Statistique

Laurent Rouvière

Septembre 2020

Table des matières

Présentation

— *Preé-requis* : Bases de **R**, probabilités, statistique et programmation

— *Objectifs* : être capable de mettre en oeuvre une démarche statistique rigoureuse pour répondre à des problèmes standards

  — estimation : ponctuelle et par intervalles

  — tests d'hypothèses

  — modèle linéaire

— *Enseignant* : Laurent Rouvière, `laurent.rouviere@univ-rennes2.fr`

  — Thèmes de recherche : statistique non-paramétrique et apprentissage statistique
  — Enseignement : probabilités, statistique et logiciels (Universités et écoles)
  — Consulting : energie (ERDF), finance, marketing.

Plan

— *Théorie* (modélisation statistique) et pratique sur machines (R).

1. Introduction à R

  — Environnement Rstudio
  — Objets R

    — Manipulation et visualisation de données

2. "Rappels" de probabilités
3. Estimation ponctuelle et par intervalle
4. Introduction aux tests.

Quelques éléments de probabilité

Introduction

Une problématique...

Exemple

Les iris de Fisher.

1. Quelle est la longueur de pétales moyenne des iris ?
2. Peut-on dire que la longueur moyenne est égale à 5.6 ?
3. Les Setosa ont-elles des longueurs de pétales plus longues que les autres espèces ? Avec quel niveau de confiance ?

Des données

Collecte de données

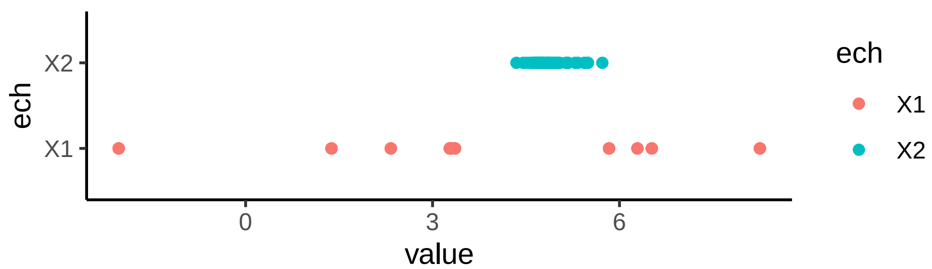
— Pour répondre à ces questions on réalise des expériences.

— Exemple : on mesure les longueurs et largeurs de sépales et pétales pour 150 iris (50 de chaque espèce).

```
> data(iris)
> summary(iris)
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
##        Species  
##  setosa    :50  
##  versicolor:50  
##  virginica :50  
##                 
##                 
##                 
```

Autre exemple

— On considères deux échantillons $X1$ et $X2$.
— *Question* : la moyenne est-elle égale à 5 ?



Remarque

Plus difficile de répondre pour $X1$ car :

— Moins d'observations ;

— Dispersion plus importante.

Un autre exemple

— Deux candidats se présentent à une élection.

— *Problématique* : qui va gagner ?

— Avec quel niveau de confiance peut-on répondre à cette question ?

```
> summary(election)
##  res
##  A:488
##  B:512
```

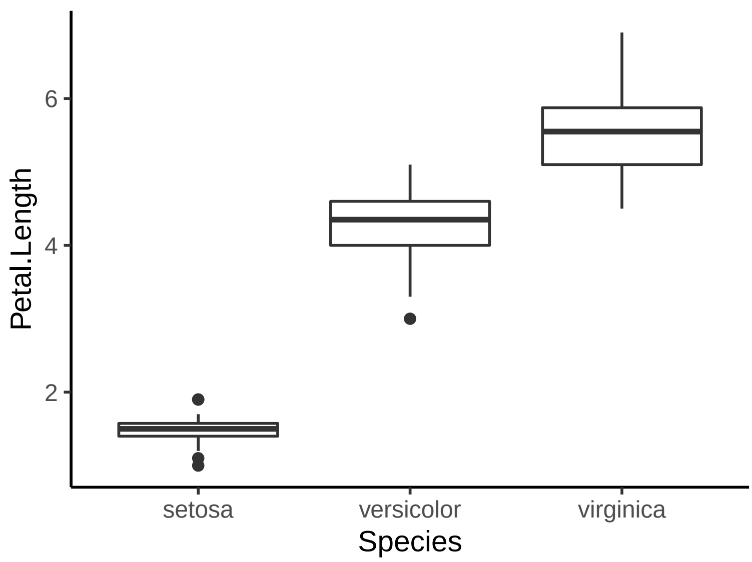
Statistiques descriptives et visualisation

Ces approches peuvent donner une intuition pour répondre.

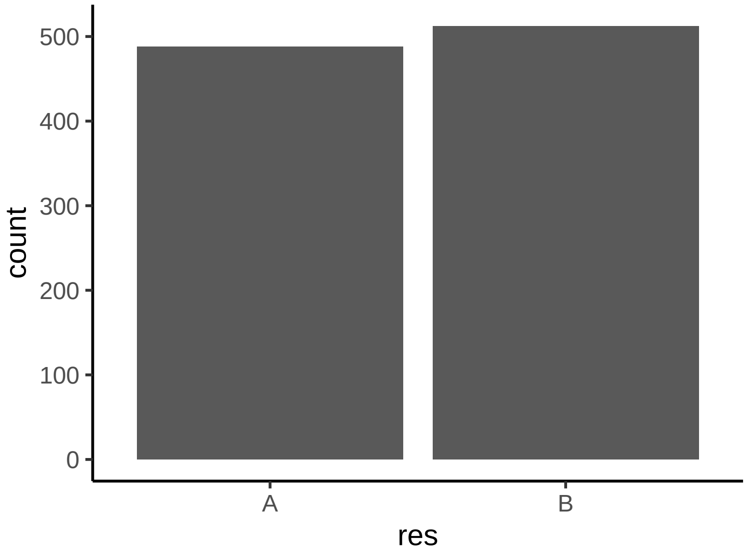
```
> iris %>% summarize(mean(Petal.Length))
##   mean(Petal.Length)
## 1              3.758
> iris %>% group_by(Species) %>% summarize(mean(Petal.Length))
## # A tibble: 3 x 2
##   Species    `mean(Petal.Length)`
##   <fct>                     <dbl>
## 1 setosa                     1.46
## 2 versicolor                 4.26
## 3 virginica                  5.55
```

```
> election %>% mutate(res_A=res=="A") %>%
+   summarize(Prop_A=mean(res_A))
##   Prop_A
## 1  0.488
```

```
> ggplot(iris)+aes(x=Species,y=Petal.Length)+geom_boxplot()
```



```
> ggplot(election)+aes(x=res)+geom_bar()
```



Hasard, aléa...

— La réponse à ces questions peut *paraître* simple.

Première réponse

— Iris : si la longueur moyenne des pétales mesurées est différente de 0.6, on répond non.

— Election : si la proportion de sondés votant pour A est supérieure à 0.5, on répond que A gagne.

Problème

— Ces réponses sont très (trop) *liées aux données observées*.

— Si je recommence l'expérience (sur d'autres iris ou d'autres électeurs), les *conclusions peuvent changer*.

— Conclusion : il faut prendre en compte cet aléa du au choix des individus ainsi que le *nombre d'observations* et la *dispersion des mesures*.

Probabilités

— *Idée* : répondre à ces questions en calculant (estimant) des probabilités.

— *Notation* : $x_1, \dots, x_n$ n observations.

Hypothèse

Les observations proviennent d'une certaine *loi de probabilité* (inconnue).

Problème

Qu'est-ce qu'une loi de probabilité ?

"Définition"

- Une loi de probabilité est un objet qui permet de mesurer la probabilité qu'un évènement se produise.
- *Mathématiquement*, il s'agit d'une fonction $\mathbf{P} : \Omega \rightarrow [0,1]$ telle que, pour un évènement $\omega \in \Omega$, $\mathbf{P}(\omega)$ mesure la "chance" que l'évènement ω se réalise.

Exemple

- *Pile ou face* : $\mathbf{P}(pile) = \mathbf{P}(false) = 1/2$.
- *Dé équilibré* : $\mathbf{P}(1) = \mathbf{P}(2) = \cdots = \mathbf{P}(6) = 1/6$.

Quelques lois de probabilités

- Une loi de probabilité permet de *visualiser/caractériser/mesurer* les valeurs que peut prendre une variable.
- On distingue *deux types* de loi de probabilité que l'on caractérise en étudiant les valeurs possibles de la variable (et donc de l'expérience).

Variable discrète

- Si l'ensemble des valeurs que peut prendre la variable est fini ou dénombrable, la variable est discrète.
- pile ou face, nombre de voitures à un feu rouge, nombre d'aces dans un match de tennis...

Variable continue

- Si l'ensemble des valeurs que peut prendre la variable est infini ($\mathbb{R}$ ou un intervalle de $\mathbb{R}$) la variable est continue.
- Duret de trajet, taille, vitesse d'un service, longueur d'un saut...

Comment définir une loi discrète ?

Pour *caractériser un loi discrète*, il faudra donner :

1. l'ensemble des valeurs possibles de la variable ;
2. la probabilité associée à chacune de ses valeurs.

Exemple

- Soit X la variable aléatoire qui représente le statut matrimonial d'une personne.
- X peut prendre 4 valeurs : célibataire, marié, divorcé, vœuf (4 valeurs donc loi discrète).
- On caractérise sa loi

$$\mathbf{P}(X = \text{cel}) = 0.20,\ \mathbf{P}(X = \text{marié}) = 0.4,\ \mathbf{P}(X = \text{div}) = 0.3,\ \mathbf{P}(X = \text{vœuf}) = 0.1.$$

Remarque

La somme des probabilités doit *toujours être égale à 1.*

Bernoulli

Définition

La loi de Bernoulli de paramètre $p \in [0, 1]$ est définie par

— Valeurs possibles : 0 et 1

— Proba : $\mathbf{P}(X = 0) = 1 - p$ et $\mathbf{P}(X = 1) = p$.

Exemple

— Modélisation de phénomènes à 2 *issues*.

— Pile ou face, ace/pas ace, acceptation/rejet, oui/non...

Binomiale

— On répète n expériences de *Bernoulli* de paramètres $p \in [0, 1]$ de façon *indépendante*.

— On note $X_1, \ldots, X_n$ les n résultats.

— $\sum_{i=1}^{n} X_i$ (qui compte le nombre de 1) suit une loi *Binomiale* $\mathcal{B}(n, p)$.

Loi binomiale

— Valeurs possibles : $\{0, 1, \ldots, n\}$.

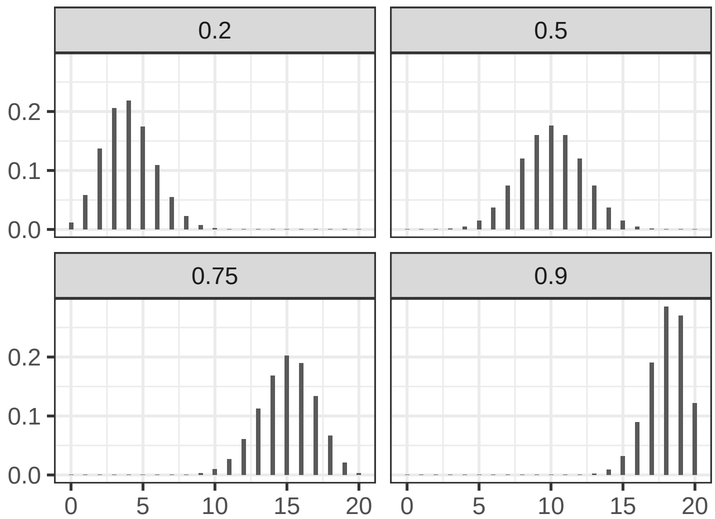
— Proba :

$$\mathbf{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Exemple

— Nombre de *succès* sur n épreuves.

— Nombre de piles, nombre d'aces sur n services.



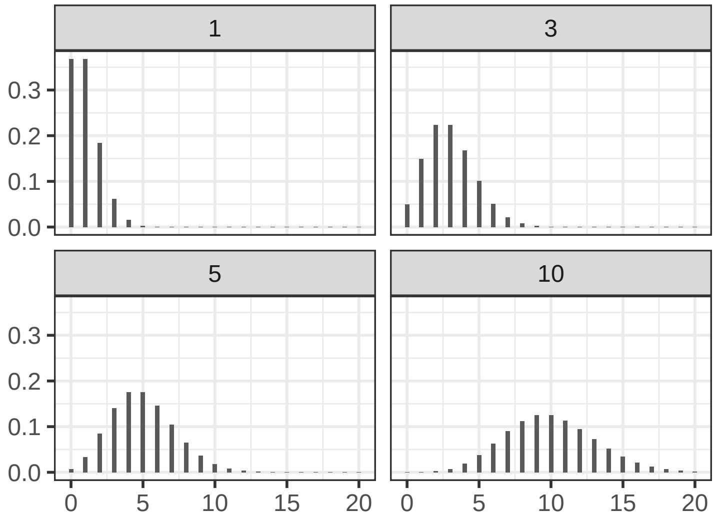
Loi de Poisson

Définition

- Valeurs possibles : $\mathbb{N}$.
- Proba :
$$\mathbf{P}(X = k) = \frac{\lambda^k \exp(-\lambda)}{k\,!}$$
où λ est un paramètre positif. On la note $\mathcal{P}(\lambda)$.

Exemple

- Données de *comptage*.
- Nombre de voitures à un feu rouge, nombre de personnes à une caisse, nombre d'admis à une épreuve...

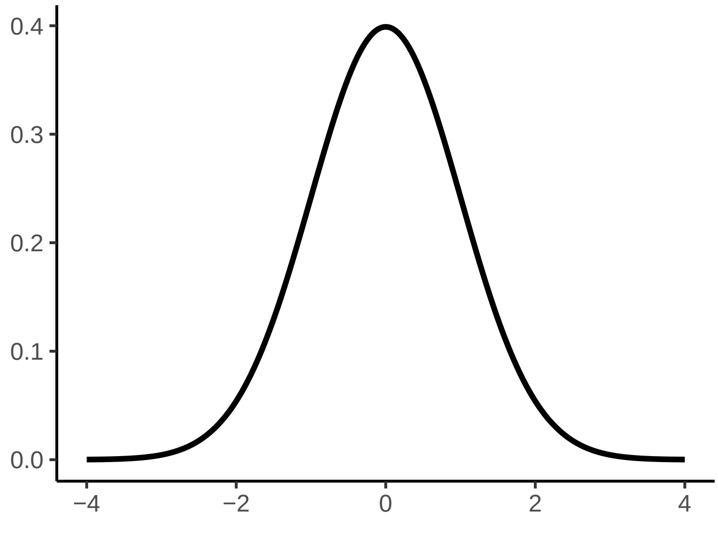


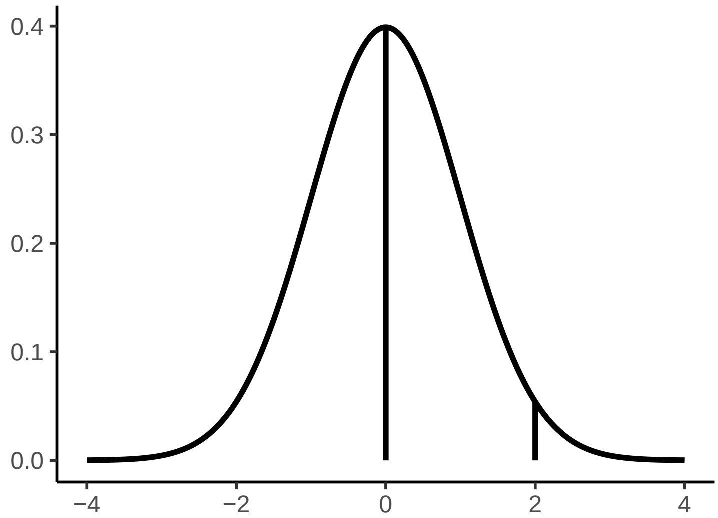
Comment définir une loi continue ?

- Une loi continue prend une *infinité* de valeurs (sur un intervalle ou sur $\mathbb{R}$ tout entier).
- Pour la caractériser on utilisera une *fonction de densité* qui permettra de *mesurer la probabilité* que la variable appartienne à un intervalle.
- Cette probabilité se déduit de l'*aire* sous la densité.

Exemple

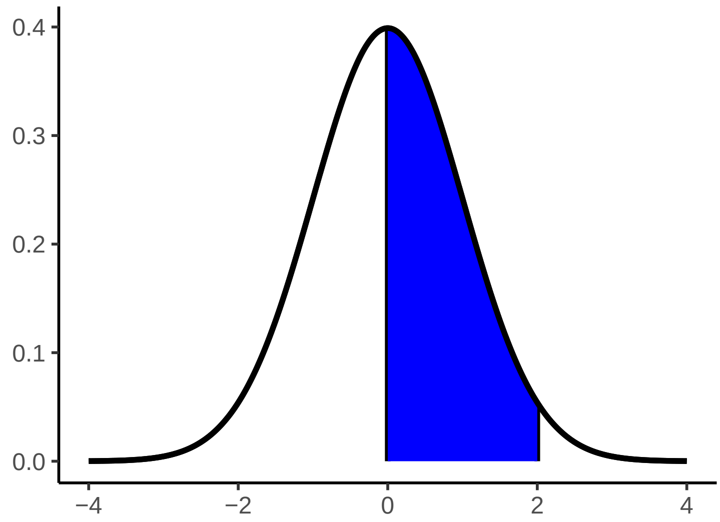
Si X admet pour densité f, alors
$$\mathbf{P}(X \in [a, b]) = \int_a^b f(x)\,\mathrm{d}x.$$





Question

$$\mathbf{P}(X \in [0,2]) = ???$$



Réponse

$$\mathbf{P}(X \in [0,2]) = \int_0^2 f(x)\,\mathrm{d}x \simeq 0.48.$$

Densité

Définition

Une densité de probabilité est donc une *fonction* $f : \mathbb{R} \to \mathbb{R}$ qui doit vérifier les trois propriétés suivantes :

1. Elle doit être positive : $f(x) \geq 0\ \forall x \in \mathbb{R}$;
2. Elle doit être intégrable.
3. Son intégrale sur $\mathbb{R}$ doit être égale à un :

$$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1.$$

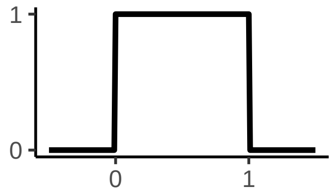
Loi uniforme

Définition

La loi uniforme sur un intervalle $[a,b]$ admet pour densité

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a,b] \\ 0 & \text{sinon.} \end{cases}$$

On la note $\mathcal{U}_{[a,b]}$.



Interprétation

Les valeurs de X sont réparties *uniformément* sur l'intervalle $[a,b]$.

La loi normale

Définition

La loi normale ou loi gaussienne de paramètre $\mu \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}^+$ admet pour densité

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

On la note $\mathcal{N}(\mu, \sigma^2)$.

Remarque

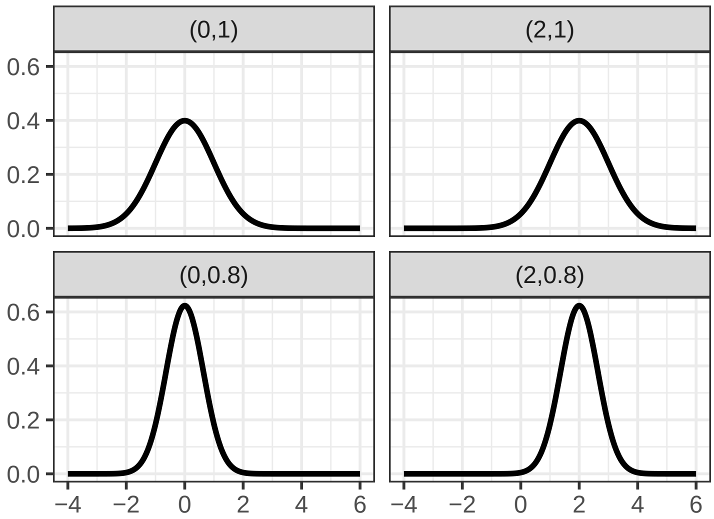
— μ représente le tendance centrale de la loi, on parle de valeur *moyenne*.

— σ^2 représente la *dispersion* de la loi autour de la valeur moyenne, on parle(ra) de *variance*.

— Elle permet de modéliser des phénomènes centrés en une valeur.

— C'est la loi limite du *théorème central limite*.

Exemples pour différents (μ, σ^2)



Loi exponentielle

Définition

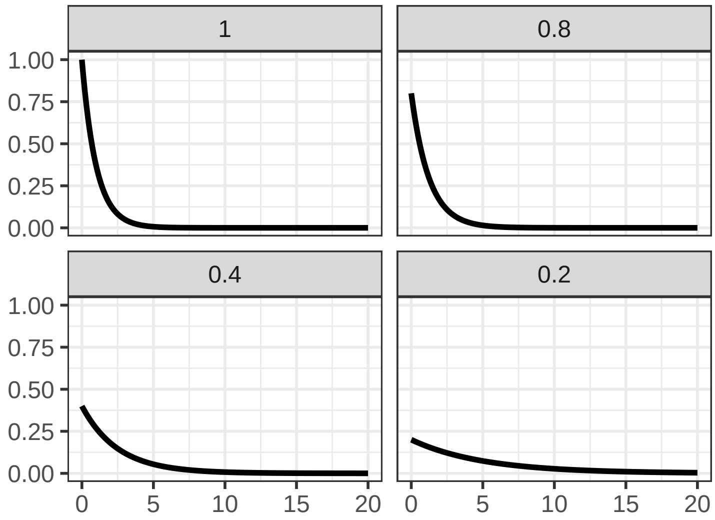
La loi exponentielle de paramètre $\lambda > 0$ admet pour densité

$$f(x) = \lambda \exp(-\lambda x), \quad x \in \mathbb{R}^+.$$

On la note $\mathcal{E}(\lambda)$.

Exemple

— Cette loi est souvent utilisée pour modéliser des *durées de vie* (composant électronique, patients atteint d'une pathologie...).



Espérance et variance

Motivations

- Loi de probabilité : *pas toujours facile à interpréter* d'un point de vue pratique.
- *Objectif* : définir des indicateurs (des nombres par exemple) qui permettent d'interpréter une loi de probabilité (tendance centrale, dispersion...).

Espérance

Définition

L'*espérance* d'une variable aléatoire X est le *réel* défini par :

$$\mathbf{E}[X] = \int_{\Omega} X(\omega)\, \mathrm{d}\mathbf{P}(\omega).$$

Interprétation

- La formule ci-dessus ne sera d'aucun intérêt pratique, elle permet juste de *comprendre l'interprétation de l'espérance.*
- L'espérance revient à intégrer les valeurs de la v.a.r. X pour chaque évènement ω pondéré par la mesure de probabilité de chaque évènement.
- Elle s'interprète ainsi en terme de valeur moyenne prise par X.

Calculs d'espérance

- Pour les *calculs d'espérance*, on distingue les cas discrets et continus.

Propriété

- Cas discret :

$$\mathbf{E}[X] = \sum_{\text{valeurs possibles de } X} x\mathbf{P}(X = x).$$

- Cas continu :

$$\mathbf{E}[X] = \int_{-\infty}^{+\infty} x f(x)\, \underset{\cdot}{\mathrm{x}}$$

où f est la densité de X.

Exemples

Loi	Espérance
$\mathcal{B}(p)$	p
$\mathcal{B}(n,p)$	np
$\mathcal{P}(\lambda)$	λ
$\mathcal{U}_{[a,b]}$	$\frac{a+b}{2}$
$\mathcal{N}(\mu,\sigma^2)$	μ

Variance

Définition

- La *variance* de X, notée $\mathbf{V}[X]$, est définie par :

$$\mathbf{V}[X] = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

- Sa racine carrée positive $\sigma[X]$ est appelée *écart-type* de X.

Interprétation

- La variance est un réel positif.
- Elle mesure l'écart entre les valeurs prises par X et l'espérance (moyenne) de $X \Longrightarrow$ interprétation en terme de dispersion.

Exemple

1. Loi de Bernoulli $\mathcal{B}(p)$: $\mathbf{V}[X] = p(1-p)$;
2. Loi uniforme sur $[0,1]$: $\mathbf{V}[X] = 1/12$;
3. Loi uniforme sur $[1/4, 3/4]$: $\mathbf{V}[X] = 1/48$;

\end{frame}

Espérance et variance de quelques lois classiques

X	$\mathbf{E}[X]$	$\mathbf{V}[X]$
$\mathcal{B}(p)$	p	$p(1-p)$
$\mathcal{B}(n,p)$	p	$np(1-p)$
$\mathcal{P}(\lambda)$	λ	λ

Lois discrètes

X	$\mathbf{E}[X]$	$\mathbf{V}[X]$
$\mathcal{U}_{[a,b]}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2
$\mathcal{E}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Lois continues

Modèle et estimation

L'exemple du décathlon

- On s'intéresse aux *performances de décathloniens* au cours de deux épreuves (jeux olympiques et décastar)

Quelques problèmes

1. Quelle est la *distribution* de la variable vitesse au 100m ?
2. Les *performances* aux decastar et aux jeux olympiques sont-elles *identiques* ?
3. Quelles sont les disciplines les plus *influentes* sur le classement ?
4. Existe t-il un *lien* entre les performances au 100m et les autres disciplines ?
5. Si oui, peut-on le *quantifier* ?

Les données

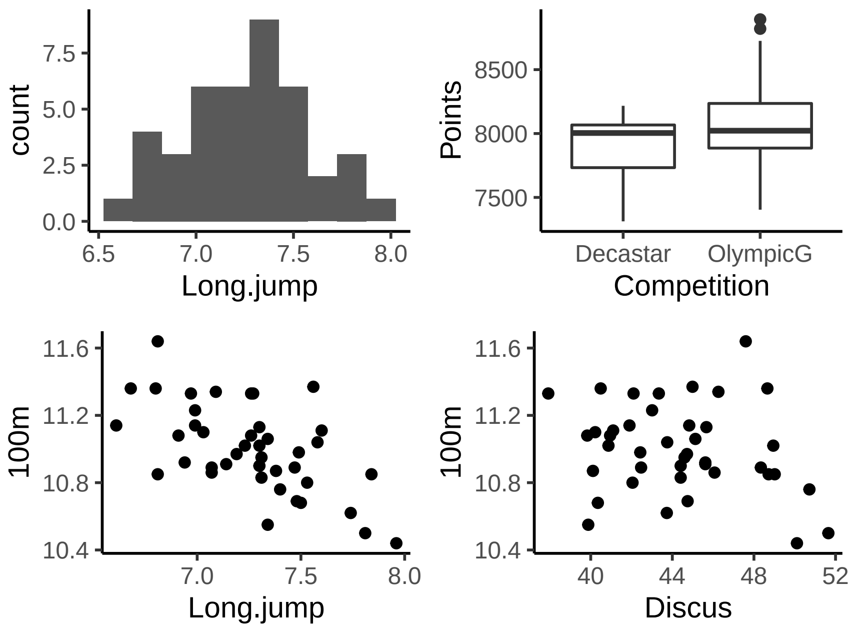
— Pour tenter de répondre à ces questions, on dispose des performances d'une vingtaine de décathloniens au cours de deux épreuves :

```
> head(decathlon)
##           100m Long.jump Shot.put High.jump  400m 110m.hurdle Discus
## SEBRLE   11.04      7.58    14.83      2.07 49.81       14.69  43.75
## CLAY     10.76      7.40    14.26      1.86 49.37       14.05  50.72
## KARPOV   11.02      7.30    14.77      2.04 48.37       14.09  48.95
## BERNARD  11.02      7.23    14.25      1.92 48.93       14.99  40.87
## YURKOV   11.34      7.09    15.19      2.10 50.42       15.31  46.26
## WARNERS  11.11      7.60    14.31      1.98 48.68       14.23  41.10
##          Pole.vault Javeline 1500m Rank Points Competition
## SEBRLE         5.02    63.19 291.7    1   8217    Decastar
## CLAY           4.92    60.15 301.5    2   8122    Decastar
## KARPOV         4.92    50.31 300.2    3   8099    Decastar
## BERNARD        5.32    62.77 280.1    4   8067    Decastar
## YURKOV         4.72    63.44 276.4    5   8036    Decastar
## WARNERS        4.92    51.77 278.1    6   8030    Decastar
```

Statistiques descriptives (capital)

```
> p1 <- ggplot(decathlon)+aes(x=Long.jump)+geom_histogram(bins=10)
> p2 <- ggplot(decathlon)+aes(x=Competition,y=Points)+geom_boxplot()
> p3 <- ggplot(decathlon)+aes_(x=as.name("Long.jump"),
+                              y=as.name("100m"))+geom_point()
> p4 <- ggplot(decathlon)+aes_(x=as.name("Discus"),
+                              y=as.name("100m"))+geom_point()
```

```
> library(gridExtra)
> grid.arrange(p1,p2,p3,p4,nrow=2)
```



Modèle statistique

— On s'intéresse d'abord uniquement à la variable 100m.
— On dispose de $n = 41$ observations $x_1, \dots, x_n$

```
> decathlon %>% summarize(moy=mean(`100m`),var=var(`100m`),
+                         min=min(`100m`),max=max(`100m`))
##        moy       var   min   max
## 1 10.99805 0.0691811 10.44 11.64
```

Question

Peut-on dire que le temps moyen au 100m pour les décathloniens est de 10.99 ?

Hazard, aléa...

— Le résultat de 10.99 dépend des *conditions* dans lesquelles l'expérience a été réalisée.

— Si on *re-mesure les performances de nouvelles compétitions*, il est fort possible qu'on n'obtienne *pas la même durée moyenne.*

Remarque

— Nécessité de prendre en compte que le résultat observé dépend des conditions expérimentales.
— Ces conditions expérimentales vont cependant être difficiles à caractériser précisément.
— On dit souvent que le hasard ou l'aléa intervient dans ces conditions.
— L'approche *statistique* prend en compte le nombre et la dispersion des observations pour apporter une réponse.

Modèle statistique

— Pour prendre en compte cet aléa, on fait l'hypothèse que les observations x_i sont issues d'une loi de probabilité $\mathbf{P}_i$ (inconnue).

Echantillon i.i.d

— Si les mesures x_i sont faites de façons indépendantes et dans des conditions identiques, on dit que $x_1, \dots, x_n$ sont n observations indépendantes et de même loi $\mathbf{P}$.
— On emploi souvent le terme *échantillon i.i.d* (indépendant et identiquement distribué).

Le problème statistique

Estimer

— La loi $\mathbf{P}$ ainsi que toutes ses quantités dérivées (espérance, variance) est et sera toujours inconnue.
— Le travail du statisticien sera d'essayer de retrouver, ou plutôt d'*estimer*, cette loi ou les quantités d'intérêt qui dépendent de cette loi.

Quelques exemples

Efficacité d'un traitement

- On souhaite tester l'efficacité d'un nouveau traitement (autorisé) sur les performances d'athlètes.
- On traite $n = 100$ patients athlètes.
- A l'issue de l'étude, 72 patients ont amélioré leurs performances.

Modélisation

- On note $x_i = 1$ si le $i^{\text{ème}}$ athlète a amélioré, 0 sinon.
- Les x_i sont issues d'une loi de Bernoulli de paramètre inconnu $p \in [0, 1]$.
- Si les individus sont choisis de manière indépendante et ont tous la même probabilité de progresser (ce qui peut revenir à dire qu'ils sont au même niveau), il est alors raisonnable de supposer que l'échantillon est *i.i.d.*

Le problème statistique

Estimer le paramètre p :

$$p = \mathbf{P}(X = 1) = \mathbf{P}(\text{"Athlète améliore"}).$$

Exemple d'estimateur

- Il parait naturel d'estimer p par la proportion d'athlètes dans l'échantillon qui ont amélioré leur performance.
- Cela revient à estimer p par la moyenne (empirique) des x_i :

$$\hat{p} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Durée de trajet

- On s'intéresse à la *durée de trajet moyenne* "domicile/travail".
- *Expérience* : je mesure la durée de trajet domicile/travail pendant plusieurs jours.
- Je récolte $n = 100$ observations :

```
> summary(duree_ht)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.62   16.42   18.46   19.37   21.88   30.20
```

Modélisation

Les données sont issues d'une loi inconnue $\mathbf{P}$.

Le problème statistique

Estimer l'espérance (moyenne) μ de la loi $\mathbf{P}$.

Exemple d'estimateur

Là encore, un estimateur naturel de μ est donné par la moyenne empirique

$$\hat{\mu} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

X	Paramètre	Estimateur
$\mathcal{B}(p)$	p	$\bar{x}_n$
$\mathcal{P}(\lambda)$	λ	$\bar{x}_n$
$\mathcal{U}_{[0,\theta]}$	θ	$2\bar{x}_n$
$\mathcal{E}(\lambda)$	λ	$1/\bar{x}_n$
$\mathcal{N}(\mu,\sigma^2)$	μ et σ^2	$\bar{x}_n$ $\frac{1}{n}\sum_{i=1}^n(x_i-\bar{x}_n)^2$

Le modèle gaussien

Cadre

- $x_1,\dots,x_n$ i.i.d. de loi $\mathcal{N}(\mu,\sigma^2)$.
- Le problème : estimer $\mu = \mathbf{E}[X]$ et $\sigma^2 = \mathbf{V}[X]$.

Exemple d'estimateurs

- *Moyenne empirique* :
$$\hat{\mu} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i.$$
- *Variance empirique* :
$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2.$$

Autres exemples

Conclusion

De nombreux estimateurs sont construits à partir de la moyenne empirique $\bar{x}_n$.

La moyenne empirique

Remarque

- De nombreux estimateurs sont construits à partir de la moyenne empirique
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i.$$
- La moyenne empirique est une variable aléatoire.
- Elle va donc posséder une loi, une espérance, une variance...

Cas gaussien

- On se place tout d'abord dans le cas où les observations suivent une *loi gaussienne*.
- On considère alors $X_1,\dots,X_n$ un échantillon i.i.d. de loi $\mathcal{N}(\mu,\sigma^2)$.

Propriété

— Dans le cas gaussien, la moyenne empirique $\bar{X}_n$ suit une loi normale $\mathcal{N}(\mu, \sigma^2/n)$.

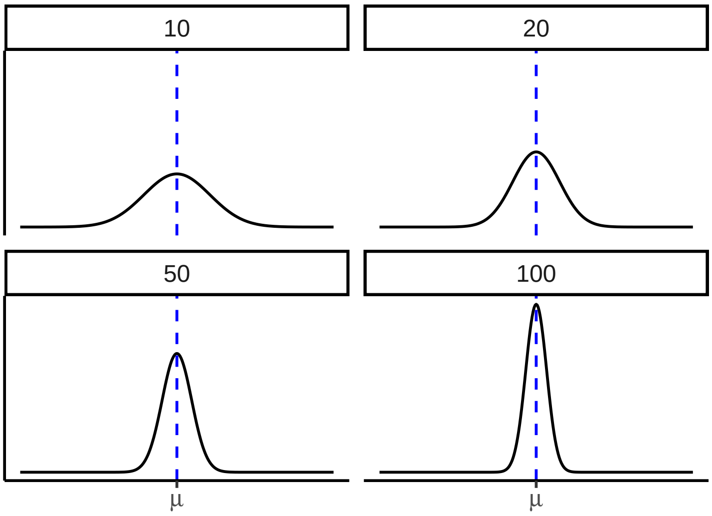
— On a ainsi

$$\mathbf{E}[\bar{X}_n] = \mu \quad \text{et} \quad \mathbf{V}[\bar{X}_n] = \frac{\sigma^2}{n}.$$

Conclusion

— $\bar{X}_n$ est centrée autour de μ.

— Sa dispersion dépend de σ^2 et n.



Biais et variance

— $\bar{X}_n$ tombe toujours en moyenne sur μ. On dit que c'est un estimateur sans biais de μ.

— Sa précision augmente lorsque :

  — σ^2 diminue (difficile à contrôler) ;

  — n augmente (lorsqu'on augmente le nombre de mesures).

Cas non gaussien

— On dispose ici d'un échantillon $X_1, \dots, X_n$ i.i.d. (de *même loi*).

— La loi est *quelconque* (discrète, continue...). On note $\mu = \mathbf{E}[X_1]$ et $\sigma^2 = \mathbf{V}[X_1]$.

Propriété

On a

$$\mathbf{E}[\bar{X}_n] = \mu \quad \text{et} \quad \mathbf{V}[\bar{X}_n] = \frac{\sigma^2}{n}.$$

Commentaires

— L'espérance et la variance de $\bar{X}_n$ sont identiques au cas gaussien.

— Les remarques faites dans le cas gaussien restent donc valables.

— Seul changement : on ne connait pas ici la loi de $\bar{X}_n$ (juste son espérance et sa variance).

- Dans de nombreuses applications (*intervalles de confiance*, *tests statistiques*), on a besoin de connaître la loi de $\bar{X}_n$.
- On rappelle que, dans le cas gaussien,
$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1).$$
- *Interprétation* : $\mathcal{L}(\bar{X}_n) = \mathcal{N}(\mu, \sigma^2/n)$.

La puissance du TCL

- Le *théorème central limite* stipule que, sous des hypothèses très faibles, on peut étendre ce résultat (pour n grand) à "n'importe quelle" suite de variables aléatoires indépendantes.
- C'est l'un des résultats les plus impressionnants et les plus utilisés en probabilités et statistique.

Le TCL

Théorème Central Limite (TCL)

Soit $X_1, \dots, X_n$ un n-échantillon i.i.d. On note $\mathbf{E}[X_i] = \mu$, $\mathbf{V}[X_i] = \sigma^2$ et $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. On a alors
$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \text{quand } n \to \infty.$$

- Les hypothèses sont *faibles* : on demande juste des v.a.r i.i.d. qui admettent une variance.
- Conséquence : si n est suffisamment grand, on pourra approcher la loi de $\bar{X}_n$ par la loi $\mathcal{N}(\mu, \sigma^2/n)$.
- On pourra écrire $\mathcal{L}(\bar{X}_n) \approx \mathcal{N}(\mu, \sigma^2/n)$ mais *pas*
$$\mathcal{L}(\bar{X}_n) \xrightarrow{\mathcal{L}} \mathcal{N}(\mu, \sigma^2/n).$$

TCL pour modèle de Bernoulli

- $X_1, \dots, X_n$ i.i.d. de loi de Bernoulli de paramètre $p \in [0,1]$.
- On a donc $\mathbf{E}[X_1] = p$ et $\mathbf{V}[X_1] = p(1-p)$.

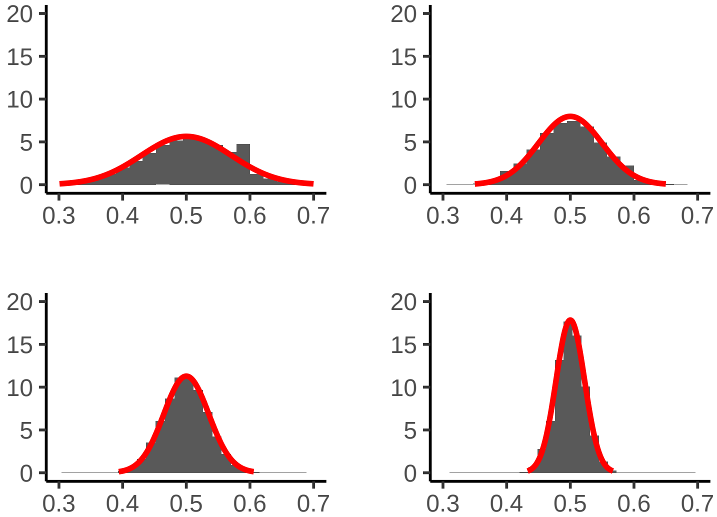
TCL

On a d'après le TCL
$$\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \text{quand } n \to \infty.$$

Conséquence

On peut donc *approcher la loi* de la moyenne empirique $\bar{X}_n$ par la loi
$$\mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

— Approximation TCL pour le modèle de Bernoulli $\mathcal{B}(1/2)$ avec $n = 50, 100, 200, 500$.



Intervalles de confiance

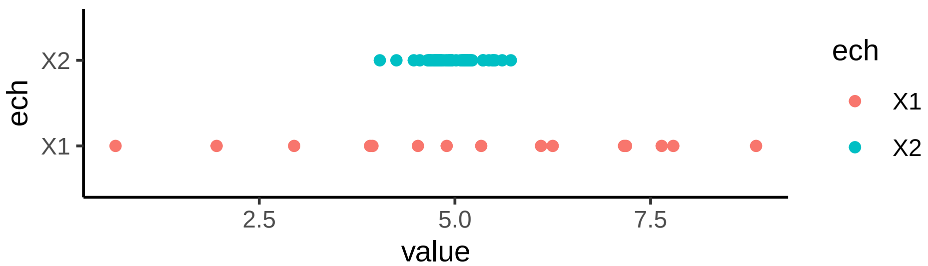
Motivations

— Donner une *seule valeur* pour estimer un paramètre peut se révéler trop ambitieux.

— *Exemple :* la performance est de 72% lorsque on prend le traitement (alors qu'on ne l'a *testé que sur 100 athlètes*).

— Il peut parfois être plus raisonnable de donner une réponse dans le genre, la performance se trouve dans l'*intervalle* $[70\%, 74\%]$ avec une *confiance* de 90%.

Un exemple



Remarque

— Ces deux échantillons sembelent avoir (à peu près) la même moyenne.

— Cependant, l'échantillon 2 semble être plus précis pour estimer cette moyenne.

— $X_1, \ldots, X_n$ un échantillon i.i.d. de loi $\mathbf{P}$ inconnue.

— Soit θ un paramètre inconnu, par exemple $\theta = \mathbf{E}[X]$.

Définition

Soit $\alpha \in]0,1[$. On appelle intervalle de confiance pour θ tout intervalle de la forme $[A_n, B_n]$, où A_n et B_n sont des fonctions telles que :

$$\mathbf{P}(\theta \in [A_n, B_n]) = 1 - \alpha.$$

Définition

Si $\lim_{n\to\infty} \mathbf{P}(\theta \in [A_n, B_n]) = 1 - \alpha$, on dit que $[A_n, B_n]$ est un *intervalle de confiance asymptotique* pour θ au niveau $1 - \alpha$.

Construction d'IC

- Un *intervalle de confiance* pour un paramètre inconnu θ se construit généralement à partir d'un estimateur de θ dont on connait la loi.

- A partir de la loi de $\hat{\theta}$, on cherche deux bornes A_n et B_n telle que

$$\mathbf{P}(\theta \in [A_n, B_n]) = 1 - \alpha.$$

Remarque

A priori, plus α est *petit*, plus l'intervalle aura un *grande amplitude*.

Exemple

- $X_1, \dots, X_n$ i.i.d. de loi normale $\mathcal{N}(\mu, 1)$.
- *Question* : IC de niveau 0.95 pour μ ?

Construction de l'IC

- Estimateur : $\hat{\mu} = \bar{X}_n$.
- Loi de l'estimateur : $\mathcal{L}(\hat{\mu}) = \mathcal{N}(\mu, 1/n)$.
- On déduit

$$\mathbf{P}\left(\hat{\mu} - q_{1-\alpha/2}\frac{1}{\sqrt{n}} \leq \mu \leq \hat{\mu} + q_{1-\alpha/2}\frac{1}{\sqrt{n}}\right) = 1 - \alpha.$$

- Un intervalle de confiance de niveau $1 - \alpha$ est donc donné par

$$\left[\hat{\mu} - q_{1-\alpha/2}\frac{1}{\sqrt{n}}, \hat{\mu} + q_{1-\alpha/2}\frac{1}{\sqrt{n}}\right].$$

Quantiles

- $q_{1-\alpha/2}$ désigne le *quantile d'ordre* $1 - \alpha/2$ de la loi normale $\mathcal{N}(0, 1)$.
- Il est défini par

$$\mathbf{P}\left(X \leq q_{1-\alpha/2}\right) = 1 - \frac{\alpha}{2}.$$

Définition

Plus généralement, le *quantile d'ordre* α d'une variable aléatoire X est défini par le réel q_α vérifiant

$$\mathbf{P}(X \leq q_\alpha) \geq \alpha \quad \text{et} \quad \mathbf{P}(X \geq q_\alpha) \geq 1 - \alpha.$$

- Les quantiles sont généralement renvoyés par les *logiciels statistique* :

```
> c(qnorm(0.975),qnorm(0.95),qnorm(0.5))
## [1] 1.959964 1.644854 0.000000
```

Une exemple à la main

— $n = 50$ observation issues d'une loi $\mathcal{N}(\mu, 1)$:

```
> head(X)
## [1] 3.792934 5.277429 6.084441 2.654302 5.429125 5.506056
```

— *Estimation* de μ :

```
> mean(X)
## [1] 4.546947
```

— *Intervalle de confiance* de niveau 95% :

```
> binf <- mean(X)-qnorm(0.975)*1/sqrt(50)
> bsup <- mean(X)+qnorm(0.975)*1/sqrt(50)
> c(binf,bsup)
## [1] 4.269766 4.824128
```

Loi normale (cas réel)

— $X_1, \dots, X_n$ i.i.d de loi $\mathcal{N}(\mu, \sigma^2)$.
— On a vu qu'un IC pour μ est donné par

$$\left[\hat{\mu} - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Problème

— Dans la vraie vie, σ est inconnu !
— L'intervalle de confiance n'est donc pas calculable.

Idée

1. Estimer σ^2 par

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

2. Et considérer l'IC :

$$\left[\hat{\mu} - q_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}, \hat{\mu} + q_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right]. \quad (1)$$

Problème

— On a bien

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

— mais

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\widehat{\sigma}} \neq \mathcal{N}(0, 1)$$

— Pour avoir la loi de

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\widehat{\sigma}} \neq \mathcal{N}(0,1)$$

avec

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

— il faut définir d'autres *lois de probabilité*.

La loi normale (Rappel)

Définition

— Une v.a.r X suit une loi *normale* de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 > 0$ admet pour densité

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

Propriétés

— $\mathbf{E}[X] = \mu$ et $\mathbf{V}[X] = \sigma^2$.

— Si $X \sim N(\mu, \sigma^2)$ alors

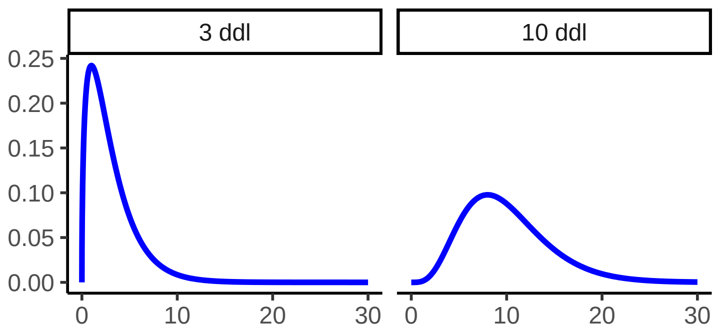
$$\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

Loi du χ^2

Définition

— Soit $X_1, \ldots, X_n$ n variables aléatoires réelles indépendantes de loi $\mathcal{N}(0,1)$. La variable $Y = X_1^2 + \ldots + X_n^2$ suit une loi du *Chi-Deux à n degrés de liberté*. Elle est notée $\chi^2(n)$.

— $\mathbf{E}[Y] = n$ et $\mathbf{V}[Y] = 2n$.



Loi de Student

Définition

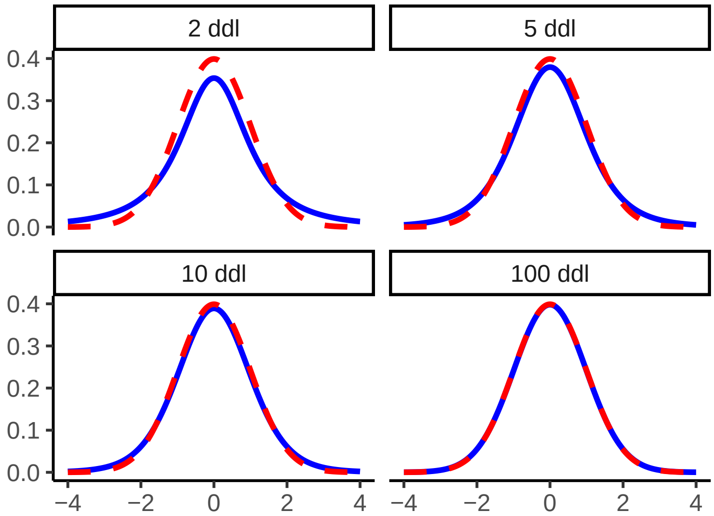
— Soient X et Y deux v.a.r. *indépendantes* de loi $\mathcal{N}(0,1)$ et $\chi^2(n)$. Alors la v.a.r.

$$T = \frac{X}{\sqrt{Y/n}}$$

suit une *loi de student à n degrés de liberté*. On note $\mathcal{T}(n)$.

— $\mathbf{E}[T] = 0$ et $\mathbf{V}[T] = n/(n-2)$.

— Lorsque *n est grand* la loi de student à n degrés de liberté peut être *approchée par la loi* $\mathcal{N}(0,1)$.



Légende

Densités des lois de student à 2, 5, 10 et 100 degrés de liberté (bleu) et densité de la loi $\mathcal{N}(0,1)$ (rouge).

Loi de Fisher

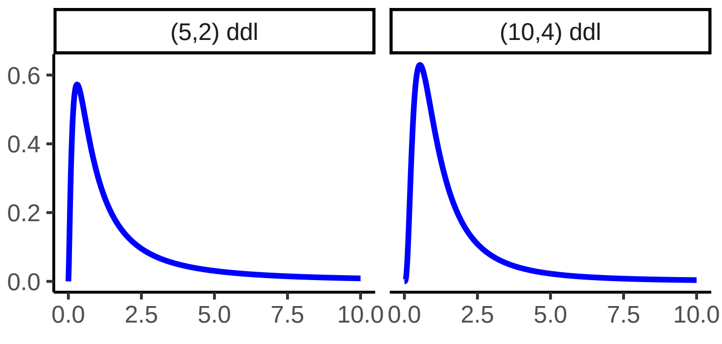
Définition

— Soient X et Y deux v.a.r *indépendantes* de lois $\chi^2(m)$ et $\chi^2(n)$. Alors la v.a.r

$$F = \frac{X/m}{Y/m}$$

suit une *loi de Fisher à m et n degrés de liberté*. On note $\mathcal{F}(m,n)$.

— Si $F \sim \mathcal{F}(m,n)$ alors $1/F \sim \mathcal{F}(n,m)$.



Théorème de Cochran

— $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$.

— On note

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

Théorème de Cochran

On a alors

1. $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$.
2. $\bar{X}_n$ et S^2 sont indépendantes.
3. On déduit
$$\sqrt{n}\frac{\bar{X}_n - \mu}{S} \sim \mathcal{T}(n-1).$$

Remarque

1 et 3 sont très importants pour construire des *intervalles de confiance.*

IC pour la loi gaussienne

IC pour μ

On déduit du résultat précédent qu'un IC de niveau $1-\alpha$ pour μ est donné par

$$\left[\bar{X}_n - t_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X}_n + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right],$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ ddl.

IC pour σ^2

Un IC de niveau $1-\alpha$ pour σ^2 est donné par

$$\left[\frac{(n-1)S^2}{\chi_{1-\alpha/2}}, \frac{(n-1)S^2}{\chi_{\alpha/2}}\right]$$

où $\chi_{1-\alpha/2}$ et $\chi_{\alpha/2}$ sont les quantiles d'ordre $1-\alpha/2$ et $\alpha/2$ de loi $\chi^2(n-1)$.

Exemple (IC pour μ)

— $n = 50$ observation issues d'une loi $\mathcal{N}(\mu, \sigma^2)$:

```
> head(X)
## [1] 3.792934 5.277429 6.084441 2.654302 5.429125 5.506056
```

— *Estimation* de μ :

```
> mean(X)
## [1] 4.546947
```

— *Estimation* de σ^2 :

```
> S <- var(X)
> S
## [1] 0.783302
```

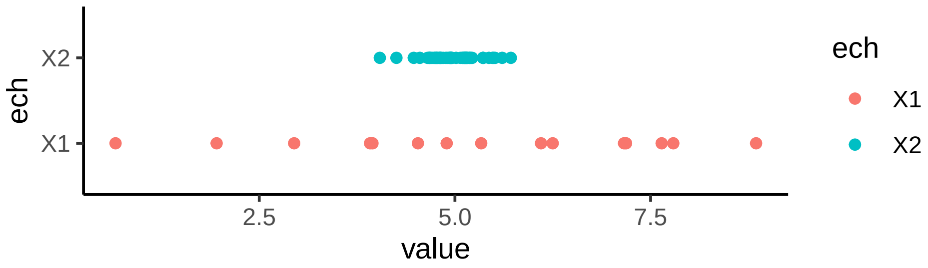
— *Intervalle de confiance* de niveau 95% :

```
> binf <- mean(X)-qt(0.975,49)*sqrt(S)/sqrt(50)
> bsup <- mean(X)+qt(0.975,49)*sqrt(S)/sqrt(50)
> c(binf,bsup)
## [1] 4.295420 4.798474
```

— On peut obtenir directement l'intervalle de confiance à l'aide de la fonction *t.test* :

```
> t.test(X)$conf.int
## [1] 4.295420 4.798474
## attr(,"conf.level")
## [1] 0.95
```

Autre exemple




```
> t.test(df1$value)$conf.int[1:2]
## [1] 3.990982 6.563659
> t.test(df2$value)$conf.int[1:2]
## [1] 4.887045 5.074667
```

Conclusion

Sans surprise, on retrouve bien qu'on est *plus précis* avec l'échantillon 2.

Exemple (IC pour σ^2)

— On obtient l'IC pour σ^2 à l'aide de la formule

$$\left[\frac{(n-1)S^2}{\chi_{1-\alpha/2}}, \frac{(n-1)S^2}{\chi_{\alpha/2}}\right]$$

— On peut donc le calculer sur R :

```
> binf <- 49*S/qchisq(0.975,49)
> bsup <- 49*S/qchisq(0.025,49)
> c(binf,bsup)
## [1] 0.5465748 1.2163492
```

Application décathlon

— IC de niveau 95% pour la longueur moyenne en saut en longueur :

```
> t.test(decathlon$Long.jump)$conf.int
## [1] 7.160131 7.359869
## attr(,"conf.level")
## [1] 0.95
```

— IC de niveau 95% pour la temps moyen au 1500m :

```
> t.test(decathlon$`1500m`)$conf.int
## [1] 275.3403 282.7094
## attr(,"conf.level")
## [1] 0.95
```

— IC de niveau 90% pour la temps moyen au 1500m :

```
> t.test(decathlon$`1500m`,conf.level=0.90)$conf.int
## [1] 275.9551 282.0946
## attr(,"conf.level")
## [1] 0.9
```

Remarque

L'IC à 95% a une amplitude plus grande que celui à 90% (c'est *normal*).

Comparer des moyennes

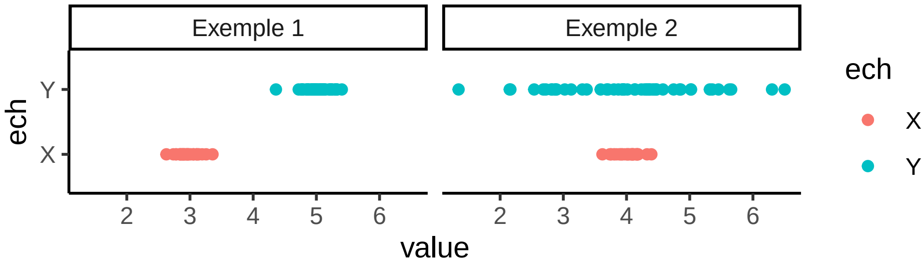
Question (fréquente)

— Peut-on dire que deux populations ont les mêmes catactéristiques ?
— Ou plus simplement que deux caractéristiques ont la même moyenne ?

Observations

— $X_1, \dots, X_{n_1}$ observations pour la population 1.
— $Y_1, \dots, Y_{n_2}$ observations pour la population 2.

Exemple



Idée

Utiliser des IC pour décider.

Comparer des moyennes.

— *Approche* : constuire un IV pour $\mu_X - \mu_Y$ et regarder si 0 est à l'intérieur de l'IC

— *Méthode* : trouver la loi de $\bar{X} - \bar{Y}$.
— *Résultat* : cette loi est proche d'un loi Gaussienne. On peut montrer plus précisément que

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}$$

suit un loi de Student à ν degrés de liberté (ν par de forme explicite pour ν).
— On déduit de ces résultats des *IC* pour $\mu_X - \mu_Y$.

Exemple

— On reprend les deux échantillons des diapos précédentes.

```
> t.test(df1$value,df2$value)
##
##  Welch Two Sample t-test
##
## data:  df1$value and df2$value
## t = -55.526, df = 81.644, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.134079 -1.986443
## sample estimates:
## mean of x mean of y
##  2.965286  5.025547
```

Conclusion

0 n'étant pas dans l'intervalle de confiance, on peut penser que les moyennes sont différentes.

— On reprend les deux échantillons des diapos précédentes.

```
> t.test(df1$value,df2$value)
##
##  Welch Two Sample t-test
##
## data:  df1$value and df2$value
## t = -55.526, df = 81.644, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.134079 -1.986443
## sample estimates:
## mean of x mean of y
##  2.965286  5.025547
```

Conclusion

0 étant dans l'intervalle de confiance, on peut penser que les moyennes sont proches.

— Les procédures de test statistique permettront de préciser cette méthodologie.