

Éléments de compréhension des statistiques

Jeffery P.

Doctorant au Laboratoire des Sciences du Numérique de Nantes (LS2N)

2019

Crédits

- ▶ Ce support est inspiré du cours de M. Jean-Philippe Babin, responsable pédagogique de Licence à l'Université de Nantes - Laboratoire de Psychologie des Pays de la Loire.
- ▶ Également, plusieurs formulations ont pu être améliorées grâce aux commentaires avisés de M. Paul Marti et M. Damien Schnebelen, tous deux doctorants au LS2N - pôle SIEL, team PACCE.

Statistiques descriptives

Préambule

La statistique descriptive nous permet de décrire l'ensemble des données recueillies

- ▶ Individu statistique : unité élémentaire pour laquelle on va recueillir les données
- ▶ Population : ensemble des individus statistiques
- ▶ Échantillon : sous-ensemble de la population
- ▶ Variable ou caractère : aspect auquel on s'intéresse

→ l'effectif est le nombre d'individus statistiques auxquels on s'intéresse, on le note en général N

Variables et Catégories

Une variable peut prendre un nombre fini ou infini de valeurs que l'on nomme **modalités**

Il existe deux grands types de variable :

- ▶ Les variables **qualitatives** : les modalités sont des mots/expressions
- ▶ Les variables **quantitatives** : les modalités sont des nombres entiers ou décimaux

Pour chacun de ces deux grands types, il existe deux sous-catégories. . .

Les variables qualitatives

Peut-on ordonner (hiérarchiser) les modalités ?

- ▶ oui : il s'agit d'une variable **qualitative ordinale** (e.g., type de logement {appartement/maison})
- ▶ non : il s'agit d'une variable **qualitative nominale** (e.g., fréquentation {rare < occasionnelle < souvent})

Les variables quantitatives

Peut-on énumérer les modalités sans en omettre ?

- ▶ oui : il s'agit d'une variable **quantitative discrète** (e.g., âge en année(s) [1, 2, 3, 4, 5, ...])
- ▶ non : il s'agit d'une variable **quantitative continue** (e.g., taille en cm (précision supposée infinie) [toute valeur supérieure à 0])

Remarque :

- ▶ En général l'énoncé et le contexte nous guident pour déterminer si on est dans le cas discret ou continu
- ▶ Dans les deux cas, il peut exister une infinité de valeur
- ▶ Une erreur fréquente est de considérer qu'une variable discrète a forcément des modalités entières !

Les variables quantitatives regroupées en classe

Lorsqu'on étudie une variable quantitative disposant d'un trop grand nombre de modalités, il est souvent utile de faire un découpage en classe, i.e. on ordonne les modalités et on opère des regroupements (de même amplitude si possible)

→ on appelle cela **regroupement en classe**

- ▶ Un regroupement en classe implique forcément une perte d'information
- ▶ Une variable pour laquelle on aura effectué un regroupement en classe nécessitera un traitement différent (calcul du mode, de la moyenne, de l'écart-type, etc.)

Les variables quantitatives regroupées en classe

- ▶ Une classe est définie par sa borne inférieure (B_{inf}) et sa borne supérieure (B_{sup})
- ▶ Les classe sont, dans la majorité des cas, des intervalles fermés à gauche et ouvert à droite i.e., une classe contient sa plus petite valeur (B_{inf}) mais pas sa plus grande (B_{sup}), on note $[B_{inf}; B_{sup}[$

Par exemple $[0; 20[$ mais pas $[0; 20]$ ou $]0; 20[$

- ▶ Pour chacune des classe de regroupement, on nomme **amplitude** la longueur de l'intervalle correspondant soit $B_{sup} - B_{inf}$. On la note souvent a

Par exemple un intervalle $[0; 20[$ à une amplitude de 20

- ▶ À chacune des classes, on associe un effectif souvent noté n et on nomme **densité** la valeur $\frac{n}{a}$

Première étape de description

La première fois que l'on aborde un problème il faut nécessairement trouver les réponses aux questions suivantes :

- ▶ Quelle est la population étudiée ?
- ▶ Qu'est-ce qu'un individu statistique ?
- ▶ Quel est l'échantillon ? quel est son effectif ?
- ▶ Quelles sont les variables ? pour chacune, à quelle catégorie appartient-elle et quelles sont les modalités ?

Exemple simplifié

Cas général

On mène une étude sur l'accès à internet en agglomération nantaise. Pour cela on enquête auprès de 1000 foyers judicieusement choisis qui ont accès à internet. Pour chacun, on demande leur débit (MB/s) ainsi que leur opérateur.

- ▶ Population : ensemble des foyers en agglomération nantaise qui ont accès à internet
- ▶ Individu statistiques : un foyer nantais ayant accès à internet
- ▶ Échantillon : 1000 foyers nantais ayant accès à internet
- ▶ Variables :
- ▶ Débits : variable quantitative discrète {1MB/s, 2MB/s, ...}
- ▶ Opérateur : variable qualitative nominale {orange, SFR, Free, Bouygues Télécom}

Exemple simplifié

Remarques

Le débit peut tout aussi bien être une variable quantitative discrète regroupée en classe.

- ▶ On peut faire des classes d'amplitude égale à 20 ($a = 20$) :
 $\{[0; 20[, [20; 40[, [40; 60[, \dots]\}$
- ▶ On peut faire des classes d'amplitude inégales :
 $\{[0; 20[, [20; 80[, [80; 150[, \dots]\}$

Deuxième étape de description

La deuxième étape de description concerne une approche par variable. Pour chacune d'elle nous allons disposer de données recueillies pour les individus statistiques de l'échantillon

Il est souvent utile d'aborder chacune des variables en déterminant si l'on dispose de données brutes ou non ?

→ Pour une variable, les données sont dites **brutes** si on dispose d'une valeur pour chacun des individus statistiques, sinon on dit qu'elles sont **traitées**

Cette question n'est pas anodine, car des données déjà traitées (e.g., listées en tableau ordonné) orientent le lecteur vers des premières conclusions

Données brutes vs. traitées (exemple)

Données brutes :

Individu	Opérateur	Débit
Foyer 1	Free	20
Foyer 2	Orange	10
⋮	⋮	⋮
Foyer 1000	SFR	75

Données pré-traitées :

Opérateur	Nombre de foyers	Débit moyen
Free	95	20
Orange	330	10
⋮	⋮	⋮
B&You	43	75

Données brutes vs. traitées (exemple)

Données pré-traitées avec regroupement en classe pour le débit. . .

Opérateur	Nombre de foyers	Débit moyen
Free	95	[20 ; 40[
Orange	330	[0 ; 20[
SFR	120	[140 ; 160[
⋮	⋮	⋮
B&You	43	[60 ; 80[

Présentation des données en tableau

Pour présenter une variable, nous introduisons quelques notations :

- ▶ (rappel) N : effectif total i.e., nombre d'individus statistiques
- ▶ p : nombre de modalités associés à la variable
- ▶ n_i : effectif d'une modalité
- ▶ n_{ic} : effectif cumulé i.e, somme cumulée des n_i
- ▶ f_i ou $\%n_i$: pourcentage d'effectif d'une modalité ($f_i = \frac{n_i}{N}$)
- ▶ f_{ic} ou $\%n_{ic}$: pourcentage cumulé ($f_{ic} = \frac{n_{ic}}{N}$)

Quantité	modalité 1	modalité 2	...	modalité p	Total
n_i	2	13		12	N
n_{ic}	2	15		N	
f_i	$\frac{2}{N}$	$\frac{13}{N}$		$\frac{12}{N}$	1
f_{ic}	$\frac{2}{N}$	$\frac{15}{N}$		1	1

Calcul du mode

Le mode est associé à une variable, il correspond à la **modalité** ayant l'effectif maximum. Il existe quelque soit la catégorie d'une variable

- ▶ Une erreur fréquente est de confondre mode et effectif associé !
- ▶ Pour une variable quantitative regroupée en classe, le mode correspond à la classe de densité maximale ($\frac{n}{a}$). Si toutes les classes sont d'amplitudes égales, le mode sera alors la classe d'effectif maximum

Calcul du mode : exemple

Imaginons une variable avec trois modalités, et un effectif total $N = 12$:

Quantité	modalité 1	modalité 2	modalité 3
n_i	2	3	7

- Le mode est donc la modalité 3

Calcul de la médiane

La médiane est la modalité qui sépare la population en deux groupes d'effectifs égaux, on la note m_e . En pratique :

1. On calcule le range médian i.e., le rang pour lequel on a autant d'individus au dessus qu'en dessous :

$$range(m_e) = (N + 1)/2$$

2. On recherche la modalité qui contient le rang médian

Calcul de la médiane : remarques

- ▶ Le rang médian ne correspond pas tout le temps à un entier. Si l'effectif total est pair le rang médian sera toujours décimal, il correspondra à un individu fictif
- ▶ Une erreur fréquente est de confondre la médiane avec le rang médian associé !
- ▶ Il est souvent pratique d'organiser les données en tableau pour déterminer la médiane, on cherche alors le rang sur **la ligne d'effectif cumulé**

Calcul de la médiane : exemple

Imaginons une variable avec trois modalités, et un effectif total $N = 12$:

Quantité	modalité 1	modalité 2	modalité 3
n_i	2	3	7
n_{ic}	2	5	12

- ▶ Le rang médian est alors
 $\text{rang}(m_e) = (N + 1)/2 = (12 + 1)/2 = 13/2 = 6,5$
- ▶ La médiane est donc la modalité 3

Calcul de la moyenne

C'est une modalité (possiblement fictive) moyenne on la note \bar{x} . Elle n'est pertinente **que pour les variables quantitatives** et se calcule différemment en fonction de 3 situations :

1. Les données sont brutes :

$$\bar{x} = \frac{\sum x_i}{N}$$

2. Les données sont présentées en tableau avec effectifs :

$$\bar{x} = \frac{1}{N} \sum x_i \times n_i$$

3. Les données sont présentées en tableau et la variable a été regroupée en classe :

$$\bar{x} = \frac{1}{N} \sum c_i \times n_i$$

Notations

- ▶ x_i représente une modalité
- ▶ $c_i = (B_{inf}^i + B_{sup}^i)/2$ est le centre d'une classe i

Calcul de la moyenne : exemple situation 1

Imaginons un échantillon de trois individus, pour lesquels on dispose du poids (kg) :

Individu	1	2	3
Modalité	68	76	72

- La moyenne est alors $\bar{x} = \sum x_i / N = (68 + 76 + 72) / 3 = 72$

Calcul de la moyenne : exemple situation 2

Imaginons un échantillon de 12 individus, pour lesquels on dispose du poids (kg) :

Poids	68	72	76
n_i	5	2	5

- La moyenne est alors

$$\bar{x} = \sum \frac{x_i \times n_i}{N} = (68 \times 5 + 72 \times 2 + 76 \times 5) / 12 = 72$$

Calcul de la moyenne : exemple situation 3

Imaginons un échantillon de 12 individus, pour lesquels on dispose du poids (kg) :

Poids	[68 ; 72[[72 ; 76[[76 ; 80[
c_i	70	74	78
n_i	5	2	5

► La moyenne est alors

$$\bar{x} = \sum \frac{c_i \times n_i}{N} = (70 \times 5 + 74 \times 2 + 78 \times 5) / 12 = 74$$

Le quantilage

Le quantilage est le découpage de la population en proportions égales. Il s'effectue nécessairement pour des **variables quantitatives**

Par exemple un effectif de 12 observations peut être divisé :

- ▶ En deux groupes de 6 modalités (quantiles d'ordre 6)
- ▶ En trois groupes de 4 modalités (quantiles d'ordre 4 nommés **quartiles**)
- ▶ ...

→ Les quantiles ne font pas nécessairement référence à des groupes entiers, c'est même rarement le cas. On peut vouloir découper un effectif de 12 en 10 groupes (quantiles d'ordre 10 nommés **déciles**)

Le quantilage

Les **quantiles** sont des modalités (possiblement non observées) qui correspondent à des séparation de l'ensemble des modalités observées.

Un quantile est toujours précisé avec deux indices :

- ▶ p : le nombre total de groupes
- ▶ r : l'index de la séparation avec $1 \leq r \leq p - 1$

Si on note Q_p^r le quantile r d'ordre p pour une variable, alors :

Il y a une proportion au moins égale à $\frac{r}{p}$ d'observation
 $\leq Q_p^r$ Et une proportion au moins égale à $1 - \frac{r}{p}$
d'observation $\geq Q_p^r$

Le quantilage

Remarques

- ▶ L'ordre p d'un quantile correspond au nombre de groupes que l'on fait
- ▶ Il y a toujours $p - 1$ quantiles d'ordre p
- ▶ Les quantiles sont ordonnés $Q_p^1 < Q_p^2 < \dots < Q_p^{p-1}$
- ▶ Bien retenir que :
 - ▶ les quantiles d'ordre 4 sont appelés **quartiles** (Q_4)
 - ▶ les quantiles d'ordre 10 sont nommée **déciles** (Q_{10})
 - ▶ les quantiles d'ordre 100 sont nommés **centiles** (Q_{100})
- ▶ Le deuxième quartile correspond à la médiane ($m_e = Q_4^2$)

Le quantilage en pratique

En pratique, on procède un peu comme pour la médiane. Il y a deux étapes :

1. On commence par déterminer le rang du quantile recherché avec la formule suivante :

$$\text{rang}(Q_p^r) = r \times \frac{N + 1}{p}$$

2. On recherche la modalité associé au rang (c'est là que ça peut se corser !)

La recherche d'une modalité associé à un rang

Cette recherche peut arriver dans deux grands cas souvent fréquents :

- ▶ On dispose des données brutes mais le rang n'est pas entier
- ▶ Les données ont été regroupées en classe

On utilise alors la méthode d'**interpolation linéaire**

L'interpolation linéaire sur données brutes

Facile ! Imaginons qu'on obtienne un rang égal à 2,7

Supposons d'autre part que l'on dispose des 3 observations ordonnées suivantes : 4, 10, 18 ...

- Il ne serait pas logique de privilégier la modalité 10 plutôt que 18 (et inversement) → on imagine bien que la modalité associée au rang 2.7 se situerait entre 10 et 18 mais plus prêt de 18 quand même...

On utilise alors le calcul d'interpolation suivant :

$$10 + (2,7 - 2) \times (18 - 10) = 15,6$$

L'interpolation linéaire sur données brutes

Le calcul énoncé :

$$10 + (2,7 - 2) \times (18 - 10) = 15,6$$

Correspond à la formule générale :

$$V_i + (\tilde{i} - i) \times (V_{i+1} - V_i)$$

où :

- ▶ V_i et V_{i+1} : sont les valeurs des rangs entiers i et $i + 1$
- ▶ \tilde{i} : est le rang du quantile recherché tel que $i < \tilde{i} < i + 1$

Remarque :

Bien que le résultat obtenu soit plus précis avec l'interpolation sur données brutes, on ne fait pas toujours ce calcul. . .souvent, on se contente de faire la moyenne entre V_i et V_{i+1}

L'interpolation linéaire sur données regroupées en classe

Supposons maintenant que l'on recherche le quantile associé au rang 8 et que l'on ait les observations suivantes :

Poids	[68 ; 72[[72 ; 76[[76 ; 80[
n_i	5	4	7
n_{ic}	5	9	16

- On sait d'après la ligne des effectifs cumulés (n_{ic}) que l'individu 8 se situe dans la classe [72; 76[mais où exactement ?

L'interpolation linéaire sur données regroupées en classe

Poids	[68 ; 72[[72 ; 76[[76 ; 80[
n_i	5	4	7
n_{ic}	5	9	12

→ L'idée est de se dire que cette classe contient 4 observations et que ces dernières sont réparties *uniformément* de manière ordonnée entre 72kg et 76kg. Cela signifie qu'ils ne sont pas tous proches de 72kg, ni tous proches de 76kg mais qu'au contraire, il y en a proches de 72kg d'autres proches de 76kg et d'autres plus proches de 74kg (centre de la classe).

En suivant cette idée l'individu 8 serait alors l'avant dernière valeur de cette classe, donc relativement proche de 76kg. . .

L'interpolation linéaire sur données regroupées en classe

Poids	[68 ; 72[[72 ; 76[[76 ; 80[
n_i	5	4	7
n_{ic}	5	9	12
n'_{ic}	1, 2, 3, 4, 5	6, 7, 8, 9	...

On utilise alors le calcul suivant :

$$72 + (8 - 5) \times \frac{(76 - 72)}{4} = 75kg$$

- ▶ On part de la borne inférieure (B_{inf}^i) de la classe, ici égale à 72
- ▶ On cherche le 8ème individu en sachant qu'il y a 5 observations $\leq B_{inf}^i$
- ▶ L'amplitude (a_i) de la classe est initialement divisée en 4 (n_i) proportions

L'interpolation linéaire sur données regroupées en classe

Avec les notations, le calcul :

$$72 + (8 - 5) \times \frac{(76 - 72)}{4}$$

Est équivalent à :

$$B_{inf}^i + (\tilde{i} - n_{(i-1)c}) \times \frac{(B_{sup}^i - B_{inf}^i)}{n_i} = B_{inf}^i + (\tilde{i} - n_{(i-1)c}) \times \frac{a_i}{n_i}$$

Où :

- ▶ \tilde{i} est le rang du quantile recherché tel que $n_{(i-1)c} < \tilde{i} < n_{ic}$
- ▶ n_i est l'effectif de la classe i
- ▶ B_{inf}^i et B_{sup}^i sont les bornes de la classe i
- ▶ a_i est l'amplitude de la classe i
- ▶ $n_{(i-1)c}$ est l'effectif cumulé de la classe inférieure

L'interpolation linéaire sur données regroupées en classe : remarques

- ▶ Le calcul précédent est intuitif (non ?)
- ▶ Le calcul reste valable peu importe la nature du rang recherché, qu'il soit entier ou décimal
- ▶ L'interpolation linéaire nécessite les deux types de valeur d'effectif (non cumulé et cumulé)

Digression

- ▶ Le calcul de la médiane est une interpolation linéaire

Boîtes de dispersion

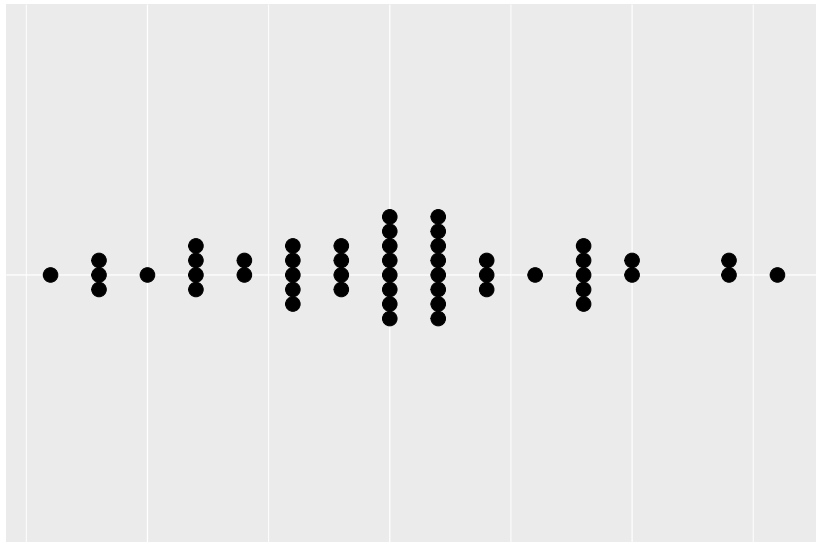
En statistique, il est souvent utile de visualiser les distributions.

On peut par exemple tracer des **boîtes de dispersion** (en anglais **boxplot**). Pour cela, on représente :

- ▶ Un rectangle délimité par les 1^{er} et 3^{ème} quartiles (Q_1 et Q_3)
- ▶ La médiane (Q_2) à l'intérieur de ce rectangle
- ▶ Un segment inférieur jusqu'au 1^{er} décile (D_1), ou jusqu'au minimum
- ▶ Un segment supérieur jusqu'au 9^{ème} décile (D_9), ou jusqu'au maximum

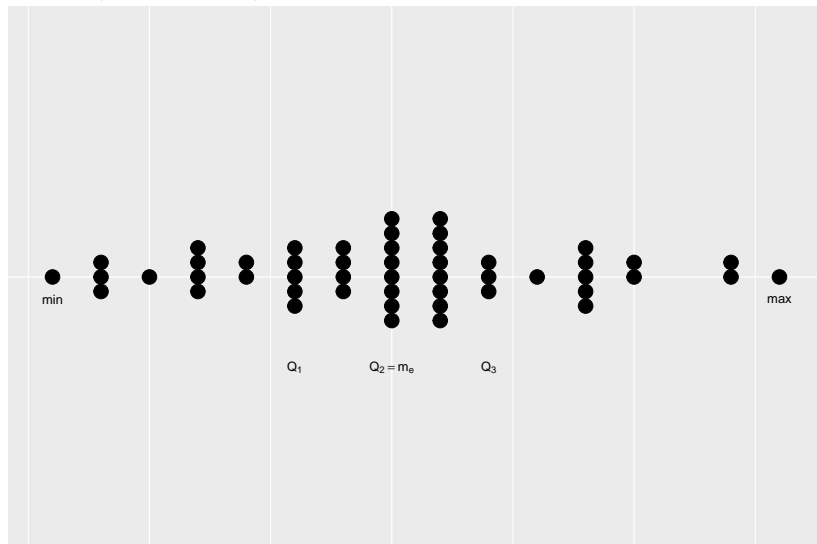
Boîtes de dispersion

Données iris (Fisher's or Anderson's)



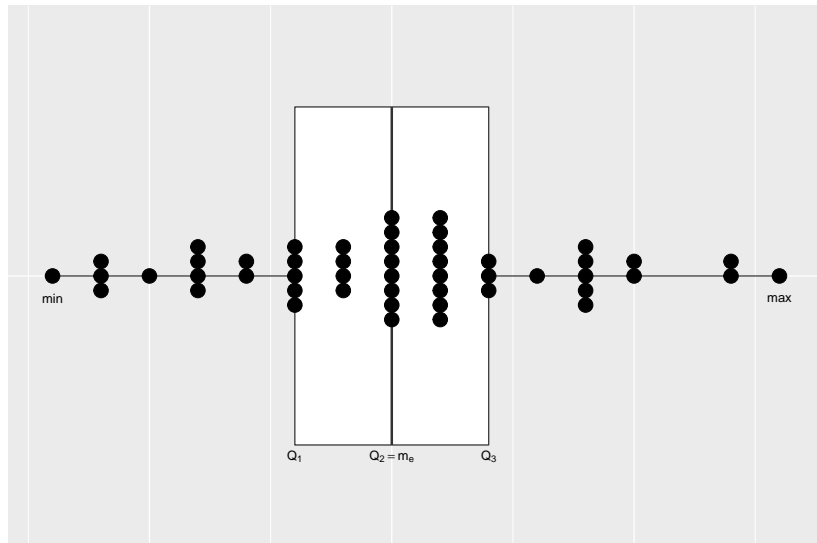
Boîtes de dispersion

Données iris (Fisher's or Anderson's)



Boîtes de dispersion

Données iris (Fisher's or Anderson's)



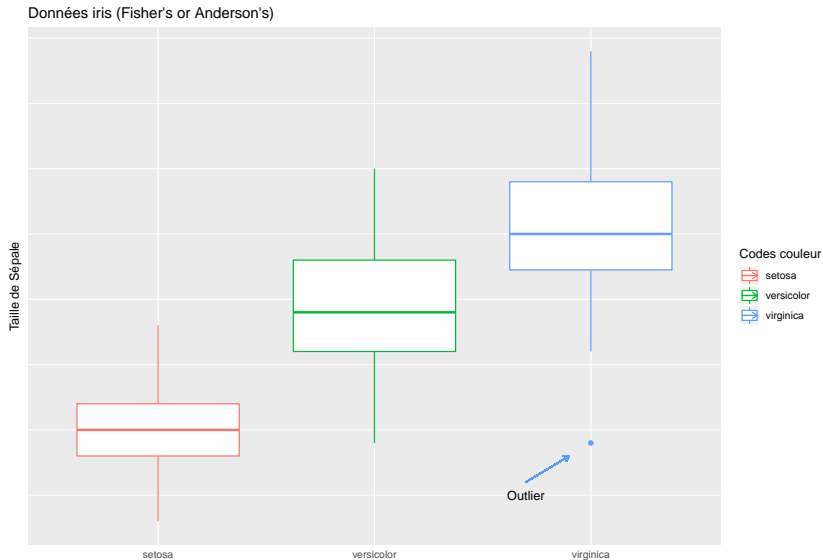
Boîtes de dispersion : quelques intérêts

Grâce aux boîtes de dispersion, on peut rapidement avoir une idée de la répartition des données

- ▶ Appréhender l'échelle
- ▶ Savoir si les données sont rassemblées autour de la médiane ou dispersées...

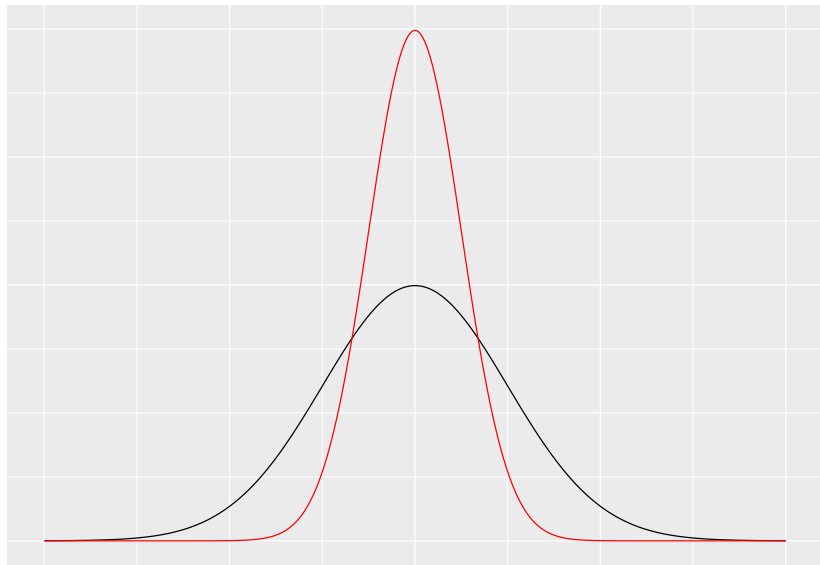
De plus, les boîtes de dispersion permettent de comparer plusieurs distributions entre elles !

Boîtes de dispersion



Mesure de dispersion

Que pensez vous des deux courbes suivantes ?



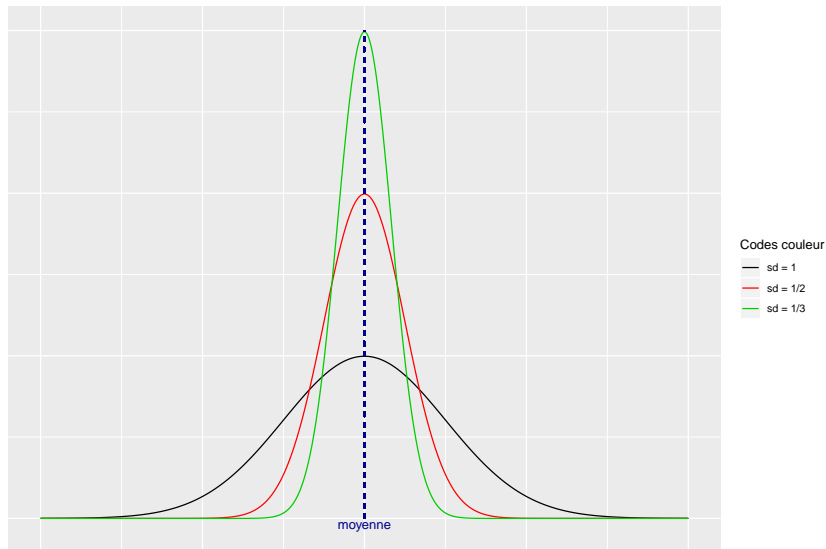
Mesure de dispersion

En statistique la dispersion se nomme **variance**. Il s'agit d'une valeur strictement positive qui ne se calcule que pour des variables **quantitatives**

- ▶ La dispersion traduit le fait que les données sont plus ou moins resserrées autour de la valeur moyenne \bar{x}
- ▶ Plus la variance est élevée, plus les données seront éloignées de leur moyenne (i.e., plus l'histogramme sera aplati)
- ▶ La variance est nulle si toutes les observations ont la même valeur

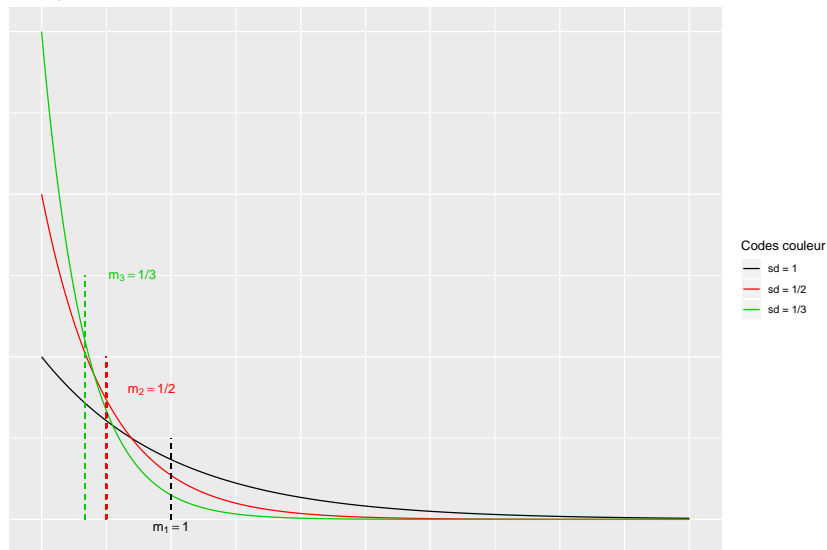
Mesure de dispersion : illustration

Lois normales



Mesure de dispersion : illustration

Lois exponentielles



Le calcul de la variance d'observation

La variance d'**observation** se note souvent σ^2 ou var. Elle se calcule différemment en fonction des cas :

1. Sur données brutes :

$$\sigma^2 = \frac{1}{N} \sum (x_i - \bar{x})^2$$

2. Sur données traitées avec effectif :

$$\sigma^2 = \frac{1}{N} \sum n_i (x_i - \bar{x})^2$$

3. Sur données traitées regroupées en classes :

$$\sigma^2 = \frac{1}{N} \sum n_i (c_i - \bar{x})^2$$

Éléments de compréhension de la variance

Exemple sur données brutes :

$$\sigma^2 = \frac{1}{N} \sum (x_i - \bar{x})^2$$

Pour obtenir la variance :

1. On calcule les écarts entre chaque observation et la valeur moyenne
2. On élève tous ces écarts au carré (puissance 2)
3. On calcule la moyenne des valeurs observées

Éléments de compréhension de la variance

On peut montrer que :

$$\sigma^2 = \frac{1}{N} \sum (x_i - \bar{x})^2 = \frac{1}{N} \sum x_i^2 - \left(\frac{1}{N} \sum x_i \right)^2$$

Ou encore :

$$\sigma^2 = \bar{x^2} - \bar{x}^2, \quad \text{où } \bar{x^2} = \frac{1}{N} \sum x_i^2$$

Donc :

$$\sigma^2 = \text{moyenne des carrés} - \text{carré de la moyenne}$$

Pourquoi élever au carré ?

Si on n'élevait pas les écarts au carré, les valeurs pourraient se compenser. Par exemple avec les observations suivantes
2, 3, 3, 4

► On a $\bar{x} := \frac{\sum x_i}{N} = \frac{2+3+3+4}{4} = 12/4 = 3$

Sans élever les écarts au carré, on trouverait :

$$\sigma^2 = \frac{1}{N} \sum (x_i - \bar{x}) = \frac{1}{4} ((2 - 3) + (3 - 3) + (3 - 3) + (4 - 3)) = 0$$

...et pourtant les observations n'ont pas toutes la même valeur

Le calcul de la variance d'échantillon

En statistique inférentielle on corrige la variance par le facteur $\frac{N}{N-1}$. On obtient ainsi une variance dite d'**échantillon**

En pratique, on applique les mêmes formules mais on divise par $N - 1$ plutôt que par N

1. Sur données brutes :

$$s^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2$$

2. Sur données traitées avec effectif :

$$s^2 = \frac{1}{N-1} \sum n_i (x_i - \bar{x})^2$$

3. Sur données traitées regroupées en classes :

$$s^2 = \frac{1}{N-1} \sum n_i (c_i - \bar{x})^2$$

L'écart-type

- ▶ L'écart type est la racine carré de la variance que l'on note naturellement σ ou s ($= \sqrt{\sigma^2}$ ou $\sqrt{s^2}$)
- ▶ Puisque qu'il existe une variance d'observation et d'échantillon, on calcule également un écart-type d'observation et un écart-type d'échantillon

Calcul de l'écart-type

1. Sur données brutes :

$$\sigma = \sqrt{\frac{1}{N} \sum (x_i - \bar{x})^2} \quad \text{ou} \quad s = \sqrt{\frac{1}{N-1} \sum (x_i - \bar{x})^2}$$

2. Sur données traitées avec effectif :

$$\sigma = \sqrt{\frac{1}{N} \sum n_i (x_i - \bar{x})^2} \quad \text{ou} \quad s = \sqrt{\frac{1}{N-1} \sum n_i (x_i - \bar{x})^2}$$

3. Sur données traitées regroupées en classes :

$$\sigma = \sqrt{\frac{1}{N} \sum n_i (c_i - \bar{x})^2} \quad \text{ou} \quad s = \sqrt{\frac{1}{N-1} \sum n_i (c_i - \bar{x})^2}$$

Calcul de l'écart-type : remarques

- ▶ Comme la variance, l'écart-type est strictement positif (attention aux calculs !)
- ▶ Puisqu'il se déduit de la variance (par fonction croissante), l'écart-type est aussi une mesure de dispersion qui s'interprète de la même façon
- ▶ L'écart-type est une mesure homogène aux données : si les observations sont des poids en kg, l'écart-type est aussi en kg
- ▶ Étant données plusieurs distributions, on peut comparer les dispersions avec la valeur des écart-types. . . mais il faut prendre quelques précautions

Le coefficient de variation

L'écart type est sensible à l'échelle des données. On utilise alors le coefficient de variation (noté cv) pour le mettre à l'échelle. Il est défini par le pourcentage suivant :

$$cv = \frac{S}{\bar{X}} \times 100$$

- ▶ Le cv est réservé aux distributions strictement positives
- ▶ Le cv est une mesure d'**homogénéité** des données, plus il est faible, plus les données sont homogènes
- ▶ En pratique si $cv < 15\%$, on dit que les données sont homogènes
- ▶ Le cv est très utile pour comparer des distributions avec des unités différentes et/ou avec des moyennes très éloignées

Statistiques inférentielles

Préambule

Imaginons deux personnes qui discutent à propos d'un téléphone...

L'un dit :

“J'ai dû le faire réparer deux fois en 1 an, je te déconseille d'en acheter car ils tombent rapidement en panne”

L'autre répond :

“Ah bon ? Pourtant je connais 10 personnes qui en ont un et ils n'ont encore jamais eu de problème depuis 1 an”

→ Que dire de la fiabilité du téléphone ?

Préambule

Imaginons deux personnes qui discutent à propos d'un téléphone. . .

L'un dit :

“J'ai dû le faire réparer deux fois en 1 an, je te déconseille d'en acheter car ils tombent rapidement en panne”

L'autre répond :

“Ah bon ? Pourtant je connais 10 personnes qui en ont un et ils n'ont encore jamais eu de problème depuis 1 an”

→ on a plutôt tendance à faire confiance à la deuxième personne non ?

Préambule

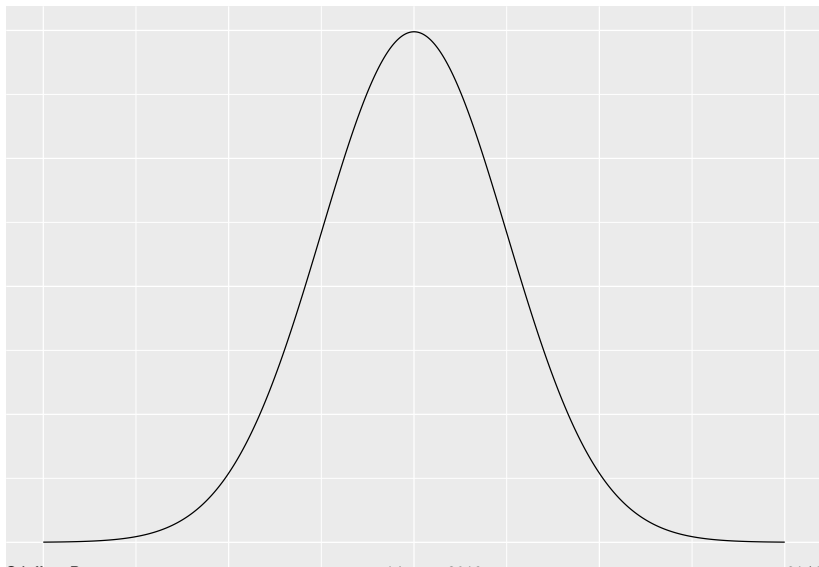
La statistique inférentielle nous permet d'inférer sur les données i.e., de généraliser des propriétés constatées sur l'échantillon à la population entière (e.g., moyenne, dispersion ou quantiles)

Dans l'exemple précédent, si on souhaite savoir combien de temps s'écoule avant d'avoir une panne

- ▶ Comment faire sachant qu'on ne pourra étudier qu'un nombre limité d'appareils à chaque fois ?
- ▶ Est-il préférable d'avoir 10 appareils ?
- ▶ Ou 1000 appareils ?

Quelques mots sur la loi normale

S'il y a bien une loi populaire en statistique, il s'agit de la loi normale. . .la célèbre courbe en cloche !

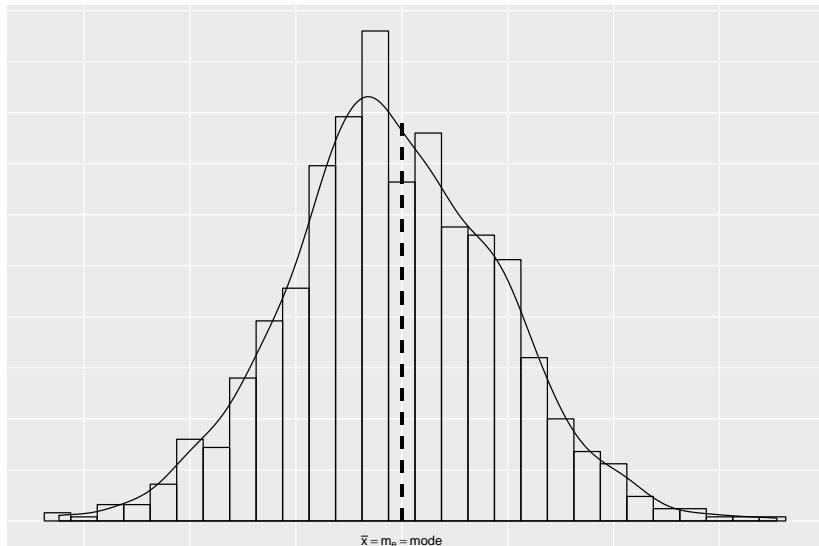


Caractéristiques de la loi normale

- ▶ La loi normale est une loi **symétrique, centrée autour de sa moyenne**
- ▶ La symétrie de la distribution implique que **la médiane est égale à la moyenne**
- ▶ C'est une loi unimodale, **son mode est égale à la moyenne**

Reconnaître une loi normale

En pratique, on peut supposer une distribution normale d'après l'histogramme, ou le diagramme en barre



Quelques mots sur la loi normale

- ▶ La loi normale est définie par deux paramètres :
- 1. Sa moyenne souvent notée μ (où *mean*)
- 2. Son écart-type notée σ (où *sd* pour “Standard Deviation”)

Sa densité de probabilité est la suivante :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

avec une variable X qui suit une loi normale de moyenne μ et d'écart-type σ

On note $X \sim N(\mu, \sigma^2)$

Illustration de la loi normale

Changer la moyenne d'une loi normale revient à translater la distribution vers la droite ou la gauche

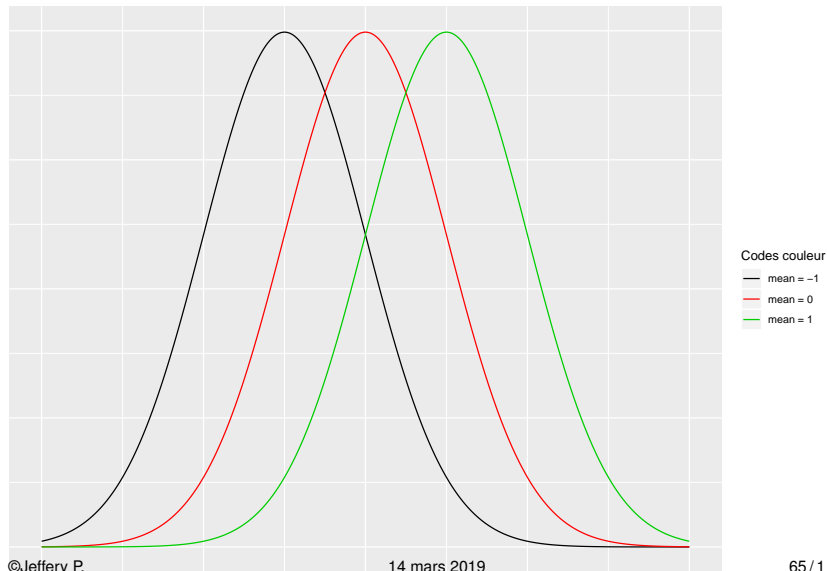
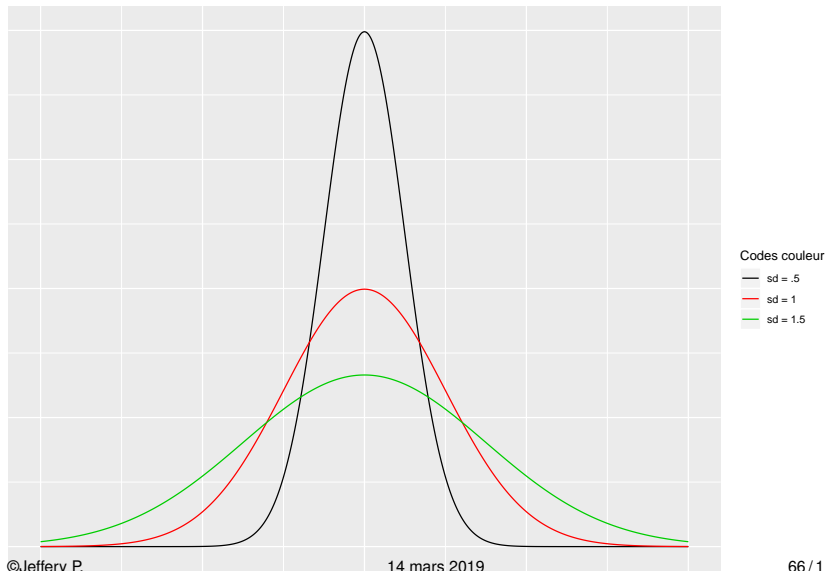


Illustration de la loi normale

Changer l'écart-type d'une loi normale revient à aplatir ou resserrer sa distribution autour de sa moyenne



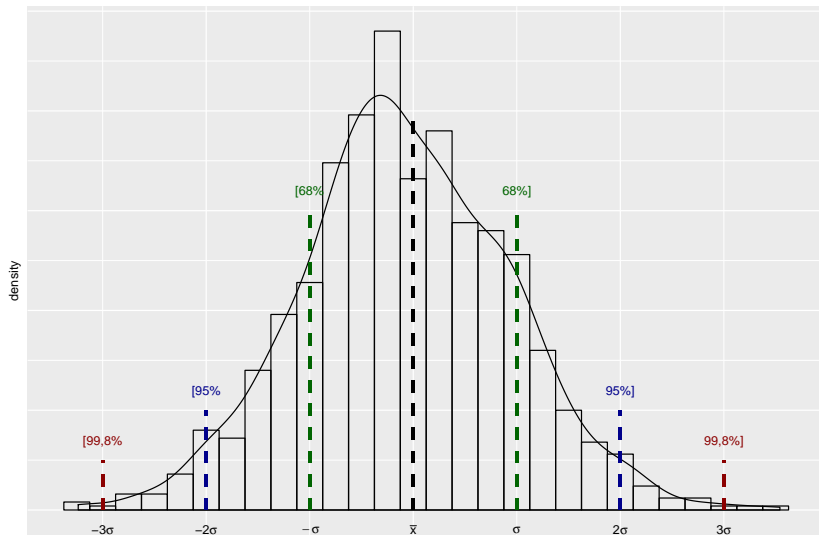
Propriété intéressante

Si l'on dispose d'observations d'une distribution $N(\mu, \sigma^2)$ alors :

- ▶ 68% des observations sont comprises dans l'intervalle $[\mu - \sigma; \mu + \sigma]$
- ▶ 95% des observations sont comprises dans l'intervalle $[\mu - 2\sigma; \mu + 2\sigma]$
- ▶ 99,8% des observations sont comprises dans l'intervalle $[\mu - 3\sigma; \mu + 3\sigma]$

Estimation graphique de μ et σ

- En pratique on estime μ en prenant le centre de la distribution, et on déduit σ en estimant l'intervalle vert



Calcul de probabilité : généralité

La probabilité pour qu'une variable aléatoire X soit inférieure à une quelconque valeur x s'écrit $P(X \leq x)$

- **Un probabilité est toujours positive et inférieure à 1**

Remarque

- Pour une variable aléatoire réelle on a :
 $P(-\infty < X < +\infty) = 1$
- Pour une variable continue, on a $P(X = x) = 0$, d'où :

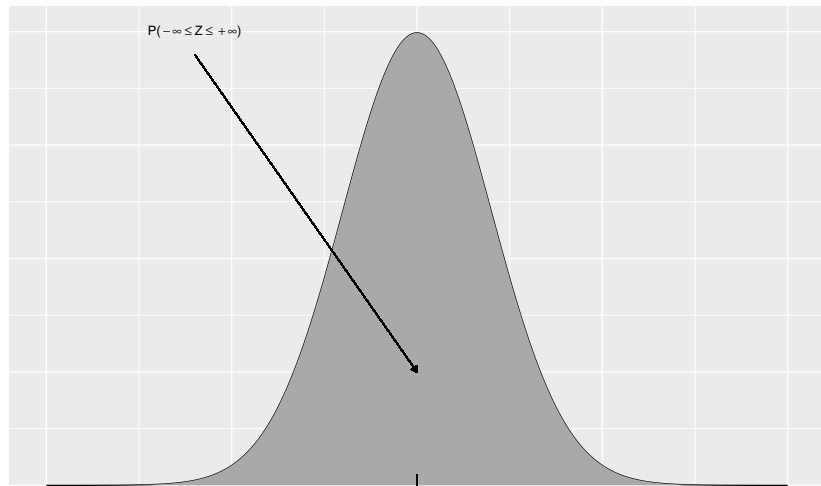
$$P(X \leq x) = P(X < x) \text{ où encore } P(X \geq x) = P(X > x)$$

Ce résultat s'explique avec un peu de théorie mathématique que l'on ne détaillera pas ici, mais il peut être utile d'avoir ces propriétés en tête pour les exercices...

Calcul de probabilité : lien avec l'aire sous la courbe de densité

Cas 0 : soit $Z \sim N(0, 1)$, l'aire sous la courbe est égale à 1

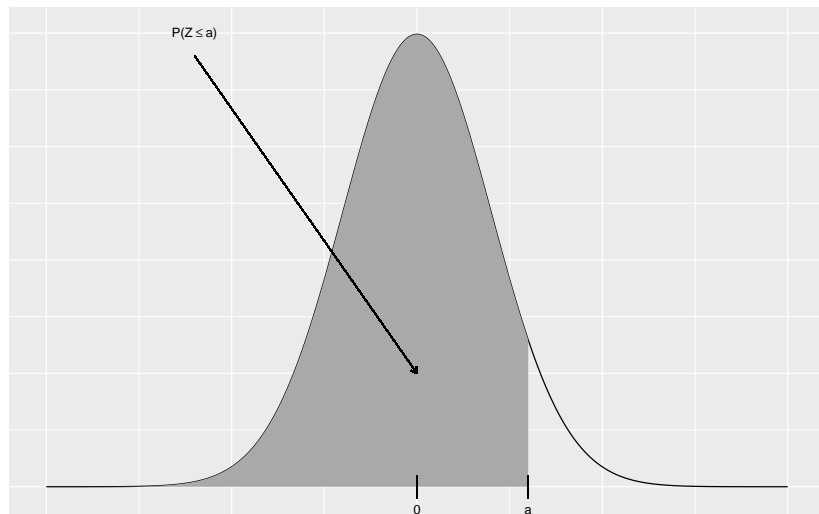
$$P(-\infty < Z < +\infty) = 1$$



Calcul de probabilité : valeurs de la table

Cas I : soient $Z \sim N(0, 1)$ et a un nombre réel **positif**

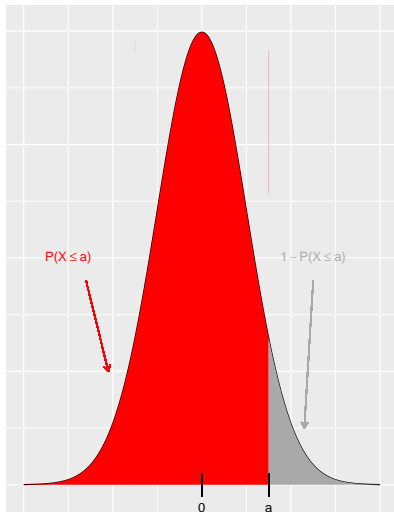
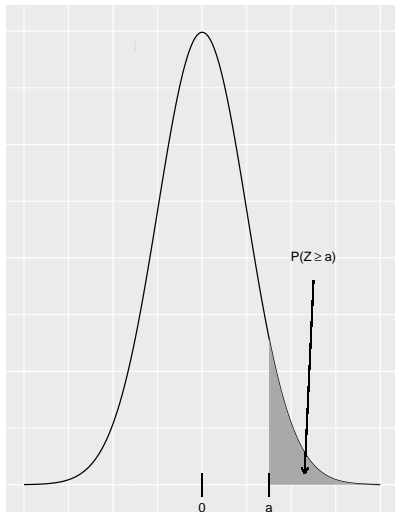
$P(Z \leq a) \rightarrow$ le résultat se trouve dans la table !



Calcul de probabilité avec la loi normale

Cas II : soient $Z \sim N(0, 1)$ et a un nombre réel **positif**

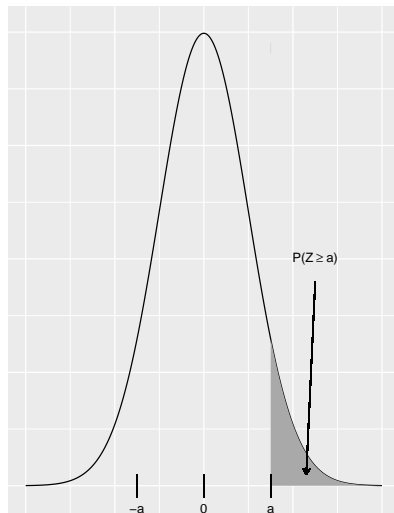
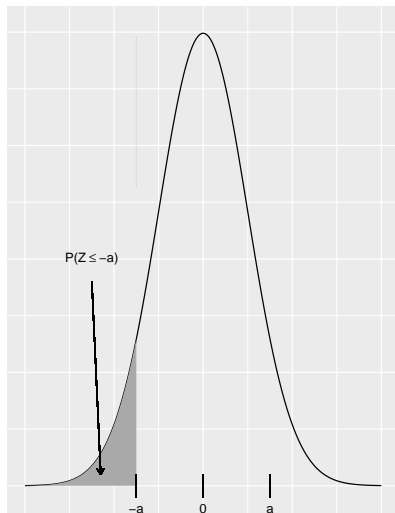
$P(Z \geq a) = 1 - P(Z \leq a) \rightarrow$ on se ramène au **cas I**



Calcul de probabilité avec la loi normale

Cas III : soient $Z \sim N(0, 1)$ et a un nombre réel **positif**

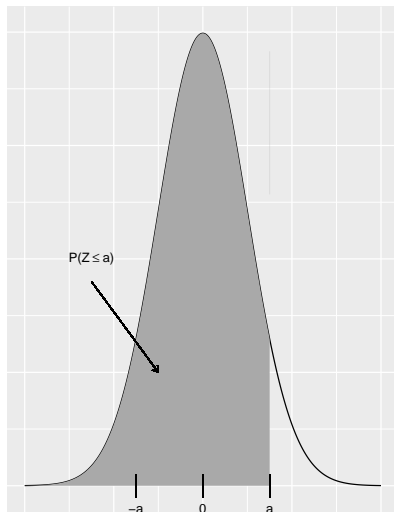
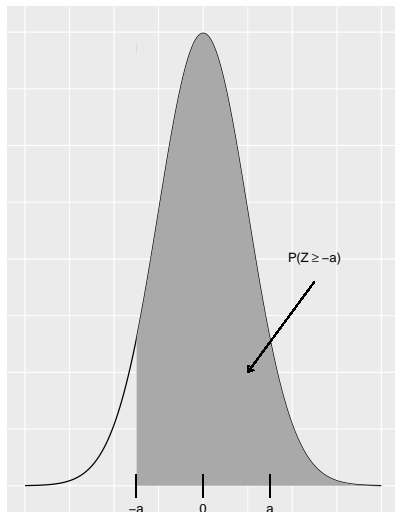
$P(Z \leq -a) = P(Z \geq a) \longrightarrow$ on se ramène au **cas II**



Calcul de probabilité avec la loi normale

Cas IV : soient $Z \sim N(0, 1)$ et a un nombre réel **positif**

$P(Z \geq -a) = P(Z \leq a) \longrightarrow$ on se ramène au **cas I**



Résumé








I	$\mathbb{P}(X \leq a)$		\Rightarrow table
II	$\mathbb{P}(X \geq a)$	 $= 1 -$ 	\Rightarrow cas I
III	$\mathbb{P}(X \leq -a)$	 $=$ 	\Rightarrow cas II
IV	$\mathbb{P}(X \geq -a)$	 $=$ 	\Rightarrow cas I

FIGURE 1 – Calcul de probabilité (*Extrait du cours de M. Gérin, Paris Ouest 2012-2013*)

Formule de calcul avec une intervalle quelconque

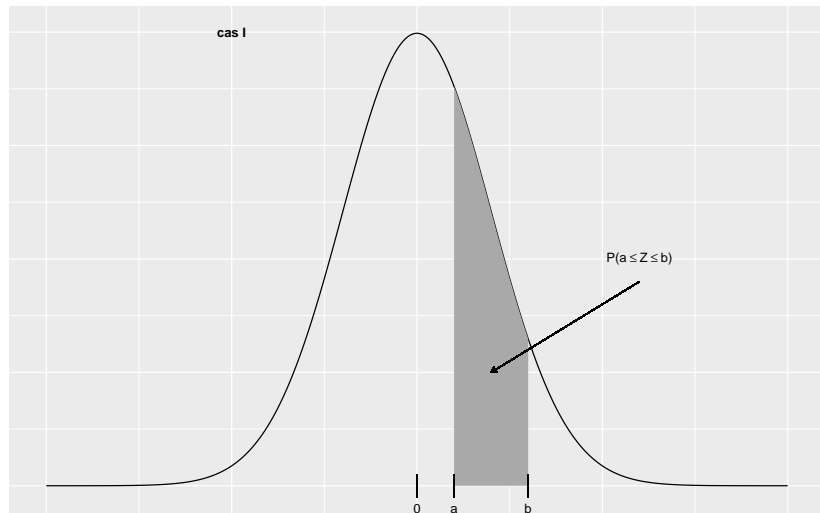
Soient $Z \sim N(0, 1)$ et u, v deux nombres réels avec $u \leq v$

$$P(u \leq Z \leq v) = P(Z \leq v) - P(Z \leq u)$$

Calcul de probabilité avec la loi normale

Cas V.1 : soient $Z \sim N(0, 1)$ et $0 \leq a \leq b$

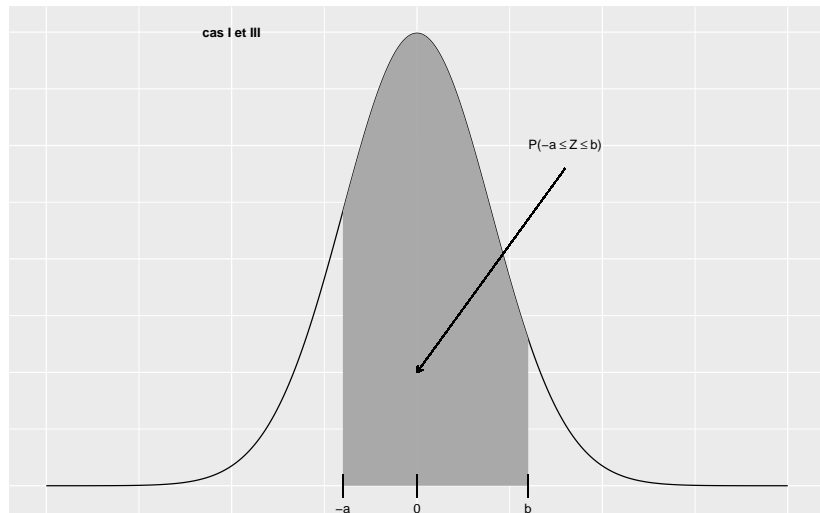
$$P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$$



Calcul de probabilité avec la loi normale

Cas V.2 : soient $Z \sim N(0, 1)$ et $0 \leq a \leq b$

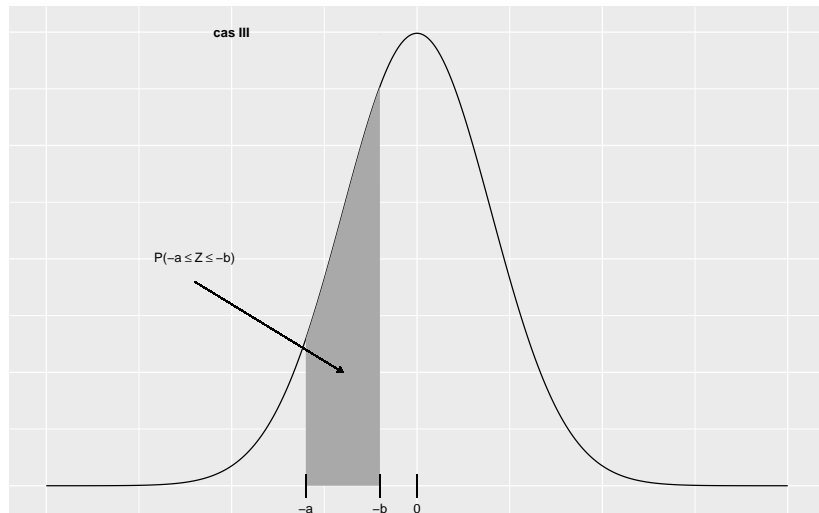
$$P(-a \leq Z \leq b) = P(Z \leq b) - P(Z \leq -a) = P(Z \leq b) + P(Z \leq a) - 1$$



Calcul de probabilité avec la loi normale

Cas V.3 : soient $Z \sim N(0, 1)$ et $-a \leq -b \leq 0$

$$P(-a \leq Z \leq -b) = P(Z \leq -b) - P(Z \leq -a) = P(Z \leq a) - P(Z \leq b)$$



Transformation de la loi normale

Toute transformation *affine* d'une loi normale est encore une loi normale i.e., quelque soit les nombre réels a et $b \neq 0$ on a :

$$X \sim N(\mu, \sigma^2) \Rightarrow aX + b \sim N(\mu + b, a^2 \times \sigma^2)$$

Par conséquent :

- ▶ si $X \sim N(\mu, \sigma^2)$ alors $X - \mu \sim N(0, \sigma^2)$ (**centrage**)
- ▶ si $X \sim N(\mu, \sigma^2)$ alors $\frac{X}{\sigma} \sim N(\mu, 1)$ (**réduction**)

En pratique, on fait souvent les deux en même temps :

- ▶ si $X \sim N(\mu, \sigma^2)$ alors $\frac{X-\mu}{\sigma} \sim N(0, 1)$
(**normalisation=centrage+réduction**)

Calcul de probabilité : en pratique

Si la variable X ne suit pas une loi normale centrée réduite, on peut toujours se ramener aux cas précédents, par exemple :

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq \frac{x - \mu}{\sigma}\right)$$

Où $Z \sim N(0, 1)$

→ Rappel : il s'agit de la **normalisation**, cette petite transformation est très utile et beaucoup utilisée !

Cas pratique : calcul de probabilité

68% des observations sont comprises dans l'intervalle $[\bar{x} - \sigma; \bar{x} + \sigma]$. . . Est-ce bien vrai ?

$$\begin{aligned}P(\bar{x} - \sigma \leq X \leq \bar{x} + \sigma) &= P(-\sigma \leq X - \bar{x} \leq \sigma), \quad \text{on centre} \\&= P(-1 \leq \frac{X - \bar{x}}{\sigma} \leq 1), \quad \text{on réduit} \\&= P(-1 \leq Z \leq 1)\end{aligned}$$

Où $Z \sim N(0, 1)$

Il faut maintenant trouver la valeur de cette probabilité sur les tables. . .

Cas pratique

$$\begin{aligned}P(\bar{x} - \sigma \leq X \leq \bar{x} + \sigma) &= P(-1 \leq Z \leq 1) \\&= P(Z \leq 1) - P(Z \leq -1) \\&= P(Z \leq 1) - (1 - P(Z \leq 1)) \\&= 2P(Z \leq 1) - 1\end{aligned}$$

Où $Z \sim N(0, 1)$

Calcul de probabilité inversé

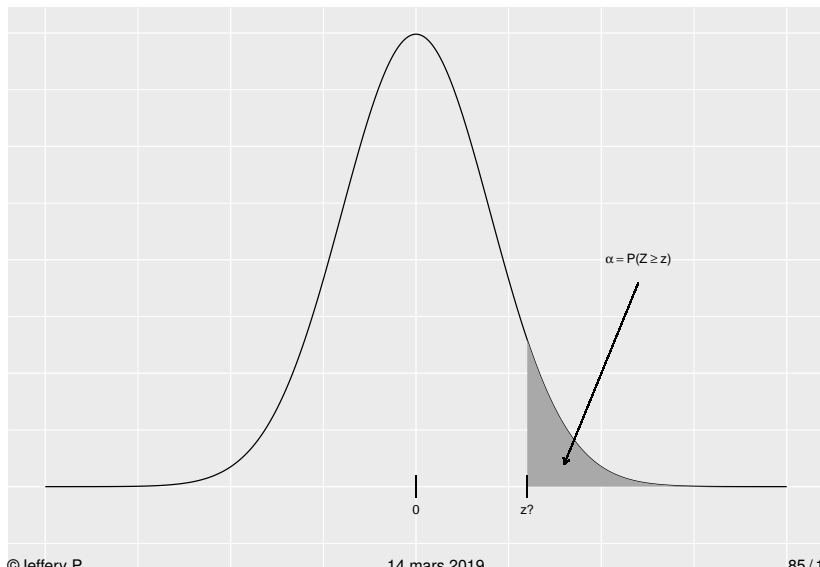
Dans certains cas, on ne cherche pas à calculer une probabilité pour un b donné (e.g., $P(Z \leq b)$) mais on cherche b telle que la probabilité soit égale à une certaine valeur :

On veut trouver b tel que $P(X \geq b) = \alpha$ où $P(|X| \geq b) = \alpha$

- ▶ Dans ce cas, on s'assure que la probabilité concerne une variable suivant une **loi normale centrée réduite**, puis on regarde dans une des tables

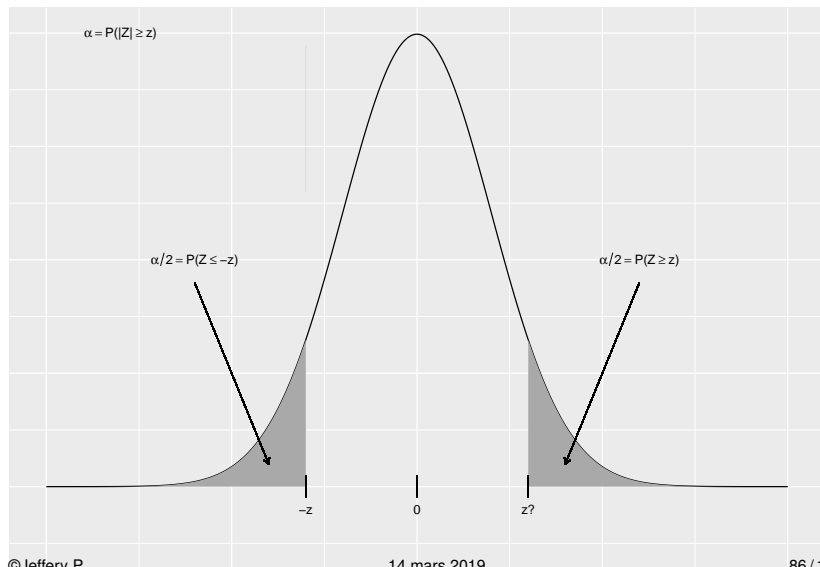
Calcul de probabilité inversé : table 1

On recherche une valeur z telle que $P(Z > z)$ avec $Z \sim N(0, 1)$:



Calcul de probabilité inversé : table 2

On recherche une valeur z telle que $P(|Z| > z)$ avec $Z \sim N(0, 1)$:



Calcul de probabilité inversé

Attention, cette problématique est faussement facile... dans la majorité des cas elle nécessite une bonne maîtrise du calcul des probabilités de la loi normale !

Cas fréquents

- ▶ On ne demande pas de trouver z tel que $P(Z \geq z) = \alpha$ mais plutôt $P(Z \leq z) = 1 - \alpha$.
- ▶ On ne demande pas de trouver z tel que $P(|Z| \geq z) = \alpha$ mais plutôt $P(|Z| \leq z) = 1 - \alpha$.
- ▶ La probabilité porte sur une variable $X \sim N(\mu, \sigma^2)$. Il faut bien penser à centrer et réduire l'évènement (ce qui se trouve dans la probabilité) e.g., $P(X \leq x) = P(Z \leq \frac{x-\mu}{\sigma})$ où $Z \sim N(0, 1)$

Intervalle de confiance : introduction

Jusqu'ici on a calculé des probabilités sur une loi normale connue.

- Imaginons maintenant que l'on dispose d'un échantillon d'observations que l'on suppose issues d'une loi normale $N(\mu, \sigma^2)$

Comment faire pour estimer la vraie moyenne μ avec les données dont on dispose ?

Intervalle de confiance : introduction

Une première intuition peut consister à supposer que la vraie moyenne μ est « à peu près » égale à la moyenne de l'échantillon dite empirique i.e., \bar{x}