Module 17

MATHEMATIQUES

C08 - PROBABILITES – STATISTIQUES No 8

Statistiques inductives

- **Quelles relations entre les caractéristiques d'une population de plusieurs milliers d'éléments et celle d'un échantillon de quelques dizaines extrait au hasard de cette population ?**

Retour sur quelques distributions

- Quelles autres distributions utiles, hormis Bernouilli (Loi binomiale), Poisson et Gauss (Loi normale)
  - Distribution Gamma
  - Distribution exponentielle
  - Distribution du Chi Carré

Distribution Gamma

- Dans l'étude de la durée de vie d'un équipement industriel ainsi que dans d'autres domaines, on rencontre souvent la distribution Gamma (du nom de la fonction mathématique Gamma)

$$f(x) = \frac{x^{\alpha - 1}\ e^{-(x/\beta)}}{\beta^{\alpha}\ \Gamma(\alpha)} \quad \text{avec } X>0$$

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha - 1}\ e^{-x}\, dx$$

Distribution Gamma

Le raisonnement mathématique (intégration par partie) démontre :

$$\Gamma(\alpha + 1) = \alpha\, \Gamma(\alpha)$$

$$\Rightarrow \Gamma(\alpha + 1) = \alpha\,! \text{ (Fonction factorielle)}$$

On démontre que la moyenne et la variance de la distribution Gamma :

$\mu = E(X) = \beta\, \alpha$ $\qquad$ $\sigma^2 = E(X^2) - \mu^2 = \beta^2\, \alpha$

Distribution Exponentielle

La distribution exponentielle est un cas particulier de la fonction Gamma où $\alpha = 1$

$$f(x) = \frac{e^{-(x/\beta)}}{\beta}$$

Loi de probabilité exponentielle

- La variable aléatoire exponentielle X, encore appelée variable $\gamma 1$, est une variable continue pouvant prendre n'importe quelle valeur entre 0 et l'infini avec la densité de probabilité $f(x)= ce^{-cx}$
- c est un paramètre positif qui est égal à l'inverse de la moyenne de la distribution
- La fonction de répartition :

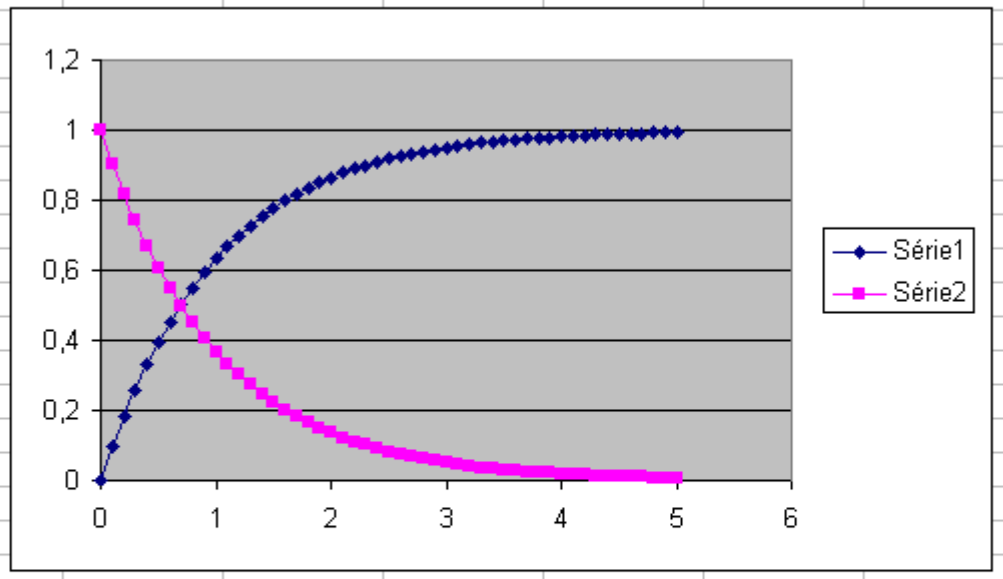
$$F(X) = \int_0^x ce^{-cx} dx = 1 - e^{-cx}$$

Loi de probabilité exponentielle réduite

- Soit le changement de variable $Z=cX$
- Z est aussi une variable aléatoire exponentielle, de paramètre c=1
- L'espérance mathématique de la variable exponentielle réduite est égale à 1

Loi de probabilité exponentielle réduite

0	0	1
0,1	0,09516258	0,90483742
0,2	0,18126925	0,81873075
0,3	0,25918178	0,74081822
0,4	0,32967995	0,67032005
0,5	0,39346934	0,60653066
0,6	0,45118836	0,54881164
0,7	0,5034147	0,4965853
0,8	0,55067104	0,44932896
0,9	0,59343034	0,40656966
1	0,63212056	0,36787944
1,1	0,66712892	0,33287108
1,2	0,69880579	0,30119421
1,3	0,72746821	0,27253179
1,4	0,75340304	0,24659696
1,5	0,77686984	0,22313016
1,6	0,79810348	0,20189652
1,7	0,81731648	0,18268352
1,8	0,83470111	0,16529889
1,9	0,85043138	0,14956862
2	0,86466472	0,13533528
2,1	0,87754357	0,12245643
2,2	0,88919684	0,11080316
2,3	0,89974116	0,10025884
2,4	0,90928205	0,09071795
2,5	0,917915	0,082085
2,6	0,92572642	0,07427358
2,7	0,93279449	0,06720551
2,8	0,93918994	0,06081006
2,9	0,94497678	0,05502322
3	0,95021293	0,04978707



Distribution Exponentielle

On a déterminé que la distribution exponentielle est un modèle approprié pour calculer la probabilité qu'une machine fonctionne convenablement pendant une durée totale de t unités de temps avant de tomber en panne.

Distribution Exponentielle

Un fabricant de matériel électronique sait par expérience que son matériel fonctionne en moyenne 2 ans sans réparation et que la durée avant d'atteindre la première panne suit une distribution exponentielle. S'il garantit son matériel pour une durée d'un an, quelle proportion de ses clients devra-t-il dépanner si ces pannes se produisent pendant la première année ?

Distribution Exponentielle

Un fabricant de matériel électronique sait par expérience que son matériel fonctionne en moyenne 2 ans sans réparation et que la durée avant d'atteindre la première partie suit une distribution exponentielle. S'il garantit son matériel pour une durée d'un an, quelle proportion de ses clients devra-t-il dépanner si ces pannes se produisent pendant la première année ?

Puisque $\beta = 2$ est la moyenne de la distribution exponentielle, la densité qui s'applique dans ce cas est $f(x) = e^{-x/2}/2$

Calculons P(X<1). En posant t = x/2

$$P(X < 1) = \int_0^1 \frac{e^{-X/2}}{2} dx = \int_0^{1/2} e^{-t} dt = 0{,}39$$

ou Loi.exponentielle(1;0,5;vrai)
ou Loi.exponentielle(0,5;1;vrai)

Même si la durée de vie moyenne est le double de la durée de vie garantie, la probabilité que l'équipement tombe en panne avant l'expiration de la garantie est forte

Distribution Exponentielle

Une autre application intéressante de la distribution exponentielle est en relation avec la distribution de Poisson.

On peut démontrer que, si le nombre de réalisations d'un évènement dans une unité de temps suit une distribution de Poisson de paramètre μ, alors le temps entre deux réalisations successives de l'évènement se distribue selon une loi exponentielle de paramètre $\beta = 1/\mu$

Distribution Exponentielle

Le nombre moyen de clients qui se présentent à une caisse d'un supermarché sur un intervalle de 5 minutes est de 10.

On suppose que le nombre de clients suit une distribution de Poisson.

Quelle est la probabilité qu'aucun client ne se présente à une caisse dans un intervalle de 2 minutes ?

Distribution Exponentielle

Le nombre moyen de clients qui se présentent à une caisse d'un supermarché sur un intervalle de 5 minutes est de 10.

On suppose que le nombre de clients suit une distribution de Poisson.

Quelle est la probabilité qu'aucun client ne se présente à une caisse dans un intervalle de 2 minutes ?

Puisque μ = 10 clients dans un intervalle de 5 minutes, μ = 2 clients dans un intervalle d'une minute.

La moyenne de la distribution exponentielle qui donne le temps en minutes entre les arrivées est donnée par $\beta = 1/\mu = 1/2$

La distribution exponentielle associée $f(x) = 2\,e^{-2x}$ avec $x > 0$

où x représente la durée en minutes entre des arrivées successives.
En effectuant la substitution t = 2 x

$$P(X >= 2) = \int_{2}^{\infty} 2e^{-2x}dx = \int_{4}^{\infty} e^{-t}dt = e^{-4} = 0{,}018$$

ou 1-Loi.exponentielle(2;2;vrai)
ou 1-Loi.exponentielle(4;1;vrai)

Distribution Exponentielle

$$P(X >= 2) = \int_{2}^{\infty} 2e^{-2x}dx = \int_{4}^{\infty} e^{-t}dt = e^{-4} = 0{,}018$$

Ce résultat se traduit : si le nombre moyen de clients qui se présentent dans ce supermarché sur une journée est 960, le nombre moyen de creux d'au moins 2 minutes (périodes sans clients) est de 960 * 0,018 = 17

Distribution du Chi Carré

La distribution du chi carré est un autre cas particulier de la distribution gamma avec de nombreuses applications statistiques

On l'obtient en choisissant $\beta = 2$ et en écrivant $\alpha = \nu / 2$

$$f(x) = \frac{x^{(\nu / 2) - 1} e^{-(x/2)}}{2^{(\nu / 2)} \Gamma(\nu / 2)} \text{ avec } x>0$$

Distribution du Chi Carré

Cette transformation est associée à divers problèmes statistiques

ν est appelé nombre de degrés de liberté

On peut calculer la moyenne et la variance d'une variable du chi carré

En posant $\beta = 2$ et $\alpha = \nu / 2$

dans les formules

$\mu = E(X) = \beta\ \alpha$ $\qquad$ $\sigma^2 = E(X^2) - \mu^2 = \beta^2\ \alpha$

$\mu = \nu$ $\qquad$ $\sigma^2 = 2\ \nu$

Exercice sur la distribution du Chi Carré

Dans un centre de traitement informatique, une armoire contient 10 unités centrales identiques qui fonctionnent H24 J7

Ces appareils sont susceptibles de tomber en panne de manière aléatoire.

Au cours d'une période de 40 semaines, on a observé le nombre d'appareils tombés en panne chaque semaine. On a obtenu la statistique suivante :

Nombre d'unités tombées en panne au cours d'une semaine (xi)	Nombre de semaines ni
0	11
1	10
2	7
3	6
4	4
5	2
6 et plus	0
Total	40

Exercice sur la distribution du Chi Carré

Nombre d'unités tombées en panne au cours d'une semaine (xi)	Nombre de semaines ni
0	11
1	10
2	7
3	6
4	4
5	2
6 et plus	0
Total	40

Ajuster cette distribution observée par une loi de Poisson

- En déterminant la valeur du paramètre de la loi
- En dressant un tableau des valeurs théoriques en regard des valeurs observées

Testez la validité de cet ajustement par la méthode du χ^2

Exercice sur la distribution du Chi Carré

La loi de Poisson de l'ajustement aura comme paramètre la moyenne qui sera égale à la moyenne $\bar{x}$ de la distribution observée.

Nombre d'unités tombées en panne au cours d'une semaine (xi)	Nombre de semaines ni	Nombre de semaines ni
0	11	0
1	10	10
2	7	14
3	6	18
4	4	16
5	2	10
6 et plus	0	0
Total	40	68

$$\overline{X} = \frac{1}{N} \sum n_i x_i$$

$$\overline{X} = \frac{68}{40} = 1{,}7$$

Exercice sur la distribution du Chi Carré

Il y a donc, en moyenne, 1,7 unités tombant en panne chaque semaine.

On ajustera donc la distribution observée par une Loi de Poisson

On détermine la loi ajustée

$$pi = P(X = xi) = \frac{e^{-m} m^{xi}}{xi!}$$

Nombre d'unités tombées en panne au cours d'une semaine (xi)	Effectif observé	Loi ajustée	Effectif théorque N*pi
0	11	0,1827	7,3073
1	10	0,3106	12,4225
2	7	0,2640	10,5591
3	6	0,1496	5,9835
4	4	0,0636	2,5430
5	2	0,0216	0,8646
6 et plus	0	0,0080	0,3200
Total	40	1,0000	40

Exercice sur la distribution du Chi Carré

On teste la validité de l'ajustement par la méthode du χ^2

Pour le calcul de la distance d du khi2, on va regrouper les classes de sorte que les effectifs soient au moins égaux à 5, ceci de façon à respecter les conditions de convergence vers la loi.

Nombre d'unités tombées en panne au cours d'une semaine (xi)	Effectif observé	Loi ajustée	Effectif théorique N*pi	Distance du khi2 (ni+-Npi)²/Npi
0	11	0,1827	7,3073	1,8660
1	10	0,3106	12,4225	0,4724
2	7	0,2640	10,5591	1,1997
3 et plus	12	0,2428	9,7111	0,5395
Total	40	1,0000	40	4,0776

Exercice sur la distribution du Chi Carré

La distance d ainsi calculée suit une loi du khi2 à v=k-r-1 degrés de liberté où k représente le nombre de classes et r le nombre de paramètres estimés.

Après regroupement, le nombre de classes est égal à 4.

Un seul paramètre a été estimé, le paramètre m de la Loi de Poisson.

Par conséquent, le nombre de degrés de liberté est égal à 2

En consultant la table du khi2 pour v=2 on lit :

$P(\chi^2<4,61)=0,90 \Rightarrow P(\chi^2>=4,61)=0,10$

La valeur 4,0776 trouvée pour d a donc une probabilité légèrement supérieure à 0,10 (très précisément 0,1302) d'être dépassée dans l'hypothèse où le phénomène suit une loi de Poisson de paramètre m=1,7

Cette probabilité est suffisamment forte pour que l'on accepte l'hypothèse (rejet si la probabilité est inférieure à 0,05)

Exercice No 3

- Dans un magasin, afin de déterminer le nombre de vendeurs nécessaires pour assurer le service de clientèle dans des conditions satisfaisantes, on a procédé à une observation statistique sur les temps qui s'écoulent entre l'arrivée de 2 clients successifs
- Sur une demi-journée, on a obtenu les résultats suivants exprimés en centièmes d'heures

Temps T qui s 'écoule entre deux arrivées (en centièmes d'heures)	0-1	1-3	3-5	5-8	8-20
Nombre de cas observés	**28**	**30**	**23**	**9**	**10**

Exercice No 3

- Calculer une valeur approchée du temps moyen qui sépare l'arrivée de 2 clients successifs
- Calculer l'écart-type correspondant

Temps T qui s 'écoule entre deux arrivées (en centièmes d'heures)	0-1	1-3	3-5	5-8	8-20
Nombre de cas observés	**28**	**30**	**23**	**9**	**10**

Exercice No 3

Temps T qui s 'écoule entre deux arrivées (en centièmes d'heures)	0-1	1-3	3-5	5-8	8-20			
Centre de classe	0,5	2	4	6,5	14			
Nombre de cas observés fi	28	30	23	9	10	100		
fi*T	14	60	92	58,5	140	364,5		
						3,645	Moy	
Xi - moy	-3,145	-1,645	0,355	2,855	10,355			
Carré de Xi - moy2	9,891025	2,706025	0,126025	8,151025	107,226025			
	276,9487	81,18075	2,898575	73,359225	1072,26025	1506,6475		
						15,066	Variance	
						3,8816	Ecart TYpe	

Exercice No 3

- Les hypothèses faites sur le phénomène « arrivées des clients » conduisent à affirmer que la variable aléatoire T suit une fonction de répartition exponentielle
- $F(t) = P(T<t) = 1 - e^{-\lambda t}$ (t en centièmes d'heures)
- Justifier la valeur 0,274 qui pourrait être attribuée au coefficient λ.

Exercice No 3

- La loi de Poisson peut être la résultante d'un Processus de Poisson
- Un processus de Poisson correspond à à la réalisation d'évènements aléatoires dans le temps : arrivée bateaux, trains, avions à destination, appels téléphoniques, clients au guichet, pannes machines
- Le processus de Poisson répond aux hypothèses suivantes :
  - Probabilité de réalisation d'un événement au cours d'une petite période infinitésimale de temps dt est proportionnelle à cette durée de temps dt. Elle tend donc vers 0 si dt tend vers 0
  - Evènements indépendants entre eux et indépendants du temps

Exercice No 3

- On peut démontrer que, si le nombre de réalisations d'un évènement dans une unité de temps suit une distribution de Poisson de paramètre m, alors le temps entre deux réalisations successives de l'évènement se distribue selon une loi exponentielle de paramètre $\lambda = 1/m$
- L'intensité du processus de Poisson est défini par λ, paramètre de la loi exponentielle
- On choisit donc comme valeur de λ l'inverse de la moyenne, soit 1/3,645 = 0,274

Exercice No 3

- Pour toute la suite, on supposera $\lambda = 0,3$, valeur approchée de la précédente
- Déterminer les valeurs de F(t) pour les bornes des classes du tableau précédent (e-0,3 = 0,74082). En déduire les effectifs théoriques des différentes classes.

Exercice No 3

- Pour toute la suite, on supposera $\lambda = 0,3$, valeur approchée de la précédente
- Déterminer les valeurs de F(t) pour les bornes des classes du tableau précédent (e-0,3 = 0,74082). En déduire les effectifs théoriques des différentes classes.

lambda	0,3
1	0,25918
3	0,59343
5	0,77687
8	0,90928
20	0,99752
1E+14	1

Exercice No 3

lambda	0,3				
			Proba théorique	Effectif théorique	Effectif empirique
1	0,25918	Int 0-1	0,25918	25,92	28,00
3	0,59343	Int 1-3	0,33425	33,42	30,00
5	0,77687	Int 3-5	0,18344	18,34	23,00
8	0,90928	Int 5-8	0,13241	13,24	9,00
20	0,99752	Int 8-20	0,08824	8,82	10,00
1E+14	1	Int 20+	0,00248	0,25	

Exercice No 3

- Déterminer à l'aide d'un test du $\chi 2$ au seuil de 5% la validité de l'ajustement ainsi réalisé

Exercice No 3

- Principe du test
- Les écarts entre la distribution observée et la distribution ajustée à la loi peuvent être de deux causes :
  - Une fluctation normale d'échantillonnage (l'échantillon est un extrait de la population) avec des écarts faibles
  - L'ajustement n'a pas lieu d'être, avec un écart supérieur avec des écarts élevés
- Cet écart va être mesuré par la **distance** existante entre la théorique ajustée et la distribution observée
- Cette distance étant une grandeur aléatoire, elle est mesurée par une loi de probabilité

Exercice No 3

- Cette loi permet de calculer la probabilité d'obtenir une distance supérieure à la distance observée
- On se fixe un seuil de probabilité α dit **seuil de confiance**
- Si la probabilité obtenue est inférieure au seuil de confiance, on rejette l'hypothèse.
- Si la probabilité obtenue est supérieure au seuil de confiance, on accepte l'hypothèse.

Exercice No 3

Effectif théorique	Effectif empirique	ecart	carré écart	carré écart/Npi
25,91817793	28	-2,081822068	4,333983124	0,167217894
33,42485609	30	3,424856094	11,72963927	0,350925647
18,34394996	23	-4,656050041	21,67880198	1,181795744
13,24122069	9	4,241220686	17,98795291	1,358481467
9,071795329	10	-0,928204671	0,861563911	0,09497171
			D=	3,1533925

Exercice No 3

- Ho : le phénomène suit une loi exponentielle de paramètre 0,3
- Variable de décision : D
- D-> $\chi 2$ (υ = k-r-1)
- k = 5 classes après regroupement
- r = 1 paramètre de la loi exponentielle
- 3 degrés de liberté
- Seuil 5% =>H0 étant vrai P(H1)=0,05
- P(H0) = 0,95
- On lit dans table P(D<7,81) =0,95
- Donc D limite 7,81 à comparer avec 3,15
- Hypothèse acceptée

Introduction à la statistique inductive

- Soient des données collectées dans le monde réel, caractéristiques d'une population : données empiriques
- Soient des données théoriques issues de modèles mathématiques, ceux des lois de probabilité
- Est-il possible de prévoir la valeur de ces caractéristiques en déterminant la loi de probabilités qui les régit à partir de l'analyse d'une échantillon de la population ?

Sondage

- Lorsqu'on doit évaluer une **caractéristique** X d'une **population** P (sexe ou taille d'êtres vivants, goûts musicaux ou culinaires de personnes, qualité de fabrication de pièces), deux méthodes peuvent être employées.
- Le **recensement** consiste à mesurer la valeur du caractère chez tous les individus.
- Le **sondage** limite l'analyse à un sous-ensemble appelé **échantillon**

Sondage

- De nombreuses raisons évidentes militent pour la seconde méthode
  - La faisabilité
  - La rapidité
  - Le coût moins élevé
- Mais la méthode a bien sûr ses faiblesses
  - Les valeurs de l'échantillon ne représentent qu'accidentellement celles de la population : ce sont des variables aléatoires
  - L'écart entre échantillon et population est d'autant plus grand que l'échantillon est atypique (erreur d'échantillonage)

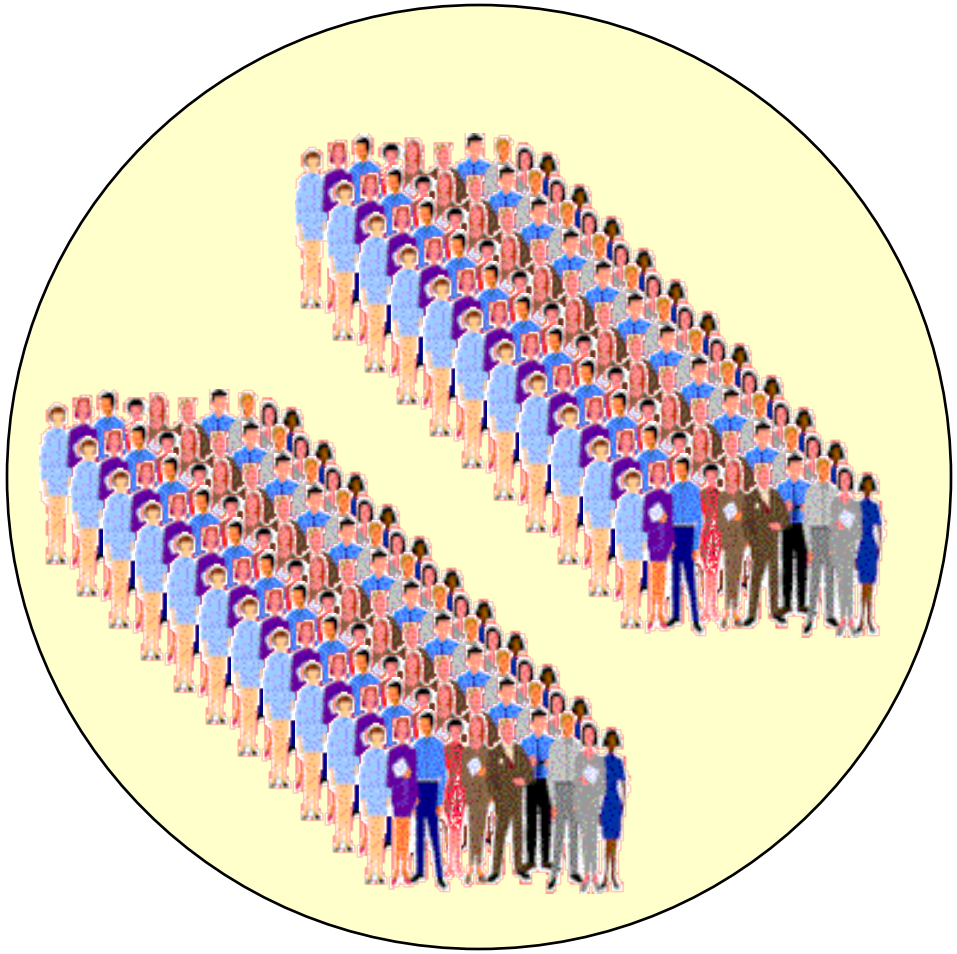
Sondage

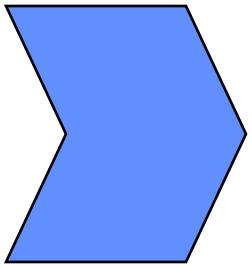
- Seule une sélection effectuée parfaitement au hasard permet d'éliminer toutes les causes de déviation systématique (**biais**)
- Exemple : dans une enquête sur les goûts alimentaires (Aimez-vous le yaourt à la vanille ?) il faut s'assurer d'équilibrer hommes et femmes, actifs et inactifs, forts et faibles pouvoir d'achat.
- Dans une population P de p individus, chacun aura la probabilité 1/p de figurer dans l'échantillon.

Sondage

- Il faut donc :
- Une « population » P
- Une variable aléatoire X associée à cette population
- Un échantillon E de P
- Exemple :
  - Soit X le nombre de pièces défectueuses repérées dans un lot E de 100 pièces soumises à un contrôle, au sein d'une production P de 5000

Echantillonage





Choisir l'échantillon de population le plus représentatif du comportement de l'ensemble

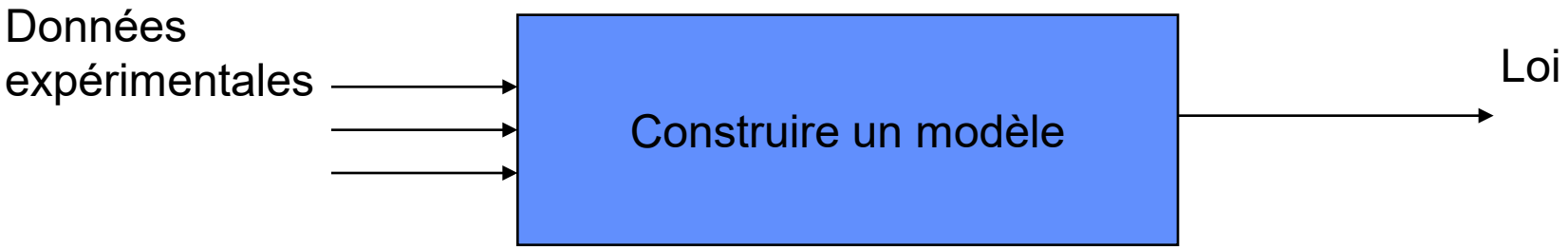
Echantillonage

- La collecte des données **empiriques** d'un échantillon pose divers problèmes
- Pertinence du choix de l'échantillon (représentativité)
- Importance ou non de l'ordre d'arrivée des données
- Classification des données
- Représentation des données (histogrammes des valeurs en répartissant les données en classes autour des centres de classes)

Echantillonage

- En statistique, la **population** des résultats désigne la totalité des résultats expérimentaux possibles
- Un **échantillon** de la population est un ensemble de données rassemblées en réalisant l'expérience un certain nombre de fois.
- **L'inférence statistique** consiste à tirer des conclusions **théoriques** au sujet d'une population au moyen d'un échantillon extrait **empiriquement** de cette population.

Inférence



Processus d'inférence inductive

Si données statistiques => Inférence statistique

Inférence et densité de probabilité

- Le choix du mathématicien
  - Modèle qui prédise les résultats associés à un tirage de 100 pièces
  - Ou
  - Modèle qui prédise la fréquence des différentes valeurs de X
- Le choix 2 conduit à retenir comme modèle des fonctions de densité de variables aléatoires et les inférences statistiques s'appliquent généralement aux fonctions de densité

Base théorique pour l'échantillonnage : Moments empiriques

- Pour étudier une distribution d'un ensemble de valeurs, les histogrammes procurent beaucoup d'informations générales
- La description mathématique fournit des informations plus précises et plus utiles
- Cette description est basée les moments
- Moments d'ordre 1, 2, 3, ..., n
- Dans la pratique 1 et 2

Moments empiriques

- Soit x1, X2, ..., Xn les valeurs observées d'un échantillon de taille n de la variable aléatoire X
- Le moment d'ordre k centré sur l'origine d'une distribution empirique est donné par

$$m_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

Moments empiriques

- Le moment d'ordre k centré sur la moyenne d'une distribution empirique est donné par

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^k$$

Moments empiriques

- Le moment d'ordre 1, $\overline{x}$, est le centre de gravité de la distribution empirique.
- Cette moyenne de l'échantillon sert à estimer la moyenne théorique μ

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} (x_i)$$

Moments empiriques

- Puisque σ^2 est le moment d'ordre 2 d'une distribution théorique, le moment d'ordre 2 d'une distribution empirique est tout naturellement associé à la variance.
- s^2 est la variance de l'échantillon
- L'écart type est s
- Noter le n-1

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

Exercice (voir les corrigés dans dossier exo)

- Soit la distribution empirique du tableau ci-dessous concernant des durées en secondes de conversations téléphoniques.
- Tracer l'histogramme, calculer la moyenne et l'écart-type de cet échantillon
- Déterminer les pourcentages approximatifs de données qui se situent dans les intervalles x-s et x+s, x-2s et x+2s

Xi	49,50	149,50	249,50	349,50	449,50	549,50	649,50	749,50	849,50	949,50
fi	6	28	88	180	247	260	133	42	11	5

Tests d'hypothèses

- Généralement, une hypothèse statistique est une affirmation sur la fonction de densité d'une variable aléatoire.
- Affirmer qu'une variable aléatoire se distribue selon une loi normale est un exemple d'hypothèse statistique
- Dans la plupart des cas on va supposer la fonction de densité connue et l'hypothèse va porter sur une affirmation concernant la valeur d'un paramètre de cette fonction de densité
- Exemple : Hypothèse que la moyenne d'une variable aléatoire de Poisson est égale à 10

Tests d'hypothèses

- Un test d'hypothèse statistique définit une procédure d'acceptation ou de rejet d'une hypothèse
- Cette définition assure une liberté au statisticien pour concevoir son test

Population vs Echantillon

- Définitions
- Caractéristiques
- Notations
- Loi de probabilité

Retour sur l'échantillonnage : définitions

- **Tirage simple**
- **Tirage exhaustif** : un individu déjà sélectionné n'est pas remis dans la population "mère" et ne peut donc être sélectionné à nouveau
- **Tirage non exhaustif** : un individu déjà sélectionné est remis dans la population mère et peut donc être tiré une nouvelle fois
- **Plan d'expérience** : Etude des méthodes d'échantillonnage et des problèmes qui s'y rattachent
- **Echantillon aléatoire** : chaque individu de la population mère a la même probabilité d'appartenir à l'échantillon
- **Distribution d'échantillonnage**. Considérons tous les échantillons de taille n tirés de la population mère et, pour chacun d'eux, calculons une caractéritique C (moyenne, variance). L'ensemble des valeurs de C donne la distribution d'échantillonnage de C

Population vs échantillon : Population

- La distribution du caractère quantitatif X dans la population mère P est caractérisée par le tableau No 1 :

Modalités	**X1**	**…**	**Xi**	**…**	**Xq**
Effectifs	**N1**	**…**	**Ni**	**…**	**Nq**
Fréquences	$P1 = \frac{N1}{N}$	**…**	$Pi = \frac{Ni}{N}$	**…**	$Pq = \frac{Nq}{N}$

Population vs échantillon : Population

Cette population est de plus caractérisée par ses moments.

La moyenne (empirique):

$$M = \sum_{i=1}^{p} (f_i X_i)$$

La variance (empirique):

$$\sigma^2 = \sum_{i=1}^{p} f_i \, (X_i - M)^2$$

Population vs échantillon : Echantillon

Le prélèvement de n individus dans P conduit à un échantillon de taille n.

Il y a C_N^n échantillons de taille n possibles

La distribution du caractère quantitatif X dans cet échantillon est caractérisée par le tableau No 2:

Modalités	**x1**	**…**	**xi**	**…**	**xq**
Effectifs	**n1**	**…**	**ni**	**…**	**nq**
Fréquences	$f1 = \frac{n1}{n}$	**…**	$f1 = \frac{ni}{n}$	**…**	$f1 = \frac{nq}{n}$

Population vs échantillon : Echantillon

Cet échantillonnage est de plus caractérisé par ses moments.

La moyenne :

$$\overline{x} = \sum_{i=1}^{n} (f_i X_i)$$

La variance :

$$s^2 = \sum_{i=1}^{n} f_i \, (X_i - \overline{X})^2$$

Population vs échantillon : Distribution d'échantillonnage

- Les C_N^n moyennes des C_N^n échantillons différents consituent la **distribution d'échantillonnage des moyennes.**

- Celle ci est aussi caractérisable par une moyenne et un écart-type

Population vs échantillon : notations

Notations	Population	Echantillon
Taille	**N**	**n**
Moyenne	**M**	$\overline{x}$
Ecart-type	σ	**s**
Fréquence	**p**	**f**

Population vs échantillon : Loi de probabilité

- L'échantillon est obtenu par n tirages successifs.
- Chacun de ces tirages représente une expérience aléatoire dont le résultat est xi

Modalités	**x1**	**…**	**xi**	**…**	**xq**
Probabilités	**P1**	**…**	**Pi**	**…**	**Pq**

- Aux n tirages de l'échantillon sont donc associées n variables aléatoires Xi de même loi de probabilité.
- Les tirages étant non exhaustifs (tirages avec remise) ces n variables sont indépendantes
- P(X1 = Xa, X2 = Xb, ...) = P(X1 = Xa)*P(X2=Xa)

Estimation et Distribution d'échantillonnage

- Suivant que la distribution du caractère X dans la population mère P est connue ou non, 2 problèmes peuvent être abordés.
- **Pb No 1** : Connaissant $\overline{x}$ et s^2, que peut on dire de la moyenne M et de la variance σ^2 de la population mère ?
- Ce problème est celui de **l'estimation.** Comment décrire la population mère à partir d'un échantillon ? La grandeur caractéristique de l'échantillon est **l'estimateur**.
- **Pb No 2** : Connaissant la distribution de X dans P et les valeurs de M et σ^2 , que peut-on dire des caractéristiques d'un échantillon tiré au hasard ?
- Ce problème est celui de la **théorie des distributions d'échantillonnage,** qui étudie les distributions de toutes les caractéristiques de l'échantillon tiré au hasard.

Distribution d'échantillonnage

- Distribution des moyennes
- Distribution des fréquences

Distribution d'échantillonnage : cas des moyennes

- Cette théorie étudie les distributions de toutes les caractéristiques de l'échantillon tiré au hasard : variables xi, moyennes, variances et fréquences
- Considérons le cas des moyennes
- Considérons le cas où **les variables aléatoires Xi sont indépendantes (tirage non exhaustif) et de même loi de probabilité normale dont l'écart-type est connu**
- $\overline{X_n}$ est défini comme X1 + X2 + ... + Xi + ... Xn / n
- Le théorème "Central limit" dit que la loi de la moyenne centrée réduite de n variables aléatoires indépendantes peut être approximée par une loi normale (centrée, réduite) avec une précision d'autant plus grande que n est grand.
- La variable centrée réduite $T = (\overline{x_n} - E(\overline{x_n})) / \sigma(\overline{x_n})$ suit donc une loi normale centrée réduite si n est assez grand

Distribution d'échantillonnage des moyennes

On démontre que :

$$E(\overline{X}) = M$$

$$\sigma(\overline{X}) = \sigma / \sqrt{n}$$

Il en résulte que la moyenne de l'échantillon suit approximativement une loi normale $\mathcal{N}$(moyenne population, écart type population divisé par racine carré de l'échantillon)

Distribution d'échantillonnage des moyennes

C'est ici qu'il faut bien comprendre les composantes du problème posé

Le problème posé est de déduire la distribution d'échantillonnage de la moyenne fondée sur un échantillon aléatoire de taille n extrait d'une population normale N.

Soit X distribué selon une loi normale de moyenne M et de variance σ^2

Nous envisageons un échantillon aléatoire de taille n prélevé dans cette population

La moyenne de cet échantillon :

$$\overline{X} = \frac{1}{n}(X1 + X2 + ... + Xn)$$

Distribution d'échantillonnage des moyennes

- Cette moyenne est une variable aléatoire parce que les Xi qui la composent sont des variables aléatoires

- Après le prélèvement, $\overline{X}$ est un nombre

- Avant le prélèvement, c'est une variable aléatoire dont les valeurs dépendent des valeurs prises par la variable de départ $\overline{X}$
- Il faut déterminer la fonction de densité de $\overline{X}$

- Nous avons vu que la variable $\overline{X}$ est une variable normale de moyenne M et de variance σ^2/n

Distribution d'échantillonnage des moyennes

- On peut donc exprimer le théorème selon lequel :
- **Si X se distribue selon une loi normale de moyenne M et de variance σ^2 et si on prélève un échantillon aléatoire de taille n, la moyenne de l'échantillon X se distribue selon une loi normale de moyenne M et de variance σ^2/n**
- Ce théorème démontre que la précision d'une moyenne d'un échantillon qui estime la moyenne d'une population augmente lorsque la taille de l'échantillon croit
- Il faut prélever un échantillon quatre fois plus important si on veut doubler la précision de l'estimateur.

Distribution d'échantillonnage des moyennes

- Nous avons considéré le cas du tirage non exhaustif dans une population dont l'écart type est connu.
- Dans le cas d'un tirage sans remise, on doit corriger l'écart type, avec un coefficient d'exhaustivité

$$\sqrt{\frac{N-n}{N-1}}$$

- Dans le cas où l'écart-type n'est pas connu, on en fait une estimation ponctuelle (Loi de Student vue plus loin)

Distribution d'échantillonnage des fréquences

- La fréquence d'échantillons est la variable aléatoire F
- Dans le cas d'échantillons indépendants (tirage avec remise)

$F = P(n,p)$

Si $n \geq 30$ et $np(1-p) < 5$

$F \Rightarrow B(n,p)$

Si $n \geq 30$ et $np(1-p) < 5$

$F \Rightarrow \mathcal{N}$ dont $E(F) = p$ et $\sigma(F) = \sqrt{(p)(1-p) / n}$

Exercice

- Un fabricant de fil synthétique de canne à pêche a déterminé après une longue période d'essai que la résistance à la rupture de son fil se distribue approximativement selon une loi normale de moyenne égale à 30kg et d'écart type égal à 4 kg.
- Il modifie son processus de fabrication pour gagner du temps.
- On prélève un échantillon de 25 pièces dans la production du nouveau processus et on mesure la moyenne de cet échantillon qui est égale à 28 kg.
- Quelle est la probabilité d'avoir une résistance moyenne à la rupture inférieure ou égale à 28 kg si le nouveau processus ne diminue pas la résistance à la rupture ?

Estimation

- Principe
- Estimation ponctuelle
- Estimation par intervalles de confiance

Estimation

- Nous venons de voir que la théorie des distributions d'échantillonnage avait pour but de déduire la connaissance des distributions des variables aléatoires de l'échantillon à partir de la connaissance de la distribution de X dans la population mère.
- La théorie de l'estimation se propose de résoudre le problème inverse.
- L'estimation est la recherche de la valeur d'une caractéristique inconnue Θ d'une population mère, à partir des observations faites sur un échantillon
- Un estimateur T de Θ est une fonction des valeurs observées sur un échantillon ayant pour but de fournir une valeur de Θ

Estimation

- Nous nous attachons plus particulièrement à trouver une estimation, autrement dit une approximation de la moyenne M et de l'écart-type σ de la population lorsque le caractère X est supposé suivre une loi de Gauss.
- Nous nous proposons aussi d'estimer une proportion p d'une modalité X dans la population mère.
- Dans ces travaux, nous déduisons toujours la valeur approchée du paramètre Θ à estimer, à partir de l'observation d'un échantillon de taille n.

Estimation : deux méthodes

- L'estimation ponctuelle détermine pour le paramètre Θ cherché une valeur approchée unique .
- L'estimation par intervalle de confiance détermine un "intervalle de confiance" qui a une grande probabilité de contenir la valeur exacte de Θ

Estimation ponctuelle

- Cette méthode utilise un estimateur ponctuel du paramètre inconnu Θ.
- Il s'agit d'une fonction à plusieurs variables Tn (X1, X2, ..., Xn) qui aux n variables aléatoires Xi de l'échantillon fait correspondre une variable aléatoire Tn appelée estimateur.
- Cette fonction est telle que si les résultats d'un sondage sont : X1=x1, X2=x2, ..., Xn = xn, la valeur numérique Tn est une valeur approchée du paramètre Θ à estimer
- L'estimation ponctuelle est la valeur unique fournie pour le paramètre Θ par l'estimateur retenu

Estimation ponctuelle

- Supposons que la moyenne M du caractère X dans la population mère puisse être estimée par la moyenne m ($\overline{x}$) des valeurs observées dans l'échantillon (m_n)
- Ceci revient à dire que m_n est un estimateur ponctuel de M
- Soit $Tn(X1, X2, ..., Xn) = 1/n \sum xi = m_n$

- Si dans un échantillon de taille 3 on trouve X1=2, X2=6, X3=4,
- La moyenne M pourra être estimée par la valeur T3(2,6,4) = 1/3(2+6+4) = 4

Estimation ponctuelle

- Pour que la valeur approchée du paramètre Θ, fournie par l'estimateur ponctuel, comporte une précision suffisante, et surtout pour que cette précision s'améliore lorsque la taille n de l'échantillon augmente, il faut que l'estimateur réponde aux conditions suivantes :
- $E(Tn) = \Theta$
- $V(Tn) \to 0$ quand $n \to \infty$
- L'estimateur est dit alors "**absolument correct**"
- L'estimateur est une variable aléatoire, souvent notée avec un accent circonflexe (^m, ^p) dont on connaît, grâce à l'échantillon, une réalisation.
- Cette réalisation constitue l'estimation.

Estimation ponctuelle

- Le meilleur des estimateurs, le plus précis, est à taille égale de l'échantillon celui dont la variance est la plus faible.
- On peut considérer la variance de l'estimateur comme un indice de sa précision.
- Ce meilleur estimateur est appelé estimateur efficace.

Estimateur de la moyenne d'une population

- La moyenne $\overline{x}$ observée sur l'échantillon est l'estimateur efficace de la moyenne M de la population

$$E\{\overline{x}\} = m$$

- La variance de cet estimateur est égale à :

$$V\{\overline{x}\} = \frac{\sigma^2}{n}$$ dans le cas de tirages avec remise

$$V\{\overline{x}\} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$ dans le cas de tirages sans remise

Estimateur de la variance d'une population

- On démontre que

$$E\left\{s_n^2\right\} = \frac{n-1}{n} . \sigma^2$$

- La variance de l'échantillon n'est donc pas un estimateur absolument correct de la variance de la population
- L'estimateur de la variance de la population est :

$$\hat{\sigma}^2 = \frac{n}{n-1} . s_n^2$$

- L'estimation de la variance inconnue de la population mère sera celle observée dans l'échantillon, multipliée par n/n-1

Estimateur d'une proportion d'une population

$$p = \frac{N_i}{N} \qquad f_n = \frac{n_i}{n}$$

- On démontre que fn est un estimateur efficace de p

$$\hat{p} = f_n$$

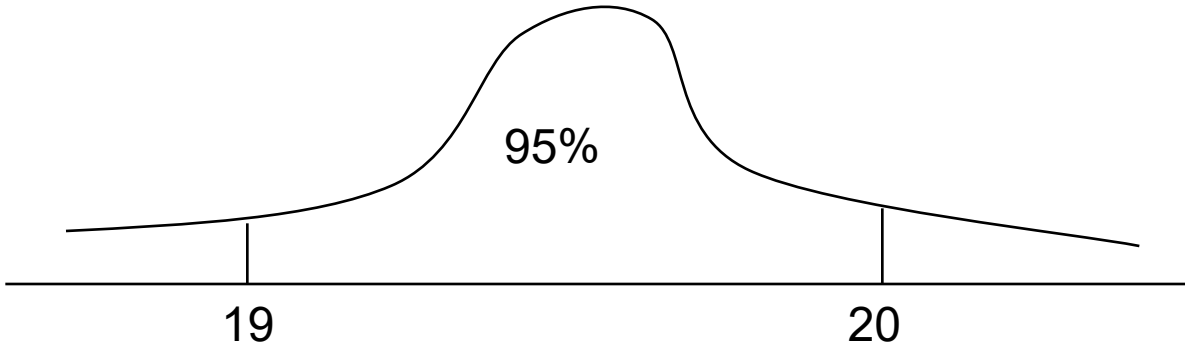
- L'estimation de la proportion inconnue de la population mère sera donc la fréquence observée dans l'échantillon

Estimation par intervalle de confiance

- L'estimation ponctuelle a pour défaut de ne fournir ni la précision de l'estimation, ni le risque d'erreur.
- La méthode d'estimation par intervalle de confiance a pour mérite de fournir l'intervalle $(\Theta - \Delta\Theta , \Theta + \Delta\Theta)$ ou la valeur vraie Θ^* a la probabilité α de se trouver.
- Cette méthode donne, outre la valeur approchée Θ, la précision de cette approximation $\Delta\Theta / \Theta$
- La précision de l'estimation est la probabilité α' de commettre une erreur relative égale à l'approximation en considérant Θ à la place de Θ^*
- α' est le **degré** ou **coefficient de confiance**
- $\alpha = 1 - \alpha'$ indique la probabilité inverse que l'intervalle de confiance ne contienne pas Θ^*. C'est le **seuil de confiance** ou **risque d'erreur**

Estimation par intervalle de confiance

- Intervalle de confiance : (19,20)
- Limites de confiance : 19 et 20
- Coefficient de confiance : 95%
- La valeur cherchée a 95 % de chances de se trouver entre 19 et 20
- 95% = P(19,5 - 0,5 < Θ^* < 19,5 + 0,5)
- Précision : 0,5 / 19,5 = 3%



Intervalle de confiance pour la moyenne d'une loi normale

- A chaque individu d'une population P, est attachée une valeur xi d'un caractère x
- La distribution de X dans P est supposée correspondre à une loi normale $\mathcal{N}(M,\sigma)$
- On se propose d'estimer M en prélevant au hasard un échantillon de taille n
- Soit $\overline{x}$ la moyenne de la variable X dans l'échantillon de taille n

Intervalle de confiance pour la moyenne d'une loi normale

- Considérons que la variance est connue
- On a démontré que si X suit une loi normale dans la population mère, m_n suit également une loi normale

- $E(\overline{X}) = M$

- $\sigma(\overline{X}) = \sigma / \sqrt{n}$

- Autrement dit, si X suit la loi $\mathcal{N}(M,\sigma)$, $\overline{x}$ suit la loi $\mathcal{N}(M,\sigma/\sqrt{n})$
- et la variable centrée réduite $(m_n-M) / \sigma/\sqrt{n}$ suit la loi $\mathcal{N}(0,1)$,

Intervalle de confiance pour la moyenne d'une loi normale

- Si on se fixe à l'avance un coefficient de confiance α' et si on cherche t tel que

- $P(-t < (m_n - M) / \sigma/\sqrt{n} < +t) = \alpha'$

- il en résulte que t est défini en fonction de α' par la relation

$$\int_{-t}^{+t} \frac{1}{\sqrt{2\pi}} \times e^{-u2/u} \times du = \alpha'$$

- et sa valeur est lue dans la table de loi normale

Intervalle de confiance pour la moyenne d'une loi normale

- Or

$$-t < \frac{m_n - M}{\sigma / \sqrt{n}} < +t \Leftrightarrow m_n - t\frac{\sigma}{\sqrt{n}} < M < m_n + t\frac{\sigma}{\sqrt{n}}$$

- Donc

$$\Pr ob\left(m_n - t\frac{\sigma}{\sqrt{n}} < M < m_n + t\frac{\sigma}{\sqrt{n}}\right) = \alpha'$$

- Autrement dit :

$$\left(m_n - t\frac{\sigma}{\sqrt{n}} < M < m_n + t\frac{\sigma}{\sqrt{n}}\right)$$

- constitue un intervalle de confiance à α% de M

Intervalle de confiance pour la moyenne d'une loi normale

- Exemple :
- On choisit α' = 95%
- Donc t = 1,96
- si un sondage de taille n = 100 a donné $\overline{x} = 3$ et a permis de supposer que $\sigma = 2$
- On peut affirmer que M a 95 chances sur 100 d'appartenir à l'intervalle :

$$\left[3 - 1{,}96.\frac{2}{10};\ 3 + 1{,}96.\frac{2}{10}\right]$$

- Soit [2,6 , 3,4]

Exercice

- La gestion de la qualité de service du RTC est fondée sur divers indicateurs comme l'indicateur TCOM : temps d'établissement des communications (délai exprimé en secondes de mise en relation entre deux abonnés)
- A l'image de la durée de communication de l'exercice précédent, on a pu déterminer que ce délai était une variable aléatoire régie par une loi normale
- Sur un échantillon de 300 communications, on observe une valeur moyenne de cet indicateur X = 15,5 secondes et un écart-type de 4 secondes.

1. Déterminer un intervalle de confiance de niveau 95% pour la valeur moyenne de TCOM
2. Quelle taille d'échantillon serait nécessaire pour estimer la moyenne m avec une précision de +/- 0,1 sec (pour le même niveau de confiance de 95% et un écart-type constant quelque soit la taille de l'échantillon)

Cas de populations quelconques

- Nous avons fait l'hypothèse d'une population "normale" (caractéristique étudiée répartie selon une loi normale)
- Dans le cas de population quelconque, la distribution de la moyenne x ne tend vers une loi normale que lorsque l'effectif n de l'échantillon tend vers l'infini
- Le principe vu est donc applicable pour de gros échantillons (n>30)
- Lorsque l'effectif de l'échantillon est petit, c.a.d. en pratique inférieur à 30 unités, la moyenne x de l'échantillon ne suit une loi normale que si la population d'origine est elle-même normale (Notre hypothèse de départ)

Cas où on ne connait pas l'écart-type de la population

- A chaque prélèvement d'échantillon est attaché une valeur de la variable aléatoire m_n (moyenne de l'échantillon), une valeur de la variable aléatoire s^2_n (variance de l'échantillon)
- Si on ne connait pas l'écart type de la population,on peut introduire un estimateur de cet écart type

$$\hat{\sigma}^2 = \frac{n}{n-1} . s_n^2$$

- puis la variable centrée réduite

$$T' = \frac{m_n - M}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

Cas où on ne connait pas l'écart-type de la population

- On démontre que T' suit une loi de Student à ν = (n-1) degrés de liberté (si X suit une loi de Gauss)
- n représente l'effectif de l'échantillon
- Cette loi est tabulée en fonction de ν (En Excel LOI.STUDENT.INVERSE(probabilité, degré liberté)
- Pour des valeurs de ν suffisamment grandes (supérieures à 30) elle est convenablement approximée par la loi normale réduite $\mathcal{N}(0,1)$

	0,9	0,5	0,3	0,2	0,1	0,05	0,02	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	31,599
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	12,924

Exemple a verifier avec sujet devoir

- A la suite d'un accident dans une centrale nucléaire, avec rejet de particules radioactives dans l'atmosphère, un échantillon aléatoire de 16 personnes a été tiré avec probabilités égales dans la ville voisine.
- Cet échantillon a été soumis pendant une année à un contrôle d'irradiation.
- On désigne par x la mesure du rayonnement reçue par une personne en un an. La variable x est normalement distribuée.
- Les résultats de l'échantillon
- moyenne $\overline{x}$ = 15,125 rem
- Ecart type s = 4,841
- Estimer le rayonnement moyen reçu par les habitants de la ville et déterminer l'intervalle de confiance à 99% de cette estimation

Estimation d'une proportion

- Considérons la distribution d'un caractère X dans une population P tel que celle-ci est composée de deux catégories d'individus en proportion p et q = 1-p.
- On estime la proportion p inconnue par la fréquence f = x/n observée sur l'échantillon.
- Les cas à considérer
  - Echantillon tiré avec remise
    - Cas d'un gros échantillon
    - Cas où p est petit, avec un échantillon assez gros
    - Cas d'un petit échantillon
  - Echantillon sans remise

Estimation d'une proportion : echantillon tiré avec remise

- La fréquence f est une variable binomiale de paramètre n et p
- f -> $\mathcal{B}$(n, p)
- Son espérance mathématique

$$E(F) = p$$

- Son écart type

$$\sigma(F) = \sqrt{\frac{p(1-p)}{n}} = \sigma_F$$

- La connaissance de la loi d'échantillonnage de f permet de déterminer l'intervalle de confiance

Estimation d'une proportion : Gros échantillon avec remise

- Lorsque l'échantillon est suffisamment grand, la loi binomiale peut être approchée par la loi normale.
- L'approximation de la loi binomiale par la loi normale est acceptable lorsque npq > 9
- Dans ces conditions f suit une loi normale de paramètres :
- M = p et $\sigma_F = \sqrt{\frac{p(1-p)}{n}}$
- p étant inconnu, σ_f l'est aussi.
- La loi d'échantillonnage de f n'est pas entièrement donnée.
- Deux possibilités existent pour déterminer l'intervalle de confiance :
  - méthode par estimation de l'écart-type
  - méthode de l'ellipse

Exercice

- On s'intéresse à la proportion d'individus achetant le journal local dans une petite ville de 10 000 habitants. Sur 100 personnes interrogées, 70 personnes déclarent acheter le journal.
- Au seuil de confiance de 80%, estimer la proportion d'individus qui achètent le journal dans la ville
- Même question au seuil de 90%
- Combien de personnes doit-on interroger au seuil de 90% pour que la précision de l'estimation soit de 5%

Détermination de la taille d'un échantillon

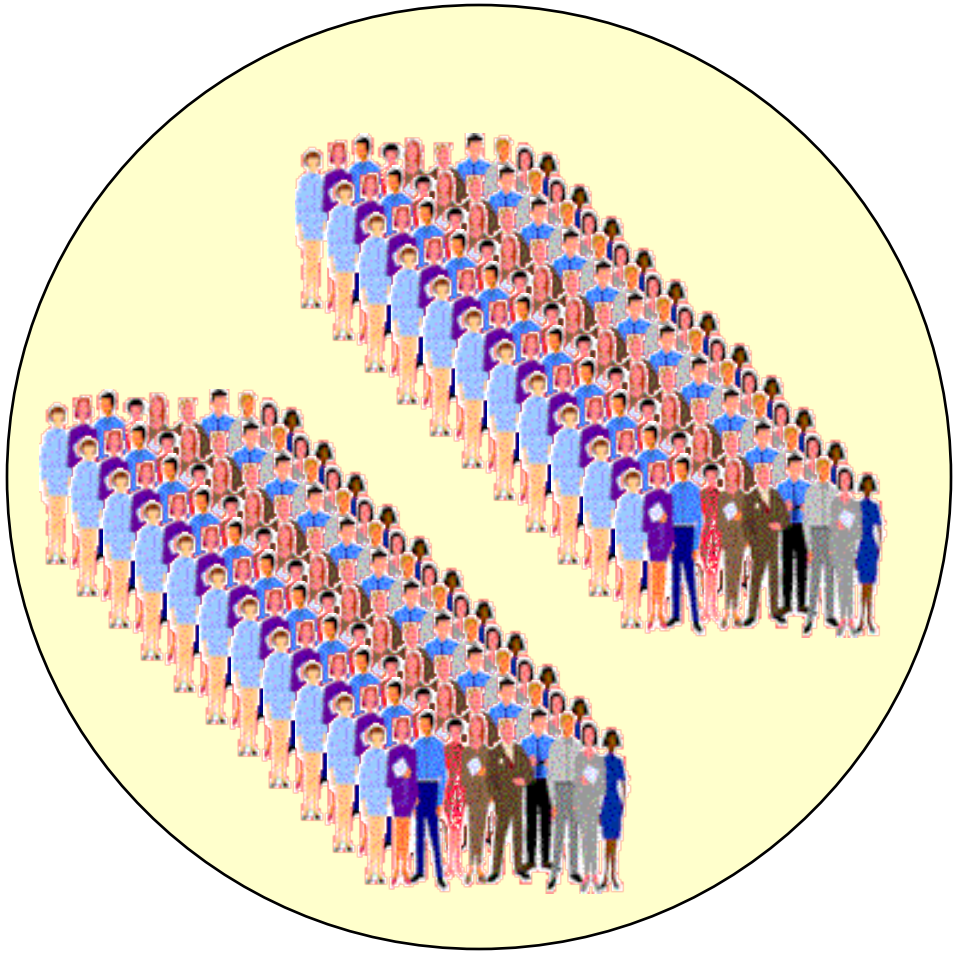
- La détermination de la taille d'un échantillon pour obtenir une précision donnée est l'inverse du calcul de l'intervalle de confiance d'une estimation

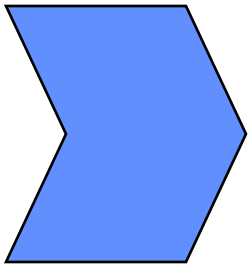
- Etant donné un seuil de probabilité 1-α fixé a priori, quel doit être l'effectif n de l'échantillon pour obtenir la précision, c'est à dire l'intervalle de confiance désiré ?

Echantillonage

- En statistique, la **population** des résultats désigne la totalité des résultats expérimentaux possibles
- Un **échantillon** de la population est un ensemble de données rassemblées en réalisant l'expérience un certain nombre de fois.
- **L'inférence statistique** consiste à tirer des conclusions **théoriques** au sujet d'une population au moyen d'un échantillon extrait **empiriquement** de cette population.

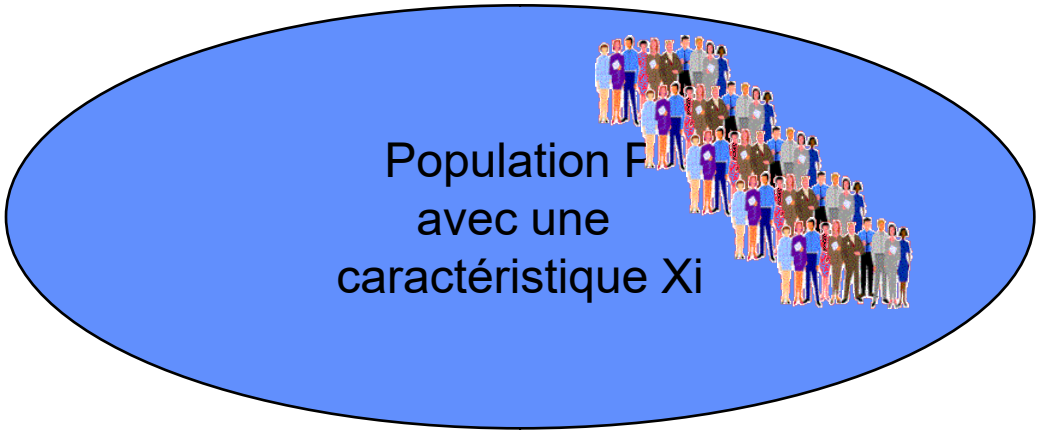
Echantillonage



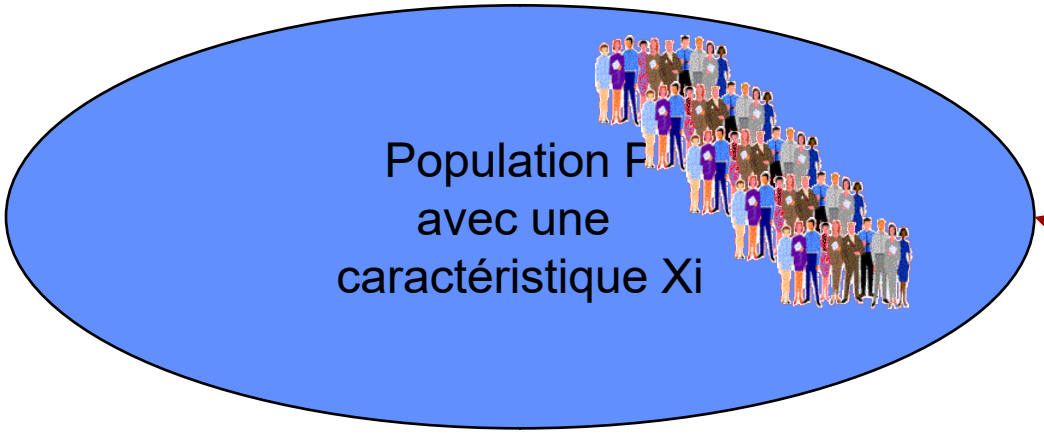


Choisir l'échantillon de population le plus représentatif du comportement de l'ensemble

Synthèse

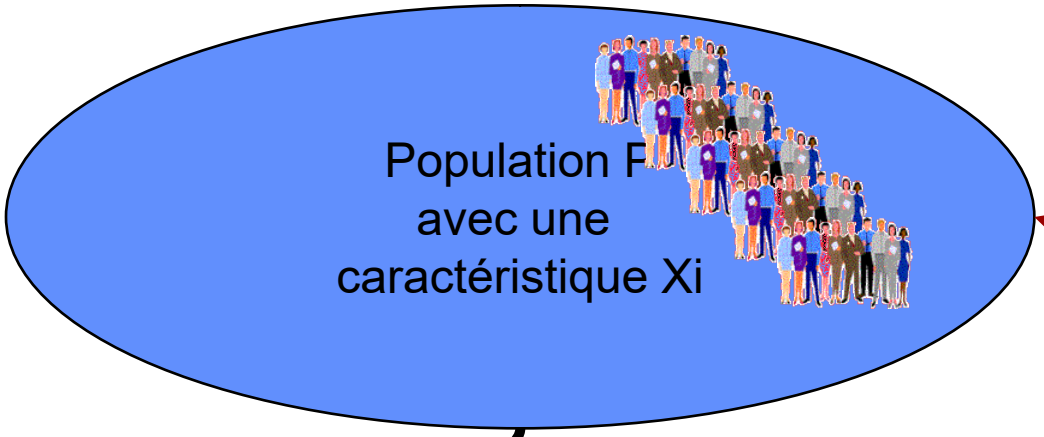


Synthèse

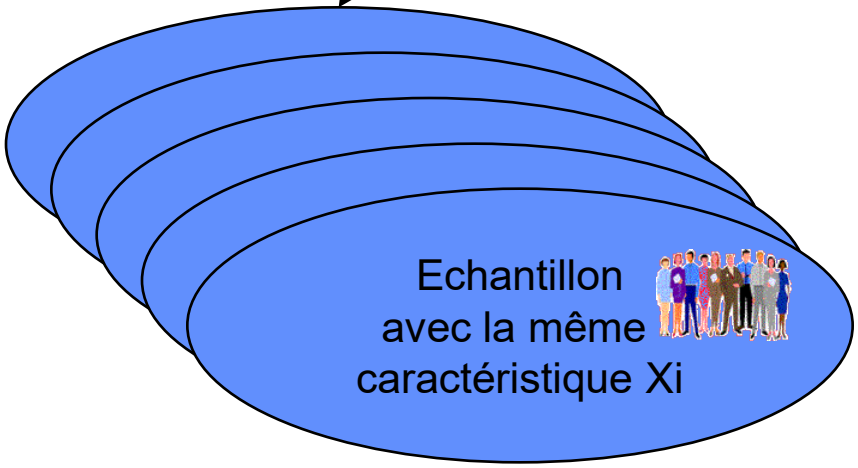


Loi de probabilité régissant la distribution de Xi dans P caractérisée par M et σ

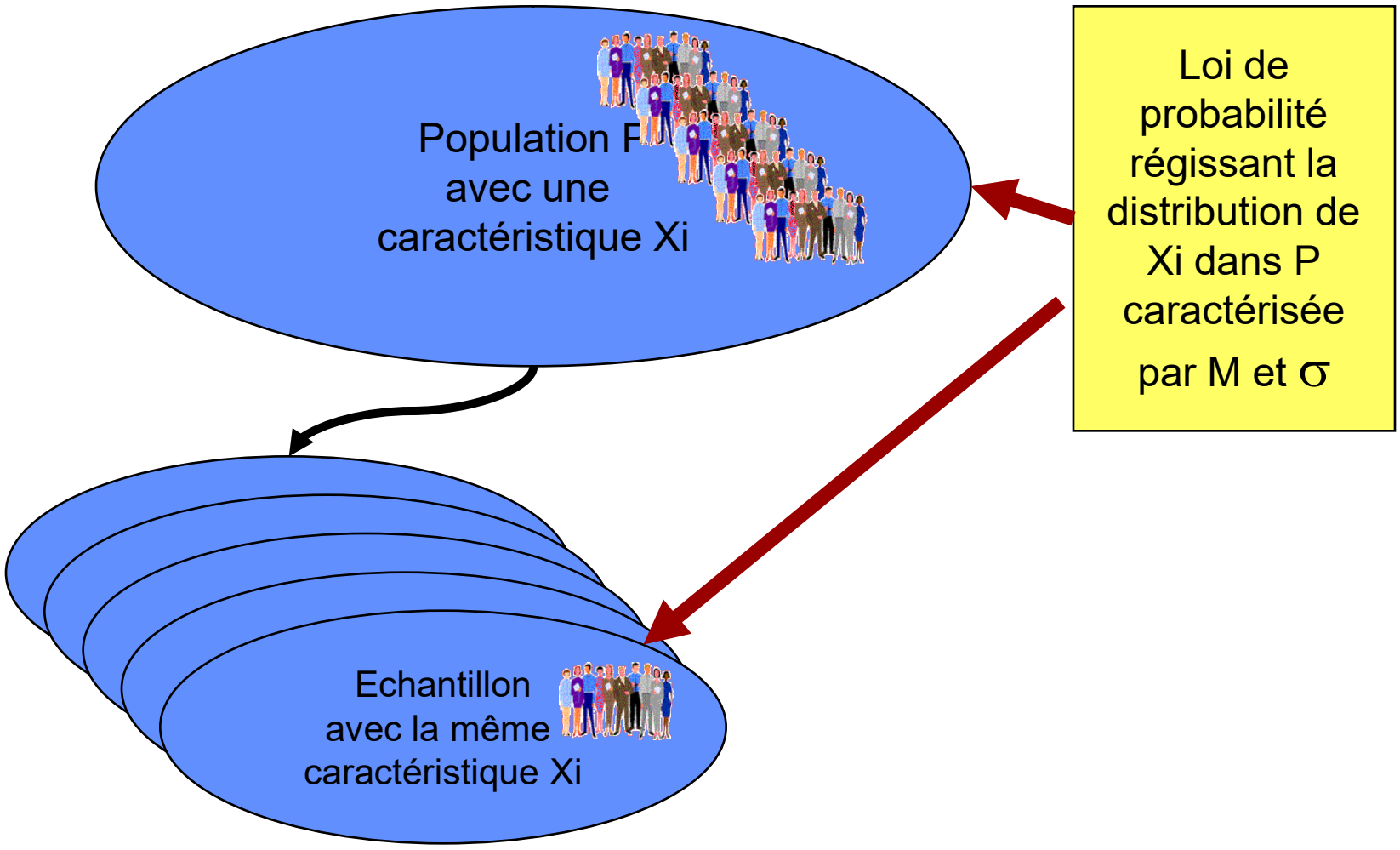
Synthèse



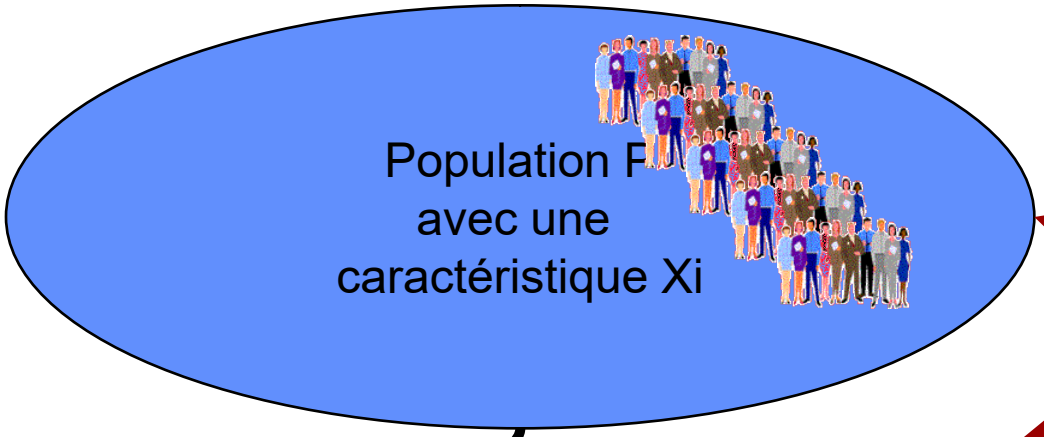
Loi de probabilité régissant la distribution de Xi dans P caractérisée par M et σ



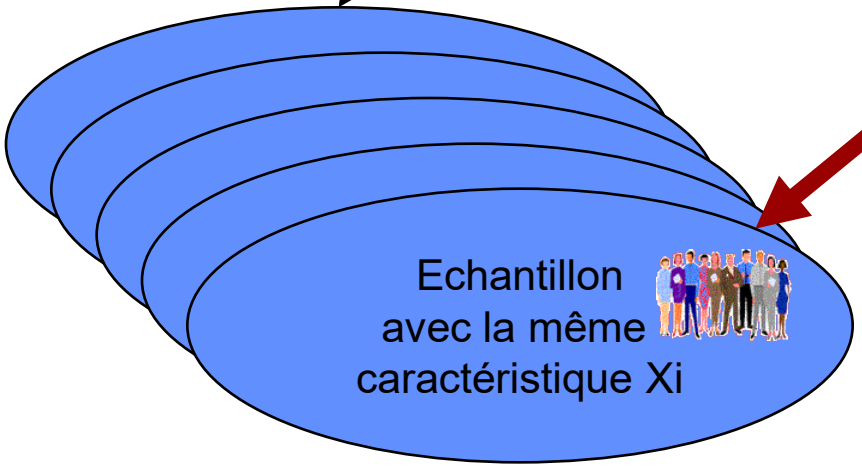
Synthèse


Population P
avec une
caractéristique Xi
Loi de
probabilité
régissant la
distribution de
Xi dans P
caractérisée
par M et σ
Echantillon
avec la même
caractéristique Xi

Synthèse

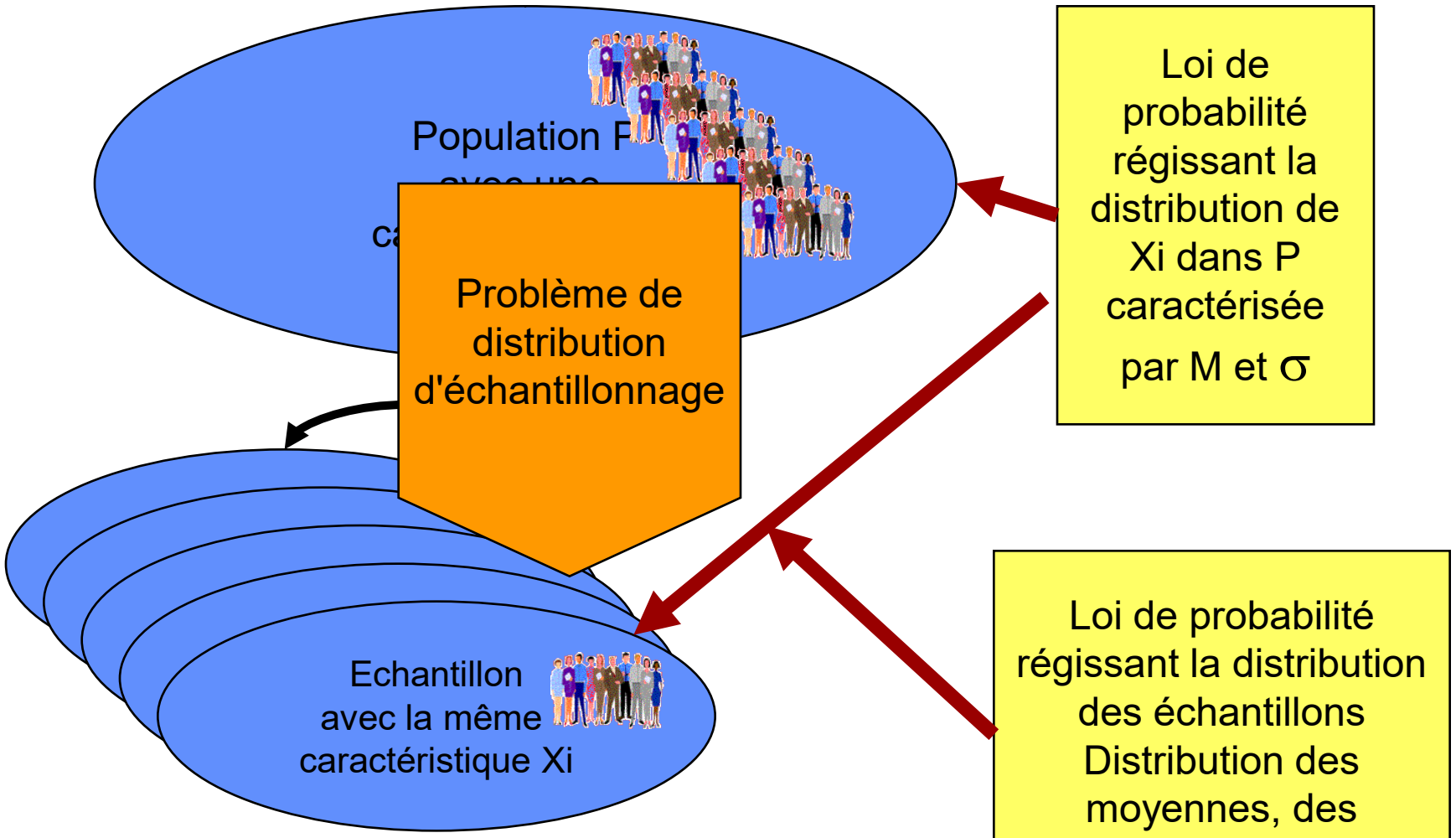


Loi de probabilité régissant la distribution de Xi dans P caractérisée par M et σ

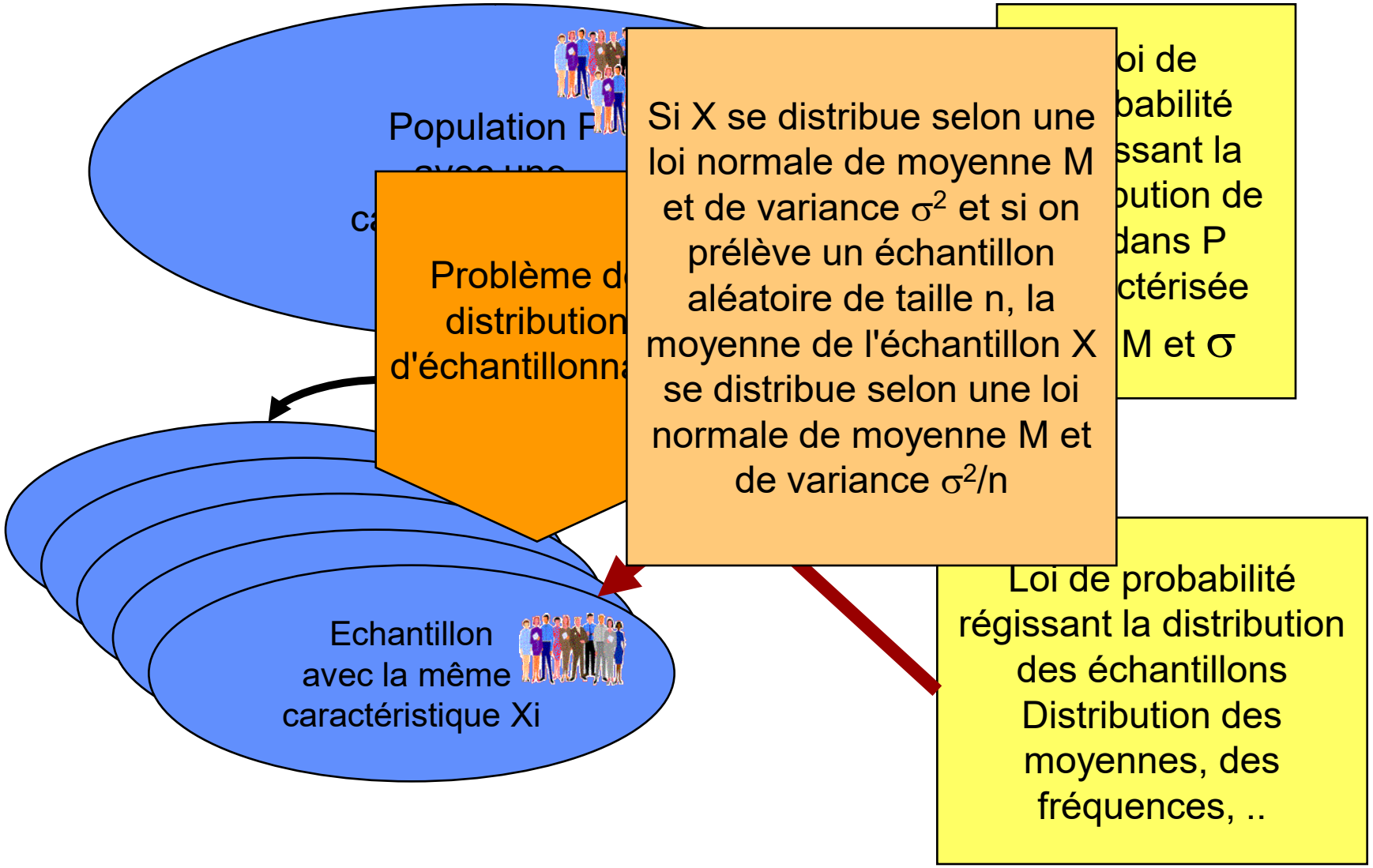


Loi de probabilité régissant la distribution des échantillons Distribution des moyennes, des fréquences, ..

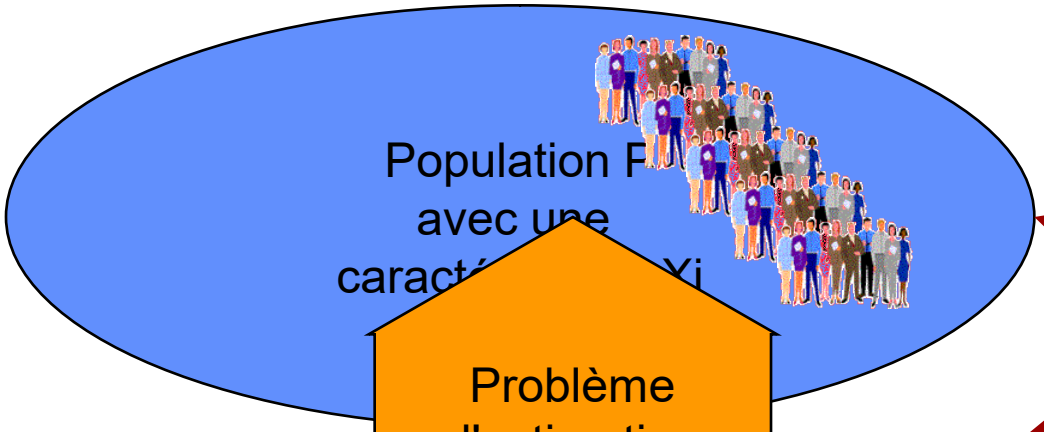
Synthèse


Population P
Problème de distribution d'échantillonnage
Loi de probabilité régissant la distribution de Xi dans P caractérisée par M et σ
Echantillon avec la même caractéristique Xi
Loi de probabilité régissant la distribution des échantillons Distribution des moyennes, des fréquences, ..

Synthèse


Population P
Problème d
distribution
d'échantillonn
Si X se distribue selon une loi normale de moyenne M et de variance σ^2 et si on prélève un échantillon aléatoire de taille n, la moyenne de l'échantillon X se distribue selon une loi normale de moyenne M et de variance σ^2/n
oi de
babilité
ssant la
bution de
dans P
ctérisée
M et σ
Echantillon
avec la même
caractéristique Xi
Loi de probabilité régissant la distribution des échantillons
Distribution des moyennes, des fréquences, ..

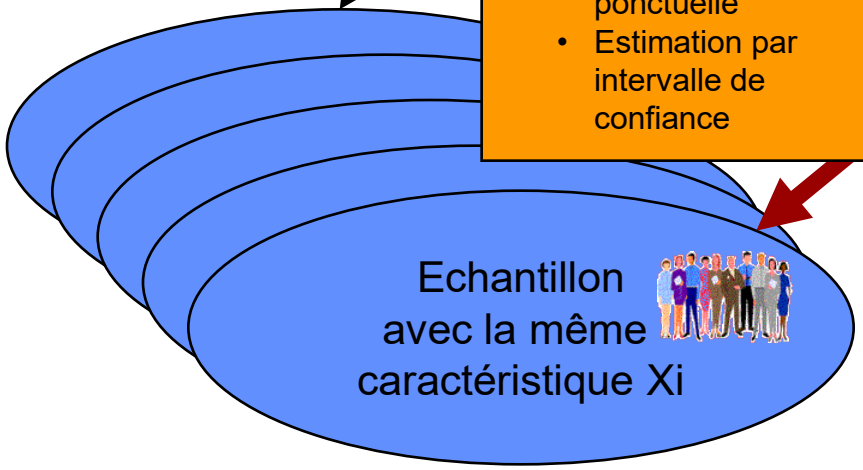
Synthèse



Population P
avec une
caract Xi

Problème d'estimation

- Estimation ponctuelle
- Estimation par intervalle de confiance

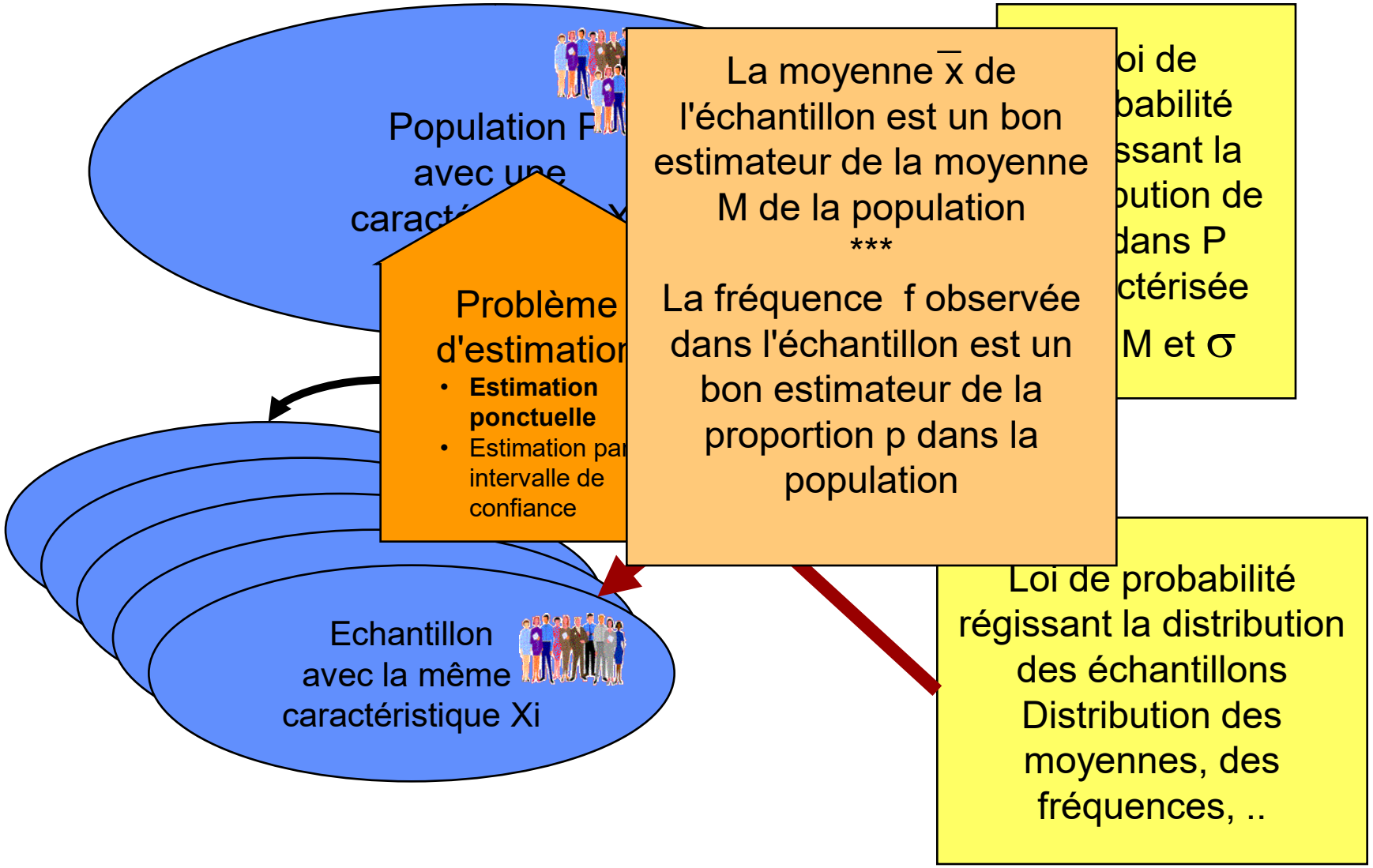


Echantillon avec la même caractéristique Xi

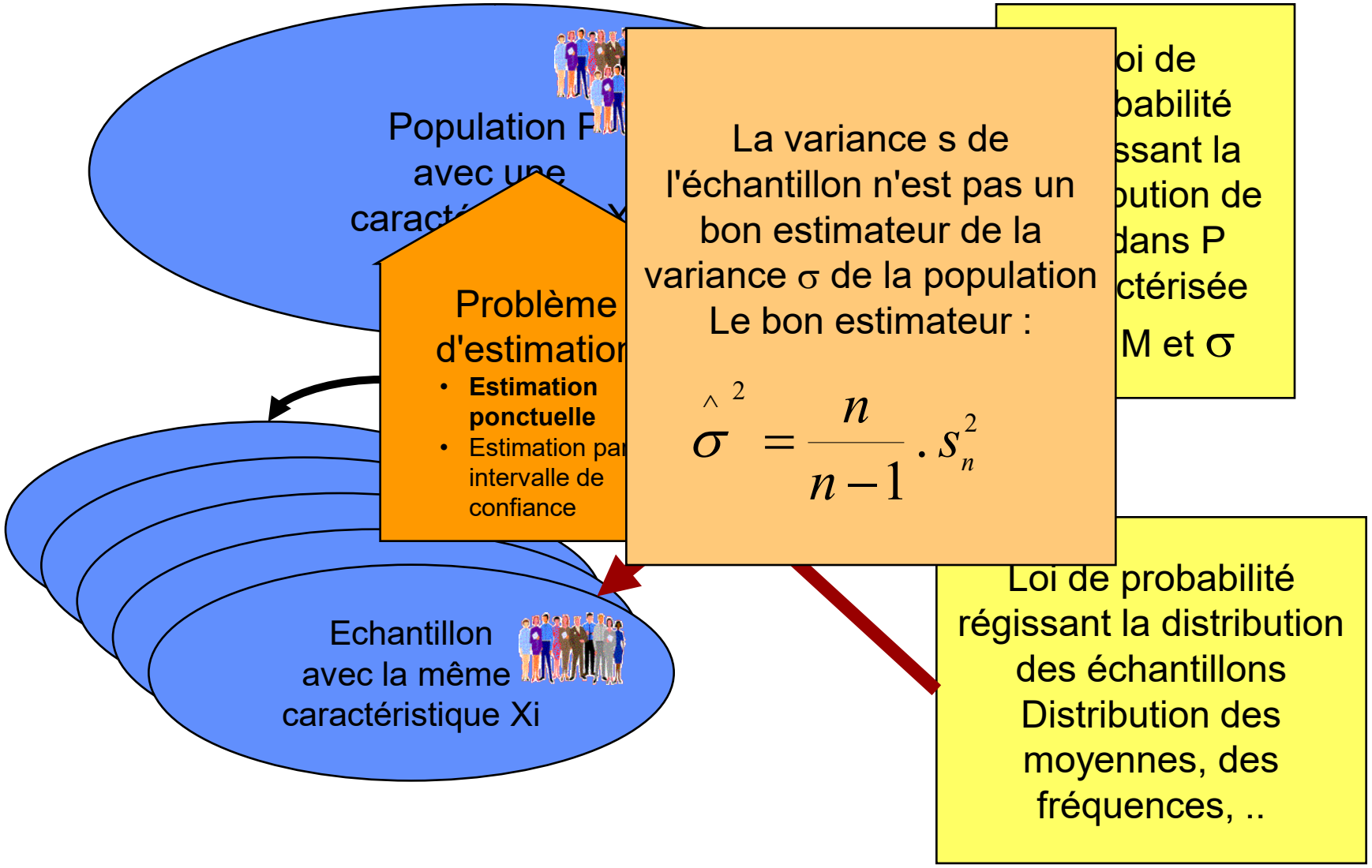
Loi de probabilité régissant la distribution de Xi dans P caractérisée par M et σ

Loi de probabilité régissant la distribution des échantillons
Distribution des moyennes, des fréquences, ..

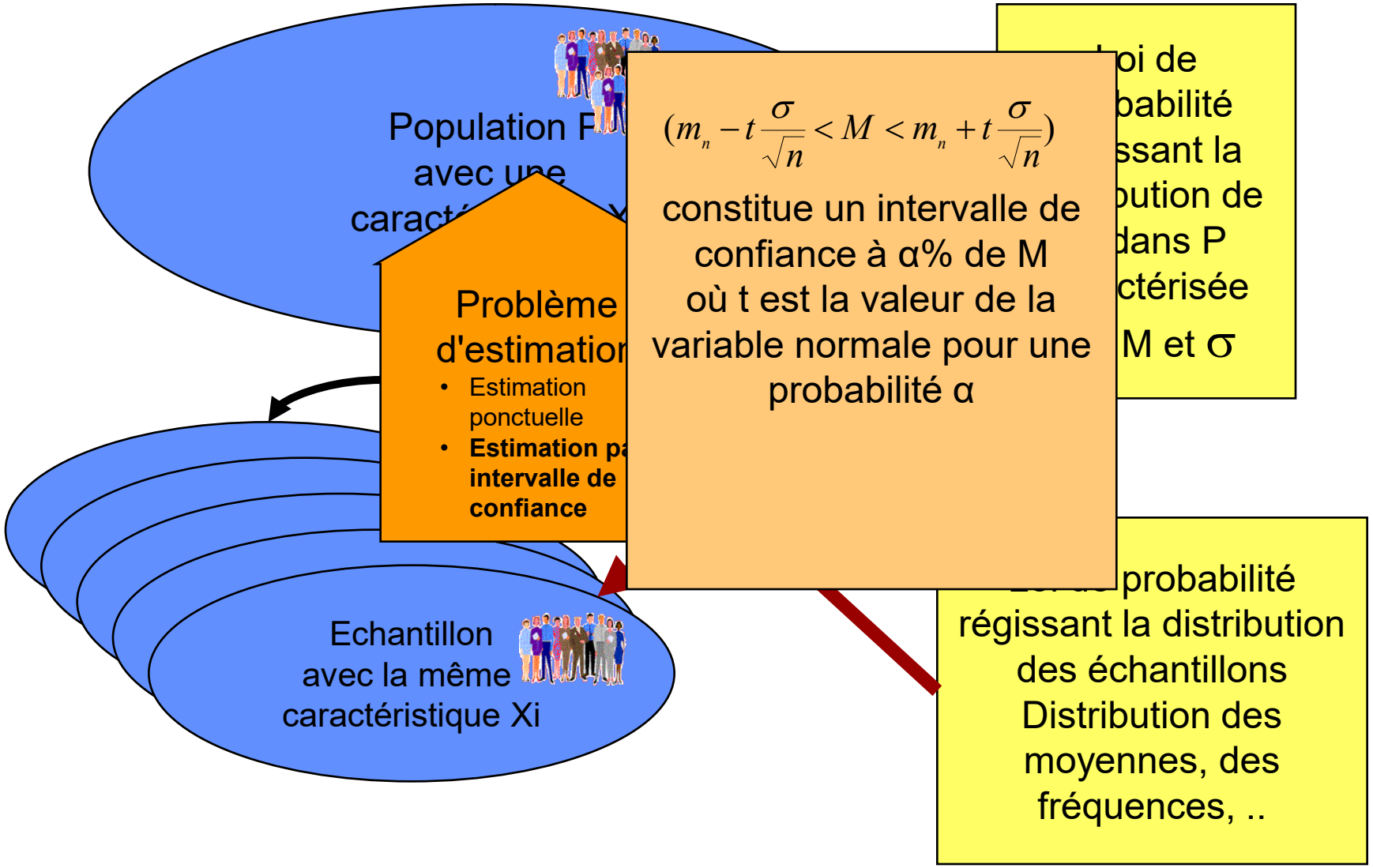
Synthèse


Population P
avec une
caract
Problème
d'estimation
• Estimation ponctuelle
• Estimation par intervalle de confiance
La moyenne $\bar{x}$ de l'échantillon est un bon estimateur de la moyenne M de la population
***
La fréquence f observée dans l'échantillon est un bon estimateur de la proportion p dans la population
oi de
babilité
ssant la
bution de
dans P
ctérisée
M et σ
Echantillon
avec la même
caractéristique Xi
Loi de probabilité régissant la distribution des échantillons
Distribution des moyennes, des fréquences, ..

Synthèse


Population P
avec une
caract
Problème
d'estimation
• Estimation ponctuelle
• Estimation par intervalle de confiance
La variance s de l'échantillon n'est pas un bon estimateur de la variance σ de la population
Le bon estimateur :
$\hat{\sigma}^2 = \frac{n}{n-1} \cdot s_n^2$
oi de
babilité
ssant la
bution de
dans P
ctérisée
M et σ
Echantillon
avec la même
caractéristique Xi
Loi de probabilité régissant la distribution des échantillons
Distribution des moyennes, des fréquences, ..

Synthèse


Population P
avec une
caract
Problème
d'estimation
• Estimation ponctuelle
• **Estimation par intervalle de confiance**
$(m_n - t\frac{\sigma}{\sqrt{n}} < M < m_n + t\frac{\sigma}{\sqrt{n}})$
constitue un intervalle de confiance à α% de M
où t est la valeur de la variable normale pour une probabilité α
oi de
babilité
ssant la
ution de
dans P
ctérisée
M et σ
probabilité
régissant la distribution
des échantillons
Distribution des
moyennes, des
fréquences, ..
Echantillon
avec la même
caractéristique Xi

Exercice

- Une machine d'une chaîne de fabrication découpe des verres de montres dont le diamètre doit être égal à 30 mm.
- Une certaine tolérance, toutefois est acceptée et le disque est considéré conforme si son diamètre est compris entre 29,950 mm et 30,050 mm.
- Les diamètres des verres de montres sont supposés suivre une loi normale.
- Le contrôle de la qualité de la production est fait par un échantillonnage : chaque jour un échantillon de 50 verres est extrait, de façon aléatoire, de la production des 1000 verres fabriqués quotidiennement.
- On obtient, lors d'un contrôle, les résultats suivants :

Diamètre	[29,90;29,95]	[29,95;29,99]	[29,99;30,01]	[30,01;30,05]	[30,05;30,10]
Nbre de verres observés	5	12	22	9	2

Exercice

- Calculez le diamètre moyen et l'écart type de ce diamètre dans l'échantillon ainsi que la proportion de verres conformes
- En déduire une estimation ponctuelle de ses trois paramètres dans la production
- Par intervalle de confiance, au seuil de 98%, estimer le diamètre moyen d'un verre dans la production. Selon la règle énoncée par la direction, si le diamètre moyen estimé est compris dans l'intervalle [29.98 ; 30,02], la qualité est décidée "bonne" sinon, la qualité de la production est décidée "mauvaise" et un réglage de la machine est immédiatement mis en place.
  Etant donné l'échantillon, quelle décision doit-on prendre ?
- Estimer, par intervalle de confiance au seuil de 95%, le nombre de verres conformes produits chaque jour.
- Quelle taille d'échantillon faudrait-il choisir pour que la précision relative de l'estimation de ce nombre de verres conformes soit égale à 10% pour le même seuil de 95%

Test d'hypothèses

- Tests de comparaison
  - à un standard
  - entre deux échantillons

Test d'hypothèse

- Nous avons défini un test d'hypothèse statistique comme une procédure d'acceptation ou de rejet d'une hypothèse
- Un test paramétrique consiste à définir une règle de décision concernant la validité d'une hypothèse portant sur la valeur d'un paramètre d'une loi de distribution dans la population
- Les tests non paramétriques sont construits à partir d'une fonction des valeurs observées sur l'échantillon, fonction indépendante de la loi de distribution dans la population. Un bon exemple de ce type de test est le test du χ^2

Test d'hypothèse non paramétrique : le $\chi 2$

- Le test du χ^2 constitue la troisième et dernière étape de la modélisation d'un phénomène statistique par une loi de probabilité.
- 1ere étape : statistique descriptive via une distribution empirique
- 2ème étape : ajustement d'une loi de probabilité à la distribution empirique
- 3ème étape : test de la validité de l'ajustement effectué. C'est ici que prend place le test du χ^2

Test d'hypothèse non paramétrique : le $\chi 2$

- Soit T1, T2, ... Tν, ν variables normales centrés réduites indépendantes
- Soit χ^2 la somme de leurs carrés
- Cette somme est elle-même une variable aléatoire qui varie entre 0 et l'infini
- Cette variable aléatoire a pour fonction de densité :

$$f(x) = \frac{x^{(\nu/2)-1}\, e^{-(x/2)}}{2^{(\nu/2)}\ \Gamma(\nu/2)} \text{ avec } x = \chi^2 > 0$$

Nous retrouvons la fonction vue en début de chapitre

$E(\chi^2)=\nu \quad V(\chi^2)=2\nu$

Test d'hypothèse non paramétrique : le $\chi 2$

- On dit que c'est une loi du χ^2 à ν degrés de liberté
- La loi du χ^2 est une distribution dissymétrique étalée vers la droite.
- Elle tend à se rapprocher de la distribution normale quand le nombre de degrés de liberté augmente
- Sous Excel, LOI.KHIDEUX(x;d) renvoie la probabilité d'une variable aléatoire x suivant une loi du χ^2 à d degrés de liberté
- KHIDEUX.INVERSE renvoie, pour une probabilité donnée, la valeur de la variable aléatoire suivant une loi du χ^2
- Construction table

Test d'hypothèse non paramétrique : le $\chi 2$

- Pour $\nu = 9$, la valeur du χ^2 a une probabilité de 75% d'être supérieure à 5,90 et de 5% d'être supérieure à 16,9

	Nb deg liberte	1	2	3	4	5	6	7	8	9
x										
0,005	0,995	3,927E-05	0,01002508	0,07172177	0,20698909	0,4117419	0,67572678	0,98925569	1,34441309	1,73493291
0,01	0,99	0,00015709	0,02010067	0,1148318	0,29710948	0,55429808	0,87209033	1,23904231	1,64649738	2,08790074
0,025	0,975	0,00098207	0,05063562	0,21579528	0,48441856	0,83121162	1,23734425	1,68986919	2,17973075	2,70038952
0,05	0,95	0,00393214	0,10258659	0,35184632	0,71072302	1,14547623	1,6353829	2,16734992	2,7326368	3,32511286
0,1	0,9	0,01579077	0,21072103	0,58437437	1,06362322	1,61030799	2,20413068	2,83310693	3,48953913	4,16815904
0,25	**0,75**	0,10153104	0,57536415	1,21253292	1,92255756	2,67460285	3,45459887	4,25485221	5,07064054	**5,898826**
0,5	0,5	0,45493643	1,38629438	2,36597389	3,356694	4,35146022	5,34812084	6,34581137	7,34412163	8,34283278
0,75	0,25	1,32330472	2,77258872	4,1083445	5,38526906	6,62567989	7,84080412	9,03714745	10,218855	11,3887515
0,9	0,1	2,70554397	4,60517019	6,25138846	7,77944034	9,23635694	10,6446407	12,0170366	13,3615661	14,6836566
0,95	**0,05**	3,84145915	5,99146455	7,81472776	9,48772904	11,0704978	12,5915872	14,0671404	15,5073131	**16,91898**
0,975	0,025	5,02388647	7,37775891	9,34840357	11,1432868	12,832502	14,4493753	16,0127643	17,5345461	19,0227678
0,99	0,01	6,63489671	9,21034037	11,3448667	13,2767041	15,0862725	16,8118938	18,4753069	20,090235	21,6659943
0,995	0,005	7,87943869	10,5966347	12,8381564	14,860259	16,7496024	18,5475842	20,2777399	21,954955	23,5893508

Test d'hypothèse non paramétrique : le $\chi 2$

- Pour $\nu = 9$, la valeur du χ^2 a une probabilité de 75% d'être supérieure à 5,90 et de 5% d'être supérieure à 16,9

	Nb deg liberte	1	2	3	4	5	6	7	8	9
x										
0,005	0,995	3,927E-05	0,01002508	0,07172177	0,20698909	0,4117419	0,67572678	0,98925569	1,34441309	1,73493291
0,01	0,99	0,00015709	0,02010067	0,1148318	0,29710948	0,55429808	0,87209033	1,23904231	1,64649738	2,08790074
0,025	0,975	0,00098207	0,05063562	0,21579528	0,48441856	0,83121162	1,23734425	1,68986919	2,17973075	2,70038952
0,05	0,95	0,00393214	0,10258659	0,35184632	0,71072302	1,14547623	1,6353829	2,16734992	2,7326368	3,32511286
0,1	0,9	0,01579077	0,21072103	0,58437437	1,06362322	1,61030799	2,20413068	2,83310693	3,48953913	4,16815904
0,25	**0,75**	0,10153104	0,57536415	1,21253292	1,92255756	2,67460285	3,45459887	4,25485221	5,07064054	**5,898826**
0,5	0,5	0,45493643	1,38629438	2,36597389	3,356694	4,35146022	5,34812084	6,34581137	7,34412163	8,34283278
0,75	0,25	1,32330472	2,77258872	4,1083445	5,38526906	6,62567989	7,84080412	9,03714745	10,218855	11,3887515
0,9	0,1	2,70554397	4,60517019	6,25138846	7,77944034	9,23635694	10,6446407	12,0170366	13,3615661	14,6836566
0,95	**0,05**	3,84145915	5,99146455	7,81472776	9,48772904	11,0704978	12,5915872	14,0671404	15,5073131	**16,91898**
0,975	0,025	5,02388647	7,37775891	9,34840357	11,1432868	12,832502	14,4493753	16,0127643	17,5345461	19,0227678
0,99	0,01	6,63489671	9,21034037	11,3448667	13,2767041	15,0862725	16,8118938	18,4753069	20,090235	21,6659943
0,995	0,005	7,87943869	10,5966347	12,8381564	14,860259	16,7496024	18,5475842	20,2777399	21,954955	23,5893508

=KHIDEUX.INVERSE(x;ν)

Test d'hypothèse non paramétrique : le $\chi 2$

- Principe du test
- Les écarts entre la distribution observée et la distribution ajustée à la loi peuvent être de deux causes :
  - Une fluctation normale d'échantillonnage (l'échantillon est un extrait de la population) avec des écarts faibles
  - L'ajustement n'a pas lieu d'être, avec un écart supérieur avec des écarts élevés
- Cet écart va être mesuré par la **distance** existante entre la théorique ajustée et la distribution observée
- Cette distance étant une grandeur aléatoire, elle est mesurée par une loi de probabilité

Test d'hypothèse non paramétrique : le $\chi 2$

- Cette loi permet de calculer la probabilité d'obtenir une distance supérieure à la distance observée
- On se fixe un seuil de probabilité α dit **seuil de confiance**
- Si la probabilité obtenue est inférieure au seuil de confiance, on rejette l'hypothèse.
- Si la probabilité obtenue est supérieure au seuil de confiance, on accepte l'hypothèse.

Test d'hypothèse non paramétrique : le $\chi 2$

- L'expérience concerne N observations classées selon k modalités (classes de valeurs)
- A chaque modalité Ci correspond un effectif Ni et la probabilité pi déterminé par la loi P de probabilité théorique
- La distance d :

$$d = \sum_{i=1}^{k} e_1^2 = \sum_{i=1}^{k} \frac{(N_i - N_{pi})^2}{N_{pi}}$$

- Avant la réalisation des observation, d est une variable aléatoire qui suit une loi du χ^2 à ν=k-1 degrés de liberté

Test d'hypothèse non paramétrique : le $\chi 2$

- Après la réalisation, on est à même de calculer d
- On connaît pas a priori la loi P. Celle-ci est ajustée d'après la distribution observée.
- Si l'ajustement de la loi théorique a nécessité l'estimation de r paramètres à partir des observations, la distance d suit, dans l'hypothèse où la distribution théorique est effectivement la loi ajustée, une loi du χ^2 à ν=k-r-1 degrés de liberté
- Tout ceci repose sur une distribution normale des écats qui implique des effectifs suffisamment grands (4,5) dans chaque modalité, d'où de possibles regroupements
- Seuil α de 2 à 5%

Test d'hypothèse non paramétrique : le $\chi 2$

- Corrigé Pb N o 5
- Doc D1 et D2

Exercice

Test d'hypothèse paramétrique

- Construire un test suppose le processus suivant :

1. Choix des hypothèses
   - H0 Hypothèse dite nulle, c'est l'hypothèse qui sera privilégiée
   - H1 contre laquelle on teste H0
2. Détermination de la variable de décision D
   - En supposant l'hypothèse H0 vraie, la loi de probabilité de la variable D doit être parfaitement déterminée
3. Choix du risque α, de 1ère espèce, ce risque correspond à la probabilité de rejeter, à tort, l'hypothèse H0
   - Détermination de la région critique , ensemble des valeurs de D conduisant au rejet de H0. Cette région dépend de l'hypothèse H1
4. Enoncé de la règle de décision

Test de comparaison d'un paramètre à une norme

- Comparaison d'une moyenne à une norme
  - On utilise la variable de décision $\overline{X}$
- Comparaison d'une proportion à une norme
  - On utilise la variable de décision F

Test de comparaison d'un paramètre à une norme

- Exemple
- On considère un échantillon d'effectif 50 dans lequel on mesure $\overline{x} = 42$, $s = 7$
- Tester H0 : M=40 contre H1 M>40
- La variable de décision adaptée est X
- X est régie par une loi normale $\mathcal{N}$ $(M; \sigma(x)=\sigma / \sqrt{n})$
- Si H0 vrai, M=40
- σ, écart-type de l'age dans la population est inconnu
- Il est estimé par $s/\sqrt{n-1} = 7/7 = 1$
- La loi $\mathcal{N}$ (40;1) puisque N>= 30

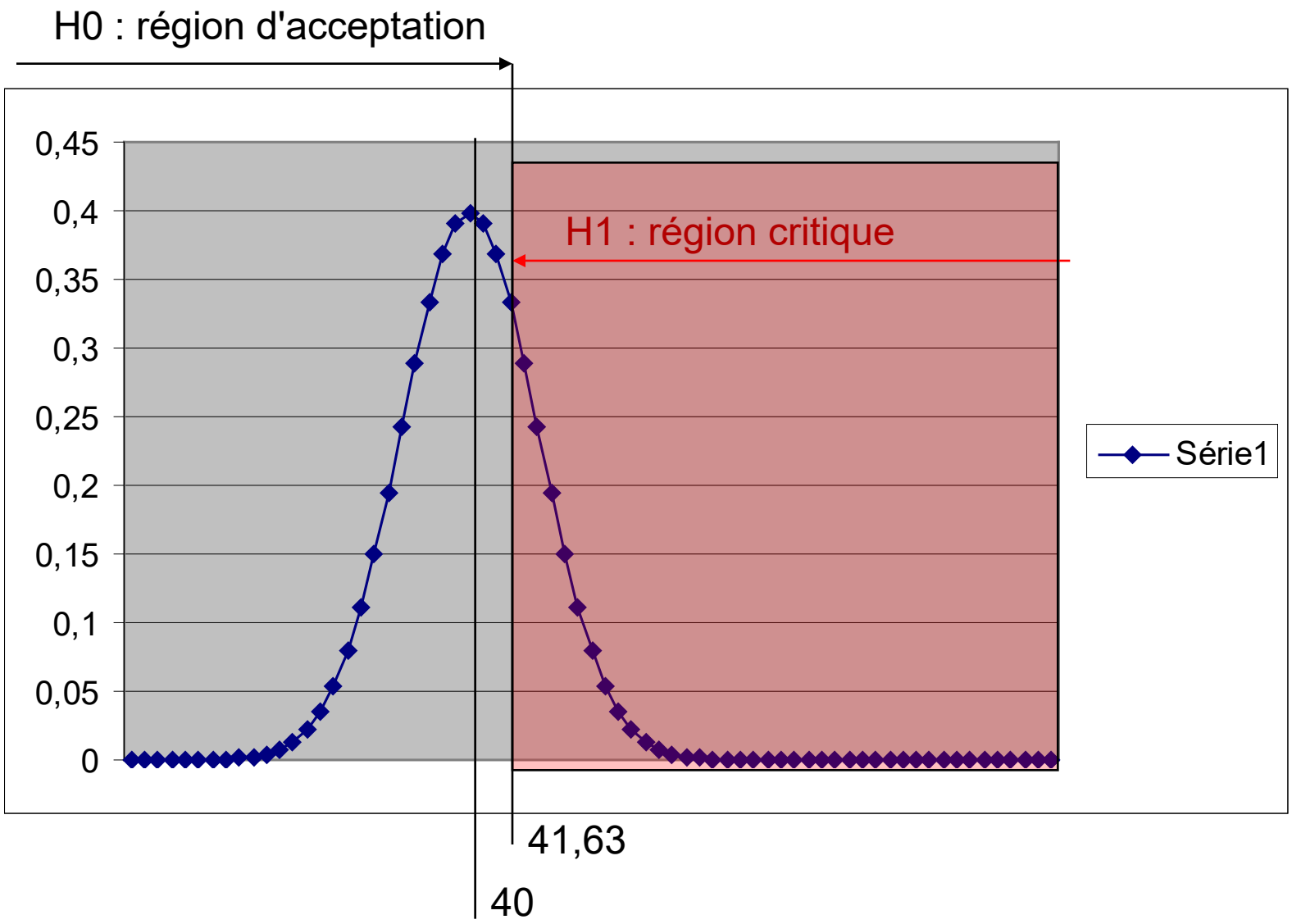
Test de comparaison d'un paramètre à une norme

- La loi $\mathcal{N}$ (40;1) gère l'expérience
- Le seuil de signification étant fixé à α = 5 %, la limite l de la région critique est donnée par :
- P(x<l / M=40) = 0,05 <=> P(T<$t_{0,05}$) = 0,05
- ou T représente la variable normale centrée réduite

$$T = \frac{\overline{x} - M}{\frac{s}{\sqrt{n}}}$$

- La valeur de $t_{0,05}$ dans la table (P(0,1) = 1,645
- D'ou t = 1,645 = l - 40/ (7/$\sqrt{50}$)
- => l = 41,63
- On rejette H0 (42 > 41,63)

Test de comparaison d'un paramètre à une norme



Comparaison d'échantillons

- Soit 2 échantillons A et B
- A : m_A, σ_A, P_A
- B : m_B, σ_B, P_B
- On considère que les échantillons sont indépendants

Comparaison d'échantillons : comparaison moyennes

- Les hypothèses :
- H0 => mA = mB
- H1 => mA $\neq$ mB
- Pour effectuer ce test, on le transforme en
- H0 => mA - mB = 0
- H1 => mA - mB $\neq$ 0
- On utilise la variable de décision $\overline{X_A}$ - $\overline{X_B}$
- qui répond à une loi normale

$$\mathcal{N}(m_A - m_B; \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}})$$

Comparaison d'échantillons : comparaison proportions

- Les hypothèses :
- H0 => pA = pB
- H1 => pA ≠ pB
- Pour effectuer ce test, on le transforme en
- H0 => pA - pB = 0
- H1 => pA - pB ≠ 0
- On utilise la variable de décision $F_A - F_B$
- qui répond à une loi normale

$$\mathcal{N}\left(p_A - p_B ; \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}\right)$$

Comparaison d'échantillons : comparaison proportions

- Dans une entreprise, on tire un échantillon de 150 personnes. 30 femmes parmi 50 ont un salaire mensuel inférieur à 2000 € alors que 65 hommes parmi 100 ont un salaire mensuel inférieur à 2000 €
- Peut on considérer, au seuil de 5%, que la proportion de salaires inférieurs à 2000 est la même chez les femmes que chez les hommes.

Test d'ajustement d'une distribution statistique par une loi de probabilité

- Principe identique
- Les hypothèses :
- H0 => X suit telle loi de probabilité
- H1 => X ne suit pas telle loi de probabilité
- On utilise la variable de décision D
- Sous H0

$$D = \sum \frac{(N_{ith} - n_{iobs})^2}{N_{ith}}$$

- Avec Nith = effectif théorique qui serait observé si H0 vrai
- Nobs = effectif observé dans distribution empirique

Test d'ajustement d'une distribution statistique par une loi de probabilité

- D suit une loi du $\chi 2(\nu)$
- avec ν : paramètre loi du $\chi 2$ = k - r - 1
- k = nombre de modalités
- r = nombre de paramètres

- Détermination d'une région critique, de seuil α, correspondant au risque de 1ere espèce (Probabilité d'avoir H1 vrai alors que H0 est considéré comme vrai, d'ou contradiction)
- Probabilité β d'avoir H0 vrai alors que H1 est considéré comme vrai (contradiction inverse : risque 2ème espèce)

- **Modèle RappelBase**

Exercices

- **Le bruit émis par les avions doit être inférieur à 80 décibels dans les zones voisines, sinon l'aéroport doit indemniser les riverains**
- **Ceux-ci affirment que le niveau de bruit atteint effectivement 80 décibels alors que l'aéroport affirme qu'il n'est que de 78 décibels.**
- **Des experts font des mesures en prélevant un échantillon de n= 100 et s2 = 49**
- **1: Que signifie le choix H0 m = 80 et H1 : m<80**
- **2. H0 m=80**
  - **Quelle région critique ?**
  - **Que faire si moyenne de l'échantillon est 79,1**