statistiques

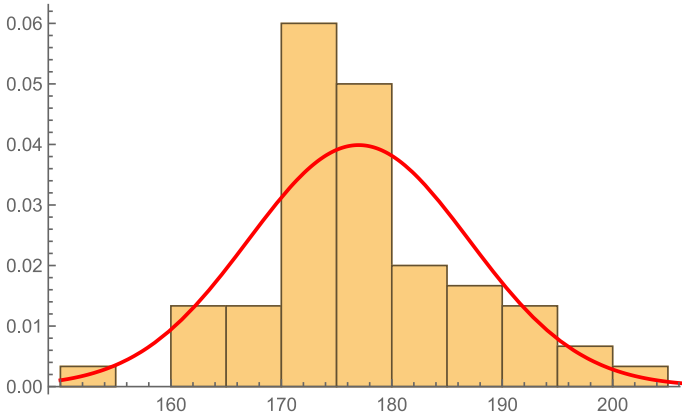
Par Élise Davignon

MAT1978

Hiver 2021

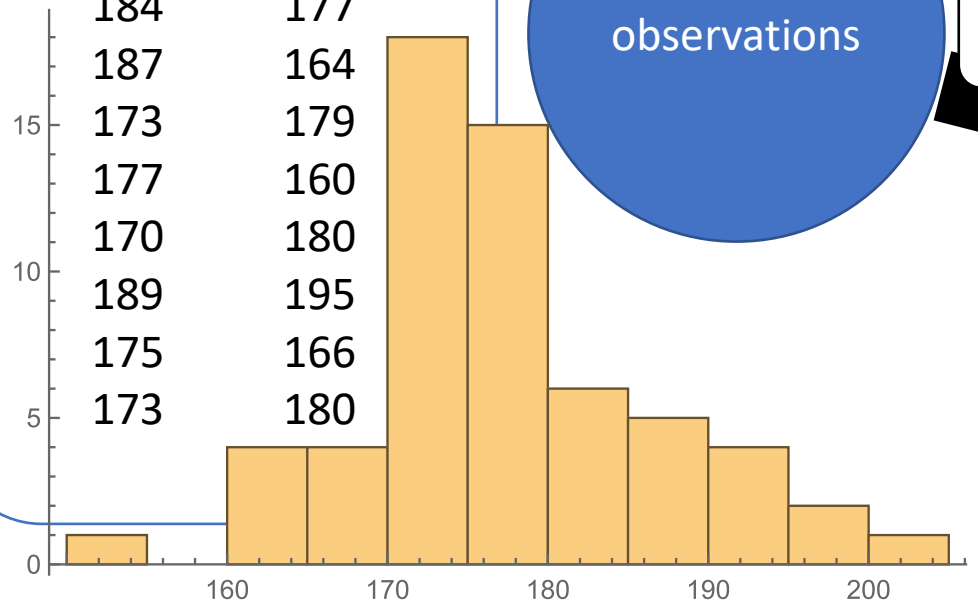
IPSES : section 2.3.
Sections 6.1 à 6.5

objectif



données

171	174
191	180
183	174
163	185
184	177
187	164
173	179
177	160
170	180
189	195
175	166
173	180

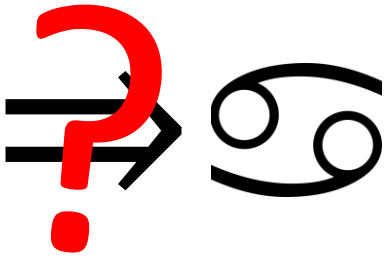


observations

statistiques

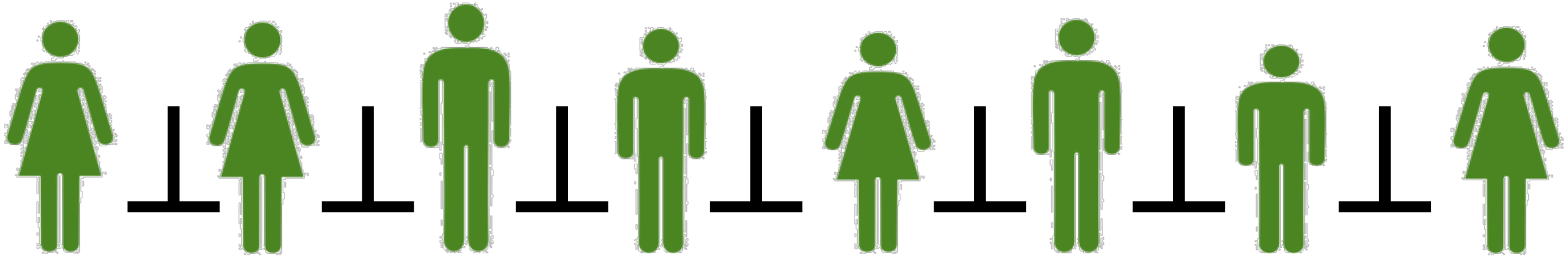
prédictions

explications

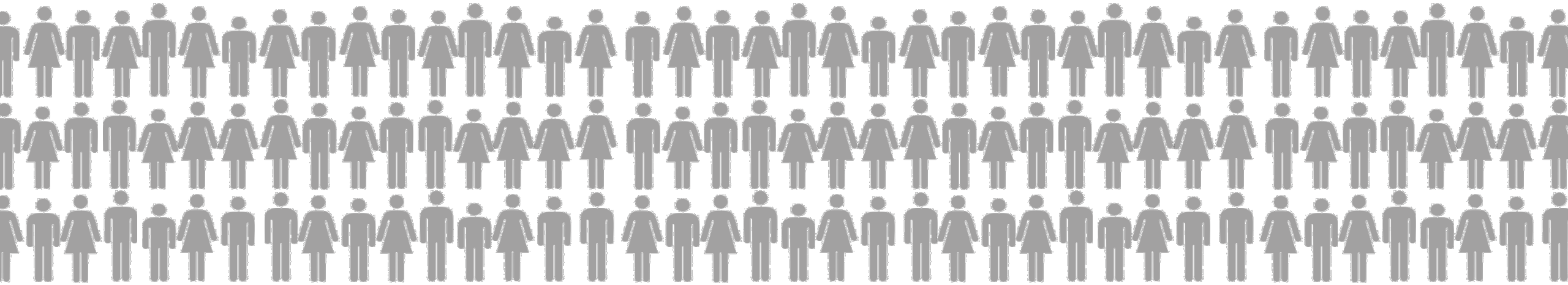


échantillon v. population

hypothèses générales



indépendance



|population| ≈ ∞

hypothèses générales

$$X_1, X_2, X_3, X_4, X_5, X_6, \ldots, X_n$$

données de l'échantillon : variables aléatoires i.i.d.

statistique :

Fonction des données de l'échantillon.

$$S = f(X_1, X_2, X_3, \ldots, X_n)$$

⇨ aussi une variable aléatoire

exemples de statistiques :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

moyenne empirique
(ou échantillonnale)

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

variance empirique
(ou échantillonnale)

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

Écart-type empirique
(ou échantillonnale)

$$X_{(1)} = \min\{X_i : 1 \leq i \leq n\},$$
$$X_{(k+1)} = \min\{X_i : X_i \geq X_{(k)}\}$$

statistiques d'ordre :
$X_{(k)}$ est le kième plus petit élément de l'échantillon.

$$M = \begin{cases} \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+1}{2}\right)}}{2}, & n \text{ est pair} \\ X_{\left(\frac{n+1}{2}\right)}, & n \text{ est impair} \end{cases}$$

la médiane empirique

$$X_{(n)} - X_{(1)}$$

l'étendue échantillonnale

exemple

Rendez-vous sur la feuille de calcul suivante : https://docs.google.com/spreadsheets/d/1E-MWquc4mXN0k5HaxF1Qr5v3QItpkNVV-AvQBPYpWU8/edit?usp=sharing

et inscrivez votre taille en centimètres, au centimètre près (valeurs entières SVP)

On va calculer toutes les statistiques précédentes.

la moyenne échantillonnale

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

distribution de la population
=
paramètres de la loi aléatoire des données de l'échantillon

$$\mu = \mathbb{E}[X_1]$$
$$\sigma^2 = \mathrm{Var}[X_1]$$

remarque : les variables aléatoires X_i (données de l'échantillon) peuvent *a priori* suivre n'importe quelle loi aléatoire.

$$\overline{\mu} = \mathbb{E}\left[\overline{X}\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \cdot n\,\mu = \mu$$

$$\overline{\sigma}^2 = \mathrm{Var}\left[\overline{X}\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

$$\overline{\sigma} = \frac{\sigma}{\sqrt{n}}$$

Par la loi des grands nombres, lorsque n tend vers l'infini,

$$\overline{X} \to \mu$$

la moyenne échantillonnale

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Par la loi des grands nombres, lorsque n tend vers l'infini,

$$\overline{X} \to \mu$$

Par le Théorème de la limite centrale, pour n grand, on a que $\overline{X}$ est de loi à peu près normale, avec espérance $\overline{\mu}$ et écart-type $\overline{\sigma}$

$$\overline{X} - \mu \approx \mathcal{N}\left(0, \frac{\sigma}{\sqrt{n}}\right)$$

$\overline{X}$ est un **estimateur** de μ

estimateur : statistique qui fournit un estimé d'un paramètre de la distribution de la population.

Intervalle de confiance

On connaît :

- Les données
- Les statistiques

$$\overline{X} = \dots$$

Du point de vue empirique, on connaît ces valeurs parce qu'on les a recueillies et mesurées...

On cherche :

- Les paramètres de la distribution de la population

Intervalle de confiance

$$\mu \in (\overline{X} - \epsilon, \overline{X} + \epsilon) \text{ avec proba } \alpha$$

… et on veut trouver les paramètres de la population à partir des données de l'échantillon.

niveau de confiance

Intervalle de confiance

On cherche ϵ t.q. $\mathbb{P}\{\overline{X} - \epsilon < \mu < \overline{X} + \epsilon\} = \alpha$

$$\Rightarrow \quad \mathbb{P}\{-\epsilon < \overline{X} - \mu < \epsilon\} = \alpha$$

$$\Rightarrow \quad \mathbb{P}\left\{-\epsilon \cdot \frac{\sqrt{n}}{\sigma} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \epsilon \cdot \frac{\sqrt{n}}{\sigma}\right\} = \alpha$$

$$\Rightarrow \quad \mathbb{P}\left\{-\epsilon \cdot \frac{\sqrt{n}}{\sigma} < \mathcal{N}(0,1) < \epsilon \cdot \frac{\sqrt{n}}{\sigma}\right\} \approx \alpha$$

$$\Rightarrow \quad 2\,\Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right) - 1 \approx \alpha$$

$$\Rightarrow \quad \epsilon \approx \frac{\sigma}{\sqrt{n}} \cdot \Phi^{-1}\left(\frac{\alpha + 1}{2}\right)$$

exemple

La moyenne de la population μ est inconnue

L'écart-type de la population est $\sigma = 3$

La taille échantillonnale est $n = 36$

La moyenne échantillonnale est $\overline{X} = 4{,}5$

Trouver les intervalles de confiance à $\alpha = 0{,}5$ (50%) et $\alpha = 0{,}95$ (95%).

Interpréter.

la variance échantillonnale $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

On a $S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$

Preuve rapide :
Si on choisit uniformément l'une des valeurs de l'échantillon au hasard, $\overline{X}$ est l'espérance de ce nombre, $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ est l'espérance de ce nombre au caré, et S^2 en est la variance. Le résultat suit.

Preuve longue :
Au tableau

la variance échantillonnale $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

On a $$\mathbb{E}[S^2] = \left(1 - \frac{1}{n}\right)\sigma^2$$

$$\mathbb{E}[S^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}\left[\overline{X}^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \overline{\sigma}^2 - \overline{\mu}^2$$

$$= \frac{1}{n}(n\sigma^2 + n\mu^2) - \frac{\sigma^2}{n} - \mu^2$$

$$= \sigma^2 - \frac{\sigma^2}{n} = \left(1 - \frac{1}{n}\right)\sigma^2$$

Donc, quand n tend vers l'infini $\mathbb{E}[S^2] \to \sigma^2$

la variance échantillonnale non biaisée

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

On a

$$\mathbb{E}\left[\hat{S}^2\right] = \sigma^2$$

S^2 est un **estimateur biaisé** de σ^2 -- ce n'est qu'à la limite que les estimations sont bonnes.

$\hat{S}^2$ est un **estimateur non-biaisé** de σ^2 car son espérance est constante.

la variance échantillonnale $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

Si les données de l'échantillon sont **normales** avec espérance μ et écart-type σ, alors

$$\overline{X} \approx \mu + \mathcal{N}\left(0, \frac{\sigma}{\sqrt{n}}\right) \quad \perp \quad S^2 \approx \frac{\sigma^2}{n} \cdot \chi^2_{n-1}$$

$$\hat{S}^2 \approx \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1}$$

Preuve :
difficile.

Les estimateurs S^2 et $\hat{S}^2$ suivent dans ce cas des lois khi-carré, et ont des écarts-types de l'ordre de $\frac{2\sigma}{\sqrt{n}}$.

la loi khi-carré (χ^2)

Paramètre : $n \in \mathbb{N}$

$$\underbrace{X_1^2 + X_2^2 + \cdots + X_n^2}_{} = \chi_n^2$$

X_i : variables aléatoires normales standard indépendantes

variable aléatoire de loi khi-carré avec n degrés de liberté.

$$\mathbb{E}[\chi_n^2] = n \qquad \mathrm{Var}[\chi_n^2] = 2\,n$$

la variance échantillonnale $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Éléments de preuve :

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\mu + \mu^2 - X_i^2 + 2X_i\overline{X} - \overline{X}^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(-2X_i\mu + \mu^2 - 2X_i\overline{X} - \overline{X}^2\right)$$

$$= -2(\mu - \overline{X})\frac{1}{n}\sum_{i=1}^{n}X_i + \mu^2 - \overline{X}^2$$

$$= -2(\mu - \overline{X})\overline{X} + \mu^2 - \overline{X}^2$$

$$= 2\overline{X}^2 - \overline{X}^2 - 2\mu\overline{X} + \mu^2$$

$$= (\overline{X} - \mu)^2$$

la variance échantillonnale $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Éléments de preuve :

$$\underbrace{\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2}_{\chi^2_n} - \frac{n}{\sigma^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

$\left(\frac{X_i-\mu}{\sigma}\right) = \mathcal{N}(0,1)$

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

$$= \frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)$$

$$= \frac{n}{\sigma^2}(\overline{X}-\mu)^2 \qquad = \underbrace{\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2}_{\chi^2_1} \approx \mathcal{N}(0,1)$$

$$\Rightarrow \quad \chi^2_n - \frac{n}{\sigma^2}S^2 \approx \chi^2_1$$

Et on complète avec de petites manipulations algébriques.

la loi khi-carré (χ^2) $F(x) = ?$

$$\overline{F_n}(z_{n,\alpha}) = \mathbb{P}\{\chi_n^2 > z_{n,\alpha}\} = \alpha$$

$$\overline{F_n}^{-1}(\alpha) = z_{n,\alpha}$$

n	$\alpha = 0.975$	$\alpha = 0.95$	$\alpha = 0.05$	$\alpha = 0.025$
5	0.83	1.15	11.07	12.83
6	1.24	1.64	12.59	14.45
7	1.69	2.17	14.07	16.01
8	2.18	2.73	15.51	17.54
9	2.70	3.33	16.92	19.02
10	3.25	3.94	18.31	20.48
11	3.82	4.58	19.68	21.92
12	4.40	5.23	21.03	23.34
13	5.01	5.89	22.36	24.74
14	5.63	6.57	23.69	26.12

Cette table donne les inverses des fonctions de répartition complémentaires pour les lois du khi-carré.

la loi du khi-carré

Les variables khi-carré sont des sommes de variables aléatoires i.i.d.

$$X_1^2 + X_2^2 + \cdots + X_n^2 = \chi_n^2$$

⇨

$$\chi_n^2 \approx \mathcal{N}\left(n, \sqrt{2n}\right)$$

lorsque n est grand, par le théorème de la limite centrale

exemple

Trouver la probabilité

$$\mathbb{P}\{\chi^2_{15} \geq 24{,}96\}$$

En utilisant l'approximation normale standard.

exemple

On reçoit une grosse cargaison d'ampoules et on nous informe que la durée de vie des ampoules a une distribution normale.

La durée de vie moyenne des ampoules est inconnue, mais on nous dit que l'écart type de la durée de vie est $\sigma = 100$ heures.

Supposons que nous testons la durée de vie de 16 ampoules. Quelle est la probabilité que l'écart-type échantillonnal non-biaisé $\hat{S} = \sqrt{\hat{S}^2}$ excède 129 heures ?

En supposant que $\hat{S} = 129$, croiriez-vous que $\sigma = 100$ heures ?

Supposons maintenant qu'on ne teste que 6 ampoules. Pour quelle valeur de s obtient-on $\mathbb{P}\{\hat{S} \leq s\} = 0{,}05$?

$\overline{X} - \mu$ si σ est inconnu

Loi T de Student

Si les données de l'échantillon sont **normales** avec espérance μ, alors

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \approx t_{n-1}$$

Lorsqu'on ne connait pas l'écart-type de la distribution de la population, mais seulement l'écart-type de l'échantillon, on a que la différence entre la moyenne de la population et la moyenne échantillonnale suit une loi T de Student.

la loi T de Student

Paramètre : $n \in \mathbb{N}$

variable normale standard

$$\frac{Z}{\sqrt{\chi_n^2/n}} = T_n$$

Variable aléatoire de loi T de Student à n degrés de liberté.

variable khi-2 à n degrés de libertés

$$\mathbb{E}[T_n] = 0 \qquad \mathrm{Var}[T_n] = \frac{n}{n-2}$$

Intervalle de confiance

On cherche ϵ t.q. $\mathbb{P}\{\overline{X} - \epsilon < \mu < \overline{X} + \epsilon\} = \alpha$

$$\mathbb{P}\{-\epsilon < \overline{X} - \mu < \epsilon\} = \alpha$$

$$\mathbb{P}\left\{-\epsilon \cdot \frac{\sqrt{n}}{S} < \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} < \epsilon \cdot \frac{\sqrt{n}}{S}\right\} = \alpha$$

$$\mathbb{P}\left\{-\epsilon \cdot \frac{\sqrt{n}}{S} < t_{n-1} < \epsilon \cdot \frac{\sqrt{n}}{S}\right\} \approx \alpha$$

$$\mathbb{P}\left\{t_{n-1} \geq \epsilon \cdot \frac{\sqrt{n}}{S}\right\} \approx \frac{1-\alpha}{2}$$

On trouve dans plusieurs ouvrages de référence des tables qui donnent des tables pour les valeurs $t_{\alpha,n}$ telles que

$$\mathbb{P}\{t_n \geq t_{\alpha,n}\} = \alpha$$

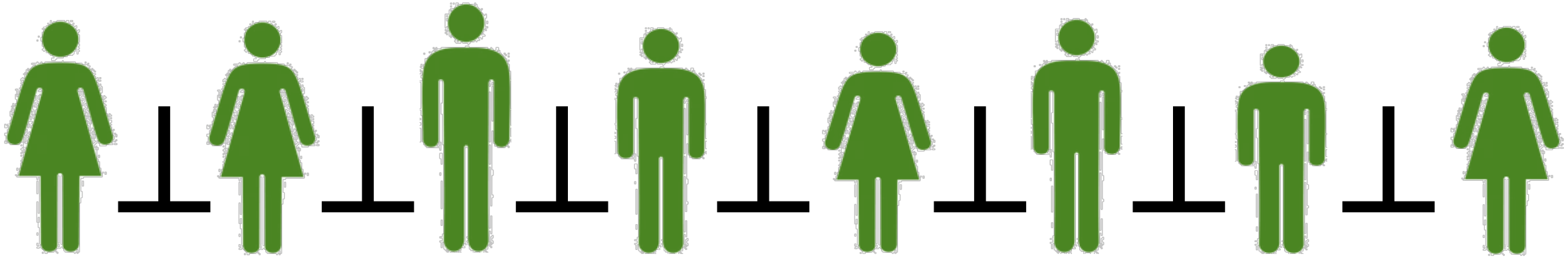
Ici, en particulier, on a

$$\epsilon \approx \frac{S}{\sqrt{n}}\, t_{\frac{1-\alpha}{2},n-1}$$

exemple

Trouver un intervalle de confiance pour la moyenne de la population à partir des données de notre échantillon pour les tailles dans la classe.

des données pour deux variables



Taille

$$X_1, X_2, X_3, X_4, X_5, X_6, \ldots, X_n$$

Âge

$$Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, \ldots, Y_n$$

Chaque donnée de l'échantillon est un vecteur (X_i, Y_i).

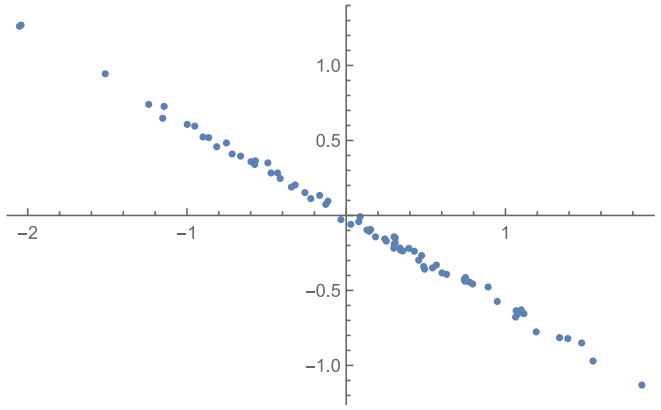
le coefficient de corrélation empirique

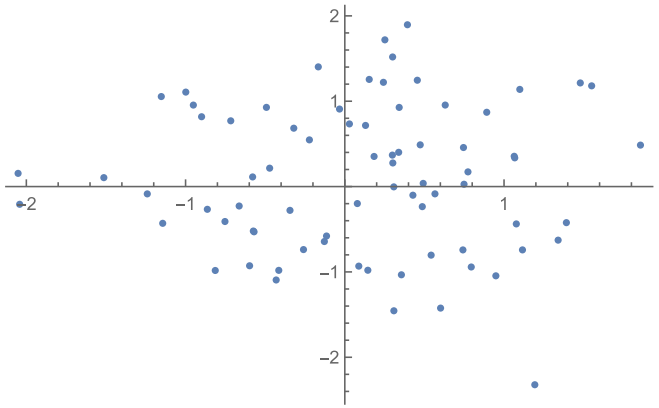
$$R_{X,Y} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}$$

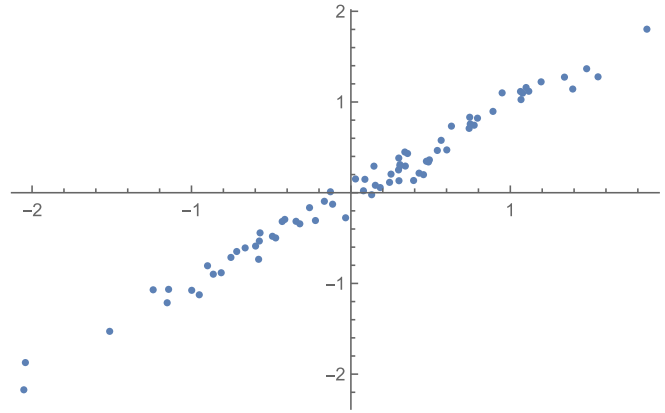
$$= \frac{1}{n} \cdot \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{S_X \cdot S_Y}$$

$R_{X,Y}$ mesure la tendance de X et Y à varier « ensemble ».

linéarité de la relation entre X et Y







$R_{X,Y} = -1$

$R_{X,Y} = 0$

$R_{X,Y} = 1$

$Si\ R_{X,Y}$ est proche de -1, les valeurs sont corrélées positivement et les données sont très près d'être linéaires.

Si $R_{X,Y} = 0$, les données n'ont pas de tendance linéaire.

$Si\ R_{X,Y}$ est proche de +1, les valeurs sont corrélées positivement et les données sont très près d'être linéaires.

relation linéaire et corrélation : analyse

$$\underbrace{Y = a\,X + b}_{\text{Y est presque une fonction linéaire de X}} + \underbrace{Z}_{\text{mais avec une perturbation aléatoire Z indépendante}}$$

On voit que, si on suppose a positif, la corrélation varie négativement avec la variance du terme de bruit. Lorsque celui-ci est nul, $\rho[X, Y] = 1$.

$$\rho[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \cdot \sigma_Y} = \frac{a\,\sigma_X^2}{\sigma_X \sqrt{a^2\sigma_X^2 + c^2\sigma_Z^2}} = \frac{\mathrm{sgn}\,a}{\sqrt{1 + \frac{\sigma_Z^2}{a^2\sigma_X^2}}}$$