MAT 2777
Probabilités et statistique pour ingénieur.e.s

Chapitre 4
Statistiques descriptives et distributions d'échantillonnage

P. Boily (uOttawa)

Hiver 2023

Aperçu

4.1 – Les descriptions de données

En un sens, la raison sous-jacente de l'analyse statistique est de parvenir à une **compréhension des données**.

Les études et les expériences donnent naissance à des **unités statistiques**.

Ces unités sont généralement décrites avec des **variables** (et des mesures).

Les variables sont soit **qualitatives** (catégoriques) soit **quantitatives** (numériques).

Les variables catégorielles prennent des valeurs (**niveaux**) dans un ensemble fini de **catégories** (ou classes).

Les variables numériques prennent des valeurs dans un ensemble (potentiellement infini) de **quantités numériques**.

Exemples:

1. L'âge est une variable numérique, mesurée en années (il est souvent rapporté à l'année entière la plus proche, ou dans une étendue d'années, auquel cas il s'agit d'une variable **ordinale**.

2. Les variables numériques comprennent la distance (m), le volume (cm^3), etc.

3. Le diagnostique de la maladie est une variable catégorielle avec (au moins) 2 catégories (positif/négatif).

4. La conformité à une norme est une variable catégorielle : il peut y avoir 2 niveaux (conforme/non conforme) ou plus (conformité, problèmes mineurs de non-conformité, problèmes majeurs de non-conformité).

5. Les variables de comptage sont des variables numériques.

Les résumés numériques

Dans un premier temps, une variable numérique peut être décrite selon 2 dimensions : la **centralité** et la **dispersion** (l'**asymétrie** et l'**aplatissement** sont aussi parfois utilisés) :

- mesures de **centralité** : la **médiane**, la **moyenne**, (le mode, moins fréquemment) ;
- mesures de **dispersion** (ou d'**étendue**) : l'**écart-type**, les **quartiles**, l'**étendue inter-quartile** (EIQ), (l'étendue, moins fréquemment).

La médiane, l'étendue, et les quartiles se calculent facilement à partir d'une liste **ordonnée** des données.

La médiane d'un échantillon

La **médiane** $\text{med}(x_1, \ldots, x_n)$ d'un échantillon de taille n est une valeur numérique qui divise les données ordonnées en 2 sous-ensembles égaux : la moitié des observations se situent en dessous de la médiane, **et** l'autre moitié au-dessus.

- Si n est **impair**, l'observation médiane est la $\frac{n+1}{2}^{\text{e}}$ observation ordonnée.
- Si n est **pair**, l'observation médiane est la moyenne des $\frac{n}{2}^{\text{e}}$ et $(\frac{n}{2}+1)^{\text{e}}$ observations ordonnées.

La procédure est simple : ordonnez les données, et suivez les règles paires/impaires **à la lettre**.

Exemples:

1. $\text{med}(4,6,1,3,7) = \text{med}(1,3,4,6,7) = x_{(5+1)/2} = x_3 = 4$. Il y a 2 observations sous 4 $(1,3)$, et 2 observations au-dessus de 4 $(6,7)$.

2. $\text{med}(1,3,4,6,7,23) = \frac{x_{6/2}+x_{6/2+1}}{2} = \frac{x_3+x_4}{2} = \frac{4+6}{2} = 5$. Il y a 3 observations sous 5 $(1,3,4)$, et 3 observations au-dessus de 4 $(6,7,23)$.

3. $\text{med}(1,3,3,6,7) = x_{(5+1)/2} = x_3 = 3$. Il semble n'y avoir que 1 observations sous 3 (1), mais 2 observations au-dessus de 3 $(6,7)$.

   Ce n'est pas tout à fait l'interprétation correcte de la médiane : **sous** et **au-dessus** dans la définition devraient être interprétés comme **après** et **avant**, respectivement. Dans cet exemple, il y a 2 d'observations $(x_1 = 1, x_2 = 3)$ avant la médiane $(x_3 = 3)$, et 2 après $(x_4 = 6, x_5 = 7)$.

La moyenne d'un échantillon

La **moyenne** d'un échantillon est simplement la moyenne arithmétique de ses observations. Pour les observations $x_1, x_2, \ldots, x_n$, la moyenne de l'échantillon est

$$\text{MA}(x_1, \ldots, x_n) = \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right).$$

D'autres moyennes existent, telles que la moyenne **harmonique** et la moyenne **géométrique** :

$$\text{MH}(x_1, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} \quad \text{et} \quad \text{MG}(x_1, \ldots, x_n) = \sqrt[n]{x_1 \cdots x_n}.$$

Exemples:

1. $\text{MA}(4, 6, 1, 3, 7) = \frac{4+6+1+3+7}{5} = \frac{21}{5} = 4.2 \approx 4 = \text{med}(4, 6, 1, 3, 7)$.

2. $\text{MA}(1, 3, 4, 6, 7, 23) = \frac{1+3+4+6+7+23}{6} = \frac{44}{6} \approx 7.3$, ce qui n'est pas aussi près de $\text{med}(1, 3, 4, 6, 7, 23) = 5$.

3. $\text{MH}(4, 6, 1, 3, 7) = \frac{5}{\frac{1}{4}+\frac{1}{6}+\frac{1}{1}+\frac{1}{3}+\frac{1}{7}} = \frac{5}{53/28} = \frac{140}{53} \approx 2.64$.

4. $\text{MG}(4, 6, 1, 3, 7) = \sqrt[5]{4 \cdot 6 \cdot 1 \cdot 3 \cdot 7} \approx \sqrt[5]{(504)} \approx 3.47$.

Si $x = (x_1, \ldots, x_n)$ et $x_i > 0$ pour tout i,

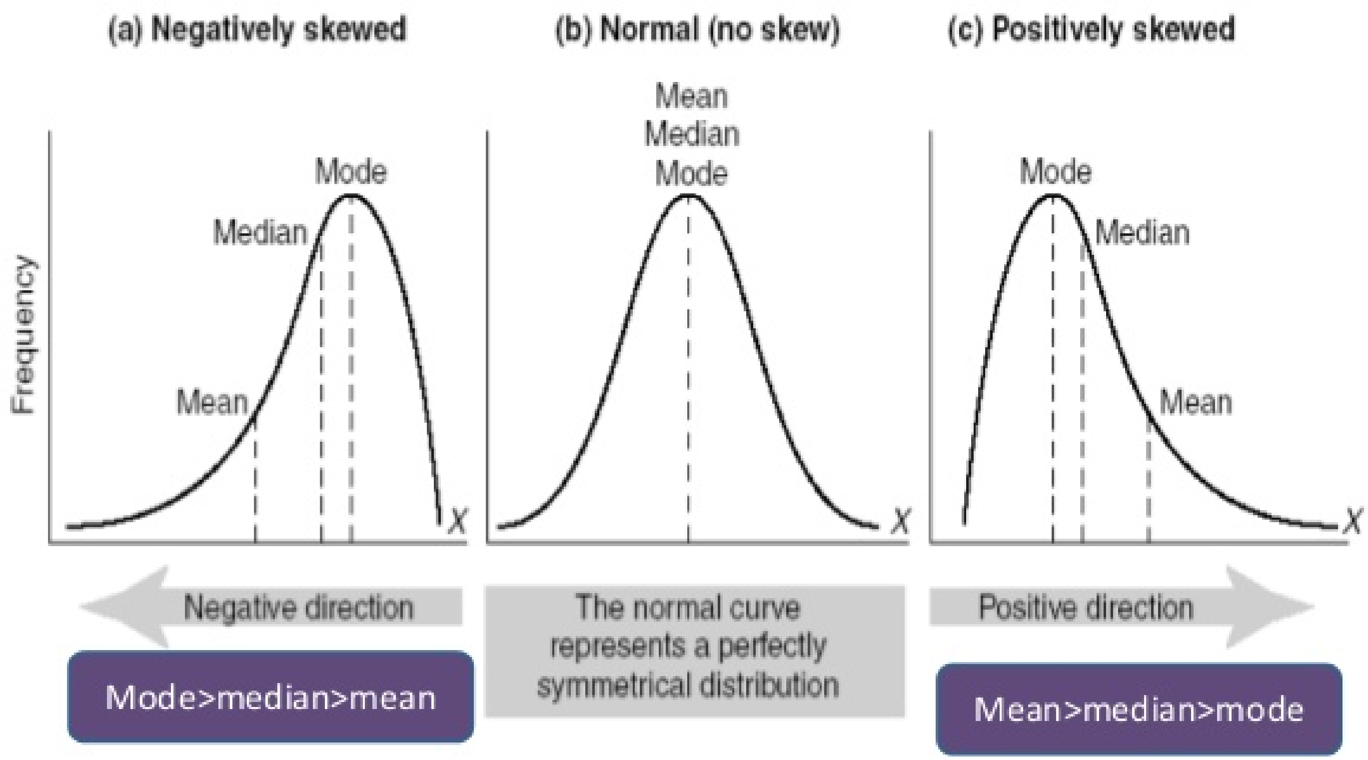
$$\min(x) \leq \text{MH}(x) \leq \text{MG}(x) \leq \text{MA}(x) \leq \max(x).$$

La moyenne ou la médiane?

Quelle mesure de centralité utilise-t-on pour rendre compte des données ?

1. La moyenne est **théoriquement supportée** (par le TLC).
2. Si la distribution des données est à peu près symétrique, les deux valeurs seront près l'une de l'autre.
3. Si la distribution des données est **asymétrique**, la moyenne est tirée vers la longue queue et donne par conséquent une vue déformée du centre. La médiane est utilisée pour les prix des maisons, les revenus, etc.
4. La médiane est **robuste** à l'encontre des valeurs aberrantes et les lectures incorrectes alors que la moyenne ne l'est pas.

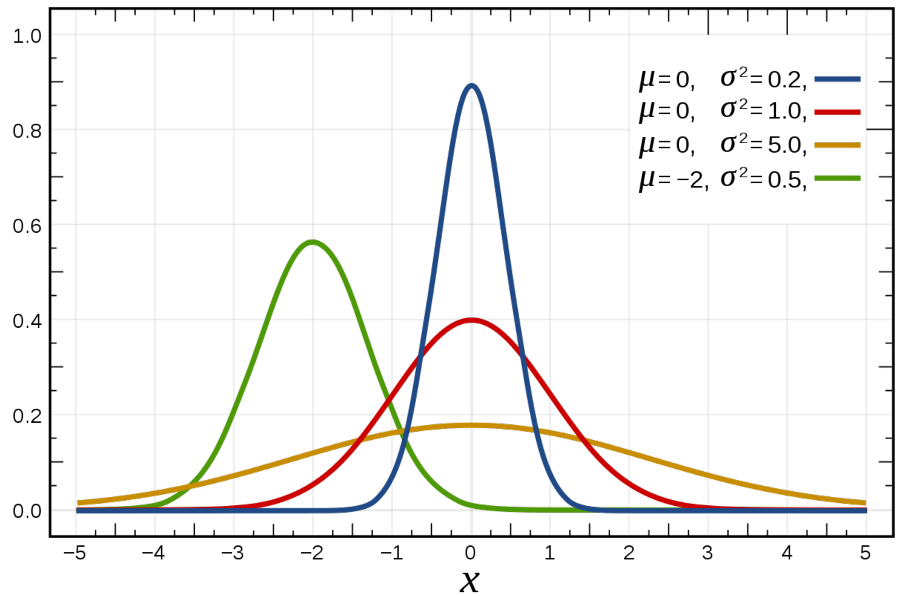
La moyenne ou la médiane?



L'écart-type (reprise)

La moyenne, la médiane, et le mode donnent une idée de l'endroit où se trouve la “masse” de la distribution.

L'écart-type donne une idée de sa dispersion.



Les quartiles d'un échantillon

On utilise également les **centiles**, **déciles**, ou **quartiles** afin de fournir des informations sur la dispersion des données.

Le **quartile inférieur** $Q_1(x_1, \ldots, x_n)$ d'un échantillon de taille n divise les données ordonnées en 2 sous-ensembles **inégaux** : 25% des observations sont inférieures à Q_1, **et** 75% des observations sont supérieures à Q_1.

Le **quartile supérieur** $Q_3(x_1, \ldots, x_n)$ divise les données ordonnées: 75% des observations inférieures à Q_3, **et** 25% des observations supérieures à Q_3.

La **médiane** peut être interprétée comme le **quartile moyen** Q_2 de l'échantillon, le **minimum** comme Q_0, et le **maximum** comme Q_4.

Les **centiles** p_i, $i = 0, \ldots, 100$ et **déciles** d_j, $j = 0, \ldots, 10$ utilisent différents pourcentages $\Longrightarrow p_{25} = Q_1, p_{75} = Q_3, d_5 = Q_2$, etc.

Triez les observations de l'échantillon $\{x_1, x_2, \ldots, x_n\}$ en **ordre croissant** :

$$y_1 \leq y_2 \leq \ldots \leq y_n.$$

Le plus petit y_1 a un **rang** de 1 et le plus grand y_n a un **rank** de n.

On calcule le quartile inférieur Q_1 selon la moyenne des observations de rang $\lfloor \frac{n}{4} \rfloor$ et $\lfloor \frac{n}{4} \rfloor + 1$.

De même, le quartile supérieur Q_3 est la moyenne des observations de rang $\lceil \frac{3n}{4} \rceil$ et $\lceil \frac{3n}{4} \rceil + 1$.

Exemples:

$$Q_1(1, 3, 4, 6, 7, 10, 12, 23) = 3.5, \quad Q_3(1, 3, 4, 6, 7, 10, 12, 23) = 11.$$

Exemple: voici le nombre quotidien d'accidents à Sydney

```
> accident
 6, 3, 2, 24, 12, 3, 7, 14, 21, 9, 14, 22, 15, 2,
 17, 10, 7, 7, 31, 7, 18, 6, 8, 2, 3, 2, 17, 7, 7,
 21, 13, 23, 1, 11, 3, 9, 4, 9, 9, 25
> sort(accident)
 1  2  2  2  2  3  3  3  3  4  6  6  7  7  7  7  7  7  8  9
 9  9  9 10 11 12 13 14 14 15 17 17 18 21 21 22 23 24 25 31
> summary(accident)
  Min.  1st quartile  Median  Mean  3rd quartile   Max.
  1.00      5.50        9.00  10.78     15.50       31.00
> var(accident)
  58.7
```

Si on remplace le 31 par 130, la moyenne devient 13.28 et la variance devient 412.4, mais la médiane **demeure la même**.

Les mesures de dispersion

L'**étendue** de l'échantillon $x_1, \ldots, x_n$ est $\max\{x_i\} - \min\{x_i\} = y_n - y_1$, où $y_1 \leq \ldots \leq y_n$ sont les données ordonnées.

L'**étendue inter-quartile** est EIQ $= Q_3 - Q_1$.

L'**écart-type** s et la **variance** s^2 de l'échantillon sont des estimations des paramètres de la distribution sous-jacente σ et σ^2.

Pour l'échantillon $x_1, x_2, \ldots, x_n$, nous avons

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right).$$

Les valeurs aberrantes

Une **valeur aberrante** est une observation qui ne se situe pas dans la tendance générale d'une distribution. Soit x une observation dans l'échantillon. It is a **valeur aberrante présumée** si

$$x < Q_1 - 1.5 \times \text{EIQ} \quad \text{ou} \quad x > Q_3 + 1.5 \times \text{EIQ},$$

où $\text{EIQ} = Q_3 - Q_1$.

Cette définition ne s'applique avec certitude qu'aux données qui **suivent une loi normale**, bien qu'elle soit souvent utilisée comme première passe lors de l'analyse des valeurs aberrantes.

Exercice: Considérons un échantillon de $n = 10$ observations affichées par ordre croissant.

$$15, 16, 18, 18, 20, 20, 21, 22, 23, 75.$$

1. Vérifiez que l'écart-type de cet échantillon est de $s = 17.81884$.

2. Vérifiez que $Q_1 = 17$ et $Q_3 = 22.5$.

3. Y a-t-il des valeurs aberrantes probables dans l'échantillon ? Si oui, indiquez leurs valeurs.

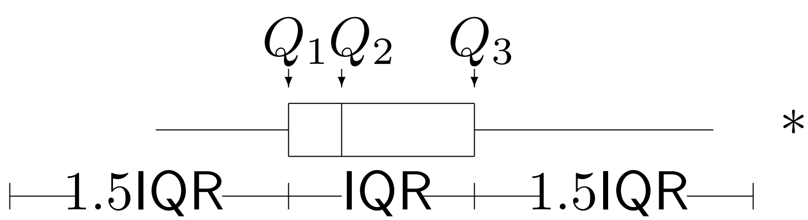
4.2 – Les résumés visuels

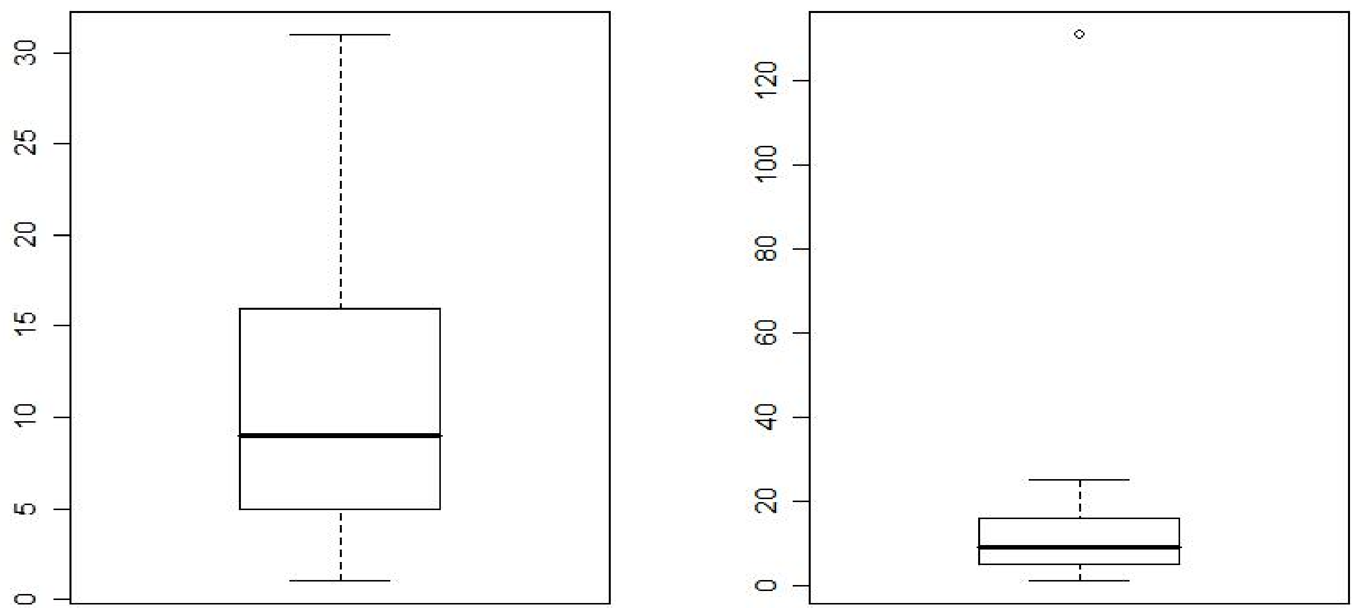
Le **diagramme à moustache (boxplot)** est un moyen rapide et facile de présenter un résumé graphique d'une distribution **univariée**.

Dessinez une boîte le long de l'axe d'observation, avec des extrémités aux **quartiles inférieur et supérieur**, et avec une "ceinture" à la **médiane**.

Ensuite, tracez une ligne s'étendant de Q_1 à la **plus petite valeur supérieure** à $Q_1 - 1.5 \times \text{EIQ}$, et de Q_3 à la **plus grande valeur inférieure** à $Q_3 + 1.5 \times \text{EIQ}$.

Toute valeur aberrante présumée est tracée séparément.




0
5
10
15
20
25
30
0
20
40
60
80
100
120

L'asymétrie

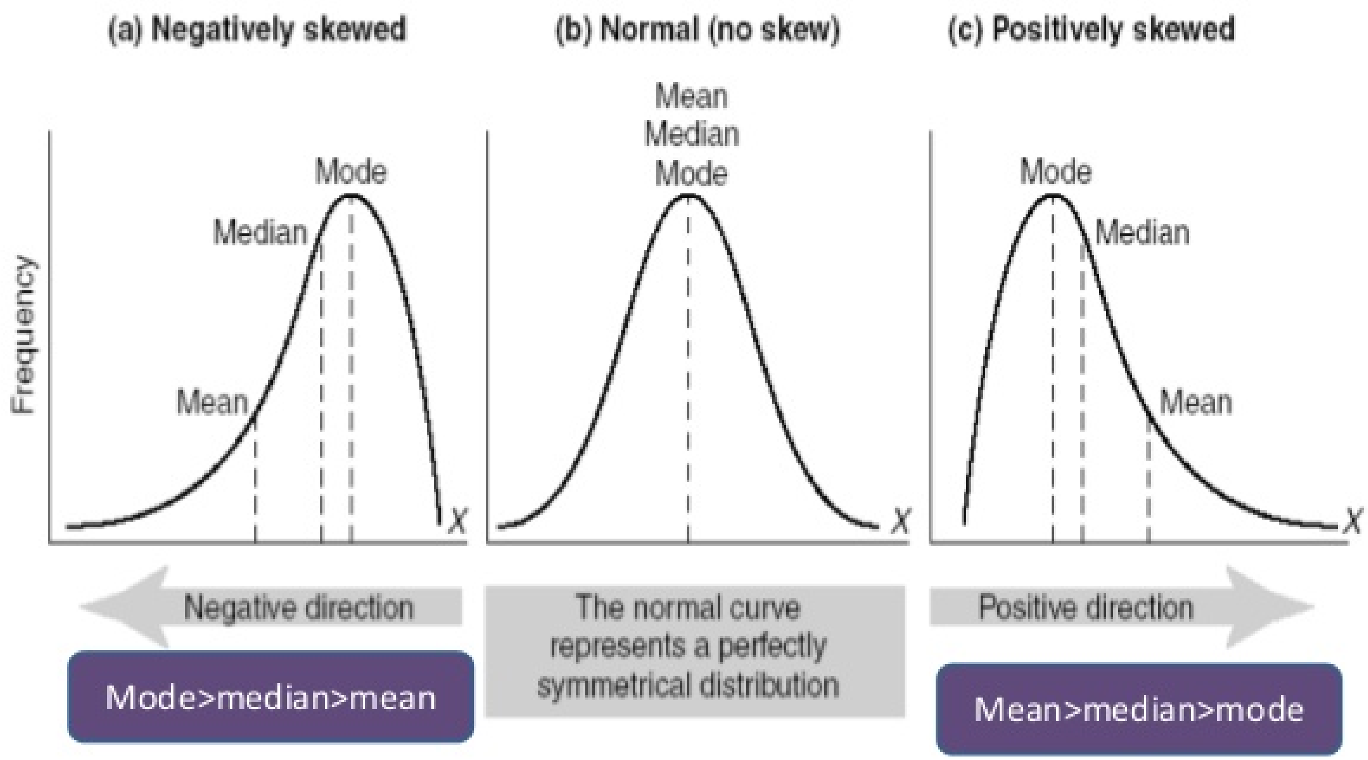
Si la distribution des données est **symétrique**, la médiane et la moyenne (de la population) sont **égales** et les premier et troisième quartiles (de la population) sont **équidistants de la médiane**.

Si $Q_3 - Q_2 \gg Q_2 - Q_1$ alors la distribution des données est **asymétrique vers la droite**.

Si $Q_3 - Q_2 \ll Q_2 - Q_1$ alors la distribution des données est **asymétrique vers la gauche**.

Dans les deux exemples de la diapositive précédente, les distributions sont asymétriques **vers la droite**.

L'asymétrie


(a) Negatively skewed
(b) Normal (no skew)
(c) Positively skewed
Frequency
Mode
Median
Mean
X
Mean
Median
Mode
X
Mode
Median
Mean
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction
Mode>median>mean
Mean>median>mode

Les histogrammes

Les **histogrammes** fournissent également une indication de la distribution d'un échantillon de taille n. Ils devraient contenir les informations suivantes :

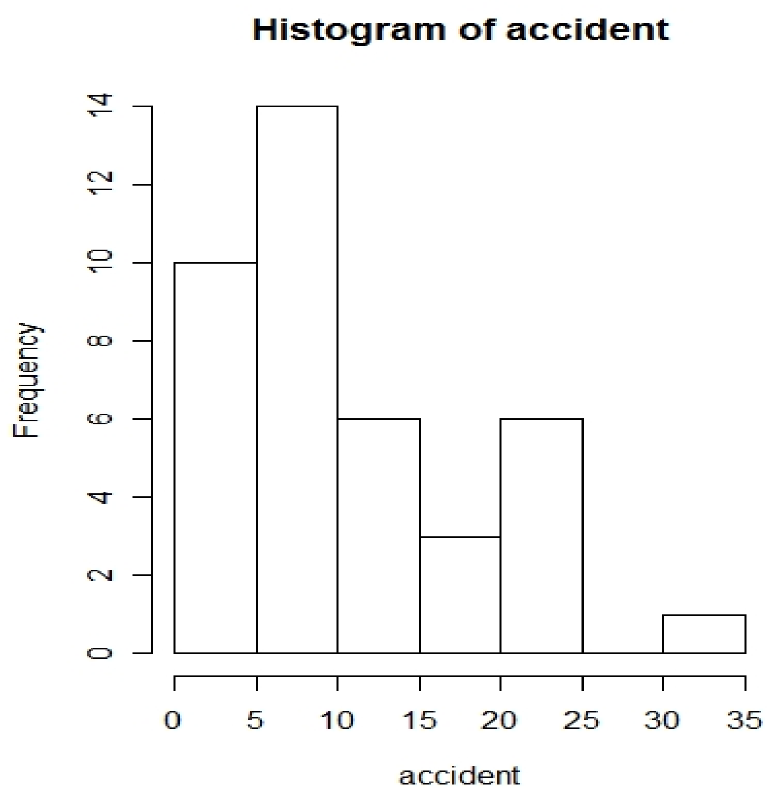
- l'**étendue** de l'histogramme est de $r = \max\{x_i\} - \min\{x_i\}$;
- le **nombre** de "bacs" devrait approcher $k = \sqrt{n}$;
- la **largeur** des "bacs" devrait approcher r/k,
- et la **fréquence des observations** dans chaque "bac" est souvent ajoutée au graphique.

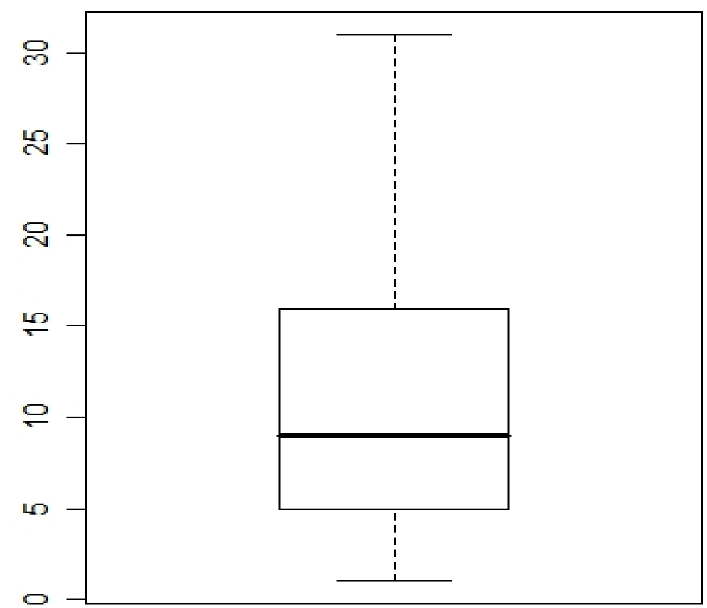
La "forme" d'un ensemble de données

Les "boxplots" constituent un moyen graphique simple pour obtenir une impression de la forme de l'ensemble de données. On utilise cette forme afin de suggérer un modèle mathématique pour la situation d'intérêt.

L'ensemble de données est **asymétrique à droite** si la boîte à moustache est étirée **vers la droite** ; il est **asymétrique à gauche** si elle est étirée **vers la gauche**.

Comme pour les boîtes à moustache, l'ensemble de données est **asymétrique à droite** si l'histogramme est étirée **vers la droite** ; il est **asymétrique à gauche** s'il est étirée **vers la gauche**.


Histogram of accident
Frequency
0
2
4
6
8
10
12
14
0 5 10 15 20 25 30 35
accident

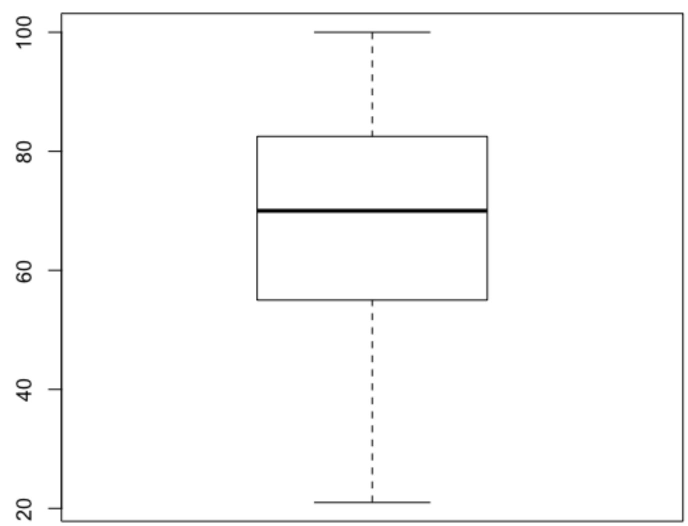

0
5
10
15
20
25
30

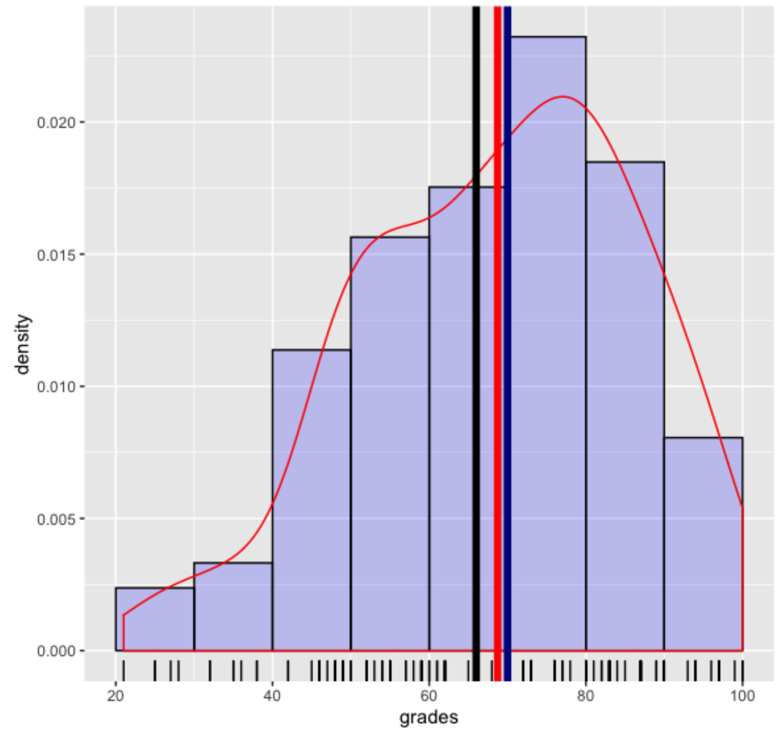
Exemple: les notes d'un examen sont indiquées ci-dessous. Discutez des résultats.

```
> grades<-c(80,73,83,60,49,96,87,87,60,53,66,83,32,80,66,90,72,
55,76,46,48,69,45,48,77,52,59,97,76,89,73,73,48,59,55,76,87,55,
80,90,83,66,80,97,80,55,94,73,49,32,76,57,42,94,80,90,90,62,85,
87,97,50,73,77,66,35,66,76,90,73,80,70,73,94,59,52,81,90,55,73,
76,90,46,66,76,69,76,80,42,66,83,80,46,55,80,76,94,69,57,55,66,
46,87,83,49,82,93,47,59,68,65,66,69,76,38,99,61,46,73,90,66,100,
83,48,97,69,62,80,66,55,28,83,59,48,61,87,72,46,94,48,59,69,97,
83,80,66,76,25,55,69,76,38,21,87,52,90,62,73,73,89,25,94,27,66,
66,76,90,83,52,52,83,66,48,62,80,35,59,72,97,69,62,90,48,83,55,
58,66,100,82,78,62,73,55,84,83,66,49,76,73,54,55,87,50,73,54,52,
62,36,87,80,80)

> hist(grades)
```

```
> # fonction qui calcule le mode
> fun.mode<-function(x){as.numeric(names(sort(-table(x)))[1])}

> library(ggplot2)
> ggplot(data=data.frame(grades), aes(grades)) +
    geom_histogram(aes(y =..density..), breaks=seq(20, 100, by = 10),
                   col="black", fill="blue", alpha=.2) +
    geom_density(col=2) + geom_rug(aes(grades)) +
    geom_vline(aes(xintercept = mean(grades)),col='red',size=2) +
    geom_vline(aes(xintercept = median(grades)),col='darkblue',size=2) +
    geom_vline(aes(xintercept = fun.mode(grades)),col='black',size=2)

> boxplot(grades)

> summary(grades)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
21.00   55.00   70.00   68.74   82.50  100.00
```

```
> library(psych)
> describe(grades)
n    mean     sd median trimmed   mad min max range  skew kurtosis  se
211 68.74 17.37     70    69.43 19.27  21 100    79 -0.37    -0.46 1.2
```





4.3 – Les distributions d'échantillonnage

Une **population** est un ensemble d'éléments similaires qui présente un intérêt par rapport à certaines questions ou expériences. Dans certaines situations, il est impossible d'observer l'ensemble des éléments qui composent une population. Dans ces cas, nous devons considérer un **échantillon** (un sous-ensemble) de la population afin de faire des inférences sur la population.

Supposons que $X_1, \ldots, X_n$ sont n v.a. **indépendantes**, chacune ayant la même f.r.c. F, c-à-d qu'elles sont **identiquement distribuées**. Alors, $\{X_1, \ldots, X_n\}$ est un **échantillon aléatoire** de taille n provenant de la population avec f.r.c. F.

Toute fonction d'un tel échantillon est une **statistique** de l'échantillon.

La distribution d'une statistique est une **distribution d'échantillonnage**.

Les propriétés de l'espérance et de la variance

Rappel: si X est une v.a., et $a, b \in \mathbb{R}$, alors

$$\begin{aligned} \mathrm{E}\left[a+bX\right] &= a+b\mathrm{E}[X], \\ \mathrm{Var}\left[a+bX\right] &= b^2\mathrm{Var}[X], \\ \mathrm{ET}\left[a+bX\right] &= |b|\mathrm{ET}[X]. \end{aligned}$$

De plus,

$$\begin{aligned} \mathrm{Var}\left[aX+bY\right] = \mathrm{E}\left[(aX+bY)^2\right] - \left(\mathrm{E}\left[aX+bY\right]\right)^2 &= a^2\left(\mathrm{E}[X^2]-\mathrm{E}^2[X]\right) \\ +2ab\left(\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]\right) &+ b^2\left(\mathrm{E}[Y^2]-\mathrm{E}^2[Y]\right) \\ = a^2\mathrm{Var}[X]+b^2\mathrm{Var}[Y]+2ab\mathrm{Cov}(X,Y) \end{aligned}$$

Les sommes de v.a. indépendantes

Pour toute v.a. X et Y, nous avons

$$\mathrm{E}[X+Y]=\mathrm{E}[X]+\mathrm{E}[Y]$$

Si **de plus** X et Y sont des v.a. **indépendantes** ($\mathrm{Cov}(X,Y)=0$), alors

$$\mathrm{Var}[X+Y]=\mathrm{Var}[X]+\mathrm{Var}[Y]\,.$$

En général, si $X_1,X_2,\ldots,X_n$ sont des v.a. **indépendantes**, alors

$$\mathrm{E}\left[\sum_{i=1}^n X_i\right]=\sum_{i=1}^n \mathrm{E}[X_i]\quad \text{et}\quad \mathrm{Var}\left[\sum_{i=1}^n X_i\right]=\sum_{i=1}^n \mathrm{Var}[X_i]\,.$$

Les v.a. indépendantes et identiquement distribuées

Un cas particulier de ce qui précède se produit lorsque tous les $X_1, \ldots, X_n$ ont **exactement la même distribution** (c-à-d la même f.c.r.). Dans ce cas, nous disons qu'elles sont **indépendantes et identiquement distribuées**, ce qui est traditionnellement abrégé par **iid**.

Si $X_1, \ldots, X_n$ sont iid,

$$\mathrm{E}\left[X_i\right] = \mu \quad \text{et} \quad \mathrm{Var}\left[X_i\right] = \sigma^2 \quad \text{pour } i = 1, \ldots, n,$$

alors

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = n\mu \quad \text{et} \quad \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = n\sigma^2\,.$$

Exemples

1. Un échantillon aléatoire de taille $n = 100$ est prélevé d'une population dont la moyenne est $\mu = 50$ et la variance est $\sigma^2 = 0.25$. Trouvez l'espérance et la variance du **total de l'échantillon** τ.

   Solution: Si $X_1, X_2, \ldots, X_{99}, X_{100}$ sont iid avec $\mathrm{E}[X_i] = \mu = 50$ et $\mathrm{Var}[X] = \sigma^2 = 0.25$ pour $i = 1, \ldots, 100$, nous cherchons $\mathrm{E}\left[\tau\right]$ et $\mathrm{Var}\left[\tau\right]$ pour $\tau = \sum_{i=1}^{n} X_i$.

   Selon les formules iid, ce sont tout simplement

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = 100\mu = 5000 \quad \text{et} \quad \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = 100\sigma^2 = 25\,.$$

2. La valeur moyenne du poids de sacs de mélange de rempotage est de 5 kg, avec un écart-type de 0.2. Si une vendeuse porte 4 sacs (choisis indépendamment dans le stock), alors quelles sont l'espérance et l'écart-type du poids total transporté ?

   Solution: la "population" des poids de sac est implicite. Soit $\{X_1, X_2, X_3, X_4\}$ un échantillon iid de taille $n = 4$, avec $\mathrm{E}[X_i] = \mu = 5$ et $\mathrm{ET}[X_i] = \sigma = 0.2$ (d'où $\mathrm{Var}[X_i] = \sigma^2 = 0.2^2 = 0.04$).

   Soit $\tau = X_1 + X_2 + X_3 + X_4$; selon les formules iid,

   $$\mathrm{E}[\tau] = n\mu = 4 \cdot 5 = 20 \quad \text{et} \quad \mathrm{Var}[\tau] = n\sigma^2 = 4 \cdot 0.04 = 0.16,$$

   d'où $\mathrm{ET}[\tau] = \sqrt{0.16} = 0.4$.

La moyenne d'échantillon (reprise)

La **moyenne d'échantillon** est une statistique typique d'intérêt :

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

Si $X_1, \ldots, X_n$ sont iid, avec $\mathrm{E}[X_i] = \mu$ et $\mathrm{Var}[X_i] = \sigma^2$ pour tout i, alors

$$\mathrm{E}\left[\overline{X}\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\left(n\mu\right) = \mu$$

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

Exemple: un ensemble de balances retourne le poids réel de l'objet objet pesé avec une erreur aléatoire moyenne de 0 et d'écart-type 0.1 g. Trouvez l'écart-type de la moyenne de 9 telles mesures d'un objet.

Solution: supposons que l'objet a un poids réel de μ. L'erreur aléatoire indique que chaque mesure $i = 1, \ldots, 9$ s'écrit sous la forme $X_i = \mu + Z_i$, où $\mathrm{E}[Z_i] = 0$, $\mathrm{ET}[Z_i] = 0.1$, et les Z_i sont iid.

Par conséquent, les X_i sont iid avec $\mathrm{E}[X_i] = \mu$ et $\mathrm{ET}[X_i] = \sigma = 0.1$. La moyenne de $X_1, \ldots, X_n$ (avec $n = 9$) est $\overline{X}$, et

$$\mathrm{E}\left[\overline{X}\right] = \mu \text{ and } \mathrm{ET}\left[\overline{X}\right] = \frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{9}} = \frac{1}{30} \approx 0.033\,.$$

Notez que nous ne connaissons pas la distribution des X_i, seulement leur moyenne et leur variance.

Les sommes de v.a. indépendantes normales

Un autre cas intéressant se produit lorsque nous faisons affaire à **plusieurs v.a. normales indépendantes**.

Supposons que $X_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ pour $i = 1, \cdots, n$, et que tous les X_i sont indépendants. Nous savons que pour la somme $\tau = X_1 + \cdots + X_n$:

$$\begin{aligned} \mathrm{E}[\tau] &= \mathrm{E}[X_1 + \cdots + X_n] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] = \mu_1 + \cdots + \mu_n\,; \\ \mathrm{Var}[\tau] &= \mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = \sigma_1^2 + \cdots + \sigma_n^2\,. \end{aligned}$$

Il s'avère que τ suit aussi une **loi normale**, c-à-d que

$$\tau = \sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2\right).$$

Si $\{X_1, \ldots, X_n\}$ est un échantillon aléatoire tiré d'une population de moyenne μ et de variance σ^2, alors

- $\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = n\mu$ et $\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = n\sigma^2$;
- $\mathrm{E}\left[\overline{X}\right] = \mu$ et $\mathrm{Var}\left[\overline{X}\right] = \sigma^2/n$;
- en outre, si la population suit une **loi normale**, alors $\sum_{i=1}^{n} X_i$ et $\overline{X}$ le font également, c-à-d que

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right) \quad \text{et} \quad \overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Exemple: supposons que la population des poids des étudiants soit normale avec une moyenne de 75 kg et un écart-type de 5 kg. Si 16 étudiants sont choisis au hasard, quelle est la distribution du poids total τ ? Quelle est la probabilité que le poids total dépasse 1250 kg ?

Solution: Si $X_1, \ldots, X_{16} \sim \mathcal{N}(75, 25)$ (iid), le total $\tau = X_1 + \cdots + X_{16}$ suit également une **loi normale**, avec

$$\tau = \sum_{i=1}^{16} X_i \sim \mathcal{N}(16 \cdot 75, 16 \cdot 25) = \mathcal{N}(1200, 400) \text{ et } Z = \frac{\tau - 1200}{\sqrt{400}} \sim \mathcal{N}(0,1).$$

Alors

$$\begin{aligned} P(\tau > 1250) &= P\left(\frac{\tau - 1200}{\sqrt{400}} > \frac{1250 - 1200}{20}\right) \\ &= P(Z > 2.5) = 1 - P(Z \leq 2.5) \approx 1 - 0.9938 = 0.0062\,. \end{aligned}$$

4.4 – Le théorème de la limite centrale

Motivation: un professeur enseigne un cours depuis les 20 dernières années. Pour chaque cours de cette période, les notes des examens de mi-session de tous les étudiants ont été enregistrées.

Soit $X_{i,j}$ la note de l'étudiant i durant l'année j. En consultant les listes de classe, le professeur constate que

$$\mathrm{E}[X_{i,j}] = 65 \quad \text{et} \quad \mathrm{ET}[X_{i,j}] = 15.$$

Cette année, il y a 49 étudiants dans la classe. À quoi le professeur devrait-il s'attendre, en terme de moyenne de la classe à l'examen de mi-session ?

Bien sûr, le professeur ne peut être certain de ce qui va se passer, mais il peut essayer l'approche suivante :

1. il simule les résultats de la classe de 49 étudiants en générant un échantillon de notes $X_{1,1},\ldots,X_{1,49}$ à partir d'une loi quelconque de moyenne 65 et de variance 15^2 (peut-être une loi normale ?) ;

2. il calcule la moyenne d'échantillon et l'enregistre sous la forme $\overline{X}_1$;

3. il répète les étapes 1-2 à m reprises et calcule l'écart-type des moyennes d'échantillon $\overline{X}_1,\ldots,\overline{X}_m$;

4. il trace l'histogramme des moyennes d'échantillon $\overline{X}_1,\ldots,\overline{X}_m$.

À votre avis, qu'est-ce qui en découle ?

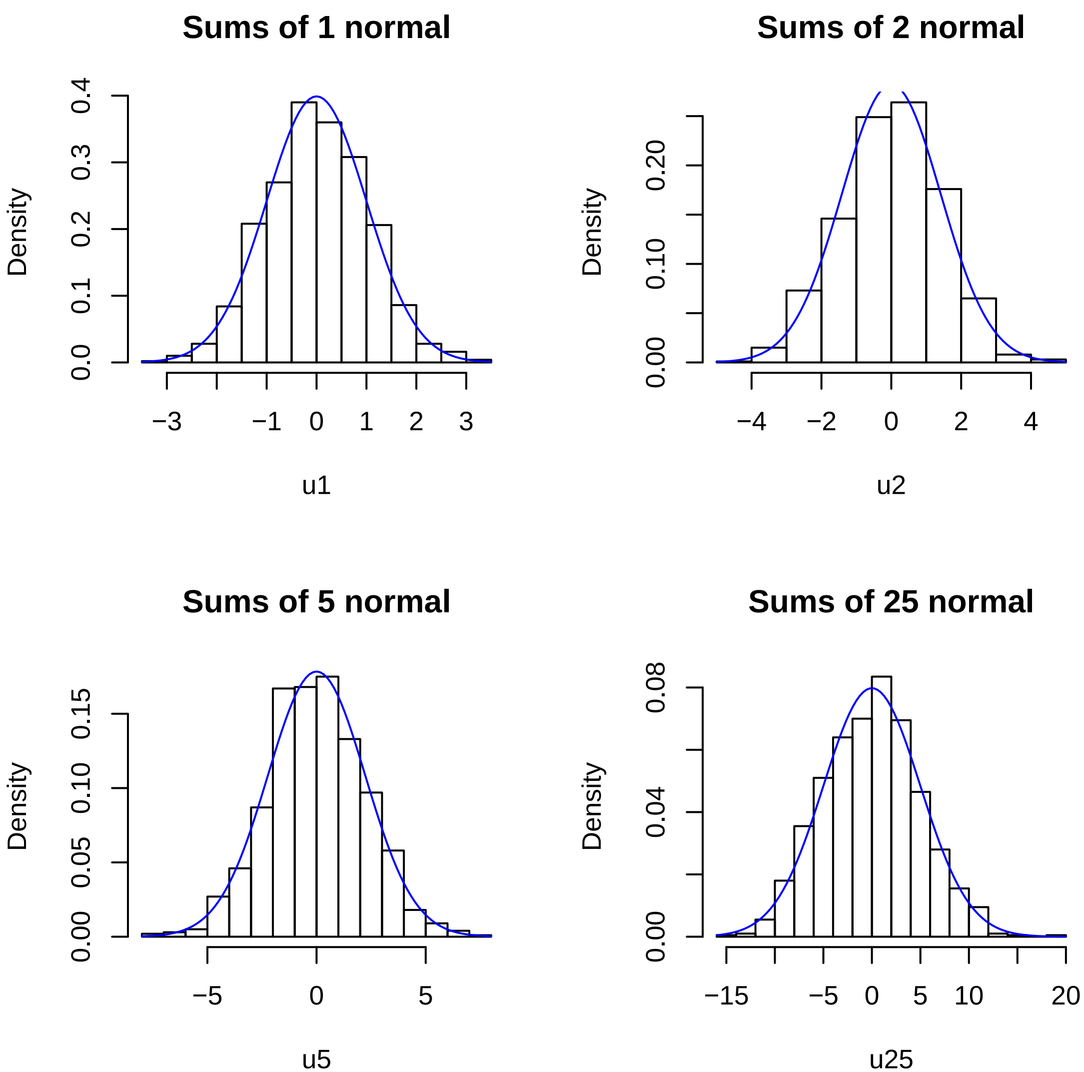
Théorème de la limite centrée: Si $\overline{X}$ est la moyenne d'un échantillon aléatoire de taille n prélevé d'une population **quelconque** de moyenne μ et de variance finie σ^2, alors $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$, suit la loi normale centrée réduite $\mathcal{N}(0,1)$ lorsque $n \to \infty$.

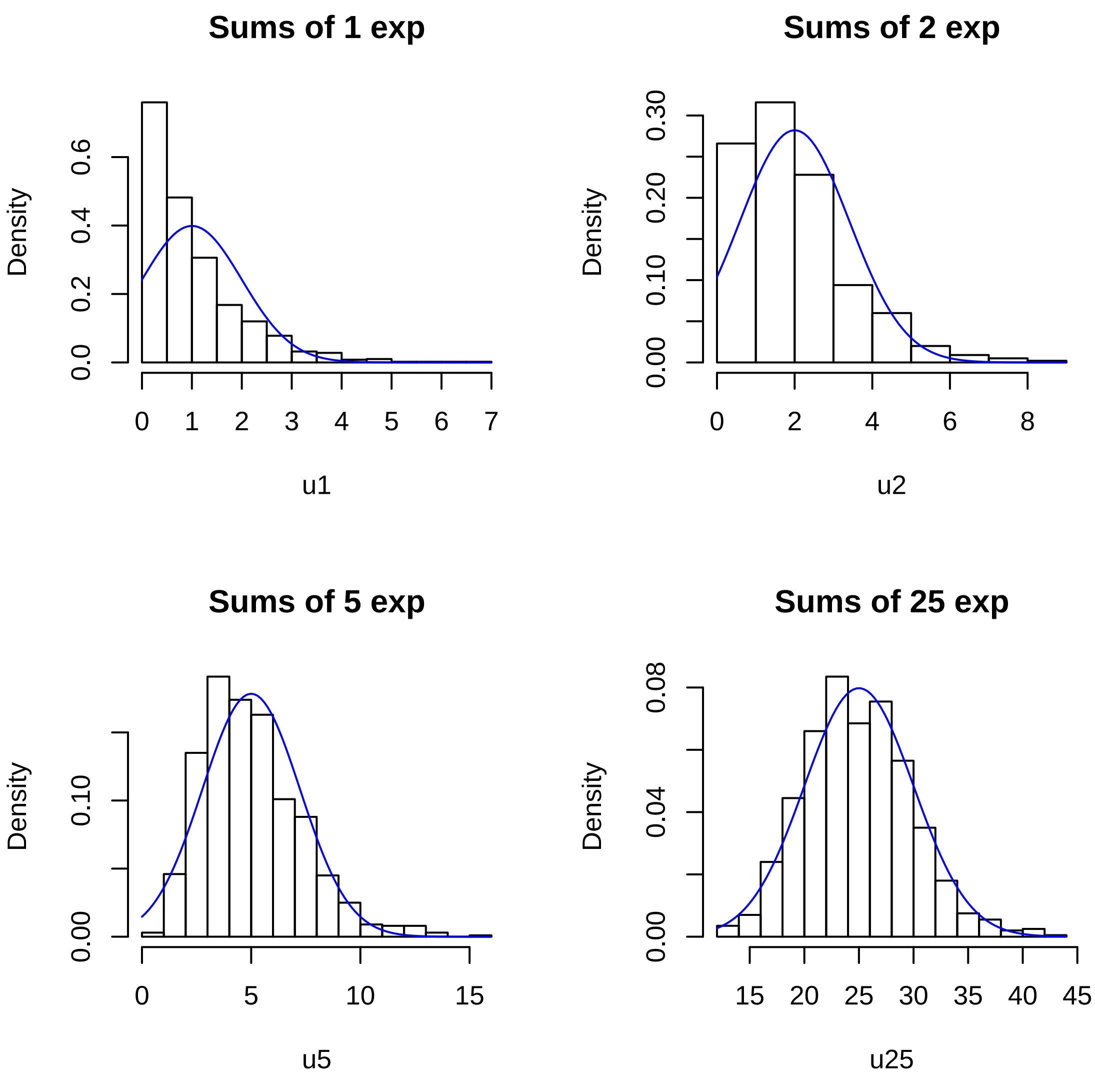
Plus précisément, le TLC est un résultat **asymptotique**. Si nous considérons les v.a.

$$Z_n = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}},$$

en tant que fonctions de n, **sans tenir compte du fait que les X_i soient normaux ou non,** pour chaque z on a

$$\lim_{n\to\infty} P\left(Z_n \leq z\right) = \Phi(z) \quad \text{et} \quad P\left(Z_n \leq z\right) \approx \Phi(z) \text{ si } n \text{ est suffisament élevé.}$$


Sums of 1 normal
Density
0.0 0.1 0.2 0.3 0.4
−3 −1 0 1 2 3
u1
Sums of 2 normal
Density
0.00 0.10 0.20
−4 −2 0 2 4
u2
Sums of 5 normal
Density
0.00 0.05 0.10 0.15
−5 0 5
u5
Sums of 25 normal
Density
0.00 0.04 0.08
−15 −5 0 5 10 20
u25


Sums of 1 exp
Density
0.0 0.2 0.4 0.6
0 1 2 3 4 5 6 7
u1
Sums of 2 exp
Density
0.00 0.10 0.20 0.30
0 2 4 6 8
u2
Sums of 5 exp
Density
0.00 0.10
0 5 10 15
u5
Sums of 25 exp
Density
0.00 0.04 0.08
15 20 25 30 35 40 45
u25

Exemples:

1. Les notes d'examen d'un cours universitaire suivent une loi dont la moyenne est de 56 et l'écart-type, 11. Dans une classe de 49 élèves, quelle est la probabilité que la note moyenne soit inférieure à 50 ? Quelle est la probabilité que la note moyenne se situe entre 50 et 60 ?

   Solution: Soient $X_1, ..., X_{49}$ les notes; supposons que les performances sont **indépendantes**. Selon le TLC,

   $$\overline{X} = (X_1 + X_2 + \cdots + X_{49})/49, \text{ avec } \mathrm{E}[\overline{X}] = 56, \mathrm{Var}[\overline{X}] = 11^2/49.$$

   Nous avons donc

   $$P(\overline{X} < 50) \approx P\left(Z < \frac{50-56}{11/7}\right) = P(Z < -3.82) = 0.0001$$

$$P(50 < \overline{X} < 60) \approx P\left(\frac{50-56}{11/7} < Z < \frac{60-56}{11/7}\right)$$
$$= \quad P(-3.82 < Z < 2.55) = \Phi(2.55) - \Phi(-3.82) = 0.9945.$$

Note: ceci ne dit rien sur la nature de la distribution des notes; Si elles sont **normales**, cependant, les $\approx$ sont remplacés par des $=$.

2. Les mesures de la pression artérielle systolique pour les femmes pré-ménopausées non enceintes âgées de 35 à 40 ans ont une moyenne de 122.6 mm Hg et un écart-type de 11 mm Hg. Un échantillon indépendant de 25 femmes est tiré de cette population cible et leur tension artérielle est enregistrée. Quelle est la probabilité que la pression artérielle moyenne soit supérieure à 125 mm Hg ? Comment la réponse changerait-elle si la taille de l'échantillon passait à 40 ?

Solution: selon le TLC, $\overline{X} \sim \mathcal{N}(122.6, 121/25)$, approximativement. Ainsi

$$P(\overline{X} > 125) \approx P\left(Z > \frac{125 - 122.6}{11/\sqrt{25}}\right)$$
$$= P(Z > 1.09) = 1 - \Phi(1.09) = 0.1378.$$

Si la taille de l'échantillon est de 40, alors

$$P(\overline{X} > 125) \approx P\left(Z > \frac{125 - 122.6}{11/\sqrt{40}}\right) = 0.0838.$$

L'augmentation de la taille de l'échantillon réduit la probabilité que la moyenne soit éloignée de l'espérance de chaque mesure originale.

3. Supposons que nous prélevons l'échantillon aléatoire $\{X_1, \ldots, X_{100}\}$ d'une population dont la moyenne est de 5 et la variance, 0.01. Quelle est la probabilité que la différence entre la moyenne de l'échantillon aléatoire et la moyenne de la population dépasse 0.027 ?

Solution: selon le TLC, nous savons que $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ suit la loi normale centrée réduite, approximativement. La probabilité recherchée est ainsi

$$\begin{aligned}
P(|\overline{X}-\mu| \geq 0.027) &= P(\overline{X}-\mu \geq 0.027 \text{ or } \mu - \overline{X} \geq 0.027) \\
&= P\left(\frac{\overline{X}-5}{0.1/\sqrt{100}} \geq \frac{0.027}{0.1/\sqrt{100}}\right) + P\left(\frac{\overline{X}-5}{0.1/\sqrt{100}} \leq \frac{-0.027}{0.1/\sqrt{100}}\right) \\
&\approx P\left(Z \geq 2.7\right) + P\left(Z \leq -2.7\right) = 2P\left(Z \geq 2.7\right) \\
&\approx 2(0.0035) = 0.007.
\end{aligned}$$

4.5 – Les distributions d'échantillonnage (reprise)
La différence de 2 moyennes

Théorème: Soit $\{X_1, \ldots, X_n\}$ un échantillon aléatoire provenant d'une population de moyenne μ_1 et de variance σ_1^2, et $\{Y_1, \ldots, Y_m\}$ un autre échantillon aléatoire, indépendant du premier, prélevé d'une population de moyenne μ_2 et de variance σ_2^2.

Si $\overline{X}$ et $\overline{Y}$ sont les moyennes respectives des échantillons, alors

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

suit la loi normale centrée réduite $\mathcal{N}(0,1)$ lorsque $n, m \to \infty$; c'est aussi un résultat **limite**.

Exemple: deux machines différentes sont utilisées pour remplir des boîtes de céréales sur une chaîne de production. La mesure critique influencée par ces machines est le poids du produit dans les boîtes. Pour les deux machines, la variance des poids est $\sigma^2 = 1$.

Chaque machine produit un échantillon de 36 boîtes. Quelle est la probabilité que la différence entre les moyennes des échantillons est < 0.2, en supposant que les vraies moyennes sont identiques ?

Solution: nous avons $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2 = 1$, et $n = m = 36$. Ainsi,

$$
\begin{aligned}
P\left(|\overline{X} - \overline{Y}| < 0.2\right) &= P\left(-0.2 < \overline{X} - \overline{Y} < 0.2\right) \\
&= P\left(\frac{-0.2 - 0}{\sqrt{1/36 + 1/36}} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{1/36 + 1/36}} < \frac{0.2 - 0}{\sqrt{1/36 + 1/36}}\right) \\
&= P(-0.8485 < Z < 0.8485) = \Phi(0.8485) - \Phi(-0.8485) \approx 0.6.
\end{aligned}
$$

La variance d'échantillonnage S^2

Théorème : si

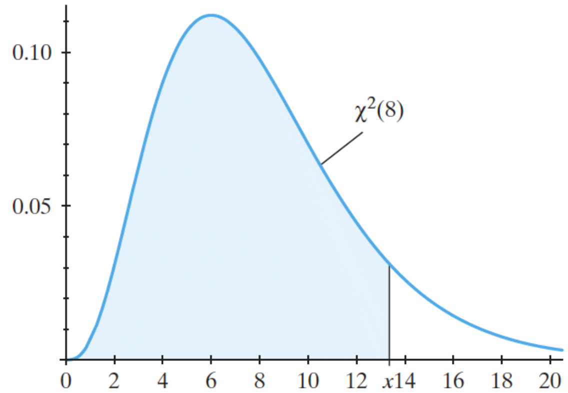
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

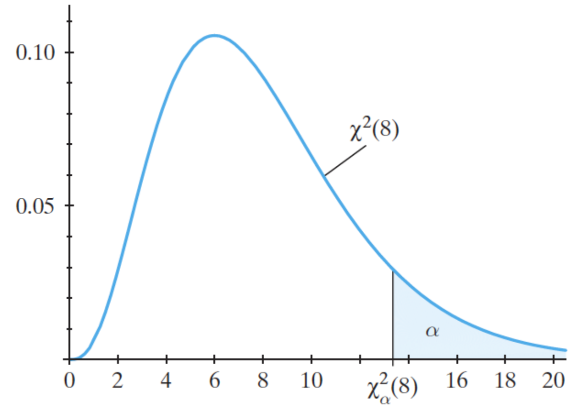
est la variance d'un échantillon aléatoire de taille n prélevé d'une population normale avec une variance σ^2, alors la statistique

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

suit une **loi χ-carré avec $\nu = n-1$ degrés de liberté**. En général, $\chi^2(\nu) = \Gamma(1/2, \nu)$.

Table IV The Chi-Square Distribution





$$P(X \le x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2} dw$$

	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21

Notation: pour $0 < \alpha < 1$ et $\nu \in \mathbb{N}^*$, $\chi^2_\alpha(\nu)$ est la **valeur critique** pour laquelle

$$P(\chi^2 > \chi^2_\alpha(\nu)) = \alpha\,,$$

où $\chi^2 \sim \chi^2(\nu)$ suit une loi chi-carré avec ν degrés de liberté. Nous pouvons trouver la valeur de $\chi^2_\alpha(\nu)$ dans des tableaux (qui seront mis à votre disposition lors de l'examen final, si nécessaire).

Par exemple, lorsque $\nu = 7$ et $\alpha = 0.95$, nous avons $\chi^2_{0.95}(7) = 2.167$, donc $P(\chi^2 > 2.167) = 0.95$, où $\chi^2 \sim \chi^2(7)$, c-à-d que χ^2 suit une loi chi-carré avec $\nu = 7$ degrés de liberté.

Autrement dit, 95% de l'aire sous la courbe de la f.d.p. de $\chi^2(7)$ se trouve à droite de 2.167.

La moyenne d'échantillonnage quand la variance est inconnue

Supposons que $Z \sim \mathcal{N}(0,1)$ et $V \sim \chi^2(\nu)$. Si Z et V sont indépendants, alors la distribution de la variable aléatoire

$$T = \frac{Z}{\sqrt{V/\nu}}$$

est une **loi t de Student avec ν degrés de liberté**, que nous désignons par $T \sim t(\nu)$.

La f.d.p. de $t(\nu)$ est

$$f(x) = \frac{\Gamma(\nu/2 + 1/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)(1 + x^2/\nu)^{\nu/2+1/2}}, \quad x \in \mathbb{R}.$$

Théorème: Soit $X_1, \ldots, X_n$ des v.a. normales indépendantes de moyenne μ et d'écart-type σ . Soit $\overline{X}$ et S^2 la moyenne et la variance de l'échantillon, respectivement. Alors la variable aléatoire

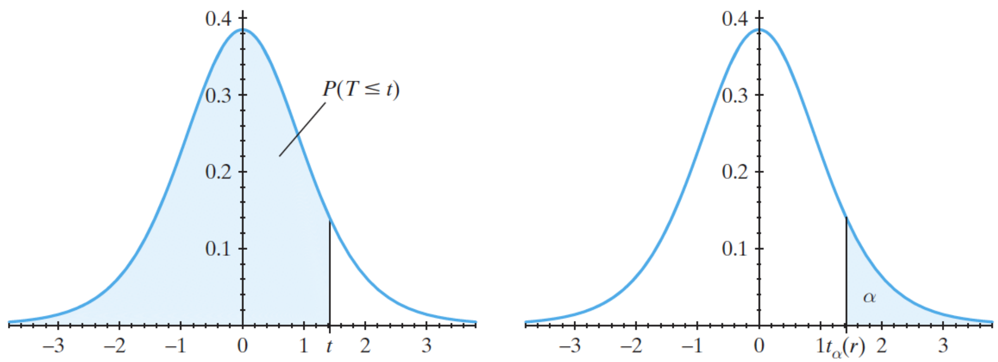
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

suit une **loi t de Student avec $\nu = n - 1$ degrés de liberté**.

Table t: soit $t_\alpha(\nu)$ la **valeur critique** de t à partir de laquelle on retrouve une surface égale à α , c'est-à-dire $P(T > t_\alpha(\nu)) = \alpha$, où $T \sim t(\nu)$.

Pour tout ν, la loi t de Student est une **distribution symétrique autour de zéro**; nous avons donc $t_{1-\alpha}(\nu) = -t_\alpha(\nu)$.

Table VI The t Distribution



$$P(T \le t) = \int_{-\infty}^{t} \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\,\Gamma(r/2)(1+w^2/r)^{(r+1)/2}}\,dw$$

$$P(T \le -t) = 1 - P(T \le t)$$

	$P(T \le t)$						
	0.60	0.75	0.90	0.95	0.975	0.99	0.995
r	$t_{0.40}(r)$	$t_{0.25}(r)$	$t_{0.10}(r)$	$t_{0.05}(r)$	$t_{0.025}(r)$	$t_{0.01}(r)$	$t_{0.005}(r)$
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169

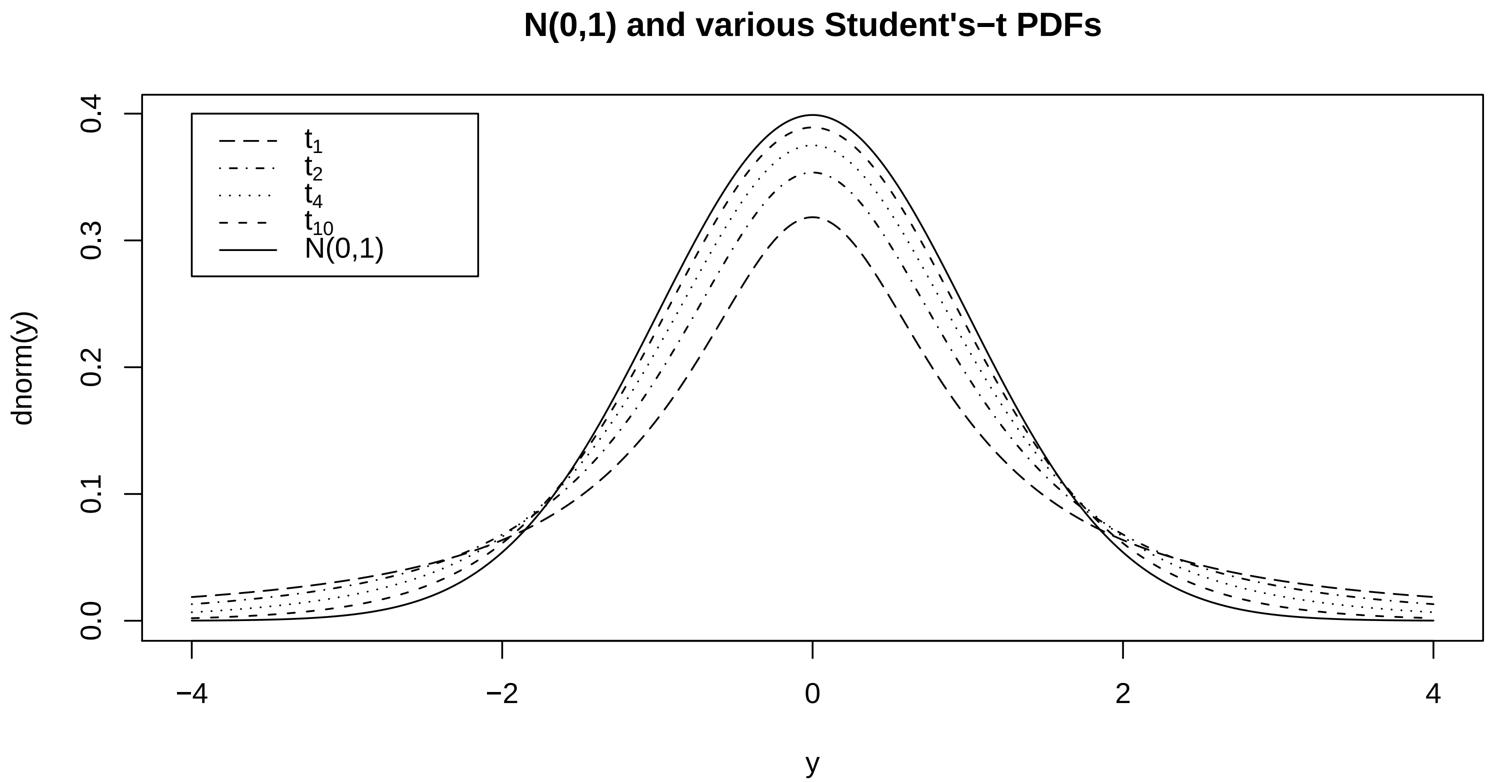
Si $T \sim t(\nu)$, alors pour tout $0 < \alpha < 1$, nous avons

$$\begin{aligned} P\left(-t_{\alpha/2}(\nu) < T < t_{\alpha/2}(\nu)\right) &= P\left(T < t_{\alpha/2}(\nu)\right) - P\left(T < -t_{\alpha/2}(\nu)\right) \\ &= 1 - P\left(T > t_{\alpha/2}(\nu)\right) - \left(1 - P\left(T > -t_{\alpha/2}(\nu)\right)\right) \\ &= 1 - P\left(T > t_{\alpha/2}(\nu)\right) - \left(1 - P\left(T > t_{1-\alpha/2}(\nu)\right)\right) \\ &= 1 - \alpha/2 - (1 - (1 - \alpha/2)) = 1 - \alpha\,, \end{aligned}$$

où la troisième égalité suit par $t_{1-\alpha}(\nu) = -t_\alpha(\nu)$.

Par conséquent,

$$P\left(-t_{\alpha/2}(n-1) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)\right) = 1 - \alpha\,.$$



Lorsque $\nu \to \infty$, $t(\nu) \to \mathcal{N}(0,1)$. Cela va de soit, puisque $S \to \sigma$ lorsque $n \to \infty$.

Exemple: à partir de la table, nous pouvons voir que si $n = 9$,

$$P(T > 2.306) = 0.025 \Longrightarrow P(T < -2.306) = 0.025\,,$$

où $T \sim t(8)$, de sorte que $t_{0.025}(8) = 2.306$ et

$$\begin{aligned} P(|T| \leq 2.306) &= P(-2.306 \leq T \leq 2.306) \\ &= 1 - P(T < -2.306) - P(T > 2.306) \\ &= 0.95\,. \end{aligned}$$

La loi t de Student sera utile lorsque viendra le temps de calculer des **intervalles de confiance** et de faire des **tests d'hypothèse**.

Les lois F

Soient $U \sim \chi^2(\nu_1), V \sim \chi^2(\nu_2)$ des v.a. indépendantes; la distribution de la v.a.

$$F = \frac{U/\nu_1}{V/\nu_2}$$

est une loi **de F avec ν_1 et ν_2 degrés de liberté**, que nous désignons par $F \sim F(\nu_1, \nu_2)$.

La f.d.p. de $F(\nu_1, \nu_2)$ est

$$f(x) = \frac{\Gamma(\nu_1/2 + \nu_2/2)(\nu_1/\nu_2)^{\nu_1/2} x^{\nu_1/2-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)(1 + x\nu_1/\nu_2)^{\nu_1/2+\nu_2/2}}, \quad x \geq 0.$$

Table VII *continued*

$$P(F \leq f) = \int_0^f \frac{\Gamma[(r_1+r_2)/2](r_1/r_2)^{r_1/2} w^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(1+r_1 w/r_2)^{(r_1+r_2)/2}}\, dw$$

		Den. d.f.	Numerator Degrees of Freedom, r_1									
α	$P(F \leq f)$	r_2	1	2	3	4	5	6	7	8	9	10
0.05	0.95	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
0.025	0.975		647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63
0.01	0.99		4052	4999.5	5403	5625	5764	5859	5928	5981	6022	6056
0.05	0.95	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
0.025	0.975		38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
0.01	0.99		98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
0.05	0.95	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
0.025	0.975		17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
0.01	0.99		34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
0.05	0.95	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
0.025	0.975		12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
0.01	0.99		21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
0.05	0.95	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
0.025	0.975		10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
0.01	0.99		16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05

Théorème: si S_1^2 et S_2^2 sont les variances empiriques de deux échantillons aléatoires indépendants de taille n et m, respectivement, prélevés de populations normales avec variances respectives de σ_1^2 et σ_2^2 , alors

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

suit une loi F **avec** $\nu_1 = n-1$ **et** $\nu_2 = m-1$ **degrés de liberté**.

Notation: pour $0 < \alpha < 1$ et $\nu_1, \nu_2 \in \mathbb{N}^*$, $f_\alpha(\nu_1, \nu_2)$ est la **valeur critique** pour laquelle $P(F > f_\alpha(\nu_1, \nu_2)) = \alpha$, si $F \sim F(\nu_1, \nu_2)$.

On peut montrer que $f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)}$.

On peut trouver les valeurs de $f_\alpha(\nu_1, \nu_2)$ dans des tables (ou avec des logiciels). Par exemple, nous avons $f_{0.95}(5, 10) = \frac{1}{f_{0.05}(10,5)} = \frac{1}{4.74} = 0.211$.