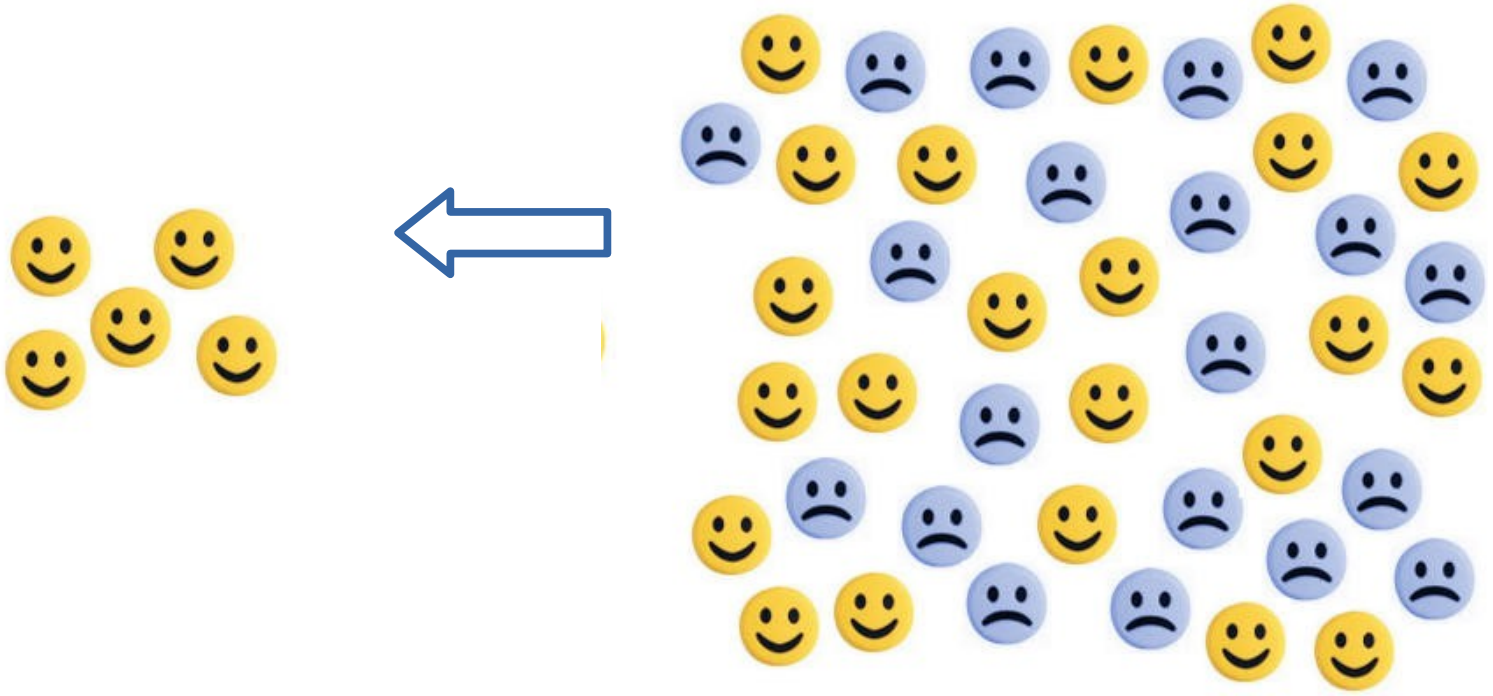
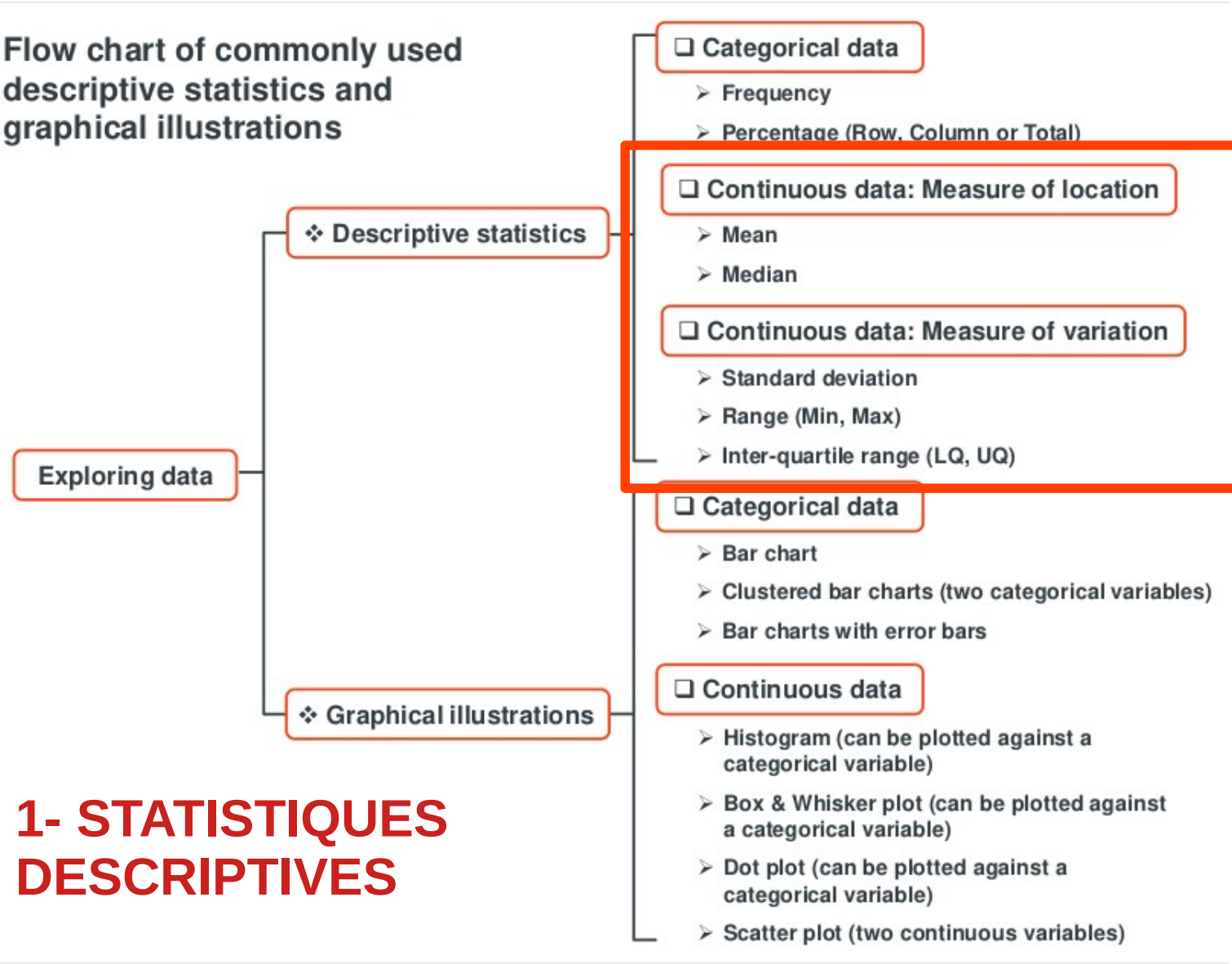
Synthétiser et visualiser des données continues: Importance de la taille de l'échantillon



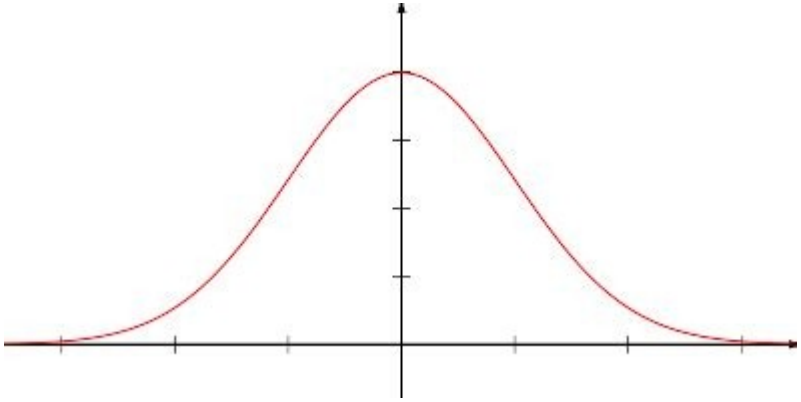


Moyenne
Médiane
Ecart-type
Min-max
IQ

1- STATISTIQUES DESCRIPTIVES

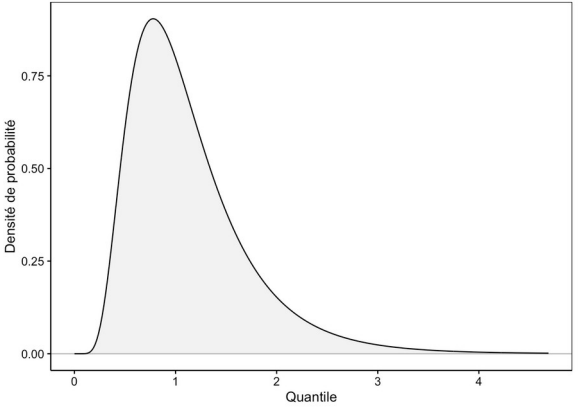
STATISTIQUES DESCRIPTIVES

- Distribution normale
paramétrique
Moyenne arithmétrique et écart-type

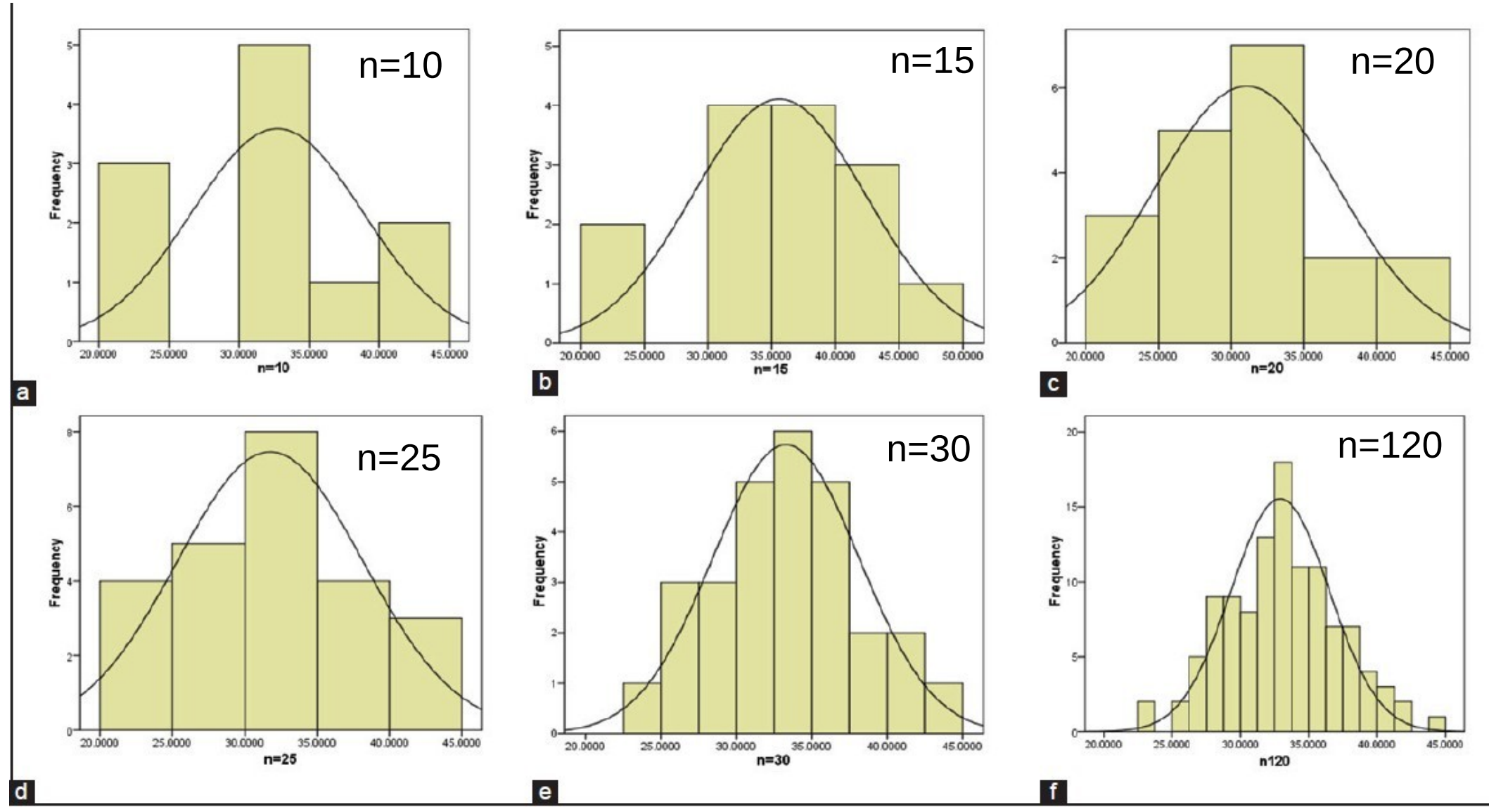


- Distribution autre
Ordinal (non-paramétrique)
Médiane et étendue min-max ou écart interquartile (25-75%)

=> petits échantillons on ne peut pas déterminer le type de distribution!



(Simpson, 2015)


n=10
Frequency
n=10
a
n=15
Frequency
n=15
b
n=20
Frequency
n=20
c
n=25
Frequency
n=25
d
n=30
Frequency
n=30
e
n=120
Frequency
n120
f

Statistiques descriptives
=>quand il y a assez de données à synthétiser

Mais qu'est-ce qu'un petit échantillon?

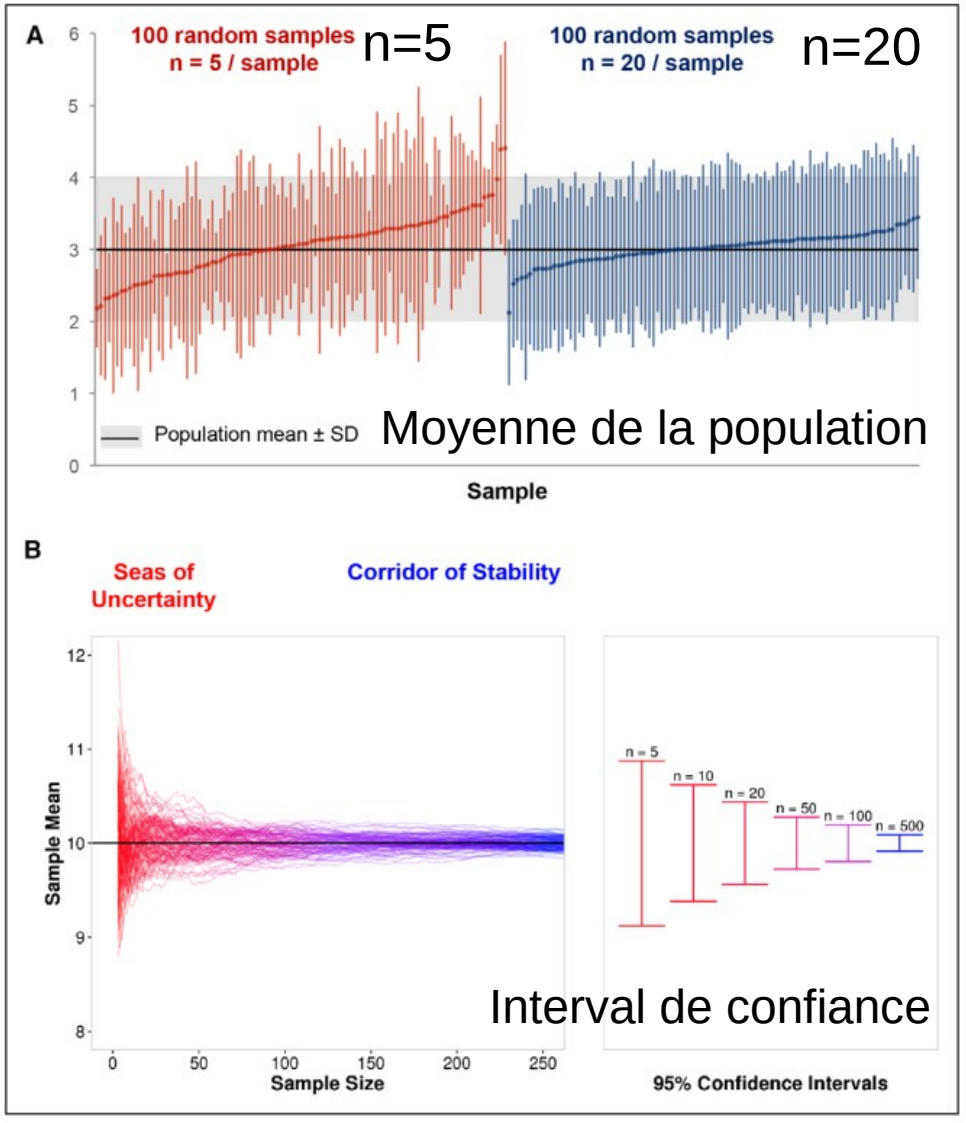
<30 selon les livres de stat

La plupart des études utilisent entre
3 et 6 échantillons en biologie expérimentale
Parfois 10 ou 15…

Compromis: temps, coût, bras

Pas de bras?

??



amary statistics are only meaningful when there are enough data to summarize

Moyenne+/- écart-type (SD)
=> variation des données

Moyenne +/- erreur standard (SE)
=>exactitude de la moyenne

SE=SD/√(n)

effectif grand=>incertitude faible

Attention à la comparaison de groupes de taille différente!=> SE

(Weissberger et al. 2015, 2019)

Intervalle de confiance de la moyenne de la population

Frontière de probabilité: (C1 ≤ moyenne de la population ≤ C2)

Hyp: la distribution suit une loi normale

Calcul de la marge d'erreur :

$Z_{a/2}$ x SD/√(n).

$Z_{a/2}$ coefficient de confiance,

avec a = degré de confiance, SD = écart type et n = taille de l'échantillon

Interval de confiance

$\overline{x} \pm Z_{a/2}$*SD/√(n).

Z Selon la table de l'écart réduit

$\overline{x}$ représente la moyenne de l'échantillon

À 99% (a=1%) $\overline{x}$ ± 2.6 x SE (2.576)

À 95% (a=5%)) $\overline{x}$ ± 2 x SE (1.960)

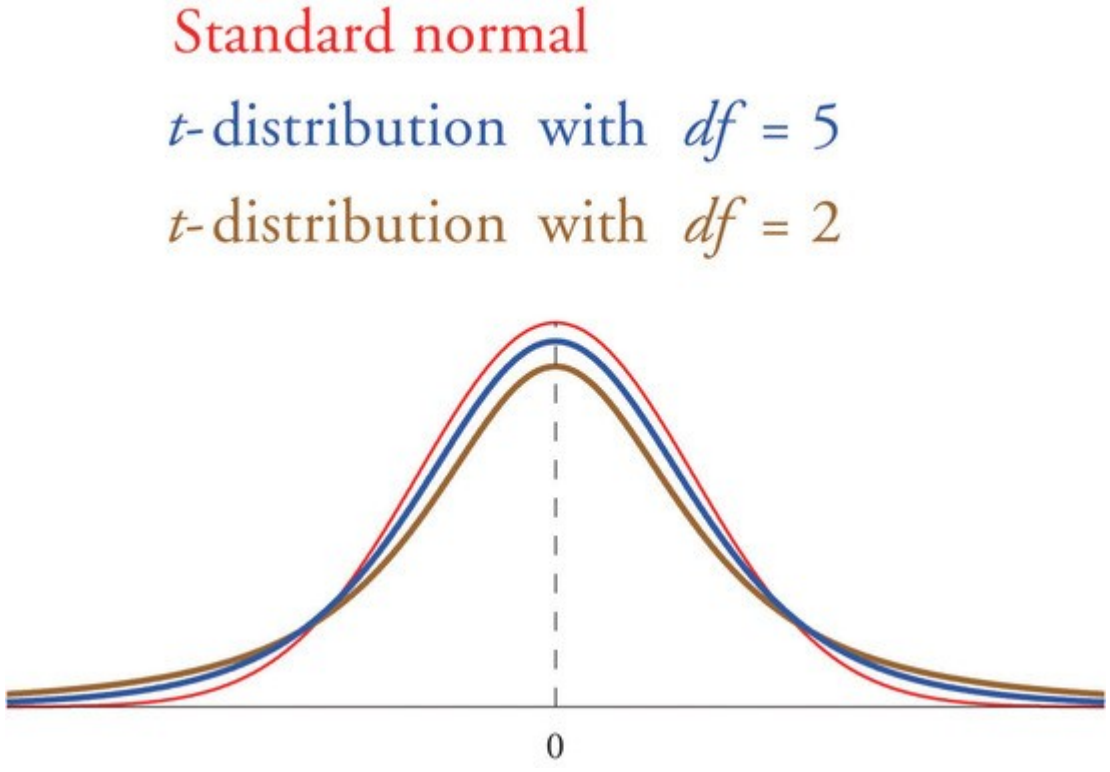
$\overline{x}$ ± SE => 68%

Interval de confiance pour un petit échantillon?

hyp: les données ont une distribution normale
=> distribution t de Student avec n-1 ddl

Centrée sur zero
mais
Plus applatie, plus large que la distribution normale

=> pas besoin de connaitre le SD de la population



Interval de confiance de la moyenne de la population pour un petit échantillon?

distribution t de Student => le facteur diminue avec l'effectif

n	Confidence Level	Multiplicative Factor
2	0.95	12.71
3	0.95	4.30
4	0.95	3.18
5	0.95	2.78
infinity	0.95	1.96
2	0.99	63.66
3	0.99	9.92
4	0.99	5.84
5	0.99	4.60
infinity	0.99	2.58

Si n=2, SD= $|(X_1-X_2)|/\sqrt{(2)}$

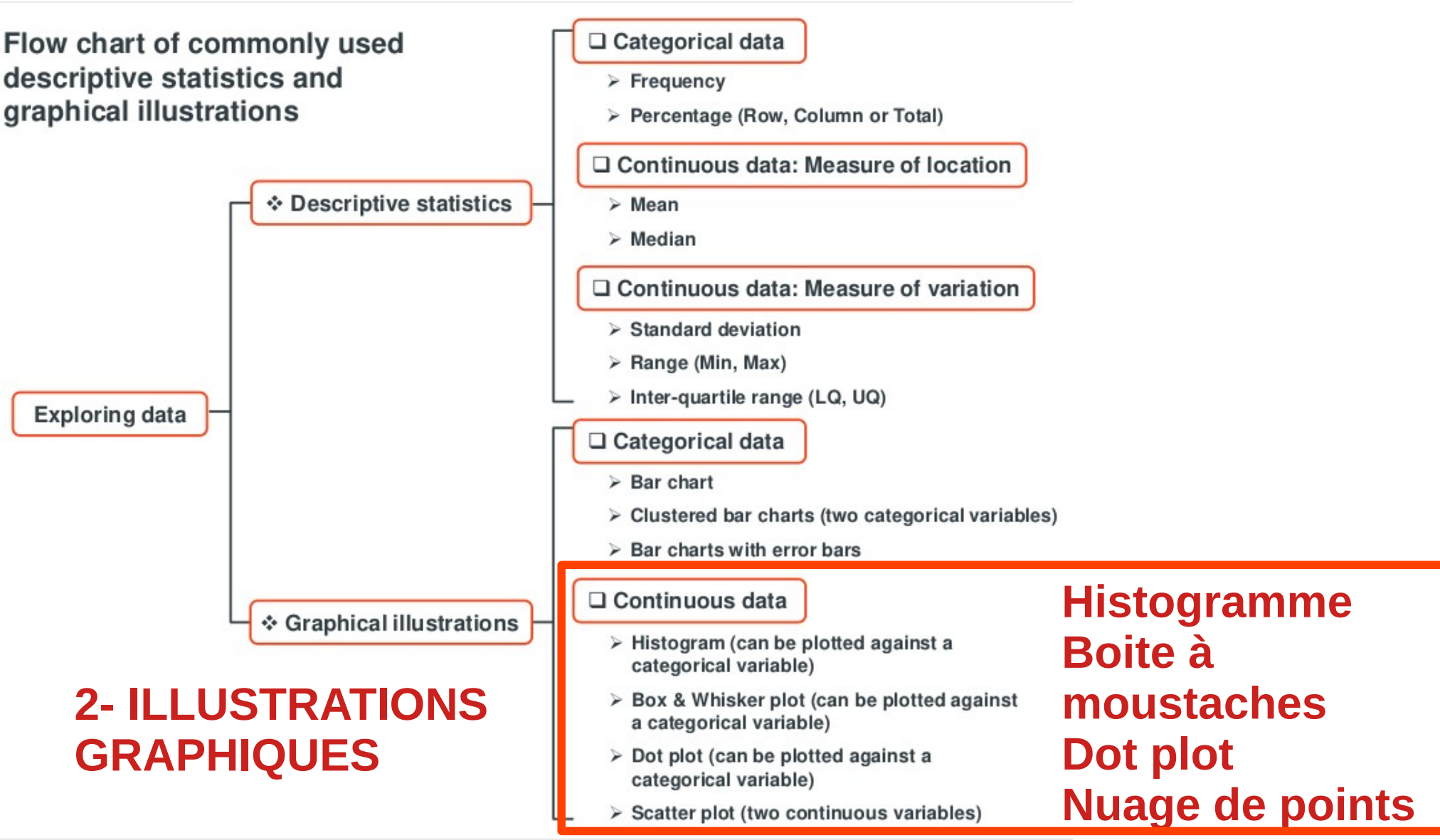
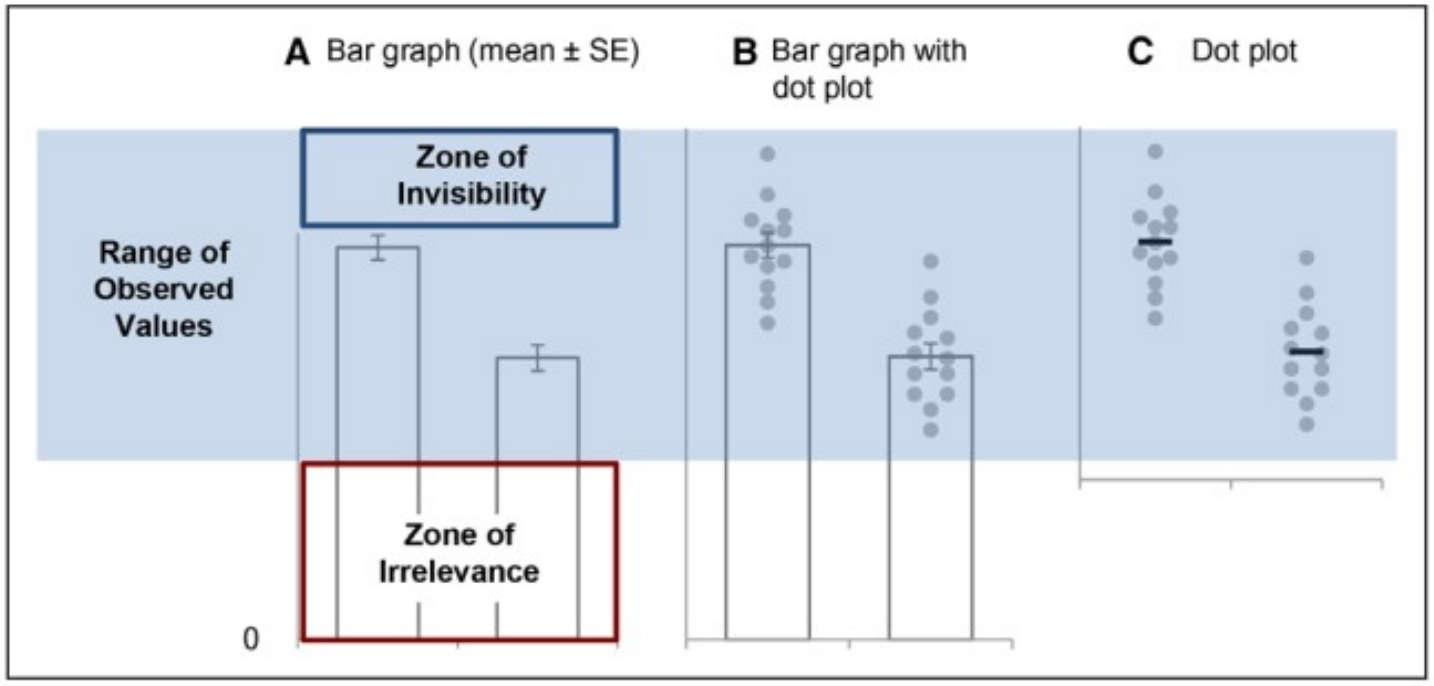

Flow chart of commonly used descriptive statistics and graphical illustrations
Exploring data
Descriptive statistics
Categorical data
Frequency
Percentage (Row, Column or Total)
Continuous data: Measure of location
Mean
Median
Continuous data: Measure of variation
Standard deviation
Range (Min, Max)
Inter-quartile range (LQ, UQ)
Graphical illustrations
Categorical data
Bar chart
Clustered bar charts (two categorical variables)
Bar charts with error bars
Continuous data
Histogram (can be plotted against a categorical variable)
Box & Whisker plot (can be plotted against a categorical variable)
Dot plot (can be plotted against a categorical variable)
Scatter plot (two continuous variables)
2- ILLUSTRATIONS GRAPHIQUES
Histogramme
Boite à moustaches
Dot plot
Nuage de points

Diagramme en barres



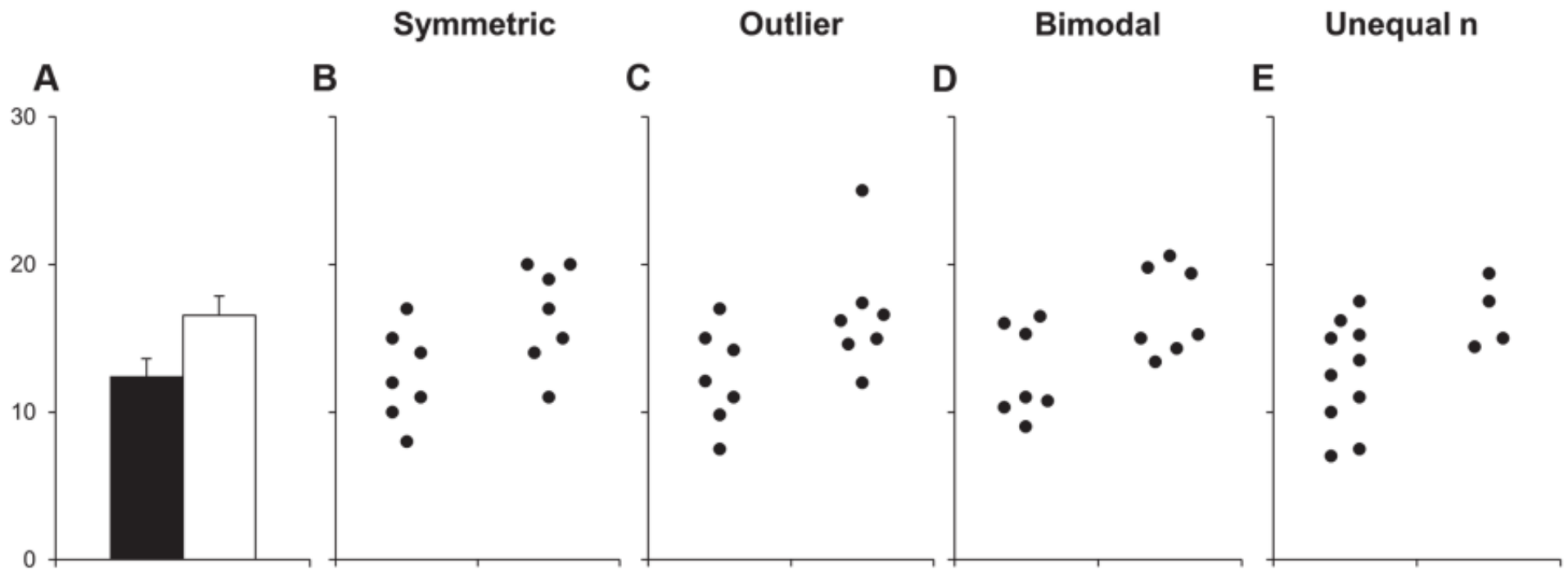
(Weissgerber et al. 2019)

Pas pour les données continues!

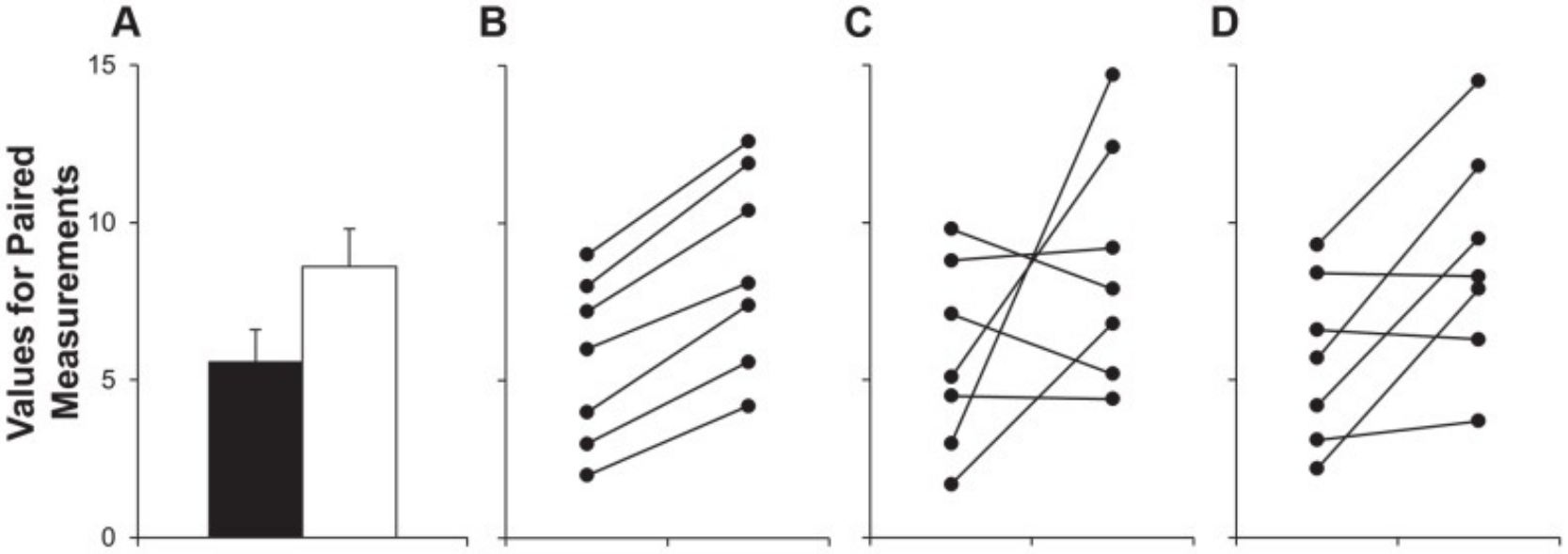
OK pour les données qualitatives

=> On ne peut pas évaluer vraiment les données

=> distribution, outliers??

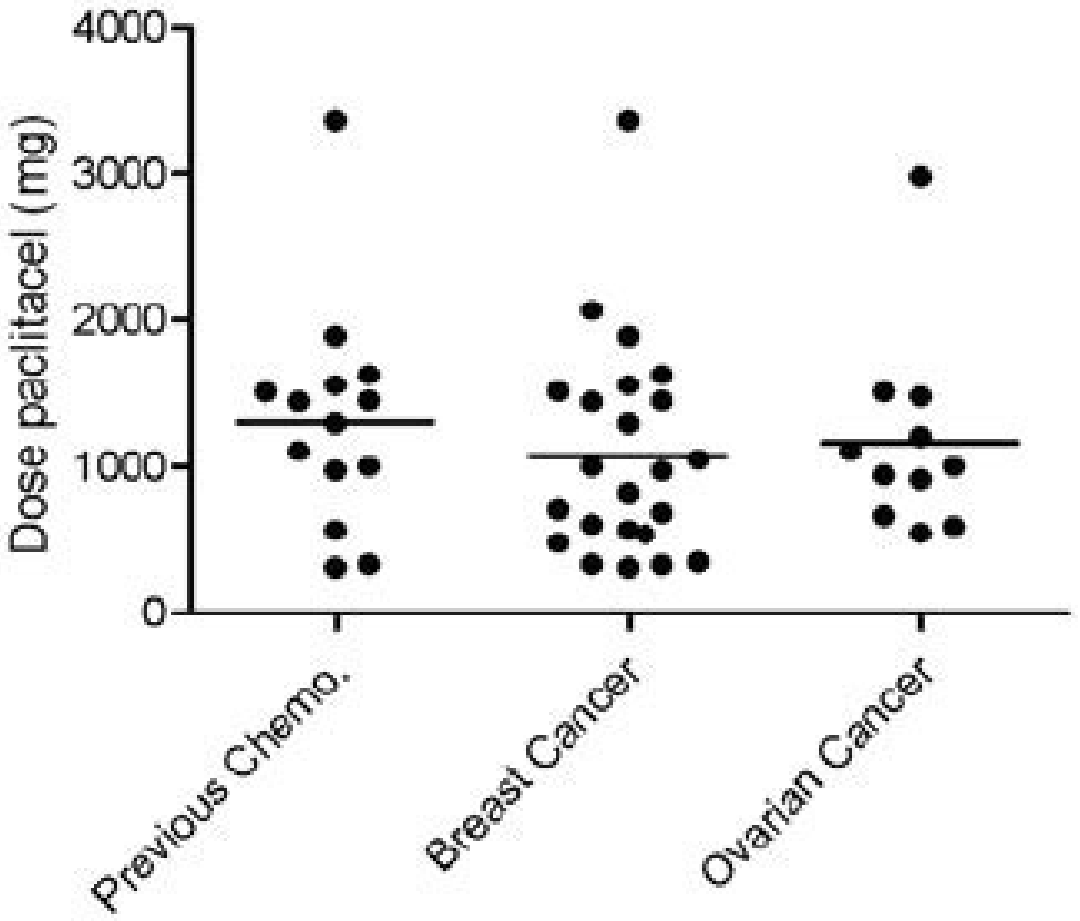


Différentes distributions => le même diagramme en barres….



Pas assez informatif pour les données appariées
=> le même graphique pour différents designs expérimentaux

Représentation par points: Dot plot

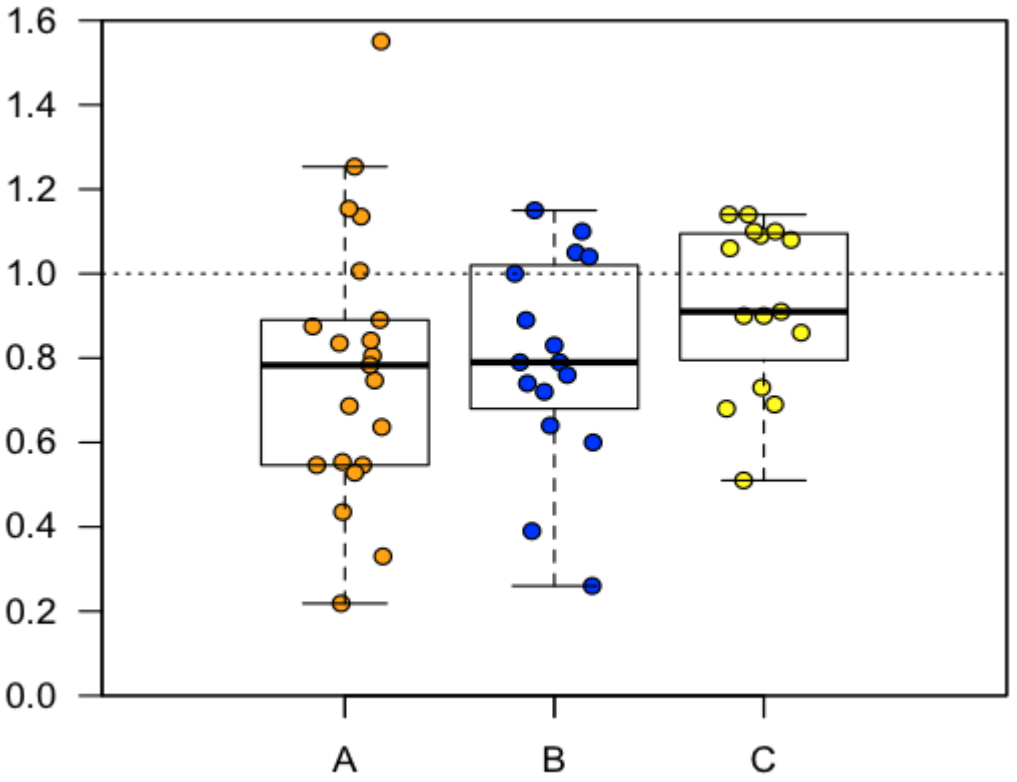


Adapté pour les petits échantillons

Ou forte hétérogénéité (outliers…)

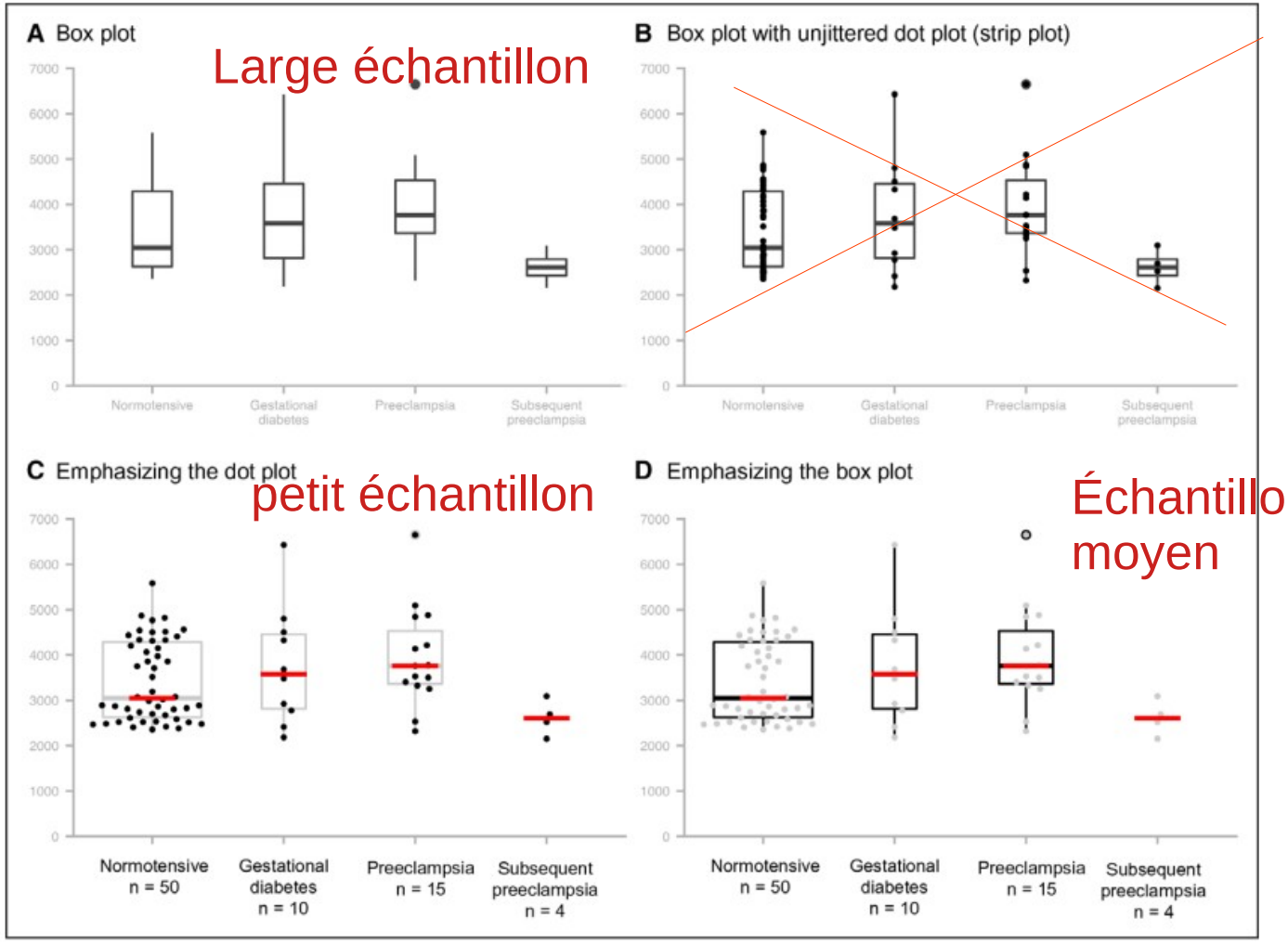
=> N'importe quelle distribution
=> Éparpiller les points
=> Moyenne ou médiane

Points et boites à moustache : Dot plot+ boxplot (ou violin plot)



Adapté pour les échantillons de taille moyenne

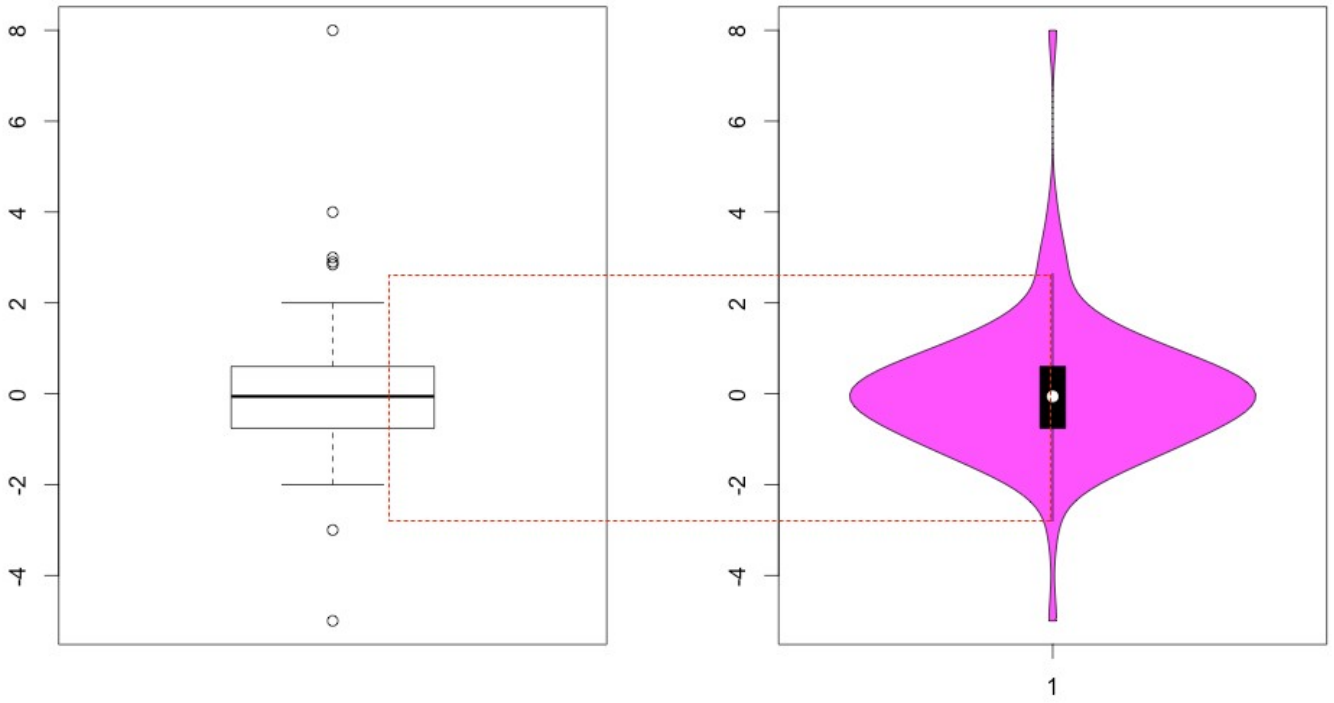
=> N'importe quelle distribution
=> Éparpiller les points
=>Médiane et quartiles
=>Spécifier dans la légende ce que représentent les moustaches



Adapter à la taille de l'échantillon

Et éparpiller les points

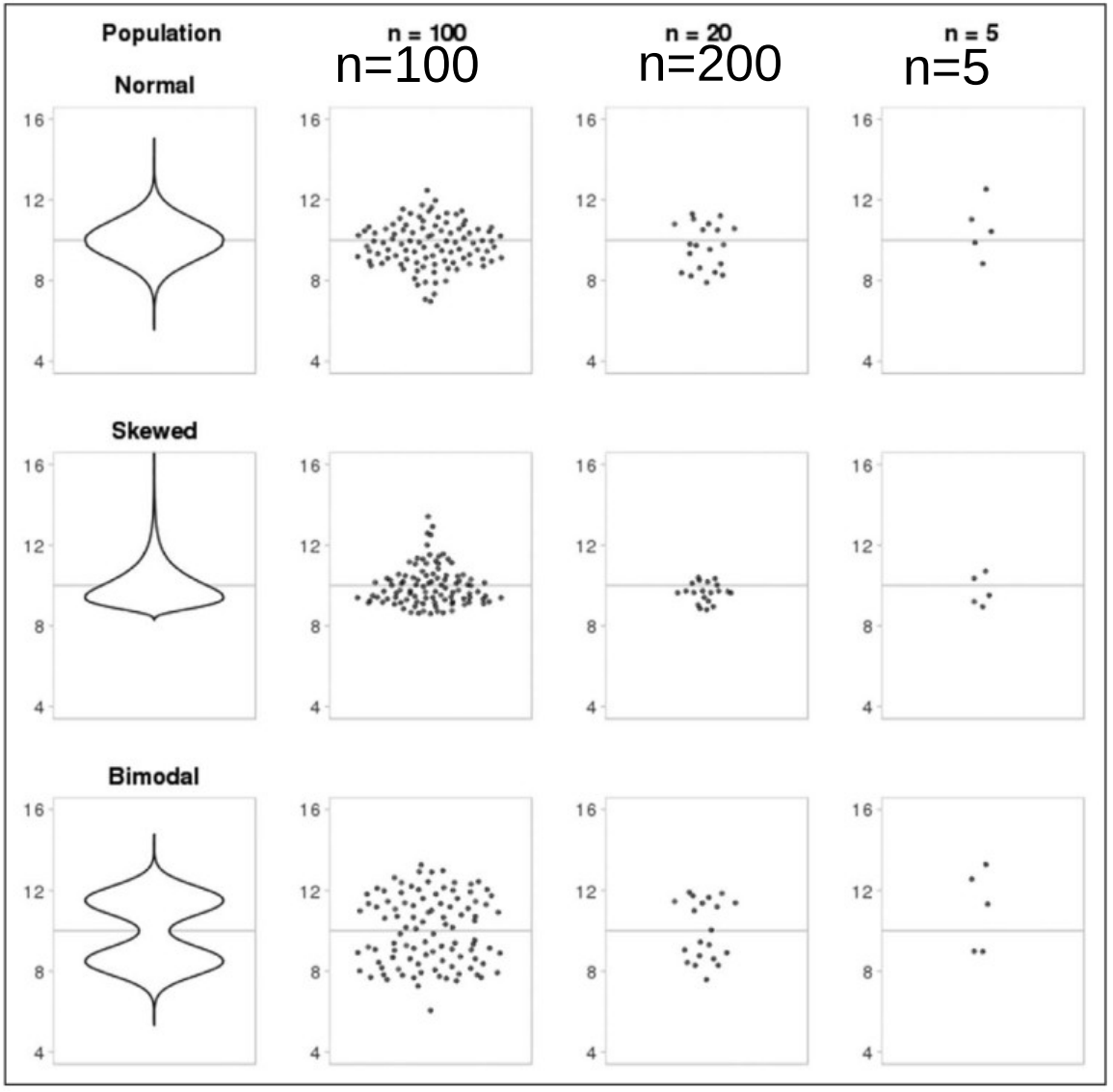
Diagramme de violon: Violin plot



Taille d’échantillon large

=> Donne une vue de la distribution

=>N’importe quelle distribution



Violin plot
Échantillon large
=> on voit la forme de la distribution

Petit échantillon=>
rien

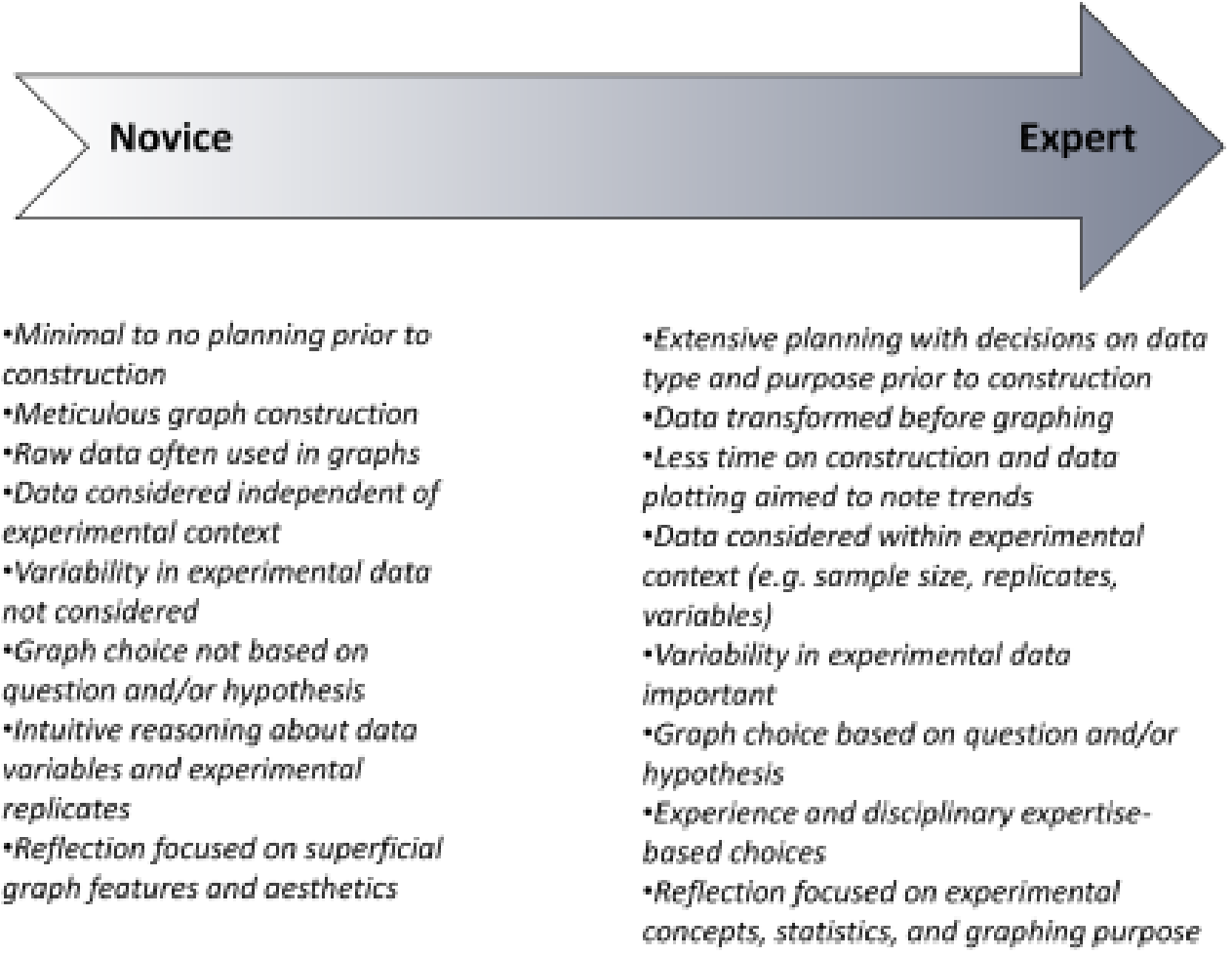
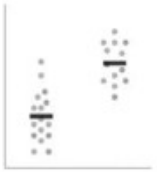
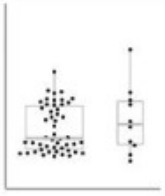
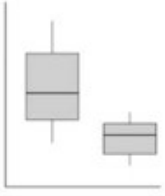
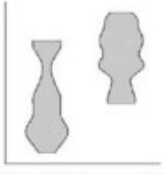
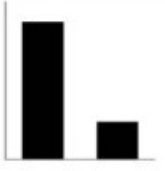


FIGURE 4. Visual summary of graph-construction reasoning, graphing behavior, and graph attribute findings with the reasoning behind graph choice and construction, graphing behaviors, and graph attributes along the novice to expert continuum.

Choix et construction des graphes par les biologistes
Un objet d'étude pour les SHS...

(Angra et Gardner 2017)

Figure Types	Example	Type of Variable	What the Plot Shows	Sample Size	Data Distribution	Best Practices
Dot plot		Continuous	Individual data points & mean or median line Other summary statistics (i.e. error bars) can be added for larger samples	Very small OR small; can also be useful with medium samples	Sample size is too small to determine data distribution OR Any data distribution	• Make all data points visible - use symmetric jittering • Many groups: Increase white space between groups, emphasize summary statistics & de-emphasize points • Only add error bars if the sample size is large enough to avoid creating a false sense of certainty • Avoid "histograms with dots"
Dot plot with box plot or violin plot		Continuous	Combination of dot plot & box plot or violin plot (see descriptions above and below)	Medium	Any	• Make all data points visible (symmetric jittering) • Smaller n: Emphasize data points and de-emphasize box plot, delete box plot and show only median line for groups with very small n • Larger n: Emphasize box plot and de-emphasize points
Box plot		Continuous	Horizontal lines on box: 75th, 50th (median) and 25th percentile Whiskers: varies; often most extreme data points that are not outliers Dots above or below whiskers: outliers	Large	Do not use for bimodal data	• List sample size below group name on x-axis • Specify what whiskers represent in legend
Violin plot		Continuous	Gives an estimated outline of the data distribution. The precision of the outline increases with increasing sample size.	Large	Any	• List sample size below group name on x-axis • The violin plot should not include biologically impossible values
Bar graph		Counts or proportions	Bar height shows the value of the count or proportion	Any	Any	• **Do not use for continuous data**