Fabio D'Andrea
LMD – 4^{e} étage
01 44 32 22 31

dandrea@lmd.ens.fr

http://www.lmd.ens.fr/dandrea/TEACH

Program

29/1 Elementary statistics – 1
5/2 Elementary statistics - 2
12/2 Exercises – Computer room

19/2 Fourier Analysis -1
26/2 Fourier Analysis -2, stochastic processes
5/3 Exercises – Computer room
12/3 Exercises – Computer room

19/3 Principal component analysis -1
26/3 Principal component analysis -2
16/4 Exercises – Computer room
/4 Exercises – Computer room

/5 Cluster analysis
/5 Exercises – Computer room

/5 Principal component analysis: Complements

/6 We'll see.

/6 Exam

Further reading:

H. Von Storch and F. Zwiers. Statistical Analysis for Climate Research
Cambridge University Press, 1999

http://www.statsoft.com/Textbook

Statistical software

R, SAS, Matlab, Pyton…..

Do it yourself or use packaged routines??

. Lesson 1.
Introduction to elementary statistics

What exactly is statistics ?

The purpose of statistics is to develop and apply methodology for extracting useful knowledge from both experiments and data. In addition to its fundamental role in data analysis, statistical reasoning is also extremely useful in data collection (design of experiments and surveys) and also in guiding proper scientific inference (Fisher, 1990).

Statistical data analysis can be subdivided into :

descriptive statistics
inferential statistics.

Descriptive statistics is concerned with exploring and describing a sample of data, whereas inferential statistics uses statistics from sample of data to make statements about the whole population.

We will see how we describe random variables, by their mean, variance, and p.d.f, and how we compare different random variables.

A random variable can be thought of as an unknown value that may change every time it is inspected. Thus, a random variable is a function mapping the sample space of a random process (a physical process) to the space of real numbers.

EXAMPLES

Discrete	**Continuous**
number of people in a car	total weight of people in a car
number of cars in a parking lot	distance between cars in a parking lot
number of phone calls to 911	time between calls to 911.

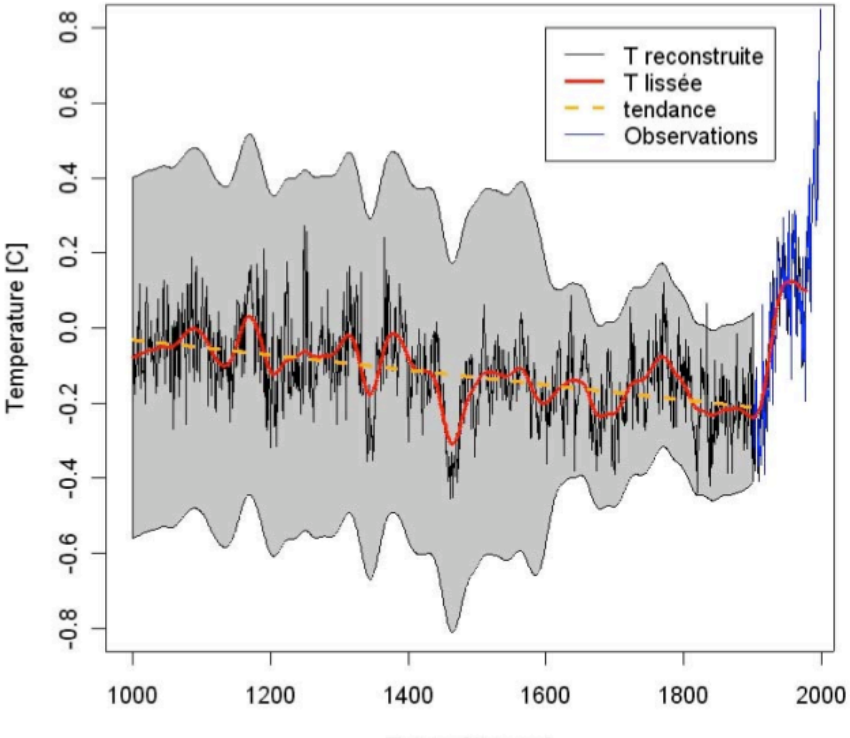
In geophysics we often have to do with data – typically with time series:

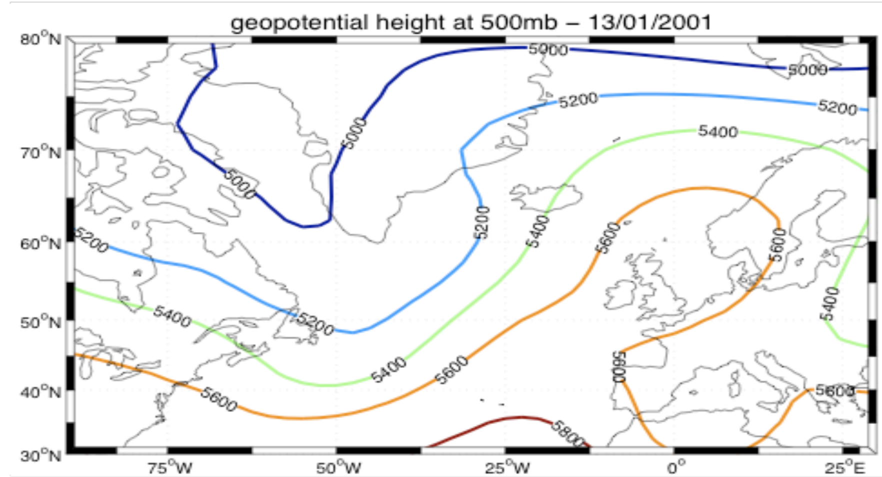
Let's call a timeseries x(t), or, in the discrete case, x_i where i=1,2,.....,N

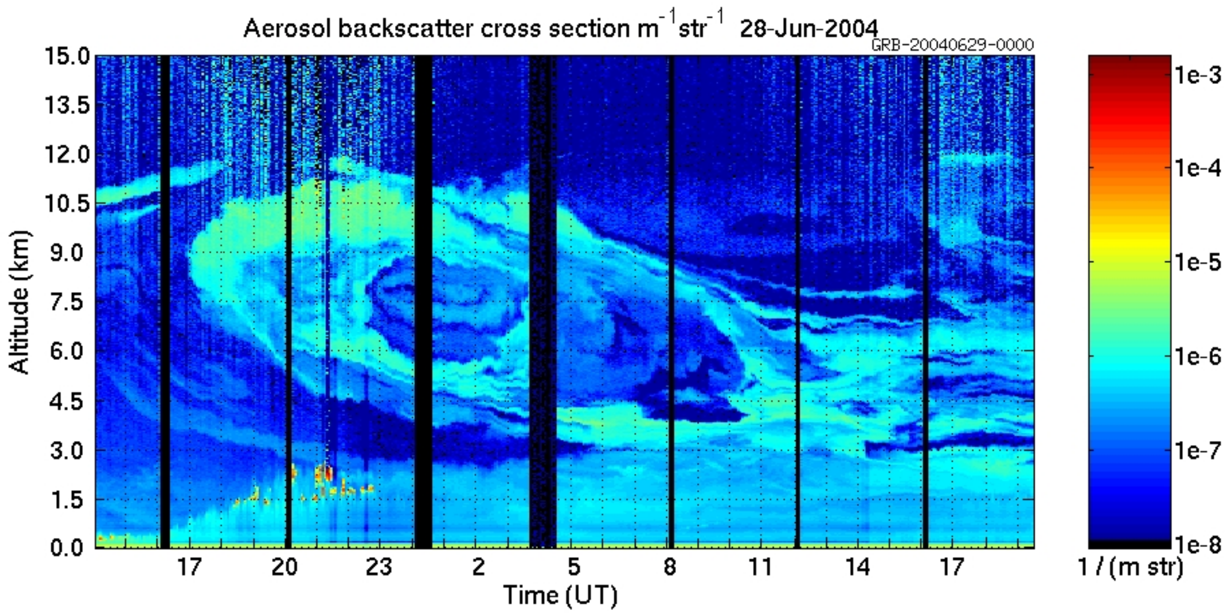
x_i can be a scalar number, or also it can be a vector. $\mathbf{x}_i \in \Re^d$

Time series analysis is a sub-field of statistics. It is declined in two main fields:

time domain methods and **frequency domain methods**.







Scalar and vector examples

Descriptive vs Inference

Descriptive Statistics

Gives numerical and graphic procedures to summarize a collection of data in a clear and understandable way

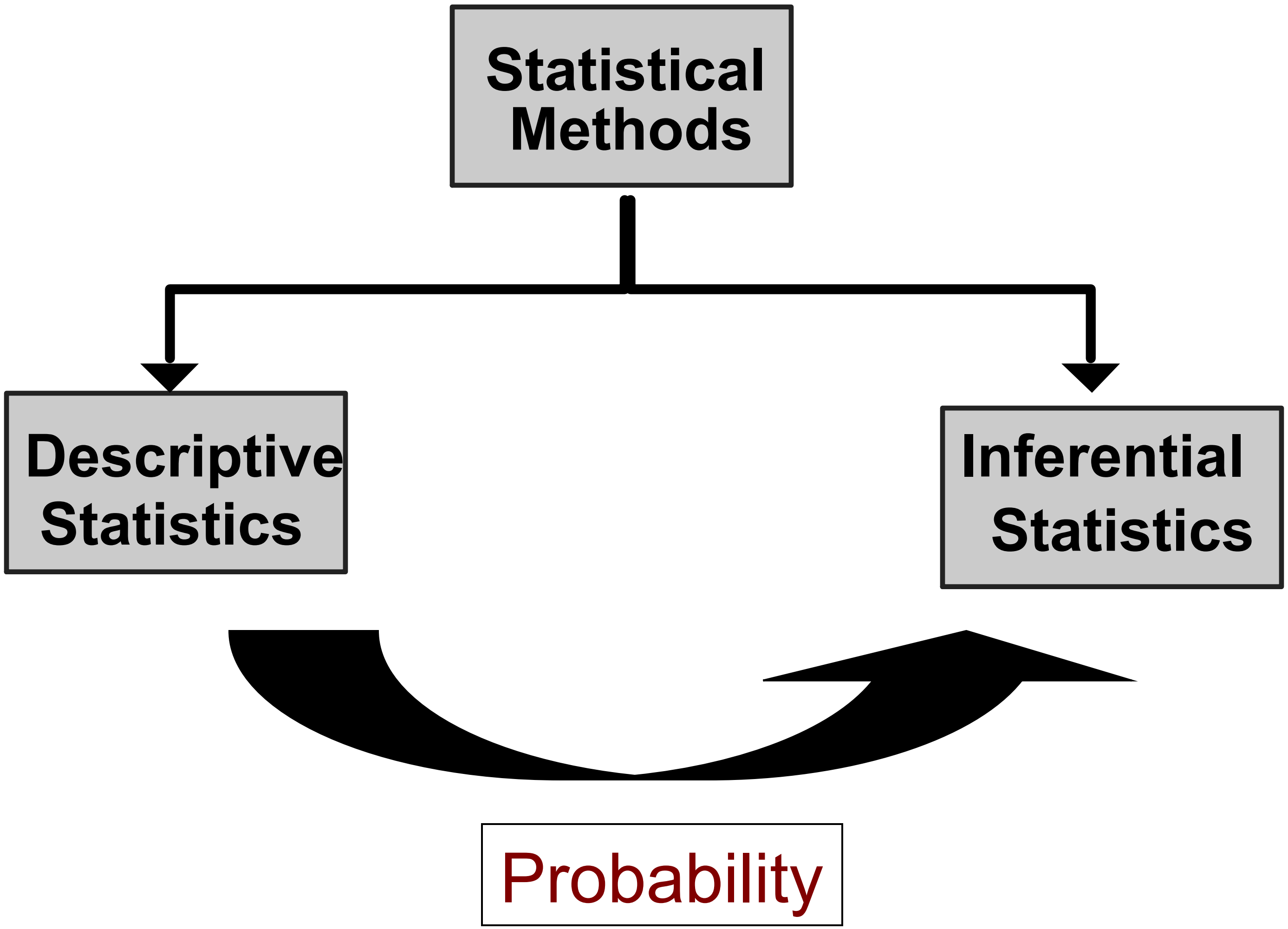
→ Finding ways to summarize the important characteristics of a dataset

Inferential Statistics

Provides procedures to draw inferences about a population from a sample

→ How and when to generalize from a sample dataset to the larger population

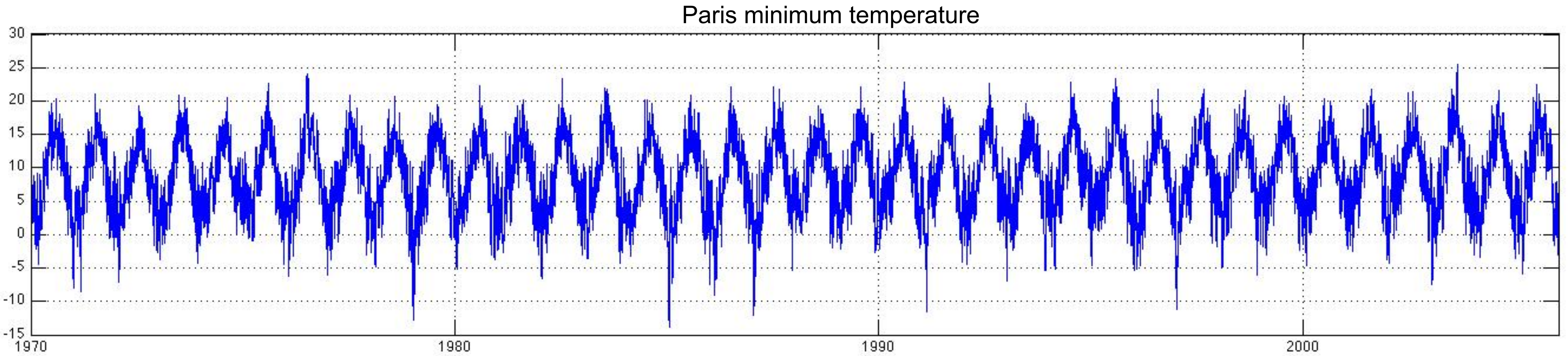
Statistical Methods



Elementary description of data: the table

	Name	Height (cm)
1	Fabio	180
2	Francois	181
3	Jean-Philippe	185
4	Loic	172
5	Ara	176
6	Alvaro	172
7	Ann'Sophie	170
8	Xavier	180
9	Guillaume	183
10	Hector	171
11	Bernard	182
12	Michael	177
13	Marta	162
14	Tonia	178

The graph:



Basic attributes of data population:

N **population size**

$$M(x) = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$ **mean**

$$V(x) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$ **variance**

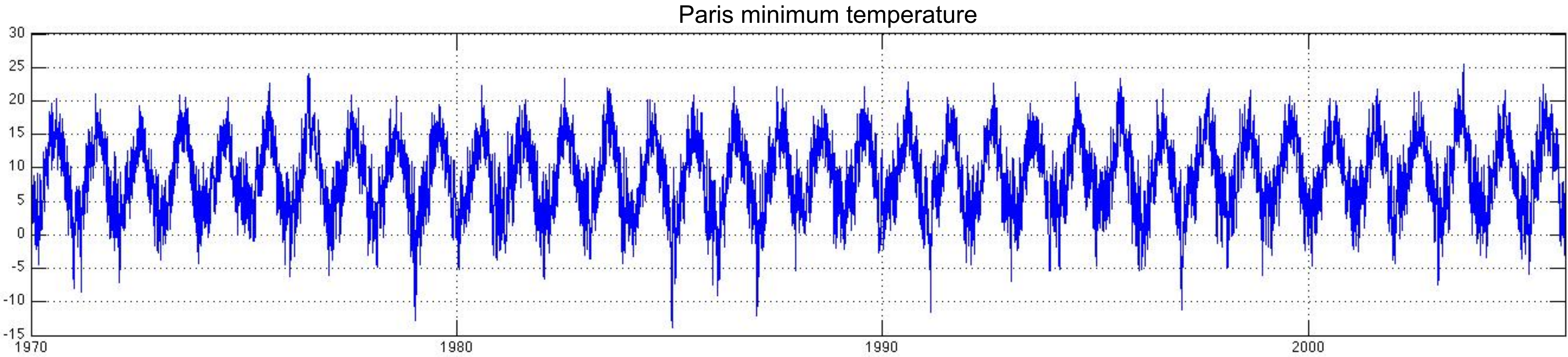
σ is the **standard deviation**

	Name	Height (cm)
1	Fabio	180
2	Francois	181
3	Jean-Philippe	185
4	Loic	172
5	Ara	176
6	Alvaro	172
7	Ann'Sophie	170
8	Xavier	180
9	Guillaume	183
10	Hector	171
11	Bernard	182
12	Michael	177
13	Marta	162
14	Tonia	178

$$N = 14$$

$$\mu = 176.36$$

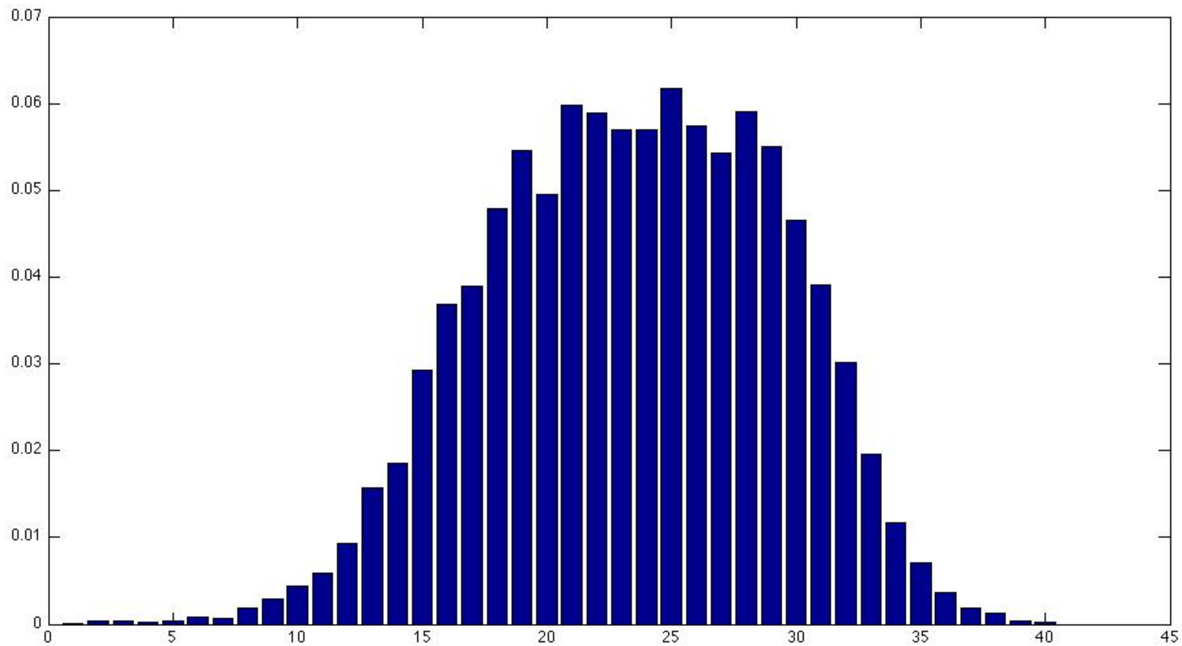
$$\sigma = 6.30$$



$$N = 13149$$

$$\mu = 8.68$$

$$\sigma = 5.70$$



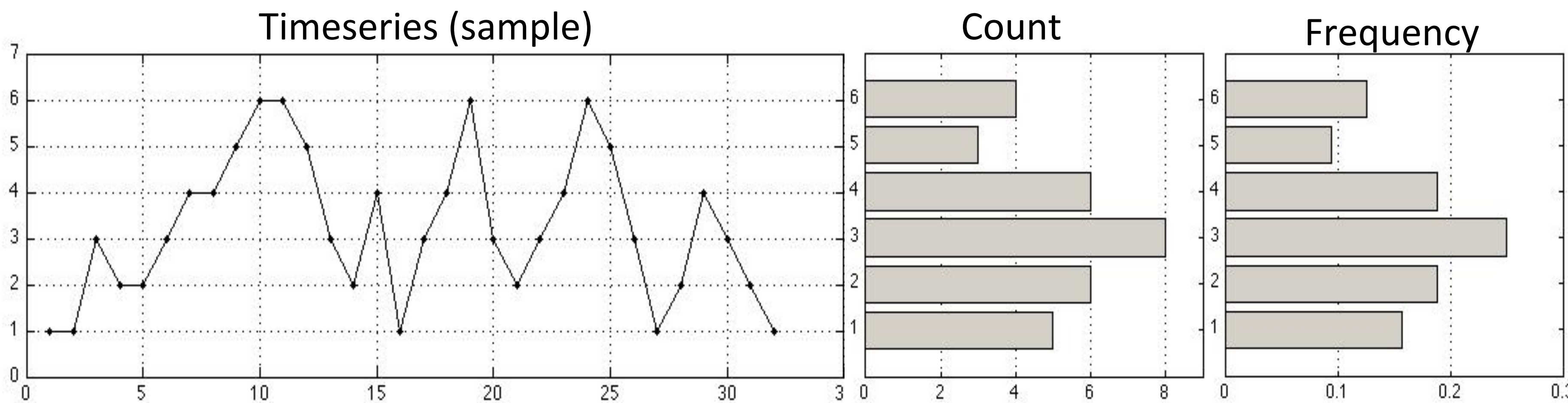
The mean:

$$M(x) = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Equivalent to

The mathematical expectation, or expected value:

$$E(x) = \int_{-\infty}^{\infty} x f(x)\,dx$$



$$\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{n=1}^{6} \hat{x}_n N(\hat{x}) = \sum_{n=1}^{6} \hat{x}_n f(\hat{x})$$

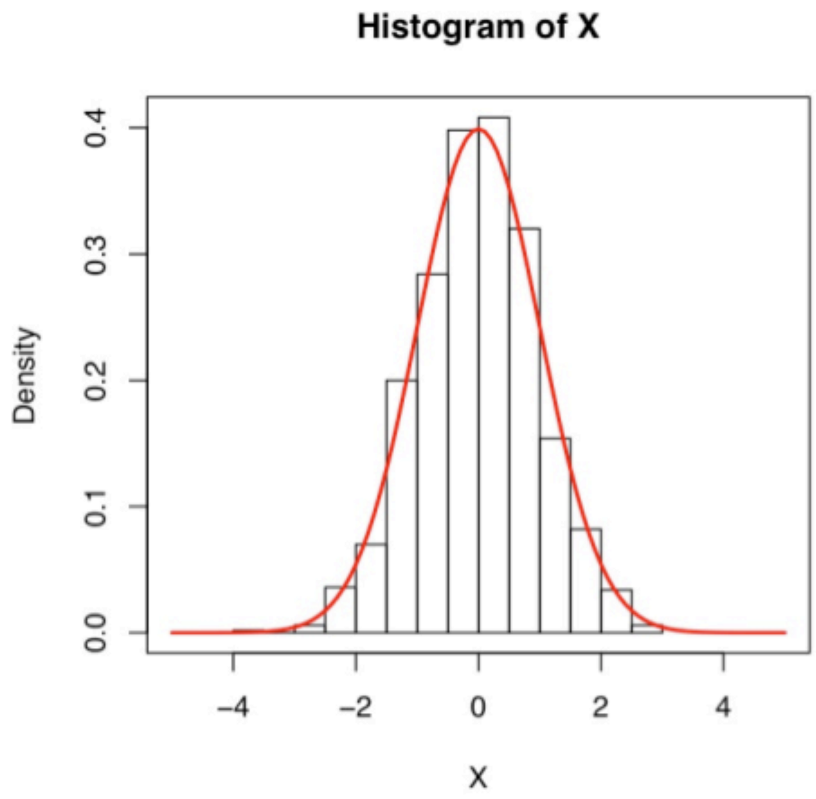
In the case of a real and continuous random variable, the probability density function (PDF) is a real function *f* for which:

$$P(a \le x \le b) = \int_a^b f(x)\,dx$$

It has the properties of a distribution, hence:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\forall x\,;\, f(x) \ge 0$$



In the same way as the mathematical expectation, or mean:

The second moment is the the variance:

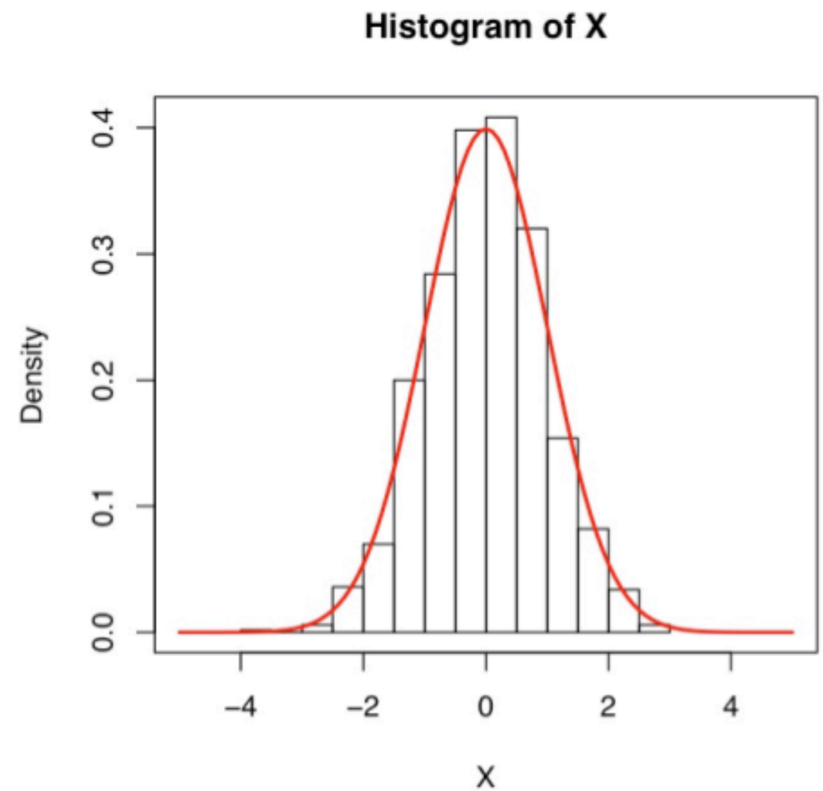
$$\mathrm{var}(x) = E((x - E(x))^2) = \int_{-\infty}^{\infty} (x - E(x))^2 f(x)\,dx$$

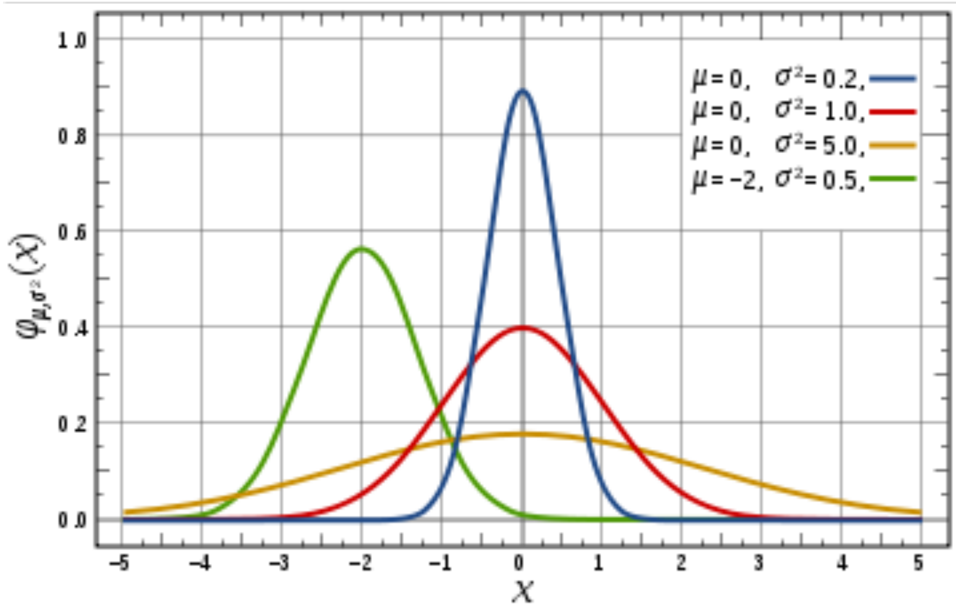
In the discrete case:

$$\mathrm{var}(x) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - E(x))^2$$

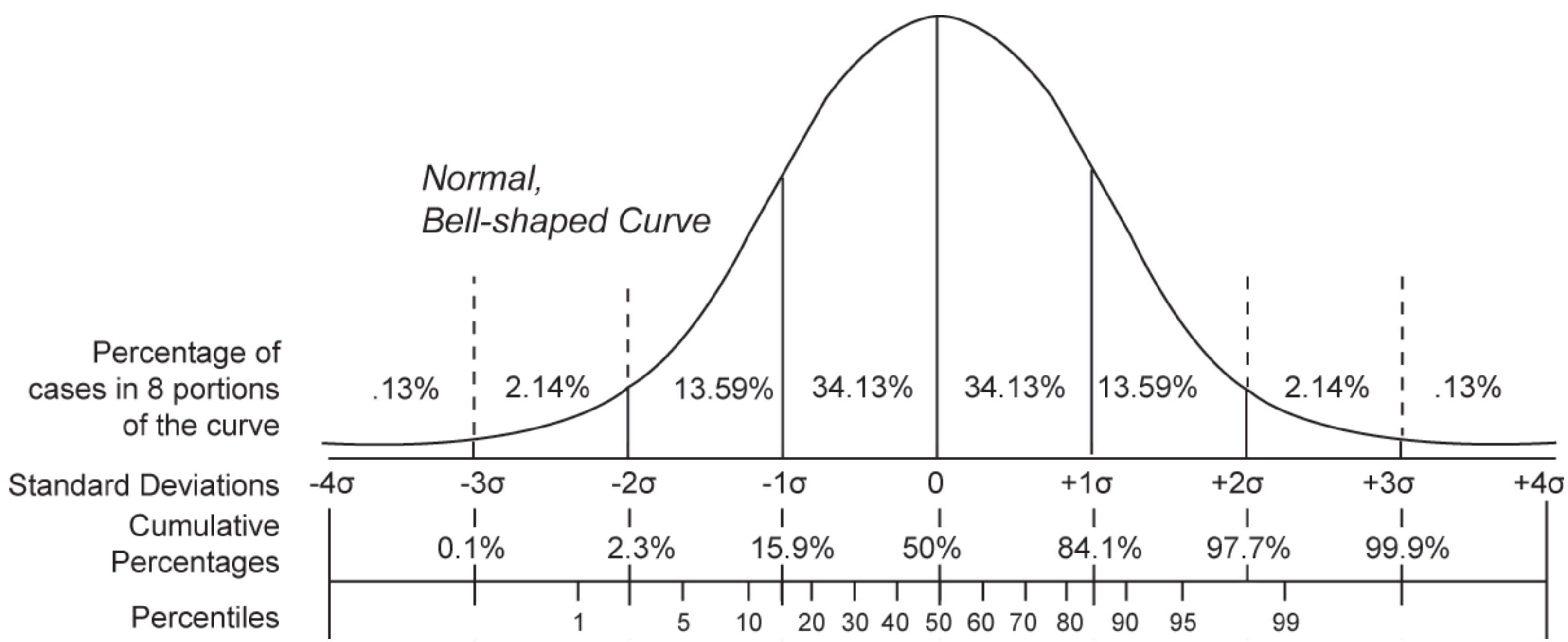
Normal distribution, or Gaussian distribution.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$





$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Higher order moments

$$m_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^r$$

$$a_3 = \frac{m_3}{\sigma^3}$$

$$a_4 = \frac{m_4}{\sigma^4}$$

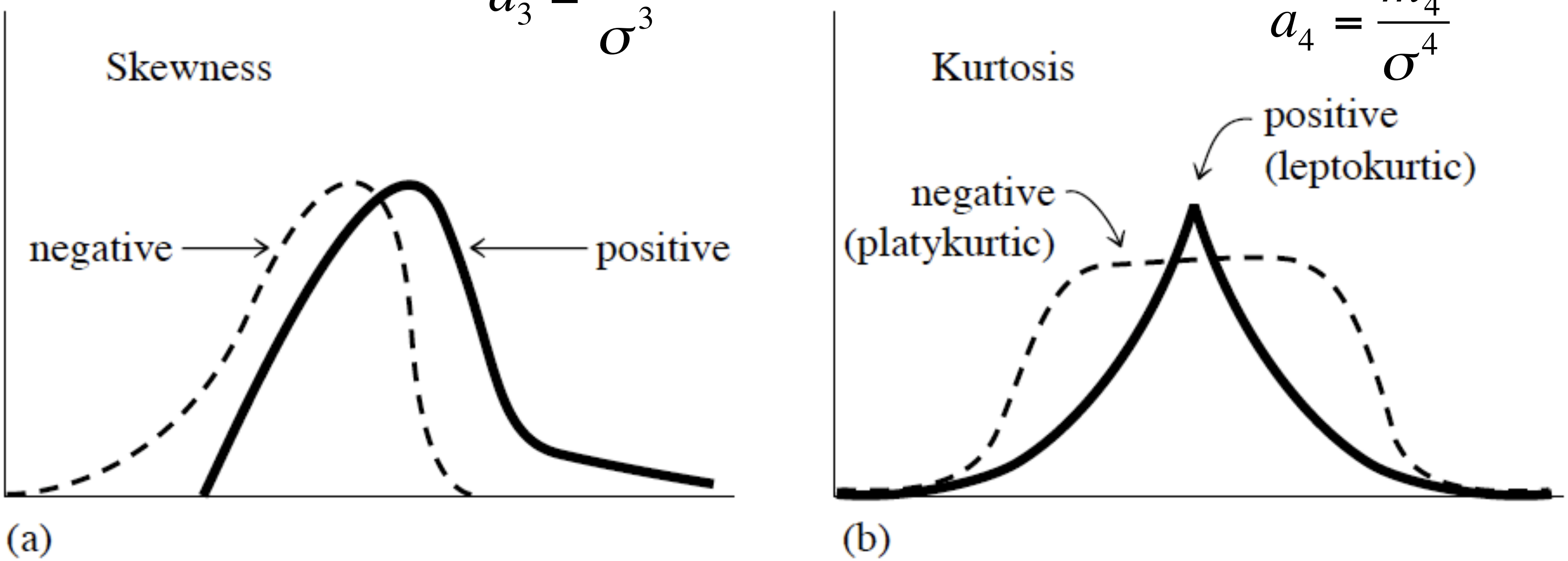
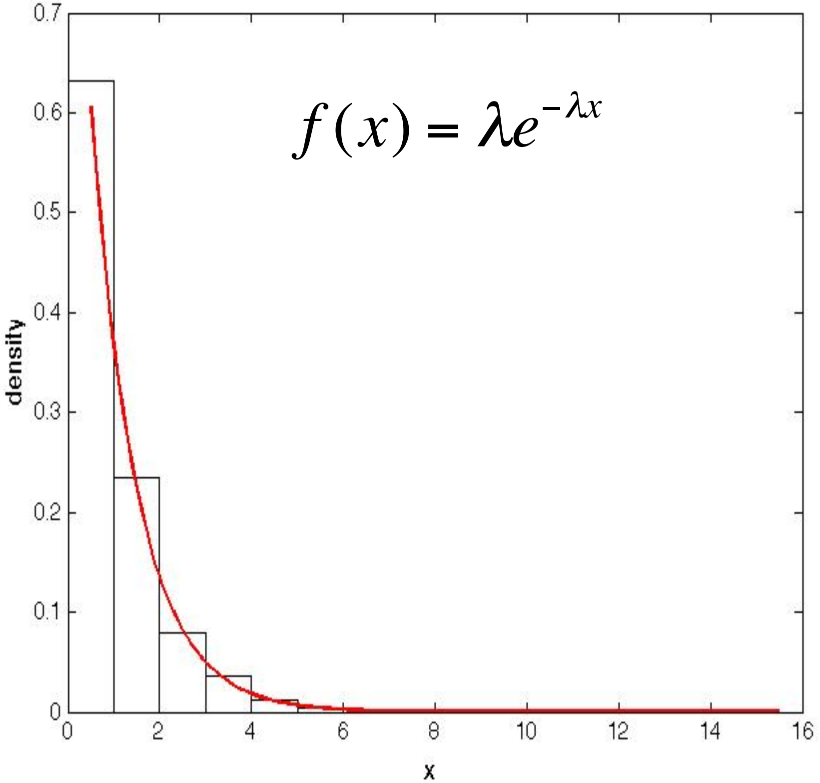


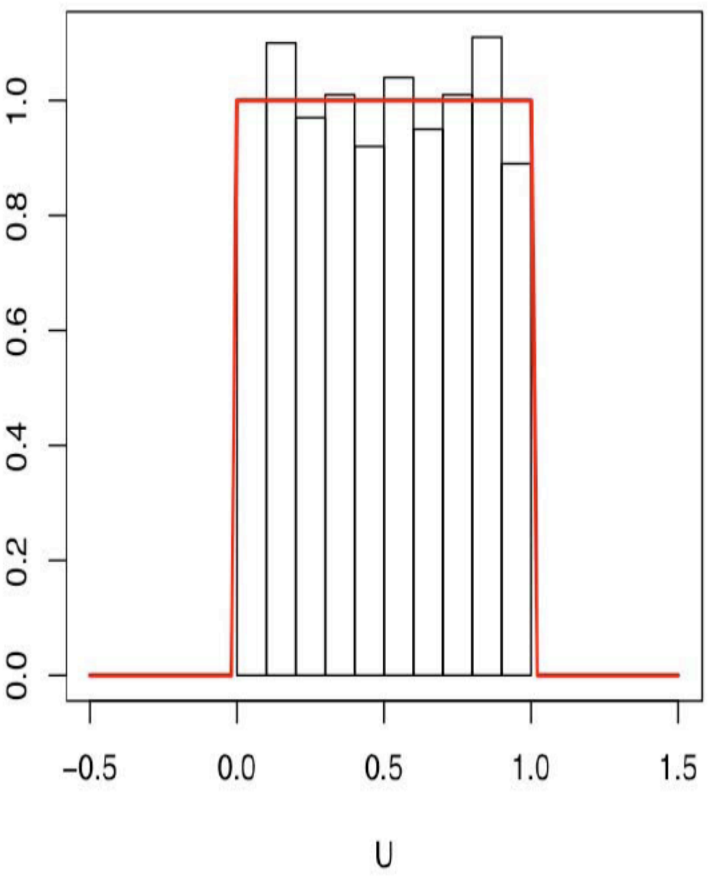
Figure 14.1.1. Distributions whose third and fourth moments are significantly different from a normal (Gaussian) distribution. (a) Skewness or third moment. (b) Kurtosis or fourth moment.

Other distributions:

Exponential



Flat



Interpretation of probability:

1) Inverse of the number of possible ways of happening of an event (classical interpretation)
2) Frequency of a repeated event (frequentist interpretation)
3) Non-frequentist subjective interpretation - Bayes
4) Mathematical definition (Kolmogorov)

An event A is a set (or group) of possible outcomes of an uncertain process

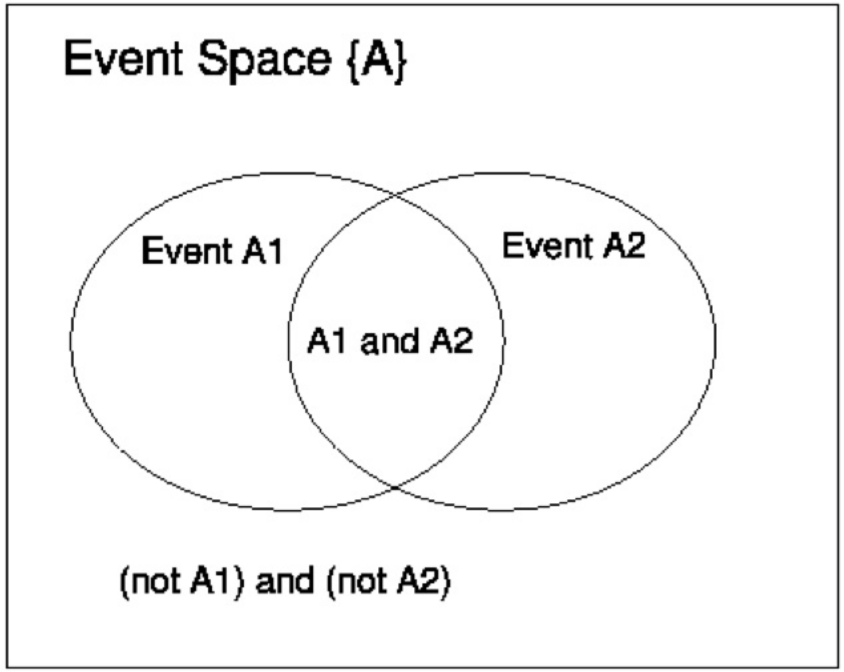
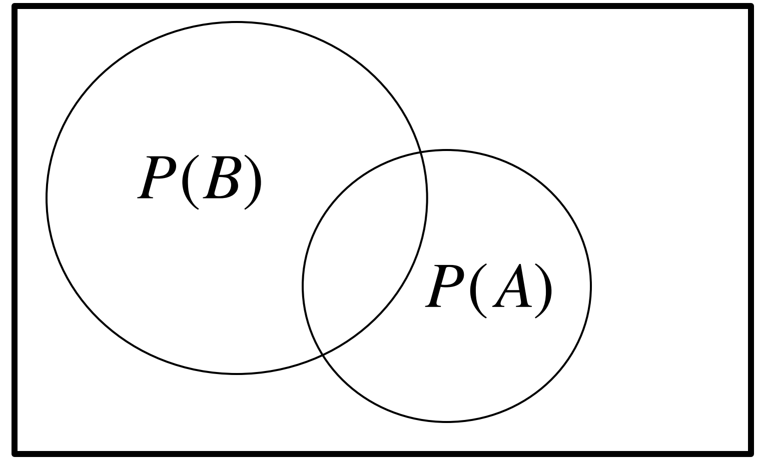


Figure 3.1: Euler diagram showing event space for 2 events.

A random variable, X, is a label allocated to a random event A.
The probability is a function defined on the event space,

$$P(A) = P(X = x) \in [0,1]$$



Some properties of probability

1) $P(\neg A) = P(A^c) = 1 - P(A)$

2) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

3) $P(A \cap B) = P(A|B)P(B)$

$P(A \cap B) = P(A)P(B)$

From (3), we can also write: (Bayes Theorem)

$$P(A|B)P(B) = P(B|A)P(A)$$

Exercise

The probability of a New York teenager owning a skateboard is 0.37, of owning a bicycle is 0.81, and of owning both is 0.36.

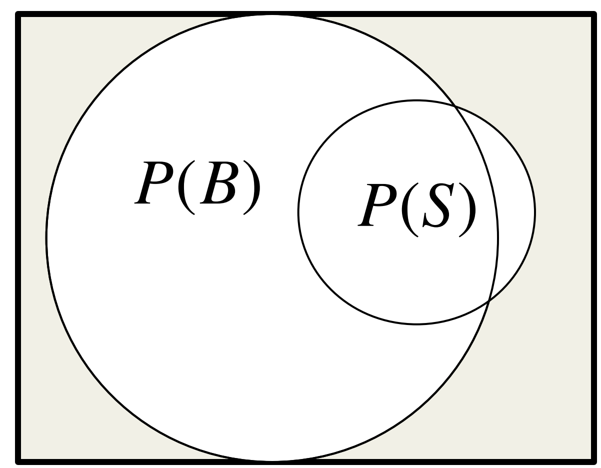
If a New York teenager is chosen at random, what is the probability that he/she does not own neither a skateboard nor a bicycle?

Exercise

The probability of a New York teenager owning a skateboard is 0.37, of owning a bicycle is 0.81, and of owning both is 0.36.

If a New York teenager is chosen at random, what is the probability that he/she does not own neither a skateboard nor a bicycle?

$$P(S \cup B) = P(S) + P(B) - P(S \cap B) =$$
$$= 0.37 + 0.81 - 0.36 = 0.82$$



$$P(\overline{S \cup B}) = 1 - P(S \cup B) =$$
$$1 - 0.82 = 0.18$$

A medical example of the Bayes theorem.

1) If an individual has the disease, the test is positive 99% of the times
2) If an individual doesn't have the disease, the test is positive 1% of the times
3) The population as a whole has probability 0.01% of having the disease

$1)\ P(B \mid A) = 0.99\ (correct\ \ positive)$

$2)\ P(B \mid A^c) = 0.01\ (false\ positive)$

$3)\ P(A) = 0.0001\ (one\ person\ out\ of\ 10000\ has\ the\ disease)$

A: event "having the disease
B: event "test is positive"

If you go to the doctor and you test positive, what is the probability that you have the disease? i.e., what is $P(A \mid B)$?

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)} = \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} =$$
$$= 0.0098 \approx 1\%$$

Don't worry and repeat the test a few times!
(*if you take the test again and you are positive, the probability is 50%, if you take a third, and are positive again, it's 99%...*)

Exercise

Among the clients of a gas station, 35% buy normal gasoline. 40% unleaded gasoline and 25% super-enhanced gasoline. Of those using normal gasoline, 60% systematically get a full tank-up, while of the others, only 30% and 50% respectively do it.

Compute:

1.The probability that next client will get a full tank-up of unleaded gasoline

2.The probability that the next client will get a full tank-up

3. If the next client gets a full tank-up, compute the probability that it is of normal gasoline

Exercise

Among the clients of a gas station, 35% buy normal gasoline. 40% unleaded gasoline and 25% super-enhanced gasoline. Of those using normal gasoline, 60% systematically get a full tank-up, while of the others, only 30% and 50% respectively do it.

$P(N) = 0.35$, $P(U) = 0.4$, $P(S) = 0.25$

$P(F|N) = 0.6$, $P(F|U) = 0.3$, $P(F|S) = 0.5$

Compute:

1.The probability that next client will get a full tank-up of unleaded gasoline

1.The probability that the next client will get a full tank-up

1. If the next client gets a full tank-up, compute the probability that it is of normal gasoline

Exercise

Among the clients of a gas station, 35% buy normal gasoline. 40% unleaded gasoline and 25% super-enhanced gasoline. Of those using normal gasoline, 60% systematically get a full tank-up, while of the others, only 30% and 50% respectively do it.

$P(N) = 0.35$, $P(U) = 0.4$, $P(S) = 0.25$
$P(F|N) = 0.6$, $P(F|U) = 0.3$, $P(F|S) = 0.5$

Compute:

1.The probability that next client will get a full tank-up of unleaded gasoline
$P(F|U)\ P(U) = 0.4*0.3 = 0.12$

1.The probability that the next client will get a full tank-up

1. If the next client gets a full tank-up, compute the probability that it is of normal gasoline

Exercise

Among the clients of a gas station, 35% buy normal gasoline. 40% unleaded gasoline and 25% super-enhanced gasoline. Of those using normal gasoline, 60% systematically get a full tank-up, while of the others, only 30% and 50% respectively do it.

$P(N) = 0.35$, $P(U) = 0.4$, $P(S) = 0.25$
$P(F|N) = 0.6$, $P(F|U) = 0.3$, $P(F|S) = 0.5$

Compute:

1. The probability that next client will get a full tank-up of unleaded gasoline
$P(F|U)\ P(U) = 0.4*0.3 = 0.12$

1. The probability that the next client will get a full tank-up
$P(F|U)\ P(U) + P(F|N)\ P(N) + P(F|S)\ P(S) = 0.4*0.3 + = 0.6*0.35 + 0.5*0.25 = 0.455$

1. If the next client gets a full tank-up, compute the probability that it is of normal gasoline

Exercise

Among the clients of a gas station, 35% buy normal gasoline. 40% unleaded gasoline and 25% super-enhanced gasoline. Of those using normal gasoline, 60% systematically get a full tank-up, while of the others, only 30% and 50% respectively do it.

P(N) = 0.35, P(U) = 0.4, P(S) = 0.25
P(F|N) = 0.6, P(F|U) = 0.3, P(F|S) = 0.5

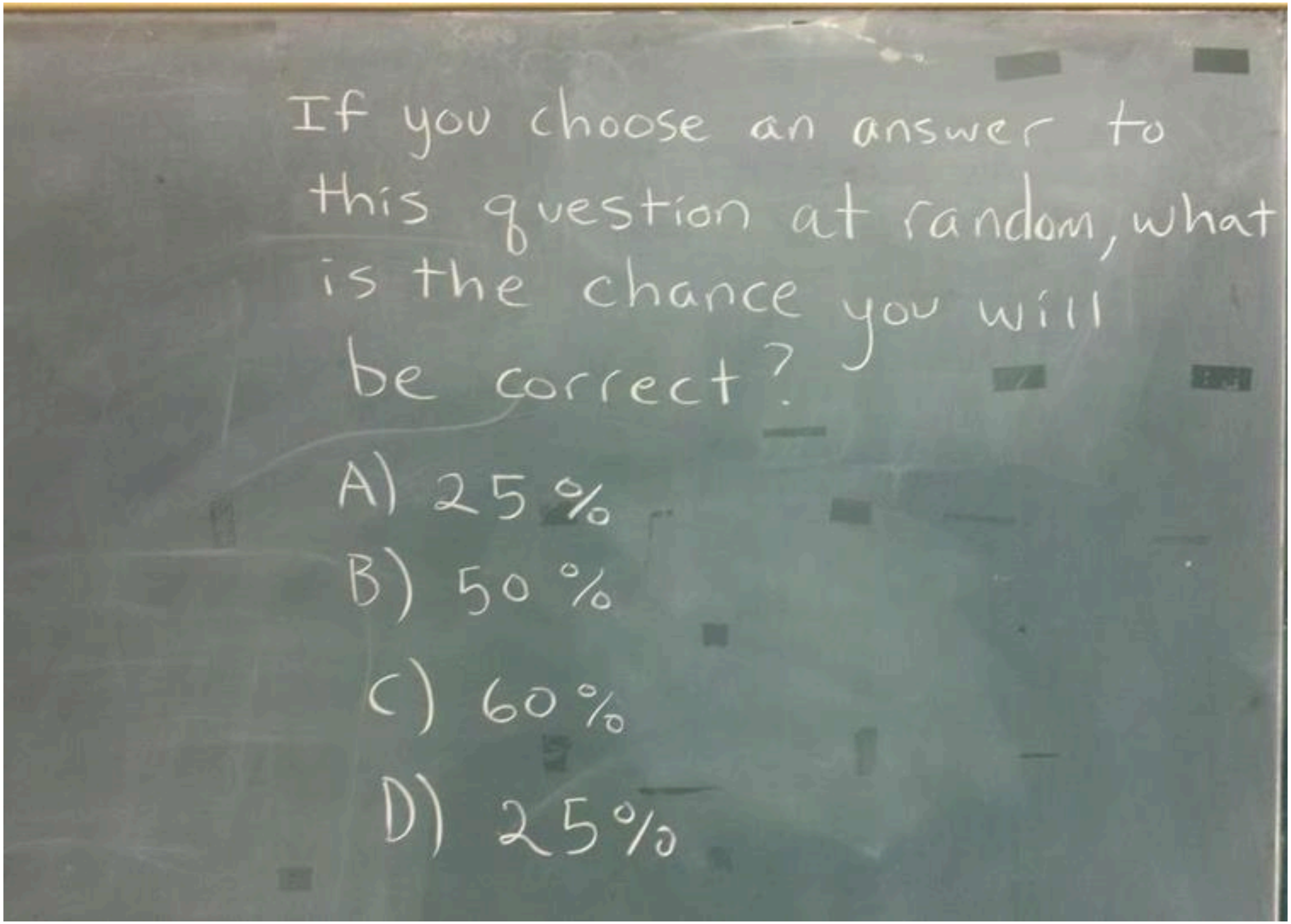
Compute:

1.The probability that next client will get a full tank-up of unleaded gasoline
P(F|U) P(U) = 0.4*0.3 = 0.12

1.The probability that the next client will get a full tank-up
P(F|U) P(U) + P(F|N) P(N) + P(F|S) P(S) = 0.4*0.3 + = 0.6*0.35 + 0.5*0.25 = 0.455

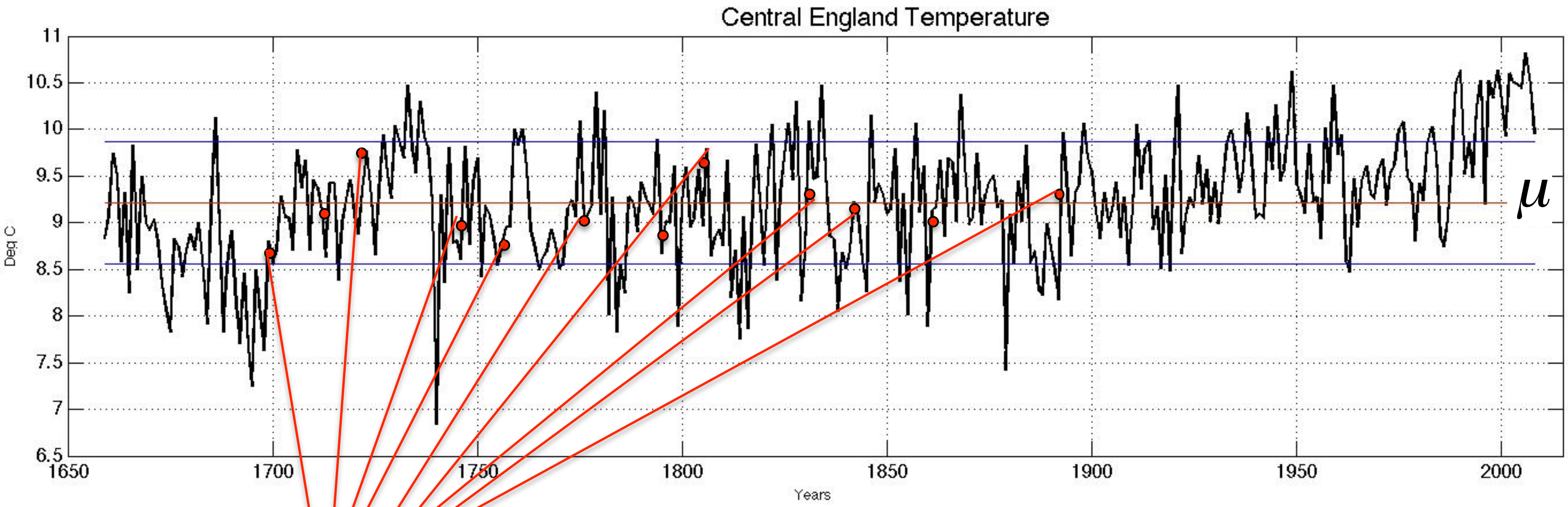
1. If the next client gets a full tank-up, compute the probability that it is of normal gasoline

$$P(F \mid N)\ P(N) = P(N \mid F)\ P(F) \Rightarrow P(N \mid F) = \frac{P(F \mid N)\ P(N)}{P(F)} = \frac{0.6 * \ 0.35}{0.455} = 0.46$$


If you choose an answer to this question at random, what is the chance you will be correct?
A) 25%
B) 50%
C) 60%
D) 25%

sampling theory

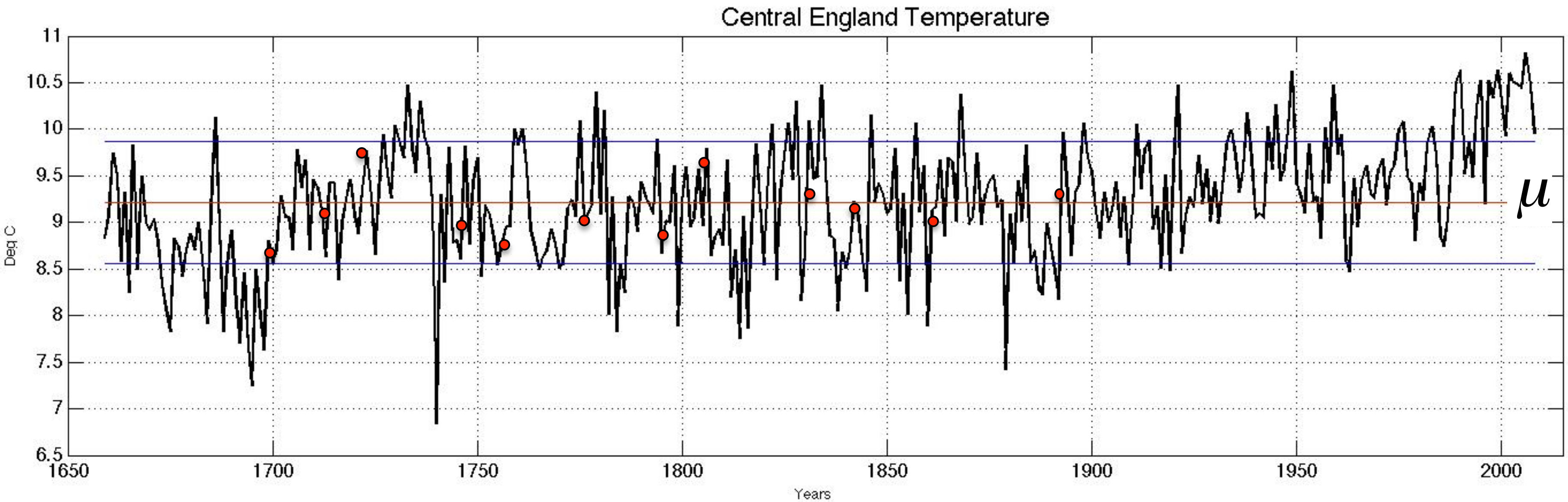
Sampling distributions.



$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

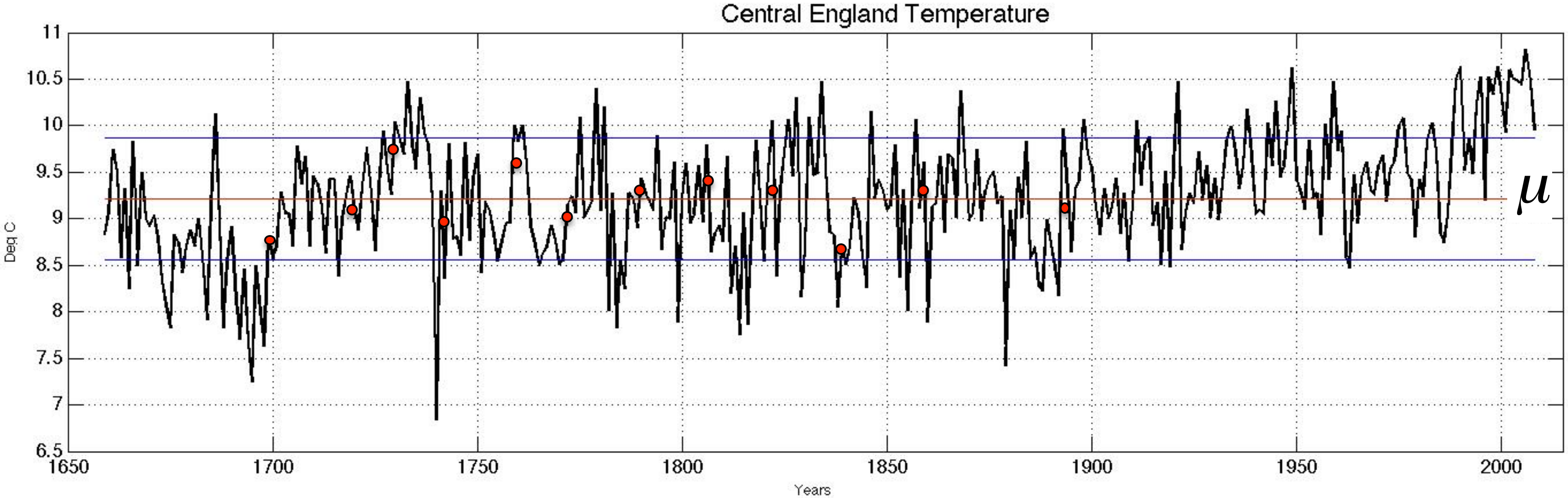
Can we say that $\bar{x} = \mu$?

Sampling distributions.



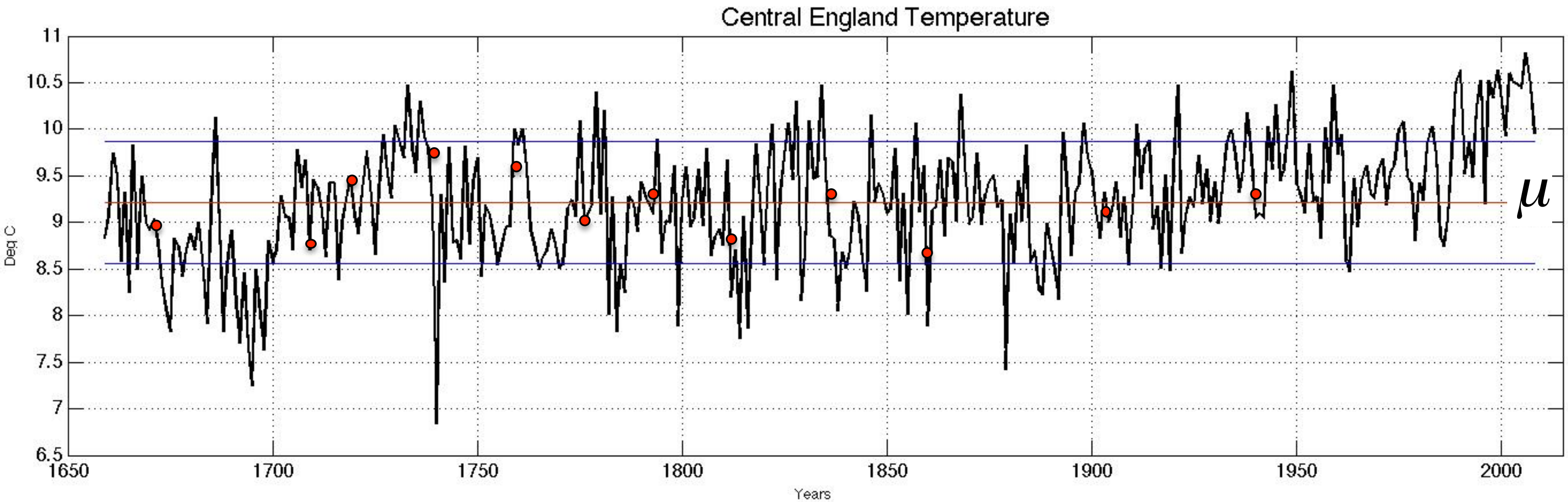
$$(\bar{x})_1$$

Sampling distributions.



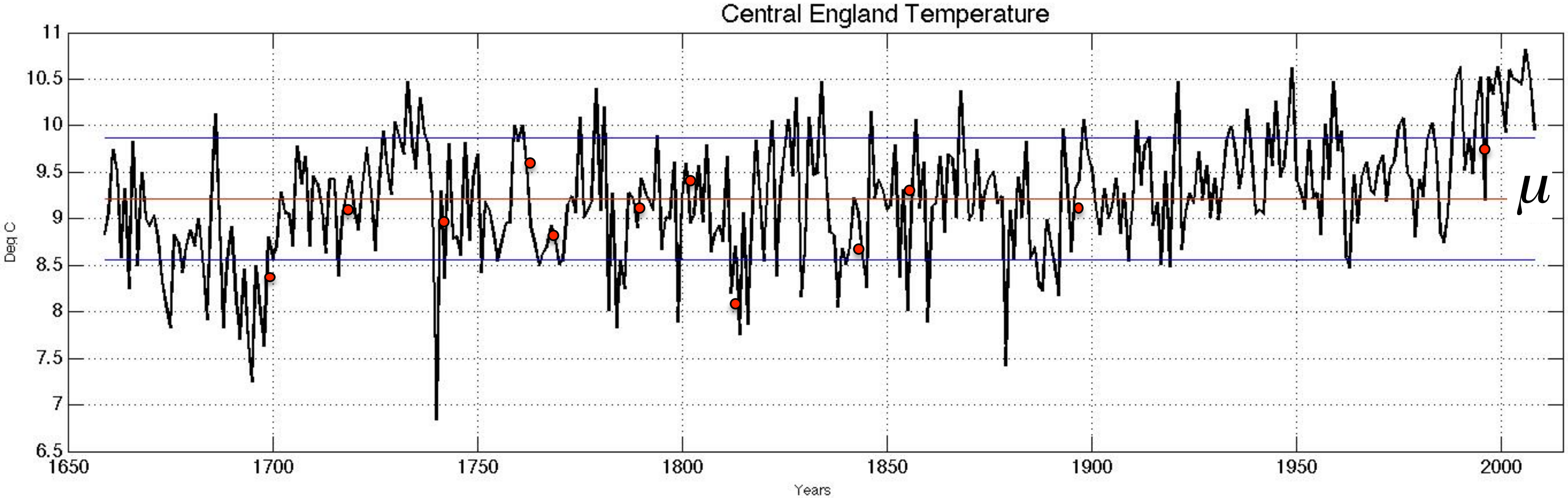
$$(\bar{x})_2$$

Sampling distributions.



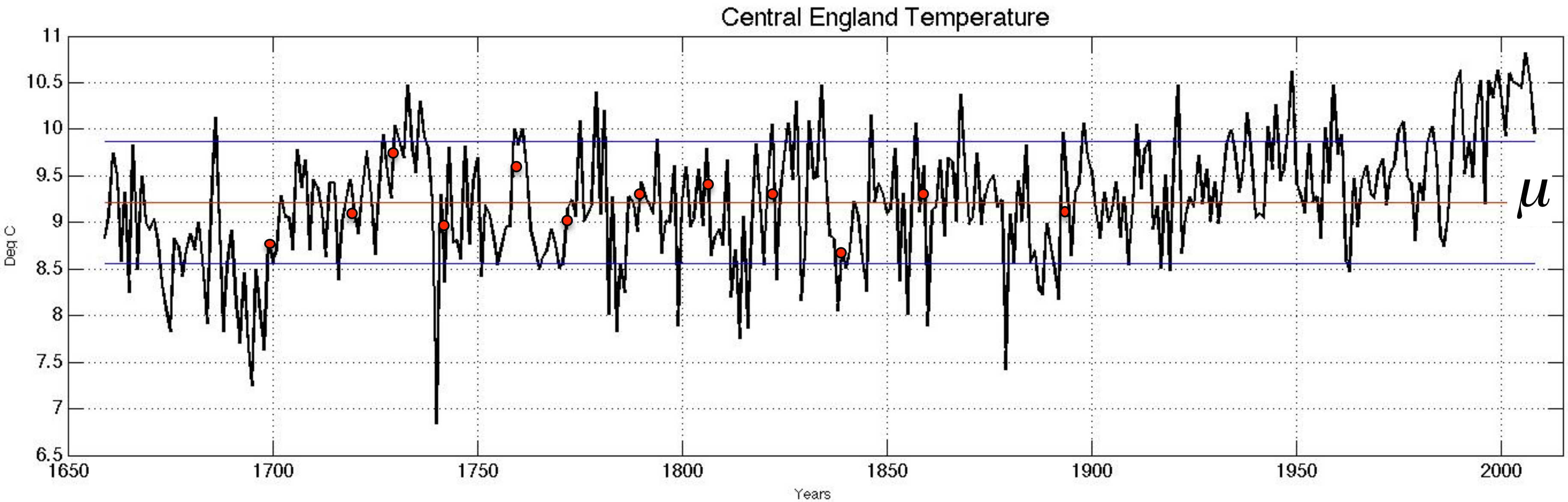
$$(\bar{x})_3$$

Sampling distributions.



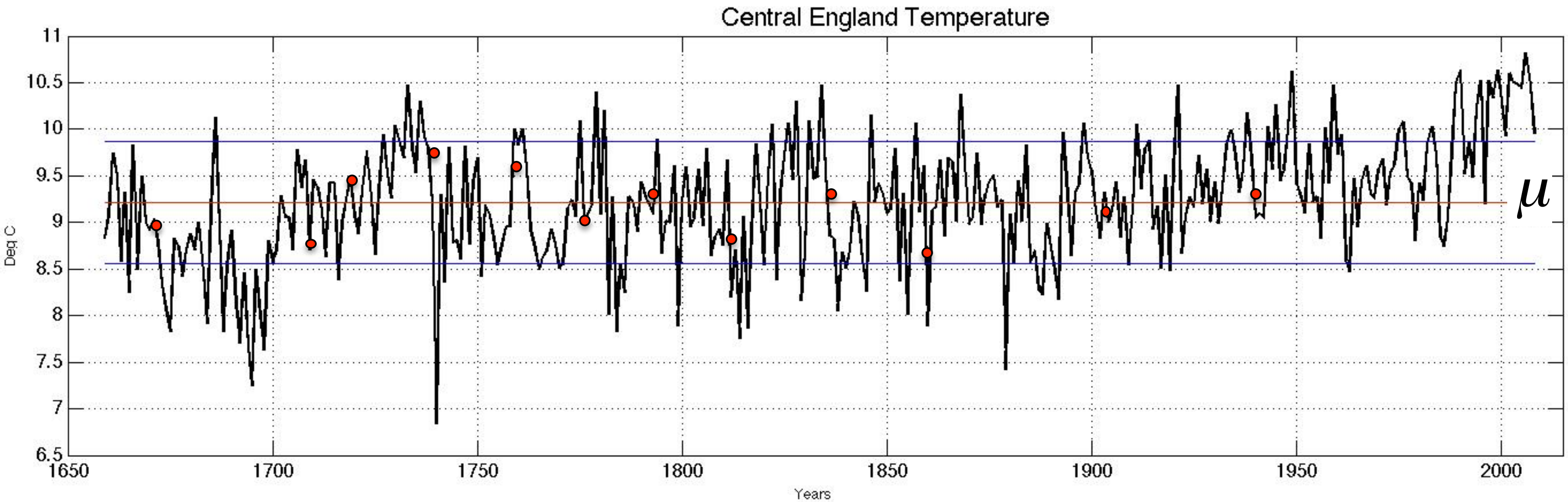
$$(\bar{x})_4$$

Sampling distributions.



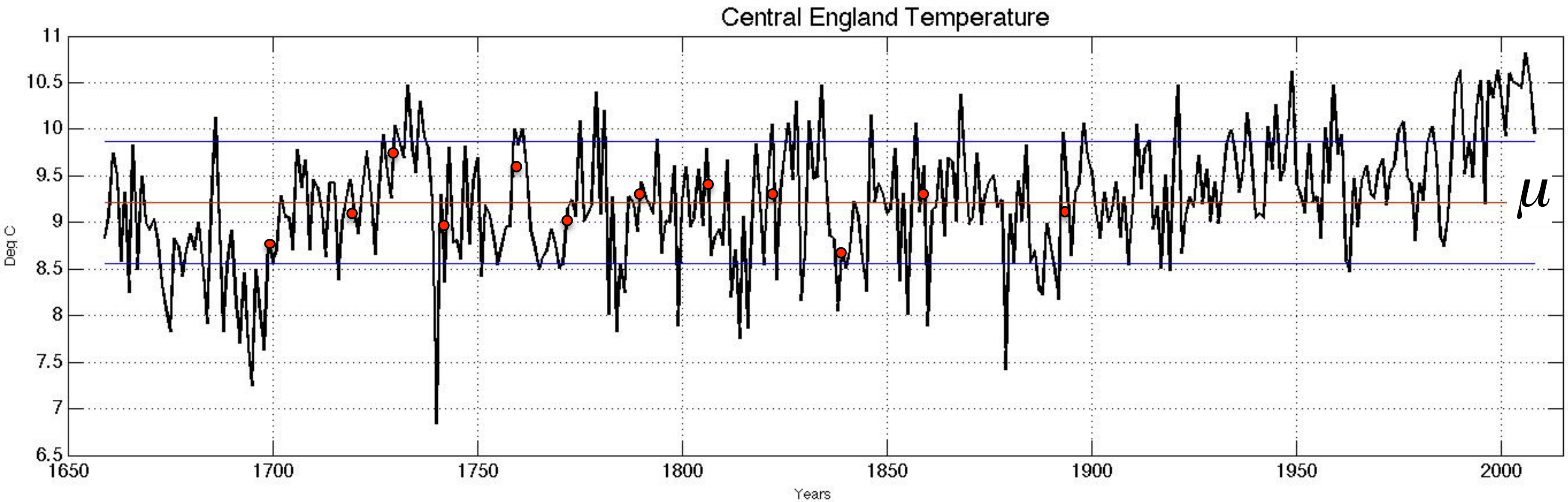
$$(\bar{x})_5$$

Sampling distributions.



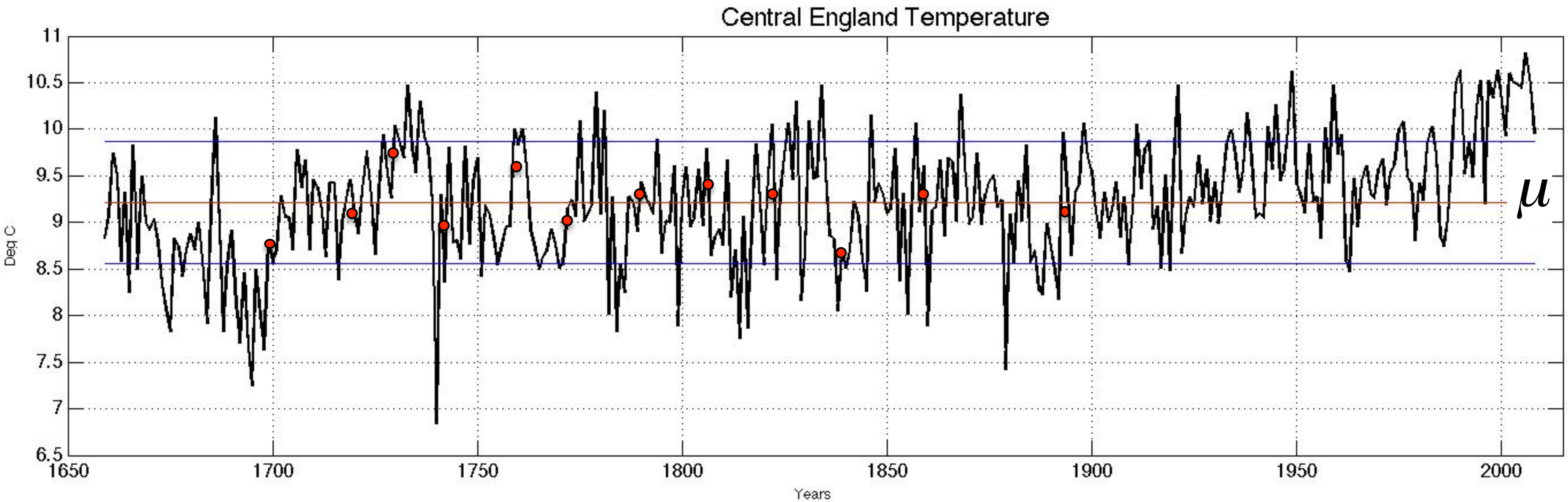
$$(\bar{x})_6$$

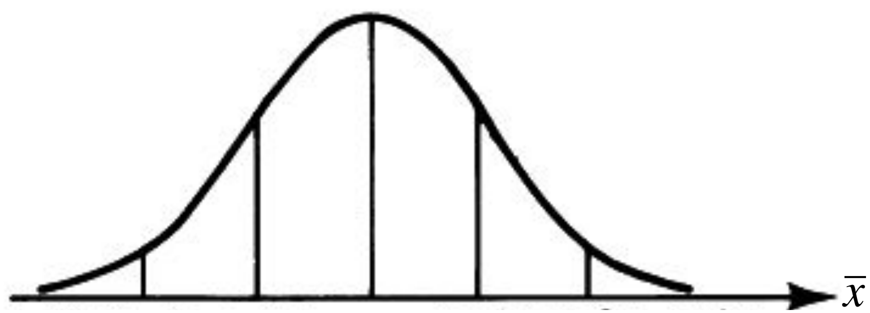
Sampling distributions.



$$(\bar{x})_7$$

Sampling distributions.





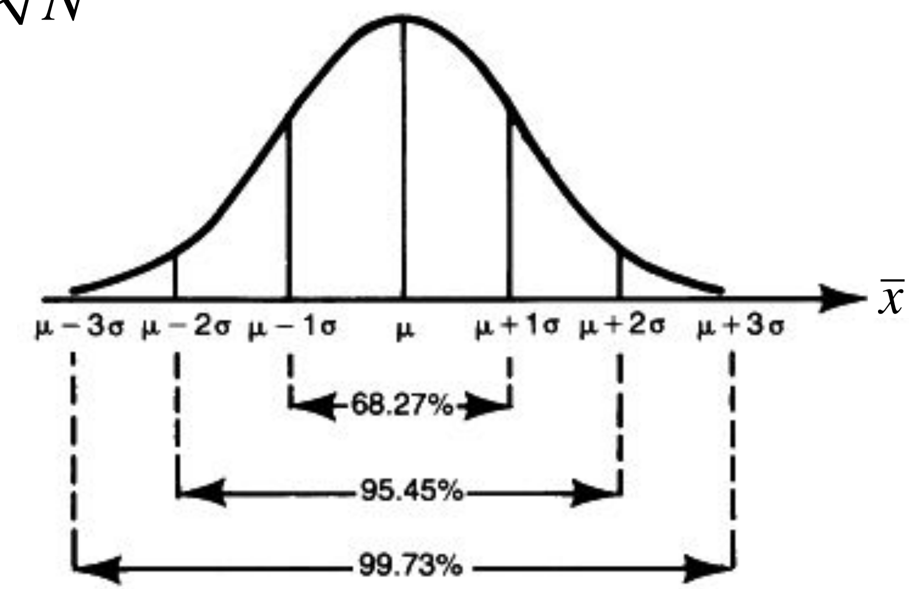
$$P(\bar{x})$$

Central limit theorem

The sampling distribution tends to a normal distribution with mean equal to μ and variance equal to σ^2 / N,

<u>independently of the distribution of the variable</u>

Consequently, whenever you have a sample of length N, it will be affected by a standard error of $\frac{\sigma}{\sqrt{N}}$



Proof of the central limit theorem.

NB, this is a “crude” proof. For anything mode formal ask your mathematicians friends (or look at a book).

i) mean $E(\bar{x}) = \mu$

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

$$= \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \frac{1}{N}\sum_{i=1}^{N} \mu =$$

$$= \mu$$

Proof of the central limit theorem.

NB, this is a “crude” proof. For anything mode formal ask your mathematicians friends (or look at a book).

i) mean $E(\bar{x}) = \mu$

$$E(\bar{x}) = E\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \int_{-\infty}^{\infty} \frac{1}{N}\sum_{i=1}^{N} x_i f(x)dx = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty} x_i f(x)dx$$

$$= \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \frac{1}{N}\sum_{i=1}^{N} \mu =$$

$$= \mu$$

ii) variance $E((\bar{x}-\mu)^2)=\frac{\sigma^2}{N}$

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

$$=\frac{1}{N^2}E\left(\left[\sum_{i=1}^{N}(x_i-\mu)\right]^2\right)=$$

$$=\frac{1}{N^2}\sum_{i=1}^{N}E\left([(x_i-\mu)]^2\right)+\frac{1}{N^2}\sum_{i\neq j}^{N}E\left(2(x_i-\mu)(x_j-\mu)\right)=$$

$$=\frac{1}{N^2}N\,\mathrm{var}(x)\;=\;\frac{\sigma^2}{N}$$

Now we can prove that the correct estimator of the variance of a population from a sample is:

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

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 =$$

$$= \sum_{i=1}^{N}(x_i - \mu - (\bar{x} - \mu))^2 =$$

$$= \sum_{i=1}^{N}(x_i - \mu)^2 + N(\bar{x} - \mu)^2 - 2(\bar{x} - \mu)\sum_{i=1}^{N}(x_i - \mu) =$$

$$= \sum_{i=1}^{N}(x_i - \mu)^2 + N(\bar{x} - \mu)^2 - 2N(\bar{x} - \mu)^2$$

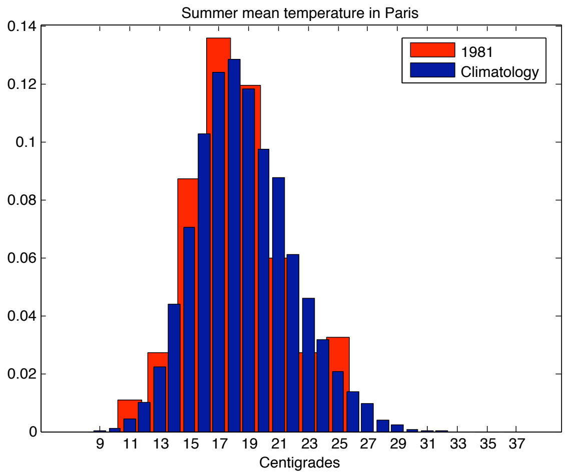
Taking the expectation value:

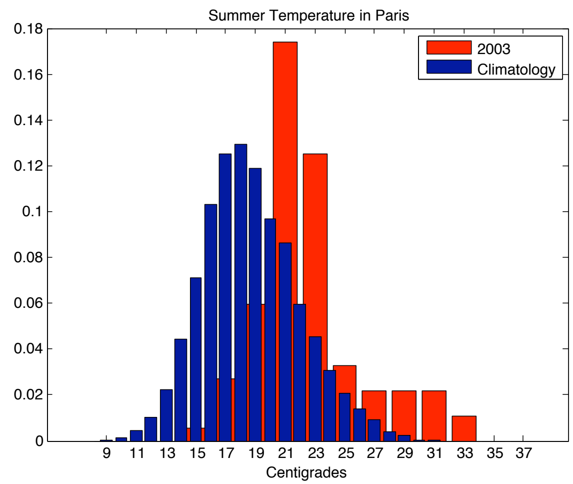
$$E(\sum_{i=1}^{N}(x_i - \mu)^2) - E(N(\bar{x} - \mu)^2) = N\sigma^2 - N\frac{\sigma^2}{N} = (N-1)\sigma^2$$

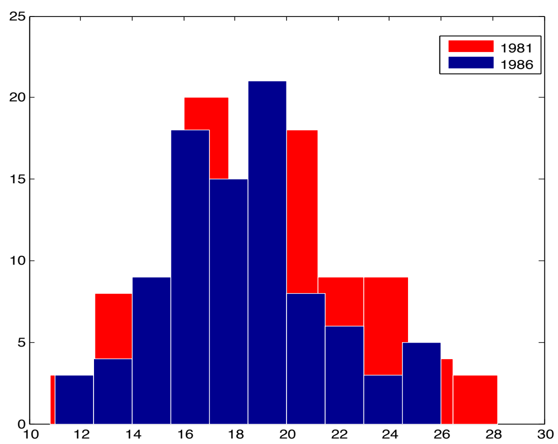
Comparing two set of data:

1) Are two samples issued from the same population?

Ex. Is the mean temperature in Paris for summer 2003 statistically different from climatology?







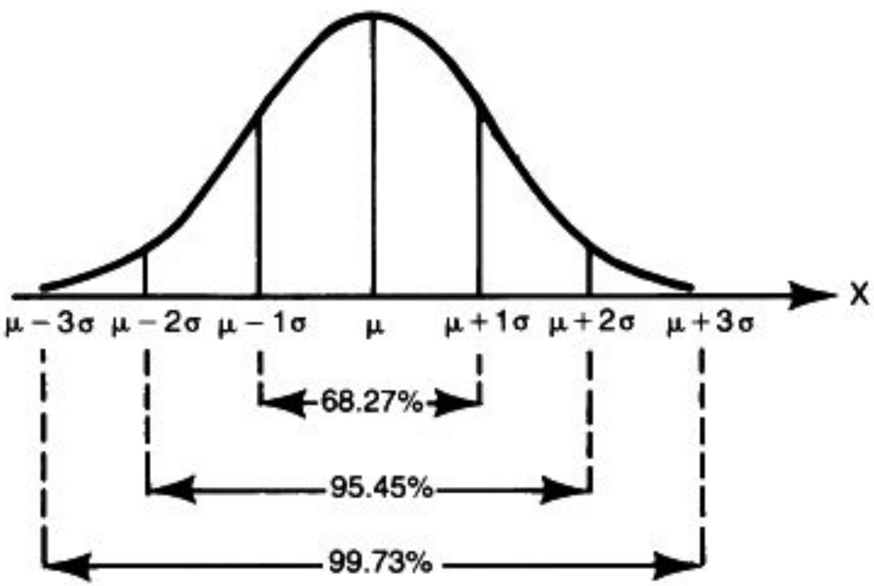
People at LMD are taller on average than the rest of the population of Paris - Or are they?

Ronald Fisher

	Name	Height (cm)
1	Fabio	180
2	Francois	181
3	Jean-Philippe	185
4	Hugo	183
5	Alexandre	175
6	Alexandra	162
7	Alessandro	172
8	Ayah	160
9	Guillaume	183
10	Hector	171
11	Marie-Christine	165
12	Michael	177
13	Pauline	175
14	Riwal	177

An example of hypothesis testing:
1) Set up a Null Hypothesis
2) Chose a test statistic
3) Chose a level of significance
4) Compute the level of probability of the sample, from the test statistic
5) If that is lower than the level of significance, reject the null hypothesis.

People at LMD are taller on average than the rest of the population of Paris - Or are they?



Paris: $\mu_0 = 170, \;\; \sigma_0 = 14$

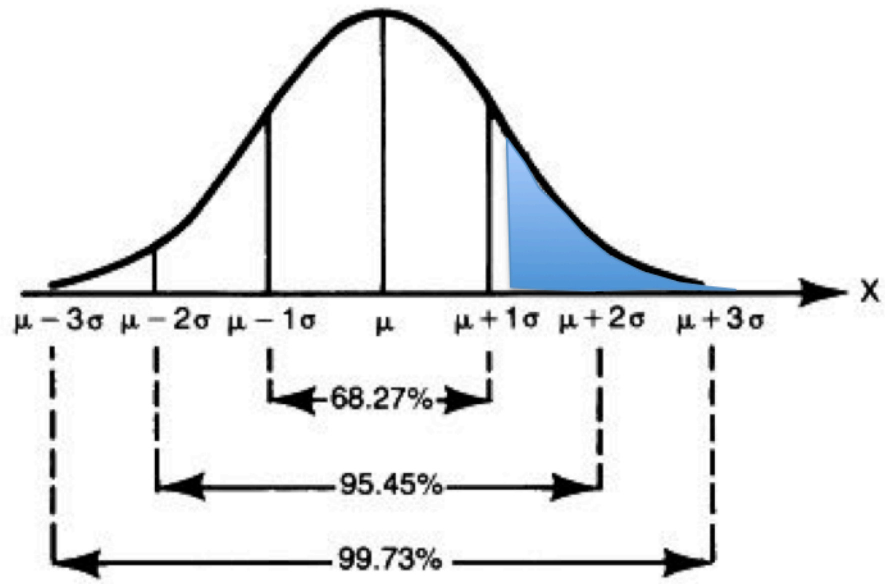
LMD $\mu = 174.7, \;\; \sigma = \sigma_0 / \sqrt{N}$

	Name	Height (cm)
1	Fabio	180
2	Francois	181
3	Jean-Philippe	185
4	Hugo	183
5	Alexandre	175
6	Alexandra	162
7	Alessandro	172
8	Ayah	160
9	Guillaume	183
10	Hector	171
11	Marie-Christine	165
12	Michael	177
13	Pauline	175
14	Riwal	177

An example of hypothesis testing:

1) Set up a Null Hypothesis
2) Chose a test statistic
3) Chose a level of significance
4) Compute the level of probability of the sample, from the test statistic
5) If that is lower than the level of significance, reject the null hypothesis.

hypothesis testing:



1)Set up a Null Hypothesis
The sample is issued from the same population:
People at LMD are not taller than the rest.

2)Chose a test statistic
The sample mean distribution

4)Chose a level of significance
95%, i.e. 2 standard errors, $p \leq 5\%$

4) Compute the level of probability of the sample, from the test statistic

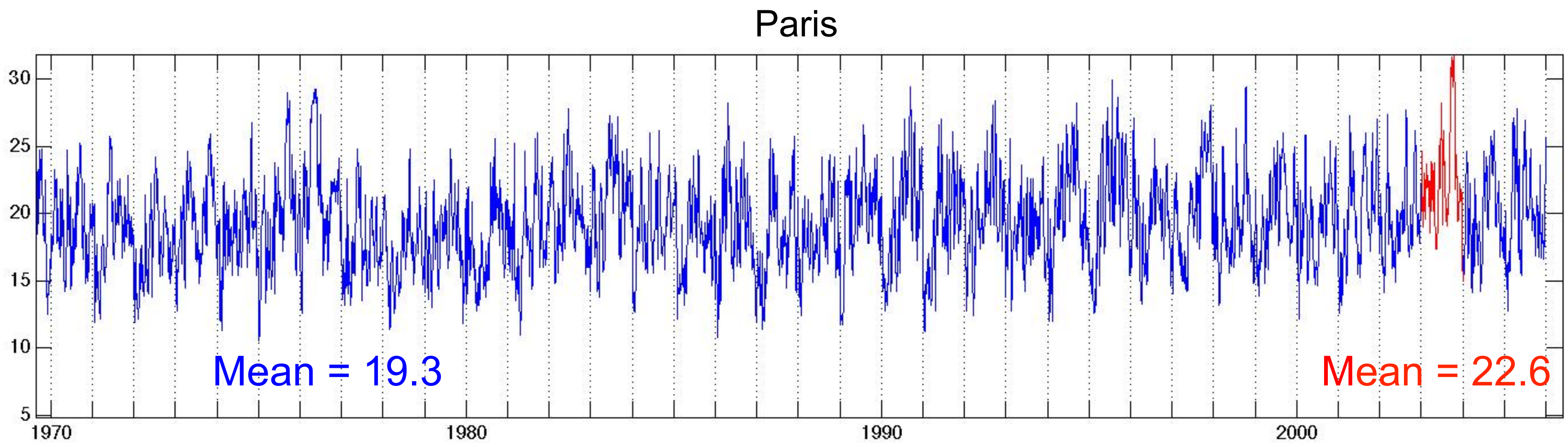
$$\frac{(\mu - \mu_0)}{\sigma / \sqrt{N}} = \frac{(174.7 - 170)}{14 / \sqrt{14}} = 1.26$$

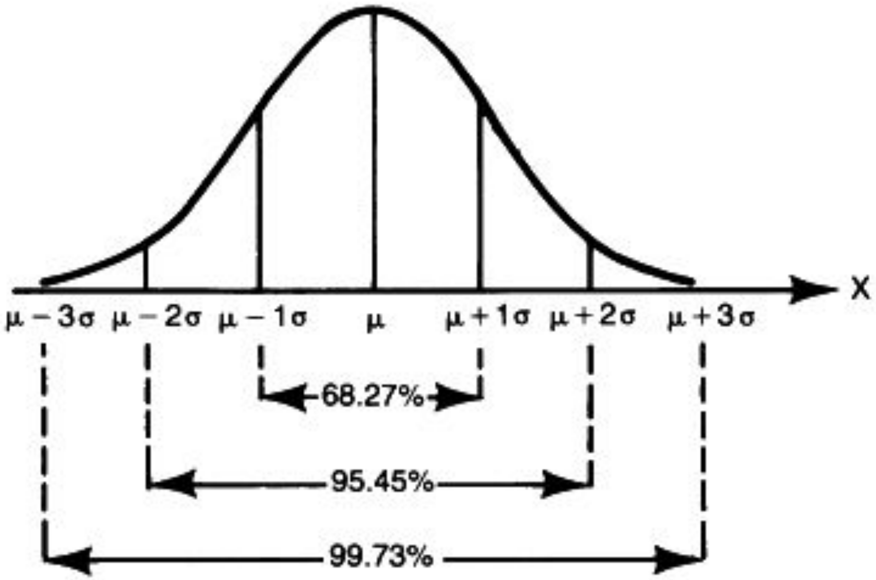
(difference of means in std error units)

$$p = P(\left\|h\right\| \geq \mu) \approx 14 - 15\%$$

5) If that is lower than the level of significance, reject the null hypothesis.
CANNOT REJECT

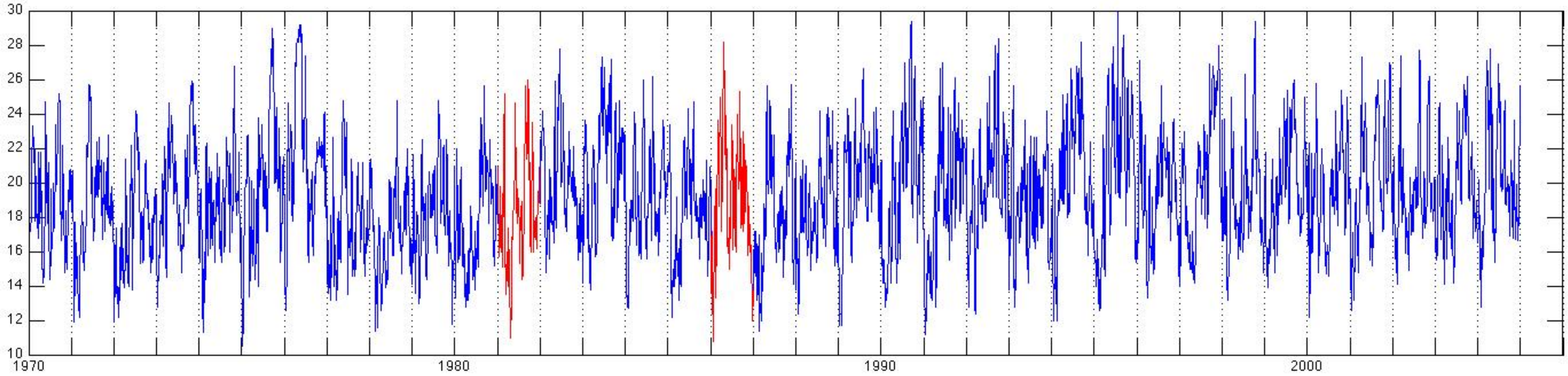
Back to the initial question



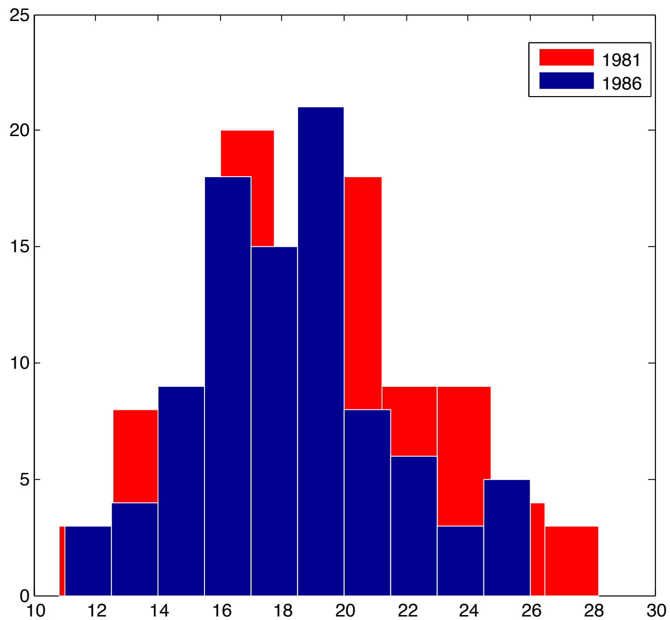


$$\frac{(\mu - \mu_0)}{\sigma / \sqrt{N}} = \frac{(22.6 - 19.3)}{3.4 / \sqrt{92}} = 9.3$$

Comparing two samples



Are the mean of these two samples different? I.e. are they issued from two distinct populations?



We are comparing two series that are both the sampling of an unknown variable.

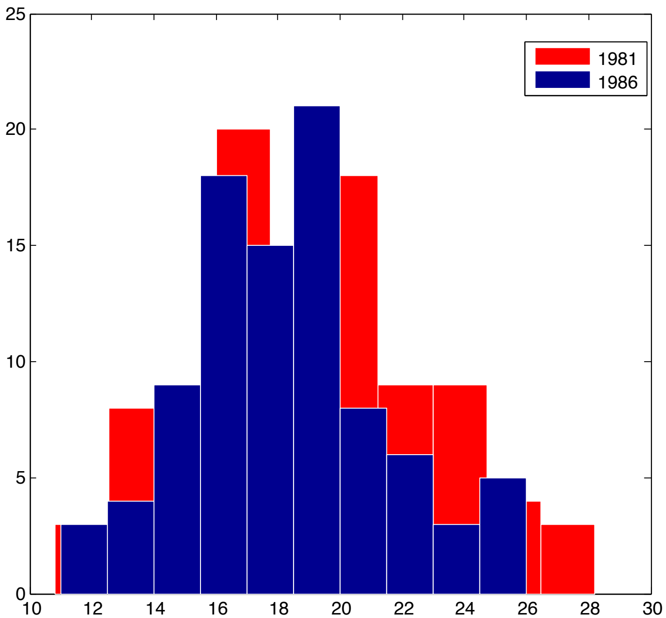
We need another test statistics

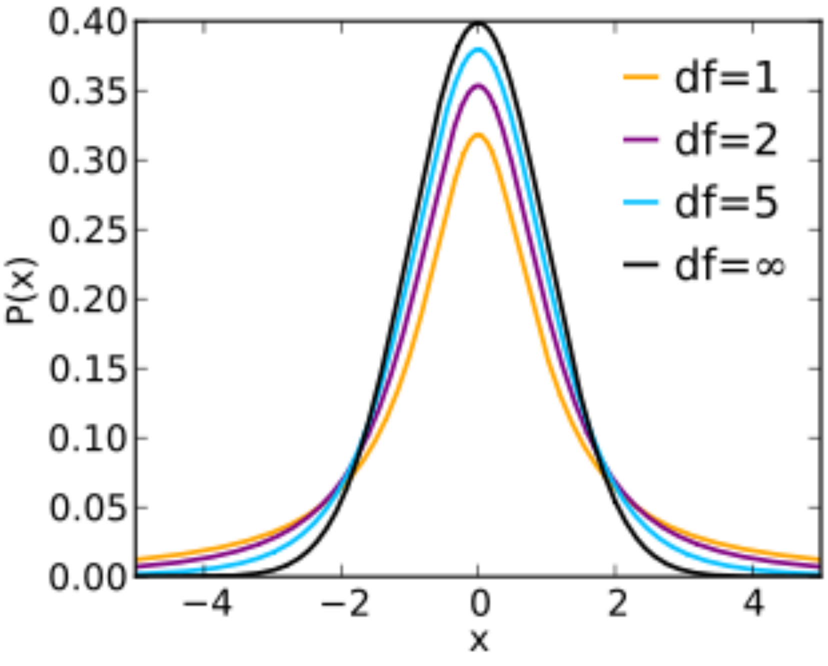
The distribution of:

$$t = \frac{\mu_a - \mu_b}{s / \sqrt{N}}, \quad \text{where} \quad s = \sqrt{\frac{(N_a - 1)\sigma_a^2 + (N_b - 1)\sigma_b^2}{N_a + N_b - 2}}$$

$$\frac{1}{N} = \left(\frac{1}{N_a} + \frac{1}{N_b} \right)$$

Is known, it is called the Student-t distribution. Its value for a given t can be found in tables, or in the most usual statistical software packages.





$$\mu_a - \mu_b = 0.72,$$

$$s = \sqrt{\frac{(N_a - 1)\sigma_a^2 + (N_b - 1)\sigma_b^2}{N_a + N_b - 2}} = 3.47$$

$$t = \frac{\mu_a - \mu_b}{s / \sqrt{N}} = 1.4$$

$$p(t) = 0.1588$$

$$\mathrm{df} = N_a + N_b - 2$$

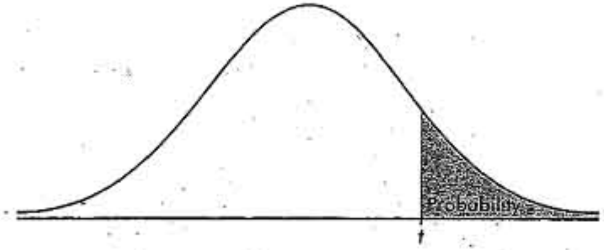


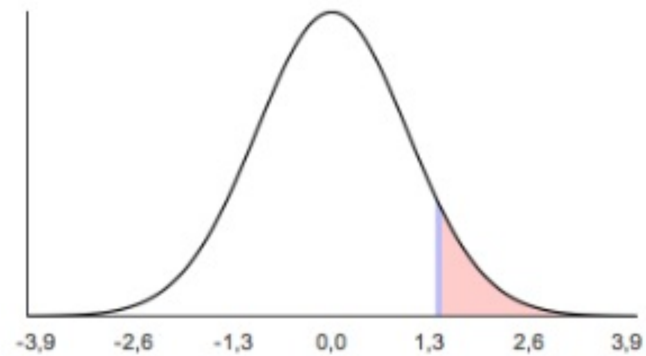
TABLE B: t-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

t-Distribution: $X \sim t_{(df)}$

df = 180

x = 1.4   P(X > x) = 0.08162



Help

©2012 Matt Bognar
Department of Statistics and Actuarial Science
University of Iowa

Degrees of freedom
In case of autocorrelation use:

$$n'_a = \frac{n_a}{1 + \sum_{i=1}^{n_a} \left(1 - \frac{i}{n_a}\right) \rho_a(i)}$$

Suppose we divide the class in two groups: one studies in a room with loud techno music all day long, the other group studies in a silent room. After the exams, these are the results of the two groups:

Group A (with music)	Group B (no music)	
18	15	
17	15	
13	10	
10	11	
14		
Ma = 14.8	Mb = 12.75	Ma-Mb = 2.05
Var(A) = 9.7	Var(B) = 6.9167	

Can we say that techno music is good for studying statistics?

$$\mu_a - \mu_b = 14.8 - 12.75 = 2.05$$

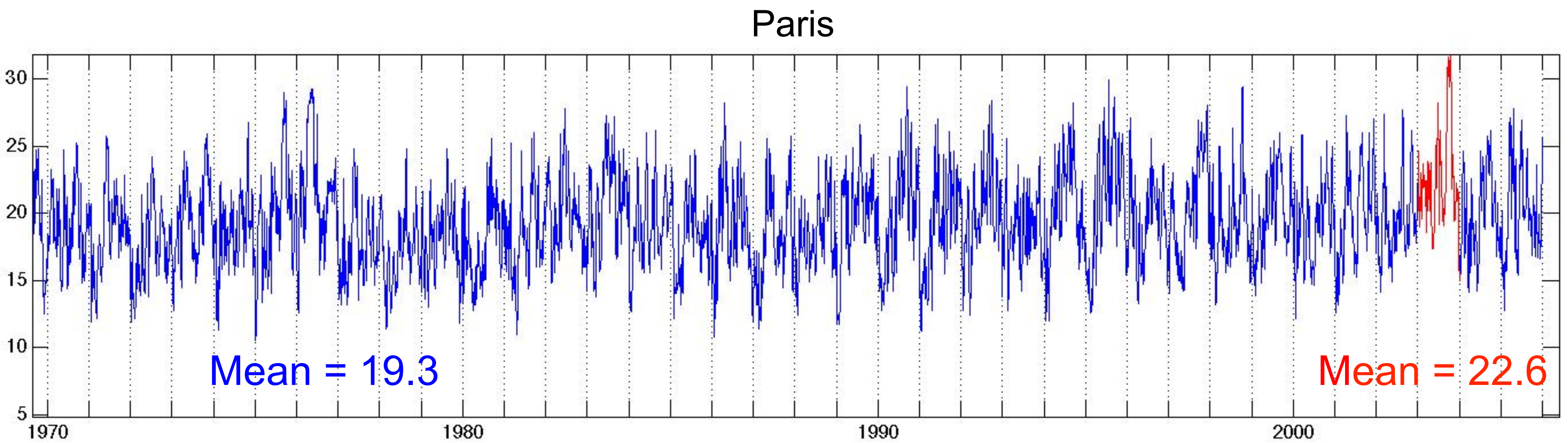
$$s = \sqrt{\frac{(N_a - 1)\sigma_a^2 + (N_b - 1)\sigma_b^2}{N_a + N_b - 2}} = \sqrt{\frac{4\sigma_a^2 + 3\sigma_b^2}{7}} = \sqrt{\frac{49.7 + 36.9167}{7}} = 2.9167$$

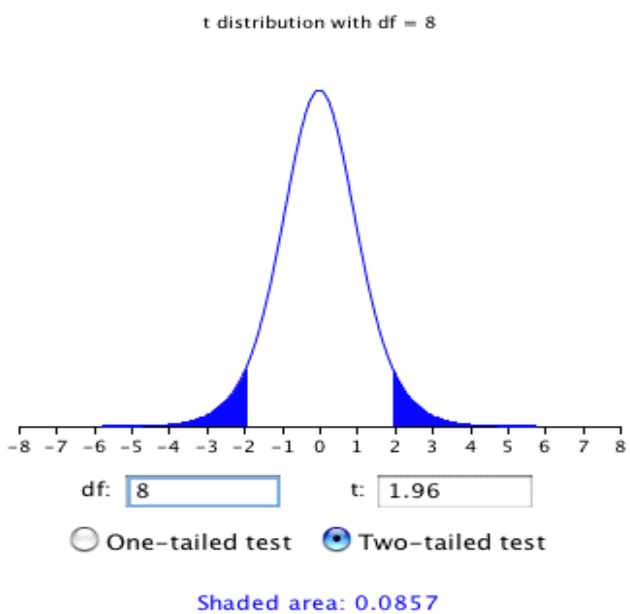
$$N = \frac{1}{\frac{1}{4} + \frac{1}{5}} = 2.222$$

$$t = \frac{\mu_a - \mu_b}{s / \sqrt{N}} = \frac{2.05}{2.9167 / \sqrt{2.222}} = 1.0477$$

$$p(t) = 0.1648$$

Applying it to the long timeseries…





$$\mu_a - \mu_b = 3.3,$$

$$s = \sqrt{\frac{(N_a - 1)\sigma_a^2 + (N_b - 1)\sigma_b^2}{N_a + N_b - 2}} = 3.25$$

$$t = \frac{\mu_a - \mu_b}{s / \sqrt{N}} = 9.4,$$

Each problem its test distribution.

1 One-Sample tests:

1.1 mean with known variance: Z-test $Z = \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0;\ 1)$

1.2 mean with unknown variance: T-test $T = \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \sim t_{n-1}$

2 Two-Samples tests:

2.1 means with unknown variances: T-test

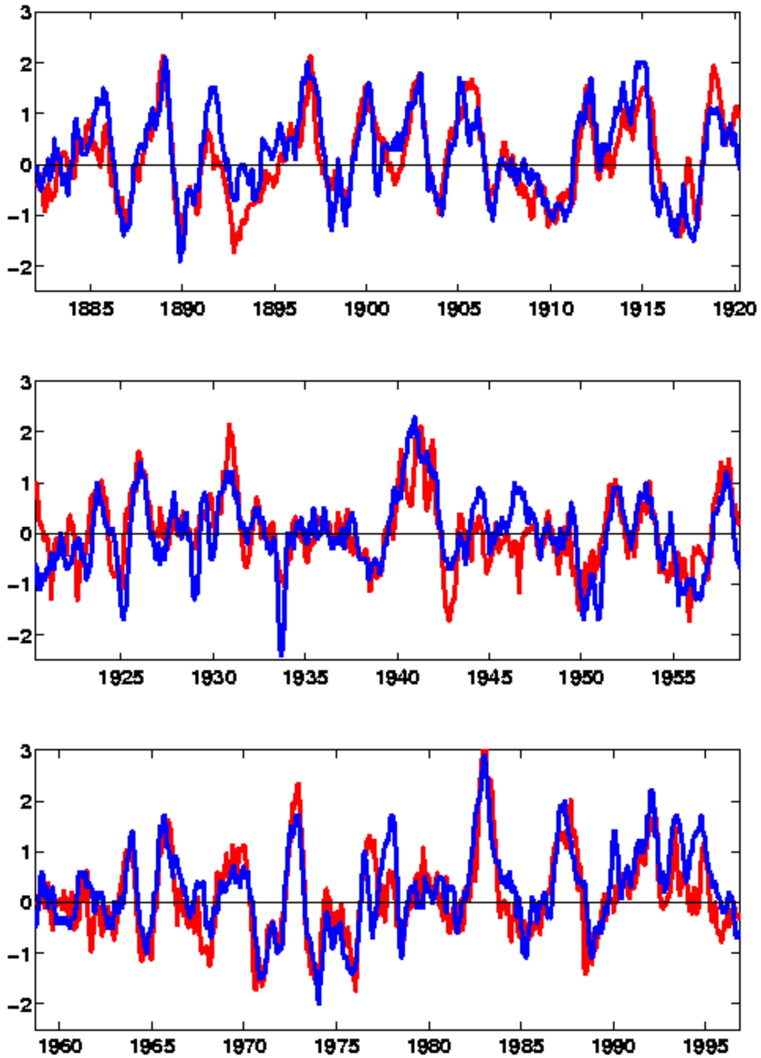
$$T = \frac{\mu_a - \mu_b}{s/\sqrt{n}} \sim t_{n_a + n_b - 2},\ s = \sqrt{\frac{(n_a - 1)\sigma_a^2 + (n_b - 1)\sigma_b^2}{n_a + n_b - 2}},\ \frac{1}{n} = \frac{1}{n_a} + \frac{1}{n_a}$$

2.2 Variances: F-test, χ^2 test

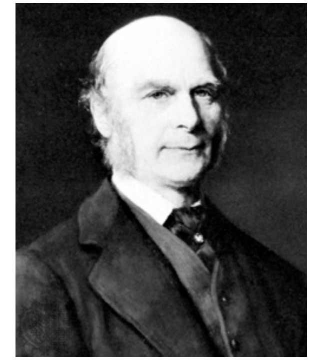
Look up your favourite statistics handbook and find your special case!!

Are two timeseries correlated?

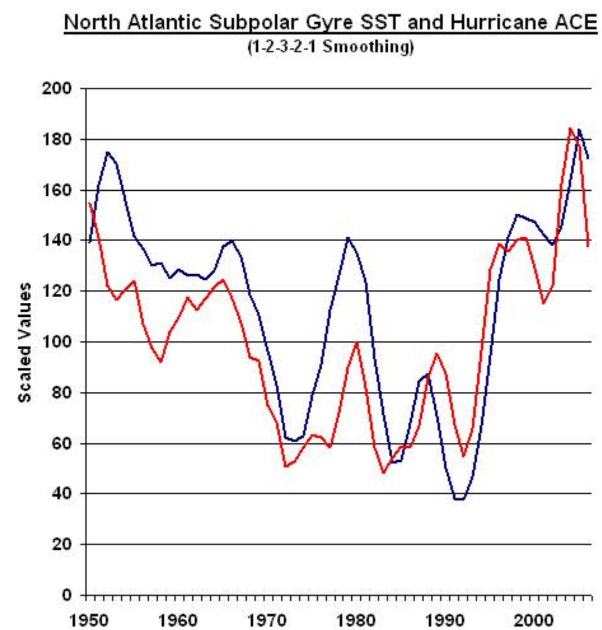
Darwin SLP NINO3 SST 1882 – 1996

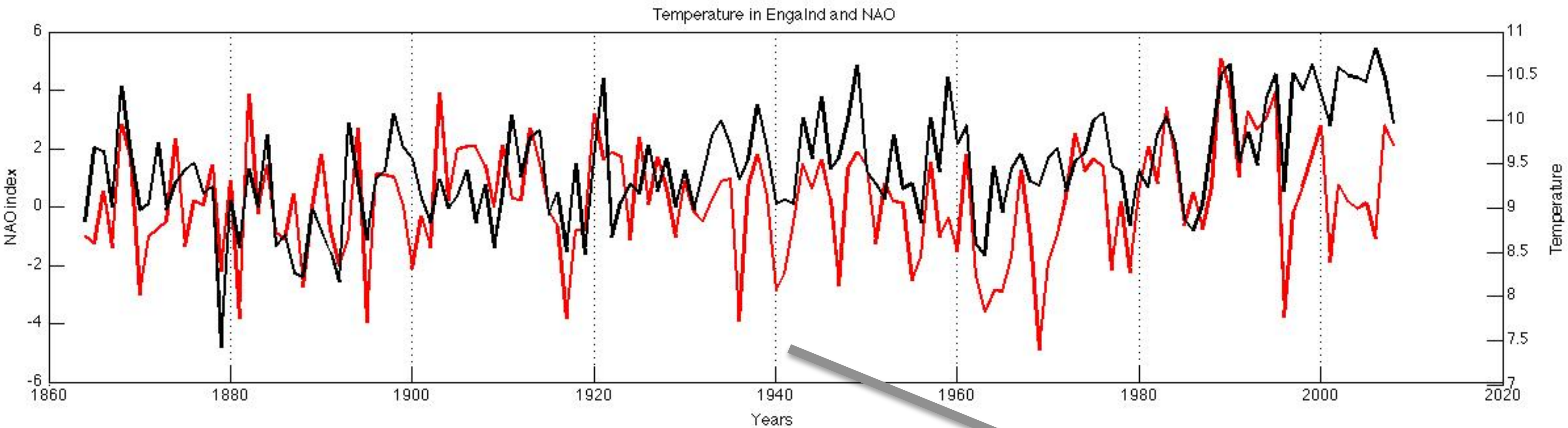


Francis
Galton

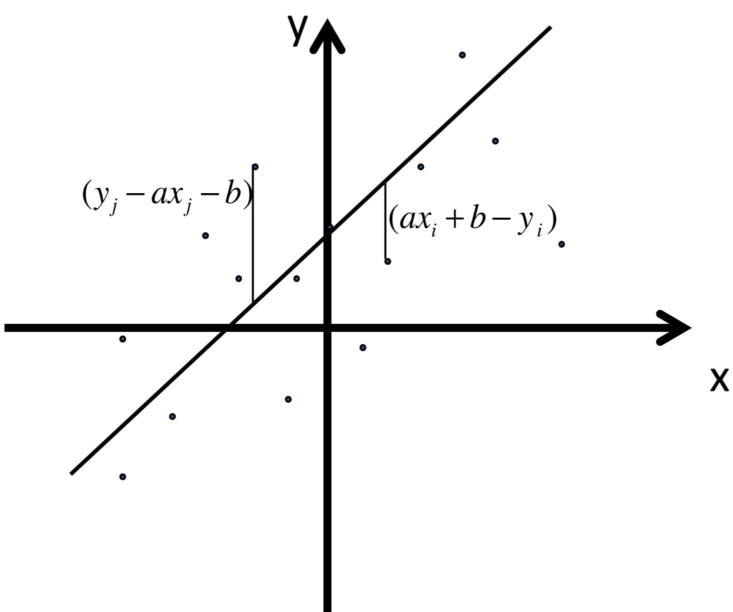


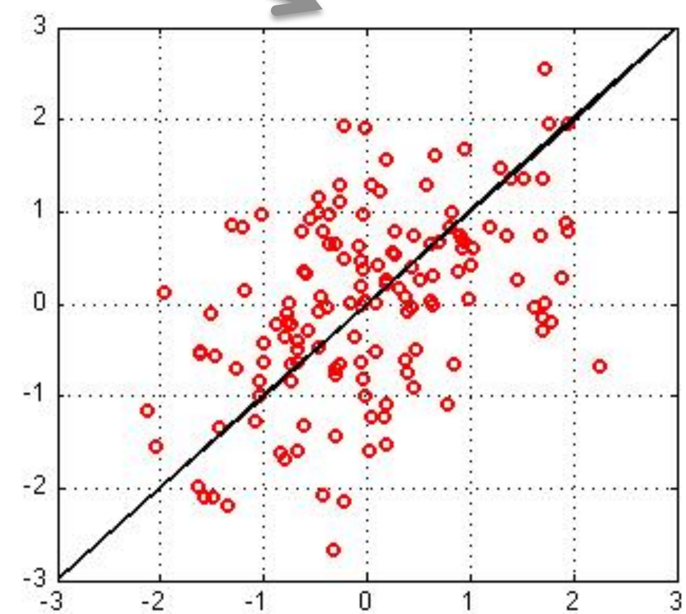
In geophysics, sometimes one wonders whether two timeseries are linked.

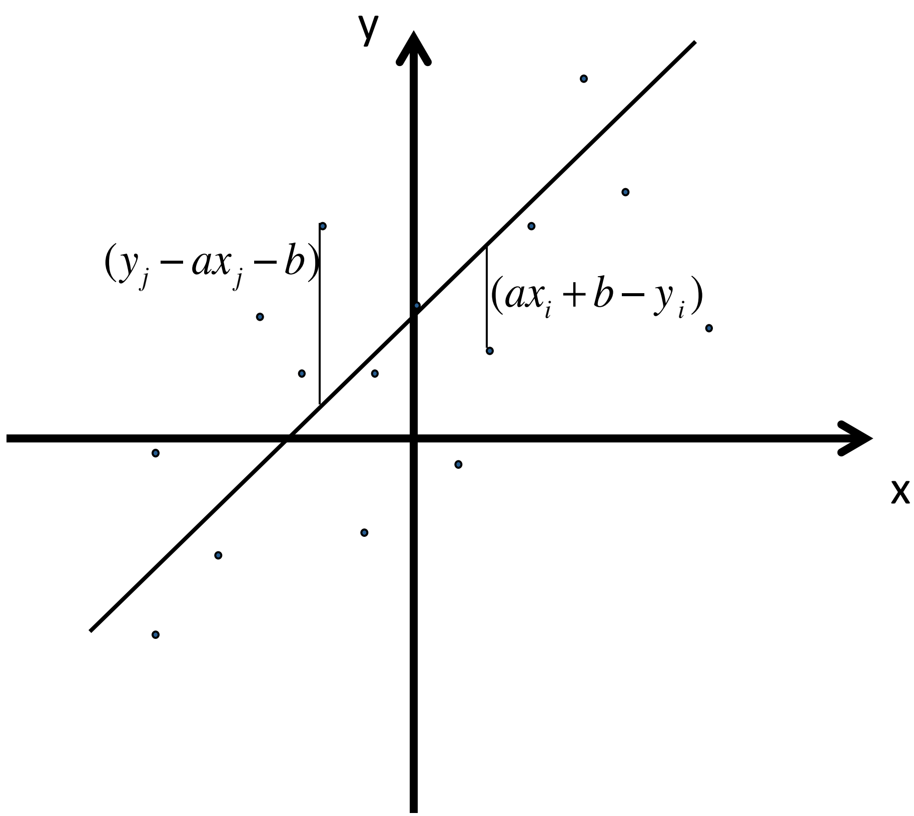




One way to put it is to estimate if one series can be obtained by linear transformation from the other: $y_i = ax_i + b$







We want to minimize:

$$\sum_{i=1}^{N} (y_i - ax_i - b)^2$$

We take the derivative with respect to *a* and *b* and we obtain the two conditions:

$$a)\quad \sum x_i (y_i - ax_i - b) = 0$$
$$b)\quad \sum (y_i - ax_i - b) = 0$$

Condition b) gives:

$$b)\quad \sum_{i=1}^{N} y_i - a \sum_{i=1}^{N} x_i - Nb = 0 \;\Rightarrow\; \boxed{b = \overline{y} - a\overline{x}}$$

Substituting b) into a) gives:

$$a)\quad \sum_{i=1}^{N} \left(y_i x_i - a x_i^2 - \overline{y} x_i - a \overline{x} x_i \right) = \sum_{i=1}^{N} \left(y_i' x_i' - a x_i'^2 \right)$$

Where we have introduced the definitions :

$$x_i = \overline{x} + x',\; y_i = \overline{y} + y'.$$

Hence

$$a = \frac{\sum_{i=1}^{N} x_i' y_i'}{\sum_{i=1}^{N} x_i'^2} = \frac{\overline{x'y'}}{\overline{x'^2}} \qquad \text{Regression}$$

$$b = \overline{y} - a\overline{x}$$

The regression is not perfect: $\hat{y}_i = ax_i + b \neq y_i$

How good is the regression? One way to answer is to compute how much of the variance of y is explained by x.

Introducing the error $y_i^* = y_i - \hat{y}_i$ We can write $y_i = ax_i + b + y_i^*$

And the variance of y becomes:

$$\overline{y'^2} = \overline{a^2 x'^2} + \overline{y^{*2}} \quad \Rightarrow \quad \boxed{\frac{\overline{a^2 x'^2} + \overline{y^{*2}}}{\overline{y'^2}} = 1}$$

explained variance + unexplained variance = 1

Substituting the value of *a* found above we find:

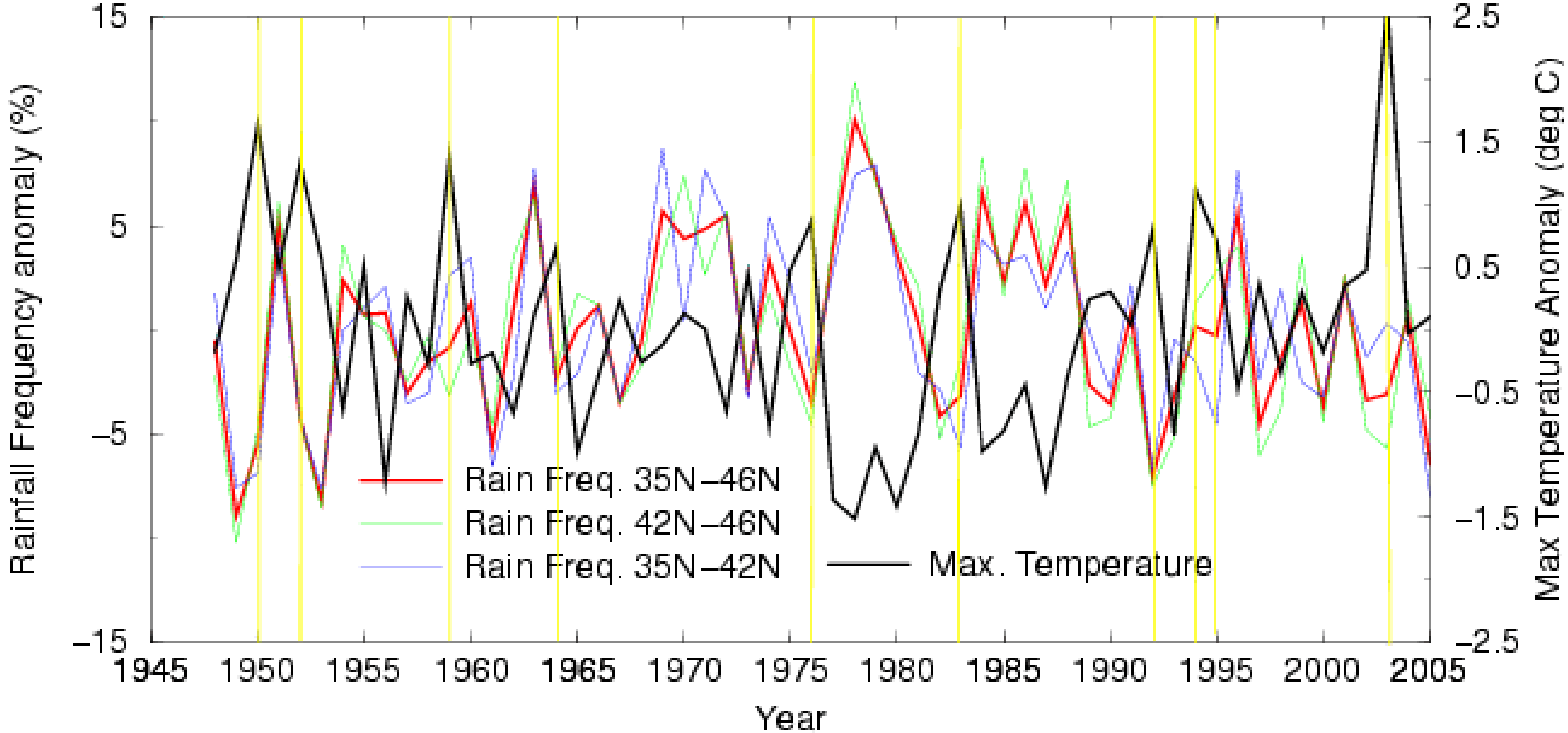
$$\frac{\overline{a^2 x_i'^2}}{\overline{y_i'^2}} = \frac{\left(\overline{x'y'}\right)^2}{\overline{x'^2}\,\overline{y'^2}} = r^2, \qquad r = \frac{\overline{x'y'}}{\sigma_x \sigma_y}$$

Is the (Pearson's) correlation coefficient

$$\boxed{r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}; \quad 1 - r^2 = \frac{\text{Unexplained Variance}}{\text{Total Variance}}}$$

Karl Pearson

Another example, summer heat in Europe and Mediterranean precipitations



Detrended summertime (JJA) daily maximum temperature anomalies, averaged over European stations, as a function of year (in black), together with the detrended anomaly of rainfall frequency averaged in the 35°N-46°N latitude band during preceding winter and early spring (January to May), in red. Temperature anomalies are in °C while precipitation frequencies anomalies are in % of days. The correlation between the two sets of values is -0.55. In order to assess the sensitivity of this latitude band for precipitation frequency, it is split into 2 latitude bands for which the time series are also calculated: 42°N-46°N (green) and 35°N-42°N. Yellow bars indicate the selected 10 hottest summers.

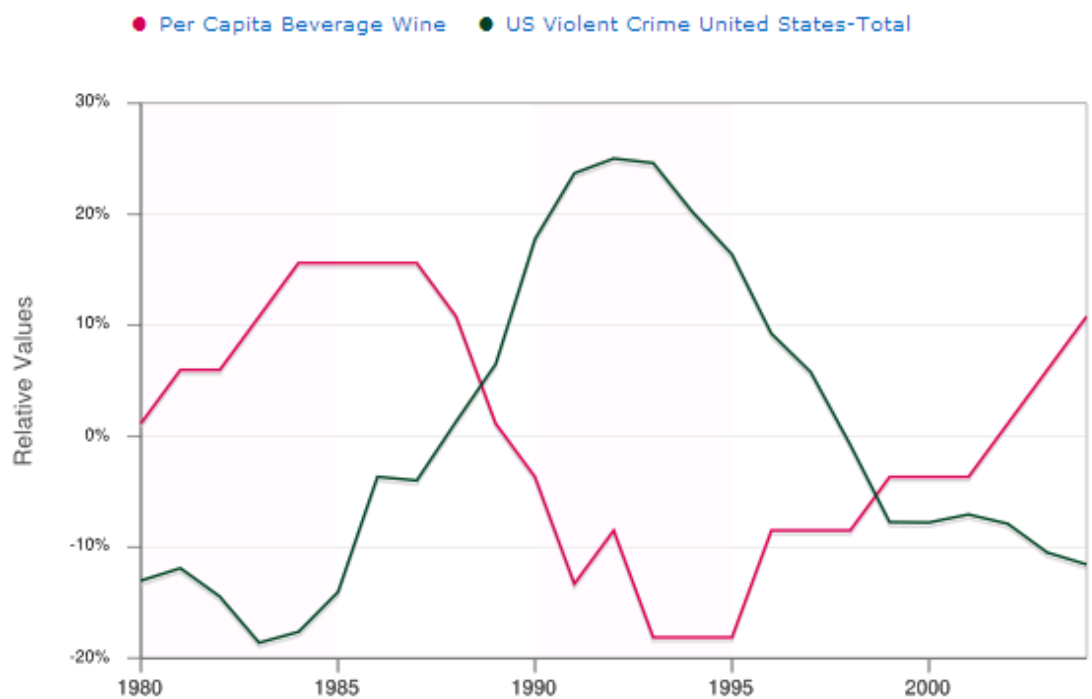
Vautard, R., P. Yiou, F. D'Andrea, N. de Noblet, N. Viovy, C. Cassou, J. Polcher, P. Ciais, M. Kageyama, and Y. Fan (2007), Summertime European heat and drought waves induced by wintertime Mediterranean rainfall deficit, Geophys. Res. Lett., 34, L07711, doi:10.1029/2006GL028001.

Two cautionary notes

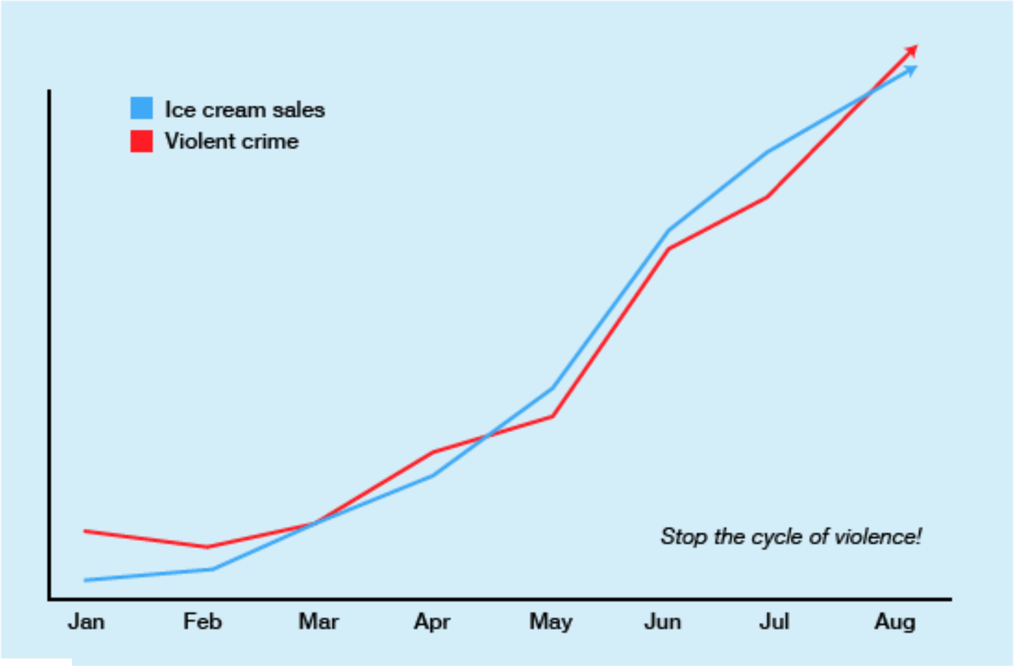
1. One should always check the statistical significance of the correlations one computes.

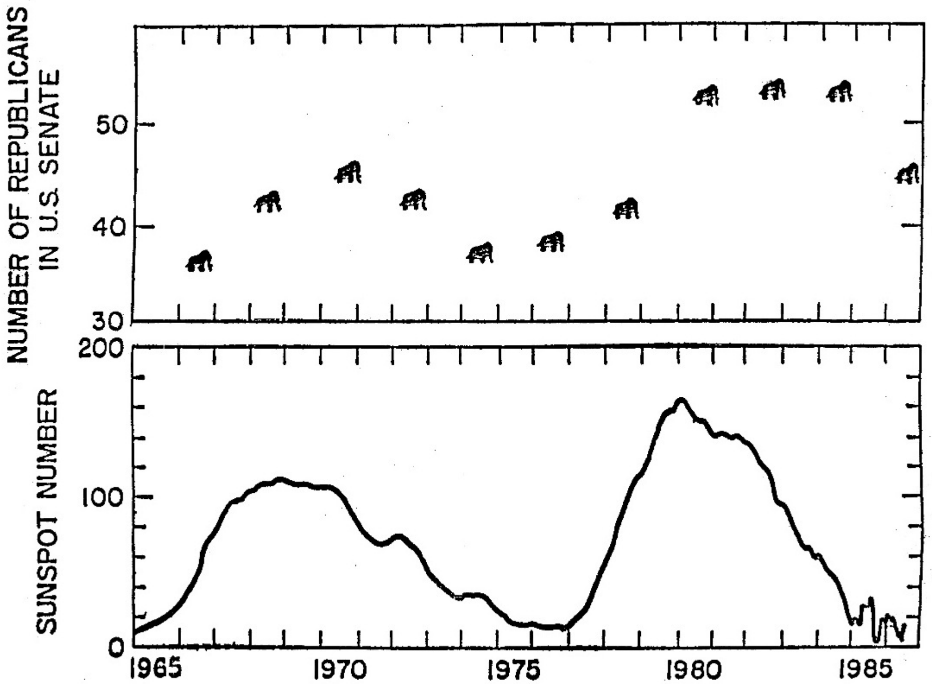
2. High correlation does NOT imply causality.

hypothesis testing:
1)Set up a Null Hypothesis
2)Chose a test statistic
3)Chose a level of significance
4)Compute the level of probability of the sample, from the test statistic
5)If that is lower than the level of significance, reject the null hypothesis.

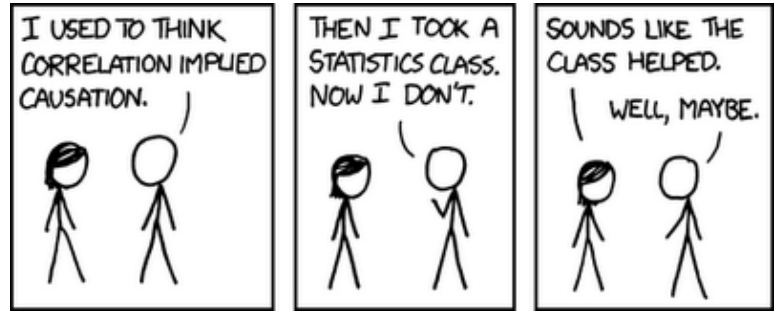


Reduce Violent crimes?
Stop eating ice cream!

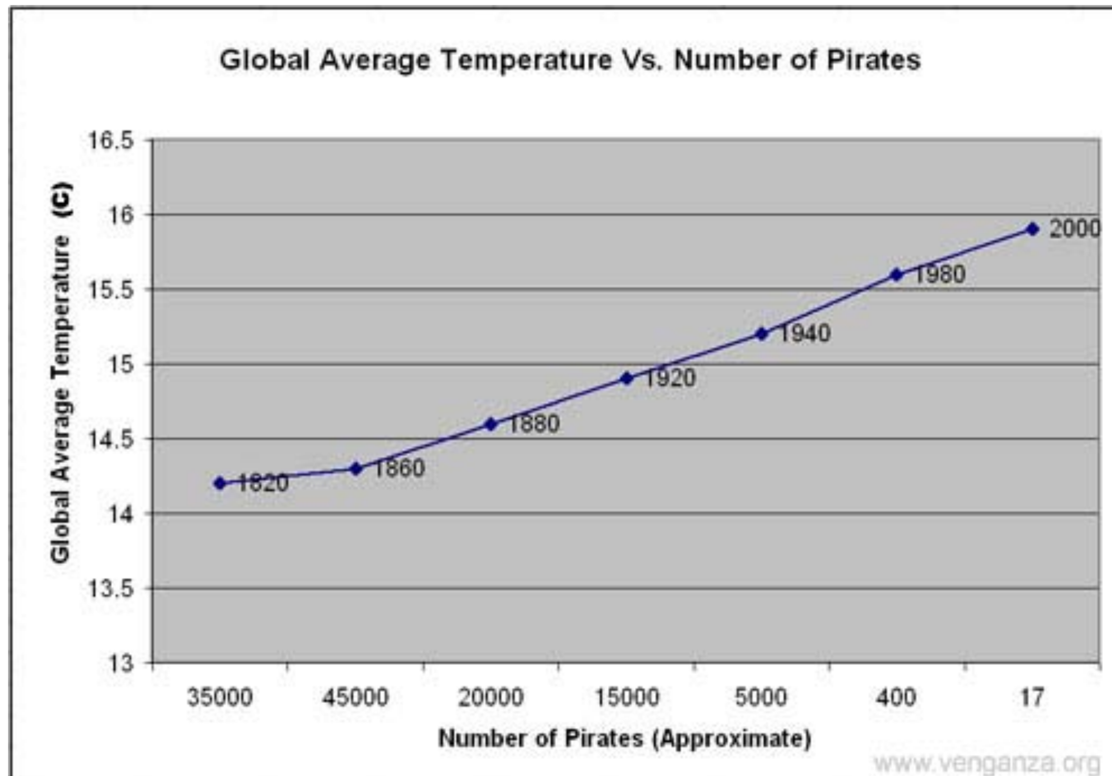




"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"



Church of the Flying Spaghetti Monster
http://www. venganza.org/

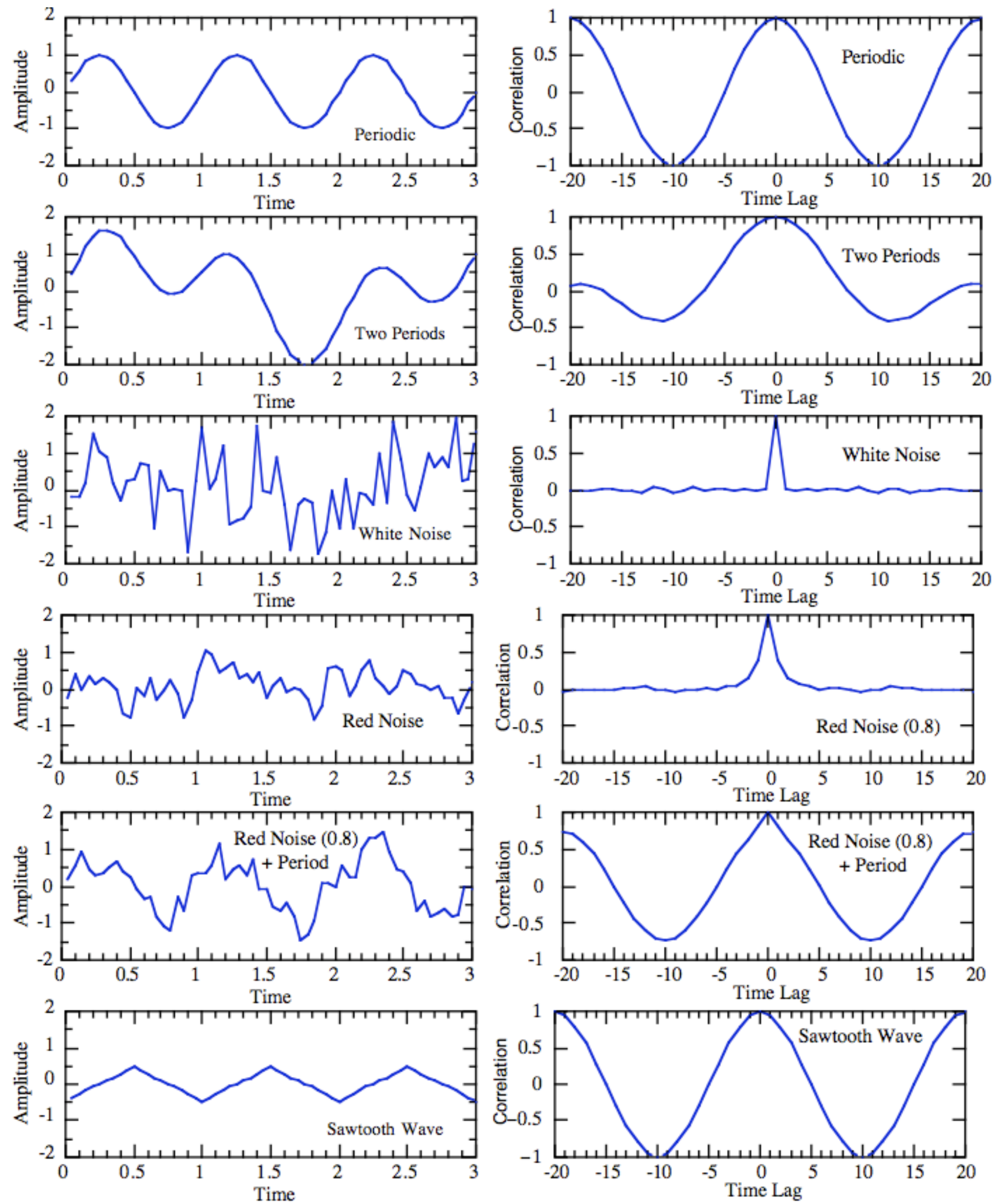


AUTOCORRELATION

Correlation of a time series with itself.

$$\phi(L) = \frac{1}{N-2L}\sum_{k=L}^{N-L} x'_k x'_{k+L} = \overline{x'_k \, x'_{k+L}}$$

$$L = 0, \pm 1, \pm 2 \ldots\ldots$$



Montecarlo methods are a class of computational algorithms that rely on repeated random sampling to computer their results.

It is useful when one doesn’t know a priori the PDF of the statistical parameter to be tested.

Example**, test the correlation of two timeseries.**

1) Set up null hypothesis

    The two series are not correlated

2) Chose a test statistics

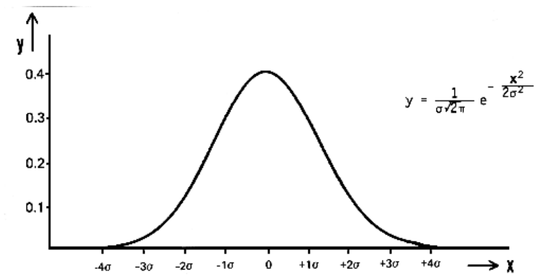
    Create 1000 random time series of the same length, mean and variance of one of the two series, estimate PDF from it.

3) Chose a level of significance
4) Compute the probability of the sample

    Compare the value of the correlation of the two series with the correlation of one series with all the random ones.

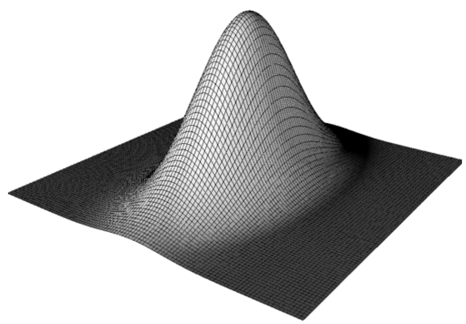
5) Reject or accept the null hypothesis.

Analysing a vector series $\mathbf{x}(t)$



Everybody is familiar with scalar time series statistics. Mean, variance, correlation, etc.

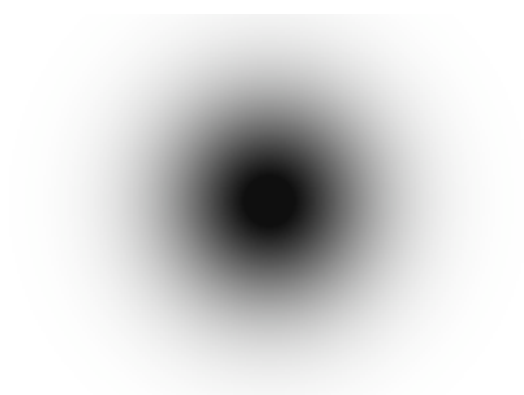
What happens with vector time series?



The mean is easy. Let's suppose $\bar{x} = 0$

But what takes the place of variance?

The covariance matrix: $\overline{xx^T}$



$$C = \begin{pmatrix} \overline{x_1 x_1} & \overline{x_1 x_2} & \ldots\ldots & \ldots\ldots & \overline{x_1 x_N} \\ \overline{x_2 x_1} & \overline{x_2 x_2} & \ldots\ldots & \ldots\ldots & \overline{x_2 x_N} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ \overline{x_N x_1} & \overline{x_N x_2} & \ldots\ldots & \ldots\ldots & \overline{x_N x_N} \end{pmatrix}$$

C gives the variance of the sample in any given direction in phase space.

So if $\mathbf{e}$ is a unitary vector,

$$\mathbf{e}^T C \mathbf{e}$$

is the variance in the direction $\mathbf{e}$.

ROADMAP for exercise 1.

1) Go to http://www.lmd.ens.fr/dandrea/TEACH/index.html

2) Get city temperature data (filenames T_jja_City.txt)
These are mean daily summer temperature data for a few cities in Europe, in the files there are two column: the date and the temperature in C.

3) Chose a city and read the data into MATLAB or Python

4) Compute mean and standard deviation of the daily data.
Do a loop, don't cheat using pre-made function like mean() or std().
Numerical trick: can you compute mean AND variance in one loop only?

5) Compute yearly temperature means of the city you chose. Then chose another city and do the same.

6) Compute the correlation of the yearly temperatures of the two cities
NB: *the temperature timerseries of the different cities may not be syncronous.*

7) Is the correlation significant? Do a montecarlo test of the correlation.
Compute the correlation of one of the two series with a high number of random series having the same mean and variances as the other timeseries.