Ce qu'il surtout savoir – Statistiques

Statistique descriptive

• Savoir calculer une moyenne et un écart type pour un échantillon de valeurs donné (statistique à une variable).

• Savoir trouver, avec une calculatrice, l'équation d'une droite de régression (méthode des moindres carrés, statistique à deux variables).

Statistique inférentielle

• La moyenne (disons m) des valeurs d'une caractéristique d'une population est estimée par celle d'un échantillon ($\bar{x}$) ; de même pour une fréquence p (ou un pourcentage) d'apparition ; l'écart type des valeurs d'une population (disons σ) est estimé par l'écart type (notons le σ_e) d'un échantillon de taille n multiplié par $\sqrt{\dfrac{n}{n-1}}$: $\boxed{\sigma \simeq \sigma_e\sqrt{\dfrac{n}{n-1}}}$.

• Pour avoir plus d'information concernant la moyenne, on peut contruire un intervalle de confiance $I = \left[\bar{x} - t\dfrac{\sigma}{\sqrt{n}}\,;\,\bar{x} + t\dfrac{\sigma}{\sqrt{n}}\right]$ centré en $\bar{x}$ devant contenir m avec un coefficient de confiance donné par l'énoncé, t dépendant du coefficient de confiance. Pour une fréquence, l'intervalle est $\boxed{I = \left[f - t\sqrt{\dfrac{f(1-f)}{n-1}}\,;\,f + t\sqrt{\dfrac{f(1-f)}{n-1}}\right]}$ où f est la fréquence constatée sur l'échantillon.

Tests de validité bilatéral

Exemple de cas type : des pièces fabriquées par une machine doivent avoir un diamètre moyen $d = 100$ avec un écart type de 2. On veut vérifier que c'est le cas en effectuant un prélévement de $n = 1000$ pièces et en mesurant leur diamètre moyen $\bar{x}$.

– **hypothèse nulle** H_0 : tout est conforme (la norme est respectée), ici $d = 100$; **hypothèse alternative** H_1 : ce n'est pas le cas, donc $d \neq 100$

– on recherche un intervalle $I = [d-a; d+a]$ centré en d devant contenir toutes les valeurs de $\bar{X}$, diamètre moyen des pièces d'un échantillon quelconque de n pièces, avec une probabilité donnée (appelé **coefficient de confiance**) ou un **seuil de risque** α %. Pour cela on utilise le fait que, sous l'hypothèse H_0, $\bar{X}$ suit approximativement la loi normale $\mathscr{N}\left(d, \dfrac{\sigma}{\sqrt{n}}\right) = \mathscr{N}\left(100; \dfrac{2}{\sqrt{1000}} \simeq 0,063\right)$.

On peut alors écrire la région d'acceptation de H_0, centrée en d :
$I = \left[d - t\dfrac{\sigma}{\sqrt{n}}\,;\,d + t\dfrac{\sigma}{\sqrt{n}}\right] = \left[100 - 1,96\dfrac{2}{\sqrt{1000}}\,;\,100 + 1,96\dfrac{2}{\sqrt{1000}}\right] \simeq$
$[\,99,876\,;\,100,124\,]$.
On peut (re)trouver cet intervalle à l'aide de Geogebra.

– on énonce la **règle de décision** :
" On prélève au hasard et avec remise un échantillon de 1000 pièces et on calcule la moyenne $\bar{x}$ de leur diamètre. Si $\bar{x}$ appartient à I alors on accepte H_0 avec le seuil de risque de α % sinon on rejette H_0 ".

– on regarde si la valeur $\bar{x}$ obtenue sur un échantillon (qui est donnée ou doit être calculée) se trouve dans cet intervalle ou au contraire dans la **région critique** et on prend une décision conformément à la règle ci-dessus.

Tests de validité unilatéral

Ici l'hypothèse alternative H_1 s'écrirait $d > 100$ (ou $d < 100$).
On ne peut plus utiliser alors la formule $I = \left[d - t\dfrac{\sigma}{\sqrt{n}}\,;\,d + t\dfrac{\sigma}{\sqrt{n}}\right]$.
Par exemple, si l'on veut tester que la moyenne est supérieure à 100 (en supposant qu'elle ne peut pas être inférieure à 100) alors H_1 est « $d > 100$ » et H_0 est « $d = 100$ ».
Si H_0 est vraie alors la moyenne observée ne devrait pas trop s'éloigner de 100 donc on cherche a tel que $P(\bar{X} \leqslant a) = 0,95$ (pour 95 % de confiance).
Sachant que sous H_0, $\bar{X}$ suit approximativement la loi normale $\mathscr{N}(100; 0,063)$, on trouve (calculatrice ou Geogebra) $a \simeq 100,1036$.
La région d'acceptation de H_0 est donc $]-\infty\,;\,100,1036\,]$.
Puis, on énonce la règle de décision, etc.