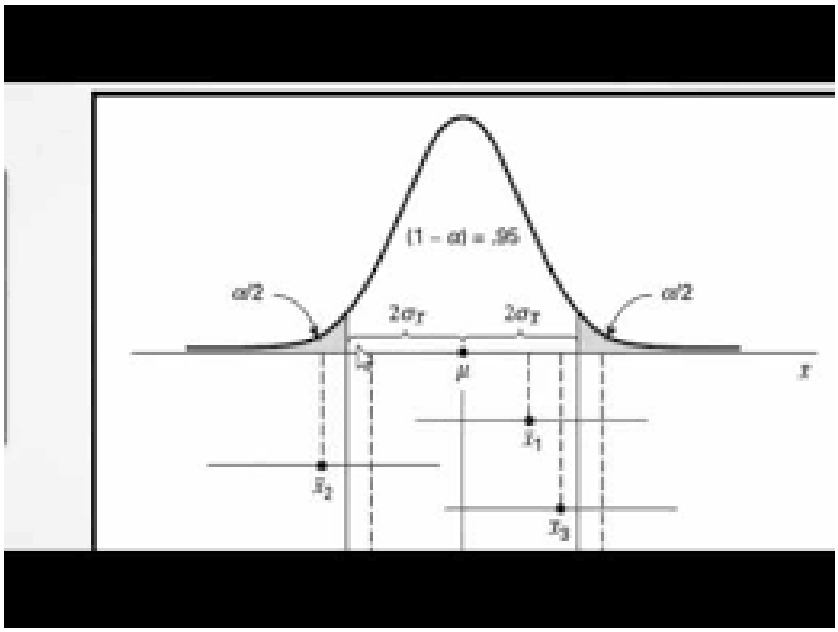
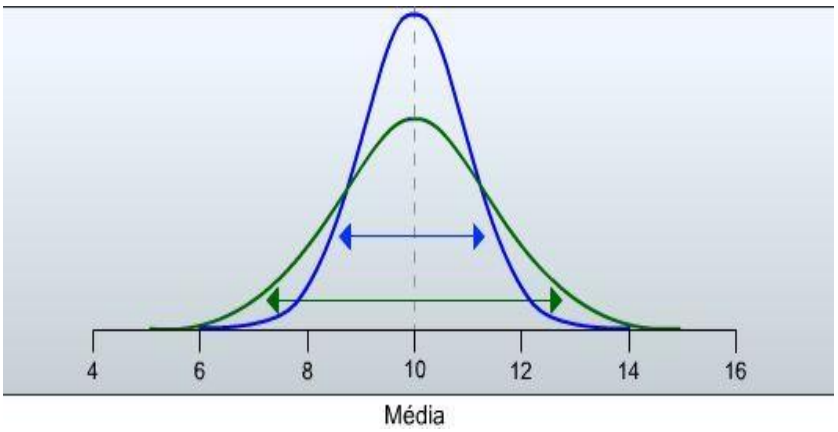
INSTITUT SUPERIEUR DES ETUDES TECHNOLOGIQUES DE RADES
DÉPARTEMENT DES SCIENCES ÉCONOMIQUES ET DE GESTION
1ère Année de Licence appliquée en CD , TCF, et MA



COURS DE STATISTIQUES INFERENTIELLES



PREPARE PAR :
Yakdhane ABASSI

Année universitaire : 2025-2026

SOMMAIRE

Introduction générale

L'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon issu de cette population. Les caractéristiques de l'échantillon, une fois connues, reflètent avec une certaine marge d'erreur possible celles de la population. Les statistiques inférentielles reposent sur divers méthodes d'échantillonnage qui permettent de fournir soit une estimation ponctuelle de paramètres inconnue de la population soit une estimation par un intervalle de confiance. Elles permettent en outre à formuler des tests d'hypothèse utilisés dans différentes optiques. En effet ces tests permettent de juger la conformité de paramètres inconnus de la population avec certaines valeurs prédéterminées (tests paramétriques). Ces tests sont aussi utilisés pour vérifier si les valeurs observées d'une population suit ou pas une loi statistique théorique susceptible d'encadrer ses probabilités de réalisation (tests d'ajustements). Ces tests permettent en outre de se prononcer sur la présence ou pas de lien entre deux variables aléatoires (test d'indépendance).

Les résultats issus des statistiques inférentielles permettent de guider la décision dans divers domaines de gestion notamment au niveau de la gestion de la production (contrôle de qualité, choix des fournisseurs …) du marketing (estimation de la proportion d'acheteurs d'un produit, estimation d'impact d'actions commerciales, similitude entre plusieurs catégories de clients..) et de la gestion financière (gestion des crédits, gestion de portefeuille titres…)

Chapitre 1 :

Notions de variables aléatoires et lois de probabilité

I-Rappel sur des notions de probabilités :

On appelle épreuve aléatoire, une épreuve dont le résultat dépend du hasard. Le résultat est donc incertain. On note Ω l'ensemble des résultats possible d'une épreuve aléatoire et Ƥ (Ω) l'ensemble des parties possibles de Ω. On appelle probabilité P sur (Ω, Ƥ (Ω)) toute application de Ƥ (Ω) dans [0 ,1] telle que P(Ω) = 1, et toute famille (A_n) $n \in \mathbb{N}$ d'éléments deux à deux disjoints de Ƥ (Ω), on a $P(\bigcup_n A_n) = \sum_n P(A_n)$.

(Ω, Ƥ (Ω) , P) forme alors un espace probabilisée. Pour tout éléments A et B

$$\forall A, B \in P(\Omega), P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$$

$$\forall A, B \in P(\Omega), P(A/B) = \frac{P(A \bigcap B)}{P(B)}$$

Si A et B sont indépendants alors

$$P(A \bigcap B) = P(A) x P(B)$$

Si tous les éléments de Ω sont équiprobables alors P(A) = CARD (A) / CARD (Ω)

II- Notion de variable aléatoire :

On appelle variable aléatoire (V.A) le résultat caractéristique d'une épreuve aléatoire. De façon conventionnelle, on notera toujours par une majuscule (exemple X) la variable aléatoire et par des minuscules (exemple x_i) les valeurs qu'elle peut prendre.

Exemple 1 :

X la V.A : « Résultat d'un jet de dé »

$\{x_i\} = \{1,2,3,4,5,6\}$

Exemple 2 :

X la V.A : « Taille d'une personne tirée dans un échantillon »

$x_i : \{[1m ; 1,5m[; [1,5m, 2m[; [2m, 3m[\}$

Dans le premier cas, la variable aléatoire est dite discrète car elle prend uniquement des valeurs isolées. Dans le second cas la variable aléatoire est dite continue du fait qu'elle prend n'importe quelle valeur réelle à l'intérieur d'un intervalle.

III- Loi de probabilité :

1- Cas d'une VA discrète :

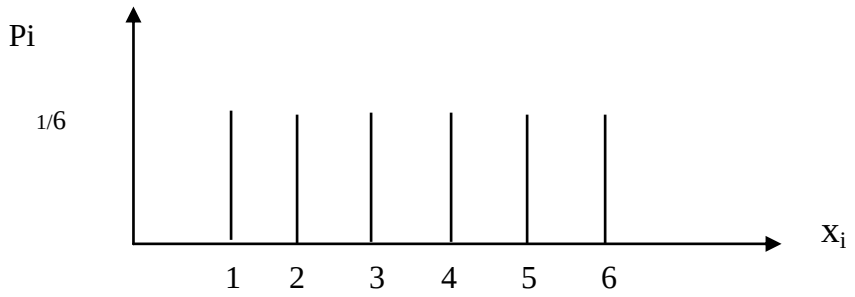
On appelle loi de probabilité d'une VA discrète la relation qui permet de déterminer la probabilité que cette variable prenne une valeur donnée.

Exemple : Soit X la VA : « Résultat d'un jet de dé »

La loi de probabilité de X peut être représentée par le tableau suivant sous réserve que le dé ne soit pas truqué :

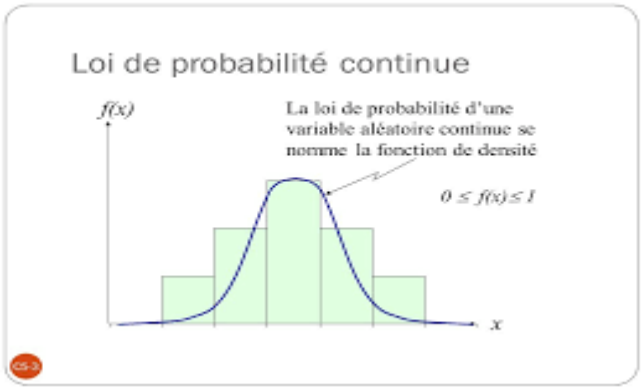
X_i	1	2	3	4	5	6
$P(X= X_i)$	1/6	1/6	1/6	1/6	1/6	1/6

On peut associer à cette loi de probabilité le diagramme en bâtons suivant :



2- Cas d'une VA continue :

On appelle loi de probabilité d'une VA continue la fonction qui permet de déterminer la probabilité que cette variable appartienne à un intervalle.
Cette loi de probabilité s'exprime par une fonction dite fonction de densité de probabilité et notée par f(x). Cette fonction peut être représentée graphiquement par une courbe :



VI- Fonction de répartition :

La fonction de répartition ou fonction cumulative F (x) d'une variable aléatoire X est la fonction qui associe à toute valeur x de X, la probabilité que X soit inférieure ou égale à x : $F(x) = P(X \leq x)$

Propriétés:

* $0 \leq F(x) \leq 1$ (car F(x) est une probabilité)
* $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1)$

Remarque:

Graphiquement la fonction de répartition d'une VA discrète sera un diagramme en escaliers alors qu'il s'agit d'une courbe continue et croissante dans le cas d'une VA continue.

V- Fonction de distribution ou de densité de probabilité :

Si la fonction de répartition d'une VA continue X est dérivable, la fonction de distribution ou de densité de probabilité sera alors la fonction dérivée de F.
$f(x) = F'(x)$ On peut alors écrire $\int_{x_1}^{x_2} f(x)\,dx = [F(x)]_{x_1}^{x_2} = F(x_2) - F(x_1) = P(x_1 < X \leq x_2)$

Conséquence : Dans le cas d'une VA continue, la probabilité que cette variable prenne une valeur particulière est nulle.
Démonstration :

$$P(X = y) = P(y \langle X \leq y) = F(y) - F(y) = 0$$

VI- Notion d'espérance et de variance :

1- Cas d'une VA discrète :

Dans le cas d'une VA discrète l'espérance et la variance sont données par :

$$E(X) = \sum_{i=1}^{N} p_i x_i$$

$$V(X) = \sum_{i=1}^{N} p_i (x_i - E(X))^2$$

$$= \sum_{i=1}^{i=N} p_i x_i^2 - E(X)^2$$

2- Cas d'une VA continue :

Dans le cas d'une VA continue l'espérance est la variance sont données par :

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

$$V(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x) dx$$

3- Signification de l'espérance mathématique :

C'est le résultat moyen que l'on doit s'attendre à obtenir sur un grand nombre d'épreuves.
Exemple : dans un jeu de pile ou face un individu A donne 4 dinars à B chaque fois que le côté pile apparaît et reçoit 2 dinars dans le cas ou c'est le côté face qui apparaît.
Soit X la VA : « Gain de l'individu A »

x_i	- 4	2
$P(X=x_i)$	0.5	0.5

E(X)= 0.5 (-4) + 0.5(2) = -1
Donc on peut conclure que l'individu A doit s'attendre à perdre en moyenne 1 dinar en jouant à ce jeu, var l'espérance mathématique de son gain est négative –1.

VII- Opérations sur les VA :

Soient deux variables aléatoires X et Y, on aura alors :
E(X+Y) = E(X)+E(Y)

E(X-Y) = E(X)-E(Y)

E(aX) = a E(X) ; E(X+b) = E(X)+b ; E(aX+b) = a E(X) + b

$V(aX) = a^2V(X)$; V(X+b) = V(X) ; $V(aX+b) = a^2V(X)$;

$V(aX+bY) = a^2V(X) + b^2 V(Y) + 2ab\ COV(X,Y)$

Si X et Y sont indépendants on aura alors:

V(X+Y) = V(X) + V(Y)

V(X-Y)= V(X) + V(Y)

Chapitre 2 : Les lois usuelles de probabilité

I- La loi binomiale :

Soit une épreuve aléatoire dans laquelle il n'existe que deux résultats possibles, l'un étant qualifié de favorable et l'autre de défavorable (échec et succès).

Soit p la proportion du cas favorable q=1-p celle du cas défavorable. On réalise n'épreuves identiques et indépendantes et on notera :

X la VA : « Nombre de résultats favorables »

La probabilité que cette variable prenne une valeur particulière x est :

$$P(X = x) = C_n^x P^x (1-P)^{n-x} = C_n^x P^x q^{n-x}$$

Cette relation définit la loi binomiale de probabilité. Elle dépend de deux paramètres n et p. Il existe différentes tables de probabilité qui indiquent la probabilité simple P(X=x) ou la probabilité cumulé $P(X \leq x)$

1- Conditions d'application de la loi :

- L'épreuve est dichotomique, elle ne comporte que deux résultats possibles.
- Les répétitions de cette épreuve sont mutuellement indépendantes, ainsi la probabilité du cas favorable p est constante.

2- Caractéristiques de la loi :

Soit X une VA qui suit une loi binomiale de paramètres n et p. On notera alors :

$X \longrightarrow B(n,p)$

Les paramètres de cette loi sont :

E(X)= np

V(X) = np (1-p)=npq

Exemple:

On jette dix fois une pièce de monnaie et on considère X la VA nombre de faces obtenues.

On aura alors $X \longrightarrow B(10, 0,5)$

D'ou: E(X) = 10x0.5=5

V(X)= 10x0.5x 0.5=2.5

II- La loi de poisson :

Lorsque le nombre des épreuves devient très grand (n), et que la probabilité du cas favorable devient très faible (p) on démontre que la loi binomiale tend vers la forme limite :

$$P(X=x)=\frac{(np)^x e^{-np}}{x!}$$

Cette nouvelle loi ainsi définie est appelée loi de poisson qui est qualifiée de loi des événements rares du fait que p est très faible.

On note souvent $\lambda = np$ avec $X \longrightarrow P(\lambda)$ d'ou :

$$P(X=x)=\frac{(\lambda)^x e^{-\lambda}}{x!}$$

1- Conditions d'application de la loi :

- Les conditions d'application de la loi binomiale sont réunies (épreuve dichotomique)
- La probabilité du cas favorable est faible (événement rare)
- Le nombre des épreuves est très grand

2- Caractéristiques de la loi :

Soit X une VA tel que : $X \quad P \longrightarrow P(\lambda)$ les paramètres de cette loi sont :

$E(X)= \lambda$

$V(X)= \lambda$

Il existe différentes tables de probabilité de la loi de poisson qui indiquent la probabilité simple($X=x$) ou la probabilité cumulé $P(X \leq x)$. Cette loi s'applique au cas d'un nombre rare d'évènement pendant une période de temps file d'attente. Si un évènement surgit avec une proportion p pendant une période de temps limité t(au point que l'évènement ne peut se produire au plus qu'une seule fois pendant t) , alors la probabilité que ledit évènement se produise x fois pendant dans une période plus longue T est régie par une loi de poisson de paramètre $\lambda = n\,p$ avec $n = T/t$

Exemple : Supposons qu'un standardiste reçoit deux appels par minute, t doit correspondre à une unité de temps plus faible que la minute soit, la seconde avec une probabilité de recevoir un appel de 2/60= 1/30, le nombre d'appels reçus pendant cinq minutes suit une loi de poisson de paramètre $\lambda = n\,p = 300/30 = 10$

III- La loi normale :

La loi normale (ou loi de Gauss- Laplace) est une loi de probabilité continue dont la densité de probabilité est donnée par l'expression :

$$f(x) = \frac{1}{\sigma\sqrt{2II}} \exp^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}$$

m et σ sont les paramètres de la loi, respectivement sa moyenne et son écart type.

Pour noter qu'une variable aléatoire X suit une loi de moyenne m et d'écart type σ on écrit : $X \longrightarrow N(m, \sigma)$

1- Caractéristiques de la loi :

- Graphiquement, la courbe représentative de la loi normale à l'allure d'une courbe en cloche : elle est symétrique par rapport à m abscisse pour lequel elle passe par un maximum et elle admet l'axe des x comme asymptote.
- La fonction y=P(x) est entièrement définie et connue dès l'instant où l'on connaît les valeurs des paramètres m et σ .

2- Conditions d'application de la loi :

La loi normale tire son importance du fait que sous certaines conditions plusieurs autres lois de probabilité convergente vers elle. On démontre notamment que la somme de variables aléatoires indépendantes obéissant à des lois de probabilité quelconques tend vers une loi normale lorsque le nombre de ces variables augmente indéfiniment (théorique central limite).

D'autre part, on peut dire qu'une variable X est distribué normalement si :

- Les facteurs de variation de X sont nombreux
- Les fluctuations de X dues à ces différents facteurs sont indépendantes les uns des autres
- Les fluctuations de X dues à ces différents facteurs doivent être suffisamment petites et approximativement du même ordre de grandeur.

3- La loi normale centrée réduite :

On considère la variable aléatoire $X \longrightarrow N(m, \sigma)$ et on veut calculer la probabilité suivante : $P(X<x_1)$. Le problème c'est qu'on ne dispose pas d'une table de probabilité pour les différentes lois normales (suivant les valeurs de m et σ) car il faudra prévoir une infinité.

Il faut alors procéder au changement de variable suivant :

$$U = \frac{X - m}{\sigma}$$

Cette nouvelle variable aléatoire U obéit à une loi normale réduite (LNCR) du fait que :

E(U)= m=0 (variable centrée)

V(U)= σ =1 (variable réduite)

Ainsi la probabilité qu'on veut calculer sera : $P(\frac{X - m}{\sigma} \langle \frac{x_1 - m}{\sigma})$

L'intérêt de ce changement de variable est qu'il permet de ramener n'importe quelle distribution normale à une même loi de probabilité qui est la loi normale centrée réduite N(0,1). Pour laquelle on dispose de tables de probabilité. Puisque cette loi est symétrique alors on a :

$U \rightarrow$ N(0,1)

Si P (U< a) , a > 0 est connue, alors

P (U< - a) = 1 - P (U< a)

P (-a < U < a) = P (U< a) - P (U< - a) = P (U< a) – [1 - P (U< a)]

= 2 P (U< a) - 1

Application :

Soit X⟶ N (100,10).

Calculer les probabilités suivantes :

- P(X<110)
- P(X>105)
- P (110<X<120)

Corrigé :

- P(X<110) = P [(X -100)/10 < (110-100)/10] = P (U < 1) = F (1) = 0.8413
- P (X > 105) = P [(X-100)/10 > (105 – 100)/10] = P (U > 0.5)
  = 1 – P (U < 0.5) = 1 – F (0.5) = 1 - 0.6915 = 0.3085
- P (110<X<120) = P [(110-100)/10 < (X-100)/10 < (120-100)/10]
  = P (1 < U < 2) = F(2) – F (1) = 0.9772 – 0.8413 = 0.1359

Table de la Loi Normale Centrée Réduite

$$X^* \rightarrow N(0,1)$$

$$P\left[X^* \prec t_\alpha\right] = 1-\alpha$$

	0,00	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997

NB : ***Pour avoir la probabilité qui correspond à $t_\alpha = 1.53$ il faut lire l'intersection de la ligne 1.5 et de la colonne 0.03 (1.53 = 1.5 + 0.03) ainsi P[X<1.53] = 0.9370***

IV- Les distributions dérivées de la loi normale :

Les lois de probabilités qui seront présentées dans cette section sont définies à partir de la loi normale

1- Loi de Khi-deux(χ^2) :

Soient n variables aléatoires indépendantes $U_1,U_2,U_3,\ldots\ldots U_n$ suivant chacune la loi normale centrée réduite.

Ainsi la somme : $$U_1^2+U_2^2+U_3^2+.....+U_n^2=\sum_{i=1}^{n}U_i^2$$

Représente une variable aléatoire dont la distribution est celle d'une χ^2 à n degrés de liberté (n ddl) car c'est la somme de n carrés de lois normales centrées et réduites. Cette loi se note $\chi^2(n)$ ou χ^2_n.

Propriétés :

- $E(\chi^2_n)=n$
- $V(\chi^2_n)=2n$

Lecture de la table de probabilité :

La variable χ^2 est tabulée en fonction du nombre de ddl. La table donne pour différentes valeurs de α, la valeur de x telle que : $P(\chi^2_{(n)}> x)= \alpha$

Application :

Calculer $P(\chi2_{(10)}>20.48)$

Une lecture de la table de khi-deux nous donne $P(\chi2_{(10)}>20.48) = 0.025$

2- Loi de Student (T) :

Soient n+1 variables aléatoires indépendantes $U_1,U_2,U_3\ldots\ldots U_n$ et U_{n+1} suivant chacune la loi normale centrée réduite. Ainsi la quantité suivante :

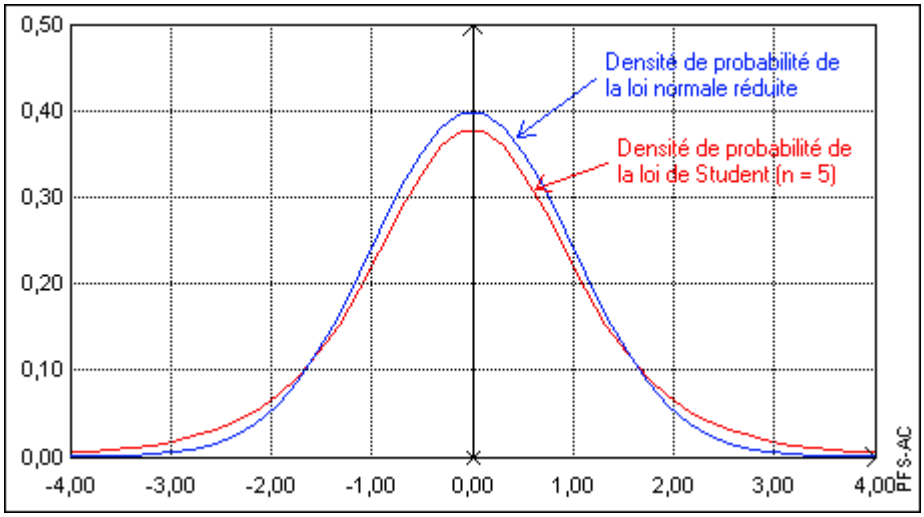
$$T=\frac{U_{n+1}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}U_i^2}}$$

Représente une variable aléatoire dont la distribution est celle d'une loi de Student à n de grés de liberté (n= ddl). Cette loi se note T(n) ou T_n.

Propriétés :

- $E(T_n)=0$
- $V(T_n)= (n/n-2)$

Graphiquement la courbe de densité de probabilité de la loi de Student est une courbe en cloche symétrique par rapport à l'axe des ordonnées, mais qui est plus évasée que celle de la LNCR.



Lecture de la table de probabilité :

La table de Student donne pour un nombre n de ddl déterminé et pour une probabilité α la valeur t_α de la variable de Student telle que :

$P(T< t_\alpha) = 1- \alpha$

Exercice : Soit la variable T qui suit une loi de Student à 20ddl :

- Déterminer la valeur de t telle que : P(T<t)=0.7
- Déterminer la valeur de t telle que P(T>t)= 0.05
- Calculer P(T>0.86)
- Déterminer la valeur de t telle que $P(|T|<t) = 0.98$

Corrigé :

Une lecture de la table de student nous conduit aux résultats suivants

- P(T<t)=0.7 → t = 0.533
- P(T>t)= 0.05 → P(T<t)=0.95 → t = 1.725
- P(T>0.86) = 1 - P(T< 0.86) = 1 - 0.8 = 0.2
- $P(|T|<t)$ =0.98 → 2 x P(T<t) – 1 = 0.98 (puisque la loi de student est symétrique) → P(T<t) = 0.99 → t = 2.528

Table de la loi de Khi-deux

$$X \rightarrow \chi^2_n$$

n : le nombre de degrés de liberté

$$P\left[X \succ \chi^2_\alpha\right] = \alpha$$

		α														
		0,995	**0,99**	**0,975**	**0,95**	**0,9**	**0,75**	**0,5**	**0,3**	**0,25**	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**	**0,001**
n	**1**	0,000	0,000	0,001	0,004	0,016	0,102	0,455	1,074	1,320	2,706	3,841	5,020	6,635	7,880	10,827
	2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,408	2,770	4,605	5,991	7,380	9,210	10,600	13,815
	3	0,072	0,115	0,216	0,352	0,584	1,210	2,366	3,665	4,110	6,251	7,815	9,350	11,345	13,000	16,266
	4	0,207	0,297	0,484	0,711	1,064	1,920	3,357	4,878	5,390	7,779	9,488	11,140	13,277	15,000	18,467
	5	0,412	0,554	0,831	1,150	1,610	2,670	4,351	6,064	6,630	9,236	11,070	12,830	15,086	16,860	20,515
	6	0,676	0,872	1,240	1,640	2,204	3,450	5,348	7,231	7,840	10,645	12,592	14,450	16,812	18,650	22,457
	7	0,989	1,240	1,690	2,170	2,833	4,250	6,346	8,383	9,040	12,017	14,067	16,010	18,475	20,370	24,322
	8	1,340	1,640	2,180	2,730	3,490	5,070	7,344	9,524	10,220	13,362	15,507	17,530	20,090	22,030	26,125
	9	1,730	2,090	2,700	3,330	4,168	5,900	8,343	10,656	11,390	14,684	16,919	19,020	21,666	23,660	27,877
	10	2,160	2,560	3,250	3,940	4,865	6,730	9,342	11,781	12,550	15,987	18,307	20,480	23,209	25,250	29,588
	11	2,600	3,050	3,820	4,570	5,578	7,580	10,341	12,899	13,700	17,275	19,675	21,920	24,725	26,820	31,264
	12	3,070	3,570	4,400	5,230	6,304	8,430	11,340	14,011	14,850	18,549	21,026	23,340	26,217	28,350	32,909
	13	3,570	4,110	5,010	5,890	7,042	9,300	12,340	15,119	15,980	19,812	22,362	24,740	27,688	29,870	34,528
	14	4,070	4,660	5,630	6,570	7,790	10,160	13,339	16,222	17,120	21,064	23,685	26,120	29,141	31,370	36,123
	15	4,600	5,230	6,260	7,260	8,547	11,030	14,339	17,322	18,250	22,307	24,996	27,490	30,578	32,850	37,697
	16	5,140	5,810	6,910	7,960	9,312	11,910	15,338	18,418	19,370	23,542	26,296	28,850	32,000	34,310	39,252
	17	5,700	6,410	7,560	8,670	10,085	12,790	16,338	19,511	20,490	24,769	27,587	30,190	33,409	35,760	40,790
	18	6,260	7,010	8,230	9,390	10,865	13,670	17,338	20,601	21,600	25,989	28,869	31,530	34,805	37,190	42,312
	19	6,840	7,630	8,910	10,120	11,651	14,560	18,338	21,689	22,720	27,204	30,144	32,850	36,191	38,620	43,820
	20	7,430	8,260	9,590	10,850	12,443	15,450	19,337	22,775	23,830	28,412	31,410	34,140	37,566	40,030	45,315
	21	8,030	8,900	10,280	11,590	13,240	16,340	20,337	23,858	24,930	29,615	32,671	35,480	38,932	41,430	46,797
	22	8,640	9,540	10,980	12,340	14,041	17,240	21,337	24,939	26,040	30,813	33,924	36,780	40,289	42,830	48,268
	23	9,260	10,200	11,690	13,090	14,848	18,130	22,337	26,018	27,140	32,007	35,172	38,080	41,638	44,210	49,728
	24	9,890	10,860	12,400	13,850	15,659	19,030	23,337	27,096	28,240	33,196	36,415	39,360	42,980	45,590	51,179
	25	10,560	11,520	13,120	14,610	16,473	19,940	24,337	28,172	29,340	34,382	37,652	40,650	44,314	46,960	52,620
	26	11,200	12,200	13,840	15,380	17,292	20,840	25,336	29,246	30,430	35,563	38,885	41,920	45,642	48,320	54,052
	27	11,840	12,880	15,570	16,150	18,114	21,750	26,336	30,319	31,530	36,741	40,113	43,190	46,963	49,670	55,476
	28	12,490	13,560	15,310	16,930	18,939	22,660	27,336	31,391	32,620	37,916	41,337	44,460	48,278	51,020	56,893
	29	13,150	14,260	16,050	17,710	19,768	23,560	28,236	32,461	33,710	39,087	42,557	45,720	49,588	52,360	58,302
	30	13,820	14,950	16,790	18,490	20,599	24,480	29,336	33,530	34,800	40,256	43,773	46,980	50,892	53,700	59,703

Table de la Loi de Student

$$X \rightarrow T(n)$$

n: nombre de degrés de liberté

$$P\left[X \prec t_{\alpha}\right] = 1\text{-}\alpha$$

		1- α							
		0,6	**0,7**	**0,8**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
N	**1**	0,325	0,727	1,376	3,078	6,314	12,710	31,820	63,660
	2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925
	3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841
	4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604
	5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032
	6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707
	7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499
	8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355
	9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250
	10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169
	11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106
	12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055
	13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012
	14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977
	15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947
	16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921
	17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898
	18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878
	19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861
	20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845
	21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831
	22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819
	23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807
	24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797
	25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787
	26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779
	27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771
	28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763
	29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756
	30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750

Chapitre3 :

Les méthodes d'échantillonnage

Pour effectuer une étude statistique (enquête, observation ou expérimentation), on se sert généralement d'un échantillon. Celui-ci doit refléter le plus exactement possible L'image de la population. En fait, choisir un échantillon, c'est mettre ensemble un certain nombre d'individus qui composeront une sorte de modèle réduit de la population à laquelle ils appartiennent. Mais comment choisir ces individus? C'est à cette question que nous répondrons dans ce chapitre, en présentant les méthodes d'échantillonnage les plus courantes.

I- Les méthodes aléatoires:

Commençons d'abord par les méthodes dites aléatoires. Le mot "aléatoire" vient du latin aléa, "jeu de dés". II est employé ici pour signifier que les méthodes abordées ont toutes comme caractéristique d'être basées sur le hasard.

1/ L'échantillonnage aléatoire simple:

L'échantillonnage aléatoire simple consiste à choisir des individus de telle sorte que chaque membre de la population ait une égale chance de figurer dans l'échantillon. Ce choix peut se faire avec ou sans remise: avec remise (on dit aussi non exhaustif), un individu peut être choisi plusieurs fois; sans remise (on dit aussi exhaustif), un individu déjà choisi ne peut l'être de nouveau. L'échantillonnage aléatoire simple est habituellement fait sans remise. En effet, l'échantillon est très souvent de petite taille par rapport à celle de la population, et la possibilité qu'un même individu soit choisi plus d'une fois est alors faible. En outre, l'échantillonnage sans remise est plus facile sur le plan opérationne1.

Pour recourir à l'échantillonnage aléatoire simple, il faut disposer d'une liste à jour de tous les individus d'une population déterminée. On choisit alors L'échantillon à l'aide de nombres aléatoires: un nombre aléatoire est un nombre dont le hasard a déterminé chacun des chiffres qui le composent. Pour y parvenir, on utilise soit une table spécialement conçue à cette fin, soit une calculatrice, soit un programme informatique. L'exemple suivant aidera à comprendre comment procéder.

Exemple :

On veut choisir par échantillonnage aléatoire simple, sans remise, 8 étudiants; parmi un groupe de 60. La sélection de L'échantillon sera faite à l'aide de la table de nombres aléatoires qui se trouve à la page suivante.

Table des nombres aléatoires

92033	**80696**	**28539**	**82033**	**85696**	**33539**	**46017**	**09513**	**27344**	**20174**
48885	**18295**	**35380**	**38885**	**23295**	**40380**	**24443**	**11793**	**12962**	**04574**
04774	**56239**	**41710**	**50226**	**61239**	**46710**	**02387**	**13903**	**16742**	**14060**
60385	**11803**	**81477**	**50385**	**16803**	**86477**	**30193**	**27159**	**16795**	**02951**
56702	**16943**	**35824**	**46702**	**21943**	**40824**	**28351**	**11941**	**15567**	**04236**
32744	**07171**	**46142**	**22744**	**12171**	**51142**	**16372**	**15381**	**07581**	**01793**
63084	**35161**	**89169**	**53084**	**40161**	**94169**	**31542**	**29723**	**17695**	**08790**
96750	**08697**	**97546**	**86750**	**13697**	**12546**	**48375**	**32515**	**28917**	**02174**
99809	**57531**	**13973**	**89809**	**62531**	**18973**	**49905**	**04658**	**29936**	**14383**
48837	**32253**	**97156**	**38837**	**37253**	**10156**	**24419**	**32385**	**12946**	**08063**
46347	**90015**	**62242**	**36347**	**95015**	**67242**	**23174**	**20747**	**12116**	**22504**
65793	**54819**	**89871**	**55793**	**59819**	**94871**	**32897**	**29957**	**18598**	**13705**
40585	**83547**	**35551**	**30585**	**88547**	**40551**	**20293**	**11850**	**10195**	**20887**
48546	**02045**	**19225**	**38546**	**07045**	**24225**	**24273**	**06408**	**12849**	**00511**
62661	**07873**	**00215**	**52661**	**12873**	**05215**	**31331**	**01072**	**17554**	**01968**
35762	**94389**	**86960**	**25762**	**99389**	**91960**	**17881**	**28987**	**08587**	**23597**
35940	**85953**	**54717**	**25940**	**90953**	**59717**	**17970**	**18239**	**08647**	**21488**
96063	**43229**	**00984**	**86063**	**48229**	**05984**	**48032**	**00328**	**28688**	**10807**
29650	**55866**	**60952**	**19650**	**60866**	**65952**	**14825**	**20317**	**06550**	**13967**
85129	**68938**	**62005**	**75129**	**73938**	**67005**	**42565**	**20668**	**25043**	**17235**
50109	**88952**	**74335**	**40109**	**93952**	**79335**	**25055**	**24778**	**13370**	**22238**
37708	**49907**	**91188**	**27708**	**54907**	**96188**	**18854**	**30396**	**09236**	**12477**
72040	**49876**	**02777**	**62040**	**54876**	**07777**	**36020**	**00926**	**20680**	**12469**
06335	**19548**	**02172**	**36650**	**24548**	**07172**	**03168**	**00724**	**12217**	**04887**
11859	**66327**	**22991**	**01859**	**71327**	**27991**	**05930**	**07664**	**00620**	**16582**
54718	**01251**	**55991**	**44718**	**06251**	**60991**	**27359**	**18664**	**14906**	**00313**
38055	**63671**	**12730**	**28055**	**68671**	**17730**	**19028**	**04243**	**09352**	**15918**
79352	**16177**	**95394**	**69352**	**21177**	**10394**	**39676**	**31798**	**23117**	**04044**
94669	**73141**	**26022**	**84669**	**78141**	**31022**	**47335**	**08674**	**28223**	**18285**
20501	**93684**	**87259**	**10501**	**98684**	**92259**	**10251**	**29086**	**03500**	**23421**
41980	**06793**	**31724**	**31980**	**11793**	**36724**	**20990**	**10575**	**10660**	**01698**
37273	**92154**	**85129**	**27273**	**97154**	**90129**	**18637**	**28376**	**09091**	**23039**
58854	**67852**	**50109**	**48854**	**72852**	**55109**	**29427**	**16703**	**16285**	**16963**
79359	**05765**	**37708**	**69359**	**10765**	**42708**	**39680**	**12569**	**23120**	**01441**
57074	**10788**	**72040**	**47074**	**15788**	**77040**	**28537**	**24013**	**15691**	**02697**
47399	**14441**	**55866**	**37399**	**19441**	**60866**	**23700**	**18622**	**12466**	**03610**
81615	**88425**	**68938**	**71615**	**93425**	**73938**	**40808**	**22979**	**23872**	**22106**
11168	**61067**	**88952**	**01168**	**66067**	**93952**	**05584**	**29651**	**00389**	**15267**
87140	**24350**	**72362**	**77140**	**29350**	**77362**	**43570**	**24121**	**25713**	**06088**
80039	**77313**	**17951**	**70039**	**82313**	**22951**	**40020**	**05984**	**23346**	**19328**
12567	**84092**	**27273**	**02567**	**89092**	**32273**	**06284**	**09091**	**00856**	**21023**
51646	**92661**	**84092**	**41646**	**97661**	**89092**	**25823**	**28031**	**13882**	**23165**
72362	**04775**	**92661**	**62362**	**09775**	**97661**	**36181**	**30887**	**20787**	**01194**
17951	**55486**	**37273**	**07951**	**60486**	**42273**	**08976**	**12424**	**02650**	**13872**
47518	**78723**	**72362**	**37518**	**83723**	**77362**	**23759**	**24121**	**12506**	**19681**
70070	**07152**	**17951**	**60070**	**12152**	**22951**	**35035**	**05984**	**20023**	**01788**
73556	**43198**	**07273**	**63556**	**48198**	**12273**	**36778**	**02424**	**21185**	**10800**

- La première étape consiste à numéroter chaque étudiant à L'aide de deux chiffres afin que chacun des étudiants ait la même chance d'être choisi: 01, 02, 03..., 59, 60.
- La deuxième étape consiste à choisir un processus de déplacement dans la table et à déterminer au hasard un premier nombre de deux chiffres (car il y a deux chiffres dans 60) qui sera le point de départ. L'échantillon sera alors constitué des huit premiers nombres de deux chiffres n'excédant pas 60, tout nombre apparaissant une deuxième fois étant éliminé.
Supposons que l'on ait choisi comme point de départ de la table les deux premiers chiffres à l'intersection de la 5ème ligne et de la 3ème colonne. Supposons aussi que L'on ait décidé d'aller de haut en bas jusqu'à la fin de la colonne en ne prenant toujours que les deux premiers chiffres de chaque nombre, puis de reprendre le même processus en haut de la colonne suivante.

L'échantillon se composerait alors des étudiants portant les numéros suivants: 35,46,13,19,54,60,02 et 22.

En résumé, L'échantillonnage aléatoire simple consiste à choisir n unités statistiques parmi N unités d'une population dont on possède la liste.

1. On numérote tous les individus de la liste avec des nombres comportant un même nombre de chiffres.

2. En utilisant une table de nombres aléatoires, une calculatrice ou un programme informatique, on obtient des nombres aléatoires comportant le nombre de chiffres désiré.

3. En rejetant les nombres qui ne font pas partie de la liste ou qui se répètent, on s'arrête après avoir sélectionné n individus.

2/ L'échantillonnage systématique:

L'échantillonnage systématique est une méthode demandant moins de manipulations que l'échantillonnage aléatoire simple. Cependant, elle suppose aussi l'existence d'une liste de la population où chaque individu est numéroté de 1 jusqu'à N. Notons n, le nombre d'individus que doit comporter L'échantillon. L'entier voisin de N/n sera noté r et appelé la raison de sondage (ou le pas de sondage). Choisissons ensuite au hasard un entier d entre 1 et N: cet entier sera le point de départ. Pour former L'échantillon, il s'agira de choisir un premier individu comme point de départ; ce sera l'individu dont le numéro correspond à : d. Pour sélectionner (les autres, il suffit d'ajouter à d la raison de sondage: les individus choisis seront alors les individus dont les numéros correspondent à: d ; d + r; d + 2r; d + 3r; etc.
II faudra reprendre au début lorsque la liste sera épuisée.

Exemple : On veut choisir par échantillonnage systématique 6 étudiants parmi un

groupe de 60. La raison de sondage (r) sera 10, car N/n = 10. (Si L'on devait trouver un échantillon de taille (n) 7, la raison serait plutôt 9 puisque 60/7 = 8,57, arrondi à 9.) Si L'origine choisie au hasard est disons, 3, les étudiants inclus dans L'échantillon correspondront aux numéros suivants:
3, c'est-à-dire le point de départ ; 13, c'est-à-dire 3 + 10 ;
23, c'est-à-dire 3 + (2 x 10) 33, c'est-à-dire 3 + (3 x 10) 43, c'est-à-dire 3 + (4 x 10) 53, c'est-à-dire 3 + (5 x 10)

Notre échantillon est maintenant complet; il sera composé des étudiants portant les numéros suivants: 3, 13,23, 33, 43, 53.
Selon la raison de sondage et le point de départ choisi, il peut arriver que le nombre obtenu se situe à l'extérieur de la liste. En pareil cas, il faut revenir au début de la liste. Si L'on reprenait l'exemple précédent et si L'on devait continuer à prélever des étudiants, on obtiendrait le numéro 63, c'est-à-dire 3 + (6 x 10). Comme ce numéro r se situe à l'extérieur de la liste, on reviendrait au début de la liste et on choisirait le numéro 3 (puisque 63 - 60 = 3).

En résumé, L'échantillonnage systématique consiste à choisir n unités statistiques parmi N unités d'une population dont on possède la liste.
1. On numérote tous les individus de la liste.

2. On calcule la raison de sondage (r), c'est-à-dire l'entier le plus proche de Nin.

3. On choisit au hasard un entier d entre 1 et N.

4. A partir de d, on ajoute autant de fois r que cela est nécessaire pour sélectionner n unités statistiques.

3/ L'échantillonnage stratifié:
Contrairement aux deux méthodes précédentes, dans L'échantillonnage stratifié, on tient compte des renseignements qu'on pourrait déjà posséder sur la population, renseignements obtenus en particulier lors d'un recensement. La méthode repose en effet sur une division de la population en groupes relativement homogènes, appelés strates, puis sur la sélection d'un échantillon dans chaque strate. C'est une méthode qui permettra d'obtenir un échantillon représentatif, c'est-à-dire un échantillon qui possédera les mêmes caractéristiques que la population dont il a été extrait. Illustrons le processus à L'aide d'un exemple.

Exemple :
On veut choisir par échantillonnage stratifié 10 élèves dans un groupe de 60, en tenant compte du fait que 50 % d'entre eux sont en première année, 30 % en 2e année et 20 % en 3e année. Chaque année sera une strate dans laquelle on ira chercher des élèves en tenant compte des pourcentages qu'occupe chaque strate dans la ' population. Ainsi, on

choisira au hasard:
5 élèves en 1ère année, puisque 10 x 50% = 5 ;
3 élèves en 2ème année, puisque 10 x 30% = 3 ;
2 élèves en 3ème année, puisque 10 x 20% = 2.
Il ne reste plus qu'à sélectionner un échantillon dans chaque strate, ce qui pourrait être fait par échantillonnage aléatoire simple ou systématique.
Dans une population peu homogène, le découpage en strates sera d'autant plus avantageux qu'existera une certaine homogénéité à L'intérieur de chaque strate.

4/ L'échantillonnage par grappes:

Dans chacune des méthodes précédentes, les unités de l'échantillon étaient choisies individuellement. L'échantillonnage par grappes consiste plutôt à choisir plusieurs individus en même temps, c'est-à-dire par groupes. Par exemple, prenons comme population les habitants d'une ville à partir desquels on désire constituer un échantillon de 600 individus. Selon les méthodes précédentes, il faudrait choisir 600 individus disséminés dans toute la ville. Suivant l'échantillonnage par grappes, on pourra choisir les 600 individus dans une vingtaine d'immeubles choisis au hasard, où tous les occupants auront été retenus.

Exemple :

Au moyen de L'échantillonnage par grappes, il s'agit de choisir 12 étudiants; dans un groupe de 60. On demande aux étudiants de se regrouper par 6. On choisit ensuite au hasard deux regroupements, par exemple les grappes numéros 4 et 7. En retenant tous les individus de ces deux grappes, on constitue un ; échantillon de 12 étudiants.
Dans l'exemple précédent, la situation était assez simple. Dans les faits, les choses sont: beaucoup plus compliquées. Par exemple, dans une ville, les quartiers et les immeubles ne sont pas composés d'un nombre égal d'individus. Comment peut-on procéder alors pour effectuer un échantillonnage par grappes? Puisque le nombre d'habitants et le nombre d'immeubles et de logements de chacun des quartiers sont généralement connus, il est possible de quadriller la ville en un grand nombre de secteurs (qu'on appelle îlots) ayant des populations à peu près semblables. Pour constituer un échantillon, il suffira alors de choisir certains de ces secteurs et, par conséquent, toutes les personnes qui y habitent.

En résumé, L'échantillonnage par grappes consiste à choisir n unités statistiques parmi N unités d'une population.

1. On subdivise la population en grappes, autant que possible nombreuses et de taille équivalente. Par exemple, si L'on doit répartir 307 individus dans 10 grappes, on pourra faire 7 grappes de 31 personnes et 3 grappes de 30 personnes.
2. On calcule combien il faut de grappes pour constituer L'échantillon.
3. On choisit au hasard les grappes qui serviront à construire L'échantillon.
4. On sélectionne tous les individus des grappes choisies.

II-Les méthodes non aléatoires :

Les méthodes de la section précédente sont dites aléatoires. Le processus de sélection est alors basé sur le hasard, et les individus faisant partie de la population ont tous une égale possibilité de faire partie de L'échantillon. On oppose aux méthodes aléatoires les méthodes dites non aléatoires: ce sont des méthodes où le concept de chances égales est absent. En utilisant des méthodes non aléatoires, il faut toujours craindre un manque de fiabilité. Un des seuls moyens de mesurer l'exactitude de leurs résultats est souvent de les contrôler par d'autres études. Doit-on alors penser que les méthodes non aléatoires sont rarement employées? Pour différentes raisons, on ne peut souvent pas faire autrement que d'y recourir. On s'en sert pour des études exploratoires; pour réduire les coûts; parce que le recours à une méthode aléatoire n'est ni possible ni même envisageable; ou encore lorsque l'homogénéité de la population est telle qu'il serait inutile d'utiliser des méthodes aléatoires.

1/ L'échantillonnage à l'aveuglette:

La personne qui déguste des échantillons de vin pour déterminer lequel est le meilleur ou le journaliste qui, au moyen d'entrevues dans (a rue, sonde L'opinion du grand public sur un sujet d'actualité, mettent en pratique la méthode de L'échantillonnage à l'aveuglette. Le grand avantage de cette méthode est qu'elle demande peu de planification. D'application restreinte, elle peut malgré tout donner de bons résultats si la population observée est homogène. Par exemple, si L'on désire évaluer la concentration d'un produit chimique dans un lac ou le taux de glycémie d'une personne, il est raisonnable de supposer que les résultats devraient être assez semblables d'un échantillon à L'autre.

2/ L'échantillonnage de volontaires:

Dans le cas d'expériences psychologiques ou médicales, d'enquêtes sur les habitudes de consommation, il ne serait pas pratique de choisir au hasard des individus dans toute la population. Comme l'enquête sera longue, exigeante, quelquefois même désagréable, on préfère réunir des volontaires, d'où le nom d'échantillonnage de volontaires. Néanmoins, il faut toujours craindre un écart entre les caractéristiques des volontaires et celles de la population, surtout en matière d'opinion. Les maisons de sondage font souvent appel à un échantillon fixe, appelé panel, formé de volontaires. Il faut savoir que cette façon de procéder peut être une source importante d'erreurs même si on essaie de l'améliorer en se servant d'un échantillon d'assez grande taille. Par exemple, on a souvent noté la tendance à être systématiquement favorables ou neutres sur certaines d'actualité, alors que les opinions de la population sur les mêmes questions étaient beaucoup plus diversifiées.

3/ L'échantillonnage par quotas:

L'échantillonnage par quotas est largement utilisé dans les enquêtes d'opinion et les études de marché, notamment lorsqu'on ne dispose pas de liste exhaustive des

individus de la population. La méthode porte également le nom d'échantillonnage dirigé ou par choix raisonné. On parle aussi d'échantillonnage représentatif. On demande en effet aux enquêteurs de faire un nombre déterminé d'entrevues dans divers groupes établis en fonction du secteur géographique, de l'âge, du sexe ou d'autres caractéristiques. Les échantillons de départ peuvent être obtenus par sélection au hasard, par exemple par téléphone, tout comme pour un échantillon aléatoire. La différence essentielle réside toutefois dans la sélection des individus à l'étape ultérieure: dans la méthode des quotas, chaque enquêteur doit suivre des instructions quant au nombre de personnes à interroger suivant le sexe, l'âge, etc.; il doit écarter certains individus, en choisir d'autres, et ce jusqu'à l'obtention de ses quotas. Ceux-ci ont été calculés à partir de données disponibles, de manière que les sexes, les groupes d'âge et les catégories socio-économiques soient correctement représentés dans l'échantillon. L'échantillon devra avoir à peu près la même composition que la population totale étudiée. On se base sur certains critères basique comme le sexe, l'âge, la catégorie socioprofessionnelle, la région, la ville, le niveau d'études… L'hypothèse est que l'échantillon étant représentatif du point de vue des critères retenus, il y de grandes chances pour qu'il le soit également concernant les caractéristiques de l'enquête.

La méthode des quotas est généralement moins coûteuse et plus facile à réaliser que celle d'un échantillonnage aléatoire.

Chapitre 4 :

Les distributions d'échantillonnage

D'une façon générale, la distribution d'échantillonnage caractérise les fluctuations d'échantillonnage de toute statistique (moyenne, proportion variance) calculée sur tous les échantillons possibles de même taille.

I- Distribution d'échantillonnage d'une moyenne :

Si nous prélevons tous les échantillons possibles de même taille d'une population, la moyenne arithmétique (X) calculée sur chaque échantillon variera d'un échantillon à l'autre. Cependant, ces moyennes auront des valeurs autour d'une valeur centrale qui est celle de la moyenne de la population.

Démonstration

$$\overline{X} = \frac{\sum_{1}^{n} xi}{n}$$

$$E(\overline{X}) = \frac{\sum_{1}^{n} E(xi)}{n} = \frac{\sum_{1}^{n} m}{n} = \frac{n.m}{n} = m$$

$$V(\overline{X}) = \frac{\sum_{1}^{n} V(xi)}{n^2} = \frac{\sum_{1}^{n} \sigma^2}{n^2} = \frac{n.\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Conclusion

Si on prélève un échantillon aléatoire de taille n, d'une population (infinie ou d'une population finis échantillonnage avec remis) dont les éléments possède caractère X qui suit une loi probabilité de moyenne E(X)=m et de variance V(X)= σ^2 alors la moyenne d'échantillon $\overline{X}$ suit une loi de probabilité de moyenne :

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

Et de variance:

Var ($\overline{X}$)= (σ^2 /n)

Remarque : Lorsque l'échantillonnage s'effectue sans remise à partir d'une population finie de taille N, on doit apporter une correction à Var ($\overline{X}$)

Var ($\overline{X}$) = (σ^2 /n) * [(N-n)/(N-1)] = (σ^2 /n) *[1- ((n-1)/(N-1))] .

(n-1/N-1) ≈ le taux de sondage(n /N), → le facteur de correction peut être ignoré si le taux de sondage est inférieur à 10%.

Théorème central limite :

On sait que X suit une variable aléatoire sans pour autant connaître cette loi de probabilité. Pour identifier cette loi il faut recourir au théorème central limite

Théorème central limite (TCL) :

Si on prélève un échantillon aléatoire de taille n, d'une population (infinie ou d'une population finie et échantillonnage avec remise) dont les éléments possèdent un caractère X qui suit une loi de probabilité de moyenne E(X)=m et de variance V(X)= σ^2 alors la distribution d'échantillonnage de la variable aléatoire $\overline{X}$ tend à se rapprocher d'une loi normale de moyenne E($\overline{X}$)=m et de variance V($\overline{X}$)= σ^2 /n

Ainsi $\overline{X} \longrightarrow N(m, \frac{\sigma}{\sqrt{n}})$

On remarque que l'écart type de $\overline{X}$ est en fonction de l'écart type de la population σ . Un problème peut ainsi se poser lorsqu'on ne connaît pas cet écart type de la population. Dans ce cas il faut estimer cet écart type par celui de l'échantillon (S) :

$$S = \sqrt{\frac{\sum_{1}^{n}(Xi - \overline{X})^2}{n-1}}$$

Cependant, dans ce cas, la loi de probabilité peut changer selon la taille de l'échantillon. Trois cas sont alors possibles :

1er cas :

σ connu alors la loi de probabilité sera :

$\overline{X} \longrightarrow N(m, \frac{\sigma}{\sqrt{n}})$ d'où $\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \longrightarrow$ N(0,1)

2ème cas :

σ inconnu et n≥30 alors la loi de probabilité sera :

$\overline{X} \longrightarrow N(m, \frac{S}{\sqrt{n}})$ d'où $\frac{\overline{X} - m}{\frac{S}{\sqrt{n}}} \longrightarrow$ N(0,1)

3ème cas :

σ inconnu et n<30 alors la loi de probabilité sera :

$$\frac{\overline{X} - m}{\frac{S}{\sqrt{n}}} \longrightarrow T(n-1)$$

II-La distribution d'échantillonnage de la différence de deux moyennes :

La distribution d'échantillonnage de la différence de deux moyennes d'un caractère X mesuré sur deux échantillons extraits de deux populations différentes en supposant que les écarts type de ce caractère au niveau de chaque population sont connus.

Si on désigne par m_1 et m_2 les moyennes de cette variable X sur deux populations 1 et 2, et par σ_1 et σ_2 ses écarts types, et si on tire un échantillon d'une taille n_1 de la première population et un échantillon d'une taille n_2 de la deuxième population et on calcule les moyennes empiriques de ces deux échantillon ($\overline{X}_1$ et $\overline{X}_2$), on aura alors :

$$\overline{X}_1 \longrightarrow N(m_1, \frac{\sigma_1}{\sqrt{n_1}})$$

$$\overline{X}_2 \longrightarrow N(m_2, \frac{\sigma_2}{\sqrt{n_2}})$$

Puisque la différence de deux lois normales est une loi normale, la statistique $X = \overline{X}_1 - \overline{X}_2$ est une loi normale de moyenne $m = m_1 - m_2$ et d'écart type

$\sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ d'où :

$$\frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0,1)$$

III- Distribution d'échantillonnage d'une proportion :

On considère une population caractérisée par une proportion p d'individus possédant un certain caractérise qualitatif. Si on prélève au hasard différents échantillons de cette population on remarque que la proportion f d'individus possédant ce caractère qualitatif varie d'un échantillon à l'autre, on peut donc lui attribuer une variable aléatoire.

On démontre que pour une taille d'échantillon $n \geq 30$:

$$f \to N(p, \sqrt{\frac{pq}{n}}) \quad d'où \quad \frac{f-p}{\sqrt{\frac{pq}{n}}} \to N(0,1)$$

IV-La distribution d'échantillonnage de la différence de deux proportions :

La différence de deux proportions ($f_1 - f_2$) relatives à un même caractère mesuré sur deux échantillons n_1 et n_2 ayant chacun une taille supérieure à 30 et extraits de deux populations différentes suit une loi normale d'une moyenne égale à la différence des deux proportions réelles de ce caractère dans les deux population $p_1 - p_2$ et d'écart type $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ d'où la statistique :

$$\frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \to N(0,1)$$

Chapitre 5 :

L'estimation statistique

Un aspect important de l'inférence statistique est celui d'obtenir à partir des résultats trouvés sur l'échantillon d'une population des estimations fiables de certain paramètre de cette population tels que la moyenne, la variance et la proportion.

Ces estimations peuvent s'exprimer soit par une seule valeur, on parlera alors d'estimation ponctuelle, soit par un intervalle, on parlera dans ce cas d'estimation par intervalle de confiance.

I- L'estimation ponctuelle :

Cette technique consiste à estimer un paramètre θ de la population à l'aide d'un seul nombre déduit des résultats de l'échantillon, ce nombre est appelé estimateur ponctuel de θ et sera noté $\hat{\theta}$.

Un estimateur dois posséder certaines qualités ou propriétés afin de fournire une bonne estimation. Cet estimateur doit être sans biais efficace et convergent.

-Estimateur sans biais :E($\hat{\theta}$)= θ

-Estimateur efficace : si sa variance est la plus faible parmi les variances des autres estimateurs sans biais

- estimateur convergent : $\lim_{n\to+\infty} V(\hat{\theta}) = 0$

Exercice :

- Montrer que $\overline{X}$ est un estimateur sans biais et convergent de m
- Montrer que f est un estimateur sans biais et convergent de p

$$E(\overline{X}) = \frac{\sum_1^n E(xi)}{n} = \frac{\sum_1^n m}{n} = \frac{n.m}{n} = m$$

$$V(\overline{X}) = \frac{\sum_1^n V(xi)}{n^2} = \frac{\sum_1^n \sigma^2}{n^2} = \frac{n.\sigma^2}{n^2} = \frac{\sigma^2}{n} \rightarrow \lim_{n\to+\infty} V(\overline{X}) = 0$$

Puisque $f \to N(p, \sqrt{\frac{pq}{n}})$, E (f) = p

Et V(f) = $\frac{pq}{n} \underset{n \to +\infty}{\to}$ limite V (f)= 0

Remarque : les estimations ponctuelles ne fournissent aucune information concernant la précision des estimations, c'est à dire qu'elles ne tiennent pas fluctuations d'échantillonnage.

II- L'estimation par intervalle de confiance :

L'estimation par intervalle d'un paramètre inconnu θ consiste à calculer à partir d'un estimateur choisi $\hat{\theta}$, un intervalle dans lequel on a un pourcentage de chance d'y trouver correspondante du paramètre θ .

L'intervalle de confiance est défini par deux limites auxquelles est associée une certaine probabilité de contenir la vraie valeur du paramètre θ .

$$P(LI \le \theta \le LS) = 1 - \alpha$$

LI : Limite inférieure de l'intervalle de confiance

LS : Limite supérieure de l'intervalle de confiance

1- Estimation d'une moyenne :

Trois cas peuvent se présenter selon la loi de probabilité de X :

1er cas : σ connu alors la loi de probabilité de X sera :

$$\overline{X} \to N(m, \frac{\sigma}{\sqrt{n}}) d'où \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$$

$$P\left(\left| \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \right| \prec t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$P\left(-t_{\frac{\alpha}{2}} \prec \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \prec t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

On détermine alors la valeur de $t_{\alpha/2}$ à partir de la table de la loi normale centrée réduite. Par la suite on pourra dégager un intervalle de confiance pour m.

$$\overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \prec m \prec \overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \; ; \; \overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

Application :

Une entreprise de conserves désire connaître le poids moyen des boites qu'elle fabrique. Des tests effectués il y 2 ans permettent de considérer que le poids d'une boite est distribué normalement avec une variance de 9 grammes.

Un test sur un échantillon de 16 boites a donné un poids moyen de 219 gr.

Estimer par un intervalle de confiance le poids moyen des boites avec un niveau de confiance de 95%.

Soit X le poids d'une boite de conserve. X→N (m, σ)

σ est connu, ($\sigma = \sqrt{9}$ = 3) d'où $\overline{X}$ → N (m, $\sigma / \sqrt{n}$)

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

$$P\left(\left| \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \right| \prec t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \; ; \; \overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

$1 - \alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \alpha_{/2} = 0.025 \rightarrow F(t_{\alpha/2}) = 0.975 \rightarrow t_{\alpha/2} = 1.96$

La limite inférieure de l'intervalle de confiance est donc :

$LI = \overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 219 - 1.96 \times 3 / \sqrt{16} = 217.53$

Et la limite supérieure de l'intervalle de confiance est donc :

$LS = \overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 219 + 1.96 \times 3 / \sqrt{16} = 220.47$

Remarque : L'expression des bornes de l'intervalle de confiance pourrait être modifiée si on se trouve dans l'obligation d'apporter un ajustement à la variance dans le cas où le tirage de l'échantillon se fait sans remise avec un taux de sondage supérieur à10%. Dans ce cas on a :

LI= $\overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$ et LS = $\overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

2ème cas : σ inconnu et $n \geq 30$ alors la loi de probabilité de X sera :

$$\overline{X} \rightarrow N(m, \frac{S}{\sqrt{n}})$$

$$P\left(\left| \frac{\overline{X} - m}{\frac{S}{\sqrt{n}}} \right| \prec t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$P\left(-t_{\frac{\alpha}{2}} \prec \left| \frac{\overline{X} - m}{\frac{S}{\sqrt{n}}} \right| \prec t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

On détermine alors la valeur de $t_{\alpha/2}$ à partir de la table de la loi normale centrée réduite. Par la suite on pourra dégager un intervalle de confiance pour m.

$$\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \prec m \prec \overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

D'ou

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \quad ; \quad \overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right]$$

Application :

Une entreprise de conserves désire connaître le poids moyen des boites qu'elle fabrique. Un test sur un échantillon de 36 boites a donné un poids moyen de 219 grammes avec un écart type de 1.5 grammes Estimer par intervalle de confiance le poids moyen des boites avec un niveau de confiance de 99%.

Soit X le poids d'une boite de conserve. X→N (m, σ)

σ est inconnu, et n = 36 >30 d'où $\overline{X}$ → N (m, $S / \sqrt{n}$)

$$\frac{\overline{X} - m}{\frac{S}{\sqrt{n}}} \rightarrow N(0,1)$$

$$P\left(\left|\frac{\overline{X}-m}{\frac{S}{\sqrt{n}}}\right| \prec t_{\frac{\alpha}{2}}\right)=1-\alpha$$

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \;;\; \overline{X} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

$1-\alpha = 0.99 \rightarrow \alpha = 0.01 \rightarrow \alpha_{/2} = 0.005 \rightarrow F(t_{\alpha/2}) = 0.995 \rightarrow t_{\alpha/2} = 2.58$

La limite inférieure de l'intervalle de confiance est donc :

LI = $\overline{X} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$ = 219 – 2.58 x 1.5 / $\sqrt{36}$ = 218.355

Et la limite supérieure de l'intervalle de confiance est donc :

LS = $\overline{X} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$ = 219 + 2.58 x 1.5 / $\sqrt{36}$ = 219.645

3ème cas : σ inconnu et n<30 alors la loi de probabilité de X sera :

$$\frac{\overline{X}-m}{\frac{S}{\sqrt{n}}} \rightarrow T(n-1)$$

$$P\left(\left|\frac{\overline{X}-m}{\frac{S}{\sqrt{n}}}\right| \prec t_{\frac{\alpha}{2}}\right)=1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2}} \prec \frac{\overline{X}-m}{\frac{S}{\sqrt{n}}} \prec t_{\frac{\alpha}{2}}\right)=1-\alpha$$

On détermine alors la valeur de $t_{\alpha/2}$ à partir de la table de la loi de Student. Par la suite on pourra dégager un intervalle de confiance pour m.

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \;;\; \overline{X} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

Application :

Une entreprise de conserves désire connaître le poids moyen des boites qu'elle fabrique. Un test sur un échantillon de 9 boites a donné un poids moyen de 219 grammes avec un écart type de 1.5 grammes

Estimer par intervalle de confiance le poids moyen des boites avec un niveau de confiance de 99%

Soit X le poids d'une boite de conserve. X→N (m, σ)

σ est inconnu, et n = 9 < 30 d'où $\frac{\overline{X} - m}{\frac{S}{\sqrt{n}}}$ → T(n-1)

S = 1.5

$$P\left(\left|\frac{\overline{X} - m}{\frac{s}{\sqrt{n}}}\right| \prec t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$m \in \left[\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \; ; \; \overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$$

1 – α = 0.99 → α = 0.01 → $\alpha_{/2}$ = 0.005 et ddl = 8 → $t_{\alpha/2}$ = 3.355

La limite inférieure de l'intervalle de confiance est donc :

LI = $\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ = 219 – 3.355 x 1.5 / $\sqrt{9}$ = 217.322

Et la limite supérieure de l'intervalle de confiance est donc :

LS = $\overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ = 219 + 3.355 x 1.5 / $\sqrt{9}$ = 220.677

2-Estimation de la différence de deux moyennes :

La différence de deux moyennes d'un même caractère X mesuré sur deux populations différentes ayant un écart type connu au niveau de chaque population peut être estimée comme suit :

Puisque

$$\frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0,1)$$

$$P\left(\left|\frac{(\overline{X}_1-\overline{X}_2)-(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right| \prec t_{\frac{\alpha}{2}}\right)=1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2}} \prec \left|\frac{(\overline{X}_1-\overline{X}_2)-(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right| \prec t_{\frac{\alpha}{2}}\right)=1-\alpha$$

On détermine alors la valeur de $t_{\alpha/2}$ à partir de la table de la loi normale centrée réduite. Par la suite on pourra dégager un intervalle de confiance pour $m_1 - m_2$.

$$\overline{X}_1-\overline{X}_2-t_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \prec m_1-m_2 \prec \overline{X}_1-\overline{X}_2+t_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

D'ou

$$m_1-m_2 \in \left[\overline{X}_1-\overline{X}_2-t_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \quad ; \quad \overline{X}_1-\overline{X}_2+t_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

3- Estimation d'une proportion :

Pour déterminer un intervalle de confiance pour une proportion on utilise la distribution suivante :

$$f \rightarrow N\left(p,\sqrt{\frac{pq}{n}}\right) \qquad d'ou \qquad \frac{f-p}{\sqrt{\frac{pq}{n}}} \rightarrow N(0,1) \qquad avec\, n \geq 30$$

$$P\left(\left|\frac{f-p}{\sqrt{\frac{pq}{n}}}\right| \leq t_{\frac{\alpha}{2}}\right)=1-\alpha$$

On détermine la valeur de $t_{\alpha/2}$ à partir de la table de la loi normale

$$-t_{\frac{\alpha}{2}} \leq \frac{f-p}{\sqrt{\frac{pq}{n}}} \leq t_{\frac{\alpha}{2}}$$

$$f-t_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} \leq p \leq f+t_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}}$$

On remarque que les bornes de l'intervalle de confiance dépendent de p, on va alors estimer p par f uniquement au niveau des deux bornes de l'intervalle.

$$f - t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} \leq p \leq f + t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$$

Exercice :

Dans une banque, on a effectué un sondage pour connaître l'opinion des clients sur un nouveau service aux agences. D'une liste de 6000 clients de la banque on extrait 150, sur ces 150 clients interrogés 45 étaient satisfaits de ce service. Déterminer un intervalle de confiance pour la vraie proportion des clients qui sont satisfaits de ce nouveau service avec un niveau de confiance de 99%.

Corrigé :

Soit p la vraie proportion des clients satisfaits et f la proportion empirique de ces clients observé sur l'échantillon

$$f \to N(p, \sqrt{\frac{pq}{n}}) \quad d'où \quad \frac{f-p}{\sqrt{\frac{pq}{n}}} \to N(0,1)$$

f = cas favorables/taille de l'échantillon= k/n = 45/150 = 0.3 → q= 1-p = 0.7

$$P\left(\left|\frac{f-p}{\sqrt{\frac{pq}{n}}}\right| \leq t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$f - t_{\frac{\alpha}{2}}\sqrt{\frac{f'(1-f)}{n}} \leq p \leq f + t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$$

$1 - \alpha = 0.99 \rightarrow \alpha = 0.01 \rightarrow \alpha/2 = 0.005 \rightarrow F(t_{\alpha/2}) = 0.995 \rightarrow t_{\alpha/2} = 2.58$

La limite inférieure de l'intervalle de confiance est donc :

$$LI = f - t_{\frac{\alpha}{2}}\sqrt{\frac{f'(1-f)}{n}} = 0.3 - 2.58 \times \sqrt{\frac{0.3x0.7}{150}} = 0.2035$$

Et la limite supérieure de l'intervalle de confiance est donc :

$$LS = f + t_{\frac{\alpha}{2}}\sqrt{\frac{f'(1-f)}{n}} = 0.3 + 2.58 \times \sqrt{\frac{0.3x0.7}{150}} = 0.396$$

3-Estimation de la différence de deux proportions :

La différence de deux proportions ($f_1 - f_2$) relatives à un même caractère mesuré sur deux échantillons n_1 et n_2 ayant chacun une taille supérieure à 30 et extraits de deux populations différentes peut être estimée comme suit :

Puisque

$$\frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \to N(0,1)$$

$$P\left(\left| \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \right| \le t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

On détermine la valeur de $t_{\alpha/2}$ à partir de la table de la loi normale

$$-t_{\frac{\alpha}{2}} \le \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \le t_{\frac{\alpha}{2}}$$

$$f - t_{\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \le p_1 - p_2 \le f + t_{\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

On remarque que les bornes de l'intervalle de confiance dépendent de p_1 et p_2, on va alors estimer p_1 par f_1 et p_2 par f_2 uniquement au niveau des deux bornes de l'intervalle.

$$(f_1 - f_2) - t_{\frac{\alpha}{2}} \sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}} \le (p_1 - p_2) \le (f_1 - f_2) + t_{\frac{\alpha}{2}} \sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}$$

III- La détermination de la taille de l'échantillon :

La taille de l'échantillon est liée à la marge d'erreur E qu'on va tolérer c'est à dire la différence en valeur absolue entre le paramètre à estimer et son estimateur.

$$E = \left| \theta - \hat{\theta} \right|$$

1- Cas d'une moyenne :

On sait que : $\overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \prec m \prec \overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Donc : $\left| m - \overline{X} \right| \leq t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Ainsi pour un niveau de confiance 1-α la marge d'erreur E est au plus égale à :

$$t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Cette valeur quantifie l'erreur attribuable aux fluctuations d'échantillonnage. Ainsi on peut fixer la marge d'erreur E qu'on ne veut pas excéder et déterminer la taille minimale requise de l'échantillon.

$$E = t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{t_{\frac{\alpha}{2}} \sigma}{E} \right)^2$$

Ainsi plus la marge d'erreur est faible plus la taille de l'échantillon est élevée.

Exemple :

On veut estimer la durée de vie moyenne d'un dispositif électronique. D'après le bureau de RD l'écart type de la durée de vie de ce dispositif serait 100 heures. Déterminer le nombre d'essais requis pour estimer avec un niveau de confiance de 95%, la durée de vie moyenne d'une grande production de ce dispositif de sorte que la marge d'erreur dans l'estimation n'excède pas 50 heures.

E= 50, σ = 100, 1-α = 0.95 → $t_{\alpha/2}$ = 1.96 → n = (1.96 x 100/50)² = 15.366≈16

Remarque :

Dans le cas particulier d'un tirage sans remise avec un taux de sondage n/N > 10 % , on a :

$$\overline{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \prec m \prec \overline{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

d'où une marge d'erreur :

$$E = t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \Rightarrow E^2 = t_{\frac{\alpha}{2}}^2 \frac{\sigma^2}{n} \frac{N-n}{N-1} \Rightarrow E^2 \frac{N-1}{t_{\frac{\alpha}{2}}^2 \sigma^2} = \frac{N}{n} - 1$$

$$\Rightarrow \frac{N}{n} = \frac{E^2(N-1) + t_{\frac{\alpha}{2}}^2 \sigma^2}{t_{\frac{\alpha}{2}}^2 \sigma^2} \Rightarrow \frac{n}{N} = \frac{t_{\frac{\alpha}{2}}^2 \sigma^2}{E^2(N-1) + t_{\frac{\alpha}{2}}^2 \sigma^2}$$

$$\Rightarrow n = \left(\frac{N t_{\frac{\alpha}{2}}^2 \sigma^2}{E^2(N-1) + t_{\frac{\alpha}{2}}^2 \sigma^2} \right)$$

Application :

On souhaite estime la consommation par ménage d'une population de 1000 ménages. Quelle taille d'échantillon faut-il retenir pour que la consommation moyenne d'un échantillon tiré **avec remise** ne s'éloigne pas de la moyenne de la population de plus de 2 DT avec une probabilité de 95% sachant que l'écart-type de la consommation σ est de 12DT ?

Corrigé :

N= 1000, E= 2, σ = 12, 1-α = 0.95 → $t_{\alpha/2}$ = 1.96

$$\rightarrow n = \left(\frac{N t_{\frac{\alpha}{2}}^2 \sigma^2}{E^2(N-1) + t_{\frac{\alpha}{2}}^2 \sigma^2} \right) = \frac{1000 x 1.96^2 x 12^2}{2^2 x 999 + 1.96^2 x 12^2} = 553190.4/4549.19 = 121.6 \approx 122$$

2- Cas d'une proportion :

Dans le cas de l'estimation d'une proportion on sait que :

$$|f - p| \leq t_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

$$E = t_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

La taille de l'échantillon serait donc :

$$n = \frac{t_{\alpha/2}^2 \; p\,(1-p)}{E^2}$$

Si on a une information sur la valeur approximative de p on va la remplacer dans la formule, sinon on remplace p par 0.5.

Application :

On veut effectuer un sondage auprès des 8000 étudiants d'une faculté pour estimer le pourcentage des fumeurs. Déterminer la taille de l'échantillon requise pour estimer avec un niveau de confiance de 95% cette proportion avec une marge d'erreur n'excédant pas 5% dans les deux cas suivants :

1- Une enquête similaire effectuée il y a 3 ans indiqua que 32% des étudiants fumaient

2- En supposant qu'on n'a aucune information préalable sur p.

Corrigé :

1- P= 0.32 → 1-p = 0.68, E =0.05

$1-\alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \alpha_{/2} = 0.025 \rightarrow F(t_{\alpha/2}) = 0.975 \rightarrow t_{\alpha/2} = 1.96$

$$\rightarrow n = \frac{t_{\alpha/2}^2 \; p\,(1-p)}{E^2} = 1.96^2 \text{x } 0.32\text{x}0.68 / 0.05^2 = 334.37 \approx 335$$

2- Si p est inconnue, on prend p =0.5 d'où $n = 1.96^2\text{x}0.5\text{x}0.5 / 0.05^2 = 384.16 \approx 385$

Chapitre 6 :

Les tests paramétriques

Pour effectuer un test statistique on doit d'abord émettre des hypothèses concernant un paramètre de la population (moyenne, proportion …) afin de vérifier par la suite la véracité de ces hypothèses.

I- La procédure du test :

Supposons qu'on veut effectuer le test suivant

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

La décision va s'effectuer sur la base des résultats d'un échantillon. Deux types de décisions erronées que l'on appellera erreurs peuvent être prises :

- On appelle erreur de première espèce le fait de rejeter à tort l'hypothèse H_0.

La probabilité correspondante à cette erreur est appelée risque de première espèce :α

$\alpha = P(\text{rejeter } H_0 / H_0 \text{ vraie}) = P(\text{décider } H_1 / \theta = \theta_0)$

- On appelle erreur de seconde espèce le fait de rejeter à tort l'hypothèse H_1. La probabilité correspondante à cette erreur est appelée risque de seconde espèce :β

$\beta = P(\text{rejeter } H_1 / H_1 \text{ vraie}) = P(\text{décider } H_0 / \theta \neq \theta_0)$

		Décision	
Réaliste		H_0 retenue	H_1 retenue
	H_0 vraie	Bonne décision (1- α)	Erreur de première espèce (α)
	H_1 vraie	Erreur de seconde espèce(β)	Bonne décision (1- β)

- On appelle puissance du test : $\eta = 1 - \beta$
- On appelle région critique d'un test, la région de rejet de l'hypothèse H_0

La démarche à suivre pour effectuer un test est la suivante :

- Enoncer les hypothèses H_0 et H_1 (H_0 est l'hypothèse à laquelle on croît le plus)
- Préciser la loi de la variable dans la population
- Préciser la statistique utilisée et sa distribution au niveau de l'échantillon
- Déterminer en fonction du risque de première espèce α la frontière de la région critique
- Si la valeur observée à partir de l'échantillon appartient à la région critique, on rejette H_0 sinon on accepte cette hypothèse.

II- Test sur une moyenne :

Pour la détermination de la région critique, on va se placer dans le cas où la variable étudiée est distribué au niveau de la population suivant une loi normale de moyenne m inconnue qu'on va tester et un écart type σ connu.

1-Test unilatéral à droite :

Les hypothèses du test se présentent sous la forme :

$\begin{cases} H_0 : m = m_0 \\ H_1 : m > m_0 \end{cases}$

La variable X étudiée au niveau de la population suit une loi normale de paramètres m et σ,avec σ connu. Ainsi la distribution de X au niveau de l'échantillon sera :

$$\overline{X} \rightarrow N(m, \frac{\sigma}{\sqrt{n}}) \quad d'où \quad \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0,1)$$

On commence par déterminer la frontière de la région critique (c)

$\alpha = P$ (rejeter H_0/H_0 vraie)

$\alpha = P$ ($m>m_0$ /$m=m_0$)

La décision se fait compte tenu de la valeur de $\overline{X}$ calculée au niveau de l'échantillon

P($\overline{X}$ >c / $m=m_0$)= α

$$P(\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}} \succ \frac{c-m}{\frac{\sigma}{\sqrt{n}}} / m=m_0)=\alpha$$

$$P(X^* \succ \frac{c-m_0}{\frac{\sigma}{\sqrt{n}}})=\alpha$$

$$P(X^* \succ \frac{c-m_0}{\frac{\sigma}{\sqrt{n}}})=1-\alpha$$

$$D'ou \rightarrow t_\alpha = \frac{c - m_0}{\frac{\sigma}{\sqrt{n}}}$$

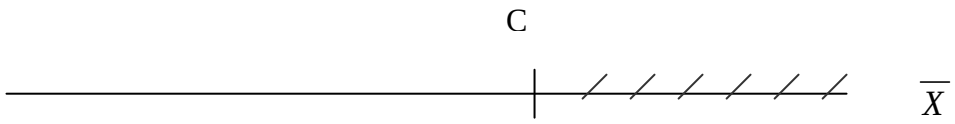
On détermine la valeur de t_α à partir de la table de la loi normale. Ainsi la frontière de la région critique aura pour expression :

$$c=m_0+t_\alpha\frac{\sigma}{\sqrt{n}}$$

Règle de décision :

- Rejet de H_0 si la valeur de $\overline{X}$ >c
- Acceptation de H_0 si la valeur de $\overline{X} \leq c$

Axe :



Acceptation H_0 : $m = m_0$

Rejet Ho : $m > m_0$

Application :

Une machine est réglée afin de verser 500 grammes de poudre de chocolat dans les sachets. Les statistiques du service de production montrent que le poids versé dans ces sachets suit une loi normale d'écart type 9 grammes. Un échantillon de 9 sachets prélevés au hasard a donné un poids net moyen de 504 grammes. Tester au seuil de 1% si la machine est en train de verser plus de 500 grammes dans les sachets.

Soit X le poids versé dans les sachets. $X \rightarrow N(m, \sigma)$,

σ =9, et n = 9 d'où $\overline{X} \rightarrow N(m, \sigma / \sqrt{n})$

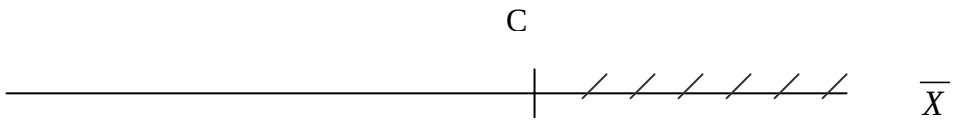
Le test d'hypothèse consiste à vérifier si la machine est en train de verser 500 grammes ou plus. ($m_0 = 500$).

Il s'agit donc d'un test à droite qui s'exprime ainsi :

$$\begin{cases} H_0 : m = 500 \\ H_1 : m > 500 \end{cases}$$

La règle de décision est donc rejet de H_0 si $\overline{X} > c$ avec $c = m_0 + t_\alpha \frac{\sigma}{\sqrt{n}}$

Axe :



Acceptation H_0 : $m = m_0$

Rejet Ho : $m > m_0$

$\alpha = 0.01 \rightarrow F(t_\alpha) = 0.99 \rightarrow t_\alpha = 2.33$ d'où $c = 500 + 2.33 \times (9/\sqrt{9}) = 506.99$

$\overline{X} = 504 < c \rightarrow$ on retient l'hypothèse H_0

2- Test unilatéral à gauche :

Les hypothèses du test se présentent sous la forme :

$$\begin{cases} H_0 : m = m_0 \\ H_1 : m < m_0 \end{cases}$$

La variable X étudiée au niveau de la population suit une loi normale de paramètres N(m,σ) avec σ connu. Ainsi la distribution de $\overline{X}$ au niveau de l'échantillon sera :

$$\overline{X} \to N(m, \frac{\sigma}{\sqrt{n}}) d'ou \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$$

Détermination de la frontière de la région critique (c)
α =P (rejeter H_0/H_0 varie)
α =P ($m<m_0/m=m_0$)
La décision se fait compte tenu de la valeur de $\overline{X}$ calculée au niveau de l'échantillon.
P($\overline{X}$ <c/ $m=m_0$)= α

$$P(\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} < \frac{c - m}{\frac{\sigma}{\sqrt{n}}} / m = m_0) = \alpha$$

$$P(X^* < \frac{c - m_0}{\frac{\sigma}{\sqrt{n}}}) = \alpha$$

Or α<0,5 d'où $\frac{c - m_0}{\frac{\sigma}{\sqrt{n}}} < 0$ on aura ainsi

$$P(X^* < \frac{m_0 - c}{\frac{\sigma}{\sqrt{n}}}) = 1 - \alpha$$

$$D'ou \to t_\alpha = \frac{m_0 - c}{\frac{\sigma}{\sqrt{n}}}$$

On détermine la valeur de t_α à partir de la table de la loi normale. Ainsi la frontière de la région critique aura pour expression :

$$c = m_0 - t_\alpha \frac{\sigma}{\sqrt{n}}$$

Règle de décision

- Rejet de H_0 si la valeur de $\overline{X}$ <c
- Acceptation de H_0 si la valeur de $\overline{X} \geq c$

Axe :

Application :
Une machine est réglée afin de verser 500 grammes de poudre de chocolat dans les sachets. Les statistiques du service de production montrent que le poids versé dans ces

sachets suit une loi normale d'écart type 9 grammes. Un échantillon de 9 sachets prélevés au hasard a donné un poids net moyen de 495 grammes. Tester au seuil de 5% si la machine est en train de verser moins que 500 grammes dans les sachets.
Soit X le poids versé dans les sachets. $X \rightarrow N(m, \sigma)$,
σ =9, et n = 9 d'où $\overline{X} \rightarrow N(m, \sigma / \sqrt{n})$

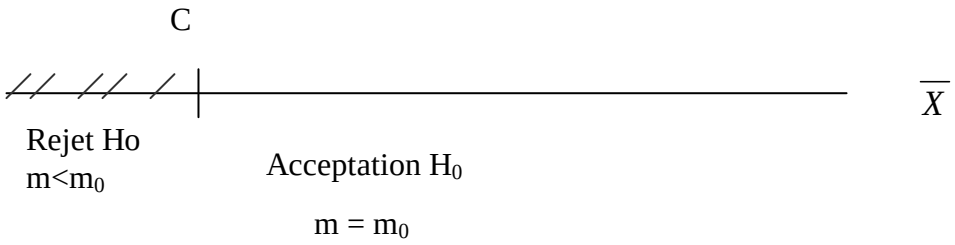
Le test d'hypothèse consiste à vérifier si la machine est en train de verser 500 grammes ou moins. ($m_0 = 500$).

Il s'agit donc d'un test à droite qui s'exprime ainsi :

$$\begin{cases} H_0 : m = 500 \\ H_1 : m < 500 \end{cases}$$

La règle de décision est donc rejet de H_0 si $\overline{X} < c$ avec $c = m_0 - t_\alpha \frac{\sigma}{\sqrt{n}}$

Axe :



$\alpha = 0.05 \rightarrow F(t_\alpha) = 0.95 \rightarrow t_\alpha = 1.65$ d'où $c = 500 - 1.65 \times (9/\sqrt{9}) = 495.05$

$\overline{X} = 495 < c \rightarrow$ on retient l'hypothèse H_1

3- Test bilatéral :

Les hypothèses du test se présentent sous la forme :

$$\begin{cases} H_0 : m = m_0 \\ H_1 : m \neq m_0 \end{cases}$$

La variable X étudiée au niveau de la population suit une loi normale de paramètre m et σ, avec σ connu .Ainsi la distribution de $\overline{X}$ au niveau de l'échantillon sera :

$$\overline{X} \rightarrow N(m, \frac{\sigma}{\sqrt{n}}) d'ou \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0,1)$$

Détermination de la région critique :
$\alpha = P$ (rejeter H_0 / H_0 vraie)
$\alpha = P(m < m_0 / m = m_0) + P(m > m_0 / m = m_0)$
$P(m < m_0 / m = m_0) = \alpha/2$
$P(m > m_0 / m = m_0) = \alpha/2$
La décision se fait compte tenu de la valeur de $\overline{X}$ calculée au niveau de l'échantillon.
$P(\overline{X} < c_1 / m = m_0) = \alpha/2$

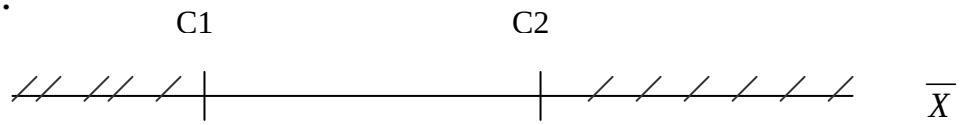
$P(\overline{X} > c_2 / m=m_0) = \alpha/2$
D'ou :

$c_1 = m_0 - t_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$c_2 = m_0 + t_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Décision : rejet de H_0 si $\overline{X} < c_1$ ou $\overline{X} > c_2$

Axe :



Rejet Ho $m<m_0$ | Acceptation H_0 | Rejet Ho $m>m_0$

Exem ~~$m = m_0$~~

Une machine est réglée afin de verser du lait dans ...s de 500 ml le volume de lait versé suit une loi normal d'écart type 5 ml .Un échantillon de 10 bouteilles a donné un volume moyen de 496, 7 ml.
Peut-on conclure au seuil de 5% que la machine est déréglée?
Soit X le poids versé dans les sachets. $X \rightarrow N(m, \sigma)$,
σ =5, et n = 10 d'où $\overline{X} \rightarrow N(m, \sigma / \sqrt{n})$

Le test d'hypothèse consiste à vérifier si la machine est en train de verser 500 grammes ou pas ($m_0 = 500$). Il s'agit donc d'un bilatéral qui s'exprime ainsi :

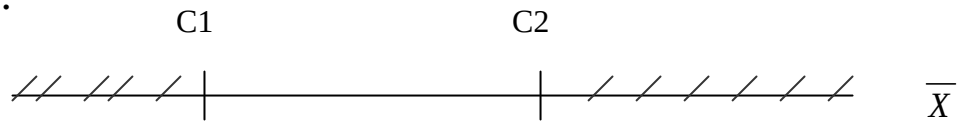
$$\begin{cases} H_0 : m=500 \\ H_1 : m \neq 500 \end{cases}$$

La règle de décision est donc rejet de H_0 si $\overline{X} < c_1$ ou $\overline{X} > c_2$ avec :

$c_1 = m_0 - t_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$c_2 = m_0 + t_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Axe :



Rejet Ho $m<m_0$ | Acceptation H_0 $m = m_0$ | Rejet Ho $m>m_0$

$1 - \alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \alpha_{/2} = 0.025 \rightarrow F(t_{\alpha/2}) = 0.975 \rightarrow t_{\alpha/2} = 1.96$ d'où :

$c_1 = m_0 - t_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 500 - 1.96 \times \frac{5}{\sqrt{10}} = 496.901$

$c_1 = m_0 + t_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 500 + 1.96 \times \frac{5}{\sqrt{10}} = 503.099$

$\overline{X} = 496.7 < c_1 \rightarrow$ on retient l'hypothèse H_1

Remarque : Les expressions des frontières des régions critiques doivent être modifiées lorsqu'on change la distribution de $\overline{X}$.

III- Test sur une proportion :

Pour effectuer un test sur une proportion on utilise la distribution suivante :

$$f \to N(p, \sqrt{\frac{p(1-p)}{n}})$$

$$\frac{f-p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0,1)$$

1- Test unilatéral à droite :

Les hypothèses du test sont :

$\begin{cases} H_0 : p = p_0 \\ H_1 : p > p_0 \end{cases}$

Il faut déterminer la frontière c de la région critique.

P (rejet H_0/ H_0 vraie)=α

$P(f>c \ p=p_o)= \alpha$

$$P(\frac{f-p}{\sqrt{\frac{p(1-p)}{n}}} \succ \frac{c-p_0}{\sqrt{\frac{p_o(1-p_o)}{n}}})=\alpha$$

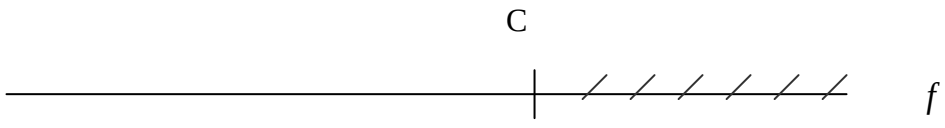
$$P(X^*<\frac{c-p_o}{\sqrt{\frac{p_o(1-p_0)}{n}}})=1-\alpha$$

$$t_\alpha = \frac{c-p_o}{\sqrt{\frac{p_o(1-p_0)}{n}}}$$

D'ou $c=p_0+t_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}$

Règle de décision : Rejet de Ho si f > c et acceptation de Ho si f ≤ c

Axe :



Acceptation H_0 : $p = p_0$

Rejet Ho : $p>p_0$

Exemple :
Aux dernières élections, un parti politique a obtenu 42% des suffrages. Un récent sondage a révélé que sur 1041 personnes interrogées 458 accordaient leur appui à ce parti. Le chef du parti déclara que a popularité de son parti était à la hausse. Tester la validité de cette affirmation au seuil de 5%.
Soit p la proportion des votes pour le parti politique et f la proportion empirique des votes observée sur l'échantillon.

$$f \to N(p, \sqrt{\frac{pq}{n}}) \quad d'où \quad \frac{f-p}{\sqrt{\frac{pq}{n}}} \to N(0,1)$$

f= k /n =458/1041 = 0.4399

Le test d'hypothèse consiste à vérifier si la proportion des votes est égale à l'ancienne valeur de 0.42 ou supérieure à cette valeur. Il s'agit donc d'un test unilatéral à droite sur la proportion qui s'exprime ainsi :

$$\begin{cases} H_0 : p = 0.42 \\ H_1 : p > 0.42 \end{cases}$$

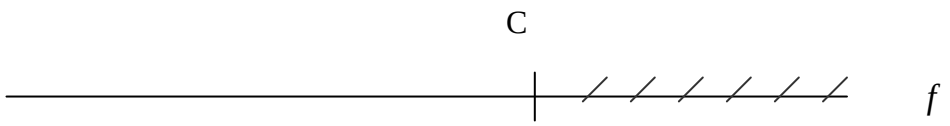
La règle de décision est donc rejet de H_0 si $f > c$.

avec $c = p_0 + t_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$

$\alpha = 0.05 \rightarrow F(t_\alpha) = 0.95 \rightarrow t_\alpha = 1.65$

$$c = 0.42 + 1.65\sqrt{\frac{0.42x0.58}{1041}} = 0.4452$$

Axe :



Acceptation H_0 : $p = p_0$

Rejet Ho : $p > p_0$

f= k /n = 0.4399<c → on rejette l'hypothèse H_1 selon laquelle la popularité de ce parti était à la hausse.

2- Test unilatéral à gauche :
Les hypothèses du test sont :

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p < p_0 \end{cases}$$

Il faut déterminer la frontière c de la région critique
P(rejet H_0/H_0 vraie)= α
$P(f<c / p=p_0)= \alpha$

$$P(X^* \prec \frac{c-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}})=\alpha$$

$$P(X^* \prec \frac{p_0-c}{\sqrt{\frac{p_0(1-p_0)}{n}}})=1-\alpha$$

d'ou $$\frac{p_0-c}{\sqrt{\frac{p_0(1-p_0)}{n}}}) = t_\alpha$$

D'ou $$c = p_0 - t_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$$

Axe et règle de décision :

Rejet Ho
$p<p_0$

Acceptation H_0

$p = p_0$

<u>3- Test bilatéral :</u>
Les hypothèses du test sont :

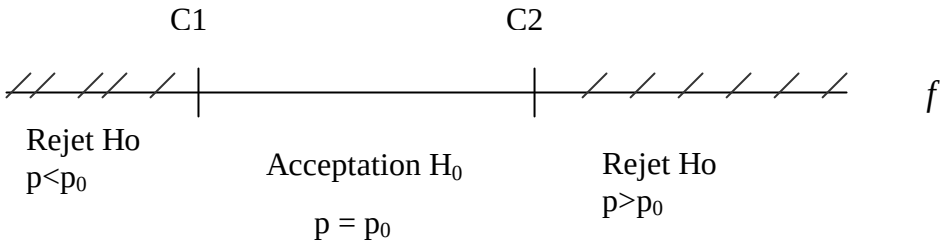
$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

P(rejet H_0/H_0 vraie)= α

$P(f>c_1 / P=P_0)+p(f<c_2 / P=P_0)= \alpha$
$P(f>c_1 / P=P_0)= \alpha/2$
$p(f<c_2 / P=P_0)= \alpha/2$
d'où :

$$c_1 = p_0 - t_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} \; ; \; c_2 = p_0 + t_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$$

Axe et règle de décision :



IV- Test sur la différence de deux moyennes :

Supposons qu'on veut effectuer un test sur la différence m_1-m_2 des moyennes de 2 populations qui sont normalement distribuées.
Si les écarts types sont connus, la statistique à utiliser est :

$$\frac{(\overline{X}_1-\overline{X}_2)-(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \rightarrow N(0,1)$$

L'expression de la région critique se détermine de la même façon que les tests précédents :

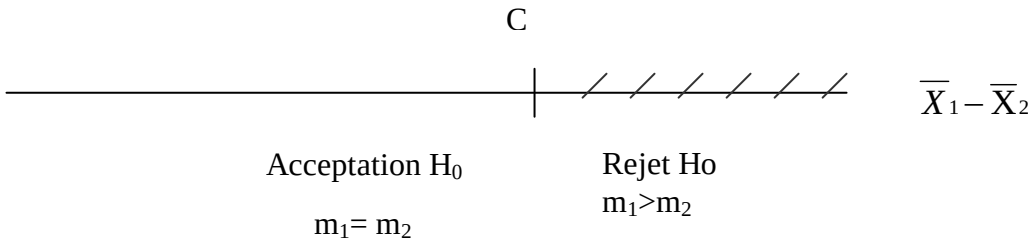
1- Test unilatéral à droite :

Les hypothèses du test sont :

$$\begin{cases} H_0 : m_1-m_2=0 \rightarrow m_1= m_2 \\ H_1 : m_1-m_2>0 \rightarrow m_1 > m_2 \end{cases}$$

$$c = t_\alpha \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Axe et règle de décision



2- Test unilatéral à gauche :

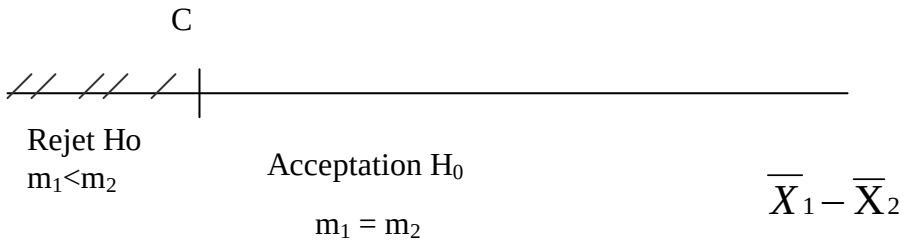
Les hypothèses du test sont :

$$\begin{cases} H_0 : :m_1-m_2=0 \rightarrow m_1= m_2 \end{cases}$$

$H_{1:}$: m_1-m_2<0 $\rightarrow m_1 < m_2$

$$c = - t_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Axe et règle de décision



3- Test bilatéral :

Les hypothèses du test sont :

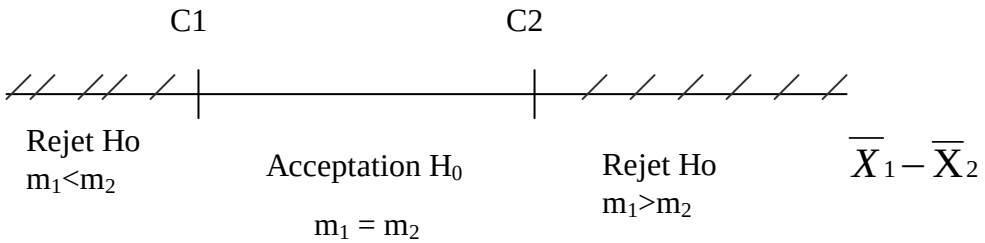
$$\begin{cases} H_0 : m_1\text{-}m_2 = 0 \rightarrow m_1 = m_2 \\ H_1 : m_1\text{-}m_2 \neq 0 \rightarrow m_1 \neq m_2 \end{cases}$$

La règle de décision est donc le rejet de H_0 si $\overline{X}_1 - \overline{X}_2 < c_1$ ou si $\overline{X}_1 - \overline{X}_2 > c_2$
Avec :

$$c_1 = - t_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$c_2 = t_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Axe et règle de décision



Remarque : Dans le cas où les variances sont inconnues, on utilise la distribution suivante :

$$\frac{(X_1 - X_2) - (m_1 - m_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \rightarrow T(n_1 + n_2 - 2)$$

et si $(n_1+n_2-2 \geq 30)$ On utilise alors l'approximation par la loi normale centrée réduite.

Application :
On a effectué des essais sur la durée de vie des ampoules de 60 Watts fabriquées par deux différentes entreprises. Les essais effectués dans les mêmes conditions sur un échantillon de 40 lampes de chaque entreprise ont donné les résultats suivants :

Entreprise 1	n_1=40	X_1=1025h	S_1=120h
Entreprise2 :	n_2= 40	X_2= 1070h	S_2= 140h

Peut-on affirmer au seuil de signification de 5% que les ampoules de l'entreprise 1 ont une durée de vie inférieure à celles de l'entreprise 2.

Soit X_1 la durée de vie des ampoules de l'entreprise 1 et X_2 la durée de vie des ampoules de l'entreprise 2. $X_1 \rightarrow N(m_1, \sigma_1)$ et $X_2 \rightarrow N(m_2, \sigma_2)$

σ_1 et σ_2 sont inconnus, et $(n_1+n_2-2 = 78 > 30)$ la statistique à utiliser est donc :

$$\frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \rightarrow N(0,1)$$

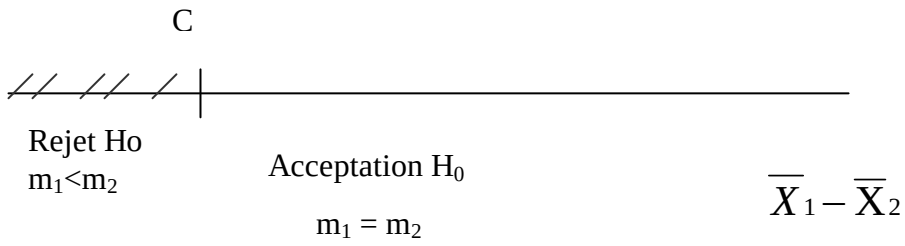
Le test d'hypothèse consiste à vérifier si la durée de vie des ampoules de l'entreprise est inférieure à celles fabriquées par l'entreprise 2. Il s'agit donc d'un test bilatéral à gauche sur la différence de deux moyennes qui s'exprime ainsi :

$$\begin{cases} H_0 : m_1 - m_2 = 0 \rightarrow m_1 = m_2 \\ H_1 : m_1 - m_2 < 0 \rightarrow m_1 < m_2 \end{cases}$$

La règle de décision est donc le rejet de H_0 si $\overline{X}_1 - \overline{X}_2 < c$, Avec

$$c = -t_\alpha \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Axe et règle de décision



$\alpha = 0.05 \rightarrow F(t_\alpha) = 0.95 \rightarrow t_\alpha = 1.65$ d'où :

$$c = -1.65\sqrt{\frac{120^2}{40}+\frac{140^2}{40}} = -48.1054$$

$\overline{X}_1 - \overline{X}_2$ = 1025-1070 = -45 > c → on accepte l'hypothèse H_0 → La durée de vie des ampoules de l'entreprise 1 n'est inférieure à celle de l'entreprise 2.

V- Test sur la différence de deux proportions

Supposons qu'on veut effectuer un test sur la différence de deux proportions p_1-p_2

La statistique à utiliser est :

$$\frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} \rightarrow N(0,1)$$

Sous l'hypothèse H_0 les valeurs p_1 et p_2 sont inconnus mais supposées égales à une valeur commune p. on obtient une estimation de cette valeur en combinant les observations dans chaque échantillon.

D'ou $p = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$

$$\frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})}} \rightarrow N(0,1)$$

L'expression de la région critique se calcule de la même façon que les tests précédents :

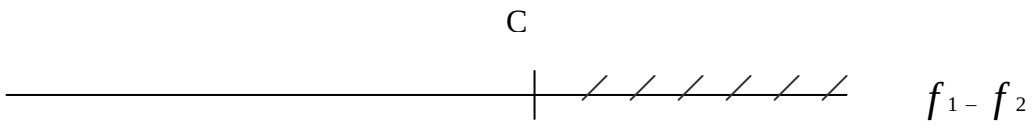
1- Test unilatéral à droite :

Les hypothèses du test sont :

$$\begin{cases} H_0 : p_1 - p_2 = 0 \\ H_1 : p_1 - p_2 > 0 \end{cases}$$

$$c = t_\alpha \sqrt{p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})}$$

Axe et règle de décision



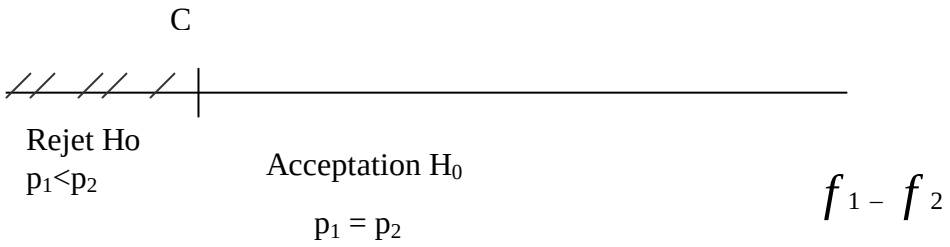
2- Test unilatéral à gauche :

Les hypothèses du test sont :

$\begin{cases} H_0 : p_1 - p_2 = 0 \rightarrow p_1 - p_2 = 0 \\ H_1 : p_1 - p_2 < 0 \rightarrow p_1 < p_2 \end{cases}$

$$c = - t_\alpha \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Axe et règle de décision



3- Test bilatéral :

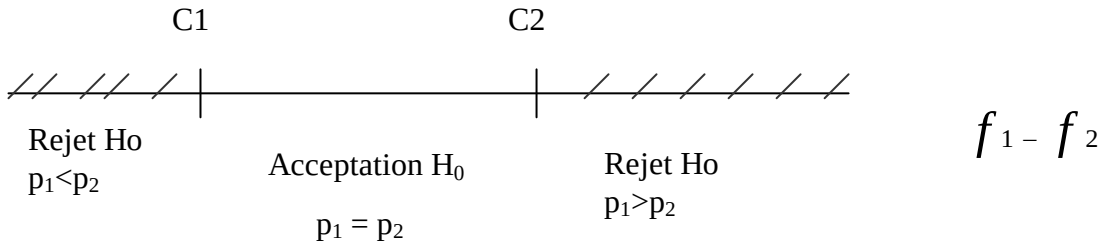
Les hypothèses du test sont :

$\begin{cases} H_0 : p_1 - p_2 = 0 \rightarrow p_1 = p_2 \\ H_1 : p_1 - p_2 \neq 0 \rightarrow p_1 \neq p_2 \end{cases}$

$$c_1 = - t_{\alpha/2} \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$c_2 = t_{\alpha/2} \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Axe et règle de décision



Exemple :

Pour la promotion de son nouveau produit, la direction commerciale d'une entreprise émet l'hypothèse selon laquelle la publicité du type A serait plus efficace que celle du type B. une enquête auprès de 125 individus ayant vu la publicité A a montré que 44 d'entre eux ont acheté le produit alors que sur 100 personnes ayant vu la publicité B, 32 ont acheté le produit.

Est- ce que les résultats de ce sondage permettent de confirmer au seuil de signification de 5%, l'hypothèse émise par la direction commerciale ?

Soit P_1 et P_2 les proportions des acheteurs exposés aux publicités A et B, et f_1 et f_2 les proportions observé des acheteurs ayant vu les publicités A et B.

$f_1 = k_1/n_1 = 44/125 = 0.352$ et $f_2 = k_2/n_2 = 32/100 = 0.32$.

$$\frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \rightarrow N(0,1)$$

Le test d'hypothèse consiste à vérifier si la publicité A est plus efficace ou pas que la publicité B. Ce test s'exprime ainsi :

$\begin{cases} H_0 : p_1 = p_2 \rightarrow p_1 - p_2 = 0 \\ H_1 : p_1 > p_2 \rightarrow p_1 - p_2 > 0 \end{cases}$

Sous l'hypothèse H_0 les valeurs p_1 et p_2 sont inconnus mais supposées égales à une valeur commune p. on obtient une estimation de cette valeur en combinant les observations dans chaque échantillon.

D'où $p = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = \frac{44+32}{125+100} = 0.3378$

$$\frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} \rightarrow N(0,1)$$

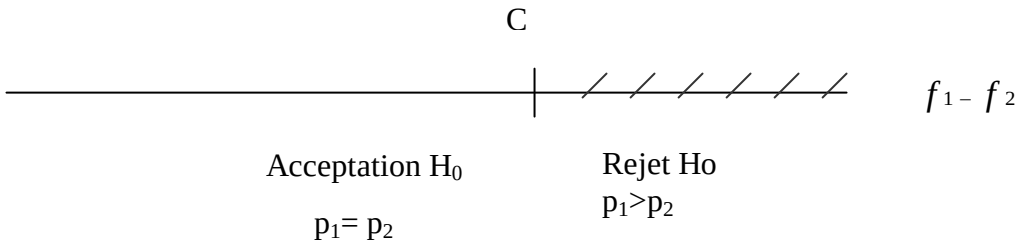
La règle de décision est donc le rejet de H_0 si $f_1 - f_2 > c$.

Avec $c = t_\alpha \sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}$

$\alpha = 0.05 \rightarrow F(t_\alpha) = 0.95 \rightarrow t_\alpha = 1.65$

$$c = 1.65 \sqrt{0.3378 x 0.6622 (\frac{1}{125} + \frac{1}{100})} = 0.1047$$

Axe et règle de décision



$f_1 - f_2 = 0.352 - 0.32 = 0.032 < c \rightarrow$ on accepte H_0 (les efficacités des deux publicités sont identiques).

Chapitre 7 :

Les tests non paramétriques

Dans ce chapitre, nous allons examiner comment nous pouvons construire un test sur la distribution d'une certaine variable X et sur la relation de cette variable avec d'autres.

I- Le test d'ajustement :

Considérons, dans une population, une certaine variable X dont on ne connaît pas la nature de la distribution .on tire un échantillon aléatoire de cette population et l'on se demande si nous pouvons accepter ou non que l'échantillon obtenu se conforme tout à fait à une distribution particulière spécifiée.
Les hypothèses du test sont :
H_0 : La distribution observée est tirée d'une population mère caractérisée par une distribution théorique f(x, θ) où θ est un paramètre connu ou inconnu.
H_1 : La distribution observée n'est pas tirée d'une population mère caractérisée par une distribution théorique f(x, θ)

Le prélèvement de l'échantillon permettra ultérieurement d'accepter ou de rejeter H_0 avec un risque d'erreur α.
Pour effectuer ce test on utilise la loi suivante :

$$\frac{\sum_{i=1}^{n}(n_i - np_i)^2}{np_i} \rightarrow \chi^2(k-l-1)$$

k : Le nombre de classes dans la distribution
l : Nombre de paramètres inconnus de la distribution (qu'on doit estimer)
p_i : Probabilité théorique de la classe i
np_i : Effectif théorique de la classe i
n_i : Effectif observé de la classe i

règle de décision :

Soit α le niveau de risque accepté. Le seuil critique χ^2_α est donné tel que :

$P(\chi^2 > \chi^2_\alpha) = \alpha$

Soit χ^2_c la statistique calculée à partir de l'échantillon :

- Si $\chi^2_c \leq \chi^2_\alpha$: On accepte l'hypothèse H_0

- Si $\chi^2_c > \chi^2_\alpha$: On rejette l'hypothèse H_0

Remarques :
- S'il existe un ou plusieurs effectifs théoriques $np_i<5$ on effectue des regroupements

- La taille de l'échantillon doit être suffisamment grande ($n \geq 30$)

Exemple1 (ajustement à une loi uniforme) :

Le responsable des jeux d'un casino veut vérifier si les dés utilisés sont bien équilibrés. Il prend au hasard d'un dé et le jette 120 fois. Les résultats obtenus sont résumés dans le tableau suivant :

Face	1	2	3	4	5	6
Effectif	14	16	28	30	18	14

Peut-il conclure au seuil de signification α =5% que le dé est bien équilibré ?

Il s'agit de tester l'hypothèse le dé est équilibré contre l'hypothèse le dé n'est pas équilibré.

H_0 : Le dé est équilibré (la répartition des résultats est uniforme)
H_1 : Le dé n'est pas équilibré (la répartition des résultats n'est pas uniforme)

Pour décider entre H_0 et H_1 il faut calculer la statistique de χ^2_c et la comparer à celle figurant sur la table χ^2_α

face	Effectif observé (ni)	Probabilité théorique (pi)	Effectif théorique ($n.p_i$)	$(n_i-n.p_i)^2/n.p_i$
1	14	1/6	20	$6^2/20=1.8$
2	16	1/6	20	$4^2/20=0.8$
3	28	1/6	20	$8^2/20=3.2$
4	30	1/6	20	$10^2/20=5$
5	18	1/6	20	$12^2/20=0.2$
6	14	1/6	20	$6^2/20=1.8$
	n=120			χ^2_c**=12.8**

ddl= k-l-1 = 6-0-1=5

$\chi^2_{5\%}(5)$ = 11.07 (d'après la table de χ^2)

$\chi^2_{5\%}(5) < \chi^2_c$

→ On retient l'hypothèse H_1 → le dé est truqué.

Rappel

Lorsqu'on veut tester l'ajustement d'une distribution donnée à une certaines loi on est parfois amené à estimer certains paramètres de la loi théorique.

Loi	Paramètres de la loi	Estimateurs
Normale	m et σ	$\overline{X}$ et S
Poisson	λ	$\overline{X}$
Binomiale	P	$\overline{X}$ /n' n' : Taille de chaque lot

Exemple2 (ajustement à une loi de Poisson) :

Dans une usine on a effectué une étude sur le nombre de pannes mensuelles des machines de production. Les données sur les cinq dernières années sont résumées dans le tableau suivant :

Nombre de pannes mensuelles	Nombre de mois
0	9
1	15
2	18
3	11
4	6
5 et plus	1

Est-ce qu'on peut dire au seuil de signification α=1% que le nombre de pannes mensuelles suit une loi de poisson ?

Il faut d'abord estimer le paramètre λ de la loi de poisson à partir de la moyenne des observations :

Nombre de pannes mensuelles x_i	Nombre de mois ni	$X_i\ n_i$
0	9	0
1	15	15
2	18	36
3	11	33
4	6	24
5 et plus	1	5
	$\sum n_i = 60$	$\sum x_i n_i = 113$

$\hat{\lambda} = \overline{X} = \sum x_i n_i / \sum n_i = 113/60 = 1.88$

Il s'agit de tester l'hypothèse que le variable aléatoire nombre de pannes mensuelles qu'on peut noter X suit une loi de poisson de paramètre

$\lambda = 1.88$ contre l'hypothèse que cette variable ne suit pas une loi de poisson.

H_0 : X ↝ P(1.88)

H1 : X ↛ P(1.88)

Nombre de pannes x_i	Effectif observé n_i	Prob théorique p_i $P(X=x_i) = e^{-1.88}1.88^x/\ x!$	Effectif théorique $e_i = n \times p_i$	Calcul de X^2 $= \sum(n_i - e_i)^2/e_i$
0	9	0.152	9.12	0.0016
1	15	0.286	17.16	0.272
2	18	0.269	16.14	0.214
3	11	0.168	10.08	0.084
4	6 } 7	0.079	4.74 } 6.528	0.0341
5 et plus	1 }	0.0298	1.788 }	
Total	**n=60**			χ^2_c **= 0.6057**

Après regroupement des effectifs théoriques inférieures à 5 on dispose de 5 classes d'où ddl= k-l-1 = 5-1-1 = 3

$\chi^2_{1\%}(3)$ = 11.345 (d'après la table de χ^2)

$\chi^2_{1\%}(3)$ > χ^2_c

→ On retient l'hypothèse H_0 → La variable aléatoire nombre de panne suit une loi de poisson de paramètre $\lambda = 1.88$

Exemple 3 (ajustement à une loi binomiale) :

Dans une entreprise fabriquant des tubes de verre, on effectue un contrôle visuel sur un échantillon de 320 lots formé chacun de 5 tubes.

Tubes défectueux	Nombre de lots
0	18
1	56
2	110
3	88
4	40
5	8
Total	320

Est-ce qu'on peut dire au seuil de signification α=5% que le nombre de tubes défectueux suit une loi binomiale?

Le nombre de tubes défectueux varie de 0 à 5, le premier paramètre de la loi n, est donc connue (n=5), quant à la proportion p des tubes défectueux p, elle peut être estimée à partir de la moyenne des observations :

Nombre de tubes défectueux x_i	Nombre de lot ni	$x_i\ n_i$
0	18	0
1	56	56
2	110	220
3	88	264
4	40	160
5	8	40
	$\sum n_i = 320$	$\sum x_i n_i = 740$

$\hat{P}=\bar{X} = \sum x_i n_i / n \sum n_i = 740 / (5x320) = 0.46$

Il s'agit de tester l'hypothèse que la variable aléatoire nombre de tubes défectueux qu'on peut noter X suit une loi de binomiale de paramètre n=5, p=0.46 contre l'hypothèse que cette variable ne suit pas une loi binomiale.

H_0 : X ↝ B(5 ; 0.46)

H1 : X ↛ B(5 ; 0.46)

Nombre de pannes x_i	Effectif observé n_i	Prob théorique p_i $P(X=x_i) = C_5^x\ 0.46^x \times 0.54^{5-x}$	Effectif théorique $e_i = n \times p_i$	Calcul de X^2 $= \sum(n_i - e_i)^2/e_i$
0	18	0.045	14.4	0.9
1	56	0.193	61.8	0.54
2	110	0.332	106.3	0.13
3	88	0.286	91.3	0.12
4	40	0.123	39.4	0.01
5	8	0.021	6.8	0.21
Total	**n=320**	$\sum$ **=1**	$\sum$ = 320	χ^2_c **= 1.91**

ddl= k-l-1 = 6-1-1 = 4

$\chi^2_{5\%}(4)$ = 9.49 (d'après la table de χ^2)

$\chi^2_{5\%}(4)$ > χ^2_c

→ On retient l'hypothèse H_0 → La variable aléatoire nombre de tubes défectueux suit une loi binomiale de paramètre n=5 et p=0.46

II- Le test d'indépendance :

On considère deux variable X et Y dans une même population. La question et de savoir si ces deux variable sont indépendantes ou liées. Le test de Khi-deux permet de répondre à cette question.

Les hypothèses à considérer se notent :

H_0 : Les deux variables X et Y étudiées dans cette population sont **indépendantes**

H_1 : Les deux variables X et Y étudiées dans cette population sont **dépendantes**

Pour effectuer ce test, on tire un échantillon de taille n de cette population. On compare les effectifs observés n_{ij} (associés aux différentes couples de modalités de X et Y) avec les effectifs théoriques n'_{ij}.

Les observations au niveau de l'échantillon sont regroupées dans un tableau de contingence :

	y_1	….	y_j	…..	y_r	n_j
X_1	n_{11}	….	n_{1j}	….	n_{1r}	n_1
….	……	…	….	…….	……	…….
X_j	n_{i1}		n_{ij}	….	n_{ir}	n_j
….	…	….	…..	…..	…..	……
X_k	n_{k1}	….	n_{kj}	…..	n_{kr}	n_k
n_j	n_1	…	$n_{.j}$	….	n_r	N

(k modalités pour la variable X)

(r modalités pour la variable Y)

Les effectifs théoriques en cas d'indépendance s'obtiennent en utilisant la propriété suivante :

$$n'_{ij} = \frac{n_{i.} * n_{.j}}{n}$$

On est donc en présence de deux distributions : L'une empirique et l'autre théorique dont on cherche l'adéquation. On utilise la statistique :

$$A = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

La statistique $A \rightarrow \chi^2[(k-1)(r-1)]$ sous les hypothèses que :

- Les k*r observations sont indépendantes
- La taille n de l'échantillon est suffisamment grande ($n \geq 30$)
- Les effectifs théoriques n'_{ij} sont suffisamment grands ($n'_{ij} \geq 5$)

Règle de décision :

Soit α le niveau de risque accepté. Le seuil critique X^2 α est donné tel que :

$P(X^2 > \chi^2_\alpha) = \alpha$

Soit χ^2_c la statistique calculée à partir de l'échantillon :

- Si $\chi^2_c \leq \chi^2_\alpha$: On accepte l'hypothèse H_0

- Si $\chi^2_c > \chi^2_\alpha$: On rejette l'hypothèse H_0

Exemple :

Une entreprise de produits cosmétiques fabrique trois types de produit. Afin d'ajuster sa stratégie marketing, elle veut connaître s'il y a ou non un lien de dépendance entre l'âge du consommateur et le type de produit préféré. Une enquête sur un échantillon de 200 personnes a donné les résultats suivants :

	Produit 1	Produit 2	Produit 3
Moins de 20 ans	13	19	25
Entre 20 et 40 ans	28	29	28
Plus que 40 ans	24	18	16

Effectuer le test adéquat pouvant aider cette entreprise (α=5%)

Si on désigne par X l'âge du consommateur et par Y le type de produit préféré, les hypothèses du test s'expriment ainsi :

H_0 : X et Y sont indépendants

H_1 : X et Y sont dépendants (le type de produit préféré dépend de l'âge du consommateur).

Tableau des effectifs observés :

	Produit 1	Produit 2	Produit 3	Totaux
Moins de 20 ans	13	19	25	57
Entre 20 et 40 ans	28	29	28	85
Plus que 40 ans	24	18	16	58
Totaux	65	66	69	200

Tableau des effectifs théoriques :

	Produit 1	Produit 2	Produit 3
Moins de 20 ans	18.525	18.81	19.665
Entre 20 et 40 ans	27.625	28.05	29.325
Plus que 40 ans	18.85	19.14	20.01

Le calcul de χ_c^2 :

χ_c^2 = (18.525-13)²/18.525 + (18.81- 19)²/18.81 + (19.665 -25)²/19.665
+ (27.625-28)²/27.625 + (28.05-29)²/28.05 + (29.325-28)²/29.325
+ (18.85 – 24)²/18.85 + (19.14- 18)²/19.14 + (20.01- 16)²/20.01
= 1.647 + 0.0019 + 1.447 + 0.005 + 0.0321 + 0.059 + 1.407 + 0.0678 + 0.836
= 5.5

ddl= (3-1) x(3-1) = 2x2 = 4

$\chi^2_{5\%}(4)$ = 9.488 (d'après la table de χ^2)

$\chi_c^2 < \chi^2_{5\%}(4)$

→ On retient l'hypothèse H_0 (le type de produit préféré est indépendant de l'âge du consommateur)

TRAVAUX

DIRIGES

TD chapitre 1 :

Notions de variables aléatoires et lois de probabilité

Exercice 1

On lance deux dés. Soit D la variable aléatoire « valeur absolue de la différence des chiffres montrés par les deux dés ».

a- Déterminer et représenter graphiquement la loi de probabilité de D.
b- Déterminer la fonction de répartition de D.
c- Calculer l'espérance et l'écart type de D.

Exercice 2

On lance trois fois une pièce de monnaie. Soit Z la variable aléatoire « nombre de fois où pile a été obtenu ».

a- Déterminer la loi de probabilité de Z.
b- Déterminer sa fonction de répartition.
c- Calculer l'espérance et l'écart type de Z.

Exercice 3

Le responsable commercial d'une entreprise estime que les ventes prévisionnelles peuvent présenter trois valeurs possibles : 5000 DT avec une probabilité de 50%, 7000 DT avec une probabilité de 30% et 8000 DT avec probabilité p.

1- Quelle est la probabilité d'avoir des ventes de 8000 DT ?
2- Déterminer les ventes moyennes.
3- Déterminer la variance des ventes moyennes.

Exercice 4

Soit X la variable aléatoire discrète dont la fonction de répartition est :

$$F(x) = \begin{cases} 0 & si \quad x < 2 \\ 0,6 & si \ 2 \le x < 4 \\ 0,85 & si \quad 4 \le x < 5 \\ 1 & si \ x \ge 5 \end{cases}$$

1. Déterminer la loi de probabilité de X.
2. Déterminer l'espérance la variance et l'écart type de X.

Exercice 5

La loi de probabilité relative à une variable aléatoire X est résumée dans le tableau suivant :

x_i	*1*	*2*	*3*	*4*
$P(X = x_i)$	0,2	a	b	0,3

1- Déterminer les valeurs de a et b sachant que E(X) = 2,5

2- Déterminer la fonction de répartition de X.

Exercice 6

Au début d'un mois M le stock initial d'un produit fini est une variable aléatoire dont l'espérance mathématique est égale à 120 et l'écart type à 20. La production de ce produit au cours de ce mois est une variable aléatoire dont l'espérance mathématique est égale à 40 et l'écart type à 9. Les ventes de ce produit au cours du même mois M est également une variable aléatoire dont l'espérance mathématique est égale à 30 et l'écart type à 12.
En supposant que ces variables aléatoires sont mutuellement indépendantes, déterminer les paramètres du variable aléatoire « stock final de ce produit fini ».

Exercice 7

Un investisseur dispose d'un portefeuille composé de 40% d'une action X et de 60 % d'une action Y. Les rendements possibles de ces actions sont comme suit :

Rx_i	*0.02*	*0.05*	*0.08*	*0.1*
$P(R_X = Rx_i)$	0.2	0.4	0.3	0.1

Ry_i	*- 0.05*	*0.04*	*0.07*	*0.12*
$P(R_Y = Ry_i)$	0.1	0.5	0.3	0.1

1- Déterminer les rendements espérés et la variance de chacune de ces actions.

2- Calculer le rendement espéré du portefeuille et sa variance sachant que la covariance entre les rendements des deux actions est $Cov(R_X, R_y) = 0.25$.

TD chapitre 2 : LES LOIS USUELLES DE PROBABILITE

Exercice 1

Une usine produit des règles en grandes quantités. La probabilité qu'une règle présente un défaut est de 0.1. On prélève au hasard et avec remise un échantillon de 5 règles dans la production d'une journée. On note X la variable aléatoire qui compte le nombre de règles présentant un défaut parmi les cinq règles prélevés.

1) Dégager la loi de probabilité de X.
2) Déterminer la fonction de répartition de X.
3) Donner l'espérance mathématique et la variance de X.

Exercice 2

En condition normale de fonctionnement, une machine produit des pièces défectueuses dans une proportion constante égale à 1%. Un client reçoit un lot de 300 pièces usinées par cette machine.
Quelle est la probabilité qu'il trouve moins de 3 pièces défectueuses dans ce lot ?

Exercice 3

Une entreprise commercialise un produit dont la demande mensuel X suit une loi normale d'espérance m=160 et d'écart type σ =15.

1) Déterminer P(130≤X≤175)
2) Déterminer P(X≤200)
3) Le gestionnaire des stocks pense que s'il garde un stock mensuel de 180 unités de produit il y aura moins de 5% de chance de tomber dans une rupture de stock. A-t-il raison ?
4) Quel stock mensuel faut-il garder pour garantir une probabilité de 90% de satisfaire la demande sans rupture de stock

Exercice 4

Un complexe sidérurgique fabrique des pipe-lines d'un diamètre moyen égal à 50 cm avec un écart type de 0.25 mm. Le cahier des charges alloue une tolérance sur le diamètre comprise entre 49.95 et 50.05 cm, sinon les canalisations seront refusées.
En admettant que le diamètre des pipe-lines est normalement distribué, quel sera le pourcentage de rejet de la production ?

Exercice 5

Une société minière exploite deux gisements. Le premier a une production journalière moyenne de 2000 tonnes avec un écart type de 150 tonnes. Le second a une production journalière moyenne de 3000 tonnes avec un écart type de 200 tonnes.
En admettant que les productions sont indépendantes et normalement distribuées, quelle est la probabilité que la production journalière de la société excède 5200 tonnes ?

Exercice 6

Un commerçant de chaussures reçoit en moyenne 8 clients par heure.

1. Dégager la loi de probabilité régissant le nombre de clients reçus en trente minutes en admettant qu'il suit une loi de poisson.
2. Calculer la probabilité de recevoir 3 clients.

TD Chapitre 3 : LES METHODES D'ECHANTILLONNAGE

Exercice 1

Hichem est un étudiant en Marketing. Suite à l'obtention de son diplôme, il a décidé de créer une entreprise dont l'objet est la commercialisation des autoradios et des accessoires autos. Afin d'évaluer les chances de succès de son nouveau projet, il envisage de mener une enquête auprès des propriétaires des voitures dans le district du Tunis. En s'adressant à l'institut national des statistiques (INS) il a pu disposer des données suivantes concernant le parc automobile du grand Tunis :

le nombre total des voitures citadines s'élèvent à 20 000 répartis comme suit :

Selon le lieu de résidence des propriétaires ;

- 7 000 dans le gouvernorat de BEN AROUS.
- 3 000 dans le gouvernorat de TUNIS.
- 6 000 dans le gouvernorat de MANOUBA.
- 4 000 dans le gouvernorat de L'ARIANA.

Selon leur usage ;

- 15 000 voitures particulières.
- 5 000 voitures utilitaires.

Selon leur puissance fiscale :

- 7 000 d'une puissance de 4 chevaux fiscaux.
- 7 000 d'une puissance de 5 chevaux fiscaux.
- 4 000 d'une puissance entre 6 et 8 chevaux fiscaux.
- 2 000 d'une puissance plus 8 chevaux fiscaux.

1) Quelle est d'après vous la base de sondage qu'on peut utiliser pour repérer la population mère ?
2) Hichem a décidé enfin d'appliquer la méthode de quota avec un taux de sondage de 1/20. illustrer à travers un tableau la structure de l'échantillon.
3) Quelles sont les autres méthodes d'échantillonnage qui vous paraissent appropriées pour mener l'enquête ?

Exercice 2

En vue d'évaluer la satisfaction vis-à-vis de ses produits un fournisseur de services internet, souhaite mener une enquête auprès d'un échantillon de 400 clients. Pour ce faire, les 6000 clients ont été numérotés de 0001 à 6000.

1- Quelles sont les différentes méthodes probabilistes (aléatoires) qu'on peut appliquer ? Quelle est la différence entre ces méthodes et les méthodes empiriques (non aléatoires) ?
2- Si on souhaite appliquer la méthode systématique et si le numéro du premier client choisi est 39, quels seront les numéros des cinq clients suivants ?

TD chapitre 5 : L'ESTIMATION

Exercice 1

On considère que les résultats obtenus par les étudiants pour un test d'aptitude en informatique sont distribués selon une loi normale de variance $\sigma^2 = 225$. Pour un échantillon aléatoire de 25 étudiants, le résultat moyen est de 70,6.

a- Quelle est l'estimation ponctuelle du résultat moyen de l'ensemble des étudiants ?
b- Estimer par intervalle de confiance résultat moyen de l'ensemble des étudiants avec un niveau de confiance de 99%.

Exercice 2

La longueur des tiges fabriquées par une entreprise est distribuée normalement avec une moyenne m et un écart type de 9 mm. Sur l'ensemble de la production mensuelle qui est de 280 tiges on effectue un contrôle de qualité en prélevant **sans remise** un échantillon de 36 tiges qui a donné une longueur moyenne de 25 cm.

a- Quelle serait l'estimation ponctuelle de m ?
b- Donner deux propriétés de cet estimateur.
c- Déterminer les limites de l'intervalle qui aurait 95 chances sur 100 d'encadrer la vraie valeur de m.
d- Quelle aurait-dû être la taille de l'échantillon si on avait fixé un niveau de confiance de 99% avec une marge d'erreur n'excédant pas 3 mm.

Exercice 3

Une entreprise tunisienne de construction mécanique fabrique une pièce de moteur de voiture pour un grand constructeur automobile européen. Le diamètre des pièces est supposé distribué normalement avec une moyenne m et un écart-type $\sigma = 2.75$ cm. En vue de contrôler le bon déroulement de la fabrication le contrôleur de qualité a prélevé avec remise un échantillon de 64 pièces et a mesuré leur diamètre ce qui a donné un diamètre moyen de 54 cm.

1- Déterminez un intervalle de confiance sur m pour un niveau de confiance $1-\alpha = 0.95$
2- Quelle est la marge d'erreur dans l'estimation obtenue.
3- Quelle aurait-dû être la taille de l'échantillon si on avait fixé un niveau de confiance de 99% avec une marge d'erreur n'excédant pas 0.5cm.
4- Quelle aurait-dû être la taille de l'échantillon a été prélevé avec remise et si on avait fixé un niveau de confiance de 99% avec une marge d'erreur n'excédant pas 0.5cm et un nombre de pièces fabriquées de 1000 unités.

Exercice 4

Pour juger de la teneur en magnésium d'une eau minérale, on a effectué 10 mesures :
248 ; 246 ; 246 ; 247 ; 247 ; 249 ; 247 ; 250 ; 248 ; 245 (mg par litre).

La teneur étudiée est supposée être une variable aléatoire normale d'espérance m et de variance σ^2.

1. Donner une estimation ponctuelle de la moyenne m.

2. Déterminez un intervalle de confiance sur m pour un niveau de confiance:
$1-\alpha = 0.95$.
3. Quelle est la marge d'erreur dans l'estimation obtenue ?

Exercice 5

Lors d'un récent sondage effectué auprès de la population étudiante, on a observé que sur un échantillon de 700 étudiants, 380 sont satisfaits de la qualité de la nourriture offerte à la cafétéria.

a- Estimer pour l'ensemble de la population étudiante, la proportion des étudiants satisfaits de la qualité de la nourriture offerte à la cafétéria avec un niveau de confiance de 95%.
b- Quelle est la marge d'erreur de ce sondage ?

Exercice 6

On veut estimer la proportion des électeurs qui vont voter pour le parti «TUNISIANO» lors des prochaines élections.
1- Quelle doit être la taille de l'échantillon afin d'estimer cette proportion avec un niveau de confiance de 99% et une marge d'erreur maximale de 5% ?
2- Sur cet échantillon, 180 ont l'intention de voter pour ce parti. Calculer un intervalle de confiance à 99% pour la proportion des voix que va recueillir ce parti

Exercice 7

Dans un échantillon de 1000 personnes résidant dans une ville A, on a observé 400 personnes abonnées au téléphone. Dans un échantillon de 1500 personnes d'une autre ville B, on a observé 675 abonnées.
Déterminer un intervalle de confiance pour la différence des proportions des abonnées dans chaque ville avec un niveau de confiance de 99%.

Exercice 8

Une multinationale possède deux filiales A et B dans deux pays différents. Le salaire des différents employés de la filiale A suit une loi normale de moyenne m_a et d'écart type 130$.
Pour la filiale B le salaire est également normalement distribué avec une moyenne m_b et un écart type de 150$.
Un échantillon de 20 employés de la filiale A a donné un salaire moyen de 450$ avec un écart type de 100$.
Un échantillon de 25 employés de la filiale B a donné un salaire moyen de 500$ avec un écart type de 120$.
Donner un intervalle de confiance à 95% pour la différence des salaires moyens des deux filiales dans les deux cas suivants:
a- Les écarts types des salaires des filiales sont connus.
b- Les écarts types des salaires des filiales sont inconnus

TD chapitre 6 : LES TESTS PARAMETRIQUES

Exercice 1

Une entreprise d'assemblage de composants électroniques utilise des plaques qui sont achetées auprès d'un fournisseur étranger. Ce dernier livre périodiquement des plaques qui doivent respecter une norme d'épaisseur moyenne de 6mm sinon le lot livré sera rejeté.
Pour vérifier si la dernière livraison est conforme à la norme, l'entreprise y a prélevé un échantillon de 16 plaques et a calculé une épaisseur moyenne de 6.1 mm.
En supposant que l'épaisseur des plaques est distribuée normalement avec une variance de 0.25 mm^2 , quelle sera la décision de l'entreprise avec un risque d'erreur de 5%.

Exercice 2

L'an dernier, le salaire hebdomadaire moyen payé par les entreprises aux spécialistes en informatique était de 475 dinars. Cette année, un échantillon aléatoire de 25 entreprises dans le domaine informatique révèle les faits suivants:

$$\overline{X} = 495 * et * \sum \left(X_i - \overline{X}\right)^2 = 9600$$

En supposant que le salaire est distribué normalement peut-on conclure au seuil de signification $\alpha = 0.05$ que le salaire hebdomadaire moyen présente une augmentation significative par rapport à l'an dernier ?

Exercice 3

Un constructeur automobile chinois souhaite vérifier le respect des normes écologiques par 60 unités d'un nouveau modèle de voitures qu'il a fabriqué. Pour ce faire, il a mesuré les émissions de monoxyde de carbone CO en grammes par kilomètre parcouru d'un échantillon de neuf voitures, ce qui a donné les résultats suivants :

2.2 ; 2 ; 2.3 ; 2.15 ; 2.35 ; 2.25 ; 2.5 ; 2.2 ; 2.3

1) En supposant que ces émission sont normalement distribuées avec une moyenne m et un écart type σ, donner une estimation un intervalle de confiance de m avec un niveau de confiance: $1-\alpha = 0.99$ dans chacun des cas suivants :
a- L'échantillon a été constitué par un tirage avec remise, et l'écart-type des émissions est supposé connu ($\sigma = 0.2$).
b- L'échantillon a été constitué par un tirage avec remise, et l'écart-type des émissions est supposé inconnu.
2) Quelle est la marge d'erreur associée à l'estimation de la question b.
3) Sachant que les normes écologiques du marché américain des voitures, exige des émissions de CO/gramme qui ne dépasse pas 2.2, peut-on conclure au seuil de signification $\alpha = 0.05$, que ce constructeur ne peut pas commercialiser ses voitures aux États-Unis en supposant les conditions de la question a (tirage avec remise et $\sigma = 0.2$).

Exercice 4

D'après une étude sur le comportement du consommateur, il semble que 2 consommateurs sur 5 sont influencés par les promotions sur les produits lors de l'achat. Le responsable marketing d'une grande surface a interrogé au hasard 200 consommateurs parmi lesquels 65 se disent influencés par les promotions.
Est-ce que ce sondage permet d'affirmer au seuil $\alpha = 0.01$ la conclusion de l'étude sur le comportement du consommateur ?

Exercice 5

1- Un bureau d'étude compte réaliser une enquête sur la proportion des tunisiens utilisant des moyens de paiement en ligne. Quelle doit être la taille de l'échantillon à envisager afin d'estimer cette proportion avec un niveau de confiance de 99% et une marge d'erreur maximale de 3,5%.
2- Sur cet échantillon on a trouvé 221 individus qui utilisent des moyens de paiement en ligne. Déterminer alors un intervalle de confiance à 99% pour la proportion relative à l'ensemble des tunisiens utilisant des moyens de paiement en ligne.
3- Peut-on conclure au seuil de 1% que la proportion des tunisiens utilisant des moyens de paiement en ligne est supérieure à 11%.

Exercice 6

On veut estimer la proportion des consommateurs qui préfèrent une marque X de Jus de fruit.

1- Quelle doit être la taille de l'échantillon afin d'estimer cette proportion avec un niveau de confiance de 99% et une marge d'erreur maximale de 5% ?
2- Sur cet échantillon, 180 ont l'intention d'acheter cette marque. Calculer un intervalle de confiance à 99% pour la proportion des consommateurs qui vont acheter cette marque.
3- Peut-on conclure au seuil de 99% que cette proportion est inférieure à 50% ?

Exercice 7

Une firme cherche à établir si la proportion des pièces acceptables est plus élevée chez un fournisseur étranger (P1) que chez un fournisseur local (P2).
L'observation réalisée sur un échantillon extrait des livraisons de chaque fournisseur a donné:

- Sur un échantillon de 100 pièces du fournisseur étranger 90% sont acceptables.
- Sur un échantillon de 80 pièces du fournisseur local 70% sont acceptables.

Tester au niveau de signification de 1% l'hypothèse:

$H0 : P1 = P2$
contre $H1 : P1 > P2$

Exercice 8

Selon les statistiques des 10 dernières années le Ministère du tourisme tunisien a constaté que les dépenses en devises des touristes allemands et français sont normalement distribuées avec un écart type respectivement de 50 et 40 dollars.
Pour mieux cibler son action promotionnelle, le Ministère désire savoir s'il y a ou non une différence significative au seuil de 5% entre les dépenses moyennes des touristes allemands et ceux français. On a prélevé ainsi un échantillon de chaque catégorie et les résultats sont résumés dans le tableau suivant :

	Allemands	Français
Taille de l'échantillon	500	400
Dépense moyenne	1200 $	1150 $

a- Quel est le test statistique pouvant répondre aux besoins du Ministère du tourisme ?
b- Quelle est la décision sur la base de ce test ?
c- Quelles seront les conséquences sur l'action promotionnelle du Ministère ?

TD chapitre 7 : LES TESTS NON PARAMETRIQUES

Exercice 1

On a réalisé un sondage auprès de 200 étudiants afin de connaître leur préférence entre cinq marques de sport. Les résultats sont résumés dans le tableau suivant :

Marque	**Adidas**	**Puma**	**Nike**	**Reebok**	**Lotto**
$\mathbf{n_i}$	50	35	60	25	30

Peut-on conclure au seuil de 5% que la distribution des préférences des étudiants est égalitaire ?

Exercice 2

Une compagnie d'assurance désire connaître la loi de la variable aléatoire X représentant le nombre d'accidents de la route par semaine dans une ville. L'observation sur une période d'une année a donné les résultats suivants :

Nombre d'accidents	0	1	2	3	4	5	6	7	8 et +
Nombre de semaines	8	16	8	3	6	5	3	2	1

Peut-on conclure que la variable aléatoire X : «nombre d'accidents de la route par semaine dans cette ville», suit une loi de poisson ? (risque : 1%)

Exercice 3

Une entreprise reçoit ses commandes d'un fournisseur par boites de contenant chacune trois ampoules. L'entreprise a contrôlé au hasard 300 lots et a noté les résultats suivants :

Nombre d'ampoules défectueuses x_i	0	1	2	3
Nombre de boites observés n_i	190	95	10	5

Soit Y la variable aléatoire : « Nombre d'ampoules défectueuses par boite».
Tester au niveau de signification de 5%, si Y est distribuée suivant une loi binomiale.

Exercice 4

Un test de produit a été mené auprès d'un échantillon de consommateurs de deux régions afin de savoir s'il existe ou non une relation entre la région et le produit préféré. Pour ce test 4 produits différents ont été testés tout en notant à chaque fois la préférence du répondant :

	Produit 1	**Produit 2**	**Produit 3**	**Produit 4**
Région A	40	80	50	40
Région B	20	30	60	80

Que peut-on conclure au seuil de 1%