Université de Nantes
UFR des Sciences et Techniques
Département de Mathématiques

2015/2016

TD STATISTIQUE DESCRIPTIVE

Frédéric Lavancier

Références

- Poly de cours :
http://www.math.sciences.univ-nantes.fr/~lavancie/enseignement/L3mathseco/CM-Stat_Desc.pdf
- "Statistique descriptive", A. Hamon et N. Jégou, dont de nombreux exercices sont extraits.

Exercice 1. Donner pour chacune des variables suivantes, son type et si cela est possible ses modalités.

1. état matrimonial
2. couleur des yeux
3. taux de cholestérol
4. âge
5. poids
6. région de naissance
7. mention au baccalauréat

Exercice 2.

1. Le tableau suivant donne les taux de chômage dans cinq bassins d'emploi :

Bassin d'emploi	Taux chômage (%)
A	$8,9$
B	$10,4$
C	$7,3$
D	$9,2$
E	$8,2$

Le taux de chômage moyen est :

- compris entre $7,3\%$ et $10,4\%$?
- $\frac{1}{5} \times (8,9\% + 10,4\% + 7,3\% + 9,2\% + 8,2\%)$ soit $8,8\%$?
- impossible à déterminer précisément avec ces seules données ?

2. Calculer le revenu annuel moyen d'un salarié à partir du tableau suivant :

Bassin d'emploi	Revenu annuel moyen net par salarié (€)	Nombre de salariés (en milliers)
A	18300	$54,5$
B	14600	$10,9$
C	21900	$10,3$
D	22100	$6,8$
E	16400	$3,1$

Exercice 3. *Effet de structure*
Les salariés d'une entreprise ont constaté que dans chacune des cinq catégories de salaires de l'entreprise, les revenus avaient baissé. Le responsable syndical se présente dans le bureau du directeur avec pour preuve le tableau ci-dessous :

	Année 2007		Année 2008	
	Effectifs	Revenus mensuels	Effectifs	Revenus mensuels
Catégorie 1	3	1000	2	900
Catégorie 2	4	1500	3	1400
Catégorie 3	6	2000	4	1900
Catégorie 4	4	2500	6	2400
Catégorie 5	3	3000	5	2900

Le directeur lui rétorque qu'au contraire, il dépense sans cesse plus pour les salaires.
Calculer le salaire moyen, médian et modal pour les années 2007 et 2008. Qui faut-il croire ?

Exercice 4. Dans un groupe de sept enfants dont deux sont des jumeaux, l'âge modal est 5 ans, l'âge médian est 6 ans, l'âge moyen est 7 ans et c'est l'âge des jumeaux. Quel est l'âge de chaque enfant ?

Exercice 5. On considère n observations $x_1, \ldots, x_n$ d'une variable quantitative X. On note $\bar{x}$ leur moyenne et σ leur écart-type. On construit la nouvelle variable Y en centrant et réduisant X, c'est à dire, pour $i = 1, \ldots, n$,

$$y_i = \frac{x_i - \bar{x}}{\sigma}.$$

Cette transformation est-elle toujours possible ? Calculer la moyenne et l'écart-type de Y.

Exercice 6. On dispose de la répartition de l'âge des employés de deux entreprises, résumée dans les tableaux suivants :

Entreprise 1 :

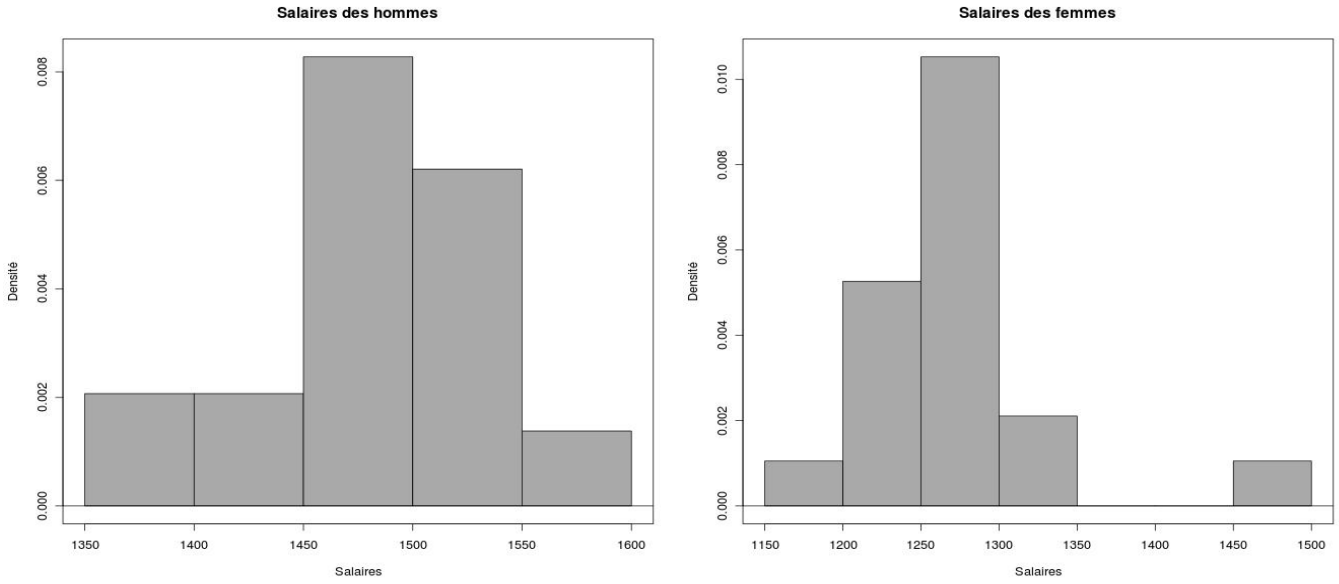
Classes d'âge	$[0;20[$	$[20;40[$	$[40;60[$	$[60;90[$
Effectifs observés	$n_1 = 6$	$n_2 = 5$	$n_3 = 5$	$n_4 = 4$

Entreprise 2 :

Classes d'âge	$[10;30[$	$[30;40[$	$[40;50[$	$[50;70[$	$[70;90[$
Effectifs observés	$n_1 = 7$	$n_2 = 4$	$n_3 = 4$	$n_4 = 3$	$n_5 = 2$

Construire l'histogramme de la distribution des âges pour chaque entreprise.

Exercice 7. On considère la distribution des salaires mensuels nets (en euros) des hommes et des femmes dans une entreprise. On précise que l'entreprise compte 30 hommes et 20 femmes.



1. Quelle est la classe modale pour le salaire des hommes ? Des femmes ?
2. Retrouver le nombre d'hommes et le nombre de femmes présents dans chacun des intervalles de salaires.
3. Représenter la distribution des salaires tous sexes confondus.

Exercice 8. On a mesuré la durée de vie (en milliers de kms) de vingt moteurs de deux marques différentes. Les résultats sont les suivants :

333	335	332	309	284	341	322	315	293	328
309	335	354	371	351	338	317	334	350	349

TABLE 1 – Durée de vie pour la marque A

334	383	356	385	345	333	337	404	292	392
376	242	320	314	364	253	274	499	372	395

TABLE 2 – Durée de vie pour la marque B

1. Donner le minimum, le maximum, la médiane, les quartiles, l'écart inter-quartile et le quantile à 90% pour chaque série d'observations.
2. Représenter la boite de dispersion des durées de vie pour chaque marque.

Exercice 9. On rappelle les deux résultats suivants :

(i) Soit un échantillon $x_1, \ldots, x_n$ de moyenne $\bar{x}_n$ et d'écart-type σ. Soit $\alpha \geq 1$ et $\mathcal{I} = [\bar{x}_n - \alpha\,\sigma,\ \bar{x}_n + \alpha\,\sigma]$, alors

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \in \mathcal{I}} \geq 1 - \frac{1}{\alpha^2}$$

(ii) Inégalité de Bienaymé-Chebyshev : soit X une variable aléatoire de carré intégrable et $\alpha > 0$, alors

$$P(|X - E(X)| \geq \alpha) \leq \frac{Var(X)}{\alpha^2}.$$

En quoi (i) est une conséquence de (ii) ?

Exercice 10. On considère n observations $x_1, \ldots, x_n$ d'une variable quantitative X.

1. Rappeler la définition de la moyenne $\bar{x}$, de l'écart-type σ et de l'écart moyen absolu EMA de X.
2. Justifier l'inégalité suivante : pour tout réel y,

$$\sigma^2 \leq \frac{1}{n}\sum_{i=1}^{n}(x_i - y)^2.$$

3. On note $c = (x_{(1)} + x_{(n)})/2$. Montrer que pour tout $i = 1, \ldots, n$,

$$|x_i - c| \leq \frac{1}{2}(x_{(1)} - x_{(n)}).$$

4. Déduire des deux questions précédentes que
$$\sigma \leq \frac{1}{2}(x_{(1)} - x_{(n)}).$$
5. Justifier que
$$EMA \leq \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|.$$
6. En utilisant l'inégalité de Schwartz, en déduire que
$$EMA \leq \sigma.$$

Exercice 11. On considère deux variables qualitatives X et Y ayant chacune deux modalités, X_1, X_2 et Y_1, Y_2.

1. On observe X et Y chez n individus. Comment est défini le tableau de contingence résumant les observations ?
2. Rappeler l'expression de la distance du chi-deux pour ce tableau. Quelles valeurs cette distance peut-elle prendre ?
3. On suppose que dans l'échantillon, il y a 20% d'individus ayant la modalité X_1 et 50% ayant la modalité Y_1. Dresser le tableau de contingence que l'on obtiendrait s'il y avait indépendance entre X et Y.
4. Donner un exemple de tableau de contingence correspondant à une dépendance totale entre X et Y, c'est à dire pour lequel la distance du chi-deux prend sa valeur maximale.

Exercice 12. On considère l'observation de deux variables quantitatives X et Y sur n individus : $(x_1, y_1), \ldots, (x_n, y_n)$.

1. Rappeler la définition de la covariance entre les observations de X et de Y. Proposer une écriture alternative et la prouver.
2. Donner la définition de la corrélation r entre X et Y. Dans quel intervalle varie-t-elle ? Pourquoi ?
3. Dans quelles situations a-t-on $r = 1$? $r = -1$?

Exercice 13. On considère l'observation de deux variables quantitatives X et Y sur n individus : $(x_1, y_1), \ldots, (x_n, y_n)$.

1. Donner l'équation de la droite des moindres carrés de Y en fonction de X.
2. Comment sont définis les résidus de l'ajustement précédent ? Que vaut leur moyenne ?
3. Donner l'équation de la droite des moindres carrés de Y en fonction de X passant par l'origine.
4. Comment sont définis les résidus de l'ajustement précédent ? Que vaut leur moyenne ?

Exercice 14. Soit s_t, $t \in \mathbb{N}$, une série temporelle centrée et périodique de période T, i.e. $s_{t+T} = s_t$ pour tout t, et $\sum_{i=1}^{T} s_i = 0$.

1. On considère la moyenne mobile centrée d'orde T : M_T. Rappeler sa définition.
2. Montrer que $M_T(s_t) = 0$ pour tout t.
3. On considère la série temporelle $x_t = at + b + s_t$ où a et b sont deux réels. Que vaut $M_T(x_t)$? Quel est l'intérêt de cette transformation ?
4. Discuter de l'intérêt d'appliquer la moyenne mobile centrée d'ordre $T + 1$ à x_t.
5. Discuter de l'intérêt d'appliquer la moyenne mobile centrée d'ordre kT à x_t, où $k \in \mathbb{N}^*$.
6. A quelle série correspond $\tilde{x}_t = x_t - M_T(x_t)$? En déduire une façon d'estimer le profil saisonnier de x_t.
7. Que deviennent les méthodes de filtrage précédent si x_t est entaché d'un bruit aléatoire, i.e. $x_t = at + b + s_t + r_t$ où r_t est une variable aléatoire centrée ?