LES ÉPREUVES DE STATISTIQUES DANS LES BTS

On observe, dans les propositions de sujets qui sont faites pour les épreuves de mathématiques des différents BTS, de nombreuses erreurs dans les exercices portant sur les statistiques. Certaines d'entre-elles se retrouvent d'ailleurs dans les épreuves proposées aux candidats, ce qui fait désordre. La crédibilité de la formation dispensée aux élèves est en cause. De plus, il est bien connu que les annales régulent l'enseignement fait. Des sujets correctement rédigés inciteront les professeurs à améliorer la qualité scientifique de leur cours. L'objet de la présente note est d'attirer l'attention des inspecteurs sur les plus courantes de ces erreurs.

I. STATISTIQUE DESCRIPTIVE

1- Le regroupement en classe des données

Quand le nombre d'observations d'une variable réelle est important, il est d'usage de faire des classes et de fournir un tableau où figurent les classes et le nombre d'observations par classe. À noter qu'il s'agit d'une première étape de la démarche statistique. On ne s'intéresse plus aux individus mais à la population en donnant une idée de la distribution de celle-ci. En contre-partie, on perd de l'information.

Dans ces conditions, demander aux élèves de **calculer médiane, moyenne, écart-type de la variable n'a pas de sens**, car on ne connaît plus les valeurs prises par la variable, mais seulement les intervalles dans lesquels elles se trouvent.

Tout au plus peut-on en calculer des valeurs approchées. La tradition veut que pour trouver une valeur approchée de la médiane on fasse l'approximation suivante : à l'intérieur d'une classe les observations sont distribuées uniformément. Cela veut dire que, si n est le nombre d'observations, la fonction de répartition dite empirique

$$F_n : x \mapsto \frac{1}{n} \times (\text{nombre d'observations inférieures à } x)$$

qui n'est connue qu'aux extrémités des classes, est approximée par une fonction continue affine par morceaux.

En revanche, pour trouver une valeur approchée de la moyenne et de l'écart-type, on procède à l'approximation suivante : on fait comme si toutes les observations d'une classe avaient comme valeur commune le centre de la classe. $F_n(x)$ est alors approximée par une fonction en escalier dont les points de discontinuité sont les centres des classes. Pour le calcul de la moyenne, les deux approximations conduisent à la même valeur. Il n'en est pas de même pour le calcul de la variance et donc de l'écart-type. La variance est systématiquement sous-estimée car on néglige la variation à l'intérieur de chaque classe. Si σ_1^2 est l'approximation de l'écart-type σ^2 de la série statistique calculée ainsi, on a $\sigma_1^2 < \sigma^2$.

Soit σ_2^2 l'approximation de σ^2 faite en supposant, comme pour la médiane, les observations uniformément distribuées à l'intérieur de chaque classe. S'il y a k classes et si I_j est la largeur de la classe j, un calcul simple montre que :

$$\sigma_2^2 = \sigma_1^2 + \frac{1}{12n} \sum n_j I_j^2.$$

σ_2^2 est en général plus proche de σ^2 que σ_1^2. Cela permet d'avoir une idée de l'erreur faite en remplaçant σ par σ_1 et **d'éviter des questions du type : les résultats sont donnés à 10^{-2} près alors que l'erreur de méthode ainsi calculée montre que l'on ne peut obtenir qu'une approximation de l'ordre de l'unité**.

2- La régression linéaire

On considère n observations bivariées (x, y). Dans de nombreux cas on a entre y et x une liaison qui peut être représentée par une relation affine aux fluctuations près. On pose alors $y_i = ax_i + b + \varepsilon_i$ où a et b sont deux coefficients à déterminer. La méthode des moindres carrés consiste à déterminer a et b tels que $\sum_{i=1}^{n} \varepsilon_i^2$ soit minimum. x est appelé variable explicative et y variable à expliquer. Cette méthode est liée à la description euclidienne des données. Si dans l'espace euclidien à n dimensions E_n, y et x sont les vecteurs de coordonnées

respectives $(y_1 \dots y_n)$, $(x_1 \dots x_n)$, $\overline{y}$ et $\overline{x}$ les vecteurs ayant toutes leurs coordonnées égales respectivement à la moyenne de y et à la moyenne de x, le vecteur $a(x-\overline{x})$ est la projection orthogonale de $(y-\overline{y})$ sur $(x-\overline{x})$. Le vecteur ε de coordonnées $(\varepsilon_1 \dots \varepsilon_n)$ est donc orthogonal à $(x-\overline{x})$.

Si on pose $x_i = \alpha y_i + \beta + \eta_i$, y devient la variable explicative et x la variable à expliquer. La même méthode consiste à déterminer α et β tels que η soit minimum. Les deux droites représentatives sont évidemment distinctes et elles se coupent en G point moyen, de coordonnées $(\overline{x}, \overline{y})$. Dans E_n on projette alors orthogonalement $(x-\overline{x})$ sur $(y-\overline{y})$.

Il est donc absurde de faire déterminer dans une première question l'équation de la droite des moindres carrés où y est la variable à expliquer puis à faire prévoir x quand y prend une valeur donnée. Il fallait alors faire la régression de x en y et non celle de y en x.

Il doit y avoir une cohérence entre le modèle et son utilisation.

La détermination de a et b (ou de α et β) nécessitait avant l'usage des calculatrices des calculs longs et pénibles. Aussi avait-on cherché des méthodes empiriques donnant un ajustement affine approximatif dans les cas où $\sum \varepsilon_i^2 \ll \sum (y_i - \overline{y})^2$. L'une des plus célèbres est la méthode de Meyer. On coupe le nuage des n points dans E_2 (le point M_i a pour coordonnées (x_i, y_i) en deux (ou trois) sous-nuages). Celui qui correspond à des abscisses x_i inférieures à t_1, celui qui correspond à des abscisses supérieures à t_2, les deux sous-nuages étant de même effectif. Si G_1 et G_2 sont les points moyens de ces deux sous-nuages, la droite représentative de la relation affine est la parallèle à G_1G_2 passant par G voire (G_1G_2). Cette méthode plaît à des professeurs de mathématiques car elle fait faire des calculs de moyennes, mais elle ne repose sur aucune modélisation. **Elle est donc à proscrire**, les calculatrices effectuant les calculs sans difficultés. À la limite autant faire ajuster à l'oeil une droite sur une représentation graphique de nuage.

Le coefficient de corrélation ρ représente le cosinus de l'angle des vecteurs $(y-\overline{y}, x-\overline{x})$ dans E_n, il est donc caractéristique de la qualité de la représentation. Dans trop de sujets $\rho > 0,98$ ce qui dans beaucoup de cas dits concrets, est trop beau pour être vrai. Ce sont des données artificielles qu'il vaut mieux éviter.

2. STATISTIQUE INDUCTIVE

1- Modèle probabiliste et statistique

Dans les BTS industriels, la statistique inductive est une partie importante du programme. Elle trouve son application en contrôle de fabrication et en fiabilité. Dans l'industrie il existe des procédures normalisées dont la description est faite dans les publications de l'AFNOR (normes ISO ou AFNOR). Pour les mettre en œuvre nul besoin de comprendre ce qu'est la statistique inductive, il suffit d'exécuter les instructions d'un algorithme. Trop souvent les sujets sont du type « faites comme on vous a appris à faire » et négligent la partie « comprendre », la plus intéressante.

Rappelons que la situation concrète est caractérisée par un modèle probabiliste dont certains paramètres sont inconnus. L'observation faite est considérée comme une réalisation de la situation concrète aléatoire modélisée. L'objet de la statistique est de dire des choses sur les paramètres inconnus du modèle donc de les mesurer au sens large du terme. **Il est donc absurde de demander aux élèves de mettre en œuvre une procédure sans spécifier le modèle pour lequel elle est adéquate**. Au niveau du BTS, sauf en maintenance, les seuls modèles considérés sont n tirages indépendants dans une urne à deux catégories ou n repétitions indépendantes d'une variable gaussienne de moyenne et/ou de variance inconnue. La crédibilité du modèle dépend des conditions expérimentales. Il importe donc de rappeler le modèle, ou bien au moins en partie, les conditions de l'expérience qui le valident. Cela est vrai en particulier pour l'indépendance des observations.

Il faut aussi être rigoureux au niveau du langage. 3,5 n'est pas une variable aléatoire et si μ est le paramètre inconnu, écrire $P(\mu < 3{,}5)$ n'a pas de sens : 3,5 est la réalisation d'une variable aléatoire suivant par exemple une loi normale de moyenne μ et d'écart-type 1. Il ne faut pas confondre une variable aléatoire et sa réalisation. Une fois la réalisation faite, il n'y a plus de probabilité, le modèle probabiliste est dans l'action.

2- Les procédures statistiques

Les deux seules procédures statistiques enseignées sont l'estimation et le test. Pour l'estimation on distingue l'estimation ponctuelle et l'estimation par intervalle. L'estimation ponctuelle ne pose pas problème ; en revanche l'estimation par intervalle est l'occasion de nombreuses fautes. Souvent on confond confiance et probabilité. L'intervalle de confiance avant résultat expérimental est aléatoire. On cherche deux variables aléatoires L et U telles que si μ est le paramètre à estimer on ait $P_\mu\big([L,U] \ni \mu\big) = 1-\alpha$ où α est fixé. En général $\alpha = 0,05$,

μ est inconnu mais fixé, P_μ est la loi qui régit le phénomène. On a un constat expérimental que l'on note ω, $L(\omega)$ et $U(\omega)$ sont les réalisations de L et de U. On dit que $L(\omega) < \mu < U(\omega)$ avec la confiance $1-\alpha$ pour rappeler que la procédure utilisée est telle que l'intervalle aléatoire dont $\big[L(\omega), U(\omega)\big]$ est une réalisation, avait une probabilité $1-\alpha$ de recouvrir la valeur μ inconnue. Il faut bien **distinguer confiance et probabilité**.

De même pour les tests. On choisit arbitrairement une hypothèse nulle. On détermine dans l'ensemble des observations une zone de probabilité supérieure ou égale à $1-\alpha$ si l'hypothèse nulle est vraie et de probabilité la plus petite possible quand elle est fausse. Si l'issue observée est dans cette zone, cela ne veut pas dire que l'hypothèse nulle est vraie, cela veut dire qu'avec cette hypothèse l'issue observée est vraisemblable au niveau $1-\alpha$ et qu'il n'est pas utile de changer l'hypothèse, celle-ci ayant été choisie en fonction de sa commodité. **Il est indispensable dans un test de préciser l'hypothèse nulle, l'alternative** (souvent la négation de la première) **et le seuil choisi**, et d'employer un vocabulaire précis : « Est-ce que $\mu = \mu_0$ » n'a pas le même sens que : « Tester l'hypothèse $\mu = \mu_0$ ». Dans une procédure statistique, seule la deuxième formulation a un sens.

Dans le but de ne pas surcharger les programmes, la notion de **paramètre nuisible** n'est pas abordée. Par contre elle apparaît dans les problèmes dans la situation suivante : $X_i \dots X_n$ sont n variables aléatoires indépendantes de même loi, la loi normale de moyenne μ et d'écart-type σ, tous deux inconnus mais l'inférence porte sur μ, σ est appelé paramètre nuisible.

On introduit la moyenne et la variance de l'échantillon :

$$\overline{X} = \frac{1}{n}\sum X_i$$

$$S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2.$$

On montre que $\overline{X}$ suit une loi normale de moyenne μ et d'écart- type $\dfrac{\sigma}{\sqrt{n}}$, $\dfrac{S^2}{\sigma^2}$ suit une loi dite du khi-deux à $n-1$ degrés de liberté. Pour trouver un intervalle de confiance pour μ ou pour exécuter un test dont l'hypothèse nulle est par exemple $\mu = \mu_0$, on a besoin de la quantité $\dfrac{\overline{X}-\mu}{S}$ dont la loi de probabilité est connue : c'est la loi de Student à $n-1$ degrés de liberté indépendante de μ et de σ. Cela permet d'exécuter la procédure sans se préoccuper de σ inconnu et nuisible. Mais la loi de Student n'est pas au programme. La procédure enseignée aux élèves est de faire comme si $\dfrac{\overline{X}-\mu}{s}$ suivait une loi normale de moyenne nulle et d'écart-type 1 où s est la réalisation de S pour les observations faites. Pour être exact, **il importe dans l'énoncé de signaler que l'on fait cette approximation** qui n'est valide que si $n \geqslant 20$.

Là aussi, on observe trop souvent que les observations numériques figurant dans les énoncés sont telles que par exemple $\overline{X}$ et le μ supposé sont très près l'un de l'autre. Là encore, cela sent les exemples fabriqués, c'est trop beau pour être vrai.

OBJECTIF POURSUIVI

Il est indispensable pour que les anciens élèves des BTS puissent suivre avec profit les cours de la formation permanente que la formation initiale soit de qualité et donc que les professeurs ne répètent pas des erreurs parce que pour eux les sujets d'examen sont *a priori* sans tache.