Psy1004 – Section 2:

Statistiques descriptives

Plan du cours:
- Varia
- 2.0: Vers une synthèse de données
- 2.1: Tendance centrale
- 2.2: Variabilité
- 2.3: Asymétrie
- 2.4: Kurtose
- 2.5: Erreur type
- Survol de SPSS

Disponible sur: http://mapageweb.umontreal.ca/cousined/home/course/PSY1004

Varia

- La page web du cours (ou en dépôt à la bibliothèque EPC)
  http://mapageweb.umontreal.ca/cousined/home/course/PSY1004/
- Le TP1 est maintenant disponible sur ce site.
  - à remettre dans deux semaines;
  - à faire seul, à deux, ou exceptionnellement à trois.
- La séance de formation pratique sur SPSS aura lieu ce jeudi, le 18 septembre, au A332, à 13h00:
  - Formation pour apprendre à se débrouiller avec un ordinateur;
- Important pour le prochain cours :
  - Imprimer les tables statistiques disponibles sur le site web.
- Questions sur la section 1?

2.0: Vers une synthèse de données (1/2)

Soit les trois échantillons:
- **X** = {86, 87, 88, 92, 93, 95, 96, 96, 97, 97, 98, 99, 101, 101, 102, 102, 102, 103, 103, 103, 103, 104, 104, 105, 107, 108, 108, 110, 113, 114}
- **Y** = {91, 91, 92, 92, 93, 93, 93, 94, 94, 95, 95, 96, 96, 97, 98, 98, 98, 98, 98, 98, 100, 101, 104, 106, 107, 114, 118, 119, 121, 131}
- **Z** = {87, 88, 89, 89, 90, 90, 91, 91, 92, 93, 94, 94, 95, 96, 96, 96, 97, 97, 99, 99, 100, 100, 100, 101, 101, 103, 104, 107, 107, 111}

Tracez le graphique des fréquences d'un de ces échantillons en utilisant des classes de tailles 5 partant à 75 (i.e. de 75 à 80, de 80 à 85, de 85 à 90, etc.).

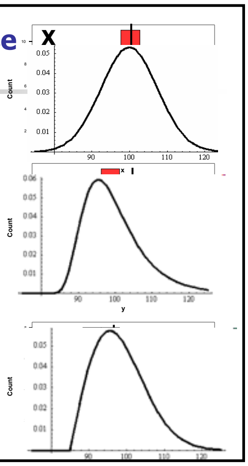
2.0: Vers une synthèse de données (2/2)

On remarque que:

- Le centre de gravité est à peu près le même dans les trois cas…
- La variabilité est aussi comparable…
- Certaines distributions de valeurs sont asymétriques…

Parfois, on utilise une courbe idéale pour représenter ces échantillons:

Des statistiques informatives doivent quantifier ces aspects.



2.0: Note sur la nomenclature (1/2)

- On note un échantillon avec une lettre majuscule de la fin de l'alphabet, tel **X**, **Y** ou **Z** (en gras);
- On note une statistique (une description) sur un échantillon par la lettre de l'échantillon avec un signe par dessus: $\overset{?}{\mathbf{X}}$
- Par exemple:
  - La moyenne des échantillons est: $\bar{\mathbf{X}}, \bar{\mathbf{Y}}, \bar{\mathbf{Z}}$
  - L'écart type (qu'on verra plus loin): $\overleftrightarrow{\mathbf{X}}, \overleftrightarrow{\mathbf{Y}}, \overleftrightarrow{\mathbf{Z}}$
  - D'autres symboles sont possibles: $\mathring{\mathbf{X}}, \tilde{\mathbf{Y}}, \overset{\frown}{\mathbf{Z}}$

2.0: Note sur la nomenclature (2/2)

- Dans le passé, il y avait confusion pour le signe d'écart type, s vs. S ou encore s_{n-1} vs. s_n.
- Pour éviter cela, $\overleftrightarrow{\mathbf{X}}$
- Comme il y a deux façon de calculer l'écart type, j'utilise une étiquette: $\overleftrightarrow{X}$

divise par *n*

- Nous remplaçons donc: par:

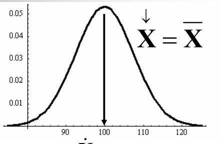
Est le plus utilisé

s s_{n-1} σ_{n-1}	${}_{n-1}\overleftrightarrow{\mathbf{X}}$
S s_n σ_n	${}_{n}\overleftrightarrow{\mathbf{X}}$

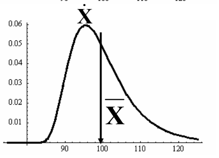
2.1: Tendance centrale (1/3)

"Où est la distribution?"

(le centre de gravité de la distribution, donné par la moyenne arithmétique); facile si symétrique ➔



Plus dur si asymétrique ➔
Pour une distribution asymétrique, on peut utiliser le mode $\dot{\mathbf{X}}$ (bruyant)



Pour pondérer moins les valeurs d'un extrême (quand $\overline{\mathbf{X}}$ est très loin du mode $\dot{\mathbf{X}}$), utiliser:

- la médiane $\breve{\mathbf{X}}$ ou
- la moyenne géométrique $\mathring{\mathbf{X}}$.

2.1: Tendance centrale (2/3)

- Les autres mesures de la tendance centrale sont rarement utilisée en psychologie (moyenne harmonique, géométrique, médiane, mode);
- La médiane est utile quand on sait que la distribution est **très** asymétrique (telle les revenues des ménages);
- La moyenne harmonique va être utilisée dans un test à la section 9.

2.1: Tendance centrale (3/3)

Soit les trois échantillons:

- **X** = {86, 87, 88, 92, 93, 95, 96, 96, 97, 97, 98, 99, 101, 101, 102, 102, 102, 103, 103, 103, 103, 104, 104, 105, 107, 108, 108, 110, 113, 114}
- **Y** = {91, 91, 92, 92, 93, 93, 93, 94, 94, 95, 95, 96, 96, 97, 98, 98, 98, 98, 98, 98, 100, 101, 104, 106, 107, 114, 118, 119, 121, 131}
- **Z** = {87, 88, 89, 89, 90, 90, 91, 91, 92, 93, 94, 94, 95, 96, 96, 96, 97, 97, 99, 99, 100, 100, 100, 101, 101, 103, 104, 107, 107, 111}

Calculer la moyenne arithmétique d'un de ces échantillons.

Solution: $\overline{\mathbf{X}} = 100.6$

$\overline{\mathbf{Y}} = 100.7$

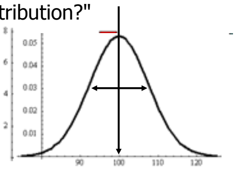
$\overline{\mathbf{Z}} = 96.6$

2.2: Variabilité (1/3)

"Quelle est l'étendue de la distribution?"

- L'écart moyen au centre➔ mais! vaut toujours zéro…

$$\boxed{\begin{aligned}\frac{1}{n}\sum_i\left(\mathbf{X}_i-\overline{\mathbf{X}}\right) &= \frac{1}{n}\sum_i \mathbf{X}_i-\frac{1}{n}\sum_i\overline{\mathbf{X}}\\ &=\overline{\mathbf{X}}-\frac{1}{n}n\overline{\mathbf{X}}\\ &=\overline{\mathbf{X}}-\overline{\mathbf{X}}=0\end{aligned}}$$



- L'écart **carré** moyen au centre: ${}_{n}\overleftrightarrow{\mathbf{X}}^{2}=\frac{1}{n}\sum_i\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)^2$
- L'écart **carré** moyen au centre corrigé pour le biais : ${}_{n-1}\overleftrightarrow{\mathbf{X}}^{2}=\frac{n}{n-1}\,{}_{n}\overleftrightarrow{\mathbf{X}}^{2}=\frac{1}{n-1}\sum_i\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)^2$

2.2: Variabilité (2/3)

- La racine carrée de la variance s'appelle l'écart type.
- Signification de l'écart type: "En prenant une donnée au hasard, elle a toute les chances d'être à ± un écart type de la moyenne des données."
- Autrement dit, l'écart type est l'écart typique entre une donnée et sa moyenne.

2.2: Variabilité (3/3)

Soit les trois échantillons:

- **X** = {86, 87, 88, 92, 93, 95, 96, 96, 97, 97, 98, 99, 101, 101, 102, 102, 102, 103, 103, 103, 103, 104, 104, 105, 107, 108, 108, 110, 113, 114}
- **Y** = {91, 91, 92, 92, 93, 93, 93, 94, 94, 95, 95, 96, 96, 97, 98, 98, 98, 98, 98, 98, 100, 101, 104, 106, 107, 114, 118, 119, 121, 131}
- **Z** = {87, 88, 89, 89, 90, 90, 91, 91, 92, 93, 94, 94, 95, 96, 96, 96, 97, 97, 99, 99, 100, 100, 100, 101, 101, 103, 104, 107, 107, 111}

Calculer l'écart type non biaisé d'un de ces échantillons.

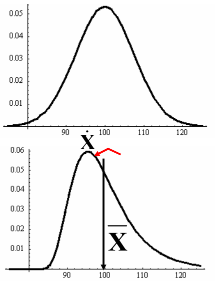
Rappel: $\overline{\mathbf{X}}=100.6$ Solution: $\overleftrightarrow{\mathbf{X}}=7.0$

$\overline{\mathbf{Y}}=100.7$ $\overleftrightarrow{\mathbf{Y}}=10.2$

$\overline{\mathbf{Z}}=96.6$ $\overleftrightarrow{\mathbf{Z}}=6.1$

2.3: Asymétrie (1/2)

- Si la moyenne correspond à la médiane ou au mode, la distribution n'est pas asymétrique ➔
- Si la moyenne diffère du mode, la distribution est asymétrique:
  - Négative si le mode est à droite de la moyenne
  - Positive si le mode est à gauche de la moyenne ➔



Par ex.: Revenu, Temps de réponses, des mesures qui peuvent être proches de zéro, mais non négatives.

- Pour quantifier l'asymétrie (*skewness*):

$$\overset{\nwarrow}{\mathbf{X}} = \frac{1}{n}\frac{\sum_i \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^3}{{}_n\overset{\leftrightarrow}{\mathbf{X}}^3}$$

2.4: Kurtose (1/2)

Quelle est l'épaisseur des "queues" par rapport au centre?

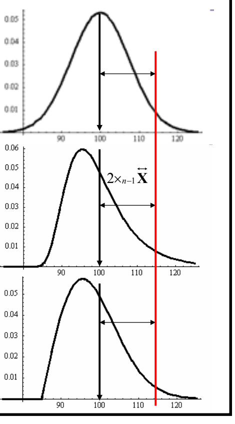
Si les queues sont plus importantes: Kurtose > 0

Ci-contre, les kurtoses sont de 3, 9.0, et 3.2 resp.

Pour la distribution Normale (qui sert de référence), la kurtose = 3.

Pour calculer la kurtose:

$$\overset{\swarrow\searrow}{\mathbf{X}} = \frac{1}{n}\frac{\sum_i \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^4}{{}_n\overset{\leftrightarrow}{\mathbf{X}}^4}$$



2.3 & 2.4: *Skewness* et Kurtose (2/2)

Soit les trois échantillons:

- **X** = {86, 87, 88, 92, 93, 95, 96, 96, 97, 97, 98, 99, 101, 101, 102, 102, 102, 103, 103, 103, 103, 104, 104, 105, 107, 108, 108, 110, 113, 114}
- **Y** = {91, 91, 92, 92, 93, 93, 93, 94, 94, 95, 95, 96, 96, 97, 98, 98, 98, 98, 98, 98, 100, 101, 104, 106, 107, 114, 118, 119, 121, 131}
- **Z** = {87, 88, 89, 89, 90, 90, 91, 91, 92, 93, 94, 94, 95, 96, 96, 96, 97, 97, 99, 99, 100, 100, 100, 101, 101, 103, 104, 107, 107, 111}

Calculer l'asymétrie et la kurtose d'un de ces échantillons.

Rappel: $\overline{\mathbf{X}} = 100.6$ $\overset{\leftrightarrow}{\mathbf{X}} = 7.0$ Solution: $\overset{\nwarrow}{\mathbf{X}} = -0.26$ $\overset{\swarrow\searrow}{\mathbf{X}} = 2.71$

$\overline{\mathbf{Y}} = 100.7$ $\overset{\leftrightarrow}{\mathbf{Y}} = 10.2$ $\overset{\nwarrow}{\mathbf{Y}} = 1.48$ $\overset{\swarrow\searrow}{\mathbf{Y}} = 4.33$

$\overline{\mathbf{Z}} = 96.6$ $\overset{\leftrightarrow}{\mathbf{Z}} = 6.1$ $\overset{\nwarrow}{\mathbf{Z}} = 0.43$ $\overset{\swarrow\searrow}{\mathbf{Z}} = 2.53$

2.5: Erreur type (1/3)

Signification de l'écart type: "En prenant une donnée au hasard, elle a *toute les chances* d'être à ± un écart type de la moyenne des données."

Supposons que vous soyez très riche et collectiez un très grand nombre *M* d'échantillons indépendants, vous obtenez un ensemble de moyennes $\left\{\overline{\mathbf{X}}_1, \overline{\mathbf{X}}_2, \ldots \overline{\mathbf{X}}_M\right\}$.

Évidemment, si *M* >>, la moyenne des moyennes est la vraie moyenne de la population (μ)

2.5: Erreur type (2/3)

Nous voudrions alors connaître l'erreur type: "En prenant une moyenne au hasard, elle a *toute les chances* d'être à ± un erreur type de la moyenne des moyennes."

Appelons erreur type (*Standard error* parfois traduit *erreur standardisée*):

$$SE_{\overline{\mathbf{X}}} = \frac{{}_{n-1}\overleftrightarrow{\mathbf{X}}}{\sqrt{n}}$$

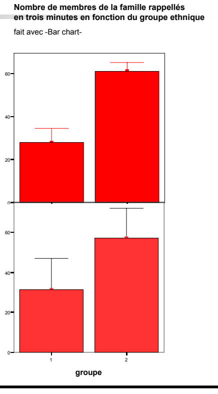
Nous avons alors *toutes les chances* que:

$$\overline{\mathbf{X}} - SE_{\overline{\mathbf{X}}} < \mu < \overline{\mathbf{X}} + SE_{\overline{\mathbf{X}}} \qquad \left(\overline{\mathbf{X}} \pm SE_{\overline{\mathbf{X}}}\right)$$

$$\left|\mu - \overline{\mathbf{X}}\right| < SE_{\overline{\mathbf{X}}}$$

2.5: Erreur type (3/3)

- L'erreur type devrait toujours être présent sur un graphe des moyennes
  - (peut être fait automatiquement avec SPSS, à la main avec EXCEL):
  - Elle donne une indication de la variabilité de chaque groupe;
  - Permet de savoir si la différence entre deux moyennes est importante;
  - A un lien important avec plusieurs tests sur la moyenne (cf. section 5).



Survol de SPSS

- Comment entrer des données à la main;
  - Démarrer SPSS et l'utiliser comme un chiffrier
- Comment exécuter une analyse sur ces données;
  - Ouvrir une fenêtre de syntaxe, écrire une commande, et l'exécuter
- Comment ouvrir un fichier de données déjà existant.
  - Ouvrir une fenêtre de syntaxe (ou utiliser celle existante), écrire la commande d'ouverture de fichier et l'exécuter