Rappel

- **<u>Cours 1:</u>**
  - Chapitre 1 : rappels (probabilités, densité de probabilité, fonction de répartition, espérance, variance), lois de probabilités discrètes et continues classiques
  - Chapitre 2 : statistiques descriptives (1 et 2 variables), représentations graphiques, résumés statistiques, covariance et corrélation
  - Devoir de maison: régression linéaire

Chapitre 3 - Estimation

Théorie de l'échantillonnage

- Population et Variable Aléatoire X:

La population E est un ensemble fini. Les éléments de E sont appelés *individus*. Chaque individu E_k a une mesure réelle x_k. On considère l'expérience aléatoire mesure qui consiste à tirer un individu au hasard dans E, l'univers de cette expérience étant E. Soit alors la variable aléatoire (VA) mesure X définie à partir de cette expérience par :

$$X : E \to \mathbf{R}$$

$$e_k \to x_k$$

On note:
avec N le nombre d'individus dans la population

$$E(X) = \frac{1}{N}\sum_{k=1}^{N} x_k = \mu$$

$$V(X) = \frac{1}{N}\sum_{k=1}^{N} (x_k - \mu)^2 = \sigma^2$$

Théorie de l'échantillonnage

- n-échantillonnage de E

On considère l'expérience aléatoire n-échantillonnage qui consiste à tirer, avec remise, n individus au hasard dans E. Soit alors, la VA définie à partir de cette expérience par :

$$X_k : \quad E^n \to \mathbf{R}$$

qui associe au n-uplet tiré la mesure du k° individu tiré. On peut alors considérer le vecteur aléatoire :

$$(X_1,\ldots,X_n) : E^n \to \mathbf{R}^n$$

$X_1,\ldots,X_n$ suivent la même loi et sont indépendantes

Théorie de l'échantillonnage

- Statistique de n -échantillonnage:

Par définition c'est une VA qui est fonction des VA X_1, X_2 …X_n. Les plus connues sont les statistiques d'échantillonnage:

Moyenne d'échantillonnage:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Variance d'échantillonnage:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

Moyenne d'échantillonnage

1 - Fluctuations de la moyenne d'échantillonnage:

Pour être en mesure d'estimer la moyenne de la population par intervalle de confiance ou encore d'effectuer un test d'hypothèse sur la moyenne , il faut connaître la VA $\bar{X}$

Distribution d'échantillonnage de $\bar{X}$*:* c'est la distribution des différentes valeurs que peut prendre la moyenne d'échantillonnage calculée sur tous les échantillons possibles, de même taille, d'une population donnée.

Pour mieux comprendre à quoi correspond $\bar{X}$ nous allons prendre un EXEMPLE

Attention dans la suite du cours nous n'aurons jamais accès à toutes les valeurs prises par $\bar{X}$!!!

Moyenne d'échantillonnage

Exercice:

Expérience d'échantillonnage avec remise dans la population:

→ détermination de la distribution de la moyenne de 2-échantillonnage.

Supposons qu'une population consiste en 5 contenants (numérotés de 1 à 5) et que le poids respectif de chacun est : x_1=332g, x_2=336g, x_3=340g, x_4=344g et x_5=348g.

1/ Déterminer la moyenne, la variance et la distribution de la population (du caractère « poids » de la population). Soit X la VA qui associe à chaque contenant son poids, alors :

Moyenne :

Variance :

Distribution :

Moyenne d'échantillonnage

Supposons maintenant que nous voulons former tous les échantillons possibles de taille n = 2 de cette population en effectuant un échantillonnage avec remise. Il y a dans ce cas 5^2 échantillons différents possibles. Chaque échantillon a donc une probabilité égale à 1/25 d'être choisi.

Echantillon n°	Contenant n°	Résultat de l'échantillonnage	Moyenne sur l'échantillon
1	(1,1)	(332,332)	332
2	(1,2)	(332,336)	334
3	(1,3)	(332,340)	336
4	(1,4)	(332,344)	338
5	(1,5)	(332,348)	340
6	(2,1)	(336,332)	334
7	(2,2)	(336,336)	336
8	(2,3)	(336,340)	338
9	(2,4)	(336,344)	340
10	(2,5)	(336,348)	342
11	(3,1)	(340,332)	336
12	(3,2)	(340,336)	338
13	(3,3)	(340,340)	340
14	(3,4)	(340,344)	342
15	(3,5)	(340,348)	344
16	(4,1)	(344,332)	338
17	(4,2)	(344,336)	340
18	(4,3)	(344,340)	342
19	(4,4)	(344,344)	344
20	(4,5)	(344,348)	346
21	(5,1)	(348,332)	340
22	(5,2)	(348,336)	342
23	(5,3)	(348,340)	344
24	(5,4)	(348,344)	346
25	(5,5)	(348,348)	348

Moyenne d’échantillonnage

2/ Déterminer ensuite la distribution de la moyenne d’échantillonnage:

Moyenne d’échantillonnage

3/ Pour finir, calculer la moyenne et la variance de la moyenne d’échantillonnage et les exprimer en fonction de la moyenne et de la variance de la population.

Moyenne :

Variance :

Moyenne d'échantillonnage

Remarque 1:

Si nous répétons cette expérience en prélevant cette fois des échantillons de taille n = 3, il y aura 5^3 = 125 échantillons possibles (tirage avec remise) et la moyenne de la distribution d'échantillonnage de sera à nouveau 340 g.

Toutefois la variance des moyennes d'échantillonnage va diminuer : on obtiendrait 32/3. De plus, la forme de la distribution d'échantillonnage de s'approchera de plus en plus de celle d'une loi normale.

Réaliser cet exemple ainsi que le précédent sur Matlab

Remarque 2 :

Dans la suite du cours et de manière plus générale en statistique vous n'aurez jamais accès à toutes les valeurs prises par la moyenne d'échantillonnage ! **Cette partie avait pour but de vous faire comprendre à quoi correspond la variable $\bar{X}$.**

Moyenne d'échantillonnage

- **Paramètres de la moyenne d'échantillonnage**

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 mesurable (réalisation d'une VA X) qui suit une loi de probabilité de moyenne $E(X) = \mu$ et de variance $V(X) = \sigma^2$, alors la moyenne d'échantillonnage $\bar{X}$ suit une loi de probabilité moyenne :

$$E(\bar{X}) = E(X) = \mu$$

et de variance :

$$V(\bar{X}) = \frac{V(X)}{n} = \frac{\sigma^2}{n}$$

A connaître car on s'en resservira tout le temps!!!

Moyenne d'échantillonnage

2 - Forme de la distribution de la moyenne d'échantillonnage

Pour caractériser complètement les fluctuations de la moyenne d'échantillonnage, il faut également être en mesure de préciser la forme probabiliste des fluctuations. Pour connaître exactement la distribution de $\bar{X}$, il faut connaître la distribution de la population qui a été échantillonnée ou alors utiliser le **théorème central limite.**

Loi suivie par $\bar{X}$ lorsque l'on considère que X suit une loi normale

$\bar{X}$ en tant que somme de n VA indépendantes toutes normales, suit une loi normale de moyenne E(X) et de variance V(X) ; alors la VA centrée réduite suit une loi normale centrée sur 0 et d'écart type 1.

Moyenne d'échantillonnage

Exercice :

Les billes de roulement fabriquées par une société ont une masse moyenne de 5.02 g avec un écart type de 0.3 g. On considère que la variable qui associe à une bille sa masse moyenne suit une loi normale.

Calculer la probabilité qu'un échantillon de 15 billes ait une masse moyenne supérieure à 5.3 g.

Moyenne d'échantillonnage

Convergence en loi de $\bar{X}$ lorsque la taille de l'échantillon n tend vers …. même si la loi suivie par X n'est pas connue.

$\forall n \in N^*$, $X_1, X_2, \ldots, X_n$ sont des VA à valeurs réelles indépendantes qui suivent toutes la même loi (celle de X) ayant pour espérance μ et pour variance σ^2

alors :

$$\bar{X}^* = \frac{\bar{X} - E(\bar{X})}{\sqrt{V(\bar{X})}}$$

converge en loi vers N(0,1) (la loi normale centrée réduite)

avec $\bar{X}$ la moyenne d'échantillonnage (taille n)

On utilisera l'approximation dès que $n \geq 30$

Moyenne d'échantillonnage

Cas important de la fréquence d'échantillonnage :

Tous les individus ont pour mesure 0 ou 1. On note p = P (X = 1) : p la **proportion d'individus de mesure 1**. X est une VA de Bernouilli de paramètre p

D'où : $E(X) = p$

$V(X) = pq = p(1-p)$

Ici la moyenne d'échantillonnage $\overline{X}$ est appelée fréquence d'échantillonnage et correspond à la fréquence de sortie du 1 dans le n -échantillon.

On pourra également noter cette variable F:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = F$$

Moyenne d'échantillonnage

Exercice :

Quelle est la probabilité pour qu'en 120 lancers d'une pièce équilibrée la proportion de faces soit comprise entre 40% et 60% ?

Proposer deux solutions : approche probabiliste et approche statistique.

Estimation

L'estimation des paramètres est l'objectif fondamental de l'échantillonnage d'une population.

Introduction

Un aspect important de l'inférence statistique est celui d'obtenir à partir de l'échantillonnage d'une population, des estimations fiables de certains paramètres de cette population. Dans ce cours, les paramètres que nous allons estimer sont **la moyenne μ** (ou **la proportion p encore notée f**). Ces estimations peuvent s'exprimer soit par une seule valeur (estimation ponctuelle), soit par un intervalle (estimation par intervalle).

Estimation ponctuelle

Lorsqu'une caractéristique d'une population est estimée par un seul nombre, déduit des résultats de l'échantillon, ce nombre est appelé une **estimation ponctuelle**.

Soient α un paramètre de la population et Y une statistique de n-échantillonnage.

Y est un estimateur non biaisé de
$\alpha \Leftrightarrow \mathrm{E}(\mathrm{Y})=\alpha$

Si y est la valeur de Y sur un *n*-échantillon donné alors on dit que y est une estimation non baisée de α.

L'estimation ponctuelle se fait à l'aide d'un estimateur. Cet estimateur est fonction des observations de l'échantillon. L'estimation est la valeur numérique que prend l'estimateur selon les observations de l'échantillon.

Estimation ponctuelle

Que dire de la moyenne d'échantillonnage par rapport à la moyenne de la population?

Estimation par intervalle de confiance

Les estimations ponctuelles ne fournissent aucune information concernant la précision des estimations. Elles ne tiennent pas compte de l'erreur possible dans l'estimation, erreur attribuable aux fluctuations d'échantillonnage.

Quelle confiance avons-nous en une valeur unique ? On ne peut répondre à cette question en considérant uniquement l'estimation ponctuelle. Il faut lui associer un intervalle qui permet d'englober avec une certaine fiabilité, la vraie valeur du paramètre correspondant.

On va ici partir de plusieurs exemples et de ce que l'on connaît (c'est-à-dire de ce qui a été vu aux chapitres précédents) donner les intervalles de confiance de l'estimation d'une moyenne ou d'une proportion.

Estimation par intervalle de confiance

Exercice :

Une université compte plus de 5000 étudiants. On tire un échantillon de 100 étudiants. Sa masse moyenne est de 67.45 kg et sa variance de 8.5275.

1/ Quelle est la masse moyenne de l'ensemble des étudiants ? On a évidemment envie de dire que la masse moyenne de tous les étudiants est d'environ 67.45 kg. Quel est le raisonnement qui permet en toute rigueur de le dire ?

2/ On veut ici bien entendu préciser le « environ ». On va donc chercher le *y* tel que :
P(67.45 - *y* < masse moyenne de tous les étudiants < 67.45 + *y*) = 0.95
Calculer y

Remarque : On dit que [67.45 - *y* ; 67.45 + *y*] est l'intervalle de confiance pour la masse moyenne de tous les étudiants au niveau de confiance N_C = 95%.

Intervalle de confiance d'une moyenne

Population normale de variance connue ou grand échantillon ($n \geq 30$)

A partir d'un échantillon aléatoire de taille n d'une population normale de variance connue σ^2 on définit, en prenant comme estimation ponctuelle de μ la moyenne de l'échantillon $\bar{x}$, un intervalle de confiance ayant un niveau de confiance N_c% de contenir la vraie valeur de comme suit :

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

où $z_{\frac{\alpha}{2}}$ est la valeur de la variable normale centrée réduite telle que la probabilité que Z soit compris entre $-z_{\frac{\alpha}{2}}$ et $z_{\frac{\alpha}{2}}$ est $1-\alpha = N_c$

Intervalle de confiance d'une moyenne

Population normale de variance inconnue ou grand échantillon ($n \geq 30$)

Dans le cas d'un grand échantillon ($n \geq 30$) provenant d'une population de variance inconnue mais estimée par la variance d'échantillon s^2 , alors l'intervalle de confiance

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Population normale de variance inconnue et $n < 30$

A partir d'un échantillon aléatoire de petite taille ($n < 30$), prélevé d'une population normale de moyenne m (inconnue) et de variance s^2 inconnue, alors on définit, en prenant comme estimation ponctuelle de m la moyenne $\bar{x}$ de l'échantillon, un intervalle de confiance ayant un niveau de confiance N_c% de contenir la valeur vraie de m comme suit :

$$\bar{x} - t_{\frac{\alpha}{2};\nu} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2};\nu} \frac{s}{\sqrt{n}}$$

Intervalle de confiance d'une moyenne

Intervalle de confiance d'un pourcentage (proportion)

A partir d'un échantillon aléatoire de taille n et en prenant comme estimation ponctuelle de f la fréquence observée $\hat{f}$ d'avoir un certain caractère qualitatif, on définit l'intervalle de confiance de p (la proportion d'éléments (individus) possédant un caractère qualitatif dans la population) avec un niveau de confiance N_c % :

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

ou

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

Cet intervalle de confiance est valable à condition que: $n\hat{f} \geq 5$ et $n(1-\hat{f}) \geq 5$

Intervalle de confiance d'une moyenne

Exercice :

Une population de 2000 individus vote. Il y a 2 candidats. Un échantillon de n = 100 électeurs indique que 55% d'entre eux ont voté pour Duchmol.

1/ Donner une estimation non biaisée de la moyenne de la population.

2/ Donner l'intervalle de confiance à 99% pour le pourcentage de la population ayant voté pour Duchmol.

3/ Au vue de l'échantillon quelle est la probabilité que Duchmol soit élu ?

Intervalle de confiance d'une moyenne

Exercice: Détermination de la taille de l'échantillon requise pour un essai de fiabilité d'un dispositif électronique

Une firme vient de développer un nouveau dispositif électronique qui entre dans la fabrication d'appareils de traitement de texte. Avant de mettre en production ce nouveau dispositif, on veut effectuer des essais préliminaires pour être en mesure d'estimer la fiabilité en terme de durée de vie. D'après le bureau d'étude de l'entreprise, l'écart type de la durée de vie de ce nouveau dispositif électronique serait de l'ordre de 100 heures.

Déterminer :
1/ Le nombre d'essais requis pour estimer, avec un niveau de confiance de 95%, la durée de vie moyenne d'une grande production de sorte que la marge d'erreur dans l'estimation n'excède pas 50 heures.

2/ Le nombre d'essais requis (pour le même niveau de confiance) pour estimer la durée de vie avec une marge d'erreur de 20 heures.