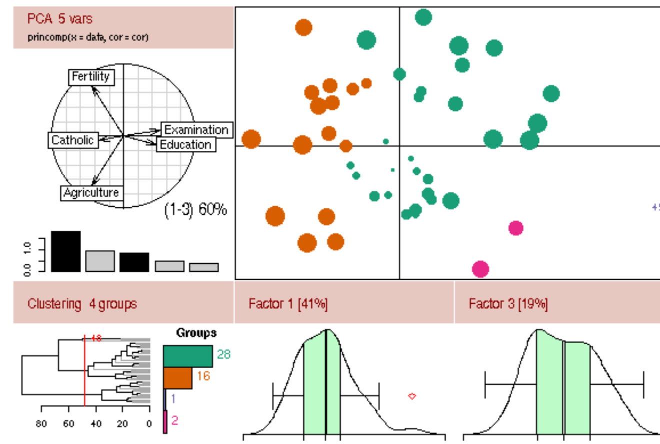
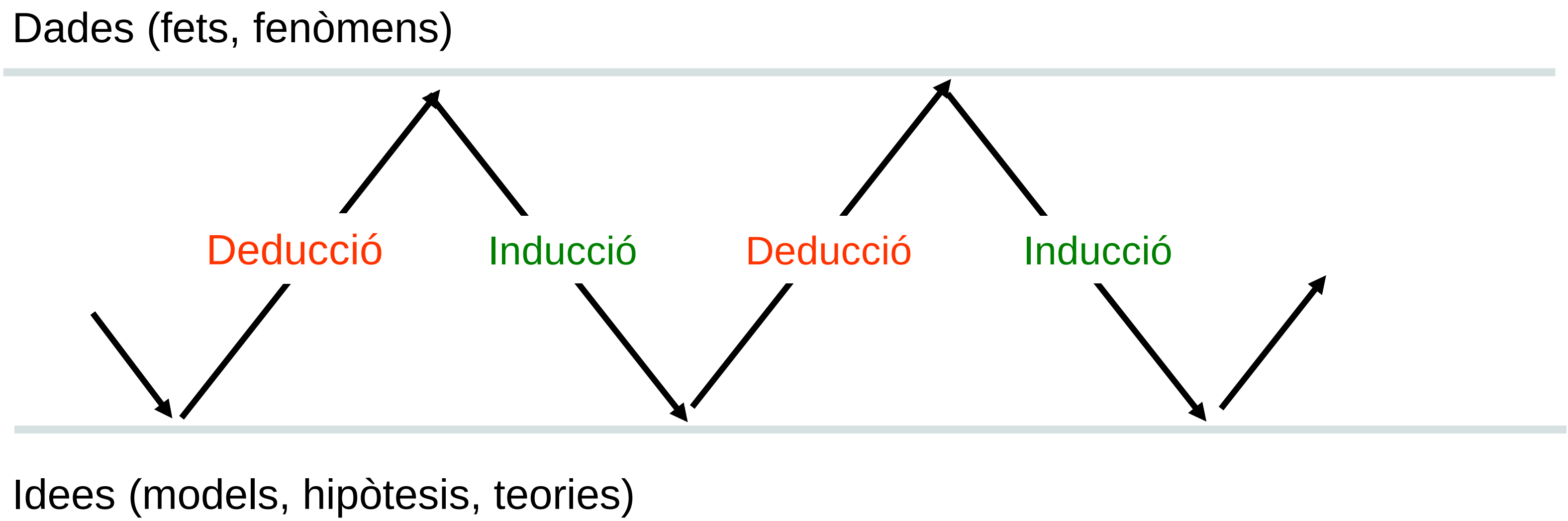
Introducció. La importància d'experimentar

$$y_{ij} = \mu + \alpha_i + e_{ij}$$



Inducció – Deducció



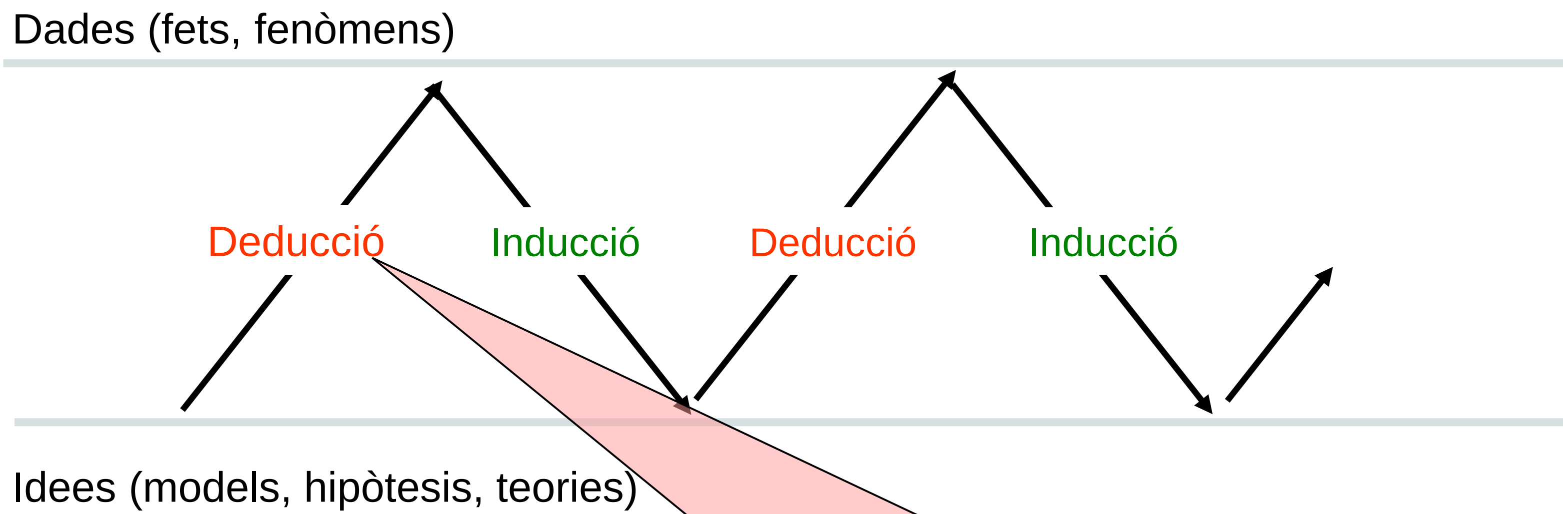
L'aprenentatge es basa en la iteració **deducció – inducció**.

Aquest és el fonament del mètode científic (que es coneix ja des de temps d'Aristòtil)

Plantegem un model, que porta unes conseqüències. Aquestes conseqüències es comparen amb el que passa en realitat (les dades).

Si les conseqüències que suposàvem no concorden amb les dades, haurem de modificar el nostre model.

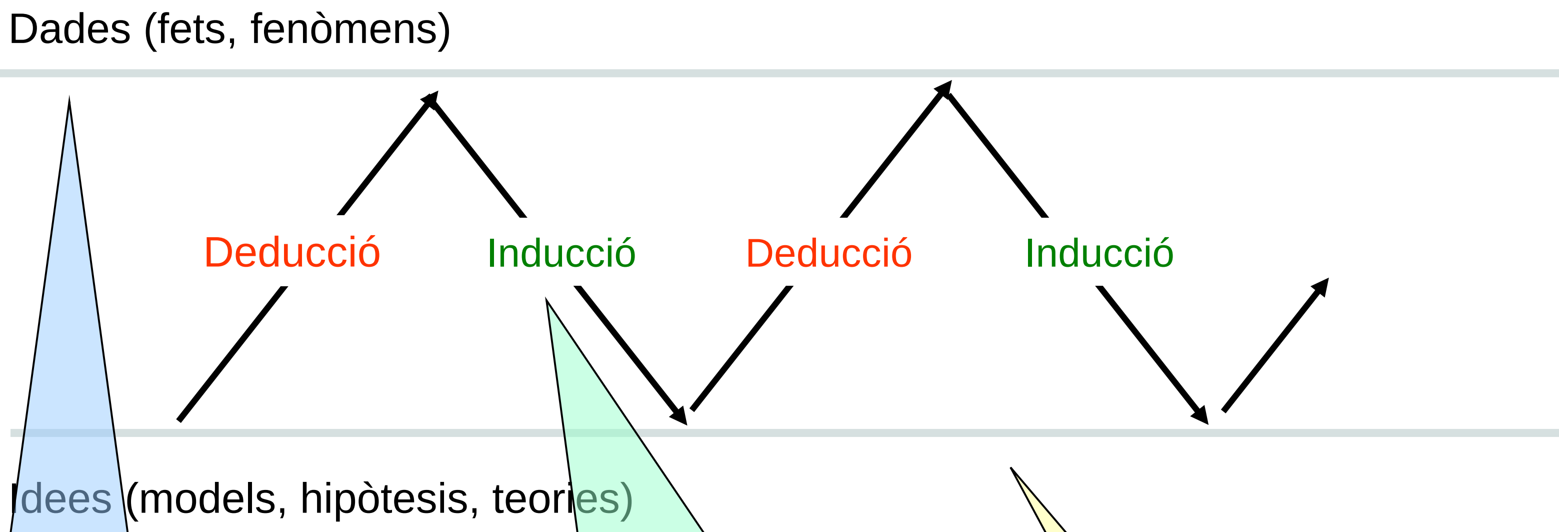
Inducció – Deducció: exemple (1)



MODEL:
La Clara ha descobert un nou catalitzador.
Segurament, la seva presència farà que el producte A es combini amb el B per formar el producte C, amb un rendiment molt elevat.

DEDUCCIÓ:
Encara que es mira a revistes i llibres, sembla que no hi ha cap referència de que aquesta reacció s'hagi provat abans.
Pels coneixements químics que té la Clara, prepara un experiment per aconseguir dades i veure si el catalitzador funciona.
Provarà de fer la reacció a una temperatura de 600ºC

Inducció – Deducció: exemple (2)



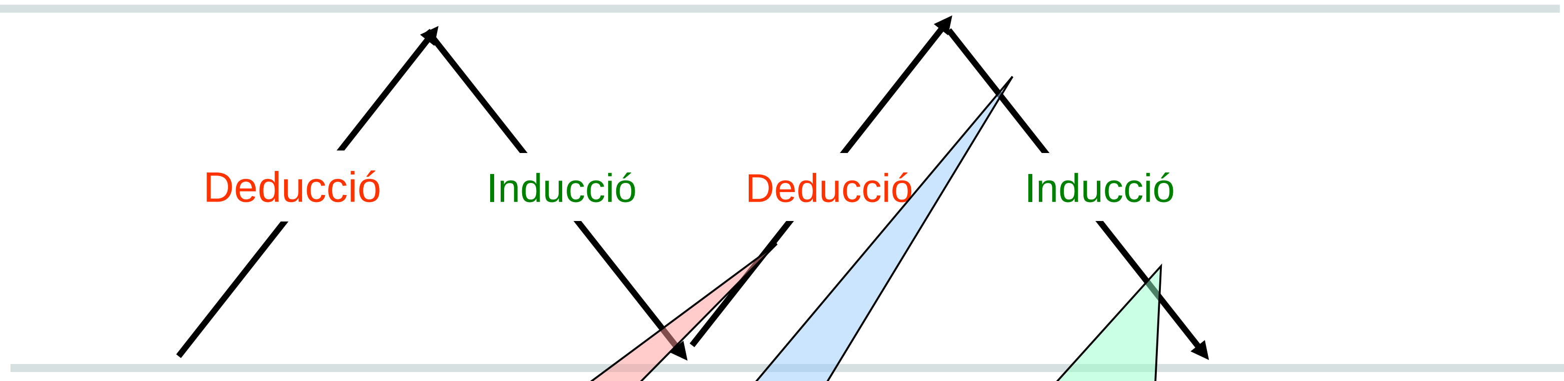
DADES:
El resultat de l'experiment és molt decebedor . El producte C hauria de ser incolor i inodor, però s'obté una mena de pasta negra amb un contingut inferior a l'1% en producte C.

INDUCCIÓ:
El model i les dades obtingudes no lliguen…
La Clara pensa que el que pot haver passat és que el producte C s'hagi obtingut amb un rendiment elevat, però que després s'hagi descomposat.

MODEL:
Segurament, les condicions del primer experiment eren massa extremes.

Inducció – Deducció: exemple (3)

Dades (fets, fenòmens)



Idees (models, hipòtesis, teories)

DEDUCCIÓ:
Potser una temperatura més baixa produeix C amb un rendiment millor. La Clara decideix provar de repetir l'experiment a $550^{O}C$ i a $500^{O}C$, aviam què passa

DADES:
El resultat de la reacció és ara millor: no és tan pastós ni tan negre. La reacció a $550^{O}C$ dóna un 4% de producte C, la reacció a $500^{O}C$ dóna un 17% de producte C.

INDUCCIÓ:
A partir d'aquests resultats, i fent servir els seus coneixements, la Clara decideix que caldrà fer més experiments baixant més encara la temperatura. També valdria la pena provar de canviar altres variables (la concentració dels reactius, el temps de reacció, la quantitat de catalitzador…)

El joc de les 20 preguntes (1)

Objectiu: endevinar la paraula que el jugador B té en una fitxa

Jugador A:

- Fa preguntes que es poden respondre amb un SÍ o un NO.
- Com a màxim en 20 preguntes ha d'endevinar l'objecte.

Jugador B:

- Respon les preguntes que li fa A amb un SÍ o un NO
- Va apuntant en un full la pregunta i la resposta.

El joc de les 20 preguntes (2)

Jugador A

1. És una persona?
2. És un objecte?
3. Serveix per transportar-se?
4. Hi cap dins de casa?
5. És natural (no creat per l'home)?
6. Es pot menjar?
7. És una fruita?
8. És una pera?
9. És un plàtan?
10. És una poma?

Jugador B

1. No
2. Sí
3. No
4. Sí
5. Sí
6. Sí
7. Sí
8. No
9. No
10. Sí

El jugador A ho ha endevinat amb només 10 preguntes… per què pot endevinar el jugador A la paraula que només sap el jugador B?

El joc de les 20 preguntes (3)

Fixeu-vos en l'analogia entre l'experimentació científica i el joc de les 20 preguntes! Per poder endevinar l'objecte cal:

Saber alguna cosa d'aquell objecte. Almenys, cal conèixer què vol dir la paraula.

Seguir una bona estratègia de preguntes per tal d'endevinar la paraula en 20 preguntes, com a màxim.

Igualment, quan experimentem ens cal:

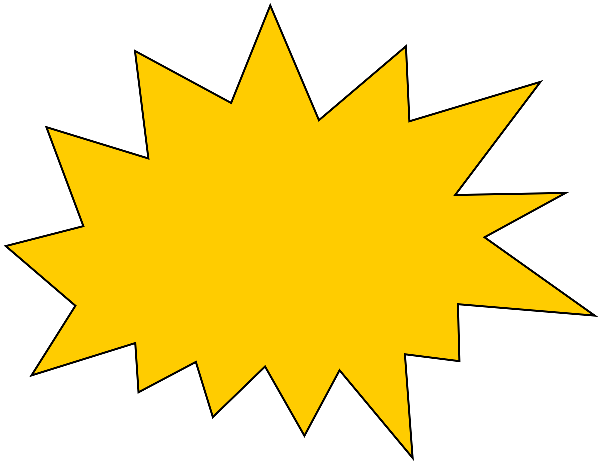
Tenir coneixements sobre el fenòmen amb el que treballem, saber el màxim possible d'aquell procés, consultar documentació tècnica sobre el procés, parlar amb experts...

Fer servir una bona estratègia d'experimentació, amb les eines estadístiques adequades. Així, descobrirem el màxim de coses amb el mínim esforç.

La col·laboració entre estadístics (experts en l'estratègia d'experimentació) i tècnics (experts en el procés amb què es treballa) és la fórmula guanyadora!

Paradigma del coneixement

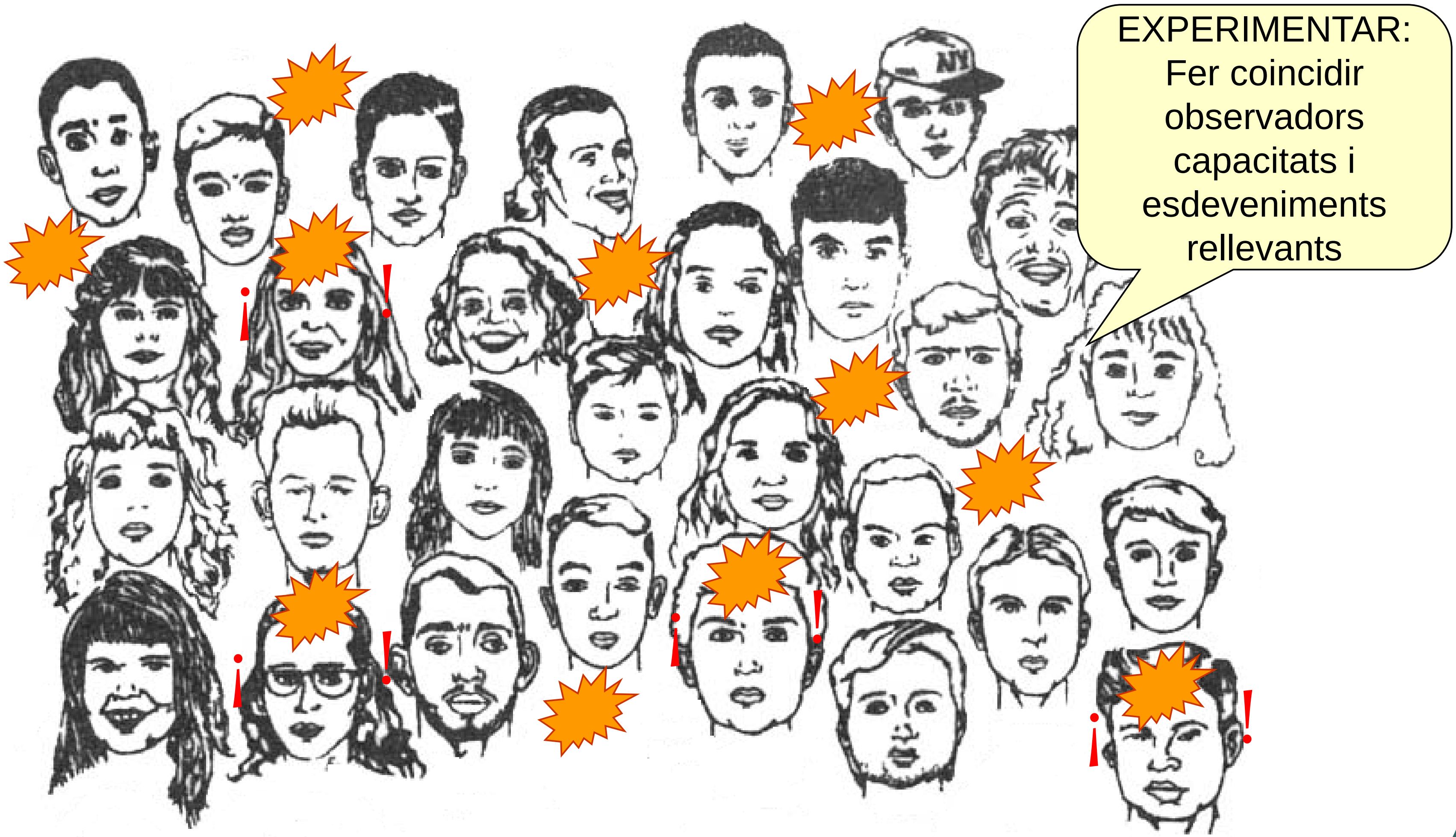
Esdeveniment rellevant



Per què avança el coneixement?

Observador capacitat

Experimentació dirigida



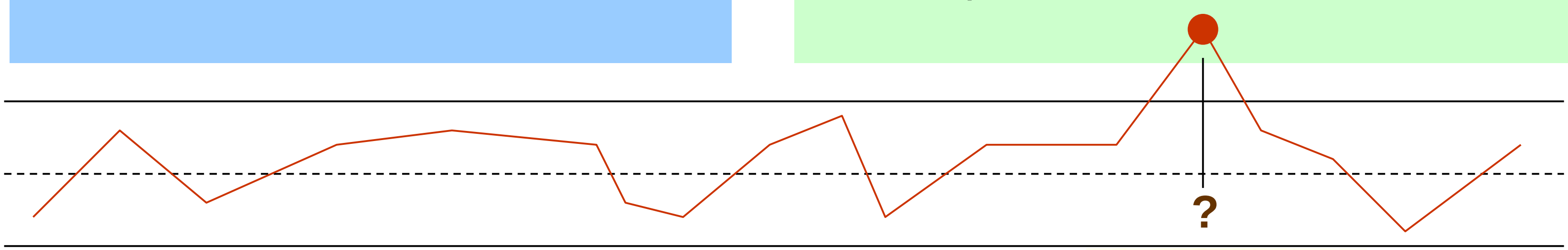
Accelerar l'adquisició de coneixement

Experimentació dirigida

Induir el fet que es succeeixin esdeveniments rellevants

Observació intel·ligent

Augmentar la probabilitat que els esdeveniments que succeeixen naturalment siguin observats per observadors capacitats



Anàlisi de dades existents

Dades inconsistents:

Envelliment

Reparacions

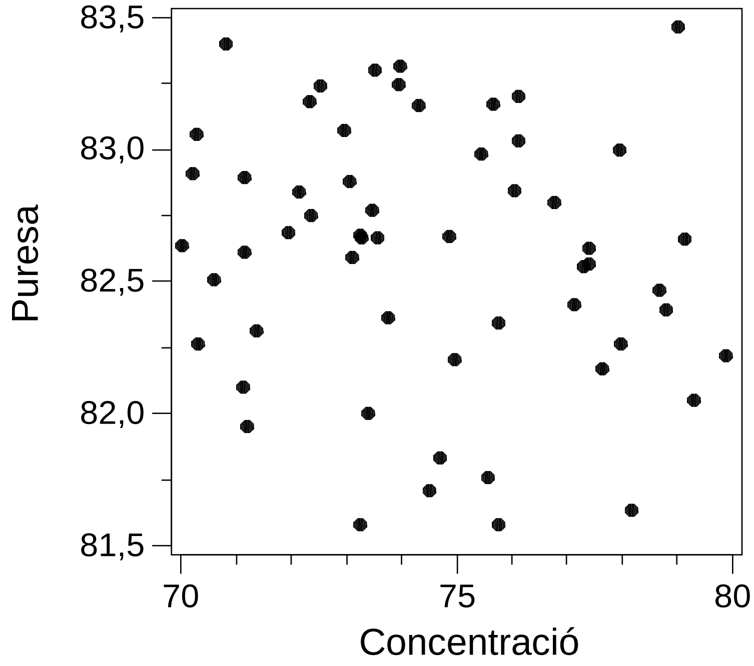
Canvis en procediments

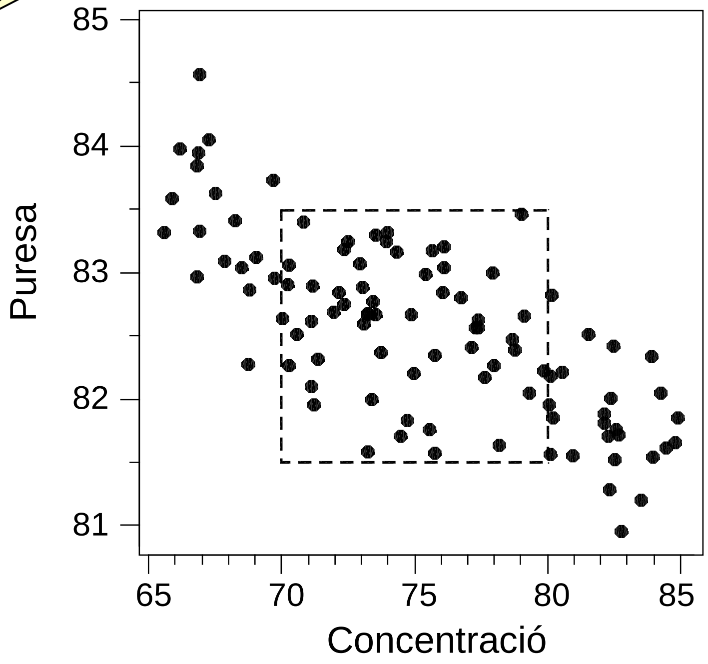
...

1	79,69	226	15,3	1,8
2	71,58	211	13,0	1,8
3	65,84	220	13,3	1,9
4	78,82	205	14,5	1,6
5	79,74	224	13,3	1,7
6	73,85	209	10,3	2,0
7	70,76	245	14,9	1,2
8	76,19	243	11,7	1,1
9	76,09	234	15,8	1,6
10	73,57	243	16,6	1,4
11	79,21	208	18,8	1,1
12	67,10	211	10,7	1,6
13	75,21	210	16,5	1,8
14	75,20	244	17,2	1,5
15	76,75	217	15,9	1,2
16	70,74	223	17,3	1,4
17	75,57	235	14,7	1,2
18	80,87	218	16,1	1,9
19	79,09	220	17,8	1,5
20	76,65	237	12,4	1,1
21	80,92	220	19,3	2,0
22	76,96	206	10,7	1,2
23	78,84	209	16,7	1,1
24	70,87	246	18,3	2,0
25	77,80	216	11,5	1,6

Anàlisi de dades existents

Rang de les variables limitat pel control

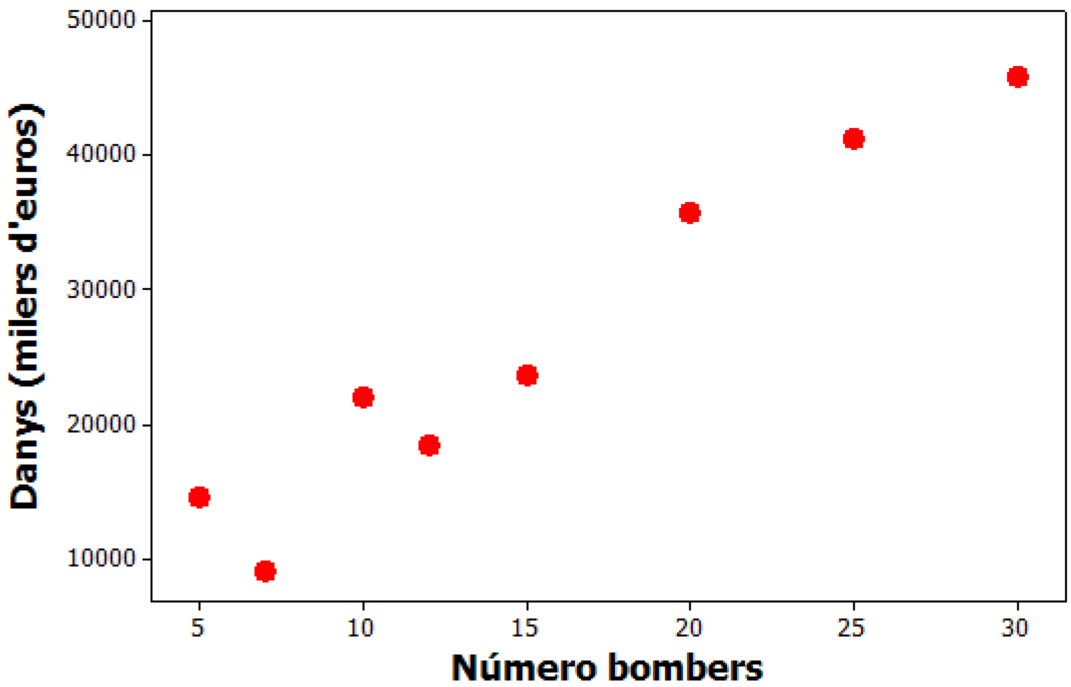




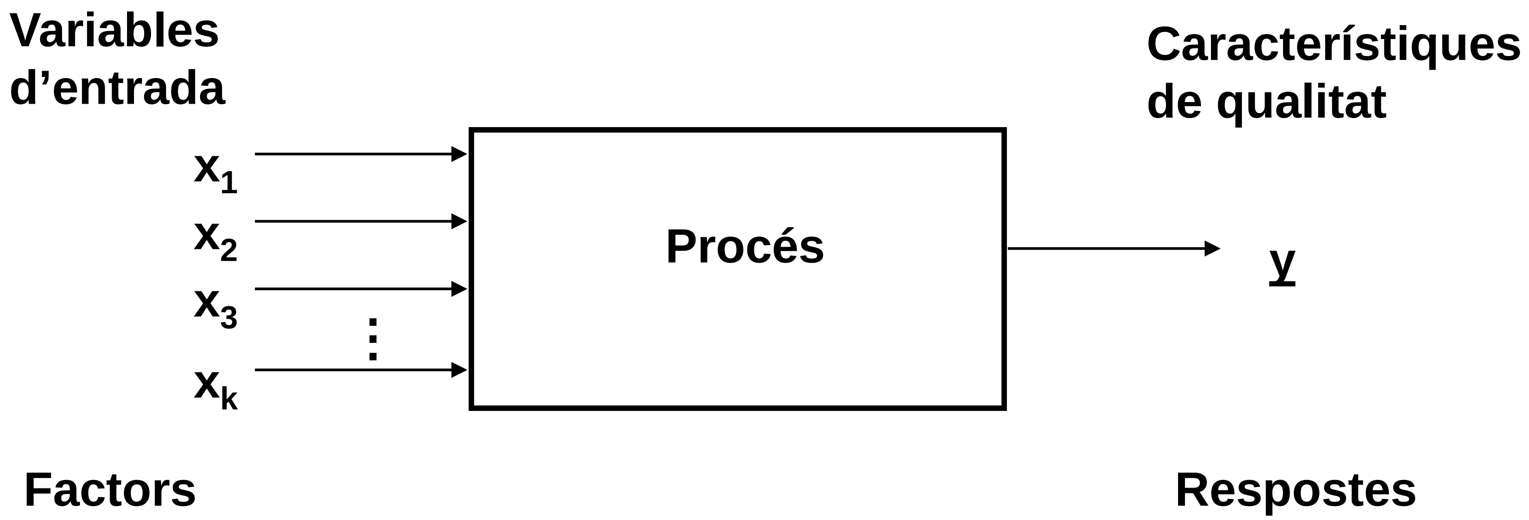
Anàlisi de dades existents

Relació entre número de bombers que s'envien a apagar un incendi i danys que provoca l'incendi (en milers d'euros)

Correlació i causalitat



La importància d'experimentar



És necessari experimentar

Nomenclatura

Resposta (Y): la variable que observem, la característica de qualitat que volem estudiar. De fet, pot haver més d'una resposta.

Factors (Xs): les variables que fixem durant l'experimentació. Volem descobrir com afecten els factors (les Xs) a la resposta.

Nivells: cadascun dels valors als que fixem un factor durant l'experimentació.

Covariables: variables que podem observar durant l'experimentació, però que no les podem controlar (=fixar al valor que volem).

Disseny: tot el conjunt d'experiments que realitzem per donar resposta a la pregunta d'interès.

Algunes idees bàsiques

Aleatoritzar: fer els experiments en ordre aleatòria per protegir-nos contra possibles fonts desconegudes de variabilitat.

Replicar: anar augmentant la grandària de la mostra per així millorar la precisió en l'estimació dels paràmetres.

Bloquejar: agrupar els experiments en blocs per eliminar causes conegudes de variabilitat.

Factors fixes: els nivells dels factors amb què experimentem són tots els que ens interessen. Fem inferència amb aquests nivells únicament.

Factors aleatoris: els nivells dels factors amb què experimentem són una mostra aleatòria de la població de possibles nivells. Volem fer inferència sobre tots els nivells de la població, encara que només experimentem amb els de la mostra.

Models

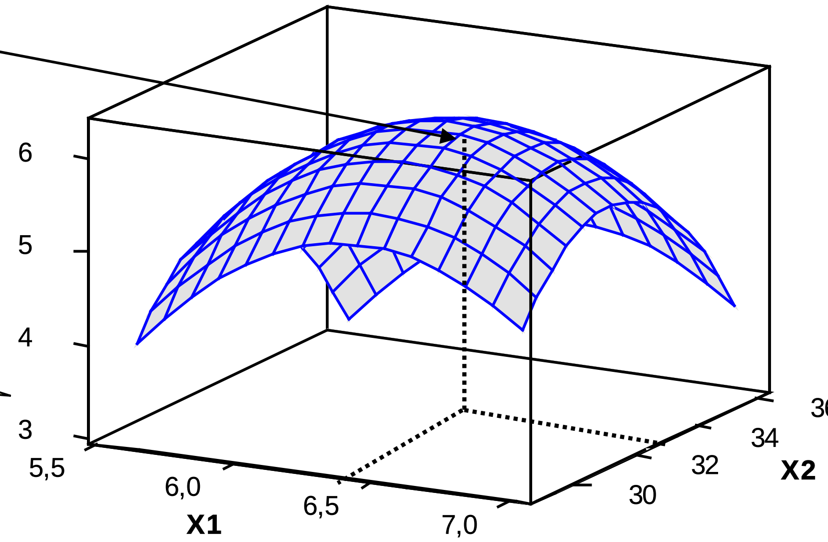
$$y = f(\underline{x}_1) + \varepsilon(\underline{x}_2)$$

$$\underline{x}_1 = \left[x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots\right]'$$

$$\underline{x}_2 = \left[x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, \ldots\right]'$$

$$y = \underbrace{f(X_1, X_2)}_{\text{Part estructural}} + \underbrace{\varepsilon}_{\text{Part estocàstica}}$$

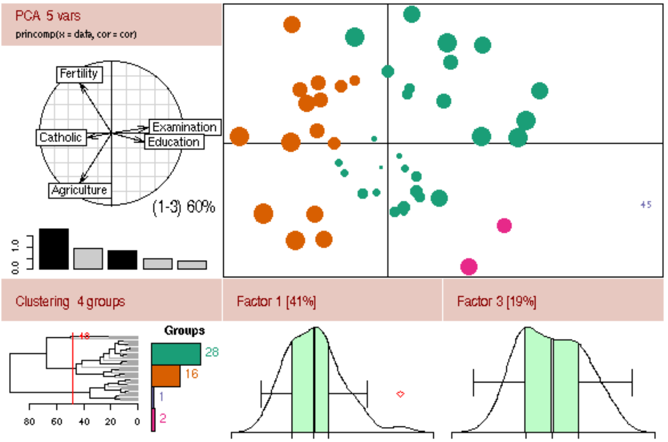
$$y = f(X_1, X_2)$$



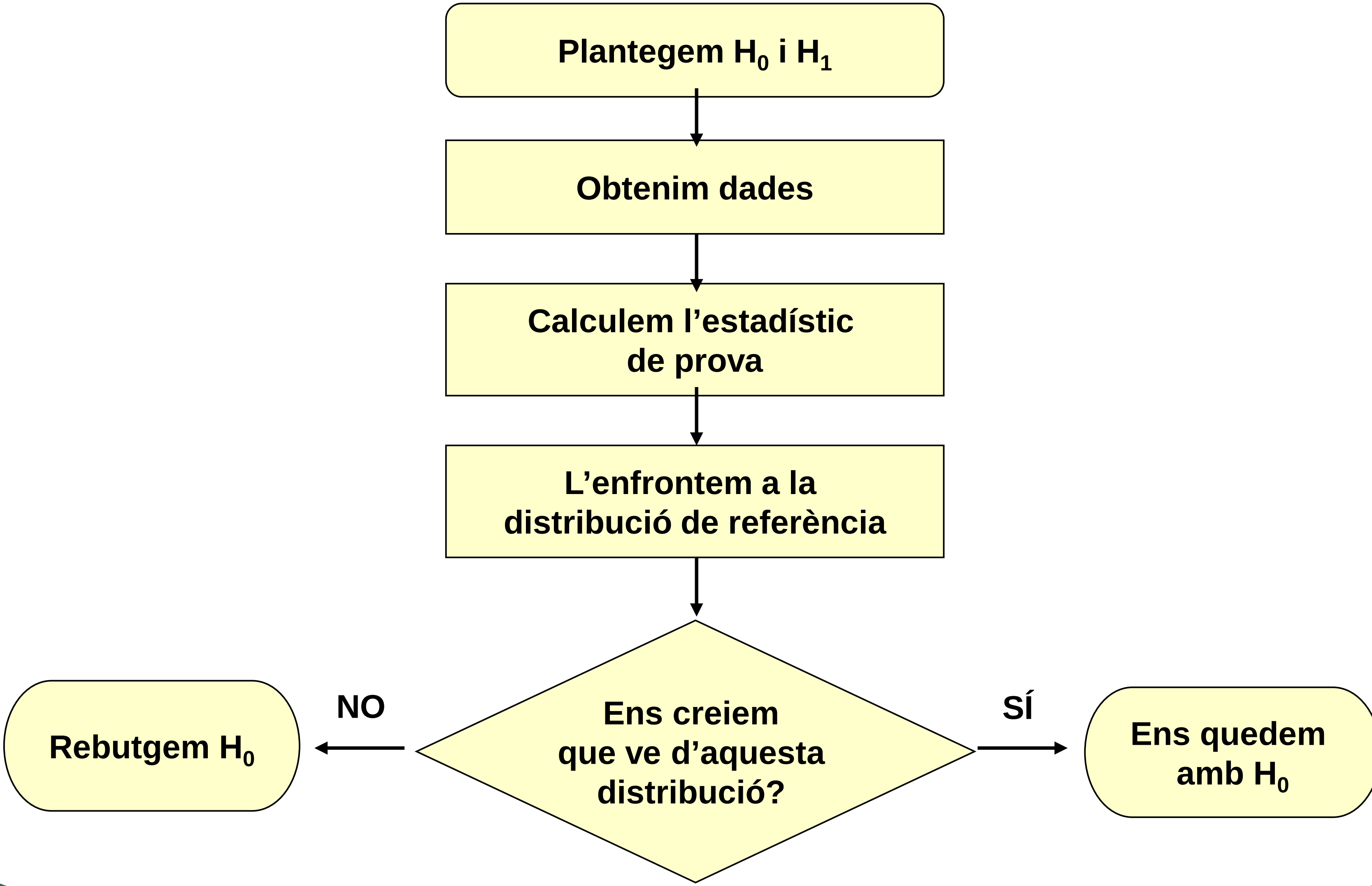
En realitat, la superfície té una "tremolor" deguda a la part estocàstica.

Repàs d'alguns conceptes bàsics. Distribucions que farem servir.

$$y_{ij} = \mu + \alpha_i + e_{ij}$$



Esquema de raonament del contrast d'hipòtesis



Error tipus I i error tipus II

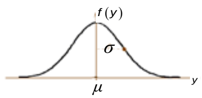
L'esquema de raonament del contrast d'hipótesis és semblant al d'un judici

H_0: Declarem a l'acusat no culpable.

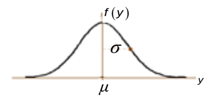
H_1: Declarem a l'acusat culpable.

		El que decidim: Tribunal declara a l'acusat	
		H_0: NO CULPABLE	**H_1:CULPABLE**
La realitat: L'acusat és	**INNOCENT**	**Encertem!**	**Error tipus I**
	CULPABLE	**Error tipus II**	**Encertem!**

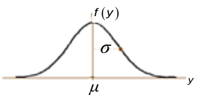
Distribució Chi-quadrat χ^2_ν



N (0;1)



N (0;1)



N (0;1)

$$Y_{11} \qquad Y_{21} \qquad \qquad Y_{\nu 1} \qquad \sum Y_{i1}^2$$

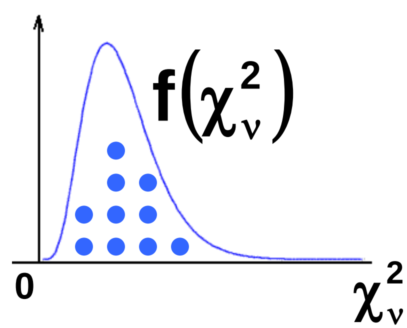
$$Y_{12} \qquad Y_{22} \qquad \qquad Y_{\nu 2} \qquad \sum Y_{i2}^2$$

$$\vdots \qquad \vdots \qquad \cdots \qquad \vdots$$

$$Y_{1i} \qquad Y_{2i} \qquad \qquad Y_{\nu i} \qquad \sum Y_{i\nu}^2$$

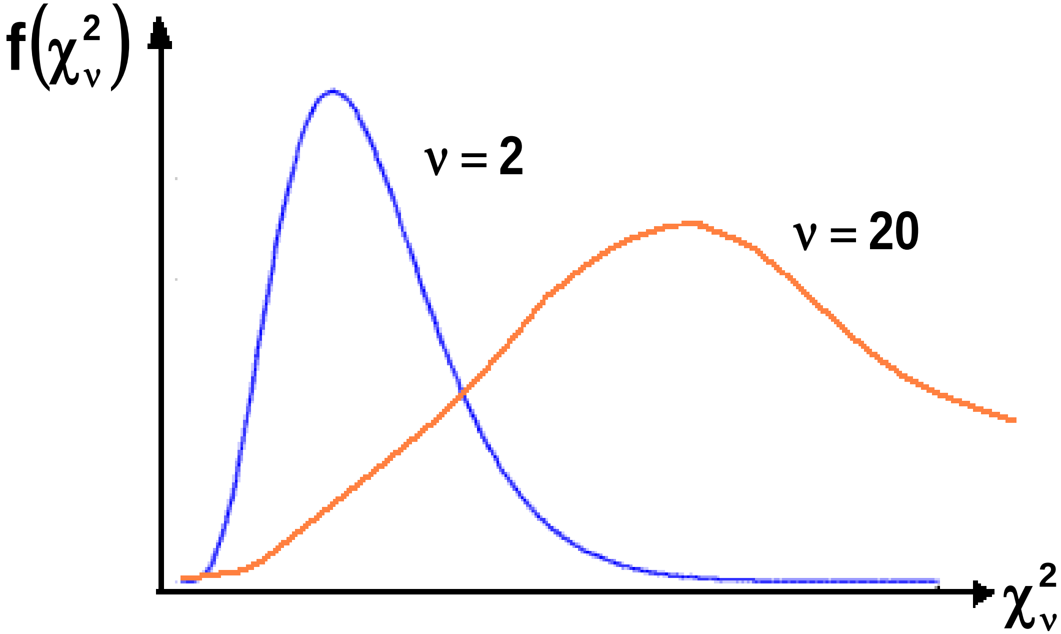
$$Y_i \sim N(0;1)$$

i independents



$$\chi^2_\nu = \sum_{i=1}^{\nu} Y_i^2$$

Distribució Chi-quadrat χ^2_ν



La forma de la densitat de χ^2 depèn de v

Quan $\nu \rightarrow \infty$ aleshores $\chi^2_\nu \rightarrow N\left(\nu; \sqrt{2\nu}\right)$

$$E\left(\chi^2_\nu\right) = \nu = \mu \qquad V\left(\chi^2_\nu\right) = 2\nu$$

t-Student

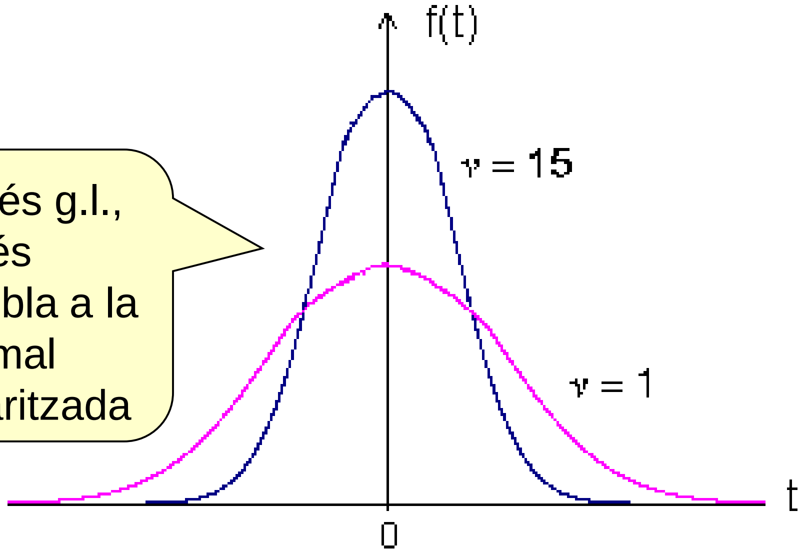
Tenim $z \sim N(0;1)$ $U \sim \chi^2_v$

z i U són independents

Aleshores:

$$F = \frac{z}{\sqrt{\frac{U}{v}}} \sim t_v$$

Com més g.l., més s'assembla a la normal estandaritzada



$$\nu \rightarrow \infty \quad f(t) \rightarrow \text{Normal}$$

$$t = \frac{Y - \mu}{s} \sim t\text{-Student}_{n-1}$$

$$t = \frac{\overline{Y} - \mu}{s/\sqrt{n}} \sim t\text{-Student}_{n-1}$$

$$E(t) = 0$$

$$Var(t) = \frac{\nu}{\nu - 2} \quad \nu > 2$$

F-Fisher (o F-Snedecor)

Tenim $V \sim \chi^2_{\nu_2}$ $U \sim \chi^2_{\nu_1}$

U i V són independents

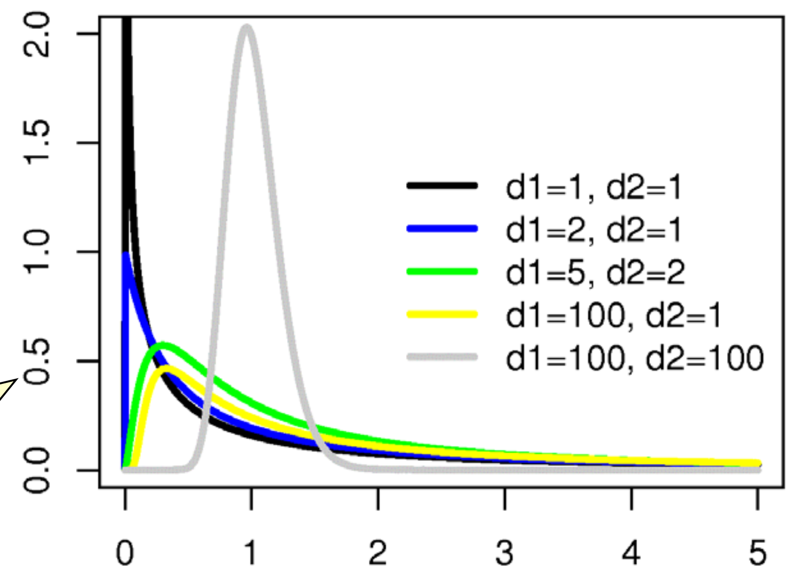
Aleshores:

$$F = \frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}} \sim F_{\nu_1;\nu_2}$$

$$F_\alpha(\nu_1;\nu_2) = \frac{1}{F_{1-\alpha}(\nu_2;\nu_1)}$$

$$[t-\text{Student}_\nu]^2 = F_{(1;\nu)}$$

- **La F-Snedecor no és en general simètrica**
- **Depèn de 2 paràmetres:**
  Els g.l. del numerador
  Els g.l. del denominador



Font: wikipedia.org

⚠ $F_{(\nu_1,\nu_2)} \neq F_{(\nu_2,\nu_1)}$

Comparar una mitjana amb un valor

Un procés omple ampolles amb un contingut que segueix una N(μ; σ).
(No sabem quant val ni μ ni σ)

L'enginyer de la planta productora vol esbrinar si el procés està centrat en μ = 200 o si està centrat en un valor menor.

1. Plantegem H_0 i H_1

   H_0 $\mu = 200$

   H_1 $\mu < 200$

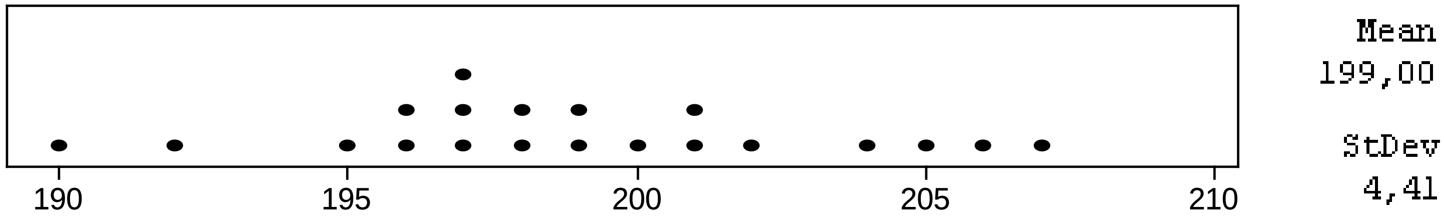
Altres possibilitats de H_1 serien:
$H_{1:}$ $\mu > 200$
$H_{1:}$ $\mu \neq 200$

2. Prenem una m.a.s. de les ampolles produïdes en el procés, i mesurem el seu contingut:

Contingut
205
202
198
196
201
207
199
195
197
192
204
206
201
200
190
196
197
197
199
198

Comparar una mitjana amb un valor

3. Fem un AED (anàlisi exploratori de dades). En aquest cas, un diagrama de punts.



4. Comprovem el supòsits en que es basa la metodologia.
   En aquest cas, cal que la mostra sigui realment una m.a.s. Gràcies al teorema central del límit, no cal normalitat estricta, ja que $\bar{y}$ és aproximadament normal.

Comparar una mitjana amb un valor

5. Calculem l'estadístic de prova

L'estadístic de prova és un número que podem calcular a partir de la nostra mostra i que:

- Ha de ser rellevant per verificar la validesa de H_0
- Ha de tenir una densitat de probabilitat (distribució de referència), coneguda com a mínim en el cas de que H_0 sigui certa

En aquest cas, l'estadístic de prova és:

$$t = \frac{\bar{Y} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t\text{-}Student_{n-1}$$

μ_0: La μ si la H_0 és certa

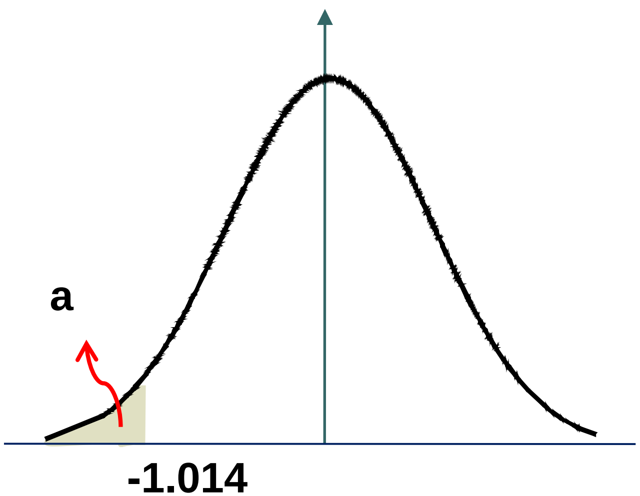
Comparar una mitjana amb un valor

Amb les dades de les ampolles...

$$t = \frac{\bar{Y} - 200}{\frac{s}{\sqrt{n}}} = \frac{199 - 200}{\frac{4,41}{\sqrt{20}}} = -1,014$$

$H_0 \quad \mu = 200$
$H_1 \quad \mu < 200$

6. Enfrontem l'estadístic de prova a la distribució de referència, i calculem el p-valor.



La distribució de referència és una t-Student amb n-1 g.l. (n–1 = 20–1=19)

p-valor = a = 0,1562

Si H_1 hagués estat $\mu \neq 200$ el p-valor seria 2a

Comparar una mitjana amb un valor

El p-valor és la probabilitat d'observar un valor de l'estadístic de l'experiment tant extrem (o fins i tot més) que l'obtingut si fos certa la H_0. El p-valor, que depèn de les dades experimentals, dóna una mesura de l'evidència inductiva contra la H_0 en l'experiment realitzat.

p-valor = 0,1562

Donat que el p-valor és força gran, no es rebutja la hipòtesi nul·la: no podem dir que $\mu < 200$

Rebutgem H_0 quan el p-valor és superior al nivell de significació α (la nostra frontera). Això és el mateix que dir que l'estadístic de prova queda a la zona de rebuig.

Al final, la decisió de si rebutgem o no la hipòtesi nul·la no és un problema estadístic: depèn del risc que estem disposats a assumir.

Comparar una mitjana amb un valor

Si tenim la situació: $H_{0:}\ \mu = \mu_0$
$H_{1:}\ \mu \neq \mu_0$

Un altre plantejament és, enlloc de calcular el p-valor, trobar un IC per la μ i veure si aquest IC inclou o no el valor $\mu_{0.}$

Hi ha una relació entre p-valor i IC:

Exemple: Amb $H_{0:}\ \mu = 50$; $H_{1:}\ \mu \neq 50$

IC del 95% per μ	[48; 52]. Inclou el 50, p-valor > 5%
IC del 97% per μ	[50; 53]. El 50 està en un extrem, p-valor = 3%
IC del 98% per μ	[51; 53]. No inclou el 50, p-valor < 2%

Per què? Recordem: IC 1-α per μ

$$\left[\bar{Y} - t_{n-1;\,\alpha/2}\frac{s}{\sqrt{n}}\,;\bar{Y} + t_{n-1;\,\alpha/2}\frac{s}{\sqrt{n}}\right]$$