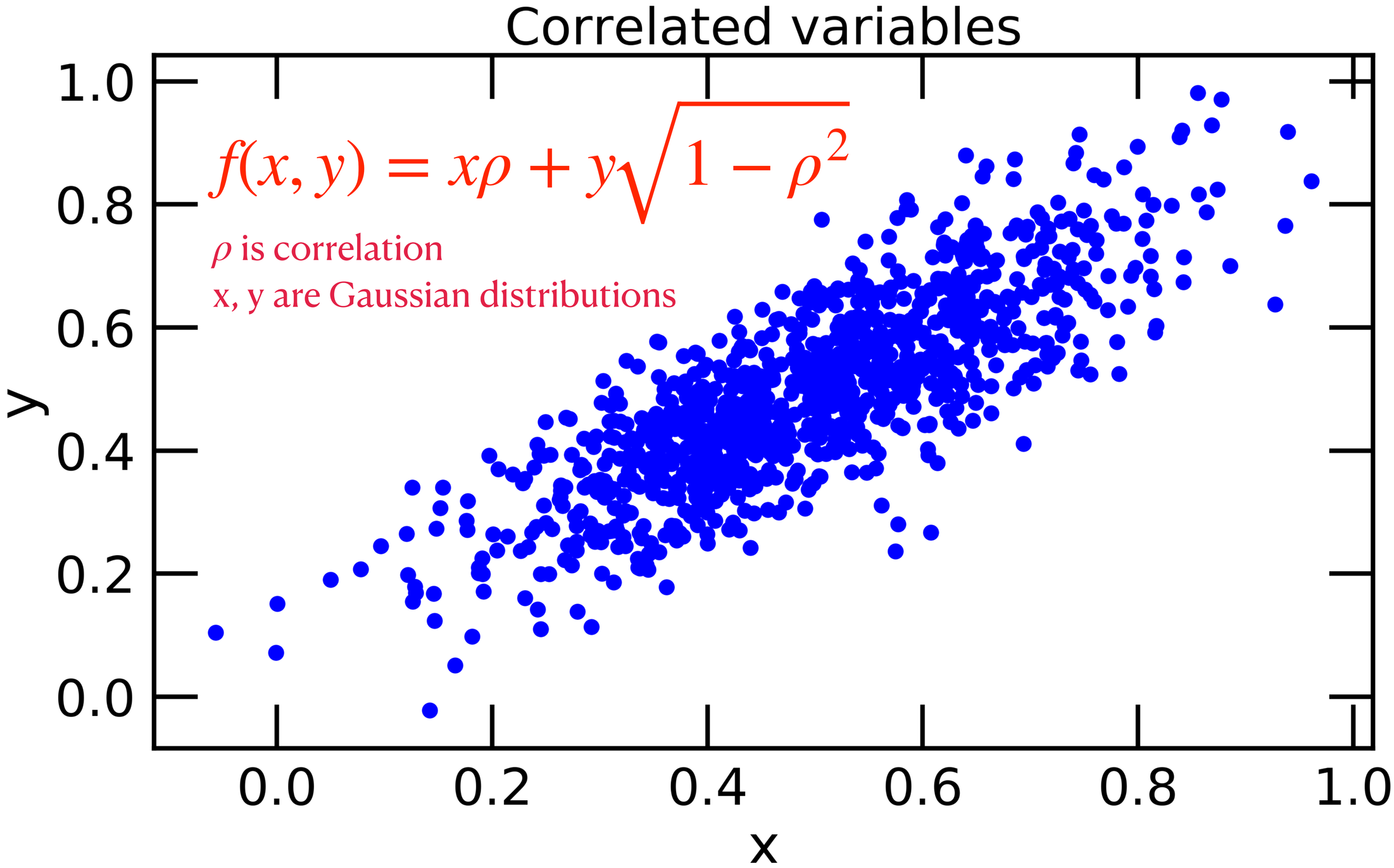
2. Correlation

In this section, we will learn how to quantity the correlation of variables.

- The correlation between variables can figure out some **underlying physical processes**.
- The variables are correlated if they share dependent statistic.

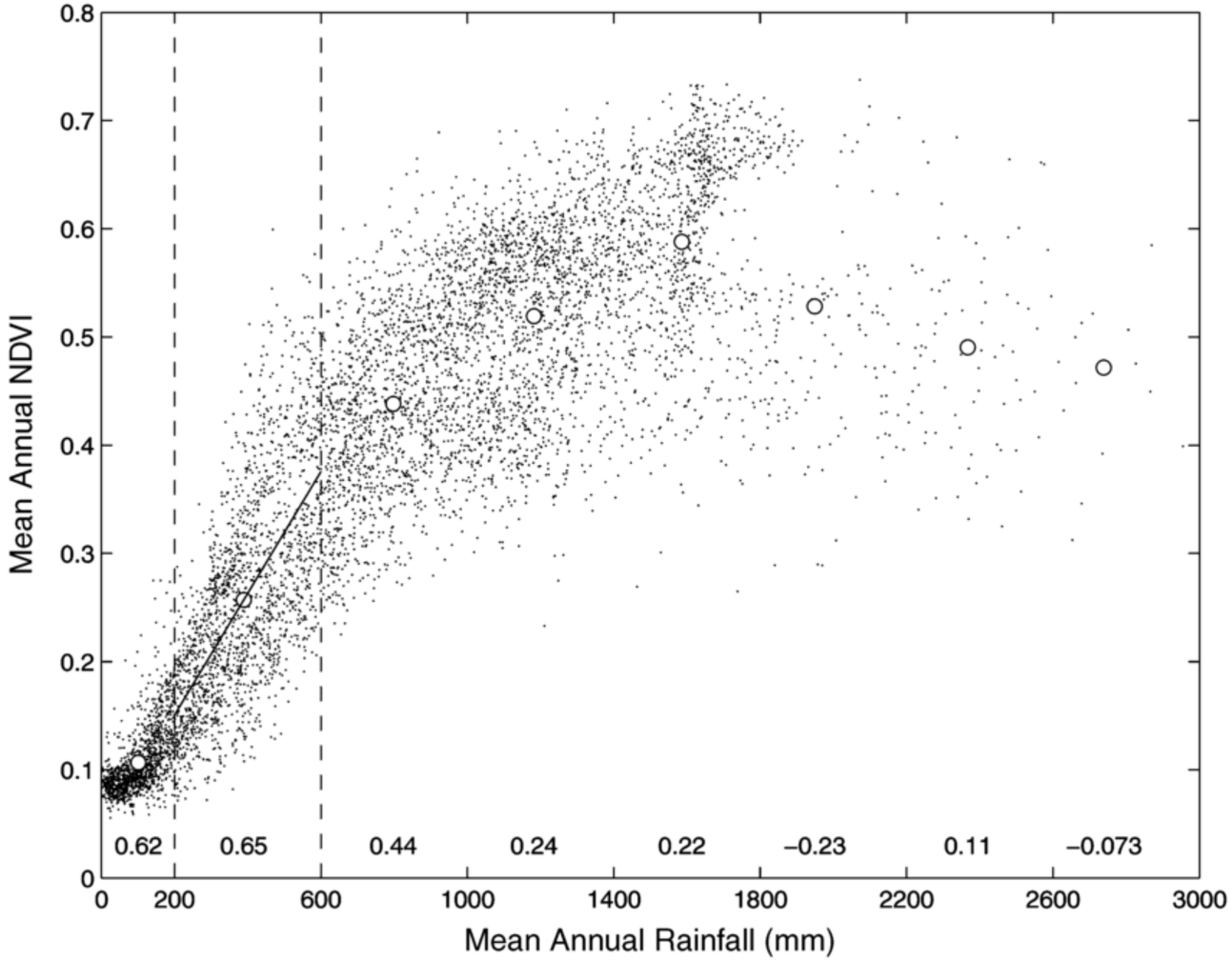
2. Correlation



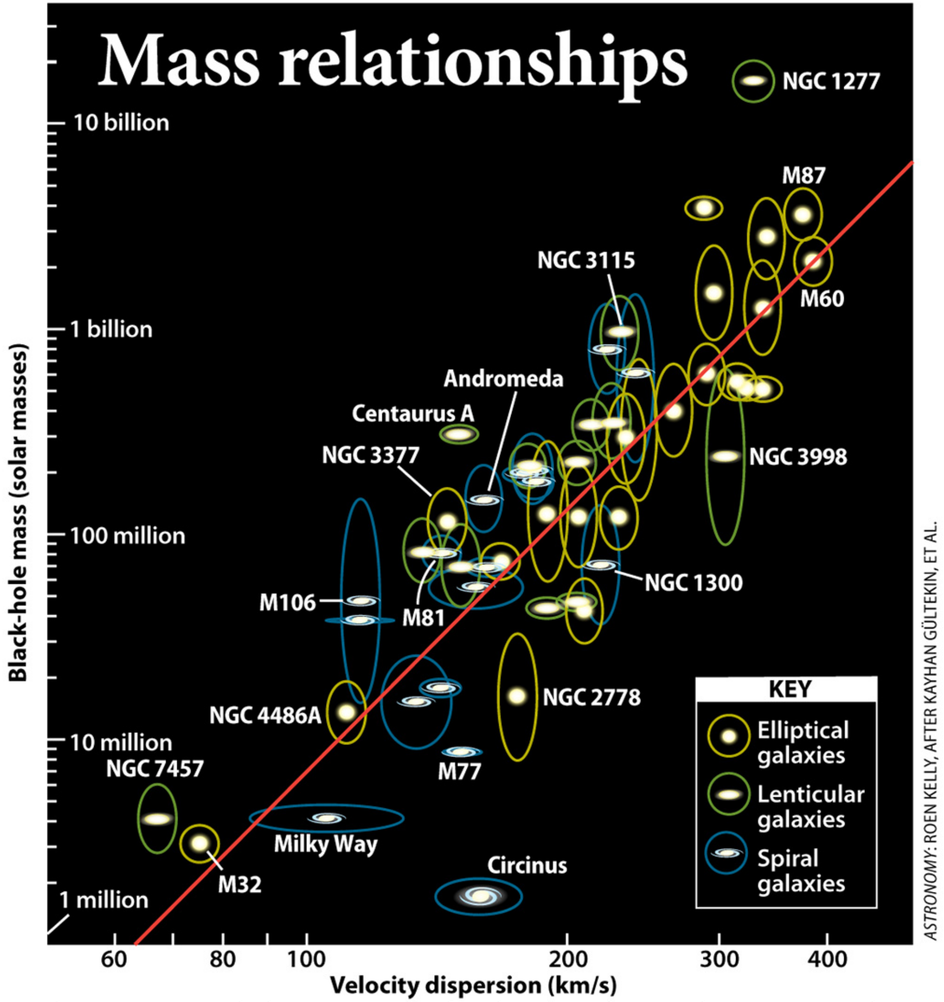

Correlated variables
$f(x,y) = x\rho + y\sqrt{1-\rho^2}$
ρ is correlation
x, y are Gaussian distributions
1.0
0.8
0.6
0.4
0.2
0.0
y
0.0
0.2
0.4
0.6
0.8
1.0
x

2. Correlation in science

Compared regimes of NDVI (Normalized difference vegetation index) and Rainfall in semi-arid regions of Africa



Black Hole mass and velocity dispersion of galaxies relation



The statistical dispersion of velocity about the mean velocity of Black Hole: M–sigma relation

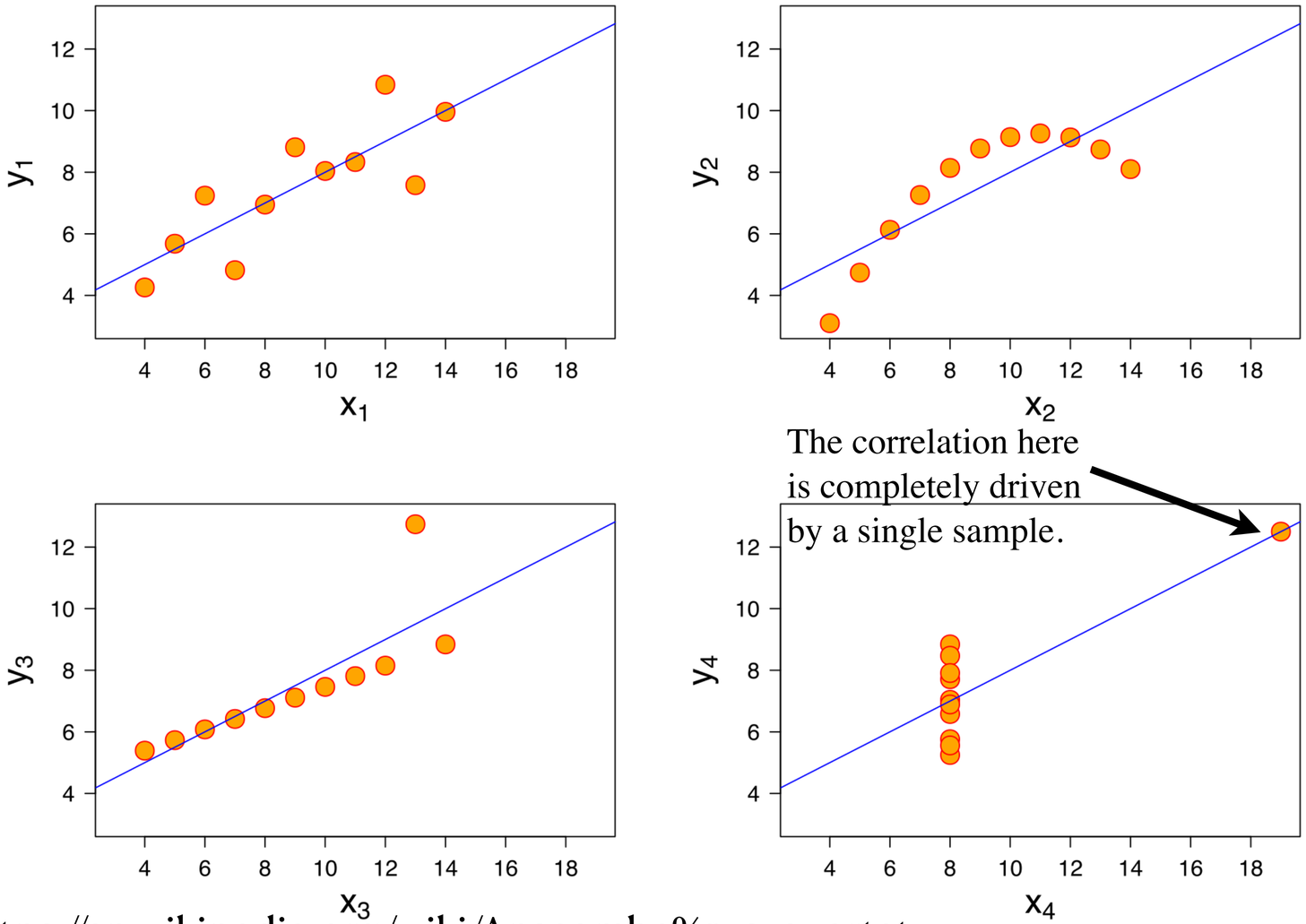
Credit: https://en.wikipedia.org/wiki/M%E2%80%93sigma_relation

2. Pitfalls in searching for correlation

Correlations can be driven by a small number sample.

The following four (x, y) datasets all have the same mean, variance, correlation coefficient and regression line:

Mean of x	9
Variance x	11
Mean of y	7.50
Variance of y	4.125
Correlation	0.816
Linear regression	y=0.500x + 3.00



Spurious Correlation example:
https://www.tylervigen.com/spurious-correlations

Credit: Anscombe's quartet: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

2.1. Correlation coefficient ρ

The correlation coefficient measures the strength of the correlation between two variables. If the variables (x,y) have mean (μ_x, μ_y) and standard deviation (σ_x, σ_y), the theoretical correlation coefficient:

$$\rho = \frac{\langle (x-\mu_x)(y-\mu_y) \rangle}{\sigma_x \sigma_y} = \frac{\langle xy \rangle - \mu_x \mu_y}{\sigma_x \sigma_y}$$

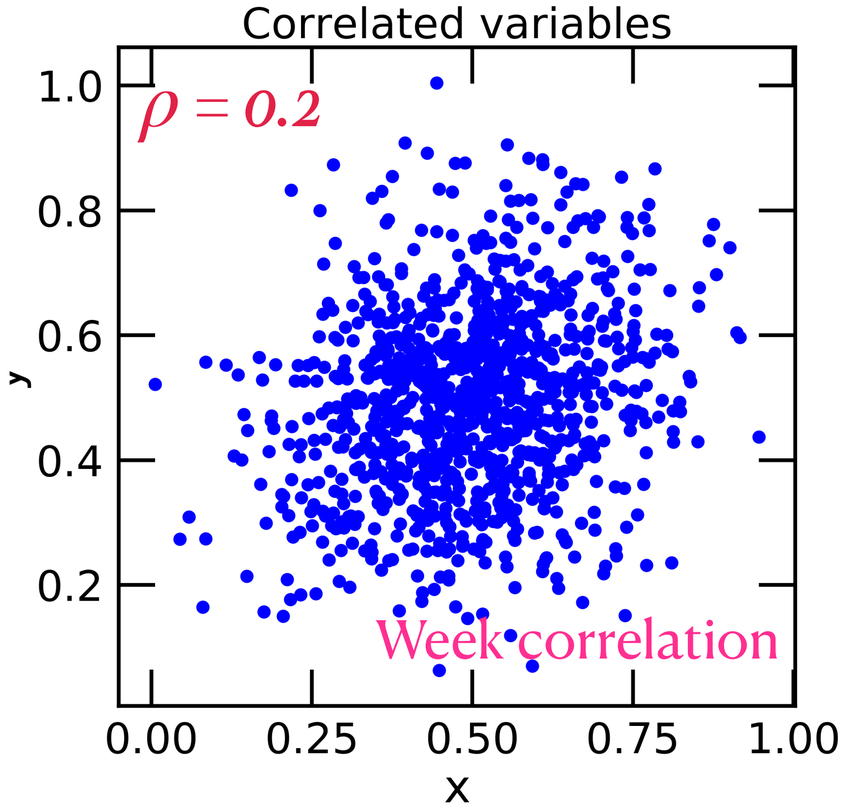
$$\langle xy \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, y\, P(x,y)\, dx\, dy$$

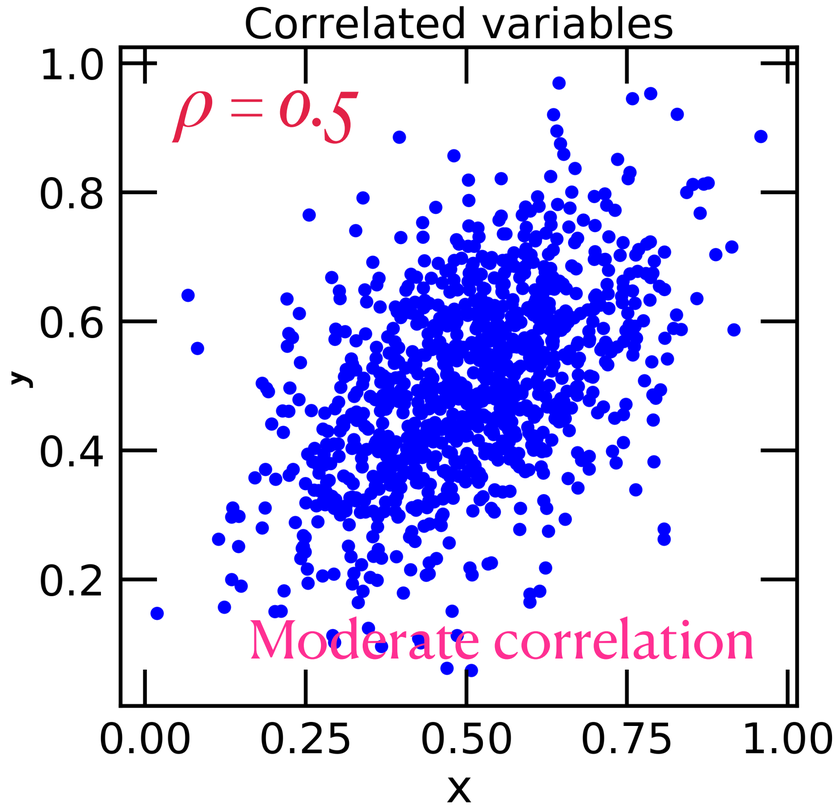
The range of value $[-1 \leq \rho \leq 1]$

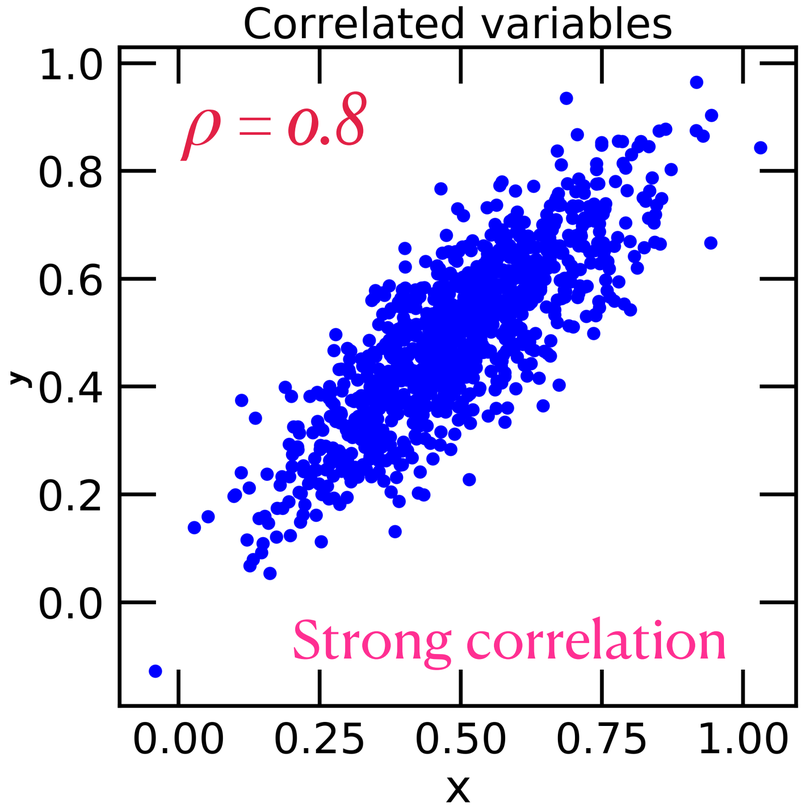
2.1. Correlation coefficient

$$\rho = \frac{\langle (x-\mu_x)(y-\mu_y)\rangle}{\sigma_x \sigma_y} = \frac{\langle xy\rangle - \mu_x\,\mu_y}{\sigma_x\,\sigma_y}$$

- $\rho = 0$: **No correlation**
- $\rho = 1$: **Complete correlation**
- $\rho = -1$: **Anti-correlation**

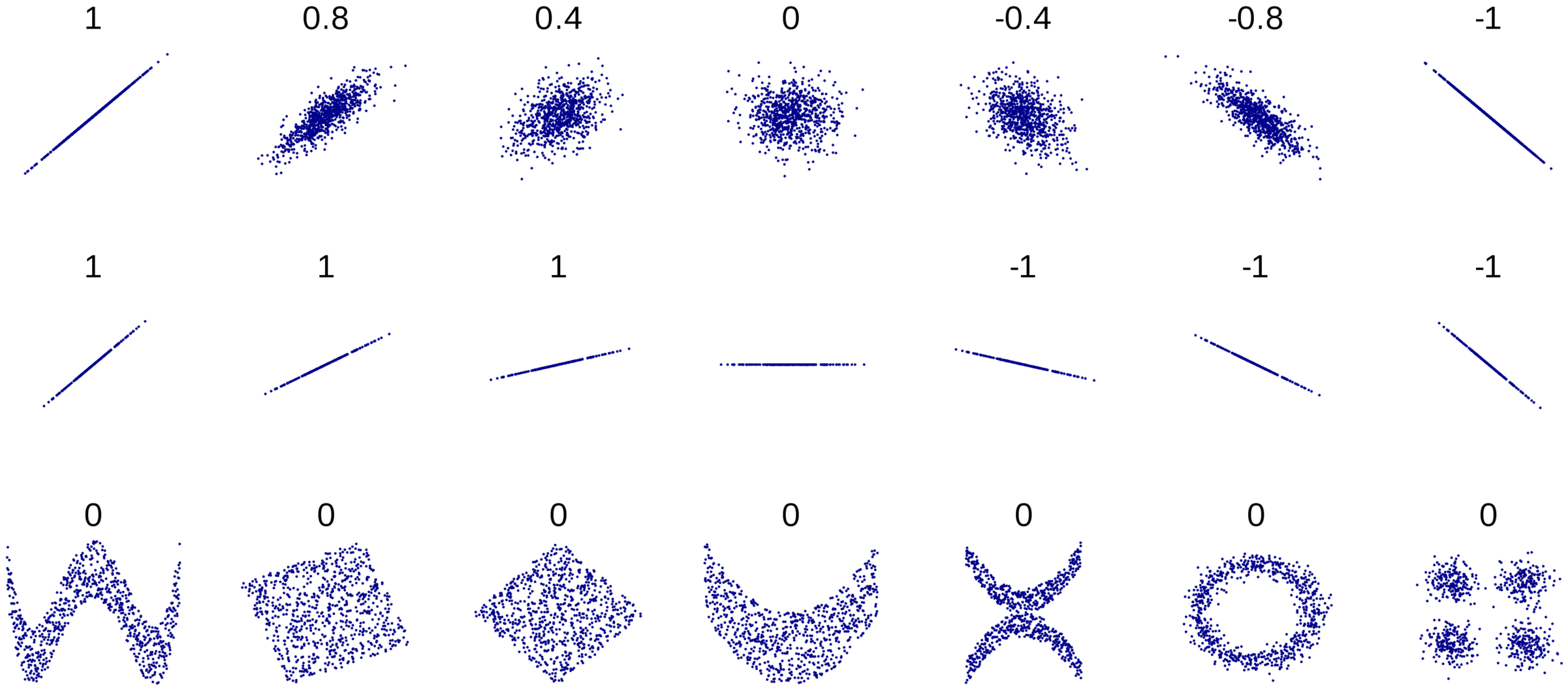






2.1. Correlation coefficient

The correlation reflects the **strength** and **direction** of a linear relationship (top row), but *not the slope* of that relationship (middle), nor many aspects of *nonlinear relationships*.



https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

2.1. Pearson product-moment correlation

- We can estimate the correlation coefficient using Pearson product moment formula:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{(N-1)\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}$$

- r is an estimator of ρ.

- The uncertainty in the measurement of r: $\sigma(r) = \sqrt{\frac{1 - r^2}{N - 2}}$

2.1. Pearson product-moment correlation

- After test correlation, we need to check the **significance of the correlation.**
- We can use hypothesis tests to t-Test statistic,

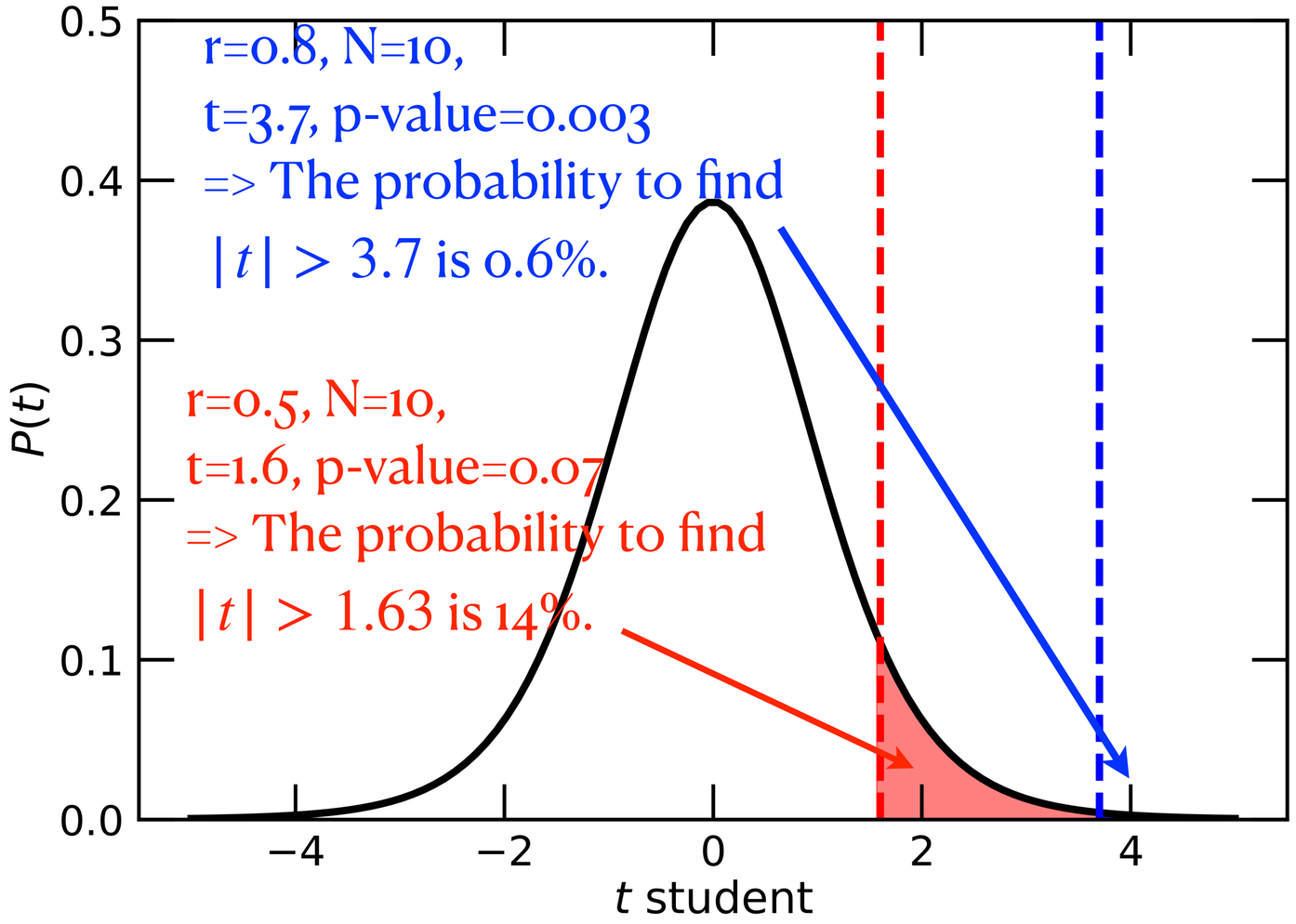
$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

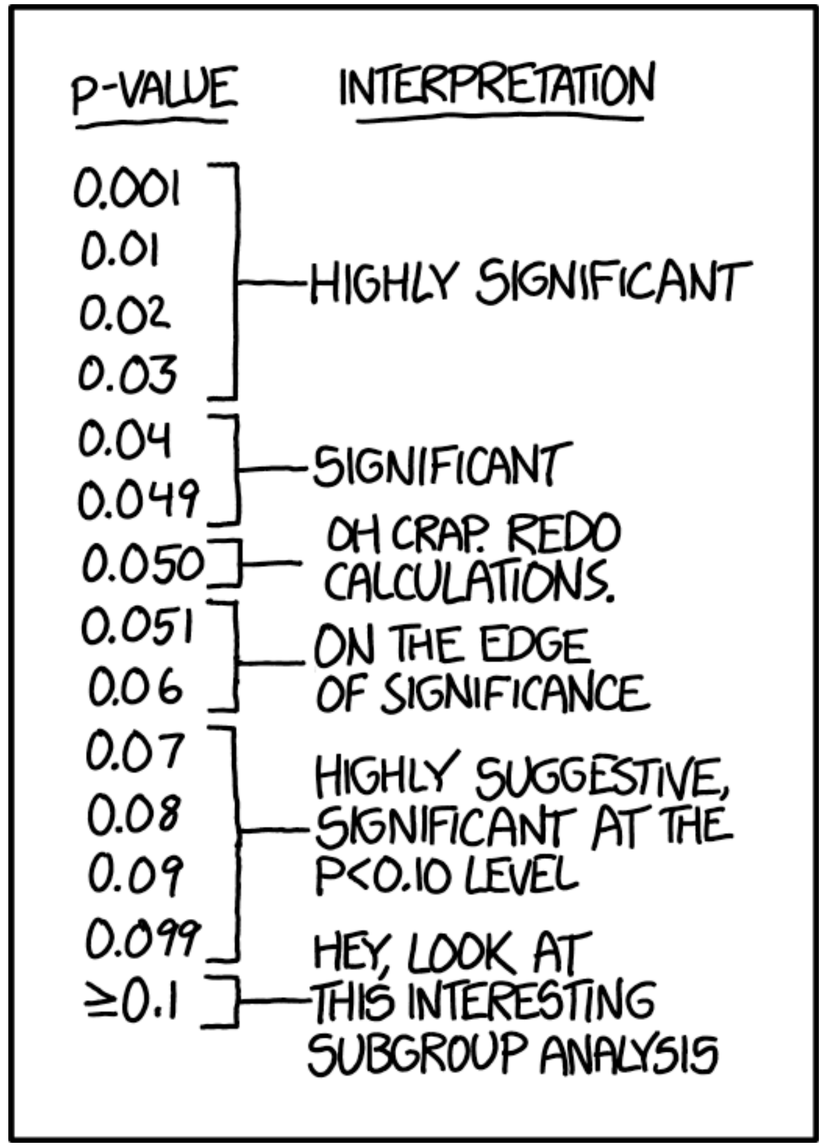
- **Example**:

We measure r = 0.8 for N = 10 data points. Is this correlation significant?

2.2 p-value

- The p-value is correspond to the significance of the result:





2. Hubble and Lemaitre datasets

Exercise: For the two datasets, determine the Pearson correlation coefficient, its error and statistical significance using t-test.