UD 5

INFERENCE

INFERENCE

Part 1: Distributions in sampling

Part 2: Inference about one population

Comparison of populations

Part 3: ANOVA (Analysis of Variance)

Part 4: Regression

UD 5 part 1

Distributions in sampling

GENERAL CONCEPTS

POPULATION

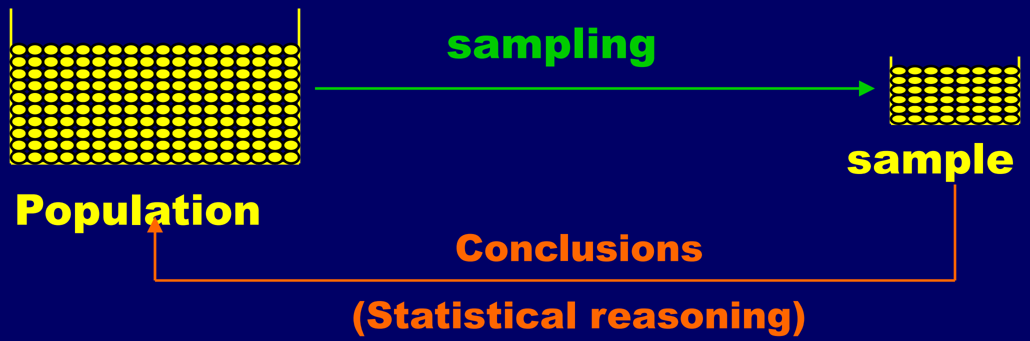
Set of objects that we are interested in obtaining conclusions.

Example: All pieces that are to be manufactured in a certain process.

SAMPLE

Subset formed by part of objects of one population.

Example: 10 pieces taken from the process.



The sample must be "representative" of the population.

Only way to guarantee "representativity": random sampling.

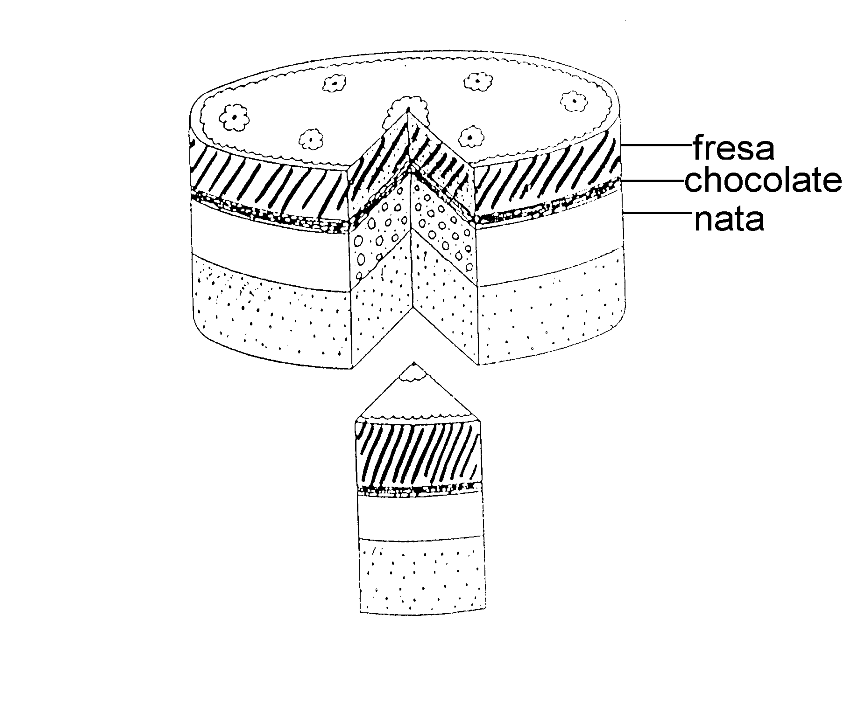
OBJECT OF SAMPLING

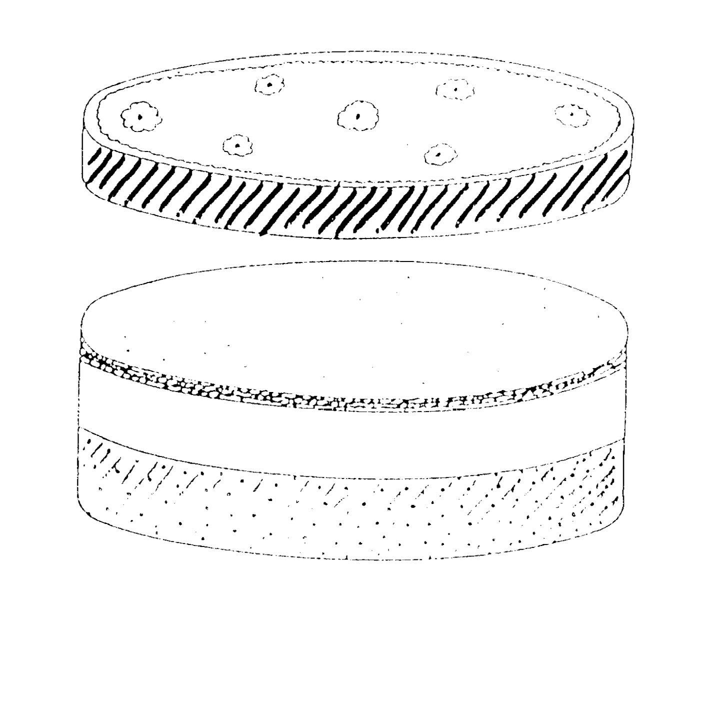
To obtain precise and reliable conclusions about the population characteristics (at the minimum cost), based on the sample analysis.

STATISTICAL INFERENCE

Process of reasoning to obtain conclusions (with a known margin of error) about the population, based on the analysis of samples taken from it.

GOLDEN RULE OF SAMPLING





THE SAMPLE MUST BE REPRESENTATIVE OF THE WHOLE SET

EXAMPLE OF A VERY BAD SAMPLING

Sampling procedure to estimate TV audience share:

http://www.elmundo.es/elmundo/2011/06/12/television/1307893647.html

CHARACTERISTICS OF SAMPLING

- **RANDOM**

  Any unit of the population must have the same probability to be chosen as part of the sample.

- **ADEQUATE SAMPLE SIZE according to:**

- **Size of the population under study**
- **Variability of the evaluated characteristic**
- **Maximum errors allowed in the estimation**

EXAMPLE: How would you select 100 people for the TV program: "I have a question for you, Mr. President" ?

Difference between: Simple random sampling (s.r.s.)
stratified random sampling (e.g. social strata)

EXAMPLE:

In order to study if the manufacturing process of a certain piece works correctly, 10 pieces have been randomly taken, being the length (in mm) the following:

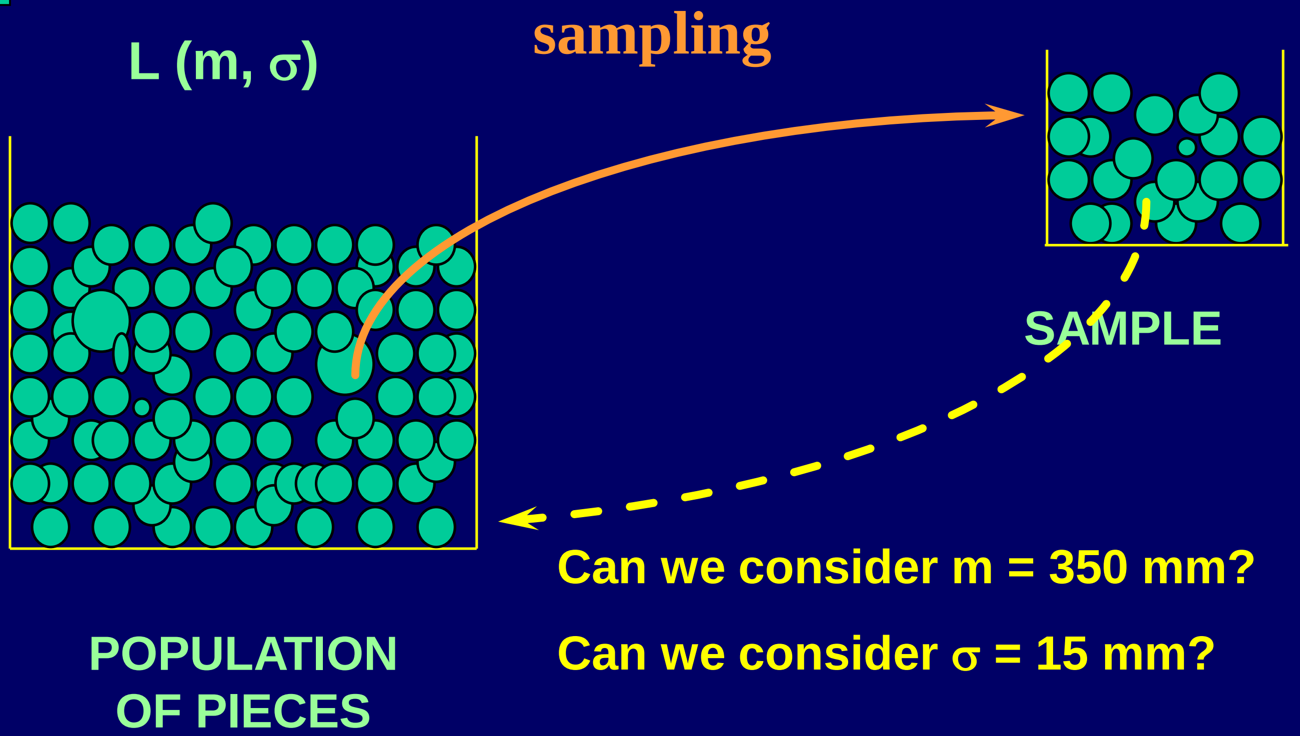
348.3 378.9 329.6 379.3 348.8 367.7 358.4 378.2 377.9 341.8

The sample mean is: $\bar{x} = 360.89 \text{ mm}$

The sample standard deviation is: $s = 18.11 \text{ mm}$

Can we consider that the population mean of the pieces' length is = 350 mm, which is the nominal value ?

Can we consider that the population standard deviation of the pieces' length is 15 mm ?



$\bar{x} = 360.89$ mm

$s = 18.11$ mm

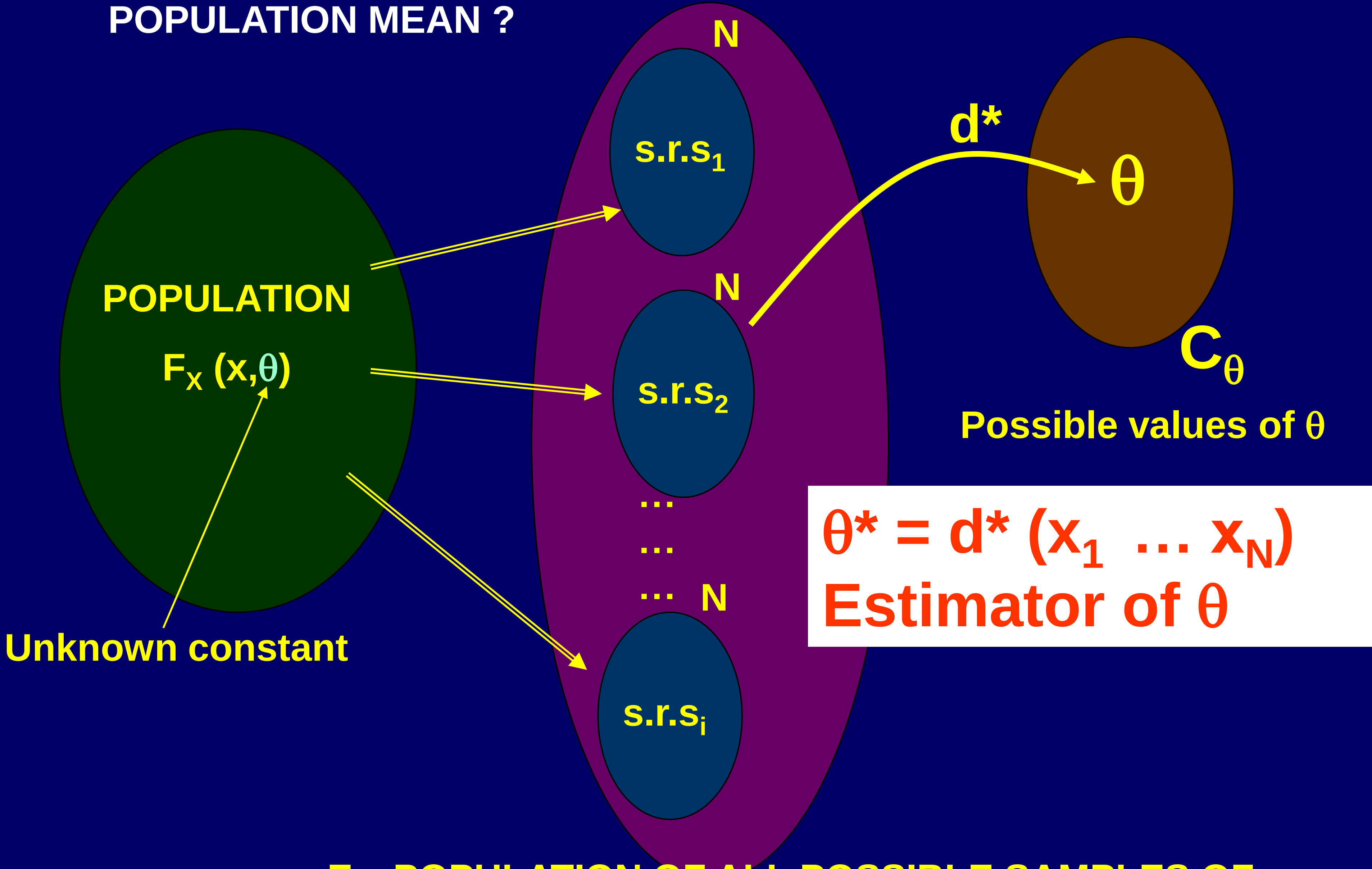
Can we consider m = 350 mm?

Can we consider σ = 15 mm?

Depends on ...

To what extent the average ($\bar{x}$) and the standard deviation (s) of one sample can differ from the average (m) and the standard deviation (σ) of the population, respectively.

IS THE SAMPLE MEAN THE BEST ESTIMATOR OF THE POPULATION MEAN ?



E_X: POPULATION OF ALL POSSIBLE SAMPLES OF SIZE N TAKEN FROM THE POPULATION

We want to know what is the average length of pieces manufactured in a certain process. For that purpose, a sample of 4 pieces is taken, and the length if each one is measured (x_i)

What is the best estimator of the population mean, m ?

θ* (estimator of θ) is unbiased if: $E(\theta^*) = \theta$

$$m^* = \frac{x_1 + x_2 + x_3 + x_4}{4}$$

Unbiased estimator, consistent of minimum variance

$$m^* = \frac{x_{min} + x_{max}}{2}$$

$$m^* = \text{median}\{x_1, x_2, x_3, x_4\}$$

Unbiased estimators
But which of all is the "best" estimator?

Unbiased if the distrib. is Normal

$$m^* = (x_1 \cdot x_2 \cdot x_3 \cdot x_4)^{1/4}$$

unbiased?

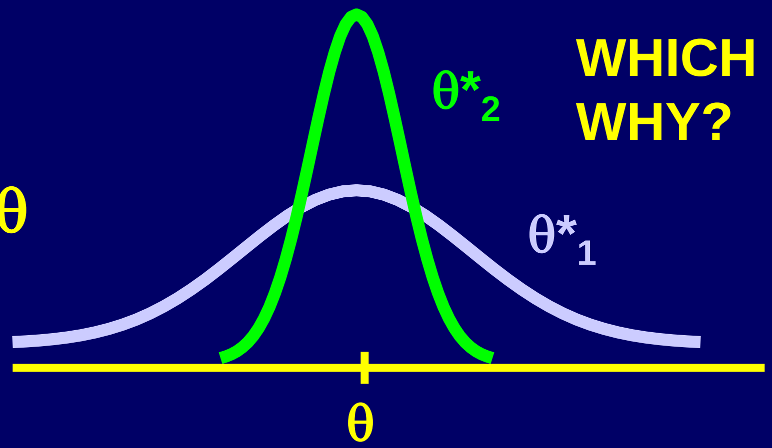
$$m^* = \min\{x_1, x_2, x_3, x_4\}$$

Biased estimator: if n is high, the deviation from m increases

Bias $(\theta^*, \theta) = E(\theta^*) - \theta$

How to assess the goodness of one estimator θ* for the estimation of θ?

If θ^*_1 and θ^*_2 are 2 estimators of θ



WHICH IS BETTER? WHY?

The best, is the estimator unbiased, of minimum variance and consistent

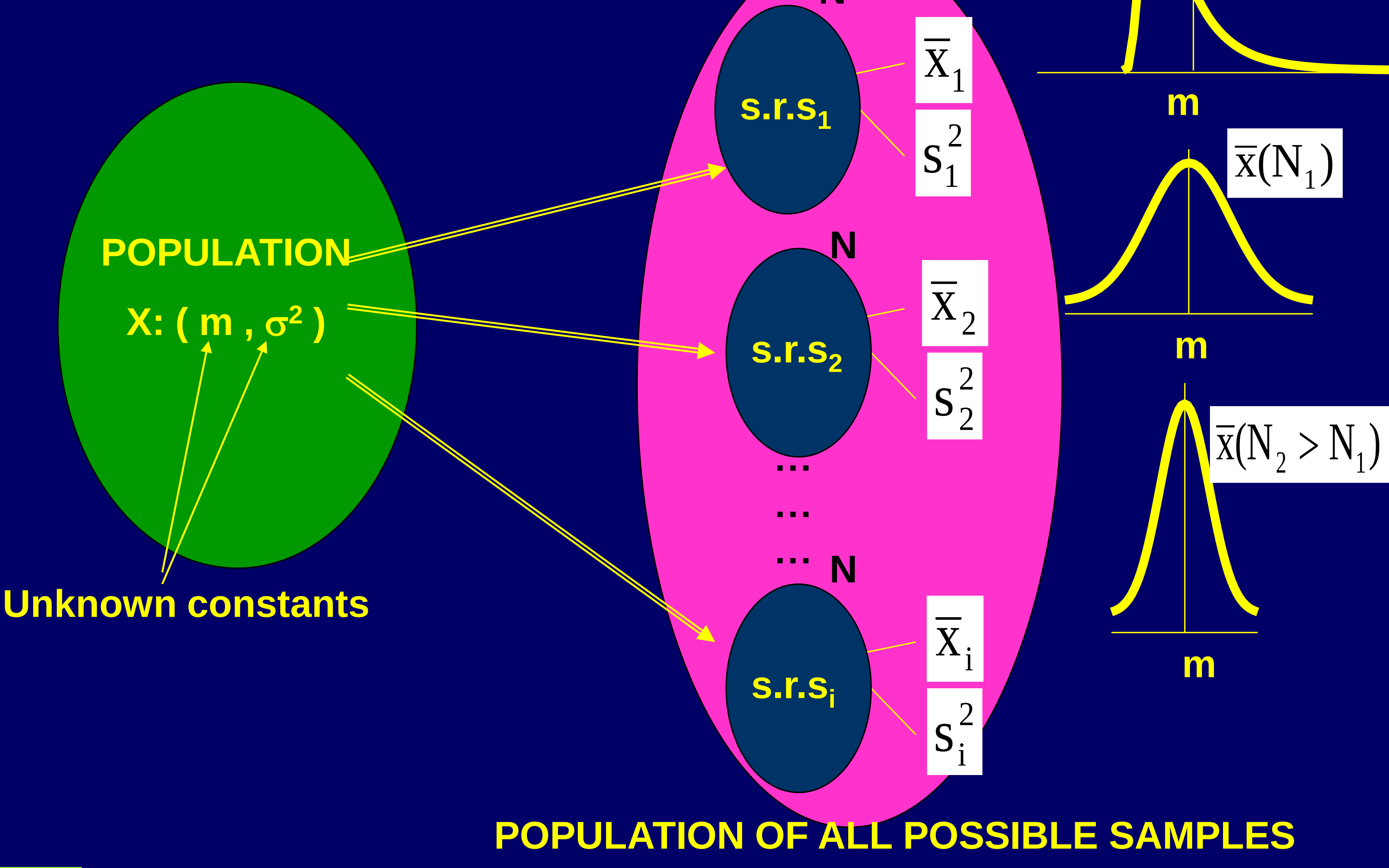
$$\lim \xrightarrow{n\to\infty} \sigma^2(\hat{\theta}) = 0$$

- the sample mean is the best estimator of m

- the sample variance is the best estimator of σ^2

$$\hat{m} = \overline{X} \qquad \hat{\sigma}^2 = S^2_{n-1} \qquad \hat{P} = p$$

SAMPLE MEAN



POPULATION
X: (m , σ^2)
Unknown constants
N
s.r.s_1
$\bar{X}_1$
S_1^2
N
s.r.s_2
$\bar{X}_2$
S_2^2
...
...
...
N
s.r.s_i
$\bar{X}_i$
S_i^2
POPULATION OF ALL POSSIBLE SAMPLES
X
m
$\bar{x}(N_1)$
m
$\bar{x}(N_2 > N_1)$
m

DISTRIBUTION OF $\overline{X}$

THE SAMPLE MEAN IS CALCULATED AS:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum X_i}{N}$$

EACH ONE OF THESE X_i THAT CONSTITUTES THE SAMPLE, WILL BE THE OBSERVED VALUE OF A RANDOM VARIABLE WITH MEAN m AND VARIANCE σ^2.

$$E(\overline{X}) = E\left(\frac{X_1 + X_2 + \ldots + X_N}{N}\right) = \frac{m + m + \ldots + m}{N} = m$$

THE AVERAGE OF SAMPLE MEAN IS THE POPULATION MEAN

(for any kind of distribution of X)

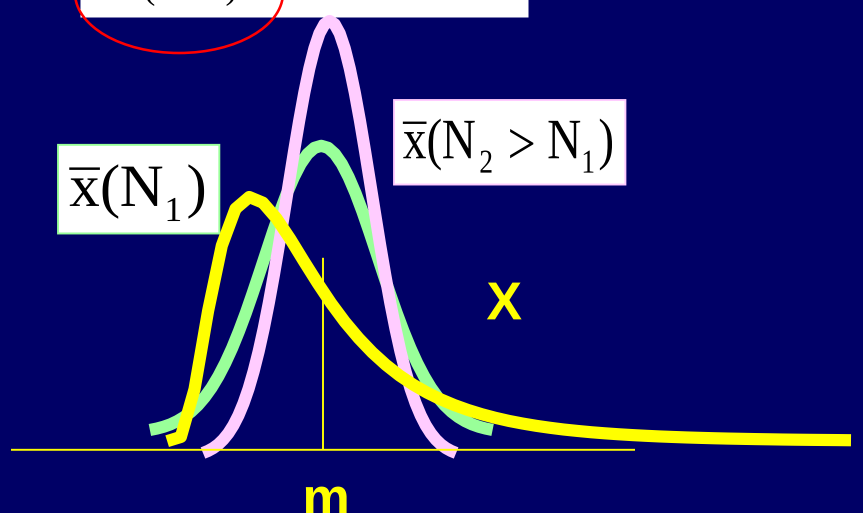
$$\sigma^2(\bar{X})=\sigma^2\left(\frac{X_1+X_2+\ldots+X_N}{N}\right)=\frac{1}{N^2}(\sigma^2(X_1)+\ldots+\sigma^2(X_N))=\frac{N\sigma^2}{N^2}=\frac{\sigma^2}{N}$$

independence

THE VARIANCE OF THE SAMPLE MEAN IS THE POPULATION VARIANCE DIVIDED BY THE SAMPLE SIZE (for any distribution)

$\bar{X}$ **is sum of independent random vars. with the same distribution**

$$\bar{X} \underset{\substack{N\to\infty\\ (\mathrm{CLT})}}{\approx} N\left(m,\frac{\sigma}{\sqrt{N}}\right)$$



EXERCISE :

In the process of car painting, the thickness of the paint layer follows a normal distribution with average 100 μm and standard deviation 5 μm. The quality control of this process is conducted by obtaining the average of 4 measurements from 4 cars randomly selected. The process is considered as correct if the mean obtained is > 95 μm. What is the probability to reject the process?

SOLUTION:

Mean of the 4 measurements:

$$\bar{X} \equiv N\left(100, \frac{5}{\sqrt{4}}\right) \equiv N(100, 2.5)$$

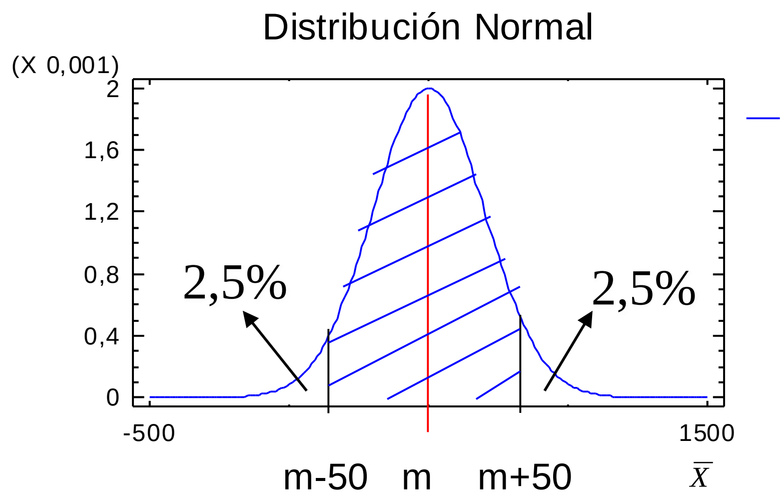
Probability to reject the process:

$$P = P(\bar{X} \leq 95) = \phi\left(\frac{95-100}{2.5}\right) = \phi(-2) = 0.0228$$

EXERCISE :

In order to know the average expenses of Spanish families in summer holidays, **N** families are randomly chosen and asked about their expenses. The population standard deviation is assumed to be $\sigma = 200$ €

What should be the value N so that the difference (in absolute value) between the sample mean obtained and the unknown population mean is < 50 € with a 95% probability ?



$$P\left(\left|(\bar{x}-m)\right| \leq 50\right) \geq 0{,}95 \quad \equiv \quad P\left(\bar{x} \leq m-50\right) \leq 0{,}025 \quad = \phi\left(\frac{(m-50)-m}{\frac{200}{\sqrt{n}}}\right) \leq 0{,}025$$

$$\left(\frac{(\mathrm{m}-50)-\mathrm{m}}{\frac{200}{\sqrt{\mathrm{n}}}}\right) = -1{,}96 \quad \Rightarrow \frac{-50\sqrt{\mathrm{n}}}{200} = -1{,}96 \Rightarrow \mathrm{n} = 62{,}14 \approx 63 \quad \text{families}$$

DISTRIBUTION OF s^2

$$s^2_{n-1} = \frac{(X_1 - \overline{X})^2 + \dots + (X_N - \overline{X})^2}{N-1} = \frac{\sum (X_i - \overline{X})^2}{N-1}$$

$$E(s^2_{n-1}) = E\left(\frac{\sum (X_i - \overline{X})^2}{N-1}\right) = \frac{1}{N-1} E\left[\sum \left[(X_i - m) - (\overline{X} - m)\right]^2\right] = \dots = \sigma^2$$

Unbiased estimator

$$E(s^2_n) = \frac{n-1}{n} \cdot \sigma^2$$

Asymptotically unbiased estimator

THE AVERAGE OF THE SAMPLE VARIANCE IS THE POPULATION VARIANCE

$$\sigma^2(s^2_{n-1}) = \frac{2\sigma^4}{n-1} \xrightarrow[N \to \infty]{} 0$$

Consistent estimator

SUM OF NORMAL VARIABLES

If X_1, X_2, ..., X_n are random variables Normally distributed with average m_x and standard deviation σ equal for all of them, then:

$$\sum_{i=1}^{n} X_i \sim N\left(n \cdot m_x, \sqrt{n} \cdot \sigma\right)$$

Consequently:

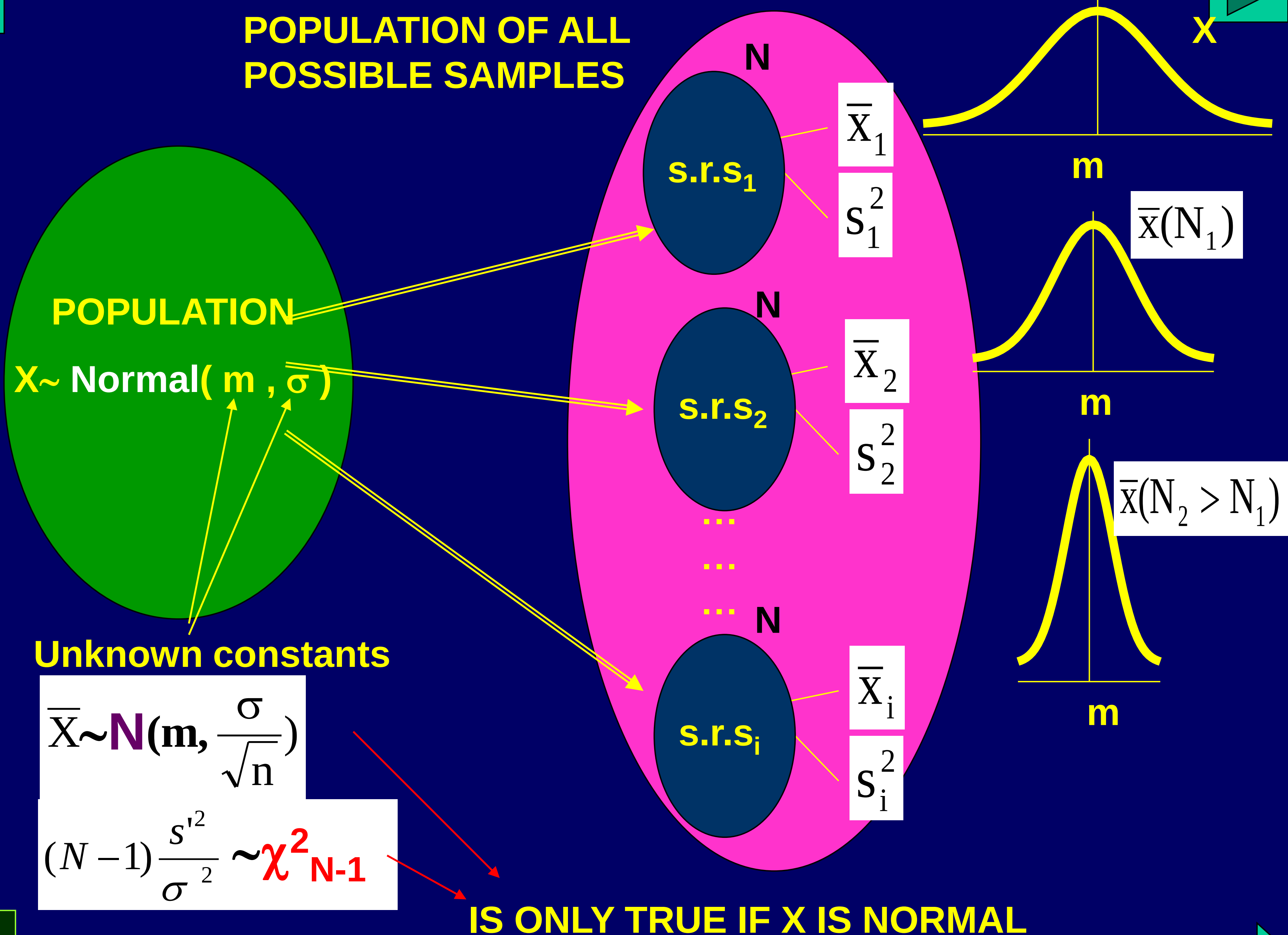
(for any value n)

$$\overline{X}_n \sim N\left(m_x, \frac{\sigma}{\sqrt{n}}\right)$$

$$\frac{\bar{x} - m}{\sigma / \sqrt{n}}_{\,n} \sim N(0;\ 1)$$

If X is not Normally distributed:

$$\overline{X} \underset{\substack{N \to \infty \\ (CLT)}}{\approx} N\left(m, \frac{\sigma}{\sqrt{N}}\right)$$



POPULATION OF ALL POSSIBLE SAMPLES
POPULATION
X~ Normal(m , σ)
Unknown constants
N
s.r.s$_1$
$\bar{X}_1$
s_1^2
N
s.r.s$_2$
$\bar{X}_2$
s_2^2
...
N
s.r.s$_i$
$\bar{X}_i$
s_i^2
X
m
$\bar{x}(N_1)$
m
$\bar{x}(N_2 > N_1)$
m
$\bar{X} \sim N(m, \frac{\sigma}{\sqrt{n}})$
$(N-1)\frac{s'^2}{\sigma^2} \sim \chi^2_{N-1}$
IS ONLY TRUE IF X IS NORMAL

To study the pattern of variability of statistical parameters that appear in the sampling of normal variables,

it is necessary to know three new probability distributions:

- **χ^2 (Pearson's chi-square distribution)**
- **Student's t**
- **Fisher's F (or F of Snedecor)**

IMPORTANT NOTE:

THESE DISTRIBUTIONS DO NOT MODEL THE PATTERN OF VARIABILITY OF ANY REAL VARIABLE; THEY APPEAR IN THE PROCESS OF STATISTICAL INFERENCE.

DISTRIBUTION χ^2

$$\chi_n^2 = \sum_{i=1}^{n} X_i^2 \quad ; \quad \mathbf{X_i} \sim N(0,1) \quad \textbf{independent}$$

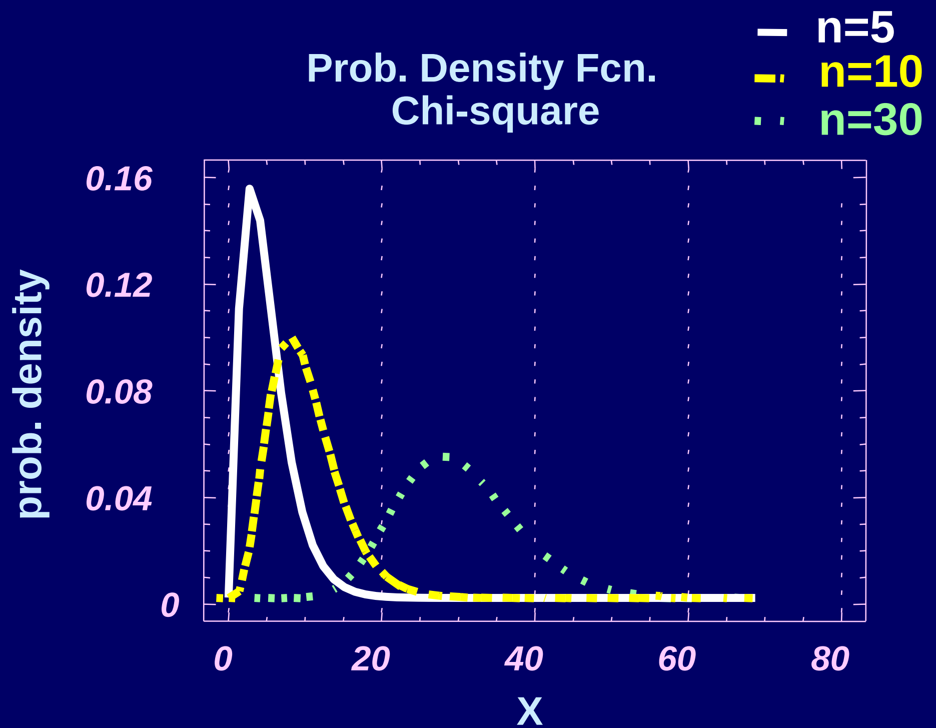
$$E(\chi_n^2) = n$$

$$\sigma^2(\chi_n^2) = 2n$$

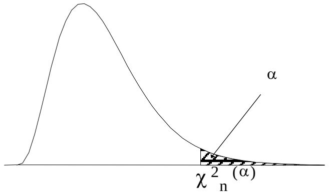
$$\frac{\chi_n^2 - n}{\sqrt{2n}} \xrightarrow{n\to\infty} N(0,1)$$

(for n>50, good approximation)

(formula table, up to n = 600)



PEARSON'S χ^2 DISTRIBUTION



n	0.9995	0.999	0.995	0.99	0.975	0.95	0.90	0.50	0.10	0.050	0.025	0.01	0.005	0.001	0.0005
1	0.000	0.000	0.000	0.000	0.001	0.004	0.016	0.455	2.706	3.842	5.024	6.635	7.879	10.827	12.115
2	0.001	0.002	0.010	0.020	0.051	0.103	0.211	1.386	4.605	5.992	7.378	9.210	10.597	13.815	15.201
3	0.015	0.024	0.072	0.115	0.216	0.352	0.584	2.366	6.251	7.815	9.348	11.345	12.838	16.266	17.731
4	0.064	0.091	0.207	0.297	0.484	0.711	1.064	3.357	7.779	9.488	11.143	13.277	14.860	18.466	19.998
5	0.158	0.210	0.412	0.554	0.831	1.146	1.610	4.352	9.236	11.071	12.833	15.086	16.750	20.515	22.106
6	0.299	0.381	0.676	0.872	1.237	1.635	2.204	5.348	10.645	12.592	14.449	16.812	18.548	22.457	24.102
7	0.485	0.599	0.989	1.239	1.690	2.167	2.833	6.346	12.017	14.067	16.013	18.475	20.278	24.321	26.018
8	0.710	0.857	1.344	1.647	2.180	2.733	3.490	7.344	13.362	15.507	17.535	20.090	21.955	26.124	27.867
9	0.972	1.152	1.735	2.088	2.700	3.325	4.168	8.343	14.684	16.919	19.023	21.666	23.589	27.877	29.667
10	1.265	1.479	2.156	2.558	3.247	3.940	4.865	9.342	15.987	18.307	20.483	23.209	25.188	29.588	31.419
11	1.587	1.834	2.603	3.054	3.816	4.575	5.578	10.341	17.275	19.675	21.920	24.725	26.757	31.264	33.138
12	1.935	2.214	3.074	3.571	4.404	5.226	6.304	11.340	18.549	21.026	23.337	26.217	28.300	32.909	34.821
13	2.305	2.617	3.565	4.107	5.009	5.892	7.042	12.340	19.812	22.362	24.736	27.688	29.819	34.527	36.477
14	2.697	3.041	4.075	4.660	5.629	6.571	7.790	13.339	21.064	23.685	26.119	29.141	31.319	36.124	38.109
15	3.107	3.483	4.601	5.229	6.262	7.261	8.547	14.339	22.307	24.996	27.488	30.578	32.802	37.698	39.717
16	3.536	3.942	5.142	5.812	6.908	7.962	9.312	15.339	23.542	26.296	28.845	32.000	34.267	39.252	41.308
17	3.980	4.416	5.697	6.408	7.564	8.672	10.085	16.338	24.769	27.587	30.191	33.409	35.718	40.791	42.881
18	4.439	4.905	6.265	7.015	8.231	9.390	10.865	17.338	25.989	28.869	31.526	34.805	37.156	42.312	44.434
19	4.913	5.407	6.844	7.633	8.907	10.117	11.651	18.338	27.204	30.144	32.852	36.191	38.582	43.819	45.974
20	5.398	5.921	7.434	8.260	9.591	10.851	12.443	19.337	28.412	31.410	34.170	37.566	39.997	45.314	47.498
21	5.895	6.447	8.034	8.897	10.283	11.591	13.240	20.337	29.615	32.671	35.479	38.932	41.401	46.796	49.010
22	6.404	6.983	8.643	9.543	10.982	12.338	14.042	21.337	30.813	33.925	36.781	40.289	42.796	48.268	50.510
23	6.924	7.529	9.260	10.196	11.689	13.091	14.848	22.337	32.007	35.173	38.076	41.638	44.181	49.728	51.999
24	7.453	8.085	9.886	10.856	12.401	13.848	15.659	23.337	33.196	36.415	39.364	42.980	45.558	51.179	53.478
25	7.991	8.649	10.520	11.524	13.120	14.611	16.473	24.337	34.382	37.653	40.647	44.314	46.928	52.619	54.948
26	8.537	9.222	11.160	12.198	13.844	15.379	17.292	25.337	35.563	38.885	41.923	45.642	48.290	54.051	56.407
27	9.093	9.803	11.808	12.879	14.573	16.151	18.114	26.336	36.741	40.113	43.195	46.963	49.645	55.475	57.856
28	9.656	10.391	12.461	13.565	15.308	16.928	18.939	27.336	37.916	41.337	44.461	48.278	50.994	56.892	59.299
29	10.227	10.986	13.121	14.256	16.047	17.708	19.768	28.336	39.088	42.557	45.722	49.588	52.336	58.301	60.734
30	10.804	11.588	13.787	14.954	16.791	18.493	20.599	29.336	40.256	43.773	46.979	50.892	53.672	59.702	62.160
40	16.906	17.917	20.707	22.164	24.433	26.509	29.051	39.335	51.805	55.759	59.342	63.691	66.766	73.403	76.096
50	23.461	24.674	27.991	29.707	32.357	34.764	37.689	49.335	63.167	67.505	71.420	76.154	79.490	86.660	89.560
60	30.339	31.738	35.534	37.485	40.482	43.188	46.459	59.335	74.397	79.082	83.298	88.379	91.952	99.608	102.697
70	37.467	39.036	43.275	45.442	48.758	51.739	55.329	69.335	85.527	90.531	95.023	100.43	104.22	112.32	115.58
80	44.792	46.520	51.172	53.540	57.153	60.392	64.278	79.334	96.578	101.88	106.62	112.32	116.32	124.84	128.26
90	52.277	54.156	59.196	61.754	65.647	69.126	73.291	89.334	107.56	113.15	118.14	124.11	128.29	137.20	140.78
100	59.895	61.918	67.328	70.065	74.222	77.929	82.358	99.334	118.49	124.34	129.56	135.81	140.17	149.45	153.16

EXERCISES:

1) Demonstrate that $E(\chi^2_n)=n$

2) Calculate the median of a χ^2_5 and of a χ^2_{50}

3) Justify intuitively that :

$$(N-1)\frac{s'^2}{\sigma^2} \sim \chi^2_{N-1}$$

4) What is the probability to obtain a sample variance > 10 when taking a sample of size 20 from a Normal population of $\sigma^2 = 5$?

t-STUDENT DISTRIBUTION

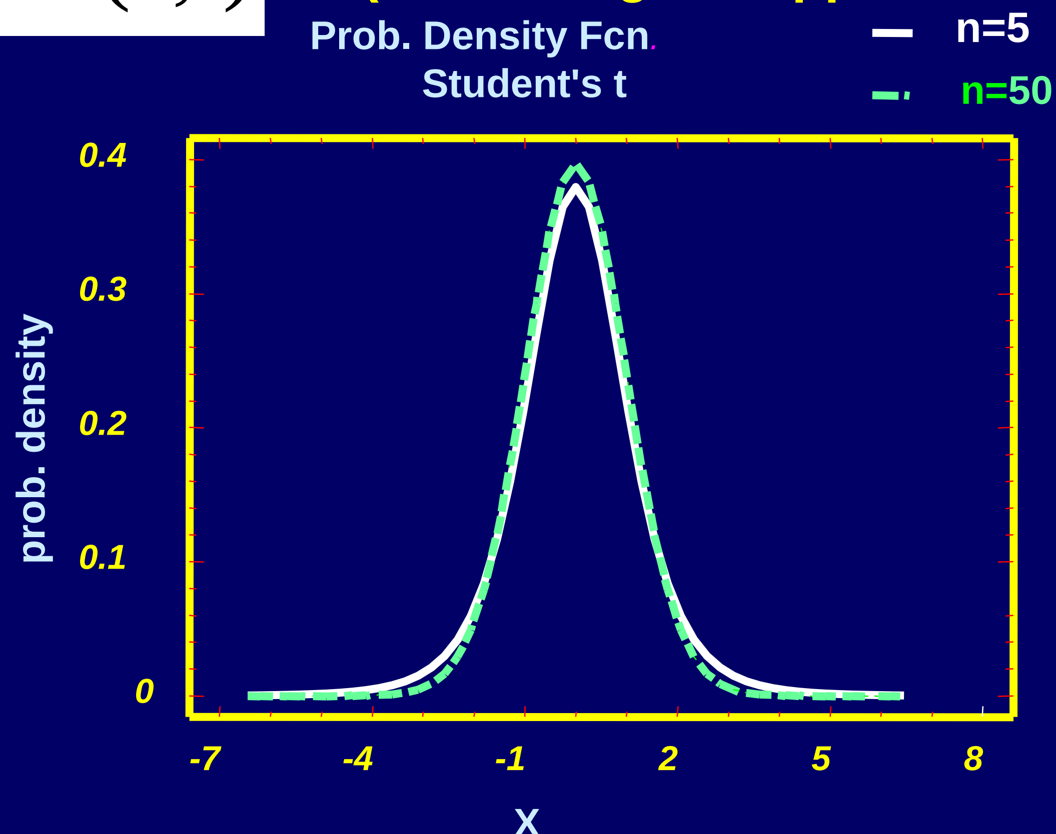
$$t_n = \frac{N(0,1)}{\sqrt{\frac{\chi_n^2}{n}}}$$ **independent**

$$E(t_n) = 0$$

$$\sigma^2(t_n) = \frac{n}{n-2} \quad (n > 2)$$

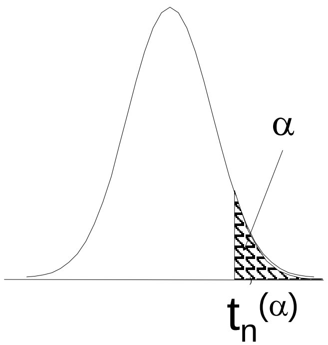
$$t_n \xrightarrow{n\to\infty} N(0,1)$$

(for n>30, good approximation)



Student’s t distribution

	Probabilidad de una cola												
n	**0.0005**	**0.001**	**0.005**	**0.01**	**0.025**	**0.05**	**0.1**	**0.2**	**0.25**	**0.3**	**0.4**	**0.45**	**0.475**
1	636.578	318.289	63.656	31.821	12.706	6.314	3.078	1.376	1.000	0.727	0.325	0.158	0.079
2	31.600	22.328	9.925	6.965	4.303	2.920	1.886	1.061	0.816	0.617	0.289	0.142	0.071
3	12.924	10.214	5.841	4.541	3.182	2.353	1.638	0.978	0.765	0.584	0.277	0.137	0.068
4	8.610	7.173	4.604	3.747	2.776	2.132	1.533	0.941	0.741	0.569	0.271	0.134	0.067
5	6.869	5.894	4.032	3.365	2.571	2.015	1.476	0.920	0.727	0.559	0.267	0.132	0.066
6	5.959	5.208	3.707	3.143	2.447	1.943	1.440	0.906	0.718	0.553	0.265	0.131	0.065
7	5.408	4.785	3.499	2.998	2.365	1.895	1.415	0.896	0.711	0.549	0.263	0.130	0.065
8	5.041	4.501	3.355	2.896	2.306	1.860	1.397	0.889	0.706	0.546	0.262	0.130	0.065
9	4.781	4.297	3.250	2.821	2.262	1.833	1.383	0.883	0.703	0.543	0.261	0.129	0.064
10	4.587	4.144	3.169	2.764	2.228	1.812	1.372	0.879	0.700	0.542	0.260	0.129	0.064
11	4.437	4.025	3.106	2.718	2.201	1.796	1.363	0.876	0.697	0.540	0.260	0.129	0.064
12	4.318	3.930	3.055	2.681	2.179	1.782	1.356	0.873	0.695	0.539	0.259	0.128	0.064
13	4.221	3.852	3.012	2.650	2.160	1.771	1.350	0.870	0.694	0.538	0.259	0.128	0.064
14	4.140	3.787	2.977	2.624	2.145	1.761	1.345	0.868	0.692	0.537	0.258	0.128	0.064
15	4.073	3.733	2.947	2.602	2.131	1.753	1.341	0.866	0.691	0.536	0.258	0.128	0.064
16	4.015	3.686	2.921	2.583	2.120	1.746	1.337	0.865	0.690	0.535	0.258	0.128	0.064
17	3.965	3.646	2.898	2.567	2.110	1.740	1.333	0.863	0.689	0.534	0.257	0.128	0.064
18	3.922	3.610	2.878	2.552	2.101	1.734	1.330	0.862	0.688	0.534	0.257	0.127	0.064
19	3.883	3.579	2.861	2.539	2.093	1.729	1.328	0.861	0.688	0.533	0.257	0.127	0.064
20	3.850	3.552	2.845	2.528	2.086	1.725	1.325	0.860	0.687	0.533	0.257	0.127	0.063
21	3.819	3.527	2.831	2.518	2.080	1.721	1.323	0.859	0.686	0.532	0.257	0.127	0.063
22	3.792	3.505	2.819	2.508	2.074	1.717	1.321	0.858	0.686	0.532	0.256	0.127	0.063
23	3.768	3.485	2.807	2.500	2.069	1.714	1.319	0.858	0.685	0.532	0.256	0.127	0.063
24	3.745	3.467	2.797	2.492	2.064	1.711	1.318	0.857	0.685	0.531	0.256	0.127	0.063
25	3.725	3.450	2.787	2.485	2.060	1.708	1.316	0.856	0.684	0.531	0.256	0.127	0.063
26	3.707	3.435	2.779	2.479	2.056	1.706	1.315	0.856	0.684	0.531	0.256	0.127	0.063
27	3.689	3.421	2.771	2.473	2.052	1.703	1.314	0.855	0.684	0.531	0.256	0.127	0.063
28	3.674	3.408	2.763	2.467	2.048	1.701	1.313	0.855	0.683	0.530	0.256	0.127	0.063
29	3.660	3.396	2.756	2.462	2.045	1.699	1.311	0.854	0.683	0.530	0.256	0.127	0.063
30	3.646	3.385	2.750	2.457	2.042	1.697	1.310	0.854	0.683	0.530	0.256	0.127	0.063
40	3.551	3.307	2.704	2.423	2.021	1.684	1.303	0.851	0.681	0.529	0.255	0.126	0.063
60	3.460	3.232	2.660	2.390	2.000	1.671	1.296	0.848	0.679	0.527	0.254	0.126	0.063
120	3.373	3.160	2.617	2.358	1.980	1.658	1.289	0.845	0.677	0.526	0.254	0.126	0.063
∞	3.290	3.090	2.576	2.326	1.960	1.645	1.282	0.842	0.674	0.524	0.253	0.126	0.063
n	**0.001**	**0.002**	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**	**0.4**	**0.5**	**0.6**	**0.8**	**0.9**	**0.95**



EXERCISE

Obtain a value x so that: $P\left(t_{10} > |x|\right) = 0.05$

IMPORTANCE OF THIS DISTRIBUTION:

If $\overline{X}$ and s are the mean and standard deviation of a sample with size N taken from a Normal population (m , σ) , the statistic:

$$\frac{\overline{X} - m}{s / \sqrt{N}} \sim t_{N-1}$$

IMPORTANT NOTE:

SEE THE ANALOGY BETWEEN:

$$\frac{\overline{X} - m}{\sigma / \sqrt{N}} \sim N(0,1)$$

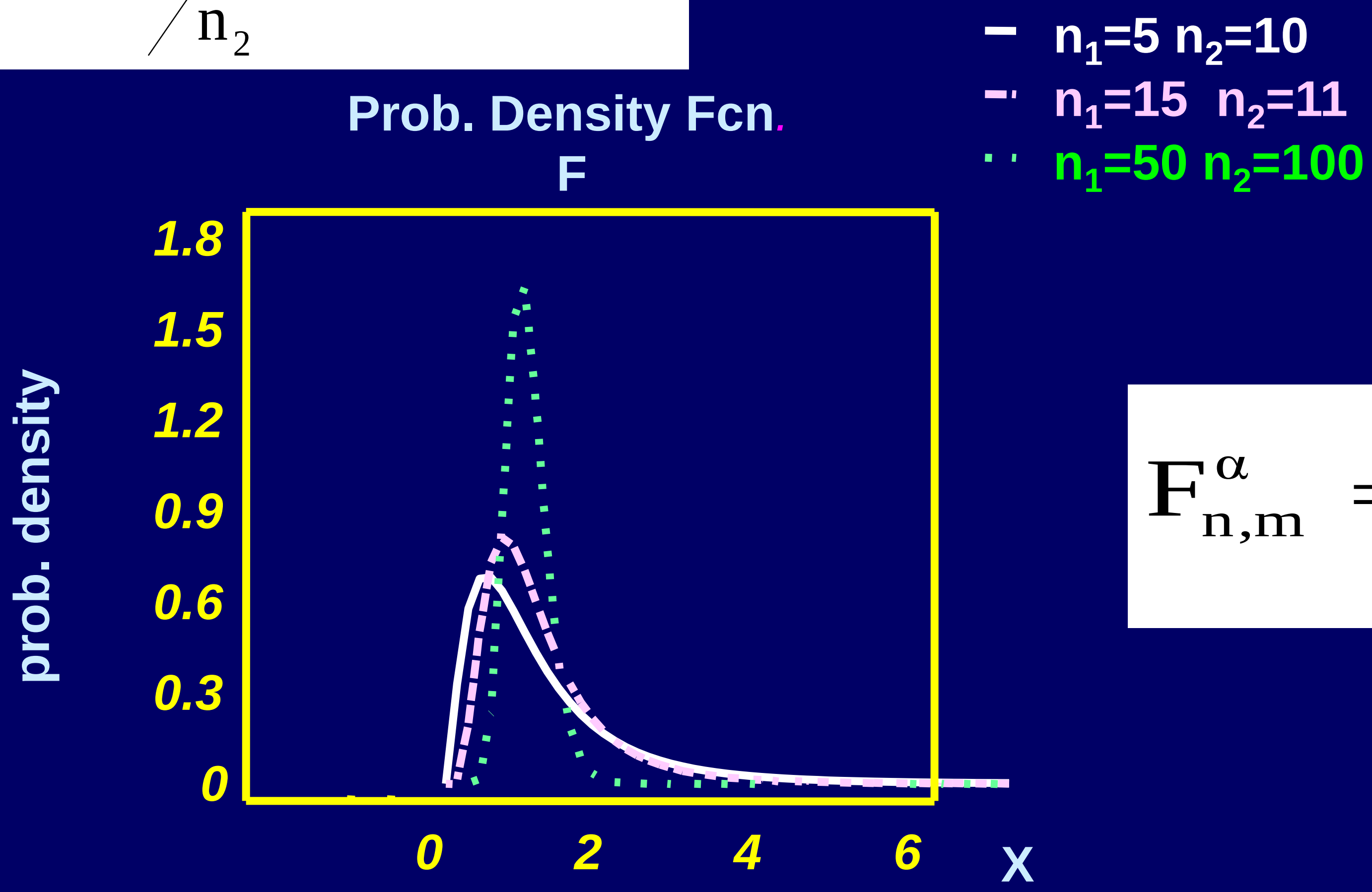
AND

$$\frac{\overline{X} - m}{s / \sqrt{N}} \sim t_{N-1}$$

Fisher's F distribution

$$F_{n_1,n_2} = \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$$ **independient**

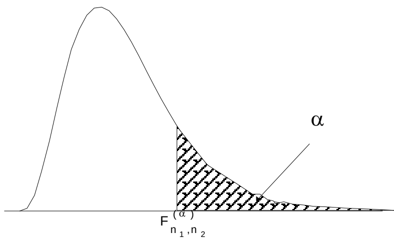
$$E(F_{n_1,n_2}) = \frac{n_2}{n_2 - 2} \quad (n_2 > 2)$$



$$F^{\alpha}_{n,m} = \frac{1}{F^{1-\alpha}_{m,n}}$$

Fisher's F distribution

	Grados de libertad de la varianza mayor															
	1		2		3		4		5		6		7		8	
p→	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
1	161.45	4052.2	199.50	4999.3	215.71	5403.5	224.58	5624.3	230.16	5763.9	233.99	5858.9	236.77	5928.3	238.88	5980.9
2	18.51	98.50	19.00	99.00	19.16	99.16	19.25	99.25	19.30	99.30	19.33	99.33	19.35	99.36	19.37	99.38
3	10.13	34.12	9.55	30.82	9.28	29.46	9.12	28.71	9.01	28.24	8.94	27.91	8.89	27.67	8.85	27.49
4	7.71	21.20	6.94	18.00	6.59	16.69	6.39	15.98	6.26	15.52	6.16	15.21	6.09	14.98	6.04	14.80
5	6.61	16.26	5.79	13.27	5.41	12.06	5.19	11.39	5.05	10.97	4.95	10.67	4.88	10.46	4.82	10.29
6	5.99	13.75	5.14	10.92	4.76	9.78	4.53	9.15	4.39	8.75	4.28	8.47	4.21	8.26	4.15	8.10
7	5.59	12.25	4.74	9.55	4.35	8.45	4.12	7.85	3.97	7.46	3.87	7.19	3.79	6.99	3.73	6.84
8	5.32	11.26	4.46	8.65	4.07	7.59	3.84	7.01	3.69	6.63	3.58	6.37	3.50	6.18	3.44	6.03
9	5.12	10.56	4.26	8.02	3.86	6.99	3.63	6.42	3.48	6.06	3.37	5.80	3.29	5.61	3.23	5.47
10	4.96	10.04	4.10	7.56	3.71	6.55	3.48	5.99	3.33	5.64	3.22	5.39	3.14	5.20	3.07	5.06
11	4.84	9.65	3.98	7.21	3.59	6.22	3.36	5.67	3.20	5.32	3.09	5.07	3.01	4.89	2.95	4.74
12	4.75	9.33	3.89	6.93	3.49	5.95	3.26	5.41	3.11	5.06	3.00	4.82	2.91	4.64	2.85	4.50
13	4.67	9.07	3.81	6.70	3.41	5.74	3.18	5.21	3.03	4.86	2.92	4.62	2.83	4.44	2.77	4.30
14	4.60	8.86	3.74	6.51	3.34	5.56	3.11	5.04	2.96	4.69	2.85	4.46	2.76	4.28	2.70	4.14
15	4.54	8.68	3.68	6.36	3.29	5.42	3.06	4.89	2.90	4.56	2.79	4.32	2.71	4.14	2.64	4.00
16	4.49	8.53	3.63	6.23	3.24	5.29	3.01	4.77	2.85	4.44	2.74	4.20	2.66	4.03	2.59	3.89
17	4.45	8.40	3.59	6.11	3.20	5.19	2.96	4.67	2.81	4.34	2.70	4.10	2.61	3.93	2.55	3.79
18	4.41	8.29	3.55	6.01	3.16	5.09	2.93	4.58	2.77	4.25	2.66	4.01	2.58	3.84	2.51	3.71
19	4.38	8.18	3.52	5.93	3.13	5.01	2.90	4.50	2.74	4.17	2.63	3.94	2.54	3.77	2.48	3.63
20	4.35	8.10	3.49	5.85	3.10	4.94	2.87	4.43	2.71	4.10	2.60	3.87	2.51	3.70	2.45	3.56
21	4.32	8.02	3.47	5.78	3.07	4.87	2.84	4.37	2.68	4.04	2.57	3.81	2.49	3.64	2.42	3.51
22	4.30	7.95	3.44	5.72	3.05	4.82	2.82	4.31	2.66	3.99	2.55	3.76	2.46	3.59	2.40	3.45
23	4.28	7.88	3.42	5.66	3.03	4.76	2.80	4.26	2.64	3.94	2.53	3.71	2.44	3.54	2.37	3.41
24	4.26	7.82	3.40	5.61	3.01	4.72	2.78	4.22	2.62	3.90	2.51	3.67	2.42	3.50	2.36	3.36
25	4.24	7.77	3.39	5.57	2.99	4.68	2.76	4.18	2.60	3.85	2.49	3.63	2.40	3.46	2.34	3.32
26	4.23	7.72	3.37	5.53	2.98	4.64	2.74	4.14	2.59	3.82	2.47	3.59	2.39	3.42	2.32	3.29
27	4.21	7.68	3.35	5.49	2.96	4.60	2.73	4.11	2.57	3.78	2.46	3.56	2.37	3.39	2.31	3.26
28	4.20	7.64	3.34	5.45	2.95	4.57	2.71	4.07	2.56	3.75	2.45	3.53	2.36	3.36	2.29	3.23
29	4.18	7.60	3.33	5.42	2.93	4.54	2.70	4.04	2.55	3.73	2.43	3.50	2.35	3.33	2.28	3.20
30	4.17	7.56	3.32	5.39	2.92	4.51	2.69	4.02	2.53	3.70	2.42	3.47	2.33	3.30	2.27	3.17
31	4.16	7.53	3.30	5.36	2.91	4.48	2.68	3.99	2.52	3.67	2.41	3.45	2.32	3.28	2.25	3.15
32	4.15	7.50	3.29	5.34	2.90	4.46	2.67	3.97	2.51	3.65	2.40	3.43	2.31	3.26	2.24	3.13
33	4.14	7.47	3.28	5.31	2.89	4.44	2.66	3.95	2.50	3.63	2.39	3.41	2.30	3.24	2.23	3.11
34	4.13	7.44	3.28	5.29	2.88	4.42	2.65	3.93	2.49	3.61	2.38	3.39	2.29	3.22	2.23	3.09
38	4.10	7.35	3.24	5.21	2.85	4.34	2.62	3.86	2.46	3.54	2.35	3.32	2.26	3.15	2.19	3.02
42	4.07	7.28	3.22	5.15	2.83	4.29	2.59	3.80	2.44	3.49	2.32	3.27	2.24	3.10	2.17	2.97
46	4.05	7.22	3.20	5.10	2.81	4.24	2.57	3.76	2.42	3.44	2.30	3.22	2.22	3.06	2.15	2.93
50	4.03	7.17	3.18	5.06	2.79	4.20	2.56	3.72	2.40	3.41	2.29	3.19	2.20	3.02	2.13	2.89
60	4.00	7.08	3.15	4.98	2.76	4.13	2.53	3.65	2.37	3.34	2.25	3.12	2.17	2.95	2.10	2.82
80	3.96	6.96	3.11	4.88	2.72	4.04	2.49	3.56	2.33	3.26	2.21	3.04	2.13	2.87	2.06	2.74
100	3.94	6.90	3.09	4.82	2.70	3.98	2.46	3.51	2.31	3.21	2.19	2.99	2.10	2.82	2.03	2.69
200	3.89	6.76	3.04	4.71	2.65	3.88	2.42	3.41	2.26	3.11	2.14	2.89	2.06	2.73	1.98	2.60
1000	3.85	6.66	3.00	4.63	2.61	3.80	2.38	3.34	2.22	3.04	2.11	2.82	2.02	2.66	1.95	2.53
∞	3.84	6.63	3.00	4.61	2.60	3.78	2.37	3.32	2.21	3.02	2.10	2.80	2.01	2.64	1.94	2.51



EXERCISE:

1) Justify intuitively that $E(F_{N_1,N_2}) \cong 1$

2) Calculate a value k so that: $P(F_{4,8} > k) = 0.05$

IMPORTANCE OF THIS DISTRIBUTION:

To compare the variability due to different sources:

If s_1^2 is the variance of a sample with size N_1 extracted from a Normal population (σ_1^2)
and s_2^2 is the variance of a sample with size N_2 extracted from a Normal population (σ_2^2)

and both samples are independent: $$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{N_1-1,N_2-1}$$

EXERCISE:

3) if two samples with size 10 are taken from the same Normal population, what is the probability to get the second sample variance double or higher than the first one?

4) We want to know if the accuracy of 2 machines filling bottles is the same. For this purpose:

- 9 bottles from machine 1 are weighted, being $s_1^2 = 180$

- 9 bottles from machine 2 are weighted, being $s_2^2 = 250$

Can we conclude that their accuracy is different?

UD 5 part 2

Inference about one population

HYPOTHESIS TESTS

Are used to decide if certain assumptions established a priori about the population are reasonable or not.

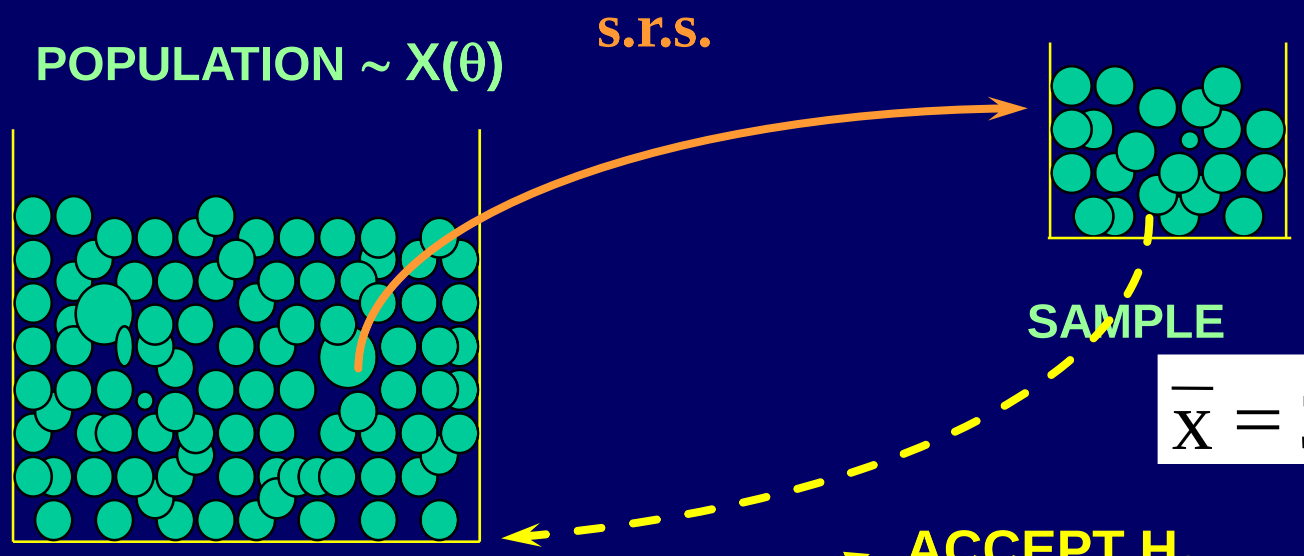
- If the assumptions are about the parameter values of the distribution: Parametric tests

- If assumptions are about other aspects, like type of distribution, independence, etc.: nonparametric tests

- **Null hypothesis (H_0): the one that we want to test (usually associated to the situation considered as correct, usual or desirable).**
- **Alternative hypothesis (H_1): the opposite to H_0 (usually associated to the situation considered as incorrect, unusual or undesirable).**

HYPOTHESIS TESTS

Can we consider that the average population length of pieces is 350 mm, which is the nominal value ?



$$\bar{x} = 360.89 \text{ mm}$$

$H_0 : \theta = \theta_0$ **m = 350**

$H_1 : \theta \neq \theta_0$ **m ≠ 350**

DECIDE → **ACCEPT H_0**

DECIDE → **REJECT H_0** → **ACCEPT H_1**

TYPES OF HYPOTHESES

SIMPLE HYPOTHESIS

H_0: m = 350

Corresponds to a single point $\theta = \theta_0$ of the parametric space C_θ

Assuming that this hypothesis is true, the population distribution is completely specified.

X ~ N (350,σ)

COMPOUND HYPOTHESIS

Corresponds to a region $\subset$ C_θ , containing more than one possible value of the parameter.

This type of hypothesis does not specify completely the population distribution.

H_0: m $\lesssim$ 350

H_1: m > 350

In this subject we will only consider the following tests:

$H_0: m = 100$ $H_0: \sigma^2 = 5$ $H_0: m_1 = m_2$ $H_0: \sigma^2{}_1 = \sigma^2{}_2$

$H_1: m \neq 100$ $H_1: \sigma^2 \neq 5$ $H_1: m_1 \neq m_2$ $H_1: \sigma^2{}_1 \neq \sigma^2{}_2$

Inference about one Normal population (first two tests)

Comparison of 2 Normal populations (last two tests)

How to study compound hypothesis tests:

$H_0: m \leq 100$

$H_1: m > 100$

1) Test the hypothesis:

$H_0: m = 100$ $H_1: m \neq 100$

2) Think with logic

E.g. $\bar{x} = 104 \longrightarrow$ accept $H_0: m=100 \longrightarrow$ accept $H_0: m \leq 100$

E.g. $\bar{x} = 109 \longrightarrow$ reject $H_0: m=100 \longrightarrow$ accept $H_1: m > 100$

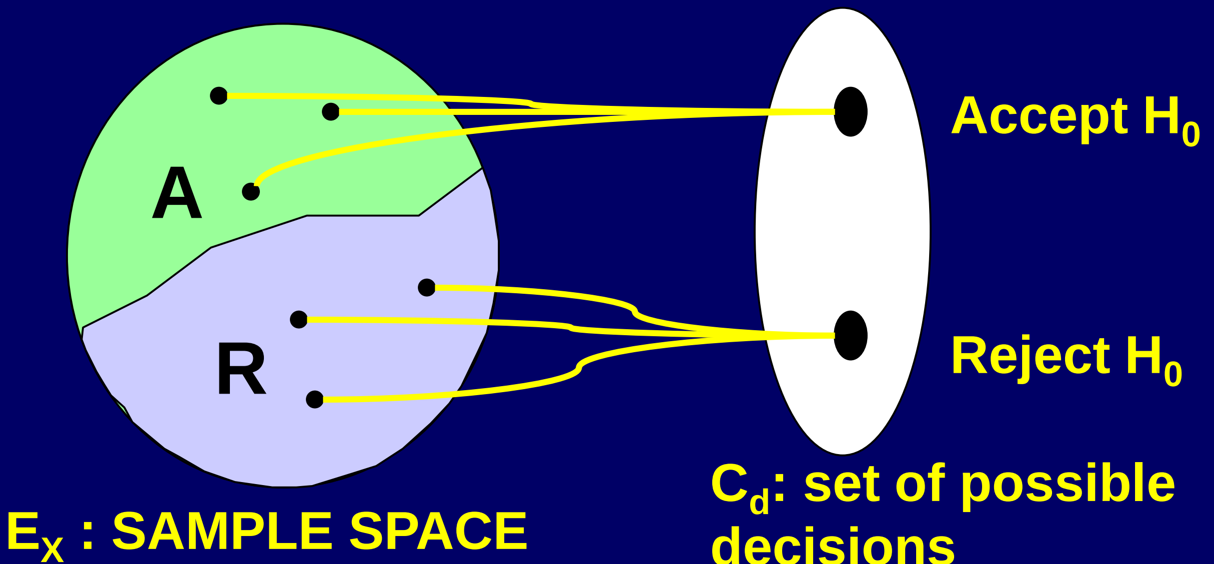
E.g. $\bar{x} = 92 \longrightarrow$ reject $H_0: m=100 \longrightarrow$ accept $H_0: m \leq 100$

CONSTRUCTION OF HYPOTHESIS TESTS

One statistical hypothesis tests implies establishing one partition of the sample space E_x (i.e. the set of all samples than can be obtained) in two regions:

- Region R of rejection: if the sample $(x_1, x_2, \dots, x_n) \in R$, the null hypothesis H_0 is rejected.
- Region A of acceptance: if the sample $(x_1, x_2, \dots, x_n) \in A$ the null hypothesis H_0 is accepted.

being A the complementary region of R in the sample space E_X



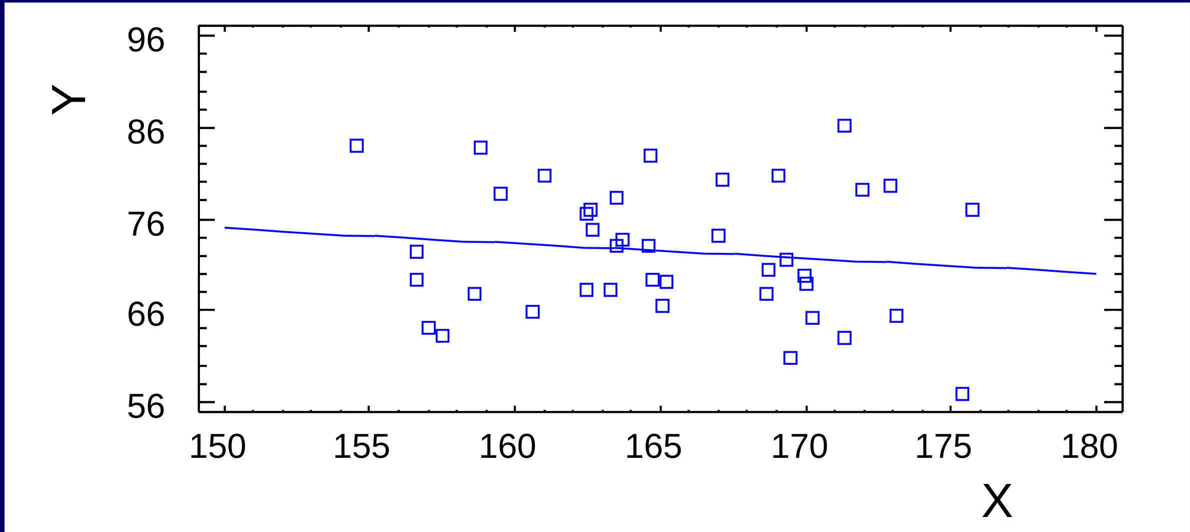
One hypothesis test is determined by:

- Sample size — E.g.: $n = 5$
- Sample type — E.g.: simple random sampling
- Statistical parameter θ^* used — E.g.: $\bar{x}$
- Region of acceptance / rejection

E.g.: if $\bar{x}$ € [2.8, 3.2] $\longrightarrow$ accept H_0: m=3

How would we test the following hypothesis?

Is there a relationship X - Y or are they independent ?



$Y = a + b \cdot X$

H_0: $b = 0$

H_1: $b \neq 0$

TYPE I, TYPE II ERRORS

When conducting a hypothesis TEST, there are two possible erroneous decisions that can be made, called:

- Type I error: To reject H_0 when it is true (error of the 1^{st} kind, α error, false positive).
- Type II error: To accept H_0 when it is false (being true H_1) (error of the 2^{nd} kind, β error, false negative).

Definitions:

- Type I risk (α): probability to make a type I error *(significance level of the test)*
- Type II risk (β): probability to make a type II error.

$1-\alpha$ = confidence level

$1-\beta$ = power of the test

Observed significance level: p-value (probability of having obtained a computed statistical parameter more unfavorable, being true H_0)

EXERCISE

Random var. X: No. of defects in one piece $X \sim Ps(\lambda)$

λ = Average number of defects in one piece

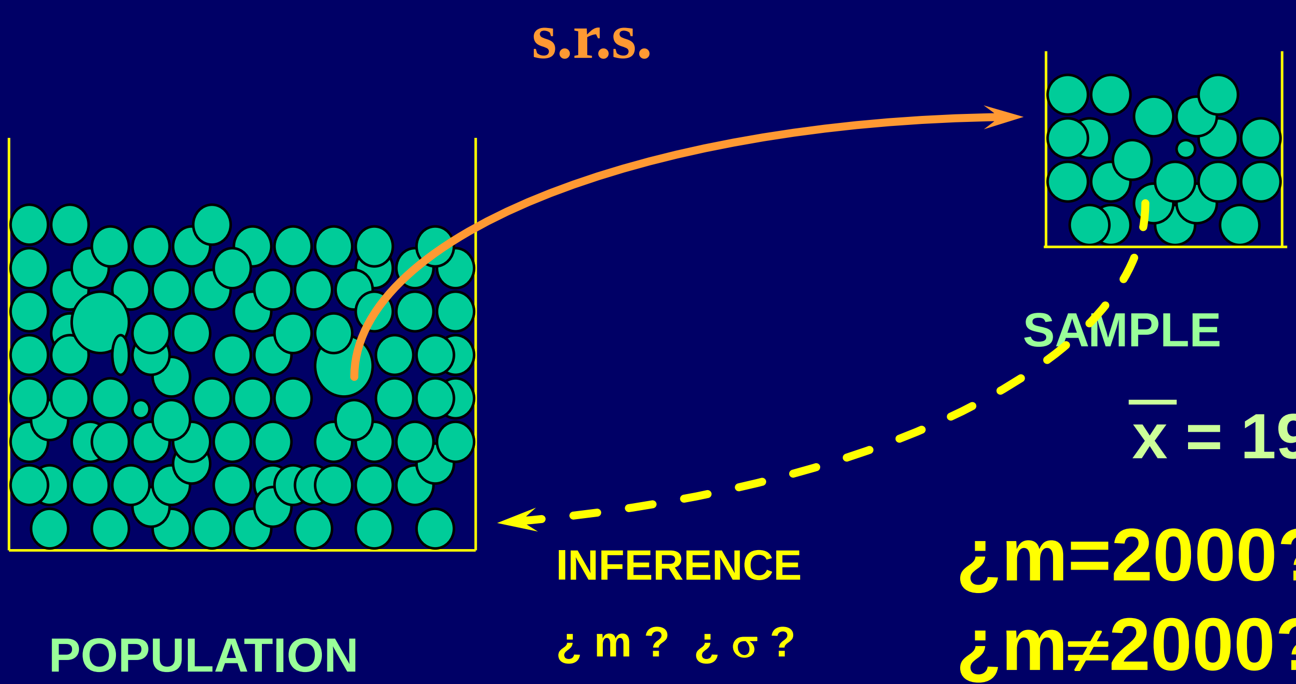
From a sample of size 10, we want to test the null hypothesis that the parameter λ of a Poisson distribution is 2 versus the alternative that is > 2. We will accept H_0 if the sample mean is ≤ 2.5 and will reject H_0 otherwise.

A) What is the type I risk of this test ?

B) What is the type II risk if λ actually is 3 ?

C) What is the type II risk if λ actually is 4 ?

INFERENCE ABOUT ONE NORMAL POPULATION



$\bar{x} = 1993.6$

¿m=2000? or
¿m≠2000?

EXAMPLE:

One machine that fills 2-liter soft drink bottles is adjusted to fill in average 2000 ml. In order to control its performance, a sample of 15 bottles is randomly taken, resulting the following data (ml filled):

1989 2015 1962 2013 1983 1989 1992 2011 1958
2023 1980 1977 1994 2017 2001

1) Estimate from the sample, the mean m and the standard deviation σ of the population under study.

2) Is there enough evidence to say that m differs significantly from 2000 and that, consequently, the machine should be readjusted?

3) Between what limits is comprised the value of m, with a reasonable confidence?

4) Between what limits is comprised the value of σ, with a reasonable confidence?

STEPS IN THE INFERENCE ABOUT A NORMAL POPULATION

1) Descriptive analysis of the sample (parameters of position and dispersion).

2) Normality of the data and detection of outliers (Histograms, Normal Probability Plot).

3) Hypothesis test: m=2000 (Student's t test).

4) Confidence interval for m (Student's t).

5) Confidence interval for σ (Chi^2).

6) Analysis with Statgraphics (Describe $\Rightarrow$ Numeric Data $\Rightarrow$ One-Variable Analysis).

1) DESCRIPTIVE ANALYSIS OF THE SAMPLE:

Variable:	Volume
Sample size	15
Average	**1993.6**
Median	1992
Mode	1989
Geometric mean	1993.51
Variance	391.971
Standard deviation	**19.7983**
Standard error	5.11189
Minimum	1958
Maximum	2023
Range	65
Lower quartile	1980
Upper quartile	2013
Interquartile range	33
Skewness	-0.256502
Standardized skewness	**-0.405564** $\in (-2,2) \Rightarrow$ CA=0
Kurtosis (CC-3)	-0.750953
Standardized kurtosis	**-0.593681** $\in (-2,2) \Rightarrow$ CC=3

Since $\overline{x}$ = 1993.6 which is different from 2000, should we readjust the machine?

NOT NECESSARILY !

The difference between $\overline{x}$ and 2000 can be by chance (due to the random sampling)

Actually, $\overline{x}$ will never be exactly 2000

2) NORMALITY OF DATA:

Most techniques of statistical inference for continuous variables assume that sampled populations are Normal.

How can we check if this previous hypothesis is acceptable in our case?

3 ways:

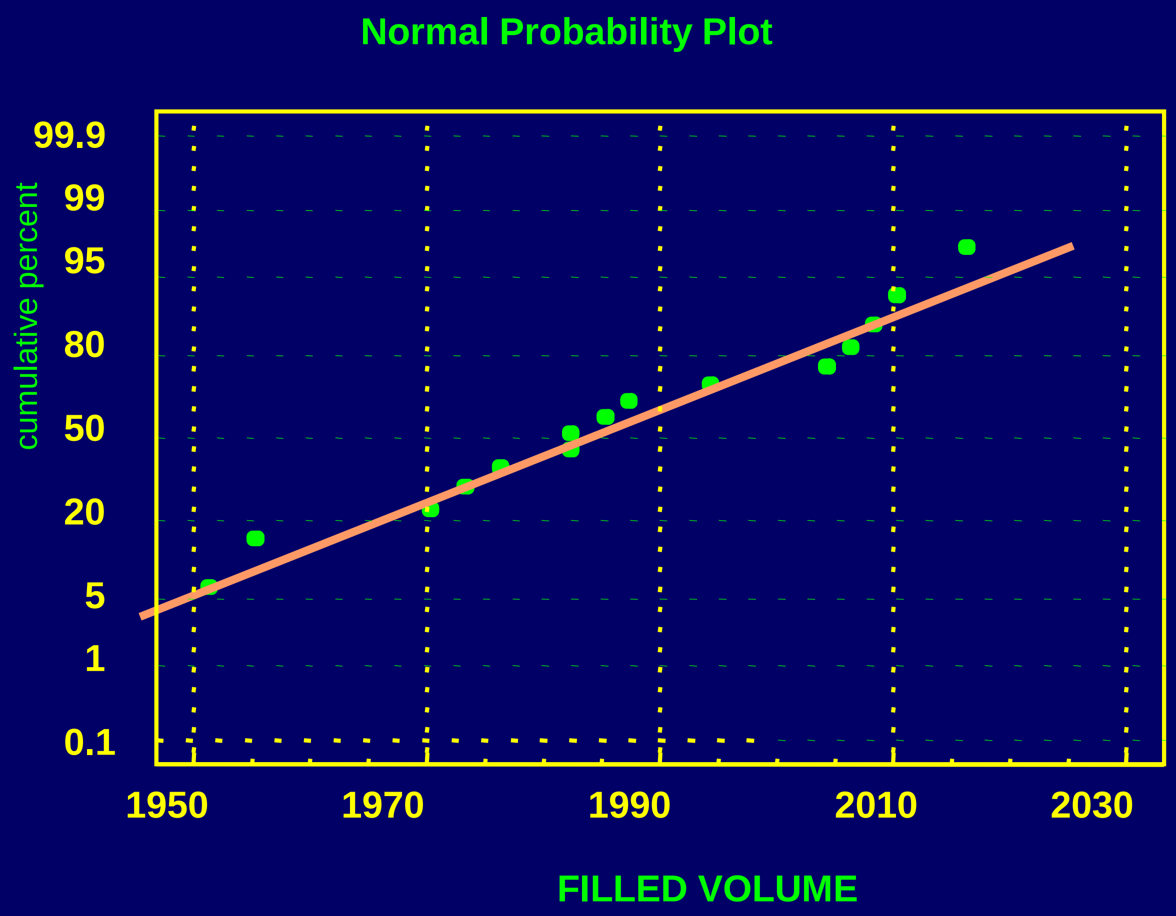
- To use formal statistical tests (they require many data in general. Not very useful in practice).

- To plot a Histogram (requires at least 40-50 data).

- To plot data on a Normal Probability Plot.

It is also convenient to check the values of the skewness and kurtosis coefficients.

CAN THE DATA BE REGARDED AS NORMAL?



NO OUTLIERS ARE OBSERVED

3) HYPOTHESIS TEST m=2000:

It is called Null Hypothesis because it reflects the previous knowledge of the situation (the machine should fill in average 2000 ml)

Intuitive reasoning: if m=2000 (H_0 true), $\bar{x}$ will be "similar" to 2000 and, hence, ($\bar{x}$ - 2000) will be similar to zero. Consequently:

- **If ($\bar{x}$ - 2000) "is similar" to zero, H_0 will be accepted (and the machine will not be readjusted).**

- **If ($\bar{x}$ - 2000) is "quite different" from zero, H_0 will be rejected: it will be admitted that m differs from 2000 (and the machine will be readjusted).**

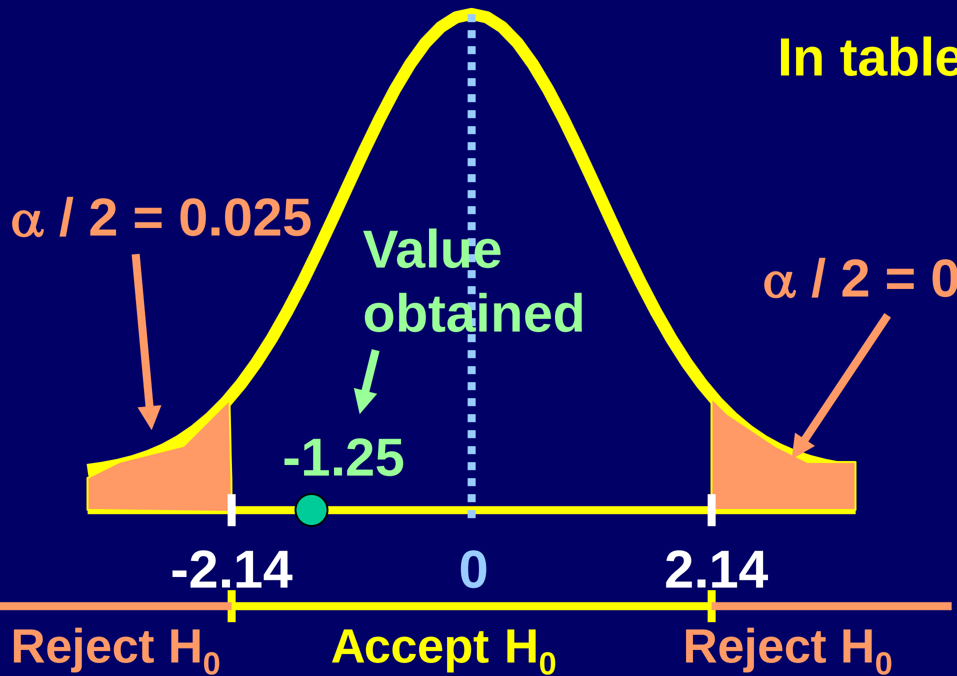
But … what should we consider as "being similar" ?

The "distance" in statistics has to be measured taking into account the variability:

$$\frac{\overline{X} - 2000}{s_{\overline{X}}} = \frac{\overline{X} - 2000}{s/\sqrt{N}}$$

$H_0 : m = 2000$ $\qquad$ $H_1 : m \neq 2000$

We know that $\dfrac{\bar{X} - m}{s/\sqrt{N}} \sim t_{N-1}$ $\quad$ If H_0 is true (m=2000) $\Longrightarrow$ $\dfrac{\bar{X} - 2000}{s/\sqrt{N}} \sim t_{N-1}$



In tables we obtain: $P(|t_{14}| > 2.14) = 0.05$

If H_0 is true: $\left|\dfrac{\bar{X} - 2000}{s/\sqrt{N}}\right| \leq 2.14$

If H_0 is false: $\left|\dfrac{\bar{X} - 2000}{s/\sqrt{N}}\right| > 2.14$

Thus, if: $\left|\dfrac{\bar{X} - 2000}{s/\sqrt{N}}\right|$ $\quad$ > 2.14 Reject H_0

≤ 2.14 Accept H_0

$$\frac{1993.6 - 2000}{19.8/\sqrt{15}} = -1.25$$

Conclusion: the hypothesis m=2000 is acceptable !

(there is not enough evidence to reject H_0)

SUMMARY OF THE TEST:

$H_0 : m = m_0$

$H_1 : m \neq m_0$

If $\left| \dfrac{\overline{X} - m_0}{s / \sqrt{N}} \right| > t_{N-1}(\alpha/2)$ critical value → **Reject H_0**

If $\left| \dfrac{\overline{X} - m_0}{s / \sqrt{N}} \right| \leq t_{N-1}(\alpha/2)$ → **Accept H_0**

Being $t_{N-1}^{\alpha/2}$ a value in Student's t table so that:

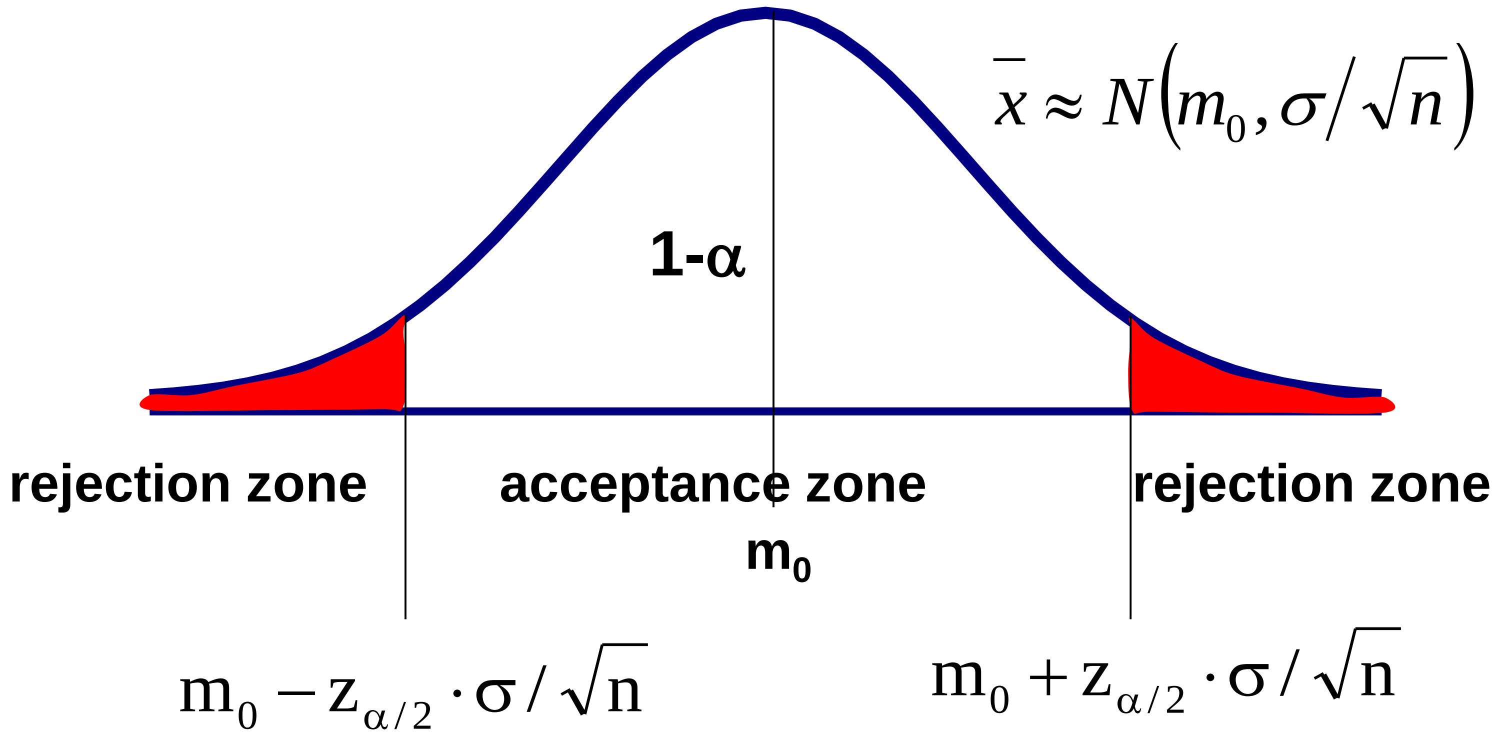
$$P\left(\left| t_{N-1} \right| > t_{N-1}^{\alpha/2} \right) = \alpha$$

If σ is known: use the N(0; 1) table (last row in t table)

- If H_0 is accepted it does not imply that it is necessarily true, it is just that we don't have enough evidence to reject it.

If σ is known:

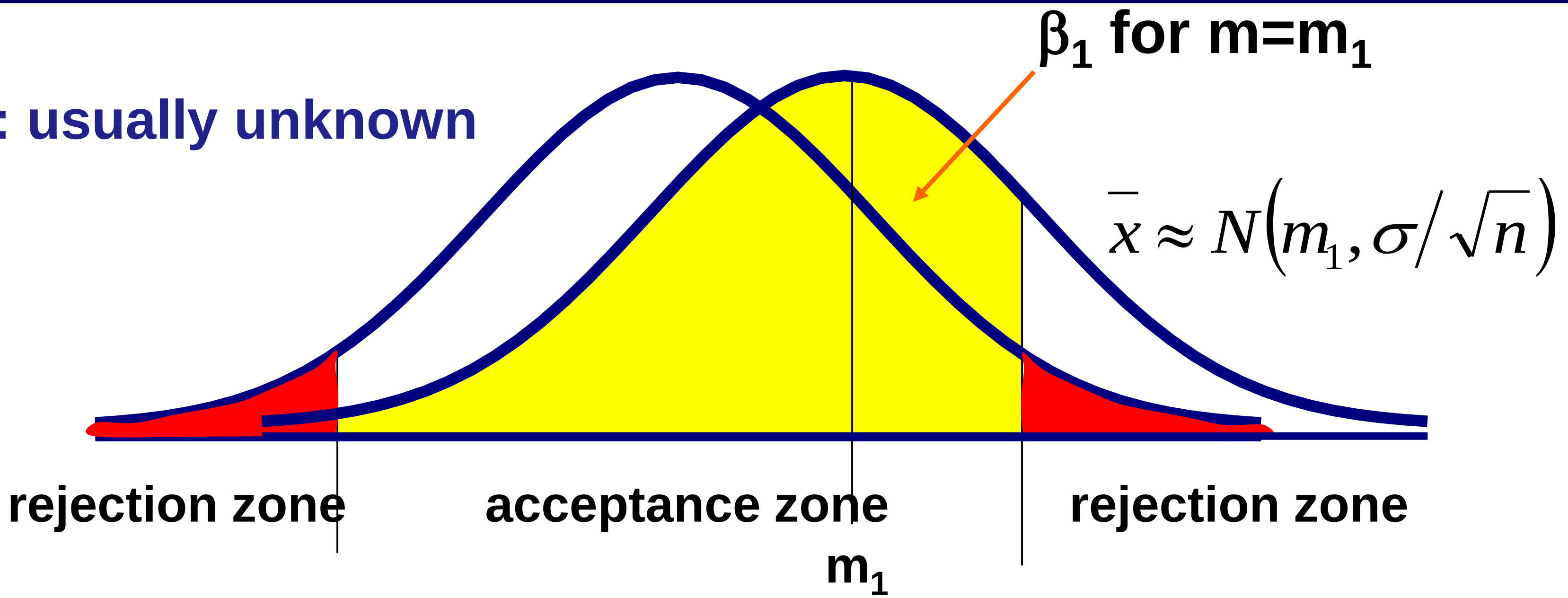
- **When H_0 is true, we select a region where it is quite likely (probability 1-α) to find that statistical parameter. This is the test acceptance region, and the complementary, the critical region.**



being $z_{\alpha/2}$ the critical value of N(0;1)

Although in fact $m \neq m_0$ we will still accept H_0 because the sample mean falls in the acceptance region with a probability β_1 (probability of type II error for $m = m_1 \neq m_0$)

β: usually unknown



α should be low, but if α decreases, β increases (and vice versa)

In order to decrease α and β we should increase the sample size (having more information about the population allows us to reduce the probability of choosing the wrong decision).

α : usually set at 0.05 or 0.01 (never, α > 0.1)

4) CONFIDENCE INTERVAL FOR m:

Is it possible, from the sample, to calculate an interval containing with a high probability (1-α) the unknown value m of the population mean?

Since: $\dfrac{\bar{X}-m}{s/\sqrt{N}} \sim t_{N-1}$

$$P(-t_{N-1}(\alpha/2) < t_{N-1} < +t_{N-1}(\alpha/2)) = 1-\alpha$$

$$P\left(-t_{N-1}(\alpha/2) < \frac{\bar{X}-m}{s/\sqrt{N}} < +t_{N-1}(\alpha/2)\right) = 1-\alpha$$

Therefore:

$$P\left(\underbrace{\bar{X} - t_{N-1}(\alpha/2)\frac{s}{\sqrt{N}}} < m < \underbrace{\bar{X} + t_{N-1}(\alpha/2)\frac{s}{\sqrt{N}}}\right) = 1-\alpha$$

This confidence interval has a probability (1-α) of containing m.

1-α: confidence level

If σ is known: use $z_{\alpha/2}$ instead of $t_{\alpha/2}$

EXAMPLE:

$$1993.6 \pm 2.14\frac{19.8}{\sqrt{15}}$$

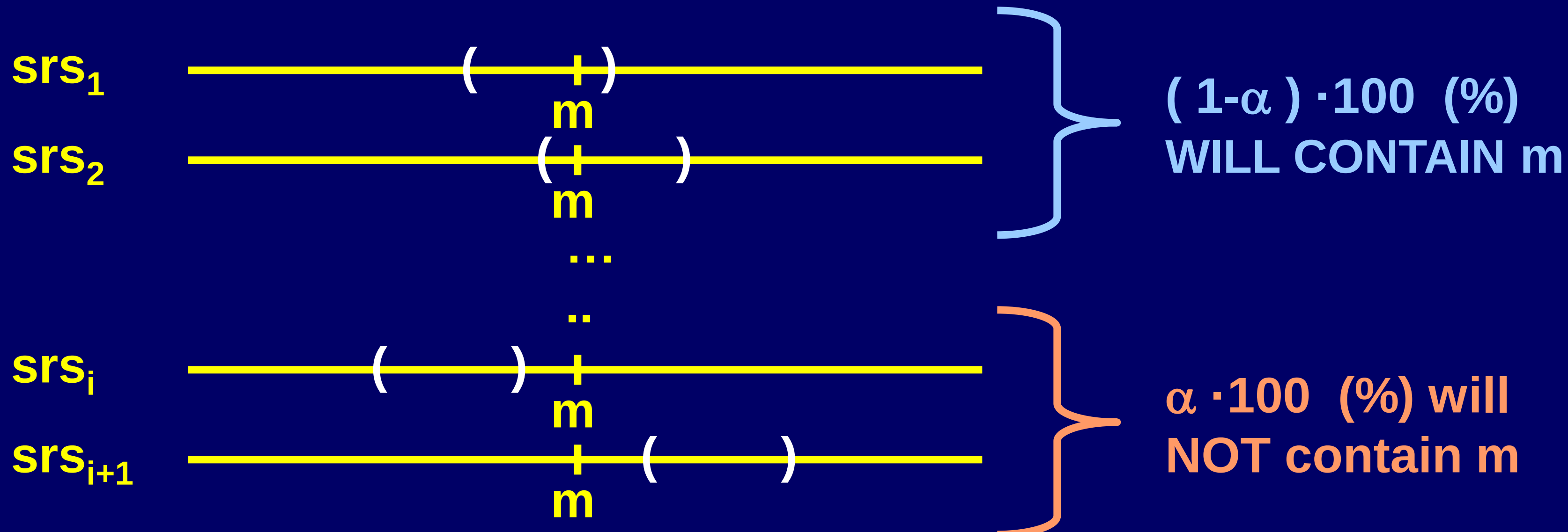
1982.7

2004.5

CONFIDENCE INTERVAL FOR m (95%) (1982.7 , 2004.5)

QUESTION:

What practical interpretation has this probability 1-α associated to a certain confidence interval?



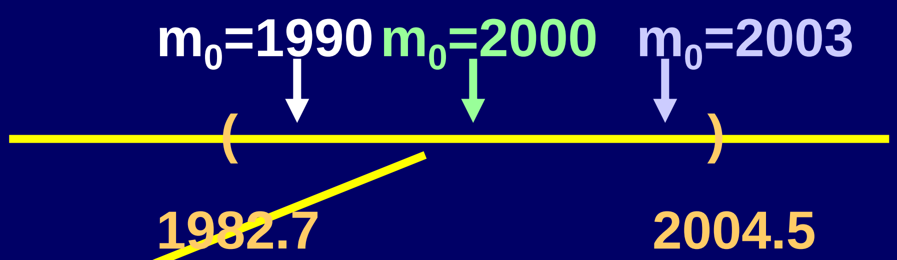
What kind of interval can we assume in this case for the computed interval (1982.7 , 2004.5)?

HYPOTHESIS TEST for m USING CONFIDENCE INTERVALS:

$H_0 : m = m_0$ $\qquad$ $H_1 : m \neq m_0$

If $m_0 \in$ Confidence Interval → Accept H_0

If $m_0 \notin$ C.I. → Reject H_0 → Accept H_1



Accept H_0 →

Is reasonable that m=2000 → Do not readjust the machine

Is reasonable that m=1990 → Readjust the machine because fills <2000

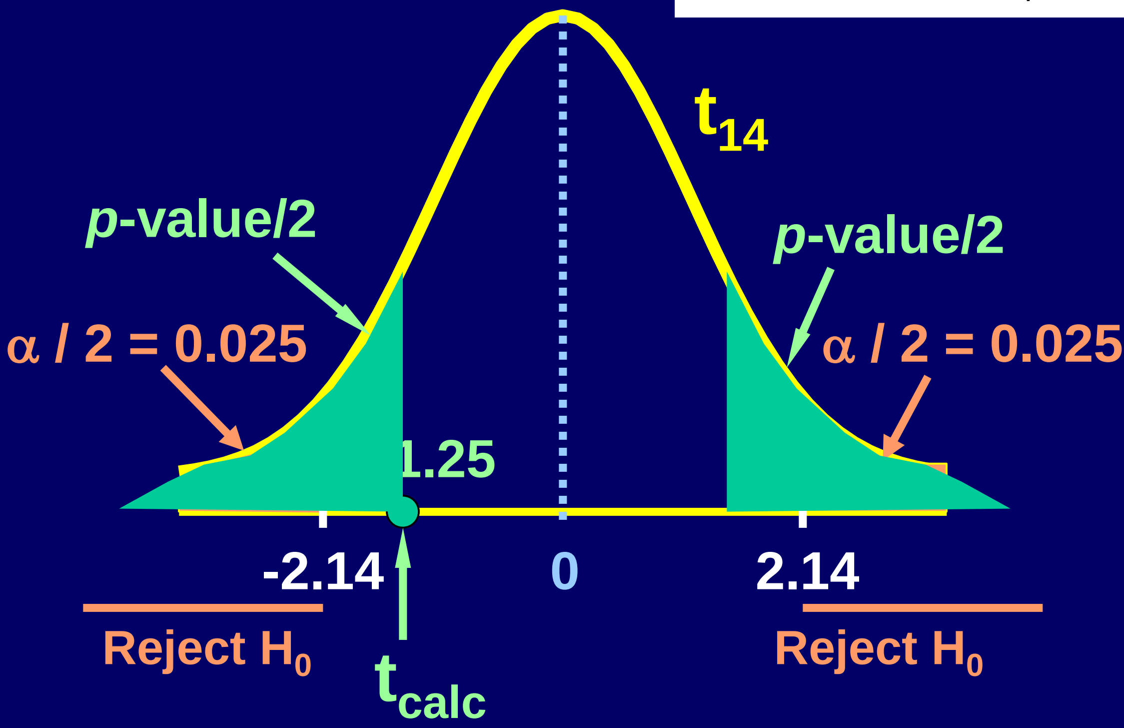
Is reasonable that m=2003 → Readjust the machine because fills >2000

Have we demonstrated that m=2000 ?

The confidence interval contains all null hypotheses consistent with the obtained sample

P-VALUE (OBSERVED SIGNIFICANCE LEVEL)

For this test: *p*-value: $p = P\left(t_{n-1} > \left|t_{calc}\right|\right)$



If *p*-value < α : reject H_0

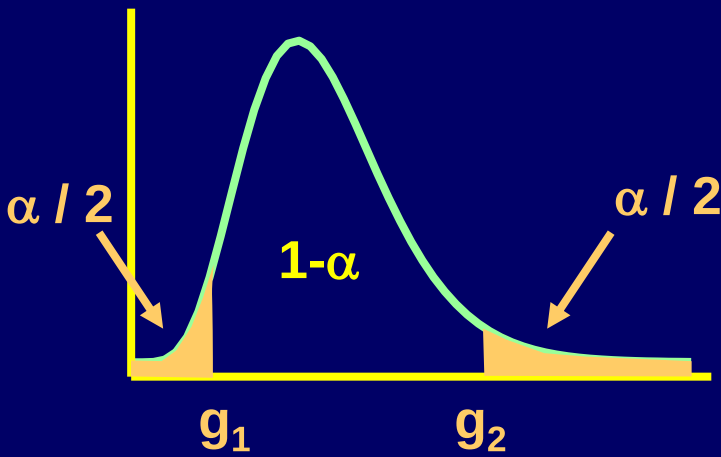
If *p*-value > α : accept H_0

For other tests, *p*-value is calculated differently but this rule is always true.

***p*-value:** probability of having obtained a computed statistical parameter more unfavorable, being true H_0

5) CONFIDENCE INTERVAL FOR σ^2:

We know that: $(N-1)\dfrac{s^2}{\sigma^2} \sim \chi^2_{N-1}$ $\qquad (n-1)\dfrac{s^2}{\sigma^2} \in [g_1,\ g_2]$



In the χ^2 table it is possible to find two values g_1, g_2 so that:

$$P(g_1 < \chi^2_{N-1} < g_2) = 1-\alpha \qquad (1)$$

For example: $P(5.63 < \chi^2_{14} < 26.1) = 1-0.05 = 0.95$

From (1) we obtain:

$$P\left(\frac{(N-1)\cdot s^2}{g_2} < \sigma^2 < \frac{(N-1)\cdot s^2}{g_1}\right) = 1-\alpha$$

Therefore:

$$\sigma^2 \in \left[\frac{(N-1)\cdot s^2}{g_2}, \frac{(N-1)\cdot s^2}{g_1}\right] \qquad \sigma \in \left[\sqrt{\frac{(N-1)\cdot s^2}{g_2}}, \sqrt{\frac{(N-1)\cdot s^2}{g_1}}\right]$$

In this example:

$$\sqrt{\frac{14\cdot 392}{5.63}} = 31.2 \qquad \sqrt{\frac{14\cdot 392}{26.1}} = 14.5$$

[14.5 , 31.2] is a confidence interval for σ

6) ANALYSIS WITH STATGRAPHICS

OPTION: “ONE-VARIABLE ANALYSIS”

```
Hypothesis Tests for VOLUME

Sample mean = 1993.6   Sample median = 1992.0

t-test

Null hypothesis: mean = 2000
Alternative: not equal

Computed t statistic = -1.25198
P-Value = 0.231089

Do not reject the null hypothesis for alpha =0.05
```

```
Confidence Intervals for VOLUME

95% confidence interval for mean:
1993.6 +/- 10.9639   [1982.64; 2004.56]

95% confidence interval for standard deviation:
[14.4948; 31.2238]
```

COMPARISON OF 2 NORMAL POPULATIONS

Two computer programs (A, B) are available to search files in a hard disk. In order to determine which one works faster, 10 files are searched with each program, and the time required to find the each file is recorded.

OBJECT OF THE STUDY, to compare two populations:

- Files to be searched by program A
- Files to be searched by program B

20 trials are conducted: → 10 with program A; → 10 with program B

What is measured in each experimental trial?

RESULTS:

											$\bar{x}$	s
prog. A.	3.4	3.7	2.9	2.5	1.6	2.8	3.7	5.9	4.8	4.3	3.56	1.23
prog. B	2.7	3.2	1.8	1.9	1.1	2.2	2.8	4.8	4.3	3.4	2.82	1.15

HOW SHOULD THESE DATA BE STATISTICALLY ANALYZED ?

FUNDAMENTALS OF THE STATISTICAL ANALYSIS:

2 Populations studied: files in the hard disk that can be searched by program A or B.

Random variable: time required to search a file.

It is assumed that the variable is normally distributed:

m_1 σ_1 m_2 σ_2

Sampling: $X_1, X_2, \ldots, X_{10}$ $X_1, X_2, \ldots, X_{10}$

Statistical parameters calculated from the samples: $\boxed{\bar{X}_1}$ $\boxed{S_1}$ $\boxed{\bar{X}_2}$ $\boxed{S_2}$

$¿\ m_1 \overset{>}{\underset{<}{=}} m_2\ ?$

$¿\ \sigma_1 \overset{>}{\underset{<}{=}} \sigma_2\ ?$

COMPARISON OF VARIANCES

$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$

If the null hypothesis is true:

- s_1^2 will be "similar" to s_2^2
- The ratio s_1^2 / s_2^2 will be similar to 1. The null hypothesis will be rejected if this ratio is clearly different to 1.

But... what should be considered as being "similar" or not?

If H_0 is true: $s_1^2 / s_2^2 \sim F_{N_1-1,\,N_2-1}$

STEPS: 1) If $s_1 > s_2$: divide s^2_1 / s^2_2. If $s_2 > s_1$: divide s^2_2 / s^2_1

2) Test if the obtained ratio is "too high" to be a $F_{n1-1,n2-1}$

ALTERNATIVE PROCEDURE:

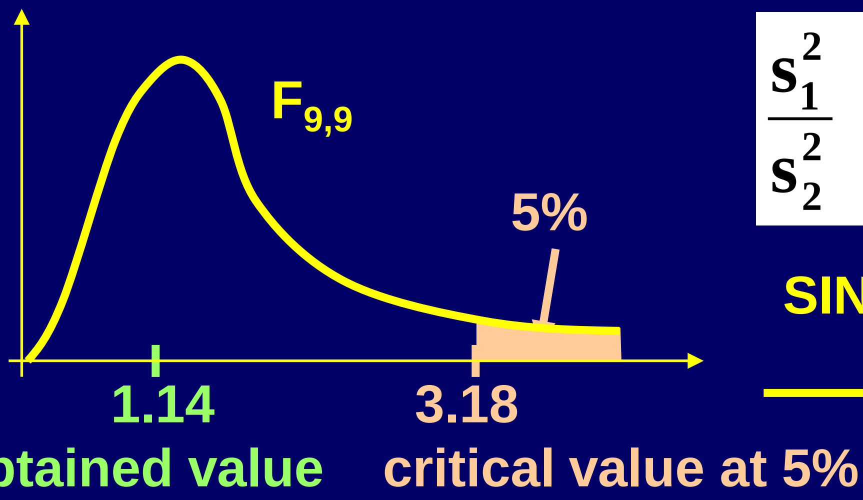
1) Obtain a confidence interval for σ^2_1 / σ^2_2 (see formula)

2) if 1 belongs to this interval: accept H_0

EXAMPLE:

$P\,(F_{9,9} > f) = 0.05 \longrightarrow f = 3.18$

IF THE NULL HYPOTHESIS IS TRUE: $\sigma_1^2 = \sigma_2^2$



$$\frac{s_1^2}{s_2^2} \sim F_{N_1-1,\,N_2-1}$$

$$\frac{1.23^2}{1.15^2} = 1.14$$

SINCE $F_{9,9}(5\%) = 3.18 > 1.14$

$\longrightarrow$ THE NULL HYPOTHESIS IS ACCEPTED

THERE IS NOT ENOUGH EVIDENCE TO AFFIRM THAT THE VARIANCE OF TIME TO SEARCH A FILE WITH PROGRAMS A or B IS DIFFERENT.

If $\sigma_1^2 \neq \sigma_2^2$ the subsequent test for mean comparison is approximate, though it is quite “robust” if the number of observations in both samples is similar.

COMPARISON OF MEANS

$H_0 : m_1 = m_2$

$H_1 : m_1 \neq m_2$

If H_0 is true:

- $\bar{X}_1$ will be "similar" to $\bar{X}_2$
- $\bar{X}_1 - \bar{X}_2$ will be "similar" to zero.

What should be considered as "being similar"?

We know that: $\dfrac{x_1 - x_2 - (m_1 - m_2)}{S_{(x_1 - x_2)}} \sim t_{N_1+N_2-2}$ (considering that $\sigma_1^2 = \sigma_2^2$)

If $m_1 = m_2$: $\dfrac{x_1 - x_2}{S_{(x_1 - x_2)}} \sim t_{N_1+N_2-2}$

$$S_{(\bar{x}_1-\bar{x}_2)} = S \cdot \sqrt{\frac{1}{N_1} + \frac{1}{N_2}} = \sqrt{\frac{(N_1-1)\cdot s_1^2 + (N_2-1)\cdot s_2^2}{N_1+N_2-2}} \cdot \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

If $N_1 = N_2$:

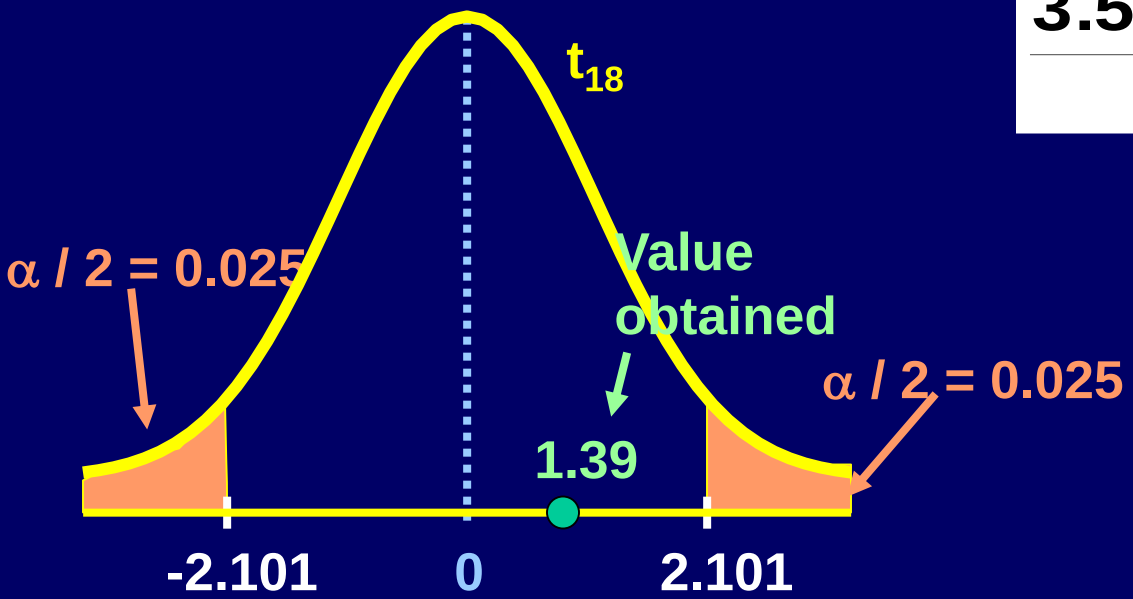
$$S_{(\bar{x}_1-\bar{x}_2)} = \sqrt{\frac{s_1^2 + s_2^2}{2}} \cdot \sqrt{\frac{2}{N}}$$

EXAMPLE:

$$\bar{x}_1 - \bar{x}_2 = 3.56 - 2.82 = 0.74$$

$$S_{(x_1 - x_2)} = \sqrt{\frac{1.23^2 + 1.15^2}{2}} \cdot \sqrt{\frac{2}{10}} = 0.53$$

$$\frac{3.56 - 2.82}{0.53} = 1.39$$



And since $t_{18}(5\%) = 2.101 > 1.39$

RESULTS ARE CONSISTENT WITH THE HYPOTHESIS $m_1 = m_2$

CONFIDENCE INTERVAL FOR m_1-m_2

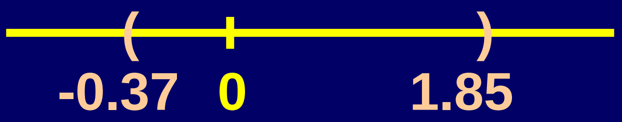
Alternative equivalent way (though more informative) of analyzing the results of this experiment:

Interval for m_1-m_2 with a confidence level $(1-\alpha)$ x 100 :

$$m_1 - m_2 \in \left[(\bar{x}_1 - \bar{x}_2) \pm t^{\alpha/2}_{N_1+N_2-2} \, S_{(\bar{x}_1 - \bar{x}_2)} \right]$$

In the example: $(3.56 - 2.82) \pm 2.101 \cdot 0.53 = [-0.37, 1.85]$

being $2.101 = t_{18}(2.5\%)$ from the t-table.



$$0 \in [-0.37, 1.85] \Rightarrow m_1 - m_2 = 0 \Rightarrow m_1 = m_2$$

We can affirm with quite confidence (95% of confidence) that the difference m_1-m_2 is comprised between -0.37 and 1.85

ANALYSIS OF RESIDUALS

General definition:

Residual= value observed - value estimated by a model

Residual is the part of the observed value due to the variability caused by factors not controlled in the experiment.

residual = value observed — value estimated

AVERAGE

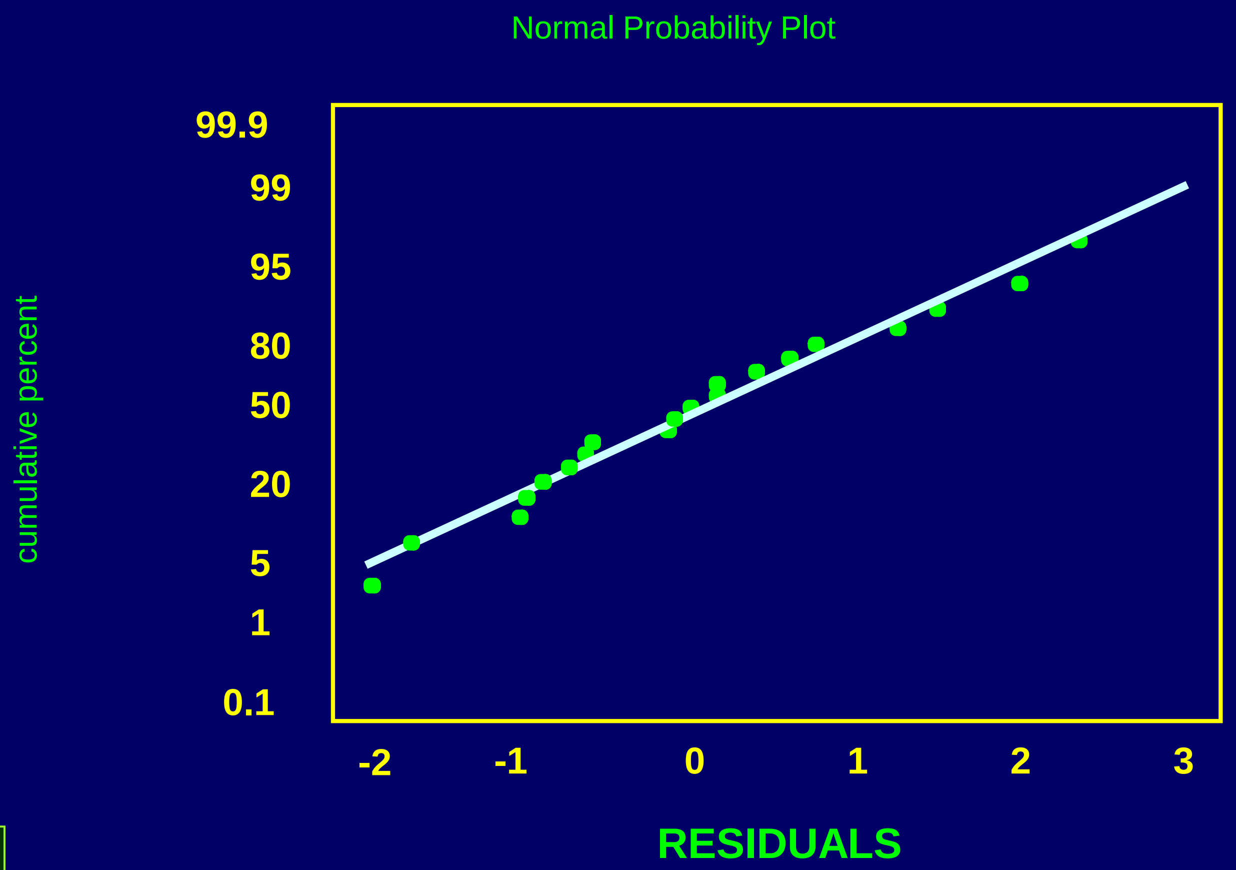
EXAMPLE: First observation from program A:

residual = 3.4 — 3.56 = -0.16

UTILITY OF RESIDUALS ANALYSIS (GRAPHICAL METHODS)

NORMAL PROBABILITY PLOT OF RESIDUALS :

* CHECK FOR NORMALITY
* DETECTION OF OUTLIERS



The average of all residuals is zero

The average is statistically different from 100

The difference between both sample means is statistically significant

There is enough evidence to say that...

- the population average is not 100.

- the population means are different.

Differences statistically significant ≠ differences important

Doing N high enough, we can detect as significant ANY difference of means, though in practice they might be irrelevant.

Actually, if $n \to \infty$ we are comparing the whole populations.

ANALYSIS WITH STATGRAPHICS:

COMPARE => 2 SAMPLES => TWO-SAMPLE COMPARISON

```
Comparison of Means

95,0% confidence interval for mean of time_A: 3,56 +/- 0,8795
95,0% confidence interval for mean of time_B: 2,82 +/- 0,82036
95,0% confidence interval for the difference between the means
   assuming equal variances: 0,74 +/- 1,117   [-0,377,1,857]

t test to compare means:
   Null hyp.: mean1 = mean2  Alt. hypothesis: mean1 NE mean2
   assuming equal variances: t = 1,39186   P-value = 0,180927
```

```
Comparison of Standard Deviations

Variance time_A: 1,51156  Variance time_B: 1,31511
Ratio of Variances = 1,14937

95,0% Confidence Intervals
     Ratio of Variances: [0,285488; 4,62738]

F-test to Compare Standard Deviations:

   H0: sigma1 = sigma2   Alt. hypothesis: sigma1 NE sigma2
   F = 1,14937   P-value = 0,839105
```

Conclusion of the test: accept $m_A = m_B$ BUT...

For all trials: $time_A > time_B$

prog. A.	3.4	3.7	2.9	2.5	1.6	2.8	3.7	5.9	4.8	4.3
prog. B	2.7	3.2	1.8	1.9	1.1	2.2	2.8	4.8	4.3	3.4

Lowest value from A and B:

Highest value from A and B:

Is it a coincidence?

For some reason this file was more difficult to be found by both programs

In this case (one two-dimensional variable), better to apply another test

STATGRAPHICS:

Compare => 2 samples => Two-sample comparison

Compare => 2 samples => Paired-sample comparison

To compare the population mean of two characteristics measured in the same individuals, a paired-sample comparison is more powerful than a two-sample comparison.

With a paired-sample test: reject H_0: $m_A > m_B$ (makes sense!)