Mesures de dispersion

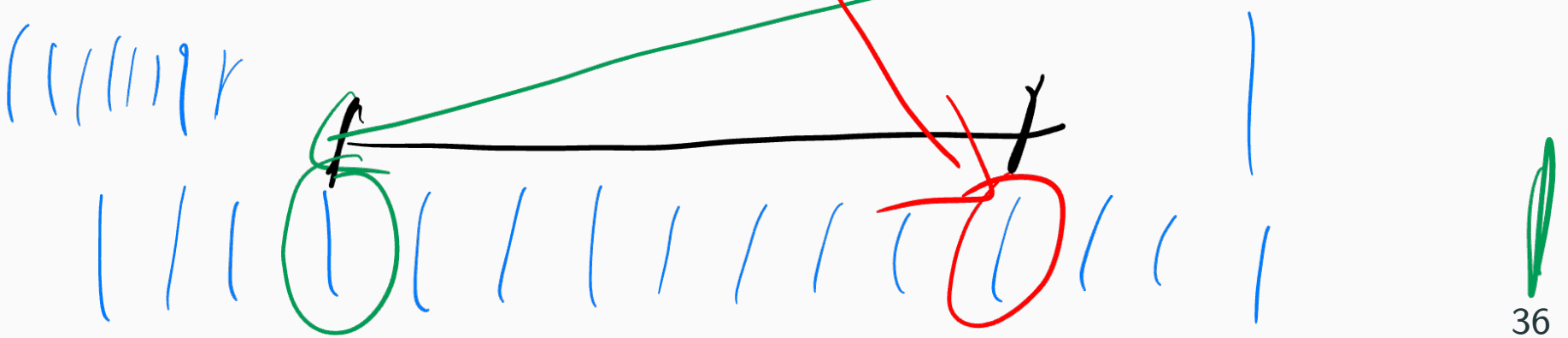
- **l'écart-type (standard deviation)**,

$$s = \left\{ \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2 \right\}^{1/2} = \left\{ \frac{1}{n-1} \left(\sum_{j=1}^{n} y_j^2 - n\bar{y}^2 \right) \right\}^{1/2},$$

où s^2 est la **variance de l'échantillon** (on verra plus tard pourquoi on divise par $n-1$)

- **l'étendue (range)**, $y_{(n)} - y_{(1)} = \max(y_1, \ldots, y_n) - \min(y_1, \ldots, y_n)$
- **l'étendue/écart interquartile (interquartile range, IQR)**,

$$\mathrm{IQR}(y) = y_{(\lceil 3n/4 \rceil)} - y_{(\lceil n/4 \rceil)}$$



Mesures de dispersion

- **l'écart-type (standard deviation)**,

$$s = \left\{ \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2 \right\}^{1/2} = \left\{ \frac{1}{n-1} \left(\sum_{j=1}^{n} y_j^2 - n\,\bar{y}^2 \right) \right\}^{1/2},$$

où s^2 est la **variance de l'échantillon** (on verra plus tard pourquoi on divise par $n-1$)

- **l'étendue (range)**, $y_{(n)} - y_{(1)} = \max(y_1, \ldots, y_n) - \min(y_1, \ldots, y_n)$
- **l'étendue/écart interquartile (interquartile range, IQR)**,

$$\mathrm{IQR}(y) = y_{(\lceil 3n/4 \rceil)} - y_{(\lceil n/4 \rceil)}$$

Dans l'exemple précédant 42, 27, 31, 45, 31, 31, 29, 36, 34, 39 on ordonne 27, 29, 31, 31, 31, 34, 36, 39, 42, 45 et trouve $\mathrm{IQR}(y) = y_{(8)} - y_{(3)} = 39 - 31 = 8$

1.4 Le boxplot (boîte à moustache)

Boxplot (boîte à moustache)

Poids (en *pounds*) de 92 étudiants d'une école américaine

140	145	160	190	155	165	150	190	195	138	160
155	153	145	170	175	175	170	180	135	170	157
130	185	190	155	170	155	215	150	145	155	155
150	155	150	180	160	135	160	130	155	150	148
155	150	140	180	190	145	150	164	140	142	136
123	155									
140	120	130	138	121	125	116	145	150	112	125
130	120	130	131	120	118	125	135	125	118	122
115	102	115	150	110	116	108	95	125	133	110
150	108									

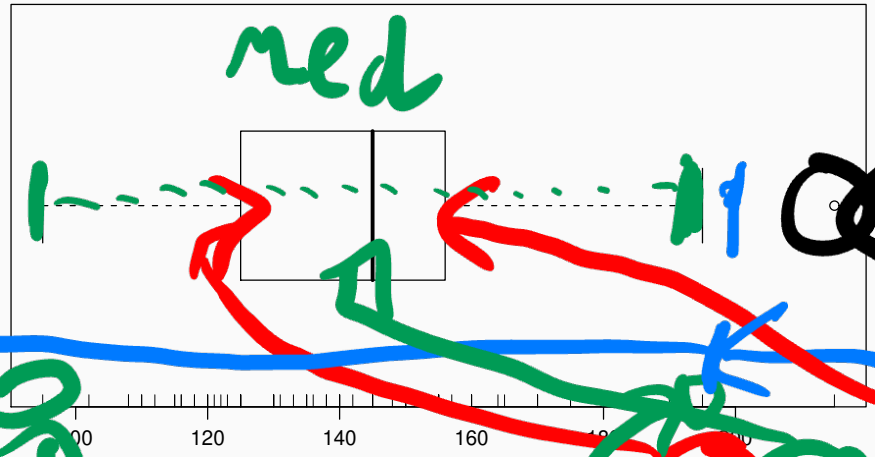
- Le **"five-number summary"** est la liste des cinq valeurs

$$y_{(1)}, \quad y_{(\lceil n/4 \rceil)}, \quad y_{(\lceil n/2 \rceil)}, \quad y_{(\lceil 3n/4 \rceil)}, \quad y_{(n)},$$

donnant un résumé numérique simple et pratique des données

- Cette liste est à la base de la **boîte à moustache (boxplot)**

Boxplot (boîte à moustache)



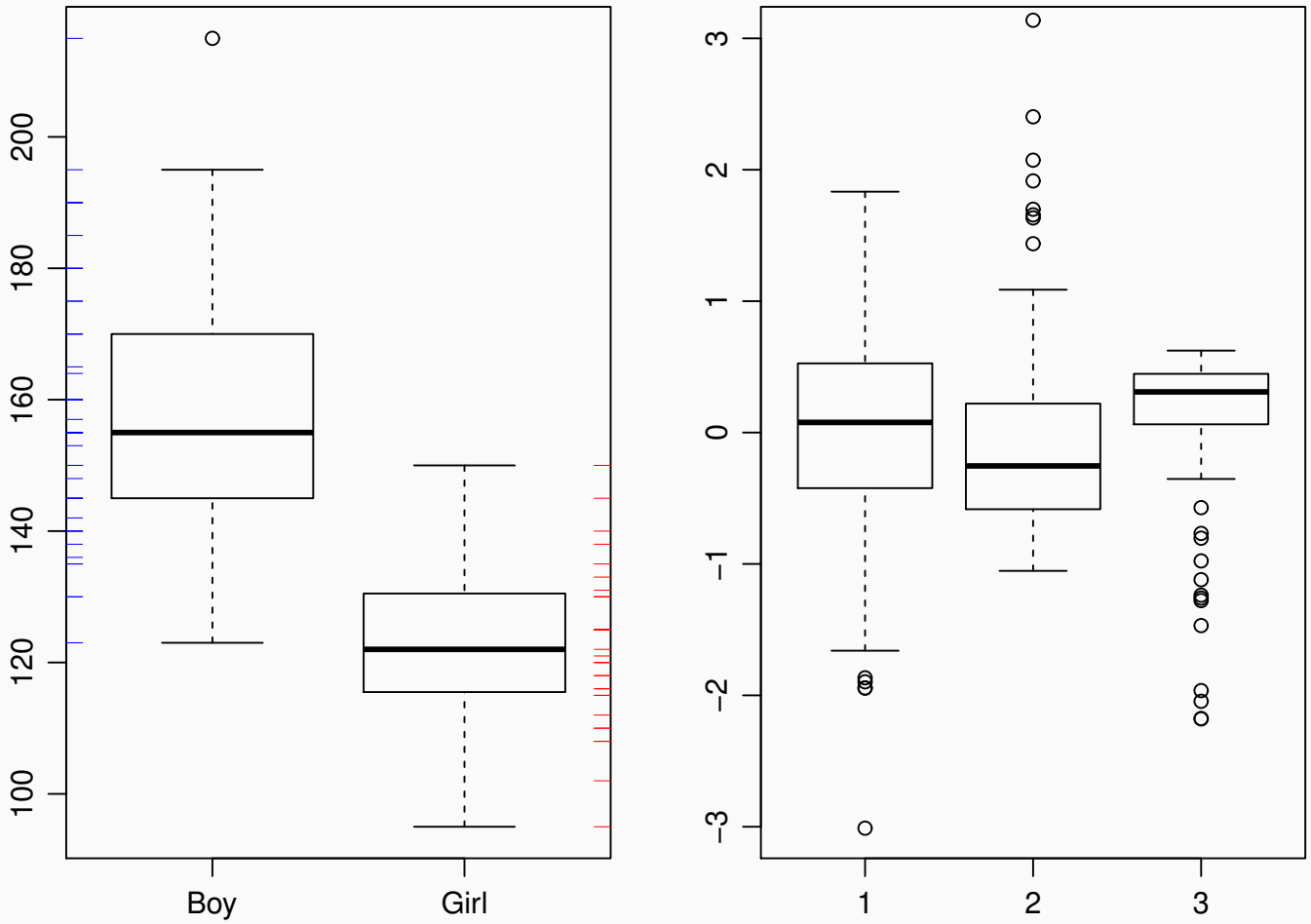
- Pour les poids, le "five-number summary" est 95, 125, 145, 156, 215, et donc

$$\begin{aligned} \mathrm{IQR}(y) &= y_{(\lceil 3n/4 \rceil)} - y_{(\lceil n/4 \rceil)} = 156 - 125 = 31 \\ C &= 1.5 \times \mathrm{IQR}(y) = 1.5 \times 31 = 46.5 \\ y_{(\lceil n/4 \rceil)} - C &= 125 - 46.5 = 78.5 \\ y_{(\lceil 3n/4 \rceil)} + C &= 156 + 46.5 = 202.5 \end{aligned}$$

- Les limites de la moustache sont les y_i les plus extrêmes qui se trouvent à l'intérieur de l'intervalle $[y_{(\lceil n/4 \rceil)} - C, y_{(\lceil 3n/4 \rceil)} + C]$
- Les y_i à l'extérieur de la moustache sont montrés individuellement

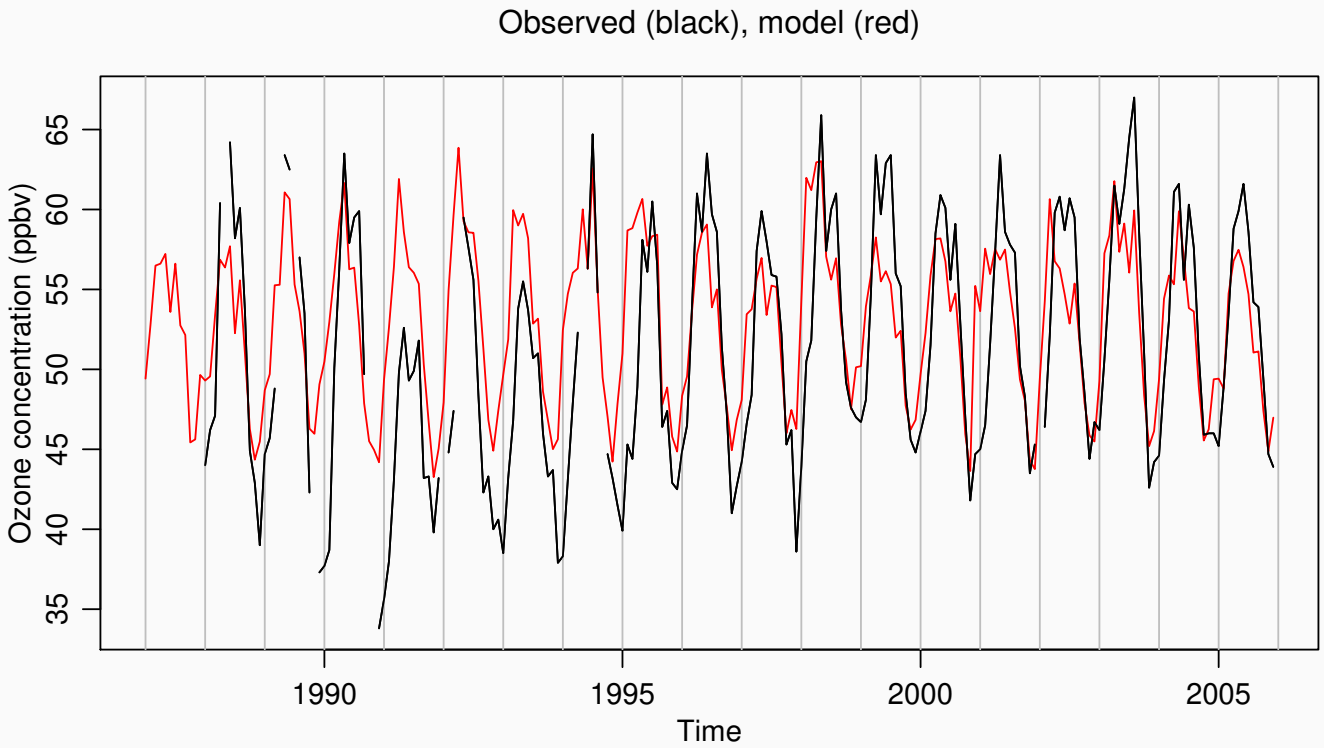
Boxplot (boîte à moustache)

- Le boxplot est utile pour la comparaison de groupes d'observations
- Boxplots du poids des étudiants selon le sexe, et de trois groupes d'observations simulées :



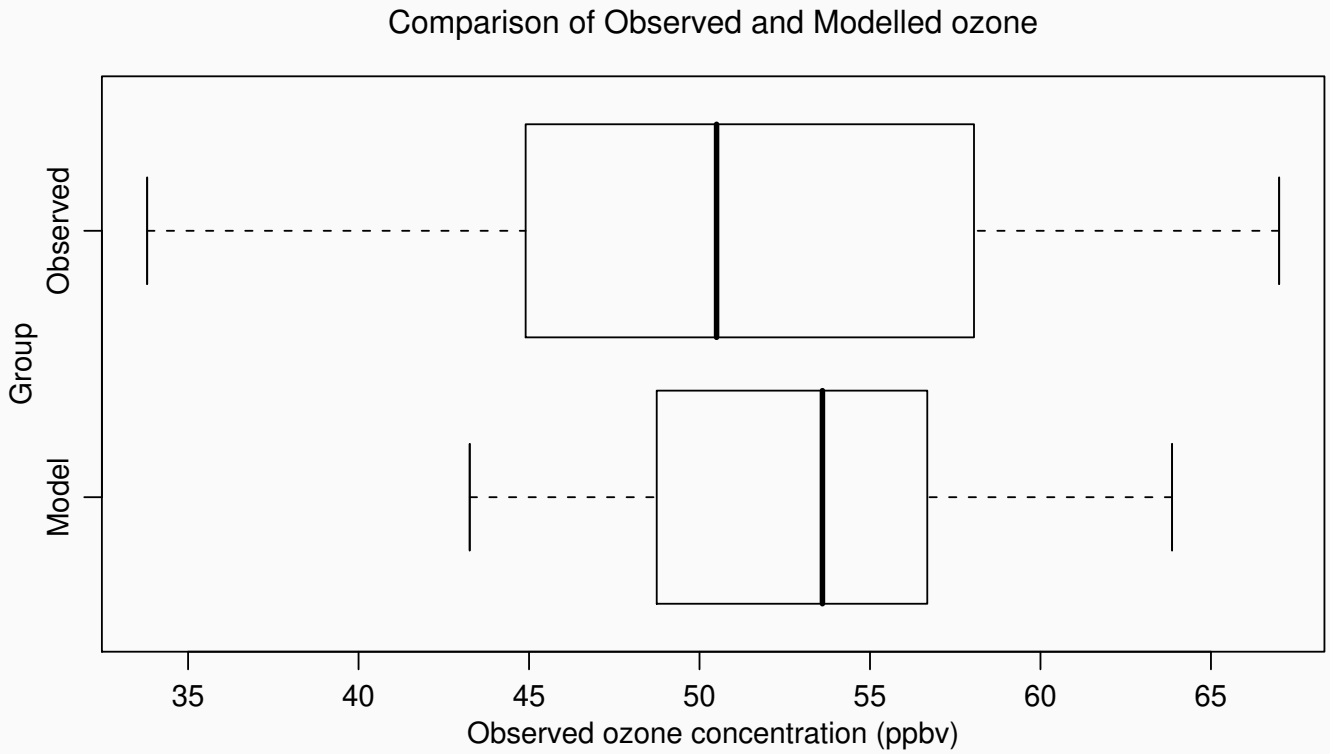
Ozone atmosphérique

Observations de la concentration de l'ozone au Jungfraujoch, de janvier 1987 à décembre 2005 (quelques valeurs manquantes), et résultats d'une modélisation



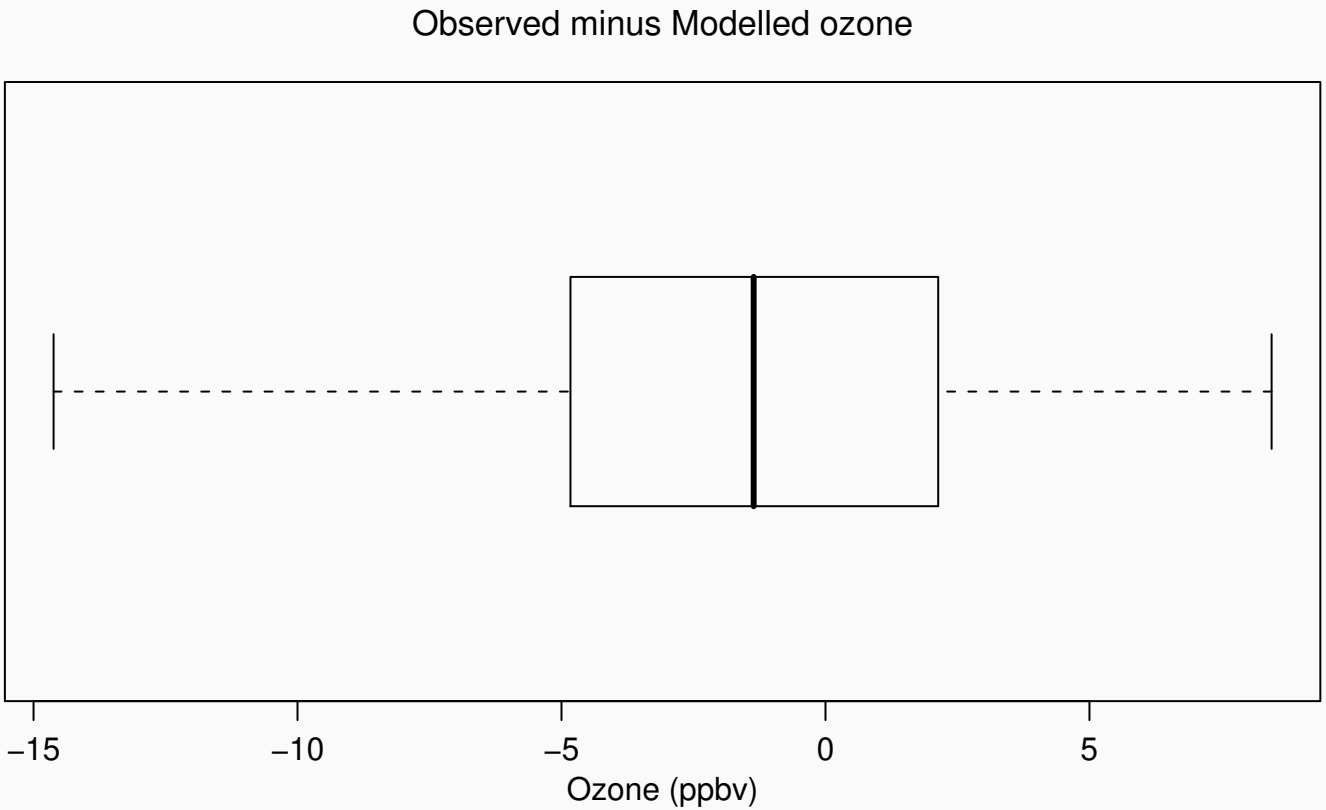
Est-ce que la modélisation est bonne ?

Ozone atmosphérique



Boxplot des données réeles et celles issues du modèle

Ozone atmosphérique



Différences des données réeles et celles issues du modèle

Commentaires

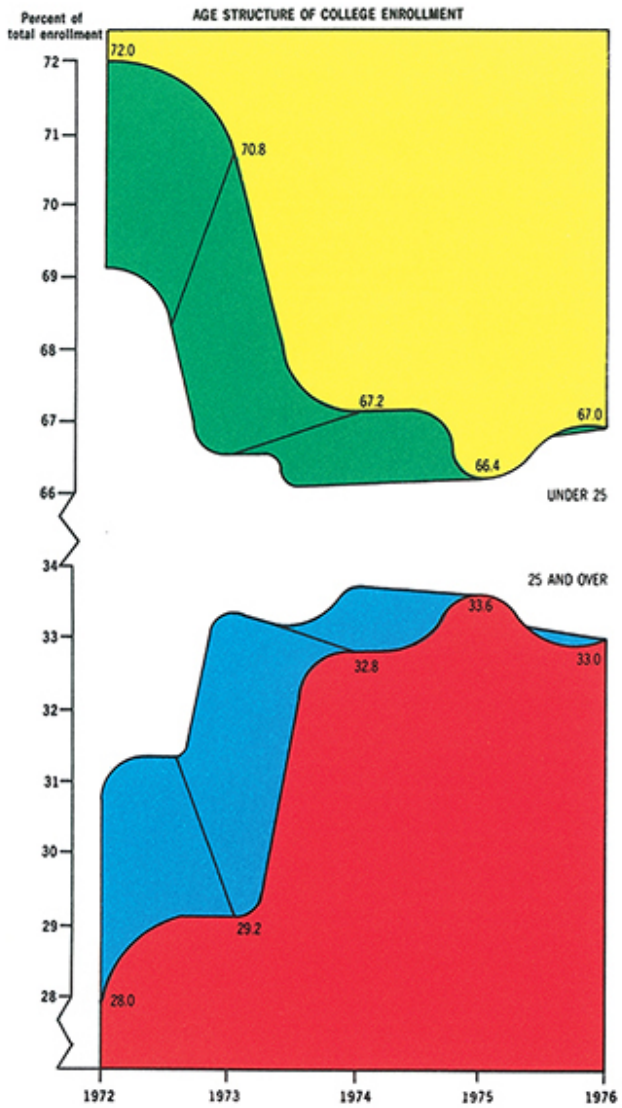
Il n'est pas toujours facile de créer de bons graphiques.

Quelques conseils :

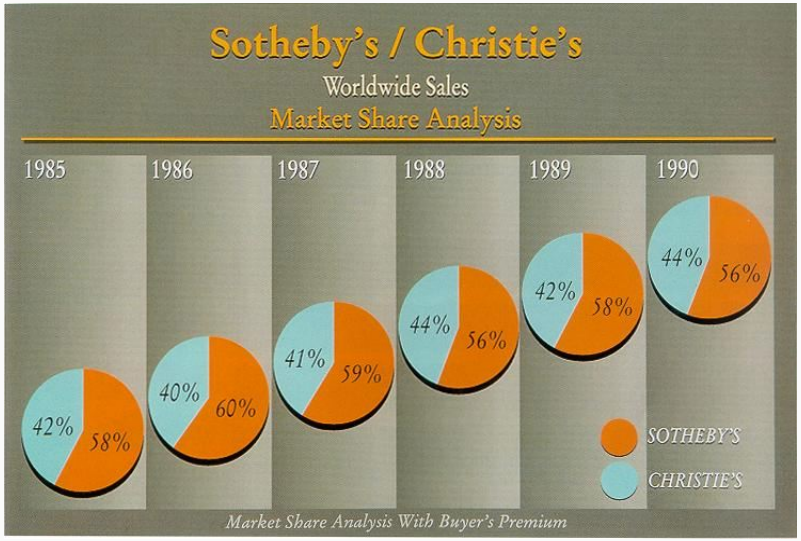
- essayer autant que possible de montrer les données elles-mêmes—pas de **fioritures/chart-junk** (couleurs/lignes/... inutiles etc.)
- mettre des unités et explications claires pour les axes et la légende
- pour comparer des quantités liées, utiliser les mêmes axes et mettre les graphiques en relation proche
- choisir les echelles telles que les relations systématiques apparaissent à un angle de $\sim 45^\circ$ des axes
- **transformer** les données peut aider à la visualisation
- dessiner le graphique de sorte que les départs du 'standard' apparaissent comme départs de la linéarité ou d'un nuage aléatoire de points

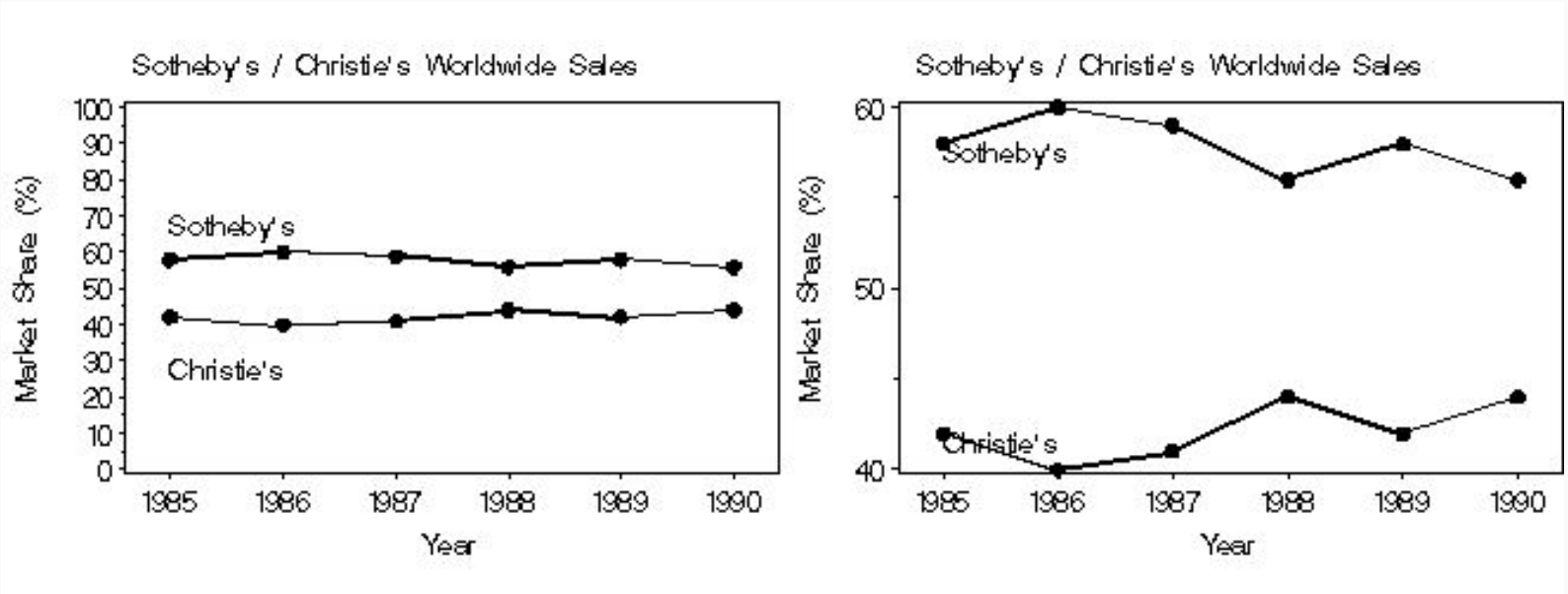
Chartjunk

Ce graphique montre 5 chiffres !



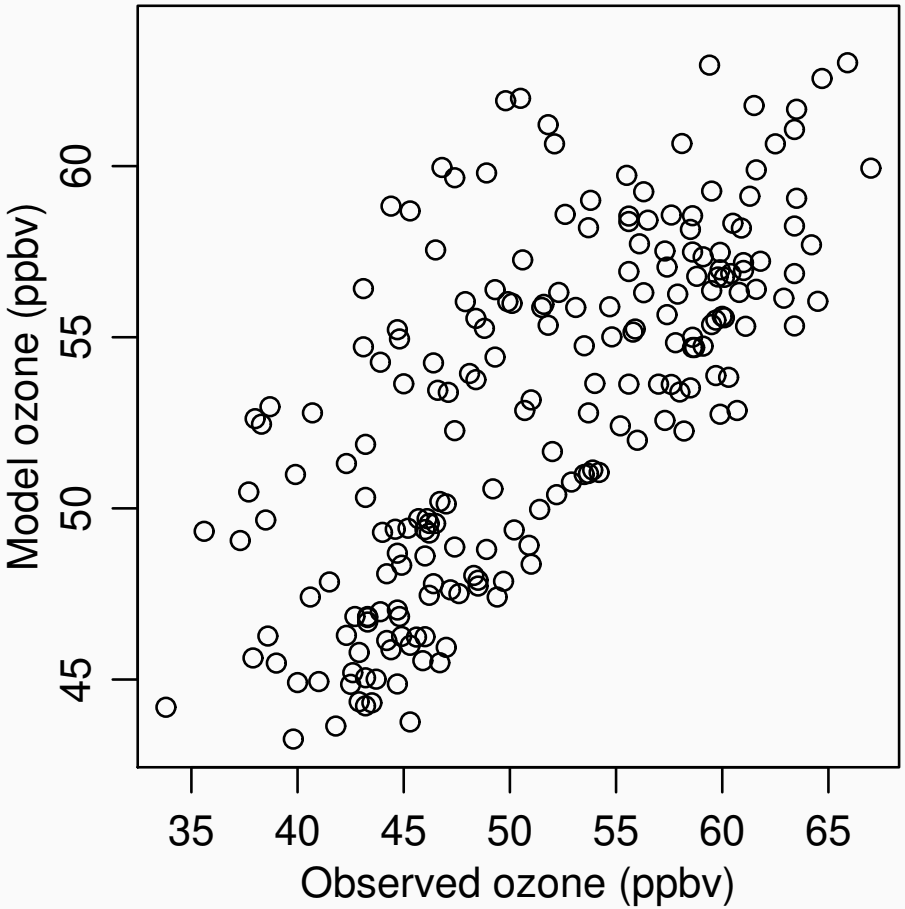
Chartjunk et échelle

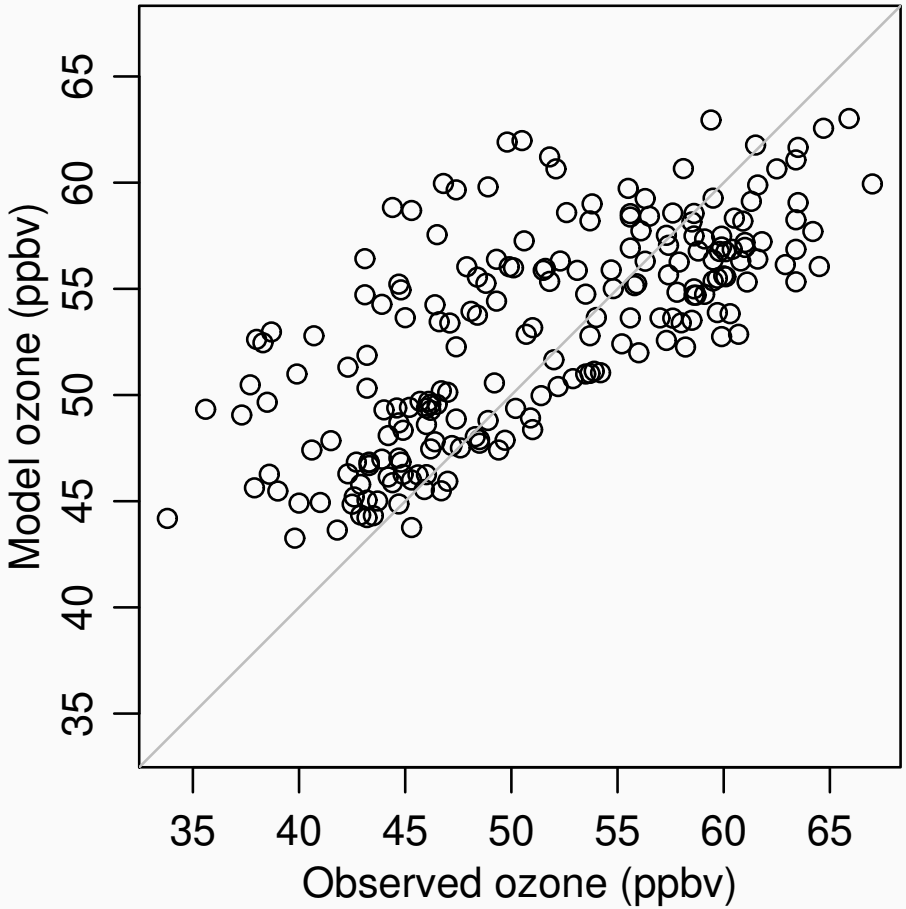




Choisir les bons axes

Effet du choix des axes sur la perception d'une relation :

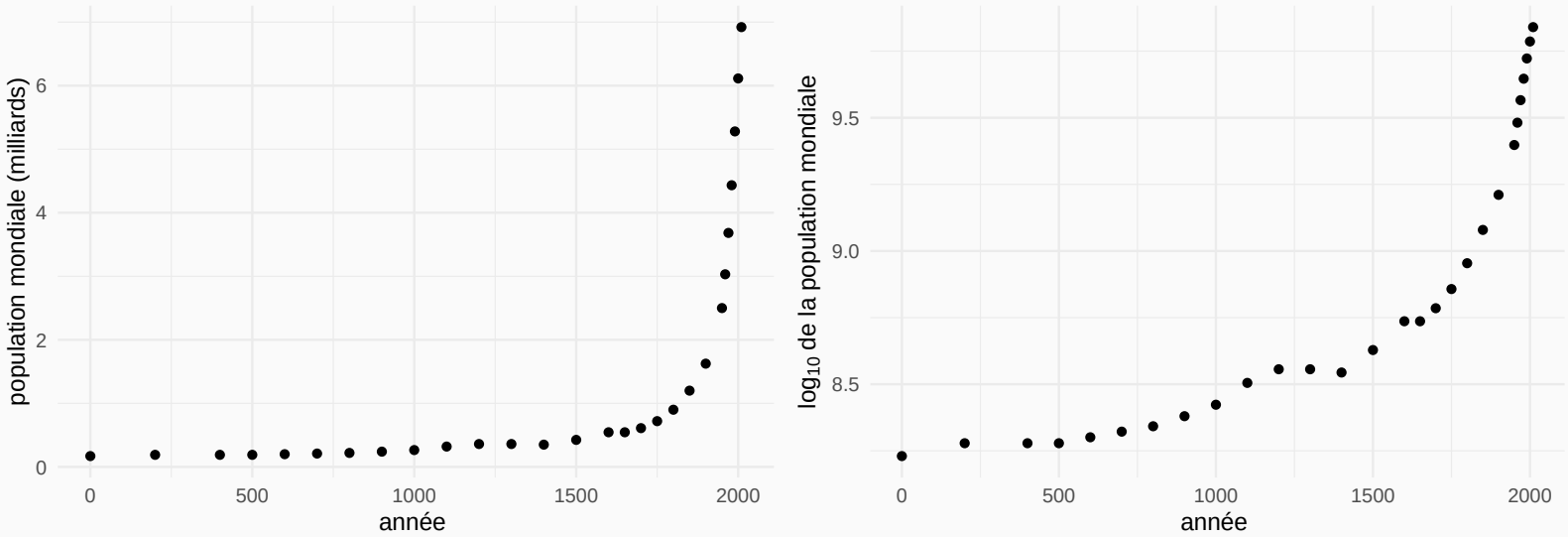




Changements d'échelles

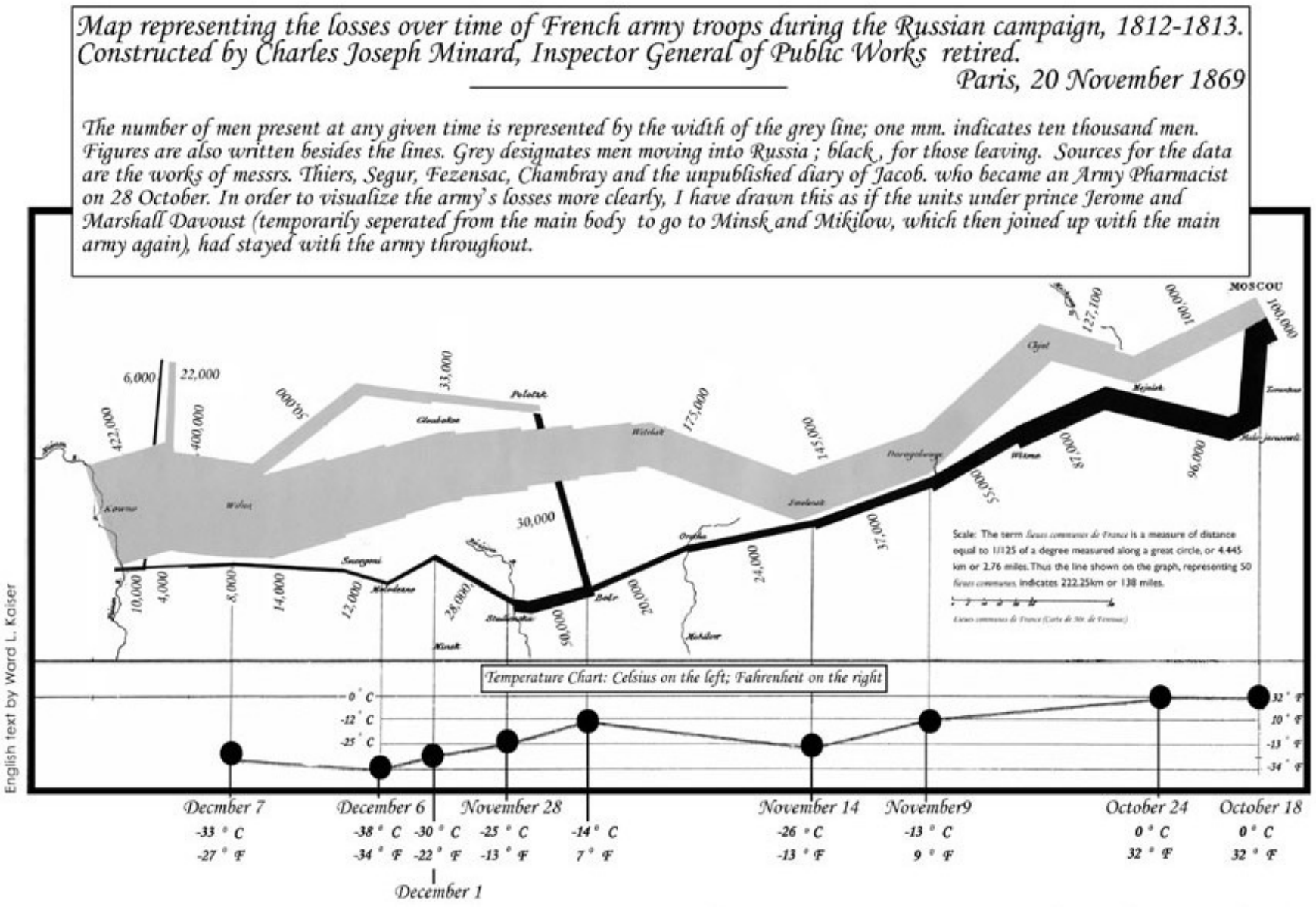
Pour certaines données, il est intéressant de les **transformer** avant de les représenter

Exemple : Population mondiale entre l'an 0 et 2000. L'échelle logarithmique permet de visualiser clairement le taux de croissance



La population en 1200 était de 360 millions, et en 1600 de 545 millions

La campagne russe de 1812



Editor s note: dates & temperatures are only referenced for the retreat from Moscow

1.5 Stratégie

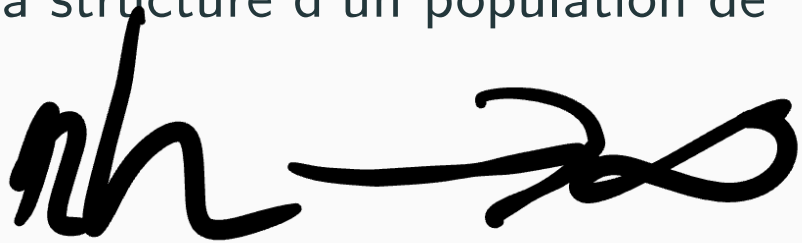
Analyse initiale des données

On a maintenant une **stratégie** pour explorer des données issues d'une variable quantitative :

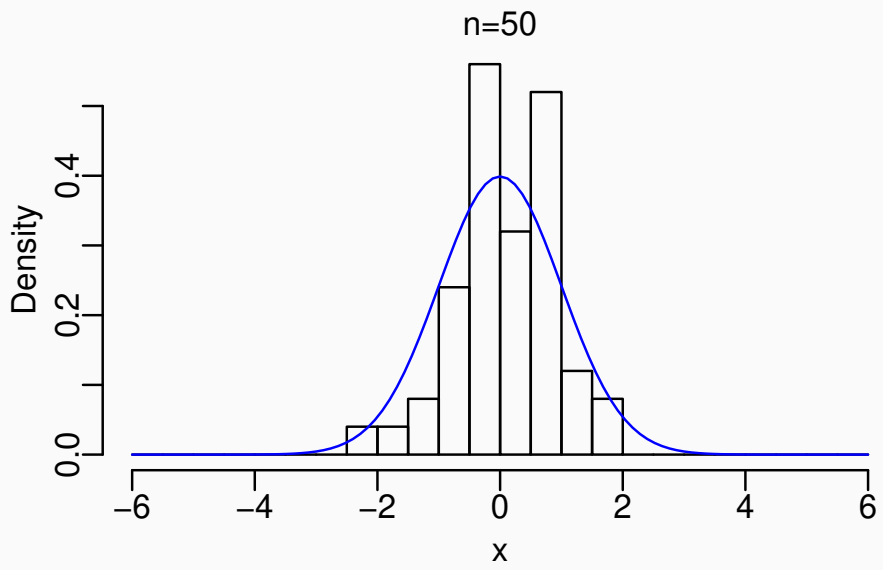
1) toujours faire des **représentations graphiques** d'abord
2) étudier la **structure globale** des données et identifier d'éventuelles valeurs atypiques / aberrantes ("outliers")—trouver pourquoi elles apparaissent
3) calculer des **synthèses numériques** pour décrire la tendance centrale (position / centre / lieu) et la dispersion (échelle)
4) souvent, la structure globale est si régulière qu'on aimerait la décrire par une courbe lisse. Cette courbe est une description mathématique pour la distribution des données

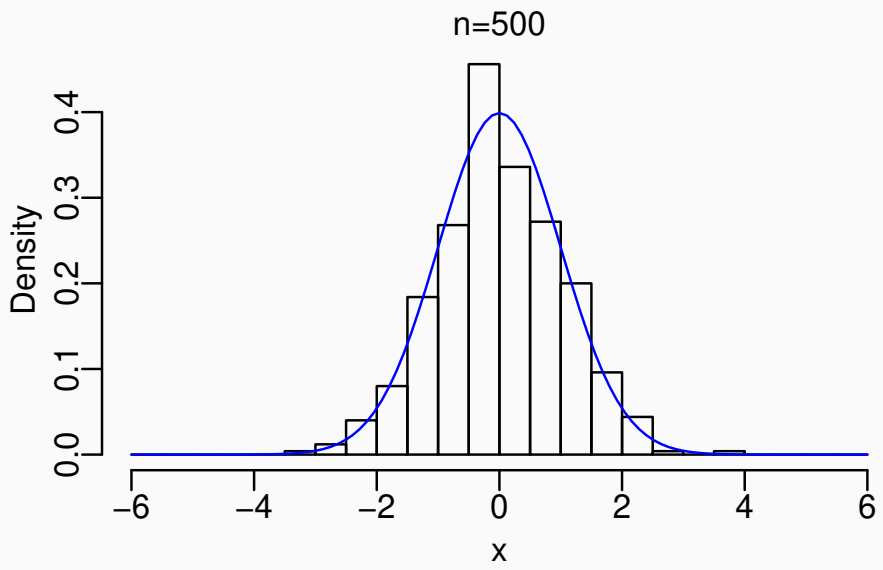
Modélisation des données

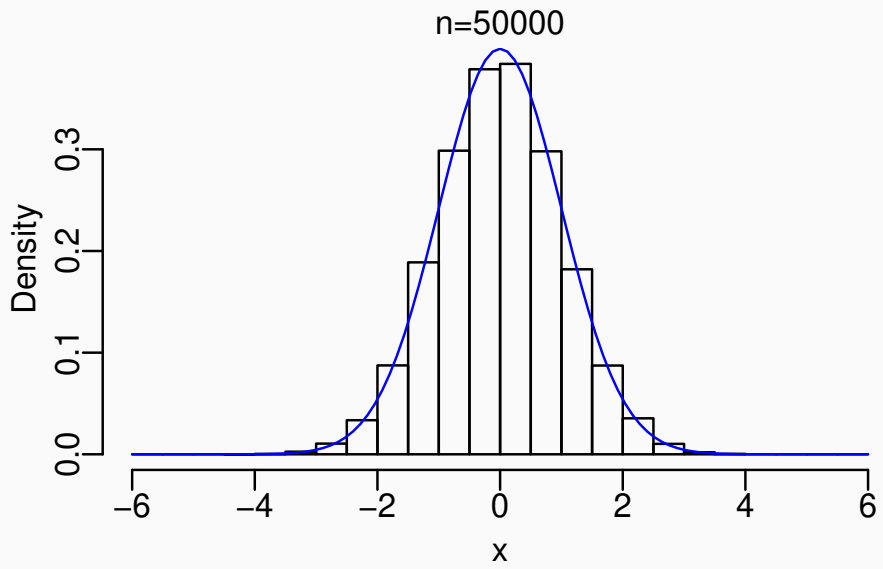
- Souvent on suppose que les données sont issues d'un échantillon aléatoire tiré d'une population d'intérêt
- Cette population est considérée comme très grande, d'une taille presque infinie
- En statistique ces modèles mathématiques sont souvent des **courbes de densité**, une fonction qui est toujours ≥ 0 et qui s'intègre à 1 ; l'aire sous cette courbe est la fréquence relative
- On peut comprendre la courbe de densité comme la limite d'un histogramme normalisé décrivant la structure d'un population de taille n, quand $n \to \infty$ et $h \to 0$

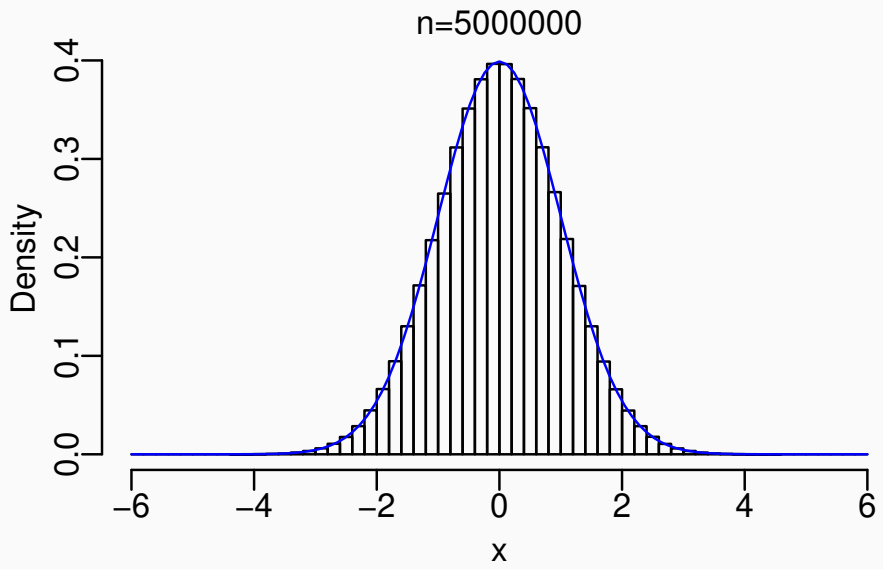


Modélisation des données, courbe de densité









1.5 La loi normale

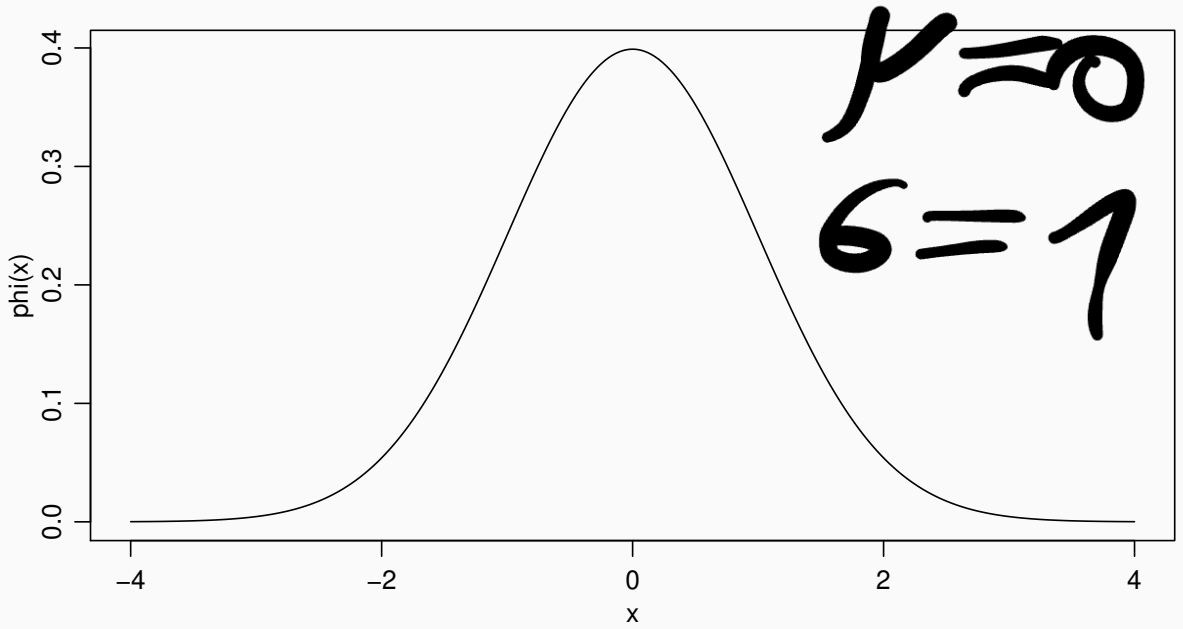
Distribution normale

Une classe particulière et importante de densités est la **densité normale (densité gaussienne)**, $\mathcal{N}(\mu, \sigma^2)$

$$f_{\mu,\sigma}(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x, \mu < \infty, \sigma > 0,$$

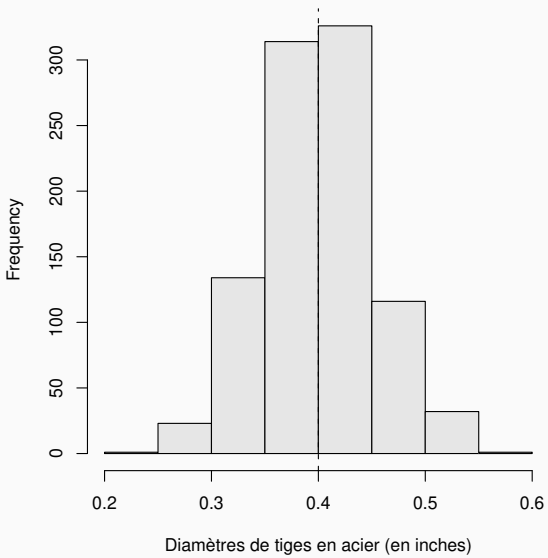
où μ est la **moyenne** et σ est l'**écart-type**

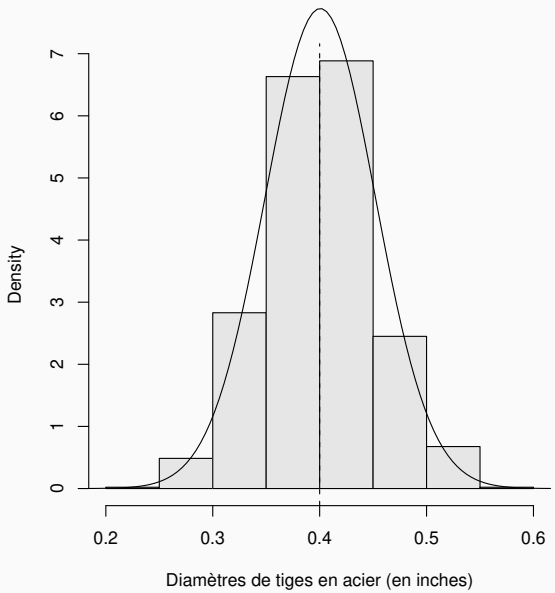
$f_{\mu,\sigma}(x)$ est la hauteur de la courbe au point x



Tiges en acier

Diamètres de 947 tiges en acier en pouces (inches)





Tiges en acier

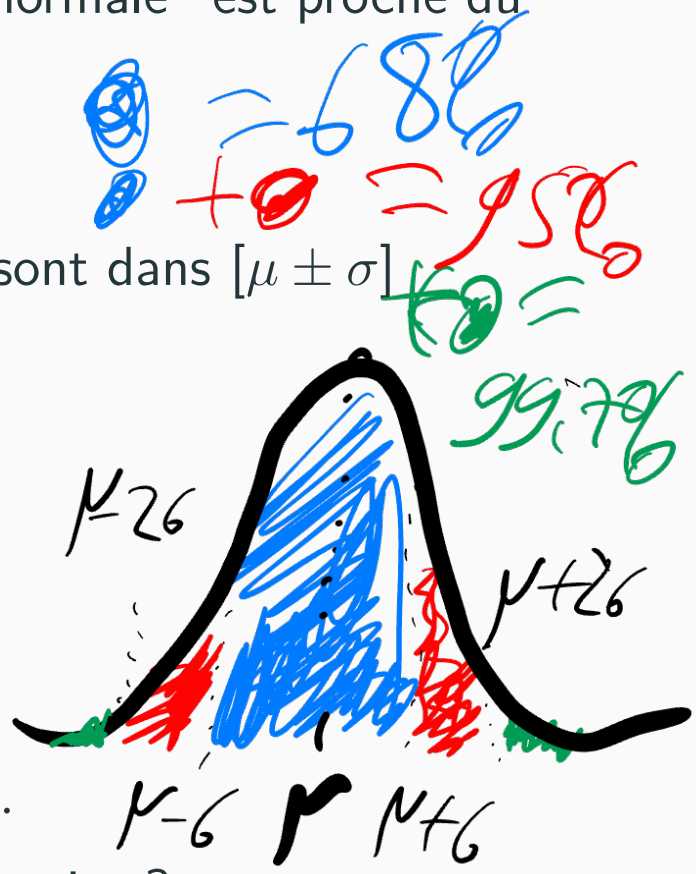
- Pour obtenir les paramètres, on calcule la moyenne $\bar{x} = 0.4$ et l'écart-type $s = 0.051$
- Courbe précédente : $\mathcal{N}(\mu = 0.40,\ \sigma^2 = 0.051^2)$
- 472 des 947 tiges ont un diamètre ≤ 0.4 inches. Donc leur fréquence relative est
$$\frac{472}{947} = 0.498$$
- L'aire correspondante de la surface sous la courbe précédente vaut 0.5 — proche de 0.498, donc donne une bonne approximation

Propriétés de $\mathcal{N}(\mu, \sigma^2)$

Il y a une infinité des densités normales selon le choix de μ et σ, mais toutes ont des propriétés communes. En voici quelques-unes :

- La majorité des observations d'une "population normale" est proche du centre μ
- La règle "68-95-99.7" :

$$\mathcal{N}(\mu, \sigma^2) \Rightarrow \begin{cases} 68\% \text{ des observations sont dans } [\mu \pm \sigma] \\ 95\% \text{ dans } [\mu \pm 2\sigma] \\ 99.7\% \text{ dans } [\mu \pm 3\sigma] \end{cases}$$



Exemple des tiges: Diamètres de 947 tiges d'acier :

69.06%	dans	$[\bar{x} \pm s]$
92.05%	dans	$[\bar{x} \pm 2s]$
99.8%	dans	$[\bar{x} \pm 3s]$.

Le modèle normal semble-t-il être une bonne approximation ?

Si oui, comment calculer ces mêmes proportions à l'aide de ce modèle ?

Standardisation

Si x est une observation issue d'une densité de moyenne μ et d'écart-type σ, alors la **valeur standardisée** de x est

$$z = \frac{x - \mu}{\sigma}$$

z est une observation issue d'une densité de moyenne 0 et d'écart-type 1

Exemple de tiges: Ici, $n = 947$, $\bar{x} = 0.400$, $s = 0.051$, et alors si on met $\mu = \bar{x}$ et $\sigma = s$, on a

$$x_{(644)} = 0.4239 \Rightarrow z_{(644)} = \frac{0.4239 - 0.400}{0.051} = 0.452$$

et de même, la tranformée $x \mapsto z = (x - \mu)/\sigma$ donne

$$\bar{x} = 0.400 \quad \Rightarrow \quad \bar{z} = 0$$

$$s_x = 0.051 \quad \Rightarrow \quad s_z = 1$$