3C - Mathématiques
Statistique inférentielle

A. Ridard

Gymnase Yverdon

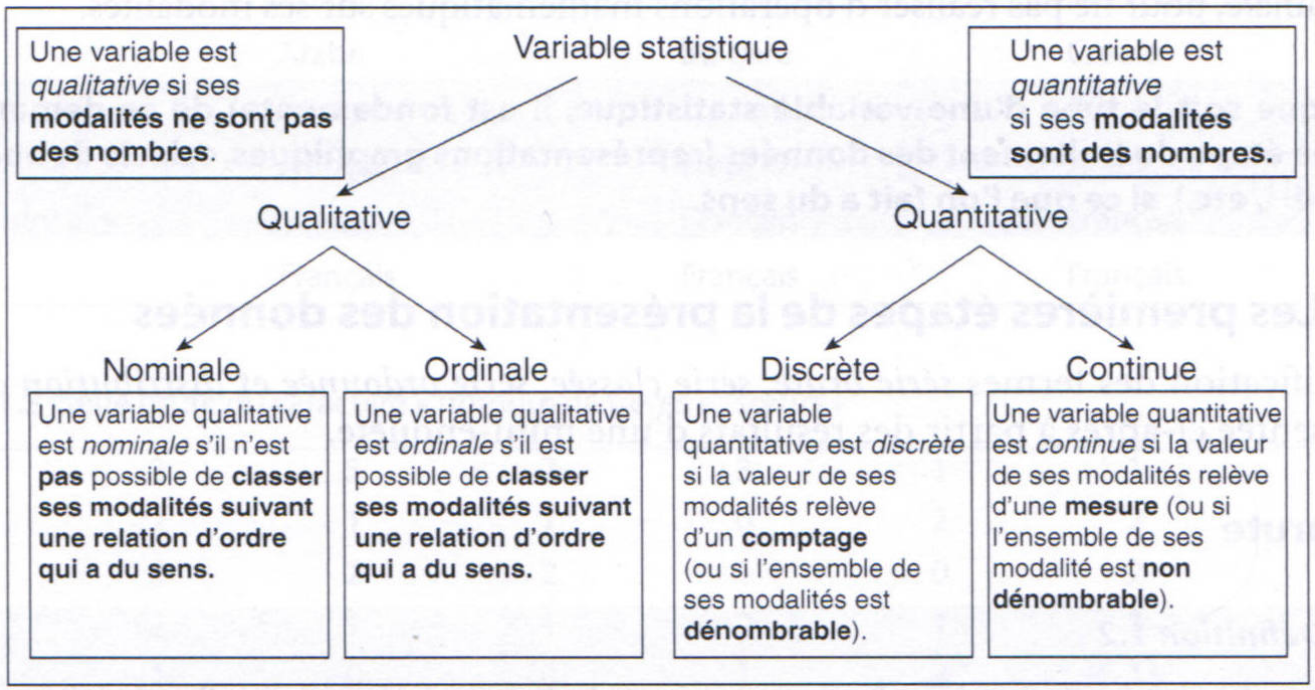
A propos de ce document

- Pour naviguer dans le document, vous pouvez utiliser :
  - le menu (en haut à gauche)
  - l'icône en dessous du logo GyYv
  - les différents liens
- Pour signaler une erreur, vous pouvez envoyer un message à l'adresse suivante : anthony.ridard@eduvaud.ch

Gymnase Yverdon

Commençons par rappeler quelques notions :

- population, individu
- variable **statistique**



Dans ce cours, nous nous limiterons aux variables quantitatives.

- données sous différentes formes :
  1. données brutes (individu par individu) de la forme :

     $$x_1,\ x_2,\ \ldots,\ x_n$$

  2. données regroupées par valeur, dans le cas discret, de la forme :

x_1	x_2	$\ldots$	x_p
n_1	n_2	$\ldots$	n_p

  3. données regroupées par classe, dans le cas continu, de la forme :

$[e_1, e_2[$	$[e_2, e_3[$	$\ldots$	$[e_p, e_{p+1}[$
n_1	n_2	$\ldots$	n_p

- effectif, fréquence et représentations graphiques associées :
  2. données regroupées par valeur : diagramme en bâtons
  3. données regroupées par classe : histogramme

Gymnase Yverdon

- indicateurs de position (tendance centrale) : moyenne et médiane
- indicateurs de dispersion : variance/écart-type et écart inter-quartile

En notant $n = \sum_{i=1}^{p} n_i$ l'effectif total, $f_i = \frac{n_i}{n}$ la fréquence associée à n_i et $c_i = \frac{e_i + e_{i+1}}{2}$ le centre de la classe $[e_i, e_{i+1}[$, on a les formules suivantes : .

Forme des données	1	2	3
Moyenne $\overline{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\frac{1}{n}\sum_{i=1}^{p} n_i x_i$	$\frac{1}{n}\sum_{i=1}^{p} n_i c_i$
Variance s^2	$\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$	$\frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \overline{x})^2$	$\frac{1}{n}\sum_{i=1}^{p} n_i (c_i - \overline{x})^2$

Gymnase Yverdon

La variance est notée s^2 et non σ^2 pour plus de cohérence par la suite.

Pour la calculer, on utilisera plutôt la formule suivante :

$$s^2 = \text{Moyenne des carrés - Carré de la moyenne}$$

Forme des données	1	2	3
Variance s^2	$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2$	$\left(\frac{1}{n}\sum_{i=1}^{p} n_i x_i^2\right) - \overline{x}^2$	$\left(\frac{1}{n}\sum_{i=1}^{p} n_i c_i^2\right) - \overline{x}^2$

L'indicateur de dispersion exprimé dans la même unité que les données est :

$$\text{Écart-type} = \sqrt{\text{Variance}}$$

$$s = \sqrt{s^2}$$

Gymnase Yverdon

Considérons la taille (en cm) de 50 élèves de 2C du gymnase d'Yverdon :

```
[159. 170. 164. 163. 167. 179. 158. 176. 184. 164. 157. 154. 179. 157.
 164. 161. 165. 168. 173. 174. 159. 168. 157. 168. 172. 162. 169. 163.
 161. 164. 173. 160. 172. 176. 157. 185. 163. 167. 162. 170. 167. 162.
 175. 148. 158. 147. 171. 167. 160. 173.]
```

1. Identifier la population, la variable statistique et la forme des données.
2. Construire le tableau des effectifs et des fréquences après avoir regroupé les données par classe [a].
3. Représenter l'histogramme des fréquences et le polygone des fréquences.
4. Ajouter les fréquences cumulées croissantes et décroissantes.
5. Représenter les fréquences cumulées croissantes et décroissantes.
6. En déduire une valeur approchée de la médiane.
7. Calculer la médiane et l'écart interquartile.
8. Représenter la boîte à moustaches.
9. Calculer la moyenne et l'écart-type.

a. On prendra des classes de largeur 10 en commençant par [140 ; 150 [

Gymnase Yverdon

Gymnase Yverdon

La Statistique **descriptive** permet d'**étudier une population** toute entière d'individus (selon un ou plusieurs caractères).

La Statistique **inférentielle** permet de **déduire des informations sur une population** de taille N **à partir d'un échantillon** « représentatif » de taille n.

En général, nous n'avons pas accès à la population toute entière, mais seulement à une partie (pensez aux sondages lors des élections). Si cette partie est représentative de la population toute entière, alors on peut espérer en déduire des informations (inconnues) concernant la population.

Gymnase Yverdon

Intéressons-nous, par exemple, à la taille X de tous les élèves des gymnases du canton de Vaud. Cette variable possède une moyenne μ et un écart-type σ. En théorie, ces deux paramètres pourraient être calculés si l'on connaissait la taille de tous les élèves, mais en pratique cela n'est pas « réalisable ».

Nous voyons donc ces deux paramètres comme des valeurs théoriques que l'on ne connaît pas, et que l'on ne connaîtra jamais.

Si nous avons accès à la taille des élèves pour un échantillon [a], comment peut-on approcher la moyenne (théorique) μ ?

a. par exemple, les 50 élèves de l'exercice précédent

Il suffit [1] de calculer la moyenne (empirique) $\overline{x}$ de l'échantillon.

On dit alors que $\overline{x}$ est une **estimation ponctuelle** de μ.

1. Cela est légitimé par la loi forte des grands nombres

Les fonctions statistiques de Python fournissent les résultats suivants :

```
[159. 170. 164. 163. 167. 179. 158. 176. 184. 164. 157. 154. 179. 157.
 164. 161. 165. 168. 173. 174. 159. 168. 157. 168. 172. 162. 169. 163.
 161. 164. 173. 160. 172. 176. 157. 185. 163. 167. 162. 170. 167. 162.
 175. 148. 158. 147. 171. 167. 160. 173.]

taille : 50
min : 147.0
max : 185.0
moyenne : 165.84
écart-type : 8.07554332537446
```

Les valeurs de la moyenne et de l'écart-type ne coïncident pas avec celles obtenues dans l'exercice, savez-vous pourquoi ?

Mais avec un autre échantillon [2], on obtient une autre estimation ponctuelle.

```
[169. 172. 163. 177. 160. 167. 173. 155. 148. 161. 162. 164. 155. 160.
 171. 160. 162. 158. 178. 163. 170. 170. 172. 169. 153. 158. 151. 159.
 160. 162. 168. 169. 166. 163. 162. 164. 179. 158. 161. 162. 163. 175.
 177. 152. 160. 171. 164. 161. 153. 159.]

taille : 50
min : 148.0
max : 179.0
moyenne : 163.78
écart-type : 7.2423476856610 52
```

Comment peut-on résoudre ce défaut ?

En cherchant un intervalle qui contient μ avec une forte probabilité, par exemple 95%, plutôt qu'une valeur approchée de μ.

On parle alors d'**estimation par intervalle de confiance**.

2. par exemple, 50 élèves de 2M du gymnase d'Yverdon

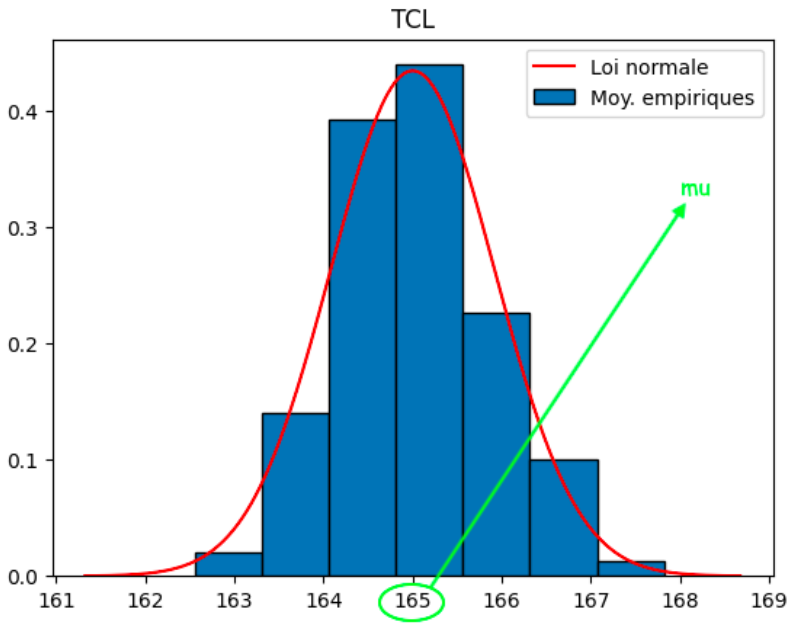
Si on dispose de 200 échantillons de taille 50, on a 200 estimations ponctuelles $\overline{x}$, toutes « proches de » μ.

```
[164.02 164.18 165.36 164.7  165.76 165.7  165.24 165.   164.5  163.4
 165.12 165.02 166.28 164.92 166.08 166.42 163.86 163.46 164.02 165.86
 165.66 165.02 164.7  165.02 165.52 164.5  163.34 165.5  165.6  164.48
 164.74 167.82 164.38 165.62 165.18 164.1  165.08 166.7  164.58 166.98
 165.36 164.58 164.7  165.42 164.68 165.22 164.36 163.8  164.62 164.66
 164.84 163.9  165.32 164.86 165.08 163.74 166.76 163.96 164.78 164.84
 166.56 166.88 165.4  165.28 165.1  164.56 164.54 164.32 164.44 166.44
 165.82 165.06 165.12 164.94 165.28 163.86 165.16 166.22 164.68 164.46
 165.38 162.56 164.44 166.2  164.52 165.22 164.46 165.   165.16 164.56
 165.34 164.24 165.36 165.12 164.98 163.96 164.48 164.92 164.62 166.
 164.04 165.98 165.56 166.88 165.48 164.56 165.86 166.28 165.98 164.5
 165.82 165.16 164.52 165.74 167.1  163.2  166.3  166.18 164.76 164.26
 164.94 165.3  164.92 165.42 164.2  166.6  166.12 165.84 165.5  165.78
 166.96 164.62 165.88 165.84 165.84 163.8  164.68 166.5  165.64 164.68
 165.06 165.92 165.06 164.64 166.02 164.68 165.36 165.38 163.62 166.16
 163.56 162.94 163.58 164.58 166.5  164.22 164.86 165.46 165.48 165.7
 166.54 164.36 166.46 164.02 164.54 165.46 164.46 164.98 164.72 164.94
 165.04 164.02 165.   163.92 165.44 164.44 164.9  164.7  165.28 164.76
 165.88 165.24 164.1  165.22 165.84 164.48 164.22 164.26 165.54 163.8
 165.7  164.46 164.2  164.12 164.84 165.16 164.46 164.5  166.46 165.38]
```

Mais comment se répartissent-elles ?

Gymnase Yverdon

En fait, la théorie des probabilités [3] nous assure que l'histogramme de ces moyennes $\overline{x}$ a toujours une forme de « cloche » :



Cette courbe rouge, qui modélise le comportement de la moyenne empirique, permet de calculer la probabilité pour que la moyenne empirique se trouve dans $[a\ ;\ b]$.
Il suffit pour cela de calculer l'aire sous la courbe entre les droites $x = a$ et $x = b$.

3. plus précisément, le théorème central limite

Ces aires approchées par des méthodes numériques[4] sont répertoriées dans une table, celle de la gaussienne[5] centrée réduite. Pour s'y ramener, on considère les $\frac{\overline{x}-\mu}{\frac{s^*}{\sqrt{n}}}$, plutôt que les $\overline{x}$, où s^* est l'écart-type corrigé défini par :

$$s^* = \sqrt{s^{*2}} = \sqrt{\frac{n}{n-1}s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

L'écart-type corrigé est noté s^* et non S pour plus de cohérence !

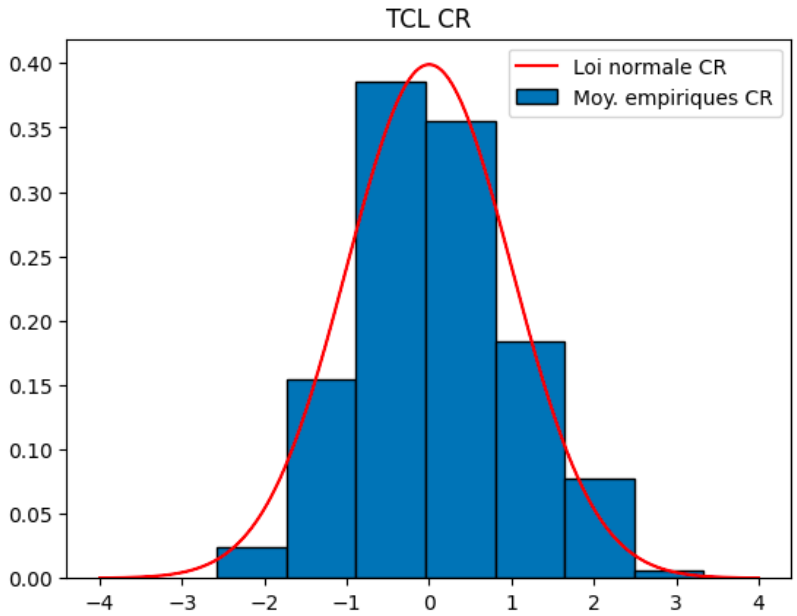
4. car impossibles à calculer de manière exacte
5. aussi appelée loi normale

L'opération de centrage et réduction réalisée sur les 200 moyennes $\overline{x}$ fournit :

```
[-1.14 -0.85  0.42 -0.35  0.86  0.87  0.24  0.   -0.57 -1.66  0.13  0.02
  1.4  -0.09  1.09  1.43 -1.46 -1.82 -1.03  0.88  0.71  0.02 -0.29  0.02
  0.49 -0.57 -1.71  0.56  0.5  -0.65 -0.26  3.33 -0.64  0.6   0.21 -1.21
  0.09  1.9  -0.53  2.12  0.43 -0.52 -0.35  0.42 -0.37  0.23 -0.64 -1.19
 -0.41 -0.33 -0.17 -1.17  0.35 -0.15  0.09 -1.47  2.   -0.99 -0.23 -0.17
  1.77  2.21  0.49  0.3   0.13 -0.43 -0.55 -0.69 -0.63  1.88  0.91  0.07
  0.14 -0.07  0.32 -1.21  0.17  1.41 -0.37 -0.66  0.41 -2.57 -0.7   1.27
 -0.44  0.23 -0.62  0.    0.18 -0.47  0.39 -0.88  0.43  0.14 -0.02 -1.12
 -0.5  -0.1  -0.4   1.39 -1.22  1.05  0.62  2.08  0.53 -0.54  0.77  1.43
  1.12 -0.58  0.88  0.17 -0.46  0.89  2.2  -1.91  1.59  1.29 -0.24 -0.85
 -0.07  0.29 -0.09  0.5  -0.93  1.91  1.09  0.98  0.51  0.67  2.27 -0.38
  0.98  0.87  0.99 -1.33 -0.37  1.47  0.71 -0.31  0.06  1.22  0.07 -0.42
  1.2  -0.36  0.49  0.38 -1.41  1.06 -1.64 -2.25 -1.66 -0.46  1.68 -1.09
 -0.14  0.56  0.5   0.88  1.71 -0.68  1.44 -0.97 -0.64  0.51 -0.48 -0.02
 -0.35 -0.08  0.05 -0.97  0.   -0.99  0.44 -0.77 -0.11 -0.32  0.29 -0.27
  0.87  0.24 -1.08  0.23  1.13 -0.54 -0.86 -1.05  0.65 -1.11  0.8  -0.58
 -0.91 -0.82 -0.17  0.15 -0.66 -0.46  1.75  0.39]
```

Gymnase Yverdon

Ces 200 moyennes centrées réduites $\frac{\overline{x}-\mu}{\frac{s^*}{\sqrt{n}}}$ ont pour histogramme :



Un extrait de la table[6] est disponible, dans le formulaire, comme suit :

Niveau de confiance	90%	95%	98%	99%
z	1,64	1,96	2,33	2,58

6. de la gaussienne centrée réduite

On en déduit qu'il y a 95% de chances pour que :

$$-1,96 < \frac{\overline{x} - \mu}{\frac{s^*}{\sqrt{n}}} < 1,96$$

ou encore :

$$\mu - 1,96\frac{s^*}{\sqrt{n}} < \overline{x} < \mu + 1,96\frac{s^*}{\sqrt{n}}$$

ce qui équivaut à :

$$\boxed{\overline{x} - 1,96\frac{s^*}{\sqrt{n}} < \mu < \overline{x} + 1,96\frac{s^*}{\sqrt{n}}}$$

1. Estimer μ [a] par intervalle de confiance à l'aide du premier échantillon [b] :
   - au niveau 95%
   - au niveau 90%
2. Que se passe-t-il lorsque l'on diminue la confiance ?

---

a. Taille moyenne de tous les élèves des gymnases du canton de Vaud

b. Celui des 50 élèves de 2C du gymnase d'Yverdon

Intéressons-nous maintenant à la proportion π de gauchers parmi tous les élèves des gymnases du canton de Vaud. Là encore, en théorie, ce paramètre pourrait être calculé si l'on connaissait cette caractéristique pour tous les élèves, mais en pratique cela n'est pas « réalisable ».

Nous voyons donc ce paramètre comme une valeur théorique que l'on ne connaît pas, et que l'on ne connaîtra jamais.

Si nous avons accès à cette caractéristique pour un échantillon [a], comment peut-on approcher la proportion (théorique) π ?

a. par exemple, le premier échantillon constitué de 50 élèves de 2C du gymnase d'Yverdon

Il suffit de calculer la proportion (empirique) c'est à dire la fréquence f de l'échantillon.

On dit alors que f est une **estimation ponctuelle** de π.

Les fonctions statistiques de Python fournissent les résultats suivants :

```
[0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0
 1 0 0 0 0 0 0 0 1 0 0 0 0]

taille : 50
fréquence : 0.16
```

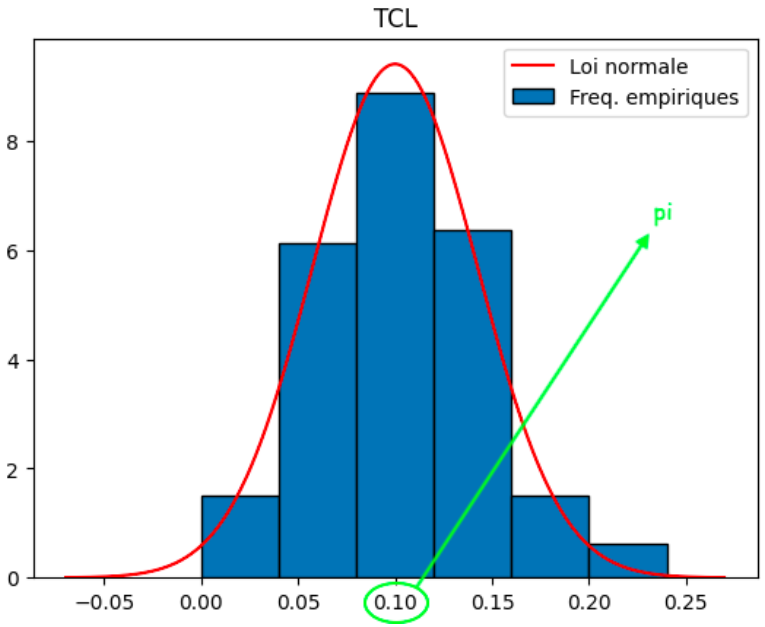
Gymnase Yverdon

Si on dispose de 200 échantillons de taille 50, on a 200 estimations ponctuelles f, toutes « proches de » π.

```
[0.1  0.22 0.16 0.1  0.08 0.18 0.04 0.14 0.06 0.1  0.06 0.08 0.08 0.14
 0.1  0.12 0.1  0.14 0.08 0.02 0.1  0.08 0.06 0.12 0.16 0.12 0.08 0.18
 0.1  0.06 0.04 0.16 0.1  0.04 0.16 0.04 0.04 0.12 0.14 0.12 0.1  0.1
 0.08 0.14 0.12 0.04 0.08 0.06 0.02 0.1  0.08 0.1  0.1  0.12 0.1  0.1
 0.12 0.02 0.1  0.06 0.08 0.12 0.1  0.04 0.1  0.1  0.1  0.04 0.04 0.
 0.08 0.1  0.08 0.04 0.14 0.06 0.02 0.12 0.12 0.06 0.06 0.06 0.16 0.12
 0.14 0.18 0.08 0.1  0.08 0.14 0.1  0.14 0.08 0.04 0.06 0.12 0.06 0.1
 0.08 0.08 0.02 0.06 0.12 0.06 0.06 0.08 0.1  0.08 0.08 0.08 0.16 0.06
 0.12 0.14 0.04 0.1  0.1  0.18 0.12 0.1  0.12 0.04 0.04 0.06 0.12 0.12
 0.1  0.1  0.06 0.14 0.1  0.12 0.06 0.14 0.08 0.1  0.02 0.1  0.14 0.24
 0.04 0.06 0.1  0.06 0.14 0.06 0.02 0.14 0.02 0.12 0.14 0.12 0.12 0.1
 0.04 0.08 0.12 0.04 0.12 0.12 0.1  0.12 0.04 0.08 0.22 0.04 0.14 0.04
 0.12 0.08 0.16 0.14 0.08 0.14 0.02 0.08 0.02 0.06 0.1  0.14 0.1  0.
 0.06 0.04 0.06 0.08 0.22 0.08 0.08 0.04 0.12 0.1  0.1  0.16 0.1  0.06
 0.06 0.14 0.2  0.12]
```

Mais comment se répartissent-elles ?

Le théorème central limite nous assure, ici encore, que l'histogramme de ces fréquences f a toujours une forme de « cloche » :

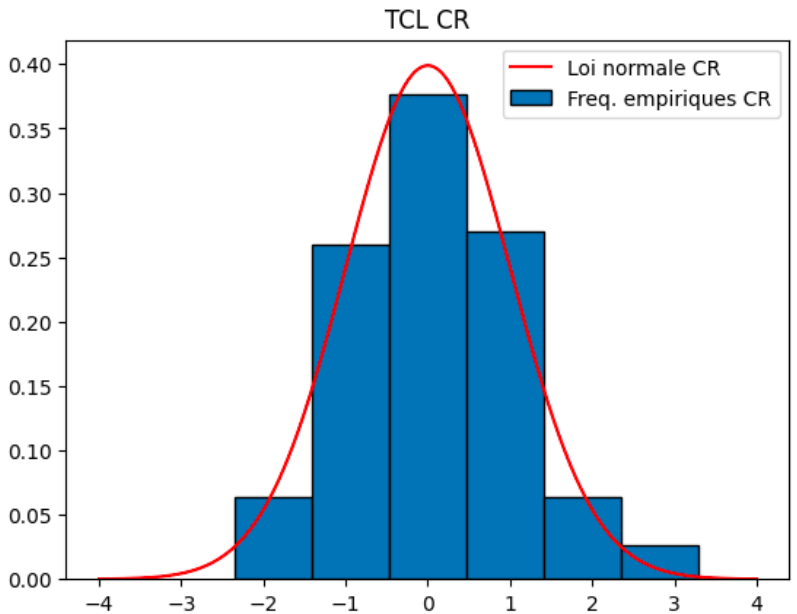


Pour se ramener à la table de la gaussienne centrée réduite, on considère les $\frac{f-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$, plutôt que les f.

L'opération de centrage et réduction réalisée sur les 200 fréquences f fournit :

```
[ 0.    2.83  1.41  0.   -0.47  1.89 -1.41  0.94 -0.94  0.   -0.94 -0.47
 -0.47  0.94  0.    0.47  0.    0.94 -0.47 -1.89  0.   -0.47 -0.94  0.47
  1.41  0.47 -0.47  1.89  0.   -0.94 -1.41  1.41  0.   -1.41  1.41 -1.41
 -1.41  0.47  0.94  0.47  0.    0.   -0.47  0.94  0.47 -1.41 -0.47 -0.94
 -1.89  0.   -0.47  0.    0.    0.47  0.    0.    0.47 -1.89  0.   -0.94
 -0.47  0.47  0.   -1.41  0.    0.    0.   -1.41 -1.41 -2.36 -0.47  0.
 -0.47 -1.41  0.94 -0.94 -1.89  0.47  0.47 -0.94 -0.94 -0.94  1.41  0.47
  0.94  1.89 -0.47  0.   -0.47  0.94  0.    0.94 -0.47 -1.41 -0.94  0.47
 -0.94  0.   -0.47 -0.47 -1.89 -0.94  0.47 -0.94 -0.94 -0.47  0.   -0.47
 -0.47 -0.47  1.41 -0.94  0.47  0.94 -1.41  0.    0.    1.89  0.47  0.
  0.47 -1.41 -1.41 -0.94  0.47  0.47  0.    0.   -0.94  0.94  0.    0.47
 -0.94  0.94 -0.47  0.   -1.89  0.    0.94  3.3  -1.41 -0.94  0.   -0.94
  0.94 -0.94 -1.89  0.94 -1.89  0.47  0.94  0.47  0.47  0.   -1.41 -0.47
  0.47 -1.41  0.47  0.47  0.    0.47 -1.41 -0.47  2.83 -1.41  0.94 -1.41
  0.47 -0.47  1.41  0.94 -0.47  0.94 -1.89 -0.47 -1.89 -0.94  0.    0.94
  0.   -2.36 -0.94 -1.41 -0.94 -0.47  2.83 -0.47 -0.47 -1.41  0.47  0.
  0.    1.41  0.   -0.94 -0.94  0.94  2.36  0.47]
```

Ces 200 fréquences centrées réduites $\frac{f-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$ ont pour histogramme :



Comme pour la moyenne, on en déduit qu'il y a 95% de chances pour que :

$$-1,96 < \frac{f-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} < 1,96$$

ou encore :

$$\pi - 1,96\sqrt{\frac{\pi(1-\pi)}{n}} < f < \pi + 1,96\sqrt{\frac{\pi(1-\pi)}{n}}$$

ce qui équivaut à :

$$f - 1,96\sqrt{\frac{\pi(1-\pi)}{n}} < \pi < f + 1,96\sqrt{\frac{\pi(1-\pi)}{n}}$$

mais aussi [7] :

$$\boxed{f - 1,96\sqrt{\frac{f(1-f)}{n}} < \pi < f + 1,96\sqrt{\frac{f(1-f)}{n}}}$$

7. Cette équivalence très technique, qui nécessite les coniques, est admise

1. Estimer π [a] par intervalle de confiance à l'aide du premier échantillon [b] :
   - au niveau 95%
   - au niveau 90%
2. Que se passe-t-il lorsque l'on diminue la confiance ?

---

a. Proportion de gauchers parmi tous les élèves des gymnases du canton de Vaud

b. Celui des 50 élèves de 2C du gymnase d'Yverdon

Gymnase Yverdon

La **confiance** est la probabilité pour que le paramètre estimé se trouve dans l'intervalle.

La **précision** est la demi-longueur de l'intervalle, on peut d'ailleurs lire parfois :

$$\pi = f \pm 1,96\sqrt{\frac{f(1-f)}{n}}$$

Quels sont les deux leviers permettant d'améliorer la précision ?

La forme de la précision nous invite à :

- diminuer la confiance
- augmenter la taille de l'échantillon

Quelle est la taille minimale de l'échantillon pour avoir une confiance de 95% et une précision n'excédant pas 8% ?

Comme nous sommes en train de régler la taille de l'échantillon à observer, nous ne disposons pas encore de la fréquence f, il s'agit donc de considérer « le pire des cas ».

Sur $[0\,;1]$, le maximum de la fonction $f(x)=x(1-x)$ est atteint en $\frac{1}{2}$, et vaut $\frac{1}{4}$.

Par conséquent, quelque soit la fréquence f (non encore observée), on est assuré d'avoir $f(1-f)\leq\frac{1}{4}$ et donc $1{,}96\sqrt{\frac{f(1-f)}{n}}\leq 1{,}96\sqrt{\frac{1}{4n}}$.

Pour que la précision n'excède pas 8%, il suffit alors d'avoir $1{,}96\sqrt{\frac{1}{4n}}\leq 0{,}08$ ce qui équivaut à $\frac{1}{4n}\leq\left(\frac{0{,}08}{1{,}96}\right)^2$ ou encore $n\geq\frac{1}{4}\left(\frac{1{,}96}{0{,}08}\right)^2$ c'est à dire $n\geq 150$

Gymnase Yverdon