Lecture nº 4.

Tests on gaussian samples (parametric).

"That's what I want to say. See if you can find some statistics to prove it."

Outline:

1. Tests on sample statistics .
2. ANOVA.

Tests sur les échantillons gaussiens.

Conditions initiales des tests:

- les données sont des valeurs numériques (autres que codes, rangs, conventions etc.);
- les variables aléatoires étudiées sont des réalisations de la loi normale,
- les variances de la variable étudiée dans les populations qu'on compare sont égales.

L'hypothèse de normalité,
sous laquelle les tests paramétriques sont valides
n'est pas toujours vérifiée,
ni même vérifiable en pratique.

Pour des échantillons de grande taille,
le théorème central limite assure la normalité asymptotique
des distributions empiriques.

Statistiques des tests.

variables permettant d'effectuer le test (et les lois qu'elles suivent)

	TEST de CONFORMITE	TEST d'HOMOGENEITE
COMPARAISON de **FREQUENCES**	$U = \dfrac{f - p}{\sqrt{\dfrac{p(1-p)}{n}}}$	$U = \dfrac{f_1 - f_2}{\sqrt{\dfrac{p(1-p)}{n_1} + \dfrac{p(1-p)}{n_2}}}$ où $p = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$
COMPARAISON de **MOYENNES**	$T = \dfrac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n}}}$	$T = \dfrac{\overline{X_1} - \overline{X_2}}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ où $\sigma = \sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}}$
COMPARAISON de **VARIANCES**	$Y^2 = \dfrac{n-1}{\sigma^2} \cdot S^2$ χ^2 (n-1)	$F = \dfrac{(S_1)^2}{(S_2)^2}$ de Snedecor F(n_1-1, n_2-1)

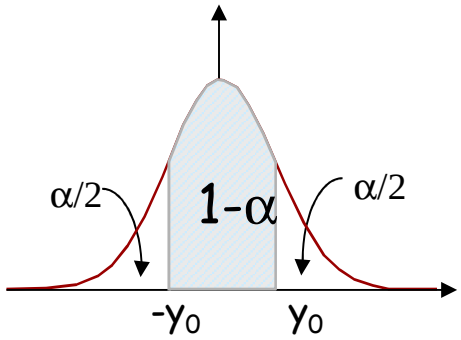
GAUSS ou STUDENT

Règles de décision:

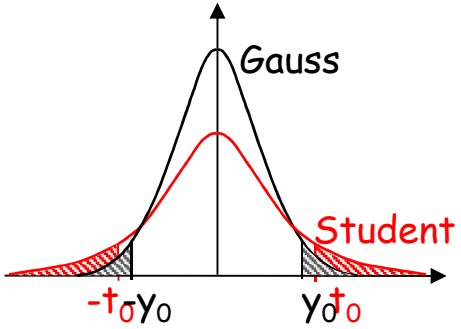
- Grands échantillons (n > 30):
  - Sous l'hypothèse H_o, quelle que soit la loi réellement suivie par la variable étudiée X, la variable U suit la loi N(0, 1).
  - Pour un niveau de risque α choisi les tables statistiques donnent la valeur y_0 telle, que

$$P(-y_0 < U < y_0) = 1 - \alpha$$

  - Si $|U| > y_0$, l'hypothèse H_0 doit être rejetée.



- Petits échantillons (n < 30):
  - Sous Ho, quelle que soit la loi suivie parX, la variable U suit la loi de Student à (n-1) degrés de liberté.



RAPPEL:
Pour le même risque α, la valeur limite à partir de laquelle H_0 sera rejetée est toujours plus large pour la distribution de Student par rapport à la loi normale.

Exemple d'application.

Considérons une machine destinée à la fabrication de comprimés devant peser 1 g. Aucun comprimé, s'il est mesuré au microgramme près, ne pèse 1 gramme exactement. Les poids des comprimés sont en moyenne de 1g, (avec un écart-type lié aux caractéristiques de la machine).

1 CONTROLE DE QUALITE DE LA PRODUCTION

n = 10 comprimés
$\overline{x}$ = 0.995 (un poids moyen)
σ = 0.01g (l'écart-type lié aux caractéristiques de la machine)

→ la statistique de test prend la valeur

$$\sqrt{\frac{n}{\sigma^2}}\,(\bar{x}-\mu)=\sqrt{\frac{10}{0.01^2}}\,(0.995-1)=-1.581$$

dont la p-valeur (loi normale) est de 0.0569.

2 CONTROLE DE LA FIABILITE DE LA MACHINE

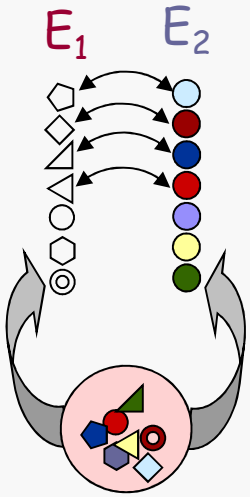
n = 10 comprimés
σ = 0.01g (caractéristique de la machine)
S = 0.013g (l'écart-type observé)

→ la statistique de test prend la valeur

$$\frac{nS^2}{\sigma^2}=\frac{10\cdot(0.013)^2}{(0.01)^2}=16.9$$

dont la p-valeur (loi χ^2) est de 0.0503.

Cas particulier 1: échantillons appariés.



ÉCHANTILLONS APPARIÉS:

chaque valeur x_{1i} de E_1 est associée à une valeur x_{2i} de E_2.

Les échantillons appariés ont la même taille.

On dispose de deux échantillons appariés, de taille n :

E → $E_1 : x_{11}, x_{12}, \ldots, \ldots, x_{1n}$ $\quad E_1(\overline{X}_1, S_1)$

E → $E_2 : x_{21}, x_{22}, \ldots, \ldots, x_{2n}$ $\quad E_2(\overline{X}_2, S_2)$.

PROBLEME: La différence entre les moyennes $(\overline{X}_1 - \overline{X}_2)$ est-elle significative ?

Hypothèse H_0 : La différence entre les échantillons n'est pas significative.

On calcule, pour chaque paire $\{ x_{1i} , x_{2i} \}$ la différence $\mathbf{d_i}$

$\mathbf{d_i} = x_{1i} - x_{2i}$

$E(\mathbf{d}) = \overline{\mathbf{d}}$ et $V(\mathbf{d}) = S^2$

Sous l'hypothèse H_0, $\overline{\mathbf{d}} = 0$

Statistique du test (variable servant à effectuer le test):

$$U = \frac{\overline{d}}{S/\sqrt{n}}$$

EXEMPLE 1 : On teste, sur six volontaires, deux somnifères : A et B. Les durées d'endormissement observées sont recueillis dans le tableau. Y a-t-il une différence significative d'efficacité entre ces deux médicaments?

	1	2	3	4	5	6
Somnifère A	15	27	38	19	45	8
Somnifère B	10	23	35	15	45	10
d_i	5	4	3	4	0	-2

$\nu = n-1 = 6-1 = 5$

$\alpha = 0.05$

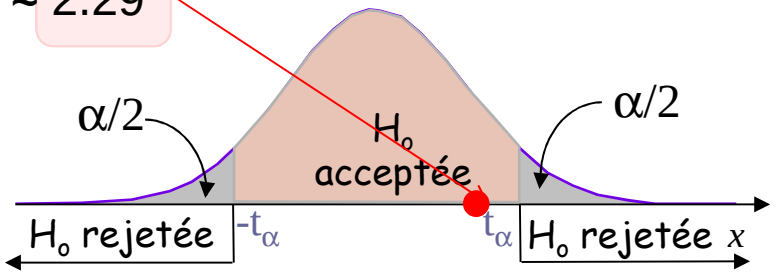
$t_\alpha(5, 0.05) = 2.571$

Hypothèse H_0: Les deux somnifères ont le même effet.

Sous l'hypothèse H_0, $\overline{d_i} = 0$

Dans l'expérience, $\overline{d_i} = 2.33$ et $S = 2.49$

La statistique du test prend la valeur $U = \dfrac{\overline{d}}{S/\sqrt{n}} \approx 2.29$



Conclusion : au risque α, l'hypothèse H_0 doit être acceptée; les deux somnifères ont le même effet.

Cas particulier 2: Analyse de la variance.

ANOVA: **AN**alyse **O**f **VA**riance

(COMPARAISON de PLUSIEURS MOYENNES EXPERIMENTALES)

- On étudie k populations P_1, P_2,P_3….P_k dans lesquelles les moyennes sont μ_1, μ_2, …, μ_k.
- De chacune des populations P_k on extrait un échantillon E_k de taille n_k.

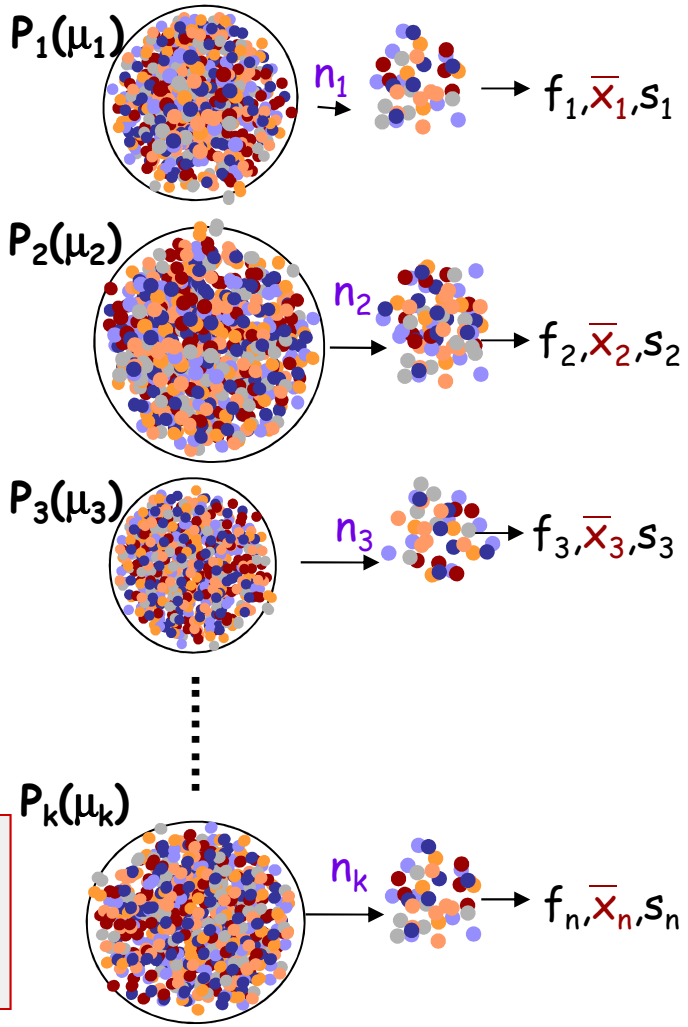
En général, les k échantillons correspondent à k (l,m,n…) modalités d'un facteur(s) étudié.

→ pour un échantillon E_i on observe n_i, $\bar{x}_i$ et $(S_i)^2$.

PROBLEME: Les moyennes des échantillons dépendent –elles du facteur(s) étudié ?

Hypothèse H_0 : Les moyennes des échantillons sont indépendantes du facteur étudié :

$\mu_1 = \mu_2 = \ldots = \mu_k = \mu$



Procédure du test.

Procédure:

Nous allons comparer les variances empiriques de chaque échantillon $(S_i)^2$ à la variance de l'échantillon global,

de taille $N = \sum_{i=1}^{k} n_i$

et de moyenne '*générale* ' $\bar{X} = \frac{1}{N} \sum_{i=1}^{k} n_i \cdot \bar{x}_i$

Variance résiduelle (intragroupe) (la moyenne des variances) :

caractérise (en moyenne) des fluctuations à l'intérieur de chaque échantillon.

$$(S_R)^2 = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1)(S_i)^2$$

Variance factorielle (intergroupe) (la variance des moyennes):

mesure la dispersion des échantillons autour de la moyenne générale :

$$(S_F)^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{X})^2$$

THEOREME: Sous l'hypothèse H_0, la variable aléatoire F

$$F = \frac{(S_F)^2}{(S_R)^2}$$

suit la loi de Snédécor à $\nu = (k - 1, N - k)$ degrés de liberté.

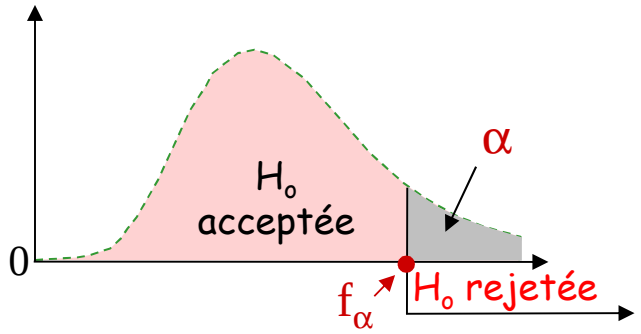
Décision :

La probabilité de rejeter Ho est conditionnée par le test du seuil:

$$P(F \geq f_\alpha) = \alpha$$

- si $F < f_a$ → on ne peut pas rejeter H_0 ;
- si $F \geq f_a$ → on écarte H_0 avec le risque α.

(autrement dit, on attribue une influence significative au facteur étudié).



Effectuer une analyse de variance
ou plusieurs tests de comparaison des moyennes (bilatéraux)
est strictement équivalent.

EXEMPLE 2 : On a dosé la teneur en calcium de trois types d'eaux issues d'origines géographiques différentes. Chaque type d'eau a fait l'objet de quatre prélèvement. Les résultats des dosages (en mg/l) sont :

Origine d'eau	dosage
A	18 20 22 25
B	15 16 17 21
C	15 20 21 25

L'origine géographique a-t-elle l'influence significative sur la teneur en calcium des eaux considérées ?

Hypothèse H_0: L'origine géographique des eaux n'a pas d'influence sur leur teneur en calcium.

Caractéristiques de l'échantillon global :

$$N = \sum_i n_i = 12$$

$$\bar{X} = \frac{1}{N}\sum_i n_i \bar{x}_i = \frac{1}{12}(4 \cdot 21.25 + 4 \cdot 17.25 + 4 \cdot 20.25) = 18.75$$

variance résiduelle : $(S_R)^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)(S_i)^2 = \frac{1}{9}(3 \cdot 2.586 + 3 \cdot 4.153 + 3 \cdot 3.573) \approx 3.437$

variance factorielle : $(S_F)^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 =$

$$= \frac{1}{2}\{(4 \cdot (21.25 - 18.75)^2 + 4 \cdot (17.25 - 18.75)^2 + 4 \cdot (20.25 - 18.75)^2\} = 21.5$$

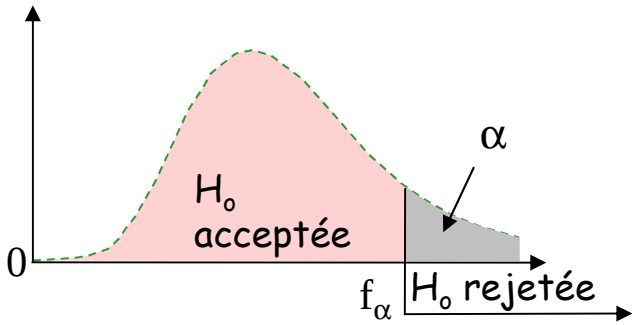
Sous l'hypothèse H_0, F suit la loi de Snédécor:

$$F = \frac{(S_F)^2}{(S_R)^2} = 6.255$$

ν = (2, 9) degrés de liberté

α = 0.1

$F > F_\alpha$ → F_α = 4.26



Conclusion : au risque α, l'hypothèse H_0 est rejetée ; l'origine géographique influence la teneur en calcium des eaux considérées.

Si l'analyse de variance accepte l'hypothèse d'égalité des moyennes, l'étude est terminée.

Mais si H_0 est rejetée, on peut souhaiter aller plus loin et comparer les effets du facteur étudié, pour des sous-ensembles de valeurs:

- en répétant des analyses de variance partielles.
- en comparant deux par deux les échantillons entre eux.

EXEMPLE 3: On étudie la durée de développement d'un parasite à l'intérieur d'un organisme hôte en fonction de la température d'élevage :

t[C]	nombre d'animaux	durée du développement	
		x	S^2
16	32	81	6.8
20	33	52	5.2
23	31	46	6.7

La température influence-t-elle le développement du parasite ?

Hypothèse H_0: la température n'influence pas la durée du développement du parasite.

Caractéristiques de l'échantillon global :

$$N = \sum_{i=1} n_i = 96$$

$$\overline{X} = \frac{1}{N}\sum_{i=1} n_i \overline{x_i} = \frac{1}{96}(32\cdot 81 + 33\cdot 52 + 31\cdot 46) \approx 59.73$$

variance résiduelle : $(S_R)^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)(S_i)^2 =$

$$= \frac{1}{93}\{\ 31\cdot(6.8)^2 + 32\cdot(5.2)^2 + 30\cdot(6.7)^2\ \} \approx 39.20$$

variance factorielle : $(S_F)^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x_i} - \overline{x})^2 =$

$$= \frac{1}{2}\{32\cdot(81-\overline{x})^2 + 33\cdot(52-\overline{x})^2 + 31\cdot(46-\overline{x})^2\} \approx 11146.48$$

$$F = \frac{(S_F)^2}{(S_R)^2} \approx 284.36$$

$\nu = (2, 93)$
$\alpha = 0.1,$
$F_\alpha = 7.3$

$F > F_\alpha$

Conclusion : au risque α, l'hypothèse H_0 est rejetée; la température a une influence significative sur la durée du développement du parasite

L'étude continue pour déterminer les paires des échantillons significativement différentes.

t[C]	nombre d'animaux	durée du développement	
		x	S^2
16	32	81	6.8
20	33	52	5.2
23	31	46	6.7

1. Hypothèse H_0: Il n'y a pas de difference de durée du développement du parasite à T=16 ºC et T= 20 ºC

Comparaison des moyennes sur des echantillons gaussiens (n > 30) :

$$\sigma = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{31\cdot 6.8 + 32\cdot 5.2}{32+33-2}} = 2.45$$

$$T = \frac{x_1 - x_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{81-52}{2.45\sqrt{\frac{1}{32}+\frac{1}{33}}} = 47.71$$

$\alpha = 0.1$,
$T_\alpha = 1.645$

$T > T_\alpha$

Conclusion : au risque α, l'hypothèse H_0 est rejetée; la température a une influence significative sur la durée du développement du parasite

L'étude continue pour déterminer les paires des échantillons significativement différentes.

t[C]	nombre d'animaux	durée du développement	
		x	S^2
16	32	81	6.8
20	33	52	5.2
23	31	46	6.7

2. Hypothèse H_0: Il n'y a pas de difference de durée du développement du parasite à T=20 ºC et T= 23 ºC

Comparaison des moyennes sur des echantillons gaussiens (n > 30) :

$$\sigma = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{32\cdot 5.2+30\cdot 6.7+}{33+31-2}} = 2.45$$

$$T = \frac{x_1-x_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{52-46}{2.45\sqrt{\frac{1}{33}+\frac{1}{31}}} = 9.79$$

$\alpha = 0.1$,
$T_\alpha = 1.645$

$T > T_\alpha$

Conclusion : au risque α, l'hypothèse H_0 est rejetée; la température a une influence significative sur la durée du développement du parasite