TD09 - STATISTIQUE INFERENTIELLE. INTERVALLES DE CONFIANCE.

Considérations théoriques

La majorité d'études statistiques ne se réalise pas sur toute la population statistique à cause d'un ou de plusieurs inconvénients :

- la dimension de la population peut être parfois très grande ;
- le temps effectif d'étude est proportionnel au nombre d'entités étudiées ;
- les coûts et les ressources allouées augmentent proportionnellement au nombre d'entités étudiées ;
- il y a des situations dans lesquelles on ne peut pas recueillir des informations sur tous les individus de la population.

Donc, pour étudier les paramètres des caractéristiques de la population statistique on procède de façon suivante :

a. on extrait un echantillon représentatif. On choisit la taille de l'échantillon de manière à ce que l'on permette une étude exhaustive des caractéristiques ;
b. en fonction de la nature des caractéristiques (quantitatives ou qualitatives) en utilisant la statistique descriptive, on détermine les principaux paramètres ;
c. en utilisant la statistique d'inférence on essaie d'estimer les paramètres pour toute la population, à partir de résultats obtenus au niveau de l'échantillon.

Exemple procédé pour une variable quantitative :

La moyenne et l'écart standard théorique d'une caractéristique pour une population ne sont pas connus.

On extrait de manière aléatoire un échantillon représentatif pour lequel on détermine la moyenne et l'écart standard.

On essaie d'estimer les paramètres de la population à partir de valeurs des paramètres observés.

L'estimation peut être ponctuelle ou par l'intermédiaire d'un intervalle de confiance.

L'estimation ponctuelle

Un estimateur d'un paramètre est une fonction qui dépend de résultats obtenus sur un échantillon extrait de manière aléatoire.

La valeur donnée par la fonction attachée à l'estimateur s'appelle **estimation ponctuelle** du paramètre et elle est une variable aléatoire.

L'estimation de l'intervalle de confiance

L'estimation ponctuelle d'un paramètre théorique fournit une valeur qui dépend de fluctuations de l'échantillonnage et donc elle peut être différente de la valeur réelle du paramètre estimé. Ainsi, l'estimation d'un paramètre théorique ne se fait pas par une seule valeur mais par un intervalle dans lequel le paramètre estimé se trouve à une probabilité importante appelée intervalle de confiance.

Il est recommandable d'estimer un paramètre théorique par un intervalle, appelé intervalle de confiance (confidence interval CI), dans lequel on peut affirmer que le paramètre estimé se trouve à une probabilité elevée.

ESTIMATION A L'AIDE DE L'INTERVALLE DE CONFIANCE

- L'intervalle de confiance est un intervalle limité de valeurs (les limites s'appellent des limites de confiance), qui inclue la moyenne de la caractéristique étudiée.
- Plus l'intervalle est large, plus on est sûr que la moyenne de la caractéristique étudiée se retrouvera dans cet intervalle.
- La taille de la confiance, la confidence, est donnée par la probabilité que la valeur (les valeurs) étudiée(s) se trouvent dans cet intervalle.
- La confiance (confidence) utilisée fréquemment est de 95% , 99% ou 99.9%

Le cas d'une variable quantitative - l'estimation d'une moyenne

Soit P une population dans laquelle la variable X a une moyenne théorique μ inconnue.
De la population P on extrait au hasard l'échantillon E représentatif.
Dans l'échantillon E pour la variable X on observe une moyenne m et on calcule une variation ponctuelle estimée s^2;
On essaie de déterminer pour la valeur inconnue de la moyenne théorique μ un intervalle de confiance au seuil α, (à l'aide de m et s^2 observés), c'est-à-dire déterminer un intervalle [a,b] dans lequel la probabilité que la moyenne théorique μ soit connue est de 1-α:
$Pr(a \leq \mu \leq b) = 1 - \alpha$

α s'appelle seuil de signification ou risque
1 – α s'appelle niveau de confiance ou niveau de confidence
a et b s'appellent limites de confiance

La détermination des intervalles de confiance se fait à base de formules de calcul présentées ci-dessous :

I. Quand **σ (déviation standard de la population) est connu**, l'intervalle de confiance pour la moyenne μ au seuil de signification α est:

$$\bar{x} - u_\alpha \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + u_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

$u_\alpha \cdot \frac{\sigma}{\sqrt{n}}$ s'appelle marge d' erreur

α=0,05 sau α=5%	u_α=1,96
α=0,01 sau α=1%	u_α=2,58
α=0,001 sau α=0,1%	u_α=3,29

II. Si **σ n' est pas connu**

a. L'échantillon a n>120, dans le calcul de l'intervalle de confiance on utilise la déviation standard de l'échantillon

$$\bar{x} - u_\alpha \cdot \frac{S}{\sqrt{n}} < \mu < \bar{x} + u_\alpha \cdot \frac{S}{\sqrt{n}}$$

$u_\alpha \cdot \frac{S}{\sqrt{n}}$ s'appelle marge d' erreur

α=0,05 sau α=5%	u_α=1,96
α=0,01 sau α=1%	u_α=2,58
α=0,001 sau α=0,1%	u_α=3,29

b. L'échantillon a n<120, dans le calcul de l'intervalle de confiance on utilise la déviation standard de l'échantillon

$$\bar{x} - t_{\alpha,n-1} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha,n-1} \cdot \frac{S}{\sqrt{n}}$$

$t_{\alpha,n-1} \cdot \frac{S}{\sqrt{n}}$ s'appelle marge d' erreur

$t_{\alpha,n-1}$ est lu du tableau de distribution "t" au niveau α et n-1 degré de liberté ou à l'aide de la fonction TINV implémentée en EXCEL.

TINV (Fonction TINV ou LOI.STUDENT.INVERSE)
Retourne l'envers de la distribution t Etudiant bi-alternative
Cette fonction a été remplacée avec une ou plusieurs nouvelles fonctions qui peuvent offrir une précision améliorée et dont les noms reflètent mieux leur utilisation .
(fonction T.INV.2T ou fonction T.INV.)

Syntaxe

TINV(probabilité,degrés_liberté)

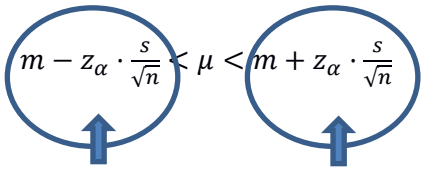
La Syntaxe de la fonction TINV a les arguments suivants :

- **Probabilité** Obligatoirement. C'est la probabilité associée à la répartition t Etudiant bi-alternative.
- **Degrés_liberté** Obligatoirement. C'est le nombre des degrés de liberté qui caractérise la répartition.

Détermination du volume n de l'échantillon

Il y a des situations dans lesquelles le chercheur est préoccupé par la taille de l'échantillon étudié de manière qu'on puisse estimer un paramètre à une certaine probabilité.

A l'estimation de la moyenne théorique μ d'une variable quantitative définie sur une population statistique P, l'intervalle de confiance à base de la moyenne observée m est calculé sur un échantillon représentatif E à la taille (volume) n.
Pour un seuil de signification α (risque) et une précision δ de l'intervalle de confiance [m-δ; $m+\delta$, tenant compte du fait que l'intervalle de confiance est exprimé:



δ δ

$$2\delta = 2z_\alpha$$

On obtient :

$$n = \left(z_\alpha \cdot \frac{s}{\delta} \right)^2$$

Donc le volume de l'échantillon peut être exprimé en fonction de:

- Précision
- Déviation standard
- probabilité

Exemple de problème résolu pour le calcul de l'intervalle de confiance

Sur un échantillon aléatoire de 100 nouveaux-nés on connait la moyenne du poids à la naissance des nouveaux-nés : 3275 g et l'écart standard: 854.
Quel est l'intervalle de confiance estimé à une probabilité de 95% pour la moyenne du poids à la naissance sur la population dont on extrait l'échantillon?

La formule qui s' applique est (on connait la déviation standard pour la population):

$$\bar{x} - u_\alpha \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + u_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

α=0,05 ou α=5% u_α=1,96

La limite inférieure de l'intervalle de confiance :

$$3275 - \frac{1{,}96 * 854}{10} = 3275 - 558 = 3107{,}6$$

Moyenne Marge d'erreur Limite inférieure

La limite supérieure de l'intervalle de confiance :

$$3275 + \frac{1{,}96 * 854}{10} = 3275 + 558 = 3442.6$$

Moyenne Marge d'erreur Limite supérieure

L'intervalle de confiance estimé à la probabilité de 95% pour la moyenne du poids à la naissance de la population étudiée est : [3107,6 ; 3442,4]

Le calcul de la valeur qui s'additionne et se soustrait de la valeur moyenne (marge d' erreur) peut se faire à base de la formule de calcul ou à l'aide des fonctions implémentées en Excel (fonction CONFIDENCE ou INTERVALLE.CONFIANCE.STUDENT).

La fonction CONFIDENCE
On retourne la marge d'erreur, c'est-à-dire la valeur qui doit être soustraite, respectivement additionnée, à la valeur moyenne pour trouver l'intervalle de confiance. La moyenne de l'échantillon, x, est au centre de cet intervalle, et l'intervalle est x ± CONFIDENCE.

CONFIDENCE(alfa,dev_standard,dimension)
La Syntaxe de la fonction CONFIDENCE a les arguments suivants :

Alfa Obligatoire. C'est le seuil de signification ou le risque utilisé pour calculer le niveau de confiance. **Dev_standard** Obligatoire. C'est la déviation standard de la population pour la zone de données et on suppose être connue.
Dimension Obligatoire. C'est la dimension de l'échantillon.

Dans l'exemple donné, l'appellation de la fonction CONFIDENCE avec les arguments suivants :
CONFIDENCE(0,05;854;100), génère la valeur : 167,3809 Moyenne= 3275
On obtient :
Limite inférieure: 3107,619
Limite supérieure: 3442,381

Objectifs:

- l'assimilation du moyen de détermination de l'**intervalle de confiance (de confidence) pour la moyenne d' un échantillon**
- La détermination du nombre d' observations que doit contenir l'échantillon pour déterminer un intervalle de confiance d'un paramètre d'une certaine dimension.

Pour le set de 100 valeurs, la valeur CI peut être calculée aussi à l'aide de l'option Statistiques Descriptives de Analyses de données.

3585	2376	3197	1919	4817	4619	3734	3042	2908	936	3519	3644	3384
2716	3179	1854	2976	4043	3229	3755	4034	3632	4159	2654	3134	2856
3311	2845	1511	3462	3755	2682	2329	3091	3045	3579	2446	3410	3560
4347	3966	4009	2541	2865	2697	2890	3336	2984	2285	3641	2987	4108
2964	3471	2652	4351	2522	3398	3274	3059	3250	3475	4106	4194	
3155	4087	3168	3553	2917	3213	4282	3795	4212	3309	2907	2798	
3196	3350	2659	2536	3586	2790	2085	3635	3135	4181	3324	2894	
5005	3598	2684	4474	4008	3209	3975	2946	3010	1823	2336	2734	

Appelant les fonctions implémentées en Excel

moyenne	=AVERAGEA(A1:J10)
dev standard	=STDEVA(A1:J10)
CONFIDENCE	=INTERVALLE.CONFIANCE.STUDENT(0,05;O4;100)
Limite inférieure	=O3-O5
Limite supérieure	=O3+O5

On obtient:

moyenne	3248,71
dev standard	709,65
CONFIDENCE	140,8104

CONFIDENCE.T

Limite inférieure	3107,87
Limite supérieure	3389,49

Appelant **Statistiques descriptives** de Data Analyses de données on obtient :
Attention! Les 100 valeurs doivent apparaitre sur une colonne dans la feuille de calcul de Excel.
La valeur de la ligne portant l'étiquette Confidence Level(95,0%) représente la valeur qui sera soustraite et additionnée à la moyenne pour obtenir la limite inférieure et supérieure de l'intervalle de confiance.

Colonne1	
Moyenne	3248,708
Erreur Standard	70,96509
Mediane	3210,982
Module	#N/A
Deviation Standard	709,6509
Variance	503604,3
Voûte	0,642855
Asymétrie	-0,22385
Plage	4069,028
Minimum	935,989
Maximum	5005,017
Somme	324870,8
Nombre d'échantillon	100
Confidence Level(95,0%)	140,8101

Exemple de problème résolu pour trouver le nombre d' observations

Trouvez le nombre d'observations qu'on doit effectuer pour estimer la taille moyenne de la population masculine adulte si l'on connait que S^2=9 cm^2 et on demande une précision δ=1 cm, qui soit assurée avec la probabilité 1-α= 0.95.

Risque α=0.05
Déviation standard S=3 cm
Précision δ=1 cm
z_α=1,96

$$n = \left(1{,}96 \cdot \frac{3}{1}\right)^2 = 35{,}574 \simeq 35$$

pour n=34
$z_{0,05;34}$=2,033

$$n = \left(2.033 \cdot \frac{3}{1}\right)^2 = 37{,}197 \simeq 37$$

37 observations sont nécessaires pour estimer la taille moyenne à la précision de 1 cm.

Problèmes proposés.

1. Calculez à une probabilité de 95%, 99% et 99.9% l'intervalle de confiance pour la valeur moyenne de TAS, pour une population à une répartition normale dont on a extrait un échantillon de 40 individus auxquels on a mesuré les valeurs suivantes de TAS.

Nr_id	TAS (mm/Hg)
1	126
2	130
3	135
4	116
5	122
6	126
7	128
8	130
9	123
10	124
11	140
12	125
13	120
14	121
15	140
16	135
17	115
18	135
19	132
20	128
21	126
22	125
23	115
24	117
25	121
26	125
27	129
28	124
29	131
30	127
31	127
32	125
33	140
34	115
35	110
36	112
37	125
38	143
39	122
40	135

2. Il y a l'hypothèse que les malades souffrant d'artrite rhumatoide ont un risque majeur de développer l'ostéoporose. On a mis sous observation 20 malades d' artrite rhumatoide, tous de sexe masculin et ayant le même âge. Pour ces patients on a mesuré le niveau de la calcitonine (HCT). Les valeurs HCT observées se trouvent ci-dessous :

Nr_id	HCT (pg/ml)
1	26
2	21
3	44
4	30
5	37
6	20
7	35
8	39
9	48
10	19
11	19
12	22
13	27
14	39
15	38
16	20
17	21
18	21
19	23
20	45

On exige :

Calculez l'intervalle de confiance pour la moyenne HCT inconnue de la population souffrant d' artrite rhumatoide, connaissant les valeurs mesurées pour un échantillon formé de 20 individus.

3. Le manager d'une compagnie pharmaceutique est préoccupé parce que les médicaments similaires sont vendus aux prix sensiblement différents dans les pharmacies de proximité. C'est pourquoi on a enregistré les prix de vente d'un même médicament dans 40 pharmacies. Les prix de vente sont:

26	34	14
32	12	28
33	17	29
12	36	24
11	25	19
19	15	21
18	24	28
21	23	20
21	29	22
40	30	27
39	30	22
27	26	27
30	17	
32	23	

Estimez à une probabilité de 95% l'intervalle de confiance du prix de vente du médicament dans les 40 pharmacies.

4. Le niveau sanguin du calcium chez 52 poulets broiler constitués dans le Lot L1, nourris dans des conditions d' apport d' aluminium en ration et le niveau sanguin du calcium chez 52 poulets broiler constitués dans le Lot témoin est présenté dans le tableau suivant:

Lot 1				Lot témoin			
17,7	8,9	11,6	9,6	13,4	15,2	13,5	9,2
9,6	9,5	15,6	10,2	9,5	13,5	14,2	15,2
12,4	9,6	14,2	11,6	9,9	14,2	10,2	13,5
11,6	10,2	16,2	15,6	9,6	9,8	11,6	14,2
11,7	11,6	14,6	14,2	10,7	10,1	14,3	9,9
10,2	15,6	12,4	16,2	10,6	10,4	10,7	11,2
11,6	14,2	11,6	14,6	15,2	11,3	15,2	10,7
12,9	12,4	11,7	12,4	12,9	10,7	13,5	10,6
9,8	11,6	10,2	11,6	11,6	10,6	14,2	15,2
8,9	11,7	14,2	11,7	10,3	15,2	11,1	12,9
9,5	10,2	12,4	10,2	9,8	12,9	15,2	11,6
9,6	8,9	11,6	10,1	8,9	11,6	13,5	10,3
10,2	9,5	11,7	10,7	15,2	10,3	14,2	12,4

Estimez à la probabilité de 95% l'intervalle de confiance pour la valeur moyenne du niveau sanguin du calcium pour le lot témoin.

5. La production de lait en millilitres chez deux lots de vaches de la race A et de la race B est présentée dans le tableau ci-dessous:

race A				race B		
2,9	5,9	4	5	3,8	5	5,1
5,7	5	4,3	2	4,7	5,4	3,6
4,6	2,4	5	2,9	4	6	5
3,9	5,1	2	4,6	6	5,4	3,8
4,2	3,6	2,9	3,9	3,6	4,1	4
3,8	5,1	5,7	4	4,2	5,2	6
4,3	3,6	4,6	2,2	3,1	5,2	3,6
3	5,4	3,9	4,3	4,3	6	4,2
5,3	3,5	4,2	5,7	4,1	4,1	3,1
2,8	4	3,8	5	5,2	5	4,3
2,5	4,2	4,3		5,2	5	4,1
4,6	5	3		4,1	5,2	
5,2	2,2	5,3		5	5,2	

Calculez l'erreur standard de la moyenne pour les échantillons de la race A et B.
Estimez à la probabilité de 95% l'intervalle de confiance pour la production de lait pour la race B.

6. Suite à une diète prescrite par le fameux diététicien Dr. C, on perd du poids hebdomadairement, en moyenne 2 kg.
Les pertes de poids chez un lot de 50 personnes qui ont suivi cette diète sont présentées dans le tableau ci-dessous :

0,3	3,1	3,3	0,6
2,2	1,2	1,5	3,7
0,8	3,7	1,1	2,4
4	1,5	3,7	2,8
0,5	3,9	0,3	3,3
1	3,3	2,4	3,8
0,5	1,6	0,4	0,5
0,6	1,8	0,4	3
1	0,8	0,9	0
3,1	1,5	2,1	1,2
0,8	2,9	3,7	0,7
0,5	0	1,1	2
3,4	0,8		

a. Déterminez le pourcentage des cas existant entre les intervalles :
$(\acute{x}-s;\acute{x}+s)$;
$(\acute{x}-2s;\acute{x}+2s)$;
b. Estimez à une probabilité de 95% l'intervalle de confiance pour la valeur moyenne de la perte de poids pour le lot étudié .

7. Combien d'observations sont nécessaires pour déterminer l'intervalle de confiance pour la moyenne du poids à la naissance à une précision de 280g (δ=2,58) et à un risque de 1% sachant que ce paramètre a une distribution normale à une déviation standard de 709,65g ?

8. Combien d'observations sont nécessaires pour déterminer l'intervalle de confiance pour la moyenne du poids perdu à la suite d'une diète, au risque de 5%, à la précision de 0,5kg sachant que ce paramètre a une distribution normale à une déviation standard de 1,2688kg (et δ=1,96)?