D - Inférence Statistique – Estimation et Tests d'hypothèses

1. Introduction – Déduction et inférence statistique
2. Fluctuations d'échantillonnage d'une proportion observée
3. Principe et résolution d'un test d'hypothèse
4. Fluctuation d'échantillonnage d'une moyenne observée
5. Comparaison de 2 variances observées - de 2 moyennes observées
6. Tests du Khi-deux
7. Autres tests non paramétriques
8. Analyses de variance - ANOVA

D - Inférence Statistique - Estimation et Tests d'hypothèses

1. Introduction – Déduction et inférence statistique

Population : ensemble total d'objets ou d'individus à étudier, à partir duquel sont extraits des échantillons.

La moyenne μ et l'écart-type σ de la population

sont des **constantes** (généralement inconnues),

exemples de **paramètres** fixes de la population ou **objectifs**.

⇨ Les probabilités **P(X)** sont utilisées dans le calcul de μ et σ

D - Inférence Statistique - Estimation et Tests d'hypothèses

1. Introduction – Déduction et inférence statistique

Echantillon : Sous ensemble de la population. Un échantillon **représentatif** est un sous ensemble **choisi au hasard dans la population**.

La moyenne $\overline{X}$ et l'écart-type **s** de l'échantillon sont des **variables aléatoires**, variant d'un échantillon à l'autre, et sont appelées **statistiques d'échantillon**, statistiques aléatoires ou **estimateurs** (ici respectivement de μ et σ)

Remarque : la médiane m_e peut être dans certains cas un meilleur estimateur de μ que $\overline{x}$

Les fréquences relatives f/n sont utilisées dans le calcul de $\overline{x}$ et s

⇨[Un estimateur T est dit biaisé si son espérance E(T) est différente de sa cible θ dans la population : biais = E(T)-θ]

Tirage d'un échantillon **avec remise** : **important** pour garantir **l'indépendance** des n observations qui le constituent (surtout dans les **petites populations**).

La proportion comme la moyenne changeant sinon à chaque tirage!

Tirage d'un échantillon **sans remise** : **sans importance dans les grandes populations**, aucune différence pratiquement que l'on remette ou non chaque individu avant le tirage suivant. Pour l'essentiel les observations sont indépendantes. Ce n'est pas le cas pour une petite population

D - Inférence Statistique - Estimation et Tests d'hypothèses

1. Introduction – Déduction et inférence statistique

Déduction :

prédire, à partir d'une population connue ou supposée connue,

les caractéristiques des **échantillons** qui en seront prélevés

Induction (inférence) :

prédire les caractéristiques d'une **population** inconnue à partir des statistiques déterminées dans un échantillon représentatif de cette population.

⇨ *Extrapolation des observations réalisées dans un échantillon à l'ensemble de la population*

Allons y ...

- Etude des fluctuations d'échantillonnage d'une **proportion** observée,
- Tests d'hypothèses (conformité, homogénéité)
- Estimations : Intervalles de pari, Intervalles de confiance
- Etude des fluctuations d'échantillonnage d'une **moyenne** observée,

  (en se servant du modèle de cahier des charges vu lors de l'étude des fluctuations d'échantillonnage d'une proportion observée)
- Intervalles de pari, Intervalles de confiance.
- Autre test utile : **Test du χ^2** de Pearson , (tests : conformité, homogénéité, ajustement à une loi)
- **Exemples ; exercices**
- **IMPORTANCE DES SCHEMAS !**

Les proportions, indicateurs parmi d'autres

- **Définition**
- **Intérêt**
- **On travaille sur un échantillon de taille quelconque pour obtenir des informations sur la population d'origine**
- **Caractéristique : Π** [population]
- **Statistique : P_o** [observée ou observable sur un n-échantillon]
- **P_o comme estimateur de Π**
- **Tests et encadrements possibles : sur Π et P_o**
- **On va travailler sur le biais : $\theta = \Pi - P_o$**

Comment traduire P(X=k) = 0.33 ?

⇨ Si l'on répétait un très grand nombre de fois l'expérience on obtiendrait X=k dans 33% des cas

⇨ Autre façon de le formuler :

33% des échantillons conduisent à X=k (nbre infini de tirages)

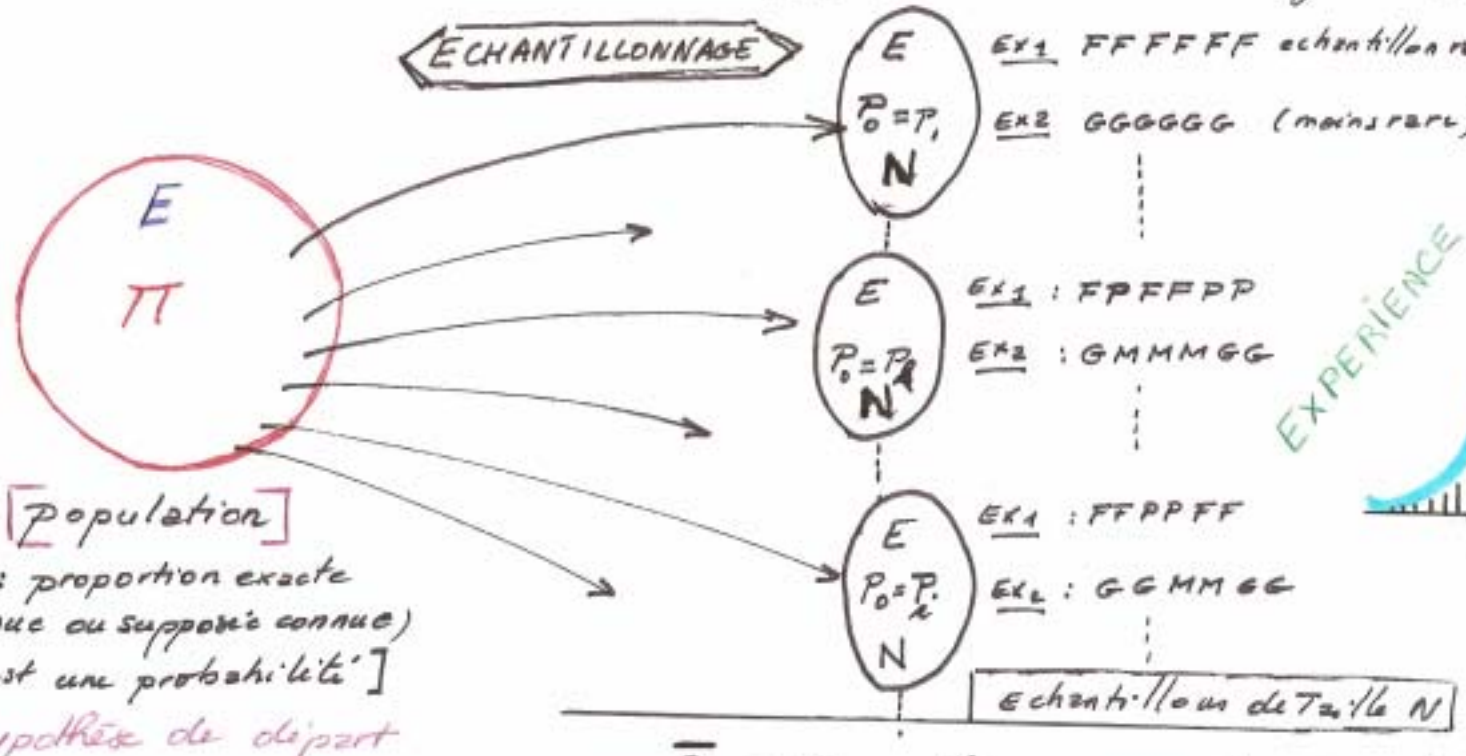
En remplaçant **X=k** par $\mathbf{P_o = k/n}$:

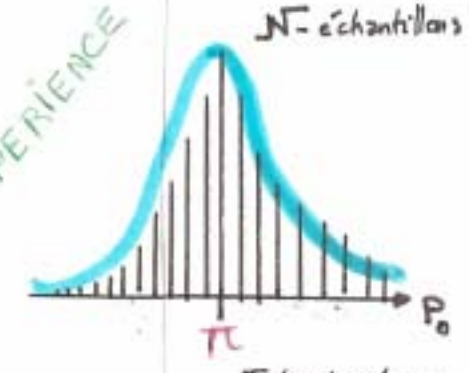
⇨ Fluctuations d'échantillonnage d'une proportion observée.

⇨ Nature : Loi Binomiale de paramètres **n** et Π (objectif dans population)

⇨ Problème continu lorsque **n** très grand.

Fluctuations d'échantillonnage de P_0





Fluctuations de P_0 autour sur π

π : proportion exacte (connue ou supposée connue)

[π est une probabilité]

Hypothèse de départ

ex1 : obtenir 'Face' (E) avec une pièce non truquée

➡ proportion de 'Face' obtenue dans une population infinie de jet d'une pièce $\pi = 0{,}5$

'F' → obtenir Face

ex2 : Efficacité d'un traitement antibiotique

pourcentage de guérison (E) constaté dans la population $\pi = 0{,}97$

$\overline{P_0} = \pi$ (en moyenne P_0 vaut π)

↳ il faut faire un très grand nombre de tirages pour le constater expérimentalement

X = " Nbre de réalisations de l'événement E dans un échantillon de taille N "

P_0 = " Proportion de réalisation de E mesurée dans un échantillon de taille N "

$X \rightsquigarrow \mathcal{B}(N, \pi)$ et $P_0 \rightsquigarrow \mathcal{B}(N, \pi)$

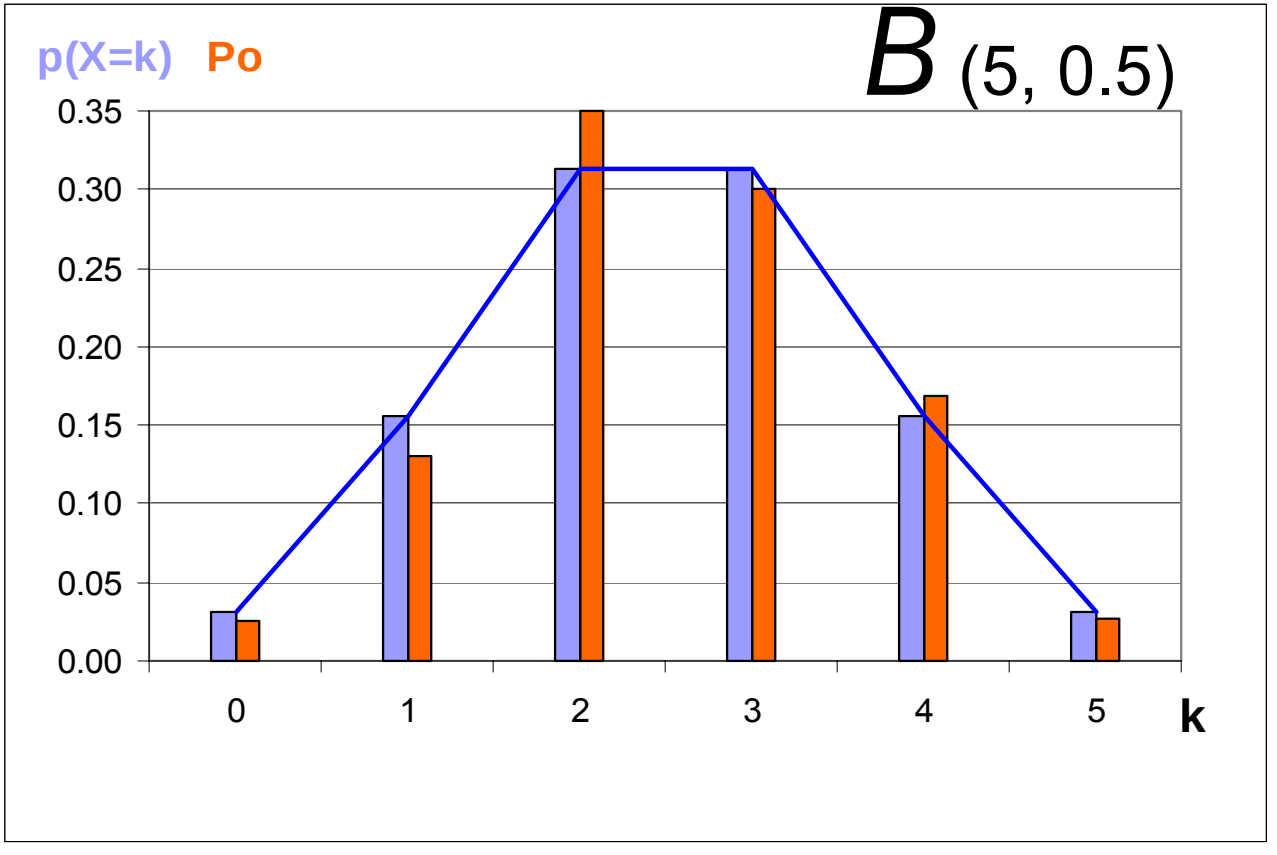
$\mu_X = N\pi$

$\mu_{P_0} = \pi$ et $\sigma_X^2 = \dfrac{\pi(1-\pi)}{N}$

$N\pi > 5$ et $N(1-\pi) > 5$

→ Fluctuation à tendance Normales $\mathcal{N}(\mu_X, \sigma_X)$

C - Lois de probabilités
Fluctuation d'échantillonage d'une proportion expérimentale (observée)



k	p(X=k)	Po
0	0.031	0.025
1	0.156	0.13
2	0.313	0.35
3	0.313	0.3
4	0.156	0.168
5	0.031	0.027
moy.	2.500	2.540

Résultat de l'échantillonnage
Expérience réalisée sur un grand nombre d'échantillons)

Loi théorique (atteinte lorsque le tirage concerne une nombre infini d'échantillons)

Fluctuations d'échantillonnage d'une proportion observée

⇨ Le caractère étudié est qualitatif
et sa répartition dans l'échantillon ou la population constitue une variable aléatoire quantitative discrète.

Lorsqu'un évènement donné E a une probabilité Π d'être observé dans une population, la proportion observée P_o de cet évènement dans des **n**-échanitillons, tirés au hasard de la population, subit des fluctuations d'échantillonnage suivant une loi binomiale de moyenne Π et d'écart type $\sigma = \sqrt{\frac{\Pi(1-\Pi)}{n}}$

En pratique (Ceci rend les calculs de probabilité aisés)

Si N Π > 5 et N(1-Π) > 5

Alors la loi binomiale est approchée par une loi Normale de même moyenne et même écart-type

$$Y \sim a \left(\Pi, \sqrt{\frac{\Pi(1-\Pi)}{n}}\right)$$

Nous serons amenés à vérifier cette condition à *priori* ou à *posteriori* !

⇨ Nous allons exploiter ces propriétés

D - Inférence Statistique - Estimation et Tests d'hypothèses
2. Fluctuations d'échantillonnage d'une proportion observée

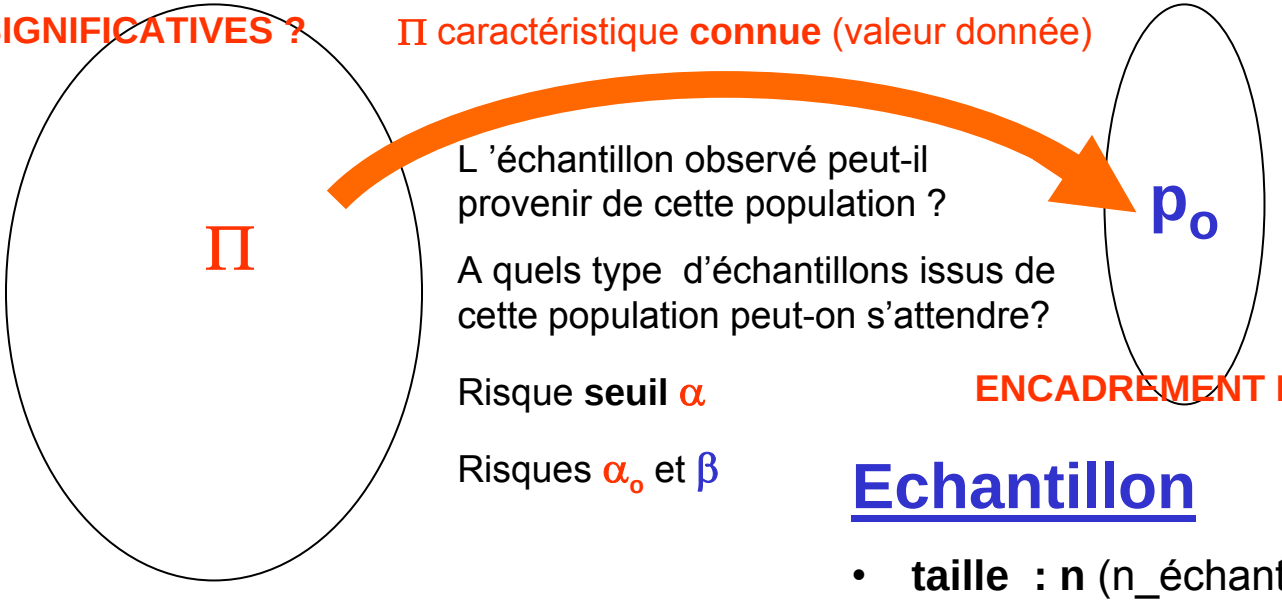
Tests de conformité

Intervalles de pari

Hypothèse nulle Ho :

DIFFERENCES (Po-Π) SIGNIFICATIVES ?

Π caractéristique **connue** (valeur donnée)



L 'échantillon observé peut-il provenir de cette population ?

A quels type d'échantillons issus de cette population peut-on s'attendre?

Risque **seuil** α

Risques α_o et β

ENCADREMENT DE Po

Echantillon

- **taille : n** (n_échantillon)
- **représentatif**
- **P_o : observée**

Population

- **taille ?**
- **Inaccessible**
- **Π : caractéristique connue ou supposée connue** (résultant par ex d'études menées sur de nombreux échantillons)

D - Inférence Statistique - Estimation et Tests d'hypothèses
2. Fluctuations d'échantillonnage d'une proportion observée

Un **intervalle de pari** , au risque (ou taux, ou seuil) α ,
encadre les valeurs de P_0 observables dans un n-échantillon lorsque celui-ci est tiré au hasard d'une population caractérisée par le paramètre exact Π

$$\Pi - \varepsilon_{\alpha/2}\sqrt{\frac{\Pi(1-\Pi)}{n}} \leq P_0 \leq \Pi + \varepsilon_{\alpha/2}\sqrt{\frac{\Pi(1-\Pi)}{n}}$$

Poser l'Hypothèse nulle (notée **Ho**)

Déduction

La probabilité de trouver P_0 dans cet intervalle est $(1-\alpha)$

Seuil standard : α = 5 %

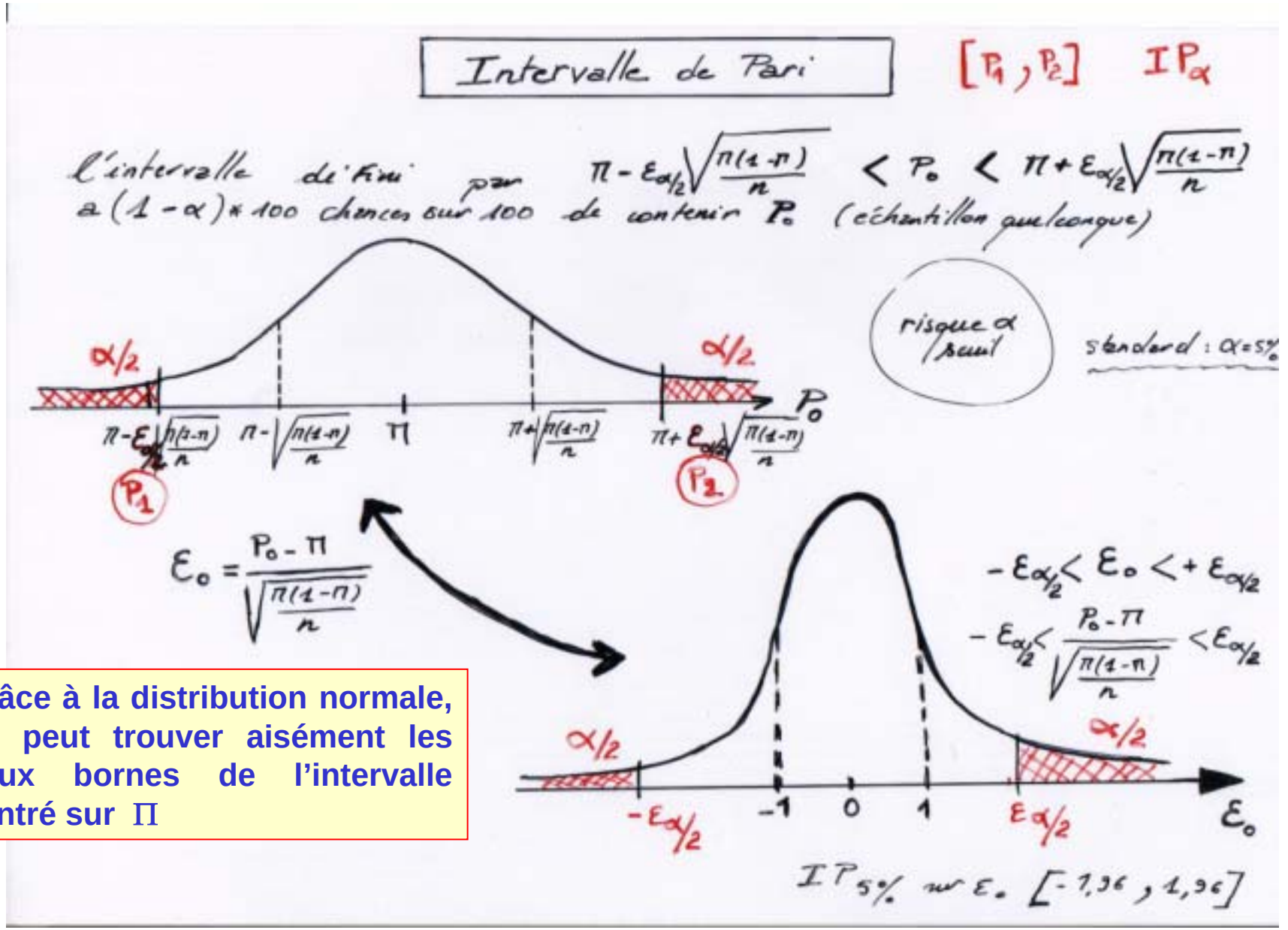
Il faut, bien sûr, être dans une situation où l'on peut approcher les fluctuations binomiales de la variable d'échantillonnage P_0 par une **loi normale**

⇨ ($N\Pi > 5$ <u>et</u> $N(1-\Pi) > 5$)

Important : On considère ε_α dans la table de l'écart réduit

mais $\varepsilon_{\alpha/2}$ dans la table $N(0,1)$ plus précise

Tirages aléatoire d'échantillons issus d'une population caractérisée par le paramètre Π



Grâce à la distribution normale, on peut trouver aisément les deux bornes de l'intervalle centré sur Π

Exercice ***"La machine déréglée "***

La production d'une machine déréglée conduit à 50 % de pièces défectueuses.

Un échantillon de 100 pièces est extrait de la production de cette machine.

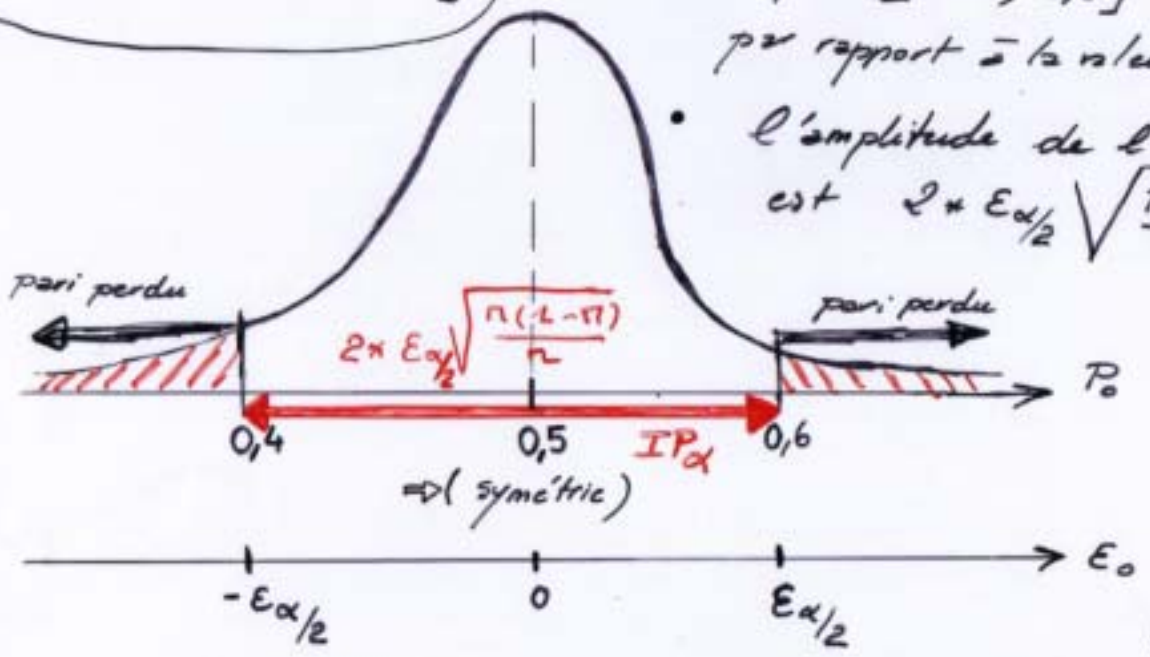
- Quelle est la probabilité d'observer moins de 30% de pièces défectueuses dans l'échantillon ?
- L'une des personnes de l'atelier voisin a parié une tablette de chocolat (suisse!) que la proportion observée de pièces défectueuses est comprise entre 40% et 60%. Quel risque a-t-elle pris de perdre son pari?
- Déterminer l'$IP_{5\%}$

D - Inférence Statistique - Estimation et Tests d'hypothèses

2. Fluctuations d'échantillonnage d'une proportion observée

On perd le pari pour P_o en dehors de $[0{,}4 , 0{,}6]$

- Soit α le risque associé à l'intervalle de pari $[0{,}4 , 0{,}6]$, symétrique par rapport à la valeur $\pi = 0{,}5$
- l'amplitude de l'intervalle de pari est $2 * \varepsilon_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}}$



$\Pi = 0{,}5$

$\sigma = 0{,}05$

$\varepsilon_o \sim N(0,1)$

Table

$$2\, \varepsilon_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}} = (0{,}6 - 0{,}4)$$

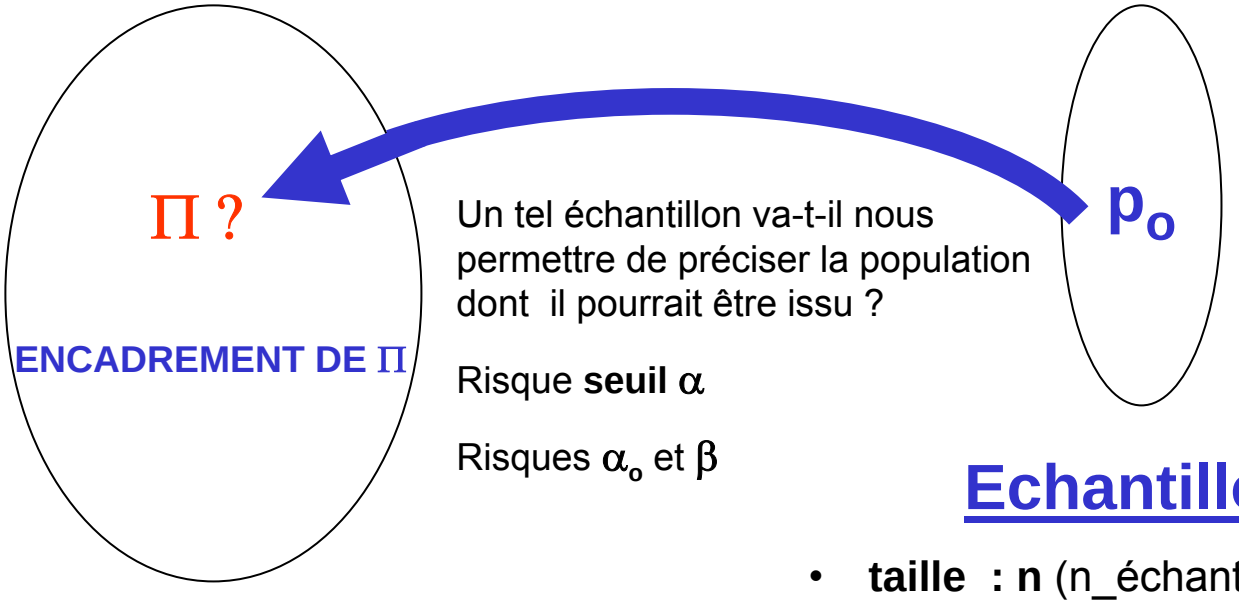
$$(\varepsilon_{\alpha/2})^2 = \frac{n \cdot 0{,}2^2}{4(\pi(1-\pi))} \quad ; \quad \varepsilon_{\alpha/2} = \sqrt{\frac{\cancel{100} \times \cancel{0{,}04}}{\cancel{4} \times 0{,}5 \times 0{,}5}}$$

$$\varepsilon_{\alpha/2} = \sqrt{\frac{100}{25}} \qquad \boxed{\varepsilon_{\alpha/2} = 2} \rightarrow \boxed{\alpha = 4{,}6\,\%}$$

probabilité d'observer entre 40% et 60% de sportifs dans l'échantillon : 95,4 %

Intervalles de confiance

On part des seules informations disponibles : p_o et n



Echantillon

- **taille : n** (n_échantillon)
- **représentatif**
- **P_o : observée**

Population

- **taille ?**
- **Inaccessible**
- **Π : caractéristique théorique ou attendue (inconnue)**

D - Inférence Statistique - Estimation et Tests d'hypothèses
2. Fluctuations d'échantillonnage d'une proportion observée

Intervalle de confiance **(IC_α)** : intervalle encadrant les valeurs de Π permettant de caractériser un ensemble de populations dont pourrait être issu l'échantillon dans lequel la valeur $\mathbf{P_0}$ a été observée

$$P_0 - \varepsilon_{\alpha/2}\sqrt{\frac{P_0(1-P_0)}{n}} \leq \Pi \leq P_0 + \varepsilon_{\alpha/2}\sqrt{\frac{P_0(1-P_0)}{n}}$$

Inférence

Validation de l 'IC$_\alpha$: Il faut finalement vérifier *à posteriori* que **nΠ** et **n(1-Π)** sont tous deux supérieurs à **5** aux deux **bornes** de l'intervalle. Dans le cas contraire, on ne peut pas valider l'intervalle de confiance.

Important : On considère ε_α dans la table de l'écart réduit

mais $\varepsilon_{\alpha/2}$ dans la table $N(0,1)$ plus précise

Exercice *" Les produits laitiers "*

Un micro sondage a révélé que sur 30 parisiens interrogés au hasard dans la rue, 18 consomment régulièrement des produits laitiers (lait, fromages, yaourts, beurre, produits dérivés du lait, …). En sachant que la ville de Paris compte 2,5 millions d'habitants, estimez le nombre de parisiens consomment régulièrement des produits laitiers aux risques de 10%, 5% et finalement 1% ?

Quels sont vos commentaires?

D - Inférence Statistique - Estimation et Tests d'hypothèses

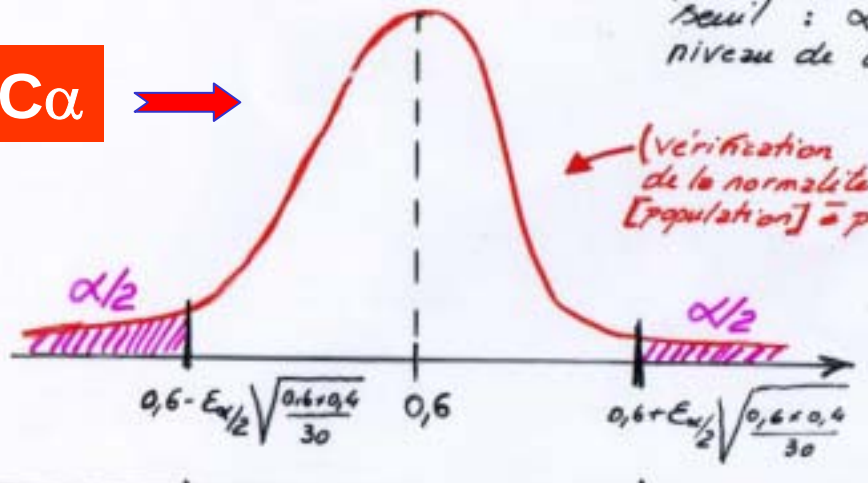
2. Fluctuations d'échantillonnage d'une proportion observée

observation : $18/30 = 0,6$ (P_0) ⇒ on construit un intervalle de confiance pour π à partir de P_0

seuil : α

niveau de confiance : $1-\alpha$

ICα



(vérification de la normalité [population] à posteriori)

⇝ Connaissant l'échantillon, peut-on obtenir une information applicable à la population totale ?

π, proportion de Parisiens consommant régulièrement des produits laitiers

N_1 N_2

N, nbre de Parisiens consommant régulièrement des produits laitiers

($N = 2,5\ 10^6 * \pi$)

	α	$E_{\alpha/2}$	$IC\alpha$	Validation de l'IC	Encadrement de N
0,1440	10%	1,64	[0,45, 0,75]	OUI $n\pi > 5$ et $n(1-\pi) > 5$ (aux bornes)	[$1,125.10^6$, $1,875.10^6$]
0,1744	5%	1,96	[0,43, 0,77]	OUI $n\pi$ et $n(1-\pi) > 5$ aux 2 bornes	[$1,075.10^6$, $1,925.10^6$] CREDIBILITE ? DU SIMPLE AU DOUBLE !
0,2074	1%	2,57	[0,37, 0,83]	OUI $(1-0,81)\times 30 > 5$ (5,7) à juste !	[$0,925\ 10^6$, $2,075\ 10^6$] L'écart se creuse !

la variable d'échantillonnage est P_0 : proportion de gens consommant régulièrement des produits laitiers dans un échantillon de 30 parisiens (tiré au hazard)

$\sqrt{\frac{0,6\times 0,4}{30}} = 0,089$

VOS REMARQUES ?

Question : A partir de quel seuil α, ne peut-on plus encadrer N (en utilisant la loi normale) ?

D - Inférence Statistique - Estimation et Tests d'hypothèses
3. Principe et résolution d'un test d'hypothèse

Cette partie va nous permettre de développer le **principe** des tests d'hypothèses et de présenter la **méthode** de résolution des problèmes qui leurs sont associés

Généralités

Les tests sur les **proportions** vont servir de **modèles** pour tous les tests d'hypothèses.

Pour résoudre un test d'hypothèses il va falloir suivre une **procédure**

Cette méthode permet d'obtenir toutes les **conclusions** que l'on peut tirer des observations dont on dispose, conclusions pouvant servir de base à une **décision**.

Il n'y a **jamais de réponse absolue** dans cette démarche !
Les conclusions des tests d'hypothèse doivent faire apparaître les **risques** de se tromper qui y sont associés (en termes de **probabilités**)

Avertissements

Tirer des conclusions à partir d 'un nombre limité d'observations (échantillon)

⇨ La certitude absolue ne peut jamais être atteinte car la population totale n 'est connue qu 'à travers l 'image imparfaite que constitue l 'échantillon.

Mais le degré de confiance que l 'on peut accorder à une conclusion peut être formulée en terme de probabilité.

D - Inférence Statistique - Estimation et Tests d'hypothèses
3. Principe et résolution d'un test d'hypothèse

Posons la problématique

Une **observation** faite sur un **échantillon** semble **peu probable** dans la **population** référence (ou incompatible avec un autre échantillon)...

⇨ Les **écarts** constatés entre l'échantillon et la population totale sont-ils **significatifs** ou imputables au **hasard** des fluctuations d'échantillonnage ?

⇨ On confronte la loi suivie dans la population à la probabilité de l'**observation** dans un échantillon tiré au hasard de la population. On est ainsi amené à poser une hypothèse (dite hypothèse nulle) pouvant rendre compte de cette observation **(voir transparent)**.

⇨ L'approche est la même pour toutes les variables d'échantillonnages étudiées (proportion, moyenne, ...) et pour tous les tests d'hypothèses (tests de conformité, d'homogénéité, d'ajustement à une loi de distribution) et toutes les lois de distributions utilisées (loi normale, Student-Fisher, Chi-deux, …).

⇨ Décision : rejet ou non de l'hypothèse nulle ; différences significatives ou dues au hasard des fluctuations d'échantillonnage.

Si les tests permettent de réfuter une hypothèse nulle avec un degré de confiance parfaitement défini, ils sont, dans l'autre sens, impuissants à démontrer la satisfaction de l'hypothèse nulle ; ils indiquent simplement que celle-ci n'est pas contredite par les faits.

Quelles vont être les grandes étapes ?

- Collecter les informations (échantillon)
- Caractériser l'**observation**
- Définir la **population**, l'**échantillon**
- Définir la **Variable Aléatoire** d'étude
- **Poser une Hypothèse nulle** (notée H_o) et son hypothèse alternative **H1**
- **Fixer** un risque **seuil** α arbitraire (généralement 5%, risque standard)
- **Calculer** α_o, probabilité d'observer **sous** H_o un tel évènement ou des évènements encore moins probables (et calculer éventuellement β) en utilisant les tables des lois de distributions connues : Binomiale, Poisson, Normale, Chi-deux, Student, Fisher, ...
- **Conclure** en comparant notamment α et α_o .

⇨ Cahier des charges de la résolution d'un test d'hypothèses

Exemple introductif : ***"Le médicament miracle"***

Un laboratoire pharmaceutique garantit que son médicament destiné au traitement d'une affection cutanée est efficace dans 80 % des cas. Un groupe indépendant a collecté les données de cabinets médicaux d'une région où 40 patients atteints par l'affection ont été traités par le médicament. Parmi eux, 28 ont été définitivement guéris. Pour les 12 autres patients le médicament a été sans effet, ils ont dû changer de traitement .

- Ces résultats sont-ils conformes à l'affirmation du laboratoire pharmaceutique ?
- En se basant sur les chiffres du laboratoire pharmaceutique, quelle est la probabilité de trouver entre 75% et 85 % de malades guéris par le médicament dans un tel échantillon ?
- Au risque seuil de 5%, donner les limites de la valeur attendue pour P_o.
- Quelle seraient ces limites si l'on avait étudié un échantillon de 150 personnes traitées par le médicament dont 105 auraient été guéries par le médicament ?

« Le médicament miracle » - Construction du test d'hypothèse

Population : l'ensemble des malades atteints de l'affection cutanée et traitées par le médicament
Echantillon : 40 patients (issus de cabinets médicaux de la région) atteints par l'affection traités par le médicament
V.A. d'étude :
P_o = "pourcentage de personnes guéries par le médicament dans un échantillon de 40 personnes"
Observation : P_o=28/40 = 70% guéris

Ho : l'échantillon provient d'une population définie par un taux de guérison par le médicament de Π**=0,8**.
Les différences entre ce paramètre et la valeur observée dans l'échantillon sont imputables au hasard des fluctuations d'échantillonnage.
Seuil : fixons un seuil α=5% à ne pas dépasser

D - Inférence Statistique - Estimation et Tests d'hypothèses
3. Principe et résolution d'un test d'hypothèse

Méthode ⇨ **(cahier des charges d'un test d'hypothèses)**

- Définir l'**échantillon** et caractériser l'**observation** (collecter les données, tri à plat)
- Définir la **variable aléatoire**
- Définir la **population** (que connaît-on de cette population ? Loi de distribution suivie par la variable aléatoire d'étude dans la population ?)
- **Poser une Hypothèse nulle** (notée H_o) permettant de raccorder l'observation faite sur un échantillon à une population donnée.
- **Le test** va alors consister à évaluer la probabilité d'obtenir, par hasard **sous l'hypothèse nulle**, l'observation dont on dispose (ainsi que les évènements encore moins probables).
  Il faut se fixer un **risque seuil** α arbitraire (généralement 5%) que l'on ne doit pas dépasser pour accepter l'hypothèse nulle. Ce seuil sera comparé à la probabilité α_o d'obtenir par le hasard sous H_o les observations dont on dispose.
- La **conclusion du test** dépendra du résultat cette comparaison

  Si $\alpha > \alpha_o$ on **rejette** l'hypothèse nulle avec un risque de première espèce α_o de se tromper en la rejetant alors qu'elle est vraie.

  Si $\alpha_o \geq \alpha$ l'hypothèse nulle est **accepatable** (acceptable jusqu'au seuil α_o)
  Dans certains cas nous pourrons définir un **risque de seconde espèce**, noté β, probabilité d'erreur d'admettre que H_o , qui est fausse, puisse être acceptable alors que c'est une **hypothèse alternative** H_1 qui est vraie.
- L'expérimentateur peut alors prendre une décision en fonction du résultat du test (+ bon sens ...)

Test de conformité à une proportion exacte

Il s'agit de confronter une valeur de la variable aléatoire P_o , observée dans un échantillon, à la valeur exacte du paramètre Π, dans la population.

Tout revient à déterminer si la différence observée entre les 2 valeurs est significative ou si elle peut s'expliquer par le hasard des fluctuations d'échantillonnage.

Le critère de test utilisé est :

$$\varepsilon_0 = \frac{P_0 - \Pi}{\sqrt{\frac{\Pi(1-\Pi)}{n}}}$$

$n\Pi > 5$ et $n(1-\Pi) > 5$

avec $\varepsilon_0 \sim N(0, 1)$

Si $\alpha_0 \geq \alpha \Leftrightarrow \varepsilon_0 \leq \varepsilon_\alpha$, H_o est acceptable

Si $\alpha_0 < \alpha \Leftrightarrow \varepsilon_0 > \varepsilon_\alpha$, H_o est rejetée

D - Inférence Statistique - Estimation et Tests d'hypothèses

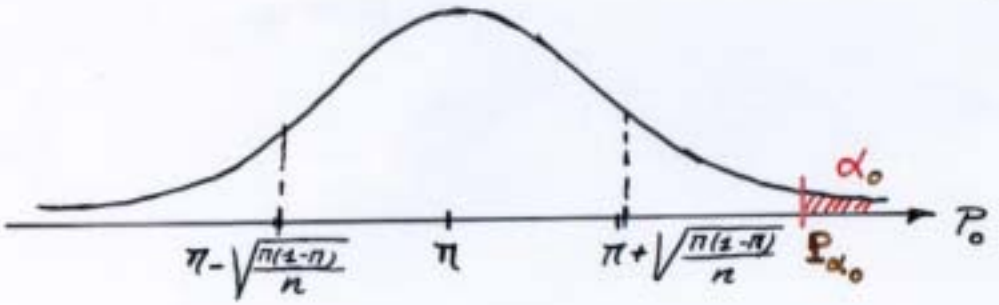
3. Principe et résolution d'un test d'hypothèse

- Fluctuation d'échantillonnage d'une proportion -

$P_0 \rightsquigarrow \mathcal{B}(n, \pi)$

$\simeq \mathcal{N}\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

$(n\pi > 5 \text{ et } n(1-\pi) > 5)$

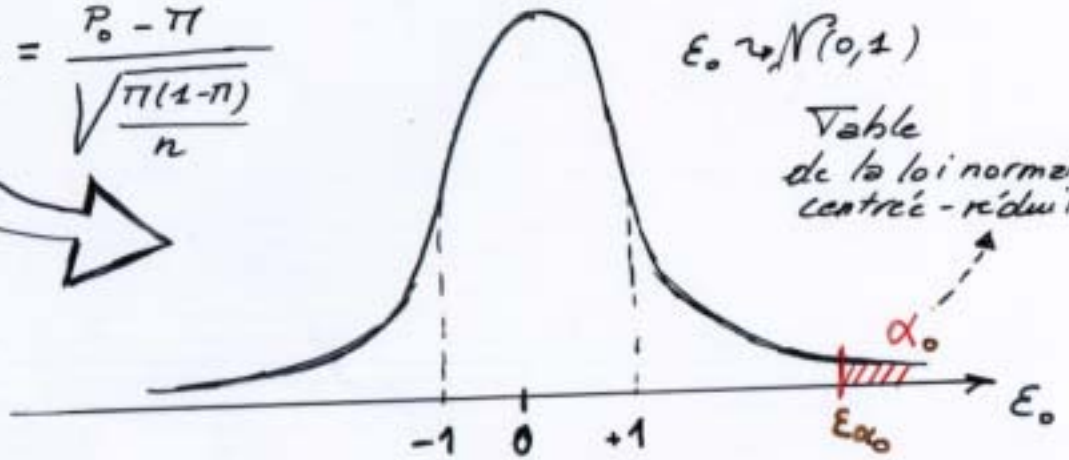


$$\varepsilon_0 = \frac{P_0 - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Changement de variable

$\varepsilon_0 \rightsquigarrow \mathcal{N}(0,1)$

Table de la loi normale centrée-réduite



« Le médicament miracle » - Construction du test d'hypothèse

Population : l'ensemble des malades atteints de l'affection cutanée
Echantillon : 40 patients (issus de cabinets médicaux de la région) atteints par l'affection traités par le médicament
V.A. d'étude :
P_o = "pourcentage de personnes guéries par le médicament dans un échantillon de 40 personnes"
Observation : P_o=28/40 = 70% guéris

Ho : l'échantillon provient d'une population définie par un taux de guérison par le médicament de **Π=0,8**.
Les différences entre ce paramètre et la valeur observée dans l'échantillon sont imputables au hasard des fluctuations d'échantillonnage.
Seuil : fixons un seuil α=5% à ne pas dépasser
Calcul de αo : sous Ho, P_o subit des fluctuations d'échantillonnage de nature binomiale B (40, 0,8) approximables par une distribution normale (0,8x40 et 0,2x40 tous deux supérieurs à 5)
de moyenne $\mu=\Pi$=0,8 et d'écart type $\sigma=(0,8x0,2/40)^{1/2}$=0,063 ; N (0,8, 0,063)

Test bilatéral : à priori, il y a autant de chances que le hasard induise des différences négatives que positives.
α_o=P((Po<0,7) ou (Po>0,9)) = 2×P(Po>0,9), probabilité d'observer un tel échantillon sous Ho
Z=(Po-0,8)/0,063 est normale centrée réduite N(0,1).
α_o=2×P(Z>(0,9-0,8)/0,063)
α_o=2×P(Z>1,58)
α_o=2×0,057053=0,114 (lecture table N(0,1))

Conclusion : $\alpha_o > \alpha$; Ho est acceptée au risque seuil α=5%
De telles différences peuvent s'expliquer par le hasard des fluctuations d'échantillonnage.
Ces résultats sont conformes à l'affirmation du laboratoire pharmaceutique.
Remarques : Aucune certitude mais la Statistique peut expliquer ces différences.
Une étude sur d'autres échantillons serait intéressante.

Exercice *"Les noix de coco"*

Un producteur déclare à son client que deux tiers des noix de coco qu'il lui fournit sont de tailles supérieures à une certaine limite qu'ils ont fixée ensemble. Le distributeur conteste cette statistique et avance que d'après les ventes qu'il a réalisées, ce sont seulement la moitié des noix de coco qui sont de taille suffisante (atteignant ou dépassant la limite fixée).

Pour en avoir le coeur net, un échantillon de 50 noix de coco est prélevé ou hasard de la production destinée au client. Le résultat tombe bientôt : 30 noix de l'échantillon sont de taille supérieure à la limite fixée.

Que peut-on en conclure ?

3. Principe et résolution d'un test d'hypothèse

Les noix de coco						
test unilatéral						
seuil α	**0.05**					
Po	**0.6**					
1ère solution						
Π sous Ho	**0.667**					
Π sous H1	**0.5**					
N	**25**	**50**	**100**	**150**	**250**	**1000**
σ (Ho)	0.094	0.067	0.047	0.038	0.030	0.015
σ (H1)	0.100	0.071	0.050	0.041	0.032	0.016
εo, critère de test sous Ho	0.707	1.000	1.414	1.732	2.236	4.472
α_0 = P(Po<0,6)	**24.0%**	**15.9%**	**7.9%**	**4.2%**	**1.3%**	**0.0%**
position limite pour α=5%	0.51	0.56	0.59	0.60	0.62	0.64
critère de test sous H1	0.120	0.811	1.787	2.536	3.724	8.995
β = P(Po<limite)	45.2%	20.9%	3.7%	0.6%	0.0%	0.0%
conclusion	***Ho acceptable***	***Ho acceptable***	***Ho acceptable***	***différences significatives***	***différences significatives***	***différences significatives***
2ème solution						
Π sous Ho	**0.5**					
Π sous H1	**0.667**					
σ (Ho)	0.100	0.071	0.050	0.041	0.032	0.0158
σ (H1)	0.094	0.067	0.047	0.038	0.030	0.015
εo, critère de test sous Ho	1.000	1.414	2.000	2.449	3.162	6.325
α_0 = P(Po>0,6)	**15.9%**	**7.9%**	**2.3%**	**0.7%**	**0.1%**	**0.0%**
position limite pour α=5%	0.66	0.62	0.58	0.57	0.55	0.53
critère de test sous H1	0.028	0.761	1.796	2.591	3.851	9.441
β = P(Po>0,6)	48.9%	22.3%	3.6%	0.5%	0.0%	0.0%
conclusion	***Ho acceptable***	***Ho acceptable***	***différences significatives***	***différences significatives***	***différences significatives***	***différences significatives***

D - Inférence Statistique - Estimation et Tests d'hypothèses

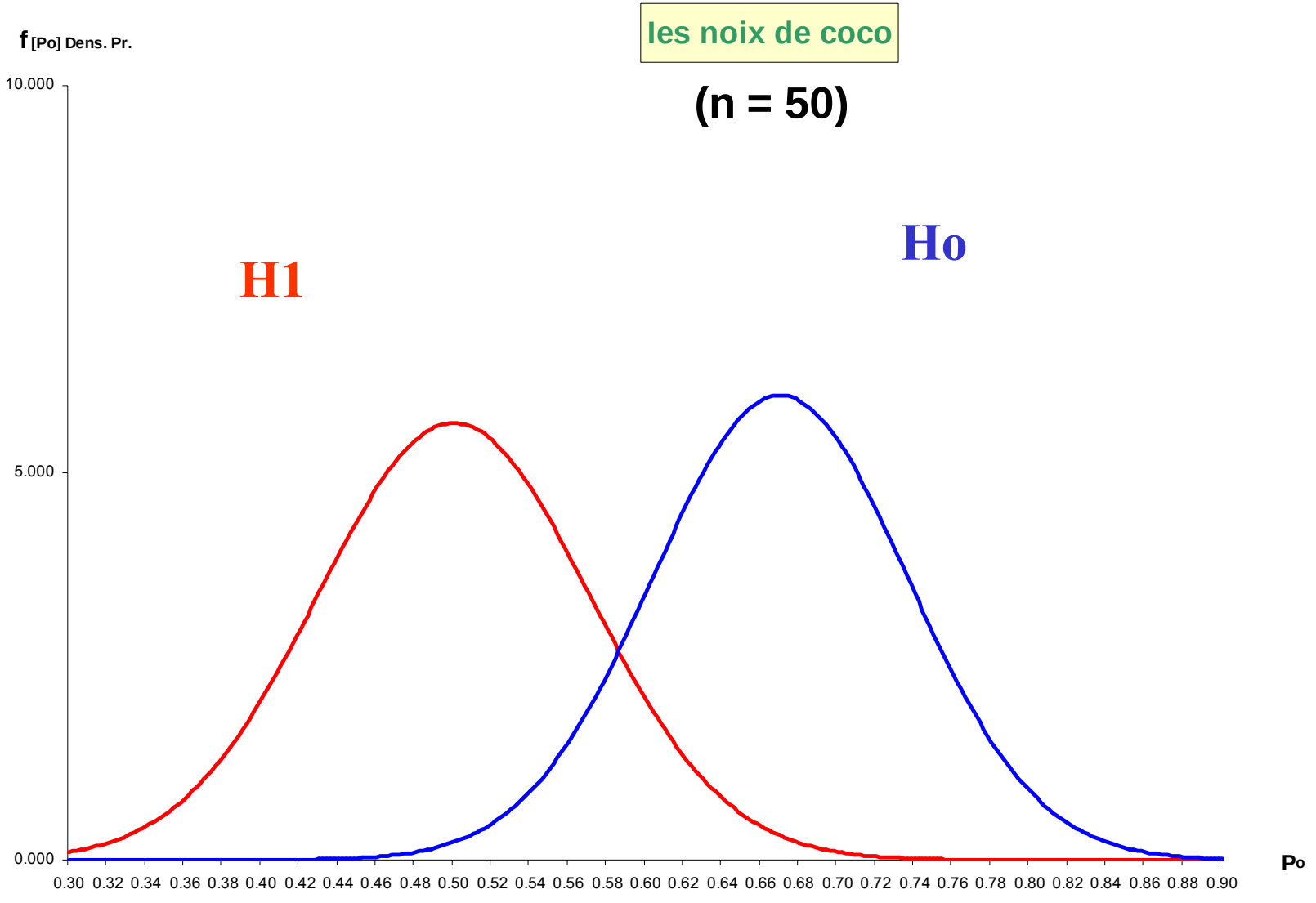
3. Principe et résolution d'un test d'hypothèse

Une première approche intuitive de la notion de maximum de vraisemblance peut-être évoquée dans le cadre des tests d'hypothèses. Elle consiste a considérer et tester plusieurs hypothèses nulles pour sélectionner, comme la plus vraisemblable, celle associée au plus fort degré de signification αo.

Dans l'exemple des « noix de coco »,
il apparaît que **(H_0 : Π = 2/3)** est l'hypothèse nulle la plus vraisemblable

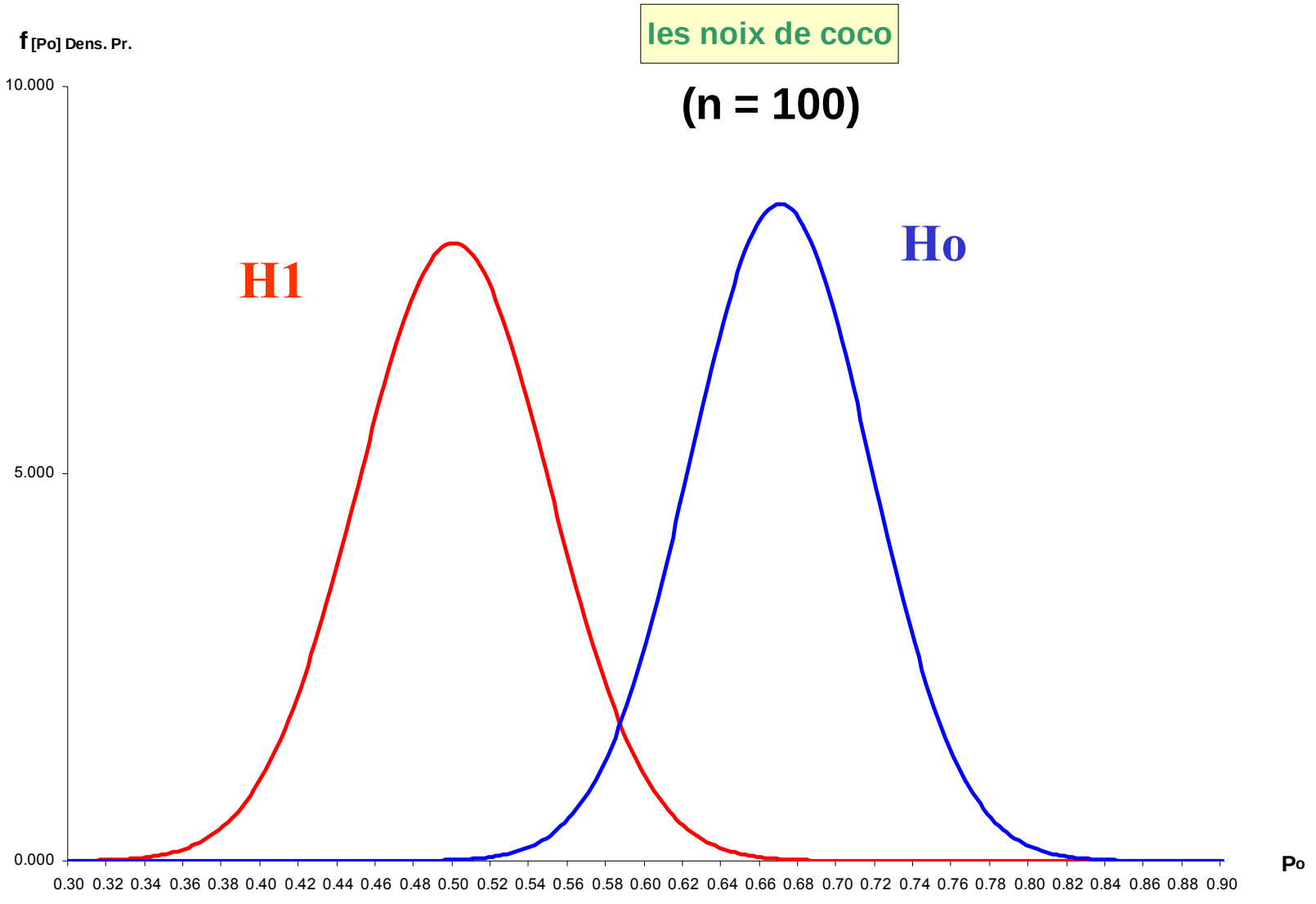
D - Inférence Statistique - Estimation et Tests d'hypothèses

3. Principe et résolution d'un test d'hypothèse



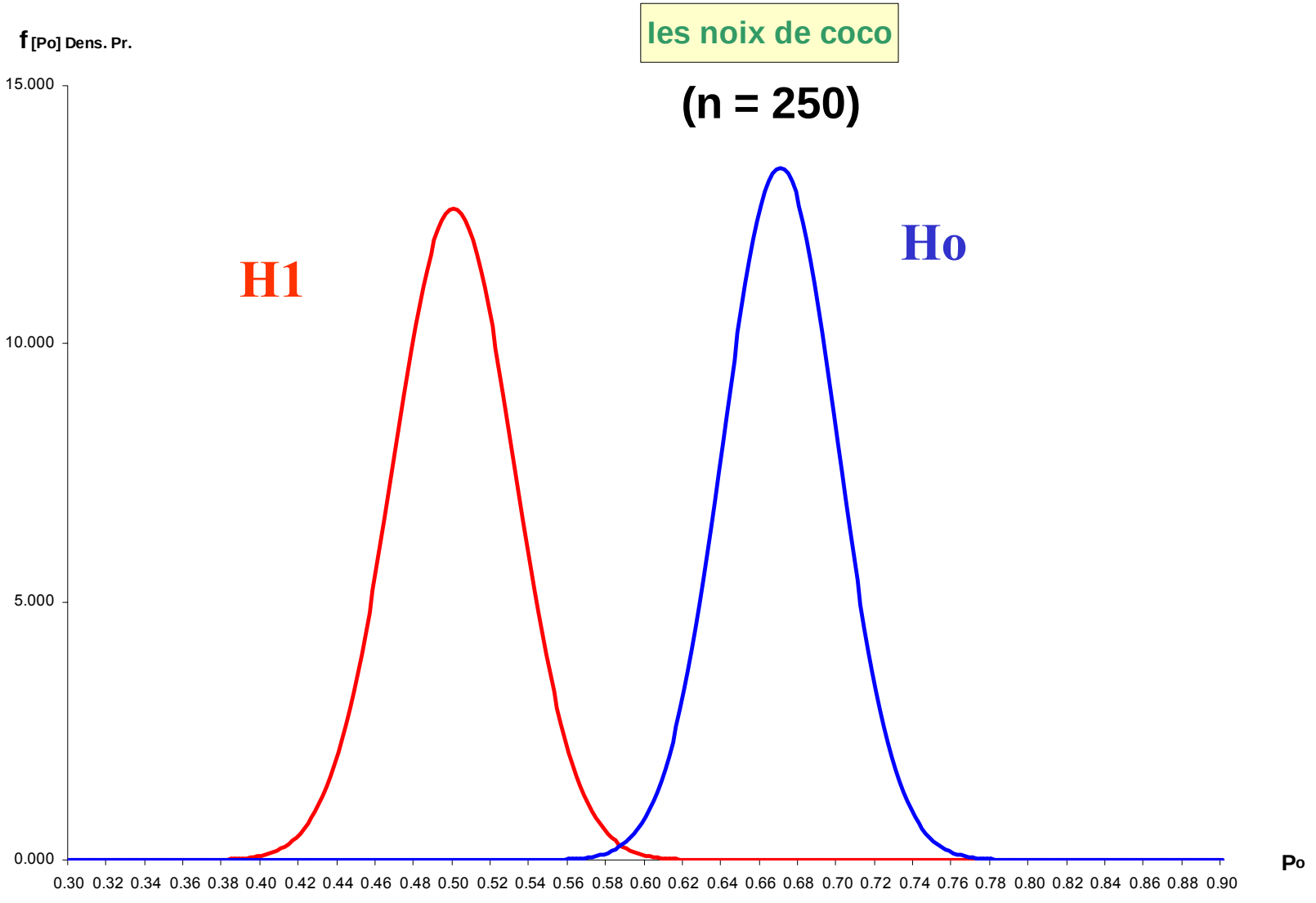
D - Inférence Statistique - Estimation et Tests d'hypothèses

3. Principe et résolution d'un test d'hypothèse



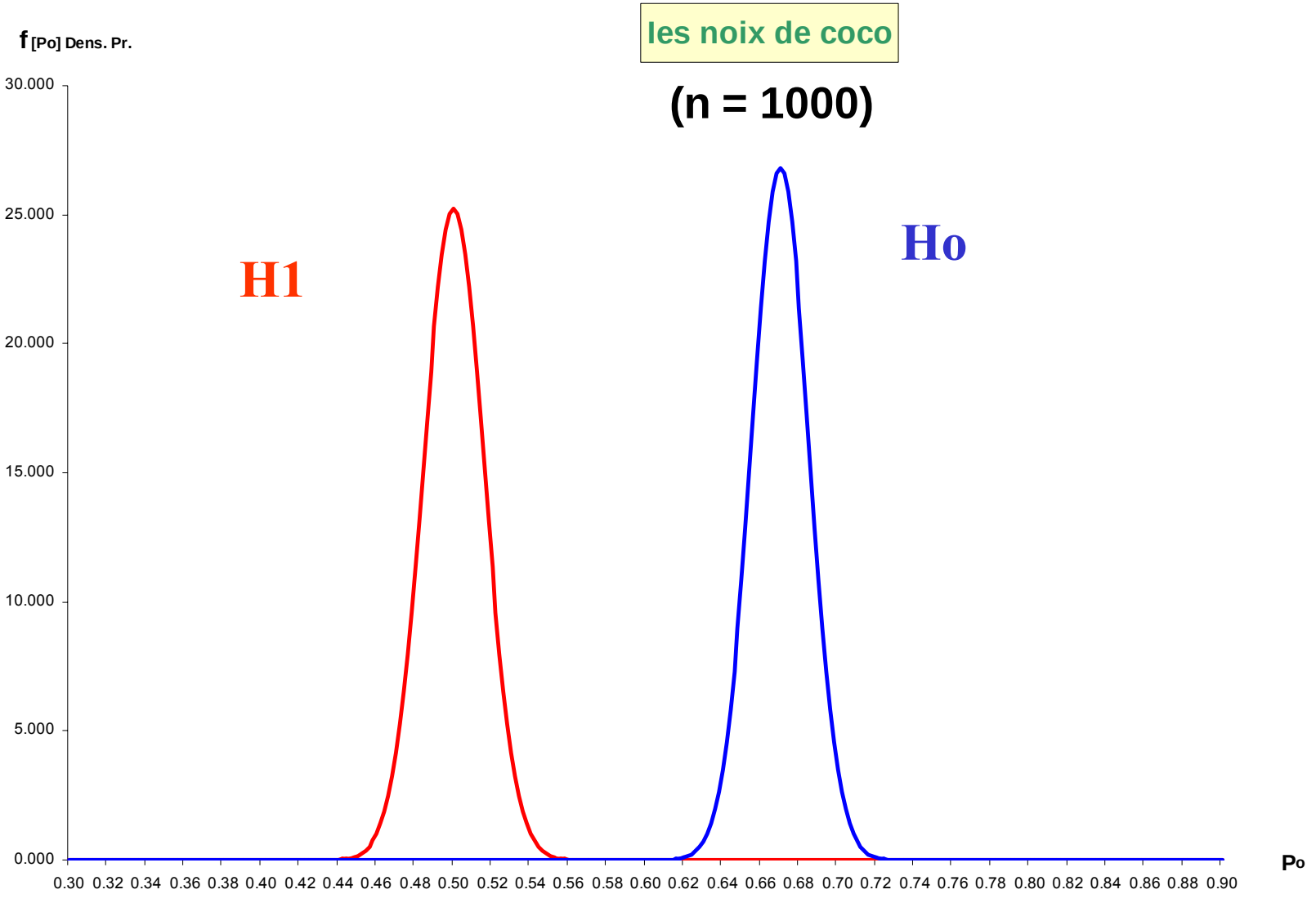
D - Inférence Statistique - Estimation et Tests d'hypothèses

3. Principe et résolution d'un test d'hypothèse



D - Inférence Statistique - Estimation et Tests d'hypothèses

3. Principe et résolution d'un test d'hypothèse



Risque de seconde espèce

β = probabilité d'accepter à tord Ho

C'est-à-dire probabilité :

- d'accepter H_o alors que c'est l'hypothèse alternative H_1 qui est vraie
- de réalisation de la V.A. sous l'hypothèse H_1 dans le domaine d'acceptation de H_o

Puissance du test

$1-\beta$ = puissance du test

- probabilité de mettre en évidence que l'hypothèse nulle H_o est fausse lorsqu'elle est effectivement fausse (voir schéma)

D - Inférence Statistique - Estimation et Tests d'hypothèses
3. Principe et résolution d'un test d'hypothèse

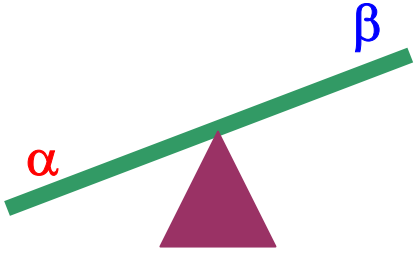
Les quatres résultats possibles d'un test d'hypothèses

Réalité	Décision	
	H_o acceptable	H_o rejetée
H_o est Vraie	***Décision correcte*** probabilité = (1 - α) seuil de confiance	***H_o rejetée à tord*** risque de 1ère espèce probabilité = α_o α seuil du test
H_o est fausse (H_1 est Vraie)	***H_o acceptée à tord*** risque de 2nde espèce probabilité = β	***Décision correcte*** probabilité = (1 - β) Puissance du test

⇨ **On peut accepter H_o jusqu'au seuil α_o**

La réduction α de entraîne l'accroissement de β !

⇨ **compromis** (voir transparent)



Analogie avec un procès : il n'y a pas de moyen de rendre α nul (assurance absolue qu'on n'accuse pas un innocent) sans que β atteigne 1 (relaxe de tout prévenu, ce qui rendrait le procès dénué de sens).

⇨ Une augmentation de la taille n de l'échantillon rends les distributions d'échantillonnage plus exactes. On peut alors réduire β sans augmenter α.

⇨ En toute rigueur le seuil α, la taille **n** de l'échantillon et l'écart entre les 2 valeurs impliquées dans les 2 hypothèses sont les 3 paramètres faisant varier β.

D - Inférence Statistique - Estimation et Tests d'hypothèses
3. Principe et résolution d'un test d'hypothèse

L'hypothèse nulle H_O permet de calculer des probabilités.
H_O est acceptable tant que l'on n'a pas acquis la quasi certitude qu'elle est fausse.

Si l'on peut démontrer que l'observation est improbable sous H_O celle-ci est rejetée, au profit de l'hypothèse alternative H_1, Sinon H_O est **acceptable** (on ne conclue pas).

En pratique, on divise la distribution en :

- ⇨ **domaine d'acceptation de H_O au seuil α,**
  région pour laquelle la V.A. d'étude peut prendre des valeurs, avec une probabilité 1-α, si H_O est vraie

- ⇨ **domaine de rejet de H_O au seuil α,**
  région pour laquelle la V.A. d'étude prend des valeurs improbables (petite probabilité α) si H_O est vraie

  Le seuil de rejet, limite entre ces deux régions, est la valeur (liée à α), **à partir de laquelle l'expérimentateur décide de rejeter H_O en faveur de H_1.**

H_O est acceptable tant que l'observation ne dépasse pas cette valeur seuil.
Le seuil de rejet est calculable par rapport à la distribution modèle considérée.

α mesure le risque que l'expérimentateur accepte de prendre en rejetant à tord H_O
β mesure le risque que l'expérimentateur accepte de prendre en acceptant à tord H_O

La valeur observée dans l'échantillon P_o est comparée au seuil de rejet P_{seuil}

Si P_o appartient au domaine d'acceptation de H_o , l'observation P_o est compatible avec la vraisemblance de H_o ⇨ l'observation n'est pas extraordinaire, elle ne permet pas d'affirmer avec une probabilité suffisamment élevée que H_o est fausse

Si P_o appartient au domaine de rejet de H_o, on préfère rejeter H_o , trop peu vraisemblable, en faveur de H_1 ⇨ L'observation nous semble suffisamment extraordinaire pour nous permettre de rejeter H_o avec une faible probabilité de se tromper

D - Inférence Statistique - Estimation et Tests d'hypothèses

3. Principe et résolution d'un test d'hypothèse

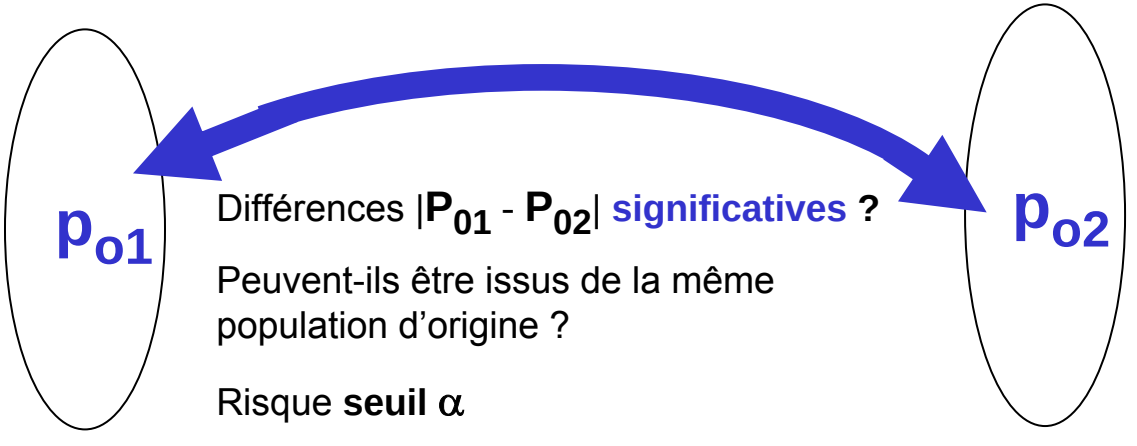
L'inférence et la déduction statistique aident l'expérimentateur à prendre une décision en considérant le risque d'erreur associé (1ère et 2ème espèce)

Autre préoccupation :

Test d'homogénéité concernant deux valeurs de proportions observées dans deux échantillons (tirés au hasard).

Tests d'homogénéité

On part des seules informations disponibles : p_{o1}, p_{o2} et n1, n2



Différences $|P_{01} - P_{02}|$ **significatives ?**

Peuvent-ils être issus de la même population d'origine ?

Risque **seuil** α

Risques α_o et β

Echantillon 1

- **taille : n1** (n1_échantillon)
- **représentatif**
- **P_{01} : observée**

Echantillon 2

- **taille : n2** (n2_échantillon)
- **représentatif**
- **P_{02} : observée**

Population d'origine commune ?

Test d'homogénéité de deux proportions observées

Il s'agit de confronter 2 valeurs de la variable aléatoire P_O, P_{O1} et P_{O2}, observées dans 2 échantillons (supposés provenir d'une population commune sous l'hypothèse nulle). Tout revient à déterminer si la différence observée entre les 2 valeurs est significative ou peut, au contraire, s'expliquer par le hasard des fluctuations d'échantillonnage.

On pose Ho : $\Pi_1 = \Pi_2$, soit : $(\Pi_1 - \Pi_2) = 0$ **H1** : $\Pi_1 \neq \Pi_2$ ***test bilatéral***

⇨ Egalité des deux paramètres exacts dans les deux populations d'origine ;
les différences observées dans les échantillons étant dues au hasard des fluctuations d'échantillonnage.

⇨ On est ainsi ramené à comparer la différence (P_{O1}-P_{O2}) à la valeur exacte 0.

Le critère de test utilisé est : $$\varepsilon_0 = \frac{(P_{01} - P_{02}) - 0}{\sqrt{P(1-P)\left(\frac{1}{n1} + \frac{1}{n2}\right)}}$$

(à condition que n1p >5 , n1(1-p) > 5 , n2p >5 et n2(1-p) > 5)

avec $\varepsilon_0 \sim N(0, 1)$ et $$P = \frac{n1.P_{01} + n2.P_{02}}{n1 + n2}$$

(la variance d'une différence est la somme des variances)

D - Inférence Statistique - Estimation et Tests d'hypothèses
3. Principe et résolution d'un test d'hypothèse

Le premier jugement que l'on tire d'un test qui permet de conclure est un **jugement de signification** (les différences sont significatives).

Le **jugement d'interprétation** permet, quant à lui, de déterminer si c'est la présence ou l'absence d'un caractère est à l'origine de ces différences. **Pour discuter de la causalité à l'issue d'un test, il faut s'assurer du contrôle d'un des deux caractères étudiés.**

Démarche / situation expérimentale

Lorsque l'expérience se conduit avec un facteur contrôlé (par l'expérimentateur).

Au cours de la constitution de l'échantillon, on reste libre d'un caractère (exemple : individu i traité ou non). On ne peut affirmer la causalité hors d'une **démarche expérimentale** qui est la seule qui permet d'assurer que les individus constituant l'échantillon sont comparables ou homogènes sauf pour ce qui concerne le caractère contrôlé (attribué par tirage au sort).

Démarche / situation d'observation

Lorsque l'expérience se conduit sur la base de deux facteurs aléatoires (non contrôlés par l'expérimentateur ; ex : surpoids et maladie rare, couleur des yeux et couleur cheveux, effets secondaires)
⇨ Il faudra faire des recherches (par ex sur les gènes) pour conclure plus finement.

Exemple typique : comparaison de deux traitements (un des traitement est la référence, constitution d'un groupe témoin recevant le traitement de référence) ou mise en évidence de l'effet d'un traitement (constitution d'un groupe témoin recevant un placebo. Procédure « d'aveugle » où les malades ne doivent pas savoir s'ils reçoivent le traitement ou le placebo. Procédure de « double aveugle » si l'attribution du traitement ou du placebo se fait en plus à l'insu du médecin).

D - Inférence Statistique - Estimation et Tests d'hypothèses
3. Principe et résolution d'un test d'hypothèse

Exercice " Traitement contre la grippe " (un grand classique!)

On veut comparer deux formes de traitement T1 et T2 contre la grippe. Dans ce but, on décide d'attribuer T1 ou T2 par **tirage au sort** aux 300 personnes qui se présentent à un service hospitalier. Après 5 jours de traitement, on constate que sur les 200 personnes qui ont suivi T1, 175 sont guéries et sur les 100 autres qui ont suivi T2, 70 sont guéries.

- **A -** Pensez-vous que la nature du traitement influe sur le taux de guérison en 5 jours ?
- **B -** Dans chacun des 2 groupes; on a noté le sexe de chaque individu.
  Les proportions de femmes observées sont les suivantes : 52 % dans le groupe T1 et 49 % dans le groupe T2. Au risque d'erreur 5%, la différence entre les 2 pourcentages observés est-elle significative ?
  Ce résultat pouvait-il être attendu ?
- **C -** Sachant que sur les 104 femmes ayant suivi le traitement T1, 90 sont guéries et que sur les 49 femmes ayant suivi le traitement T2, 40 sont guéries, pensez-vous que la guérison soit liée au sexe, quel que soit le traitement suivi?

Exercice *Traitement contre la grippe* **tests bilatéraux**

❍ **A –** P_{01} =175/200 = 0,875 ; P_{02} =70/100 = 0,700 ;

P=(200x0.875+100x0.7)/(200+100)=0.817

(la plus petite quantité à comparer à 5 étant 0.183x100=18.3 est bien > 5,
on peut appliquer l'approximation normale des fluctuations d'échantillonnage)

$$\varepsilon_0 = \frac{0.875 - 0.700}{\sqrt{0.817 \times 0.183\,(\frac{1}{200} + \frac{1}{100})}} = 3.695 \quad , \alpha_O = 2\text{x}0.0001$$

conclusion : si l'on pose α=5% et même α=1% ⇨ **H_O rejetée** ;

au seuil de 1% on peut dire, pratiquement sans aucun risque de se tromper (en rejetant H_o),
que **la nature du traitement influe sur le taux de guérison (jugement de signification)**;

les différences étant significatives (jugement de signification) jusqu'au seuil α=0.02 %
(⇨ auquel on peut encore rejeter de H_o)

Situation expérimentale

❍ **B –** P_{01} = 0,52 ; P_{02} = 0.49 ; P=0.51 ,

(la plus petite quantité à comparer à 5 étant 0.49x100=49 est bien > 5)

$$\varepsilon_0 = \frac{0.52 - 0.49}{\sqrt{0.51 x 0.49 \left(\frac{1}{200} + \frac{1}{100}\right)}} = 0.490 ; \quad \alpha_0 = 2x0.312 = 0.624$$

décision : α=5% ⇨ **H_o acceptable** ; (jusqu'à α=62.4%!)

La différence entre les 2 pourcentages n'est pas significative au seuil de 5 %

❍ **C –** Quelque soit le traitement suivit ⇨ On regroupe les données des 2 traitements T1 et T2. Les deux échantillons considérés sont les hommes et les femmes sélectionnés pour être traités (indifféremment parT1 ou T2).

P_{01} =130/153 = 0,850 (90+40)=130 femmes guéries sur (104+49)=153 femmes au total

P_{02} =115/147= 0,782 (175+70)-(90+40)=115 hommes guéris sur (300-104+49)=147 hommes

P=0.817 , ε_o=1.52 ; α_o=0.064x2=0.128 (la table de la loi normale centrée réduite donne α_o/2)

(la plus petite quantité à comparer à 5 étant 0.183x100=18.3 est bien > 5)

$$\varepsilon_0 = \frac{0.850 - 0.782}{\sqrt{0.817 x 0.183 \left(\frac{1}{153} + \frac{1}{147}\right)}} = 1.52 ; \quad \alpha_0 = 2x0.064 = 0.128$$

décision : α=5% ⇨ **H_o acceptable** (jusqu'à 12.8 %!!);

Au seuil de 5 %, les différences ne sont pas assez significatives pour que l'on puisse dire que le taux de guérison soit lié au sexe quelque soit le traitement suivi

D - Inférence Statistique - Estimation et Tests d'hypothèses
4. Fluctuations d'échantillonnage d'une moyenne observée

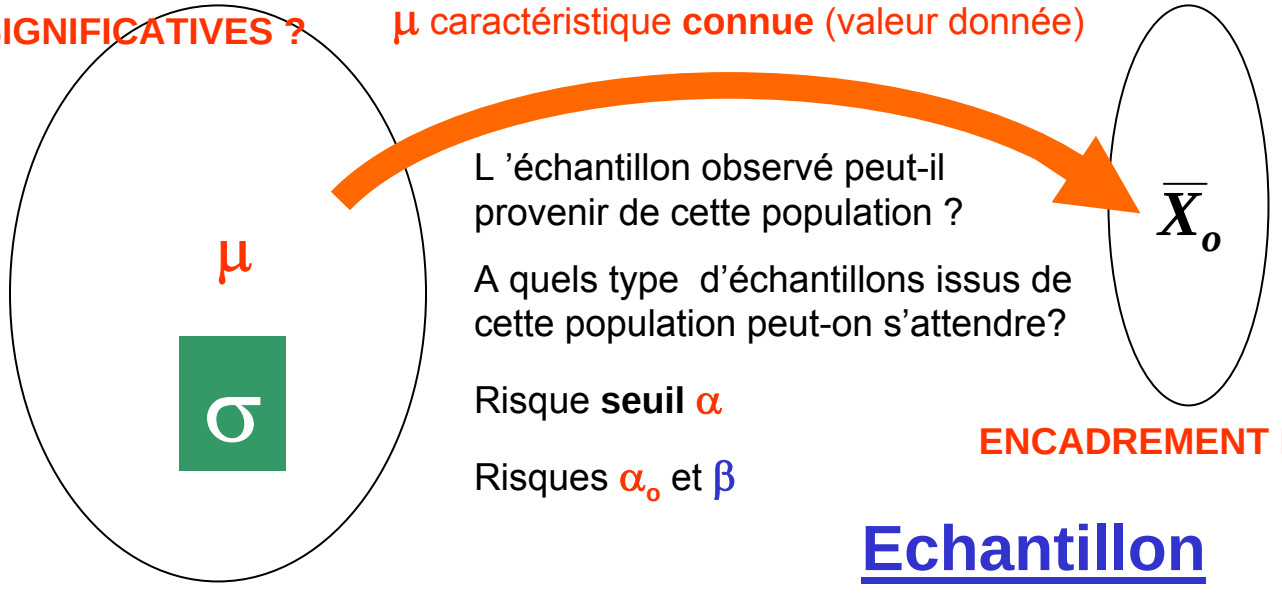
Tests de conformité

Intervalle de pari

DIFFERENCES ($\overline{X}o$-μ) SIGNIFICATIVES ?

Hypothèse nulle Ho :

μ caractéristique **connue** (valeur donnée)



L 'échantillon observé peut-il provenir de cette population ?

A quels type d'échantillons issus de cette population peut-on s'attendre?

Risque **seuil** α

Risques α_o et β

ENCADREMENT DE $\overline{X}o$

Echantillon

- **taille : n** (n_échantillon)
- **représentatif**
- $\overline{X}_o$ **: observée**

Population

- **taille ?**
- **Inaccessible**
- μ **: caractéristique connue ou supposée connue**
- (résultant par ex d'études menées sur de nombreux échantillons)

Plusieurs cas possibles …

- La variable étudiée dans la population ou l'échantillon est X
- La variable d'échantillonnage est sa moyenne $\overline{X_0}$:
- Loi suivie par $\overline{X_0}$: il sera nécessaire de vérifier que $\overline{X_0}$ (le cas échéant X) est distribuée normalement.
- Connaissance de σ (population) ?
- Théorème « central limit »

D - Inférence Statistique - Estimation et Tests d'hypothèses
4. Fluctuations d'échantillonnage d'une moyenne observée

Dans tous les tests sur la moyenne,

il est souvent utile de démontrer que $\overline{X}_0$ suit une loi normale

Théorème « central limit »

Pour appliquer le théorème de la limite centrale il faut que σ soit connu!

Si X, variable aléatoire d'étude, est distribuée de façon quelconque dans une population de moyenne μ et écart type σ, paramètres exacts **connus**,

$\overline{X}_0$, moyenne observée dans un échantillon de taille n, prélevé dans cette population tend vers une distribution normale de même moyenne μ et d'écart-type $\sqrt{\sigma^2/n}$.

⇨ Lorsque la loi de distribution suivie par la variable d'étude X est normale dans la population, $\overline{X}_0$ est réellement distribuée normalement

⇨ Lorsque la loi de distribution suivie par la variable d'étude X n'est pas connue dans la population, on peut considérer que $\overline{X}_0$ est distribuée normalement, dès que $n \geq 30$

⇨ Si $n < 30$: il est parfois possible d'émettre l'hypothèse que la variable d'étude X est supposée distribuée normalement dans la population (souvent le cas en biologie) ou au moins que la distribution est symétrique. Dans le cas contraire, on ne pourra pas traiter le problème en utilisant la loi normale.

Cas où σ, écart type de la population, est connu

Intervalle de pari ,

au risque α , encadrant les valeurs de $\bar{x}_0$ observables dans un échantillon tiré au hasard d'une population caractérisée par le paramètre exact μ (*caractéristique*) :

$$\mu - \varepsilon_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq \bar{X}_0 \leq \mu + \varepsilon_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

La probabilité de trouver $\bar{x}_0$ dans cet intervalle est (1-α)

Seuil standard : α = 5 %

Exemple introductif *"Simulation d'échantillonnage"*

Une expérience d'échantillonnage consiste à simuler le prélèvement de 100 échantillons de 20 individus dans une population de sujets âgés de moins de 30 ans.

Chez ces sujets, le taux d'une enzyme y est supposée distribuée normalement

avec une moyenne exacte μ = 1 mg/ml et un écart-type σ = 0,3 mg/ml.

1) Quel est l'intervalle de pari sur $\overline{X}_0$ au risque α = 1% ?
2) Sur les 100 valeurs observées, combien de valeurs observées de la moyenne vous attendez-vous à trouver dans cet intervalle ?
3) Donnez l'intervalle de pari de la variable centrée et réduite au risque α = 1%.
4) Sur les 100 ICα construits au risque α, combien en moyenne recouvrent la valeur exacte μ = 1 mg/ml ?

D - Inférence Statistique - Estimation et Tests d'hypothèses
4. Fluctuations d'échantillonnage d'une moyenne observée

Intervalle de pari de la variable d'échantillonage $\overline{X}_0$

Cas où σ, écart type de la population, n'est <u>pas connu</u>

<u>Intervalle de pari</u> ,

au risque α , encadrant les valeurs de $\overline{X}_0$ observables dans un échantillon tiré au hasard d'une population caractérisée par le paramètre exact μ (*caractéristique*) :

$$\mu - t_{\alpha/2}\sqrt{\frac{\sigma_0^{\,2}}{n}} \leq \overline{X}_0 \leq \mu + t_{\alpha/2}\sqrt{\frac{\sigma_0^{\,2}}{n}}$$

La probabilité de trouver $\overline{X}_0$ dans cet intervalle est (1-α)

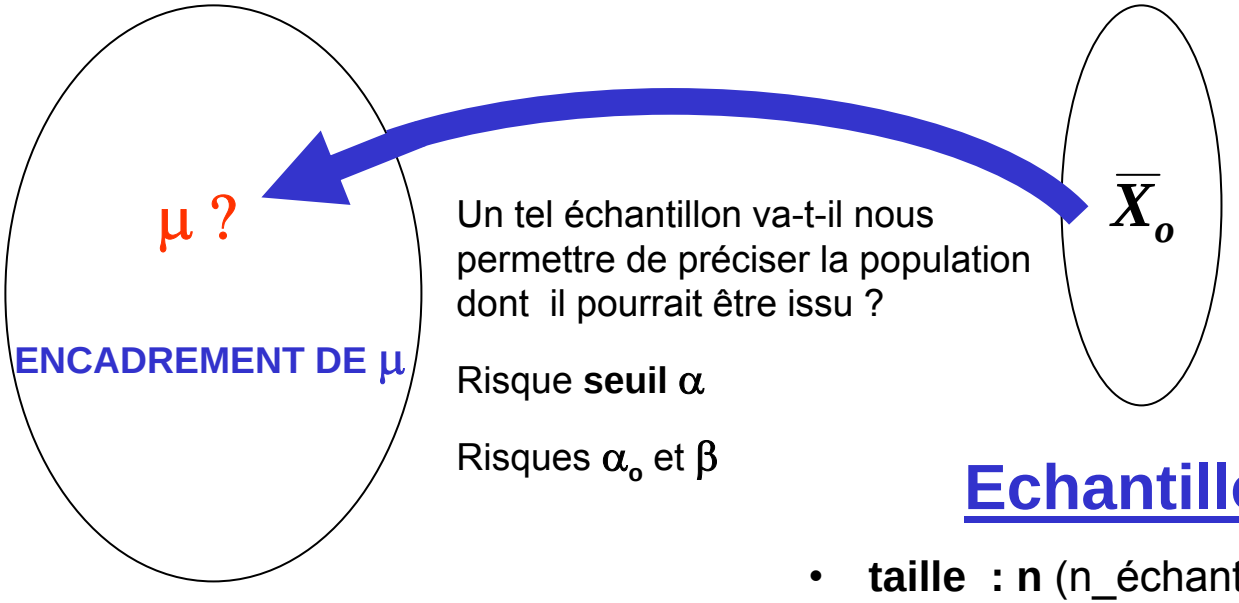
[seuil standard : α = 5 %]

<u>On se sert des tables de la loi de Student à n-1 ddl</u>

<u>Condition d'application</u> :
(Il faut que la V.A. d'étude **X** suive une loi normale dans la population)

Intervalles de confiance

On part des seules informations disponibles : $\overline{X}_o$ et n



Un tel échantillon va-t-il nous permettre de préciser la population dont il pourrait être issu ?

Risque **seuil** α

Risques α_o et β

Echantillon

- **taille : n** (n_échantillon)
- **représentatif**
- $\overline{X}_o$ **: observée**

Population

- **taille ?**
- **Inaccessible**
- μ **: caractéristique théorique ou attendue**

Cas où σ, écart type de la population, est connu

Situation rare car lorsque μ est inconnue, σ est généralement inconnu lui aussi !

Intervalle de confiance , au risque α , encadrant les valeurs attendues pour le paramètre exact μ (*caractéristique*) , à partir de la connaissance de $\overline{X}_0$ observée dans un échantillon supposé représentatif :

$$\overline{x}_0 - \varepsilon_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{x}_0 + \varepsilon_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

La probabilité de trouver μ dans cet intervalle est $(1-\alpha)$

Seuil standard : α = 5 %

⇨ On définit ainsi plusieurs populations pouvant rendre compte de l'échantillon observé.

Cas où σ, écart type de la population, est inconnu

Intervalle de confiance , au risque α , encadrant les valeurs attendues pour le paramètre exact μ (*caractéristique*) , à partir de la connaissance de $\overline{X}_0$ observée dans un échantillon supposé représentatif :

$$\overline{X}_0 - t_{\alpha/2}\sqrt{\frac{\sigma_0^2}{n}} \leq \mu \leq \overline{X}_0 + t_{\alpha/2}\sqrt{\frac{\sigma_0^2}{n}}$$

La probabilité de trouver μ dans cet intervalle est $(1-\alpha)$

Seuil standard : α = 5 %

⇨ On définit ainsi plusieurs populations pouvant rendre compte de l'échantillon observé.

Test de conformité à une moyenne exacte μ

dans le cas où <u>σ, écart type de la population est connu</u>

Critère de test : $$\varepsilon_0 = \frac{(\overline{X}_0 - \mu)}{\sqrt{\frac{\sigma^2}{n}}}$$, suivant la loi normale centrée réduite

⇨ **Même raisonnement que pour l'étude des fluctuations d'échantillonnage d'une proportion observée**

Exemple introductif ***"Vers une vie éternelle"***

Un laboratoire de recherche étudie, sur une nouvelle espèce de vers *(C. marginalus)*, les gènes pouvant être impliqués dans la mort programmée. Chez ces vers, la durée de vie est caractérisée par une moyenne exacte μ = 250 heures et un écart-type exact σ = 24 heures.

Après **modification d'un gène**, supposé intervenir dans la mort programmée, on a relevé dans un échantillon de 80 vers une durée de vie moyenne de 263 heures et un écart type de 27 heures.

Que peut-on en conclure?

Situation d'expérience

L'écart-type de l'échantillon n'est pas une donnée utilisée

D - Inférence Statistique - Estimation et Tests d'hypothèses

4. Fluctuations d'échantillonnage d'une moyenne observée

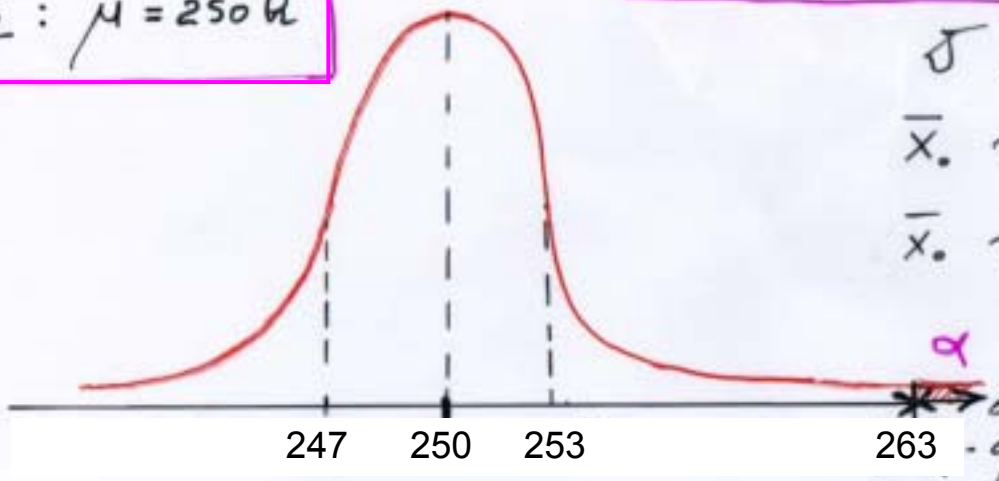
$H_0 : \mu = 250$ h

Test de conformité à une moyenne exacte

σ : connu

$\bar{X}_0 \rightsquigarrow N(\mu, \sigma/\sqrt{n})$

$\bar{X}_0 \rightsquigarrow N(250, 2{,}7)$



$\bar{X}_0$ durée de vie moyenne observée en heure. [espèce de vers] dans un échantillon de 80 vers

H1 : $\mu > 250$ heures

test unilatéral

$$\varepsilon_0 = \frac{\bar{X}_0 - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

Test unilatéral

$$\varepsilon_0 = \frac{263 - 250}{\sqrt{\frac{576}{80}}}$$

$$\varepsilon_0 = 4{,}845 !$$

on a découvert un gène impliqué dans le processus de mort programmée

$\alpha_0 = 6 \cdot 10^{-7} \ll 5\%$

conclusion

H_0 rejetée avec un risque de 1ère espèce très faible, quasiment nul

⇒ La modification du gène a induit des différences SIGNIFICATIVES

Test de conformité à une moyenne exacte μ,

Cas où σ, écart type de la population, n'est pas connu

Dans ce cas, on se base sur la seule connaissance de σ_O (calculé à partir des données de l'échantillon). Contrairement à σ^2/n, σ_O^2/n n'est plus une constante mais une variable aléatoire suivant une loi de distribution de **Student**. Le critère de test utilisé sera t_o, à (n-1) degré de liberté. On se servira de la **table** échantillonnée **de Student** pour trouver les valeurs de la probabilité α_O connaissant t_o ou celles de t_o connaissant α_O. La distribution de Student est symétrique ; elle pourra dans la pratique être approchée, avec peu d'erreur, par une distribution normale lorsque n est suffisamment grand (n> 30, théorème central limite).

Critère de test :

$$t_0 = \frac{(\overline{X}_0 - \mu)}{\sqrt{\frac{\sigma_0^2}{n}}}$$

Condition d'application :
(Il faut que la V.A. d'étude **X** suive une loi normale dans la population)

$\mathbf{t_0}$ suivant une loi de **Student** à n-1 degrés de libertés (ddl)

D - Inférence Statistique - Estimation et Tests d'hypothèses
4. Fluctuations d'échantillonnage d'une moyenne observée

Exemple introductif ***"Vers une vie éternelle"***

Un laboratoire de recherche étudie, sur une nouvelle espèce de vers *(C. marginalus)*, les gènes pouvant être impliqués dans la mort programmée. Chez ces vers, la durée de vie est caractérisée par une moyenne exacte μ = 250 heures.

Après **modification d'un gène**, supposé intervenir dans la mort programmée, on a relevé dans un échantillon de 80 vers une durée de vie moyenne de 263 heures et un écart type de 27 heures.

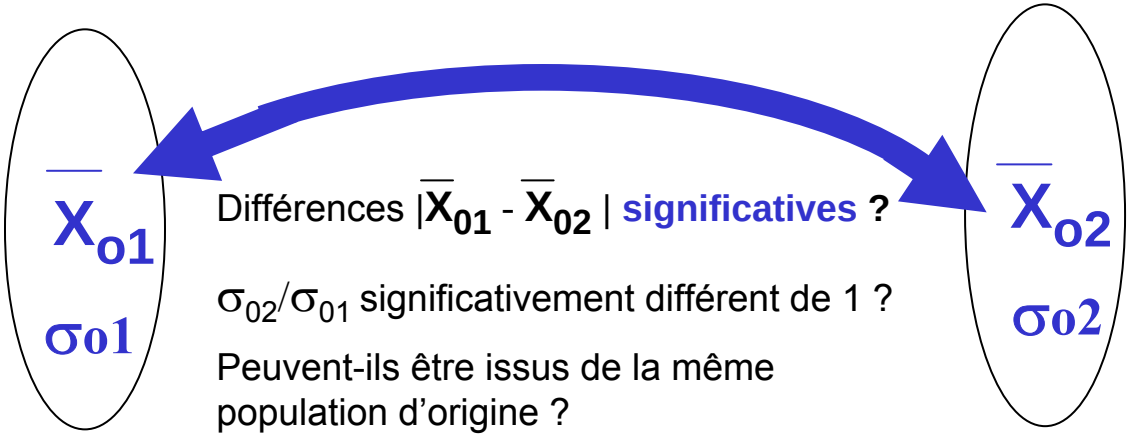
- Que peut-on en conclure?
- Même question si l'observation avait été faite sur un échantillon de 25 vers ?

Situation d'expérience

L'écart-type de l'échantillon est utilisé en absence de la connaissance de σ (population) => test t

Tests d'homogénéité

On part des seules informations disponibles : $\overline{X}_{o1}$, $\overline{X}_{o2}$ et n1, n2



Différences $|\overline{X}_{01} - \overline{X}_{02}|$ **significatives ?**

σ_{02}/σ_{01} significativement différent de 1 ?

Peuvent-ils être issus de la même population d'origine ?

Risque **seuil** α

Risques α_o et β

Echantillon 1

- **taille : n1** (n1_échantillon)
- **représentatif**
- **X_{01} : observée**

Echantillon 2

- **taille : n2** (n2_échantillon)
- **représentatif**
- **X_{02} : observée**

Existe-t-il une Population d'origine commune ?

Test de comparaison de 2 variances observées

Hypothèse nulle **H_o : $\sigma_1^2 = \sigma_2^2$**

⇨population commune / différences dues au hasard

On se servira de la **table** échantillonnée **de Fisher** pour trouver les valeurs de la probabilité α_o connaissant F ou celles de F connaissant α_o .

Critère de test :

$$F = \frac{\sigma_{01}^2}{\sigma_{02}^2}$$

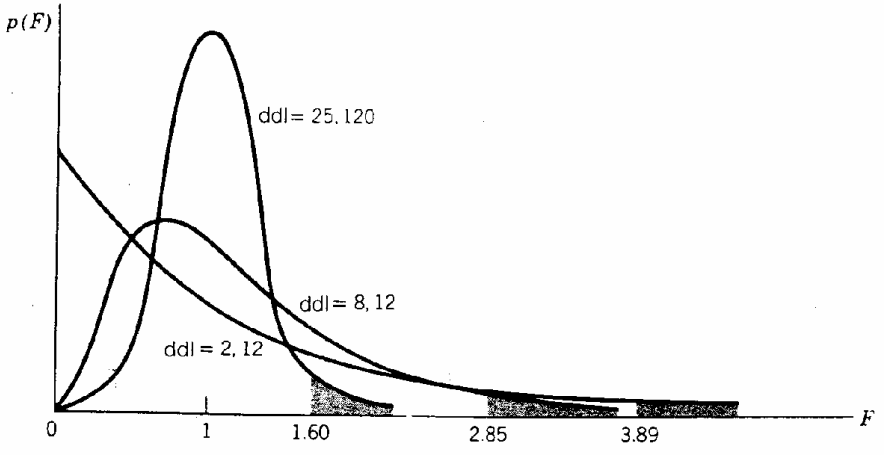
F suivant une loi de **Fisher** à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de libertés (ddl)

Test F

Pour mesurer la crédibilité de Ho, on calcule sa probabilité critique (probabilité de la queue de distribution de F située au delà de la valeur observée).

Lecture de la table de Fisher dépendant des degrés de libertés de la variance du numérateur (n_1-1) et de ceux de la variance du dénominateur (n_2-1)

Table F à double entrée ddl : colonne n_1-1 et lignes n_2-1



Quelques distributions particulières de *F*, avec divers d.d.l. au numérateur et au dénominateur. A noter comment le point critique de 5 % (au-delà duquel H_0 est habituellement rejetée) se déplace vers la gauche, vers la valeur 1, quand d.d.l. augmente.

Test d'homogénéité de 2 moyennes observées

Cas où n_1 et n_2 sont tous deux supérieurs à 30

Ho : $\mu_1 - \mu_2 = 0$

⇨population d'origine commune / différences dues au hasard

On se servira de la **table** échantillonnée **de la loi normale centrée réduite** pour trouver les valeurs de la probabilité α_O connaissant ε_o ou celles de ε_o connaissant α_O .

Critère de test :

$$\varepsilon_0 = \frac{(\overline{X}_{01} - \overline{X}_{02}) - 0}{\sqrt{\frac{\sigma_{01}^2}{n_1} + \frac{\sigma_{02}^2}{n_2}}}$$

ε_O suivant la loi N (0,1)

D - Inférence Statistique - Estimation et Tests d'hypothèses

5. Comparaison de 2 variances observées / de 2 moyennes observées

Test d'homogénéité de 2 moyennes observées

Petits échantillons n_1 ou/et n_2 inférieur(s) à 30

Ho : $\mu_1 - \mu_2 = 0$

⇨population d'origine commune / différences dues au hasard

Critère de test :

$$t_0 = \frac{(\overline{X}_{01} - \overline{X}_{02}) - 0}{\sqrt{\sigma_0^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

On se sert de la **table** échantillonnée **de la loi de student** pour trouver les valeurs de la probabilité α_o connaissant t_o ou celles de t_o connaissant α_o.

Attention : le nombre de degrés de liberté est ici : (n_1+n_2-2) ddl

⇨**Test F préalable** sur les variances observées pour pouvoir calculer la variance commune si elle existe (Ho : $\sigma_1=\sigma_2$, hypothèse d'égalité des variances des 2 populations, acceptable)

L'approximation de la variance supposée commune des variables étudiées est alors :

$$\sigma_0^2 = \frac{(n_1 - 1)\sigma_{01}^2 + (n_2 - 1)\sigma_{02}^2}{(n_1 - 1) + (n_2 - 1)}$$

Si le test F échoue, on ne peut pas conclure à l'égalité des 2 moyennes car le test t ne peut être réalisé !

Exercice **"Une production machine-dépendante ?"**

On veut comparer la production de deux machines artisanales. Le volume de parfum déposé par ces machines dans des flacons destinés à la vente a été mesuré en ml. Les statistiques figurent ci-dessous. Le laborantin qui en est responsable soupçonne une production dépendante de la machine.

Machine 1	Machine 2
47	54
53	50
49	51
50	51
46	49

Qu'en pensez-vous?

Exercice **"Une production machine-dépendante ?"**

$\overline{X}_{o1}$=49 et $\overline{X}_{o2}$=51 Attention, il s'agit de petits échantillons :

les tailles des 2 échantillons sont inférieures à 30.

⇨ **Test F** préalable comparant les variances observées

Ho : $\sigma^2_1/\sigma^2_2 = 1$ (ou $\sigma^2_2/\sigma^2_1 = 1$);

si les variances sont les mêmes (même population d'origine),
le rapport des deux variances observées doit tendre vers 1.

σ^2_{o1} = [(47-49)^2+(53-49)^2+(49-49)^2+(50-49)^2+(46-49)^2] / (5-1) = 30/4 = 7.5

σ^2_{o2} = 14/4 = 3.5

Fo = 7.5/3.5 = 2.14

Cette valeur est plus petite que $\mathbf{F_{0.10}}$ (4.11) lue sur la **table de Fisher**
à 4 ddl au numérateur (intervenant dans le calcul de la première variance)
et 4 ddl au dénominateur (intervenant dans le calcul de la deuxième variance).
Ho est donc largement acceptable (jusqu'au risque seuil de 10 % et même presque 25 %)
et l'on peut calculer une variance commune :

$$\sigma_0^2 = \frac{4x7.5 + 4x3.5}{(5-1)+(5-1)} = 5.5$$

Exercice **"Une production machine-dépendante ?"**
(suite)

Le test sur les variance étant positif, on peut poursuivre le test sur les moyennes

Test t sur les deux moyennes observées

(on supposera que la production est distribuée normalement dans les 2 populations)

Ho : $\mu 1 - \mu 2 = 0$

$$t_0 = \frac{51 - 49}{\sqrt{5.5\,(\frac{1}{5} + \frac{1}{5})}} = 1.35 \text{ ; à } (5 + 5 - 2) = 8 \text{ ddl}$$

(la variance commune calculée est injectée dans le calcul de t_o)

$t_o < 1.86$ (à 8 ddl)

(test bilatéral, il y autant de chance d'observer des écarts positifs que des écarts négatifs)

$\alpha o > 10\%$

Ho est acceptable, au moins jusqu'au seuil de 10%, et même 20 % (to pour 20% est 1.39)

les différences sont imputables au hasard des fluctuations d'échantillonnage.

On ne peut affirmer sans risque de se tromper que la production est machine dépendante.

Exercice **"Une production machine-dépendante ?"**

On veut comparer maintenant la production de la première machine avec celle d'une autre machine. Les statistiques figurent ci-dessous. Le laborantin soupçonne toujours une production dépendante de la machine.

Machine 1	Machine 3
47	55
53	54
49	58
50	61
46	52

Qu'en pensez-vous?

Exercice **"Une production machine-dépendante ?"**

$\overline{X}o_1$=49 et $\overline{X}o_2$=56 Attention, il s'agit de petits échantillons :

les tailles des 2 échantillons sont inférieures à 30.

⇨ **Test F** préalable comparant les variances observées

Ho : $\sigma_1/\sigma_2 = 1$; si les variances sont les mêmes (même population d'origine), le rapport des deux variances observées doit tendre vers 1.

σ^2_{o1} = [(47-49)^2+(53-49)^2+(49-49)^2+(50-49)^2+(46-49)^2] / (5-1) = 30/4 = 7.5

σ^2_{o2} = 50/4 = 12.5

Fo = 12.5/7.5 = 1.67

Cette valeur est plus petite que $F_{0.25}$ (2.06) lue sur la **table de Fisher** à 4 ddl au numérateur (intervenant dans le calcul de la première variance) et 4 ddl au dénominateur (intervenant dans le calcul de la deuxième variance).

Ho est donc largement acceptable (jusqu'au risque seuil de 25 %) et l'on peut calculer une variance commune :

$$\sigma_0^2 = \frac{4x7.5 + 4x12.5}{(5-1)+(5-1)} = 10$$

Exercice **"Une production machine-dépendante ?"**
(suite)

Test t sur les deux moyennes observées
(on supposera que la production est distribuée normalement dans les 2 populations)

Ho : $\mu 1 - \mu 2 = 0$

$$t_0 = \frac{56 - 49}{\sqrt{10\left(\frac{1}{5} + \frac{1}{5}\right)}} = \frac{7}{2} = 3.5 \text{ ; à } (5 + 5 - 2) = 8 \text{ ddl}$$

(la variance commune calculée est injectée dans le calcul de t_o)

$t_o > 2.31$ (à 8 ddl)
(test bilatéral, il y autant de chance d'observer des écarts positifs que des écarts négatifs)
$\alpha o < 5\%$ (et même $\alpha o < 0.1\%$!!!)

Ho est rejetée, au seuil de 0.1%,
les différences sont significatives au seuil de 1 pour mille.
Au seuil de 0.1% on peut dire que la production est machine dépendante.

Comparaison de 2 moyennes sur séries appariées

Cas où n est supérieur à 30

Ho : $\mu = 0$

⇨population d'origine commune / différences dues au hasard

Critère de test :

$$\varepsilon_0 = \frac{\bar{d} - 0}{\sqrt{\frac{\sigma_0^2}{n}}}$$

On se sert de la **table** échantillonnée **de la loi normale centrée réduite** pour trouver les valeurs de la probabilité α_o connaissant ε_o ou celles de ε_o connaissant α_o.

On utilise la variance expérimentale des différences , n est le nombre de paires

$$\sigma^2_0 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

Permet de contrôler le paramètre toutes choses égales par ailleurs
⇨ évite les effets de variances inter échantillon (puisqu'on travaille avec le même échantillon)

Comparaison de 2 moyennes sur séries appariées

Petits échantillons n inférieur à 30

Ho : $\mu = 0$

⇨population d'origine commune / différences dues au hasard

Critère de test :

$$t_0 = \frac{\bar{d} - 0}{\sqrt{\frac{\sigma_0^2}{n}}}$$

On se sert de la **table** échantillonnée **de la loi de student** pour trouver les valeurs de la probabilité α_o connaissant t_o ou celles de t_o connaissant α_o.

Attention : le nombre de degrés de liberté est ici : (n_1-1) ddl

Validité : la variable d doit être normale dans la population d'origine

Permet de contrôler le paramètre toutes choses égales par ailleurs
⇨ évite les effets de variances inter échantillon (puisqu'on travaille avec le même échantillon)

Comparaison de 2 moyennes sur séries appariées

N° de l'animal	Taux de l'enzyme avant traitement [mg/ml]	Taux de l'enzyme au cours du traitement [mg/ml]	différences d_i
1	2.75	2.36	-0.39
2	3.14	2.88	-0.26
3	3.01	2.62	-0.39
4	3.4	3.14	-0.26
5	3.53	3.01	-0.52
6	3.14	2.75	-0.39
7	3.01	2.75	-0.26
8	3.66	3.01	-0.65
9	3.27	2.88	-0.39
10	2.88	2.62	-0.26
11	3.92	3.27	-0.65
12	3.66	3.14	-0.52
13	2.75	2.36	-0.39
14	3.14	2.88	-0.26
15	3.01	2.62	-0.39
		moyenne	-0.40
		écart-type	0.13
	Test d'homogénéité :		
		t_0	11.50
		α	5.0%
	Ho rejetée au seuil 5%	α_0	0.0%
		ddl	14